Automated Media Framing Analysis in Corporate Fraud Scandals

Nipun Angkavichai
Language and Information Technologies,
Faculty of Arts, Chulalongkorn University
6440126922@student.chula.ac.th

Abstract

Media framing—the process by which news outlets select and emphasize particular aspects of a story—shapes public understanding of corporate scandals, yet manual annotation is time-consuming and limited in scope. Automated framing classification in fraud contexts remains under-explored, particularly for multilingual and low-resource settings. Prior work has focused on English political news using topic models or monolingual transformers, and the Wirecard fraud study by Oelrich and Siebold (2024) provided a richly annotated German–English dataset but no automated methods. To address this gap, we fine-tune multilingual transformers (mBERT, XLM-R) on the Wirecard dataset and explore instruction-tuned LLMs (GPT-4o-mini, GPT-3.5, LLaMA 3.1B) via zero- and one-shot prompt engineering. Evaluated against majority and logistic regression baselines, mBERT achieved the best balance of accuracy (0.58) and Macro-F1 (0.58) on Wirecard and maintained reasonable generalization (0.46 accuracy, 0.30 Macro-F1) on new fraud cases. LLMs demonstrated competitive top-k performance (up to 0.90 Top-3 accuracy) and provided interpretable reasoning, though they struggled with minority frames such as moral evaluation. Our results suggest that supervised multilingual transformers currently offer the most reliable automated framing analysis, while LLMs hold promise for low-data or interactive settings. Our automated media framing anaysis in fraud scandals approach supports future research in computational social science.

1 Introduction

Media framing—the selection and emphasis of certain story aspects through rhetoric and style—shapes public perception of corporate scandals (Entman, 1993; Goffman, 1974). Manual framing analysis is labor-intensive and limits scalability across outlets and languages.

Automating this task would enable large-scale studies of how frames (e.g., factualization, failure attribution, moral evaluation (Oelrich and Siebold, 2024)) evolve during fraud scandals and vary between media. Such tools are vital for computational social science, informing media effects and accountability research.

Framing classification is challenging due to overlapping frame semantics, rare minority frames, and multilingual reporting. Existing NLP work focuses largely on English political news—using topic models (Roberts et al., 2014) or monolingual BERT (Mendelsohn et al., 2021)—and does not address fraud domains or other languages.

We present a two-pronged approach: (1) fine-tuning multilingual transformers (mBERT, XLM-R) on Oelrich and Siebold's frame-annotated Wirecard dataset—mostly in German language and some English language—and (2) applying instruction-tuned LLMs (GPT-4o-mini, GPT-3.5, LLaMA 3.1B) via prompt engineering. Our best model, mBERT, achieves 0.58 accuracy and Macro-F1 on Wirecard, generalizes to new fraud cases, and outperforms baselines. We also analyze errors, frame distributions, and top-k performance for our approach.

Contributions

- The first **automated media framing classifier** for corporate fraud scandals, combining supervised multilingual transformers and instruction-tuned LLMs. (Section 3).

- Demonstration that **mBERT fine-tuning** yields the best balance of accuracy and Macro-F1 across frames and languages (Section 5).

- Comprehensive **error analysis** (confusion matrix) revealing overlapping frame challenges (Section 6).

2 Related works

Our work builds on foundational theories of framing from social sciences. Entman (Entman, 1993) defined framing as the selection and salience of information to promote specific interpretations, while Goffman (Goffman, 1974) emphasized the role of cognitive schemas in shaping perception. Building upon these theoretical foundations, we leverage the annotation framework developed in the Wirecard media framing study by Oelrich and Siebold (Oelrich and Siebold, 2024), which used abductive analysis to identify six distinct framing categories across 795 manually coded news articles on the Wirecard scandal. In our research, we adopt this richly annotated Wirecard dataset to develop and evaluate automated media framing classification models, aiming to scale framing analysis to broader contexts of corporate fraud scandals.

Several annotated corpora have been developed for computational framing. The Media Frames Corpus (Card et al., 2015) labels U.S. policy news with general frame dimensions (economics, morality, capacity, etc.), while Liu et al. (Liu et al., 2019) and Akyürek et al. (Akyürek et al., 2020) extended frame classification to specific topics and languages. However, these resources focus mostly on English and political discourse. Prior corporate framing studies (e.g., (Clemente and Gabbioneta, 2017)) remain qualitative and domain-specific.

Early NLP methods for framing used topic modeling (Roberts et al., 2014) to discover latent "frames" in political news. More recently, supervised neural models like BiLSTMs (Naderi and Hirst, 2017) and transformer-based models (e.g., BERT, XLM-R) have become standard (Mendelsohn et al., 2021). For multilingual framing, approaches include cross-lingual projection (Field et al., 2018), translation-based methods and fine-tuning multilingual BERT variants (Akyürek et al., 2020).

Recent work also explores prompting large language models (LLMs). Pastorino et al. (Pastorino et al., 2024) showed that GPT-4 can detect framing bias in headlines using one-shot prompting, though performance varies by task and model size. Still, most LLM studies center on political content and short texts.

Otmakhova et al. (Otmakhova et al., 2024) provide a comprehensive survey of computational approaches to media framing, highlighting the need for models that capture the nuanced aspects of framing beyond mere topic categorization. Their typology underscores the importance of integrating cognitive, linguistic, and communicative dimensions in framing analysis.

Akter and Anastasopoulos (Akter and Anastasopoulos, 2024) address the challenge of scaling multilingual framing analysis. They introduce the Student-sourced Noisy Frames Corpus (SNFC), covering domains like immigration and same-sex marriage in languages such as Bengali and Portuguese. Their work demonstrates the viability of crowdsourcing and machine translation for creating multilingual framing datasets, and they find that task-specific fine-tuning of LLMs outperforms using larger, non-specialized models.

Our work contributes by automating framing analysis in corporate fraud scandals, applying multilingual transformers and instruction-tuned LLMs to a novel domain. We also emphasize generalization across languages. Unlike prior work, our frame schema is grounded in framing theory and media rhetoric, not only topic categorization.

3 Methodology

The task of media framing classification in fraud scandals involves interpreting how news narratives emphasize certain angles, actors, or consequences of events. This is a highly contextual and language-sensitive task, requiring models that can capture not only surface-level keywords but also nuanced framing strategies that vary across cultures and media outlets. The primary dataset used in this study is mostly in German, with some English data offers potential for broader multilingual exploration. For this reason, we require models that can understand and generalize across multiple languages to ensure robustness and applicability beyond a single linguistic context.

To achieve our objectives, we employ two main strategies:

- **Transformer-based fine-tuning:** We fine-tune pre-trained transformer models (such as mBERT and XLM-R) on our labeled dataset to capture deep, contextual semantic representations. These models are well-suited to the framing task due to their ability to model long-range dependencies and multilingual texts, which is essential for capturing framing across various media and languages.

- **Large Language Models (LLMs):** We explore both in-context learning and fine-tuning approaches with LLMs. For in-context learning, we design prompts that reflect the framing task and provide a one-shot format to guide the model's output. For fine-tuning, we use lightweight adaptation methods such as LoRA on models like LLaMA to improve task-specific performance in a parameter-efficient way.

To evaluate the effectiveness of our approaches, we compare them against two baselines:

- A **majority class baseline** that always predicts the most frequent frame in the training data (*Failure Attribution*).
- A **logistic regression model** trained on TF-IDF weighted bag-of-words features to represent the text in a simpler, interpretable manner.

4 Experiment

4.1 Dataset

This study utilizes a dataset comprising 795 newspaper articles on the 2020 Wirecard fraud scandal from prior research (Oelrich and Siebold, 2024) . Each article is annotated with one of six distinct media frames, identified based on the principles of selection and salience.

These principles govern how certain aspects of a story are selected through bounding (i.e., which aspects of a perceived reality are focused on or kept out of focus, thereby regulating what the audience sees) and contextualization (i.e., how certain aspects of a perceived reality are considered within a setting in which they occur and are given meaning) and make selected aspects more salient through articulation (i.e., how selected aspects of a perceived reality are communicated to an audience) and directionality (i.e., how selected aspects of a perceived reality are used to elicit or prime the audience for a specific interpretation or argument of presented events).

The six annotated frames are briefly defined as follows:

1. **Failure Attribution:** Emphasizes oversight bodies' malpractice in accounting fraud. Articles often clarify and incriminate, using rhetorical devices to highlight failure in fraud prevention.

   For example: *"The affair surrounding the crooked financial firm Wirecard can be summed up in one word: TOTAL SCREW UP! [...] Germany as a financial center has disgraced itself to the bone."*

2. **Treatment Recommendation:** Focuses on reforms for oversight bodies and auditors. Articles are action-oriented, quoting experts and advocating changes to prevent future fraud.

   For example: *"Tighter access for the financial regulator BaFin. Greater liability for auditors. Much more transparency by companies in their balance sheet disclosures."*

3. **Factualization:** Centers on the key facts about the fraud, presenting sober and objective reporting supported by financial data and expert analysis.

