

Article

LLM-Based Merger Withdrawal Prediction and Model–Market Disagreement Around Climate Risk

Muntasir Shohrab , Zhibo Ye, Yanguang Liu * and Dantong Yu

Martin Tuchman School of Management, New Jersey Institute of Technology, Newark, NJ 07102, USA

* Correspondence: yl224@njit.edu

Abstract

Predicting merger withdrawals is challenging because failed deals can impose large costs on firms and investors, while withdrawals are rare and difficult to identify using structured deal variables alone. We propose MergerLLM, a large-language-model-based framework that predicts merger completion using serialized deal characteristics and text-based firm information. Using a sample of U.S. mergers and acquisitions (M&A) deals, we compare MergerLLM with machine-learning and deep-learning benchmarks. MergerLLM improves detection of withdrawn deals, especially in recall and F1-score, while maintaining competitive ranking performance. Its predicted merger-success probabilities also contain economic information in post-announcement return tests. We then construct HAIDiff, a proxy for model–market disagreement, defined as the difference between MergerLLM’s predicted merger-success probability and a transformed announcement-return-based market signal. Climate exposure is positively associated with this disagreement measure, and the result is robust to an alternative measure based on target cumulative abnormal returns. The evidence suggests that climate-exposed deals are settings in which model-implied completion probabilities and market-based signals diverge more strongly.

Keywords: artificial intelligence; machine learning; fintech; merger and acquisition; model–market disagreement; climate risk

JEL Classification: G34; G14; C45; Q54

1. Introduction

Mergers and acquisitions (M&As) are central corporate decisions that can reshape firm boundaries, market structure, and shareholder value (Andrade et al., 2001; Cookson et al., 2026; Renneboog & Vansteenkiste, 2019). Although announced mergers are often motivated by expected synergies, diversification, market expansion, or strategic repositioning, a nontrivial fraction of announced deals are not completed (Becher et al., 2015). Deal withdrawals create important financial risks for acquirers, targets, investors, and other market participants. They may involve termination fees, legal and advisory costs, reputational damage, financing disruptions, and sharp stock price reactions (Harford, 2005; Officer, 2003). Predicting whether an announced deal will be completed is, therefore, an economically important problem.

Merger completion risk is difficult to predict because it depends on several types of information at the same time. Some risks are reflected in structured deal terms, such as payment method, percentage of shares sought, target public status, hostile bids, competing bidders, and challenged deals. Other risks are embedded in textual information, such as the business descriptions of the acquirer and target, the strategic relation between the two firms,



Academic Editor: Thanasis Stengos

Received: 21 April 2026

Revised: 13 June 2026

Accepted: 13 June 2026

Published: 1 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and the economic meaning of the transaction. In addition, merger outcomes are affected by broader regulatory, financing, and macroeconomic conditions. These features make merger-withdrawal prediction a natural setting for methods that can combine structured transaction variables with text-based firm information.

Climate exposure represents a complex, long-horizon source of uncertainty that may shape how merger-related information is assessed. Climate-related exposure can enter deal assessment through several channels. First, it may increase regulatory and policy-review uncertainty when a transaction involves carbon-intensive firms, environmentally sensitive assets, or operations that are likely to be affected by climate regulation (Krueger et al., 2020; Lu, 2022; Seltzer et al., 2022). Second, it may increase valuation uncertainty because expected synergies, stakeholder value, and long-run cash flows are harder to assess when firms face material environmental exposure and when ESG assessments differ across information providers (Berg et al., 2022; Deng et al., 2013; Gomes, 2019). Third, it may be related to financing conditions if lenders and investors price climate-related risk into loan spreads, bond yields, or the cost of capital (Chava, 2014; Delis et al., 2024; Javadi & Masum, 2021). Fourth, firms exposed to the green transition may face stranded assets, abatement costs, carbon-pricing risk, or changes in product demand (Bolton & Kacperczyk, 2021, 2023; Ilhan et al., 2021). Fifth, physical climate risk may affect insurance costs, asset insurability, or the collateral value of assets exposed to extreme weather and natural-disaster risk (Garbarino & Guin, 2021; Ge & Lam, 2022; Kousky, 2019). Finally, climate-related news, investor attention, and market learning may affect market reactions to announced deals, especially when investors have difficulty pricing long-horizon climate exposure (Choi et al., 2020; Engle et al., 2020; Hong et al., 2019). These channels suggest that climate exposure may affect the information environment in which merger-completion risk is assessed. We, therefore, use climate exposure as a setting for studying when model-implied completion probabilities and announcement-return-based market signals diverge.

In this paper, we propose MergerLLM, a large-language-model-based framework for predicting merger withdrawals. The method converts each tabular deal observation into a natural-language representation by serializing feature names and feature values into text. This representation allows the model to process numerical variables, categorical variables, and text fields within a common language-based input. We then fine-tune the model for the binary classification task of predicting whether an announced merger will be completed or withdrawn. This design is well suited to our empirical setting for two reasons. First, merger withdrawals are rare, which makes the classification problem highly imbalanced. Second, important information appears both in structured deal characteristics and in textual descriptions that may lose context when converted into simple engineered features.

We evaluate MergerLLM on a large sample of U.S. M&A deals from SDC Platinum. The final prediction sample contains 55,363 announced deals, including 52,252 completed deals and 3111 withdrawn deals. We compare MergerLLM with several machine-learning and deep-learning baselines, including gradient-boosted tree models, random forests, neural-network models, and tabular deep-learning models. The results show that MergerLLM improves minority-class detection relative to strong baselines. In the test sample, MergerLLM achieves an accuracy of 0.955, a precision of 0.606, a recall of 0.440, an F1-score of 0.510, an ROC AUC of 0.905, and a PR AUC of 0.465. Withdrawn deals remain difficult to classify because they are rare and economically heterogeneous. In this setting, false negatives are especially important because a withdrawn deal that is classified as likely to complete may expose investors to substantial downside risk. The main result is that MergerLLM provides a more balanced prediction of rare withdrawal events, especially in recall and F1-score, while maintaining competitive ranking performance.

We also examine the economic information content of MergerLLM's predicted completion probabilities. Deals are sorted into quintiles based on predicted merger-success probabilities. Acquirers in the highest predicted-success quintile earn higher post-announcement returns than acquirers in the lowest predicted-success quintile. A long-short strategy that buys acquirers in the highest quintile and shorts acquirers in the lowest quintile generates a positive three-month return spread. We interpret this evidence as showing that MergerLLM's predictions contain economically meaningful information about merger outcomes. Because actual trading performance depends on transaction costs, short-selling constraints, liquidity, bid-ask spreads, deal duration, delisting, and other market frictions, we interpret the portfolio results as evidence on the economic information content of the predictions instead of as a complete trading implementation.

After establishing the predictive content of MergerLLM, we use the model to study disagreement between model-implied merger-success probabilities and market-based assessments around merger announcements. This comparison is motivated by prediction policy problems, where algorithmic predictions may help decision makers when the outcome is uncertain and the relevant information is complex (Kleinberg et al., 2015). The M&A setting is useful for this comparison because announcement returns aggregate market participants' assessments of the transaction, but they also reflect expected synergies, financing conditions, risk premia, information asymmetry, arbitrage constraints, and short-term trading pressure. We, therefore, use announcement returns as a market-based signal of deal assessment rather than as a clean measure of investor beliefs about merger completion. Comparing this signal with MergerLLM's predicted completion probabilities allows us to study model-market disagreement around merger announcements.

To implement this comparison, we construct HAIDiff as a proxy for model-market disagreement. Cumulative abnormal returns around merger announcements are often used as a proxy for market reactions to deal announcements (Dasgupta et al., 2024; Luo, 2005; Tunyi, 2021). However, announcement returns are not clean measures of perceived deal-completion probability. They also reflect expected post-merger performance, overpayment concerns, financing terms, governance issues, and risk premia. To reduce this concern, we residualize acquirer announcement returns with respect to post-merger operating performance and then scale the residual-based measure between zero and one. HAIDiff is defined as the difference between MergerLLM's predicted probability of merger success and this scaled market-based signal. This measure captures the distance between a model-implied completion probability and a transformed announcement-return-based signal.

We then examine whether climate exposure is associated with larger model-market disagreement. Climate exposure is useful for this analysis because it captures a source of uncertainty that is multi-dimensional, long-horizon, and difficult to summarize using a small set of structured variables. Following Sautner et al. (2023), we use firm-level climate change exposure to capture attention to climate-related opportunities, physical risks, and regulatory shocks. We find that climate exposure is positively associated with HAIDiff. This result suggests that climate-exposed deals are settings in which model-implied completion probabilities and announcement-return-based market signals differ more strongly.

To provide additional evidence, we examine model-market disagreement around Hurricane Ian, a period of heightened attention to physical climate risk. We classify treatment firms based on pre-Ian firm-level climate exposure and compare changes in HAIDiff before and after the event. The results show larger post-Ian increases in HAIDiff for firms with higher pre-existing climate exposure. Event-study, balance, matching, and inverse-probability-weighted tests provide additional evidence on this pattern.

The paper offers three connected contributions. The primary contribution is to develop and evaluate MergerLLM as an LLM-based framework for merger-withdrawal prediction.

This contribution addresses a difficult rare-event prediction problem in corporate finance by combining structured deal variables with text-based firm information. The second contribution is to use the model's predicted completion probabilities to study model–market disagreement around merger announcements. This comparison relates model-implied probabilities to a transformed announcement–return-based signal that reflects market assessments of the deal as well as other economic forces. The third contribution is to examine climate exposure as a setting in which model-implied completion probabilities and market-based signals may diverge. This analysis links the prediction exercise to climate finance by showing that climate-exposed deals are associated with larger model–market disagreement around merger announcements.

Together, these contributions connect the prediction exercise to broader economic questions in M&A and climate finance. The prediction model provides the foundation for estimating merger-completion risk. The model–market comparison examines the economic content of these predictions. The climate-risk analysis then studies a setting in which complex long-horizon uncertainty may make model–market disagreement more pronounced. This structure connects the paper to the M&A prediction literature, the machine-learning-in-finance literature, and the climate finance literature.

2. Literature

Relation to the Literature

We organize the literature around the paper's central problem: predicting merger withdrawals when relevant information is dispersed across structured deal variables, textual firm information, market reactions, and broader economic risks. Existing merger-outcome models identify important deal- and firm-level predictors, but merger withdrawals remain difficult to predict because they are rare and because relevant information is often embedded in text fields and broader economic conditions. This setting connects three areas of research. The first studies merger outcomes and deal-completion risk. The second develops machine-learning methods for financial prediction problems involving structured and textual data. The third examines climate-related financial risks, which provide a setting in which long-horizon and multidimensional uncertainty may affect merger-completion assessments.

Prior M&A research provides the starting point by identifying deal- and firm-level determinants of merger completion. M&A transactions are major corporate events that can affect firm value, market structure, and strategic positioning (Andrade et al., 2001; Cookson et al., 2026; Renneboog & Vansteenkiste, 2019). Prior research shows that merger completion depends on deal characteristics, firm characteristics, regulatory conditions, and market reactions. Earlier studies often use financial and managerial variables to explain merger outcomes, including market-to-book ratios, cash flow, managerial inefficiency, industry conditions, and other observable firm characteristics (Pasiouras et al., 2008). Other work studies related corporate failure and bankruptcy prediction problems (Cho et al., 2010; Olson et al., 2012). More recent studies apply data-mining and machine-learning methods to M&A settings, including merger target prediction (Xiang et al., 2012). These studies show that prediction methods can be useful in corporate finance, but merger withdrawal remains a difficult rare-event classification problem because withdrawn deals account for only a small share of announced transactions.

A central challenge in merger-withdrawal prediction is that economically relevant information is not limited to numerical variables. Deal terms such as payment method, target public status, challenged-deal indicators, competing bidders, hostility, and percentage of shares sought are important structured predictors. However, important information may also appear in text fields, including acquirer and target business descriptions, consideration descriptions, and other deal narratives. These text fields may contain information about

business relatedness, regulatory exposure, industry similarity, technological fit, and long-horizon risks. If these variables are reduced to simple engineered features, some economic context may be lost. Therefore, the prediction problem requires a method that can preserve both the structure of the tabular data and the semantic meaning of text-based variables.

Machine-learning research in finance provides the methodological benchmark for this prediction problem. Machine-learning methods have been applied to stock return prediction, risk management, consumer credit risk, textual analysis, and other financial prediction tasks (Butaru et al., 2016; Gu et al., 2020; Khandani et al., 2010; Netzer et al., 2019). In tabular prediction tasks, tree-based models such as XGBoost, LightGBM, and CatBoost are often strong benchmarks because they handle nonlinearities, interactions, missing values, and heterogeneous variable types well. Accordingly, these models provide demanding benchmarks for evaluating MergerLLM. At the same time, standard tree-based models do not directly process natural-language inputs. Text variables typically need to be converted into bag-of-words, TF-IDF, embeddings, or other engineered representations before being used by these models. This preprocessing can be effective, but it may separate text from the economic meaning of feature names and from the structured variables in the original deal record.

Large language models provide a different way to represent this prediction problem. Recent work in financial language models, including BloombergGPT and FinGPT, shows that language models can be adapted to financial text and financial tasks (Wu et al., 2023). Related work on tabular LLMs shows that tabular rows can be serialized into natural-language strings and then processed by language models (Hegselmann et al., 2023). This design is especially relevant for our setting because each merger record contains both structured variables and text fields. By serializing feature names and feature values into natural language, MergerLLM preserves the semantic meaning of variables such as “payment method,” “challenged deal,” and “target business description.” It also allows the model to process categorical, numerical, and textual information within one input representation. In this sense, the model is designed for the specific structure of merger data: it combines deal terms, firm descriptions, and task-specific prompts in a common language-based format.

Our work differs from existing tabular LLM applications in two ways. First, we apply this approach to a high-stakes financial rare-event prediction problem, where the minority class is economically important. Merger withdrawals are infrequent, but they can have significant financial consequences for investors and firms. Second, we examine whether the model’s predicted completion probabilities contain economic information beyond standard classification metrics. This connects the prediction exercise to merger arbitrage returns and to market-based assessments of deal completion. Thus, the paper is not only a technical application of LLMs to tabular data. It also studies whether language-model predictions are useful in an economically meaningful corporate-finance setting.

A related literature on prediction policy and market-based assessments helps motivate our model–market comparison. A growing literature examines prediction policy problems, where machine-learning methods can help improve decisions in settings such as hiring, bail, and financial decision-making (Cao et al., 2024; Chalfin et al., 2016; Kleinberg et al., 2015). The M&A setting provides a useful environment for this question because stock price reactions around merger announcements aggregate market participants’ assessments of deal value and completion risk. Because market reactions reflect multiple economic forces, they provide an informative but noisy signal of deal assessment. Announcement returns may reflect expected synergies, financing terms, risk premia, liquidity, arbitrage constraints, information asymmetry, and short-term trading pressure. Prior M&A research uses announcement returns to study investor reactions and perceived deal quality (Dasgupta et al., 2024; Luo, 2005; Tunyi, 2021), while prediction-market research emphasizes that markets can aggregate dispersed infor-

mation when participants trade on uncertain outcomes (Arrow, 1998; Fama, 1970; Wolfers & Zitzewitz, 2004). Building on this idea, we compare MergerLLM's predicted completion probabilities with market-based signals to study model–market disagreement around merger announcements.

