Writing Quality and Soft Information in the GenAI Age: Evidence from Online Credit Markets*

Lin William Cong Yanhong Guo Xin Zhao Wenjun Zhou

December 2024

Abstract

We document systematically how the various dimensions of writing quality matter for different borrower groups on a dominant online credit platform, and would be reshaped by the use of large language models (LLMs). Utilizing a Writing Quality Index (WQI) that effectively predicts online credit outcomes, we find that LLMs (e.g., ChatGPT) significantly enhance the writing and perceived quality of loan applications, especially for high-score borrowers who treat good credit and quality writing as substitutes in their applications, whereas these are historically complements. We further build proprietary BERT-based multi-modal models of lenders' decision-making that accommodate lenders' constraints, and introduce "deep" Heckman correction for sample selection bias. We demonstrate that ChatGPT adoption decreases soft information conveyed and increases credit misallocation, likely due to the convergence in writing; when lenders respond to borrowers' LLM adoption, they rely more on hard information, mitigating the misallocation. We leverage generative modeling to characterize the counterfactual equilibrium with endogenous borrowers' LLM adoption and lenders' responses. Our findings provide insights into the evolving role of soft information and potential impacts of GenAI in lending.

Keywords: Generative AI, Heckman Correction, Lending, LLMs, Misallocation, Multi-Modal.

*Cong is at Cornell University SC Johnson College of Business (Johnson) and NBER. Guo and Zhao are at Dalian University of Technology School of Economics and Management. Zhou is at Haslam College of Business, The University of Tennessee, Knoxville. We thank Simon Xu, Allen Hu, and conference and seminar participants at CUHK Shenzhen, Fudan University, the 2024 Generative AI for Finance Conference (Montréal), and Lingnan University for constructive comments. Send correspondence to Cong at will.cong@cornell.edu.

1 Introduction

It pays to write well (greater readability, persuasiveness, clarity, etc.) in corporate disclosure, loan applications, and other financial settings (Iyer et al., 2016; Hwang and Kim, 2017; Gao et al., 2023). However, a systematic understanding of the various dimensions of "writing well" and how they link to the soft information conveyed is missing in the literature. The recent breakthroughs in generative AI (GenAI), especially large language models (LLMs), not only illuminate these issues, but are also believed to dramatically change how people write in practice. We investigate how LLMs affect writing quality in financial settings, how borrowers adopt and lenders respond, and what all these imply for the transmission of soft information and allocative efficiency in financial intermediation.[1]

To these ends, we use (i) historical data from Prosper, the first marketplace lending platform in the United States, and still the most prominent one, (ii) loan applications aided by ChatGPT, a representative LLM, and (iii) proprietary multimodal lender decision AI models. In the process, we introduce, for the first time in finance, a Heckman-like correction for deep-learning models and an empirical implementation of a new procedure for analyzing counterfactual equilibrium (Cong, 2024), both of which constitute methodological contributions.

We demonstrate based on our data sample that LLM adoption significantly alters credit allocation dynamics: while ChatGPT could have increased approval rates by up to 9.35 percentage points for high-credit borrowers, it may also raises false positive rates (FPR) by 5.62pp under small-scale LLM adoption. When lenders adapt to widespread AI usage, misallocation gets mitigated, and whether allocative efficiency remains significantly below pre-GenAI levels depends on the lenders' constraints (e.g., lack of access to powerful models, or behavioral biases). As borrowers' adoption of LLMs increases and lenders respond to the reduction in soft information by focusing more on hard information, lenders' return on investment (ROI) dropped from 18.63% (no GenAI) to 15.68% (small-scale adoption), recovering marginally to 16.78% with large-scale adoption. These findings underscore the dual-edged impact of LLMs in online lending markets.

Specifically, following prior studies on online credit markets (e.g., Netzer et al., 2019; Gao

[1] On the one hand, better writing can more effectively convey soft information, which is often communicated through unstructured channels (texts, conversations, etc.). On the other hand, better writing could also be more persuasive for getting the recipient act in favor of the sender even with the same informational content.

et al., 2023), we use rich loan-level lending and performance information from the Prosper platform, which predate the introduction of any LLMs, allowing us to distinguish human-produced and LLM-generated writing. To generate counterfactual loan applications when borrowers adopt LLMs, we rewrite historical loan descriptions using ChatGPT, in a way similar to how borrowers would have done. Following OpenAI's tutorial, we design a prompt consisting of the role, task, and rule settings. The rule setting encompasses specific instructions on how the borrower wishes to change the original loan description, including tone, style, and formatting. Our main results are based on a prompt that aims to improve writing quality and loan fundability without advanced knowledge from research about what specifically would improve loan fundability—what borrowers naturally would ask—and are robust to rerunning randomized generations of each prompt version multiple times and employing alternative prompt versions.

We first verify how writing quality is associated with lending outcomes historically. We identify the major categories of textual features such as "Spelling Errors" and "Tone" based on prior linguistic and finance studies, and document that "Text Length," "Formality," and "Readability" to be most positively associated with loans being funded whereas "Grammatical Errors" is significantly associated with rejections. However, the associations can be heterogeneous across individuals with different credit scores: more spelling errors by medium-score borrowers and less formal writing by low-credit borrowers are associated with better funding outcomes. Taking ChatGPT rewrites as benchmarks for high-quality writing, we further propose a Writing Quality Index (WQI) based on the cosine similarity between the text embeddings of the original and the rewritten texts. We find that WQI is higher for low-score borrowers and predicts loan outcomes beyond any individual textual feature, especially among high-score borrowers.

To analyze the impacts of LLMs on writing quality, we first employ human-based evaluations to confirm that ChatGPT indeed improves the perceived writing quality and subjects' willingness to lend significantly. Concerning textual features, rewrites are more detailed, formal, and contain fewer grammatical and spelling errors. Because the difference between the original description and the rewrite represents ChatGPT's impact on the credit market at the loan level, we focus in subsequent analyses on WQI, which contains information beyond what the textual features capture and measures the improvement of writing quality coming from LLMs. By improving the writing quality, ChatGPT seems to improve a high-credit-score borrower's chances of securing funding.

That said, LLMs may alter the soft information that lenders rely on to make lending decisions, especially for the online credit market where loan descriptions constitute the primary channel soft information is conveyed to the lenders.[2] To quantify soft information in text, we assess lenders' decision quality when using both structured and textual features versus using structured features alone. Decision quality is measured by classification accuracy[3] or return on investment (ROI). We employ advanced language models, including Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), and incorporate a multimodal Transformer for a unified lender model, which also accommodates lenders' constraints in reduced form.

The results reveal a significant decline in the predictive value of texts. Using cosine similarity to measure within-group homogeneity, we find that loan descriptions of low-score borrowers are more similar to each other, and those of high-score borrowers are less so, despite that homogeneity increases in both groups after ChatGPT rewriting. In other words, ChatGPT reduces variability in writing, which may explain the diminished soft information. This convergence makes it harder to discern the true loan quality. Moreover, this reduction in the effective usage of soft information turns out to result in greater credit misallocation.

To mitigate the potential sample selection bias arising from solely relying on funded loans (for which loan quality and repayment outcomes are observed), we devise a "deep Heckman correction" for our multimodal transformer modal in the spirit of Heckman (1979) and building from the specification of Kahng et al. (2023). Compared with the pre-GenAI scenario, if borrowers have adopted LLMs while lenders lend as before, there would be a significant increase in credit misallocation, as measured by the false positive rate (FPR) and false negative rate (FNR) when juxtaposing the funding decision and the eventual loan outcome (repayment or default).

The above analyses are under a "partial equilibrium" view because we have not accounted for the fact that lenders can adapt their decisions in a marketplace anticipating LLM usage. We are

[2] Iyer et al. (2016) find that lenders can effectively gauge borrower quality using soft information in loan descriptions, such as word choice, writing clarity, and tone. This allows lenders to outperform traditional credit scores in predicting defaults. Similarly, Netzer et al. (2019) report that peer-to-peer lending borrowers' text descriptions offer insights into past behavior and future intentions, enhancing loan outcome predictions.

[3] Following Netzer et al. (2019), we use the area under the curve (AUC) as a measure of decision model accuracy. In particular, when a model is trained to predicting whether a loan would become default or not, we create a receiver-operating characteristic (ROC) curve based on the true positive rates and false positive rates at various cutoff values of the model's prediction. The AUC values range between 0 and 1, with 1 indicating perfect prediction.

thus interested in expanding the scope of our research by considering lender response. Specifically, with large-scale LLM adoption by the borrowers, lenders would update their decision model by training it with newer data, which includes descriptions produced by ChatGPT for some loans. In this counterfactual scenario, we retrain the multimodal lender model using generated loan descriptions to maximize some combination of conformity to historical lending practice and lenders' expected payoff out-of-sample. The misallocation is somewhat mitigated, because soft information is partially recovered and the lenders also rely more on hard information.

In a counterfactual equilibrium (as defined in Cong, 2024), borrowers' LLM adoption level solves a fixed point problem: Lenders take LLM adoption and usage by borrowers as given, and use both historical data and applications generated under LLM adoption to train their model. Borrowers take the lenders equilibrium decision as given, and make optimal decisions on LLM adoption and usage. The aggregation of their adoption decisions in turn pins down the equilibrium adoption level. Guided by historical patterns, we conjecture that borrower adoption is correlated with credit rating, search over the possible adoption levels, and verify whether each level is an equilibrium. With small adoption costs of LLM, we find that a full adoption of LLM is the unique equilibrium. We discuss soft information and misallocation in this counterfactual equilibrium.

Our study is the earliest to systematically verify the effectiveness of LLMs in improving the writing quality and fundability of loan applications. We show that although LLMs can help enhance fundability, it leads to convergence in writing, reduces soft information available to help lender decisions, and leads to greater credit misallocation. We are also the first empirical study to analyze the long-run counterfactual equilibria under endogenous GenAI adoption, while existing works focus on the impact of LLMs as decision makers (Bowen et al., 2024) without considering responses by counterparties and equilibrium implications. While low-credit borrowers lose their historical advantage in conveying soft information, they still benefit from ChatGPT's improvement of writing quality. In such a unique equilibrium, borrower loan descriptions converge in writing and lenders rely more on hard information.

We also make several methodological contributions to the literature. First, this is one of the few studies that build multimodal Transformer models in economics, and is the first one for the online credit market (Gu and Budhkar, 2021).[4] The multi-modal modeling of lending decisions allow

[4] Cong et al. (2019) is the first application of Transformer in Finance; Chen et al. (2022) also recently apply

modality interactions as well as capture richer non-linearity and information content from various types of data, achieving greater out-of-sample performance. Second, we extend the Heckman correction to account for selection biases in deep learning models (Heckman, 1979). Third, our usage of generated/synthetic data to study the counterfactual equilibrium also illustrate empirically how in settings where historical data are not sufficient because the data-generating process could have changed, generative modeling comes to rescue to enable counterfactual equilibrium analysis in a data-driven manner, complementing reduced-form theory and structural models.

2 Institutional Background, Data, and Prompt Design

We exploit historical data from a prominent online credit market to (i) understand how writing quality matters (before the advent of LLMs) and how LLMs improve writing quality, (ii) extract the underlying distribution of borrower types which is useful when considering counterfactual credit allocation, and (iii) generate loan applications in order to characterize equilibrium outcomes when adoption and usage of LLMs in the online credit market are endogenous.[5]

2.1 Online Credit Data from Prosper Marketplace

Prosper Marketplace, established in 2005 and began operations in 2006, is recognized as the first marketplace lending platform in the U.S.[6] On the platform, despite different loan origination dates, the lending process is consistent across all loans and follows several stages, as illustrated in Figure 1. In the pre-funding stage, borrowers submit loan requests and open the listings for lenders to bid. They specify the amount requested, and the maximum interest rate they are willing to accept, the length of the bidding period (up to two weeks), the minimum fulfillment rate (e.g., 80%) for the listing to become a loan, and structured and unstructured information about their credit profiles. Then in the bidding stage (Stage 1), lenders specify the bidding amount and the minimum interest rate they are willing to lend for this amount, all within the bidding period as specified in

transformer-based language models to asset pricing.