   For example: *"The scandal caused an 80 percent plunge in the company's stock price over the last two days."*

4. **Consequences:** Highlights the financial and reputational impacts on the company and stakeholders, often employing emotional narratives to underscore significant losses.

   For example: *"Worst affected: Small investors! All those who have invested their hard-earned savings in order to use the returns to pay for their retirement."*

5. **Personification:** Examines the criminal behavior of top managers, with sensational reporting that compares events to popular culture or films.

   For example: *"Entertaining and action-packed, but also pretty incredible. [...] Now he [COO Jan Marsalek] has pulled off a cinematic escape of his own."*

6. **Moral Evaluation:** Critiques the lack of morality among stakeholders, using judgmental rhetoric to convey outrage and condemnation.

   For example: *"He [CEO Markus Brown] did what he had presumably also done before as Wirecard boss: He lied! [...] He tricks, he ruthlessly pursues his own advantage."*

For model development, the Wirecard dataset was divided into three subsets: 64% for the training set (508 rows), 16% for the development set (128 rows), and 20% for the test set (160 rows).

The frame distribution in the dataset is as follows: Failure Attribution (27.01%), Factualization (19.09%), Treatment Recommendation (17.58%), Personification (11.05%), and Moral Evaluation (6.40%).

In addition to the Wirecard dataset, we also used a separate dataset consisting of 35 frame-annotated articles from four different fraud cases for testing purposes: OneCoin (20 articles), Benko (8 articles), Schlecker (5 articles), and Skania (2 articles). This dataset serves to evaluate the generalizability of the model to new, unseen cases. The frame distribution in this dataset is as follows: Personification (40%), Factualization (23%), Consequences (17%), Failure Attribution (14%), Treatment Recommendation (3%), and Moral Evaluation (3%).

A detailed breakdown of the frame distribution of the datasets is shown in Table 1.

The Wirecard dataset predominantly consists of German articles (95%), with only 5% in English. This highlights the need for the model to generalize well across languages, as future test cases need to include articles in other languages.

The articles are notably lengthy, with a median word count of 390.50 and a mean of 457.35. Additionally, the word count shows high variability, with a standard deviation of 289.65, and numerous outliers further skewing the distribution.

4.2 Preprocessing methods

In Wirecard training data, beyond fine-tuning on raw article texts, we also explored feature engineering techniques separately to customize the input representation in the following ways:

4.2.1 Translation-Based Data Augmentation

We generated additional training samples by translating German articles into English. This translation technique aims to introduce lexical and syntactic variations in the training data.

The underlying hypothesis is that exposing the model to such linguistic diversity can improve its ability to generalize across different writing styles and phrasings.

4.2.2 Paragraph/Sentence-Level Classification

We experimented with splitting articles into smaller units, such as paragraphs and sentences. Paragraphs were identified using newline characters—ideally double newlines—but since only around 59 articles used them consistently, we also considered single newlines for more comprehensive segmentation.

For sentence splitting, we used the NLTK library. To assign frame labels to these segments, we leveraged the `exemplary_quote` column in the Wirecard dataset. If a paragraph or sentence contained part of the quote, it was labeled with the article's original frame; otherwise, it was labeled as `no signal`.

The hypothesis is that this finer granularity could help the model more accurately capture localized framing signals, especially in complex articles with multiple perspectives.

4.2.3 Excluding the First Paragraph

We investigated whether removing the first paragraph—typically dominated by factual or summary content—could improve classification performance by emphasizing more frame-specific sections. Paragraphs were detected using double newline separation.

An analysis of first-paragraph lengths revealed a mean of 11 lines and a median of 8 lines. Based on this, we conducted two experimental setups: one excluding the first 11 lines and another excluding the first 8 lines from each article during training.

To support this decision, we referenced the `exemplary_quote` column in the dataset, which contains excerpts from the article that reflect the annotated frame. We observed that these quotes rarely appear in the opening paragraph, reinforcing the idea that the first paragraph often contains general information and may dilute the framing signal.

The rationale behind this approach is that filtering out less informative content may allow the model to focus on parts of the article where distinct framing cues are more likely to appear.

4.2.4 LLM Data Augmentation

We use GPT-4o-mini to generate synthetic articles for training by leveraging structured factual inputs (see Table 9). Specifically, we prepare six distinct fact lists, three related to the scandal period and three to the post-scandal period.

For each fact list, we prompt the LLM to write six articles, each reflecting a different media frame. This results in 36 additional articles (6 frames × 6 fact lists), expanding the training dataset with contextually coherent but diverse examples.

The hypothesis is that this technique allows us to control the framing of input data and introduce

Frame	Train	Dev	Test	Other Fraud Cases
Failure Attribution	137	35	43	5
Treatment Recommendation	96	24	30	1
Factualization	97	24	31	8
Consequences	89	23	28	6
Personification	56	14	18	14
Moral Evaluation	33	8	10	1
Total	**508**	**128**	**160**	**35**

Table 1: Frame distribution across our primary Wirecard splits (training, development, test) and the supplementary "Other Fraud Cases" dataset (35 articles).

high-quality, frame-specific samples to support low-resource frame categories.

4.3 Experimental conditions

4.3.1 Transformer-based model

We fine-tune several pretrained multilingual and German-specific transformer architectures to capture rich contextual representations:

- **mBERT** (Pires et al., 2019): a multilingual BERT model pretrained on Wikipedia in 104 languages, which provides strong cross-lingual embeddings and baseline performance across diverse language settings.

- **German BERT** (Darji et al., 2023): Google's German-specific BERT variant, pretrained on large German news and web corpora, designed to capture nuanced German vocabulary and syntax.

- **XLM-RoBERTa** (Conneau et al., 2019): a RoBERTa-based model trained on 2.5TB of CommonCrawl data spanning 100 languages, offering enhanced robustness for low-resource languages and domain adaptation.

- **XLM-Long** (Sagen, 2021): an extension of XLM-R using the Longformer attention mechanism. This model is based on the master's thesis work that adapts multilingual models like XLM-R into long-context variants without retraining from scratch on long-sequence multilingual corpora. Instead, the model leverages Longformer-style pretraining applied to the original XLM-R, primarily on English data, while preserving cross-lingual capabilities. This adaptation allows the model to process sequences of up to 4096 tokens.

Each model is fine-tuned on our Wirecard training set to optimize performance on the media framing classification task.

4.3.2 Large language models (LLMs)

We leverage large language models through two key approaches: (1) *in-context learning*, where prompt engineering guides predictions in zero-shot and one-shot settings (Brown et al., 2020); and (2) *parameter-efficient fine-tuning*, where models are further trained on labeled data using LoRA (Low-Rank Adaptation) (Hu et al., 2021) to improve performance on the media framing classification task.

For in-context learning, we utilize OpenAI's GPT-3.5, GPT-4o-mini, and the o1 model (OpenAI, 2023). For both in-context learning and fine-tuning, we employ Meta's LLaMA 3.1 8B Instruct model (Grattafiori, 2024). All models are accessed through the Hugging Face Transformers library.

We designed several prompt engineering strategies to explore the capabilities of large language models for frame classification:

- **Zero-shot prompt**: A standard prompt including detailed criteria—*bounding*, *contextualization*, *articulation*, and *directionality*—for each of the six media framing categories. The model is instructed to select and return the single most relevant category (see Table 3).

- **Zero-shot + Chain-of-thought (CoT)**: An extended version of the zero-shot prompt, encouraging the model to reason step-by-step by explaining how media framing dimensions apply to the article, using the phrase "Let's think step by step" to induce structured reasoning. The model is asked to return dominant and sub-dominant frames with supporting rationale. (see Table 4 for the prompt and Table 8 for the output).

- **Zero-shot voting prompt**: Four separate prompts are constructed—each focusing solely on one framing dimension (e.g., bounding-only prompt). The model is asked to return dominant and sub-dominant frames with supporting rationale per criterion. A majority vote determines the final predicted frame. In cases of ties, multiple top-ranked predictions are retained (see Table 5).

- **Zero-shot Top-k prompt**: Based on the standard zero-shot setting, this prompt explicitly instructs the model to rank and return the top three most relevant frames for an article—enabling evaluation using Top-k metrics (see Table 6).

- **One-shot prompt**: We select a representative article for each of the six frames using the highest-confidence predictions from a logistic regression model trained on the Wirecard dataset. These examples are then prepended to the prompt as demonstrations to guide the model's prediction on new examples (see Table 7).

4.4 Implementation

All pretrained models are sourced from the Hugging Face library. The fine-tuning process is conducted using Google Colab Pro+ with an NVIDIA A100 GPU, allowing for efficient training on long documents and computationally intensive models. All training progress reports to Weights & Biases (W&B) for experiment tracking.

4.4.1 Hyperparameter Tuning for Transformer Models

We fine-tuned transformer-based models using random search over the following hyperparameter ranges:

- Learning rates: `1e-5`, `3e-5`, `5e-5`, `7e-5`
- Batch sizes: `4`, `8`, `16`, `32`, `64`
- Number of epochs: `10`, `20`, `30`

We selected the best model checkpoint based on accuracy on the development set. All models were trained with a maximum sequence length of 512 tokens, except for XLM-long, which was configured for up to 1024 tokens.

