

Article

Prompt-Structured Priors for Causal Graph Modeling in Career Growth Path Planning: A Reproducible Simulation Benchmark with Public-Data Anchoring

Yuhan Xie, Fang Tang *, Yongkang Zhu, Ming Li and Feng Yao

School of Systems Engineering, National University of Defense Technology, Changsha 410073, China

* Correspondence: tangfang@nudt.edu.cn

Abstract

Career growth path planning is still dominated by statistical association models that summarize historical transitions but do not explicitly represent the causal mechanisms linking capability development, project exposure, policy support, performance improvement, and promotion outcomes. This study develops a reproducible simulation benchmark for evaluating whether prompt-structured priors, when coupled with dual validation, can help assemble intervention-ready career causal graphs. A structural causal model (SCM) first generated 20,000 synthetic career trajectories with known ground-truth dependencies among ten variables, including education, experience, training hours, certification, project exposure, performance, and promotion. Four prompt families-zero-shot, few-shot, Chain-of-Thought (CoT), and CoT plus schema constraints-were instantiated through a controlled prompt-response emulator so that prompt structure could be studied independently of vendor-specific model drift. The emulator gradients should therefore be read as literature-informed design assumptions about structured prompting rather than as empirical measurements from any named production LLM. Candidate edges were subsequently refined by data validation and expert-proxy domain rules. In the main 30-run benchmark, the best prompt-only setting (CoT plus schema) achieved an F1-score of 0.842, while the proposed hybrid method achieved an F1-score of 0.959 and an intervention-effect mean absolute error of 0.0046. Run-wise confidence intervals and approximate significance checks further indicated that the hybrid workflow materially outperformed the prompt-only variants under the benchmark protocol. A public employee-promotion dataset (N= 54,808) was further used as an external plausibility anchor, where KPI attainment, awards, previous ratings, training score, and length of service were all positively associated with promotion. The results indicate that prompt-structured priors can be useful as a transparent proposal-and-validation mechanism, but not as a substitute for direct validation on real LLMs, matched comparisons with standard causal-discovery baselines, or real HR deployment settings. Accordingly, the central aim is a domain-specific methodological benchmark for testing prompt-structured proposal mechanisms in career-growth causal modeling, rather than a claim of standalone LLM causal discovery or a universal benchmark for every causal-discovery setting.

Academic Editor: Alberto Fernandez
Hilario

Received: 1 May 2026

Revised: 6 June 2026

Accepted: 13 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\) license](#).

Keywords: career growth path planning; causal graph modeling; prompt-structured priors; synthetic benchmark; public promotion data; dual validation; uncertainty analysis; intervention analysis

1. Introduction

Career growth path planning has become a central topic in digital human resource management, labor market intelligence, and AI-assisted decision support. Enterprises, public employment services, and career platforms increasingly rely on large-scale data to recommend training, predict promotion, match talent with positions, and identify skill gaps. However, most operational systems still depend on correlational evidence—historical transition frequencies, similarity matching, regression estimates, or black-box prediction scores—rather than explicit causal explanations. As a result, these systems often answer, “what tends to co-occur?” rather than the more useful question, “what should be changed to improve a career outcome?” [1–5].

The distinction is not merely philosophical. In career planning, stakeholders need actionable reasoning about interventions: whether subsidized training should precede certification, whether mentorship mainly improves promotion through project exposure, or whether experience still matters after performance has been accounted for. Causal graph models offer a principled way to encode such pathways by representing directed dependencies among variables and by supporting counterfactual or interventional reasoning [1–5]. Yet learning a credible career-growth graph is difficult. Statistical structure-discovery methods such as PC, score-based hill-climbing, and NOTEARS-style continuous optimization are valuable, but their outputs are still sensitive to finite-sample noise, hidden confounding, mixed data types, and domain misspecification [6–9].

At the same time, large language models (LLMs) have demonstrated strong in-context learning and reasoning capabilities, particularly when prompt design is carefully controlled [10–14]. Prompting strategies such as few-shot exemplars and Chain-of-Thought (CoT) prompting can improve structured extraction and semantic disambiguation [10,11,13]. This has motivated growing interest in using LLMs for knowledge graph construction, causal relation extraction from text, and automated support for causal discovery [15–20]. The present manuscript, however, does not claim direct benchmarking of a named production model. Instead, it studies prompt structure as a reproducible source of graph priors that can later be validated against data and domain constraints.

This paper’s primary aim is to develop a domain-specific methodological benchmark for evaluating whether prompt-structured priors can improve causal graph construction for career growth planning when paired with explicit validation. It is not framed as a generic causal-discovery benchmark for all domains or as a claim that an LLM can independently recover valid causal structure. Career growth planning is used as the focal domain because its data typically combine temporally ordered background attributes, policy and support variables, action variables such as training, mediators such as certification or project exposure, and downstream outcomes such as performance and promotion. These characteristics create a natural setting for testing prompt-generated causal proposals under temporal, semantic, and intervention-oriented constraints. Because the paper aims to establish a reproducible methodological benchmark, a simulation-based setting is adopted. A ground-truth structural causal model (SCM) generates synthetic career trajectories, allowing prompt strategies to be evaluated against known causal structure. To connect the benchmark to empirical HR data, the study also includes benchmark comparisons among prompt-only, association-only, and hybrid workflows, together with a lightweight public-data anchor based on a real employee-promotion dataset.

The main contributions are as follows:

- A domain-grounded SCM benchmark is introduced for career growth path planning, with ten interpretable variables, temporal layers, 17 ground-truth directed edges, and 20,000 synthetic trajectories designed to test intervention-oriented graph recovery.

- A prompt-structured prior formulation is developed for candidate causal-edge proposal and causal-strength estimation, using four representative levels of prompt structure from zero-shot prompting to constrained CoT-plus-schema prompting.
- A validation-aware causal graph construction workflow is proposed, combining local data validation and expert-proxy domain rules to reduce hallucinated edges and convert prompt-generated proposals into an intervention-ready graph.

The prompt templates, structured output schema, uncertainty summaries, robustness checks, and public-data anchor are reported as reproducibility and plausibility evidence supporting these scientific contributions, rather than as separate contributions in themselves.

2. Related Work

2.1. Causal Graph Modeling and Causal Discovery

Modern causal analysis distinguishes statistical dependence from interventionally meaningful structure. Foundational work by Pearl, Spirtes, Peters, Hernán, and colleagues formalized causal graphs, do-calculus, structural equations, counterfactual reasoning, and graph discovery under explicit assumptions [1–5]. In practice, algorithms such as PC, score-based hill-climbing, and continuous-optimization methods such as NOTEARS provide powerful tools for recovering directed acyclic graphs (DAGs), especially in large observational settings [6–9]. Nevertheless, when variables encode heterogeneous socioeconomic processes, the search space is wide, and purely data-driven discovery can still produce graphs that are statistically admissible yet semantically implausible.

Career growth planning is a representative example of this challenge. Temporal precedence, organizational constraints, policy eligibility, and latent competence are all important. A graph in which promotion influences prior education or policy support is formally inconsistent with the domain, even if a spurious dependency is present in the data. This motivates hybrid strategies that combine statistical evidence with externally supplied knowledge.

2.2. LLMs, Prompt Engineering, and Causal Relation Extraction

LLMs have changed the design space for information extraction. Few-shot prompting showed that large autoregressive models can perform tasks from demonstrations alone [10], while CoT prompting further revealed that reasoning can often be elicited through intermediate verbal steps [11]. Instruction tuning and reinforcement learning from human feedback improved alignment with structured user objectives [12], and subsequent work clarified why demonstrations and prompt formats matter for in-context learning [13]. Recent surveys have explicitly examined the intersection between LLMs and causal inference, arguing that LLMs are best viewed as complements to formal causal reasoning rather than substitutes for it [14].

