Measuring Economic Sentiment from Open-Ended Survey Comments Using Large Language Models

Pascal Seiler*
ETH Zurich

September 2025

Abstract

This article develops a novel economic sentiment indicator (LLM-ESI) by applying large language models to open-ended responses from Swiss business tendency surveys. Using a BERT-based transformer model, it extracts firm-level sentiment from free-text survey comments and aggregates it into a high-frequency indicator of macroeconomic conditions. The LLM-ESI closely tracks the business cycle and performs on par with, or better than, traditional benchmarks in nowcasting GDP. These results highlight the potential of large language models and open-ended survey responses to deliver timely and nuanced signals for real-time economic analysis.

JEL classification: C55, C53, E32, E37, E66
Keywords: Economic Sentiment, Large Language Model, Business Tendency Surveys, Survey Comments, Textual Analysis, Forecasting

*Correspondence to: ETH Zurich, KOF Konjunkturforschungsstelle, Leonhardstrasse 21, 8092 Zurich, Switzerland. Telephone: +41 44 632 89 44. E-mail address: seiler@kof.ethz.ch. Webpage: https://www.pascalseiler.ch/.

1 Introduction

Economic sentiment indicators are essential for monitoring macroeconomic conditions in real time. Traditionally, these indicators are constructed from closed-ended survey questions (e.g., Abberger et al., 2014; Wegmueller and Glocker, 2024) or from textual data such as news articles, analyzed using word counts (e.g., Baker et al., 2016; Altig et al., 2020) or sentiment dictionaries (e.g., Shapiro et al., 2022). These approaches are well established and provide timely signals of business and consumer sentiment.

An alternative, yet largely underutilized[1], source of sentiment information lies in open-ended survey responses–voluntary free-text comments submitted by firms alongside standard questionnaires. These comments offer rich insights into firms' expectations and first-order concerns (Ferrario and Stantcheva, 2022). While traditional text analysis methods can be applied to such unstructured inputs, recent advances in natural language processing (NLP) offer more powerful alternatives. In particular, large language models (LLMs) can capture tone, context, and meaning far beyond the capabilities of earlier tools, making them especially well-suited for extracting sentiment from free-text comments. Their adoption in economics is growing rapidly, including applications in central bank communication (Gorodnichenko et al., 2023; Hansen and Kazinnik, 2024) and economic forecasting (Yu et al., 2023; Faria-e Castro and Leibovici, 2024).

This paper introduces a novel LLM-based Economic Sentiment Indicator (LLM-ESI), constructed from voluntary free-text comments in Swiss business tendency surveys. To my knowledge, this is the first systematic effort to quantify firm-level sentiment from open-ended survey responses using LLMs. The LLM-ESI is built using a state-of-the-art transformer model, enabling it to extract nuanced signals from a previously underexploited data source. It updates in near real time and spans a long historical window, offering high-frequency insights into the business cycle.

I show that the LLM-ESI closely tracks macroeconomic conditions and exhibits strong cyclical properties. In a real-time pseudo out-of-sample forecasting exercise, the indicator performs on par with or better than traditional benchmarks and an AR(1) model–particularly for current-quarter nowcasts. These results highlight the potential of the LLM-ESI as a timely and robust tool for economic analysis and forecasting.

By linking qualitative survey responses to quantitative analysis, this approach demonstrates how LLMs can unlock high-frequency economic information from free-text in-

[1] Exceptions remain scarce, with only a handful of studies exploring such comments systematically (e.g., Yotzov et al., 2021; Gerardin and Ranvier, 2021).

puts. The resulting indicator combines immediacy, forward-looking content, and long historical coverage–making it a valuable addition to the real-time monitoring toolkit of economists and policymakers.

2 Methodology

2.1 Firm-level Comments in Business Tendency Surveys

This paper uses open-ended, voluntary comments submitted by firms responding to the KOF Business Tendency Surveys, which cover most of the private sector in Switzerland. Conducted monthly or quarterly depending on the sector, the surveys receive around 4,500 responses per wave (60 percent response rate), completed online or on paper in German, French, Italian, or English.

