Unsupervised Discovery of Causal Models from Interview Transcripts

Anonymous ACL submission

Abstract

This paper introduces four new causal model datasets and evaluates a novel framework for the unsupervised extraction of Fuzzy Cognitive Maps (FCMs)—signed, directed causal models—from unstructured text using Large Language Models (LLMs). While FCMs offer a powerful mechanism for integrating diverse stakeholder perspectives, their construction is traditionally manual and difficult to scale. We address this challenge by systematically evaluating a suite of LLM-driven extraction methods, including in-context learning and an LLM-plus-ant-colony-optimization (ACO) approach. We benchmark these methods on a new corpus of semi-structured interviews spanning marine biodiversity, offshore wind development, food systems, and fisheries. Since interviews contain disfluencies and non-linear narratives, we treat them as a challenging lower bound test case compared with more polished sources. Results show that in-context learning outperforms ACO. We also release a dataset of interview transcripts paired with expert-generated causal models to facilitate future research in automated mental model extraction.

1 Introduction

Understanding complex social-ecological systems requires the integration of diverse, often conflicting, human perspectives. Fuzzy Cognitive Mapping (FCM) has emerged as a preeminent tool for this task, providing a graphical representation of the "mental models" held by individuals or groups regarding system dynamics. By representing concepts as nodes and causal influences as signed, directed edges, FCMs allow researchers to simulate policy interventions and visualize the interconnectedness of socioeconomic and environmental variables (Figure 1).

However, a fundamental challenge persists: the construction of high-fidelity FCMs is typically a labor-intensive, manual process. It requires experts to conduct semi-structured interviews, manually transcribe responses, and iteratively code causal relationships—often returning to participants for validation. While this approach yields high validity and accuracy, the manual bottleneck severely limits the scalability of FCM research. Furthermore, it is prone to recency bias and prevents the rapid, thorough integration of numerous perspectives needed to inform time-sensitive policy and management decisions (Özesmi and Özesmi, 2004).

The rapid advancement of Large Language Models (LLMs) presents a significant opportunity to automate FCM extraction. Unlike traditional relation extraction (RE) models, LLMs can leverage broad world knowledge to infer causality even when explicit syntactic markers are absent. Yet, the accuracy of automated causal extraction from unstructured qualitative data remains insufficiently assessed. Specifically, while LLMs show promise in structured information extraction, their ability to navigate the nuances of human discourse—where causal links are often implied or buried in narrative—is an open research question.

In this paper, we address this gap by asking: Can LLMs reliably identify causal relationships and relevant concepts in text to faithfully extract FCMs? To establish a robust lower bound on performance, we focus on extracting FCMs from challenging interview transcripts rather than polished research or news articles. We conjecture that if an LLM can extract coherent causal structure from the informal, often disjointed speech of an interview, its performance on cleaner, more structured textual data will be substantially higher.

We systematically evaluate three modern Natural Language Processing (NLP) methods for extracting causal structure and compare the resulting AI-generated models to expert-crafted "ground-truth" FCMs. Our study introduces four new datasets spanning marine biodiversity, food security, and offshore wind energy to support the generalizabil-

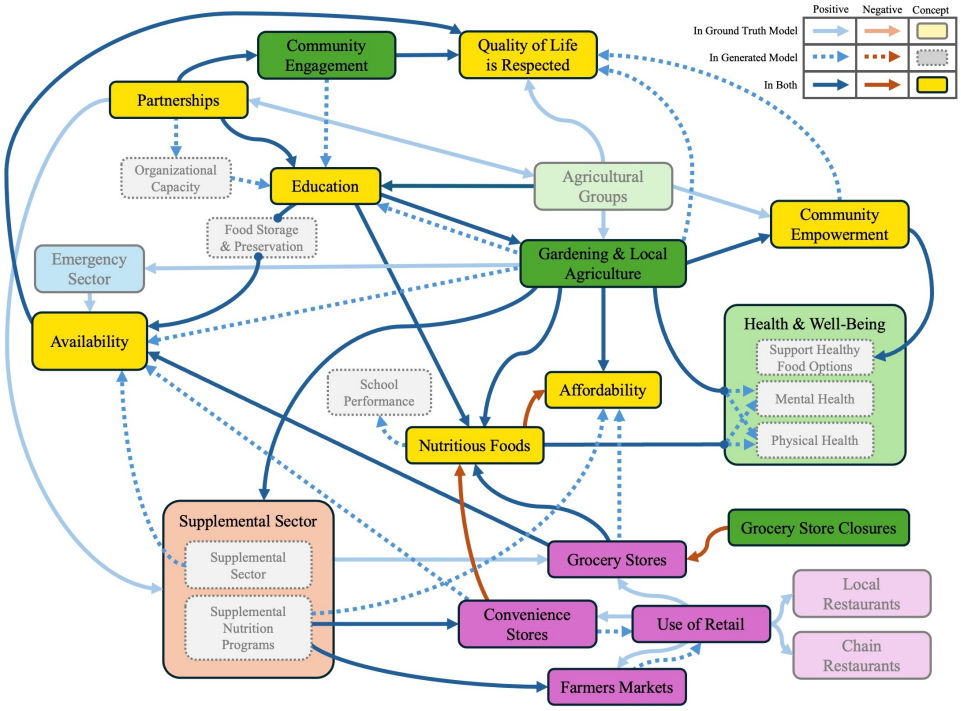

Figure 1: Comparison of a high-F1, AI-generated FCM and the corresponding grounded-truth for the Flint, MI food system; see key in upper-right. Supplemental and emergency sectors have been slightly simplified for legibility.

ity of our findings and encourage future research.

