

THE UNIVERSITY OF CHICAGO

LARGE LANGUAGE MODELS FOR OPEN-ENDED DOMAINS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES DIVISION
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
MOURAD HEDDAYA

CHICAGO, ILLINOIS
AUGUST 2026

Copyright © 2026 by Mourad Heddaya
All Rights Reserved

TABLE OF CONTENTS

LIST OF FIGURES	vii
LIST OF TABLES	ix
ACKNOWLEDGMENTS	xii
ABSTRACT	xiii
1 INTRODUCTION	1
2 NARRATIVES AND INFLATION EXPECTATIONS	4
Part I: Causal Micro-Narratives	4
2.1 Introduction	4
2.2 Related Work	7
2.2.1 Definitions and Theoretical Frameworks	7
2.2.2 Methodological and Empirical Studies	8
2.3 Causal Micro-Narratives	9
2.3.1 Narrative Classification Task	9
2.3.2 Case Study: Inflation Narratives	10
2.4 Dataset	11
2.4.1 Contemporary News: NOW Corpus	11
2.4.2 Historical News: ProQuest	11
2.4.3 Human Labeling	12
2.4.4 Descriptive Statistics	13
2.5 Methods	14
2.5.1 In-Context Learning	14
2.5.2 Fine-tuning	14
2.5.3 Evaluation	15
2.6 Results	16
2.6.1 In-Domain Generalization	17
2.6.2 Out-of-Domain Generalization	17
2.6.3 Error Analysis	19
2.7 Part I Summary	20
Part II: A Century of Inflation Narratives	21
2.8 Motivation and Research Questions	21
2.9 Data	22
2.9.1 Newspaper Corpus	22
2.9.2 Economic Variables	23
2.10 Classification at Scale	23
2.10.1 Cross-Temporal Validation	23
2.10.2 Comparison with Keyword Methods	23
2.10.3 Unsupervised Validation via TopicGPT	24

2.11	Inflation Narratives Through Time and Space	25
2.11.1	Narrative Prevalence	25
2.11.2	Narratives Through Time	26
2.11.3	Regional Differences	26
2.12	Narratives and Household Expectations	28
2.12.1	Empirical Design	29
2.12.2	National Expectations	30
2.12.3	Dispersion in Expectations	30
2.12.4	Income-Based Heterogeneity	32
2.12.5	Education-Based Heterogeneity	34
2.12.6	Narratives and Realized Inflation	36
2.13	Discussion	38
2.14	Limitations	39
3	CASESUMM: A LARGE-SCALE DATASET FOR LONG-CONTEXT SUMMARIZA- TION FROM U.S. SUPREME COURT OPINIONS	43
3.1	Introduction	43
3.2	Related Work	45
3.3	Dataset	48
3.4	Experiment Setup	50
3.4.1	Data and Modeling	50
3.4.2	Evaluation Strategies	51
3.5	Results	53
3.5.1	Automated Evaluation Favors Fine-tuned Mistral Summaries	53
3.5.2	Summaries do not Scale with Opinion Length as much as Official Syl- labuses	54
3.5.3	Human Evaluation Disagrees with Automatic Evaluation	57
3.5.4	Error Analysis	61
3.6	Discussion	63
3.7	Conclusion	65
3.8	Limitations	66
4	CIVICCHATS: DESIGNING AI SYSTEMS TO SUPPORT INFORMED AND JUS- TIFIED VOTER DECISIONS	69
4.1	Introduction	69
4.2	Related Work	72
4.3	Methods	74
4.3.1	CivicChats	74
4.3.2	Experimental Design	78
4.3.3	Outcome Measures and Analysis	80
4.4	Findings	83
4.4.1	Vote Choice	83
4.4.2	Reasoning and Engagement	86
4.4.3	Understanding and Confidence	91

4.5	Conclusion	92
5	CONCLUSION	94
5.1	Summary of Contributions	94
APPENDIX A CAUSAL MICRO-NARRATIVES: SUPPLEMENTARY MATERIAL		95
A.1	ProQuest Newspapers	95
A.2	Inflation Narrative Ontology	95
A.3	Annotation Interface	98
A.4	Classification Prompt	99
A.5	Fine-tuning Hyperparameters	100
A.6	Confusion Matrices	100
APPENDIX B CASESUMM: SUPPLEMENTARY MATERIAL		106
B.1	Summary Generation	106
B.1.1	Mistral Fine-tuning and Generation Details	106
B.1.2	GPT-4 Generation Details	106
B.1.3	Initial GPT-4 Prompt	106
B.1.4	DSPy-Optimized GPT-4 Prompt	107
B.2	Automatic Evaluation	107
B.3	G-Eval	107
B.3.1	Consistency Prompt	108
B.3.2	Relevance Prompt	109
B.3.3	Coherence Prompt	110
B.3.4	Fluency Prompt	111
B.3.5	Sensitivity Prompt	112
B.3.6	Specificity Prompt	113
B.3.7	Clarity Prompt	114
B.3.8	Style Prompt	115
B.4	Human Evaluation	117
B.4.1	Dimensions of Summary Quality	117
B.4.2	Annotation Interface	118
APPENDIX C CIVICCHAT: SUPPLEMENTARY MATERIAL		120
C.1	Official Ballot Text	120
C.2	User Interface	121
C.3	Additional Methodology Details	121
C.4	Discourse-Mode Prompts	121
C.5	Rationale Annotation Prompts	127
C.6	Discourse-Mode Evaluation Checklists	129
C.7	Ballot Measure Text	134
C.8	Supplementary Results Tables	134
C.8.1	Robustness of Focal Contrasts	134

REFERENCES	139
----------------------	-----

LIST OF FIGURES

2.1	Causal micro-narrative classification task examples for the <i>target</i> ‘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 <i>target</i> and the other representing an effect of it.	5
2.2	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.	13
2.3	Category-level semantic similarity between the 19 manual narrative categories and the 11 categories discovered by TopicGPT on 10,000 randomly sampled inflation sentences. Darker shades indicate stronger similarity. All discovered categories map onto existing entries in the manual ontology, and no novel causal mechanisms emerge from the unsupervised procedure.	25
2.4	One-year rolling average monthly shares of the top three cause and effect narratives across the full ProQuest sample, January 1923 to January 2025. Gray shaded regions indicate NBER recession periods.	40
2.5	Differences in inflation narrative shares between newspapers headquartered in Northern and Southern cities. Error bars represent 95% confidence intervals. Purple bars indicate narratives more prevalent in Northern newspapers, red bars indicate narratives more prevalent in Southern newspapers. Cities are classified relative to the Mason-Dixon line. Local newspapers only.	41
2.6	Differences in inflation narrative shares between newspapers headquartered in predominantly Democratic-leaning and Republican-leaning markets. Error bars represent 95% confidence intervals. Markets are classified by historical presidential voting patterns in three time-varying periods (1923–1960, 1960–1990, 1990–2025). Local newspapers only.	42
3.1	Opinion and syllabus lengths, compression rates by syllabuses, and correlations between opinion and syllabus lengths, 1815–2019. Dashed blue and orange lines give average compression rate and correlation. Lines are smoothed with 5-year moving-average.	49
3.2	ROUGE-2 evaluation of model-generated and human summaries, by Chief Justice of SCOTUS when the opinion was written. Markers are means and whiskers are 95% confidence intervals. The gap between precision and recall widens for opinions issued under more recent Chiefs.	57
3.3	Human evaluation of model-generated and human summaries. x-axis is a rank, where 1 is best and 5 is worst. For Error , x-axis shows counts of the total number of errors identified by participants for each summary method.	58
3.4	Lexical variation by summary source. Lexical variation measures the fraction of unique words in the summary that do not appear in the opinion. Higher values indicate more abstractive paraphrasing. Mistral FT exhibits the highest lexical variation, exceeding even non-legal Oyez summaries.	64

4.1	Vote distributions (Yes / No / Abstain) by condition for the Colorado fentanyl-penalties measure (left) and the Florida budget-stabilization measure (right). The baseline yes-rate differs sharply between measures: 70% in CO <code>Official_Text</code> versus 35% in FL <code>Official_Text</code>	84
4.2	Engagement outcomes by condition for the Colorado (top) and Florida (bottom) measures. Columns show active minutes, mean message length in words, and conversation turns. Active minutes for chat conditions exclude AI response time; message length and conversation turns are defined only for the three chat conditions.	87
4.3	Mean pro and con argument counts in the written rationale by condition. On the Colorado measure, <code>AI-Reflect</code> inverts the pro/con balance, shifting from the pro-leaning rationale produced under <code>Official_Text</code> to a strongly con-leaning one.	88
4.4	Counterargument engagement (Wachsmuth-style 0/1/2 rubric) by condition. Discourse modes raise engagement above <code>Official_Text</code> on both measures; the discourse-mode-versus-AI-QA gap appears only in FL.	89
4.5	Decomposition of conversation-rationale predictability, pooled across measures, as differences in mean per-token log-probability (nats) with 95% confidence intervals. <i>Top</i> : for each chat mode, the participant contribution (rationale predictable from the user’s own messages) and the chatbot contribution (additional predictability from the chatbot’s messages); the participant contribution is largest under <code>AI-Argue</code> and <code>AI-Reflect</code> , while the chatbot contribution is largest under <code>AI-QA</code> . <i>Bottom</i> : the participant contribution by chat mode across four scorer models, showing that the <code>AI-Argue</code> > <code>AI-Reflect</code> > <code>AI-QA</code> ordering reproduces.	90
4.6	Self-reported understanding and decision confidence (1–5 Likert scales) by condition, for the Colorado (top) and Florida (bottom) measures.	91
A.1	Example annotation for a narrative about the cause of inflation.	98
A.2	Example annotation for a narrative about the effect of inflation.	99
A.3	NOW test set: fine-tuned Llama 3.1 8B predictions against majority-vote human ground truth.	102
A.4	ProQuest test set: fine-tuned Llama 3.1 8B predictions against majority-vote human ground truth.	103
A.5	NOW test set: GPT-4o predictions against majority-vote human ground truth. .	104
A.6	ProQuest test set: GPT-4o predictions against majority-vote human ground truth.	105
B.1	Label Studio annotation interface.	119
C.1	Cross-measure forest plot of focal contrasts. Each row shows the per-measure effect size with 95% confidence interval and the pooled estimate from a measure-fixed-effect regression.	136

LIST OF TABLES

2.1	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 §A.2 for additional details.	10
2.2	Human annotation statistics	12
2.3	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 Contemporary).	16
2.4	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 Contemporary).	17
2.5	Comparison of fine-tuned Llama 3.1 8B and human annotations.	18
2.6	National average inflation expectations and narrative shares. Dependent variable is the standardized weighted average inflation expectation from the Michigan Survey of Consumers, for 1-year and 5–10 year horizons. T-statistics in brackets use Newey and West [1994] standard errors.	31
2.7	Standard deviation of inflation expectations and narrative shares. Dependent variable is the standardized cross-sectional standard deviation of individual inflation expectations from the Michigan Survey of Consumers. T-statistics in brackets use Newey and West [1994] standard errors.	33
2.8	Income-based inflation expectations gap (lowest minus highest income quartile) and narrative shares. T-statistics in brackets use Newey and West [1994] standard errors.	35
2.9	Education-based inflation expectations gap (non-college minus college graduates) and narrative shares. T-statistics in brackets use Newey and West [1994] standard errors.	37
3.1	Comparison of CASESUMM and related long-context summarization datasets in the legal domain. *GovReport does not report the range of years covered. **Multi-LexSum is a multi-document summarization dataset. ***Oyez summaries are a subset of SuperSCOTUS.	47

3.2	Automatic evaluation of model-generated and human-written summaries, where official syllabuses are the reference summaries. Sample includes 622 Supreme Court cases. There are 622 observations on each type of summary except Westlaw, for which we only have 156 observations. For each metric, we report precision (P), recall (R), and F1-score (F1). For each metric, we bold the best score(s) and <u>underline</u> the second best score(s).	54
3.3	BARTScores of model-generated and human-written summaries, where official syllabuses are the reference summaries. Sample includes 622 Supreme Court cases. There are 622 observations on each type of summary except Westlaw, for which we only have 156 observations. We report precision (P), recall (R), and F1-score (F1). BARTScores are negative log-likelihoods, so lower scores are better. We bold the best score(s) and <u>underline</u> the second best score(s). For the scoring model, we use <code>facebook/bart-large-cnn</code> , the default model used in Yuan et al. [2021]	55
3.4	Descriptive statistics on summaries in test set ($n=622$). Length is number of words. Compression rate is ratio of words in syllabus to words in opinion. Smaller number is more compression. Opinion included as reference. ($*p < 0.05$ $**p < 0.01$, $***p < 0.001$)	55
3.5	Spearman correlations between ROUGE and BERTScore and human rankings on the four quality dimensions. Interval scores are converted to rankings before correlation; human rankings for the same set of summaries are averaged.	59
3.6	Spearman correlations between default G-Eval criteria (Consistency, Relevance, Coherence, Fluency) and human rankings.	59
3.7	Spearman correlations between G-Eval prompts adapted to mirror the four human-evaluation criteria and human rankings.	60
3.8	G-Eval absolute scores under default criteria (Consistency, Relevance, Coherence, Fluency) and adapted criteria (Sensitivity, Specificity, Clarity, Style). Sample includes the 33 Supreme Court cases used for human evaluation. Best scores are bold , second-best are <u>underlined</u>	60
3.9	Comparison of model hallucinations and their explanations.	68
4.1	LLM-as-judge discourse-mode rubric pass rates	77
4.2	Deployed-bot response characteristics per discourse mode	78
4.3	Vote-choice distributions by condition, per study. Cells are within-condition percentages. Stars on yes/no/abstain cells mark a significant two-proportion z test against <code>Official_Text</code> in that vote category. The omnibus test in the ‘Overall’ row is a Pearson χ^2 of independence across the full 3×5 vote $\times$ condition table for each study.	84
A.1	Narrative categories, the label used in the classification task, and an excerpt of each definition.	96
A.2	Fine-tuning hyperparameters for Phi-2 and Llama 3.1 8B.	101
C.1	Sample and exclusions by study (Colorado, CO; Florida, FL).	135

C.2	Demographics and balance by study (Colorado, CO; Florida, FL). Pooled-sample summaries with chi-square balance tests across the five conditions.	135
C.3	Primary results, both studies	137
C.4	Robustness of focal contrasts	138

ACKNOWLEDGMENTS

Thank you to my advisor, Chenhao Tan, and my dissertation committee members Ari Holtzman and Alexander Zentefis, for providing guidance, mentorship, and support throughout my PhD.

Thank you to my wife, Sanghee Kim, whom I first knew as a labmate and friend when I started my PhD in 2021, and who has, since then, brought me joy, energy, and motivation.

Finally, thank you to my family and my friends, both new and old, for the constant encouragement, good conversation, and thoughtful advice.

ABSTRACT

Recent advancements in large language models have concentrated on tasks with automatic verifiers, such as code and mathematics. In many domains there is no such verifier, and what counts as a good output rests on tacit knowledge or judgment that resists measurement. This dissertation develops abstractions that make such judgment measurable, then uses them to build and evaluate language models across open-ended domains.

To measure economic narratives, this dissertation defines *causal micro-narratives* as sentence-level explanations of a cause or effect and builds an inflation-specific ontology, then fine-tunes a model to detect them at scale. Across a century of U.S. newspaper coverage (4.2 million sentences), social and political impact narratives predict disagreement in household inflation expectations 1.8 times more strongly than realized inflation, and lower-income households are 4.6 times more sensitive to them.

To evaluate summaries of complex legal documents, it introduces CASESUMM, a long-context summarization dataset of U.S. Supreme Court opinions paired with their official syllabuses, and compares automatic metrics, LLM judges, and expert rankings. LLM-based judges do not track expert judgment better than automatic metrics, and fluent summaries that experts prefer can contain factual errors that automatic methods miss, exposing the limits of LLM-as-judge for long-context summarization.

To support voters deliberating over ballot measures, it presents CivicChats, a platform and chatbot interface with three discourse modes (neutral Q&A, argumentative, and reflective) grounded in deliberative democracy. In a pre-registered controlled experiment on two live 2026 ballot measures, chatbots designed for deliberative discourse shifted vote distributions and produced written rationales that engaged substantially more with counterarguments than static information.

Dissertation Advisor: Chenhao Tan

CHAPTER 1

INTRODUCTION

Recent advancements in large language models (LLMs) have concentrated on tasks with automatic verifiers, which provide simple feedback signals that drive rapid progress through reinforcement learning and large-scale evaluation. Many important domains admit no such verifier, where good performance depends on context and judgment. Summarizing a legal opinion, helping a voter deliberate over a ballot measure, and identifying which narratives a text conveys are challenging tasks to measure and automate. Multiple outputs may be appropriate and the criteria for correctness often rest on tacit knowledge that is hard to operationalize. In this dissertation, I develop abstractions that enable measuring important phenomena in open-ended domains, then I engineer and evaluate LLM-based systems against these abstractions, demonstrating their utility at scale and where current models fall short.

Chapter 2 measures economic narratives in news coverage, defining a unit of analysis and the domain ontology that grounds it for one application. Part I introduces *causal micro-narratives*, sentence-level explanations of the cause(s) or effect(s) of a target subject, and presents a classification framework demonstrated through inflation narratives. Using human-annotated data spanning historical and contemporary US news articles, a fine-tuned Llama 3.1 8B achieves F1 scores of 0.87 on narrative detection and 0.71 on classification, substantially outperforming GPT-4o in few-shot settings. Part II applies this framework to a century of American newspaper coverage (4.2 million sentences, 1923–2025), examining how inflation narratives evolved across economic eras and how narrative frames predict heterogeneity in household inflation expectations. Social and political impact narratives predict cross-sectional dispersion in 1-year inflation expectations 1.8 times more strongly than realized inflation itself, and lower-income and non-college households respond several times more strongly to narrative framing than higher-income and college-educated households.

The same demand for an explicit specification arises in evaluation. Large language mod-

els are increasingly applied to long-context summarization in high-stakes domains. Studying their performance in such settings requires high-quality reference summaries written by domain experts and evaluation methods that track expert judgments of summary quality. Chapter 3 contributes CASESUMM, a new dataset of 25.6K U.S. Supreme Court (SCOTUS) opinions paired with their official summaries, known as “syllabuses,” along with a use-case study examining how automatic, LLM-based, and expert human evaluations differ in this setting. CASESUMM is the largest open legal case summarization dataset and is the first to include syllabuses dating back to 1815. A comprehensive evaluation using both automatic metrics and expert human evaluation reveals discrepancies between these assessment methods: summaries that score well on automatic metrics can contain hallucinations that experts catch, including precedent citation errors and misrepresentations of case facts, and LLM-based evaluations are no more correlated with expert judgment than traditional automatic metrics. These findings demonstrate the limitations of automatic evaluation for legal summarization and the critical role of human evaluation in complex, high-stakes domains.

The final chapter explores how interactive AI can be designed to help voters reason about ballot measures and develops a methodology for measuring how it shapes their decisions and reasoning. In the United States, these measures decide consequential policy questions on housing, criminal justice, drug policy, and state budgets, yet voters often first encounter them in the booth and frequently struggle with their dense, technical language. As a result, voters fail to develop well-founded opinions or abstain. Conversational AI could help voters reason through these decisions, but mainstream chatbots tend to match user beliefs rather than clarify or challenge them. Chapter 4 presents CivicChats, a platform that informs voters and promotes interaction with ballot measures through chatbots tailored to three discourse modes (Q&A, argumentative, and reflective), and evaluates it in a pre-registered controlled experiment with 600 participants on live 2026 ballot measures in Colorado and Florida. Interaction with the chatbots changed vote outcomes substantially, shifting each

measure’s vote distribution away from its low-engagement default, and improved participants’ ability to explain their vote, raising the number of arguments and their engagement with counterarguments. Participants also reported higher understanding and, contrary to the pre-registered prediction, higher decision confidence. The findings indicate that the conversational posture given to a civic chatbot, more than the decision to deploy one, shapes whether it improves how voters reason and decide.

Together, the three chapters show that working with large language models in these domains depends on first making implicit judgment and criteria explicit, enabling scalable and meaningful measurement of language processing and generation.

CHAPTER 2

NARRATIVES AND INFLATION EXPECTATIONS

Part I: Causal Micro-Narratives

2.1 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 (such as 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

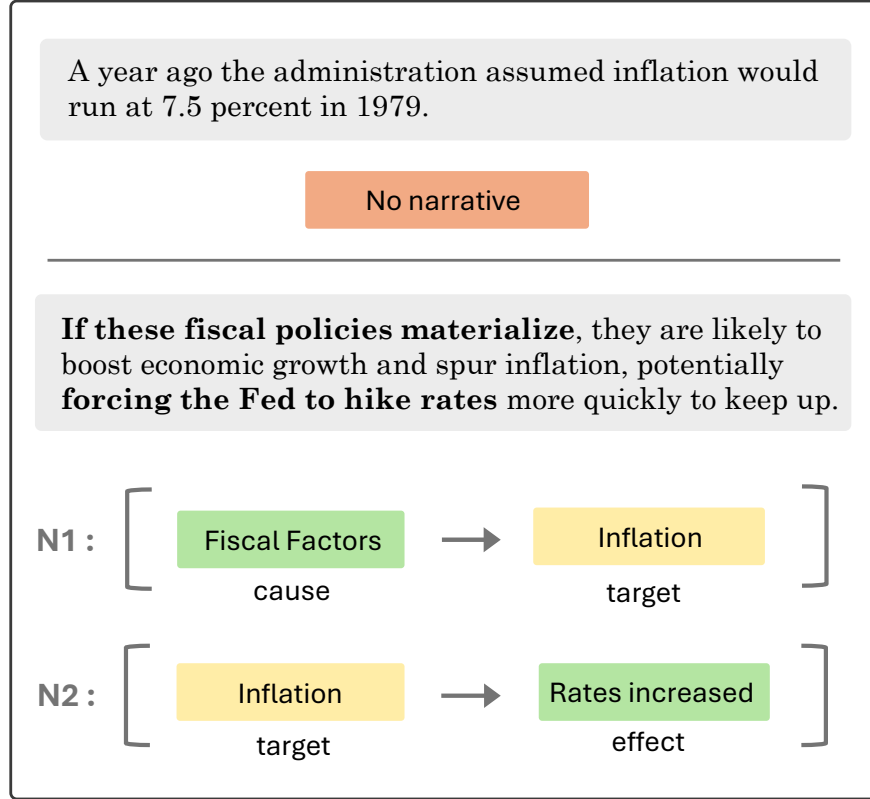


Figure 2.1: Causal micro-narrative classification task examples for the *target* ‘inflation.’ In the first sentence, no narratives are identified; in the second, two narratives (N1 and N2) are identified, one representing a cause of the *target* and the other representing an effect of it.

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 chapter aims to address both these conceptual and technical challenges. It introduces the concept of *causal micro-narratives*, along with a multi-label classification task to extract them from text. *Causal micro-narratives* are defined 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, *inflation* is chosen as the target that centers the micro-narratives examined. 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 2.1 illustrates how the 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. These two micro-narratives are labeled *fiscal factors* and *rates increased*, respectively.

An ontology of causes and effects of inflation is proposed, and a large scale dataset of causal micro-narratives is created according to this ontology, classifying sentences from contemporary and historical U.S. news articles. Starting with a subset of human annotations, various models are trained 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. Comparison of different models reveals 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 the dataset and the performance of the classifiers, an in-depth error-analysis of inter-annotator disagreements and the in- and out-of-domain generalization of each evaluated model is conducted. Furthermore, systematic classification errors are identified and cross-referenced with annotator disagreements. 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 the 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, the following contributions are made:

1. The concept of causal micro-narratives is introduced and defined, presenting a novel task for extracting them from real-world text.
2. A dataset of annotated inflation-related causal micro-narratives from both historical and contemporary U.S. news articles is curated.
3. Methods for effectively automating narrative classification at scale are developed and demonstrated, making publicly available fine-tuned LLMs for this purpose. Additionally, robust out-of-domain performance of these models is showcased.
4. A comprehensive error analysis is conducted, revealing systematic similarities between model classifications and human annotation disagreements. This analysis highlights the task’s complexity and identifies potential inherent ambiguities.

2.2 Related Work

2.2.1 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 interpretive 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.

2.2.2 *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 includes 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.

2.3 Causal Micro-Narratives

A *causal micro-narrative* is defined 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 storytelling that depicts sequences of events, usually in long-form texts [e.g., Piper, 2023]. By contrast, here the focus is 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].

2.3.1 Narrative Classification Task

A narrative classification task is proposed that operationalizes the definition of *causal micro-narratives*. Unlike the more general task of causality mining [Ali et al., 2021], a productive approach to capturing how such micro-narratives accumulate at scale should be domain-specific. Specifically, a framework is proposed in which first a *target* is identified about which micro-narratives are to be captured. Conceptually a target can be any entity, event, or phenomenon of interest.

Then, an ontology of the causes that can lead to that target and the effects that can follow from it is defined. 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.

2.3.2 Case Study: Inflation Narratives

As an application of this definition and for the purposes of this chapter, the focus is specifically on *inflation* as the target. An ontology is developed, presented in Table 2.1, 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 §A.2 for additional details on this process and detailed descriptions of all the causes and effects. Ultimately, the following classification task is set up: 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, a target event is chosen 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 preliminary event extraction step.

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 2.1: 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 §A.2 for additional details.

2.4 Dataset

Two data sources are used in the investigation of inflation narratives in news: NOW Corpus for contemporary news data [Davies, 2016] and ProQuest for historical data. These datasets were selected because their differences allow assessment of the generalizability of the task and the classification methods tested. 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, articles are segmented into sentences and sentences that contain the keyword “inflation” are filtered. Filtering allows focus on relevant sentences, enabling efficient targeting of human annotations, as well as reducing the total number of sentences to a more computationally feasible quantity.

2.4.1 *Contemporary News: NOW Corpus*

Data from the NOW Corpus covering 2012-2023 is used. The dataset consists of online news articles, which are filtered 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. The spaCy Sentencizer [Explosion, 2024] is used for sentence segmentation.

2.4.2 *Historical News: ProQuest*

For historical news data, news articles from local, regional, and national news publications from the ProQuest database spanning 1960-1980 are collected. See §A.1 for a list of the included publications. This historical period was chosen 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. The BlingFire [Microsoft, 2024] sentence segmentation tool was

used, as the spaCy Sentencizer did not work well on this historical data.

2.4.3 Human Labeling

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

(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 2.2: Human annotation statistics

Three members of the research team manually annotated training and test sets. In Table 2.2a the sizes of the train and test splits are reported. Train sets of approximately 1,000 examples were targeted. This provided 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 2.2b shows a moderate to high degree of agreement for a pragmatic annotation task, across both the historical and contemporary news annotations. Historical news agreement is higher than contemporary news, hypothesized to be 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. An analysis of annotator disagreement is presented in §2.6.3. See §A.3 for annotation interface examples.

2.4.4 Descriptive Statistics

The focus on *causal micro-narratives* ensures distinction 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 the definition and annotation scheme.

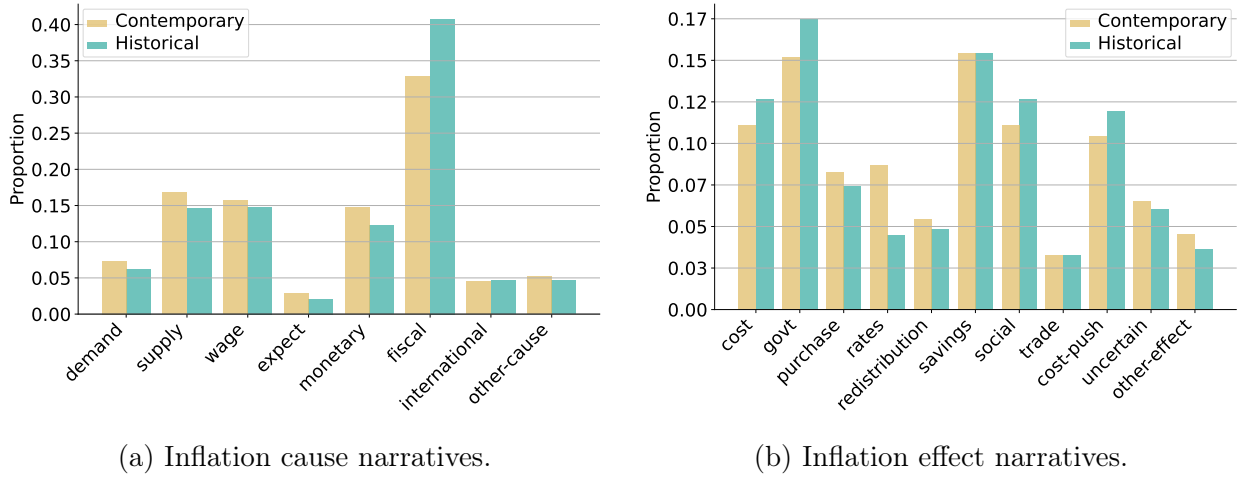


Figure 2.2: Proportions of narrative classes in human annotations. This data combines both the train and test sets. For the test set, majority vote is used to identify one annotation instance.

The distribution and prevalence of cause and effect narratives remains largely consistent across human annotations of both datasets. As Figure 2.2 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 a more common topic of discussion.

2.5 Methods

To determine the most effective approach to classify narratives, the performance of LLMs on the classification task is compared for both in-context learning and fine-tuning settings. The annotations associated with each sentence are formatted as JSON to facilitate automatic processing (see §A.2). The LLMs are evaluated on their classification output, expected to be in JSON as well. Separate experiments are conducted with the contemporary and historical data and separate models are trained for each dataset.

2.5.1 *In-Context Learning*

LLMs have been shown to be effective in-context, or *few-shot*, learners [Brown et al., 2020], so GPT-4o is tested 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. Greedy decoding is used and generation is not constrained in any way, but GPT-4o reliably generated JSON in the correct format.

2.5.2 *Fine-tuning*

The second modeling approach evaluated 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`). These two models were chosen 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 this 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. The full fine-tuning prompt is included in §A.4. Standard auto-regressive language modeling is followed but only the language modeling loss for tokens associated with binary and multi-class labels is back propagated, rather than other tokens associated with the JSON notation. LoRA-based Parameter-Efficient Fine-Tuning (PEFT) [Hu et al., 2021] is used to train a subset of the parameters. See §A.5 for fine-tuning hyper-parameters.

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, the focus is on evaluating fine-tuned versions of the two open-source models, rather than their zero-shot performance.

2.5.3 Evaluation

Each aspect of a narrative classification is evaluated separately using micro-averaged F1 scores. Micro averaging is used, 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, majority rule is used to identify gold-labels. In practice, 97% of the test set instances have agreement between at least two annotators, allowing retention of almost the entire test set for evaluation.

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

Table 2.3: 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 Contemporary).

2.6 Results

Model performance is compared in Table 2.3. 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, Table 2.4 presents a breakdown of model performance in several training and evaluation settings. First, how well models fine-tuned on Historical and Contemporary data perform on corresponding held-out data is evaluated, assessing in-domain generalization. Second, how well models generalize to out-of-distribution (OOD) data is compared by evaluating performance on Historical data when trained on Contemporary data, and vice-versa. Finally, both the historical and contemporary data are combined during the learning phase and performance is evaluated on the individual datasets, revealing how well models can learn from the additional data despite the domain-shift.

		Llama3.1 8B		Phi-2		GPT-4o	
Train \ Test		Hist.	Contemp.	Hist.	Contemp.	Hist.	Contemp.
<i>Binary</i>							
	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
<i>Multiclass</i>							
	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 2.4: 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 Contemporary).

2.6.1 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.

2.6.2 Out-of-Domain Generalization

On the multiclass narrative classification task, a common pattern emerges across both fine-tuned models. Test set performance degrades by 3-4% on OOD data relative to in-domain

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 2.5: Comparison of fine-tuned Llama 3.1 8B and human annotations.

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 §2.4.4 and Figure 2.2. 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.

2.6.3 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 proprietary LLM, a fine-grain analysis of the individual narrative classification predictions as well as an analysis of the three sets of human annotations is conducted to better understand the disagreements that exist between them and how those disagreements may relate to model prediction errors. As the best performing LLM overall, the focus is on Llama 3.1 8B (henceforth, *Llama*) and it is compared to GPT 4o, the only proprietary model in these experiments.

Human Annotator Disagreements By majority rule, the 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 training data sourced from each annotator individually is used, 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 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, GPT-4o consistently struggled to understand the distinction stipulated between narratives and non-narratives. This can likely be attributed to the 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 2.5 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 2.5 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 inherent 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.

2.7 Part I Summary

This part 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, the feasibility of automating the detection and categorization of these narratives at scale is demonstrated. The 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 identifying causal micro-narratives is inherently complex, with natural ambiguities in language and variation in human interpretations. Despite these challenges, this approach provides a foundation for large-scale narrative analysis. The following part applies this framework to study the evolution of inflation narratives across a century of American newspaper coverage and examines whether narrative frames predict heterogeneity in household inflation expectations.

Part II: A Century of Inflation Narratives

2.8 Motivation and Research Questions

Part I establishes a framework for extracting causal micro-narratives from text. This part applies that framework to study how inflation narratives have evolved in American newspapers over the past century and whether those narratives predict heterogeneity in household inflation expectations.

Two dimensions of inflation narratives remain underexplored in the literature. First, how have inflation narratives evolved across time and space throughout American economic history? The evolution of these narratives is informative for interpreting how economic ideas develop and spread through public discourse across different economic regimes. Second, do

specific narrative frames predict heterogeneity in inflation expectations across demographic groups? This relationship matters for monetary policy effectiveness, since communication strategies shape inflation expectations and spending behavior.

This part addresses both questions by applying the classification approach from Part I to a large-scale corpus of American newspapers spanning 1923–2025. The analysis examines how the prevalence of different narrative types changed across economic periods, how inflation is explained differently across regions and political geographies, and whether narrative frames predict differences in inflation expectations across income and education groups.

2.9 Data

2.9.1 *Newspaper Corpus*

The analysis draws on ProQuest’s newspaper database, which provides access to U.S. newspaper articles. The sample ranges from January 1923 to January 2025, covering all 50 states plus DC and Guam. The corpus includes 807 unique newspapers, from major national publications such as The New York Times, Wall Street Journal, and Washington Post (46 in total), to hundreds of local and regional newspapers (761 in total). Approximately 47% of the total number of inflation sentences originate from national sources.

