Predicting from Heterogeneous Texts: A Source-preserving Prompt-based Approach

Zongxiao Wu[a], Yizhe Dong[a,*], Yaoyiran Li[b,c], Baofeng Shi[d]

[a] Business School, University of Edinburgh, Edinburgh, EH8 9JS, United Kingdom

[b] Google, London, N1C 4AG United Kingdom

[c] Language Technology Lab, University of Cambridge, Cambridge, CB3 9DA, United Kingdom

[d] College of Economics and Management, Northwest A&F University, Yangling, 712100, China

Abstract

Textual data used in decision-making is often produced by multiple agents with distinct informational roles and incentives. Existing approaches typically either analyze each text source in isolation or combine heterogeneous texts into shared representations that obscure source-specific meaning and reduce interpretability. We propose a source-preserving, prompt-based framework that converts each text source independently into an interpretable propensity signal using cloze-style masked language modelling. We apply the framework to a unique dataset of micro and small enterprise loans containing 15 heterogeneous text variables. Across off-the-shelf and fine-tuned settings, our approach consistently outperforms structured variable-only models and established natural language processing benchmarks. We find that loan officer narratives contain greater predictive information than borrower self-reports, with direct implications for how lenders collect and weight soft information. Profit analysis shows that the predictive gains translate into economically meaningful improvement in lending decisions. These findings contribute to the literature on text-based financial decision-making by showing how heterogeneous narratives can be integrated without collapsing their source-specific information, and by providing an interpretable design for high-stakes predictive analytics.

Keywords: Natural language processing; Language model; Prompt approach; Credit risk; Text mining; Deep learning; mSE loans

* Corresponding Author.
Email addresses: Zongxiao.Wu@ed.ac.uk, Yizhe.Dong@ed.ac.uk, yl711@cam.ac.uk, shibaofeng@nwsuaf.edu.cn

The authors would like to thank Yu An, Martin Becker, Donald Bowen, Cristián Bravo, Winston Wei Dou, Yongmiao Hong, Zhiyong Li, Alejandro Lopez Lira, Belen Martin-Barragan, Sovan Mitra, Hongwei Mo, María Óskarsdóttir, Sotirios Sabanis, Feng Shen, Zhao Wang, Echo Yang, Yi Yang, Zhipeng Zhang, Huimin Zhao, Fanyin Zhou, and conference and seminar participants at 2025 China International Conference in Finance, 2025 Finance and Accounting Annual Research Symposium, 18th International Behavioral Finance Conference, Credit Scoring and Credit Control Conference XIX, Credit Scoring and Credit Rating Conference III, Heriot-Watt University, Shanghai Jiao Tong University, University of Chinese Academy of Sciences, and Xiamen University, for their helpful feedbacks.

1. Introduction

Textual information has become increasingly important in decision-making across finance, banking, operations, and information systems. A growing body of research shows that unstructured text can be used to construct economically meaningful signals that improve prediction and reveal latent dimensions of firm behavior and risk, including innovation, misconduct risk, cyber risk, supply chain exposure, and fund flows (e.g., Bellstam et al., 2021, Campbell & Shang, 2022, Hillert et al., 2025, Jiang et al., 2024). At the same time, advances in machine learning and natural language processing (NLP) have enabled researchers to extract richer information from text. Recent studies move beyond dictionary-based methods toward neural NLP models and large language models (LLMs), which are better able to capture contextual meaning and have improved the measurement of textual information in economic and finance settings (e.g., Cong et al., 2025, Loughran & McDonald, 2011, Siano, 2025).

Despite this progress, most existing studies either focus on a single type of document or implicitly treat textual information as if it came from a homogeneous source. In contrast, many real-world decision settings involve heterogeneous textual data produced by multiple agents with different informational roles, incentives, and perspectives. In credit markets, borrower self-descriptions, loan officer assessments, and platform-generated narratives coexist and jointly inform lending decisions. In equity analysis, analysts read not only the management commentary but also the audit opinion, footnotes, and earnings call exchanges between executives and questioners, each reflecting a different informational position. In fraud detection, compliance officers draw simultaneously on transaction narratives, counterparty disclosures, and internal risk assessments. While prior work has developed sophisticated tools for extracting information from individual text types, it remains an open question how to integrate such heterogeneous narratives into a unified predictive framework without distorting their source-specific informational content.

Two challenges arise when heterogeneous multi-perspective texts are integrated for predictive modelling, and neither has been satisfactorily resolved in the existing literature. The first is what we call fusion ambiguity. Heterogeneous texts differ not only in topic but also in stance, tone, incentive structure, and factual emphasis. When texts with these structural differences are combined through concatenation or embedded into a shared latent space, their signals may interact in complex and non-transparent ways, and generate semantic noise that can obscure the underlying predictive information (Wang et al., 2023; Zhang et al., 2016). Wang et al. (2021) demonstrate in a study of Chinese corporate risk prediction that multi-source soft information integration can significantly improve predictive performance; however, they rely on a pooling-based architecture that collapses source-specific contributions into a shared representation, making it difficult to identify which source drives the model's predictions.

The second challenge is interpretability. When multiple text sources are processed using neural NLP models, such as fine-tuned bidirectional encoder representations from transformers (BERT),

attention pooling, or other dense representations methods, the final output is typically a function of high-dimensional latent representations. Although these representations can be predictive, their individual dimensions usually do not have direct interpretable semantic meaning (Wu et al., 2025). This creates a gap between predictive performance and transparency. Post-hoc explanation tools such as local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) attempt to address this limitation by approximating model behavior after estimation (Siano, 2025; Stevenson et al., 2021; Vaca et al., 2024). They approximate the model's attribution after the fact by perturbing inputs and measuring output changes. However, as Lipton (2018) argues, post-hoc explanation approaches reconstruct rather than reveal the model's reasoning. In high-stakes financial decisions such as credit approval, fraud detection, and regulatory assessment, this limitation carries real consequences. Decision-makers require predictions that can be traced to interpretable inputs to ensure accountability, regulatory compliance, and effective model governance. Rudin (2019) argues that, in such settings, interpretability should be built into the model by design rather than imposed ex post. However, achieving source-level interpretability in the presence of heterogeneous textual inputs remains unmet in the financial NLP literature.

In this study, we propose a source-preserving, prompt-based framework for predictive analytics with heterogeneous multi-perspective text to fill these gaps in the literature. Rather than encoding all text sources into a shared latent representation, we process each text source independently through its own cloze-style prompt. The prompt instructs a pre-trained masked language model (MLM) to complete an outcome-relevant sentence, and the probability assigned to outcome-relevant label words at the [MASK] position is directly transformed into a propensity score for that text source. In this way, the propensity score makes source attribution a structural property of the representation itself, thereby satisfying the interpretability-by-design objective that current multi-source NLP approaches have yet to achieve (Lipton, 2018; Rudin, 2019). The framework can operate in two modes: an off-the-shelf mode that applies pre-trained MLMs without adaptation and is suitable for data-scarce settings, and a fine-tuned mode that retains the same architecture while incorporating supervised optimization through a two-stage training pipeline to deliver stronger predictive performance.

We demonstrate the framework in the context of credit risk prediction for micro and small enterprises (mSEs). This empirical setting is both practically significant and theoretically well-suited to the challenges the framework is designed to address. The mSEs represent a disproportionately large share of employment and economic output yet face persistent exclusion from formal credit markets (Lee et al., 2024). This financing gap is a direct consequence of information asymmetry: the mSE knows the nature of its business and its genuine capacity to repay; the lender, lacking the means to verify these directly, cannot confidently assign risk from structured data alone (Akerlof, 1970; Stiglitz & Weiss, 1981). To partially bridge this gap, lenders generally collect two complementary types of textual narrative during site visits conducted prior to loan approval. The first consists of borrowers' application form or self-reported statements, which cover business goals, loan purposes, household assets, and

repayment intentions. These statements provide subjective but granular accounts written under strong incentives to present the borrower favorably. The second consists of loan officers' independent field assessments, covering borrower credibility, business viability, repayment capacity, and external risk factors. It is a professional evaluation written under accountability to accuracy rather than to approval. These two sources are structurally heterogeneous in the sense the framework is designed to handle. They originate from agents with different informational incentives and differ in tone and epistemic status.

Our empirical analyses draw on a unique dataset of agricultural mSE loans from a major Chinese bank, comprising 15 heterogeneous text variables generated by borrowers and loan officers alongside standard credit features. Across off-the-shelf and fine-tuned settings, four MLM architectures, six classifiers, and three discrimination metrics, the framework consistently and significantly outperforms structured data baselines and established NLP benchmarks, with the largest gains realized under the fine-tuned mode. A series of correlation and feature importance analyses show that textual features provide complementary rather than redundant predictive signals relative to structured variables. We also find that loan officers' assessments show more predictive than borrowers' self-reports, which is consistent with the signaling theory prediction and has direct implications for how lenders collect, standardize, and weight soft information.[1] Profit-based evaluation shows that the framework's predictive gains translate directly into economic improvements in lending decisions. In addition, ablation results demonstrate that both the [MASK]-head prompting design and the two-stage training pipeline contribute independently to performance. Replications across three LLMs (GPT-3.5, GPT-4, and DeepSeek-Chat) and on an alternative dataset further confirm that the results are not driven by a particular model or dataset and help identify the boundary conditions under which prompt-based approaches outperform standard fine-tuning in domain-specific, low-resource financial applications.

This study makes three contributions to the literature on text-based financial decision-making and AI-assisted predictive analytics. First, we develop a predictive analytics framework for heterogeneous multi-perspective text where each source is processed independently through a dedicated cloze-style prompt. This approach produces a scalar propensity score whose attribution to a named source is a structural property of the model rather than a post hoc inference. In this way, the framework delivers interpretability by design at the level of source attribution and addresses a longstanding gap in applying neural NLP methods to high-stakes financial and managerial decision-making (Cong et al., 2025; Lipton, 2018; Rudin, 2019). It also addresses a specific challenge in multi-source text modelling: existing neural NLP approaches typically integrate heterogeneous text through shared latent representations or attention-pooling mechanisms, which blur source boundaries and make it difficult to identify the origin of predictive signals (Wang et al., 2021, 2023; Wu et al., 2025; Yang et al., 2023). By converting each

[1] Signaling theory suggests that third-party assessments from informed, incentive-aligned intermediaries carry greater informational efficiency than self-disclosures susceptible to strategic misrepresentation (Spence, 1973).

source into an economically meaningful propensity score, our framework extends the literature on scalar text-based measures of financial outcomes (Bao & Datta, 2014; Bellstam et al., 2021; Buehlmaier & Whited, 2018; Campbell & Shang, 2022; Loughran & McDonald, 2011). We show that such measures can be derived from prompted pre-trained language models (PLMs), which offer a more semantically grounded alternative to dictionaries, topic models, and bespoke neural architectures. Second, we establish the *pre-train, prompt, and predict* paradigm as a practical and data efficient solution for domain-specific financial analytics in a low-resource setting. We show that the framework achieves superior predictive performance relative to benchmarks and that its architectural design extends naturally from MLMs to LLMs. These findings contribute to the emerging literature on the use of foundation models in domain-specific decision contexts (Cong et al., 2025; De Kok, 2025; Siano, 2025). Third, we provide empirical evidence on the differential predictive contribution of text sources generated by agents with structurally divergent informational incentives in the credit market. The finding that loan officers' assessments outperform borrowers' self-reports across all models offers an empirical test of signaling theory in the mSE lending context and contributes to the broader alternative data literature on the informational value of human intermediary assessments (Campbell & Shang, 2022; Chi et al., 2025; Yang et al., 2023).

The remainder of the paper is organized as follows. Section 2 reviews the related literature on financial textual analysis and heterogeneous text integration. Section 3 provides background on MLM pre-training and Section 4 introduces the proposed prompt-based framework. Section 5 describes the dataset and Section 6 reports the main empirical results. Section 7 validates the framework design and examines its generalizability across LLMs and an alternative loan dataset. Section 8 concludes and discusses implications and limitations.

2. Literature Review

2.1 Text as an Information Source in Financial Decision-making

Text has become an important source of information in financial decision-making because many economically relevant signals are difficult to observe from structured variables alone (Zhou et al., 2024). Financial ratios, balance-sheet items, and transaction records summarize realized outcomes, but they reveal less about managerial explanations, organizational frictions, soft information about borrowers, emerging risks, and market interpretations (Cong et al., 2025). Text helps fill this gap by recording how economic agents describe circumstances, justify actions, communicate expectations, and express uncertainty. This role becomes particularly important when hard information is incomplete, delayed, or noisy. In such settings, textual disclosures and narratives can help investors, lenders, analysts, and regulators infer underlying fundamentals and risks that structured data cannot capture (Loughran and McDonald, 2011, Wang et al., 2026, Wu et al., 2025).

Existing studies use text in financial decision-making in two broad ways. One stream uses text to

construct measures of latent firm characteristics, beliefs, and risks that are difficult to quantify directly (Wu, 2024). Bellstam et al. (2021), for example, develop a text-based measure of corporate innovation from analyst reports and show that it captures innovative activity beyond conventional proxies such as patents and R&D. Campbell & Shang (2022) use employee reviews to construct a forward-looking measure of corporate misconduct risk. They find that employees possess information about firms' operating practices, control environments, and cultural conditions that outside observers cannot easily see. Wu (2024) combines bag-of-word counting with seeded word embeddings to construct a measure of supply chain risk exposure from earnings calls. Hillert et al. (2025) document that the language used in mutual fund shareholder letters predicts subsequent fund flows and managerial behavior. In each case, text is transformed into a measurable construct that proxies for an underlying economic attribute that structured variables capture only incompletely.

The second stream uses text to extract predictive signals for future financial outcomes. Loughran & McDonald (2011) develop a finance-specific dictionary and show that it outperforms general-purpose dictionaries in predicting trading volume, return volatility, fraud, and unexpected earnings. Subsequent studies show that textual content in corporate disclosures (Frankel et al., 2022), analyst reports (Chi et al., 2025), news (Fan et al., 2025), conference call transcripts (De Kok, 2025) and borrower narratives (Stevenson et al., 2021) improves the prediction of market reactions, corporate distress, and credit outcomes. More recent studies show that machine learning models can capture richer predictive signals from financial text without relying entirely on manually specified lexicons (Wu et al., 2025; Zhou et al., 2024). Taken together, these two streams establish that text contains economically relevant information, and that the value of this information depends not only on the source from which it is drawn, but also on how textual meaning is represented and integrated into downstream prediction.

2.2 Existing Approaches to Financial Text Representation

Existing studies have followed two broad methodological routes to process text for financial decision-making (Frankel et al., 2022). One relies on conventional NLP approaches, such as dictionaries, bag-of-words representations, and topic models, that convert text into structured inputs for downstream regression or classification tasks. The other relies on neural NLP methods, including word embeddings, recurrent and convolutional networks, and PLMs such as BERT. The two routes differ in how they represent text, and each carries distinct advantages and limitations.

Conventional NLP methods are transparent, scalable, and relatively easy to implement (Kriebel & Stitz, 2022; Netzer et al., 2019). Researchers can usually trace a result back to identifiable words, phrases, or topics. Jiang et al. (2019) uses latent Dirichlet allocation to extract features from borrowers' descriptions for credit risk prediction. Bao & Datta (2014) apply a similar approach to discover and quantify risk types from textual risk disclosures in 10-K filings. Because these methods represent text through observable components, the resulting features are relatively easy to describe and link to economic constructs (Wang et al., 2020). Their main limitation is that they represent text through word

counts, co-occurrences, or topic prevalence rather than through full linguistic context (Bochkay et al., 2023; Donovan et al., 2021). As a result, they often struggle to capture meaning that depends on surrounding language, local syntax, or the position from which a statement is made (Frankel et al., 2022; Siano, 2025). Campbell & Shang (2022) explicitly separate employee reviews into "pros", "cons", and "advice", arguing that the same word can convey different information depending on the section in which it appears. When "ethical" appears in the pros (cons) of a firm, it is likely to mean the firm has ethical (unethical) practices. This limitation becomes more serious when texts come from heterogeneous sources, where the same word or topic may carry different implications depending on who produces the text and in what context it appears.

