Assessing change in (cognitive models of) systems over time

19 Jan 2026

The Architect and the Clerk: Analysing Central Bank speeches with GenAI-supported causal mapping

Abstract

Using Generative AI (GenAI) as a "mere assistant" to code text according to a pre-defined thematic codebook is possible but relatively uninteresting, because the AI is not involved in theory building. Many qualitative researchers are using GenAI for collaborative meaning-making, with and without a codebook. This paper presents another way, based on causal mapping, a well-established approach in social research which codes *only* causal claims within text. We use a "zero-shot" approach with no codebook, using only "In vivo" labels for the identified causes and effects. This results in heaps of causal claims containing large numbers of cause and effect labels. Making sense of these heaps is then done in two phases: firstly applying Large Language Models (LLMs) for semantic clustering and secondly using non-GenAI causal mapping techniques to visualise overall and divergent causal narratives within the text(s). This procedure is quite highly standardised and yet still depends on creative and iterative human input at key points: Qualitative judgement (what are the main cause/effect labels and how are they best organised in order to build an interesting and useful theory?) remains central while many of the other tasks become more reproducible, checkable, and scalable.
We demonstrate this approach with a corpus of speeches from leaders of Central Banks, asking: what drives what in the national and global economy, in the opinion of these experts? And how do these opinions change over time?

Introduction: Beyond the Conversation

GenAI can be used to automatically code themes according to a pre-defined codebook (Xiao et al. 2023). But this is a relatively uninteresting use of AI as it is used purely as a "clerk" and is not directly involved in theory building.

Recent debates about AI-assisted qualitative research invite us to give AI a bigger role. One way to do this is the "conversational" paradigm, where AI acts as a co-researcher in a dialogic, hermeneutic process (Friese 2025; n.d.; Dai et al. 2023; n.d.; Nguyen-Trung 2025). That approach leverages the AI's capacity to identify meanings within larger passages of text and thus in various ways to take part in a conversation around theory building.

We propose a complementary, yet distinct, path. It is based on causal mapping, a social research approach technique that identifies and visualizes beliefs about "what causes what" (Nadkarni & Shenoy 2004; Scavarda et al. 2006; Ackermann et al. 2004; Eden et al. 1992; Axelrod 1976). In causal mapping, we code *only* the causal claims within a text. This can be done with a codebook, but more interesting is with a zero codebook, using only "In vivo" labels for causes and effects.

This preliminary task can be done by a "clerk" and does not need an "architect".

Causal mapping is a very good fit for the initial step of almost-automated coding because a) causal coding is quite surprisingly easy to automate on a page-by-page basis (Studdiford & Lupyan 2025; n.d.; Veldhuis et al. 2024; Powell et al. 2025), being a much more highly determined task than, say, the task of "identify themes within this text"; and b) extracted *causal* narratives are usually more interesting and closer to answering actual research questions than lists of themes (Britt et al. 2025; n.d.).

Making sense of — and building a theory around — the resulting heaps of causal claims (containing large numbers of cause and effect labels) is then done in two phases: first, applying LLM-supported semantic clustering and then using established (non-GenAI) causal mapping techniques to visualise the overall and divergent causal narratives within the text(s). Qualitative judgement (what are the main cause/effect labels and how are they best organised?) remains a central theory-building step. Most of the other tasks are reproducible, checkable, and easy to scale.

Case study dataset: Central Bank speeches (1996–2023)

We used a corpus of central bank speeches (1996–2023) Campiglio et al. (2025): 1,354 speeches spanning 1996–2023, sampling up to 20 speeches per year, resulting in a sample of 522 speeches, equivalent to 3934 pages at 500 words/page.
We use this corpus purely as an illustrative worked example; our contribution is methodological rather than domain-substantive.

Causal coding

We applied causal coding to this dataset. The coding prompts were quite restrictive. We did not, for example, ask the AI to "summarize" or "discuss themes." We instructed it simply to identify every specific instance where the text claimed X causally influences Y, and record only cause, effect and a supporting verbatim quote, without using a pre-existing codebook. We call this minimalist or "barefoot" causal coding: capturing explicit causal claims using the source's own vocabulary.

In practice, we iterated the extraction prompt and ran basic quality checks (spot-checking pages and links) until we were satisfied that the outputs were consistent with these instructions.