The surveys collect qualitative assessments of recent and expected developments in business indicators such as demand, costs, or employment.[2] At the end of each survey, firms may optionally provide additional comments in an open-ended text box. These free-text responses allow them to elaborate on their situation, add nuance, or raise firm-specific or policy-related concerns not captured by closed-ended questions (Ferrario and Stantcheva, 2022).

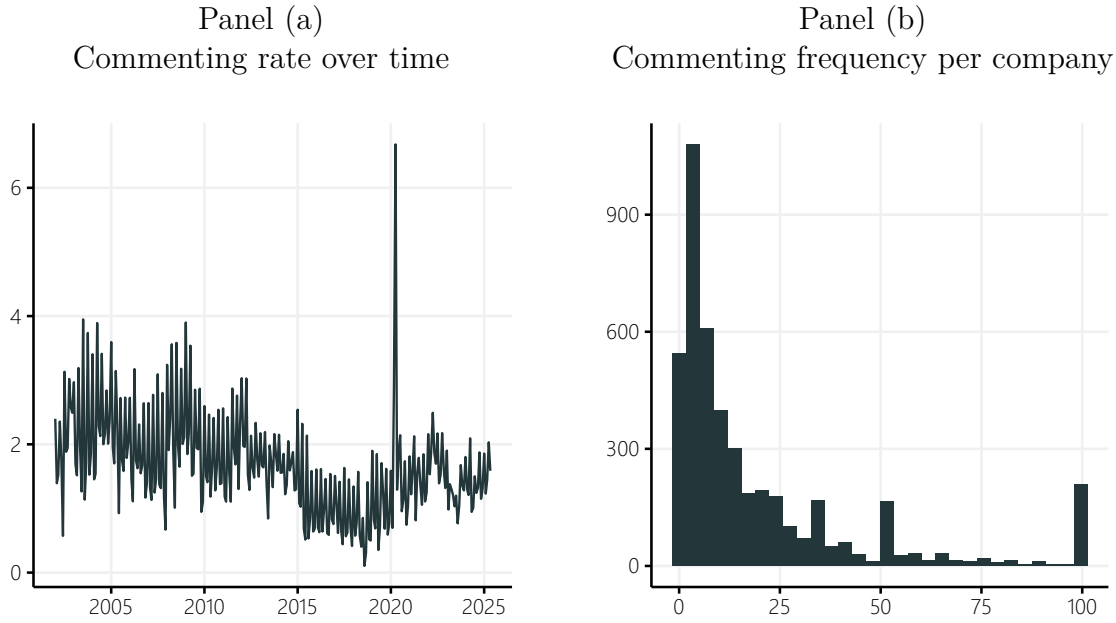
The dataset includes all comments submitted online between January 2002 and May 2025. Out of over 460,000 responses, 22,397 include a comment, resulting in an average commenting rate of 4.8 percent. As shown in Panel (a) of Figure 1, this rate varies over time and increases notably during periods of economic turmoil (e.g., 2008 financial crisis, COVID-19 pandemic).

To focus on substantive content, I exclude boilerplate phrases (e.g., "kind regards"), administrative notes, and entries under three characters. I then apply standard pre-processing steps: converting text to lowercase, removing digits and special characters, and tokenizing by whitespace. The final cleaned dataset consists of 19,862 comments. Comments range from one-word notes to multi-sentence explanations (mean: 19 words), and typically address firm-specific conditions, including sales, costs, and macroeconomic developments (e.g., "EU" or "COVID"). Figure A.1 in Online Appendix A illustrates the most frequently used terms by language-specific word clouds, confirming the relevance of these comments for tracking firm sentiment in real time. The comments originate from 4,552 firms. While most never comment (67 percent),

[2] An example of the monthly questionnaire used in the manufacturing sector is provided in Online Appendix A.

among those who do, the average commenting frequency is 19.4 percent. Panel (b) of Figure 1 shows that commenting is broadly distributed across firms, although some are particularly vocal. Most comments are written in German (77 percent), followed by French (17 percent), Italian (6 percent), and English (below 1 percent). Table A.1 in Online Appendix A summarizes the distribution of comments by firm characteristics.

