

# Reasoning Strategies and Robustness in Language Models: A Cognitive View

Contrasting Reasoning in Humans and Large Language Models On a  
Causal Benchmark

## Master's Thesis

submitted in partial fulfillment of the requirements for the degree of  
*Master of Science in Machine Learning*  
Department of Computer Science  
University of Tübingen

by

**H. M. Dettki**

[hmd8142@nyu.edu](mailto:hmd8142@nyu.edu)

New York, NY  
2025

Department of Computer Science at the University of Tübingen.

Date of Defense:        October 15, 2025

**University of Tübingen:**

**New York University:**

1. Berichterstatter:	Prof. Charley M. Wu	External Advisor:	Prof. Brenden M. Lake (now Princeton)
2. Berichterstatter:	Prof. Jakob Macke	Collaborator:	Prof. Bob Rehder

# Abstract

Large Language Models (LLMs) have made significant strides in natural language processing and exhibit impressive capabilities in natural language understanding and generation. However, *how* they reason, and to what degree they *align with human reasoning* remains underexplored. To this end, we evaluate over 20 LLMs on 11 causal reasoning tasks formalized by a collider graph and compare their performance to that of humans on the same tasks. We also evaluate to what extent causal reasoning depends on knowledge about common causal relationships in the natural world, on the degree of irrelevant information in the prompt, and on the reasoning budget.

We find that most LLMs are aligned with human reasoning up to ceiling effects with chain of thought increasing alignment for LLMs that were misaligned under single-shot prompting. We further find that most LLMs reason in a more deterministic, rule-like way than humans. Chain-of-thought prompting increases reasoning consistency and, under noisy prompts, also pushes some LLMs' reasoning from a probabilistic to a more deterministic regime. Reasoning is robust to replacing real-world content with abstract placeholders but degrades when prompts are injected with irrelevant text; in these cases, chain-of-thought recovers much of the lost performance. Across experiments, many models exceed human baselines on qualitative reasoning signatures such as explaining-away and Markov compliance, which humans typically exhibit only weakly or violate.

Together, we find a divergence in reasoning style: humans rely more on probabilistic judgments, whereas many LLMs default to near-deterministic rules. Such determinism can enhance reliability and augment human reasoning by providing stable, rule-like outputs. Yet it also risks failure in real-world settings where uncertainty is intrinsic, underscoring the need to better characterize LLM reasoning strategies to guide their safe and effective application.



# Acknowledgements

I would like to express my deepest gratitude to my advisors, Prof. Charley Wu, Prof. Brenden Lake, and Prof. Bob Rehder for their invaluable guidance, support, and encouragement throughout my research journey. Their expertise and insights have been instrumental in shaping this thesis.

I would like to thank Bob for the human data and Brenden for letting me use his API keys making it possible to run all the experiments with all the LLMs. I would like to thank the Department of Psychology at NYU for providing 1000\$ in computing credits to run the experiments.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>List of Figures</b>	<b>iv</b>
<b>List of Tables</b>	<b>vi</b>
<b>Notation</b>	<b>vii</b>
1 Large Language Models . . . . .	vii
2 Basic Mathematical Objects and Abbreviations . . . . .	vii
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 This Thesis . . . . .	2
1.2.1 Contributions . . . . .	3
1.2.2 Code and Reproducibility . . . . .	3
1.3 Related Work . . . . .	4
<b>2 Background</b>	<b>7</b>
2.1 Causal Bayesian Networks (CBNs) . . . . .	7
2.2 A Simple, Interpretable Model of Causal Reasoning . . . . .	8
2.3 Collider Graph-specific Reasoning Signatures . . . . .	8
2.3.1 The Effect is More Likely if More Causes are Present . . . . .	8
2.3.2 Explaining Away (EA) . . . . .	8
2.3.3 Markov Violation (MV) . . . . .	9
2.4 Why Collider Graphs? . . . . .	9
2.5 Large Language Models (LLMs) . . . . .	9
<b>3 Causal Reasoning Benchmark</b>	<b>11</b>
3.1 Causal Inference Tasks from Rehder and Waldmann [1] . . . . .	11
3.1.1 Causal Inference Tasks. . . . .	11
3.1.2 Cover Stories and Knowledge Domains . . . . .	12
3.1.3 Experimental Protocol for Humans . . . . .	13
3.2 A Causal Reasoning Benchmark for LLMs with A Human Baseline . . . . .	13

## Contents

3.2.1	Experimental Protocol for LLMs . . . . .	13
3.2.2	Prompt and Content Manipulations . . . . .	14
3.2.3	Software Package . . . . .	15
<b>4</b>	<b>Analysis and Experimental Results</b>	<b>17</b>
4.1	Q1: Do Agents Reason Differently Across Domains? . . . . .	19
4.2	Q2: Are Humans and LLMs Aligned? . . . . .	21
4.3	Q3: Do Humans and LLMs Reason Normatively? . . . . .	22
4.3.1	Operational Definition of Normativity . . . . .	22
4.3.2	Fitting Causal Bayesian Networks to Likelihood Judgments . . . . .	23
4.3.3	Most Agents Are Described Well By A Causal Bayesian Network . . . . .	24
4.4	Q4: Reasoning Consistency Across Experiment-Prompt Conditions . . . . .	27
4.4.1	Reasoning Consistency . . . . .	27
4.4.2	Chain-of-Thought improves reasoning consistency and helps mitigate the impact of distracting information. . . . .	28
4.5	Q5: What Kind Of Cognitive Strategies Do Agents Use? . . . . .	28
4.5.1	Probabilistic vs. Deterministic Reasoning: Leak-Adjusted Determinacy . . . . .	28
4.5.2	Most LLMs reason more deterministically than humans, some reason more probabilistically than humans. . . . .	29
4.5.3	Qualitative Measures of Reasoning: Explaining Away and Markov Compliance . . . . .	30
4.5.4	Explaining Away and Markov Compliance in LLMs . . . . .	30
4.6	Q6: Do LLMs Reason Robustly Under Content Manipulations? . . . . .	33
4.6.1	Robustness Across Prompt and Content Manipulations . . . . .	33
4.6.2	Findings . . . . .	34
<b>5</b>	<b>Conclusion</b>	<b>39</b>
5.1	Summary . . . . .	39
5.2	Limitations . . . . .	40
5.3	Future Work . . . . .	40
	<b>Bibliography</b>	<b>43</b>
<b>A</b>	<b>Analysis – Details</b>	<b>47</b>
A.1	Overloaded / Noisy Prompt Generation and Prompt Variants . . . . .	47
A.1.1	Overloaded and Abstract Prompts . . . . .	47
A.1.2	Abstract prompts (contrast) . . . . .	51
A.1.3	Prompt-categories (LLM-Output Instructions Chain-of-Thought vs. numeric) . . . . .	52
A.2	Human-LLM Alignment Details . . . . .	53
A.3	Derivation of Noisy Or Model’s Predicted Probability for each Task I-XI . . . . .	53
A.3.1	Notation . . . . .	53
A.3.2	Predictive Inference . . . . .	54
A.3.3	Independence of Causes . . . . .	54



A.3.4	Diagnostic Inference – Effect Present . . . . .	55
A.3.5	Diagnostic Inference – Effect Absent . . . . .	56
A.4	LLM Details . . . . .	56
<b>B</b>	<b>Additional Results</b>	<b>59</b>
B.1	Domain differences per experiment and prompt-style . . . . .	60
B.1.1	RW17 . . . . .	60
B.2	Human-LLM alignment: Domain-wise breakdowns for Chain-of-Thought prompts in comparison to Numeric prompts . . . . .	65
B.3	Additional Results for the distribution of likelihood judgements . . . . .	67
B.4	Additional Results for the effect of Chain-of-Thought prompts on causal reason- ing across experiments . . . . .	69
B.5	Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explainaing Away, and Markov Violation . . . . .	73
B.5.1	Metrics by Release Date of LLMs . . . . .	81
B.6	Most and Least changing LLMs across prompt-category & content manipulations	83
B.6.1	Experiment-wise changes with fixed prompt-style (Numeric or CoT) . .	83
B.6.2	Prompt-wise changes with fixed experiment (e.g., RW17 or Abstract) .	89
B.7	Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments . . . . .	95
B.7.1	RW17 and Abstract, Numeric Prompts . . . . .	95
B.7.2	RW17 and Abstract, CoT Prompts . . . . .	98
B.7.3	RW17-Overloaded, Numeric and CoT Prompts . . . . .	101
B.7.4	Abstract-Overloaded, Numeric and CoT Prompts . . . . .	103
<b>C</b>	<b>Declaration of Generative AI Usage</b>	<b>105</b>

# List of Figures

2.1	Collider graph encoding $p(C_1 = 1 \mid C_2 = 0)$ (task V in Figure 4.2(c)). . . . .	7
3.1	Dataset Illustration . . . . .	12
4.1	Human-LLM alignment . . . . .	22
4.2	Agents vs. causal Bayes net predictions . . . . .	24
4.3	Normativity – Causal Bayes net error metrics: loss . . . . .	25
4.4	Normativity – Causal Bayes net error metrics: RMSE . . . . .	26
4.5	Normativity – Causal Bayes net error metrics: MAE . . . . .	26
4.6	Reasoning Consistency Across conditions (LOOCV $R^2$ ) . . . . .	27
4.7	By release date: Leak-Adjusted Determinacy (LAD) . . . . .	29
4.8	Collider-induced metrics for RW17: explaining away and Markov compliance levels . . . . .	32
4.9	Leak-Adjusted Determinacy (LAD) levels . . . . .	35
4.10	Collider and CBN-induced metrics for RW17 . . . . .	36
B.1	Human-LLM alignment with domain breakdowns and CoT effects . . . . .	66
B.2	Distribution of Likelihood Judgements by Domain: empirical cumulative distribution (ECDF) . . . . .	68
B.3	Explaining Away (EA) levels by agent and experiment and prompt-category . .	70
B.4	Markov Violation (MV) levels by agent and experiment and prompt-category .	71
B.5	Reasoning consistency levels (LOOCV- $R^2$ ) by agent and experiment and prompt-category . . . . .	72
B.6	Leak-Adjusted Determinacy (LAD) levels Claude 1 . . . . .	74
B.7	Leak-Adjusted Determinacy (LAD) levels Claude 2 . . . . .	75
B.8	Leak-Adjusted Determinacy (LAD) levels Gemini . . . . .	76
B.9	Leak-Adjusted Determinacy (LAD) levels GPT mix . . . . .	77
B.10	Leak-Adjusted Determinacy (LAD) levels GPT-5-mini . . . . .	78
B.11	Leak-Adjusted Determinacy (LAD) levels GPT-5-nano . . . . .	79
B.12	Leak-Adjusted Determinacy (LAD) levels OpenAI Reasoning Models . . . . .	80
B.13	By release date: Explaining Away (EA) . . . . .	81
B.14	By LLM release date: Markov Violation (MV) . . . . .	82
B.15	By LLM release date: LOOCV $R^2$ . . . . .	82
B.16	Pairwise experiment-wise comparisons (RW17 $\rightarrow$ Abstract), fixed prompt-category	84

## List of Figures

B.17	Pairwise experiment-wise comparisons (RW17 $\rightarrow$ Abstract), fixed prompt-category	85
B.18	Pairwise experiment-wise comparisons (RW17 $\rightarrow$ Over RW17), Numeric . . .	86
B.19	Pairwise experiment-wise comparisons (RW17 $\rightarrow$ Over RW17), CoT . . . . .	87
B.20	Pairwise prompt-category comparisons, RW17 . . . . .	90
B.21	Pairwise prompt-category comparisons, Abstract . . . . .	91
B.22	Pairwise prompt-category comparisons, RW17 Overloaded . . . . .	92
B.23	Pairwise prompt-category comparisons, Abstract Overloaded . . . . .	93
B.24	Parameter Values of best fitting causal Bayes Nets (CBN) for Numeric Prompts in RW17 and Abstract experiments . . . . .	95
B.25	Parameter Values of best fitting causal Bayes Nets (CBN) for chaing-of thought prompts in RW17 and Abstract experiments . . . . .	98
B.26	Parameter Values of best fitting causal Bayes Nets (CBN) for chaing-of thought prompts in RW17 and Abstract experiments . . . . .	101
B.27	Parameter Values of best fitting causal Bayes Nets (CBN) for chaing-of thought prompts in abstract and Abstract experiments . . . . .	104

# List of Tables

4.1	Agent-domain differences tested by Kruskal-Wallis test . . . . .	20
A.1	LLM Details: Release Date & Context Window Size . . . . .	57
B.1	Kruskal–Wallis across agents within each domain, RW17-Numeric prompts. . .	60
B.2	Kruskal–Wallis across agents within each domain. (RW17, Numeric prompts) .	60
B.3	Kruskal–Wallis across domains within each agent. (RW17, Numeric prompts) .	61
B.4	Kruskal–Wallis across agents within each domain. (RW17, CoT prompts) . . .	61
B.5	Kruskal–Wallis across domains within each agent. (RW17, Numeric prompts) .	62
B.6	Kruskal–Wallis across agents within each domain. RW17-overloaded, Numeric	62
B.7	Kruskal–Wallis across domains within each agent. RW17-overloaded, Numeric	63
B.8	Kruskal–Wallis across domains within each agent. RW17-overloaded, CoT . .	63
B.9	Kruskal–Wallis across agents within each domain. RW17-overloaded, CoT . .	63
B.10	Kruskal–Wallis across agents within each domain. RW17-overloaded, CoT . .	64
B.11	Kruskal–Wallis across agents within each domain. RW17-overloaded, Numeric	64
B.12	Kruskal–Wallis across domains within each agent, Abstract Overloaded, Numeric.	64
B.13	Kruskal–Wallis across agents within each domain. RW17-overloaded, CoT . .	64
B.14	Kruskal–Wallis across domains within each agent. RW17-overloaded, CoT . .	65
B.15	Human–LLM alignment, single-shot/numeric prompting . . . . .	67
B.16	Human–LLM alignment, chain-of-thought prompting . . . . .	69
B.17	Top 3 most and least changing LLMs by prompt-content . . . . .	88
B.18	Top 3 most and least changing LLMs by prompt-category . . . . .	94
B.19	CBN-Fit metrics for RW17 numeric . . . . .	96
B.20	CBN-Fit metrics for Abstract plain Numeric . . . . .	97
B.21	CBN-Fit metrics for Abstract CoT . . . . .	99
B.22	CBN-Fit metrics for Abstract-Plain CoT . . . . .	100
B.23	CBN-Fit metrics for RW17-overloaded Numeric . . . . .	102
B.24	CBN-Fit metrics for RW17-overloaded CoT . . . . .	103
B.25	CBN-Fit metrics for abstract-overloaded numeric . . . . .	104
C.1	Generative AI programs and their version numbers used in this work. . . . .	105

# Notation

## 1 Large Language Models

### OpenAI GPT Models

OpenAI’s GPT models follow a naming convention that indicates the model version, and size (e.g., `gpt-4.1-mini`).

**GPT-5 family naming scheme** The GPT-5 family uses different runtime control fields that replace the conventional “temperature” parameter from its predecessors.

We introduce a compact, self-descriptive model name that encodes the model variant together with its two unique runtime control fields.

The canonical pattern is:

`gpt-5-<variant>-v-<verbosity>-r-<effort>`

Where: - `<variant>` denotes the specific GPT-5 variant (e.g., “`gpt-5-nano`”). - `v` / `verbosity` controls the length/detail of the model’s output. Accepted values: `low`, `medium`, `high`. - `r` / `reasoning_effort` controls the model’s internal reasoning effort. Accepted values: `minimal`, `low`, `medium`, `high`.

**OpenAI’s Reasoning Models** Reasoning models follow a different naming convention, starting with an “o” followed by their version and size (e.g., `o1`, `o3-mini`).

### Claude Models

Claude models end with a date stamp indicating a specific snapshot of the model. `claude-sonnet-4-20250514` identifies the Sonnet-4 variant from May 14, 2025.

## 2 Basic Mathematical Objects and Abbreviations

## Notation

### Probability Theory

$P(\cdot)$	Probability of an event
$P(\cdot   \cdot)$	Conditional probability of an event
$\mathbb{E}(\cdot)$	Expectation of a random variable
$\text{Cov}(\cdot, \cdot)$	Covariance between two random variables
$\text{Var}(\cdot)$	Variance of a random variable

Distributions and densities.

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian probability density function

### Machine Learning

$n$	Number of training data
$\mathbf{X}$	Training data inputs
$\mathbf{y}$	Training data targets
$\theta$	Parameters of a model
$\ell$	Loss function

### Acronyms & Abbreviations

AI	Artificial intelligence
LLM	Large language model
API	Application programming interface
BR	Bayes' rule
PR	Product rule
IA	Independence assumption
M	Marginalization
LLM	Large language model
CoT	Chain of Thought
CBN	Causal Bayesian network
GPU	Graphics processing unit
i.i.d.	independent and identically distributed
MD	Model Definition
MAP	Maximum a posteriori estimation
MAE	Mean absolute error
L-BFGS	Limited memory BFGS, a type of quasi-Newton method
RMSE	Root mean square error
RW17	Rehder and Waldmann [1] tasks
RW17 Over / RW17 Overloaded	Overloaded version of the Rehder and Waldmann [1] tasks

## 2 Basic Mathematical Objects and Abbreviations

SOTA	State of the art
Abstract Over / Abstract Overloaded	Abstract overloaded version of the Rehder and Waldmann <a href="#">[1]</a> tasks





# Introduction

---

1.1	Motivation . . . . .	1
1.2	This Thesis . . . . .	2
1.2.1	Contributions . . . . .	3
1.2.2	Code and Reproducibility . . . . .	3
1.3	Related Work . . . . .	4

---

## 1.1 Motivation

**Courtroom 4.** The judge’s stomach growls—it’s been hours since their last meal. This is the 34<sup>th</sup> case since the court opened at 8:00 a.m., and the third borderline one in a row. Eyes gritty from decision fatigue, she defaults to the status quo when uncertain. She ticks the box for continued detention. “Next case.” Recess in six minutes.

**Primary Care Clinic.** Nine hours into her shift, the physician opens the chart of the 39<sup>th</sup> patient today. Congestion, fever, clear lungs. The guideline says watchful waiting, but fatigue clouds judgment. The “just-in-case” antibiotic order goes in with two clicks. “Next patient.”

*Same inputs, varying outputs: human decision-making can be severely impaired by hunger, fatigue and cognitive depletion—metabolic constraints that everybody is subject to.*

These examples illustrate well-documented effects of how metabolic states such as fatigue and hunger impact human judgment and decision making. Judges’ leniency declines as sessions progress [2], while inappropriate antibiotic prescriptions rise hour-by-hour in medical settings

[3]. Sleep loss shifts risk sensitivity, mild dehydration impairs working memory, and decision fatigue systematically biases choices toward default options [4–6].

Such findings motivate AI-assisted decision-making. In principle, computational systems could provide consistent, fatigue-resistant analysis to complement human judgment. In fact, Large Language Models (LLMs) are increasingly deployed for a wide variety of decision-making tasks in high-stakes scenarios – for example in the court system [7] and in medical institutions [8]. However, successful deployment requires understanding whether AI systems exhibit genuine reasoning capabilities or rely on sophisticated but brittle pattern matching [9]. This distinction becomes critical in high-stakes domains where genuine causal understanding is required and purely associative pattern matching is not sufficient for good decision-making.

### 1.2 This Thesis

Human decision-making suffers from systematic biases and metabolic constraints, motivating AI assistance. However, deploying AI systems requires understanding *how* they reason about cause and effect, not just whether they produce correct answers. To that end, we need diagnostic tools that distinguish genuine causal understanding from associative reasoning.

In this work, we introduce a causal reasoning benchmark with paired human data and a cognitively-grounded framework to evaluate normative reasoning. This framework is built on simple causal Bayesian networks with interpretable parameters which expose reasoning strategies: normative reasoning is defined as the existence of a set of parameters in a causal Bayes net that closely fits an agent’s judgments, while associative shortcuts occur when an agent relies on superficial correlations rather than underlying causal reasoning. We identify reasoning signatures that Large Language Models (LLMs) seem to operate by and compare them to human reasoning patterns on the exact same tasks. This framework helps understand to what degree agents are able to solve causal reasoning tasks, for example revealing where on the spectrum between deterministic and probabilistic reasoning LLMs fall, whether they are robust to content manipulations, and *how alike they are to human reasoning and biases*.

**Research Questions** More specifically, this thesis aims to empirically address the following research questions:

- Q1 Domain differences.** Do agents reason differently across knowledge domains (economics, meteorology, sociology)?
- Q2 Human-LLM alignment.** How do LLM likelihood judgments correlate with human judgments on matched causal inference tasks and does chain-of-thought prompting affect alignment?
- Q3 Normative reasoning.** Are LLMs well-described by causal Bayes nets and consequently can be said to reason normatively?
- Q4 Reasoning consistency.** Do LLMs generalize across related causal tasks (measured via cross-validated performance)?

**Q5 Cognitive strategies.** What parameter signatures distinguish different reasoning approaches — deterministic vs. probabilistic?

**Q6 Robustness and genuineness.** Does performance transfer across content manipulations (semantic vs. abstract prompts, clean vs. overloaded contexts), indicating reasoning beyond surface-level pattern matching?

### 1.2.1 Contributions

This work makes both methodological and empirical contributions toward measuring causal reasoning in language models under different manipulations and toward understanding the alignment with human and normative reasoning.

- **Causal reasoning benchmark.** We introduce a causal reasoning benchmark with paired human data centered on common-effect relationships, for which humans are known to exhibit cognitive biases.
- **Diagnostic framework.** We propose a causal Bayesian network–based evaluation framework that produces interpretable parameter profiles and consistency measures, enabling systematic distinction between normative and associative reasoning strategies in LLMs.
- **Human–LLM alignment.** We demonstrate that state-of-the-art models reach ceiling-level alignment with human judgments (Spearman  $\rho \approx 0.85$ ), the apparent upper bound given human response variability.
- **Reasoning regimes.** We identify two characteristic modes of inference: deterministic, rule-like reasoning often occurring in frontier models, and probabilistic, association-driven reasoning frequently present in smaller or older models.
- **Conditional prompting effects.** We show that chain-of-thought prompting primarily stabilizes reasoning consistency; under noisy prompts it additionally shifts models toward more normative causal patterns, with greatest benefits for weaker models.
- **Robustness to content manipulations.** We find that reasoning is largely preserved under semantic abstraction, but degrades under overload from irrelevant information—losses that chain-of-thought can recover for many models.

Overall, we demonstrate that frontier language models can reason causally about common-effect relationships, even under prompt manipulations, and that they are largely aligned with human reasoning, at times surpassing it.

### 1.2.2 Code and Reproducibility

As new LLMs are released at a rapid pace, it is crucial to have open-source tools that allow researchers to test new models on established benchmarks and create new ones. As part of this thesis we provide [CAUSAIGN](#), a software package that was developed alongside this thesis and allows to a) compare new LLMs with humans on our established benchmarks, b) fit causal Bayes nets to their responses, and c) easily create new collider based causal inference tasks with custom

content manipulations that are directly comparable to our existing collider reasoning benchmarks and human baseline.



<https://github.com/hmd101/causAlign>

All code and data to reproduce the results in this thesis are publicly available in the GitHub repository above.

### 1.3 Related Work

**Cognitive Interpretability of AI Systems** This work contributes to the emerging field of *Cognitive Interpretability*—understanding not just whether AI systems succeed at tasks, but *how* they succeed [10]. We combine *behavioral accounts* that compare human and LLM reasoning patterns with *processing accounts* that give rise to potential latent computational strategies underlying their behavior. Our study provides one of the few direct human-LLM comparisons on matched cognitive tasks, joining recent work on social reasoning [11], logical inference [12], causal strength judgments [13] and extending Dettki et al. [14]. We extend this line by introducing a formal framework for inferring reasoning strategies from compact model fits—revealing whether models employ normative computation, associative shortcuts, or hybrid approaches across different conditions.

**Evidence For and Against Causal Reasoning in LLMs** While Dettki et al. [14] have found some evidence of causal reasoning in collider graphs and alignment with human-like reasoning, Willig et al. [9] have argued that LLMs lack genuine causal understanding beyond pattern matching, describing LLMs as causal parrots. They argue that LLMs are not causal and their apparent successes arise from correlations among textual causal facts encoded in a proposed *meta*-structural causal models (SCM). These two studies represent complementary perspectives on LLM causal reasoning rather than contradictory ones. Dettki et al. [14] conducts a targeted *human-LLM* comparison on collider graphs, a simpler causal structure than those studied by Willig et al. [9], but a well studied one in humans, making it particularly suitable for human-LLM comparison. In this work we build on Dettki et al. [14] by expanding the set of models tested, introducing a measure of reasoning consistency and a more cognitively grounded analysis. We also investigate robustness to content manipulations inspired by Mirzadeh et al. [15] and Jin et al. [16] as discussed below.

**Robustness of LLMs to Content Manipulations** Shi et al. [17] and Mirzadeh et al. [15] demonstrated that introducing irrelevant context can drastically alter the outputs of LLMs. Jin et al. [16] introduced a systematic benchmark for causal reasoning in LLMs, by evaluating performance on a plethora of reasoning tasks embedded in a variety of causal graph topologies with content manipulations, replacing real world scenarios with nonsensical abstract ones. However, they did

### 1.3 Related Work

not include direct human comparison data and their tasks would require college level introductory Math classes. We build on this line of work by evaluating whether LLMs maintain normative causal reasoning patterns under content manipulations, replacing real-world scenarios with abstract placeholders and akin to Mirzadeh et al. [15], we dilute the signal-to-noise ratio by injecting irrelevant information.



# Background

2.1	Causal Bayesian Networks (CBNs)	7
2.2	A Simple, Interpretable Model of Causal Reasoning	8
2.3	Collider Graph-specific Reasoning Signatures	8
2.3.1	The Effect is More Likely if More Causes are Present	8
2.3.2	Explaining Away (EA)	8
2.3.3	Markov Violation (MV)	9
2.4	Why Collider Graphs?	9
2.5	Large Language Models (LLMs)	9

## 2.1 Causal Bayesian Networks (CBNs)

Pearl’s framework of Causal Bayesian Networks (CBNs) provides the foundational formalism for normative causal reasoning [18]. Causal Bayes nets encode causal relationships through directed acyclic graphs where nodes represent random variables and edges represent direct causal influences. In this thesis, we present agents with a simple induced causal structure with two binary causes  $C_1, C_2 \in \{0, 1\}$  and one binary effect  $E \in \{0, 1\}$  and model the agents’ causal judgments with a  $(C_1 \rightarrow E \leftarrow C_2)$  collider graph-induced causal structure with a causal Bayes net. Numerous studies have shown that causal Bayes nets (CBNs) provide a good model of human causal reasoning [19–21]. An important question we ask in this study is whether causal judgements by Large Language Models (LLMs) can also be well-modeled by causal Bayes nets.

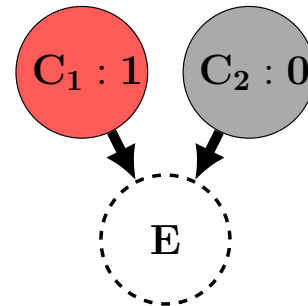


Figure 2.1: Collider graph encoding  $p(C_1 = 1 \mid C_2 = 0)$  (task V in Figure 4.2(c)).

## 2.2 A Simple, Interpretable Model of Causal Reasoning

We adopt a *leaky noisy-OR* link to separate background propensity from causal strengths. For causes  $C_1, C_2 \in \{0, 1\}$  and effect  $E$ ,

$$\Pr(E=1 \mid C_1, C_2) = 1 - (1 - b) (1 - m_1)^{C_1} (1 - m_2)^{C_2}, \quad (2.1)$$

with leak / background propensity  $b \in [0, 1]$  and causal strengths  $m_1, m_2 \in [0, 1]$ . We also fit priors and enforce them to be symmetric  $p(C_1) = p(C_2)$  and consider a three and four parameter tying scheme ( $m_1=m_2$  vs. free). Parameters live in  $[0, 1]$  and support cognitive interpretations: low  $b$ /high  $m \Rightarrow$  deterministic regimes; elevated  $b$ /attenuated  $m \Rightarrow$  probabilistic regimes.

As opposed to Dettki et al. [14], who use a logistic link, we here use a leaky noisy-OR link function for  $E$  for the following reasons: Relative to a logistic link, the noisy-OR affords (i) cognitive grounding via Cheng’s causal power theory [22], and (ii) interpretable parameters separating *background propensity* ( $b$ ) from *causal sensitivity* ( $m_i$ ), while scaling naturally to multiple parents [23–25].

## 2.3 Collider Graph-specific Reasoning Signatures

We follow the RW17 collider family (eleven tasks I–XI) spanning predictive inference, independence checks, and diagnostic inference. Three key signatures of collider / common-effect graph reasoning are discussed next.<sup>1</sup>

### 2.3.1 The Effect is More Likely if More Causes are Present

Numerous studies have shown that humans know that in a collider structure, the effect is more likely to occur if more causes are present [20, 26–28]. Dettki et al. [14]<sup>2</sup> show that LLMs also demonstrate this basic property of collider-based reasoning (see Figure 4.2(b) for a subset of LLMs and humans).

### 2.3.2 Explaining Away (EA)

Explaining away (EA) is a key signature of collider/common-effect graph reasoning. Evidence for one cause reduces belief in the alternative cause in the presence of the effect:

$$\text{EA iff } \Pr(C_1=1 \mid E=1) - \Pr(C_1=1 \mid E=1, C_2=1) > 0.$$

<sup>1</sup>From here on forward, we use the terms collider and common-effect graph interchangeably.

<sup>2</sup>An earlier version of this work was published at the 47th Annual Meeting of the Cognitive Science Society in San Francisco [14].



## 2.4 Why Collider Graphs?

**Example for Explaining Away** Consider the well-known example of a burglar alarm [29]. Suppose the alarm can be triggered ( $E$ ) either by a burglary ( $C_2$ ) or by an earthquake ( $C_1$ ). If we hear the alarm ( $E = 1$ ), both  $C_2$  and  $C_1$  become more likely. However, once we learn that an earthquake ( $C_1 = 1$ ) has in fact occurred, the probability of a burglary decreases: the earthquake *explains away* the burglary. Formally,

$$P(C_2 = 1 \mid E = 1) > P(C_2 = 1 \mid E = 1, C_1 = 1).$$

Visually, explaining away is represented by a positive slope in Figure 4.2(d).

### 2.3.3 Markov Violation (MV)

Independence of causes in a collider means that the state of one cause should not inform belief in the other cause in light of the absence evidence about the effect:

$$\text{MV small iff } |\Pr(C_1=1 \mid C_2=1) - \Pr(C_1=1 \mid C_2=0)| \approx 0.$$

We compute EA/MV on normalized raw likelihood judgments (not model predictions).

**Example for Markov condition** Using again one of the classic examples by Judea Pearl [25]: If an automatic sprinkler runs on a fixed timer, Rain ( $R$ ) and Sprinkler ( $S$ ) are independent causes of Wet grass ( $W$ ). Let  $R$  (rain) and  $S$  (sprinkler) be causes of wet grass  $W$  with  $R \rightarrow W \leftarrow S$ . Under independence of causes,  $\Pr(S=1 \mid R=1) = \Pr(S=1 \mid R=0)$  and  $\text{MV} = |\Pr(S=1 \mid R=1) - \Pr(S=1 \mid R=0)| \approx 0$ . When the sprinkler is suppressed when it rains, this yields a Markov violation  $\text{MV} > 0$  (negative association).

In our tasks, we always state that causes independently generate the effect, so agents should show  $\text{MV} \approx 0$ . Visually, Markov violation is represented by a flat slope in Figure 4.2(c).

## 2.4 Why Collider Graphs?

**Humans Show Weak Explaining Away and Markov Violations** It has been repeatedly observed that humans show *too little* or no explaining away [19, 30]. Humans who are asked to judge the likelihood of one cause have repeatedly shown to be influenced by the absence or presence of an alternative cause [1, 31–33] (Markov violation). Whether LLMs show similar human biases is an open question we address in this thesis (see Section 4.5).

## 2.5 Large Language Models (LLMs)

Large language models (LLMs) are neural networks that model the probability of text sequences at the token level. Concretely, given tokens  $x_{1:n}$ , an autoregressive LLM factorizes the sequence likelihood as

$$p_\theta(x_{1:n}) = \prod_{t=1}^n p_\theta(x_t \mid x_{<t}), \quad (2.2)$$

## Chapter 2 Background

and is trained to minimize next-token prediction loss

$$\mathcal{L}(\theta) = - \sum_{t=1}^n \log p_{\theta}(x_t \mid x_{<t}). \quad (2.3)$$

Modern LLMs are built on the Transformer architecture [34], which replaces recurrence with *self-attention*. For queries  $Q$ , keys  $K$ , and values  $V$ , a single attention head computes

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V, \quad (2.4)$$

with a *look-ahead mask*<sup>3</sup> ensuring each position attends only to past tokens in generation. Stacks of multi-head attention and feed-forward layers, plus residual connections and layer normalization, yield scalable sequence models.

Although masked-language models (e.g., BERT; [35]) are widely used for understanding tasks, this thesis focuses on *autoregressive* LLMs because they are the standard models such as GPT, Gemini and Claude models used by practitioners. Two empirical regularities contextualize capabilities: (i) scaling laws that relate performance to model size, data, and compute [36], and (ii) compute-optimal training recommending more tokens per parameter than previously typical [37]. In practice, base models are adapted via prompting (in-context learning) [38], instruction tuning and reinforcement learning from human feedback (RLHF) [39], and sometimes external retrieval to ground outputs in evidence [40].

---

<sup>3</sup>Also known as masked-self-attention or a causal mask, it prevents attending to future tokens.

# Causal Reasoning Benchmark

---

3.1	Causal Inference Tasks from Rehder and Waldmann [1] . . . . .	11
3.1.1	Causal Inference Tasks. . . . .	11
3.1.2	Cover Stories and Knowledge Domains . . . . .	12
3.1.3	Experimental Protocol for Humans . . . . .	13
3.2	A Causal Reasoning Benchmark for LLMs with A Human Baseline . . . . .	13
3.2.1	Experimental Protocol for LLMs . . . . .	13
3.2.2	Prompt and Content Manipulations . . . . .	14
3.2.3	Software Package . . . . .	15

---

This chapter details our experimental methods for collecting and analyzing causal inference judgments from both humans and large language models (LLMs) on causal reasoning tasks embedded in a collider graph structure. We describe the dataset we compiled based on existing human data comprised of causal inference tasks, how those tasks were verbalized into prompts suitable for LLMs and how the data was collected from both humans and LLMs.

## 3.1 Causal Inference Tasks from Rehder and Waldmann [1]

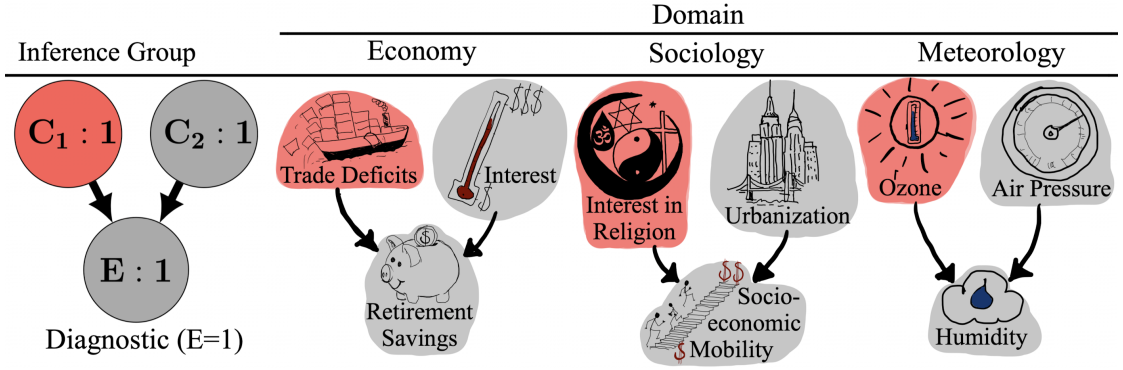
We use human data from Rehder and Waldmann [1] (Experiment 1, Model-Only condition,  $N = 48$  undergraduate students at NYU), who collected likelihood judgments of causal inference tasks verbalizing a collider graph  $C_1 \rightarrow E \leftarrow C_2$ , and compare these to LLM judgments on the same tasks.<sup>1</sup>

### 3.1.1 Causal Inference Tasks.

The collider structure was instantiated in 11 different causal inference tasks (I-XI) grouped into four diagnostic groups (see [Figure 4.2](#) for an overview). Each task differed in which nodes

---

<sup>1</sup>Subsequently, we will refer to the original dataset from Rehder and Waldmann [1] as RW17.



**Figure 3.1: Dataset Illustration of RW17.** The left most graph represents task X from the diagnostic inference group. The nodes are colored according to: ● → latent (query node); ● → observed  $\in \{0, 1\}$ .

were observed and which node was queried ●<sup>2</sup> for which both humans and LLMs were asked to provide a likelihood judgment on a continuous scale (0-100) given the observations. Importantly, variables were always binary and there is *no ground truth likelihood* for the query node given the observations, as the causal strengths and priors were not specified to subjects. Furthermore, humans were explicitly instructed that each cause can bring about the effect independently, which we also reflect in the prompts given to LLMs.

Having *no ground truth likelihoods* is a key feature of this dataset, as it allows us to study the *qualitative* patterns of causal inference judgments across different agents, for example, whether agents lean more towards *probabilistic* or *deterministic* reasoning behavior, or whether they exhibit *explaining away* behavior which humans typically do only weakly [1].

### 3.1.2 Cover Stories and Knowledge Domains

Rehder and Waldmann [1] embed the collider causal structure  $C_1 \rightarrow E \leftarrow C_2$  in one of three cover stories from three different knowledge domains (meteorology, economics, and sociology), allowing for a natural language description of the causal structure (see Figure 3.1). The three domains were chosen because the undergraduate subjects were expected to be relatively unfamiliar, such that their causal inferences would reflect the causal structure given to them and not idiosyncratic prior knowledge. Nevertheless, as an additional safeguard, the direction of each variable was *counterbalanced* (e.g., in the domain of sociology, some subjects were told that *high* urbanization causes *high* socio-economic mobility, others that it causes *low* socio-economic mobility, etc). In fact, Rehder and Waldmann [1] did not find significant effects of domain or the counterbalancing factor, suggesting that subjects' inferences were not strongly influenced by domain knowledge.

<sup>2</sup>Rehder and Waldmann [1] only inquired about a query node ● being present, i.e., 1.

