UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Sandor Vunk

Causal Information Extraction Using Large Language Models

Bachelor's Thesis (9 ECTS)

Supervisor:
Giacomo Magnifico, MA

Tartu 2025

Causal Information Extraction Using Large Language Models

Abstract:

This thesis investigates the ability of Large Language Models (LLMs) for causal information extraction, an important task for high-level natural language comprehension. In a controlled experiment of eight flagship models of leading AI organizations — including OpenAI's GPT-o3, Anthropic's Claude 3.7 Sonnet, xAI's Grok-3, and others — this study examines both their ability to extract cause-effect pairs from text and their performance at evaluating such extractions.

A purpose-designed multi-domain dataset was generated to serve this end, with controlled causal relations hidden in contexts with diverse complexity levels, covering economics, environmental science, and technology domains. The dataset incorporates a number of difficult variations achieved through the use of cue masking and pair shuffling methods.

By applying a zero-shot approach with standardized prompting, a twin evaluation framework is employed that uses traditional human evaluation with a model-based semantic scoring system, in which LLMs score other LLM's extractions. This provides a more informative model performance evaluation.

Results revealed impressive causal extraction capabilities across all models, with leading models, outperforming smaller models. Especially notable were OpenAI's GPT-o3, Antropic's Claude 3.7 Sonnet and xAI's Grok-3, outperforming its counterparts. Overall, models demonstrated semantic understanding beyond reliance on explicit linguistic markers, though pair shuffling showed some dependence on pre-trained associations.

This research illuminates the capabilities of state-of-the-art LLMs in causal information extraction, establishing a foundation for enhanced causal reasoning systems across diverse domains.

Keywords: Causal information extraction, large language models, natural language processing, artificial intelligence, zero-shot evaluation

CERCS: P176 Artificial intelligence, P170 Computer science, numerical analysis, systems, control

Põhjusliku Informatsiooni Ekstraheerimine Kasutades Suuri Keelemudeleid

Lühikokkuvõte:

Käesolev töö uurib suurte keelemudelite (LLM-ide) võimekust põhjusliku informatsiooni eraldamisel, mis on oluline ülesanne kõrgetasemeliseks loomuliku keele mõistmiseks. Kontrollitud eksperimendis, mis hõlmab kaheksat juhtivate tehisintellekti organisatsioonide lipulaevmudelit — sealhulgas OpenAI GPT-o3, Anthropic'u Claude 3.7 Sonnet, xAI Grok-3 ja teisi — uuritakse nii nende võimet ekstraheerida tekstist põhjus-tagajärg paare kui ka hinnata selliseid ekstraktsioone.

Selle eesmärgi tarbeks loodi spetsiaalne mitut valdkonda hõlmav andmestik, kus kontrollitud põhjuslikud seosed on peidetud erineva keerukusastmega kontekstidesse, hõlmates majanduse, keskkonnateaduse ja tehnoloogia valdkondi. Andmestik sisaldab mitmeid keerukaid variatsioone, mis on saavutatud vihjete maskeerimise ja paaride segamise meetodite abil.

Rakendades nullõppe (ingl. *zero-shot*) lähenemisviisi standardiseeritud küsimustega, kasutatakse kahekordset hindamisraamistikku, mis kombineerib traditsioonilist inimhindamist mudelipõhise semantilise skoorimissüsteemiga, kus keelemudelid hindavad teiste keelemudelite ekstraktsioone. See võimaldab mudelite jõudlust informatiivsemalt hinnata.

Tulemused näitasid muljetavaldavat põhjuslike seoste eraldamise võimekust kõigi mudelite puhul, kusjuures juhtivad mudelid edestavad väiksemaid mudeleid. Eriti märkimisväärsed olid OpenAI GPT-o3, Anthropic'u Claude 3.7 Sonnet ja xAI Grok-3, mis ületasid teiste jõudlust. Üldiselt demonstreerisid mudelid semantilist arusaamist, mis ületab pelgalt eksplitsiitsete keeleliste märgete tuvastamise, kuigi paaride segamistestid näitasid sõltuvust eeltreenitud teadmiste osas.

Antud uuring selgitab tipptasemel suurte keelemudelite võimekust põhjusliku informatsiooni eraldamisel, luues aluse täiustatud põhjusliku arutlussüsteemide arendamiseks erinevates valdkondades.

Võtmesõnad: Põhjuslik informatsiooni ekstraheerimine, suured keelemudelid, loomuliku keele töötlus, tehisintellekt, nullõppe hindamine

CERCS: P176 Tehisintellekt, P170, Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Table of Contents

1. Introduction

Understanding causality is crucial for several tasks within natural language processing (NLP), such as predictive analytics, question answering, and decision-support systems. Causal information extraction (CIE) is an emerging subfield of NLP that focuses on identifying and extracting cause-and-effect relations from unstructured text to reveal how one event or entity influences another (Feder et al., 2022). By focusing on such relations rather than surface associations found in correlation-based text mining, CIE enables the construction of interpretable causal models and domain-knowledge graphs (Yang et al., 2022).

As NLP pipelines benefit from the reliability and interpretability provided by CIE, applications already span multiple domains. In biomedicine, causal extraction helps uncover treatment effects, disease aetiology and adverse drug interactions, forming knowledge graphs that support evidence-based practice (Yang et al., 2022).

Despite rapid progress, it remains unclear how reliably recent large-language models (LLMs) extract causal relations; this holds true especially when explicit lexical cues are missing or cause–effect pairs are embedded in distracting context. Therefore our research question is: *How accurately do modern large-language models extract causal information from text?*

To address the gap in current research defined by the aforementioned question, the present work introduces a multi-tier benchmark dataset and an automated zero-shot evaluation pipeline. The framework assesses eight state-of-the-art LLMs under increasingly challenging settings and quantifies their strengths and failure modes. The resulting benchmark provides a reproducible baseline for future research on causal information extraction.

The remainder of the thesis is organised as follows: Chapter 2 reviews related work; Chapter 3 describes the creation of the dataset and its properties Chapter 4 provides the evaluation methodology along with the experimental setup; Chapter 5 reports the results of the evaluation; Chapter 6 and 7 discuss the results of the evaluation, as well as presenting an overall conclusive analysis of the work presented in the thesis.

2. Background and Related Work

2.1 Pattern-Based and Rule-Based Approaches

Causal information extraction targets three primary forms of causality: explicit intra-sentential causality with overt lexical cues like "because" or "leads to" within a single sentence; implicit causality where connections lack explicit markers and must be inferred from context; and inter-sentential causality where cause and effect span multiple sentences (Yang et al., 2022). Early CIE systems addressed these through manually engineered linguistic patterns and syntactic rules (Asghar, 2016). While these deterministic approaches offered interpretability—extractions could be traced to specific triggers—they demonstrated significant limitations in recall performance, particularly with implicit causality, and insufficient cross-domain robustness. As textual corpora expanded in volume and diversity, these predominantly domain-constrained implementations revealed inherent scalability limitations that necessitated alternative methodologies (Asghar, 2016; Yang et al., 2022).

2.2 Statistical Models and Neural Networks

Methodologies for CIE evolved from knowledge-based approaches to statistical machine learning and neural systems in the mid-2010s (Asghar, 2016). Early ML systems automated causal indicator discovery but remained dependent on handcrafted features and expert knowledge, offering improved domain flexibility while still suffering from labor-intensive engineering and error propagation (Yang et al., 2022). Deep neural networks subsequently revolutionized CIE through distributed text representations, with architectures like RNNs, CNNs, LSTMs, and transformers automatically learning features to capture complex causal patterns. Modern transformer-based systems leverage self-attention mechanisms to identify sophisticated causal relations without explicit human-crafted indicators, achieving state-of-the-art performance (Ali et al., 2023). Despite these advances, neural approaches face significant challenges: substantial annotated data requirements, considerable computational resources, and reduced interpretability compared to knowledge-based methods. Ongoing research addresses these limitations through data augmentation, transfer learning, and architectural optimization to balance performance with resource requirements (Yang et al., 2022).