Finally, this paper relates to the climate finance literature. Climate risk is increasingly recognized as a financially relevant source of uncertainty (Addoum et al., 2020; Bernstein et al., 2019). Firm-level climate exposure can reflect attention to climate-related opportunities, physical risks, and regulatory shocks (Sautner et al., 2023). These risks may affect corporate transactions through several channels, including regulatory approval uncertainty, post-merger valuation uncertainty, financing costs, transition costs, insurance costs, and investor attention. Prior work shows that institutional investors and financial markets increasingly consider climate risk (Krueger et al., 2020; Seltzer et al., 2022), that climate-related information can be difficult to measure because ratings and assessments may disagree (Berg et al., 2022), and that climate risk can affect financing conditions and the cost of capital (Chava, 2014; Delis et al., 2024; Javadi & Masum, 2021). Other studies show that carbon-transition risk, physical risk, and investor attention to climate news can be priced in financial markets (Bolton & Kacperczyk, 2021, 2023; Choi et al., 2020; Engle et al., 2020; Hong et al., 2019; Ilhan et al., 2021).

This climate literature is important for our empirical design because it motivates climate exposure as a setting in which merger-related information may be assessed differently by models and market participants. Climate exposure can be related to expected cash flows, financing conditions, regulatory scrutiny, and investor attention, all of which may affect the information environment around announced deals. Because climate risk is long-horizon, multidimensional, and difficult to summarize with a small number of variables, it is a natural setting in which model-implied probabilities and market-based signals may diverge. Hurricane Ian provides a period of heightened physical climate-risk salience. We, therefore, use it to examine whether model–market disagreement changes more for firms with higher pre-existing climate exposure. This setting connects the climate-risk analysis to the broader research question: when merger-related uncertainty is difficult to summarize with standard variables, language-model predictions may differ from market-based signals.

These streams of the literature inspire the paper's empirical design. Because merger withdrawal is economically important but difficult to predict, we develop MergerLLM to combine structured deal variables with text fields while preserving the economic meaning of feature names. We then use climate exposure as a setting for studying differences between model-implied completion probabilities and announcement-return-based market signals.

3. Materials and Methods

3.1. Mergers and Acquisitions Sample

We compile a dataset of mergers and acquisitions (M&A) transactions from the SDC Platinum database, which covers deals announced between 1965 and 2023. The dataset contains key transaction information, including the announcement date, deal value, percentage of shares sought, and deal outcome.

We construct the prediction sample at the deal-announcement level. The unit of observation is a unique SDC transaction identifier. We begin with announced M&A transactions from SDC Platinum and impose several filters. First, we require a nonmissing acquirer cumulative abnormal return around the announcement date. This restriction requires a successful match between the SDC acquirer and CRSP return data and, therefore, limits the final sample to transactions with publicly traded acquirers that have sufficient return data around the announcement.

For predictor variables used in the withdrawal classifier, we retain observations with missing values when the variable is otherwise available in SDC, because missingness may itself reflect the information environment of the transaction. Tree-based baseline models handle missing values internally, and MergerLLM serializes missing entries using a common missing-value token. We do not use outcome-realized variables or post-merger operating variables as inputs to the withdrawal classifier.

In addition, we restrict the sample to transactions involving U.S. acquirers. We also retain only transactions with a final SDC status that can be mapped clearly into completed or withdrawn outcomes. Specifically, transactions with status code C are classified as completed, while transactions with status codes W, SW, and IW are classified as withdrawn. In the SDC status definitions, C denotes completed transactions, W denotes withdrawn transactions, and SW denotes seeking-buyer withdrawn transactions. The code IW appears in the raw SDC data but is not separately listed in the manual available to us. We group these observations with withdrawn transactions because they are not completed transactions in the SDC extract.

We remove duplicate transaction records using the unique SDC deal identifier. Multiple-bid transactions are not removed from the sample because competing bids are economically meaningful for deal-completion risk. Instead, we retain these observations and include the SDC multiple-bidder indicator, *Multibidder*, as one of the input variables. Similarly, if a withdrawn transaction is later reannounced as a separate SDC transaction with a new deal identifier and announcement date, we treat the reannounced deal as a separate transaction attempt. This coding reflects the prediction problem at the announcement date because the first transaction attempt is correctly labeled as withdrawn if it fails to complete, while a later reannouncement represents a new deal attempt with its own information set and outcome.

After applying these filters, the final prediction sample contains 55,363 M&A deals involving U.S. acquirers, including 52,252 completed deals and 3111 withdrawn deals. In line with prior studies (Bates & Lemmon, 2003; Betton et al., 2014; Boone & Mulherin, 2007; Cookson et al., 2026; Luo, 2005), we augment the SDC data with additional deal characteristics known to be related to merger outcomes, especially withdrawals. These characteristics include the method of payment, the involvement of a white knight, the presence of anti-takeover provisions, and rumors before the deal announcement. These variables are used as input variables in the prediction analysis and as controls in the model–market disagreement tests. Detailed definitions of the deal characteristics from SDC Platinum are provided in Appendix A.

The prediction task is defined at the merger announcement date. We, therefore, use deal characteristics that are recorded in SDC Platinum and that describe the announced transaction, including payment method, percentage of shares sought, target public status, white knight, hostile deal, rumor, challenged deal, and acquirer and target business descriptions. We exclude variables that are realized only after the deal outcome, such as final completion-related information beyond the dependent variable, and post-merger operating performance, from the withdrawal classifier. At the same time, some SDC fields may be revised as additional information becomes available during the deal process. This timing issue is most relevant for variables that can evolve after announcement, such as challenged-deal status or revised transaction terms. We, therefore, interpret the prediction results as based on the transaction information recorded in SDC for the announced deal while taking care not to use ex post return or operating-performance variables as classifier inputs.

We calculate cumulative abnormal returns (CARs) for acquiring firms over the three-day event window around the M&A announcement, covering trading days $[-1, 1]$. CARs are estimated using the Fama–French three-factor model and stock return data from CRSP.

The parameters of the Fama–French three-factor model are estimated over trading days $[-212, -10]$ relative to the announcement date.

3.2. Macroeconomic Factors

Following [Chen et al. \(2024\)](#), we incorporate nine macroeconomic variables to capture different dimensions of macroeconomic risk: (1) CON: the monthly growth in real personal consumption expenditures on nondurable goods and services per capita; (2) TFP: the quarterly percentage change in total factor productivity; (3) IPG: the monthly growth rate of industrial production; (4) TERM: the yield spread between 20-year and 1-year Treasury bonds; (5) DEF: the default premium, calculated as the monthly change in the yield spread between BAA-rated and AAA-rated corporate bonds; (6) DEI: the monthly change in expected inflation; (7) SENT: Baker and Wurgler (BW) sentiment measure on various market anomalies [Baker and Wurgler \(2006\)](#); (8) MKT: the value-weighted excess returns of the stock market; and (9) LAB: the logarithmic growth rate of nominal labor income per capita. Additionally, we include firm-level climate change exposure (CC Expo) based on [Sautner et al. \(2023\)](#), which reflects the attention given by earnings call participants to firms' climate change risks and opportunities. The CC Expo measure accounts for exposures to opportunities, physical risks, and regulatory shocks associated with climate change.

3.3. Variable Construction and Summary Statistics

Figure 1 illustrates the trends in completed and withdrawn merger cases over time. The figure shows substantial time-series variation in merger withdrawal rates. Some periods with higher withdrawal rates coincide with changes in regulatory scrutiny, including periods of stronger antitrust enforcement, as well as macroeconomic conditions and broader market uncertainty. The withdrawal rate also increased in 2022, a period that includes Hurricane Ian and heightened attention to physical climate risk. We treat this pattern as descriptive. The formal Hurricane Ian analysis below examines whether model–market disagreement changes more for firms with higher pre-existing climate exposure after this climate-risk-salience event.

Table 1 presents annual merger withdrawal statistics, including the total number of completed and withdrawn cases, as well as the yearly percentage of withdrawals. Table 2 shows that the final sample consists of 55,363 deals, of which 52,252 were completed and 3111 were withdrawn. This corresponds to a deal withdrawal rate of approximately 5.62%, consistent with the prior literature (e.g., [Luo, 2005](#)). Most M&A transactions in the sample are full takeovers, with a median percentage of shares sought at 100% (mean of 83.67% for completed deals and 83.43% for withdrawn deals). The differences between completed and withdrawn samples, as shown in Table 2, align with findings in prior research on M&A completion. For instance, acquirer CARs are higher for completed deals compared to withdrawn deals, consistent with findings by ([Ang et al., 2021](#); [Liu & McConnell, 2013](#); [Luo, 2005](#)).

Table 1. Merger withdrawal statistics by year. This table presents the withdrawal statistics by year, showing the total number of completed and withdrawn cases, along with the percentage of withdrawals each year.

Year	Completed	Withdrawn	Percentage Withdrawn (%)
1978	13	3	18.75
1979	12	3	20.00
1980	43	7	14.00
1981	445	58	11.53
1982	690	91	11.65

Table 1. *Cont.*

Year	Completed	Withdrawn	Percentage Withdrawn (%)
1983	1080	98	8.32
1984	1432	144	9.14
1985	630	87	12.13
1986	964	129	11.80
1987	986	131	11.73
1988	1024	166	13.95
1989	1511	221	12.76
1990	1644	200	10.85
1991	1617	178	9.92
1992	1945	181	8.51
1993	2410	211	8.05
1994	3019	242	7.42
1995	3327	197	5.59
1996	4201	234	5.28
1997	5121	230	4.30
1998	5716	227	3.82
1999	4742	205	4.14
2000	4219	310	6.84
2001	2858	164	5.43
2002	2419	109	4.31
2003	2366	75	3.07
2004	2611	81	3.01
2005	2894	63	2.13
2006	2940	85	2.81
2007	2925	89	2.95
2008	2389	107	4.29
2009	1686	72	4.10
2010	2116	83	3.77
2011	2254	89	3.80
2012	2275	86	3.64
2013	2094	74	3.41
2014	2464	128	4.94
2015	2353	114	4.62
2016	2015	75	3.59
2017	2081	76	3.52
2018	2039	87	4.09
2019	1693	87	4.89
2020	1556	81	4.95
2021	2286	137	5.65
2022	1842	140	7.06
2023	1322	50	3.64

Table 2. Summary statistics. This table presents the summary statistics of key variables. Detailed variable definitions are in Table A1.

	Completed (N = 52,252)		Withdrawn (N = 3111)		Diff.
	Mean	SD	Mean	SD	
CAR Acq [−1, 1]	0.0104	0.0870	0.0098	0.1222	0.0006
log(deal value)	3.6588	2.0194	4.5826	2.2084	−0.9238
white knight	0.0010	0.0312	0.0055	0.0737	−0.0045
Pct. Shares Sought	83.6743	32.4675	83.4281	32.7189	0.2462
Rumored Deal (0/1)	0.0203	0.1411	0.0360	0.1863	−0.0157
Competing Bidder (0/1)	0.0071	0.0839	0.1350	0.3418	−0.1279
Hostile Deal (0/1)	0.0025	0.0500	0.0759	0.2648	−0.0734
Target Public (0/1)	0.2570	0.4370	0.6050	0.4889	−0.3480
Acquirer Market Value (billion)	10.7049	31.4263	9.0012	28.2181	1.7037

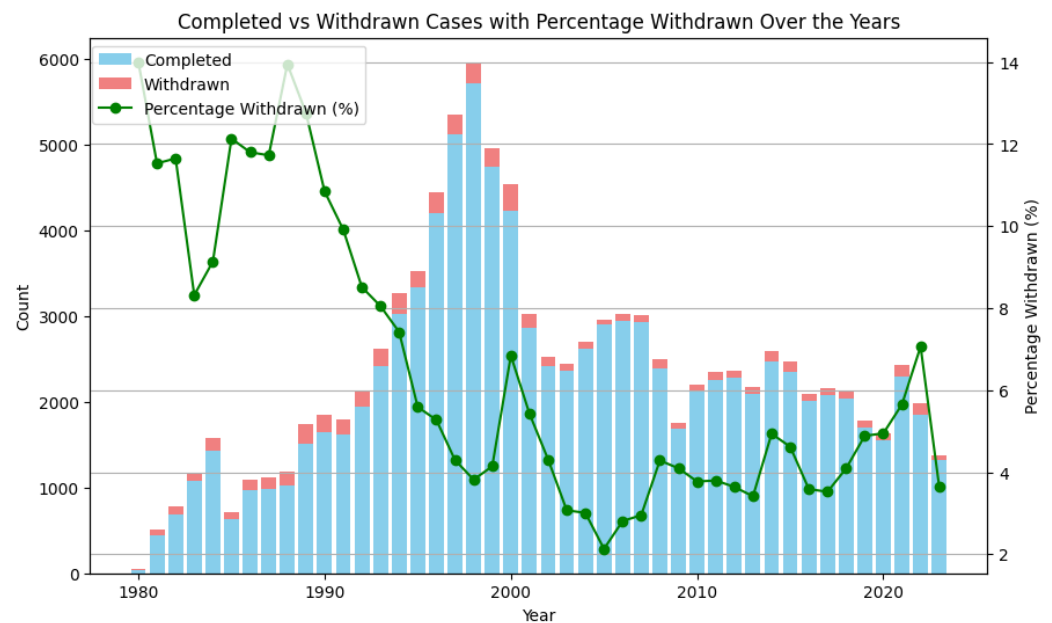


Figure 1. Annual trends in completed and withdrawn mergers. This figure presents the completed vs. withdrawn cases. The x-axis contains the year of the merger. The left side of the y-axis shows the total number of mergers, and the right side shows the percentage of withdrawals. The height of the bar plot represents the total number of merger deals in the year. The blue part shows the completed mergers, and the red part shows the withdrawn mergers. The line plot represents the percentage of mergers that were withdrawn in that year.

3.4. Information and Variables as Inputs to LLM

Figure 2 illustrates the overall architecture of our approach. First, we serialize feature names and values into a natural language string. This serialized string is combined with a task-specific prompt. The LLM is fine-tuned using labeled examples and then employed to generate predictions for unlabeled examples. The LLM output is replaced with a Feedforward Neural Network (FFN) to predict the label of each row, and cross-entropy loss is used to back-propagate gradients and fine-tune the LLM weights. For fine-tuning, we apply LoRA adaptation that updates only a subset of the LLM’s original weights.

To apply LLMs to tabular data, it must first be transformed into a natural text representation (Dinh et al., 2022). Various serialization techniques can be employed to convert a table row into text, such as text templates, list templates, table-to-text, Text-T0, and GPT-3. Among these techniques, we adopt the text template due to its superior effectiveness (Hegselmann et al., 2023). This serialized natural text is then input into the LLMs to fine-tune them for specific tasks.

This step is essential because LLMs are pre-trained on extensive text corpora, with textual knowledge embedded within the model. Our goal is to leverage the existing knowledge base of LLMs. The mergers and acquisitions dataset is initially stored in a tabular format, which is unsuitable for LLM processing. We convert the tabular data into a simple text format, allowing LLMs to process it directly and infer the probability of success.

Given a tabular dataset with n rows and d columns, we represent it as $D = \{(x_i, y_i)\}_{i=1}^n$, where each x_i is a d -dimensional feature vector. We formulate this prediction as a classification problem, where $y_i \in C$, with C representing a set of classes. The column (attribute) names are denoted as $F = \{f_1, \dots, f_d\}$. Each f_i is a natural-language string, such as “percentage of shares sought,” “target public status,” or “method of payment.” We define a function, $serialize(F, x)$, that takes as input the column names F and feature values x for a row and outputs a textual representation of the data. We also use a prompt to describe the classification problem. The prompt that we use is “will the M&A deal be completed

or withdrawn? Answer:". We merge both the serialized text and the task-specific prompt $((\text{serialize}(F, x), p))$ and feed the combined prompt to the selected LLM.