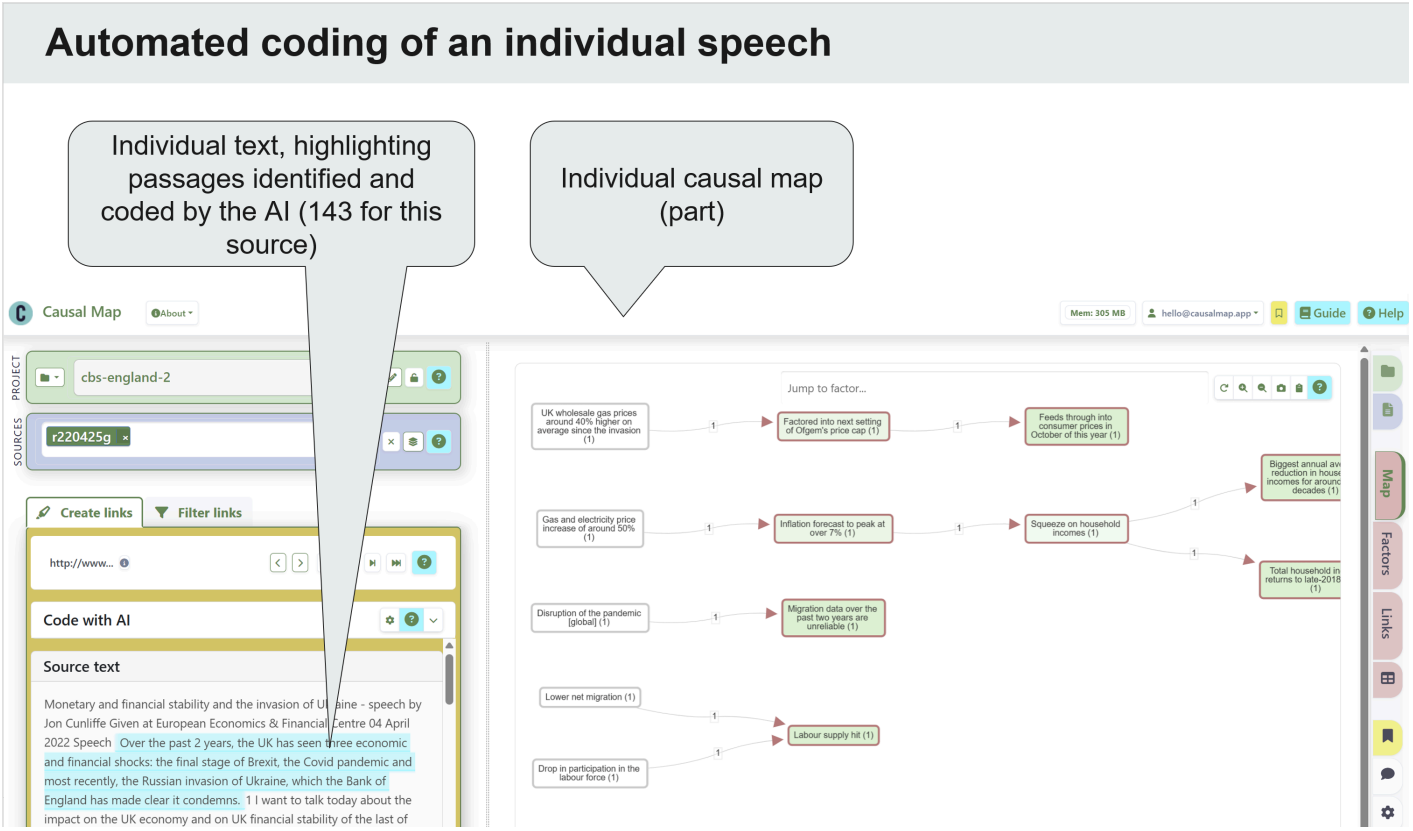


Figure 1: A screenshot of the Causal Map app after automatically coding one source.

The Creative Challenge: Magnetic Clustering as (part of) theory building

The output of this extraction is a "hairball" of thousands of links between thousands of unique factor labels often with overlapping meaning (e.g., "rising prices," "inflation," "cost of living increase"). It is here that the AI-supported process confronts the core challenge of standard qualitative inquiry: Theory building.

In most forms of qualitative coding, the researcher must decide which codes belong together to form clusters. This is a highly under-determined task. It is not completely arbitrary, because some sets of clusters will better cover the raw codes than others. But it is not well specified, because many alternative "theories" may cover the raw codes equally well. The assignment of codes to clusters as well as assessing the goodness of fit of this assignment is carried out through Magnetic Clustering, as follows.

1. **Identifying initial clusters:** Sorting thousands or, as in this case, tens of thousands of labels into clusters of similar meaning is a task which in principle humans can carry out but in practice the scale is overwhelming. So, each label (raw label and magnet) is represented as an **embedding**: a numerical vector encoding of meaning. In NLP this general idea underpins modern semantic search and vector retrieval, including transformer-based representations (e.g., BERT-style embeddings (Devlin et al. 2019)) and retrieval-augmented (RAG) pipelines that use vector indices (Lewis et al. 2021). Semantic similarity is measured by cosine similarity (the angle between embedding vectors). We then use a clustering algorithm (k-means) to create clusters of raw labels with similar meaning (Grootendorst 2022).
2. **Defining the Labels:** The next task is to generate labels for each cluster. This is a more feasible task for humans to do (usually there are 5-50 clusters with a handful of exemplars for each cluster) but in practice even this is tiresome over many iterations, so we employ GenAI to help but usually inspect and adjust the labels as our theoretical ideas of the subject matter evolve.
3. **The Attraction:** these first two steps are relatively standard procedure. But our next step differs: An algorithm "attracts" the thousands of raw, granular labels to these suggested cluster labels, which we call "Magnets", based on semantic similarity: each raw code is assigned to that Magnet to which it is semantically most similar. An assignment fails if the closeness of code to theme is below some similarity

