# `MindGraph`: Measuring Causal Mental Models from Text with Large Language Models*

Andrea Guido
Paris School of Business

**Abstract**

Causal mental models, people's beliefs about which factors produce an outcome and how those factors are linked, are increasingly measured from open-ended text, but extracting them has required trained human coders. This paper documents `MindGraph`, a workflow in which a large language model (LLM) codes each text into a directed acyclic graph over an *a priori* taxonomy of factors and is validated as a third coder against two trained humans. On 60 open-ended explanations of climate change in English and French, three-way Fleiss' $\kappa$ is 0.66 and 0.62 for causal links and 0.80 and 0.75 for factors, and average model–human agreement exceeds human–human agreement in both languages. Agreement on factors is stable across four models; agreement on links falls with the smallest.

## 1 Introduction

A growing literature treats beliefs as causal models. In theory, agents reason with subjective directed acyclic graphs (DAGs) that may omit or misdirect links (Spiegler, 2016; Eliaz and Spiegler, 2020); empirically, households' explanations of inflation and investors' explanations of stock returns are elicited in open text and coded into such graphs (Andre et al., 2026, 2024). Management research has long studied the same object under the name of causal or cognitive maps (Axelrod, 1976; Carley, 1997), and the theory-based view casts decision makers as theorists whose causal models direct search and investment (Felin and Zenger, 2017; Camuffo et al., 2024).

The binding constraint is measurement. Mapping a text into a causal graph requires deciding which factors it names, which of them it links, and in which direction. Trained human coders make these judgments, which limits samples to what a team can read and languages to those its coders master, and leaves coding decisions hard to reproduce.

This paper documents `MindGraph`, a workflow that delegates this coding to an LLM under three constraints. The model codes *structure*, a respondent-level DAG over a fixed taxonomy, rather than labels. It is validated as a *coder*, in a three-rater design against trained humans, because no gold standard exists for a person's subjective causal model. And every annotation is tied to a hash of the prompt, the taxonomy and the model, so that results are traceable. In a benchmark on climate-change narratives, model–human agreement is at least as high as human–human agreement in two languages, at a marginal cost below $0.001 per response.

*Code, prompts, taxonomies, the annotation application and synthetic examples on which every program runs: `https://github.com/andreaguido/mindgraph` (MIT license). Survey responses, human codes and model outputs are not released. Draft of November 27, 2026; comments welcome.

## 2 Related approaches

Existing measures differ in who chooses the concepts and who decides the links. In management, causal maps are coded by hand from interviews and documents (Axelrod, 1976; Markóczy and Goldberg, 1995; Carley, 1997), while Pathfinder networks are derived from respondents' ratings of how related each pair of researcher-selected concepts is (Schvaneveldt, ed, 1990); links are then undirected and the rating burden grows with the square of the number of concepts (Lim and Klein, 2006; DeChurch and Mesmer-Magnus, 2010).

Computational approaches extract structure from text without human coding. `RELATIO` (Ash et al., 2024) uses semantic role labelling to extract agent–verb–patient statements, clusters entities into coherent groups, and maps the resulting statement networks in the U.S. Congressional Record. Its relations are explicit statements rather than causal claims. Lange et al. (2022) extend `RELATIO` with coreference resolution and link pairs of statements into causal narratives when a causal connective joins them; implicit causality is missed, and clauses without agent–verb–patient structure are lost.

LLMs have shifted the frontier. Gehring and Grigoletto (2025) classify more than one million tweets on climate policy by characters and their roles as heroes, villains or victims, a narrative dimension distinct from causal structure. Garg and Fetzer (2025) build graphs of concepts and claimed relations from 44,852 economics papers with a multi-stage LLM workflow and label each edge by whether a causal-inference design supports it. Li et al. (2025) generate event-level causal graphs from chapters and short stories by combining LLM summarization with a linguistically informed classifier, and report gains over prompting frontier models directly. Closest in spirit, Powell and Caldas Cabral (2025) validate an LLM's identification of causal claims in evaluation interviews against expert coding. More broadly, LLMs can match or exceed crowd workers on annotation tasks (Gilardi et al., 2023), but they follow codebook definitions and exclusion rules imperfectly (Halterman and Keith, 2026; Xiao et al., 2023), and benchmarks of causal-graph reasoning find only partial understanding (Chen et al., 2024). Competence on a given coding task must therefore be shown, not assumed.

`MindGraph` differs from these tools in combining a closed vocabulary, so that graphs aggregate across respondents and languages without post hoc clustering, with a respondent-level DAG whose sink is the outcome of interest, and with validation against several human coders at the level of both factors and links.

## 3 The workflow

**Taxonomy.** Before coding, the researcher fixes the factors a respondent may name as causes of the outcome, each with a definition, example statements and boundary rules that separate adjacent factors. The taxonomy is the only domain-specific input. In the application below, it has 25 factors in five groups: natural causes, structural and systemic human causes, personal and cumulative choices, political and institutional responsibility, and skepticism or misinformation. It was written before data collection from pilot responses, the climate-science literature, technical reports and press coverage.