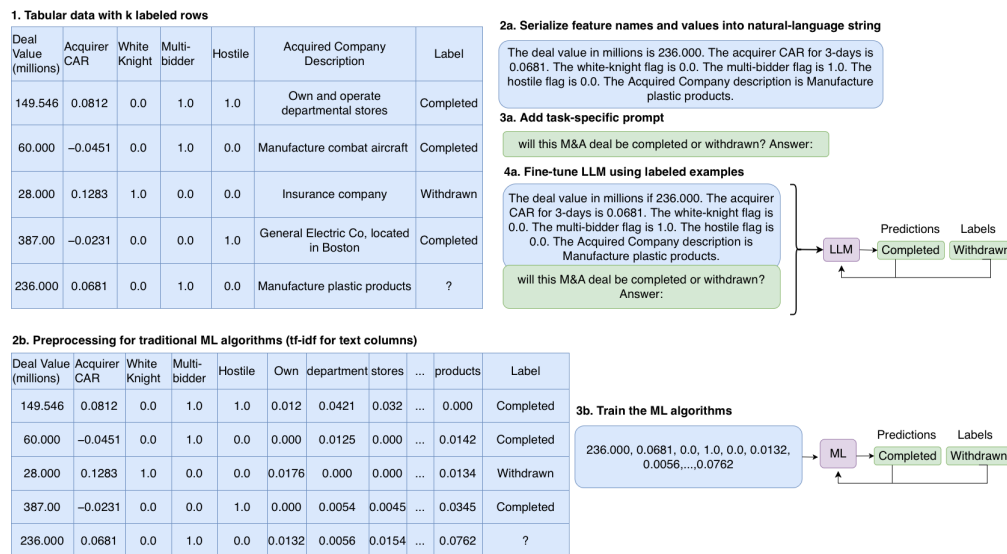


Figure 2. MergerLLM model overview. This figure summarizes the MergerLLM framework. Panel (2a) shows the serialization process for historical M&A data. Each row is converted into a text representation and combined with the task-specific prompt, “Will this M&A deal be completed or withdrawn? Answer:". The serialized data are then used to fine-tune the LLM with the Hugging Face Trainer. Panel (2b) shows the baseline pipeline, where text columns are represented using TF-IDF before training the machine-learning benchmark models.

3.5. Large Language Model Fine-Tuning

Our framework is compatible with a wide range of LLMs designed to generate responses from natural language inputs. In this study, we employ autoregressive LLMs with billions of parameters (ranging from 2 to 70 billion) trained on extensive text corpora. These models are pre-trained to predict subsequent words based on the input text. Let LLM denote a large language model with vocabulary V . The output of LLM includes a classification head aligned with the number of labels. For instance, $LLM((\text{serialize}(F, x), p)) \in V^*$ represents the prompted output of the LLM, which is mapped to a valid class C . For fine-tuning, the LLM is updated using training examples $(\text{serialize}(F, x), p)|(x, y) \in D_k$, where D_k represents the domain of all labeled examples. The LLM generates text in the vocabulary space V^* , which must be assigned to a valid binary class: $C = \text{Withdrawn}, \text{Completed}$. Several approaches to this problem have been proposed in the literature. For instance, a verbalizer defines a mapping between LLM output tokens and the discrete labels within the label space (Schick & Schütze, 2021). Verbalizers can be manually defined or automatically learned (Cui et al., 2022). However, verbalizers are more appropriate for few-shot learning (Nguyen et al., 2024) where the model parameters are not updated. In our framework, we use a classification head on top of the last-hidden layer of the LLM.

One of the most important components of our framework is the LLM. Our framework is agnostic to the LLM implementation and any particular fine-tuning method applied on the LLM. In our case, we employ Llama3, a decoder transformer that masks the future word tokens in relation to the current step and predicts the target tokens w_i using the preceding tokens $w_{<i}$ in a sequence. The training objective of the LLM is to maximize the likelihood:

$$\mathcal{L}_{LM}(w) = \sum_{i=1}^n \log P(w_i | w_{<i}). \quad (1)$$

On the base model of the LLM, we use a classification head. The model produces logits, z , for the possible classes, C . The classification head maps these logits to the discrete label space of “completed” and “withdrawn”. We optimize the model by minimizing the cross-entropy loss for our domain-specific task:

$$\mathcal{L} = - \sum_{i=1}^n \sum_{c \in C} \mathbf{1}[y_i = c] \log p(c | x_i), \quad (2)$$

where $p(c | x_i)$ is the predicted probability for class c , and $\mathbf{1}[y_i = c]$ is an indicator function for the true label. This loss function penalizes incorrect predictions and helps fine-tune the LLM weights by back-propagating the gradient, improving alignment with the classification task.

Fine-tuning all the weights of an LLM is challenging due to the need to load all weights into GPU memory. The size of LLMs can be substantial, depending on the number of weights. To fine-tune our LLM efficiently, we use Low-Rank Adaptation (LoRA). LoRA freezes all pre-trained model weights and adds trainable low-rank decomposition matrices into different layers of the Transformer architecture (Hu et al., 2021). This approach significantly reduces the number of trainable parameters. As a result, the model can be trained with lower GPU memory usage, which is important because GPU memory is a major constraint in LLM fine-tuning. Specifically, for a weight matrix, $W_0 \in \mathbb{R}^{d \times k}$, in the model, LoRA constrains updates by representing it as $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with $r \ll \min(d, k)$. During training, W_0 remains frozen, while only A and B are updated. This method simulates bypass and residual learning, enabling the model to avoid catastrophic forgetting and adapt effectively to specific tasks without requiring full parameter fine-tuning.

To implement LoRA, we follow these steps:

1. Initialize the low-rank matrices A and B . A is initialized randomly, and B is initialized to zero.
2. Train only the low-rank matrices A and B while keeping the pre-trained weights W_0 fixed.
3. Combine the updated low-rank matrices with the original weight matrix W_0 during inference to obtain the final adapted weights $W = W_0 + BA$.

In our model implementation, we replace the LLM’s output layer with a classification head to produce logits for each label associated with a data entry, instead of simply outputting the next word (“Completed” and “Withdrawn”). The replacement transforms the LLM task for next word prediction into a standard classification problem, which optimizes the fine-tuning process and improves model accuracy for the classification task. The classification head parameters are also updated during this training process.

3.6. Implementation Details and Reproducibility

In this subsection, we describe the implementation choices behind MergerLLM in sufficient detail that an independent researcher can reproduce our results. Our primary backbone is Llama 3 with 8 billion parameters (Grattafiori et al., 2024), a decoder-only autoregressive transformer pretrained on a large general-purpose text corpus. We adopt Llama 3 for three reasons: its 8 billion-parameter count fits comfortably on a single 80 GB accelerator under LoRA fine-tuning, its open weights make the entire pipeline reproducible from publicly available checkpoints, and its context window accommodates the serialized M&A rows used in our sample. As a robustness exercise, we also fine-tune an encoder backbone, ModernBERT (Warner et al., 2025), a 149-million-parameter bidirectional encoder;

the cross-architecture comparison is reported in our companion technical paper and reaches qualitatively similar conclusions.

Because merger waves and antitrust regimes evolve over time, randomly shuffling deals into folds would allow information about future regulatory environments to leak into the training set. We, therefore, sort all deals by announcement date and partition the sample chronologically, allocating the earliest 80% to training, the next 10% to validation, and the most recent 10% to testing. This temporal split mirrors the operational setting in which a model is trained on historical deals and deployed on future ones. It prevents future deals from entering the training sample and therefore reduces one important source of look-ahead bias. Within each training run, we set the random seed to 42 for PyTorch, NumPy, and the Hugging Face data loader, and we use the same seed across MergerLLM and every baseline so that differences in performance reflect modeling choices rather than initialization noise. We do not report multi-seed averages in the main results because the dominant source of evaluation variance is the choice of test window, which is held fixed by the chronological split. The exact seeds used to generate every table and figure are documented in the public code repository.

Each row is first converted into natural language using the deterministic serializer described in Section 3.4 and is then concatenated with the task-specific prompt “Will the M&A deal be completed or withdrawn? Answer:”. Because the serializer copies attribute values verbatim from the source row, and the language model is only ever asked to classify the resulting string, the input pipeline cannot hallucinate values. We tokenize the combined sequence with the native Llama 3 tokenizer and cap each row at 2048 tokens. This budget accommodates the structured deal variables and most serialized text fields in our sample. For observations that exceed the token budget, typically because the acquirer’s or target’s business description is unusually long, we truncate the longest free-text field first while preserving all numeric and categorical attributes.

For parameter-efficient fine-tuning, we follow Hu et al. (2021) and insert LoRA adapters into the query and value projection matrices of every transformer block, with rank $r = 8$, scaling factor $\alpha = 16$, and adapter dropout 0.05. Under these settings, the number of trainable parameters is roughly 1% of the base model, while the remaining 99% of pretrained weights remain frozen. The LoRA adapters and the classification head are optimized jointly using AdamW (Loshchilov & Hutter, 2019) under the cross-entropy loss in Equation (2), with a learning rate of 2×10^{-5} , weight decay 0.01, and a per-device batch size of 16. We deliberately omit a learning-rate schedule, gradient clipping, and label smoothing from the main specification so that the reported gains can be attributed to the serialization and adaptation strategy rather than to additional optimization tricks.

We train for four epochs and evaluate validation loss after every epoch, retaining the best-performing checkpoint as the final model. This serves as our early-stopping rule and protects against overfitting on a dataset where the withdrawal class accounts for only 5.62% of observations. To handle the resulting class imbalance at inference time, we select the binary decision threshold that maximizes the validation F1-score and apply this same threshold, unchanged, to the held-out test set. We do not use oversampling, undersampling, or synthetic minority-class generation in the main specification; instead, we address class imbalance through validation-based threshold selection and by reporting recall, F1-score, PR AUC, threshold sensitivity, cost-sensitive loss, and calibration in the evaluation. The identical threshold-tuning protocol is applied to every baseline model in Section 4, so threshold selection cannot bias the comparisons in favor of MergerLLM. All training and inference runs are executed on a single NVIDIA A100-SXM4-80GB GPU using PyTorch, the Hugging Face transformers library, and the peft library for LoRA; a full four-epoch

training pass completes in approximately seven hours, and inference on the test set takes under fifteen min.

To limit look-ahead concerns, we define the prediction date as the merger announcement date and construct classifier inputs from variables that describe the announced transaction. We exclude variables that are mechanically realized only after completion or withdrawal, as well as post-merger operating performance, from the withdrawal classifier. Post-merger operating performance is used only later to construct the model–market disagreement measure. Some transaction-database fields may be updated during the deal process, especially variables related to challenged-deal status or revised transaction terms. We, therefore, interpret the prediction results as being based on the announced-deal information recorded in SDC while recognizing that future work using historical database snapshots could further refine the timing of information availability.

3.7. Baseline Models

We implement the conventional machine-learning baselines using AutoGluon with its default tabular training pipeline. The evaluated models include LightGBM, XGBoost, CatBoost, random forests, extra-trees models, multilayer perceptrons, FastAI neural networks, and k -nearest-neighbor classifiers. AutoGluon identifies variable types and applies the pre-processing and missing-value handling procedures associated with its tabular models. We use the same chronological training, validation, and test samples for all model comparisons.

The dataset also contains text fields, including deal-type descriptions and business descriptions for the acquiring and target firms. To incorporate these fields into the conventional baseline models, we represent them using TF-IDF features and combine the resulting text features with the structured deal variables.

Beyond the conventional AutoGluon baselines, we also evaluate SAINT, a transformer architecture designed for tabular data (Somepalli et al., 2021), and TabNet, which uses sequential attention and feature selection for tabular prediction (Arik & Pfister, 2021). We do not include TabPFN because the implementation available for this study could not accommodate a training sample of our size.

4. Results

4.1. Comparison Between Models

In this study, we evaluated the performance of MergerLLM, a novel framework leveraging large language models (LLMs) for the classification of tabular data. We compared MergerLLM against several traditional machine learning models, including LightGBM, XGBoost, Random Forest, CatBoost, and neural network models. Our evaluation was based on several metrics, each of which provides different insights into the model's performance.

- Accuracy represents the proportion of correct predictions made by the model, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP and TN are true positives and true negatives, and FP and FN are false positives and false negatives, respectively.

- Precision measures the proportion of true positive predictions among all predicted positives, and it is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall refers to the proportion of actual positives correctly identified by the model, and it is given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-score is the harmonic mean of precision and recall, offering a balance between the two when there is an uneven class distribution, and it is calculated as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- ROC AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which illustrates the trade-off between true positive and false positive rates. A higher ROC AUC indicates better model performance.
- PR AUC refers to the area under the precision–recall curve, which is useful in cases of imbalanced datasets, highlighting the trade-off between precision and recall.

The performance metrics for each model are summarized in Tables 3 and 4, while the detailed test results for MergerLLM are visualized using the confusion matrix in Figure 3.

The dataset used in this study is highly imbalanced, with a ratio of withdrawn deals to completed deals being approximately 1:18. This imbalance presents a significant challenge for traditional classification models, which are often biased toward the majority class. The precision–recall area under the curve (PR AUC) is especially informative in this setting because it focuses on performance for the minority withdrawal class. Similarly, the F1-score summarizes the trade-off between precision and recall, which is useful when the minority class is economically important. These two metrics (PR AUC and F1 score) are well suited for imbalanced datasets, offering clearer insights into the model’s ability to handle the minority class compared to traditional metrics like accuracy.

Table 3. Model performance on validation data. This table reports validation-set performance for MergerLLM and the baseline models. The metrics include accuracy, precision, recall, F1-score, ROC AUC, and PR AUC. MergerLLM achieves the highest recall and F1-score on the validation set, suggesting stronger detection of withdrawn deals. Several baseline models remain competitive in ROC AUC and PR AUC, indicating that MergerLLM’s main advantage is concentrated in minority-class classification.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC	PR AUC
LightGBM–Extra Trees	0.963	0.674	0.210	0.320	0.934	0.486
XGBoost	0.963	0.637	0.252	0.361	0.936	0.510
LightGBM–Large	0.962	0.621	0.249	0.356	0.922	0.478
MLP	0.962	0.655	0.180	0.283	0.906	0.412
LightGBM	0.962	0.630	0.201	0.305	0.922	0.465
Random Forest–Gini	0.962	0.770	0.109	0.190	0.922	0.502
Random Forest–Entropy	0.961	0.788	0.095	0.169	0.933	0.524
Extra Trees–Gini	0.961	0.750	0.104	0.183	0.912	0.436
Extra Trees–Entropy	0.961	0.774	0.095	0.169	0.915	0.450
Neural Network–FastAI	0.961	0.616	0.159	0.253	0.925	0.437
CatBoost	0.959	0.800	0.018	0.036	0.902	0.336
K-Nearest Neighbors–Distance Weights	0.955	0.381	0.129	0.193	0.677	0.163
K-Nearest Neighbors–Uniform Weights	0.955	0.370	0.125	0.187	0.673	0.114
SAINT	0.961	0.579	0.286	0.383	0.912	0.419
TabNet	0.958	0.471	0.076	0.131	0.910	0.370
MergerLLM (Ours)	0.962	0.544	0.538	0.541	0.934	0.463

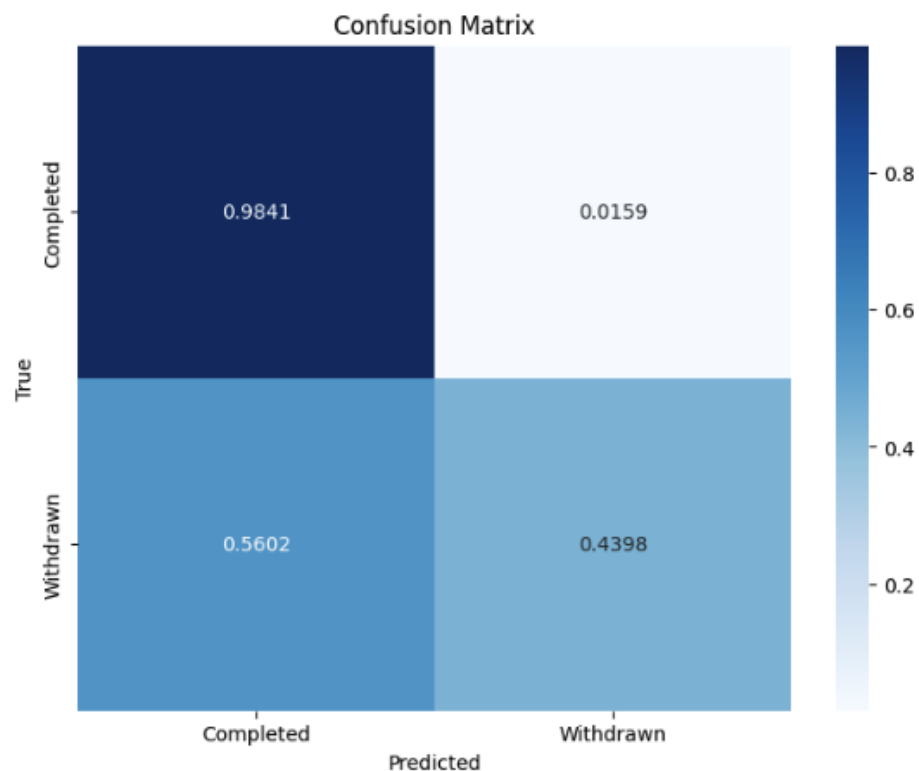


Figure 3. Confusion matrix on the test set. This figure reports the row-normalized confusion matrix for MergerLLM on the test set. The model classifies most completed deals correctly, but it also misses a meaningful share of withdrawn deals. This motivates the threshold-sensitivity, cost-sensitive-loss, and calibration analyses reported below.