2 Background

2.1 Fuzzy Cognitive Maps and Their Applications

FCMs are semi-quantitative, graph-based models used to represent the causal structure of complex systems. First introduced by Kosko (1986), an FCM consists of a set of natural language concepts $C = \{c_1, c_2, \ldots, c_n\}$ represented as nodes, and causal relationships represented as directed, weighted edges E. The weight w_{ij} of an edge from c_i to c_j typically ranges from $[-1, 1]$, where:

- $w_{ij} > 0$ indicates a positive causal relationship (an increase in c_i leads to an increase in c_j).
- $w_{ij} < 0$ indicates a negative causal relationship (an increase in c_i leads to a decrease in c_j).
- $w_{ij} = 0$ indicates no causal relationship.

FCMs have been widely adopted in Social-Ecological Systems (SES), extreme weather adaptation, and public health to capture stakeholder knowledge (Gray et al., 2014). Unlike traditional Bayesian Networks or System Dynamics models, FCMs excel in data-scarce environments where subject-matter expertise is the primary source of information. By aggregating multiple individual maps into an aggregate, collective-intelligence model, researchers can simulate the downstream effects of specific policy interventions or environmental stressors (Knox et al., 2023; Papageorgiou and Salmeron, 2013).

2.2 Manual vs. Automated FCM Extraction

The standard protocol for manually building FCMs typically follows one of two paths:

Participatory Modeling. A researcher builds the map in real time with a stakeholder during an interview. While highly accurate, this is difficult to scale and can be subject to interviewer bias.

Post-hoc Coding. Researchers manually code causal relationships from interviews or literature. This method is prone to interviewer bias and knowledge siloing, where specific phrasing used by a participant may be reinterpreted by the coder.

The manual bottleneck prevents FCMs from being used for large-scale collective-intelligence projects that require synthesizing thousands of viewpoints. Automated extraction using NLP offers a potential solution by rapidly scanning text for causal signals, yet identifying these signals in unstructured and semi-structured discourse remains a significant challenge (Dodouras and James, 2007).

2.3 NLP Approaches to Causal Relation and Concept Extraction

Extracting causal relationships from text is a core sub-task of RE. As detailed in comprehensive surveys of the field, traditional approaches to *causal relation extraction* (CRE) can be mapped into three

foundational methodological streams: knowledge-based systems relying on hand-crafted rules, statistical machine learning models utilizing semantic features, and deep learning architectures ranging from recurrent neural networks to modern transformer models (Ali et al., 2021; Yang et al., 2022). Methodological splits in traditional systems also depend heavily on the linguistic representation of causality, notably differentiating between explicit and implicit expressions, as well as handling localized intra-sentential versus complex, cross-boundary inter-sentential causal dynamics (Yang et al., 2022). Furthermore, the interaction between NLP and causal inference extends beyond simple sequence labeling; it encompasses a dual-track paradigm that 1) utilizes text as variables for downstream causal effect estimation (e.g., text as a treatment) or 2) applies formal causal frameworks to diagnose the robustness, fairness, and interpretability of NLP models (Feder et al., 2022).

Causal Discovery and LLMs. Recent advancements in generative architectures have shifted the extraction focus toward context-aware, autonomous discovery. Frameworks like Causal-LLM (Roy et al., 2025) demonstrate that Large Language Models (LLMs) can perform one-shot graph discovery by leveraging internal world knowledge to identify structural dependencies that statistical methods miss. Recent literature surveys formalize the intersection of causal discovery and LLMs as a bidirectional pipeline leveraging LLMs to execute qualitative causal inference tasks (such as structural discovery and counterfactual reasoning), or utilizing structural causal metrics to audit, evaluate, and debias the alignment of generative models (Liu et al., 2025; Ma, 2025).

Agentic and Hybrid Methods. While LLMs demonstrate powerful few-shot reasoning capabilities, open-loop inference frequently suffers from structural instability when confronted with complex domain boundaries. Recent literature points out that generative inference heavily benefits from external wrappers, agentic state tracking, or optimization routines to stabilize graph generation (Liu et al., 2025; Ma, 2025). For instance, the "agentic leash" framework lets LLMs autonomously retrieve and process localized text patches to iteratively refine FCM dynamical parameters (Panda et al., 2025). Others couple LLMs with evolutionary algorithms (e.g., CMA-ES) to reconcile sparse generative priors against observational data constraints, learning sparse, interpretable FCMs (Schuerkamp and Giabbanelli, 2025).

A core challenge in automated FCM construction is node reconciliation—ensuring that "rising ocean temperatures" and "warming seas" are mapped to the same concept. Soft measures were introduced in Berijanian et al. (2024) for evaluating extracted causal intelligence models on the basis of edge semantics. Their work established human-aligned evaluation measures of concept matching in causal graphs.

Despite these advances, most literature focuses on clean data like research abstracts or news articles. This paper extends previous work by evaluating methods on noisy data sources: raw interview transcripts. Motivated by the need to increase efficiencies in SES applications, we evaluate extraction by in-context learning and LLM-plus-optimization to determine how well they can recover expert-validated structures.

3 Methods

3.1 Datasets

We introduce four interview-based datasets paired with expert-derived FCMs that were built using participatory systems modeling software[1]. Each dataset originates from a separate SES study with no overlap among interviewers/annotators/participants; interviewers/annotators are evaluators and coauthors of the present study.

Marine Biodiversity captures perspectives on marine and coastal biodiversity governance, ecosystem services, stressors, and management actions from marine resource managers, researchers, and users with attention to biodiversity impacts upon participant lives. Locations were selected for comparison with each reflecting a unique system with high biodiversity, a history of environmental degradation, and residents who possess a high level of sense of place. Interviews were conducted in-person and via video conferencing software. Chain referral sampling with multiple points of entry was used for recruitment. A pair of social scientists from the Smithsonian and the University of South Alabama conducted interviews with one facilitator and one FCM annotator.

