🌻 Research on the ability of LLMs to detect causal claims

11 Dec 2025

Summary

This note explains why modern Large Language Models (LLMs) — especially since instruction-tuned chat models — often seem to have a "native" ordinary-language grasp of causation: they can spot, generate, and elaborate "A influences B" talk even when it is informal, implicit, or socially framed.

The core claim is **hybrid**:

- **Pre-training (scale) supplies the raw capacity**: these models learn by reading huge amounts of text and trying to predict the next word. Because people constantly write about reasons, consequences, and mechanisms, doing well at prediction pushes the model to absorb many common causal patterns (e.g., "fell" → "broke").
- **"Chat training" makes that capacity usable in conversation**: newer models are additionally trained on lots of question-and-answer style instructions ("why did this happen?", "what led to that?"), and then refined using human ratings that reward answers that are clear and coherent (often called RLHF). The result is not just better *spotting* of causal language, but better *explaining* and *rephrasing* causal claims on demand.

Historically, this is framed as a transition from **causality-as-extraction** to **causality-as-generation**:

- **2015-2019 (pre-generative)**: "causal understanding" was largely operationalised as Causal Relation Extraction (CRE) — classifying or tagging explicit cause/effect relations in sentences or documents. Methods were feature-heavy (SVMs), then neural (RNN/CNN), and later rule "sieves" (e.g., CATENA). These systems were brittle for **implicit causality** (where the link is inferred rather than marked by words like "caused").
- **2019-2021 (contextual encoders)**: BERT-era models improved extraction by using deep context, but remained primarily *discriminative* (understand/tag) rather than *generative* (explain/construct narratives).

- **2022-2025 (instruction-following generators)**: chat-oriented models excel at producing causal explanations, counterfactual-style answers, and “influence” talk because their training regimes contain huge amounts of “why/how/what caused what” interactions, plus preference tuning for coherent explanation.

To ground “ordinary language causation,” this note uses two cognitive-linguistic lenses that match what chat models often do well:

- **Force Dynamics (Talmy)**: causal meaning as roles/forces (agonist vs antagonist, enabling vs preventing), which connects to how models handle “let,” “despite,” “kept,” etc.
- **Implicit Causality (IC) verbs**: verbs carry pragmatic biases about who is responsible (NP1 vs NP2 bias), and LLM continuations often match human patterns in these explanation contexts.

This note is also explicit about limits and failure modes that matter for “detecting causal claims” in text:

- **Causal fluency ≠ causal truth**: RLHF can encourage plausible-sounding causal stories even when the premise is wrong (“hallucinated causality”) or when a correlation is being misread as a cause.
- **Temporal — causal confusion**: models are biased to treat temporal sequence as causal sequence (post hoc fallacy), because narratives in training data often align “then” with “therefore.”
- **Surface competence vs intervention-level reasoning**: performance can drop on “fresh” causal probes; this supports the view that some behaviour is pattern-based association rather than robust interventionist reasoning.

Current frontiers (2024-2025) are framed as attempts to make causal reasoning more checkable and structured, including addressing the above shortcomings: “reasoning models” that perform intermediate causal checks (often hidden) and pipelines where LLMs **extract** candidate causal edges into explicit graphs (DAG-like representations) for downstream formal analysis.

1. Introduction: The Emergence of "Native" Causal Fluency

The capacity of Large Language Models (LLMs) to identify, generate, and reason about causal relationships in ordinary language is a notable (and still debated) development in artificial intelligence over the last decade. Since the release of ChatGPT (based on GPT-3.5) and its successors, these systems have often appeared able to process prompts involving influence, consequence, and mechanism without extensive few-shot examples or rigid schema engineering

that characterised previous generations of Natural Language Processing (NLP). This report investigates the trajectory of this capability from 2015 to 2025, asking how much is a by-product of scale versus the result of specific (often implicit) training choices.

Furthermore, the report explores the philosophical and linguistic dimensions of this capability, using frameworks such as Leonard Talmy's Force Dynamics and the theory of Implicit Causality (IC) verbs to benchmark LLM performance against human cognitive patterns. The evidence suggests that while LLMs can often handle the *linguistic interface* of causality — the "language game" of cause and effect — significant questions remain regarding the grounding of these symbols in a genuine world model.

