

Can LLMs Reliably Identify Causal Conclusions in Scientific Literature? A Multidisciplinary Evaluation

Young Soo Ko, Nadia Sharief, Xuan Lu

University of Arizona

{youngsooko, nadiasharief, luxuan}@arizona.edu

Abstract

Large language models (LLMs) hold promise for automating causal knowledge extraction, a fundamental component of scientific discovery, yet their ability to reliably identify causal claims in scientific literature remains poorly understood. Using 1,000 manually annotated abstracts spanning 20 disciplines, we evaluate eight frontier LLMs under four prompting strategies on a causal conclusion identification task. More elaborate prompting strategies do not necessarily improve performance, while open-weight models achieve competitive performance relative to proprietary models. LLMs systematically produce false-positive causal predictions on non-causal abstracts, with GPT-5 showing the most extreme bias, and Few-Shot prompting provides the most consistent calibration benefit. Performance varies substantially across disciplines with stronger and more stable results in biomedical fields. Our findings provide practical insights for using LLMs in scientific causal analysis and highlight calibration and domain variation as key challenges for advancing causal reasoning in scholarly text.

1 Introduction

Causal knowledge underpins scientific understanding, yet manually extracting causal claims—especially the *causal conclusions* that authors draw from their own findings and that inform systematic reviews, new hypotheses, and policy decisions—is infeasible given the exponential growth of the literature (Girju, 2003; Do et al., 2011; Jonnalagadda et al., 2015; Higgins et al., 2019). Causal language in scientific contexts is often hedged, implicit, or discipline-specific (Hyland, 1996; Yu et al., 2019), making it difficult to distinguish truly causal conclusions from descriptive, correlational, or mechanistic ones. Large Language Models (LLMs) offer a scalable alternative (Gilardi et al., 2023), but existing evaluations of LLM-based causal extraction often rely on limited domains, narrow model

coverage, or synthetic examples (Lee et al., 2025; Miliani et al., 2025; Lithgow-Serrano et al., 2025), leaving unclear how reliably frontier LLMs handle real scholarly prose across disciplines and model families.

In this work, we present an extensive evaluation of state-of-the-art LLMs for identifying causal conclusions in scientific literature, providing practical insights for scientists seeking to use LLMs for scientific discovery and for AI researchers developing causal reasoning systems. We define causal conclusions as cause-and-effect judgments that authors assert on the basis of their study’s evidence. Such conclusions are especially important because they often shape subsequent citation, replication, evidence synthesis, and policy uptake (Haber et al., 2022; Cofield et al., 2010). Automatically determining whether a scientific article reports a causal conclusion is therefore a critical first step toward scalable evidence synthesis (Jonnalagadda et al., 2015; Higgins et al., 2019). For finer-grained tasks, such as disentangling causes and effects at scale, reliable conclusion-level identification can further reduce the search space and help prioritize relevant documents. We focus on article abstracts because they are widely accessible, structured summaries of scientific contributions (Dernoncourt and Lee, 2017), and because their rhetorical context—including hedging, study-design cues, and author stance—can provide important evidence for judging whether a conclusion is genuinely causal (Hyland, 1998; Yu et al., 2019).

To capture the spectrum of causal expressions across scientific fields, we selected 20 disciplines spanning biomedicine, physical and materials sciences, and ecology and environmental science. We curated a dataset of 1,000 abstracts from the AI4Science corpus (Xie et al., 2024) and manually annotated each abstract for its main conclusion and whether that conclusion was causal. We then use this dataset to evaluate eight LLMs, including

six proprietary and two open-weight models, under four prompting strategies for causal conclusion identification. To understand where errors arise, we further decompose the end-to-end task into two complementary settings: conclusion extraction and classification given human-annotated conclusions. In each setting, we compare LLM performance with non-LLM baselines. Finally, we conduct a discipline-level analysis to examine cross-field heterogeneity.

To the best of our knowledge, this is the first work to systematically evaluate LLMs for causal conclusion identification in scientific literature using human-annotated abstracts spanning multiple disciplines.¹ Our findings provide a comprehensive view of current LLM capabilities for this task, offering practical insights into model and prompting choices while identifying calibration and domain variation as key remaining challenges.

2 Related Work

2.1 Causal Language in Scientific Text

Causal claims in scientific writing are not binary but exist along a continuum of commitment—from strong assertions (*X causes Y*) through hedged implications (*X may contribute to Y*) to purely correlational reports (*X is associated with Y*) (Yu et al., 2019, 2020). Models that ignore this gradedness systematically conflate correlation with causation, particularly in observational or exploratory studies. The continuum is further complicated by genre conventions: scientific authors intentionally avoid categorical causal assertions, relying instead on hedging devices and rhetorical structuring to modulate the strength of their claims (Hyland, 1996). These conventions vary markedly across disciplines—biomedical literature employs overt causal markers such as *induces* or *inhibits*, whereas physical-science fields more frequently encode causality through mechanistic descriptions or quantitative relationships without explicit causal vocabulary (Hernán and Robins, 2010). These findings motivate our study design, which adopts a binary annotation scheme that can be applied consistently across disciplines while facilitating in-depth analysis of boundary cases across evaluation settings.

Annotation studies confirm that this linguistic

complexity translates directly into labeling difficulty. Annotated biomedical corpora reveal that scientific causal claims are often mechanistic, probabilistic, and span multiple clauses; even with detailed guidelines, inter-annotator agreement remains moderate, reflecting genuine ambiguity in scientific causal reasoning rather than annotation noise (Mihăilă et al., 2013). Our evaluation framework examines whether LLMs can navigate the same distinctions that pose challenges for human annotators.

2.2 LLM-based Causal Extraction

LLMs have demonstrated strong performance on explicit text classification tasks and high internal consistency as annotators (Gilardi et al., 2023), making them natural candidates for scalable causal extraction. Prompting strategies such as Chain-of-Thought (CoT), which elicit step-by-step reasoning before a final judgment, have been shown to improve performance on tasks requiring multi-step inference (Wei et al., 2022). However, fundamental questions remain about whether LLMs perform genuine causal reasoning or merely exploit surface-level co-occurrence patterns.

Critical analyses suggest that LLMs function as “causal parrots,” relying on distributional regularities memorized during pre-training rather than performing true causal inference (Zečević et al., 2023; Wu et al., 2024; Garcia-Corral et al., 2025). This limitation manifests concretely in evaluation: on the ExpliCa benchmark, state-of-the-art LLMs frequently confound temporal relations with causal ones and exhibit a lower threshold for establishing causality than humans (Milianni et al., 2025). In domain-specific settings, the picture is equally concerning—evaluations on biomedical texts show that LLMs approach random-guessing accuracy when classifying causal relation types, with instruction-tuned models displaying severe miscalibration and overconfidence in incorrect predictions (Lithgow-Serrano et al., 2025).

Despite these known vulnerabilities, existing benchmarks have evaluated LLM causal reasoning in isolated domains such as clinical guidelines or economics (Lee et al., 2025), and have focused exclusively on proprietary models, leaving open questions about generalizability across diverse scientific disciplines and across the proprietary–open divide. Our work addresses these gaps by systematically evaluating eight LLMs—including two frontier open-weight models—across 20 disciplines,

¹The annotated data, prompt templates, raw predictions, and analysis scripts are available at: <https://github.com/kosue153/Identify-Causal-Conclusion>

alongside non-LLM baselines that contextualize the contribution of LLMs.

2.3 Non-LLM Approaches to Scientific Information Extraction

Prior to LLMs, scientific information extraction relied on supervised classifiers trained on domain-specific corpora. Yu et al. (2019) trained traditional classifiers on PubMed abstracts to distinguish causal from correlational claims, achieving F_1 around 0.75 with hand-engineered features. Beltagy et al. (2019) demonstrated that domain-adapted BERT variants substantially outperform general-domain models on scientific tasks. More recent work explores hybrid approaches combining fine-tuned encoders with LLM-based reasoning (Gkountouras et al., 2025), yet direct comparisons between fine-tuned encoders and frontier LLMs on causal classification remain rare. Our work addresses this gap by systematically benchmarking SciBERT against eight frontier LLMs.

3 Dataset

We construct a benchmark from abstracts spanning 20 disciplines drawn from the AI4Science corpus (Xie et al., 2024), a large-scale collection of scientific papers aggregated from major publication venues across the natural and biomedical sciences. The disciplines are selected based on their volumes in the corpus, excluding computer science and AI-related fields. Discipline labels are inherited from the corpus and were generated through data-driven clustering rather than a predefined taxonomy, enabling finer-grained analysis of variation in causal expression across fields that may appear closely related but are semantically distinct. Further details are provided in Appendix A. We randomly sample 50 abstracts per discipline, yielding 1,000 abstracts in total. Abstracts are used instead of full papers to control document length while preserving the core claims.

