

# From Experts to Open-Weight Models: A Validated LLM Pipeline for Institutional Economic News Annotation

Michele Petrocelli<sup>1,3</sup>, Andrea Rollin<sup>1</sup>, Matteo Berta<sup>2\*</sup>, Francesca Zafonte<sup>2</sup>,  
Simone Monaco<sup>2</sup>, Daniele Apiletti<sup>2</sup>, Tania Cerquitelli<sup>2</sup>

<sup>1</sup> Ministry of the Economy and Finance, Roma, Italy

<sup>2</sup> Politecnico di Torino,

Department of Control and Computer Engineering (DAUIN), Torino, Italy

<sup>3</sup> Guglielmo Marconi University, Roma, Italy

<sup>1</sup>{name.surname}@mef.gov.it    <sup>2</sup>{name.surname}@polito.it

## Abstract

Economic sentiment plays a central role in shaping perceptions of economic conditions, influencing economic behavior, and policy decisions. Recent advances in large language models have made large-scale sentiment extraction from news increasingly feasible, driving its growing use in economic analysis and forecasting systems across national public-sector institutions. Existing approaches scale mainly through automated or weakly supervised labeling, sacrificing expert alignment and domain nuance, and treat sentiment in isolation, overlooking complementary annotations, such as topic, needed to capture structured economic signals for downstream forecasting. To address these limitations, we introduce a human-annotated dataset of Italian economic news labeled by domain experts for sentiment and topic, and develop a statistically validated pipeline that uses few-shot prompting of commercial large language models to expand coverage while preserving expert alignment at scale. We fine-tune multiple open-weight models on the resulting corpus, yielding language-specific annotation models that maintain high expert agreement across architectures and satisfy the reproducibility and governance requirements of national public-sector institutions. Starting from 800 expert-labeled articles, the pipeline produces approximately 9,200 additional expert-aligned labels and trains local models that pass the Alternative Annotator Test for both sentiment and topic classification. A subset of the annotated dataset is publicly released to support further research<sup>1</sup>.

## 1 Introduction

Textual information from economic news is widely used by fiscal and monetary authorities to construct indicators of expectations and perceptions, complementing traditional macroeconomic variables in

forecasting and policy analysis. The reliability of sentiment and topic indicators depends critically on annotation quality: signals must reflect expert judgment to support meaningful downstream use in operational forecasting pipelines. Producing such indicators at scale for non-English economic news, for which domain-specific linguistic resources are scarce, remains an open practical challenge. The central difficulty is scaling annotation while maintaining alignment with domain experts. This is particularly acute for Italian economic news, where linguistic nuance, implicit expectations, and technical register make annotation inherently subjective and lead to non-negligible inter-annotator disagreement.

Existing approaches rely mainly on automated or weakly supervised labeling, often sacrificing expert alignment and using English-centric resources (Loughran and McDonald, 2011; Shapiro et al., 2022) that transfer poorly to the Italian economic domain. Recent work employs large language models (LLMs) to automate annotation, and typically evaluating performance through classification metrics or agreement measures such as Cohen’s or Fleiss’  $\kappa$  (Artstein and Poesio, 2008; He et al., 2024; Pavlović and Poesio, 2024). However, these metrics are descriptive and do not establish whether a model can reliably substitute human annotators.

To address these limitations, we adopt the Alternative Annotator Test (Calderon et al., 2025), which evaluates whether model–human agreement falls within the range of human variability. Building on this criterion, we propose a pipeline that combines expert annotation, LLM-based label propagation, and distillation into instruction-tuned open-weight models. The use of open-weight models reflects an operational requirement of institutional forecasting infrastructures, which demand reproducibility and independence from external API policies; confidentiality and data-residency considerations further make institutions such as central

\*Corresponding author.

<sup>1</sup><https://github.com/matteoberta/Economic-News-Annotation>

banks and fiscal authorities reluctant to transmit non-public economic data through third-party commercial systems. We introduce a dataset of Italian economic news annotated for both sentiment and topic. Empirically, the validated LLM acts as a substitute annotator, enabling large-scale label propagation with expert-level agreement, and the resulting open-weight models replicate this behavior across architectures; the error analysis further shows that most prediction errors occur in cases of human disagreement, indicating that model uncertainty reflects intrinsic ambiguity rather than systematic misalignment.

## 2 Related Work

We review prior work on economic sentiment analysis, human annotation reliability, and the use of large language models (LLMs) for scalable annotation. Together, these strands highlight the challenge of producing high-quality, scalable annotations for domain-specific economic text.

### 2.1 Economic news sentiment and forecasting

Economic sentiment is widely used to capture expectations and complement traditional macroeconomic indicators, with media tone shown to predict financial markets and relate to firm fundamentals (Tetlock, 2007; Tetlock et al., 2008; Young and Soroka, 2012). Early lexicon-based methods offered transparent, replicable measures (Loughran and McDonald, 2011; Garcia, 2013; Shapiro et al., 2022), while more recent supervised and neural models improve predictive performance at the cost of greater dependence on labeled data (Gentzkow et al., 2019; Chong et al., 2022; Sharpe et al., 2023; Eugster and Uhl, 2024; Réka Ónozó et al., 2024; Petrocelli et al., 2026). These approaches rely critically on annotation quality, yet many datasets are based on heuristic, crowd-sourced, lexicon-derived, or algorithmically assigned labels (Loughran and McDonald, 2011; Shapiro et al., 2022; Eugster and Uhl, 2024), and expert-annotated resources remain scarce, particularly in domain-specific and non-English settings. One of the few exceptions is our prior work Petrocelli et al. (2026), which was conducted on a smaller scale and without formal human–LLM agreement validation.

### 2.2 Human annotation and agreement

Sentiment annotation is inherently subjective, especially when tone is implicit or context-dependent, leading to substantial inter-annotator variability

(Wiebe et al., 2005). Agreement measures, such as Cohen’s  $\kappa$ , are commonly used to quantify reliability (Landis and Koch, 1977), and model performance should be interpreted relative to human agreement rather than in absolute terms (Artstein and Poesio, 2008).

The quality of the annotations also depends on the annotators’ expertise. Although crowd-sourced labels can be effective in simple settings (Snow et al., 2008), domain-specific tasks often require expert knowledge, and disagreement may reflect genuine ambiguity rather than noise (Passonneau and Carpenter, 2014). Despite this, few datasets explicitly benchmark model performance against expert-level agreement.