---

2. The Pre-Generative Landscape (2015-2019): Causality as Extraction

To appreciate the "native" fluency of 2025-era models, one must first analyse the fragmented and rigid methodologies that dominated the field between 2015 and 2019. During this period, the "ordinary language concept of causation" was operationalised not as a generative understanding, but as a classification task known as Causal Relation Extraction (CRE).

2.1 The Legacy of SemEval-2010 Task 8

For much of the decade, the benchmark defining the field was **SemEval-2010 Task 8**, which framed causality as a relationship between two nominals marked by specific directionality. Systems were tasked with identifying whether a sentence like "The fire was triggered by the spark" contained a *Cause-Effect(e2, e1)* relationship.

Research from this era was characterised by a heavy reliance on feature engineering and pipeline architectures. Early approaches used Support Vector Machines (SVMs) and later, Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). These models did not "understand" causality in any holistic sense; rather, they learned to detect explicit lexical triggers — words like "caused," "led to," or "resulted in."

The limitation of this paradigm was its inability to handle **implicit causality** — relationships where the causal link is inferred from world knowledge rather than stated explicitly. For instance, in the sentence "The rain stopped; the sun came out," a human reader infers a temporal and potentially causal sequence. Pre-transformer models, lacking a comprehensive probabilistic model of how events co-occur in the world, consistently failed to identify such links, achieving F1 scores that rarely exceeded 0.60 on implicit datasets. This era treated causality as a syntactic puzzle rather than a semantic reality.

2.2 The Shift to Event-Centric Resources: EventStoryLine and Causal-TimeBank

Between 2015 and 2018, the research community began to move beyond sentence-level extraction toward document-level understanding, driven by the creation of corpora like the **EventStoryLine Corpus** and **Causal-TimeBank**.

- **EventStoryLine:** This corpus was designed to evaluate "StoryLine Extraction," requiring systems to connect disparate event mentions (e.g., "shooting," "hospitalisation," "death") into a coherent narrative structure. The annotation scheme introduced specific classes like `ACTION_CAUSATIVE` to distinguish events that initiate change from those that merely describe states.
- **Causal-TimeBank:** Research using this corpus highlighted the inextricable link between *temporality* and *causality*. The "Causal-TempBank" approach demonstrated that knowing "A happened before B" significantly improved the classification of "A caused B".

Despite these richer datasets, the methods remained fundamentally discriminative. Systems like **CATENA** (2016) used "sieves" — rule-based filters — to extract causal links. These systems could identify likely causal passages, but they did so through rigid, handcrafted logic rather than conversational explanation. They could not generate an explanation or reason about counterfactuals; they could only point to where a human annotator might say a cause existed.

2.3 The BERT Revolution and Contextual Embeddings

The release of **BERT (Bidirectional Encoder Representations from Transformers)** in 2018 marked a pivotal transition. BERT introduced deep contextual embeddings, allowing models to distinguish the semantic nuance of causal words based on their surrounding text.

Comparative studies from this period show a dramatic jump in performance. Fine-tuned BERT models (such as BioBERT) achieved F1-scores of approximately 0.72 on medical causality tasks, significantly outperforming previous architectures. BERT represents a "careful reader" — a model that can attend to the entire sentence simultaneously to resolve ambiguities.

However, BERT was still an encoder-only architecture. It was designed to *understand* (classify/tag), not to *speak*. While it could identify causal passages with greater accuracy than ever before, it lacked the autoregressive capability to generate causal narratives. The "native ordinary language concept" requires not just recognition, but the ability to formulate causal thoughts — a capability that would only emerge with the Generative Pre-trained Transformer (GPT) series.

3. The Generative Era (2020-2025): Structural Induction of Causal Logic

The observation that models "since around ChatGPT 3.5" (released late 2022) exhibit a distinct causal proficiency aligns with the industry's shift toward **Instruction Tuning (IT)** and **Reinforcement Learning from Human Feedback (RLHF)**. The analysis of research data suggests that this proficiency is not just a coincidence, but is materially shaped by training methodologies that (often unintentionally) act as a large "causal curriculum."