Each abstract is annotated by two authors. During annotation, each annotator determines whether the abstract contains a causal conclusion—a statement indicating that one phenomenon *contributes to*, *enables*, *produces*, or *influences* another—and provides a summary of the main conclusion. Both explicit and implicit causal expressions are included. Pre-adjudication Cohen’s κ for the causal/non-causal classification is 0.638, indicating substantial inter-annotator agreement. Disagree-

ments are resolved through discussion: annotators reach consensus on causal labels and meaningfully different extracted conclusions, while paraphrastic differences are resolved by selecting the clearer formulation. Ambiguous cases are retained when supported by a reasonable interpretation, considering that causal language in scientific writing often exists on a continuum rather than fitting into a strict binary distinction (Mihăilă et al., 2013). In the final dataset, 74.0% of abstracts ($n=740$) are labeled *causal*, and 26.0% ($n=260$) are labeled *non-causal*. A full per-discipline breakdown is provided in Appendix A.

4 Causal Conclusion Identification

The task for the LLMs includes two steps. First, the model extracts the concluding claim(s) from a given abstract. Second, it classifies whether the extracted conclusion is causal or non-causal (Yes/No). A robustness check (Section 4.5) provides results for direct causality classification of abstracts. We evaluate eight LLMs on our *end-to-end* pipeline: GPT-4o, GPT-4o mini, and GPT-5 from OpenAI; Claude Sonnet 4.5 and Claude Opus 4.6 from Anthropic; Gemini 3.1 Pro from Google; and two frontier open-weight models—Llama 3.3 70B Instruct from Meta and Qwen 3.5 397B A17B from Alibaba. All models were accessed via APIs with default decoding parameters, and no fine-tuning was performed. Implementation details are described in Appendix B.3. For brevity, we omit version details in subsequent discussion when the intended model is clear from context.

We further conduct diagnostic evaluations for each step. The *extraction-only* evaluation compares model-extracted conclusions with human-generated conclusions, while the *classification-only* evaluation uses human-generated conclusions as input to isolate causality classification performance. We also include non-LLM baselines for both tasks.

Each LLM model is paired with four prompting strategies. The *Basic* prompt provides concise instructions without exemplars or scaffolding; the *Few-Shot (FS)* prompt provides two causal and two non-causal labeled examples drawn from a held-out set; the *Chain-of-Thought (CoT)* prompt asks for step-by-step reasoning before a final judgment; and the *Self-Refinement (SR)* prompt asks the model to verify and optionally revise its initial answer. All four strategies share an identical

Table 1: **Weighted F_1 by model and prompting strategy for the end-to-end pipeline.** The best result for each model is underlined. (*) denotes open-weight models.

Model	Basic	FS	CoT	SR
Gemini 3.1 Pro	<u>0.853</u>	0.828	0.843	0.845
Claude Opus 4.6	0.834	0.836	0.835	<u>0.846</u>
Llama 3.3 70B (*)	<u>0.831</u>	0.825	0.830	0.813
Qwen 3.5 397B (*)	0.830	<u>0.836</u>	0.825	0.776
Claude Sonnet 4.5	0.820	0.822	0.801	<u>0.823</u>
GPT-4o	0.810	0.782	<u>0.827</u>	0.811
GPT-5	<u>0.805</u>	0.803	<u>0.798</u>	0.791
GPT-4o mini	0.799	0.793	<u>0.799</u>	0.711

base prompt structure—definitions, instructions, and output schema—differing only in their strategy-specific additions, ensuring fair comparison. This design yields 32 primary configurations (eight models $\times$ four strategies). The prompt templates are provided in Appendix D.

4.1 Overall Model Performance

We use *weighted F_1* as the primary evaluation metric because it accounts for the imbalanced causal/non-causal label distribution while remaining sensitive to per-class precision and recall. Full results, including complementary metrics *accuracy* and *macro F_1* as well as bootstrap estimates, are provided in Appendix B.

As shown in Table 1, among the Basic configurations, Gemini achieves the highest performance with a weighted F_1 of 0.853, followed by Claude Opus (0.834), Llama (0.831), and Qwen (0.830). Notably, Claude Opus performs best under SR, achieving a weighted F_1 score of 0.846. The next tier comprises Claude Sonnet (0.820) and GPT-4o (0.810), whose best configurations reach 0.823 (SR) and 0.827 (CoT), respectively. The remaining GPT models show comparatively lower performance, with GPT-5 reaching 0.805 and GPT-4o mini reaching 0.799.

Two main observations emerge from the results. First, non-basic prompting strategies do not necessarily improve performance on causality identification. The Basic prompt achieves the best performance for three of the eight models (Gemini, Llama, and GPT-5), and ties with CoT as the best for GPT-4o mini. Among models where alternative prompts help, the type of effective prompt varies as well as the largest gain: Claude Opus improves from 0.834 (Basic) to 0.846 (SR), GPT-4o improves from 0.810 (Basic) to 0.827 (CoT), and

Qwen improves from 0.830 (Basic) to 0.836 (FS). This finding suggests that more elaborate prompting does not consistently improve causality identification and may sometimes introduce additional reasoning or revision steps that do not translate into better final judgments.

Second, the open-weight models, Qwen and Llama, achieve competitive performance relative to proprietary models. Under Basic prompting, Llama ranks third and Qwen fourth; under Few-Shot prompting, Qwen ties Claude Opus for the best result across all models, with a weighted F_1 of 0.836. These results suggest that frontier open-weight models are comparable to proprietary models in identifying causal conclusions from scientific text. More broadly, the task remains equally challenging across model families rather than being solved by access to proprietary models alone.

This difficulty is further underscored by comparison with standard causality benchmarks and general-domain datasets with more explicit causal language, such as SemEval-2010 Task 8 (Hendrickx et al., 2010) and sentence classification in FinCausal (Mariko et al., 2020), where modern LLMs and fine-tuned models often achieve F_1 scores above 0.90 or 0.95. In contrast, our results suggest that identifying causal conclusions from scientific literature is substantially harder. This challenge is also reflected in recent domain-specific benchmarks, which show that LLMs struggle in specialized settings; for example, recent studies report GPT-4 achieving F_1 around 0.60 on clinical practice guidelines (Gopalakrishnan et al., 2025) and an average of only 0.535 on the ReCITE benchmark for real-world academic text (Saklad et al., 2025).

4.2 Class-level Performance

To better understand this challenge, we examine model performance by class using per-class F_1 scores, which assess performance on causal and non-causal conclusions separately. We further report the false-positive rate, defined as the proportion of truly non-causal conclusions incorrectly classified as causal, and the missed rate, defined as the proportion of truly causal conclusions incorrectly classified as non-causal. Table 2 reports these metrics for each model under the Basic prompt.

The F_1 score on truly causal conclusions ranges from 0.867 to 0.901, whereas that on truly non-causal conclusions ranges from 0.581 to 0.715, with gaps between the two metrics ranging from

0.186 for Gemini to 0.303 for GPT-5. These results indicate a systematic asymmetry in model performance: LLMs identify causal conclusions relatively well but struggle to distinguish non-causal conclusions from causal ones. The false-positive rate is also high, ranging from 29.2% for Gemini (95% CI [23.7%, 34.5%]) to 51.5% for GPT-5 (95% CI [45.8%, 57.5%]). Per-metric bootstrap CIs (Appendix Table B.3) provide further evidence that cross-tier model differences are statistically robust.

Several factors may contribute to this tendency. First, causal language is common in scientific writing, and models may learn to associate conclusion-oriented abstracts with causal interpretation even when the evidence is correlational or descriptive (see case study for GPT-5 in Section 4.4.1). This pattern is consistent with the “causal parrots” hypothesis (Zečević et al., 2023), which observes that LLMs may reproduce the rhetorical patterns of causal language present in their training data. Second, instruction tuning and RLHF may encourage models to provide decisive answers, making them less willing to reject causality under ambiguity. Third, scientific rhetoric often frames findings as meaningful, explanatory, or actionable, which can blur the boundary between causal claims and non-causal conclusions (Hyland, 1996; Yu et al., 2019). These factors suggest that improving causal extraction from scientific text requires not only stronger models but also better calibration: models must learn when to withhold causal judgments rather than merely identify plausible causal language. We therefore next examine whether prompting strategies can mitigate this false-positive tendency by shifting the balance between causal detection and non-causal rejection.

4.3 Prompting Effects on False-Positive Causal Predictions

Few-Shot Prompting. Although Few-Shot prompting does not improve aggregate weighted F_1 , it has a consistent calibration effect by reducing the systematic false-positive causal predictions identified above. Table 2 reports the change in the false-positive rate from Basic to Few-Shot. Across all eight models, Few-Shot examples reduce false positives by an average of 13.3 percentage points, with Claude Sonnet showing the largest reduction (−28.0 pp). Six of the eight models also improve their non-causal F_1 under Few-Shot prompting (mean $\Delta = +0.026$, Table B.1), suggesting that

the reduction in false positives is not merely an artifact of indiscriminate label flipping. Llama exhibits the smallest calibration effect (−1.5 pp in the false-positive rate), likely because it already shows a relatively balanced classification under Basic.

