OP-FED: Opinion, Stance, and Monetary Policy Annotations on FOMC Transcripts Using Active Learning

Alisa Kanganis
Williams College
abk2@williams.edu

Katherine A. Keith
Williams College
kak5@williams.edu

Abstract

The U.S. Federal Open Market Committee (FOMC) regularly discusses and sets monetary policy, affecting the borrowing and spending decisions of hundreds of millions of people. In this work, we release OP-FED, a dataset of 1044 human-annotated sentences and their contexts from FOMC transcripts that captures monetary policy stance—specifically, whether an individual FOMC member expresses support for tightening or loosening policy. We faced two major technical challenges in dataset creation: imbalanced classes—we estimate fewer than 8% of sentences express a non-neutral stance toward monetary policy—and inter-sentence dependence—65% of instances require context beyond the sentence-level to annotate. To address these challenges, we developed a five-stage hierarchical schema to isolate aspects of opinion, monetary policy, and stance toward monetary policy, as well as the level of context needed. Second, we selected instances to annotate using active learning, approximately doubling the number of positive instances across all schema aspects. Using OP-FED, we found a top-performing, closed-weight LLM achieves 0.80 zero-shot accuracy in opinion classification but only 0.61 zero-shot accuracy classifying stance toward monetary policy—below our human baseline of 0.89. We expect OP-FED to be useful for future model training, confidence calibration, and as a seed dataset for future annotation efforts.

1 Introduction

The twelve voting members of the United States' Federal Open Market Committee (FOMC) cast votes approximately every six weeks to decide the target federal funds rate. These decisions impact hundreds of millions of people—affecting economic conditions from mortgage rates to loan availability—and ultimately influence unemployment and inflation levels. Yet, there remain open questions about the internal decision making of the FOMC, e.g., majority dynamics, consensus building, influence of the Chairman, partisan preferences, etc. (Chappell Jr et al., 2004).

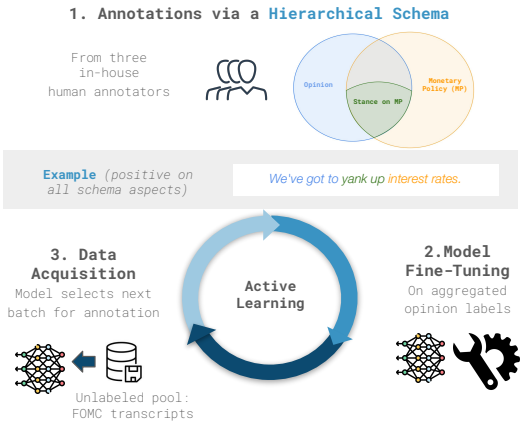


Figure 1: **Overview of OP-FED dataset creation**. **(1)** In batches, humans annotate sentences from FOMC transcripts using our hierarchical schema. **(2&3)** We fine-tune a pre-trained model on the annotated batch, and acquire a new batch via an AL acquisition function.

Although there is a large body of prior work that analyzes FOMC communications, many empirical studies are limited by their use of dictionary-based or bag-of-words approaches (Acosta and Meade, 2015; Mazis and Tsekrekos, 2017; Ruman, 2023; Bordo et al., 2024; Aruoba and Drechsel, 2024). Other work is limited by their focus on certain FOMC texts—such as press releases and meeting minutes—which are carefully curated by the FOMC prior to their public release (Shah et al., 2023).

In contrast, our work focuses on verbatim meeting transcripts—which the FOMC releases after a five-year delay—because studies suggest the transcripts contain economic signals not present in the FOMC's other released communications. In particular, Meade (2005) found the committee members' formal votes were almost always consensual de-

Findings of the Association for Computational Linguistics: ACL 2026, pages 676–699

spite their notable disagreements in the transcripts; and Fischer et al. (2023) estimated real-time access to FOMC transcripts could substantially increase market forecasting accuracy.

Under ideal circumstances, economists studying FOMC dynamics would have a (perfectly accurate) classifier with an utterance from a FOMC member as input and the member's stance toward monetary policy as output. For example, the utterance *"We've got to yank up interest rates"* would be classified as a positive stance toward tightening monetary policy, and the utterance *"We don't want to keep the funds rate biased upward for too long"* would be classified as a negative stance.

Prior to this work, training supervised classifiers or benchmarking large language models (LLMs) as zero-shot classifiers for this task was infeasible because an annotated dataset for individuals' stances in FOMC transcripts did not exist.[1] In this work, we conceptualize the task and create the first dataset for it. However, in doing so, we faced two major technical challenges:

Challenge I: Inter-sentence dependence. Because the FOMC verbatim transcripts contain rich, extemporaneous discourse, there is often inter-sentence dependence. For example, resolving the referent of "your proposal" in "Yes, I'm fine with your proposal"[2] requires tracing back 19 sentences across 10 utterances to "the 25 basis point increase in the funds rate," spoken by another FOMC member. We estimate 65% of transcript sentences require context beyond a single sentence to classify stances (see §3.3).

Challenge II: Imbalanced classes. In the transcripts, FOMC members discuss many topics other than policy actions, e.g., the state of the economy or wording of public communications. We estimate fewer than 8% of all transcript sentences express a non-neutral stance toward monetary policy (see §3.3). Given this low proportion and our finite annotation budget, if we sampled uniformly at random, we would likely obtain too few non-neutral instances to benchmark or train a supervised classifier.

To address these technical challenges, we (1) create a new hierarchical schema to isolate aspects of opinion, monetary policy, and stance toward monetary policy as well as the context needed beyond the sentence-level, and (2) deploy *active learning*, a model-and-human-in-the-loop technique that improves on uniform random sampling of unlabeled instances (Zhang et al., 2022). Figure 1 provides an overview of our active learning and annotation pipeline. In summary, our primary contributions are:

- We release OP-FED[3], a dataset of 1044 sentences plus their surrounding context that are human-annotated with our hierarchical schema of opinion, monetary policy, and stance toward monetary policy (§3.2).
- We evaluate the zero-shot accuracy of top-performing, closed-weight LLMs (§6). The best performing model, *Claude Opus 4.1*, achieves 0.80 accuracy classifying opinion but only 0.61 accuracy for classifying stance toward monetary policy—below our human baseline of 0.89.[4]

These results suggest OP-FED is likely a useful dataset for future model training or as a seed dataset for future annotation efforts. Next, we describe our data and schema (§3), AL pipeline (§4), and describe how economists might use our dataset in the future (§7).

2 Related Work

2.1 Analysis of FOMC Communications

Several studies have examined FOMC texts other than transcripts, such as press releases (Acosta and Meade, 2015), FOMC statements (Rohlfs et al., 2016), speeches (Zirn et al., 2015), or Federal Reserve staff documents (Aruoba and Drechsel, 2024). Other economists have attempted to analyze transcripts manually—e.g., Chappell Jr et al. (2004) inferred individual members' policy reaction functions based on their stated preferences for the federal funds rate in the transcripts. However, these manual analyses are not in a machine-readable format and do not have the scale necessary to train or benchmark NLP models. Other studies have applied NLP to FOMC transcripts using simple dictionary-based or bag-of-word approaches (Mazis and Tsekrekos, 2017; Ruman, 2023; Bordo et al., 2024).

The two works most closely related to ours are Shah et al. (2023) and Peskoff et al. (2023).

[1] To the best of our knowledge there are no machine-readable, publicly available datasets for this specific task.

[2] Spoken by Thomas Hoenig, President of Federal Reserve Bank of Kansas, 1991-2011; Transcript ID 20050630 in Chang et al. (2020)

[3] The dataset name is a wordplay on "op-ed" in newspapers.

[4] We provide our dataset and code at https://github.com/kakeith/op-fed and https://huggingface.co/datasets/kakeith406/op-fed.

Shah et al. created an annotated dataset of FOMC speeches, minutes, and press conference transcripts for hawkish-dovish classification. However, they use keywords (e.g., "inflation" and "fund rate") to filter the initial dataset, resulting in a sample that is biased by keyword filtering. Additionally, they operationalized their *dovish* label as "any sentence that indicates future monetary policy easing," e.g., *"In addition, U.S. inflation remains muted"* is classified as dovish. This is much more coarse-grained than our schema in which we extract FOMC members' explicit opinions and stances. Peskoff et al. estimate an F1 score of 0.57 for zero-shot predictions from LLMs (GPT-3 and GPT-4) using a small manually-annotated sample on the same hawkish-dovish classification task as Shah et al. Our dataset, OP-FED, differs from these existing annotated FOMC transcript datasets in several key ways: (1) we annotate the personal stances of individual members to enable more fine-grained downstream inference, (2) we incorporate context beyond the sentence-level, and (3) we address class imbalance by deploying active learning.

2.2 Opinion Mining & Stance Detection

Our schema draws inspiration from the tasks of opinion mining (Pang et al., 2008; Stoyanov et al., 2005; Wilson et al., 2017) and stance detection. Stance detection aims to identify speakers' positions toward a specific target, claim, or proposition, and stance datasets have been released for a variety of domains including U.S. politics (Mohammad et al., 2016), companies' mergers and acquisitions (Conforti et al., 2020), and judicial opinions (Gupta et al., 2025). As Allaway and McKeown (2020) discuss, there are two different operationalizations of stance detection. In the most common operationalization—*topic-phrase stance*—a text is classified with stance labels {*pro, con, neutral*} toward a topic which is typically a noun-phrase, e.g., "gun control" (Küçük and Can, 2020). In the second operationalization—*topic-position stance*—stance with labels {*agree, disagree, discuss, unrelated*} is classified between a text and a topic that is an entire position statement, e.g., *"We should disband NATO"*. Our hierarchical schema (§3.2) builds upon this latter topic-position stance operationalization.