Neural NLP approaches address some of these limitations by representing text through dense, context-sensitive embeddings. Compared with conventional methods, they capture semantics, syntax, and domain-specific language use more effectively because they model words in context rather than treating text primarily as a collection of isolated tokens (Fitzpatrick & Mues, 2021). Li et al. (2021) use word embeddings to measure corporate culture from earnings call transcripts. Stevenson et al. (2021) and Wu et al. (2025) fine-tune BERT on loan officers' assessments for small business default prediction. Kim & Nikolaev (2024) apply a BERT-based model to 10-K filings to classify future performance changes. Siano (2025) further shows that BERT-based representations capture the information content of earnings disclosures more effectively than dictionary-based and earlier machine learning methods.

Yet stronger semantic capacity does not by itself resolve two challenges that arise when text is used for high-stakes financial decisions. The first concerns the integration of multiple text sources, which we discuss in detail in Subsection 2.3. The second concerns interpretability. Neural NLP models produce high-dimensional latent representations whose internal dimensions do not have direct semantic meaning. Existing studies typically rely on post-hoc explanation tools, including feature attribution, attention visualization, LIME, and SHAP, to inspect model behavior after inference (Luo et al., 2024; Stevenson et al., 2021; Vaca et al., 2024). Lipton (2018) and Rudin (2019) argue that these post-hoc methods reconstruct rather than reveal a model's reasoning. Therefore, interpretability should be built into the model by design rather than imposed ex post. Recent empirical work has further documented that attention weights can be empirically uncorrelated with gradient-based attribution, raising questions about whether attention-based explanations reliably identify which input drives a prediction (Jain & Wallace, 2019). These concerns are particularly significant in financial settings, where decision-makers need to know not only whether text improves prediction, but also which source contributes to that prediction and how that contribution enters the decision process.

2.3 Integrating Heterogeneous Multi-source Text

A growing body of work in finance and information systems has examined how multi-source textual data can be jointly used in predictive modelling. Wang et al. (2021) develop a self-adaptive method for financial risk prediction that integrates heterogeneous sources, including corporate filings,

news, and social media, and demonstrate significant gains over single-source approaches in a large-scale study of Chinese listed firms. Wang et al. (2023) extend this line of work to fine-grained attention over multi-section financial reports, showing that different sections of the same document carry differentially valuable information for fraud detection. In a related design-science study, Yang et al. (2023) develop a theory-driven deep learning framework that combines voice and text with structured features for financial risk prediction, demonstrating that human-generated signals carry incremental predictive content beyond structured data. Cao (2025) examine the joint use of image and text information through a deep learning approach that integrates the two modalities for prediction, showing that multimodal learning improves performance in marketing decision tasks. These studies suggest that integrating diverse signals improves prediction and that the diversity of perspectives represented in heterogeneous data is a primary source of incremental informational value.

Existing approaches to multi-source text integration share, however, a common architectural limitation. Each combines sources through pooled latent representations, attention-weighted aggregation, or shared embeddings (Augenstein et al., 2019; Izacard & Grave, 2021), in which the contribution of any individual source becomes difficult to identify after the model has been estimated. These designs support flexible information fusion, but they also make it harder to determine whether the model has preserved the distinct informational roles of different sources or simply blended them into a common representation (Yu et al., 2022; Zhang et al., 2016). This issue matters because heterogeneous texts are often produced by actors with different incentives, viewpoints, and levels of credibility. Borrowers describe themselves strategically; managers disclose under legal and reputational constraints; analysts interpret disclosures from a more evaluative standpoint; the media selectively amplifies some aspects of events while downplaying others; and loan officers evaluate borrowers under professional accountability to accuracy.

This creates a gap that existing approaches have not fully resolved. The integration of heterogeneous text typically involves a trade-off: analyzing each source separately preserves source attribution but fails to model the joint information that motivates multi-source designs, whereas pooling sources into a shared representation captures cross-source information but obscures attribution. The interpretability problem and the integration problem therefore reinforce each other. What is needed is a framework that keeps heterogeneous texts distinct while still integrating their predictive content, so that source-level interpretability is built into the model rather than reconstructed after the fact.

3. Background

Among PLMs, MLMs such as BERT are particularly well suited to language understanding tasks, including text classification, natural language inference, and extractive question answering. MLMs are trained to predict masked tokens from surrounding context, an objective that makes them well suited to settings in which the goal is to infer meaning from text rather than to generate free-form language (De

Kok, 2025, Liu et al., 2023, Siano, 2025). We describe the MLM pre-training mechanism here because it provides the foundation for our prompt-based framework introduced in Section 4.

MLMs differ from earlier sequence-based neural models in their use of self-attention, which models relationships across all tokens in an input simultaneously rather than sequentially (He et al., 2023). Each token is represented in light of every other token in the sequence, and the resulting contextual representations capture meaning that depends on linguistic context rather than position alone (Devlin et al., 2019). Formally, let $x = \{w_1, w_2, \dots, w_{|x|}\}$ denote an input text sequence. During pre-training, the sequence is augmented with two special tokens to form $x' = \{w_{[\mathrm{CLS}]}, w_1, w_2, \dots, w_{|x|}, w_{[\mathrm{SEP}]}\}$, where [CLS] and [SEP] represent the start of the sequence and a sentence separator respectively. For example, the model may take an input text x' = "[CLS] These movies are great as they are well made. [SEP]". A randomly selected proportion α of the non-special tokens are then masked, which produces a corrupted sequence $\tilde{x}$. In our example, the model might mask the word "great" or "well". The corrupted sequence is mapped into dense token embeddings $\boldsymbol{e}$ that encode lexical and positional information, and these embeddings, $\{\boldsymbol{e}_{[\mathrm{CLS}]}, \boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_{|x|}, \boldsymbol{e}_{[\mathrm{SEP}]}\}$, are passed through a transformer encoder. After several layers of self-attention and feed-forward transformations, the encoder produces contextualized hidden representations $\{\boldsymbol{h}_{[\mathrm{CLS}]}, \boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_{|x|}, \boldsymbol{h}_{[\mathrm{SEP}]}\}$, where $\boldsymbol{h}_i$ denotes the hidden vector associated with the i-th token position. The MLM with parameters θ is then trained to recover the masked tokens by maximizing

$$\max_{\theta} \log p_{\theta}(x|\tilde{x}) = \max_{\theta} \sum_{i \in C} \log p_{\theta}(\widetilde{w}_i = w_i|\tilde{x}) \quad (1)$$

where C is the index set of the masked tokens in the sequence, $\widetilde{w}_i$ is the model prediction for the i-th token, and w_i is the ground-truth token. By predicting masked words from their full surrounding context rather than from local n-gram patterns, MLMs acquire substantial semantic and syntactic knowledge over the course of pre-training. This knowledge is encoded jointly in the token embeddings and in the parameters of the transformer encoder, and it is the resource that prompting subsequently exploits. When a downstream task is formulated as a cloze-style completion at a masked position, the model can recover task-relevant information through the same prediction mechanism used during pre-training. Figure 1 summarizes the pre-training process.

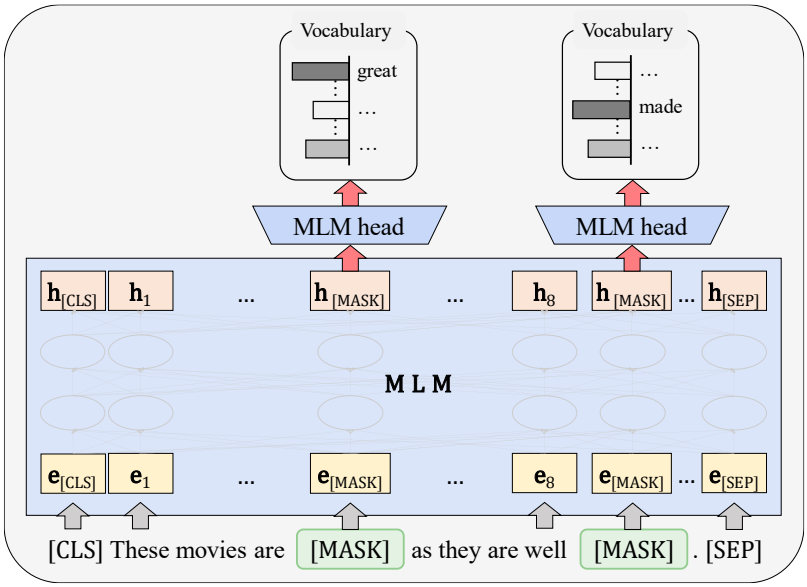


Figure 1 The example of MLM pre-training

4. Prompt-based Framework

4.1 Design Rationale

A common way to use MLMs for downstream prediction is to fine-tune them with a task-specific classification head built on a pooled hidden representation, typically the $\boldsymbol{h}_{[\mathrm{CLS}]}$ vector shown in Figure 1. The pooled representation summarizes the entire input in a single high-dimensional vector. This *pre-train and fine-tune* paradigm is effective in many settings, but it has two limitations that matter for our purposes. First, the dimensions of $\boldsymbol{h}_{[\mathrm{CLS}]}$ have no inherent semantic meaning, which makes the resulting features hard to interpret and compare across heterogeneous text sources. Second, the supervised objective at the [CLS] head is structurally different from the masked-token prediction objective the model was pre-trained to solve, and prior work suggest that this gap can weaken the alignment between pre-trained knowledge and downstream prediction (Han et al., 2022; Sun et al., 2022).

We therefore adopt the *pre-train, prompt, and predict* paradigm (Liu et al., 2023). Rather than training a new classification head on a pooled representation, we reformulate the downstream task as a cloze completion problem. Specifically, the original text is inserted into a task-specific template containing a single [MASK] token, the MLM then scores candidate completions at the masked position, and these scores are converted into a scalar propensity that summarizes the model's directional assessment of the text. The reformulation rests on a specific argument that we refer to as task-pre-training alignment, which shows two advantages over the *pre-train and fine-tune* paradigm.

The first advantage is representational. During pre-training, the MLM learns to infer masked tokens from their surrounding context across large and heterogeneous corpora. A cloze-style prompt places the downstream task in the same form: a sentence with a masked position, where the model estimates the

conditional probability of candidate label words. The downstream task therefore sits within the input distribution for which MLM's pre-trained representations were optimized. It does not require a shift from masked-token prediction to a separate supervised objective, nor does it require learning a new mapping from a pooled high-dimensional representation to an outcome class. It also avoids relying on a classification head to recover from limited supervised data knowledge that is already encoded in the MLM. The second advantage is informational. Because the pre-trained representations are accessed through their native interface rather than through a learned projection of an opaque latent vector, the supervised data are required to do less work. In the *pre-train and fine-tune* paradigm, the supervised signal must teach the classification head both how to extract task-relevant information from a latent space and how to map that information to an outcome class. Under the *pre-train, prompt, and predict* paradigm, by contrast, the prompt provides part of the extraction structure through the cloze format and the verbalizer, while the supervised data are used mainly to refine calibrate or refine the resulting scores. This mechanism helps explain why prompt-based approaches achieve competitive performance when labelled data are limited, as documented by Schick & Schütze (2021) and Liu et al. (2023) across a range of tasks.

In our heterogeneous text setting, the alignment argument has a further implication. Because the prompt is applied to each text source separately, the MLM's linguistic knowledge is queried independently for each source. This avoids forcing heterogeneous sources into a shared representation where source-specific information may become entangled. The framework therefore preserves source independence at the representational level, not through an additional architectural constraint, but as a direct consequence of processing each source through its own cloze-style completion.

4.2 Prompting Templates

We define a task-specific prompt that consists of two components: a template $T(\cdot)$ and a set of label words V. The template maps each input text x into a prompted output, $x_{prompt} = T(x)$, which preserves the original text and appends a cloze-style statement containing a single [MASK] token. As an illustration, consider our empirical task of credit default prediction. Suppose the input text $x =$ "My business is experiencing some liquidity issues". We use the template $T(\cdot) =$ " $\cdot$ So the probability of default for this borrower is [MASK].". This yields the prompted output: x_{prompt} = "My business is experiencing some liquidity issues. So the probability of default for this borrower is [MASK]." The prompted sequence is then passed through the MLM. The model generates a contextualized hidden representation for the [MASK] position, denoted by $\boldsymbol{h}_{[\mathrm{MASK}]}$. For any candidate label word $v \in V$, the MLM assigns a probability that v is the appropriate token to fill the masked position:

$$p([\mathrm{MASK}] = v | x_{prompt}) = \frac{\exp\left(\boldsymbol{e}_v^{\mathrm{T}} \cdot \boldsymbol{h}_{[\mathrm{MASK}]}\right)}{\sum_{\tilde{v} \in V} \exp\left(\boldsymbol{e}_{\tilde{v}}^{\mathrm{T}} \cdot \boldsymbol{h}_{[\mathrm{MASK}]}\right)}, \tag{2}$$

where $\boldsymbol{e}_v$ is the token embedding of v. The prompted task takes the same form the MLM was trained to solve. Rather than learning a new prediction rule from scratch, the model evaluates how its existing

linguistic and contextual representations already encoded during pre-training to ranks task-relevant verbal completions. If the input text contains information associated with increased default risk, the prompt completion "the probability of default ... is [MASK]." should assign greater conditional probability to label words that indicate high risk than to those associated with low risk. Prompting therefore uses MLM as a structured probe of task-relevant knowledge (Liu et al., 2023).

4.3 Mapping Label Words to Task-relevant Signals

To connect masked-token prediction to the downstream task, we define a verbalizer $\phi: Y \to V$ that maps the target space Y to one or more label words in V. This step is important because MLMs predict words, whereas downstream applications require task-relevant signals. The verbalizer provides the semantic bridge between the two. In our empirical setting, $Y = \{0,1\}$, where 1 denotes default and 0 denotes non-default. We associate the default class with the words "high" and "large," and the non-default class with the words "low" and "small."[2] The verbalizer ϕ then links the probability distribution over target values to the probability distribution over label words at the masked position, i.e., $p(y|x) = p([\text{MASK}] = \phi(y)|x_{prompt})$. This design keeps the downstream task aligned with the MLM's pre-training objective. Rather than predicting an abstract target directly, the model evaluates the likelihood of semantically meaningful words that express the target through natural language. Figure 2 illustrates the prompting mechanism for the running example.