[5] The implicit assumption is that the distribution of borrower quality is stationary; should this fail to hold, resampling from the historical distribution guided by stipulated evolution of borrower quality can be used.

[6] https://www.prosper.com/about

the listing.[7] Once the total requested amount is fulfilled within the bidding period, the listing becomes a loan; otherwise, the loan application is considered rejected. Stage 2 starts when a loan is funded and lasts to its maturity. We consider a loan as having "defaulted" if its status in Prosper is "Charge-off," "Defaulted (Bankruptcy)," or "Defaulted (Delinquency)." Conversely, we categorize a loan as "paid" when it is marked as "Paid in full," "Settled in full," or simply "Paid." Since the text component of their loan requests is written in Stage 1, it contains information potentially useful in predicting both the likelihood of funding and the likelihood of default.

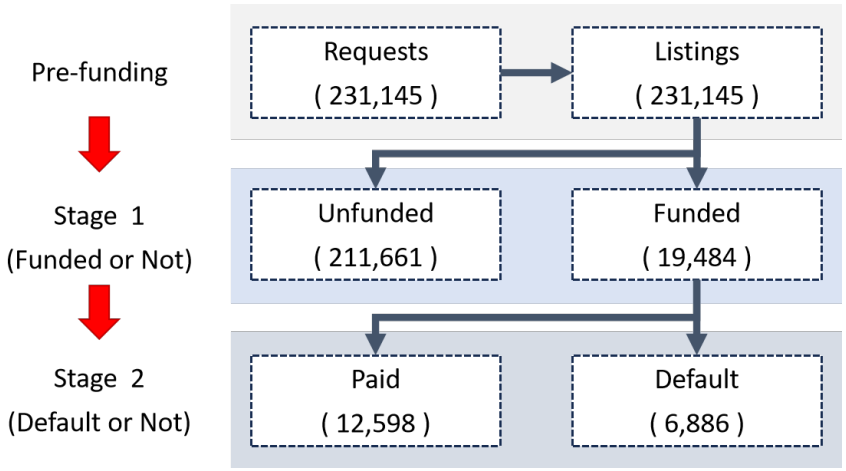


Figure 1: Life Cycle of Online Lending

We download all loan requests posted between April 2007 and October 2008, with a total of 231,145 listings. We focus on this early episode of Prosper for multiple reasons: (i) Data during this sample period was once publicly available, and is richer and more diverse than data from later episodes (Netzer et al., 2019) after regulations tightened. (ii) The period predate the widespread use of language models, so that we can isolate the effect of how LLMs improve writing. Among these listings, 19,484 (8.4%) were successfully funded. The longest loan term was 36 months, and all loans were mature at the time of data collection, with observable repayment outcomes. In summary, 6,886 (35.3%) of the funded loans defaulted, and the other 12,598 (64.7%) were repaid. For rejected listings, we extract an equal number of samples based on the credit distribution of funded loans, and refer to them as unfunded loans. Consequently, our subsequent analysis focuses on the dataset of 38,968 observations, comprising 19,484 funded loans and 19,484 unfunded loans.[8] This

[7] To effectively spread limited funds of individual lenders across different loans for diversification, the lending amount is usually much smaller than the requested amount, with a minimum of $25.

[8] We do not take the full unfunded sample because rewriting with ChatGPT is costly.

suffices for analyzing how writing mattered historically for online lending. However, when discussing lenders' responses and misallocataion, using this sample implies a corresponding belief of the underlying borrower quality. That said, our results are robust when using samples proportional to the historical credit-score distribution of the unfunded loans.

2.2 Structured Features

Borrowers' financial and demographic information is available to lenders mostly as structured data (i.e., hard or verifiable information Caldieraro et al., 2018). These variables include the loan request information, the borrower's credit profile, demographic information, and an overall credit rating by Prosper. Borrowers' credit profile include their debt-to-income ratios (*DebtToIncomeRatio*), the number of delinquencies (*DelinquenciesLast7Years*), and the number of inquires in the past six months (*InquiresLast6Months*). They are commonly used in traditional risk models (Guo et al., 2016; Fitzpatrick and Mues, 2021). Their demographics information include a categorical variable indicating the current employment status (*EmploymentStatus*), the tenure (in years) in the current employment status (*LengthStatusMonths*), and a dummy variable indicating whether the borrower is a homeowner (*IsBorrowerHomeowner*).

Loan request information include the amount requested in U.S. dollars (*AmountRequested*) and the lenders' interest rate (*LenderRate*).[9] Similar to existing studies, we include a dummy variable of whether the user has a head shot photo in their user profile (Li and Xie, 2020; Ho et al., 2022). In addition, the credit score of the borrower is an ordinal variable with seven possible levels, ranging from AA (best credit) to HR (high risk). Appendix A contains more details of these structured features, including definitions and descriptive statistics.

2.3 Textual Features

Writing quality is a well-studied topic in fields such as linguistics, education, and cognitive psychology. Building upon the existing literature, we consider the key dimensions of textual features, before proposing a composite metric for writing quality using advanced language models and the concept of semantic similarity.

[9] Since the data was collected in a period when interest rates were determined by lenders bidding on the loan requests, lenders' interest rate reflects the extent to which the request is appealing from the perspective of investors.

Text length. Texts should convey the right amount of information with concise and coherent language. Longer text may convey more detailed information, but it may reduce readability or even suggest deception (e.g., Ott et al., 2012; Hancock et al., 2007). Meanwhile, shorter profiles often signal dishonesty on online dating sites, as liars may avoid elaborating on certain subjects (Toma and Hancock, 2012). In online lending, the relationship between length and loan quality remains unclear (Toma and Hancock, 2012; Hancock et al., 2007). That said, the free-text box for borrowers to enter a description at Prosper.com has a limit of up to 4,000 characters (Ryan et al., 2007), borrowers are prevented from writing longer texts and often elect to write much shorter than this limit. To measure text length in loan applications, existing studies use the word count (Netzer et al., 2019; Siering, 2023) or character count (Dorfleitner et al., 2016; Larrimore et al., 2011) of loan descriptions. We measure text length as the number of words in the loan description.

Spelling and grammatical errors. Linguistic conformity entails reducing spelling errors or grammar mistakes. Few spelling errors directly enhance text readability by contributing to its clarity and coherence (Ghose and Ipeirotis, 2010; Loughran and McDonald, 2014). Dorfleitner et al. (2016) reports that spelling mistakes often act as indicators of traits associated with lower income. Consequently, Harkness (2016) finds that in traditional lending channels, spelling errors are associated with a lower probability of loan approval. In peer-to-peer lending, Netzer et al. (2019) accounts for the number of spelling errors in loan application texts. Gao et al. (2023) considers not only spelling errors but also grammatical errors as components of readability, with the count of each divided by the number of words in the loan request and then standardized to zero mean, unit variance. We measure (the lack of) conformity as the percentage of spelling errors (the number of spelling errors divided by the number of words in the loan request) and the percentage of grammatical errors (the number of grammatical errors divided by the number of words in the loan request).

Formality. According to Tausczik and Pennebaker (2010), this metric is typically employed to assess complex language, educational level, and social status. Nyhus and Webley (2001) found that individuals with higher education often have higher income and financial literacy, reducing their likelihood of loan default compared to those with less education. However, the use of sophis-

ticated language can also be risky if readers perceive it to be artificially or frivolously complex, suggesting that language with higher complexity is likely used deceptively (Oppenheimer, 2006). Iyer et al. (2016) considers the extent of advanced vocabulary use by measuring the average number of characters in each word. Similarly to Netzer et al. (2019), we measure the sophistication by the percentage of words with six or more letters.

Readability. Although general writing quality can be assessed in various dimensions, recent research in finance has primarily focused on readability (Hwang and Kim, 2017; Cao et al., 2023; Gao et al., 2023), the ease with which a reader can understand the text. Lower readability in firms' public disclosures has been linked to a significant discount in the firms' value as it can reduce investor trust and quality perception (Hwang and Kim, 2017). Similarly, in the context of peer-to-peer lending, higher readability is linked to more investor enthusiasm and lower interest rates (Gao et al., 2023). Several readability metrics have been used in the finance literature, to capture the years of education required for a reader to understand the text at first glance (e.g., Kim et al., 2019; Cao et al., 2024; Gao et al., 2023). We follow Netzer et al. (2019) to use a measure typically referred to as the "simple measure of gobbledygook," or the SMOG index (McLaughlin, 1969).

Tone. Writing quality may also be measured by engagement and style. One of the engagement elements is related to the emotional beat, often measured as tone or sentiment (Loughran and McDonald, 2011). Tone is an important aspect in communicating financial information, given investors often respond favorably to positive tone or sentiment (Tetlock, 2007; Tetlock et al., 2008), and entrepreneurs who demonstrate higher positivity have a greater likelihood of success (Zott and Huy, 2007). In online credit markets, Gao et al. (2023) similarly finds that loan descriptions with positive sentiment are more likely to secure funding. We measure the tone of loan descriptions using sentiment scores provided by a pre-trained sentiment-BERT model, with higher values indicating more positive tones.

These features are computed using various Python packages. Specifically, for text length and formality, we use fundamental string operations in Python; for spelling error, we use the "spellchecker" package; for grammatical error, we utilize the "language_tool_python" package;

for readability, we apply the "textstat" package; and for tone, we employ a pre-trained sentiment-BERT model from HuggingFace. https://huggingface.co/michellejieli/emotion_text_classifier. We later use these textual features across the six dimensions to assess the writing of loan applications/descriptions.

Semantic similarity. Finally, we later define a composite metric capturing the similarity between two textual documents. Using state-of-the-art deep learning techniques, we can convert any text into a numeric vector, i.e., an embedding, and use the cosine coefficient to measure the semantic similarity between them. Specifically, given two numeric vectors $\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $\mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, the cosine similarity is calculated as:

$$CosineSimilarity(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|} \tag{1}$$

where $\mathbf{x}_1 \cdot \mathbf{x}_2 = \sum_{j=1}^{n} x_{1j}x_{2j}$ is the dot product of the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and $\|\mathbf{x}_i\| = \sqrt{\sum_{j=1}^{n} x_{ij}^2}$ is the norm of vector $\mathbf{x}_i$. This cosine similarity can measure the semantic similarity between any pair of texts in a holistic way. We employ the Siamese-BERT (SBERT) model to encode loan descriptions, which is more efficient and effective than BERT in capturing textual information for semantic similarity analysis that involves the cosine similarity between the encoded descriptions (Reimers and Gurevych, 2019).

2.4 GPT Prompt Design

To use LLMs to help with loan applications, borrowers would input prompts into them. As econometricians, we can mimic such behavior in order to generate loan application data for characterizing counterfactual equilibria where borrows endogenous adopt and use LLMs. OpenAI guidelines prescribe that an effective prompt should be specific and can include role setting, task setting, and rule setting. Role settings enable ChatGPT to comprehend the context, task setting defines the task for ChatGPT, and rule setting clarifies the rules it should follow. In the baseline prompt design, we ask ChatGPT to improve the writing and general fundability of the loan description.

