

FX sentiment analysis with large language models

Daniele Ballinari, Jessica Maly

SNB Working Papers

11/2025



EDITORIAL BOARD SNB WORKING PAPER SERIES

Marc-Antoine Ramelet
Enzo Rossi
Rina Rosenblatt-Wisch
Pascal Towbin
Lukas Frei

DISCLAIMER

The views expressed in this paper are those of the author(s) and do not necessarily represent those of the Swiss National Bank. Working Papers describe research in progress. Their aim is to elicit comments and to further debate.

COPYRIGHT©

The Swiss National Bank (SNB) respects all third-party rights, in particular rights relating to works protected by copyright (information or data, wordings and depictions, to the extent that these are of an individual character).

SNB publications containing a reference to a copyright (© Swiss National Bank/SNB, Zurich/year, or similar) may, under copyright law, only be used (reproduced, used via the internet, etc.) for non-commercial purposes and provided that the source is mentioned. Their use for commercial purposes is only permitted with the prior express consent of the SNB.

General information and data published without reference to a copyright may be used without mentioning the source. To the extent that the information and data clearly derive from outside sources, the users of such information and data are obliged to respect any existing copyrights and to obtain the right of use from the relevant outside source themselves.

LIMITATION OF LIABILITY

The SNB accepts no responsibility for any information it provides. Under no circumstances will it accept any liability for losses or damage which may result from the use of such information. This limitation of liability applies, in particular, to the topicality, accuracy, validity and availability of the information.

ISSN 1660-7716 (printed version)
ISSN 1660-7724 (online version)

© 2025 by Swiss National Bank, Börsenstrasse 15,
P.O. Box, CH-8022 Zurich

FX Sentiment Analysis with Large Language Models

Daniele Ballinari* and Jessica Maly†

August 15, 2025

Abstract

We enhance sentiment analysis in the foreign exchange (FX) market by fine-tuning large language models (LLMs) to better understand and interpret the complex language specific to FX markets. We build on existing methods by using state-of-the-art open source LLMs, fine-tuning them with labelled FX news articles and then comparing their performance against traditional approaches and alternative models. Furthermore, we test these fine-tuned LLMs by creating investment strategies based on the sentiment they detect in FX analysis articles with the goal of demonstrating how well these strategies perform in real-world trading scenarios. Our findings indicate that the fine-tuned LLMs outperform the existing methods in terms of both the classification accuracy and trading performance, highlighting their potential for improving FX market sentiment analysis and investment decision-making.

JEL Classification: F31, G12, G15

Keywords: Large Language Models, Sentiment Analysis, Fine-tuning, Text Classification, Natural Language Processing, Foreign Exchange, Financial Markets.

*Swiss National Bank (SNB); daniele.ballinari@snb.ch

†University of St. Gallen and Thurgau Institute for Digital Transformation at the University of Konstanz and at the HTWG Konstanz (TIDIT); jessica.maly@tidit.ch

We thank Francesco Audrino and the entire Technology and Data Science team of the Money Market and Foreign Exchange division at the Swiss National Bank (SNB) as well as seminar and conference participants at the SNB and Columbia University for helpful comments and discussions. The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the SNB. The SNB takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.

1 Introduction

The foreign exchange (FX) market, with a daily trading volume exceeding 7 trillion US Dollars (USD) (BIS, 2022), is the largest financial market globally, playing a pivotal role in the global financial landscape. The language used in FX market analyses is exceptionally complex and features specialized jargon and the inherent relativity of currency pairs, such as the price of one euro in terms of US dollars (EUR/USD), which poses significant challenges for accurate sentiment classification. Traditional methods often struggle to capture these nuances, but recent advancements in natural language processing (NLP), particularly with large language models (LLMs), offer transformative potential. LLMs, with their ability to process and generate human-like text, provide innovative tools to automate and enhance the interpretation of FX market language, paving the way for more precise and actionable sentiment analysis.

Despite the success of sentiment analysis in equity and other financial markets, its application in the FX market remains relatively unexplored. This research aims to address this gap by fine-tuning and evaluating the performance of state-of-the-art LLMs for FX sentiment analysis. The research questions addressed in this work are as follows: How do fine-tuned LLMs, such as Meta’s Llama 3.1 (Grattafiori et al., 2024), perform when analysing sentiment in FX markets compared to traditional methods and existing financial models like FinBERT (Araci, 2019)? How can sentiment scores generated by fine-tuned LLMs be effectively utilized to construct trading strategies, and how do these strategies perform in real-world scenarios? The results demonstrate that fine-tuned LLMs offer strong potential for sentiment analysis in the FX market. Leveraging both human-labelled and return-labelled FX news data, this study addresses the unique challenges of currency-specific sentiment. The fine-tuned LLMs achieve high classification accuracy while maintaining computational efficiency during training, and the models’ sentiment scores enable the construction of trading strategies that deliver robust performance, as measured by annualized returns, volatility, and Sharpe ratios, underscoring the practical utility of LLMs in the financial domain.

The implications of this research are profound. By fine-tuning LLMs specifically for the FX market, this paper highlights their potential to interpret complex financial language and market dynamics. This work underscores the importance of integrating LLMs into financial analysis workflows, offering researchers, analysts, and market participants a powerful tool to better understand and respond to market sentiment. Analysing sentiment in FX news articles is particularly relevant because these texts

often serve as key drivers of market perception and trading behaviour. News outlets such as FXStreet and Investing.com provide timely and detailed coverage of market developments, offering critical insights into currency trends and economic conditions. The nuanced language and technical jargon used in these articles reflect both the prevailing sentiment among market participants and expectations for future movements. By distinguishing between past and future sentiment, this research captures the temporal dynamics of FX market language, and the proposed approach provides valuable tools for interpreting sentiment-driven market behaviour, making it highly applicable to both academic research and real-world financial decision-making.

The first contribution of this study is an evaluation that involves fine-tuning a LLM for a specialized financial domain, specifically the FX market, by using limited labelled data. We explore two fine-tuning approaches: human labelling for high-quality annotations and distant labelling to scale data generation. Furthermore, we introduce a novel framework for distinguishing between past and future sentiment labels in FX news articles, effectively addressing the temporal dynamics of currency-specific sentiment. To address the scarcity of labelled datasets, we combine human-labelled data with distant labelling, which links sentiment to observed returns around the publication dates of news articles, enabling efficient and scalable dataset creation.

The fine-tuned LLM achieves notable improvements in FX sentiment analysis, outperforming traditional lexicon-based methods such as *Valence Aware Dictionary and sEntiment Reasoner* (VADER) (Hutto & Gilbert, 2014), financial models such as FinBERT and untuned LLMs in terms of both accuracy and F1 scores. The distinct classification between past and future sentiment enhances the model’s ability to capture temporal nuances in FX market language. Furthermore, the results demonstrate the effectiveness of combining high-quality human annotations with scalable distant labelling approaches, highlighting the potential of advanced fine-tuning techniques in terms of their ability to address the complexities of FX sentiment.

Another key contribution of our approach is the simultaneous classification of sentiment for the ten most traded currencies (commonly referred to as the G10 currencies). Traditional methods typically assign a single sentiment score to an entire article, whereas we leverage the text generation capabilities of LLMs, such as Llama 3.1, to generate sentiment labels for each G10 currency mentioned. Furthermore, as previously noted, the LLM provides both forward- and backward-looking sentiment scores for each currency, offering a more comprehensive perspective.

A third contribution is our demonstration of the effectiveness of trading with the signals obtained from the LLM. Our results show that the fine-tuned Llama 3.1 model with 8 billion parameters outperforms traditional sentiment analysis methods in terms of predicting FX movements, particularly when using articles published on DailyFX and Investing.com. The model consistently generates positive cumulative returns, suggesting it effectively captures analysts’ views and the underlying textual patterns that precede market shifts. However, performance varies across data sources, with dictionary-based approaches performing competitively in some cases. These findings highlight the importance of fine-tuning and domain adaptation when leveraging LLMs for financial sentiment analysis.

Literature review: The evolution of sentiment analysis within the financial domain has witnessed significant advancements, particularly with the integration of LLMs. Traditional lexicon-based approaches, such as VADER and the Loughran-McDonald dictionary (Loughran & McDonald, 2011), have been widely used but often struggle with context sensitivity and financial terminology (Mishev, Gjorgjevikj, Vodenska, Chitkushev, & Trajanov, 2020). The emergence of transformer-based models has introduced more sophisticated methods for handling financial texts. FinBERT, an adaptation of BERT specifically designed for financial sentiment analysis, has shown substantial improvements over traditional methods. By pre-training on financial documents and fine-tuning for sentiment tasks, FinBERT achieves higher accuracy and better understanding of financial jargon (Araci, 2019). However, despite its capabilities, this approach faces certain challenges. It struggles with sensitivity to numerical data, and its classification accuracy tends to decline as sentence complexity increases, which might be attributable to its relatively modest size of 110 million parameters.

Recent research has focused on making these models more capable and tailored for financial applications. The FinGPT and Instruct-FinGPT models, for example, use small LLMs, such as the Llama 1 and Llama 2 models, as a base and fine-tune them with financial instructions (Liu, Wang, & Zha, 2023; Yang, Liu, & Wang, 2023; Zhang, Yang, & Liu, 2023). The FinLlama model builds on Llama 2 with 7 billion parameters and employs parameter-efficient fine-tuning techniques like low-rank adaptation (Hu et al., 2021) to reduce computational demands while maintaining high performance in sentiment classification (Konstantinidis, Iacovides, Xu, Constantinides, & Mandic, 2024). In addition to these advancements, Bloomberg also developed an LLM, BloombergGPT, which is a large-scale model developed specifically for finance that combines a vast corpus of financial and general-purpose texts to enhance its capabilities (Wu et al., 2023). Moreover, multimodal financial language models have ad-

vanced rapidly, enabling simultaneous reasoning over visual, tabular, numerical, and textual inputs. A notable example is the FinTral model family, which extends a Mistral-7B model with finance-specific pre-training, instruction tuning, and reinforcement learning from AI feedback to support diverse tasks ranging from chart interpretation to document Q&A. It consistently outperforms ChatGPT-3.5 and matches or surpasses GPT-4 on several finance benchmarks (Bhatia, Nagoudi, Cavusoglu, & Abdul-Mageed, 2024).

These models represent significant advances in financial AI, yet they face critical limitations when applied to FX sentiment analysis. For instance, models such as BloombergGPT and FinGPT are not open source, which restricts their accessibility and customization by the broader research community. Furthermore, while models like FinBERT and FinLlama are trained for sentiment analysis, they are not specifically tailored for the FX market. Most critically, multimodal LLMs, despite their impressive general financial capabilities, are not designed to handle the unique challenges of FX sentiment analysis, including the relative nature of currency pairs and specialized FX jargon of currency-specific news. This gap highlights the need for models that are not only open source but are also specifically trained for sentiment analysis in the FX domain.¹

This research contributes to the existing literature by advancing the application of LLMs in FX sentiment analysis. Specifically, we build on previous studies by fine-tuning the state-of-the-art model Llama 3.1 with 8 billion parameters and use labelled FX news data to address the unique challenges inherent to FX markets, such as the relative nature of currency pairs and the specialized jargon of FX analysis. Unlike existing approaches, which often rely on generic sentiment models or lexicon-based methods, we tailor LLMs to the FX domain, enhancing their ability to interpret nuanced language and relative sentiment within currency pairs.

Recognizing the scarcity of openly accessible labelled FX news datasets, we contribute to the literature by constructing a novel dataset sourced from three major FX platforms. To achieve this objective, we employ a hybrid labelling approach: (1) human labelling, which ensures high-quality annotations for a limited number of examples, and (2) distant labelling, which scales the dataset by assigning sentiment labels based on the returns observed before and after each article’s publication. This dual approach balances annotation quality with scalability. Furthermore, we distinguish past and future sentiment in FX analysis articles to provide a more nuanced understanding of temporal

¹The fine-tuned Llama 3.1 model weights used in this study are available upon request from the authors.

dynamics in currency-specific sentiment. By separately labelling and analysing past and future sentiment, we enhance the interpretability of the model’s outputs, allowing for a clearer differentiation between historical market conditions and forward-looking expectations in FX news. Moreover, we utilize Quantized Low-rank Adaptation (QLoRA) (Dettmers, Pagnoni, Holtzman, & Zettlemoyer, 2023) for fine-tuning, a method that significantly reduces computational requirements while maintaining high performance. This strategy allows for the fine-tuning of large models with accessible resources, contributing to the democratization of advanced sentiment analysis techniques in the financial domain.