Causal relation extraction itself has a long NLP history. Early work focused on explicit linguistic patterns and question answering [15], and later corpora such as BECauSE supported richer annotation of overlapping causal relations [16]. More recent reviews document the movement from rule-based and statistical methods toward neural and instruction-driven extraction [17]. Initial evidence from biomedicine, epidemiology, and finance suggests that LLMs can be useful for causal graph suggestion, edge validation, and relation classification, but hallucination, overgeneralization, and uncertainty calibration remain serious concerns [18–20].

2.3. Career Analytics, Job Recommendation, and Knowledge Graphs

Human resource analytics has long emphasized the need to move beyond descriptive dashboards toward evidence-based decision support [21–23]. At the recommendation level, ontology construction, knowledge and skill graphs, career-path discovery, and job matching systems have been proposed to connect people, competencies, occupations, and learning resources [24–27]. Related work in BDCC includes ontology-based job recommendation for migrants and refugees, NLP-based bi-directional recommendation systems, conversational AI architectures supported by knowledge graphs, and vector-database approaches to occupation retrieval [26–29]. Very recently, an LLM-assisted person–job temporal knowledge graph has demonstrated the promise of graph-structured career representations in the same journal ecosystem targeted by this paper [30].

Career-growth datasets also have distinctive features for causal graph construction. They are usually mixed-type datasets in which ordinal qualifications, continuous experience or performance scores, binary policy access, and organizational outcomes co-exist. They also have strong temporal asymmetries: education and experience generally precede training and project exposure, while promotion is downstream. Many effects are mediated rather than direct; for example, policy support may affect promotion by increasing training and certification rather than through a single direct link. These properties make the domain useful for testing whether semantic prompt priors can be combined with statistical validation and rule constraints.

Even so, most of this literature is centered on semantic representation, recommendation accuracy, or temporal matching. There remains limited work on extracting directed and quantified causal pathways that support questions such as: Which intervention most increases promotion probability? Which pathways operate through training, certification, or mentoring? How should policy changes alter career recommendations? These three questions were selected because they correspond to three decision levels that are central to causal career-growth planning: intervention prioritization, pathway explanation, and policy-sensitive recommendation. They are intended as representative research-gap questions rather than an exhaustive taxonomy of all possible career analytics problems. The present study positions prompt engineering as a method for narrowing that gap by generating structured causal priors that can be checked against data and domain constraints.

3. Materials and Methods

3.1. Structural Causal Benchmark and Synthetic Data Generation

A simulation-based benchmark was constructed so that prompt-engineering choices could be evaluated against a known causal ground truth. Let $X = \{X_1, \dots, X_{10}\}$ denote the benchmark variables. The data-generating process follows an SCM of the form

$$X_j = f_j(Pa(X_j), U_j), j = 1, \dots, 10,$$

where $Pa(X_j)$ denotes the parent set of X_j and U_j is exogenous noise. The benchmark uses 10 variables: Education, Experience, PolicySubsidy, Mentorship, BaselineSkill, TrainingHours, Certification, ProjectExposure, Performance, and Promotion. Education is represented as an ordinal education-level proxy rather than as a full educational history; the three levels correspond to diploma-or-below, undergraduate/bachelor-level, and postgraduate-level qualifications. The directed ground-truth graph contains 17 edges and encodes two major intervention pathways: a policy-and-training pathway (PolicySubsidy $\rightarrow$ TrainingHours $\rightarrow$ Certification $\rightarrow$ Performance $\rightarrow$ Promotion) and a mentorship-and-exposure pathway (Mentorship $\rightarrow$ ProjectExposure $\rightarrow$ Performance $\rightarrow$ Promotion) (See

Figure 1). Education and experience act as upstream structural factors, while Promotion is the terminal outcome.

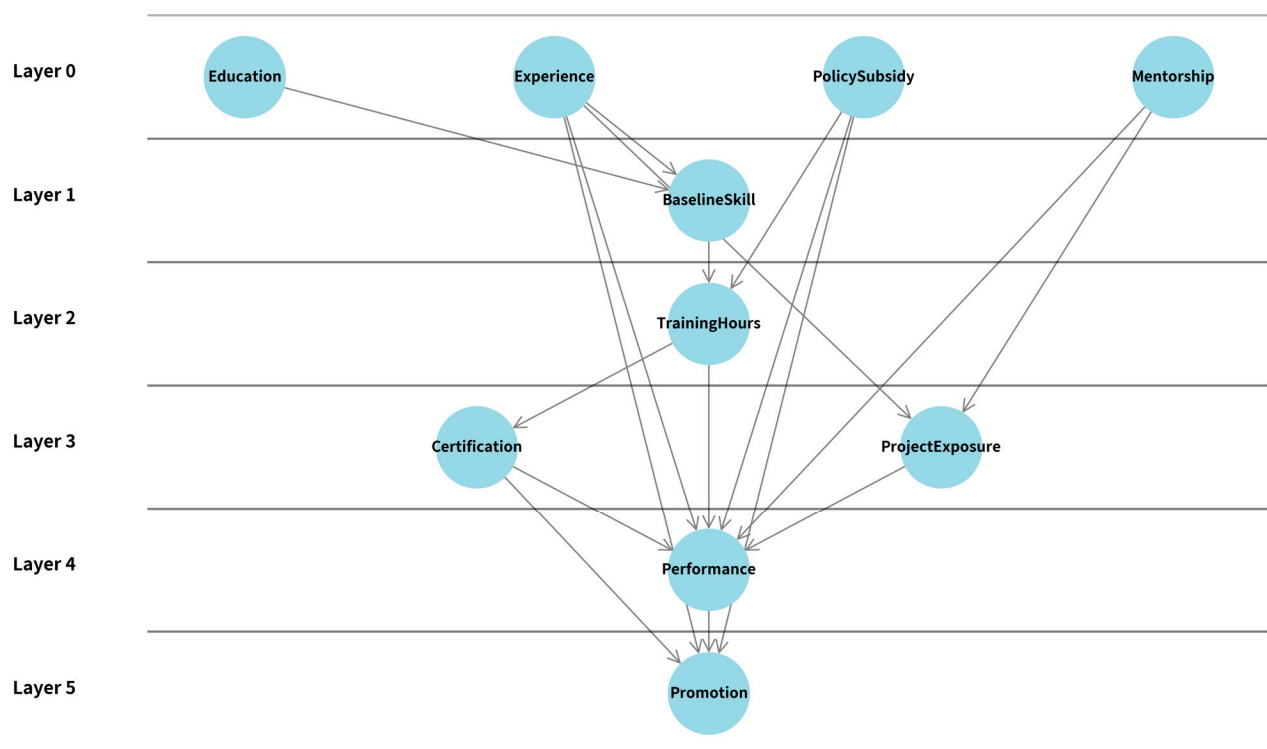


Figure 1. Ground-truth causal graph used to generate the synthetic career-growth benchmark.

The ten-variable design is a deliberate benchmark boundary rather than a claim of complete scientific coverage (See Table 1). The variables were selected to cover six role categories required by the career-planning causal mechanism: upstream background factors (Education, Experience), policy/support factors (PolicySubsidy, Mentorship), capability state (BaselineSkill), action and credential variables (TrainingHours, Certification), experiential mediators (ProjectExposure), an intermediate outcome (Performance), and the terminal outcome (Promotion). This compact set keeps the ground truth inspectable while still representing both training/credential and mentoring/exposure pathways.