Table 4. Model performance on test data. This table reports test-set performance for MergerLLM and the baseline models. The metrics include accuracy, precision, recall, F1-score, ROC AUC, and PR AUC. MergerLLM achieves the highest recall and F1-score among the evaluated models, indicating stronger detection of the rare withdrawal class. Its ROC AUC and PR AUC remain competitive with the strongest baseline models, although the PR AUC improvement over LightGBM–Large and SAINT is modest.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC	PR AUC
LightGBM–Extra Trees	0.950	0.555	0.232	0.327	0.914	0.430
XGBoost	0.947	0.490	0.219	0.303	0.906	0.408
LightGBM–Large	0.950	0.578	0.217	0.316	0.906	0.455
MLP	0.948	0.562	0.075	0.132	0.893	0.395
LightGBM	0.949	0.548	0.168	0.257	0.904	0.437
Random Forest–Gini	0.948	0.639	0.042	0.079	0.901	0.430
Random Forest–Entropy	0.948	0.769	0.018	0.036	0.908	0.436
Extra Trees–Gini	0.948	0.688	0.020	0.039	0.886	0.384
Extra Trees–Entropy	0.948	0.833	0.018	0.036	0.894	0.391
Neural Network–FastAI	0.949	0.662	0.078	0.140	0.896	0.390
CatBoost	0.948	0.750	0.016	0.032	0.798	0.308
K-Nearest Neighbors–Distance Weights	0.934	0.275	0.153	0.197	0.583	0.119
K-Nearest Neighbors–Uniform Weights	0.934	0.275	0.153	0.197	0.592	0.110
SAINT	0.950	0.563	0.277	0.372	0.890	0.451
TabNet	0.947	0.494	0.148	0.228	0.895	0.385
MergerLLM (Ours)	0.955	0.606	0.440	0.510	0.905	0.465

As presented in Table 3, MergerLLM performs strongly on the validation set, especially in detecting the minority withdrawal class. Its accuracy is comparable to the strongest baseline models, while its recall and F1-score are higher than those of the other models. This pattern suggests that MergerLLM is better able to identify withdrawn deals, which are rare but economically important. The PR AUC results also show that several tree-based models remain highly competitive in ranking observations by withdrawal risk. Overall, the validation results indicate that MergerLLM's strongest advantage is in minority-class detection.

Table 4 reports the corresponding test-set results. MergerLLM performs strongest on threshold-based minority-class metrics. It achieves the highest recall and F1-score among the evaluated models, indicating better detection of withdrawn deals at the validation-selected threshold. Its ROC AUC and PR AUC remain competitive with the strongest tree-based and tabular deep-learning benchmarks, although the PR AUC difference is modest. These results suggest that MergerLLM's main advantage lies in classifying rare withdrawal events more effectively. But it does not uniformly dominate all models across every ranking metric.

The baseline models provide useful benchmarks. LightGBM, XGBoost, and SAINT achieve strong AUC and PR AUC results, confirming that standard tabular models remain competitive in ranking merger-withdrawal risk. The comparison, therefore, suggests that MergerLLM's clearest gains appear in recall and F1-score, while ranking-based performance remains closer to the strongest baselines.

Figure 3 reports the normalized confusion matrix for the test set. MergerLLM classifies most completed deals correctly and identifies a sizable portion of withdrawn deals. At the same time, some withdrawn deals are classified as completed, reflecting the difficulty of predicting rare withdrawal events in a highly imbalanced sample. In the context of merger-risk prediction, this pattern highlights the importance of considering both types of classification errors. False negatives correspond to withdrawn deals classified as completed, while false positives correspond to completed deals classified as withdrawn. The relative importance of these errors may vary across applications, especially for investors with different risk preferences.

4.2. Threshold Sensitivity, Cost-Sensitive Loss, and Calibration

The confusion-matrix evidence motivates a more decision-oriented evaluation. In merger-risk prediction, the relative cost of false negatives and false positives may differ across applications. A false negative corresponds to a withdrawn deal classified as likely to complete, while a false positive corresponds to a completed deal classified as withdrawn. Investors who are especially concerned about avoiding failed deals may, therefore, prefer a lower decision threshold than the default threshold of 0.50.

Table 5 reports MergerLLM's test-set performance across alternative classification thresholds. The results show a clear precision–recall trade-off. At the 0.50 threshold, precision is 0.620 and recall is 0.378, with a false-negative rate of 0.622. Lowering the threshold increases recall and reduces false negatives. For example, at the 0.10 threshold, recall increases to 0.493 and the false-negative rate declines to 0.507, while the false-positive rate increases only from 0.013 to 0.020. At the 0.05 threshold, recall further increases to 0.542 and the false-negative rate falls to 0.458, although precision declines to 0.519. These results indicate that the default threshold is conservative in this imbalanced setting and that lower thresholds may be preferable when missing withdrawn deals is costly.

Table 5. Threshold sensitivity analysis. This table reports the test-set performance of MergerLLM across alternative classification thresholds. The predicted probability is the model-implied probability of merger withdrawal. A deal is classified as withdrawn when the predicted withdrawal probability is greater than or equal to the threshold. FPR is the false-positive rate, defined as completed deals incorrectly classified as withdrawn. FNR is the false-negative rate, defined as withdrawn deals incorrectly classified as completed. Pred. Withdraw is the share of test observations classified as withdrawn.

Threshold	Precision	Recall	F1	FPR	FNR	Pred. Withdraw
0.05	0.519	0.542	0.530	0.028	0.458	0.055
0.10	0.573	0.493	0.530	0.020	0.507	0.045
0.20	0.605	0.438	0.508	0.016	0.562	0.038
0.30	0.612	0.412	0.493	0.015	0.588	0.035
0.40	0.619	0.394	0.482	0.014	0.606	0.034
0.50	0.620	0.378	0.469	0.013	0.622	0.032
0.60	0.620	0.349	0.446	0.012	0.651	0.030
0.70	0.633	0.327	0.431	0.011	0.673	0.027
0.80	0.647	0.241	0.351	0.007	0.759	0.020
0.90	0.674	0.113	0.194	0.003	0.887	0.009

Table 6 reports the normalized expected loss under different false-negative to false-positive cost ratios. When false negatives and false positives receive equal weight, thresholds between 0.20 and 0.40 generate similar expected loss. However, when false negatives receive greater weight, the lower threshold performs better among the thresholds reported in the table. For example, when a false negative is assigned ten times the cost of a false positive, the expected loss is 0.2867 at the 0.10 threshold, compared with 0.3402 at the 0.50 threshold. This pattern is consistent with the economic intuition that investors with greater concern about missed withdrawals may prefer a more conservative classification rule that flags more deals as risky.

Table 6. Cost-sensitive expected loss. This table reports normalized expected loss for MergerLLM under alternative classification thresholds and different false-negative to false-positive cost ratios. False negatives are withdrawn deals incorrectly classified as completed, and false positives are completed deals incorrectly classified as withdrawn. Expected loss is defined as $(c_{FN}FN + c_{FP}FP)/N$, where c_{FN} is the false-negative cost, c_{FP} is the false-positive cost, and N is the number of test observations.

Threshold	$c_{FN}/c_{FP} = 1$	$c_{FN}/c_{FP} = 2$	$c_{FN}/c_{FP} = 5$	$c_{FN}/c_{FP} = 10$
0.10	0.0461	0.0728	0.1530	0.2867
0.20	0.0447	0.0743	0.1632	0.3113
0.30	0.0447	0.0757	0.1686	0.3234
0.40	0.0447	0.0766	0.1724	0.3321
0.50	0.0450	0.0778	0.1762	0.3402

Finally, we examine the calibration of MergerLLM's predicted withdrawal probabilities. Figure 4 and Table 7 compare average predicted withdrawal probabilities with realized withdrawal rates across probability bins. The model provides useful ranking information: the highest predicted-risk bin has an actual withdrawal rate of 0.3577, much higher than the unconditional withdrawal rate in the test sample. However, the raw probabilities are not perfectly calibrated. In most bins, the realized withdrawal rate is higher than the average predicted withdrawal probability. For example, in the highest-risk bin, the mean predicted withdrawal probability is 0.3086, while the realized withdrawal rate is 0.3577. The Brier score is 0.0388, while the ROC AUC and PR AUC are 0.9046 and 0.465, respectively. Overall, these results suggest that MergerLLM is useful for ranking merger-withdrawal

risk and identifying higher-risk transactions, but its raw predicted probabilities should be interpreted primarily as risk scores unless additional probability calibration is applied.

Table 7. Calibration of predicted withdrawal probabilities. This table reports calibration results for MergerLLM on the test set. Observations are sorted into ten bins based on predicted withdrawal probabilities. Mean Pred. Prob. is the average predicted probability of merger withdrawal in each bin. The Actual Withdraw Rate is the realized share of withdrawn deals in the same bin. The Calibration Error is defined as Mean Pred. Prob. minus the Actual Withdraw Rate. Negative values indicate that the model underpredicts the realized withdrawal rate in that bin.

Probability Bin	Mean Pred. Prob.	Actual Withdraw Rate	Calib. Error
[0.000000, 0.000223]	0.0002	0.0038	−0.0037
(0.000223, 0.000314]	0.0003	0.0058	−0.0055
(0.000314, 0.000407]	0.0004	0.0029	−0.0025
(0.000407, 0.000533]	0.0005	0.0077	−0.0072
(0.000533, 0.000706]	0.0006	0.0058	−0.0052
(0.000706, 0.000976]	0.0008	0.0115	−0.0107
(0.000976, 0.001440]	0.0012	0.0135	−0.0123
(0.001440, 0.002860]	0.0020	0.0375	−0.0355
(0.002860, 0.012600]	0.0058	0.0808	−0.0749
(0.012600, 0.980000]	0.3086	0.3577	−0.0491

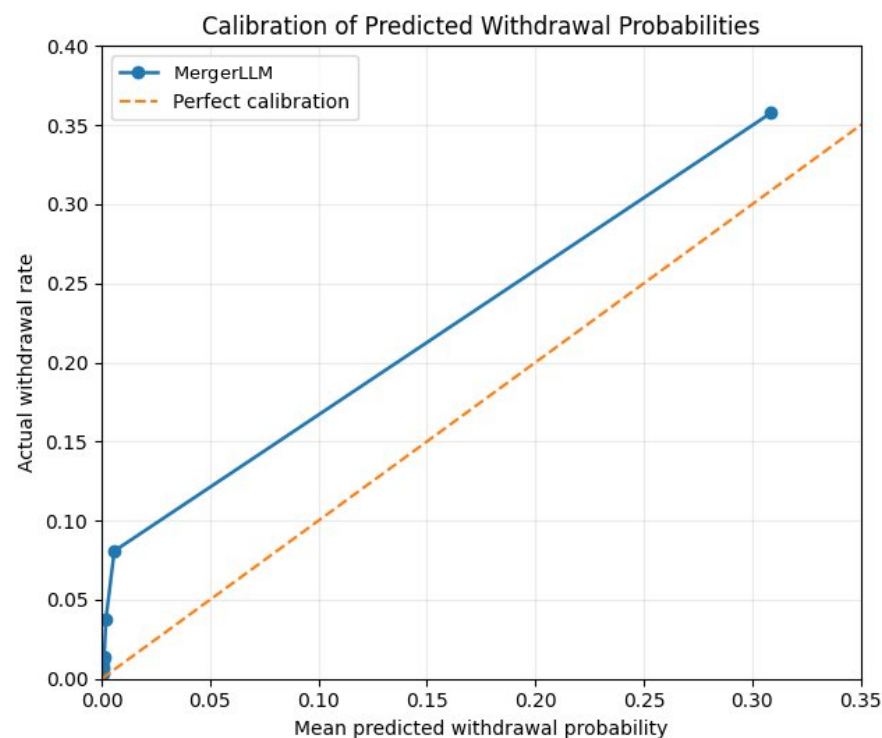


Figure 4. Calibration of predicted withdrawal probabilities. This figure plots the average predicted withdrawal probability against the realized withdrawal rate by probability bin. The dashed line represents perfect calibration. The figure is zoomed in to the range where most predicted probabilities are concentrated.

Overall, the results indicate that MergerLLM is particularly useful for detecting the minority withdrawal class. Its main advantage appears in recall and F1-score, which are especially relevant when withdrawn deals are rare but economically important. The threshold-sensitivity and cost-sensitive analyses further show that different decision thresholds may be preferable depending on the relative cost of false negatives and false positives. Thus, MergerLLM provides more balanced detection of withdrawn deals while maintaining ranking performance comparable to strong machine-learning and deep-learning baselines.

Figure 5 plots the precision–recall curves for MergerLLM and the strongest baseline models on the test set, providing a visual comparison of minority-class ranking performance.

