RUHR
ECONOMIC PAPERS

Tobias Schmidt
Kai-Robin Lange
Matthias Reccius
Henrik Müller
Michael Roos
Carsten Jentsch

Identifying Economic Narratives in Large Text Corpora – An Integrated Approach Using Large Language Models

technische universität dortmund
RUHR UNIVERSITÄT BOCHUM RUB
#1163

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Ludger Linnemann
Technische Universität Dortmund, Department of Business and Economics
Economics – Applied Economics
Phone: +49 (0) 231/755-3102, e-mail: Ludger.Linnemann@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83 - 3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Almut Balleer, Prof. Dr. Manuel Frondel,
Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49 - 213, e-mail: presse@rwi-essen.de

Editorial Office

Niels Oelgart
RWI, Phone: +49 (0) 201/81 49-213, e-mail: niels.oelgart@rwi-essen.de

Ruhr Economic Papers #1163

Responsible Editor: Carsten Jentsch

ISSN 1864-4872 (online) – ISBN 978-3-96973-348-6

Ruhr Economic Papers #1163

Tobias Schmidt, Kai-Robin Lange, Matthias Reccius,
Henrik Müller, Michael Roos, and Carsten Jentsch

Identifying Economic Narratives in Large Text Corpora – An Integrated Approach Using Large Language Models

technische universität dortmund

RUHR UNIVERSITÄT BOCHUM RUB

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

https://dx.doi.org/10.4419/96973348
ISSN 1864-4872 (online)
ISBN 978-3-96973-348-6

Tobias Schmidt, Kai-Robin Lange, Matthias Reccius, Henrik Müller,
Michael Roos, and Carsten Jentsch*

Identifying Economic Narratives in Large Text Corpora – An Integrated Approach Using Large Language Models

Abstract

As interest in economic narratives has grown in recent years, so has the number of pipelines dedicated to extracting such narratives from texts. Pipelines often employ a mix of state-of-the-art natural language processing techniques, such as BERT, to tackle this task. While effective on foundational linguistic operations essential for narrative extraction, such models lack the deeper semantic understanding required to distinguish extracting economic narratives from merely conducting classic tasks like Semantic Role Labeling. Instead of relying on complex model pipelines, we evaluate the benefits of Large Language Models (LLMs) by analyzing a corpus of Wall Street Journal and New York Times newspaper articles about inflation. We apply a rigorous narrative definition and compare GPT-4o outputs to gold-standard narratives produced by expert annotators. Our results suggests that GPT-4o is capable of extracting valid economic narratives in a structured format, but still falls short of expert-level performance when handling complex documents and narratives. Given the novelty of LLMs in economic research, we also provide guidance for future work in economics and the social sciences that employs LLMs to pursue similar objectives.

JEL-Codes: C18, C55, C87, E70

Keywords: Economic narratives; natural language processing; large language models

July 2025

*Tobias Schmidt, TU Dortmund University; Kai-Robin Lange, TU Dortmund University; Matthias Reccius, Ruhr-University Bochum; Henrik Müller, TU Dortmund University; Michael Roos, Ruhr-University Bochum; Carsten Jentsch, TU Dortmund University. – This study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University and the Narrative Economic Alliance Ruhr (NEAR) project, supported by the Mercator Research Center Ruhr (MERCUR) with project number Ko-2022-0015. It was also partially funded by the Reality Check incubator project at the Research Center for Trustworthy Data Science and Security. – All correspondence to: Kai-Robin Lange, TU Dortmund University, e-mail: kalange@statistik.tu-dortmund.de

1. Introduction

Macroeconomic policy decisions are not made in a vacuum. More and more, central banks and other institutions recognize the role that public discourse and widely shared narratives play in shaping expectations and thus economic behavior. The central bank communications literature in particular has described the active role of policy makers in explaining their policies to great detail (Gorodnichenko et al., 2023; Gurkaynak et al., 2005; Hansen et al., 2017, 2019). But neither fiscal nor monetary policy discourses are top-down processes. People's economic views and expectations are increasingly shaped by the media (Fiore et al., 2025). Particularly in times of uncertainty, stories about rising inflation, looming recessions, or job market disruptions spread rapidly and uncontrollably.

As economic dynamics are driven by the beliefs and expectations of households and firms, monitoring the narratives that dominate public discourse is crucial. Quantifying economic narratives in mass media can be challenging, however, as the amount of information transmitted is vast and it's rhetorical packaging diverse. Many researchers perform qualitative analyses, which require expert evaluations of texts. While such procedures produce reliable results, they are not scalable to large text corpora. Increasing numbers of articles, reports, opinion pieces and even comments on social media are released every day, so purely relying on qualitative research is not feasible when analyzing the narratives that circulate in an economy. Therefore, economists need efficient, quantitative methods of extracting and presenting narratives from texts.

State-of-the-art language models such as BERT (Devlin et al., 2019) and its variations have shown considerable success in tasks like sentiment analysis, named entity recognition, and semantic role labeling, which are seen by many as foundational to the identification of economic narratives. However, while these models excel at narrow linguistic tasks, they often fall short when tasked with the deeper, more nuanced understanding required to distinguish between complex economic narratives and simpler textual structures. Pipeline approaches that mix different methods also introduce compounding error risks due to their multiple processing stages. Hence, traditional NLP models address the scalability problem but lack the integrated language comprehension needed to fully capture the contextual meaning and economic relevance of narratives.

This gap presents an opportunity for exploring alternative approaches that can offer greater contextual understanding. The advent of Large Language Models (LLMs), with both commercial models such as GPT (Brown et al., 2020), Claude Sonnet and Google Gemini and open source models like Llama (Dubey et al., 2024), has opened up en-

tirely new possibilities for narrative extraction. LLMs have demonstrated remarkable proficiency in a wide array of tasks, excelling especially in those that require a deeper comprehension of language and context (Shahriar et al., 2024). These models are trained on vast amounts of data and can process complex narratives with a level of understanding closer to that of human readers. But how should economists leverage LLMs to accurately measure complex phenomena such as narratives? And can LLMs match expert annotators in identifying and extracting economic narratives from newspaper articles?

In this paper, we use a set of complex instructions and expert-labeled examples to evaluate the proficiency of LLMs at extracting economic narratives. We utilize `GPT-4o` (OpenAI et al., 2024b), the leading commercial LLM at the time of analysis, for the extraction of inflation narratives from a corpus of Wall Street Journal and New York Times newspaper articles. We further demonstrate how complex concepts like economic narratives can be operationalized for human coding, and then adapted for coding by LLMs. While such models enable new avenues for quantitative economic research, the litmus test for their usefulness lies in the rigorous validation of their outputs. Hence, we compare the results of the LLMs' analysis against a gold-standard set of narratives, representing the consensus of three trained annotators with domain expertise. A detailed and rigorous annotator codebook was developed through multiple iterations and refinements. We find that expert annotators reach notably different results extracting narratives, pointing to the inherent subjectivity involved in narrative sense-making. To refine and adjust our evaluation of GPTs' performance to the subjectivity of the task, we not only consider the gold-standard data set as a set of optimal answers, but also the deviations made by the expert annotators, which we call "expected" deviations. We also compare the results attainable through a set of different LLM prompting strategies recently found to be the most effective by research in Artificial Intelligence (AI).

The rest of the paper is structured as follows. In Section 2, we shed light on previous research that has aimed to define economic narratives or automatically extract them from texts. We also provide a short overview of inflation-related literature, as inflation is the center of our evaluation strategy. In Section 3, we introduce the data as well as the LLM we use in our experiments. Section 4 covers our narrative extraction codebook as well as a description of the process of creating a gold-standard data set. In Section 5 we explain our prompting strategy and compare it to other approaches used in the field of AI to maximize LLM performance. The results of the experiments are then displayed and interpreted in Section 6 where we also provide an outlook for future research. Finally, we conclude in Section 7.

2. Related Work

2.1. Defining economic narratives

The field of narrative economics has gained remarkable prominence in recent years (Roos and Reccius, 2024). Only eight years after the well-known paper by Shiller (2017) on this topic, it has become widely accepted in the economic literature that individuals do not simply react to economic data, but interpret the world through personal or publicly shared interpretations of what is going on. Such stories, or narratives, are expected to have the potential to shape personal expectations and even guide collective behavior, when widely shared (Bénabou et al., 2018; Flynn and Sastry, 2024; Larsen and Thorsrud, 2019). This relationship appears particularly relevant in times of uncertainty, when traditional expectation formation models struggle to capture real-world behavior. As King and Kay (2020) argue, under conditions of radical uncertainty, economic agents are unable to maximize utility in the conventional sense. Instead, they turn to shared sense-making structures to reduce complexity and navigate decision-making.

Despite the growing recognition that narratives constitute a promising field of research, the definition of what precisely qualifies as an "economic narrative" remains fragmented, with researchers highlighting different aspects depending on their analytical objectives. Early contributions, most notably Shiller (2017), describe economic narratives as broad stories that convey interpretations of economic events, morals, or simplified theories. While Shiller's intuitive approach to narratives captures their communicative power, high-level concepts like "moral" or "interpretation" resist the consistent formal structure required for empirical analysis. More recent studies, therefore, aim to articulate definitions that are both theoretically rigorous and operationalizable.

A central contribution to the formalization of economic narratives comes from Eliaz and Spiegler (2020, 2024), who conceptualize narratives as simplified causal models represented by directed acyclic graphs (DAGs). Drawing on Bayesian network theory, their 2020 model shows how individuals adopt competing narrative-policy pairs that interpret long-run correlations to maximize anticipatory utility. Narratives, in their framework, are not neutral representations of data but selective causal stories that distort objective relationships by omitting relevant variables or misattributing causal directions. In their 2024 extension, the authors apply this logic to media markets, showing how media platforms strategically supply both biased information and empowering narratives to boost consumer engagement. The empirical study by Andre et al. (2024) follows a similar logic. Specifically, the authors define narratives as backward-looking causal accounts of recent

events, or explanations that people construct in order to make sense of what has happened, and that in turn shape their forward-looking expectations. The DAG-structure of narratives that is pervasive in this line of research can be viewed as a key contributor to the simplifying and organizing function narratives fulfill for individuals trying to make sense of information. Roos and Reccius (2024) synthesize these and other perspectives to propose a definition of collective economic narratives. They argue that economically relevant narratives are not just individual stories, but shared sense-making structures that arise in a social context, explain economic events, and suggest collective action. For a more detailed overview over different definitions of economic narratives, we refer readers to this work.

A recurring insight across the literature is that causality is the core ingredient of any narrative. It is what distinguishes a narrative from adjacent concepts such as topics or frames. While a topic identifies what is being discussed, a narrative links who did what to whom, and with what consequence. It attributes responsibility, defines trajectories, and frames problems in a way that invites action. Strikingly, "who is to blame for the problem" (Crow and Jones, 2018) is central to understanding how narratives influence public opinion and policymaking. Against this background, the present study puts particular emphasis on recognizing this specific property of narratives. In order to automatically identify narratives in large text data, detecting causal connections between events—implicit as well as explicit connections—is paramount.

Building on this literature, we define economic narratives as causal connections between two temporally and semantically distinct events, formulated in the structure *A causes B* or *A is caused by B*, that reflect an interpretative framing of economic developments. For example:

1. the prices for cucumbers rose by 100% this month - causes - people stop buying cucumbers
2. increasing money supply - causes - prices for real estate go up
3. the FOMC raised the policy rate - causes - turmoil on Wall Street

Crucially, not every causal claim qualifies as a narrative in our framework. Narratives, as we understand them, are more than just descriptions of factually accurate chains of events; they often imply a perspective, or a selective emphasis that helps make sense of economic developments. To this end, for example, we distinguish between entities that qualify as genuine *events* (such as "inflation is on the rise") and those that do not (such

as "high prices", which has no temporal dimension). For a detailed operationalization of our definition, see Section 4.

2.2. Extracting economic narratives from texts

Historically, narrative extraction has been performed qualitatively. Hoping to create a more scalable solution, researchers have switched their focus to quantitative NLP methods to extract narratives from texts automatically. To be able to extract complex narratives however, classic NLP tasks such as topic modeling or sentiment analysis do not suffice. Each of these methods extracts information that is only a small part of most narrative definitions. For instance, topic modeling extracts latent topics from texts, enabling researchers to make an educated guess as to what the documents are about. This can be interpreted as the setting for a story, containing all necessary places and characters, but falls short of connecting these individual parts to a coherent sense-making story.

As a result, researchers have increasingly turned to pipeline-based approaches. Ash et al. (2024) propose such a pipeline, named RELATIO, that heavily relies on semantic role labeling to extract narrative actors. They define narratives as a connection between so called "agents" and "patients", where agents actively perform an action that affects patients. Notably, the causal component inherent in most theoretical approaches to narratives is not considered. For example, Ash et al. (2024) consider "ECB raise interest rate" to be a narrative, with "ECB" as the agent, "interest rate" as the patient and "raise" as the corresponding verb that defines the connection between the two. Lange et al. (2022a) extend this concept by incorporating additional NLP tasks into the pipeline, such as coreference resolution and causal discovery, with the aim of connecting agent-patient pairs to form a sense-making story. However, they also acknowledge that complex pipelines with many interdependent components can cause even minor errors to cascade into serious downstream failures. This sensitivity has motivated the search for integrated models capable of handling these tasks within a unified framework.

Fixing this major issue of narrative extraction techniques, while simultaneously not watering down the definition, requires NLP models of a different architecture. We therefore propose to use LLMs to extract economic narratives from texts. For conceptual clarity, we use the term LLM only when referring to generative language models that have undergone instruction tuning. This limits the scope of the term to models like `GPT-4o`, while excluding families of discriminative language models like BERT. Given the "world knowledge" and language understanding capabilities encoded in LLMs, it is possible to prompt

a model with the definition of an economic narrative and extract those narratives directly without the need to run the text through an error prone pipeline. Previous works have shown that LLMs are capable of solving related tasks, such as summarization, speaker attribution and Retrieval Augmented Generation (Bornheim et al., 2024; Fan et al., 2024; Zhang et al., 2025).

Recent research further underscores the potential of LLMs for content analysis. Studies in computational social science have shown that LLMs can match or even surpass human coders in annotating political, social, and economic texts (Ziems et al., 2024). Mellon et al. (2024) reports that LLMs achieved 95% agreement with expert annotators when analyzing British election statements. Similarly, Gilardi et al. (2023) demonstrates that LLMs like `GPT-3.5` could classify tweet content, author stances, and narrative frames more accurately than trained crowd workers. The applicability of LLMs to economic narratives is highlighted by Gueta et al. (2025), who use `GPT-3.5` to extract macroeconomic narratives from Twitter data. Although their operationalization of narratives relies heavily on sentiment analysis and topic modeling, their work demonstrates that LLMs are capable of handling unstructured textual data effectively. Schmidt (2025) makes use of the large language model (LLM) `Claude 3.5 Sonnet` to identify predefined inflation narratives in the German media coverage in 2022. Using a detailed codebook and exhaustive prompting techniques, the author was able to detect the same inflation narratives proposed by Andre et al. (2024) and track their appearance in coverage, demonstrating that LLM prompting is a promising approach to extract backward-looking "blame" narratives from text. Lange et al. (2025) showed that combining scalable methods such as topic modeling with LLMs can lead to additional benefits and can be utilized to extract shifts in complex economic narratives over time. Instead of analyzing every available document with an LLM, the authors propose to only extract narratives at points in time in which a change in narrative is suspected. They detect change points in the word-topic distributions of the dynamic topic model RollingLDA (Rieger et al., 2021) using bootstrap percentile tests (Lange et al., 2022b; Rieger et al., 2022). Afterwards, the authors utilize the LLM `Llama 3.1 8B` (Dubey et al., 2024) to categorize the changes according to the Narrative Policy Framework from political science (Jones and McBeth, 2010; Schlaufer et al., 2022; Shanahan et al., 2018). Such an analysis can be adapted to work with other change detection methods (e.g. Benner et al., 2022), allowing researchers to use a detection method suited for their purposes. This approach of Lange et al. (2025) does, however, also reveal the limitations of `Llama 3.1 8B`, as it infers narratives from most inputs, even when none are present. In our approach, which covers a more complex notion of an economic narrative, we therefore opt for a high-performing model to observe what kind of results a state of the art model can produce.

Given the increasing feasibility of narrative extraction through LLMs, a natural next step is to explore substantive domains where narratives matter most. The topic of inflation is particularly instructive in this regard, as few economic issues are more closely tied to public sentiment, expectation formation, and media framing.

2.3. Inflation-related literature

Inflation is particularly interesting in the context of narrative extraction. First and foremost, the topic is widely researched due to its formal relation to personal beliefs. In standard New Keynesian frameworks, expectations are not merely passive reflections of current conditions but actively shape future inflation trajectories (Werning, 2022). Agents form beliefs about inflation that, in turn, influence their consumption, pricing, and wage-setting behavior (Bachmann et al., 2015; Burch and Werneke, 1975; Juster et al., 1972). In this context, economic narratives play an important role, as narratives help individuals understand the world around them and make decisions. This tight linkage between belief formation and macroeconomic outcomes is one reason why inflation narratives have attracted more research attention than, say, labor market narratives. Central banks in particular have shown strong interest in the topic. In the modern forward guidance era, they aim to actively shape the inflation expectations of households and firms (Blinder et al., 2008). Given this policy regime, it is unsurprising that a growing number of central bank researchers are engaged in studying the role of narratives in the expectation formation process (Kalamara et al., 2020; Nyman et al., 2021; Ter Ellen et al., 2022; Tuckett et al., 2020).

One central finding from inflation expectation research is that media coverage appears to play an important role in expectation formation processes. As Conrad et al. (2022) show in the German context, traditional media consumption is associated with more accurate perceptions of past and expected inflation, particularly when media coverage is intensive. Similarly, Lamla and Lein (2014) find that intensive inflation reporting increases the forecast accuracy of households, while negative or sensationalist framing can lead to exaggerated inflation perceptions. Ter Ellen et al. (2022) further show that narrative monetary policy shocks (i.e., stories about interest rate decisions) can affect real macroeconomic variables if successfully disseminated.

Another central contribution in this domain is provided by Andre et al. (2024), who analyze the narratives individuals construct to explain the recent surge in inflation. Drawing on open-ended survey responses from thousands of U.S. households, the authors find that

while expert narratives predominantly attribute inflation to demand-side factors (e.g., fiscal and monetary expansion), household and manager narratives are more heterogeneous and often emphasize supply-side shocks or political mismanagement. Through randomized priming experiments, the authors show that media narrative exposure causally shifts beliefs, highlighting the importance of inflation narratives in expectation formation.

Other lines of research suggests, however, that responsiveness of expectations to media coverage varies over time. Building on the rational inattention framework (Reis, 2006; Sims, 2003), Coibion and Gorodnichenko (2015) and Bracha and Tang (2025) provide evidence that attention to inflation is cyclical. As the authors show, media attention to inflation spikes in periods of high inflation and declines otherwise. Recent research by Schmidt et al. (2023) references to the same effect by analyzing how varying levels of inflation reporting affect inflation expectations. The authors make use of the Inflation Perception Indicator (IPI), which tracks thematic shifts in German inflation coverage, to analyze narrative shifts in German inflation reporting. Employing a threshold VAR framework, the authors show that the influence of media coverage on inflation expectations is regime-dependent: only during high-inflation periods does narrative intensity significantly affect expectations and real variables.

In what follows, we put these insights into practice. Using inflation as a well-studied and socially salient topic, we apply an LLM-based narrative extraction approach to a corpus of media reports. Although inflation serves as the empirical domain of this paper, our methodology is generalizable to other economic issues. Our goal is not only to assess the performance of LLMs in a high-stakes context, but also to demonstrate how economic narratives can be systematically identified in large text corpora.

3. Model and data

In our experiment, we aim to get the best performance possible given the current generation of LLMs. In this section, we argue why the model `GPT-4o` (OpenAI et al., 2024b) is our model of choice to accomplish this and describe how we have created the data set for our experiment.

3.1. Choosing an LLM

At the time of analysis, `GPT-4o` (OpenAI et al., 2024b) is the newest iteration in the 4th generation of the Generative Pre-trained Transformer family of models by OpenAI. We chose this particular LLM—model snapshot `gpt-4o-2024-11-20` —over other commercial models, such as `GPT-4` (OpenAI et al., 2024a), as it offers comparable or superior performance while being more scalable due to lower financial costs. Its performance in human reasoning and language understanding tasks is especially impressive (Shahriar et al., 2024), as these are crucial components of narrative extraction, which requires deep understanding of complex economic circumstances. In particular, we choose to use `GPT-4o` over the more recent class of reasoning models, such as `OpenAI-o1` (OpenAI, 2024), as our prompt requires Chain-of-Thought-prompting (Wei et al., 2022) tailored specifically to our narrative definition. As recent research shows that prompting reasoning models with additional Chain-of-Thought sequences does not improve performance (DeepSeek-AI, 2025; Wang et al., 2024), we opt for our own specialized prompting over the generic reasoning offered by models like `OpenAI-o1`. Our prompting strategy is described in detail in Section 5.

While analyzing the potential of large commercial models to extract economic narratives is valuable, their use also entails notable downsides. For one, the cost of using such models on large data sets is very high and might not be feasible for universities, research centers or companies with low financial backing. Additionally, since these models are not publicly available and operate on proprietary hardware, their long-term availability is uncertain, as companies like OpenAI are likely to prioritize profit over maintaining access to older models. Therefore, all evaluations performed by older models will ultimately lose their reproducibility, undermining a cornerstone of good scientific practice. For this reason, the use of the commercial model `GPT-4o` in this paper should be understood as a theoretical example of what state-of-the-art LLMs are currently capable of, and what open-source models may achieve in the future as their development progresses.

3.2. The text corpus

We evaluate the proficiency of `GPT-4o` on the task of economic narrative extraction using a curated corpus of Wall Street Journal and New York Times news paper articles published between 01-01-1985 and 27-09-2023. We filtered for articles containing the words “inflation” or “price (increase|hike|surge)”. This time period was selected because it includes both high- and low-inflation phases. Focusing on inflation-related documents allows us

to work with a thematically coherent corpus, enabling aggregation and comparison of narratives, even within a relatively small dataset.

To further narrow down the potential economic narratives in these documents, we provide the model with a short excerpt rather than the full article. Each excerpt includes the sentence containing the filter terms, along with the two preceding and two following sentences. This allows us to focus only on the topic of inflation while still providing enough context to interpret potential narratives, including those that span multiple sentences.

From this corpus, we randomly sample 100 documents to create gold-standard responses for the LLM. While a larger annotated dataset is always desirable, we prioritize a thorough annotation of a smaller subset over a broader but less precise coverage. The annotation task is complex, demands sustained focus, and is prone to errors when done hastily or without sufficient care.

4. The codebook and the gold-standard

When trying to operationalize rich concepts like economic narratives for empirical research, we often struggle to translate verbal definitions into a rigorous and workable codebook. Aspects of narratives that seem sensible enough in a verbal explanation are sometimes hard or even impossible to identify in an empirical setting. This problem gets exacerbated when a codebook is required to apply to LLM-based extraction in addition to human coders. While LLMs are trained to appear as human-like partners in interactions, they differ substantially from humans in the way they process information when performing tasks (Mondorf and Plank, 2024).

Therefore, we will proceed by clarifying what defining aspects of economic narratives we consider to be important in an empirical setting. Based on these aspects, we create a codebook designed for human coders to create a consensus gold-standard data-set. We then translate the human codebook into a prompt designed to maximize the LLM's performance, accounting for the model's peculiarities relative to human annotators.

4.1. A narrative codebook

Translating any of the narrative conceptualizations presented in Section 2 to empirical work entails some fundamental challenges. These definitions operate with high-level concepts such as *event*, *story* and *moral*. As a result, when two experts are given only a definition to guide text annotation, their results are likely to differ significantly. Because a narrative is tied to language, domain understanding, and the reader's interpretation of the author's intent, no narrative extraction can be truly objective. Additionally, from a purely linguistic perspective, narrative extraction is not a simple span detection task in which one specific part of a text must be marked. Instead, key contextual information given in the text must be extracted and synthesized into a coherent structure. These structures must also be amenable to aggregation when comparing narratives across large corpora.

We propose a codebook that both narrows down the type of narratives that are to be annotated and outlines a clear structure for the annotation process. The codebook is the end product of multiple workshops that were held with experts and non-experts. These sessions aimed to develop clear guidelines for extracting economic narratives, minimizing common coding errors and sources of ambiguity wherever possible. The entire codebook can be found in Section A in the appendix. The most important aspects of the codebook can be summarized as follows.

- **Goal**: The main task is to accurately extract narratives from newspaper excerpts, defined as a causal link between two consecutive events. Multiple researchers should code the same texts to minimize subjective bias.

- **Target Form**: Each narrative must be reduced to one of two forms: *event 1 - causes - event 2* or *event 1 - is caused by - event 2*. To preserve the natural order of the text, the first event in the source should remain as *event 1* in the coded narrative.

- **Event Types**: Permissible events include occurrences, activities, conditions, future events, plans and policies. The coded event should remain as close as possible to the original wording, avoiding synonyms or abstractions.

- **Coreference Resolution**: When entities are referred to indirectly (e.g., with the use of pronouns), the coder is to replace them with the correct entity to maintain clarity (e.g., "{He|President Biden}").

- **Statements**: When an event consists of a statement made by an individual, coders must distinguish whether the causal link to another event pertains to the content of the statement or to the statement itself. Typically, the focus should be on the content. However, if the statement itself triggers a reaction, it should be coded as the event. Typical examples include forward-looking statements in corporate disclosures or forward guidance issued by central banks.

- **Embedded Clauses**: Coders should ignore non-essential elements within events (e.g., appositions, parentheses) unless they contribute meaningfully to the narrative.

- **Chained narratives**: Each event in a chained narrative (e.g. "event 1 causes event 2; event 2 causes event 3") should be coded separately.

- **Narrative forks**: When an event is caused by multiple events or an event causes multiple other events, each causal connection has to be noted as a separate narrative unless the combination of events is integral to the narrative.

- **Positive Causality Only**: Only positive causal links should be coded, negative ones (e.g., "does not cause") should be excluded.

- **Economic Focus**: Only narratives with an explicit economic context should be coded. Non-economic narratives should be ignored.

- **Maintain Full Meaning**: The full message of the original narrative is to be retained without omitting crucial information, even if some parts seem superfluous.

- **Edge Cases**: When a code is unsure if a sentence contains a causal narrative, they code it anyway and resolve uncertainties later with other coders.

4.2. Creating a gold-standard data set

To annotate our 100 randomly sampled documents, three expert annotators —familiar with the definition of economics narratives, experienced in applying it and proficient in English at a near-native level—were provided with the codebook. Each annotator was tasked with coding all narratives from the 100 documents independently, without discussing their results with the other annotators. To address any remaining language barriers, such as idiomatic expressions or culturally specific references, annotators were

given access to a built-in DeepL API, allowing them to translate the text into their native language if needed. In such cases, narratives were still coded based on the original English version.

In regular intervals during the annotation process, the annotators met to revise their understanding of the codebook and discuss edge cases they were uncertain about.

After coding all 100 documents individually, the annotators met in multiple sessions to discuss the results. Each individually coded narrative was discussed by all three annotators. If they unanimously agreed it was correctly coded, the narrative was added to the list of gold-standard narratives for the corresponding document. In this process, it did not matter whether a narrative had been identified by one, two, or all three annotators. Given the subjectivity of the task and the possibility that even experts may overlook narratives without considering multiple perspectives, inclusion of a narrative in the gold standard was based solely on substantive discussion grounded in the codebook. Such procedures have been shown to produce more valid and reliable results than simply applying majority voting (Burla et al., 2008; Neidhardt, 2010).

The individual annotations by each annotator, however, were not discarded. Instead, we use them to compute an *expected deviation* from the gold-standard, representing the level of variation that is acceptable among expert annotators. The output of the LLM is then compared to this baseline. We also document the reasons for each deviation from the gold-standard labels. Suppose an event has multiple consequences—i.e. *Event A* causes three distinct events: *B*, *C* and *D*. Annotator 1 did not split these into three individual narratives but instead coded *Event A* as causing the combined outcome of *B*, *C* and *D*. This would be classified as a *minor deviation*, with the reason noted as "missing multiple consequences". This approach of differentiating major from minor deviations allows us to formalize complex patterns within our codebook while still comparing human and AI performance effectively. *Major deviations* either completely alter the meaning of a narrative, identify a narrative that does not exist or miss an existing narrative. Minor deviations, by contrast, occur when the overarching narrative is correctly identified but is either formally incorrect (e.g. due to missing coreference resolution) or contains subjective elements that differ from the gold-standard annotation.

In total, the gold-standard data set contains 291 narratives in 100 documents.

5. Prompting LLMs to extract narratives

Prompting language models to extract and annotate data in research is a relatively new venture. We feel strongly that LLMs open entirely new avenues for quantitative economic research. Korinek (2023) was among the first to urge economists to explore this potential. However, as of the time of this writing, the major use case for LLMs in economic research is sentiment analysis (Examples include Jeong and Ahn (2025) and Han et al. (2024)), for which high-quality models exist that do not require Generative AI. Due to the novelty of LLMs, no best practices for crafting prompts have been established to date in any of the social sciences. However, recent literature in AI and computational linguistics has introduced a set of general strategies that can enable researchers to make informed choices. We hope that our task can provide a useful example for related studies, and we will cite relevant work throughout. Using `GPT-4o`, we explore multiple approaches to extract economic narratives informally. We then narrow down the more useful principles which we subject to a formal evaluation. We require narratives to be extracted in a consistent format that can be parsed by machines. Therefore, we always use the "JSON" option in the OpenAI API, which guarantees that the model will output consistent Java Script Object Notation (JSON) files.

5.1. Prompt optimization: strategies and challenges

The least data-intensive yet most heuristic prompting strategy is zero-shot learning (Kojima et al., 2022; Wei et al., 2021). Generally speaking, LLM-assistants like ChatGPT are fine-tuned for instruction-following, which is the reason users can interact with them in a way that feels natural and human-like. By using zero-shot prompting, researchers can exploit this property by only supplying the model with the research question and task instruction along with the text to perform inference on. To set a benchmark for comparison to human performance, we use this zero-shot strategy, supplying our entire unabridged codebook as an instruction. The results are very inconsistent, both regarding the form of the outputs and in their faithfulness to the codebook.

We also explore a variation of this strategy, using a condensed version of our codebook. While current LLMs use interpolation techniques (Zhong et al., 2025) to expand hard context windows beyond the length of any common-sense codebook, it remains unclear how well performance is maintained as context size increases. Due to the intransitivity of LLM architectures, few theoretical guarantees can be considered for applied work. However, empirical investigations have documented phenomena such as the lost-in-the-

middle effect (Liu et al., 2024), which suggests that increasing the length of inputs tends to result in subtle forms of information loss, mostly in the middle of input sequences. The distribution of information in prompts appears to be a relevant criterion as well. Tian et al. (2024) suggest that a higher relative distance between crucial pieces of information in a prompt adversely affects information retrieval rates, which could compromise tasks like ours. Hence, ceteris paribus, limiting prompt length is advisable. However, the level of detail in our codebook must still cover the intricacies of the coding task. To find the sweet spot regarding this prompt-length trade-off, we condense our codebook for several iterations, emphasizing different aspects in each version. We evaluate by probing the outputs that result from using different codebook variations.

Generally, we find that increasing the information density of the codebook tends to be beneficial for performance. However, a priori, it is not obvious what kinds of information should be compressed and what concepts must be dealt with extensively in the instructions. For example, several passages in the codebook deal with the definition of an event, including detailed elaborations on the types of events we consider to be the most relevant constituents of economic narratives. These include the implementation of a policy or the future-facing intentions of firms or policy makers. We endowed human coders with a great level of detail here partly because issues involving the delineation of events had come up regularly during the codebook workshops. However, as it turns out, `GPT-4o` defaults to a conceptualization of events that closely corresponds to our definition. Hence, simply mentioning that narratives consist of two events turns out to be the optimal level of information we can provide about events. By contrast, the ordering of events in the output, which we require to remain unchanged from the source text, turns out to be a problem that needs explicit mentioning in multiple spots of the instructions.

In general, simply adding more requirements to prompts does not reliably increase LLM-performance. Additional aspects often introduce competing constraints that the model must resolve implicitly (Yang et al., 2025). Researchers should therefore prioritize which concepts to explain in detail. Morover, since an LLM's conceptual understanding is shaped by its training data and the alignment techniques used during training — both of which are unknown to researchers — practitioners should always probe the model's sensitivity to the way information is provided, even for seemingly uncontroversial concepts.

In addition to descriptive instructions, various sections of the codebook (see Section A) contain synthetic examples of full narratives and of narrative components. These examples are meant to highlight specific aspects of the extraction task for human coders to focus on when familiarizing themselves with the coding process. Such partial demonstrations are essential for designing high-quality codebooks, as they have shown to be

helpful for human-coders (Saldaña, 2016). Conversely, they do not contribute positively to LLM performance on our task. The model appears to struggle with synthesizing these aspect-driven examples into a consistent, multi-aspect narrative definition, which adversely affects model performance.

5.2. Few-shot Chain-of-Thought prompting

Given the results of the initial explorations, we decide to fully adapt our original codebook for LLM inference. Instead of using instructions that mirror the structure of the original codebook outlined in Section 4, we only provide the model with a succinct description of the key concepts and processes that it contains. Crucially, along with these instructions, we supply hand-labeled input-output examples. This widely used method is called few-shot learning in the AI literature (see Song et al. (2023) for a review). Unlike the supervised learning paradigm that most empirical economists will be familiar with, effective few-shot learning only requires a handful—rather than hundreds—of hand-labeled training examples. In contrast to the strategy outlined in Section 5.1, where distributed examples throughout the codebook each highlight a specific aspect, few-shots are complete input-to-output demonstrations that are provided in a dedicated block after the verbal instructions.

The few-shot paradigm allows the model to infer information about the desired output from full-fledged examples rather than from lengthy prosaic instructions. Consequently, the selection of few-shots and their total number are crucial. Few shots should be representative of the full data-set, ideally without adding redundancies. During this final round of prompt engineering, we use rigorous formal validation: From our set of 100 hand-annotated, gold-standard documents, we randomly select 20 for cross-validation. Any of these 20 examples can be used either as few-shots in prompts or for evaluating the performance of a candidate-prompt. The rest of the gold-standard documents are held out entirely for testing the performance of the final prompt. Given our validation set, we explore using between 1 and 9 few-shots in a single prompt. After evaluating the performance on the remaining validation example, holding constant the rest of the instructions, we settle for using 7 few-shots. Finally, we run evaluations with rotating sets of seven examples to identify the few-shot combination that yields the best evaluation results. Intuitively, the optimal set captures the most relevant narrative characteristics for the model to learn the task effectively.

We combine few-shot learning with the Chain-of-Thought (CoT) prompting paradigm, which means that we specify intermediate steps for the model to take when constructing an answer (Wei et al., 2022), with the few-shot examples mirroring this step-wise format. When dealing with problems that require complex forms of reasoning, dividing up a task into sub-problems turns out to be advantageous. This is because LLMs can only "reason" about a problem insofar as they can produce tokens about it. The CoT-logic of transforming narrative extraction into a set of sequential tasks allows the model to devote more computational power to each step in the reasoning chain. Every step must be framed as a discrete action and output separately by the model before moving on to the next step. In addition to enhancing performance, CoT-prompting also increases transparency, as errors in LLM inference can be traced to particular steps in the chain.

Conceptually, the CoT paradigm is reminiscent of classic pipelines approaches to economic narrative extraction (see Section 2), where sub-tasks are handled by different language models rather than a unified LLM (Ash et al., 2024; Lange et al., 2022a). However, the CoT pipeline merely imitates a full separation into sub-task. The LLM always has access to the prior steps since they are stored in its context. Since our codebook contains inter-dependent tasks that require economic judgment and adherence to linguistic principles, this feature is highly desirable.

5.3. An integrated narrative extraction pipeline

Our final prompt relies heavily on the few-shot CoT paradigm introduced in the previous section. A detailed description can be found in Section C in the appendix. The prompt is divided into two main parts, *Codebook* and *Examples*, where the former subsumes the written instructions and the latter only encompasses the few-shots.

The first subsection of the codebook, *Basic Idea*, is designed to prime the model by broadly outlining the task and the source medium of the input texts. Subsection 2, *Definition*, briefly states the crucial structural elements needed in every extracted narrative:

An economic narrative consists of exactly two events
and a causal connection that is asserted between those events.

Subsection 3 is called *Event structures*. It introduces three distinct types of linguistic constructions in which causally linked events are typically embedded when they are discussed

within the article excerpts. We build on the formalism of graphical models introduced by Pearl (2009), which is widely used in the AI field and in parts of economics, to distinguish *direct causal effects* from causal *chains* and *forks*. A causal chain occurs if an *Event A* causes an *Event B* which in turn causes an *Event C*. A causal fork occurs if an *Event B* is a common cause for two Events, *A* and *C*. Embracing this terminology enables us to represent causal structures as directed acyclic graphs (DAGs), as is common practice in the literature on economic narratives. Using this established formal language also allows us to optimally leverage `GPT4o's` pre-training, because we do not have to explain concepts in the instructions to the LLM as long as they are encapsulated in the model's world knowledge. As stated in Section 4, we require chains and forks to be coded as separate narratives by the model. While some of the few-shots contain examples of such causal constellations, the model turned out to struggle with them, which prompted us to incorporate them into the *Codebook* section. Hence, apart from raising awareness about the aforementioned causal construct, subsection 3 also sets up expectations on how the LLM should deal with these structures when they arise.

Subsections 4 and 5, *Rules* and *Target forms*, precisely establish the set of allowable forms that extracted narratives are expected to conform to. We assert that the expression denoting the causal connection between events in a narrative must be classified into either "causes" or "caused by", depending on the direction of the causal relationship. As laid out in Section 2, causality is a core ingredient of economic narratives, because causal connections enable individuals to make sense of why events have unfolded. Many economic choices by households and firms are based on this sense-making inference. Quite commonly, however, causal connections that appear obvious to human readers are not written down explicitly in the source text. One of the main advantage of using LLMs for our task is their ability to recognize a diverse set of explicit and implicit causal connections from a given context. While extracting implicit causality—when related events are not connected using a specific causal cue—is challenging, it is equally important. For example, when events occur at different points in time, authors often imply causal relationships through temporal cues. Distinguishing whether the author intends to signal causality or merely describes a sequence of independent events often depends on contextual knowledge. LLMs excel at such tasks, which require deep language understanding and nuanced interpretation. This capability sets them apart from dictionary-based methods that rely on explicit cue words to detect causality or language models like BERT (Devlin et al., 2019) that require fine-tuning.

In the final subsection, *Extraction process*, we describe the CoT that the model should produce for every input. Figure 1 displays this procedure for an example text.

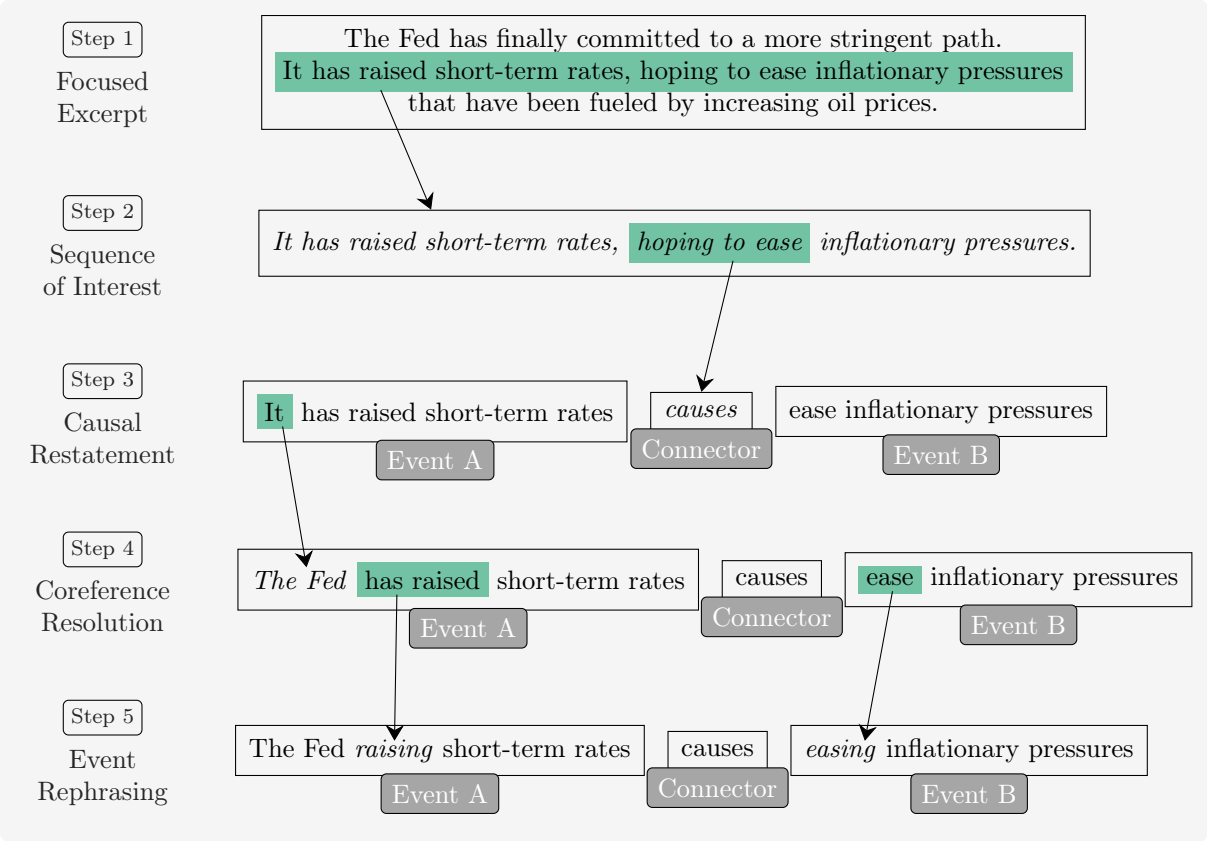

Figure 1: A diagram depicting the Chain-of-Thought transformations that our prompting strategy induces. The chain extracts an economic narrative from an example document and molds it into a standardized form step by step. Green highlighting indicates segments to be changed in the subsequent step, arrows map the source segments to their corresponding results. Results are cursive.

As mentioned before, LLM performance hinges on steering the model's attention mechanism to relevant sections of the inputs. Step 1, therefore, consists of extracting a *focused excerpt* from the input, which is an article fragment that features a term deemed indicative for an inflation discourse. Since the input texts vary significantly in their narrative density, it is helpful to first let the model disregard any sections from it that do not contain any events or causal relationship. This initial recognition of a narrative sequence is already non-trivial. In our example, the events are the Fed's realized policy measure—raising interest rates—and it's intention of easing inflationary pressures in the economy. Their causal connection is hypothetical because the outcome event has not occurred and is not guaranteed to occur at all. These nuances must be recognized by the LLM in the first CoT step, at least to the degree that it will include the segment in the focused excerpt instead of disregarding it altogether.

Step 2 delineates a single, previously unextracted narrative from the focused excerpt by producing a *Sequence of interest*. The task requires the LLM to isolate exactly two events that are causally connected, while disregarding any unrelated elements found in the text. Sequences of interest are essentially full narratives in their rawest form. In our example, the focused excerpt actually features two more events—the Fed committing to a more stringent path and increasing oil prices. While either of those might be important for further narratives, Step 2 requires the LLM to not attend to them for the time being. The *Causal restatement* in Step 3 serves to separate the events from the causal connector and to make the latter explicit. Collectively, these first 3 CoT steps perform the "heavy lifting" of narrative recognition.

Steps 4 and 5 only apply for a subset of extracted narratives. Steps 4 replaces pronouns in the sequence of interest with the entities they refer to. This task is known as *coreference resolution* in the NLP literature and is generally handled with high accuracy by the LLM. If required, this step is crucial because entities that act or are acted upon are essential component of many narratives (Gehring and Grigoletto, 2023). Step 5 consists of some final grammatical adaptations: For the sake of event aggregation, we require events to be grammatically interchangeable. *Event rephrasing* standardizes grammatical structure, allowing for easier extraction, comparison, and inference over events. This step mostly entails nominalization, in which a verb phrase is rephrased as a noun phrase.

After Step 5, we redirect the LLM to the focused excerpt to check for further narratives.

To validate our prompting strategy, we use between 1 and 9 few-shots from the validation sample and evaluate performance on the remaining 11 examples. In our experiments, performance peaks with 7 few-shots. Subsequently, we test and evaluate `GPT-4o` with our chosen prompt on the remaining 80 documents with gold-standard labels. We set the LLMs temperature hyperparameter, which governs the randomness or creativity of its responses, to 0.2. This setting makes responses more deterministic than the default while retaining enough variability to allow the model to explore the space of possible answers. For the purpose of reproducibility, eliminating any randomness by using a temperature setting of 0 would be preferable. However, a fully deterministic setting is known to degrade performance of generative language models, causing what is known as the *likelihood trap* (Huang et al., 2023; Zhang et al., 2021).

5.4. Narrative abstraction and event clustering

After extracting all narratives in the form of *Event A causes Event B* from the 80 test documents, we apply a series of post-processing steps aimed at enabling large-scale abstraction and aggregation of recurring economic patterns. The goal of this procedure is to move beyond individual instances and identify broader narrative structures. The following steps represent an initial attempt to structure and cluster narrative content in a principled way. While the primary focus of this paper lies in the extraction of narrative pairs, we briefly outline this pipeline to illustrate how the extracted data can be further processed and organized for downstream applications.

Event decomposition. First, we split compound events into distinct atomic propositions. Events such as "Energy and food prices are on the rise" were forked into two independent entries, such as "Energy prices are on the rise" and "food prices are on the rise". This decomposition is essential, as early tests revealed that our narrative extraction model occasionally fails to fork such multi-entity constructions in the desired way (see above). Accordingly, this step functions as a correction layer that may become redundant in future implementations if the model's extraction capabilities improve. We implement this step using a few-shot prompt with `GPT-4o-mini`, designed to detect conjunctions and rewrite them into separate, grammatically coherent event statements. This approach ensures that each entry represents a clearly interpretable unit of analysis. By contrast, a simple regex-based "and"-splitting approach would often yield semantically incomplete fragments, such as "Energy" in the upper example, that lack narrative coherence.

Valence and topic assignment. Next, each atomic event is passed through another LLM-based classification step to extract both its semantic *topic* (e.g., "inflation", "interest rates") and its directional *valence* (e.g., "rising", "falling", "high"). The prompt follows a structured format with few-shot examples and requires the model to output a JSON object containing both events A and B and their respective valences, thereby ensuring consistent outputs across all extractions.

Topic normalization and abstraction. Since we prompted our initial narrative extraction model to stick as closely to the original wording of the text as possible during inference, many of the extracted narratives vary lexically while describing the same or similar phenomena (e.g., "interest rates" vs. "borrowing costs"). Therefore, we implement a semi-automated abstraction step to cluster semantically similar topics. First, we embed all extracted topic strings using the `all-MiniLM-L6-v2` sentence-transformer model. Second, we map each topic embedding to a controlled set of embeddings representing different

macroeconomic categories (e.g., "government spending", "interest rates", "inflation"). To compare the topic embeddings and the predefined anchor terms, we use cosine similarity metrics, as they are invariant to the high number of embedding dimensions. Previous research has shown that clustering by anchor terms can serve as an alternative to fine-tuning based classification approaches in an unsupervised setting and help to interpret embeddings (Lange et al., 2024; Mathew et al., 2020). We take care to preserve economically meaningful distinctions, e.g., "inflation expectations" and "inflation" remain separate due to their distinct roles in economic theory and policy, and even in public understanding. We then manually refine the mappings by carefully inspecting cluster compositions. In cases where embedding similarity alone appears insufficient, we use domain knowledge and manually add recurrent topic synonyms to the respective topic cluster or remove misclassified terms. Via pattern-based string matching we then replace all topics that were assigned to a certain cluster with the respective topic label. This hybrid approach ensures that topic abstraction remain both semantically robust and economically meaningful. An example of this mapping is provided in Figure 2.

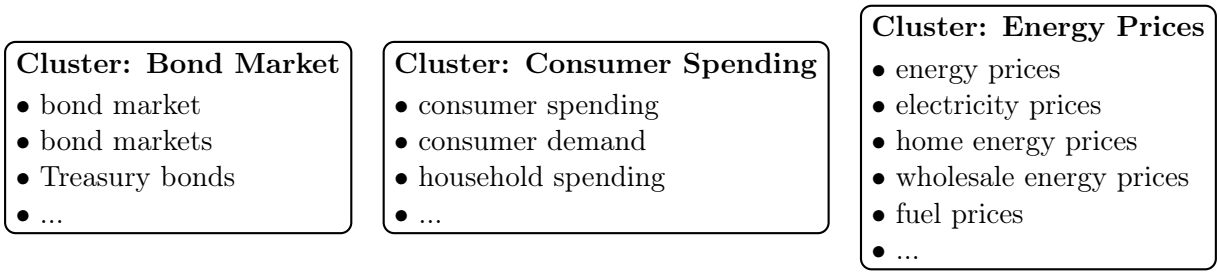

Figure 2: Example illustrating how events are grouped by a cluster topic.

Finally, we translate all valence expressions indicating an increase or general high level of something with an upwards-pointing arrow (↑) and all expressions indicating low levels or a decrease with a downwards pointing arrow (↓). In cases where implicit negations are involved, these are reversed to ensure that, for example, *rising economic vulnerability* and *rising economic stability* are represented by different arrows, even if the term "rising" is used in both contexts.

We are well aware of the information loss this step inevitably entails and the economic limitations that come with it. For instance, real interest rates differ fundamentally from nominal interest rates, which in turn are not the same as borrowing costs. However, we argue that such simplifications are justified given the inherent nature of narrative formation (i.e., the tendency to reduce complexity) and the context of our corpus, which consists of general-interest newspaper articles. Distinctions that are highly relevant in academic economics (e.g., between monetary aggregates or types of interest rates) may

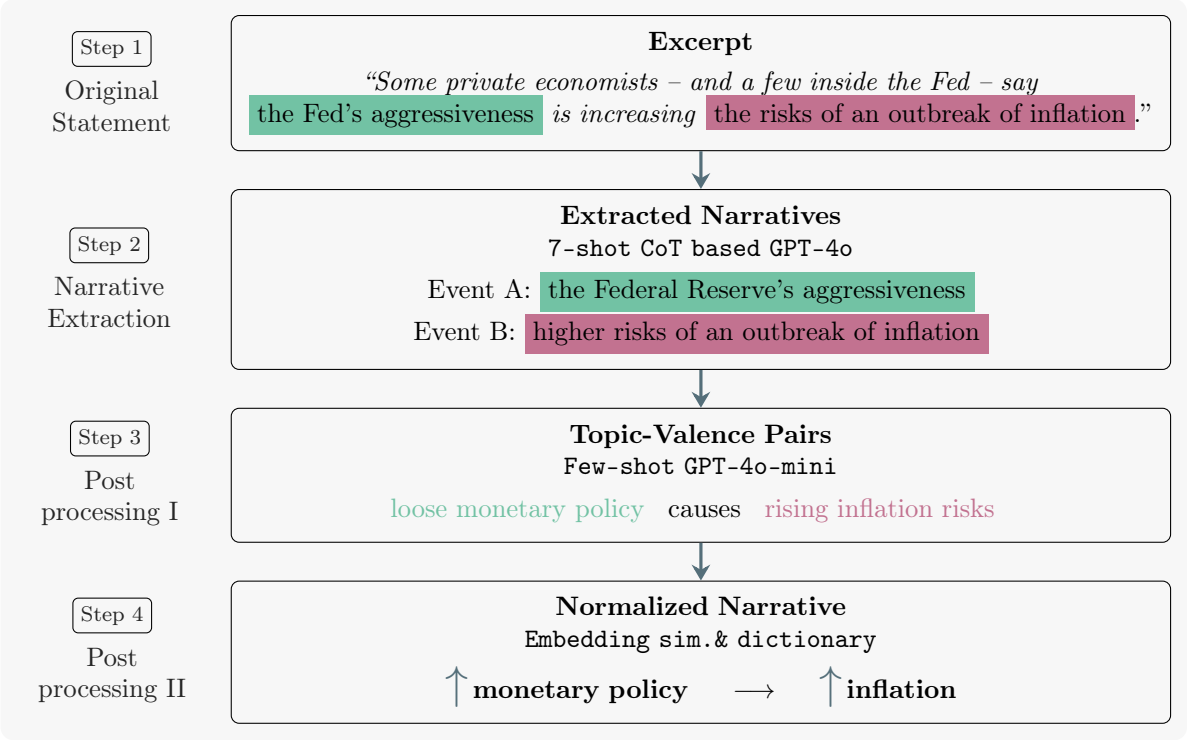

Figure 3: Illustration of our narrative abstraction pipeline using a real example from the corpus.

be less salient or even irrelevant to the typical audience of U.S. dailies. In some cases, however, this loss of specificity results in tautological or arbitrary narrative formulations, such as *rising interest rates cause higher interest rates*. While such artifacts are rare in our dataset, they illustrate one of the limitations of semantic abstraction: when the model reduces different expressions to a shared label, it may inadvertently collapse distinct concepts into self-referential loops. We acknowledge this as a trade-off inherent to our current aggregation strategy, though it affects only a small number of cases and might be mitigated through more refined post-processing in future applications. Figure 3 visualizes the final extraction pipeline using a real example from our corpus.

6. Results

Using the codebook presented earlier, we let `GPT-4o` process the remaining 80 documents that were neither used for evaluation nor for model prompting. The results offer a nuanced picture of the model's ability to extract economic narratives from text. We will first discuss the results that we have attained through prompting the LLM and then turn to the critical post-processing step of aggregating the extracted narratives. Despite the

complexity involved in creating the gold-standard, comparing `GPT-4o` to this benchmark turned out to be straight-forward, since essentially all disagreements had been resolved beforehand.

The dataset containing test and evaluation documents, including the gold standard as well as the model's and individual experts' annotations, is publicly available and can be found on GitHub.

6.1. Narrative identification

On a structural level, the model performs remarkably well. In almost all cases, `GPT-4o` precisely follows the formatting instructions and consistently produces valid JSON outputs. In Section B of the appendix, we present three examples comparing the gold standard narratives to those generated by the model, alongside the original excerpt for reference.

The LLM also closely adheres to the CoT-procedure laid out in Figure 1. In Steps 1 and 2, `GPT-4o` typically paraphrases relevant parts of the source text in a semantically faithful and economically coherent manner. The model also correctly reproduces event order and directions of causality in most narratives. The CoT logic likely aids this process, as it requires the model to state the narrative sequence as continuous text twice before imposing causal structure and determining event order in Step 3. Even when the model misorders events, it typically preserves correct causal directionality, making such errors relatively minor and more grammatical than substantive. Misorderings tend to favor the causal connector *causes* over *caused by*. The reason for this preference is speculative. `GPT-4o` may favor stating causes first and effects second because such constructions occur more frequently in its training data.

In Table 1, we compare the model's performance to that of our expert annotators' using basic summary statistics. Notably, the LLM identifies a similar number of narratives per document as the human coders, on average. This suggests that the overall level of performance is not compromised by persistent over- or under-identification of narratives. However, a closer look at the distribution reveals important limitations. While human annotators show substantial variance in how many narratives they identify per document (standard deviation of 2.03, on average) the model's output is remarkably stable, averaging 2.32 narratives per document with a standard deviation of only 1.15. This points to a potential inductive bias: the model appears to have an implicit prior about how many narratives it "expects" to find, leading to systematic over-identification in narrative-sparse

Measure	Expert 1	Expert 2	Expert 3	Gold	Model
Average	2.22	2.36	2.29	2.91	2.32
Standard deviation	1.95	2.12	2.04	2.35	1.15

Table 1: Narrative counts per document for human experts and `GPT-4o`.

texts and under-identification in narrative-dense ones. This pattern is also evident in edge cases. In four documents where the gold standard specifies no narrative, `GPT-4o` nonetheless extracts at least one narrative in each case. This suggests that the model may be prone to overfitting the 7-shot examples provided in the prompt.

To better understand this pattern, it is useful to examine some specific challenges the model faces. The lack of variability in the number of narratives is partly due to the model struggling with causal chains and forks. Even with explicit instructions and a dedicated CoT step in place, the model tends to return most narratives in a non-forked fashion, even in clear-cut cases where all human experts agree. Rather than disentangling forked or chained narratives into multiple atomic phrases, the model frequently compresses them into a single generic statement (see Figure 4). This issue is related to stylistic differences across documents. Some of the articles break down inflation-related phenomena in a technical style, even referencing the channels of monetary policy that are thought to be at work. In such articles, causal chains and forks may intermix, even within a single sentence, yielding multiple narratives from a short text span. When `GPT-4o` struggles with disentangling these structures and coding all narratives separately, the model's narrative count gets notably deflated for the respective article. Conversely, narrative-sparse articles tend to focus less on textbook economic explanations and more on human interest and storytelling. Given the latter, the model tends to infer inaccurate or speculated narratives much more liberally than human coders. This lack of consistency is related to the phenomenon of hallucinations, a term describing the well-known tendency of LLMs to generate plausible but inaccurate answers (Huang et al., 2023).

In general, it appears that the model performs particularly well on short documents with relatively straightforward narratives, while struggling with longer, more complex documents. `GPT-4o` sometimes fails to detect crucial narrative structures, especially when narratives are implicitly stated or span multiple sentences. As the performance metrics in Table 2 illustrate, these characteristics come with substantial downstream effects: `GPT-4o` averages 1.25 major deviations per document overall, more than twice the rate of any expert coder. This translate to an "unexpected major-deviation rate" of 0.91 narratives per document——a metric that captures how many more major deviation the model made per case compared to the human experts. Taken together, these false positives and false neg-

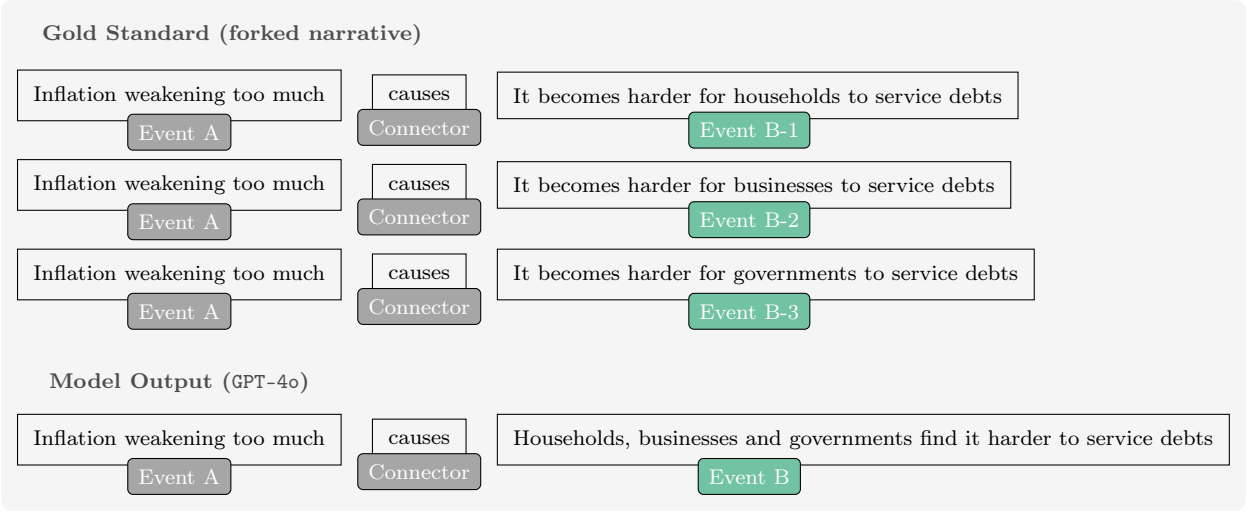

Figure 4: Comparison of a forked narrative (as indicated by the gold standard) and a non-forked narrative (as returned by the LLM).

atives drive down the model's overall accuracy to 44%, compared to the 67–74% accuracy range achieved by individual annotators.

It is important to note that not all errors that we classify as major deviations are created equal. Some of the deviations `GPT4o` produces represent grave misunderstandings and violations of the codebook while others are quite subtle. Occasionally, the economic expertise encoded in the model appears to collide with the aim of faithfully representing the contents of the documents. For example, `GPT-4o` encodes "plans to slow an influx of hard currency that is fueling rapid money-supply growth and pushing up inflation" as a narrative chain, asserting that *money-supply growth* is said to *push up inflation*. While economically plausible, the text clearly asserts the *influx of hard currency* as the cause for inflation. While both variants are related, the model's reading shifts the focal point and chain of causality, resulting in a distorted interpretation of the document.

Measure	**Expert 1**	**Expert 2**	**Expert 3**	**Model**
Major-deviation rate	0.40	0.35	0.49	1.25
Unexpected major deviations	—	—	—	0.91
Accuracy (vs. gold)	0.72	0.74	0.67	0.44
Jaccard similarity	0.59	0.60	0.59	0.40

Table 2: Annotator agreement and deviation metrics for human experts and `GPT-4o`.

In an additional evaluation, we benchmark the narrative extraction capabilities using a Jaccard similarity metric. This merely quantitative approach quantifies lexical overlap between predicted and reference token sets on a scale from 0 (no overlap) to 1 (perfect match). Matching the impressions described earlier, the human experts outperform the model, achieving mean scores across all documents of J = 0.59 or 0.6, whereas the model reaches J = 0.40. This gap again points towards some limitations, despite the measure being rather lenient in nature, as it rewards partial lexical matches and ignores semantic coherence, causal directionality, or narrative plausibility.

6.2. Narrative aggregation

Given the partially promising yet still limited extraction capabilities of `GPT-4o`, any downstream analysis of the extracted narratives must be interpreted with caution. At this stage, results should be seen as indicative rather than conclusive. Nonetheless, as outlined earlier, our goal is to develop a comprehensive narrative extraction framework which also requires us to consider how the extracted data can be processed and aggregated for future applications and econometric analysis. Below, we present the results of this aggregation step, even though its implications remain exploratory at this point.

The abstraction and harmonization of extracted events enable us to compare and group narrative elements across documents. This, in turn, allows for the identification of frequently recurring causal patterns, such as the link between government spending and inflation, or between interest rate hikes and reduced economic growth. Figure 5 illustrates a selection of such simplified narrative arcs that appeared multiple times across our corpus.

One narrative that recurs notably in our sample is the link between government spending and inflation (n=3). This causal association aligns with theoretical arguments suggesting that expansionary fiscal policy—especially if debt-financed—can trigger inflationary pressure, particularly in environments of constrained monetary policy or supply-side rigidity (Sargent, Wallace, et al., 1981). Notably, such narratives are not only central to formal macroeconomic models but also resonate in journalistic discourse (Andre et al., 2024; Schmidt, 2025).

If such narratives appear frequently in the media, they are expected to influence how individuals interpret macroeconomic developments, especially in periods of heightened uncertainty. As discussed in Section 2, narratives are not just post-hoc explanations, but

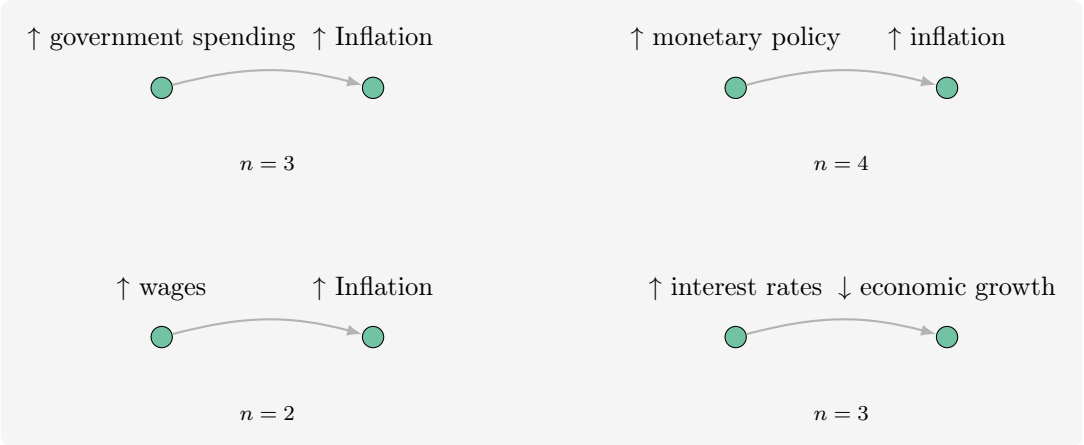

Figure 5: Simplified narratives derived from our LLM based narrative extraction pipeline. "↑" in the context of monetary policy translates to "loose m.p.". n=x indicates how often this narrative can be found in our sample.

sense-making stories that shape how people form expectations and make decisions. This might certainly also apply to the kinds of narratives presented here. For instance, if media audiences are repeatedly exposed to a narrative like "higher wages cause rising inflation" (n=2 in our sample), it is reasonable to assume that such frames may shape their thinking, e.g., influencing how they approach future wage negotiations. Although recent research points in this direction (Andre et al., 2024), it remains unclear how often a reader needs to encounter a given narrative for it to have a measurable effect. With most prominent narrative types in our sample occurring in just 2–4 of 80 articles, the strength and reach of such effects remain speculative and warrant further investigation.

Other frequently recurring narratives include the effect of loose monetary policy on inflation rates (n=4)—which is also a prominent argument in both theory and public discussion—, and interest rate hikes on economic growth (n=3). All four of these narrative arcs point toward classical macroeconomic concerns that have been widely discussed in both academic and policy debates. The most frequently occurring narrative-parts (valence-topic combinations) in our sample are presented in Table 3.

Further analysis of the directional roles played by different events in the extracted narratives reveals additional information on the structure of economic storytelling in the media. For instance, the topic *stock market* appears predominantly as an "effect event" (90% of all instances) suggesting that (business) journalism often frames stock prices as a barometer reacting to other economic developments. In contrast, narratives in which stock market movements are framed as a causal force influencing other domains are relatively rare. Conversely, *government spending* is predominantly discussed as a "cause event" (91% of instances) highlighting its frequent use as an explanatory device in inflation reporting.

Event (example)	Valence	Sym.	Topic	Count
higher inflation	rising	↑	Inflation	58
economic stability in the us	positive	↑	Economy	29
higher interest rates	rising	↑	Interest rates	27
the Fed's aggressiveness	loose	↑	Monetary policy	25
the ECB tightening its m.p.	tight	↓	Monetary policy	23
economy plunges toward hard landing	falling	↓	Economy	22
lower inflation rates	falling	↓	Inflation	21
higher gas prices	rising	↑	Energy prices	19
stocks are getting attractive again	rising	↑	Stock Market	15

Table 3: Most frequent valence-topic combinations in extracted narratives. The presented *Events* (left column) are sampled from all events that translate to the respective valence-topic combination.

This asymmetry mirrors typical theoretical priors in economics: while public spending is often viewed as a policy lever that sets economic processes in motion, the stock market tends to be treated as an outcome variable, shaped by shifts in policy, sentiment, or macroeconomic conditions.

However, the aggregation process also reveals important limitations. Despite careful normalization and abstraction, many extracted narratives remain highly context-specific and cannot be assigned to larger clusters. This is partly due to the large combinatorial space of possible valence-topic pairs: the number of possible combinations of clusters A and B and valences A and B is very high, leading to overall rare instances where different texts lead to the same simplified narrative statement. After decomposition and standardization, our 80 evaluation documents yielded a total of 409 distinct narratives, all made up of two valences and two events. Many of them only appear once in the whole dataset, which explains why we count only two or three instances of our most prominent complete narratives.

This lack of convergence, however, is not considered as a methodological weakness. Instead, it reflects a central feature of narratives as they appear in public discourse, and also in the media. Most newspaper articles tend to convey causal claims that are tailored to specific events, actors, and settings, making generalization inherently difficult. Additionally, our strict definition of narratives as precise, two-event causal claims may contribute to this fragmentation, as it emphasizes nuance and specificity over generalization. It is also important to note that not all causal relationships extracted by the model are economically meaningful. Some narratives form around very specific events, locations

or protagonists, which cannot be translated to a broader economic category and may not even be of notable relevance.

All in all, our results highlight both the potential and the limitations of LLM-based narrative aggregation and clustering. While the described framework is capable of detecting economically plausible causal structures and producing interpretable output, further methodological refinement is needed to move from fragmented extractions toward more robust generalizations.

7. Discussion

Our study provides evidence that general-purpose language models can be meaningfully used to extract economic narratives from text. The results achieved by `GPT-4o` on our complex annotation task are promising in that they show considerable improvement compared to earlier attempts using previous generations of LLMs. Identifying economic narratives from text requires abstract reasoning, domain-specific knowledge, and the ability to resolve ambiguity and co-reference. We show that a general-purpose, instruction-tuned LLM is able to approach this task with some consistency without any task-specific fine-tuning. This illustrates meaningful development in the field of narrative economics. The progress we observe here reflects ongoing advances in language model capabilities, particularly in tasks that require instruction-following and context-sensitive reasoning.

We also see considerable promise in combining LLM-based extraction with scalable methods of abstraction and clustering. Our initial experiments show that harmonizing extracted narratives into macroeconomic categories enables the identification of recurring narrative patterns. Although most extracted narratives remain too specific to be aggregated meaningfully, the potential for scalable aggregation is clear. Recent work by Reccius (2025, forthcoming), for example, introduces a polar embedding-based approach that could be used to measure the distance between narratives along theoretically meaningful dimensions. Integrating such techniques with LLM-based extraction appears promising for both theoretical and empirical work on economic narratives.

Our findings also reflect some critical shortcomings of LLMs. While tracing the exact origins of each LLM-misstep is currently not possible, it is worth remembering that LLMs are pre-trained to produce coherent text and then fine-tuned to follow instructions. There is some evidence that, as a side effect of this training, models occasionally try to oversatisfy prompts, placing helpfulness over accuracy and thus avoiding rejections (Chen et al.,

2025). This tendency results in the model trying to 'squeeze blood from a stone', yielding outputs that are unfaithful to instructions. These drawbacks are especially problematic in research settings and must be addressed.

Importantly, our evaluation approach also highlights the challenges of defining and operationalizing the concept of an economic narrative in practice. Even among expert annotators, we observed substantial variation in annotator opinion about which narratives were identified, how events were segmented, and how causal relations were interpreted. This variability stresses the subjective nature of the sense-making process that occurs when people consume narratives through media. The extraction of economic narratives from text is inherently interpretive, even when formal definitions and structured codebooks are applied. Consequently, fully automating the process remains an ambitious goal.

Looking ahead, we envision a future in which economic narrative research will increasingly be shaped by hybrid workflows—pairing the scalability of language models with the interpretive capacity of human experts. Rather than replacing researchers, LLMs will act as amplifiers of their judgment. Human coders will remain central in defining what constitutes a narrative in the first place, developing precise annotation frameworks, and crafting high-quality prompts that translate conceptual definitions into machine-readable tasks. One key component of this collaboration will be the creation of exemplary few-shot examples, which LLMs rely on for task adaptation.

Accordingly, the time saved by automating the annotation of thousands of documents will not come without cost. It will need to be reinvested in the evaluation process itself. Unlike more standardized tasks such as sentiment classification, narrative extraction requires a deep understanding of semantics, context, and domain-specific framing to evaluate if a models' results are valid. That is, quick eyeballing or resorting to simple accuracy scores will not suffice. Instead, each output must be assessed on its interpretive merit, demanding a level of scrutiny that (for now) only human coders can provide. In that sense, our study underscores the continued importance of human involvement in content analysis and social science research (Haim et al., 2023). LLMs may extend what is possible; but only when embedded in workflows that preserve interpretive oversight and methodological rigor.

To conclude, our results suggest that while narrative extraction via LLMs is still in its early stages, it offers potential for scalable, interpretable, and theoretically informed analysis of economic discourse. With further refinement, including improved prompts, fine-tuned models, and downstream validation, the approach has the potential to contribute meaningfully to the emerging field of narrative economics.

7.1. Outlook

While LLMs will further develop and excel even more at language understanding in the future, we believe that the performance of the current generation of models can be improved on with more specialized training, such as (parameter-efficient) fine-tuning. As best practices for prompting evolve alongside model architectures, we see a need to continuously refine both our extraction and aggregation techniques.

The aggregation framework presented in this paper reflects only an initial stage. Future work should explore more advanced methods of narrative clustering and abstraction. A promising direction is offered by Reccius (2025, forthcoming), who proposes a polar embedding-based approach to improve the alignment of semantically similar narrative elements. Incorporating such techniques may allow us to systematically group related narratives across documents, even when surface-level variation is high. Once greater reliability in extraction and aggregation is achieved, a natural next step is to scale the method to larger corpora. This would enable a more systematic assessment of narrative prevalence and diffusion over time and across domains.

However, while the performance of `GPT-4o` on our task showed that Large Language Models have great potential for narrative extraction, using them requires good hardware or financial backing. In addition to that, the environmental impact of running such models should not be understated. We therefore also aim to develop methods that cleverly combine scalable NLP methods with LLMs, enabling a thorough analysis of large corpora with a minimal amount of resources.

Ultimately, this paper focused on establishing the feasibility of LLM-based narrative extraction. Future research should move toward substantive applications. In particular, linking extracted narratives to external economic indicators—such as inflation expectations, consumer sentiment, or financial market volatility—could offer novel insights into the role of narrative framing in macroeconomic dynamics. Such connections would not only extend the methodological contribution of this paper, but also advance our theoretical understanding of how economic stories shape the economy itself.

8. Conclusion

Narratives play an crucial role in everyday economic decision-making. All actors in an economy, regardless of their level of sophistication or macroeconomic importance, are influenced by the economic narratives circulating in their environment. Extracting such narratives from mass media enables us to trace their origins, understand how they spread through society, and examine their development over time. For economists, this is a crucial yet complex task.

To enable a quantitative analysis of economic narratives, we present an annotation codebook as a stepping stone into extracting economic narratives from a corpus of texts. We utilize this codebook to let three expert annotators code narratives from a total of 100 documents and form a gold-standard data set out of these coded narratives. We then prompt `GPT-4o` to perform the same task and evaluate it on our gold-standard annotations.

We categorize deviations from the gold-standard annotations into minor and major deviations, where minor deviations cover formal or small subjective deviations and major deviations cover cases in which a narrative has been coded incorrectly. We also compare the resulting deviations to the ones that resulted from our expert annotators.

Our results indicate that the level of language comprehension and economic expertise encoded in contemporary Large Language Models (LLMs) enables them to extract economic narratives from newspaper articles effectively. We further find that few-shot Chain-of-Thought prompting—a state-of-the-art AI strategy—is well-suited for LLM-based text retrieval tasks in economics, as it combines high performance with strong transparency in complex annotation tasks.

Our findings also underscore some limitations. Most notably, `GPT-4o` exhibits biases in narrative density, at least when prompted as presented, under-identifying narratives in some contexts while over-identifying them in others. The model struggles with structurally complex narratives, particularly those involving nuanced causal structures like forks or chains. It also shows limited adaptability when encountering challenging edge cases or hidden and implicit causal links between events that may span multiple sentences.

These shortcomings suggest that, while LLMs possess strong general capabilities, economic narrative extraction is a specialized task where expert judgment cannot (and should not) be fully replaced by general-purpose LLMs. Achieving human-level reliability will require more capable foundational models and methodological refinements.

Acknowledgments

This study is part of a project of the Dortmund Center for data-based Media Analysis (DoCMA) at TU Dortmund University and the Narrative Economic Alliance Ruhr (NEAR) project, supported by the Mercator Research Center Ruhr (MERCUR) with project number Ko-2022-0015. It was also partially funded by the Reality Check incubator project at the Research Center for Trustworthy Data Science and Security.

References

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart (2024). *Narratives about the Macroeconomy*. SAFE Working Paper 426. DOI: 10.2139/ssrn.4947636.

Ash, Elliott, Germain Gauthier, and Philine Widmer (2024). "Relatio: Text Semantics Capture Political and Economic Narratives". In: *Political Analysis* 32.1, pp. 115–132. DOI: 10.1017/pan.2023.8.

Bachmann, Rüdiger, Tim O Berg, and Eric R Sims (2015). "Inflation expectations and readiness to spend: Cross-sectional evidence". In: *American Economic Journal: Economic Policy* 7.1, pp. 1–35. DOI: 10.1257/pol.20130292.

Bénabou, Roland, Armin Falk, and Jean Tirole (2018). *Narratives, Imperatives, and Moral Reasoning*. NBER Working Paper 24798. DOI: 10.3386/w24798.

Benner, Niklas, Kai-Robin Lange, and Carsten Jentsch (2022). *Named Entity Narratives*. Ruhr Economic Papers 962. DOI: 10.4419/96973126.

Blinder, Alan S, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen (2008). "Central bank communication and monetary policy: A survey of theory and evidence". In: *Journal of economic literature* 46.4, pp. 910–45. DOI: 10.1257/jel.46.4.910.

Bornheim, Tobias, Niklas Grieger, Patrick Gustav Blaneck, and Stephan Bialonski (2024). "Speaker Attribution in German Parliamentary Debates with QLoRA-adapted Large Language Models". In: *Journal for Language Technology and Computational Linguistics* 37.1, pp. 1–13. DOI: 10.21248/jlcl.37.2024.244.

Bracha, Anat and Jenny Tang (2025). "Inflation levels and (in) attention". In: *Review of Economic Studies* 92.3, pp. 1564–1594. DOI: 10.1093/restud/rdae063.

Brown, Tom B. et al. (2020). "Language models are few-shot learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Curran Associates Inc. ISBN: 9781713829546. DOI: 10.5555/3495724.3495883.

Burch, Susan W and Diane Werneke (1975). "The stock of consumer durables, inflation, and personal saving decisions". In: *The Review of Economics and Statistics*, pp. 141–154.

Burla, Laila, Birte Knierim, Jurgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel (2008). "From text to codings: intercoder reliability assessment in qualitative content analysis". In: *Nursing research* 57.2, pp. 113–117. DOI: 10.1097/01.NNR.0000313482.33917.7d.

Chen, Shan, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle S Bitterman (2025). "When Helpfulness Backfires: LLMs and the Risk of Misinformation Due to Sycophantic Behavior". In: *Research Square*, rs–3.

Coibion, Olivier and Yuriy Gorodnichenko (2015). "Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation". In: *American Economic Journal: Macroeconomics* 7.1, pp. 197–232. DOI: 10.1257/mac.20130306.

Conrad, Christian, Zeno Enders, and Alexander Glas (2022). "The role of information and experience for households' inflation expectations". In: *European Economic Review* 143, p. 104015. DOI: 10.1016/j.euroecorev.2021.104015.

Crow, Deserai and Michael Jones (2018). "Narratives as tools for influencing policy change". In: *Policy & Politics* 46.2, pp. 217–234. DOI: 10.1332/030557318X15230061022899.

DeepSeek-AI (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dubey, Abhimanyu et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783v2.

Eliaz, Kfir and Ran Spiegler (2020). "A model of competing narratives". In: *American Economic Review* 110.12, pp. 3786–3816. DOI: 10.1257/aer.20191099.

Eliaz, Kfir and Ran Spiegler (2024). "News media as suppliers of narratives (and information)". In: arXiv: 2403.09155 [econ.TH].

Fan, Wenqi, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li (2024). *A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models*. arXiv: 2405.06211 [cs.CL].

Fiore, Fiorella De, Alexis Maurin, Andrej Mijakovic, and Damiano Sandri (2025). *Monetary policy in the news: The FOMC's media coverage and inflation expectations*. URL: https://cepr.org/voxeu/columns/monetary-policy-news-fomcs-media-coverage-and-inflation-expectations (visited on 06/02/2025).

Flynn, Joel P. and Karthik A. Sastry (2024). *The Macroeconomics of Narratives*. NBER Working Paper 32602. National Bureau of Economic Research. DOI: 10.3386/w32602.

Gehring, Kai and Matteo Grigoletto (2023). *Analyzing Climate Change Policy Narratives with the Character-Role Narrative Framework*. CESifo Working Paper 10429. DOI: 10.2139/ssrn.4456361.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). "ChatGPT outperforms crowd workers for text-annotation tasks". In: *Proceedings of the National Academy of Sciences* 120.30. ISSN: 1091-6490. DOI: 10.1073/pnas.2305016120.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (Feb. 2023). "The Voice of Monetary Policy". In: *American Economic Review* 113.2, pp. 548–584. DOI: 10.1257/aer.20220129.

Gueta, Almog, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart (2025). "Can LLMs Learn Macroeconomic Narratives from Social Media?" In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, pp. 57–78. ISBN: 979-8-89176-195-7. URL: https://aclanthology.org/2025.findings-naacl.4/.

Gurkaynak, Refet S., Brian P. Sack, and Eric T. Swanson (2005). "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements". In: *International Journal of Central Banking* 1, pp. 55–93. DOI: 10.2139/ssrn.633281.

Haim, Mario, Valerie Hase, Johanna Schindler, Marko Bachl, and Emese Domahidi (2023). "(Re) Establishing quality criteria for content analysis: A critical perspective on the field's core method". In: *Studies in Communication and Media (SCM)* 12, pp. 277–288.

Han, Di, Wei Guo, Han Chen, Bocheng Wang, and Zikun Guo (2024). "LEST: Large language models and spatio-temporal data analysis for enhanced Sino-US exchange rate forecasting". In: *International Review of Economics & Finance* 96, p. 103508.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2017). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach". In: *The Quarterly Journal of Economics* 133.2, pp. 801–870. DOI: 10.1093/qje/qjx045.

Hansen, Stephen, Michael McMahon, and Matthew Tong (2019). "The long-run information effect of central bank communication". In: *Journal of Monetary Economics* 108, pp. 185–202. DOI: 10.1016/j.jmoneco.2019.09.002.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu (Nov. 9, 2023). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43.2, pp. 1–55. ISSN: 1558-2868. DOI: 10.1145/3703155. arXiv: 2311.05232 [cs.CL].

Jeong, Minhyuk and Kwangwon Ahn (2025). "Energy organization sentiment and oil return forecast". In: *Energy Economics* 141, p. 108105.

Jones, Michael D. and Mark K. McBeth (2010). "A Narrative Policy Framework: Clear Enough to Be Wrong?" In: *Policy Studies Journal* 38.2, pp. 329–353. DOI: 10.1111/j.1541-0072.2010.00364.x.

Juster, F Thomas, Paul Wachtel, Saul Hymans, and James Duesenberry (1972). "Inflation and the Consumer". In: *Brookings Papers on Economic Activity* 1972.1, pp. 71–121. DOI: 10.2307/2534115.

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2020). "Making text count: economic forecasting using newspaper text". In: *Journal of Applied Econometrics*.

King, Mervyn and John Kay (2020). *Radical uncertainty: Decision-making for an unknowable future*. Hachette UK.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large language models are zero-shot reasoners". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc. ISBN: 9781713871088. DOI: 10.5555/3600270.3601883.

Korinek, Anton (2023). "Generative AI for Economic Research: Use Cases and Implications for Economists". In: *Journal of Economic Literature* 61.4, pp. 1281–1317. ISSN: 0022-0515. DOI: 10.1257/jel.20231736.

Lamla, Michael J and Sarah M Lein (2014). "The role of media for consumers' inflation expectation formation". In: *Journal of Economic Behavior & Organization* 106, pp. 62–77.

Lange, Kai-Robin, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael Roos, and Carsten Jentsch (2022a). *Towards Extracting Collective Economic Narratives from Texts*. Ruhr Economic Papers 963. DOI: 10.4419/96973127.

Lange, Kai-Robin, Jonas Rieger, Niklas Benner, and Carsten Jentsch (2022b). "Zeitenwenden: Detecting changes in the German political discourse". In: *Proceedings of the 2nd Workshop on Computational Linguistics for the Political and Social Sciences*, pp. 47–53. URL: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022/254133848-1662996909/cpss-2022-proceedings.pdf.

Lange, Kai-Robin, Jonas Rieger, and Carsten Jentsch (2024). "Lex2Sent: A bagging approach to unsupervised sentiment analysis". In: *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*. Association for Computational Linguistics, pp. 281–291. URL: https://aclanthology.org/2024.konvens-main.28/.

Lange, Kai-Robin, Tobias Schmidt, Matthias Reccius, Henrik Müller, Michael Roos, and Carsten Jentsch (2025). "Narrative Shift detection: A hybrid approach of Dynamic Topic Models and Large Language Models". In: *Proceedings of the Text2Story'25 Workshop*. URL: https://www.di.ubi.pt/~jpaulo/Text2Story2025/paper6.pdf.

Larsen, Vegard H and Leif Anders Thorsrud (2019). *Business cycle narratives*. CESifo Working Paper 7468. DOI: 10.2139/ssrn.3338822.

OpenAI (2024). URL: https://openai.com/index/learning-to-reason-with-llms/ (visited on 05/07/2025).

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). "Lost in the Middle: How Language Models Use Long Contexts". In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173. DOI: 10.1162/tacl_a_00638. URL: https://aclanthology.org/2024.tacl-1.9/.

Mathew, Binny, Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier (2020). "The POLAR Framework: Polar Opposites Enable Interpretability of Pre-Trained Word Embeddings". In: *Proceedings of The Web Conference 2020*. WWW '20. Taipei, Taiwan:

Association for Computing Machinery, pp. 1548–1558. ISBN: 9781450370233. DOI: 10.1145/3366423.3380227.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman (2024). "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale". In: *Research & Politics* 11.1. DOI: 10.1177/20531680241231468.

Mondorf, Philipp and Barbara Plank (2024). "Comparing Inferential Strategies of Humans and Large Language Models in Deductive Reasoning". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 9370–9402. DOI: 10.18653/v1/2024.acl-long.508.

Neidhardt, Friedhelm (2010). "Selbststeuerung der Wissenschaft: Peer Review". In: *Handbuch Wissenschaftspolitik*, pp. 280–292. DOI: 10.1007/978-3-531-91993-5_19.

Nyman, Rickard, Sujit Kapadia, and David Tuckett (2021). "News and narratives in financial systems: exploiting big data for systemic risk assessment". In: *Journal of Economic Dynamics and Control* 127, p. 104119. DOI: 10.1016/j.jedc.2021.104119.

OpenAI et al. (2024a). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

OpenAI et al. (2024b). *GPT-4o System Card*. arXiv: 2410.21276 [cs.CL].

Pearl, Judea (2009). *Causality*. Cambridge University Press. ISBN: 9781139632997.

Reis, Ricardo (2006). "Inattentive producers". In: *The Review of Economic Studies* 73.3, pp. 793–821. DOI: 10.1111/j.1467-937X.2006.00396.x.

Rieger, Jonas, Carsten Jentsch, and Jörg Rahnenführer (2021). "RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, pp. 2337–2347. DOI: 10.18653/v1/2021.findings-emnlp.201.

Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch (2022). "Dynamic change detection in topics based on rolling LDAs". In: *Proceedings of the Text2Story'22 Workshop*, pp. 5–13. URL: http://ceur-ws.org/Vol-3117/paper1.pdf.

Roos, Michael and Matthias Reccius (2024). "Narratives in Economics". In: *Journal of Economic Surveys* 38.2, pp. 303–341. DOI: 10.1111/joes.12576.

Saldaña, Johnny (2016). *The coding manual for qualitative researchers.* 3E. SAGE. ISBN: 9781473902480.

Sargent, Thomas J, Neil Wallace, et al. (1981). "Some unpleasant monetarist arithmetic". In: *Federal reserve bank of minneapolis quarterly review* 5.3, pp. 1–17.

Schlaufer, Caroline, Johanna Kuenzler, Michael D Jones, and Elizabeth A Shanahan (2022). "The narrative policy framework: a traveler's guide to policy stories". In: *Politische Vierteljahresschrift* 63.2, pp. 249–273. DOI: 10.1007/s11615-022-00379-6.

Schmidt, Tobias (2025). *Narrating inflation: How German economic journalists explain post-covid price rises*. DoCMA Working Paper 14. DOI: 10.17877/DE290R-25380.

Schmidt, Torsten, Henrik Müller, Jonas Rieger, Tobias Schmidt, and Carsten Jentsch (2023). *Inflation perception and the formation of inflation expectations*. Ruhr Economic Papers 1025. DOI: 10.4419/96973191.

Shahriar, Sakib, Brady Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool (2024). *Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency*. arXiv: 2407.09519v1.

Shanahan, Elizabeth A, Michael D Jones, Mark K McBeth, and Claudio M Radaelli (2018). "The narrative policy framework". In: *Theories of the policy process*. Routledge, pp. 173–213. DOI: 10.4324/9780429494284.

Shiller, Robert J. (2017). "Narrative Economics". In: *American Economic Review* 107.4, pp. 967–1004. ISSN: 0002-8282. DOI: 10.1257/aer.107.4.967.

Sims, Christopher A (2003). "Implications of rational inattention". In: *Journal of monetary Economics* 50.3, pp. 665–690. DOI: 10.1016/S0304-3932(03)00029-1.

Song, Yisheng, Ting Wang, Puyu Cai, Subrota K. Mondal, and Jyoti Prakash Sahoo (July 2023). "A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities". In: *ACM Computing Surveys* 55.13s, pp. 1–40. ISSN: 1557-7341. DOI: 10.1145/3582688.

Ter Ellen, Saskia, Vegard H Larsen, and Leif Anders Thorsrud (2022). "Narrative monetary policy surprises and the media". In: *Journal of Money, Credit and Banking* 54.5, pp. 1525–1549.

Tian, Runchu, Yanghao Li, Yuepeng Fu, Siyang Deng, Qinyu Luo, Cheng Qian, Shuo Wang, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Huadong Wang, and Xiaojiang Liu (2024). *Distance between Relevant Information Pieces Causes Bias in Long-Context LLMs*. arXiv: 2410.14641 [cs.CL].

Tuckett, David, Douglas Holmes, Alice Pearson, and Graeme Chaplin (2020). *Monetary policy and the management of uncertainty: a narrative approach*. Bank of England working papers 870. Bank of England. URL: https://ideas.repec.org/p/boe/boeewp/0870.html.

Wang, Guoqing, Zeyu Sun, Zhihao Gong, Sixiang Ye, Yizhou Chen, Yifan Zhao, Qingyuan Liang, and Dan Hao (2024). "Do advanced language models eliminate the need for prompt engineering in software engineering?" In: arXiv: 2411.02093 [cs.SE].

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (Sept. 3, 2021). *Finetuned Language Models Are Zero-Shot Learners*. Tech. rep. arXiv: 2109.01652 [cs.CL].

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). "Chain-of-thought prompting elicits rea-

soning in large language models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. Curran Associates Inc. ISBN: 9781713871088. DOI: 10.5555/3600270.3602070.

Werning, Iván (2022). *Expectations and the Rate of Inflation*. NBER Working Paper 30260. DOI: 10.3386/w30260.

Yang, Chenyang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kästner, and Tongshuang Wu (2025). *What Prompts Don't Say: Understanding and Managing Underspecification in LLM Prompts*. Tech. rep. arXiv: 2505.13360 [cs.CL]. URL: https://arxiv.org/abs/2505.13360.

Zhang, Haopeng, Philip S. Yu, and Jiawei Zhang (2025). "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models". In: *ACM Comput. Surv.* Just Accepted. ISSN: 0360-0300. DOI: 10.1145/3731445.

Zhang, Hugh, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan (2021). "Trading Off Diversity and Quality in Natural Language Generation". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Ed. by Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina. Online: Association for Computational Linguistics, pp. 25–33. URL: https://aclanthology.org/2021.humeval-1.3/.

Zhong, Meizhi, Chen Zhang, Yikun Lei, Xikai Liu, Yan Gao, Yao Hu, Kehai Chen, and Min Zhang (2025). "Understanding the RoPE Extensions of Long-Context LLMs: An Attention Perspective". In: *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 8955–8962. URL: https://aclanthology.org/2025.coling-main.600.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang (2024). "Can Large Language Models Transform Computational Social Science?" In: *Computational Linguistics* 50.1, pp. 237–291. DOI: 10.1162/coli_a_00502.

Appendix

A. Codebook

1. Basic Idea

```
The goal of this coding task is to extract all narratives from the provided
texts as accurately and comprehensively as possible.  The texts are excerpts
from newspaper articles published in the NYT or the Wall Street Journal.[a]
```

We define a narrative as a causal connection between two consecutive events.

Simple examples of a narrative:

- Russia's war on Ukraine - causes - People leaving Ukraine
- rising prices - causes - lifting of global bond yields
- rapid money-supply growth - is caused by - influx of hard currency

These narratives are sometimes explicitly mentioned in the text and easy to identify. Often, however, they require some interpretation and abstraction. That is why it is crucial for our research that multiple coders read and annotate the same texts. This ensures that results are not biased by individual interpretations and that the extracted narratives are recognizable by other coders as well.

2. Target Format

We aim to train a language model to reduce all narratives in a text to one of two possible formats:

- Format A: event 1 - causes - event 2
- Format B: event 1 - is caused by - event 2

Each narrative must follow one of these formats. The distinction between A and B preserves the order of events as they appear in the original text. The event mentioned first must be event 1 in your coded narrative.

Examples:

- *Source text:* The Titanic struck an iceberg on 14 April 1912. As a result, the Titanic sank.
  the Titanic struck an iceberg - causes - the Titanic sank
- *Source text:* People have a hard time finding jobs because the Fed keeps interest rates high.
  people have a hard time finding jobs - is caused by - the Fed keeps interest rates high

These examples show: The narrative does not have to be a coherent sentence. The task is to reduce running text into a "bullet point" format that strips away all information not part of the causal chain.

3. Events

3.1 Types of Events

Events can be categorized into events and activities as well as states and circumstances.

- ‘‘the collapse of the Berlin Wall in 1989’’ (event)
- ‘‘Mexico grew rapidly’’ (activity)
- ‘‘Trump signed the bill into law’’ (activity)

- ‘‘inflation’’ (state)
- ‘‘lack of competitive innovation’’ (circumstance)

States and circumstances can also include properties of people or other entities:

- ‘‘Stubborn Indian inflation’’
- ‘‘The prime minister’s arrogance’’

Events and activities can also include future events or plans.

- ‘‘The ECB will probably raise rates by next year’’ (plan)
- ‘‘There might be more earthquakes in Germany in the future’’ (future event)

Policy measures also count as events:

- ‘‘The Schuldenbremse’’
- ‘‘The Affordable Care Act’’

3.2 Unchanged Coding of Events

Events should generally be coded in the exact form used in the source text. The closer the extracted narrative is to the source text, the faster the model will learn.

Specifically:

- Do not use synonyms for terms used in the text
- Keep the same verb tense (present stays present, past stays past, etc.)
- Do not abstract semantically or economically (e.g., do not interpret ‘‘rising prices’’ as ‘‘inflation’’)
- Exception: Explanatory clauses (e.g., appositions) can usually be left out (see 3.5)

3.3 Coreference Resolution

An exception to 3.2 applies when entities (e.g., people, countries) are only indirectly mentioned.

Example:
‘‘She talked a lot about North Korea. ’The country is really poor, therefore many citizens suffer from hunger’, she said.’’
-> Instead of coding ‘‘The country’’, include the resolved entity:

- {The country|North Korea} is really poor - causes - citizens suffer from hunger

Similarly:
‘‘President Biden discussed the matter today. He downplayed the probability of a government shutdown which has calmed the markets.’’

- {He|President Biden} downplayed the probability of a government shutdown - causes - the markets were calmed

3.4 Opinions

Sometimes a causal link is part of a third-party opinion or assessment. In most cases, the narrative refers to the content of the opinion, not the act of expressing it.

Example:
"Analysts believe that bond yields might rise even further which would put even more strain on pension funds."

- bond yields might rise even further - causes - more strain on pension funds

But in some cases, the act of expression itself is causal:

"ECB president Lagarde made it clear that no further rate raises were on the horizon. Markets responded calmly."

- ECB president Lagarde made it clear that no further rate raises were on the horizon - causes - Markets responded calmly

3.5 Parentheses and Side Clauses

Explanatory side information set off by commas, dashes, or parentheses should be left out unless essential for understanding the narrative.

Example:
"The highest inflation in decades and the war in Ukraine - the first major military conflict in Europe in over 30 years - raised concerns..."

- highest inflation in decades - causes - concerns about a possible recession
- war in Ukraine - causes - concerns about a possible recession

3.6 Chained Events

Some statements include multiple events connected by "and" or "as well as". These should be split into separate narratives.

Example:
"She wants you to forget that federal spending contributed to soaring prices, as well as to the labor shortages across the economy."

- federal spending - causes - soaring prices
- federal spending - causes - labor shortages across the economy

4. Only Positive Causal Links

Please code only positive causal relationships. Ignore negative causality.

Example:
Putin invading Ukraine - does not cause - President Zelensky fleeing Ukraine *(do not code)*

5. Multiple Narratives per Text

A document may contain multiple narratives. Please write each narrative on a new line. Do not use enumerations.

Correct:
- the prices for cucumbers rose by 100% this month - causes - people stop buying cucumbers
- huge money supply - causes - prices for real estate go up

Wrong:
1. the prices for cucumbers rose ... | 2. huge money supply ...

6. Thematic Scope

6.1 Explicit Economic Reference

Please only code narratives that have an explicit economic relevance. You can ignore all other narratives. If you are unsure, write down the narrative and add a note (uncertain).

Examples:

- Putin invading Ukraine - causes - rising energy prices *(code)*
- Putin invading Ukraine - causes - enormous military support by the United States *(do not code)*

6.2 Focus on Inflation Narratives

We are primarily interested in inflation narratives. Therefore, please mark all narratives without explicit reference to inflation or price changes with an "(x)". As a reminder: narratives that have no economic relevance at all should not be coded in the first place.

Examples:

- Putin invading Ukraine - causes - disruption of agricultural supply chains (x) *(code)*
- Putin invading Ukraine - causes - gas price increases in Germany *(code)*
- Putin invading Ukraine - causes - enormous military support by the United States *(do not code)*

The example from 3.6 should therefore be:

- federal spending - causes - soaring prices
- federal spending - causes - labor shortages across the economy (x)

7. Preserve Content

This point is similar to point 3.2. Sometimes, there is a strong temptation to omit parts of a sentence in order to make the narrative more pointed. In some

cases, that's acceptable; however, the rule should be to include all parts of the sentence that are necessary to preserve the core message of the narrative. Please preserve all essential content in the narrative.

Example 1:

- wage inflation going up - causes - the risk of yields on longer-dated bonds rising at a faster pace compared to those on shorter-dated notes (correct)
- wage inflation going up - causes - longer-dated bonds rising faster than shorter-dated notes (incorrect)

Example 2:

- The highest inflation in decades - causes - concerns about a possible recession (correct)
- inflation - causes - recession (incorrect)

8. Edge Cases

If you are unsure whether something really constitutes a narrative or not, code it anyway to be on the safe side. You can later compare all such cases with the other coders or with us.

Example 1:
"Semi"-causal cues (such as "as", "suggesting", ...) can be interpreted as either temporal or causal.

John Sicher, editor and publisher of Beverage Digest, said that soft-drink sales have shown some improvement this year, estimating that they rose about 3.5% in January as price increases stabilized.

In such cases code the narrative just to be safe, and compare your results with other coders afterward.

Example 2:
If you interpret the "as" in the example above as a causal cue, a second ambiguity arises: you could shorten the passage in two ways:

John Sicher, editor and publisher of Beverage Digest, said that soft-drink sales have shown some improvement this year, estimating that they rose about 3.5% in January as price increases stabilized.

- price increases stabilized - causes - soft-drink sales showing some improvement (x)
- price increases stabilized - causes - soft-drink sales rose about 3.5% in January (x)

Again, if you're unsure which version to choose, write down both to be on the safe side.

```
If unsure, include both options.
```

[a]Since the texts are excerpts, some passages may seem out of context. In most cases, however, the meaning should still be clear.

B. Annotated Examples

2004-05-27 | WSJ | China Targets Auto Loans, Speculators

At the same time, overinvestment in the industry has led to rising car inventories. GM says its own inventory levels are healthy and for some models it has waiting lists. Separately, China's foreign-exchange regulator announced plans to slow an influx of hard currency that is fueling rapid money-supply growth and pushing up inflation. Speculators are betting that Beijing will revalue its currency, the yuan, and are parking money in Chinese banks to take advantage of interest rates higher than those in the U.S.The new rules also aim to reduce a run-up in foreign debt, as some borrowers switch to foreign-currency loans that are cheaper than yuan-denominated loans. As part of the tighter measures to come into force on July 1, foreign-invested companies must state a specific use for any funds greater than $200,000 entering China, said a statement posted on the State Administration of Foreign Exchange's Web site.

Narratives

- overinvestment in the industry - causes - rising car inventories (x)
- an influx of hard currency - causes - rapid money-supply growth
- an influx of hard currency - causes - pushing up inflation
- run-up in foreign debt - is caused by - some borrowers switch to foreign-currency loans that are cheaper than yuan-denominated loans
- The new rules - causes - reducing a run-up in foreign debt. (x)

Model Output

- overinvestment in the car industry - causes - rising car inventories
- an influx of hard currency - causes - rapid money-supply growth
- rapid money-supply growth - causes - pushing up inflation
- Speculators are betting that Beijing will revalue its currency, the yuan - causes - speculators parking money in Chinese banks to take advantage of interest rates higher than those in the U.S.
- foreign-currency loans are cheaper than yuan-denominated loans - causes - some borrowers switch to foreign-currency loans

2023-03-03 | WSJ | Business News: Bud Brewer's Volume Fell as Prices Jumped

And people are eating and drinking more at home, a setting that tends to favor beer over other drinks, he said:̈ Beer remains resilient,M̈r. Doukeris said. Ï'll never say that beer is immune to the inflation and everything that is happening out there, but [it] remains a very resilient category:̈ The brewer has for years seen its mainstream brands Bud and Bud Light decline, partly amid a broader shift by American drinkers from beer toward wine and spirits. In response, it has

worked to increase its exposure to premium offerings such as Michelob Ultra and Stella Artois.

Narratives

- people are eating and drinking more at home, a setting that tends to favor beer over other drinks - causes - Beer remains resilient (x)
- Bud and Bud Light decline - is caused by - a broader shift by American drinkers from beer toward wine and spirits (x)
- Bud and Bud Light decline - causes - {it|the brewer} has worked to increase its exposure to premium offerings such as Michelob Ultra and Stella Artois (x)

Model Output

- people are eating and drinking more at home - causes - beer being favored over other drinks
- a broader shift by American drinkers from beer toward wine and spirits - causes - Bud and Bud Light decline
- Bud and Bud Light decline - causes - the brewer worked to increase its exposure to premium offerings such as Michelob Ultra and Stella Artois

2023-03-03 | NYT | Market Gains After 3 Days of Losses

Shares of Salesforce soared 11.5 percent for one of the market's biggest gains after it topped forecasts for profit and revenue last quarter. It also gave a stronger-than-expected forecast for upcoming results. Expectations have been coming down recently for profits at big U.S. companies given still-high inflation and interest rates. But several joined Salesforce in rising Thursday after posting encouraging results. Macy's rose 11.1 percent after reporting stronger profit and revenue for the holidays than analysts expected.

Narratives

- Shares of Salesforce soared 11.5 percent for one of the market's biggest gains - is caused by - {it|Salesforce} topped forecasts for profit and revenue last quarter. (x)
- Shares of Salesforce soared 11.5 percent for one of the market's biggest gains - is caused by - {it|Salesforce} gave a stronger-than-expected forecast for upcoming results (x)
- {it|Salesforce} gave a stronger-than-expected forecast for upcoming results (x) - is caused by - {it|Salesforce} topped forecasts for profit and revenue last quarter. (x)
- Expectations have been coming down recently for profits at big U.S. companies - is caused by - still-high inflation
- Expectations have been coming down recently for profits at big U.S. companies - is caused by - still-high interest rates

- {Several|U.S. companies} joined Salesforce in rising Thursday - is caused by - {U.S. companies} posting encouraging results (x)
- Macy's rose 11.1 percent - is caused by - reporting stronger profit and revenue for the holidays than analysts expected (x)

Model Output

- Shares of Salesforce soared 11.5 percent for one of the market's biggest gains - is caused by - Salesforce topped forecasts for profit and revenue last quarter
- Salesforce gave a stronger-than-expected forecast for upcoming results - causes - Shares of Salesforce soared 11.5 percent for one of the market's biggest gains
- Macy's rose 11.1 percent - is caused by - Macy's reported stronger profit and revenue for the holidays than analysts expected

C. Prompt

```
# Codebook:

## Basic Idea:
I will provide you with an excerpt from newspaper articles that have appeared in
the New York Times or The Wall Street Journal.
Your task is to extract all economic narratives that occur in the excerpt.  This
task is a well-defined multi-step extraction process.  The term "narrative"
is frequently used with different meanings and in a variety of contexts.  The
multi-step extraction process is designed to make sure that your work respects
a very specific definition of economic narratives.  For each step that outlined
below, it is imperative that you only make the changes that are indicated.  I will
provide you with this definition next.

## Definition:
An economic narrative consists of exactly two events and a causal connection that
is asserted between those events.

## Event structures:
In the terminology of causal inference, the definition corresponds to finding the
"direct causal effect" of Event A on Event B. The causal pathways you encounter
in the excerpts may be more complex.  Therefore, in addition to direct causal
effects, find and deconstruct the following causal structures as follows:
1) Fork:  If Event A is said to be a common cause of Events B and C, code both
pathways as separate narratives.
2) Chain:  If Event B is said to be a mediator between Event A and Event C, also
code both pathways as separate narratives.
```

```
## Rules:
You also have to follow a couple of hard-and-fast rules:
1) Retain the order of the two events from the source text at all times.  The
event that appears first in the text must be coded as "Event A", the event that
appears second must be coded as "Event B".
2) When you state the narrative in its "Target Form", always represent the causal
connection by using one of the following phrasings:  i) "causes":  Use this causal
connector when Event A is the cause and Event B is an effect of Event A. ii) "is
caused by":  Use this causal connector when Event B is the cause and Event A is an
effect of Event B.

## Target Forms:
{
"Event A":  "[...]",
"causal connector":  "causes",
"Event B":  "[...]"
}
{
"Event A":  "[...]",
"causal connector":  "caused by",
"Event B":  "[...]"
}

## Extraction Process:
1) "Focused Excerpt":  The excerpts vary in length and in their narrative
density.  Parts of each excerpt may obviously not contain any narrative.  To
focus on relevant parts of the excerpt, start by repeating it, but leaving out the
sentences that you are sure no narrative appears in.
2) "Sequence of Interest":  Go through the "Focused Excerpt" and state the first
narrative sequence you find, that is a sequence that contains two events and what
you consider to be a causal connection between them.
3) "Causal Restatement":  Based on the current "Sequence of Interest", restate the
narrative in the appropriate target form.  Make sure to repeat the events verbatim
without rephrasing.
4) "Coreference Resolution":  If necessary, rephrase one or both events to clearly
identify the entities that occur in the narrative.  For the most part, this will
involve replacing personal pronouns with the entity itself.
5) "Event Rephrasing":  If necessary, rephrase one or both events again so that
the narrative uses correct language.
After every narrative, continue by circling back to 1) and restate the "Focused
Excerpt".  Do so even if you did not previously detect more narratives in it.  Use
it to refocus your attention and double-check if more economic narratives occur in
the excerpt.

# Example:
```

```
## Excerpt:
"[...]"
{
"Focused Excerpt": "[...]",
"Sequence of interest": "[...]",
"Causal Restatement":
{
"Event A": "[...]",
"causal connector": "causes",
"Event B": "[...]"
},
"Coreference Resolution":
{
"Event A": "[...]",
"causal connector": "causes",
"Event B": "[...]"
},
"Event Rephrasing":
{
"Event A": "[...]",
"causal connector": "causes",
"Event B": "[...]"
}
},
{
"Focused Excerpt": [...],
[...]
}

# Your Work
## Excerpt:
"[...]"
```