Gulf Offshore Wind (OSW) targets perceived offshore-wind energy impacts (environmental, social, economic) and potential policy responses in Gulf systems. Stakeholders with an interest in the development of offshore wind technology in the

[1] https://www.mentalmodeler.com/

Dataset	N	Hours	Tokens	Nodes	Edges	+ Edge Density	- Edge Density
Biodiversity	42	50	6958 ± 2047	27.3 ± 10.1	52.2 ± 32.8	**9.0% ± 13.5%**	**3.8% ± 6.0%**
FLPP	51	**75**	5954 ± 2375	28.7 ± 9.4	51.4 ± 19.5	5.9% ± 3.2%	1.5% ± 0.7%
Gulf OSW	31	50	4882 ± 1392	**38.6 ± 8.1**	**60.5 ± 17.4**	4.3% ± 1.3%	0.0% ± 0.1%
Red Snapper	**52**	45	**7189 ± 2888**	15.7 ± 5.0	11.7 ± 4.7	3.7% ± 1.8%	1.8% ± 1.1%
All	176	220	6370 ± 2466	26.2 ± 11.4	41.1 ± 28.4	5.7% ± 7.1%	1.9% ± 3.2%

Table 1: Summary statistics for all datasets (mean ± SD; **greatest** / least).

Dataset	Participant Groups	Primary Locales	Interview / Transcription Workflow	Notes for Interpretation
Biodiversity	Managers, researchers, and resource users (e.g., fishers, ecotourism operators)	Puget Sound (WA), Mobile Bay (AL), Chesapeake Bay (MD/VA), broader U.S. coastal contexts	Semi-structured interviews; two-researcher workflow with live map elicitation and post-interview coding; manual transcription	Strong participatory structure with iterative coder reconciliation and community-level map aggregation
FLPP	Community partners and domain experts with food-system experience	Flint, Michigan	Video interviews with facilitator-built live models; snowball recruitment; OtterAI transcripts validated and anonymized by researchers	Standard seed concepts plus participant expansion; supports cross-map concept standardization
Gulf OSW	Stakeholders and subject-matter experts with interests in offshore wind impacts	U.S. Gulf of America (formerly Mexico) region	Phone/video/self-administered interactions; convenience + chain-referral sampling; partial Google Meet auto-transcription (permission-dependent)	DPSIR-oriented protocol emphasizes impacts and responses; structure affects edge elicitation conventions
Red Snapper	Commercial, recreational, and charter-for-hire fishers	Gulf of America, cross-sector fishery communities	Semi-structured phone interviews; model expanded sequentially to saturation; transcription by GMR Transcription	Fisher knowledge captured across sectors; sequential expansion supports broad concept discovery

Table 2: Qualitative comparison of dataset construction protocols and contextual characteristics.

Gulf and its potential positive and negative effects on Gulf systems were recruited via convenience and snowball sampling. National Oceanic and Atmospheric Administration (NOAA) federal and affiliate social scientists led scoping, participant identification, and facilitation. A NOAA affiliate social scientist constructed models through interviews facilitated using a semi-structured protocol linked to DPSIR framing (Maxim et al., 2009).

Flint, Michigan Food System (FLPP) focuses on interactions among retail, supplemental, and emergency food sources and their relationships to community-defined values (Zareei et al., 2025; Helgeson et al., 2025; Knox et al., 2023). Addressing food insecurity in urban areas like Flint, MI remains a major policy and supply chain challenge that requires causal knowledge integration across the public and private domains (Knox et al., 2023). Workshops with Flint community members and consultation with the Community Consultative Panel (CCP) heavily informed development of the interview protocol. A University of Michigan social scientist facilitated live mapping and performed transcript validation and concept standardization to improve cross-participant comparability.

Gulf of America Red Snapper Fishermen synthesizes interviews across commercial, private recreational, and charter-for-hire sectors with a focus on the economically and culturally significant red snapper fish species in the Gulf. A NOAA affiliate scientist conducted semi-structured phone interviews according to the methods of Clay and Colburn (2020) and expanded a shared conceptual model interview-by-interview until saturation.

The heterogeneous nature of the datasets introduces several dimensions of variability that challenge the robustness of automated extraction. Technical variability arises from differences in transcript quality—ranging from manual transcriptions to automated Google Meet outputs—and the specific interview protocols used, such as the structured DPSIR framework. Cognitive and contextual variability is driven by the diverse backgrounds of both the interviewers and interviewees whose varying levels of education, disciplinary biases, and professional relationships to the system (e.g., manager vs. resource user) significantly influence how causal relationships are articulated. Finally, the systemic variability across the four domains, spanning the biological complexity of marine fisheries to the socioeconomic interactions of the Flint, Michigan food system, ensures that the extraction methods are tested against a wide spectrum of causal structures and linguistic contexts.

Across all four domains, ground truth FCMs vary substantially in structural complexity. Red Snapper FCMs are notably compact, while OSW FCMs are the most elaborate. Edge densities are highest in Biodiversity, reflecting richer interconnection, while OSW FCMs notably lack negative relations. Graph statistics appear in Table 1; Table 2 summarizes differences among interviews.

3.2 Unsupervised FCM Extraction

We introduce three FCM extraction approaches: two in-context learning approaches (with and without explanation) and one bio-inspired. To ensure model coverage, each approach is implemented via commercial and open-weight chat completion models of various sizes: `gpt-5-mini`, `gpt-5.2`, `gemini-3-flash-preview`, `gemini-2.5-flash`, `qwen2.5-14B-instruct`, and `mistral-small-3.1-24B-instruct-2503`. Qwen and Mistral models were locally served via vLLM on 8 Quadro RTX 5000 GPUs with a 32k token context window using half-precision format.