2.3 Hybrid Models

Contemporary causal information extraction methodologies synthesize knowledge-driven heuristics with data-driven paradigms through integrative frameworks that transcend singular methodological constraints. Sorgente et al. (2018) exemplified this convergence through a bifurcated architecture juxtaposing linguistic pattern recognition with Bayesian probabilistic filtering, thereby preserving interpretative transparency while quantifying relational ambiguities.

Knowledge-augmented frameworks substantively enhance extraction fidelity by incorporating ontological resources to preclude contradictions with established epistemological constructs. Seminal implementations leveraged lexical taxonomies (Girju & Moldovan, 2002), whereas contemporary architectures embed domain knowledge directly into computational graphs, as evidenced by knowledge-infused convolutional networks that assimilate exogenous semantic patterns (Ali et al., 2023). These epistemologically-enriched neural architectures demonstrate superior precision metrics compared to their autonomous counterparts.

These hybridized methodological approaches constitute a theoretical reconciliation that preserves structured linguistic insights while maintaining algorithmic adaptability. Extant literature indicates that the trajectory of causal information extraction resides in such integrative frameworks that harmonize linguistic formalism with empirical inference to extract high-fidelity causal representations (Ali et al., 2023; Sorgente et al., 2018).

2.4 Introduction to LLMs

Large Language Models (LLMs) are massive neural networks, typically Transformer-based, trained on enormous text corpora that demonstrate unprecedented performance across NLP tasks through scale and advanced architectures (Wan et al., 2025). Pre-training on vast datasets equips LLMs with broad world knowledge and linguistic pattern recognition, endowing them with strong reasoning abilities and cross-domain generalization capabilities (Wan et al., 2025). Models like GPT-4 perform complex reasoning steps and integrate pre-training knowledge, enabling more human-like understanding than earlier NLP approaches (Wan et al., 2025).

LLMs have been applied to causal relation extraction using frameworks like natural language inference and question-answering. Unlike traditional systems relying on explicit linguistic cues, LLMs can infer causality by evaluating contextual meaning and leveraging world

knowledge (Liu et al., 2025). In natural language inference setups, models assess whether hypotheses like "X causes Y" are entailed by given premises, achieving high accuracy in determining causal relations (Liu et al., 2025). LLMs function as automated domain experts, using broad real-world understanding to infer causality beyond superficial textual patterns (Wan et al., 2025).

Empirical studies demonstrate LLMs' effectiveness, with some methods correctly determining cause-effect direction between event pairs with up to 97% accuracy (Liu et al., 2025). These models can handle both explicit causality marked by connectives and implicit causality requiring comprehension and commonsense reasoning. LLMs effectively extract causal variables and event sequences from unstructured text by synthesizing linguistic and background knowledge (Xiong et al., 2024), significantly improving causal pair extraction over previous methods (Liu et al., 2025). Research by Ma (2025) confirms that LLMs effectively answer causal questions by leveraging embedded commonsense knowledge, uncovering relationships missed by more formulaic approaches.

3. Dataset Creation

This section describes the methodology for constructing a dedicated evaluation dataset to test large language models (LLMs) on causal information extraction under controlled, unbiased conditions. The goal was to create textual data that probes whether models can identify cause–effect relationships without relying on superficial cues or prior exposure. To achieve this, a three-step approach was followed: first generating domain-specific cause–effect pairs using an external LLM, and then embedding these pairs into contexts. Lastly, transformations were applied (masking and shuffling) to produce multiple dataset variants of increasing difficulty. The following subsections detail each stage of this dataset creation process.

Existing public datasets for textual causality were reviewed, but none met the requirements of this study. Most available resources frame causality as a classification or recognition task (often a yes/no decision or a multi-class relation label) rather than a direct extraction task. For example, the well-known SemEval-2010 Task 8 corpus (Hendrickx et al., 2010) treats cause–effect identification as a multi-way classification problem (assigning a relationship label to a given sentence) (Anuyah et al., 2025), instead of asking the model to pinpoint the cause phrase and the effect phrase in the text. Since the aim was to evaluate the extraction of causal information (i.e. identifying the exact cause and effect span in a passage), a new dataset had to be constructed from scratch.

3.1 Cause-Effect Pair Generation

The dataset construction began with the generation of 75 domain-specific cause–effect pairs (25 per domain), separated into three different domains: economics, environment and technology. These pairs (each consisting of a distinct cause and a corresponding effect) were created using the GPT-4.5 large language model through prompt engineering (see appendix I for reference). The use of an LLM for synthetic data generation was deemed a suitable solution, since it allowed the creation of custom-tailored examples in chosen domains. The prompt-engineering process produced plausible and diverse pairs, that were realistic for the domains, laying a solid foundation for the evaluation dataset. GPT-4.5 was chosen specifically because it is not among the models being evaluated in the experiments (see Section 4.1: Model selection). By using a model, which is excluded from evaluation, the process aims to avoid any unfair advantage or bias that might occur if an evaluated model were exposed to the answers in advance.

3.2 Context Generation

After obtaining the raw cause–effect pairs, the next step was to embed each pair into a surrounding textual context. For every cause–effect pair, a descriptive paragraph was generated to incorporate the cause and effect in a natural way. Each paragraph was generated to be either neutral in tone or intentionally convoluted (potentially even misleading) to not overtly signal the causal relationship.

Therefore, the prompt given to the LLM instructed it to produce a coherent paragraph or short narrative that includes both the cause and the effect, but without using explicit causal connectives (see appendix II for reference). Words and phrases such as "because", "due to", "therefore", or "as a result" were explicitly disallowed in these prompts to present the cause–effect relation in a covert or implicit manner. Where some paragraphs were straightforward and neutral (mentioning cause and effect in the same scenario), other paragraphs contained additional details or events that could act as distractors (mentioning unrelated factors or information) and mislead a superficial reading.

By avoiding obvious cue phrases and mixing the cause–effect pair with extra information, the idea was to test an LLM's ability to infer causality from meaning and context, rather than from surface patterns. Ultimately, the contexts required recognizing implicit causality, which is known to demand deeper comprehension compared to explicit causal statements (Anuyah et al., 2025). The hope for the dataset was to mimic the complexity of real-world text, providing a challenging yet fair ground for evaluation.

3.3 Dataset Alterations for Robust Evaluation

To ensure a rigorous evaluation of causal extraction capabilities, two systematic modifications, each addressing a specific vulnerability in large language model evaluation, were applied to the base dataset using automated scripts, which are also made available on Github (Sandor, 2025).This resulted in four distinct evaluation scenarios. The first alteration, cue masking, systematically replaced explicit causal indicators with neutral or ambiguous alternatives throughout the dataset texts. Although the context generation instructions aimed to minimize explicit causality markers, some causal cues inevitably remained in the generated paragraphs. Through the automated masking process, phrases like "X happened because Y occurred" were transformed to "X happened while Y occurred", eliminating direct causal cues. This modification forces models to infer causality purely from meaning and content rather than relying on surface-level linguistic signals (Anuyah et al., 2025).

The second modification, pair shuffling, addressed a more subtle evaluation challenge. Since both the cause-effect pairs and their contexts were generated by GPT-4.5, there might exist inherent associations in the model's knowledge between these elements (Li et al., 2024). By reassigning cause-effect pairs to different contexts where the terms still appear but in non-causal relationships, the shuffling process breaks these potential knowledge links. This tests whether models are genuinely extracting causal relationships from the given text or merely parroting pre-trained associations. Importantly, pairs were only shuffled into contexts that already contained the relevant terms, creating plausible but incorrect associations.

