UPPSALA UNIVERSITET

UPTEC STS 26046

Examensarbete 30 hp

Juni 2026

Narrating Market Movements: An LLM-Based Approach to Explaining Stock Price Movements from Financial News

An expert-evaluated study of Swedish listed stocks

Maja Danielsson Kadestam

Civilingenjörsprogrammet i System och Teknik i Samhället

UPPSALA
UNIVERSITET

Narrating Market Movements: An LLM-Based Approach to Explaining Stock Price Movements from Financial News

Maja Danielsson Kadestam

Abstract

Fluctuations in the financial markets are fundamentally driven by information, and financial news articles constitute a primary mechanism through which such information is distributed to market participants. Whereas previous research on applying Natural Language Processing (NLP) to link financial news to market movements mainly focuses on predictive tasks, this thesis takes an explanatory approach by considering a price movement and related articles *after* it occurred. This thesis examined the capability of a Large Language Model (LLM) to generate a narrative explanation of significant market movements of Swedish listed stocks using financial news articles from Dagens Industri (DI), published between the 2nd of January 2026 and the 20th of April 2026. The methodological approach implemented an algorithm in three steps. Firstly, significant market movements were identified using a volatility-normalised z-score. Secondly, financial news articles were linked to Swedish listed stocks through Named Entity Linking (NEL) performed by Claude Haiku 4.5. Thirdly, Claude Sonnet 4.6 was prompted with the significant market movement and its matched articles and (i) score the articles relevance, (ii) generate a narrative explanation based on the relevant articles, and (iii) assign a verbalised confidence label to signal the quality of the explanation. The results were lastly evaluated through a structured form by financial professionals. The results demonstrated that the quality of the narrative explanations generated by the general-purpose LLM to a high degree aligned with the opinions of the financial professionals. However, the LLM tended to overestimate the relevance of the matched set of articles to the market movement.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala/Visby

Handledare: Daniel Roos Ämnesgranskare: Yi Wang

Examinator: Jens Engström

Table of Contents

Populärvetenskaplig sammanfattning

Att förstå och förklara prisrörelser på aktiemarknaden har länge varit centralt för både institutionella investerare och privatsparare. Ett grundläggande antagande är att aktiemarknadspriser speglar den information som finns tillgänglig om en aktie och dess framtida utveckling, såsom kvartals- och årsrapporter, prognoser och regulatoriska beslut. Tillgången till och distributionen av sådan information kan därför påverka investerares förväntningar och därmed även aktiekurser och aktivitet på finansmarknaden. Dagens Industri (DI) är Nordens största affärstidning och utgör en viktig källa till finansiell och företagsrelaterad information för en stor del av aktiva investerare på den svenska börsen, och man kan därför anta att den information som förmedlas genom de nyheter som DI publicerar har en påverkan på den svenska aktiemarknaden och dess rörelser.

Relationen mellan nyheter och aktiemarknaden har studerats ingående inom den akademiska litteraturen. Språkteknologi (*eng. Natural Language Processing, NLP*), som kombinerar maskininlärning med lingvistik, utgör ett område inom vilket sådana analyser är vanliga. Bland annat har man studerat hur sentiment i nyhetsartiklar, det vill säga *tonen* som en text förmedlar, påverkar aktiepriser. Metoderna inom traditionell språkteknik har dock haft stora begränsningar med avseende på den kontextuella betydelsen av ord och språk. Den senaste tidens snabba utveckling inom generativ AI har därför öppnat upp för stora möjligheter inom fältet, och det är precis mot denna bakgrund som arbetet tar sin utgångspunkt.

Arbetet har utförts i samarbete med Njorda, en liten fintech-startup, som noterat ett intresse bland privatsparare för att förstå marknadsrörelser med hjälp av nyheter. Arbetet har därför kretsat kring hur man kan använda generativ AI för att generera kvalitativa förklaringar av signifikanta prisrörelser på den svenska börsen med nyhetsartiklar från DI som underlag. Den förklarande infallsvinkeln skiljer arbetet från tidigare forskning, som framför allt fokuserat på att *förutsäga* prisrörelser.

Trots de stora möjligheter som generativ AI medför finns det även betydande begränsningar och utmaningar. Det som kallas för hallucinationer, att AI genererar felaktiga påståenden, är särskilt problematiskt i en kontext där resultat kan komma att påverka finansiellt beslutsfattande. Vidare är kommersiella AI-modeller, som ChatGPT och Claude, closed-source, vilket innebär att användare inte har tillgång till den teknik som ligger bakom de genererade svaren. Generativ AI kan därmed betraktas som så kallade svarta lådor, vilket medför begränsningar när det kommer till tillit och transparens. Ett sätt att motarbeta detta är en teknik som kallas för verbaliserad konfidens (*eng. verbalised confidence*) som går ut på att synliggöra sannolikheter från den bakomliggande modellen i det genererade svaret genom att prompta modellen att ange en säkerhet i samband med sitt svar.

Mot denna bakgrund implementerades en algoritm i tre steg, där nyhetsartiklar från DI och prisdata från 70 svenska aktier under fyra månaders tid användes som dataunderlag. Inledningsvis identifierades signifikanta prisrörelser med en statistisk metod som tog

hänsyn till aktiens volatilitet under perioden. Därefter parades respektive nyhet samman med den eller de aktier som artikeln berörde med hjälp av den stora språkmodellen (*eng. Large Language Model, LLM*) Claude Haiku 4.5. Därefter sammanställdes ett dataset bestående av signifikanta prisrörelser och de relevanta nyhetsartiklar som hade publicerats under veckan före respektive prisrörelse. Datasetet användes sedan som indata i algoritmens sista steg, där kvalitativa förklaringar genererades med hjälp av den stora språkmodellen Claude Sonnet 4.6. För att utvärdera kvaliteten på de genererade förklaringarna användes två metoder: modellens verbaliserade konfidens, som angavs av modellen i samband med varje förklaring, samt expertintervjuer med tre personer som alla hade yrkeserfarenhet inom finanssektorn.

Resultaten visade på en korrelation mellan AI-modellens verbaliserade konfidens och experternas bedömning av förklaringarnas kvalitet. Förklaringar som modellen bedömde med hög säkerhet fick med andra ord även högre kvalitetsbedömningar av experterna. Däremot visade resultaten att AI-modellen tenderade att överskatta nyhetsartiklarnas relevans till den signifikanta prisrörelsen.

Arbetet och resultaten öppnar för flera intressanta vägar framåt för framtida studier. Många av de signifikanta prisrörelser som identifierades inledningsvis saknade förklaring bland de slutliga resultaten. En möjlig väg framåt vore därför att inkludera fler typer av finansiella dokument i det underliggande materialet, som mer direkta finansiella källor såsom kvartals- och årsrapporter, men även bredare makroekonomiska nyheter som påverkar aktiepriser indirekt. Vidare finns med generativ AI större möjligheter för ämnesanalys, en metod inom språkteknologi för att identifiera övergripande ämnen i textdokument, som hade kunnat tillföra djupare insikter om hur olika ämnen påverkar prisrörelser på aktiemarknaden.

List of abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
DI	Dagens Industri
EMH	Efficient Market Hypothesis
GPT	Generative Pretrained Transformer
ISIN	International Securities Identification Number
JSON	JavaScript Object Notation
LLM	Large Language Model
NEL	Named Entity Linking
NLP	Natural Language Processing
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
XML	eXtensible Markup Language

List of financial terms

Asset	Anything of monetary value, in the context of this thesis, a financial asset, i.e. something that can be sold or bought on the financial market. Examples include equity stocks and mutual funds.
Retail investor	An individual that participates in the financial market by buying and selling financial assets.
Closing price	The final price at which a stock trades on a given trading day.
Intraday prices	The fluctuation of an asset price throughout the trading day.
Drift	In the EMH context, the pace of adjustment of prices after new information becomes available to market participants.
Large cap/mid cap/small cap	A classification of stocks by market capitalisation, where large cap stocks are worth at least $5 billion, mid cap between $1 billion and $5 billion and small cap less than $1 billion.
Bag-of-words	Text representation that disregards the order of words.
Heavy-tailed (distribution)	Statistical property of financial returns where extreme values appear more often than predicted under Gaussian assumptions.
Volatility	A measure of risk using the standard deviation of an asset return.
ISIN	A globally recognised unique identifier for a financial asset.
Investor/market sentiment	The emotion or attitude of investors towards specific assets or the financial market.
Sentiment analysis	Analysis of the opinion or emotion conveyed in a textual document.

List of terms related to NLP and LLMs

Hallucination	The tendency of LLMs to generate factually incorrect statements.
Chain-of-thought	A prompting technique that forces step-by-step reasoning.
Few-shot/Zero-shot/One-shot learning	In-context learning, "shots" refer to the number of input-output examples included in the prompt.
Fine-tuning	Further training of a pretrained model on domain-specific data.
In-context learning	Improving an LLM at a certain task via examples of input-output pairs within the prompt rather than weight updates.
Token	The smallest possible unit of a language model, and corresponds to words, sub-words or characters.
Temperature	An LLM parameter controlling output randomness.
Word embedding	A numerical vector representation of words capturing semantic relationships.
Transformer	The neural architecture underlying modern LLMs.
Knowledge cutoff	The date past which the LLM has no training data.
Corpus	A large collection of written text.
Topic analysis	Identifying and categorising broader themes in text.
RLHF	A technique used to train LLMs to conform outputs to human preferences.

1. Introduction

Fluctuations in financial markets are fundamentally driven by information. The way in which information is interpreted and incorporated into asset prices has therefore long been a central concern for economists, financial institutions, and investment professionals. At the same time, a growing number of individuals actively participate in financial markets (Euroclear Sweden, 2026) to manage their savings, investments, and pensions, increasing the demand for accessible explanations of why stock prices fluctuate. Understanding the relationship between information flows and market movements is thus of importance not only within academic finance but also for the broader society.

A foundation within classical financial theory is that stock prices reflect the information available to investors about an asset. According to the Efficient Market Hypothesis (EMH), prices adjust rapidly and efficiently as new information becomes available to market participants. However, Fama (1970) argues that markets in practice cannot live up to this ideal due to frictions such as unequal access to information, transaction costs, and disagreement among market participants on the implications of available information for an asset. Consequently, the process through which information becomes incorporated into asset prices is more complex than the assumptions of perfectly efficient markets suggest.

Within this process, financial news plays a particularly important role. News media can be said to reflect the state of the economy (Bybee et al., 2024) and constitute one of the primary mechanisms through which information is distributed to market participants and influence both institutional and retail investors. Historical research has demonstrated the causal relationship between news coverage and fluctuations in stock prices, trading volume, and investor behaviour (Tetlock, 2007; Engelberg and Parsons, 2011). However, not all news carries equal informational weight, as they reflect media and audience biases, speculative market sentiments and interpretive noise (Bybee et al., 2024). Beaudry and Portier (2014) further distinguish between news that convey fundamental information about an asset, and news that capture investor attention without affecting the underlying value of an asset. The former produces persistent effects on prices, whereas the latter causes temporary noise that fades as investors reevaluate their interpretations (Beaudry and Portier, 2014).

The technical opportunities for analysing the content of news articles have evolved along with recent developments within the field of Natural Language Processing (NLP), a subfield of machine learning concerned with the computational analysis of human language (Khurana et al., 2023). In particular, the emergence of Large Language Models (LLMs) has opened up new methodological possibilities within the financial domain, as they allow for more contextual interpretations of textual information, as opposed to previous approaches, which often relied on relatively simple techniques such as dictionary-based sentiment analysis. These advances are expected to significantly

influence financial analysis and investment decision-making processes for both institutional and retail investors (Du et al., 2025).

Previous research applying NLP and machine learning to financial markets has predominantly focused on predictive tasks, such as forecasting stock returns or market sentiment (Araci, 2019; Lopez-Lira and Tang, 2025). While predictive performance has received considerable attention, comparatively less emphasis has been placed on explanatory applications that aim to interpret and contextualise market movements *after* they occur. This thesis addresses that gap by applying a qualitative and explanatory approach. Rather than predicting future price changes, the study investigates whether LLMs can be used to explain observed stock market movements through the analysis of news articles published prior to those movements.

The project was conducted in collaboration with a small company in the financial technology sector and was motivated by a growing demand among retail investors for interpretable explanations of portfolio performance. While LLMs offer new opportunities for financial applications, their outputs are generated through processes that are not transparent to the end user and may include factually incorrect statements known as hallucinations (Du et al., 2025; Kadre et al., 2025). The use of LLMs to provide financial information, therefore, raises not only methodological questions about its capability of generating an explanation but also ethical questions regarding trust, accuracy, and the appropriate role of automated reasoning in personal investment decisions.

Commercial LLMs are typically closed-source and accessed through APIs, which means that no information about their parameters or the process through which their outputs are generated is available to the user and can therefore be argued to constitute black boxes (Xiong et al., 2024). In a financial decision-making context, where the generated outputs may influence investment decisions, a lack of transparency is particularly problematic. One mechanism to address this limitation is *verbalised confidence*, in which the model is prompted to express a confidence level as part of its output, thus surfacing a probability within the output of a closed-source model (Xiong et al., 2024). Previous research has demonstrated that verbalised confidence is consistent with other measures of output quality (Tian et al., 2023; Kadavath et al., 2022). To the author's knowledge, it remains unexplored whether this holds for an explanatory task within the financial domain.

1.1 Project Aim

The aim of this thesis is to examine the capability of a Large Language Model (LLM) to generate explanations of significant market movements of Swedish stocks using financial news articles from Dagens Industri (DI), the largest business newspaper in the Nordic region. Although previous research primarily focuses on predictive approaches, this thesis aims to examine the capabilities of LLMs to link financial news articles to stock prices *after* they occur. Further, the thesis aims to explore whether verbalised confidence accurately signals the quality of the output. To achieve the thesis aim, the method implements an algorithm that (i) identifies significant price movements through statistical anomaly detection, (ii) links them to financial news articles, and (iii) utilises an LLM to generate a narrative explanation and a verbalised confidence label for each movement. As no objective truth exists for which articles explain a given movement, the outputs are evaluated in relation to the judgement of financial professionals.

1.1.1 Research Questions

- **RQ1**: To what extent can a general-purpose LLM use financial news articles to generate a narrative explanation of a significant market movement?
- **RQ2**: Does an LLM's verbalised confidence in its explanations align with the explanation quality as judged by financial professionals?

2. Prerequisites: Financial time series, statistical anomalies and LLMs

This chapter aims to establish the theoretical and technical foundation for the thesis. The method implemented in Chapter 4 combines statistical anomaly detection on financial time series with LLM-based analysis of news articles, and therefore builds on three separate fields. Section 2.1 introduces financial time series and their statistical properties, and methods for detecting significant price movements. Section 2.2 introduces Natural Language Processing (NLP), Large Language Models (LLMs), and prompting techniques, along with known limitations. Section 2.2.3 describes how NLP and LLMs have historically been applied in the financial domain.

2.1 Financial Time Series and Price formations

A central component of the thesis method is the identification of significant price movements among Swedish listed stocks, which the LLM is subsequently asked to explain. This requires a theoretical understanding of how prices form and respond to information, how stock returns are defined and behave statistically, and how anomalies in such returns can be identified. Section 2.1.1 outlines the Efficient Market Hypothesis as a theoretical basis for interpreting price responses to news. Section 2.1.2 introduces the definitions of returns used throughout the analysis. Section 2.1.3 discusses volatility, and section 2.1.4 presents the empirical properties of financial time series, and section 2.1.5 presents statistical methods for anomaly detection.

2.1.1 Price formations

The Efficient Market Hypothesis (EMH) is a theoretical basis for the assumption that market prices respond to new information, and serves as a foundation for this thesis (Fama, 1970). The financial market essentially serves as a mechanism for the distribution of ownership of financial capital, and prices are intended to signal the allocation of this distribution (Fama, 1970). This implies that prices in an efficient market reflect all available information about an asset (Fama, 1970). Market efficiency thus serves as a measure of the speed at which prices reflect new information, where slow adjustments are commonly referred to as “drift” (Fama, 1970). Fama (1970) further resonates around market conditions that may help or hinder such market efficiency. He argues that perfect market efficiency only arises in a state where (i) all market participants have complete access to all available information, (ii) there are no transaction fees, and (iii) all participants agree on the implications that the available information may have on an instrument (Fama, 1970). All three factors, unless perfect, create friction within the efficiency of a market (Fama, 1970).