**Coder.** Each text is coded in one stateless call to a reasoning model (`gpt-6.1-terra` at its default reasoning effort). The system prompt contains the taxonomy and three ordered tasks: identify the factors the text names; link them, with the outcome as the unique sink and no factor added at this step; and check the graph for coherence. Eleven disambiguation rules, written from disagreements in pilot and benchmark coding, play the role of the decision rules of a qualitative codebook (MacQueen

Table 1: Agreement between the LLM coder and two human coders

| | Links (edge level) | | | | Factors (node level) | | | |
|---|---|---|---|---|---|---|---|---|
| | English | | French | | English | | French | |
| | $\kappa$ | PPA | $\kappa$ | PPA | $\kappa$ | PPA | $\kappa$ | PPA |
| *Panel A. Pairwise agreement,* `gpt-6.1-terra` | | | | | | | | |
| Model vs. Coder 1 | 0.60 | 0.66 | 0.59 | 0.65 | 0.76 | 0.79 | 0.72 | 0.75 |
| Model vs. Coder 2 | 0.72 | 0.77 | 0.65 | 0.70 | 0.86 | 0.87 | 0.83 | 0.85 |
| Coder 1 vs. Coder 2 | 0.63 | 0.69 | 0.60 | 0.65 | 0.76 | 0.78 | 0.69 | 0.72 |
| *Panel B. Three-way Fleiss' $\kappa$ (model, Coder 1, Coder 2), by model* | | | | | | | | |
| `gpt-6.1-terra` | 0.66 | | 0.62 | | 0.80 | | 0.75 | |
| `gpt-5.4` | 0.64 | | 0.64 | | 0.78 | | 0.75 | |
| `gpt-5.4-mini` | 0.64 | | 0.62 | | 0.79 | | 0.73 | |
| `gpt-5-mini` | 0.57 | | 0.56 | | 0.77 | | 0.73 | |

*Notes:* 30 responses per language. Items: 514 (English) and 772 (French) directed pairs; 750 response–factor pairs per language. PPA $= 2n_{11}/(2n_{11} + n_{10} + n_{01})$. `gpt-6.1-terra` at default reasoning effort, other models at medium effort; same prompt and taxonomy throughout.

et al., 1998). Three principles govern the coding. *Parsimony*: a link is coded only when the respondent states it, and mechanisms are never inferred, because the target is the respondent's model, not the correct one. *Independence*: no other response enters the context. *Language*: texts are coded in the language in which they were written, without translation. The output is strict JSON, a list of nodes and a list of directed edges; an empty graph is a valid output.

**Auditability.** Every annotation is stored under a `coding_version`, a hash of the full prompt, the taxonomy and the model identifier. A change to any of the three invalidates earlier codes rather than silently mixing them, and each version can be traced to an exact prompt under version control.

**Validation.** Two trained coders code a benchmark sample independently, without access to model output, in an annotation application that uses the same taxonomy and returns graphs in the same format. Agreement is computed at two levels. At the *node* level, each respondent contributes a binary vector over the fixed set of factors. At the *edge* level, each respondent contributes all directed pairs among the factors used by any rater, plus links to the outcome, each scored present or absent. For each pair of raters I report Cohen's $\kappa$ (Cohen, 1960) and positive percent agreement (PPA), the share of positive codes on which the two agree; because graphs are sparse, raw agreement and negative agreement are mechanically high and PPA is the demanding statistic (Feinstein and Cicchetti, 1990). Three-way agreement is Fleiss' $\kappa$ (Fleiss, 1971). The acceptance criterion is that model–human agreement is at least as high as human–human agreement.

## 4 Benchmark

The benchmark consists of 30 English responses from an online panel (United States and United Kingdom) and 30 French responses from a laboratory session, each answering in their own words what causes climate change. Two coders, both economists working on climate policy, one of whom is the author (Coder 1), coded all 60 responses; the model coded the same items. Table 1 reports the results.

Three results stand out. First, the model meets the acceptance criterion. Averaged over the two coders, model–human $\kappa$ is 0.66 for links in English against 0.63 between the humans, and 0.62 against 0.60 in French; for factors, 0.81 against 0.76 and 0.77 against 0.69. The model agrees more with Coder 2 than the coders agree with each other, and slightly less with Coder 1 on links (by 0.03 in English and less than 0.01 in French). By the conventions of Landis and Koch (1977), three-way agreement is substantial on both links and factors.

Second, agreement on factors exceeds agreement on links for every pair of raters. Raters largely agree on which causes a respondent names and disagree on how they are connected. Inspection of the most disputed edges points to two sources: whether a cause is routed through greenhouse gases or linked directly to climate change, and the choice among overproduction, industry and economic growth for generic references to industry, capitalism or production; the disambiguation rules target both. Agreement is lower in French for the human pair as well, and the model's margin over that pair holds in both languages: the coder introduces no language-specific effect.

Third, the result does not hinge on one model (Panel B). Two further models are within 0.02 of the main model on every measure, a difference of the size expected from 30 items per language. The smallest model names the same causes (within 0.03) but agrees less on links (0.06 to 0.09 lower): coding causal structure, not identifying factors, is what demands model capacity.