3.2.1 In-Context Extraction

FCM nodes and edges were extracted from source documents using an LLM-based pipeline. Documents in plain text, JSON, and DOCX formats were ingested, cleaned, and submitted individually to models. A structured two-part prompt instructed the model to identify FCM nodes—defined as quantitatively or qualitatively varying concepts expressed as noun phrases—and directed causal edges labeled with a sign (positive, negative, or ambiguous). For each edge, the model also returned a confidence score $\in (0, 1)$ reflecting how explicitly the relationship was stated in the source text and a verbatim supporting text snippet. One approach additionally requested a brief causal explantion statement for each edge (denoted *E*; no explanation, *NE*). Model responses were returned as JSON and validated against a structured schema; malformed responses triggered automatic retries with exponential back-off. Documents were processed concurrently, and results were stored in a PostgreSQL database, with processing logs enabling incremental re-runs that skip already-completed files. See A.2 for details.

3.2.2 Bio-Inspired Extraction

Current automated causal-graph extraction methods do not scale well because the number of possible relationships among N concepts grows quadratically, making exhaustive LLM-based verification too slow and expensive. This problem is compounded by context-window limits, "lost-in-the-middle" phenomena (Liu et al., 2024), and hallucinations, which reduce reliability in long texts.

To address this, we propose a novel, hybrid framework that combines LLMs with Ant Colony Optimization (ACO) (Dorigo et al., 1996; Dorigo and Blum, 2005), treating causal discovery as a heuristic search rather than a brute-force process: instead of checking every possible edge, "computational ants" (LLM agents) probabilistically explore candidate relationships, reinforce semantically validated edges, and allow weak or unsupported ones to fade, thereby making graph discovery from unstructured qualitative data more efficient.

The proposed system operates in two stages: first, it extracts stable concepts from raw text using LLM-based meta-prompting and sliding-window chunking to preserve context while maximizing domain adaptability across transcripts; second, it performs causal discovery through an ACO-inspired process in which all possible directed edges are initialized equally, promising candidates are probabilistically sampled, LLMs verify directionality, polarity, and confidence, and pheromone updates iteratively reinforce reliable relationships while unsupported ones evaporate until the graph converges into a stable causal structure (see Algorithm 1).

Mathematical Formulation. In standard ACO, ants transition between nodes based on a probability function. Because our approach applies ACO to discover the topology itself, the "ants" are selecting candidate *edges* to evaluate.

The probability $P_{ij}(t)$ of selecting a candidate edge between concepts i and j at iteration t is defined as:

$$P_{ij}(t) = \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{(u,v)\in C}[\tau_{uv}(t)]^{\alpha}[\eta_{uv}]^{\beta}}, \qquad (1)$$

where $\tau_{ij}(t)$ is the accumulated pheromone level on the edge (i, j) at iteration t; η_{ij} is the heuristic desirability of the edge before LLM verification (it can be a pre-filtering metric like co-occurrence frequency, semantic similarity, or sentence proximity); C is the set of all currently unverified candidate edges; α and β are parameters controlling the relative influence of the pheromone trace vs. the heuristic information. The hyperparameters used in this procedure are summarized in Table 4.

Once the LLM evaluates the selected candidate edges, the global pheromone matrix is updated. The standard evaporation and reinforcement rule is applied to every edge:

$$\tau_{ij}(t+1) = (1-\rho)\big[\tau_{ij}(t) + \Delta\tau_{ij}(t)\big], \qquad (2)$$

where $\rho \in (0, 1)$ is the evaporation rate. A higher ρ causes unvalidated edges to lose their pheromones faster. $\Delta\tau_{ij}(t)$ is the pheromone deposited during

Algorithm 1 ACO-LLM Guided Edge Discovery

Require: Set of concept nodes V, text corpus T, evaporation rate ρ, max Iterations M, convergence threshold ϵ
Ensure: Stabilized FCM Adjacency Matrix

```
Initialize pheromone matrix τ_ij = τ_0 for all i, j ∈ V
Calculate heuristic matrix η_ij based on text snippet co-occurrence
for t = 1 to M do
    C ← edges not yet permanently verified or discarded
    Select candidate edges probabilistically using Eq 1
    for each selected edge (i, j) do
        Retrieve source contexts S_ij from T
        Resp ← QueryLLM(i, j, S_ij)
        if Resp.causal == True then
            Δτ_ij ← Resp.confidence × |Resp.weight|
            Record causality direction and polarity
        else
            Δτ_ij ← 0
        end if
    end for
    for each edge (i, j) ∈ V × V do
        τ_ij(t + 1) ← (1 − ρ)[τ_ij(t) + Δτ_ij(t)]   ▷ Eq 2
    end for
    if |τ(t + 1) − τ(t)| < ε then
        break                                   ▷ Convergence reached
    end if
end for
    return Edges where |W_ij| > 0
```

the current iteration based on the LLM's verification. The LLM deposit can be defined as:

$$\Delta\tau_{i,j}(t) = \begin{cases} C_{ij} \times |W_{ij}| & \text{if LLM validates } i \to j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Here, C_{ij} is the confidence score returned by the LLM, and $W_{ij} = \pm 1$ is the causal polarity.