For mBERT, the best-performing configuration was a learning rate of `7e-5`, batch size of `32`, and training for `10` epochs, with evaluation and checkpoint saving every `10` steps.

4.4.2 Prompt Engineering for LLMs

To extract structured outputs from instruction-following LLMs, we employed prompt engineering techniques and used `pydantic` along with Python's built-in `typing` library (`Dict`, `Optional`, `List`) to enforce structured response formatting in JSON-like outputs.

4.4.3 Fine-Tuning LLaMA 3.1 with LoRA

We briefly experimented with fine-tuning LLaMA 3.1 using LoRA (Low-Rank Adaptation) and HuggingFace's `SFTTrainer`. We followed standard configurations based on QLoRA, using a learning rate of `2e-4`, `LoRA rank = 64`, and training for 10 epochs with gradient checkpointing enabled. Due to the high cost of training large LLMs, this was done only for exploratory purposes. Interestingly, further increasing the number of training epochs (e.g., to 20 or 40) led to a drop in accuracy, possibly due to overfitting on the small training set.

4.5 Evaluation

To evaluate model performance on the multi-class media framing classification task, we use three primary metrics: Macro-F1 Score, Accuracy, and Top-k Accuracy.

Macro-F1 Score: The F1 score is the harmonic mean of precision and recall. Since our task involves imbalanced frame categories, we adopt the Macro-F1 score, which computes the F1 score independently for each class and then averages them. This ensures that all frames are treated equally, regardless of their frequency in the dataset, making it well-suited for evaluating performance across both majority and minority frames.

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Accuracy: We also report accuracy, which measures the proportion of correctly predicted frames out of all predictions. While accuracy gives an overall sense of correctness, it may be biased toward the majority class in imbalanced datasets, and therefore should be interpreted alongside Macro-F1.

Top-k Accuracy: In top-k evaluation, the model is considered correct if the true frame appears

Method	Wirecard Test Set		Other Fraud Cases	
	Accuracy	Macro F1	Accuracy	Macro F1
Baseline				
Majority baseline	0.27	–	0.40	–
LR	0.59	0.53	0.17	0.05
LR + LLM Data Augmentation	**0.60**	0.54	**0.17**	**0.11**
LR + Exclude First Paragraph (Train set)	0.54	0.48	–	–
LR + Exclude First Paragraph (All dataset)	0.46	0.41	–	–
LR + Translation Data Augmentation	0.55	**0.55**	–	–
Transformer Models				
mBERT	0.58	**0.58**	0.46	0.30
mBERT + LLM Data Augmentation	0.47	0.43	0.37	0.23
mBERT + Exclude First Paragraph (Train set)	0.46	0.46	–	–
mBERT + Exclude First Paragraph (All dataset)	0.40	0.42	–	–
mBERT + Translation Data Augmentation	0.55	0.54	–	–
XLM-Long (window=1024)	0.56	0.49	0.51	**0.37**
BERT-German	**0.61**	0.54	0.37	0.23
XLM-R-Large	0.57	0.50	**0.60**	**0.37**
LLMs				
GPT-4o mini (Zero-shot bounding prompt)	0.41	0.32	0.20	0.10
GPT-4o mini (Zero-shot contextualizing prompt)	0.41	0.35	0.20	0.11
GPT-4o mini (Zero-shot articulation prompt)	0.42	0.31	**0.31**	0.20
GPT-4o mini (Zero-shot directionality prompt)	0.41	0.26	0.29	0.19
GPT-4o mini (Zero-shot voting)	0.56	0.46	**0.31**	**0.21**
GPT-4o mini (Chain-of-thought & Zero-shot)	0.52	0.43	–	–
GPT-4o (Zero-shot)	0.54	0.47	–	–
GPT-3.5 (Zero-shot)	0.49	0.35	–	–
GPT-3.5 (One-shot)	0.32	0.26	–	–
OpenAI o1 (Zero-shot)	**0.57**	**0.52**	–	–
LLaMA 3.1 8B (Zero-shot)	0.39	0.32	–	–
LLaMA 3.1 8B (LoRA fine-tuned)	0.44	0.36	–	–

Table 2: Model performance on the Wirecard and other fraud cases. Accuracy and Macro-F1 are reported. Dashes indicate results not available.

among the top-k highest scoring frames. For models like mBERT, this is determined by the top-k logits or probabilities, while for LLMs, it is based on the top-k most relevant or likely frames returned from in-context responses. This metric helps evaluate how well the model ranks the correct frame, even if it is not the top-1 prediction.

5 Results

5.1 Evaluation of All Approaches: Accuracy and Macro-F1 Metrics

Table 2 summarizes model performance on both the primary Wirecard test set and the supplementary "Other Fraud Cases" test set. Across all metrics, fine-tuned transformer models outperform both traditional baselines and in-context LLM approaches.

Among the baselines, logistic regression (LR) with TF-IDF features yields 0.59 accuracy and 0.53 Macro-F1 on the Wirecard test set. Augmenting LR with LLM-generated synthetic data marginally improves performance to 0.60 accuracy and 0.54 Macro-F1. However, when evaluated on the Other Fraud Cases dataset, LR's performance drops drastically to just 0.17 accuracy and 0.05 Macro-F1, indicating poor generalization. This sharp decline suggests that the keyword-based approach overfits to case-specific vocabulary and fails to capture deeper framing structures that differ across contexts. In contrast, transformer-based models that leverage contextual embeddings are better equipped to generalize across varying fraud narratives.

Within the transformer family, mBERT strikes the best balance between overall accuracy (0.58) and, crucially, Macro-F1 (0.58) on the Wirecard set—demonstrating robust, balanced performance across all six frame classes. Although BERT-German achieves the highest accuracy (0.61), its Macro-F1 (0.54) lags behind mBERT, indicating less consistency on minority frames. While XLM-R-Large and XLM-Long generalize well in other fraud cases, with XLM-Large obtaining 0.60 accuracy/0.37 Macro-F1 and XLM-Long reaching 0.51 accuracy/0.37, their Macro-F1 also lags behind mBERT in Wirecard test set.

For Paragraph/Sentence-Level Classification method. the model consistently failed to assign any of the six target frames, instead classifying all

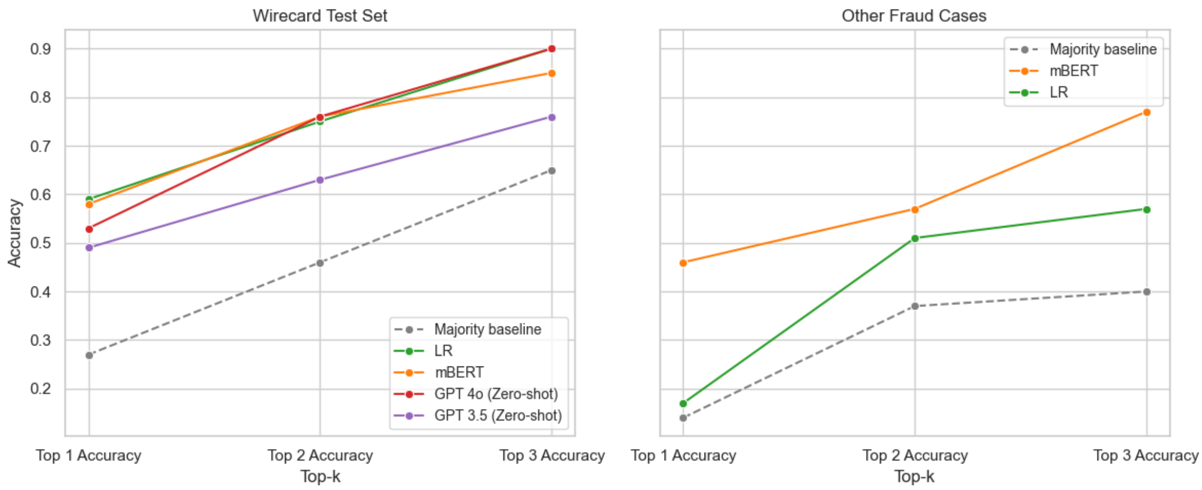

Figure 1: Top-k accuracy of selected models on the Wirecard test set and Other Fraud Cases. Each line chart shows the accuracy when allowing the model to select the correct frame within its top-1, top-2, or top-3 predictions.

units as "none" or irrelevant. Due to this persistent lack of signal, we exclude these results from further reporting.