Figure 1. Firm-level commenting behavior



Notes: Panel (a) shows the monthly share of survey respondents who submit a comment (in percent). Panel (b) shows the distribution of firms' commenting frequency, measured as the share of times a firm leaves a comment (in percent), conditional on having commented at least once.

2.2 Large Language Model: Selection and Specification

To extract sentiment from firms' survey comments, I use the pre-trained XLM-T model (Barbieri et al., 2022), a sentiment classification variant of the XLM-RoBERTa base model (Conneau et al., 2019), trained on 200 million tweets from the social media platform X (formerly Twitter) across 30 languages and fine-tuned for sentiment analysis in eight of them, including German, French, Italian, and English.Because it was trained on social-media texts, XLM-T is well suited to handle the informal, short-form, and often noisy nature of voluntary business survey comments, which–like tweets–frequently contain typos, nonstandard phrasing, and idiosyncratic grammar.

Moreover, relying on a pre-trained model rather than training one from scratch or fine-tuning on proprietary data offers both practical and methodological advantages: such models already encode extensive linguistic and domain knowledge, including economic terminology, and avoid the risk of overfitting or memorizing sensitive business data–important considerations for reproducible academic research.

XLM-T is built on the RoBERTa architecture (Liu et al., 2019), which itself extends BERT, a bidirectional transformer encoder (Vaswani et al., 2017). This architecture captures full sentence context via multi-headed attention and positional encoding, allowing it to differentiate between semantically distinct uses of words based on surrounding tokens.

To classify sentiment, the model transforms each input comment into a contextualized representation and outputs a probability distribution across three sentiment classes: positive, neutral, and negative. Formally, for each comment i at time t, the model produces probabilities $P^{+}_{i,t}$, $P^{0}_{i,t}$, and $P^{-}_{i,t}$, which quantify the likelihood that the comment expresses positive, neutral, or negative sentiment, respectively.

2.3 Construction of the LLM-Based Economic Sentiment Indicator (LLM-ESI)

To quantify economic sentiment from the business survey comments, I assign each comment a sentiment score based on the LLM's output probabilities. Specifically, each score $S_{i,t}$ is computed as a weighted average of the predicted sentiment probabilities:

$$S_{i,t} = 1 \cdot P^{+}_{i,t} + 0 \cdot P^{0}_{i,t} + (-1) \cdot P^{-}_{i,t}. \tag{1}$$

This results in a continuous sentiment measure ranging from -1 (fully negative) to $+1$ (fully positive), capturing the tone of each individual comment. Table A.2 in Online Appendix A presents the comments with the most extreme sentiment scores.