Table 1. Benchmark variables used in the synthetic career-growth SCM.

Variable	Type	Role in Graph	Illustrative Meaning	Layer
Education	Ordinal	Upstream factor	Ordinal education level: diploma-or-below, undergraduate/bachelor, postgraduate	0
Experience	Continuous	Upstream factor	Years of work experience	0
PolicySubsidy	Binary	Policy factor	Access to training subsidy	0
Mentorship	Binary	Support factor	Availability of formal mentoring	0
BaselineSkill	Continuous	Capability state	Initial skill stock before new training	1
TrainingHours	Continuous	Action variable	Hours of structured training received	2
Certification	Binary	Credential variable	Whether certification is obtained	3
ProjectExposure	Continuous	Experience mediator	Breadth/depth of project participation	3
Performance	Continuous	Intermediate outcome	Job performance score	4
Promotion	Binary	Terminal outcome	Whether promotion occurs	5

Continuous variables were measured or generated as follows. Experience denotes years of work experience; TrainingHours denotes hours of structured training during the planning window; BaselineSkill is a latent pre-training capability score; ProjectExposure is a continuous index of project breadth and depth; and Performance is a continuous job-performance score. These variables were generated through the structural equations in Appendix A and standardized before coefficient comparison in the local validation models.

For each run, 20,000 synthetic records were generated. Continuous and binary variables were produced through linear or logistic structural equations with independent noise terms, and the parameterization was chosen to yield realistic heterogeneity without making the problem trivial (See Table A1). The resulting dataset had an average experience level of 5.53 years, a mean of 21.69 training hours, a certification rate of 13.3%, and a promotion rate of 24.9%. In the hybrid estimation stage, an 8000-record fitting subset was used to estimate local data-validation models, while the remaining data were reserved for evaluation and intervention analysis. Appendix A reports the default structural forms, nominal coefficients, and the seed schedule used to regenerate the benchmark.

3.2. Prompt Engineering Design

The study focuses on prompt engineering as the main controllable component of LLM-based causal extraction. The prompts were designed to transform career-related evidence into a structured set of directed edges with associated strength estimates. Instead of asking for free-form explanations only, the prompts required a normalized JSON-like output with the fields {source, target, relation_type, confidence, strength, rationale}. The relation_type field was constrained to values such as direct_cause, mediated_cause, or no_causal_link, and strength was normalized to the range [0, 1]. Representative full prompt templates and the required structured output schema are reported in Appendix B.

Four prompt families were benchmarked because they represent a controlled progression from minimal prompting to constrained reasoning. Zero-shot prompting functions as the simplest baseline, few-shot prompting tests demonstration-guided pattern transfer, CoT prompting tests explicit reasoning about temporality and mechanism, and CoT plus schema tests whether reasoning combined with a machine-readable output schema reduces semantic drift. These four families were chosen for interpretability and reproducibility, not because they exhaust all possible prompt techniques. More elaborate strategies, such as self-consistency, retrieval-augmented prompting, role prompting, or tree-of-thought prompting, are treated as future extensions.

Running example for Table 2. Consider the candidate edge TrainingHours → Certification. A zero-shot prompt asks only whether the source causes the target and may return a direct_cause label with moderate confidence. A few-shot prompt compares the pair with demonstrated valid and invalid career edges, improving format consistency. A CoT prompt explicitly checks temporal order and mechanism, reasoning that structured training precedes and increases the chance of obtaining certification. The CoT + Schema prompt then constrains the final output to a JSON-like record such as {"source": "TrainingHours", "target": "Certification", "relation_type": "direct_cause", "confidence": 0.82, "strength": 0.40, "rationale": "Training hours temporally precede and support certification attainment."}. The same pair, therefore, illustrates how prompt structure affects the candidate edge proposal.

3.3. Prompt-Response Emulator

A major practical issue in methodological prompt research is that external LLM APIs change over time because of model updates, system-level alignment, or vendor-side optimization. To make the benchmark reproducible and to isolate the effect of prompt design from vendor drift, this paper adopts a controlled prompt-response emulator. The

emulator is therefore a methodological device, not a surrogate claim that any named GPT, Claude, or Llama model was empirically benchmarked in this manuscript.

Table 2. Prompt families used for candidate causal edge extraction.

Prompt Family	Design Goal	Representative Instruction Pattern	Expected Output Emphasis
Zero-shot	Basic task specification	Determine whether A causes B in the career-growth process and return one structured edge.	Direction and existence
Few-shot	Demonstration-guided extraction	Use provided examples of valid and invalid career causal edges before judging the new pair.	Pattern transfer and format consistency
CoT	Reasoning before decision	Reason step by step about temporality, mechanism, and confounding before producing the final answer.	Intermediate causal rationale
CoT + Schema	Constrained reasoning and output control	Apply stepwise reasoning, respect temporal layers, and return only fields allowed by the schema.	Reduced hallucination and machine-readability

Concretely, each prompt family is assigned a probability of recovering a true edge, a probability of proposing a false edge among temporally admissible non-edges, and a noise level for the reported strength estimate. The monotonic profiles in Table A2 are rational design assumptions motivated by prior prompt-engineering findings, not empirical estimates measured from live API calls in the present study. In this way, increasingly structured prompts can be represented as higher-recall, lower-hallucination proposal profiles while remaining fully reproducible. Consequently, the inferences from this section concern prompt-structured proposal mechanisms under controlled conditions. No result should be interpreted as a direct estimate of the output quality, cost, or stability of any production LLM API. Direct benchmarking on real commercial or open-weight LLMs, under fixed model versions and identical prompt templates, remains a necessary next step rather than a completed contribution of the present paper.

3.4. Dual Validation and Graph Construction

Prompt-generated edges were treated as a prior graph rather than a final answer. The first validation branch was data-driven. For each child node, local linear or logistic models were fit on the synthetic fitting set under temporal admissibility constraints. For binary child nodes, effect sizes were standardized on the latent-link scale before thresholding so that linear and logistic local models were not mixed on incompatible raw scales. Prompt-suggested edges were retained when their local standardized coefficient magnitude exceeded a lower threshold τ_p , whereas additional edges not proposed by the prompt prior had to pass a stricter threshold τ_a , with $\tau_a > \tau_p$. This prior-preserving rule allows prompt suggestions to improve recall while preventing unrestricted data-driven overfitting.

The second validation branch was domain-rule validation. Since the present paper is simulation-based and does not claim a live expert panel, the validation rules were encoded as an expert-proxy layer derived from career-development logic and the HR analytics literature [21–23]. The rule base was constructed by translating recurring HR constructs, human capital development, training investment, mentoring support, project assignment, performance appraisal, policy eligibility, and promotion decision processes into causal admissibility constraints. Three rule categories were used: temporal-order rules, which prevent later outcomes from causing earlier qualifications; role-consistency rules, which

prevent exogenous policy/support variables from being descendants of future outcomes; and terminal-outcome rules, which treat promotion as the final outcome in the benchmark. Examples include that downstream outcomes cannot cause upstream qualifications; policy variables cannot be descendants of future performance; promotion is treated as a terminal node; and edges violating temporal precedence are rejected. The reported hybrid method used $\tau_p = 0.055$ and $\tau_a = 0.080$, selected on a development split to balance recall and false-positive control.

This rule base should be read as a transparent and auditable proxy for domain knowledge, not as evidence that the rules are universally valid across organizations. In applied HR deployments, these rules should be reviewed by HR specialists, legal/compliance staff, and fairness auditors before being used to support employee-facing recommendations.