Among the instruction-following large language models (LLMs), OpenAI o1 stands out with 0.57 accuracy and 0.52 Macro-F1—competitive with the best supervised models—even though it is evaluated in a zero-shot setting. This suggests that high-capacity instruction-tuned models can internalize generalizable frame semantics without direct task-specific training. Interestingly, GPT-3.5 (zero-shot) outperforms its one-shot variant (0.49 vs. 0.32 accuracy), indicating that one-shot prompting can sometimes introduce misleading inductive biases, especially in framing tasks where subtle contextual reasoning is required. Moreover, although LLaMA 3.1 8B, when fine-tuned with LoRA on labeled data, slightly improves over its zero-shot variant (0.44 vs. 0.39 accuracy), the gain is marginal—suggesting that providing an example of frame-annotated articles does not substantially benefit LLMs, potentially due to the limited or case-specific nature of the training data.

While LLMs like GPT-4o mini combining multiple prompting strategies via voting show promise (achieving 0.56 accuracy and 0.46 Macro-F1), they still lag behind supervised transformer models like mBERT. Notably, the voting-based approach not only improves performance but also produces reasoning outputs that align well with the predicted frames, making its predictions more interpretable and justifiable upon qualitative inspection.

However, the relatively low Macro-F1 scores across most LLM configurations on the Wirecard dataset—and even lower accuracy and F1 on the Other Fraud Cases—underscore the difficulty of using raw in-context learning to address media framing classification, particularly in settings with highly imbalanced label distributions. Additionally, we observe that LLMs rarely predict the moral evaluation frame, with even the best-performing instruction-following model, OpenAI o1, predicting this frame only once out of 160 samples—despite 10 ground truth instances in the test set. This highlights a key limitation in LLMs' ability to detect nuanced normative framing in zero-shot settings.

Because the "Other Fraud Cases" dataset comprises only 35 articles with uneven frame distribution, its results should be interpreted cautiously. Nonetheless, mBERT's stable Macro-F1 on Wirecard, combined with its strong accuracy, makes it our preferred model for media-framing classification in fraud scandals.

5.2 Top-k Evaluation Performance

To better evaluate the models' ability to suggest plausible frames beyond their top prediction, we conduct a Top-k evaluation (k = 1, 2, 3), where a prediction is considered correct if the true label appears within the top-k candidates.

Figure 1 illustrates the top-k performance for both the Wirecard test set and the Other Fraud Cases, shown as two line charts in a single figure. In the Wirecard set, supervised models such as Logistic Regression and mBERT, which learn from labeled training data, show strong Top-k improvements. For instance, mBERT improves from 58%

(Top-1) to 76% (Top-2) and 85% (Top-3) accuracy. Similarly, Logistic Regression reaches 90% Top-3 accuracy, demonstrating that even linear models can perform well with informative features.

In contrast, LLMs like GPT-4o and GPT-3.5, which operate purely through instruction-based prompts without task-specific training (i.e., in a zero-shot setting), approach similar Top-3 performance—GPT-4o achieves 76% Top-2 and 90% Top-3 accuracy—despite lower Top-1 scores. This underscores their capacity to understand the task as a form of reading comprehension, selecting plausible frames directly from textual guidance.

For the Other Fraud Cases, although the dataset is relatively small and suffers from class imbalance, mBERT demonstrates substantially stronger performance compared to baseline methods. It achieves 77% Top-3 accuracy, outperforming both the majority baseline (40%) and a Logistic Regression (LR) model (57%) trained on the same data. The poor performance of the majority baseline highlights a key limitation: it relies on the label distribution of the Wirecard training set, which does not align with the distribution in other fraud cases. Specifically, the top-3 frames in Wirecard are Failure Attribution, Factualization, and Treatment Recommendation, while in other fraud cases, the top-3 frames shift to Personification, Factualization, and Consequences. This mismatch renders the majority baseline ineffective for generalization. Similarly, the LR model, while slightly better, still underperforms due to its limited capacity to capture contextual semantics. In contrast, mBERT's contextualized embeddings allow it to generalize framing patterns beyond surface-level correlations, making it more robust for frame detection in unseen fraud narratives.

These results suggest that while supervised models maintain stronger precision when trained on task-specific data, LLMs demonstrate promising capabilities in zero-shot scenarios, especially when flexibility in predictions is permitted—making them suitable for assisting annotation or multi-label suggestions in frame analysis.

In particular, our **Zero-Shot + Chain-of-Thought** prompt with GPT-4o-mini, which asks the model to identify both dominant and sub-dominant frames, yields insightful qualitative outputs (see Table 8). For example, when analyzing a Wirecard article whose gold frame is "Consequences," GPT-4o-mini accurately highlights the key passages (formatted in **bold**) that correspond to the annotated consequence frame. However, it nevertheless labels the dominant frame as "Failure Attribution," likely due to overlapping language cues. Despite this misclassification, the model's extracted quotations and step-by-step reasoning closely mirror the annotator's own focus, demonstrating its capacity to attend to the same salient information. This suggests that finer-grained annotation—at the paragraph or sentence level—could be leveraged to aggregate local frame signals and improve document-level prediction, especially in long articles where multiple frames co-occur.

6 Discussion

6.1 Frame-Level Insights and Model Confusions from mBERT Prediction

While the supervised mBERT model shows strong overall performance on the media framing classification task, a closer look reveals that its effectiveness varies significantly across different frame types. On the Wirecard test set, mBERT achieves strong F1 scores for Personification (80%), followed by Treatment Recommendation (63%), Consequences (61%), Factualization (57%), Failure Attribution (51%), and finally Moral Evaluation (36%). This wide range highlights varying levels of difficulty in identifying certain frames, particularly more abstract or normative ones like Moral Evaluation.

On the Other Fraud Cases test set, a drop in performance is observed across nearly all frames. F1 scores for this set are: Personification (64%), Factualization (53%), Consequences (43%), Failure Attribution (22%), and 0% for both Moral Evaluation and Treatment Recommendation—the latter likely due to the extreme class imbalance, as only two instances of each appear in the dataset. Initially, we were concerned that the model might overfit to Personification in the Wirecard case, since this frame often refers to top executives, whose identities vary across fraud cases. When tested with keyword-based models like Logistic Regression using TF-IDF feature, it was found that the name of Wirecard's COO, Jan Marsalek, emerged as a key indicator strongly associated with this frame. This reliance on case-specific identifiers implies that the model may struggle to generalize when applied to other fraud cases involving different individuals. However, the relatively strong generalization of Personification to unseen fraud cases supports the model's ability to capture the general semantics of this frame beyond named entities.

In terms of predicted frame distribution, mBERT's predictions are generally aligned with the ground truth distribution reported in Table 1. Specifically, the predicted proportions are: Factualization (29%), Failure Attribution (25%), Treatment Recommendation (15%), Consequences (13%), Personification (11%), and Moral Evaluation (7%). While the top two frames are swapped in order compared to the actual distribution, this close match suggests the model does not heavily overproduce or underproduce particular frames, which is encouraging for downstream interpretability and fairness.

Further analysis of the confusion matrix in Figure 2 reveals systematic misclassifications between certain frames. For instance, Consequences is often misclassified as Factualization (7 out of 28 times), and Factualization is misclassified as Failure Attribution (9 out of 35 times). Additionally, Treatment Recommendation is frequently confused with both Failure Attribution (7 errors) and Factualization (4 errors). These confusions suggest that factual or causal reasoning often gets blurred in the model's learned representation space. Given that Factualization, Failure Attribution, and Consequences all involve statements about causes, effects, or assertions of truth, it is reasonable that the model occasionally struggles to cleanly separate them.

Overall, the findings highlight key strengths and remaining challenges. mBERT handles factual and actor-centered frames relatively well, and it generalizes personification effectively to unseen fraud cases. However, the model still struggles with subtle distinctions between overlapping frames and performs poorly on rare or normatively loaded frames, especially Moral Evaluation, where both F1 scores and classification counts remain low. Addressing these issues may require richer contextual understanding, targeted data augmentation, or frame-aware loss functions to improve separation between semantically similar classes.

To demonstrate the practical utility of our approach, we applied the fine-tuned mBERT model to a new corpus of 768 news articles covering the CumEx Fraud Scandal—one of the largest tax fraud schemes in Europe involving banks, investors, and legal advisors who exploited dividend refund loopholes to reclaim taxes never paid. Our focus was on the period between 2019 and 2020, during which

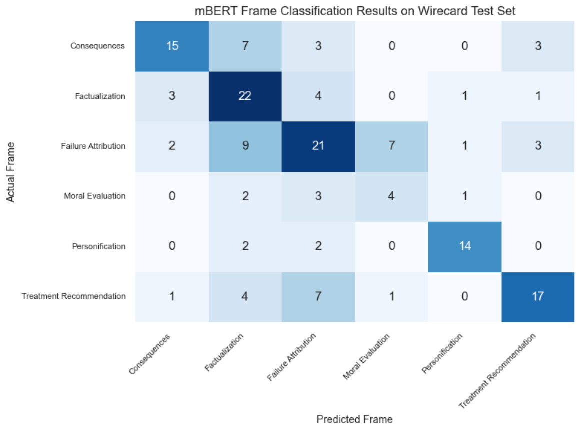

Figure 2: mBERT Frame Classification Results on the Wirecard Test Set. The confusion matrix shows the number of documents predicted for each frame (columns) versus the actual frame labels (rows).

public discourse shifted from initial revelations to intensified legal scrutiny and prosecution. By classifying these articles at the document level, we can examine how media framing evolved over time. As illustrated in Figure 3, the model enables temporal and cross-sectional analysis of dominant frames, raising questions such as: How do different framing categories (e.g., failure attribution or treatment recommendation) gain prominence at various stages of a scandal? How does framing differ across media outlets or political orientations? This case study underscores the value of automated framing tools in enhancing large-scale content analysis for journalism, media studies, and public policy research.