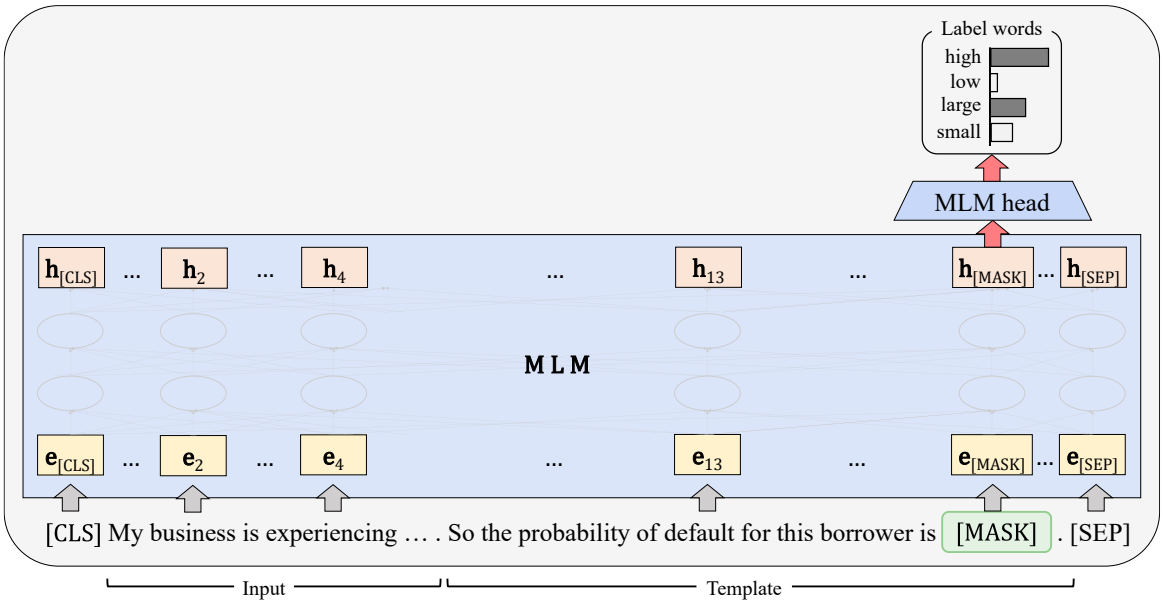


Figure 2 The example of MLM prompting

[2] The selection of label words is itself a design decision and merits explicit attention. Suitable label words should satisfy three criteria. First, they should have stable and unambiguous representations in the pre-trained MLM, which is best achieved by selecting tokens that occur with sufficient frequency in the pre-training corpus. Second, they should be directionally aligned with the outcome of interest, such that the probability mass assigned to each word can be interpreted as evidence for the corresponding class or risk state. Third, candidate label-word pairs should be selected to maximize discriminative separation on a held-out validation subset.

4.4 Constructing the Source-level Propensity Score

Prompt outputs can be sensitive to the precise wording of the template (Han et al., 2022). To reduce this sensitivity, we use m semantically similar templates that express the same decision question with different wording. For each template, the MLM produces the logits that precede the predicted probabilities of label words at the masked position. We then map these label-word logits into a scalar propensity score. Let $z(x, T_j)$ denote the vector of logits associated with the selected label words under template T_j. For each template, the template-specific propensity score is

$$S_{temp,j} = W^{\mathrm{T}} z(x, T_j) + b, \tag{3}$$

where W and b are the parameters of a linear mapping. In our setting, this score summarizes how strongly the prompted text shifts the model toward label words associated with higher default risk relative to those associated with lower default risk. A higher value of $s_{temp,j}$ indicates that, under template T_j, the text makes default-associated completions more likely than non-default-associated completions. A lower value indicates the opposite. Because the score is constructed directly from a small, semantically interpretable set of label words rather than from the full hidden state, it remains readable as the model's directional assessment of the text along a task-relevant dimension defined ex ante by the verbalizer. We then average the template-specific scores across the m templates:

$$Score = \frac{1}{m} \sum_{i=1}^{m} S_{temp,i}. \tag{4}$$

We apply this prompting procedure separately to each text source and obtain one propensity score per source. The score has three design properties that each provides a distinct advantage over conventional pooled representation approaches. The first property is semantic grounding. The interpretation of the propensity score is determined by the prompt design before the model is run. Specifically, the verbalizer establishes a semantic correspondence between label words and outcome classes, and the score represents the model's estimate of the input's likelihood under that correspondence. As a result, a reader can interpret each score directly without relying on external explanation tools. The second property is source attribution. The propensity score for each source is a function exclusively of the text in that source. Because each source is processed through its own independent prompt and forward pass, the score assigned to the given source is not affected by information from other sources. This invariance preserves source separation and can be verified without perturbation tests, gradient inspection, or attention analysis. The last property is comparability. Propensity scores from different sources are scalar quantities on a common scale because all sources share the same prompt-and-verbalizer construction and differ only in their input text. The scores are therefore comparable across sources and can be aggregated in downstream classifiers as commensurable inputs.

4.5 Off-the-shelf and Fine-tuned Implementations

We implement the proposed MLM prompting approach in two modes that differ in whether the MLM's parameters are adapted to the supervised task. The off-the-shelf mode uses a pre-trained MLM without any task-specific re-training. The model produces template-level logits at the masked position, and we construct propensity scores directly from those logits using equations (3) and (4). This mode keeps the original pre-trained model and is particularly attractive when labelled data are scarce, since prompt-based methods can extract substantial signal from the pre-trained knowledge alone (Liu et al., 2023).

The fine-tuned mode retains the same prompting architecture but allows the model to adapt to the supervised task. We place a multilayer perception (MLP) on top of the prompting outputs, i.e. the logits associated with the label words. We then estimate the model using a binary cross-entropy loss:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)), \tag{5}$$

where y_i is the observed class label and $\hat{y}_i$ denotes the predicted probability of default for observation i. The key difference from a conventional fine-tuned baseline is that our fine-tuned mode still centers the representation on prompted label-word logits rather than on a pooled [CLS] embedding. Fine-tuning, therefore, adapts a task-specific prompting structure whose outputs are still tied to interpretable label words, preserving source-level interpretability while allowing the decision boundary to adjust to the data. Figure 3 outlines the workflow of our prompting MLM approach in both off-the-shelf and fine-tuned modes.[3]

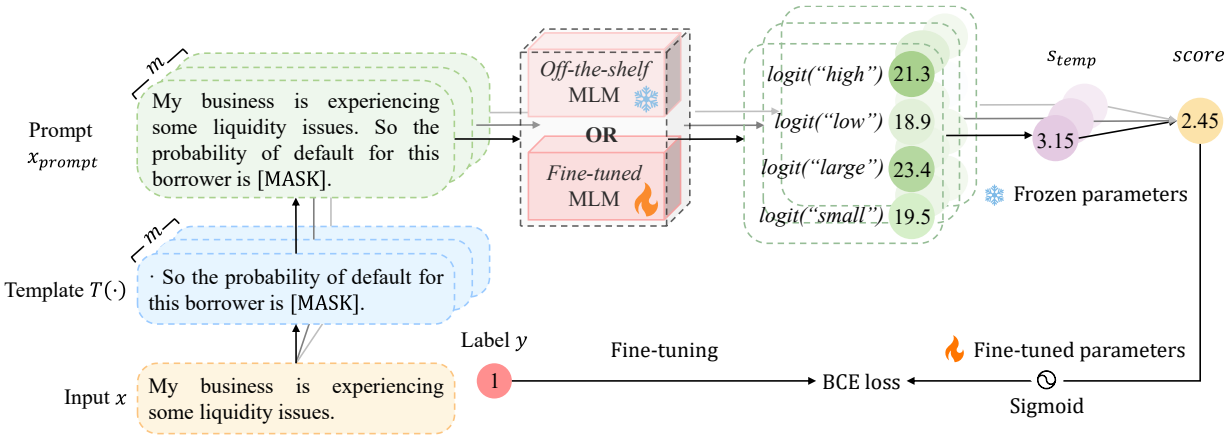


Figure 3 The MLM prompting architecture

[3] Although we illustrate the framework on binary default prediction, it is not limited to this setting. Other applications can adapt the same logic by redefining the verbalizer and the mapping from prompted outputs to the target outcome of interest.

5. Data and Model Specification

5.1 Setting and Sample

We evaluate the framework using credit default prediction for mSE loans, a setting that fits the problem our framework is designed to address. Information asymmetry between borrowers and lenders is particularly severe in this segment, because many mSE borrowers lack audited financial statements, formal credit histories, or other verifiable records of financial performance (Lee et al., 2024). As a result, lenders often rely on site visits to collect soft information that structured data miss. Loan officers write this information down as text. These texts are often inherently heterogeneous. Borrowers and loan officers provide different accounts of the same loan application, shaped by different incentives, information access, and evaluative roles. This structural divergence between borrower- and officer-produced texts provides an ideal empirical setting for assessing whether source-preserving text representations can improve prediction while retaining source-level attribution.

The loan dataset was provided by a Chinese bank operating on a national basis and originally includes 4,500 short term loans granted primarily to agricultural mSEs to finance cash flow, inventory purchases and other operating needs. After removing records with significant missing variables, the final sample consists of 4,172 loans with durations ranging from 1 to 9 months. Of these, 4,108 borrowers fully repaid at or before maturity, and 64 defaulted, where default is defined as 30 days past payment due. Each loan has 16 categorical and 14 numeric structured variables, 9 texts variables provided by borrowers themselves, 6 texts produced by loan officers, and a class label (i.e., default or non-default).

5.2 Variables

Structured Variables

The 30 structured variables cover borrower's demographics, business details, loan particulars, and so forth. For numeric variables, missing values are imputed with feature means and for categorical variables, missing values are assigned to a separate new category. Appendix A of the supplementary materials presents the definitions and summary statistics of all structured variables. To ensure consistency and comparability of input features, we follow standard credit scoring practice and encode each variable using the weight of evidence (WoE) method (Jiang et al., 2019). The WoE value in each category/bin for a given feature is calculated as the logarithm of the proportion of defaulters to the proportion of non-defaulters in that particular category/bin. A large positive value corresponds to a higher default risk and vice versa. We then use the information value (IV) and variance inflation factor (VIF) measures to remove redundancy, reduce the multicollinearity of WoE-encoded features and select an appropriate subset of features for inclusion in our models. We retain WoE-encoded features with IV between 0.01 and 0.50 and VIF no greater than 10 (Wu et al., 2025). After these preprocessing steps, 22 structured variables are retained for our subsequent modelling.

Heterogeneous Text Variables

In our mSE lending, loan officers conduct site visits to verify application information and to gather first-hand evidence about the borrower's business activities, household assets, and stated repayment intentions. During these visits, loan officers record two categories of textual narrative. The first consists of statements provided by the borrowers themselves, covering self-descriptions, business goals, loan purposes, household assets and liabilities, and repayment intentions. The second consists of independent assessments produced by the loan officer, including the officer's impression of the borrower, the borrower's neighborhood reputation, repayment intention, credit history, overall profile, and loan application. Loan underwriters then thoroughly review these documents alongside the borrower's application, financial information, and any available credit histories to make loan approval decisions. The definitions of the 15 text variables, along with summary statistics on their word lengths, are reported in Table 1.

Table 1 Definitions of text variables and statistical information on their text lengths

Text type	Definition	Summary statistics for text length			
Borrower		*Mean*	*SD*	*Min*	*Max*
Self-description	Brief introduction provided by the borrower about themselves	33.15	25.31	2.00	171.00
Personal interests	Borrower's personal interests and leisure activities	7.80	6.64	2.00	80.00
Loan purpose	Primary reason behind the borrower's loan application	6.30	3.73	3.00	31.00
Repayment plan	Borrower's planned method or source for loan repayment	5.66	2.20	3.00	49.00
Current asset overview	Statement of household current assets	20.57	10.64	2.00	35.00
Fixed asset overview	Statement of household fixed assets	26.02	34.14	2.00	257.00
Liability overview	Statement of household liabilities	14.17	25.84	2.00	257.00
Bank savings	Details of the borrower's bank savings	13.18	21.07	2.00	210.00
Supplementary information	Supplementary details provided to support loan approval	71.06	45.14	2.00	256.00
Loan officer					
Impression	Initial impression of the borrower	19.50	8.16	2.00	28.00
Neighborhood	Assessment of borrower's neighborhood relationship	11.20	11.05	2.00	104.00
Repayment intention	Analysis of the borrower's repayment intention	22.03	6.97	2.00	31.00
Credit review	Review of the borrower's credit history	13.65	11.95	2.00	110.00
Borrower evaluation	Comprehensive evaluation of the borrower	15.23	11.36	2.00	154.00
Application review	Overall assessment of the loan application	25.65	5.06	2.00	35.00

The two categories of text differ in several important respects. First, they differ in their incentive and accountability structures. Borrowers often describe themselves and their businesses with strong incentives to present their circumstances favorably, whereas loan officers prepare their assessments under professional accountability to the lending institution. Second, the two types of text differ in form. Borrower statements are usually more detailed, while loan officer assessments are shorter and more evaluative. Third, the two groups have different informational positions. Borrowers have private

knowledge of their own circumstances but may have limited understanding of how these circumstances relate to credit outcomes. Loan officers, by contrast, bring sector-specific experience in interpreting signals of repayment capacity. Below, we present four typical extracts from our dataset. The original texts are in Chinese and have been translated into English for clarity. Additional examples for both defaulters and non-defaulters are provided in Appendix B of the supplementary materials.

Extract 1 (Self-description, borrower): *"I am raising pigs on my contracted land with the help of my wife. One of my sons was married, and he is currently assisting me in my plastics business."*

Extract 2 (Personal interests, borrower): *"I don't have any bad hobbies like gambling."*

Extract 3 (Repayment intention, loan officer): *"The borrower demonstrates a strong willingness to repay, supported by a stable income and no obvious factors affecting his willingness to repay."*

Extract 4 (Borrower evaluation, loan officer): "*The borrower appears to be quite honest and eager in his loan application, showing a strong commitment to borrow and a proactive stance in collaborating with our procedures."*

Before conducting the prediction analyses, we first examine whether the source-level propensity scores derived from each text source contain useful information about default risk. We convert each of the 15 text variables into a propensity score using the framework described in Section 4, encode each score using the same WoE transformation applied to the structured variables, and estimate two logistic regressions (LRs) for each text variable on the training and validation data.

The first model regresses the default indicator only on the WoE-encoded propensity score:

$$\Pr(Default_i = 1) = \Lambda\left(\alpha + \beta_j Score_{ij}^{WoE}\right), \tag{6}$$

where $Default_i$ equals one if borrower i defaults and zero otherwise, $Score_{ij}^{WoE}$ debotes the WoE-encoded propensity score for borrower i from text type j, and $\Lambda(\cdot)$ denotes the logistic cumulative distribution function. The second specification adds the structured variables X_i as controls:

$$\Pr(Default_i = 1) = \Lambda\left(\alpha + \beta_j Score_{ij}^{WoE} + \gamma' X_i\right), \tag{7}$$

Table 2 reports the estimated coefficients and standard errors for both models. The results provide initial empirical support for the semantic grounding of our framework. In both specifications, all coefficients are positive and statistically significant at least at the 5% levels. This indicates that each of the 15 propensity scores contains default-relevant information and that the prompt-based procedure captures this information from the underlying text. The estimates also show some variation across text sources. Among borrower narratives, the fixed-asset overview, repayment plan, and liability overview have the largest coefficients. Among loan officer assessments, impression, neighbourhood, and borrower evaluation are the strongest predictors. This variation is consistent with the signalling theory prediction discussed previously and motivates the source-level analyses in Section 6.