## 5 Discussion

The workflow has four limitations. The taxonomy shapes every graph, so its coverage must be checked on pilot answers, and inductive construction of a taxonomy from text is a separate problem that `MindGraph` does not address. The benchmark is small, 30 responses per language, and validates English and French only; other languages require their own human benchmark. One of the two human coders is the author. Finally, the main model is proprietary and outputs vary across calls: versioning detects a change of model but cannot prevent one, which is why every change is re-benchmarked and an open-weights replication is the next step. Whether agreement carries over from short lay explanations to denser organizational text (founder pitches, shareholder letters, employee surveys) remains to be tested.

## References


**Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart**, "Narratives about the Macroeconomy," *Review of Economic Studies*, 2026. Forthcoming.

**_ , Philipp Schirmer, and Johannes Wohlfart**, "Mental Models of the Stock Market," 2024. Working paper.

**Ash, Elliott, Germain Gauthier, and Philine Widmer**, "RELATIO: Text Semantics Capture Political and Economic Narratives," *Political Analysis*, 2024, *32* (1), 115–132.

**Axelrod, Robert**, *Structure of Decision: The Cognitive Maps of Political Elites*, Princeton, NJ: Princeton University Press, 1976.

**Camuffo, Arnaldo, Alfonso Gambardella, and Andrea Pignataro**, "Theory-Driven Strategic Management Decisions," *Strategy Science*, 2024, *9* (4), 382–396.

**Carley, Kathleen M.**, "Extracting Team Mental Models through Textual Analysis," *Journal of Organizational Behavior*, 1997, *18* (S1), 533–558.

**Chen, Sirui, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, and Chaochao Lu**, "CLEAR: Can Language Models Really Understand Causal Graphs?," in "Findings of the Association for Computational Linguistics: EMNLP 2024" 2024.

**Cohen, Jacob**, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, 1960, *20* (1), 37–46.

**DeChurch, Leslie A. and Jessica R. Mesmer-Magnus**, "Measuring Shared Team Mental Models: A Meta-Analysis," *Group Dynamics: Theory, Research, and Practice*, 2010, *14* (1), 1–14.

**Eliaz, Kfir and Ran Spiegler**, "A Model of Competing Narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

**Feinstein, Alvan R. and Domenic V. Cicchetti**, "High Agreement but Low Kappa: I. The Problems of Two Paradoxes," *Journal of Clinical Epidemiology*, 1990, *43* (6), 543–549.

**Felin, Teppo and Todd R. Zenger**, "The Theory-Based View: Economic Actors as Theorists," *Strategy Science*, 2017, *2* (4), 258–271.

**Fleiss, Joseph L.**, "Measuring Nominal Scale Agreement among Many Raters," *Psychological Bulletin*, 1971, *76* (5), 378–382.

**Garg, Prashant and Thiemo Fetzer**, "Causal Claims in Economics," 2025. arXiv:2501.06873.

**Gehring, Kai and Matteo Grigoletto**, "Virality: What Makes Narratives Go Viral, and Does It Matter?," CESifo Working Paper 12064, CESifo 2025.

**Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli**, "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks," *Proceedings of the National Academy of Sciences*, 2023, *120* (30), e2305016120.

**Halterman, Andrew and Katherine A. Keith**, "Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts," *Political Analysis*, 2026, *34* (2), 188–204.

**Landis, J. Richard and Gary G. Koch**, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 1977, *33* (1), 159–174.

**Lange, Kai-Robin, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch**, "Towards Extracting Collective Economic Narratives from Texts," Ruhr Economic Papers 963, RWI – Leibniz-Institut für Wirtschaftsforschung 2022.

**Li, Zehan, Ruhua Pan, and Xinyu Pi**, "Beyond LLMs: A Linguistic Approach to Causal Graph Generation from Narrative Texts," 2025. arXiv:2504.07459.

**Lim, Beng-Chong and Katherine J. Klein**, "Team Mental Models and Team Performance: A Field Study of the Effects of Team Mental Model Similarity and Accuracy," *Journal of Organizational Behavior*, 2006, *27* (4), 403–418.

**MacQueen, Kathleen M., Eleanor McLellan, Kelly Kay, and Bobby Milstein**, "Codebook Development for Team-Based Qualitative Analysis," *Cultural Anthropology Methods*, 1998, *10* (2), 31–36.

**Markóczy, Livia and Jeff Goldberg**, "A Method for Eliciting and Comparing Causal Maps," *Journal of Management*, 1995, *21* (2), 305–333.

**Powell, Steve and Gabriele Caldas Cabral**, "AI-Assisted Causal Mapping: A Validation Study," *International Journal of Social Research Methodology*, 2025, pp. 1–20.

**Schvaneveldt, Roger W., ed.**, *Pathfinder Associative Networks: Studies in Knowledge Organization*, Norwood, NJ: Ablex, 1990.

**Spiegler, Ran**, "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

**Xiao, Ziang, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer**, "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding," in "Companion Proceedings of the 28th International Conference on Intelligent User Interfaces" ACM 2023, pp. 75–78.