3.1 The "Coincidence" of Pre-training: Implicit World Models

Before discussing specific training, one must acknowledge the foundation: pre-training on web-scale corpora (The Pile, Common Crawl, C4). The primary objective of these models is next-token prediction.

Theoretical research suggests that optimising for prediction error on a diverse corpus forces the model to learn a compressed representation of the data generating process — effectively, a "world model". Because human language is intrinsically causal (we tell stories of *why* things happen), a model trained to predict the next word in a narrative must implicitly model causal physics.

- *Example:* To predict the token "shattered" following the context "The vase fell off the shelf and...", the model must encode the causal relationship between *falling (gravity)* and *shattering (impact)*.

Recent theoretical work on **Semantic Characterization Theorems** argues that the latent space of these models evolves to map the topological structure of these semantic relationships. Thus, the "native" understanding is partially a coincidence of the data's nature: the model learns causality because causality is the glue of human discourse.

3.2 The Instruction Tuning Hypothesis: Specific Training via Templates

The transition from "text completer" (GPT-3) to "helpful assistant" (ChatGPT) was mediated by **Instruction Tuning**. This process involves fine-tuning the model on datasets of (Instruction, Output) pairs. An analysis of major instruction datasets — **FLAN**, **OIG**, and **Dolly** — reveals that they are saturated with causal reasoning tasks.

3.2.1 The FLAN Collection: The Template Effect

The **FLAN (Finetuned Language Net)** project was instrumental in this development. Researchers took existing NLP datasets (including causal extraction datasets) and converted them into natural language templates.

- **The Mechanism:** A classification task from the *COPA (Choice of Plausible Alternatives)* dataset, which asks for the cause of an event, was transformed into prompts like: *"Here is a premise: The man broke his toe. What was the cause?"*
- **The Scale:** FLAN 2022 aggregated over 1,800 tasks. By training on millions of examples where the input is a scenario and the output is a causal explanation, the model explicitly learned the linguistic patterns of *identifying influence*.
- **Mixed Prompting:** Crucially, FLAN mixed **Chain-of-Thought (CoT)** templates (which require intermediate reasoning steps using "therefore," "because," "so") with standard prompts. This trained the model not just to guess the answer, but to *generate the causal logic* leading to it.

This contradicts the idea that the capability is purely coincidental. The models were specifically drilled on millions of "causal identification" exercises, disguised as instruction following.

3.2.2 Open Instruction Generalist (OIG) and Dolly

The **OIG** and **Dolly** datasets expanded this to open-domain interactions. These datasets contain thousands of "brainstorming" and "advice" prompts.

- *Data Evidence:* An entry from the OIG dataset reads: *": I'm having trouble finding a good job, what can I do to improve my chances? : One thing a person could do is...".*
- *Implication:* To answer this, the model must access a causal chain: *Action (revise resume) -> Effect (better chances)*. The prevalence of "how-to" and "why" questions in these datasets forces the model to organise its internal knowledge into causal structures (Means-End reasoning).

3.3 Reinforcement Learning from Human Feedback (RLHF): The Coherence Filter

The final layer of "specific training" is **RLHF**. In this phase, human annotators rank model outputs based on preference.

- **Preference for Logic:** Research indicates that humans have a strong bias for **causal coherence**. A narrative that flows logically (Cause A -> Effect B) is rated higher than one that is disjointed.
- **Length and Explanation Bias:** RLHF has been shown to induce a "length bias," where models produce longer, more detailed explanations to secure higher rewards. In the context of causality, this encourages the model to generate elaborate causal chains.

- **Sycophancy:** However, this training can also lead to "hallucinated causality." If prompted with a leading question that implies a false causation (e.g., "Why does the moon cause earthquakes?"), an RLHF-aligned model might generate a plausible-sounding but scientifically incorrect causal explanation, prioritising "helpfulness" over "truth".

Conclusion on Training vs. Coincidence: The capability is a hybrid. The *potential* to understand causality is a coincidence of pre-training scale (World Models), but the *ability to natively identify and articulate* it in response to a prompt is the result of specific Instruction Tuning and RLHF regimens that prioritise causal templates and coherent explanation.