6.2 Preprocessing Methods Do Not Outperform Using Raw Article

Various preprocessing and data augmentation techniques were explored to improve frame classification, but none consistently outperformed using the full raw articles as input for mBERT.

(1) Translation-Based Augmentation added machine-translated sentences to double the dataset size, but this led to lower performance. MaxEnt's accuracy dropped from 0.59 to 0.55, suggesting that the added lexical variety introduced noise—particularly keyword overlap between competing frames—rather than improving generalization.

(2) Paragraph and Sentence-Level Classification approaches also underperformed. Models trained on individual paragraphs or sentences predicted nearly everything as 'no signal'. This failure likely stems from inconsistent labeling—many

paragraphs lack annotations even if they carry framing cues—making the 'exemplary_quote' column unsuitable as a standalone supervision target.

(3) Removing the First Paragraph or Initial Lines also did not help. In mBERT, omitting the first 11 lines reduced accuracy to 45.6%, while even partial removal (e.g., 8 lines) only slightly improved results (49.0%). Further removal consistently hurt performance, showing that the early parts of the article contain essential context, often factual but still predictive of downstream framing.

(4) LLM-based Data Augmentation marginally improved Logistic Regression (accuracy from 0.59 → 0.60; Macro-F1 from 0.53 → 0.54), especially for rare frames like Treatment Recommendation. However, mBERT performance degraded with this augmentation (accuracy 0.58 → 0.47; F1 0.58 → 0.43), suggesting that mBERT already captures sufficient semantics from raw data and is more sensitive to noisy or inconsistent synthetic examples.

Together, these results indicate that preprocessing strategies aimed at simplifying or restructuring the input often harm model performance, particularly for transformer-based models like mBERT. Full raw articles remain the most reliable format, as they preserve the narrative and semantic context crucial for accurate frame identification.

7 Conclusion

In this paper, we address the problem of automating media framing classification in corporate fraud scandals, with a focus on multilingual and generalizable models.

To tackle this, we fine-tuned transformer-based models such as mBERT and leveraged instruction-tuned large language models (LLMs) for both supervised and zero-shot settings. Our best-performing model, mBERT, achieved the best balance between accuracy and F1-score, demonstrating strong performance across diverse framing categories and language inputs. Notably, both mBERT and LLMs exhibited high Top-3 accuracy, which may be particularly useful in practical applications that allow for multiple framing interpretations per article.

For future work, we recommend expanding the frame-annotated datasets to include a wider range of fraud cases and languages. Improving frame definitions to reduce ambiguity and increasing annotated examples for minority frames would help the model learn more effectively. Additional data would also allow for more robust testing of cross-case and cross-lingual generalization. Furthermore, future research could explore framing classification at a finer granularity, such as at the paragraph or sentence level, to better capture nuanced and localized framing within longer articles.

Acknowledgements

I would like to express my deepest gratitude to Associate Professor Attapol Rutherford from the Department of Linguistics, Chulalongkorn University, for his invaluable guidance, support, and mentorship throughout every stage of this research project. His insights and encouragement were instrumental in shaping the direction and execution of this work.

I also sincerely thank Sebastian Oelrich and Nicole Siebold, the creators of the Wirecard dataset and authors of the foundational study on media framing in corporate fraud, for their generous support. Their assistance in clarifying the annotation framework and providing access to extended versions of the dataset greatly enriched the scope and quality of this research.

References

Syeda Sabrina Akter and Antonios Anastasopoulos. 2024. A study on scaling up multilingual news framing analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4156–4173. Association for Computational Linguistics.

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Dallas Card, Amber Boydstun, Justin Gross, Philip Resnik, and Noah Smith. 2015. The media frames corpus: Annotations of frames across issues. 2:438–444.

Marco Clemente and Claudia Gabbioneta. 2017. How does the media frame corporate scandals? the case of german newspapers and the volkswagen diesel scandal. *Journal of Management Inquiry*, 26:105649261668930.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German bert model for legal named entity recognition. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, page 723–728. SCITEPRESS - Science and Technology Publications.

Robert Entman. 1993. Framing: Toward clarification of a fractured paradigm. *The Journal of Communication*, 43:51–58.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press, Cambridge, MA.

Jauhri et al. Grattafiori, Dubey. 2024. The llama 3 herd of models.

Edward J. Hu, Yelong Shen, et al. 2021. Lora: Low-rank adaptation of large language models.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Sebastian Oelrich and Nicole Siebold. 2024. Media framing in wirecard's fraud scandal: Facts, failures, and spying fraudster fantasies. *Critical Perspectives on Accounting*, 100:102755.

OpenAI. 2023. Gpt-4 technical report.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428. Association for Computational Linguistics.

Valeria Pastorino, Jasivan Sivakumar, and Nafise Sadat Moosavi. 2024. Decoding news narratives: A critical analysis of large language models in framing bias detection. *ArXiv*, abs/2402.11621.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Markus Sagen. 2021. Large-context question answering with cross-lingual transfer. Master's thesis, Uppsala University, Department of Information Technology.

Appendix

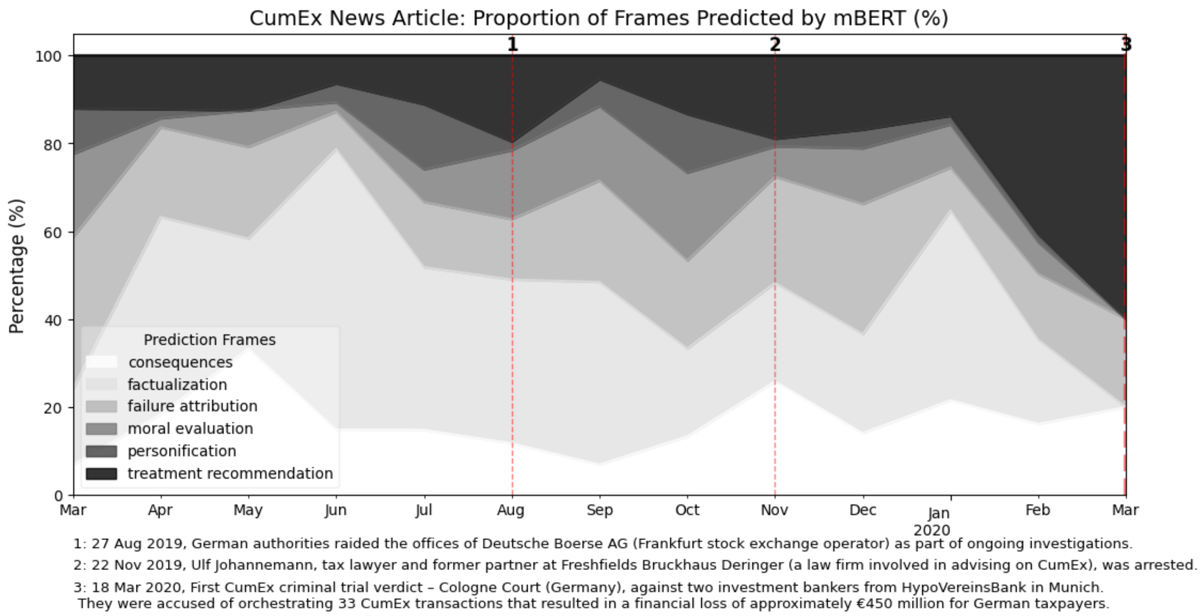

Figure 3: Stacked area chart showing the application of the mBERT model to analyze 768 news articles on the CumEx Fraud Scandal. The scandal, a significant European tax fraud involving banks, investors, and lawyers exploiting tax loopholes, is examined during the critical period of 2019–2020, highlighting the transition from initial exposure to legal prosecution.