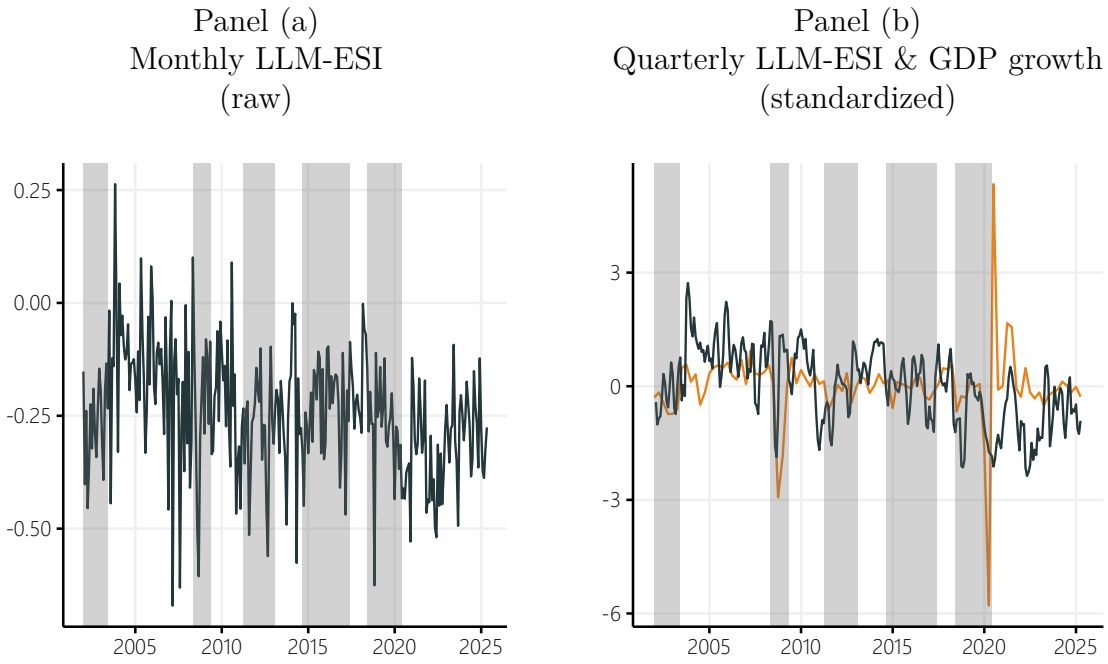
To construct the aggregate LLM-ESI, I first compute monthly weighted averages of firm-level sentiment scores within sectors–industry, construction, retail, and other services–using full-time equivalent employment weights. Then, I aggregate these averages to create the monthly composite LLM-ESI, using the European Commission's fixed sector weights (45% for industry, 10% for construction, 10% for retail, and 35% for other services), aligning with ESI methodology (European Commission, 2025).[3]

[3]The aggregation follows a two-step procedure for two reasons: (i) the European Commission weights are based on sectoral representativeness (typically measured by value-added) and their ability to track GDP growth, and (ii) this approach is standard in the construction of comparable indicators. Retaining this method ensures methodological consistency and comparability of the LLM-ESI with

3 Empirical Analysis of the LLM-ESI

3.1 In-sample Properties and Cyclical Properties

Figure 2. The LLM-based Economic Sentiment Indicator (LLM-ESI)



Notes: Time series of the LLM-based Economic Sentiment Indicator (LLM-ESI) derived from firms' comments in the KOF Business Tendency Surveys. Panel (a) shows the raw monthly LLM-ESI. Panel (b) shows the quarterly LLM-ESI and compares it to real quarter-on-quarter GDP growth. Both series have been standardized to have a mean of 0 and a standard deviation of 1. The sample period is 2002:01–2025:05. The shaded areas depict recessions as dated by the OECD.

Figure 2 shows the LLM-ESI from January 2002 to May 2025. Panel (a) shows the monthly raw series, and Panel (b) shows the quarterly standardized series together with quarter-on-quarter GDP growth. Conditional time series moments are reported in Table B.1 in Online Appendix B. On average, sentiment is negative, reflecting a tendency among firms to comment when dissatisfied rather than satisfied. The indicator has low variance and moderate volatility (as measured by the coefficient of variation). Skewness and kurtosis are close to Gaussian, suggesting that the LLM-ESI primarily captures regular business cycle variation rather than rare, extreme

established benchmarks.

shocks. The first-order autocorrelation implies moderate persistence, with sentiment innovations fading after about one month.

The LLM-ESI tracks key macroeconomic events over the sample period. It declines following the dot-com crash, the Global Financial Crisis, the European debt crisis, the Swiss National Bank's abandonment of the minimum exchange rate, the COVID-19 pandemic, and the Russian invasion of Ukraine. This responsiveness confirms that the indicator reflects firms' contemporaneous assessments of economic conditions

To assess cyclicality more formally, I regress the standardized LLM-ESI on several business cycle measures, including an OECD recession dummy, the KOF Economic Barometer, and real GDP growth. As shown in Table 1, sentiment is significantly lower during recessions and increases with GDP growth or the KOF Barometer. These results confirm the cyclicality of the LLM-ESI and its strong co-movement with conventional measures of economic activity.