### 2.3 Large language models for annotation

LLMs are increasingly used to scale annotation, often achieving performance comparable to human annotators under few-shot prompting (He et al., 2024; Pavlović and Poesio, 2024; Gilardi et al., 2023). However, their alignment with human judgments varies, and systematic biases have been observed in subjective tasks (Movva et al., 2024; Shen et al., 2023; Chen et al., 2024), influenced by alignment procedures such as RLHF (Ouyang et al., 2022; Bai et al., 2022). Recent work explores human-in-the-loop pipelines and formal validation frameworks: Calderon et al. (2025) propose a statistical test comparing model–human and human–human agreement to assess when LLMs can act as substitute annotators, while LLMs are also used to generate supervision for smaller models through distillation, enabling scalable deployment (Hsieh et al., 2023; Zhang et al., 2025; Anikina et al., 2025). This work addresses these gaps within a unified framework.

## 3 Methodology

The objective of this work is to construct a high-quality human-annotated dataset of Italian economic news and to validate a scalable pipeline for transferring expert judgments to a large corpus suitable for supervised fine-tuning of open-weight language models. The complete pipeline is shown in Figure 1. The framework follows a two-stage supervised transfer process: from expert-defined reference labels to a validated commercial large language model (LLM) as substitute annotator, and from the aligned LLM to an open-weight model trained on the expanded dataset. In production,

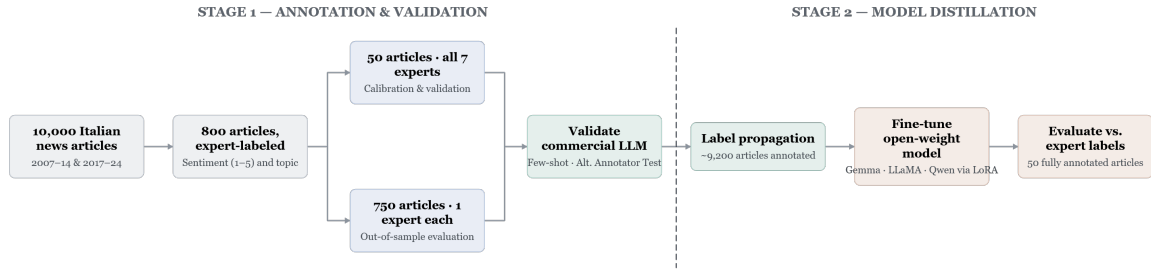


Figure 1: Expert-labeled articles are used to validate a commercial LLM, which then propagates labels to a large unlabeled corpus. The resulting dataset is used to fine-tune open-weight language models, whose performance is evaluated against expert annotations.

the open-weight model can be versioned and periodically re-evaluated against newly annotated expert samples, enabling drift monitoring in media language and macroeconomic regime changes. A key component of this framework is the construction of a fully cross-annotated reference subset, which serves to calibrate annotators, estimate human agreement, and provide a statistical basis for validating the LLM as a substitute annotator.

### 3.1 Data collection and expert annotation

The study relies on a corpus of 10,000 Italian economic news articles collected from major national newspapers through the Factiva database, using a predefined set of economy-related keywords developed in collaboration with domain experts (Appendix B). The dataset spans two periods, 2007–2014 and 2017–2024, equally distributed over years to ensure balanced temporal coverage and capture heterogeneous macroeconomic phases. From the full corpus, 800 articles were selected for manual annotation, preserving both yearly distribution and article length. Seven domain experts independently labeled the articles at the document level. The annotation effort is distributed across two subsets: a fully cross-annotated subset of 50 articles, annotated by all seven experts, used for calibration, human agreement estimation, and statistical validation; and a larger singly annotated subset of 750 articles, divided equally among the seven annotators, in which each articles received a label from a domain expert drawn from the same validated pool. This design maximizes corpus coverage under a fixed annotation budget, while ensuring all reference labels were provided by experts belonging to the same calibrated pool. Sentiment is defined on a five-point ordinal scale (1 = very negative, 5 = very

positive), with annotators instructed to evaluate the tone and framing of the article, not the objective state of economic conditions. The topic labels were defined in collaboration with domain experts and comprise six mutually exclusive categories: *Monetary Policy and Central Banks*, *Fiscal Policies and Taxes*, *Financial Markets*, *Geopolitics and Social Impact*, *Inflation and Prices*, and *Labour Market*. Each article is assigned a single dominant topic that reflects its primary economic focus. Agreement statistics are computed on the fully cross-annotated subset. For sentiment, the average pairwise  $\kappa$  is 0.240, with values ranging from 0.057 to 0.412, and the average pairwise RMSE is 0.832 on 1–5 scale. The moderate value of  $\kappa$  should be interpreted with caution: Cohen’s  $\kappa$  treats sentiment as a set of unordered categories and penalizes any disagreement equally, regardless of whether two annotators differ by 0.5 or by 4 points; with 0.5 increments on the 1–5 scale, it is also chance-corrected against a baseline computed over nine possible values, so annotators who are consistently close to each other but rarely pick the identical value can still produce a low  $\kappa$ . RMSE, by contrast, respects the ordinal structure of the scale: a 0.5-point difference contributes a small error, while a 4-point difference contributes a large one. Section 4.3 returns to this point with the full distribution of pairwise disagreements. For topic classification, the average pairwise  $\kappa$  is 0.492, with values between 0.391 and 0.682. To ensure consistency, detailed written guidelines were provided, and a calibration phase was conducted prior to large-scale labeling.

### 3.2 Target construction

Downstream applications require a single target label for each document. On the 50 fully annotated

|           | Max   | Min   | Avg   | RMSE  |
|-----------|-------|-------|-------|-------|
| Sentiment | 0.412 | 0.057 | 0.240 | 0.832 |
| Topic     | 0.682 | 0.391 | 0.492 | –     |

Table 1: Pairwise inter-annotator agreement on the 50 fully annotated articles. For sentiment, both categorical agreement ( $\kappa$ ) and disagreement magnitude (RMSE on the 1–5 scale) are reported; for topic,  $\kappa$  is shown.

articles, sentiment is aggregated as the arithmetic mean across annotators, rounded to the nearest 0.5. Given the symmetric five-point scale, the mean preserves intensity and yields a continuous supervision signal. This aggregation choice is empirically supported by the inter-annotator disagreement profile: 93% of pairwise comparisons differ by at most one point on the 1–5 scale, indicating a tightly clustered rating distribution in which the arithmetic mean closely approximates both the median and ordinal aggregation estimates (Passonneau and Carpenter, 2014). The topic is aggregated through majority voting. For the 750 articles individually annotated, the available human label serves as the reference target.

### 3.3 Validation of the LLM as substitute annotator

Before extending the annotations, we evaluate whether a commercial LLM can approximate expert judgments and act as a substitute annotator. The model is assessed in a controlled few-shot setting without task-specific optimization.

Few-shot examples are constructed from the fully cross-annotated subset, pairing full articles with their aggregated labels, which the model is prompted to predict (mean sentiment and majority-vote topic). Sentiment is predicted on a 1–5 scale with 0.5 increments to match the continuous supervision signal.