Through the combination of these two alterations, four dataset variants were created: (1) the base dataset with original cause-effect pairs in their contexts; (2) a masked version with causal cues neutralized; (3) a shuffled version with reassigned pairs to prevent knowledge leakage; and (4) a masked-shuffled version combining both modifications. This multi-faceted approach enables evaluation across increasingly challenging scenarios that systematically control for different potential shortcuts models might exploit. All dataset alterations were performed using automated scripts to ensure consistency, and all the used scripts and resulting datasets are publicly available on GitHub (Sandor, 2025).

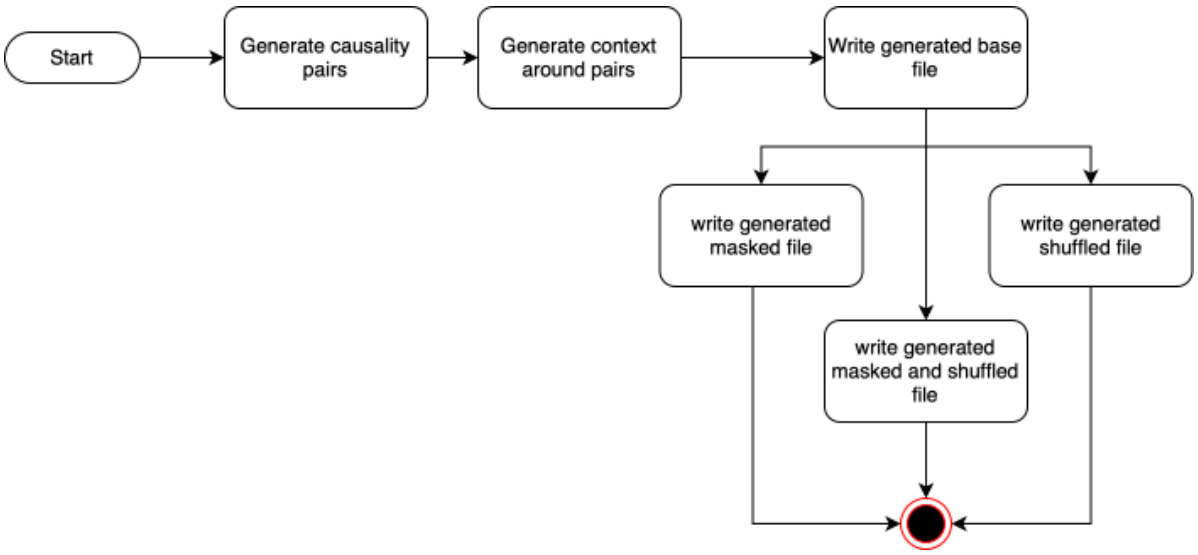


Figure 1. Dataset generation pipeline

4. Causal Information Extraction and Evaluation

4.1 Model Selection

The present study focuses on employing large language models (LLMs) for causal information extraction. To ensure robust reasoning capabilities, the GPQA benchmark (Rein et al., 2023)—a challenging graduate-level question-answering test—was used as the selection criterion. While not specifically designed for causal reasoning, GPQA performance serves as a useful proxy for general language understanding and reasoning ability.

One flagship model from each of eight major AI organizations was selected based on data from llm-stats.com (*LLM Leaderboard 2025 - Verified AI Rankings*, 2025). This community-maintained repository was chosen for its standardized testing conditions, current performance metrics, and alignment with established evaluation methodologies. To verify reliability, the GPQA scores were compared with metrics published in official model documentation and technical reports by each organization, confirming consistency where such information was available.

The selected models represent diverse approaches to LLM development across the industry. At the top end of the performance spectrum, the proprietary systems OpenAI o3 (GPQA = 83.3%), Anthropic Claude 3.7 Sonnet (84.8%), and Google Gemini 2.5 Pro (84.0%) define the current state of the art. Meta Llama 4 Maverick, the leading open-source contender, achieves 80.5%, demonstrating that transparent architectures can now approach the closed-model frontier. To ensure coverage of smaller model classes, the study further incorporates Qwen3-30B-A3B (65.8%), Mistral-Small-3-24B-Instruct (45.3%), and the mid-sized DeepSeek-R1 (71.5%). XAI Grok-3 contributes an additional high-performing reference point at 84.6%.

This diverse set encompasses various architectures and sizes, enabling comprehensive evaluation across different scales of LLMs.

4.2 Causal Information Extraction Pipeline

After selecting appropriate models for evaluation, a robust pipeline was implemented to systematically extract causal information from text using multiple large language models. This section describes the technical implementation of the constructed pipeline, developed to standardize interactions with heterogeneous LLM APIs, as well as ensuring methodological

consistency between evaluation conditions.The architecture of the causal extraction pipeline is illustrated in Figure 2, a simplified flow chart of the system's operation. A more technical flow chart with additional details can be found in the GitHub repository (Sandor, 2025). The pipeline is implemented as a modular Python framework, and the complete source code is publicly available on GitHub (Sandor, 2025).

This approach addresses several key technical challenges in cross-model extraction. Firstly, it standardizes interactions across APIs with vastly different implementation patterns, from Google's Vertex AI to Anthropic's Messages API. Secondly, it implements consistent prompt engineering across all models. Thirdly, it implements robust error handling and rate limitation compliance to prevent API throttling on large-scale extraction tasks. Finally, it provides structured output formats optimized for quantitative evaluation.

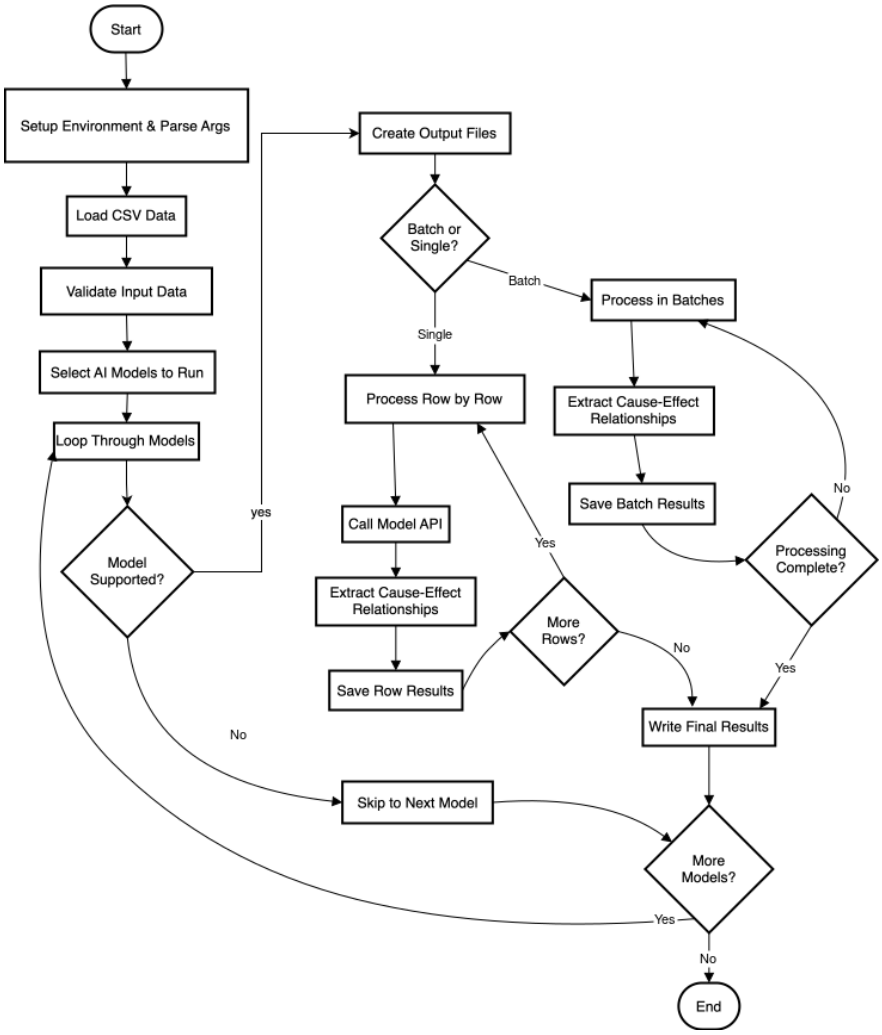


Figure 2. Simplified pipeline for causal-information extraction

Unified Prompt Design for Zero-Shot Extraction. Cross-model comparison validation required crafting a template for a standardized prompt. Importantly, the study utilizes a zero-shot approach, providing no demonstrations or examples of correctly extracted causes and effects. This methodological choice tests each model's ability to understand causal relationships without task-specific training (Liakhovets & Schlarb, 2022).