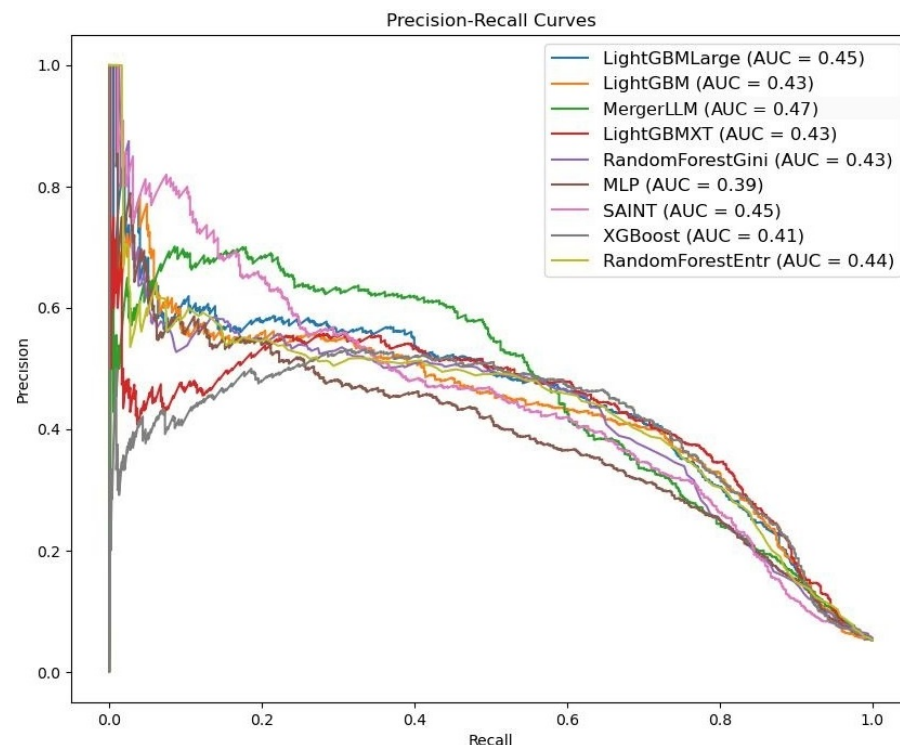


Figure 5. Precision–recall curves. This figure plots precision–recall curves for MergerLLM and the top baseline models on the test set. MergerLLM achieves a PR AUC of approximately 0.47, which is competitive with the strongest baseline models.

4.3. Feature Importance

We next examine which input variables are most closely associated with MergerLLM’s predictions. Because MergerLLM operates on serialized text instead of a conventional feature matrix, standard permutation-based feature-importance methods cannot be applied to the model in the same direct way as they are applied to tree-based or other tabular models. We, therefore, use a surrogate feature-importance approach. Specifically, we fit an ensemble model using the original deal-level variables as covariates to approximate MergerLLM’s predicted outcomes and then compute permutation-based feature importance using AutoGluon. The resulting importance scores provide an interpretable summary of which original variables are most closely related to MergerLLM’s predictions.

The permutation-shuffling method measures the reduction in predictive performance when the values of a given feature are randomly shuffled across observations, while all other variables remain unchanged. A larger decline in performance indicates that the feature contains more information for explaining the model’s predictions. The importance weights are multiplied by 100 to improve readability. This approach provides an interpretable summary of the variables most closely related to MergerLLM’s predictions, while the ablation tests in the next subsection provide additional evidence on the role of text features and feature-name semantics.

Table 8 reports the top seven features from this analysis and compares their ranks with those from the baseline ensemble model. The highest-ranked variables are the payment method, the target public status, the percentage of shares sought, the acquirer market value, the target business description, the acquirer business description, and the challenged deal. These variables are economically plausible predictors of merger withdrawal and are consistent with prior M&A research. For example, payment method may affect completion

risk because stock-based offers are more sensitive to market movements and valuation uncertainty, which can increase the likelihood of renegotiation or withdrawal (Bates & Lemmon, 2003). The target public status may also matter because acquisitions of public targets often face greater regulatory scrutiny and shareholder approval requirements, increasing completion risk (Officer, 2003).

Table 8. Surrogate feature importance for MergerLLM predictions. This table reports the top seven variables from a surrogate feature-importance analysis of MergerLLM predictions. The surrogate model uses the original deal-level variables as covariates, and feature importance is computed using permutation-based importance from AutoGluon. The table also reports the corresponding ranks and weights from the baseline ensemble model for comparison. The importance weights are multiplied by 100 for readability.

Feature	Surrogate Rank	Surrogate Weight	Ensemble Rank	Ensemble Weight
Payment Method	1.0	3.400	3.0	0.200
Target Public Status	2.0	3.200	1.0	0.500
Percent Shares Sought	3.0	3.000	5.0	0.100
Acquirer Market Value	4.0	0.900	28.0	−0.100
Target Business Description	5.0	0.300	7.0	0.000
Acquirer Business Description	6.0	0.300	30.0	−0.400
Challenged Deal	7.0	0.300	2.0	0.300

The feature rankings also highlight the role of textual information. Target and acquirer business descriptions appear among the top features associated with MergerLLM's predictions, suggesting that firm-level textual descriptions contain useful information for predicting deal completion. This finding is consistent with the design of MergerLLM, which combines structured transaction variables with text-based firm information through natural-language serialization. The next subsection further examines this point using ablation tests that remove text features and permute feature names.

4.4. Ablation Analysis and Feature-Name Sensitivity

To better understand the sources of MergerLLM's predictive performance, we conduct two ablation tests. First, we remove text-based features to examine whether firm descriptions contribute to the model's performance. Second, we permute feature names while keeping feature values unchanged to evaluate whether MergerLLM uses the semantic information embedded in column names. These tests complement the feature-importance analysis by providing evidence that is more directly tied to the model's input structure.

4.4.1. Removing Text Features

Table 9 reports the ablation results after removing text-based features, including deal-level descriptions and the acquirer and target business descriptions. Compared with the full MergerLLM specification, the no-text version shows lower PR AUC, precision, F1-score, and ROC AUC. Specifically, PR AUC decreases from 0.465 to 0.400, precision decreases from 0.606 to 0.351, F1-score decreases from 0.510 to 0.408, and ROC AUC decreases from 0.905 to 0.808. These results suggest that deal narratives and firm-level business descriptions provide useful information for distinguishing completed deals from withdrawn deals. Interestingly, recall increases from 0.440 to 0.487 after removing text features, indicating that the no-text model identifies a larger share of withdrawn deals but does so with lower precision. Overall, the ablation results suggest that text-based deal and firm information improves the model's classification quality and ranking performance, while structured transaction variables alone still contain meaningful predictive information.

Table 9. Effect of removing text features. This table compares the full MergerLLM model with a no-text version that removes deal-level text descriptions and acquirer and target business descriptions. Model performance is reported across five evaluation metrics: PR AUC, precision, recall, F1-score, and ROC AUC.

Serialization Strategy	PR AUC	Precision	Recall	F1-Score	ROC AUC
MergerLLM (Ours)	0.465	0.606	0.440	0.510	0.905
MergerLLM (No Text)	0.400	0.351	0.487	0.408	0.808

4.4.2. Permuting Column Names

Figure 6 compares the performance of MergerLLM with a version in which feature column names are permuted while the underlying feature values remain unchanged. This comparison evaluates the extent to which the model uses the semantic information contained in column names when making predictions. The results show that permuting column names reduces performance on several key metrics, including precision, recall, F1-score, ROC AUC, and PR AUC, while accuracy changes only slightly. These findings suggest that semantic information embedded in feature names contributes to MergerLLM’s classification and ranking performance.

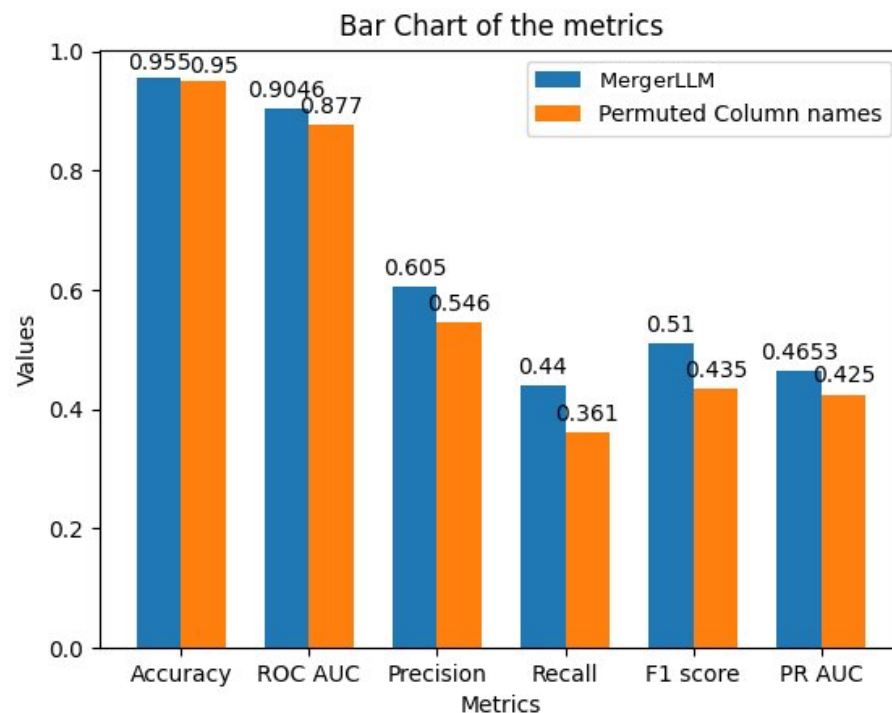


Figure 6. Performance under original and permuted column names. This figure compares MergerLLM with a version in which feature column names are permuted while the underlying feature values remain unchanged. The comparison is reported across six evaluation metrics: accuracy, ROC AUC, precision, recall, F1-score, and PR AUC. The results show that permuting column names reduces several classification and ranking metrics, suggesting that meaningful feature names provide useful semantic information to the LLM.

Column names in tabular datasets often contain useful contextual information. For example, variables such as “rumored,” “challenged,” and “payment method” convey economic meaning beyond their numerical or categorical values. Permuting these names removes this semantic context while preserving the underlying data structure. The observed decline in classification and ranking metrics suggests that meaningful feature names help MergerLLM connect transaction attributes with their economic interpretation.

The column-name permutation test also helps clarify the source of MergerLLM’s performance. Because the model is pretrained on large text corpora, meaningful feature names may allow it to use general semantic knowledge in addition to financial relationships learned during fine-tuning. For example, terms such as “challenged,” “rumored,” and “payment method” help define the economic meaning of the corresponding values. This is consistent with the purpose of serialization, which is to preserve the semantic content of tabular inputs rather than treating feature names as arbitrary labels. The results suggest that MergerLLM’s predictions may reflect both patterns learned from the M&A data and pretrained language knowledge about financial and deal-related terms.

A further test would separate feature-name semantics from variable-definition information by replacing column names with neutral labels while providing detailed variable descriptions separately. This design would help distinguish the effect of semantic feature names from the effect of explicit variable definitions. We leave this additional prompt-design research for future work.

4.5. Economic Utility

Merger withdrawals can pose significant financial risks, including substantial financial penalties and market backlash (Officer, 2003). They can also damage a company’s reputation, increase debt burdens, elevate the cost of capital (Harford, 2005), and lead to higher legal and consulting expenses. Forecasting merger withdrawals is relevant for merger-arbitrage and event-driven trading strategies, which depend on the relation between deal-completion risk and post-announcement returns. If MergerLLM’s predicted success probabilities contain economic information, acquirers in the highest predicted-success quintile (Q5) should earn higher subsequent returns than acquirers in the lowest predicted-success quintile (Q1).

Given MergerLLM’s strong performance in predicting merger success and withdrawal, we examine whether its predicted success probabilities are associated with subsequent acquirer returns. Table 10 reports raw post-announcement stock returns for acquirers in the highest and lowest predicted-success quintiles. Acquirers in Q5 earn 0.65% in the first month, 1.20% in the third month, and 1.91% over the three months following the announcement. In contrast, acquirers in Q1 earn −0.32% in the first month, −1.03% in the third month, and −1.43% over the same three-month window. The Q5-minus-Q1 spread is, therefore, 3.34% over three months. Because these are raw returns and do not account for risk adjustment, transaction costs, short-selling constraints, liquidity, bid–ask spreads, or other implementation frictions, we interpret this evidence as an initial indication that MergerLLM’s predicted success probabilities contain economically meaningful information.

Table 10. Merger arbitrage strategy. This table reports post-announcement returns for acquiring firms in the highest and lowest MergerLLM predicted-success quintiles. Returns are reported for the first month, the third month, and the full three-month period after the announcement. A *t*-test is used to assess whether the mean return difference between the two quintiles is statistically significant.

Company	Time Frame	Mean (Highest Quintile)	Mean (Lowest Quintile)	Difference	t-Statistics	p-Value
Acquiring Company	First Month	0.0065	−0.0032	0.0097	2.3519	0.0187
	Third Month	0.0120	−0.0103	0.0223	5.2313	<0.0001
	Three Months	0.0191	−0.0143	0.0334	4.8039	<0.0001

Table 11 compares the economic performance of MergerLLM with representative benchmark models from Table 4. Across the first-month, third-month, and three-month horizons, MergerLLM generates the largest return spread between the highest predicted-success quintile (Q5) and the lowest predicted-success quintile (Q1). This result suggests

that MergerLLM’s predicted probabilities provide economically meaningful rankings of merger outcomes relative to selected baseline models. However, these return spreads are based on raw post-announcement returns and do not account for common risk factors or portfolio implementation issues. We, therefore, complement the raw return analysis with calendar-time portfolio tests that report risk-adjusted alphas, drawdowns, and turnover.

Table 11. Model comparisons on portfolio analysis. This table compares the portfolio performance of MergerLLM with selected benchmark models. Deals are sorted into quintiles based on predicted merger-success probabilities. The table reports returns for the highest and lowest quintiles across different holding horizons, along with return differences, t-statistics, and *p*-values. Bold values indicate the largest return difference within each holding horizon among the models reported in the table.

Model	Time Frame	Mean (Highest Quintile)	Mean (Lowest Quintile)	Difference	t-Statistic	<i>p</i> -Value
MergerLLM	First Month	0.0065	−0.0032	0.0097	2.3519	0.0187
	Third Month	0.0120	−0.0103	0.0223	5.2313	<0.0001
	Three Months	0.0191	−0.0143	0.0334	4.8039	<0.0001
LightGBM–Extra Trees	First Month	0.0007	0.0049	−0.0042	−1.0586	0.2898
	Third Month	0.0079	−0.0020	0.0098	2.3608	0.0182
	Three Months	0.0258	0.0053	0.0205	3.3865	0.0007
Random Forest–Entropy	First Month	0.0139	0.0048	0.0090	1.9551	0.0506
	Third Month	0.0116	−0.0040	0.0120	2.9027	0.0037
	Three Months	0.0223	0.0019	0.0204	3.2634	0.0011
K-Nearest Neighbors–Uniform Weights	First Month	−0.0055	0.0070	−0.0124	−3.0223	0.0025
	Third Month	0.0060	−0.0027	0.0087	2.0881	0.0369
	Three Months	0.0089	0.0102	−0.0012	−0.1957	0.8448
Neural Network–FastAI	First Month	0.0071	0.0005	0.0065	1.6199	0.1053
	Third Month	0.0091	−0.0005	0.0097	2.4233	0.0154
	Three Months	0.0256	−0.0043	0.0299	5.3053	<0.0001
SAINT	First Month	0.0031	0.0070	−0.0039	−0.8902	0.3734
	Third Month	0.0042	0.0011	0.0030	0.7422	0.4580
	Three Months	0.0118	0.0099	0.0019	0.2849	0.7756

Risk-Adjusted Calendar-Time Portfolio Performance

To further evaluate the economic value of MergerLLM’s predictions, we construct calendar-time long-short portfolios. In each month, we take long positions in acquirers in the highest predicted-success quintile (Q5) and short positions in acquirers in the lowest predicted-success quintile (Q1). The primary specification uses a three-month post-announcement holding period. Specifically, positions associated with each deal remain in the portfolio during the first, second, and third months following the announcement. Because new deals enter the portfolio in each calendar month, this design produces overlapping monthly cohorts. We also examine one-month and third-month holding horizons to assess the sensitivity of the results to the timing of post-announcement returns. Standard errors for the alpha regressions are adjusted using Newey–West corrections with three lags.

