Economic News and Personal Income Expectations

PRELIMINARY

Franziska Bender & Johannes Binswanger

August 2025

Abstract

We study how subjective expectations about personal income are shaped by economic news provided by the media. We address two questions: First does publicly available information in the form of news topics affect households income expectations? Second, do households use this information efficiently? To answer this, we combine household survey data from the Michigan Survey of Consumers with a corpus of roughly 3 million articles from *The Wall Street Journal* and *The New York Times*. We use a Latent Dirichlet Allocation (LDA) model to construct monthly time series for the prominence and tone of 70 economic news topics. Our analysis yields two key findings. First, we establish that news content has significant predictive power for households' future income expectations. Second, and our main contribution, we find compelling evidence that households use this information inefficiently. While household forecasts suffer from large general biases, a model that optimally processes news topics can reduce the remaining forecast error for realized income substantially. This belief distortion stems from households over-relying on their private information or judgment and mis-weighting information from news topics. Our results help explain the origins of systematic errors in personal income expectations, identifying the inefficient processing of public information from the media as a significant driver. This suggests that models of expectation formation may need to incorporate deviations from rational processing to fully capture household behavior.

Keywords: Individual income expectations, media, news, topic model.

JEL Codes: D84, E20, E71.

1 Introduction

Expectations play a pivotal role in macroeconomics and intertemporal decision-making. A rich body of literature has established that household expectations for aggregate variables, such as inflation or GDP, systematically deviate from the full-information rational expectations (FIRE) benchmark. In response, the literature has moved beyond the FIRE assumption along two dimensions. The first relaxes the assumption of full information, positing that individuals do not observe the economy's fundamental states directly but instead form beliefs based on noisy signals. This more realistic assumption can account for many of the empirically observed deviations (Coibion and Gorodnichenko, 2015; Maćkowiak et al., 2023). The second alternative deviates from the assumption of rationality itself. A growing body of work shows that expectations are shaped by behavioral biases, such as an over-reliance on personal experiences Malmendier and Shen (2024) or an overreaction to recent events (Bordalo et al., 2020).

While the literature has extensively studied expectations for aggregate variables, a household's expectations about their own income are arguably more consequential as the primary driver of consumption and savings decisions. In a notable gap in the literature, however, this crucial area has received significantly less attention. The few existing studies confirm that personal income expectations also diverge from the FIRE benchmark, exhibiting systematic biases that may be consistent with over-extrapolation (Massenot and Pettinicchi, 2019; Cocco et al., 2019) and a striking pattern of pessimism among low-income households and optimism among their high-income counterparts (Rozsypal and Schlafmann, 2023). The precise channels that drive the formation of, and the systematic errors in, household personal income expectations remain an important open question.

Our study explores one plausible channel through which these expectations are formed: the news media. Given that households rely at least to some extent on external signals to form their outlook, the news media represents a source of information about the economy. Previous literature has already established that news media is a powerful driver for expectations about aggregate outcomes like inflation (Lamla and Lein, 2014; Larsen et al., 2021). It is not self-evident however, that news would affect personal income expectations in the same way. For personal income, private signals such as conditions at their workplace are likely to be more important than for aggregate variables. However, the news media report on a wide array of specific topics. For certain households, news about a specific industry, a new tax policy, or local employment trends may be just as relevant as their own private signals.

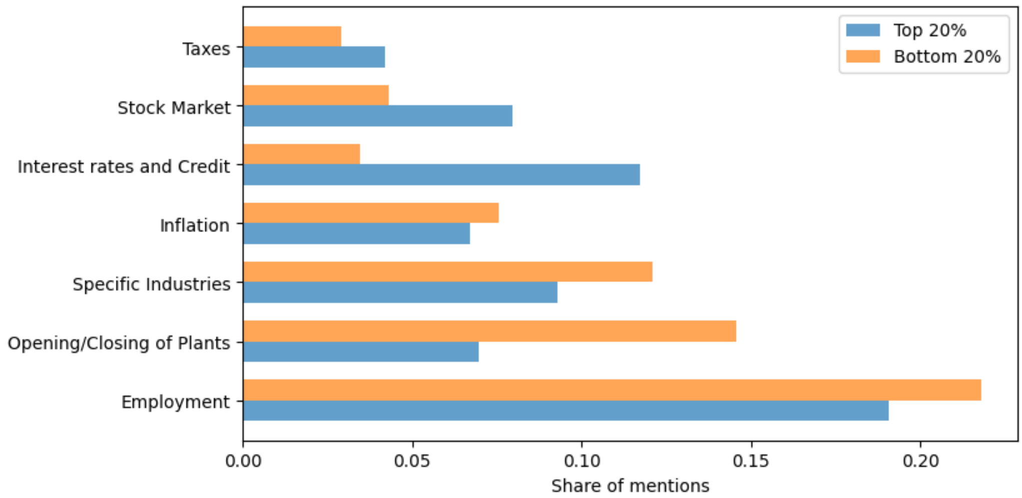
Motivating evidence from the Michigan Survey of Consumers (MSC) shows that households' attention to economic news is heterogeneous. When asked about recent news on business conditions, households' responses vary systematically with their income. As shown in Figure 1, the top income quintile is more likely to recall news related to 'Taxes' or the 'Stock Market', while the bottom quintile focuses more on 'Employment' or the 'Opening/Closing of Plants'. This divergence is intuitive and lends itself to the interpretation that households selectively attend to news topics most relevant to their personal financial situations. While such selective attention may be a rational response to cognitive constraints, it could also be a key mechanism behind systematic forecast errors[1] if it leads households to inefficiently process the full set of publicly available information.

This paper systematically investigates the link between the news media and household personal income expectations by addressing two fundamental questions. First, do household income expectations react to information presented in the economic news? Second, do households use this public information efficiently, or can a significant portion of their forecast errors be attributed to a systematic mis-weighing of news content?

To answer these questions, our analysis combines two datasets. We construct a quantitative measure of news content from a corpus of roughly 3 million articles published in *The New York Times* and *The Wall Street Journal* between 1984 and 2023. Using a Latent Dirichlet Allocation (LDA) topic model, we identify 70 recurring economic topics and measure both their monthly prominence and journalistic tone. We then connect this news data to household-level

[1] It is well documented that households expectation errors regarding various variables differ predictably with socioeconomic characteristics (Souleles, 2004; Das et al., 2020).

Figure 1: Attention to News Topics for Top and Bottom Income Quintile



Note: The bars represent respondents from either the top 20% or the bottom 20% of the income distribution in the MSC sample that have mentioned a particular topic. The percentage numbers indicated by the bars are relative to the number of respondents who have heard *any* news.

expectations from the Michigan Survey of Consumers (MSC), for which we construct matched expectation-realization pairs using the imputation methodology of Rozsypal and Schlafmann (2023).

Our empirical strategy proceeds in two parts. To address our first question, we design an out-of-sample forecasting exercise to test whether the news topics have incremental predictive power for future income expectations over standard controls. While a predictive link does not identify a causal effect, it provides important evidence consistent with the idea that households react to news. We then shift the analysis to our second question by predicting realized income growth itself. Here, we test whether a forecasting model augmented with our news topics can systematically improve upon households' own survey expectations. By construction, any improvement from our model implies that households are not efficiently incorporating the publicly available information contained in the news.

Our analysis yields two main findings that address our guiding questions. In response to our first question, we find that publicly available information in news reports has significant predictive power for households' future income expectations, containing information not captured by a wide range of standard macroeconomic indicators. This suggests that households do incorporate information from the news media when forming their personal financial outlook. In response to our second question, we find compelling evidence that households use this public information inefficiently. While their forecasts suffer from large general biases, a substantial portion of the remaining forecast error can be attributed to the mis-weighting of news topics. This inefficiency appears to stem from households overweighting their private information or judgment. We also show that the specific information households mis-weigh is not static; the set of relevant news topics changes dynamically over time, often aligning with major economic events like the Global Financial Crisis and the dot-com bubble. This underscores the importance of using dynamic models to understand how households process information in an ever-changing economic environment.

The remainder of the paper is structured as follows. Section 2 reviews the related literature on household income expectations and the economic impact of the news media. Section 3 describes the two primary datasets we construct: the household survey data from the Michigan Survey of Consumers, for which we create matched expectation-realization pairs, and the measurement of topics and their sentiment from Newspaper articles. In Section 4, we conduct an out-of-sample forecasting exercise to test whether news topics have predictive power for households' income expectations. Section 5 investigates the efficiency of this process by testing whether a forecasting model augmented with news topics can systematically improve upon households' own forecasts of their realized income growth. Section 6 concludes.

2 Related Literature

Our paper contributes to two distinct but related strands of economic literature: the study of household income expectations and the role of the media in shaping economic beliefs.