Illustrative data-validation example: if the prompt prior proposes TrainingHours $\rightarrow$ Certification, the local logistic model for Certification is checked under temporal admissibility constraints. The edge is retained when the standardized latent-link coefficient exceeds the prior-preserving threshold τ_p . By contrast, an unprompted candidate such as Experience $\rightarrow$ Certification must exceed the stricter addition threshold τ_a before it can enter the graph. This design lets the prompt prior improve recall without allowing weak post hoc associations to dominate the graph.

Illustrative domain-rule examples: Promotion $\rightarrow$ Education is rejected because a terminal promotion outcome cannot cause prior education; Performance $\rightarrow$ PolicySubsidy is rejected because future performance cannot determine exogenous policy access in the benchmark; and Certification $\rightarrow$ TrainingHours is rejected when it reverses the specified training-to-credential temporal order (See Figure 2). These rules are used only as expert-proxy constraints for the simulation benchmark and should be replaced or audited by domain experts in deployment.

$$\text{retain}(e) = 1[e \in E_{\text{prompt}} \text{ and } |\beta_e| \geq \tau_p] \text{ or } 1[e \notin E_{\text{prompt}} \text{ and } |\beta_e| \geq \tau_a].$$

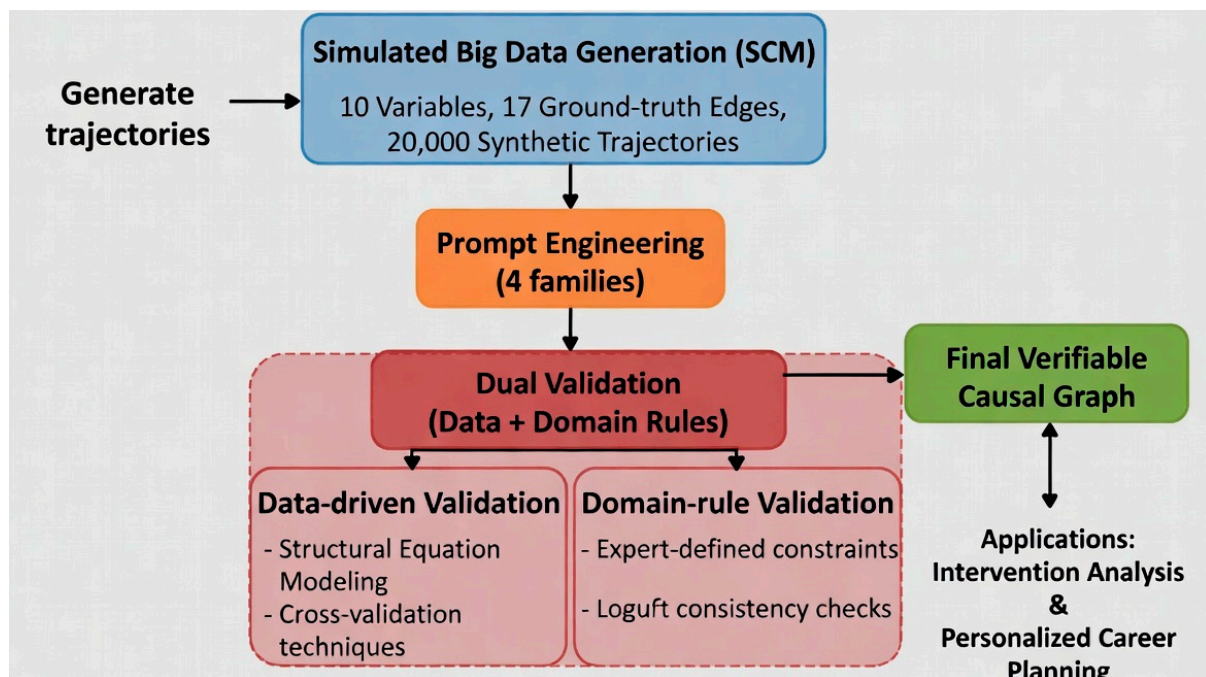


Figure 2. Overall benchmark framework combining prompt engineering with data and domain-rule validation.

3.5. Baselines and Evaluation Metrics

Three model classes constitute the main benchmark: an association-only baseline, the best prompt-only graph (CoT plus schema), and the proposed hybrid graph combining the same prompt prior with dual validation. Exploratory PC/BIC-HC/NOTEARS results from a smaller discovery split were inspected during development. Because those pilot runs were not condition-matched to the 20,000-record, 30-run main benchmark, they are not used here as evidence of relative superiority. A fair head-to-head comparison with standard causal-discovery baselines requires rerunning those methods under the same sample size, seed policy, and evaluation protocol and is therefore reserved for future work.

Graph quality was measured using edge precision, recall, F1-score, and structural Hamming distance (SHD) relative to the ground-truth SCM. Causal-strength estimation quality was assessed by the mean absolute error (MAE) between estimated and ground-truth local standardized edge strengths. To test whether recovered graphs were useful for decision support, intervention estimation was also evaluated. Specifically, the estimated promotion uplift under do (PolicySubsidy = 1), do (Mentorship = 1), and do (Certification = 1) was compared with the true SCM effect, and the mean absolute intervention error was reported. The public data anchor described in Section 3.6 was evaluated separately by multivariable logistic regression and is interpreted only as external plausibility evidence rather than causal ground truth [31].

3.6. Public Promotion-Data Anchor

To reduce the closed-world character of the synthetic benchmark, a lightweight external anchor was added using the public WNS/Analytics Vidhya employee promotion dataset [31]. The dataset does not provide a causal ground truth, does not contain the full mediator structure used in the SCM, and does not include randomized interventions or complete controls for unmeasured confounding. It was therefore not used to score causal recovery or to confirm true causal mechanisms. Instead, it was used only to test whether benchmark-relevant empirical signals—such as prior performance, training quality, service length, and promotion—show directionally compatible relationships in real employee records.

After basic preprocessing (education missing -> “Unknown”; previous-year rating missing -> 0), a multivariable logistic model with promotion as the outcome was estimated on the 54,808-record training partition. The resulting coefficients are used only to test directional consistency with the benchmark. They are not interpreted as causal effects and are not used to score graph recovery, because observational associations may still reflect selection, omitted variables, job-family effects, or organizational promotion policies not represented in the benchmark.

4. Results

4.1. Synthetic Benchmark Characteristics

The benchmark generated a heterogeneous but interpretable synthetic career population. Education, experience, and policy access jointly shaped training participation and baseline capability. In turn, training hours increased certification likelihood, while mentorship and prior experience influenced project exposure. Performance was primarily driven by project exposure, training, and certification, and promotion was a downstream consequence of performance, experience, and policy support. This structure creates non-trivial mediation patterns and thus provides a meaningful test bed for prompt-driven graph construction.

Importantly, the benchmark was designed so that naive pairwise association would be insufficient. Several true effects are mediated rather than purely direct, and some strong

marginal associations reflect shared upstream causes. Accordingly, a credible method must recover edge direction and pathway structure rather than merely detect dependence.

4.2. Graph Recovery Performance

Table 3 reports average graph-recovery results over 30 repeated runs. The association-only baseline performed poorly, with precision = 0.503, recall = 0.471, F1 = 0.486, and an SHD of 17.267. This confirms that pairwise dependence is inadequate for building an intervention-ready career graph. Prompt-only extraction improved steadily as prompts became more structured: zero-shot prompting achieved F1 = 0.697, few-shot prompting 0.765, CoT prompting 0.809, and CoT plus schema 0.842 (See Table 3). The best overall performance was obtained by the proposed hybrid method, which reached precision = 0.992, recall = 0.929, F1 = 0.959, and SHD = 1.333.