Table 2 Propensity scores and realized default

Text type	Score only	Score + structured controls

Borrower	Coef.	SE	Coef.	SE
Fixed asset overview	0.974***	0.220	0.934***	0.245
Repayment plan	0.899***	0.278	0.770**	0.307
Liability overview	0.876***	0.324	0.991***	0.331
Bank savings	0.824***	0.246	0.582**	0.277
Personal interests	0.795***	0.180	0.756***	0.205
Loan purpose	0.783***	0.192	0.632***	0.217
Supplementary information	0.754***	0.171	0.833***	0.189
Self-description	0.685***	0.224	0.705***	0.255
Current asset overview	0.586***	0.171	0.656***	0.190
Loan officer				
Impression	1.015***	0.195	0.919***	0.211
Neighbourhood	0.891***	0.210	0.915***	0.221
Borrower evaluation	0.827***	0.181	0.816***	0.194
Application review	0.810***	0.174	0.644***	0.196
Repayment intention	0.636***	0.209	0.555**	0.223
Credit review	0.632***	0.196	0.420**	0.212
N	2,920		2,920	
Structured controls	No		Yes	

Notes: This table reports the results of LRs estimated on the development sample. The dependent variable equals one for defaulters and zero otherwise. The "Score only" model includes only the WoE-encoded propensity score, whereas the "Score + structured controls" model additionally includes the structured variables. Standard errors are reported in the SE columns. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

5.3 Model Specifications and Performance Evaluation

To formally evaluate the effectiveness of the proposed prompt-based approach, we organize the prediction pipeline into two stages. In the first stage, we use MLMs to convert each text variable into a propensity score. This text representation stage is implemented in both an off-the-shelf mode and a fine-tuned mode. In the second stage, these text-derived signals are used as inputs to downstream credit scoring models, either alone or in combination with structured variables, to predict default outcomes. To ensure that our findings are not driven by a particular text encoder or classifier, we consider multiple model combinations in both stages of the pipeline.

For the first stage, we use four representative MLMs: BERT, enhanced representation through knowledge integration (ERNIE), a lite BERT (ALBERT), and robustly optimized BERT pretraining approach (RoBERTa).[4] These models are selected because they represent widely used transformer-based architectures with different design features and have shown strong performance in language understanding tasks (Liu et al., 2023). Because the texts in our dataset are originally written in Chinese, we use Chinese-compatible versions of these models throughout the empirical analysis. BERT serves

[4] Detailed information about the MLMs used in this study can be found on their respective Hugging Face pages:
BERT: https://huggingface.co/bert-base-multilingual-uncased
ERNIE: https://huggingface.co/nghuyong/ernie-1.0-base-zh
ALBERT: https://huggingface.co/uer/albert-base-chinese-cluecorpussmall
RoBERTa: https://huggingface.co/uer/chinese_roberta_L-12_H-768

as the benchmark MLM that is trained using masked language modelling and next sentence prediction objectives (De Kok, 2025; Devlin et al., 2019). Unlike traditional neural NLP models that process text sequentially, BERT's bidirectional mechanism allows simultaneous consideration of all tokens in a sequence. ERNIE extends this framework by masking phrases and entities rather than individual tokens, which allows it to better capture prior knowledge at the phrase and entity levels (Sun et al., 2019). Considering the challenges posed by the transformer's model size, ALBERT provides a more parameter-efficient architecture through factorized embedding parameterization and cross-layer parameter sharing (Lan et al., 2019). RoBERTa differs from BERT by removing the next sentence prediction objective and adopting dynamic masking during training; it is also trained on a larger text corpus (Liu et al., 2019; Siano, 2025). Using this set of MLMs allows us to examine whether the usefulness of the proposed framework generalizes across alternative text encoders rather than depending on a single pre-trained backbone.

In the second stage, we use six commonly applied downstream classifiers: LR, light gradient-boosting machine (LGBM), extreme gradient boosting (XGB), random forest (RF), support vector machine (SVM), and MLP. These classifiers span linear, tree-based, kernel-based, and neural-network-based approaches and are widely used in credit scoring research (Hurlin et al., 2024). We use them to assess whether the predictive value of the text-derived propensity signals is robust across alternative decision rules. For brevity, detailed descriptions of the downstream classifiers are omitted and can be found in Lessmann et al. (2015) and Gunnarsson et al. (2021). Following our empirical design, the downstream classifiers are estimated using three feature configurations: structured-only features, text-only features, and combined features that integrate both structured and text-based inputs. This yields 6 models in the structured-only setting and 24 models each in the text-only and combined settings.

We divide our dataset into three distinct subsets – training, validation, and test (Kriebel & Stitz, 2022; Wu et al., 2025). We use the stratified random sampling method to select 70% of the samples as the training set, 20% of the training set as the validation set, and the remaining 30% as the test set. We tune each classifier with a grid search on the validation set. Appendix C of the supplementary materials report the hyperparameter search space for each model. We evaluate model performance on the test set with three discrimination measures. They are the area under the receiver operating characteristic curve(AUC), the Kolmogorov-Smirnov (KS) statistic, and the H-measure (Chen et al., 2023; Lee et al., 2024; Lessmann et al., 2015). Higher values of all three measures suggest better discrimination performance. Because the default rate is only 1.53 percent, we repeat the whole train-validation-test procedure 25 times and report the average. For brevity, the main text reports AUC results only, whereas the corresponding KS and H-measure results, which point the same way, are presented in the appendices. the KS and H-measure results,

6. Empirical Results

6.1 Prediction Results of the Prompt-based Approach

We begin with the main question of whether the proposed prompt-based approach improves credit default prediction compared to models that use structured variables only. Table 3 reports the out-of-sample AUC results for both the off-the-shelf and fine-tuned modes of the method, across the four representative MLMs and six classifiers.[5] Because the structured-only models do not use MLM-generated features, their results are the same across all four MLM panels.

Table 3 Predictive performance of the prompt-based approach

		Prompt (off-the-shelf)			Prompt (fine-tuned)		
MLM	Classifier	Structure	Text	Combine	Structure	Text	Combine
BERT	LR	0.641	0.662	0.670	0.641	0.766	0.765
	LGBM	0.567	0.765	0.720	0.567	0.680	0.746
	XGB	0.642	0.645	0.708	0.642	0.677	0.732
	RF	0.632	0.658	0.685	0.632	0.738	0.741
	SVM	0.610	0.672	0.674	0.610	0.756	0.753
	MLP	0.648	0.582	0.656	0.648	0.733	0.756
	Avg.	0.623	0.664	0.686	0.623	0.725	0.749
	Diff.	—	**0.041** **	**0.063** ***	—	**0.102** ***	**0.125** ***
ERNIE	LR	0.641	0.718	0.736	0.641	0.712	0.708
	LGBM	0.567	0.670	0.662	0.567	0.655	0.682
	XGB	0.642	0.675	0.736	0.642	0.704	0.712
	RF	0.632	0.626	0.683	0.632	0.696	0.712
	SVM	0.610	0.594	0.574	0.610	0.714	0.695
	MLP	0.648	0.660	0.699	0.648	0.709	0.704
	Avg.	0.623	0.657	0.682	0.623	0.698	0.702
	Diff.	—	0.034 **	0.059 ***	—	0.075 ***	0.079 ***
ALBERT	LR	0.641	0.660	0.677	0.641	0.703	0.709
	LGBM	0.567	0.632	0.665	0.567	0.715	0.709
	XGB	0.642	0.636	0.654	0.642	0.739	0.723
	RF	0.632	0.683	0.685	0.632	0.720	0.715
	SVM	0.610	0.695	0.697	0.610	0.683	0.702
	MLP	0.648	0.662	0.700	0.648	0.668	0.706
	Avg.	0.623	0.661	0.680	0.623	0.705	0.711
	Diff.	—	0.038 ***	0.057 ***	—	0.081 ***	0.087 ***
RoBERTa	LR	0.641	0.683	0.666	0.641	0.686	0.682
	LGBM	0.567	0.600	0.642	0.567	0.669	0.704
	XGB	0.642	0.646	0.638	0.642	0.707	0.711
	RF	0.632	0.652	0.663	0.632	0.674	0.659

[5] We further analyze the performance of our proposed method using individual templates. The unreported results show that different MLMs display distinct preferences for specific templates. However, predictions based on the average results across all templates consistently outperform those using individual templates for all MLMs. These findings suggest that employing a diverse set of templates enhances the robustness and accuracy of MLM prompting outcomes.

SVM	0.610	0.710	0.684	0.610	0.685	0.692
MLP	0.648	0.635	0.667	0.648	0.692	0.704
Avg.	0.623	0.654	0.660	0.623	0.685	0.692
Diff.	—	0.031***	0.037***	—	0.062***	0.069***

Notes: This table reports the average AUC for predicting the default risk of borrowers under the off-the-shelf and fine-tuned settings of the proposed prompt-based approach. Results are presented for four MLMs and six classifiers. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average AUC across classifiers. Diff. reports the difference in average AUC relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The off-the-shelf results show that text-derived features generated by pre-trained MLMs contains meaningful predictive information for default prediction. Across all classifiers, the average AUC of the structured-only models is 0.623. Text-only models perform better across all MLMs, with average AUC values ranging from 0.654 to 0.664, and every improvement over the structured-only benchmark is statistically significant at the 5% level. These findings suggest that, even without task-specific fine-tuning, MLM prompting can extract informative signals from heterogeneous textual data for identifying borrower default risk. Predicative performance improves further when structured and text-derived features are combined, increasing the average AUC to between 0.660 and 0.686. All gains are statistically significant at the 1% level. These results indicate that the two feature sets contain complementary information. Structured variables capture standard borrower characteristics, while prompted text provides additional signals that improve discrimination between defaulters and non-defaulters. Although the four MLMs differ somewhat in predictive performance, they yield a consistent conclusion that text-only models outperform structured-only models, and combined models generally achieve the strongest performance.

The fine-tuned implementation strengthens these findings. After fine-tuning MLMs, the average AUC of the text-only models increases to between 0.685 and 0.725, while the average AUC of the combined models rises to between 0.692 and 0.749. All average improvements relative to the structured-only benchmark remain positive and statistically significant at the 1% level. These findings further suggest that fine-tuning enables MLMs to adapt more effectively to the downstream task. For instance, the combined BERT-LR model improves from an AUC of 0.670 in the off-the-shelf setting to 0.765 after fine-tuning. More generally, average AUC values increase for both the text-only and combined specifications across all four MLMs, which demonstrate that fine-tuning helps the models capture more subtle and task-relevant patterns in credit-related text.

Overall, Table 3 shows that the proposed prompt-based approach is effective in both settings. In the off-the-shelf mode, prompted textual features improve prediction relative to structured-only models. In the fine-tuned mode, these gains become larger, and the strongest results are generally obtained when textual and structured features are used jointly. Taken together, the evidence indicates that heterogeneous textual data contains valuable credit-relevant information and that the proposed framework can effectively extract and integrate this information for default prediction. The KS and H-

measure results reported in Appendix D of the supplementary material are consistent with the AUC findings.

6.2 How Text Shapes Default Predictions

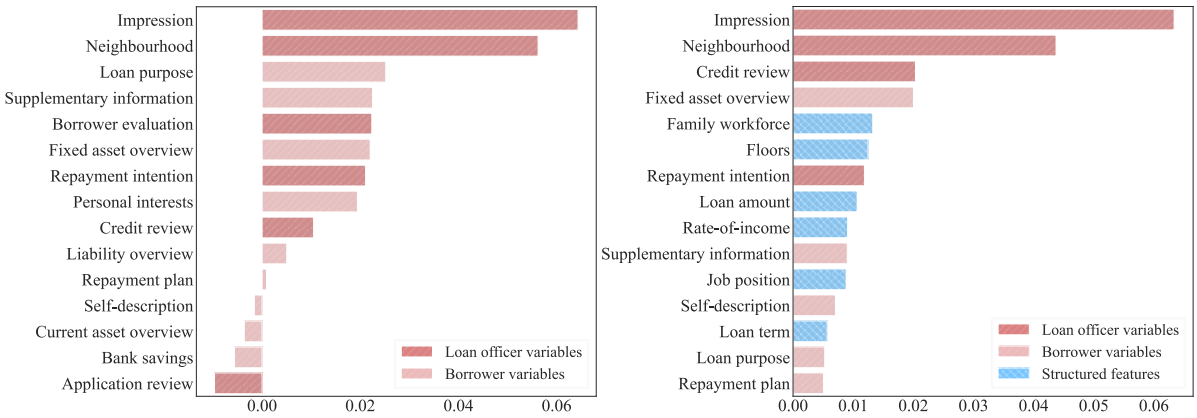
The results in Section 6.1 show that text-derived propensity scores improve default prediction. We next examine how textual information changes the model's ranking of borrower risk and which text sources contribute most to the predictions. Because the BERT-LGBM specification achieves strong predictive performance after incorporating text, we use it as the focal model in this analysis.

We first compare the predicted default probabilities from the structured-only, text-only, and combined models using Spearman rank correlations (Stevenson et al., 2021). Table 4 shows that the correlation between the structured-only and text-only predictions is close to zero: 0.035 in the off-the-shelf setting and 0.055 in the fine-tuned setting. These low correlations suggest that the two models rank borrowers in significantly different ways. Thus, the textual features do not simply reproduce the information already contained in structured credit variables; instead, they capture a distinct source of default-relevant information. The combined models are more strongly correlated with the text-only models than with the structured-only models. In the off-the-shelf setting, the combined model has a correlation of 0.477 with the text-only model, compared with 0.184 with the structured-only model. In the fine-tuned setting, the corresponding correlations are 0.567 and 0.269. This pattern suggests that, once text-derived propensity scores are included, they play an important role in shaping the model's risk ranking. The higher correlations in the fine-tuned setting further suggest that supervised adaptation helps align the text-derived signals more closely with the default prediction task.

Table 4 Spearman's rank correlation between predicted default probabilities (test set)

Mode	Subset	Structure	Text	Combine
Prompt (off-the-shelf)	Structure	1.000		
	Text	0.035	1.000	
	Combine	0.184	0.477	1.000
Prompt (fine-tuned)	Structure	1.000		
	Text	0.055	1.000	
	Combine	0.269	0.567	1.000

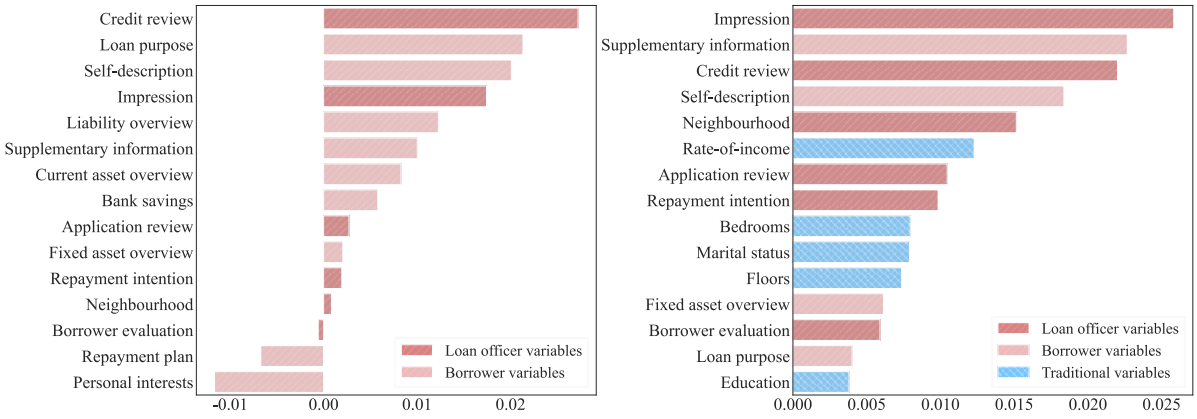
We then examine which inputs drive predictions using permutation importance. Figures 4 and 5 report most important features for the text-only and combined models in the off-the-shelf and fine-tuned settings. As the patterns observed using Shapley value importance are consistent with those using the permutation importance, we provide the corresponding results in Appendix E of the supplementary materials.



(a) (b)

Figure 4 Permutation importance of (a) text-only, (b) combined model in off-the-shelf mode

The results of the text-only model in Figure 4(a) show that the 'Impression' and 'Neighborhood' features play the most influential role in predicting borrowers' default. This finding suggests the model's strong reliance on the loan officer's holistic view of the borrower and their evaluation of the borrower's neighborhood relationship. Features such as 'Loan purpose', 'Supplementary information', and 'Borrower evaluation' also contribute meaningfully to the model's prediction performance. When we examine the combined model results displayed in Figure 4(b), we observe that 'Impression' and 'Neighborhood' remain the most important predictors for forecasting outcomes. However, in contrast to the text-only model, the combined model attributes greater importance to 'Credit review' and 'Fixed asset overview'. This shift suggests that the integration of structured data modifies the model's reliance on certain textual features, likely due to informational overlaps or complementarity between the two modalities.