2.3 Active Learning for NLP

There is a large literature that combines active learning (AL) with NLP; see Settles (1995); Zhang et al. (2022) for surveys. Previous work has proposed NLP-specific AL strategies such as using the pre-training loss as a proxy for uncertainty (Yuan et al., 2020), using AL after domain-adaptive pre-training (Margatina et al., 2022), or using parameter efficient fine-tuning (PEFT) within the AL loop (Jukić and Šnajder, 2023). Most closely related to our AL setting, Dor et al. (2020) found active learning with BERT models outperformed random sampling by the largest margin for datasets with imbalanced class distributions. One possible concern of using AL is that "training a successor model with an actively-acquired dataset does not consistently outperform training on i.i.d. sampled data" (Lowell et al., 2019). However, Margatina and Aletras (2023) argue that AL evaluations with simulated data from clean, curated datasets—which typically have balanced class proportions—may only be a *lower* bound to their effectiveness on real (noisy, class-imbalanced) data.

3 Data & Annotation Schema

3.1 Data & Pre-processing

We use the *FOMC Corpus* provided by the Cornell Conversational Analysis Toolkit (Zhang and Danescu-Niculescu-Mizil, 2018). This is comprised of meetings held between January 1977 and December 2008 and totals 286 transcripts and 364 distinct speakers. We used spaCy for sentence segmentation (Honnibal and Johnson, 2017). We excluded sentences from utterances with fewer than four white-space tokens (19,221 total) as these were often repetitive, discourse cues such as *"thank you"*. We also removed utterances—a single turn of a speaker in a conversation—exceeding 500 white-space tokens in length (3,955 total) as these were almost always factual briefings about economic conditions rather than substantive policy discussions. After these filters, we retained 85,328 utterances (280,975 sentences).

3.2 Five-Stage Hierarchical Schema

Consider the following motivating downstream example: an economist wants to measure when an FOMC member expresses a directional stance toward monetary policy—specifically, whether they support contractionary actions (e.g., raising rates) or expansionary actions (e.g., lowering rates) by the Federal Reserve—and how these expressed stances differ from the member's formal votes and how they change over the member's tenure.

Sentence	Schema Aspect				
	Opinion	MP	MP Ctxt.	StanceNLI	StanceNLI Ctxt.
"But the discrepancies in the numbers were even larger than we had."	no	*	*	*	*
"If the Congress feels comfortable with the arrangement, I'm certainly on board in this context."	yes	no	*	*	*
"I don't know how important all of this is."	yes	yes	-5 sentences	neutral	sentence
"That sounds like a good idea."	yes	yes	utterance	contradiction	utterance
"I would go with "B" also."	yes	yes	-5 sentences	entailment	-5 sentences
"Well, I'm clearly not in tune with the other members of the FOMC."	yes	yes	utterance	entailment	utterance

Table 1: **Challenging examples from OP-FED and their gold-standard labels.** For this table, we manually selected shorter sentences (for display reasons) and highlighted phrases that require context beyond the sentence-level to be resolved; their full context is in Table 9. Here, $*$ indicates the label is not applicable due to the schema hierarchy.

To develop an annotation schema for this motivating task, we (the authors) first conducted several rounds of qualitative exploration of the data via uniform random sampling of sentences. However, we found it difficult to simultaneously assess all aspects needed to classify a speaker's stance. We decided a **hierarchical annotation schema** could potentially reduce annotators' cognitive load and isolate individual aspects for a model to classify within our AL loop. Hierarchical annotation schemas have shown to be useful in other NLP tasks, e.g., annotating clinical notes (Gao et al., 2022), argumentative discourse structures (Stab and Gurevych, 2014), and ideology detection (Liu et al., 2024). Second, we found many sentences required inter-sentence coreference resolution (**Challenge I**); see Table 1 for examples. Short responses such as "*Me too*" also lacked sufficient context for labeling. Thus, our schema also includes the amount of context a sentence required in order to be classified.

To operationalize stance, we use natural language inference (NLI), the task of determining the logical relationship between a premise sentence and hypothesis sentence (Bowman et al., 2015; Williams et al., 2018).[5] We use the hypothesis—equivalent to the "stance position statement" (§2.2)—"*We should tighten monetary policy*" because an *entailment* label would capture all three aspects of interest: stance, monetary policy, and tightening direction (e.g., raising target rates). A *contradiction* label would capture a stance toward loosening monetary policy (e.g., lowering target rates).

In summary, our schema has annotators proceed hierarchically through the following five stages:

1. **Opinion:** Annotators first determine whether the speaker of the sentence expresses an *opinion* (toward any subject), which we define as *the expression of a subjective perspective, recommendation, or position.* Possible labels are *yes*, *no*, and *ambiguous*.
2. **Monetary Policy (MP):** If an opinion is present, the annotators then assess whether the instance referenced monetary policy tools or actions (e.g., the federal funds rate or the money supply). Possible labels are *yes*, *no*, and *ambiguous*.
3. **MP Context:** If the MP label was positive, annotators record how much contextual information is required to determine the label. Possible labels are *sentence*, *utterance*, *-5 sentences* (previous five sentences), and *-200 tokens* (previous two hundred tokens, rounding up to the nearest full sentence).
4. **StanceNLI:** Given *yes* labels for Stages 1 and 2, annotators then choose whether the target sentence *entails*, *contradicts*, is *neutral*, or is *ambiguous* in reference to the hypothesis statement *"We should tighten monetary policy."*
5. **StanceNLI Context:** Annotators then indicate the amount of context required for the NLI label. Possible labels are the same as Stage 3.

3.3 Prototype Annotation Round

To estimate class proportions, we sampled 200 sentences uniformly at random from our pool of transcript sentences, and one of the authors annotated the sentences given our five-stage schema. We

[5]NLI has been used by other computational social science work to classify events in news articles (Halterman et al., 2021) and political texts (Burnham et al., 2025). We primarily chose NLI because our annotators were familiar with the NLI task from other projects.

found for Stage 3, 54% of instances required additional context (beyond the sentence-level) to determine MP, and for Stage 5, 65% required additional context for the NLI label, confirming **Challenge I**. We also observed the following (second column; Table 2): for Stage 1, 24.5% of the sentences exhibited an opinion; for Stage 2, 13.5% contained both an opinion and a reference to MP; and for Stage 4, only 7.5% expressed a non-neutral stance on MP.

4 Active Learning Pipeline

4.1 AL for OP-FED Set-Up

Given the extreme class imbalance observed in our prototype annotation round (**Challenge II**) and our limited annotation budget, we deployed active learning (AL) with the hope of increasing the number of non-neutral samples annotated. Rather than classifying all five stages in our hierarchical schema, the model within our AL loop only classifies Stage 1 (opinions) for several reasons. First, we observed in our prototype annotation round that all opinion labels could be inferred using sentences alone (avoiding **Challenge I**). This increases the chance of AL being successful since sentence-level models typically perform better than longer context models (Mishra et al., 2021; Zhang et al., 2024). Second, we hypothesized increasing the number of positive opinion instances (the first stage in our five-stage hierarchical schema) would increase positives for all other aspects. See Table 2 for an empirical confirmation of this hypothesis. We leave to future work multi-task learning of all stages within the AL loop; see Rotman and Reichart (2022).

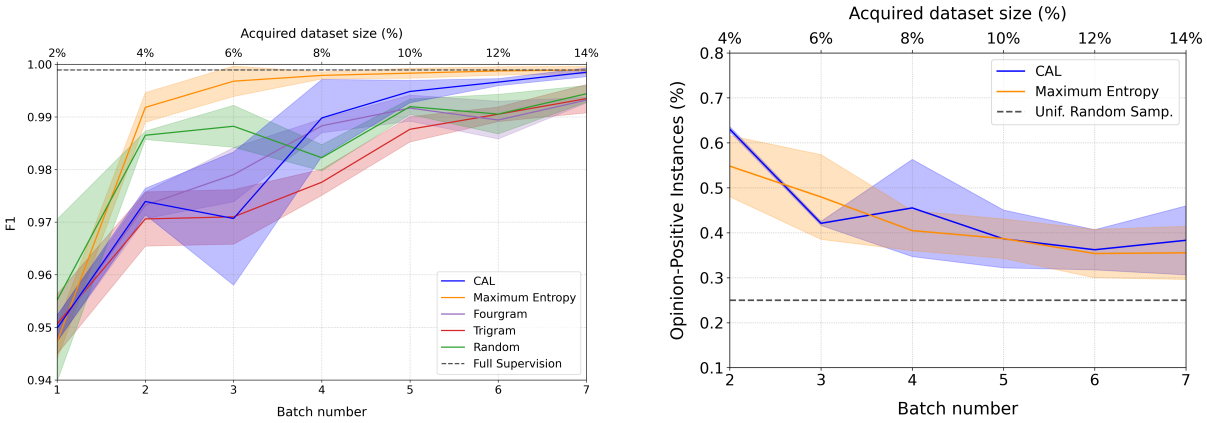
4.2 AL Overview

We implement pool-based, cold-start active learning. Following the definition of active learning provided by Margatina and Aletras (2023), let C be a corpus of text. Let $A \subset C$ and $B \subset C$ both be labeled datasets and $|A| = |B| = \mathcal{D}$ where $\mathcal{D}$ is the annotation budget. We denote a model class M and the data on which it has been trained with subscripts M_A and M_B. Let $T \subset C$ be a labeled test set. We say A is a "better" dataset if $\delta(M_A(T)) > \delta(M_B(T))$ for the chosen performance metric δ. Otherwise, B is better. In pool-based active learning, the model $\mathcal{M}$ is first trained on the initial subset of labeled data $\mathcal{D}_{\text{lab}}$. Following prior work, our annotators label a uniform random sample for the initial labeled set. Then, using the model in-the-loop, an acquisition function F samples a batch $\mathcal{Q} \subset \mathcal{D}_{\text{pool}}$ of the "best" sentences in $\mathcal{D}_{\text{pool}}$. Here, "best" is typically a trade-off between informativeness (e.g., selecting instances with the highest uncertainty or disagreement) and representativeness (e.g., selecting instances that are different from already labeled instances). $\mathcal{Q}$ is subsequently labeled by human annotators, appended to $\mathcal{D}_{\text{lab}}$, and removed from $\mathcal{D}_{\text{pool}}$. We then re-train the model $\mathcal{M}$ on the updated $\mathcal{D}_{\text{lab}}$ and we record its performance on a held-out test set $\mathcal{D}_{\text{test}}$.

