Unsupervised Discovery of Causal Models from Interview Trascripts

Anonymous ACL submission

Abstract

This paper introduces and evaluates a novel framework for the unsupervised extraction of Fuzzy Cognitive Maps (FCMs)—a class of signed, directed causal models—from unstructured textual data using Large Language Models (LLMs). While FCMs offer a powerful mechanism for integrating diverse stakeholder perspectives into systemic causal models, their construction has traditionally been a bottlenecked, manual process. We address this challenge in scalability of FCM development and use by systematically evaluating a suite of LLM-driven extraction methods, including in-context learning (with and without explanation) and a bio-inspired LLM-plus-ant-colony optimization. To establish a rigorous performance benchmark, we test these methods on a newly compiled corpus of semi-structured interview transcripts across four domains: marine biodiversity, offshore wind development, food systems, and fisheries. We posit that interview transcripts—characterized by disfluencies and non-linear narratives—represent a lower bound for performance compared to polished technical literature. Our findings demonstrate that in-context learning with explicit explanation generation outperforms baseline methods. We additionally contribute to the field by publishing a comprehensive dataset of interview transcripts paired with expert-generated ground-truth 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 skilled facilitators 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 hypothesize that if LLMs can extract coherent causal structure from the informal, often disjointed speech of an interview, their performance on clearer textual data will be substantially higher.