A rich literature has established that household expectations for aggregate variables like inflation or GDP deviate systematically from the full-information rational expectations (FIRE) benchmark (see D'Acunto and Weber (2024) for an excellent review). In contrast, household expectations about personal income – a key determinant of consumption and saving decisions – have received significantly less attention. Recently, however, a number of contributions have begun to fill this gap. This emerging research provides evidence that expectations for personal income also deviate from the FIRE benchmark (Massenot and Pettinicchi, 2019; D'Haultfoeuille et al., 2021) and exhibit systematic biases and heterogeneity across households. In particular, studies find patterns of excessive pessimism among low-income households and optimism among high-income households, which may be consistent with an overpersistence bias in how individuals form their beliefs (Rozsypal and Schlafmann, 2023).

A primary reason for the focus in the literature on aggregate variables such as inflation is data availability; since the ex-post realizations are publicly known, computing forecast errors is straightforward. In contrast, studying expectations for *personal income* has been challenging. This requires matching a households' forecasts with their own subsequent realized income, a data linkage that is often unavailable in standard surveys. A key methodological advance to overcome this data limitation comes from Rozsypal and Schlafmann (2023), who introduce a procedure to construct matched expectation-realization pairs for personal income in the Michigan Survey of Consumers (MSC).

Having established that household income expectations are systematically biased, we now turn to a second strand of literature that offers a potential explanation for these distortions: the role of the news media in shaping economic beliefs. Numerous studies have shown that the news media is a powerful driver of expectations for aggregate outcomes (Doms and Morin, 2004; Soroka, 2006; Medovikov, 2016). The tone and intensity of reporting have been found to influence the accuracy and disagreement of inflation expectations (Lamla and Maag, 2012; Lamla and Lein, 2014). Methodologically, our work is closely related to Larsen et al. (2021), who use a Latent Dirichlet Allocation (LDA) topic model to show that news topics can predict both inflation and inflation expectations.

This paper bridges the gap between these two literatures. We provide the first systematic test of whether narratives presented in the news media drive households *personal income* expectations in a similar way as they do for expectations about aggregate variables. Furthermore, we assess the efficiency of this process by testing whether the mis-weighting of public information from news topics can explain a significant portion of households' systematic forecast errors.

3 Data

This section describes the two primary datasets constructed for our analysis. First, we detail the use of micro-level data from the Michigan Survey of Consumers to obtain measures of household income expectations and their corresponding ex-post realizations. To create a complete panel, we follow the imputation methodology of Rozsypal and Schlafmann (2023). Second, we introduce a novel dataset on news media content, constructed from a corpus of roughly 3 million articles from *The New York Times* and *The Wall Street Journal*. We explain the multi-step process used to measure the prevalence and tone of specific economic topics, which involves data cleaning, a Latent Dirichlet Allocation (LDA) topic model, and a tone-adjustment procedure. The following subsections describe the construction of each dataset in detail.

3.1 Household Survey Data

We use micro-level data on households expectations from the Michigan Survey of Consumers (MSC). This survey interviews a representative cross-section of about 500 households every month. It gathers data on consumer attitudes and expectations regarding the economy, personal finances, business conditions, and overall economic outlook. It also has information on the socioeconomic characteristics of the respondents, such as education, sex, age, race, and income.

In this paper we focus on households' expectations about their own personal income. These are elicited with the following two questions:

- *During the next 12 months, do you expect your income to be higher or lower than during the past year?*
- *By about what percent do you expect your income to (increase/decrease) during the next 12 months?*

We combine these questions to construct a numeric variable for household's subjective expectations about the growth rate of their personal income in the next 12 months ($\mathbb{E}_t^i[\Delta Y_{t+12}^i]$). Throughout the paper, we refer to this variable as income expectations, unless otherwise clarified.

A key requirement for our analysis is a dataset containing both household income expectations and their subsequent realizations. A significant challenge, however, is that no single survey a long-running panel with directly matched ex-ante expectation and ex-post realization pairs for the same households.[2] .

We overcome this limitation by adopting the imputation methodology developed by Rozsypal and Schlafmann (2023) for the Michigan Survey. Their method addresses the imperfect overlap between the period of income expectation (next 12 months) and the period of reported realized income (previous calendar year). The procedure uses the subset of households for which realizations are observable to estimate a model of income growth as a function of household characteristics. This model is then used to impute a corresponding income realization for every stated expectation in the sample.

Following this approach, we obtain our two primary variables of interest: the household's stated expectation for income growth ($\mathbb{E}_t^i[\Delta Y_{t+12}^i]$), and the corresponding realized income growth (and ΔY_{t+12}^i). The realized income growth is either based on a direct report or imputed via the Rozsypal and Schlafmann (2023) procedure, ensuring we have a complete set of expectation-realization pairs for our analysis.

3.2 News Media Data

The news data consists of full-text news articles from Proquest[3], for the period of January 1984 to December 2023. Our news media corpus consists of roughly 3 million news articles, written in English, from two major US news outlets, *The New York Times* and *The Wall Street Journal*.

3.2.1 Data Cleaning and Preprocessing

In a first step, we clean the data, by removing articles that obviously do not contain any economically relevant content[4], as well as articles that are not useful for estimating the topic model; this includes, for example, articles that consist primarily of tables without much text. The structure and style differ between the two newspapers. We therefore have to define different criteria to exclude articles not suitable for our sample.

For the *The New York Times* we exclude articles that:

- Do not originate from the national, foreign, or business and financial desks.[5].
- Contain title patterns indicating non-news content[6]

For the *Wall Street Journal* we exclude articles that:

- Have page-citation tags corresponding to any section other than A, B, C, digit or missing.
- Correspond to weekend editions.
- Have a majority of subject tags associated with non-economic content, such as sports.

[2] The New York Fed's Survey of Consumer Expectations (SCE), introduced in June 2013, does feature a rotating panel where respondents can remain for up to 12 months, allowing for some direct matching of expectations and realizations. However, its time series is significantly shorter than the one available from the Michigan Survey of Consumers, making the MSC more suitable for our long-run analysis.

[3] https://tdmstudio.proquest.com

[4] We are very conservative here and keep all articles that do not clearly fall into this category

[5] If we would exclude the foreign desk as well we wouldn't get all the international topics, which might actually be better.

[6] For example "Quotation of the day", or"The Listings".

- Contain headline patterns associated with data tables or regular columns on sports, leisure, or books.
- Originate from document sections associated with obviously non-economic content[7].

For both newspapers we exclude all articles with less than 100 or more than 4000 words. The resulting set of articles is the data we will work with. We remove stopwords and lemmatize the words. To build the vocabulary we remove tokens that appear in less than 0.05% of documents and tokens that appear in more than 40% of documents, to remove both rare, potentially idiosyncratic words and overly common words that carry little specific meaning. We convert articles into a vector of counts for each term in the vocabulary. This processed corpus forms the basis for our topic analysis, which we detail in the following section.

3.2.2 Topic Model

To get a measurement for topics in our data we use a Latent Dirichlet Allocation (LDA), (Blei, 2003)) topic model. The LDA is one of the most popular topic models in the Natural Language Processing (NLP) literature. It has been widely used in recent economic research using text data, including for example business cycle and monetary policy analysis (Hansen et al., 2018; Hansen and McMahon, 2016; Bybee et al., 2024; Larsen and Thorsrud, 2019; Larsen et al., 2021).

The LDA is an unsupervised algorithm, it can discover an underlying structure in the data without being given any labeled samples to learn from. At its core, LDA assumes that documents are composed of a mixture of topics, and that each topic can be described by a group of words that frequently appear together. For instance, in a collection of newspaper articles, one topic might revolve around "elections", with words like "campaign", "candidate", and "vote" while another might focus on "employment" with terms like"worker", "job", and"labor." The model identifies these patterns by examining how words co-occur across the collection, essentially clustering words into coherent themes. The clusters are optimized so that relatively few clusters preserve as much of the meaning in the original corpus as possible by best explaining the variation in terms across articles. These clusters of terms are the topics we are interested in. The number of topics the model will identify is a parameter chosen by the researcher. We choose to extract 70 topics $K = 70$, which is in the range of K that is typically used in other work that uses data from newspapers (Larsen and Thorsrud, 2019; Thorsrud, 2020). The topics are labeled subjectively. The word clusters as well as our subjective labels are reported in table 3 in the appendix.

To get a measure of the monthly frequency of these identified topics in the data, we apply the trained model to an article to estimate the proportion of the text that is dedicated to each topic. Let $a_{i,t}$ denote article i written in period t, with $i \in (0, ..., N_t)$ where N_t is the number of articles written in period t. This article can be represented as a probability distribution over the total number of K topics

$$a_{i,t} = (p_{0,i,t}, ..., p_{K,i,t}),$$

where $p_{k,i,t}$ the proportion of article a_{it} that is associated with topic k. Typically the probabilities are zero or close to zero for most of the topics, with 2 or 3 topics being dominant. We

[7] Our approach parallels Bybee et al. (2024): in an early working-paper version they retained all sections and a later version applied a more restrictive filter, we adopt an intermediate screen.

get the monthly topic time series by summing these probabilities over all articles written in month t and dividing be the number of articles, that is for topic k we have

$$\mathrm{NT}_{k,t}^{freq} = \frac{\sum_i^{N_t} p_{k,i,t}}{N_t}$$

While $\mathrm{NT}_{k,t}^{freq}$ is not exactly equivalent to the frequency of articles that report about a certain topic since articles can be, and in fact typically are, mixtures of topics. Nonetheless this aggregation method effectively captures the monthly prevalence of each topic across the entire news corpus.