3.3 Evaluation

3.3.1 Quantitative (Semantics & Structure)

Similarity between AI-generated and ground-truth FCMs is quantified with soft F1 scores between edge sets to balance precision and recall, which penalize overly verbose and insufficient responses, respectively. We adapt the method in Berijanian et al. (2024) to compute node similarities via cosine similarity between `qwen3-embedding-0.6B` textual embeddings of node names. Nodes with cosine similarities above a threshold (T) are deemed matches. We count a signed, directed edge, (source, target, sign), as a true positive (TP) if (1) the cosine similarity of both source and target exceeds T for a ground-truth edge, and (2) the sign matches. Following Berijanian et al. (2024), a "partial positive" (PP) is counted if the sign differs. The soft F1 ($\mathbf{K}$) between an AI-generated FCM (F) and a ground-truth model (G) is computed as:

$$\mathbf{K}(F, G) = \frac{2TP + \alpha PP}{2TP + \alpha PP + FP + FN}. \tag{4}$$

Hyperparameters $T = 0.6$, $\alpha = 0.6$ are tuned following Berijanian et al. (2024). The Hungarian algorithm pairs the highest-scoring edges according to node embedding cosine similarity.

In addition to assessing individual FCM *semantics* via soft F1, LLM methods are globally evaluated on the basis of *graph structural correlations* with ground-truth. This includes *graph-size-aware* measures of node and edge counts and *size-agnostic* measures of positive and negative edge density. For each method, Pearson correlations between the LLM-generated and ground-truth graph statistics are computed across the entire dataset. Semantics and structure must be considered together.

3.3.2 Qualitative

To qualitatively assess AI-extracted models, the scientists from each study manually compared ground-truth FCMs to corresponding high-, median-, and low-F1 FCMs produced by the top-performing extraction method as determined by quantitative results. Evaluators were instructed to appraise models for node/edge absence/presence, semantics, granularity (more/less general), hallucination, and graph structural similarity.

4 Results

4.1 Quantitative

In-context methods outperformed ACO variants overall (Figure 2–A). `gpt-5-mini` achieved the highest mean F1 ($M = 0.69$), followed by `gpt-5.2` ($M = 0.68$) and `gemini-2.5-flash` ($M = 0.66$), with negligible differences between explanation and no-explanation prompting. `mistral` ($M = 0.59$) and `gemini-3-flash` ($M \approx 0.52$) performed at intermediate levels, while `qwen` ranked lowest ($M \approx 0.48$). Among ACO configurations, only `aco-gpt-5.2` had competitive performance ($M = 0.66$); `qwen` and `mistral` reached $M = 0.54$–0.57, the Gemini variants dropped to $M \approx 0.36$–0.37, and `aco-gpt-5-mini` failed entirely ($M = 0.04$), producing near-empty graphs.

Structural fidelity, assessed via Pearson correlations between ground-truth and AI-generated graph properties (Panel B), broadly mirrored F1 rankings. Node and edge count correlations were strongest for `gpt-5-mini` ($r_{\text{nodes}} = 0.58$, $r_{\text{edges}} = 0.59$) and `aco-gpt-5.2` ($r_{\text{nodes}} = 0.65$, $r_{\text{edges}} = 0.63$). However, edge density correlations were near zero for all ACO methods—including `aco-gpt-5.2` despite its strong edge-count correlation. Direct-prompting methods, by contrast, produced more

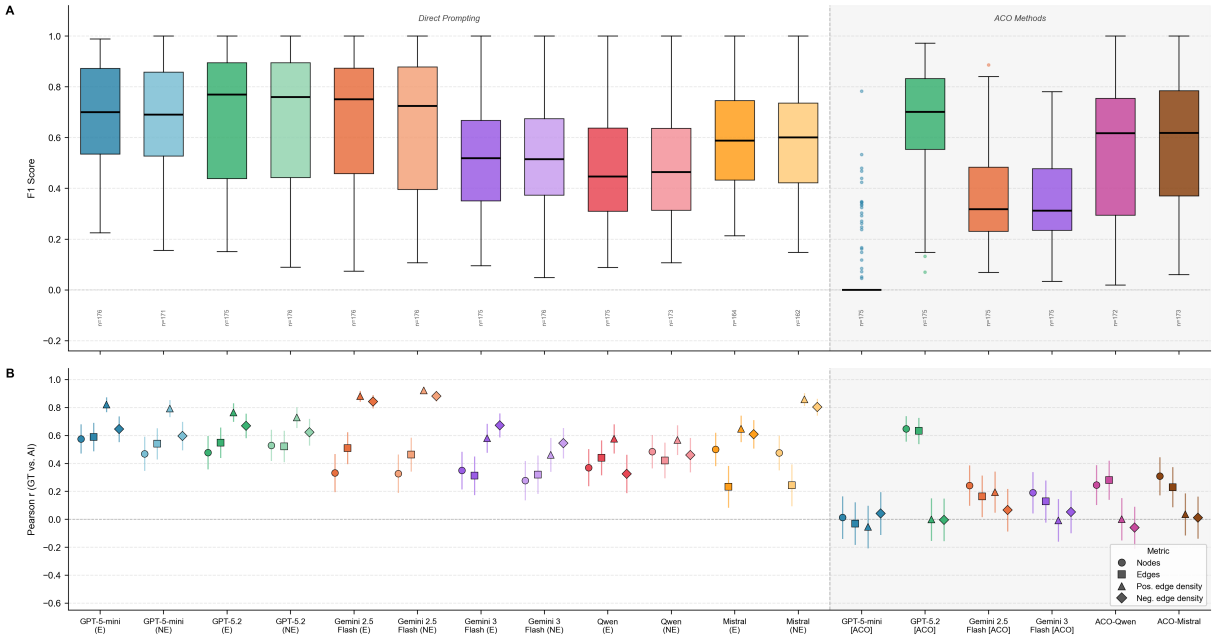

Figure 2: Quantitative performance summary across datasets and methods, including soft F1 comparisons and node/edge count correlations between AI-generated and ground-truth FCMs.

proportionate node and edge counts and exhibited consistently positive, if modest, density correlations.