We also contribute to the growing literature employing neural networks in asset pricing (e.g. Audrino, Gentner, and Stalder, 2024; Chen, Kelly, and Xiu, 2022; Cheng and Chin, 2025; Schuettler, Audrino, and Sigris, 2024). Kelly, Kuznetsov, Malamud, and Xu (2025) implanted a transformer inside the stochastic discount factor, cutting the pricing errors relative to earlier machine-learning models. Giglio, Kelly, and Xiu (2022) surveyed how machine learning reshapes factor models, and Chen, Pelger, and Zhu (2024) showed that deep-learning stochastic discount factors built under no-arbitrage dominate linear competitors in terms of their out-of-sample Sharpe ratios. Moreover, Lopez-Lira and Tang (2023) documented that ChatGPT’s zero-shot headline sentiment predicts daily equity returns and partially closes market-efficiency gaps.

Finally, we bridge the gap between theoretical advancements and practical applications by evaluating the real-world utility of fine-tuned LLMs. By constructing trading portfolios based on the generated sentiment scores, we demonstrate how these models can inform investment strategies. By using financial metrics such as the annualized return, annualized volatility, and Sharpe ratio, we provide a rigorous assessment of the models’ ability to manage risk and generate returns. This study not only advances methodological approaches in FX sentiment analysis but also highlights their practical implications for market participants and financial analysts.

The remainder of this paper is structured as follows: Section 2 provides a detailed description of the dataset utilized in this study. Next, Section 3 outlines the model selection, fine-tuning process, and evaluation of the model’s performance. Subsequently, Section 4 presents a trading application that leverages the fine-tuned models, and Section 5 assesses the robustness of the results. Finally, the study is concluded in Section 6.

2 Data

We collected publicly available FX articles from three well-known sources: Investing.com, DailyFX, and FXStreet. These platforms are recognized for their high-quality financial content and extensive reach among FX professionals. For Investing.com we specifically scraped the analysis section of the FX category obtained from <https://www.investing.com/analysis/forex/>. This section features detailed insights on FX markets, authored by market analysts, economists, and independent contributors. The FX analysis section of Investing.com adheres to editorial guidelines to ensure relevance, accuracy, and originality, with a focus on technical analysis, macroeconomic trends, and trading strategies. DailyFX provides timely FX news and analysis written by analysts with expertise in FX markets. We sourced these articles from <https://www.dailyfx.com/archive/>. The FXStreet articles were retrieved from <https://www.forexcrunch.com/blog/author/fxstreet/> with contributions from staff writers, independent analysts, and guest contributors. FXStreet is visited by more than 1 million individual users per month (see <https://about.fxstreet.com/become-contributor/>). Each of these three sources ensures quality through experienced authorship and editorial oversight, making them reliable datasets for FX market research.

To ensure data quality, we applied a series of filtering steps to the scraped articles. First, articles containing fewer than 20 words were removed to exclude content with insufficient informational value. Articles exceeding the CSV file character limit of 32,767 characters were truncated at this limit to ensure compatibility with our processing pipeline, as extremely long articles would require excessive computational resources when given to an LLM. Only a very small number of articles reached this limit. Additionally, duplicate articles were eliminated by checking for repetition across their titles, full text, and URLs. These filtering steps resulted in a final dataset comprising 77,789 articles from Investing.com, 45,593 articles from DailyFX, and 128,463 articles from FXStreet, as summarized in Table 1. The average daily article count is significantly higher for FXStreet (114 articles) compared to both Investing.com (17 articles) and DailyFX (15 articles). The datasets have only marginal differences in terms of the average length of individual articles, as highlighted in Table 1. DailyFX articles are the longest on average at 403 words, followed closely by Investing.com at 394 words and FXStreet at 374 words. Furthermore, the variability in article length is lowest in the FXStreet dataset, and the word count remains relatively consistent over time across all of the datasets (see Figure A.2).

		Mean	Median	Std. Dev.	Min	Max	Sum
Article Count per Day	Investing.com	17	11	14	1	60	77,789
	DailyFX	15	16	8	1	39	48,593
	FXStreet	114	145	70	2	233	128,436
Word Count per Article	Investing.com	394	308	337	20	5,303	27,923,462
	DailyFX	403	371	231	20	5,473	19'587'092
	FXStreet	374	355	126	185	2,859	48,038,153

Table 1: Summary statistics of the Investing.com, DailyFX and FXStreet text datasets used in this study.

The sources have varying coverage periods, as visualized in Figure 1: Investing.com spans from October 2011 to June 2024, DailyFX covers January 2015 to June 2024, and the FXStreet dataset includes articles published between October 2018 and June 2021. The short sample period for FXStreet is due to the limitations of the archive we scraped from (<https://www.forexcrunch.com/blog/author/fxstreet/>), which only contains articles from October 2018 to June 2021. We could not find a comprehensive archive on the FXStreet website itself to extend the dataset. While FXStreet maintains consistently high volumes, both the DailyFX and Investing.com samples show a decline in the later years. One notable change in the publication patterns is observed for DailyFX; its publication volume declined, particularly in 2023 and 2024. DailyFX was acquired by the IG Group family in 2016, and as of September 4, 2024, DailyFX no longer operates independently, with all its articles now released under the IG brand. The reduction in publication volume may reflect changes in strategic focus under IG Group ownership. Conversely, the decline in Investing.com article volume beginning in 2019 lacks an obvious explanation.

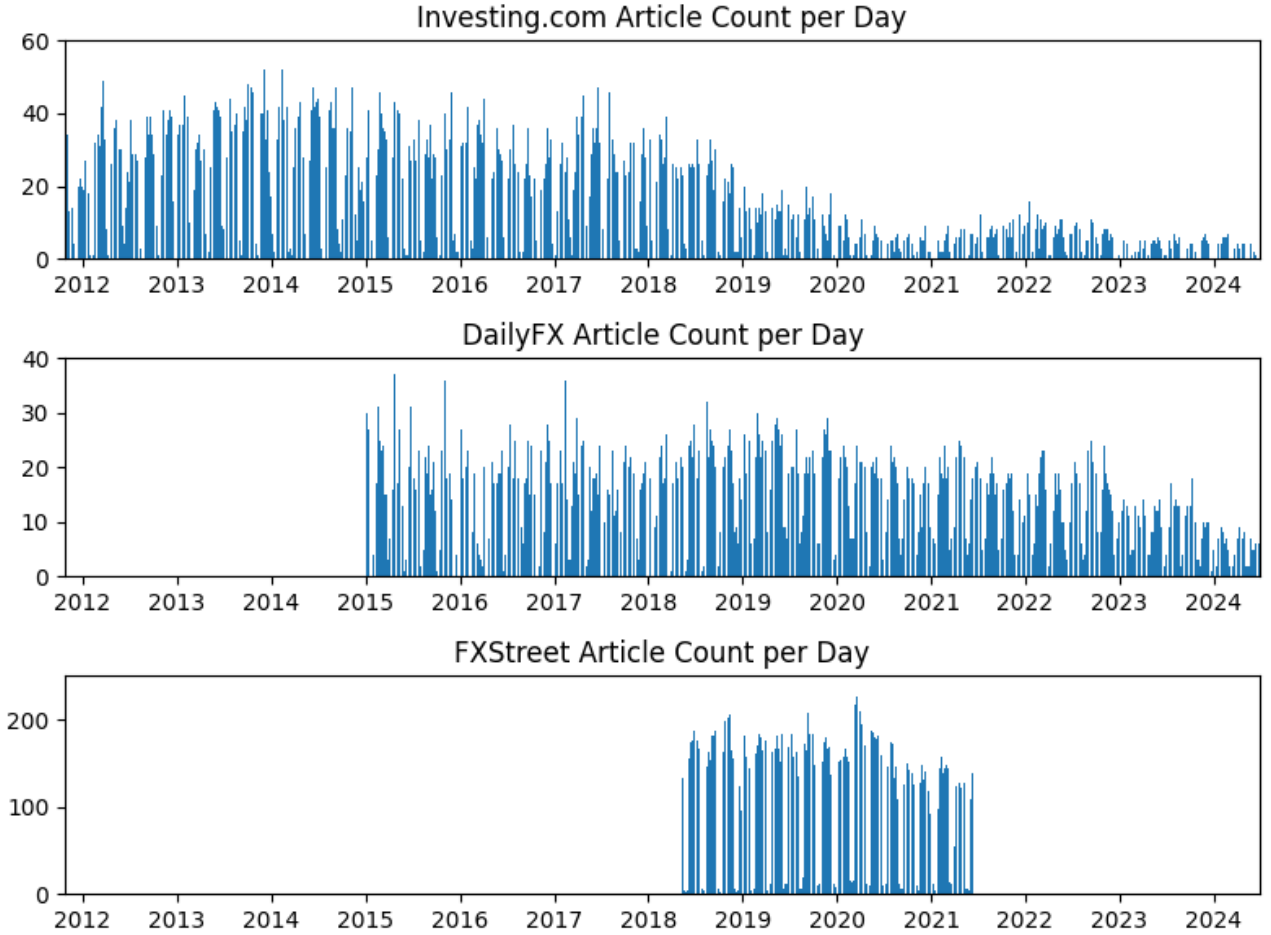


Figure 1: Daily number of articles for the Investing.com, DailyFX and FXStreet datasets over time.

The frequency of G10 currency² mentions across the datasets aligns with expectations, with the euro (EUR) and the USD being the most frequently mentioned due to their dominance in global financial markets (see Figure A.3). Other major currencies, such as the Japanese yen (JPY), British pound (GBP), and the Swiss franc (CHF), also appear prominently. Lesser-traded currencies are mentioned less frequently, which is consistent with expectations.

3 LLM selection, fine-tuning and evaluation

Training an LLM from scratch is not feasible due to the immense computational resources required. The process involves massive datasets and significant computational power, often necessitating thousands of GPUs and substantial time, translating into costs that can reach millions of dollars. Moreover, training from scratch is unnecessary given the availability of numerous pre-trained open-source LLMs. These models, which were developed by leading AI research institutions, have already undergone ex-

²The G10 currencies, i.e. the ten most traded currencies in the world, are the US dollar (USD), euro (EUR), Japanese yen (JPY), British pound (GBP), Swiss franc (CHF), Swedish krona (SEK), Norwegian krona (NOK), New Zealand dollar (NZD), Australian dollar (AUD) and the Canadian dollar (CAD).

tensive training on diverse datasets. They offer robust performance across a wide range of natural language processing tasks and can be fine-tuned for specific applications, including FX sentiment analysis. Utilizing these pre-trained models allowed us to leverage state-of-the-art capabilities without the prohibitive costs and complexity of training from scratch. However, one drawback of using open-source pre-trained models is that we have no control over their knowledge cut-off, meaning they may incorporate information from sources that extend beyond the intended historical context, potentially creating a so-called look-ahead-bias (Noguer i Alonso, 2019).

3.1 LLM selection

We selected Meta’s Llama 3.1 with 8 billion (8B for short) parameters for the FX sentiment classification of news articles due to several key advantages:

- **Performance and Accuracy:** Llama 3.1 8B has demonstrated high performance in natural language understanding tasks and offers significant advancements in various metrics when compared to other open-source LLMs. Its performance nearly matches the best-performing closed-source LLMs such as GPT-4o as visible in many LLM leaderboards (Chiang et al., 2024; Trustbit, 2024).
- **Advanced Architecture:** Llama 3.1 8B utilizes an advanced transformer architecture and incorporates grouped-query attention, which enhances its ability to process information efficiently (Meta, 2024).
- **Open-Source Access:** Llama 3.1 8B is accessible to the research and developer communities via HuggingFace’s model repository³ under Meta’s licensing. This open access allows fine-tuning to our specific use case, FX sentiment analysis.
- **Resource Efficiency:** Despite its powerful performance, Llama 3.1 8B is designed to be relatively resource-efficient. It is recognized for its balance between performance and resource usage, making it a cost-effective choice for intensive computational tasks without requiring exorbitant computational power.
- **Multilingual Capabilities:** FX markets are global and involve data in multiple languages. Llama 3.1 8B excels in multilingual support, having been trained on data in over 30 languages. This feature ensures that sentiment analysis can be effectively performed across diverse linguistic datasets, capturing sentiments from various sources around the world (Meta, 2024).