threshold, in which case the raw code is not assigned to any Magnet. The percentage of raw codes which succeed in being reassigned or "magnetised", for any given closeness threshold, can be called "coverage". Note **we do not actually use the original cluster solution other than to provide exemplars in order to create the magnetic labels, which then create their own clusters**: the clusters of raw labels attracted to each magnet will be similar to the original solution but not identical to it. This is what makes our approach different from most clustering procedures: we **re-cluster the raw labels all over again** on the basis of their similarity to these Magnets, a process which can also be called "**soft recoding**". This means that we can tweak the Magnet labels on the fly, individually or severally, and watch as the raw labels quickly recluster themselves. We follow the principle of Braun and Clarke's (Braun & Clarke 2021) Reflexive Thematic Analysis that themes must show internal homogeneity (by using k-means clustering) and external heterogeneity (our labelling procedure, whether conducted by AI or ourselves, is explicitly designed to find labels which are distinct from one another in meaning). The original AI-generated labels provide a good start, but they are usually only the beginning of a process. For example, we might want to substitute one label with a pair of labels which are relatively similar in meaning — and so would not have been produced in the initial phase — but which represent for us an important theoretical distinction.
4. The Refinement: We iteratively tweak the "magnetism" and experiment with different lists of magnets.

Qualitative scholars usually distinguish between mere clusters of codes and *themes* which may form part of a more meaningful theory and which have a closer relationship to the research questions. In our case, it would be perfectly *possible* to use more theory-driven labels, but these would in general not function as well as magnets as they would usually be further away semantically from the original raw labels. Ideally we might discover labels which are both more interesting theoretically and nevertheless also fit the original language of the respondents. Failing this, we can use an additional filter to provide convenience labels for the original Magnets which describe them in terms of our evolving theory, for example we might relabel "Interest rate cuts" as part of a hierarchical system: "Monetary policy; interest rate cuts", in such a way that only "interest rate cuts" is given to the clustering algorithm but the full label is shown in user-facing outputs such as tables[1].

This process of Magnet selection is related to the codebook development phase in, say, Thematic Analysis, but it is a soft (temporary) rather than a hard coding procedure. It is also possible to go back to the start and "hard-recode" the source texts again from scratch, using the Magnets list as a codebook. In practice, "soft-recoding" with magnetic labels is our preferred approach because we can test and compare the "fit" of different theoretical models to the data more or less in real time.

Deciding which Magnets to use is a substantial qualitative decision that shapes the entire model and which falls under the researcher's theory-building responsibility even when using AI-assisted causal mapping.

However the results of the "magnetisation" process still does not reach the level of declarative conclusions which one would expect from a theory-building analysis. That is because the immediate result of the magnetic is not a single text but a query-able qualitative model, as discussed in the next section.

Results: A Query-able Qualitative Model

The result of this process is not a static narrative, but a dynamic causal map ,(i.e. a query-able database of causal evidence) : a large number of links between a set of common labels (the "Magnets"). It is possible to simply list the "Magnets" and the frequency with which they were mentioned, or show a map with only the most frequently mentioned Magnets. While we can create standard frequency tables (see figure 2), the most interesting part of this process is being able to interrogate this qualitative model to answer more interesting questions, by applying a library of pre-set filters individually or in chains, as discussed in the next section.

Factor	Citation Count	Source Count	Citation Count: In	Citation Count: Out
Major global events [global]	2384	295	1005	1379
Inflation	1230	221	747	483
Financial market conditions	821	258	403	418
Lack of economic growth	818	242	514	304
Financial instability	783	267	406	377
Underlying economic factors	515	228	293	222
The macroeconomic policy environment	471	179	153	318
Long-term economic trends	386	167	224	162
Financial stability	366	161	191	175
Financial crisis management	293	134	129	164

Figure 2: The most frequently mentioned causal factors, in descending order of citation count. Also showing "Source Count" (number of sources mentioning the factor at least once) and "Citation Count: In" aka "Indegree" (number of mentions of the factor as an effect) and "Outdegree" (number of mentions of the factor as a cause).[2](https://app.causalmap.app/?bookmark=966) 2026-01-12 16:44_] As an aside: Researchers used to most forms of systems modelling will find this kind of presentation confusing because some factors appear with both positive and negative variants (Financial stability and Financial instability) and some do not. But this simply reflects the sources' causal narratives. Causal mapping does not usually attempt to model a logical world of facts: it attempts to model the (often not very logical) mental models implicit within texts. So, if two opposed concepts appear separately within the text, by default we will simply use them as-is. Other techniques are available to combine opposed pairs of ideas, but they will not be discussed here. Combining opposites, sentiment.