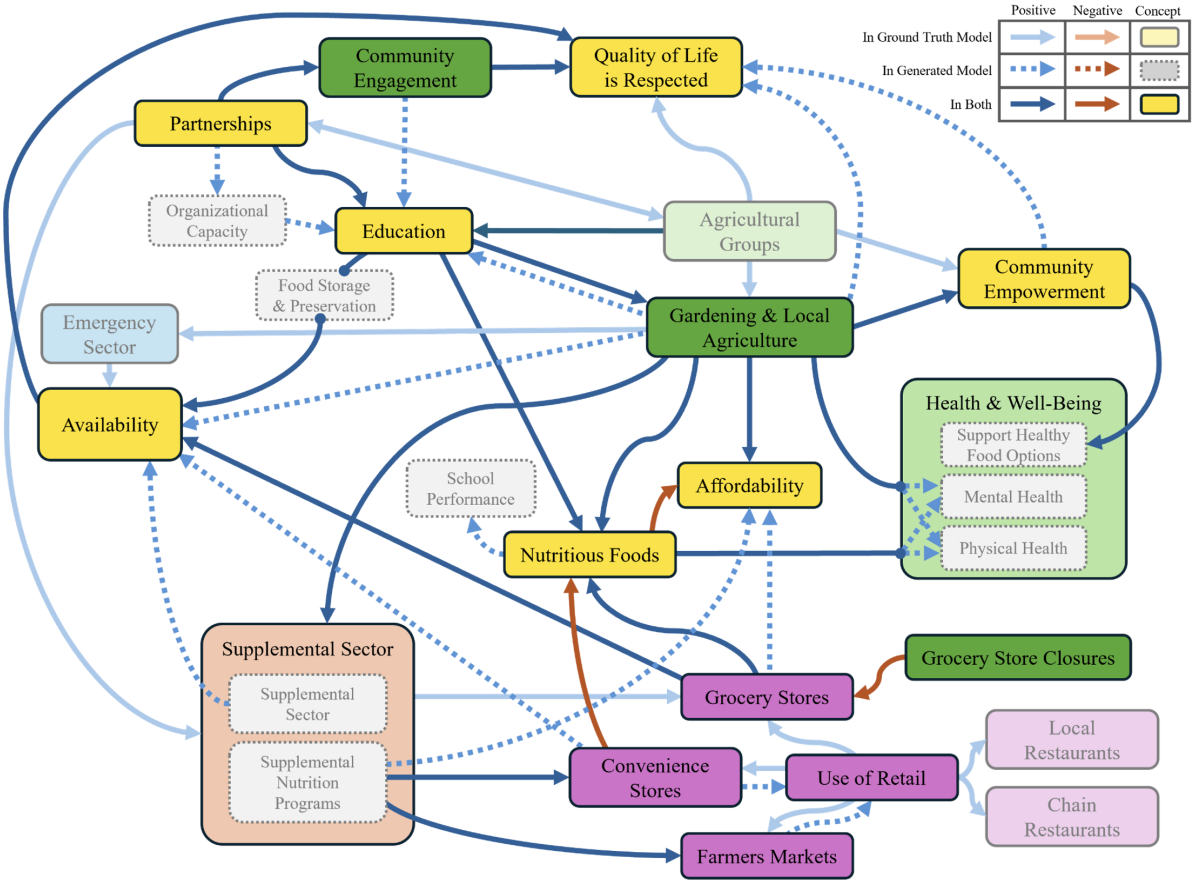

Figure 1: Comparison of a high-F1, AI-generated FCM and the corresponding grounded-truth FCM for the Flint, MI food system; see key in upper-right. Supplemental and emergency sectors have been slightly simplified for legibility.

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 generalizability 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 building FCMs involves laborious manual extraction. This 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 interview transcripts 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 sub-task of Relation Extraction, a well-defined NLP task. Traditional approaches to Causal Relation Extraction (CRE) consisted of rule-based systems that functioned by identifying linguistic "causal markers" (e.g., "causes," "leads to") (Blanco et al., 2008). More recently, approaches to CRE have shifted to discriminative deep learning models, leveraging deep learning frameworks as well as contextual word embeddings produced by transformer-based language models (Anuyah et al., 2025). Advances in generative models have enabled more nuanced, context-aware extraction, such as:

Causal Discovery from Text Recent frameworks like Causal-LLM (Roy et al., 2025) demonstrate that LLMs can perform "one-shot" full-graph discovery by leveraging their internal world knowledge to identify dependencies that statistical methods might miss.

Agentic and Hybrid Methods As shown in **?**, an "agentic leash" framework lets LLMs autonomously fetch and process text to iteratively refine a map's dynamical structure. Others have combined LLMs with Evolutionary Algorithms (e.g., CMA-ES) to learn sparse, interpretable FCMs that fit observational data while maintaining causal validity (Schuerkamp and Giabbanelli, 2025).

Additionally, extracting FCMs from text requires manipulation of the CRE outputs in order to produce a densely connected, informative graph. 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 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 this work by evaluating methods on noisy data sources: raw interview transcripts. We evaluate extraction strategies—in-context learning and LLM-plus-optimization—to determine how well they recover the expert-validated structures from four new datasets.

3 Methods

3.1 Datasets

We introduce four interview-based datasets paired with expert-derived ground-truth FCMs. FCM graph statistics are reported in Table 1; Table 2 summarizes qualitative differences in participants, collection context, and transcription workflow. All identifiable information was manually removed from interviews by data holders.

3.1.1 Marine Biodiversity

This dataset captures participatory interviews on marine and coastal biodiversity governance, ecosystem services, stressors, and management actions. Interviewers co-developed respondent maps during sessions, then applied iterative coding to harmonize concepts across interviews before building aggregate community models.

3.1.2 Gulf Offshore Wind (OSW)

This dataset targets perceived offshore-wind impacts (environmental, social, economic) and potential policy responses in Gulf systems. Data collection mixes interviewer-led and self-administered formats, with partial auto-transcription and protocol constraints linked to DPSIR framing (Maxim et al., 2009).