The robustness of this Few-Shot calibration effect across both proprietary and open-source models suggests that in-context examples can help calibrate causal judgments across different training pipelines. This effect is largely neutral in aggregate weighted F_1 (mean $\Delta = -0.007$), because the gains in non-causal F_1 are offset by small reductions in F_1 on truly causal conclusions. Indeed, three models (Claude Opus, Claude Sonnet, and Qwen) show small gains in weighted F_1 under Few-Shot, while the remaining models show modest losses. Few-Shot prompting therefore appears to function primarily as a bias-correction strategy, producing more class-balanced causal classifications. This is particularly valuable for applications where false-positive causal assertions have high downstream cost, including causal knowledge base construction, systematic review support, and scientific claim verification.

Chain-of-Thought Prompting. CoT improves weighted F_1 for only two of the eight models (Claude Opus and GPT-4o), with marginal gains of +0.002 and +0.017 respectively, but its effects are uneven. It notably degrades Claude Sonnet (−0.019 in weighted F_1 , Table 1), mainly through over-correction: its missed rate triples from 0.077 under Basic to 0.236 under CoT (see Table B.1). This suggests that step-by-step reasoning can sometimes make models more conservative, but this conservatism does not always translate into better calibration. Instead, CoT may shift the error profile from false-positive to false-negative causal predictions, depending on the model.

Self-Refinement Prompting. As shown in Table 2, SR effects are even more model-dependent. Qwen exhibits the largest SR benefit, reducing false positives by 32.7 percentage points, more than twice the next-largest reduction. By contrast, GPT-5 is the only model for which SR increases false positives (+8.1 pp), suggesting that self-verification can sometimes reinforce rather than correct initial causal judgments, a pattern related to confirmation drift (Huang et al., 2024; Madaan et al., 2023). GPT-4o mini flips 40.1% of predictions but achieves no net improvement (0.0 pp), in-

Table 2: **Class-level asymmetry and prompting effects on false-positive causal predictions in the end-to-end pipeline.** F_{1C} and F_{1NC} denote per-class F_1 under the Basic prompt. FPR, Miss, and Flip are reported as percentages. Changes under Few-Shot (FS), Chain-of-Thought (CoT), and Self-Refine (SR) are relative to Basic. $\Delta F_{1,w}$ denotes the absolute change in weighted F_1 , and ΔFPR is reported in percentage points. Flip denotes the percentage of first-pass predictions reversed by SR.

Model	Basic				FS		CoT		SR		
	F_{1C}	F_{1NC}	FPR	Miss	$\Delta F_{1,w}$	ΔFPR	$\Delta F_{1,w}$	ΔFPR	$\Delta F_{1,w}$	ΔFPR	Flip
Gemini 3.1 Pro	0.901	0.715	29.2	9.6	-0.025	-9.2	-0.010	-0.8	-0.008	-7.7	6.2
GPT-4o	0.867	0.647	31.9	14.9	-0.028	-13.1	+0.017	-12.3	+0.001	-13.8	15.9
Claude Opus 4.6	0.895	0.665	39.9	7.4	+0.002	-12.4	+0.001	-5.4	+0.012	-4.3	5.8
Llama 3.3 70B	0.892	0.655	40.0	8.1	-0.006	-1.5	-0.001	0.0	-0.018	-13.5	17.8
Qwen 3.5 397B	0.895	0.646	43.1	6.8	+0.006	-16.2	-0.005	-2.7	-0.054	-32.7	27.4
Claude Sonnet 4.5	0.888	0.625	44.5	7.7	+0.002	-28.0	-0.019	-30.7	+0.003	-14.1	17.1
GPT-4o mini	0.867	0.607	40.8	12.6	-0.006	-11.2	0.000	-1.5	-0.088	0.0	40.1
GPT-5	0.884	0.581	51.5	6.5	-0.002	-15.0	-0.007	+1.9	-0.014	+8.1	7.5

Table 3: **Classification-only performance: LLMs under the Basic prompt vs. non-LLM baselines.** Methods are ranked by weighted $F_1(F_{1,w})$. F_{1NC} denotes non-causal F_1 , and FPR denotes false-positive rate.

Model	$F_{1,w}$	F_{1NC}	FPR
SciBERT (fine-tuned)	0.842	0.682	0.358
Gemini 3.1 Pro	0.820	0.665	0.300
Llama 3.3 70B	0.816	0.658	0.308
Claude Opus 4.6	0.815	0.673	0.265
Qwen 3.5 397B	0.806	0.659	0.254
GPT-4o	0.800	0.667	0.192
Claude Sonnet 4.5	0.799	0.654	0.239
GPT-5	0.799	0.611	0.396
GPT-4o mini	0.749	0.630	0.127
TF-IDF + LogReg	0.778	0.571	0.427
TF-IDF + SVM	0.775	0.552	0.469
TF-IDF + RandomForest	0.734	0.365	0.750

dicating that frequent revision alone is insufficient without reliable second-pass judgment. Together, these patterns suggest that self-refinement depends not only on whether a model revises its answer, but also on whether its second-pass judgment is better calibrated than its initial response.

4.4 Pipeline Decomposition and Non-LLM Baselines

The end-to-end task combines two distinct steps: extracting the concluding claim(s) and classifying whether the extracted conclusion is causal. To better understand where errors arise, we conduct two complementary analyses. First, we decompose end-to-end errors into extraction-associated and classification-associated errors. Second, we compare LLMs with non-LLM baselines for each pipeline stage.

Extraction quality is measured using SBERT

Table 4: **Error decomposition under the Basic strategy.** All values are percentages. Full = correct extraction and classification; Cls. = correct extraction but incorrect classification; Lucky = imperfect extraction but correct classification; Ext. = imperfect extraction and incorrect classification; Share = share of end-to-end errors associated with imperfect extraction.

Model	Full	Cls.	Lucky	Ext.	Share
GPT-5	64.4	13.4	17.4	4.8	26.4
Gemini 3.1 Pro	64.0	10.8	21.3	3.9	26.5
Qwen 3.5 397B	66.9	11.7	16.9	4.5	27.8
Claude Sonnet 4.5	63.1	13.1	17.8	6.0	31.4
GPT-4o	58.7	13.2	22.0	6.1	31.6
Llama 3.3 70B	61.5	11.1	22.1	5.3	32.3
Claude Opus 4.6	66.5	12.1	15.3	6.1	33.5
GPT-4o mini	54.8	11.7	25.3	8.2	41.2

cosine similarity (Reimers and Gurevych, 2019) between the model-extracted conclusion and the human-annotated conclusion. A human evaluation shows a moderate and significant correlation between SBERT cosine similarity and human judgments (see Appendix B.4), supporting its use as a scalable evaluation signal, although it does not necessarily capture causal-content preservation. Across all models, using a threshold of 0.7 to identify high-similarity extractions, 26–41% of end-to-end errors are associated with imperfect extraction, while 59–74% occur despite high-similarity extractions. This indicates that end-to-end errors arise from both the extraction and classification stages, but classification accounts for the larger share under this decomposition (Table 4).

We also observe a complementary “lucky recovery” pattern: in 15–25% of cases, models produce imperfect extractions but still assign the correct end-to-end label. This suggests that exact extrac-

tion fidelity is not always required for correct causal classification, likely because imperfect extractions may still preserve enough semantic information for the model to infer the correct label. Notably, providing models with human-annotated conclusions does *not* consistently improve aggregate performance over the end-to-end setting; weighted F_1 decreases by 0.022 on average across the eight LLMs. This counterintuitive pattern suggests that LLMs adapt their classification to their own extraction style, and that semantic equivalence between extracted and gold conclusions matters more than verbatim accuracy. Per-model human-annotated-conclusion results are reported in Appendix Table B.2.

We further contextualize each pipeline stage using non-LLM baselines (Table 3). For conclusion extraction, rule-based heuristics substantially underperform LLM extraction, with the best heuristic reaching 0.686 mean SBERT cosine similarity compared with 0.782 for the best LLM (Qwen; see Appendix B.5). For causal classification, however, the picture is more nuanced: fine-tuned SciBERT (Beltagy et al., 2019) achieves the highest aggregate weighted F_1 of 0.842, surpassing all eight LLMs by 2.2 to 9.3 percentage points. The strongest LLM, Gemini, reaches 0.820, while GPT-4o mini ranks lowest at 0.749. Classical TF-IDF classifiers achieve 0.734–0.778, with Logistic Regression (0.778) and SVM (0.775) even outperforming GPT-4o mini (0.749). Implementation details are provided in Appendix B.6.