Fama (1970) categorised historical attempts to evaluate or test market efficiency into three main forms. *Weak form* tests refer to the assumption that prices merely reflect the historical price movement of a single asset. This implies that the expected return of an

asset is fully explained by its historical value, and price movements are assumed to be independent and normally distributed (Fama, 1970). The *semi-strong form* of evaluation incorporates all publicly available information about an asset, such as press releases or news articles (Fama, 1970). The *strong form* assumes all available information about an asset is reflected within its price, and especially limited information such as monopolistic, insider information (Fama, 1970). The *strong form* assumption would, in practice, imply that prices move instantaneously as new information becomes available, an assumption rejected by Fama (1970) in his original paper, and thus the strong form tests only serve a theoretical ideal for which other tests are benchmarked (Fama, 1970).

In his consecutive article, Fama (1991) suggests a foundational limitation of the Efficient Market Hypothesis, which he refers to as the *joint hypothesis problem*. This limitation arises due to anomalous returns being evaluated in relation to the expected return as predicted by a chosen asset pricing model, which determines what is considered a "normal" return. Thus, he argues, it is not possible to know whether anomalous returns are truly anomalous or simply reflect a poorly chosen asset pricing model (Fama, 1991). The *joint hypothesis problem,* therefore, implies that all tests within the Efficient Market Hypothesis are always restricted to the chosen definition of an expected return (Fama, 1991).

2.1.2 Financial Time Series Analysis

Financial time series consist of asset returns over time and demonstrate several characteristics that distinguish them from other time series in analysis (Tsay, 2010, p. 1). Returns are generally preferred over prices when examining financial time series, because they are independent of the asset's price and have beneficial statistical properties (Tsay, 2010, p. 2). Let P_t denote the price at time t, then the *one-period simple gross return* is defined as

$$1 + R_t = \frac{P_t}{P_{t-1}},$$

which can equivalently be written as $P_t = P_{t-1}(1 + R_t)$ (Tsay, 2010, p. 3). The corresponding *one-period simple net return* is then defined as

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}},$$

and the *k-period simple net return* (also referred to as compound return) between t − k and t is given by

$$R_t[k] = \frac{P_t - P_{t-k}}{P_{t-k}}.$$

The natural logarithm of the simple gross return is the *log return*, defined as

$$r_t = \ln(1 + R_t) = \ln \frac{P_t}{P_{t-1}} = p_t - p_{t-1},$$

where $p_t = \ln(P_t)$ and is also referred to as continuously compounded returns (Tsay, 2010, p. 5). Log returns are preferred over simple net returns due to their additive and symmetric characteristics (Tsay, 2010, p. 5).

2.1.3 Volatility

Volatility is a way to describe the inherent uncertainty of the expected returns and is a central component of financial time series (Tsay, 2010, p. 20). Volatility can be defined as the standard deviation of asset returns conditioned on the historical returns up to t and is commonly denoted by σ_t (Tsay, 2010, p. 98). As demonstrated by several authors, volatility tends to cluster over time (Engle, 1982; Tsay, 2010, p. 20; Cont, 2001), meaning high volatility tends to follow periods of high volatility.

2.1.4 Statistical properties of financial time series

Financial time series share some statistical properties that have implications for statistical analysis (Cont, 2001). Firstly, financial time series are heavy-tailed, which means extreme values will appear more often in financial time-series data than in what is expected under a Gaussian distribution (Tsay, 2010, p. 14; Cont, 2001). Additionally, due to what's called the *leverage effect*, returns and volatility of an asset are negatively correlated, meaning negative asset return correlate with increased volatility, whereas positive asset returns correlate with decreased volatility (Cont, 2001; Tsay, 2010, p. 99). Lastly, volatility is positively correlated with trading volume (Cont, 2001). The inherent uncertainty of volatility is the main reason financial time series analysis differs from traditional time series analysis (Tsay, 2010, p.1).

2.1.5 Anomaly Detection Methods

Anomaly detection, also referred to as outlier detection, is a field within statistical theory concerned with identifying data instances that might be considered abnormal (Tan et al., 2020, p. 1115). Anomalies are commonly distinguished from noise, which is the random variation within a time series, and is usually not of interest as an object of analysis (Tan et al., 2020, p. 1115). Anomalies can be considered abnormal in both a global and local context (Tan et al., 2020, p. 1125). For instance, a market movement considered unusually high for a low-risk mutual fund might be an anomaly in the local context of the specific fund, whereas the same movement would be considered normal in the context of a high-risk stock.

Statistical approaches for anomaly detection utilise probability distributions to model the normal class, and can be either parametric or non-parametric (Tan et al., 2020, p. 1127). A common parametric approach assumes a Gaussian distribution and requires estimating

the mean and standard deviation, from which the z-score can be derived (Tan et al., 2020, p. 1129). The mean and standard deviation are estimated from training data by using the sample mean and sample standard deviation (Tan et al., 2020, p. 1129). Statistical methods are beneficial when there is prior knowledge about the data, and they enable the use of confidence intervals to set anomaly thresholds (Tan et al., 2020, p. 1140-1141). However, if the wrong distribution is chosen, normal data is likely to be falsely identified as abnormal (Tan et al., 2020, p. 1141). Applied to financial returns, the Gaussian assumption underlying the z-score has some limitations, which are further discussed in relation to the methodological choices.

2.2 Natural Language Processing and Large Language Models

Natural Language Processing (NLP) is a scientific field that combines linguistics and machine learning to analyse natural language and textual data using computers (Kadre et al., Ch. 1, 2025; Nigam et al., 2000). The recent advancement in generative AI primarily stems from research within NLP and improved computational capabilities (Kadre et al., 2025). NLP has been widely applied in finance, in fields such as intelligent investment research and risk assessment (Gao et al., 2021), and its applications within the financial domain are further discussed in section 2.2.3. Within the field of NLP, language models are statistical models that predict the probability distribution of a specific word sequence (Li et al., 2023). Early versions were built on Markov processes but later evolved to Recurrent Neural Networks (RNNs), building on neural networks that had the capability of remembering contexts across an increased amount of data (Li et al., 2023).

For computers to process natural language, textual data needs to be represented numerically (Mikolov et al., 2013). Word embeddings were introduced by Mikolov et al. (2013) and refer to the technique where words are represented by numerical vectors in such a way that their semantic relationships and contextual meanings are captured, and were a significant improvement as opposed to traditional word counting techniques (Mikolov et al., 2013; Young et al., 2017). One of the main limitations in building effective NLP algorithms includes the highly context-dependent meanings of words and sentences in natural language (Kadre et al., 2025). The field of NLP, therefore, demonstrated rapid advancements along with the emergence of the transformer technology (Kadre et al., 2025; Li et al., 2023). Transformers utilise self-attention mechanisms to consider words in parallel rather than sequentially as in the previous RNNs (Vaswani et al., 2017), which enables contextual embeddings where the representation of a word changes based on its surrounding context (Kadre et al., 2025; Li et al., 2023).

The transformer technology further contributed to the Generative Pretrained Transformer (GPT) models. Pre-trained models refer to deep learning models trained on huge corpora of textual data available on the internet, to learn natural language, grammar and factual statements (Kadre et al., 2025). They further contain billions of learned parameters, such as bias and weights (Kadre et al., 2025). By allowing computations to be parallelised across input sequences (Vaswani et al., 2017), the transformer architecture allowed

language models to scale in both parameters and training data. This expansion in scale marked the transition from language models (LMs) to what is now commonly referred to as Large Language Models (LLMs), where increasing the number of model parameters enhanced the capabilities of the language models (Kaplan et al., 2020). Through pre-training, LLMs utilising the transformer technology have achieved state-of-the art results across a diversity of NLP tasks (Li et al., 2023; Radford et al., 2018). GPT models are widely referred to as Generative AI models, as they can generate natural language textual data, including but not limited to, summaries, translations and answers to questions (Kadre et al., 2025). BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019), is an alternative transformer-based pretrained model designed primarily for understanding text and classification, rather than text generation (Devlin et al., 2019).

Utilising LLMs in applications can be done with either an open-source model or a commercial LLM provider (Li et al., 2023). Large companies such as OpenAI, Google, and Microsoft (Li et al., 2023), and Anthropic (Anthropic, n.d.a), provide general-purpose pretrained LLMs accessed through APIs. Open-source models such as LLaMA, BLOOM, and Flan-T5 are available through Hugging Face Transformers, but require self-hosting (Li et al., 2023). However, they do offer greater flexibility, as parameters such as weights and biases are accessible (Li et al., 2023). In contrast, commercial closed-source models accessed via APIs can be argued to be black boxes, as no parameters or other information about how output is generated is available (Xiong et al., 2024).

While LLMs perform well on general tasks, there are several approaches to adapting them to specific domains. Fine-tuning is one such approach, in which the model is further trained on domain-specific data to improve its performance within that domain (Brown et al., 2020). However, fine-tuning is computationally expensive and requires substantial amounts of data (Brown et al., 2020). Brown et al. (2020) introduced in-context learning with GPT-3, and demonstrated that the model was capable of performing a wide range of NLP tasks if the prompt included a task description along with a few examples of input-output pairs. Brown et al. (2020) further established the terms *zero-shot*, *one-shot* and *few-shot* learning, where "shots" refer to the number of input-output examples included in the prompt, and can be considered a prompt engineering technique. Prompts are essentially task-specific instructions used as input to an LLM to generate a desired output (Sahoo et al., 2024).

2.2.1 Prompt engineering

There are several prompting approaches to improve the output of an LLM. Firstly, structured task descriptions in prompts have been proven by several authors to improve performance in LLM outputs (Mishra et al., 2022; Sclar et al., 2024). Examples include dividing complex tasks into multiple smaller subtasks or breaking down instructions into steps (Mishra et al., 2022). Closely related is a technique called Chain-of-thought, where models are forced to divide complex tasks into separate reasoning steps (Wei et al., 2022). Secondly, assigning a role to an LLM is a standard practise among commercial AI

applications (Zheng et al., 2024). Kong et al. (2024) demonstrated that assigning a GPT model a role improves its reasoning capabilities when compared to zero-shot prompting. The roles can be that of a social identity or a domain expert, depending on the task (Salewski et al., 2023). Salewski et al. (2023) found that assigning a persona, such as a domain expert, increased performance on tasks within the specific domain, whereas social identities risked increasing bias. Lastly, prompt-level output constraining forces the output to specific formats, usually JSON or XML (Beurer-Kellner et al., 2023). The method has been proven to improve performance in LLM outputs (Mishra et al., 2022), improve adherence to task instructions (Dolphin et al., 2024) and enable integration with existing applications (Beurer-Kellner et al., 2023).

2.2.2 Limitations of LLMs

LLMs demonstrate several limitations. A primary concern is hallucinations, where models generate factually incorrect statements (Du et al., 2025; Kadre et al., 2025; Chiang & Lee, 2023), which raises concerns regarding truth, trust and privacy (Kadre et al., 2025). Additionally, the increased adoption of LLMs raises concerns regarding algorithmic bias, data privacy, job displacement and copyright issues (Kadre et al., 2025).

Commercial LLMs such as ChatGPT and Claude are finetuned with a method referred to as reinforcement learning from human feedback (RLHF) (Kadavath et al., 2022), which encourages helpful and harmless outputs, that align with user intent and expectations (Bai et al., 2022; Ouyang et al., 2022). This may, however, come at the cost of factual correctness, as the model is encouraged to produce outputs that are preferred by the human asking rather than responses that are strictly accurate (Kadavath et al., 2022). A related concern is the stochasticity of LLM responses across runs, which makes research on LLMs difficult to reproduce (Wang et al., 2024).

Further limitations are particularly relevant to the financial context of this thesis. LLMs are not inherently designed to handle time-series data (Gruver et al., 2023), which could prove to be a significant limitation when dealing with financial time series. Look-ahead bias is another limitation when using LLMs to reason about historical events, as the model may have been exposed to the data during training (Glasserman & Lin, 2023). Lastly, Lai et al. (2023) suggest that ChatGPT has limitations in performance in languages other than English, which is likely a consequence of the training data composition, as 93 % of GPT-3's training data was in English (Brown et al., 2020). This could imply a limitation in performance when using Swedish textual data as input. All three limitations are further discussed in relation to the methodological design.

2.2.3 NLP and LLMs Applied to Financial Text

NLP has been applied extensively within the financial domain, and typical tasks include text classification, named entity recognition, word similarity, and question answering chatbots (Kadre et al., 2025). The set of applications has expanded with recent advancements to include summarisation and reading comprehension (Radford et al.,

2019). However, NLP applications within a financial context are complicated due to domain-specific language (Loughran and McDonald, 2016), which has shaped the techniques developed and applied in the field.

Prior to the emergence of LLMs, bag-of-words techniques were dominant in financial NLP, where the order and sequence of words in a document were disregarded (Loughran and McDonald, 2016). Loughran and McDonald (2011) contributed with a dedicated dictionary for NLP tasks in financial contexts, motivated by the observation that words typically classified as negative in general language, such as *tax*, *cost*, or *liability*, may be considered neutral or positive in a financial context (Loughran and McDonald, 2016). The Loughran-McDonald dictionary represented an important step in specialising general NLP techniques to the financial domain. The underlying bag-of-words approach, however, remained limited in capturing meaning across different contexts, as it disregarded word sequence and order (Loughran and McDonald, 2016). Further, classical deep learning techniques applied to financial decision-making faced limitations regarding interpretability, as the underlying models constituted black boxes, which is particularly problematic in a context where transparency is essential (Tong et al., 2024).

The recent advancements of NLP applications in finance have primarily been driven by the increasing availability of textual data in various formats, such as text documents and news articles (Du et al., 2025). Modern NLP is widely applied in sentiment analysis, information extraction and question answering chatbots, and shows potential within financial regulation, compliance, fraud detection and credit scoring (Du et al., 2025). Generative AI is further assumed to impact wealth management, portfolio management and sustainable finance (Du et al., 2025). Within the financial domain, it is particularly important that generative AI possesses the capability of reasoning about textual data combined with numerical data (Du et al., 2025). In parallel, language models specifically pre-trained or fine-tuned on financial corpora have been developed, such as FinGPT (Yang et al., 2025) and BloombergGPT (Wu et al., 2023).

3. Related work

The relationship between news and stock market movements has been studied extensively by utilising the different techniques within NLP. Dictionary- and sentiment-based methods were foundational within the field, and methods have evolved within the emergence of transformer-based models and general-purpose LLMs. This thesis contributes within the field of linking news articles to stock market movements utilising general purpose LLMs, but shifts the angle from *predicting* market movements to *explaining* them.

3.1 NLP Approaches to linking news articles with stock market movements prior to LLMs.

Previous approaches to relating news articles to stock market movements centred around sentiment analysis, which is when words in text are analysed to determine the emotion of the text (Liu, 2012), usually in a binary output such as “positive” and “negative” (Kadre et al., 2025). The sentiment of news articles has been examined in relation to stock market movements prior to the emergence of LLMs by several authors, by utilising bag-of-words or deep learning methods (Araci, 2019). Tetlock (2007) constructed a daily pessimism measure from a Wall Street Journal column using dictionary-based sentiment analysis and found that high pessimism had a measurable negative effect on the American stock market. Further, Fang and Peress (2009) demonstrated that instruments with no media coverage yield higher returns than instruments with high media coverage. Engelberg and Parsons (2011) proved the causal effect of local news coverage to local markets, and argues that the same pattern could be expected of national news coverage.

Apart from sentiment, topic analysis has been used to study the same relationship. Latent Dirichlet Allocation (LDA) as described by Blei et al. (2003) has been implemented to conduct topic analysis, with notable results. Bybee et al. (2024) proved using LDA that topic in economic news articles highly coincide with numerical measures of the economy. Furthermore, Feuerriegel and Pröllochs (2021) demonstrated a discrepancy regarding the effect that different topics in 8K-filings (announcements that American companies are obliged to report following a major event that might impact the company (Kenton, 2024)) had on the stock market. Topics that had significant implications for asset returns include earning results, credit rating changes, business strategy announcements and acquisitions (Feuerriegel and Pröllochs, 2021). A study of the effects that the Norwegian business newspaper *Dagens Næringsliv* (DN) had on the fluctuations of the Norwegian stock market found the most important news topic to impact returns is news related to development in the financial market, credit and borrowing (Larsen and Thorsrud, 2019).

Araci (2019) argues that bag-of-words and dictionary-based methods are limited in their ability to capture context and word order, and proposed FinBERT, a model finetuned from BERT for financial sentiment analysis. However, despite the methodological advancement, the main tasks remained focused on sentiment classification rather than reasoning over the underlying content. The consequent emergence of general-purpose LLMs extended the field by enabling models to reason over textual data and generate natural-language outputs.