3.1.3 Flint, Michigan Food System (FLPP)

This dataset focuses on interactions among retail, supplemental, and emergency food sources and their relationships to community-defined values. Facilitated live mapping, transcript validation, and

Dataset	N Interviews/FCMs	Nodes (mean ± SD)	Edges (mean ± SD)	Pos. Edge Density (mean ± SD)	Neg. Edge Density (mean ± SD)
Biodiversity	42	27.3 ± 10.1	52.2 ± 32.8	9.0% ± 13.5%	3.8% ± 6.0%
FLPP	51	28.7 ± 9.4	51.4 ± 19.5	5.9% ± 3.2%	1.5% ± 0.7%
Gulf OSW	31	38.6 ± 8.1	60.5 ± 17.4	4.3% ± 1.3%	0.0% ± 0.1%
Red Snapper	52	15.7 ± 5.0	11.7 ± 4.7	3.7% ± 1.8%	1.8% ± 1.1%
All	178	26.2 ± 11.4	41.1 ± 28.4	5.7% ± 7.1%	1.9% ± 3.2%

Table 1: FCM summary statistics for all datasets. Large SD of edge density is driven by Biodiversity datas.

Dataset	Participant Groups	Primary Locales	Interview / Transcription Workflow	Notes for Interpretation
Biodiversity	Managers, researchers, and resource users (e.g., fishers, eco-tourism 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 SMEs with interests in offshore wind impacts	U.S. Gulf 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.

concept standardization were designed to improve comparability across participant models.

3.1.4 Gulf of America Red Snapper Fishermen

This dataset synthesizes interviews across commercial, private recreational, and charter-for-hire sectors. NOAA scientists expanded a shared conceptual model interview-by-interview until saturation, providing a domain-specific corpus with strong practitioner grounding.

Across all four domains, ground truth FCMs vary substantially in structural complexity. Red Snapper FCMs are notably compact (15.7 ± 5.0 nodes; 11.7 ± 5.0 edges,), while Gulf Offshore Wind FCMs were the most elaborate (38.6 ± 8.1 nodes; 60.6 ± 17.4 edges). Biodiversity and FLPP FCMs occupied intermediate structural complexity. Positive edge densities are highest in biodiversity ($9.0 \pm 13.5\%$), reflecting richer interconnection among concepts, while Gulf OSW FCMs have a strikingly lack of negative relations (mean negative edge density < 0.01%). Across datasets, heterogeneity in elicitation protocol, transcript generation quality, participant backgrounds, and domain complexity creates a deliberately challenging benchmark for automated FCM extration pipelines.

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 served locally via vLLM on 8 Quadro RTX 5000 GPUs with 32k token context 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 (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 expo-

nential back-off. Documents were processed concurrently, and results were stored in a PostgreSQL database (nodes, edges, and source documents in separate tables), with processing logs enabling incremental re-runs that skip already-completed files. Details appear in the appendix.

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), 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 (see Algorithm 1).

The proposed system operates in two sequential 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.

3.3 Evaluation

3.3.1 Quantitative

Similarity between AI-generated and ground-truth FCMs is quantified with soft F1 scores between edge sets. We adapt the method in Berijanian et al. (2024) to compute node similarities via cosine similarity between `qwen3-embedding-0.6M` 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:

$$\mathbf{K}(F, G) = \frac{2TP + \alpha PP}{2TP + \alpha PP + FP + FN}. \quad (1)$$