³<https://huggingface.co/meta-llama/Llama-3.1-8B>

3.2 Fine-tuning

Datasets for fine-tuning

To the best of our knowledge, there are no publicly available datasets that contain full-text articles labelled for FX sentiment. This absence presents a significant obstacle for advancing research in this domain, as labelled data is essential for training models that are capable of accurately capturing and analysing FX market sentiment. To address this challenge, we constructed two datasets: one based on human annotation and another generated through a distant labelling methodology. These datasets complement each other by balancing the trade-off between annotation quality and scalability. Both datasets comprise sampled full text articles from the data described in Section 2 up to the end of 2019.

The distant labelling dataset is inspired by methodologies proposed by Chen et al. (2022) and Ke, Kelly, and Xiu (2019). Building on the approach developed by Chen et al. (2022), who assigned sentiment labels based on the sign of the asset returns over a three-day window surrounding news release dates, and Ke et al. (2019), who derived sentiment labels by identifying the sentiment-charged words that frequently co-occurred with stock returns of the same sign with a supervised learning framework, we adopt a similar methodology to constructing sentiment labels in our analysis. We begin by computing daily log returns for G10 currencies from J.P. Morgan’s tradable currency index obtained from Bloomberg. Multi-day cumulative returns are calculated to capture meaningful market movements while smoothing short-term noise, and we assign sentiment labels based on the distribution of these multi-day returns. Specifically, we define the label for future events based on the sum of log returns over the next five trading days and for past events based on the sum of log returns over the previous five trading days. An article published on day t is then assigned a forward-looking label for each G10 currency as follows: “appreciation” if the cumulative return of a currency from $t + 1$ to $t + 5$ is among the 30% largest currency returns for that time window, “depreciation” if it is among the 30% lowest currency returns, and “unchanged” otherwise. The same strategy is used for cumulative returns from $t - 5$ to $t - 1$ for backward-looking labels. This approach allows us to systematically categorize significant market movements while minimizing the influence of smaller, less relevant fluctuations. We choose to assign 30% of the currencies with the “appreciation” (“depreciation”) label to ensure a balanced distribution among the classes (three “appreciation”, three “depreciation” and four “unchanged”). For future sentiment labels, if an article’s publication date does not coincide with a trading day, we use the next available trading day. Conversely, for past sentiment labels, we use the most recent trading day preceding the article’s publication date. Articles in which a currency is not

explicitly mentioned retain a default value of “unaffected” for that currency. By using this approach, we apply distant labelling to 30,000 randomly selected articles before 2020. This number was chosen because processing a larger dataset would not have been computationally feasible given the resources at hand. To ensure diversity and coverage across years and sources, we preprocess the articles to identify mentions of G10 currencies by using regex-based patterns and stratify the sampling process by currency mentions, publication year and source.

The human-labelled dataset serves as a high-quality dataset, and is constructed through a rigorous annotation process. To select articles for human labelling, we initially sample 1,068 articles selected randomly from the dataset by using a sampling process similar to that used for the distant labeling dataset. A batch of 100 of these articles are labelled by the authors of this study. From the 100 articles labelled by the authors, we assess three annotators by having each label the same 10 randomly selected articles. Their annotations are compared to the authors’ labels, and the annotator with the highest agreement is selected to label the remaining 868 articles. Detailed guidelines on FX sentiment classification and domain-specific nuances were provided to the annotator, and the authors were in close contact with the annotator during the labelling process to eliminate any ambiguity. To ensure consistency and accuracy, we implement extensive sanity checks on the human-labelled dataset. For example, if an article discussed a currency pair (e.g. EUR/USD) and the base currency (e.g. EUR) is labelled as appreciating, we expect the price currency (e.g. USD) to be labelled as depreciating, and vice versa. These checks reveal an error rate of only 2.7%, indicating a high level of reliability in the annotated labels. The authors also conduct spot checks on the labeled data to further validate its quality.

By combining distant labelling for scale and human labelling for precision, we create a robust dataset that serves as the foundation for fine-tuning the large language model. The distant-labelled dataset ensures sufficient coverage of the FX market, while the human-labelled dataset captures the complexities and nuances of FX sentiment.

Training process

To adapt the open source LLM model for FX-specific language and sentiment analysis, we employ QLoRA (Quantized Low-Rank Adaptation), a fine-tuning method that combines 4-bit quantization with LoRA-style parameter-efficient adaptation (Dettmers et al., 2023). This approach enables substantial memory and computational savings without compromising the model’s ability to learn task-specific features. The main model weights are quantized to 4 bits, reducing the GPU memory require-

ments for the checkpoint storage and training overhead. Simultaneously, LoRA introduces low-rank adaptation matrices into the model’s linear layers, updating only a subset of the parameters that are crucial for fine-tuning. For more details, we refer to Dettmers et al. (2023) and Hu et al. (2021).

Before training, we exclude a set of 200 human-labelled examples, 100 labelled by the authors and 100 by the professional annotator, to serve as an evaluation set after training. The remaining data is shuffled and split into training (80%) and validation (20%) subsets to ensure robust learning and rigorous evaluation. The training process integrates a detailed prompt to guide the model in learning FX-specific tasks. The prompt specifies the task objective, including the desired output format and sentiment labels (“appreciation,” “depreciation,” or “unchanged”). It also incorporates currency synonyms, such as “greenback” for USD or “sterling” for GBP, to account for real-world linguistic variations. This structured prompt ensures that the model generates consistent and interpretable outputs tailored to financial contexts (Figure A.1 details the complete prompt). For each query, the model is provided with a single newspaper article and tasked with assigning a label to each of the G10 currencies. Due to the instruction tuning applied during fine-tuning, the resulting models are expected to generate responses in the prescribed format without difficulty.

By using QLoRA, we inject low-rank adaptation matrices into all linear layers of the model. This configuration enables memory-efficient fine-tuning by freezing the quantized original weights and updating only the lightweight adaptation matrices. The LoRA hyperparameters are set with a rank of 8, a scaling factor of 16, and a dropout probability of 0.1. Training is performed over three epochs for the distant labelled dataset and ten epochs for the human-labelled dataset. We employ a tuned learning rate, small per-device batch sizes, and gradient accumulation to simulate a larger effective batch size. A paged AdamW optimizer (Loshchilov & Hutter, 2019) manages the weight updates, and an early-stopping criterion halts training if the validation loss stagnates (see Table A.1 for an overview of all training parameters). Throughout training, the performance metrics are logged by using Weights & Biases, and intermediate checkpoints are saved to enable progress monitoring and potential resumption.

By embedding instruction tuning directly within the QLoRA framework, we unify memory-efficient model adaptation with task-specific learning. This approach allows the model to efficiently learn nuanced linguistic and market-specific features, enabling accurate and interpretable FX sentiment analysis tailored to the demands of financial applications.

3.3 Evaluation

This section presents a comprehensive comparison of our fine-tuned LLMs with several established baselines for sentiment extraction in FX news articles. We compare the fine-tuned transformer-based models to its non-fine-tuned versions as well as to another fine-tuned language model for financial purposes (FinBERT, as outlined in Section 1), a lexicon-based method (VADER), and a dictionary-based method (Loughran-McDonald dictionary). VADER (Valence Aware Dictionary for Sentiment Reasoning) is a lexicon and rule-based sentiment analysis tool that is designed to work well on text from various domains but particularly excels in handling text from online platforms (Hutto & Gilbert, 2014). The Loughran-McDonald dictionary (LM-dictionary) is a widely used dictionary for sentiment analysis in the context of financial and economic texts. The dictionary is tailored specifically to address the unique language and sentiment found in financial documents, such as SEC filings (Loughran & McDonald, 2011). Most LLMs specialized in financial market language are not openly available, as outlined in Section 1, making a direct comparison to these models difficult.

We evaluate each model’s ability to classify both past and future sentiment, focusing on three distinct classes: “appreciation,” “unchanged,” and “depreciation.” We are primarily interested in evaluating a currency’s potential future value rather than its past movements, which is why we distinguish past and future sentiment. The comparisons draw on a human-labelled dataset of 200 randomly selected articles covering the period from 2011 to 2019. Half of these were labelled by the authors, and the other half by the same human labeller who provided the training annotations. In 0.05% of the cases, the untuned Llama 3.1 model fails to generate a meaningful response, and we exclude these cases from the evaluation. While all currencies and classification labels are represented in the dataset, there is a higher frequency of labels for more prominent currencies, such as the USD and EUR, which aligns with expectations given their dominant role in global financial markets (see Table A.2). We benchmark performance by using two widely adopted metrics, the classification accuracy and the F1 score. Accuracy measures the proportion of correctly classified instances among all instances, providing a straightforward assessment of overall correctness. However, it may not be reliable for imbalanced datasets in which one class dominates. The F1 score is the harmonic mean of the precision and recall, balancing the trade-off between false positives and false negatives. It is particularly useful when the dataset is imbalanced, as it ensures that both precision (correctness of positive predictions) and recall (coverage of actual positives) are considered.

Because the benchmark models such as FinBERT and VADER cannot differentiate between past and future sentiments and are unable to classify entire articles, we adopt the following approach for sentiment analysis: First, we apply FinBERT to each sentence that mentions a G10 currency in an article. It tokenizes the text and classifies sentiment into appreciation, depreciation, or neutral via a softmax layer. Second, we use VADER, a lexicon-based method, to compute a compound sentiment score for the same sentences, and predefined thresholds map these scores to the three sentiment categories. Third, we utilize the LM-dictionary, a finance-specific lexicon, to count positive and negative terms and determine the overall sentiment direction. The sentiment scores from each of these methods are aggregated for each currency, and the final classification is based on the net balance of sentiment. In cases in which a currency appears primarily as a quote currency rather than a base currency, we invert its sentiment score to align with market conventions. Since we are unable to distinguish between past and future sentiments, this procedure yields a single, general sentiment score per article. By contrast, leveraging large language models such as Llama 3.1 enables an entire article to be analysed at once, eliminating the need for sentence-level aggregation and distinguishing between past and future sentiments.

Model	Classification Task	Comparison	Accuracy	F1 Score
Llama (DL & HL)	Past Sentiment	Past Sentiment	0.60	0.62
	Future Sentiment	Future Sentiment	0.56	0.56
Llama	Past Sentiment	Past Sentiment	0.57	0.60
	Future Sentiment	Future Sentiment	0.53	0.55
FinBERT	General Sentiment	Past Sentiment	0.50	0.51
	General Sentiment	Future Sentiment	0.43	0.45
VADER	General Sentiment	Past Sentiment	0.47	0.48
	General Sentiment	Future Sentiment	0.44	0.44
LM-dictionary	General Sentiment	Past Sentiment	0.44	0.46
	General Sentiment	Future Sentiment	0.38	0.40

Table 2: Accuracy and average F1 score benchmarked against a human-labelled dataset. The models output the sentiment classes “appreciation”, “unchanged”, and “depreciation” for each article and for each currency mentioned in an article. “DL” stands for “distant labelling” and “HL” for “human labelling”. The sample includes 200 randomly selected articles from 2011 to 2019. Half of the sample articles were classified by the authors and half by the labeller that classified the training data.

Table 2 reports the accuracy and F1 score (averaged across the three classes) that result from classifying past sentiment as well as future sentiment. The models under comparison include the fine-tuned version of our Llama 3.1 8B model, denoted by Llama (DL & HL), which incorporates both distant labelling and human labelling, alongside the non-fine-tuned Llama 3.1 8B model, FinBERT, VADER, and the LM-dictionary, allowing for a comprehensive comparison of model performance in terms of

FX sentiment analysis. Notably, Llama (DL & HL) shows the strongest performance across most tasks, with an accuracy of 0.60 and an F1 score of 0.62 for past sentiment, while also achieving 0.56 for both accuracy and F1 score on future sentiment. The non-fine-tuned Llama model closely follows in terms of overall performance. Thus, distant plus human-labelled training data prove beneficial for improving overall classification metrics.