(a) (b)

Figure 5 Permutation importance of (a) text-only, (b) combined model in fine-tuned mode

Considering the fine-tuned mode, Figure 5(a) shows that 'Credit review', 'Loan purpose', and 'Self-description' emerge as the most influential features of the text-only model for predicting default. These findings suggest that the loan officer's assessment of the borrower's credit history, the borrower's stated purpose for the loan, and their self-presentation play critical roles in determining default risk. Notably, certain features such as 'Borrower evaluation' and 'Personal interests', which were initially

important in the off-the-shelf mode, become less influential in the fine-tuned mode. This shift underscores the role of contextual and domain-specific learning in shaping the model's interpretation of relevant features and, by extension, its predictive behavior. For the combined model, Figure 5(b) reinforces our earlier conclusion that textual features play a dominant role in predictive performance compared to structured features. Textual features such as 'Impression' and 'Credit review' continue to rank among the most important predictors.

Taken together, the feature-importance results show that predictive information is distributed across multiple text sources, rather than being concentrated in a single field. They also reveal a systematic difference between source types. Loan officer assessments generally have greater predictive importance than borrower self-reports, consistent with the view that third-party evaluations contain more disciplined and decision-relevant soft information. At the same time, borrower narratives remain informative, especially after fine-tuning, indicating that self-reported text also contains useful credit-risk signals.

6.3 Performance Comparison with Benchmarks

We next compare the proposed prompt-based method with established NLP benchmarks. This analysis examines whether the gains documented in Section 6.1 reflect the proposed representation design rather than the use of textual information alone. Because BERT achieves the strongest average performance among the MLM backbones in Section 6.1, we use it as the focal MLM in this comparison.

We construct benchmarks that correspond to the two implementation modes of our framework. In the off-the-shelf setting, we compare our method with Ada-002[6], a pre-trained text embedding model that maps each text input into a 1,536-dimensional dense vector (Wu et al., 2025). These embeddings are then used as inputs to the same downstream classifiers as in the main analysis. Ada-002 therefore represents a standard off-the-shelf approach that requires no task-specific fine-tuning. In the fine-tuned setting, we compare our method with a conventional fine-tuned BERT model, denoted as BERT-CLS (Stevenson et al., 2021; Kriebel & Stitz, 2022). This benchmark follows the standard *pre-train and fine-tune* paradigm: the pooled [CLS] representation from BERT is fed into a prediction head (a single-layer neural network, NN) and trained directly for default classification. For the combined specification, the [CLS] representation is concatenated with structured variables before being passed to the prediction head. Unlike our prompt-based method, BERT-CLS does not first convert each text source into a separate scalar propensity score that can then be used by alternative downstream classifiers. It is therefore evaluated as an end-to-end fine-tuned model rather than through classifier-specific second-stage models. Table 5 reports the out-of-sample AUC results for both comparisons.

[6] Ada-002, known as text-embedding-ada-002, is a transformer-based text representation model proposed by OpenAI in December 2022 that can produce sequence embeddings for text inputs. It outperforms all of OpenAI's previous embedding models on text search, code search, and sentence similarity tasks and achieves a comparable performance on text classification (Wu et al., 2025).

Table 5 Performance comparison with NLP benchmarks

Method	Subset	LR	LGBM	XGB	RF	SVM	MLP	Avg.	Diff.
Panel A. Off-the-shelf mode									
Ada-002	Text	0.642	0.569	0.707	0.639	0.634	0.617	0.635	—
Prompt	Text	0.662	0.765	0.645	0.658	0.672	0.582	0.664	0.029*
Ada-002	Combine	0.651	0.699	0.665	0.685	0.648	0.655	0.667	—
Prompt	Combine	0.670	0.720	0.708	0.685	0.674	0.656	0.686	0.018***
Panel B. Fine-tuned mode									
BERT-CLS	Text	—	—	—	—	—	—	0.623	—
Prompt	Text	0.766	0.680	0.677	0.738	0.756	0.733	0.725	**0.102****
BERT-CLS	Combine	—	—	—	—	—	—	0.689	—
Prompt	Combine	0.765	0.746	0.732	0.741	0.753	0.756	0.749	**0.060****

Notes: This table reports average AUC values. Panel A compares the Prompt (off-the-shelf) method with Ada-002, and Panel B compares the Prompt (fine-tuned) method with BERT-CLS. Text denotes text-only models, and Combine denotes models using both structured and text-derived features. Classifier-specific entries are not applicable for BERT-CLS because it is trained as an end-to-end fine-tuned model. Avg. reports the average AUC across classifiers where applicable. Diff. reports the difference in average AUC relative to the corresponding benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 5 shows that the prompt-based method outperforms the corresponding NLP benchmarks in both implementations. In the off-the-shelf setting, the prompt approach improves the average AUC relative to Ada-002 from 0.635 to 0.664 in the text-only specification and from 0.667 to 0.686 in the combined specification. These gains suggest that the prompted propensity-score representation is more effective than using generic dense embeddings for this credit default prediction task. The improvement is larger in the fine-tuned setting. Compared to BERT-CLS, the prompt approach increases the average AUC from 0.623 to 0.725 in the text-only specification and from 0.689 to 0.749 in the combined specification. This comparison is particularly informative because both approaches use BERT as the underlying MLM. The performance difference therefore suggests that the gains are not driven solely by fine-tuning BERT, but also by how textual information is represented. Instead of relying on a pooled [CLS] representation, the proposed method maps each text source into a task-specific propensity score before downstream prediction.

We also observe that the gains are more pronounced in the text-only specification than in the combined specification. This pattern is consistent with the idea that structured variables already provide useful credit information, leaving less incremental room for textual representations to improve performance when both feature sets are used jointly. Our prompt approach continues to outperform the benchmarks after structured variables are included, which indicates that its text-derived signals remain complementary to standard credit features. Overall, these results show that the proposed framework improves over both off-the-shelf embeddings and conventional fine-tuned BERT representations. The supplementary KS and H-measure results in Appendix F provide broadly consistent evidence.

6.4 Economic Value in Lending Decisions

The preceding analyses show that the proposed prompt-based method improves statistical

prediction. We next examine whether these predictive gains translate into economically meaningful improvements in lending decisions. This step is important because a credit scoring model is valuable to lenders only if better risk discrimination leads to better approval and rejection decisions. We focus on the fine-tuned BERT-LGBM combined model, which uses both structured variables and text-derived propensity scores, and compare its realized profit with that of the corresponding structured-only model.

We follow Kozodoi et al. (2025) to evaluate lending profitability. For each loan, let L_i denote the loan amount and assume an interest rate of $r = 10\%$. In the event of a default (denoted as $D_i = 1$), the lender recovers $L_i \times (1 + r) \times (1 - LGD)$, where we assume a loss given default (LGD) of 0.9 to reflect the high cost of lending to defaulters (Wu et al., 2025). For a non-defaulter, the expected revenue is $L_i \times (1 + r)$. To assess overall profitability, we rank all borrowers in the test set in descending order based on their predicted probability of default and use Equation (8) to calculate total profits when rejecting a certain number of borrowers.

$$Profit = \sum_{i}^{N} [D_i \times L_i \times (1 + r) \times (1 - LGD) + (1 - D_i) \times L_i \times (1 + r) - L_i] \quad (8)$$

Figure 6 plots the profit difference between the combined model and the structured-only model across rejection thresholds. Positive values indicate that the combined model yields higher realized profits than the structured-only model. Since borrowers are rejected in descending order of predicted default probability, the x-axis represents a progressive exclusion of the riskiest borrowers first, followed by medium-risk borrowers, and finally, low-risk borrowers. As shown in Figure 6, the combined model generates higher profit over a broad range of rejection thresholds. The profit difference increases as more high-risk borrowers are rejected and reaches a peak of approximately ¥370,000 when around 800 borrowers are excluded. This pattern indicates that incorporating textual information helps the model identify costly default risk that is not fully captured by structured credit variables. After the peak, the profit advantage remains relatively stable, which suggests that the combined model continues to preserve profitability even when the rejection threshold becomes more conservative.

Overall, the profit analysis shows that the predictive gains from heterogeneous textual information have direct economic value. By improving the ranking of borrower risk, the prompt-based model enables lenders to avoid more costly defaults while preserving profitable lending opportunities. These results complement the statistical evidence in Sections 6.1–6.3 and show that the proposed framework can improve both predictive accuracy and lending profitability.

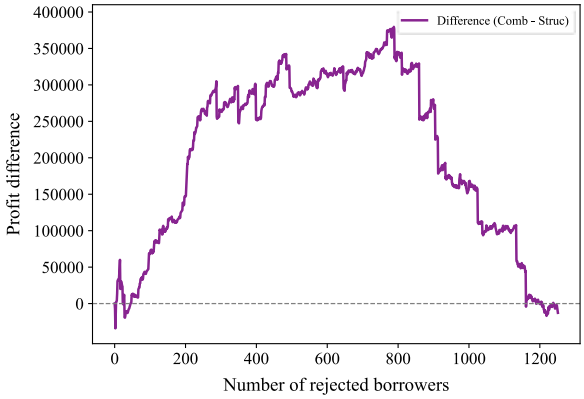


Figure 6 Profit difference curve for the combined model in fine-tuned mode

7. Validating the Framework Design and Generalizability

Section 7 further validates the proposed approach along three dimensions. First, we conduct a design ablation study to identify the contribution of the prompt-based mechanism and the two-stage learning procedure. Second, we examine model-level generalizability by applying the framework to three LLMs and assessing whether text-derived signals remain predictive across different model families. Third, we evaluate data-level generalizability using an alternative loan dataset.

7.1 Design Ablation: The Role of Prompting and Two-Stage Learning

The preceding analyses show that the proposed prompt-based approach improves default prediction compared to structured-variable models and established NLP benchmarks. We next examine whether these gains come from the two core design artifacts of the framework: the prompt-based masked-token representation and the two-stage learning procedure. We focus on fine-tuned implementation because these design choices are defined relative to standard supervised MLM fine-tuning. The off-the-shelf mode does not involve task-specific training and therefore does not provide a comparable setting for separating single-stage and two-stage learning effects.

As in Section 6.3, we use BERT as the focal MLM and compare three BERT-based specifications. The first is BERT-CLS, the conventional fine-tuned BERT benchmark. It uses the pooled [CLS] representation as the text representation, passes it to a single-layer NN, and trains the model directly for default classification. This specification follows the standard *pre-train and fine-tune* paradigm but does not use a cloze-style prompt or label-word mapping. The second specification is BERT-MASK, a prompted single-stage BERT model. It uses the same cloze-style templates and verbalizer as our proposed approach. For each text input, BERT-MASK extracts the logits assigned to the label words at the [MASK] position and passes these logits directly to a single-layer NN for end-to-end default classification. Thus, BERT-MASK preserves the prompt-based [MASK] representation but does not use the two-stage learning procedure. The third specification is the proposed Prompt model in the fine-tuned mode. Like BERT-MASK, it extracts label-word logits from the MLM prediction head at the [MASK] position. It then converts these prompted outputs into source-level default propensity scores

and then trains downstream classifiers on these scores, either alone or jointly with structured variables. The Prompt model therefore combines prompt-based representation with the two-stage learning procedure. Figure 7 summarizes the relationship among the three specifications, and Table 6 reports their out-of-sample AUC values.

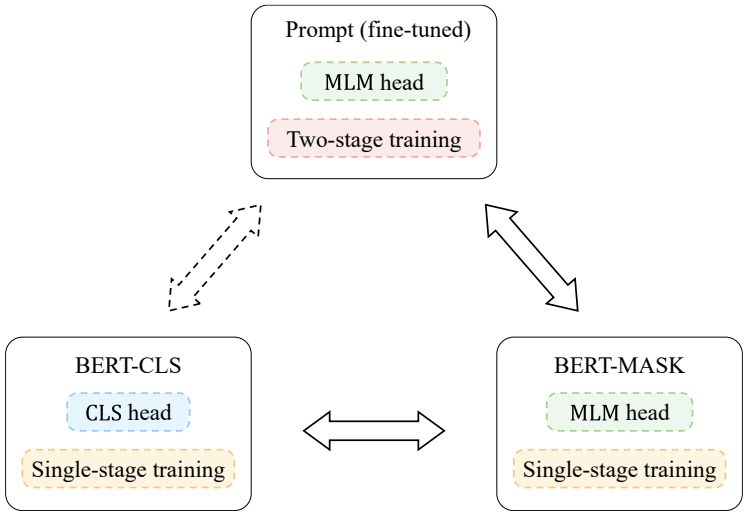


Figure 7 Comparison of model training strategies

Table 6 shows that both design artifacts contribute to predictive performance. First, BERT-MASK outperforms BERT-CLS in both feature settings: the text-only AUC increases from 0.623 to 0.689, and the combined AUC increases from 0.689 to 0.703. Because the two specifications use the same BERT backbone and single-stage training architecture, this comparison isolates the value of replacing the pooled [CLS] representation with label-word logits generated at the [MASK] position. The results suggest that framing default prediction as a masked-token completion task yields a more task-aligned textual representation than a generic pooled sentence embedding. Second, the Prompt model further improves on BERT-MASK. The average AUC increases from 0.689 to 0.725 in the text-only setting and from 0.703 to 0.749 in the combined setting, with the gains significant at the 5% and 10% levels, respectively. Because Prompt and BERT-MASK use the same cloze-style templates, verbalizer, and [MASK]-based label-word logits, this comparison identifies the incremental value of the two-stage learning procedure. The evidence suggests that converting prompted outputs into source-level propensity scores, rather than using them only within a single end-to-end classifier, helps downstream models integrate information across heterogeneous text sources and structured variables. Overall, these results suggest that our proposed prompt-based approach's gains are not driven solely by fine-tuning BERT. They arise from combining a task-aligned prompt representation with a source-preserving two-stage prediction pipeline. The supplementary KS and H-measure results in Appendix G provide the consistent findings.

Table 6 Performance comparison of different model training strategies

Method	Text representation	Training architecture	Text	Combine
BERT-CLS	[CLS] pooled representation	Single-stage; single-layer NN	0.623	0.689
BERT-MASK	[MASK] label-word logits	Single-stage; single-layer NN	0.689	0.703

Prompt (fine-tuned)	Source-level propensity scores from [MASK] logits	Two-stage; six classifiers, avg.	0.725	0.749
Diff.			0.036**	0.046*

Notes: This table reports average AUC values for the design ablation in the fine-tuned setting. Text denotes text-only models, and Combine denotes models using both structured and text-derived features. For Prompt, reported values are averaged across the six downstream classifiers. Diff. reports the difference in average AUC between Prompt and the stronger benchmark (BERT-MASK) in the corresponding feature setting, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

7.2 Generalizability Across Large Language Models

We also examine whether the proposed prompt-based approach can generalize beyond MLM backbones used in the main analysis. In recent years, LLMs are increasingly used as general-purpose tools for textual analysis in financial and managerial applications. This raises the question: whether our source-preserving prompt logic can be transferred to from MLMs to LLMs. This extension is nontrivial because our primary implementation is based on cloze-style prompting, which is naturally aligned with the masked-token prediction objective of MLMs, whereas LLMs are typically optimized for autoregressive generation and instruction following. We evaluate three LLMs: GPT-3.5, GPT-4, and DeepSeek-Chat.[7] Because these models are not fine-tuned on our mSE loan dataset, the analysis focuses on the off-the-shelf mode only. For each text source, we apply the same source-preserving prompt logic to obtain a text-derived propensity score and then evaluate its predictive value using the same six downstream classifiers as in the main analysis. Table 7 reports the out-of-sample AUC results.