3.2 Utilising general-purpose LLMs in financial NLP

With the emergence of general-purpose Large Language Models, the relationship between news articles and stock market movements has been examined through new methodological approaches. Several authors have demonstrated that general-purpose LLMs perform competitively on financial NLP tasks despite not being explicitly trained

on financial data. Fatouros et al. (2023) compared ChatGPT to traditional sentiment classification models on financial text and found that ChatGPT achieved competitive results without any domain-specific training. Kirtac and Germano (2024) further demonstrated that sentiment signals extracted by LLMs can be used as input to a trading strategy and outperform dictionary-based approaches. Pelster and Val (2024) examined whether ChatGPT can assist in stock selection and found that LLM-generated stock recommendations yielded positive returns. Ardekani et al. (2024) compared general-purpose LLMs to a fine-tuned, domain-specific financial sentiment model and concluded that general-purpose models perform comparably to specialised models on financial sentiment tasks using few-shot learning. Additionally, Guo and Hauptmann (2024) demonstrated that fine-tuning LLMs on financial news outperforms conventional sentiment-based methods for stock return prediction. Fine-tuning, however, requires substantially more computational resources than utilising general-purpose models directly (Brown et al., 2020).

Apart from sentiment scoring, LLMs have further been utilised to extract structured information from financial text and to reason over textual data. Dolphin et al. (2024) demonstrated that LLMs can effectively extract structured metadata, such as ISIN codes and company names, from unstructured financial news articles, which enables integration with existing financial applications. Lopez-Lira and Tang (2025) utilised the general-purpose model GPT-4 to predict the direction of stock market movements based on news article headlines, and achieved around 90 % correct directional predictions. Lopez-Lira and Tang (2025) concluded that the general-purpose GPT-4 performed well on financial texts despite not being explicitly trained to do so, which raises the question of whether domain-specific fine-tuning is necessary for financial NLP tasks.

3.3 LLMs and market efficiency

The findings of Lopez-Lira and Tang (2025) can be interpreted in the context of the Efficient Market Hypothesis. Lopez-Lira and Tang (2025) utilised the general-purpose model GPT-4 to predict the direction of stock market movements based on news article headlines, and achieved approximately 90 % correct directional predictions. Prices were observed to drift in the predicted direction for an additional two to three trading days following publication (Lopez-Lira and Tang, 2025), which under the *semi-strong* form of the Efficient Market Hypothesis (Fama, 1970) represents the friction of the speed of which prices respond to new, publicly available information. Through topic analysis, Lopez-Lira and Tang (2025) further demonstrated that markets adjust efficiently to quantitative and transparent information, such as earnings reports and strategic partnerships, but adjust less efficiently to qualitative information, such as conference presentations or insider stock transactions, where prices drift for a longer period after publication. This elaborates the semi-strong form by suggesting that the speed of price adjustment varies across different types of contents in articles, but depends on how difficult it is to interpret the implications of the information for an asset.

Lopez-Lira and Tang (2025) further observe that the Sharpe ratio of an LLM-based trading strategy decreased over the sample period, and suggest that the increased availability of LLMs during the time might have enabled investors to analyse and act on new information faster, thereby increasing market efficiency. This reflect Fama's (1970) original argument that the frictions that prevent perfect efficiency are reduced as information becomes more accessible to market participants.

Lopez-Lira and Tang (2025) examine whether an LLM can act on news before prices adjust, which sits within the *semi-strong* form of the Efficient Market Hypothesis. This thesis takes the opposite approach, examining whether an LLM can identify, after a price movement has occurred, the news articles that explain it. Both questions sit within the semi-strong form, but address opposite sides of the price adjustment process.

4. Method

To address the research questions, the method was structured in four separate sections. First, daily closing prices for a set of Swedish listed companies were assessed for significant movements using a volatility-normalised z-score. Second, financial news articles were linked to companies through a Named Entity Linking step performed by an LLM. Third, for each identified movement, articles matched to the company within a seven-day window were passed to a second LLM call, which scored article relevance, generated a natural-language explanation, and attached a verbalised confidence level. Finally, since no objective truth exists for which articles explain a given movement, the outputs were evaluated by financial professionals through a structured form, which also provided a benchmark against which the LLM's verbalised confidence could be compared.

4.1 LLM Providers

Although LLMs explicitly trained in financial contexts does exist, this project implemented a commercial general-purpose model, as they have been proven to perform well in financial applications (Li et al., 2023). There are multiple commercial providers of general-purpose models, and the main provider used for this specific project was Anthropic. The choice of Anthropic as a primary provider was mainly motivated by practical reasons, such as access through an API by the partnering company, as well as integration with their technological stack.

Given the pace of development in the field, the set of available models updates frequently. At the time of the project, the most recent models released by Anthropic were Claude Haiku 4.5, Claude Sonnet 4.6, and Claude Opus 4.7 (Anthropic, n.d.a). Haiku being the smallest and fastest, Sonnet intermediate, and Opus the largest and slowest (Anthropic, n.d.a). Performance on complex reasoning tasks scales correspondingly from Haiku to Opus (Anthropic, n.d.a).

There are several parameters used to engineer LLM outputs. The model's *temperature* sits on a scale of 0 to 1 and controls the randomness in the model's response (Anthropic, n.d.b). A higher temperature will increase creativity and diversity in outputs, but lower determinism (Anthropic, n.d.b). However, a temperature of 0 will not yield fully deterministic outputs (Anthropic, n.d.b). Further, the maximum number of tokens in the output can be chosen. For Claude, one token corresponds to 3.5 characters (Anthropic, n.d.b). The knowledge cutoff date for Claude Sonnet 4.6 is May 2025 (Anthropic, 2026), which essentially removes the risk of look-ahead bias, as the input data sample spans November 2025 – April 2026. No information about the knowledge cutoff date was available for Claude Haiku 4.5, but as the model was released prior to Claude Sonnet 4.6, it is assumed to have an earlier cutoff date. Further, the task that Claude Haiku 4.5 was asked to perform is not at risk of being affected by the look-ahead bias effect.

As discussed in the preliminaries, the performance of general-purpose LLMs has been proven to decay in tasks involving languages other than English. Anthropic provides performance data of their models across a set of different languages, and although Swedish is not included, it shows that performance decays for other languages than English (Anthropic, n.d.d). Performance in non-English languages ranges from 79.7 % to 98.2 % relative to English performance (100%) (Anthropic, n.d.d). The span is quite large, and does not provide more than an indication of the performance of the model when prompted with Swedish text, and thus remains a limitation of the methodological choice. Another concern regarding the performance and trust in LLM output is that of reinforcement learning from human feedback (RLHF), where models prioritise responses that receive the most "reward" by a human (Kadavath et al., 2022).

4.2 Data

There were two main data types used to examine the thesis aim: (i) newspaper articles and (ii) historical stock price data. In the following section, the selection of data sources, collection methods and data processing steps is outlined. The time span that is of interest for the study is between January 2026 and April 2026, and the news articles used in the study is therefore published between the 2nd of January and the 20th of April. Stock market prices span 30 days prior to provide the rolling volatility window needed for the statistical anomaly detection method.

4.2.1 News articles

Financial newspaper articles were provided by Uppsala University, and the set consisted of printed articles in *Dagens Industri* (DI) published between the 2nd of January 2026 and the 20th of April 2026. DI is the largest business newspaper in the Nordic region, and provides coverage across several industries (Dagens Industri, 2025), and was therefore selected as the news source for the project. The articles were stored in JSON format, containing the fields *publication date*, *article title*, and *article content*. As all articles were originally printed, they were timestamped at 00.00 on the day of publication. Since they became available to the public on the morning of publication, they were included for analysis in relation to identified anomalous price movements dated on the same day. The total number of articles in the data set was 2 612. Due to licensing restrictions, article titles and body text from DI are not included in the thesis. Where individual articles are mentioned, they are referenced by their matched company and publication date.

4.2.2 Market data and stock selection

Historical stock market data was provided by the partnering company through an API, with data attributes *date*, *closing price*, *ISIN* and *volume*. Stock market data was collected for 70 Swedish-listed companies. The full list of selected companies can be found in Appendix A, tables 16-22. The selection of companies was primarily driven by news coverage in the available set of news articles to produce enough data points for the subsequent LLM analysis task and ensure meaningful evaluation. It should be noted that

this constitutes a delimitation, as the analysis is conditioned on companies with available news coverage. This is consistent with the primary aim of the project, which is to examine the capabilities of the LLM to generate a narrative explanation. Further, priority was given to stocks commonly traded by retail investors on Avanza, the biggest investment platform in Sweden (Avanza Bank Holding AB, 2026). The real-estate sector was excluded because its price movements are primarily driven by macroeconomic factors, particularly interest rate movements and corporate bond market conditions (Sveriges Riksbank, 2025, pp. 11–12). Instead, industries such as industrials, consumer retail, and technology were prioritised, where stocks tend to have a larger retail investor ownership (Avanza, n.d.) and thus higher trading activity driven by public sentiment (Barber & Odean, 2008).

4.2.3 Identifying significant market movements

Significant market movements needed to be identified and used as input for the LLM explanation task. Based on what was found in the background section, the volatility-normalised z-score was used to identify such movements. For each stock, daily log returns were computed as $r_t = \ln(P_t/P_{t-1})$. Both the mean μ_w and standard deviation σ_w of returns were calculated from an estimation window of 30 trading days before t.

The standardised return on a day t is

$$z_t = (r_t - \mu_w)/\sigma_w,$$

and a return was flagged as significant if $| z_t | \geq k$.

The use of a rolling estimation window rather than a static one was motivated by volatility clustering as demonstrated by Engle (1982), which means periods of high volatility tend to follow periods of high volatility. A static estimate of σ would therefore misclassify periods of high volatility as a single extreme market movement and miss outliers within calmer market periods. The 30-day window was considered short enough to accurately reflect the clustered volatility of the period, while still being long enough to provide a stable estimate of σ. The threshold k was set to 1.96, which would identify 5% of returns as anomalous given a Gaussian distribution. However, as proven by Cont (2001), financial returns are heavy-tailed, so a Gaussian z-score systematically underestimates the number of extreme events. Among the 70 selected companies, the statistical detector identified 547 significant price movements. Given the sample size of 5 128 price points during the sample period, 10.6 % of movements were identified as significant, which reflects the heavy-tailed property of the financial time series. The threshold was kept at 1.96, which is a deliberately permissive choice that aligns with the main objective of the thesis to explore the capabilities of an LLM to generate a narrative explanation rather than the precision of anomaly detection. Further, the method does not account for general market movements. This means that even if a ±9 % move is large in absolute terms, the stock might simply have been following the market on a turbulent trading day. A model

that adjusts returns to general market movements could be useful in isolating company-specific events, but it is out of scope for this thesis.

4.2.4 Company-to-article matching

It was essential to link each article to one or more of the 70 stocks in the data set. This step can be considered a Named Entity Linking (NEL) problem; given mentions of company names in unstructured Swedish news text, link them to one of the stocks being assessed within the thesis. Hence, the goal of this step was to identify, for every article, the company that the article was primarily about, as well as any other companies mentioned in the article.

Two separate approaches were evaluated. The first approach utilised naive substring matching of company aliases against article text, but the approach proved to be limited in several aspects. As an example, the alias *abb* (for ABB Ltd) is a substring of several common Swedish words such as *snabb* ("fast"), *drabbat* ("affected"), and *rabatt* ("discount"). The alias *rusta* (for Rusta AB) appears within the Swedish verb *rustar*. This resulted in a high number of false positives among the linked articles. This motivated the implementation of the second approach. As demonstrated by Dolphin et al. (2024), LLMs are effective in extracting structured metadata such as ISIN and company names from financial news articles.

The implementation called Claude Haiku 4.5 at a temperature 0 for each article, to perform the NEL task. Anthropic's smallest and least capable model was chosen, as the task was well-defined by a preset list of company ISINs and a constrained output schema, and therefore didn't require the reasoning capabilities of a larger model. The input consisted of two system prompts and one user prompt. The first system prompt, as found in Appendix B, listing 4, provides context and task scope, whereas the second system prompt, found in Appendix B, listing 5, provides the list of 70 companies, commonly used abbreviations, ISIN, and sector. The user prompt changes for every call and contains the article title, the first 1 500 characters of the article body, and a constrained JSON response schema, as found in Appendix B, listing 6. The purpose of using of a constrained response schema was to enable reproducibility and technical integration. Further, ISINs generated by the LLM that could not be found in the list of companies were discarded to eliminate hallucinated ISINs. The second approach further enabled the separation between primary articles mentioned in the articles and articles only mentioned in passing.

The articles were then matched to each anomaly by company and a temporal window of 168 hours (seven days), ending at the time of market close on the anomaly day. Only articles in which the LLM identified the company as the primary subject were included in the matching. Of the 547 anomalies flagged in the statistical detection, 67 had at least one matched article within this window. The distribution of anomalies across sectors, along with their mean z-score and mean number of matched articles is presented in Table 1. Among the matched anomalies, there were a total of 91 unique articles. These matches formed the input to the subsequent LLM explanation task. Note that re-running the

resolution step on the same data might not yield identical results, due to the stochastic nature of LLM outputs, despite a temperature at 0. This is a known limitation of the method.

Table 1. Distribution of matched anomalies across sectors, mean z-score and mean number of matched articles.

Sector	Anomalies	Mean z-score	Mean # of matched articles
Finance and Investment	18	−0.5	2.17
Industrials	16	−0.36	1.31
Tech, digital and media	13	0.45	1.38
Consumer, retail and services	11	0.28	1.91
Mining	6	−1.82	2.5
Healthcare, pharma and care	3	0.86	1

4.3 LLM Analysis task

Section 4.2.3 resulted in a set of stock market anomalies, and Section 4.2.4 matched each anomaly to a set of news articles, linked by NEL and a temporal window of seven days. The following section aims to describe the main LLM-analysis task, which can be divided into four high-level steps:

1) Reason over each company-to-article match.
2) For each article, provide a score signalling the article's relevance to the identified price movement.
3) Produce a natural-language, narrative explanation based on the matched set of articles.
4) Verbalise a confidence level for each anomaly explanation.

The model chosen to perform the analysis task was Claude-Sonnet 4.6, which has demonstrated higher reasoning capabilities than Claude Haiku 4.5. For each company-to-article match, the model was invoked with one system prompt and one user prompt. The system prompt applied role prompting to encourage the model to reason like “a senior financial analyst specialising in Swedish listed equities”, and can be found in Appendix C, listing 7. The technique of role prompting was motivated by the improved performance in specific domains of expertise as demonstrated by Salewski et al. (2023). The user prompt provided structured task instructions, along with the relevant data about the company-to-article matching: Company name, date and size of the price movement, and all articles that were matched in the previous step, as well as their temporal proximity to the movement. The structured task description contained three tasks:

i) Return a relevance score on a 0-10 scale for each matched article, where 10 = the article directly explains the move, 5 = the article is related but not the clear driver, and 0 = the article is unrelated or coincidentally appeared within the time window.

ii) Generate a 3-5 sentence long explanation of the price movement, citing the articles which received the highest relevance score.

iii) Set a confidence label in {high, medium, low, none}, based on the generated explanation's quality.

Finally, the user prompt constrained the output by forcing it to a structured JSON. The user prompt with the structured task description can be found in Appendix C, listing 8, and the appended constrained JSON output is presented in Appendix C, listing 9.

The model was invoked through an API call once per price movement. This design implied a larger prompt and a smaller number of API calls, which enabled the model to compare candidate articles against each other. To save computational power and lower costs, each article was truncated after 2 000 characters. The temperature of the model was set to 0.2, and a maximum response length of 4 000 tokens. The temperature was restricted to enable reproducibility, while still enabling a certain degree of creativity in the outputs. To further promote reproducibility, the outputs were constrained to a structured JSON format.

4.3.1 Prompt engineering

The prompting techniques chosen for the LLM-analysis task followed the principles discussed in section 2.2.1, as well as general prompting guidelines published by Anthropic. The main technique used was the structured task description, where context and data was included at the beginning of the prompt, and the task was described last, which aligns with the recommended prompting technique by Anthropic when including large data in prompts (Anthropic, n.d.c). The task was then divided into subtasks, motivated by the findings of Mishra et al. (2022). The subtasks enforced separate "reasoning" steps, effectively implementing *chain-of-thought* as proven successful by Wei et al (2022). Further, role prompting was motivated by both the demonstrated benefits of Salewski et al. (2023), as well as the guidelines as published by Anthropic (Anthropic, n.d.c). Additionally, the prompts were kept clear and direct. The output was constrained at the user prompt level to produce a structured JSON schema. The prompts can be found in Appendix C, listings 7-9.