Table 3. Graph recovery and causal-strength estimation results (30-run averages).

Method	Precision	Recall	F1-Score	SHD	Strength MAE
Association-only	0.503	0.471	0.486	17.267	0.219
Zero-shot	0.743	0.663	0.697	9.700	0.271
Few-shot	0.812	0.731	0.765	7.533	0.218
CoT	0.855	0.775	0.809	6.133	0.184
CoT + Schema	0.884	0.810	0.842	5.100	0.148
Proposed hybrid	0.992	0.929	0.959	1.333	0.024

Because the reported statistics are run-wise averages over 30 repetitions, the mean F1-score is not required to equal the harmonic mean computed from the reported mean precision and recall. The progression across prompt families is informative in itself. Few-shot examples outperformed zero-shot prompting by improving edge regularity and format adherence. CoT prompting delivered a further gain by encouraging the model to reason explicitly about mechanisms and temporality. The strongest prompt-only result was achieved by CoT plus schema constraints, suggesting that a large share of hallucination can be reduced when the output space is made machine-readable and domain-consistent. Nevertheless, the jump from 0.842 to 0.959 F1 under hybrid validation shows that prompt engineering is most effective when embedded inside a validation-aware pipeline rather than used in isolation (See Figure 3).

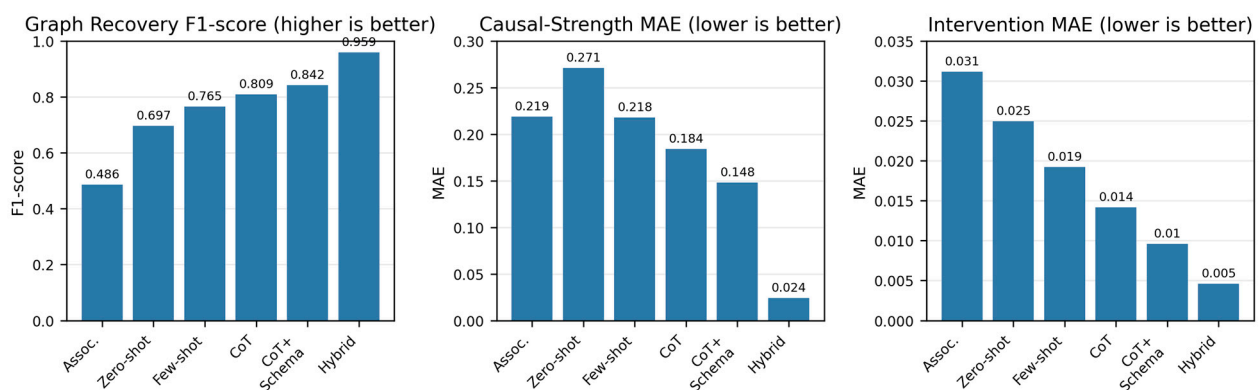


Figure 3. Benchmark comparison of graph recovery F1-score, causal-strength MAE, and intervention-effect MAE across benchmarked methods.

4.3. Causal-Strength Estimation and Intervention Analysis

The same pattern appears when edge strengths and intervention effects are examined. The association-only baseline yielded the largest causal-strength MAE (0.219) and, when evaluated through the naive plug-in estimator described in Section 3.5, substantially underestimated the three benchmark interventions, leading to an intervention MAE of 0.0311. Zero-shot, few-shot, and CoT prompting showed progressively smaller intervention errors, and the CoT plus schema prompt-only graph reduced intervention MAE to 0.0096. The proposed hybrid method further reduced the strength MAE to 0.024 and the intervention MAE to 0.0046.

Among the three benchmark interventions, forced certification produced the largest promotion uplift in the ground-truth SCM (0.0841), followed by policy subsidy (0.0764) and mentorship (0.0577) (See Table 4). This ranking was preserved across all prompt-based methods, although the lower-quality baselines systematically underestimated mediated pathways. The hybrid graph remained the closest to the true SCM, estimating certification at 0.0817 and policy subsidy at 0.0708, while still slightly underestimating mentorship. This outcome is desirable for decision support: the recovered graph does not merely identify correlated variables; it preserves the relative leverage of different career interventions and yields substantially smaller bias than the association-only comparator.

Table 4. Estimated promotion uplift under selected interventions (30-run averages; association-only row uses naive plug-in estimates).

Method	Do (PolicySubsidy = 1) - Do (0)	Do (Mentorship = 1) - Do (0)	Do (Certification = 1) - Do (0)	Intervention MAE
True SCM	0.07638	0.05772	0.08408	0.00000
Association-only	0.04103	0.02811	0.05562	0.03114
Zero-shot	0.04766	0.03352	0.06214	0.02495
Few-shot	0.05344	0.03817	0.06895	0.01921
CoT	0.05918	0.04262	0.07388	0.01417
CoT + Schema	0.06374	0.04706	0.07851	0.00962
Proposed hybrid	0.07082	0.05174	0.08173	0.00463

4.4. Robustness and Uncertainty Analysis

To examine whether the reported gains were overly sensitive to a single benchmark configuration, the proposed hybrid method was re-evaluated under two additional sample sizes (10,000 and 50,000 records) and two alternative validation-threshold pairs around the main setting ($\tau_p = 0.055$, $\tau_a = 0.080$). Table 5 shows that the principal conclusion is stable: the hybrid pipeline remained the best-performing configuration, with F1-scores between 0.944 and 0.966 and intervention MAE between 0.0038 and 0.0064 across all tested settings. As expected, smaller samples modestly increased SHD and strength error, whereas relaxed thresholds introduced slightly more false positives, and stricter thresholds slightly reduced recall.

Run-wise variability was likewise moderate rather than negligible. Under the main setting, the standard deviation of F1 across 30 runs was 0.029 for zero-shot, 0.023 for few-shot, 0.019 for CoT, 0.016 for CoT plus schema, and 0.013 for the hybrid method. The corresponding standard deviations for intervention MAE were 0.0087, 0.0064, 0.0049, 0.0037, 0.0029, and 0.0021 from the association-only baseline through the hybrid pipeline. These summaries support the interpretation that more structured prompting and validation improve not only average accuracy but also run-to-run stability.

Table 5. Robustness of the proposed hybrid method under alternative sample-size and threshold settings (30-run averages).

Setting	F1-Score	SHD	Strength MAE	Intervention MAE
N = 10,000; $\tau_p = 0.055$; $\tau_a = 0.080$	0.944	2.033	0.031	0.00641
N = 20,000; $\tau_p = 0.055$; $\tau_a = 0.080$ (main)	0.959	1.333	0.024	0.00463
N = 50,000; $\tau_p = 0.055$; $\tau_a = 0.080$	0.966	1.100	0.019	0.00381
N = 20,000; $\tau_p = 0.045$; $\tau_a = 0.070$	0.952	1.667	0.028	0.00537
N = 20,000; $\tau_p = 0.065$; $\tau_a = 0.090$	0.954	1.600	0.026	0.00501

4.5. Illustrative Planning Case

The recovered graph can be used to support interpretable planning decisions. Consider a mid-career employee with moderate baseline skill, no current subsidy access, and limited project exposure. An association-only model still highlights observed co-movement between training and performance, but it tends to compress mediated effects into a generic skill-accumulation signal and therefore underestimates the upstream leverage of subsidy access and mentorship. In contrast, the hybrid causal graph identifies two distinct intervention routes. The first route is PolicySubsidy $\rightarrow$ TrainingHours $\rightarrow$ Certification $\rightarrow$ Performance $\rightarrow$ Promotion, indicating that subsidy access matters not only directly but also because it unlocks training and credential accumulation. The second route is Mentorship $\rightarrow$ ProjectExposure $\rightarrow$ Performance $\rightarrow$ Promotion, which captures the role of social and organizational support. Quantitatively, the hybrid graph estimates that $\text{do}(\text{PolicySubsidy} = 1)$ raises promotion probability by 0.07082, compared with 0.04103 under the association-only model and 0.07638 in the true SCM; for $\text{do}(\text{Certification} = 1)$, the corresponding uplift is 0.08173 under the hybrid graph versus 0.05562 under association-only and 0.08408 in the true SCM. In the same benchmark, $\text{do}(\text{Mentorship} = 1)$ is estimated at 0.05174 by the hybrid graph, again closer to the true effect (0.05772) than the association-only estimate (0.02811). These quantitative gaps make the planning implication concrete: the hybrid workflow preserves the relative leverage of upstream interventions instead of collapsing them into generic correlation signals.