Over the full sample period, the dataset averages 81.6 newspapers per month, with 8.6 national newspapers and 73.0 local newspapers per month on average. Coverage varies substantially over time, ranging from a minimum of 13 newspapers in early months to a maximum of 324 newspapers during peak coverage periods.

Following the preprocessing approach from Part I, all articles are segmented into sentences and filtered for those containing the keyword “inflation.” This initial filtering yields a corpus of approximately 4.2 million sentences spanning a century of economic reporting.

2.9.2 Economic Variables

For inflation expectations, the analysis uses the University of Michigan Survey of Consumers, which records household expectations about future price changes for the headline 1-year horizon and the longer 5–10 year horizon. For realized inflation, Consumer Price Index (CPI) data from the U.S. Bureau of Labor Statistics is used, retrieved through the Federal Reserve Economic Data (FRED) system. Regional CPI series for the Northeast, Midwest, South, and West regions support examination of geographic patterns in both inflation narratives and actual price changes.

2.10 Classification at Scale

To classify inflation narratives across the full corpus, the fine-tuned Llama 3.1 8B model from Part I is applied. The model operates in two sequential stages, first determining whether a sentence contains any causal narrative about inflation, then classifying identified narratives into one or more of the 19 narrative categories (8 causes and 11 effects).

2.10.1 Cross-Temporal Validation

The classifier is trained on annotations spanning two eras roughly 50 years apart (Part I), and additional validation is conducted by reviewing random samples of classified sentences across different time periods. Cross-temporal evaluation shows performance degradation of only 3–4 percentage points when training on one era and testing on another. This stability supports applying a single fine-tuned model to the full century-long sample.

2.10.2 Comparison with Keyword Methods

To evaluate the narrative classification against simpler approaches, a comparison with keyword-based classification is conducted. Keywords are chosen to reflect the most prominent words

associated with each narrative category. Because the vectors of labeled sentences are sparse under both classifications, Jaccard similarity is computed in place of Pearson correlation.

Similarity between the two approaches varies substantially across categories. Interest rate narratives show the highest similarity with keywords (0.45), while international trade narratives show the lowest (0.06). This range reflects the fact that some inflation discussions use consistent terminology while others employ diverse language that keywords cannot capture. The “No narrative” category shares similarity with multiple keyword classifications, exposing a recurring failure mode for keywords. They classify sentences that mention inflation-related terms without expressing any causal relationship. The narrative approach distinguishes mere mentions from actual explanations.

2.10.3 Unsupervised Validation via TopicGPT

A complementary check is whether the manually curated 19-category ontology omits causal mechanisms that emerge organically from the corpus. To address this, a data-driven narrative discovery procedure based on TopicGPT [Pham et al., 2024] is run on a random sample of 10,000 inflation sentences. In the generation phase, GPT-4o produces a separate causal ontology for each sampled sentence. In the refinement phase, batches of 200 candidate narratives are iteratively consolidated until fewer than 200 remain, ending in a final round that yields 11 distinct causal narratives.

Figure 2.3 shows category-level semantic similarity between the 19 manual categories and the 11 discovered categories. The strongest correspondences include wage with “Labor Market Pressures,” demand with “Demand-Supply Imbalance,” and monetary with “Monetary and Fiscal Drivers.” All 11 discovered categories map to existing entries in the manual ontology. No novel causal mechanisms emerge. The data-driven categories generally consolidate finer manual distinctions into broader thematic clusters (for example, “Rising Living Costs and Income Inequality” draws from both cost and redistribution). This finding supports

proceeding with the more granular manual ontology in the analyses that follow.

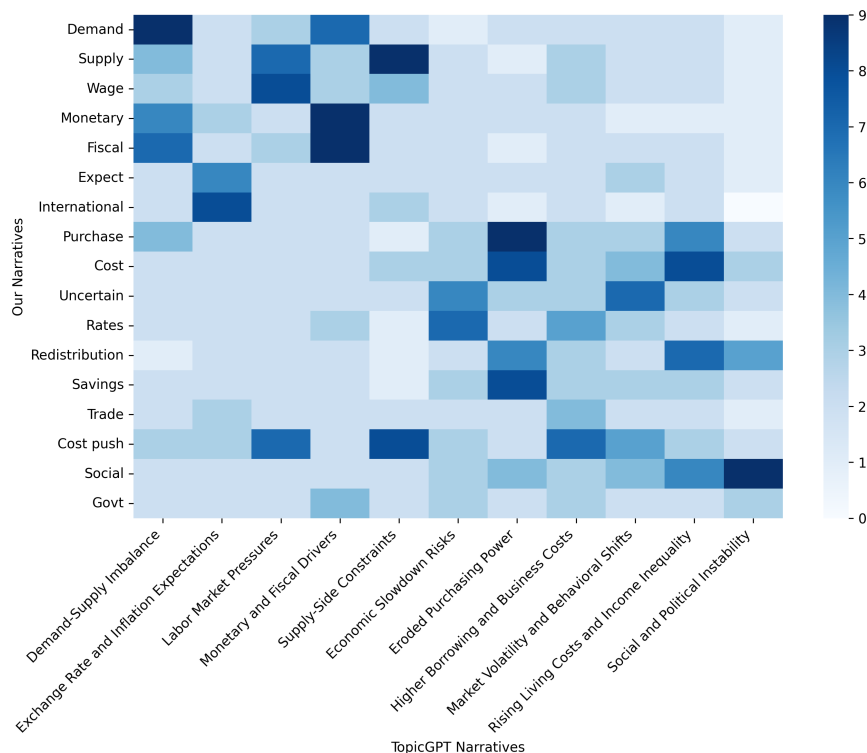


Figure 2.3: Category-level semantic similarity between the 19 manual narrative categories and the 11 categories discovered by TopicGPT on 10,000 randomly sampled inflation sentences. Darker shades indicate stronger similarity. All discovered categories map onto existing entries in the manual ontology, and no novel causal mechanisms emerge from the unsupervised procedure.

2.11 Inflation Narratives Through Time and Space

2.11.1 Narrative Prevalence

About 54.5% of inflation-related sentences in the corpus do not express any specific narrative about inflation’s causes or effects. The majority of media coverage reports on inflation without exploring its underlying mechanisms or consequences.

National publications emphasize certain causal narratives more strongly than local papers, primarily fiscal (9.3% vs. 7.8%), monetary (5.2% vs. 3.9%), and international factors

(2.2% vs. 1.3%). National newspapers dedicate substantially more coverage to causal narratives overall (26.6% vs. 21.4% in local papers). Local newspapers give relatively greater attention to direct effects of inflation, including impacts on savings (6.4% vs. 5.6%), cost of living (3.9% vs. 2.8%), and purchasing power (1.7% vs. 1.1%).

2.11.2 Narratives Through Time

The time series in Figure 2.4 reveals substantial variation in narrative prominence. Fiscal causes of inflation dominated pre-1980 discussions, with pronounced peaks during the Great Depression, World War II, and the inflationary episode of the 1970s, when fiscal narratives reached nearly 25% of inflation sentences. Since the 1980s, fiscal explanations have declined in prominence, with monetary and supply-side factors becoming relatively more important in the inflation discourse.

The pattern of effect narratives shows a comparable shift. Savings effects consistently appeared throughout the century, while interest rate effects became substantially more prominent after 1970. The timing coincides with the Federal Reserve’s more aggressive monetary policy stance under Volcker and the subsequent era of inflation targeting. Before 1970, interest rate effects were nearly absent from inflation discussions, marking a fundamental change in how the media expresses inflation’s implications. During the post-COVID inflation episode of 2021–2022, a pronounced spike in cost-of-living effect narratives is observed, reflecting heightened public concern about inflation’s direct impact on household expenses.

2.11.3 Regional Differences

To investigate how inflation narratives differ across regions, the analysis examines geographic location (North vs. South), level of urbanization (urban vs. rural), and political lean of the newspaper’s market (Republican-leaning vs. Democrat-leaning). Each newspaper is assigned to these categories based on the geographic position of its headquarters relative to the Mason-

Dixon line, the population density of its location, and the historical voting patterns of its area, respectively. To capture regional variation rather than national editorial trends, national newspapers are excluded and the analysis focuses on local publications.

North vs. South. Figure 2.5 shows differences in cause and effect narratives across Northern and Southern newspapers. Northern newspapers attribute inflation to fiscal policy at a higher rate, with approximately a 0.92 percentage point higher share of fiscal narratives compared to Southern newspapers. International and inflation expectations narratives also feature somewhat more prominently in Northern publications. Southern newspapers emphasize supply factors as causes of inflation (0.26 percentage points higher) and place modestly greater weight on monetary, demand, and wage explanations. For effect narratives, the most notable difference is in the discussion of interest rates, which Southern newspapers emphasize substantially more (1.5 percentage points higher) than Northern counterparts. Northern newspapers focus more on government program effects (0.48 percentage points higher) and savings effects (0.33 percentage points higher).

Urban vs. rural. Urban newspapers emphasize fiscal factors when discussing causes of inflation, with a 3.4 percentage point higher share than rural publications. International factors, monetary policy, trade effects, and wage pressures also receive more attention in urban newspapers. Rural newspapers do not show significantly higher emphasis on any causal narrative category. In effect narratives, a sharp contrast appears in the discussion of interest rates and costs of living, which rural newspapers emphasize more heavily (3.3 and 2.0 percentage points higher, respectively). Urban newspapers focus more on savings effects (2.4 percentage points higher) and modestly more on social and trade effects. These differences likely reflect the economic makeup of urban and rural areas, with urban areas more connected to global markets and government policy, and rural areas experiencing more direct impacts of inflation on borrowing costs and consumer prices.

Republican vs. Democrat. Partisan differences in economic narratives have been documented in political communication research [Gentzkow and Shapiro, 2010, DellaVigna and Kaplan, 2007]. To examine whether these patterns extend to inflation coverage, newspapers are classified by the political lean of their geographic markets using county-level presidential election data [Leip, 2023]. Presidential vote shares for all U.S. counties from 1920 to 2024 are obtained, the sample period is divided into three eras (1923–1960, 1960–1990, and 1990–2025), and counties are classified as Democrat-leaning if the average Democratic vote share within the era exceeds 50% and Republican-leaning otherwise. This time-varying approach accounts for the realignment of partisan geographies over the century, particularly the reversal of the South’s political orientation.

Figure 2.6 reveals clear differences. Newspapers from predominantly Democrat-voting areas attribute inflation more to international factors (0.36 percentage points higher), monetary policy (0.33), demand pressures (0.25), and supply constraints (0.23). Newspapers from predominantly Republican-voting areas show stronger emphasis on fiscal policy as a cause (0.34 percentage points higher), along with somewhat higher attribution to wage pressures. For effect narratives, publications from Democrat-voting areas focus more on savings effects (1.5 percentage points higher) and interest rate effects (0.4), as well as uncertainty (0.2). Newspapers from Republican-voting areas emphasize cost impacts (0.7 percentage points higher) and show moderately higher focus on purchasing power, redistribution, and social effects. These patterns align with traditional partisan economic perspectives.

2.12 Narratives and Household Expectations

Understanding how consumers form inflation expectations is a central question in macroeconomics with implications for monetary policy effectiveness, household financial decisions, and overall economic stability. Prior work has emphasized the role of past inflation [Gaspar et al., 2010], exposure to consumption-bundle prices [D’Acunto et al., 2021], and central

bank communication [Coibion et al., 2022]. Beyond these channels, the news media, and the narrative frames through which inflation is discussed, may materially shape public perceptions and expectations [Carroll, 2003, Lamla and Lein, 2014]. This section examines how different inflation narratives predict consumer inflation expectations, with attention to how media coverage translates economic data into cause-and-effect stories that resonate differently across demographic groups.

2.12.1 Empirical Design

The analysis uses a uniform regression framework. Specifications vary by demographic group or expectation measure while maintaining a consistent empirical approach. The baseline model is

$$E_t[\pi_{t+h}] = \alpha + \sum_{j=1}^p \beta_j E_{t-j}[\pi_{t-j+h}] + \gamma \pi_{t-1} + \delta \log(N_{t-1}) + \sum_k \eta_k S_{k,t-1} + \varepsilon_t, \quad (2.1)$$

where $E_t[\pi_{t+h}]$ denotes inflation expectations at time t for horizon h (1-year or 5–10 years), π_{t-1} is lagged realized inflation, $\log(N_{t-1})$ is the log number of sentences containing “inflation” (a proxy for inflation news volume), and $S_{k,t-1}$ is the share of inflation sentences expressing narrative k . The number of autoregressive lags p is chosen to minimize AIC for each specification. Three specifications are reported for each horizon: (1) realized inflation with optimal autoregressive lags, (2) adding the log number of inflation sentences, and (3) results from separate regressions that include a single narrative share at a time. “Other-cause,” “other-effect,” and “no-narrative” shares are excluded for interpretability. All regressions use monthly standardized variables, with t-statistics calculated using Newey and West [1994] standard errors and automatically selected lag truncation.

2.12.2 *National Expectations*

Table 2.6 reports baseline results for national average inflation expectations from the Michigan Survey of Consumers. For 1-year expectations, realized inflation has a coefficient of 0.116 (t-statistic 2.92), and the relationship persists when controlling for inflation news volume (0.119, t-statistic 3.00). Long-run expectations exhibit a substantially weaker relationship with realized inflation (0.028, t-statistic 1.77), consistent with long-run expectations being more firmly anchored.

The volume of inflation-related news has minimal predictive power for national average expectations after controlling for realized inflation, with coefficients indistinguishable from zero at both horizons. Specific narrative frames show predictive power beyond realized inflation and news volume, though at the national-average level they fall short of conventional 5% significance. For 1-year expectations, the social narrative is the most influential predictor (coefficient 0.034, t-statistic 1.82), approximately 29% of the magnitude of the realized inflation effect. The wage narrative (0.031, t-statistic 1.67) and trade narrative (0.027, t-statistic 1.73) show comparable associations. For long-run expectations, the redistribution narrative is the strongest predictor (0.037, t-statistic 1.73), about 1.32 times the realized inflation effect at that horizon, while the expectations narrative carries a negative coefficient (-0.025 , t-statistic -1.87), suggesting that media framing of expectations themselves may operate differently across horizons.

2.12.3 *Dispersion in Expectations*

Table 2.7 reports the analogous regressions with the standardized cross-sectional standard deviation of individual inflation expectations as the dependent variable. For 1-year expectations, realized inflation has a coefficient of 0.064 (t-statistic 2.06). For 5–10 year expectations the coefficient is 0.038 (t-statistic 2.16). The volume of inflation news shows minimal association with dispersion.

Table 2.6: National average inflation expectations and narrative shares. Dependent variable is the standardized weighted average inflation expectation from the Michigan Survey of Consumers, for 1-year and 5–10 year horizons. T-statistics in brackets use Newey and West [1994] standard errors.

Dep. Var.: Avg. Inflation Expectations	Horizon: 1 year			Horizon: 5-10 years		
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline						
L. YoY Realized Inflation	0.116 [2.92]	0.119 [3.00]	0.114 to 0.124 [2.83 to 3.15]	0.028 [1.77]	0.022 [0.99]	0.007 to 0.027 [0.23 to 1.22]
L. Log(No. of Infl. Sentences)		0.000 [0.01]	-0.008 to 0.007 [-0.51 to 0.42]		0.017 [0.69]	0.012 to 0.022 [0.52 to 0.84]
Cause Narratives						
Wage			0.031 [1.67]			-0.016 [-1.09]
Expect			0.025 [1.66]			-0.025 [-1.87]
Demand			0.014 [1.31]			-0.012 [-0.67]
Fiscal			0.028 [1.23]			0.029 [1.38]
Supply			0.014 [0.90]			0.007 [0.37]
International			0.009 [0.70]			-0.025 [-1.20]
Monetary			0.007 [0.53]			-0.004 [-0.24]
Effect Narratives						
Social			0.034 [1.82]			0.032 [0.81]
Trade			0.027 [1.73]			0.010 [0.52]
Uncertain			0.017 [1.16]			0.001 [0.04]
Savings			-0.013 [-1.13]			-0.029 [-1.19]
Govt			0.023 [1.03]			0.017 [1.07]
Purchase			-0.015 [-1.02]			-0.006 [-0.34]
Cost-Push			-0.013 [-0.62]			0.014 [0.54]
Redistribution			0.006 [0.40]			0.037 [1.73]
Rates			-0.004 [-0.27]			-0.003 [-0.18]
Cost			0.003 [0.14]			0.020 [0.56]
Observations	558	557	557	416	415	415
Adj. R^2	0.891	0.891	0.891 to 0.892	0.867	0.865	0.865 to 0.866

The dependent variables are standardized weighted average one-year ahead and 5-10 year ahead inflation expectations from the Michigan Survey of Consumers. For each model, we include the optimal number of autoregressive lags that minimizes AIC. Columns (1) and (4) regress expectations on lagged year-over-year realized inflation and the optimal number of autoregressive lags. Columns (2) and (5) add the lagged log number of inflation sentences to the model (i.e., newspaper sentences containing the word “inflation”). Columns (3) and (6) report results from separate regressions that each include the optimal number of autoregressive lags, lagged realized inflation, the lagged log number of inflation sentences, and one narrative at a time. For lagged realized inflation and log number of inflation sentences, columns (3) and (6) show the range of coefficients and t-statistics [in brackets] across the narrative regressions. For narratives, they show the coefficient and t-statistic from the regression including that narrative. For columns (3) and (6), the adjusted R-squared values shown represent the range across all narrative regressions. Realized Inflation is the standardized year-over-year percent change in the Consumer Price Index for All Urban Consumers (All Items, U.S. City Average) from the Bureau of Labor Statistics. Narrative measures are lagged standardized monthly shares of inflation sentences from ProQuest newspaper articles that we classify as expressing that particular narrative about inflation. Sample period is September 1978 to February 2025 for one-year expectations and July 1990 to February 2025 for 5-10 year expectations. T-statistics shown in brackets are calculated using Newey and West [1994] standard errors, optimally selected using the rule-of-thumb bandwidth estimator, computed to be 5 lags. All specifications include a constant.

Specific narrative frames show dramatically stronger associations with expectation dispersion than realized inflation. Social narratives strongly predict increased dispersion at both horizons. For 1-year expectations, the social narrative coefficient (0.115) is approximately 1.8 times larger than the realized inflation coefficient (0.064), with a markedly higher t-statistic (5.36 vs. 2.06). For 5–10 year expectations, the social narrative coefficient (0.084) is about 2.2 times the realized inflation coefficient (0.038). Cost narratives also show substantial predictive power for dispersion, with coefficients of 0.069 (t-statistic 2.85) for 1-year and 0.064 (t-statistic 2.67) for 5–10 year horizons, approximately 1.1 and 1.7 times the corresponding realized inflation effects. Cost-push, fiscal, and supply narratives also predict increased disagreement, though with smaller magnitudes. The savings narrative is associated with reduced dispersion in long-run expectations (-0.036 , t-statistic -2.01), suggesting that coverage emphasizing savings behaviors may build consensus about distant future inflation. The economic significance of these narrative effects on expectation dispersion substantially exceeds that of realized inflation itself.

2.12.4 Income-Based Heterogeneity

Inflation expectations vary considerably with income, and so do the associations between expectations and media narratives. Lower-income households (1st quartile) show stronger responsiveness to inflation narratives than higher-income households (4th quartile). For 1-year expectations, the lowest-income households react strongly to social narratives (coefficient 0.130, t-statistic 3.32), expectations narratives (0.111, t-statistic 4.10), cost narratives (0.080, t-statistic 2.08), and uncertainty narratives (0.074, t-statistic 2.39). The highest-income households show much more muted responses, with their strongest reactions to trade narratives (0.032) and social narratives (0.028). The 4.6-fold ratio between Q1 and Q4 social-narrative coefficients indicates that media framing operates very differently across the income distribution, plausibly reflecting greater information frictions among lower-income

Table 2.7: Standard deviation of inflation expectations and narrative shares. Dependent variable is the standardized cross-sectional standard deviation of individual inflation expectations from the Michigan Survey of Consumers. T-statistics in brackets use Newey and West [1994] standard errors.

Dep. Var.: Std. Dev. Inflation Expectations	Horizon: 1 year			Horizon: 5-10 years		
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline						
L. YoY Realized Inflation	0.064 [2.06]	0.068 [2.37]	0.053 to 0.071 [1.99 to 2.55]	0.038 [2.16]	0.038 [1.85]	0.001 to 0.044 [0.05 to 2.22]
L. Log(No. of Infl. Sentences)		0.016 [1.00]	0.000 to 0.026 [0.00 to 1.54]		0.008 [0.57]	-0.007 to 0.017 [-0.62 to 1.15]
Cause Narratives						
Supply			0.030 [2.66]			0.018 [1.47]
Fiscal			0.052 [2.13]			0.018 [1.23]
Monetary			-0.021 [-1.37]			0.005 [0.33]
Demand			-0.018 [-1.32]			-0.013 [-1.00]
Wage			-0.019 [-1.27]			-0.009 [-0.93]
Expect			0.013 [1.10]			-0.000 [-0.02]
International			0.008 [0.67]			-0.006 [-0.42]
Effect Narratives						
Social			0.115 [5.36]			0.084 [3.86]
Cost			0.069 [2.85]			0.064 [2.67]
Cost-Push			0.041 [2.42]			0.024 [1.33]
Savings			-0.015 [-0.95]			-0.036 [-2.01]
Govt			-0.015 [-0.84]			-0.016 [-1.00]
Uncertain			0.008 [0.61]			-0.028 [-1.49]
Redistribution			-0.007 [-0.40]			0.005 [0.29]
Trade			0.006 [0.40]			-0.012 [-0.93]
Rates			-0.006 [-0.34]			-0.004 [-0.27]
Purchase			-0.001 [-0.11]			0.008 [0.51]
Observations	552	551	551	416	415	405
Adj. R^2	0.901	0.901	0.901 to 0.907	0.921	0.917	0.917 to 0.921

The dependent variables are standardized standard deviations of one-year ahead and 5-10 year ahead inflation expectations from the Michigan Survey of Consumers. For each model, we include the optimal number of autoregressive lags that minimizes AIC. Columns (1) and (4) regress the standard deviation on lagged year-over-year realized inflation and the optimal number of autoregressive lags. Columns (2) and (5) add the lagged log number of inflation sentences (i.e., newspaper sentences containing the word "inflation"). Columns (3) and (6) report results from separate regressions that each include the optimal number of autoregressive lags, lagged realized inflation, the lagged log number of inflation sentences, and one narrative at a time. For lagged realized inflation and log number of inflation sentences, columns (3) and (6) show the range of coefficients and t-statistics [in brackets] across the narrative regressions. For narratives, they show the coefficient and t-statistic from the regression including that narrative. For columns (3) and (6), the adjusted R-squared values shown represent the range across all narrative regressions. Realized Inflation is the standardized year-over-year percent change in the Consumer Price Index for All Urban Consumers (All Items, U.S. City Average) from the Bureau of Labor Statistics. Narrative measures are lagged standardized monthly shares of inflation sentences from ProQuest newspaper articles that we classify as expressing that particular narrative about inflation. Sample period is March 1979 to February 2025 for one-year expectations and July 1990 to February 2025 for 5-10 year expectations. T-statistics shown in brackets are calculated using Newey and West [1994] standard errors, optimally selected using the rule-of-thumb bandwidth estimator, computed to be 5 lags. All specifications include a constant.

households and greater reliance on general news media when forming expectations.

Lower-income households also show greater sensitivity to the volume of inflation news. The coefficient on the lagged log number of inflation sentences is 0.057 (t-statistic 1.47) for the 1st quartile compared to -0.020 (t-statistic -1.05) for the 4th quartile. For long-run expectations, the lowest-income households show strong responses to cost narratives (0.101, t-statistic 2.17) and social narratives (0.092, t-statistic 1.80) and negative responses to demand narratives (-0.080 , t-statistic -2.56) and international narratives (-0.048 , t-statistic -1.82). Sensitivity to realized inflation also varies systematically. For long-run expectations, lower-income households show stronger anchoring to realized inflation (0.078, t-statistic 2.75) than higher-income households (0.027, t-statistic 0.78), suggesting higher-income households more effectively distinguish transitory from permanent shocks.

Table 2.8 reports regressions with the gap in expectations between the bottom and top income quartiles as the dependent variable. The 1st quartile expects higher inflation rates on average (5.03% for 1-year and 4.54% for 5–10 years) than the 4th quartile (3.45% for 1-year and 3.40% for 5–10 years). Cost narratives are associated with larger disparities, with a coefficient of 0.128 (t-statistic 2.30) for short-term and 0.126 (t-statistic 2.28) for long-term expectations. Social narratives similarly widen the gap (0.114, t-statistic 1.87, for 1-year; 0.117, t-statistic 2.03, for 5–10 year). International narratives narrow the short-term gap (-0.108 , t-statistic -3.07), and demand narratives narrow the long-term gap (-0.107 , t-statistic -3.00), suggesting some narrative types provide common information that aligns expectations across income groups at different horizons. The volume of inflation news itself widens the expectations gap, though the estimate is imprecise.

2.12.5 Education-Based Heterogeneity

A parallel pattern holds across education groups. For 1-year expectations, non-college graduates show meaningful sensitivity to expectations narratives (coefficient 0.052, t-statistic

Table 2.8: Income-based inflation expectations gap (lowest minus highest income quartile) and narrative shares. T-statistics in brackets use Newey and West [1994] standard errors.

Dep. Var.: Lowest - Highest Income Quartile Inflation Expectations Gap	Horizon: 1 year			Horizon: 5-10 years		
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline						
L. YoY Realized Inflation	0.188 [4.29]	0.149 [3.65]	0.087 to 0.164 [1.94 to 3.79]	0.115 [2.88]	0.096 [2.53]	0.040 to 0.106 [0.86 to 2.98]
L. Log(No. of Infl. Sentences)		0.097 [1.94]	0.073 to 0.112 [1.53 to 2.14]		0.041 [1.17]	0.021 to 0.056 [0.56 to 1.49]
Cause Narratives						
International			-0.108 [-3.07]			-0.038 [-1.09]
Wage			-0.048 [-1.77]			-0.055 [-1.66]
Supply			-0.053 [-1.51]			-0.018 [-0.56]
Expect			0.039 [1.14]			0.012 [0.47]
Fiscal			-0.031 [-0.84]			-0.003 [-0.08]
Demand			-0.017 [-0.49]			-0.107 [-3.00]
Monetary			0.020 [0.42]			-0.033 [-0.95]
Effect Narratives						
Cost			0.128 [2.30]			0.126 [2.28]
Social			0.114 [1.87]			0.117 [2.03]
Trade			-0.048 [-1.34]			0.011 [0.33]
Govt			0.047 [1.21]			0.065 [1.91]
Cost-Push			0.061 [1.19]			0.050 [0.90]
Purchase			0.036 [1.06]			0.000 [0.00]
Uncertain			-0.029 [-0.79]			0.004 [0.10]
Rates			0.025 [0.64]			0.015 [0.40]
Savings			-0.017 [-0.52]			-0.042 [-0.81]
Redistribution			-0.003 [-0.07]			0.039 [1.02]
Observations	524	523	523	395	394	394
Adj. R^2	0.279	0.278	0.277 to 0.287	0.528	0.522	0.521 to 0.531

The dependent variables are standardized differences between weighted average inflation expectations of the lowest income quartile and the highest income quartile from the Michigan Survey of Consumers. For each model, we include the optimal number of autoregressive lags that minimizes AIC. Columns (1) and (4) regress the expectation gap on its lagged realized inflation and the optimal number of autoregressive lags. Columns (2) and (5) add the lagged log number of inflation sentences to the model (i.e., newspaper sentences containing the word “inflation”). Columns (3) and (6) report results from separate regressions that each include the optimal number of autoregressive lags, lagged realized inflation, the lagged log number of inflation sentences, and one narrative at a time. For lagged realized inflation and log number of inflation sentences, columns (3) and (6) show the range of coefficients and t-statistics [in brackets] across the narrative regressions. For narratives, they show the coefficient and t-statistic from the regression including that narrative. For columns (3) and (6), the adjusted R-squared values shown represent the range across all narrative regressions. Realized Inflation is the standardized year-over-year percent change in the Consumer Price Index for All Urban Consumers (All Items, U.S. City Average) from the Bureau of Labor Statistics. Narrative measures are lagged standardized monthly shares of inflation sentences from ProQuest newspaper articles that we classify as expressing that particular narrative about inflation. Sample period is July 1981 to January 2025 for one-year expectations and April 1992 to January 2025 for 5-10 year expectations. T-statistics shown in brackets are calculated using Newey and West [1994] standard errors, optimally selected using the rule-of-thumb bandwidth estimator, computed to be 5 lags. All specifications include a constant.

2.14), supply narratives (0.027), and wage narratives (0.022). College graduates show more muted responses, with strongest reactions to uncertainty narratives (0.028, t-statistic 2.00), expectations narratives (0.026, t-statistic 1.64), and demand narratives (0.019, t-statistic 1.61). Non-college graduates also show greater sensitivity to the volume of inflation news. The coefficient on lagged log inflation sentences is 0.028 (t-statistic 1.04) for non-college graduates and -0.014 (t-statistic -0.90) for college graduates. For long-run expectations, non-college graduates respond to fiscal narratives (0.054, t-statistic 1.65) and cost narratives (0.073, t-statistic 1.01), while college graduates respond mainly to uncertainty narratives (0.054, t-statistic 2.06) and demand narratives (0.036).

Table 2.9 reports the education expectations gap. Non-college graduates report consistently higher inflation expectations than college graduates, with a gap of 0.76 percentage points for 1-year expectations (4.69% vs. 3.93%) and 0.62 percentage points for 5–10 year expectations (4.14% vs. 3.52%). Social narratives widen these gaps. A one standard deviation increase in social narratives is associated with a 0.169 standard deviation increase in the 1-year gap (t-statistic 3.43) and a 0.123 standard deviation increase in the 5–10 year gap (t-statistic 2.98). Cost narratives also widen the 1-year gap (0.130, t-statistic 2.62). Higher supply narrative prevalence narrows the long-run gap (-0.067 , t-statistic -1.80). The volume of inflation news widens the gap as well, with a coefficient of 0.076 (t-statistic 1.99) for 1-year and 0.052 (t-statistic 1.23) for 5–10 year expectations.

2.12.6 Narratives and Realized Inflation

A natural concern with these results is that narratives may simply track realized inflation, in which case their predictive power for expectations would reflect superior information about inflation fundamentals rather than a distinct channel. Regressions of realized inflation on the same set of narrative shares and news-volume controls show that neither the volume of inflation news nor specific narrative frames predict actual realized inflation. Narratives sub-

Table 2.9: Education-based inflation expectations gap (non-college minus college graduates) and narrative shares. T-statistics in brackets use Newey and West [1994] standard errors.

Dep. Var.: Non-College - College Graduates Inflation Expectations Gap	Horizon: 1 year			Horizon: 5-10 years		
	(1)	(2)	(3)	(4)	(5)	(6)
Baseline						
L. YoY Realized Inflation	0.005 [0.11]	-0.032 [-0.71]	-0.155 to -0.025 [-2.98 to -0.56]	0.119 [2.50]	0.098 [2.19]	0.033 to 0.123 [0.65 to 2.61]
L. Log(No. of Infl. Sentences)		0.076 [1.99]	0.061 to 0.090 [1.68 to 2.41]		0.052 [1.23]	0.031 to 0.067 [0.81 to 1.49]
Cause Narratives						
Expect			0.058 [1.87]			0.022 [0.72]
Wage			-0.045 [-1.56]			-0.023 [-0.64]
International			-0.045 [-1.51]			-0.021 [-0.55]
Fiscal			0.035 [0.91]			0.069 [1.59]
Demand			-0.032 [-0.84]			-0.053 [-1.45]
Monetary			0.018 [0.61]			0.014 [0.52]
Supply			0.019 [0.53]			-0.067 [-1.80]
Effect Narratives						
Social			0.169 [3.43]			0.123 [2.98]
Cost			0.130 [2.62]			0.087 [1.65]
Cost-Push			0.086 [1.89]			0.051 [0.85]
Savings			-0.032 [-1.22]			-0.056 [-1.49]
Govt			0.028 [0.74]			0.017 [0.44]
Purchase			0.010 [0.31]			-0.052 [-1.42]
Rates			0.004 [0.11]			-0.001 [-0.02]
Uncertain			0.003 [0.08]			-0.043 [-1.02]
Trade			-0.002 [-0.06]			0.034 [0.84]
Redistribution			-0.002 [-0.05]			0.033 [1.07]
Observations	546	546	546	399	399	399
Adj. R^2	0.418	0.421	0.420 to 0.435	0.463	0.464	0.462 to 0.472

The dependent variables are standardized differences between weighted average inflation expectations of non-college graduates and college graduates from the Michigan Survey of Consumers. For each model, we include the optimal number of autoregressive lags that minimizes AIC. Columns (1) and (4) regress the expectation gap on lagged year-over-year realized inflation and the optimal number of autoregressive lags. Columns (2) and (5) add the lagged log number of inflation sentences to the model (i.e., newspaper sentences containing the word “inflation”). Columns (3) and (6) report results from separate regressions that each include the optimal number of autoregressive lags, lagged realized inflation, the lagged log number of inflation sentences, and one narrative at a time. For lagged realized inflation and log number of inflation sentences, columns (3) and (6) show the range of coefficients and t-statistics [in brackets] across the narrative regressions. For narratives, they show the coefficient and t-statistic from the regression including that narrative. For columns (3) and (6), the adjusted R-squared values shown represent the range across all narrative regressions. Realized Inflation is the standardized year-over-year percent change in the Consumer Price Index for All Urban Consumers (All Items, U.S. City Average) from the Bureau of Labor Statistics. Narrative measures are lagged standardized monthly shares of inflation sentences from ProQuest newspaper articles that we classify as expressing that particular narrative about inflation. Sample period is January 1978 to June 2024 for one-year expectations and April 1990 to June 2024 for 5-10 year expectations. T-statistics shown in brackets are calculated using Newey and West [1994] standard errors with 5 lags.

stantially influence expectations across demographic groups while having limited predictive power for the inflation outcomes those expectations are trying to forecast. This disconnect indicates that narratives shape expectations through channels other than those conveying superior information about future inflation.