Limitations of Prompt engineering Salewski et al. (2023) found that hidden biases within LLMs risked being amplified when assigning roles to LLMs. Additionally, Zheng et al. (2024) demonstrates that role prompting, despite being standard practice in commercial applications, have negligible effects on the quality of the LLM output. Further, output constraining has been demonstrated to degrade the reasoning capabilities

of LLMs when performing complex tasks (Tam et al., 2024). The choice of constraining the output was thus primarily motivated by reproducibility, rather than improving the reasoning capabilities.

4.3.2 Verbalised confidence

To address RQ2, the model is required to report a verbalised confidence level to every generated explanation among four labels: {high, medium, low, none}. The labels were defined with an explanatory sentence to guide the choice:

- **High** - The explanation is well-supported by the articles, factually grounded, and adequately accounts for the move
- **Medium** - The explanation is plausible but partial: incomplete coverage, weaker supporting articles, or unverified detail
- **Low** - The explanation is weakly supported, speculative, or relies on weak article connections
- **None** - No meaningful explanation can be given from these articles

This is the verbalised confidence method as studied by Tian et al. (2023) and Xiong et al. (2024), in which the model is prompted to output a confidence label as part of its output. Tian et al. (2023) proved the confidence rating to be well-calibrated when compared to other performance measurements of LLM outputs. Kadavath et al. (2022) demonstrates that LLMs for self-assessed probability are well calibrated for a selection of tasks, which motivates exploring whether this holds for a reasoning task within the financial domain as well. Further, verbalised confidence is a method to counteract the black box of commercial, closed-source LLMs, as it surfaces probabilities within the generated outputs (Xiong et al., 2024). In Section 4.4.3, the distribution of verbalised confidence across runs was compared with expert relevance ratings on a selected sample of explanations and across 3 separate runs.

4.4 Evaluation of results

Both research questions rely on the judgement of financial professionals, as there is no objectively true answer to whether a given news article explains a price movement. Human evaluations of LLM-based systems have been argued for by several authors (Chiang & Lee, 2023; Van der Lee et al., 2021), and especially when outputs include factual claims to counter the known limitations of hallucinations in LLM output (Chiang & Lee, 2023). The overall evaluation method consists of a form sent out to financial professionals. The results of the form were then compared systematically to the article ratings and verbalised confidence in the output of the LLM Analysis task.

4.4.1 Sampling of anomalies for expert evaluation

A total of 10 anomalous price movements with matched articles were included in the form for evaluation by human experts. Out of the 10 anomalies, 3 were rated by the LLM with

high confidence, 3 with *medium* confidence, 3 with *low* confidence and 1 with *none*. For each anomaly, at most 3 articles were included for the experts to assess. 5 of the anomalies were a positive movement, and 5 were a negative movement. All anomalies included in the form can be found in Table 2.

Table 2. Anomalies included in the evaluation form for the financial professionals.

Company	Date	Price movement	LLM Confidence	Matched Articles
Hexatronic Group	2026-02-06	+10.66 %	high	1
Electrolux B	2026-02-02	+7.33 %	high	3
Boliden	2026-03-19	−9.00 %	high	3
Acast	2026-02-12	−4.30 %	medium	1
Hoist Finance	2026-01-16	+3.55 %	medium	1
Volvo B	2026-03-19	−5.80 %	medium	1
Fagerhult Group	2026-01-13	+2.66 %	low	1
H&M B	2026-02-04	+3.64 %	low	3
Fagerhult Group	2026-01-14	−2.92 %	low	1
Humana	2026-01-26	−2.78 %	none	1

4.4.2 Expert evaluation form

For the qualitative evaluation of the LLM outputs, a form was sent out to financial professionals to provide their expert opinion on the output. For each of the 10 anomalies presented, there were three main sections, all presented in Appendix D, figures 6-9. Firstly, the expert was presented with background context: the company, the date, the magnitude of the anomalous price movement, and up to three news articles matched to the anomaly (Figure 6). For each article, the expert was asked to provide a relevance score on a 0-10 scale (Figure 7), where 0 indicates the article is an unrelated coincidence, and 10 indicates that the article directly explains the price movement. For the third and last step, the experts were asked to assess the explanation quality across three dimensions on a 1-5 scale: whether the explanation is *factually correct*, whether the articles cited in the explanation are *relevant*, and whether the explanation would be *useful* to a regular retail investor trying to understand why their holding moved (Figure 8). Optional free-text comments for each anomaly were made available across the form to capture anything missing in the questions (Figure 9).

Importantly, the LLM's verbalised confidence in its explanation, the LLM's 0-10 relevance score for each article, and the LLM's ranking of the articles were intentionally hidden from the experts during evaluation. This was to ensure that the expert's judgements

were not affected by the model's own assessments, which would otherwise compromise the comparison between expert and LLM scores.

4.4.3 Reproducibility across runs

As discussed in Section 2.2.2, the variability of LLM outputs is a known limitation (Wang et al., 2024), and thus follows that a temperature of 0 does not guarantee fully deterministic outputs (Anthropic, n.d.b). To assess the stability of the LLM Analysis task, it was run three separate times on the same set of 67 anomalies, using identical prompts, input data, and model parameters as described in Section 4.3. Two aspects of the output were compared across runs. Firstly, the 0-10 relevance scores assigned to each article were compared across runs to assess the stability of scoring. Secondly, the verbalised confidence label was compared per anomaly to assess whether the same explanation received the same confidence label across runs. The results of the comparison are presented in Section 5.

5. Results

The following chapter presents the results of the algorithm described in Chapter 4. Section 5.1 presents an overview of the results from the statistical anomaly detection, company-to-article matching, and the LLM analysis task. Section 5.2 further presents the results of the expert evaluation form and compares the LLM's verbalised confidence to the expert ratings. Section 5.3 reports the reproducibility of the LLM analysis task across three identical runs.

5.1 Overview

The method consisted of three main steps, and this section reports the output of each step. the statistically anomalous price movements identified by the statistical anomaly detector, the company-to-article matching produced by Claude Haiku 4.5, and the relevance scores, narrative explanations, and verbalised confidence labels as produced by Claude Sonnet 4.6.

5.1.1 Significant market movements

The statistical z-score detector with threshold k = 1.96 identified a total of 547 statistically anomalous price movements, which constitutes 10.6 % of the 5 128 data points in the available set. At least one significant movement was identified in each of the 70 stocks included in the dataset. The distribution of statistically anomalous price movements across the different sectors can be shown in Figure 1. Note that the distribution aligns with the distribution of the stocks in the set across the sectors.

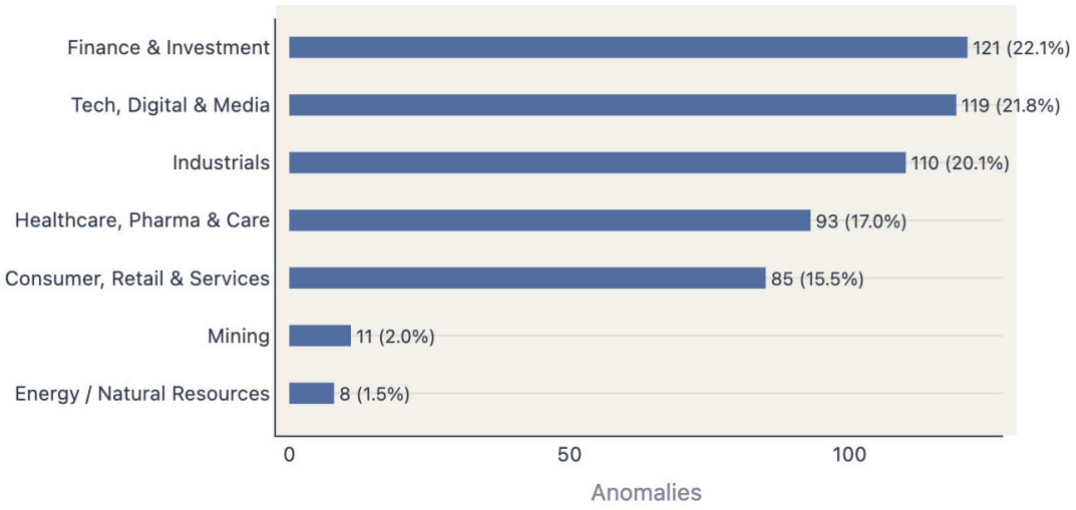

Figure 1. All detected anomalies by sector.

5.1.2 Company-to-article matching

Claude Haiku 4.5 at temperature 0 was called to perform the company-to-article matching step among the 2 612 DI news articles and match them to one of the 70 companies. Out of the 2 612 DI articles, 428 article–company pairs were produced. The temporal-matching step then restricted the output to only include articles published within 168 h (7 days) of an identified price movement, which lowered the number of unique articles included in the subsequent analysis step to 91. Because 26 articles matched to multiple anomalies, this produced 117 company-to-article pairs across 67 anomalies. The temporal distribution of matched news articles can be found in Figure 2. Note that 37.6 % of matched articles were published on the same trading day as the observed price movement. No matched articles were published more than 4 days prior to the identified price movement, even though the temporal window allowed up to seven days. The distribution of number of articles per price movement can be viewed in Table 3. Note that 40.3 % of the price movements had 2 or more articles matched.

Table 3. Distribution of articles per anomaly

Number of articles	Anomalies
1	40
2	12
3	9
4	4
5	2

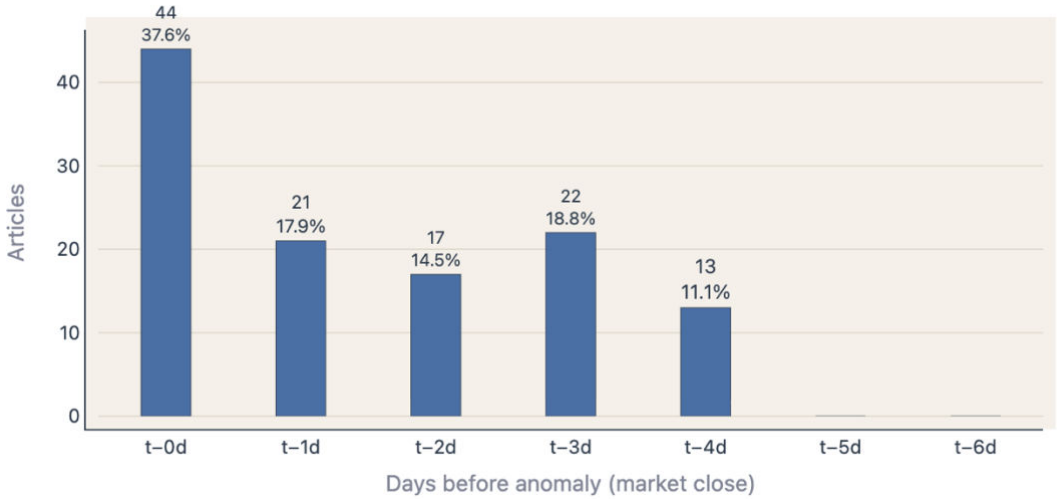

Figure 2. The temporal distribution of matched news articles.

5.1.3 LLM-analysis task

The main analysis task produced three outputs per anomaly:

i) A 0-10 relevance score for each matched article.

ii) A 3-5 sentence narrative explanation of the price movement.

iii) A verbalised confidence label in {high, medium, low, none}.

Article relevance scoring Figure 3 presents the distribution of relevance scores (0-10) assigned to the 117 company-to-article pairs, of which 48 (41 %) received a score higher than 5. This indicates that for a majority of matched articles, the model did not consider the article a clear driver of the movement.

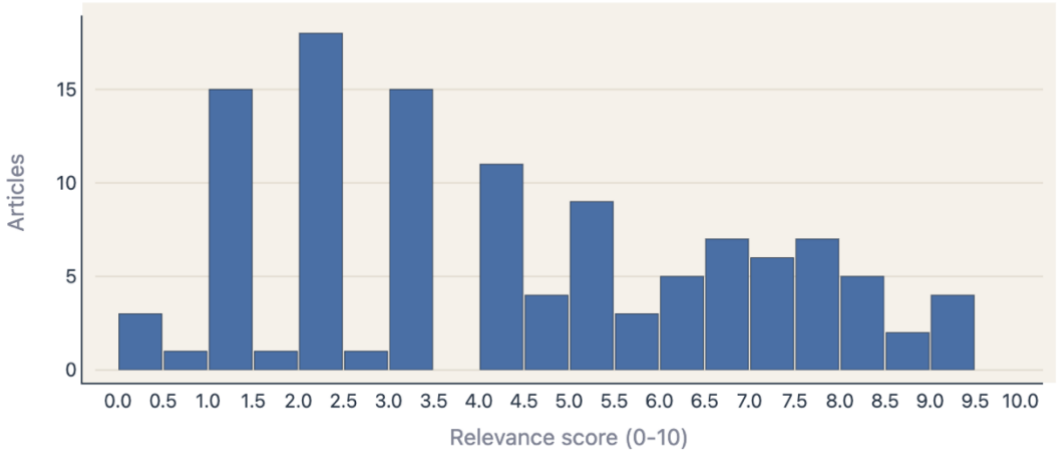

Figure 3. Distribution of article relevance scores.

Narrative explanation All 67 anomalies received a 3-5 sentence explanation. Three examples are shown in Listings 1-3. All three were included in the expert evaluation sample described in Section 4.4.1.

Company: Humana
Date: 2026-01-26
Price movement: -2.78%

Confidence: NONE

No meaningful explanation for Humana's -2.78% decline on 2026-01-26 can be derived from the available article. [Article 1] is a general market recap from 2026-01-24 covering OMXS30 and various individual stocks, but Humana is not mentioned at all. The price move may have been driven by company-specific news (e.g., earnings, guidance, or regulatory developments) or broader sector sentiment in Swedish healthcare/elderly care, but none of these potential drivers are evidenced in the provided article.

Listing 1. LLM-generated explanation for Humana (−2.78 %, 2026-01-26), assigned NONE confidence.

Company: Boliden
Date: 2026-03-19
Price movement: -9.00%

Confidence: HIGH

Boliden's -9% decline on 2026-03-19 was primarily driven by the forced shutdown of its most profitable mine, Garpenberg, following an unusually strong seismic event over the preceding weekend [Article 3]. Garpenberg accounted for roughly one-third of Boliden's total operating profit (4.4 of 13.2 billion SEK in the prior year), making any prolonged halt highly material [Article 3]. As of the capital markets day on March 19, the mine remained offline with no confirmed restart date, and there was unresolved uncertainty about whether the mine's shaft infrastructure had sustained damage – a scenario the CEO acknowledged could cause a significant production setback [Article 1]. The financial bleed was estimated at approximately 10 million SEK per day with no ability to recover lost output [Article 1]. Although Boliden simultaneously announced 5.5 billion SEK in new investments, the positive long-term signal was overwhelmed by near-term operational risk, and the stock closed sharply lower [Article 2].

Listing 2. LLM-generated explanation for Boliden (−9.00 %, 2026-03-19), assigned HIGH confidence.

```
Company: Acast
Date: 2026-02-12
Price movement: -4.30%

Confidence: MEDIUM

Acast's Q4 earnings report, covered in [Article 1], appears to
be the primary driver of the -4.30% price decline on 2026-02-
12. While the company reported strong operational metrics —
including a more than doubled operating result and revenue
growth from 576 to 733 million SEK — the bottom line came in
negative due to currency headwinds. Investors likely reacted
negatively to the net loss, which overshadowed the otherwise
encouraging listener growth figures (50% in North America, 16%
in Europe). The market may have expected the strong operational
momentum to translate into net profitability, and the currency-
driven loss disappointed those expectations. However, the
article is a brief print summary and does not provide full
detail on guidance or market consensus, leaving some
uncertainty about the complete picture.
```

Listing 3. LLM-generated explanation for Acast (−4.30 %, 2026-02-12), assigned MEDIUM confidence.

Verbalised confidence The distribution of verbalised confidence labels across the 67 anomalies is presented in Figure 4. Five explanations (7.5 %) received the label *high*, 22 (32.8 %) *medium*, 19 (28.4 %) *low*, and 21 (31.4 %) *none*.

