Right for the Right Reasons? A Chain-of-Thought Analysis of Commonsense Causal Reasoning

Serena Auriemma[1,*], Martina Miliani[1], Alessandro Bondielli[1,2], Lucia Passaro[1] and Alessandro Lenci[1]

[1]CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa, Pisa, Italy.
[2]Department of Computer Science, University of Pisa, Pisa, Italy.

Abstract

Chain-of-thought (CoT) prompting is widely used to improve model performance, but it remains unclear whether longer reasoning traces reflect genuinely more elaborate reasoning or merely more verbose output. We present a fine-grained analysis of CoT quality in the Qwen3.5 model family (0.8B–27B parameters) on commonsense causal reasoning, using an LLM-as-judge framework to annotate reasoning traces at two levels: trace-level validity and chunk-level reasoning strategies, spanning both reasoning strategies (e.g. reasoning mechanism articulation, evaluation of alternatives, counterfactual reasoning) and CoT artifacts. Across two task formats, four causal-plausibility variants, and three thinking-token budgets, we find that reasoning validity and answer accuracy are tightly coupled only above a model-scale threshold, below which CoT is actively counterproductive. Models rely on a fixed, largely budget-invariant strategy schema dominated by reasoning mechanism articulation and almost entirely lacking counterfactual reasoning, unlike typical human reasoning on this task. We further identify a distinct dissociation between reasoning validity and answer correctness specifically on implausible causal scenarios, and show that CoT artifacts systematically increase when models struggle. Our results suggest that larger thinking budgets primarily help larger models execute an already-fixed reasoning schema to completion, rather than inducing qualitatively different reasoning.

Keywords

Chain-of-Thought, Commonsense causal reasoning, LLM-as-judge, Reasoning mechanism

1. Introduction

For humans, the faculty of language presupposes a capacity for reasoning, whereas large language models (LLMs) and so-called reasoning models invert this relationship. Models first learn to model language via pre-training, and only through additional training stages do they learn to "reason" over complex problems by decomposing them into a step-by-step procedure expressed in natural language. A model's reasoning is therefore tightly coupled to its training and the data it is trained to reason over. Such ability is susceptible to the well-known problem of out-of-distribution generalization: chain-of-thought (CoT) reasoning that is effective on in-distribution problems degrades under even moderate distribution shifts, suggesting that models often recognize patterns from the training distribution rather than acquiring general, transferable reasoning capabilities [1, 2]. This raises the question of whether models possess a genuine capacity for critical reasoning, or whether such reasoning is an *illusion of thinking* that dissolves when problems can no longer be matched to familiar patterns [3].

Commonsense causal reasoning is an ideal testbed for this debate. Recognizing cause-effect relations is a hallmark of human-like intelligence [4], requiring models to go beyond pattern matching by distinguishing causation from association and reasoning counterfactually. Among the various forms of causal reasoning (e.g., over mathematical data, code, or physical simulations) commonsense causality requires grasping causal relations grounded in everyday world knowledge, and it is here that LLMs reveal their most telling gaps. On the one hand, much of the relevant world knowledge is so taken for

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy
*Corresponding author.
serena.auriemma@fileli.unipi.it (S. Auriemma); martina.miliani@fileli.unipi.it (M. Miliani); alessandro.bondielli@unipi.it (A. Bondielli); lucia.passaro@unipi.it (L. Passaro); alessandro.lenci@unipi.it (A. Lenci)
0009-0006-6846-5826 (S. Auriemma); 0000-0003-1124-9955 (M. Miliani); 0000-0003-3426-6643 (A. Bondielli); 0000-0003-4934-5344 (L. Passaro); 0000-0001-5790-4308 (A. Lenci)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4315/paper10.pdf

granted by humans that it rarely appears explicitly in training data (everyone knows that bananas are yellow, without needing to state it), a phenomenon known as *reporting bias* [5]. On the other hand, the sheer variability of scenarios to which commonsense causality may apply can push models in opposite directions: either to over-generalize, imputing a causal link between events that merely resemble causal ones, or to confront an unfamiliar scenario for the first time, testing their ability to tell genuine causes from spurious ones. Crucially, because a causal relation can be perturbed in a controlled way – through plausibility manipulations of candidate causes – commonsense causality can also be used to probe the coherence and validity of a model's reasoning: verifying that a model is actually reasoning causally requires looking beyond task accuracy to the *type* of reasoning steps that led to a given answer and to their robustness across input perturbations [6, 7, 8, 9].

This paper addresses precisely this question: whether models actually reason over cause–effect relations, which strategies they apply most frequently, and whether the strategies they apply lead to correct or incorrect outputs. Drawing on the taxonomy of reasoning strategies introduced by [10], we simplify and adapt their scheme into an annotation framework amenable to automatic annotation by an LLM-as-a-judge, which we use to label the CoT produced by models on a commonsense causal reasoning task. Each item asks whether a causal link holds between two sentences, $sent\ A$ and $sent\ B$, where $sent\ A$ instantiates one of four experimental conditions: a causal sentence (BASE); a semantically similar causal sentence (CAUSAL-EQUIVALENT); a sentence describing an atypical or fantastical scenario that nonetheless preserves the causal link to $sent\ B$ (CAUSAL-IMPLAUSIBLE); and a non-causal sentence (NON-CAUSAL). Together, these conditions span a graded plausibility continuum along which we can probe the coherence of a model's reasoning. We further operationalize the task in two formats: a yes/no question, a multiple-choice non-cause selection,so as to separate variation in the CoT and in the final answer that is attributable to the response format itself.

Our experiments cover a single model family, namely `Qwen3.5`, at four scales (from 0.8B to 27B), in both thinking and non-thinking modes, which lets us trace how the performances and CoT changes with model scale. Within the thinking mode we additionally vary the reasoning budget across three settings: 1) a maximum length of 1.5k tokens; 2) 3k tokens, i.e., twice as much as setting 1; 3) 25k tokens, a substantially larger margin. We do so to assess how trace length affects the models' ability to carry their reasoning through to completion. Beyond task performance, the CoT produced in each setting is automatically annotated by `Qwen3.5-122B-A10B`, a larger Mixture-of-Experts (MoE) model from the same family and therefore aligned with the same reasoning regime as the evaluated sibling models, following a purpose-built annotation scheme. The scheme (cf. Section 3.4) operates at two levels: a trace-level annotation of reasoning validity, and a chunk-level annotation of the reasoning strategies used in a trace. Building on this annotation, we address the following research questions:

RQ1 Does CoT quality (reasoning validity) track task performance? In particular, how frequent is the dissociation between a correct answer and invalid reasoning (and the reverse)?

RQ2 How are reasoning strategies distributed, and does this distribution differ by task format (yes/no vs. multiple-choice), by $sent\ A$ plausibility condition, and by model scale?

RQ3 Which reasoning strategies are most associated with correct answers and with valid reasoning?