What are the main consequences of Brexit according to the sources?

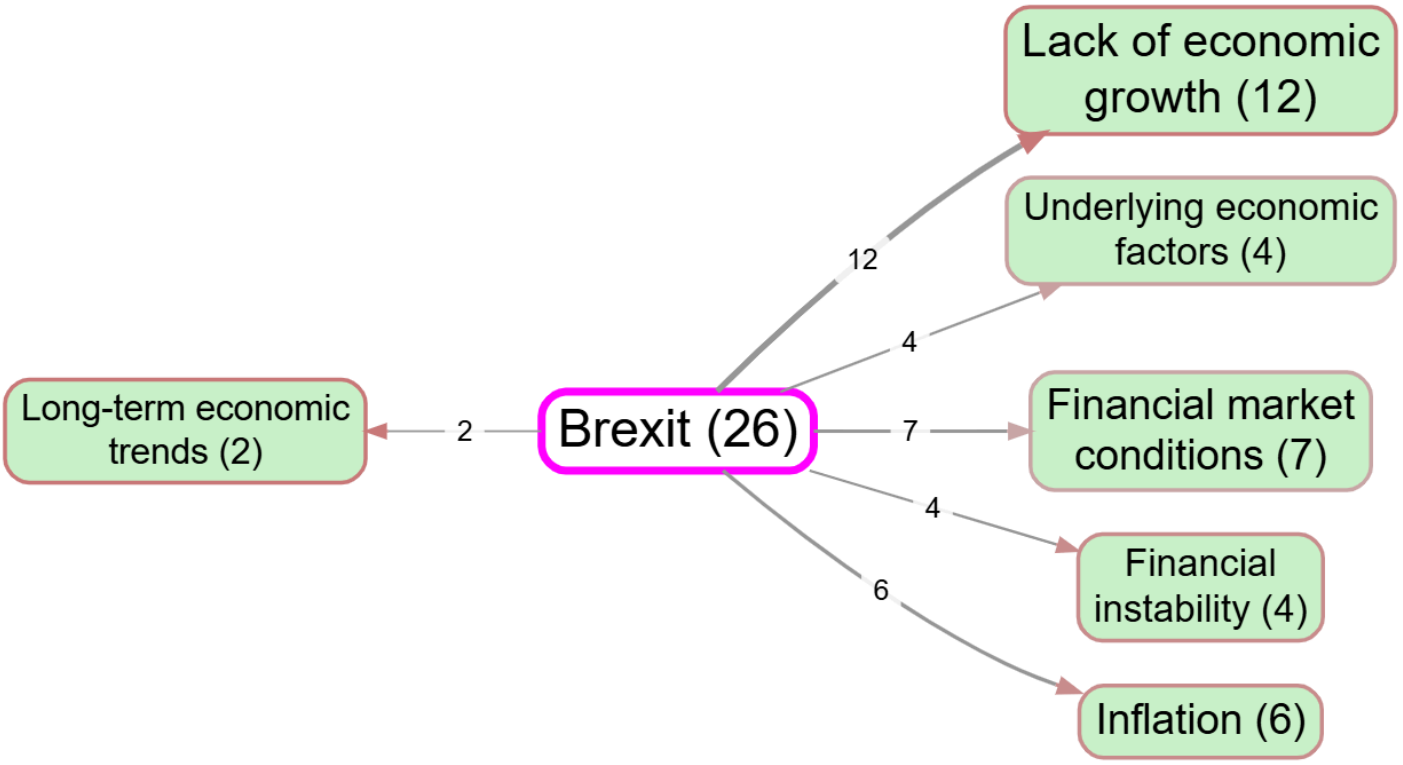


Figure 3: A screenshot of the Causal Map app after automatically coding one source. [3] In this case we apply a simple filter to explore the consequences of "Brexit": any Magnets listed as its immediate effects. We can explore the resulting map further, e.g. by inspecting the verbatim quotes associated with each link.

What are the top-level causal narratives in each decade (specifically, as explanations of growth) and how do they differ from one another?

Here, we use the same set of magnets for each decade, filter to show only paths leading to Economic growth or Lack of economic growth and then again show the most frequently mentioned links in each case.