3.2.3 Tone Adjustment

Other than the frequency of reports on a topic, the tone is likely to matter as well, particularly for expectations. For example if there are more reports on '*Employment*' it is relevant whether these news convey a positive or negative sentiment. We therefore follow the methodology used in Thorsrud (2020) and Larsen et al. (2021) to create tone-adjusted time series.

To measure the sentiment of an article we construct a sentiment score based on the number of positive and negative words in the article. We use the Loughran and McDonald Dictionary (Loughran and McDonald, 2024) that defines whether certain terms are positive or negative. It is specialized for economic and financial contexts. When we count the number of positive and negative words in an article we also check for negation, i.e. whether negation words such as 'not' precede a word with a positive or negative tone and adjust the count accordingly.

$$\mathrm{Pos}_{i,t} = \frac{\#\text{positive words}}{\#\text{total words}} \qquad \mathrm{Neg}_{i,t} = \frac{\#\text{negative words}}{\#\text{total words}},$$

As a result we get a measure for the sentiment of every article. The sentiment of article i in period t is calculated as follows

$$S_{i,t} = \mathrm{Pos}_{i,t} - \mathrm{Neg}_{i,t}$$

A challenge we face when we want to measure the sentiment of a topic is that an article by assumption of the LDA typically consists of more than one topic. A possibility is that the sentiment of one topic is primarily positive, whereas the sentiment of another topic in this article is primarily negative, which would give us a neutral sentiment for the whole article. To measure sentiment of topics and ultimately create the tone-adjusted time series we follow Larsen et al. (2021). Formally, for each topic k on day t, we find the article i^* that is most representative for a given topic[8]. We then measure the sentiment of the topic as the sentiment of this representative article.

$$S_{k,t} = S_{i^*,t} \quad \text{for} \quad i^* = \arg \max_{i \in \{0,\ldots,N_t\}} p_{k,i,t},$$

where $S_{k,t}$ is our measure of sentiment of topic k on day t. We construct the tone-adjusted

[8] 'Most representative' means the article has the highest probability associated with topic k among all articles published on day t.

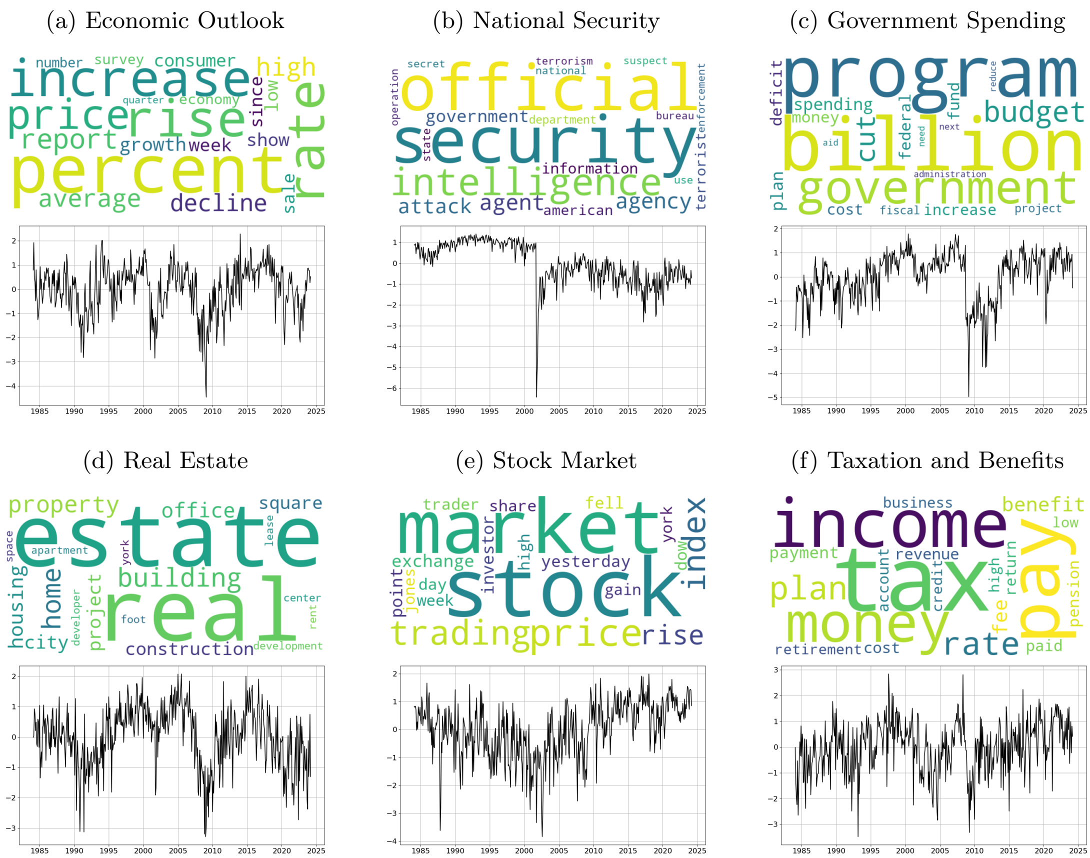
time series by multiplying the frequency of the topic with the sentiment of the topic

$$\text{NT}_{k,t} = S_{k,t} \cdot \text{NT}_{k,t}^{freq}$$

Finally we aggregate these series to a monthly frequency. Positive values of $\text{NT}_{k,t}$ are interpreted as positive news about topic k at time t, an increase in $\text{NT}_{k,t}$ can come either from an increase in the sentiment or – given a positive sentiment – from an increase in the number of articles on that topic. Similarly negative values of $\text{NT}_{k,t}$ are interpreted as negative news about topic k at time t, and a further decrease happens when the news either become more negative, or more articles with an overall negative sentiment are written about the topic.

Figure 2 illustrates the topic time series for six of the 70 topics (The full list of the estimated topics is given in table 3) The distribution of words for each topic are illustrated using word clouds, where a bigger font indicates a higher probability of the word occurring in the topic. The tone-adjusted time series, i.e. how much a topic is written about at any given point in time, and its tone, is illustrated below the word clouds. Progressively more positive (negative) values means that the media writes more about this topic, and/or that the tone of reporting on this topic is positive (negative).

Figure 2: Tone-adjusted Topics over Time



Note: For each topic the size of the word in the word cloud reflects the probability of this word occurring in the topic. Each word cloud only contains a subset of all words in the topic distribution. Topic labels are assigned on a subjective basis. The topic time series are standardized.

4 News-driven Income Expectations?

The first question we want to address is whether households income expectations react to news reported in the media. Prior literature has shown that households inflation expectations do react to topics reported in the media (see for example; Lamla and Lein, 2014; Larsen et al., 2021). While related, inflation and income expectations are conceptually different. Inflation is an aggregate price phenomenon, whereas income expectations pertain to a household's personal financial trajectory. It is not self-evident that households would update their personal income outlook based on the same set of public signals that influence their views on aggregate variables such as inflation.

To shed light on this question, we test whether information reported in newspapers contains predictive power for households' income expectations. We employ an out-of-sample forecasting framework to test if models augmented with sentiment-adjusted news topics can more accurately predict future income expectations than a baseline model. This approach allows us address whether news provide genuine, forward-looking informational content that is incorporated by households. While we cannot establish a causal link, we do interpret a high predictive power of news for income expectations as suggestive evidence consistent with the the idea that households react to reported news content.

This section proceeds as follows: Section 4.1 details our empirical methodology, including the forecasting models and evaluation criteria. Section 4.2 presents the results for both the average household and for different terciles of the income distribution, followed by a discussion of the implications.

4.1 Econometric Framework

To test the hypothesis that household income expectations are associated with news media content, we design an out-of-sample forecasting exercise. This approach moves beyond simple in-sample correlation to assess whether news topics provide incremental predictive power for future expectations. Our strategy relies on a comparison between three competing models, estimated over a 60-month rolling window to account for potential time-variation in the underlying relationships.

Our dependent variable $\mathbb{E}_t^j[\Delta Y_{t+12}]$ is the average expectation of households income growth over the next 12 months for a specific group j at time t. We specify three linear models to predict these expectations for the following month out-of-sample.

Our baseline model exclusively controls for variables derived from the Michigan Survey of Consumers, including age, race, sex, education and lagged realized income growth. The baseline regression equation is specified as

$$\text{Baseline Model:} \qquad \mathbb{E}_t^j[\Delta Y_{t+12}] = \alpha_0 + \beta' \mathbf{X}_{jt} + \epsilon_{jt}, \tag{1}$$

where $\mathbf{X}_{jt}$ is the vector of control variables from the MSC.