A borrower using LLMs to write loan applications would want to get help from a professional. We thus describe the writer role as: "*You are a practitioner who has been in the lending business*

for many years and are well versed in matters of applying for loans." The task component communicates the action needed. Since we want the ChatGPT to conduct a rewrite, we describe the task of ChatGPT as: "*Please help me rewrite the following narrative in a loan application*." Our template also includes a section of quoted text to capture the original text for ChatGPT to rewrite. Figure 2 shows an example, in which the rule setting is: *Please improve the clarity and readability of the text, and emphasize anything that helpful for lenders' decision-making. Please make the reason compelling, and make myself sound trustworthy, so lenders willing to fund it.*

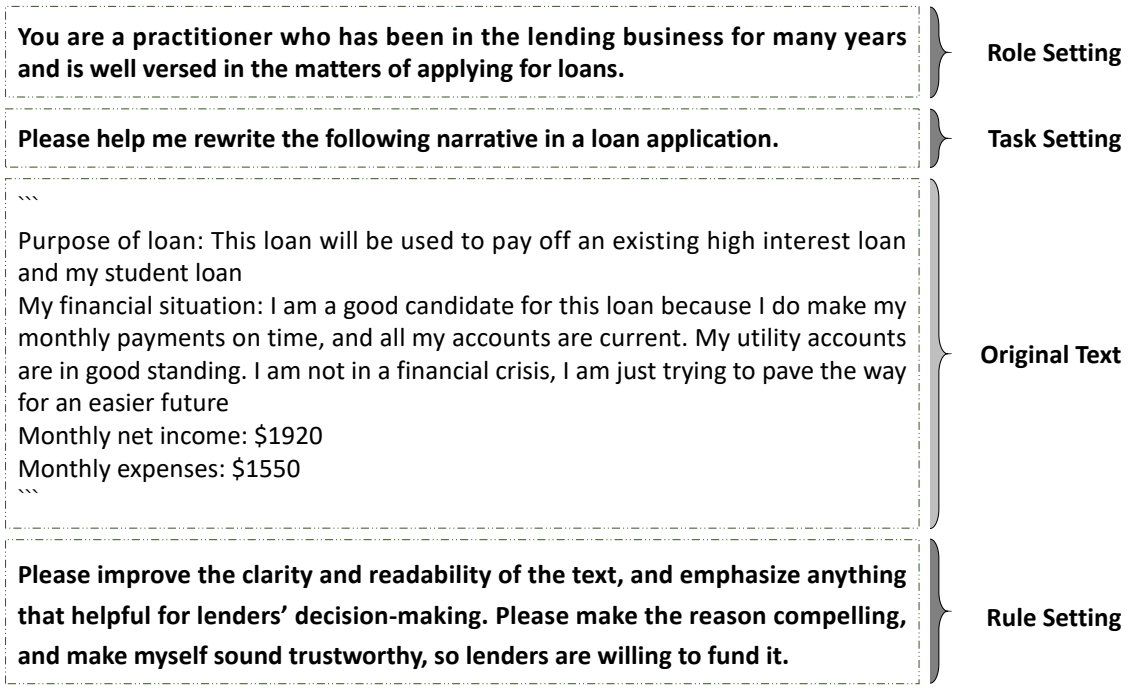


Figure 2: An Example ChatGPT Prompt.

Guided by this prompt template, the loan descriptions can be generated based on the input. For example, for the original text in Figure 2, one re-written version is the following: "*I am seeking this loan to consolidate my existing high interest loan and my student loan. To illustrate my financial situation, I am consistently making all of my monthly payments on time, and all of my accounts are up to date. My utility accounts are also in good standing. I would like to clarify that I am not experiencing any financial difficulties; rather, I am considering this loan as a means to create a smoother future. As for my income, I earn a monthly net income of $1920. With regards to my expenses, they total $1550 per month.*" It is evident that there are significant differences in the

description of financial status (monthly net income) before and after the rewrite, with the rewritten ones better in writing quality and credibility. Appendix B offers more examples for illustration.

We used OpenAI's API to the GPT4o-mini model for text generation, leverage its default parameters. In particular, we set the temperature parameter as 1, which strikes a balance between randomness and consistency.[10] In addition, to mitigate possible bias in framing the rule settings in the prompt, before generating rewrites for the loan descriptions, we first generated variations of the sentences in the rule setting. For each sentence, we generate 10 versions. Then, when putting together the prompt to generate rewrites for the loan descriptions, we randomly pick one of the versions of each sentence. In Appendix C, we document the variations of rule-setting sentences and report results under alternative prompts, including those where borrowers possess advanced knowledge based on research about what improves fundability.

3 Textual Features, Writing Quality, and Online Lending

We first conduct a human perception test to see if the writing quality and fundability of loan descriptions are indeed improved by ChatGPT, before we examine the specific textual features of the original loan descriptions and how LLMs alter each of them. We then propose a writing quality index (WQI) as a composite measure of writing quality. We demonstrate its relationship with text features from linguistic studies, and its effectiveness in capturing heterogeneous credit quality and funding status.

3.1 LLM-Generated Loan Descriptions

Human-verified improvements by ChatGPT. Recent literature has shown that LLMs such as ChatGPT have surpassed humans in producing high-quality writing (Herbold et al., 2023; Deveci et al., 2023). We first validate this claim in the context of lending using an online experiment involving human raters. We ask human raters recruited on Prolific platform to either rate the perceived writing quality or their willingness to lend based on the text, all on the scale of 1-10. Appendix D contains more details about the settings.

[10] The temperature parameter ranges from 0 to 2. Larger values correspond to a more random answer, whereas smaller values correspond to a more consistent answer. See https://platform.openai.com/docs/

Table 1: Human Rater Results

	DV: Writing Quality	DV: Willingness to Lend
Rewrite	**2.342*****	**1.450*****
	(0.223)	(0.249)
Constant	5.283***	4.500***
	(0.158)	(0.176)
Listing Fixed Effect	Yes	Yes
Rater Fixed Effect	Yes	Yes
Observations	240	240
Pseudo R^2	73%	72%

Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses.

From Table 1, we see that human-perceived writing quality of loan applications improves when rewritten by ChatGPT. Similarly, loan descriptions generated by ChatGPT through rewriting the original loans received higher funding interests. All these are consistent with our expectations.

LLM-generated changes in textual feature. Table 2 summarizes the textual features of loan descriptions before and after ChatGPT rewriting. We note that ChatGPT significantly modifies all textual features, making the rewritten texts longer, with fewer errors, and with a slightly more positive tone. On average, following a ChatGPT rewriting, the number of words in loan descriptions increases by 38.8% on average, the percentage of spelling and grammatical errors decreases by 74.3% and 56.3%, respectively, and the tone increases by 1.2%. In addition, formality increases by 53.6% and readability increases by 50.7%, on average.

Table 2: Textual Features Before and After Rewrite

	Original		Rewrite		Change	
	Mean	(SD)	Mean	(SD)	DIF[a]	(%)
Text Length	202.5	(132.2)	281.1	(99.44)	78.6***	(+38.8%)
Spelling Errors	0.035	(0.021)	0.009	(0.007)	-0.026***	(-74.3%)
Grammatical Errors	0.064	(0.041)	0.028	(0.023)	-0.036***	(-56.3%)
Formality	0.239	(0.058)	0.367	(0.035)	0.128***	(+53.6%)
Readability	10.45	(3.775)	15.75	(1.684)	5.30***	(+50.7%)
Tone	0.745	(0.157)	0.754	(0.144)	0.009***	(+ 1.2%)

[a] Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

`api-reference/chat/create` for a list of all parameters and their default values.

3.2 Writing Quality Index: A Composite Measure

The change in each textual feature reflects the influence of LLM rewriting on the text in a specific dimension as discussed in the literature on writing quality. We propose a composite metric, the Writing Quality Index (WQI), to holistically capture the gap between an original description and an "ideal" one rewritten by ChatGPT.

We define WQI as:

$$WQI(\mathbf{x}) = CosineSimilariy(emb(\mathbf{x}), emb(\mathbf{x}')), \tag{2}$$

where $\mathbf{x}$ represents the original text, $\mathbf{x}'$ represents its rewrite, and $emb()$ represents the conversion from raw text to its numeric embedding. The assumption is that by default, GPT writes well and the prompt produces an improved version, as supported in Section 3.1.

WQI: interpretation and link to textual features. WQI measures the similarity of loan descriptions before and after ChatGPT rewrites. When we consider the ChatGPT-rewritten version as the optimal version, it assesses the writing quality of the original text by measuring how close it is to the optimal benchmark (i.e., ChatGPT rewrite). Higher values represent higher writing quality of the original description, as the original version is more similar to the rewrite version, and lower values indicate lower writing quality and great potential impact on writing quality using ChatGPT.

Table 3 reveals that the textual features explaining 11.0% of WQI in the full model (see R^2 in Model 7). This suggests that WQI captures some textual feature-related information and beyond.

3.3 Credit Score, WQI, and Online Lending

Using WQI as an effective measure of writing quality, we find that historically, borrowers treat their credit score and the writing quality of loan applications as substitutes, whereas they can be complements for getting the loans funded, especially for borrowers with high credit scores.

Application heterogeneity and WQI. How is WQI related to individual credit scores and the funding status? We conduct a two-way ANOVA analysis that attributes the WQI by two factors: credit scores (factor A, with 3 levels—high, medium, low) and funding status (factor B, with 2

Table 3: Relationship Between Textual Features and WQI

	DV: WQI						
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Text Length	0.135***						0.114***
	(0.003)						(0.003)
Spelling Errors		-0.804***					-0.393***
		(0.024)					(0.028)
Grammatical Errors			-0.195***				0.082***
			(0.012)				(0.014)
Formality				-0.203***			-0.050***
				(0.009)			(0.010)
Readability					0.090***		0.060***
					(0.003)		(0.003)
Tone						-0.010***	-0.025***
						(0.003)	(0.003)
Constant	0.737***	0.800***	0.785***	0.821***	0.738***	0.780***	0.758***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.003)
Observations	38,968	38,968	38,968	38,968	38,968	38,968	38,968
R-squared	0.084	0.043	0.010	0.021	0.024	0.000	0.110

Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses.

levels—funded, unfunded). In particular, consider a linear model as follows:

$$WQI_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{3}$$

where WQI_{ijk} represents the k-th observed value of WQI for a sample under the i-th level of factor A and the j-th level of factor B, μ represents the overall mean, α_i represents the effect of the i-th level of factor A, reflecting the deviation of this level from the overall mean, β_j represents the effect of the j-th level of factor B, reflecting the deviation of this level from the overall mean, γ_{ij} represents the interaction effect between the i-th level of factor A and the j-th level of factor B, and ϵ_{ijk} is the random error term (assumed to be independently and identically distributed, following a normal distribution with a mean of 0 and variance σ^2).