4.2 Qualitative

Performance and structural outputs varied depending on the complexity of the ground-truth models, the communication style of the interviewees, and the LLM's literal interpretation of the transcripts. Notably, the LLM consistently acted as a high-fidelity transcriber, which led to high recall of explicit statements but resulted in structural divergences such as redundant concepts, non-measurable nodes, and the loss of complex causal motifs when compared to the functional synthesis provided by human facilitators. A detailed summary of the qualitative observations specific to each dataset is provided in Table 3. A visual comparison with an exemplar appears in Figure 1. We elaborate with specific instances of qualitative comparison in the discussion.

5 Discussion

Across the qualitative analyses of the Biodiversity, FLPP, OSW, and Red Snapper datasets, a central theme emerges regarding the capabilities of LLMs in extracting FCMs: LLMs currently function primarily as high-resolution transcribers, though their capacity for functional system modeling varies significantly based on input complexity. While human facilitators naturally synthesize narratives, aggregate concepts, and filter noise to build cohesive systems, the AI generally defaults to mapping the text verbatim. This fundamental difference in approach drives many of the structural anomalies and discrepancies observed in the generated models.

Because the LLM often maps literal text, it frequently struggles with conceptual aggregation, leading to redundancy and an over-representation of granular anecdotes. In the Flint study, this manifested as redundant nodes, such as splitting a single "Availability" concept into "Availability (Overall)" and "Availability (Retail)." Similarly, in the Red Snapper analysis, the AI treated highly specific, unrelated anecdotes—such as an individual fishing guide's weekend bookings—with the same weight as key systemic drivers, resulting in generated models that were substantially larger and more complex than their ground-truth counterparts. Interestingly, the OSW dataset provided a notable exception: the LLM demonstrated a capacity for conceptual binning, successfully resolving minor, related sub-concepts (e.g., "chemicals/new contaminants" and "spills/contaminants") into a broader, synthesized node ("water chemical disturbance"). This suggests that while LLMs lean toward literal extraction, they possess latent aggregation capabilities that may be sensitive to the specific domain or context.

Despite occasionally capturing granular details or successfully binning minor nodes, the AI consistently demonstrated a lack of complex systems reasoning. Generated FCMs frequently lacked ad-

Dataset	Performance & Structural Characteristics	Key Divergences from Ground Truth
Biodiversity	- Higher F1 scores when ground truth edge-to-node ratio was near 1 or maximal. - Generated models generally contained fewer nodes and edges.	- Converted complex motifs (e.g., indirect/moderated effects) into direct effects. - Inability to handle "unknown" (0-strength) relationships resulted in forced connections, removed concepts, or disconnected clusters.
FLPP	- Higher accuracy on smaller ground-truth networks with fewer connections. - More accurate extraction from participants who explicitly stated connections versus those using narrative storytelling.	- Included redundant variables rather than aggregating similar concepts. - Retained highly specific, non-research-focused details (e.g., "box weight") that human facilitators filtered. - Produced disconnected clusters and isolated nodes, unlike fully connected ground-truth models.
Gulf OSW	- Generally strong performance; over 50% of outputs achieved F1 > 0.66. - Successfully captured all major nodes and edges. - Lower F1 scores (F1 < 0.5) correlated with transcripts featuring slang and rapid topic-jumping.	- Extracted fewer nodes and edges than humans for complex, non-linear interviews. - Occasionally merged minor, related edges into single, broader nodes differently than human architects.
Red Snapper	- Demonstrated high literal fidelity to transcripts, operating as an exhaustive mapper. - Generated models frequently had *more* nodes and connections than ground truth. - Occasionally captured socioeconomic nuances that humans simplified.	- Prioritized literal mapping over mechanistic accuracy and functional parsimony. - Included non-measurable concepts, biological impossibilities, and irrelevant anecdotes.

Table 3: Summary of specific qualitative comparisons and structural divergences between LLM-generated (`gpt-5-mini` with explanation) and human-generated ground-truth FCMs across the four datasets.

vanced network motifs, such as indirect, bidirectional, or moderated effects, defaulting instead to simpler, direct causal links. Furthermore, while ground-truth FCMs are typically cohesive, the AI-generated models frequently suffer from structural fragmentation, containing disconnected clusters or nodes. This often stems from the LLM's inability to recognize higher-level conceptual bridges or resolve ambiguous, context-dependent relationships. For example, the Biodiversity analysis highlighted that the AI processed ambiguous "unknown" relationships into simplified causal links, merged concepts inappropriately, or omitted them.

The analyses also underscored the profound impact of the participant's communication style on model accuracy and structure. Both the Flint and OSW studies revealed that the LLM performs best when extracting from simpler, more direct interactions. In the Flint study, fidelity dropped when participants employed narrative formats. Similarly, the lowest-scoring models in the OSW dataset were generated from transcripts of outlier subjects who used slang and jumped rapidly between unrelated concepts. For these more erratic transcripts, the LLM resolved far fewer nodes and edges compared to the ground-truth, struggling to bridge the conversational gaps that a human facilitator naturally navigates.

Ultimately, LLMs provide a highly robust first pass at capturing the major themes and explicit linkages present in interviews. The AI consistently captured the core nodes without missing critical systemic drivers. However, because they lack the contextual discernment and systems-thinking capabilities inherent to domain-expert modelers, LLMs likely require subsequent human refinement to filter irrelevancies, consistently aggregate redundancies, and restore complex mechanistic nuances necessary for rigorous system-level analysis. This suggests the potential value of a two-stage AI system that first extracts and then refines concepts and linkages in a manner similar to, but more refined than, our bio-inspired approach.