---

4. Linguistic Frameworks: Analysing "Ordinary" Causation

This note emphasises the "native ordinary language concept of causation." To understand this, we must look beyond computer science to **Cognitive Linguistics**. Recent research has benchmarked LLMs against human linguistic theories, particularly **Talmy's Force Dynamics** and **Implicit Causality (IC)**.

4.1 Force Dynamics: Agonists and Antagonists in Latent Space

Leonard Talmy's theory of **Force Dynamics** posits that human causal understanding is rooted in the interplay of forces: an **Agonist** (the entity with a tendency towards motion or rest) and an **Antagonist** (the opposing force).

- *Linguistic Patterns:* "The ball kept rolling despite the grass" (Agonist: Ball; Antagonist: Grass). "He let the book fall" (Removal of Antagonist).
- *LLM Evaluation:* Recent studies have tested LLMs on translating and explaining these force-dynamic constructions.
    - **Findings:** GPT-4 often shows a solid grasp of these concepts. When translating "He let the greatcoat fall" into languages like Finnish or Croatian, the model can select verbs that convey "cessation of impingement" (allowing) rather than "onset of causation" (pushing).
    - **Implication:** This suggests that LLMs have acquired a **schematic semantic structure** of causality. They do not merely predict words; they map the *roles* of entities in a physical interaction. However, this capability degrades in abstract social contexts. For example, in the sentence "Being at odds with her father made her uncomfortable,"

models sometimes misidentify the Agonist/Antagonist relationship, struggling to map "emotional force" as accurately as "physical force".

4.2 Implicit Causality (IC) Verbs

Another major area of inquiry is **Implicit Causality (IC)**, which refers to the bias native speakers have regarding *who* is the cause of an event based on the verb used.

- *NP1-Bias (Subject):* "John **upset** Mary." (Why? Because *John* is annoying).
- *NP2-Bias (Object):* "The teacher **scolded** Mary." (Why? Because *Mary* did something wrong).

In this sense, "bias" means the useful working expectation of which part of the sentence is the cause.

Benchmarking Results: Research comparing LLM continuations to human psycholinguistic data reveals a high degree of alignment.

- **Coreference:** When prompted with "John amazed Mary because...", LLMs overwhelmingly generate continuations referring to John, matching human NP1 bias.
- **Coherence:** Humans tend to provide *explanations* following these verbs. LLMs mirror this "explanation bias," prioritising causal connectives over temporal or elaborative ones in these contexts.
- **Significance:** This indicates that LLMs have encoded the **pragmatics of blame and credit** inherent in ordinary language. They "know" that "apologizing" implies the subject caused a negative event, while "thanking" implies the object caused a positive one. This is crucial for the "native" feel of their interactions — they navigate the social logic of causality fluently.

4.3 The Limits of "Native" Understanding: The Causal Parrot Debate

Despite these successes, a vigorous debate persists regarding whether this constitutes "understanding" or merely "stochastic parroting".

- **The "Parrot" Argument:** Critics argue that LLMs fail when the linguistic surface form is stripped away. On benchmarks like **CausalProbe** , which uses fresh, non-memorised data, model performance drops significantly. This suggests that LLMs rely on **Level 1**

(Association) reasoning — pattern matching seen examples — rather than **Level 2 (Intervention)** reasoning.
- **The "Simulacrum" Argument:** Conversely, the **Semantic Characterization Theorem** proposes that the model's high-dimensional space creates a functional topology that is mathematically equivalent to a discrete symbolic system. Even if the model has never "seen" a glass break, its representation of "glass" and "break" are topologically linked in a way that allows it to simulate the causal outcome efficiently.

A caution about dated negative findings: many "LLMs cannot do causal reasoning" results from around 2020-2022 are best read as results about a specific model family and evaluation setup (often base models, short prompts, and narrow benchmarks). Newer instruction-tuned models (and more careful prompting protocols) can reduce some of these gaps on standard tests, but the picture remains mixed and sensitive to benchmark design, leakage, and what is being counted as "causal reasoning" versus plausible explanation.

