HICAR: Human-in-the-Loop Causal Alignment and Reconciliation for Education via Verdict-Guided Retrieval

Anonymous ACL submission

Abstract

Causal knowledge in education is distributed across observational data and a large body of prior research. Causal discovery methods can infer structure from longitudinal data, but are often sensitive to modeling assumptions and unobserved confounding. At the same time, retrieval-augmented generation (RAG) can extract causal claims from text, but such claims may be context-dependent, directionally ambiguous, or weakly grounded in the dataset of interest.

We present HICAR, a human-in-the-loop framework for aligning and reconciling causal evidence from text and data. HICAR constructs a data-derived causal graph from observational data and uses it to anchor retrieval over the education literature, while preserving text-derived and data-derived graphs as distinct evidential views. Their agreement is treated as a signal of reliability, while disagreement triggers targeted causal queries for direct effects, reverse direction, confounding, and mediation.

We instantiate HICAR on the Education Longitudinal Study of 2002 and a corpus of education research over 22 concept nodes and 9,084 students. Compared with unguided retrieval, G_{data}-anchored retrieval reduces the candidate pair space by 87.9% and improves extraction precision by 32.4 percentage points. Human-in-the-loop reconciliation further expands validated coverage while explicitly preserving unresolved uncertainty.

HICAR reframes causal RAG from extracting isolated claims toward validating and reconciling causal evidence across text, data, and expert reasoning.

1 Introduction

Understanding causal relationships is central to advancing data-driven decision-making in complex domains such as education, where policies and interventions rely on understanding relationships between factors. For example, education policy decisions often depend on causal claims: which student, school, and institutional factors affect long-term educational outcomes, and under what conditions. However, reliable causal knowledge in education is difficult to construct (Pearl, 2009; Spirtes et al., 2000; Morgan and Winship, 2015; Feng, 2024). Observational education datasets are often incomplete, confounded, and limited in their measurement of social and institutional mechanisms, while causal evidence in the literature is fragmented across studies, populations, interventions, and contexts. As a result, neither data-driven causal discovery (CD) nor text-derived causal knowledge alone is sufficient for building reliable causal representations.

This challenge is evident in longitudinal education datasets such as the Education Longitudinal Study of 2002 (ELS:2002) (Ingels et al., 2011), a nationally representative study that follows U.S. students from high school into postsecondary education and the workforce. ELS contains rich information about students, families, schools, academic preparation, aspirations, and educational outcomes, making it valuable for studying educational trajectories. At the same time, like many real-world education datasets, it contains missing variables, correlated covariates, limited sample sizes for some subgroups, and unobserved social or institutional factors. Consequently, CD methods, which aim to recover causal structure from observational data under assumptions such as causal sufficiency and faithfulness (Spirtes et al., 2000; Chickering, 2002; Peters et al., 2017), often produce multiple plausible graphs rather than a uniquely identifiable causal structure. This limitation helps explain why prior education research has largely focused on estimating specific causal effects, such as the effects of financial aid (Liu et al., 2025b; Dynarski, 2003; Nguyen et al., 2019), academic preparation (Dynarski et al., 2023), school context (Kraft et al.,

2016), and student engagement (Kuh et al., 2008), instead of learning and validating broader causal structure.

Scientific literature provides a complementary source of causal evidence. Recent advances in large language models (LLMs) and retrieval-augmented generation (RAG) make it possible to extract structured causal relationships from research articles and policy reports (Lewis et al., 2020; Guu et al., 2020). In education, such text encodes evidence from randomized trials, quasi-experimental studies, policy interventions, and qualitative analyses. However, text-derived causal claims are also imperfect: they may be context-specific, based on implicit assumptions, or not directly grounded in the empirical dataset of interest (Kıcıman et al., 2023; Zhou et al., 2025). Therefore, the key challenge is not simply to extract more causal relations from text, or to learn another graph from data, but to validate and reconcile causal evidence across imperfect sources.

Recent work on causal RAG has begun to couple retrieval with causal reasoning, including systems that retrieve passages to support causal claims (Wang et al., 2025), use reinforcement-learning-driven query refinement over causal pathways (Khatibi et al., 2025), construct entity graphs from text via LLM (Edge et al., 2024), and perform multi-hop retrieval over logical passage graphs (Liu et al., 2025a). Iterative RAG systems such as Self-RAG (Asai et al., 2024), CRAG (Yan et al., 2024), and Iter-RetGen (Shao et al., 2023) trigger re-retrieval from model self-assessment or prior generation rather than from typed human verdicts. A fundamental limitation unites all of these systems: none anchors retrieval to a statistically validated causal graph, and none encodes typed human-expert verdicts as causal-role-specific retrieval directives.

Systems that combine HITL with RAG over causal graphs come closer but still fall short. Cora (Barres et al., 2025) implements iterative RAG grounded in a formal causal model with typed retrieval roles, but those roles are derived by the LLM from extracted text rather than from typed human verdicts on individual edges, and user edits do not generate new typed corpus queries. CausalChat (Zhang et al., 2025) elicits mediator, confounder, and latent-variable probes from users, but queries GPT-4's parametric knowledge rather than a text corpus. To our knowledge, HICAR is among the first frameworks to encode typed expert verdicts on causal edges as retrieval directives for subsequent corpus search.

This paper asks: *How can a statistically validated causal graph anchor text-based causal extraction, and how can a one-to-one mapping from causal verdict types to causal query types encode expert judgment as targeted retrieval directives that close a self-referential validation loop?*

We address this question with HICAR (Human-in-the-Loop Causal Alignment and Reconciliation), a self-referential causal RAG framework in which a data-grounded causal graph G_{data} anchors and iteratively steers text-based causal extraction. Our framework treats text and data as two complementary but potentially inconsistent views of an underlying causal structure. We first construct a text-derived causal graph G_{text} by adapting a G_{data}-anchored CD-RAG pipeline (Zhou et al., 2025) over education literature, and a data-derived causal graph G_{data} using an ensemble of CD algorithms applied to ELS:2002 data. Rather than immediately merging these graphs, we preserve their distinction and compare their edge-level agreement and disagreement. Edges supported by both modalities provide stronger candidates for reliable causal relationships, while modality-specific edges expose conflicts that may arise from latent confounding, measurement limitations, variable-definition mismatch, or context-specific mechanisms.

Human expertise is central to our framework. Cross-modal agreement alone cannot resolve whether a disagreement reflects a false discovery, a missing variable, a context-dependent causal effect, or a limitation of the dataset. We therefore introduce an iterative HITL reconciliation procedure grounded in a one-to-one mapping from causal verdict types to causal query types. A *needs-evidence* verdict fires a mechanism query into the corpus; a *wrong-direction* verdict fires a reversal query; a *both-plausible* verdict fires a feedback query; and a confirmed edge suspected of mediation fires a mediation query. Because each verdict type classifies the *nature* of the causal ambiguity — not merely its existence — the system retrieves precisely the passages needed to resolve that specific ambiguity, rather than generic additional evidence.

We instantiate our framework in education by integrating causal knowledge extracted from literature with CD on ELS:2002. We further propose a cross-modal agreement metric that formally tests whether text–data agreement predicts causal reliability, and show that agreement edges exhibit

stronger causal effects than data-only edges. More broadly, our work shifts the focus from CD alone to causal validation: instead of asking a single modality to recover the causal graph, we provide a framework for validating and reconciling causal evidence across text, data, and expert reasoning.