RQ4 How do thinking budgets (1.5k / 3k / 25k tokens) affect the reasoning traces – do larger budgets yield better reasoning or merely longer and more repetitive traces?

2. Related Work

The rise of LLMs has renewed the interest in causality and its linguistic expression, where the central question is whether fluent causal language reflects genuine reasoning or mere pattern recombination over training data [11, 12, 9]; recent overviews map this landscape across causal discovery, counterfactual, and actual-causality tasks [13, 14, 15, 16]. Causal reasoning has been probed in several domains including code and mathematics [17], domain-specific observational discovery [18] and commonsense,

or empirical, causal reasoning [19]. Our focus lies on the latter, and specifically on Pairwise Causal Discovery (PCD), the task of discerning pairwise causal relations from observational data [14, 20, 21, 22]. The dataset we use, AITIA, builds on [22] by pairing each base causal item with three controlled variants, namely CAUSAL-EQUIVALENT, CAUSAL-IMPLAUSIBLE, and NON-CAUSAL, spanning a plausibility continuum that probes whether models track the underlying causal structure or are misled by surface form and topical similarity.

Probing reasoning through input perturbations. The literature has shown that causal discovery performances degrade on counterfactual task variants [6, 7]. Studies on stress testing reasoning models against input perturbations [8, 9] have shown sharp accuracy drops e.g. on syllogisms with fantastical antecedents [23]. Regarding causal reasoning tasks, CoT prompting has yielded mixed results: it only partially helps [24] and does not consistently improve PCD [13]. Moreover, athours in [25] show that models fall back on pretraining-stored causal associations when narrative evidence conflicts with them, taking a a shortcut that neither CoT nor in-context learning resolves.

The AITIA dataset we use allows us for probing causal reasoning by manipulating the input rather than the model, testing whether a judgment survives controlled perturbations of the candidate cause. Counterfactual and implausible variants are especially diagnostic in this context.

Analyzing the chain of thought beyond accuracy. Beyond perturbation studies that still read reasoning off task accuracy, another line argues that the final answer reveals little about the process behind it, and evaluates the reasoning trace directly. Mondorf and Plank [26] frame this shift from reasoning performance to reasoning behavior, organizing methods into conclusion-based, rationale-based, interactive, and mechanistic evaluation; our study is rationale-based. Closest to this work, Mondorf and Plank [10] draw on cognitive psychology to characterize the inferential strategies LLMs use in propositional logic (supposition following, chain construction, and a symbolic strategy absent in human reasoners), manually annotating both strategy and logical soundness. They observe that accuracy is a poor proxy for reasoning validity, since models are often "right for the wrong reasons," and sound reasoning sometimes leads to wrong answers [26, 10].

In this work, we adopt this perspective but depart from prior work in two ways. We operationalize a simplified strategy taxonomy for an LLM-as-a-judge protocol, and we probe *commonsense causal* reasoning, where cause plausibility can be perturbed in a controlled way, rather than neutral-content deduction. This lets us test not just whether a trace is internally valid but whether it stays robust as the causal signal weakens.

LLMs as judges. The *LLM-as-a-judge* paradigm prompts a language model to assess whether an output meets a given criterion, offering a scalable alternative to human annotation [27]. It is now widely used across NLP benchmarks and even as a training reward signal, valued for approximating human judgment at scale [28]. Its reliability, however, is not unqualified: despite high reported agreement with human raters [27], LLM judges show systematic biases, including verbosity, authority, and self-enhancement effects, that persist even in strong judges [29], and position bias that intensifies when candidates are of similar quality [30].

Given these findings, we use an LLM-as-a-judge for its scalability, but validate its output against manual inspection on a subset of chunks and traces (Section 3.5). Since this check did not reveal systematic annotation failures, we retain the judge's labels for the full dataset, while treating the manual inspection as a sanity check rather than a full human-agreement study.

3. Experimental Setup

3.1. Data

We evaluate the models on AITIA, a contrastive dataset for commonsense causal reasoning in LLMs, built as an extension to ExpliCa [22]. Each item is a causal schema pairing a Sentence A (a cause) with

Condition	Sentence A	Sentence B
CAUSAL	It started raining.	The driver turned the wipers on.
CAUSAL-EQUIVALENT	There was a sudden downpour.	
CAUSAL-IMPLAUSIBLE	A flying whale released water over the city.	
NON-CAUSAL	The clouds turned pink at sunset.	

Table 1
An example of four Sentence A conditions in AITIA for a single item, all paired with the same Sentence B.

a Sentence B (its probable effect). Starting from 107 causal pairs selected from ExpliCa, each pair is enriched with three contrastive Sentence A variants that hold Sentence B fixed: a CAUSAL-EQUIVALENT sentence describing a similar event that preserves the causal link; a CAUSAL-IMPLAUSIBLE sentence describing a fantastical event that nonetheless maintains it; and a NON-CAUSAL sentence for which no causal link with Sentence B holds but there is some form of thematic similarity. See Table 1 for an example. The final dataset includes 107 quadruples (428 sentences in total). This contrastive structure probes whether models track causal relations consistently as premise plausibility varies.

3.2. Tasks

Task 1: Yes/No Causality Judgement In the first task, the model is asked to decide whether the event in Sentence A plausibly causes the event in Sentence B, answering only with *Yes* or *No*. Each quadruple is queried four times, once per Sentence A variant. We measure accuracy, i.e. the share of items with the expected answer: *Yes* for CAUSAL, CAUSAL-EQUIVALENT, and CAUSAL-IMPLAUSIBLE, *No* for NON-CAUSAL; when leveraging thinking mode, items left unanswered because the CoT exhausts the token budget are counted as errors.

Task 2: Multiple-Choice (Non-Cause Selection) In the second task, given Sentence B and the four Sentence A variants as options, the model must select the one that is *not* a cause of Sentence B, i.e., the NON-CAUSAL variant. Variant-to-option (A–D) assignment is randomized per item to control for position bias. Accuracy is the proportion of items where the NON-CAUSAL variant is correctly chosen, with unanswered thinking-mode items again counted as errors.

3.3. Models

We run all experiments on a single model family, `Qwen3.5` [31]. We chose this family for two reasons: first, it has shown strong results on ExpliCa [22]; second, it includes models at different parameter scales, covering a broad portion of the size spectrum. We consider only dense (non-MoE) models of the family, and evaluate the `-0.8B`, `-2B`, `-9B`, and `-27B` variants. This gives us a uniform lineup of variants across parameter scale of models that can be run on a single, consumer-grade GPU, with a direct toggle to enable/disable "thinking" (CoT) mode.

Each model is evaluated, using the same prompt in two ways. In *non-thinking mode* the model emits a direct answer, with generation capped at 8 tokens. In *thinking mode* the model first produces an explicit reasoning chain and only then commits to a final answer. To enable the thinking mode, we simply append the <think> token to the prompt. We use three different context budgets, namely 1.5k, 3k, and 25k tokens.

As for the judge model, we employ `Qwen3.5-122B-A10B-FP8`. It is the second largest model of the family. It is a MoE model with 122B total parameters and 10B active ones. We use the quantized fp8 variant due to memory limitations.

We used HuggingFace models via the vLLM inference engine. All experiments were conducted on a GPU node with 4 Nvidia A100 80GB GPUs. All models were used at their default precision level (fp16 for the tested models, fp8 for the judge). We used the default chat template and default generation parameters. Tested models could be fit onto a single GPU. For the judge model we used

all available compute to speed-up the computation. Overall, the experiment required roughly 30 GPU hours, including obtaining answers from tested models and labelling the CoT with the judge.

Once collected, models' responses were parsed using simple regular expressions, e.g., to isolate either "yes" or "no" for Task 1, or one of the "A", "B", "C", or "D" options for Task 2. Given the relatively small size of the datasets, we also manually revised the output of the regular expressions to ensure they matched the actual models' responses. We note that all models, including the smaller `-0.8B` variant were able to follow instructions and output only one of the expected responses.

3.4. CoT Annotation Framework

We analyze the chains of thought along two complementary axes, each implemented as a separate LLM-as-a-judge pass. The first operates over the **trace** as a whole, while the second operates over individual reasoning steps (henceforth referred as **chunks**) and labels the *strategies* they employ. In neither case the judge model has access to the dataset's gold labels, so that the annotation reflects the trace itself rather than the correctness of the final answer.

At the trace level we annotate the classical logical notion of *validity*, adapted from the *validity and soundness* categories of Mondorf and Plank [10]. The judge identifies the premises the model relies on and the conclusion it draws, and returns a binary label: an argument is valid when, assuming its premises were true, the conclusion follows with no logical gap, non-sequitur, or contradiction, irrespective of whether the premises are actually true. Because causal reasoning is typically non-deductive, we relax strict deductive validity and treat a step as valid when its premises, if true, adequately warrant the conclusion. Unlike Mondorf and Plank [10], we do not additionally annotate *soundness*, because it is a notion proper to deductive reasoning: it requires an argument to be valid *and* to rest on premises that are factually true of the world, which in practice means checking the premises against the judge's own world knowledge. This notion does not transfer cleanly to our setting, because our dataset by design includes premises that are not meant to be world-true: the CAUSAL-IMPLAUSIBLE variant describes hypothetical, physically or causally unrealistic scenarios that are constructed to be structurally valid but not factually grounded. Scoring such items for soundness would conflate our deliberate plausibility manipulation with judge error, and would make soundness depend on the judge's own (unverifiable) world knowledge rather than on properties of the reasoning itself.

At the chunk level, we characterize *how* a model reasons rather than whether its conclusion holds. Each trace is segmented deterministically into chunks by splitting on blank lines, and the judge classifies each chunk independently, with a sliding context window of up to two preceding and one following chunk, assigning a multi-label, binary set of strategies. We build this inventory by simplifying and adapting the scheme of Mondorf and Plank [10], condensing categories that are fine-grained for human analysis but hard for an LLM judge to discriminate reliably, and adding categories specific to causal reasoning and to model-generated CoT behavior. The resulting inventory comprises five reasoning strategies: **Reasoning-mechanism articulation** (explaining *how* the event in one statement produces the event in the other, describing the intermediate process or mechanism that links them, and collapsing Mondorf and Plank [10]'s *chain construction, compound strategies*, and *concatenation* into one label), **Evaluation of alternatives** (weighing at least one alternative before committing to a causal link, subsuming Mondorf and Plank [10]'s *supposition following*), **Counterfactual reasoning** (reasoning about an alternative or removed condition, added for its centrality to genuine causal reasoning), **Self-correction / backtracking** (explicitly reversing an earlier position, cued by e.g. "wait", "actually"), and **Uncertainty / hedging** (expressing doubt without reversing a prior position), plus one non-reasoning category, **CoT artifact**, for chunks that merely restate task instructions, copy back input statements, or repeat the expected answer format, tracked separately to quantify how much of a trace is boilerplate rather than genuine reasoning.

We ask the judge model to provide independent labels for each of the reasoning strategies. A chunk may carry several strategies at once (e.g. restating the task and then reasoning). One caveat applies to self-correction and, to a lesser extent, evaluation of alternatives: because each chunk is judged within a bounded context window, a reversal or comparison that refers back to material outside the window

Category	Example	Cond./Format
Reasoning strategies		
Reasoning-mechanism articulation	*Does getting a promotion typically lead to celebration? Yes, promotions are positive life events that are commonly celebrated. Is it likely? Yes, it is a very common social norm and behavioral response to receive a promotion to share the good news and celebrate with loved ones. While it's not guaranteed (some might not celebrate, or might celebrate differently), the question asks if it is "likely". In common sense reasoning, a promotion is a strong trigger for celebration.*	Causal / Yes–No
Evaluation of alternatives	*Wait, I need to check if there is a scenario where getting excellent grades causes scolding. Maybe the child got excellent grades, and the mother scolded the child for not doing something else? No. Maybe the child got excellent grades, and the mother scolded the child for something else? Yes, but does A cause B? No. Okay, I'm confident.*	Non-causal / Yes–No
Counterfactual reasoning	*Wait, I need to check if there is a possibility that the answer is "No" because the action in B is not caused by the state in A. For example, if the man turned down the volume before the radio was on, then A doesn't cause B. But A says "The radio was on". So the radio is on. So the man cannot hear. So the man turns down the volume. This is a standard cause-effect. So, Yes.*	Causal / Yes–No
Self-correction / backtracking	*Wait, could it be C because of the tense? "Trains" vs "Won". If I say "He won the race because he trains every day", it's acceptable English. If I say "He won the race because a centaur trained him", it's factually false. Factual falsehood is a stronger reason for "NOT describing a cause" in a general context than tense agreement. Also, in many logic puzzles, "Centaur" is a clear marker for "False/Impossible".*	N/A / MCQ
Uncertainty / hedging	*Wait, I need to be careful. Is there any scenario where "went to the pharmacy" is considered a cause? Maybe if the pharmacy had a specific trigger? No. So A is the correct answer.*	N/A / MCQ
Non-reasoning strategy		
CoT artifact	1. *Analyze the Request: Input: Two sentences (A and B). Sentence A: "Silvia is allergic to flour." Sentence B: "Silvia cannot touch the dough." Task: Determine if the event in A causes (brings about) the event in B. Constraint: Answer only with "Yes" or "No".*	Causal-Equivalent / Yes–No

Table 2
Representative examples for each category in the chunk-level annotation inventory: five reasoning strategies and one non-reasoning category. The *Cond./Format* column reports the condition and task format of the trace each example is drawn from.

may be missed. In Table 2 we summarize the full inventory with examples.

3.5. Metrics and Judge Reliability Assessment

For the evaluation of both Task 1 and Task 2, we adopt a standard **accuracy** metric, defined with respect to the expected response under each causal variant. For Task 1, accuracy is the proportion of items for which the model produces the following response: *Yes* for the CAUSAL, CAUSAL-EQUIVALENT, and CAUSAL-IMPLAUSIBLE variants, and *No* for the NON-CAUSAL variant. For Task 2, the expected response is always the NON-CAUSAL option among the presented alternatives. For both tasks, whenever the model exhausts its token budget before producing a final answer, we count that response as incorrect.

Alongside accuracy, we also report a **valid reasoning rate**. It represents the proportion of items for which the trace-level judge assigned a positive validity label, computed as the number of valid traces divided by the total number of items, yielding a score between 0 and 1. Trace-level validity annotation is restricted to items for which the model completed its CoT and produced a final answer; interrupted traces, i.e. cases where the model exhausted its token budget before concluding the reasoning and generating a response, are treated by default as invalid, since validity is a judgment about a completed argument and cannot be meaningfully assessed on a truncated one. This mirrors how no-response items are handled for accuracy, so that the two rates are directly comparable. Incomplete traces are, however, still annotated at the chunk level to identify the reasoning strategies the model applied up to the point of truncation. This allows us to estimate how much of the allotted token budget is actually spent on substantive reasoning versus non-substantive content (e.g. CoT artifacts or repetition) before the trace is cut off, which is informative for characterizing budget usage independently of whether the model manages to reach a final answer.

At the chunk level, each reasoning strategy is a binary flag assigned independently to every chunk. To characterize strategy use at the trace level, we compute **density**: the mean of the chunk-level binary flag across all chunks of a trace, yielding a continuous value between 0 and 1 that reflects how pervasively a given strategy is used throughout the reasoning process. We adopt density rather than a simple presence indicator because longer CoT traces mechanically contain more chunks and are therefore

more likely to include at least one instance of any strategy.

To assess the reliability of the LLM-as-judge annotation, we conducted a manual qualitative inspection on two levels. At the chunk level, we randomly sampled 60 chunks from the annotated dataset and manually checked the correctness of the assigned reasoning-strategy labels against the annotation guidelines. At the trace level, we inspected 10 full CoT traces to verify the correctness of the judge's validity annotation. Manual inspection suggested that the judge applied the annotation guidelines consistently enough for large-scale annotation, although this remains a qualitative validation.

4. Results and Discussion

4.1. RQ1: CoT Quality and Task Performance

Table 3 reports, for each model and CoT length, the global and per-variant accuracy alongside the valid reasoning rate. The clearest pattern in the table is a scale effect. Above the 9B parameter threshold, models achieve both the highest accuracy and the highest valid reasoning rates, and the two quantities track each other closely: **as the valid reasoning rate increases, accuracy increases correspondingly**, on both tasks. For these larger models, accuracy also improves as the CoT budget grows, indicating that they are able to make productive use of additional tokens rather than merely padding the trace. Below this threshold, models show markedly lower accuracy on both tasks, but this is driven almost entirely by a high no-response rate.

Additionally, the table reveals **a minor dissociation between validity and correctness**, orthogonal to the no-response effect discussed above. In most rows the valid reasoning rate is slightly higher than accuracy, indicating a residual set of cases in which the model completes a structurally valid chain of reasoning but still arrives at an incorrect final answer: the model is, in this sense, "right in the process but wrong in the conclusion". This gap is modest and roughly constant across the CAUSAL, CAUSAL-EQUIVALENT, and NON-CAUSAL variants, but it is markedly larger, and grows with model scale and CoT budget, **on the CAUSAL-IMPLAUSIBLE variant**. For Qwen3.5-27B at 25k tokens, the valid reasoning rate on CAUSAL-IMPLAUSIBLE reaches 0.991, nearly matching the near-ceiling rates observed on the other variants, yet accuracy on this variant remains at 0.822, a gap of 16.9 points, the largest such gap we observe (Qwen3.5-9B at 25k shows an analogous gap of 11.3 points). This indicates that for the largest models, the bottleneck on implausible scenarios is no longer the ability to construct a structurally valid argument, which is achieved in virtually all cases, but a further difficulty specific to this variant: **the model reasons validly about a hypothetical, counter-to-expectation causal scenario yet still fails, more often than on any other variant, to convert that valid reasoning into the response the task expects**. A plausible explanation is a training-induced prior: implausible or fantastical causal scenarios are typically underrepresented as positive causal examples in naturally occurring text and in the kind of supervised or preference data used to train these models, where they more often appear as distractors or as negative/incorrect instances. Models may therefore carry over a latent bias that treats implausibility itself as evidence against a causal link, independently of the explicit reasoning chain that precedes the answer. Under this account, the model can correctly articulate, step by step, why the causal link holds in the hypothetical scenario, and still overturn that conclusion at the final-answer step in favor of the more familiar heuristic that "implausible sounding scenarios are usually the wrong answer." An alternative, non-mutually-exclusive explanation is that our judge evaluates the surface text of the reasoning trace rather than the model's underlying computation, so a chain of thought that reads as valid may not fully reflect the process that actually determined the final answer; we leave a closer investigation of this distinction, for instance through targeted interventions such as contrastive prompting or activation-level analysis, to future work.

4.2. RQ2: Distribution of Reasoning Strategies

Figure 1 reports the density of each reasoning strategy broken down by task format, causal variant, and model scale. Across all three comparisons, CoT ARTIFACT is by far the most prevalent category,

	Qwen3.5-0.8B				Qwen3.5-2B				Qwen3.5-9B				Qwen3.5-27B			
Metric	Non-think.	1.5k	3k	25k	Non-think.	1.5k	3k	25k	Non-think.	1.5k	3k	25k	Non-think.	1.5k	3k	25k
Task 1																
Acc. – Global	0.675	0.093	0.126	0.110	0.556	0.311	0.383	0.416	0.790	0.836	0.843	0.914	0.907	0.864	0.879	**0.925**
Acc. – Causal	0.860	0.047	0.140	0.112	0.523	0.439	0.514	0.570	0.869	0.944	0.944	0.991	**1.000**	0.953	0.953	0.991
Acc. – Causal-Eq.	0.813	0.121	0.112	0.121	0.430	0.430	0.449	0.495	0.766	0.841	0.841	0.916	**0.953**	0.860	0.879	0.916
Acc. – Causal-Impl.	0.710	0.037	0.056	0.047	0.327	0.159	0.234	0.262	0.607	0.701	0.710	**0.822**	0.776	0.692	0.738	**0.822**
Acc. – Non-Causal	0.318	0.168	0.196	0.159	0.944	0.215	0.336	0.336	0.916	0.860	0.879	0.925	0.897	0.953	0.944	**0.972**
Valid – Global	N/A	0.138	0.187	0.161	N/A	0.332	0.409	0.437	N/A	0.874	0.890	0.977	N/A	0.902	0.916	**0.991**
Valid – Causal	N/A	0.075	0.187	0.150	N/A	0.449	0.514	0.579	N/A	0.944	0.953	**1.000**	N/A	0.953	0.953	**1.000**
Valid – Causal-Eq.	N/A	0.206	0.215	0.224	N/A	0.439	0.467	0.514	N/A	0.888	0.897	**0.972**	N/A	0.897	0.916	**0.972**
Valid – Causal-Impl.	N/A	0.103	0.150	0.112	N/A	0.196	0.299	0.290	N/A	0.766	0.785	0.935	N/A	0.776	0.822	**0.991**
Valid – Non-Causal	N/A	0.168	0.196	0.159	N/A	0.243	0.355	0.364	N/A	0.897	0.925	**1.000**	N/A	0.981	0.972	**1.000**
No-response	N/A	0.839	0.785	0.808	N/A	0.666	0.586	0.549	N/A	0.124	0.110	0.012	N/A	0.098	0.079	**0.000**
Task 2																
Acc. – Global	0.150	0.028	0.000	0.019	0.308	0.252	0.402	0.383	0.813	0.748	0.794	0.925	0.944	0.813	0.860	**0.972**
Valid – Global	N/A	0.037	0.000	0.019	N/A	0.271	0.430	0.439	N/A	0.766	0.813	**1.000**	N/A	0.822	0.879	**1.000**
No-response	N/A	0.963	0.991	0.981	N/A	0.729	0.570	0.561	N/A	0.234	0.187	**0.000**	N/A	0.178	0.121	**0.000**

Table 3
Task 1 and Task 2 results by model and CoT length. **Acc.** = accuracy; **Valid** = valid reasoning rate. No-response items are counted as both incorrect and reasoning-invalid by default. Valid reasoning rate and no-response rate are not applicable (N/A) in the non-thinking condition, since no chain of thought is produced. Bold marks the best value in each row across all models and CoT lengths (for the no-response rows, best means lowest); ties are bolded jointly.

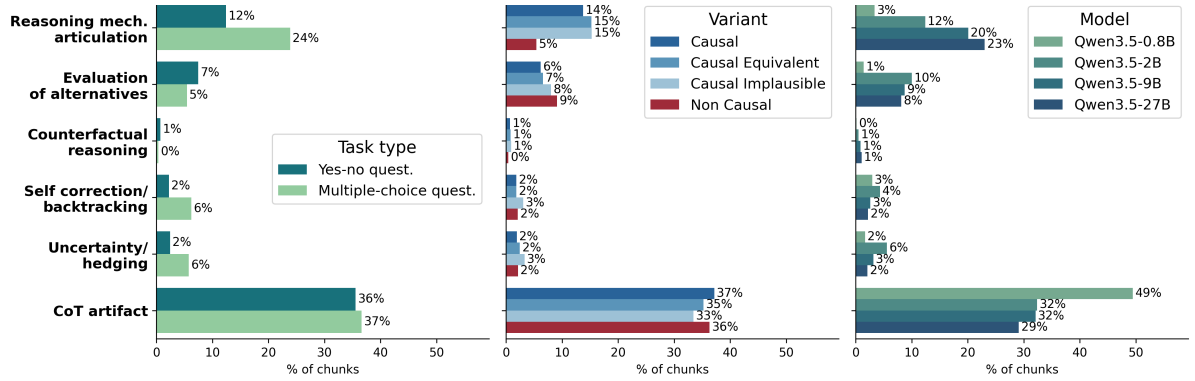


Figure 1: Distribution of reasoning strategies (density) by task format (left), causal variant for Task 1 (center), and model scale (right). Density is the mean of the chunk-level binary flag across all chunks of a trace, averaged across traces within each group.

accounting for roughly a third to a half of all chunks; substantive reasoning strategies are comparatively rare, with Reasoning-mechanism articulation and Evaluation of alternatives as the two most frequent, and Counterfactual reasoning essentially absent (at or below 1% in every condition). This confirms, **at the strategy level**, that **a large share of the generated CoT consists of non-substantive content regardless of task format, causal variant, or model scale**. Moreover, the strategy repertoire is uneven: models lean heavily on decomposing the causal chain into steps and weighing alternative explanations, while counterfactual reasoning, i.e., constructing a hypothetical scenario in which the premise does not hold and the consequence accordingly does not follow, is almost entirely absent. This is worth noting because the counterfactual move is what distinguishes a genuine cause from a mere correlate [32], and our tasks ask precisely about whether one specific event brings about another.

Regarding task format, the distribution of strategies differs substantially between the two formats, with the exception of the CoT-artifact category: multiple-choice traces show roughly double the density of Reasoning-mechanism articulation and triple the density of both Self-correction / backtracking and Uncertainty /hedging relative to yes/no traces, while Evaluation of alternatives is somewhat more frequent in the yes/no format. This pattern is consistent with the **multiple-choice format** being **more demanding to resolve: with more than two candidate answers to adjudicate, models**

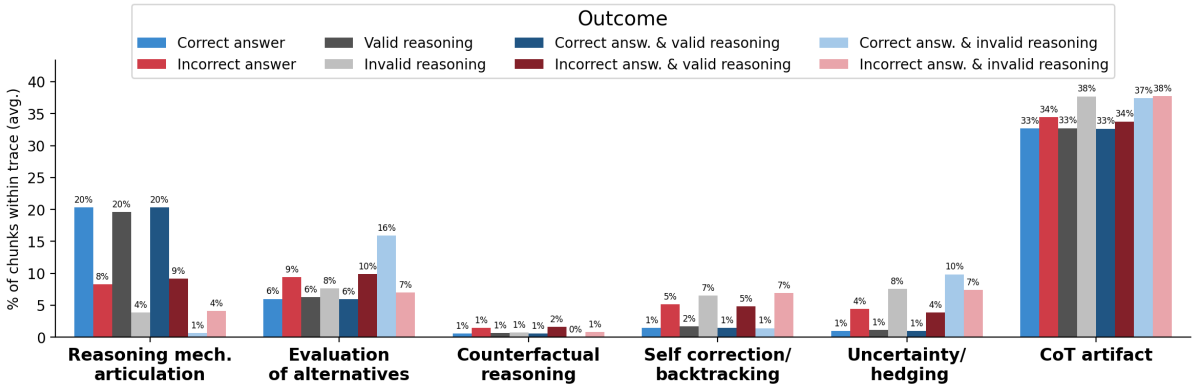


Figure 2: Association between reasoning strategies and answer correctness and reasoning validity, restricted to completed CoT traces with a final answer and validity annotation.

articulate the causal mechanism more explicitly and revise or hedge their position more often, whereas a binary yes/no judgment leaves comparatively less to backtrack on. This aligns with the higher token cost of the multiple-choice format observed in the accuracy results (Table 3), where Task 2 shows a higher no-response rate than Task 1 at constrained budgets.

Within Task 1, the clearest effect concerns Reasoning-mechanism articulation, which is flagged in 14–15% of chunks for the causal, causal-equivalent, and causal-implausible variants, but drops to just 5% for the non-causal variant. This is expected, since articulating a causal mechanism presupposes a causal link to reconstruct step by step, so traces concluding there is no causal link naturally show fewer occurrences of this strategy. Evaluation of alternatives, by contrast, increases monotonically from causal through causal-equivalent and causal-implausible to non-causal, suggesting that models weigh alternative explanations (reverse causation, confounding, coincidence) more heavily when the causal status of the pair is less straightforward to affirm outright. The causal-implausible variant also shows the highest density of both Self-correction / backtracking and Uncertainty / hedging, a small but consistent signal that **models are comparatively less confident when reasoning about implausible or fantastical scenarios**. This is in line with the accuracy results reported in Table 3, where causal-implausible is the hardest variant.

Finally, **model scale shows the strongest and most monotonic effect of the three comparisons: Reasoning-mechanism articulation increases steadily with scale, while CoT artifact decreases correspondingly**. The bulk of this shift occurs between the 0.8B and 2B models, with comparatively smaller gains at larger scales, suggesting a threshold effect whereby even a modest increase in scale substantially reduces boilerplate and increases substantive reasoning, after which returns diminish. Evaluation of alternatives follows a similar pattern, jumping from 1% at 0.8B to 8–10% at 2B and above. Notably, Self-correction / backtracking and Uncertainty / hedging both peak at 2B and decline at 9B and 27B. This suggests that backtracking and hedging behavior is most prominent at intermediate scale, where models are capable enough to notice tension or uncertainty in their own reasoning but not yet reliable enough to resolve it directly; the largest models appear to reach correct or confident conclusions with less visible self-revision.

4.3. RQ3: Strategy–Outcome Associations

Figure 2 reports strategy density across eight outcome groups, restricted – as detailed in Section 3.5 – to completed traces with a final answer and a trace-level validity annotation. **Reasoning-mechanism articulation is the strategy most clearly associated with both a correct answer and valid reasoning**: its density is essentially identical whether we condition on a correct answer alone, on valid reasoning alone, or on their conjunction, and it drops sharply for incorrect answers and for correct answers paired with invalid reasoning. This strategy is thus the clearest marker of a trace that is both right and right for the right reasons, and its near-total absence when the answer is correct but the

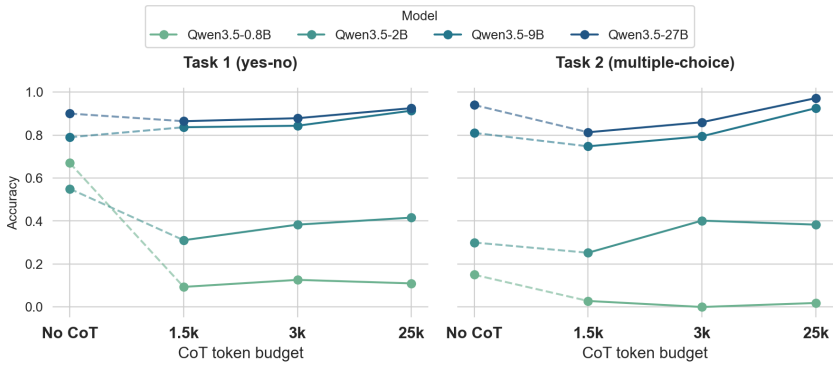


Figure 3: Accuracy as a function of CoT token budget (1.5k, 3k, 25k), by model and task format.

reasoning is invalid suggests that these traces reach a correct answer largely without building an explicit causal mechanism – more consistent with a fortuitous or heuristic-driven answer than with genuine causal reasoning.

Evaluation of alternatives and Uncertainty / hedging show a different pattern: both are denser in incorrect than in correct answers, denser in invalid than in valid reasoning, and peak by a wide margin in the correct-answer/invalid-reasoning group. This suggests that **when the model is more uncertain and more inclined to weigh alternative explanations, it frequently drifts toward an invalid line of reasoning** – even though this does not necessarily surface in the final answer, consistent with the validity–accuracy dissociation already discussed for RQ1 (Section 4.1).

Self-correction / backtracking follows a related but distinct pattern: rather than peaking in the correct-answer/invalid-reasoning group, it remains uniformly low (around 1%) across every subgroup in which the answer is correct, regardless of reasoning validity, and uniformly elevated (5–7%) across every subgroup in which the answer is incorrect. This indicates that self-correction tracks primarily with getting the final answer wrong, rather than specifically with the dissociation between valid reasoning and correctness: **a model that backtracks during its CoT does not typically recover into a correct answer**, and self-correction is better read as a symptom of item difficulty than as a corrective mechanism that reliably pays off.

CoT artifact increases slightly across the board in every incorrect and invalid subgroup relative to their correct/valid counterparts, suggesting that its prevalence is not simply a fixed stylistic habit, but **tracks — at least in part — how much the model is struggling with the item**. Counterfactual reasoning remains negligible throughout and does not differentiate any outcome group, consistent with its near-absence already noted in RQ2.

4.4. RQ4: Effects of Thinking Budget on Reasoning Traces

Figure 3 shows that accuracy increases monotonically with token budget for every model and on both tasks, but the size of the gain is highly scale-dependent. The two smaller models (0.8B, 2B) improve only marginally as the budget grows from 1.5k to 25k tokens and remain far below the larger models at every budget, whereas Qwen3.5-9B and Qwen3.5-27B show a clear, steady climb, especially on Task 2, with the largest jump occurring between the 3k and 25k conditions. This indicates that **additional thinking tokens are converted into higher accuracy mainly by models that are already large enough to make use of them, rather than uniformly across scale**.

This asymmetry also clarifies when CoT helps versus hurts overall, relative to the non-thinking baseline. The pattern, however, is not uniform across task format. On Task 1, Qwen3.5-9B improves monotonically over the non-thinking baseline at every budget, while Qwen3.5-27B shows a more nuanced trajectory: the intermediate budgets (1.5k, 3k) actually *underperform* the baseline, and only the largest budget (25k) yields a net gain. On Task 2, this nuanced pattern extends to Qwen3.5-9B as well: both intermediate budgets fall below the non-thinking baseline, and only 25k tokens recovers

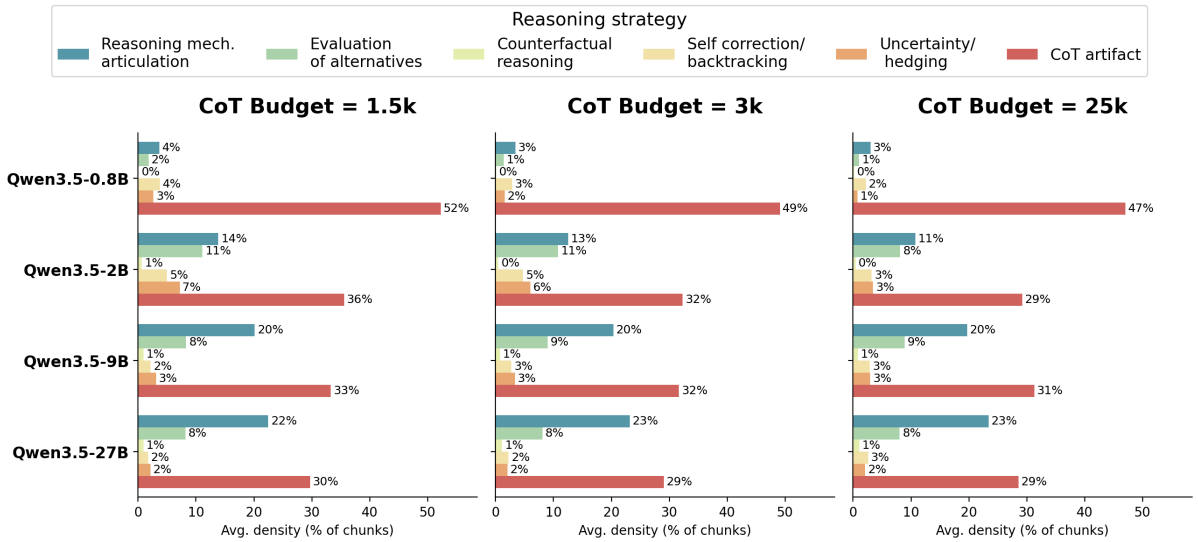


Figure 4: Reasoning strategy density by model and CoT token budget.

and then exceeds it, mirroring what Qwen3.5-27B shows on both tasks. Thus, even for the two largest models in our pool, CoT is not unconditionally beneficial: it pays off only once the budget is large enough to support a complete, valid argument, and the threshold at which this happens can shift with task format. For the smallest model (Qwen3.5-0.8B), CoT is uniformly counterproductive on both tasks, with accuracy under every thinking budget falling below the non-thinking baseline, since the model is consistently unable to complete the reasoning process within the allotted tokens. Qwen3.5-2B shows a similar, uniformly negative pattern on Task 1, but on Task 2 it behaves more like the larger models: only the smallest budget (1.5k) underperforms the baseline, while 3k and 25k both yield a net gain.

Figure 4 breaks this pattern down by reasoning strategy. CoT ARTIFACT dominates the token budget in every setting, but its share shrinks slightly as the budget grows and more markedly as model scale increases. This shrinkage, however, does not translate into a proportional gain for more substantive reasoning strategies: for Qwen3.5-0.8B and Qwen3.5-2B, the density of REASONING-MECHANISM ARTICULATION and EVALUATION OF ALTERNATIVES stays essentially flat, or even decreases slightly, as the budget increases, so the tokens freed up from artifact content are not redirected into deeper reasoning strategies. For Qwen3.5-9B and Qwen3.5-27B, by contrast, the reasoning strategies remain stable in density across budgets while CoT ARTIFACT continues to shrink, meaning that the additional tokens available at 25k are spent on some combination of more numerous reasoning chunks and reduced boilerplate, not on a qualitatively different reasoning style. Taken together with the accuracy trends, this points to a consistent reasoning profile that each model applies regardless of budget: **larger budgets give already-present reasoning strategies more room to run to completion, but the underlying combination of strategies stays essentially fixed** – what changes with budget is how much of that same schema the model manages to complete, not which strategies it draws on. For the smaller models, additional tokens mostly reproduce this same fixed mix at greater length, consistent with longer but not qualitatively better reasoning.

4.5. Key Findings

Reasoning validity and accuracy are a privilege of scale. Reasoning validity is directly tied to task performance, but both accuracy and valid reasoning are effectively out of reach for models below a certain scale threshold: smaller models are consistently unable to bring a commonsense causal argument to completion, to the point that they are better deployed in non-thinking mode than with an explicit chain of thought.

Models differ from expected causal reasoning strategies and rely on a fixed schema regardless of context or budget. Counterfactual reasoning, a central strategy for causal judgments, is nearly absent; models instead overwhelmingly favor reasoning-mechanism articulation, i.e., explicitly decomposing the causal chain into the intermediate steps linking the two statements, likely a schema inherited from CoT training itself rather than one genuinely tailored to causal reasoning. More broadly, the entire strategy mix is stable in its qualitative composition, with the relative ranking of strategies essentially unchanged across task format, causal variant, and token budget. Larger budgets thus do not change which strategies are used, only whether models above the scale threshold can carry this fixed schema through to a correct conclusion.

Implausible scenarios expose a distinct validity–accuracy gap. The CAUSAL-IMPLAUSIBLE variant, the atypical causal scenario models are presumably least exposed to during training, shows a validity–accuracy dissociation not observed at the same scale elsewhere: reasoning here remains largely valid, yet is disproportionately more likely to end in an incorrect answer than for any other variant. Unlike the accuracy gap observed at small model scale, which stems from incomplete reasoning, this gap emerges specifically where reasoning is otherwise reliable, pointing to a distinct failure mode tied to the plausibility of the scenario itself rather than to the model's capacity to reason at all.

Struggle, not deliberate padding, drives non-substantive content. A large share of every model's token budget is spent on CoT artifact rather than substantive reasoning, and this share grows precisely when models struggle: incorrect answers and invalid reasoning are systematically associated with more CoT artifact, more self-correction and backtracking, and more evaluation of alternatives and hedging, mirroring the pattern already observed for validity and accuracy.

5. Conclusion and Future Work

This paper examined the quality and composition of CoT reasoning in the Qwen3.5 model family on commonsense causal reasoning, using an LLM-as-judge annotation scheme checked against manual inspection. Our results show that reasoning validity and task accuracy are tightly coupled above a scale threshold, that models rely on a largely budget-invariant strategy schema dominated by mechanism articulation, that implausible causal scenarios expose a distinct validity–accuracy dissociation, and that non-substantive CoT artifact grows precisely when models struggle.

Several directions remain for future work. First, we plan to relate surface-level CoT annotations to indicators of the model's underlying computation, for instance using measures such as those proposed by Scalena et al. [33], to identify which chunks and, in turn, which items are computationally harder for the model, and to test whether the CoT faithfully reflects that computation or diverges from it. Second, we intend to extend the analysis to additional model families, to assess whether the strategy distribution observed here is a general property of LLM reasoning or an idiosyncrasy of this family's CoT training data and procedure. Finally, we plan a deeper validation of the annotation pipeline, via manual CoT annotation with inter-annotator agreement against the judge, alongside a finer-grained, clustering-based analysis directly on CoT chunks to test whether categories beyond our human-derived reasoning-strategy schema emerge from the data.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: improve writing style (English text revision and proofreading); formatting assistance (LaTeX tables and bibliography); and assistance with code writing for running the experiments. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Zhao, Z. Tan, P. Ma, D. Li, B. Jiang, Y. Wang, Y. Yang, huan liu, Is chain-of-thought reasoning of LLMs a mirage? a data distribution lens, in: First Workshop on Foundations of Reasoning in Language Models, 2025. URL: https://openreview.net/forum?id=o2AoLPIjle.

[2] Y. Neuhaus, C. Tanchotikul, N. Flammarion, M. Hein, F. Croce, In- and out-of-distribution generalization of reasoning in multimodal LLMs for simple visual planning tasks, in: ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, 2026. URL: https://openreview.net/forum?id=LAzMmAeHyJ.

[3] P. Shojaee, I. Mirzadeh, M. Horton, S. Bengio, M. Farajtabar, et al., The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, Advances in Neural Information Processing Systems 38 (2026) 108018–108059.

[4] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Basic Books, 2018.

[5] J. Gordon, B. Van Durme, Reporting bias and knowledge acquisition, in: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 25–30. doi:10.1145/2509558.2509563.

[6] Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, Y. Kim, Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1819–1862. doi:10.18653/v1/2024.naacl-long.102.

[7] F. Hoppe, F. Ilievski, J.-C. Kalo, Investigating the robustness of deductive reasoning with large language models 413 (2025) 1776–1783. doi:10.3233/FAIA251007.

[8] J. Li, G. Li, Y. Chang, Y. Wu, Don't take the premise for granted: Evaluating the premise critique ability of large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 836–869. doi:10.18653/v1/2025.findings-emnlp.44.

[9] D. Akiba, The illusion of causality in llms: A developmentally grounded analysis of semantic scaffolding and benchmark–capability mismatches, Machine Learning and Knowledge Extraction 8 (2026). doi:10.3390/make8030057.

[10] P. Mondorf, B. Plank, Comparing inferential strategies of humans and large language models in deductive reasoning, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 9370–9402. doi:10.18653/v1/2024.acl-long.508.

[11] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, Dissociating language and thought in large language models, Trends in Cognitive Sciences 28 (2024) 517–540. doi:https://doi.org/10.1016/j.tics.2024.01.011.

[12] K. Yamin, S. Gupta, G. R. Ghosal, Z. C. Lipton, B. Wilder, Failure modes of llms for causal reasoning on narratives, in: ICML 2025 Workshop on Scaling Up Intervention Models, 2025. URL: https://openreview.net/forum?id=00zPdjUZY3.

[13] E. Kıcıman, R. Ness, A. Sharma, C. Tan, Causal reasoning and large language models: Opening a new frontier for causality, Transactions on machine learning research (2024). URL: https://openreview.net/forum?id=mqoxLkX210.

[14] S. Chen, B. Peng, M. Chen, R. Wang, M. Xu, X. Zeng, R. Zhao, S. Zhao, Y. Qiao, C. Lu, Causal evaluation of language models, ArXiv abs/2405.00622 (2024).

[15] G. Wan, Y. Lu, Y. Wu, M. Hu, S. Li, Large language models for causal discovery: Current landscape and future directions, in: J. Kwok (Ed.), Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 10687–10695. doi:10.24963/ijcai.2025/1186.

[16] A. Wu, K. Kuang, M. Zhu, Y. Wang, Y. Zheng, K. Han, B. Li, G. Chen, F. Wu, K. Zhang, Causality

for large language models, arXiv preprint arXiv:2410.15319 (2024).
[17] Z. Wang, CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models, in: K.-F. Wong, M. Zhang, R. Xu, J. Li, Z. Wei, L. Gui, B. Liang, R. Zhao (Eds.), Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 143–151.
[18] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks, J. Mach. Learn. Res. 17 (2016) 1103–1204.
[19] S. Cui, Z. Jin, B. Schölkopf, B. Faltings, The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 16722–16763. doi:10.18653/v1/2024.emnlp-main.932.
[20] L. Du, X. Ding, K. Xiong, T. Liu, B. Qin, e-CARE: a new dataset for exploring explainable causal reasoning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, Dublin, Ireland, 2022, pp. 432–446. doi:10.18653/v1/2022.acl-long.33.
[21] A. Srivastava, A. Rastogi, A. Rao, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, Transactions on Machine Learning Research (2023).
[22] M. Miliani, S. Auriemma, A. Bondielli, E. Chersoni, L. Passaro, I. Sucameli, A. Lenci, ExpliCa: Evaluating explicit causal reasoning in large language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 17335–17355. URL: https://aclanthology.org/2025.findings-acl.891/. doi:10.18653/v1/2025.findings-acl.891.
[23] A. R. Joishy, I. B. Balappanawar, V. K. Bonagiri, M. Gaur, K. Thirunarayan, P. Kumaraguru, Flying pigs, far and beyond: Evaluating llm reasoning in counterfactual worlds, 2026. URL: https://arxiv.org/abs/2505.22318. arXiv:2505.22318.
[24] J. Gao, X. Ding, B. Qin, T. Liu, Is ChatGPT a good causal reasoner? a comprehensive evaluation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 11111–11126. doi:10.18653/v1/2023.findings-emnlp.743.
[25] K. Yamin, S. Gupta, G. R. Ghosal, Z. C. Lipton, B. Wilder, LLMs struggle to balance reasoning and world knowledge in causal narrative understanding, in: The Fourteenth International Conference on Learning Representations, 2026.
[26] P. Mondorf, B. Plank, Beyond accuracy: evaluating the reasoning behavior of large language models–a survey, arXiv preprint arXiv:2404.01869 (2024).
[27] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in neural information processing systems 36 (2023) 46595–46623.
[28] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al., A survey on llm-as-a-judge, The Innovation 7 (2026).
[29] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, et al., Justice or prejudice? quantifying biases in llm-as-a-judge, in: International Conference on Learning Representations, volume 2025, 2025, pp. 102351–102390.
[30] L. Shi, C. Ma, W. Liang, X. Diao, W. Ma, S. Vosoughi, Judging the judges: A systematic study of position bias in llm-as-a-judge, in: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 2025, pp. 292–314.
[31] Q. Team, Qwen3.5: Accelerating productivity with native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.
[32] D. Lewis, Causation, The journal of philosophy 70 (1973) 556–567.
[33] D. Scalena, L. Zotos, E. Fersini, M. Nissim, A. Üstün, Eager: Entropy-aware generation for adaptive inference-time scaling, arXiv preprint arXiv:2510.11170 (2025).