---

5. Benchmarking the "Informal": From Social Media to Counterfactuals

The evaluation of causal understanding has evolved from F1 scores on extraction tasks to sophisticated benchmarks that test the model's ability to handle the messy, informal causality of the real world.

5.1 CausalTalk: Informal Causality in Social Media

The **CausalTalk** dataset focuses on "passages where one thing influences another" in informal contexts.

- *The Challenge:* In social media (e.g., Reddit), causality is often expressed without explicit markers. "I took the vaccine and now I feel sick" contains no "because," yet the causal assertion is clear.
- *Findings:* LLMs often perform well at identifying these **implicit causal claims**, sometimes outperforming traditional supervised models. They can detect "gist" causality — the overall causal assertion of a post — even when it is buried in sarcasm or non-standard grammar.
- *Application:* This is critical for **misinformation detection**. Models are being used to identify exaggerated causal claims in science news (e.g., reporting a correlation as a causation). However, LLMs sometimes struggle to distinguish between someone *reporting* a correlation and *asserting* a causation, highlighting a nuance gap in informal causal language.

5.2 Explicit vs. Temporal Confusion (ExpliCa)

The **ExpliCa** benchmark investigates a specific failure mode: the confusion of time and cause.

- *The Fallacy: Post hoc ergo propter hoc* ("After this, therefore because of this").
- *LLM Behaviour:* Research shows that LLMs are prone to this fallacy. When events are presented in chronological order ("The sun set. The streetlights turned on."), models are statistically more likely to infer a causal link than humans, who might see it as mere sequence. This suggests that the "native" understanding is heavily biased by the **narrative structure** of training data, where chronological sequencing often implies causality.

Again, a caution about dated negative findings: while these weaknesses are interesting, frontier models in 2025 are much less likely to display them.

5.3 Counterfactuals and "What If" (CRASS)

The **CRASS** (Counterfactual Reasoning Assessment) benchmark tests the model's ability to reason about what *didn't* happen.

- *Task:* "A man drinks poison. What would have happened if he drank water?"
- *Results:* While base models perform adequately, fine-tuning with techniques like **LoRA (Low-Rank Adaptation)** significantly boosts performance. This reinforces the "training hypothesis" — the capacity for causal reasoning is latent in the weights but requires specific activation (instruction tuning) to be robustly deployed.

---

6. Philosophical Dimensions: Symbol Grounding and World Models

The impressive performance of LLMs on causal tasks raises profound philosophical questions about the nature of meaning. Can a system that has never physically interacted with the world truly understand "force," "push," or "cause"?

6.1 The Symbol Grounding Problem

Cognitive scientists have long argued that human concepts are **grounded** in sensorimotor experience. We understand "heavy" because we have felt gravity.

- **The Disembodied Mind:** LLMs are disembodied. Their understanding of "force" is purely distributional — "force" is defined by its mathematical proximity to "push," "move," and

"impact" in vector space.

- **Cognitive Alignment:** Research using the **Brain-Based Componential Semantic Representation (BBSR)** shows that LLM representations align well with human cognition for concrete concepts but diverge for embodied experiences (e.g., olfaction, gustation) and spatial cognition.

- **Functional Understanding:** However, proponents of the "Functionalist" view argue that if an LLM can answer "What happens if I drop this?" indistinguishably from a human, it possesses a **functional understanding** of causality. The **Semantic Characterization Theorem** supports this by demonstrating that continuous learning dynamics can give rise to stable, discrete semantic attractors that behave like symbolic rules.

---

7. Current Frontiers (2024-2025): Reasoning Models and Future Directions

The field is currently undergoing another shift with the introduction of "Reasoning Models" (e.g., OpenAI's o1/o3 series, DeepSeek R1).

7.1 Chain-of-Thought Monitoring and "Thinking" Tokens

Newer models are trained to produce hidden "chains of thought" before generating a final answer.

- *Impact on Causality:* This allows the model to perform **intermediate causal checks**. Instead of predicting the effect immediately, the model can "reason" silently: *Premise -> Mechanism -> Potential Confounders -> Conclusion*.