Contributions. (1) Verdict-guided causal retrieval. To our knowledge, HICAR is among the first causal RAG systems to establish a one-to-one functional mapping from causal verdict types (needs evidence, wrong direction, both plausible, mediation suspected) to causal query types (mechanism, reversal, feedback, mediation), closing a self-referential loop in which the graph under construction guides its own evidence gathering. Unlike systems that trigger re-retrieval from model self-assessment (Self-RAG, CRAG) or RL reward (CDF-RAG), or that consume expert input as DAG constraints, HICAR exploits the *type* of the expert's causal judgment to retrieve the structurally relevant passages.

(2) Graph-anchored causal extraction from text. Anchoring RAG to a data-derived causal graph reduces the candidate pair space by 87.9% relative to the all-pairs directed grid and improves extraction precision by +32.4 percentage points over unguided text retrieval, while eliminating direction-blind NLI artifacts by design.

(3) Cross-modal validation with HITL reconciliation. We propose a cross-modal agreement metric — a 2×2 matrix of NLI entailment scores and SEM β^* coefficients — that formally tests whether cross-modal agreement predicts causal reliability, and couple this metric with an iterative expert-feedback procedure in which each verdict triggers a targeted re-retrieval pass and expert judgments accumulate across rounds.

2 Related Work

Causal discovery from observational data. CD aims to recover causal structure from observational data under assumptions such as causal sufficiency, faithfulness, and acyclicity. Classical approaches include constraint-based methods such as PC and FCI (Spirtes et al., 2000), which rely on conditional independence testing, and score-based methods such as GES (Chickering, 2002), which optimize a graph score over candidate structures. More recent work formulates CD as a continuous optimization problem (Zheng et al., 2018). Comprehensive treatments of causal inference are provided by (Pearl, 2009; Peters et al., 2017). While these methods provide a principled foundation for learning causal structure from data, they are often underdetermined in practice and sensitive to modeling assumptions and unobserved confounding.

Causal inference in education. Education research has extensively studied causal relationships using randomized experiments, quasi-experimental designs, and longitudinal analyses. For example, Dynarski (2003) and Angrist et al. (2009) examine the causal effects of financial aid and incentives on college outcomes, while Nguyen et al. (2019) provides a meta-analysis of grant aid impacts on persistence and attainment. Structural and theoretical perspectives on student success and attrition are developed by Tinto (1993) and Perna (2006). Additional work studies the roles of expectations, peer effects, and inequality in shaping educational outcomes (Jacob and Wilder, 2011; Sacerdote, 2011; Herbaut, 2020). Despite this rich literature, causal knowledge in education remains fragmented across studies and is not systematically extracted or integrated into a unified causal representation.

Iterative and causal RAG. Retrieval-augmented generation (RAG) has emerged as a powerful approach for grounding large language models in external knowledge sources (Lewis et al., 2020; Guu et al., 2020). Recent work explores the ability of LLMs to extract and reason about causal relationships from text (Kıcıman et al., 2023). Iterative RAG systems trigger re-retrieval from model self-assessment: Self-RAG (Asai et al., 2024) learns inline reflection tokens; CRAG (Yan et al., 2024) uses a T5-based retrieval evaluator; FLARE (Jiang et al., 2023b) re-retrieves on low-confidence token predictions; and Iter-RetGen (Shao et al., 2023) concatenates the model's prior output as the next query. All four rely on automated signals — not human verdicts — and none types queries by causal role. Dedicated causal RAG systems impose additional structure: CausalRAG (Wang et al., 2025) retrieves passages to support causal claims but is single-shot and data-free; CDF-RAG (Khatibi et al., 2025) iteratively refines queries using a PPO-trained agent over causal pathways but has no human and no causal-role query typing; GraphRAG (Edge et al., 2024) constructs entity graphs from text via LLM with no causal semantics; and HopRAG (Liu et al., 2025a) performs multi-hop retrieval where hops are lexically rather than causally directed. A common limitation unites these sys-

tems: none anchors retrieval to a statistically validated causal graph. Barres et al. (2025) (Cora) implement iterative RAG grounded in a formal causal model with typed retrieval roles (Qualitative Process Theory: Influences-direct, Influences-inverse, Triggers), but the role typing is LLM-derived from extracted text rather than from typed human verdicts on individual edges, and user edits do not generate new typed corpus queries. Most related to our evaluation design, Hassanzadeh (2024) present WikiCausal, the first evaluation framework for causal knowledge graph construction from text, measuring precision via LLM-based evaluation without manual annotation. We adopt their LLM-judged precision paradigm and extend it with a cross-modal agreement metric that jointly evaluates NLI entailment and SEM coefficient significance.

Knowledge-augmented and multimodal causal discovery. Several recent approaches incorporate prior knowledge, metadata, or expert input into CD (Ban et al., 2023; Zečević et al., 2023). These methods treat external knowledge as a prior or constraint rather than as a separate source of causal evidence to be validated. Zhang et al. (2025) (CausalChat) use LLMs to iteratively probe mediator, confounder, and latent-variable relationships through multi-perspective prompting, but queries are directed at GPT-4's parametric knowledge rather than a text corpus via RAG retrieval. CauScientist (Yong et al., 2025) synergizes LLMs as hypothesis-generating agents with probabilistic statistics as verifiers. Unlike CauScientist, HICAR applies to a real longitudinal dataset without ground truth and involves a domain expert to resolve cross-modal conflicts.

Human-in-the-loop causal discovery. Human-in-the-loop (HITL) methods incorporate expert feedback to improve reliability and alignment in machine learning systems (Wu et al., 2022). In CD, prior work elicits expert input on edges as add, remove, or orient operations (Ban et al., 2023; Ghai and Mueller, 2022; Ankan and Textor, 2025), as binary confirmations in Bayesian DAG posteriors (Bonilla et al., 2026), or as noisy judgments in ancestral graph learning (Tóth et al., 2023). Ghai and Mueller (2022) encode these as formal graph operations — a pattern we adopt and extend. Crucially, all prior HITL causal systems consume expert input as DAG constraints or parametric priors — none encodes typed verdicts as causal-role-specific retrieval directives fired back into a text corpus. To the best of our knowledge, HICAR is the first to close this loop.

Positioning. Prior work has largely treated CD from data, causal extraction from text, and human-in-the-loop refinement as separate problems. This work connects these directions by focusing on the validation and reconciliation of causal knowledge across modalities. The key distinction is the one-to-one functional mapping from causal verdict types to causal query types. Prior HITL systems consume verdicts as DAG constraints or Bayesian priors; prior iterative RAG systems use model tokens or RL rewards to trigger re-retrieval. HICAR is, to the best of our knowledge, the first system to exploit the *nature* of the expert's causal judgment — direction dispute, mechanism gap, bidirectionality, mediation — to generate structurally targeted corpus queries, closing a self-referential loop in which the graph under construction guides its own evidence gathering.

3 Methodology

We present **HICAR**, a framework for **H**uman-in-the-loop **C**ausal **A**lignment and **R**econciliation across text and data. HICAR is designed for settings in which causal evidence is distributed across imperfect sources: observational data provide statistical constraints on causal structure, while scientific literature provides contextual and mechanistic claims that may not be directly identifiable from the dataset alone. Rather than merging these sources into a single model, HICAR preserves them as distinct evidential views and uses their agreement, disagreement, and expert resolution to construct a validated causal graph.