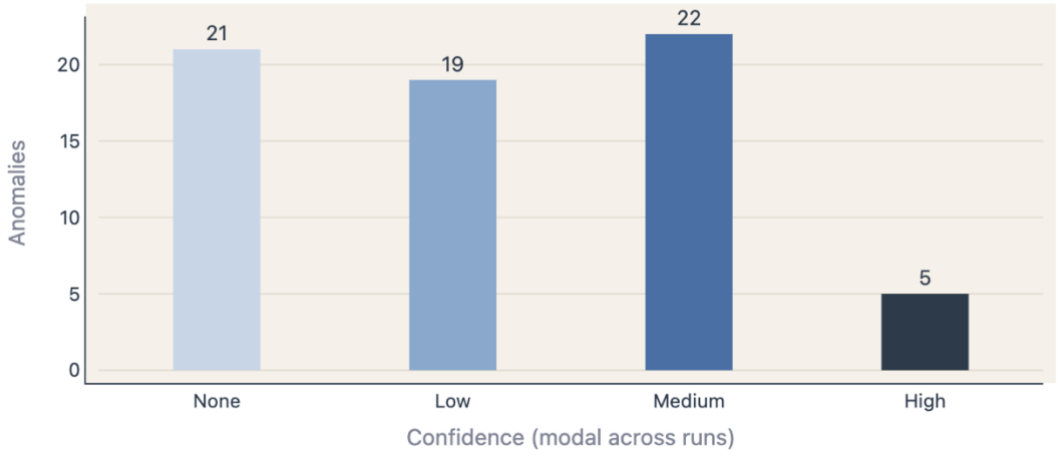

Figure 4. Distribution of verbalised confidence labels.

Assessing the composition of anomalies across confidence labels reveals that *none* labels are not primarily a consequence of the low threshold of k = 1.96, which could be expected to over-flag movements reflecting general market activity rather than company-specific,

significant movements. As shown in Table 4, *none*-labelled anomalies show a mean absolute return of 4.11 % and a mean absolute z-score of 3.16, comparable to the *medium* and *low* labels and above the detection threshold. Instead, the share of anomalies matched to only a single article rises from 40 % for *high* to 86 % for *none*, suggesting that lower confidence labels reflect a smaller article base rather than a weaker underlying price signal. Notably, the largest price movement in the entire dataset (Volvo Car B, +25.53 % on 6 February 2026) was assigned a *none* confidence label, indicating that the absence of an explanation reflects the lack of relevant matched articles rather than noise in the detected movement.

Table 4. Anomaly mean and median movement by confidence label

Confidence	n anomalies	Mean |Δ%|	Median |Δ%|	Max |Δ%|	Mean |z|	Share of anomalies with 1 article
high	5	9.15 %	9.00 %	16.98 %	3.76	40% (2/5)
medium	22	4.65 %	3.99 %	13.62 %	3.02	41 % (9/22)
low	19	3.81 %	3.59 %	7.49 %	2.76	58% (11/19)
none	21	4.11 %	2.89 %	25.53 %	3.16	86% (19/21)

5.2 Expert Evaluation

The expert evaluation form described in Section 4.4.2 was completed by 3 financial professionals and consisted of 10 anomalies with a total of 16 matched articles, as displayed in Table 2, section 4.4.1. The empirical results are reported in three parts: Section 5.2.1 presents the article-level relevance scores and compares them with the corresponding LLM-scores, Section 5.2.2 presents the experts' overall and 1-5 scale judgements of the generated explanations and compares them with the LLM's verbalised confidence labels, and Section 5.2.3 summarises themes from the free-text comments.

5.2.1 Article relevance scoring

The experts' 0-10 article relevance scores were compared to the relevance scores assigned by the LLM for the 16 articles cited across the 10 evaluation anomalies, as demonstrated in Figure 5. Points above the identity line are articles the experts judged more relevant than the LLM itself did, and similarly, points below the dotted line are articles that the experts judged less relevant. The two sets of scores broadly corresponded, as articles assigned a low relevance score by the LLM were generally also rated low by the experts. However, more articles fell below the identity line than above it, indicating that the LLM tended to assign higher relevance scores than the experts. The disagreement was largest among articles that the LLM scored highly, several of which the experts considered substantially less relevant to the price movement, whereas comparatively few articles were judged more relevant by the experts than by the LLM.

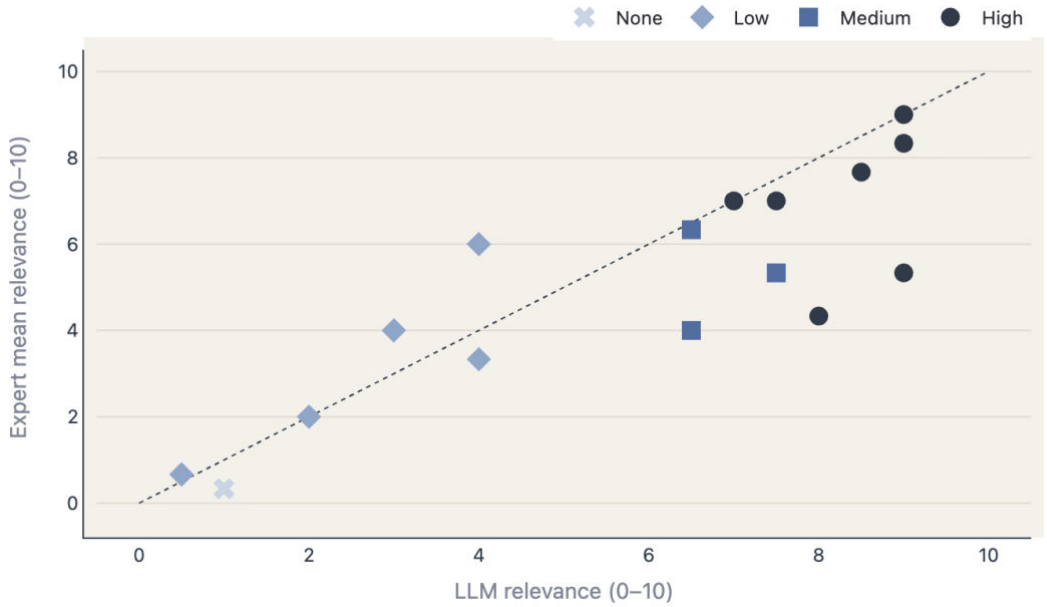

Figure 5. LLM-assigned vs expert article relevance for the 16 articles cited across the 10 evaluation anomalies. Each marker is one article. The dotted line is the identity $y = x$.

5.2.2 Explanation quality evaluations

After reviewing and evaluating the articles, the experts were presented with the generated explanation and asked to rate it on a scale of 1-5 across three dimensions: *factual correctness*, *relevance* of the cited articles, and *usefulness* for helping retail investors understand why a holding had moved. The ratings provided by the three experts for each of the ten evaluated anomalies are presented in Tables 5–14, grouped according to the confidence label assigned by the LLM to the corresponding explanation. The results are reported below for each confidence category, from *high*-confidence to *none*-confidence.

For explanations assigned the high-confidence label, which the LLM therefore considered to be of the highest explanatory quality, the 1-5 ratings were consistently high. *Factual correctness* received the highest overall ratings, while *relevance* and *usefulness* were never rated lower than 4. Across the three anomalies in this category, the mean ratings were 4.78 for *factual correctness*, 4.33 for *usefulness*, and 4.22 for *relevance* of the cited articles. These results indicate that the experts consistently judged the explanations to be factually accurate, well supported by the cited articles, and useful for helping retail investors understand the observed price movements.

Table 5. Anomaly 1 - Hexatronic Group, 2026-02-06, +10.66%, LLM confidence: high

Expert	Factual correctness	Relevance	Usefulness
1	5	5	5
2	4	4	4
3	5	4	5
Mean	**4.67**	**4.33**	**4.67**

Table 6. Anomaly 2 - Electrolux B, 2026-02-02, +7.33%, LLM confidence: high

Expert	Factual correctness	Relevance	Usefulness
1	5	5	5
2	5	4	4
3	5	4	4
Mean	**5.00**	**4.33**	**4.33**

Table 7. Anomaly 3 - Boliden, 2026-03-19, −9.00%, LLM confidence: high

Expert	Factual correctness	Relevance	Usefulness
1	5	4	4
2	5	4	4
3	4	4	4
Mean	**4.67**	**4.00**	**4.00**

Among the anomalies that were labelled with medium-confidence, the 1-5 ratings were lower, with means of 3.56 for *factual correctness*, 3.44 for *relevance*, and 3.11 for

usefulness. The ratings also displayed greater disagreement between the experts, most notably for Volvo B, which was rated from 5 to 2 across all three dimensions. *Usefulness* was the lowest-rated dimension in this group, indicating that even where the experts found the explanations factually correct, they did not consider them valuable to a retail investor.

Table 8. Anomaly 4 - Acast, 2026-02-12, −4.30%, LLM confidence: medium

Expert	Factual correctness	Relevance	Usefulness
1	3	3	3
2	4	4	3
3	3	3	2
Mean	**3.33**	**3.33**	**2.67**

Table 9. Anomaly 5 - Hoist Finance, 2026-01-16, +3.55%, LLM confidence: medium

Expert	Factual correctness	Relevance	Usefulness
1	4	4	4
2	4	4	3
3	4	3	3
Mean	**4.00**	**3.67**	**3.33**

Table 10. Anomaly 6 - Volvo B, 2026-03-19, −5.80%, LLM confidence: medium

Expert	Factual correctness	Relevance	Usefulness
1	5	5	5
2	3	3	3
3	2	2	2
Mean	**3.33**	**3.33**	**3.33**

Among the *low* confidence anomalies as displayed in tables 11-13, *factual correctness* remained relatively high (mean 3.78), while *relevance* (mean 2.78) and the *usefulness* of the explanation (mean 2.44) were rated lower. The pattern was most evident for H&M B (table 12), with a mean *factual correctness* of 4.00 and a mean *usefulness* of 1.67. This indicates that the experts considered the explanations factually correct, yet not relevant to the price movement and thus the resulting explanation of limited use.

Table 11. Anomaly 7 - Fagerhult Group, 2026-01-13, +2.66%, LLM confidence: low

Expert	Factual correctness	Relevance	Usefulness
1	4	3	4
2	4	4	4
3	4	3	3
Mean	**4.00**	**3.33**	**3.67**

Table 12. Anomaly 8 - H&M B, 2026-02-04, +3.64%, LLM confidence: low

Expert	Factual correctness	Relevance	Usefulness
1	2	2	2
2	5	3	1
3	5	2	2
Mean	**4.00**	**2.33**	**1.67**

Table 13. Anomaly 9 - Fagerhult Group, 2026-01-14, −2.92%, LLM confidence: low

Expert	Factual correctness	Relevance	Usefulness
1	2	2	2
2	4	4	2
3	4	2	2
Mean	**3.33**	**2.67**	**2.00**

The results of the *none*-confidence anomaly, for Humana, received the lowest ratings for *relevance* and *usefulness*, both of which were rated 1 by all three experts. This is consistent with the LLM's own assessment, as the explanation explicitly stated that no meaningful reason behind the movement could be given from the available article. *Factual correctness*, in contrast, was rated unevenly by the experts, where one rating of 1 and two ratings of 5 resulted in a mean rating of 3.67, which may reflect differing interpretations of factual correctness for an explanation that correctly declines to explain the movement rather than fabricating an inaccurate explanation.

Table 14. Anomaly 10 - Humana, 2026-01-26, −2.78%, LLM confidence: none

Expert	Factual correctness	Relevance	Usefulness
1	1	1	1
2	5	1	1
3	5	1	1
Mean	**3.67**	**1.00**	**1.00**

Considering the results of the ratings across the confidence labels, the expert ratings generally aligned with the LLM's confidence labels, with the strongest alignment for the *relevance* of the cited articles and the *usefulness* of the explanation, and weaker for *factual correctness*, which remained relatively high even among the *low*- and *none*-labelled explanations. The results of the 1-5 ratings across the three dimensions are complemented by the experts' free-text comments, which offered a more nuanced picture of the individual explanations and are presented in the following section.

5.2.3 Themes from free-text comments

The experts were given the option to provide free-texts comments about each anomaly, which provided more nuanced results about the quality of the explanations and the relevance of the matched articles. The comments revealed several themes among the explanations and their quality, which are presented below.

Among the explanations that the LLM rated with *high* confidence, the experts noted the LLM drew heavily on the available set of articles, which, despite being relevant, might not explain the full picture of the movement. Regarding Boliden's 9% decline on March 19, one expert observed that although the explanation accurately reflected the three matched articles, the articles themselves did not capture the broader macroeconomic and geopolitical conditions influencing copper and gold prices during the period, despite Boliden's strong exposure to these markets. Further, Electrolux yielded a +7.33 % return on the 2nd of February, drawing on 3 different articles. One of the experts noted that the price gain could likely be explained completely by the Q4-report mentioned in two of the articles, and that the generated explanation is a bit careful in conveying the explanatory importance of the Q4 earnings report, and draws a bit more heavy on the third article, which mentions the CEO's optimism following the tariffs in the US.

Among the anomalies that received a *medium* confidence label by the LLM, the experts indicated the explanations fell short in several aspects. A -4.30% price drop in Acast was identified on the 12th of February 2026. The anomaly had one matched article that contained three sentences, which noted increased turnover in the fourth quarter, listener growth in North America and Europe, and a negative impact on net profits from currency effects. However, despite the limited article contents, the LLM produced a relatively long explanation, which was presented in Listing 3. The explanation was considered speculative by one of the experts, who also noted that the LLM generated a long explanation based on little underlying information. He further noted that Acast's exposure to currency effects is well known to informed investors and would already be priced in, weakening the LLM's implicit statement that the currency-driven loss came as a surprise to market participants. Another one of the experts notes that the LLM lacks evidence to support its interpretation in the same explanation. Another *medium*-labelled explanation that an expert considered speculative was the one that explained the +3.55% price gain of Hoist Finance on the 16th of January 2026. The expert did however, note that the LLM includes the weakness in its own explanation: "to its (the LLMs) defence […] it points out that it could just as well have been other things". Furthermore, one of the experts commented on the +2.66% gain for Fagerhult Group on the 13th of January, which was labelled *low* confidence, that the move was too small and likely a normal daily fluctuation of the price.

5.3 Reproducibility

The LLM analysis task was run three times on the same input data using identical prompts and parameters, as described in Section 4.4.3. Two aspects of the output were compared across runs: the 0-10 article relevance scores and the verbalised confidence label per anomaly. In the article relevance scoring task, all 117 company-to-article pairs were assessed across three runs. 81 articles (69.2 %) received an identical score in all three runs, while 36 (30.8 %) received at least one varying score. Of these 36 articles, 29 varied by no more than ±1 point on the 0-10 scale. Furthermore, the LLM verbalised confidence of each explanation was assessed across three runs, and the stability of confidence scoring is presented in Table 15.

Table 15. Stability of confidence labels across three runs.

Change in confidence labelling	n	%
0 - the anomaly was labelled with the same confidence label	58	86.6 %
1 - the anomaly was labelled one level apart	9	13.4 %
2 - the anomaly was labelled two labels apart	0	0.00 %
3 - the anomaly was labelled three labels apart	0	0.00%

To conclude, the LLM analysis task is largely reproducible with minor fluctuations in the article relevance scoring and some degree of variance in verbalised confidence labelling. The variance in the output is consistent with the known stochasticity of LLM outputs at non-zero temperature (Wang et al., 2024).

6. Discussion

This chapter discusses the results presented in Chapter 5 in relation to the project aim of examining whether a general-purpose LLM can use news articles to explain significant Swedish stock price movements. Section 6.1 discusses the methodological choices that shaped these results and their limitations. Section 6.2 interprets the empirical findings from each pipeline step and returns to the research questions stated in Section 1.1.1. Section 6.3 considers the broader implications when implementing LLMs in financial applications that might impact decision making.

6.1 Methodological discussion

The method produced three main sets of empirical results: the share and distribution of statistically anomalous price movements, news articles matched to those movements, and the LLM's relevance scores, explanations and verbalised confidence labels and the expert evaluation of these outputs by financial professionals. The methodological choices and their implications on the results will be discussed in the following section. The main results and addressment of the research questions (i.e. output of the last step) will be discussed in section 6.2.

6.1.1 Identifying significant market movements

Significant market movements were identified using the volatility-normalised z-score, which served as the foundational dataset for the subsequent analysis steps. A rolling volatility estimation window of $w = 30$ days was used to account for clustered periods of volatility in the market, and the threshold $| z_t | \geq k$ was set at a permissive 1.96. For a Gaussian distribution, this threshold would flag 5 % of returns as extreme. However, 10.6 % of movements were flagged as anomalous, which clearly reflects the heavy-tailed characteristic of financial time series as proven by Cont (2001). This overestimation indicates that some flagged movements were noise rather than true anomalies, which was also raised by one of the financial professionals during the evaluation.

Further, the efficient market hypothesis is limited by the joint hypothesis problem due to price movements only being considered anomalous in relation to a chosen definition of expected return, for which there is no objective truth about which definition is correct (Fama, 1991). For the statistical method chosen here, anomalies were evaluated in relation to the expected return μ_w (the mean of the last 30 days) and normalised by the volatility σ_w of the last 30 days, which establishes a clear definition of what was considered "normal" within the method.