Table 1. Business cycle properties of the LLM-ESI

	LLM-ESI		
	(1)	(2)	(3)
Recession dummy	-0.289^{***}		
	(0.057)		
KOF Economic Barometer		0.019^{***}	
		(0.003)	
Real GDP growth			0.096^{***}
			(0.021)
Constant	0.168^{***}	-1.894^{***}	-0.170^{***}
	(0.038)	(0.305)	(0.059)
Observations	1,245	1,405	465
R^2	0.020	0.027	0.047
Adjusted R^2	0.020	0.026	0.045
Residual Std. Error	0.997 (df = 1243)	0.985 (df = 1403)	0.969 (df = 463)
F Statistic	25.776^{***} (df = 1; 1243)	38.816^{***} (df = 1; 1403)	22.825^{***} (df = 1; 463)

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Each column presents the results of a regression of the LLM-ESI on various measures of the business cycle, including a recession dummy based on OECD-dated recessions (in column 1), the KOF Economic Barometer (in column 2), and real GDP growth (in column 3). The sample period is 2002:01–2025:05. The LLM-ESI is standardized. For the quarterly GDP data, the regression uses the sentiment indicator aggregated to quarterly frequency.

I also explore heterogeneity in sentiment expression across firm characteristics and the timing of responses (Online Appendix B). Sentiment is more negative among smaller firms, those in construction, and comments in French or Italian; male respondents also express slightly more negative sentiment. Temporally, sentiment dips midweek and is

more negative in early afternoon responses. These patterns indicate that sentiment reflects not only macroeconomic conditions but also firm-level and behavioral factors.

3.2 Pseudo out-of-sample analysis

To assess the usefulness of the LLM-ESI for nowcasting and short-term forecasting of economic activity, I follow Burri and Kaufmann (2020) and conduct a pseudo out-of-sample forecast evaluation for quarterly real GDP growth using real-time data vintages. Specifically, I estimate:

$$y_{\tau+h} = \alpha_h + \beta_{h,1} f_{\tau|t} + \beta_{h,2} f_{\tau-1} + \nu_{\tau+h}, \tag{2}$$

where $y_{\tau+h}$ denotes quarterly GDP growth at horizon h, with τ indexing calendar quarters. The variable $f_{\tau|t}$ represents the current-quarter LLM-ESI, constructed as the average of monthly LLM-ESI values up to month t within quarter τ. The lagged sentiment term $f_{\tau-1}$ captures information from the previous quarter, and $\nu_{\tau+h}$ is an error term.

Forecasts are generated in real time using only the information available when a new quarterly GDP becomes available each month. This rolling procedure produces nowcasts ($h = 0$) and one- to four-quarter-ahead forecasts ($h \in \{1, 2, 3, 4\}$). I benchmark the LLM-ESI against an AR(1) model and four leading indicators of the Swiss economy: the KOF Economic Barometer, the KOF Economic Sentiment Indicator, the OECD Composite Leading Indicator, and the SECO Swiss Economic Confidence.[4] Forecast accuracy is evaluated using root-mean-squared forecast errors (RMSE) with the most recent available GDP vintage. Table 2 summarizes the results.

[4] Table B.2 and Figure B.1 in Online Appendix B provide descriptions and plots of these indicators. Furthermore, Figure B.2 shows that the LLM-ESI exhibits significant leading relationships with most of the indicators in cross-correlation analyses.