Decade 1 1996-2005

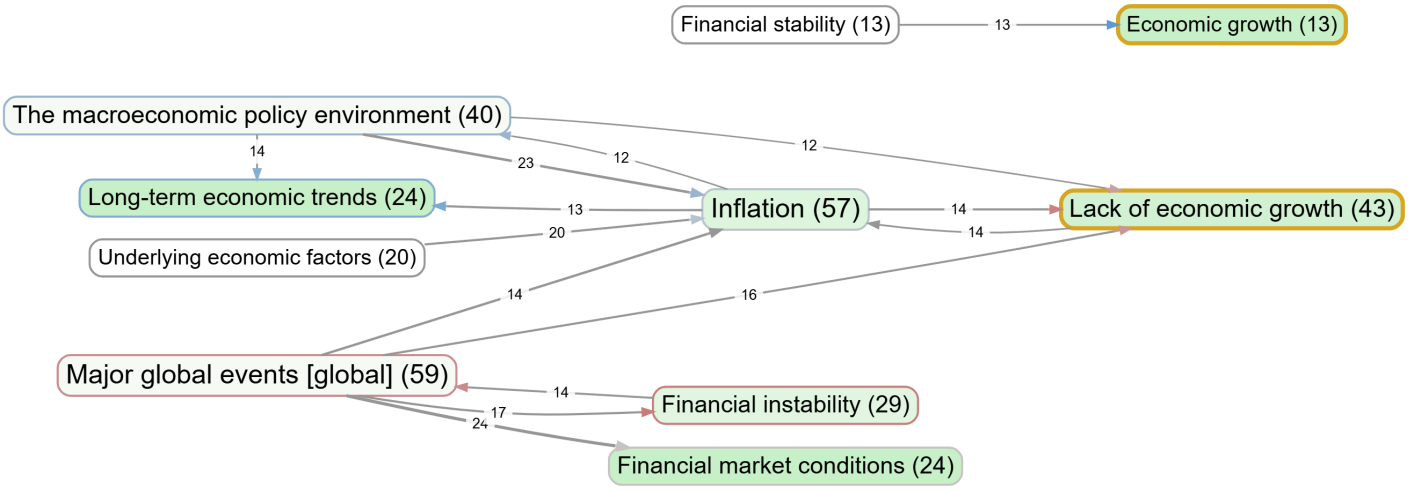
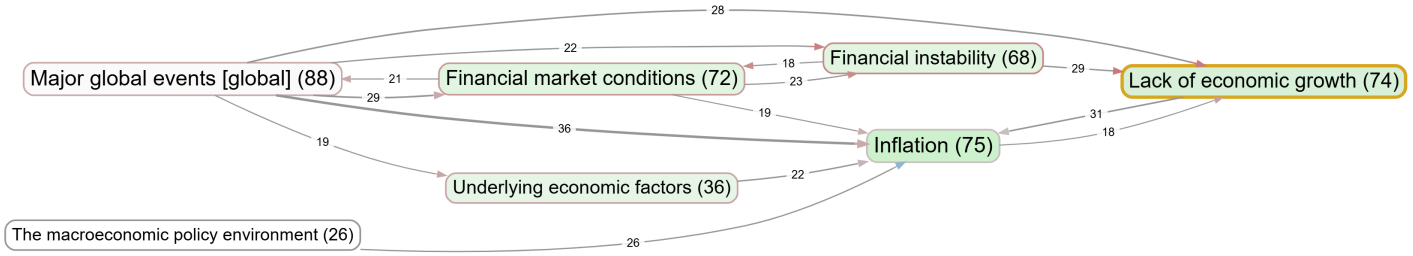


*Figure x * Filters applied: Soft recode: 18 magnets, similarity>=0.62; Source Groups: Decade=1; Path: to "Growth", 2 steps; Link frequency: top 19 by source count.*

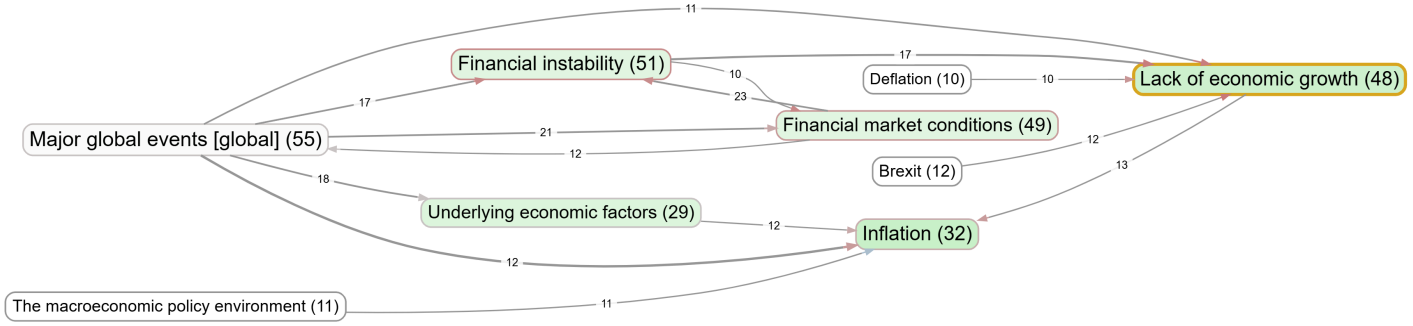
Decade 2 2006-2015



Bookmark #962 Filters applied: Soft recode: 18 magnets, similarity>=0.62; Source Groups: Decade=2; Path: to "Growth", 2 steps; Link frequency: top 19 by source count.