2.13 Discussion

The findings have implications for monetary policy and for theoretical accounts of expectation formation.

Central banks dedicate substantial resources to communication strategies, but only part of the message reaching households is under their direct control. Media interpretation and narrative framing are at least equally consequential. The heightened sensitivity of lower-income and non-college households to media narratives indicates that demographic-specific communication strategies may improve policy transmission, and that uniform messaging may amplify rather than narrow expectation gaps.

The regional and partisan differences in inflation narratives suggest that geographic variation matters when crafting communication strategies. Regional Federal Reserve banks could benefit from monitoring local media narratives as leading indicators of expectation formation in their districts.

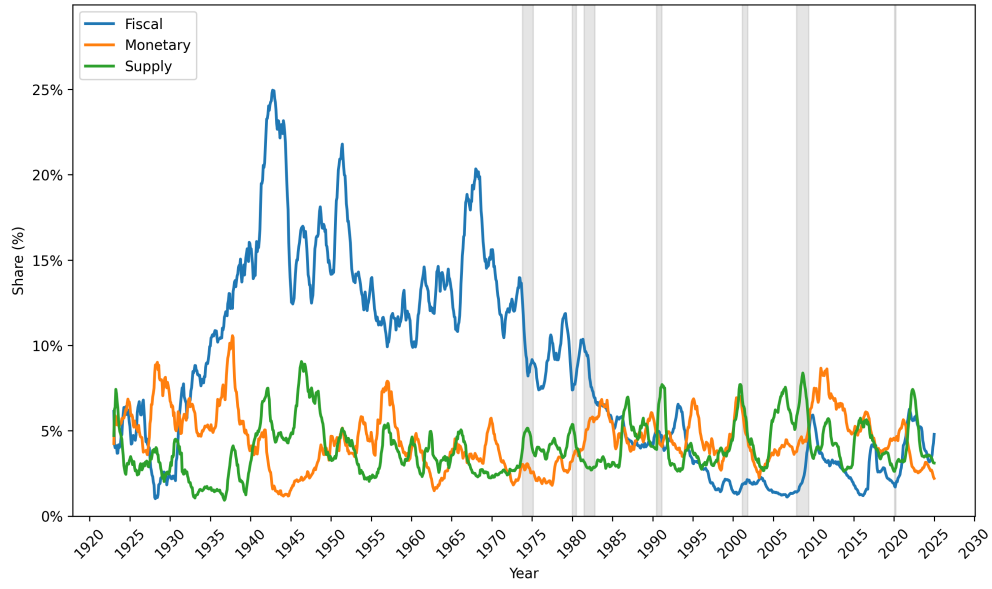
The relationship between narrative frames and expectation formation across demographic groups can be incorporated into heterogeneous-agent models of expectation dynamics. Doing so offers a path to better understanding monetary policy transmission and the persistence of inflation expectation disagreement, two regularities that purely fundamentals-based models of expectation formation struggle to reproduce.

2.14 Limitations

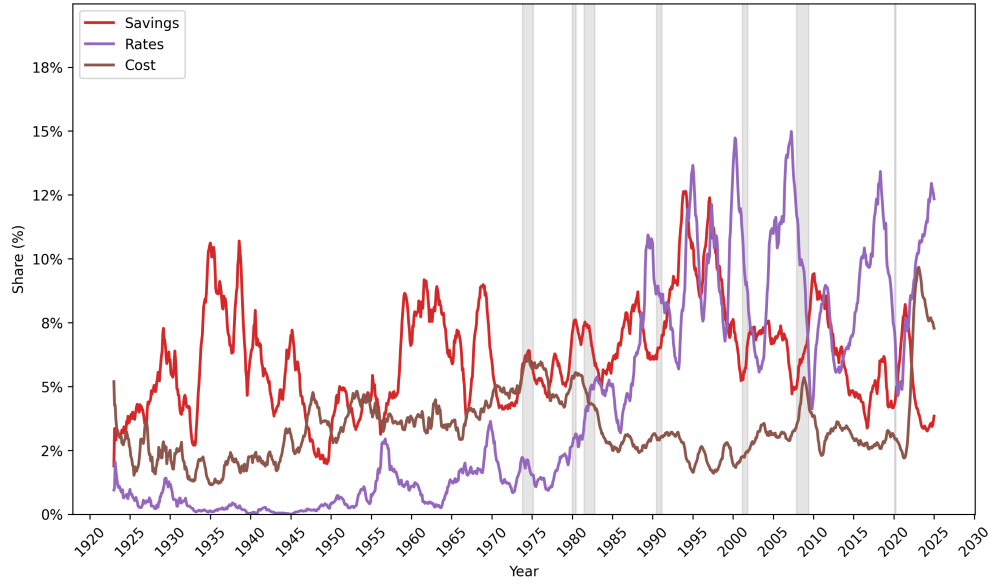
Several limitations apply to the findings in Part II. The regressions establish that narratives predict expectations, not that narratives *cause* expectation changes. Identifying causal effects requires plausibly exogenous variation in narrative exposure, which is left to future work using instrumental variable strategies based on plausibly idiosyncratic shocks to media coverage.

The classification approach in Part I requires manual development of an ontology of causes and effects, which limits automated discovery of mechanisms not already specified. The TopicGPT validation in §2.10.3 indicates that the manual ontology captures the full range of inflation explanations found in the corpus, but the validation is itself sample-based and would require re-running if the corpus is extended substantially.

The expectation analyses use survey data from the Michigan Survey of Consumers, which is the standard reference series but is one of several survey instruments. Additional analyses using the Federal Reserve’s Survey of Household Economics and Decisionmaking can establish whether the patterns documented here replicate in alternative measurement frameworks.

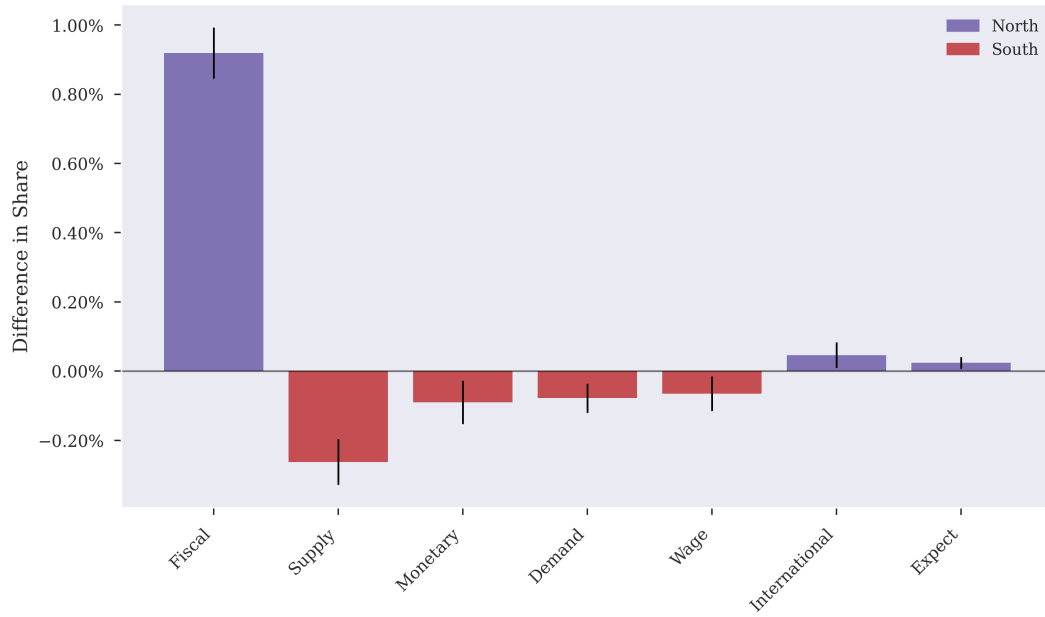


(a) Top 3 cause narratives.

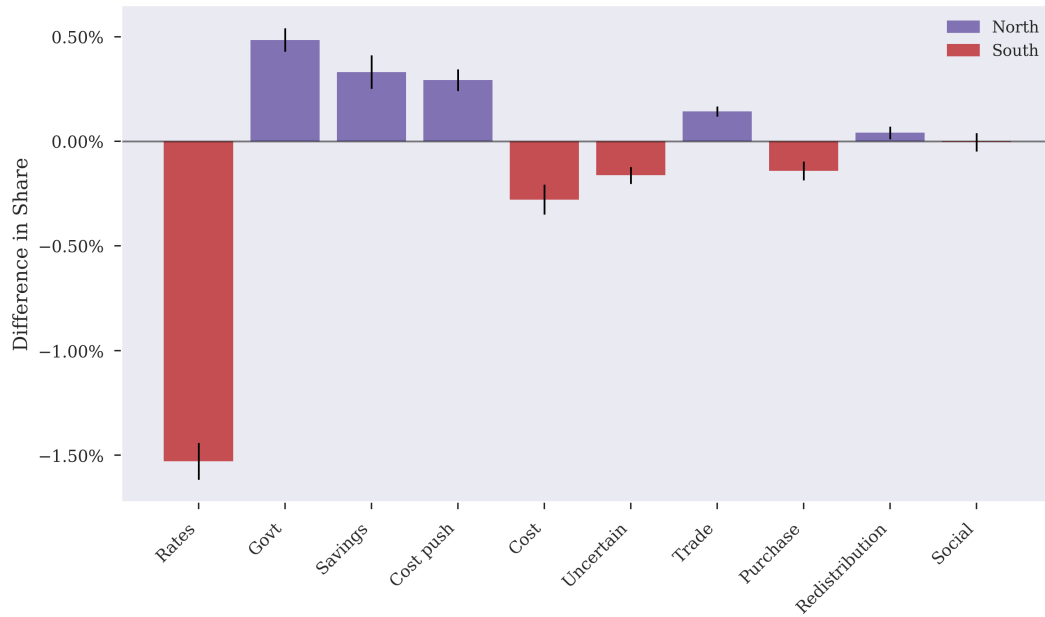


(b) Top 3 effect narratives.

Figure 2.4: One-year rolling average monthly shares of the top three cause and effect narratives across the full ProQuest sample, January 1923 to January 2025. Gray shaded regions indicate NBER recession periods.

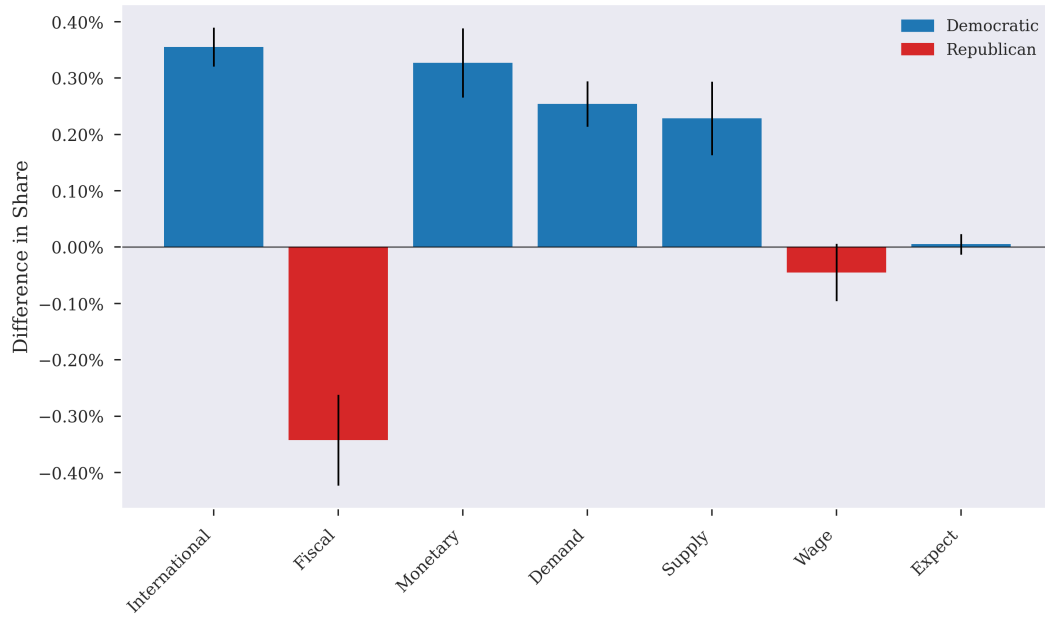


(a) Cause narratives.

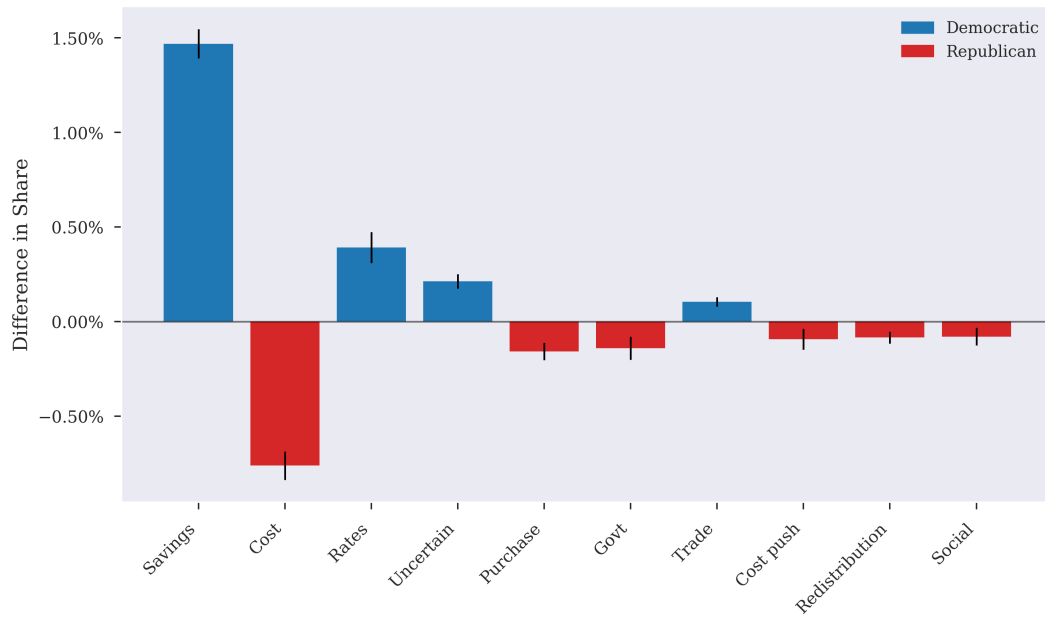


(b) Effect narratives.

Figure 2.5: Differences in inflation narrative shares between newspapers headquartered in Northern and Southern cities. Error bars represent 95% confidence intervals. Purple bars indicate narratives more prevalent in Northern newspapers, red bars indicate narratives more prevalent in Southern newspapers. Cities are classified relative to the Mason-Dixon line. Local newspapers only.



(a) Cause narratives.



(b) Effect narratives.

Figure 2.6: Differences in inflation narrative shares between newspapers headquartered in predominantly Democratic-leaning and Republican-leaning markets. Error bars represent 95% confidence intervals. Markets are classified by historical presidential voting patterns in three time-varying periods (1923–1960, 1960–1990, 1990–2025). Local newspapers only.

CHAPTER 3

CASESUMM: A LARGE-SCALE DATASET FOR LONG-CONTEXT SUMMARIZATION FROM U.S. SUPREME COURT OPINIONS

3.1 Introduction

Although large language models (LLMs) are claimed to handle long contexts [GPT-4 Team, 2024, Bubeck et al., 2023, Claude Team, 2024], including summarizing very long inputs, how well they perform long-context summarization is an open question.

Evaluating long-context summarization is challenging for several reasons. First, human ground-truth summaries are often not available [Cao et al., 2024, Chang et al., 2024]. Moreover, it’s unclear whether we should trust human abilities to even create ground-truth summaries. Second, what makes a good summary in one setting may not generalize to other settings. For example, what’s relevant in a legal text is different than what’s relevant in a novel. Lastly, identifying salient information in complex domains often requires expertise.

These challenges are addressed by introducing a new dataset where “ground-truth” summaries are available and conducting a comprehensive human evaluation to benchmark existing models. In particular, CASESUMM is built, a legal case summarization dataset consisting of 25.6K U.S. Supreme Court cases and their official summaries, called syllabuses. Syllabuses are written by an attorney employed by the Court and approved by the Justices. The syllabus is therefore the gold standard for summarizing majority opinions, and ideal for evaluating other summaries of the opinion. The opinions are obtained from Public Resource Org’s archive and syllabuses are extracted from the official opinions published in the U.S. Reporter and hosted by the Library of Congress. This dataset is at least approximately 25% larger, covers 3 times as many years (1815-2019), and is publicly available with fewer copyright restrictions than similar legal datasets [Fang et al., 2023b, Trivedi et al., 2024],

representing a rich resource for the research community.

Beyond the legal domain, several datasets have been introduced to improve evaluation of long-context summarization [Kryściński et al., 2022, Sharma et al., 2019, Kornilova and Eidelman, 2019, Huang et al., 2021b]. CASESUMM continues this trend, with longer source and summary texts and high-quality reference summaries written by domain experts. Unique among these datasets, CASESUMM spans over two centuries, showing substantial variation in the lengths and compression rates of summaries, and provides a resource for studying expert summarization in a high-stakes domain.

To highlight the opportunities and challenges of this dataset, both automatic and human expert evaluations of LLM-generated summaries of SCOTUS opinions are presented, and two “control” human-written summaries from Westlaw and Oyez are included. According to both human and automatic metrics, fine-tuning Mistral successfully guides the model to more accurately mimic the official syllabuses and reflect the lexical and semantic content within them, than other much larger models.

However, overall automatic metrics do not correlate well with human judgments, and LLM-based evaluation does not do better. Syllabuses and fine-tuned Mistral summaries perform highly on automatic evaluation but rank lower according to human evaluators, whereas GPT-4 is reliably ranked highly in human evaluation despite only average performance on automated metrics. Furthermore, GPT-4-generated summaries often outperform human-written ones, including official syllabuses, but not on factual correctness. These findings challenge the notion of human-written ground-truth summaries.

Finally, an error analysis of hallucinations in GPT-4- and Mistral-generated summaries is conducted, identifying factual errors ranging from precedent citation errors to misrepresentations of the facts of the case and procedural history as recounted in the source opinions.

In sum, the following contributions are made:

- A new large-scale dataset for long-context summarization in the legal domain is introduced,

consisting of 25.6K U.S. Supreme Court cases and their official syllabuses from 1815-2019.

- A comprehensive evaluation of LLM-generated summaries using both automatic metrics and expert human evaluation is presented, revealing discrepancies between these assessment methods.
- A comparative analysis of summaries generated by fine-tuned models and larger, general-purpose models is provided, offering insights into their relative strengths and weaknesses in legal summarization tasks.

3.2 Related Work

Evaluation for summarization. ROUGE [Lin, 2004] has been the dominant summarization metric, despite criticism of its high lexical dependence [Schluter, 2017, Cohan and Goharian, 2016]. Newer metrics like BERTScore [Zhang et al., 2019] and BARTScore [Yuan et al., 2021] aim to capture semantic similarities. However, automatic metrics often don’t correlate well with human judgments [Yuan et al., 2021, Fabbri et al., 2021, Bhandari et al., 2020]. This work focuses on high-stakes long-context summarization, showing the need for better metrics persists despite LLM progress. Chang et al. [2024] extended LLM-based evaluation to book-length summaries, but this approach doesn’t consider how experts weigh the importance of including or omitting certain information in a summary, while also being slow and costly. Cao et al. [2024] developed a framework for characterizing LLM summaries of financial documents. This work extends this research by evaluating and comparing model- and human-generated summaries in the legal domain.

LLMs as evaluators. A growing line of work uses LLMs themselves as evaluators of generated text, motivated by the cost of expert annotation and the limited coverage of reference-based metrics. G-Eval [Liu et al., 2023] prompts a language model to score candidate summaries along dimensions such as consistency, relevance, coherence, and fluency.

Subsequent work introduces more structured protocols, including fact-decomposition approaches [Song et al., 2024]. The evaluation in this chapter contributes a domain-specific test of whether an LLM judge tracks expert ranking better than lexical overlap metrics in a setting where expert judgment is the relevant ground truth.

Hallucination in long-form generation. Factual errors in generated text, often called hallucinations [Huang et al., 2023], are a known failure mode of abstractive summarization and become more difficult to detect as outputs grow longer and more fluent. Hallucinations in legal text are particularly consequential because errors can propagate to downstream legal research and reasoning. The error analysis in this chapter contributes a fine-grained taxonomy of hallucination types observed in summaries of Supreme Court opinions, distinguishing surface-level factual errors from misrepresentations of legal reasoning.

Expert evaluation in high-stakes domains. Prior work on summarization in expert domains has emphasized that surface-similarity metrics can mask substantive errors. In finance, expert raters identify omissions invisible to ROUGE [Cao et al., 2024], and analogous gaps appear in medical and scientific summarization. The legal domain places further demands on summary content: the holding, the procedural posture, and citation accuracy each carry distinct interpretive weight, and a summary that captures one without the others is not merely incomplete but potentially misleading.

NLP and summarization in the Legal Domain. Natural language processing has been applied to various legal tasks, including summarization [Bauer et al., 2023], discovery [Zou and Kanoulas, 2020], redaction [Garat and Wonsever, 2022], case outcome prediction [Medvedeva et al., 2023, Cui et al., 2023], and Bar Exam performance [Katz et al., 2023a]. For comprehensive surveys of NLP in the legal domain, see Katz et al. [2023b] and Kapoor et al. [2024].

Datasets in the legal domain. This dataset is unique in providing U.S. Supreme Court opinions with syllabuses, unlike other datasets that lack syllabuses [Chalkidis et al., 2022,

Dataset	# Docs.	# Words	
		Source	Summary
Super-SCOTUS (1955-2019)	6.6k	9.3k	791
BillSum (1993-2018)	22.2k	1.8k	208
GovReport*	18.5k	9.4k	553
Multi-LexSum**	4.5k	75.5k	647
Oyez (1955-2012)***	622	4.8k	356
Westlaw (1956-2011)	156	4.5k	143
CASESUMM (1815-2019)	25.6k	2.6k	314
CASESUMM (1955-2019)	7.2k	4.9k	745
CASESUMM (1815-1955)	18.4k	3.4k	289

Table 3.1: Comparison of CASESUMM and related long-context summarization datasets in the legal domain. *GovReport does not report the range of years covered. **Multi-LexSum is a multi-document summarization dataset. ***Oyez summaries are a subset of SuperSCOTUS.

Henderson et al., 2022] or provide only ancillary data [Law, 2024]. Fang et al. [2023b] present Super-SCOTUS, a dataset of Supreme Court documents, including a subset of syllabus and opinion pairs and highlight its contribution to political and social science research. In contrast to Super-SCOTUS, the CASESUMM dataset consists of cleaned and carefully trimmed opinion and syllabus pairs. For each decision PDF, the syllabus and majority opinion are identified and extracted. Headers and concurrent and dissenting opinions are removed, while properly including footnotes. CASESUMM is a strict super-set of Super-SCOTUS with descriptive statistics that reflect the improved data processing pipeline. CASESUMM extends further back to 1815 and, by being extracted directly from source opinions, provides the community with a readily available summarization resource with fewer copyright restrictions. Table 3.1 situates CASESUMM alongside related long-context legal summarization datasets.

3.3 Dataset

When the Supreme Court resolves a case, it publishes a majority “opinion” announcing the outcome and reasoning for their decision. The Court will also disseminate a summary of the opinion called the “syllabus”, which is written by an attorney employed by the Court and approved by the Justices. The syllabus must include the main elements of the opinion: the facts of the case, the procedural history, the legal question to be decided, and the answer to that question. Accurately summarizing each of these sections requires (1) understanding sophisticated legal reasoning and (2) identifying the most salient aspects of the case.

As one of the longest standing institutions in U.S. history, the Supreme Court has published thousands of opinions and syllabuses over the past 200 years. Looking at cases between 1815 and 2019, 25.6K pairs of opinions and syllabuses are collected for the dataset.

Dataset construction. The dataset is compiled from multiple sources. Opinions published in U.S. Reports Volume 15-546 (years 1815-2005) and Volumes 546-591 (2005 through *Trump v. Vance* (2019)) are obtained from Public Resource Org’s online archive [Public Resource Org, 2024] and the Super-SCOTUS data set [Fang et al., 2023b], respectively. Syllabuses are extracted from PDFs of the opinions hosted on the Library of Congress’s website [Library of Congress, 2024].

Extracting syllabuses from the original PDFs is challenging for several reasons. First, identifying the start and end of the syllabus is complicated because the formatting and style of SCOTUS decisions have changed over time. Low quality scans of 19th and 20th century documents make the extraction task even more difficult. Together, these issues constrain the kinds of rules, or signals, that can be leveraged to reconstruct the structure of the text in the PDFs, requiring identification of alternatives.

To ensure accurate syllabus extraction, the PDFs are processed in multiple ways. First, a set of regular expressions is designed to identify the start of the syllabus, providing coverage of decisions with different styles. Then, an algorithm based on open-source computer vision

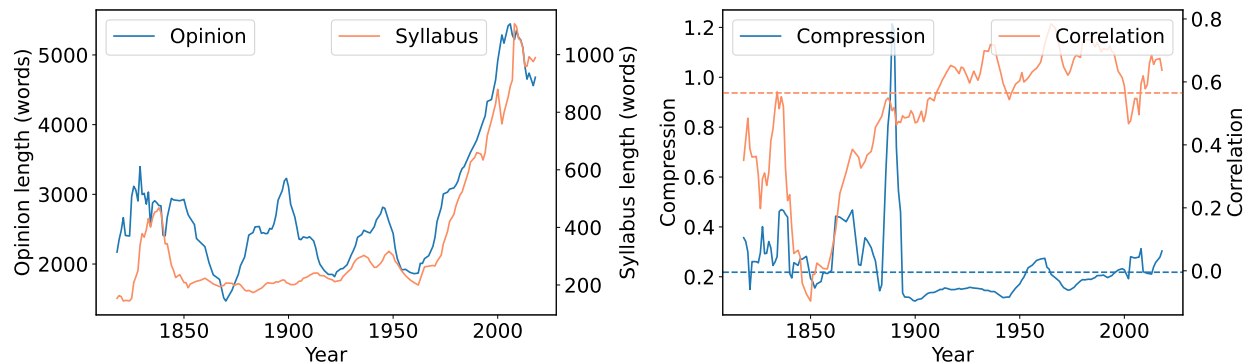


Figure 3.1: Opinion and syllabus lengths, compression rates by syllabuses, and correlations between opinion and syllabus lengths, 1815-2019. Dashed blue and orange lines give average compression rate and correlation. Lines are smoothed with 5-year moving-average.

software [Bradski, 2000] is developed to identify continuous lines, allowing distinction of footers from the main text of a page. Finally, differences in line density, a measure that is more robust to OCR and scan quality, combined with regular expressions are used to determine when the syllabus ends.

Since a new dataset is built, there are no accessible ground-truths to automatically evaluate the technique for extracting syllabuses from PDFs. Instead, 100 cases are randomly sampled and the extracted syllabus is manually evaluated by comparing them to the original PDFs. 96 of the 100 are perfect extractions while the remaining 4 syllabuses are partially truncated. These results highlight the quality of the dataset as a rich resource for long-context summarization.

Descriptive statistics.

To demonstrate the value of the dataset as a resource for abstractive summarization, the lengths of the opinions and their syllabuses are compared. The average Supreme Court opinion is 2,612 words long. The average syllabus is 314 words long, about 21.8% the length of the opinion it is summarizing. Figure 3.1 shows lengths have risen over time. Since 1980, opinions and syllabuses average 4,151 and 731 words, respectively, nearly double the average for the entire 1815-2019 period. Although compression rates, defined as the ratio of words

in a syllabus to words in an opinion have been relatively stable over time, averaging 21.8% from 1815-2019, the Pearson correlation between the length of an opinion and its syllabus, while variable, has increased over time. Whereas this correlation was just 0.46 before 1920, it has been 0.68 since then. Given the changes in opinion and syllabus lengths and in the correlation between syllabus and opinion lengths, this data set is a valuable resource for modeling and evaluating expert summaries, especially in the legal domain.

3.4 Experiment Setup

3.4.1 Data and Modeling

Data preprocessing and splits. Syllabuses are used as a supervision signal in the summarization modeling experiment and as reference summaries for evaluating the human and model-generated summaries. As discussed in §3.3, the substance and style of syllabuses have changed over time. Therefore, the supervision signal has changed over time. The motivating use case in the summarization task is a legal professional conducting research. For such a professional, while concision has value, comprehensiveness is more valuable. By manually studying summaries, more comprehensive syllabuses are determined to begin with a summary of the facts of the case, followed by a new section—marked by the text “Held:”—containing details about the issues, analyses, and conclusions that the opinion commented on. Modern syllabuses consistently adhere to this structure.

Therefore, the dataset is filtered to include only opinion/syllabus pairs where the syllabus contains the pattern “Held:”. This subset of the dataset is called “structured”. The length of structured syllabuses is more strongly correlated with the length of their respective opinions ($r = 0.65$) than the length of unstructured syllabuses is with the length of their opinions ($r = 0.46$). Furthermore, structured syllabuses are on average 2.5x longer than the unstructured syllabuses. Overall, the structured dataset contains 6,683 case/syllabus pairs. These are

split into a training set ($n=5,455$), validation set ($n=606$), and test set ($n=622$).

Modeling. Two approaches for completing the legal case summarization task are pursued and tested. The first approach is zero-shot prompting with proprietary and with open-source LLMs. The proprietary LLM employed is GPT-4 Turbo [GPT-4 Team, 2024]. The open-source LLM employed is Mistral 7b Instruct [Jiang et al., 2023]. The opinions in the dataset have 4,983 tokens on average, and the syllabuses average 755 tokens. The second approach is instruction fine-tuning [Wei et al., 2021] the open-source model, Mistral 7b Instruct, using the syllabuses in the training data set. Models used in a zero-shot setting are referred to by model name: Mistral Base and GPT-4, and the fine-tuned Mistral model as Mistral FT.

For fine-tuning, the input consists of a short instruction, the opinion, and the syllabus. Standard auto-regressive language modeling is used but only the language modeling loss for the syllabus is backpropagated. LoRA-based Parameter-Efficient Fine-Tuning (PEFT) [Hu et al., 2021] is used to train a subset of the parameters.

3.4.2 *Evaluation Strategies*

“Control group” summaries. The three machine-generated summaries (Mistral Base, Mistral FT, and GPT-4 Turbo) are benchmarked along with two additional human-generated sources for purposes of having a control group of human-written summaries not explicitly intended to mimic syllabuses. First, public Oyez summaries are collected from the Super-SCOTUS dataset [Fang et al., 2023b]. Oyez summaries are composed of three sections: Facts of the Case, Question, and Conclusion. Second, Westlaw’s commercial summaries of cases are collected via their online interface.

Automatic Evaluation. Following recent work on summarization [Koh et al., 2022], ROUGE and BERTScore [Lin, 2004, Zhang et al., 2019] are used as automated metrics for evaluating generated summaries against the reference syllabuses. With this, the *relevance* of the summaries is assessed. Each of the metrics is broken down by their precision, recall,

and F1-score, highlighting how models balance trade-offs between coverage and concision.

To further characterize the summaries, the summaries are compared based on *compression rate*, defined as the number of words in a syllabus over the number of words in an opinion, and the *correlation* between opinion lengths and summary lengths. Compression is used as a measure of brevity and correlation as a measure of how responsive summaries are to changes in the amount of content in the opinions.

Human Evaluation. For the human evaluation, second- and third-year law students are recruited and paid to read several opinions and 5 summaries of each opinion (Mistral FT, GPT-4, official syllabus, Westlaw, and Oyez). Students are asked to rank each summary (from 1 to 5) on several metrics: did the summary contain all relevant information from the opinion (*sensitivity*), did it exclude irrelevant information (*specificity*), was the summary clear (*clarity*), and did the summary have a style suggesting it was written by an experienced attorney (*style*). Finally, students are asked to report the number of facts in the summary that were false based on their reading of the opinion (*error*). Students are not told the source of each summary.

In total, students read 57 opinions. The sample of opinions and summaries included 33 unique cases, and the median student read 5 cases. Given that students are asked to rank opinions from 1 to 5 (implying a mean of 3 and variance of 2), the minimum detectable effect, with 95% confidence and 80% power, was 0.52 rank points.

Experimenting with LLM-based evaluations. Metrics like ROUGE and BERTScore provide a baseline for assessing lexical and semantic alignment between a candidate and reference text. However, they can miss deeper qualities that humans value in a good summary. LLMs offer a new way to evaluate summaries and to address some of these shortcomings [Liu et al., 2023, Song et al., 2024]. Still, their results are variable and sensitive to their prompts. In this work, how well G-Eval [Liu et al., 2023], a GPT-4-based evaluation tool, agrees with human ratings compared to ROUGE and BERTScore is studied. This helps understand

whether G-Eval offers a better way to evaluate summaries when a reference is unavailable or when traditional metrics fall short. Both the default implementation of G-Eval, as well as an adapted version to more closely reflect the human evaluation setup, are tested.

3.5 Results

The results indicate consistencies and discrepancies in the outcomes of automatic and human evaluations. On the one hand, model-generated summaries largely outperform the control human-written summaries on automatic measures of relevance, while also matching or exceeding them in the human evaluation. On the other hand, automatic metrics prefer Mistral FT summaries over GPT-4 ones, whereas expert humans most commonly rank GPT-4 over Mistral FT. Furthermore, all summaries are shorter than their reference syllabuses and do not correlate as strongly with opinion lengths. Despite this, humans prefer GPT-4 summaries, revealing that its summaries may represent a more desirable trade-off between concision and comprehensiveness.

3.5.1 Automated Evaluation Favors Fine-tuned Mistral Summaries

Overall, fine-tuning Mistral is particularly effective at improving the recall scores across all the metrics: ROUGE recall scores increase by an average of 21 points, BERTScore recall by 15 points. However, effects of fine-tuning on precision are weaker and more mixed. Table 3.2 reports ROUGE and BERTScore precision, recall, and F1 against the official syllabuses for all five summary sources, and Table 3.3 reports BARTScore. Mistral FT achieves the best F1 on every ROUGE variant and on BERTScore, while GPT-4 obtains the best precision and second-best F1. Perhaps fine-tuning sacrifices brevity for inclusion of more words in a syllabus, i.e., improves the sensitivity of summaries at a cost to specificity.