Table 12 reports the calendar-time portfolio results. The main three-month strategy earns an average monthly long-short return of 0.98%, corresponding to an approximate annualized return of 11.82%. The strategy has an annualized Sharpe ratio of 1.16, a maximum drawdown of −12.51%, and an average monthly turnover of 33.69%. The CAPM alpha is 0.81% per month and is statistically significant. The Fama–French three-factor alpha is positive but not statistically significant, while the Fama–French five-factor alpha is

0.56% per month and is marginally significant. These results indicate that the MergerLLM-based long-short portfolio generates positive calendar-time returns, although the estimated risk-adjusted performance varies across factor specifications.

Table 12. Calendar-Time Long-Short Portfolio Performance. This table reports calendar-time performance for Q5-minus-Q1 long-short portfolios based on MergerLLM predicted merger-success probabilities. Q5 contains deals with the highest predicted success probabilities, and Q1 contains deals with the lowest predicted success probabilities. Mean returns, annualized returns, maximum drawdowns, turnover, and alphas are reported in percentage points. CAPM, Fama–French three-factor, and Fama–French five-factor alphas are monthly alphas. Newey–West standard errors with three lags are used for alpha regressions. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Strategy	Months	Mean	Ann. Ret.	Sharpe	Max DD	Turnover	CAPM α	FF3 α	FF5 α
First Month	47	1.13	13.51	0.98	−13.04	91.94	0.86 **	0.59	0.73 *
Third Month	45	1.80	21.57	1.89	−6.69	92.13	1.73 ***	1.31 ***	1.22 ***
Three Months	47	0.98	11.82	1.16	−12.51	33.69	0.81 **	0.48	0.56 *

The alternative horizons provide additional evidence on the timing of the return spread. The first-month portfolio generates a positive average return and a statistically significant CAPM alpha, although the Fama–French alphas are weaker. The third-month portfolio exhibits the strongest risk-adjusted performance, with positive and statistically significant alphas under the CAPM, Fama–French three-factor, and Fama–French five-factor models. This pattern suggests that MergerLLM’s predicted success probabilities are associated with post-announcement return differences that are not limited to the first month after the announcement.

Table 13 reports the subperiod performance for the calendar-time long-short portfolios. The three-month strategy performs more strongly in the later subperiod. During 2018–2020, the strategy earns a small positive average monthly return of 0.28%, but the risk-adjusted alphas are not statistically significant. During 2021–2023, the average monthly return increases to 1.61%, and the CAPM, Fama–French three-factor, and Fama–French five-factor alphas are all positive and statistically significant. The first-month and third-month strategies show a similar pattern, with stronger performance in 2021–2023 than in 2018–2020. These results indicate that the economic value of MergerLLM’s predictions is stronger in the later part of the test period.

Table 13. Subperiod Calendar-Time Long-Short Portfolio Performance. This table reports subperiod performance for Q5-minus-Q1 long-short portfolios based on MergerLLM predicted merger-success probabilities. Q5 contains deals with the highest predicted success probabilities, and Q1 contains deals with the lowest predicted success probabilities. Mean returns, annualized returns, maximum drawdowns, turnover, and alphas are reported in percentage points. CAPM, Fama–French three-factor, and Fama–French five-factor alphas are monthly alphas. Newey–West standard errors with three lags are used for alpha regressions. * and *** denote statistical significance at the 10% and 1% levels, respectively.

Strategy	Subperiod	Months	Mean	Ann. Ret.	Sharpe	Max DD	Turnover	CAPM α	FF3 α	FF5 α
First Month	2018–2020	22	0.48	5.75	0.35	−13.04	93.24	0.36	−0.38	0.00
First Month	2021–2023	25	1.69	20.33	1.84	−7.51	90.84	1.36 ***	1.31 ***	1.21 ***
Third Month	2018–2020	20	0.74	8.93	0.88	−6.69	93.26	0.89 *	−0.05	0.36
Third Month	2021–2023	25	2.64	31.69	2.70	−5.71	91.23	2.41 ***	2.26 ***	2.02 ***
Three Months	2018–2020	22	0.28	3.31	0.33	−12.51	35.18	0.29	−0.17	0.54
Three Months	2021–2023	25	1.61	19.30	1.94	−8.68	32.44	1.32 ***	1.22 ***	1.03 ***

We next compare the MergerLLM calendar-time strategy with an external merger-arbitrage benchmark. We use MRGR, a merger-arbitrage ETF, as an investable benchmark

over the same active monthly sample period. MRGR provides a closer reference point than a broad equity index because it reflects merger-arbitrage performance. The benchmark serves as an external reference point, highlighting the distinct mechanics of our approach. While standard merger-arbitrage strategies often rely on target-side deal spreads and specific hedging, leverage, or implementation rules, the MergerLLM strategy specifically sorts acquirers by model-implied deal-success probabilities.

Table 14 reports the full-sample benchmark comparison. Over the overlapping sample from July 2018 to November 2023, the MergerLLM Q5–Q1 portfolio earns an average monthly return of 0.98%, compared with 0.29% for MRGR. The return difference is 0.70% per month, or approximately 8.39% per year. The MergerLLM strategy has a CAPM alpha of 0.81% per month and a Fama–French five-factor alpha of 0.56% per month, while the Fama–French three-factor alpha is positive but not statistically significant. The direct MergerLLM-minus-MRGR spread is also positive, although it is not statistically significant under the factor models. These results suggest that MergerLLM predictions contain economically meaningful information, although the statistical evidence for benchmark-adjusted outperformance is weaker after factor adjustment.

Table 14. Calendar-Time Portfolio Performance Relative to Merger-Arbitrage Benchmark. This table compares the MergerLLM Q5-minus-Q1 calendar-time long-short portfolio with MRGR, a merger-arbitrage ETF, over the overlapping monthly sample period. MRGR is used as an investable benchmark for merger-arbitrage performance. The MergerLLM-minus-MRGR row reports the monthly return difference between the MergerLLM strategy and MRGR. Mean returns, annualized returns, maximum drawdowns, and alphas are reported in percentage points. CAPM, Fama–French three-factor, and Fama–French five-factor alphas are monthly alphas. Newey–West standard errors with three lags are used for alpha regressions. * and ** denote statistical significance at the 10% and 5% levels, respectively.

Portfolio	Months	Mean	Ann. Ret.	Sharpe	Max DD	CAPM α	FF3 α	FF5 α
MergerLLM Q5–Q1	47	0.98	11.82	1.16	−12.51	0.81 **	0.48	0.56 *
MRGR ETF	47	0.29	3.42	0.37	−6.17	0.04	−0.01	0.04
MergerLLM minus MRGR	47	0.70	8.39	0.81	−15.41	0.62	0.34	0.35

Table 15 reports the corresponding subperiod benchmark comparison. MergerLLM earns higher average monthly returns than MRGR in each subperiod, but the strength of the benchmark-adjusted performance varies over time. During 2018–2019, MergerLLM earns 1.11% per month, compared with 0.67% for MRGR. During 2020–2021, the difference is smaller: MergerLLM earns 0.47% per month, compared with 0.35% for MRGR. The strongest benchmark-adjusted performance occurs in 2022–2023. During this period, MergerLLM earns 1.35% per month, while MRGR earns −0.05% per month. The MergerLLM-minus-MRGR spread is 1.40% per month and remains statistically significant under CAPM, Fama–French three-factor, and Fama–French five-factor adjustments. Overall, the benchmark results indicate that MergerLLM predictions have economic information content, with the strongest evidence appearing in the later part of the sample.

These portfolio tests are designed to evaluate the economic information content of MergerLLM’s predicted merger-success probabilities. Building a fully implementable trading strategy would require transaction-level data on trading costs, short-selling availability, borrowing fees, bid–ask spreads, liquidity, deal duration, and delisting returns. Because these microstructural data are not consistently available across our full sample, we focus our empirical scope on whether MergerLLM predictions are associated with economically meaningful post-announcement return differences, leaving the execution constraints of a fully implementable strategy for future work.

Table 15. Subperiod Performance Relative to Merger-Arbitrage Benchmark. This table reports subperiod performance for the MergerLLM Q5-minus-Q1 calendar-time long-short portfolio, MRGR, and the MergerLLM-minus-MRGR return spread. MRGR is used as an investable benchmark for merger-arbitrage performance. Mean returns, annualized returns, maximum drawdowns, and alphas are reported in percentage points. CAPM, Fama–French three-factor, and Fama–French five-factor alphas are monthly alphas. Newey–West standard errors with three lags are used for alpha regressions. The number of months refers to active portfolio months with nonmissing MergerLLM strategy returns and benchmark returns. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Subperiod	Portfolio	Months	Mean	Ann. Ret.	Sharpe	B	CAPM α	FF3 α	FF5 α
2018–2019	MergerLLM Q5–Q1	13	1.11	13.31	1.73	−4.75	0.94 ***	1.10 ***	0.95 **
2018–2019	MRGR ETF	13	0.67	8.02	2.06	−0.26	0.48 **	0.11	−0.08
2018–2019	MergerLLM minus MRGR	13	0.44	5.29	0.64	−7.76	0.28	0.81 *	0.84 *
2020–2021	MergerLLM Q5–Q1	16	0.47	5.62	0.47	−12.51	0.67	0.07	0.50
2020–2021	MRGR ETF	16	0.35	4.16	0.73	−4.20	−0.13	−0.09	0.31 ***
2020–2021	MergerLLM minus MRGR	16	0.12	1.46	0.12	−15.41	0.77	0.15	0.17
2022–2023	MergerLLM Q5–Q1	18	1.35	16.24	1.57	−6.13	1.24 ***	1.18 ***	0.99 ***
2022–2023	MRGR ETF	18	−0.05	−0.55	−0.97	−4.48	−0.30	−0.34	−0.34
2022–2023	MergerLLM minus MRGR	18	1.40	16.79	1.61	−5.13	1.29 ***	1.27 **	1.09 ***

5. Model–Market Disagreement

5.1. HAIDiff as a Proxy for Model–Market Disagreement

Stock price reactions around M&A announcements aggregate market participants’ assessments of deal value and completion risk. Prior M&A research often uses announcement returns to study market reactions to deal announcements (Dasgupta et al., 2024; Luo, 2005; Tunyi, 2021). These reactions may contain information about expected deal completion, but they also reflect expected synergies, financing terms, overpayment concerns, governance issues, liquidity, arbitrage constraints, information asymmetry, and risk premia. Therefore, announcement returns should not be interpreted as a clean measure of investor beliefs about merger completion.

We use announcement returns to construct a market-based signal of deal assessment and compare this signal with MergerLLM’s predicted probability of merger success. This comparison is motivated by prediction-policy problems, where algorithmic predictions may provide useful information when outcomes are uncertain and the relevant information is complex (Kleinberg et al., 2015). It allows us to study whether model-implied completion probabilities differ systematically from market-based assessments, without treating either the model or the market as a bias-free benchmark.

Formally, we first estimate the following auxiliary regression:

$$CAR_{i[-1,+1]}^{Acq} = \alpha + \delta ROA_i + u_i, \quad (3)$$

where $CAR_{i[-1,+1]}^{Acq}$ is the acquirer’s three-day announcement cumulative abnormal return and ROA_i captures operating performance. The residual u_i captures the component of acquirer announcement returns that is orthogonal to the included operating-performance measure. We then transform the residual component into a scaled market-based signal:

$$MarketSignal_i = \frac{u_i - \min(u)}{\max(u) - \min(u)}. \quad (4)$$

Finally, HAIDiff is defined as:

$$HAIDiff_i = \hat{P}_i^{LLM}(Completed) - MarketSignal_i, \quad (5)$$

where $\hat{P}_i^{LLM}(Completed)$ is MergerLLM’s predicted probability that deal i will be completed. Thus, HAIDiff captures disagreement between a model-implied completion probability and a transformed market-based signal.

HAIDiff is designed to capture the distance between two different assessments of the announced deal. The first component is the model-implied probability of deal completion. The second component is a transformed market-based signal derived from announcement returns. Because announcement returns reflect expected synergies, financing terms, overpayment concerns, risk premia, liquidity, and arbitrage constraints, the market-based component should be interpreted as a broad signal of market assessment rather than a direct measure of completion beliefs.

5.2. Determinants of Model–Market Disagreement in Merger Completion Assessment

In the following analysis, we examine whether climate exposure is associated with model–market disagreement, as measured by HAIDiff. We include firm-level climate change exposure (CC Exposure) as our main variable of interest. Following Sautner et al. (2023), this variable measures the attention paid by earnings-call participants to firms' climate-related opportunities, physical risks, and regulatory shocks. Climate exposure is relevant for this test because it captures a source of uncertainty that is multidimensional, long-horizon, and difficult to summarize in a small set of transaction variables. If climate-related information is processed differently by the model and by market participants around the announcement date, climate-exposed deals should exhibit a larger HAIDiff.

Table 16 presents regression results on the determinants of model–market disagreement in merger completion assessment. The dependent variable is HAIDiff, defined as MergerLLM's predicted merger-success probability minus the scaled market-based signal derived from acquirer announcement returns. Across all specifications, the coefficient on CC Exposure is positive and statistically significant. This result indicates that climate exposure is associated with larger positive differences between model-implied completion probabilities and market-based signals. Although some macroeconomic controls are also significantly associated with HAIDiff, the coefficient on CC Exposure remains positive and statistically significant across all specifications. Overall, the results show that climate-exposed deals exhibit larger differences between model-implied completion probabilities and announcement-return-based market signals.

Table 16. Determinants of Model–Market Disagreement in Merger Completion Assessment. This table reports determinants of model–market disagreement in merger completion assessment. The dependent variable is HAIDiff, defined as MergerLLM's predicted merger-success probability minus the scaled market-based signal derived from acquirer announcement returns. CC Exposure is climate change exposure. Variable definitions are provided in Appendix A. Continuous variables, except state-level variables, are winsorized at their 1st and 99th percentiles. Standard errors are clustered at the year level, with t-statistics reported in parentheses. ** and *** denote statistical significance at the 5% and 1% levels, respectively.

Variables	(1) HAIDiff	(2) HAIDiff	(3) HAIDiff	(4) HAIDiff
CC Exposure	3.848 ** (2.45)	2.905 ** (2.08)	3.785 ** (2.40)	2.950 ** (2.20)
log(deal value)	−0.010 *** (−3.53)	−0.009 *** (−3.44)	−0.010 *** (−3.48)	−0.009 *** (−3.36)
con			−0.094 (−0.69)	−0.153 (−1.26)
ipg			−0.083 (−0.37)	−0.033 (−0.18)
tfp			0.321 *** (3.80)	0.273 *** (3.40)
term			−0.680 (−0.75)	−0.751 (−0.88)
def			−11.733 ** (−2.23)	−11.661 ** (−2.32)

Table 16. Cont.

Variables	(1) HAIDiff	(2) HAIDiff	(3) HAIDiff	(4) HAIDiff
dei			0.657 (0.15)	−0.340 (−0.07)
mkt			0.192 *** (3.39)	0.172 *** (3.08)
lab			−0.272 (−1.60)	−0.250 (−1.41)
sent			−0.004 (−0.46)	0.002 (0.23)
Observations	3689	3689	3689	3689
Adjusted R-squared	0.013	0.047	0.016	0.050
Industry FE	N	Y	N	Y
Year FE	Y	Y	Y	Y
Controls	Y	Y	Y	Y

One possible interpretation is that climate-related information is reflected differently in model-implied probabilities and in announcement-return-based market signals. Although MergerLLM does not directly use climate exposure as an input in the withdrawal classifier, it may capture related information indirectly through deal characteristics, firm descriptions, industry context, market value, and financing structure. For example, prior work suggests that environmental performance can affect information asymmetry and payment-method choices in M&A deals (Hussaini et al., 2023). These channels may help explain why climate exposure is associated with a larger positive HAIDiff. Given that HAIDiff is constructed from a transformed market signal, this interpretation isolates structural model–market disagreement, leaving the attribution of investor error outside the scope of this test.