AL Acquisition Functions. We implement and experiment with several acquisition functions that trade-off informativeness and representativeness. As our baseline for representativeness, we use **n-gram density** which calculates the number of unseen n-grams, normalized by the sentence length (Eck et al., 2005). For informativeness, we implement **maximum entropy** (Schröder et al., 2022) which selects "uncertain" instances, those for which the model's predictive distribution is closer to uniform. Additionally, we implement **contrastive active learning (CAL)** (Margatina et al., 2021)—a hybrid approach that targets "contrastive samples", i.e., instances that are similar in the model feature space and yet close to the model decision boundary. See §E for more details.

4.3 Models

In our pipeline, we use open-weight, encoder-only pre-trained models. Open-weight models are necessary because we need to update the model weights within the AL loop. We use encoder-only models because many of the AL acquisition functions require predicted probabilities, and extracting probabilities from generative, decoder-only models can be ad-hoc and require prompt engineering; see Tian et al. (2023). Within the AL loop, we fine-tune ModernBERT with our opinion (Stage 1) annotations. ModernBERT has outperformed other competitive encoder-only models on benchmarks like GLUE while being faster and more memory efficient at inference time (Warner et al., 2025). In each AL iteration, we fine-tune the model's weights from its pretraining checkpoint. We fix all hyperparameters because at true inference time—when we deploy the AL loop for annotators—we do not have a separate validation set so searching over hyperparameters is infeasible. Other simulation AL work found "ignoring the dev data and setting a constant number of epochs yielded qualitatively similar, albeit noisier, results" (Dor et al., 2020). See Table 11 for the full list of hyperparameters.



(a) **Acquisition functions** (colors) are compared via macro F1 on the test set of VAST+AN (y-axis) for the model trained on $\mathcal{D}_{\text{lab}}$ associated with that AL batch (x-axis).

(b) **Binary class distribution** of $\mathcal{D}_{\text{lab}}$ (y-axis) across active learning batches (x-axis).

Figure 2: **Active learning simulations** with ModernBERT and VAST+AN as $\mathcal{D}_{\text{pool}}$ with 75% *opinion-negative* and 25% *opinion-positive* labels ($n = 7459$), and held-out test set ($n = 3185$). The mean (solid line) and standard deviation (bands) is across 5 runs with different random seeds.

In our simulations, we also compare ModernBERT to RoBERTa (Liu et al., 2020) and a baseline of bag-of-words logistic regression (see §F).

4.4 AL Simulations

For AL pipelines, the "best" acquisition function is typically task- and model-dependent (Zhang et al., 2022). To decide on the acquisition function for our (one-shot) deployment, we *simulate* active learning by using an existing annotated dataset as a pool of unlabeled data and assign their existing labels as the "oracle" annotations within the AL loop, as is standard in AL (Margatina and Aletras, 2023).

Opinion Simulations. For our simulations, we use VAST, an annotated dataset of comments from the New York Times' "Room for Debate" section, where readers post short responses to policy issues or social questions (Allaway and McKeown, 2020). Each comment in VAST is labeled with respect to multiple topic phrases as having a *pro*, *con*, or *neutral* stance. To transform VAST into a binary classification task that matches our set-up at inference time (§4.1), we collapse any comment in VAST with at least one *pro* or *con* label into an *opinion* label, and collapse comments with only *neutral* labels into a *no-opinion* label.

To match the 25% positive and 75% neutral class distribution observed in our prototype annotation round, we augmented VAST with additional neutral examples of randomly selected sentences from New York Times articles (Tumanov, 2021). These augmented examples are "silver-standard" labels because they do not have direct human verification. To estimate the false negative rate, we manually reviewed a random sample of 20 instances and found only one indicated an opinion, providing sufficient confidence in the quality for our simulations. We refer to the resulting binary opinion dataset as VAST+AN, where "AN" denotes the inclusion of augmented neutrals, and divide it into a train ($n = 7459$) and test splits ($n = 3185$).

Simulation Results. We compared the following acquisition functions: maximum entropy, CAL, n-gram density with tri-grams and four-grams, and baseline of uniform random sampling. We use an AL query batch size of 150. Figure 2(a) shows that maximum entropy has both the steepest acquisition curve and reaches the level of full supervision after seeing approximately 8% of $\mathcal{D}_{\text{pool}}$, outperforming all other methods. Figure 2(b) isolates CAL and maximum entropy and shows both acquisition functions acquire a much higher percentage of stance-positive examples than the 25% we would expect from randomly sampling $\mathcal{D}_{\text{pool}}$. In Appendix F, we experiment with the following: changing the model from ModernBERT to RoBERTa, changing the batch size, changing the task from classifying opinion to classifying monetary policy, and changing the class proportions to 10%-90%. We found our choice of acquisition function was relatively robust to these changes.

We hypothesize maximum entropy works well as an acquisition function for this domain because it is simple yet effective: it relies on selecting the exam-

Aspect	Unif. Rand.		OP-FED (w/ AL)	
Opinion	24.5%	(49/200)	52.3%	(546/1044)
→MP	13.5%	(27/200)	26.1%	(272/1044)
→→ StanceNLI	7.5%	(15/200)	13.5%	(141/1044)

Table 2: **Active learning (AL) increases rates of positive labels** compared to uniform random sampling. For StanceNLI, positive labels are Entail. ∪ Contr. Here, → is a reminder that the schema is *hierarchical*.

ples that the trained model is least certain about. In an unbalanced dataset (like ours) a binary classifier is often the least certain about the minority class and thus oversamples this class.

5 OP-FED Creation

Annotator training. Prior to the annotation period, our three human annotators—all undergraduate economics majors at a private institution in the U.S.—participated in two training sessions. The first session introduced the annotation schema and codebook. In the second session, the annotators worked together to discuss and annotate 20 example sentences drawn from the prototype rounds; see §B for additional details on the annotators.

AL deployment. We then deployed our AL pipeline with maximum entropy as our acquisition function, ModernBERT as our model, an annotation budget of 1050 examples, an AL batch size of 150, and Stage 1 (opinion) as the classification task within the AL loop. On each of our seven consecutive annotation days, the three human annotators labeled the AL-selected batch of 150 instances independently and without visibility of each other's responses. Each instance was labeled beginning at Stage 1 (opinion), with later stages completed only if the relevant preceding stage received a positive label. The average annotation time per instance was 50 seconds; see Table 8 for additional statistics. For each instance and each annotation stage, we use majority vote among the three annotators to determine the final label. For opinion (Stage 1), if the majority label was *ambiguous* or if there was no majority label, the instance was not used for training in the next iteration of the AL loop.

Agreement Metrics. In Table 3, we provide descriptive statistics of the final dataset—after filtering to instances that received a majority vote—and we report chance-adjusted agreement scores of Fleiss' Kappa (κ) (Fleiss et al., 2003) and Krippendorff's Alpha (α) (Krippendorff, 2004). OP-FED achieved moderate agreement levels for Stage 1 opinion ($\kappa = 0.57$), Stage 2 MP ($\alpha = 0.62$), Stage 3 MP context ($\alpha = 0.72$), Stage 4 StanceNLI ($\alpha = 0.55$), and Stage 5 StanceNLI Context ($\alpha = 0.67$). While these agreement rates may seem low compared to simple NLP tasks, they are on par with other complex tasks in computational social science, e.g., Thalken et al. (2023) reported $\alpha = 0.63$ for classifying "grand" versus "formal" legal reasoning in U.S. Supreme Court opinions.

Unlabeled Pool, $\mathcal{D}_{\text{pool}}$		n
Num. Transcripts		286
Num. Sentences		280,975
OP-FED, Final $\mathcal{D}_{\text{lab}}$		n
1-Opinion	Yes	546
($\kappa = 0.57$)	No	496
	Ambiguous	2
2-MP	Yes	272
($\alpha = 0.62$)	No	214
	Ambiguous	1
3-MP Context	Sentence	99
($\alpha = 0.72$)	Utterance	119
	-5 sents.	20
	-200+ toks.	4
4-StanceNLI	Entailment	47
($\alpha = 0.55$)	Neutral	18
	Contradiction	94
	Ambiguous	15
5-StanceNLI Context	Sentence	24
($\alpha = 0.67$)	Utterance	111
	-5 sents.	22
	-200+ toks.	5

Table 3: **Descriptive statistics** of OP-FED—after taking the majority vote from the three human annotators—including Fleiss' Kappa (κ) and Krippendorff's Alpha (α) for this subset.

For Stages 3 and 5, most instances required at least utterance-level context (Figure 4), confirming **Challenge I**. Table 7 breaks down the Fleiss' Kappa scores for Stage 1 by AL batch and annotator pair. The agreement scores decrease with each AL batch. We hypothesize this is due to the most uncertain examples—selected by the maximum entropy acquisition function—also being the most challenging examples for human annotators.

AL analysis. Our AL pipeline successfully addressed our goal of mitigating the class imbalance (**Challenge II**). In Table 2, we find the number of positive instances roughly doubled across the three target aspects when comparing our prototype annotation round—sampled uniformly at random—to OP-FED—sampled via AL.

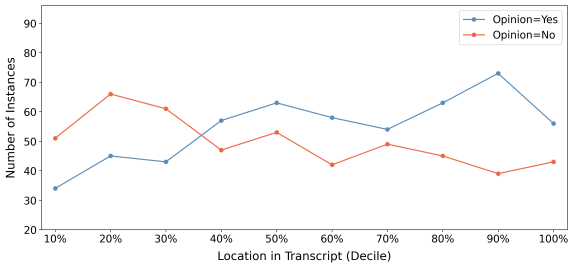


Figure 3: **Opinions are expressed later in transcripts.** Distribution of OP-FED gold-standard opinion labels by their location in FOMC transcripts.