As mentioned before, it proves quite difficult to compare these Llama-based models with FinBERT, VADER, and the LM-dictionary because the latter methods are unable to distinguish between past and future sentiment. As a result, we rely on a general sentiment classification task for FinBERT, VADER, and the LLM-dictionary and then compare these general sentiment labels to the separate the past and future labels in our dataset. Compared to the Llama-based models, FinBERT exhibits moderate performance, particularly for past sentiment, and its performance drops more sharply for future sentiment predictions. VADER and the LM-dictionary trail behind all of the transformer-based methods, indicating the limitations of purely lexicon- or dictionary-driven approaches when handling highly contextual financial texts. VADER’s performance is comparable to FinBERT’s performance for past and future sentiment classifications, whereas the LM-dictionary appears least capable of handling the nuances of FX sentiment in the included articles, especially for future-oriented labels.

The discrepancy in performance between past and future sentiment classification can be attributed to several factors. First, past sentiment is directly observable in historical data, making it easier for models to learn clear patterns from explicit textual cues that describe past currency movements. Financial news articles often contain definitive statements about past events, using explicit language such as “the Euro appreciated following the ECB’s announcement,” which provides clear signals for sentiment classification. In contrast, future sentiment relies on implicit reasoning, probabilistic forecasts, and nuanced language, making it more challenging to classify accurately. Additionally, future sentiment often includes speculative elements, conditional statements, and uncertainty markers (e.g., “the dollar may strengthen if the Fed raises rates”), which introduce ambiguity. These linguistic complexities make it harder for models to distinguish between confident predictions and uncertain speculation. While the fine-tuned model improves performance in this regard, it still struggles with the inherent uncertainty in future sentiment prediction.

A more granular class-level breakdown appears in Table 3, focusing on how each model classifies the “appreciation,” “depreciation,” and “unchanged” labels for future sentiment. This breakdown re-

veals that appreciation and depreciation are predicted with higher accuracy and F1 scores than the “unchanged” class across all classification models. The Llama models stand out by better balancing the performance in terms of the appreciating and depreciating classes while still struggling with “unchanged” predictions. Specifically, Llama (DL & HL) demonstrates high accuracy scores of 0.65 for appreciation and 0.56 for depreciation, whereas the unchanged class lags behind at 0.31 in terms of accuracy, with a corresponding F1 score of 0.29. The same difficulty arises for the other models. However, the poorer performance for the unchanged class might also be explained by the fewer observations in that class in the training and evaluation sample (see Table A.2 and Table A.3). Furthermore, (unreported) currency level results show that the models demonstrate consistent performance across most currencies, with the exception of the CHF, for which the accuracy is notably lower across all models. Given the relatively limited number of articles per currency, we refrain from reporting detailed results, as the sample size constraints may limit the representativeness and reliability of these findings when broken down further.

Model	Classification Class	Accuracy	F1 Score
Llama (DL & HL)	Appreciation	0.65	0.62
	Depreciation	0.56	0.60
	Unchanged	0.31	0.29
Llama	Appreciation	0.54	0.59
	Depreciation	0.57	0.61
	Unchanged	0.37	0.26
FinBERT	Appreciation	0.47	0.50
	Depreciation	0.48	0.51
	Unchanged	0.22	0.17
VADER	Appreciation	0.56	0.52
	Depreciation	0.39	0.45
	Unchanged	0.28	0.24
Loughran-McDonald Dictionary	Appreciation	0.36	0.41
	Depreciation	0.43	0.48
	Unchanged	0.27	0.17

Table 3: Accuracy and average F1 score benchmarked individually per classification class against a human-labelled dataset, only for the classification of the “future” labels. The models output the “appreciation”, “unchanged”, and “depreciation” sentiment classes for each article and for each currency mentioned in an article. “DL” stands for “distant labelling” and “HL” for “human labelling”. The sample includes 200 randomly selected articles from 2011 to 2019. Half of the sample articles were classified by the authors and half by the labeller that classified the training data.

Taken together, these findings underscore the importance of using advanced transformer-based archi-

tructures for FX sentiment analysis. Even a pre-trained LLM can directly classify multiple currencies and distinguish between past and future views, highlighting the versatility of large language models for capturing context in FX data. However, fine-tuning remains critical for enabling these advanced models to learn the nuances of currency-related discourse, as shown by the performance of our fine-tuned Llama 3.1 8B model. In contrast, FinBERT’s comparatively weaker results illustrate the need for adapting even specialized financial language models to the FX language. The consistently lower performance on the “unchanged” class suggests that specialized sampling or more sophisticated modelling approaches may be required to better capture neutral signals, which are often only implied rather than explicitly stated in professional financial reporting. Overall, these results emphasize the gains achievable through domain-specific fine-tuning and highlight the particular challenges of predicting future-oriented sentiment in the context of FX markets.

4 Financial application

In the following, we construct portfolios based on the models’ sentiment classification by using articles published on DailyFX, Investing.com and FXStreet from January 2020 until June 2024. To generate daily sentiment scores, the sentiment from each text is aggregated to compute an average sentiment for each day per currency. The definition of a day is aligned with that of the FX returns, namely midnight to midnight New York time. The daily sentiment score for each currency is calculated based on the difference between the count of appreciation and depreciation signals. We focus on appreciation and depreciation signals, as the evaluation in the previous chapter demonstrated that the LLMs perform much worse in the classification of unchanged signals. Following Antweiler and Frank (2004), the daily sentiment score for currency i is computed as:

$$S_{i,t} = \log(1 + \text{CountAppreciation}_{i,t}) - \log(1 + \text{CountDepreciation}_{i,t}) \quad (1)$$

where $\text{CountAppreciation}_{i,t}$ ($\text{CountDepreciation}_{i,t}$) is the number of articles published on day t for which a model assigns the future label of currency i to “appreciation” (“depreciation”). Finally, we round the signals as follows:

$$\hat{S}_{i,t} = \begin{cases} +1, & \text{if } S_{i,t} > 0, \\ 0, & \text{if } S_{i,t} = 0, \\ -1, & \text{if } S_{i,t} < 0. \end{cases}$$

We apply rounding to the sentiment score so that currencies with fewer mentions do not automatically receive reduced exposure due to sparse article counts.

Trading strategies have a long position from day t to $t + 1$ in those currencies with a positive daily sentiment score $S_{i,t}$ and a short position in those with a negative daily sentiment score. To ensure that the constructed portfolios are zero-cost portfolios, we define the weights of the long positions and the short positions as follows:

$$w_{i,t:t+1} = \begin{cases} \frac{\widehat{S}_{i,t}}{\sum_{j:\widehat{S}_{j,t}>0} \widehat{S}_{j,t}}, & \text{if } \widehat{S}_{i,t} > 0 \text{ and } \sum_{j:\widehat{S}_{j,t}>0} \widehat{S}_{j,t} > 0 \\ -\frac{\widehat{S}_{i,t}}{\sum_{j:\widehat{S}_{j,t}<0} |\widehat{S}_{j,t}|}, & \text{if } \widehat{S}_{i,t} < 0 \text{ and } \sum_{j:\widehat{S}_{j,t}<0} |\widehat{S}_{j,t}| > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This definition of the daily sentiment score and portfolio weights ensures that the long (short) position in a currency increases with the number of published positive (negative) articles. If the construction of a zero-cost portfolio is not possible on a given day (e.g. all currencies have a positive daily sentiment score), all positions are set to zero. Note that the weights defining the portfolio held from the close of day t to the close of day $t + 1$ are based on articles published on day t , ensuring that no future information is used. Furthermore, the signals are aligned with the daily FX log returns, ensuring proper temporal synchronization. If no new sentiment signal emerges for a given currency on a subsequent day, the previous day's signal is retained. This approach aims to reduce excessive turnover and improve signal stability, reflecting the possibility that market sentiment does not shift significantly on a daily basis. The portfolios are constructed separately for DailyFX, Investing.com and FXStreet, which allows us to disentangle the informativeness of the three news article providers. All reported returns are presented before the transaction fees, as these costs vary depending on the currency pair and trading conditions. However, we report the annualized portfolio rebalancing frequency to provide insight into potential transaction costs. The performance of the strategy is evaluated by using several key metrics, including the annualized return, annualized volatility, and Sharpe ratio, which measures risk-adjusted returns. These metrics provide a comprehensive understanding of the strategy's effectiveness in terms of managing risk and generating returns.

The cumulative returns are presented in Figure 2. Notably, the fine-tuned Llama model is the only one that consistently maintains a positive return throughout the sample period. Its performance is

particularly strong when using DailyFX signals, achieving a cumulative return exceeding 10% from 2020 to mid-2024 and outperforming all other models. The model also generates positive returns across the entire sample for FXStreet and Investing.com, though dictionary-based methods such as VADER and the LM dictionary perform comparably well for these sources. However, as shown in Figure 1, the number of Investing.com articles declines significantly after 2020, potentially affecting performance. Similarly, the limited availability of FXStreet articles means that the model encountered relatively few of them during fine-tuning, which may explain its weaker performance in this context.

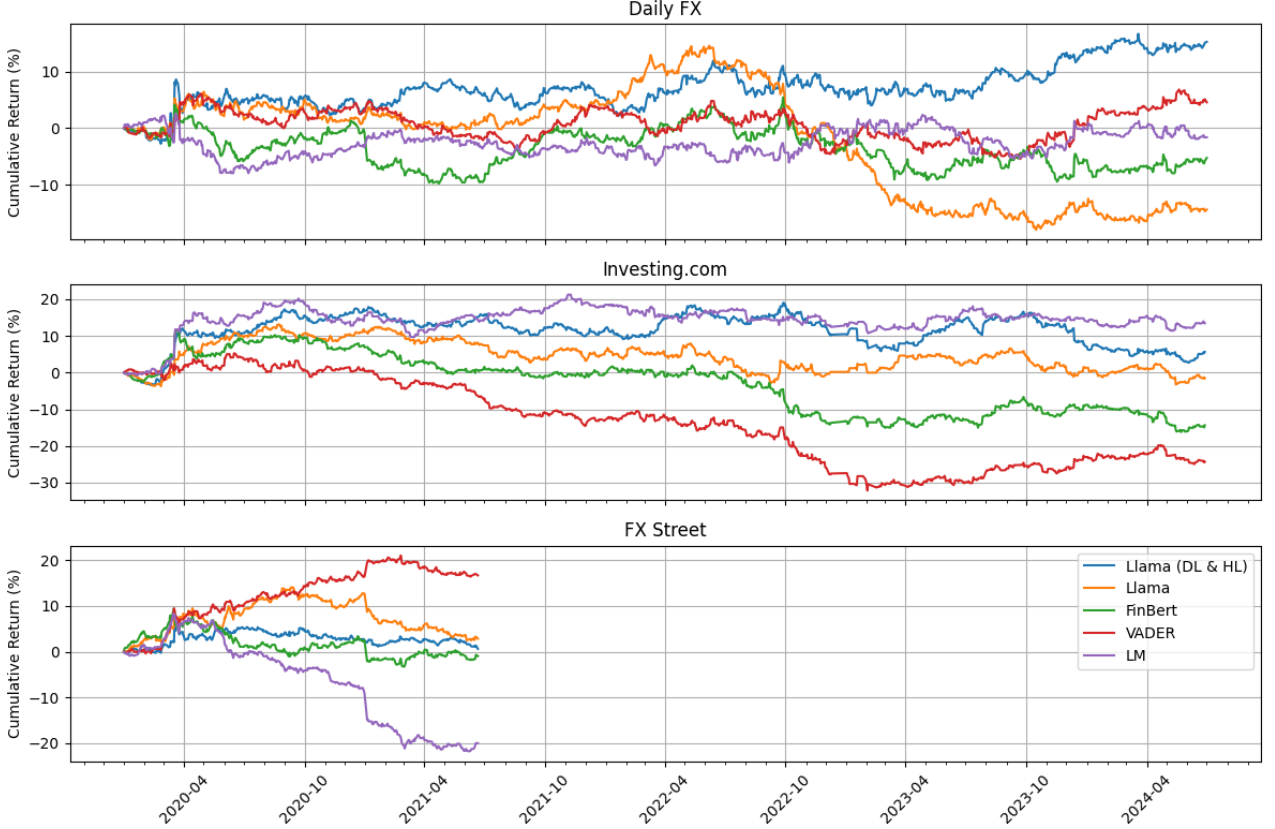


Figure 2: Cumulative returns of a G10 sentiment trading strategy based on sentiments generated by various methods and models split by data source.