To investigate whether expectations react to topics reported in the news media we augment the baseline model with the sentiment-adjusted news topics (see section 3.2.3). Given the high dimensionality of the topics, we use the elastic net regularization method, proposed by Zou and Hastie (2005), for the selection of news topics. It performs variable selection by shrinking the coefficients of less important variables towards zero, effectively removing them from the model. The selection is performed on the full set of news topics and the baseline MSC controls. The news topics are lagged one period to avoid look-ahead bias. The model is estimated in two stages. In the first stage, the elastic net selects a subset of news variables, $\mathbf{N}^*_{t-1}$, and MSC controls, $\mathbf{X}^*_{jt}$. In the second stage, we run an OLS regression using the selected variables[9]. The resulting "News Model" is specified as:

$$\text{News Model:} \qquad E_j[\Delta Y_{t+12}] = \alpha_1 + \beta'(\mathbf{X}^*_{jt}) + \gamma'(\mathbf{N}^*_{t-1}) + \nu_{jt} \tag{2}$$

Comparing the out-of-sample mean squared error of the baseline model to the news model allows us to assess whether the news topics have predictive power for households income expectations. In the spirit of Larsen et al. (2021), who have estimated the predictive power of news topics for inflation expectations , we similarly interpret our results as tentative evidence of income expectations reacting to information reported by the news media.

A significant caveat is that the news media is just one of many channels through which households receive economic information. Consequently, any observed correlation between news content and expectations might arise not from the media's unique influence, but because news topics simply act as a proxy for the underlying economic fundamentals that households also observe through other means. To assess the predictive power of news in light of this concern, we directly compare the forecasting accuracy of our News Model against a benchmark that explicitly accounts for these underlying fundamentals. We estimate a third model where we replace the news-topics with a comprehensive set of macroeconomic variables $\mathbf{M}_{t-1}$ from the FRED-MD database[10]. The FRED-MD compiled by McCracken and Ng (2016) is a widely used dataset containing over 120 leading economic indicators covering the stock market, interest rates, prices, income, consumption and the labor market. We again use elastic net to select a subset of the FRED-MD variables $\mathbf{M}^*_{t-1}$ and MSC control variables $\mathbf{X}^*_{jt}$, and then, in a second stage, run an OLS regression using the selected variables. The resulting "FRED Model" is specified as:

$$\text{FRED Model:} \qquad E_j[\Delta Y_{t+12}] = \alpha_2 + \beta'(\mathbf{X}^*_{jt}) + \gamma'(\mathbf{M}^*_{t-1}) + \nu_{jt} \tag{3}$$

This allows us to compare the predictive power of news topics to the predictive power of FRED-MD variables. The idea is not that households track these individual series, which is clearly unrealistic, but rather that they are correlated with other signals (private or public) that households get and use when forming their expectations. The predictive power of the FRED variables is well established, it therefore serves as a deliberately challenging benchmark. If the News Model performs comparably to or better than the FRED Model, it would suggest that the narratives curated by the media are at least as effective at capturing the drivers of household expectations.

[9]See Belloni and Chernozhukov (2013) for results on refitting OLS after sparse, regularized selection, and Belloni et al. (2014) for a broader overview of high-dimensional selection and refitting.

[10]See also Larsen et al. (2021), who include FRED Variables to assess the independent relevance of news topics for inflation expectations.

4.2 Results

We now turn to the results of our out-of-sample forecasting exercise, which was designed to test whether news topics have predictive power for household income expectations. The key metrics for this evaluation are the ratios of the Mean Squared Error (MSE) from the augmented models (News and FRED) to the MSE of the baseline model, where a ratio below one indicates an improvement in predictive accuracy.

Table 1 presents results for the full sample alongside a breakdown by household income terciles[11]. The first two rows evaluate the News-Augmented Model, showing the reduction in MSE as both a ratio and a percentage improvement over the baseline. The last two rows provide the same metrics for the FRED-Augmented Model, allowing for a direct comparison of their predictive power.

Table 1: Out-of-Sample MSE-Ratios for Income Expectations

		Income Group		
	Full Sample	Bottom	Middle	Top
$MSE_{\text{News}}/MSE_{\text{Baseline}}$	0.92	0.85	0.96	0.89
Improvement News	7.8%	14.6%	4.0%	10.9%
$MSE_{\text{FRED}}/MSE_{\text{Baseline}}$	0.95	0.87	1.01	0.90
Improvement FRED	4.9%	13.0 %	-0.7%	10.2%

Note: The table displays the ratio of the out-of-sample Mean Squared Error (MSE) from the News-Augmented Model (Equation 2) and the FRED-Augmented Model (Equation 3) relative to the Baseline Model (Equation 1). The columns "Bottom", "Middle", and "Top" refer to households grouped by income terciles. Ratios below one indicate an improvement in forecast accuracy over the baseline model. The "Improvement" row shows this reduction in MSE as a percentage.

Our first result suggests that news topics contain some predictive information for household expectations. The News-Augmented Model consistently outperforms the baseline model across all household groups. As shown in the first panel of Table 1, the $MSE_{\text{News}}/MSE_{\text{Baseline}}$ ratio is 0.92 for average households income expectations, which corresponds to a 7.8% reduction in the models prediction error. This demonstrates that news topics contain significant predictive power for future income expectations, beyond what is captured by standard demographic controls. Notably, the predictive power of news is not uniform across the income distribution, and the pattern is non-monotonic. The effect is strongest for the bottom income tercile, where the news model reduces the MSE by a substantial 14.6%. The improvement is smallest for the middle tercile (4.0%) and rises again for the top tercile (10.9%). This U-shaped pattern would suggests that households at the lower and upper ends of the income spectrum are more responsive to the information contained in news topics when forming their expectations.

An alternative explanation for the non-monotonic, U-shaped pattern of forecast improvement relates to the composition of the income groups themselves. It may be that the middle-income tercile is more heterogeneous in its economic exposures than the bottom and top terciles. If this group contains a diverse mix of professions – for instance, public sector employees,

[11] Appendix C.1 shows additional results for other household groups, including income quintiles. The results for income quintiles are qualitatively similar.

skilled manufacturing workers, and healthcare professionals – then these subgroups might react strongly, but to very different news topics. In contrast, the bottom and top income terciles could be more homogeneous in their primary sources of income and thus more uniform in the news that is most relevant to their financial outlook. Our model seeks a single set of news topics to predict the average expectation for an entire group; if the middle tercile is indeed a composite of distinct subgroups, their unique reactions would effectively average out, leading to the muted overall signal we observe. Thus, the modest improvement for this group may not reflect a lack of reaction, but rather a diversity of reactions[12].

This predictive power suggests that households do incorporate information from the media when forming their expectations. This finding, however, does not establish a causal link, as news topics are likely correlated with underlying economic fundamentals, about which households may learn from a variety of sources. To test whether news content is merely a proxy for this broader economic information, our next step is to directly account for these fundamentals.

To that end we pit the News Model against the FRED Model, a benchmark representing a comprehensive set of official macroeconomic data. As shown in Table 1, the News Model consistently offers better predictive performance than the FRED Model across all household groups. For the average household, the News Model reduces the forecast error by 7.8%, while the FRED model reduces it by only 4.9%. This gap is evident across the income distribution. The News Model outperforms the FRED model for the bottom (14.6% vs 13.0%) and top (10.9% vs 10.2%) terciles. Most strikingly, for the middle income tercile, the FRED model provides no improvement over the baseline (a -0.7% change in MSE), whereas the News Model still improves the forecast by 4.0%. While the margin of outperformance is modest, the superior performance of the News Model across all groups suggests that narratives curated by the media contain information relevant to household expectations that is not captured by a wide range of standard macroeconomic indicators.

In summary our results show that a model based on news topics consistently outperforms a deliberately challenging benchmark. While the margin of outperformance is modest for some household groups, the superior predictive power of the News Model across all groups is a significant finding. It is particularly compelling because the assumption that households track official economic statistics is highly implausible. News, in contrast, is an arguably far more accessible source of information, making it a more likely driver of expectations. Overall the findings are suggestive for the hypothesis that news play a significant role in expectation formation. It is highly likely that individuals – or their peers – rely on information from news at least to some degree. A natural next question is therefore whether households do so in an efficient or optimal way.

[12] In Appendix C.1 we show additional results for different household groups that also indicates that homogeneity within the group with respect to the expectation formation may be an important factor to consider.

5 Mis-Weighing of Information from News

The preceding analysis provided significant evidence that news topics have predictive power for households' personal income expectations, suggesting that households incorporate information from the media when forming their outlook. This finding naturally leads to a deeper question regarding the efficiency of this process. For households to use news information efficiently, two conditions should be met: first, the news must contain genuine, predictive information for future realized income, and second, households' expectations must react appropriately to that information. If households systematically mis-weigh these public signals, either because they overreact or because they underreact, their forecasts will be suboptimal. This section therefore investigates whether a part of households' expectation errors can be attributed to an inefficient use of the information contained in news topics.