Face Validity. To assess the *face validity*—whether measurements "look plausible" in the social sciences (Jacobs and Wallach, 2021)—of the opinion labels, Figure 3 shows that opinion-positive statements are more likely to occur later in the transcripts, with the highest being in the 80-90% decile. This pattern is consistent with the known structure of FOMC meetings—in the latter portion of the meeting, participants engage in the "policy go-around," in which each member outlines their preferred policy stance—and lends face validity to our sampling and annotation efforts.

6 Baseline Experiments

6.1 Zero-Shot LLM Performance

LLMs have achieved relatively high performance for zero-shot classification tasks, i.e., performing an unseen task with no supervised examples (Brown et al., 2020; Wei et al., 2021; Sanh et al., 2021, *inter alia*). With the creation of OP-FED, we can answer one of the motivating questions for this work: what is the zero-shot performance of LLMs for the FOMC tasks we have described with our schema?

In our experimental set-up, we subset to OP-FED instances with non-ambiguous labels.[6] We compare OpenAI's *GPT-5* and *GPT-5 nano* models, Anthropic's *Claude Opus 4.1*, and *DeepSeek-V3.1*. We use the recommended structured output format (e.g., JSON or XML) for each model; see §G for all prompting details. For comparison, we also implement baselines of predicting the majority class; four-fold cross validation accuracy (aggregating across instances in every inference fold) with a bag-of-words logistic regression model (see §D.1 for details); and estimates of the minimum and maximum human baselines, i.e., the accuracy achieved by one (out of three) annotators when compared against the aggregated gold-standard.

Results. Examining Table 4, LLMs' zero-shot accuracy is relatively high on Task 1, $0.72 - 0.80$, and Task 2, $0.73 - 0.76$, but still slightly below the maximum human baselines on the two tasks with 0.95 and 0.89 respectively. For Task 4, accuracy is relatively low across all models with the highest accuracy (0.61) from *Claude Opus 4.1*, which only slightly outperforms predicting the majority class. Notably, *DeepSeek-V3.1* on Task 4 has a lower accuracy than our bag-of-words logistic regression baseline, 0.29 versus 0.31 respectively. We hypothesize this low performance on Task 4 is because the task is linguistically challenging and often requires very long contexts; see Table 9 for examples with their full context. All LLMs perform poorly on predicting whether more context is needed (Tasks 3 and 5) with lower accuracy than both the majority class and bag-of-words logistic regression baselines. Thus, our recommendation to economists wanting to use NLP for Task 4 (our motivating downstream example) is to fine-tune LLMs on gold-standard data; see Halterman and Keith (2025) for instruction-tuning guidelines for computational social scientists.

6.2 Confidence Score Calibration

Downstream FOMC analyses likely also need *confidence scores* on individual instances in addition to class predictions; see Angelopoulos et al. (2023) for a general inference framework. We prompt the zero-shot LLMs in our experiments to generate "verbalized confidence scores"—values between 0.0 and 1.0 expressed in token-space—since Tian et al. (2023) found these are often better calibrated than model conditional probabilities. We evaluate with **expected calibration error** (ECE), which partitions the confidence score into M equally sized bins (Guo et al., 2017).[7] In Table 5, we report ECE for Tasks 1 and 2 (those with decent zero-shot accuracy), and we find the zero-shot models are moderate to poorly calibrated; *Claude Opus 4.1* achieves the lowest calibration error on Task 1 (ECE=0.09) and *GPT-5 nano* achieves the lowest calibration error on Task 2 (ECE=0.06), but error can be substantial, e.g., ECE=0.20 on Task 1 for *GPT-5*. For context, this is much higher error than

[6] This makes Task 4 a multi-class (with three classes) and the other tasks binary. For Tasks 2 and 5, we provide the models with maximum context for the target sentence: the utterance of target sentence and the 200 previous tokens to that utterance.

[7] We use $M = 20$. We leave to future work alternative calibration estimation methods (Nguyen and O'Connor, 2015).

	Model	Checkpoint	1-Opinion ($n = 1042$)	2-MP ($n = 485$)	3-MP Ctxt. ($n = 242$)	4-StanceNLI ($n = 159$)	5-Stance Ctxt. ($n = 162$)
Baselines	Majority Class	–	0.52	0.56	0.59	0.59	0.85
	BOW+LogReg	–	0.61	0.60	0.80	0.31	0.73
	Human min.*	–	0.85	0.85	0.86	0.79	0.85
	Human max.*	–	0.95	0.89	0.89	0.89	0.93
Zero-Shot	GPT-5	`gpt-5-2025-08-07`	0.72	0.74	0.28	0.47	0.28
	GPT-5 nano	`gpt-5-nano-2025-08-27`	0.75	0.73	0.33	0.46	0.28
	Claude Opus 4.1	`claude-opus-4-1-20250805`	0.80	0.74	0.24	0.61	0.29
	DeepSeek-V3.1	`deepseek-chat-2025-08-21`	0.75	0.76	0.47	0.29	0.15

Table 4: **Accuracy of baseline models.** Stages 3 and 5 are each a binary classification task for whether context is needed beyond the sentence. Here, n is the number of examples after removing instances with an *ambiguous* label. Corresponding F1 scores are in Table 10 and prompts in §G.

Model	1-Opinion	2-MP
GPT-5	0.20	0.14
GPT-5 nano	0.13	0.06
Claude Opus 4.1	0.09	0.17
DeepSeek-V3.1	0.13	0.13

Table 5: **Expected calibration error (ECE)** of each LLM's verbalized (generated) confidence scores.

Tian et al. (2023)'s findings that *GPT-3.5 Turbo*'s verbalized confidence scores can achieve ECE as low as 0.05 on standard NLP benchmarks. We view improving confidence score calibration as an area of future work where our dataset could be useful, e.g., Platt scaling (Platt et al., 1999) with our gold-standard labels.

Error analysis. In Table 13 we provide preliminary manual error analysis on some of *GPT-5*'s top miscalibrated examples for classifying OPINION. From this manual analysis, we find much of the error stems from *GPT-5*'s confidence in lexical phrases: "*I don't know*" as (almost always) not an opinion and "*I think*" as (almost always) an opinion. *GPT-5* also misclassified with high confidence subtle linguistic indications of an opinion, such as, "*we're going to have to deliver*."

7 Conclusion and Future Work

In this work, we release OP-FED, a human-annotated dataset of FOMC transcripts with hierarchical labels for opinion, monetary policy, stance toward monetary policy, as well as the context needed beyond the sentence-level. Using active learning substantially increased the yield of positive instances in our dataset. Future work could likely improve LLM accuracy and confidence score calibration by shifting to supervised fine-tuning or other iterative inference techniques. Other work could use our hierarchical schema and OP-FED as the seed dataset in a new AL loop to create an even larger dataset of instances. Finally, economists could combine our gold-standard annotations with zero-shot LLM predictions to improve economic analyses, i.e., Angelopoulos et al. (2023)'s method.

Acknowledgments

We thank Yanni Kakouris and Aidan Casey for their work on the annotations. We thank Mark Hopkins for feedback on an early draft; Kenneth Kuttner for conversations about the FOMC; and anonymous ARR reviewers for helpful suggestions. KK received support from a YI Grant from Allen Institute for Artificial Intelligence (AI2) and National Science Foundation under Grant No. 2451403. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AI2 or the National Science Foundation.

Limitations

Given that these transcripts are publicly available government data, we see very few potential risks with this project or dataset. We identify FOMC speakers by name, but these speakers are public figures whose consent is implicit given their professional roles.

We acknowledge several limitations to our work. First, given our fixed annotation budget, we successfully increased the class proportions by using active learning; however, we believe this could be improved by future work with larger annotation budgets. Second, we used three "in-house" trained annotators who were undergraduate students majoring in economics. We hypothesized undergraduates with domain training would have higher annotation quality compared to crowd-workers. However, we also hypothesize that domain-experts—

who have even greater expertise with the Federal Reserve and monetary policy—may annotate instances with even higher quality (albeit with the tradeoff of requiring greater compensation for annotations). Third, we rely on raw data scraped and processed by the Cornell Conversational Analysis Toolkit (Chang et al., 2020) whose collected transcripts end in 2008, limiting the temporal coverage of our annotation. Future work could expand this upstream dataset ingestion for more recent FOMC transcripts. Our final limitation is our use of proprietary, closed-weight LLMs on OP-FED. As Palmer et al. (2024) advise, "using proprietary language models in academic research requires explicit justification." Proprietary models lack details about their training data, which limits transparency and reproducibility. However, we use these models as baselines because they are likely to be the first choice of applied researchers (i.e., economists), so we believe it is important to benchmark their accuracy.

References

Miguel Acosta and Ellen E Meade. 2015. Hanging on every word: Semantic analysis of the fomc's post-meeting statement. Technical report, Board of Governors of the Federal Reserve System (US).

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931. Association for Computational Linguistics.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. 2023. Prediction-powered inference. *Science*, 382(6671):669–674.

S Borağan Aruoba and Thomas Drechsel. 2024. The long and variable lags of monetary policy: Evidence from disaggregated price indices. *Journal of Monetary Economics*, 148:103635.

Michael D Bordo, Klodiana Istrefi, and Humberto Martínez. 2024. Rules vs. discretion: Decoding fomc policy deliberations. Technical report, National Bureau of Economic Research.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Michael Burnham, Kayla Kahn, Ryan Wang, and Rache Peng. 2025. Political debate: Efficient zero-shot and few-shot classifiers for political text. *(Forthcoming) Political Analysis*.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60.

Henry W Chappell Jr, Rob Roy McGregor, and Todd Vermilyea. 2004. *Committee decisions on monetary policy: Evidence from historical records of the Federal Open Market Committee*. MIT press.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: an empirical study. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7949–7962.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based on n-gram coverage. In *Proceedings of Machine Translation Summit X: Papers*, pages 227–234.