### 3.1.3 Experimental Protocol for Humans

The experiment for *humans* consisted of two phases. In the *learning phase* subjects were presented and tested on the domain knowledge, including the causal mechanisms. In the *testing phase* they were presented with each of the inference tasks in random order on a sequence of four computer screens. A graphical representation of the collider structure remained on the screen during testing.

## 3.2 A Causal Reasoning Benchmark for LLMs with A Human Baseline

A key contribution of this work is the compilation of a causal inference task benchmark as described in Section 3.1.2, enabling direct comparisons between human causal inference judgments collected in Rehder and Waldmann [1] and LLMs. The benchmark is designed to replicate the experimental conditions of Rehder and Waldmann [1] (Experiment 1, Model-Only condition) as closely as possible with notable differences and extensions described below.

Part of this benchmark was published during the course of this thesis in the following work:

- [14] H. M. Dettki et al. “Do Large Language Models Reason Causally Like Us? Even Better?”  
In: *Annual Conference of the Cognitive Science Society* (2025)

### 3.2.1 Experimental Protocol for LLMs

In contrast to humans for the *LLMs* each textual prompt included both the content of the training and testing phase in a single prompt without the actual testing prompt that humans received and no graphical representation. The different domains and inference tasks were presented within each of the four counterbalancing groups. Whereas humans provided their probability judgments using a slider ranging from 0 to 100 with default setting=50.0, LLMs were instructed to provide a numerical answer  $\in [0.0, 100.0]$ .

**Example Prompt** Below is an *example prompt* from the sociology domain, matching the visualization in Figure 3.1 and diagnostic task X in Figure 4.2(e), where the query node (●) is  $C_1 = 1$  and  $C_2$  and the effect  $E$  are known to be absent. Note that only the *italicized text* following “:” was presented to LLMs in one piece.

- **Domain introduction:**
  - *Sociologists seek to describe and predict the regular patterns of societal interactions. To do this, they study some important variables or attributes of societies. They also study how these attributes are responsible for producing or causing one another.*
- **Causal mechanism:**
  - *Assume you live in a world that works like this:*
    - \*  $C_1 \rightarrow E$ : *High urbanization causes high socio-economic mobility.*
      - **Explanation:** *Big cities provide many opportunities for financial and social improvement.*
    - \*  $C_2 \rightarrow E$ : *Also, low interest in religion causes high socio-economic mobility.*

## Chapter 3 Causal Reasoning Benchmark

- **Explanation:** *Without the restraint of religion-based morality, the impulse toward greed dominates and people tend to accumulate material wealth.*
- **Observation:**
  - *Now suppose you observe the following: low socio-economic mobility and low urbanization.*
- **Inference task, here X:**
  - *Your task is to estimate how likely it is that low interest in religion is present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Note that each of the causes can bring about the effect independently. Please provide your answer as a single number between 0 and 100, where 0 means very unlikely and 100 means very likely. Do not include any explanations or additional text.*

**Prompting Configuration and Reasoning Effort** We test a range of LLMs from different providers (OpenAI, Anthropic, and Google) and model families (reasoning, non-reasoning, different model sizes) via their respective APIs. These models differ in how they can be configured to use additional budget for reasoning.

For all models, we use *zero-shot* (Numeric prompt-category) and *chain-of-thought (CoT)* prompting, where we ask the model to “think step by step” before answering, i.e. providing a likelihood judgment of the query node being 1. Where applicable, we set the temperature to 0.0 to get the most deterministic responses. For the GPT-5 family, we test different levels of *reasoning effort* configurable via the `reasoning_effort` parameter in the API. Another parameter unique to GPT-5 is `verbosity`, which we set to low.

### 3.2.2 Prompt and Content Manipulations

We also evaluate to what extent causal reasoning depends on knowledge about common causal relationships in the natural world, on the degree of irrelevant information in the prompt, and on the reasoning budget.

To this end, we extend the original RW17 prompts to *abstract domains* to test whether causal reasoning is influenced by domain knowledge, and to *overloaded prompts* to assess whether irrelevant information injected into the prompt degrades performance. A model reasoning robustly would be expected to demonstrate the same performance across these manipulations. For both manipulations, we adhere to the original prompt scaffolding from RW17 and test two prompting strategies: single-shot numeric prompts ● as in RW17 and chain-of-thought (CoT) prompts where LLMs are instructed to “think step by step” before providing their likelihood judgment ●.

**Abstract Domain** Besides the three original knowledge domains, we introduce an *abstract domain* to test whether causal reasoning is influenced by domain knowledge. While still adhering to the original prompt structure from RW17, we created three new abstract domains stripped of any real-world context, where each variable is now a randomly generated 10 character string assembled from a mix of letters, numbers, and symbols (e.g., “XJ3\_!9Pq2#”). This allows us to

## 3.2 A Causal Reasoning Benchmark for LLMs with A Human Baseline

test whether LLMs can apply their causal reasoning capabilities in a completely abstract setting without any real-world knowledge.<sup>3</sup>

**Overloaded Prompts** To test how easily distracted LLMs are by irrelevant information, we created *overloaded* versions of both the original RW17 and Abstract prompts by appending irrelevant information to reduce the signal-to-noise ratio in the prompt.

For details and examples of the abstract and overloaded prompts as well as single shot and chain-of-thought instructions, we refer the reader to [Section A.1](#).

### 3.2.3 Software Package

Our causal reasoning and human alignment benchmark is publicly available as part of a [Python package on GitHub](#). It can be used, for example, to benchmark additional LLMs or to algorithmically create custom prompts adhering to the RW17 scaffolding beyond the ones in this thesis. Subsequently, all analyses and figures presented in this thesis in [Chapter 4](#) can be reproduced using the package as well as extended to additional models or prompt variants.



<https://github.com/hmd101/causAlign>

---

<sup>3</sup>Note that, while these abstract prompts are intentionally designed to have not been seen during training, we cannot rule out that some LLMs may have been post-hoc trained on permutations that make them robust to nonsensical, abstract variables.





# Analysis and Experimental Results

---

4.1	Q1: Do Agents Reason Differently Across Domains? . . . . .	19
4.2	Q2: Are Humans and LLMs Aligned? . . . . .	21
4.3	Q3: Do Humans and LLMs Reason Normatively? . . . . .	22
4.3.1	Operational Definition of Normativity . . . . .	22
4.3.2	Fitting Causal Bayesian Networks to Likelihood Judgments . . . . .	23
4.3.3	Most Agents Are Described Well By A Causal Bayesian Network . . . . .	24
4.4	Q4: Reasoning Consistency Across Experiment-Prompt Conditions . . . . .	27
4.4.1	Reasoning Consistency . . . . .	27
4.4.2	Chain-of-Thought improves reasoning consistency and helps mitigate the impact of distracting information. . . . .	28
4.5	Q5: What Kind Of Cognitive Strategies Do Agents Use? . . . . .	28
4.5.1	Probabilistic vs. Deterministic Reasoning: Leak-Adjusted Determinacy . . . . .	28
4.5.2	Most LLMs reason more deterministically than humans, some reason more probabilistically than humans. . . . .	29
4.5.3	Qualitative Measures of Reasoning: Explaining Away and Markov Compliance . . . . .	30
4.5.4	Explaining Away and Markov Compliance in LLMs . . . . .	30
4.6	Q6: Do LLMs Reason Robustly Under Content Manipulations? . . . . .	33
4.6.1	Robustness Across Prompt and Content Manipulations . . . . .	33
4.6.2	Findings . . . . .	34

---

### TL;DR (Results at a glance)

**LLMs can follow causal rules, often more rigidly than people.** Across tasks and domains, most models apply the intended causal rules in a rule-like, repeatable way—*more deterministic than humans* (i.e., they give the same pattern of answers across items and downweight surface associations). A minority behave *more probabilistically than humans*, leaning on context and associative cues.

**Step-by-step prompting mainly boosts reliability; under noise it also improves reasoning quality.** Chain-of-thought (asking for intermediate steps) makes models' answers more consistent (*consistency = same pattern of answers across items*) and, when prompts are noisy or distracting, shifts behavior toward the intended causal rules (*normative behavior = follows the benchmark's causal structure*).

**Changing the context domain doesn't matter much; adding irrelevant text does.** Swapping real-world content for abstract placeholders leaves reasoning intact. But appending irrelevant sentences makes models less consistent and more driven by associations. Chain-of-thought recovers much of this loss.

**Two behavioral regimes emerge.** A *deterministic, rule-following* regime (low reliance on associations) and a *probabilistic, association-heavy* regime. Frontier instruction-tuned models cluster in the former; smaller/earlier models more often fall into the latter.

**Scope and limits.** Findings are behavioral and provide some insights as to *how* agents reason: as LLMs are proprietary (weights, training data, undisclosed parameter counts), we cannot make any claims about which features drive a certain behavior. Patterns should be interpreted as consistent behavioral trends under our tasks, providing some insight as to *how* they might be reasoning.

## Roadmap

We organize results by the six questions in Chapter 1: (Q1) *Domain differences* via significance tests; (Q2) *Human-LLM alignment* via Spearman correlation  $\rho$ ; (Q3) *Normative reasoning* via causal Bayes net fitting metrics; (Q4) *Reasoning consistency* via task-level LOOCV  $R^2$ ; (Q5) *Cognitive strategies* via causal Bayes net parameter signatures and collider induced reasoning signatures explaining away (EA) and Markov violation (MV); (Q6) *Genuine reasoning* as measured by robustness to content manipulations.

We benchmark over 20 LLMs from OpenAI, Anthropic, and Google in two prompting configurations: single-shot prompts (numeric ●) as in RW17, and chain-of-thought (CoT) ● prompts where LLMs are instructed to “think step by step” before providing their likelihood judgment. The dependent variable is the likelihood judgments (0–100) for a query variable being present given a set observations that emerge in the collider structure, amounting to eleven distinct structural causal reasoning tasks (see Figure 4.2 for collider graph visualizations of the 11 causal tasks).

In total, we evaluate 8 experimental conditions (2 prompting strategies  $\times$  4 content manipulations) where the content manipulations diverge from the original RW17 prompts by either stripping it of

## 4.1 Q1: Do Agents Reason Differently Across Domains?

real-world context and replacing it with abstract placeholders (Abstract), or by injecting irrelevant information into the prompt (Overloaded), for both the original RW17 and Abstract prompts.

Naming conventions of LLMs are provided in [Section 1](#) and a full list of models and some of their publicly available features such as release data and context window size is given in [Table A.1](#).<sup>1</sup>

### 4.1 Q1: Do Agents Reason Differently Across Domains?

We begin by testing whether our data supports the hypothesis that some agents exhibit domain-dependent reasoning. More precisely, we test for each agent, whether their likelihood distributions differ across domains, i.e. we have the following null and alternative hypothesis:

$$H_0 : \text{all domain distributions are identical} \quad \text{vs.} \quad H_1 : \text{at least one domain differs.}$$

We use a Kruskal–Wallis  $H$ -test and adjusted the resulting p-values via Benjamini–Hochberg correction to control the false discovery rate, since we are in a multiple comparisons setting. The results are given in [Table 4.1](#).

*Across domains within each agent*, agents showed no statistically significant differences in their likelihood distributions after a Benjamini–Hochberg correction indicating that the specific domain does not systematically affect reasoning behavior. For this reason, we pool domains for all subsequent analyses. For full results see [Section B.1](#).

**What about differences across agents within a domain?** Conversely, *across agents within each domain*, we see statistically significant differences after a multiple comparisons correction for each experiment and prompt condition. To disentangle which agents drive the differences within a domain we compared agents pairwise via Mann–Whitney U tests with Benjamini–Hochberg FDR correction. This showed that only a few agents, namely gpt-3.5-turbo and some of the gpt-5 variants drive these differences. For the purpose of this thesis, we won’t further investigate this and instead pool all domains and focus on cross-domain agent performance. See [Section B.1](#) for full results. [Figure B.2](#) shows the empirical cumulative distribution functions (ECDFs) for an illustrative set of agents across RW17 domains.

---

<sup>1</sup>Note that none of the LLMs’ providers tested here disclose model architectures, number of parameters or information about training data and post-training refinements limiting conclusions we can draw. When we refer to smaller models, we go by the providers’ naming conventions. For example, OpenAI’s gpt-4.1-mini is a smaller variant of gpt-4.1.

**Table 4.1:** Kruskal–Wallis test results for all agents.  $k$  = number of domains,  $H$  = test statistic,  $df$  = degrees of freedom,  $n_{\text{total}}$  = sample size,  $p_{\text{FDR-BH}}$  =  $p$ -value after Benjamini–Hochberg correction across agents.

Agent	$k$	$H$	$df$	$p$ -value	$n_{\text{total}}$	$p_{\text{FDR-BH}}$
claude-3-5-haiku-20241022	3	5.2021	2	0.0742	240	0.7420
claude-3-7-sonnet-20250219	3	0.6062	2	0.7385	240	0.9852
claude-3-haiku-20240307	3	1.5285	2	0.4657	240	0.9852
claude-3-sonnet-20240229	3	5.3585	2	0.0686	240	0.7420
claude-opus-4-1-20250805	3	2.0709	2	0.3551	232	0.9852
claude-opus-4-20250514	3	1.3777	2	0.5022	240	0.9852
claude-sonnet-4-20250514	3	0.1026	2	0.9500	240	0.9852
gemini-1.5-pro	3	1.0573	2	0.5894	2640	0.9852
gemini-2.5-flash	3	0.4403	2	0.8024	230	0.9852
gemini-2.5-flash-lite	3	1.5403	2	0.4630	240	0.9852
gemini-2.5-pro	3	0.4648	2	0.7926	2622	0.9852
gpt-3.5-turbo	3	10.6529	2	<b>0.0049</b>	240	0.1458
gpt-4	3	0.0749	2	0.9632	240	0.9852
gpt-4.1	3	0.1498	2	0.9279	240	0.9852
gpt-4.1-mini	3	0.4145	2	0.8128	240	0.9852
gpt-4o	3	0.8397	2	0.6572	2640	0.9852
gpt-5-mini-v_low-r_high	3	0.0334	2	0.9835	240	0.9852
gpt-5-mini-v_low-r_low	3	0.1805	2	0.9137	240	0.9852
gpt-5-mini-v_low-r_medium	3	0.4362	2	0.8040	240	0.9852
gpt-5-mini-v_low-r_minimal	3	3.4228	2	0.1806	240	0.9852
gpt-5-nano-v_low-r_high	3	1.2224	2	0.5427	240	0.9852
gpt-5-nano-v_low-r_low	3	0.1585	2	0.9238	240	0.9852
gpt-5-nano-v_low-r_medium	3	0.3626	2	0.8342	240	0.9852
gpt-5-nano-v_low-r_minimal	3	3.5389	2	0.1704	240	0.9852
gpt-5-v_low-r_low	3	0.0514	2	0.9746	240	0.9852
gpt-5-v_low-r_medium	3	0.3306	2	0.8476	240	0.9852
gpt-5-v_low-r_minimal	3	0.9507	2	0.6217	240	0.9852
humans	3	0.9405	2	0.6248	240	0.9852
o3	3	0.1413	2	0.9318	240	0.9852
o3-mini	3	0.0299	2	0.9852	240	0.9852

## 4.2 Q2: Are Humans and LLMs Aligned?

### TL;DR (Human–LLM Alignment (Q2))

*State-of-the-art (SOTA) models are aligned* with human reasoning. For smaller and older models *chain-of-thought* prompting substantially improves human–LLM alignment up to ceiling effects established by SOTA models. For those the impact of chain-of-thought is minimal.

For each agent  $a$  and domain  $d$ , we quantify human–LLM alignment by computing Spearman’s rank correlation coefficient  $\rho_{a,d}$  between human likelihood judgments and model predictions. Let  $\mathcal{I}_{a,d}$  denote the index set of all matched items for agent  $a$  in domain  $d$ , with  $h_i$  and  $m_i$  denoting the human and model judgments for item  $i \in \mathcal{I}_{a,d}$ . Then

$$\rho_{a,d} = \text{corr}\left(\text{rank}(\{h_i\}_{i \in \mathcal{I}_{a,d}}), \text{rank}(\{m_i\}_{i \in \mathcal{I}_{a,d}})\right).$$

See [Section A.2](#) for details on confidence intervals.

[Figure 4.1](#) shows human–LLM alignment measured via Spearman correlations ( $\rho$ ) between human and LLM judgments (95% CIs via 2000 bootstrap resamples), under both Numeric and CoT prompting. The numeric prompting-strategy refers to a single likelihood estimate per task as output, while CoT prompts elicit step-by-step reasoning before the final likelihood estimate. Full per-model results are reported in [Section B.2](#).

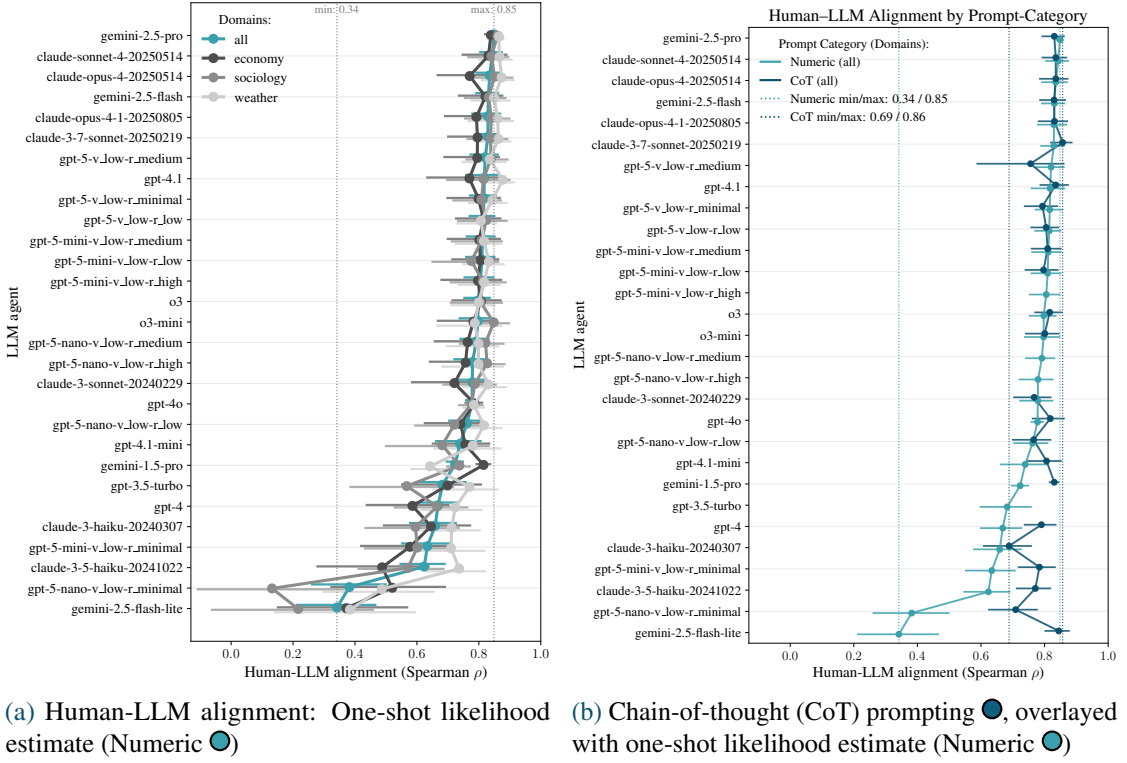
**Alignment Increases Under Chain-of-Thought Prompting** [Figure 4.1](#) shows that across most models, CoT prompting generally increases alignment. Numeric prompting yields correlations in the range  $\rho = 0.31$ - $0.85$  (see [Figure 4.1\(a\)](#)), while CoT boosts the overall range to  $\rho = 0.54$ - $0.85$ .

**Best-Performing Models Form a Saturated Cluster** State-of-the-art models (e.g., `gemini-2.5-pro`, `claude-sonnet-4-20250514`) reach the ceiling of observed human–LLM alignment, converging at  $\rho \approx 0.84$ - $0.85$  under both prompt types. For these models, CoT adds little or no further improvement.

Lighter and older models benefit most from CoT (see [Figure 4.1\(b\)](#)). For instance, `gemini-2.5-flash-lite` shows a dramatic increase of  $\Delta\rho = +0.503$ , from  $\rho = 0.342$  to  $\rho = 0.845$ . Other compact models show gains in the range of  $+0.15$  to  $+0.45$ .

While absolute  $\rho$  values vary by domain (with *weather* typically easiest and *economy* harder), the relative model rankings remain broadly stable. [Section B.2](#) reports domain-wise breakdowns in [Figure B.1](#) and exact numbers with confidence intervals in [Table B.15](#) and [Table B.16](#).

## Chapter 4 Analysis and Experimental Results



**Figure 4.1: Human-LLM alignment: CoT boosts alignment.** Each panel reports human-LLM alignment; per domain (shades of gray) and pooled domains (shade of blue) with 95% bootstrapped confidence intervals sorted from highest to lowest  $\rho$ . Vertical dashed lines indicate the minimum and maximum pooled  $\rho$  values across agents.

### 4.3 Q3: Do Humans and LLMs Reason Normatively?

Since the causal reasoning tasks all share the same underlying causal graph, i.e. a common-effect/collider graph ( $C_1 \rightarrow E \leftarrow C_2$ ), we can assess their degree of normative reasoning by fitting a causal Bayesian network (CBN) with a leaky noisy-OR parametrization (see also Sections 2.1 and 2.2) to each agent’s domain-pooled likelihood judgments per prompt condition and comparing the model fits. The better a causal Bayes net fits the predictions of an agent, the more normative its reasoning is.

#### 4.3.1 Operational Definition of Normativity

In the experiments by Rehder and Waldmann [1] there is no ground truth likelihood for the query node given the observations, as the causal strengths and priors were not specified to subjects, which doesn’t allow us to compute accuracy metrics against a ground truth. Instead, we define an *operational normativity rule* based on how well a leaky noisy-OR causal Bayesian network fits an agent’s judgments.

### 4.3 Q3: Do Humans and LLMs Reason Normatively?

We say an agent is *normative* iff there exists a set of causal Bayes net parameters  $\theta$  that describe the agent’s judgments well, as measured by error metrics (RMSE, MAE, loss).

#### 4.3.2 Fitting Causal Bayesian Networks to Likelihood Judgments

**Model and Parameters** For each experiment-prompt combination, we fit a causal Bayesian network (CBN) to the likelihood judgments of all tasks and domains jointly *per agent*. We represent collider structures using the *leaky noisy-OR* model. For binary causes  $C_1, C_2 \in \{0, 1\}$ , effect  $E$ , leak parameter  $b \in [0, 1]$ , and causal strengths  $m_1, m_2 \in [0, 1]$ , the conditional probability of the effect is

$$\Pr(E = 1 \mid C_1, C_2) = 1 - (1 - b)(1 - m_1 C_1)(1 - m_2 C_2). \quad (4.1)$$

This formulation captures that each active cause independently increases the likelihood of the effect, while the leak  $b$  accounts for background activations in the absence of any active causes. Additionally, priors on the causes  $\Pr(C_i = 1) = p(C_i)$  specify the baseline frequency of each cause. To obtain an unconstrained optimization problem, parameters are mapped to the unit interval via a sigmoid transform, e.g.  $p(C_i) = \sigma(\theta_{pC_i})$ ,  $b = \sigma(\theta_b)$ ,  $m_i = \sigma(\theta_{m_i})$ .

**Model Fitting** For each task  $t$  and query node  $Q$ , the model gives a predictive probability  $\hat{y}_t(\theta) \in [0, 1]$  for  $Q = 1$  given the task’s observed parent states. We normalize likelihood judgments to  $[0, 1]$  and define residuals  $r_t = \hat{y}_t(\theta) - y_t$ . We fit the causal Bayesian networks via the following regression objective over tasks:

$$\min_{\theta \in [0, 1]^d} \sum_{t \in \mathcal{T}} \ell(r_t) \quad \text{s.t.} \quad \theta \in [0, 1]^d,$$

where  $\ell(\cdot)$  is either mean-squared error (MSE),  $\ell(r) = \frac{1}{2}r^2$ , or the Huber loss

$$\ell_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta, \end{cases}$$

In our implementation,  $\delta$  is fixed at  $\delta = 1.0$ ; we did not tune  $\delta$  via validation. Empirically, we tried both losses and often observed more stable training with Huber, consistent with occasional heavy-tailed errors and bounded/extreme ratings (some tasks place substantial mass at the scale endpoints).<sup>2</sup>

**Optimization details.** We fit three and four parameter CBNs by constrained gradient-based optimization with  $R$  stochastic restarts (distinct random seeds) to mitigate local minima. Each restart  $r$  yields  $(\theta_r^*, L_r)$  with final loss  $L_r$  and derived per-restart metrics. Selection of the representative (“best”) model per parameter-tying scheme is via lowest loss  $L_r$  but can be changed to other metrics in our code base `causAIgn`. We use either L-BFGS (default) with a

<sup>2</sup>See Section B.3 for empirical cumulative distributions of likelihood judgments, illustrating the prevalence of endpoint clustering and heavy tails.

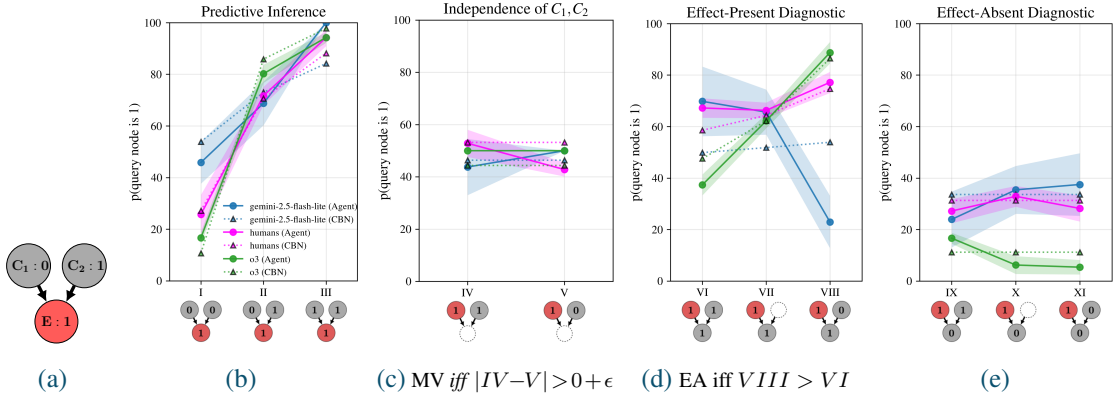
## Chapter 4 Analysis and Experimental Results

fixed maximum number of iterations or Adam with a fixed epoch budget; we do not employ early stopping. Parameters are kept in-range via smooth re-parameterizations. We fit per agent and prompt condition where per experiment all domains are pooled, i.e., treated as one.

**Model selection: Which CBN describes an agent best?** For each experiment and prompt-category, we select a single CBN-*winner* per agent amongst the different parameter tying schemes we fit. This happens *after* fitting all schemes with  $R$  restarts each and having selected the best restart per scheme. The primary criterion is to maximize pooled LOOCV- $R^2$  when available.<sup>3</sup>

### 4.3.3 Most Agents Are Described Well By A Causal Bayesian Network

Figure 4.2 visually displays causal Bayes net fits (dashed lines) for an illustrative subset of agents across all 11 collider tasks (I-XI) in the RW17 independent causes domain with numeric prompts.



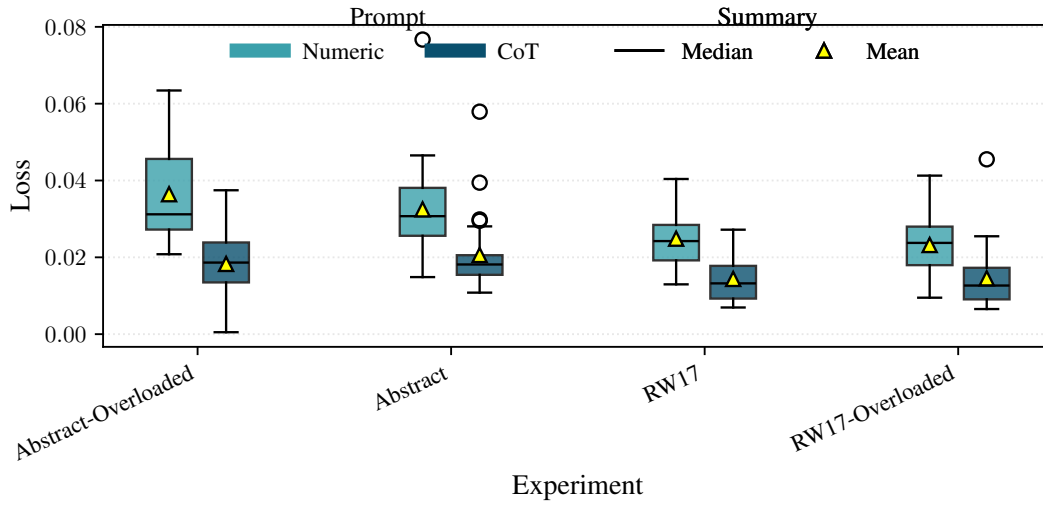
**Figure 4.2: Agents vs. Causal Bayes Net Predictions across 11 collider induced inference tasks (I-XI) (RW17, Numeric ●) for an illustrative subset of agents.** *Some agents are well described by their CBN model in dashed lines (e.g., o3), others not so much (e.g., gemini-2.5-flash-lite).* Likelihood judgments that query node ● has value  $1 \in \{0, 100\}$  of agents' predictions vs. their respective CBN model predictions with bootstrapped 95% confidence intervals for agents ordered by reasoning category Figures 4.2(b) to 4.2(e). **Plot details:** Graphs on the x-axis visualize the conditional probability of the causal inference tasks (I-XI) where the nodes are colored according to: ● → query node that the question is asked about; ● → observed  $\in \{0, 1\}$ ; and ○ → no information on. **Panel descriptions:** Figure 4.2(a) shows the reference graph for task II. Figure 4.2(c) shows *Markov violations (MV)* for humans and gemini-2.5-flash-lite, as  $|IV - V| > 0 + \epsilon$ , visualized by non-horizontal lines, where  $\epsilon$  is 0.05 in our study. o3 shows no Markov violations and perfect independence of causes. Figure 4.2(d) brings about *explaining away (EA)*, iff  $VIII > VI$ , visualized by a positive slope. o3 displays *perfect EA*, whereas gemini-2.5-flash-lite shows no EA and *humans show weak EA*. Experiment: Semantically meaningful (RW17) content, numeric prompt.

<sup>3</sup>If all LOOCV- $R^2$  values are missing/NaN in a group, we fall back to (in order): cross-validated  $R^2$  if present, in-sample  $R^2$ , then information criteria (prefer lower BIC, then lower AIC). Ties within a small epsilon on the primary metric, here LOOCV- $R^2$ , are broken by preferring: lower LOOCV-RMSE, then lower BIC, lower AIC, lower training loss.



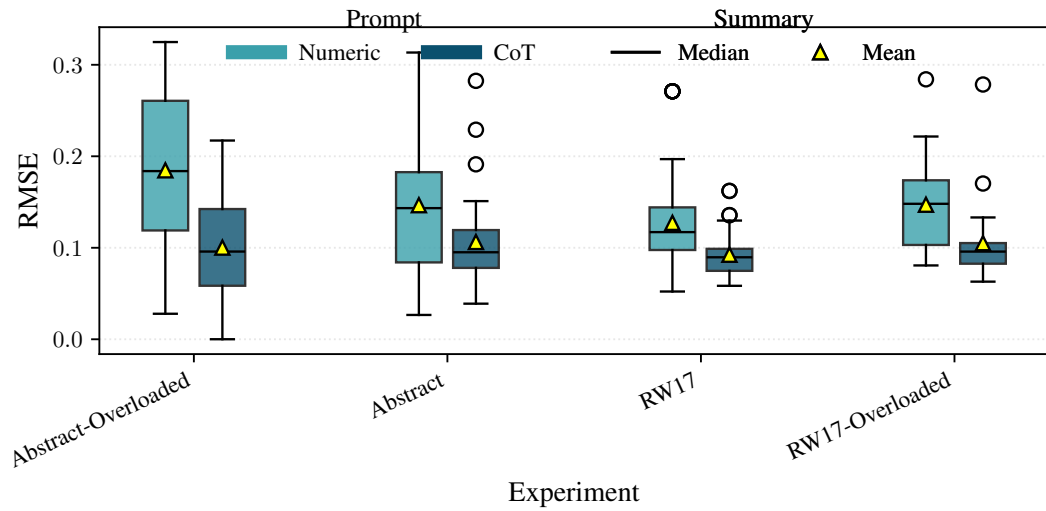
### 4.3 Q3: Do Humans and LLMs Reason Normatively?

**Chain-of-Thought Prompting Improves Normative Reasoning** We evaluate causal Bayes net fits via three metrics: Mean Absolute Error ( $MAE \in [0, 1]$ ), Root Mean Squared Error ( $RMSE \in [0, 1]$ ),  $R^2 \in [-\inf, 1]$ , and and Huber loss (with  $\delta = 1$ ,  $\in [0, 05]$ , see [Section 4.3.2](#)). Low loss and low causal Bayes net errors (MAE, RMSE) indicate that there is a set of parameters  $\theta$  within the leaky noisy-OR parameterization that describes an agent’s reasoning behavior well across all 11 collider tasks and can be said to reason normatively. [Figures 4.3 to 4.5](#) show that chain-of-thought (CoT) prompting  $\bullet$  generally improves causal Bayes net fits compared to numeric prompting  $\bullet$  across all three goodness of fit metrics. Most agents are well described by a causal Bayes net ( $MAE \in [0.00004, 0.2887]$ ,  $MAE_{median} = 0.0710$ ,  $RMSE \in [0.00005, 0.32455]$ ,  $RMSE_{median} = 0.0950$ , Huber loss  $l \in [0.0005, 0.0579]$ ,  $l_{median} = 0.0124$ ; and  $R^2 \in [0.039, 0.995]$ , with  $R^2_{median} = 0.84$ ).

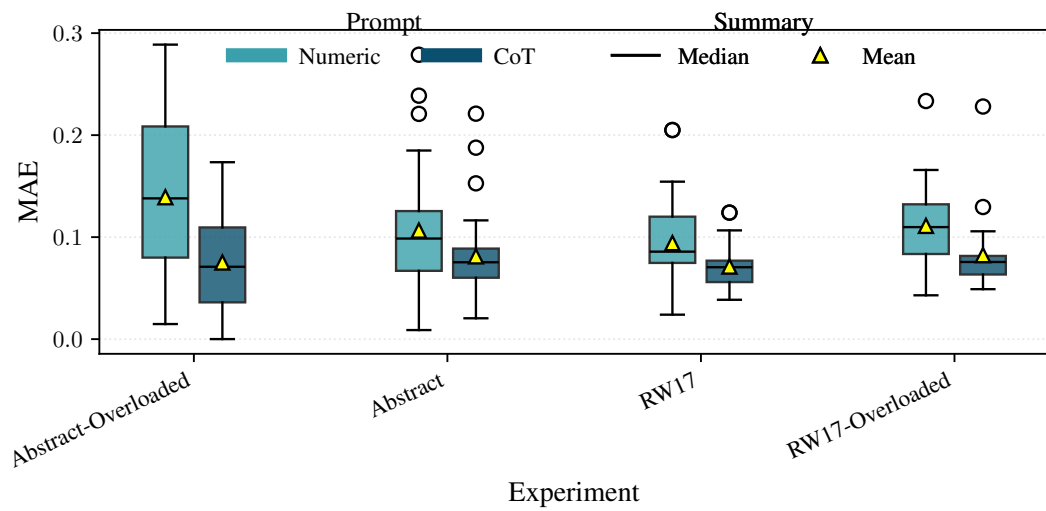


**Figure 4.3: Chain-of-thought (CoT) improves causal Bayes net fits.** Chain-of-thought (CoT) prompting  $\bullet$  generally reduces loss compared to numeric prompting  $\bullet$ . Huber loss ( $\in [0, 05]$ , with  $\delta = 1$ , see [Section 4.3.2](#)).

**Least normative LLMs** LLMs with the highest loss and error metrics are primarily `claude-3-haiku`, `claude-3-opus`, `claude-3-sonnet`, `gpt-3.5-turbo`, `gpt-4.1-mini`, and the smaller `gpt-5 mini` and `nano` variants, particularly under numeric prompting conditions where reasoning effort is minimal. Conversely, `gemini-2.5-pro`, `claude-opus-4-1`, `o3-mini`, and `o3` are among the best fitting models with low error metrics and low loss across experiments and prompt conditions. For a complete overview, see [Section B.7 Tables B.19 to B.25](#) for full results, including the three or four parameter tying scheme.



**Figure 4.4: CBN error metrics: RMSE** Error is higher in overloaded versions, in particular when stripped of semantically meaningful context (Abstract). Chain-of-thought (CoT) prompting generally reduces RMSE compared to numeric prompting.



**Figure 4.5: CBN error metrics: MAE** Error is higher in overloaded versions, in particular when stripped of semantically meaningful context (Abstract). Chain-of-thought (CoT) prompting generally reduces MAE compared to numeric prompting.

## 4.4 Q4: Reasoning Consistency Across Experiment-Prompt Conditions

### TL;DR

Across all experiments, *chain-of-thought (CoT) increases reasoning consistency* (LOOCV  $R^2$ ) and *tightens dispersion*, with *largest gains under overloaded prompts*; several SOTA LLMs match or exceed the human consistency benchmark reaching ceiling performance.

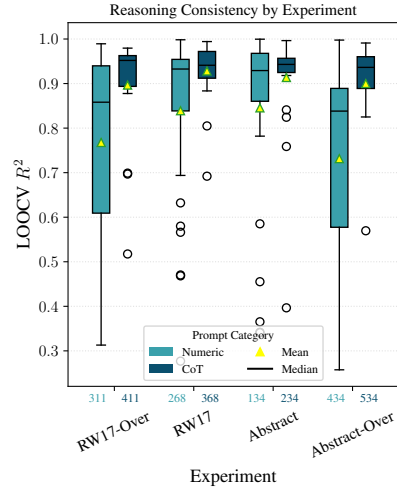
Next, we ask how consistently an agent applies a reasoning strategy across related problems within the collider family.

### 4.4.1 Reasoning Consistency

We define *reasoning consistency* as the task-level LOOCV- $R^2$  from a leaky noisy-OR CBN fit across the 11 tasks (fit on 10; predict the held-out task; average over folds). This score is agnostic to explaining away (EA) and Markov violation levels beyond the shared responses and captures how consistently an agent applies a strategy across related problems.