Following the logic of recent literature, in particular Bianchi et al. (2019), we explore two potential sources of inefficiency. First, we test for general biases, such as persistent optimism or a tendency to overweight private signals. Second, we examine the hypothesis, that households systematically mis-weight predictable information contained within news topics. To test these hypotheses, we shift our focus from predicting expectations to predicting realized income growth. We construct a forecasting model that optimally processes information from news topics and compare its accuracy against the raw survey expectations. If our model, using publicly available news, can improve upon households' own forecasts, it would imply an inefficient use of that information. As we will demonstrate, the results indicate that a significant portion of households' forecast errors can be explained by both persistent biases and the inefficient processing of news topics.

5.1 Econometric Framework

The core logic of our framework is to test whether we can systematically improve upon households' own forecasts using publicly available information. To do so, we construct two benchmark models and compare their out-of-sample forecasting accuracy for realized income growth against the households' raw survey forecasts. We begin with a baseline model where household expectations are the sole predictor of realized income. We then add our news topics as additional predictors. If these topics demonstrably enhance the model's forecasting accuracy, it serves as evidence that the information they contain was not efficiently incorporated into the original household expectations.

The first model corrects the raw survey forecast for simple, systematic biases, such as persistent optimism or pessimism, and for any stable tendency to overweight private information. This model regresses realized income growth on households' expectations:

$$\Delta Y_{t+12} = \alpha + \beta \mathbb{E}_t[\Delta Y_{t+12}] + \epsilon_{t+12} \tag{4}$$

If households were perfectly rational we would expect $\alpha = 0$, and $\beta = 1$. A finding of $\alpha \neq 0$ would indicate a persistent optimistic or pessimistic bias. Following the logic of Bianchi et al. (2022) a coefficient of $\beta < 1$ suggests that households place too much weight on their own forecast – which reflects private signals or judgment – relative to what would be optimal.

Our main model tests whether households efficiently incorporate the information contained in the news media. It augments the baseline model with our full set of $K = 70$ news topics:

$$\Delta Y_{t+12} = \alpha + \beta \mathbb{E}_t[\Delta Y_{t+12}] + \sum_{k=1}^{K} \gamma_k \text{NewsTopic}_{k,t-1} + \epsilon_{t+12} \tag{5}$$

If households efficiently use all public information from the news, then the additional topics should have no predictive power, and we would find $\gamma_k = 0$ for all k. Conversely, an improvement in the model's out-of-sample forecast accuracy after including the news topics would serve as evidence that households are systematically mis-weighting this public information when forming their expectations. Given the large number of predictors, this model is estimated using an Elastic Net (EN) regularized regression designed to handle data-rich environments by optimally selecting relevant predictors[13].

5.1.1 Estimation and Evaluation

We estimate the Baseline and News-Augmented models using a rolling window of 60 months. This dynamic approach is crucial because the set of news topics that are informative for household income may change over time. Some topics might be highly relevant during certain economic periods (e.g., a recession or a housing boom) but less so in others. Our framework captures this by allowing the Elastic Net to dynamically select the most predictive subset of news topics for each specific 60-month window. This is motivated by the finding in Bianchi et al. (2022), who show that even when sparse models are optimal, the precise information utilized by an efficient forecasting algorithm changes from period to period.

For each 60-month estimation window, we generate a genuine out-of-sample prediction for realized income growth 12 months ahead. This process is then repeated by rolling the estimation window forward one period at a time, which generates a full time series of forecasts for our evaluation sample. This out-of-sample validation is critical, as it ensures our findings are not based on the benefit of hindsight and that the machine model can only differ from the survey forecast if it demonstrably improves prediction based on information available ex ante.

We compare the accuracy of all three forecasts (survey expectations, baseline model, and news-augmented model) by calculating their respective Mean Squared Errors (MSEs) over the evaluation period. A comparison of the MSEs allows us to determine whether correcting for an average bias (via the Baseline Model) or for the inefficient use of news (via the News-Augmented Model) leads to statistically and economically meaningful improvements in predictive accuracy over households' own forecasts.

5.2 Results

The results from our out-of-sample forecasting exercise are presented in Table 2 for the full sample and disaggregated by income terciles. The top row shows the Mean Absolute Error (MAE) of the raw survey forecasts, illustrating that the scale of households' forecast errors is quite large – 2.5 percentage points for the average forecast in the full sample and even larger

[13] The parameters of the elastic net are chosen to minimize 'BIC' in the training data.

for certain subgroups like the bottom income tercile. The subsequent rows then evaluate the performance of our corrective models. We first compare the Mean Squared Error (MSE) of our Baseline model to the raw survey forecast ($MSE_{\text{Baseline}}/MSE_{\text{MSC}}$), and then assess the marginal improvement from news by comparing the News-Augmented model to the already-corrected Baseline model ($MSE_{\text{News}}/MSE_{\text{Baseline}}$).

Table 2: Out-of-Sample MSE-Ratios for Realized Income Growth

		Income Group		
	Full Sample	Bottom	Middle	Top
MAE MSC	0.025	0.038	0.029	0.028
$MSE_{\text{Baseline}}/MSE_{\text{MSC}}$	0.52	0.34	0.30	0.36
Improvement Baseline	48.2%	66.3%	69.8%	64.3%
$MSE_{\text{News}}/MSE_{\text{Baseline}}$	0.81	0.80	0.78	0.70
Improvement News	19.2%	19.9%	22.3%	29.5%

Note: This table displays the out-of-sample Mean Squared Error (MSE) ratios of the raw household survey forecast (MSC) compared to our Baseline model (Equation 4) and of the Baseline model compared to the News-Augmented model (Equation 5). The columns "Bottom", "Middle", and "Top" refer to households grouped by income terciles.

First, we find strong evidence that household expectations suffer from large, systematic biases unrelated to their use of specific news information. The top panel of Table 2 compares the accuracy of our Baseline model, which corrects for a constant bias and mis-weighing of expectations, to the raw survey forecasts. The MSE ratio $MSE_{\text{Baseline}}/MSE_{\text{MSC}}$ is 0.52 for the full sample, indicating that this simple correction process reduces the mean squared forecast error by nearly half (48.2%). This substantial improvement holds across all income terciles. This demonstrates that a significant portion of households' prediction errors can be attributed to persistent biases, such as generalized optimism or pessimism, or a tendency to overweight private judgment. This finding is consistent with prior literature on household expectations (cite cite). However, quantifying this effect is a crucial first step in our analysis. By correcting for these general distortions in our Baseline model, we establish a more challenging benchmark. This allows us to better disentangle the impact of general biases from the specific effect of how households process news, meaning any additional predictive power from the News-Augmented model can be more clearly attributed to the inefficient use of information from the media.

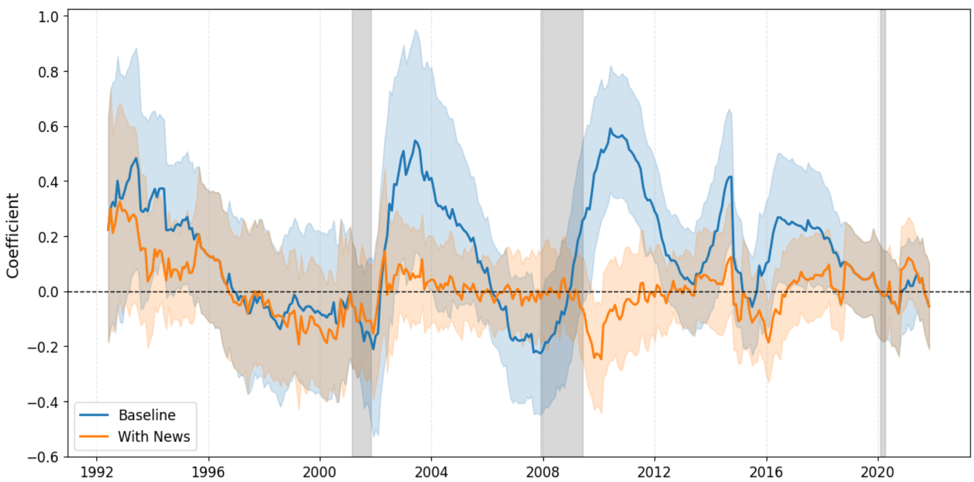
Second, turning to our main result, we find that on top of these general biases, households inefficiently process publicly available information from news media. The bottom panel of Table 2 shows the marginal improvement of the News-Augmented model over the already-corrected Baseline model. The $MSE_{\text{News}}/MSE_{\text{Baseline}}$ ratio of 0.81 indicates that by systematically processing information from 70 news topics, our model can reduce the remaining forecast error by an additional 19.2% on average. This result provides direct, out-of-sample evidence for our central hypothesis: a significant component of belief distortion stems from the mis-weighing of public information.