Eric Fischer, Rebecca McCaughrin, Saketh Prazad, and Mark Vandergon. 2023. Fed transparency and policy expectation errors: A text analysis approach. *FRB of New York Staff Report*.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*, 3 edition. John Wiley & Sons, Hoboken, NJ.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M Churpek, and Majid Afshar. 2022. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *International Conference on Language Resources & Evaluation (LREC)*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ankita Gupta, Douglas Rice, and Brendan O'Connor. 2025. -stance: A large-scale real world dataset of stances in legal argumentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31450–31467.

Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. 2021. Corpus-level evaluation for event qa: The indiapoliceevents corpus covering the 2002 gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253.

Andrew Halterman and Katherine A Keith. 2025. Codebook llms: Evaluating llms as measurement tools for political science concepts. *Political Analysis*.

Matthew Honnibal and Mark Johnson. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–90. Association for Computational Linguistics.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Josip Jukić and Jan Šnajder. 2023. Parameter-efficient language model tuning with active learning in low-resource settings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5061–5074.

Daniel Jurafsky and James H. Martin. 2025. Vector semantics and embeddings. In *Speech and Language Processing*, chapter 6. Draft of January 12, 2025. Available at https://web.stanford.edu/~jurafsky/slp3/6.pdf.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. Sage Publications, Thousand Oaks, CA.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Songtao Liu, Bang Wang, Wei Xiang, Han Xu, and Minghua Xu. 2024. Encoding hierarchical schema via concept flow for multifaceted ideology detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2930–2942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach. *ICLR*.

David Lowell, Zachary C Lipton, and Byron C Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30.

Katerina Margatina and Nikolaos Aletras. 2023. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419. Association for Computational Linguistics.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663. Association for Computational Linguistics.

Panagiotis Mazis and Andrianos Tsekrekos. 2017. Latent semantic analysis of the fomc statements. *Review of Accounting and Finance*, 16(2):179–217.

Ellen E Meade. 2005. The fomc: preferences, voting, and consensus. *Federal Reserve Bank of St. Louis Review*, 87(March/April 2005).

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1322–1336.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598.

Alexis Palmer, Noah A Smith, and Arthur Spirling. 2024. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Denis Peskoff, Adam Visokay, Sander V Schulhoff, Benjamin Wachspress, Brandon M Stewart, et al. 2023. Gpt deciphering fedspeak: Quantifying dissent among hawks and doves. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Christopher Rohlfs, Sunandan Chakraborty, and Lakshminarayanan Subramanian. 2016. The effects of the content of fomc communications on us treasury rates. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2096–2102.

Guy Rotman and Roi Reichart. 2022. Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*, 10:1209–1228.

Asif M Ruman. 2023. A comparative textual study of fomc transcripts through inflation peaks. *Journal of International Financial Markets, Institutions and Money*, 87:101822.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203. Association for Computational Linguistics.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *CoRR*.

Burr Settles. 1995. Active learning literature survey. *Science*, 10(3):237–304.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Georgia Tech Scheller College of Business Research Paper No. 4447632.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-perspective question answering using the opqa corpus. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 923–930.

Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. 2023. Modeling legal reasoning: Lm annotation at the edge of human agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9252–9265.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Alexander Tumanov. 2021. New york times articles data. Accessed: April 2025.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Claire Cardie. 2017. Mpqa opinion corpus. In *Handbook of Linguistic Annotation*, pages 813–832. Springer.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7935–7948.

Jonathan P. Zhang and Cristian Danescu-Niculescu-Mizil. 2018. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 20th Conference on Computational Natural Language Learning*, pages 193–203. Association for Computational Linguistics.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, et al. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37:60755–60775.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190. Association for Computational Linguistics.

Cäcilia Zirn, Robert Meusel, and Heiner Stuckenschmidt. 2015. Lost in discussion? tracking opinion groups in complex political discussions by the example of the fomc meeting transcriptions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 747–753.

A Artifacts: Citations and Licenses

See Table 6 for artifact citations, versions, and licenses. The use of all artifacts is consistent with their intended use.

B Datasheets for Datasets

In this section, we report our answers to Datasheets for Datasets (Gebru et al., 2021).

B.1 Motivation for Dataset Creation

Why was the dataset created?
This dataset was created to analyze the opinions and stances of U.S. Federal Open Market Committee (FOMC) members in their verbatim meeting transcripts. Existing datasets on FOMC communications either lack speaker attribution, rely on simple dictionary-based or bag-of-words approaches, or are limited to curated sources such as press releases and meeting minutes. OP-FED addresses these limitations by annotating transcripts with a novel hierarchical annotation schema that captures opinion, monetary policy references, and stance directionality—while preserving speaker identity and contextual nuance.

What (other) tasks could the dataset be used for?
This dataset could be used to support research on group decision-making, hierarchical multi-label classification, and opinion classification and stance detection in complex policy settings.

Has the dataset been used for any tasks already?
At the time of publication, only the original paper.

Who funded the creation of the dataset?
Funding was provided via the senior author's start-up funds from their institution.

B.2 Dataset Composition

What are the instances?
Instances are individual sentences extracted from raw FOMC meeting transcripts. Each instance is annotated using a five-stage hierarchical schema that captures whether the speaker expresses (1) an opinion, (2) whether that opinion refers to monetary policy, (3) how much context (beyond the sentence-level) was needed to determine the monetary policy reference, (4) whether the sentence entails, contradicts, or is neutral toward the hypothesis "I think we should tighten monetary policy," and (5) how much context (beyond the sentence-level) was needed to assign that inference label.

Are relationships between instances made explicit in the data?
Not explicitly; while context (e.g., preceding sentences) is tracked, no graph or linkage structure is included.

How many instances of each type are there?
There are 1044 instances in OP-FED and 1050 instances in the unaggregated dataset (that is, the one that releases the labels for each of the three annotators).

What data does each instance consist of?
Each instance consists of a target sentence, associated context (including the full utterance, the five preceding sentences, and up to 200+ preceding tokens—rounded to full sentences), and labels from the five-stage hierarchical schema. Annotations are provided both as aggregate (majority vote) labels and as raw labels from each of the three annotators.

Is everything included or does the data rely on external resources?
Everything is included.

Are there recommended data splits or evaluation measures?
No official splits are provided.

Artifact	Version (date)	License	Citation
spaCy	3.8.7	MIT	Honnibal and Johnson (2017)
ConvoKit	3.4.1	MIT	Chang et al. (2020)
ModernBERT	ModernBERT-base (2024-12-19)	Apache 2.0	Warner et al. (2025)
RoBERTa	roberta-base (v1, 2019)	MIT	Liu et al. (2020)
GPT-5	`gpt-5-2025-08-07`	Proprietary (API)	No publication
GPT-5-nano	`gpt-5-nano-2025-08-27`	Proprietary (API)	No publication
Claude Opus 4.1	`claude-opus-4-1-20250805`	Proprietary (API)	No publication
DeepSeek-V3.1	`deepseek-chat-2025-08-21`	Proprietary (API)	No publication

Table 6: Latest version, citation, and license for NLP artifacts used.

What experiments were initially run on this dataset?
We experiment with zero-shot classification using closed-weight proprietary LLMs.

B.3 Data Collection Process

How was the data collected?
Raw transcripts were sourced from the ConvoKit version of the FOMC corpus. Sentences were sampled for annotation using active learning.

Who was involved in the data collection process?
The dataset was annotated by three undergraduate economics majors from a college in the United States, two of whom were hired for the project and the third being an author on this paper. The hired annotators were paid $30/hour with total compensation amounting to $420 per annotator.

Over what time-frame was the data collected?
Annotation occurred over seven consecutive days, from March 31 to April 6, 2025, following annotator training.

How was the data associated with each instance acquired?
The data was acquired via human annotation in batches of 150 instances, using a codebook to guide labeling decisions.

Does the dataset contain all possible instances?
No. It is a sample from a larger set of 280,975 sentences.

If the dataset is a sample, then what is the population?
All sentences in the FOMC transcripts between 1977 and 2008, after preprocessing.

Is there information missing from the dataset and why?
No data is missing.

Are there any known errors, sources of noise, or redundancies in the data?
No.

B.4 Data Preprocessing

What preprocessing/cleaning was done?
Utterances with fewer than 4 or more than 500 tokens were excluded from the unlabeled pool. We used sentence segmentation to split sentences.

Was the "raw" data saved in addition to the preprocessed/cleaned data?
Yes, the raw data exists from ConvoKit.

Is the preprocessing software available?
No.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
Yes.

B.5 Dataset Distribution

How is the dataset distributed?
The dataset is distributed via the authors' GitHub, `https://github.com/kakeith/op-fed` and HuggingFace `https://huggingface.co/datasets/kakeith406/op-fed`.

When will the dataset be released/first distributed?
At the time of publication.

What license (if any) is it distributed under?
We will release a license at the time of publication.

Are there any fees or access/export restrictions?
No.

B.6 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?
The last author on this paper.

Will the dataset be updated?
No updates are planned.

If the dataset becomes obsolete how will this be communicated?
This will be posted on the project's GitHub page.

Is there a repository to link to any/all papers/systems that use this dataset?
No.

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?
They may download and build upon it freely.

B.7 Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?
Not applicable. The data are public-domain government transcripts.

If it relates to other ethically protected subjects, have appropriate obligations been met?
Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals?
No.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?
Not applicable.

If it relates to people, could this dataset expose people to harm or legal action?
Not applicable.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?
Not applicable.

If it relates to people, were they provided with privacy guarantees?
Not applicable.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)?
Not applicable. The data are public-domain government transcripts.

Does the dataset contain information that might be considered sensitive or confidential?
No.

Does the dataset contain information that might be considered inappropriate or offensive?
No.