We compare multiple prompting configurations, including standard few-shot and structured variants incorporating expert motivations (He et al., 2024), and select the one that maximizes agreement with human annotations. Evaluation is conducted on the remaining 750 articles using the same metrics adopted for inter-annotator agreement.

To assess whether the LLM can act as a substitute annotator, we apply the Alternative Annotator Test (Calderon et al., 2025), which compares model–human agreement to the empirical range of human variability, and use it as an operational gate: the commercial LLM annotates the remaining corpus only if its agreement with experts falls

within the empirical range of human inter-annotator variability.

The fully cross-annotated subset of 50 articles provides an empirical basis for this procedure for this procedure. The empirical distribution of human variability is estimated from all pairwise annotator comparisons on this subset: with seven annotators, this yields  $\binom{7}{2} = 21$  annotator pairs, producing 1,050 pairwise agreement observations for sentiment. This sample is suitable for characterizing the human agreement distribution, consistent with the sample size guidance provided by Calderon et al. (2025). Furthermore, the Alt-Test results reported in Section 4 are decisive rather than marginal (the model achieves  $\omega \geq 0.71$  at all tested cost-benefit margins  $\varepsilon$  for both tasks), indicating that the conclusion is robust to uncertainty in the estimated agreement distribution.

### 3.4 Large-scale label propagation

Once validated, the LLM is used to annotate the remaining unlabeled portion of the corpus using the selected prompting configuration (joint sentiment–topic few-shot, without motivations; see Section 4.1), extending expert-aligned annotation behavior to a substantially larger set of Italian economic news articles. The final training dataset uses these LLM-labeled articles. Evaluation is conducted exclusively on held-out human-annotated articles, ensuring performance is measured solely against expert judgments and remains directly comparable to human–human agreement levels.

### 3.5 Supervised fine-tuning of the open-weight model

In the final stage, instruction-tuned open-weight language model is fine-tuned on the expanded dataset using supervised learning. Sentiment prediction is formulated as a regression task using the aggregated score as the target, whereas topic prediction is treated as a standard multiclass classification task.

This stage enables the deployment of an independent, domain-specific model for Italian economic sentiment and topic classification. Locally controlled open-weight models provide stability and reliability over time, as its behavior is not affected by external updates. This ensures consistency of predictions over time, facilitates reproducibility, and opens the possibility of maintaining full governance over model versioning and performance.

The overall pipeline therefore consists of two

| Strategy                               | RMSE                                | MAE                                 |
|----------------------------------------|-------------------------------------|-------------------------------------|
| Few-shot (no motivations)              | <b><math>0.308 \pm 0.112</math></b> | <b><math>0.210 \pm 0.129</math></b> |
| Motivations (output)                   | $0.401 \pm 0.071$                   | $0.270 \pm 0.076$                   |
| Motivations (input, two-step pipeline) | $0.398 \pm 0.123$                   | $0.280 \pm 0.104$                   |

Table 2: 5-fold cross-validation on the 50 fully annotated articles (mean  $\pm$  std). Lower values indicate better alignment with expert consensus.

sequential transfers: from human experts to a statistically validated LLM annotator, and from the aligned LLM to a domain-specific open-weight model, with validation at each stage. It was designed with institutional domain experts and is currently undergoing proof-of-concept validation prior to integration into institutional forecasting infrastructure.

## 4 Results

This section evaluates the proposed pipeline, including prompting strategies, statistical validation of the commercial LLM, and distillation into an open-weight model.

### 4.1 Prompting strategy and model alignment

We compare three prompting configurations: (i) standard few-shot, (ii) few-shot with generated motivations, and (iii) a two-step approach with motivations provided as input (He et al., 2024). Prompts are reported in Appendix A.

We evaluate sentiment prediction on the 50 fully annotated articles using 5-fold cross-validation (Table 2). The standard few-shot approach achieved the lowest error, whereas rationale-based strategies do not improve alignment.

**Out-of-sample evaluation.** We evaluated the same configurations on the 750 singly annotated articles and consider a joint sentiment–topic setup (Table 4). Few-shot prompting again yields the lowest sentiment error. The joint configuration preserves sentiment performance while enabling topic prediction and was therefore used for large-scale annotation.

### 4.2 Comparison of commercial LLMs

We compared gpt-5-mini-2025-08-07 and o4-mini-2025-04-16 under the same few-shot configurations described in Section 4.1. The objective was to select the model providing the best trade-off between alignment with expert annotations and annotation cost for large-scale label propagation.

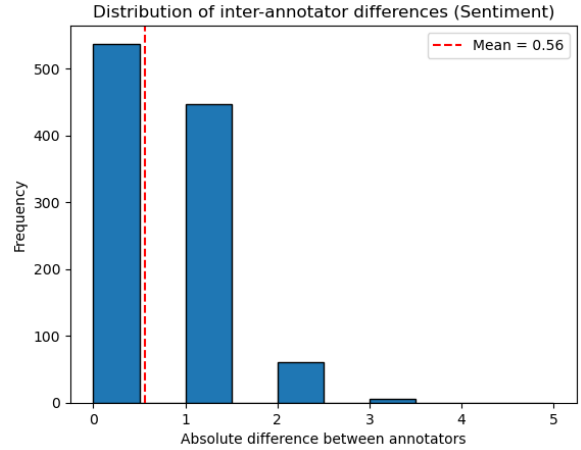


Figure 2: Distribution of absolute inter-annotator differences for sentiment on the 50 fully annotated articles. The dashed line indicates the mean absolute difference.

| Task      | $\varepsilon$ | $\rho$ (AP) | $\omega$ (WR) | Pass |
|-----------|---------------|-------------|---------------|------|
| Sentiment | 0.10          | 0.7800      | 0.7143        | ✓    |
| Sentiment | 0.15          | 0.7800      | 1.0000        | ✓    |
| Sentiment | 0.20          | 0.7800      | 1.0000        | ✓    |
| Topic     | 0.10          | 0.9541      | 1.0000        | ✓    |
| Topic     | 0.15          | 0.9541      | 1.0000        | ✓    |
| Topic     | 0.20          | 0.9541      | 1.0000        | ✓    |

Table 3: Alt-Test results for the selected few-shot configuration. Following Calderon et al. (2025),  $\varepsilon$  is the tolerance for comparing model–human and human–human agreement,  $\rho$  the Advantage Probability (likelihood that the model’s agreement with humans is comparable to inter-annotator agreement), and  $\omega$  the Winning Rate (proportion of cases where model–human agreement matches or exceeds human variability). The model passes when  $\omega \geq 0.5$ .

As shown in Table 5, gpt-5-mini-2025-08-07 substantially outperforms o4-mini-2025-04-16 in sentiment prediction, reducing RMSE by approximately 32% and MAE by approximately 36% on the fully annotated subset.