Table 7 Predictive performance of the prompt-based approach (LLMs)

LLM	Subset	LR	LGBM	XGB	RF	SVM	MLP	Avg.	Diff.
GPT-3.5	Structure	0.641	0.567	0.642	0.632	0.610	0.647	0.623	—
	Text	0.580	0.663	0.607	0.597	0.538	0.620	0.601	-0.023
	Combine	0.650	0.656	0.563	0.676	0.645	0.665	0.642	0.019*
GPT-4	Structure	0.641	0.567	0.642	0.632	0.610	0.647	0.623	—
	Text	0.627	0.559	0.588	0.539	0.686	0.607	0.601	**-0.023**
	Combine	0.644	0.649	0.678	0.619	0.647	0.655	0.649	0.025***
DeepSeek-Chat	Structure	0.641	0.567	0.642	0.632	0.610	0.647	0.623	—
	Text	0.595	0.648	0.611	0.613	0.555	0.520	0.590	-0.033
	Combine	0.628	0.606	0.695	0.687	0.654	0.670	0.657	**0.033*** **

Notes: This table reports average AUC values for the prompt-based approach using LLMs in the off-the-shelf setting. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average AUC across the six downstream classifiers. Diff. reports the difference in average AUC relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 7 shows a consistent pattern across the three LLMs. Text-only models outperform random guess for classification, but their average AUC values remain below those of the structured-only models. This pattern differs from the MLM results in Section 6.1, where text-derived features alone generally

[7] The GPT-3.5 model denotes GPT-3.5-turbo, updated as of May 27, 2023. The GPT-4 model is updated as of November 8, 2023. The DeepSeek-Chat model is updated as of July 17, 2025.

outperform structured variables. The weaker stand-alone performance of LLM-derived text signals is consistent with the difference between MLM and LLM pre-training objectives: our prompts are designed around masked-token completion and are therefore more directly aligned with MLMs than with off-the-shelf autoregressive LLMs. The combined models, however, outperform the structured-only models for all three LLMs, and the gains are statistically significant in each case. This result suggests that, although LLM-derived text signals do not outperform structured variables on their own, they still contain incremental default-relevant information when combined with standard credit features. This finding is consistent with the broader evidence in our study that textual information and structured variables capture complementary dimensions of borrower risk. Overall, these results support model-level generalizability by showing that the source-preserving prompt logic can be transferred to generative LLMs, although its stand-alone predictive strength is weaker than in the MLM setting. The KS and H-measure results reported in Appendix H are broadly consistent with the AUC findings.

7.3 External Validation on an Alternative Dataset

We further validate data-level generalizability by replicating the main analysis on an alternative loan dataset. This dataset is drawn from the same banking context but differs from the primary sample in loan type, sample composition, and available text sources. It therefore provides an additional setting in which to assess whether the proposed approach can extract default-relevant signals from heterogeneous borrower and loan officer texts. Appendix I describes the construction and characteristics of the alternative dataset. We apply the same experimental design as in the main analysis and Table 8 reports the out-of-sample AUC results.

Table 8 Predictive performance of the prompt-based approach (alternative dataset)

		Prompt (off-the-shelf)			Prompt (fine-tuned)		
MLM	Classifier	Structure	Text	Combine	Structure	Text	Combine
BERT	LR	0.549	0.762	0.726	0.549	0.731	0.692
	LGBM	0.625	0.684	0.744	0.625	0.716	0.711
	XGB	0.636	0.675	0.675	0.636	0.741	0.715
	RF	0.641	0.694	0.716	0.641	0.741	0.711
	SVM	0.537	0.538	0.713	0.537	0.645	0.702
	MLP	0.541	0.658	0.676	0.541	0.676	0.668
	Avg.	0.588	0.669	0.708	0.588	0.708	0.700
	Diff.	—	0.081 ***	**0.120** ***	—	0.120 ***	0.112 ***
ERNIE	LR	0.549	0.736	0.685	0.549	0.753	0.731
	LGBM	0.625	0.769	0.708	0.625	0.676	0.648
	XGB	0.636	0.748	0.704	0.636	0.752	0.714
	RF	0.641	0.746	0.788	0.641	0.749	0.736
	SVM	0.537	0.707	0.715	0.537	0.745	0.714
	MLP	0.541	0.668	0.613	0.541	0.709	0.696
	Avg.	0.588	0.729	0.702	0.588	0.731	0.706
	Diff.	—	**0.141** ***	0.114 ***	—	0.142 ***	0.118 ***

ALBERT	LR	0.549	0.668	0.642	0.549	0.760	0.750
	LGBM	0.625	0.670	0.757	0.625	0.749	0.730
	XGB	0.636	0.735	0.682	0.636	0.776	0.757
	RF	0.641	0.736	0.745	0.641	0.756	0.753
	SVM	0.537	0.636	0.735	0.537	0.704	0.749
	MLP	0.541	0.566	0.625	0.541	0.718	0.704
	Avg.	0.588	0.668	0.698	0.588	0.744	0.741
	Diff.	—	0.080 ***	0.109 ***	—	**0.156** ***	**0.152** ***
RoBERTa	LR	0.549	0.668	0.642	0.549	0.739	0.734
	LGBM	0.625	0.650	0.611	0.625	0.697	0.686
	XGB	0.636	0.623	0.656	0.636	0.719	0.712
	RF	0.641	0.693	0.644	0.641	0.736	0.725
	SVM	0.537	0.674	0.664	0.537	0.632	0.696
	MLP	0.541	0.657	0.618	0.541	0.683	0.683
	Avg.	0.588	0.661	0.639	0.588	0.701	0.706
	Diff.	—	0.073 ***	0.051 ***	—	0.113 ***	0.118 ***

Notes: This table reports the average AUC for predicting the default risk of borrowers under the off-the-shelf and fine-tuned settings of the proposed prompt-based approach. Results are presented for four MLMs and six classifiers. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average AUC across classifiers. Diff. reports the difference in average AUC relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 8 shows that the findings generalize to the alternative dataset. The average AUC of the structured-only models is 0.588. In the off-the-shelf setting, text-only models outperform this benchmark for all four MLMs, with average AUC values ranging from 0.661 to 0.729. Combined models also improve on the structured-only benchmark, with average AUC values ranging from 0.639 to 0.708. All differences relative to the structured-only benchmark are statistically significant at the 1% level. These results demonstrate that prompted MLMs extract default-relevant textual signals even without task-specific fine-tuning. The fine-tuned setting yields a similar, and slightly stronger, pattern. The average AUC ranges from 0.701 to 0.744 for text-only models and from 0.706 to 0.741 for combined models. Again, all improvements over the structured-only benchmark are significant at the 1% level. This evidence is consistent with the main-sample results and suggest that supervised adaptation improves the alignment between prompted representations and the default prediction task. Overall, results from the alternative dataset support the data-level generalizability of the proposed approach. The KS and H-measure results reported in Appendix J are consistent with the AUC findings.

8. Conclusion and Discussion

This study develops a source-preserving, prompt-based framework for predictive analytics with heterogeneous textual data. The framework addresses a common challenge in financial decision-making where multiple text sources are produced by different agents, may contain potentially distinct information and need to be integrated without losing their source-specific meanings. Rather than combining all texts into a shared latent representation, the proposed approach converts each text source

into an interpretable default propensity score through cloze-style prompting. These source-level scores are then used as inputs to downstream credit scoring models, either alone or together with structured variables. Our proposed framework can be implemented in an off-the-shelf mode, which uses pre-trained MLMs without task-specific training, or in a fine-tuned mode, which adapts the prompted representations to the prediction task through supervised learning.

To evaluate the effectiveness of our framework, we use a unique dataset of agricultural mSE loans containing multiple borrower- and loan officer-provided text variables. The empirical results show that text-derived propensity scores significantly improve default prediction compared to structured variables alone, with the strongest performance obtained under the fine-tuned mode. The prompt-based approach also outperforms established NLP benchmarks in both off-the-shelf and fine-tuned settings. Additional analyses show that textual information is not simply a substitute for structured variables but provides complementary information about credit risk. In particular, loan officers' assessments contribute more predictive information than borrower self-reports. This is consistent with the idea that professionally provided assessments contain more reliable credit-relevant signals. Profit analysis further shows that the predictive gains can translate into economically meaningful improvements in lending decisions. Finally, the ablation study, LLM-based validation, and alternative-dataset analysis provide additional evidence on the robustness and generalizability of the framework.

Our study contributes to the growing literature on advanced NLP tools in finance and information systems and has broader implications for the design and use of text-based predictive analytics. First, the proposed approach can be extended to other domains and tasks that involve heterogeneous textual sources generated by different agents, such as management disclosures, analyst reports, audit comments, customer reviews, platform narratives, regulatory filings, and internal risk assessments. In these settings, researchers and practitioners often need to understand not only whether text improves prediction, but also which text source contributes to the prediction. Our framework provides a flexible way to construct task-specific, source-level propensity scores from heterogeneous texts. These scores can be used as predictors in downstream models, as outcome variables that measure latent textual constructs, or as intermediate signals for comparing the informational content of different text sources. Compared with neural NLP approaches that map texts into high-dimensional latent representations, this source-level design provides a more transparent way to incorporate multiple text variables into predictive models, especially in high-stakes decision-making environments. Moreover, our findings suggest that organizations should treat textual data as differentiated information rather than as a homogeneous input. Texts produced by different agents can contain different levels of predictive content. In our setting, loan officer assessments are more informative than borrower self-reports, which suggests that the value of textual data depends on who produces the text, under what informational role, and with what incentives. For organizations that increasingly collect narrative information, this finding highlights the importance of designing data collection and modelling pipelines that preserve source identity, rather than simply aggregating all available narratives into a single text representation.

This work inevitably has limitations that point to several directions for future research. First, our empirical analysis focuses on credit default prediction using two loan datasets from a specific institutional setting. Although the alternative-dataset analysis provides additional support for generalizability, future research could examine whether the framework performs similarly in other financial and managerial contexts. Second, our proposed framework relies on manually designed prompt templates and label-word pairs. While the results are robust across several models and settings, prompt design remains an important modelling choice. Future work could systematically compare alternative prompt templates, verbalizers, and automated prompt-search procedures. Third, while the proposed approach improves source-level interpretability, it does not fully explain the linguistic mechanisms through which each text affects predictions. Combining source-preserving propensity scores with more granular textual diagnostics may further improve transparency.

References

Appendix

A. Overview of Structured Variables

Table A1 Definitions and statistical information of structured features

Feature	Definition	Summary statistics			
Continuous		*Mean*	*SD*	*Min*	*Max*
Age §	The age of the business owner	40.83	8.43	20.00	64.00
Annual expense §	Total annual expense (¥)	18,306.00	38,162.44	5,456.70	330,000.00
Annual revenue	Total annual revenue (¥)	179,133.58	344,940.88	32,400.00	2,364,000.00
Bedrooms §	Number of bedrooms in the borrower's house	4.84	2.59	1.00	16.00
Family members §	Number of borrower's family members	3.57	0.93	1.00	9.00
Family workforce §	Number of workforces available in borrower's family	2.48	0.77	1.00	8.00
Floors §	Number of floors in the borrower's house	1.72	0.82	1.00	5.00
House area §	Borrower's residential land area (m^2)	174.85	132.93	50.00	1,064.70
Loan amount §	Loan amount (¥)	39,368.12	15,665.30	1,000.00	100,000.00
Loan term §	Loan term (months)	3.85	0.69	1.00	9.00
Monthly revenue §	Average monthly revenue (¥)	15,804.74	12,303.22	7,135.00	91,999.92
Rate-of-income §	The ratio of monthly repayment to income (%)	12.65	33.27	2.00	408.90
Rest amount	Remaining loan amount outstanding (¥)	482.46	4,504.41	0.00	80,000.00
Rest interest	Remaining loan interest outstanding (¥)	24.03	282.34	0.00	6,770.93
Categorical		*Number of categories*		*Values*	
Business type §	Types of borrowers' business	3		{1, 2, 3}	
Credit rating §	Borrower's credit rating	6		{1, 2, 3, 4, 5, Other}	
Customer type	Whether the borrower is a new or existing customer of the bank	2		{New_cust, Old_cust}	
Degree	Borrower's academic degree	4		{0, 4, 5, 9}	
Education §	Borrower's educational background	10		{10, 20, 30, 40, 50, 60, 70, 80, 90, 99}	
Homeownership §	Borrower's homeownership	6		{1, 2, 3, 4, 5, Other}	
House type	House type	2		{Flat, House}	
Housekeeping	Housekeeping condition	3		{Bad, Moderate, Good}	
Job position §	Borrower's job position	6		{1, 2, 3, 4, 5, Other}	
Job title §	Borrower's job title	6		{1, 2, 3, 4, 5, Other}	

License type §	Types of business license	11	{A, C, E, F, H, O, P, Q, S, Z, Other}
Marital status §	Borrower's marital status	7	{10, 21, 22, 23, 30, 40, 90}
Occupation §	Borrower's occupation	10	{0, 1, 3, 4, 5, 8, 9, X, Y, Z}
Postcode	The postcode of the business location	12	{153200, 065300, 164100, 325700, …}*
Repay type §	Types of loan repayment	3	{A, B, C}
Verified ID §	Borrower's citizenship category	7	{1, 2, 3, 4, 5, 6, Other}

Notes: We conduct a data desensitization process on the original structured data to protect private and business-sensitive information. * indicates that some infrequent categories are omitted, in the interest of space. § indicates the final features selected for our credit scoring models.

B. Examples of Defaulter and Non-defaulter's Text Variables

Table B1 Examples of defaulter and non-defaulters' text variables

Text type	Defaulter	Non-defaulter
Borrower		
Self-description	Right now, it's just the three of us at home. I'm divorced, and I have two kids who are both in college.	My spouse and I are running a farm together. Our son is doing his PhD and plans to work at the Agricultural Bureau after graduation. Our daughter is studying for her bachelor's degree in Changsha and has already secured a job—she'll start working right after she graduates.
Personal interests	I don't really have any hobbies.	I don't have any bad habits.
Loan purpose	I'd like to expand the farm operation.	I need money to support the production and day-to-day operations of the farm.
Repayment plan	I can use my business income to repay the loan.	I plan to repay the loan using the income from the farm's livestock and overall operations.
Current asset overview	I don't have much cash savings at home.	I have some accounts receivable from clients, totaling around 50,000 yuan. I also have about 5,000 yuan in cash savings at home.
Fixed asset overview	I have one refrigerator, one air conditioner, and two televisions. I also have a small four-wheel tractor, purchased in 1993 for 12,000 yuan.	I bought a laptop in 2006 for 7,800 yuan; a motorcycle in 2007 for 4,700 yuan; a desktop computer in 2008 for 5,000 yuan; and a refrigerator in 2008 for 3,500 yuan.
Liability overview	I have never taken out a loan before.	I don't have any outstanding personal loans or bank loans.
Bank savings	I have 3,000 yuan in my bank account.	I have 10,000 yuan in my bank account.
Supplementary information	Both of my children are doing well in their studies. My older daughter is studying at a women's college, and my son is attending a college of information engineering.	Our family has a very harmonious relationship. The children are doing well in their studies, and both of my parents are still in good health.
Loan officer		
Impression	The borrower is honest and sincere, embodying the simplicity and hardworking spirit typical of rural individuals.	The borrower is dependable and steady, with a strong work ethic and a willingness to endure hardship.
Neighborhood	The borrower maintains an average relationship with their neighbors.	According to conversations with the neighbors, the borrower maintains a very good and harmonious relationship with them.
Repayment intention	The borrower was cooperative during the loan investigation and showed a strong willingness to repay the loan.	The borrower has a clear understanding of the bank's loan policies and demonstrates a very positive attitude toward repayment.
Credit review	The borrower seems to be fairly honest and straightforward.	According to the loan investigation, the borrower has a good credit history and is classified as an AA-rated customer.
Borrower evaluation	The borrower is a warm and genuine person, someone who is practical and reliable in getting things done.	The borrower is honest and trustworthy. Neighbors have a positive impression of the couple, and others in the village also speak well of them.
Application review	The borrower has plenty of experience in farming. All things considered, I'd recommend approving the loan.	The borrower has been operating the farm for many years and possesses extensive experience. With a solid credit history and after reviewing the relevant documents and field visits, I recommend approving the loan for this borrower.