	Annotators			
Batch	All	1 and 2	2 and 3	1 and 3
1	0.647	0.663	0.558	0.716
2	0.572	0.499	0.420	0.808
3	0.681	0.639	0.684	0.715
4	0.541	0.539	0.492	0.589
5	0.512	0.584	0.413	0.538
6	0.457	0.620	0.308	0.433
7	0.411	0.430	0.202	0.609
Total	0.567	0.586	0.463	0.649

Table 7: **Fleiss' Kappa scores for opinion** (Stage 1) by AL batch and annotator pair.

C Additional Analysis of OP-FED

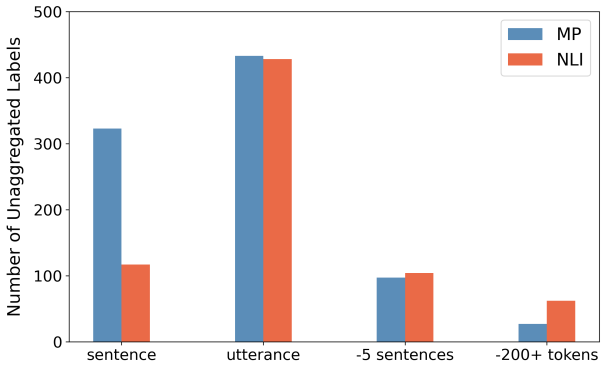


Figure 4: Context label distribution across the three annotators for Stage 3 - MP Context and Stage 5 - StanceNLI Context.

In this section, we provide additional analyses of OP-FED. Table 3 provides the label distribution breakdown of the dataset and inter-annotator statistics of the dataset. These inner-annotator agreement rates are for the subset reported in the table (i.e., after discarding instances that did not achieve a 2/3 majority vote by annotators).

Table 7 shows the Fleiss' Kappa scores for Stage 1 (opinion) by AL batch number and by annotator pairs. Table 8 reports the annotation time per annotator.

D Additional Modeling and Compute Details

All experiments are run on a local server that has the following NVIDIA GPUs: 16 × A6000s and 4 × A100s. See Table 11 for hyperparameters used for ModernBERT.

	Annotator		
Batch	1	2	3
1	90	120	145
2	135	105	180
3	90	135	105
4	120	120	120
5	130	140	130
6	120	135	135
7	130	120	120
Total	815	875	935

Table 8: Time to complete annotations (in minutes) by annotator and batch.

D.1 Details for BOW+LogReg experiments

For our bag-of-words logistic regression baseline (Table 4), we use 4-fold cross validation and report the overall accuracy across all instances across every inference fold. We implement logistic regression using scikit-learn (Pedregosa et al., 2011) with a balanced class weight loss function and an L2 penalty with $C = 1.0$. We obtain bag of word representation of the text using unigram tokenizer, removing stop words, setting a minimum document frequency to 5, and the maximum number of features (unigrams) to 1000.

E Acquisition functions

The effectiveness of an active learning pipeline depends heavily on the acquisition function used to select instances for labeling. We evaluated three types of acquisition functions which Zhang et al. (2022) bucket into categories of informativeness, representativeness, and hybrid. Methods under informativeness may select challenging but redundant examples, while methods under representativeness may favor easy yet uninformative ones. To address these shortcomings, we also implement contrastive active learning (CAL), a hybrid strategy that tradeoffs both informativeness and representativeness (Margatina et al., 2021). We elaborate on the details of these acquisition functions in the subsections below.

E.1 N-Gram Density

As a baseline acquisition function, we use n-gram density, which selects the next batch of instances for annotation based on unseen n-gram counts (Eck et al., 2005). As a representativeness approach, it prioritizes sentences that introduce new lexical content, encouraging the model to select underrepresented regions of the input space during training. The basic algorithm for a single active learning iteration is as follows:

1. For each unlabeled instance $x_i \in \mathcal{D}_{\text{pool}}$ not already in the sorted list:
   - Calculate the weight of x_i via Equation 1.
2. Select the instance with the highest weight and add it to the sorted list.
3. Repeat until the length of the sorted list reaches the desired batch size b.

The set of observed n-grams is cumulative and persists across active learning iterations. The weight of each unlabeled instance x_i is determined by the following equation:

$$\text{weight}_{i,j} = \frac{\sum_{n=1}^{j} \text{count}(\text{unseen } n\text{–grams})}{|\text{sent}|} \quad (1)$$

where j is the maximum n-gram length considered (e.g., $j = 4$ includes n-grams up to fourgrams); $\text{count}(\text{unseen } n\text{–grams})$ is the total number of n-grams of length n in the sentence that have not appeared in any previously labeled sentence; and $|\text{sent}|$ is the token length of the sentence. Repeated occurrences of the same unseen n-gram within a sentence are counted multiple times when computing the weight.

E.2 Maximum Entropy

Maximum entropy is an informativeness acquisition strategy that selects instances for which the model's predictive distribution is most uncertain

(Schröder et al., 2022). This approach relies on the output of a supervised model, and selects instances where the predicted class probabilities are closest to uniform, indicating maximal uncertainty. For instance, in binary classification, the largest maximum entropy scores correspond to instances where the predicted probability for each class is close to 0.5.

Formally, for a given unlabeled instance $x_i \in \mathcal{D}_{\text{pool}}$, where x_i is the input and we have c possible classes, we select instances that have the maximum entropy:

$$\arg\max_{x_i} \left[-\sum_{j=1}^{c} P(y_i = j \mid x_i) \log P(y_i = j \mid x_i) \right] \tag{2}$$

Here, $P(y_i = j \mid x_i)$ is the model's predicted probability that instance x_i belongs to class j, and the index i ranges over all unlabeled instances in the pool. The acquisition function selects the top b instances with the highest entropy scores.

E.3 Contrastive Active Learning (CAL)

Contrastive active learning (CAL) is a hybrid acquisition function that combines informativeness and representativeness by identifying "contrastive" instances: unlabeled instances whose model predictions diverge significantly from those of their labeled nearest neighbors (in the model feature space) (Margatina et al., 2021). We re-implement Margatina et al.'s CAL algorithm, which we report in Algorithm 1.

In our experiments, we use the `[CLS]` token embedding of either RoBERTa or ModernBERT as our encoder $\Phi(\cdot)$. For each unlabeled candidate $x_p \in \mathcal{D}_{\text{pool}}$, we identify its k nearest neighbors in the labeled set $\mathcal{D}_{\text{lab}}$, forming a neighborhood:

$$\mathcal{N}_{x_p} = \{x_p, x_l^{(1)}, \ldots, x_l^{(k)}\}, \quad x_l^{(i)} \in \mathcal{D}_{\text{lab}}$$

Consistent with the authors' setup, we fix $k = 100$. However, while the authors use Euclidean distance to compute neighborhood similarity, we use cosine distance (Jurafsky and Martin, 2025). We then use the model $\mathcal{M}$ to compute the predicted class distributions $p(y \mid x)$ for all $x \in \mathcal{N}_{x_p}$. The final score s_{x_p} for the candidate x_p is computed as the average Kullback–Leibler (KL) divergence between its predicted class distribution and those of its k labeled neighbors.

Finally, the top b instances with the highest scores are selected for labeling in each active learning iteration. Future work could explore alternative hyperparameter choices (e.g., $k = 100$, cosine distance) or incorporate manual inspection of nearest neighbors for more nuanced semantic similarity judgments.

F Additional AL Simulation Experiments

F.1 Additional Task for Simulations

For AL simulation, we also experiment with the Trillion Dollar Words (TDW) dataset which includes sentences from FOMC speeches, meeting minutes, and press conference transcripts (Shah et al., 2023). Each sentence is labeled as either *hawkish*, *dovish*, or *neutral*. These labels correspond to whether the sentence advocates for policy tightening (hawkish), easing (dovish), or neither (neutral). For example, the following sentence is classified as *dovish*: *"Recent declines in payroll employment and industrial production, while still sizable, were smaller than those registered earlier in 2009"*. In contrast, the following sentence is classified as *hawkish*: *"Yields on longer-term inflation-indexed Treasury securities, which are relatively illiquid, rose more sharply than did those on nominal securities."* Notably, hawkish-dovish classification differs from our desired task because it does not require the stance to be attributed to an individual speaker. Yet, we believe it may be useful for our AL simulations given that the dataset is in the same domain of Federal Reserve communications.

To match the 25% positive and 75% class proportions observed in our prototype annotation rounds, we augmented both the TDW dataset with additional neutral examples. To do so, we use the observation that TDW dataset was originally constructed using a domain-specific keyword filter, and so we randomly sampled sentences from our FOMC transcript corpus that contained keywords the original authors had used as exclusion criteria as our "neutrals". For simplicity, we refer to the resulting augmented datasets as TDW+AN, where "AN" denotes the inclusion of augmented neutrals.

F.2 Full supervision

In Table 12, we report baseline model performance under full supervision, where models are trained on the entire gold-standard training set. This serves as an upper bound on performance for AL methods.