Preliminary experiments comparing Google and OpenAI models on the same task revealed systematic distributional differences: Google models tended toward more extreme sentiment scores, while OpenAI models clustered closer to the center of the scale. This suggests that the choice of commercial model can leave a modest imprint on the propagated labels, although both remained within the bounds of human variability.

Beyond predictive performance, LLM-based propagation is also substantially less expensive than human annotation: our estimate puts it at roughly 400–1,600 $\times$  cheaper (a 99.8–99.9% reduction) over the full  $\sim$ 9,200-article corpus, an order-

| Strategy               | RMSE                                | MAE                                 | Topic Acc.        |
|------------------------|-------------------------------------|-------------------------------------|-------------------|
| Few-shot               | <b>0.641 <math>\pm</math> 0.077</b> | <b>0.504 <math>\pm</math> 0.063</b> | –                 |
| Few-shot + motivations | 0.653 $\pm$ 0.052                   | 0.532 $\pm$ 0.042                   | –                 |
| Few-shot joint         | 0.656 $\pm$ 0.091                   | 0.506 $\pm$ 0.072                   | 0.643 $\pm$ 0.049 |

Table 4: Out-of-sample evaluation on 750 singly annotated articles (mean  $\pm$  std across annotators). Sentiment error is computed against individual bias-normalized expert labels; topic accuracy: joint configuration only.

Table 5: Comparison between commercial models on sentiment prediction (5-fold cross-validation on 50 fully annotated articles).

| Model                 | RMSE          | MAE           |
|-----------------------|---------------|---------------|
| o4-mini-2025-04-16    | 0.4533        | 0.3300        |
| gpt-5-mini-2025-08-07 | <b>0.3082</b> | <b>0.2100</b> |

of-magnitude gap robust to the underlying cost assumptions (Appendix A.2). Considering both the agreement metrics and the estimated cost differential, gpt-5-mini-2025-08-07 was selected as the commercial annotator for large-scale label propagation.

### 4.3 Validation as alternative annotator

We assess whether the selected LLM can serve as a substitute annotator by comparing model-human agreement to the natural variability observed among experts. Human inter-annotator agreement (Section 3.1) shows average pairwise  $\kappa = 0.240$  for sentiment and  $\kappa = 0.492$  for topic. The average pairwise RMSE for sentiment is 0.832 on the 1–5 scale. A finer-grained view of the magnitude of the disagreement is shown in Figure 2. Across all annotator pairs, 50% of comparisons show perfect agreement, and 93% differ by at most one point. The mean absolute difference is 0.56, while disagreements of three or more points occur in less than 1% of cases. Sentiment agreement on the subtle and implicit economic tone is known to be only moderate, even among trained experts (Wiebe et al., 2005; Artstein and Poesio, 2008). Despite moderate categorical  $\kappa$ , annotators thus exhibit strong quantitative convergence, with disagreements largely confined to adjacent categories, providing a natural tolerance band against which model errors can be interpreted. Since model predictions are evaluated against the aggregated mean of seven annotators, whose variance is lower than individual judgments, model RMSE values should be interpreted relative to this baseline.

We apply the Alternative Annotator Test (Alt-Test) (Calderon et al., 2025). For sentiment, we use the Alt-Test with  $S = -\text{RMSE}$  under a leave-one-

annotator-out setup, as appropriate for continuous annotations. For topic (1–6), we use accuracy as the alignment function. Table 3 shows that the model passes the Alt-Test for both tasks across all tested cost-benefit margins  $\varepsilon$ . We therefore proceed with large-scale label propagation. The substitution is also economically favorable: large-scale propagation through LLM costs roughly two to three orders of magnitude less than equivalent expert annotation (Appendix A.2).

### 4.4 Distillation into the open-weight model

After validating the commercial LLM as an admissible annotator, we transfer its annotation behavior to an open-weight model through parameter-efficient fine-tuning. gemma-2-9b-it is adapted via LoRA using the  $\sim 9,200$  LLM-labeled articles, while evaluation is conducted exclusively on the 50 fully human-annotated documents.

Table 6 shows that LoRA adaptation substantially improves sentiment prediction across all architectures and enables all models to satisfy the Alternative Annotator Test for both sentiment and topic, whereas base checkpoints consistently fall outside the range of human variability. The three backbones do not adapt identically, however: for topic, Gemma reaches  $\omega = 1.00$  while LLaMA and Qwen reach  $\omega = 0.86$ ; for sentiment, RMSE ranges from 0.359 (Gemma) to 0.397 (LLaMA), with Gemma adapting slightly more closely to expert judgment than the other two backbones on this data.

These results show that LoRA-adapted models reproduce the commercial LLM’s expert-aligned behavior without direct access to human annotations during training, yielding independent open-weight models whose predictions lie within the empirical range of human agreement.

The LoRA-adapted models are further evaluated on the 750 singly annotated articles. This subset supports standard error-based evaluation but not the Alt-Test, as they lack the multi-annotator ground truth required for the statistical test. Table 7 reports RMSE, MAE, and topic accuracy for the three backbones. These results are broadly consistent with those on the 50 fully annotated articles (Table 6): Gemma retains the lowest sentiment RMSE, while no single backbone dominates across all metrics. These results provide broader evidence that the annotation behavior transferred from the commercial LLM generalizes beyond the cross-annotated subset.

| (a) Predictive performance |         |               |               |               | (b) Alternative Annotator Test |            |               |             |      |
|----------------------------|---------|---------------|---------------|---------------|--------------------------------|------------|---------------|-------------|------|
| Model                      | Variant | MSE           | RMSE          | MAE           | Task                           | Model      | $\omega$ (WR) | $\rho$ (AP) | Pass |
| Gemma                      | Base    | 0.4274        | 0.6538        | 0.4986        | Sentiment                      | Gemma Base | 0.00          | 0.55        | ×    |
|                            | LoRA    | <b>0.1289</b> | <b>0.3590</b> | <b>0.2814</b> |                                | Gemma LoRA | 1.00          | 0.76        | ✓    |
| LLaMA                      | Base    | 1.1737        | 1.0834        | 0.8850        |                                | LLaMA Base | 0.00          | 0.55        | ×    |
|                            | LoRA    | <b>0.1474</b> | <b>0.3970</b> | <b>0.3171</b> |                                | LLaMA LoRA | 1.00          | 0.80        | ✓    |
| Qwen                       | Base    | 1.1735        | 1.0833        | 0.9453        |                                | Qwen Base  | 0.00          | 0.55        | ×    |
|                            | LoRA    | <b>0.1319</b> | <b>0.3692</b> | <b>0.3301</b> |                                | Qwen LoRA  | 1.00          | 0.79        | ✓    |
|                            |         |               |               |               | Topic                          | Gemma Base | 0.00          | 0.71        | ×    |
|                            |         |               |               |               |                                | Gemma LoRA | <b>1.00</b>   | 0.91        | ✓    |
|                            |         |               |               |               |                                | LLaMA Base | 0.00          | 0.75        | ×    |
|                            |         |               |               |               |                                | LLaMA LoRA | 0.86          | 0.87        | ✓    |
|                            |         |               |               |               |                                | Qwen Base  | 0.00          | 0.75        | ×    |
|                            |         |               |               |               |                                | Qwen LoRA  | 0.86          | 0.87        | ✓    |