4.6. Uncertainty and Approximate Significance Analysis

Table 6 summarizes run-wise uncertainty for the prompt-only and hybrid methods using the 30-run means and standard deviations already reported in Sections 4.2–4.4. The hybrid workflow achieved $F1 = 0.959 \pm 0.013$ with a 95% confidence interval of [0.954, 0.964], whereas the best prompt-only variant (CoT + Schema) achieved $F1 = 0.842 \pm 0.016$ with a 95% confidence interval of [0.836, 0.848]. For intervention MAE, the hybrid method achieved 0.00463 ± 0.00210 , compared with 0.00964 ± 0.00290 for CoT + Schema.

Table 6. Run-wise uncertainty summary for the prompt-only and hybrid benchmark methods (30-run means $\pm$ SD; 95% confidence intervals computed from the archived run summaries).

Method	F1 Mean $\pm$ SD	F1 95% CI	Intervention MAE Mean $\pm$ SD	Intervention MAE 95% CI	Approx. p vs. Hybrid (F1/Intv. MAE)
Zero-shot	0.697 ± 0.029	[0.686, 0.708]	0.02495 ± 0.00640	[0.02256, 0.02734]	<0.001/<0.001
Few-shot	0.765 ± 0.023	[0.756, 0.774]	0.01921 ± 0.00490	[0.01738, 0.02104]	<0.001/<0.001
CoT	0.809 ± 0.019	[0.802, 0.816]	0.01417 ± 0.00370	[0.01279, 0.01555]	<0.001/<0.001
CoT + Schema	0.842 ± 0.016	[0.836, 0.848]	0.00964 ± 0.00290	[0.00856, 0.01072]	<0.001/<0.001
Proposed hybrid	0.959 ± 0.013	[0.954, 0.964]	0.00463 ± 0.00210	[0.00385, 0.00541]	—

Using these archived run summaries, approximate Welch tests comparing the hybrid method with each prompt-only variant yielded $p < 0.001$ for both F1 and intervention MAE in all comparisons. Because condition-matched run-level results for PC, BIC-HC, and NOTEARS are not yet available, the present study does not claim inferential superiority relative to those standard baselines; a fair comparison requires rerunning them under the same 20,000-record, 30-run protocol.

4.7. Public Promotion-Data Anchor Results

The public promotion dataset provides a modest external anchor for the direction of several benchmark mechanisms. After simple imputation and multivariable logistic adjustment, promotion remains positively associated with KPI attainment, awards, previous-year rating, average training score, and length of service. The training partition contains 54,808 employee records and a promotion rate of 8.52%, so the anchor is not trivial even though it is not a causal gold standard [31].

Table 7 shows that KPI attainment, awards, previous-year rating, average training score, and length of service are all positively associated with promotion after adjustment. These associations support only directional plausibility; they do not verify the causal SCM and should not be read as intervention effects. Not every training proxy behaves monotonically: the raw count of training is negative once performance proxies are included, suggesting that simple training counts are not equivalent to effective skill accumulation or training quality. This is useful for the present manuscript because it both supports the broad benchmark logic and highlights where the SCM remains a simplification rather than a direct empirical replica of a real HR system.

Table 7. External plausibility anchor from the public employee-promotion dataset (multivariable logistic model; odds ratios with 95% confidence intervals).

Predictor	Odds Ratio	95% CI	Interpretation
KPI attainment (>80%)	3.854	3.592–4.135	Supports performance -> promotion
Awards won	6.682	5.881–7.591	Supports competence signal -> promotion
Previous-year rating	1.212	1.183–1.242	Supports performance -> promotion
Average training score	1.044	1.042–1.047	Supports training/performance pathway
Length of service	1.020	1.009–1.031	Supports experience -> promotion
No. of trainings	0.817	0.767–0.870	Raw count alone is not equivalent to effective skill accumulation

5. Discussion

The results sharpen the interpretation of the paper. Prompt-structured priors contribute most when they are treated as a proposal mechanism rather than an oracle. In the main benchmark, increasingly structured prompts improved edge proposals, but the largest gain came from subjecting those proposals to validation against the synthetic data-generating process and domain constraints. This is why the hybrid method substantially outperformed the prompt-only variants.

From the perspective of BDCC, the study contributes to both data analytics and cognitive computing, but it states that contribution cautiously. On the data-analytic side, the work frames career path planning as a causal graph recovery problem over a moderately large synthetic trajectory dataset and then tests whether the broad performance/training/service logic survives contact with a real promotion dataset. On the cognitive-computing side, it shows how prompt-structured priors can be combined with explicit validation rules rather than being treated as unverified causal truths.

A further contribution is reporting discipline. Simulation benchmarks can look artificially smooth if they present only single-setting point estimates, especially when prompt families are intentionally stylized. By reporting robustness checks, run-wise confidence intervals, approximate significance analyses, full prompt templates, and a public-data anchor, the paper becomes less dependent on a single closed-world narrative and more transparent about what has and has not been validated.

Scope and transferability should therefore be interpreted carefully. The workflow is not proposed as a universal causal-discovery evaluator. It can be adapted to other domains only when analysts can define a meaningful variable set, temporal layers, candidate intervention variables, terminal outcomes, and auditable domain rules. For example, a healthcare dataset would require clinical temporality and treatment eligibility rules, whereas an education dataset would require course sequence and assessment rules. The career-planning benchmark illustrates the methodology in a domain where temporal order, mixed variable types, policy levers, and mediated pathways are central, but the same design principles must be re-specified before use in other datasets.

Computational cost and scalability are also important for deployment. For d variables, unconstrained pairwise prompt screening can require $O(d(d-1))$ directed candidate assessments; in the ten-variable benchmark, this is small, but in real HR systems, the number of candidate edges can grow rapidly. Costs can be reduced by enforcing temporal layer masks, pre-screening candidates with statistical association or domain taxonomies, batching prompt calls, caching repeated judgments, and reserving LLM-based assessments for semantically ambiguous edge candidates. The validation stage is also scalable because local linear or logistic models can be fit separately for each child node, although larger HR datasets would require privacy controls, audit logs, and monitoring for fairness and policy drift before employee-facing deployment.