The gains from leveraging news information are substantial across all income groups and follow an upward trend with income. The forecast improvement is 19.9% for the bottom tercile, 22.3% for the middle tercile, and rises to a remarkable 29.5% for the top tercile. This pattern suggests that while all households fail to efficiently process news, the information contained in the media is particularly relevant for higher-income households, who appear to mis-weigh this valuable public information to a greater degree. This finding is noteworthy, as prior literature on other belief distortions often finds biases such as excessive pessimism to be more pronounced among low-income (or low education) households[14]. Our results therefore suggest that the inefficient processing of public information from the media is a distinct phenomenon, evident across all three income-terciles and most significant for the highest-earning group.

5.2.1 Over-weighing of Private Information or Judgment

Figure 3 displays the time-varying coefficient of income expectations (β) from our rolling-window estimations. This provides a dynamic view of how our models weigh household forecasts over time.

Figure 3: Estimated Coefficient of Income Expectations



Note: Estimates of the coefficient of income expectations $\hat{\beta}$ in the baseline model (blue) and news-augmented model (orange) with 95%-Confidence Intervals. The model is estimated using the full sample.

The coefficient from the baseline model (blue line) is persistently and significantly below one, providing dynamic evidence that households consistently overweight their private information or judgment when forming expectations. The inclusion of news topics in the news-augmented model causes the coefficient on expectations (orange line) to drop to nearly zero for the entire sample period. This suggests that the predictive content of household expectations is largely subsumed by the information contained within the news topics.

The figure hints at a potential cyclical pattern. The baseline model's coefficient on expectations (blue line) appears to rise during the recessions of the early 2000's and 2008-2010. While this pattern is not uniform, one tentative interpretation is that the perceived informational content of households' judgment may become more valuable during uncertain times. However, even during these episodes, the corresponding coefficient in the news-augmented model (orange

[14]See for example Rozsypal and Schlafmann (2023) for income expectation, but similarly studies find that belief distortion with respect to other variables like inflation are larger among low-income or low-education households (D'Acunto et al., 2023).

line) remains near zero, suggesting that an efficient processing of the information in news topics would still be the more robust forecasting strategy.

5.2.2 Inefficient use of Information in News-Topics

Figure 4 illuminates *which* public information households use inefficiently by showing the specific news topics our algorithm selects over time. The selection of a topic ($\gamma_k \neq 0$) implies that it held predictive power for realized income that was not captured by households' expectations in an efficient way.

Figure 4: Selected Topics Over Time

Note: Some of the selected topics of the news-augmented model. The date on the x-axis represents the prediction-period for which a topic was selected, therefore it was selected based on the training period of the previous 5 years.

The plot provides a compelling justification for our dynamic, rolling-window approach, clearly showing that the set of topics relevant in our model is not static but changes in response to major economic and social events. We can observe distinct shifts in which topics are selected that align closely with major historical episodes and support the *narrative realism* of our approach. In the aftermath of the Global Financial Crisis of 2008, the"Real Estate" topic, previously ignored by the model, became a highly persistent and important predictor of household income. Similarly, during the dot-com bubble of the early 2000s, topics such as "Internet Technology" and "Stock Market" were frequently selected by the model. The model also demonstrates a sensitivity to non-economic shocks. The "National Security" topic, for example, was selected almost exclusively in the years immediately following the events of September 11, 2001.

5.2.3 Differentiating between Over- and Under-reaction to News

While our results clearly suggests that part of the observed forecast errors is due to a misweighting of information from news topics, we cannot distinguish between over- and underreaction to news. In both cases, either if households react stronger to news topics than they should, or if they react less to news topics than they should, our estimated model would downweigh households income expectations and give positive weight to the respective news topic. Both over- and underreaction are plausible and they can be present simultaneously in the sense that households may overreact to some topics while they underreact to other topics.

6 Conclusion

Households' expectations for their own income are a cornerstone of economic decision-making, yet the source of the well-documented distortions in these beliefs remains a critical open question. This paper identifies a potential source: the inefficient processing of public information from the news media. While prior literature has established that news media can shape expectations for aggregate variables like inflation, it is not self-evident that the same holds for personal income, where private signals are likely more important. This paper addresses this gap by not only investigating whether news media shapes personal income expectations, but also assessing whether households incorporate this information efficiently.

Our empirical analysis sheds light on these questions. First, we establish that news media content is indeed a significant predictor of households' personal income expectations. In an out-of-sample forecasting exercise, a model augmented with news topics consistently outperforms a demographic baseline and offers predictive performance that is competitive with, and marginally better than, a challenging benchmark model comprised of over 120 standard macroeconomic indicators. The significant predictive power of news suggests that households' expectations do react to news content, although our analysis does not admit a causal interpretation.

Building on this finding, our analysis then addresses the question of efficiency of expectations. We demonstrate that households incorporate information provided by the news media in a systematically inefficient manner. While their forecasts are subject to large general biases, we show that a substantial portion of the remaining forecast error can be explained by mis-weighting of news topics. This inefficiency appears to stem from households overweighting their private information or judgment; the predictive content of their expectations is largely subsumed once publicly available news is optimally processed. Finally, we show that the nature of this inefficiency is dynamic. The efficient forecasting weights assigned to news topics change over time, with our model selecting a shifting set of topics that often align with major economic events, such as "Real Estate" after the 2008 financial crisis and "Internet Technology" during the dot-com bubble.

Our findings have important implications for both household finance and macroeconomic theory. Given that personal income expectations are a critical driver of consumption and savings decisions, the large, systematic forecast errors are of significant concern, as they may lead to suboptimal choices by households. This paper contributes to understanding the source of these biases by identifying a specific channel: the inefficient processing of publicly available information from the news media, which is arguably more accessible to households than the official economic statistics. For economic theory, our results add to a growing literature that has moved beyond the full-information rational expectations benchmark. We show that while households do use public information like news as signals to form beliefs, they do so inefficiently. This suggests that simply relaxing the full-information assumption may be insufficient; models may also need to incorporate deviations from rational processing to fully capture how expectations are formed in reality.

This study is subject to several limitations that offer avenues for future work. First, our analysis is based on the content of two major national newspapers, *The New York Times* and *The Wall Street Journal*. The findings may only generalize to the extent that the content in these outlets is correlated with other media sources like local news, television, or social media. Second, while we demonstrate that a mis-weighting of public information explains a substantial part of households' forecast errors, our model cannot distinguish between over-reaction and under-reaction to specific news topics. Finally, our empirical framework establishes a strong predictive relationship but does not identify a causal link between news consumption and expectation formation, as news reports may be correlated with other unobserved signals that households receive.

Our findings point to several promising avenues for future research. Foremost among them is a deeper investigation into the heterogeneity of belief formation. While our results indicate that households' reactions to news are not uniform, future work could investigate whether there are distinct "types" of households, each characterized by a different pattern of belief formation. For example, some types might systematically overreact to certain topics while others underreact to different ones. A crucial next step would be to determine which observable household characteristics – such as income, education, or media consumption habits – best predict these patterns. Addressing the limitations of this study, future research could also use survey experiments to isolate the causal effect of different news topics or sentiment on income expectations or extend the textual analysis to a wider array of media sources. Together, these lines of inquiry would provide a clearer understanding of *why* different households process information the way they do, a necessary prerequisite for deriving any potential policy implications.

References

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2).

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bianchi, F., Kung, H., and Morales, G. (2019). Growth, slowdowns, and recoveries. *Journal of Monetary Economics*, 101:47–63.

Bianchi, F., Ludvigson, S. C., and Ma, S. (2022). Belief Distortions and Macroeconomic Fluctuations. *American Economic Review*, 112(7):2269–2315.

Blei, D. M. (2003). Latent Dirichlet Allocation.

Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2020). Overreaction in Macroeconomic Expectations. *American Economic Review*, 110(9):2748–2782.

Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024). Business News and Business Cycles. *The Journal of Finance*, 79(5):3105–3147.

Cocco, J. F., Gomes, F., and Lopes, P. (2019). Evidence on Expectations of Household Finances. *SSRN Electronic Journal*.

Coibion, O. and Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, 105(8):2644–2678.

D'Acunto, F., Malmendier, U., and Weber, M. (2023). What do the data tell us about inflation expectations? In *Handbook of Economic Expectations*, pages 133–161. Elsevier.

Das, Kuhnen, C. M., and Nagel, S. (2020). Socioeconomic Status and Macroeconomic Expectations. *The Review of Financial Studies*, 33(1):395–432.

D'Haultfoeuille, X., Gaillac, C., and Maurel, A. (2021). Rationalizing rational expectations: Characterizations and tests. *Quantitative Economics*, 12(3):817–842.

Doms, M. and Morin, N. (2004). Consumer Sentiment, the Economy, and the News Media.

D'Acunto, F. and Weber, M. (2024). Why Survey-Based Subjective Expectations Are Meaningful and Important. *Annual Review of Economics*, 16:329–357.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*. *The Quarterly Journal of Economics*, 133(2):801–870.