The relationship between aggregate performance and calibration reveals a fundamental trade-off. While SciBERT achieves the highest aggregate F_1 , its false-positive rate (0.358) is *higher* than that of seven of the eight LLMs. Classical TF-IDF classifiers exhibit the highest false-positive rates (0.427–0.750). This pattern indicates that no single approach dominates on both axes: supervised fine-tuning on encoder models can improve aggregate accuracy, while specific instruction-tuned LLMs offer better calibration on the minority non-causal class. Thus, the choice between LLMs and fine-tuned encoders depends critically on which objective—aggregate F_1 or class-wise calibration—is prioritized for the downstream application.

4.4.1 Case Study: Persistent False Positives in GPT-5

To better understand the false-positives produced by LLMs, we select GPT-5 for a case study through pipeline decomposition. We separate false posi-

tives into extraction-induced (errors resolved when the human-annotated conclusion is provided) and classification-inherent (errors that persist), and apply this analysis to the 117 non-causal abstracts that GPT-5 consistently misclassifies as causal under both Basic and CoT prompting. In 43.6% of these persistent false positives, the error disappears when human-annotated conclusions are provided, confirming that the model-extracted conclusions introduced a causal framing not present in the original gold annotations (extraction-induced). The remaining 56.4% persist even with human-annotated conclusions, often involving transformation, capability, or mechanistic language that invites causal interpretation (classification-inherent). Few-Shot prompting corrects 29.1% of these persistent cases, suggesting that in-context examples mitigate some false positives but do not fully resolve ambiguous causal boundaries. See details in Appendix C.

4.5 Direct Classification without Extraction

To test whether our findings depend on the two-stage extraction-then-classification design, we conducted a direct classification experiment in which four models (GPT-5, Claude Opus 4.6, Gemini 3.1 Pro, and Qwen 3.5 397B) classified 200 randomly sampled abstracts as causal or non-causal without first extracting a conclusion. False-positive rates diverge more sharply than in the end-to-end pipeline, with GPT-5 remaining the most biased and Qwen the most calibrated. This model-dependent heterogeneity is consistent with our end-to-end results and reinforces the central message that false-positive causal predictions are a model-specific property of instruction-tuning rather than an artifact of the two-stage design. Full per-model results are reported in Appendix B.7.

5 Discipline-Specific Evaluation

We next examine whether model performance varies across scientific disciplines. Given that causal and non-causal samples are highly imbalanced in some disciplines (e.g., Organic Chemistry contains only one non-causal sample; see Table A.1), we interpret the results at the level of discipline clusters. Figure 1 shows substantial discipline-specific variation in weighted F_1 across model-prompt configurations. Biomedical disciplines, including Cancer Biology, Immunology, Developmental Biology, Oncology, Molecular Biology / Genetics, Cell Biology, and Neuroscience,

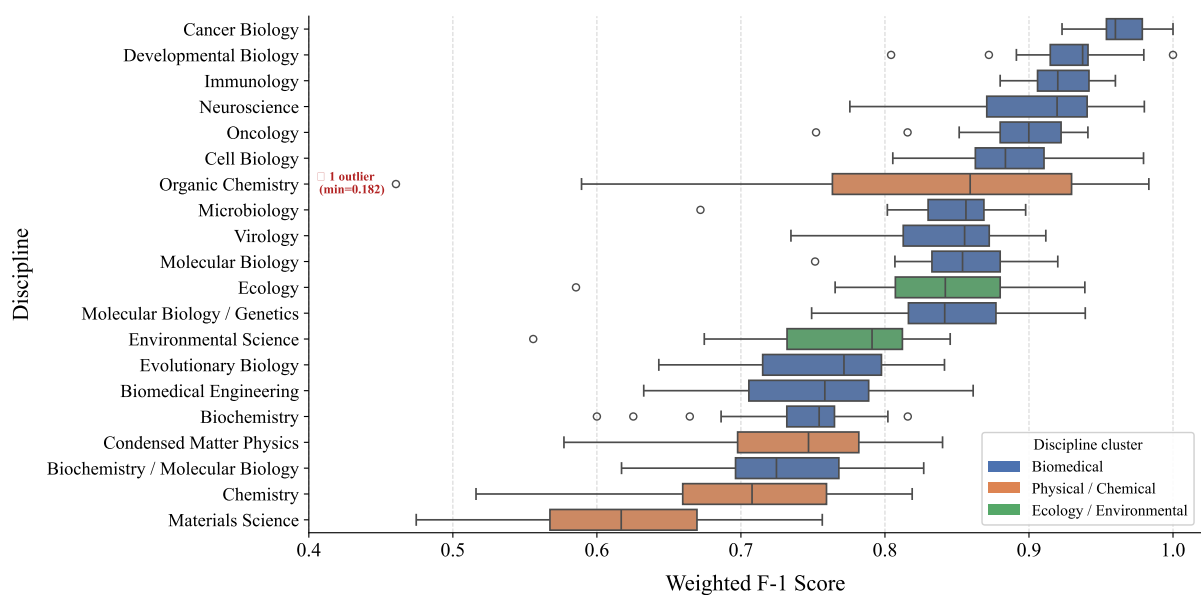


Figure 1: **Distribution of weighted F_1 scores across scientific disciplines.** Each box aggregates 32 model–prompt configurations (8 models $\times$ 4 prompting strategies).

achieve the highest median weighted F_1 scores, approaching or exceeding 0.9 across model–prompt configurations. These fields also show relatively compact interquartile ranges, suggesting robustness to model and prompt choice. By contrast, several physical-science disciplines, including Materials Science, Biochemistry/Molecular Biology, Chemistry, and Condensed Matter Physics, show lower median scores and wider variation, suggesting greater sensitivity to model and prompt choice.

To disentangle the sources of this variation, we compute causal recall, or *sensitivity*, and non-causal recall, or *specificity*, for each model–prompt configuration within each discipline. We then perform separate hierarchical clustering analyses for sensitivity and specificity using average linkage (UPGMA) with Euclidean distance. The full heatmaps are shown in Figures B.2 and B.3 in the Appendix.

The sensitivity heatmap reveals clear discipline-level stratification. Biomedical disciplines form a high-sensitivity cluster, indicating that causal conclusions in these fields are consistently recognized across models and prompting strategies. In contrast, several physical-science disciplines, including Organic Chemistry, Chemistry, Condensed Matter Physics, and Materials Science, form a low-sensitivity cluster. This pattern suggests that causal conclusions in these domains may be expressed less directly, often through mechanisms, material properties, experimental effects, or domain-specific

terminology, making them harder for prompt-based LLM extraction. Notably, open-weight models such as Llama and Qwen under Basic and CoT prompts cluster with stronger proprietary configurations, reinforcing the earlier finding that open-weight models can be competitive on this task.

The specificity heatmap reveals a different source of variation. Low-specificity configurations are concentrated among GPT-family models, consistent with their higher false-positive rates in Table 2. This produces an important inversion: some configurations with high sensitivity, such as GPT-5 under Basic, CoT, and SR prompts, show weaker specificity, whereas more conservative configurations, such as Qwen (SR) and Claude Sonnet (SR), achieve stronger specificity. In other words, the best models for detecting causal conclusions are not always the best models for rejecting non-causal ones.

Together, these patterns show that discipline-specific performance differences are not only about overall accuracy, but also about the balance between sensitivity and specificity. Biomedical disciplines appear easier because causal conclusions are more consistently detected, whereas physical-science disciplines expose sensitivity deficits. At the same time, GPT-family models exhibit stronger tendencies toward causal interpretation, which improves recall on causal cases but weakens rejection of non-causal conclusions. More specifically, biomedical disciplines achieve mean specificity of

0.669 across model–prompt configurations, compared with 0.588 for physical and chemical disciplines. The corresponding sensitivity gap is smaller, indicating that the biomedical advantage is driven primarily by more reliable rejection of non-causal conclusions rather than by better detection of causal ones. The optimal model–prompt configuration therefore depends on the downstream application’s tolerance for false positives versus false negatives. For *biomedical or cancer-related literature screening*, where missing causal findings during expert review can be costly, high-sensitivity configurations such as GPT-5 (Basic) or Claude Opus (SR) may be preferred. For *causal knowledge graph construction or systematic review support*, where incorrectly propagating non-causal findings as causal can be harmful, high-specificity configurations such as Claude Sonnet (FS) or Qwen (SR) may be more suitable.

6 Discussion

Calibration is a key source of task difficulty.

Causal conclusion identification remains challenging because LLMs struggle to calibrate causal judgments. All eight models systematically produce false-positive causal predictions, misclassifying non-causal conclusions more frequently than missing true causal ones. GPT-5 exemplifies this with the highest false-positive rate (51.5%) and lowest missed rate (6.5%). Conversely, smaller LLMs (GPT-4o mini: 0.127, GPT-4o: 0.192, Claude Sonnet 4.5: 0.239) achieve substantially lower false-positive rates on classification-only inputs, demonstrating calibration is independent of model scale. Furthermore, while fine-tuned SciBERT achieves the highest aggregate F_1 (0.842), it retains a false-positive rate (0.358) higher than that of seven LLMs. Thus, supervised fine-tuning improves aggregate accuracy but does not automatically resolve calibration. Progress requires targeted interventions like Few-Shot calibration, task-specific fine-tuning, or class-balanced objectives.