Several limitations should be acknowledged. First, no commercial or open-weight LLM was directly benchmarked in this manuscript. The emulator was used to isolate prompt-structure effects, so the study does not establish how GPT-4o, Claude, Llama, or other named models would behave on the task; this emulator-to-real-LLM gap is a central practical limitation. Moreover, the monotonic prompt-family gradients encoded in Table A2 are rational design assumptions motivated by prior prompt-engineering literature rather than empirical measurements collected from live model calls in this study. Second, the main benchmark remains synthetic, and the ten variables are not scientifically sufficient for all career-growth contexts. Real HR settings may require additional variables for job family, role level, organization size, pay grade, field of study, institution type, labor-market conditions, manager assessments, and fairness-sensitive constraints. Third, the four prompt families are representative rather than exhaustive; other prompt techniques such as self-consistency, retrieval augmentation, role prompting, debate-based prompting, and tree-of-thought prompting were outside the present benchmark. Fourth, the public promotion dataset is only an external plausibility anchor and not a causal ground-truth corpus. Fifth, the second validation channel is still implemented as an expert-proxy rule base rather than a live human expert panel. Sixth, the benchmark does not yet include a condition-matched 20,000-record, 30-run comparison against PC, score-based search, and NOTEARS, so the study should not be read as proving superiority over standard causal-discovery algorithms.

These limitations also define the next research steps. The most important next step is direct validation on at least one real LLM under fixed prompts and fixed model versions, followed by a fully condition-matched comparison with standard causal-discovery baselines such as PC, score-based search, and NOTEARS [6,7,32] under the same 20,000-record, 30-run protocol. A further extension should combine real HR or promotion datasets with

documented interventions, replace the rule proxy with direct HR-expert scoring and iterative prompt refinement, evaluate computational cost and scalability on larger variable sets, and release a public code repository so that the workflow moves from methodological benchmark to deployment-ready decision support.

6. Conclusions

This paper presents a reproducible simulation benchmark in which prompt-structured priors and dual validation are used to construct a career-growth causal graph. Under the main 30-run setting, increasingly structured prompting improved candidate edge extraction, and the hybrid workflow produced high graph-recovery and intervention-estimation accuracy.

The scope of this claim is deliberately limited. The primary contribution is a domain-specific benchmark and workflow for career-growth causal modeling, not a universal causal-discovery evaluator, and not evidence that a named production LLM can recover valid career causal graphs out of the box. The workflow can be transferred to other datasets only when analysts define suitable variables, temporal layers, intervention targets, validation models, and auditable domain rules for the new domain. The public promotion dataset is used only as an external plausibility anchor and not as causal ground truth. On that basis, prompt-structured priors should be understood as one interpretable component in a broader causal-discovery workflow whose next necessary steps are direct validation on real LLMs, fully matched comparisons with standard baselines, richer real HR datasets, computational-cost evaluation, and public release of executable code.

Author Contributions: Conceptualization, F.T. and Y.X.; methodology, Y.X. and F.T.; software, F.T.; validation, Y.X. and Y.Z.; formal analysis, M.L.; investigation, M.L.; resources, Y.Z.; data curation, F.T.; writing—original draft preparation, F.T. and Y.X.; writing—review and editing, F.Y.; visualization, Y.X.; supervision, F.T.; project administration, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable because this study used synthetic data only and involved no human participants.

Informed Consent Statement: Not applicable.

Data Availability Statement: The benchmark specification, nominal SCM coefficients, emulator profiles, representative full prompt templates, structured output schema, and seed policy are reported in Appendices A and B. The public-data anchor relies on the publicly accessible WNS/Analytics Vidhya employee-promotion dataset. Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Default Benchmark Configuration and Reproducibility Notes

This appendix summarizes the nominal benchmark configuration used for this study. All reported main-text results were averaged over 30 paired runs with synthetic-data seed $20,250 + r$ and emulator seed $30,250 + r$, where $r = 1, \dots, 30$. Continuous variables were standardized before local coefficient comparison; for binary child nodes, standardized latent-link coefficients from logistic models were used so that local effect sizes were compared on a scale-consistent basis rather than by mixing raw linear and logistic coefficients.

These nominal equations reproduce the intended causal ordering and approximate descriptive characteristics reported in Section 3.1. The ground-truth graph contains

17 directed edges; independent noise draws create run-to-run variation without changing the parent sets.

Table A1. Default structural equations and nominal parameterization of the synthetic career-growth benchmark.

Node	Structural Form	Parents	Nominal Parameterization
Education	Ordinal draw	—	Levels coded 0/1/2 with probabilities {0.28, 0.49, 0.23}, corresponding, respectively, to diploma-or-below, undergraduate/bachelor-level, and postgraduate-level qualifications.
Experience	Truncated normal	—	$\mu = 5.5$, $\sigma = 2.4$, truncated to $[0, 15]$.
PolicySubsidy	Bernoulli	—	$p = 0.36$.
Mentorship	Bernoulli	—	$p = 0.43$.
BaselineSkill	Linear-Gaussian	Education, Experience	$0.55 \cdot \text{Education} + 0.22 \cdot \text{Experience} + \varepsilon$, with $\varepsilon \sim N(0, 0.70^2)$.
TrainingHours	Linear-Gaussian	PolicySubsidy, Education, BaselineSkill	$12 + 5.2 \cdot \text{PolicySubsidy} + 2.1 \cdot \text{Education} + 3.7 \cdot \text{BaselineSkill} + \varepsilon$, with $\varepsilon \sim N(0, 5.5^2)$.
Certification	Logistic Bernoulli	TrainingHours, Education, BaselineSkill	$\text{logit}(p) = -5.30 + 0.09 \cdot \text{TrainingHours} + 0.40 \cdot \text{Education} + 0.60 \cdot \text{BaselineSkill}$.
ProjectExposure	Linear-Gaussian	Experience, Mentorship, BaselineSkill	$0.8 + 0.34 \cdot \text{Experience} + 1.75 \cdot \text{Mentorship} + 0.70 \cdot \text{BaselineSkill} + \varepsilon$, with $\varepsilon \sim N(0, 1.8^2)$.
Performance	Linear-Gaussian	TrainingHours, Certification, ProjectExposure	$52 + 0.17 \cdot \text{TrainingHours} + 5.2 \cdot \text{Certification} + 2.6 \cdot \text{ProjectExposure} + \varepsilon$, with $\varepsilon \sim N(0, 6.0^2)$.
Promotion	Logistic Bernoulli	Performance, Experience, PolicySubsidy	$\text{logit}(p) = -5.20 + 0.050 \cdot \text{Performance} + 0.090 \cdot \text{Experience} + 0.32 \cdot \text{PolicySubsidy}$.

Table A2. Default prompt-response emulator profiles used for the four prompt families.

Prompt Family	True-Edge Proposal Probability	Admissible False-Edge Proposal Probability	Strength-Noise SD	Constraint Emphasis
Zero-shot	0.66	0.19	0.18	Task description only.
Few-shot	0.73	0.14	0.14	Task description plus demonstrations.
CoT	0.78	0.11	0.11	Mechanism and temporality reasoning.
CoT + Schema	0.81	0.08	0.09	Reasoning plus schema, layer, and output constraints.

The proposed hybrid method uses the CoT + Schema profile as its prompt prior and then applies the dual validation procedure described in Section 3.4. No run-specific retuning was applied beyond stochastic noise and seed variation. Representative full prompt templates are provided in Appendix B.

Appendix B. Representative Prompt Templates and Structured Output Schema

This appendix provides representative prompt text so that the benchmark interface can be inspected and reimplemented. The templates below are written for pairwise edge as-

assessment using placeholder tokens in braces. They are not presented as optimal production prompts; rather, they document the structured prompt families abstracted by the emulator.

Appendix B.1. Zero-Shot Template

Task: Evaluate whether {SOURCE} is a cause of {TARGET} in the career-growth process.

Available variables: Education, Experience, PolicySubsidy, Mentorship, BaselineSkill, TrainingHours, Certification, ProjectExposure, Performance, Promotion.