The performance metrics presented in Table 4 further highlight the relative effectiveness of different sentiment models. The fine-tuned Llama model achieves the highest annualized return for DailyFX at 3.29%, with a Sharpe ratio of 0.42, outperforming all other models on this dataset. However, the raw Llama model exhibits a negative annualized return of -3.10% and a significantly higher maximum drawdown of 32.58%, suggesting that fine-tuning is crucial for extracting meaningful signals from DailyFX data. For Investing.com, the fine-tuned Llama model also performs well, generating a positive annualized return of 1.27%, though it is outperformed by the Loughran-McDonald Dictionary, which achieves a return of 3.02% with a Sharpe ratio of 0.45. These results suggest that traditional dictionary-based approaches remain competitive, particularly when handling FX news with a lower

article count. Conversely, VADER and FinBERT perform poorly on Investing.com data, both generating negative returns and high drawdowns, with VADER suffering the worst drawdown of 37.43%. For FXStreet, the best-performing model is VADER, which achieves an annualized return of 10.99% and a Sharpe ratio of 1.70, outperforming all other models. Interestingly, the fine-tuned Llama model still delivers a positive return on FXStreet data, but its performance is weaker than that on DailyFX, potentially due to the limited number of FXStreet articles in the training set. The much smaller time frame for FXStreet data than that of the other data sources makes a direct comparison challenging and potentially influences the performance metrics. Moreover, we tested the differences in the Sharpe ratios between the fine-tuned LLaMA (DL & HL) model and the benchmark models across different data sources. While some differences were statistically significant at the 5% and 10% levels, they became insignificant after applying multiple testing corrections. A possible explanation is the relatively modest sample size starting from 2020.

Across all data sources, the rebalancing frequency is consistently high across models. It is calculated as the annualized average number of days with at least one position change (a signal change in one of the currencies), which is obtained by multiplying the mean daily change frequency by 252. All of the models react frequently to sentiment shifts, but VADER, FinBERT, and the LM-dictionary adjust the positions daily (252 times per year), indicating highly reactive signals. In contrast, the fine-tuned Llama model rebalances slightly less often (approximately 232 times per year), suggesting more persistent signals while still maintaining frequent adjustments. Thus, the highest transaction costs are associated with VADER, FinBERT, and the LM-dictionary, given their near-daily trading frequency, whereas the Llama models, while still highly active, incurs slightly lower costs. These results imply that our strategy based on the Llama models would be less eroded by trading frictions, providing a distinct economic advantage by preserving a larger portion of the gross returns in a real-world implementation. The precise magnitude of these frictions is difficult to quantify, as real-world transaction costs in currency markets depend dynamically on several factors, including the liquidity of the specific currency pair, trade size, and overall market volatility, particularly during periods of stress (e.g. Breedon, Chen, Ranaldo, and Vause (2023) and Menkhoff, Sarno, Schmeling, and Schrimpf (2012)).

Meta trained Llama 3.1 8B by using the data up to December 2023, meaning that part of the generated signals fall within its training period. This circumstance raises concerns about potential look-ahead bias since the model might leverage prior knowledge of returns when generating sentiment signals. However, our results suggest that the extent of FX-specific knowledge obtained from Meta’s training

	Model	Annualized Return (%)	Annualized Volatility (%)	Sharpe Ratio	Max. Drawdown (%)	Annualized Rebalancing Frequency
DailyFX	Llama (DL & HL)	3.29	7.74	0.42	7.97	232.18
	Llama	-3.10	6.98	-0.44	32.58	212.83
	FinBert	-1.13	7.48	-0.15	14.91	252.00
	VADER	0.98	6.63	0.15	11.58	252.00
	LM	-0.35	6.97	-0.05	10.50	252.00
Investing.com	Llama (DL & HL)	1.27	7.52	0.17	16.35	232.18
	Llama	-0.34	6.75	-0.05	16.30	212.83
	FinBert	-3.22	7.28	-0.44	27.84	252.00
	VADER	-5.45	6.75	-0.81	37.43	252.00
	LM	3.02	6.75	0.45	10.51	252.00
FXStreet	Llama (DL & HL)	0.41	7.20	0.06	7.17	232.18
	Llama	1.88	7.45	0.25	11.61	212.83
	FinBert	-0.64	7.60	-0.08	12.70	252.00
	VADER	10.99	6.47	1.70	4.66	252.00
	LM	-13.19	7.79	-1.69	30.06	252.00

Table 4: Annualized return, annualized volatility, Sharpe ratio, maximum drawdown, and annualized average number of position changes (annualized rebalancing frequency) for each model’s construction of the G10 sentiment trading strategy.

data is rather limited. The untuned Llama model fails to generate positive returns for both DailyFX and Investing.com, indicating that its baseline performance does not benefit from substantial pre-existing domain expertise. In contrast, the fine-tuned Llama model achieves a positive annualized return for DailyFX and the least negative return for Investing.com, where all models yield negative annualized returns for 2024.

Next, we focus on a currency pair that is particularly sensitive to sentiment shifts: EUR/JPY. The Japanese yen is widely regarded as a safe-haven currency, while the Euro is typically seen as a risk-on currency, often appreciating during periods of economic optimism. This contrast makes EUR/JPY an ideal candidate for analysing the impact of sentiment changes, as shifts in market sentiment can lead to pronounced movements in the exchange rate. Figure 3 illustrates the cumulative returns of the trading strategy based on the signals generated by various models and methods. No position is taken when both the EUR and JPY signals have the same sign, while a long position is entered when the base currency signal exceeds the quote currency signal, provided they don’t have the same sign. Conversely, a short position is taken when the base currency signal is lower than the quote currency signal and both don’t have the same sign. While the fine-tuned Llama model experiences periods of negative cumulative returns within the sample, its performance accelerates especially in later years,

ultimately achieving more than 10% cumulative returns for DailyFX and nearly 20% for Investing.com, significantly outperforming all other models and methods. As observed earlier, VADER again delivers the best results for FXStreet.

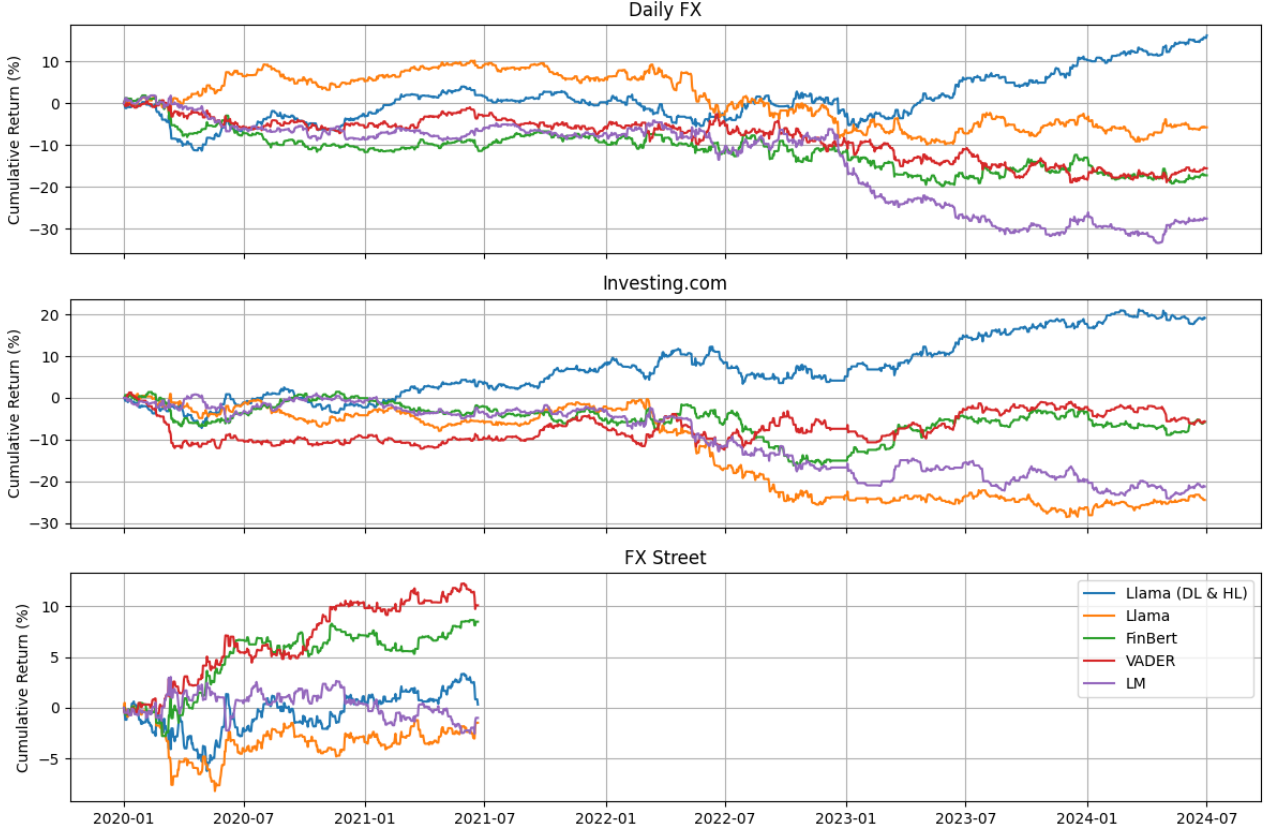


Figure 3: Cumulative returns of EUR/JPY trading strategies based on sentiments generated by various methods and models split by data source.

Table 5 provides a detailed comparison of the statistical performance of each model across different data sources. The Llama model fine-tuned with distant and human-labelled data (DL & HL) achieves the highest annualized returns of 2.49% and 3.59% for DailyFX and Investing.com, respectively, while also maintaining favourable Sharpe ratios of 0.43 and 0.56. Moreover, it exhibits lower maximum drawdowns compared to the baseline Llama model, which struggles with negative annualized returns of -0.89% and -4.58%. Interestingly, for FXStreet, traditional sentiment analysis methods such as VADER and FinBERT outperform Llama-based approaches, achieving Sharpe ratios above 0.9 and maintaining minimal drawdowns of 2.69% and 3.27%, respectively. Similar to that when testing the difference between the Sharpe ratios for G10 trading strategies, some differences in the Sharpe ratios between the fine-tuned LLaMA (DL & HL) model and the benchmark models are statistically significant at the 10% level. However, after applying multiple conservative testing corrections, these differences become insignificant.⁴

⁴The test for Sharpe ratio differences involves the asymptotic properties of the method-of-moments estimators and

Compared to the G10 trading strategy, the rebalancing frequency is generally lower across all models when trading only EUR/JPY, which is intuitive given that signal changes are assessed for just two currencies instead of ten. Among the models, the fine-tuned Llama model exhibits the lowest rebalancing frequency for EUR/JPY, rebalancing less than half as often as most of the others. As a result, the transaction costs for the fine-tuned Llama model are significantly lower than those of the other models.

	Model	Annualized Return (%)	Annualized Volatility (%)	Sharpe Ratio	Max. Drawdown (%)	Annualized Rebalancing Frequency
DailyFX	Llama (DL & HL)	2.49	5.82	0.43	11.62	92.02
	Llama	-0.89	6.34	-0.14	20.03	181.69
	FinBert	-2.65	5.75	-0.46	21.63	197.26
	VADER	-2.40	6.25	-0.38	19.57	193.01
	LM	-4.25	6.31	-0.67	35.22	168.47
Investing.com	Llama (DL & HL)	3.59	6.36	0.56	8.89	92.02
	Llama	-4.58	6.18	-0.74	29.52	181.69
	FinBert	-1.09	5.87	-0.18	17.65	197.26
	VADER	-1.06	6.13	-0.17	13.88	193.01
	LM	-3.98	6.17	-0.64	25.16	168.47
FXStreet	Llama (DL & HL)	0.15	5.86	0.03	6.87	92.02
	Llama	-0.69	4.85	-0.14	8.72	181.69
	FinBert	4.00	4.42	0.90	3.27	197.26
	VADER	4.75	4.64	1.03	2.69	193.01
	LM	-0.47	5.16	-0.09	5.59	168.47

Table 5: Annualized return, annualized volatility, Sharpe ratio, maximum drawdown, and annualized average number of position changes (annualized rebalancing frequency) for each model’s construction of the EUR/JPY sentiment trading strategy.