**Evaluating Cross-task Generalization** We perform leave-one-out cross-validation (LOOCV) over the collider tasks: for each held-out task, fit on the remaining tasks and predict the held-out with the winning model’s configuration. Our LOOCV metrics are computed by pooling predictions across folds rather than averaging fold-wise metrics. Concretely, let  $\hat{y}^{(i)}$  and  $y^{(i)}$  be the prediction and actual for fold  $i$ . We form concatenated vectors  $\hat{\mathbf{y}} = [\hat{y}^{(1)}, \dots, \hat{y}^{(K)}]$  and  $\mathbf{y} = [y^{(1)}, \dots, y^{(K)}]$ , then compute  $\text{RMSE} = \sqrt{\frac{1}{K} \sum_i (\hat{y}^{(i)} - y^{(i)})^2}$ ,  $\text{MAE} = \frac{1}{K} \sum_i |\hat{y}^{(i)} - y^{(i)}|$ , and  $R^2 = 1 - \text{SS}_{\text{res}}/\text{SS}_{\text{tot}}$  with  $\text{SS}_{\text{res}} = \sum_i (\hat{y}^{(i)} - y^{(i)})^2$  and  $\text{SS}_{\text{tot}} = \sum_i (y^{(i)} - \bar{y})^2$ , where  $\bar{y}$  is the mean of *all* held-out actuals. Higher pooled LOOCV- $R^2$  indicates better cross-task generalization of a single CBN for that agent/condition.

The fitted parameters and LOOCV  $R^2$  for all agents are reported in [Section B.7](#).



**Figure 4.6:** Reasoning consistency measured by LOOCV  $R^2$  grouped by experiment and prompt category. CoT helps improve  $R^2$  scores most within overloaded conditions (see left- and right most box plot pair), mirroring them  $R^2$  closer the plain experiments (RW17 and Abstract) where CoT has a smaller but still positive effect. Small numbers represent average prompt length (in tokens) per experiment-prompt-condition.

#### 4.4.2 Chain-of-Thought improves reasoning consistency and helps mitigate the impact of distracting information.

We report reasoning consistency via task-level leave-one-out cross-validated (LOOCV)  $R^2$  in Figure 4.6 grouped by experiment and prompt condition. Chain-of-thought ● increases median reasoning consistency relative to single-likelihood estimates (Numeric ●) and narrows inter-quartile ranges in all four settings, real-world domains (*RW17*), *Abstract* prompts, real-world domains with overloaded prompts (*RW17-Over*), and abstract overloaded prompts (*Abstract-Over*). Dispersion reduces most in the overloaded conditions, indicating that Chain-of-Thought mitigates distraction effects induced by irrelevant content.

See Figure B.5 for agent breakdowns by experimental conditions.

**Human benchmark.** Humans (*RW17*, Numeric) achieve LOOCV- $R^2 = 0.937$ , which several models match or exceed (see Figure B.5).

**Implication for reasoning consistency.** Chain-of-Thought generally *improves reasoning consistency*—especially when prompts are overloaded—where it seems like stepwise reasoning with chain-of-thought ● facilitates models to mediate the diluted signal-to-noise ratios in overloaded conditions.

### 4.5 Q5: What Kind Of Cognitive Strategies Do Agents Use?

**TL;DR** (Cognitive Strategies in LLMs (Q5))

**Most LLMs reason more deterministically than humans, with low leak and strong causal strength parameters.** Chain-of-Thought can increase determinism and explaining away in smaller models, but has mixed effects overall. Markov compliance is common, but a few LLMs—like humans—show associative biases.

#### 4.5.1 Probabilistic vs. Deterministic Reasoning: Leak-Adjusted Determinacy

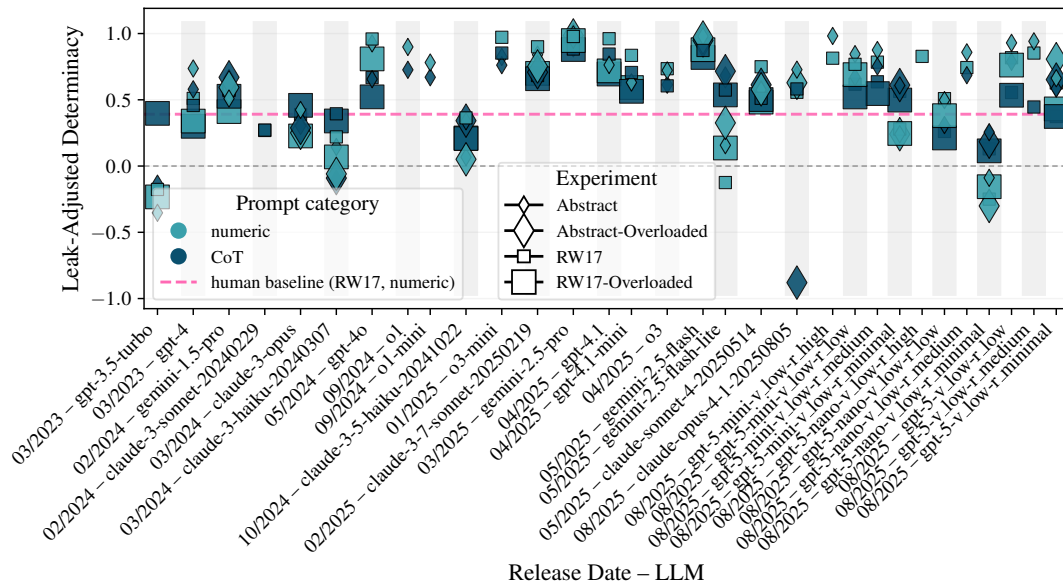
To quantify the degree of deterministic versus probabilistic reasoning in a noisy-OR model, we define the *Leak-Adjusted Determinacy* (LAD)  $\in [-1, 1]$  as

$$\text{LAD}_{\text{agent}} = \overline{m}_{\text{agent}} - b_{\text{agent}} \quad (4.2)$$

where  $\overline{m}_{\text{agent}}$  is the mean of the causal strength parameters of  $m_{1,2}$  and  $b_{\text{agent}}$  is the leak parameter of the fitted CBN for that agent. Positive values of LAD  $\rightarrow 1$  indicate that average causal strength exceeds background leak, reflecting more deterministic reasoning, whereas negative values indicate leak-dominated (more probabilistic) reasoning LAD  $\rightarrow -1$ .

### 4.5.2 Most LLMs reason more deterministically than humans, some reason more probabilistically than humans.


Figure 4.7 shows Leak-Adjusted Determinacy (LAD, see Equation (4.2)) per experiment and prompt category by LLM release date. Higher Leak-Adjusted Determinacy (LAD) indicates more deterministic reasoning while lower LAD indicates more probabilistic reasoning. Most agents reason more deterministically than humans, indicated by more points above the pink human threshold (RW17, Numeric:  $LAD \approx 0.45$ ), with some agents (e.g., gemini-2.5-pro and gemini-2.5-flash) reaching near-perfect determinism ( $LAD \approx 1.0$ ). A handful of agents reason more probabilistically than humans, some even with negative Leak-Adjusted Determinacy values (scatters below 0) indicating they assume the effect is often present even in the absence of causes (higher leak  $b$  and/or lower causal strengths  $m$ ) than humans. Figure 4.7 also shows



**Figure 4.7: Leak-Adjusted Determinacy (LAD) per experiment and prompt category by LLM release date.** Higher Leak-Adjusted Determinacy (LAD) indicates more deterministic reasoning while lower LAD indicates more probabilistic reasoning.

effects of experiment and prompt category on Leak-Adjusted Determinacy (LAD) indicated by shape, size and color of scatter. This is what we will explore next in more detail in Section 4.6. For a parameter breakdown per agent and experimental condition see Section B.7.

**Chain-of-Thought has mixed effects on Leak-Adjusted Determinacy (LAD).** Figure 4.10(b) shows Leak-Adjusted Determinacy (LAD) for the RW17 experiment and the effects of Chain-of-Thought. Chain-of-Thought ● tends to increase Leak-Adjusted Determinacy (LAD) relative to single direct likelihood estimate prompts ● for some LLMs, smaller ones (e.g., gpt-5-nano-v low-r minimal, gemini-2.5-flash-lite). Conversely, LLMs that already reason very

deterministically (e.g., `gemini-2.5-pro`, `o3-mini`) show slightly decreased Leak-Adjusted Determinacy (LAD) levels under Chain-of-Thought . Again, this figure illustrates, that most LLMs reason more deterministically than humans (more points to the right of the pink horizontal human benchmark line), with some exceptions (e.g., `gpt-3.5-turbo`, `claude-3-sonnet`).

### 4.5.3 Qualitative Measures of Reasoning: Explaining Away and Markov Compliance

In collider structures, two key qualitative signatures emerge that have been widely studied in humans: they are (a) *explaining away* (EA) and (b) *Markov violation/compliance* (MV/MC). It has been repeatedly shown that humans often fail to exhibit these signatures robustly [1], which is why we ask whether LLMs exhibit these signatures more robustly than humans. For this, we define qualitative signatures for explaining away (EA) and Markov violation (MV) based on the agents' raw likelihood judgments (not their CBN predictions).<sup>4</sup>

We evaluate agents against qualitative collider norms and a model-based consistency check.

**Explaining Away (EA)** Explaining away in a collider graph occurs when evidence for one cause reduces the belief in the alternative cause (see Section 2.3.2), i.e.

$$\text{EA iff } \Pr(C_1=1 \mid E=1) - \Pr(C_1=1 \mid E=1, C_2=1) > 0. \quad (4.3)$$

Visually explaining away is represented as a positive slope in Figure 4.2(d).

**Markov Violation and Compliance (MV/MC)** A Markov violation occurs when the presence or absence of one cause affects the belief in another cause, violating the independence assumption in a collider structure (see Section 2.3.3). Formally, we define the Markov violation signature as

$$\text{MC iff MV} = |\Pr(C_1=1 \mid C_2=1) - \Pr(C_1=1 \mid C_2=0)| \approx 0.$$

Markov compliance (MC), meaning no violation occurs  $\text{MV} \approx 0$ , is visually represented as a flat line in Figure 4.2(c).

### 4.5.4 Explaining Away and Markov Compliance in LLMs

Figure 4.9 shows Explaining Away (EA) and Markov Violation (MV) levels for three illustrative agents in the middle and right panel respectively across all experiments and prompt categories. For an agent level breakdown per experiment see Figure 4.8(a) and Figure B.3 for explaining away and Figure 4.8(b) and Figure B.4 for Markov violations.

*High explaining away levels* occur when the presence of one cause explains away the presence of an alternative cause drastically reducing the likelihood of the alternative cause (represented by scatters clustered on the right side in Figure 4.8(a) and in the center panels of Figure 4.9).

<sup>4</sup>To our knowledge, there are no canonical numeric standards for EA or MV magnitudes; norms in the causal literature are typically qualitative (EA should be positive; MV should be practically zero).

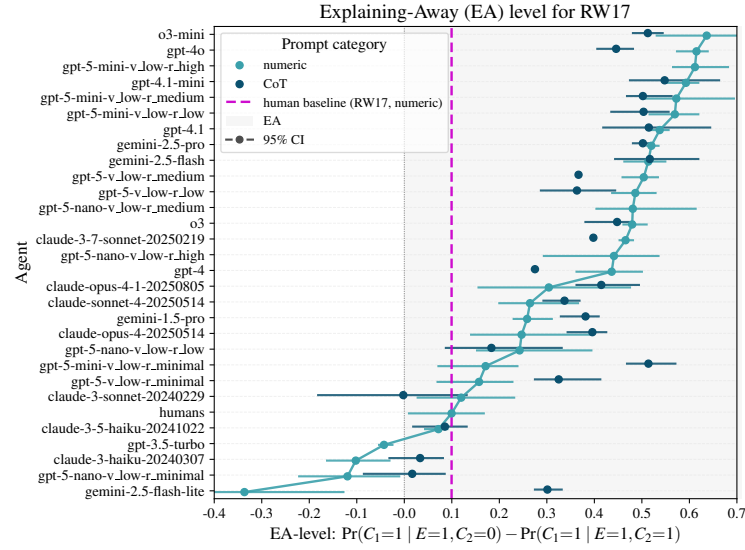
## 4.5 Q5: What Kind Of Cognitive Strategies Do Agents Use?

*Markov compliance (MC)* occurs when the presence of one cause does not affect the likelihood of another cause (represented by scatters clustered around 0 in the right panels).

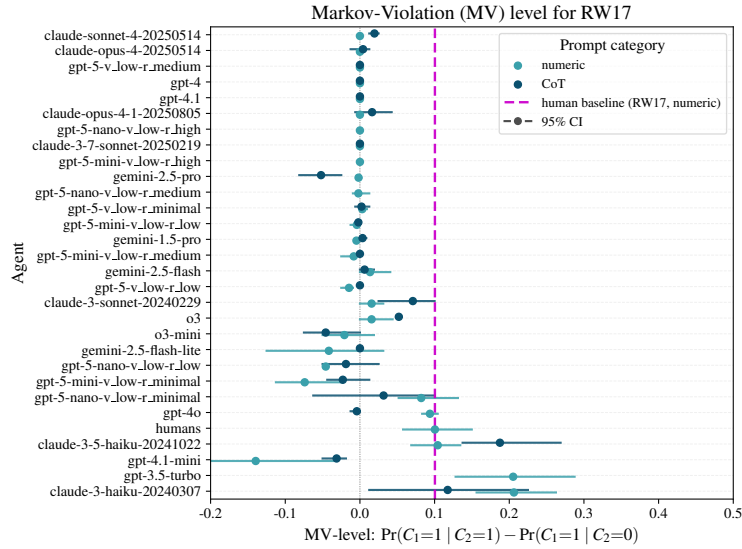
**Most LLMs exhibit substantially stronger explaining away than humans.** Figure 4.8(a) displays the explaining away levels for the RW17 experiment and shows things: (1) most LLMs exhibit explaining away (scatters in the gray shaded area right of 0); (2) most LLMs exhibit stronger explaining away compared to humans (scatters to the right of pink horizontal line, RW17, Numeric:  $EA \approx 0.099$ ); and (3) Chain-of-Thought ● tends to increase explaining away levels relative to single direct likelihood estimate prompts ● for some LLMs. Those effects are most pronounced for smaller models (e.g., gpt-5-nano-v low-r minimal, gemini-2.5-flash-lite and gpt-5-mini-v low-r minimal). See Figure B.3 for an agent level breakdown for the remaining three experiments.

**Most LLMs exhibit Markov compliance, while some show associative biases like humans.** Figure 4.8(b) displays the Markov violation levels (when scatters are away from 0) for the RW17 experiment and shows things: (1) most LLMs exhibit Markov compliance (scatters around 0); (2) some LLMs exhibit associative biases like humans or greater (scatters to the right of pink horizontal line), and (3) Chain-of-Thought ● tends to decrease Markov violation levels (pushes scatters closer to 0) relative to single direct likelihood estimate prompts ● for some LLMs. Those effects are most pronounced for smaller models (e.g., gpt-5-nano-v low-r minimal, gemini-2.5-flash-lite and gpt-5-mini-v low-r minimal). See Figure B.4 for an agent level breakdown for the remaining three experiments.

## Chapter 4 Analysis and Experimental Results



(a) *Explaining Away (EA)*: Most agents exceed the human baseline ( $EA > 0.099$ ), and CoT effects are mixed. Under CoT, all agents display some level of explaining away ( $EA > 0$ ). Some agents increase EA levels drastically, namely those with lower EA under Numeric, while others slightly decrease, mostly those with higher EA under Numeric.



(b) *Markov Violation (MV)*: Most agents respect cause independence ( $MV$  near 0), while humans reason associatively ( $MV > 0$ ). CoT pushes some agents closer to 0, namely mostly those with higher  $|MV|$  under Numeric.

**Figure 4.8: Impacts of CoT and comparison to human baseline across explaining away and Markov compliance levels for RW17.** Agents that exceed human baselines (higher EA, lower  $|MV|$ ) in the Numeric prompt-condition mostly belong to reasoning models are for example gemini-2.5-pro, the reasoning model o3. Smaller models (e.g., gpt-5-nano and -mini and gemini-2.5-flash-lite) seem to benefit most from CoT prompting.



## 4.6 Q6: Do LLMs Reason Robustly Under Content Manipulations?

**TL;DR** (Robustness to Content and Prompting Manipulations (Q6))

**CoT prompting enhances reasoning robustness, especially under noisy or abstract conditions.** Most LLMs maintain stable causal structure across a subset of content manipulations, but overloaded prompts shift towards more probabilistic (increased leak, reduced causal strength). Gemini-2.5-pro is most robust and deterministic across all conditions, while smaller models benefit most from CoT (e.g., Gemini-2.5-flash-lite).

### 4.6.1 Robustness Across Prompt and Content Manipulations

We finally ask how robustly LLMs reason across content and prompt manipulations across all metrics (determinism, explaining away, Markov compliance, and reasoning consistency) and in comparison to the human baseline (RW17-Numeric). To refresh the reader’s mind, we briefly recap the experimental manipulations and evaluation metrics and provide a guide to how to interpret the scatter plots in (Figure 4.9) in Section 4.6.1 that follow and contain the robustness analysis at a glance; otherwise skip to Section 4.6.2 for the main findings.

**Experimental Manipulations (Recap)** We assess robustness across two orthogonal content manipulations: (A) *Content reduction* — rendering inputs semantically abstract (RW17 → Abstract); (B) *Noise injection/Overloading* — diluting signal with irrelevant information (RW17 → RW17-Over, and Abstract → Abstract-Over). We also evaluate the effect of prompting strategy (Numeric vs. CoT) on reasoning robustness.

**Evaluation Metrics (Recap)** We assess: (i) *normative reasoning* via a set of causal Bayes net parameters that describes an agent’s behavior well (low error & loss); (ii) *Reasoning consistency* via LOOCV  $R^2$  (Figure 4.6); (iii) *probabilistic vs. deterministic reasoning* via Leak-Adjusted Determinacy (LAD, see Equation (4.2)); and (v) common-effect/collider graph specific *Explaining away (EA)* and *Markov violation (MV)* using raw normalized judgments (Figures B.3 and B.4).

### A guide to interpreting Figure 4.9: Robustness across all metrics at a glance

Figure 4.9 summarizes the remaining metrics (ii-iv) for each agent’s performance across content and prompting manipulations. Each agent’s scatter plot (e.g., Figure 4.9(a)) visualizes:

- *Y-axis*: probabilistic vs. deterministic reasoning via Leak-Adjusted Determinacy (LAD) ( $LAD = \bar{m} - b$ ), where higher values (top) indicate more deterministic reasoning (high causal strength  $m$ , low leakage  $b$ ) and lower values (bottom) indicate more probabilistic reasoning (low  $m$ , high  $b$ );

## Chapter 4 Analysis and Experimental Results

- *X-axes:*  $R^2$ , EA, and MV respectively from left to right, higher values (to the right) indicate higher reasoning consistency and higher explaining away, while values away from zero indicate Markov violation and conversely, values near zero indicate Markov compliance.
- *Symbols:* Experimental conditions ( $\square$  for RW17,  $\diamond$  for Abstract), enlarged for overloaded variants;
- *Colors:* Prompt types (Numeric:  $\bullet$ , CoT:  $\bullet$ ).

Clusters indicate robustness across the respective conditions.

### 4.6.2 Findings

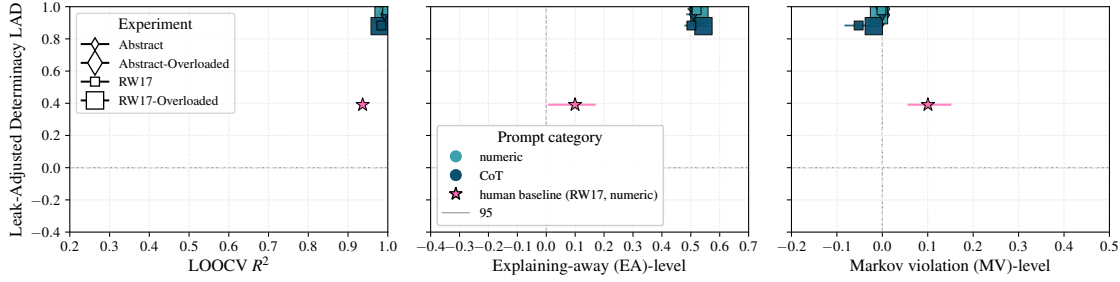
**Gemini-2.5-flash and pro are most deterministic and robust across metrics and conditions.** Figures B.8(c) and 4.9(a) show Gemini-2.5-flash and pro, revealing clusters for all 8 experimental manipulations in the top right for  $R^2$  and EA, and MV near 0 with high LAD, indicating high robustness as well as strong deterministic reasoning, and explaining-away and Markov compliance. Conversely, Gemini-2.5-flash-lite (Figure 4.9(b)) shows clear effects of prompt-category (Numeric  $\bullet$  vs. chain-of-thought  $\bullet$ ) for  $R^2$  and EA, indicated by the spread of triangles across the x-axis. In particular, chain-of-thought ( $\bullet$ ) leads to higher  $R^2$  and EA levels, while MV levels are similar across prompt-categories.

**Content Effects (RW17 vs. Abstract) are less systematic than prompting effects (Numeric vs. CoT) for probabilistic vs deterministic reasoning.** Clear separation along the y-axis indicates a shift from probabilistic to deterministic reasoning and is observed in a number of agents such as gpt-5-mini-v-low-r-minimal (Figure B.10(d)), gpt-5-nano-v-l-r-minimal and gpt-3.5-turbo in Figure 4.9(c). It seems that the prompt-type (Numeric  $\bullet$  vs. chain-of-thought  $\bullet$ ) has a greater effect on the level of deterministic reasoning (LAD levels) than content manipulations (RW17  $\square$  vs. Abstract  $\diamond$ ). This is revealed by the fact that the clustering along the y-axis is more prevalent for darkblue  $\bullet$  (CoT) vs. lighter blue (Numeric,  $\bullet$ ) indicating that chain-of-thought prompting has a greater effect on shifting reasoning in the probabilistic vs deterministic regime (LAD levels) but effects remain mixed depending on the agent.

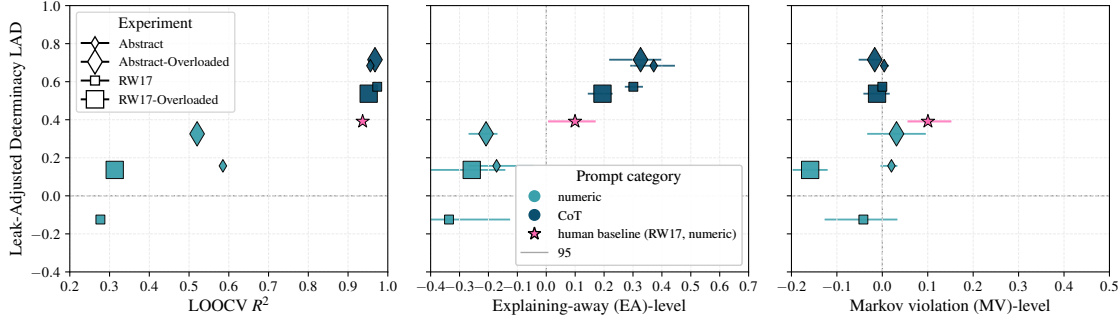
**Summary of Findings** Gemini-2.5-pro and flash and exhibit the most consistent and normative reasoning across conditions indicated by a single cluster in Figures B.8(c) and 4.9(a). Other models vary: some benefit from chain-of-thought (e.g., for reasoning consistency and explaining away), while content manipulations exert less consistent effects. Where agents do not already perform at ceiling, chain-of-thought prompting improves reasoning consistency and explaining away for most agents, with minimal cost to Markov violations and some improvements. These findings suggest that, while some general trends seem to emerge, meaningful insights are best drawn at the agent level, since effects vary widely across models and within most models and the interpretation of results depends on the specific use case or interest in how a given LLM is desired to behave relative to humans under experimental manipulations across these metrics.

## 4.6 Q6: Do LLMs Reason Robustly Under Content Manipulations?

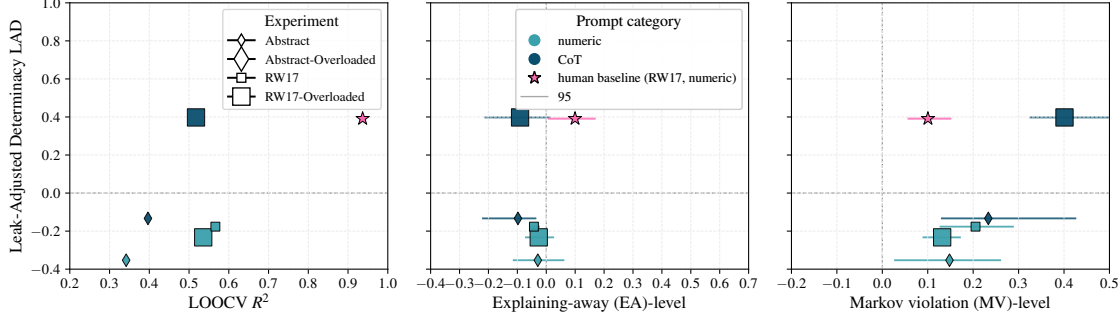
Leak-Adjusted Determinacy ( $LAD = \bar{m} - b$ ) vs  $R^2$ /EA/MV



(a) *Gemini-2.5-pro* reasons most normatively high LAD and robustly across experimental conditions. From left to right panel, high LOOCV  $R^2$  indicates high reasoning consistency, high EA levels, and respecting independence of causes (MV near 0).



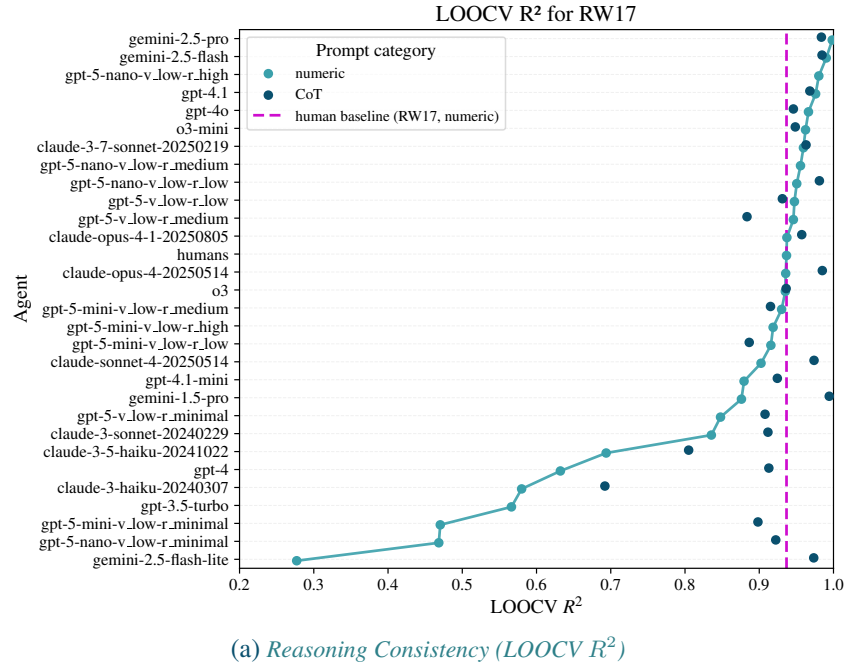
(b) *Gemini-2.5-flash-lite* shows clear effects of prompt-category (Numeric  $\bullet$  vs. CoT  $\bullet$ ) where  $\bullet$  leads to higher  $\Psi_{norm}$ ,  $R^2$  and EA levels, while MV levels are pushed to 0. Abstract conditions (diamonds) tend to score slightly higher / closer to 0 for MV than RW17 (squares), indicating a slight sensitivity to prompt content.



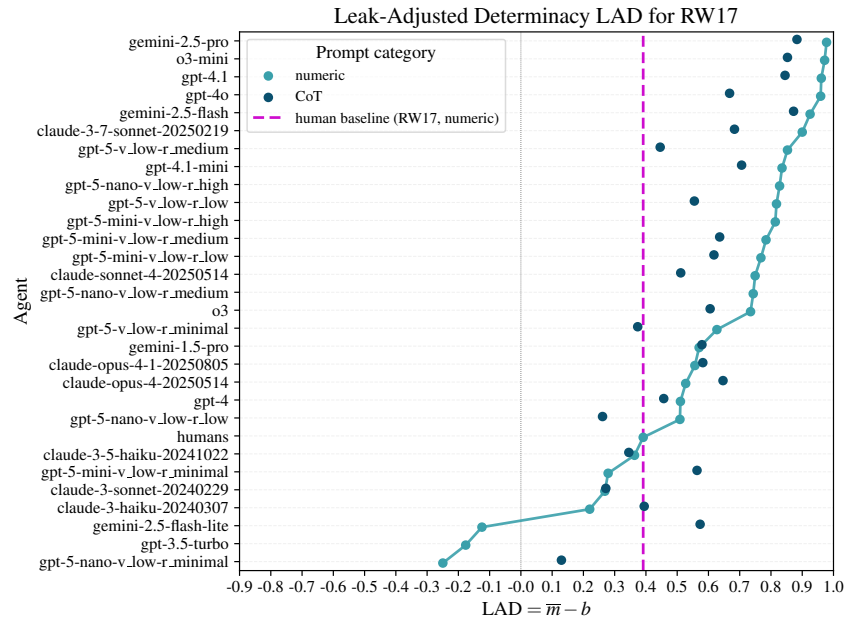
(c) *GPT-3.5-turbo* shows great prompt-category effects (Numeric  $\bullet$  vs. CoT  $\bullet$ ) for LAD, where  $\bullet$  leads to higher levels, while MV levels are greater than 0 across all conditions indicating associative reasoning.

**Figure 4.9: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning. For the LOOCV  $R^2$  panels, points closer to the right indicate higher reasoning consistency. For the EA panels, points closer to the top indicate higher EA levels, i.e., the presence of one cause explains away the presence of an alternative cause to have brought about the effect. For the MV panels, points closer to 0 indicate higher Markov compliance (respecting independence of causes). Additionally, for MV, points above 0 indicate positive MV (overestimating the effect of the cause on the effect) and points below 0 indicate associative reasoning, i.e., that the presence of once cause increases another one.

## Chapter 4 Analysis and Experimental Results



(a) Reasoning Consistency (LOOCV  $R^2$ )



(b) Leak-Adjusted Determinacy LAD.

**Figure 4.10: Impacts of CoT and comparison to human baseline across Collider and CBN-induced metrics for RW17.** The top row shows metrics based on normalized raw likelihood judgments (EA, MV), the bottom row shows CBN-based metrics (LOOCV  $R^2$ , Leak-Adjusted Determinacy LAD). Agents that exceed human baselines (higher EA, lower  $|MV|$ , higher  $R^2$ ) and higher (LAD) in the Numeric prompt-condition mostly belong to reasoning models are for example gemini-2.5-pro, the reasoning model o3. Smaller models (e.g., gpt-5-nano and -mini and gemini-2.5-flash-lite) seem to benefit most from CoT prompting.

#### 4.6 Q6: Do LLMs Reason Robustly Under Content Manipulations?

Additional agent results highlighting chain-of-thought impact and humans baseline comparison can be found in [Figures B.3 to B.5](#) [Section B.4](#) in Appendix. For a full comparison across experimental conditions additional scatter plots per agents can be found in [Section B.5](#) in [Figures B.6 to B.12](#). Parameter shifts by experiment (e.g., RW17  $\rightarrow$  Abstract) are plotted in [Section B.6](#); horizontal lines indicate robustness. For prompt-type transitions (Numeric  $\rightarrow$  chain-of-thought) can be found in [Section B.6.1](#) in [Figures B.16 to B.19](#) and [B.21 to B.23](#) and [Tables B.17 and B.18](#).



# Conclusion

---

5.1	Summary . . . . .	39
5.2	Limitations . . . . .	40
5.3	Future Work . . . . .	40

---

## 5.1 Summary

Large language models are increasingly integrated into decision-making workflows in high-stakes scenarios, such as in medical, judicial and financial domains. This necessitates thorough assessment of the causal reasoning capabilities of such models, both to understand the limits of their applicability and to prevent associative reasoning fallacies.

In this work, we introduced a causal reasoning benchmark with human comparison data and a cognitively-grounded diagnostic framework for evaluating causal reasoning in LLMs. Our cognitive lens—fitting simple causal Bayesian networks to model causal judgments—provides a compact way to understand to what degree LLMs are able to reason causally rather than merely associatively and how the prompt manipulations shape that trade-off. Our framework yields interpretable parameters (leak  $b$ , causal strengths  $m_1, m_2$ , priors  $p(C_i)$ ) that distinguish reasoning strategies along a spectrum from deterministic rule-following to probabilistic association-driven inference.

By analyzing responses from 20+ LLMs and humans on eleven matched collider inference tasks, we addressed research questions about domain effects, human-LLM alignment, normative reasoning, reasoning consistency, cognitive strategies, and robustness to content and prompting manipulations. Overall, we find that most *frontier LLMs reason more normatively than humans*, also exhibiting stronger explaining away and better Markov compliance, while a minority show more probabilistic patterns than humans.

Specifically, we find that modern LLMs can implement causal computations that are stable across domains (Q1) and align more closely with humans when prompted with chain-of-thought (Q2).

## Chapter 5 Conclusion

Fitting simple, interpretable causal Bayesian networks captures most models' behavior well (Q3), while chain-of-thought increases reasoning consistency (Q4). Parameter signatures in causal Bayes nets reveal two regimes (Q5): a *normative/deterministic* regime (low leak, high causal strengths, symmetric fits) and an *associative/hedging* regime (higher leak, weaker causal strengths, asymmetric fits). The key robustness result of our work (Q6) is that content abstraction preserves consistency, while adding irrelevant information generally hurts reasoning performance. Chain-of-thought is broadly beneficial and is especially helpful when irrelevant information is included in the prompt, where it improves both normativity and consistency.

The combination of stable causal Bayes net fits, abstraction-agnostic reasoning, and the resilience to prompt noise provides evidence that Gemini-2.5-pro and Gemini-2.5-flash perform *genuine causal computations on our common-effect causal inference tasks* rather than surface pattern matching. Other LLMs tested showing slightly lower clusters in Figures B.6 to B.12 and 4.9, suggest to have at least some degree of genuine causal computations on our common-effect causal inference tasks. However, the degradation under irrelevant information highlights a robustness gap that must be part of any future benchmark in causal reasoning and our findings suggest that chain-of-thought prompting helps reverse this trend for many models.

### 5.2 Limitations

Perhaps the biggest limitation of our work is the sole focus on the common-effect graph, however this is for good reason as collider structures give rise to interesting causal reasoning patterns such as explaining away and it has been well established in decades of research that humans reveal systematic biases in precisely these structures. An important question we answered in this thesis is whether LLMs exhibit similar biases and the answer is mostly no. Alternative graph topologies such as chains, forks, and more complex graphs remain subject to future research.

While we do provide human baseline data compiled from Rehder and Waldmann [1] as part of our causal reasoning benchmark, we do not have human responses to either abstract prompts or prompts with irrelevant information added. This would be an exciting and relatively straightforward extension of this work.

Finally, the architecture, activations, training datasets and any (post-)training adjustments are kept proprietary by OpenAI, Anthropic and Google whose LLMs we consider. Therefore, it is challenging to disentangle confounders like whether some models have been explicitly trained on abstract reasoning tasks or permutations facilitating generalization to abstract content.

### 5.3 Future Work

Our results and the limitations above suggest concrete next steps for future work. Our causal benchmark with the associated Python package can be applied to benchmark LLMs on additional causal reasoning tasks for which human data exists. For example on other graph topologies such as chains and forks. Given that chain-of-thought significantly improves smaller and older models and helps particularly in the setting where the prompt is overloaded with irrelevant content, it would be highly-interesting to investigate the activations of an open-source model during the forward pass with and without chain-of-thought. Finally, evaluating human reasoning more



### 5.3 Future Work

closely in settings that mirror more closely what the weaknesses of current frontier LLMs are, for example the case of overloaded prompts, could reveal interesting differences between them and further inform when LLMs are useful to augment human decision-making.



# Bibliography

- [1] B. Rehder and M. R. Waldmann. “Failures of explaining away and screening off in described versus experienced causal learning scenarios”. In: *Memory & Cognition* 45 (2017), pp. 245–260 (cit. on pp. i, viii, ix, 9, 11–13, 22, 30, 40).
- [2] S. Danziger, J. Levav, and L. Avnaim-Pesso. “Extraneous factors in judicial decisions”. In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 6889–6892. DOI: [10.1073/pnas.1018033108](https://doi.org/10.1073/pnas.1018033108) (cit. on p. 1).
- [3] J. A. Linder, J. N. Doctor, M. W. Friedberg, H. Reyes Nieva, C. Birks, D. Meeker, and C. R. Fox. “Time of Day and the Decision to Prescribe Antibiotics”. In: *JAMA Internal Medicine* 174.12 (2014), pp. 2029–2031. DOI: [10.1001/jamainternmed.2014.5225](https://doi.org/10.1001/jamainternmed.2014.5225) (cit. on p. 2).
- [4] V. Venkatraman, Y. M. L. Chuah, S. A. Huettel, and M. W. L. Chee. “Sleep Deprivation Elevates Expectation of Gains and Attenuates Response to Losses Following Risky Decisions”. In: *Sleep* 30.5 (2007), pp. 603–609. DOI: [10.1093/sleep/30.5.603](https://doi.org/10.1093/sleep/30.5.603) (cit. on p. 2).
- [5] C. M. Barnes, B. C. Gunia, and D. T. Wagner. “Sleep and moral awareness”. In: *Journal of Sleep Research* 24.2 (2015), pp. 181–188. DOI: [10.1111/jsr.12231](https://doi.org/10.1111/jsr.12231) (cit. on p. 2).
- [6] M. T. Wittbrodt and M. Millard-Stafford. “Dehydration Impairs Cognitive Performance: A Meta-analysis”. In: *Medicine & Science in Sports & Exercise* 50.11 (2018), pp. 2360–2368. DOI: [10.1249/MSS.0000000000001682](https://doi.org/10.1249/MSS.0000000000001682) (cit. on p. 2).
- [7] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho. “Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models”. In: *Journal of Legal Analysis* 16.1 (Jan. 2024). ISSN: 2161-7201, 1946-5319. DOI: [10.1093/jla/laae003](https://doi.org/10.1093/jla/laae003). URL: <http://arxiv.org/abs/2401.01301> (cit. on p. 2).
- [8] L. Riedemann, M. Labonne, and S. Gilbert. “The path forward for large language models in medicine is open”. In: *npj Digital Medicine* 7.1 (2024), p. 339 (cit. on p. 2).
- [9] M. Willig, M. Zecevic, D. S. Dhami, and K. Kersting. “Causal parrots: Large language models may talk causality but are not causal”. In: *arXiv preprint arXiv:2308.13067* 8 (2023) (cit. on pp. 2, 4).
- [10] E. Bigelow, J. Hu, E. S. Lubana, K. Gandhi, L. Ruis, T. Fel, E. Pavlick, and N. Goodman. *CogInterp: Interpreting Cognition in Deep Learning Models*. <https://coginterp.github.io/neurips2025/>. Workshop at NeurIPS 2025. 2025 (cit. on p. 4).