Table 2. Pseudo real-time evaluation of the LLM-ESI

	Horizon	RMSE	Relative RMSE LLM-ESI / Benchmark	DMW test (p-value) LLM-ESI < Benchmark
LLM-ESI	$h = 0$	1.28		
	$h = 1$	1.30		
	$h = 2$	1.32		
	$h = 3$	1.33		
	$h = 4$	1.33		
(a) AR(1)	$h = 0$	1.41	0.91	0.043
	$h = 1$	1.31	0.99	0.394
	$h = 2$	1.31	1.00	0.534
	$h = 3$	1.36	0.97	0.277
	$h = 4$	1.36	0.98	0.163
(b) KOF Barometer	$h = 0$	1.26	1.02	0.637
	$h = 1$	1.37	0.95	0.054
	$h = 2$	1.33	0.99	0.377
	$h = 3$	1.33	0.99	0.366
	$h = 4$	1.32	1.02	0.643
(c) KOF ESI	$h = 0$	1.79	0.77	0.041
	$h = 1$	1.45	0.90	0.061
	$h = 2$	1.41	0.93	0.577
	$h = 3$	1.41	0.94	0.517
	$h = 4$	1.70	0.78	0.114
(d) OECD CLI	$h = 0$	1.24	1.03	0.811
	$h = 1$	1.38	0.95	0.061
	$h = 2$	1.33	1.00	0.496
	$h = 3$	1.34	0.99	0.420
	$h = 4$	1.33	1.01	0.653
(e) SECO SEC	$h = 0$	1.34	0.96	0.272
	$h = 1$	1.34	0.97	0.061
	$h = 2$	1.33	1.00	0.499
	$h = 3$	1.35	0.98	0.309
	$h = 4$	1.34	1.00	0.504

Notes: This table summarizes the results of the pseudo out-of-sample analysis. The root-mean-squared errors (RMSE) are reported for forecasts made on the release dates of new quarterly GDP data. The horizon $h = 0$ refers to the current-quarter forecast, while $h \in \{1, 2, 3, 4\}$ correspond to one- to four-quarter-ahead forecast. I compare the LLM-ESI (in the top row) to five benchmark models: (a) an AR(1) model, (b) the KOF Economic Barometer, (c) the KOF Economic Sentiment Indicator, (d) the OECD Composite Leading Indicator, and (e) the SECO Swiss Economic Confidence. The Diebold-Mariano-West (DMW) evaluates whether differences in forecast accuracy relative to the benchmark are statistically significant, based on a quadratic loss function (Diebold and Mariano, 2002; West, 1996). The sample period is 2002:01–2025:05.

Panel (a) shows that the LLM-ESI outperforms the AR(1) model for current-quarter forecasts, reducing RMSE by 9% (the Diebold-Mariano-West test is significant at the 5% level). For forecasts over longer horizons, RMSEs are similar and not statistically

different.

Panels (b) through (e) compare the LLM-ESI to standard leading indicators. For nowcasts ($h = 0$), performance is broadly comparable, though the LLM-ESI significantly outperforms the KOF ESI (23% lower RMSE). For one-quarter-ahead forecasts ($h = 1$), the LLM-ESI yields lower RMSEs than all benchmarks, with improvements ranging from 1–5%. These differences are statistically significant for the KOF ESI (5%) and marginally so for the Barometer and OECD CLI (10%). For forecasts over longer horizons, the LLM-ESI does not significantly improve over the benchmarks.

Overall, the LLM-ESI outperforms an autoregressive model and delivers forecasting accuracy on par with, or slightly better than, established business cycle indicators–especially for one-quarter-ahead forecasts.

4 Conclusion

This paper introduced a novel LLM-based Economic Sentiment Indicator (LLM-ESI) derived from open-ended business survey comments. Using a state-of-the-art transformer model the approach captures nuanced firm sentiment at scale and in near real-time. The resulting indicator aligns closely with macroeconomic developments and, in a pseudo-real-time forecasting exercise, outperforms an AR(1) model while matching or exceeding the performance of established leading indicators, particularly for current-quarter forecasts.

By linking qualitative survey responses to quantitative analysis, this approach shows how LLMs can unlock high-frequency, forward-looking information from free-text inputs. The resulting indicator combines immediacy with macroeconomic relevance–making it a compelling tool for real-time economic analysis and policy monitoring.