We find that credit scores, funding outcomes, and their interactions are significantly associated with WQI. According to Figure 3, the average WQI is decreasing for low-, medium-, and high-score borrowers for both funded (in dark red) and unfunded (in dark blue) listings. Take the funded loans for instance, the WQI for low-, medium-, and high-score borrowers are expected to be 0.780, 0.7710, 0.760, respectively. Among the unfunded loans, the WQI shows the same

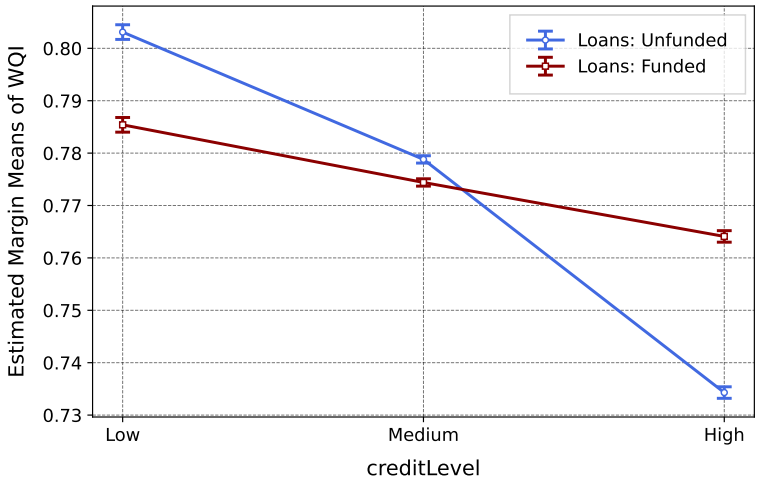


Figure 3: WQI by Credit Score Category and Funding Outcomes.

ordering, with a mean WQI of 0.790, 0.779, 0.726 for low-, medium-, and high-score borrowers, respectively. Low-score borrowers produced significantly higher WQI, which indicates that low-score borrowers' loan descriptions exhibit better writing quality. Low-score borrowers, due to their poor credit profiles (i.e., hard information), often face difficulties when applying for conventional loans. When they apply for loans in online marketplace, they have the opportunity of disclosing helpful soft information via the textual explanations. Unlike credit scores, loan descriptions are entirely crafted by the borrowers and serve as a communication tool with lenders. Consequently, these low-score borrowers historically invested more effort in crafting informative and persuasive narratives to enhance their chances of obtaining funding. In contrast, high-score borrowers are less concerned about explanations given their good credit profiles, and thus provide writing of lower quality.

The lines in Figure 3 are not parallel. For low- and medium-credit borrowers, the WQI is slightly higher for unfunded loans (0.803 and 0.779) than funded loans (0.785 and 0.774), whereas this pattern is reversed for high-score borrowers—the WQI for unfunded loans (0.734) is much lower than funded loans (0.764). This indicates that good credit profiles did compensate for the lack of good writing to an extent. High-score borrowers need to write more poorly to not get funded.

WQI and fundability. To check the relationship between writing quality and getting funded, we employ a logistic regression model and select previously mentioned structured features as control

variables to conduct the significance test. The model may be specified as follows:

$$\begin{aligned}\text{logit}(Outcome_i) =& \beta_0 + \beta_1 \times IsHighCredit_i + \beta_2 \times IsLowCredit_i + \beta_3 \times WQI_i \\ &+ \beta_4 \times WQI_i \times IsHighCredit_i \\ &+ \beta_5 \times WQI_i \times IsLowCredit_i + \boldsymbol{\gamma} \cdot \boldsymbol{c}_i + \varepsilon_i, \end{aligned} \tag{4}$$

where $Outcome_i$ represents whether loan i is funded, $\boldsymbol{c}_i$ represents the vector of control variables, $\boldsymbol{\gamma}$ is the corresponding coefficient vector, and ε_i represents the random error. With this model specification, the estimated effect of WQI would be $\hat{\beta}_3$ for medium credit borrowers, $\hat{\beta}_3 + \hat{\beta}_4$ for high-score borrowers, and $\hat{\beta}_3 + \hat{\beta}_5$ for low-score borrowers.[11]

Table 4: Regression Analysis of Writing Quality and Funding Outcome

	DV: Funding (Funded: 1, Unfunded: 0)			
	Model 1	Model 2	Model 3	Model 4
WQI	**0.916*****	-0.352	-0.289	**-1.789*****
	(0.146)	(0.194)	(0.154)	(0.203)
WQI × High-Score Borrower		**5.256*****		**5.709*****
		(0.350)		(0.354)
WQI × Low-Score Borrower		**-1.682*****		**-1.414*****
		(0.387)		(0.394)
Text Length			**0.363*****	**0.370*****
			(0.015)	(0.015)
Spelling Errors			-0.000	0.017
			(0.014)	(0.014)
Grammatical Errors			**-0.112*****	**-0.129*****
			(0.014)	(0.014)
Formality			**0.053*****	**0.050*****
			(0.013)	(0.013)
Readability			**0.057*****	**0.048*****
			(0.012)	(0.012)
Tone			0.022	0.018
			(0.012)	(0.012)
Structured Features Controlled	Yes	Yes	Yes	Yes
Intercept	Yes	Yes	Yes	Yes
Observations	38,968	38,968	38,968	38,968
Pseudo R^2	15.3%	15.9%	17.3%	18.0%

Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses.

[11] Following Netzer et al. (2019), we classify credit scores into three groups: Low, Medium, and High. Low = low credit scores: E, HR; Medium = medium credit scores: B, C, D; High = high credit scores: AA, A.

Results are reported in Table 4. All models include an intercept and controls for the structured features. In Models 1 and 2, we use WQI alone to predict the funding outcome. In Models 3 and 4, we also control for textual features while using WQI to predict fundability. Models 2 and 4 includes interaction terms between the WQI and the credit scores to help us investigate the heterogeneous effect of WQI for borrowers with different credit scores. We can see that overall the WQI has a significant positive association with fundability (Model 1), indicating that loans with higher WQI are more likely to be funded. When considering heterogeneous effects by credit scores (Model 2), we can see that WQI is positively associated with fundability for high-score borrowers, and negatively associated for low-score borrowers.

When controlling for textual features, the effect of WQI becomes insignificant (Model 3). This seems to suggest that the writing quality captured in WQI is explained away by the textual features considered. After controlling for textual features, the heterogeneous effects remain statistically significant (Model 4). There is a negative association between WQI and fundability for medium ($\hat{\beta}_3 = -1.789$, $p < 0.001$) and low-score borrowers ($\hat{\beta}_3 + \hat{\beta}_5 = -3.203$, $p < 0.001$), with a significant difference between them. High-score borrowers still show a positive association between WQI and fundability ($\hat{\beta}_3 + \hat{\beta}_4 = 3.920$, $p < 0.001$). Figure 4 visualizes the relationship between WQI and fundability based on Model 4 in Table 4. This result seems to suggest that writing better (i.e., with a higher WQI) is beneficial for high-score borrowers, but counterproductive for medium and low-score borrowers.

As a comprehensive writing quality measure, we find that WQI better predicts fundability than any individual text feature. Details can be found in Appendix E. There, we also discuss how writing better in each of the textual features can help or hurt fundability of loan applications. Interestingly, having more spelling errors by medium-score borrowers or writting less formally by low-score borrowers are associated with a greater likelihood of being funded, *ceteris paribus*. We contribute to the literature by documenting such nuances. Robustness of our findings to altertive prompt designs can be found in Appendix F.

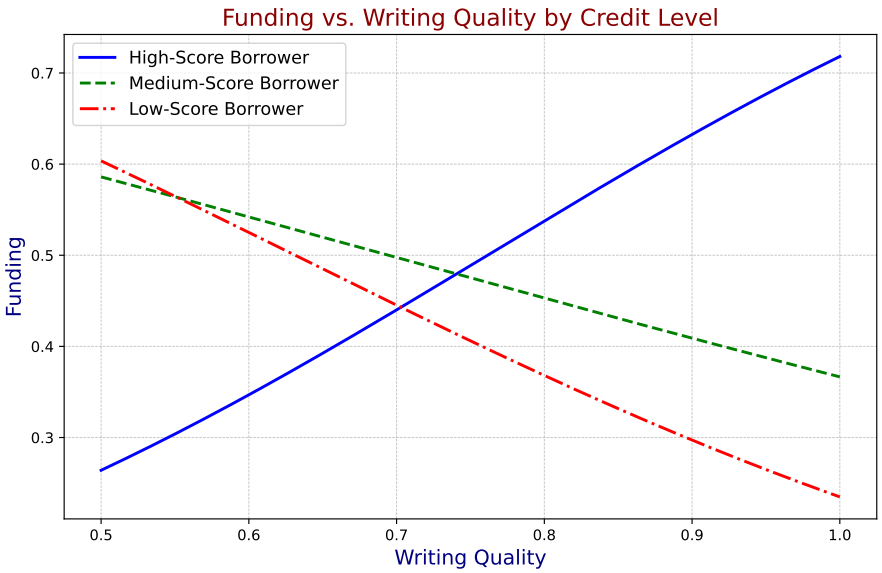


Figure 4: Heterogeneous Incremental Effect of WQI on Fundability.

4 Fundability and Soft Information in the GenAI Age

Lenders are the decision-makers when it comes to funding loan applications. Consequently, they determine the allocation of credit and how soft information is received and utilized. We next describe our construction of the lender model using a multimodal Transformer adapted with Heckman correction. This model helps us examine the increase in fundability and the loss of soft information when loan descriptions are generated by LLMs. Both the adaption of a multimodal setup and the Heckman correction for AI models in Finance constitute novel technical contributions.

4.1 A "Lender Model" with Heckman Correction

To analyze counterfactuals in the GenAI age, we need a model to mimic lenders' decisions. It is possible to reproduce lenders' decision process using a well trained machine model. This strategy has been adopted in existing studies (Bali et al., 2023; Cao et al., 2024). First, we recognize that lenders face both structured and textual data from the loan applications. To jointly model text and structured features as predictors, we adopt the multimodal Transformer (Gu and Budhkar, 2021) as the base architecture. More technical details can be found in Appendix G.

It is also important to note that our data is truncated: if a listing was not funded historically, we would not know whether the loan would have been repaid if funded. This poses a new challenge

when using models trained on a selected sample. Traditionally, Heckman selection can be used to adjust for sample selection bias (Heckman, 1979). Specifically, suppose that at the population level, there is a prediction function f and a selection model g, specified as

$$y_i = f(\mathbf{x}_i) + U_{1i}, \tag{5}$$

$$z_i = g(\mathbf{x}_i) + U_{2i}, \tag{6}$$

respectively. Here, $\mathbf{x}_i$ represents sample i's feature vector, y_i represents its observed outcome (e.g., repaid or not), and z_i represents whether the sample is selected (e.g., being funded). U_{1i} and U_{2i} are the error terms, which are correlated at the sample level (and are independent across samples) with a covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

In the online lending context, the prediction model f takes loan features $\mathbf{x}_i$ as input, and predicts its quality y_i, represented as the propensity of repayment. The selection model g would also take loan features $\mathbf{x}_i$ as input, and predicts z_i, the selection status indicating whether the loan would be selected for funding. Only those selected for funding would allow us to observe the repayment status. In this case, if we attempt to use all listing samples to fit Eqn. (5), there would be missing values for y_i for all unfunded ones. Alternatively, if we use only the funded loans to fit Eqn. (5), its model estimates may be subject to a selection bias.

In simple regressions, Heckman (1979) showed that by assuming that U_{1i} and U_{2i} follow a joint bivariate normal distribution, it is possible to derive a sample-level Inverse Mills Ratio (IMR) from the selection model g, which would then be included as an extra term in the prediction model f to account for the sample selection bias. In particular, this IMR is computed as $\lambda_i = \frac{\phi(Z_i)}{\Phi(-Z_i)}$, where ϕ and Φ are the density and distribution functions of a standard normal variable, and $Z_i = -\frac{g(\mathbf{x}_i)}{\sqrt{\sigma_2^2}}$ represents sample i's standardized propensity of selection. In this case, Eqn. (5) would become $y_i = f(\mathbf{x}_i) + \lambda_i + U_{1i}$, whose estimates obtained for model f using the funded loans only would be unbiased due to the IMR adjustments.