C. Search Space for Model Hyperparameters

Table C1 Hyperparameter search spaces

Model	Search space
	MLM
BERT	epoch = 7, learning_rate = (2e-5, 1e-5), batch_size = (16, 32, 48)
ERNIE	epoch = 7, learning_rate = (2e-5, 1e-5), batch_size = (16, 32, 48)
ALBERT	epoch = 7, learning_rate = (2e-5, 1e-5), batch_size = (16, 32, 48)
RoBERTa	epoch = 7, learning_rate = (2e-5, 1e-5), batch_size = (16, 32, 48)
	Classifier
LR	C = (0.01, 0.1, 1), solver = ("liblinear")
LGBM	learning_rate = (0.01, 0.1), max_depth = (2, 6, 8, 10), min_samples_splt = (2, 6, 8, 10), min_samples_leaf = (2, 6, 8, 10), n_estimators = (30, 50, 70, 90)
XGB	learning_rate = (0.01, 0.1), max_depth = (2, 6, 8, 10), n_estimators = (30, 50, 70, 90), scale_pos_weight = (1, 60)
RF	max_depth = (2, 6, 8, 10), min_samples_splt = (2, 6, 8, 10), min_samples_leaf = (2, 6, 8, 10), n_estimators = (30, 50, 70, 90)
SVM	C = (0.1, 1, 2, 3, 4), class_weigth = (None, "balanced"), kernel = ("rbf", "poly"), gamma = ("scale", "auto")
MLP	hidden_layer_sizes = (5, 10, 25, 50), learning_rate_init = (0.0001, 0.001, 0.01), max_iter = 100

D. Predictive Performance Using Alternative Metrics

Table D1 Predictive performance of the prompt-based approach (KS)

		Prompt (off-the-shelf)			Prompt (fine-tuned)		
MLM	Classifier	Structure	Text	Combine	Structure	Text	Combine
BERT	LR	0.305	0.448	0.396	0.305	0.459	0.456
	LGBM	0.245	0.512	0.387	0.245	0.354	0.445
	XGB	0.299	0.250	0.431	0.299	0.331	0.399
	RF	0.289	0.312	0.366	0.289	0.417	0.415
	SVM	0.224	0.338	0.322	0.224	0.457	0.442
	MLP	0.277	0.226	0.345	0.277	0.420	0.436
	Avg.	0.273	0.347	0.374	0.273	0.406	0.432
	Diff.	—	**0.074** **	**0.101** ***	—	**0.133** ***	**0.159** ***
ERNIE	LR	0.305	0.379	0.450	0.305	0.372	0.366
	LGBM	0.245	0.307	0.310	0.245	0.356	0.348
	XGB	0.299	0.306	0.429	0.299	0.391	0.378
	RF	0.289	0.298	0.341	0.289	0.367	0.372
	SVM	0.224	0.215	0.172	0.224	0.391	0.343
	MLP	0.277	0.328	0.368	0.277	0.376	0.362
	Avg.	0.273	0.306	0.345	0.273	0.376	0.361
	Diff.	—	0.033 **	0.072 ***	—	0.103 ***	0.088 ***
ALBERT	LR	0.305	0.309	0.318	0.305	0.368	0.356
	LGBM	0.245	0.317	0.298	0.245	0.391	0.368
	XGB	0.299	0.298	0.298	0.299	0.418	0.393
	RF	0.289	0.324	0.357	0.289	0.394	0.385
	SVM	0.224	0.345	0.357	0.224	0.359	0.352
	MLP	0.277	0.327	0.380	0.277	0.350	0.373
	Avg.	0.273	0.320	0.335	0.273	0.380	0.371
	Diff.	—	0.047 ***	0.062 ***	—	0.107 ***	0.098 ***
RoBERTa	LR	0.305	0.329	0.300	0.305	0.348	0.350
	LGBM	0.245	0.191	0.305	0.245	0.338	0.359
	XGB	0.299	0.293	0.289	0.299	0.358	0.367
	RF	0.289	0.317	0.291	0.289	0.317	0.299
	SVM	0.224	0.382	0.313	0.224	0.331	0.350
	MLP	0.277	0.314	0.318	0.277	0.348	0.366
	Avg.	0.273	0.304	0.303	0.273	0.340	0.349
	Diff.	—	0.031 **	0.030	—	0.067 ***	0.076 ***

Notes: This table reports the average KS statistic for predicting the default risk of borrowers under the off-the-shelf and fine-tuned settings of the proposed prompt-based approach. Results are presented for four MLMs and six classifiers. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average KS statistic across classifiers. Diff. reports the difference in average KS statistic relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

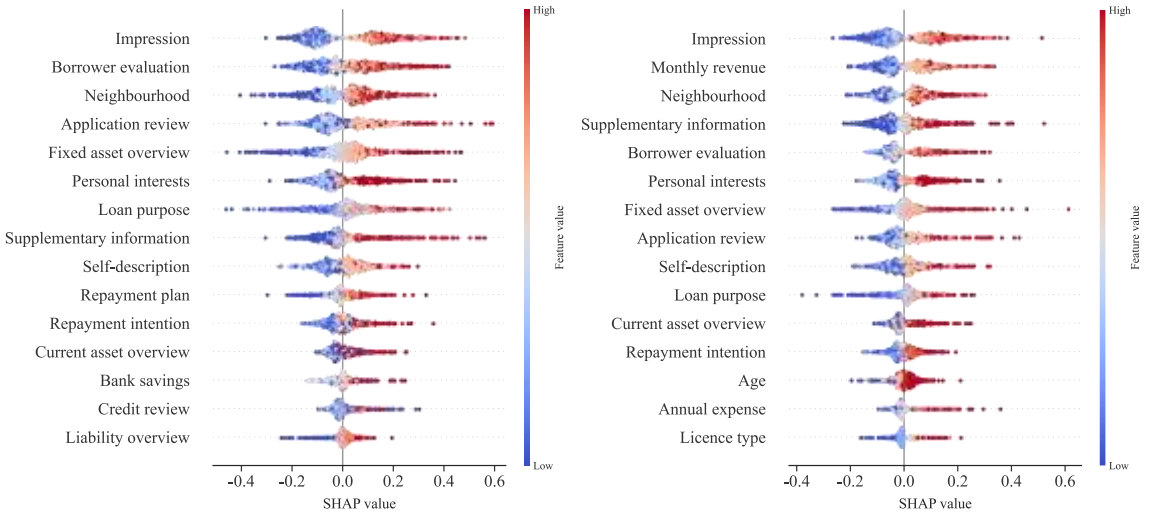
Table D2 Predictive performance of the prompt-based approach (H-measure)

		Prompt (off-the-shelf)			Prompt (fine-tuned)		
MLM	Classifier	Structure	Text	Combine	Structure	Text	Combine
BERT	LR	0.134	0.281	0.231	0.134	0.302	0.299
	LGBM	0.197	0.340	0.285	0.197	0.253	0.282
	XGB	0.196	0.146	0.314	0.196	0.246	0.269
	RF	0.143	0.222	0.276	0.143	0.295	0.297
	SVM	0.121	0.214	0.214	0.121	0.296	0.286
	MLP	0.180	0.115	0.200	0.180	0.266	0.286
	Avg.	0.162	0.220	0.254	0.162	0.276	0.287
	Diff.	—	**0.058** **	**0.092** ***	—	**0.114** ***	**0.125** ***
ERNIE	LR	0.134	0.193	0.249	0.134	0.272	0.266
	LGBM	0.197	0.227	0.248	0.197	0.257	0.257
	XGB	0.196	0.245	0.319	0.196	0.277	0.271
	RF	0.143	0.216	0.251	0.143	0.261	0.263
	SVM	0.121	0.130	0.145	0.121	0.266	0.251
	MLP	0.180	0.165	0.232	0.180	0.258	0.258
	Avg.	0.162	0.196	0.241	0.162	0.265	0.261
	Diff.	—	0.034 ***	0.079 ***	—	0.103 ***	0.099 ***
ALBERT	LR	0.134	0.173	0.211	0.134	0.251	0.251
	LGBM	0.197	0.235	0.255	0.197	0.266	0.258
	XGB	0.196	0.248	0.263	0.196	0.278	0.265
	RF	0.143	0.235	0.250	0.143	0.271	0.270
	SVM	0.121	0.228	0.195	0.121	0.243	0.241
	MLP	0.180	0.160	0.249	0.180	0.240	0.248
	Avg.	0.162	0.213	0.237	0.162	0.258	0.255
	Diff.	—	0.051 ***	0.075 ***	—	0.096 ***	0.094 ***
RoBERTa	LR	0.134	0.192	0.234	0.134	0.249	0.250
	LGBM	0.197	0.134	0.210	0.197	0.241	0.260
	XGB	0.196	0.205	0.141	0.196	0.263	0.264
	RF	0.143	0.208	0.243	0.143	0.240	0.235
	SVM	0.121	0.230	0.250	0.121	0.225	0.235
	MLP	0.180	0.142	0.205	0.180	0.239	0.246
	Avg.	0.162	0.185	0.214	0.162	0.243	0.248
	Diff.	—	0.023 *	0.052 ***	—	0.081 ***	0.086 ***

Notes: This table reports the average H-measure for predicting the default risk of borrowers under the off-the-shelf and fine-tuned settings of the proposed prompt-based approach. Results are presented for four MLMs and six classifiers. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average H-measure across classifiers. Diff. reports the difference in average H-measure relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

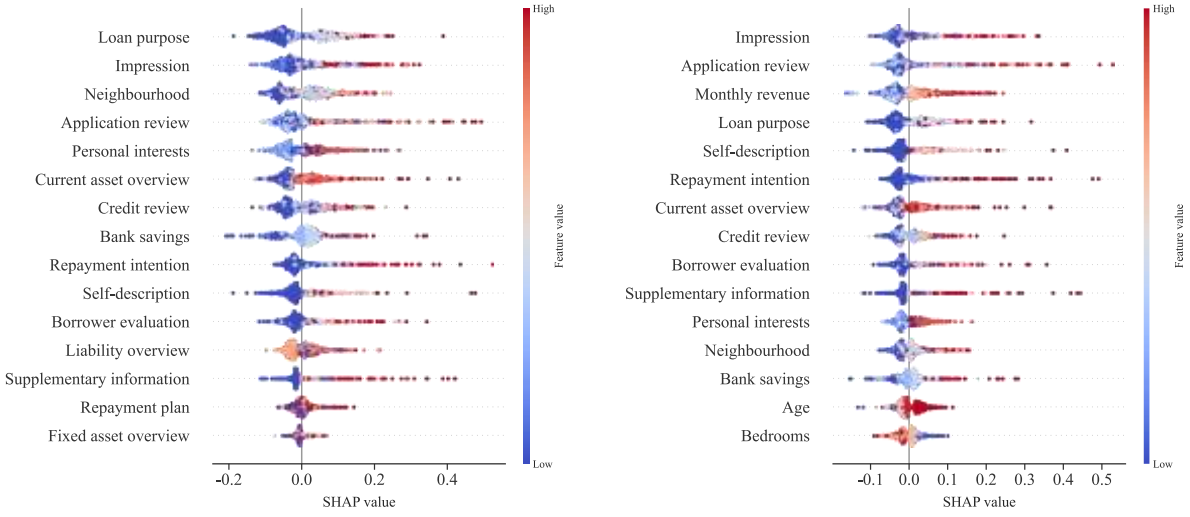
E. Robustness Checks on Feature Importance Analysis

To further verify the robustness of our feature importance analysis, we additionally use the Shapley value importance measures to evaluate the contribution of textual features to the model's predictive performance. The results, presented in Figures E1 and E2, demonstrate that incorporating structured data into the model can change the relative importance of textual features. Notably, across off-the-shelf and fine-tuned settings, textual features consistently exhibit greater predictive importance than structured features in combined models. Furthermore, the analysis highlights that loan officers' texts play a more substantial role in default prediction than borrowers' texts. Overall, these findings are consistent with our earlier conclusions and highlight the critical value of incorporating textual features in credit risk assessment.



(a) (b)

Figure E1 Shapley value importance of (a) text-only, (b) combined model in off-the-shelf mode



(a) (b)

Figure E2 Shapley value importance of (a) text-only, (b) combined model in fine-tuned mode

F. Benchmark Comparisons Using Alternative Metrics

Table F1 Performance comparison with NLP benchmarks (KS)

Method	Subset	LR	LGBM	XGB	RF	SVM	MLP	Avg.	Diff.
Panel A. Off-the-shelf mode									
Ada-002	Text	0.306	0.305	0.366	0.295	0.371	0.336	0.330	—
Prompt	Text	0.448	0.512	0.250	0.312	0.338	0.226	0.347	0.018
Ada-002	Combine	0.307	0.350	0.330	0.345	0.307	0.299	0.323	—
Prompt	Combine	0.396	0.387	0.431	0.366	0.322	0.345	0.374	0.051***
Panel B. Fine-tuned mode									
BERT-CLS	Text	—	—	—	—	—	—	0.278	—
Prompt	Text	0.459	0.354	0.331	0.417	0.457	0.420	0.406	**0.128****
BERT-CLS	Combine	—	—	—	—	—	—	0.351	—
Prompt	Combine	0.456	0.445	0.399	0.415	0.442	0.436	0.432	**0.081****

Notes: This table reports average KS statistic values. Panel A compares the Prompt (off-the-shelf) method with Ada-002, and Panel B compares the Prompt (fine-tuned) method with BERT-CLS. Text denotes text-only models, and Combine denotes models using both structured and text-derived features. Classifier-specific entries are not applicable for BERT-CLS because it is trained as an end-to-end fine-tuned model. Avg. reports the average KS statistic across classifiers where applicable. Diff. reports the difference in average KS statistics relative to the corresponding benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table F2 Performance comparison with NLP benchmarks (H-measure)

Method	Subset	LR	LGBM	XGB	RF	SVM	MLP	Avg.	Diff.
Panel A. Off-the-shelf mode									
Ada-002	Text	0.213	0.167	0.303	0.176	0.286	0.244	0.231	—
Prompt	Text	0.281	0.340	0.146	0.222	0.214	0.115	0.220	-0.012
Ada-002	Combine	0.158	0.223	0.281	0.211	0.152	0.228	0.209	—
Prompt	Combine	0.231	0.285	0.314	0.276	0.214	0.200	0.254	0.045***
Panel B. Fine-tuned mode									
BERT-CLS	Text	—	—	—	—	—	—	0.165	—
Prompt	Text	0.302	0.253	0.246	0.295	0.296	0.266	0.276	**0.111****
BERT-CLS	Combine	—	—	—	—	—	—	0.241	—
Prompt	Combine	0.299	0.282	0.269	0.297	0.286	0.286	0.287	**0.046**