## Bibliography

- [11] K. Gandhi et al. “Human-like affective cognition in foundation models”. In: *arXiv preprint arXiv:2409.11733* (2024) (cit. on p. 4).
- [12] A. K. Lampinen et al. “Language models, like humans, show content effects on reasoning tasks”. In: *PNAS Nexus* 3.7 (July 2024), pgae233. ISSN: 2752-6542. DOI: [10.1093/pnasnexus/pgae233](https://doi.org/10.1093/pnasnexus/pgae233). URL: <https://doi.org/10.1093/pnasnexus/pgae233> (cit. on p. 4).
- [13] A. Keshmirian, M. Willig, B. Hemmatian, U. Hahn, K. Kersting, and T. Gerstenberg. “Biased Causal Strength Judgments in Humans and Large Language Models”. In: *ICLR 2024 Workshop on Representational Alignment*. 2024. URL: <https://openreview.net/forum?id=544P6YidFk> (cit. on p. 4).
- [14] H. M. Dettki, B. M. Lake, C. M. Wu, and B. Rehder. “Do Large Language Models Reason Causally Like Us? Even Better?” In: *Annual Conference of the Cognitive Science Society* (2025) (cit. on pp. 4, 8, 13).
- [15] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. “Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models”. In: *arXiv preprint arXiv:2410.05229* (2024) (cit. on pp. 4, 5).
- [16] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. G. Adauto, M. Kleiman-Weiner, M. Sachan, et al. “Cladder: A benchmark to assess causal reasoning capabilities of language models”. In: *arXiv preprint arXiv:2312.04350* (2023) (cit. on p. 4).
- [17] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. “Large Language Models Can Be Easily Distracted by Irrelevant Context”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 31210–31227. URL: <https://proceedings.mlr.press/v202/shi23a.html> (cit. on p. 4).
- [18] J. Pearl. “From Bayesian networks to causal networks”. In: *Mathematical models for handling partial knowledge in artificial intelligence*. Springer, 1995, pp. 157–182 (cit. on p. 7).
- [19] B. M. Rottman and R. Hastie. “Reasoning about causal relationships: Inferences on causal networks.” In: *Psychological bulletin* 140.1 (2014), p. 109 (cit. on pp. 7, 9).
- [20] M. Waldmann. *The Oxford handbook of causal reasoning*. Oxford University Press, 2017 (cit. on pp. 7, 8).
- [21] M. R. Waldmann. “Knowledge-based causal induction”. In: *Psychology of Learning and Motivation*. Vol. 34. Elsevier, 1996, pp. 47–88 (cit. on p. 7).
- [22] P. W. Cheng. “From covariation to causation: A causal power theory.” In: *Psychological review* 104.2 (1997), p. 367 (cit. on p. 8).
- [23] M. Henrion. “Some Practical Issues in Constructing Belief Networks”. In: *Uncertainty in Artificial Intelligence 3*. Ed. by L. N. Kanal and J. F. Lemmer. Amsterdam: North-Holland, 1989, pp. 161–173 (cit. on p. 8).

- [24] F. J. Díez. “Parameter Adjustment in Bayes Networks: The Generalized Noisy OR-Gate”. In: *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence (UAI-93)*. Washington, DC: Morgan Kaufmann, 1993, pp. 99–105 (cit. on p. 8).
- [25] J. Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009 (cit. on pp. 8, 9).
- [26] T. L. Griffiths and J. B. Tenenbaum. “Structure and Strength in Causal Induction”. In: *Cognitive Psychology* 51.4 (2005), pp. 334–384. DOI: [10.1016/j.cogpsych.2005.05.004](https://doi.org/10.1016/j.cogpsych.2005.05.004) (cit. on p. 8).
- [27] C. G. Lucas and T. L. Griffiths. “Learning the form of causal relationships using hierarchical Bayesian models”. In: *Cognitive Science* 34.1 (2010), pp. 113–147 (cit. on p. 8).
- [28] M. R. Waldmann. “Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules”. In: *Cognitive science* 31.2 (2007), pp. 233–256 (cit. on p. 8).
- [29] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., 2018 (cit. on p. 9).
- [30] P. M. Fernbach and B. Rehder. “Cognitive shortcuts in causal inference”. In: *Argument & Computation* 4.1 (2013), pp. 64–88 (cit. on p. 9).
- [31] N. Ali, N. Chater, and M. Oaksford. “The mental representation of causal conditional reasoning: Mental models or causal models”. In: *Cognition* 119.3 (2011), pp. 403–418 (cit. on p. 9).
- [32] R. Mayrhofer and M. R. Waldmann. “Sufficiency and necessity assumptions in causal structure induction”. In: *Cognitive science* 40.8 (2016), pp. 2137–2150 (cit. on p. 9).
- [33] J. Park and S. A. Sloman. “Mechanistic beliefs determine adherence to the Markov property in causal reasoning”. In: *Cognitive psychology* 67.4 (2013), pp. 186–216 (cit. on p. 9).
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. 2017 (cit. on p. 10).
- [35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of NAACL-HLT*. 2019 (cit. on p. 10).
- [36] J. Kaplan, S. McCandlish, T. Henighan, et al. “Scaling Laws for Neural Language Models”. In: *arXiv preprint arXiv:2001.08361* (2020) (cit. on p. 10).
- [37] J. Hoffmann, S. Borgeaud, A. Mensch, et al. “Training Compute-Optimal Large Language Models”. In: *arXiv preprint arXiv:2203.15556* (2022) (cit. on p. 10).
- [38] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. 2020 (cit. on p. 10).
- [39] L. Ouyang, J. Wu, X. Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *Advances in Neural Information Processing Systems*. 2022 (cit. on p. 10).

## Bibliography

- [40] P. Lewis, E. Perez, A. Piktus, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP”. In: *Advances in Neural Information Processing Systems*. 2020 (cit. on p. 10).
- [41] P. W. Cheng. “From Covariation to Causation: A Causal Power Theory”. In: *Psychological Review* 104.2 (1997), pp. 367–405. DOI: [10 . 1037 / 0033 - 295X . 104 . 2 . 367](https://doi.org/10.1037/0033-295X.104.2.367) (cit. on pp. 35, 74–80).

# Analysis – Details

---

A.1	Overloaded / Noisy Prompt Generation and Prompt Variants . . . . .	47
A.1.1	Overloaded and Abstract Prompts . . . . .	47
A.1.2	Abstract prompts (contrast) . . . . .	51
A.1.3	Prompt-categories (LLM-Output Instructions Chain-of-Thought vs. numeric) . . . . .	52
A.2	Human-LLM Alignment Details . . . . .	53
A.3	Derivation of Noisy Or Model’s Predicted Probability for each Task I-XI . . . .	53
A.3.1	Notation . . . . .	53
A.3.2	Predictive Inference . . . . .	54
A.3.3	Independence of Causes . . . . .	54
A.3.4	Diagnostic Inference – Effect Present . . . . .	55
A.3.5	Diagnostic Inference – Effect Absent . . . . .	56
A.4	LLM Details . . . . .	56

---

## A.1 Overloaded / Noisy Prompt Generation and Prompt Variants

### A.1.1 Overloaded and Abstract Prompts

This section documents how we construct overloaded variants of the RW17 prompts and how these compare to domain-agnostic abstract prompts.

**RW17 baseline scaffolding (recap)** The original RW17 prompts follow a fixed scaffold: (1) a short domain introduction, (2) detailed descriptions for the three variables  $C_1$ ,  $C_2$ , and  $E$  (e.g., “Interest rates are ...”), (3) a causal mechanism paragraph introduced by “Here are the causal relationships:” with two edge explanations (one for  $C_1 \rightarrow E$ , one for  $C_2 \rightarrow E$ ), and (4) an inference task instruction with counterbalancing. We preserve this scaffold exactly in all

## Appendix A Analysis – Details

overloaded variants; only additional material is appended at specific attachment points described below.

### Overloaded RW17 prompts

We generate overloaded variants by injecting irrelevant content to a clone of the base RW17 domain and append either cross-domain content (e.g content from the sociology domain to the weather domain) or neutral filler (from the lorem ipsum vocabulary<sup>1</sup>). Clones are named to reflect the manipulation and source domain, e.g., `econ_ovl_d=soc` (economy with sociology content appended to detailed fields), or `weath_ovl_e=econ`. Control variants use length-matched neutral filler and are suffixed with `=ctl` (e.g., `soc_ovl_de=ctl`).

**Attachment points and variants** CAUSAIIGN supports three injection points to inject irrelevant text and also supports tailored content.

- **d** (detailed fields only): Appends content to the *variable descriptions* for  $C_1$ ,  $C_2$ , and  $E$ . Concretely, this text appears in the block *before* the sentence “Here are the causal relationships:” and after the baseline RW17 description sentences for each variable. The causal mechanism sentences are left unchanged.
- **e** (edge explanations only): Appends content to the *causal mechanism sentences* that follow “Here are the causal relationships:”. The verbalizer, which reads from  $C_{1,2} \rightarrow E$ , picks up the added text and appends it to the causal mechanism native to that domain. Variable descriptions remain unchanged.
- **de** (both detailed and edge explanations): Applies both of the above simultaneously: variable descriptions (d) and edge explanations (e) receive appended content.

In this work, we only report results for the e condition, as we have the most LLM data for this condition, referred to as RW17-Over in the main text.

**Content vs. length controls** For each of d/e/de we provide control clones (`=ctl`) that append *neutral filler* instead of cross-domain content. Filler is domain-agnostic and generated from a lorem-ipsum style vocabulary to match the requested word length. It is sentence-cased and ends with a period to minimally disturb surface formatting. Thus, `=ctl` variants control for increased text *length* without introducing extraneous *content*. This allows clean separation of distraction-by-content from mere text-length effects.

---

<sup>1</sup><https://www.lipsum.com>



## A.1 Overloaded / Noisy Prompt Generation and Prompt Variants

### Example: de control (econ\_ovl\_de=ctl)

This example shows a *de-control* prompt in the Economy domain. It preserves the baseline RW17 scaffold and appends neutral, length-matched filler (in italics) to both variable descriptions (d) and edge explanations (e). Only the *italized* text is part of the actual prompt. The **bold** text is there to help the reader understand the structure of the prompt. **Domain introduction** *Economists seek to describe and predict the regular patterns of economic fluctuation. To do this, they study some important variables or attributes of economies. They also study how these attributes are responsible for producing or causing one another.*

**Variable descriptions (with *d* distractions) X (Interest rates).** *Interest rates are the rates banks charge to loan money. Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut.*

**Y (Trade deficit).** *A country's trade deficit is the difference between the value of the goods that a country imports and the value of the goods that a country exports. Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut.*

**Z (Retirement savings).** *Retirement savings is the money people save for their retirement. Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut.*

**Causal relationships (with *e* distractions)** *Here are the causal relationships:*

- *Low interest rates causes high retirement savings. Low interest rates stimulate economic growth, leading to greater prosperity overall, and allowing more money to be saved for retirement in particular. Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut enim ad minim veniam quis nostrud exercitation ullamco laboris nisi.*
- *Small trade deficits causes high retirement savings. When the economy is good, people can cover their basic expenses and so have enough money left over to contribute to their retirement accounts. Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut enim ad minim veniam quis nostrud exercitation ullamco laboris nisi.*

**Observations** *You are currently observing: high retirement savings and small trade deficits.*

**Task instruction** *Your task is to estimate how likely it is that low interest rates are present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Note that each of the causes can bring about the effect independently. Please provide your answer as a single number between 0 and 100, where 0 means very unlikely and 100 means very likely. Do not include any explanations or additional text.*

### Example: **de cross-domain** (weath\_ovl\_de=econ), **numeric, collider**

This example shows a *de cross-domain* prompt in the Weather domain with Economy content appended to variable descriptions (d) and edge explanations (e) serving as irrelevant distractions. It preserves the baseline RW17 scaffold and appends cross-domain, length-matched content (marked with ) to both variable descriptions (d) and edge explanations (e). Only the *italicized* text is part of the actual prompt. The **bold** text is there to help the reader understand the structure of the prompt.

**Domain introduction** *Meteorologists seek to describe and predict the regular patterns that govern weather systems. To do this, they study some important variables or attributes of weather systems. They also study how these attributes are responsible for producing or causing one another.*

**Variable descriptions (with *d* distractions) X (Ozone).** *Ozone is a gaseous allotrope of oxygen (O<sub>3</sub>) and is formed by exposure to UV radiation.*

*Some systems have high ozone levels. Others have normal ozone levels.*

*Interest rates are the rates banks charge to loan money.*

**Y (Air pressure).** *Air pressure is force exerted due to concentrations of air molecules.*

*Some systems have high air pressure. Others have normal air pressure.*

*A country's trade deficit is the difference between the value of the goods that a country imports and the value of the goods that a country exports.*

**Z (Humidity).** *Humidity is the degree to which the atmosphere contains water molecules.*

*Some systems have low humidity. Others have normal humidity.*

*Retirement savings is the money people save for their retirement.*

**Causal relationships (with *e* distractions)** *Here are the causal relationships:*

- *High ozone levels causes low humidity. Ozone attracts extra oxygen atoms from water molecules, creating a concentration of water vapor in that region. The good economic times produced by the low interest rates leads to greater confidence and less worry about the future, so people are less concerned about retirement.*
- *High air pressure causes low humidity. When air pressure is high, water vapor condenses into liquid water (rain), and the atmosphere is left with little moisture. The loss of local manufacturing jobs means that there are people out of work, and contributions to retirement accounts decreases.*

**Observations** *You are currently observing: normal humidity and high air pressure.*

**Task instruction** *Your task is to estimate how likely it is that high ozone levels are present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Note that each of the causes can bring about the effect independently. Please provide your answer as a single number between 0 and 100, where 0 means very unlikely and 100 means very likely. Do not include any explanations or additional text.*

### A.1.2 Abstract prompts (contrast)

As a complementary control, we generate *abstract* prompts that drop realistic RW17 domain semantics entirely while preserving causal structure and response formatting used in the RW11 variants.

#### Example: abstract prompt (numeric, collider; id 62)

**Domain introduction** *In abstract reasoning studies, researchers examine relationships between symbolic variables  $u8jzPde0Ig$ ,  $xLd6GncfBA$ , and  $epfJBd0Kh8$ .*

**Variable descriptions** *Some systems have high  $u8jzPde0Ig$ . Others have low  $u8jzPde0Ig$ . Some systems have weak  $xLd6GncfBA$ . Others have strong  $xLd6GncfBA$ . Some systems have weak  $epfJBd0Kh8$ . Others have powerful  $epfJBd0Kh8$ .*

**Causal relationships** *Here are the causal relationships:*

- *High  $u8jzPde0Ig$  causes weak  $epfJBd0Kh8$ .*
- *Weak  $xLd6GncfBA$  causes weak  $epfJBd0Kh8$ .*

**Observations** *You are currently observing: weak  $epfJBd0Kh8$  and high  $u8jzPde0Ig$ .*

**Task instruction** *Your task is to estimate how likely it is that weak  $xLd6GncfBA$  is present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Note that each of the causes can bring about the effect independently. Please provide your answer as a single number between 0 and 100, where 0 means very unlikely and 100 means very likely. Do not include any explanations or additional text.*

#### Example: Abstract Overloaded prompt (DE)

This example shows a *de overloaded* prompt in the *abstract* setting which is the one reported in the main text and referred to as Abstract-Over. It preserves the abstract scaffold and appends *neutral, length-matched filler* to both variable descriptions (d) and edge explanations (e). Only the *italicized* text is part of the actual prompt.<sup>2</sup>

**Domain introduction** *In abstract reasoning studies, researchers examine relationships between symbolic variables  $u8jzPde0Ig$ ,  $xLd6GncfBA$ , and  $epfJBd0Kh8$ .*

**Variable descriptions (with *d* distractions) X** ( $u8jzPde0Ig$ ). *Some systems have high  $u8jzPde0Ig$ . Others have low  $u8jzPde0Ig$ . Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.*

**Y** ( $xLd6GncfBA$ ). *Some systems have weak  $xLd6GncfBA$ . Others have strong  $xLd6GncfBA$ . Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.*

<sup>2</sup>The **bold** text is there to help the reader understand the structure of the prompt. Only the italicized text is part of the actual prompt.

## Appendix A Analysis – Details

**Z** ( $\text{epfJBd0Kh8}$ ). Some systems have weak  $\text{epfJBd0Kh8}$ . Others have powerful  $\text{epfJBd0Kh8}$ . *Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.*

**Causal relationships (with  $e$  distractions)** Here are the causal relationships:

- High  $\text{u8jzPd0Ig}$  causes weak  $\text{epfJBd0Kh8}$ . *Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut enim ad minim veniam.*
- Weak  $\text{xLd6GncfBA}$  causes weak  $\text{epfJBd0Kh8}$ . *Lorem ipsum dolor sit amet consectetur adipiscing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua ut enim ad minim veniam.*

**Observations** You are currently observing: weak  $\text{epfJBd0Kh8}$  and high  $\text{u8jzPd0Ig}$ .

**Task instruction** Your task is to estimate how likely it is that weak  $\text{xLd6GncfBA}$  is present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Note that each of the causes can bring about the effect independently. Please provide your answer as a single number between 0 and 100, where 0 means very unlikely and 100 means very likely. Do not include any explanations or additional text.

### A.1.3 Prompt-categories (LLM-Output Instructions Chain-of-Thought vs. numeric)

We use two prompt-categories that differ only in the required output format and whether the model should explain its reasoning. All categories keep the same causal scaffold and observations; only the output instruction changes.

**Zero-shot Answer (numeric ●)** Output: a single number between 0 and 100 indicating the likelihood. No explanation or extra text is allowed.

**Example instruction** Your task is to estimate how likely it is that strong  $B$  is present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Please provide your answer as a single number between 0 and 100, where 0 means very unlikely and 100 means very likely. Do not include any explanations or additional text.

**CoT (Chain-of-Thought ●)** Output: step-by-step explanation in an XML format followed by the likelihood judgment.

**Example instruction** Your task is to estimate how likely it is that weak  $\text{h\#Bel3liEl}$  is present on a scale from 0 to 100, given the observations and causal relationships described. 0 means completely unlikely and 100 means completely likely. Note that each of the causes can bring about the

effect independently. First, think through this step by step and explain your reasoning. Then provide your likelihood estimate. Return your response as raw text in one single line using this exact XML format: `<response><explanation>YOUR_STEP_BY_STEP_REASONING</explanation><likelihood>YOUR_NUMERIC_RESPONSE_HERE</likelihood></response>`.

Replace `YOUR_STEP_BY_STEP_REASONING` with your concise reasoning process.

Replace `YOUR_NUMERIC_RESPONSE_HERE` with your likelihood estimate between 0 (very unlikely) and 100 (very likely). DO NOT include any other information, explanation, or formatting outside the XML. DO NOT use Markdown, code blocks, quotation marks, or special characters.

## A.2 Human-LLM Alignment Details

For each agent  $a$  and domain  $d$ , we quantify human-LLM alignment by computing Spearman’s rank correlation coefficient  $\rho_{a,d}$  between human likelihood judgments and model predictions.

To assess the robustness of these correlations, we estimate 95% confidence intervals via non-parametric bootstrapping with  $B = 2000$  resamples. Specifically, for each bootstrap replicate  $b = 1, \dots, B$ , we draw a resampled index set  $\mathcal{I}_{a,d}^{(b)}$  of size  $|\mathcal{I}_{a,d}|$  by sampling with replacement from  $\mathcal{I}_{a,d}$ , and recompute

$$\rho_{a,d}^{(b)} = \text{corr}\left(\text{rank}(\{h_i\}_{i \in \mathcal{I}_{a,d}^{(b)}}), \text{rank}(\{m_i\}_{i \in \mathcal{I}_{a,d}^{(b)}})\right).$$

The percentile method then yields

$$\text{CI}_{95\%}(\rho_{a,d}) = \left[ \text{quantile}_{0.025}(\{\rho_{a,d}^{(b)}\}_{b=1}^B), \text{quantile}_{0.975}(\{\rho_{a,d}^{(b)}\}_{b=1}^B) \right].$$

In addition to domain-specific alignment for RW17, we report pooled alignment across all domains for each agent. Let  $\mathcal{D}$  denote the set of domains; the pooled index set is

$$\mathcal{I}_{a,\text{pool}} = \bigcup_{d \in \mathcal{D}} \mathcal{I}_{a,d},$$

with  $\rho_{a,\text{pool}}$  and its confidence interval computed in the same manner.

## A.3 Derivation of Noisy Or Model’s Predicted Probability for each Task I-XI

### A.3.1 Notation

- BR: Bayes rule
- PR: Product rule
- M: Marginalization

$$p(C_1) = \sum_{C_2} p(C_1, C_2)$$

## Appendix A Analysis – Details

- IA: Independence assumption

$$p(C_1, C_2) = p(C_1)p(C_2)$$

- MD: model definition (Noisy Or)

$$p(E = 1|C_1, C_2) = 1 - (1 - b)(1 - m_1^{C_1})(1 - m_2^{C_2})$$

### A.3.2 Predictive Inference

#### Task I:

$$p(E = 1|C_1 = 0, C_2 = 0)$$

$$p(E = 1|C_1 = 0, C_2 = 0) \stackrel{MD}{=} b \quad (\text{A.1})$$

#### Task II:

$$p(E = 1|C_1 = 0, C_2 = 1)$$

$$p(E = 1|C_1 = 0, C_2 = 1) \stackrel{MD}{=} 1 - (1 - b)(1 - m_2) \quad (\text{A.2})$$

#### Task III:

$$p(E = 1|C_1 = 1, C_2 = 1)$$

$$p(E = 1|C_1 = 1, C_2 = 1) \stackrel{MD}{=} 1 - (1 - b)(1 - m_1)(1 - m_2) \quad (\text{A.3})$$

### A.3.3 Independence of Causes

#### Task IV:

$$p(C_1 = 1|C_2 = 1)$$

$$p(C_1 = 1|C_2 = 1) \stackrel{BR}{=} \frac{p(C_1 = 1, C_2 = 1)}{p(C_2 = 1)} \quad (\text{A.4})$$

$$\stackrel{IA}{=} \frac{p(C_1 = 1)p(C_2 = 1)}{p(C_2 = 1)} \quad (\text{A.5})$$

$$\stackrel{MD}{=} p(C_1) \quad (\text{A.6})$$

## A.3 Derivation of Noisy Or Model's Predicted Probability for each Task I-XI

### Task V:

$$p(C_1 = 1 | C_2 = 0)$$

$$p(C_1 = 1 | C_2 = 0) \stackrel{BR}{=} \frac{p(C_1 = 1, C_2 = 0)}{p(C_2 = 0)} \quad (A.7)$$

$$\stackrel{IA}{=} \frac{p(C_1 = 1)(1 - p(C_2))}{1 - p(C_2)} \quad (A.8)$$

$$\stackrel{MD}{=} p(C_1) \quad (A.9)$$

### A.3.4 Diagnostic Inference – Effect Present

#### Task VI:

$$p(C_1 = 1 | E = 1, C_2 = 1)$$

$$p(C_1 = 1 | E = 1, C_2 = 1) \stackrel{BR}{=} \frac{p(E = 1 | C_1 = 1, C_2 = 1)p(C_1 = 1 | C_2 = 1)}{p(E = 1 | C_2 = 1)} \quad (A.10)$$

$$\stackrel{M+IA}{=} \frac{p(E = 1 | C_1 = 1, C_2 = 1)p(C_1 = 1)}{p(E = 1 | C_1 = 1, C_2 = 1)p(C_1 = 1) + p(E = 1 | C_1 = 0, C_2 = 1)p(C_1 = 0)} \quad (A.11)$$

$$\stackrel{MD}{=} \frac{[1 - (1 - b)(1 - m_1)(1 - m_2)]p(C_1)}{[1 - (1 - b)(1 - m_1)(1 - m_2)]p(C_1) + [1 - (1 - b)(1 - m_2)]p(C_1 = 0)} \quad (A.12)$$

#### Task VII:

$$p(C_1 = 1 | E = 1)$$

$$p(C_1 = 1 | E = 1) \stackrel{BR}{=} \frac{p(E = 1 | C_1 = 1)p(C_1 = 1)}{p(E = 1)} \quad (A.13)$$

$$\stackrel{M+IA}{=} \frac{p(C_1 = 1) \sum_{C_2} p(E = 1 | C_1 = 1, C_2)p(C_2)}{\sum_{C_1} \sum_{C_2} p(E = 1 | C_1, C_2)p(C_1)p(C_2)} \quad (A.14)$$

$$\stackrel{MD}{=} \frac{p(C_1)[p(C_2)(1 - (1 - b)(1 - m_1)(1 - m_2)) + (1 - p(C_2))(1 - (1 - b)(1 - m_1))]}{Z} \quad (A.15)$$

where  $Z$  is the evidence summing over all combinations normalizing the fraction:

$$Z = p(E = 1) = \sum_{C_1} \sum_{C_2} p(E = 1 | C_1, C_2)p(C_1)p(C_2) \quad (A.16)$$

$$= p(E = 1 | C_1 = 1, C_2 = 1)p(C_1 = 1)p(C_2 = 1) + \quad (A.17)$$

$$p(E = 1 | C_1 = 1, C_2 = 0)p(C_1 = 1)p(C_2 = 0) + \quad (A.18)$$

$$p(E = 1 | C_1 = 0, C_2 = 1)p(C_1 = 0)p(C_2 = 1) + \quad (A.19)$$

$$p(E = 1 | C_1 = 0, C_2 = 0)p(C_1 = 0)p(C_2 = 0) \quad (A.20)$$

Substituting our model parameterization:

$$Z = [1 - (1 - b)(1 - m_1)(1 - m_2)]p(C_1)p(C_2) + \quad (A.21)$$

$$[1 - (1 - b)(1 - m_1)]p(C_1)(1 - p(C_2)) + \quad (A.22)$$

$$1 - (1 - b)(1 - m_2)p(C_2) + \quad (A.23)$$

$$b(1 - p(C_1))(1 - p(C_2)) \quad (A.24)$$

## Appendix A Analysis – Details

### Task VIII:

$$p(C_1 = 1 | E = 1, C_2 = 0)$$

$$p(C_1 = 1 | E = 1, C_2 = 0) \stackrel{BR}{=} \frac{p(E = 1 | C_1 = 1, C_2 = 0)p(C_1 = 1 | C_2 = 0)}{p(E = 1, C_2 = 0)} \quad (\text{A.25})$$

$$\stackrel{M+IA}{=} \frac{p(E = 1 | C_1 = 1, C_2 = 0)p(C_1 = 1)}{p(E = 1 | C_1 = 1, C_2 = 0)p(C_1 = 1) + p(E = 1 | C_1 = 0, C_2 = 0)p(C_1 = 0)} \quad (\text{A.26})$$

$$\stackrel{MD}{=} \frac{[1 - (1 - b)(1 - m_1)]p(C_1)}{[1 - (1 - b)(1 - m_1)]p(C_1) + b(1 - p(C_1))} \quad (\text{A.27})$$

### A.3.5 Diagnostic Inference – Effect Absent

#### Task IX:

$$p(C_1 = 1 | E = 0, C_2 = 1)$$

$$p(C_1 = 1 | E = 0, C_2 = 1) \stackrel{BR}{=} \frac{p(E = 0 | C_1 = 1, C_2 = 1)p(C_1 = 1)}{p(E = 0, C_2 = 1)} \quad (\text{A.28})$$

$$\stackrel{M}{=} \frac{[1 - p(E = 1 | C_1 = 1, C_2 = 1)]p(C_1 = 1)}{[1 - p(E = 1 | C_1 = 1, C_2 = 1)]p(C_1 = 1) + [1 - p(E = 1 | C_1 = 0, C_2 = 1)]p(C_1 = 0)} \quad (\text{A.29})$$

$$\stackrel{MD}{=} \frac{[(1 - b)(1 - m_1)(1 - m_2)]p(C_1)}{[(1 - b)(1 - m_1)(1 - m_2)]p(C_1) + [(1 - b)(1 - m_2)](1 - p(C_1))} \quad (\text{A.30})$$

#### Task X:

$$p(C_1 = 1 | E = 0)$$

$$p(C_1 = 1 | E = 0) \stackrel{BR}{=} \frac{p(E = 0 | C_1 = 1)p(C_1 = 1)}{p(E = 0)} \quad (\text{A.31})$$

$$\stackrel{M+IA}{=} \frac{p(C_1 = 1) \sum_{C_2} p(E = 0 | C_1 = 1, C_2)p(C_2)}{\sum_{C_1} \sum_{C_2} p(E = 0 | C_1, C_2)p(C_1)p(C_2)} \quad (\text{A.32})$$

$$\stackrel{MD}{=} \frac{p(C_1)[p(C_2)(1 - b)(1 - m_1)(1 - m_2) + (1 - p(C_2))(1 - b)(1 - m_1)]}{Z} \quad (\text{A.33})$$

where  $Z$  is again the evidence normalizing the fraction (see Equation (A.16)).

#### Task XI:

$$p(C_1 = 1 | E = 0, C_2 = 0)$$

$$p(C_1 = 1 | E = 0, C_2 = 0) \stackrel{BR}{=} \frac{p(E = 0 | C_1 = 1, C_2 = 0)p(C_1 = 1 | C_2 = 0)}{p(E = 0 | C_2 = 0)} \quad (\text{A.34})$$

$$\stackrel{M+IA}{=} \frac{[1 - p(E = 1 | C_1 = 1, C_2 = 0)]p(C_1 = 1)}{[1 - p(E = 1 | C_1 = 1, C_2 = 0)]p(C_1 = 1) + [1 - p(E = 1 | C_1 = 0, C_2 = 0)]p(C_1 = 0)} \quad (\text{A.35})$$

$$\stackrel{MD}{=} \frac{[(1 - b)(1 - m_1)]p(C_1)}{[(1 - b)(1 - m_1)]p(C_1) + (1 - b)(1 - p(C_1))} \quad (\text{A.36})$$

## A.4 LLM Details



**Table A.1:** LLM release dates and context window sizes. Parameters are not publicly disclosed for. Context window sizes are as reported by the providers, with some variation depending on the specific version or subscription plan (e.g., enterprise plans). The context windows of all models exceed the maximum input length of our tasks (approximately 500 tokens). Note that some models have different context window sizes for input and output, which is indicated where applicable.

Provider	LLM-name	Context Window Size	Release date
Anthropic	claude-3-5-haiku-20241022	200K	2024-10-22
Anthropic	claude-3-7-sonnet-20250219	200K	2025-02-24
Anthropic	claude-3-haiku-20240307	200K	2024-03-13
Anthropic	claude-3-opus	200K	2024-03-04
Anthropic	claude-3-sonnet-20240229	200K	2024-02-29
Anthropic	claude-opus-4-1-20250805	200K (500K on Enterprise)	2025-08-05
Anthropic	claude-opus-4-20250514	200K (500K on Enterprise)	2025-05-22
Anthropic	claude-sonnet-4-20250514	200K (500K on Enterprise)	2025-05-22
Google	gemini-1.5-pro	1M	2024-02-15
Google	gemini-2.5-flash	1M	2025-05-14
Google	gemini-2.5-flash-lite	1M	2025-05-14
Google	gemini-2.5-pro	1M	2025-03-25
OpenAI	gpt-3.5-turbo	16K	2023-03-01
OpenAI	gpt-4	8K-128K (version dependent)	2023-03-14
OpenAI	gpt-4.1	1M	2025-04-14
OpenAI	gpt-4.1-mini	1M	2025-04-14
OpenAI	gpt-4o	128K	2024-05-13
OpenAI	gpt-5-mini-v_low-r_high	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-mini-v_low-r_low	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-mini-v_low-r_medium	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-mini-v_low-r_minimal	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-nano-v_low-r_high	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-nano-v_low-r_low	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-nano-v_low-r_medium	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-nano-v_low-r_minimal	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-v_low-r_low	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-v_low-r_medium	400K (272K input + 128K output)	2025-08-07
OpenAI	gpt-5-v_low-r_minimal	400K (272K input + 128K output)	2025-08-07
OpenAI	o1	8k-128k not clearly disclosed	2024-09-12
OpenAI	o1-mini	8k-128k not clearly disclosed	2024-09-12
OpenAI	o3	8k-200k not clearly disclosed	2025-04-16
OpenAI	o3-mini	8k-200k not clearly disclosed	2025-01-31



## Additional Results

---

B.1	Domain differences per experiment and prompt-style . . . . .	60
B.1.1	RW17 . . . . .	60
B.2	Human-LLM alignment: Domain-wise breakdowns for Chain-of-Thought prompts in comparison to Numeric prompts . . . . .	65
B.3	Additional Results for the distribution of likelihood judgements . . . . .	67
B.4	Additional Results for the effect of Chain-of-Thought prompts on causal reasoning across experiments . . . . .	69
B.5	Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explainaing Away, and Markov Violation . . . . .	73
B.5.1	Metrics by Release Date of LLMs . . . . .	81
B.6	Most and Least changing LLMs across prompt-category & content manipulations	83
B.6.1	Experiment-wise changes with fixed prompt-style (Numeric or CoT) . .	83
B.6.2	Prompt-wise changes with fixed experiment (e.g., RW17 or Abstract) .	89
B.7	Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments . . . . .	95
B.7.1	RW17 and Abstract, Numeric Prompts . . . . .	95
B.7.2	RW17 and Abstract, CoT Prompts . . . . .	98
B.7.3	RW17-Overloaded, Numeric and CoT Prompts . . . . .	101
B.7.4	Abstract-Overloaded, Numeric and CoT Prompts . . . . .	103

---

## B.1 Domain differences per experiment and prompt-style

**Kruskal–Wallis Test.** The Kruskal–Wallis test is a nonparametric method to compare the central tendencies of  $k \geq 3$  independent groups. All observations across groups are pooled and ranked from 1 to  $N$  (with average ranks for ties). Let  $R_i$  denote the sum of ranks in group  $i$  of size  $n_i$ . The test statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1), \quad (\text{B.1})$$

where  $N = \sum_{i=1}^k n_i$  is the total sample size.

Under the null hypothesis that all  $k$  domains come from the same distribution,  $H$  approximately follows a  $\chi^2$  distribution with  $k - 1$  degrees of freedom. Large values of  $H$  indicate that at least one group median differs from the others.

**Pairwise.** For every domain pair within an agent, we ran a two-sided Mann–Whitney  $U$  test. Null: the two domain distributions are identical; alternative: they differ. We report the rank-biserial effect size  $r = \frac{2U}{n_1 n_2} - 1 \in [-1, 1]$  (positive indicates the first domain tends higher). P-values were BH–FDR adjusted *within each agent* across all its domain pairs (default  $\alpha = 0.05$ ).

### B.1.1 RW17

Table B.1: Kruskal–Wallis across agents within each domain, RW17-Numeric prompts.