Table 6: Base vs. LoRA-adapted models on the 50 fully annotated articles: (a) sentiment regression performance; (b) Alt-Test results at  $\varepsilon = 0.15$  (results are consistent across  $\varepsilon \in \{0.10, 0.15, 0.20\}$ ), metrics as defined in Table 3.

| Model | RMSE          | MAE    | Topic Acc. |
|-------|---------------|--------|------------|
| Gemma | <b>0.6764</b> | 0.5522 | 62.07%     |
| LLaMA | 0.6891        | 0.5074 | 62.98%     |
| Qwen  | 0.7279        | 0.4567 | 60.78%     |

Table 7: Out-of-sample evaluation of the LoRA-adapted models on the 750 singly annotated articles. Errors are computed against individual expert labels.

## 4.5 Error analysis

We analyze prediction errors in relation to human disagreement. Errors increase at the extremes of the sentiment scale, while predictions are more accurate for neutral or moderate cases, leading to a concentration around central values. This reflects both training on aggregated annotations and higher annotator variability for extreme cases.

Prediction errors are strongly associated with annotator disagreement: articles with mixed or ambiguous sentiment exhibit higher error, indicating that model uncertainty aligns with intrinsic ambiguity in the data. Appendix D.1 illustrates two cases of human sentiment disagreement.

LoRA adaptation reduces architecture-specific biases, with fine-tuned models showing consistent error patterns across backbones.

For topic classification, errors primarily occur between semantically related categories, mirroring human annotation patterns (see Appendix D.2). In general, errors largely reflect human disagreement rather than systematic model bias.

## 5 Resource availability

We release the annotation guidelines, the prompting templates used for the commercial LLMs (Appendix A.1), and the training and LoRA configurations (Appendices A.3 and A.4). A subset of the

annotated dataset, including sentiment and topic labels, persistent article identifiers, such as publication source, publication date, and Factiva ID, is available in this [repository](#). The full article texts cannot be released because the corpus was collected through the Factiva database, whose licensing terms prohibit the redistribution of full-text content.

## 6 Conclusion and Future Work

We introduced a scalable annotation pipeline for Italian economic news, combining expert labeling, Alt-Test validation of a commercial LLM as a substitute annotator, and distillation into an open-weight instruction-tuned model aligned with expert judgments. The results show that standard few-shot prompting achieves the strongest alignment with expert annotations, that the selected LLM operates within the range of human disagreement, and that parameter-efficient fine-tuning transfers this behavior to open-weight models across architectures, providing a practical solution for domain-specific settings with limited labeled data.

**Future work** will explore alternative target construction methods that explicitly model annotator disagreement and will evaluate additional model backbones. It will also assess the macroeconomic relevance of the resulting sentiment indicators in downstream tasks, such as inflation forecasting, for which preliminary results are presented in Appendix E. This assessment will include a rigorous econometric comparison with existing sentiment measures such as Loughran–McDonald-style lexicons (Loughran and McDonald, 2011) under a common forecasting setup.

## 7 Limitations

This work presents a scalable, statistically grounded pipeline for annotating Italian economic news, with the following scope constraints.

The corpus covers only two periods (2007–2014, 2017–2024); extending it to additional years and periodicals would strengthen temporal representativeness.

Topic annotation assigns a single dominant category per article, so articles in which multiple themes are equally prominent are not captured; multi-label schemes are left for future work.

The fully cross-annotated subset (50 articles, 7 annotators; 21 pairs, 1,050 pairwise observations) is small in absolute terms, though consistent with the sample-size guidance in [Calderon et al. \(2025\)](#). Appendix C shows Fleiss’  $\kappa$  is stable across group sizes 2–7, and Table 3 shows the Alt-Test verdict holds at every tested margin ( $\omega \geq 0.71$  for both tasks). This reflects a deliberate tradeoff under a fixed expert-hour budget, prioritizing coverage (750 singly annotated articles) over a larger cross-annotated subset, which we leave for future work.

Validation focuses on a single primary commercial annotator (gpt-5-mini-2025-08-07); Section 4.2 compares it against one alternative only. Broader testing across commercial and open-weight models is left for future work and would require re-running the Alt-Test gate (Section 4.3), not redesigning the pipeline.

The limited amount of expert-labeled data available for fine-tuning (800 articles) and long source articles (2,000–8,000 words, exceeding standard encoder context windows) motivated instruction-tuned LLMs over smaller encoder-only classifiers; whether fine-tuned encoders can match LoRA-adapted LLMs in lower-resource or shorter-document settings remains open.

The pipeline is designed to be language-agnostic beyond the Italian keyword list (Appendix B) and few-shot prompts (Appendix A), requiring only domain-expert annotators, a cross-annotated calibration subset, and the Alt-Test gate (Section 4.3). However, such cross-lingual transfer has not yet been empirically validated and would require native-speaker domain experts for the target setting.

Finally, the dataset is drawn only from major Italian national newspapers; broadening it to other outlets, wire services, or social media would increase stylistic diversity and support more general-

purpose models.

## References

- Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. [A rigorous evaluation of LLM data generation strategies for low-resource languages](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8282–8303, Suzhou, China. Association for Computational Linguistics.
- Ron Artstein and Massimo Poesio. 2008. [Survey article: Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. [Constitutional AI: Harmlessness from AI feedback](#). *Preprint*, arXiv:2212.08073.
- Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of ACL 2025*, pages 16051–16081.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of EMNLP 2024*, pages 8301–8327.
- Eilyn Chong, Chiung Ching Ho, Zhong Fei Ong, and Hong H Ong. 2022. [Using news sentiment for economic forecasting: a Malaysian case study](#). In Bank for International Settlements, editor, *Machine learning in central banking*, volume 57 of *IFC Bulletins chapters*. Bank for International Settlements.
- Patrick Eugster and Matthias W. Uhl. 2024. [Forecasting inflation using sentiment](#). *Economics Letters*, 236:111575.
- Diego Garcia. 2013. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.
- Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. [Text as data](#). *Journal of Economic Literature*, 57(3):535–74.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu-Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of NAACL-HLT 2024 (Industry Track)*, pages 165–190.

- Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Dis-tilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.
- Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of EMNLP 2024*, pages 9048–9062.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). Preprint, arXiv:2203.02155.
- Rebecca J. Passonneau and Bob Carpenter. 2014. [The benefits of a model of annotation](#). *Transactions of the Association for Computational Linguistics*, 2:311–326.
- Maja Pavlović and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of NLPerspectives 2024 @ LREC-COLING*, pages 100–110.
- Michele Petrocelli, Andrea Rollin, Matteo Berta, Francesca Zafonte, Simone Monaco, Salvatore Lo Sardo, Daniele Apiletti, Daria Scacciatelli, and Tania Cerquitelli. 2026. [From human to scalable annotation: Teaching LLMs to mimic experts labeling on economic news sentiment](#). In *Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference*, volume 4192 of *CEUR Workshop Proceedings*, Tampere, Finland. CEUR-WS.org.
- Lívia Réka Ónozó, Frigyes Viktor Arthur, and Bálint Gyires-Tóth. 2024. [Leveraging llms for financial news analysis and macroeconomic indicator nowcasting](#). *IEEE Access*, 12:160529–160547.
- Adam Hale Shapiro, Moritz Sudhof, and Daniel J. Wilson. 2022. [Measuring news sentiment](#). *Journal of Econometrics*, 228(2):221–243.
- Steven A. Sharpe, Nitish R. Sinha, and Christopher A. Hollrah. 2023. [The power of narrative sentiment in economic forecasts](#). *International Journal of Forecasting*, 39(3):1097–1121.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of EMNLP 2023*, pages 4215–4233.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.
- Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms’ fundamentals. *The Journal of Finance*, 63(3):1437–1467.
- Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. [Annotating expressions of opinions and emotions in language](#). *Language Resources and Evaluation*, 39(2/3):165–210.
- Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.
- Hengran Zhang, Minghao Tang, Keping Bi, Jiafeng Guo, Shihao Liu, Daiting Shi, Dawei Yin, and Xueqi Cheng. 2025. [Utility-focused LLM annotation for retrieval and retrieval-augmented generation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1683–1702, Suzhou, China. Association for Computational Linguistics.

## A Experimental setup

This appendix details the experimental configurations used for both the commercial LLM annotation stage and the open-weight fine-tuning stage. We report prompting templates, data splitting strategy, training hyperparameters, hardware configuration, and evaluation protocols to ensure full reproducibility of the proposed two-stage transfer pipeline. All implementation choices are documented explicitly to allow independent replication and comparison.

### A.1 Prompting configuration for commercial LLMs

All commercial LLM experiments were conducted using structured system–user prompt templates. We evaluated four prompting configurations: (i) sentiment-only few-shot prompting, (ii) few-shot prompting with motivations provided as input, (iii) few-shot prompting with motivations generated in output, and (iv) joint sentiment–topic prediction without motivations.

The prompts were originally written in Italian. For transparency, we provide English translations generated using GPT-5.2.

#### A.1.1 Few-shot sentiment classification (without motivations)

The baseline prompting strategy uses few-shot examples containing only article text and sentiment labels:

##### System Prompt (Sentiment Only)

You are an expert annotator of Italian economic news articles.

Read each article carefully and assign a Sentiment score based only on linguistic tone (not factual content).

Scale: 1 (very negative) to 5 (very positive), with 0.5 increments allowed.

Use integer values when tone is clearly in a category. Use intermediate values when tone shows mixed adjacent characteristics.

The response must be exclusively a valid JSON object:

```
{"Sentiment": number between 1 and 5 (step 0.5)}
```

#### A.1.2 Motivation generation (two-step annotation pipeline)

For the two-step pipeline configuration proposed by [He et al. \(2024\)](#), motivations were first generated for few-shot examples using the following system prompt:

##### System Prompt (Motivation Generation)

You are an expert annotator of Italian economic news articles.

You will be given articles already labeled with a sentiment score (1–5 with 0.5 increments).

**Reference Scale:** 1 = very negative

2 = negative

3 = neutral

4 = positive

5 = very positive

Intermediate values (1.5, 2.5, 3.5, 4.5) represent shades between adjacent categories.

Your task is to generate a concise explanation (maximum 100 words) justifying the assigned sentiment score {sentiment}, simulating the reasoning of a human annotator.

Requirements: - Identify specific textual elements (words, tone, framing) - If the value is an integer, explain why the tone clearly belongs to that category - If the value is intermediate, explain the elements that create the nuance - Do not propose alternative labels - Output must be valid JSON:

```
{"Motivation": "..."} 
```

### A.1.3 Few-shot sentiment with motivation in the output

In this configuration, the model was instructed to output both the sentiment and a short explanation:

#### System Prompt (Sentiment + Motivation Output)

Classify the article sentiment as described above.

Output format:

```
{ "Sentiment": number between 1 and 5 (step 0.5), "Motivation": short explanation (max 100 words) }
```

### A.1.4 Joint sentiment and topic classification

For multi-task prompting, sentiment and topic labels were predicted jointly in a single inference call:

#### System Prompt (Sentiment + Topic)

Classify the article along two independent dimensions:

1. Topic (based on content): 1 = Labor Market 2 = Fiscal Policy and Taxes 3 = Financial Markets 4 = Geopolitics and Social Impact 5 = Monetary Policy and Central Banks 6 = Inflation and Prices
2. Sentiment (based only on tone): 1–5 with 0.5 increments allowed.

If multiple topics are present, select the dominant one.

Output format:

```
{ "Sentiment": number between 1 and 5 (step 0.5), "Topic": integer between 1 and 6 }
```

In all configurations:

- Only valid JSON responses were accepted
- Few-shot examples were fixed across runs to ensure comparability

## A.2 Annotation cost estimate

We estimated the relative annotation cost of LLM-based propagation versus human expert annotation. Human cost assumes an average article length of  $\sim 5,000$  words (the midpoint of the observed 2,000–8,000 word range) and a reading-plus-annotation rate of  $\sim 200$  words/minute ( $\sim 25$  minutes/article), at an hourly rate of 10–20 €/h for the seven domain experts. LLM cost assumes an Italian word-to-token ratio of  $\sim 1.5$  and current gpt-5-mini-2025-08-07 pricing ( $\$0.25/\text{M}$  input,  $\$2.00/\text{M}$  output tokens); since the number of few-shot examples in the prompt was not logged, we vary  $k$  between 2 and 5. Table 8 reports the resulting range for both methods over the full  $\sim 9,200$ -article corpus. Across this range, LLM-based

Table 8: Estimated cost of annotating the  $\sim 9,200$ -article corpus. Human cost varies with assumed annotator rate (10/15/20 €/h); LLM cost varies with assumed few-shot count ( $k=2/3/5$ ), since exact prompt size was not logged.

| Method        | Low  | Central | High |
|---------------|------|---------|------|
| Human experts | €30k | €45k    | €60k |
| gpt-5-mini    | €96  | €128    | €192 |

propagation is approximately  $400\text{--}1,600\times$  cheaper than human annotation, a reduction of roughly 99.8–99.9%. This range reflects each method’s own uncertain parameter rather than a joint worst/best case across both, so it is wider than dividing the table’s Low and High columns directly. We report this as an order-of-magnitude estimate rather than a measured cost: individual assumptions are approximate, but the gap between methods is large enough that the qualitative conclusion is robust to reasonable variation in any one of them.