- *Research Findings:* Snippet discusses "CoT Monitoring," showing that these internal reasoning traces can be monitored to detect "reward hacking" or deceptive alignment. This suggests a move toward making the model's implicit causal reasoning **explicit** and **verifiable**.

7.2 Causal Graph Construction

Recent work has moved back to structure, using LLMs to *extract* and *construct* **Causal Graphs** (DAGs) from unstructured text, a process sometimes known as causal mapping.

- *Method:* Rather than asking the LLM to just "answer," researchers prompt it to output a graph: `Nodes:, Edges:`.

- *Result:* This leverages the LLM's linguistic fluency to structure knowledge, which can then be processed by formal causal inference algorithms, bridging the gap between "informal ordinary language" and "formal causal calculus."

---

8. Conclusion

The research of the last decade suggests that the "native" causal understanding of LLMs is a constructed capability, developed through large-scale training on human text and refined by human preference signals. It is not just a coincidence, but a plausible consequence of optimising models to predict a world that is described in strongly causal terms.

1. **Origin:** The capability originates in **pre-training**, where the model learns the distributional "shadow" of causation cast by billions of human sentences.
2. **Development:** It is sharpened by **Instruction Tuning** (FLAN, Dolly), which explicitly teaches the model the "language game" of explanation and consequence through millions of templates.
3. **Refinement:** It is polished by **RLHF**, which imposes a human preference for logical coherence and narrative flow, effectively pruning non-causal outputs.
4. **Nature:** This understanding is **linguistic and schematic**. It often mirrors the force dynamics and implicit biases of human language, but can remain brittle when faced with novel physical interactions or rigorous counterfactual logic.

Overall, these systems can simulate many of the linguistic patterns humans use when describing causes and effects. That makes them useful for drafting, paraphrase, and extraction, but it should not be treated as evidence of intervention-level causal knowledge.

---

9. Comparative Data Tables

Table 1: Evolution of Causal Tasks and Metrics (2015-2025)

Era	Primary Focus	Methodology	Dominant Datasets	Typical Metric	"Native" Capability
2015-2018	Relation Classification	SVM, RNN, Sieves	SemEval-2010 Task 8, EventStoryLine	F1 Score (~0.50-0.60)	None (Pattern Matching)

Era	Primary Focus	Methodology	Dominant Datasets	Typical Metric	"Native" Capability
2019-2021	Span/Context Extraction	BERT, RoBERTa	Causal-TimeBank, BioCausal	F1 Score (~0.72)	Contextual Recognition
2022-2025	Generative Reasoning	GPT-4, Llama, Instruction Tuning	CausalTalk, CRASS, ExpliCa	Accuracy, Human Eval	Generative/Schematic

Table 2: Performance on Causal Benchmarks (Selected Studies)

Benchmark	Task Description	Model Class	Performance Note	Source
SemEval Task 8	Classify relation between nominals	BERT-based (BioBERT)	~0.72-0.80 F1 (High accuracy on explicit triggers)	
CRASS	Counterfactual "What if" reasoning	GPT-3.5 / Llama	Moderate baseline; significantly improved with LoRA/PEFT	
CausalProbe	Causal relations in *fresh* (unseen) text	GPT-4 / Claude	Significant drop compared to training data; suggests memorisation	
Implicit Causality	Predicting subject/object bias (John amazed Mary)	GPT-4	High alignment with human psycholinguistic baselines	
Force Dynamics	Translating "letting/hindering" verbs	GPT-4	High accuracy in preserving agonist/antagonist roles	

Table 3: Key Instruction Tuning Datasets Influencing Causal Capability

Dataset	Content Type	Causal Relevance	Mechanism of Training	Source
FLAN	NLP Tasks -> Instructions	High (COPA, e-SNLI templates)	Explicitly maps "Premise" -> "Cause/Effect" in mixed prompts	
OIG	Open Generalist Dialogues	High (Advice, How-to)	Teaches Means-End reasoning (Action -> Result)	
Dolly	Human-generated Q&A	High (Brainstorming, QA)	Reinforces human-like explanatory structures	
CausalTalk	Social Media Claims	High (Implicit assertions)	Captures "gist" causality in informal discourse	