6.1.2 Company-to-article matching step

Linking each article to one or more companies in the dataset was essential, as the resulting matches served as input to the subsequent LLM analysis task. The Named Entity Linking task was performed by Claude Haiku 4.5, the smallest and least capable model available

from Anthropic at the time of the experiment. Although some limitations were observed, notably the incorrect article matches for companies with similar names (e.g. Volvo B and Volvo Car B), the approach was more reliable than the naïve substring matching. These limitations could likely be mitigated through further iteration on prompt design or by using a higher-capacity model on the same dataset.

A notable finding was the gap in company-to-article matches across the sampled stocks. Of the 547 identified anomalous price movements, only 67 received a match among the available articles. The gap could potentially be explained by multiple factors. Firstly, statistical noise identified as anomalous price movements by the z-score method might have impacted the number of matches, as such movements are less likely to have received coverage among the financial news articles. Second, the anomalies might have been covered in other news sources than the printed DI. Third, some anomalous price movements may have been covered by DI but not captured in the matched dataset by the NEL step, which, as previously demonstrated, was limited. The temporal matching window of 168 hours (7 days), however, appears to have been large enough, as there were no matches more than 4 days prior to an anomalous price movement. Unfortunately, it is not possible to distinguish between these explanatory factors in a reliable way with the available data.

Out of the 428 company-to-article pairs that the NEL step resulted in, only 91 unique articles fell within seven days of an anomaly for their matched company, indicating that the majority of news articles about the sampled companies were not temporally clustered around anomalous price movements. Furthermore, the temporal distribution of matched articles demonstrated that 37.6 % were published on the same trading day as the identified anomalous price movement, and no matched article was published more than four days prior to the movement. This same-day clustering is consistent with the semi-strong form of the Efficient Market Hypothesis (Fama, 1970), and suggests that prices adjust fast to publicly available information, with limited subsequent drift. The results could further be argued to align with the recent suggestive findings by Lopez-Lira and Tang (2025) that the growing use of LLMs in financial analysis may inflate the speed at which public information is reflected in prices. However, the methodological approach and results of this thesis are insufficient to draw any conclusions regarding increased market efficiency, and thus remain an open question for future research.

A further notable result of the company-to-article matching step was that 59.7 % of matched anomalies (40 of 67) were linked to only a single article. This constrains the subsequent LLM analysis task, as the model has limited evidence over which to reason, which in turn may increase the risk of unsupported claims or hallucination. The risk was illustrated by one of the expert-evaluated explanations, in which the expert observed that the generated explanation drew broader conclusions than the underlying article supported, despite the article containing little information directly relevant to the company (Acast).

6.2 Interpretation of results of the main LLM Analysis task

The LLM analysis task produced three outputs per anomaly: a 0-10 relevance score for each matched article, a 3-5 sentence narrative explanation, and a verbalised confidence label. The article relevance scoring and the distribution of confidence labels are discussed in this section based on the model outputs, while the quality of the narrative explanations and confidence labels against expert ratings are further discussed in Section 6.2.4 alongside the results of the expert evaluation form.

6.2.1 Article relevance scoring

Out of the 117 company-to-article pairs produced by the company-to-article matching task, 59 % were scored lower than 5 on the relevance scoring task, indicating that for a majority of articles, the model did not consider the article the clear driver of the move. The distribution of scores cannot reveal why certain articles were ranked lower, and it is therefore not possible to determine whether there exists a systematic bias in what the model considers explanatory of a price movement. Future work could combine the relevance scores with topic analysis, which has previously been demonstrated to capture relationships between news content and market movements that sentiment-based methods miss (Chen, 2025; Feuerriegel and Pröllochs, 2021).

6.2.2 Verbalised confidence

After the 3-5 sentence narrative explanation had been generated, the model was prompted to assign a confidence label to the explanation. Five explanations (7.5 %) received the label high, 22 (32.8 %) medium, 19 (28.4 %) low, and 21 (31.4 %) none. The permissive anomaly threshold (k = 1.96) could be expected to over-flag movements that reflect general market activity and therefore produce a large share of *none* labels for which no explanatory article exists within the temporal window. However, the matched and explained anomalies across the confidence labels do not support this interpretation. The none-labelled anomalies demonstrate a mean return of 4.11 % and a mean z-score of 3.16, which is comparable to the medium and low labels and well above the detection threshold of 1.96. This indicates that the *none*-label is not primarily a consequence of a weak or noisy price signal, but rather of a limited article base, as the share of anomalies matched to one article rises from 40 % for high to 86 % for none. The interpretation is further supported by the largest price movement in the entire dataset (Volvo Car B, +25.53 % on February 6th 2026), which was assigned a none label, as no relevant article was matched to the movement. This suggests that the absence of an explanation reflects the lack of matched articles rather than noise in the detected movement.

On the other hand, explanations labelled with high confidence are demonstrated a mean return of 9.15 %, notably larger than the 4.65 %, 3.81 % and 4.11 % of the medium, low and none labels respectively. This could indicate that larger movements are more likely related to major corporate events that received coverage in the financial news. However, as several high confidence explanations were supported by only a single article, and

several low confidence explanations relied on multiple matched articles, the relationship cannot be directly proven. The available data is therefore not sufficient to draw conclusions about the relationship between the number of matched articles and the resulting confidence label. This suggests that the model's verbalised confidence reflects the relevance of the article base more than the size of the price movement, or the number of articles.

6.2.3 Reproducibility

To assess the algorithm for reproducibility, the analysis task was ran three separate times. The article relevance scoring task and verbalised confidence were directly comparable across runs. At temperature 0.2, 69% of article relevance scores were consistent across the three runs, and 94% were within ±1 point on the 0-10 scale. For the verbalised confidence task, the results were equally promising, as 86.6 % of anomalies received the same confidence label across the three separate runs. To conclude, the LLM analysis task is largely reproducible with minor fluctuations in the article relevance scoring and some degree of variance in verbalised confidence labelling. The variance in the output is consistent with the known stochasticity of LLM outputs at non-zero temperature (Wang et al., 2024).

6.2.4 Expert evaluation form

The expert evaluation of the AI-generated explanations broadly aligned with the confidence labels assigned by the LLM. Explanations labelled with *high* confidence received consistently high ratings across all three dimensions, with no individual rating below 4 and a mean *factual correctness* of 4.78, whereas ratings were consistently lower for the *medium* and *low* confidence groups, particularly for the *relevance* and *usefulness* dimensions. The single explanation labelled *none* received the lowest possible mean scores for both *relevance* and *usefulness*. However, the alignment between expert ratings and LLM confidence was not consistent across the three dimensions. The *relevance* of the cited articles and the *usefulness* of the explanation largely corresponded with the verbalised confidence label, whereas *factual correctness* remained relatively high even for explanations labelled *low* or *none*. This indicates that the model's verbalised confidence reflected the strength and relevance of the underlying evidence more than the factual accuracy of the explanation, which is consistent with the observation that a *none*-labelled explanation can still be considered factually correct as it transparently admits that the underlying evidence is insufficient, rather than fabricating an incorrect cause. To conclude, the results provide support for RQ2, suggesting that the LLM's verbalised confidence aligns with the explanation quality as judged by financial professionals, although the strength of the results is limited by the small sample of ten anomalies evaluated by three experts.

The article relevance scores assigned by the experts and the LLM broadly corresponded, with the closest agreement at the lower end of the scale, which indicates that the experts and the model largely agreed on which articles were unrelated to a given price movement.

However, the LLM tended to assign higher relevance scores than the experts, and the discrepancy was most pronounced among the articles that the LLM scored highly, several of which the experts considered substantially less relevant to the price movement. This indicates a tendency of the LLM to overestimate explanatory relevance to the matched articles, which is consistent with the free-text observation that the model produced explanations exceeding what the underlying articles provided evidence for.

Although the expert ratings generally aligned with the LLM's confidence labels on average, there were some disagreements between the three experts on individual explanations. One example was the medium-confidence explanation for Volvo B (−5.80 % on March 19th 2026), which was rated from 5 to 2 by the three experts on all three dimensions. The same pattern appeared for the none-confidence explanation for Humana (−2.78 % on 26 January 2026), where factual correctness received one rating of 1 and two ratings of 5. This likely indicates different interpretations of what constitutes a factually correct explanation. While the ratings capture the overall alignment between expert judgement and the model's output, the experts' free-text comments offer a more nuanced view of the strengths and limitations of individual explanations, which are examined in the following section.

6.2.5 Free-text comments

Among the explanations labelled with *high* confidence and supported by more than one matched article, the experts noted that the LLM based its explanation closely on the available articles. For Boliden's −9.00 % movement on 19 March 2026, the explanation relied entirely on the matched articles and did not account for the broader macroeconomic and geopolitical conditions affecting copper and gold prices during the early spring of 2026, despite Boliden's strong exposure to precious metals. Two factors could potentially explain this limitation. First, the knowledge cutoff of Claude Sonnet 4.6 is May 2025 (Anthropic, 2026), which means the model could not have drawn on factual knowledge of price developments of gold and copper during the sample period, and any such reasoning would have been speculative. Second, no article covering these macroeconomic conditions was included in the matched set, as the company-to-article matching step only identified articles with the company as a primary subject. A more complete explanation would thus have required broader article coverage, which falls outside the scope of this thesis and remains an open question for future research.

A related pattern was observed for Electrolux B (+7.33 % on 2 February 2026), where one expert noted that the Q4 earnings report referenced in two of the three matched articles likely explained the move completely, while the LLM also drew on a third article describing the CEO's optimism following US tariff developments. In this case, the model did not weight the articles by their explanatory relevance but instead distributed its reasoning across all matched articles. Furthermore, the articles mentioned a +15 % gain for Electrolux on Friday, 30 January, suggesting that the +7.33 % movement on 2 February may reflect a continued price reaction on the following trading day rather than a new event. A similar drift was identified for Volvo B (−5.80 % on 19 March 2026),

where the matched article referenced a −2.3 % loss on the previous trading day. This raises a broader methodological limitation of the company-to-article matching step, as it cannot distinguish between articles explaining the movement on the day they were published and articles explaining a delayed price reaction, which is consistent with the concept of drift discussed by Fama (1970) in the context of slow price adjustment to new information.

Among the medium confidence explanations, the experts noted that the LLM tended to produce explanations that exceeded what the underlying articles provided evidence for. For Acast (−4.30 % on February 12th 2026), one expert characterised the explanation as speculative and observed that the LLM generated a long explanation from a three-sentence article. The same expert further noted that Acast's exposure to currency effects is likely known to informed investors and would therefore already be reflected in the price, which weakens the LLM's claim that the currency-driven loss came as a surprise to the market. This was additionally noted by another one of the experts. A second *medium*-labelled explanation, the +3.55 % gain for Hoist Finance on 16 January 2026, was considered speculative by one of the experts, who, however, noted that the LLM acknowledged this uncertainty within the explanation itself by pointing out that other factors could have driven the move. The latter is consistent with the model's instruction to assign a *medium* label when the explanation is reasonable but partial and further suggests that the model's verbalised confidence could reflect the strength of the underlying evidence.

Among the explanations labelled with *low* confidence, one expert commented on the +2.66 % gain for Fagerhult Group on 13 January 2026, observing that the move was small and likely reflected normal daily fluctuation rather than a company-specific event. This observation aligns with the discussion of the permissive anomaly threshold, which by design flags a relatively large share of movements as anomalous and is therefore expected to include some statistical noise.

6.3 Implications of results and future research

The results of the separate steps in the method, together demonstrate the diverse set of roles a general-purpose LLM might have in financial applications. A central aspect of the method is that a general-purpose LLM was applied to perform several different tasks that each has both general and finance-specific applications, without domain-specific fine-tuning. In the company-to-article matching step, the model extracted structured company attributes such as name and ISIN from unstructured Swedish news text, which is consistent with the findings of Dolphin et al. (2024) that LLMs are effective in extracting structured metadata from financial news articles. The article relevance scoring constitutes a text classification problem, a task that has long been central to financial NLP (Kadre et al., 2025) for both sentiment categorisation and other tasks. Within the method of this thesis, the model is required to assess relevance in relation to a specific price movement, which is arguably more complex than predefined sentiment categories. The generation of a narrative explanation further required the model to reason in its outputs about market mechanisms and how different types of information are reflected in asset prices, which connects to the semi-strong form of the Efficient Market Hypothesis (Fama, 1970) and the speed of price adjustment across information types as demonstrated by Lopez-Lira and Tang (2025). Finally, the verbalised confidence label revealed a probability measure within the otherwise inaccessible output of a closed-source model, which partially addresses the black-box limitation of commercial LLMs (Xiong et al., 2024). This demonstrates the capabilities of a general-purpose LLM to perform several separate tasks within one solution and aligns with previous findings that general-purpose LLMs perform competitively on financial NLP tasks despite not being explicitly trained or fine-tuned in the domain (Fatouros et al., 2023; Ardekani et al., 2024).

Previous NLP techniques, such as bag-of-words and dictionary-based methods, were limited to determining the sentiment of a text (Loughran and McDonald, 2016), and transformer models, such as FinBERT finetuned in the financial domain, remained focused on text classification (Araci, 2019) rather than "reasoning" over the underlying content. Although no effort was made to assess the performance of LLM as compared to previous techniques within the scope of this thesis, the explanatory task as developed in this thesis would not have been within the capacity of previous technology.

An observation from the expert evaluation was that the model tended to overestimate the relevance of the matched articles and to generate explanations exceeding what the underlying articles supported. This could have two potential explanations. Firstly, the commercial LLMs tends to favour outputs that have a higher probability of aligning with the user's expectations and opinions due to reinforcement learning from human feedback (RLHF). The design of the task itself, where a price movement is provided together with a set of articles, might introduce such an expectation for which the model adjusts its outputs to. This is a limitation of the chosen explanatory rather than predictive approach, since presenting the model with a known price movement introduces an expectation that a predictive framing could easily avoid.

Further, it could be because the model is unaware of broader macroeconomic and geopolitical conditions that may have caused a price movement, as illustrated by the Boliden explanation, and therefore does not consider any external factors not covered by the matched articles. On the contrary, this reduces the risk of look-ahead bias (Glasserman & Lin, 2023) that potentially otherwise could have impacted the results. This further raises a question for future research to extend the underlying article base beyond company-specific content to include coverage of broader macroeconomic and geopolitical factors, which may affect different sectors and industries in separate ways. Additionally, several large movements detected in the anomaly detection step were left unexplained (including the largest move during the time period, Volvo Car B of +25.53%). A bigger set of news articles with broader coverage, such as quarter reports, tweets or indirectly related news, such as geopolitical and macroeconomic, could have captured the broader, underlying information that drove significant movements.

The results further have broader implications for the use of LLM-based algorithms in financial applications that target both institutional and retail investors. As the generated explanations have the potential to impact investment decisions, the trust and explainability of the outputs become central, particularly for retail investors who may lack the knowledge to verify the outputs themselves. In this context, the risk of hallucinations is especially apparent, and the verbalised confidence label therefore constitutes an important mechanism to mitigate this risk. As the model's verbalised confidence was shown to align with the expert-rated relevance and usefulness of the explanations to a retail investor. Further, the explanations that were labelled with none- or low-confidence often included a statement that the evidence was insufficient to draw further conclusions.

7. Conclusion

The relationship between information and financial markets has long been studied extensively (Fama, 1970), and financial news constitutes a primary mechanism through which such information is distributed (Bybee et al., 2024). This thesis investigates the relationship between news and fluctuations in the financial market in an explanatory way rather than predictive, which has constituted the main approach among the related work in the area. Consequently, the aim of this thesis was to examine the capability of a general-purpose LLM to use news articles to generate a narrative explanation of significant market movements among Swedish listed stocks. Two research questions guided the thesis:

- **RQ1**: To what extent can a general-purpose LLM use financial news articles to generate a narrative explanation of a significant market movement?
- **RQ2**: Does an LLM's verbalised confidence in its explanations align with the explanation quality as judged by financial professionals?

The method to answer the research questions consisted of a technical experimental setup that handled two types of input data, data processing and the main explanation step in four separate steps. Firstly, a volatility-normalised z-score with anomaly threshold of 1.96 was used to identify 10.6 % of movements as significant. Secondly, financial news articles from DI were linked to Swedish listed companies through a Named Entity Linking step performed by Claude Haiku 4.5. Third, for each identified movement, Claude Sonnet 4.6 was prompted with articles matched to the company within a seven-day window, and asked to score each article's relevance, generate a narrative explanation, and assign a verbalised confidence label. Finally, since no objective truth exists for which articles explain a given movement, the outputs were evaluated by three financial professionals through a structured form where they were asked to provide an article relevance score on a scale of 0-10, and rate the explanation along three dimensions: factual correctness, the article’s relevance to the observed price movement, and usefulness to a retail investor.