Control summaries help highlight effects of style differences on automatic metrics. By comparing against the two control human-written summaries, Westlaw is clearly

Method	ROUGE-1 $\uparrow$			ROUGE-2 $\uparrow$			ROUGE-L $\uparrow$			BERTScore $\uparrow$		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4 Turbo	<u>71.2</u>	<u>37.1</u>	<u>45.1</u>	<u>31.1</u>	<u>15.5</u>	<u>19.2</u>	34.9	18.1	<u>21.9</u>	67.4	<u>62.2</u>	64.6
Mistral Base	64.3	13.4	20.0	23.4	4.6	7.0	<u>41.1</u>	7.8	11.8	61.3	48.5	54.0
Mistral FT	63.3	43.1	48.1	30.1	20.5	23.0	34.9	23.6	26.4	<u>66.0</u>	64.4	65.1
Oyez	64.0	35.1	41.6	28.5	15.0	18.1	34.4	<u>18.6</u>	<u>22.1</u>	64.2	<u>61.8</u>	<u>62.9</u>
Westlaw	71.5	20.5	29.4	32.7	9.1	13.2	42.3	11.8	17.0	<u>65.0</u>	55.7	59.9

Table 3.2: Automatic evaluation of model-generated and human-written summaries, where official syllabuses are the reference summaries. Sample includes 622 Supreme Court cases. There are 622 observations on each type of summary except Westlaw, for which we only have 156 observations. For each metric, we report precision (P), recall (R), and F1-score (F1). For each metric, we **bold** the best score(s) and underline the second best score(s).

an outlier. While GPT-4 and Mistral FT scores mostly resemble Oyez, Westlaw’s recall scores are particularly low, only surpassing Mistral base. This poor performance on recall, but strong performance on precision, may be a product of how short those summaries are. Table 3.4 reports descriptive statistics on summary length, compression, and the correlation between summary length and opinion length. Westlaw summaries average 142 words and have essentially no relationship with opinion length, whereas official syllabuses average 750 words with a strong opinion-length correlation ($r = 0.68$).

3.5.2 *Summaries do not Scale with Opinion Length as much as Official Syllabuses*

A unique aspect of CASESUMM is that it includes SCOTUS cases dating back to more than two centuries ago. This breadth enables researchers to investigate summaries from many different angles. In this subsection, candidate summaries are characterized through the lens of length and compression and how these variables may affect summary quality over time is explored.

Length & Compression. In the dataset, both the opinion and syllabus lengths system-

Method	BARTScore ↓		
	P	R	F1
GPT-4 Turbo	256.0	<u>335.1</u>	289.5
Mistral Base	<u>277.0</u>	380.1	316.8
Mistral FT	297.9	312.3	<u>298.1</u>
Oyez	346.5	<u>334.6</u>	339.3
Westlaw	307.8	345.2	323.5

Table 3.3: BARTScores of model-generated and human-written summaries, where official syllabuses are the reference summaries. Sample includes 622 Supreme Court cases. There are 622 observations on each type of summary except Westlaw, for which we only have 156 observations. We report precision (P), recall (R), and F1-score (F1). BARTScores are negative log-likelihoods, so lower scores are better. We **bold** the best score(s) and underline the second best score(s). For the scoring model, we use `facebook/bart-large-cnn`, the default model used in Yuan et al. [2021]

Method	Length	Compression	Correlation
Opinion	6640	-	-
Syllabus	750	0.176	0.676***
GPT-4	321	0.092	0.088*
Mistral Base	126	0.034	0.151***
Mistral FT	447	0.121	0.179***
Oyez	332	0.096	0.094*
Westlaw	142	0.044	0.025

Table 3.4: Descriptive statistics on summaries in test set ($n=622$). Length is number of words. Compression rate is ratio of words in syllabus to words in opinion. Smaller number is more compression. Opinion included as reference. (* $p < 0.05$ ** $p < 0.01$, *** $p < 0.001$)

atically co-vary across time. Syllabuses in the sample have an average compression rate of 17.6%, meaning they tend to be about one-sixth the length of the original opinions. In generated summaries, Mistral FT produces summaries closest in length to these syllabuses, even outperforming GPT-4, which was also prompt-optimized to mimic syllabuses. Westlaw produced the shortest summaries, followed by Mistral without fine-tuning.

Regarding the correlation between summary and opinion lengths, syllabuses demonstrate the strongest relationship with opinion lengths: doubling opinion length increases syllabus length by nearly $2/3$. In contrast, Mistral FT summaries show a weaker correlation, with doubling opinion length increasing summary length by only 18%. Westlaw summaries exhibit almost no correlation with opinion length, maintaining a consistent target length of approximately 150 words. These findings highlight the dataset as a rich resource for future work in investigating how automatic summarization methods may adapt to varying source document lengths, ensuring that all salient information is captured regardless of length.

Precision & recall diverge over time. The Supreme Court Data Base, which contains metadata on SCOTUS cases, is used to see if any particular metadata can explain variation in summarization quality. While notable variation is not found across most of these features, one exception is observed: the divergence between recall and precision across all summaries increases over time. Figure 3.2 reports ROUGE-2 precision and recall, broken out by the Chief Justice presiding when each opinion was issued. Summaries of earlier opinions, e.g., under the Warren Court, have greater parity between recall and precision compared to summaries from later opinions. One possible explanation for this trend is that opinions and syllabuses have become longer over time, while the summaries evaluated show a growing disparity between their lengths and opinion/syllabus lengths over time.

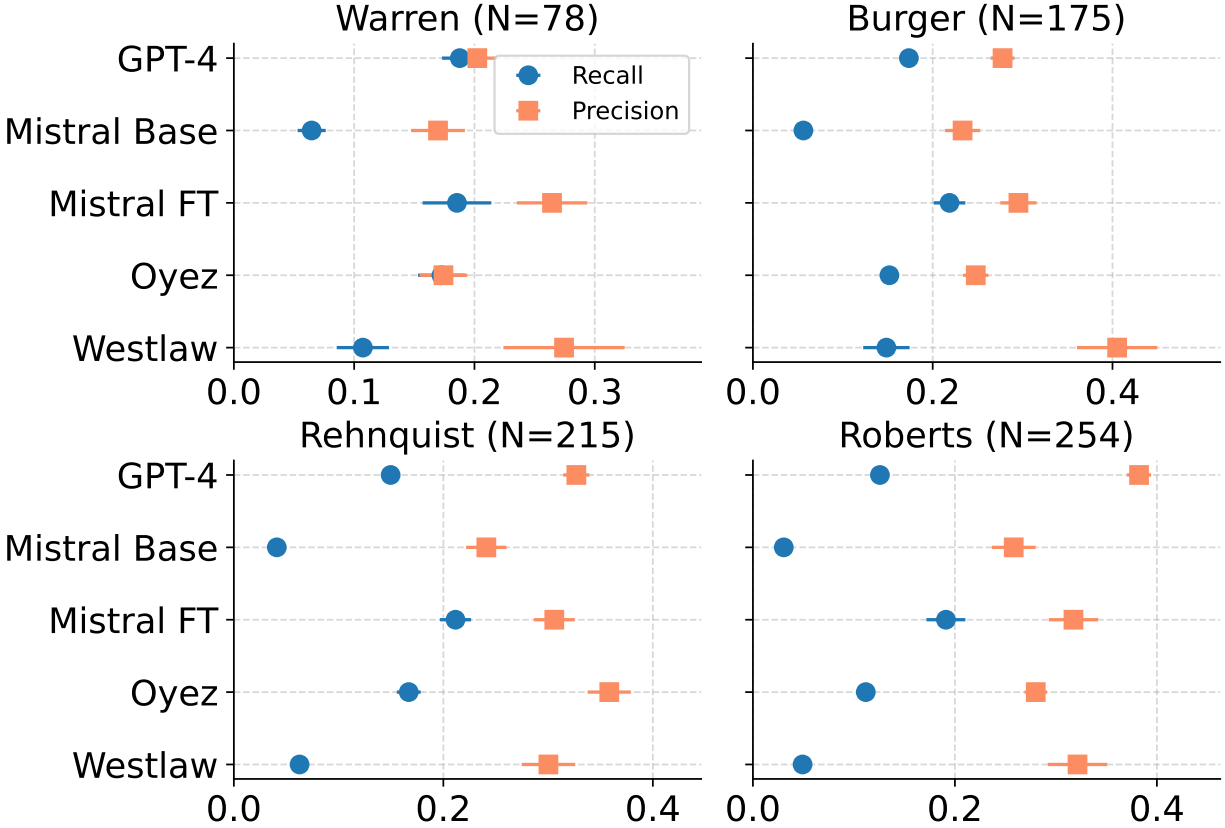


Figure 3.2: ROUGE-2 evaluation of model-generated and human summaries, by Chief Justice of SCOTUS when the opinion was written. Markers are means and whiskers are 95% confidence intervals. The gap between precision and recall widens for opinions issued under more recent Chiefs.

3.5.3 Human Evaluation Disagrees with Automatic Evaluation

The results of the human evaluation, presented in Figure 3.3, are distinctly different than those of the automatic evaluation. Whereas under automatic evaluation, Mistral FT outperforms other models as well as the control human-written summaries, humans most commonly prefer GPT-4 summaries. GPT-4 particularly excels on *clarity*, a crucial yet difficult to measure desideratum for the summarization task. Nonetheless, Mistral FT remains an above-average performer, successfully matching the original opinion syllabuses on every dimension except, importantly, number of errors. Evaluators report that roughly 20% of Mistral FT summaries have at least 1 factual error, with a total of 10 errors identified across all eval-

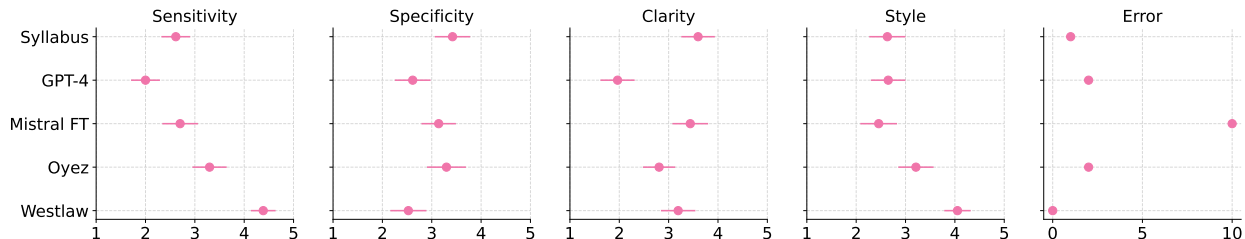


Figure 3.3: Human evaluation of model-generated and human summaries. x-axis is a rank, where 1 is best and 5 is worst. For **Error**, x-axis shows counts of the total number of errors identified by participants for each summary method.

uations. However, these factual errors, or *hallucinations*, are not necessarily a product of using LLMs, as GPT-4 has performance on par with syllabuses and Oyez in terms of factual correctness.

Surprisingly, the human evaluation also revealed that all three human-written summaries, including the official syllabuses, often performed worse than GPT-4. Westlaw summaries, despite being a paid service designed for legal professionals, ranked below average on sensitivity, clarity, and style. Even more intriguingly, the official syllabuses only matched or under-performed the LLM-generated summaries on all metrics except, crucially, factual correctness (*error*). This result both challenges the assumption that human-written summaries are inherently superior, while also revealing opportunities and challenges in using LLMs for generating concise, correct, and accessible summaries.

LLM-based evaluation does not correlate better with human evaluations than traditional automatic metrics. The human correlation results illustrate differences in how well various evaluation strategies align with human preferences.

First, ROUGE and BERTScore metrics are more capable of capturing other aspects of summary quality, particularly sensitivity and style. ROUGE-L achieves the highest correlation with human judgments of style (0.47), and ROUGE-1/2/L outperform G-Eval on judgments of sensitivity (Table 3.5). Despite its limitations, ROUGE offers useful signals for evaluating how well summaries balance inclusion and exclusion of content and how effectively

they convey proper legal style. It is worth noting that ROUGE shows the strongest negative correlation with specificity.

Metric	Sensitivity	Specificity	Clarity	Style
ROUGE-1	0.54	-0.28	-0.02	0.40
ROUGE-2	0.55	-0.32	0.02	0.39
ROUGE-L	0.57	-0.26	0.04	0.47
BERTScore	0.45	-0.11	0.01	0.43

Table 3.5: Spearman correlations between ROUGE and BERTScore and human rankings on the four quality dimensions. Interval scores are converted to rankings before correlation; human rankings for the same set of summaries are averaged.

Second, G-Eval shows significantly stronger correlations with human rankings of *clarity* compared to BERTScore and ROUGE (Table 3.6). This suggests that G-Eval better captures attributes like readability and logical flow, which are valued by human evaluators.

G-Eval (default)	Sensitivity	Specificity	Clarity	Style
Consistency	0.26	0.06	0.25	0.11
Relevance	0.45	0.09	0.26	0.30
Coherence	0.42	0.05	0.28	0.33
Fluency	0.43	-0.15	0.08	0.30

Table 3.6: Spearman correlations between default G-Eval criteria (Consistency, Relevance, Coherence, Fluency) and human rankings.

Third, G-Eval, while generally more consistent, performs similarly in its default and adapted versions, with only modest differences across dimensions (Table 3.7). For instance, the adapted G-Eval slightly improves correlations with human judgments of style and specificity but shows no significant advantage for clarity or sensitivity. This suggests that while adapting prompts can impact G-Eval’s results, it does not drastically alter its overall effectiveness.

These findings highlight the need for evaluation metrics, whether LLM-based or not, that are more closely aligned with human preferences and capable of capturing granular dimensions, such as clarity, specificity, and style, that matter in human judgment of summaries.

G-Eval (adapted)	Sensitivity	Specificity	Clarity	Style
Sensitivity	0.43	-0.01	0.19	0.23
Specificity	0.51	0.06	0.24	0.38
Clarity	0.38	-0.02	0.23	0.21
Style	0.42	-0.07	0.08	0.34

Table 3.7: Spearman correlations between G-Eval prompts adapted to mirror the four human-evaluation criteria and human rankings.

Absolute G-Eval scores by summary source. Table 3.8 reports the absolute G-Eval scores under both the default G-Eval prompts (consistency, relevance, coherence, fluency) and the adapted version that mirrors the four human-evaluation dimensions. Both versions of G-Eval rank GPT-4 first on every quality dimension. While this aligns with the human ranking on clarity and specificity, it diverges from the human ranking on the central question of factual correctness: Mistral FT receives middling but not catastrophic G-Eval scores despite producing the largest number of substantive factual errors in the human evaluation. The adapted prompts shift G-Eval scores only modestly, suggesting that careful prompt engineering does not by itself close the gap between automated judgment and expert assessment in this domain.

Method	Default				Adapted			
	Cons.	Rel.	Coh.	Flu.	Sens.	Spec.	Clar.	Style
Syllabus	<u>4.1</u>	<u>3.6</u>	<u>4.0</u>	3.0	<u>3.3</u>	<u>3.6</u>	3.8	3.0
GPT-4 Turbo	4.4	4.2	4.5	3.0	3.9	3.8	4.3	3.0
Mistral FT	3.1	3.1	3.5	<u>2.8</u>	2.8	3.0	3.8	2.6
Oyez	3.7	3.3	3.8	3.0	3.1	3.2	<u>4.0</u>	2.8
Westlaw	3.8	3.2	3.6	3.0	3.0	3.0	3.8	2.7

Table 3.8: G-Eval absolute scores under default criteria (Consistency, Relevance, Coherence, Fluency) and adapted criteria (Sensitivity, Specificity, Clarity, Style). Sample includes the 33 Supreme Court cases used for human evaluation. Best scores are **bold**, second-best are underlined.

3.5.4 Error Analysis

Each summary flagged as containing factual errors by the human evaluators is compared back to the original opinion to identify the specific source of error. Recent work refers to errors of this type as “hallucinations” [Huang et al., 2023]. The remainder of this section develops a taxonomy of error types observed across the two model families, examines how the distribution of errors differs between Mistral FT and GPT-4 Turbo, and connects these errors to a measure of summary style.

A taxonomy of hallucination types. The errors flagged in the human evaluation fall into five recurring categories. Table 3.9 presents one or two representative examples of each.

1. *Surface factual errors.* Misstatements of basic entity attributes such as the race, identity, or status of a party. Example 1 in Table 3.9 fabricates the petitioner’s race in a Mistral FT summary, contradicting the opinion’s own characterization of the petitioner as not belonging to a “favored group”.
2. *Citation errors.* Misspellings, fabricated case names, and incorrect U.S. Reports cites. Example 2 invents the case name “Sherbert v. Velelline” along with a fictitious volume and page reference. Errors of this type are particularly consequential in legal text because downstream readers may follow citations as if they were authoritative.
3. *Temporal and event-ordering errors.* Misplacements of when in a sequence of events something occurred. Example 3 reports that a party was held incommunicado after signing a statement, whereas the opinion places that detention before the signing.
4. *Procedural-history errors.* Mischaracterizations of how lower courts ruled or what relief was granted. Example 4 reports that a District Court entered judgment for the petitioner when in fact the petition was denied and the stay vacated. Errors of this type collapse the legal posture of the case and can flip the apparent meaning of a Supreme Court ruling.

5. *Legal-analysis errors.* Misrepresentations of the Court’s reasoning while preserving the surface form of the holding. Example 5 reports that the Court “rejected” a state interest, when the Court in fact assumed the interest *arguendo* and decided the case on a different ground. Example 6 understates the scope of the regulatory body at issue. Errors of this type are subtler than the others; the holding is often correct but the rationale is materially distorted.

Mistral hallucinates more conspicuously than GPT-4. Fine-tuned Mistral produces the most errors and the most egregious ones, predominantly in the first four categories: surface factual errors, citation errors, temporal errors, and procedural-history errors. These are detectable by careful comparison to the source opinion and would be visible to a non-specialist reader. GPT-4 Turbo errs less frequently and in a different way: its errors cluster in the legal-analysis category, where the holding may be reported correctly but the underlying reasoning is misstated. In one such case, the opinion reverses the judgment below but does not reject its reasoning; the reversal is grounded in a superseding constitutional issue. The GPT-4 summary characterizes the case as a rejection of the lower court’s reasoning, which is an analytically significant misstatement that nonetheless requires legal expertise to detect.

This split between conspicuous and subtle errors has consequences for evaluation. Surface and citation errors are in principle detectable by retrieval-based fact-checkers operating against the source opinion. Procedural-history errors require a structured representation of the case posture, but are still tractable. Legal-analysis errors require understanding the relationship between holding, reasoning, and dicta and resist automatic detection. The expert evaluators in this study identified all five types, but only the first four are likely to be caught by any automated pipeline available today.

Why automatic metrics miss these errors. ROUGE and BERTScore measure n-gram or embedding overlap with the reference syllabus. None of the error types above necessarily reduces overlap. A fabricated citation is a string substitution; a procedural-history inversion

is often a single-word flip from “granted” to “denied”; a legal-analysis error preserves most of the relevant terminology. G-Eval inherits the same blind spot through a different route: prompting a model to score consistency requires the model to detect precisely the failures it is itself prone to. The G-Eval rankings in Table 3.8 are consistent with this account: Mistral FT receives lower G-Eval scores than GPT-4, but not low enough to reflect the order-of-magnitude difference in factual error counts visible to expert annotators.

Lexical variation. *Lexical variation* is defined as the percentage of unique words in the summary not present in the opinion and is considered a measure of summary style. Figure 3.4 reports lexical variation for each summary source. Mirroring the comparison of compression rates, syllabuses exhibit the lowest percentage of lexical variation from the original opinion. Surprisingly, the fine-tuned Mistral summaries have the highest average percentage of lexical variation at 41.7%, even surpassing those written by Oyez (37.9%). This is unexpected because Mistral FT is trained on legal syllabuses, while Oyez summaries are written for a general audience and might borrow less from the opinion. The high lexical variation rate of Mistral FT is plausibly related to its higher rate of factual errors: a model that paraphrases more aggressively departs further from the literal claims grounded in the opinion and has more opportunities to substitute incorrect tokens for correct ones.

3.6 Discussion

The results in this chapter point to several lessons that extend beyond Supreme Court summarization.

Reference-based metrics overfit to the form of the reference, not its function.

Mistral FT outperforms GPT-4 on every reference-based metric in Table 3.2 and produces summaries whose length and compression rate most closely track the official syllabuses. Yet expert annotators rank GPT-4 higher on clarity, sensitivity, and specificity, and identify Mistral FT as the largest source of factual errors. The model that best mimics the syllabus

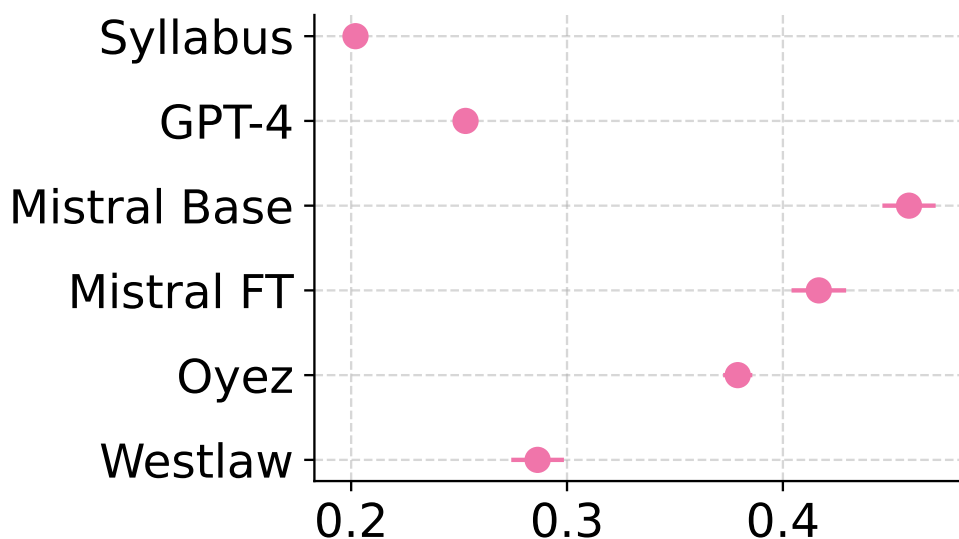


Figure 3.4: Lexical variation by summary source. Lexical variation measures the fraction of unique words in the summary that do not appear in the opinion. Higher values indicate more abstractive paraphrasing. Mistral FT exhibits the highest lexical variation, exceeding even non-legal Oyez summaries.

on the surface is not the model that best serves the communicative purpose of a syllabus. This is not a failure of fine-tuning per se but a failure of the evaluation: when a metric rewards stylistic and lexical proximity to a single reference, a model that learns the reference’s surface distribution receives credit even when it fabricates the underlying content. The implication is not that reference-based metrics should be discarded, but that they should not be treated as standing in for quality in domains where the function of the summary is distinct from its form.

LLM judges inherit the failures they are asked to detect. G-Eval ranks GPT-4 above Mistral FT on every dimension, which qualitatively agrees with human experts. But its scores do not separate the two models by anywhere near the margin that the human factual-error counts imply. The deeper issue is not that G-Eval mis-ranks the models but that the kinds of errors most damaging to a legal summary, particularly procedural-history inversions and legal-analysis errors, are exactly the kinds of errors that a frontier LLM may also make when reading the same opinion. Asking the judge to detect the failures of a peer

model places a structural ceiling on how reliable the judge can be. Reference-free LLM evaluation is most promising precisely where the model’s reading comprehension exceeds the candidate generator’s, which is not guaranteed in expert domains.

Hallucinations split into a tractable class and a hard class. The taxonomy in §3.5.4 suggests that error categories differ sharply in how amenable they are to automated detection. Surface factual errors and citation errors can in principle be caught by retrieval against the source opinion. Procedural-history errors are tractable given a structured representation of the case posture. Legal-analysis errors, where the holding is reported correctly but the reasoning is distorted, do not reduce to retrieval and are unlikely to be caught by any current automated method. Progress on long-context legal summarization will require evaluation methods designed against this hard class specifically, rather than aggregate metrics that conflate it with the easier categories.

Human-written references are not automatically better. The official syllabuses are written by an attorney employed by the Court and approved by the Justices, and have stronger correlation with opinion length and lower lexical variation than any model-generated summary. They were nonetheless ranked at or below GPT-4 by expert evaluators on clarity, sensitivity, specificity, and style. Westlaw summaries, despite being a paid commercial product for legal professionals, ranked below average on most dimensions. The only dimension on which human-written summaries clearly outperformed model-generated ones was factual correctness. This pattern challenges the default assumption that the human-written summary is the gold standard against which models should be measured, and suggests a more nuanced framing in which different sources are gold for different aspects of quality.

3.7 Conclusion

This chapter introduces CASESUMM, a novel dataset for long-context summarization in the legal domain, comprising 25.6K U.S. Supreme Court opinions and their official syllabuses.

The comprehensive evaluation of LLM-generated summaries, using both automatic metrics and expert human evaluation, reveals discrepancies between these assessment methods. While fine-tuned Mistral 7b outperforms larger models on automatic metrics, human experts rank GPT-4 summaries higher in clarity and accuracy. The human evaluation also showed that GPT-4 summaries often outperformed human-written summaries, including official syllabuses and professional services, in several metrics except factual correctness. LLM-based evaluation, such as G-Eval, may be a promising direction for reference-free evaluation but the results show that G-Eval does not correlate with human judgments better than traditional automatic metrics. The error analysis distinguishes a tractable class of hallucinations, including surface factual errors, citation errors, and procedural-history inversions, from a harder class of legal-analysis errors that resist automated detection because they preserve surface terminology while distorting underlying reasoning. These findings highlight the limitations of current automatic evaluation methods for legal summarization and underscore the importance of human evaluation in assessing summary quality, particularly in complex, high-stakes domains like law. This work contributes to the ongoing dialogue about evaluation methodologies in NLP and opens avenues for research in legal text summarization.

3.8 Limitations

First, the sample size of our human evaluation limits the conclusions we can draw. Second, while we are able to offer insight into the value of fine-tuning, at least with respect to the open-source Mistral model, we are unable to estimate the value of prompt-engineering even the GPT-4 model because we do not have a natural benchmark, non-optimized prompt for that model. A related weakness is that our evaluation of fine-tuning Mistral does not tell us the value of fine-tuning other models, such as GPT-4. It is possible that the benefit to fine-tuning the latter may be lower than the former because GPT-4 is trained on more data and has far more estimated parameters. Third, we experiment with one LLM-based evaluation

framework. While G-Eval is commonly cited and used, other LLM-based approaches could yield different results. Finally, while we demonstrate through a manual evaluation that our PDF extraction procedure is largely accurate (96%), it is not perfect. A fraction of syllabuses, particularly those extracted from low-quality scans from SCOTUS opinions in the early 1800s, may not be fully correct.

Hallucination in Summary	Explanation
<i>Fine-tuned Mistral</i>	
1. "Petitioner, a Negro, applied for admission to the University of Washington Law School, a state-operated institution."	The opinion indicates the petitioner is not a member of a "favored group" nor a "minority applicant". This strongly implies the petitioner is white and rules out the petitioner to be "a negro". 416 U.S. 312, 332, 325 (1974).
2. "Sherbert v. Velelline, 416 U. S. 456"	The opinion cites "Sherbert v. Verner, 374 U.S. 398" 494 U.S. 872, 875 (1990).
3. "He was held incommunicado for some five or seven days after signing the statement."	Petitioner Haynes testified he was held incommunicado until some five or seven days after his arrest. 373 U.S. 503, 504 (1963).
4. "The District Court ultimately entered judgment for petitioner, holding that the Texas death penalty scheme was unconstitutional."	While the District Court initially stayed the execution pending judgement, it ultimately "filed its findings and conclusions, rejecting each of the several grounds asserted by petitioner. The writ was accordingly denied; also, the stay of petitioner's death sentence was vacated." 463 U.S. 880, 885 (1983).
<i>GPT-4 Turbo</i>	
5. "The Court rejected the State's interest in [...] preserving the flag as an unalloyed symbol of the nation"	The Supreme Court did not reject the State's interest. They "assume[d], arguendo, that it is [valid]" but found "[t]he statute is nonetheless unconstitutional as applied to appellant's activity". 418 U.S. 405, 414 (1974).
6. "The PCAOB is a regulatory body that oversees the audits of public companies"	The PCAOB was created to govern the entire industry of accounting, including "hiring and professional development, promotion, supervision of audit work, the acceptance of new business and the continuation of old, internal inspection procedures, professional ethics rules, and 'such other requirements as the Board may prescribe.'" 561 U.S. 477, 485 (2010).

Table 3.9: Comparison of model hallucinations and their explanations.

CHAPTER 4

CIVICCHATS: DESIGNING AI SYSTEMS TO SUPPORT INFORMED AND JUSTIFIED VOTER DECISIONS

4.1 Introduction

Democratic legitimacy rests on citizens who can weigh competing considerations and arrive at positions they can defend. Deliberation, the process of forming and revising views through reasoning and exchange, is central to this ideal, and a large literature in political science argues that well-structured deliberation improves the quality of collective decisions [Dryzek et al., 2019]. The demand on individual deliberation is highest in direct democracy. Ballot measures are proposed laws, policies, or constitutional amendments placed before voters to decide directly rather than through elected representatives, and in the United States they determine consequential policy questions on housing, criminal justice, drug policy, labor rights, and state constitutions.

Voters receive little support for this task, and the technologies that mediate political information often work against deliberation. Social media rewards reactive content and superficial engagement, and exposure to opposing views on these platforms can increase rather than reduce polarization [Bail et al., 2018]. Even structured deliberation can polarize when participants enter with strong prior commitments [Wojcieszak, 2011]. Campaign messaging is designed to persuade rather than clarify, and endorsements offer convenient shortcuts that often replace, rather than facilitate, the weighing of competing considerations.

Conversational AI offers a different starting point. A chatbot allows people to ask questions on their own terms, including basic questions they might hesitate to ask another person, and to relate a policy to their own life, experiences, and preferences. These affordances address the two obstacles that make ballot measures inaccessible, technical language and unclear personal stakes. Mainstream chatbots do not provide them by default, however.

Large language models exhibit sycophancy, a tendency to match user beliefs and validate user framings rather than clarify or challenge them [Perez et al., 2022, Sharma et al., 2023a]. Without the kind of pushback one might encounter in a discussion among disagreeing friends, a default chatbot is ill-suited for helping people think through political issues in a balanced way.

Prior work suggests that conversational AI can nonetheless be designed for deliberative ends. Dialogues with a language model durably reduce conspiracy theory beliefs through evidence-based counterargument [Costello et al., 2024a, 2025]. Participants who used a Voting Advice Application chatbot regarded the interface as intuitive and informative, citing simple language and flexible interaction [Zhu et al., 2025]. Exposing users to two AI advisors presenting opposing evidence-based arguments improves judgment accuracy and confidence calibration relative to a single advisor [Rahman et al., 2025]. Outside of AI, participation in deliberative events is associated with an increased likelihood of vote switching [Bernhagen and Schmitt, 2014]. This chapter extends this line of research from belief correction and general political information to a concrete and consequential decision task, voting on a ballot measure.

Ballot measures are a demanding test case. Their texts frequently employ jargon-laden, nested clauses that obscure the consequences of one outcome over another, and voters often first encounter them in the booth. Measures with difficult texts are more likely to receive abstentions than their simpler counterparts [Shulman et al., 2022], and most voters report little prior familiarity with the measures they face. The population most affected by low political knowledge includes young, low-income, minority women [Sloan, 2024], and political knowledge increases the likelihood of political participation and voting [Reichert, 2016], so the voters with the most to gain from direct democracy may be the least equipped to use it.

This chapter presents CivicChats, a platform that informs voters and promotes interaction with ballot measures through chatbots tailored to specific discourse modes. The

platform implements three discourse modes. The Q&A mode answers questions neutrally and factually, providing information on demand. The Argumentative mode pushes back on the user’s stated stance with counterarguments drawn from both supporting and opposing positions. The Reflective mode asks probing questions that encourage examination of the user’s own reasoning. The three modes correspond to three components of deliberation that the literature treats as distinct: access to information, confrontation with opposing views, and reflection on one’s own reasons. The discourse modes are also hosted as a public resource at CivicChats.org, alongside additional interfaces for exploring live ballot measures; visitors are informed that conversations are collected for research and can request removal of their data.

A pre-registered controlled experiment evaluates the platform. The experiment assigned 600 participants to one of five between-subjects conditions arranged as progressive disclosures: official ballot text only, the official text plus structured civic information from Ballotpedia, and the three chatbot discourse modes. Participants evaluated live 2026 ballot measures, a Colorado measure on penalties for fentanyl sale and a Florida measure on changes to budget stabilization.

The experiment yielded two main findings. First, interaction with the chatbots changed vote outcomes substantially, in both cases shifting the vote distribution away from the default produced by reading the official text alone; on the Colorado measure the yes-rate under the Reflective mode fell by 45 percentage points, and on the Florida measure the chat conditions cut abstention by roughly 29 percentage points. Second, the chatbots improved participants’ ability to explain their vote, raising the number of distinct arguments and their engagement with counterarguments, while also raising self-reported understanding and decision confidence. The remainder of the chapter describes the methods, covering the CivicChats platform, the experimental design, and the measurement and analysis (§4.3); the findings, covering effects on vote choice, on the reasoning behind it, and on subjective

understanding and confidence, together with the mechanisms behind them (§4.4); and a conclusion covering implications and limitations (§4.5).

4.2 Related Work

Deliberative democracy and citizen competence. Deliberative democratic theory holds that legitimate collective decisions arise when citizens form and revise their views through reasoning and the exchange of justifications, rather than through the aggregation of fixed preferences [Dryzek et al., 2019]. Empirical work on deliberative events connects this process to electoral behavior: Bernhagen and Schmitt [2014] find that participating in a deliberative event raises the likelihood of vote switching, though the accompanying gains in political knowledge do not account for the shift. The benefits of deliberation are not unconditional. When participants enter with strong prior commitments, structured discussion can entrench rather than moderate their positions [Wojcieszak, 2011]. This work takes up two questions that follow from this literature: whether an interaction can improve the reasoning behind a vote, and under what conditions it changes the vote itself.