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

Lamla, M. J. and Maag, T. (2012). The Role of Media for Inflation Forecast Disagreement of Households and Professional Forecasters. *Journal of Money, Credit and Banking*, 44(7):1325–1350.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Larsen, V. H., Thorsrud, L. A., and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117:507–520.

Loughran, T. and McDonald, B. (2024). Loughran-McDonald Master Dictionary w/ Sentiment Word Lists.

Malmendier, U. and Shen, L. S. (2024). Scarred Consumption. *American Economic Journal: Macroeconomics*, 16(1):322–355.

Massenot, B. and Pettinicchi, Y. (2019). Can households see into the future? Survey evidence from the Netherlands. *Journal of Economic Behavior & Organization*, 164:77–90.

Maćkowiak, B., Matějka, F., and Wiederholt, M. (2023). Rational Inattention: A Review. *Journal of Economic Literature*, 61(1):226–273.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Medovikov, I. (2016). When does the stock market listen to economic news? New evidence from copulas and news wires. *Journal of Banking & Finance*, 65:27–40.

Rozsypal, F. and Schlafmann, K. (2023). Overpersistence Bias in Individual Income Expectations and Its Aggregate Implications. *American Economic Journal: Macroeconomics*, 15(4):331–371.

Soroka, S. N. (2006). Good News and Bad News: Asymmetric Responses to Economic Information. *The Journal of Politics*, 68(2):372–385.

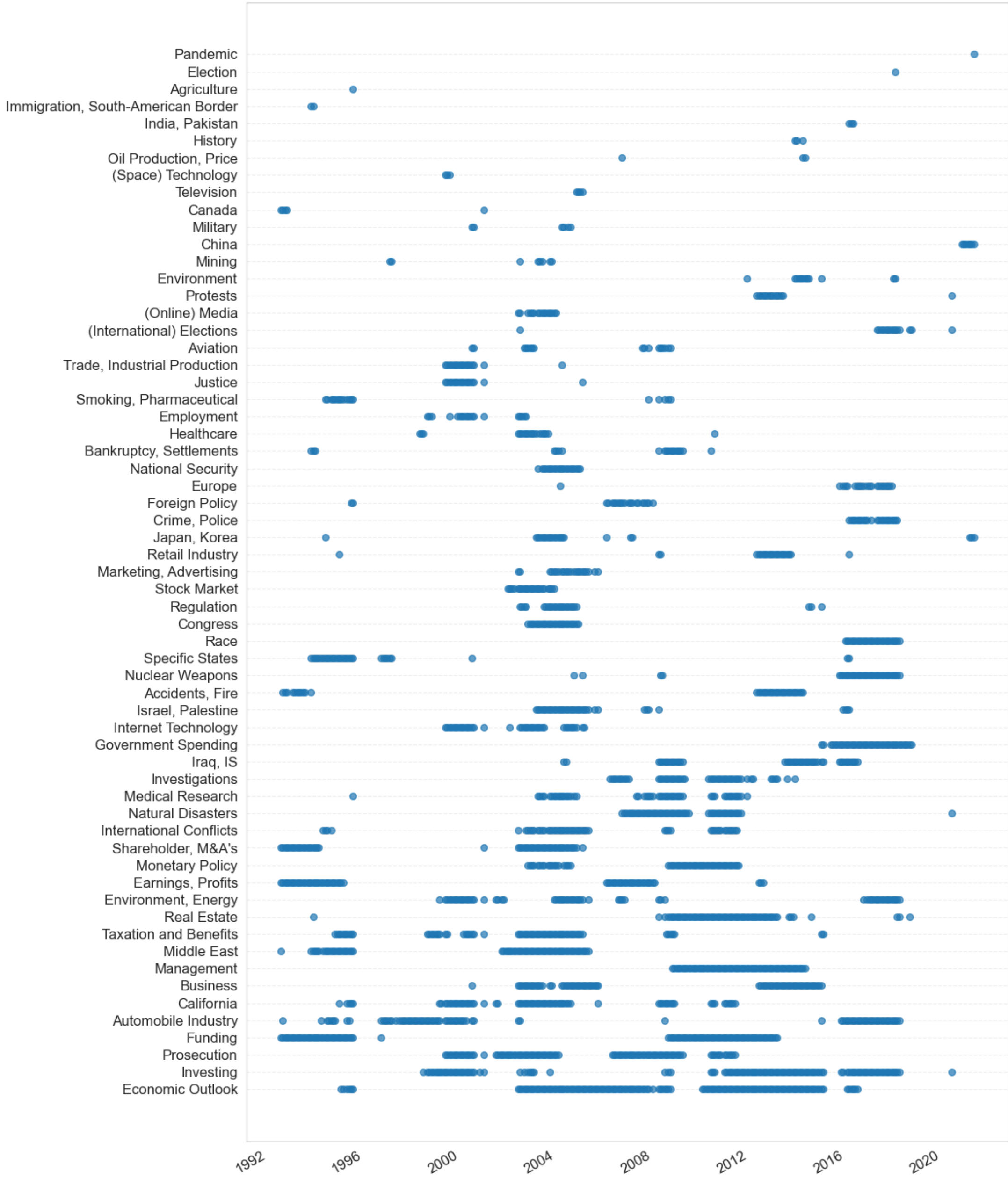
Souleles, N. S. (2004). Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys. *Journal of Money, Credit, and Banking*, 36(1):39–72.

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

A Additional Figures

Figure 5: Selected Topics Over Time



Note: These are the selected topics for the realized income regression, average.

B Tables

Table 3: News Topics.
Subjective labelling and the most important words

ID	Label	Top words
0	Congress	house, republican, senate, bill, democrat, committee, congress
1	Healthcare	health_care, insurance, hospital, medical, cost, medicare, coverage
2	Environment	water, ship, land, environmental, waste, sea, forrest, river, fish
3	Russia	soviet_union, russia, gorbachev, communist, nato, putin, kremlin
4	Justice	court, judge, law, justice, appeal, decision, ruling, legal
5	California	california, los_angeles, san_francisco, pacific, silicon_valley, bay_area
6	Europe	european_union, europe, german, french, france, euro, swiss, italy
7	Iraq, IS	iraq, iraqi, islamic_state, syria, hussein, syrian, war, baghdad
8	Employment	worker, union, job, employee, work, labor, strike, wage, pay
9	Earnings, Profits	quarter, share, earnings, sale, profit, revenue, loss, analyst, result
10	Real Estate	real_estate, building, home, property, office, housing, construction, development
11	Shareholder, M&A's	share, stock, offer, deal, plan, shareholder, sell, merger, acquisition
12	Funding	bank, bond, loan, debt, billion, yield, credit, treasury, interest_rate
13	Pandemic	test, virus, health, disease, vaccine, pandemic, infection, spread, outbreak
14	Education	school, student, university, college, education, program, teach
15	Marketing, Advertising	ad, advertising, agency, marketing, campaign, brand, commercial
16	Economic Outlook	percent, rate, rise, decline, growth, consumer, economy, survey
17	Trade, Industrial Production	trade, industry, steel, export, import, production, american, manufacturer, factory
18	Foreign Policy	united_states, president, administration, meeting, nation, foreign_policy, agreement
19	Immigration, South-American Border	mexico, government, immigration, latin_america, border, foreign
20	National Security	security, intelligence, agent, agency, attack, information, terrorist, secret_service
21	Environment, Energy	power, plant, energy, utility, electric, gas, environmental, emission, cost, air, coal
22	Random*	go, even, many, like, get, people, think, much, well, way
23	Agriculture	farm, price, farmer, food, crop, corn, agriculture, grain, wheat
24	Nuclear Weapons	nuclear, iran, weapon, missile, arm, sanction, treaty, test
25	(Online) Media	site, online, internet, magazine, newspaper, web, publisher, book, google
26	Israel, Palestine	israel, israeli, palestinian, jewish, peace, west_bank, arab, gaza, jerusalem
27	Management	executive, president, chief, chairman, board, director, management
28	Race	city, black, white, new_york, community, minority, hispanic, racial, african, race
29	Accidents, Fire	safety, fire, problem, accident, crash, cause, agency, report
30	Taxation and Benefits	tax, pay, income, money, plan, benefit, payment, pension, taxpayer
31	Random**	go, get, like, day, people, old, come, work, back, even
32	Smoking, Pharmaceutical	drug, tobacco, product, patent, pharmaceutical, johnson, cigarette, fda, smoking
33	Canada[15]	british, canada, canadian, london, britain, australia, toronto, ireland, england
34	History	world_war, american, book, history, century, life, art, museum, society
35	Regulation	rule, law, commission, require, regulation, government, industry, policy
36	Bankruptcy, Settlements	firm, file, bankruptcy, settlement, lawsuit, claim, creditor, settle
37	Middle East	saudi_arabia, gulf, kuwait, persian, middle_east, country, emirate
38	Monetary Policy	interest_rate, dollar, economy, central_bank, fed, inflation, price, monetary
39	Election	campaign, republican, president, candidate, election, democratic, voter, party, poll
40	Television	television, cable, network, tv, news, show, film, program, movie, disney, channel
41	Stock Market	stock_market, price, trading, index, rise, exchange, investor, dow_jones, share, trader
42	China	china, chinese, hong_kong, beijing, asia, taiwan, singapore, economic
43	Investing	fund, investor, investment, firm, stock, money, security, asset, mutual
44	Oil Production, Price	oil, price, gas, energy, barrel, production, pipeline, opec, drilling
45	Turkey, Greece	turkey, eu, turkish, greece, steelmaker, cyprus
46	Retail Industry	store, sale, retailer, food, product, consumer, sell, retail, customer, wal_mart
47	Japan, Korea	japan, japanese, yen, korea, tokyo, foreign, ministry, nissan
48	(Catholic) Church	church, hotel, catholic, casino, religious, pope, bishop, las_vegas, vatican
49	Prosecution	charge, case, prosecutor, trial, lawyer, prison, sentence, jury, criminal
50	Family[16]	woman, child, family, old, parent, abortion, life, father, sexual
51	Announcements	week, tuesday, wednesday, monday, thursday, plan, deal, comment, statement
52	Investigations	report, investigation, department, official, information, document, review
53	Government Spending	program, government, budget, cut, spending, deficit, fund, cost, fiscal
54	Military	military, force, defense, army, general, war, pentagon, troop, navy, soldier
55	India, Pakistan	india, pakistan, state, gandhi, tribe, delhi, hindu, minister, sikh, tamil
56	Business	business, market, big, industry, executive, deal, venture, firm, analyst
57	Sport	game, card, sport, club, team, player, league, football, baseball
58	(Space) Technology	system, technology, space, satellite, design, test, research, launch, shuttle
59	Protests[17]	government, right, protest, country, demonstration, movement, anti, vietnam
60	International Conflicts	rebel, africa, united_nation, war, troop, refugee, serb, guerilla, peace, nato
61	Automobile Industry	car, auto, motor, ford, vehicle, sale, chrysler, truck, gm
62	Natural Disasters[18]	mile, city, area, storm, town, flood, railroad, damage, rail, hurricane
63	Internet Technology	computer, software, service, technology, microsoft, phone, customer, apple, chip
64	(International) Elections	party, election, government, prime_minister, vote, opposition, power
65	Medical Research	dr, study, patient, research, cancer, disease, researcher, medical, treatment
66	Specific States	state, county, texas, florida, governor, new_york, local, national, ohio
67	Mining	gold, mine, mining, italian, ounce, metal, copper, silver, coal, commodity
68	Crime, Police	police, kill, officer, people, death, attack, gun, man, arrest, violence
69	Aviation	airline, flight, plane, carrier, pilot, airport, boeing, travel, continental