We extend this classic Heckman correction procedure beyond regression models to the deep learning context. In particular, instead of a linear specification, functions f and g are nonlinear (e.g., multimodal transformer). Using a deep learning model for selection, we obtain samples'

selection propensity from its output, and then follow the same logic as the classic Heckman model. To make this process more streamlined, we employ a custom, integrated learning algorithm inspired by a recent publication in the machine learning field (Kahng et al., 2023). More specifically, suppose that both the selection and prediction models need to be estimated, we propose a negative log-likelihood loss function, which simultaneously fits the models f and g, and provide estimates of the covariance matrix following a maximized likelihood estimation framework. The learning objective can be specified as:

$$\arg\min_{f,g,\Sigma} \sum_i \left[s_i \Lambda(f(x_i), g(x_i); y_i, s_i; \Sigma) - \{ s_i \log \Phi(g(x_i)) + (1 - s_i) \log \Phi(-g(x_i)) \} \right], \quad (7)$$

where f and g are two multimodal Transformer models (for prediction and selection, respectively), $\Lambda(f(x_i), g(x_i); y_i, s_i; \Sigma) = -\log p(y_i | s_i = 1, x_i)$ is the conditional negative log probability of y_i given $s_i = 1$. Note that in this formulation, the covariance structure Σ is also a learnable parameter and thus does not require users to set it.

We randomly partition the loans into training and test sets at the 4:1 ratio, develop the lender model based on the training set, and assess its performance on the test set. Using cross-validation to optimize hyperparameters, we arrive at a multimodal Transformer with "deep Heckman correction" as our lender model. The lender model can be trained on any data the lender faces to maximize a given objective of the lenders (e.g., ROI), or it can be trained with historical actions as the correct labels, as we discuss later.[12]

4.2 Modeling Lender Responses

In subsequent analysis, we consider three scenarios with different levels of lender response to borrowers' usage of LLMs and different loan application data.

- No GenAI: We train the lender model using the training set from the historical data, and then report its performance on the test set from the historical data.
- Small-Scale Adoption: Some borrowers adopt ChatGPT to help generate loan descriptions, but lenders do not respond because they perceive the adoption as too sporadic to have a

[12] Cong (2024) further discusses how behavioral and computational constraints can be incorporated.

material effect. Therefore, the lender model is trained on the historical training set, but we report its performance on the counterfactual test set which is essentially the historical test set with GPT rewrites.

- Large-Scale Adoption: Lenders believe there is large-scale or full adoption of LLMs by the borrowers. They therefore adjust the lending model by training one using a counterfactual training set, which is essentially the historical training set with GPT rewrites. We then report its performance on the counterfactual test set.

In sum, the lender model is trained using data from a period without GenAI tools (i.e., using the original text as the input), which applies to both "No GenAI" and "Small-Scale Adoption" scenarios. In reality, lenders can update their decision model, especially when GenAI adoption is widespread. With large-scale adoption of GenAI, the decision-making processes of lenders are likely to be adjusted. In other words, the predictive model for simulating lender decisions should be primarily trained on ChatGPT-generated data. Figure 5 illustrates the three cases of adoption.

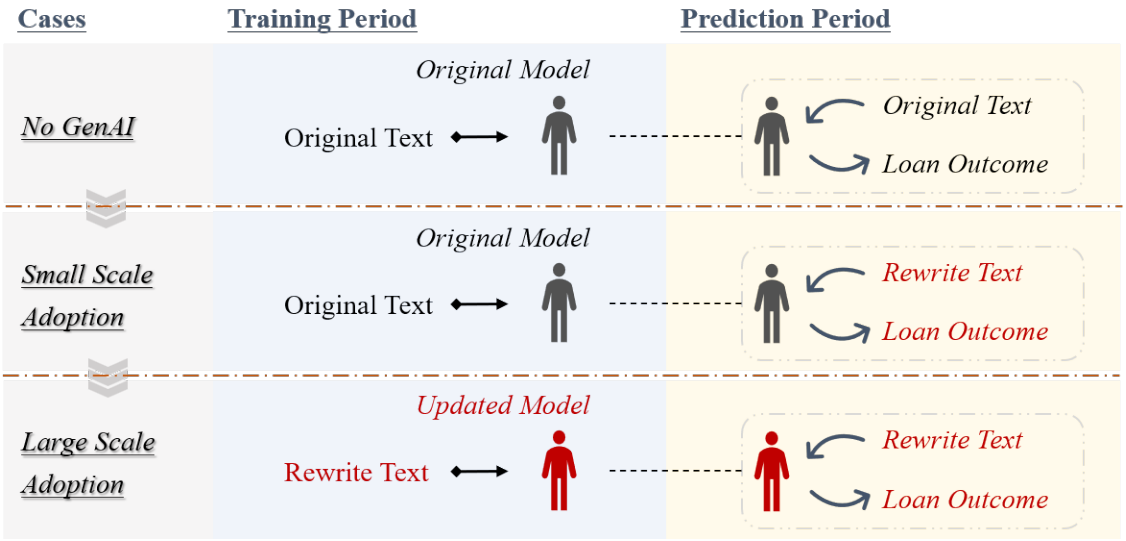
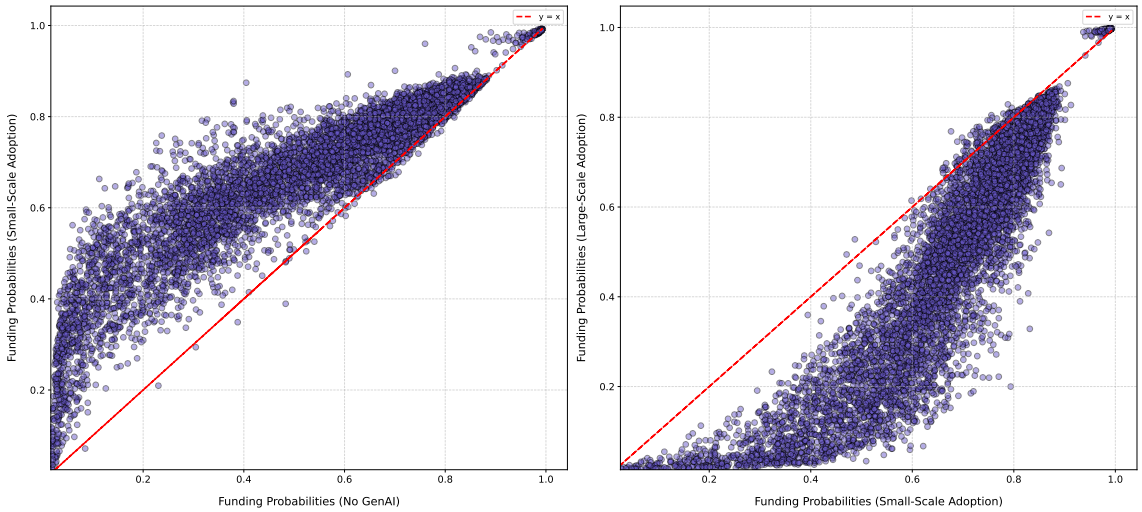


Figure 5: GenAI Adoption Scenarios.

Since most existing studies focus on the "No GenAI" scenario, the remainder of this section focuses on the "Small-Scale Adoption" scenario to highlight preliminary impact of GenAI on lending. We will discuss the "Large-Scale Adoption" scenario in Section 5.

4.3 LLM-Generated Descriptions and Loan Fundability

As GenAI helps produce high-quality writing, it may help loan applicants enhance fundability. To see if the probability of obtaining funding would increase after rewriting, we compute the predicted funding probability using the rewritten vs. the original text as the input. Results are shown in Figure 6, where each point represents a loan. The x-axis represents the predicted probability of funding using the original text, and the y-axis represents that using the text rewrites.



(a) Small-Scale Adoption (b) Large-Scale Adoption

Figure 6: Impact of ChatGPT on Fundability Under Different Levels Adoptions.

Figure 6(a) shows the comparison under the "Small-Scale Adoption" scenario. We find that using rewritten text as the input, our model generally produces a higher predicted probability of funding, as most data points are above the 45-degree line. Figure 6(b) shows the comparison under the "Large-Scale Adoption" scenario. We observe that most points falling below the 45-degree line, indicating that the predicted probability of funding has decreased. This indicates that as widespread adoption improves writing quality of loan applications for all borrowers, lenders adjust how they use the texts and funding probability drops back to be close to the pre-LLM era. Lenders' endogenous response matters.

Attribution test. Our next question then is which feature(s) led to greater improvements of fundability when loan description text is rewritten by ChatGPT? To investigate this question, we run an attribution test using the feature importance approach. Specifically, we first compute the increased

fundability as the difference between fundability using the rewrite text and that using the original text. Then, we fit an predictive model using the increased fundability as the target variable, textual features of the original text, change in textual features between the original and the rewrite text, WQI, and the structured features as predictors.[13] Based on this model, we evaluate the feature importance, and a list of the most important features are visualized in Figure 7. We can see that WQI is among the most important variables in predicting the improvement of fundability.

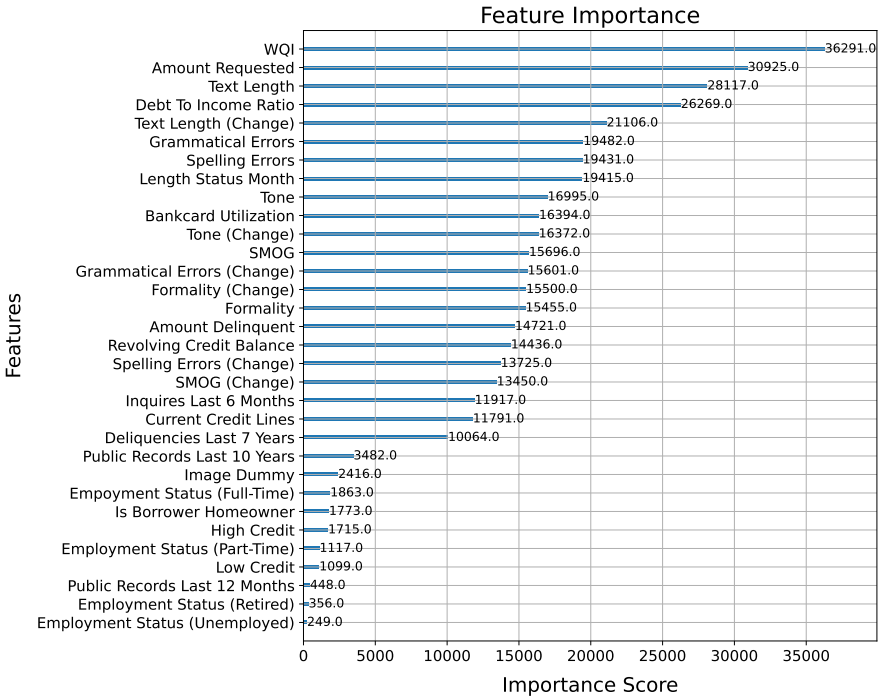


Figure 7: Feature Importance When Predicting Increased Fundability.

To picture the relationship between WQI and the improvement in fundability, we employ SHAP-based explanation (Lundberg and Lee, 2017) based on the joint predictive model. Specifically, we plot the partial dependence of the expected improvement of fundability by WQI in Figure 8. We can see that higher WQI values correspond to smaller fundability improvements. Since larger WQI values indicate that the original and the rewrite are more similar, the implication is there is less room for improvement in fundability by improving the text. In addition, for lower WQI values (approximately 0.78 or less), the corresponding expected fundability improvement stays somewhat the same (at around 9.35 percentage points). The same analysis based on the text features can be found in Appendix H.

[13] We used the XGBoost model (Chen and Guestrin, 2016) as it is widely known as one of the best performing models.