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

Algorithm 1 ACO-LLM Guided Causal 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 ← subset of edges not yet permanently verified or discarded
    Select K candidate edges probabilistically using Equation 1
    for each selected edge (i, j) do
        Retrieve source contexts S_ij from T
        Response ← QueryLLM(i, j, S_ij)
        if Response.causal == True then
            Δτ_ij ← Response.confidence × |Response.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    ▷ Equation 2
    end for
    if |τ(t + 1) − τ(t)| < ε then
        break                ▷ Convergence reached
    end if
end for
    return Edges where τ_ij > Threshold
```

3.3.2 Qualitative

To qualitatively assess AI-extracted models, SMEs from government and academia compared ground-truth FCMs from each dataset to corresponding high-, median-, and low-F1 FCMs produced by the top-performing extraction method as determined by quantitative results. Evaluation criteria include node/edge absence/presence, semantics, and granularity (more/less general), hallucination, and structural similarity.

4 Results

4.1 Quantitative

Direct-prompting methods outperformed ACO variants overall (Figure 2, Panel A). `gpt-5-mini` achieved the highest mean F1 ($M = 0.69$), followed closely by `gpt-5.2` ($M = 0.68$) and `gemini-2.5-flash` ($M = 0.66$), with negligible differences between explanation and no-explanation prompting conditions. `mistral` ($M = 0.59$) and `gemini-3-flash` ($M \approx 0.52$) performed at intermediate levels, while `qwen` ranked lowest among direct-prompting methods ($M \approx 0.48$). Among ACO configurations, only `aco-gpt-5.2` approached competitive performance ($M = 0.66$); `qwen` and `mistral` reached $M = 0.54$–0.57, the Gemini ACO 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, positive and negative edge density correlations were near zero or negative for nearly all ACO methods—including `aco-gpt-5.2` despite its strong edge-count correlation. Direct-prompting methods, by contrast, produced more proportionate node and edge counts and exhibited consistently positive, if modest, density correlations.

4.2 Qualitative

Across the four datasets, qualitative evaluations compared the LLM-generated FCMs against manual, ground-truth models. 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 divergencessuch 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.

5 Discussion

Across the qualitative analyses of the Biodiversity, Flint Food System, Red Snapper, and Gulf Offshore Wind 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 Gulf Offshore Wind dataset provided a notable exception to this rule. In this domain, 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 across the studies fre-

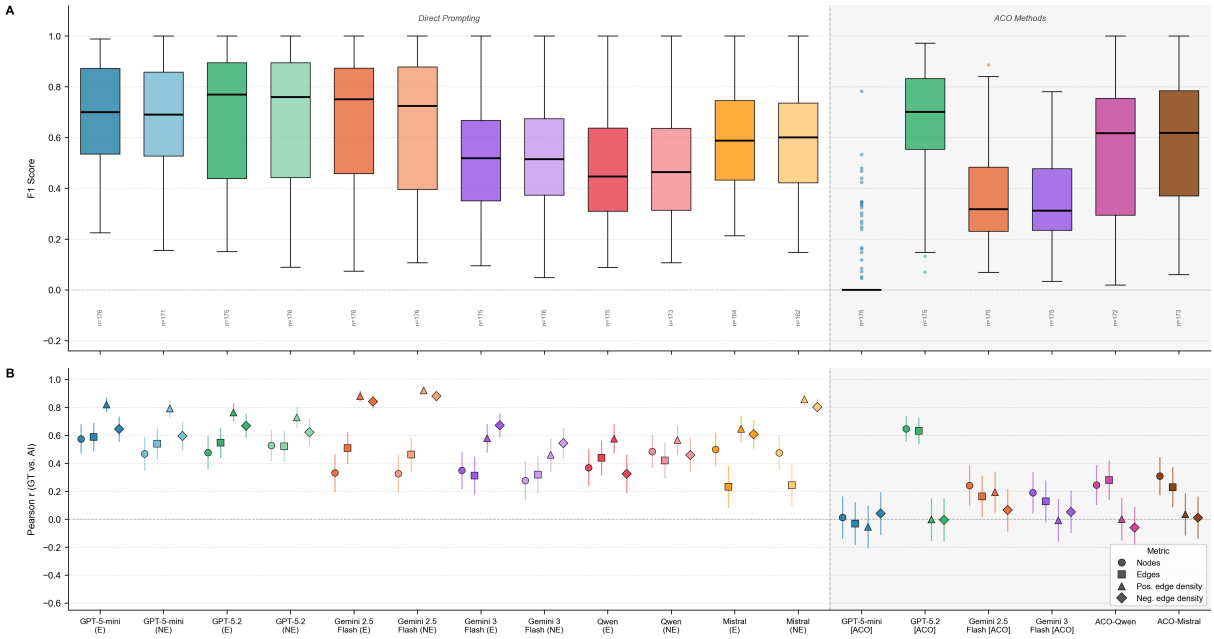

Figure 2: Quantitative performance summary across datasets and methods, including soft F1 comparisons and node/edge count correlations between AI-generated and ground-truth FCMs.

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. - Edge weights were restricted to ± 1, losing the resolution of human-coded weights.	- 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 granular 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.

quently lacked advanced network motifs, such as indirect, bidirectional, or moderated effects, defaulting instead to simpler, direct causal links. Furthermore, while ground-truth FCMs are typically cohesive networks, the AI-generated models frequently suffer from structural fragmentation, containing disconnected clusters or isolated nodes. This fragmentation 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 could not process "unknown" relationships. Faced with this ambiguity, the model forced simplified causal links, merged concepts inappro-

priately, or deleted them entirely.