The results indicate that the fine-tuned Llama model generally outperforms other sentiment analysis approaches, particularly when applied to DailyFX and Investing.com data. This superior performance might stem from two key factors. First, the model may accurately extract the sentiment conveyed by analysts in financial articles, effectively capturing their assessments of currency movements. If analysts tend to be correct in their forecasts, this alone could explain the model’s success in generating profitable trading signals. Second, the model’s training process may have identified latent textual patterns that frequently precede significant market events. Rather than merely reflecting the explicit opinions of analysts, the model might implicitly capture linguistic cues or narrative structures that historically correlate with future currency movements. This pattern recognition capability could contribute

the delta method. We control the family-wise error rate by applying the Holm-Bonferroni correction to the p-values (Holm, 1979).

to the model’s advantage over traditional dictionary-based methods, which rely on predefined word associations rather than contextual learning. However, performance varies across data sources. The model excels with DailyFX, where it maintains a positive cumulative return throughout the sample period, but performs less consistently with FXStreet, where alternative methods such as VADER show stronger results. This discrepancy may be due to differences in the data volume, article structure, or the nature of the sentiment expressed in each source. Furthermore, while the fine-tuned Llama model demonstrates strong results overall, its raw (non-fine-tuned) version struggles, underscoring the importance of domain-specific fine-tuning.

Although concerns about look-ahead bias exist due to the model’s training period overlapping with part of our sample, both the academic literature and our own results suggest this issue is unlikely to be the primary driver of the returns. Crucially, recent studies show that while LLMs can recall pre-cutoff data, they fail to retrieve information released afterward, meaning that evaluations on post-cutoff periods neutralize memorization (Lopez-Lira, Tang, & Zhu, 2025). Complementing this conclusion, the chronologically pure architectures of He, Lv, Manela, and Wu (2025) achieve comparable performances to unconstrained models while explicitly forbidding access to future information, confirming that foreknowledge is not required for alpha. Our own findings also align with this conclusion: the untuned Llama model’s poor performance underscores that predictive power stems from our domain-specific fine-tuning, where we apply strict temporal separation of the data used for fine-tuning and financial application, not pre-existing knowledge. Furthermore, the niche focus of FX articles makes it improbable that these specific articles were heavily weighted in the base model’s general pre-training corpus. While Meta’s undisclosed training data prevent absolute certainty, these findings, taken together, suggest that genuine pattern recognition drives our strategy’s performance.

5 Robustness

To ensure the robustness of our results, we conduct a series of validation steps aimed at evaluating the consistency and reliability of our findings across various configurations and methodologies. A key component of this process involves testing an alternative LLM, Mistral-NeMo-Instruct, which was developed by Mistral and NVIDIA (Mistral AI Team, 2024). With 12 billion parameters, this model is slightly larger than Llama 3.1 8B and has demonstrated strong performance across multiple benchmarks. Leveraging its capabilities, we fine-tune Mistral-NeMo-Instruct by using the same dataset and a methodology that is closely aligned with that applied to Llama 3.1 8B. The evaluation of the labelled

outputs reveals comparable levels of accuracy and F1 scores between the two models, reinforcing the validity of our approach and demonstrating that the observed patterns are robust to the deployment of alternative open-source LLMs.

To explore whether domain-specific pre-training could enhance the performance of Llama 3.1 8B, we conduct an additional experiment involving a preliminary fine-tuning phase by using 30,000 unlabelled examples from the target domain. This phase intends to adapt the base model to the domain-specific context prior to supervised fine-tuning with labelled data. However, our evaluation indicates no significant improvement in the key performance metrics, such as the accuracy and F1 score. One possible explanation for this outcome is that the pre-training dataset, while sizable, might be insufficient in terms of significantly altering the model’s representations, especially given its scale and pre-existing training on extensive datasets. Additionally, the unlabelled nature of the data might limit its utility, as the model lacked explicit task-specific guidance during this phase.

We also undertake a comprehensive optimization of prompt design to enhance the quality of the inputs provided to the models. This process involves experimenting with various prompt structures to identify those that maximizes the accuracy and consistency of outputs. By using a small randomly selected validation subset, we systematically evaluate each prompt’s performance and select the format that demonstrated superior results. Querying the model separately for each currency did not lead to superior results compared to querying all currencies in one prompt. Furthermore, the iterative refinement ensures that the model’s outputs were not overly sensitive to arbitrary choices in prompt formulation.

Finally, we explore two variations of the trading strategy. First, we modify the signal calculation method: instead of rounding the signals, we directly use the logarithmic difference computed in Equation 4 to determine the weights. Second, rather than using log counts, we compute the signals based on the difference in the daily share of “appreciation” and “depreciation” articles:

$$S_{i,t} = \frac{\text{CountAppreciation}_{i,t} - \text{CountDepreciation}_{i,t}}{\text{CountAppreciation}_{i,t} + \text{CountDepreciation}_{i,t} + \text{CountUnchanged}_{i,t}}.$$

With this definition, only the relative proportion of positive and negative articles, rather than their absolute numbers, influences the strategy’s weights. Despite these modifications, the results remain largely unchanged, indicating that the G10 trading strategy is robust to these variations (see Figure A.4

and Figure A.5). We do not present these variations for the EUR/JPY trading strategy, as the results are identical to those obtained with the original approach. This invariance arises because, when trading a single currency pair, there is no need to weigh or round signals, as any non-zero signal directly translates into either a long or short position in the pair. In contrast, for the G10 strategy, weighting adjustments matter because the portfolio must allocate across multiple currencies simultaneously.

6 Conclusion

This study contributes to the growing literature on applying LLMs to financial sentiment analysis, specifically within the FX market. By fine-tuning Meta’s Llama 3.1 with 8 billion parameters on a dataset of labelled FX news articles, we demonstrate that domain-specific adaptation of LLMs improves sentiment classification accuracy and enhances the predictive utility of sentiment signals in trading applications. Our results show that fine-tuned LLMs outperform traditional sentiment analysis methods, including lexicon-based approaches (VADER and Loughran-McDonald Dictionary) and financial-domain models such as FinBERT—both in terms of the classification metrics (accuracy and F1 score) and the performance of the trading strategies derived from sentiment signals.

One of the key contributions of this research is the distinction between past and future sentiment in FX news articles. Unlike prior studies that primarily extract general sentiment from financial text, our approach enables a more granular understanding of how news articles influence market expectations. We show that fine-tuned LLMs capture these nuances effectively, particularly when trained on a combination of human-labelled and distant-labelled datasets. The evaluation results highlight that sentiment predictions for past movements achieve higher accuracy than those for future movements, reflecting the inherent uncertainty of financial forecasting and the complexities of language used in forward-looking statements.

Furthermore, our trading strategy analysis provides strong evidence of the practical utility of LLM-based sentiment classification in financial markets. The fine-tuned Llama model generates profitable trading signals, particularly when applied to news articles published by DailyFX and Investing.com, achieving superior Sharpe ratios and lower drawdowns compared to alternative methods. These findings suggest that domain-adapted LLMs offer an advantage over pre-trained financial models that have not been fine-tuned for the FX market’s unique linguistic and structural challenges.

While our results are robust across multiple datasets and evaluation metrics, several limitations warrant further investigation. First, the scarcity of labelled FX sentiment datasets remains a constraint, despite our efforts to combine human and distant labelling approaches. Second, while our model achieves strong results on sentiment-driven trading strategies, additional research is needed to assess its applicability in more complex trading frameworks, such as higher-frequency trading. Finally, concerns about potential look-ahead bias are mitigated by the absence of performance spikes around the model’s knowledge cutoff, but future studies could further refine backtesting methodologies to ensure robustness.

Overall, this study highlights the potential of fine-tuned LLMs as a powerful tool for sentiment analysis in FX markets. By advancing the methodological rigor of text-based sentiment classification and demonstrating its applicability to trading strategies, we contribute to both the academic literature on financial natural language processing and the broader discussion on the role of AI in financial decision-making. Future research should continue to refine and expand these techniques by leveraging larger datasets and alternative LLM architectures.

References

- Anthropic. (2023). Claude: Large language models. <https://claude.ai/>
- Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of finance*, 59(3), 1259–1294.
- Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. <https://arxiv.org/abs/1908.10063>
- Audrino, F., Gentner, J., & Stalder, S. (2024). Quantifying uncertainty: A new era of measurement through large language models. <https://ssrn.com/abstract=4998319>
- Bhatia, G., Nagoudi, E. M. B., Cavusoglu, H., & Abdul-Mageed, M. (2024). Fintral: A family of gpt-4 level multimodal financial large language models. <https://arxiv.org/abs/2402.10986>
- BIS. (2022). *Triennial Central Bank Survey of foreign exchange and Over-the-counter (OTC) derivatives markets in 2022*. Bank for International Settlements, Monetary and Economic Department. https://www.bis.org/statistics/rpfx22_fx_annex.pdf
- Breedon, F., Chen, L., Ranaldo, A., & Vause, N. (2023). Judgment day: Algorithmic trading around the swiss franc cap removal. *Journal of International Economics*, 140, 103713.
- Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2), 714–750.
- Chen, Y., Kelly, B. T., & Xiu, D. (2022). Expected returns and large language models. <https://ssrn.com/abstract=4416687>
- Cheng, J., & Chin, P. (2025). Empirical asset pricing with large language model agents. <https://arxiv.org/abs/2409.17266>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. <https://arxiv.org/abs/2403.04132>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36, 10088–10115.
- Giglio, S., Kelly, B., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14(1), 337–368.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., . . . Ma, Z. (2024). The Llama 3 Herd of Models. <https://arxiv.org/abs/2407.21783>

- He, S., Lv, L., Manela, A., & Wu, J. (2025). Chronologically consistent large language models. <https://arxiv.org/abs/2502.21206>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. <https://arxiv.org/abs/2106.09685>
- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, 8(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). Predicting returns with text data. <https://www.nber.org/papers/w26186>
- Kelly, B. T., Kuznetsov, B., Malamud, S., & Xu, T. A. (2025). Artificial intelligence asset pricing models. <https://www.nber.org/papers/w33351>
- Konstantinidis, T., Iacovides, G., Xu, M., Constantinides, T. G., & Mandic, D. (2024). FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications. <https://arxiv.org/abs/2403.12285>
- Liu, X.-Y., Wang, G., & Zha, D. (2023). FinGPT: Democratizing internet-scale data for financial large language models. <https://arxiv.org/abs/2307.10485>
- Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? Return predictability and large language models. <https://arxiv.org/abs/2304.07619>
- Lopez-Lira, A., Tang, Y., & Zhu, M. (2025). The memorization problem: Can we trust llms’ economic forecasts? <https://arxiv.org/abs/2504.14765>
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. <https://arxiv.org/abs/1711.05101>
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, 66(1), 35–65.
- Menkhoff, L., Sarno, L., Schmeling, M., & Schrimpf, A. (2012). Currency momentum strategies. *Journal of financial economics*, 106(3), 660–684.
- Meta. (2024, June). Introducing Meta Llama 3: The most capable openly available LLM to date. <https://ai.meta.com/blog/meta-llama-3/>
- Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8, 131662–131682.

- Mistral AI Team. (2024). Mistral NeMo: A State-of-the-Art 12B Model with 128k Context Length. <https://mistral.ai/en/news/mistral-nemo>
- Noguer i Alonso, M. (2019). Look-Ahead Bias in Large Language Models (LLMs): Implications and Applications in Finance. <https://ssrn.com/abstract=5022165>
- OpenAI. (2023). ChatGPT: Large language models. <https://chat.openai.com/chat>
- Schuetzler, J., Audrino, F., & Sigrist, F. (2024). Does sentiment help in asset pricing? A novel approach using large language models and market-based labels. <https://ssrn.com/abstract=4905533>
- Trustbit. (2024, May). Benchmarks for ChatGPT and Co. <https://www.trustbit.tech/en/llm-leaderboard-mai-2024>
- Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. <https://arxiv.org/abs/2303.17564>
- Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. <https://arxiv.org/abs/2306.06031>
- Zhang, B., Yang, H., & Liu, X.-Y. (2023). Instruct-finGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. <https://arxiv.org/abs/2306.12659>

A Additional figures and tables

Parameter	Distant Labelling Fine-Tuning	Human Labelling Fine-Tuning
Number of epochs	3	10
Training batch size	1	1
Gradient accumulation steps	32	16
Evaluation batch size	1	1
Evaluation accumulation steps	32	16
Optimizer	Paged AdamW 32-bit	Paged AdamW 32-bit
Learning rate	1e-5	1e-5
LR scheduler	CosineAnnealing	CosineAnnealing
Warmup ratio	0	0
Weight decay	0.1	0.1
LoRAra r	8	8
LoRA α	16	16
LoRA dropout	0.1	0.1
Max token length	8192	8192
GPU	A100 (80GB)	A100 (80GB)

Table A.1: Training parameters used for both distant labelling and human labelling fine-tuning.