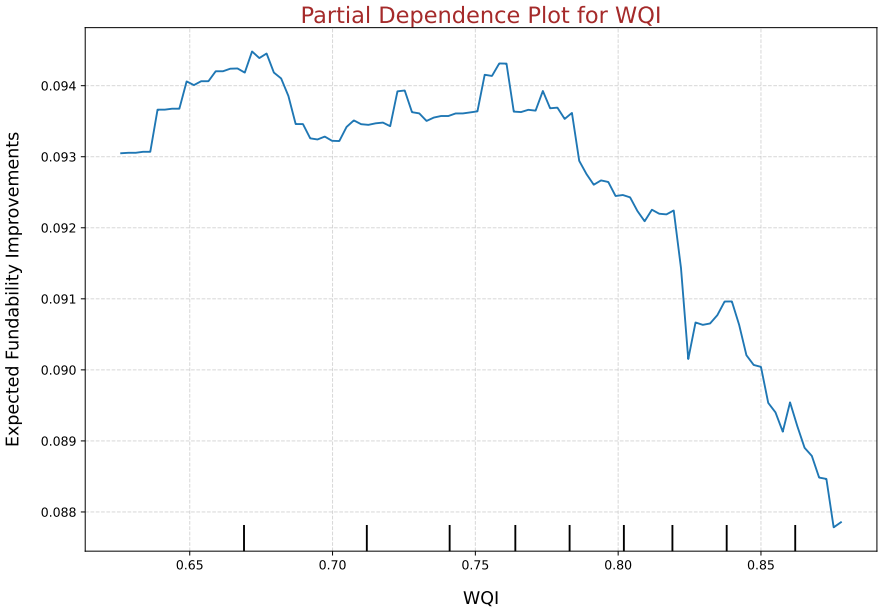


Figure 8: Partial Dependence Plot for WQI.

4.4 Soft Information in the GenAI Age

A utility-based measurement of soft information. We measure soft information as the additional predictability of textual data on loan quality/outcomes on top of other non-textual information. In Eqn. (8), the utility function U would be a performance metric such as AUC or ROI assessed on the test set using a model trained on the training set.

$$\text{soft-information} = U(\text{text}, \text{structured}) - U(\text{structured}) \tag{8}$$

Results are reported in Table 5. When using AUC as the measure of utility, the best performing model using structured features only achieves an AUC of 72.5%.[14] When both text and structured data are used, the AUC reaches 74.7% by our lender model in the "No GenAI" scenario, showing an increase of 2.2 percentage points (or a 2.9% relative increase), which may be considered as soft information quantified by AUC.

In addition to considering the risk-return trade-off between defaulted and repaid loans, we also examine the soft information as measured by ROI.[15] The best performing model when using

[14] See Appendix I for model benchmarking details.

[15] We adopt the method described in Netzer et al. (2019) to calculate ROI, which incorporates interest rates in its calculation, making it possible to account for the associated risk. Specifically, we first compute the expected profit of

structured features only is the ensemble model, achieving an ROI of 16.02%. When both text and structured data are used, the ROI reaches 18.63% in the "No GenAI" scenario, showing an increase of 2.61 percentage points (or a 14% relative increase). This increase of ROI when using text in the predictive model may be considered as soft information as quantified by ROI.

Table 5: Measuring Soft Information Using AUC and ROI

	AUC			ROI		
Scenario	Model 1A (Structured)[a]	Model 1B (Structured + Text)	Diff.[b] (Soft Info.)	Model 2A (Structured)[a]	Model 2B (Structured + Text)	Diff.[b] (Soft Info.)
No GenAI		74.7%	2.2%***		18.63%	2.61%***
Small-Scale Adoption	72.5%	73.8%	1.3%***	16.02%	15.68%	-0.34
Large-Scale Adoption		74.0%	1.5%***		16.78%	0.76

[a] The best performing model using structured features only is the ensemble model. More details can be found in Appendix I. The performance metric does not change by LLM adoption since the model does not use text.
[b] Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$. All AUCs and ROIs are averaged across 15 replications of the ten-fold mean.

The "No GenAI" scenario is straightforward and consistent with existing literature, which reports positive information value of texts in online lending (Netzer et al., 2019; Gao et al., 2023).

Soft information lost in LLM adoption. The second row of Table 5 shows results in the "Small-Scale Adoption" scenario. Specifically, our lender model reports a 73.8% AUC and 15.68% ROI. Compared with the utility without leveraging text, the quantified soft information would be reduced by a 1.3 percentage points by AUC and nearly none by ROI (a 0.34 percentage points decline without statistical significance). In other words, comparing the "Small-Scale Adoption" scenario with the "No GenAI" scenario, we observe a decline in utility, i.e., a reduction in soft information.

The reduction is somewhat mitigated once we allow the lenders to respond to large-scale LLM adoption. This can be seen in the third row of Table 5. Specifically, the AUC recovered to 74.0%

each loan as:

$$E(profit)_j = [1 - P_j] \times IE_j + [P_j] \times [-0.75 \times AG_j + 0.25 \times IE_j],$$

where $E(profit)_j$ is the expected profit of loan j, P_j is loan j's probability of default based on the corresponding predictive model, IE_j is the interest rate on the loan's term, AG_j is loan j's principal amount, which we assume \$100 for each loan. Following Netzer et al. (2019), we also assume that when a loan becomes default, an average of 25% of the principal is repaid, which is based on data published by crowdfunding consulting agencies. We then rank the loans by their expected profits, and select the top 1,000 loans with the highest expected profit to create an investment portfolio. Assuming \$100 is invested in these 1,000 loans, we calculate the actual profit (or loss) of each loan using their observed default status. In summary, assuming a total investment of \$100,000 (1,000 loans at \$100 per loan), the ROI is first computed as the ratio between the net profit (actual total profit minus the total investment) and the total investment, and then converted to an annual ROI based on the loan's term.

and the ROI increased to 16.78%, reflecting increases of 0.2 percentage points and 1.1 percentage points, respectively, compared to the "Small-Scale Adoption" scenario. At a first glance, this suggests that some "lost information" has been recaptured. That is not the case. Lenders' response in shifting to use more hard information drives the result.

Convergence in writing and soft vs. hard information. The reduction of soft information may be partially attributed to the convergence of writing (Doshi and Hauser, 2023). To demonstrate this point, we compute a homogeneity index based on the text descriptions' pairwise similarity by credit scores and compare them before and after rewrites.[16] Results are in Figure 9, from which we can see several interesting patterns. First, borrowers with better credit profiles show much less homogeneity than those with worse credit profiles before the rewrites. This indicates that high-score borrowers write in a more diverse manner than low-score borrowers. Second, after rewriting, the homogeneity of each group increased substantially, with the high-score groups increased relatively more than the low-score groups, narrowing the gap in homogeneity across credit scores. This finding suggests that the application of ChatGPT increases the homogeneity of writing. The convergence in writing may explain the reduction in soft information when lenders attempt to distinguish high quality loans from low quality ones.

Further, we show evidence that moving from the "Small-Scale Adoption" to the "Large-Scale Adoption" scenario, lenders increasingly rely on hard information than soft information. Specifically, we measure the aggregated feature importance (AFI) among structured features and textual features in the lender model.[17] Results are shown in Figure 10. We can see that the AFI score of

[16] The homogeneity index is computed as the average cosine similarity between any pair of texts in the group. Namely,

$$\text{Group Homogeneity} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{Cosine Similarity}(\text{emb}(\mathbf{x}_i), \text{emb}(\mathbf{x}_j)), \tag{9}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ represent the text description of loan i and loan j, respectively and n represents the total number of samples in this group. Due to the substantial computational expense of calculating similarity for all sample pairs, we employ a bootstrap sampling approach (Hernández et al., 2017) to estimate the group homogeneity. Specifically, we randomly sample 20% of the data within each group for computation with 10 repetitions.

[17] We compute the AFI based on estimates of feature importance aggregated by feature group. In particular, we first compute the Shapley value (Štrumbelj and Kononenko, 2014) of each feature in the lender model. Then, the aggregated feature importance is computed as

$$\text{AFI}(\mathcal{S}) = \frac{\sum_{f \in \mathcal{S}} \text{SHAP}(f)}{\sum_{f \in \mathcal{A}} \text{SHAP}(f)}, \tag{10}$$

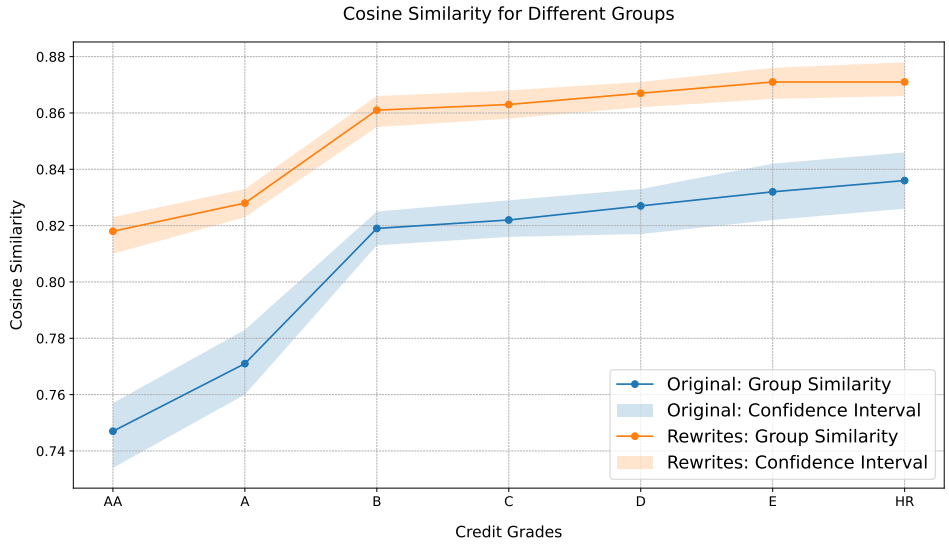


Figure 9: Text Similarity by Credit Score Categories Before and After Rewriting.

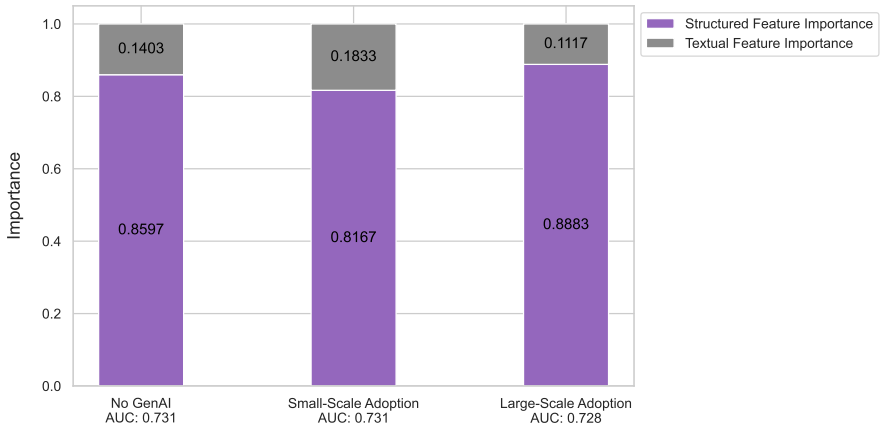


Figure 10: Aggregated Feature Importance (AFI) Scores in Different Scenarios.

structured features is 0.82 in the "Small-Scale Adoption" scenario, which increased to 0.89 in the "Large-Scale Adoption" scenario.

5 Misallocation in Generative Equilibrium

We now discuss the aggregate implications of writing quality and LLM adoption in the economy, starting with the efficiency of credit allocation before and after the introduction of LLMs. We then discuss how lenders' cognitive or computation-constraints affect the results. Finally, we establish the existence and uniqueness of the counterfactual equilibrium in the GenAI age, in addition to characterizing its aggregate properties.