Upon numerous iterations of adjustments to maximize consistency, the final prompt composition used several essential elements (see Figure 3. Zero-shot causal information extraction prompt).

```
You are a causal-extraction specialist. Identify the cause and effect
in the text.

Output ONLY a JSON object with these fields:

- "cause":   the exact cause phrase from the text

- "effect":  the exact effect phrase from the text

Your response must ONLY be the JSON—no additional text or explanation.

If no clear cause-effect relation exists, return {"cause": "",
"effect": ""}.

<paragraph>

{context}

</paragraph>
```

Figure 3. Zero-shot causal information extraction prompt

For models supporting system prompts, an additional parameter was introduced ("Return ONLY a JSON object with cause and effect fields"), in order to ensure consistent guidance for all the APIs while maintaining the primary prompt focused on the extraction task.

The specificity of demanding the extraction of "exact phrase" was a deliberate design choice intended to evaluate models' ability to identify causal spans and excluding paraphrasing or inferring unmentioned causal relationships.

4.2.1 Model-Specific Implementation

Each of the LLMs required custom integration due to significant variations in API implementations, authentication requirements, and response formats. The pipeline employs a provider-based architecture that masks these differences behind a unified interface while handling model-specific optimizations.

Default Configurations and Temperature Settings. The default, out-of-the-box configurations were used for all models to maximize reproducibility and facilitate fair comparison. Temperature settings, which control the randomness in model outputs, were standardized across models to minimize variability in responses to identical inputs (Peeperkorn et al., 2024). For all models except OpenAI's o3, the temperature parameter was set to 0, which produces the most deterministic outputs by selecting the highest probability tokens at each step of generation.

Preliminary testing with OpenAI's O3 revealed that a temperature setting of 0 caused technical issues with the model's output, therefore a temperature value of 1 was used exclusively for this model.

Response Processing and Extraction. In practice, as inconsistencies were encountered in LLM response formatting despite explicit JSON output instructions, the final solution uses a multi-stage extraction process with progressive fallback mechanisms to maximize data recovery.

The pipeline first attempts standard JSON parsing of well-formed responses. When this is not feasible due to incorrect JSON (a common occurrence despite explicit instructions), the system resorts to regular expression-based pattern matching tailored to common error patterns found during development. The extraction logic also supports model-specific nuances such as quotation normalization, reasoning/thinking segments bounded by tags, and partial JSON constructs. Implemented robust extraction approach achieved high reliability across diverse model outputs, with particular emphasis on edge cases that would otherwise result in data loss.

Sequential Processing Implementation. While the pipeline includes asynchronous batch processing capabilities for APIs that support it (primarily OpenAI), the final experimental evaluation exclusively used sequential processing with individual requests for all models. This

methodological choice ensured consistency across all evaluated systems, prioritizing experimental rigor over time and cost.

The pipeline generates complementary output formats optimized for different analytical purposes. For each processed text, the system creates both a structured JSONL record containing the row identifier, extracted cause-effect predictions and the raw API response, as well as a tabular CSV representation with appropriate content for reviewing the results with standard analysis tools.

To ensure resilience against potential problems during extended extraction runs, the system implements progressive file writing, saving results approximately every 10 processed items. This approach proved valuable during development, allowing recovery from occasional API interruptions.

4.3 Evaluation Pipeline

Following the extraction of causal relationships, an evaluation framework was implemented to gain insight into the extraction quality. This section describes the evaluation pipeline that combines LLM-based automated assessment with human evaluation to create a benchmarking system.

4.3.1 LLM-as-Judge Evaluation Approach

Traditional evaluation metrics for information extraction often rely on exact string matching, failing to capture semantic equivalence when different phrasings express the same relationship. To address this limitation, the evaluation framework employs a methodology, where large language models assess extraction quality using a three-point scale:

- 1.0: Exact match or semantically identical meaning
- 0.5: Partial match (core concept similar but missing details)
- 0.0: No match or incorrect extraction

This approach enables recognition of partially correct extractions that would be unfairly penalized by binary metrics, providing more nuanced performance assessment.

The evaluation prompt presents structured information about ground truth and predicted pairs, with explicit scoring instructions emphasizing semantic evaluation rather than literal matches. Each judge model receives identical prompts with instructions to return scores in JSON format (see Figure 4. LLM-judge prompt for evaluation).

```
"""Task: Evaluate how well the predicted cause-effect pair matches the
ground truth pair.

id: {id}

predicted_by: {source_model}

GROUND TRUTH PAIR:

Cause: {orig_cause}

Effect: {orig_effect}

PREDICTED PAIR:

Cause: {pred_cause}

Effect: {pred_effect}

EVALUATION INSTRUCTIONS:

1. Compare the predicted cause with the ground truth cause

2. Compare the predicted effect with the ground truth effect

3. Assign a score to each:

  - 1.0: Exact match or semantically identical meaning

  - 0.5: Partial match (core concept is similar but missing important
details)

  - 0.0: No match or incorrect

You MUST respond with ONLY a valid JSON object in this exact format:

{{"cause_score": 0|0.5|1, "effect_score": 0|0.5|1}}

Your response MUST ONLY contain this JSON. No other text or explanation.

"""
```

Figure 4. LLM-judge prompt for evaluation

As shown in Figure 5. Automated grading pipeline, the evaluation pipeline reuses the same panel of judge models employed in the extraction stage; all models run with a temperature of 0 to maximise determinism, except GPT-o3, which retains its default sampling temperature. The open-source code—available on GitHub (Sandor, 2025)—implements adaptive rate-

limiting and exponential-backoff retries to absorb API failures, ensuring robust evaluation over large datasets.

For each evaluated extraction model, the pipeline generates outputs in both JSONL and CSV formats, mirroring the output structure of the extraction pipeline for consistency. These outputs include per-sample scores for both cause and effect components, combined average scores, original and predicted pairs, and raw judge responses for verification.