RQ1 was addressed through the generated explanations, the article relevance scoring, and the experts' quality judgements. The experts judged the explanations to be factually correct, and those supported by stronger article evidence and therefore labelled with high confidence were also rated relevant to the price movement and useful to a retail investor by the experts. The LLM's article relevance scores broadly aligned with the ratings of the experts, although the LLM tended to assign higher relevance scores in general. These results imply that a general-purpose LLM can generate narrative explanations about significant market movements based on financial news articles that align with expert judgement to a meaningful extent. Notably, factual correctness was consistently rated high across confidence labels, despite the demonstrated risk of hallucinations in previous research.

The discussion of the results surfaced two limitations regarding the capability of a general-purpose LLM to generate the narrative explanation. Firstly, the LLM tended to produce explanations that exceeded what the underlying articles provided evidence for.

This could be due to reinforcement learning from human feedback (RLHF), as presenting the model with a known price movement together with a set of articles introduces an expectation that an explanation exists, for which the model is, by design, encouraged to comply with. This is a limitation of the explanatory approach and could be mitigated by taking a predictive one or by iterating on prompt design. Secondly, the explanations were limited to be based on company-specific articles and thus did not account for broader macroeconomic and geopolitical conditions, which resulted in few explanations of high confidence among the sampled stocks.

RQ2 examined whether the LLM's verbalised confidence aligned with the explanation quality as judged by financial professionals and was addressed by comparing the confidence labels against the expert ratings. The results demonstrated that explanations labelled with high confidence received the highest expert-ratings across all three dimensions, and ratings declined for the medium- and low-labelled explanations, and the none-labelled explanation received the lowest mean ratings for relevance and usefulness. The alignment was strongest for relevance and usefulness, and weakest for factual correctness, which remained relatively high even among explanations labelled low or none. The results thus provide support for RQ2, although the strength of this conclusion is limited by the small sample of ten anomalies evaluated by three experts.

The discussion further identified several areas for future research. Firstly, the articles used could be expanded to include additional types of financial documents, such as earning reports, to capture the underlying information that drove some of the significant movements that were left unexplained within the scope of this thesis. Further, broader news coverage such as geopolitical or macroeconomic events could be included to capture a more complete picture of why some stocks move that are closer related to such factors. Further, the article relevance scoring could be combined with LLM-powered topic analysis, which has previously been demonstrated to capture relationships between news content and market movements that sentiment-based methods miss (Feuerriegel and Pröllochs, 2021; Chen, 2025).

References

Anthropic (2026) *Anthropic's transparency hub*. Anthropic. Available at: https://www.anthropic.com/transparency/model-report [Accessed: 2026-05-26].

Anthropic (n.d.a) *Models overview.* Anthropic. Available at: https://platform.claude.com/docs- /en/about-claude/models/overview [Accessed: 2026-05-12].

Anthropic (n.d.b) *Glossary*. Anthropic. Available at: https://platform.claude.com/docs/en/about-claude/glossary [Accessed: 2026-05-14].

Anthropic (n.d.c) *Prompting best practices.* Anthropic. Available at: https://platform.claude.com- /docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices [Accessed: 2026-05-15].

Anthropic (n.d.d) *Multilingual Support*. Anthropic. Available at: https://platform.claude.com/docs/en-/build-with-claude/multilingual-support [Accessed: 2026-05-26].

Araci, D. (2019) Finbert: Financial sentiment analysis with pre-trained language models [Preprint]. *arXiv:1908.10063*.

Ardekani, A. M. et al. (2024) FinSentGPT: A universal financial sentiment engine? *International Review of Financial Analysis*. 94.

Avanza (n.d.) *Handla aktier*. Avanza. Available at: https://www.avanza.se/aktier/handla.html/-kategorier. [Accessed: 2026-05-14]

Avanza Bank Holding AB (2026) *Interim Report January–March 2026*. Stockholm: Avanza Bank Holding AB (publ). Available at: https://investors.avanza.se/files/mfn/dc7ed913-ffbc-484b-a861-17c36b6c8295/avanza-bank-holding-ab-publ-interim-report-januarymarch-2026.pdf [Accessed: 2026-05-14]

Bai, Y. et al. (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback [Preprint]. *arXiv:2204.05862.*

Barber, B. M. & Odean, T. (2008) All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*. 21(2), pp. 785–818.

Beaudry, P. & Portier, F. (2014) News-driven business cycles: Insights and challenges. *Journal of economic literature.* 52(4), pp. 993–1074.

Beurer-Kellner, L., Fischer, M. & Vechev, M. (2023) Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*. 7(PLDI), pp. 1946-1969.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) Latent dirichlet allocation. *Journal of machine learning research*. 3(Jan), pp. 993-1022.

Brown, T. B. et al. (2020) Language models are few-shot learners [Preprint]. *arXiv:2005.14165.*

Bybee, L. et al. (2024) Business news and business cycles. *The journal of finance*. 79(5), pp. 3105-3147.

Chen, Q. (2025) A framework for measuring how news topics drive stock movement [Preprint]. *arXiv:2510.06864.*

Chiang, C. H. & Lee, H. Y. (2023) Can large language models be an alternative to human evaluations? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Volume 1: Long Papers. pp. 15607-15631.

Cont, R. (2001) Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*. 1(2), pp. 223–236.

Dagens Industri (2025) *Om oss*. Dagens Industri. Available at: https://www.di.se/nyheter/om-oss/ [Accessed: 2026-05-14].

Devlin, J. et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding [Preprint]. *arXiv:1810.04805.*

Dolphin, R. et al. (2024) Extracting structured insights from financial news: An augmented LLM driven approach [Preprint]. *arXiv:2407.15788.*

Du, K. et al. (2025) Natural language processing in finance: A survey. *Information fusion*. 115.

Engelberg, J. E. & Parsons, C. A. (2011) The causal impact of media in financial markets. *The journal of finance.* 66(1), pp. 67–97.

Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica.* 50(4), pp. 987–1007.

Euroclear Sweden (2026) *Aktieägarstatistik Q1 2026*. Stockholm: Euroclear Sweden. Available at: https://www.euroclear.com/dam/ESw/Brochures/-Documents_in_Swedish/Euroclear_Swedens_Aktieagarstatistik-Q1-2026.pdf [Accessed: 2026-06-01].

Fama, E. F. (1970) Efficient capital markets: A review of theory and empirical work. *The journal of finance*. 25(2), pp. 383-417.

Fama, E. F. (1991) Efficient capital markets: II. *The journal of finance*. 46(5), pp. 1575-1617.

Fang, L. & Peress, J. (2009) Media coverage and the cross-section of stock returns. *The journal of finance*. 64(5), pp. 2023-2052.

Fatouros, G. et al. (2023) Transforming sentiment analysis in the financial domain with ChatGPT. *Machine learning with applications*. 14.

Feuerriegel, S. & Pröllochs, N. (2021) Investor reaction to financial disclosures across topics: An application of latent dirichlet allocation. *Decision sciences.* 52(3), pp. 608–628.

Gao, R. et al. (2021) ‘A review of natural language processing for financial technology’, in Shota Nakashima et al. (eds.). 2021 SPIE. pp. 118840U-118840U – 16.

Glasserman, P. & Lin, C. (2023) Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis [Preprint]. *arXiv:2309.17322*.

Gruver, N. et al. (2023) Large language models are zero-shot time series forecasters. *Advances in neural information processing systems.* 36, pp. 19622-19635.

Guo, T. & Hauptmann, E. (2024) Fine-tuning large language models for stock return prediction using newsflow [Preprint]. *arXiv:2407.18103.*

Kadavath, S. et al. (2022) Language models (mostly) know what they know [Preprint]. *arXiv:2207.05221*.

Kadre, S., Kadre, S. & Dey, S. (2025) *Mastering Text Analytics: A Hands-on Guide to NLP Using Python*. Berkeley, CA: Apress.

Kaplan, J. et al. (2020) Scaling laws for neural language models [Preprint]. *arXiv:2001.08361*.

Kenton, W. (2024) *Understanding SEC Form 8-K: Key Events and Filing Guide.* Investopedia. Available at: https://www.investopedia.com/terms/1/8-k.asp. [Accessed: 2026-05-26].

Khurana, D. et al. (2023) Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*. 82(3), pp. 3713–3744.

Kirtac, K. & Germano, G. (2024) Sentiment trading with large language models. *Finance research letters*. 62.

Kong, A. et al. (2024) Better zero-shot reasoning with role-play prompting [Preprint]. *arXiv:2308.07702.*

Lai, V. D. et al. (2023) ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning [Preprint]. *arXiv:2304.05613.*

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

Li, Y. et al. (2023) ‘Large language models in finance: A survey’, in *Proceedings of the fourth ACM international conference on AI in finance*. 27 November 2023 New York, NY, USA: ACM. pp. 374–382.

Liu, B. (2012) *Sentiment analysis and opinion mining*. San Rafael, Calif. Morgan & Claypool.

Lopez-Lira, A. & Tang, Y. (2025) Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models [Preprint]. *arXiv:2304.07619.*

Loughran, T. & McDonald, B. (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The journal of finance*. 66(1), pp. 35-65.

Loughran, T. & McDonald, B. (2016) Textual analysis in accounting and finance: A survey. *Journal of accounting research*. 54(4), pp. 1187–1230.

Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space [Preprint]. *arXiv:1301.3781.*

Mishra, S. et al. (2022) Reframing instructional prompts to gptk's language [Preprint]. *arXiv:2109.07830*.

Nigam, K. et al. (2000) Text classification from labelled and unlabelled documents using EM. *Machine learning*. 39(2–3), pp. 103–134.

Ouyang, L. et al. (2022) Training language models to follow instructions with human feedback. *Advances in neural information processing systems.* 35, pp. 27730-27744.

Pelster, M. & Val, J. (2024) Can ChatGPT assist in picking stocks? *Finance research Letters*. 59.

Radford, A. et al. (2018) *Improving language understanding by generative pre-training.* OpenAI. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [Accessed: 2026-04-13].

Radford, A. et al. (2019) *Language models are unsupervised multitask learners.* OpenAI. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [Accessed: 2026-04-03].

Sahoo, P. et al. (2024) A systematic survey of prompt engineering in large language models: Techniques and applications [Preprint]. *arXiv:2402.07927.*

Salewski, L. et al. (2023) In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems.* 36, pp. 72044-72057.

Sclar, M. et al. (2024) Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting [Preprint]. In *International Conference on Learning Representations.* Vol. 2024, pp. 25055-25083. *arXiv:2310.11324.*

Sveriges Riksbank (2025) *Finansiell stabilitet 2025:2*. Stockholm: Sveriges Riksbank. Available at: https://www.riksbank.se/globalassets/media/rapporter/fsr-/svenska/2025/251113/finansiell-stabilitetsrapport-2025_2.pdf [Accessed: 2026-05-14].

Tam, Z. R. et al. (2024) Let me speak freely? A study on the impact of format restrictions on performance of large language models [Preprint]. *arXiv:2408.02442*.

Tan, P. N. et al. (2020) *Introduction to data mining*. 2nd ed. Harlow: Pearson.

Tetlock, P. C. (2007) Giving content to investor sentiment: The role of media in the stock market. *The journal of finance*. 62(3), pp. 1139-1168.

Tian, K. et al. (2023) Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback [Preprint]. *arXiv:2305.14975.*

Tong, H. et al. (2024) Ploutos: Towards interpretable stock movement prediction with financial large language model [Preprint]. *arXiv:2403.00782.*

Tsay, R. S. (2010) *Analysis of financial time series*. 3rd ed. Hoboken, N.J: Wiley.

Van der Lee, C. et al. (2021) Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language.* 67.

Vaswani, A. et al. (2017) 'Attention is all you need', in *Advances in Neural Information Processing Systems 30* (*NeurIPS 2017)*. 4–9 December 2017 Long Beach, CA, USA: Curran Associates, Inc. pp. 5998–6008.

Wang, M., Izumi, K. & Sakaji, H. (2024) LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction [Preprint]. *arXiv:2406.10811*.

Wei, J. et al. (2022) 'Chain-of-thought prompting elicits reasoning in large language models', in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022).* 28 November–9 December 2022 New Orleans, LA, USA: Curran Associates, Inc. pp. 24824–24837.

Wu, S. et al. (2023) BloombergGPT: A large language model for finance [Preprint]. *arXiv:2303.17564*.

Xiong, M. et al. (2024) Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs [Preprint]. *arXiv:2306.13063.*

Yang, H. et al. (2025) FinGPT: Open-source financial large language models [Preprint]. *arXiv:2306.06031.*

Young, T. et al. (2017) Recent trends in deep learning based natural language processing [Preprint]. *arXiv:1708.02709.*

Zheng, M. et al. (2024) When ”a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models [Preprint]. *arXiv:2311.10054.*

Appendices

A – Selected companies per industry

ISIN	COMPANY NAME	SECTOR
SE0023313762	Apotea	Consumer, Retail & Services
SE0009888738	Boozt	Consumer, Retail & Services
SE0002626861	Cloetta	Consumer, Retail & Services
SE0016589188	Electrolux B	Consumer, Retail & Services
SE0000106270	H&M B	Consumer, Retail & Services
SE0002110064	MEKO	Consumer, Retail & Services
SE0015245535	Nelly Group	Consumer, Retail & Services
SE0020848356	Rusta	Consumer, Retail & Services
SE0005999760	Scandi Standard	Consumer, Retail & Services
SE0000163594	Securitas B	Consumer, Retail & Services
SE0006422390	Thule Group	Consumer, Retail & Services
SE0016844831	Volvo Car B	Consumer, Retail & Services

Table 16. Consumer, Retail & Services

ISIN	COMPANY NAME	SECTOR
CA5889141019	Meren Energy	Energy / Natural Resources
SE0000825820	Orrön Energy	Energy / Natural Resources

Table 17. Energy / Natural Resources

ISIN	COMPANY NAME	SECTOR
SE0015661236	Creades A	Finance & Investment
SE0012853455	EQT AB	Finance & Investment
SE0007100599	Handelsbanken A	Finance & Investment
SE0006887063	Hoist Finance	Finance & Investment
SE0000936478	Intrum	Finance & Investment
SE0015811963	Investor B	Finance & Investment
FI4000297767	Nordea Bank	Finance & Investment
SE0017831795	Norion Bank	Finance & Investment
SE0000111940	Ratos B	Finance & Investment
SE0000148884	SEB A	Finance & Investment
SE0017161458	Svolder B	Finance & Investment
SE0000242455	Swedbank A	Finance & Investment
SE0025666969	TF Bank	Finance & Investment
SE0014428835	VNV Global	Finance & Investment
SE0008321608	Öresund	Finance & Investment

Table 18. Finance & Investment

ISIN	COMPANY NAME	SECTOR
SE0009663826	Ambea	Healthcare, Pharma & Care
SE0010468116	Arjo	Healthcare, Pharma & Care
GB0009895292	AstraZeneca	Healthcare, Pharma & Care
SE0007666110	Attendo	Healthcare, Pharma & Care
SE0017769995	BioGaia B	Healthcare, Pharma & Care
SE0009858152	Bonesupport	Healthcare, Pharma & Care
SE0000683484	CellaVision	Healthcare, Pharma & Care
SE0002148817	Hansa Biopharma	Healthcare, Pharma & Care
SE0008040653	Humana	Healthcare, Pharma & Care
SE0000135485	RaySearch Laboratories B	Healthcare, Pharma & Care
SE0007577895	Vicore Pharma Holding	Healthcare, Pharma & Care
SE0004840718	Xvivo Perfusion	Healthcare, Pharma & Care

Table 19. Healthcare, Pharma & Care

ISIN	COMPANY NAME	SECTOR
CH0012221716	ABB	Industrials
SE0017486897	Atlas Copco B	Industrials
SE0000862997	Billerud	Industrials
SE0010048884	Fagerhult Group	Industrials
SE0017483506	Instalco	Industrials
SE0017160773	NCAB Group	Industrials
SE0011204510	Nederman Holding	Industrials
SE0015988019	Nibe Industrier B	Industrials
SE0000112385	Saab B	Industrials
SE0000667891	Sandvik	Industrials
SE0003756758	Sdiptech	Industrials
SE0000115446	Volvo B	Industrials