Zero-Shot Prompt for Media Framing Classification

```
You are a media framing classifier.
You have to classify news articles into six categories:
failure attribution, treatment recommendation, factualization,
consequences, personification, and moral evaluation.
Please classify it into the single most relevant framing category
and return only the name of that category as the response.

Here are the definitions for each category:
1) failure attribution
bounding:  Focus is placed on the malpractice of oversight
bodies and the auditing profession.
contextuationlization:  The selected aspects of accounting fraud
are considered as a consequence of the lack of appropriate
action of oversight bodies and auditors in preventing fraud.
articulation:  News is reported in a clarifying, incriminating manner;
rhetorical questions, sarcasm, etc.  used to illustrate malpractice.
directionality:  Oversight bodies and auditors failed to act
and are thus defined as the main problem in not preventing fraud.

2) treatment recommendation
bounding:  Focus is placed on ramifications for oversight bodies
and the auditing profession.
contextuationlization:  Fraud is considered a consequence of
lack of appropriate rules and reforms of oversight and auditing.
articulation:  News is forward-looking, action-oriented;
experts call for reforms.
directionality:  Power and resources of oversight bodies and auditors
are recommended to be increased to prevent future fraud.

3) factualization
bounding:  Focus is placed on providing key facts and info
about the fraud.
contextuationlization:  Fraud is considered a consequence of
internal failures leading to the company's demise.
articulation:  News is sober, fact-based; experts and insiders
are quoted.  Financial data is used as evidence.
directionality:  The emergence of fraudulent statements from
a formerly successful company is depicted as a surprise.

4) consequences
bounding:  Focus is placed on the consequences for the
company and key stakeholders.
contextuationlization:  Fraud is considered severe wrongdoing
imposing financial and reputational harm on stakeholders.
articulation:  News is emotional, dramatic; stakeholder stories
highlight personal losses.
directionality:  Financial and reputational losses of the defrauding
company, oversight bodies, and auditors are described as extensive.

5) personification
bounding:  Focus is placed on individual top managers
responsible for the accounting fraud.
contextuationlization:  Fraud is considered the criminal behavior
of the defrauding company's top managers.
articulation:  News is sensational, entertaining; references
to popular culture (thrillers, spies) are used.
directionality:  The criminality and depravity of top managers
are portrayed as mystifying and cinematic.

6) moral evaluation
bounding:  Focus is placed on the morality of key stakeholders.
contextuationlization:  Fraud is considered the consequence
of a lack of morality and unethical behavior.
articulation:  News is judgmental, with rhetorical questions
or sarcasm.
directionality:  The lack of morality and unethical behavior
of managers, lobbyists, and oversight bodies are shocking.

News article:  {article}
Framing category:
```

Table 3: Zero-Shot prompt for in-context learning in LLMs for media framing classification.

Zero-Shot + CoT Prompt for Media Framing Classification

```
You are tasked with analyzing the **media framing** of a given article.
Media framing refers to how media select aspects of a perceived reality through
bounding and contextualization and make selected aspects more salient through
articulation and directionality to shape public perception.

Task
Analyze the article and identify the **dominant** and **sub-dominant** criteria frames.
Let's think step by step.  This media framing framework differentiates frame selection
through bounding (i.e., which aspects of a perceived reality are focused on or kept
out of focus) and contextualization (i.e., how certain aspects of a perceived reality
are considered within a setting).  Salience is captured through articulation (i.e., how
selected aspects of a perceived reality are communicated) and directionality (i.e., how
selected aspects are used to prime a specific interpretation).
Only include:
One **dominant frame**
One or more **sub-dominant frames** (if applicable)

Criteria Frames
1.  Factualization
 Bounding:  Focus on provision of key facts and info about the fraud.
 Contextualization:  Considered a consequence of internal failures leading
 to the company's demise.
 Articulation:  Sober, fact-based reporting; experts and insiders quoted;
 financial data used as evidence.
 Directionality:  Fraudulent statements from a once-successful company
 portrayed as a surprise.

2.  Consequences
 Bounding:  Focus on consequences for the company and key stakeholders.
 Contextualization:  Severe wrongdoing imposes financial and reputational
 harm on stakeholders.
 Articulation:  Emotional, dramatic manner; personal loss stories used.
 Directionality:  Extensive financial and reputational losses; victims'
 losses depicted as disastrous.

3.  Moral evaluation
 Bounding:  Focus on the morality of key stakeholders.
 Contextualization:  Fraud is a consequence of unethical behavior.
 Articulation:  Judgmental manner; rhetorical questions, sarcasm, exclamations.
 Directionality:  Lack of morality of managers, lobbyists, oversight bodies
 judged as shocking or outrageous.

4.  Failure attribution
 Bounding:  Focus on malpractice of oversight bodies and auditors.
 Contextualization:  Fraud is due to lack of appropriate action by oversight bodies
 and auditors.
 Articulation:  Clarifying, incriminating manner; rhetorical questions, sarcasm.
 Directionality:  Oversight bodies defined as main problem in not preventing fraud.

5.  Treatment recommendation
 Bounding:  Focus on ramifications for oversight bodies and auditors.
 Contextualization:  Fraud is a consequence of missing reforms of oversight and
 auditing.
 Articulation:  Forward-looking, action-oriented approach; experts calling for reforms.
 Directionality:  Strengthening oversight powers to prevent future fraud is recommended.

6.  Personification
 Bounding:  Focus on individual top managers responsible for the fraud.
 Contextualization:  Fraud is criminal behavior of the company's top managers.
 Articulation:  Sensational, entertaining; references to popular culture.
 Directionality:  Criminality and depravity of top managers are mystified and cinematic.

Output Format
Return a JSON object containing:
- `Dominant_frame`:  The name of the dominant frame (e.g., "Factualization")
- `Sub_dominant_frames`:  A list of sub-dominant frame names (e.g., ["Consequences",
"Personification"])
  - `Quotations`:  A dictionary where the key is the frame name and the value is a list of
quote-reason pairs.  Each quote should be a direct excerpt from the article, and each reason should
explain why it reflects the frame.
Article:
{article}
```

Table 4: Zero-shot + CoT prompt for in-context learning in LLMs for media framing classification

Zero-Shot Voting (Bounding Criteria) Prompt for Media Framing Classification

```
You are tasked with analyzing the **media framing** of a given article.
Media framing refers to how media select aspects of a perceived reality
through bounding and contextualization and make selected aspects more salient
through articulation and directionality to shape public perception.

Task
Analyze the article and identify the **dominant** and **sub-dominant** criteria frames
of selection through **bounding** (i.e., which aspects of a perceived reality are focused
on or kept out of focus, thereby regulating what the audience sees).

Only include:
- One **dominant frame**
- One or more **sub-dominant frames** (if applicable)

Criteria Frames of Selection Through Bounding
1.  Factualization:  Focus is placed on the provision of key facts and
 information about the fraud.
2.  Consequences:  Focus is placed on the consequences for the
 company and key stakeholders.
3.  Moral evaluation:  Focus is placed on the morality of key stakeholders.
4.  Failure attribution:  Focus is placed on the malpractice of oversight bodies
 and the auditing profession.
5.  Treatment recommendation:  Focus is placed on the ramifications
 for oversight bodies and the auditing profession.
6.  Personification:  Focus is placed on individual top managers responsible
 for the accounting fraud.

Output Format
Return a JSON object containing:
- Dominant_frame:  The name of the dominant frame (e.g., "Factualization")
- Sub_frames:  A list of sub-dominant frame names
 (e.g., ["Consequences", "Personification"])
- Quotations:  A dictionary where the key is the frame name and the value
 is a list of quote-reason pairs.  Each quote should be a direct excerpt
 from the article, and each reason should explain why it reflects the frame.

Article:
{article}
```

Table 5: Zero-Shot Voting (Bounding Criteria) prompt for in-context learning in LLMs for media framing classification.

Zero-Shot Top-K Prompt for Media Framing Classification

```
You are a media framing classifier.
You classify news articles into six categories: failure attribution,
treatment recommendation, factualization, consequences,
personification, and moral evaluation.
For each news article, please classify it into the top 3 most relevant
framing categories, ordered by relevance.
Output the result in JSON format with the following structure:
{
 "top_3_frames": [
 "frame_1",
 "frame_2",
 "frame_3"
 ]
}

Here are the definitions for each category:
1) failure attribution
 bounding: Focus is placed on the malpractice of oversight bodies
 and the auditing profession.
 contextuationlization: The selected aspects of accounting fraud
 are considered as a consequence of the lack of appropriate
 action of oversight bodies and auditors in preventing
 accounting fraud.
 articulation: News is reported in a clarifying, incriminating
 manner. Stylistic devices such as rhetorical questions, sarcasm,
 capitalization, and exclamation marks illustrate malpractice.
 directionality: Oversight bodies and auditors failed to oversee
 and audit the defrauding company and are defined as the main
 problem in not preventing the fraud.

2) treatment recommendation
 bounding: Focus is placed on the ramifications for oversight
 bodies and the auditing profession.
 contextuationlization: Accounting fraud is a consequence of a
 lack of appropriate rules and reforms of oversight and auditing.
 articulation: News is reported in a forward-looking, action-oriented
 manner. Experts who call for reforms are quoted, and conditional
 sentences are used to express recommendations for change.
 directionality: The authority, power, and resources of oversight
 bodies and auditors are recommended to be increased to prevent
 future accounting fraud.

3) factualization
 bounding: Focus is placed on the provision of key facts and
 information about the fraud.
 contextuationlization: Fraud is considered a consequence of
 internal failures leading to the company's demise.
 articulation: News is reported in a sober, fact-based manner;
 case experts and insiders are quoted to provide insights and
 leave an impression of objectivity. Financial numbers
 are used as reference points.
 directionality: The emergence of fraudulent financial statements
 of a once-successful company with a positive reputation
 is portrayed as a surprise.

4) consequences
 bounding: Focus is placed on the consequences for the company
 and key stakeholders.
 contextuationlization: Fraud is considered severe wrongdoing that
 imposed financial and reputational consequences on stakeholders
 of the defrauding company.
 articulation: News is reported in an emotional, dramatic manner;
 stakeholder stories of individual losses illustrate desperation
 and emotional outrage.
 directionality: The financial and reputational losses of the
 defrauding company, oversight bodies, and auditors are extensive;
 victims' losses are disastrous.

5) personification
 bounding: Focus is placed on individual top managers responsible
 for the accounting fraud.
 contextuationlization: Fraud is considered the criminal behavior
 of top managers.
 articulation: News is reported in a sensational, entertaining manner;
 references to thriller movies, secret agents, or spies express fascination.
 directionality: The criminality and depravity of top managers are
 portrayed as mystifying and cinematic.

6) moral evaluation
 bounding: Focus is placed on the morality of key stakeholders.
 contextuationlization: Fraud is considered a consequence of the
 lack of morality and unethical behavior.
 articulation: News is reported in a judgmental manner; rhetorical
 questions, sarcasm, and exclamations are used.
 directionality: The lack of morality and unethical behavior of managers,
 lobbyists, and oversight bodies are depicted as shocking and outrageous.

News article: {article}
```