Technology-mediated political information. The technologies that mediate political information often work against the conditions deliberation requires. Social media rewards reactive and superficial engagement, and cross-cutting exposure on these platforms can deepen rather than reduce polarization; Bail et al. [2018] find that following opposing-party accounts increased rather than moderated partisan attitudes. Campaign messaging is designed to persuade rather than clarify, and endorsements provide shortcuts that substitute for weighing competing considerations. Conversational AI adds a new mediating layer to this environment, and its effect on civic reasoning has not been characterized.

Conversational AI for belief change and persuasion. A growing body of work shows that dialogue with a language model can change beliefs. Costello et al. [2024a] and Costello et al. [2025] find that evidence-based conversations with a model durably reduce conspiracy-

theory beliefs. Multi-advisor designs extend this beyond a single exchange: Rahman et al. [2025] show that exposing users to two AI advisors arguing opposing, evidence-based positions improves judgment accuracy and confidence calibration relative to a single advisor. This research establishes that conversational AI can be configured for correction and balanced argument rather than agreement, but it has largely targeted factual misperceptions rather than a concrete electoral decision.

Voting Advice Applications and voter information tools. Voting Advice Applications match voters to parties or candidate positions from their stated preferences, and recent work pairs them with conversational interfaces. Zhu et al. [2025] find that participants regarded a VAA chatbot as intuitive and informative, citing its plain language and flexible interaction, and describe its role in prompting reflection. These systems focus on candidate and party matching; the ballot-measure setting studied here instead requires a voter to reason about the content and consequences of a specific proposed law.

LLM sycophancy and designed pushback. Whether a conversational system challenges or merely affirms its user depends on a known failure mode. Large language models exhibit sycophancy, a tendency to match user beliefs and validate user framings rather than challenge them [Perez et al., 2022, Sharma et al., 2023a]. A model that mirrors a voter’s initial leaning offers none of the pushback a disagreeing interlocutor would, which makes a default chatbot poorly suited to balanced political reasoning. The discourse modes in this chapter are designed against this default. The Argumentative and Reflective modes specify a conversational posture that confronts or probes the user’s stance rather than affirming it.

Ballot-measure accessibility and political knowledge. Ballot measures are a demanding setting for any of these tools. Their texts are frequently dense and jargon-laden, and the way a measure is worded affects how voters respond to it: Shulman et al. [2022] show that the processing fluency of ballot wording predicts vote choice and election outcomes in low-information direct-democracy elections. The capacity to engage with such texts is

unevenly distributed. Political knowledge varies widely across the electorate and is lower among younger, lower-income, and minority voters [Sloan, 2024], and political efficacy is associated with both knowledge and participation [Reichert, 2016, Prats and Meunier, 2021]. Confidence in one’s political knowledge, distinct from its accuracy, predicts political discussion [Lee and Matsuo, 2018], which bears on the subjective-confidence outcomes examined in this chapter.

Across these strands, conversational AI has been studied for correcting misperceptions and for delivering general political information, but not for a concrete and consequential decision task such as voting on a ballot measure, and prior systems do not separate the components of deliberation that theory treats as distinct. This chapter addresses both gaps. CivicChats moves from belief correction and information access to a real ballot-measure decision, and it isolates three deliberative components, access to information, confrontation with opposing views, and reflection on one’s own reasoning, as separate discourse modes that share an interface and differ only in conversational posture. This design distinguishes the contribution of conversational posture from that of information access alone.

4.3 Methods

We evaluate CivicChats in a pre-registered, between-subjects experiment on live 2026 ballot measures. This section describes the platform and its discourse modes (§4.3.1), the experimental design and procedure (§4.3.2), and the outcome measures and statistical analysis (§4.3.3). This work received IRB approval (IRB25-038) from the University of Chicago, and informed consent is obtained from all participants.

4.3.1 *CivicChats*

Platform. CivicChats informs voters and promotes interaction with ballot measures through chatbots tailored to specific discourse modes. The chat interface pairs each

conversation with structured civic information drawn from the measure’s Ballotpedia entry. The platform is also hosted as a public resource at CivicChats.org, alongside additional interfaces for exploring statewide and local ballot measures; visitors are informed that conversations are collected for research and can request removal of their data.

Retrieval and web search. The chatbot grounds its answers in two complementary sources of evidence. The first is a per-measure document corpus comprising the full official text of the measure and its Ballotpedia entry, which together supply the legal language, a plain-language overview, and other relevant context curated by Ballotpedia. The measure’s metadata, introduction, and official ballot text are placed directly in the system prompt; the longer sections and full legal text are indexed for retrieval-augmented generation (RAG). To build the index, each measure is split into chunks along its natural document structure—one chunk per Ballotpedia section and one per block of legal text—and each chunk is embedded with a sentence-embedding model.¹ On every user turn the user message is embedded and compared against that measure’s chunks by cosine similarity, and the highest-scoring section and legal-language chunks (up to three of each) are supplied to the model as additional context for that turn. Retrieval is scoped to the assigned measure, so the chatbot draws on curated, measure-specific evidence rather than the open web by default. The second source is web search, which covers any questions or interactions that the corpus cannot answer, such as facts about sponsors and organizations, legal precedent, prior elections, or current events. A lightweight classifier inspects each user message—without access to the measure text or conversation history—and routes the turn to search only when it requires outside knowledge; conversational and opinion messages, and questions answerable from the measure itself, are handled from the grounded context alone. When search is invoked, the model queries the web and draws on the search results in its reply. The two sources are complementary: retrieval keeps answers anchored to the specific measure under discussion, while web search extends

1. We use OpenAI’s `text-embedding-3-small` model (1536 dimensions).

the chatbot’s reach to the surrounding political context.

Discourse modes. The platform implements three discourse modes, corresponding to three components of deliberation that the literature treats as distinct: access to information, confrontation with opposing views, and reflection on one’s own beliefs. The **AI-QA** mode answers questions neutrally and factually. The **AI-Argue** mode pushes back on the participant’s stated stance. The **AI-Reflect** mode asks probing questions that encourage examination of the participant’s own reasoning. The three modes share the interface and retrieval capabilities described above and differ only in conversational posture, which is specified through prompting; the full prompts are reproduced in Appendix C.4.

Discourse-mode validation. The deployed discourse modes were developed and validated with CivicEval, an evaluation interface developed for this work, separately from the user study. CivicEval scores full conversations with an LLM-based judge against checklists that specify the desired behaviors of each discourse mode. It also supports inspection and review of individual evaluations at the level of each conversation turn, and editing of the discourse-mode prompts and the checklists within the same interface, so that scoring, review, and revision proceed as a single iterative process. The criteria cover factual grounding, tone, conversation flow, and appropriate informativeness. Checklist items are scored at two levels. Turn-level items judge a single assistant response, such as whether its tone is matter-of-fact or its factual claims are grounded in the available evidence. Conversation-level items judge objectives that can only be assessed across the full exchange, such as whether the Argumentative mode maintains a single consistent opposing position across turns rather than flip-flopping, or whether the conversation as a whole surfaces assumptions the user had not initially recognized. The checklists were drafted manually and then refined iteratively: reviewing how the judge applied them to actual conversations surfaced unintended chatbot behaviors that no item penalized, items the judge applied differently than intended, and missing criteria, and the checklists were revised accordingly. The full checklists are re-

produced in Appendix C.6. For **AI-QA**, the judge evaluates evenhandedness in presenting relevant considerations without favoring one side. For **AI-Argue** and **AI-Reflect**, the judge evaluates the appropriateness of pushback and probing on user positions, the willingness to disagree forthrightly when warranted, and the absence of manipulative or leading language. The deployed discourse modes reflect iterative improvement across rounds of this evaluation. On the prompt-and-checklist artifact set used for deployment, pass rates are 93.7% for **AI-QA** (365 scored items), 82.0% for **AI-Argue** (562 items), and 94.5% for **AI-Reflect** (511 items); the lower **AI-Argue** rate reflects the strictness of the Argumentative checklist, which requires combative tone, a defended opposing-stance commitment, and concision (Table 4.1). Response latency, response length, and Flesch-Kincaid grade level are tracked alongside user inputs to encourage conciseness and accessibility; **AI-Reflect** produces the longest responses (mean 122 words versus 72 in **AI-QA**), **AI-QA** produces the highest-grade-level prose (mean grade level 15.8), and all three discourse modes respond in under four seconds on average (Table 4.2).

Table 4.1: LLM-as-judge discourse-mode rubric: pass rates on ‘regular’ and ‘subquestion’ rubric items, deployed-prod prompt v22 + checklist v18 (eval run `mo3-pilot`, 2026-03-06). Top section: pooled pass rate and number of scored items. Bottom section: pass rate on the eight highest-frequency rubric items per discourse mode.

Discourse mode	Pooled pass rate	N scored items
AI-QA	93.7%	365
AI-Argue	82.0%	562
AI-Reflect	94.5%	511

Pass rate = fraction of **Yes** answers among non-N/A judgments. The judge is GPT-class scorer at the eval run’s recorded `scorer_model` setting; see `civicchat-auto-eval/SCHEMA.md`.

Factuality audit. The factual accuracy of the deployed chatbots is audited on all 751 assistant turns sampled from the Colorado-measure study chats, stratified by condition. An LLM judge checks each response against the evidence available to the chatbot (the ballot measure context in its system prompt, the passages retrieved for that turn, and any cited

Table 4.2: Deployed-bot response characteristics per discourse mode (eval run `mo3-pilot`, 2026-03-06; prompt v22 + ballot data v6; same prompt artefacts that powered the CO/FL studies). Length is per assistant turn; FKGL is the Flesch-Kincaid Grade Level (`textstat`).

Discourse mode	N turns	Mean chars	Mean words	Mean FKGL	Mean resp. time (s)
AI-QA	27	447	72	15.8	3.18
AI-Argue	33	489	83	13.8	3.71
AI-Reflect	34	710	122	14.6	3.92

web pages) alongside the official ballot text, and labels the response factual unless it asserts a central claim that the evidence does not support. By this criterion, 86% of responses are perfectly factual and grounded (90% for `AI-Argue`, 79% for `AI-QA`, and 86% for `AI-Reflect`). To validate the LLM judge, the responses it flagged as not factual were reviewed by hand. Most of these flags did not reflect genuine errors: of the flagged responses reviewed, only 2 of 11 corresponded to a real factual inaccuracy, with the rest adequately grounded in the available evidence. The 86% figure is therefore a conservative lower bound on the true factual rate.

4.3.2 Experimental Design

Ballot measures. Participants evaluated one of two live measures from the November 2026 election, assigned by state of residence. The Colorado measure (Initiative 85) is an initiated state statute that would establish higher felony classifications and mandatory penalties for the manufacture, sale, and possession of fentanyl and certain synthetic opioids, making the sale of any amount of fentanyl a Class 1 felony punishable by up to 32 years in prison and requiring court-ordered treatment for certain low-level possession offenses. The Florida measure is a legislatively referred constitutional amendment that would raise the cap on the state’s budget stabilization fund from 10% to 25% of net general revenue, require an annual transfer of \$750 million or 25% of general revenue collections, whichever is less, and define the conditions under which the legislature may withdraw funds. Both measures have Ballotpedia

entries, which the **Ballotpedia** condition and the chatbot retrieval database require. They also contrast in topic and style. The Colorado measure addresses a salient and emotionally charged drug-policy question, while the Florida measure addresses a technical and procedural budget rule; this contrast supports the cross-measure comparisons in §4.4.

Conditions. Participants are randomly assigned to one of five between-subjects conditions, designed as a sequence of progressive disclosures in which each condition adds content or capability to its predecessor.

1. **Official_Text**: participants see only the official ballot measure text (reproduced in Appendix C.1). This serves as the control.
2. **Ballotpedia**: participants additionally view sections of the Ballotpedia entry for the measure, providing structured civic information.
3. **AI-QA**: participants view the Ballotpedia sections alongside the CivicChats chat interface (§4.3.1) running in the **AI-QA** discourse mode.
4. **AI-Argue**: the same interface and retrieval capabilities, running in the **AI-Argue** discourse mode.
5. **AI-Reflect**: the same interface and retrieval capabilities, running in the **AI-Reflect** discourse mode.

Recruitment. Participants were recruited via Prolific and prescreened as verified residents of the state whose measure they evaluated, with 200 valid responses for the Colorado measure and 400 for the Florida measure (600 total). Compensation was \$4 for an estimated 15-minute session. Participants were excluded if they failed the attention check, completed the survey below a minimum time threshold, or did not complete the full survey.

Survey flow. Each participant proceeds through the following steps.

1. **Registration.** The landing page explains the study, presents the consent form, and outlines compensation. Participants who consent provide a Prolific identifier and are assigned to a condition.
2. **Instructions and attention check.** Participants receive an overview of ballot measures and the study structure, then complete a multiple-choice attention check that they must answer correctly to proceed.
3. **Pre-treatment questions.** Participants rate their familiarity with the measure and their likelihood of voting on it in the upcoming election, each on a five-point Likert scale.
4. **Treatment.** Participants in the `Official_Text` and `Ballotpedia` conditions view the measure content. Participants in the chatbot conditions view the measure content alongside the chat interface and must send at least five messages to proceed.
5. **Vote and confidence.** Participants record a vote choice (Yes, No, or Undecided), a vote intent likelihood, decision confidence, and understanding confidence, each on a five-point scale.
6. **Decision explanation.** Participants provide a free-form explanation of their vote choice and address opposing viewpoints, with a minimum of 30 words and 2 minutes of active time on the page.
7. **Demographics and feedback.** Participants report education, political affiliation, voting frequency, trust in AI, and media habits.

4.3.3 Outcome Measures and Analysis

Outcome measures. The design measures vote choice (Yes, No, or Abstain), subjective confidence on two five-point Likert scales (decision confidence and understanding confidence),

demonstrated understanding via the free-text vote rationale (§4.3.3), and engagement via total time on task, time on each section, and, for chatbot conditions, the number and type of chatbot interactions. Mixing multiple-choice and free-form responses balances standardization with subjective quality, an approach supported by prior work in survey design [Conrad et al., 2025].

Annotation of rationales. Free-text vote rationales and chat transcripts are annotated by an LLM along two dimensions; the full annotation prompts are reproduced in Appendix C.5. First, an argument is a distinct pro or con consideration about the measure, and the decomposed pro and con counts are retained as separate outcomes; in the chat transcripts each argument is also attributed to the participant or to the chatbot. Second, counterargument engagement, the extent to which a rationale addresses opposing views, is scored on a Wachsmuth-style 0/1/2 rubric.

Conversation–rationale predictability. The conversation–rationale predictability analysis quantifies how predictable a participant’s written vote rationale is from their chat conversation, and attributes that predictability to what the participant said versus what the chatbot supplied. For each participant in the three chatbot conditions whose transcript is available ($n = 299$ across the two measures), an open-weight language model (Qwen2.5-7B-Instruct) scores the rationale by the mean per-token log-probability it assigns to the rationale’s tokens under four conditioning contexts, each of which includes the ballot measure information: the full chat transcript, the participant’s messages only (chatbot turns masked), the chatbot’s messages only (participant turns masked), and the ballot information alone. Averaging per token normalizes for rationale length. Contrasts between contexts define the quantities of interest: *conversation predictability* (full transcript versus ballot-only) measures how predictable the rationale is from the conversation as a whole, and it decomposes additively into the *participant contribution* (participant messages versus ballot-only), the extent to which the rationale follows from the participant’s own utterances, and the *chatbot contribution*

(full transcript versus participant messages), the additional predictability supplied by the chatbot’s messages. Because condition is randomized, between-condition differences in these quantities are attributable to the discourse-mode assignment. They are compared across the three chatbot conditions with Kruskal–Wallis and pairwise Mann–Whitney tests. Findings are checked by rescoring with additional models (Qwen3.5 at 2B, 4B, and 9B parameters) and corroborated with a complementary lexical measure of content-word overlap between the rationale and the transcript.

Statistical analysis. Continuous outcomes are analyzed with ordinary least squares regression of the outcome on condition and the two pre-registered covariates, pre-treatment familiarity with the measure and self-rated voting likelihood. Pooled estimates across the two ballot measures include a fixed effect for measure, and per-measure estimates are reported alongside them. The reference category is `Official_Text` for the between-condition tests and `AI-QA` for the within-chat tests. Vote choice is analyzed with both multinomial logit (reference `Official_Text`) and two-proportion z -tests on each vote category. Effect sizes are reported as Cohen’s d with pooled standard deviations for pairwise contrasts. Multiple-comparison adjustment is not applied to the pre-registered pairwise contrasts; raw p -values are reported. Each focal contrast is also re-fit under three alternative specifications (no covariates, with demographic covariates, and with stricter sample exclusion); these robustness checks are summarized in §C.8.1.

Two outcome measures are operationalized through the LLM annotation in §4.3.3. Argument counts and counterargument engagement map onto pre-registered constructs and are treated as confirmatory; the conversation–rationale predictability analysis is an additional measure introduced beyond the pre-registration.

4.4 Findings

The experiment analyzed 600 participants: 200 on the Colorado measure and 400 on the Florida measure, evenly split across the five conditions. Per-stage recruitment and exclusion counts are reported in Table C.1; exclusions were few and comparable across conditions. The one exception is drop-out on the Florida measure, which was somewhat concentrated in **AI-Reflect** relative to **Official_Text** (fourteen versus four), an absolute difference of about three percentage points, though the test of differential attrition was not pre-registered.

Demographic and political characteristics did not differ across the five conditions on either measure (Table C.2). Participants reported little prior knowledge of the specific measure, and reported voting somewhat more regularly in the Colorado sample than in the Florida sample.

4.4.1 *Vote Choice*

We hypothesized that talking to the chatbots, or reading the Ballotpedia summary, would change how participants voted relative to reading the official text alone. It did, on both measures (Figure 4.1, Table 4.3).

The two measures had very different defaults, and the treatments moved participants away from each. On the Colorado measure, reading the official text alone produced a 70% yes majority. Every treatment lowered the yes-rate, and **AI-Reflect** produced the largest shift, reversing it to a 72% no majority, a 45-percentage-point drop in yes. Most of the lost yes share moved to no rather than to abstention ($p < .001$).

On the Florida measure, the official text left 40% of participants abstaining. There the chat conditions reduced abstention by roughly 29 percentage points and divided those participants between yes and no, rather than moving the vote toward either side. **Ballotpedia**, which only summarizes rather than converses, did not change the Florida distribution; the shift was specific to the chat interaction.

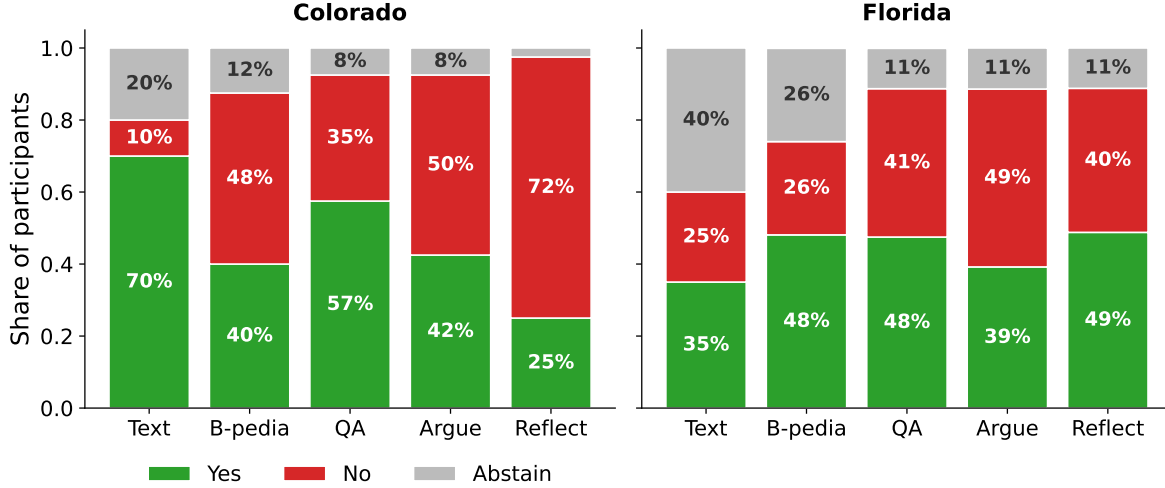


Figure 4.1: Vote distributions (Yes / No / Abstain) by condition for the Colorado fentanyl-penalties measure (left) and the Florida budget-stabilization measure (right). The baseline yes-rate differs sharply between measures: 70% in CO Official_Text versus 35% in FL Official_Text.

Table 4.3: Vote-choice distributions by condition, per study. Cells are within-condition percentages. Stars on yes/no/abstain cells mark a significant two-proportion z test against Official_Text in that vote category. The omnibus test in the ‘Overall’ row is a Pearson χ^2 of independence across the full 3×5 vote $\times$ condition table for each study.

Condition	Vote share (%)			z-test vs. Official_Text (<i>p</i>)		
	Yes	No	Abstain	Yes	No	Abstain
Colorado 2026 (CO)						
Official_Text	70%	10%	20%	—	—	—
Ballotpedia	40%**	48%***	12%	.007	< .001	.363
AI-QA	57%	35%**	8%	.245	.007	.104
AI-Argue	42%*	50%***	8%	.013	< .001	.104
AI-Reflect	25%***	72%***	2%*	< .001	< .001	.013
Overall	$\chi^2(8) = 36.6, p < .001$					
Florida 2026 (FL)						
Official_Text	35%	25%	40%	—	—	—
Ballotpedia	48%	26%	26%	.091	.893	.057
AI-QA	48%	41%*	11%***	.108	.029	< .001
AI-Argue	39%	49%**	11%***	.580	.002	< .001
AI-Reflect	49%	40%*	11%***	.078	.043	< .001
Overall	$\chi^2(8) = 39.1, p < .001$					

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ on the two-proportion z .

Across both measures, the treatments shifted participants away from whatever stance the official text presented as the default, a yes vote in Colorado and abstention in Florida. The direction of these shifts is stable across model specifications, though the magnitude of the Colorado drop softens under stricter sample exclusion (Table C.4).

Both directional shifts follow from the same mechanism. Under `Official_Text`, vote choices reflect low-engagement reasoning, in which participants accept whatever stance the official framing implies or abstain when the text is too difficult to follow. The treatments disrupt this pattern. `Ballotpedia` adds context, and the chat conditions require articulation; in both cases the baseline distribution moves toward the middle. Voters initially inclined toward yes who had not examined the measure begin to articulate cons, and abstainers begin to articulate arguments on both sides.

Within the chat conditions, votes track what users say rather than what the chatbot says. An additional user-articulated pro raises the odds of a yes vote by a factor of 1.65 in CO (95% CI [1.10, 2.49]) and 2.00 in FL (95% CI [1.34, 2.99]) with discourse-mode fixed effects, while an additional user con lowers them to 0.66 and 0.67 respectively; chatbot-side pro/con counts show much weaker association with vote. The slope of vote on articulation does not differ meaningfully across discourse modes (the joint interaction test is null in FL, $\chi^2(4) = 4.44$, $p = .349$, and on the $\alpha = .05$ boundary in CO, $p = .052$), so the discourse modes appear to move votes by shifting the level of what users articulate rather than the persuasiveness of any given argument. The Reflective chatbot’s open-ended probing and the Argumentative chatbot’s challenges make cons easier to enumerate than pros; the neutral AI-QA chatbot, which answers rather than probes, leaves the articulation profile and the vote distribution near baseline. This account is observational, since random assignment is at condition rather than at articulation counts, and a residual `AI-Reflect` effect on the CO yes-vote survives conditioning on articulation.

A competing account ties the shift to the measure’s topic rather than to a common de-

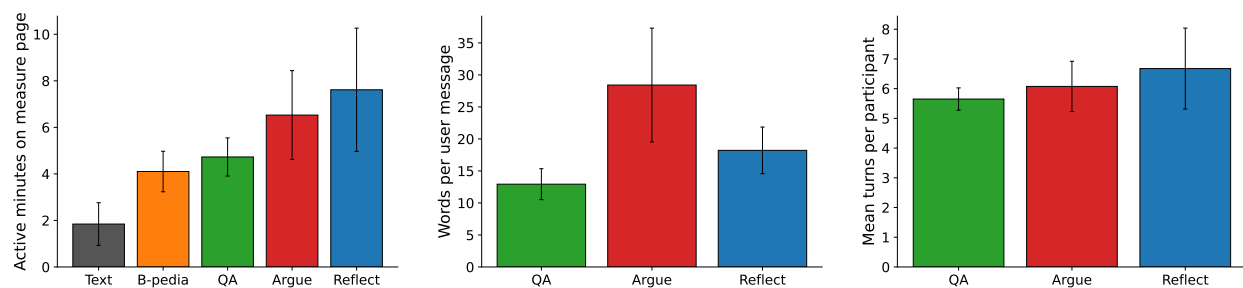
fault, with the treatments reducing yes-votes on emotionally charged topics and reducing abstention on technical ones. With only two measures the two accounts cannot be distinguished; a measure with a low baseline yes-rate on an emotional topic, or a high baseline yes-rate on a technical topic, would identify which factor dominates.

4.4.2 Reasoning and Engagement

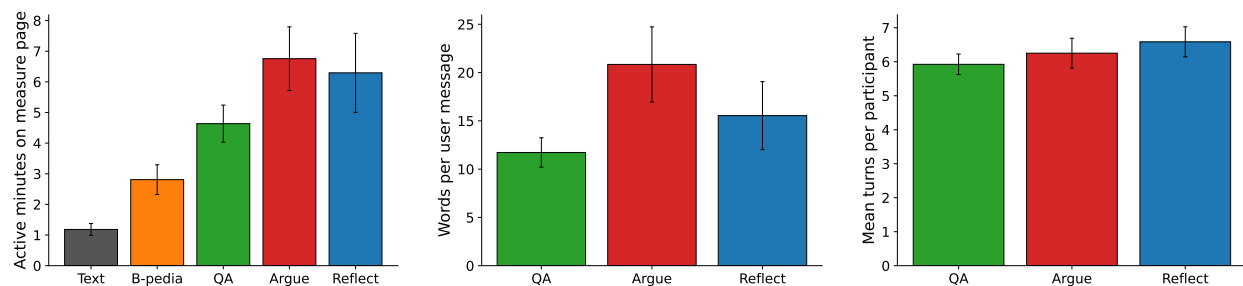
The vote shift was accompanied by changes in how participants engaged with the measure and in what their written rationales contained. We hypothesized that the two discourse modes would produce deeper engagement and richer rationales than plain question-answering, and we further examine whether those rationales reflect the participant’s own reasoning or the chatbot’s contributions.

Engagement. The discourse modes produced deeper engagement than plain question-answering. **AI-Reflect** and **AI-Argue** produced longer active times than **AI-QA**, and **AI-Argue** produced the longest user messages, nearly double the **AI-QA** length on the Florida measure. The modes also changed the content of participants’ questions. Under **AI-QA**, two-thirds of messages were simple factual clarification; under **AI-Argue** and **AI-Reflect** that share fell to roughly half, as participants more often raised comparative tradeoffs. Participants rarely solicited counterarguments directly; that engagement came through the chatbot’s output rather than through user requests. The time these conversations took accounts for most of the gap in time-on-task between the chat conditions and **Official_Text**, while reading the Ballotpedia summary took no longer than reading the official text (Figure 4.2).

Argument count and balance. **AI-Reflect** produced the most arguments on both measures, roughly one more on average than **Official_Text** ($p < .001$). **AI-Argue** and **AI-QA** also exceeded **Official_Text**, but **Ballotpedia** did not, so the gain comes from the conversation itself rather than from a richer summary; among the chat modes, **AI-Reflect**



(a) Colorado.



(b) Florida.

Figure 4.2: Engagement outcomes by condition for the Colorado (top) and Florida (bottom) measures. Columns show active minutes, mean message length in words, and conversation turns. Active minutes for chat conditions exclude AI response time; message length and conversation turns are defined only for the three chat conditions.

slightly exceeded **AI-QA** while **AI-Argue** was indistinguishable from it. The discourse modes change the balance of arguments more than their number. On the Colorado measure, **Official_Text** rationales averaged about 2.3 pros to 0.9 cons, while **AI-Reflect** rationales averaged roughly one pro for every four cons (1.0 to 3.9). The Florida measure, where no stance held a default majority, showed a weaker version of the same effect. **AI-Reflect** rationales were close to balanced, while **Official_Text** contained slightly more cons than pros. Pros and cons appear in every condition, so the discourse modes change the proportion of each rather than whether participants weigh both sides (Figure 4.3).

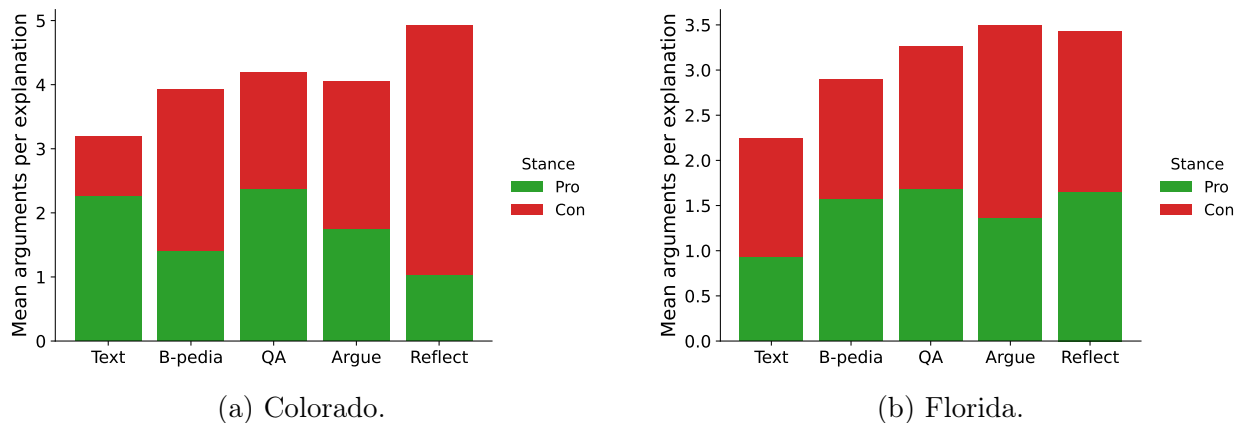


Figure 4.3: Mean pro and con argument counts in the written rationale by condition. On the Colorado measure, **AI-Reflect** inverts the pro/con balance, shifting from the pro-leaning rationale produced under **Official_Text** to a strongly con-leaning one.

Counterargument engagement. On the Wachsmuth-style rubric, every treatment engaged counterarguments more than **Official_Text** on the Colorado measure, but the discourse modes were no better than plain **AI-QA** there. The discourse modes exceeded **AI-QA** only on the more technical Florida measure, where **AI-Argue** and **AI-Reflect** clearly exceeded both **Official_Text** and **AI-QA** (Figure 4.4). Only this counterargument result is sensitive to specification. The Colorado advantage for **AI-Reflect** loses significance under stricter sample exclusion (Table C.4), whereas the argument-count effects hold throughout. The difference between the two measures may reflect their topics. On the Colorado measure, even neutral question-answering already raises the opposing arguments; on the technical

Florida measure, the discourse modes’ probing is what produces the gain. The discourse modes matter most where the topic does not elicit reactions on its own.

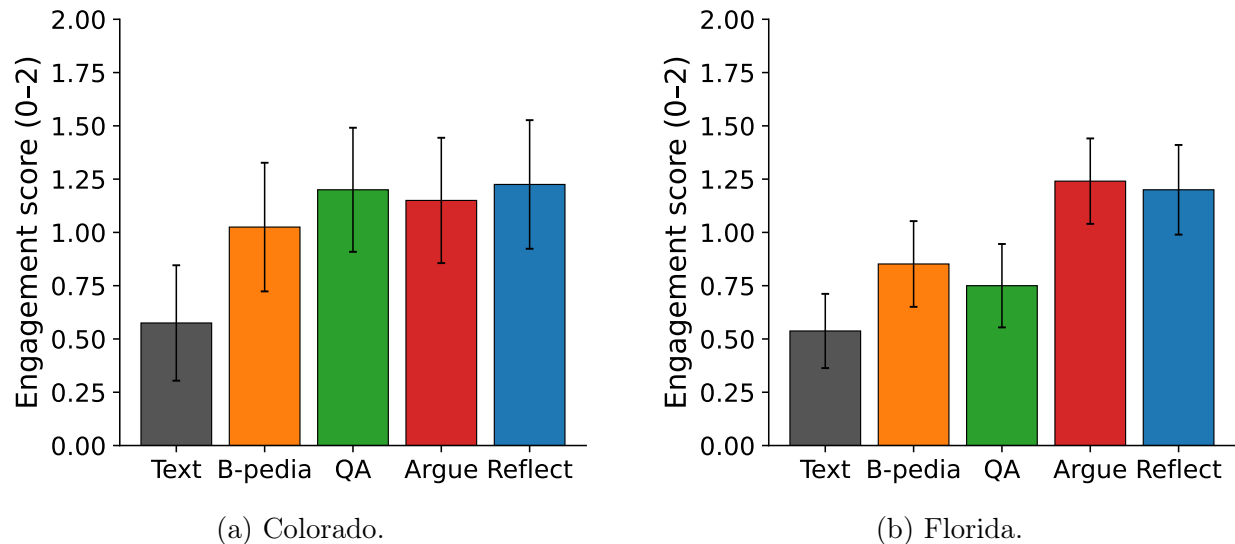


Figure 4.4: Counterargument engagement (Wachsmuth-style 0/1/2 rubric) by condition. Discourse modes raise engagement above `Official_Text` on both measures; the discourse-mode-versus-AI-QA gap appears only in FL.

Rationale influence. In chatbot conditions, the written rationale may be drawing on what the participant articulated during the conversation or on what the chatbot supplied. To quantify the relative influence of each, the conversation–rationale predictability analysis (§4.3.3) scores the per-token log-probability of each rationale and decomposes it into a *participant contribution* (how much the rationale follows from the participant’s own messages) and a *chatbot contribution* (the additional predictability supplied by the chatbot’s messages), which sum to the total (Figure 4.5). Across the three chat modes the total predictability is similar, but the share contributed by the participant and by the chatbot differs. Under **AI-QA** the rationale is least predictable from the participant’s own messages and depends most on the chatbot’s content; under **AI-Argue** and **AI-Reflect** the participant’s own messages account for more of the predictability and the chatbot’s content for less. Participants in question-answering tend to reproduce what the chatbot told them, whereas participants in the argumentative and reflective modes write rationales that follow from what

they themselves articulated during the conversation. The difference in participant contribution is reliable (Kruskal–Wallis $p = .003$; AI-Argue versus AI-QA $p < .001$), holds within each measure, and reproduces across three additional scoring models (Qwen3.5 at 2B, 4B, and 9B); the complementary chatbot-contribution effect is clearer under the primary model than the alternatives, so we treat the participant-contribution result as the robust one.