Algorithm 1 Single iteration of CAL Margatina et al. (2021)

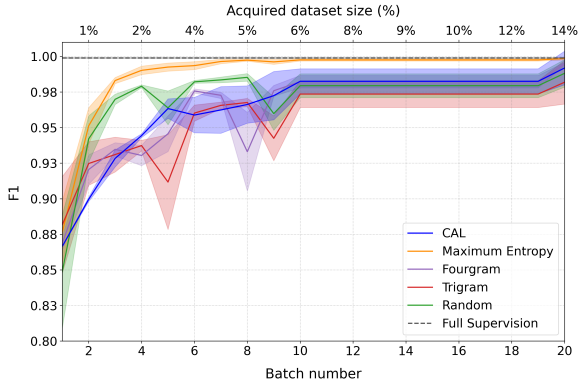
Input: labeled data $\mathcal{D}_{\text{lab}}$, unlabeled data $\mathcal{D}_{\text{pool}}$, acquisition batch size b, model $\mathcal{M}$, number of neighbors k, model representation (encoding) function $\Phi(\cdot)$

for $x_p \in \mathcal{D}_{\text{pool}}$ **do**
  $\{(x_l^{(i)}, y_l^{(i)})\}, i = 1, \ldots, k \leftarrow \text{KNN}(\Phi(x_p), \Phi(\mathcal{D}_{\text{lab}}), k)$ ▷ find neighbors in $\mathcal{D}_{\text{lab}}$
  $p(y \mid x_l^{(i)}) \leftarrow \mathcal{M}(x_l^{(i)}), i = 1, \ldots, k$ ▷ compute probabilities
  $p(y \mid x_p) \leftarrow \mathcal{M}(x_p)$
  $s_{x_p} = \frac{1}{k} \sum_{i=1}^{k} \text{KL}(p(y \mid x_l^{(i)}) \| p(y \mid x_p)), i = 1, \ldots, k$ ▷ compute divergence score
end for
end
$\mathcal{Q} = \underset{x_p \in \mathcal{D}_{\text{pool}}}{\arg\max}\, s_{x_p}, |\mathcal{Q}| = b$ ▷ select batch

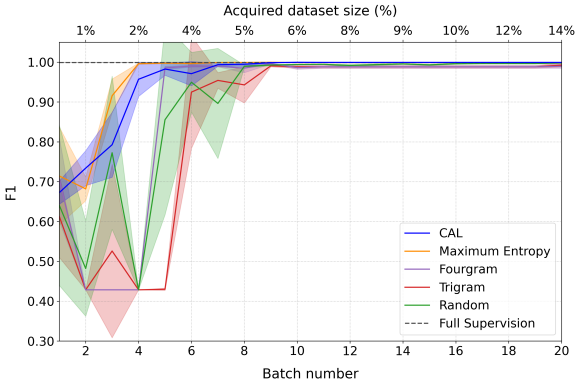
Output: $\mathcal{Q}$

F.3 Additional Results

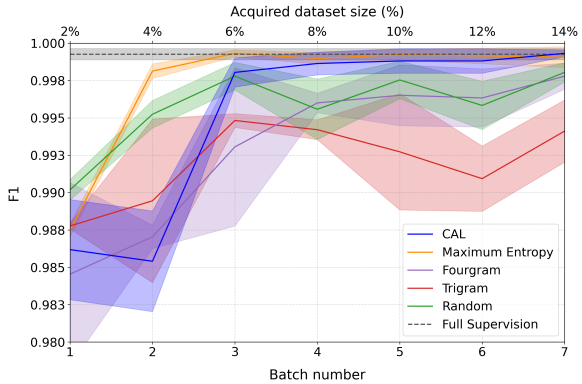
In Figure 5, we show the results of our AL pipeline after varying one decision at a time: the batch size from 50 to 150, the model from ModernBERT to RoBERTa, the simulation dataset from VAST+AN to TDW+AN, and the class distribution from 25%-75% to 10%-90% (positive-neutral classes, respectively). Four out of five of these settings reinforced our choice of using maximum entropy as the acquisition function in our one-shot deployment of AL, as it had an acquisition curve that was steeper than or equal to the other methods.



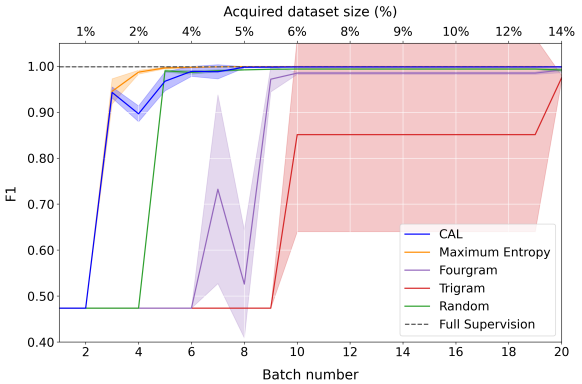
(a) Test set of VAST+AN using ModernBERT, batch size 50, and class imbalance 25% positive–75% neutral.



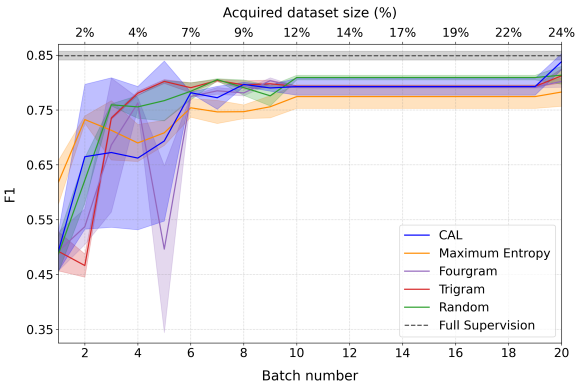
(b) Test set of VAST+AN using RoBERTa, batch size 50, and a class imbalance of 25% positive–75% neutral.



(c) Test set of VAST+AN using RoBERTa, batch size 150, and a class imbalance of 25% positive–75% neutral.



(d) Test set of VAST+AN using RoBERTa, batch size 50, and a class imbalance of 10% positive–90% neutral.



(e) Test set of TDW+AN using RoBERTa, batch size 50, and a class imbalance of 25% positive–75% neutral.

Figure 5: **Additional active learning simulation results.** We evaluate model performance on the held-out test sets of VAST+AN ($n = 3185$) and TDW+AN ($n = 4028$) under varying conditions. Each plot reports test macro F1 (y-axis) for models trained on $\mathcal{D}_{\text{lab}}$ at each AL batch (x-axis). Solid lines indicate the mean performance across 5 runs with different random seeds; shaded regions represent the standard deviation. We vary model architecture (ModernBERT vs. RoBERTa), batch size (50 vs. 150), dataset (VAST+AN vs. TDW+AN), and class imbalance (25%–75% vs. 10%–90%).

Unique id	Sentence	MP Context or StanceNLI Context
19811222_230_4	*"But the discrepancies in the numbers were even larger than we had."*	n/a
9910206_54_6	*"If the Congress feels comfortable with the arrangement, I'm certainly on board in this context."*	n/a
19810203_410_1	*"I don't know how important all of this is."*	**-5 sentences:** 'speaker': 'MR. GRAMLEY.', 'text': "It seems to me, however, that the market is going to interpret your suggestion in a very different way. If the fed funds rate on the day of the Committee meeting is 17.42 percent and we say 300 basis points on either side, the newspapers would say that they have figured out that the Fed's range is 14.42 to 20.42 percent. It makes a lot more sense to say it's 14 to 20 percent. I think that's what they're going to do. Then if we start playing games and say we're going to narrow the range down to 200 basis points on either side because we don't want rates to go up too far, they're surely going to say we are playing the rates instead of the money supply."
19791006_304_14	*"That sounds like a good idea."*	**Utterance:** A 1 percentage point increase. Under those circumstances, let me modify my proposal slightly. We work with a 4 . 6 or 4.5 percent target for M1, recognizing that people would rather see M1 growth come in somewhat lower, certainly, than higher. But that to us is a satisfactory target. In an uncertain world, all other things equal, the money market nice and equitable, expectations changing nicely, things settling down, it's possible we would feel better if it came in below than if we were embarrassed by it being too high. We have a discount rate of 12 percent. We have the reserve requirement change, which I think will be 8 percent, on a basket of managed liabilities. That is about equivalent to an added cost of 1 percent on those [liabilities]. That's the effect that has, mechanically. I don't know what that does to the prime rate, but [the added cost to banks] is just marginal. If the banks want to be mean–I've got to speak to the MA–they raise the prime rate and say that's our marginal cost of funds. If they want to be reasonable, they don't because that cost is only going to apply to a very small amount at the margin. I would be inclined to tell them at the ABA that they shouldn't reach forever on the interest rate. That sounds like a good idea. In and of itself, I think [all] that means some increase in the federal funds rate. How much of an increase? Although this scenario only puts the discount rate at about or slightly above the current federal funds rate, it nonetheless, all other things equal, puts some upward pressure on the federal funds rate. In terms of the range, Phil suggested 11-1/2 to 14-1/2 percent. As I told you, it makes me nervous to think of [the rate] going up and getting locked in at a higher level. But if you want to put the upper end higher, I reserve the right to consult if the funds rate were to begin getting up that high. We could have that understanding. I am not talking about for a day; I think it would be fine if it went there for a day or a bit beyond.
19850213_474_1	*"I would go with "B" also."*	**-5 sentences:** 'speaker': 'MR. MARTIN.', 'text': 'It seems to me that the odds are that M1 will exceed 8 percent. My only discomfort is that 8-1/4 percent may not be the correct fed funds rate for the short run; maybe it should be 8-1/2 percent. But I take it that there is enough flexibility built into our procedures [to allow for] an 8-1/2 percent rate if we need it to get 8 percent [M1 growth] or whatever that is–less than the 10 percent and 12 percent that we've been experiencing in recent months. Given that assumption, I would go along with "B."', 'speaker': 'CHAIRMAN VOLCKER.', 'text': 'Mr. Boykin.'
19930707_543_1	*"Well, I'm clearly not in tune with the other members of the FOMC."*	**Utterance:** Well, I'm clearly not in tune with the other members of the FOMC. I don't see any need to wait for any information. The markets provide indications every day as to whether or not we've provided more liquidity than is called for. And when we lowered the fed funds rate from 4 percent to 3 percent, my guess is that made very, very little difference in the rate of real GDP growth. The destabilizing factors that have led people to hold, shift, or use their balances as much as they have was, I think, a drag against the stimulus that was already in place. A 4 percent fed funds rate already was providing a lot of stimulus. We already had pegged the fed funds rate well below the natural rate of interest. We now have evidence, and we see it in our own staff's forecast, that the expected rate of inflation has moved up 0.6 or 0.7 percent just [since the May FOMC] meeting. Now, in that environment, policy is not stable. We do not have the same policy that we had at the last meeting because the real rate of interest by our best estimate has fallen to even more sharply negative territory. We clearly see in the price of gold that people are making bets out there. And we continue to lock in to a fed funds rate that will only aggravate that kind of speculation and it only detracts from the role of the U.S. currency as the stable currency for the world. The cost to the world of the United States pursuing inflationary policies in the late 1960s and the 1970s is unbelievable. We're still paying the cost because many other central banks with no confidence in us think they have to be Rambo-like in beating their chests because in some sense they've got a track for the inflation-induced environment that the Federal Reserve as the world's reserve currency provides. This is the time for us to move real interest rates back at least to zero. I would be satisfied to do a measly 1/4 percentage point, which would not get us back to zero, but there would be intervention value in that. There wouldn't be any harm in it. Is there anyone who really believes the U.S. economy, regardless of what is done on the fiscal [side], is going to suffer because the funds rate goes up from 3 to 3-1/4 percent? Now, I read Henry Kauffman, as many of you must have, and it's just absurd. Well, I must be the one that's absurd! Thank you.