Table 6: Zero-shot top-k prompt for in-context learning in LLMs for media framing classification.

One-Shot Prompt for Media Framing Classification

```
You are a media framing classifier.
You classify news article into six categories: failure attribution, treatment
recommendation, factualization, consequences, personification, and moral evaluation.

1) failure attribution
 bounding: Focus is placed on the malpractice of oversight bodies and the
 auditing profession.
 contextuationlization: The selected aspects of accounting fraud are considered as
 a consequence of the lack of appropriate action of oversight bodies and auditors.
 articulation: News is reported in a clarifying, incriminating manner; rhetorical
 questions, sarcasm, capitalization are used to illustrate malpractice.
 directionality: Oversight bodies and auditors are defined as the main problem
 in not preventing fraud.

2) treatment recommendation
 bounding: Focus is placed on the ramifications for oversight bodies and
 the auditing profession.
 contextuationlization: Fraud is considered a consequence of a lack of
 appropriate rules and reforms of oversight and auditing.
 articulation: News is forward-looking, action-oriented; experts calling for reforms
 and conditional sentences express recommendations for change.
 directionality: Authority, power, and resources of oversight bodies and auditors
 are recommended to be increased to prevent future fraud.

3) factualization
 bounding: Focus is placed on the provision of key facts and information
 about the fraud.
 contextuationlization: Fraud is considered a consequence of internal failures that
 led to the company's demise.
 articulation: News is sober and fact-based; case experts and insiders are quoted to
 convey objectivity. Financial numbers serve as reference points for evidence.
 directionality: Fraudulent financial statements from a formerly successful company
 are depicted as a surprise.

4) consequences
 bounding: Focus is placed on the consequences for the company and key
 stakeholders.
 contextuationlization: Fraud is considered severe wrongdoing imposing financial
 and reputational consequences on key stakeholders.
 articulation: News is emotional, dramatic; stakeholder stories of individual losses
 illustrate desperation. Aggregated numbers validate the event's impact.
 directionality: Financial and reputational losses of the defrauding company,
 oversight bodies, and auditors are extensive; victims' losses are disastrous.

5) personification
 bounding: Focus is placed on individual top managers responsible for the fraud.
 contextuationlization: Fraud is considered the criminal behavior of the top
 managers.
 articulation: News is sensational, entertaining, referencing popular culture
 (thrillers, spies) to express fascination.
 directionality: Criminality and depravity of top managers are portrayed as
 mystifying and cinematic.

6) moral evaluation
 bounding: Focus is placed on the morality of key stakeholders.
 contextuationlization: Fraud is considered a consequence of the lack of morality
 and unethical behavior.
 articulation: News is reported in a judgmental manner; rhetorical questions, sarcasm,
 and exclamations express moral evaluation.
 directionality: Lack of morality and unethical behavior of managers, lobbyists, and
 oversight bodies are portrayed as shocking and outrageous.

Article: Handelsblatt print: Nr. 161 vom 21.08.2020 Seite 030 / Finanzen
QUARTALSZAHLEN Der Wirecard-Effekt...Millionen Euro Nettoumsatz machte Adyen...
Frame: failure attribution

Article: Handelsblatt print: Nr. 161 vom 21.08.2020 Seite 030 / Finanzen
QUARTALSZAHLEN Der Wirecard-Effekt...Die Ermittler gehen davon aus, dass es
Mitwisser in der deutschen Unternehmenszentrale gab...
Frame: treatment recommendation

Article: Handelsblatt online vom 01.07.2020 Finanzen Banken + Versicherungen...
Beim insolventen Zahlungsdienstleister Wirecard ist am Mittwoch die
Staatsanwaltschaft angerückt...
Frame: factualization

Article: Handelsblatt print: Nr. 145 vom 30.07.2020 Seite 013 / Politik
BILANZSKANDAL Das Wirecard-Netzwerk...
Frame: consequences

Article: Handelsblatt print: Nr. 167 vom 31.08.2020 Seite 006 / Politik
FLUCHTHILFE Geheimsache Wirecard...Oppositionspolitiker hegen Zweifel...
Frame: personification

Article: Handelsblatt print: Nr. 193 vom 06.10.2020 Seite 030 / Finanzen
LEHREN AUS DEM FALL WIRECARD...Prozent könnte der MDax...
Frame: moral evaluation
```

Table 7: One-Shot prompt for in-context learning in LLMs for media framing classification.

Part 1: News Article (Gold Frame: Consequences)

Die gesamte deutsche Finanzelite hat sich mit dem Fall Wirecard blamiert Wirtschaftsprüfer, Aufseher, Ratingagenturen: Sie alle haben trotz massiver Zweifel an Wirecard weggesehen. Diese kollektive Ignoranz wird zunehmend zu einer Gefahr für den Standort Deutschland. Kommentar von Meike Schreiber Zwei Stunden und vierzehn Minuten wird das Thema sorgsam umschifft, als sich am Montag die Finanzelite in Frankfurt zur ersten leibhaftigen Branchenkonferenz seit Beginn der Corona-Krise trifft. Auf dem Podium geht es um die Pandemie, Nachhaltigkeit, Home-Office-Erfahrungen, solche Dinge. Und Wirecard? Spricht erst mal niemand an. Dabei ist seit diesem Montag klar, dass sich Wirecard zum wohl schlimmsten Bilanzskandal in Deutschland auswächst. Dass Wirecard eben nicht das Opfer gieriger Spekulanten ist, als was es Vorstandschef Markus Braun und sein Umfeld es noch bis dieses Wochenende verkauft haben. **Und damit steht fest, dass sich das ganze deutsche Finanzsystem blamiert hat: Die Aufseher haben versagt, die Ratingagenturen, die Wirtschaftsprüfer, die Banken und Fondsgesellschaften, welche Milliarden an Privatanlegergeldern verbrannt haben, ohne nachzufragen**- und das, obwohl Journalisten, allen voran jene der Financial Times, bereits seit 2015 gut begründete, massive Zweifel am Geschäftsmodell von Wirecard aufgebracht hatten. Immerhin, Felix Hufeld, Chef der deutschen Finanzaufsicht Bafin, bricht auf der Konferenz nach mehr als zwei Stunden das Schweigen zu Wirecard. Der Fall sei eine "Schande" für das Land; die Lage "entsetzlich", die Kritik an der Rolle der Aufsichtsbehörden - inklusive der Bafin - absolut verständlich. Hufeld sprach sogar Journalisten, Analysten und Hedge-Fonds seine Anerkennung aus, welche tief gegraben und die richtigen Fragen gestellt hätten. Es ist daher wenigstens mal ein Lichtblick in dem Desaster, dass sich zumindest Hufeld seiner Verantwortung und Fehler bekennt - anders als sein Dienstherr Bundesfinanzminister Olaf Scholz, der auf derselben Konferenz doch tatsächlich herumdruckst, die Aufseher hätten mit Blick auf Wirecard "sehr hart" gearbeitet. Scholz steht damit vielen in der deutschen Business-Elite in nichts nach. Wenn etwas schief geht, übernimmt man dort selten Verantwortung. **Solche Fälle haben inzwischen Folgen für den Standort Deutschland: Nach all den Skandalen bei Volkswagen, Deutscher Bank, Bayer nun also auch noch Wirecard! Internationale Investoren sind zunehmend misstrauisch mit Blick auf deutsche Unternehmen.** Deutschland erscheint als ein Land, in dem sich befreundete Vorstände, Politiker, Aufseher, Prüfer, zuweilen auch Gewerkschafter und Strafverfolger gegenseitig schützen. Ein Land, in dem Kritiker so lange als dubiose Konzernschrecks abgeschüttelt werden, bis es nicht mehr anders geht. Deutschland muss endlich lernen, dass eine starke und sowohl personell als auch vom gesetzlichen Rahmen her gut ausgestattete Finanzaufsicht ein Standortvorteil ist: Amerikas Banken sind auch deshalb so stark, weil die Aufseher dort viel härter durchgreifen, weil Fehlverhalten schnell und brutal geahndet wird. Erst im Mai 2020 nahm die Bafin Wirecard in den Blick Tatsächlich hatte sich die Aufsicht bei Wirecard lange Zeit redlich bemüht wegzuschauen, ... Und natürlich hätte sie, wie Hufeld selbst einräumt, die Aufsicht über den gesamten Zahlungsverkehrs-Konzern, und nicht nur die Bank, aktiv an sich ziehen müssen. Das alles muss nun schleunigst besser werden.