Notes: This table reports average H-measure values. Panel A compares the Prompt (off-the-shelf) method with Ada-002, and Panel B compares the Prompt (fine-tuned) method with BERT-CLS. Text denotes text-only models, and Combine denotes models using both structured and text-derived features. Classifier-specific entries are not applicable for BERT-CLS because it is trained as an end-to-end fine-tuned model. Avg. reports the average H-measure across classifiers where applicable. Diff. reports the difference in average H-measure relative to the corresponding benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

G. Design Ablation Results Using Alternative Metrics

Table G1 Performance comparison of different model training strategies (KS)

Method	Text representation	Training architecture	Text	Combine
BERT-CLS	[CLS] pooled representation	Single-stage; single-layer NN	0.278	0.351
BERT-MASK	[MASK] label-word logits	Single-stage; single-layer NN	0.337	0.362
Prompt (fine-tuned)	Source-level propensity scores from [MASK] logits	Two-stage; six classifiers, avg.	0.406	0.432
Diff.			0.069*	0.070

Notes: This table reports average KS statistic values for the design ablation in the fine-tuned setting. Text denotes text-only models, and Combine denotes models using both structured and text-derived features. For Prompt, reported values are averaged across the six downstream classifiers. Diff. reports the difference in average KS statistics between Prompt and the stronger benchmark (BERT-MASK) in the corresponding feature setting, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table G2 Performance comparison of different model training strategies (H-measure)

Method	Text representation	Training architecture	Text	Combine
BERT-CLS	[CLS] pooled representation	Single-stage; single-layer NN	0.165	0.241
BERT-MASK	[MASK] label-word logits	Single-stage; single-layer NN	0.245	0.249
Prompt (fine-tuned)	Source-level propensity scores from [MASK] logits	Two-stage; six classifiers, avg.	0.276	0.287
Diff.			0.031*	0.038

Notes: This table reports average H-measure values for the design ablation in the fine-tuned setting. Text denotes text-only models, and Combine denotes models using both structured and text-derived features. For Prompt, reported values are averaged across the six downstream classifiers. Diff. reports the difference in average H-measure between Prompt and the stronger benchmark (BERT-MASK) in the corresponding feature setting, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

H. Predictive Performance with LLMs Using Alternative Metrics

Table H1 Predictive performance of the prompt-based approach (KS)

LLM	Subset	LR	LGBM	XGB	RF	SVM	MLP	Avg.	Diff.
GPT-3.5	Structure	0.305	0.245	0.299	0.289	0.223	0.277	0.273	—
	Text	0.206	0.329	0.208	0.215	0.176	0.269	0.234	-0.039
	Combine	0.319	0.278	0.267	0.329	0.284	0.308	0.297	0.024*
GPT-4	Structure	0.305	0.245	0.299	0.289	0.223	0.277	0.273	—
	Text	0.335	0.235	0.207	0.181	0.341	0.241	0.257	**-0.016**
	Combine	0.287	0.272	0.423	0.264	0.288	0.293	0.305	0.032**
DeepSeek-Chat	Structure	0.305	0.245	0.299	0.289	0.223	0.277	0.273	—
	Text	0.224	0.307	0.261	0.27	0.193	0.148	0.234	-0.039
	Combine	0.264	0.263	0.383	0.344	0.295	0.311	0.31	**0.037*** **

Notes: This table reports average KS statistic for the prompt-based approach using LLMs in the off-the-shelf setting. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average KS statistic across the six downstream classifiers. Diff. reports the difference in average KS statistic relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table H2 Predictive performance of the prompt-based approach (H-measure)

LLM	Subset	LR	LGBM	XGB	RF	SVM	MLP	Avg.	Diff.
GPT-3.5	Structure	0.134	0.197	0.196	0.143	0.121	0.18	0.162	—
	Text	0.044	0.149	0.069	0.136	0.112	0.111	0.104	-0.058
	Combine	0.149	0.222	0.184	0.213	0.133	0.196	0.183	**0.021*** **
GPT-4	Structure	0.134	0.197	0.196	0.143	0.121	0.18	0.162	—
	Text	0.145	0.067	0.07	0.086	0.17	0.127	0.111	**-0.051**
	Combine	0.131	0.159	0.245	0.149	0.136	0.187	0.168	0.006
DeepSeek-Chat	Structure	0.134	0.197	0.196	0.143	0.121	0.18	0.162	—
	Text	0.078	0.178	0.083	0.101	0.108	0.031	0.097	-0.065
	Combine	0.115	0.097	0.265	0.187	0.147	0.176	0.164	0.003

Notes: This table reports average H-measure for the prompt-based approach using LLMs in the off-the-shelf setting. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average H-measure across the six downstream classifiers. Diff. reports the difference in average H-measure relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

I. Overview of Alternative Dataset

The alternative dataset sourced from the same Chinese bank, originally comprises 4,000 commercial and industrial loans primarily issued to mSEs to address immediate operational needs. After removing records lacking textual information and those with significant missing variables, the final sample consists of 3,727 loans with durations ranging from 1 to 9 months. All the loans are closed – 3,667 loans were fully paid off at maturity or early, and 60 borrowers defaulted. Consistent with our primary dataset, default is defined as a payment delay exceeding 30 days past the due date. Each borrower has 17 nominal and 13 numeric attributes, 6 types of texts provided by borrowers themselves, 6 types of texts produced by loan officers, and a final class label (i.e., defaulter or non-defaulter). Table I1 presents the definitions and summary statistics of all standard variables in the alternative dataset, and Table I2 provides the definitions of text variables and summary statistics of their text lengths.

Table I1 Definitions and statistical information of structured features

Feature	Definition	Summary statistics			
Continuous		*Mean*	*SD*	*Min*	*Max*
Loan amount §	Loan amount (¥)	91373.26	21001.91	3000.00	100000.00
Rest amount	Remaining loan amount outstanding (¥)	1271.83	10504.09	0.00	100000.00
Rest interest	Remaining loan interest outstanding (¥)	58.20	598.01	0.00	11322.52
Firm age §	Number of years in operation	5.51	4.39	0.00	50.00
Space	Office space (m^2)	964.62	2755.73	0.00	33300.00
Employees §	Number of employees	10.60	19.56	0.00	500.00
Rate-of-income	The ratio of monthly repayment to income (%)	17.19	22.90	0.00	94.33
Family members §	Number of borrower's family members	3.33	1.03	0.00	11.00
Family workforce §	Number of workforces available in borrower's family	2.22	0.81	0.00	7.00
Procurement §	Annual procurement frequency	3.20	6.36	0.00	23.00
Loan rate §	Loan interest rates (%)	15.49	0.46	3.00	15.84
Loan term	Loan term (months)	3.07	0.46	1.00	9.00
Owner age §	The age of the business owner	38.80	8.24	19.00	63.00
Categorical		*Number of categories*		*Values*	
Business type §	Types of borrowers' business	5		{A, B, C, D, E}	
Repay type	Types of loan repayment	3		{1, 2, 3}	
Ethnicity §	Borrower's ethnicity	12		{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}	

Education	Borrower's educational background	10	{10, 20, 30, 40, 50, 60, 70, 80, 90, 100}
Marital status §	Borrower's marital status	7	{1, 2, 3, 4, 5, 6, 7}
Occupation	Borrower's occupation	9	{1, 2, 3, 4, 5, 6, 7, 8, 9}
Job position §	Borrower's job position	5	(A, B, C, D, E}
Job title	Borrower's job title	5	{0, 1, 2, 3, 4}
Degree §	Borrower's academic degree	5	{0, 1, 2, 3, 4}
Postcode	Postcode of the business location	13	{200000, 214400, 215200, 628200, …}*
Verified ID §	Borrower's citizenship category	6	{0, 1, 2, 3, 4, 5}
Industry	Industry of the borrower's business	12	{A, D, E, F, G, H, I, J, K, M, N, P}
Homeownership	Borrower's homeownership	7	{1, 2, 3, 4, 5, 6, 7}
Customer type §	Whether the borrower is a new or existing customer of the bank	2	{new_cust, old_cust}
Credit rating §	Borrower's credit rating	5	{1, 2, 3, 4, 5}
Loan purpose §	Types of loan purpose	8	{1, 2, 3, 4, 5, 6, 7, 8}
License type §	Types of business license	3	{0, 1, 2}

Notes: We conduct a data desensitization process on the original structured data to protect private and business-sensitive information. * indicates that some infrequent categories are omitted, in the interest of space. § indicates the final features selected for our credit scoring models.

Table I2 Definitions of text variables and statistical information on their text lengths

Text type	Definition	Summary statistics for text length			
Borrower		*Mean*	*SD*	*Min*	*Max*
Repayment plan	Borrower's planned method or source for loan repayment	5.68	1.54	3.00	20.00
Main business	The firm's main business	10.36	5.35	2.00	29.00
Loan purpose	The primary reason behind the borrower's loan application	7.67	4.63	3.00	61.00
Self-description	Borrower's brief introduction about themselves	42.79	22.31	2.00	168.00
Operating history	Description of the borrower's firm operating history	95.05	41.78	3.00	189.00
Sourcing & purchase frequency	Description of the borrower's firm's sourcing and frequency of purchases	51.70	28.83	2.00	174.00
Loan officer					
Impression	Initial impression of the borrower	47.94	29.71	3.00	159.00
Repayment intention	Repayment intention analysis of borrower	56.06	34.80	3.00	166.00
Application review	Overall loan application assessment	95.35	18.78	3.00	125.00
Industry seasonality analysis	Analysis of off-peak and peak seasons in the firm's industry	27.98	19.85	2.00	165.00
Profit analysis	Analysis of the borrower's firm profits	63.53	42.95	2.00	199.00
Asset & liability verification	Verification of borrower's assets and liabilities	123.42	39.36	3.00	229.00

J. External Validation: Predictive Performance Using Alternative Metrics

Table J1 Predictive performance of the prompt-based approach (KS)

		Prompt (off-the-shelf)			Prompt (fine-tuned)		
MLM	Classifier	Structure	Text	Combine	Structure	Text	Combine
BERT	LR	0.160	0.431	0.371	0.160	0.410	0.378
	LGBM	0.348	0.330	0.449	0.348	0.404	0.409
	XGB	0.238	0.337	0.337	0.238	0.415	0.414
	RF	0.294	0.372	0.390	0.294	0.433	0.410
	SVM	0.092	0.278	0.372	0.092	0.338	0.370
	MLP	0.215	0.318	0.369	0.215	0.365	0.373
	Avg.	0.225	0.344	0.381	0.225	0.394	0.393
	Diff.	—	0.120***	0.157***	—	0.170***	0.168***
ERNIE	LR	0.160	0.410	0.364	0.160	0.420	0.408
	LGBM	0.348	0.453	0.401	0.348	0.385	0.324
	XGB	0.238	0.419	0.310	0.238	0.441	0.379
	RF	0.294	0.415	0.454	0.294	0.437	0.434
	SVM	0.092	0.372	0.469	0.092	0.445	0.447
	MLP	0.215	0.310	0.303	0.215	0.404	0.379
	Avg.	0.225	0.396	0.383	0.225	0.422	0.395
	Diff.	—	**0.172*****	**0.159*****	—	0.198***	0.171***
ALBERT	LR	0.160	0.373	0.389	0.160	0.444	0.440
	LGBM	0.348	0.328	0.418	0.348	0.457	0.408
	XGB	0.238	0.467	0.312	0.238	0.510	0.475
	RF	0.294	0.443	0.440	0.294	0.481	0.448
	SVM	0.092	0.287	0.415	0.092	0.428	0.443
	MLP	0.215	0.225	0.313	0.215	0.417	0.409
	Avg.	0.225	0.354	0.381	0.225	0.456	0.437
	Diff.	—	0.129***	0.157***	—	**0.232*****	**0.213*****
RoBERTa	LR	0.160	0.297	0.301	0.160	0.429	0.400
	LGBM	0.348	0.288	0.248	0.348	0.381	0.402
	XGB	0.238	0.331	0.301	0.238	0.398	0.394
	RF	0.294	0.348	0.300	0.294	0.432	0.421
	SVM	0.092	0.350	0.355	0.092	0.345	0.388
	MLP	0.215	0.330	0.309	0.215	0.363	0.353
	Avg.	0.225	0.324	0.302	0.225	0.391	0.393
	Diff.	—	0.100***	0.078***	—	0.167***	0.168***

Notes: This table reports the average KS statistic for predicting the default risk of borrowers under the off-the-shelf and fine-tuned settings of the proposed prompt-based approach. Results are presented for four MLMs and six classifiers. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average KS statistic across classifiers. Diff. reports the difference in average KS statistic relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table J2 Predictive performance of the prompt-based approach (H-measure)

		Prompt (off-the-shelf)			Prompt (fine-tuned)		
MLM	Classifier	Structure	Text	Combine	Structure	Text	Combine
BERT	LR	0.062	0.283	0.280	0.062	0.266	0.252
	LGBM	0.032	0.271	0.243	0.032	0.251	0.253
	XGB	0.038	0.244	0.202	0.038	0.261	0.257
	RF	0.041	0.253	0.248	0.041	0.263	0.265
	SVM	0.011	0.274	0.279	0.011	0.248	0.253
	MLP	0.026	0.111	0.111	0.026	0.178	0.215
	Avg.	0.035	0.239	0.227	0.035	0.245	0.249
	Diff.	—	**0.204** ***	**0.192** ***	—	0.210 ***	0.214 ***
ERNIE	LR	0.062	0.227	0.227	0.062	0.246	0.245
	LGBM	0.032	0.237	0.227	0.032	0.252	0.248
	XGB	0.038	0.235	0.229	0.038	0.257	0.256
	RF	0.041	0.238	0.240	0.041	0.252	0.250
	SVM	0.011	0.240	0.238	0.011	0.220	0.229
	MLP	0.026	0.147	0.109	0.026	0.222	0.215
	Avg.	0.035	0.221	0.212	0.035	0.242	0.241
	Diff.	—	0.186 ***	0.177 ***	—	0.207 ***	0.206 ***
ALBERT	LR	0.062	0.224	0.225	0.062	0.283	0.270
	LGBM	0.032	0.223	0.229	0.032	0.268	0.258
	XGB	0.038	0.218	0.224	0.038	0.285	0.268
	RF	0.041	0.238	0.243	0.041	0.291	0.289
	SVM	0.011	0.221	0.233	0.011	0.255	0.253
	MLP	0.026	0.078	0.072	0.026	0.193	0.213
	Avg.	0.035	0.200	0.204	0.035	0.262	0.259
	Diff.	—	0.165 ***	0.169 ***	—	**0.228** ***	**0.224** ***
RoBERTa	LR	0.062	0.202	0.191	0.062	0.260	0.253
	LGBM	0.032	0.151	0.194	0.032	0.277	0.255
	XGB	0.038	0.227	0.223	0.038	0.263	0.250
	RF	0.041	0.208	0.205	0.041	0.259	0.254
	SVM	0.011	0.120	0.141	0.011	0.221	0.236
	MLP	0.026	0.169	0.196	0.026	0.206	0.196
	Avg.	0.035	0.180	0.192	0.035	0.248	0.241
	Diff.	—	0.145 ***	0.157 ***	—	0.213 ***	0.206 ***

Notes: This table reports the average H-measure for predicting the default risk of borrowers under the off-the-shelf and fine-tuned settings of the proposed prompt-based approach. Results are presented for four MLMs and six classifiers. Structure, Text, and Combine denote models using structured variables only, text-derived features only, and both feature sets jointly. Avg. reports the average H-measure across classifiers. Diff. reports the difference in average H-measure relative to the corresponding structured-only benchmark, and we assess its statistical significance using paired t-tests. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.