Table 9: This table provides the context for Table 1 and adds the maximum context needed for either the MP or StanceNLI label.

	Model	Checkpoint	1-Opinion ($n = 1042$)	2-MP ($n = 485$)	3-MP Ctxt. ($n = 242$)	4-StanceNLI ($n = 159$)	5-Stance Ctxt. ($n = 162$)
Baselines	Majority Class	–	0.52	0.56	0.59	0.44	0.85
	BOW+LogReg	–	0.61	0.60	0.80	0.31	0.75
	Human min.$*$	–	0.85	0.85	0.86	0.80	0.85
	Human max.*	–	0.95	0.89	0.89	0.89	0.93
Zero-Shot	GPT-5	`gpt-5-2025-08-07`	0.72	0.74	0.28	0.52	0.28
	GPT-5 nano	`gpt-5-nano-2025-08-27`	0.75	0.73	0.33	0.48	0.28
	Claude Opus 4.1	`claude-opus-4-1-20250805`	0.80	0.74	0.24	0.63	0.29
	DeepSeek-V3.1	`deepseek-chat-2025-08-21`	0.75	0.76	0.47	0.31	0.15

Table 10: **F1 scores of baseline models.** These are for comparison to the accuracy scores from Table 4. Task 4 is a weighted F1 score, all the rest are (micro) F1 scores for a binary classification task.

Model and Training Hyperparameters			
Dropout	0.1	Epochs	5
Train batch size	50	Optimizer	AdamW
Learning rate	3e-5	Weight decay	0.05
Tokenizer Settings			
Padding side	Right	Truncation side	Right
Padding strategy	Max length	Truncation strategy	Only first
Max sequence length	512		

Table 11: Fixed hyperparameter settings used across all AL simulations and deployment.

	VAST+AN		TDW+AN	
	F1 Train ($n = 7459$)	F1 Test ($n = 3185$)	F1 Train ($n = 4028$)	F1 Test ($n = 996$)
LogReg + BoW	0.981	0.953	0.844	0.747
RoBERTa-large	1.000	0.999	0.980	0.856
ModernBERT-base	1.000	0.999	0.998	0.826

Table 12: Model performance comparison across our gold-standard datasets.

G Prompting Details

In this section, we report the prompt templates as input into the LLM APIs for each of the five tasks from our five-stage hierarchical schema. Each prompt specifies the classification task (using natural language), defines relevant terms, and enforces a labeling constraint. For context-dependent tasks (MP and StanceNLI), the variable {$CONTEXT} refers to the previous 200 tokens (plus any additional tokens required to complete the preceding sentence) followed by the full utterance. This matches the context available to human annotators during the annotation process. In these tasks, the {$SENTENCE} variable also includes the speaker's name, allowing the model to identify whether the speaker is the same as in the preceding context.

To increase generated answer compliance, we used best practices described by each AI company for each of the LLMs. For the *GPT-5* models, we specified a JSON schema (in dictionary form) that listed the valid labels for the `response_format` argument. For *DeepSeek*, we specified an argument for JSON outputs and listed the valid labels in the prompt itself. For *Claude*, we listed the schemas in their XML format and asked for the response in an XML format.

Answer parsing. We found the LLMs were 100% compliant with our constraints of the label output space. So we used **deterministic answer extraction** and match the surface form of the label with the output of the LLM (since the output is formatted as a JSON object or XML object) (Schulhoff et al., 2024).

Prompt templates. Below, we provide the prompt templates for *DeepSeek* since these explicitly listed the label constraints in the prompt. As we mention above, label constraints were passed in via a dictionary for *GPT-5* and in XML format for *Claude*.

Prompt template for predicting *opinion* (Stage 1):

```
Is the author expressing an opinion in the sentence below? Here, opinion is defined as the expression of a subjective perspective, recommendation, or position.

Label constraints: ["yes", "no"]

Output in the following JSON format:
"label": your label here,
"explanation": Give a short explanation.
"confidence": Include a numerical confidence score between 0 and 1

Sentence: {$SENTENCE}

JSON output:
```

Prompt for predicting *monetary policy* (Stage 2):

```
Does the speaker of the sentence refer to monetary policy, given the context? Here, monetary policy is defined as actions the U.S. Federal Reserve may take such as adjusting the federal funds rate, the money supply, or the discount rate.

Label constraints: ["yes", "no"]

Output in the following JSON format:
"label": your label here,
"explanation": Give a short explanation.
"confidence": Include a numerical confidence score between 0 and 1

Context: {$CONTEXT}

Sentence: {$SENTENCE}

JSON output:
```

Prompt for monetary policy "needs context?" task (Stage 3):

Given the sentence, do you need more context beyond the sentence to determine whether the sentence is about monetary policy? Here, monetary policy is defined as actions the U.S. Federal Reserve may take such as adjusting the federal funds rate, the money supply, or the discount rate.

Label constraints: ["yes", "no"]

Output in the following JSON format:
"label": your label here,
"explanation": Give a short explanation.
"confidence": Include a numerical confidence score between 0 and 1

Sentence: {$SENTENCE}

JSON output:

Prompt for StanceNLI (Stage 4):

Given the context, does the sentence entail, contradict, or have a neutral relationship to the hypothesis: "I think we should tighten monetary policy"? Here, tightening monetary policy is defined as actions the U.S. Federal Reserve may take to increase the federal funds rate, such as reducing the money supply or raising the discount rate.

Label constraints: ["entailment", "contradiction", "neutral"]

Output in the following JSON format:
"label": your label here,
"explanation": Give a short explanation.
"confidence": Include a numerical confidence score between 0 and 1

Context: {$CONTEXT}

Sentence: {$SENTENCE}

JSON output:

Prompt for StanceNLI "needs context?" task (Stage 5):

Given the sentence, do you need more context beyond the sentence to determine whether the sentence entails, contradicts, or is unrelated to the hypothesis: "I think we should tighten monetary policy"? Here, tightening monetary policy is defined as actions the U.S. Federal Reserve may take to increase the federal funds rate, such as reducing the money supply or raising the discount rate.

Label constraints: ["yes", "no"]

Output in the following JSON format:
"label": your label here,
"explanation": Give a short explanation.
"confidence": Include a numerical confidence score between 0 and 1

Sentence: {$SENTENCE}

JSON output:

H Annotation Codebook

Upon publication, we will provide the codebook provided to annotators as part of their training process in our data and code repository. The codebook outlines the annotation tasks, label definitions, and examples used to guide consistent and accurate labeling across our five stages. Annotators were instructed to refer back to this document throughout the annotation process.

Unique ID	Sentence	Opinion	GPT-5 Pred.	GPT-5 Conf.	GPT-5 Explanation
1978032127015	" I don't know whether that bill is absolutely going to pass."	yes	no	0.90	It reports uncertainty about the bill's passage, not a subjective judgment, recommendation, or stance on it.
198005204003	" I don't know for sure whether it will."	yes	no	0.90	It expresses uncertainty about an outcome, not a subjective stance, judgment, or recommendation.
1978062017	"Whether or not it would be wise, possible, desirable, or agreeable to temporarily suspend the ranges for M1 until we can find out what's happening, I don't know ."	yes	no	0.86	The speaker explicitly withholds a stance ("I don't know") and makes no recommendation or evaluation; they express uncertainty rather than an opinion.
199312212661	"Well, I'm really raising the issue: Is the October 15th tape and transcript one that if somebody comes in and asks for it under the Freedom of Information Act we're going to have to deliver ."	yes	no	0.83	The sentence raises a procedural question about FOIA obligations rather than stating a subjective judgment, recommendation, or stance.
19891219581	"Let me just say, and Larry referred to this, that we have had underway a comprehensive examination of this issue; had it not been for the overload we reached when you requested this briefing, we would have had it done by now."	yes	no	0.79	It describes actions and circumstances (ongoing examination, workload causing delay) rather than a subjective stance or recommendation.
200710312679	"Bill, I think you asked a good question."	no	yes	0.97	It expresses a subjective evaluation—"I think" and calling it a "good question" reflect the author's personal judgment.
19961113283	"Given that we feel that GDP growth may have been underestimated, it is not inconceivable that a part of those revenues could be in consumption expenditures."	no	yes	0.95	It expresses a subjective belief ("we feel") and speculation ("not inconceivable") about GDP growth and revenue allocation, indicating an opinion.
200803181052	" I think there was a pretty big change this morning at least in risk perceptions today across a bunch of markets."	no	yes	0.95	The phrase "I think" and the evaluative term "pretty big" show a subjective assessment about changes in market risk perceptions.
20080130991	"Well, that is a hard question, actually."	no	yes	0.94	It expresses a subjective judgment that the question is hard.
197907111374	"We are not rebasing; I think everyone understands that."	no	yes	0.94	The phrase "I think everyone understands that" expresses the speaker's subjective belief about others' understanding, indicating an opinion.

Table 13: **Error analysis of *GPT-5* confidence scores on OPINION.** Here we report 10 examples where the opinion label from our human annotators differed from the predicted label from *GPT-5*, sorted by descending confidence score. Note, many of the differences come from *GPT-5*'s confidence in lexical phrases: "I don't know" as not an OPINION and "I think" as an OPINION. We also highlight in blue some subtle phrases indicative of opinion/perspective that *GPT-5* misclassified.