Part 2: GPT-4o mini (Chain-of-Thought & Zero-shot) Analysis Output

Dominant Frame: Failure attribution

Quote: Die gesamte deutsche Finanzelite hat sich mit dem Fall Wirecard blamiert
English translation: The entire German financial elite has embarrassed itself with the Wirecard case
Reason: This statement highlights the collective failure of oversight bodies and auditors, attributing the malpractices directly to their negligence.

Quote: Die Aufseher haben versagt, die Ratingagenturen, die Wirtschaftsprüfer, die Banken und Fondsgesellschaften... ohne nachzufragen
English translation: The supervisors have failed, the rating agencies, the auditors, the banks and fund companies... without asking
Reason: This articulates the specific failures of oversight bodies, assigning responsibility for the fraud to those who were supposed to protect investors.

Quote: Die Aufsicht bei Wirecard lange Zeit redlich bemüht hat, wegzuschauen
English translation: The supervisory authority at Wirecard has long tried to look the other way
Reason: This emphasizes the inaction of regulatory bodies, framing them as complicit in the fraud due to their failure to act on clear warning signs.

Sub-dominant Frame: Moral evaluation

Quote: Der Fall sei eine "Schande" für das Land
English translation: It is a "shame" for the country
Reason: This quote employs a moral judgment on the situation, expressing outrage at the failure of responsibility among financial authorities.

Quote: Wenn etwas schief geht, übernimmt man dort selten Verantwortung
English translation: When something goes wrong, people rarely take responsibility
Reason: This reflects a moral critique of the corporate and regulatory culture in Germany, highlighting a perceived lack of accountability.

Sub-dominant Frame: Consequences

Quote: Internationale Investoren sind zunehmend misstrauisch mit Blick auf deutsche Unternehmen
English translation: International investors are increasingly suspicious of German companies
Reason: This points to the broader repercussions of the Wirecard scandal, framing it as damaging not only to the company but also to Germany's international business reputation.

Quote: Deutschland erscheint als ein Land, in dem sich befreundete Vorstände, Politiker, Aufseher... gegenseitig schützen
English translation: Germany appears to be a country in which friendly executives, politicians, supervisors... protect each other
Reason: This articulates the negative consequences for the integrity of the financial system in Germany, presenting it as a community that shields its members from accountability.

Table 8: Part 1 presents a Wirecard news article annotated with the gold standard frame (Consequences), where bold text highlights quotes reflecting the Consequences frame as identified by the dataset owner. Part 2 provides the GPT-4o mini analysis output using a Chain-of-Thought Zero-Shot approach, with color-coded quotes indicating the model's frame interpretations: purple for Failure Attribution, orange for Consequences, and blue for Moral Evaluation.

Part 1: Prompt for Data Augmentation (Moral Evaluation Frame)

Instructions:
You are a journalist tasked with writing a news article about Wirecard. Your article must employ the **moral evaluation** frame of media framing. Media framing is the process of selecting and emphasizing specific aspects of an event to shape public perception.

1. Frame Details:
- **Moral evaluation Frame:**
- News is reported in a judgmental manner. Stylistic devices such as rhetorical questions, sarcasm, and exclamations are used to express moral evaluation.
- News articles can become accusatory pieces by highlighting a lack of morality of individual actors and invoke underlying cultural norms to create emotional resonance with readers.

2. Key Facts and Information:
- **Fraud Details:**
- Wirecard faked multi-billion euro profits, particularly through its Middle East subsidiary.
- The public prosecutor's office is investigating several managers for fraud, breach of trust, accounting falsification, and market manipulation.
- EY is under scrutiny for failing to detect Wirecard's fraudulent financial statements.
- Regulatory oversight failures contributed to the prolonged fraud.

- **Event Timeline:**
- **February 2019:** Olaf Scholz learned about Bafin's investigation into Wirecard for market manipulation.
- **June 21, 2020:** Jan Marsalek mentioned in a Telegram chat that he didn't deny the allegations.
- **July 17, 2020:** The first Wirecard manager confessed to fraud in Munich.
- **2020:** Wirecard declared insolvency after a €1.9 billion hole was discovered in its balance sheet.

- **Key Individuals:**
- **Jan Marsalek (Fugitive ex-board member):** On the run since the scandal broke, admitted in private chats that he did not deny the allegations. Bragged about intelligence contacts.
- **Markus Braun (Former CEO):** Resigned and surrendered to authorities.
- **Nicolas Frühsorger (Lawyer):** Represents the former head of Wirecard's Cardsystems Middle East subsidiary, who admitted involvement.
- **Olaf Scholz (Finance Minister, SPD):** Knew about Wirecard irregularities since February 2019.
- **Kay Scheller (President of the Federal Audit Office):** Criticized the lack of regulatory oversight.

- **Institutions Involved:**
- **Wirecard AG:** German fintech company that collapsed due to a €1.9 billion balance sheet hole.
- **Cardsystems Middle East (Wirecard subsidiary):** Played a key role in the fraudulent Asian business.
- **Bafin (German Financial Regulator):** Faced criticism for inadequate oversight.
- **Federal Ministry of Finance:** Under scrutiny for its handling of Wirecard allegations.
- **German Audit Office for Accounting (DPR):** Responsible for checking company balance sheets, not Bafin.
- **EY (Ernst & Young):** Audited Wirecard since 2009 but failed to detect fraud.

3. Article Requirements:
- Write a compelling and professional news article that adheres to the **moral evaluation** frame.
- Ensure the selected frame is reflected in the tone, language, and focus of the article.
- Use the provided key facts and information to support the narrative.
- Structure the article with a clear headline, lead, body, and conclusion.
- Highlight the moral implications of the Wirecard case for clarity and impact.

Part 2: GPT-4o mini Generated Article (Moral Evaluation Frame)

The Corporate Scandal That Shook Germany: Wirecard's Fall from Grace

In a shocking display of corporate deception, Wirecard AG—the once-celebrated German fintech giant—has proven that greed knows no boundaries. How did a company touted as a national success story end up entangled in a web of lies, deceit, and a staggering €1.9 billion fraud? It appears the moral compass at Wirecard was not just off course but completely shattered.

The scandal began to unravel when reports surfaced revealing that Wirecard had been inflating profits through its Middle East subsidiary—a revelation that has left investors and the public reeling in disbelief. How could such blatant manipulation go undetected for so long? The German financial regulator Bafin is now under a microscope for its glaring failures in oversight. It seems the watchdog was more like a sleeping dog, ignoring the mounting evidence of wrongdoing that Olaf Scholz, the country's Finance Minister, had been made aware of as early as February 2019!

As if this wasn't enough, the former CEO, Markus Braun, whose name was once synonymous with innovation, has surrendered to authorities, but that hardly assuages the anger of those deceived by his charisma. Meanwhile, Jan Marsalek, the fugitive ex-board member who brazenly admitted via a Telegram chat that he did not deny the allegations, has slipped through the cracks, making a mockery of the justice system. One can't help but ask, where is the accountability?

And let's not forget Ernst & Young (EY), the auditing firm that had been signing off on Wirecard's accounts for over a decade. Their failure to detect years of fraud raises questions about their ethical practices. How many auditors have turned a blind eye, prioritizing profits over principles? It's a scandal that doesn't just speak to Wirecard's malfeasance but echoes a deeper issue within corporate culture in Germany.

As investigations swirl and more admissions of guilt come to light, Wirecard serves as a grim reminder of how easily morality can be cast aside for financial gain. The Federal Audit Office has now criticized the lack of regulatory oversight that allowed this colossal fraud to fester. How much longer will the system protect those who prey on it? The time has come for a reckoning—not just for Wirecard, but for all who turn a blind eye to deceit and dishonesty in the pursuit of power and profit.

In this derailing saga, one thing is crystal clear: trust once earned can be lost in an instant, and Wirecard's legacy will be remembered as a cautionary tale of greed, negligence, and moral failure.

Table 9: Part 1 provides a prompt for generating a Wirecard news article using the moral evaluation frame, based on key facts and events. Part 2 shows the GPT-4o mini-generated article employing the moral evaluation frame, highlighting the ethical failures and systemic issues surrounding the Wirecard scandal.