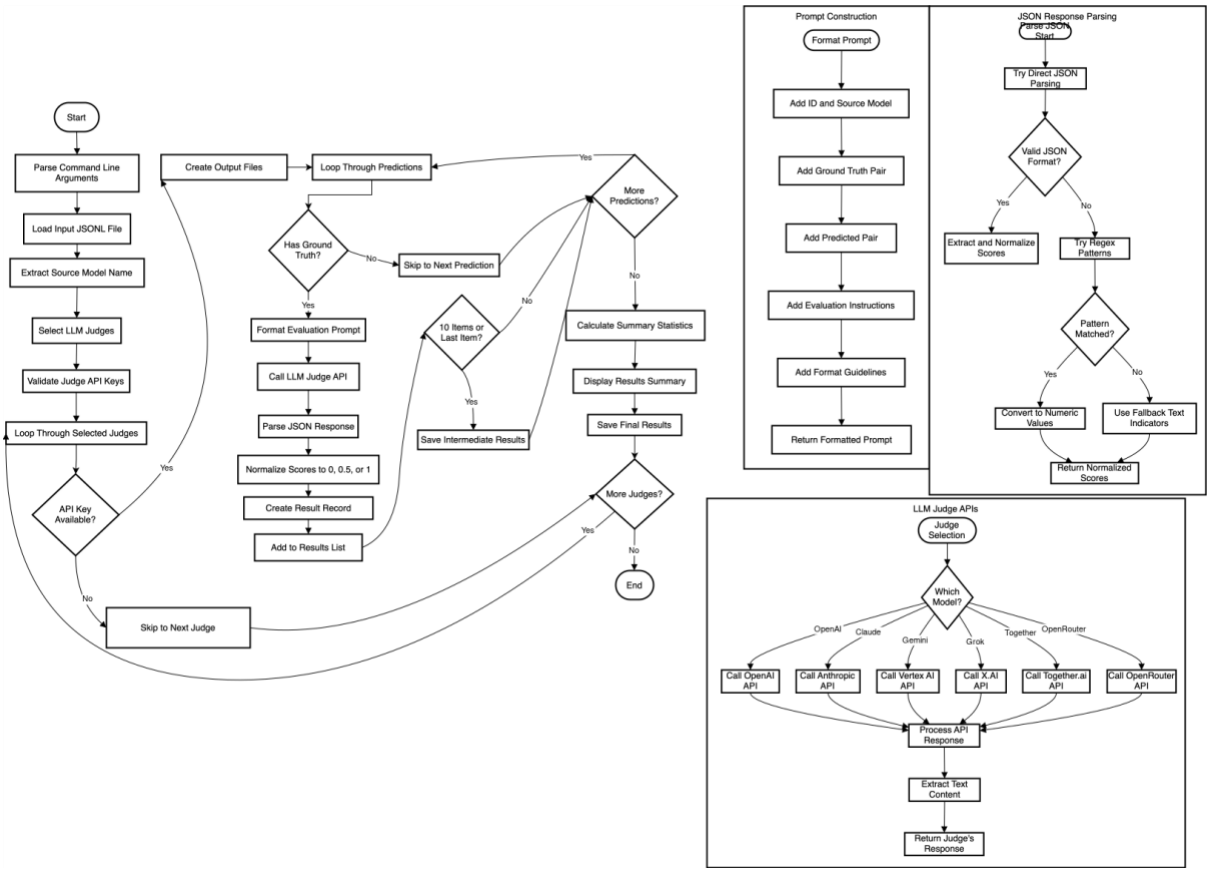


Figure 5. Automated grading pipeline

4.3.2 Human Evaluation Component

In addition to automated LLM-based evaluation, a human evaluation component was implemented to establish gold-standard benchmarks and identify true positives for subsequent analysis (Sandor, 2025). This process involved manual assessment of all LLM extractions using the same three-point scoring scale applied by the automated judges.

The human evaluation process revealed several methodological challenges. First, consistent application of scoring criteria proved difficult even for a single evaluator, particularly for borderline cases where extraction quality fell between defined score points. Second, the evaluation process was time-intensive, limiting the quality of human assessment. Third, distinguishing between semantically equivalent phrasings and substantively different extractions required domain knowledge that varied across dataset topics.

Similar to the extraction pipeline described in Section 4.2, a significant implementation focus was on reliable parsing of LLM outputs, this time focusing on score extraction from judge responses. The evaluation pipeline implements a multi-stage parsing approach with fallback mechanisms: beginning with standard JSON parsing, then applying regex-based pattern matching for common variations, and finally normalizing scores to the standardized scale. This robust extraction logic echoes the techniques developed for the extraction pipeline.

5. Results

This section presents the empirical findings from evaluating chosen large language models on causal extraction tasks across multiple dataset variants. The analysis examines performance patterns, model resilience to increasingly challenging conditions, and extraction characteristics as measured by traditional metrics.

5.1 Model Performance Overview

Significant variations in causal extraction capabilities were observed across the evaluated models. Figure 6 illustrates how models performed on the fundamental task of extracting causes versus effects across all datasets.

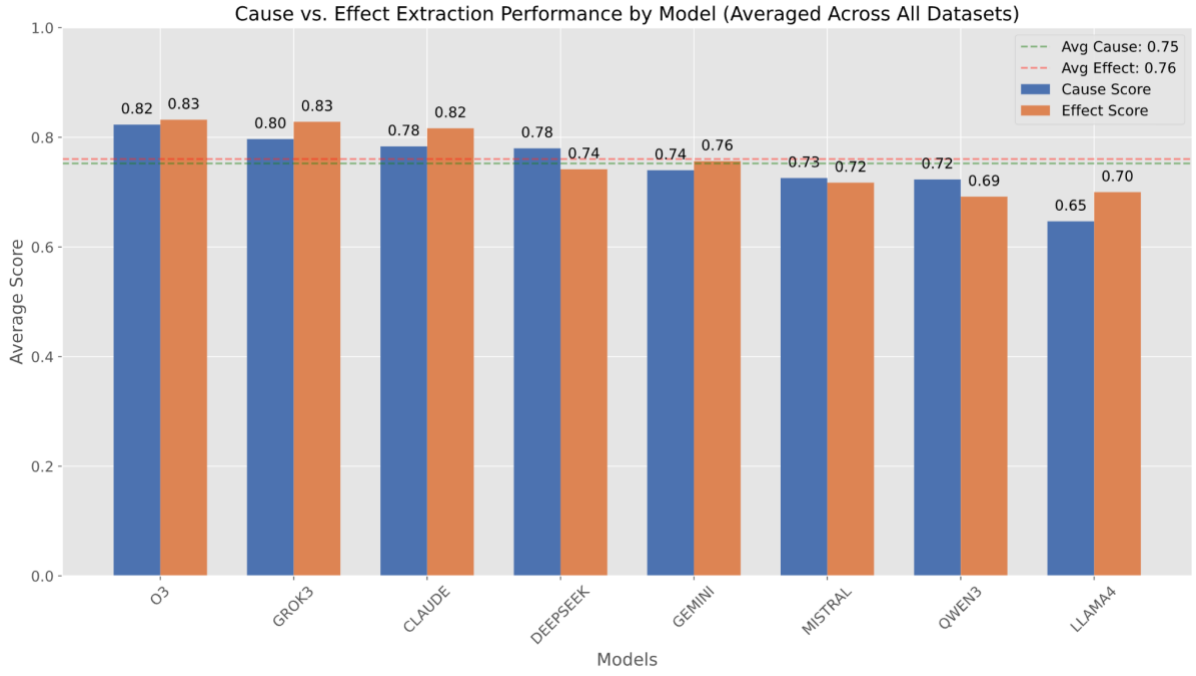


Figure 6. Cause vs. effect extraction performance across all datasets

The comparative analysis revealed a consistent pattern. Effect extraction (average score 0.76) slightly outperformed cause extraction (average score 0.75) across the systems tested. This pattern was not uniform, however. Several models, including Deepseek, Mistral and Qwen 3, actually showed stronger performance on cause extraction than effect extraction. Llama 4 demonstrated the largest disparity in the opposite direction, with cause extraction scoring 0.65 compared to effect extraction at 0.70. GPT-o3 demonstrated both superior overall performance and balanced extraction capabilities, scoring 0.82 for causes and 0.83 for effects.

The varying balance between cause and effect extraction across models indicates different underlying approaches to causal reasoning. The inconsistent pattern suggests that architectural decisions and training methodologies may influence whether a model develops stronger capabilities in identifying causes or effects.