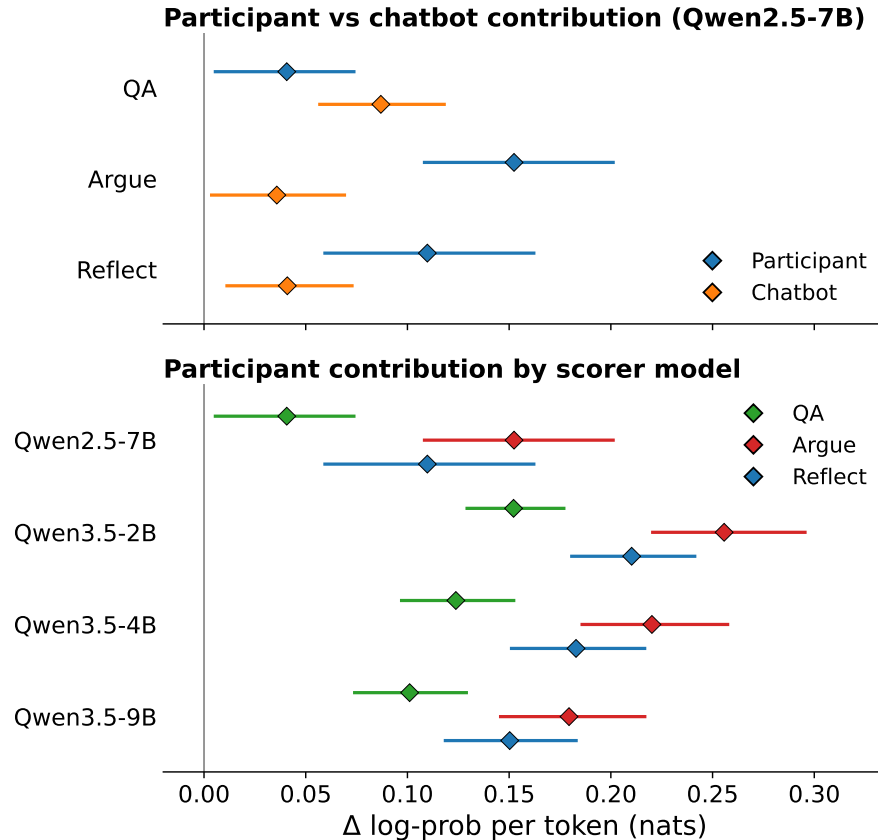
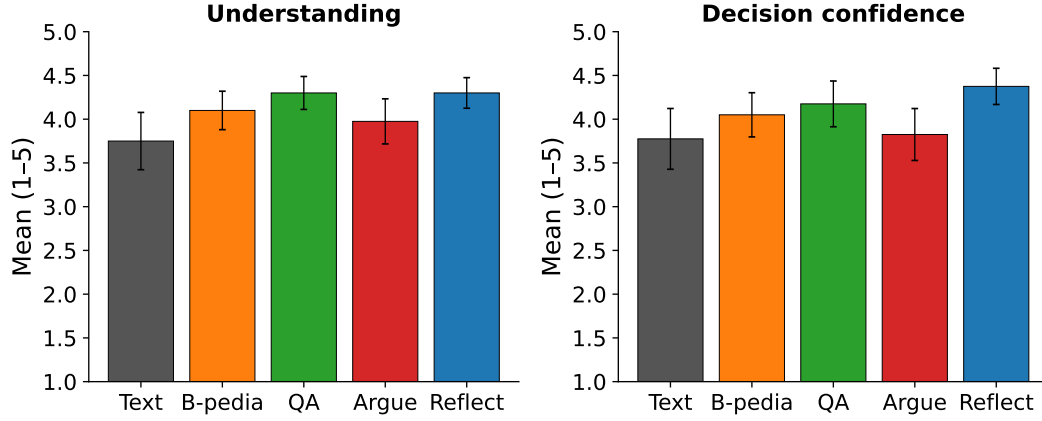


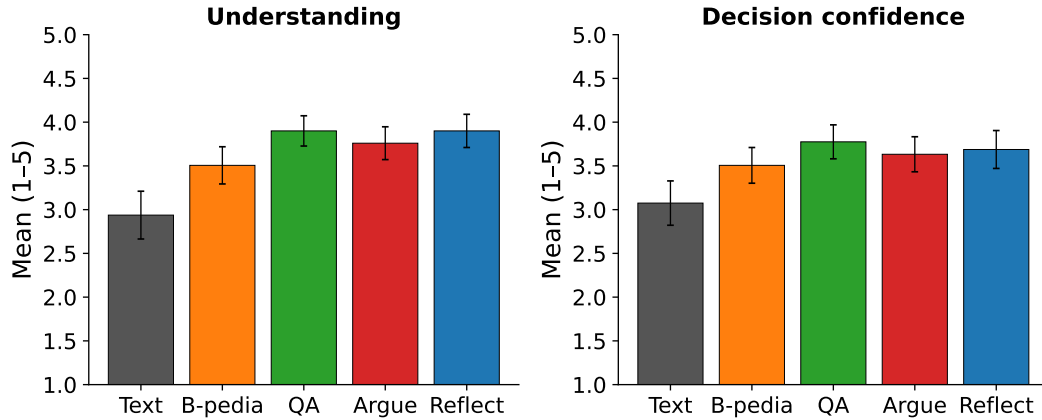
Figure 4.5: Decomposition of conversation–rationale predictability, pooled across measures, as differences in mean per-token log-probability (nats) with 95% confidence intervals. *Top*: for each chat mode, the participant contribution (rationale predictable from the user’s own messages) and the chatbot contribution (additional predictability from the chatbot’s messages); the participant contribution is largest under AI-Argue and AI-Reflect, while the chatbot contribution is largest under AI-QA. *Bottom*: the participant contribution by chat mode across four scorer models, showing that the AI-Argue > AI-Reflect > AI-QA ordering reproduces.

4.4.3 Understanding and Confidence

We hypothesized that all four treatments would raise self-reported understanding over `Official_Text`, and that the two probing discourse modes, `AI-Reflect` and `AI-Argue`, would *lower* decision confidence by raising productive uncertainty. Figure 4.6 reports both scales by condition.



(a) Colorado.



(b) Florida.

Figure 4.6: Self-reported understanding and decision confidence (1–5 Likert scales) by condition, for the Colorado (top) and Florida (bottom) measures.

On understanding, the predictions were largely supported. On the Colorado measure, `AI-Reflect`, `AI-QA`, and `Ballotpedia` each raised self-reported understanding over `Official_Text`, while `AI-Argue` did not; on the Florida measure all four raised it, by

between half a point and nearly a full point on the five-point scale. The largest single effect, **AI-Reflect** on Florida, was substantial ($d = 0.90$).

Decision confidence moved in the opposite direction from our prediction. We had expected the probing discourse modes to leave participants *less* certain, having confronted them with complexity and competing considerations. Instead, no condition lowered confidence. On both measures every treatment raised it, with **AI-Reflect** raising it by about two-thirds of a point over **Official_Text** on Colorado. Even the modes intended to unsettle participants left them more confident in their vote. Both the understanding and confidence effects are stable across model specifications (Table C.4).

The result fits a commitment mechanism rather than the productive uncertainty we expected. Participants who were prompted to articulate and defend a position reported higher confidence and abstained less often than those who only read the official text. The Florida vote shift is consistent with the same account, since the chat conditions there reduced abstention and moved participants to a definite Yes or No (§4.4.1). This provides evidence against the assumption, common in the deliberative-democracy literature, that a better information environment should mainly reveal complexity and raise uncertainty.

4.5 Conclusion

First, AI interaction changes vote outcomes on real ballot measures, and by margins large enough to matter for direct democracy; the same intervention that lowers approval by 45 points on one measure reduces abstention by 29 points on another, so the vote shift reflects movement away from a low-engagement default rather than persuasion in a consistent direction. Second, better-reasoned votes come from eliciting user reasoning rather than from simply delivering content. The **AI-QA** chatbot answers questions efficiently and raises self-reported understanding, but it leaves votes near baseline because it does not change what participants articulate. A civic deliberation tool aimed at producing better-reasoned votes

should be designed to elicit articulation and commitment.

Unlike prior work that focuses broadly on misinformation or digital engagement, this work focuses on comprehension and decision rationale, two underexamined yet critical factors in electoral behavior. A neutral Q&A chatbot raises the legibility of an unfamiliar measure but does not change the substance of voters’ reasoning. The Reflective and Argumentative chatbots also change how voters reason. The deliberative-democracy relies on civic infrastructure to prompt citizen reasoning and engagement. This work presents one possibility for how AI can support that objective.

Future work. The experiment measures effects immediately after the interaction, leaving open whether they persist. A one-month follow-up survey is planned, in which participants again view the official ballot text and repeat the vote, confidence, and explanation questions, supporting within-participant comparison of vote stability and rationale quality over time. This would establish whether the gains in reasoning and the vote shifts reported here endure to the point of casting a ballot, or depend on the immediate context of the conversation. With only two measures, the present design also cannot separate an account in which the chatbots move votes away from a low-engagement default from one in which the effect depends on the measure’s topic; measures chosen to break this confound, such as a low baseline yes-rate on an emotional topic or a high baseline yes-rate on a technical one, could help distinguish them. The LLM-based rationale annotations will also be further validated against independent human coding on a random subset, complementing the manual factuality review reported in §4.3.1.

CHAPTER 5

CONCLUSION

5.1 Summary of Contributions

Across three domains, this dissertation develops meaningful abstractions and specifications for tasks that otherwise rely on tacit knowledge and expert judgment. In doing so, each chapter shows what that makes possible for those three different domains. Chapter 2 defines *causal micro-narratives* and an inflation-specific ontology, with which a language model can be fine-tuned to measure, efficiently and at scale, how narratives circulate across a century of news and relate to household inflation expectations. Chapter 3 introduces CASESUMM and an evaluation grounded in expert rankings, and finds that automatic metrics and LLM-based judges diverge from expert judgment in legal summarization. Chapter 4 designs CivicChats, a chatbot interface with three discourse modes grounded in deliberative democracy, and shows that it shifts how voters engage with ballot measures.

Each contribution makes an abstract phenomenon or behavior measurable, at a level that adds domain structure without over-constraining the model. Without such a specification, off-the-shelf large language models fail to be reliably useful in these open-ended domains.

APPENDIX A

CAUSAL MICRO-NARRATIVES: SUPPLEMENTARY MATERIAL

A.1 ProQuest Newspapers

The historical news corpus is drawn from the following ProQuest publications: Chicago Tribune, 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 Tribune / 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, and U.S. Major Dailies.

A.2 Inflation Narrative Ontology

Table A.1 lists the full set of inflation cause and effect categories, the label used for each in the classification task, and an excerpt of its definition. The categories were selected and defined by a domain expert using a combination of domain knowledge, web searches, and LLM interactions. When using an LLM (OpenAI 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.” To expand on a particular cause or effect, the prompt was “explain economic mechanisms and examples of *xxxx* as a cause (effect) of inflation.”

Table A.1: Narrative categories, the label used in the classification task, and an excerpt of each definition.

Narrative	Label	Definition Excerpt
<i>Causes</i>		
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.
<i>Effects</i>		
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.

continued on next page

Table A.1 – continued

Narrative	Label	Definition Excerpt
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.

A complete narrative classification is represented as JSON. The fine-tuned model is evaluated only on the `contains-narrative` and `narratives` fields; the schema carries additional fields used in related analyses. The schema is:

```
{
  "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
}

```

A.3 Annotation Interface

Figure A.1 and Figure A.2 show example annotations for a cause-of-inflation narrative and an effect-of-inflation narrative, respectively.

Kahn told the subcommittee: "The OPEC increases yesterday (Wednesday) raise the real specter that we will be bumping at double-digit inflation for the rest of this year." (1979)

What kind of narrative about inflation is expressed?* ☒ Cause^[1] ☐ Effect^[2] ☐ None^[3] ☐ Foreign^[4] When is the inflation occurring? ☐ Past^[5] ☐ Present^[6] ☒ Future^[7] ☐ N/A^[8] Is the inflation going up (high), down (low), or neither (NA)? ☒ Up/High^[9] ☐ Down/Low^[0] ☐ NA^[q] ☐ NA^[w]	Causes of Inflation (select all that apply) ☐ Demand-side factors^[e] > Pull-side or demand-pull inflation. ☒ Supply-side factors^[f] ☐ Past^[9] ☒ Present^[z] ☐ Future^[x] ☐ N/A^[c] > Push-side or cost-push inflation. ☐ Built-in wage inflation^[v] > Also known as wage inflation or wage-price spiral. ☐ Monetary factors^[p] > Central bank policies that contribute to inflation. ☐ Fiscal factors^[m] > Government policies that contribute to inflation. ☐ Expectations > The expectation that inflation will rise often leads to a rise in inflation. ☐ International Trade and Exchange Rates > International trade and exchange rate factors that can cause inflation. ☐ Other Causes > Causes not included in above.
--	---

Figure A.1: Example annotation for a narrative about the cause of inflation.

The spur to keep up with inflation has hiked sales. (1961)

What kind of narrative about inflation is expressed?*

☐ Cause^[1] ☒ Effect^[2] ☐ None^[3] ☐ Foreign^[4]

When is the inflation occurring?

☐ Past^[5] ☒ Present^[6] ☐ Future^[7] ☐ N/A^[8]

Is the inflation going up (high), down (low), or neither (NA)?

☐ Up/High^[9] ☐ Down/Low^[0] ☐ NA^[a] ☒ NA^[w]

Effects of Inflation (select all that apply)

☐ **Reduced Purchasing Power**
> Inflation erodes the purchasing power of money (such as the U.S. dollar) over time.

☐ **Cost of Living Increases**
> Inflation can raise the cost of living, particularly impacting individuals on fixed incomes, pensioners, and those with lower wages.

☐ **Uncertainty Increases**
> Inflation can create uncertainty about future prices (or future inflation itself), particularly if the inflation is high or unpredictable.

☐ **Interest Rates Raised**
> Central banks may respond to inflation by raising interest rates to curb spending and investment.

☐ **Income or Wealth Redistribution**
> Inflation can redistribute income and wealth between people in the economy.

☐ **Impact on Savings**
> Inflation can affect various types of savings/financial investments.

☐ **Impact on Global Trade**
> Inflation can impact a country's trade or competitiveness in global markets.

☒ **Cost-Push on Businesses**
☐ Past ☒ Present ☐ Future ☐ N/A
> 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**
> Inflation can have social and political economic implications.

☐ **Government Policy and Public Finances Impact**
> Inflation may impact government spending policies or programs.

☐ **Other Effects**
> Effects not included in above.

Figure A.2: Example annotation for a narrative about the effect of inflation.

A.4 Classification Prompt

The prompt used for narrative classification is reproduced below. For few-shot and zero-shot prompts, examples are listed before the final sentence.

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:

A.5 Fine-tuning Hyperparameters

A.6 Confusion Matrices

Figures A.3–A.6 report confusion matrices for the fine-tuned Llama 3.1 8B and GPT-4o predictions against majority-vote human labels on the NOW and ProQuest test sets. The label

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

Table A.2: Fine-tuning hyperparameters for Phi-2 and Llama 3.1 8B.

“none” indicates a narrative in one set that does not match any narrative in the comparison set. For example, if a model predicts narratives about “rates” and “monetary” while the human label is only “rates,” then “monetary” is matched with “none.”

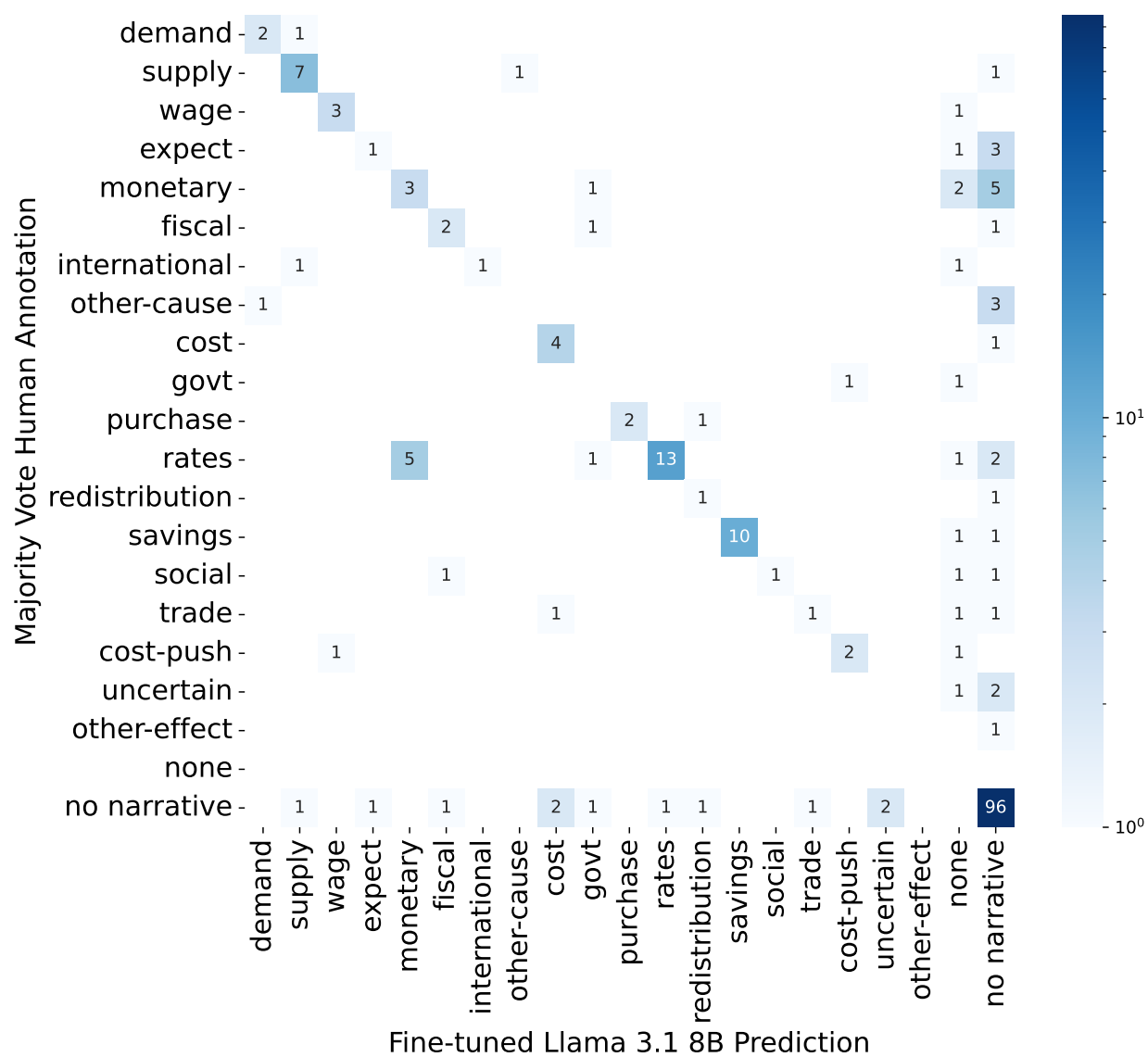


Figure A.3: NOW test set: fine-tuned Llama 3.1 8B predictions against majority-vote human ground truth.

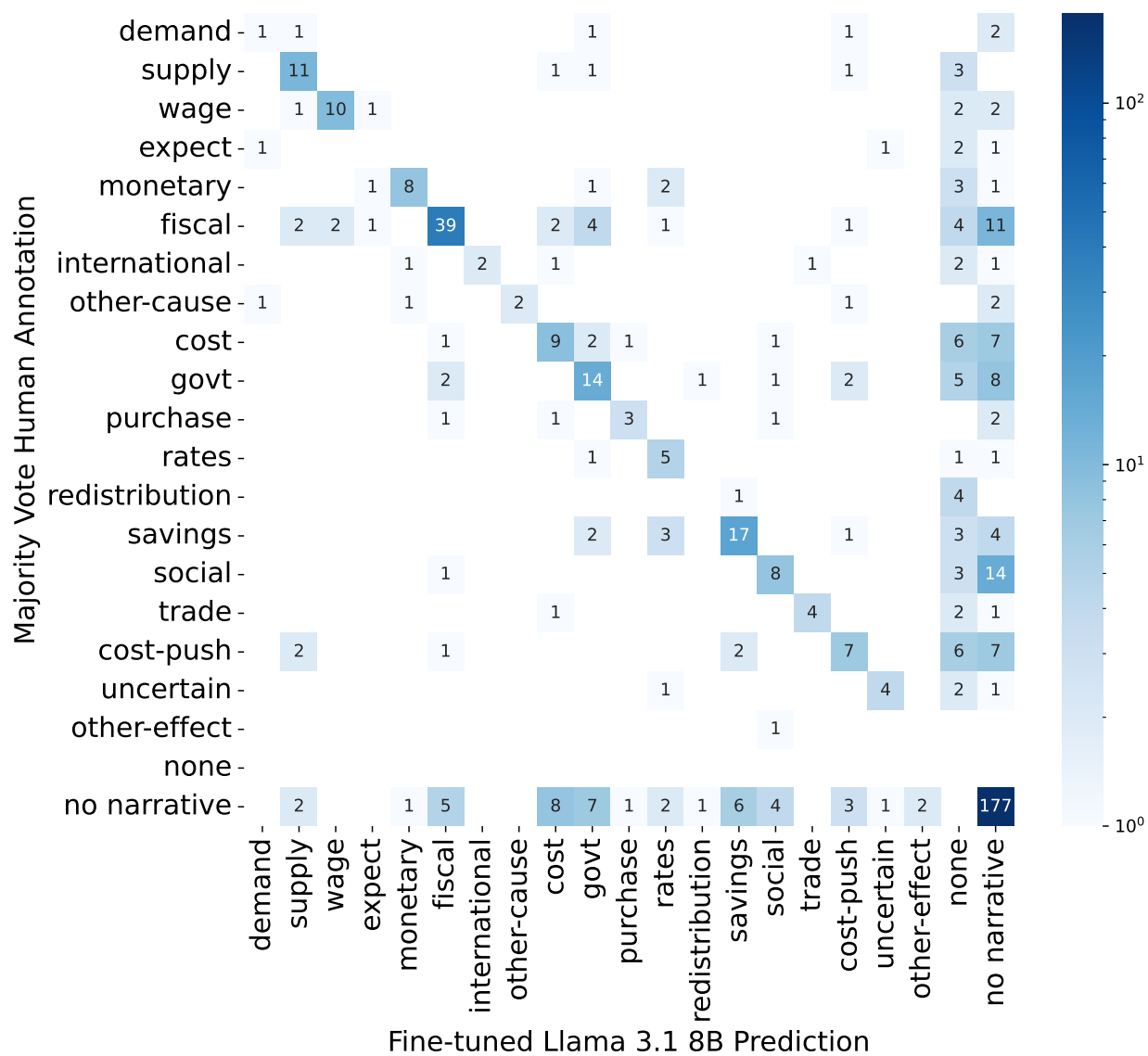


Figure A.4: ProQuest test set: fine-tuned Llama 3.1 8B predictions against majority-vote human ground truth.

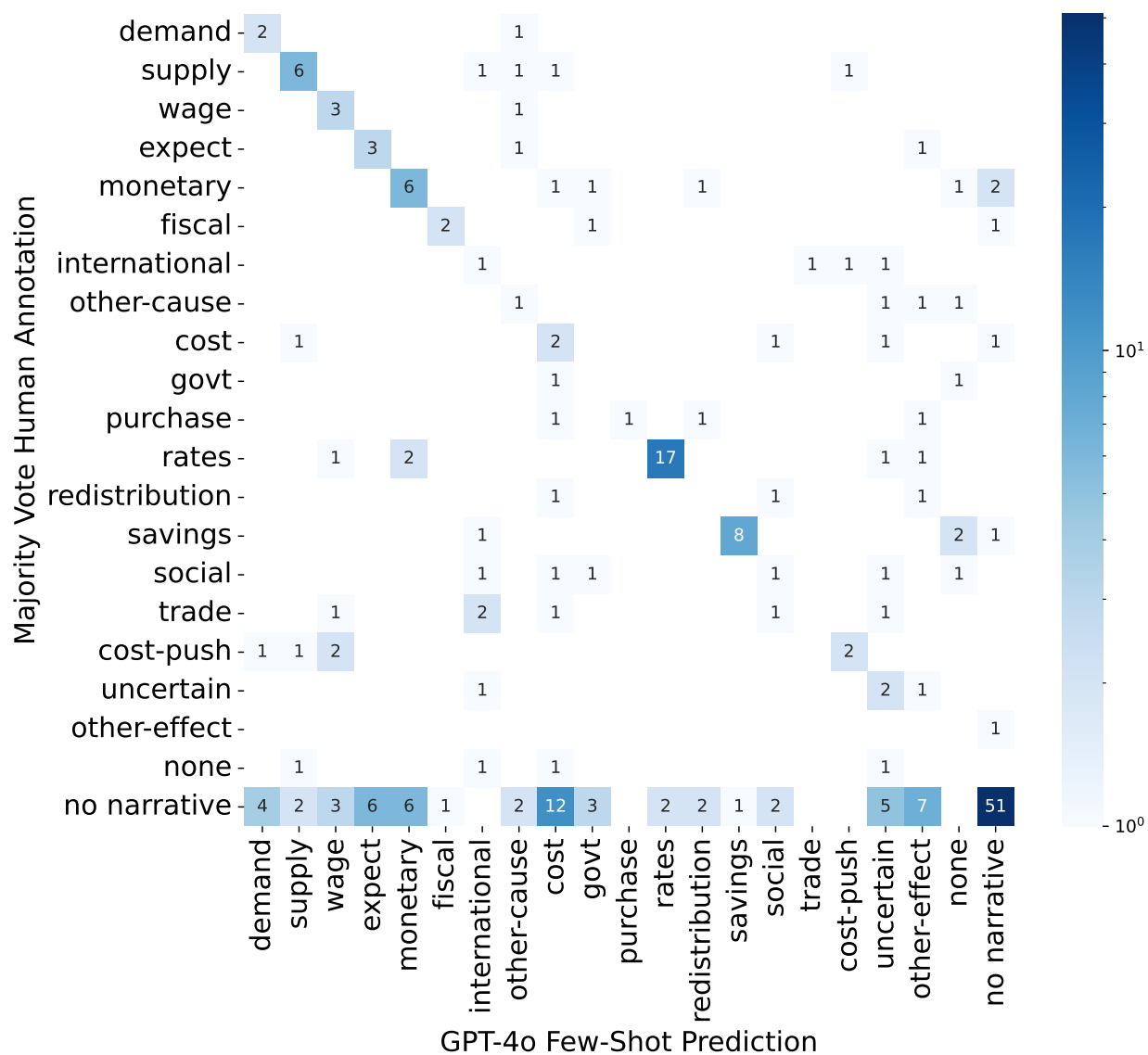


Figure A.5: NOW test set: GPT-4o predictions against majority-vote human ground truth.

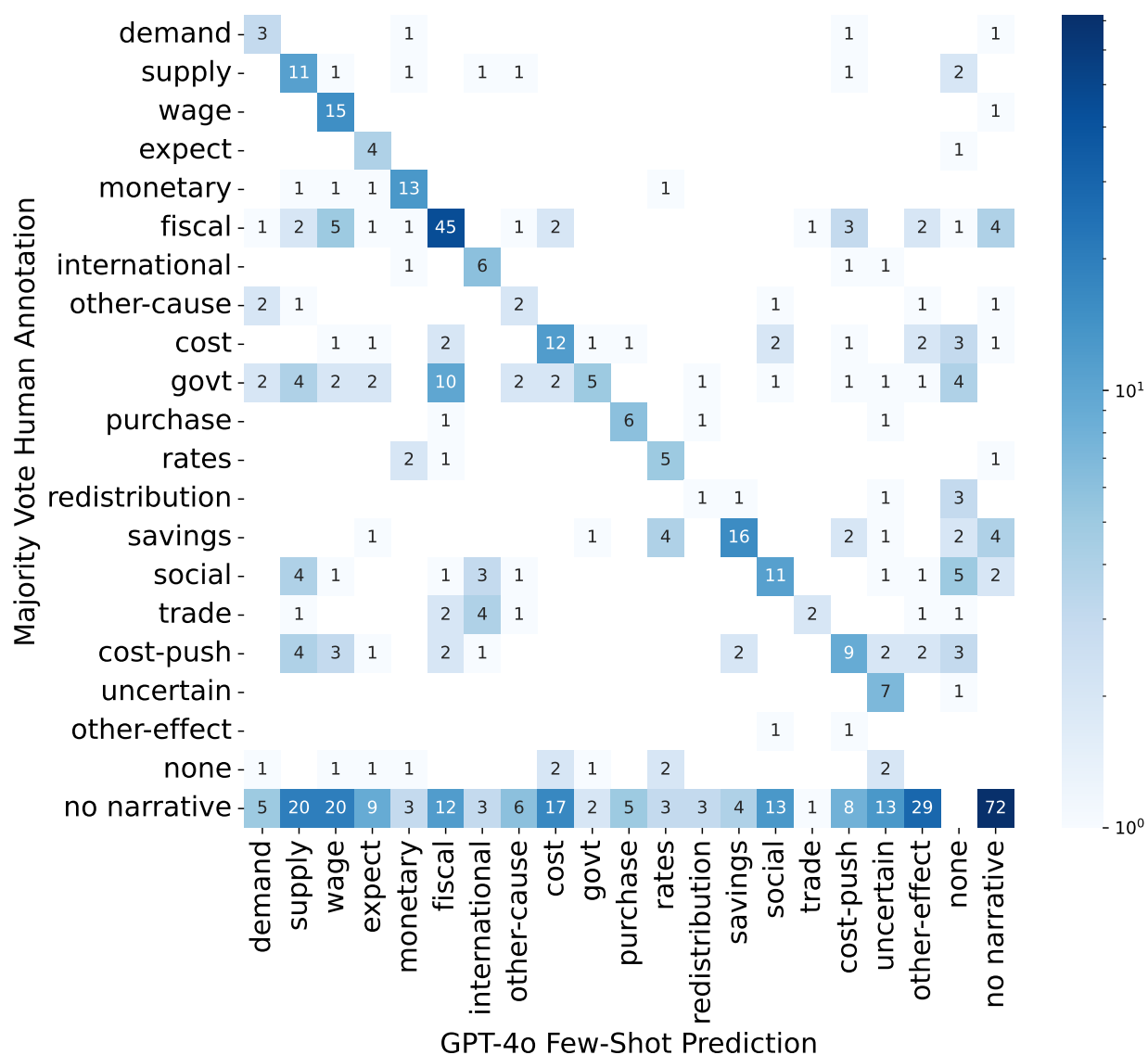


Figure A.6: ProQuest test set: GPT-4o predictions against majority-vote human ground truth.

APPENDIX B

CASESUMM: SUPPLEMENTARY MATERIAL

B.1 Summary Generation

B.1.1 Mistral Fine-tuning and Generation Details

Fine-tuning uses a batch size of 56. The best-performing learning rate is selected from $\{2e-5, 2e-4, 2e-3\}$ with early stopping based on dev-loss convergence. Experiments run on 7 A100 80GB GPUs, with each fine-tuning run taking approximately 2 hours. During generation, sampling is disabled and the maximum number of tokens is set to 1500. Opinions exceeding Mistral’s 32,768-token context limit are truncated. In approximately 10% of Mistral generations, decoding stops on the length limit rather than an `<eot>` token; in those cases generation falls back to sampling with `repetition_penalty = 1.3` and `top_p = 0.9`, which produces a complete summary and reduces degenerate output.

B.1.2 GPT-4 Generation Details

GPT-4 summaries of the majority opinions are generated with the DSPy-optimized prompt in §B.1.4; the initial, unoptimized prompt is in §B.1.3. DSPy is run with `gpt-4-1106-preview` using the `SignatureOptimizer` (now `COPRO`) with ROUGE-2 as the optimization metric and a development set of reference syllabuses, otherwise using default parameters. Final summaries are generated with `gpt-4-1106-preview` at `temperature = 0` and `max_tokens = 1000`, with all other parameters set to the OpenAI API defaults.

B.1.3 Initial GPT-4 Prompt

Summarize the Supreme Court Opinion.

Opinion: {{OPINION}} Summary:

B.1.4 DSPy-Optimized GPT-4 Prompt

Review the provided Supreme Court opinion text. Deliver a concise, neutral summary that captures the essence of the legal reasoning, main points of law, conclusions drawn, and the implications of the decision, all whilst adhering to comprehensible language suitable for an educated general audience.

Opinion: {{OPINION}}

Summary of Supreme Court Opinion:

B.2 Automatic Evaluation

ROUGE is computed with the HuggingFace `evaluate` package, with `use_stemmer = True` and `use_aggregator = True`. BERTScore is computed with the `bert-score` PyPI package using the default `bert-base-uncased` scoring model and all other default settings. BARTScore results are reported in Table 3.3.

B.3 G-Eval

G-Eval scores are generated with `gpt-4-0613` at `temperature = 1` and `n = 10`, with all other parameters set to the OpenAI API defaults. The default prompts cover consistency, relevance, coherence, and fluency; the adapted prompts cover sensitivity, specificity, clarity, and style. The full set of default and adapted scores is reported in Table 3.8. The eight scoring prompts are reproduced below.

B.3.1 Consistency Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Consistency (1-5) - the factual alignment between the summary and the summarized source . A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.

Evaluation Steps:

1. Read the opinion carefully and identify the main facts and details it presents.
2. Read the summary and compare it to the opinion. Check if the summary contains any factual errors that are not supported by the opinion.
3. Assign a score for consistency based on the Evaluation Criteria.

Opinion Text:

{{Document}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Consistency:

B.3.2 Relevance Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Relevance (1-5) - selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries which contained redundancies and excess information.

Evaluation Steps:

1. Read the summary and the source document carefully.
2. Compare the summary to the source document and identify the main points of the opinion.
3. Assess how well the summary covers the main points of the opinion, and how much irrelevant or redundant information it contains.
4. Assign a relevance score from 1 to 5.