HICAR observes two partial views of an unknown causal graph $G{=}(V, E)$: a data-derived graph G_{data} inferred from observational data, and a text-derived graph G_{text} extracted from a corpus $\mathcal{T}$ (Kıcıman et al., 2023).

HICAR proceeds in three stages (Fig. 1). **Stage 1: Ground** constructs G_{data} from observational data. **Stage 2: Extract** constructs G_{text} via G_{data}-anchored retrieval and multi-perspective causal prompting. **Stage 3: Iterative Validation** alternates between two steps — **3a Query**, which encodes expert verdicts as typed retrieval directives and re-retrieves targeted corpus evidence, and **3b Validate**, which routes the updated evidence to expert reconciliation. Stages 3a and 3b form a loop: each Validate pass issues verdicts that trig-

ger a new Query pass, and the updated corpus evidence is returned to Validate. The output is a validated graph $G^\star$ integrating cross-modal agreement, expert-resolved discrepancies, and explicit residual uncertainty.

3.1 Problem Formulation

The central object in HICAR is not a single discovered graph, but an evidence-aligned causal representation. Given G_{data} and G_{text}, we define the cross-modal agreement set

$$E_\cap = E_{\text{data}} \cap E_{\text{text}}. \tag{1}$$

Edges in $E_\cap$ are supported by both statistical structure and textual evidence, and are therefore treated as high-reliability candidates. Edges that appear in only one modality are treated as diagnostically useful discrepancies: they may indicate latent confounding, insufficient measurement, contextual variation, direction ambiguity, or causal claims supported in the literature but not identifiable from the observed data.

The goal is to construct a validated graph $G^\star{=}(V, E^*)$ by combining three sources of evidence: statistical support from G_{data}, textual support from G_{text}, and expert resolution of ambiguous or conflicting edges.

3.2 Stage 1 — Ground: Data-Derived Causal Evidence

The data-derived graph G_{data} represents causal structure supported by observational data. HICAR constructs G_{data} through an evidence-aggregation strategy rather than relying on one algorithmic view. We consider three complementary CD principles instantiated using FGES (Chickering, 2002; Ramsey et al., 2017), PC-stable (Colombo and Maathuis, 2014), and FCI (Spirtes et al., 2000; Zhang, 2008). These methods operate under standard structural assumptions including the Causal Markov condition, faithfulness, and acyclicity (Spirtes et al., 2000; Jaber et al., 2022). Temporal ordering in the longitudinal setting provides a natural basis for restricting admissible directions. FGES and PC-stable assume causal sufficiency; FCI relaxes this to allow latent confounding.

The outputs are aggregated into a high-confidence edge set:

$$E_{\text{HC}} = \left\{ (v_i \to v_j) : \sum_{\text{alg}} \mathbf{1}\big[(v_i \to v_j) \in \hat{G}_{\text{alg}}\big] \geq 2 \right\} \tag{2}$$

We define $G_{\text{data}}{=}(V, E_{\text{HC}})$. For edges retained in G_{data}, effect strengths are estimated via a linear SEM; standardized coefficients β^* populate the cross-modal agreement matrix (details in Appendix A.3).

3.3 Stage 2 — Extract: Text-Derived Causal Evidence

G_{text} is constructed by retrieving evidence passages from $\mathcal{T}$ for candidate variable pairs. Unlike prior causal RAG systems that retrieve freely over all pairs, Stage 2 is G_{data}-anchored: retrieval is focused on pairs in E_{HC} and their neighbors. This anchoring reduces the candidate search space and focuses retrieval on statistically plausible node pairs, while leaving causal judgments to be made from retrieved textual evidence. Each candidate edge $(v_i \to v_j)$ receives a corpus support score via NLI entailment averaging,

$$s_{ij} = \frac{1}{k} \sum_{c=1}^{k} p_{\text{entail}}(c,\, v_i \to v_j), \tag{3}$$

and edges above threshold τ form E_{text}. Multi-perspective prompts (direct, reverse, confounder, mediator) assign structured edge statuses used in alignment and expert reconciliation (Appendix A.4).

3.4 Stage 3a — Query: Verdict-Typed Re-Retrieval

Stage 3a (Query) closes the self-referential loop. After expert verdicts are issued in Stage 3b (Validate), each verdict is encoded as a typed retrieval directive matched to the nature of the causal ambiguity: NEEDS-EVIDENCE fires a mechanism query, WRONG-DIRECTION fires a reversal query, BOTH-PLAUSIBLE fires a feedback query, and a confirmed edge suspected of mediation fires a mediation query (see Table 1; full query templates in Appendix A.7).

Retrieved passages are re-scored via Eq. (3); edges crossing τ are promoted to E_{text} and returned to Stage 3b for the next Validate pass.

3.5 Cross-Modal Causal Alignment

The alignment stage compares G_{data} and G_{text} at the edge level, classifying each candidate edge into five categories (Appendix A.6) and ranking by a priority score combining corpus uncertainty, directional conflict, ensemble support, and structural ambiguity (Appendix A.7).