The analyses also underscored the profound impact of the participant's communication style on model accuracy and structure. Both the Flint and Gulf Offshore Wind studies revealed that the LLM performs best when extracting from simpler, more direct interactions. In the Flint study, accuracy dropped significantly when participants employed storytelling or narrative formats. Similarly, the lowest-scoring models in the Wind 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 human-generated FCMs, 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 interview transcripts. Across the datasets, the AI consistently captured the core, major nodes without missing critical systemic drivers. However, because they lack the consistent contextual discernment and advanced systems-thinking capabilities inherent to human modelers, they 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 social-ecological systems: the resource-intensive, manual construction of FCMs. By systematically evaluating a suite of LLM extraction methods against a novel corpus of unstructured, real-world stakeholder interviews spanning four distinct domains, we have established a robust benchmark and evaluation framework for automated mental model extraction. Our findings indicate that while navigating the disfluencies, slang, and non-linear narratives of interview transcripts represents a rigorous test of AI capabilities, methods such as in-context learning with explicit explanation generation can successfully extract foundational causal structures. By publishing this comprehensive dataset of transcripts paired with expert-generated ground-truth models, we provide the research community with a vital resource for future advancements in automated FCM generation.

Our evaluation of LLM behavior reveals a critical distinction between literal transcription and functional systems modeling. The AI excels at high-resolution mapping, consistently capturing major nodes and occasionally surfacing granular socioeconomic and environmental nuances that human coders might overlook. Furthermore, we observed instances where the LLM successfully performed conceptual aggregation, binning related sub-topics into broader thematic nodes. However, LLMs currently struggle with the consistent resolution of ambiguity, the interpretation of erratic communication styles, and the complex network reasoning required to build cohesive models. The AI tends to generate networks that lack advanced causal motifs and are highly sensitive to the structural simplicity of the underlying dialogue.

Consequently, we conclude that fully automated, unsupervised FCM extraction is not yet a full replacement for human facilitation when high-fidelity, system-level analysis is required. Instead, our findings strongly advocate for a "Human-in-the-Loop" (HITL) paradigm. LLMs are highly effective at providing a rapid, foundational extraction—identifying major concepts and explicit linkages, which, as demonstrated in our offshore wind analysis, can achieve utility with minor quality control. Moving forward, research should focus on optimizing these collaborative workflows, where the LLM handles the initial, labor-intensive extraction from raw qualitative data, while human experts (or dedicated AI systems) focus on higher-order refinement: resolving narrative disfluencies, standardizing node labels, and injecting the complex mechanistic reasoning necessary to translate a literal transcript map into a viable model 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.

8 References

References

Sam Anuyah, J. Vanschaik, P. Jain, S. Lehman, and S. Chakraborty. 2025. An empirical study of causal relation extraction transfer: Design and data. *arXiv preprint arXiv:2503.06076*.

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.

Eduardo Blanco, N’uria Castell, and Dan Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 310–313. European Language Resources Association (ELRA).

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

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.

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.

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.

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*.

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 Automated FCM Extraction Pipeline

A.1.1 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.1.2 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.1.3 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.1.4 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.1.5 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.1.6 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.