Title: "{row['Title']}"
Text: "{row['Full Text']}"

Instructions:

Objective: For each mentioned currency, answer the following questions:

- What has been the current/past movement of the currency (appreciation, depreciation, or unchanged)?
- What is the future expectation for the currency (appreciation, depreciation, or unchanged)?

You must answer these two questions for each of the following currencies mentioned in the article:

EUR_past: "appreciation, depreciation, or unchanged",
EUR_future: "appreciation, depreciation, or unchanged",
USD_past: "appreciation, depreciation, or unchanged",
USD_future: "appreciation, depreciation, or unchanged"

Output Format:

- Important: Provide your answer in separate rows for each currency as shown above. Do not combine multiple currencies in the same row. Each currency should have its own line with "_past" or "_future" specified.

Example:

- If the article states, "The EUR is expected to appreciate," the output should be:
EUR_past: "unchanged",
EUR_future: "appreciation"
- If the article states, "EUR/USD depreciated last week," the output should be:
EUR_past: "depreciation",
USD_past: "appreciation"
- If only future movements are mentioned for a currency, the past movement should be labelled as "unchanged" and vice versa.

Currency Pair Interpretation:

- If currencies are discussed in pairs, interpret as follows:
 - If "EUR/USD appreciated," label EUR_past as "appreciation" and USD_past as "depreciation".
 - If "EUR/USD depreciated," label EUR_past as "depreciation" and USD_past as "appreciation".

Synonyms:

- Recognize the following synonyms for each currency:
 - **EUR**: EUR, Euro
 - **USD**: USD, Dollar, Dollars, US Dollar, US-Dollar, U.S. Dollar, US Dollars, US-Dollars, U.S. Dollars, Greenback
 - **JPY**: JPY, Yen, Japanese Yen
 - **GBP**: GBP, Pound, Pounds, Sterling, British Pound, British Pounds
 - **AUD**: AUD, Australian Dollar, Australian Dollars, Aussie
 - **CAD**: CAD, Canadian Dollar, Canadian Dollars
 - **CHF**: CHF, Swiss Franc, Swiss Francs, Swissie
 - **NZD**: NZD, New Zealand Dollar, New Zealand Dollars, Kiwi
 - **NOK**: NOK, Norwegian Krone, Norwegian Kroner
 - **SEK**: SEK, Swedish Krona, Swedish Kronor

Figure A.1: Example prompt used for inference and training. In this example prompt, only the Euro and USD are mentioned in the underlying article.

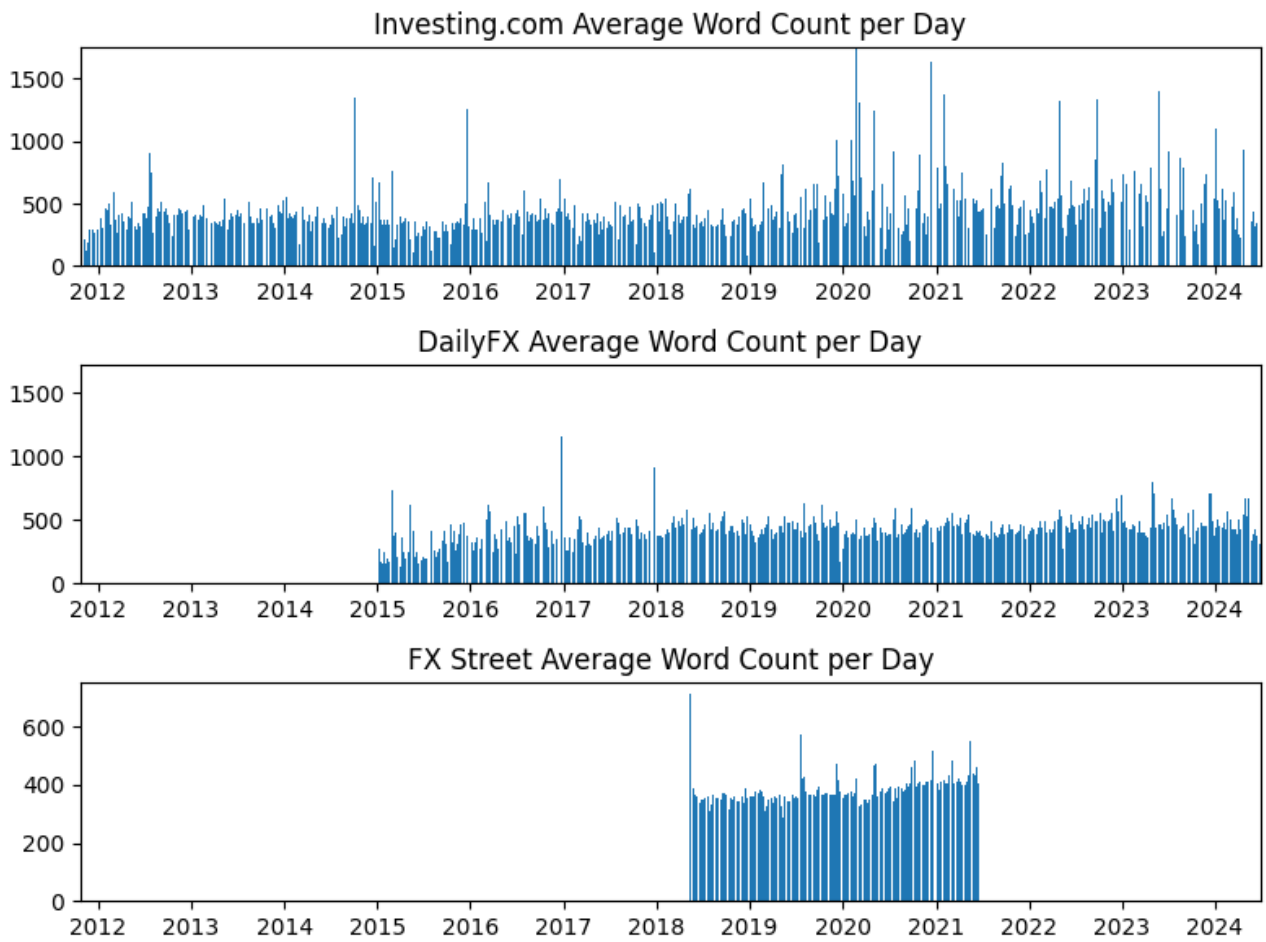


Figure A.2: Average word count per article over time for the Investing.com, DailyFX and FXStreet datasets.

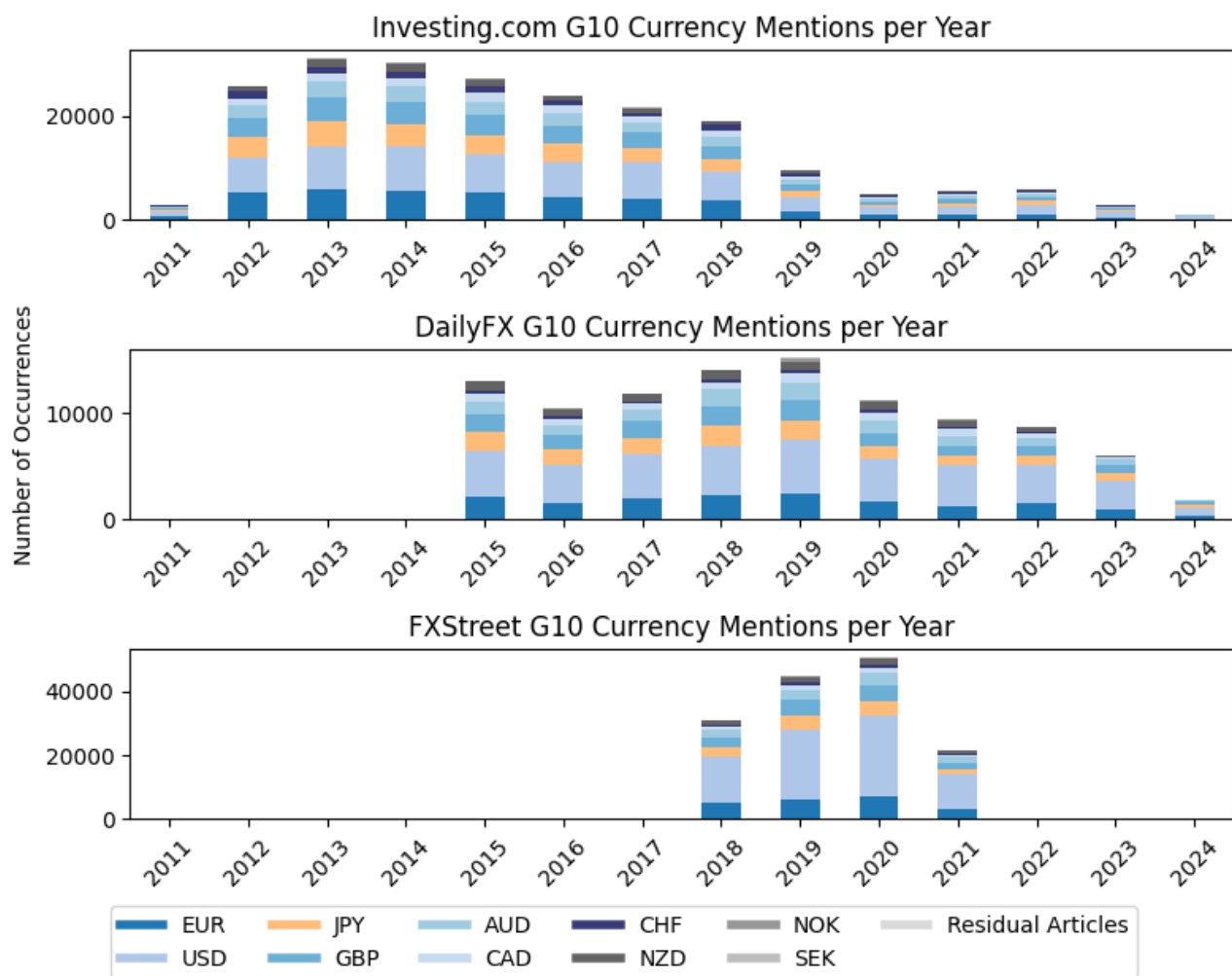


Figure A.3: How often each of the G10 currencies is mentioned in the Investing.com, DailyFX and FXStreet datasets per year.

Time Horizon	Currency	Count Appreciation	Count Unchanged	Count Depreciation
Past	USD	66	19	59
	EUR	30	19	41
	GBP	25	9	40
	JPY	24	7	32
	CAD	30	9	24
	AUD	29	5	31
	NZD	18	5	26
	CHF	17	4	19
	NOK	10	1	11
	SEK	7	1	10
Future	USD	54	30	57
	EUR	32	18	48
	GBP	34	9	32
	JPY	20	14	37
	CAD	26	5	29
	AUD	28	9	30
	NZD	25	4	25
	CHF	17	11	19
	NOK	16	5	9
	SEK	13	2	10

Table A.2: Counts of the appreciation, depreciation and unchanged labels in the human evaluation dataset. The sample includes 200 randomly selected articles from 2011 to 2019. Half of the sample articles are classified by the authors and half by the labeller that classified the training data. One article can refer to several different currencies.

Time Horizon	Currency	Count Appreciation	Count Unchanged	Count Depreciation
Past	USD	238	75	323
	EUR	124	59	244
	GBP	112	51	178
	JPY	123	60	147
	CAD	118	34	162
	AUD	119	33	171
	NZD	107	33	140
	CHF	99	28	105
	NOK	59	18	68
	SEK	44	24	88
Future	USD	265	106	319
	EUR	155	72	235
	GBP	141	62	183
	JPY	133	72	152
	CAD	137	45	148
	AUD	145	45	162
	NZD	137	34	141
	CHF	114	44	94
	NOK	81	13	73
	SEK	69	29	89

Table A.3: Counts of the appreciation, depreciation and unchanged labels in the training dataset. The sample included 869 randomly selected articles from 2011 to 2019. One article can refer to several different currencies.