### A.3 Instruction-tuned models and fine-tuning configuration

**Data splitting** The corpus consists of 10,000 Italian economic news articles. Of these, 800 were annotated by human experts and reserved exclusively for evaluation. The remaining approximately 9,200 articles were annotated using the validated commercial LLM and used for training the open-weight models.

The LLM-labeled subset was split into 85% training and 15% validation. Human-annotated articles were never used during training.

### A.4 Hyperparameters

Fine-tuning was performed using Low-Rank Adaptation (LoRA).

Table 9: LoRA configuration parameters.

| Parameter                        | Value                           |
|----------------------------------|---------------------------------|
| LoRA rank ( $r$ )                | 32                              |
| LoRA scaling factor ( $\alpha$ ) | 64                              |
| LoRA dropout                     | 0.05                            |
| Bias                             | none                            |
| Task type                        | Causal LM                       |
| Target modules                   | Attention and projection layers |

Table 10: Supervised fine-tuning configuration.

| Parameter                   | Value              |
|-----------------------------|--------------------|
| Epochs                      | 3                  |
| Per-device batch size       | 1                  |
| Gradient accumulation steps | 16                 |
| Effective batch size        | 16                 |
| Learning rate               | $1 \times 10^{-4}$ |
| Warmup ratio                | 0.03               |
| Max gradient norm           | 1.0                |
| Maximum sequence length     | 16,384             |
| Packing                     | Disabled           |
| Gradient checkpointing      | Enabled            |

**Training infrastructure** Training was conducted on  $2 \times$  NVIDIA RTX A6000 GPUs using the HuggingFace Accelerate framework.

**Training objective** Training followed a completion-only causal language modeling objective.

**Evaluation metrics** Sentiment performance was evaluated using RMSE and MAE. Topic performance was evaluated using accuracy.

## B Keyword list for data collection

The corpus was constructed using a predefined set of economy-related keywords in Italian. The selection was developed in collaboration with domain experts to ensure broad coverage of macroeconomic conditions, monetary policy, financial markets, taxation, and distributional issues, while avoiding overly narrow or sentiment-driven queries.

For transparency, an English translation of the keyword list, generated using GPT-5.2, is reported below:

Inflation, Prices, Rising cost of living, Monetary policy, Interest rates, Deflation, Consumer Price Index, Unemployment, Phillips curve, Expectations, Stagflation, Core inflation, Anti-inflationary, Wage indexation, Bank of Italy, ECB (European Central Bank), Price increases, Eurozone, Euro area, Inflation in Italy, Purchasing power, Recession, Savings, Tax wedge, Mortgages, Credit, Loans, Financial markets, FTSE MIB, GDP (Gross Domestic Product), Economy, Stock market, Commodities, Shock, Crisis, Inequality, Tax burden, Subsidies, Taxation, Trade restrictions, Tariffs, Excise duties, VAT, Supply chain crisis, US-China trade

tensions, Currency devaluation and instability, Central banks, Fed (Federal Reserve), FOMC, “Consumer squeeze”, “Domino effect of prices”, “Risk of stagnation”, “Income of Italians”, “Inflationary spiral”, “Wage stagnation”, “Making ends meet”, “Alarming peaks”, “Shopping basket”, “Price pressure”.

The English version is provided solely for interpretative clarity, as the retrieval procedure was conducted exclusively using the original Italian terms.

### C Detailed agreements between human annotators

This section reports the full distribution of inter-annotator agreement across all possible combinations of annotators. For both sentiment and topic annotation, Fleiss’  $\kappa$  was computed for every subset of annotators with group size ranging from 2 to 7. Reporting agreement by group size provides a detailed view of variability across annotator subsets and illustrates the stability of consensus as the number of annotators increases.

Table 11: Summary of Fleiss’  $\kappa$  by annotator group size (2–7)

| <b>Sentiment</b> |       |       |       |       |
|------------------|-------|-------|-------|-------|
| Group size       | Max   | Min   | Avg   | #comb |
| 2                | 0.412 | 0.057 | 0.240 | 21    |
| 3                | 0.353 | 0.145 | 0.241 | 35    |
| 4                | 0.311 | 0.178 | 0.242 | 35    |
| 5                | 0.289 | 0.213 | 0.243 | 21    |
| 6                | 0.260 | 0.226 | 0.243 | 7     |
| 7                |       |       | 0.243 | 1     |
| <b>Topic</b>     |       |       |       |       |
| Group size       | Max   | Min   | Avg   | #comb |
| 2                | 0.682 | 0.391 | 0.491 | 21    |
| 3                | 0.654 | 0.402 | 0.493 | 35    |
| 4                | 0.588 | 0.439 | 0.491 | 35    |
| 5                | 0.540 | 0.463 | 0.495 | 21    |
| 6                | 0.494 | 0.480 | 0.487 | 7     |
| 7                |       |       | 0.496 | 1     |

## D Error analysis

### D.1 Qualitative examples of annotator disagreement

We report two representative cases of expert disagreement, showing how annotator variability correlates with textual ambiguity.

**Example 1:** a newspaper article summarizing the Bank of Italy’s quarterly bulletin combines warning signs, tightening credit conditions, falling competitiveness, and pessimistic business expectations with reassuring figures on public finances and household debt. Annotators scored the article between 2 and 5, pointing to different weightings of the negative economic signals against the reassuring fiscal and household debt figures reported in the same piece.

**Example 2:** on regional governments’ liquidity measures for businesses during the coronavirus lockdown, coverage sits between two registers: a backdrop of closures and economic distress, and a series of funding announcements framed as decisive institutional action. Scores ranged from 1 to 4, suggesting that annotators differed in whether they read the article primarily through the crisis it describes or through the response it reports.