As an additional robustness test, we construct an alternative model–market disagreement measure based on target announcement returns. Specifically, we define TargetCARDiff as MergerLLM’s predicted probability of merger success minus the scaled target cumulative abnormal return over the announcement window $[-1, +1]$. Target CAR provides a useful alternative market-based signal because target shareholders are directly affected by the expected takeover premium and the probability that the announced transaction will close. This alternative measure does not rely on the residualized acquirer-CAR transformation used in the baseline HAIDiff construction. The results are reported in Table 17. The coefficient on CC Exposure remains positive and statistically significant, indicating that the main finding is not driven only by the specific construction of HAIDiff. Instead, climate exposure continues to be associated with larger differences between model-implied completion probabilities and the alternative market-based signal.

Table 17. Determinants of Model–Market Disagreement: Target CAR Measure. This table reports determinants of model–market disagreement using the target-CAR-based measure. The dependent variable is TargetCARDiff, defined as machine predicted merger success minus scaled target CAR over the announcement window $[-1, +1]$. CC Exposure is climate change exposure. Variable definitions are provided in Appendix A. Continuous variables, except state-level variables, are winsorized at their 1st and 99th percentiles. Standard errors are clustered at the year level (t-statistics are in parentheses). ** and *** denote statistical significance at the 5% and 1% levels, respectively.

Variables	(1) TargetCARDiff	(2) TargetCARDiff	(3) TargetCARDiff	(4) TargetCARDiff
CC Exposure	14.894 *** (4.10)	11.389 *** (2.81)	14.599 *** (3.89)	11.212 *** (2.75)
log(deal value)	0.018 ** (2.49)	0.016 ** (2.25)	0.018 ** (2.56)	0.016 ** (2.31)

Table 17. Cont.

Variables	(1) TargetCARDiff	(2) TargetCARDiff	(3) TargetCARDiff	(4) TargetCARDiff
con			−0.733 (−0.88)	−0.713 (−0.90)
ipg			0.291 (0.30)	0.177 (0.19)
tfp			0.871 (1.20)	0.872 (1.22)
term			1.950 (0.60)	2.128 (0.70)
def			−12.874 ** (−2.27)	−15.218 ** (−2.52)
dei			−8.698 (−0.63)	−9.883 (−0.76)
mkt			0.257 (1.58)	0.244 (1.57)
lab			−0.076 (−0.18)	−0.179 (−0.40)
sent			−0.000 (−0.01)	0.002 (0.07)
Observations	3399	3399	3399	3399
Adjusted R-squared	0.173	0.216	0.175	0.218
Industry FE	N	Y	N	Y
Year FE	Y	Y	Y	Y
Controls	Y	Y	Y	Y

Arbitrage spreads, deal spreads, and option-implied signals would also be useful proxies for market-implied completion probabilities, but these measures are not consistently available for the full sample used in our analysis. We, therefore, use target CAR as the main feasible alternative market signal and interpret both HAIDiff and TargetCARDiff as proxies for model–market disagreement.

5.3. Model–Market Disagreement Around Hurricane Ian

Hurricane Ian provides a setting in which physical climate risk became more salient to investors and firms. We use this event to examine whether model–market disagreement changed differently for firms with higher pre-existing climate exposure. The treatment variable is based on firm-level climate exposure measured before the event, so the analysis captures differences across firms with different climate-risk profiles around a period of heightened climate-risk attention.

We define the treatment group using firm-level climate exposure measured before Hurricane Ian. Specifically, for each acquirer, we compute pre-Ian climate exposure as the average value of firm-level climate change exposure, $CCExposure_i^{pre}$, using observations before 28 September 2022. We then define an indicator $HighCC_i$ equal to one if the acquirer’s pre-Ian climate exposure is above the median of the pre-Ian distribution, and zero otherwise. The control group consists of acquirers with pre-Ian climate exposure at or below the median. The post period, $PostIan_t$, equals one for deals announced on or after 28 September 2022 and zero otherwise. Thus, treatment assignment is based on pre-shock firm-level climate exposure instead of post-shock outcomes or geographic hurricane exposure.

We estimate the following binary-treatment DiD specification:

$$HAIDiff_{i,t} = \beta HighCC_i \times PostIan_t + \gamma X_{i,t} + \alpha_j + \alpha_t + \varepsilon_{i,t}, \quad (6)$$

where $HAIDiff_{i,t}$ is the difference between the MergerLLM-predicted probability of deal success and the transformed market-based signal. $X_{i,t}$ includes deal-level controls, macroeconomic controls, and log deal value. α_j denotes acquirer-industry fixed effects, and α_t denotes year fixed effects. Standard errors are clustered by year. We also estimate a continuous-exposure specification:

$$HAIDiff_{i,t} = \beta CCEXposure_i^{Pre} \times PostIan_t + \gamma X_{i,t} + \alpha_j + \alpha_t + \varepsilon_{i,t}. \quad (7)$$

Table 18 reports the results. In the binary DiD specification, the coefficient on $HighCC_i \times PostIan_t$ is positive and statistically significant. The estimate ranges from 0.036 to 0.037, indicating that HAIDiff increases by 3.6 to 3.7 percentage points more for high pre-Ian climate-exposure firms than for low pre-Ian climate-exposure firms. The continuous-exposure specification gives similar evidence. The coefficient on $CCEXposure_i^{Pre} \times PostIan_t$ is positive and statistically significant. Because the climate-exposure variable is small in raw units, the economic magnitude is better interpreted using a one-standard-deviation change in pre-Ian climate exposure. Based on the sample standard deviation of pre-Ian climate exposure, the estimate in column (4) implies that a one-standard-deviation increase in pre-Ian climate exposure is associated with an approximately 0.5 percentage-point larger post-Ian increase in HAIDiff.

To examine the parallel-trends assumption, we estimate an event-study specification around Hurricane Ian. Figure 7 plots the coefficients on interactions between the high-climate-exposure indicator and event-time indicators, with the year before Hurricane Ian as the omitted benchmark. Table 19 reports the corresponding coefficients. The coefficients at event times -5 , -4 , and -3 are statistically insignificant, suggesting no clear differential trend in the earlier pre-period. However, the coefficient at event time -2 is positive and statistically significant. This result indicates that the parallel-trends assumption is not fully supported. We, therefore, interpret the Hurricane Ian analysis as event-based evidence on climate-risk salience instead of a standalone identification strategy. The pattern is consistent with a larger increase in model–market disagreement among firms with higher pre-existing climate exposure.

Table 18. Hurricane Ian, Pre-Ian Climate Exposure, and Model–Market Disagreement. This table reports difference-in-differences regressions of HAIDiff on Hurricane Ian treatment variables. $HighCC$ equals one if the acquirer’s pre-Ian climate exposure is above the median of the pre-Ian distribution and zero otherwise. $PostIan$ equals one for deals announced on or after 28 September 2022. $CCEXposure^{Pre}$ is firm-level climate exposure measured before Hurricane Ian. Macroeconomic controls include consumption growth, industrial production growth, total factor productivity, term spread, default premium, expected inflation, market return, labor income growth, and investor sentiment. Deal controls include available deal characteristics, such as challenged deal, hostile deal, rumor, cash payment, stock payment, and the percentage of shares sought. Standard errors clustered by year are reported in parentheses. *** indicate significance at the 1% level.

Variables	Binary Treatment		Continuous Exposure	
	(1)	(2)	(3)	(4)
$HighCC \times PostIan$	0.0362 *** (0.0121)	0.0374 *** (0.0122)		
$CCEXposure^{Pre} \times PostIan$			3.4477 *** (1.2219)	2.9106 *** (0.9759)
Observations	4375	4375	4375	4375
Log deal value	Yes	Yes	Yes	Yes
Macroeconomic controls	Yes	Yes	Yes	Yes
Deal controls	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Industry fixed effects	No	Yes	No	Yes

Table 19. Event-study coefficients around Hurricane Ian. This table reports coefficients from the event-study specification. The omitted benchmark period is event time -1 , the year before Hurricane Ian. The dependent variable is HAIDiff. The regression includes log deal value, macroeconomic controls, deal controls, year fixed effects, and acquirer-industry fixed effects. Standard errors are clustered by year. The insignificant coefficients at event times -5 , -4 , and -3 suggest no clear differential trend in the earlier pre-period. The significant coefficient at event time -2 indicates that the parallel-trends assumption is not fully supported. *** indicate significance at the 1% level.

Event Time	Coefficient	Std. Error	t-Statistic	p-Value
-5	0.0039	0.0032	1.248	0.212
-4	-0.0041	0.0029	-1.395	0.163
-3	-0.0051	0.0045	-1.141	0.254
-2	0.0238 ***	0.0031	7.695	<0.001
0	-0.0176 ***	0.0037	-4.812	<0.001
+1	0.0490 ***	0.0049	9.957	<0.001

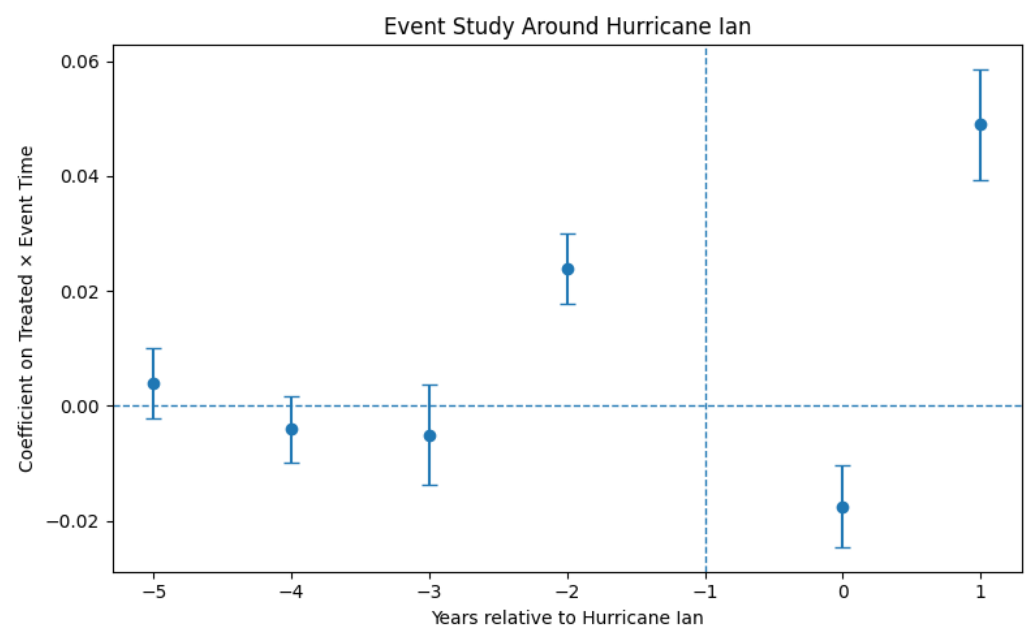


Figure 7. Event study around Hurricane Ian. This figure plots coefficients from an event-study specification that interacts the high pre-Ian climate-exposure indicator with event-time indicators around Hurricane Ian. The event time is measured in years relative to 2022. The omitted benchmark period is event time -1 , the year before Hurricane Ian. The vertical bars represent 95% confidence intervals based on standard errors clustered by year. The horizontal dashed line marks zero, indicating no differential change in HAIDiff relative to the omitted benchmark period. The vertical dashed line marks the omitted benchmark period at event time -1 .

We further examine whether treated and control firms differ before Hurricane Ian. Table 20 reports balance tests using pre-Ian observations. The results show that treated and control firms differ along several observable dimensions, including deal size, macroeconomic conditions, sentiment, hostility, and rumor status. These differences indicate that high- and low-climate-exposure firms are not identical along observable pre-shock characteristics, motivating the matched and weighted robustness tests. We complement the baseline specification with two robustness tests. First, we use propensity-score matching based on pre-shock firm, macroeconomic, industry, and deal characteristics. Second, we estimate an inverse-probability-weighted DiD model using ATT weights with propensity scores trimmed at the 5th and 95th percentiles. Table 21 shows that the coefficient on $HighCC_i \times PostIan_t$ remains positive and statistically significant in both tests, although the magnitude is smaller than in the baseline specification. These robustness tests show

that the positive DiD estimate remains after adjusting for observable differences between high- and low-climate-exposure firms through matching and reweighting. Together, the baseline DiD, event-study, balance, matching, and weighting results show a consistent positive association between climate-risk salience and model–market disagreement among firms with higher pre-existing climate exposure.

Table 20. Pre-Ian balance between high- and low-climate-exposure firms. This table compares treated and control firms using pre-Ian observations. Treated firms are acquirers with pre-Ian climate exposure above the median. Control firms are acquirers with pre-Ian climate exposure at or below the median. The standardized mean difference is the difference in group means divided by the pooled standard deviation. The imbalance in climate exposure is mechanical because climate exposure defines the treatment status. The table shows that treated and control firms also differ along several observable dimensions, motivating the matched and weighted robustness tests reported in Table 21.

Variable	Treated Mean	Control Mean	Std. Mean Diff.	t-Statistic
Log deal value	4.5918	3.4147	0.6012	48.038
Climate exposure	0.0015	0.0001	0.8191	66.312
Consumption growth	0.0016	0.0017	−0.0193	−1.785
Industrial production growth	0.0010	0.0024	−0.1683	−16.089
Total factor productivity	0.0023	0.0029	−0.0659	−6.364
Term spread	0.0019	0.0004	0.3387	31.354
Default premium	0.0000	0.0000	0.0035	0.351
Expected inflation	0.0000	−0.0000	0.0255	2.669
Market return	0.0056	0.0060	−0.0098	−1.058
Labor income growth	0.0031	0.0037	−0.0893	−8.557
Investor sentiment	0.0357	0.3368	−0.4560	−49.031
Hostile deal	0.0014	0.0045	−0.0576	−7.490
Rumor	0.0182	0.0087	0.0820	7.864

Table 21. Matched and weighted DiD robustness tests. This table reports robustness tests for the binary Hurricane Ian DiD specification. Column (1) reports the baseline specification with year and acquirer-industry fixed effects. Column (2) reports a propensity-score matched DiD specification. Matching is based on pre-shock firm, macroeconomic, industry, and deal characteristics. Column (3) reports an inverse-probability-weighted DiD specification using ATT weights, with propensity scores trimmed at the 5th and 95th percentiles. The dependent variable is HAIDiff. Standard errors clustered by year are reported in parentheses. *** and ** indicate significance at the 1% and 5% levels, respectively.

Variables	Baseline DiD (1)	Matched DiD (2)	IPW DiD (3)
<i>HighCC × PostIan</i>	0.0374 *** (0.0122)	0.0136 *** (0.0037)	0.0155 ** (0.0061)
Observations	4375	1424	4375
Log deal value	Yes	Yes	Yes
Macroeconomic controls	Yes	Yes	Yes
Deal controls	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Industry fixed effects	Yes	Yes	Yes

6. Discussion

The results reveal two levels of interpretation. The prediction results provide the strongest evidence in the paper. MergerLLM improves the detection of withdrawn deals relative to several benchmark models, especially in recall and F1-score. This finding is important because merger withdrawals are rare but economically meaningful events.

The model–market disagreement results provide a different type of evidence. HAIDiff compares a model-implied completion probability with a transformed announcement–return-based signal. The climate-exposure results show that this disagreement measure is

larger for firms with greater climate exposure. This pattern suggests that climate-exposed deals are settings in which model-based and market-based assessments differ more strongly.