C Additional Results

This section provides supplementary analysis to support the main findings of the paper. We extend the investigation to a broader range of household subgroups to test the robustness of our results and explore heterogeneity in how households' income expectations relate to news media.

C.1 News-Driven Income Expectations?

In this section, we expand on the analysis from Section 4.2, which demonstrated that news topics have predictive power for the income expectations of different income terciles. We test whether this finding holds across more granular income quintiles and other key demographic groups, including education level, sex, and age. The results are presented in Tables 4 and 5.

Table 6 shows the out-of-sample MSE ratios when the sample is divided into income quintiles. The results are consistent with the main findings, showing that the News-Augmented Model improves upon the baseline for all quintiles. Interestingly, the improvement is most pronounced for the lowest income quintile (Q1), with a 17.9% reduction in MSE. The predictive power of news then generally decreases as income rises, with the smallest improvement seen for the highest quintile (Q5) at 6.2%. This more granular view refines the U-shaped pattern observed in the tercile analysis (Table 1). While the bottom end of the income distribution consistently shows strong responsiveness to news, the upper end appears less uniform. For both news and the FRED benchmark, the improvement is stronger for Q1 than any other group.

Table 4: MSE-Ratios for Income Expectations - **Income Quintiles**

	Full Sample	Q1	Q2	Q3	Q4	Q5
$MSE_{\text{News}}/MSE_{\text{Baseline}}$	0.92	0.82	0.88	0.89	0.91	0.94
Improvement News	7.8%	17.9%	12.4%	11.3%	9.0%	6.2%
$MSE_{\text{FRED}}/MSE_{\text{Baseline}}$	0.95	0.81	0.90	0.89	0.92	0.94
Improvement FRED	4.9%	19%	9.8%	11.2%	8.0%	5.9%

Note: The table reports the ratio of out-of-sample mean squared errors and relative improvements, calculated as the percentage change in MSE. The three models are indicated by subscripts 'Baseline' (equation 1), 'News' (equation 2) and 'FRED' (equation 3).

Table 5 presents the forecasting results for households grouped by education, sex, and age. The News model consistently outperforms the FRED model, providing further evidence that media narratives contain information beyond standard macroeconomic indicators. The improvement from news is most significant for males (11.0%) and households with a college degree (8.1%). In contrast, the improvement for females (2.9%) and those without a college degree (5.9%) is more modest. As discussed in the main text regarding the middle-income tercile, this does not necessarily mean these latter groups are less reactive to news. Instead, groups like 'Female' or 'No College' might be more heterogeneous in their economic situations and the specific news topics that are relevant to them. When the diverse reactions within such a group are averaged, the overall predictive signal our model can detect is weaker.

Table 5: MSE-Ratios for Income Expectations - **Other Household Groups**

Metric	College	No College	Male	Female	Under 45	Over 45
$MSE_{\text{News}}/MSE_{\text{Baseline}}$	0.92	0.94	0.89	0.97	0.91	0.94
Improvement News	8.1%	5.9%	11.0%	2.9%	9.4%	6.1%
$MSE_{\text{FRED}}/MSE_{\text{Baseline}}$	0.92	0.97	0.92	0.95	0.92	0.98
Improvement FRED	7.6%	2.4%	7.7%	4.7%	8.0%	1.8%

Note: The table reports the ratio of out-of-sample mean squared errors and relative improvements, calculated as the percentage change in MSE. The three models are indicated by subscripts 'Baseline' (equation 1), 'News' (equation 2) and 'FRED' (equation 3).

C.2 Mis-Weighing of Information from News

We show additional out-of-sample forecasting results, examining the performance of our models across different income quintiles (Table 6) and other demographic groups (Table 7). The results broadly confirm the main findings presented in the main text. For all groups analyzed — including by education, sex, and age —the News-Augmented model provides a clear improvement over the Baseline model, which has already corrected for general systematic biases. This suggests that the inefficient processing of news is a robust phenomenon across different household types.

While these results reinforce our main conclusion, the heterogeneity in the magnitude of the improvements should be interpreted with caution. A smaller measured improvement for a particular group does not necessarily imply that the group is more rational or less prone to mis-weighing news. There are at least two factors that could lead to lower estimates for certain groups.

First, as we disaggregate the sample into smaller demographic groups, the number of monthly observations per group decreases. This can increase the noise in the dependent variable (average group expectations), which may mechanically attenuate the measured predictive power of our model.

Second, and perhaps more importantly, some of the demographic groupings may be more heterogeneous than others in terms of their news-reaction patterns. Our model can best detect a signal if a group represents a coherent 'type' of household that reacts uniformly to a specific set of news topics. However, if a defined group is actually a mix of different underlying types, their distinct reactions could average out, leading to a muted signal and a lower estimated improvement. For example, the relatively low improvements for the 'Under 45' (12.9%) and 'Over 45' (12.2%) age groups could be a result of this effect. These are extremely diverse categories, each containing households with vastly different occupations, income sources, and life stages. It is plausible that the true underlying 'types' that react to news are not perfectly captured by these broad demographic labels, leading to an underestimation of the true effect of news mis-weighting within these groups.

Table 6: MSE-Ratios for Realized Income Growth - **Income Quintiles**

	Full Sample	Q1	Q2	Q3	Q4	Q5
$MSE_{\text{Baseline}}/MSE_{\text{MSC}}$	0.52	0.33	0.31	0.36	0.31	0.47
Improvement Baseline	48.2%	66.9%	68.7%	64.5%	69.0%	52.6%
$MSE_{\text{News}}/MSE_{\text{Baseline}}$	0.81	0.92	0.97	0.91	0.82	0.95
Improvement News	19.2%	8.4%	2.9%	8.6%	18.2%	4.8%

Table 7: MSE-Ratios for Realized Income Growth - **Other Household Groups**

	College	No College	Male	Female	Under 45	Over 45
$MSE_{\text{Baseline}}/MSE_{\text{MSC}}$	0.40	0.50	0.49	0.40	0.66	0.28
Improvement Baseline	59.7%	50.5%	51.3%	60.2%	33.7%	72.4%
$MSE_{\text{News}}/MSE_{\text{Baseline}}$	0.82	0.84	0.91	0.77	0.87	0.88
Improvement News	18.4%	15.5%	9.3%	23.1%	12.9%	12.2%