### D.2 Topic confusion matrices

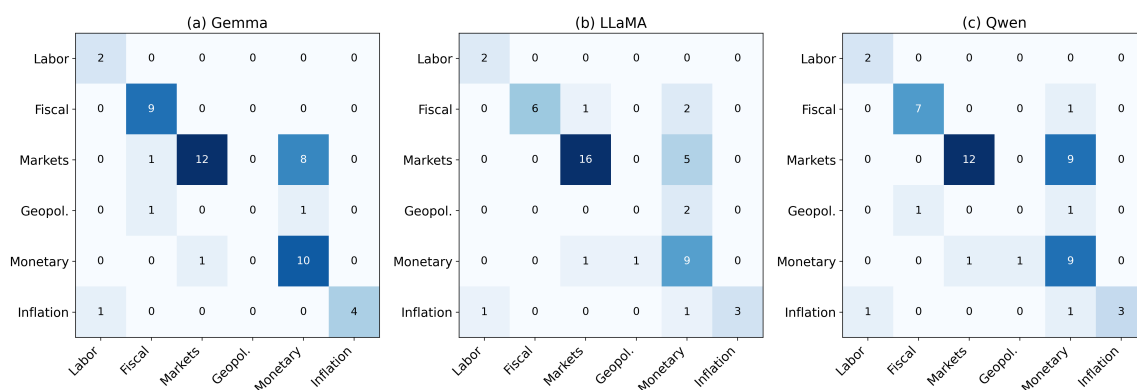


**Figure 3.** Topic confusion matrices for LoRA-adapted models (Gemma, LLaMA, Qwen). Rows correspond to ground-truth labels and columns to predictions. Errors are concentrated between semantically related categories. Labels: Labor = Labor Market; Fiscal = Fiscal Policy and Taxes; Markets = Financial Markets; Geopol. = Geopolitics and Social Impact; Monetary = Monetary Policy and Central Banks; Inflation = Inflation and Prices.

## E Downstream economic application

To illustrate the potential macroeconomic relevance of the validated sentiment signal, we report a descriptive comparison between monthly aggregated sentiment and Italian CPI inflation (year-over-year). The figures below are purely illustrative and do not constitute a formal forecasting exercise.

Figure 4 provides a descriptive comparison between the constructed sentiment indicator and Italian CPI inflation. For visualization purposes, sentiment is linearly transformed from the original 1–5 scale to a symmetric interval  $[-2, 2]$ , facilitating comparison with inflation dynamics.

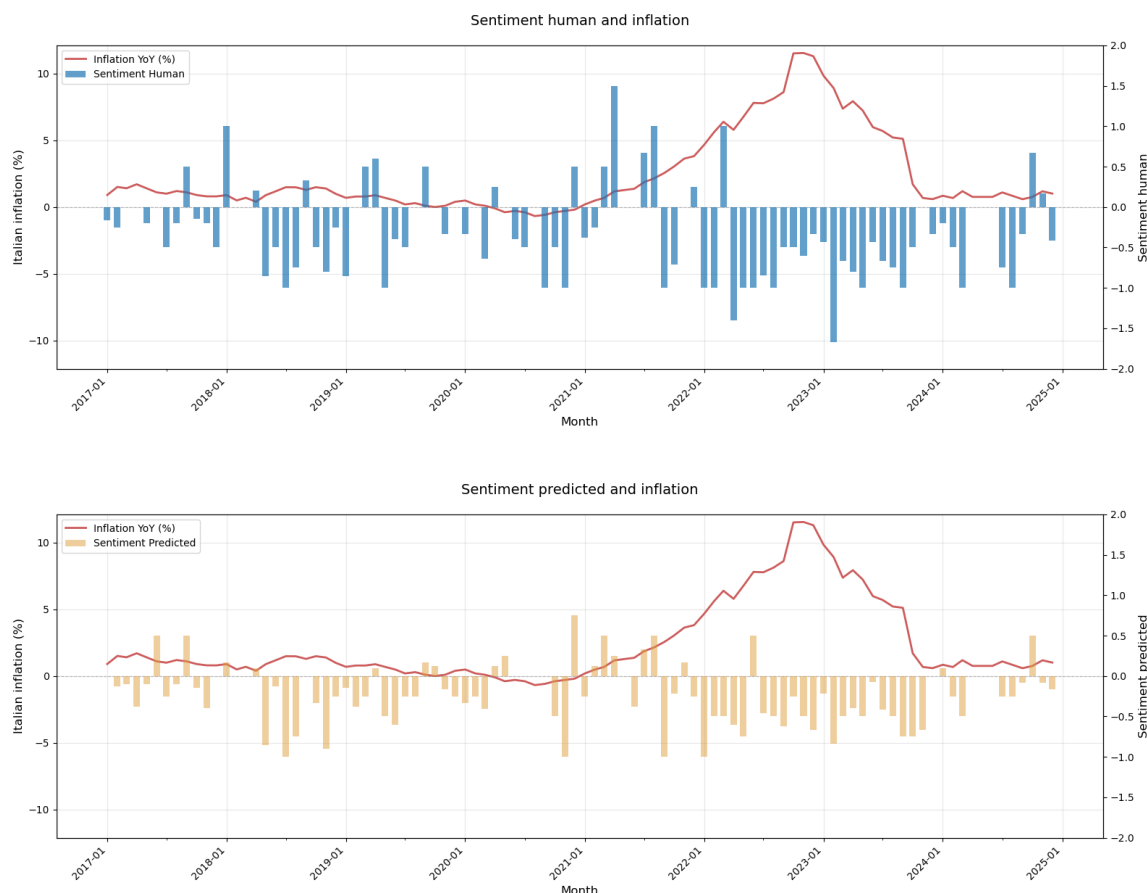


**Figure 4.** Monthly sentiment and Italian CPI inflation (year-over-year). Top: human-aggregated sentiment. Bottom: LLM-predicted sentiment. The figures are descriptive and do not constitute a formal forecasting analysis.

The series suggests a non-linear relationship between tone and price developments. Following prolonged periods of low inflation or mild deflation, an initial increase in inflation is often accompanied by positive sentiment, potentially reflecting improved macroeconomic expectations and normalization of economic activity. By contrast, when inflation accelerates further and remains persistently elevated, sentiment turns markedly negative. These patterns are illustrative and intended to highlight the potential economic relevance of the validated sentiment measure.

Beyond this descriptive comparison, we also conducted preliminary experiments testing whether the sentiment signal improves inflation forecasts when combined with standard macroeconomic predictors, across several model families and a range of feature sets. Table 12 reports the change in forecast error from adding sentiment for three representative models, evaluated on the feature-set configuration where the signal was most informative: a compact set of macroeconomic indicators. The contribution of the sentiment signal depended on both the model and the forecast horizon: it was concentrated at longer horizons, from six months onward, and strongest for the time-series foundation model, more modest for the linear benchmarks. When we broadened the candidate predictor set to include a much larger number

| Model                      | $h = 3$ | $h = 6$ | $h = 9$ |
|----------------------------|---------|---------|---------|
| Ridge                      | +0.01   | -0.05   | -0.15   |
| Lasso                      | -0.00   | +0.03   | -0.14   |
| TimesFM (foundation model) | -0.05   | -0.24   | -0.58   |

Table 12: RMSE change from adding sentiment to inflation forecasts, by model and horizon (months), on the feature-set configuration where the sentiment signal was most informative. Negative values indicate sentiment improved the forecast. Expanding-window backtest, preliminary results.

of macroeconomic indicators, the contribution of sentiment diminished across all models, consistent with increasing overlap between the two sources of information. We report this pattern as a preliminary signal, not a validated result. A formal econometric assessment of predictive content is left for future work.