The interpretation of HAIDiff depends on the construction of the market-based signal. Announcement returns are useful because they are observed around the deal announcement and reflect market reactions to the transaction. However, they also capture expected synergies, financing terms, overpayment concerns, liquidity, arbitrage constraints, and risk premia. For this reason, HAIDiff can be read as a proxy for disagreement between model-implied completion probabilities and announcement-return-based signals.

The Hurricane Ian analysis supports this interpretation from an event-based perspective. Around a period of heightened physical climate-risk salience, firms with higher pre-existing climate exposure experience larger increases in model–market disagreement. Because climate exposure is not randomly assigned and because treated and control firms may differ along other dimensions, the event results should be interpreted as suggestive evidence on climate-risk salience. In addition, the event-study results do not fully rule out differential pre-Ian dynamics. Matching and inverse-probability weighting reduce observable differences between high- and low-climate-exposure firms, but they do not eliminate all identification concerns.

7. Conclusions

In this paper, we propose MergerLLM, an LLM-based framework for predicting merger withdrawals from serialized deal characteristics and text-based firm information. The framework converts each M&A record into a natural-language representation and fine-tunes a language model for the binary classification of completed versus withdrawn deals. The empirical results show that MergerLLM improves the detection of the rare withdrawal class relative to several machine-learning and deep-learning benchmarks, especially in recall and F1-score, while maintaining competitive ROC AUC and PR AUC. At the same time, merger withdrawals remain difficult to predict: the model still misses a meaningful share of withdrawn deals, and its raw predicted probabilities should be interpreted cautiously unless further calibration is applied.

We also examine whether MergerLLM’s predicted merger-success probabilities contain economic information. Sorting deals by predicted success probabilities produces positive return spreads between high- and low-predicted-success acquirers. Calendar-time portfolio tests show positive returns and some positive risk-adjusted alphas, although the strength of the evidence varies across factor specifications and subperiods. Because the portfolio tests do not fully account for transaction costs, short-selling constraints, liquidity, bid–ask spreads, deal duration, delisting, and other implementation frictions, we interpret these results as evidence that MergerLLM predictions contain economic information, not as definitive evidence of implementable merger-arbitrage profitability.

The climate-risk analysis extends the prediction exercise by studying when model-implied completion probabilities and market-based signals diverge. We find that firm-level climate exposure is positively associated with HAIDiff, and similar results appear when we use a target-CAR-based alternative disagreement measure. Around Hurricane Ian, model–market disagreement increases more for firms with higher pre-existing climate exposure. These findings suggest that climate-exposed deals are associated with larger gaps between model-based and market-based assessments around merger announcements.

Overall, the paper’s main contribution is predictive. It shows that serialized tabular and textual M&A data can improve rare-event merger-withdrawal prediction. The model–market disagreement analysis provides an economic application of these predictions, and the climate-exposure results identify one setting in which model-based and market-based assessments differ more strongly.

Author Contributions: Conceptualization, Y.L. and M.S.; methodology, M.S., Z.Y. and Y.L.; software, M.S. and Z.Y.; validation, M.S., Z.Y. and Y.L.; formal analysis, Z.Y. and M.S.; investigation, M.S., Z.Y. and Y.L.; resources, Y.L.; data curation, Z.Y. and Y.L.; writing—original draft preparation, M.S. and Z.Y.; writing—review and editing, M.S., Z.Y., Y.L. and D.Y.; visualization, Z.Y. and M.S.; supervision, Y.L. and D.Y.; project administration, Y.L. and D.Y.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from third-party data providers and are available from the authors upon reasonable request and with the permission of the relevant third-party data providers.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Key variable definitions.

Variable	Description
acqCAR3	Acquirers' cumulative abnormal returns over three days.
DEAL VALUE	The total volume (i.e., transaction value) of the M&A deal in \$ billion.
Whiteknight	Indicator equal to one if the acquirer is a white knight.
RUMOR	Rumor Flag: "Y" indicates that the transaction is currently or originally began as a rumor.
Multibidder	Multibidder flag: 1 if the number of bidders is more than 1; 0 otherwise.
Hostile	Hostile deal flag; "Hostile" means the board officially rejects the offer but the acquirer persists with the takeover.
TPUBLIC	Public status code of the target firm at the time of the announcement. This variable identifies whether the target is publicly listed, privately held, a subsidiary, or has another ownership status.
AMV	Acquirer Market Value: Calculated by multiplying the total number of acquirer shares outstanding times the acquirer's stock price 4 weeks prior to announcement date (\$mil).
CHA	Challenging Bid Flag: Yes/No flag set to 'Y' where a third party launched an offer for the target while this original bid was pending.
TBUSINESSL	Target business description.
ABUSINESSL	Acquirer business description.
PSOUGHT	Percentage of common, or common equivalent, shares outstanding sought by the acquirer in this transaction.
CONSID	Consideration Description: Text description of consideration offered and sought.
con	The monthly growth of real personal consumption expenditures on nondurable goods and services per capita.
tfp	The quarterly percentage change in total factor productivity.
ipg	The monthly growth rate of industrial production.
term	the yield spread between the 20-year and 1-year Treasury bonds.
def	the default premium measured as the monthly change in the yield spread between the BAA-rated and AAA-rated corporate bonds.
dei	The change in monthly expected inflation estimated following Fama and Gibbons.
sent	Baker and Wurgler (BW) sentiment measure on various market anomalies
mkt	the value-weighted excess returns of the stock market.
lab	The log growth rate in nominal labor income per capita.
HAIDiff	MergerLLM's predicted probability of merger success minus the scaled market-based signal derived from residualized acquirer announcement returns.
TargetCARDiff	MergerLLM's predicted probability of merger success minus the scaled target cumulative abnormal return over the announcement window $[-1, +1]$.
CC Exposure	Firm-level climate change exposure based on earnings-call discussions of climate-related opportunities, physical risks, and regulatory shocks.
HighCC	Indicator equal to one if the acquirer's pre-Ian climate exposure is above the median of the pre-Ian distribution.
PostIan	Indicator equal to one for deals announced on or after 28 September 2022.

Table A2. Descriptive statistics of variables in model–market disagreement analysis.

Variable	Mean	SD	p50	p25	p75
HAIDiff	0.79617	0.16966	0.83608	0.83066	0.83916
CC Exposure	0.00081	0.00195	0.00029	0.00012	0.00069
con	0.00158	0.00760	0.00141	−0.00040	0.00300
ipg	0.00103	0.00909	0.00139	−0.00235	0.00530
tfp	0.00209	0.00936	0.00177	−0.00224	0.00581
term	0.00123	0.00524	0.00000	−0.00103	0.00185
def	0.00002	0.00110	−0.00000	−0.00043	0.00040
dei	0.00123	0.00373	0.00073	0.00011	0.00174
mkt	0.00577	0.04245	0.01084	−0.01847	0.03100
lab	0.00304	0.00762	0.00302	0.00124	0.00434
sent	−0.00736	0.59262	−0.15562	−0.33513	0.14945

References

- Addoum, J. M., Ng, D. T., & Ortiz-Bobea, A. (2020). Temperature shocks and establishment sales. *The Review of Financial Studies*, 33(3), 1331–1366. [\[CrossRef\]](#)
- Andrade, G., Mitchell, M., & Stafford, E. (2001). New evidence and perspectives on mergers. *Journal of Economic Perspectives*, 15(2), 103–120. [\[CrossRef\]](#)
- Ang, J. S., Hsu, C., Tang, D., & Wu, C. (2021). The role of social media in corporate governance. *The Accounting Review*, 96(2), 1–32.
- Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. [\[CrossRef\]](#)
- Arrow, K. J. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, 12(2), 91–100. [\[CrossRef\]](#)
- Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4), 1645–1680. [\[CrossRef\]](#)
- Bates, T. W., & Lemmon, M. L. (2003). Breaking up is hard to do? An analysis of termination fee provisions and merger outcomes. *Journal of Financial Economics*, 69(3), 469–504. [\[CrossRef\]](#)
- Becher, D. A., Cohn, J. B., & Juergens, A. L. (2015). Do stock analysts influence merger completion? An examination of postmerger announcement recommendations. *Management Science*, 61(10), 2430–2448. [\[CrossRef\]](#)
- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6), 1315–1344. [\[CrossRef\]](#)
- Bernstein, A., Gustafson, M. T., & Lewis, R. (2019). Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics*, 134(2), 253–272. [\[CrossRef\]](#)
- Betton, S., Eckbo, B. E., Thompson, R., & Thorburn, K. S. (2014). Merger negotiations with stock market feedback. *The Journal of Finance*, 69(4), 1705–1745. [\[CrossRef\]](#)
- Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, 142(2), 517–549. [\[CrossRef\]](#)
- Bolton, P., & Kacperczyk, M. (2023). Global pricing of carbon-transition risk. *The Journal of Finance*, 78(6), 3677–3754.
- Boone, A. L., & Mulherin, J. H. (2007). Do termination provisions truncate the takeover bidding process? *The Review of Financial Studies*, 20(2), 461–489.
- Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72, 218–239. [\[CrossRef\]](#)
- Cao, S., Jiang, W., Wang, J., & Yang, B. (2024). From man vs. machine to man+ machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160, 103910. [\[CrossRef\]](#)
- Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5), 124–127. [\[CrossRef\]](#)
- Chava, S. (2014). Environmental externalities and cost of capital. *Management Science*, 60(9), 2223–2247. [\[CrossRef\]](#)
- Chen, Z., Lu, A., & Zhu, X. (2024). Investor sentiment and the pricing of macro risks for hedge funds. *Management Science*, 71, 1623–1645.
- Cho, S., Hong, H.-S., & Ha, B.-C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the mahalanobis distance: For bankruptcy prediction. *Decision Support Systems*, 52(2), 3482–3488. [\[CrossRef\]](#)
- Choi, D., Gao, Z., & Jiang, W. (2020). Attention to global warming. *The Review of Financial Studies*, 33(3), 1112–1145. [\[CrossRef\]](#)
- Cookson, J. A., Niessner, M., & Schiller, C. (2026). Can social media inform corporate decisions? Evidence from merger withdrawals. *Journal of Finance*, 81, 91–142.

- Cui, G., Hu, S., Ding, N., Huang, L., & Liu, Z. (2022). Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7014–7024). Association for Computational Linguistics. [CrossRef]
- Dasgupta, S., Harford, J., Ma, F., Wang, D., & Xie, H. (2024). *Strategic communication in M&A conference calls: Topic modelling analysis of transcripts*. Available at SSRN 3528016. Available online: <https://ssrn.com/abstract=3528016> (accessed on 5 June 2026).
- Delis, M. D., de Greiff, K., Iosifidi, M., & Ongena, S. (2024). Being stranded with fossil fuel reserves? Climate policy risk and the pricing of bank loans. *Financial Markets, Institutions & Instruments*, 33(3), 239–265. [CrossRef]
- Deng, X., Kang, J.-K., & Low, B. S. (2013). Corporate social responsibility and stakeholder value maximization: Evidence from mergers. *Journal of Financial Economics*, 110(1), 87–109. [CrossRef]
- Dinh, T., Zeng, Y., Zhang, R., Lin, Z., Gira, M., Rajput, S., Yong Sohn, J., Papailiopoulos, D., & Lee, K. (2022). Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35, 11763–11784.
- Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. *The Review of Financial Studies*, 33(3), 1184–1216. [CrossRef]
- Fama, E. F. (1970). Efficient capital markets. *Journal of Finance*, 25(2), 383–417. [CrossRef]
- Garbarino, N., & Guin, B. (2021). High water, no marks? Biased lending after extreme weather. *Journal of Financial Stability*, 54, 100874. [CrossRef]
- Ge, S., & Lam, A. (2022). Does the threat of insurer liability for “Silent” cyber and climate risks affect insurers’ pricing? *The Geneva Papers on Risk and Insurance—Issues and Practice*, 47(1), 2–40.
- Gomes, M. (2019). Does CSR influence M&A target choices? *Finance Research Letters*, 30, 153–159. [CrossRef]
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., . . . Jia, J. (2024). The Llama 3 herd of models. *arXiv*, arXiv:2407.21783.
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273. [CrossRef]
- Harford, J. (2005). What drives merger waves? *Journal of Financial Economics*, 77(3), 529–560. [CrossRef]
- Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., & Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics* (pp. 5549–5581). PMLR.
- Hong, H., Li, F. W., & Xu, J. (2019). Climate risks and market efficiency. *Journal of Econometrics*, 208(1), 265–281. [CrossRef]
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *Iclr*, 1(2), 3.
- Hussaini, M., Rigoni, U., & Perego, P. (2023). The strategic choice of payment method in takeovers: The role of environmental, social and governance performance. *Business Strategy and the Environment*, 32(1), 200–219.
- Ilhan, E., Sautner, Z., & Vilkov, G. (2021). Carbon tail risk. *The Review of Financial Studies*, 34(3), 1540–1571.
- Javadi, S., & Masum, A.-A. (2021). The impact of climate change on the cost of bank loans. *Journal of Corporate Finance*, 69, 102019. [CrossRef]
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491–495. [CrossRef] [PubMed]
- Kousky, C. (2019). The role of natural disaster insurance in recovery and risk reduction. *Annual Review of Resource Economics*, 11, 399–418. [CrossRef]
- Krueger, P., Sautner, Z., & Starks, L. T. (2020). The importance of climate risks for institutional investors. *The Review of Financial Studies*, 33(3), 1067–1111. [CrossRef]
- Liu, B., & McConnell, J. J. (2013). The role of the media in corporate governance: Do the media influence managers’ capital allocation decisions? *Journal of Financial Economics*, 110(1), 1–17. [CrossRef]
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations (iclr)*. ICLR.
- Lu, J. (2022). Green mergers and acquisitions and innovation: Evidence from China. *Energy Economics*, 114, 106305.
- Luo, Y. (2005). Do insiders learn from outsiders? Evidence from mergers and acquisitions. *The Journal of Finance*, 60(4), 1951–1982. [CrossRef]
- Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research*, 56(6), 960–980. [CrossRef]
- Nguyen, Q. A., Tomeh, N., Lebbah, M., Charnois, T., Azzag, H., & Muñoz, S. C. (2024). Manual verbalizer enrichment for few-shot text classification. *arXiv*, arXiv:2410.06173. Available online: <https://arxiv.org/abs/2410.06173> (accessed on 5 June 2026).
- Officer, M. S. (2003). Termination fees in mergers and acquisitions. *Journal of Financial Economics*, 69(3), 431–467. [CrossRef]

- Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, 52(2), 464–473. [\[CrossRef\]](#)
- Pasiouras, F., Gaganis, C., & Zopounidis, C. (2008). Regulations, supervision approaches and acquisition likelihood in the asian banking industry. *Asia-Pacific Financial Markets*, 15, 135–154. [\[CrossRef\]](#)
- Renneboog, L., & Vansteenkiste, C. (2019). Failure and success in mergers and acquisitions. *Journal of Corporate Finance*, 58, 650–699. [\[CrossRef\]](#)
- Sautner, Z., Van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-level climate change exposure. *The Journal of Finance*, 78(3), 1449–1498.
- Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 255–269). Association for Computational Linguistics. [\[CrossRef\]](#)
- Seltzer, L. H., Starks, L. T., & Zhu, Q. (2022). *Climate regulatory risk and corporate bonds*. NBER Working Paper No. 29994. National Bureau of Economic Research.
- Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv*, arXiv:2106.01342.
- Tunyi, A. A. (2021). Revisiting acquirer returns: Evidence from unanticipated deals. *Journal of Corporate Finance*, 66, 101789. [\[CrossRef\]](#)
- Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., & Poli, I. (2025). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 2526–2547). Association for Computational Linguistics. [\[CrossRef\]](#)
- Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107–126. [\[CrossRef\]](#)
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv*, arXiv:2303.17564.
- Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1), 607–610. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.