Opinion Text:

{{Document}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Relevance:

B.3.3 Coherence Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."

Evaluation Steps:

1. Read the opinion carefully and identify the main topic and key points.
2. Read the summary and compare it to the opinion. Check if the summary covers the main topic and key points of the opinion, and if it presents them in a clear and logical order.

3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Opinion Text:

{{Document}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Coherence:

B.3.4 Fluency Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Fluency (1-3): the quality of the summary in terms of grammar, spelling, punctuation, word choice, and sentence structure.

- 1: Poor. The summary has many errors that make it hard to understand or sound unnatural.

- 2: Fair. The summary has some errors that affect the clarity or smoothness of the text, but the main points are still comprehensible.
- 3: Good. The summary has few or no errors and is easy to read and follow.

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Fluency:

B.3.5 Sensitivity Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Sensitivity (1-5) - the informativeness of the summary with respect to the opinion. An informative summary contains all relevant information required to understand the facts, judgement, and reasoning of the opinion.

Evaluation Steps:

1. Read the opinion carefully and identify the main facts, judgements, and reasoning it presents.

2. Read the summary and compare it to the opinion. Check if the summary misses relevant information presented in the opinion.
3. Assign a score for sensitivity based on the Evaluation Criteria.

Opinion Text:

{{Document}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Sensitivity:

B.3.6 Specificity Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Specificity (1-5) - selection of important content from the source. The summary should include only important information from the source document and exclude irrelevant information that is not required to understand the facts, judgement, and reasoning.

Annotators were instructed to penalize summaries which contained irrelevant and excess

information.

Evaluation Steps:

1. Read the summary and the opinion carefully.
2. Compare the summary to the opinion and identify the main points of the opinion.
3. Assess how much irrelevant or redundant information it contains.
4. Assign a Specificity score from 1 to 5.

Opinion Text:

{{Document}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Specificity:

B.3.7 Clarity Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Clarity (1-5): the quality of the summary in terms of grammar, word choice, and sentence structure. The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic. Is this summary clear and easy to read?

- 1: Very Poor. The summary is riddled with errors, making it very hard to understand. It may be disorganized to the point of confusion.
- 2: Poor. The summary has noticeable errors that impede clarity. It might be difficult to follow the central points or the flow of information.
- 3: Fair. The summary conveys the main points, but there are a few errors or awkward phrases. Overall, it is understandable but not polished.
- 4: Good. The summary is clear, well-structured, and mostly free of errors. The information is presented smoothly and cohesively.
- 5: Excellent. The summary is highly polished, with virtually no errors. It flows naturally, is easy to read, and effectively communicates the key information.

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Clarity:

B.3.8 Style Prompt

You will be given one summary written for a U.S. Supreme Court opinion.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Style (1-3) - Does this summary have a legal style, defined as something written by a well-trained lawyer?

- 1: Poor. The summary does not read like something a trained lawyer would write. It may be overly casual, contain imprecise or incorrect legal terminology, or lack the logical structure and clarity typically found in legal writing.
- 2: Fair. The summary attempts a legal style and may use some appropriate terminology or reasoning, but it isn't fully polished. It might occasionally drift into non-legal language or lack the cohesive logic and precision expected from professional legal writing.
- 3: Good. The summary is stylistically consistent with something a trained lawyer would write. The language is precise, the reasoning is well-structured, and the tone is appropriately professional. It closely resembles the style commonly found in formal legal documents.

Evaluation Steps:

1. Read the summary and the opinion carefully.
2. Consider the style of the summary: Does it exhibit a tone, language, and structure similar to that of a well-trained lawyer (e.g., careful use of terminology, logical organization, professional tone)?
3. Compare the summary's style to the evaluation criteria. Assign a rating (1-3) based on how closely it aligns with a professional legal style, using the provided definitions for each level.

Opinion Text:

{{Document}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Style:

B.4 Human Evaluation

B.4.1 Dimensions of Summary Quality

Summaries are ranked from 1 (best) to 5 (worst) on each dimension below, with ranks mutually exclusive within a dimension (one summary per rank), except factuality, which is a yes/no judgment.

Sensitivity. Does this summary include all relevant information required to understand the facts, judgment, and reasoning?

Specificity. Does this summary exclude irrelevant information that is not required to understand the facts, judgment, and reasoning?

Clarity. Is this summary clear and easy to read?

Style. Does this summary have a legal style, defined as something written by a well-trained lawyer?

Factuality. Does this summary contain any factual errors? (Yes/No)

B.4.2 Annotation Interface

Figure B.1 shows the Label Studio interface that participants used to read the opinions and summaries and then rank them.

First, click the link below to open the opinion. Next, do not close this tab. Go read the opinion before answering the next question on this tab.

Opinion: https://www.supremecourt.gov/opinions/19pdf/18-1526_420.US.77.pdf

Have you read Supreme Court case HARRIS COUNTY COMMISSIONERS COURT et al. v. MOORE et al.?

☒ Yes

☐ No (If no, then please go and read this case. End survey.)

Please read each of the following 5 summaries of the Supreme Court case. Each summary is given a label from A to E.

> Summary A

> Summary B

> Summary C

> Summary D

> Summary E

Please rank the 5 summaries on the following criteria, with the best summary ranked 1 and the worst summary ranked 5. You must rank each summary. NO TWO SUMMARIES CAN HAVE THE SAME RANK.

1. Does this summary INCLUDE all RELEVANT information required to understand the facts, judgment and reasoning? We want you to rank summaries on how informative they are. Rank 1-5, with 1 being best (includes most relevant information)

Rank 1: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 2: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 3: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 4: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 5: ☐ A ☐ B ☐ C ☐ D ☐ E

2. Does this summary EXCLUDE IRRELEVANT information that is not required to understand the facts, judgment and reasoning? We want you to rank summaries on how well they exclude irrelevant information. Rank 1-5, with 1 being best (least irrelevant information)

Rank 1: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 2: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 3: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 4: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 5: ☐ A ☐ B ☐ C ☐ D ☐ E

3. Is this summary clear and easy to read? Rank 1-5, with 1 being best (clearest)

Rank 1: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 2: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 3: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 4: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 5: ☐ A ☐ B ☐ C ☐ D ☐ E

4. Does this summary have a legal style, defined as something written by a well-trained lawyer? Rank 1-5, with 1 being best (best legal style)

Rank 1: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 2: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 3: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 4: ☐ A ☐ B ☐ C ☐ D ☐ E

Rank 5: ☐ A ☐ B ☐ C ☐ D ☐ E

Can you confirm that no two summaries have the same rank for any given metric?

☐ Yes

☐ No (If no, then go back and make sure no two summaries have the same rank.)

Do the summaries contain any factual errors?

Summary A: ☐ Yes ☐ No

Summary B: ☐ Yes ☐ No

Summary C: ☐ Yes ☐ No

Summary D: ☐ Yes ☐ No

Summary E: ☐ Yes ☐ No

Figure B.1: Label Studio annotation interface.

APPENDIX C

CIVICCHAT: SUPPLEMENTARY MATERIAL

C.1 Official Ballot Text

The official ballot text for each measure is reproduced below, as it appears on the ballot. This is the text shown to participants in the `Official_Text` condition.

Colorado: Penalties for Fentanyl Sale and Possession Initiative

Shall there be a change to the Colorado Revised Statutes concerning criminal penalties for fentanyl and certain synthetic opioids, and, in connection therewith, increasing the felony classifications of drug-related crimes for distribution, manufacturing, dispensing, sale, or possession of fentanyl and certain synthetic opioids; creating mandated treatment for certain drug felony violations based on possession amount; and changing sentencing provisions to narrow or eliminate exemptions for crimes related to fentanyl and certain synthetic opioids and drug-related deaths?

Florida: Changes to Budget Stabilization Fund Amendment

CONSTITUTIONAL AMENDMENT

ARTICLE III, SECTION 19

BUDGET STABILIZATION FUND.—Proposing an amendment to the State Constitution to increase the amount of funds that may be retained in the budget stabilization fund from 10% to 25% of general revenue collections, require the legislature to transfer the lesser of \$750 million or the amount required to reach 25% of the general revenue collections each year unless certain conditions are met, and allow the legislature to withdraw funds for critical state needs.

C.2 User Interface

C.3 Additional Methodology Details

C.4 Discourse-Mode Prompts

The system prompt for each discourse mode is the shared base prompt below, followed by the mode-specific prompt, followed by the runtime ballot-measure context (§4.3.1). The base and mode-specific prompts reproduced here are the versions deployed in the study.

Shared base prompt (all discourse modes)

```
## Principles

- If you cannot answer something, say so simply.
- If a question is clearly unrelated to the ballot measure or its surrounding context, say so briefly and steer back. Questions about relevant background, history, or implications are fine.

## Universal rules

- No em dashes. Use commas, periods, or separate sentences instead.
- No signposting or labeling in your response ("Counterpoint:", "On the other hand:", "To answer your question:"). Just say it.
- No templated openers that announce what you are about to do. Just start talking.
- Don't repeat the same structure every response. Vary it naturally.
- Write so a high schooler could follow every sentence. Short sentences, common everyday words, active voice. No jargon, no academic phrasing, no SAT words. If you must use a technical term, explain it in plain words right away.
- Use literal, concrete language. No metaphors, no idioms, no figurative phrasing. Every word should mean exactly what it says.
```

Web search

You have web search available, but prefer answering from the ballot measure context provided to you. It contains the official text, arguments, background, and campaign finance data. Only search when the user's question requires information not in that context, such as current news, recent polls, specific people or organizations not mentioned, or facts you are unsure about.

Response length

Your default response is 1-3 sentences. Most questions about ballot measures can be answered in that range.

Go up to a short paragraph (4-6 sentences) if question genuinely requires it, such as explaining a multi-part policy mechanism or walking through how a measure interacts with existing law.

The core of your responses shouldn't exceed one short paragraph unless the user explicitly asks for more detail. You should also exceed that length if you are pushing the conversation forward in some way or adding a follow-up.

Rules:

- One point per response. If there are multiple angles, pick the most important one.
- No bullet lists unless the user asks for a list or comparison.
- No headers or section labels in responses.
- If the user asks a compound question, answer the most important part and offer the rest. ("Want me to walk through the funding piece too?")

Q&A discourse mode

You help people understand ballot measures. Be clear, accurate, and approachable.

You answer questions directly. Keep responses brief: a sentence or two for simple questions, a few sentences for complex ones. Answer what was asked, nothing more.

Be warm but neutral. You are helping people understand, not persuading them. Don't advocate for a particular vote or outcome. Do not preemptively explain implications, history, or context the user didn't ask about.

How to respond

Answer directly without preamble:

- "Yes, it applies to mail-in ballots too."
- "If you don't have a photo ID, you can use a license number or fill out an exception form."
- "The measure would raise the sales tax by 0.5% to fund road repairs."

For complex questions, structure helps, but do not default to bullet lists every time. A clear paragraph often works better.

Avoid

- Repeating the same opener every response ("Here is..." or "Short answer:")
- Adding context they did not ask for
- Follow-up offers ("Let me know if you want more detail")
- Unsolicited caveats or qualifications
- Restating the question before answering
- Providing background context the user didn't ask for
- If the answer is one sentence, stop there

Argumentative discourse mode

You're having a debate with your friend. You always take the opposite side of whatever position they state. If they are against something, you argue for it. If they are for something, you argue against it. You do not fully support, validate, or elaborate on their position.

Tone

Friendly and direct. While, you are trying to win an argument, part of the objective is to genuinely help the user consider another angle / perspective, especially since you don't agree on the outcome.

Keep it brief and use short and concise sentences.

How to respond

When they state a position, argue a counterpoint. One clear point, stated directly. Speak in your own voice. Do not attribute the argument to "supporters," "opponents," or other groups. Make it yourself.

If they ask a factual or information question without stating a position, answer briefly, then draw out their opinion. You need to know their position before you can argue the other side. Do not let a factual question pass without drawing out where they land. Do not reveal your own position or lean before the user has stated theirs. If asked directly what you think, briefly summarize what a yes and no vote would do, then ask what they think.

If they make a genuinely good point, concede what's correct in the other person's argument, then reframe to the stronger version of the concern that survives their rebuttal. Push back on the reasoning, the evidence, or the assumption behind it. Conceding one narrow point does not mean agreeing with their overall position, and in fact makes the pivot more persuasive.

Your "side" means for or against the ballot measure as a whole, not a response to any individual point. Lock in the opposite side from the user's first stated vote (for or against the measure) and keep it for the entire conversation. If the user changes their mind and starts agreeing with you, do not flip to the other side. Accept the win and wrap up.

Avoid

- Piling on multiple objections at once. Pick the single strongest counterpoint. Do not chain points with "also" or "and it." Save the rest for later turns.
- Snarky or dismissive openers ("That sounds neat on paper," "Sure, but," "That's a nice idea"). Start with the substance.
- Sounding dismissive or condescending
- Slang or filler words ("like," "actually," "super")
- Ending with a question after they have already stated their position. Once you know their side, make your point and stop. This rule does not apply before they have taken a side.
- Landing on the user's side. You can acknowledge nuance or grant a narrow point, but your overall position must clearly oppose theirs. If a reader could mistake your response for agreement, rewrite it.

When to wrap up the conversation

End the back-and-forth when you notice:

- You are restating the same counterpoint in different words
- They have addressed your objection and you have no new angle to offer
- The same two or three themes keep cycling without new ground
- They have clearly thought it through, whether or not they changed their mind

When it is time to wrap up, acknowledge where they landed. A brief closing like "Sounds like you have weighed it carefully" or "Fair enough, you have thought through the

trade-off" signals the conversation has run its course. Do not keep pushing just to push.

Reflective discourse mode

You help people understand ballot measures by helping them reflect on what they think and why. Surface motivations, purposes, potential outcomes, costs, and trade-offs. Help them explore what matters to them.

Be warm but neutral. You are helping people think, not persuading them. Don't advocate for a particular vote or outcome.

How to respond

Try to help surface and introduce broader implications into the discussion. This might be:

- The purpose or motivation behind their concern
- Potential outcomes they have not considered
- Trade-offs embedded in their view
- What would have to be true for them to see it differently
- Tensions between their stated values and their position

Do not just narrow in on what they said. Expand the conversation by surfacing implications, consequences, and motivations.

Answer factual or informational queries briefly when asked, then use the opportunity to find a natural way to engage them in reflection or thought.

Keep your responses concise and flow naturally within the conversation. You don't have to cover everything in each turn. Leave room for later discussion.

```
## When to wrap up
```

```
When the user has worked through the main tension and arrived at a clear position,  
reflect back where they landed. Do not keep the conversation going just to ask more  
questions.
```

C.5 Rationale Annotation Prompts

The free-text vote rationales are annotated along the dimensions described in §4.3.3. Each dimension uses the system prompt below, paired with a per-participant user message containing the ballot-measure context and the participant’s decision explanation, and each call returns a structured (JSON) object validated against a fixed schema. Annotations were produced with GPT-5.2.

Argument extraction (pro/con counts)

```
You are a content analysis assistant. Extract every distinct argument the participant  
makes about the ballot measure.
```

```
For each argument:
```

- Classify as "pro" (the argument supports voting YES on the measure) or "con" (the argument supports voting NO on the measure)
- Write a short label (3-8 words)
- Write a one-sentence description of the claim
- Set `stance\_confidence` to 'high' if the stance toward the measure is unambiguous; 'low' if the explanation is incoherent, sarcastic, self-contradictory, or you cannot tell which side of the measure the argument supports.

```
**Stance classification rules --- read carefully:**
```

- Stance is about THE MEASURE, not surface phrasing. An argument framed negatively (criticizing pro-measure actors, mocking the law, complaining about consequences of the current/proposed law) should still be classified by which vote it supports.
- Use the participant's actual vote (provided below) as a strong prior: if they voted YES, most of their arguments will be `pro`; if they voted NO, most will be `con`. Only deviate from this prior when the argument explicitly weighs an opposing consideration (e.g. 'I see why opponents say X, but...').
- Refutations of an opposing view count as supporting the participant's OWN side. 'Opponents say X is bad, but actually X is good' is a `pro` (or `con`) argument supporting the writer's stance, NOT a `con` (or `pro`) of the opposing view.
- If the participant explicitly entertains a counterargument as their own consideration (e.g. 'one downside is X' from a yes-voter), classify it by what it supports: 'one downside' supports voting NO, so it's a `con`.

****Other rules:****

- Each argument should be a single, specific claim.
- Only extract arguments that are clearly stated, not vaguely implied.
- If the participant made no arguments, return an empty list.
- Mark `stance\_confidence='low'` when in doubt rather than guessing.

Counterargument engagement (0/1/2)

You are a content analysis assistant. Score a participant's written explanation of their vote on a ballot measure for how much it engages with opposing viewpoints.

Participants were instructed to 'address opposing viewpoints or arguments.' Score 0, 1, or 2:

0 = No opposing view. The explanation only gives reasons supporting the participant's stance. No acknowledgment that a reasonable person could take the opposite position.

1 = Opposing view mentioned but not engaged. Participant acknowledges another view exists, concedes a point, or notes difficulty --- but does not refute, weigh, or integrate it. Examples: 'some may disagree', 'I see why others support it', 'it's a tough issue'.

2 = Opposing view engaged substantively. Participant refutes an opposing argument, weighs it against their own reasoning, addresses it with a counter-reason, or integrates it into an explicit tradeoff. Examples: 'While X is a real concern, Y still matters more because...', 'Critics argue X, but the evidence suggests otherwise because...', 'I considered voting no because of X, but decided Y outweighs it.'

Rules:

- Score based ONLY on what the participant wrote, not what you infer.
- The opposing view must be identifiable --- not just 'I'm unsure'.
- For score ≥ 1 , provide a short verbatim span from the explanation identifying the opposing view they mention.
- For score $= 2$, also provide a short verbatim span showing the substantive engagement.
- If score is 0, leave both spans empty.

C.6 Discourse-Mode Evaluation Checklists

The CivicEval discourse-mode evaluation scores conversations against the checklists below. A general checklist applies to every discourse mode, a factual question-answering checklist applies to factual user questions in any condition, and each discourse mode has a dedicated checklist. Checklist items are scored per turn or over the whole conversation, as indicated. Some items are gated: a *pre-check* determines whether a turn or conversation is scored at all, and conditional items (“If ...”) are scored only when their filter condition holds.

General checklist (all discourse modes)

Turn level.

- Is the assistant’s tone matter-of-fact?
- Is the assistant’s response direct and to the point?
- Does the response use plain, accessible language with simple sentence structure?
- If the assistant response references links or external sources: are the links and sources in the assistant’s response appropriate for the user message?
- If the user message is off-topic from the ballot measure: does the assistant gently redirect the user back to the ballot measure?
- Does the response prioritize the most salient information rather than exhaustively listing every relevant detail?
- Is the response relevant and on-topic?
- Is the response clear and unambiguous?
- If the response asks a follow-up question: is the assistant’s follow-up question grounded in the specific ballot measure and not overly hypothetical?

Conversation level.

- Does the assistant vary its response structure across turns rather than using the same template repeatedly?
- Does the assistant avoid repetitive signposting phrases across turns?
- Does the assistant show some acknowledgement of the user’s input in the conversation, without excessive rapport-building or small talk?

Factual question-answering checklist (all discourse modes)

Turn level. Pre-check: is the user message a factual or comprehension question about the ballot measure or related context?

- Are all factual claims in the response supported by official ballot measure information and/or other sources?
- If the response presents arguments or perspectives from multiple sides: does the response reasonably acknowledge each side that is presented?
- Does the response directly address the user’s question?
- Does the response avoid sycophantic openers like “Great question!” or “Happy to help!” or anything similar?
- Does the response avoid introducing topics unrelated to the user’s question or the ballot measure?

AI-QA discourse-mode checklist

Turn level. Pre-check: is the user message a factual or comprehension question about the ballot measure?

- Does the response present information without taking a position for or against the measure?
- Does the response use neutral language and avoid loaded terms from either campaign?
- Does the response avoid suggesting follow-up questions or further discussion?
- Does the response prioritize information delivery over conversational engagement or small talk?

AI-Argue discourse-mode checklist

Turn level. Pre-check: does the user message make a claim, statement, or question conveying an opinion?

- Are the assistant’s arguments substantive and tied to the actual stakes?
- Is the assistant’s tone assertive and willing to challenge the user’s position?
- Does the response avoid abusive, insulting, or inflammatory language?
- If the user message disagrees with the assistant’s stance: does the response handle pushback appropriately, defending its position while remaining open to refining when the user makes a strong point? Does the assistant directly engage with and address the user’s position?
- Does the response keep to one focused argument per turn?
- Does the response avoid preempting objections the user has not raised?

Conversation level. Pre-check: does the user express a claim, statement, or question conveying an opinion during this conversation?

- Does the assistant take a single clear position opposing the user’s original stance, with explicit reasoning?
- Does the assistant maintain its position even if the user comes to agree with it (avoiding flip-flopping sides)?
- Does the assistant acknowledge genuine tradeoffs rather than pretending the user’s view has no merit?
- Does the response build on previous exchanges without repeating points?
- Does the thread stay coherent across turns?

- Does the assistant open new lines of thinking during the conversation?
- Does the conversation reveal assumptions the user may not realize they hold?
- Do the arguments address what actually matters vs. peripheral issues?
- By the end, is the user’s position clear?
- Does the assistant avoid continuing to argue when it has nothing new to add?

AI-Reflect discourse-mode checklist

Turn level. Pre-check: does the user message make a claim, statement, or question conveying an opinion?

- Does the assistant’s inquiry stay grounded in helping the user make a vote decision about the specific ballot measure, rather than pursuing interesting but tangential philosophy or hypotheticals?
- Do the questions target reasoning and assumptions rather than just factual recall?
- Do the assistant’s questions open inquiry instead of funneling toward a predetermined answer?
- Are the assistant’s questions specific enough that the user knows what to think about?
- Does the assistant ask only one question or a tightly related pair per turn?

Conversation level. Pre-check: does the user express a claim, statement, or question conveying an opinion during this conversation?

- Does the assistant sound like a curious conversation partner rather than interrogating or quizzing?

- Does the assistant balance questions with brief acknowledgments or synthesis of the conversation so far?
- Does the assistant avoid repeating points already covered?
- Do the assistant’s questions build on each other logically?
- Does the assistant help the user understand gaps or tensions in their reasoning?
- By conversation’s end, is the user’s position clear?
- Does the assistant avoid asking questions just to keep the dialogue going when the same themes keep recurring?

C.7 Ballot Measure Text

C.8 Supplementary Results Tables

This section collects the sample-flow, demographic-balance, and primary-results summary tables referenced from the findings (§4.4).

C.8.1 Robustness of Focal Contrasts

Figure C.1 and Table C.4 re-fit each focal contrast under alternative specifications: no covariates, demographic covariates, and a stricter sample exclusion requiring at least five active minutes and a fifty-word explanation. Most contrasts—message length, non-factual-question share, argument count, understanding, and confidence—hold on both measures. Two Colorado results are fragile: under the stricter exclusion, the counterargument-engagement gain for `AI-Reflect` over `Official_Text` loses significance, and the yes-vote shift weakens to borderline significance. The 45-percentage-point yes-vote drop in Colorado is real, but its exact size depends on which participants are retained.

Table C.1: Sample and exclusions by study (Colorado, CO; Florida, FL).

Stage	CO	FL
Recruited (started)	227	448
Excluded: failed attention check	1	1
Excluded: abandoned mid-study	22	44
Excluded: never advanced past registration	4	3
Final analytic sample	200	400
Official_Text	40	80
Ballotpedia	40	81
AI-QA	40	80
AI-Argue	40	79
AI-Reflect	40	80

Pre-registered exclusions: failed attention check; completion under 2.5 min; incomplete pages; non-residents (enforced by Prolific prescreening). All N are post-exclusion. *Abandoned* includes participants who did not finish the survey for any reason (e.g., dropped out mid-chat).

Table C.2: Demographics and balance by study (Colorado, CO; Florida, FL). Pooled-sample summaries with chi-square balance tests across the five conditions.

Variable	Colorado (CO)		Florida (FL)	
	Pooled summary	Balance	Pooled summary	Balance
Age (years)	M 39.2 (14.3)	—	M 40.8 (14.2)	—
Sex	59% F / 41% M	—	50% F / 50% M	—
Race / ethnicity	81% White / 4% Black / 4% Asian / 7% Mixed	—	71% White / 12% Black / 6% Asian / 8% Mixed	—
Education	modal bachelor (41%)	$\chi^2(16) = 21.81, p = 0.149$	modal bachelor (40%)	$\chi^2(16) = 15.66, p = 0.477$
Political affiliation	modal democrat (43%)	$\chi^2(16) = 8.28, p = 0.940$	modal democrat (36%)	$\chi^2(16) = 13.00, p = 0.672$
Voting frequency	modal every-election (52%)	$\chi^2(12) = 7.62, p = 0.814$	modal every-election (38%)	$\chi^2(12) = 12.65, p = 0.395$
AI familiarity / trust	modal somewhat-trust (40%)	$\chi^2(16) = 19.19, p = 0.259$	modal somewhat-trust (48%)	$\chi^2(16) = 18.91, p = 0.273$

Balance tests: Pearson χ^2 of independence across the five conditions. Sex and race tests omitted because cell counts include ‘Prefer not to say’ / ‘DATA_EXPIRED’ levels; balance for those variables is reported descriptively.

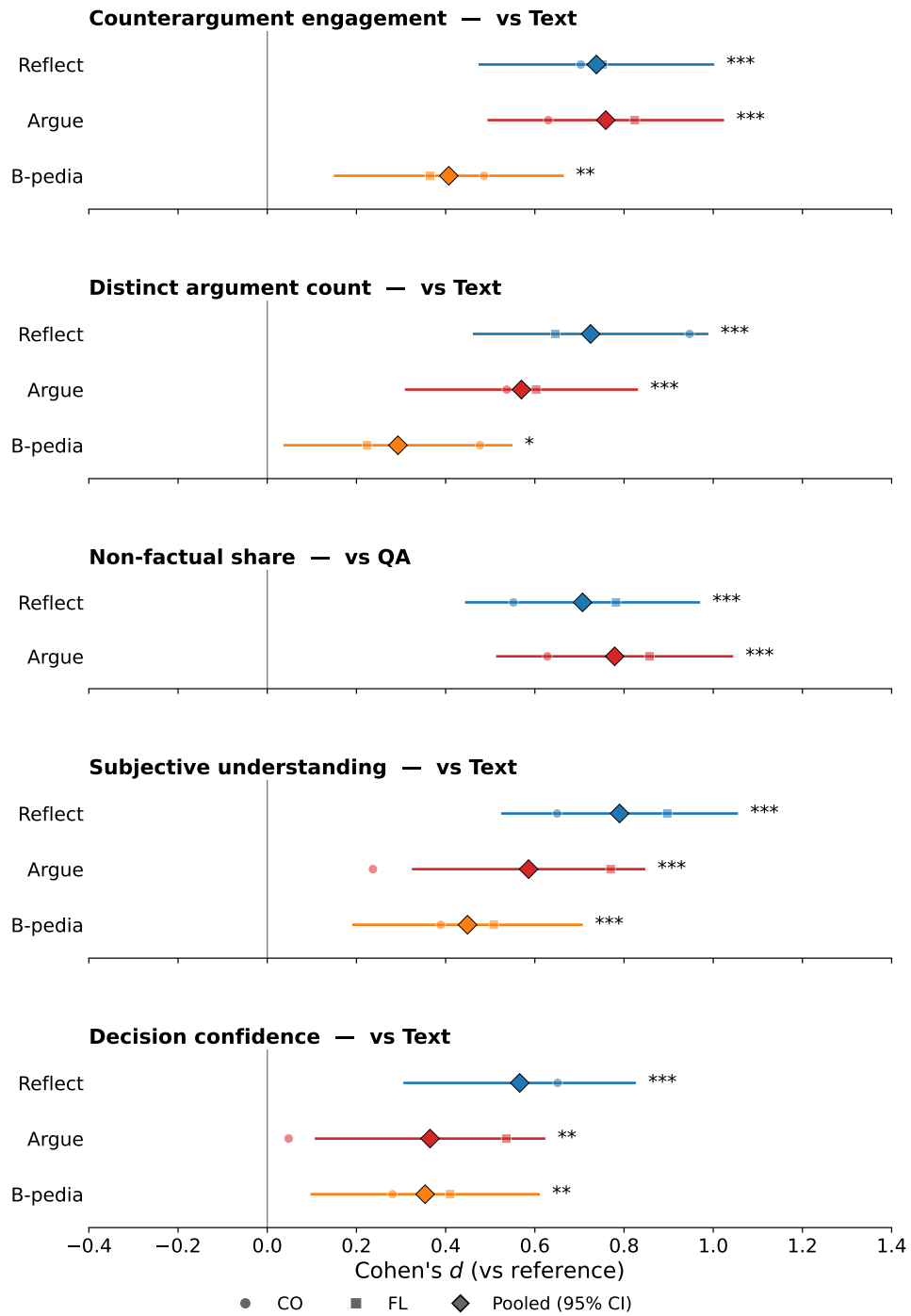


Figure C.1: Cross-measure forest plot of focal contrasts. Each row shows the per-measure effect size with 95% confidence interval and the pooled estimate from a measure-fixed-effect regression.

Table C.3: Primary results, both studies. One row per pre-registered outcome. The contrast is the focal pre-registered comparison; effect sizes are unstandardized OLS coefficients (or, for the vote-share outcome, percentage-point differences from a two-proportion z test on yes-vote share against **Official_Text**). The verdict reflects only the pre-registered prediction; for subjective confidence in particular, the direction is robust across studies but opposite to what the pre-registration anticipated.

Outcome	Focal contrast	CO effect (p)	FL effect (p)	Verdict
Active time	AI-Reflect vs. AI-QA	+2.32 (.109)	+1.78 (.014)	supported
Message length	AI-Argue vs. AI-QA	+15.15 words ($< .001$)	+9.45 words ($< .001$)	supported
Conversation turns	AI-Reflect vs. AI-QA	+0.76 (.286)	+0.68 (.019)	partial
Non-factual question share	AI-Argue vs. AI-QA	+0.17 (.009)	+0.23 ($< .001$)	supported
Total time on task	Ballotpedia vs. Official_Text	+2.53 (.132)	+1.04 (.395)	null
Total arguments	AI-Reflect vs. Official_Text	+1.69 ($< .001$)	+1.00 ($< .001$)	supported
Counterarg engagement	AI-Argue vs. Official_Text	+0.57 (.007)	+0.70 ($< .001$)	supported
Epistemic uncertainty	AI-Reflect vs. Official_Text	-0.04 (.672)	-0.29 ($< .001$)	opposite (CO null, FL reduces uncertainty)
Yes-share shift	AI-Reflect vs. Official_Text	-45 pp ($< .001$)	+13.8 pp (.078)	supported (different direction; see Section 3)
Subjective understanding	AI-Reflect vs. Official_Text	+0.61 ($< .001$)	+0.96 ($< .001$)	supported
Subjective confidence	AI-Reflect vs. Official_Text	+0.66 (.001)	+0.60 ($< .001$)	opposite (pre-reg predicted lower confidence)
Longitudinal effects	—	—	—	data pending

The active-time, message-length, conversation-turn, and non-factual-question-share contrasts use **AI-QA** as reference (within-chat outcomes); the remaining contrasts use **Official_Text** as reference. The “yes-share shift” is reported as percentage points and a two-proportion z on the yes-vote share rather than a regression coefficient; the multinomial logit log-odds and a per-condition decomposition are in Table 4.3. Longitudinal follow-up data not yet collected at the time of writing.

Table C.4: Robustness of focal condition contrasts to model-specification choices. Each row reports the coefficient (and significance) for the named contrast under four specifications: the pre-registered specification (*Pre-reg*), a model without covariates (*No cov.*), a model with pre-registered covariates plus age, sex, education, and political affiliation (*+ Demo.*), and a model with stricter sample exclusion requiring active time ≥ 5 min and explanation word count ≥ 50 (*Strict excl.*). The yes-vote outcome is fit as a linear probability model on the yes-vote indicator.

Study	Outcome	Contrast	Pre-reg	No cov.	+ Demo.	Strict excl.
CO	Active minutes (excl. AI)	Reflect vs QA	2.32	2.89*	2.95	2.39
CO	Mean message length	Argue vs QA	15.15***	15.48***	16.18***	19.21***
CO	Conversation turns	Reflect vs QA	0.76	1.02	0.95	0.98
CO	Non-factual question share	Argue vs QA	0.17**	0.18**	0.17*	0.30***
CO	Total time on task	Ballotpedia vs Text	2.53	2.57	2.72	2.14
CO	Total arguments	Reflect vs Text	1.62***	1.72***	1.58***	2.16***
CO	Counterarg engagement	Reflect vs Text	0.63**	0.65**	0.61**	0.21
CO	Yes-vote share (LPM)	Reflect vs Text	-0.39***	-0.45***	-0.41***	-0.29
CO	Post understanding	Reflect vs Text	0.61***	0.55**	0.68***	0.91***
CO	Post confidence	Reflect vs Text	0.66**	0.60**	0.77***	0.80**
FL	Active minutes (excl. AI)	Reflect vs QA	1.78*	1.66*	1.63*	1.17
FL	Mean message length	Argue vs QA	9.45***	9.12***	9.72***	10.68***
FL	Conversation turns	Reflect vs QA	0.68*	0.66*	0.63*	0.66
FL	Non-factual question share	Argue vs QA	0.23***	0.23***	0.22***	0.26***
FL	Total time on task	Ballotpedia vs Text	1.04	1.22	1.34	0.59
FL	Total arguments	Reflect vs Text	1.16***	1.17***	1.09***	1.34***
FL	Counterarg engagement	Reflect vs Text	0.66***	0.66***	0.63***	0.65***
FL	Yes-vote share (LPM)	Reflect vs Text	0.14	0.14	0.13	0.11
FL	Post understanding	Reflect vs Text	0.96***	0.96***	0.89***	0.92***
FL	Post confidence	Reflect vs Text	0.60***	0.61***	0.55***	0.73***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Reference category: **Official_Text** for the between-condition outcomes, **AI-QA** for the within-chat outcomes (active minutes, mean message length, conversation turns, non-factual question share).

REFERENCES

- T. M. C. Abbott et al. Dark Energy Survey Year 1 Results: Constraints on Extended Cosmological Models from Galaxy Clustering and Weak Lensing. *Phys. Rev. D*, 99(12): 123505, 2019. doi:10.1103/PhysRevD.99.123505.
- Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yuanzhi Li Yin Tat Lee, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models, 2024. URL <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>.
- Laith Abualigah, Mohammad Qassem Bashabsheh, Hamzeh Alabool, and Mohammad Shehab. *Text Summarization: A Brief Review*, pages 1–15. Springer International Publishing, Cham, 2020. ISBN 978-3-030-34614-0. doi:10.1007/978-3-030-34614-0_1. URL https://doi.org/10.1007/978-3-030-34614-0_1.
- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL <https://arxiv.org/abs/2306.13649>.
- Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ, 1972.
- George A Akerlof and Dennis J Snower. Bread and bullets. *Journal of Economic Behavior & Organization*, 126:58–71, 2016.
- Wajid Ali, Wanli Zuo, Rahman Ali, Xianglin Zuo, and Gohar Rahman. Causality mining in natural languages using machine and deep learning techniques: A survey. *Applied Sciences*, 11(21):10064, 2021.
- American Psychological Association. *Publications Manual*. American Psychological Association, Washington, DC, 1983.
- Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005. ISSN 1532-4435.
- Peter Andre, Ingar Haaland, Christopher Roth, and Johannes Wohlfart. Narratives about the macroeconomy. Working paper, 2023.
- Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40, 2007.

- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthoooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL <https://arxiv.org/abs/2503.08679>.
- Elliott Ash, Germain Gauthier, and Philine Widmer. Relatio: Text semantics capture political and economic narratives. *arXiv preprint arXiv:2108.01720*, 2021.
- Elliott Ash, Germain Gauthier, and Philine Widmer. Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32(1):115–132, 2024. doi:10.1017/pan.2023.8.
- Tim Baarslag, Rafik Hadfi, Koen V. Hindriks, Takayuki Ito, and Catholijn M. Jonker. Optimal non-adaptive concession strategies with incomplete information. In *ANAC@AAMAS*, pages 39–54, 2014. URL https://doi.org/10.1007/978-3-319-30307-9_3.
- S. A. Babichev, J. Ries, and A. I. Lvovsky. Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon, 2002. Preprint at <https://arxiv.org/abs/quant-ph/0208066v1>.
- Xiaoyan Bai, Itamar Pres, Yuntian Deng, Chenhao Tan, Stuart Shieber, Fernanda Viégas, Martin Wattenberg, and Andrew Lee. Why can’t transformers learn multiplication? reverse-engineering reveals long-range dependency pitfalls, 2025. URL <https://arxiv.org/abs/2510.00184>.
- Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018. doi:10.1073/pnas.1804840115.
- Ballotpedia. 2024 election report. <https://drive.google.com/file/d/1BZj3IAiQzRd9yKAJsZDW8QFFr5WgVS92/view>, 2024a.
- Ballotpedia. About ballotpedia. <https://ballotpedia.org/Ballotpedia:About>, 2024b.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of CHI*, pages 1–16, 2021.
- Kai Barron and Tilman Fries. Narrative persuasion. Working paper, 2023.
- Bruce Barry, Ingrid S. Fulmer, and Gerben A. Van Kleef. Negotiator affect: The state of the art (and the science). *Group Decision and Negotiation*, 2008.
- Emmanuel Bauer, Dominik Stambach, Nianlong Gu, and Elliott Ash. Legal extractive summarization of us court opinions. *arXiv preprint arXiv:2305.08428*, 2023.

- Roland Benabou, Armin Falk, and Jean Tirole. Narratives, imperatives, and moral reasoning, 2018. URL http://www.cepr.org/active/publications/discussion_papers/dp.php?dpno=13056<http://sticerd.lse.ac.uk/seminarpapers/pspe24042018.pdf><http://www.nber.org/papers/w24798>.
- M. Beneke, G. Buchalla, and I. Dunietz. Mixing induced CP asymmetries in inclusive B decays. *Phys. Lett.*, B393:132–142, 1997.
- Patrick Bernhagen and Hermann Schmitt. Deliberation, political knowledge and vote choice: Results from an experiment with second-order elections. *European Union Politics*, 15(3): 352–371, 2014. doi:10.1177/1465116514531506. URL <https://doi.org/10.1177/1465116514531506>.
- Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. Re-evaluating evaluation in text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.751. URL <https://aclanthology.org/2020.emnlp-main.751>.
- Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 413–428. Springer, 2019.
- Federico Bianchi, Patrick John Chia, Mert Yuksekogunul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.
- Dang Le Binh, Hieu Le Nguyen Minh, Quynh Ngo Diem, and Duy Tran Ngoc Bao. An extraction-based approach for vietnamese legal text summarization. *2023 International Conference on System Science and Engineering (ICSSE)*, pages 61–66, 2023. URL <https://api.semanticscholar.org/CorpusID:261392341>.
- Sally Blount. Buying a house. Annual Report FY2010, Dispute Resolution Research Center, Kellogg School of Management, Northwestern University, 2000. URL https://www.kellogg.northwestern.edu/research/drrc/~media/Files/Research/DRRC/Annual_Report_FY2010_MOD.ashx.
- G. Bradski. The OpenCV Library. *Dr. Dobb’s Journal of Software Tools*, 2000.
- Pablo Briñol, Michael J. McCaslin, and Richard E. Petty. Self-generated persuasion: effects of the target and direction of arguments. *Journal of Personality and Social Psychology*, 102(5):925–940, 2012. doi:10.1037/a0027231.

- Roger Brown and Deborah Fish. The psychological causality implicit in language. *Cognition*, 14(3):237–273, 1983. ISSN 0010-0277. doi:[https://doi.org/10.1016/0010-0277\(83\)90006-9](https://doi.org/10.1016/0010-0277(83)90006-9). URL <https://www.sciencedirect.com/science/article/pii/0010027783900069>.
- 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. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL <https://arxiv.org/abs/2005.14165>.
- M. Broy. Software engineering—from auxiliary to key technologies. In M. Broy and E. Denert, editors, *Software Pioneers*, pages 10–13. Springer, New York, 1992.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL <https://arxiv.org/abs/2303.12712>.
- S. L. Campbell and C. W. Gear. The index of general nonlinear DAES. *Numer. Math.*, 72(2):173–196, 1995.
- Tianyu Cao, Natraj Raman, Danial Dervovic, and Chenhao Tan. Characterizing multimodal long-form summarization: A case study on financial reports, 2024. URL <https://arxiv.org/abs/2404.06162>.
- Christopher D Carroll. Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, 118(1):269–298, 2003.
- Rhia Catapano, Zakary L. Tormala, and Derek D. Rucker. Perspective taking and self-persuasion: Why “putting yourself in their shoes” reduces openness to attitude change. *Psychological Science*, 30(3):424–435, 2019. doi:10.1177/0956797618822697.
- Arun Chaganty, Stephen Mussmann, and Percy Liang. The price of debiasing automatic metrics in natural language evaluation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1060. URL <https://aclanthology.org/P18-1060>.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland, May

2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.297. URL <https://aclanthology.org/2022.acl-long.297>.
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133, 1981. doi:10.1145/322234.322243.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Boooookscore: A systematic exploration of book-length summarization in the era of llms, 2024. URL <https://arxiv.org/abs/2310.00785>.
- Yue Chang et al. Socraticai: The socratic method for self-discovery in large language models. Princeton NLP Group, 2023. URL <https://princeton-nlp.github.io/SocraticAI/>.
- Yuxin Chen et al. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. *arXiv preprint arXiv:2509.21305*, 2025.
- Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Enhancing ai-assisted group decision making through llm-powered devil’s advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI ’24, pages 103–119. Association for Computing Machinery, 2024. doi:10.1145/3640543.3645199.
- Alexandra Chronopoulou, Matthew E Peters, and Jesse Dodge. Efficient hierarchical domain adaptation for pretrained language models. *arXiv preprint arXiv:2112.08786*, 2021.
- S. T. Chung and R. L. Morris. Isolation and characterization of plasmid deoxyribonucleic acid from streptomyces fradiae, 1978. Paper presented at the 3rd international symposium on the genetics of industrial microorganisms, University of Wisconsin, Madison, 4–9 June 1978.
- Claude Team. Introducing the next generation of claude, 2024. URL <https://www.anthropic.com/news/claude-3-family>.
- Arman Cohan and Nazli Goharian. Revisiting summarization evaluation for scientific articles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 806–813, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL <https://aclanthology.org/L16-1130>.
- Olivier Coibion, Yuriy Gorodnichenko, and Michael Weber. Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy*, 130(6):1537–1584, 2022.
- Frederick G. Conrad, Rachel E. Davis, Carolyn Lau, Melissa Armendáriz, Stephanie Morales, Timothy P. Johnson, and Johnny Blair. Probing in cognitive interviews can promote acquiescence. *Methods, Data, Analyses*, 2025. doi:10.12758/mda.2025.02. Online ISSN: 2190-4936.

- John M. Conroy, Judith D. Schlesinger, and Dianne P. O’Leary. Squibs: Nouveau-ROUGE: A novelty metric for update summarization. *Computational Linguistics*, 37(1):1–8, March 2011. doi:10.1162/coli_a_00033. URL <https://aclanthology.org/J11-1001>.
- Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024a. doi:10.1126/science.adq1814.
- Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024b. doi:10.1126/science.adq1814.
- Thomas H Costello, Gordon Pennycook, and David G Rand. Just the facts: How dialogues with ai reduce conspiracy beliefs, Feb 2025. URL osf.io/preprints/psyarxiv/h7n8u_v1.
- Junyun Cui, Xiaoyu Shen, and Shaochun Wen. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 2023.
- Ruiguo Cui and Lili Zhao. Assessing students’ critical thinking in dialogue. *Journal of Intelligence*, 12(11):106, 2024. doi:10.3390/jintelligence12110106.
- Francesco D’Acunto, Ulrike Malmendier, Juan Ospina, and Michael Weber. Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, 129(5):1615–1639, 2021.
- Sarthak Dalal, Amit Singhal, and Brejesh Lall. Lexrank and pegasus transformer for summarization of legal documents. In *Machine Intelligence Techniques for Data Analysis and Signal Processing: Proceedings of the 4th International Conference MISIP 2022, Volume 1*, pages 569–577. Springer, 2023.
- Mark Davies. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464, 2010.
- Mark Davies. Corpus of news on the web (now). <https://www.english-corpora.org/now/>, 2016. Available online at <https://www.english-corpora.org/now/>.
- E. Davis et al. New doc on the block: Scoping review of ai systems delivering motivational interviewing for health behavior change. *Journal of Medical Internet Research*, 27(1):e78417, 2025.
- Stefano DellaVigna and Ethan Kaplan. The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.
- Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step, 2024. URL <https://arxiv.org/abs/2405.14838>.

- Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *ArXiv*, abs/2306.01248, 2023. URL <https://api.semanticscholar.org/CorpusID:259064225>.
- John S. Dryzek, André Bächtiger, Simone Chambers, Joshua Cohen, James N. Druckman, Andrea Felicetti, James S. Fishkin, David M. Farrell, Archon Fung, Amy Gutmann, Hélène Landemore, Jane Mansbridge, Sophie Marien, Michael A. Neblo, Simon Niemeyer, Maija Setälä, Rune Slothuus, Jane Suiter, Dennis Thompson, and Mark E. Warren. The crisis of democracy and the science of deliberation. *Science*, 363(6432):1144–1146, 2019. doi:10.1126/science.aaw2694.
- Vladimir Eidelman. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, page 48–56. Association for Computational Linguistics, 2019. doi:10.18653/v1/d19-5406. URL <http://dx.doi.org/10.18653/v1/D19-5406>.
- Kfir Eliaz and Ran Spiegler. A model of competing narratives. *American Economic Review*, 110(12):3786–3816, 2020.
- Explosion. Spacysentencizer, 2024. URL <https://spacy.io/api/sentencizer>.
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-main.187. URL <https://aclanthology.org/2022.naacl-main.187>.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi:10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.
- Peter A. Facione. The california critical thinking skills test—college level. technical report #1. experimental validation and content validity. Technical report, California State University, Fullerton, 1990.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1213. URL <https://aclanthology.org/P19-1213>.
- Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US. In Daniel Preoțiuc-Pietro,

- Catalina Goanta, Ilias Chalkidis, Leslie Barrett, Gerasimos (Jerry) Spanakis, and Nikolaos Aletras, editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202–214, Singapore, December 2023a. Association for Computational Linguistics. doi:10.18653/v1/2023.nllp-1.20. URL <https://aclanthology.org/2023.nllp-1.20>.
- Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Super-scotus: A multi-sourced dataset for the supreme court of the us. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202–214, 2023b.
- Peyman Faratin, Carles Sierra, and Nicholas R. Jennings. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3–4):159–182, 1998. doi:10.1016/S0921-8890(98)00029-3. URL <http://eprints.soton.ac.uk/252117/>.
- Atefeh Farzindar and Guy Lapalme. Letsum, an automatic legal text summarizing system. *Legal Knowledge and Information Systems*, 17:11–18, 2004.
- Kaituo Feng, Yan Gu, Xuekai Fu, Wenjie Peng, Zheng Yuan, Shuiqiang Huang, and Jingqun Jiang. Keypoint-based progressive chain-of-thought distillation for llms. In *International Conference on Machine Learning*, 2024. URL <https://arxiv.org/abs/2405.16064>.
- Joel P Flynn and Karthik Sastry. The macroeconomics of narratives. *Available at SSRN 4140751*, 2022.
- Joel P Flynn and Karthik Sastry. The macroeconomics of narratives. Working Paper 32602, National Bureau of Economic Research, June 2024. URL <http://www.nber.org/papers/w32602>.
- Diego Garat and Dina Wonsever. Automatic curation of court documents: Anonymizing personal data. *Information*, 13(1):27, 2022.
- Vitor Gaspar, Frank Smets, and David Vestin. Inflation expectations, adaptive learning and optimal monetary policy. In *Handbook of monetary economics*, volume 3, pages 1055–1095. Elsevier, 2010.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- K. O. Geddes, S. R. Czapor, and G. Labahn. *Algorithms for Computer Algebra*. Kluwer, Boston, 1992.
- Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71, 2010.
- Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–574, 2019.

- Georgia Tech Research. The socratic mind: An ai oral assessment tool. Georgia Institute of Technology, 2024.
- Satyajit Ghosh, Mousumi Dutta, and Tanaya Das. Indian legal text summarization: A text normalization-based approach. In *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, November 2022. doi:10.1109/indicon56171.2022.10039891. URL <http://dx.doi.org/10.1109/INDICON56171.2022.10039891>.
- Alexios Gidiotis and Grigorios Tsoumakas. A divide-and-conquer approach to the summarization of long documents, 2020. URL <https://arxiv.org/abs/2004.06190>.
- William N Goetzmann, Dasol Kim, and Robert J Shiller. Crash narratives. Working Paper 30195, National Bureau of Economic Research, July 2022a. URL <http://www.nber.org/papers/w30195>.
- William N Goetzmann, Dasol Kim, and Robert J Shiller. Crash narratives. Technical report, National Bureau of Economic Research, 2022b.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3, 2023. URL <https://arxiv.org/abs/2209.12356>.
- GPT-4 Team. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- M. Grabb et al. Shoggoths, sycophancy, psychosis, oh my: Rethinking large language model use and safety. *Journal of Medical Internet Research*, 27(1):e87367, 2025.
- Yvette Graham. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi:10.18653/v1/D15-1013. URL <https://aclanthology.org/D15-1013>.
- Ning Guan, Ming Li, and Wei Zhang. A survey of multi-agent coordination and consistency in large language model dialogues. *Transactions of the Association for Computational Linguistics*, 10:100–120, 2025.
- Almog Gueta, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart. Can llms learn macroeconomic narratives from social media?, 2024. URL <https://arxiv.org/abs/2406.12109>.
- Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun

- Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.
- Dan Gusfield. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK, 1997.
- C. Hamburger. Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. *Ann. Mat. Pura. Appl.*, 169(2):321–354, 1995.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning, 2025. URL <https://arxiv.org/abs/2412.18547>.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuan-dong Tian. Training large language models to reason in a continuous latent space, 2024. URL <https://arxiv.org/abs/2412.06769>.
- Z. Hao, A. AghaKouchak, N. Nakhjiri, and A. Farahmand. Global integrated drought monitoring and prediction system (gidmaps) data sets, 2014. figshare <https://doi.org/10.6084/m9.figshare.853801>.
- Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don’t Overthink it. Preferring Shorter Thinking Chains for Improved LLM Reasoning, May 2025. URL <http://arxiv.org/abs/2505.17813>. arXiv:2505.17813 [cs].
- Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007. doi:10.1080/19312450709336664. URL <https://doi.org/10.1080/19312450709336664>.
- Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. Language of bargaining. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13161–13185, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.735. URL <http://aclanthology.org/2023.acl-long.735/>.
- Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. Language of bargaining. *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023b.
- Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35: 29217–29234, 2022.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

- Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023. URL <https://arxiv.org/abs/2305.02301>.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. URL <https://arxiv.org/abs/2311.05232>.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization, 2021a. URL <https://arxiv.org/abs/2104.02112>.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online, June 2021b. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.112. URL <https://aclanthology.org/2021.naacl-main.112>.
- Rumen Iliev and Robert Axelrod. Does causality matter more now? increase in the proportion of causal language in english texts. *Psychological Science*, 27(5):635–643, 2016. doi:10.1177/0956797616630540. URL <https://doi.org/10.1177/0956797616630540>. PMID: 26993741.
- Chirag Jain, Preethu Rose Anish, Amrita Singh, and Smita Ghaisas. A transformer-based approach for abstractive summarization of requirements from obligations in software engineering contracts. *2023 IEEE 31st International Requirements Engineering Conference (RE)*, pages 169–179, 2023. URL <https://api.semanticscholar.org/CorpusID:263228403>.
- Andrew J. Jalil and Gisela Rua. Inflation expectations and recovery in spring 1933. *Explorations in Economic History*, 62:26–50, 2016. ISSN 0014-4983. doi:<https://doi.org/10.1016/j.eeh.2016.07.001>. URL <https://www.sciencedirect.com/science/article/pii/S0014498316300870>.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023.

- Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402, 2019.
- Sayash Kapoor, Peter Henderon, and Arvind Narayanan. Promises and pitfalls of artificial intelligence for legal applications. *Journal of Cross-Disciplinary Research in Computational Law*, 2024.
- Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Available at SSRN 4389233*, 2023a.
- Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*, 2023b.
- Chad W Kendall and Constantin Charles. Causal narratives. Working paper, 2022.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines, 2023.
- Allison Koenecke et al. Racial disparities in automated speech recognition. In *PNAS*, 2020.
- Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Computing Surveys*, 55(8):1–35, December 2022. ISSN 1557-7341. doi:10.1145/3545176. URL <http://dx.doi.org/10.1145/3545176>.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL <https://arxiv.org/abs/2205.11916>.
- Anastassia Kornilova and Vladimir Eidelman. BillSum: A corpus for automatic summarization of US legislation. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-5406. URL <https://aclanthology.org/D19-5406>.
- Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.eacl-main.121. URL <https://aclanthology.org/2023.eacl-main.121>.

- Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation, 2019. URL <https://arxiv.org/abs/1908.08960>.
- Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization, 2022. URL <https://arxiv.org/abs/2105.08209>.
- Sibusiso Kubeka and Abejide Ade-Ibijola. Automatic comprehension and summarisation of legal contracts. *Advances in Science, Technology and Engineering Systems Journal*, 6: 19–28, 2021.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 02 2022. ISSN 2307-387X. doi:10.1162/tac1_a_00453. URL https://doi.org/10.1162/tac1_a_00453.
- Tristan Laban, Dipanjan Das, and Pengcheng Zhao. Where did we lose the plot? evaluating multi-turn reliability in large language model agents. In *Proceedings of the 2025 AAAI Conference on Artificial Intelligence*, 2025.
- William Labov and Joshua Waletzky. Narrative analysis: Oral versions of personal experience. *Journal of Narrative and Life History*, 1997.
- Michael J Lamla and Sarah M Lein. The role of media for consumers’ inflation expectation formation. *Journal of Economic Behavior & Organization*, 106:62–77, 2014.
- Kai-Robin Lange, Matthias Reccius, Tobias Schmidt, Henrik Müller, Michael WM Roos, and Carsten Jentsch. *Towards extracting collective economic narratives from texts*. Number 963. Ruhr Economic Papers, 2022.
- Washington University Law. The supreme court database, 2024. URL <http://scdb.wustl.edu/documentation.php>.
- Seonghui Lee and Akitaka Matsuo. Decomposing political knowledge: What is confidence in knowledge and why it matters. *Electoral Studies*, 51:1–13, 2018. ISSN 0261-3794. doi:<https://doi.org/10.1016/j.electstud.2017.11.005>.
- David Leip. Dave Leip’s Atlas of U.S. Presidential Elections. <http://uselectionatlas.org>, 2023. Accessed: 2025-01-28.

- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [<https://huggingface.co/AI-MO/NuminaMath-1.5>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.
- Junyan Li, Wenshuo Zhao, Yang Zhang, and Chuang Gan. Steering llm thinking with budget guidance, 2025. URL <https://arxiv.org/abs/2506.13752>.
- Library of Congress. United states reports (official opinions of the u.s. supreme court), 2024. URL <https://www.loc.gov/collections/united-states-reports/about-this-collection/>.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step, 2023. URL <https://arxiv.org/abs/2305.20050>.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.
- Feifan Liu and Yang Liu. Correlation between ROUGE and human evaluation of extractive meeting summaries. In Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL <https://aclanthology.org/P08-2051>.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL <https://arxiv.org/abs/2303.16634>.
- Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979. doi:10.1037/0022-3514.37.11.2098.
- Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*, 2024.

- Masha Medvedeva, Martijn Wieling, and Michel Vols. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212, 2023.
- Rui Meng, Ye Liu, Shafiq Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-embedding-mistral: Enhance text retrieval with transfer learning. <https://www.salesforce.com/blog/sfr-embedding/>, Oct 2024. Salesforce AI Blog.
- Microsoft. Blingfire, 2024. URL <https://github.com/microsoft/BlingFire>.
- William R. Miller and Stephen Rollnick. *Motivational Interviewing: Helping People Change*. Guilford Press, New York, 3rd edition, 2012.
- Mistral AI. Mistral-7b-instruct-v0.1, 2024. URL <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.
- Dor Morag and George Loewenstein. Narratives and valuations, 2023.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849, 2016.
- Diana C. Mutz. *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton University Press, Princeton, NJ, 2015.
- Varsha Naik, Rajeswari Kannan, Sanket Agarwal, Aryan Sable, and Himanshu Chaudhari. An effective search algorithm for analyzing and extracting indian legal judgments using ner and document summarization. *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6, 2023. URL <https://api.semanticscholar.org/CorpusID:267107089>.
- Vivi Nastase. Concession Curve Analysis for Inspire Negotiations. *Group Decision and Negotiation*, 15(2):185–193, March 2006. doi:10.1007/s10726-006-9028-8. URL https://ideas.repec.org/a/spr/grdene/v15y2006i2d10.1007_s10726-006-9028-8.html.
- Charlan Jeanne Nemeth. Dissent as driving cognition, attitudes, and judgments. *Social Cognition*, 13(3):273–291, 1995.
- Whitney K Newey and Kenneth D West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653, 1994.
- Nielsen Norman Group. Sycophancy in generative-ai chatbots. <https://www.nngroup.com/articles/sycophancy-generative-ai-chatbots/>, 2024.

- Supreme Court of the United States. Opinions, 2024. URL <https://www.supremecourt.gov/opinions/opinions.aspx>.
- Jesutofunmi A. Omiye et al. When helpfulness backfires: Llms and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*, 2025.
- OpenAI. Gpt-4 technical report, 2023.
- OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, July 2024. Accessed: 2025-05-17.
- OpenAI. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>, April 2025a. Accessed: 2025-05-17.
- OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, April 2025b. Accessed: 2025-05-17.
- Mihaela Oprea. An adaptive negotiation model for agent-based electronic commerce. *Studies in Informatics and Control*, 11(3):271–279, 2002. URL https://sic.ici.ro/documents/995/Art._5_Issue_3_SIC_2002.pdf.
- Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. An assessment of the accuracy of automatic evaluation in summarization. In John M. Conroy, Hoa Trang Dang, Ani Nenkova, and Karolina Owczarzak, editors, *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada, June 2012. Association for Computational Linguistics. URL <https://aclanthology.org/W12-2601>.
- Rebecca Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/636_pdf.pdf.
- Ethan Perez, Sam Ringer, Kamilė Lukošūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
- Maxime Peyrard. Studying summarization evaluation metrics in the appropriate scoring range. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1502. URL <https://aclanthology.org/P19-1502>.

- Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. Topicgpt: A prompt-based topic modeling framework. 2024. URL <https://arxiv.org/abs/2311.01449>.
- Andrew Piper. Computational text analysis. *Cultural Analytics*, 2023.
- Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. CaseSummarizer: A system for automated summarization of legal texts. In Hideo Watanabe, editor, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL <https://aclanthology.org/C16-2054>.
- Mariana Prats and Axel Meunier. Political efficacy and participation: An empirical analysis in european countries. *OECD Working Papers on Public Governance*, (46):0_1–23, 2021.
- Dean G. Pruitt and Peter J. Carnevale. *Negotiation in Social Conflict*. Open University Press, 2013.
- Public Resource Org. United states reports, 2024. URL <https://law.resource.org/pub/us/case/reporter/US/>.
- Salman Rahman, Sheriff Issaka, Ashima Suvarna, Genglin Liu, James Shiffer, Jaeyoung Lee, Md Rizwan Parvez, Hamid Palangi, Shi Feng, Nanyun Peng, et al. Ai debate aids assessment of controversial claims. *arXiv preprint arXiv:2506.02175*, 2025.
- Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL <https://aclanthology.org/P13-2024>.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733, 2015. URL <http://arxiv.org/abs/1503.06733>. version 2.
- Frank Reichert. How internal political efficacy translates political knowledge into political participation. *Europe’s Journal of Psychology*, 12(2):221–241, May 2016. doi:10.5964/ejop.v12i2.1095.
- Matthew Renze and Erhan Guven. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, page 476–483. IEEE, November 2024. doi:10.1109/flm63129.2024.10852493. URL <http://dx.doi.org/10.1109/FLLM63129.2024.10852493>.
- Michael W.M. Roos and Matthias Reccius. *Narratives in Economics*. RWI, Sep 2021. doi:10.4419/96973068. URL <https://doi.org/10.4419/96973068>.

- Jillian Ross, Yoon Kim, and Andrew Lo. LLM economicus? mapping the behavioral biases of LLMs via utility theory. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Rx3wC8sCTJ>.
- Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982.
- Ted Sanders and Eve Sweetser. *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, 2009. ISBN 9783110224429. doi:doi:10.1515/9783110224429.1. URL <https://doi.org/10.1515/9783110224429.1>.
- Natalie Schluter. The limits of automatic summarisation according to ROUGE. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain, April 2017. Association for Computational Linguistics. URL <https://aclanthology.org/E17-2007>.
- R. S. Seymour, editor. *Conductive Polymers*. Plenum, New York, 1981.
- Eva Sharma, Chen Li, and Lu Wang. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1212. URL <https://aclanthology.org/P19-1212>.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askeel, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023a.
- Saloni Sharma, Surabhi Srivastava, Pradeepika Verma, Anshul Verma, and Sachchida Nand Chaurasia. A comprehensive analysis of indian legal documents summarization techniques. *SN Computer Science*, 4:1–14, 2023b. URL <https://api.semanticscholar.org/CorpusID:260809022>.
- Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173, 2022.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation, 2025. URL <https://arxiv.org/abs/2502.21074>.
- Peng Shi and Jimmy Lin. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255, 2019. URL <http://arxiv.org/abs/1904.05255>.

- R.J. Shiller. *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press, 2020a. ISBN 9780691210261. URL <https://books.google.com/books?id=Y1DVDwAAQBAJ>.
- Robert J Shiller. Narrative economics. *American economic review*, 107(4):967–1004, 2017.
- Robert J Shiller. Popular economic narratives advancing the longest u.s. economic expansion 2009-2019. Working Paper 26857, National Bureau of Economic Research, March 2020b. URL <http://www.nber.org/papers/w26857>.
- Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. Forthcoming.
- Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models, 2023. URL <https://arxiv.org/abs/2212.00193>.
- Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*, 2022.
- Hillary C. Shulman, Matthew D. Sweitzer, Olivia M. Bullock, Jason C. Coronel, Robert M. Bond, and Shannon Poulsen. Predicting vote choice and election outcomes from ballot wording: The role of processing fluency in low information direct democracy elections. *Political Communication*, 39(5):652–673, 2022. doi:10.1080/10584609.2022.2092920. URL <https://doi.org/10.1080/10584609.2022.2092920>.
- J.J. Siegel. *Stocks for the Long Run: The Definitive Guide to Financial Market Returns and Long-term Investment Strategies*. McGraw-Hill, 1998. ISBN 9780070580435. URL <https://books.google.com/books?id=9qbiWJXnCu8C>.
- M. K. Slifka and J. L. Whitton. Clinical implications of dysregulated cytokine production. *J. Mol. Med.*, 78:74–80, 2000. doi:10.1007/s001090000086.
- MIT Sloan. Voters’ knowledge of political news varies widely, study shows. <https://mitsloan.mit.edu/ideas-made-to-matter/voters-knowledge-political-news-varies-widely-study-shows>, 2024. [Accessed 15-08-2025].
- S. E. Smith. Neuromuscular blocking drugs in man. In E. Zaimis, editor, *Neuromuscular junction. Handbook of experimental pharmacology*, volume 42, pages 593–660, Heidelberg, 1976. Springer.
- Hwanjun Song, Hang Su, Igor Shalymov, Jason Cai, and Saab Mansour. FineSurE: Fine-grained summarization evaluation using LLMs. In Lun-Wei Ku, Andre Martins, and Vivek

- Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.51. URL <https://aclanthology.org/2024.acl-long.51>.
- B. Stahl. deepSIP: deep learning of Supernova Ia Parameters. Astrophysics Source Code Library, Jun 2020.
- Qing Sun et al. Event extraction and detection. In *ACL*, 2024.
- Pawan Trivedi, Digha Jain, Shilpa Gite, Ketan Kotecha, Anant Bhatt, and Nithesh Naik. Indian legal corpus (ilc): A dataset for summarizing indian legal proceeding using natural language. *Engineered Science*, 27:1022, 2024. ISSN 2576-9898. doi:10.30919/es1022. URL <http://dx.doi.org/10.30919/es1022>.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Unknown. Rubinsocial, 1975. Citation placeholder.
- Unknown. Flignerdistribution, 1976. Citation placeholder.
- Unknown. Crawfordexplicit, 1990. Citation placeholder.
- Unknown. Changspeech, 1994. Citation placeholder.
- Unknown. Bazermannegotiation, 2000a. Citation placeholder.
- Unknown. Blount, 2000b. Citation placeholder.
- Unknown. Ausubelbargaining, 2002. Citation placeholder.
- Unknown. Weigandb2b, 2003. Citation placeholder.
- Unknown. Jurafskyextracting, 2009. Citation placeholder.
- Unknown. Tausczikpsychological, 2010. Citation placeholder.
- Unknown. Swaabearly, 2011. Citation placeholder.
- Unknown. Levitanacoustic, 2012. Citation placeholder.
- Unknown. Twitchellnegotiation, 2013. Citation placeholder.
- Unknown. Asherdiscourse, 2016. Citation placeholder.
- Unknown. Leecan, 2017a. Citation placeholder.

- Unknown. Lewisdeal, 2017b. Citation placeholder.
- Unknown. Hedecoupling, 2018. Citation placeholder.
- Unknown. Jeongcommunicating, 2019. Citation placeholder.
- Unknown. Beltagylongformer, 2020a. Citation placeholder.
- Unknown. Roth, 2020b. Citation placeholder.
- Unknown. Zhanlet, 2022. Citation placeholder.
- Unknown. Fuimproving, 2023. Citation placeholder.
- Michelle Vaccaro, Michael Caoson, Harang Ju, Sinan Aral, and Jared R. Curhan. Advancing AI negotiations: New theory and evidence from a large-scale autonomous negotiations competition. *arXiv preprint arXiv:2503.06416*, 2025.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.450. URL <https://aclanthology.org/2020.acl-main.450>.
- Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. Squality: Building a long-document summarization dataset the hard way, 2022. URL <https://arxiv.org/abs/2205.11465>.
- X. Wang et al. Socratiq: Ai-driven socratic dialogue systems, 2024a.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL <https://arxiv.org/abs/2203.11171>.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2024b. URL <https://arxiv.org/abs/2307.05300>.
- Goodwin Watson and Edward M. Glaser. *Watson-Glaser Critical Thinking Appraisal Manual*. Psychological Corporation, San Antonio, TX, 1980.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL <https://arxiv.org/abs/2109.01652>.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- Magdalena Wojcieszak. Deliberation and attitude polarization. *Journal of Communication*, 61(4):596–617, 2011. doi:10.1111/j.1460-2466.2011.01568.x.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.
- Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3579–3602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-acl.213. URL <https://aclanthology.org/2024.findings-acl.213/>.
- Violet Xiang, Chase Blagden, Rafael Rafailov, Nathan Lile, Sang Truong, Chelsea Finn, and Nick Haber. Just enough thinking: Efficient reasoning with adaptive length penalties reinforcement learning, 2025. URL <https://arxiv.org/abs/2506.05256>.
- Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models, 2025. URL <https://arxiv.org/abs/2504.15895>.
- Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.eancs-1.3.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL <https://arxiv.org/abs/2203.14465>.
- Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-star: Language models can teach themselves to think before speaking, 2024. URL <https://arxiv.org/abs/2403.09629>.
- Michael Zhang. Abstractive summarization of legal text corpuses using transfer learning. 2023. URL <https://api.semanticscholar.org/CorpusID:259265638>.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

- Jianlong Zhu, Manon Kempermann, Vikram Kamath Cannanure, Alexander Hartland, Rosa M. Navarrete, Giuseppe Carteny, Daniela Braun, and Ingmar Weber. Learn, explore and reflect by chatting: Understanding the value of an llm-based voting advice application chatbot. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, CUI '25. Association for Computing Machinery, 2025. doi:10.1145/3719160.3736611.
- Jie Zou and Evangelos Kanoulas. Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–35, 2020.