With these filters, retaining only the most frequent links, there is no more mention of: Long-term economic trends or of Financial stability leading to economic growth.

Decade 3 2016-2023



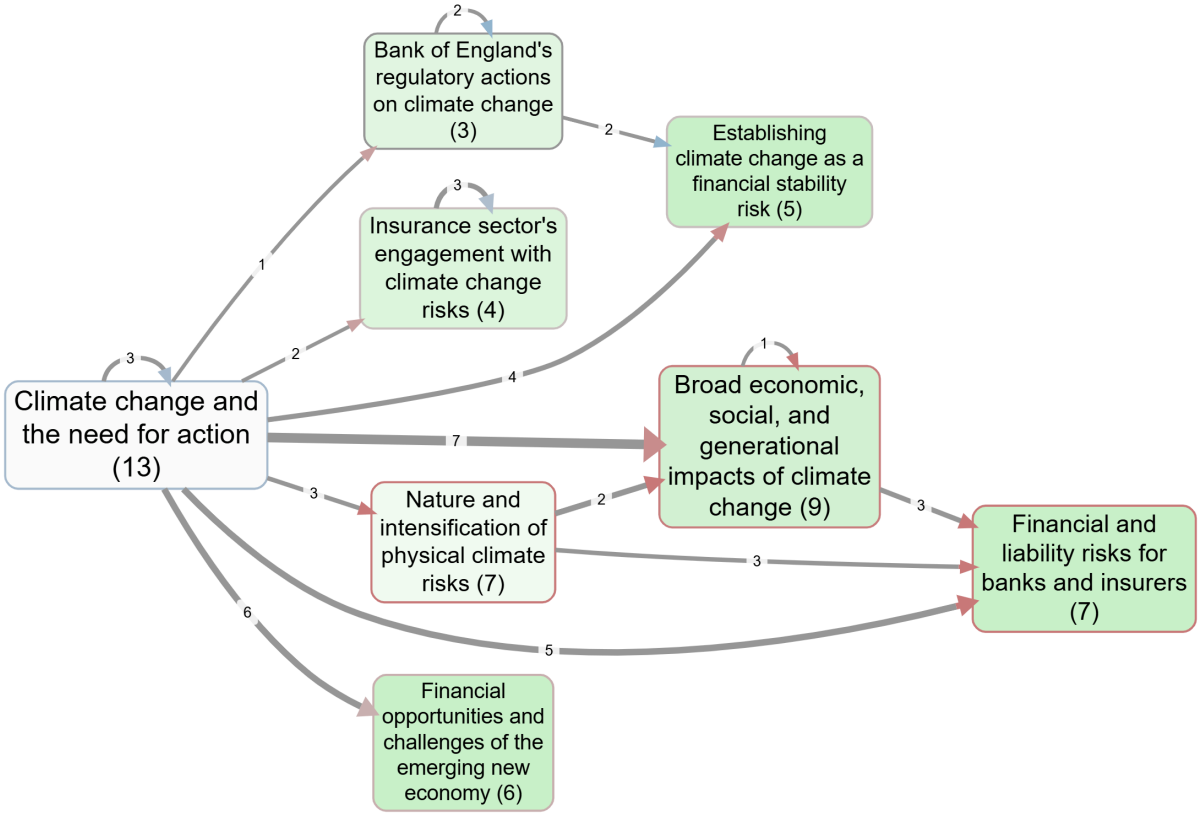
Bookmark #963 2026-01-13 09:12_

Retaining only the most frequent factors and links, now we see the first mentions of Brexit, Deflation and Climate change.

This technique allows us to systematically compare maps across groups or time-points. This is something which is much harder to do reliably using human coders. Concepts such as inter-rater reliability are often treated as a distraction by qualitative researchers and there are good reasons for that: but using the techniques described here we can have our cake and eat it: get involved in in-depth theory construction and have reliable and reproducible coding too, at the scale of hundreds of sources.

What are the narratives around climate change?

In this case, climate change and its specific causes and effects did not appear frequently enough to appear much amongst the most important Magnets. So we simply filtered for all mentions of climate change (as causes and effects) amongst the *raw* labels and created new Magnets to cover just this material.


Bank of England's regulatory actions on climate change (3)
Establishing climate change as a financial stability risk (5)
Insurance sector's engagement with climate change risks (4)
Climate change and the need for action (13)
Broad economic, social, and generational impacts of climate change (9)
Nature and intensification of physical climate risks (7)
Financial and liability risks for banks and insurers (7)
Financial opportunities and challenges of the emerging new economy (6)

Filename: cbs-england-2-filtered-2. Citation coverage 0% of all sources: 59 citations shown out of 37648. Factors — size: citation count; numbers: source count; colour: outcomeness; border: avg incoming sentiment (blue=positive, grey=neutral, red=negative). Links — width: citation count; labels: source count; arrowheads: effect sentiment (blue=positive, grey=neutral, red=negative). Filters applied: Sources included: All sources. Labels: climate chang, climate crisis, global warm; Soft recode+: 10 magnets, sim=0.62; Link freq: minimum 2 citations. Bookmark #967 2026-01-13 09:24

This is also a good place to showcase and additional way to involve GenAI in this kind of research: the entirely automated production of narrative vignettes which describe any given map and the data behind it, using a standardised prompt. The results are as follows.