Causality in scientific text may be graded rather than binary. The systematic false-positive pattern suggests causal meaning in abstracts lies on a continuum, often describing mechanisms or associations that invite causal interpretation without being purely causal. While binary classification serves as a useful first-pass screening for large-scale analysis and systematic reviews (Higgins et al., 2019), future work must consider multi-

level causality schemas tailored to specific domains. Distinguishing explicit causal claims from indirect meanings is vital, as false-positive causal predictions propagate erroneous evidence into literature maps, causal knowledge graphs, and hypothesis-generation systems.

7 Conclusion

We introduced a manually annotated dataset of 1,000 scientific abstracts across 20 disciplines to evaluate eight frontier LLMs on causal conclusion identification, compared with non-LLM baselines on subtasks. Results show that all LLMs systematically produce false-positive causal predictions on non-causal abstracts, while also indicating a trade-off between aggregate accuracy and calibration: higher aggregate F_1 is often associated with higher false-positive rate. Performance also varies substantially across fields, underscoring the need for calibrated, domain-aware methods for scientific causal analysis.

8 Limitations

Scope and scale. Our evaluation is based on 1,000 English-language abstracts spanning 20 disciplines. Although this provides broad cross-disciplinary coverage, abstracts may not capture causal reasoning distributed across full paper sections (Teufel, 1999). The exclusion of social sciences and humanities also limits generalizability, especially because causal language and evidentiary norms may differ substantially across these fields.

Annotation framework. Our binary causal/non-causal scheme, which aims to support consistent evaluation across disciplines, does not capture the graded nature of scientific causal claims (Yu et al., 2019), representing an important direction for future work. Inter-annotator agreement ($\kappa = 0.638$) likely reflects genuine boundary ambiguity rather than annotation error alone (Mihăilă et al., 2013). Such ambiguity may introduce uncertainty into the gold-standard labels.

API stability and refusals. All LLMs were accessed via commercial or hosted APIs with default parameters. Because model versioning and internal updates are beyond our control, results may vary over time or across API versions. We also observed API-level safety refusals for Claude Opus 4.6 on approximately 2.6% of biomedical abstracts; these rows were excluded from metric computation. Other models exhibited refusal rates below 0.5%.

End-to-end supervised pipeline. The higher aggregate F_1 achieved by the fine-tuned SciBERT model for conclusion extraction suggests the potential of an end-to-end pipeline that combines a supervised conclusion extractor with a supervised classifier, providing a natural direction for future work.

9 Ethical Considerations

LLMs' systematic tendency to over-predict causality poses risks if their outputs are propagated into evidence syntheses, literature maps, causal knowledge graphs, or policy documents. Our results show that this risk persists even among newer frontier models, suggesting that improved aggregate capability alone may not ensure reliable causal calibration. LLM-based causal extraction should therefore be treated as a screening tool requiring human verification, rather than as a replacement for expert judgment, particularly in high-stakes settings such as systematic reviews and policy analysis.

References

- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.
- Stacey S Cofield, Rachel V Corona, and David B Allison. 2010. Use of causal language in observational studies of obesity and nutrition. *Obesity facts*, 3(6):353–356.
- Franck Dernoncourt and Ji Young Lee. 2017. [PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Quang Do, Yee Seng Chan, and Dan Roth. 2011. [Minimally supervised event causality identification](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- P. Garcia-Corral, H. D. Bechara, K. Manohara, and S. Jankin. 2025. Causally fooled in the name of being honest? evaluating causal extraction in llms for political text. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025)*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).
- Roxana Girju. 2003. [Automatic detection of causal relations for question answering](#). In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics.
- John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. 2025. Language agents meet causality—bridging LLMs and causal world models. In *International Conference on Learning Representations*, volume 2025, pages 37311–37336.
- Seethalakshmi Gopalakrishnan, Luciana Garbayo, and Wlodek Zadrozny. 2025. Causality extraction from medical text using large language models (LLMs). *Information*, 16.
- Noah A Haber, Sarah E Wieten, Julia M Rohrer, Onyebuchi A Arah, Peter W G Tennant, Elizabeth A Stuart, Eleanor J Murray, Sophie Pilleron, Sze Tung Lam, Emily Riederer, Sarah Jane Howcutt, Alison E Simmons, Clémence Leyrat, Philipp Schoenegger, Anna Booman, Mi-Suk Kang Dufour, Ashley L O'Donoghue, Rebekah Baglini, Stefanie Do, and 30 others. 2022. Causal and associational language in observational health research: a systematic evaluation. *American Journal of Epidemiology*, 191(12):2097–2107.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. [SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals](#). In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.
- Miguel A Hernán and James M Robins. 2010. *Causal inference*. CRC Boca Raton, FL.
- Julian P T Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*, 2 edition. John Wiley & Sons.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large language models cannot self-correct reasoning yet](#). *The Twelfth International Conference on Learning Representations*.
- Ken Hyland. 1996. Writing without conviction? hedging in science research articles. *Applied Linguistics*, 17(4):433–454.
- Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins Publishing.

- Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, 4(1):78.
- D. Lee, S. Park, Y. Hwang, H. Kim, H. Oh, J. Kim, and J. Kim. 2025. Benchmarking LLM causal reasoning with scientifically validated relationships. *arXiv preprint arXiv:2510.07231*.
- Oscar William Lithgow-Serrano, Vani Kanjirang, and Alessandro Antonucci. 2025. Causal understanding by LLMs: The role of uncertainty. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 208–228.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.
- Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. [The financial document causality detection shared task \(FinCausal 2020\)](#). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.
- Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14:2.
- Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia C. Passaro, Irene Sucameli, and Alessandro Lenci. 2025. ExpliCa: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics 2025*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Ryan Saklad, Aman Chadha, Oleg Pavlov, and Raha Moraffah. 2025. [Can large language models infer causal relationships from real-world text?](#) *Preprint*, arXiv:2505.18931.
- Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh Edinburgh, Scotland.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* 35.
- Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.
- Yutong Xie, Yijun Pan, Hua Xu, and Qiaozhu Mei. 2024. Bridging AI and science: implications from a large-scale literature analysis of ai4science. *arXiv preprint arXiv:2412.09628*.
- Bei Yu, Yingya Li, and Jun Wang. 2019. [Detecting causal language use in science findings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China. Association for Computational Linguistics.
- Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. [Measuring correlation-to-causation exaggeration in press releases](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

A Dataset Details

The dataset includes abstracts spanning 20 disciplines drawn from the AI4Science corpus (Xie et al., 2024). We ranked all disciplines represented in the corpus by their paper count and selected the 20 highest-volume fields to ensure sufficient coverage and sampling diversity. Computer science and AI fields were excluded because LLMs are predominantly pretrained on large quantities of text from these domains, which could inflate performance through memorization or distributional familiarity rather than genuine causal reasoning. Fifty abstracts were randomly sampled per discipline (1,000 in total). The labels are inherited verbatim from the AI4Science corpus and reflect data-driven semantic clusters rather than a curated taxonomy. Closely related labels are retained because they capture distinct rhetorical patterns: for example, *Cancer Biology* papers tend to emphasize molecular mechanisms of tumorigenesis, whereas *Oncology* papers more often report clinical outcomes

and treatment responses. Similar distinctions apply between *Molecular Biology* and *Biochemistry / Molecular Biology*.

Table A.1 reports the per-discipline distribution of causal and non-causal labels. Overall, 74.0 % ($n=740$) of the abstracts were labelled *causal* and 26.0 % ($n=260$) *non-causal*.

Table A.1: **Per-discipline distribution of causal and non-causal labels.** Fifty abstracts were sampled per discipline.

Discipline	Causal	Non-causal	Causal %
Biochemistry	32	18	64
Biochemistry / Molecular Biology	36	14	72
Biomedical Engineering	35	15	70
Cancer Biology	46	4	92
Cell Biology	41	9	82
Chemistry	26	24	52
Condensed Matter Physics	30	20	60
Developmental Biology	37	13	74
Ecology	40	10	80
Environmental Science	36	14	72
Evolutionary Biology	31	19	62
Immunology	43	7	86
Materials Science	25	25	50
Microbiology	38	12	76
Molecular Biology	43	7	86
Molecular Biology / Genetics	38	12	76
Neuroscience	31	19	62
Oncology	39	11	78
Organic Chemistry	49	1	98
Virology	44	6	88
Total	740	260	74.0

B Supplementary Classification Results

Table B.1 reports accuracy, per-class F_1 (Yes/No), macro F_1 , and weighted F_1 for all model-prompt combinations on the End-to-End benchmark, grouped by model. The best result among the four primary configurations (Basic, FS, CoT, SR) for each model is underlined.