Future research could extend the LLM-ESI framework by incorporating structured survey responses alongside free-text comments, and by applying topic modeling to better capture the themes underlying the sentiments expressed.

References

Abberger, Klaus, Michael Graff, Boriss Siliverstovs, and Jan-Egbert Sturm (2014) "The KOF economic barometer, version 2014: a composite leading indicator for the Swiss business cycle," Working Paper 353, KOF Swiss Economic Institute.

Altig, Dave, Scott Baker, Jose Maria Barrero, Nicholas Bloom, Philip Bunn, Scarlet Chen, Steven J Davis, Julia Leather, Brent Meyer, Emil Mihaylov et al. (2020) "Economic uncertainty before and during the COVID-19 pandemic," Journal of Public Economics, Vol. 191, p. 104274.

Baker, Scott R, Nicholas Bloom, and Steven J Davis (2016) "Measuring economic policy uncertainty," The Quarterly Journal of Economics, Vol. 131, pp. 1593–1636.

Barbieri, Francesco, Luis Espinosa Anke, and Jose Camacho-Collados (2022) "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond," in Proceedings of the Language Resources and Evaluation Conference, pp. 258–266, Marseille, France: European Language Resources Association, June.

Burri, Marc and Daniel Kaufmann (2020) "A daily fever curve for the Swiss economy," Swiss Journal of Economics and Statistics, Vol. 156, p. 6.

Faria-e Castro, Miguel and Fernando Leibovici (2024) "Artificial Intelligence and Inflation Forecasts," Federal Reserve Bank of St. Louis Review, Vol. 106, pp. 1–14.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019) "Unsupervised cross-lingual representation learning at scale," Preprint 1911.02116, arXiv.

Diebold, Francis X and Robert S Mariano (2002) "Comparing predictive accuracy," Journal of Business & Economic Statistics, Vol. 20, pp. 134–144.

European Commission (2025) "The Joint Harmonised EU Programme Of Business And Consumer Surveys: User Guide," Directorate-General For Economic And Financial Affairs.

Ferrario, Beatrice and Stefanie Stantcheva (2022) "Eliciting people's first-order concerns: Text analysis of open-ended survey questions," AEA Papers and Proceedings, Vol. 112, pp. 163–169.

Gerardin, Mathilde and Martial Ranvier (2021) "Enrichissement de l'Enquête Mensuelle de Conjoncture de la Banque de France: enseignements de l'analyse textuelle des commentaires des chefs d'entreprise," Working Paper 821, Banque de France.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (2023) "The voice of monetary policy," American Economic Review, Vol. 113, pp. 548–584.

Hansen, Anne Lundgaard and Sophia Kazinnik (2024) "Can ChatGPT decipher Fedspeak?" Working Paper 4399406, SSRN.

Indergand, Ronald and Stefan Leist (2014) "A real-time data set for Switzerland," Swiss Journal of Economics and Statistics, Vol. 150, pp. 331–352.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019) "RoBERTa: A robustly optimized BERT pretraining approach," Preprint 1907.11692, arXiv.

Neusser, Klaus (2016) Time series econometrics, Vol. 1: Springer.

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J Wilson (2022) "Measuring news sentiment," Journal of Econometrics, Vol. 228, pp. 221–243.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017) "Attention is all you need," Advances in Neural Information Processing Systems, Vol. 30.

Wegmueller, Philipp and Christian Glocker (2024) "Capturing Swiss economic confidence," Swiss Journal of Economics and Statistics, Vol. 160, p. 3.

West, Kenneth D (1996) "Asymptotic inference about predictive ability," Econometrica, pp. 1067–1084.

Yotzov, Ivan, Nick Bloom, Philip Bunn, Paul Mizen, Pawel Smietanka, and Greg Thwaites (2021) "What matters to firms? New insights from survey text comments," Bank Underground Blog Post, Bank of England.

Yu, Xinli, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu (2023) "Temporal data meets LLM–explainable financial time series forecasting," Preprint 2306.11025, arXiv.