Domain	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
economy	29	165.86	28	$1.6 \times 10^{-21}$	$4.8 \times 10^{-21}$
sociology	29	157.78	28	$4.8 \times 10^{-20}$	$7.2 \times 10^{-20}$
weather	29	70.72	28	$1.5 \times 10^{-5}$	$1.5 \times 10^{-5}$

Table B.2: Kruskal–Wallis across agents within each domain. (RW17, Numeric prompts)

Domain	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
abs_all_10	28	93.27	27	$3.2 \times 10^{-9}$	$3.2 \times 10^{-9}$
abs_alnum_10	30	107.91	29	$5.1 \times 10^{-11}$	$7.7 \times 10^{-11}$
abs_num_symb_10	28	128.56	27	$3.5 \times 10^{-15}$	$1.0 \times 10^{-14}$

**Numeric prompts.**

**Numeric prompts.**

**CoT prompts.**

## B.1 Domain differences per experiment and prompt-style

Table B.3: Kruskal–Wallis across domains within each agent. (RW17, Numeric prompts)

Agent	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
claude-3-5-haiku-20241022	3	2.23	2	0.327	0.994
claude-3-7-sonnet-20250219	3	0.11	2	0.947	0.994
claude-3-haiku-20240307	3	1.50	2	0.472	0.994
claude-3-opus	3	3.84	2	0.146	0.994
claude-sonnet-4-20250514	3	1.17	2	0.557	0.994
gemini-1.5-pro	3	0.13	2	0.937	0.994
gemini-2.5-flash	3	0.17	2	0.918	0.994
gemini-2.5-flash-lite	3	0.12	2	0.940	0.994
gemini-2.5-pro	3	0.08	2	0.960	0.994
gpt-3.5-turbo	3	3.65	2	0.161	0.994
gpt-4	3	0.35	2	0.838	0.994
gpt-4.1	3	0.44	2	0.801	0.994
gpt-4.1-mini	3	0.05	2	0.974	0.994
gpt-4o	3	0.15	2	0.930	0.994
gpt-5-mini-v_low-r_high	3	0.06	2	0.970	0.994
gpt-5-mini-v_low-r_low	3	0.16	2	0.924	0.994
gpt-5-mini-v_low-r_medium	3	0.01	2	0.993	0.994
gpt-5-mini-v_low-r_minimal	3	0.11	2	0.946	0.994
gpt-5-nano-v_low-r_low	3	0.91	2	0.633	0.994
gpt-5-nano-v_low-r_medium	3	0.04	2	0.978	0.994
gpt-5-nano-v_low-r_minimal	3	5.79	2	0.055	0.994
gpt-5-v_low-r_low	3	0.24	2	0.887	0.994
gpt-5-v_low-r_medium	3	0.01	2	0.994	0.994
gpt-5-v_low-r_minimal	3	0.02	2	0.992	0.994
o1	3	0.09	2	0.955	0.994
o1-mini	3	0.02	2	0.989	0.994
o3	3	0.33	2	0.846	0.994
o3-mini	3	0.19	2	0.910	0.994

Table B.4: Kruskal–Wallis across agents within each domain. (RW17, CoT prompts)

Domain	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
abs_all_10	25	118.45	24	$1.8 \times 10^{-14}$	$2.7 \times 10^{-14}$
abs_alnum_10	26	114.64	25	$1.9 \times 10^{-13}$	$1.9 \times 10^{-13}$
abs_num_symb_10	25	126.32	24	$7.2 \times 10^{-16}$	$2.2 \times 10^{-15}$

## Appendix B Additional Results

Table B.5: Kruskal–Wallis across domains within each agent. (RW17, Numeric prompts)

Agent	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
claude-3-5-haiku-20241022	3	0.89	2	0.641	0.988
claude-3-7-sonnet-20250219	3	0.06	2	0.971	0.988
claude-3-haiku-20240307	3	5.33	2	0.070	0.988
claude-3-opus	3	0.94	2	0.624	0.988
claude-sonnet-4-20250514	3	0.08	2	0.960	0.988
gemini-1.5-pro	3	0.13	2	0.937	0.988
gemini-2.5-flash	3	0.09	2	0.958	0.988
gemini-2.5-flash-lite	3	0.22	2	0.894	0.988
gemini-2.5-pro	3	0.02	2	0.988	0.988
gpt-3.5-turbo	3	2.71	2	0.258	0.988
gpt-4	3	0.55	2	0.758	0.988
gpt-4.1	3	0.05	2	0.975	0.988
gpt-4.1-mini	3	0.08	2	0.963	0.988
gpt-4o	3	0.09	2	0.958	0.988
gpt-5-mini-v_low-r_low	3	0.40	2	0.819	0.988
gpt-5-mini-v_low-r_medium	3	0.06	2	0.972	0.988
gpt-5-mini-v_low-r_minimal	3	0.15	2	0.927	0.988
gpt-5-nano-v_low-r_low	3	1.46	2	0.483	0.988
gpt-5-nano-v_low-r_minimal	3	1.54	2	0.463	0.988
gpt-5-v_low-r_low	3	0.36	2	0.834	0.988
gpt-5-v_low-r_minimal	3	0.08	2	0.961	0.988
o1	3	0.18	2	0.916	0.988
o1-mini	3	0.30	2	0.861	0.988
o3	3	0.03	2	0.986	0.988
o3-mini	3	0.06	2	0.970	0.988

Table B.6: Kruskal–Wallis across agents within each domain. RW17-overloaded, Numeric

Domain	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
econ_ovl_e=ctl	20	188.39	19	$6.8 \times 10^{-30}$	$4.1 \times 10^{-29}$
econ_ovl_e=soc	20	184.71	19	$3.7 \times 10^{-29}$	$1.1 \times 10^{-28}$
soc_ovl_e=ctl	20	169.17	19	$4.1 \times 10^{-26}$	$6.2 \times 10^{-26}$
soc_ovl_e=econ	20	177.50	19	$9.6 \times 10^{-28}$	$1.9 \times 10^{-27}$
weath_ovl_e=ctl	20	124.52	19	$1.6 \times 10^{-17}$	$1.9 \times 10^{-17}$
weath_ovl_e=econ	20	124.11	19	$1.9 \times 10^{-17}$	$1.9 \times 10^{-17}$

## B.1 Domain differences per experiment and prompt-style

Table B.7: Kruskal–Wallis across domains within each agent. RW17-overloaded, Numeric

Agent	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
claude-3-5-haiku-20241022	6	5.19	5	0.393	0.999
claude-3-7-sonnet-20250219	6	2.62	5	0.758	0.999
claude-3-haiku-20240307	6	10.15	5	0.071	0.474
claude-3-opus	6	1.99	5	0.850	0.999
claude-sonnet-4-20250514	6	3.50	5	0.623	0.999
gemini-1.5-pro	6	2.78	5	0.734	0.999
gemini-2.5-flash	6	1.18	5	0.946	0.999
gemini-2.5-flash-lite	6	1.19	5	0.946	0.999
gemini-2.5-pro	6	0.24	5	0.999	0.999
gpt-3.5-turbo	6	10.54	5	0.061	0.474
gpt-4	6	1.62	5	0.899	0.999
gpt-4.1	6	1.01	5	0.961	0.999
gpt-4.1-mini	6	3.82	5	0.576	0.999
gpt-4o	6	0.56	5	0.990	0.999
gpt-5-mini-v_low-r_low	6	1.49	5	0.914	0.999
gpt-5-mini-v_low-r_minimal	6	13.54	5	0.019	0.377
gpt-5-nano-v_low-r_low	6	0.30	5	0.998	0.999
gpt-5-nano-v_low-r_minimal	6	0.91	5	0.970	0.999
gpt-5-v_low-r_low	6	0.68	5	0.984	0.999
gpt-5-v_low-r_minimal	6	3.72	5	0.591	0.999

Table B.8: Kruskal–Wallis across domains within each agent. RW17-overloaded, CoT

Domain	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
econ_ovl_e=ctl	19	90.17	18	$1.3 \times 10^{-11}$	$2.7 \times 10^{-11}$
econ_ovl_e=soc	21	81.49	20	$2.2 \times 10^{-9}$	$3.3 \times 10^{-9}$
soc_ovl_e=ctl	19	93.59	18	$3.2 \times 10^{-12}$	$1.9 \times 10^{-11}$
soc_ovl_e=econ	21	94.64	20	$1.1 \times 10^{-11}$	$2.7 \times 10^{-11}$
weath_ovl_e=ctl	18	64.64	17	$1.8 \times 10^{-7}$	$1.8 \times 10^{-7}$
weath_ovl_e=econ	21	71.06	20	$1.2 \times 10^{-7}$	$1.5 \times 10^{-7}$

Table B.9: Kruskal–Wallis across agents within each domain. RW17-overloaded, CoT

## Appendix B Additional Results

Table B.10: Kruskal–Wallis across agents within each domain. RW17-overloaded, CoT

Agent	$k$	$H$	df	$p$ (raw)	$p_{FDR}$
claude-3-5-haiku-20241022	6	3.60	5	0.609	0.988
claude-3-7-sonnet-20250219	6	0.79	5	0.977	0.988
claude-3-haiku-20240307	6	0.85	5	0.973	0.988
claude-3-opus	6	2.03	5	0.844	0.988
claude-sonnet-4-20250514	6	1.33	5	0.932	0.988
gemini-1.5-pro	6	2.40	5	0.791	0.988
gemini-2.5-flash	6	1.16	5	0.949	0.988
gemini-2.5-flash-lite	6	0.78	5	0.978	0.988
gemini-2.5-pro	6	0.60	5	0.988	0.988
gpt-3.5-turbo	6	2.82	5	0.728	0.988
gpt-4	3	0.92	2	0.630	0.988
gpt-4.1	6	0.77	5	0.979	0.988
gpt-4.1-mini	6	5.64	5	0.342	0.988
gpt-4o	6	2.29	5	0.808	0.988
gpt-5-mini-v_low-r_low	5	2.93	4	0.570	0.988
gpt-5-mini-v_low-r_medium	6	2.11	5	0.833	0.988
gpt-5-mini-v_low-r_minimal	6	4.71	5	0.452	0.988
gpt-5-nano-v_low-r_low	6	0.89	5	0.971	0.988
gpt-5-nano-v_low-r_minimal	6	6.33	5	0.276	0.988
gpt-5-v_low-r_low	3	0.07	2	0.963	0.988
gpt-5-v_low-r_minimal	6	0.96	5	0.966	0.988

Domain	$k$	$H$	df	$p$ (raw)	$p_{FDR}$
abs_all_10_overloaded_de	12	42.67	11	$1.2 \times 10^{-5}$	$1.2 \times 10^{-5}$
abs_alnum_10_overloaded_de	14	54.69	13	$4.6 \times 10^{-7}$	$6.9 \times 10^{-7}$
abs_num_symb_10_overloaded_de	12	77.60	11	$4.3 \times 10^{-12}$	$1.3 \times 10^{-11}$

Table B.11: Kruskal–Wallis across agents within each domain. RW17-overloaded, Numeric

Table B.12: Kruskal–Wallis across domains within each agent, Abstract Overloaded, Numeric.

Agent	$k$	$H$	df	$p$ (raw)	$p_{FDR}$
claude-3-5-haiku-20241022	3	11.16	2	0.004	0.045
claude-3-7-sonnet-20250219	3	0.16	2	0.925	0.994
claude-3-haiku-20240307	3	2.29	2	0.318	0.994
claude-3-opus	3	1.63	2	0.442	0.994
claude-sonnet-4-20250514	3	0.14	2	0.933	0.994
gemini-1.5-pro	3	0.12	2	0.942	0.994
gemini-2.5-flash	3	0.29	2	0.864	0.994
gemini-2.5-flash-lite	3	0.42	2	0.809	0.994
gemini-2.5-pro	3	0.04	2	0.981	0.994
gpt-5-mini-v_low-r_minimal	3	1.18	2	0.553	0.994
gpt-5-nano-v_low-r_minimal	3	2.14	2	0.344	0.994
gpt-5-v_low-r_minimal	3	0.01	2	0.994	0.994

Table B.13: Kruskal–Wallis across agents within each domain. RW17-overloaded, CoT

Domain	$k$	$H$	df	$p$ (raw)	$p_{FDR}$
abs_all_10_overloaded_de	12	57.96	11	$2.2 \times 10^{-8}$	$3.7 \times 10^{-8}$
abs_alnum_10_overloaded_de	12	51.63	11	$3.2 \times 10^{-7}$	$3.2 \times 10^{-7}$
abs_num_symb_10_overloaded_de	12	57.71	11	$2.5 \times 10^{-8}$	$3.7 \times 10^{-8}$



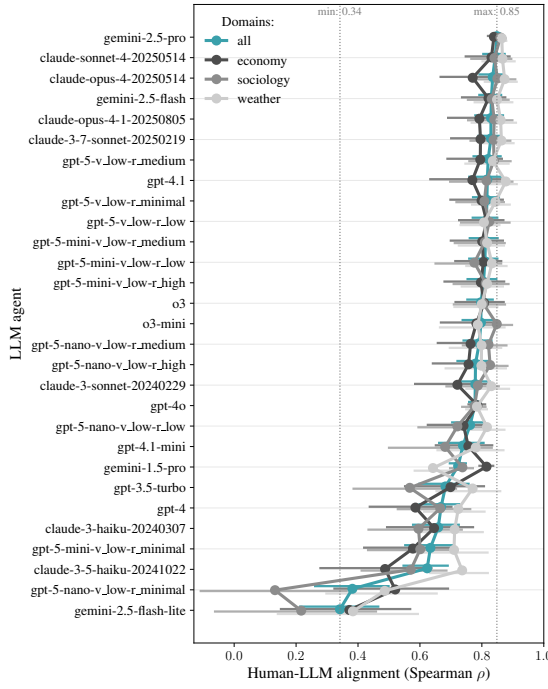
## B.2 Human-LLM alignment: Domain-wise breakdowns for Chain-of-Thought prompts in comparison to Numeric prompts

Table B.14: Kruskal–Wallis across domains within each agent. RW17-overloaded, CoT

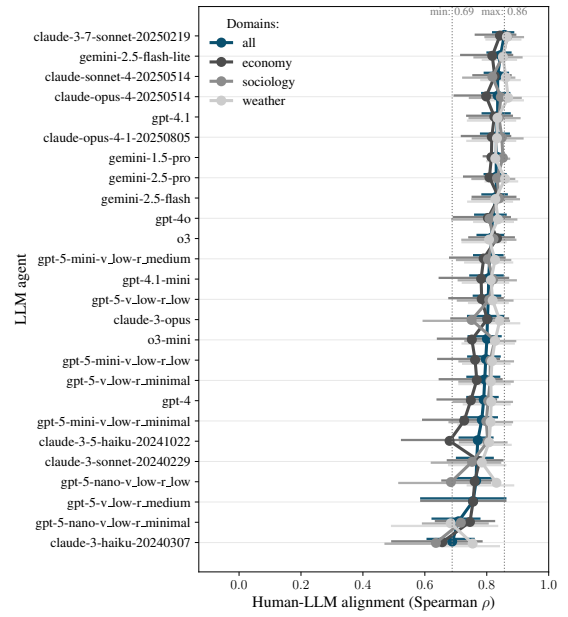
Agent	$k$	$H$	df	$p$ (raw)	$p_{\text{FDR}}$
claude-3-5-haiku-20241022	3	0.75	2	0.688	0.992
claude-3-7-sonnet-20250219	3	0.21	2	0.902	0.992
claude-3-haiku-20240307	3	0.76	2	0.683	0.992
claude-3-opus	3	0.55	2	0.758	0.992
claude-sonnet-4-20250514	3	0.10	2	0.953	0.992
gemini-1.5-pro	3	0.21	2	0.899	0.992
gemini-2.5-flash	3	0.04	2	0.982	0.992
gemini-2.5-flash-lite	3	0.04	2	0.979	0.992
gemini-2.5-pro	3	0.04	2	0.978	0.992
gpt-5-mini-v_low-r_minimal	3	0.02	2	0.992	0.992
gpt-5-nano-v_low-r_minimal	3	2.02	2	0.364	0.992
gpt-5-v_low-r_minimal	3	0.07	2	0.968	0.992

## B.2 Human-LLM alignment: Domain-wise breakdowns for Chain-of-Thought prompts in comparison to Numeric prompts

## Appendix B Additional Results



(a) Single Numeric Likelihood estimate only (Numeric prompts)



(b) Chain of Thought (CoT) prompts

**Figure B.1: Human-LLM alignment.** Each panel reports human-LLM alignment per domain (shades of gray) and pooled domains (red) with 95% bootstrapped confidence intervals sorted from highest to lowest  $\rho$ . Vertical dashed lines indicate the minimum and maximum pooled  $\rho$  values across agents.

### B.3 Additional Results for the distribution of likelihood judgements

**Table B.15:** Human–LLM alignment (Spearman rho) across domains for single-shot/numeric prompting. Agents are ordered by pooled domain alignment. Each cell reports the bootstrapped Spearman  $\rho$  and 95% confidence interval [lower, upper]. Uncertainty reflects the range of  $\rho$  values obtained by nonparametric bootstrapping (2,000 resamples) for each agent–domain pair.

Agent	Domain			
	pooled	economy	sociology	weather
gemini-2.5-pro	<b>0.849</b> [0.838, 0.858]	<b>0.839</b> [0.820, 0.855]	<b>0.860</b> [0.842, 0.877]	0.864 [0.848, 0.879]
claude-sonnet-4-20250514	0.843 [0.805, 0.874]	0.831 [0.748, 0.889]	0.841 [0.766, 0.895]	0.866 [0.802, 0.906]
claude-opus-4-20250514	0.835 [0.792, 0.870]	0.770 [0.667, 0.847]	0.850 [0.770, 0.908]	0.873 [0.810, 0.912]
gemini-2.5-flash	0.831 [0.792, 0.862]	0.821 [0.736, 0.876]	0.832 [0.755, 0.886]	0.849 [0.777, 0.898]
claude-opus-4-1-20250805	0.829 [0.779, 0.868]	0.792 [0.691, 0.864]	0.840 [0.756, 0.899]	0.861 [0.767, 0.910]
claude-3-7-sonnet-20250219	0.829 [0.789, 0.861]	0.796 [0.701, 0.856]	0.838 [0.764, 0.892]	0.863 [0.801, 0.902]
gpt-5-v_low-r_medium	0.821 [0.773, 0.860]	0.795 [0.689, 0.865]	0.839 [0.760, 0.893]	0.835 [0.747, 0.886]
gpt-4.1	0.818 [0.759, 0.862]	0.769 [0.634, 0.860]	0.815 [0.698, 0.899]	<b>0.877</b> [0.817, 0.913]
gpt-5-v_low-r_minimal	0.816 [0.772, 0.855]	0.799 [0.698, 0.868]	0.808 [0.718, 0.871]	0.844 [0.769, 0.890]
gpt-5-v_low-r_low	0.815 [0.771, 0.850]	0.815 [0.727, 0.869]	0.824 [0.733, 0.889]	0.807 [0.731, 0.857]
gpt-5-mini-v_low-r_medium	0.812 [0.761, 0.851]	0.801 [0.700, 0.868]	0.809 [0.713, 0.874]	0.817 [0.726, 0.871]
gpt-5-mini-v_low-r_low	0.810 [0.759, 0.850]	0.805 [0.714, 0.862]	0.776 [0.650, 0.862]	0.832 [0.750, 0.880]
gpt-5-mini-v_low-r_high	0.805 [0.754, 0.847]	0.797 [0.679, 0.872]	0.813 [0.709, 0.886]	0.817 [0.735, 0.868]
o3	0.797 [0.753, 0.835]	0.807 [0.715, 0.870]	0.805 [0.710, 0.875]	0.799 [0.711, 0.850]
o3-mini	0.797 [0.738, 0.846]	0.782 [0.668, 0.857]	0.848 [0.777, 0.897]	0.787 [0.665, 0.873]
gpt-5-nano-v_low-r_medium	0.792 [0.741, 0.831]	0.764 [0.658, 0.832]	0.821 [0.739, 0.879]	0.799 [0.697, 0.862]
gpt-5-nano-v_low-r_high	0.779 [0.721, 0.825]	0.757 [0.642, 0.832]	0.827 [0.750, 0.883]	0.799 [0.683, 0.877]
claude-3-sonnet-20240229	0.779 [0.723, 0.825]	0.721 [0.585, 0.815]	0.787 [0.686, 0.854]	0.831 [0.750, 0.888]
gpt-4o	0.778 [0.759, 0.796]	0.786 [0.757, 0.810]	0.776 [0.737, 0.812]	0.784 [0.750, 0.816]
gpt-5-nano-v_low-r_low	0.762 [0.704, 0.809]	0.741 [0.626, 0.818]	0.721 [0.595, 0.817]	0.816 [0.719, 0.873]
gpt-4.1-mini	0.739 [0.662, 0.805]	0.756 [0.651, 0.833]	0.681 [0.500, 0.832]	0.780 [0.657, 0.870]
gemini-1.5-pro	0.723 [0.697, 0.748]	0.815 [0.792, 0.837]	0.737 [0.697, 0.771]	0.642 [0.583, 0.694]
gpt-3.5-turbo	0.683 [0.598, 0.757]	0.699 [0.552, 0.807]	0.567 [0.386, 0.716]	0.770 [0.654, 0.859]
gpt-4	0.668 [0.599, 0.727]	0.585 [0.438, 0.702]	0.666 [0.529, 0.762]	0.724 [0.614, 0.809]
claude-3-haiku-20240307	0.659 [0.578, 0.725]	0.646 [0.494, 0.771]	0.595 [0.434, 0.735]	0.712 [0.598, 0.803]
gpt-5-mini-v_low-r_minimal	0.634 [0.553, 0.706]	0.577 [0.420, 0.704]	0.601 [0.433, 0.721]	0.710 [0.572, 0.818]
claude-3-5-haiku-20241022	0.623 [0.547, 0.690]	0.487 [0.279, 0.637]	0.571 [0.412, 0.686]	0.736 [0.633, 0.819]
gpt-5-nano-v_low-r_minimal	0.382 [0.262, 0.499]	0.520 [0.324, 0.691]	0.132 [-0.108, 0.363]	0.486 [0.298, 0.654]
gemini-2.5-flash-lite	0.342 [0.213, 0.465]	0.372 [0.151, 0.569]	0.217 [-0.062, 0.458]	0.383 [0.141, 0.593]

### B.3 Additional Results for the distribution of likelihood judgements

Likelihood judgements are often clustered around 0, 100, and 50 motivating the Huber loss for CBN fitting

## Appendix B Additional Results

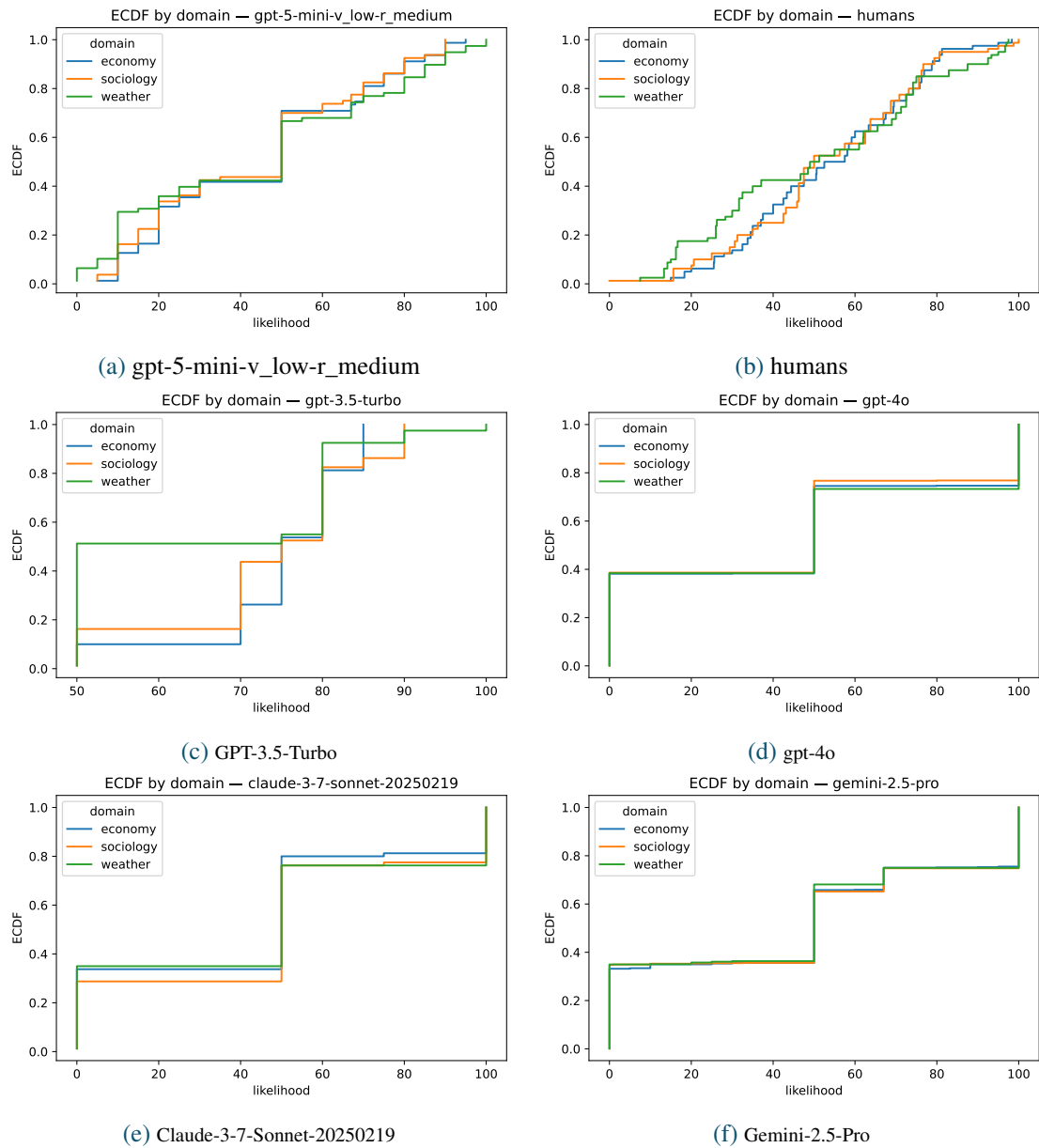


Figure B.2: Distribution of Likelihood Judgements by Domain, RW17 experiment.

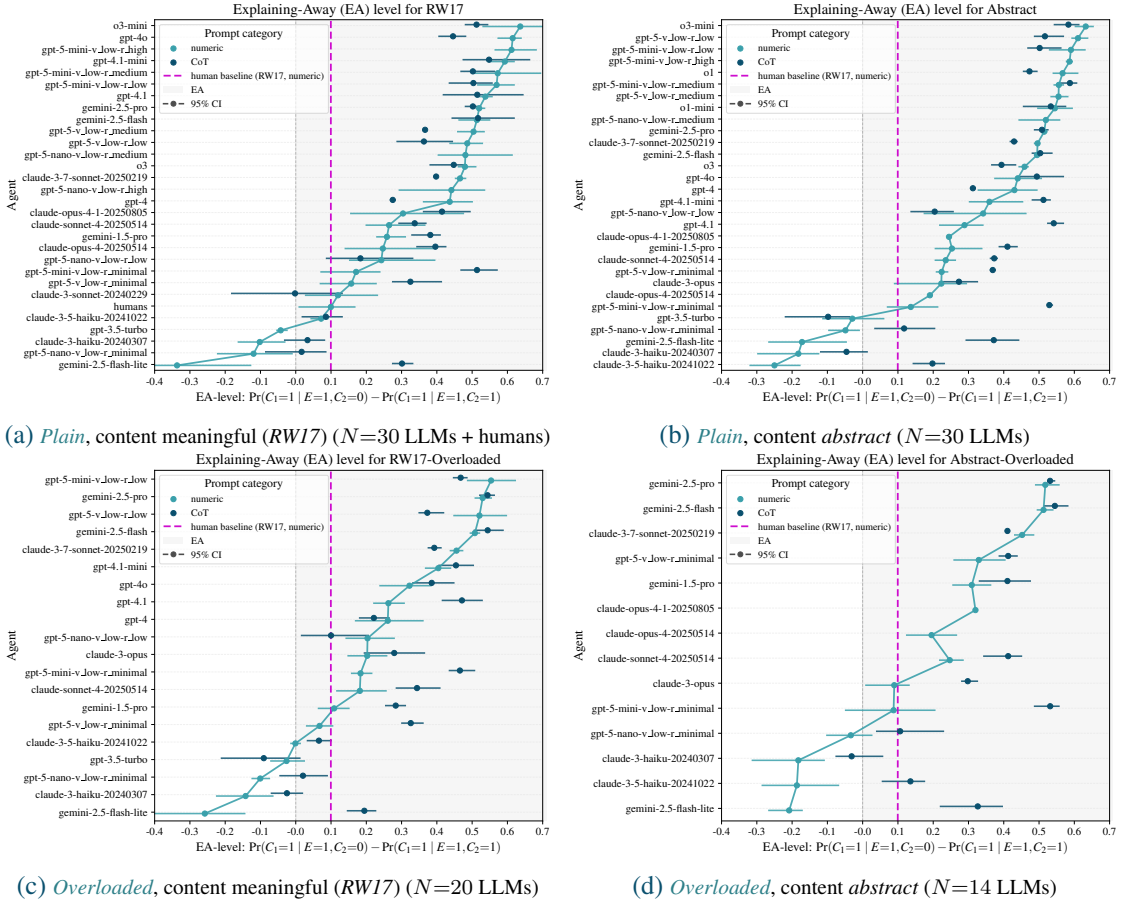
## B.4 Additional Results for the effect of Chain-of-Thought prompts on causal reasoning across experiments

**Table B.16:** Human–LLM alignment (Spearman rho) across domains for chain-of-thought prompting. Agents are ordered by pooled domain alignment. Each cell reports the bootstrapped Spearman  $\rho$  and 95% confidence interval [lower, upper]. Uncertainty reflects the range of  $\rho$  values obtained by nonparametric bootstrapping (2,000 resamples) for each agent–domain pair.

Agent	Domain			
	pooled	economy	sociology	weather
claude-3-7-sonnet-20250219	<b>0.857</b> [0.821, 0.885]	<b>0.843</b> [0.765, 0.892]	<b>0.867</b> [0.796, 0.917]	0.865 [0.799, 0.908]
gemini-2.5-flash-lite	0.845 [0.802, 0.877]	0.818 [0.717, 0.882]	0.852 [0.760, 0.912]	0.852 [0.787, 0.894]
claude-sonnet-4-20250514	0.835 [0.794, 0.868]	0.826 [0.757, 0.877]	0.818 [0.724, 0.889]	0.862 [0.782, 0.906]
claude-opus-4-20250514	0.835 [0.785, 0.873]	0.798 [0.696, 0.858]	0.845 [0.746, 0.909]	<b>0.869</b> [0.794, 0.916]
gpt-4.1	0.835 [0.787, 0.874]	0.824 [0.737, 0.881]	0.839 [0.745, 0.906]	0.834 [0.739, 0.891]
claude-opus-4-1-20250805	0.831 [0.782, 0.871]	0.816 [0.720, 0.875]	0.849 [0.755, 0.915]	0.833 [0.737, 0.892]
gemini-1.5-pro	0.831 [0.816, 0.843]	0.815 [0.790, 0.836]	0.852 [0.830, 0.872]	0.828 [0.802, 0.851]
gemini-2.5-pro	0.831 [0.793, 0.861]	0.809 [0.727, 0.865]	0.833 [0.754, 0.888]	0.859 [0.796, 0.898]
gemini-2.5-flash	0.830 [0.786, 0.865]	0.839 [0.755, 0.892]	0.840 [0.754, 0.904]	0.828 [0.740, 0.881]
gpt-4o	0.818 [0.763, 0.861]	0.804 [0.694, 0.875]	0.809 [0.689, 0.895]	0.836 [0.759, 0.884]
o3	0.816 [0.770, 0.853]	0.833 [0.744, 0.887]	0.822 [0.721, 0.892]	0.807 [0.722, 0.861]
gpt-5-mini-v_low-r_medium	0.810 [0.759, 0.851]	0.790 [0.682, 0.857]	0.804 [0.704, 0.876]	0.827 [0.730, 0.882]
gpt-4.1-mini	0.806 [0.747, 0.851]	0.782 [0.648, 0.869]	0.820 [0.709, 0.894]	0.813 [0.723, 0.869]
gpt-5-v_low-r_low	0.805 [0.758, 0.843]	0.784 [0.679, 0.851]	0.808 [0.707, 0.883]	0.819 [0.743, 0.868]
claude-3-opus	0.802 [0.740, 0.853]	0.802 [0.686, 0.869]	0.751 [0.596, 0.872]	0.842 [0.750, 0.905]
o3-mini	0.801 [0.741, 0.845]	0.752 [0.642, 0.825]	0.826 [0.731, 0.892]	0.827 [0.722, 0.887]
gpt-5-mini-v_low-r_low	0.797 [0.740, 0.842]	0.762 [0.643, 0.839]	0.811 [0.712, 0.884]	0.817 [0.724, 0.873]
gpt-5-v_low-r_minimal	0.793 [0.738, 0.840]	0.768 [0.648, 0.848]	0.811 [0.712, 0.884]	0.813 [0.712, 0.874]
gpt-4	0.790 [0.737, 0.835]	0.749 [0.641, 0.830]	0.804 [0.693, 0.881]	0.814 [0.726, 0.874]
gpt-5-mini-v_low-r_minimal	0.784 [0.719, 0.833]	0.727 [0.595, 0.820]	0.798 [0.680, 0.881]	0.812 [0.708, 0.877]
claude-3-5-haiku-20241022	0.771 [0.712, 0.818]	0.680 [0.526, 0.791]	0.803 [0.713, 0.864]	0.808 [0.708, 0.877]
claude-3-sonnet-20240229	0.767 [0.704, 0.819]	0.777 [0.674, 0.851]	0.753 [0.623, 0.843]	0.784 [0.678, 0.860]
gpt-5-nano-v_low-r_low	0.766 [0.700, 0.818]	0.761 [0.657, 0.832]	0.684 [0.518, 0.811]	0.831 [0.739, 0.885]
gpt-5-v_low-r_medium	0.756 [0.588, 0.860]	0.756 [0.588, 0.860]	–	–
gpt-5-nano-v_low-r_minimal	0.709 [0.625, 0.776]	0.746 [0.636, 0.823]	0.716 [0.594, 0.803]	0.684 [0.494, 0.833]
claude-3-haiku-20240307	0.688 [0.609, 0.757]	0.656 [0.495, 0.783]	0.636 [0.473, 0.762]	0.754 [0.634, 0.839]

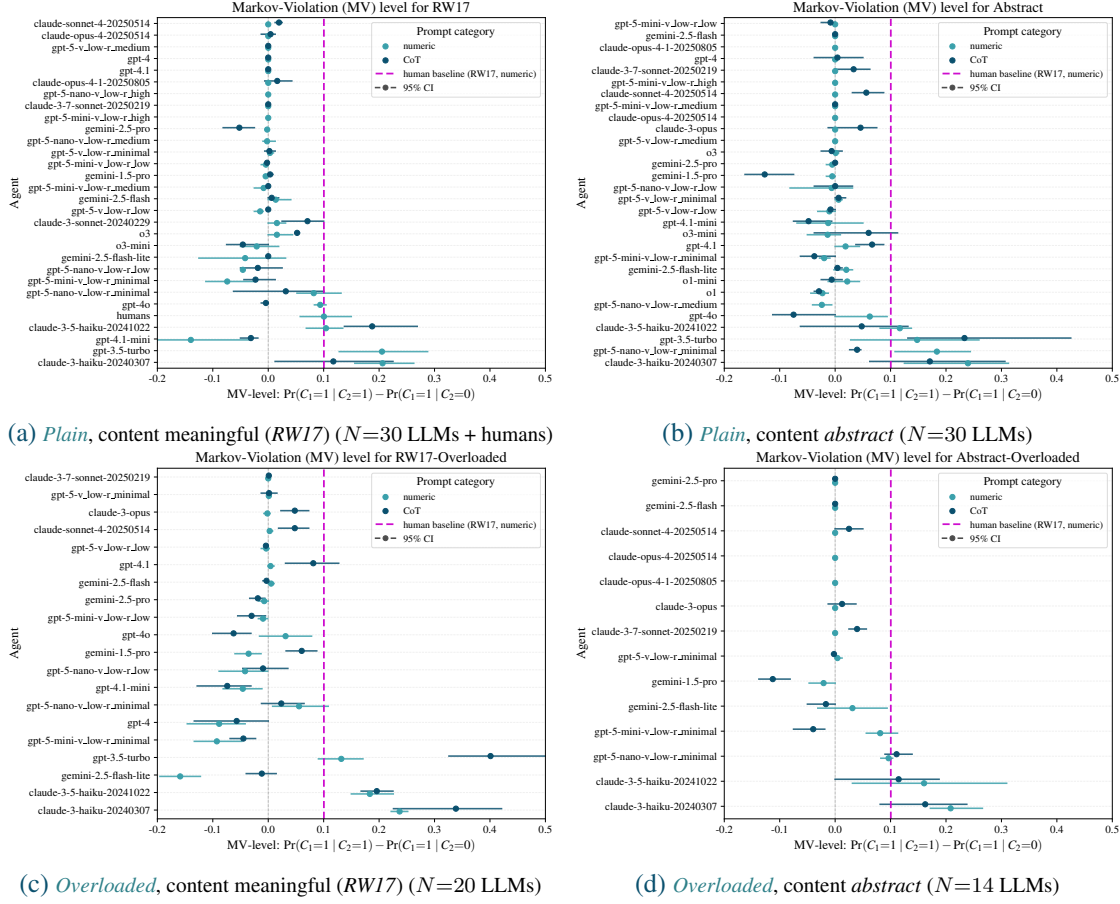
## B.4 Additional Results for the effect of Chain-of-Thought prompts on causal reasoning across experiments

## Appendix B Additional Results



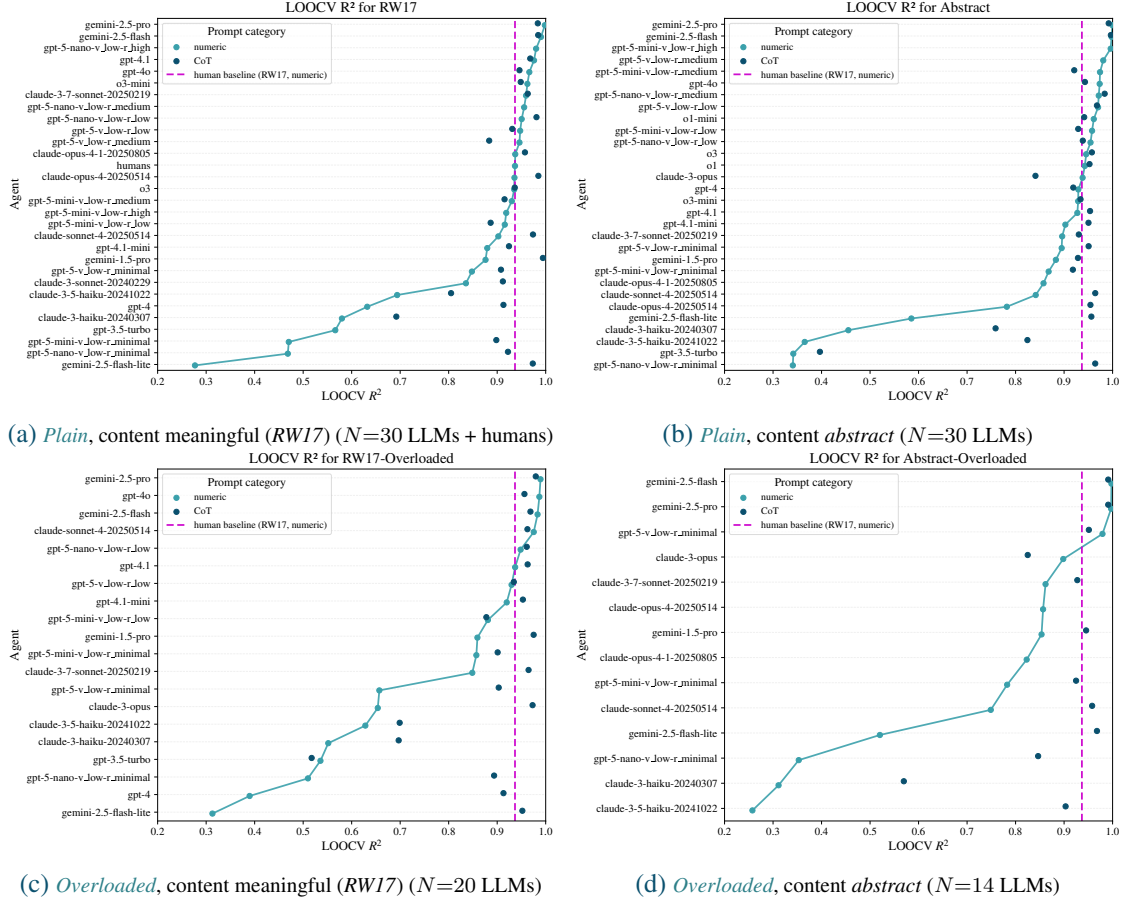
**Figure B.3: Explaining Away (EA) levels by agent on *semantically meaningful* content (RW17) (Figures B.3(a) and B.3(c)) and *abstract, semantically meaningless* content Figures B.3(b) and B.3(d) and the effects of CoT.** The left column represents the *plain* versions and the right column the *overloaded* versions (irrelevant text injected) of the respective experiments. EA-levels are computed on normalized raw agents' likelihood judgements (not their respective CBN predictions). **EA is prevalent:** overall across all conditions (see Figures B.3(a) to B.3(d)), 90.6% of LLMs exceed the human EA baseline (RW17, numeric)  $EA_{\text{human}} \approx 0.100$  and 81.3% exceed the stricter threshold that we set to  $EA > 0.3$ . *CoT generally increases EA-pass rates* relative to numeric in every setting (e.g., RW17 plain: 58.6%→69.2% vs. our threshold; 79.3%→84.6% vs. human; Abstract plain: 60.0%→76.0% and 80.0%→92.0%). *Overloaded prompts reduce EA-levels but rates remain high* (RW17 human-threshold: 73.3%→61.9%; our threshold: 53.3%→38.1%; Abstract human-threshold: 80.0%→64.3%; our threshold: 56.7%→42.9%). *Abstract mirrors meaningful content* (numeric threshold: 58.6% [RW17] vs. 60.0% [Abstract]; CoT: 69.2% vs. 76.0%; vs. human: 79.3%/84.6% vs. 80.0%/92.0%).

## B.4 Additional Results for the effect of Chain-of-Thought prompts on causal reasoning across experiments



**Figure B.4: Markov Violation (MV) levels across experiments by agent on *semantically meaningful* content (RW17) (Figures B.4(a) and B.4(c)) and *abstract, semantically meaningless* content (Figures B.4(b) and B.4(d)) and the effects of CoT.** The left column represents the *plain* versions and the right column the *overloaded* versions (irrelevant text injected) of the respective experiments. MV-levels are computed on normalized raw agents’ likelihood judgements (not their respective CBN predictions). When agents fall within the gray shaded area (representing  $|MV| \leq 0 + \epsilon$ , where  $\epsilon = 0.05$ ), we deem them as respecting independence of causes (i.e., Markov compliant). **Generally high Markov compliance (respecting independence of causes):** MV magnitudes are typically small. Overall, 93.8% of agents meet our MV-threshold  $|MV| \leq 0.05$  at least once across all conditions and 93.8% are at or below the human baseline. Typical  $|MV|$  across agents is 0.010 (95% CI [0.002, 0.031]). Human baseline (RW17, numeric)  $|MV_{\text{human}}| \approx 0.100$ .

## Appendix B Additional Results



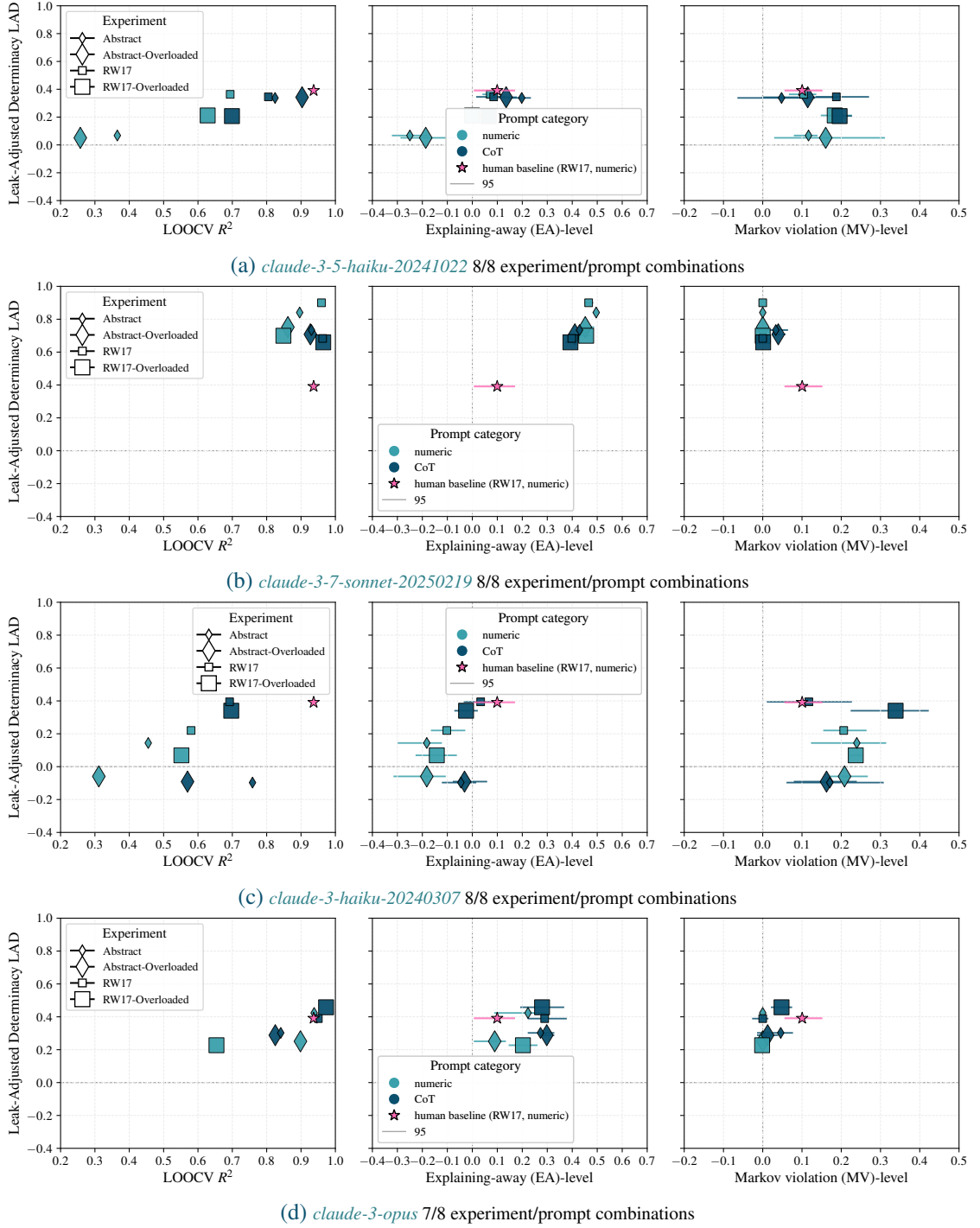
**Figure B.5: Reasoning Consistency (LOOCV-R2) across experiments by agent** on *semantically meaningful* content (RW17) (Figures B.5(a) and B.5(c)) and *abstract, semantically meaningless* content (Figures B.5(b) and B.5(d)) and the effects of CoT. The left column represents the *plain* versions and the right column the *overloaded* versions (irrelevant text injected) of the respective experiments. MV-levels are computed on normalized raw agents' likelihood judgements (not their respective CBN predictions). When agents fall within the gray shadowed aream, we deem them as respecting independence of causes (i.e., Markov compliant). Human baseline (RW17, numeric)  $\text{LOOCV}R^2 \approx 0.937$ .



## **B.5 Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explainaing Away, and Markov Violation**

## Appendix B Additional Results

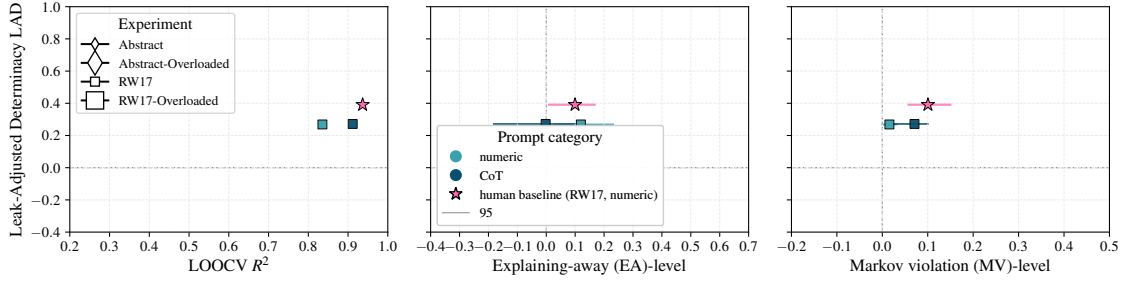
Leak-Adjusted Determinacy (LAD =  $\bar{m} - b$ ) vs  $R^2$ /EA/MV



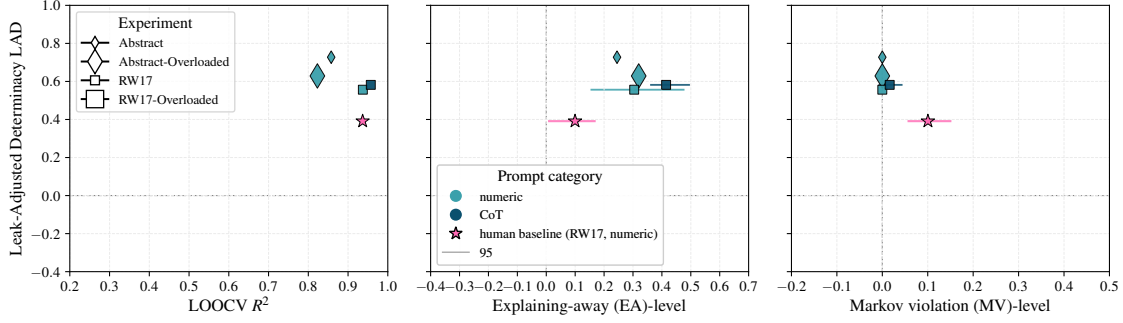
**Figure B.6: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.

## B.5 Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explaining Away, and Markov Violation

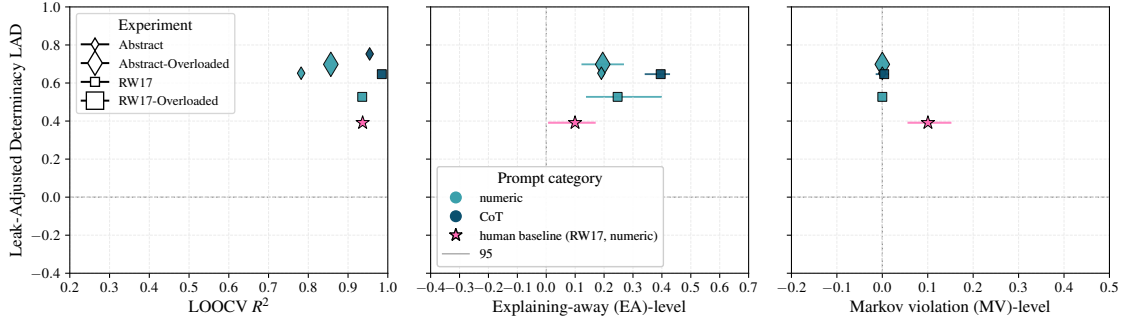
Leak-Adjusted Determinacy (LAD =  $\bar{m} - b$ ) vs  $R^2$ /EA/MV



(a) *claude-3-sonnet-20240229* 2/8 experiment/prompt combinations

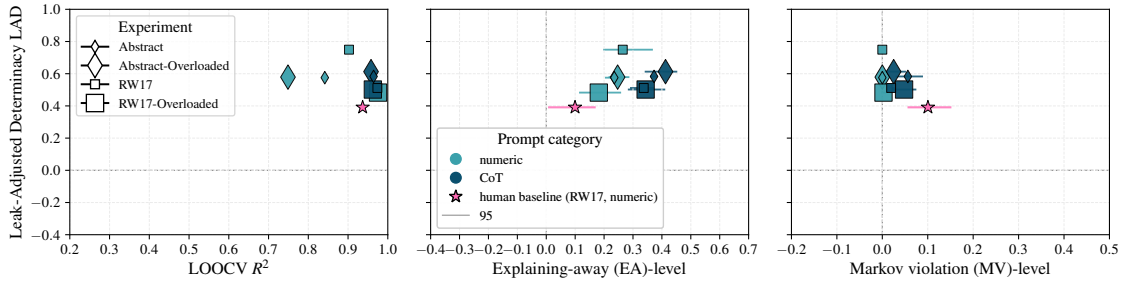


(b) *claude-opus-4-1-20250805* 5/8 experiment/prompt combinations



(c) *claude-opus-4-20250514* 6/8 experiment/prompt combinations

Leak-Adjusted Determinacy (LAD =  $\bar{m} - b$ ) vs  $R^2$ /EA/MV

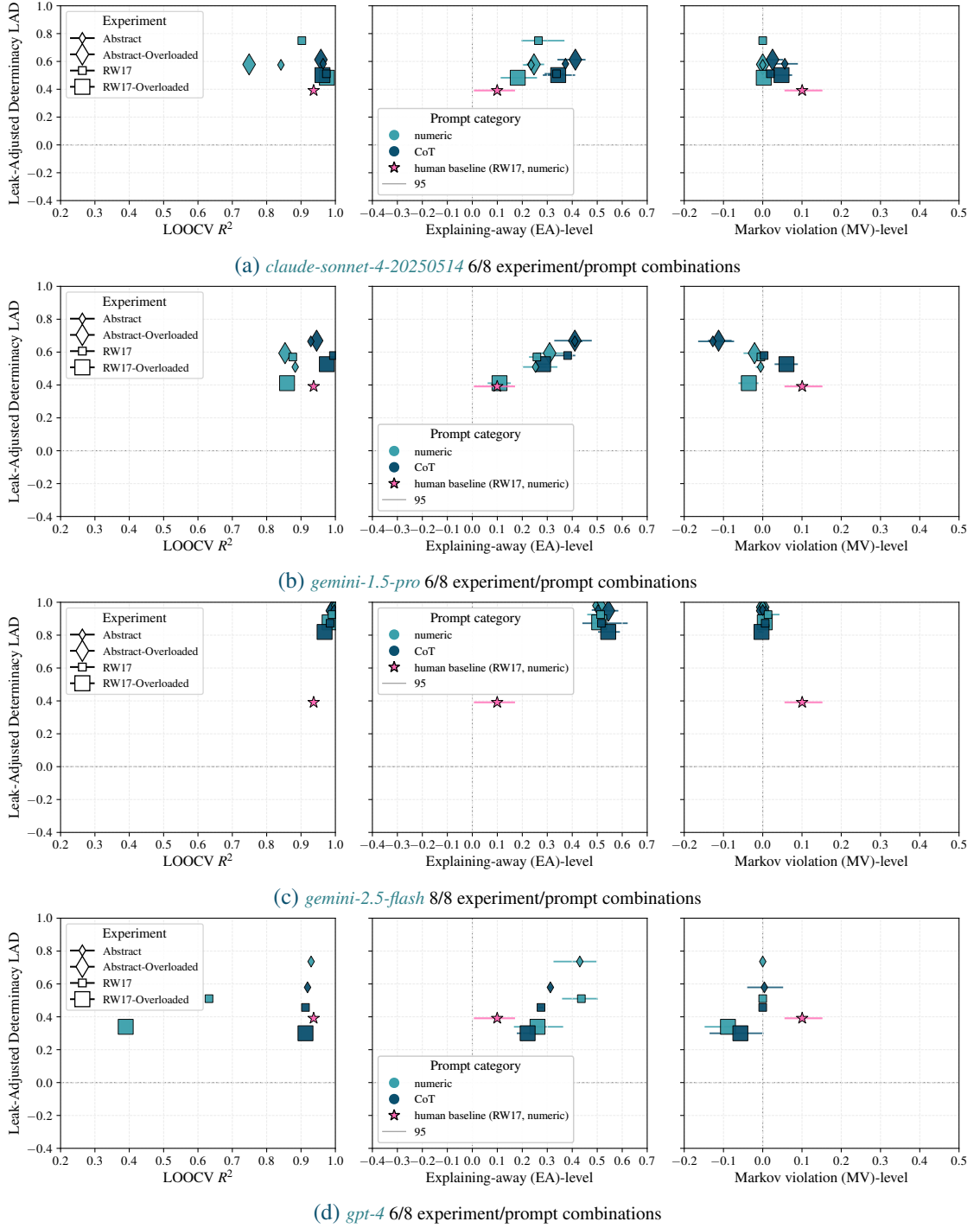


(d) *claude-sonnet-4-20250514* 8/8 experiment/prompt combinations

**Figure B.7: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.

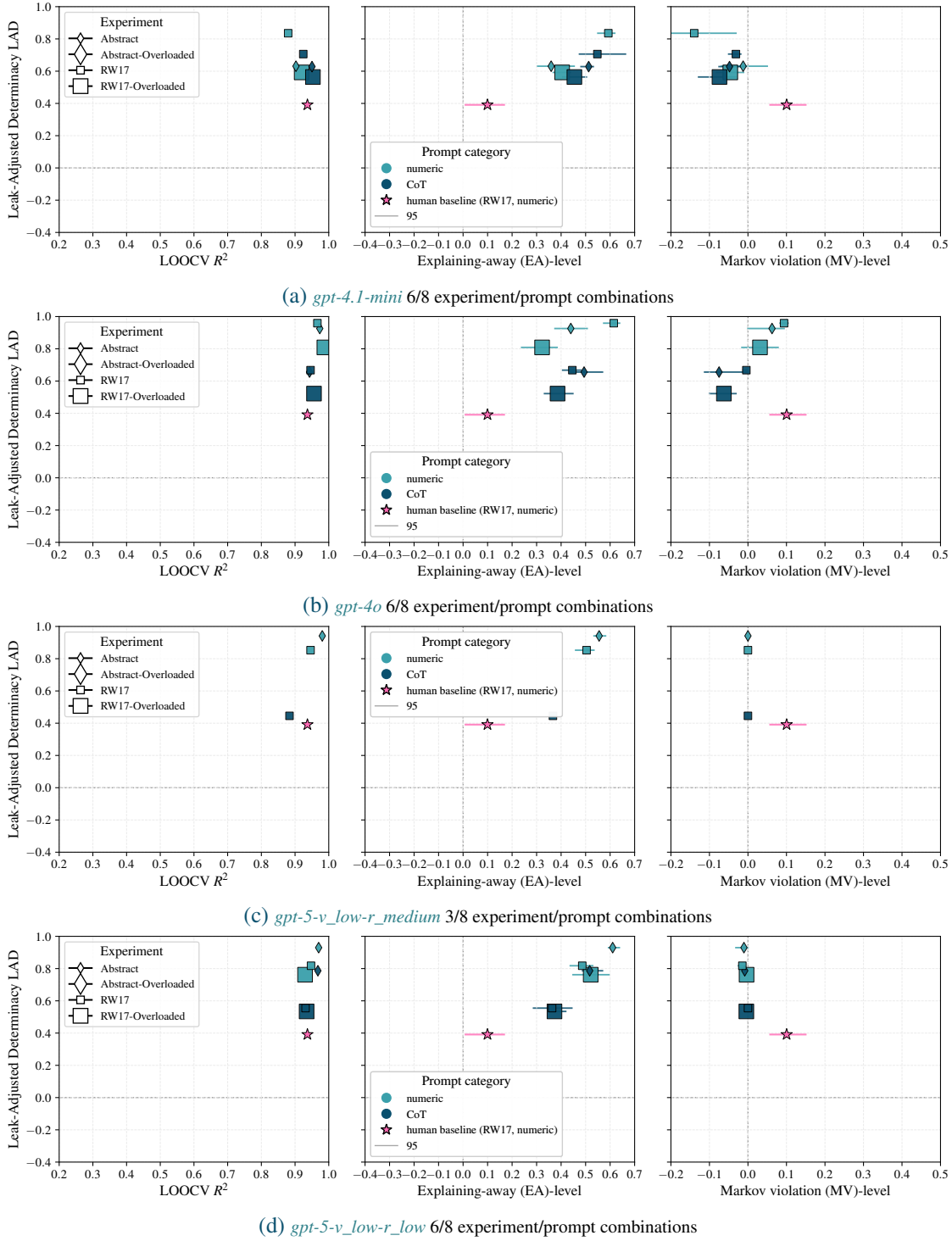
## Appendix B Additional Results

Leak-Adjusted Determinacy (LAD =  $\bar{m} - b$ ) vs  $R^2$ /EA/MV



**Figure B.8: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.

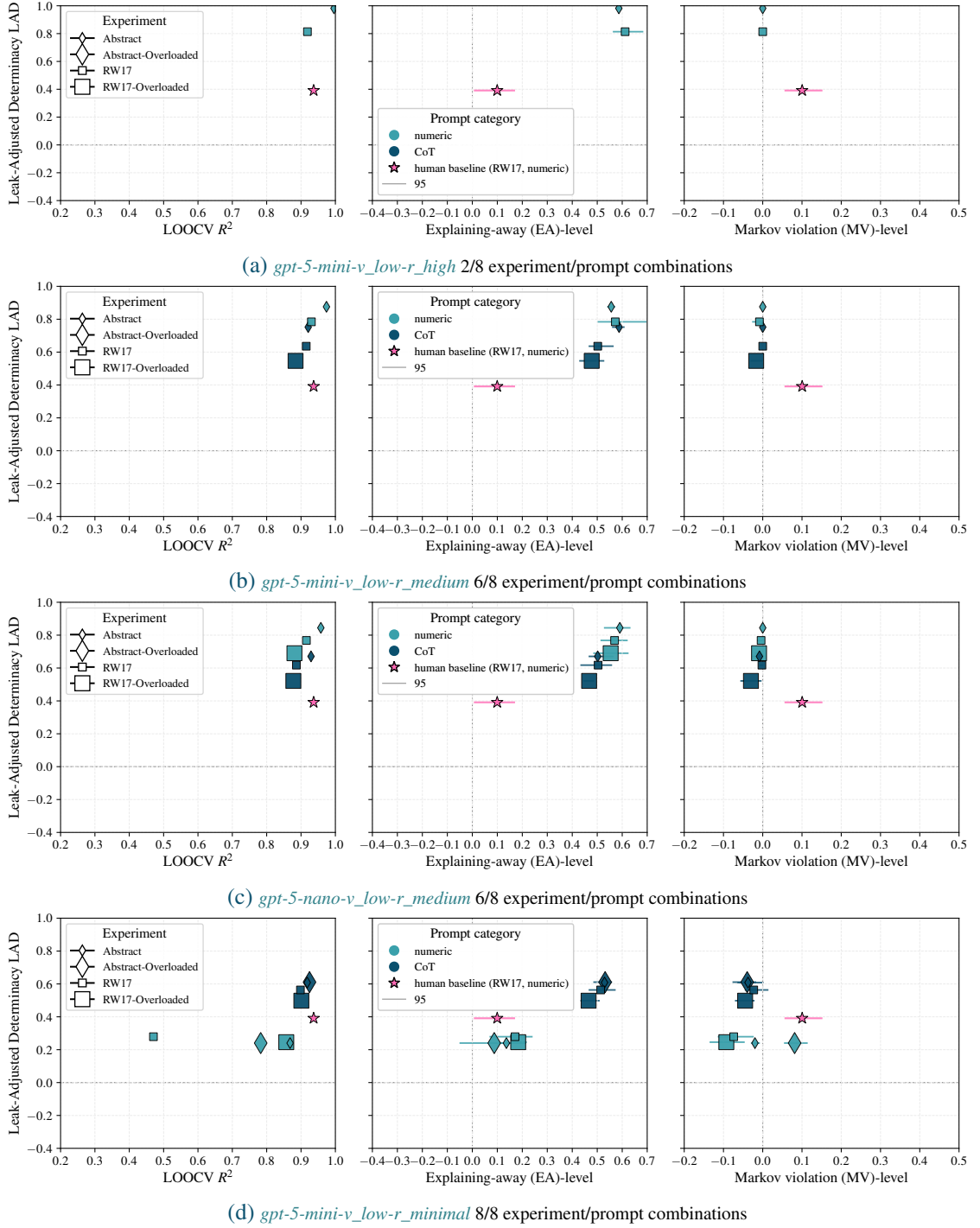
## B.5 Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explaining Away, and Markov Violation



**Figure B.9: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.

## Appendix B Additional Results

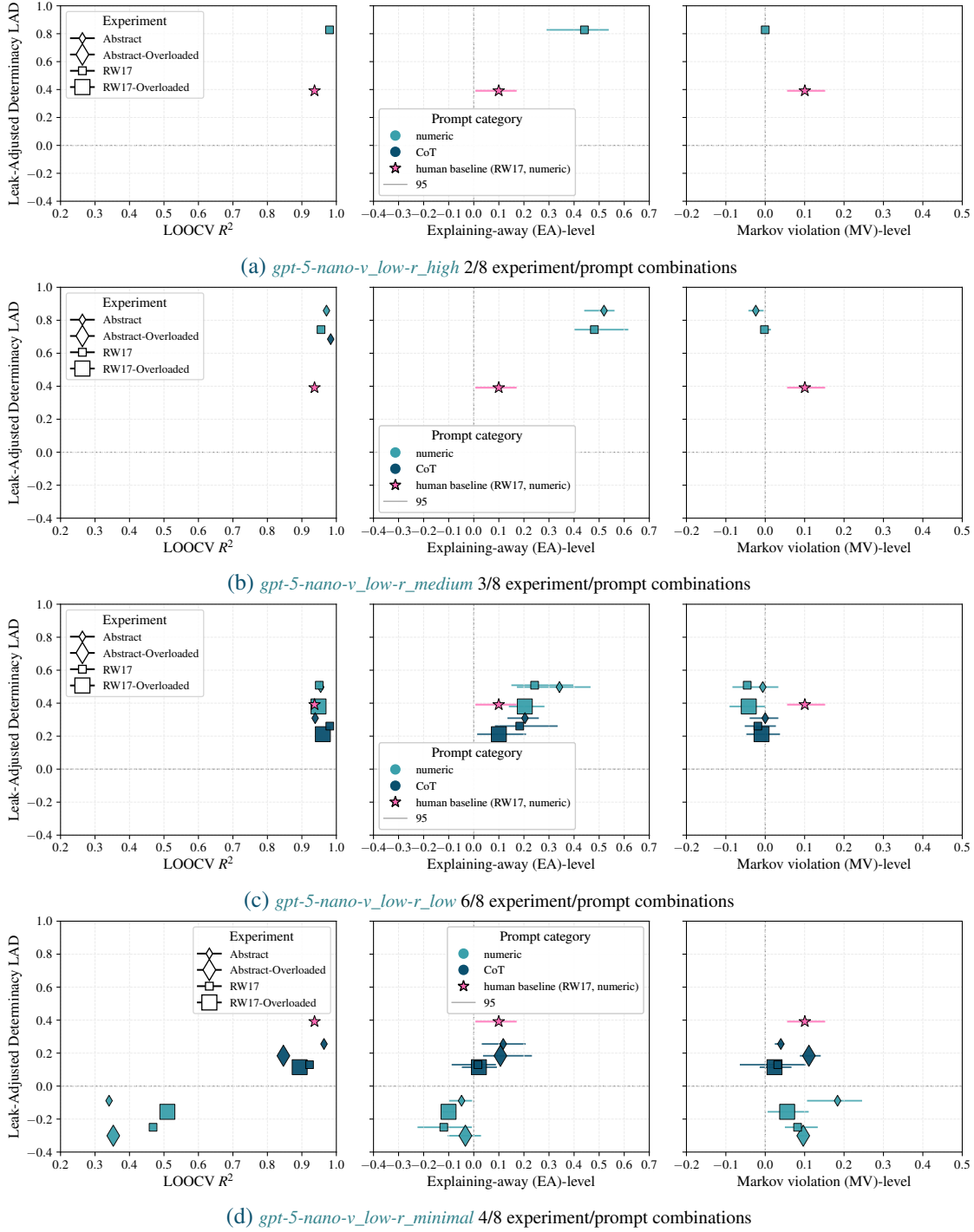
Leak-Adjusted Determinacy ( $LAD = \bar{m} - b$ ) vs  $R^2$ /EA/MV



**Figure B.10: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.

## B.5 Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explaining Away, and Markov Violation

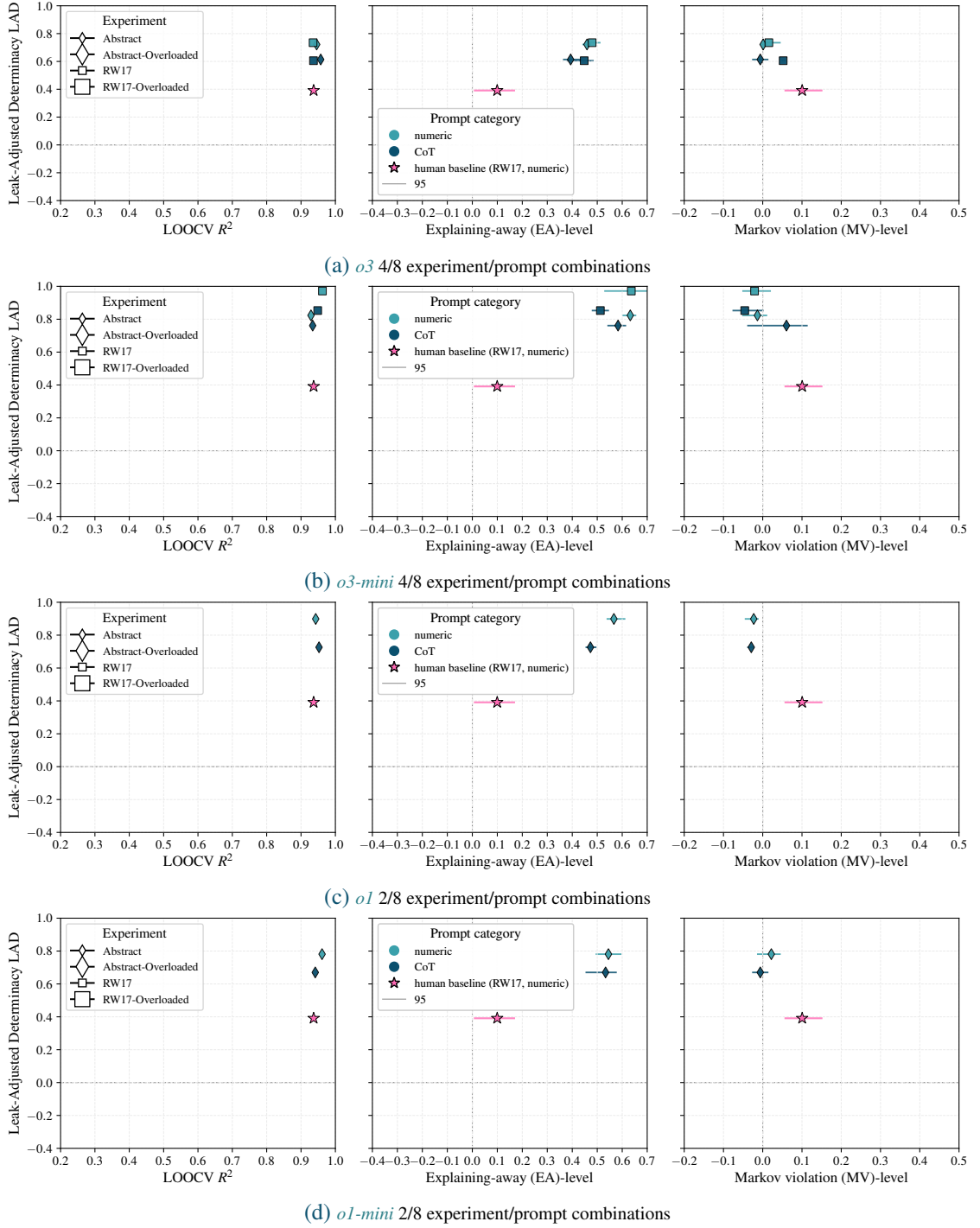
Leak-Adjusted Determinacy (LAD =  $\bar{m} - b$ ) vs  $R^2$ /EA/MV



**Figure B.11: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.

## Appendix B Additional Results

Leak-Adjusted Determinacy (LAD =  $\bar{m} - b$ ) vs  $R^2$ /EA/MV



**Figure B.12: Leak-Adjusted Determinacy (LAD) levels  $LAD = \bar{m} - b$  vs.  $R^2$  EA, and MV levels vary widely across agents and experiment / prompt category manipulations.** The figure illustrates: (1) how consistent the agent is under different prompts and content manipulations — closer scatter clustering signals greater robustness to content manipulations and prompt variations; (2) how deterministic an agent is — more scatters in the top ( $LAD \rightarrow 1$ ) indicate higher determinism, namely high causal strength  $m$  and low  $b$  consistent with Cheng [41] causal power theory. Conversely, lower  $LAD \rightarrow -1$  values indicate more probabilistic reasoning.



## B.5 Causal reasoning in Collider Graphs: Reasoning Determinacy versus Reasoning Robustness, Explaining Away, and Markov Violation

### B.5.1 Metrics by Release Date of LLMs

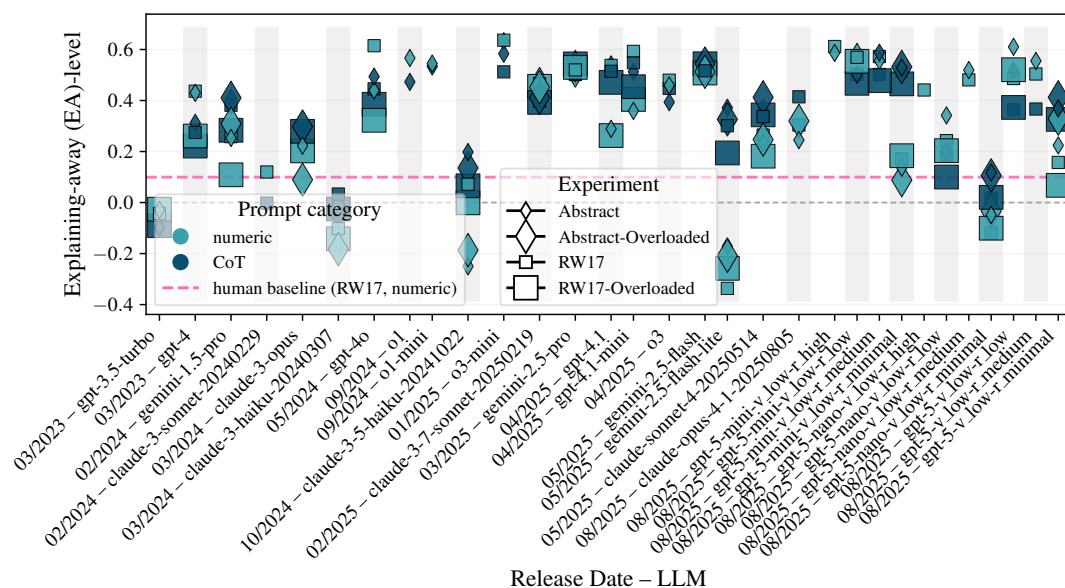


Figure B.13: Explaining away (EA) per experiment and prompt category by LLM release date.

## Appendix B Additional Results

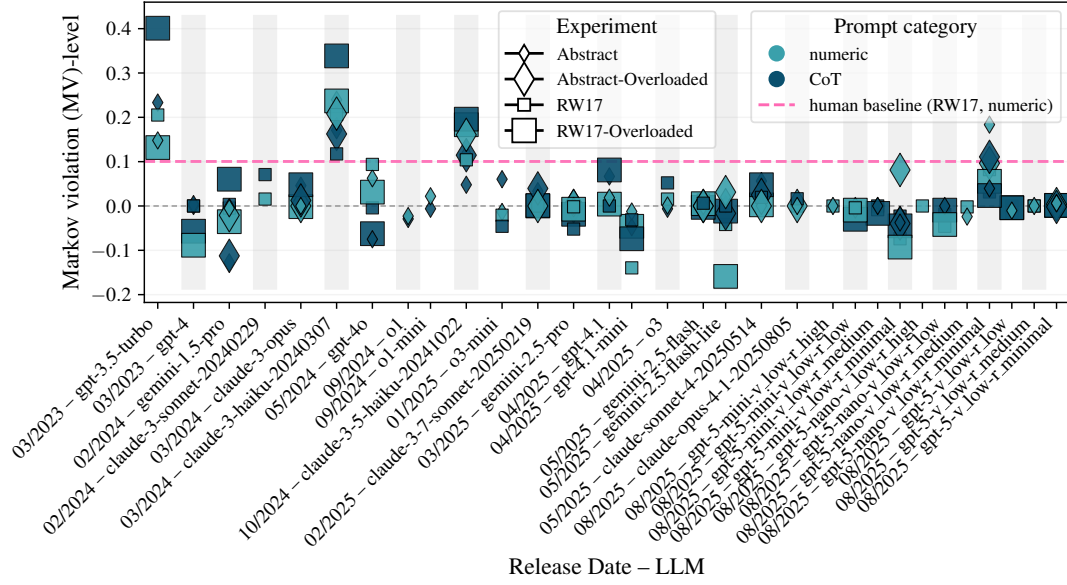


Figure B.14: Markov Violation (MV) per experiment and prompt category by LLM release date.

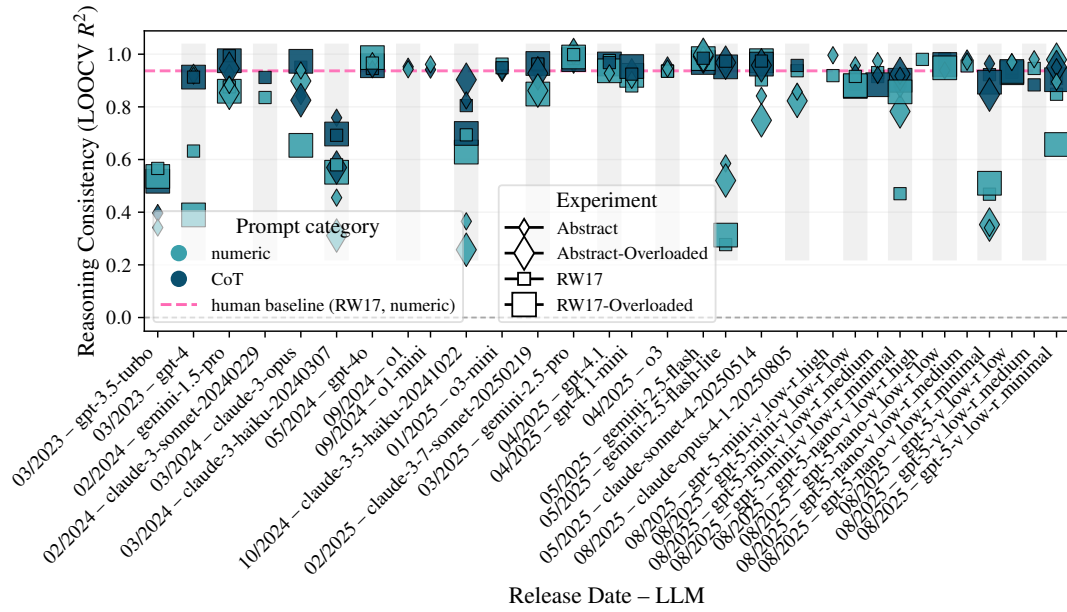


Figure B.15:  $R^2$  per experiment and prompt category by LLM release date.

## **B.6 Most and Least changing LLMs across prompt-category & content manipulations**

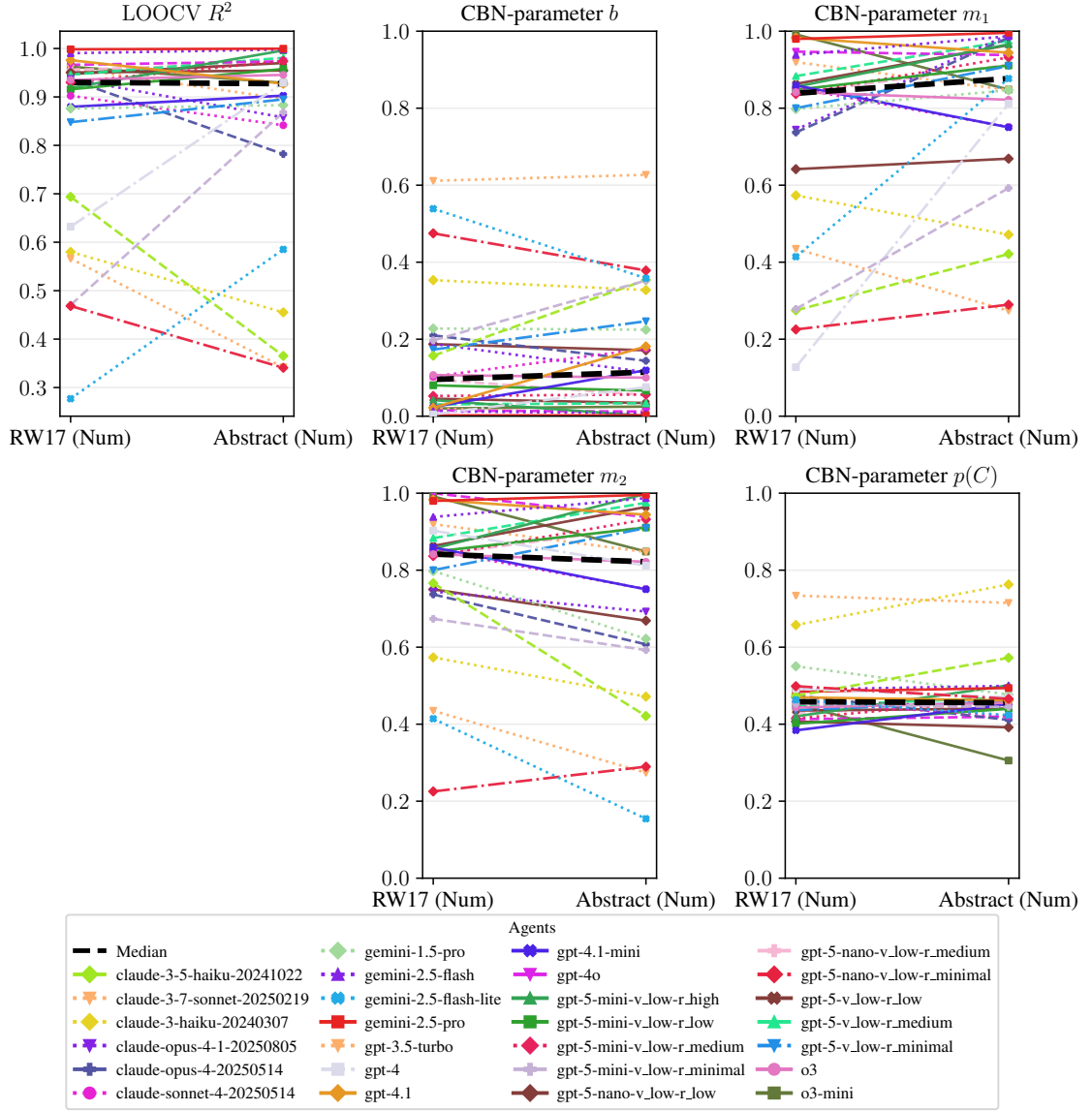
The subsequent plots and tables summarize the most and least changing agents between conditions (experiments or prompt-styles) per CBN-parameter and LOOCV  $R^2$ . These results complement Section 4.5, Section 4.5 and Section 4.6 in Chapter 4.

### **B.6.1 Experiment-wise changes with fixed prompt-style (Numeric or CoT)**

See also Table B.17 for the top 3 most and least changing agents per experiment pair and a fixed prompt-style for each CBN-parameter.

## Appendix B Additional Results

RW17 (Num)  $\rightarrow$  Abstract (Num); N=27 matched agents



**Figure B.16: Pairwise experiment-wise comparisons, Numeric.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to visually identify outliers and the most changing agents between conditions (i.e., agents with a slope).

## B.6 Most and Least changing LLMs across prompt-category & content manipulations

RW17 (CoT)  $\rightarrow$  Abstract (CoT); N=23 matched agents

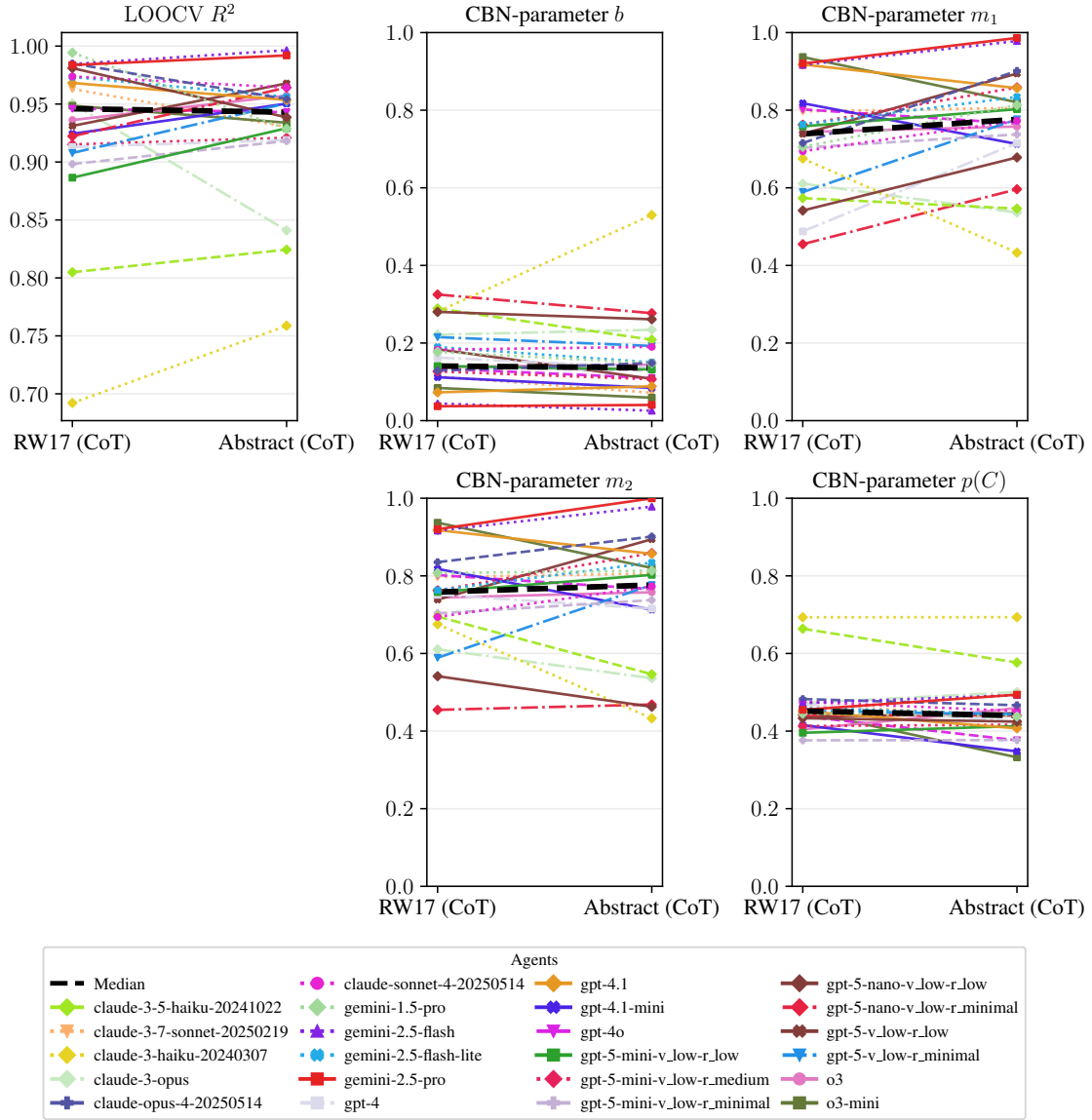
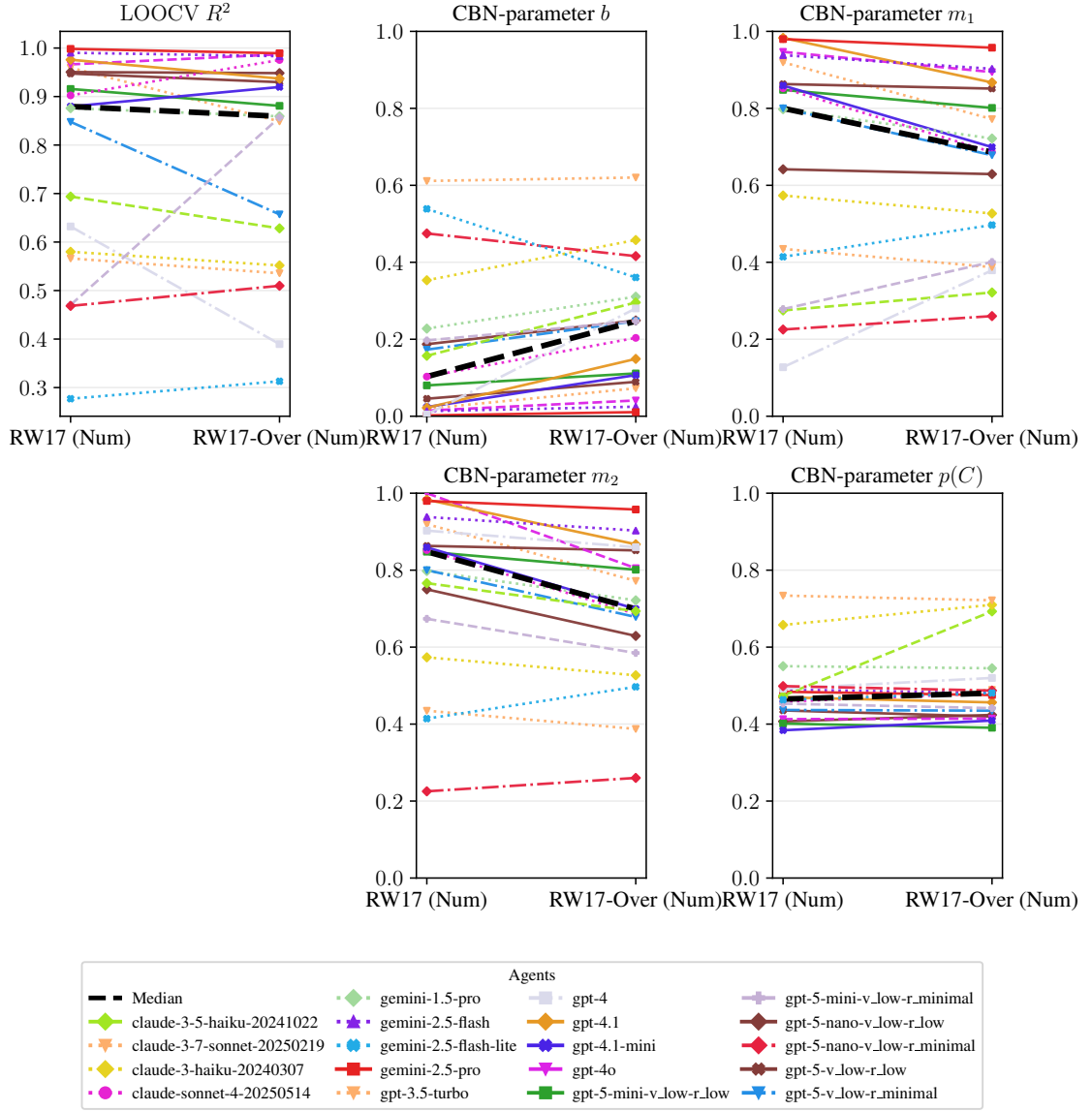


Figure B.17: **Pairwise experiment-wise comparisons, CoT.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## Appendix B Additional Results

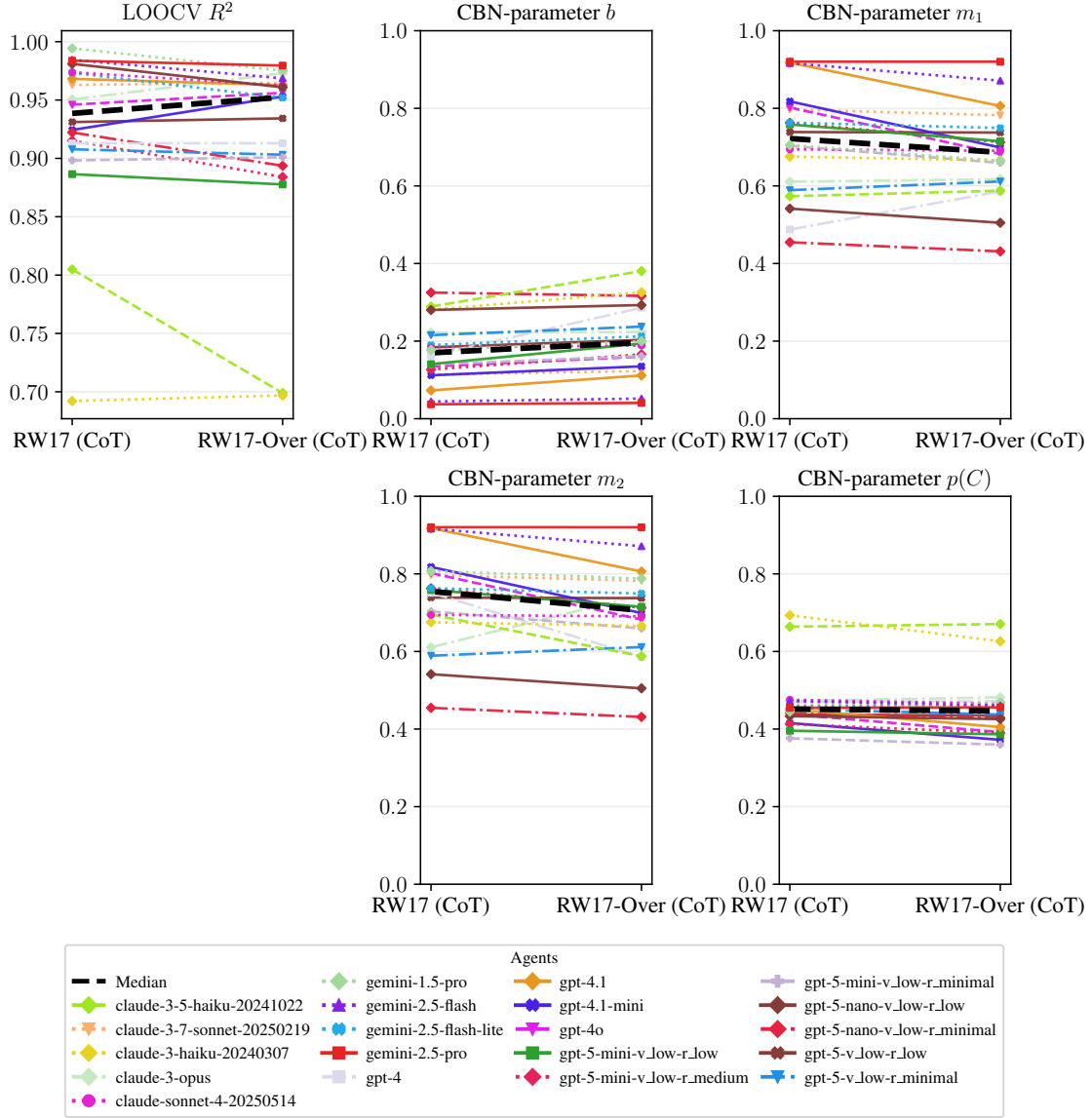
RW17 (Num)  $\rightarrow$  RW17-Over (Num); N=19 matched agents



**Figure B.18: Pairwise experiment-wise comparisons, Numeric.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## B.6 Most and Least changing LLMs across prompt-category & content manipulations

RW17 (CoT)  $\rightarrow$  RW17-Over (CoT); N=20 matched agents



**Figure B.19: Pairwise experiment-wise comparisons, CoT.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## Appendix B Additional Results

**Table B.17:** Top-3 most/least-changing agents per experiment-pair condition (by signed  $\Delta$ ;  $\Delta = A - B$ ) for given prompt category (PC) and parameter (Param).

Condition A–B	PC	Param	Most	Least
RW17 – Abstract-Over	CoT	$b$	claude-opus-4-20250514 (+0.868); claude-opus-4-1-20250805 (+0.838); claude-3-haiku-20240307 (+0.291)	gemin-2.5-pro (-0.001); gpt-5-mini-v_low-r_minimal (-0.012); gemini-2.5-flash (-0.013)
RW17 – Abstract-Over	CoT	$m_1$	claude-opus-4-1-20250805 (-0.624); claude-opus-4-20250514 (-0.596); gpt-5-v_low-r_minimal (+0.231)	claude-3-7-sonnet-20250219 (-0.003); gpt-5-nano-v_low-r_minimal (+0.009); gpt-5-mini-v_low-r_minimal (+0.036)
RW17 – Abstract-Over	CoT	$m_2$	claude-opus-4-20250514 (-0.716); claude-opus-4-1-20250805 (-0.624); gpt-5-v_low-r_minimal (+0.231)	claude-3-7-sonnet-20250219 (-0.003); gpt-5-nano-v_low-r_minimal (+0.009); gemini-1.5-pro (+0.012)
RW17 – Abstract-Over	CoT	$p(C)$	claude-opus-4-1-20250805 (+0.057); claude-3-5-haiku-20241022 (-0.039); claude-3-7-sonnet-20250219 (-0.032)	gemin-2.5-flash (+0.007); claude-3-haiku-20240307 (+0.009); gemini-2.5-flash-lite (-0.009)
RW17 – Abstract-Over	Num	$b$	gemin-2.5-flash-lite (-0.240); claude-3-5-haiku-20241022 (+0.216); gpt-5-mini-v_low-r_minimal (+0.153)	gemin-2.5-flash (+0.004); gpt-5-v_low-r_minimal (-0.008); gemini-2.5-pro (+0.012)
RW17 – Abstract-Over	Num	$m_1$	gpt-5-mini-v_low-r_minimal (+0.312); claude-opus-4-20250514 (+0.250); claude-opus-4-1-20250805 (+0.241)	gemin-1.5-pro (-0.010); gemini-2.5-pro (+0.020); gpt-5-nano-v_low-r_minimal (+0.042)
RW17 – Abstract-Over	Num	$m_2$	claude-3-5-haiku-20241022 (-0.342); claude-3-7-sonnet-20250219 (-0.231); gpt-5-nano-v_low-r_minimal (-0.225)	claude-opus-4-20250514 (-0.003); gemini-1.5-pro (-0.010); gemini-2.5-pro (+0.020)
RW17 – Abstract-Over	Num	$p(C)$	claude-3-5-haiku-20241022 (+0.203); gemini-1.5-pro (-0.085); claude-sonnet-4-20250514 (-0.074)	gemin-2.5-pro (+0.003); claude-3-7-sonnet-20250219 (-0.005); gemini-2.5-flash (+0.006)
RW17 – Abstract	CoT	$b$	claude-3-haiku-20240307 (+0.249); claude-3-5-haiku-20241022 (-0.080); gpt-5-v_low-r_low (-0.076)	gemin-2.5-pro (+0.003); o3 (+0.005); claude-sonnet-4-20250514 (+0.007)
RW17 – Abstract	CoT	$m_1$	claude-3-haiku-20240307 (-0.242); gpt-4 (+0.228); gpt-5-v_low-r_minimal (+0.187)	claude-3-7-sonnet-20250219 (+0.009); o3 (+0.014); claude-3-5-haiku-20241022 (-0.027)
RW17 – Abstract	CoT	$m_2$	claude-3-haiku-20240307 (-0.242); gpt-5-v_low-r_minimal (+0.187); gpt-5-v_low-r_low (+0.156)	gemin-1.5-pro (+0.006); claude-3-7-sonnet-20250219 (+0.009); gpt-5-nano-v_low-r_minimal (+0.014)
RW17 – Abstract	CoT	$p(C)$	o3-mini (-0.120); claude-3-5-haiku-20241022 (-0.087); gpt-4.1-mini (-0.068)	claude-3-haiku-20240307 (+0.000); gpt-5-mini-v_low-r_minimal (+0.001); gpt-5-nano-v_low-r_minimal (-0.002)
RW17 – Abstract	Num	$b$	claude-3-5-haiku-20241022 (+0.197); gemini-2.5-flash-lite (-0.181); gpt-4.1 (+0.159)	gemin-2.5-pro (+0.000); gpt-4o (-0.002); gemini-1.5-pro (-0.003)
RW17 – Abstract	Num	$m_1$	gpt-4 (+0.685); gemini-2.5-flash-lite (+0.463); gpt-5-mini-v_low-r_minimal (+0.314)	gpt-4o (-0.009); gemini-2.5-pro (+0.016); o3 (-0.020)
RW17 – Abstract	Num	$m_2$	claude-3-5-haiku-20241022 (-0.345); gemini-2.5-flash-lite (-0.260); gemini-1.5-pro (-0.176)	gemin-2.5-pro (+0.016); o3 (-0.020); gpt-4.1 (-0.039)
RW17 – Abstract	Num	$p(C)$	o3-mini (-0.147); claude-3-haiku-20240307 (+0.105); claude-3-5-haiku-20241022 (+0.099)	gpt-5-v_low-r_low (+0.005); gpt-4o (+0.007); gpt-4.1 (-0.008)
RW17 – RW17-Over	CoT	$b$	gpt-4 (+0.123); claude-3-5-haiku-20241022 (+0.092); gpt-5-mini-v_low-r_low (+0.053)	claude-3-opus (+0.002); gemini-2.5-pro (+0.004); claude-sonnet-4-20250514 (+0.006)
RW17 – RW17-Over	CoT	$m_1$	gpt-4o (-0.120); gpt-4.1-mini (-0.119); gpt-4.1 (-0.112)	gemin-2.5-pro (-0.000); gpt-5-v_low-r_low (-0.001); claude-sonnet-4-20250514 (-0.004)
RW17 – RW17-Over	CoT	$m_2$	gpt-4 (-0.166); claude-3-opus (+0.135); gpt-4o (-0.120)	gemin-2.5-pro (-0.000); gpt-5-v_low-r_low (-0.001); claude-sonnet-4-20250514 (-0.004)
RW17 – RW17-Over	CoT	$p(C)$	claude-3-haiku-20240307 (-0.067); gpt-4o (-0.047); gpt-4.1 (-0.046)	claude-3-7-sonnet-20250219 (-0.001); gemini-2.5-pro (+0.001); gpt-5-v_low-r_low (-0.002)
RW17 – RW17-Over	Num	$b$	gpt-4 (+0.275); gemini-2.5-flash-lite (-0.178); claude-3-5-haiku-20241022 (+0.139)	gemin-2.5-pro (+0.009); gpt-3.5-turbo (+0.009); gemini-2.5-flash (+0.012)
RW17 – RW17-Over	Num	$m_1$	gpt-4 (+0.252); claude-sonnet-4-20250514 (-0.166); gpt-4.1-mini (-0.160)	gpt-5-v_low-r_low (-0.012); gpt-5-nano-v_low-r_low (-0.013); gemini-2.5-pro (-0.022)
RW17 – RW17-Over	Num	$m_2$	gpt-4o (-0.195); claude-sonnet-4-20250514 (-0.166); gpt-4.1-mini (-0.160)	gpt-5-v_low-r_low (-0.012); gemini-2.5-pro (-0.022); gpt-5-nano-v_low-r_minimal (+0.035)
RW17 – RW17-Over	Num	$p(C)$	claude-3-5-haiku-20241022 (+0.219); claude-3-haiku-20240307 (+0.052); gpt-4 (+0.028)	gpt-4o (+0.001); gpt-5-v_low-r_minimal (-0.002); gemini-1.5-pro (-0.005)



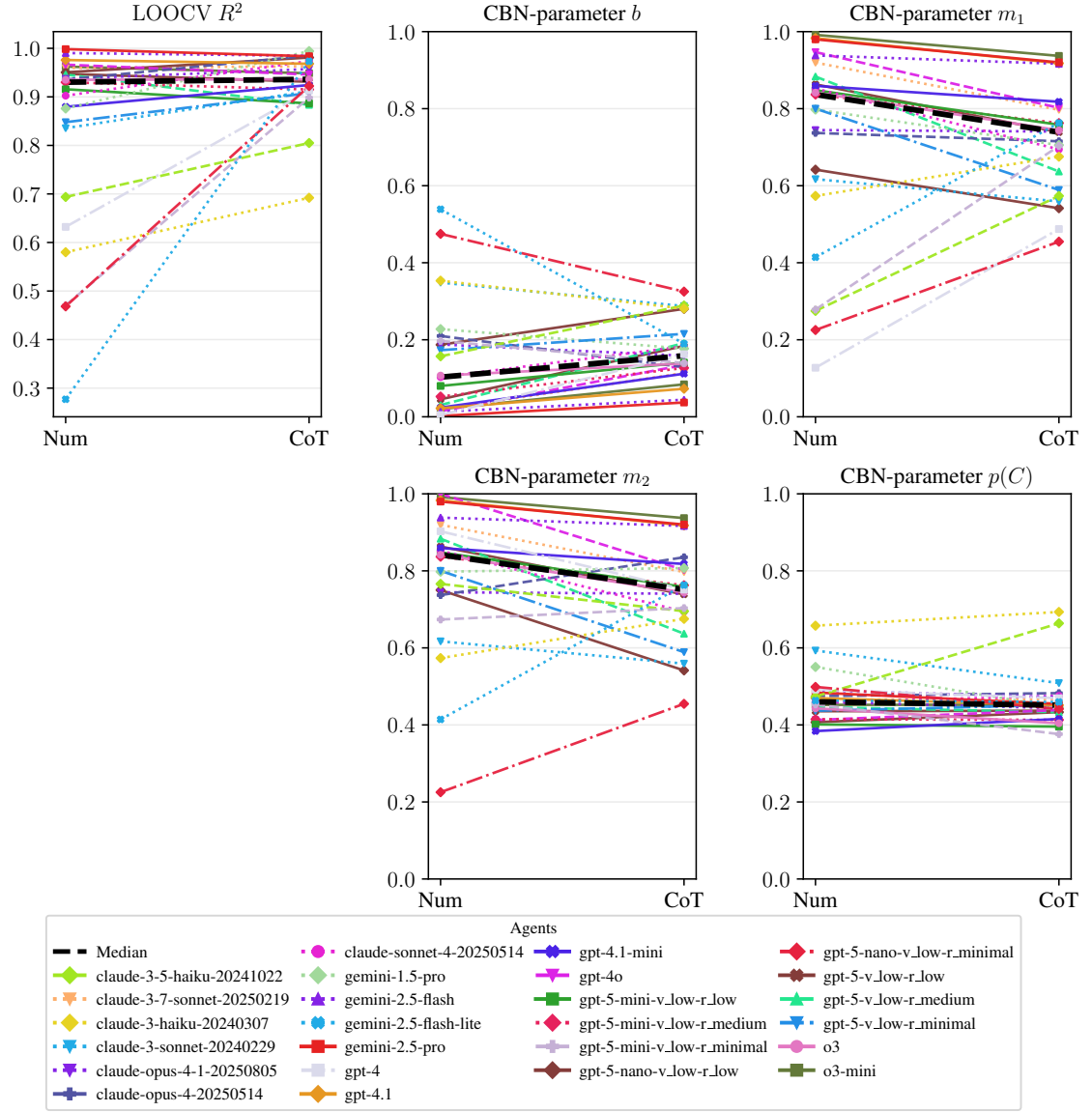
## B.6 Most and Least changing LLMs across prompt-category & content manipulations

### B.6.2 Prompt-wise changes with fixed experiment (e.g., RW17 or Abstract)

The subsequent figures show prompt-wise changes within a fixed experiment (RW17 or Abstract). These figures allow to visually identify outliers and the most changing agents between prompt styles (Numeric vs CoT) within a given experiment. See also [Table B.18](#) for a summary of the top-3 most/least changing agents per prompt-category (PC) comparisons given a experiment and CBN parameter (Param).

## Appendix B Additional Results

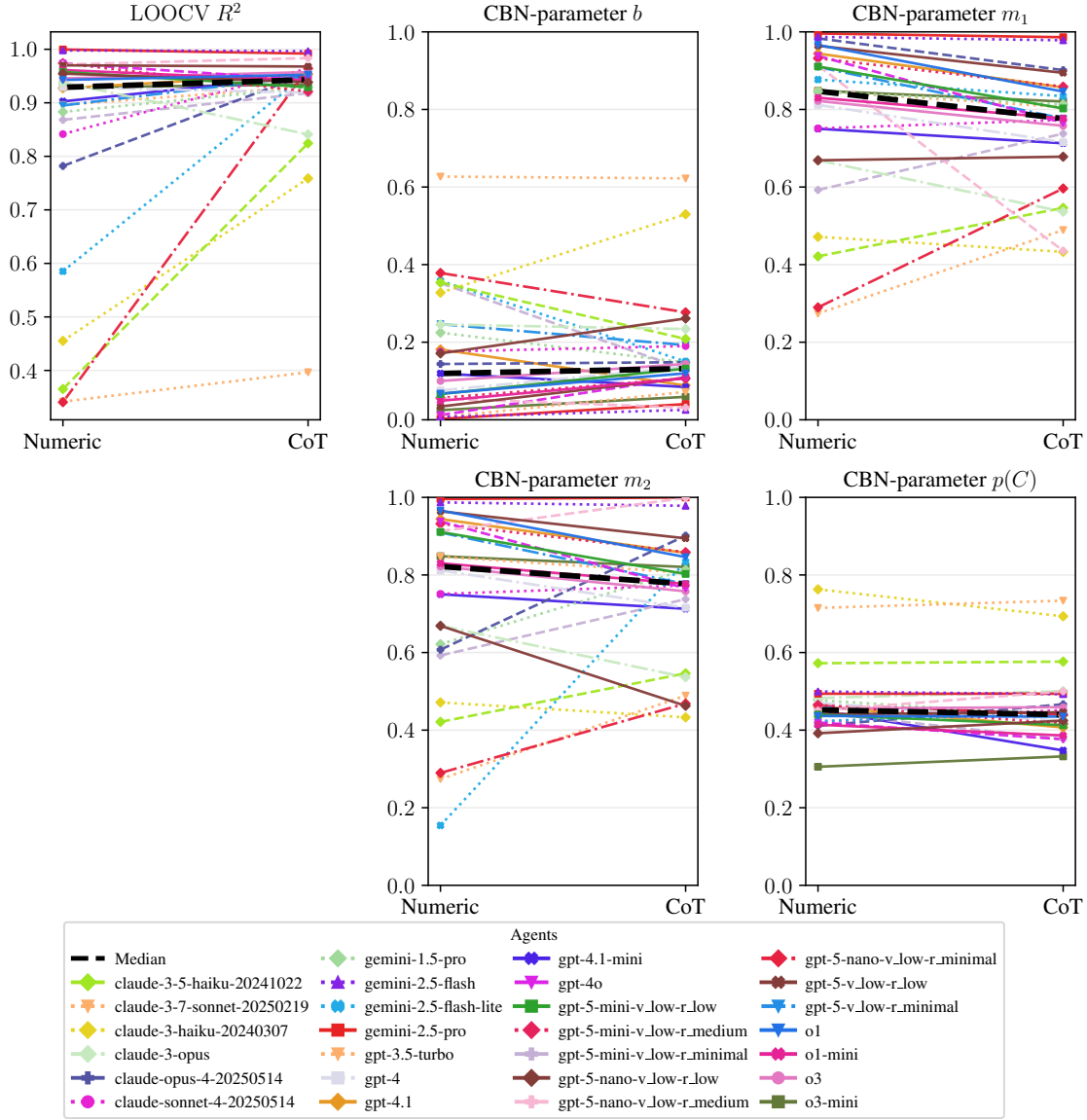
Experiment RW17: Numeric  $\rightarrow$  CoT; N=25 matched agents



**Figure B.20: Pairwise prompt-category comparisons with fixed experiment, here RW17.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## B.6 Most and Least changing LLMs across prompt-category & content manipulations

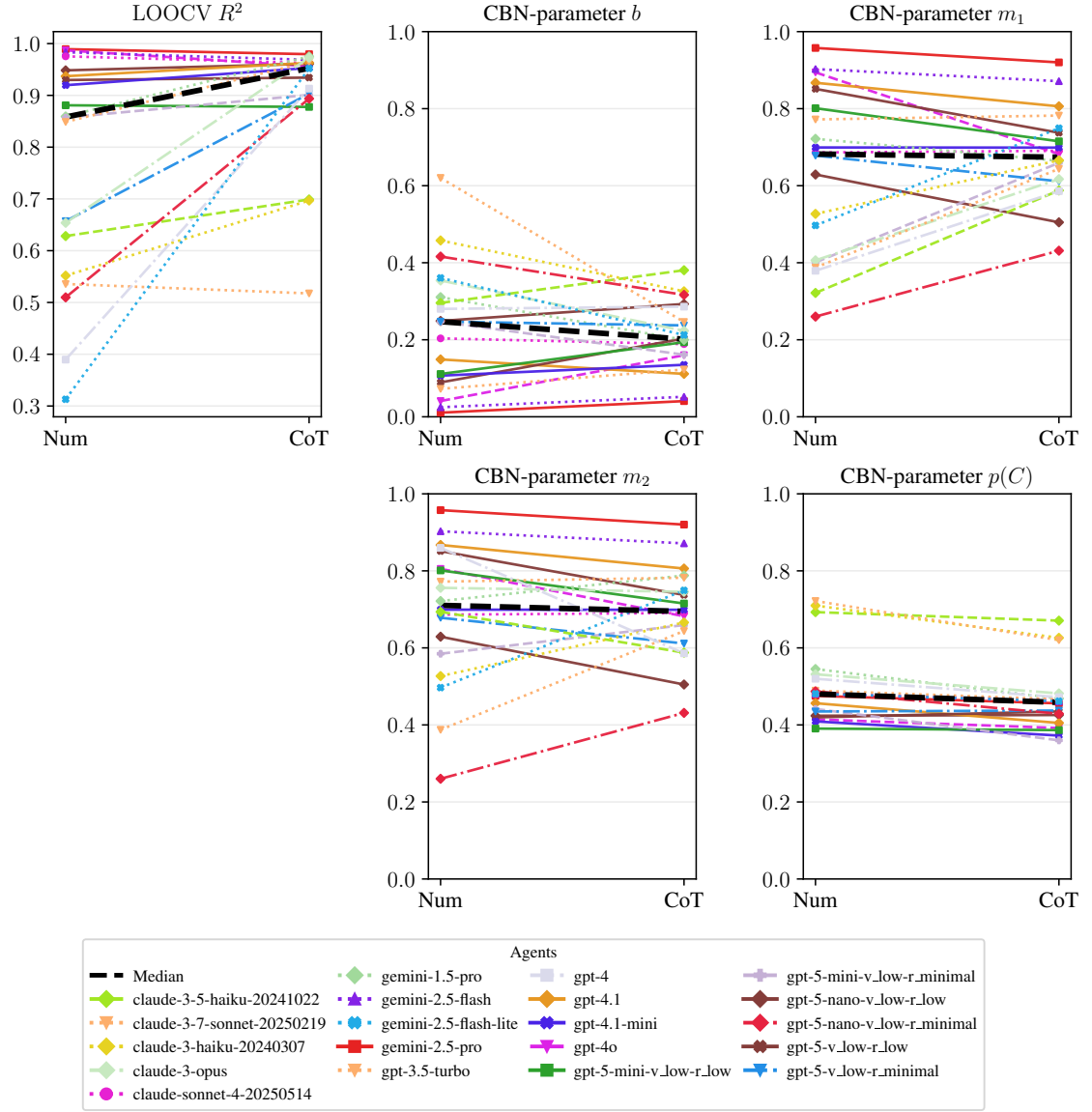
Experiment Abstract: Numeric  $\rightarrow$  CoT; N=27 matched agents



**Figure B.21: Pairwise prompt-category comparisons with fixed experiment, here abstract.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## Appendix B Additional Results

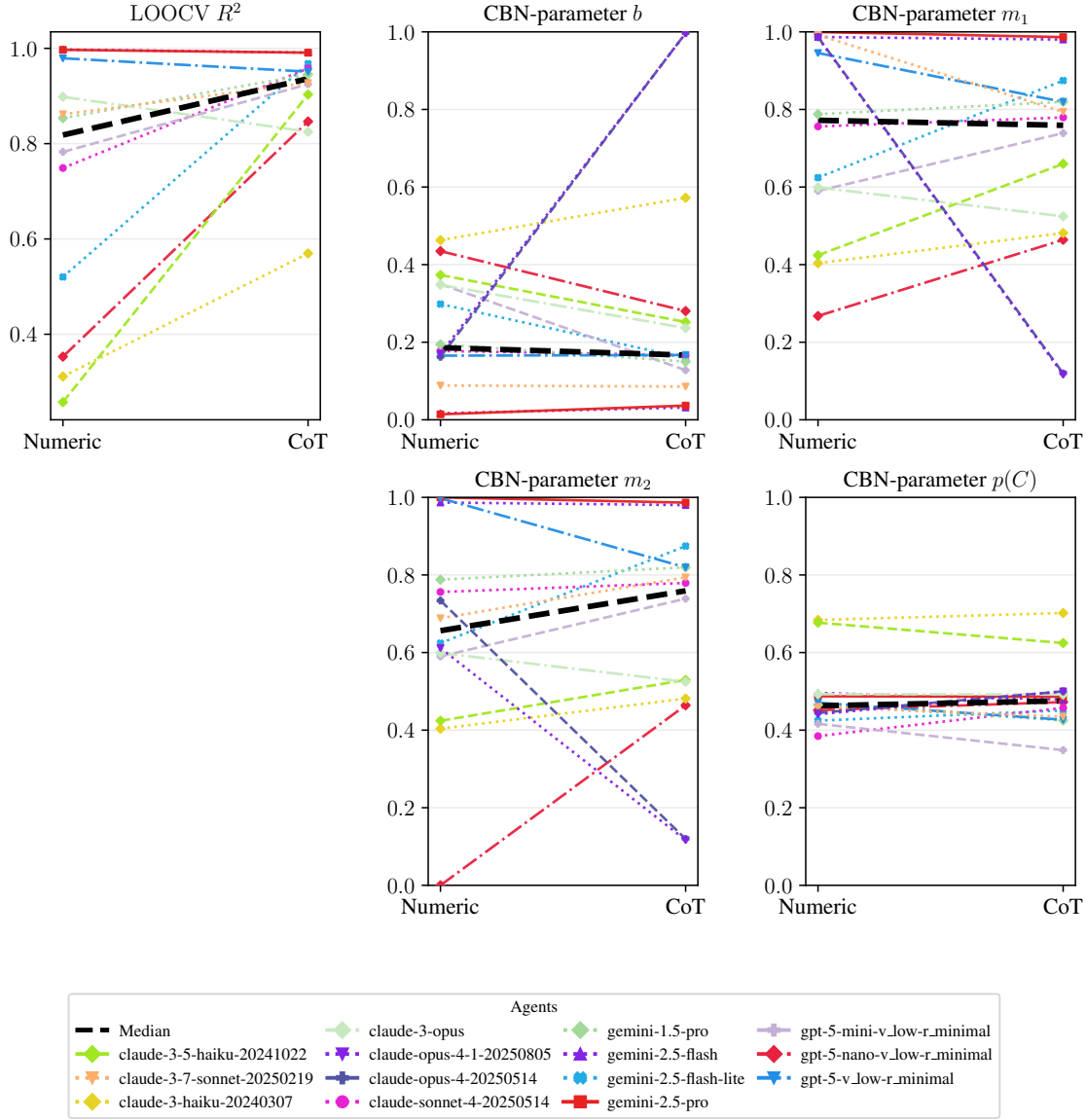
Experiment RW17-Over: Numeric  $\rightarrow$  CoT; N=20 matched agents



**Figure B.22: Pairwise prompt-category comparisons with fixed experiment, here RW17-overloaded.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## B.6 Most and Least changing LLMs across prompt-category & content manipulations

Experiment Abstract-Overloaded: Numeric  $\rightarrow$  CoT; N=14 matched agents



**Figure B.23: Pairwise prompt-category comparisons with fixed experiment, here abstract-overloaded.** Each panel shows LOOCV  $R^2$  (top left) and CBN parameters (rest) for each agent (color-coded) in two conditions (left/right per panel). These plots allow to *visually identify outliers and the most changing agents between conditions (i.e., agents with a slope)*.

## Appendix B Additional Results

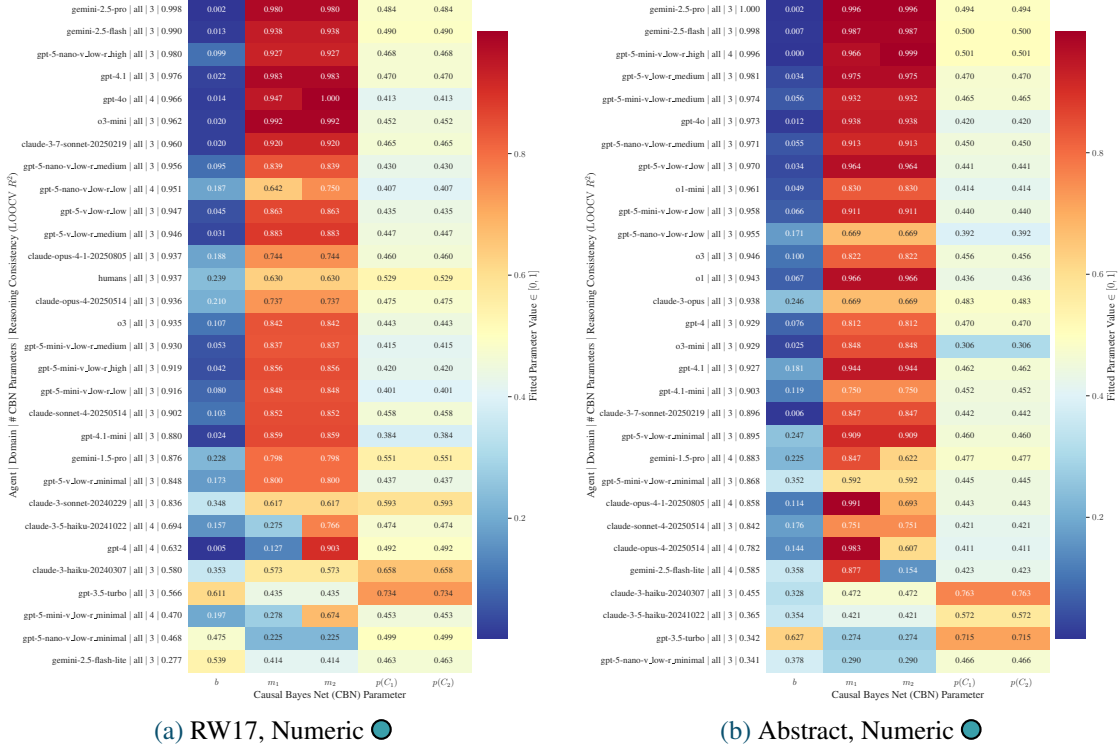
**Table B.18:** Top-3 most/least-changing agents per prompt-category (PC) comparisons (by signed  $\Delta$ ;  $\Delta = A - B$ ) for given experiment and CBN parameter (Param).

Experiment	PC Pair	Param	Most	Least
Abstract-Over	Num $\rightarrow$ CoT	$b$	claude-opus-4-20250514 (+0.835); claude-opus-4-1-20250805 (+0.826); gpt-5-mini-v_low-r_minimal (-0.222)	gpt-5-v_low-r_minimal (+0.002); claude-3-7-sonnet-20250219 (-0.003); claude-sonnet-4-20250514 (-0.011)
Abstract-Over	Num $\rightarrow$ CoT	$m_1$	claude-opus-4-1-20250805 (-0.869); claude-opus-4-20250514 (-0.867); gemini-2.5-flash-lite (+0.250)	gemini-2.5-flash (-0.007); gemini-2.5-pro (-0.013); claude-sonnet-4-20250514 (+0.023)
Abstract-Over	Num $\rightarrow$ CoT	$m_2$	claude-opus-4-20250514 (-0.615); claude-opus-4-1-20250805 (-0.495); gpt-5-nano-v_low-r_minimal (+0.464)	gemini-2.5-flash (-0.007); gemini-2.5-pro (-0.013); claude-sonnet-4-20250514 (+0.023)
Abstract-Over	Num $\rightarrow$ CoT	$p(C)$	claude-sonnet-4-20250514 (+0.073); gpt-5-mini-v_low-r_minimal (-0.068); claude-opus-4-1-20250805 (+0.059)	claude-3-opus (-0.001); gemini-2.5-pro (-0.001); gemini-2.5-flash (-0.017)
Abstract	Num $\rightarrow$ CoT	$b$	gpt-5-mini-v_low-r_minimal (-0.223); gemini-2.5-flash-lite (-0.208); claude-3-haiku-20240307 (+0.202)	claude-opus-4-20250514 (+0.005); gpt-3.5-turbo (-0.005); claude-3-opus (-0.011)
Abstract	Num $\rightarrow$ CoT	$m_1$	gpt-5-nano-v_low-r_medium (-0.477); gpt-5-nano-v_low-r_minimal (+0.307); gpt-3.5-turbo (+0.215)	gemini-2.5-flash (-0.009); gpt-5-nano-v_low-r_low (+0.009); gemini-2.5-pro (-0.010)
Abstract	Num $\rightarrow$ CoT	$m_2$	gemini-2.5-flash-lite (+0.680); claude-opus-4-20250514 (+0.294); gpt-3.5-turbo (+0.215)	gemini-2.5-pro (+0.004); gemini-2.5-flash (-0.009); claude-sonnet-4-20250514 (+0.022)
Abstract	Num $\rightarrow$ CoT	$p(C)$	gpt-4.1-mini (-0.105); claude-3-haiku-20240307 (-0.069); gpt-5-mini-v_low-r_minimal (-0.068)	gemini-2.5-pro (-0.000); o1 (-0.002); claude-3-7-sonnet-20250219 (-0.003)
RW17	Num $\rightarrow$ CoT	$b$	gemini-2.5-flash-lite (-0.349); gpt-5-v_low-r_medium (+0.161); gpt-4 (+0.158)	claude-opus-4-1-20250805 (-0.029); gemini-2.5-flash (+0.031); o3 (+0.031)
RW17	Num $\rightarrow$ CoT	$m_1$	gpt-5-mini-v_low-r_minimal (+0.425); gpt-4 (+0.360); gemini-2.5-flash-lite (+0.349)	claude-opus-4-1-20250805 (-0.004); claude-opus-4-20250514 (-0.022); gemini-2.5-flash (-0.022)
RW17	Num $\rightarrow$ CoT	$m_2$	gemini-2.5-flash-lite (+0.349); gpt-5-v_low-r_medium (-0.247); gpt-5-nano-v_low-r_minimal (+0.229)	claude-opus-4-1-20250805 (-0.004); gemini-1.5-pro (+0.010); gemini-2.5-flash (-0.022)
RW17	Num $\rightarrow$ CoT	$p(C)$	claude-3-5-haiku-20241022 (+0.190); gemini-1.5-pro (-0.105); claude-3-sonnet-20240229 (-0.085)	o3-mini (+0.000); gpt-5-v_low-r_low (+0.001); gpt-5-mini-v_low-r_medium (-0.002)
RW17-Over	Num $\rightarrow$ CoT	$b$	gpt-3.5-turbo (-0.374); gemini-2.5-flash-lite (-0.148); claude-3-haiku-20240307 (-0.132)	gpt-4 (+0.006); gpt-5-v_low-r_minimal (-0.010); claude-sonnet-4-20250514 (-0.015)
RW17-Over	Num $\rightarrow$ CoT	$m_1$	claude-3-5-haiku-20241022 (+0.266); gpt-5-mini-v_low-r_minimal (+0.258); gpt-3.5-turbo (+0.256)	gpt-4.1-mini (-0.000); claude-sonnet-4-20250514 (+0.004); claude-3-7-sonnet-20250219 (+0.010)
RW17-Over	Num $\rightarrow$ CoT	$m_2$	gpt-4 (-0.274); gpt-3.5-turbo (+0.256); gemini-2.5-flash-lite (+0.252)	gpt-4.1-mini (-0.000); claude-sonnet-4-20250514 (+0.004); claude-3-7-sonnet-20250219 (+0.010)
RW17-Over	Num $\rightarrow$ CoT	$p(C)$	gpt-3.5-turbo (-0.102); claude-3-haiku-20240307 (-0.084); gpt-5-mini-v_low-r_minimal (-0.082)	gpt-5-nano-v_low-r_low (+0.002); gpt-5-v_low-r_minimal (+0.002); gpt-5-mini-v_low-r_low (-0.004)

## B.7 Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments

### B.7 Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments

#### B.7.1 RW17 and Abstract, Numeric Prompts



**Figure B.24: Parameter Values of best fitting causal Bayes Nets (CBN) for Numeric Prompts in RW17 and Abstract experiments.** Panels (a) and (b) show RW17 prompts (human baseline available) and Abstract prompts, respectively. Each row is an agent ordered according to reasoning consistency (LOOCV  $R^2 \in [-\infty, 1]$ ). Columns (left to right) are: leak (background probability of effect  $p(E = 1 | C_1 = 0, C_2 = 0)$ ), causal strengths  $m_1$  and  $m_2$  (larger values = stronger/more deterministic influence), and prior probabilities of causes. Parameter values live in  $[0, 1]$ .

## Appendix B Additional Results

**Table B.19:** Best CBN fits per agent for RW17 in Figure B.24 (loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 0.5]$  with  $\delta = 1$

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-2.5-pro	all	3	<b>0.024</b>	<b>0.052</b>	0.032	<b>0.981</b>	<b>0.998</b>	<b>0.016</b>
gemini-2.5-flash	all	3	0.042	0.077	0.028	0.955	0.990	0.036
gpt-5-nano-v_low-r_high	all	3	0.085	0.126	0.028	0.874	0.980	0.049
gpt-4.1	all	3	0.042	0.091	0.033	0.944	0.976	0.060
gpt-4o	all	4	0.074	0.125	0.037	0.897	0.966	0.071
o3-mini	all	3	0.053	0.102	0.036	0.930	0.962	0.076
claude-3-7-sonnet-20250219	all	3	0.074	0.117	0.032	0.897	0.960	0.074
gpt-5-nano-v_low-r_medium	all	3	0.099	0.115	0.022	0.869	0.956	0.066
gpt-5-nano-v_low-r_low	all	4	0.083	0.115	<b>0.013</b>	0.803	0.951	0.056
gpt-5-v_low-r_low	all	3	0.076	0.089	0.024	0.927	0.947	0.077
gpt-5-v_low-r_medium	all	3	0.075	0.088	0.026	0.934	0.946	0.081
claude-opus-4-1-20250805	all	3	0.096	0.128	0.016	0.792	0.937	0.067
claude-opus-4-20250514	all	3	0.090	0.125	0.014	0.790	0.936	0.067
o3	all	3	0.076	0.089	0.019	0.916	0.935	0.079
gpt-5-mini-v_low-r_medium	all	3	0.080	0.095	0.023	0.912	0.930	0.086
gpt-5-mini-v_low-r_high	all	3	0.086	0.101	0.025	0.907	0.919	0.096
gpt-5-mini-v_low-r_low	all	3	0.083	0.098	0.023	0.904	0.916	0.094
claude-sonnet-4-20250514	all	3	0.088	0.102	0.021	0.897	0.902	0.103
gpt-4.1-mini	all	3	0.122	0.177	0.040	0.776	0.880	0.124
gemini-1.5-pro	all	3	0.147	0.190	0.025	0.666	0.876	0.103
gpt-5-v_low-r_minimal	all	3	0.120	0.144	0.023	0.784	0.848	0.120
claude-3-sonnet-20240229	all	3	0.136	0.166	0.017	0.533	0.836	0.081
claude-3-5-haiku-20241022	all	4	0.121	0.167	0.018	0.529	0.694	0.123
gpt-4	all	4	0.065	0.129	0.019	0.726	0.632	0.145
claude-3-haiku-20240307	all	3	0.154	0.190	0.026	0.400	0.580	0.136
gpt-3.5-turbo	all	3	0.089	0.109	0.027	0.363	0.566	0.068
gpt-5-mini-v_low-r_minimal	all	4	0.106	0.139	<b>0.013</b>	0.565	0.470	0.145
gpt-5-nano-v_low-r_minimal	all	3	0.142	0.197	0.023	0.108	0.468	0.071
gemini-2.5-flash-lite	all	3	0.205	0.271	0.037	0.231	0.277	0.187



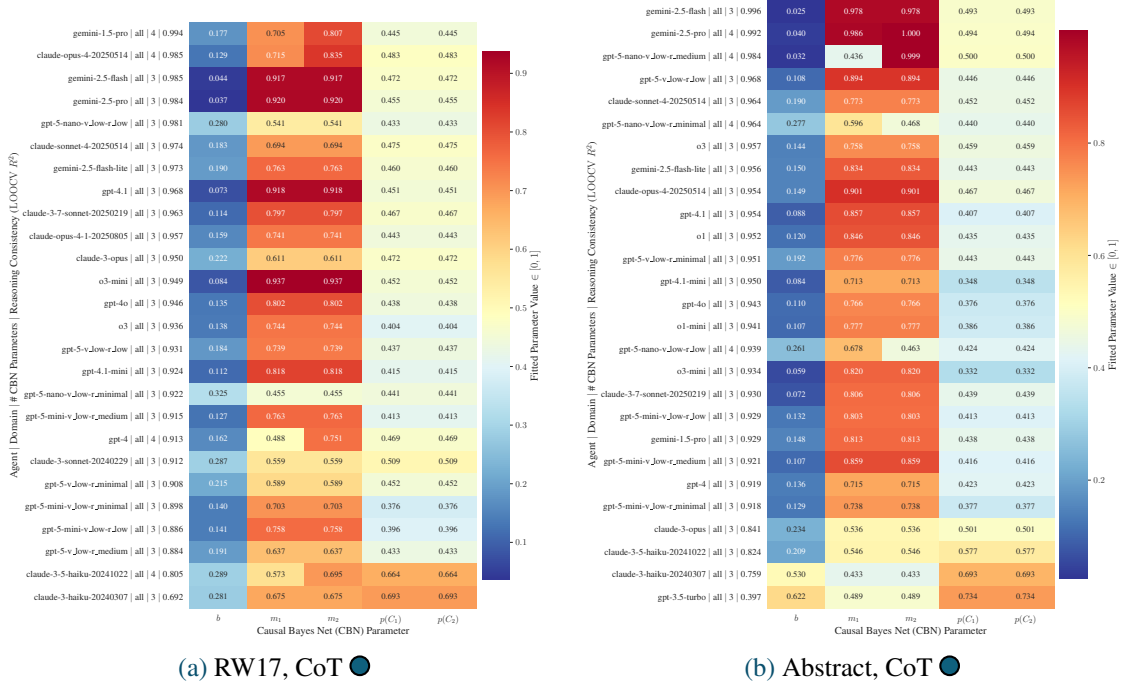
## B.7 Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments

**Table B.20:** Best CBN fits per agent for Abstract in Figure B.24 (loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 05]$  with  $\delta = 1$

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-2.5-pro	all	3	<b>0.009</b>	0.030	0.032	0.994	<b>1.000</b>	<b>0.008</b>
gemini-2.5-flash	all	3	0.012	0.038	0.033	0.990	0.998	0.018
gpt-5-mini-v_low-r_high	all	4	0.015	<b>0.027</b>	0.031	<b>0.995</b>	0.996	0.025
gpt-5-v_low-r_medium	all	3	0.043	0.067	0.030	0.967	0.981	0.053
gpt-5-mini-v_low-r_medium	all	3	0.056	0.078	0.026	0.951	0.974	0.057
gpt-4o	all	3	0.103	0.184	0.044	0.792	0.973	0.062
gpt-5-nano-v_low-r_medium	all	3	0.075	0.099	0.027	0.919	0.971	0.059
gpt-5-v_low-r_low	all	3	0.061	0.080	0.030	0.953	0.970	0.065
o1-mini	all	3	0.085	0.123	0.026	0.862	0.961	0.064
gpt-5-mini-v_low-r_low	all	3	0.067	0.083	0.024	0.940	0.958	0.071
gpt-5-nano-v_low-r_low	all	3	0.102	0.145	0.018	0.715	0.955	0.053
o3	all	3	0.067	0.085	0.017	0.920	0.946	0.071
o1	all	3	0.066	0.084	0.029	0.947	0.943	0.088
claude-3-opus	all	3	0.112	0.167	0.018	0.625	0.938	0.060
gpt-4	all	3	0.128	0.177	0.029	0.730	0.929	0.084
o3-mini	all	3	0.110	0.144	0.040	0.847	0.929	0.097
gpt-4.1	all	3	0.085	0.127	0.028	0.868	0.927	0.093
gpt-4.1-mini	all	3	0.166	0.240	0.039	0.547	0.903	0.092
claude-3-7-sonnet-20250219	all	3	0.095	0.143	0.034	0.843	0.896	0.116
gpt-5-v_low-r_minimal	all	3	0.104	0.144	0.025	0.817	0.895	0.105
gemini-1.5-pro	all	4	0.143	0.216	0.033	0.601	0.883	0.098
gpt-5-mini-v_low-r_minimal	all	3	0.119	0.163	<b>0.015</b>	0.572	0.868	0.078
claude-opus-4-1-20250805	all	4	0.081	0.123	0.035	0.885	0.858	0.131
claude-sonnet-4-20250514	all	3	0.116	0.158	0.023	0.731	0.842	0.117
claude-opus-4-20250514	all	4	0.083	0.122	0.035	0.873	0.782	0.156
gemini-2.5-flash-lite	all	4	0.185	0.253	0.045	0.491	0.585	0.187
claude-3-haiku-20240307	all	3	0.279	0.313	0.077	0.140	0.455	0.141
claude-3-5-haiku-20241022	all	3	0.239	0.292	0.047	0.160	0.365	0.158
gpt-3.5-turbo	all	3	0.168	0.216	0.045	0.049	0.342	0.058
gpt-5-nano-v_low-r_minimal	all	3	0.221	0.271	0.040	0.090	0.341	0.089

## Appendix B Additional Results

### B.7.2 RW17 and Abstract, CoT Prompts



**Figure B.25: Parameter Values of best fitting causal Bayes Nets (CBN) for CoT Prompts in RW17 and Abstract experiments.** Panels (a) and (b) show RW17 prompts (human baseline available) and Abstract prompts, respectively. Each row is an agent ordered according to reasoning consistency (LOOCV  $R^2 \in [-\inf, 1]$ ) Columns (left to right) are: leak (background probability of effect  $p(E = 1 \mid C_1 = 0, C_2 = 0)$ ), causal strengths  $m_1$  and  $m_2$  (larger values = stronger/more deterministic influence), and prior probabilities of causes. Parameter values live in  $[0, 1]$ .

## B.7 Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments

**Table B.21:** Best CBN fits per agent for RW17 in Figure B.25; (loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 0.5]$  with  $\delta = 1$

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-1.5-pro	all	4	0.066	0.088	0.011	0.892	<b>0.994</b>	<b>0.020</b>
claude-opus-4-20250514	all	4	<b>0.039</b>	0.070	0.010	0.922	0.985	0.033
gemini-2.5-flash	all	3	0.047	0.071	0.025	<b>0.958</b>	0.985	0.044
gemini-2.5-pro	all	3	0.056	0.075	0.026	0.955	0.984	0.046
gpt-5-nano-v_low-r_low	all	3	0.076	0.099	<b>0.007</b>	0.751	0.981	0.026
claude-sonnet-4-20250514	all	3	0.045	<b>0.058</b>	<b>0.007</b>	0.938	0.974	0.039
gemini-2.5-flash-lite	all	3	0.071	0.093	0.013	0.887	0.973	0.045
gpt-4.1	all	3	0.050	0.069	0.023	<b>0.958</b>	0.968	0.063
claude-3-7-sonnet-20250219	all	3	0.053	0.066	0.014	0.946	0.963	0.057
claude-opus-4-1-20250805	all	3	0.056	0.067	0.010	0.932	0.957	0.055
claude-3-opus	all	3	0.075	0.104	0.009	0.778	0.950	0.049
o3-mini	all	3	0.069	0.099	0.027	0.920	0.949	0.080
gpt-4o	all	3	0.070	0.090	0.016	0.904	0.946	0.070
o3	all	3	0.073	0.085	0.014	0.899	0.936	0.070
gpt-5-v_low-r_low	all	3	0.077	0.091	0.012	0.880	0.931	0.069
gpt-4.1-mini	all	3	0.069	0.089	0.018	0.911	0.924	0.084
gpt-5-nano-v_low-r_minimal	all	3	0.091	0.115	0.008	0.617	0.922	0.046
gpt-5-mini-v_low-r_medium	all	3	0.076	0.090	0.015	0.896	0.915	0.083
gpt-4	all	4	0.064	0.088	0.008	0.835	0.913	0.066
claude-3-sonnet-20240229	all	3	0.098	0.130	0.010	0.646	0.912	0.059
gpt-5-v_low-r_minimal	all	3	0.065	0.078	<b>0.007</b>	0.850	0.908	0.064
gpt-5-mini-v_low-r_minimal	all	3	0.081	0.103	0.015	0.843	0.898	0.085
gpt-5-mini-v_low-r_low	all	3	0.079	0.093	0.016	0.883	0.886	0.095
gpt-5-v_low-r_medium	all	3	0.077	0.094	0.009	0.812	0.884	0.078
claude-3-5-haiku-20241022	all	4	0.107	0.136	0.019	0.595	0.805	0.084
claude-3-haiku-20240307	all	3	0.124	0.162	0.025	0.564	0.692	0.122

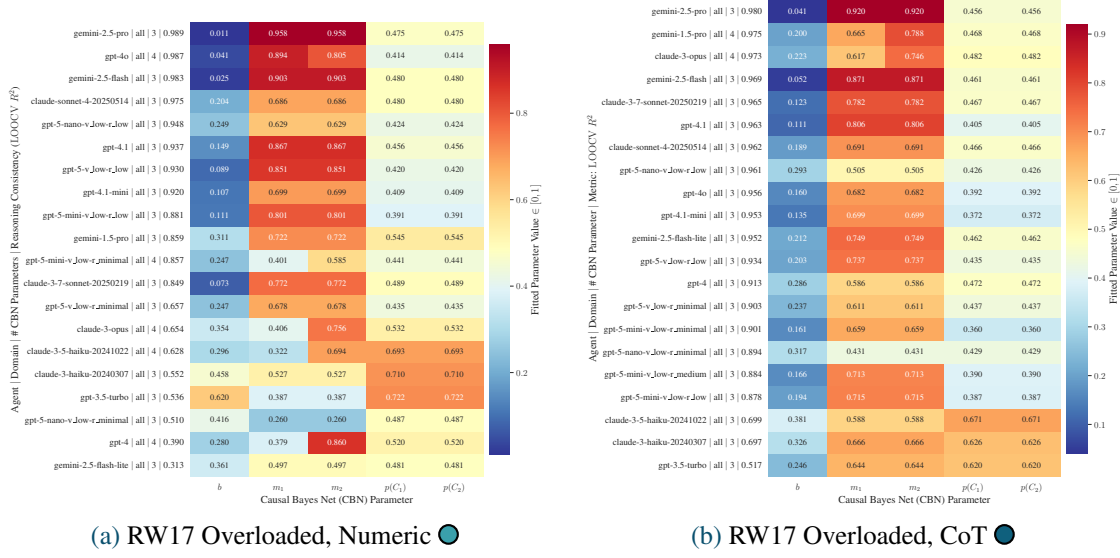
## Appendix B Additional Results

**Table B.22:** Best CBN fits per agent for Abstract in Figure B.25; (loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 0.5]$  with  $\delta = 1$

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-2.5-flash	all	3	0.024	0.047	0.030	0.984	<b>0.996</b>	<b>0.023</b>
gemini-2.5-pro	all	4	<b>0.020</b>	<b>0.039</b>	0.030	<b>0.989</b>	0.992	0.035
gpt-5-nano-v_low-r_medium	all	4	0.025	0.043	0.014	0.945	0.984	<b>0.023</b>
gpt-5-v_low-r_low	all	3	0.066	0.081	0.020	0.938	0.968	0.060
claude-sonnet-4-20250514	all	3	0.058	0.076	<b>0.011</b>	0.920	0.964	0.052
gpt-5-nano-v_low-r_minimal	all	4	0.094	0.123	<b>0.011</b>	0.681	0.964	0.037
o3	all	3	0.053	0.065	<b>0.011</b>	0.941	0.957	0.057
gemini-2.5-flash-lite	all	3	0.075	0.098	0.018	0.899	0.956	0.065
claude-opus-4-20250514	all	3	0.042	0.056	0.019	0.971	0.954	0.068
gpt-4.1	all	3	0.060	0.080	0.021	0.936	0.954	0.071
o1	all	3	0.060	0.075	0.017	0.939	0.952	0.068
gpt-5-v_low-r_minimal	all	3	0.065	0.080	0.012	0.911	0.951	0.062
gpt-4.1-mini	all	3	0.069	0.086	0.017	0.904	0.950	0.064
gpt-4o	all	3	0.077	0.105	0.019	0.871	0.943	0.070
o1-mini	all	3	0.079	0.099	0.018	0.883	0.941	0.071
gpt-5-nano-v_low-r_low	all	4	0.091	0.133	0.015	0.692	0.939	0.054
o3-mini	all	3	0.088	0.111	0.028	0.887	0.934	0.085
claude-3-7-sonnet-20250219	all	3	0.065	0.087	0.020	0.921	0.930	0.084
gpt-5-mini-v_low-r_low	all	3	0.073	0.090	0.017	0.905	0.929	0.079
gemini-1.5-pro	all	3	0.090	0.131	0.020	0.825	0.929	0.080
gpt-5-mini-v_low-r_medium	all	3	0.077	0.095	0.020	0.908	0.921	0.090
gpt-4	all	3	0.084	0.116	0.017	0.828	0.919	0.080
gpt-5-mini-v_low-r_minimal	all	3	0.078	0.100	0.016	0.867	0.918	0.080
claude-3-opus	all	3	0.116	0.151	0.014	0.552	0.841	0.081
claude-3-5-haiku-20241022	all	3	0.153	0.191	0.024	0.443	0.824	0.086
claude-3-haiku-20240307	all	3	0.188	0.229	0.039	0.141	0.759	0.056
gpt-3.5-turbo	all	3	0.221	0.282	0.058	0.109	0.397	0.105

## B.7 Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments

### B.7.3 RW17-Overloaded, Numeric and CoT Prompts



**Figure B.26: Parameter Values of best fitting causal Bayes Nets (CBN) for CoT Prompts in RW17 and Abstract experiments.** Panels (a) and (b) show RW17 prompts (human baseline available) and Abstract prompts, respectively. Each row is an agent ordered according to reasoning consistency (LOOCV  $R^2 \in [-\inf, 1]$ ) Columns (left to right) are: leak (background probability of effect  $p(E = 1 \mid C_1 = 0, C_2 = 0)$ ), causal strengths  $m_1$  and  $m_2$  (larger values = stronger/more deterministic influence), and prior probabilities of causes. Parameter values live in  $[0, 1]$ .

## Appendix B Additional Results

**Table B.23:** Best CBN fits per agent for RW17 Overloaded in Figure B.26; (Numeric, loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 0.5]$  with  $\delta = 1$

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-2.5-pro	all	3	<b>0.043</b>	<b>0.081</b>	0.031	<b>0.953</b>	<b>0.989</b>	<b>0.039</b>
gpt-4o	all	4	0.135	0.188	0.040	0.754	0.987	0.040
gemini-2.5-flash	all	3	0.051	0.082	0.026	0.944	0.983	0.045
claude-sonnet-4-20250514	all	3	0.059	0.099	0.010	0.840	0.975	<b>0.039</b>
gpt-5-nano-v_low-r_low	all	3	0.085	0.117	0.011	0.754	0.948	0.051
gpt-4.1	all	3	0.100	0.125	0.023	0.852	0.937	0.081
gpt-5-v_low-r_low	all	3	0.080	0.092	0.021	0.915	0.930	0.086
gpt-4.1-mini	all	3	0.146	0.187	0.028	0.648	0.920	0.079
gpt-5-mini-v_low-r_low	all	3	0.085	0.105	0.019	0.875	0.881	0.104
gemini-1.5-pro	all	3	0.135	0.182	0.020	0.616	0.859	0.096
gpt-5-mini-v_low-r_minimal	all	4	0.125	0.162	0.016	0.481	0.857	0.070
claude-3-7-sonnet-20250219	all	3	0.127	0.167	0.025	0.728	0.849	0.117
gpt-5-v_low-r_minimal	all	3	0.131	0.162	0.019	0.657	0.657	0.156
claude-3-opus	all	4	0.101	0.125	<b>0.009</b>	0.611	0.654	0.104
claude-3-5-haiku-20241022	all	4	0.103	0.134	0.027	0.442	0.628	0.101
claude-3-haiku-20240307	all	3	0.120	0.171	0.029	0.336	0.552	0.108
gpt-3.5-turbo	all	3	0.079	0.095	0.024	0.372	0.536	0.063
gpt-5-nano-v_low-r_minimal	all	3	0.166	0.222	0.028	0.108	0.510	0.070
gpt-4	all	4	0.117	0.162	0.016	0.533	0.390	0.159
gemini-2.5-flash-lite	all	3	0.233	0.284	0.041	0.262	0.313	0.208

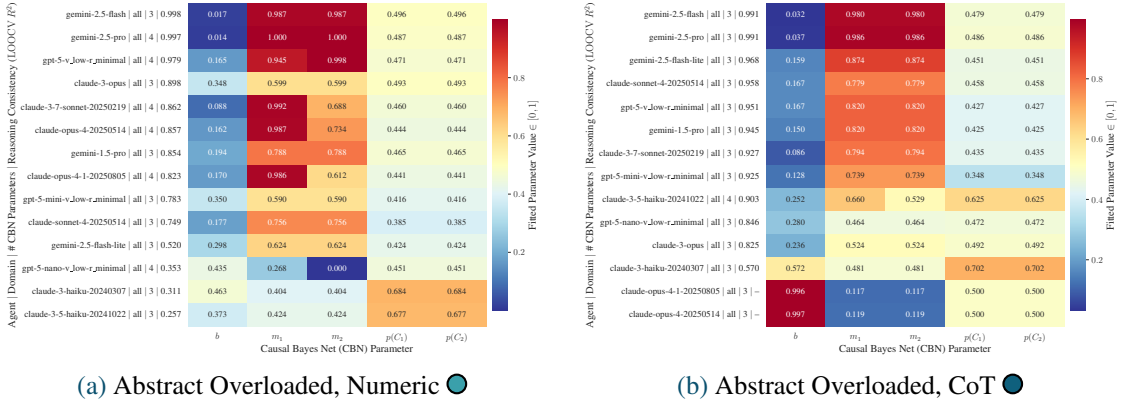
## B.7 Fitting Metrics and Parameter Values for Causal Bayesian Network (CBN) for all Agents and Experiments

Table B.24: Best CBN fits per agent for RW17 Overloaded in Figure B.26; (CoT, loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 0.5]$  with  $\delta = 1$

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-2.5-pro	all	3	0.057	0.075	0.025	<b>0.953</b>	<b>0.980</b>	0.051
gemini-1.5-pro	all	4	0.067	0.096	0.010	0.860	0.975	0.041
claude-3-opus	all	4	0.076	0.105	0.009	0.804	0.973	0.038
gemini-2.5-flash	all	3	0.057	0.080	0.022	0.941	0.969	0.059
claude-3-7-sonnet-20250219	all	3	0.051	<b>0.063</b>	0.013	0.948	0.965	0.054
gpt-4.1	all	3	0.063	0.083	0.017	0.922	0.963	0.059
claude-sonnet-4-20250514	all	3	<b>0.049</b>	0.067	0.008	0.919	0.962	0.047
gpt-5-nano-v_low-r_low	all	3	0.081	0.103	0.008	0.711	0.961	<b>0.035</b>
gpt-4o	all	3	0.081	0.098	0.013	0.852	0.956	0.054
gpt-4.1-mini	all	3	0.073	0.089	0.013	0.884	0.953	0.057
gemini-2.5-flash-lite	all	3	0.071	0.094	0.012	0.879	0.952	0.060
gpt-5-v_low-r_low	all	3	0.070	0.084	0.011	0.890	0.934	0.067
gpt-4	all	3	0.076	0.108	0.008	0.735	0.913	0.061
gpt-5-v_low-r_minimal	all	3	0.063	0.078	<b>0.007</b>	0.862	0.903	0.068
gpt-5-mini-v_low-r_minimal	all	3	0.076	0.093	0.013	0.850	0.901	0.078
gpt-5-nano-v_low-r_minimal	all	3	0.088	0.111	0.008	0.615	0.894	0.052
gpt-5-mini-v_low-r_medium	all	3	0.083	0.099	0.014	0.852	0.884	0.090
gpt-5-mini-v_low-r_low	all	3	0.082	0.097	0.013	0.855	0.878	0.091
claude-3-5-haiku-20241022	all	3	0.106	0.133	0.018	0.556	0.699	0.098
claude-3-haiku-20240307	all	3	0.130	0.170	0.020	0.557	0.697	0.121
gpt-3.5-turbo	all	3	0.228	0.278	0.046	0.334	0.517	0.178

### B.7.4 Abstract-Overloaded, Numeric and CoT Prompts

## Appendix B Additional Results



**Figure B.27: Parameter Values of best fitting causal Bayes Nets (CBN) for CoT Prompts in abstract and Abstract experiments.** Panels (a) and (b) show abstract prompts (human baseline available) and Abstract prompts, respectively. Each row is an agent ordered according to reasoning consistency ( $LOOCV R^2 \in [-\infty, 1]$ ). Columns (left to right) are: leak (background probability of effect  $p(E = 1 | C_1 = 0, C_2 = 0)$ ), causal strengths  $m_1$  and  $m_2$  (larger values = stronger/more deterministic influence), and prior probabilities of causes. Parameter values live in  $[0, 1]$ .

**Table B.25: Best CBN fits per agent for Abstract Overloaded in Figure B.27; (Numeric, loss: huber; optimizer: lbfgs; link-function: noisy-or; learning rate: 0.100). MAE and RMSE  $\in [0, 1]$ , Huber loss  $\sim \in [0, 05]$  with  $\delta = 1$**

Agent	Domain	num params	MAE	RMSE	loss	$R^2$	LOOCV $R^2$	LOOCV RMSE
gemini-2.5-flash	all	3	0.016	<b>0.028</b>	0.031	<b>0.995</b>	<b>0.998</b>	<b>0.019</b>
gemini-2.5-pro	all	4	<b>0.015</b>	0.049	0.033	0.984	0.997	0.022
gpt-5-v_low-r_minimal	all	4	0.079	0.127	0.028	0.875	0.979	0.051
claude-3-opus	all	3	0.148	0.198	<b>0.021</b>	0.468	0.898	0.068
claude-3-7-sonnet-20250219	all	4	0.073	0.115	0.031	0.893	0.862	0.130
claude-opus-4-20250514	all	4	0.082	0.143	0.031	0.829	0.857	0.129
gemini-1.5-pro	all	3	0.149	0.221	0.034	0.595	0.854	0.112
claude-opus-4-1-20250805	all	4	0.084	0.116	0.027	0.870	0.823	0.138
gpt-5-mini-v_low-r_minimal	all	3	0.145	0.200	0.022	0.486	0.783	0.108
claude-sonnet-4-20250514	all	3	0.131	0.169	0.027	0.709	0.749	0.152
gemini-2.5-flash-lite	all	3	0.236	0.302	0.050	0.344	0.520	0.193
gpt-5-nano-v_low-r_minimal	all	4	0.289	0.325	0.061	0.039	0.353	0.067
claude-3-haiku-20240307	all	3	0.228	0.274	0.051	0.110	0.311	0.128
claude-3-5-haiku-20241022	all	3	0.273	0.316	0.063	0.115	0.257	0.164



## Declaration of Generative AI Usage

This work made use of generative AI for grammar and spell checking and to support software development such as plotting code, in accordance with points 3.2 and 3.3 of the attached statement of authorship. [Table C.1](#) provides the used programs together with their version numbers.

**Table C.1:** Generative AI programs and their version numbers used in this work.

Program	Version
LLM	gpt-4o
LLM	Claude-Sonnet-4
LLM	Claude-Sonnet-3.7
LLM	Claude-Opus-4.1