Figure B.1 visualizes false-positive rates across all 32 model-prompt configurations. Consistent with the aggregate patterns reported in Section 4.2, GPT-family models exhibit the highest false-positive rates, while Few-Shot prompting consistently reduces false positives across most models.

B.1 Hierarchical Clustering Heatmaps

Figures B.2 and B.3 show the hierarchical clustering heatmaps for model sensitivity and specificity, respectively.

B.2 Bootstrap Confidence Intervals

To quantify uncertainty in our reported metrics, we computed 95% bootstrap confidence intervals using 1,000 resamples with replacement over the 1,000-abstract dataset. For each resample, we recomputed

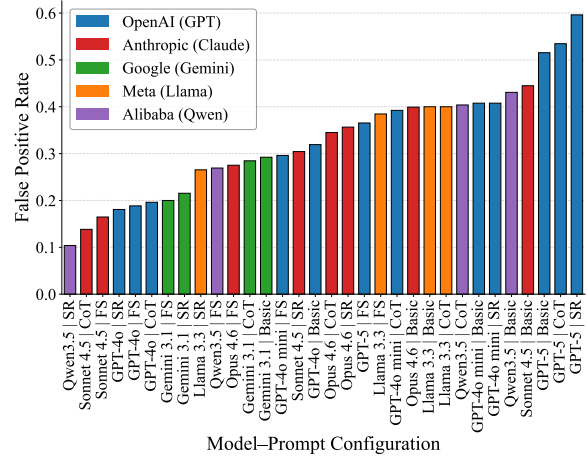


Figure B.1: **False-positive rates across all model-prompt configurations.** A false positive occurs when the model classifies a non-causal abstract as causal, reflecting a tendency to over-predict causality.

weighted F_1 , per-class F_1 , false-positive rate, and missed rate for every model-prompt configuration.

Aggregate findings are statistically robust: false-positive rates span 29.2% (Gemini, CI: [23.7%, 34.5%]) to 51.5% (GPT-5, CI: [45.8%, 57.5%]) with *non-overlapping* CIs, indicating that systematic false positives and cross-tier model differences are not artifacts of sample composition. Weighted F_1 differences between top-tier configurations are within overlapping CIs, consistent with the small point-estimate gaps reported in Table 1.

Per-discipline CIs are naturally wider due to smaller within-discipline sample sizes (50 abstracts per field, with 1–25 non-causal cases). We therefore treat discipline-level claims as trends supported by clustering rather than as precise per-field estimates. Disciplines with fewer than 10 non-causal samples are explicitly flagged in the corresponding analyses.

B.3 Reproducibility Details

Table B.4 lists exact model identifiers, API providers, and inference configurations for all evaluated LLMs. For OpenAI, Anthropic, and Google models, we used the API’s default decoding parameters. For Together AI models, we set temperature to 0.0 for reproducibility, with the following exception: Qwen 3.5 was called with temperature 0.7, top- p 0.8, and reasoning mode disabled (enable_thinking=False), following the vendor-recommended configuration for non-thinking inference. Maximum output tokens were 4,096 for Anthropic, 1,024 for Together AI (2,048 for

Table B.1: Full End-to-End classification results across all 8 models and 4 prompting strategies.

Model	Prompt	Accuracy	F ₁ (Yes)	F ₁ (No)	F ₁ (Macro)	F ₁ (Weighted)
GPT-4o	Basic	0.807	0.867	0.647	0.757	0.810
	FS	0.770	0.829	0.647	0.738	0.782
	CoT	0.820	0.872	0.699	0.785	0.827
	SR	0.802	0.856	0.683	0.769	0.811
GPT-4o mini	Basic	0.801	0.867	0.607	0.737	0.799
	FS	0.787	0.850	0.632	0.741	0.793
	CoT	0.800	0.865	0.612	0.739	0.799
	SR	0.698	0.783	0.505	0.644	0.711
GPT-5	Basic	0.818	0.884	0.581	0.732	0.805
	FS	0.802	0.865	0.625	0.745	0.803
	CoT	0.812	0.880	0.563	0.722	0.798
	SR	0.813	0.883	0.529	0.706	0.791
Claude Sonnet 4.5	Basic	0.827	0.888	0.625	0.756	0.820
	FS	0.814	0.865	0.700	0.782	0.822
	CoT	0.790	0.843	0.680	0.762	0.801
	SR	0.821	0.877	0.668	0.773	0.823
Claude Opus 4.6	Basic	0.840	0.895	0.665	0.780	0.834
	FS	0.834	0.885	0.698	0.792	0.836
	CoT	0.838	0.891	0.681	0.786	0.835
	SR	0.850	0.901	0.695	0.798	0.846
Gemini 3.1 Pro	Basic	0.853	0.901	0.715	0.808	0.853
	FS	0.821	0.873	0.699	0.786	0.828
	CoT	0.842	0.893	0.702	0.797	0.843
	SR	0.841	0.889	0.720	0.804	0.845
Llama 3.3 70B	Basic	0.836	0.892	0.655	0.774	0.831
	FS	0.828	0.886	0.650	0.768	0.825
	CoT	0.835	0.892	0.654	0.773	0.830
	SR	0.808	0.865	0.666	0.765	0.813
Qwen 3.5 397B	Basic	0.838	0.895	0.646	0.771	0.830
	FS	0.834	0.886	0.696	0.791	0.836
	CoT	0.830	0.888	0.646	0.767	0.825
	SR	0.762	0.816	0.662	0.739	0.776

Table B.2: Classification-only performance (human-annotated conclusions as input) under the Basic prompt, compared to end-to-end weighted F₁. Δ denotes the change from the end-to-end setting.

Model	E2E	HA input	Δ
Gemini 3.1 Pro	0.853	0.820	−0.033
Claude Opus 4.6	0.834	0.815	−0.019
Llama 3.3 70B	0.831	0.816	−0.015
Qwen 3.5 397B	0.830	0.806	−0.024
Claude Sonnet 4.5	0.820	0.799	−0.020
GPT-4o	0.810	0.800	−0.010
GPT-5	0.805	0.799	−0.006
GPT-4o mini	0.799	0.749	−0.050
Mean	0.823	0.801	−0.022

Qwen 3.5), and default for other providers. Approximately 2.6% of Claude Opus 4.6 outputs on biomedical abstracts were excluded due to API-level safety refusals; other models exhibited refusal rates below 0.5%.

Inference was conducted for all models between

Table B.3: Bootstrap 95% confidence intervals for weighted F₁ and false-positive rate (FPR) under Basic prompting (1,000 resamples with replacement). Non-overlapping FPR CIs between top-tier and lower-tier models indicate meaningful model-family separation in calibration behavior.

Model	Weighted F ₁ [95% CI]	FPR (%) [95% CI]
Gemini 3.1 Pro	0.853 [0.831, 0.874]	29.2 [23.7, 34.6]
Claude Opus 4.6	0.834 [0.810, 0.860]	39.9 [34.1, 46.6]
Llama 3.3 70B	0.831 [0.806, 0.855]	40.0 [34.5, 46.2]
Qwen 3.5 397B	0.830 [0.805, 0.856]	43.1 [36.8, 48.7]
Claude Sonnet 4.5	0.820 [0.794, 0.844]	44.5 [38.2, 50.4]
GPT-4o	0.810 [0.788, 0.833]	31.9 [26.4, 37.4]
GPT-5	0.805 [0.777, 0.830]	51.5 [45.8, 57.5]
GPT-4o mini	0.799 [0.777, 0.826]	40.8 [34.7, 46.7]

[05-13-2026] and [05-26-2026]. To support full reproducibility, we will publicly release upon publication: (i) the annotated dataset of 1,000 abstracts, including per-annotator labels, adjudication notes, and final consensus judgments; (ii) raw JSON predictions for all 32 model–prompt configurations on all 1,000 abstracts; (iii) all prompt templates

Table B.4: **Model identifiers, providers, and inference configurations.**

Display Name	API Identifier	Provider
GPT-4o	gpt-4o-2024-11-20	OpenAI
GPT-4o mini	gpt-4o-mini-2024-07-18	OpenAI
GPT-5	gpt-5	OpenAI
Claude Sonnet 4.5	claude-sonnet-4-5	Anthropic
Claude Opus 4.6	claude-opus-4-6	Anthropic
Gemini 3.1 Pro	gemini-3.1-pro-preview	Google
Llama 3.3 70B	meta-llama/llama-3.3-70B-Instruct-Turbo	Together AI
Qwen 3.5 397B	Qwen/Qwen3.5-397B-A17B	Together AI

and Few-Shot exemplars; (iv) non-LLM baseline implementations (SciBERT fine-tuning, TF-IDF classifiers, and rule-based extractors); and (v) evaluation scripts, including bootstrap CI computation, hierarchical clustering, and per-discipline metric aggregation.