6 Conclusions

This study addresses a critical bottleneck in the modeling of complex SES by systematically evaluating LLM extraction methods against a novel corpus of stakeholder interviews spanning four domains. By releasing this dataset paired with ground-truth models, we establish a robust benchmark for automated FCM generation. Our analysis reveals a crucial distinction between literal transcription and functional systems modeling: while LLMs excel at high-resolution transcription and some conceptual aggregation, they currently struggle to resolve narrative ambiguities and lack the complex network reasoning required to generate advanced causal motifs. Consequently, fully automated, unsupervised causal model extraction is not yet a viable replacement for human facilitation in high-fidelity, system-level analysis. Instead, we advocate for a human-in-the-loop paradigm where LLMs perform the labor-intensive initial extraction of major concepts and explicit linkages. This collaborative workflow frees human experts to focus on higher-order refinement–resolving disfluencies, standardizing node labels, and injecting the mechanistic reasoning necessary to translate these foundational maps into viable models for policy intervention.

7 Limitations

We evaluated commercial and open-weight LLMs whose availability is subject to change as models are updated and phased out. This limits the temporal generalizability of these results. Yet, the use of widely available commercial and open-weight models enables application of the same methods by other researchers and practitioners with varying levels of technical expertise. Interviews were conducted to construct FCMs, so participants and interviewers explicitly discussed concepts and causal relations. Interview format specificity limits the generalizability to general interview formats, yet it provides a baseline measurement on challenging dialogue. Textual embeddings are poor at capturing antonym relations (e.g., the cosine similarity between embeddings of "food security" and "food insecurity" is high). For antonyms in two otherwise identical FCMs, signs of all incoming and outgoing edges should be reversed. We did not observe–and do not expect to observe–a large number of antonyms between ground-truth and AI-generated FCMs, and partial positives partially account for this mismatch. However, we acknowledge the possibility of over/undercounting matches.

References

Wajid Ali, Wanli Zuo, Rahman Ali, Xianglin Zuo, and Gohar Rahman. 2021. Causality mining in natural languages using machine and deep learning techniques: A survey. *Applied Sciences*, 11(21):10064.

Maryam Berijanian, Spencer Dork, Kuldeep Singh, Michael Riley Millikan, Ashlin Riggs, Aadarsh Swaminathan, Sarah L Gibbs, Scott E Friedman, and Nathan Brugnone. 2024. Soft measures for extracting causal collective intelligence. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 99–116.

Patricia M Clay and Lisa L Colburn. 2020. A practitioner's handbook for fisheries social impact assessment.

Stefanos Dodouras and Philip James. 2007. Fuzzy cognitive mapping to appraise complex situations. *Journal of Environmental Planning and Management*, 50(6):823–852.

M. Dorigo, V. Maniezzo, and A. Colorni. 1996. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(1):29–41.

Marco Dorigo and Christian Blum. 2005. Ant colony optimization theory: A survey. *Theoretical Computer Science*, 344(2):243–278.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, and 1 others. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Steven A. Gray, Shelby Gray, J. L. De Kok, Ariella Helfgott, Brendan O'Dwyer, Rebecca Jordan, and Anyango Nyaki. 2014. Using fuzzy cognitive mapping as a participatory approach to analyze change, preferred states, and perceived resilience of social-ecological systems. *Ecology and Society*, 19(2):25.

Jennifer Helgeson, Steven Gray, Mahdi Zareei, Carissa Knox, Richard Sadler, Laura Schmitt Olabisi, Chelsea Wentworth, and Nathan Brugnone. 2025. Community-informed decisions for efficient, cost-effective, and integrative disaster resilience planning (co-decidr): Implementation with natural language processing and large language models.

CB Knox, Steven Gray, Mahdi Zareei, Chelsea Wentworth, Payam Aminpour, Renee V Wallace, Jennifer Hodbod, and Nathan Brugnone. 2023. Modeling complex problems by harnessing the collective intelligence of local experts: New approaches in fuzzy cognitive mapping. *Collective Intelligence*, 2(4):26339137231203582.

Bart Kosko. 1986. Fuzzy cognitive maps. *International Journal of Man-Machine Studies*, 24(1):65–75.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, and 1 others. 2025. Large language models and causal inference in collaboration: A comprehensive survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7668–7684.

Jing Ma. 2025. Causal inference with large language model: A survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5886–5898.

Laura Maxim, Joachim H Spangenberg, and Martin O'Connor. 2009. An analysis of risks for biodiversity under the dpsir framework. *Ecological economics*, 69(1):12–23.

Akash Kumar Panda, Olaoluwa Adigun, and Bart Kosko. 2025. The agentic leash: Extracting causal feedback fuzzy cognitive maps with llms. *arXiv preprint arXiv:2601.00097*.

Elpiniki I. Papageorgiou and Jose L. Salmeron. 2013. A review of fuzzy cognitive maps research during the last decade. *IEEE Transactions on Fuzzy Systems*, 21(1):66–79.

Amartya Roy, N Devharish, Shreya Ganguly, and Kripabandhu Ghosh. 2025. Causal-llm: A unified one-shot framework for prompt-and data-driven causal graph discovery. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:8259–8279.

T. Schuerkamp and P. J. Giabbanelli. 2025. Hybrid large-language-model and evolutionary optimization for sparse, interpretable fuzzy cognitive maps. *arXiv preprint arXiv:2502.00000*.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Mahdi Zareei, Carissa Knox, Jennifer Helgeson, Steven A Gray, Richard C Sadler, Laura Schmitt Olabisi, and Chelsea Wentworth. 2025. Community-informed decisions for equitable, cost-effective, and inclusive disaster resilience planning (co-decidr): A modeling approach. *Socio-Environmental Systems Modelling*, 7:18759–18759.