5.1 Credit Misallocation

To analyze the impact of GenAI adoption on lenders' misallocation, we evaluate misallocation using FPR and FNR as the metrics.[18] Results are summarized in Table 6.

We can see that without GenAI, the baseline FPR and FNR are at 17.50% and 20.86%, respectively. In the context of "Small-Scale Adoption," FPR and FNR significantly increase to 23.12% and 23.28%, reflecting an increase of 5.62 and 2.42 percentage points (or a 32.1% and 11.6% relative increase), respectively. This indicates a substantial increase in misallocation if borrowers leverage GenAI to improve writing while lenders have not updated their decision model. When lenders do update their decision model, as in the large-scale adoption scenario, FPR becomes 20.52%, which is still much higher than the "No GenAI" scenario but reflects a substantial (2.6 percentage points, or a 11.2% relative) reduction compared with the small-scale adoption scenario.

where $\mathcal{A}$ represents the set of all features in the model, $\mathcal{S} \subseteq \mathcal{A}$ represents a subset of features (e.g., all structured features) that we could like to aggregate, and SHAP(f) represents the Shapley value of feature f. The AFI, ranging from 0 to 1, can be interpreted as the relative weight of a group of features in a model. To make the Shapley value computation efficient, we use the XGBoost model (Chen and Guestrin, 2016) as a surrogate model of the multimodal Transformer. Using a surrogate model is a common technique when assessing the Shapley values in deep learning models (Schoch et al., 2023).

[18] Considering repayment as the positive outcome, a case is considered positive (P) if the model classifies it as repaid and negative (N) if the model classifies it as defaulted. A false positive (FP) case is one that the model classifies as subsequently repaid and it actually defaults, and a false negative (FN) case is one that the model classifies as subsequently defaulted but it would actually be repaid. The false positive rate is calculated as the ratio between the number of false positive cases and the number of positive cases (i.e., FPR = FP/P). Similarly, the false negative rate is calculated as ratio between the number of false negative cases and the number of negative cases (i.e., FNR = FN/N).

Moreover, FNR becomes 20.28%, which is equivalent to the "No GenAI" scenario (i.e., no statistical difference found) and reflects a substantial (3.0 percentage points, or a 12.9% relative) reduction compared with the small-scale adoption scenario. These trends indicate that if the lenders respond to borrowers' LLM adoption by updating their lending model, FPR and FNR partially recover.

Table 6: Misallocation in Different GenAI Adoption Scenarios

Scenario	FPR	DIFF.[a]	FNR	DIFF.[a]
No GenAI	17.50%		20.86%	
Small-Scale Adoption	23.12%	**+5.62%*****	23.28%	**+2.42%*****
Large-Scale Adoption	20.52%	**+3.02%*****	20.28%	- 0.58%

[a.] The differences are relative to the "No GenAI" scenario. Significance tests are based on 15 repetitions of the experiments.
[b.] Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

How do the misallocation errors reflect in the ROI? The third row of Table 5 shows the ROI results under large-scale adoption. Specifically, our lender model reports a 74.0% AUC and 16.78% ROI. Compared with the utility without leveraging text, the quantified soft information would be reduced to a 1.5 percentage points by AUC and nearly none by ROI (a 0.76 percentage points increase without statistical significance). Comparing the "Large-Scale Adoption" with the "Small-Scale Adoption" scenario, we observe an improvement in the lender's utility, which suggests that soft information is somewhat recaptured and the financial loss is partially mitigated.

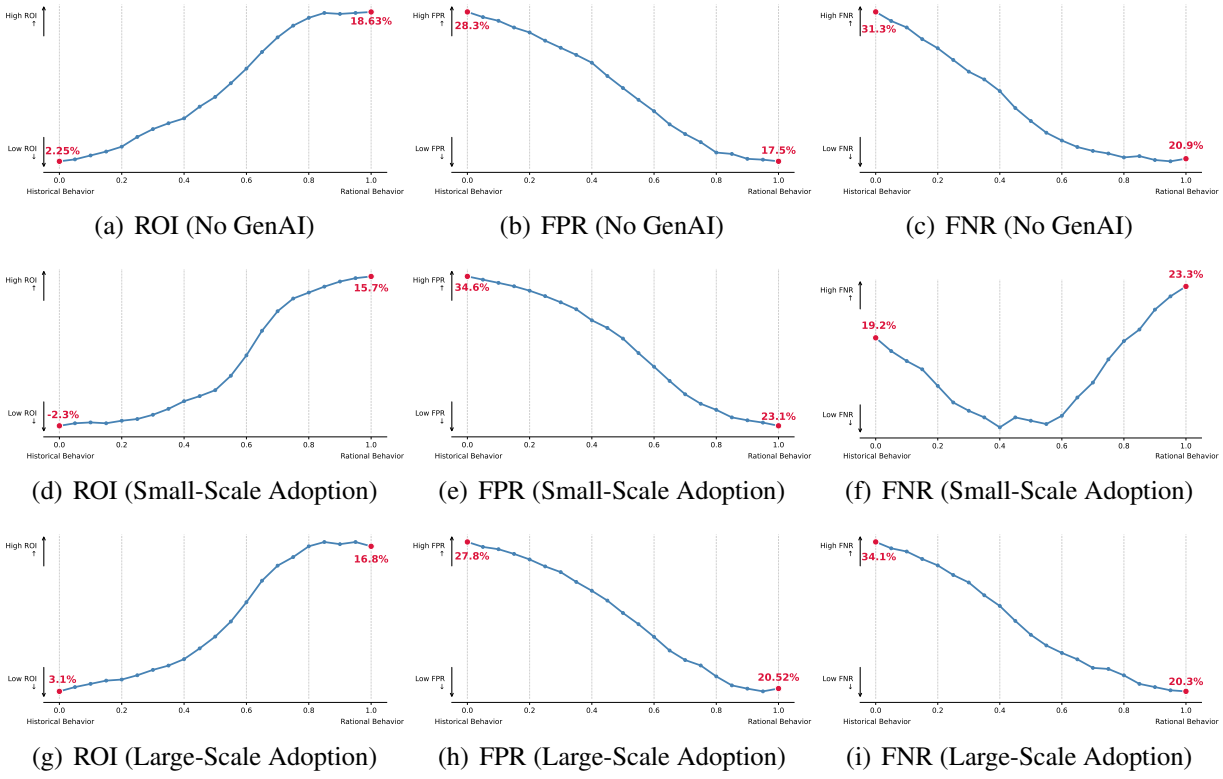
5.2 Constrained Lenders

So far, we have built one of the most cutting-edge multi-modal AI model for the lenders, in order to examine lenders' decision-making in a world where borrowers are adopting LLMs. However, in our sample period, lenders did not have access to such powerful tools. In addition, they may face various cognitive constraints leading to behavioral biases. To model lender decision-making in the online credit market more generally, we first define two decision-making modes, one is the constrained lender mode $\mathcal{M}_{cl}(\boldsymbol{x})$ and the other is the intelligent lender mode $\mathcal{M}_{il}(\boldsymbol{x})$. $\mathcal{M}_{cl}(\boldsymbol{x})$ is trained to mimic historical lending behavior whereas $\mathcal{M}_{il}(\boldsymbol{x})$ is trained to predict the loan whether to repay or not given the current information, and then use the prediction to maximize ROI. Again, we consider all three scenarios, one is without GenAI, another is with small-scale adoption of GenAI, and the other is with large-scale adoption of GenAI.

We stipulate that lenders' decisions in practice follow a combination of the two modes, which effectively captures that financial constraints, behavioral biases, technological limitations, and computational obstacles prevent the lenders from rationally maximizing financial returns with to-date the most powerful machine learning tool. The general lender model would be specified as:

$$\mathcal{M}(\boldsymbol{x}) = \alpha \mathcal{M}_{il}(\boldsymbol{x}) + (1-\alpha)\mathcal{M}_{cl}(\boldsymbol{x}). \tag{11}$$

Here, hyperparameter $\alpha \in [0, 1]$ denotes the level of unconstrainedness.



(a) ROI (No GenAI) (b) FPR (No GenAI) (c) FNR (No GenAI)

(d) ROI (Small-Scale Adoption) (e) FPR (Small-Scale Adoption) (f) FNR (Small-Scale Adoption)

(g) ROI (Large-Scale Adoption) (h) FPR (Large-Scale Adoption) (i) FNR (Large-Scale Adoption)

Figure 11: Misallocation Estimates in Different Counterfactual Scenarios.

We repeat the earlier analyses based on the general specification, and report the results inFigure 11. Each subfigure illustrates the pattern of a performance metric (ROI, FPR, and FNR, respectively, by column) in a different market scenarios (by row) when α in Eqn. (11) varies. We observe several interesting patterns. First, ROI increases as the weight of the intelligent lender mode increases, but FPR decreases with this weight. This implies that constrained lenders tend to leave money on the table and have greater misallocation. But the overall pattern that lender model adjustment mitigates

misallocation carries through, which is consistent with Sec 4 and Table 6 in Sec 5.1. In fact, when lenders are constrained (i.e., do not have access to a powerful multi-modal AI model to aid their lending decisions), lenders' adjustment in response to the data generated with LLM usage more or less restore the level of soft information usage and allocation efficiency to the pre-GenAI era.

Second, while a similar pattern was expected for the FNR, we find that the FNR in the "Small-Scale Adoption" scenario first decreases and then increases as the weight of the intelligent lender model increases. This phenomenon may be explained by the fact that the intelligent lender model gets more conservative in the "Small-Scale Adoption" scenario. It is more likely to decline funding loans than the constraint model in this challenging investment context, which has helped improve the overall ROI. Reduced FNR for medium level of α may be explained by a reduced level of agreement between the intelligent lender mode and the constraint lender mode.

5.3 A Generative Counterfactual Equilibrium

Recall that our ultimate goal is to analyze the counterfactual equilibrium where borrowers endogenously adopt LLMs and lenders train their model based on the counterfactual data generated. We have thus far analyzed two extreme cases where lenders do not react to LLM adoption and lenders adjust their lending decisions based on counterfactual loan applications assuming borrowers' full adoption of LLMs. We now follow Cong (2024) to leverage the power of GenAI itself to analyze a counterfactual general equilibrium.

Borrowers as counterfactual data generator. Each borrower $i \in G$ chooses a high-level action:

$$a_i^H \in \{\text{use ChatGPT}, \text{not use ChatGPT}\} \tag{12}$$

If ChatGPT is used, the loan application text a_i^L is drawn from a generative distribution:

$$a_i^L \sim G_i(\cdot|a_i^H). \tag{13}$$

Borrowers maximize their expected funding probability:

$$\max_{a_i^H} \mathbb{E}\left[P_{\text{funding}}(a_i^H, a_i^L, \sigma_j^*).\right] \tag{14}$$

Lenders as data-driven decision-makers. Lenders $j \in D$ form beliefs about borrower creditworthiness (likelihood of repayment), θ, upon observing the counterfactual data D^C:

$$\mu_j(\theta|D^C). \tag{15}$$

Given AI-driven changes in writing patterns, lenders shift toward relying more on hard information such as credit scores and financial history.

Lenders update their decision function σ_j^* by solving:

$$\sigma_j^* = \arg\max_{\sigma_j} \mathbb{E}_{\mu_j}[u_j(\sigma_j(D^C), a^H, a^L, \omega)] \tag{16}$$

General search for equilibria. To solve for an equilibrium, one can follow several steps:

1. Initialize borrower strategies by conjecturing that a particular subset of borrowers use ChatGPT.

2. Simulate Loan Applications: Use generative AI to create counterfactual loan texts based on borrower strategies, and combine this with structural information about the borrowers to generate counterfactual loan applications.