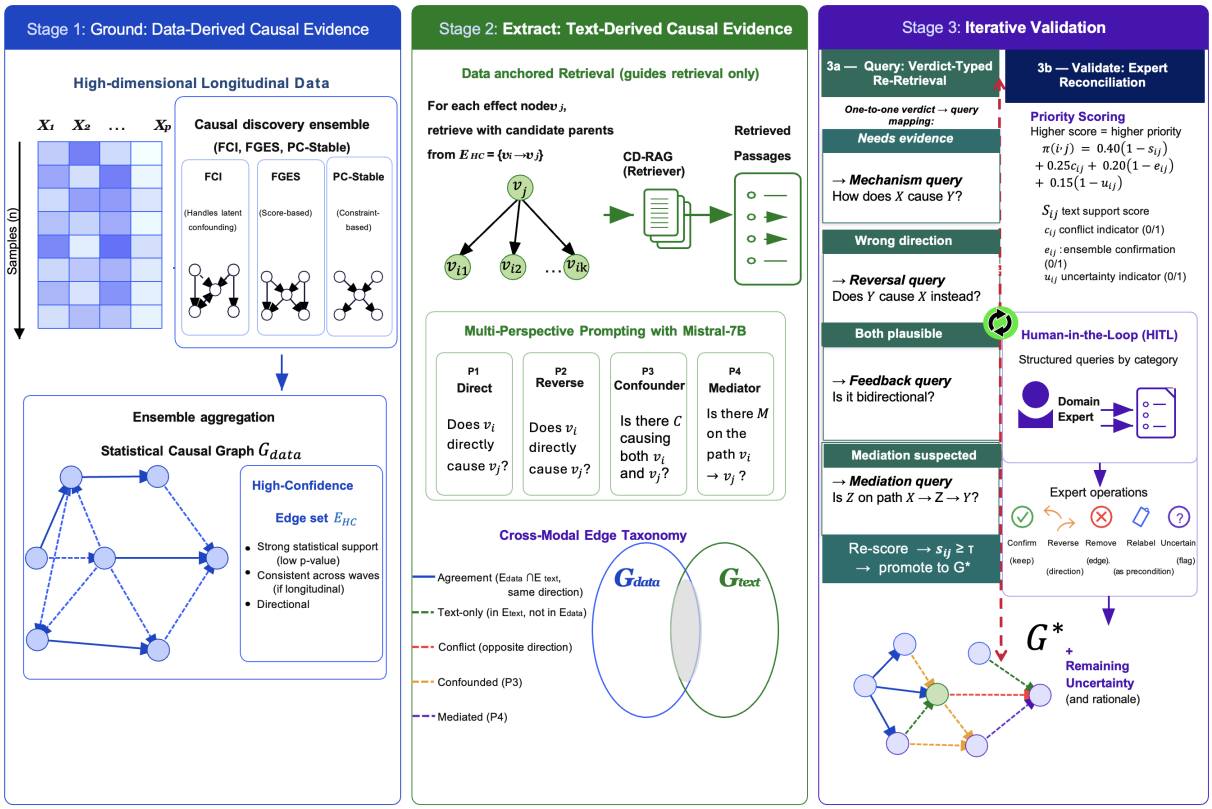

Figure 1: Overview of HICAR. **Stage 1 (Ground)** constructs G_{data} from observational data using an ensemble of CD algorithms. **Stage 2 (Extract)** constructs G_{text} via G_{data}-anchored retrieval and multi-perspective causal prompting. **Stage 3 (Iterative Validation)** alternates between **3a Query** — which encodes expert verdicts as typed retrieval directives (mechanism, reversal, feedback, mediation) and re-retrieves targeted corpus evidence — and **3b Validate** — which presents updated evidence to the domain expert and collects verdicts that trigger the next Query pass. The $\rightleftharpoons$ arrow between 3a and 3b denotes the self-referential loop; the loop terminates when no verdicts trigger further re-retrieval.

3.6 Stage 3b — Validate: Expert Reconciliation

Human expertise resolves causal claims that neither text nor data can determine alone. Expert responses are encoded as formal graph operations (CONFIRM, REVERSE, REMOVE, RELABEL, UNCERTAIN; Appendix A.7).

Reconciliation iterates with *verdict memory*: edges receiving explicit confirm/reject verdicts are not re-presented in subsequent rounds, and each Validate pass triggers a new Query pass that surfaces fresh evidence for unresolved edges. The final output is a validated graph $G^\star = (V, E^*)$ containing cross-modally supported edges, expert-resolved discrepancies, and explicitly marked uncertain relationships.

Causal identifiability. Cross-wave edges in $G^\star$ are directionally constrained by temporal ordering; within-wave edges are annotated by the domain expert as plausibly sequential or simultaneous, producing a two-tier identifiability table (see Appendix A.10). Standard graphical criteria are applied where a valid adjustment set exists (Pearl, 2009); full details are in Appendix A.7.

Instantiation on ELS:2002. We instantiate HICAR on ELS:2002 (Ingels et al., 2011) (n=9,084, $|V|$=22 concept nodes, 4 waves). The corpus comprises 3,439 chunks from 101 education research sources (58 full-text PDFs and 43 abstract-only documents). G_{text} is constructed at the operational threshold τ=0.03; post-hoc calibration against expert verdicts identifies τ=0.07 as F1-optimal (F1 = 0.892, Appendix A.9), with a permutation-based fake-edge falsification test (Eulig et al., 2025) setting a conservative upper bound τ_{fake}=0.1288. Full details in Appendix A.

4 Results

This section follows the three-stage structure of Fig. 1. *Stage 1* (Ground) builds the statistical layer

Table 1: Verdict-typed retrieval directives in HICAR.

Expert verdict	Directive	Purpose
NEEDS-EVIDENCE	Mechanism	Causal mechanism for $v_i \to v_j$
WRONG-DIRECTION	Reversal	Evidence for $v_j \to v_i$
BOTH-PLAUSIBLE	Feedback	Bidirectional / context evidence
MEDIATION-SUSP.	Mediation	Evidence for $v_i \to m \to v_j$

G_{data} and E_{HC}; *Stage 2* (Extract) constructs G_{text} via G_{data}-anchored retrieval and multi-perspective prompting; *Stage 3* (Iterative Validation) aligns modalities and runs iterative expert reconciliation to obtain $G^\star$. Table 2 summarises the progressive build.

4.1 Stage 1 — Ground: Statistical Layer

We combine FGES, PC-stable, and FCI on ELS:2002, yielding $|E_{\text{HC}}|$=56 directed edges among $|V|$=22 concept nodes after listwise deletion (n=9,084; Appendix A.1). The fitted structural equation model shows acceptable global fit (CFI=0.935, RMSEA=0.055) (Hu and Bentler, 1999). FCI contributes 16 additional orientations under latent confounding (Spirtes et al., 2000; Zhang, 2008). Of 35 cross-wave edges in E_{HC}, 32 (91%) are significant at p<0.05 on a held-out split.

4.2 Stage 2 — Extract: Retrieval and Prompting

CD-RAG proposes 67 text-supported directed candidates after G_{data}-anchored retrieval; edges above τ=0.07 form E_{text}. For comparison, unguided retrieval (no E_{HC} injection) yields $|E_\cap|$=14 at **20.6%** precision against FGES.

G_{data}-anchoring reduces the candidate pair space by **87.9%** (462 → 56 pairs) and improves extraction precision by +**32.4 percentage points** over the unguided Mistral free-extraction baseline (guided 52.9% vs. unguided 20.6% at τ=0.03 against FGES); a post-hoc audit attributes 2 of 13 unguided edges above τ=0.1288 to direction-blind NLI artifacts that guided retrieval eliminates by design. Full breakdown in Appendix A.8.

With G_{data}-anchored retrieval (P1), $|E_\cap|$ rises to 15 while $|E_{\text{text}}|$ stays nearly unchanged (67 vs. 68 unguided), so the gain is alignment precision rather than graph size. Data-graph coverage $|E_\cap|/|E_{\text{HC}}|$=15/56=26.8%.

P2–P4 leave $|E_\cap|$ at 15 but surface cross-modal tension: P2 raises direction conflicts to 2 (PSE enrollment → Financial aid; School context → Race/Ethnicity), and P3 flags two agreement edges as potentially confounded: Financial resources → Academic preparation (confounder: parental education) and School engagement → Academic achievement (confounder: SES background). The NLI threshold τ=0.07 used throughout is the F1-optimal value from expert-verdict calibration (F1 = 0.892; Appendix A.9).

4.3 Cross-Modal Agreement as a Causal Reliability Signal

We construct a 2×2 matrix indexed by G_{data} and G_{text} membership, reporting mean NLI score $\bar{s}$ and mean standardized SEM coefficient $\bar{\beta}^*$ per cell (Table 3).

4.4 Stage 3 — Iterative Validation and $G^\star$

Stage 3 runs three iterative reconciliation rounds; after each Validate pass, expert verdicts trigger a Stage 3a Query pass that re-retrieves targeted corpus evidence before the next round. Per-round outcomes (presented, confirmed, removed, retention): Round 1 (13, 4, 1, 31%); Round 2 (44, 30, 13, 68%, with one direction-corrected edge); Round 3 (10, 7, 4, 70%). Across all three rounds, 67 edges were presented; 41 confirmed and 18 removed, for an overall retention rate of **61.2%**. The rising per-round retention (31% → 68% → 70%) reflects the self-referential loop: each Query pass surfaces better-supported edges, reducing the proportion the expert must reject.