Instructions:

1. Consider temporal precedence and mechanism.
2. Decide one of {direct_cause, mediated_cause, no_causal_link}.
3. Return only the required fields; do not add commentary outside the schema.

Required output:

```
{“source”:”{SOURCE}”, “target”:”{TARGET}”, “relation_type”:”...”,
“confidence”:0.00, “strength”:0.00, “rationale”:”one sentence”}
```

Appendix B.2. Few-Shot Exemplar Template

Example 1

Input pair: TrainingHours -> Certification

Expected judgment: direct_cause

Rationale: More structured training increases the probability of obtaining certification.

Example 2

Input pair: Promotion -> Education

Expected judgment: no_causal_link

Rationale: Promotion cannot cause prior educational attainment because the temporal direction is reversed.

Now evaluate the new pair {SOURCE} -> {TARGET}. Use the same JSON schema as above and keep the rationale concise.

Appendix B.3. CoT + Schema Template

You are screening a candidate edge for a career-growth causal graph.

- Step 1: Check whether the source can temporally precede the target. Reject any edge that makes a downstream variable the cause of an upstream qualification or makes Promotion a parent of earlier variables.
- Step 2: Reason about a plausible mechanism linking the source to the target. Distinguish direct causation from a relation that is plausible only through mediators.
- Step 3: Check whether the relation is more consistent with shared antecedents than with a directed effect.
- Step 4: Assign confidence in [0, 1] and normalized strength in [0, 1].

Allowed relation_type values: direct_cause, mediated_cause, no_causal_link.
 Output constraints: return JSON only; no markdown; no extra explanation outside the rationale field.

JSON schema:

```
{“source”:”{SOURCE}”, “target”:”{TARGET}”, “relation_type”:”...”,
“confidence”:0.00, “strength”:0.00, “rationale”:”<=25 words”}
```

Appendix B.4. Structured Output Fields

source: candidate parent variable; target: candidate child variable; relation_type: one of direct_cause, mediated_cause, or no_causal_link; confidence: decimal in [0, 1]; strength: normalized effect magnitude in [0, 1]; rationale: brief natural-language justification. In the benchmark implementation, increasingly constrained prompt families map to higher true-edge proposal probabilities and lower admissible false-edge proposal probabilities, but that mapping remains a benchmark assumption rather than an empirical API measurement.

References

- Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
- Spirtes, P.; Glymour, C.N.; Scheines, R. *Causation, Prediction, and Search*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2000.
- Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press: Cambridge, MA, USA, 2017.
- Hernán, M.A.; Robins, J.M. *Causal Inference: What If*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2020.
- Glymour, M.; Pearl, J.; Jewell, N.P. *Causal Inference in Statistics: A Primer*; Wiley: Chichester, UK, 2016.
- Kalisch, M.; Bühlmann, P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **2007**, *8*, 613–636.
- Tsamardinos, I.; Brown, L.E.; Aliferis, C.F. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **2006**, *65*, 31–78.
- Zanga, A.; Ozkirimli, E.; Stella, F. A survey on causal discovery: Theory and practice. *Int. J. Approx. Reason.* **2022**, *151*, 101–129. [\[CrossRef\]](#)
- Assaad, C.K.; Devijver, E.; Gaussier, E. Survey and evaluation of causal discovery methods for time series. *J. Artif. Intell. Res.* **2022**, *73*, 767–819. [\[CrossRef\]](#)
- Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [\[CrossRef\]](#)
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [\[CrossRef\]](#)
- Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 11048–11064.
- Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; et al. Large language models and causal inference in collaboration: A comprehensive survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 7683–7699.
- Girju, R. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, Sapporo, Japan, 11 July 2003*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 76–83.
- Dunietz, J.; Levin, L.; Carbonell, J.G. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop, Valencia, Spain, 4 April 2017*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 95–104.
- Yang, J.; Han, S.C.; Poon, J. A survey on extraction of causal relations from natural language text. *Knowl. Inf. Syst.* **2022**, *64*, 1161–1186. [\[CrossRef\]](#)

18. Gopalakrishnan, S.; Garbayo, L.; Zadrozny, W. Causality extraction from medical text using large language models (LLMs). *Information* **2025**, *16*, 13.
19. Aziz, M.; Brookhart, M.A. Can contemporary large language models provide the domain knowledge needed for causal inference? Evaluating automated causal graph discovery through an ASCVD case study. *Clin. Epidemiol.* **2025**, *17*, 863–873. [[CrossRef](#)] [[PubMed](#)]
20. Chatwal, P.; Agarwal, A.; Mittal, A. Enhancing causal relationship detection using prompt engineering and large language models. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, Abu Dhabi, United Arab Emirates, January 2025; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 248–252.
21. Cho, W.; Choi, S.; Choi, H. Human resources analytics for public personnel management: Concepts, cases, and caveats. *Adm. Sci.* **2023**, *13*, 41. [[CrossRef](#)]
22. Lee, J.Y.; Lee, Y. Integrative literature review on people analytics and implications from the perspective of human resource development. *Hum. Resour. Dev. Rev.* **2024**, *23*, 58–87.
23. Wang, L.; Zhou, Y.; Sanders, K.; Marler, J.H.; Zou, Y. Determinants of effective HR analytics implementation: An in-depth review and a dynamic framework for future research. *J. Bus. Res.* **2024**, *170*, 114312.
24. Fettach, Y.; Bahaj, A.; Ghogho, M. JobEdKG: An uncertain knowledge graph-based approach for recommending online courses and predicting in-demand skills based on career choices. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107779.
25. Boškoski, P.; Redek, T.; Perne, M.; Boshkoska, B.M. Career path discovery through bipartite graphs. *J. Decis. Syst.* **2024**, *33*, 140–153. [[CrossRef](#)]
26. Ntioudis, D.; Masa, P.; Karakostas, A.; Meditskos, G.; Vrochidis, S.; Kompatsiaris, I. Ontology-based personalized job recommendation framework for migrants and refugees. *Big Data Cogn. Comput.* **2022**, *6*, 120. [[CrossRef](#)]
27. Alsaif, S.A.; Sassi Hidri, M.; Ferjani, I.; Eleraky, H.A.; Hidri, A. NLP-based bi-directional recommendation system: Towards recommending jobs to job seekers and resumes to recruiters. *Big Data Cogn. Comput.* **2022**, *6*, 147.
28. Zafar, A.; Parthasarathy, V.B.; Van, C.L.; Shahid, S.; Khan, A.I.; Shahid, A. Building trust in conversational AI: A review and solution architecture using large language models and knowledge graphs. *Big Data Cogn. Comput.* **2024**, *8*, 70. [[CrossRef](#)]
29. Gonzalez-Garcia, L.; Sicilia, M.-A.; García-Barriocanal, E. Using vector databases for the selection of related occupations: An empirical evaluation using O*NET. *Big Data Cogn. Comput.* **2025**, *9*, 175. [[CrossRef](#)]
30. Zhang, Z.; Wang, J.; Li, B.; Lin, X.; Liu, M. Construction of a person–job temporal knowledge graph using large language models. *Big Data Cogn. Comput.* **2025**, *9*, 287. [[CrossRef](#)]
31. Analytics Vidhya. WNS Analytics Wizard 2018: HR Analytics Challenge. Available online: <https://www.analyticsvidhya.com/datathon/hackathon/wns-analytics-hackathon-2018/> (accessed on 8 April 2026).
32. Zheng, X.; Aragam, B.; Ravikumar, P.; Xing, E.P. DAGs with NO TEARS: Continuous optimization for structure learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9472–9483.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.