Central Bankers Warn: Climate Crisis Poses Significant Threat to UK Financial Stability

This analysis, derived from a highly filtered view of UK central bankers' speeches, outlines their perspectives on climate change, focusing on its implications for the financial system and broader economy.

The discourse among UK central bankers, as captured in this data, predominantly frames **Climate change and the need for action** (40 citations) as a foundational premise. This central theme is consistently reported to lead to a range of significant negative outcomes across the economy and financial sector.

A primary pathway identified is the link between **Climate change and the need for action** and **Broad economic, social, and generational impacts of climate change** (11 citations, average sentiment -0.727). The sources frequently highlighted the severe negative implications for society and future generations. Furthermore, it was observed that climate change directly contributes to **Financial and liability risks for banks and insurers** (6 citations, average sentiment -1), indicating a clear concern for the financial sector's direct exposure.

The **Nature and intensification of physical climate risks** (11 citations) is another critical factor, with reports indicating its direct connection to **Climate change and the need for action** (3 citations, average sentiment -1). These physical risks, in turn, were mentioned as exacerbating both **Financial and liability risks for banks and insurers** (3 citations, average sentiment -1) and the **Broad economic, social, and generational impacts of climate change** (5 citations, average sentiment -1). This suggests a cascading effect where physical climate events translate into tangible financial and societal burdens.

The role of regulatory bodies, specifically the **Bank of England's regulatory actions on climate change** (8 citations), is also a notable theme. These actions were reported to contribute to **Establishing climate change as a financial stability risk** (2 citations, average sentiment 0.5), underscoring the central bank's proactive stance in integrating climate considerations into financial oversight. The sources also mentioned that **Climate change and the need for action** itself prompts these regulatory responses (2 citations, average sentiment -0.5).

While the focus is largely on risks, the speeches also touched upon **Financial opportunities and challenges of the emerging new economy** (6 citations, average sentiment -0.333), indicating an awareness of the dual nature of the transition to a greener economy.

The analysis reveals a strong emphasis on the *effects* of climate change rather than its scientific *causes*. The factor **Climate change and the need for action** serves as the primary causal node, acting as a given and urgent reality. From this starting point, the discourse largely explores the subsequent impacts on financial stability, economic performance, and societal well-being. The central bankers' speeches, therefore, function more as an assessment of the consequences and necessary responses to climate change, rather than an exploration of its origins. The most cited links consistently describe negative effects, such as the "broad economic, social, and generational impacts" and "financial and liability risks for banks and insurers," highlighting a clear concern for the downstream implications of a changing climate.

Discussion: Complementarity and the Architect

The “dialogue” described here is not between human and GenAI directly, but between human and the causal mapping process, mediated by GenAI which does the original coding, suggests magnetic labels and provides summary vignettes (and by the LLMs which provide the label embeddings and underly the clustering process). Although the procedure presented here is quite highly standardised in outline, each step may be iterated several times either alone, or going back one or more steps to revise preceding steps as well. What causal mapping adds is an almost universally applicable way to extract and visualise *what led to what*: a key, theory-adjacent aspect of almost any set of narratives.

The "conversational paradigm" makes liberal use of GenAI as a human-like research assistant who is asked questions like "identify the themes" or "summarise this document", even at the very beginning of the research, following an argument that coding (Nguyen-Trung & Friese 2025) is a perhaps redundant, "skeuomorphic" remnant of the age before AI. But giving such immense degrees of freedom to any interpretative task exposes it to large, arbitrary and poorly defined influences from the analyst, whether human or machine. At least we can hope that the human analyst may be at least partially aware of their own influence on such tasks, due to mood, tiredness or positionality or whatever. Humans' ability to actually understand the influence of these factors on their analysis is of course limited and error-prone. Machines are less likely to be influenced by situational factors but their performance is of course massively influenced by their architecture and training data in a way that they are unlikely to be actually aware of, regardless of what they say if we ask them. Coding is one way to reduce this exposure to arbitrary influences.

This paper complements the conversational paradigm and shows that standardising and automating many parts of the workflow does not mean dumbing it down. By delegating the massive cognitive load of extraction and clustering mostly to the AI, we free the researcher to focus on the architecture of meaning. The "Architect" is asked for creative input in a well-defined way at well-defined, specific points in the workflow. But it remains a demanding task. Getting the clusters "right" and applying the right filters to create relevant maps requires deep, iterative engagement with the research question. The AI provides the bricks; the human must still design the house.