The final $G^\star$ contains 55 edges: 15 agreement edges, 33 expert-confirmed in Rounds 1–2 (including one direction-corrected), and 7 in Round 3; two Stage 2 direction conflicts are resolved by expert judgment and one was retracted following the Round 3 verdict.

Round 3 substantive finding. The domain expert rejected School context → Financial aid (postsec.) and surfaced a mediated alternative: School context → PSE enrollment → Financial aid, suggesting institutional access interventions operate via enrollment rather than direct aid allocation. Figure 2 (Appendix A.11) shows the full $G^\star$; of 55

Step	$\lvert E_{\mathrm{HC}}\rvert$	$\lvert E_{\mathrm{text}}\rvert$	$\lvert E_{\cap}\rvert$	Conf.	Cnf	Rm	Ret.%	$\lvert E^{*}\rvert$
1. FGES+PC+FCI ensemble ($\geq$2/3 vote)	56	—	—	—	—	—	—	—
2. + Unguided text (baseline)	56	68	14	—	—	—	—	—
3. + G_{data}-anchored retrieval (P1)	56	67	15	—	—	—	—	—
4. + P2–P4 (reverse/confounder/mediator)	56	67	15	2	—	—	—	—
5. + Stage 3 Round 1 HITL	56	67	15	2	4	1	31%	19
6. + Stage 3 Round 2 HITL	56	67	15	2	30	13	68%	48
7. + Stage 3 Round 3 HITL (final)	56	67	15	2	7	4	70%	**55**

Table 2: **Progressive HICAR build** on ELS:2002. Each row adds the named component to the configuration above. $\lvert E_{\mathrm{HC}}\rvert$: high-confidence data edges; $\lvert E_{\mathrm{text}}\rvert$: thresholded text edges ($s_{ij} \geq \tau$); $\lvert E_{\cap}\rvert$: cross-modal agreement. **Conf.**: direction conflicts before HITL. **Cnf/Rm**: expert CONFIRMations and REMOVEals per round. **Ret.%**: expert retention rate per round. Rows 5–7 show the three iterative validation rounds; each round's verdicts triggered a new Stage 3a Query pass before the next round.

Cell	n	$\bar{s}$	$\lvert\bar{\beta}^{*}\rvert$	% sig.
Agreement ($E_{\cap}$)	15	0.095	**0.103**	**61.5%**
Data-only	43	—	0.076	55.8%
Text-only	58	>0.07	—	—

Table 3: Cross-modal agreement matrix. Agreement edges have 35% larger standardized causal effects and 5.7 pp more statistically significant coefficients than data-only edges, supporting the claim that text–data convergence is a reliable signal of causal validity. Two of the 15 agreement edges lack β^{*} because their endpoints are not jointly observed in the SEM-fitted subgraph; the agreement mean is over the 13 edges with β^{*}.

edges, 36 are cross-wave (formally identified) and 19 within-wave, with 100% cross-wave and 73.7% overall directional agreement (Appendix A.10).

4.5 Interpretation

Eight of the 15 agreement edges are cross-wave (inheriting longitudinal ordering, Pearl, 2009); the other seven lie within Wave 1 and should be read as mutually consistent rather than fully identified.

(i) Equity chain. A near-complete agreement chain runs from Race/Ethnicity through school context, academic preparation, PSE enrollment, to degree completion (3/4 hops in $E_{\cap}$; the school-context-to-preparation hop added by Round 2) (Chetty et al., 2014).

(ii) Parallel channels. Financial resources relate to degree completion through both academic and financial-aid pathways in $G^{\star}$; the academic route is largely agreement edges, while the financial-aid route carries one unresolved directional conflict.

(iii) Confounding and mediation. Two agreement edges are P3-flagged as potentially confounded: Financial resources $\rightarrow$ Academic preparation (parental education) and School engagement $\rightarrow$ Academic achievement (SES background) — consistent with intergenerational educational transmission (Chetty et al., 2014; Nguyen et al., 2019). The Round 3 mediation candidate (School context $\rightarrow$ PSE enrollment $\rightarrow$ Financial aid) similarly suggests institutional access operates via enrollment rather than direct aid allocation.

Limitations

We acknowledge several limitations of HICAR and the education instantiation reported here.

Dataset specificity. Our evaluation uses a single longitudinal dataset (ELS:2002) with 22 concept nodes and 9,084 students. The specific edges in $G^\star$ are properties of this cohort and this node design; generalisation to other education datasets (ECLS-K, NLSY, HSLS) requires re-running discovery and re-annotating with domain experts.

Corpus heterogeneity. A small number of large documents dominate top-k retrieval for some edges. The top 4 documents in our 3,439-chunk corpus contribute roughly 19% of all chunks; for edges whose topic is only loosely related to these documents, retrieval can return weakly relevant evidence (a pattern the domain expert flagged in Round 2 review). Our two-tier audit — direction-blind NLI detection plus permutation-based falsification (Appendix A.9) — catches the most severe artifacts, but the broader risk of retrieval imbalance from corpus heterogeneity remains.

Single domain expert. Reconciliation used one domain expert in education research. Multiple-annotator studies with inter-rater agreement would strengthen the HITL validation claims, and we leave this for future work.

Causal extraction model. Text-side extraction uses Mistral-7B-Instruct-v0.2; larger or instruction-tuned causal-extraction models may produce different candidate sets. The contribution of HICAR (verdict-typed retrieval directives + G_{data} anchoring) is orthogonal to the choice of extractor.

Identifiability of LM-derived edges. Unlike FGES, PC-stable, and FCI, text-based causal extraction lacks formal identifiability guarantees. HICAR mitigates this via cross-modal agreement and expert reconciliation, but does not formally close the identifiability gap for purely text-supported edges.

References

Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. 2019. Learning high-dimensional directed acyclic graphs with mixed data-types. In *Proceedings of the 2019 AAAI Workshop on Explainable AI*.

Joshua Angrist, Daniel Lang, and Philip Oreopoulos. 2009. Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–163.

Aahlad Ankan and Johannes Textor. 2025. Expert-in-the-loop causal discovery. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 286 of *Proceedings of Machine Learning Research*, pages 172–183.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*.

Taehyun Ban and 1 others. 2023. Human-in-the-loop causal discovery under latent confounding. *arXiv preprint arXiv:2309.12032*.

Victor Barres, Clifton McFate, Aditya Kalyanpur, Kailash Karthik Saravanakumar, Lori Moon, Nati Seifu, and Abraham Bautista-Castillo. 2025. From generating answers to building explanations: Integrating multi-round RAG and causal modeling for scientific QA. In *Proceedings of NAACL*.

Nicolás Bonilla, Shenyang Zhao, and Daniel Steinberg. 2026. Causal preference elicitation. *Preprint*, arXiv:2602.01483.

Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *Quarterly Journal of Economics*.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Diego Colombo and Marloes H Maathuis. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782.

Susan Dynarski. 2003. Does aid matter? measuring the effect of student aid on college attendance and completion. *American Economic Review*, 93(1):279–288.

Susan Dynarski, Aizat Nurshatayeva, Lindsay C Page, and Judith Scott-Clayton. 2023. Addressing nonfinancial barriers to college access and success: Evidence and policy implications. In *Handbook of the Economics of Education*, volume 6, pages 319–403. Elsevier.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A GraphRAG approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Elias Eulig, Atalanti A. Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. 2025. Toward falsifying causal graphs using a permutation-based

test. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 26778–26786.

Jiannan Feng. 2024. Introduction to causal graphs for education researchers. *Asia Pacific Education Review*.

Bhavya Ghai and Klaus Mueller. 2022. D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias. In *IEEE Transactions on Visualization and Computer Graphics*, volume 29, pages 473–482.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Oktie Hassanzadeh. 2024. WikiCausal: Corpus and evaluation framework for causal knowledge graph construction. In *Posters, Demos, and Industry Tracks at the 23rd International Semantic Web Conference (ISWC 2024)*, volume 3828 of *CEUR Workshop Proceedings*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Estelle Herbaut. 2020. What works to reduce inequalities in higher education? a systematic review. *Economics of Education Review*.

Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55.

Steven Ingels, Daniel Pratt, and James Rogers. 2011. Education longitudinal study of 2002 (els:2002): Base-year to third follow-up data file documentation. Technical report, National Center for Education Statistics.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. 2022. Causal identification under markov equivalence: Calculus, algorithm, and completeness. In *Advances in Neural Information Processing Systems*, volume 35.

Brian Jacob and Tamara Wilder. 2011. Educational expectations and attainment. *Journal of Human Resources*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023a. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7969–7992.

Elahe Khatibi, Ziyu Wang, and Amir M. Rahmani. 2025. CDF-RAG: Causal dynamic feedback for adaptive retrieval-augmented generation. *Preprint*, arXiv:2504.12560.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. arxiv 2023. *arXiv preprint arXiv:2305.00050*.

Matthew A Kraft, William H Marinell, and Darrick Shen-Wei Yee. 2016. School organizational contexts, teacher turnover, and student achievement: Evidence from panel data. *American Educational Research Journal*, 53(5):1411–1449.

George D Kuh, Ty M Cruce, Rick Shoup, Jillian Kinzie, and Robert M Gonyea. 2008. Unmasking the effects of student engagement on first-year college grades and persistence. *The journal of higher education*, 79(5):540–563.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Hao Liu, Zhiyuan Liu, and Maosong Sun. 2025a. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation. *Preprint*, arXiv:2502.12442.

Yuhan Liu, Qiongjiang Song, and Qiongxuan Zheng. 2025b. A systematic review and meta-analysis of the effects of need-based aid on students' academic achievement: insights from causal evidence. *Oxford Review of Education*, pages 1–28.

Georgy Meshcheryakov and Anna A Igolkina. 2021. semopy: A python package for structural equation modeling. *Human Heredity*, 86(4):236–240.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2 edition. Cambridge University Press.

Tuan Nguyen, Jeff Kramer, and Brendan Evans. 2019. The effects of grant aid on student persistence and degree attainment: A meta-analysis. *Review of Educational Research*.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. 2018. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62.

Laura Perna. 2006. Studying college access and choice: A proposed conceptual model. *Higher Education: Handbook of Theory and Research*.

Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance (SID) for evaluating causal graphs. volume 27, pages 771–799.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129.

Bruce Sacerdote. 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far? *Handbook of the Economics of Education*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. MIT Press.

Vincent Tinto. 1993. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press.

Bence Tóth, David Martínez-Rubio, Quentin Berthet, Arthur Gretton, Siddharth Bhatt, and Dino Sejdinovic. 2023. Active bayesian causal inference. *Preprint*, arXiv:2309.12032.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. Causalrag: Integrating causal graphs into retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22680–22693.

Yifan Wu and 1 others. 2022. A survey of human-in-the-loop machine learning. *ACM Computing Surveys*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *Preprint*, arXiv:2401.15884.

Zheng-Xin Yong, Zhaocheng Zhu, Ehsan Nezhad, and 1 others. 2025. CauScientist: Teaching LLMs to reason like causal scientists. *Preprint*, arXiv:2601.13614.

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*.

Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16–17):1873–1896.

Xingbo Zhang, Sai Kota, Eric Papenhausen, and Klaus Mueller. 2025. CausalChat: Interactive causal model development and refinement using large language models. *IEEE Transactions on Visualization and Computer Graphics*. ArXiv:2410.14146.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024. Causal-learn: Causal discovery in Python. *Journal of Machine Learning Research*, 25(60):1–8. ArXiv:2307.16405.

Xiaofan Zhou, Liangjie Huang, Pinyang Cheng, Wenpen Yin, Rui Zhang, Wenrui Hao, and Lu Cheng. 2025. Accelerating causal network discovery of alzheimer disease biomarkers via scientific literature-based retrieval augmented generation. In *IEEE BigData*.

A Additional Implementation Details

A.1 Dataset and Variable Construction

ELS:2002 contains 4,012 variables across four waves (10th grade 2002; 12th grade 2004; post-secondary entry 2006; long-run outcomes 2012). We map these to 22 concept nodes following four principles: (i) *theoretical coherence* — all variables within a node measure the same construct; (ii) *wave alignment* — all variables come from the same longitudinal wave, preserving temporal causal ordering; (iii) *intervention relevance* — nodes correspond to constructs meaningful in educational policy and advising; and (iv) *practical coverage* — nodes with $> 80\%$ missing values are excluded. Continuous and ordinal variables are z-scored and averaged into composites; binary variables are aggregated as proportions. After listwise deletion the analysis set contains n=9,084 complete cases. The node design was validated by the domain expert prior to causal discovery.

A.2 ELS Concept Node Mapping

Table 4 lists the 22 concept nodes by wave, together with the primary ELS variable codes that compose each node.

A.3 Stage 1 — Ground: Causal Discovery Details

All variables are standardized prior to causal discovery. FGES, PC-stable, and FCI are implemented via `causal-learn` (Zheng et al., 2024), with temporal ordering enforced as forbidden-edge background knowledge. FGES uses a Conditional Gaussian BIC score (Andrews et al., 2019); PC-stable and FCI use Fisher's Z at α=0.05 with Benjamini–Hochberg FDR correction (Colombo and Maathuis, 2014). The three algorithms produce the following directed edge counts before majority voting: FGES: 54; PC-stable: 66; FCI: 52. The $\geq 2/3$ majority vote yields $|E_{\mathrm{HC}}|$=56 high-confidence edges. FCI contributes 16 orientations not recoverable by FGES and PC-stable alone, reflecting its ability to orient edges under latent confounding (Spirtes et al., 2000; Zhang, 2008).

Effect strengths are estimated via a linear structural equation model. For each node $v_i \in V$,

$$v_i = \sum_{v_j \in \mathrm{Pa}(v_i)} \beta_{ji} v_j + \varepsilon_i, \tag{4}$$

where $\mathrm{Pa}(v_i)$ denotes the parent set of v_i in G_{data} and β_{ji} is the standardized effect of v_j on v_i. A linear SEM is fit via FIML using `semopy` (Meshcheryakov and Igolkina, 2021) on all n=9,084 observations; model fit is acceptable (CFI=0.935, RMSEA=0.055) (Hu and Bentler, 1999). Edge-level temporal validity is additionally checked via a separate 30% held-out OLS regression: of 35 cross-wave edges in E_{HC}, 32 are significant (p<0.05) on the held-out split; the remaining three are carried forward as weak-evidence edges.

A.4 Stage 2 — Extract: Text Retrieval and Causal Extraction

Documents are chunked into 512-token windows (64-token overlap), embedded using `text-embedding-3-small`, and indexed in ChromaDB; top-k=10 chunks are retrieved per query. Corpus support scores (Eq. 3) are computed with DeBERTa-v3-large (He et al., 2021) fine-tuned on MultiNLI and SNLI; edges with $s_{ij}<\tau$ are excluded from G_{text}. The pipeline uses an operational threshold τ=0.03 to construct G_{text} (Eq. 3); this is the heuristic value used during initial discovery and is the threshold reflected in all G_{text} counts in the main paper. Post-hoc calibration against expert verdicts across all three reconciliation rounds §A.9 identifies τ=0.07 as F1-optimal (F1 = 0.892, vs. F1 = 0.851 at τ=0.03). A permutation-based fake-edge falsification test (Eulig et al., 2025) establishes a conservative upper bound at the 95th percentile of implausible-edge scores (τ_{fake}=0.1288), used internally for audit but not as the filtering threshold. Mistral-7B-Instruct-v0.2 (Jiang et al., 2023a) is used for causal extraction via the CD-RAG pipeline of Zhou et al. (2025).

Multi-perspective prompt routing: P2 (reverse-direction) is applied to conflict edges and text-only edges (edges in G_{text} but not in G_{data}); P3 (confounder) to E_{HC} edges flagged as causally paradoxical with zero text-grounding score; P4 (mediator) to E_{HC} edges that are non-significant in cross-wave validation.

Corpus sources: NBER, ERIC, Brookings, Blueprint Labs, CEPA Stanford, arXiv, and Europe PMC. Each candidate edge is assigned a structured status label: `confirmed_direct`, `bidirectional_conflict`, `potentially_confounded`, `confirmed_mediated`, or `not_confirmed`. These labels drive alignment-stage routing and priority scoring.

Wave	Node	Primary ELS variables
1 (2002)	SES_background	BYSES1, BYSES2, BYFAMINC
	Race_ethnicity†	BYRACE, BYRACED
	First_gen_status†	BYPARED, BYMOTHED, BYFATHED
	School_context	BYSCTRL, BYURBAN, BY10FLP, BYA23K
	Academic_achievement_W1	BYTXMSTD, BYTXRSTD, BYGPAVER
	Ed_expectations_W1	BYSTEXP, BYSTLOCS
	Parental_aspirations	BYPARASP
	School_engagement_W1	BYEXTCUR, BYATTNDR, BYHMWRK
	Financial_resources_W1	BYINCOME, BYOWNHOME
2 (2004)	Academic_preparation_W2	F1TXMSTD, F1GRADE, F1HIMATH
	College_aspirations_W2	F1STEXP
	PSE_intentions_W2	F1PSEPLN, F1COLINF, F1ENRFIN
	Employment_W2	F1WRKHRS, F1COMPHR
	Financial_aid_W2	F1S52B
3 (2006)	Financial_aid_W3	F2PS1AID, F2PS1GRT, F2PELLCUM
	PSE_enrollment	F2EVRATT, F2ENRL06
	Employment_W3	F2EVRJOB
	Academic_standing_W3	F2EDLEVL, F2HIGRDE
	Credential_attainment	F2EVRGED, F2HSSTAT
	Postsecondary_sector	F2PS1LVL, F2PS1SEC, F2SECTOR
4 (2012)	Degree_completion_F3	F3ATTAINMENT, F3EDSTAT, F3TZANYDEGRE
	Labor_market_F3	F3EMPSTAT, F3JUNEMPSTAT, F3JOBPERSIST

Table 4: The 22 concept nodes by wave with primary ELS variable codes. Nodes marked † are excluded from intervention targets (immutable attributes).

A.5 Corpus Document List

The full list of 101 documents used to construct $\mathcal{T}$ (58 full-text PDFs and 43 abstract-only sources), including titles and retrieval dates, is available in the anonymized repository at https://anonymous.4open.science/r/HICAR-self-referential-RAG/.

A.6 Cross-Modal Edge Taxonomy

Table 5 defines the five categories used to partition directed candidate edges in HICAR.

A.7 Stage 3 — Iterative Validation Protocol

Graph operations. Expert responses are encoded as one of five formal graph operations: CONFIRM retains an edge as validated; REVERSE changes its direction; REMOVE deletes an unsupported or non-causal edge; RELABEL marks an edge as mediated, confounded, or context-specific; and UNCERTAIN preserves the edge with explicit unresolved status. These operations make the effect of expert feedback transparent and auditable.

Identifiability criteria. For validated edges where the graph structure admits a valid adjustment set, causal effects are identified using the backdoor criterion (Pearl, 2009). Under possible latent confounding, we apply generalized adjustment criteria (Perković et al., 2018). Edges whose effects cannot be identified are retained as causal hypotheses and flagged for additional validation.

Priority score. Edges are ranked for expert review by a priority score:

$$\pi(v_i \to v_j) = w_1(1-s_{ij}) + w_2\, c_{ij} + w_3(1-e_{ij}) + w_4(1-u_{ij}), \quad (5)$$

where s_{ij} is the corpus support score, $c_{ij} \in \{0,1\}$ indicates directional conflict, $e_{ij} \in \{0,1\}$ indicates support from G_{data}, and $u_{ij} \in \{0,1\}$ indicates structural uncertainty. Priority score weights are set to w_1=0.40, w_2=0.25, w_3=0.20, w_4=0.15. Conflict edges receive π=1.0 regardless of other scores.

Review protocol. Review proceeds iteratively: the initial pass covers agreement, conflict, and text-only edges from Stage 2 (Extract); subsequent passes incorporate edges surfaced by the Stage 3a (Query) typed re-retrieval. Structured queries are generated automatically per edge taxonomy category and presented to the domain expert in priority order.

A.8 Unguided Baseline: Full Breakdown

We compare G_{data}-anchored retrieval against an unguided Mistral free-extraction baseline that pro-

Category	Definition	Interpretation
AGREEMENT	$(v_i \to v_j) \in E_{\text{data}} \cap E_{\text{text}}$, same direction	Supported by both statistical and textual evidence
TEXT-ONLY	$(v_i \to v_j) \in E_{\text{text}}, (v_i \to v_j) \notin E_{\text{data}}$	Literature-supported but not statistically recovered
DATA-ONLY	$(v_i \to v_j) \in E_{\text{data}}, (v_i \to v_j) \notin E_{\text{text}}$	Statistically supported but not found in retrieved literature
CONFLICT	Data and text support opposite directions	Directional ambiguity requiring reconciliation
CAUSALLY AMBIGUOUS	Evidence suggests possible confounding or mediation	Relationship may be indirect, contextual, or non-causal

Table 5: Cross-modal edge taxonomy in HICAR. Categories partition the set of directed candidate edges and determine routing through Stage 3b (Validate).

poses candidate edges without E_{HC} injection. Anchored retrieval restricts NLI scoring to the 56 E_{HC} pairs — an 87.9% reduction of the 462-pair all-directed-pairs space — while the unguided extractor scores all 68 Mistral-proposed candidates without this constraint. Measured against FGES (54 edges) at the operational threshold τ=0.03, anchored precision is 52.9% (9/17 above-threshold edges in G_{data}) versus 20.6% for unguided (14/68), a +32.4 percentage-point gain. A post-hoc audit of the 13 unguided edges above the recalibrated threshold τ=0.1288 reveals that 2 of 13 are *direction-blind NLI artifacts* — the forward and reverse directions of the same pair scoring within 0.002 of each other. Anchored retrieval eliminates this class of artifact entirely by design: no anchored candidate pair has a competing reverse-direction candidate above threshold, because E_{HC} encodes statistically grounded orientation.

A.9 NLI Threshold Calibration

We calibrate the NLI entailment threshold τ using expert verdicts from all three reconciliation rounds as a ground-truth set, sweeping τ from 0.01 to 0.30. The expert-optimal threshold is τ=0.07 (F1 = 0.892), outperforming the heuristic τ=0.03 (F1 = 0.851). We report results at τ=0.07 in the body. The production pipeline filters G_{text} at the operational threshold τ=0.03; a permutation-based fake-edge falsification test (Eulig et al., 2025) establishes a conservative upper bound at the 95th percentile of implausible-edge scores (τ_{fake}=0.1288), used for internal audit but not as the filtering threshold.

A.10 Directional Agreement and Identifiability

We evaluate directional agreement between G_{data} and G_{text} for the 19 shared skeleton edges using a Structural Intervention Distance-inspired metric (Peters and Bühlmann, 2015). Cross-wave edges achieve **100%** directional agreement (11/11): temporal ordering established by ELS:2002's longitudinal design ensures both sources agree on direction when an earlier wave precedes a later wave. Within-wave edges achieve 37.5% agreement (3/8), consistent with genuine simultaneity in variables measured within the same wave. Overall directional agreement is **73.7%** (14/19).

This finding motivates a two-tier identifiability table for $G^\star$. Of 55 edges, 36 are cross-wave and formally identified by temporal ordering; 19 are within-wave (all in Wave 1), of which 11 are annotated as plausibly sequential or simultaneous. The remaining 8 within-wave edges are cross-modal agreement edges between Wave-1 social-background variables that were not part of the expert's R2 annotation batch; they are retained in $G^\star$ on the strength of cross-modal agreement alone.

We further note that 8 simultaneous-direction edges between social-background variables were annotated by the expert in R2 but flagged as wrong-direction in a separate Surprising Findings note; these were removed from $G^\star$ during R2 reconciliation and are not part of the 19 within-wave edges above.

A.11 Validated Causal Graph $G^\star$

Figure 2 shows the full validated causal graph $G^\star$ on the 22 ELS:2002 concept nodes after all three reconciliation rounds. Edges are colored by source: cross-modally agreeing edges ($|E_\cap|$=15, teal); edges added during Rounds 1 and 2 of expert reconciliation (n=33, blue); and edges added in Round 3 (n=7, red). The dashed gold arrows highlight the mediation hypothesis surfaced by the domain expert's Round 3 rejection of School Context → Financial Aid: the more plausible path runs School Context → PSE Enrollment → Financial Aid (§4.4). Light gray arrows show the full high-confidence data layer E_{HC} for context; edges that are in E_{HC} but did not survive

cross-modal alignment or expert review are not part of $G^\star$.

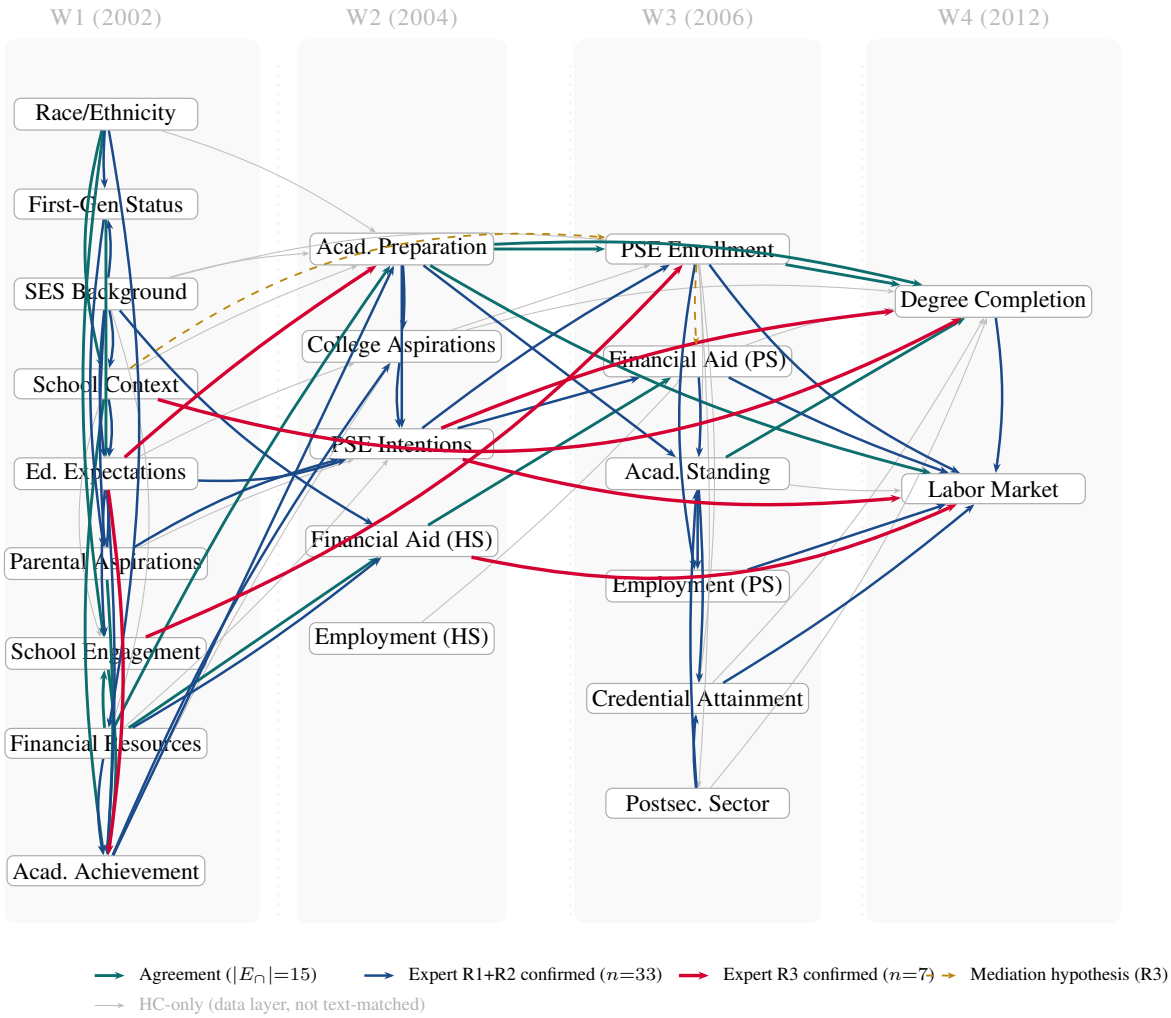

Figure 2: Validated causal graph $G^\star$ on ELS concept nodes (55 edges across four waves W1–W4). Teal: cross-modal agreement edges ($|E_\cap|$=15); blue: expert-confirmed in Rounds 1–2 (n=33); red: expert-confirmed in Round 3 (n=7); dashed gold: mediation hypothesis surfaced by Round 3 expert rejection of School Context → Financial Aid, suggesting the indirect path through PSE Enrollment; light gray: full high-confidence data layer E_{HC} shown for reference. Cross-wave edges (36) are temporally identified by design; within-wave edges (19) are annotated as plausibly sequential or simultaneous (§A.10).