B.4 Human Validation of SBERT Cosine Similarity

To assess whether SBERT cosine similarity meaningfully reflects causal-content preservation, we conducted a human validation on 100 extraction pairs from GPT-4o (Basic prompt), stratified across four cosine ranges to ensure coverage: $[0.00, 0.50]$ ($n=22$), $(0.50, 0.70]$ ($n=26$), $(0.70, 0.85]$ ($n=26$), and $(0.85, 1.01]$ ($n=26$).

Each pair was rated on a 5-point scale for how well the LLM extraction preserves the causal content of the human-annotated conclusion: 5 = causal claim preserved identically or accurately paraphrased; 4 = causal claim preserved with minor wording differences; 3 = topic preserved but causal relationship vague; 2 = causal claim distorted (e.g., causal converted to associational); 1 = causal claim entirely missed or fabricated.

Mean human scores increase monotonically with SBERT cosine (Table B.5), with a statistically significant Spearman correlation of $\rho = 0.397$ ($p < 0.0001$) and Pearson $r = 0.412$ ($p < 0.0001$). The moderate correlation magnitude reflects a ceiling effect: LLM extractions preserve causal content well in the majority of cases (mean score 4.44/5, with 79% of pairs receiving a score of 4 or 5), which compresses correlation coefficients. Notably, no pair received a score of 1, indicating that catastrophic extraction failures (complete fabrication) are absent in the sampled data.

These results support SBERT cosine similarity as a valid proxy for extraction quality: higher cosine reliably indicates better causal-content preservation. The metric is nevertheless sensitive to phrasing variation rather than a direct measure of causal

semantics, which we note as a caveat when interpreting extraction-only results.

Table B.5: **Human validation of SBERT cosine similarity.** Mean human score for causal-content preservation increases monotonically with SBERT cosine similarity.

SBERT cosine range	n	Mean human score
$[0.00, 0.50]$	22	4.00
$(0.50, 0.70]$	26	4.50
$(0.70, 0.85]$	26	4.65
$(0.85, 1.01]$	26	4.88
Overall	100	4.44

B.5 Extraction Quality: LLMs vs. Rule-based Baselines

Table B.6 reports mean SBERT cosine similarity between model-extracted conclusions and human-annotated conclusions, comparing LLMs (Basic prompt) against four rule-based heuristics. LLM extractions consistently outperform all rule-based baselines, with the best LLM (Qwen) exceeding the best heuristic (last 30% of abstract) by 0.096 in mean cosine similarity.

Table B.6: **Extraction quality: LLMs under the Basic prompt vs. rule-based baselines.** Methods are ranked by Cos., which denotes mean SBERT cosine similarity to the human-annotated conclusion. ≥ 0.7 denotes the proportion of extractions with SBERT cosine similarity at least 0.7.

Method	Cos.	≥ 0.7
Qwen 3.5 397B	0.782	0.786
GPT-5	0.772	0.778
Gemini 3.1 Pro	0.768	0.748
Llama 3.3 70B	0.766	0.726
Claude Sonnet 4.5	0.762	0.762
Claude Opus 4.6	0.761	0.786
GPT-4o	0.759	0.719
GPT-4o mini	0.740	0.665
Last 30% of abstract	0.686	0.512
Last two sentences	0.666	0.439
Causal-verb sentence	0.606	0.318
Last sentence	0.592	0.299

B.6 Non-LLM Baseline Implementation

TF-IDF baselines. We use scikit-learn’s TfidfVectorizer with unigram and bigram features, minimum document frequency of 2, maximum document frequency of 0.95, sublinear TF scaling, and English stopword removal. Logistic Regression and SVM are trained with

balanced class weights to address the 70:30 class imbalance; Random Forest uses 300 estimators with balanced class weights. All classifiers are evaluated via 5-fold stratified cross-validation with random seed 42.

SciBERT fine-tuning. We fine-tune allenai/scibert_scivocab_uncased on the human-annotated conclusions for 5 epochs per fold, with learning rate 2e-5, batch size 32, maximum sequence length 128, warmup ratio 0.1, AdamW optimizer, and class-weighted cross-entropy loss to address class imbalance. Evaluation uses 5-fold stratified cross-validation.

Rule-based extraction. We implement four heuristics using NLTK sentence tokenization: (i) *last sentence*: the final sentence of the abstract; (ii) *last two sentences*: the last two sentences concatenated; (iii) *causal-verb-weighted sentence*: the sentence containing the highest count of pre-defined causal verbs (e.g., causes, leads to, induces, regulates, controls), with fallback to the last sentence when no causal verbs are present; (iv) *last 30% of abstract*: the concatenation of sentences in the final 30% of the abstract.

B.7 Direct Classification Experiment

As a robustness check, we evaluated four representative models (GPT-5, Claude Opus 4.6, Gemini 3.1 Pro, Qwen 3.5 397B) on a direct causal/non-causal classification task, prompting the model to output a Yes/No label given only the abstract, without a conclusion-extraction step. The subset consisted of 200 abstracts sampled uniformly at random from the full 1,000-abstract benchmark (stratified by discipline). Results confirm the model-dependent heterogeneity observed in the end-to-end pipeline: GPT-5 continues to exhibit the highest false positives rate, whereas Qwen achieves the most balanced classification. This supports our interpretation that observed false positives patterns are a property of the instruction-tuned models themselves rather than an emergent effect of the two-stage pipeline.

C Case Study: Persistent False Positives in GPT-5

To better understand causal false positives, we examine the 117 non-causal abstracts that GPT-5 consistently misclassifies as causal under both Basic and CoT prompting. GPT-5 has the highest

Table B.7: **Categories of GPT-5 persistent false positives.** HA = corrected by human-annotated conclusions; FS = corrected by Few-Shot prompting. Overall, HA corrects 43.6% and FS corrects 29.1% of the 117 cases.

Root Cause	<i>n</i> (%)	HA	FS
Extraction-induced	51 (43.6%)	51/51	22/51
Classification-inherent	66 (56.4%)	0/66	12/66
Total	117	51/117	34/117

false positives rate among the evaluated models under both prompts (50.0% under Basic and 53.5% under CoT). Comparing these errors with the classification-only setting using human-annotated (HA) conclusions, we separate them into extraction-induced and classification-inherent false positives (Table B.7).

Extraction-induced false positives. In 51 cases, false positives are resolved when provided with HA conclusions, showing the errors stem from conclusion extraction. A common pattern is *causal verb injection*, where the model introduces causal verbs absent in the original text. For instance, in a Fe/S cluster synthesis study (PMID: 29097656), the human conclusion reports “crystal structures” and “SAXS analysis,” whereas GPT-5 extracts that the work “links” Fe/S cluster synthesis with lipid synthesis, imposing an ungrounded causal framing.

Classification-inherent false positives. The remaining 66 cases persist under HA conclusions, reflecting inherent classification errors. These often involve transformation, capability, or mechanistic phrasing prone to causal misinterpretation. For example, in a PbI₂ study (PMID: 32041958), the description that “PbI₂ transformed from a 1T structure to a 1H structure” is misclassified as causal. Other borderline cases include mechanistic explanations using causal phrasing without stating novel causal findings.

Recovery under Few-Shot prompting. Few-Shot prompting resolves 34 of the 117 persistent false positives (29.1%), recovering 22 extraction-induced cases (43.1%) versus 12 classification-inherent cases (18.2%). Consistent with our calibration findings, in-context examples effectively mitigate extraction-induced wording shifts, but offer limited recovery when conclusions lie near the causal boundary.