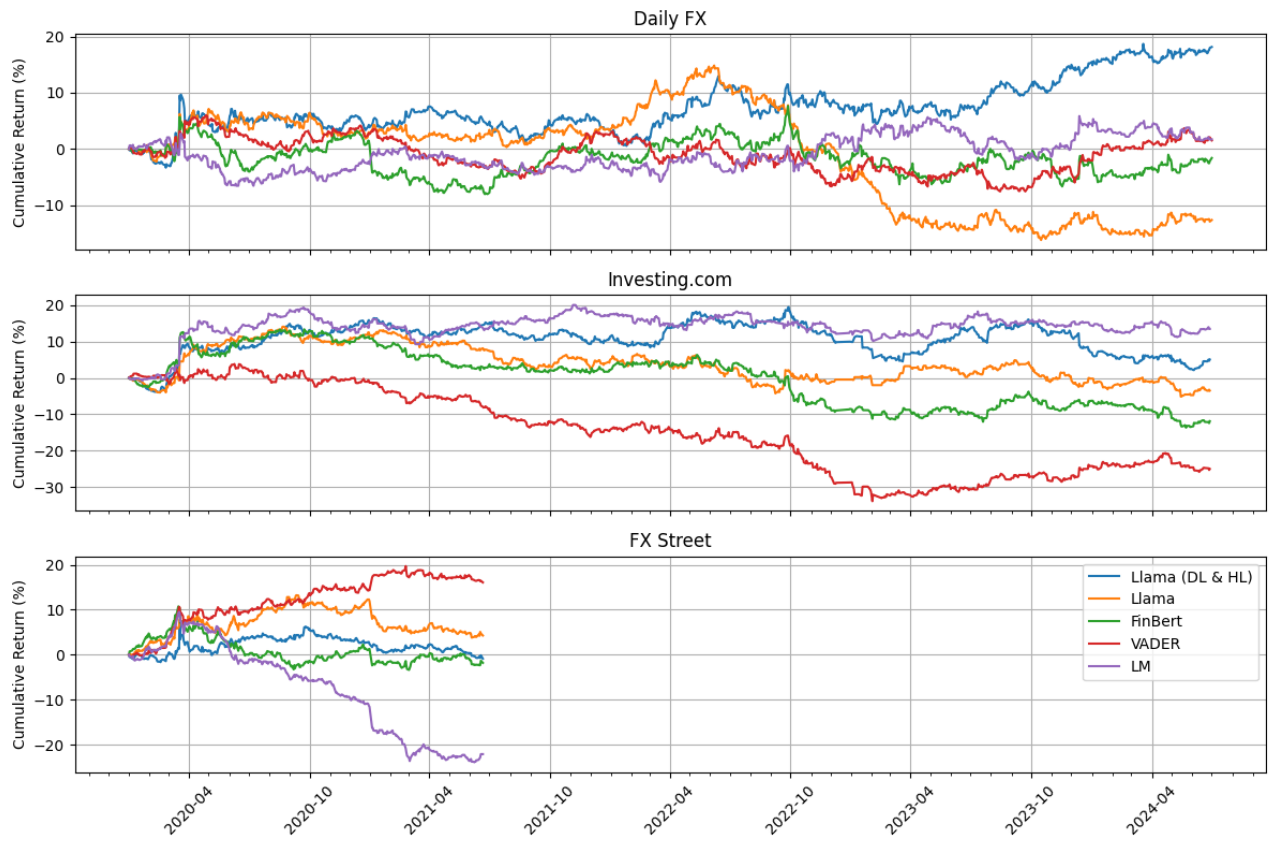


Figure A.4: Cumulative returns of the G10 sentiment trading strategies based on the sentiments generated by various methods and models split by data source. Rather than rounding the signals, we use the signals from Equation 4 to calculate the weights directly.

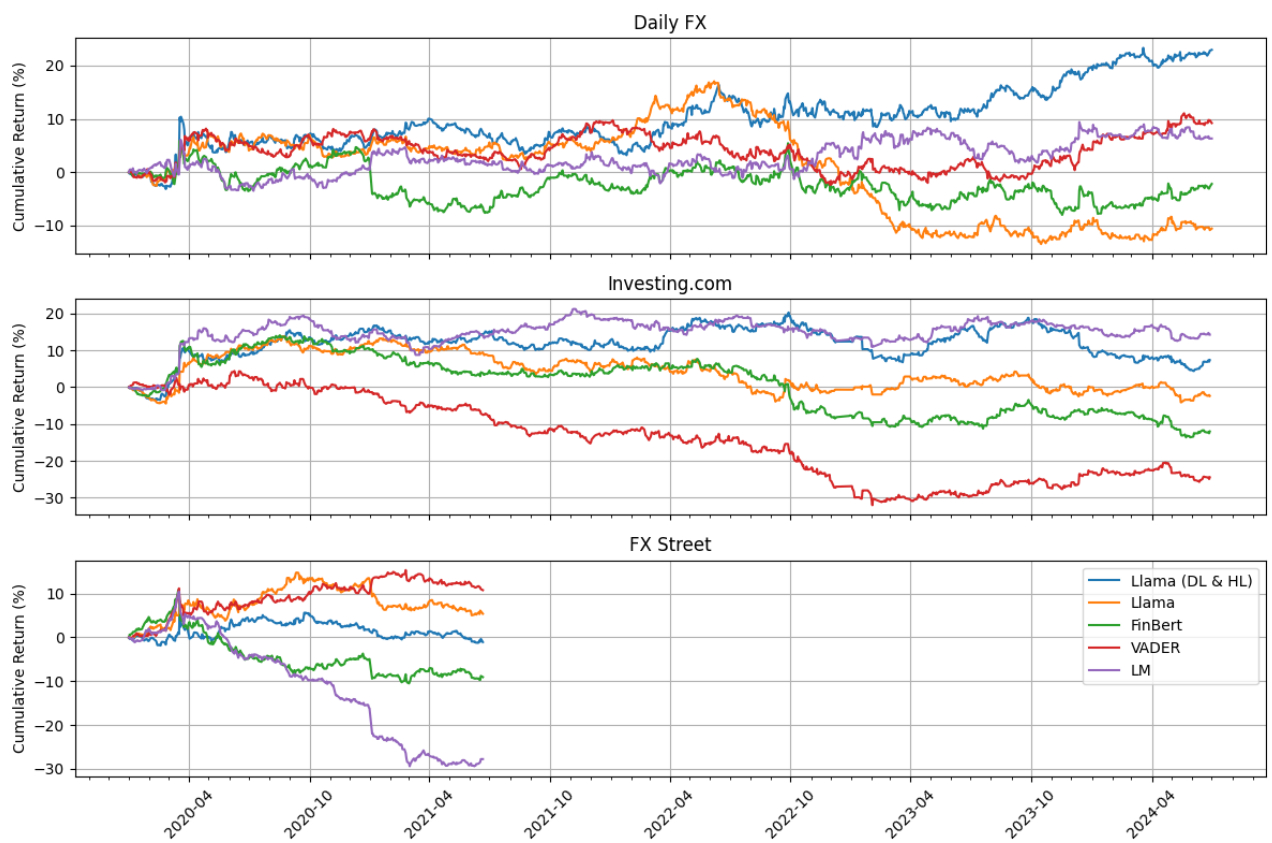
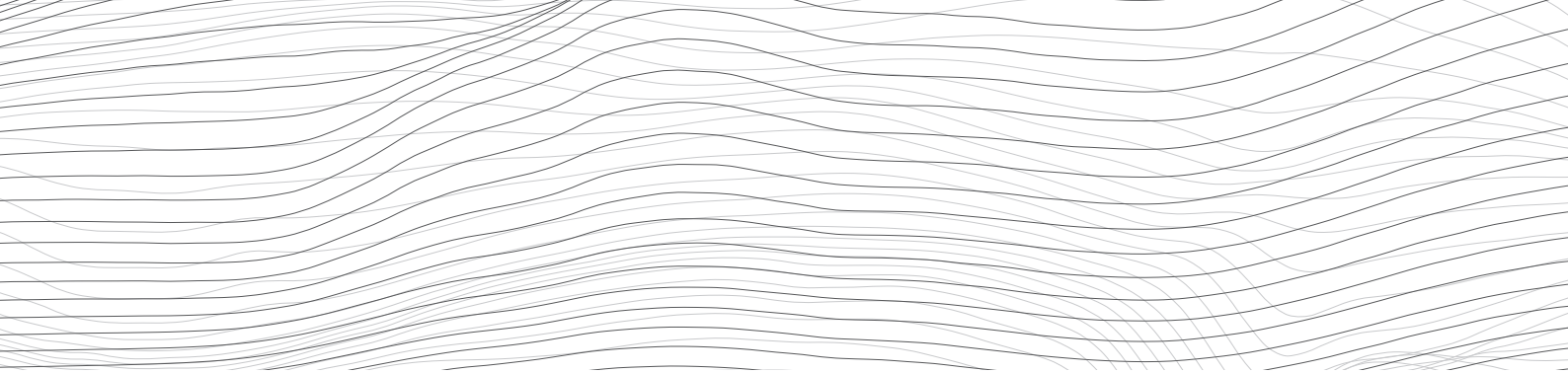


Figure A.5: Cumulative returns of the G10 sentiment trading strategies based on the sentiments generated by various methods and models split by data source. Only the relative rather than the absolute number of positive (negative) signals is considered for the strategy's weights.

Recent SNB Working Papers

2025-11	Daniele Ballinari, Jessica Maly: FX sentiment analysis with large language models	2024-12	Francesco Audrino, Jessica Gentner, Simon Stalder: Quantifying uncertainty: a new era of measurement through large language models
2025-10	Hubert János Kiss, Alfonso Rosa García, Lukas Voellmy: Redemption fees and gates in the lab	2024-11	Marc-Antoine Ramelet, Anna Zeitz: Oil price shocks and household heterogeneity: the income side
2025-09	Laura Felber: Exchange rates and cross-border consumer spending: Evidence from retail payments data	2024-10	Jayson Danton, Terhi Jokipii: A decade of low interest rates: impact on Swiss bank profitability
2025-08	Miriam Koomen, Laurence Wicht: Granularity in the current account	2024-09	Anders Brownworth, Jon Durfee, Michael Junho Lee, Antoine Martin: Regulating decentralized systems: evidence from sanctions on Tornado Cash
2025-07	Elliot Beck, Michael Wolf: Forecasting inflation with the hedged random forest	2024-08	Valentin Grob, Gabriel Züllig: Corporate leverage and the effects of monetary policy on investment: a reconciliation of micro and macro elasticities
2025-06	Jessica Leutert, Rolf Scheufele, Selina Schön: Wage-price pass-through in Switzerland	2024-07	Thomas Nitschka: Evidence on the international financial spillovers of the New York Bankers' Panic of 1907
2025-05	Dirk Bezemer, Richard Senner: Asset pricing and the Covid-19 deposit glut: an application of Liquidity Preference Theory	2024-06	Milen Arro-Cannarsa, Rolf Scheufele: Nowcasting GDP: what are the gains from machine learning algorithms?
2025-04	Lukas Altermatt, Hugo van Buggenum, Lukas Voellmy: Money creation in a neoclassical economy: equilibrium multiplicity and the liquidity trap	2024-05	Jessica Gentner: The role of hedge funds in the Swiss franc foreign exchange market
2025-03	Filippo Cavaleri, Angelo Ranaldo, Enzo Rossi: The demand for safe assets	2024-04	Tobias Cwik, Christoph Winter: FX interventions as a form of unconventional monetary policy
2025-02	Marius Faber, Kemal Kilic, Gleb Kozliakov, Dalia Marin: Global value chains in a world of uncertainty and automation	2024-03	Lukas Voellmy: Decomposing liquidity risk in banking models
2025-01	David Borner: Central bank information and pure monetary policy surprises in Switzerland		
2024-13	Aurel Ruben Mäder, Matthias Jüttner, Daniel Gatica-Perez: You are how you pay: understanding and identifying the payment behavior of sociodemographic groups		



SCHWEIZERISCHE NATIONALBANK
BANQUE NATIONALE SUISSE
BANCA NAZIONALE SVIZZERA
BANCA NAZIUNALA SVIZRA
SWISS NATIONAL BANK