5.2 Resilience Across Dataset Variants

The multi-variant dataset design enabled systematic evaluation of model resilience to increasingly challenging extraction conditions. Figure 7 presents performance across all four dataset variants: base, masked, shuffled, and masked-shuffled.

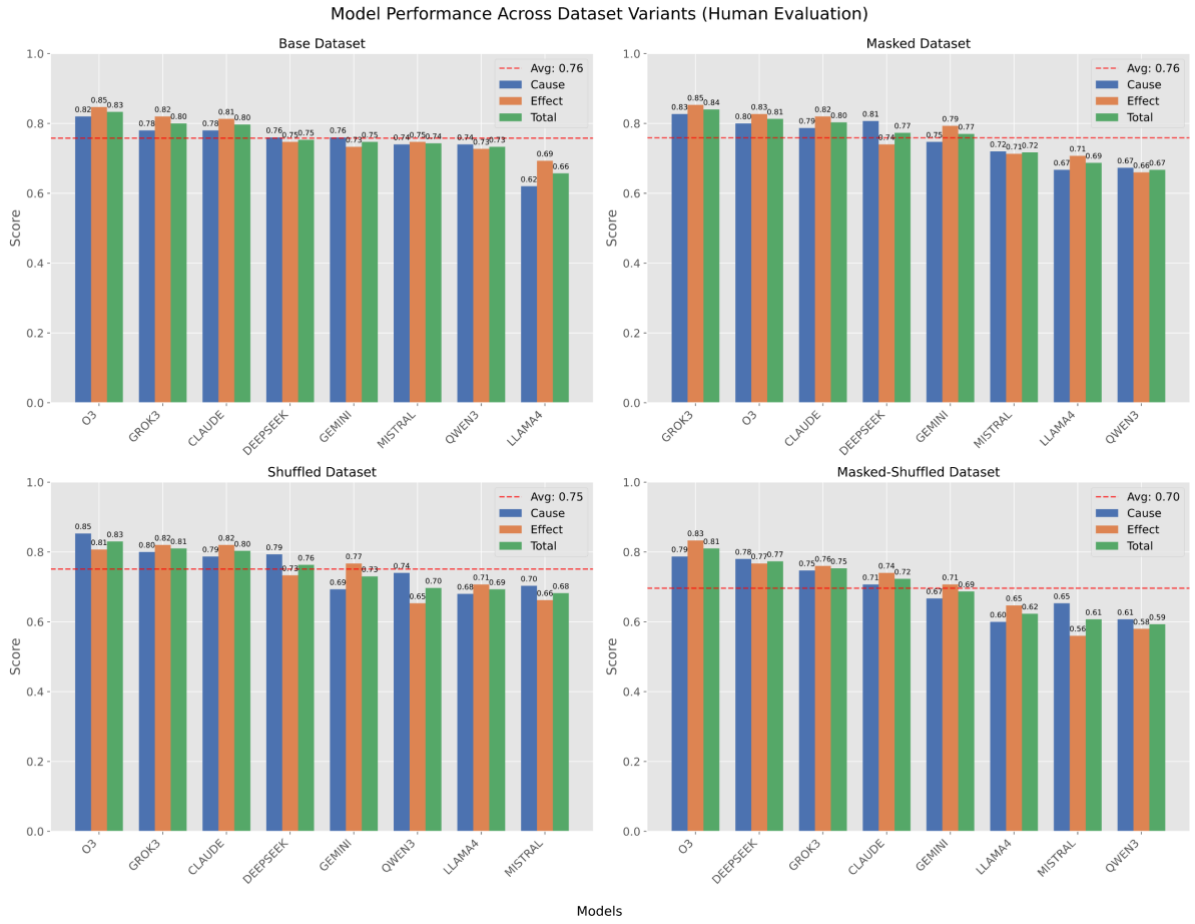


Figure 7. Model performance by dataset variant

Base dataset performance established the benchmark capability level, with an average total score of 0.76 across all models. The introduction of cue masking (removing explicit causal indicators) produced a minimal performance change, with the average score remaining at 0.76.

This could imply that contemporary LLMs do not primarily rely on explicit linguistic markers when extracting causal relationships.

The shuffled variant, which reassigned cause-effect pairs to create plausible but incorrect associations, produced a slight performance decrease to 0.75, suggesting once again minimal reliance on pre-existing term associations. The most challenging masked-shuffled variant resulted in an average performance of 0.70, representing an 8% reduction from the base condition.

Model rankings remained relatively consistent across variants, with o3 and Grok 3 maintaining top positions throughout all conditions. Notably, performance disparities between leading and trailing models widened in more challenging scenarios. GPT o3 demonstrated particular resilience in the masked-shuffled variant (0.81 score), while Qwen 3 (0.60) and Mistral (0.62) showed more substantial degradation.

This differential response to dataset modifications provides insight into extraction strategies. The minimal impact of cue masking suggests that current LLMs employ genuine semantic understanding rather than pattern matching based on causal cue words. Conversely, the more pronounced effect of pair shuffling some reliance on pre-trained term associations, particularly in models with lower overall performance.

5.3 Precision, Recall and F1 Analysis

Beyond semantic evaluation scores, traditional metrics offer additional insights into extraction characteristics. A specialized methodology converted the three-point scoring system (0, 0.5, 1) into binary classification for precision, recall, and F1 calculation, with carefully defined criteria for True/False Positives and Negatives.

The conversion methodology classified True Positives as cases where the model extracted causality, human evaluation was ≥0.5, and a majority of LLM judges assigned scores ≥0.5. False Positives occurred when either human evaluation or LLM judges scored <0.5 despite the model extracting causality. True Negatives represented cases where no causality was present and the model correctly made no extraction. False Negatives were instances where causality was present (human evaluation >0) but the model failed to extract it. Figure 8 presents these metrics across all models and dataset variants.

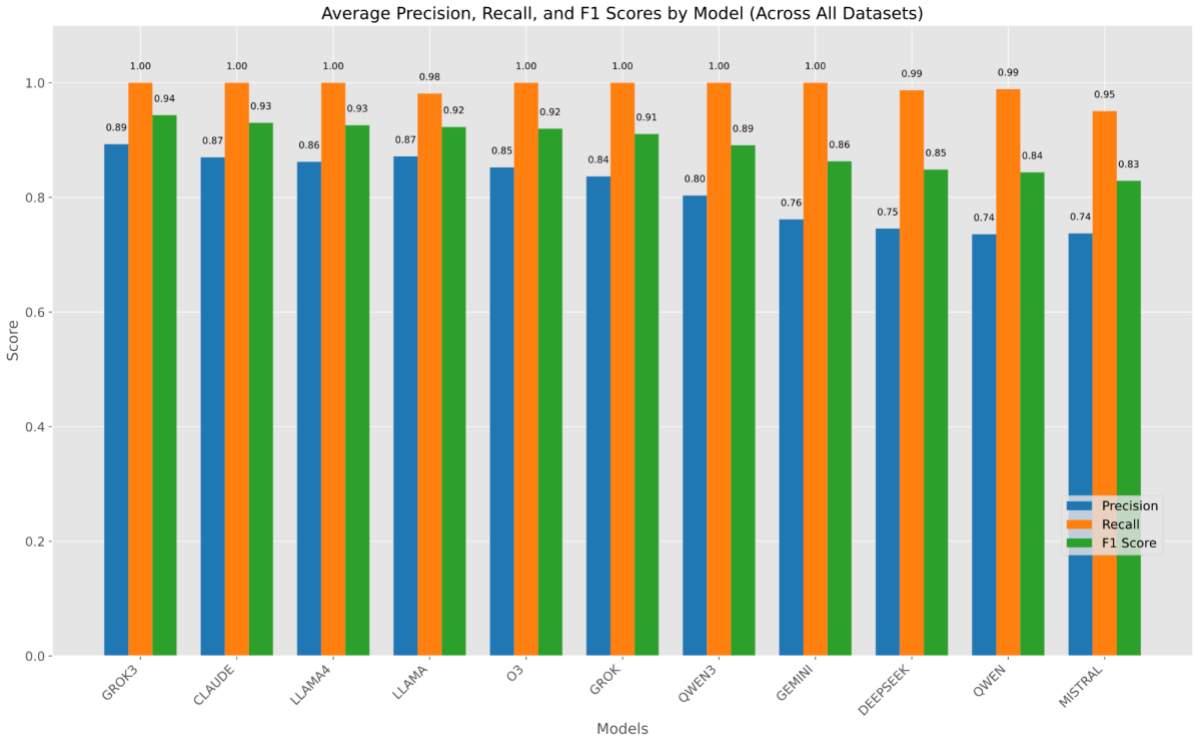


Figure 8. Average precision, recall and F1 by model

Possibly as a direct consequence of the conversion methodology being skewed towards positive evaluation when meeting minimum criteria, the recall scores elevated between 0.95 and 1.00 for all models. Precision scores demonstrated greater variability, ranging from 0.89 (Grok3) to 0.74 (Mistral and Qwen 3).

The F1 scores ranged from 0.94 (GROK3) to 0.83 (MISTRAL), with the high values across models reflecting the substantial influence of methodology-driven high recall even when precision varied. Across dataset variants, precision declined more substantially than recall in the most challenging conditions, suggesting models tend toward false positive errors rather than false negative errors as task difficulty increases.

6. Discussion and Future Work

Results indicate strong capabilities in causal extraction across LLMs (0.76 average score), with significant performance disparities between leading models (O3, Claude, Grok 3) and others (Qwen 3, Mistral). This suggests architectural design, training methodology, and scale remain influential factors in causal reasoning.

The slight advantage in effect extraction (0.76) over cause extraction (0.75) aligns with research indicating effects are typically more explicitly stated than causes (Feder et al., 2022). However, inconsistent manifestation across models, with some showing stronger cause extraction (Deepseek, Mistral, Qwen 3), suggests architectural differences affect causal directionality processing.

The generated dataset captures only a subset of causal relationship types, providing a constrained view that may not fully generalize. The dataset transformation approach revealed performance patterns but introduced artifacts. Masking explicit cause-effect cues sometimes produced linguistically awkward constructions, while shuffled variants created implausible scenarios. These modifications successfully tested reliance on explicit cues and pre-trained associations, though at the cost of real world scenarios.

The dual evaluation system provided complementary perspectives but revealed methodological complications. The three-tier scoring system (0, 0.5, 1) recognized partial extractions fairly, yet created challenges when converting to traditional metrics for literature comparison.

Conversion to binary classification produced unrealistically high recall scores (0.95-1.00), suggesting either overly lenient criteria or a need for framework refinement. Additionally, LLM judges evaluating without context access were limited to term matching, contrasting with human evaluations that incorporated contextual understanding—potentially explaining systematic differences between human- and LLM-assessments.

Using GPQA as a proxy for general reasoning proved largely successful (see appendix III for reference). Models with higher GPQA scores demonstrated superior causal extraction, suggesting correlation between general knowledge reasoning and causal abilities. This implies improvements in foundational capabilities translate to enhanced performance in specialized reasoning tasks without requiring specialized training.

The extraction pipeline successfully standardized cross-API interactions, though design decisions influenced results. Zero-shot evaluation assesses innate extraction-abilities without examples but represents only one paradigm. The minimalist prompt approach selected for compatibility may have favored certain interface patterns.

Models showed varying compliance with output format instructions — some incorporating extensive reasoning before answers, others following JSON-only format precisely. This raises questions about whether "loud chain-thinking" provided advantages or whether strict instruction-following demonstrated superior capabilities.

6.1 Limitations and Future Directions

Key limitations include dataset size and diversity, zero-shot evaluation constraints, and evaluation framework metrics. Future research should develop larger datasets spanning more domains and linguistic constructions, investigate few-shot learning benefits, and refine evaluation metrics through standardized benchmarks with expert-verified annotations.

Additional avenues include studying how prompt formulations affect extraction quality and exploring relationships between causal extraction and other reasoning capabilities to better understand causal cognition in LLMs.

7. Conclusion

This thesis has investigated the causal information extraction capabilities of eight flagship large language models across diverse AI organizations. Results demonstrate that modern LLMs possess substantial ability to identify cause-effect relationships, achieving an average performance score of 0.76 across all models and test conditions.

The most significant finding is that contemporary LLMs do not primarily rely on explicit linguistic markers when extracting causal relationships, as evidenced by the minimal impact of cue masking on performance. This suggests these models have developed semantic understanding capabilities that transcend simple pattern matching. The more pronounced effect of pair shuffling could indicate some reliance on pre-trained term associations, particularly in models with lower overall performance.

Notable performance disparities between leading models (GPT-o3, Claude 3.7 Sonnet, Grok 3) and others (Qwen-3, Mistral) confirm that architectural design, training methodology, and scale influence causal reasoning abilities. The effect extraction (0.76) slightly outperformed cause extraction (0.75), though this pattern varied across models, suggesting differences in how causal directionality is processed.

The methodological contributions — multi-variant dataset design, zero-shot extraction pipeline, and dual human-LLM evaluation framework — provide valuable tools for future causal extraction research. While challenges remain in dataset construction, evaluation metrics, and extraction methodologies, this research provides tools and insights for evaluating and comparing LLM performance on causal information extraction tasks.

References

Ali, W., Zuo, W., Ying, W., Ali, R., Rahman, G., & Ullah, I. (2023). Causality extraction: A comprehensive survey and new perspective. *Journal of King Saud University - Computer and Information Sciences*, *35*(7), 101593. https://doi.org/10.1016/j.jksuci.2023.101593

Anuyah, S., Vanschaik, J., Jain, P., Lehman, S., & Chakraborty, S. (2025). *An Empirical Study of Causal Relation Extraction Transfer: Design and Data* (No. arXiv:2503.06076). arXiv. https://doi.org/10.48550/arXiv.2503.06076

Asghar, N. (2016). Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey. *ArXiv*. https://www.semanticscholar.org/paper/Automatic-Extraction-of-Causal-Relations-from-A-Asghar/e115ec4c138cc5157ab24a6e462f1fc74902d53f

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V., & Yang, D. (2022). Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, *10*, 1138–1158. https://doi.org/10.1162/tacl_a_00511

Girju, R., & Moldovan, D. (2002). *Text Mining for Causal Relations*.

Hendrickx, I., Kim, S., Kozareva, Z., Nakov, P., Padó, S., Pennacchiotti, M., Romano, L., & Szpakowicz, S. (2010). *SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals*. 33–38.