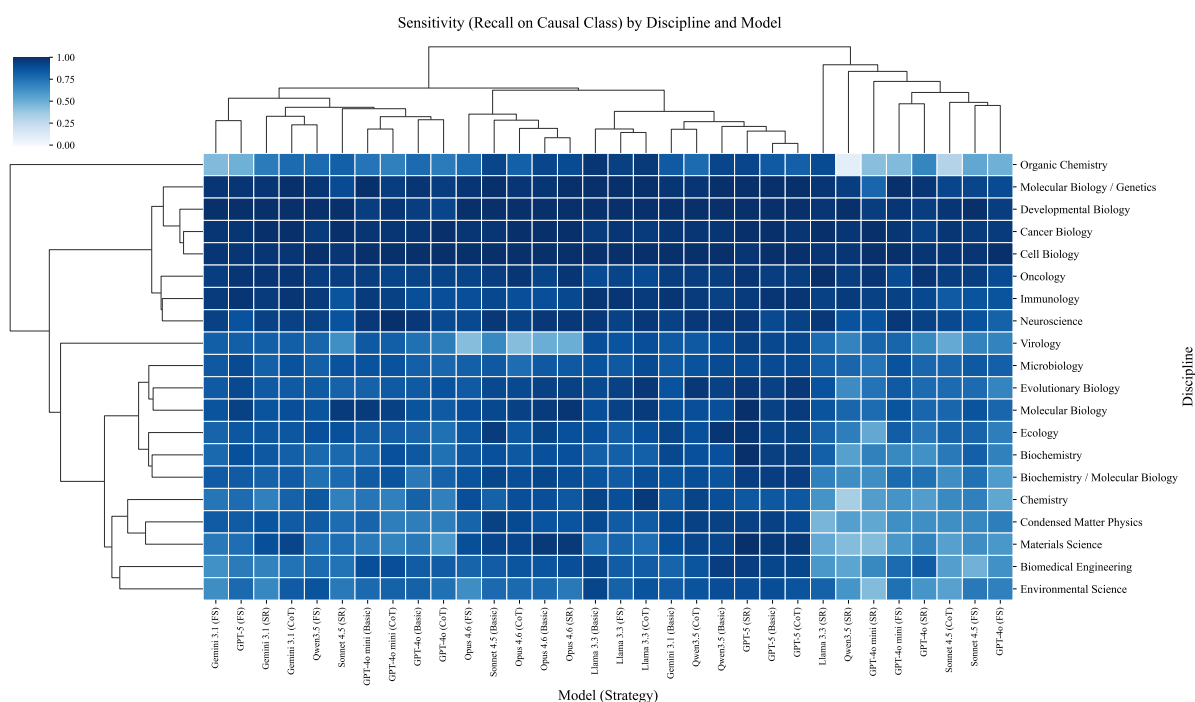


Figure B.2: Hierarchical clustering of causal-class sensitivity across disciplines and model-prompt configurations. Darker blue indicates higher recall on truly causal conclusions. Biomedical disciplines form a high-sensitivity cluster, indicating that causal conclusions in these fields are consistently recognized across models and prompts. In contrast, several physical-science disciplines, including Organic Chemistry, Chemistry, Condensed Matter Physics, and Materials Science, show lower sensitivity. Open-weight models such as Llama and Qwen under Basic and CoT prompts cluster with strong proprietary configurations.

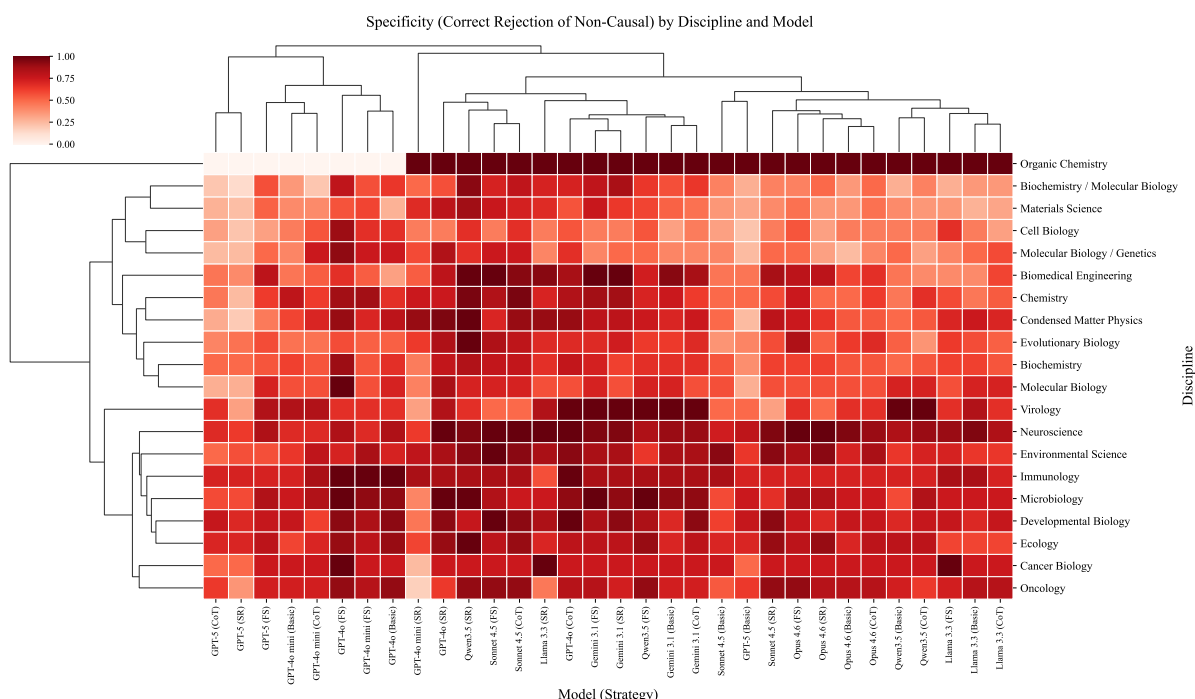


Figure B.3: Hierarchical clustering of non-causal specificity across disciplines and model-prompt configurations. Darker red indicates higher specificity, or more accurate rejection of truly non-causal conclusions. Low-specificity configurations are concentrated among GPT-family models, consistent with their higher false-positive rates. The heatmap also reveals a sensitivity–specificity tradeoff: some high-sensitivity configurations, such as GPT-5 under Basic, CoT, and SR prompts, show weaker specificity, whereas more conservative configurations, such as Qwen SR and Claude Sonnet SR, achieve higher specificity.

D LLM Prompts for Causal Claim Classification

This appendix presents the four prompt templates used in our evaluation pipeline. All four strategies share an identical base structure for fair comparison, differing only in their strategy-specific additions: Basic provides no scaffolding; Few-Shot adds in-context examples; Chain-of-Thought adds step-by-step reasoning instructions; and Self-Refinement operates as a binary verification of the first-pass prediction. The Few-Shot examples were constructed as illustrative cases representative of typical causal and non-causal patterns; we verified by both exact title matching and keyword-based search over the corpus that none of these examples overlap with the 1,000 abstracts in our evaluation dataset, eliminating data leakage.

D.1 Basic Prompt

Basic Prompt Template

You must decide if a paper's conclusion contains a *causal claim*.

Definition (use strictly):

- Causal = X changes or determines Y (verbs like: causes, leads to, induces, mediates, is required for, enables, abolishes, prevents, necessary for, results in).
- Non-causal = correlation/association /description/hypothesis/method/theory without directional change.

Task:

- 1) Write one-sentence conclusion from TITLE+ABSTRACT.
- 2) Decide causal (Yes/No) under the definition.
- 3) If No, relation_type in ["correlation", "association", "description", "hypothesis", "mechanistic/theory", "measurement/method", "other"].

Return strict JSON:

```
{"conclusion": "...", "causal": "Yes|No", "relation_type": <label or null>}
```

TITLE: {title}

ABSTRACT: {abstract}

JSON ONLY:

D.2 Few-Shot Prompt

Few-Shot Prompt Template

You are classifying whether a scientific conclusion contains a CAUSAL claim.

[Same definitions as Basic prompt]

Examples:

CAUSAL example:

TITLE: ATP depletion enhances condensate formation

ABSTRACT: ATP depletion increased phase separation of IDR proteins.

Return:

```
{"conclusion": "ATP depletion increases phase separation of IDR proteins.", "causal": "Yes", "relation_type": null}
```

[Three additional examples covering CAUSAL and NON-CAUSAL cases]

Task:

[Same as Basic prompt]

TITLE: {title}

ABSTRACT: {abstract}

JSON ONLY:

D.3 Chain-of-Thought Prompt

Chain-of-Thought Prompt Template

You are given a scientific paper's title and abstract. Think through the analysis step-by-step before providing your final output.

Step 1: Analyze the Text
(Internal Reasoning)

[Same definitions as Basic prompt]
Determine if the study design allows for causal inference (e.g., randomized control trial vs. observational study) or if it simply describes a phenomenon.

Step 2: Generate the Output
Based on your reasoning in Step 1, return the JSON specified below.

Return strict JSON:
{
 "conclusion": "...",
 "causal": "Yes|No",
 "relation_type": <label or null>
}

TITLE: {title}
ABSTRACT: {abstract}
JSON ONLY:

For end-to-end and classification benchmarks, the previous prediction's causal label is provided in the prompt. A "No" response flips the label; a "Yes" response retains the first-pass prediction.

D.4 Self-Refinement Prompt

The Self-Refinement strategy employs a two-step process. In the first step, the Basic prompt is used to obtain a first-pass prediction. In the second step, the model is asked to verify whether its first-pass prediction is correct; a "No" response triggers a label flip for the causal classification.

Self-Refinement Prompt (Second Pass)

A scientific paper was previously analyzed and classified as either CAUSAL or NON-CAUSAL based on its title and abstract.

[Definitions same as Basic prompt]

Your task: verify whether the previous classification is correct according to the definitions above.

Return strict JSON with exactly one field: {"correct": "Yes" or "No"}

"Yes" = the previous classification is correct.
"No" = the previous classification is wrong (the label should be flipped).

TITLE: {title}
ABSTRACT: {abstract}

PREVIOUS CLASSIFICATION: {label}

JSON ONLY: