Causal Micro-Narratives

Mourad Heddaya
University of Chicago
mourad@uchicago.edu

Qingcheng Zeng
Northwestern University
qingchengzeng2027@u.northwestern.edu

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Rob Voigt
Northwestern University
robvoigt@northwestern.edu

Alexander Zentefis
Hoover Institution, Stanford University
zentefis@stanford.edu

Abstract

We present a novel approach to classify *causal micro-narratives* from text. These narratives are sentence-level explanations of the cause(s) and/or effect(s) of a target subject. The approach requires only a subject-specific ontology of causes and effects, and we demonstrate it with an application to inflation narratives. Using a human-annotated dataset spanning historical and contemporary US news articles for training, we evaluate several large language models (LLMs) on this multi-label classification task. The best-performing model—a fine-tuned Llama 3.1 8B—achieves F1 scores of 0.87 on narrative detection and 0.71 on narrative classification. Comprehensive error analysis reveals challenges arising from linguistic ambiguity and highlights how model errors often mirror human annotator disagreements. This research establishes a framework for extracting causal micro-narratives from real-world data, with wide-ranging applications to social science research.[1]

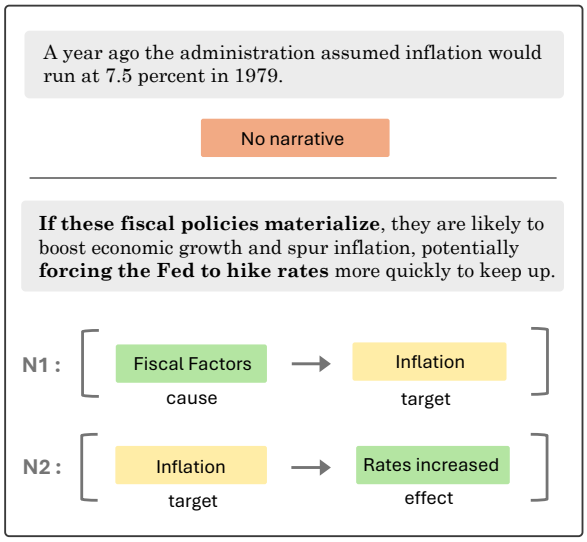


Figure 1: Causal micro-narrative classification task examples for the *target* 'inflation.' In the first sentence, no narratives are identified; in the second, two narratives (N1 and N2) are identified, one representing a cause of the *target* and the other representing an effect of it.

1 Introduction

In recent years, social scientists have increasingly recognized the power of narratives (i.e., popular stories about economic, political, or social topics) to shape individual and collective behavior. These narratives can influence people's beliefs and decisions—like when to invest in the stock market, buy a home, or pursue higher education—and can quickly spread through the collective consciousness. Nobel Prize-winning economist Robert Shiller argues that if we fail to consider and understand the properties of narratives, "we remain blind to a very real, very palpable, very important mechanism for economic change, as well as a crucial element for economic forecasting" (Shiller, 2017).

While the importance of narratives has become well recognized, formulating an operational definition remains challenging. Recent work in economics and psychology has proposed definitions based on how narratives affect people's sentiment or moral reasoning (Flynn and Sastry, 2022; Benabou et al., 2018), while other research in these fields has proposed definitions based on a *causal* account of events (Akerlof and Snower, 2016; Eliaz and Spiegler, 2020; Kendall and Charles, 2022; Morag and Loewenstein, 2023; Andre et al., 2023; Barron and Fries, 2023). These works capture important aspects of narratives, but they do not propose methods to uncover narratives from real-world data. Because narratives are disseminated to broad audiences through free-form formats like text and speech (e.g., printed, television, or web media), it is challenging to systematically extract them and quantify their prevalence and influence.

This paper aims to address both these conceptual and technical challenges. We introduce the concept of *causal micro-narratives*, along with a multi-

[1]Data is available at https://mheddaya.com/research/narratives

Proceedings of the The 6th Workshop on Narrative Understanding, pages 67–84

label classification task to extract them from text. We define *causal micro-narratives* as sentence-level explanations of the cause(s) and/or effect(s) of a target subject (e.g., an event, occurrence, emotion, phenomenon). These micro-narratives are pervasive in everyday communication. When people speak and write, they often explicitly or implicitly propose causal relations between entities and outcomes that reflect their understanding of how the world works. For instance, if someone were to say, "Jane is tired, so she won't make it to the show tonight," they implicitly propose a "micro" story that frames Jane's tiredness as the cause and her absence as the effect.

As an application of this concept, we choose *inflation* as the target that centers the micro-narratives we examine. Inflation is a popular and salient topic in news media, and can be clearly summarized by a single word, which aids in data filtering. Figure 1 illustrates how our framework distinguishes a sentence conveying a micro-narrative about inflation and one that does not. The top sentence simply reports factual news about inflation, whereas the bottom one presents two causal claims: (1) "fiscal policies" will cause inflation, and (2) the Federal Reserve will increase interest rates in response to (i.e., as an effect of) inflation. We label these two micro-narratives *fiscal factors* and *rates increased*, respectively.

We propose an ontology of causes and effects of inflation, and we create a large scale dataset of causal micro-narratives according to this ontology, classifying sentences from contemporary and historical U.S. news articles. We start with a subset of human annotations, and then use them to train various models for classifying these narratives at scale. The best model achieves F1 scores as high as 0.71, despite the difficulty of the task, having 18 classes that in some cases are semantically similar. Our comparison of different models reveal that smaller fine-tuned large language models (LLMs) outperform larger models like GPT-4o, while also being more scalable and cost efficient.

To better characterize our dataset and the performance of our classifiers, we conduct an in-depth error-analysis of inter-annotator disagreements and the in- and out-of-domain generalization of each evaluated model. Furthermore, we identify and cross-reference systematic classification errors with annotator disagreements. We find that the best-performing fine-tuned LLMs have a small performance degradation on out-of-domain data, but overall are robust to domain shifts across texts that are written 50 years apart. The errors produced by LLMs that are fine-tuned on our human-annotated data reflect the natural disagreements between annotators to a far greater extent than the errors produced by GPT-4o in a few-shot, in-context learning setting.

In summary, we make the following contributions:

1. We introduce and define the concept of causal micro-narratives, presenting a novel task for extracting them from real-world text.
2. We curate a dataset of annotated inflation-related causal micro-narratives from both historical and contemporary U.S. news articles.
3. We develop and demonstrate methods for effectively automating narrative classification at scale, making publicly available fine-tuned LLMs for this purpose. Additionally, we showcase robust out-of-domain performance of these models.
4. We conduct a comprehensive error analysis, revealing systematic similarities between model classifications and human annotation disagreements. This analysis highlights the task's complexity and identifies potential inherent ambiguities.

2 Related Work

2.1 Definitions and Theoretical Frameworks

Early work by Labov and Waletzky (1997) defined narratives as temporal accounts of event sequences, providing a formal framework for analyzing personal narratives. Building on this, Akerlof and Snower (2016) expanded the definition to include causally linked events and their underlying sources, emphasizing the role of narratives in decision-making processes.

More recent work has further refined these concepts. Eliaz and Spiegler (2020) represent narratives as directed acyclic graphs (DAGs), drawing on Bayesian Networks to model the equilibrium of narratives. Shiller (2017) likened narratives to viral phenomena, defining them as interpretive stories about economic events that spread contagiously. Benabou et al. (2018) focused on the persuasive aspect of narratives in moral decision-making, while Flynn and Sastry (2022) emphasized their contagious nature in belief formation.

Morag and Loewenstein (2023) and Barron and Fries (2023) both highlight the causal and inter-

pretive aspects of narratives. The former defines narratives as stories that establish causal links between events on a timeline, while the latter views them as subjective explanations of datasets, particularly in the context of persuasion.

2.2 Methodological and Empirical Studies

Studies have proposed different methodologies to empirically measure economic narratives. Jalil and Rua (2016) analyze word frequency in newspapers and forecasts to study inflation expectations during the Great Depression. More advanced NLP techniques have been applied as well. Lange et al. (2022) extended the RELATIO method of Ash et al. (2021) to extract narratives based on Roos and Reccius (2021)'s definition. Gueta et al. (2024) try to leverage LLMs to extract and summarize economic narrative from tweets. However, they do not clearly define *economic narrative* nor do they evaluate the LLM's performance. Flynn and Sastry (2022) utilize sentiment analysis on firm 10-K filings to build a macro model explaining economic fluctuations.

Andre et al. (2023) use open-ended surveys and DAGs to study narratives around recent high U.S. inflationary period. They contrast the narratives that households and experts write down, finding that household narratives significantly shape expectations. Their work also include experiments manipulating narratives to measure their impact on inflation expectations.

Ali et al. (2021) survey the broader field of causality extraction from text. Most causality extraction tasks are general domain, but existing methods are not very robust to complex sentence structures. Recent work by Sun et al. (2024) proposes a promising prompt-based technique with large language models to extract causal relationships in fictional stories instead of news text.

3 Causal Micro-Narratives

We define a *causal micro-narrative* as

> *a sentence-level explanation of the cause(s) and/or effect(s) of a target subject.*

The term "narrative" is most commonly applied to the discourse-level conception of story-telling that depicts sequences of events, usually in long-form texts (e.g., Piper, 2023). By contrast, here we focus on narrative fragments within individual sentences, which can capture stories about implicit and explicit cause-effect relationships that people express as they speak or write, sometimes in subtle or subconscious ways. Recent work in cognitive science highlights the prevalence of causal connectives in English and how they reveal the importance of causal relationships in the way we think and express ourselves (Iliev and Axelrod, 2016; Brown and Fish, 1983; Sanders and Sweetser, 2009).

3.1 Narrative Classification Task

We propose a narrative classification task that operationalizes our definition of *causal micro-narratives*. Unlike the more general task of causality mining (Ali et al., 2021), we suggest that a productive approach to capturing how such micro-narratives accumulate at scale should be domain-specific. Specifically, we propose a framework in which we first identify a *target* about which we hope to capture micro-narratives. Conceptually a target can by any entity, event, or phenomenon of interest.

Then, we define an ontology of the causes that can lead to that target and the effects that can follow from it. Thus, the narrative classification task is to identify, according to the ontology, sentences that express a narrative about the target subject and to predict the particular cause(s) and/or effect(s) related to the target that are present.

3.2 Case Study: Inflation Narratives

As an application of this definition and for the purposes of this paper, we focus specifically on *inflation* as the target. We develop an ontology, presented in Table 1, consisting of 8 causes of inflation and 11 effects that could follow from inflation. The causes and effects were curated by an expert economist based on domain knowledge and researching relevant resources online. See Appendix B for additional details on this process, and detailed descriptions of all the causes and effects. Ultimately, we setup the following classification task: given a sentence, identify (1) whether the sentence expresses a narrative about inflation, and (2) the expressed cause(s) and/or effect(s) of the inflation.

For this case study, we choose a target event that is fairly unambiguously summarized by a single word, *inflation*, which allows for straightforward data filtering. Nonetheless, the causal micro-narrative classification task could be applied to target events or phenomena that are expressed in more varied ways, but this would introduce more complicated filtering strategies or an additional prelimi-

nary event extraction step.

4 Dataset

We use two data sources in our investigation of inflation narratives in news: NOW Corpus for contemporary news data (Davies, 2016) and ProQuest for historical data. We selected these datasets because their differences allow us to assess the generalizability of our task and the classification methods we test. The articles in each dataset were written roughly 50 years apart and the NOW corpus includes a high degree of stylistic variation, as the articles are sourced from a range of online sources.

For each dataset, we segment articles into sentences and filter sentences that contain the keyword "inflation". Filtering allows us to focus on relevant sentences, enabling us to efficiently target our human annotations, as well as reduce the total number of sentences to a more computationally feasible quantity.

4.1 Contemporary News: NOW Corpus

We use data from the NOW Corpus covering 2012-2023. The dataset consists of online news articles, which we filter to only include U.S. articles written in English. The final filtered dataset, including "inflation" keyword filtering, contains 118,383 articles and 284,220 sentences. We use the spaCy Sentencizer (Explosion) for sentence segmentation.

4.2 Historical News: ProQuest

For historical news data, we collect news articles from local, regional, and national news publications from the ProQuest database spanning 1960-1980. See Appendix A for a list of the included publications. We chose this historical period because of the high levels of inflation that occurred throughout it, presenting an interesting opportunity to explore inflation narratives. The final dataset, including "inflation" keyword filtering, contains 392,475 articles and 751,380 sentences. We used the BlingFire (Microsoft) sentence segmentation tool, as the spaCy Sentencizer did not work well on this historical data.

4.3 Human Labeling

Three members of our team manually annotated training and test sets. In Table 2a we report the sizes of our train and test splits. We targeted train sets of approximately 1,000 examples. This provided us with sufficient training data for model fine-tuning. For the test sets, all three annotators label the same subset of data. For ProQuest, annotators initially labeled a test set of 500 sentences, however, this is reduced to 488 after filtering out texts longer than 150 words when the sentence segmentation failed.

Table 2b shows a moderate to high degree of agreement for a pragmatic annotation task, across both the historical and contemporary news annotations. We hypothesize that historical news agreement is higher than contemporary news due to (1) annotators having had more experience with the annotation since the historical annotation came second, and (2) less variation in the sourcing of historical news. The historical ProQuest news dataset primarily contains a collection of professional news publications, which results in less linguistic novelty and variation. In contrast, the contemporary news in the NOW corpus comes from a far greater variety of online sources. This variation could cause a more difficult annotation task. We present an analysis of annotator disagreement in section F. See Appendix C for annotation interface examples.

4.4 Descriptive Statistics

We focus on *causal micro-narratives* to ensure that we distinguish between general mentions of inflation in news text and a more targeted framing that presents causal stories about inflation. Analysis of the human annotations reveals that 49% and 47% of the contemporary and historical news sentences, respectively, were labeled as non-narratives. Given that these sentences are already keyword-filtered to include *inflation*, this amounts to a significant fraction of them and supports the intent of our definition and annotation scheme.

The distribution and prevalence of cause and effect narratives remains largely consistent across human annotations of both datasets. As Figure 2 shows, there are only small variations between most labels. Exceptions include *fiscal* and *govt*, which are more prevalent in historical news, and *rates*, which occurs more frequently in the contemporary data. These outliers reflect overall differences between inflation-related news in the 1960s and 1970s compared to the 2010s. These particular differences can likely be attributed to the fact that interest rate adjustment as a response to inflation did not become a significant tool deployed by the Federal Reserve until Paul Volcker's tenure as Chairman of the Fed in the 1980s (Siegel, 1998). As such, during the 60s and 70s, government spending and its relationship to inflation (*fiscal*, *govt*) was

Causes (label)	Effects (label)
Demand-side Factors (demand)	Reduced Purchasing Power (purchase)
Supply-side Factors (supply)	Cost of Living Increases (cost)
Built-in Wage Inflation (wage)	Uncertainty Increases (uncertain)
Monetary Factors (monetary)	Interest Rates Raises (rates)
Fiscal Factors (fiscal)	Income or Wealth Redistribution (redistribution)
Expectations (expect)	Impact on Savings (savings)
International Trade & Exchange Rates (international)	Impact on Global Trade (trade)
Other Causes (other-cause)	Cost-Push on Businesses (cost-push)
	Social and Political Impact (social)
	Government Policy & Public Finances Impact (govt)
	Other Effects (other-effect)

Table 1: Inflation Narrative Causes and Effects. The **label** in parentheses refers to the abbreviated name used during classification in both few-shot and fine-tuning experiments. See Appendix 6 for additional details.

	Historical	Contemporary
Train / Test	999 / 488	1,119 / 201
Median Words Per Sentence	26	25

(a) Human annotation train and test set sizes, and median sentence lengths.

Dataset	Binary	Multi-class
Contemporary	0.67	0.59
Historical	0.80	0.66

(b) Test set Inter-annotator agreement: Krippendorff's alpha using MASI distance weighting (Hayes and Krippendorff, 2007)

Table 2: Human annotation statistics

a more common topic of discussion.

5 Methods

To determine the most effective approach to classify narratives, we compare the performance of LLMs on our classification task for both in-context learning and fine-tuning settings. We focus on these two settings We format the annotations associated with each sentence as JSON to facilitate automatic processing (see Appendix D). The LLMs are evaluated on their classification output, expected to be in JSON as well. We conduct separate experiments with the contemporary and historical data and train separate models for each dataset.

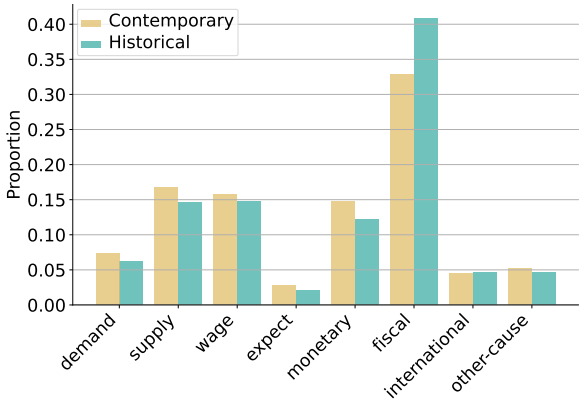
5.1 In-Context Learning

LLMs have been shown to be effective in-context, or *few-shot*, learners (Brown et al., 2020), so we tested GPT-4o in this setting by providing definitions for all the labels along with 24 narrative classification examples, one for each distinct cause and effect, as well as 5 examples of non-narratives. We use greedy decoding and do not constrain the generation in any way, but find that GPT-4o reliably generated JSON in the correct format.

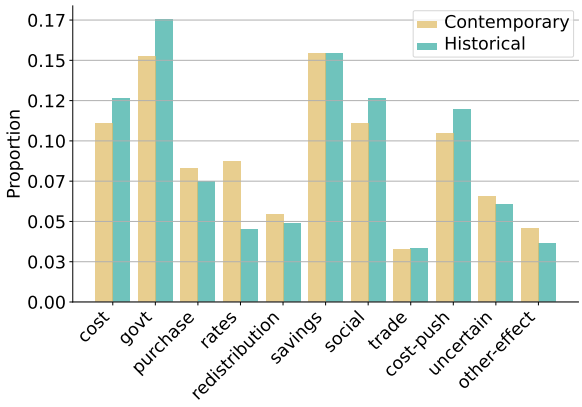
5.2 Fine-tuning

The second modeling approach we evaluate is fine-tuning two open-source, pre-trained LLMs: Llama 3.1 8B (`meta-llama/Meta-Llama-3.1-8B`) and Phi-2 (`microsoft/phi-2`). We chose these two models because they represent high quality LLMs that have performed well on LLM benchmarks. Additionally, because of their relatively smaller parameter counts compared to other recent LLMs, they are well suited for efficient inference at scale. Indeed, while this classification task test set is relatively small, the ultimate aim of our work is to enable researchers to do complex narrative classification tasks at the scale of millions of sentences from news articles across long time horizons.

For fine-tuning, the input consists of the possible causes and effects, their definitions, and a brief instruction. We include the full fine-tuning prompt in Appendix D. We follow standard auto-regressive language modeling but only back propagate the language modeling loss for tokens associated with binary and multi-class labels, rather than other tokens associated with the JSON notation. We use LoRA-based Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2021) to train a subset of the parameters. See Appendix E for fine-tuning hyper-parameters.



(a) Inflation cause narratives.



(b) Inflation effect narratives.

Figure 2: Proportions of narrative classes in human annotations. This data combines both the train and test sets. For the test set, majority vote is used to identify one annotation instance.

In few- and zero- shot experiments both models achieved extremely low F1 scores (0.12 or lower). As a result, for the purposes of this work, we focus on evaluating fine-tuned versions of the two open-source models, rather than their zero-shot performance.

5.3 Evaluation

We evaluate each aspect of a narrative classification separately using micro-averaged F1 scores. We use micro averaging, rather than weighted- or macro-averaging to get an overall picture of model performance across all instances, including less represented classes. Micro-averaged scores use the standard binary-F1 score formula, but, importantly, the precision and recall scores are based on true and false positives across all instances, irrespective of individual class distinctions. Because each sentence could have narratives with multiple causes and/or effects, micro-averaged F1 differs from a regular accuracy score.

To resolve disagreements between annotators in the test set, we use majority rule to identify gold-labels. In practice, 97% of the test set instances have agreement between at least two annotators, allowing us to retain almost the entire test set for evaluation.

6 Results

	Llama3.1	Phi-2	GPT-4o
Binary			
Hist.	0.78	**0.83**	0.47
Contemp.	**0.87**	0.79	0.63
Multiclass			
Hist.	**0.62**	0.60	0.46
Contemp.	**0.71**	0.65	0.57

Table 3: Summary F1 scores for the inflation narrative classification task on Historical (Hist.) and Contemporary (Contemp.) datasets. Phi-2 and Llama 3.1 8B are fine-tuned on a combined dataset totalling 2,118 instances. F1 uses micro-averaging for multi-class and binary for narrative detection. All scores are calculated using majority vote between the three annotators as ground truth. 14 test set instances with no majority annotation are ignored in this score. **Bolded** values indicate the best performing model on each task (binary and multiclass) and each test set (Historical and Comptemporary).

We compare model performance in Table 3. Fine-tuned Llama 3.1 8B performs the best and, along with Phi-2, outperforms GPT-4o. GPT-4o particularly suffers on Historical data and the binary narrative detection overall.

To better understand how models trained on these datasets may generalize to news from other periods, we present in Table 4 a breakdown of model performance in several training and evaluation settings. First, we evaluate how well models fine-tuned on Historical and Contemporary data perform on corresponding held-out data, assessing in-domain generalization. Second, we compare how well models generalize to out-of-distribution (OOD) data by evaluating performance on Historical data when trained on Contemporary data, and vice-versa. Finally, we combine both the historical and contemporary data during the learning phase and evaluate performance on the individual datasets, revealing how well models can learn from the additional data despite the domain-shift.

Train \ Test	Llama3.1 8B		Phi-2		GPT-4o	
	Hist.	Contemp.	Hist.	Contemp.	Hist.	Contemp.
Binary						
Hist.	0.64	0.75	0.75	0.82	0.47	0.70
Contemp.	0.73	0.82	0.75	0.83	0.51	0.63
Hist. + Contemp.	0.78	**0.87**	**0.83**	0.79	0.39	0.43
Multiclass						
Hist.	0.55	0.59	0.57	0.63	0.46	0.60
Contemp.	0.52	0.63	0.53	0.66	0.48	0.57
Hist. + Contemp.	**0.62**	**0.71**	0.60	0.65	0.43	0.46

Table 4: F1 scores for the inflation narrative classification task on Historical (Hist.) and Contemporary (Contemp.) Datasets. Phi-2 and Llama 3.1 8B are fine-tuned. F1 uses micro-averaging for multi-class and binary for narrative detection. All scores are calculated using majority vote between the three annotators as ground truth. 14 test set instances with no majority annotation are ignored in this score. Columns specify the datasets used for training; and rows, the results on test sets. **Bolded** values indicate the best performing model and training data combination for each task (binary and multiclass) and each test set (Historical and Comptemporary).

6.1 In-Domain Generalization

When trained and evaluated on the same individual dataset, Phi-2 outperforms other models. Interestingly, however, Llama 3.1 8B is better able to learn from both the Historical and Contemporary datasets, exhibiting impressive improvements of up to 14%, despite the 50-year gap between the news in the two datasets. In contrast, Phi-2 struggles and even degrades in performance on Contemporary data multi-class classification. All models perform better on contemporary data, likely because recent text and language from 2012-2023 are more prevalent in their pre-training corpora than historical newspaper data.

6.2 Out-of-Domain Generalization

On the multiclass narrative classification task, a common pattern emerges across both fine-tuned models. We observe that test set performance degrades by 3-4% on OOD data relative to in-domain data. This represents a moderate drop in performance and could be attributed to changes in the distribution of narratives across the Historical and Contemporary datasets, as explained in Section 4.4 and Figure 2. In contrast, the binary prediction task reveals a different effect. Phi-2 performs the same regardless of which dataset is used for training and which is used for testing but Llama 3.1 8B achieves up to an 11% improvement on narrative detection in Historical news sentences when trained on the Contemporary data. In the reversed setting, Llama 3.1 8B performance degrades by 7%. This pattern suggests that training Llama on Contemporary data is more successful than Historical data.

6.3 Error Analysis

To better understand model performance on this task and the variation between fine-tuning a smaller LLM and few-shot prompting a large propriertary LLM, we conduct a fine-grain analysis of the individual narrative classification predictions as well as an analysis of the three sets of human annotations to better understand the disagreements that exist between them and how those disagreements may related to model prediction errors. As the best performing LLM overall, we focus on Llama 3.1 8B (henceforth, *Lllama*) and compare it to GPT 4o, the only propriertary model in our experiments.

Human Annotator Disagreements By majority rule, our three human annotators find partial agreement on 474 out of 488 test set instances, and full agreement on 471. While this is a higher rate of majority agreement, there are nonetheless non-negligible disagreements between individual annotators. Since we use training data sourced from each annotator individually, understanding these disagreements can contextualize how model performance is impacted. Most annotator disagreements stem from differing judgments on narrative presence, not category assignment. Annotators rarely clash over which specific narrative category to apply, but often diverge on whether a narrative exists in the text at all. Furthermore, certain annotators are systematically more likely to detect narratives

Sentence	Llama 3.1 8b	Majority Annotation
"The corrosive effects of inflation eat away at the ties that bind us together as a people," said President Carter Thursday in the third of the messages–the budget, the State of the Union, and the Economic Report–that make up the traditional January triad.	no-narrative	social
But he acknowledged that the Administration-projected rate of 6.5% to 7% inflation this year still made it the nation s worst domestic problem.	no-narrative	social
He said inflation was every American's problem and that the nation's economic, military and spiritual strength depended on solving it.	no-narrative	social
'They have and will cause Inflation to accelerate in the state and the Chicago area, destroy jobs that otherwise would be available, lower family income, and increase taxes,"he said.	fiscal	govt, purchase, cost-push
"Inflation has slowed, but people's perception of that changes," he said.	no-narrative	expect
Carter finally became convinced that inflation was the No. 1 problem.	no-narrative	govt
Consequently, increases in valuation due to inflation do indeed raise the number of actual dollars in property taxes owed.	govt	savings

Table 5: Comparison of fine-tuned LLama 3.1 8B and human annotations.

than others, driving this specific form of disagreement.

Hallucinating Narratives Fine-tuning is effective at teaching a model to distinguish between narratives and non-narratives, compared to in-context learning. GPT-4o, which was not fine-tuned, correctly classifies roughly 47% and 60% fewer non-narratives in the contemporary NOW and historical ProQuest test sets, respectively, than Llama. Despite extensive experimentation with different prompts, we consistently observed that GPT-4o struggled to understand the distinction we stipulate between narratives and non-narratives. We can likely attribute this to our precise definition of narrative, such that these otherwise highly capable LLMs have limited in-context demonstration data to draw on to learn this capability.

Natural Variation & Ambiguity in Language Table 5 presents several instances where Llama predictions did not match the human labels. The first three examples illustrate that Llama's impressive 0.87 F1 score on binary narrative detection comes at the cost of false negative predictions. In fact, these three instances of failing to predict *Social & Political Impact (social)* are representative of the most common type of false negative error in Llama predictions. Interestingly, annotating *social* or not is the most common disagreement of this type among the annotators. Nonetheless, the three examples in Table 5 show failures of Llama to identify the implied, yet clear, references to inflation's social and political impact.

In contrast, the final four examples demonstrate the natural ambiguity and difficulty inherit in this task. Consider the fourth sentence. While to a human, it may be quite natural to understand this sentence as inflation being the cause of the job destruction, lower family income, and increased taxes, it is not explicit in the sentence. In fact, the more explicit mention of causation in the sentence is "they have and will cause inflation". Llama predicts a *cause of inflation* narrative ("fiscal"), whereas the reference labels are *effects of inflation* ("govt, purchase, cost-push"). In practice, this sentence does not mention who "they" is referring to, so the prediction, while a reasonable guess, is not supported. The final three examples show scenarios where the Llama predictions and human annotations could both be considered correct, depending on one's perspective. All these examples illustrate the challenging nature of the task and the natural variation that is inherent to it.

7 Conclusion

This paper proposes a *causal micro-narrative* classification task. By developing a comprehensive classification scheme and leveraging both fine-tuned and few-shot prompted large language models, we demonstrate the feasibility of automating the detection and categorization of these narratives at scale. Our results show that fine-tuned models, particularly Llama 3.1 8B, outperform few-shot prompted models in distinguishing between narrative and non-narrative content, while maintaining competitive performance in classifying specific narrative types.

The error analysis reveals that the task of iden-

tifying causal micro-narratives is inherently complex, with natural ambiguities in language and variation in human interpretations. Despite these challenges, our approach provides a foundation for future research in narrative analysis within the social sciences. By enabling the systematic extraction of causal narratives from large-scale textual data, this work opens up new possibilities for studying the evolution and impact of narratives over time, potentially offering valuable insights for policymakers, economists, and social scientists alike.

8 Limitations

The method we propose for extracting and classifying *causal micro-narratives* requires the manual development of an ontology of causes and effects for any new target. This limits automated data-driven discovery of new narratives (i.e., causes and effects not already pre-established). However, the binary micro-narrative detection task included in this paper may be helpful in filtering a large corpus into a smaller dataset of sentences that contain narratives. This may facilitate discovering new narratives, either manually, or with an automated method. In this paper, we do not evaluate this use-case but we believe this to be a good direction for future work.

References

George A Akerlof and Dennis J Snower. 2016. Bread and bullets. *Journal of Economic Behavior & Organization*, 126:58–71.

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

Peter Andre, Ingar Haaland, Christopher Roth, and Johannes Wohlfart. 2023. Narratives about the macroeconomy. Working paper.

Elliott Ash, Germain Gauthier, and Philine Widmer. 2021. Relatio: Text semantics capture political and economic narratives. *arXiv preprint arXiv:2108.01720*.

Kai Barron and Tilman Fries. 2023. Narrative persuasion. Working paper.

Roland Benabou, Armin Falk, and Jean Tirole. 2018. Narratives, imperatives, and moral reasoning.

Roger Brown and Deborah Fish. 1983. The psychological causality implicit in language. *Cognition*, 14(3):237–273.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Mark Davies. 2016. Corpus of news on the web (now). `https://www.english-corpora.org/now/`. Available online at `https://www.english-corpora.org/now/`.

Kfir Eliaz and Ran Spiegler. 2020. A model of competing narratives. *American Economic Review*, 110(12):3786–3816.

Explosion. Spacysentencizer.

Joel P Flynn and Karthik Sastry. 2022. The macroeconomics of narratives. *Available at SSRN 4140751*.

Almog Gueta, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart. 2024. Can llms learn macroeconomic narratives from social media? *Preprint*, arXiv:2406.12109.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Rumen Iliev and Robert Axelrod. 2016. Does causality matter more now? increase in the proportion of causal language in english texts. *Psychological Science*, 27(5):635–643. PMID: 26993741.

Andrew J. Jalil and Gisela Rua. 2016. Inflation expectations and recovery in spring 1933. *Explorations in Economic History*, 62:26–50.

Chad W Kendall and Constantin Charles. 2022. Causal narratives. Working paper.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative and Life History*.

Kai-Robin Lange, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael WM Roos, and Carsten Jentsch. 2022. *Towards extracting collective economic narratives from texts*. 963. Ruhr Economic Papers.

Microsoft. Blingfire.

Dor Morag and George Loewenstein. 2023. Narratives and valuations.

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39, Singapore. Association for Computational Linguistics.

Michael W.M. Roos and Matthias Reccius. 2021. *Narratives in Economics*. RWI.

Ted Sanders and Eve Sweetser. 2009. *Introduction: Causality in language and cognition – what causal connectives and causal verbs reveal about the way we think*, pages 1–18. De Gruyter Mouton, Berlin, New York.

Robert J Shiller. 2017. Narrative economics. *American economic review*, 107(4):967–1004.

J.J. Siegel. 1998. *Stocks for the Long Run: The Definitive Guide to Financial Market Returns and Long-term Investment Strategies*. McGraw-Hill.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A ProQuest Newspapers

Chicago Tribute, Chicago Defender, Los Angeles Times, Los Angeles Sentinel, Atlanta Daily World, Cleveland Call and Post, Detroit Free Press, Indianapolis Star, Kansas City Call, Louisville Courier Journal, Louisville Defender, Michigan Chronicle, Minneapolis Star Tribune, New York Amsterdam News, New York Tribute / Herald Tribune, Norfolk Journal and Guide, Philadelphia Tribune, Pittsburgh Courier, Pittsburgh Post-Gazette, San Francisco Chronicle, St. Louis American, St. Louis Post Dispatch, The Baltimore Afro-American, The Boston Globe, The Christian Science Monitor, The Cincinnati Enquirer, The Nashville Tennessean, The New York Times, The Wall Street Journal, The Washington Post, U.S. Newsstream, U.S. Major Dailies.

B Classification Task

Narrative	Label	Definition Excerpt
Causes		
Demand-side Factors	demand	Pull-side or demand-pull inflation.
Supply-side Factors	supply	Push-side or cost-push inflation.
Built-in Wage Inflation	wage	Also known as wage inflation or wage-price spiral.
Monetary Factors	monetary	Central bank policies that contribute to inflation.
Fiscal Factors	fiscal	Government policies that contribute to inflation.
Expectations	expect	The expectation that inflation will rise often leads to a rise in inflation.
International Trade & Exchange Rates	international	International trade and exchange rate factors that can cause inflation.
Other Causes	other-cause	Causes not included in above.
Effects		
Reduced Purchasing Power	purchase	Inflation erodes the purchasing power of money (such as the U.S. dollar) over time.
Cost of Living Increases	cost	Inflation can raise the cost of living, particularly impacting individuals on fixed incomes, pensioners, and those with lower wages.
Uncertainty Increases	uncertain	Inflation can create uncertainty about future prices (or future inflation itself), particularly if the inflation is high or unpredictable.
Interest Rates Raises	rates	Central banks may respond to inflation by raising interest rates to curb spending and investment.
Income or Wealth Redistribution	redistribution	Inflation can redistribute income and wealth between people in the economy.
Impact on Savings	savings	Inflation can affect various types of savings/financial investments.
Impact on Global Trade	trade	Inflation can impact a country's trade or competitiveness in global markets.
Cost-Push on Businesses	cost-push	Rising costs of production due to inflationary pressures can squeeze business profits, potentially leading to reduced investment, job cuts and unemployment, or higher prices for consumers.
Social and Political Impact	social	Inflation can have social and political economic implications.
Government Policy & Public Finances Impact	govt	Inflation may impact government spending policies or programs.
Other Effects	other-effect	Effects not included in above.

Table 6: Narrative categories, their label used in the classification task, and an excerpt of their definitions. These categories were selected and define by a domain expert, using a combination of domain knowledge, google searches, and LLM interactions. When using a LLM (Open AI ChatGPT 3.5, Google Bard/Gemini, Anthropic Claude), the prompt was "what are the causes (effects) of inflation? Describe the economic mechanisms and give examples". If we wanted to expand on a cause (effect), the prompt was "explain economic mechanisms and examples of xxxx as a cause (effect) of inflation". We also relied on Google searches of "causes (effects) of inflation".

C Annotation Interface

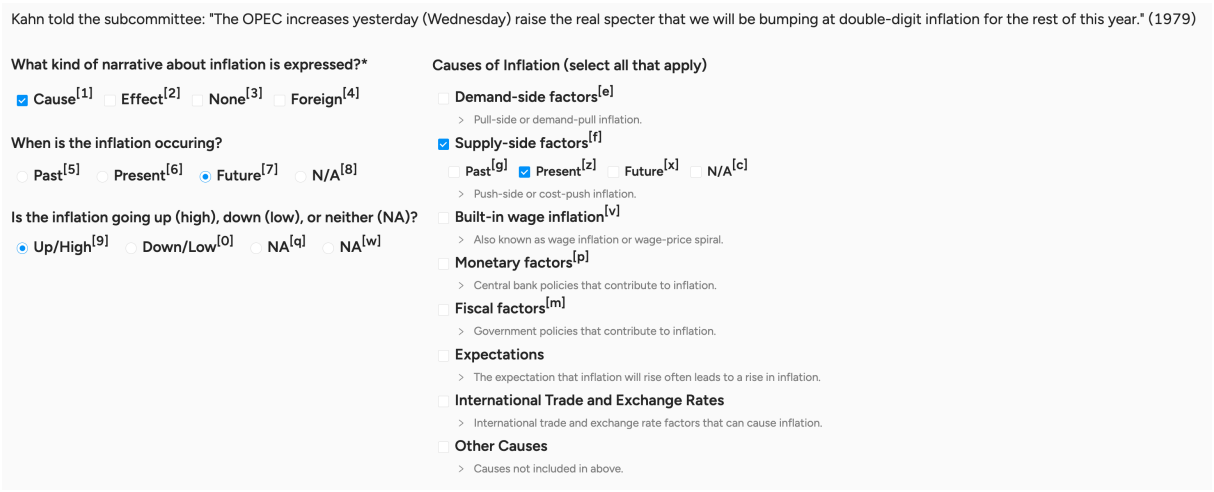


Figure 3: Example of an annotation for a narrative about the cause of inflation.

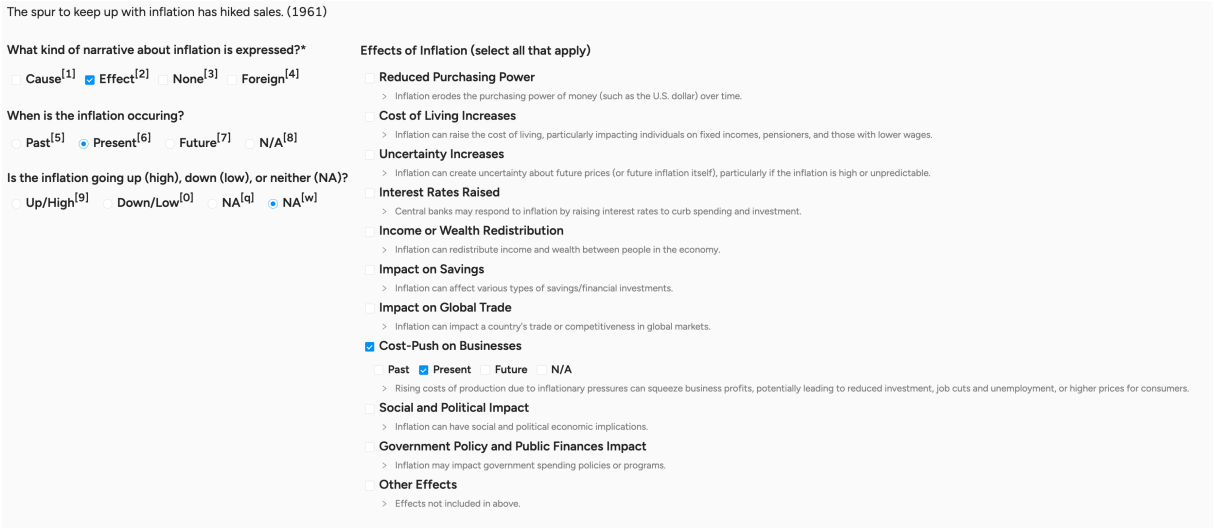


Figure 4: Example of an annotation for a narrative about the effect of inflation.

D LLM Prompts and Inputs

Due to the hierarchical multi-label classification task, we represent a complete narrative classification as JSON. This paper focuses only on the prediction results; i.e., the values associated with the fields “contains-narrative" and “narratives". However, our task includes additional information which we will discuss in future work. We define the JSON schema as follows:

```
{
  "foreign": true|false,
  "contains-narrative": true|false,
  "inflation-narratives": [
    "inflation-time": "past"|"present"|"future"| "na",
    "inflation-direction": "down"|"up"|"na",
    "narratives": [
        {"causes"|"effect": category, "time": "past"|"present"|"future"| "na"},
        ...
    ]
  ] | null
}
```

```
Below are lists of causes and effects of inflation.

Causes of inflation:
[demand] Demand-side factors: Pull-side or demand-pull inflation.
[supply] Supply-side factors: Push-side or cost-push inflation.
[wage] Built-in wage inflation: Also known as wage inflation or wage-price spiral.
[monetary] Monetary factors: Central bank policies that contribute to inflation.
[fiscal] Fiscal factors: Government policies that contribute to inflation.
[expect] Expectations: The expectation that inflation will rise often leads to a rise in inflation.
[international] International Trade and Exchange Rates: International trade and exchange rate factors that can cause inflation.
[other-cause] Other Causes: Causes not included in above.

Effects of inflation:
[purchase] Reduced Purchasing Power: Inflation erodes the purchasing power of money (such as the U.S. dollar) over time.
[cost] Cost of Living Increases: Inflation can raise the cost of living, particularly impacting individuals on fixed incomes, pensioners, and those with lower wages.
[uncertain] Uncertainty Increases: Inflation can create uncertainty about future prices (or future inflation itself), particularly if the inflation is high or unpredictable.
[rates] Interest Rates Raised: Central banks may respond to inflation by raising interest rates to curb spending and investment.
[redistribution] Income or Wealth Redistribution: Inflation can redistribute income and wealth between people in the economy.
[savings] Impact on Savings: Inflation can affect various types of savings/financial investments.
[trade] Impact on Global Trade: Inflation can impact a country's trade or competitiveness in global markets.
[cost-push] Cost-Push on Businesses: Rising costs of production due to inflationary pressures can squeeze business profits, potentially leading to reduced investment, job cuts and unemployment, or higher prices for consumers.
[social] Social and Political Impact: Inflation can have social and political economic implications.
[govt] Government Policy and Public Finances Impact: Inflation may impact government spending policies or programs.
[other-effect] Other Effects: Effects not included in above.

Identify all causes and effects of inflation that are expressed in the sentence:
% \{SENTENCE\}
```

Figure 5: Causal Micro-Narrative classification prompt. For few-shot with GPT-4o examples are listed before the final sentence.

E Hyperparameters

Max Steps	Effective Batch Size	Optimizer	Learning Rate	LoRA r, α
600	16	AdamW	1e-4	16, 32

Table 7: Fine-tuning hyper-parameters for Phi-2 and Llama 3.1 8B.

F Confusion Matrices

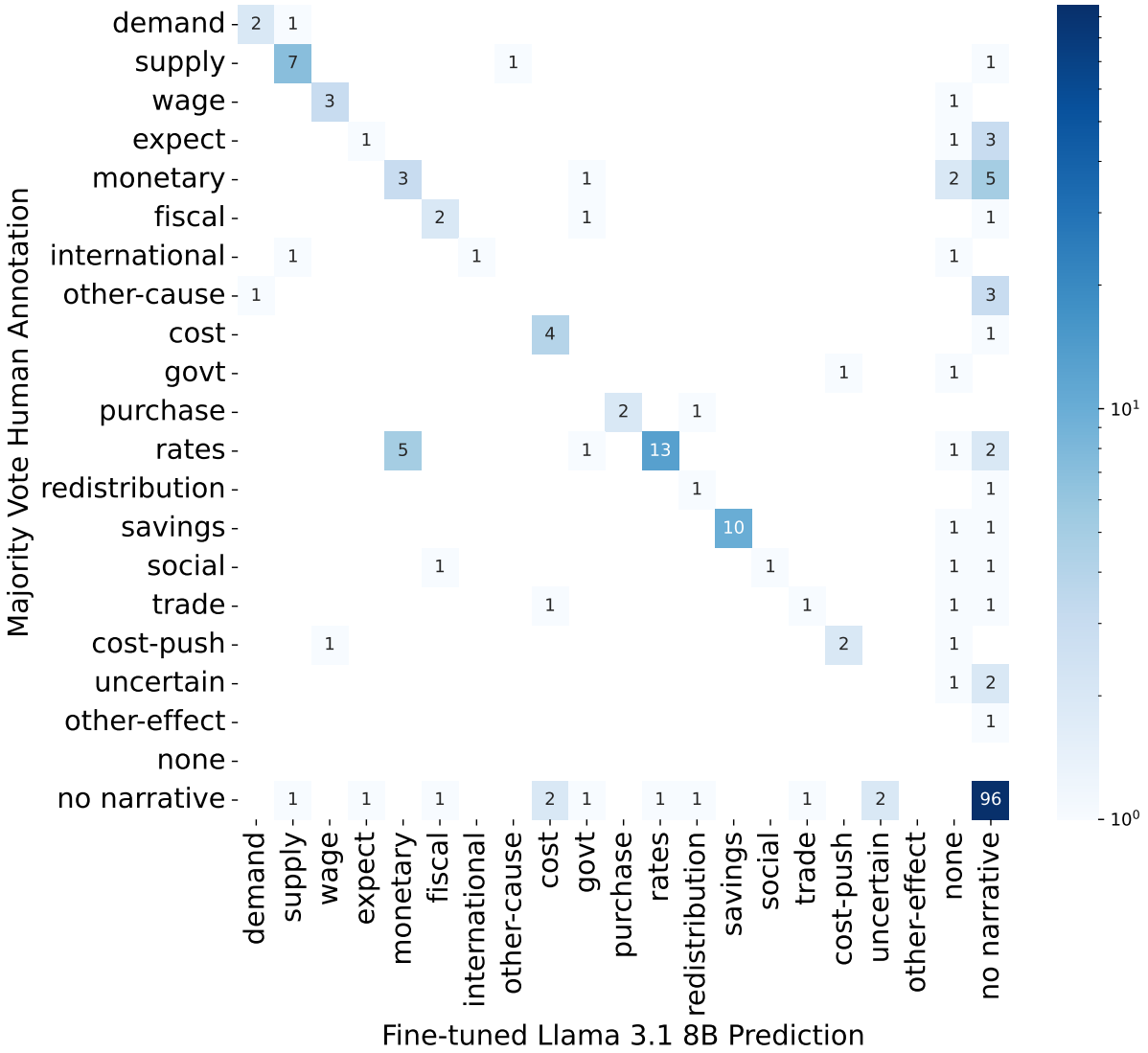


Figure 6: Confusion matrix: NOW Test set fine-tuned Llama 3.1 8B predictions against majority vote human ground-truths. Label “none" indicates when a narrative does not match any of the narratives in the comparison set. For example, if a model prediction is that a sentence contains a narrative about “rates" and one about “monetary" and the human label is “rates", then “monetary" would be matched with “none".

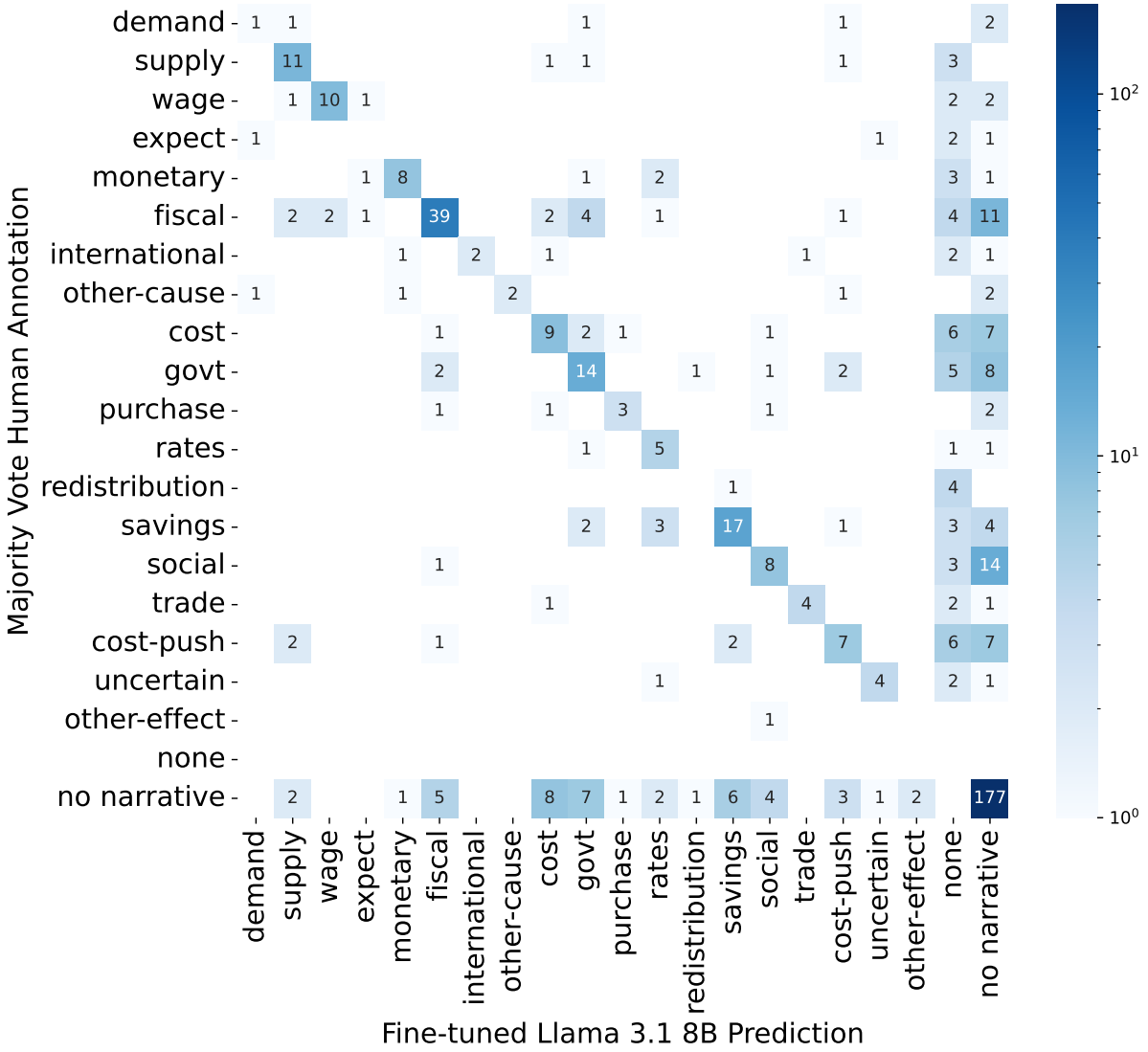


Figure 7: Confusion matrix: ProQuest Test set fine-tuned Llama 3.1 8B predictions against majority vote human ground-truths. Label “none" indicates when a narrative does not match any of the narratives in the comparison set. For example, if a model prediction is that a sentence contains a narrative about “rates" and one about “monetary" and the human label is “rates", then “monetary" would be matched with “none".

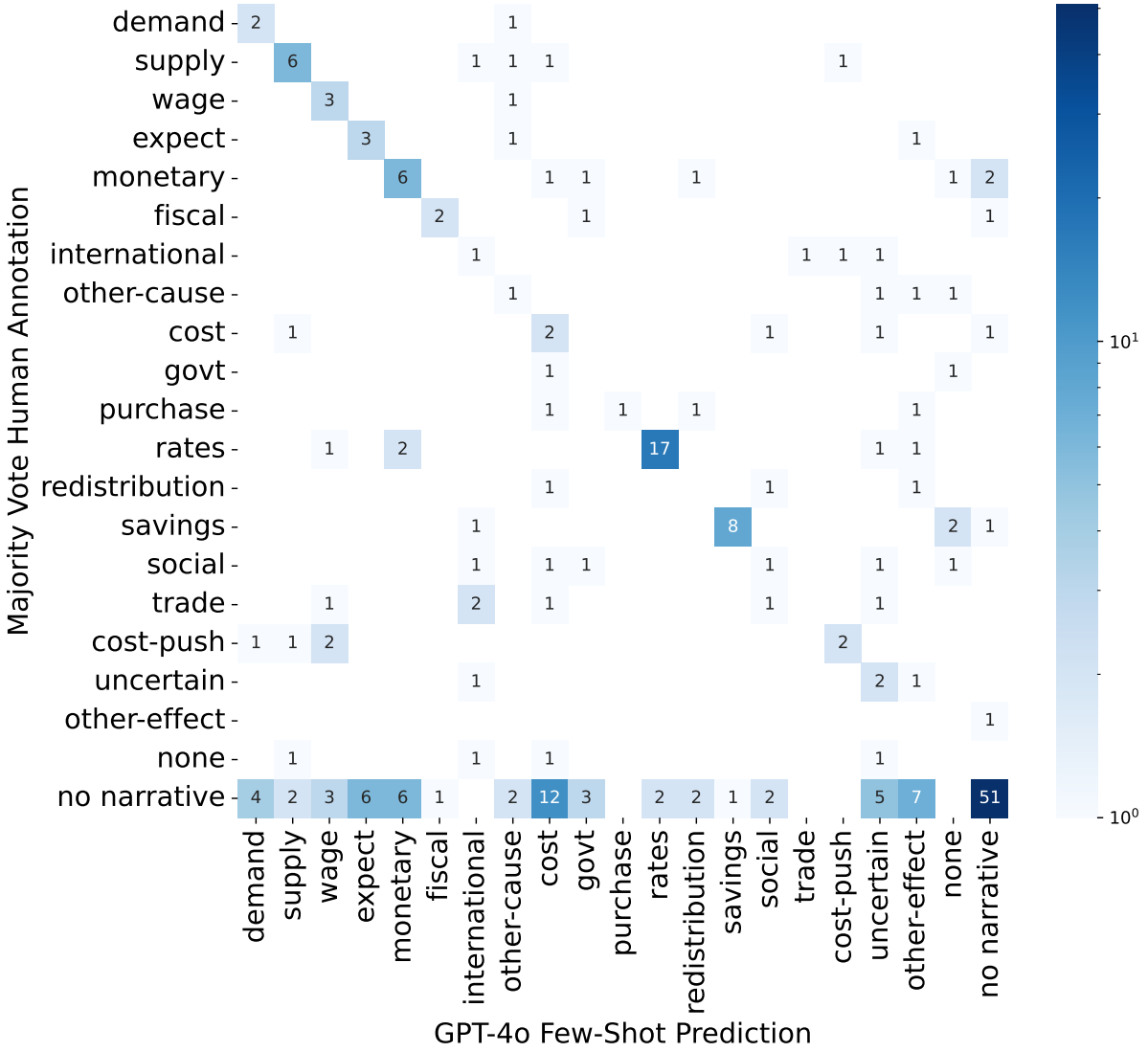


Figure 8: Confusion matrix: NOW Test set GPT-4o predictions against majority vote human ground-truths. Label “none" indicates when a narrative does not match any of the narratives in the comparison set. For example, if a model prediction is that a sentence contains a narrative about “rates" and one about “monetary" and the human label is “rates", then “monetary" would be matched with “none".

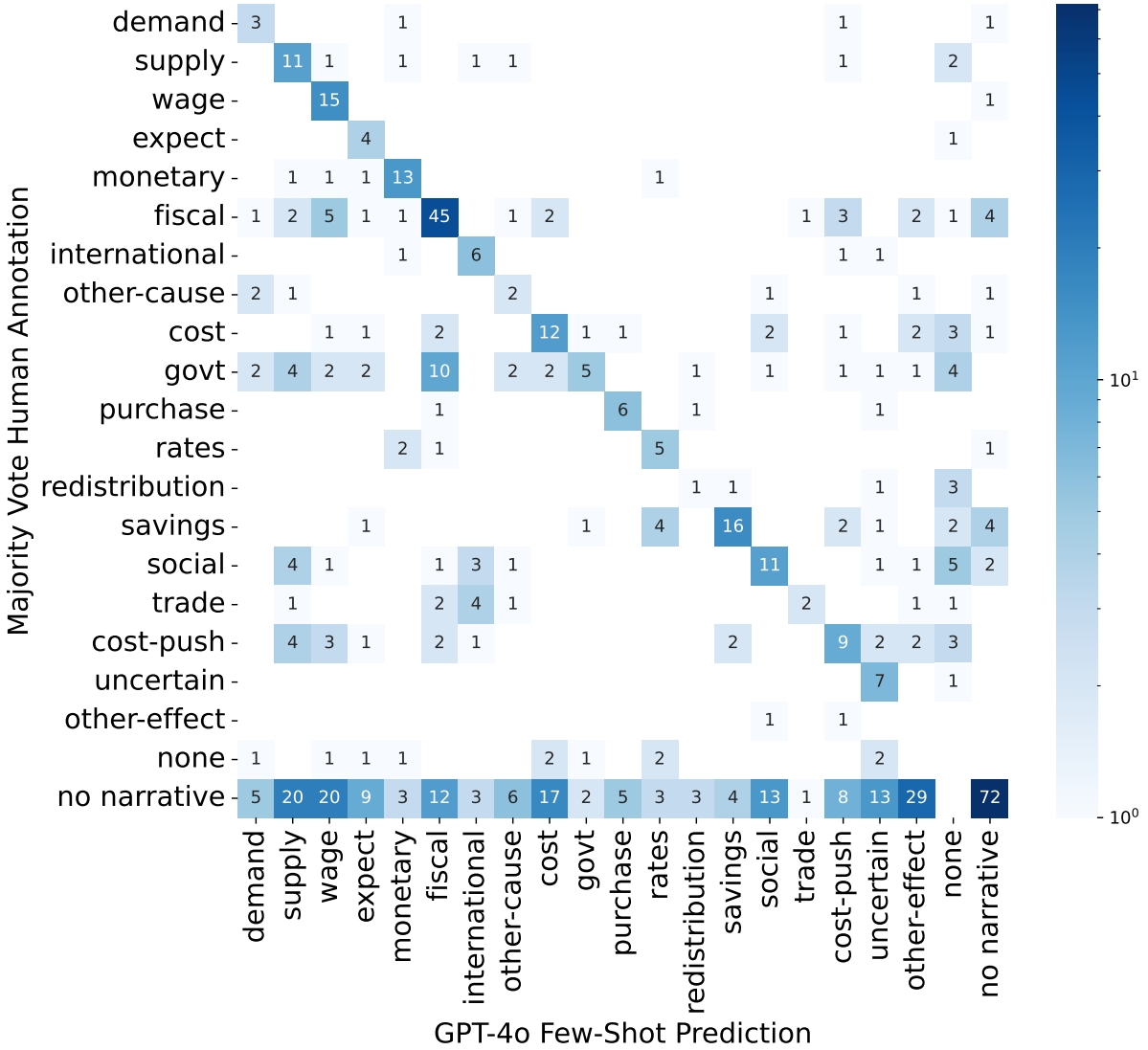


Figure 9: Confusion matrix: ProQuest Test set GPT-4o predictions against majority vote human ground-truths. Label “none" indicates when a narrative does not match any of the narratives in the comparison set. For example, if a model prediction is that a sentence contains a narrative about “rates" and one about “monetary" and the human label is “rates", then “monetary" would be matched with “none".