Li, Y., Guo, Y., Guerin, F., & Lin, C. (2024). An Open-Source Data Contamination Report for Large Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 528–541. https://doi.org/10.18653/v1/2024.findings-emnlp.30

Liakhovets, D., & Schlarb, S. (2022). *Zero-shot Event Causality Identification with Question Answering*. CLIB. https://www.semanticscholar.org/paper/Zero-shot-Event-Causality-Identification-with-Liakhovets-Schlarb/cf215070d2383b33bdba170744f9d58492b7f782

Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H., Yu, T., McAuley, J., Ai, W., & Huang, F. (2025). *Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey* (No. arXiv:2403.09606). arXiv. https://doi.org/10.48550/arXiv.2403.09606

LLM Leaderboard 2025—Verified AI Rankings. (2025, May 12). Archive.Ph. https://archive.ph/lylpD

Ma, J. (2025). *Causal Inference with Large Language Model: A Survey* (No. arXiv:2409.09822). arXiv. https://doi.org/10.48550/arXiv.2409.09822

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). *Is Temperature the Creativity Parameter of Large Language Models?* https://doi.org/10.48550/ARXIV.2405.00492

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (No. arXiv:2311.12022). arXiv. https://doi.org/10.48550/arXiv.2311.12022

Sandor, V. (2025). *CIE* [Computer software]. https://github.com/koodikirjutaja/CIE

Sorgente, A., Vettigli, G., & Mele, F. (2018). *A Hybrid Approach for the Automatic Extraction of Causal Relations from Text* (C. Lai, A. Giuliani, & G. Semeraro, Eds.; Vol. 746, pp. 15–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-68392-8_2

Wan, G., Lu, Y., Wu, Y., Hu, M., & Li, S. (2025). *Large Language Models for Causal Discovery: Current Landscape and Future Directions* (No. arXiv:2402.11068). arXiv. https://doi.org/10.48550/arXiv.2402.11068

Xiong, S., Chen, D., Wu, Q., Yu, L., Liu, Q., Li, D., Chen, Z., Liu, X., & Pan, L. (2024, October 22). *CausalEval: Towards Better Causal Reasoning in Language Models*. https://www.semanticscholar.org/paper/CausalEval%3A-Towards-Better-Causal-Reasoning-in-Xiong-Chen/ae208e3cc77feafd90547de0a4691f2ef92af5ae

Yang, J., Han, S. C., & Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, *64*(5), 1161–1186. https://doi.org/10.1007/s10115-022-01665-w

Appendices

I. Cause-effect pair generation prompt

Generate a structured and academically rigorous set of 75 distinct cause-effect pairs explicitly formatted as "Cause: [specific cause]; Effect: [specific effect]". The pairs should be evenly distributed across these three clearly labeled thematic categories:

Economics & Finance (25 pairs)
Environment & Climate (25 pairs)
Technology & Computer Science (25 pairs)
Detailed Requirements:
Each cause-effect pair must represent realistic, precise, and domain-specific causal relationships suitable for inclusion in a dataset for causal inference research in a bachelor's thesis in computer science.
Prioritize specificity, clarity, and plausibility; avoid compound or chained causal statements.
Clearly state the causal relationship without any intermediate steps or additional causes.
Each pair should be fully self-contained, understandable without supplementary context, and avoid overly general or common textbook examples.
Ensure thematic diversity within each category, avoiding repetition or overlap of closely related causal relationships.
Include causal relationships at multiple scales (individual, organizational, systemic), clearly indicating the involved scale within each cause-effect pair when applicable.
Emphasize contemporary relevance by focusing primarily on developments, trends, or phenomena occurring within the past five years.
Example Format and Expectations:
Economics & Finance:
Cause: Government increases corporate tax rates; Effect: Business investment declines.
Environment & Climate:
Cause: Melting of polar ice caps; Effect: Rise in global sea levels.
Technology & Computer Science:
Cause: Implementation of stronger encryption standards; Effect: Increased data security.
Adhere strictly to this refined format and ensure a balanced, detailed, and insightful representation across all categories.

II. Cause-effect context generation prompt

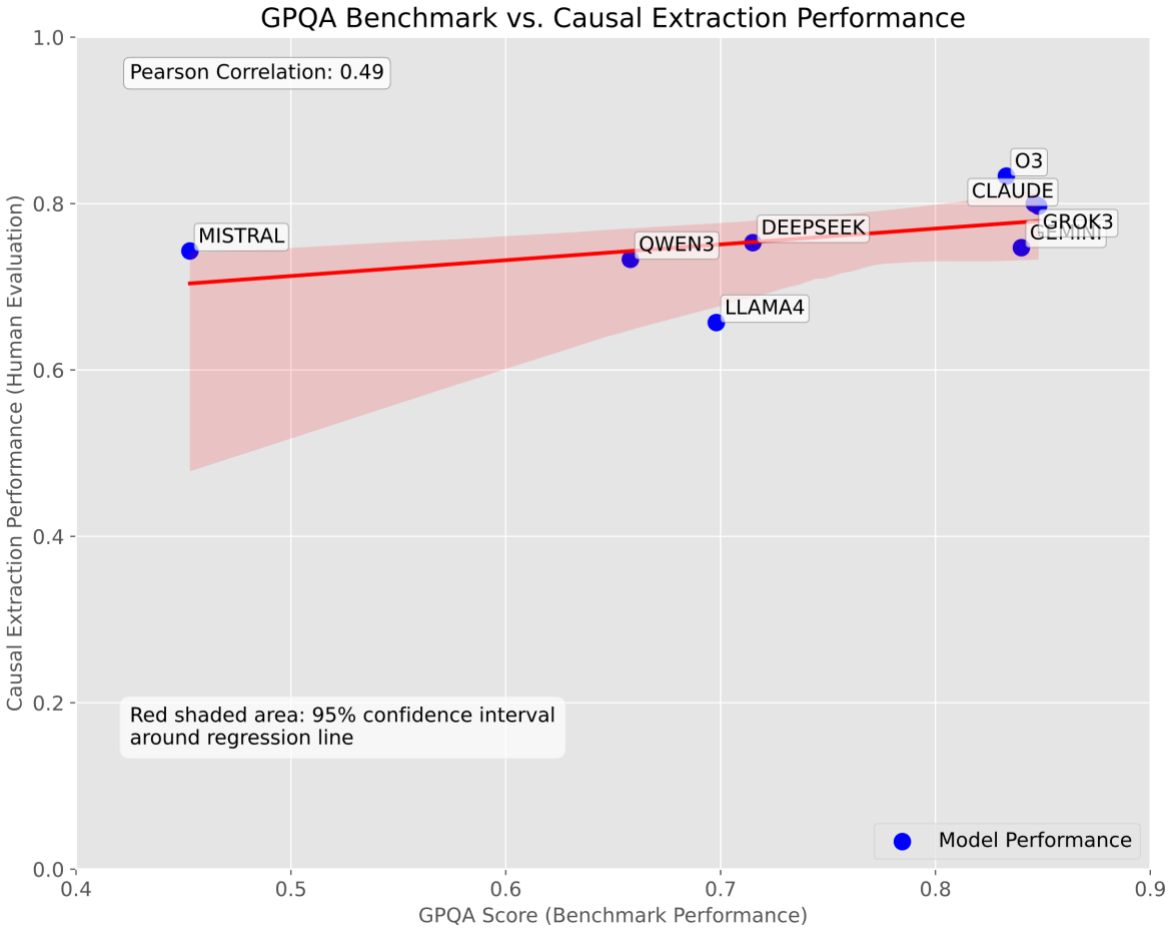
Given a text (.txt) file containing distinct cause-effect pairs formatted as "Cause: [specific cause]; Effect: [specific effect]", generate a CSV file with the following columns:

- "Cause": original cause statement
- "Effect": original effect statement
- "Context": A concise paragraph (1-3 sentences) subtly embedding the cause-effect pair, intentionally crafted to mislead causal extraction algorithms. Avoid explicit causal indicators such as "because," "due to," "as a result," or similar phrases. Introduce plausible, contextually relevant distracting or ambiguous details to obscure the direct causal relationship.

Requirements:

- Each context paragraph should remain logically coherent and readable despite intentional ambiguity.
- Domain-specific language must align appropriately with the topic implied by the cause-effect

III. Corellation between GPQA score and found results



License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Sandor Vunk ,
(author´s name)

1. grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

   Causal Information Extraction Using Large Language Models ,
   (title of thesis)

   supervised by Giacomo Magnifico ;
   (supervisor´s name)

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Sandor Vunk

15/05/2025