AI Contribution Disclosure Checklist

- Research Design: Human led (definition of Causal Mapping logic).
- Data Collection: Human/Existing Data.
- Data Analysis (Prompt engineering): human.
- Data Analysis (Coding): AI (Radical Zero-Shot extraction).
- Data Analysis (Clustering): Collaborative (AI performed clustering; Human defined "Magnets" and iteratively refined structure).
- Data Analysis (Answering questions): Mostly human, leveraging existing non-AI algorithms/filters.
- Initial drafting of Paper: AI (Gemini), based on human-provided structural constraints and source files.
- Refining and Editing: Human.

1. These labels, like every label in causal mapping, can also form a hierarchical structure — "Improved economic conditions; improved investor confidence" as well as "Improved economic conditions", but we do not use this extensively in the current study). ↩

2. _Filename: cbs-england-2-filtered-2. Citation coverage 8% of all sources: 3065 citations shown out of 37648. Filters applied: Sources included: All sources. Soft recode+: 18 magnets, sim=0.62; Link freq: top 40 by source count. Bookmark [#966 ↩

3. Legend: Factors — size: citation count; numbers: source count; colour: outcomeness (darker=more incoming links); border colour: average incoming sentiment (blue=positive, grey=neutral, red=negative). Links — width: citation count; labels: source count; arrowheads: sentiment of effect (blue=positive, grey=neutral, red=negative). Filters applied: Soft recode: 18 magnets, similarity>=0.62; Path: from "Brexit", 1 steps; Link freq: minimum 2 sources. ↩

References

Ackermann, Eden, & Cropper (2004). *Getting Started with Cognitive Mapping*.

Axelrod (1976). *The Analysis of Cognitive Maps*. In *Structure of Decision : The Cognitive Maps of Political Elites*.

Braun, & Clarke (2021). *Thematic Analysis : A Practical Guide*. SAGE Publications Ltd. https://www.torrossa.com/it/resources/an/5282292.

Britt, Powell, & Cabral (2025). *Strengthening Outcome Harvesting with AI-assisted Causal Mapping*. https://5a867cea-2d96-4383-acf1-7bc3d406cdeb.usrfiles.com/ugd/5a867c_ad000813c80747baa85c7bd5ffaf0442.pdf.

Campiglio, Deyris, Romelli, & Scalisi (2025). *Warning Words in a Warming World: Central Bank Communication and Climate Change*. https://doi.org/10.1016/j.euroecorev.2025.105101.

Dai, Xiong, & Ku (2023). *LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis*. https://arxiv.org/abs/2310.15100v1.

Devlin, Chang, Lee, & Toutanova (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805.

Eden, Ackermann, & Cropper (1992). *The Analysis of Cause Maps*. https://onlinelibrary.wiley.com/doi/10.1111/j.1467-6486.1992.tb00667.x.

Friese (2025). *Conversational Analysis with AI - CA to the Power of AI: Rethinking Coding in Qualitative Analysis*. https://doi.org/10.2139/ssrn.5232579.

Grootendorst (2022). *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*. https://doi.org/10.48550/arXiv.2203.05794.

Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, Küttler, Lewis, Yih, Rocktäschel, Riedel, & Kiela (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. http://arxiv.org/abs/2005.11401.

Nadkarni, & Shenoy (2004). *Nadkarni and Shenoy 2004 -A Causal Mapping Approach.Pdf*.

Nguyen-Trung, & Friese (2025). *On Methodological Incongruence in Applying Generative AI in Qualitative Data Analysis*. https://doi.org/10.2139/ssrn.5874482.

Nguyen-Trung (2025). *ChatGPT in Thematic Analysis: Can AI Become a Research Assistant in Qualitative Research?*. https://doi.org/10.1007/s11135-025-02165-z.

Powell, Cabral, & Mishan (2025). *A Workflow for Collecting and Understanding Stories at Scale, Supported by Artificial Intelligence*. SAGE PublicationsSage UK: London, England. https://doi.org/10.1177/13563890251328640.

Scavarda, Bouzdine-Chameeva, Goldstein, Hays, & Hill (2006). *A Methodology for Constructing Collective Causal Maps**. https://doi.org/10.1111/j.1540-5915.2006.00124.x.

Studdiford, & Lupyan (2025). *Contextual Effects in LLM and Human Causal Reasoning*. https://openreview.net/forum?id=BMHkg3BL6e.

Veldhuis, Blok, family=Boer, Kalkman, Bakker, & family=Waas (2024). *From Text to Model: Leveraging Natural Language Processing for System Dynamics Model Development*. https://doi.org/10.1002/sdr.1780.

Xiao, Yuan, Liao, Abdelghani, & Oudeyer (2023). *Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding*. In *28th International Conference on Intelligent User Interfaces*. https://doi.org/10.1145/3581754.3584136.