Table 20. Industrials

ISIN	COMPANY NAME	SECTOR
SE0020050417	Boliden	Mining

Table 21. Mining

ISIN	COMPANY NAME	SECTOR
SE0015960935	Acast	Tech, Digital & Media
DK0060952240	Better Collective	Tech, Digital & Media
SE0000108656	Ericsson B	Tech, Digital & Media
SE0012673267	Evolution	Tech, Digital & Media
SE0018040677	Hexatronic Group	Tech, Digital & Media
SE0012323715	Karnov Group	Tech, Digital & Media
SE0000421273	Knowit	Tech, Digital & Media
SE0011870195	Lime Technologies	Tech, Digital & Media
SE0016074249	MilDef Group	Tech, Digital & Media

SE0018012494	MTG B	Tech, Digital & Media
SE0010948711	Ovzon	Tech, Digital & Media
SE0006425815	PowerCell Sweden	Tech, Digital & Media
SE0015346135	Stillfront Group	Tech, Digital & Media
SE0000667925	Telia Company	Tech, Digital & Media
SE0016787071	Truecaller	Tech, Digital & Media
SE0012116390	Viaplay Group B	Tech, Digital & Media

Table 22. Tech, Digital & Media

B – Prompts for company-to-article matching

```
You are a precise classifier for Swedish financial news. You will be
given a list of tracked publicly listed companies and one news article.
Identify which of the tracked companies the article actually references.
Distinguish PRIMARY-SUBJECT mentions (the article is about the company,
its results, its strategy, its leadership, or a direct corporate event)
from PASSING mentions (the company is named only as context, in a list,
or as an example).



Output rules:

Return ONLY tracked companies — never make up an ISIN.

An article may reference 0, 1, or several tracked companies.

Provide a short evidence quote (≤120 chars) lifted verbatim from the
article for each company you list.

A passing mention of a Swedish word that happens to start with a
company's letters (e.g. 'snabb' / 'drabbat' / 'rusta') does NOT count
as a mention of ABB or Rusta. Match the actual company, not
coincidental letter sequences.

Output strictly valid JSON matching the schema. No prose.
```

Listing 4. System prompt 1.

```
TRACKED COMPANIES (ISIN | Symbol | Name | Sector):

CA5889141019 | AOI    | Meren Energy      | Energy / Natural Resources

CH0012221716 | ABB    | ABB               | Industrials

DK0060952240 | BETCO  | Better Collective | Tech, Digital & Media

FI4000297767 | NDA SE | Nordea Bank       | Finance & Investment

GB0009895292 | AZN    | AstraZeneca       | Healthcare, Pharma & Care

SE0000106270 | HM B   | H&M B             | Consumer, Retail &
Services

... 64 more lines ...

SE0025666969 | TFBANK | TF Bank           | Finance & Investment
```

Listing 5. System prompt 2.

```
ARTICLE

Published: {published}

Source: {source}



Title: {title}



Body:

{body}



RESPONSE SCHEMA (JSON only)

{

  "companies": [

    {

      "isin": "<ISIN of a tracked company>",

      "is_primary_subject": <true|false>,

      "evidence_quote": "<≤120 chars lifted verbatim from the
article>"

    }

  ]

}

If the article references no tracked company, return {"companies":
[]}.
```

Listing 6.User prompt and constrained JSON output.

C – Prompts for the LLM Explanation task

```
You are a senior financial analyst specialising in Swedish listed
equities. Respond only with valid JSON.
```

Listing 7. System prompt for the LLM Explanation task.

```
A stock price anomaly was detected for {company} ({ticker}) on {date}.

The stock {direction} by {change:+.2f}% (z-score: {z_score}.

Below are {n} news articles matched to this anomaly.

--- Article {idx} ---

Title: {title}

URL: {url}

Source: {source}

Published: {published_at}

Time relative to anomaly: {time_delta_hours:.1f} hours

Content:

{content} ← truncated to 2000 chars, suffixed "... [truncated]" if cut

Your task:

1. Score each article's relevance to this specific price movement on a
scale of 0-10.

   - 10 = directly explains the price move (e.g. earnings surprise,
major deal announced)

   - 5  = related but not the clear driver (e.g. general sector news)

   - 0  = unrelated or a time-window coincidence

2. Write a concise 3-5 sentence narrative explaining why the stock
moved, citing the most relevant articles by their index number (e.g.
[Article 2]).

3. Assign an overall confidence level for THE EXPLANATION you just
wrote — considering whether it is factually accurate, adequately
supported by the cited articles, and an adequate account of what
likely drove the move:

   - "high"   — the explanation is well-supported by the articles,
factually grounded, and adequately accounts for the move

   - "medium" — the explanation is plausible but partial: incomplete
coverage, weaker supporting articles, or unverified detail

   - "low"    — the explanation is weakly supported, speculative, or
relies on tenuous article connections

   - "none"   — no meaningful explanation can be given from these
articles
```

Listing 8. User prompt template. Curly-braced names are placeholders substituted at runtime, i.e. "direction" resolves to "rose" or "fell" depending on the sign of the daily change.

```
 Return ONLY the following JSON object, no other text.

The article_evaluations array must include ALL {n} articles, sorted by
rank ascending.



{
  "article_evaluations": [
    {
      "rank": 1,
      "relevance_score": 9.0,
      "reasoning": "One sentence explaining the score.",
      "url": "https://...",
      "title": "Article title here"
    }
  ],
  "explanation": "Narrative text citing [Article N] ...",
  "confidence": "high"
}
```

Listing 9. Constrained JSON output schema appended to the user prompt. The confidence field must be one of {high, medium, low, none}.

D – Evaluation form

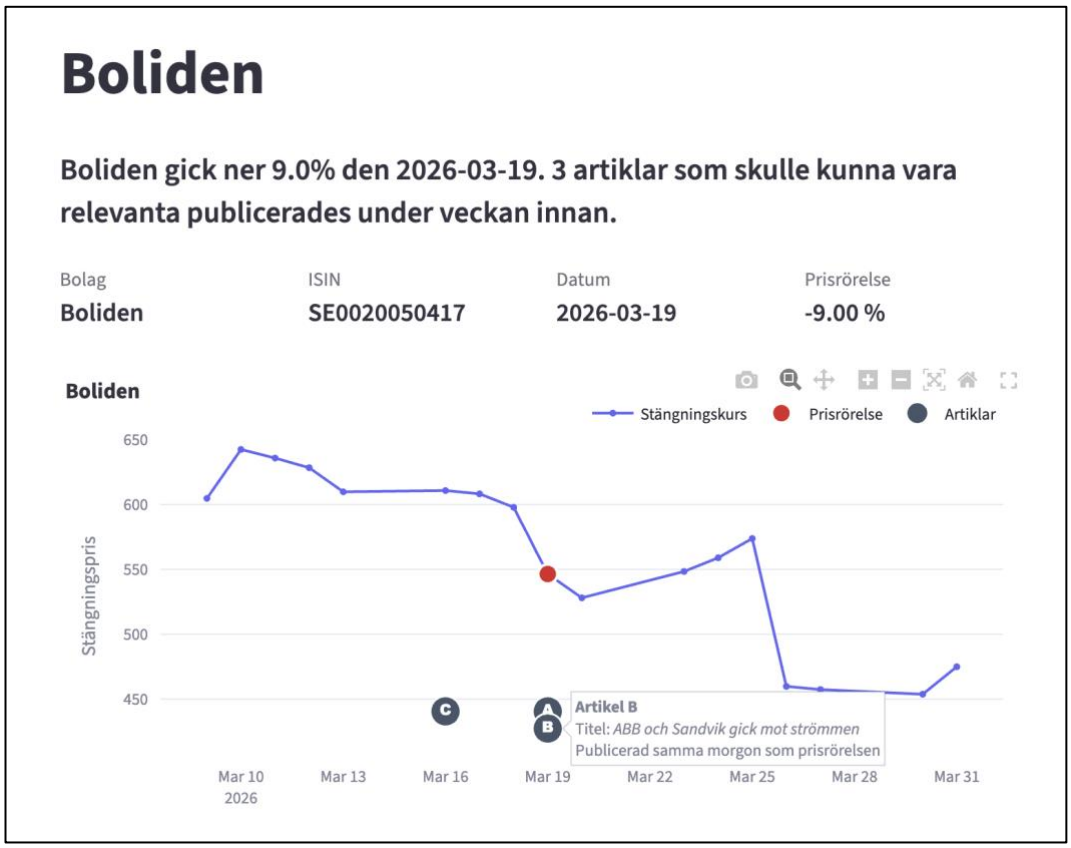

Figure 6. Example of the anomaly context shown to experts in the evaluation form, providing the company, date, and price movement for one of the 10 evaluated anomalies.

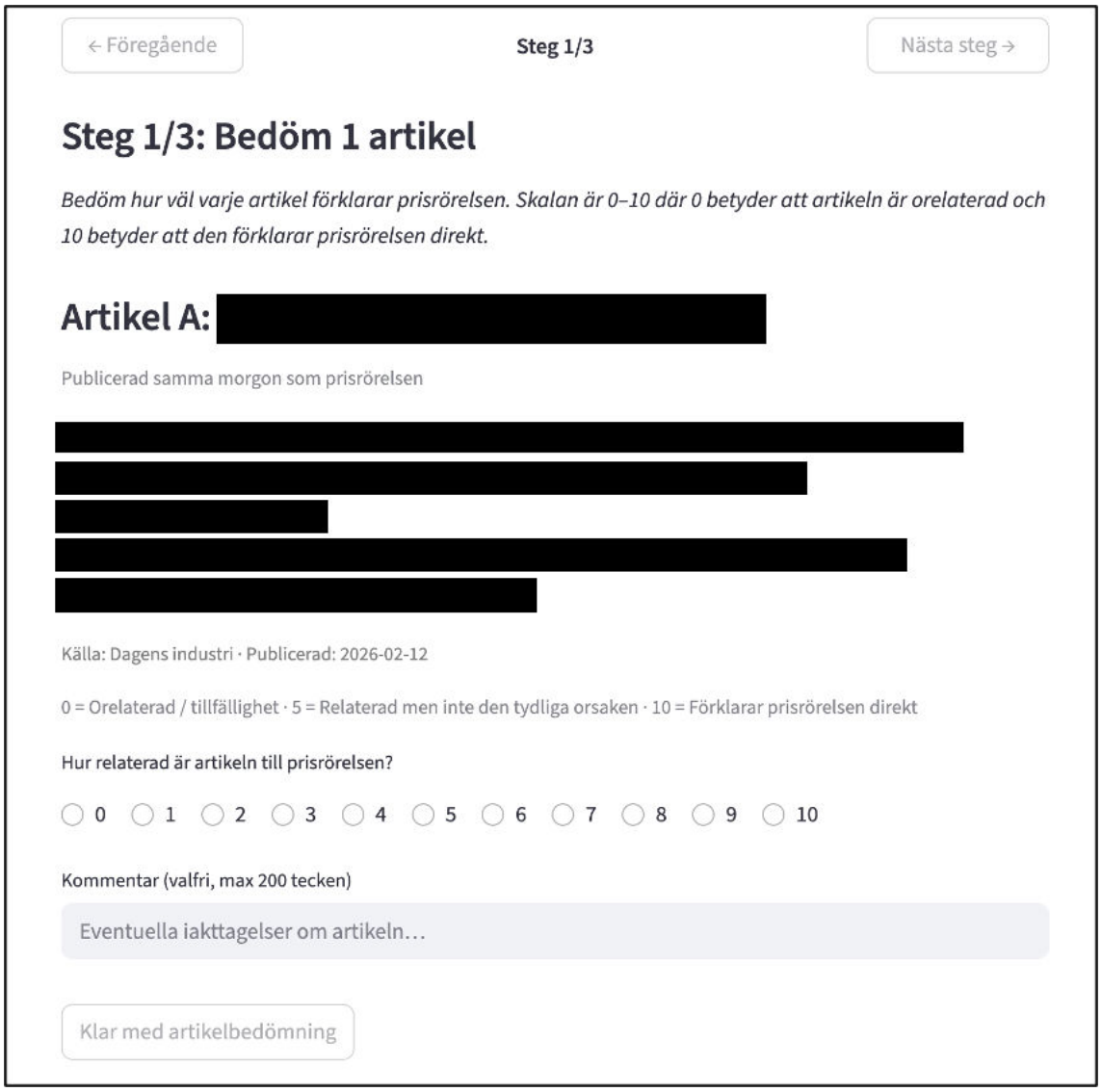

Figure 7. Example item from the expert evaluation form, in which a financial professional rates the relevance of a news article to an identified stock price movement. The article text shown to the financial experts during evaluation has been hidden due to licensing restrictions.

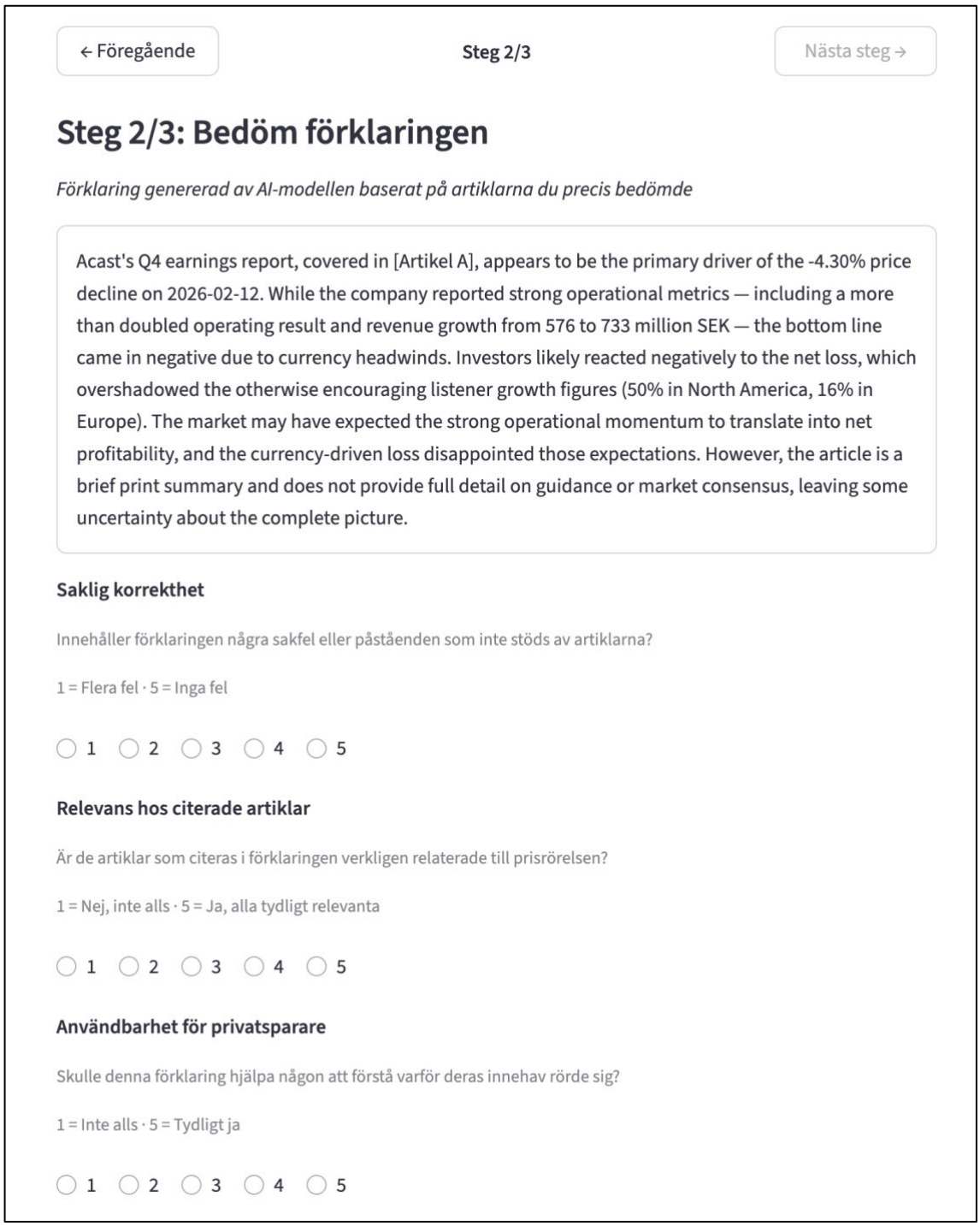

Figure 8. Example item from the expert evaluation form, in which a financial professional rates the AI-generated explanation on a 1-5 scale in 3 dimensions.

Figure 9. The optional free-text comment step of the expert evaluation form, in which a financial professional could note anything missing or incorrect in the AI-generated explanation.