U. Özesmi and S. L. Özesmi. 2004. Ecological models based on people's knowledge: a multi-step fuzzy cognitive mapping approach. *Ecological Modelling*, 176(1–2):43–64.

A Appendix

A.1 ACO Hyperparameters

Hyperparameter	Value	Notes
Max iterations (M)	5	# ACO iterations
Samples per iteration	20	# edges sampled per iter
Evaporation rate (ρ)	0.2	
Initial pheromone (τ_0)	0.1	uniform prior, all edges
Convergence threshold	0.03	
Pheromone effect (α)	1	
Heuristic effect (β)	0	(simplified variant)

Table 4: ACO Hyperparameters.

In this implementation, $\beta = 0$ and no heuristic term η_{ij} is used. This prevents shallow textual cues, such as co-occurrence, from biasing the search, so edge selection is driven only by pheromone values reinforced through LLM validation.

A.2 In-Context Extraction Pipeline

A.2.1 Overview

Fuzzy Cognitive Maps (FCMs) were extracted from raw text documents using a large language model (LLM)-based natural language processing pipeline. The pipeline ingests text corpora in multiple formats (plain text, JSON, or DOCX), queries an OpenAI chat completion model to identify causal concepts and relationships, and persists the structured output to a PostgreSQL relational database.

A.2.2 Document Ingestion and Pre-processing

Source documents were recursively discovered within a configured data directory. Three file types were supported: '.txt' (plain text transcripts), '.json' (structured records with embedded text fields), and '.docx' (Word documents). JSON files were parsed to extract the primary text body along with any available document metadata (e.g., speaker, date, document type). All text was passed through a cleaning step that normalized whitespace and removed extraneous formatting prior to inference. An optional character limit could be applied to cap the length of text submitted per document.

A.2.3 LLM-Based Triple Extraction

Each pre-processed document was submitted to an OpenAI chat completion endpoint (configurable per dataset). The model was prompted with a two-part message — a system prompt defining the extraction task, and a user prompt supplying the document text and requesting a structured JSON response.

System Prompt "' You are an expert at analyzing text and extracting causal relationships

for Fuzzy Cognitive Maps (FCMs).

An FCM consists of:

1. NODES: Concepts that can increase or decrease (quantitatively or

qualitatively)

2. EDGES: Directed causal relationships showing how an increase in one

concept affects another

For edges, use these signs:

- "+" means an increase in source INCREASES the target

- "-" means an increase in source DECREASES the target

- "?" means the relationship is ambiguous or context-dependent

Examples:

- ("rainfall", "crop yield", "+") - more rain increases crop yield

- ("pollution", "air quality", "-") - more pollution decreases air quality

- ("economic growth", "employment", "+") - growth increases employment

- ("food insecurity", "health outcomes", "-") - more food insecurity

decreases health

Focus on substantive causal claims, not mere correlations or associations.

Extract concepts at a consistent level of abstraction.

Include confidence scores (0.0-1.0) based on how explicitly the

relationship is stated.

’’’

User (Extraction) Prompt The placeholder ’text’ is replaced at runtime with the cleaned document content.

’’’

Analyze the following text and extract all concepts (nodes) and causal

relationships (edges) for a Fuzzy Cognitive Map.

TEXT:

text

Please respond with a JSON object in this exact format:

"nodes": [

"name": "concept name",

"description": "brief 5 to 10 word description of what

this concept represents",

"confidence": 0.95

],

"edges": [

"source": "source concept name",

"target": "target concept name",

"sign": "+",

"confidence": 0.9,

"reasoning": "brief 5 to 10 word explanation of why this

causal relationship exists",

"orig_text": "original text snippet supporting this relationship"

]

Important:

1. Node names should be noun phrases (e.g., "water quality" not

"improving water quality")

2. Be consistent with node naming - use the same name when referring

to the same concept

3. Only include causal relationships that are explicitly or strongly

implied in the text

4. Edge confidence should reflect how explicitly the relationship is stated

5. Include ALL relevant concepts and relationships, even if there are many

Respond ONLY with the JSON object, no additional text.

’’’

A.2.4 Structured Output Validation

Model responses were parsed as JSON and validated against Pydantic data models (’ExtractionPayload’, ’FCMNode’, ’FCMEdge’). Malformed responses triggered automatic retries with exponential back-off (up to a configurable number of attempts; default 3), catching JSON parse errors and schema validation failures.

A.2.5 Concurrency and Throughput

Documents were processed concurrently using Python’s ’asyncio’ framework. A semaphore bounded the number of simultaneous API requests (default 5; up to 50 for the interview corpus) to respect rate limits and manage cost. Each document’s processing lifecycle (pending → processing → success/failed) was logged to a ’file_processing’ table, enabling incremental re-runs that skip already-successfully-processed files.

A.2.6 Database Storage

Extracted data were stored in three PostgreSQL tables:

- fcm_documents — one row per source document, including file path, format, and LLM metadata (model name, token usage).

- fcm_nodes — one row per extracted concept, linked to its source document, including concept name, description, and confidence score.

- fcm_edges — one row per directed causal relationship (source concept → target concept), with sign, confidence score, brief reasoning, and a verbatim supporting text snippet from the source document.

Duplicate nodes and edges (matched on ’document_id’ + ’name’, or ’document_id’ + ’source’ + ’target’ + ’sign’, respectively) were handled via upsert logic to ensure idempotency across re-runs.