3. Train Lender Models: Train lenders on both historical and generated data, updating their beliefs about loan qualities utilizing the Deep Heckman Correction we develop.

4. Update Borrower Decisions: Given lender responses, borrowers re-optimize their decision to use ChatGPT or not.

5. Iterate Until Convergence: Repeat until borrower adoption and lender responses reach a fixed point, which gives one equilibrium.

Rating-sorted equilibrium. Empirical results show that high-credit borrowers benefit more from AI-enhanced applications. This suggests a natural sorting structure where higherscore borrowers are more likely to adopt ChatGPT. So we focus on rating-sorted equilibria.

We conjecture an GenAI adoption criterion and level. Each borrower has a certain credit profile, takes the equilibrium GenAI adoption rate and thus the lender's model as given, and makes a decision based on the utility from her use of GenAI. Aggregating all borrowers' adoption decision and the loan applications generated, we then train the proprietary lender's model. Finally, we verify that the equilibrium adoption is indeed the conjectured value and the borrowers' adoption decisions are rational. We repeat the process until all the equilibria are identified.

Specifically, we define borrowers' utility as the improvement in fundability based on the rewrite when compared with the fundability using the original text. We also conjecture that high-score borrowers have the highest incentives to adopt GenAI. We then check whether the top 1% credit-score borrowers adopt and then evaluate the **marginal borrower condition**:

$$\mathbb{E}[\text{Fundability}(\text{ChatGPT}, \sigma^*)] - \mathbb{E}[\text{Fundability}(\text{No ChatGPT}, \sigma^*)] = C_{\text{ChatGPT}} \tag{17}$$

where C_{ChatGPT} is the cost of AI adoption, which we set to zero given how cheap it is for individual borrowers to access LLMs these days. If the net benefit of adopting is strictly positive, we check the top 2%, and so on, until we find an adoption level that make the equality hold.

The results of searching by marginal borrower are shown in Figure 12. We can see that as the adoption rate increases from 0 to about 0.4, there is a rapid increase in the fundability improvements for all borrowers, but the largest increase is for high-score borrowers. As the adoption rate continues to increase from about 0.4 to 1, this improved funding rate begins to taper off, but at the same rate of adoption, high-score borrowers still benefit the most; as the probability of adoption nears 1, the increased gains for all borrowers reach a minimum and are essentially similar. Despite some heterogeneity (e.g., higher-credit borrowers receive a higher utility, mid-level market adoption leads to higher personal utility) almost all borrowers, irrespective of their credit scores or assumed market adoption rates, receive a positive utility from GenAI adoption. Therefore, if borrowers are rational and the cost of GenAI use is minimal, all borrowers are motivated to adopt GenAI, resulting a unique full-adoption long-run equilibrium.

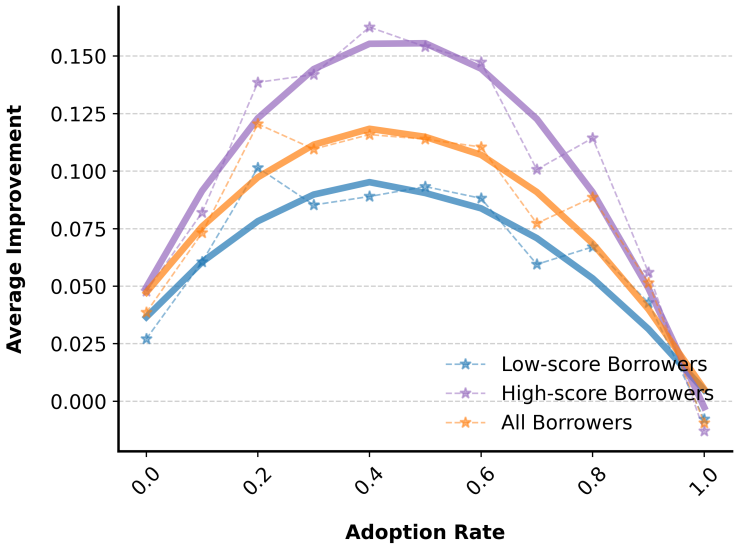


Figure 12: Market Equilibrium. (Dashed lines are results computed at increments of 10% market adoption rate, and solid lines are smoothed curves.)

In this unique credit-sorted equilibrium, every borrower adopts GenAI for online loan applications, and lenders' decision model is data-driven and trained on the synthetic rewrites. This is exactly the Large-Scale Adoption scenario. As we have shown earlier, the loss of soft information is somewhat mitigated relative to the partial equilibrium outcomes, but the lender still relies more on hard information compared to the case without GenAI.

6 Conclusion

We utilize both a commercially available LLM and proprietary multi-modal AI models to investigate how writing quality of online loan applications and the emergence of LLMs affect lending decisions and credit allocation. We simulate borrowers' use of LLMs to generate counterfactual loan applications, and then model lender decisions with a multi-modal transformer that accounts for sample observation bias using a novel Deep Heckman Correction and incorporates lenders' cognitive or computational constraints in reduced form. We also characterize the unique general equilibrium where borrowers' adoption of LLMs are monotone in their credit rating scores.

We find that LLMs can help significantly improve the writing quality and loan fundability. Low credit-score borrowers write better, as measured by the Writing Quality Index (WQI), than high credit-score borrowers. Even with advanced lender models, ChatGPT-generated applications in

aggregate reduce the predictive power of texts and thus the utility of soft information in the market, likely due to the convergence in writing. Moreover, borrowers' strategic use of LLMs could lead to a systematic gap between the perceived and actual loan risk, increasing the likelihood of credit misallocation. However, once lenders endogenously respond to applications generated by LLMs, they focus more on the hard information and the misallocation are mitigated.

Our findings have the following implications: First, writing matters in online lending and borrowers, especially those with high credit scores, should improve the writing quality of their loan applications (e.g., using ChatGPT). Second, lenders and regulatorsshould anticipate an increase in writing quality and a decrease in information value from the text-based loan description. If borrower-generated text can no longer reliably signal behavioral traits, lenders will need to reevaluate the weight they assign to textual information to focus more on the hard information or rely on alternative forms of soft information, such as digital footprints or social network data. The platforms can also mandate disclosure of LLM usage, allowing lenders to recalibrate trust in textual narratives.[19] More generally, our framework for analyzing counterfactual responses and general equilibria can be applied to other settings.

[19] Studies have developed methods to detect LLM usage. See, e.g., Cong and Zhu (2024). A metric like WQI could also prove useful in such detections.

References

Turan G Bali, Bryan T Kelly, Mathis Mörke, and Jamil Rahman. Machine forecast disagreement. Technical report, National Bureau of Economic Research, 2023.

III Bowen, Donald E., S. McKay Price, Luke C.D. Stein, and Ke Yang. Measuring and mitigating racial disparities in large language model mortgage underwriting. *Available at SSRN 4812158*, 2024.

Fabio Caldieraro, Jonathan Z Zhang, Marcus Cunha Jr, and Jeffrey D Shulman. Strategic information transmission in peer-to-peer lending markets. *Journal of Marketing*, 82(2):42–63, 2018.

Sean Cao, Wei Jiang, Baozhong Yang, and Alan L Zhang. How to talk when a machine is listening: Corporate disclosure in the age of ai. *The Review of Financial Studies*, 36(9):3603–3642, 2023.

Sean Cao, Wei Jiang, Junbo Wang, and Baozhong Yang. From man vs. machine to man+ machine: The art and ai of stock analyses. *Journal of Financial Economics*, 160:103910, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Yifei Chen, Bryan T Kelly, and Dacheng Xiu. Expected returns and large language models. *Available at SSRN 4416687*, 2022.

Lin William Cong. Data-driven generative equilibrium. Technical report, Working Paper, 2024.

Lin William Cong, Ke Tang, Jingyuan Wang, and Yang Zhang. Alphaportfolio: Direct construction through reinforcement learning and interpretable ai. *Social Science Research Network*, Working Paper, 2019.

Can Deniz Deveci, Jason Joe Baker, Binyamin Sikander, and Jacob Rosenberg. A comparison of cover letters written by ChatGPT-4 or humans. *Danish Medical Bulletin (Online)*, 70(11), 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Gregor Dorfleitner, Christopher Priberny, Stephanie Schuster, Johannes Stoiber, Martina Weber, Ivan de Castro, and Julia Kammler. Description-text related soft information in peer-to-peer lending–evidence from two leading european platforms. *Journal of Banking & Finance*, 64: 169–187, 2016.

Anil R Doshi and Oliver Hauser. Generative artificial intelligence enhances creativity. *Available at SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4535536, 2023.

Trevor Fitzpatrick and Christophe Mues. How can lenders prosper? comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research*, 294(2):711–722, 2021.

Qiang Gao, Mingfeng Lin, and Richard Sias. Words matter: The role of readability, tone, and deception cues in online credit markets. *Journal of Financial and Quantitative Analysis*, 58(1): 1–28, 2023.

Anindya Ghose and Panagiotis G Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, 2010.

Ken Gu and Akshay Budhkar. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.maiworkshop-1.10. URL https://www.aclweb.org/anthology/2021.maiworkshop-1.10.

Yanhong Guo, Wenjun Zhou, Chunyu Luo, Chuanren Liu, and Hui Xiong. Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, 249(2):417–426, 2016.

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007.

Sarah K Harkness. Discrimination in lending markets: Status and the intersections of gender and race. *Social Psychology Quarterly*, 79(1):81–93, 2016.

James Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161,

1979.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1):18617, 2023.

Wilmar Hernández, Carlos Calderón-Córdova, Vicente González-Posada, Ángel Parra-Cerrada, José Luis Jiménez, José Enrique González-Garcia, José G Zato, and Oleg Yu Sergiyenko. Bootstrap-based frequency estimation method. *Measurement*, 95:193–200, 2017.

Juei-Ling Ho, Kuan-Ying Chen, Lan-Hsun Wang, Shih-Shuo Yeh, and Tzung-Cheng Huan. Exploring the impact of social media platform image on hotel customers' visit intention. *International Journal of Contemporary Hospitality Management*, 34(11):4206–4226, 2022.

Byoung-Hyoun Hwang and Hugh Hoikwang Kim. It pays to write well. *Journal of Financial Economics*, 124(2):373–394, 2017.

Rajkamal Iyer, Asim Ijaz Khwaja, Erzo FP Luttmer, and Kelly Shue. Screening peers softly: Inferring the quality of small borrowers. *Management Science*, 62(6):1554–1577, 2016.

Hyungu Kahng, Hyungrok Do, and Judy Zhong. Domain generalization via heckman-type selection models. In *The 11th International Conference on Learning Representations (ICLR)*, 2023.

Chansog Kim, Ke Wang, and Liandong Zhang. Readability of 10-k reports and stock price crash risk. *Contemporary accounting research*, 36(2):1184–1216, 2019.

Laura Larrimore, Li Jiang, Jeff Larrimore, David Markowitz, and Scott Gorski. Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1):19–37, 2011.

Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671, 2014.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, Red Hook, NY, USA, 2017.

G Harry McLaughlin. SMOG grading—a new readability formula. *Journal of Reading*, 12(8): 639–646, 1969.

Oded Netzer, Alain Lemaire, and Michal Herzenstein. When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research*, 56(6):960–980, 2019.

Fernando Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014. URL https://github.com/bayesian-optimization/BayesianOptimization.

Ellen K Nyhus and Paul Webley. The role of personality in household saving and borrowing behaviour. *European Journal of Personality*, 15(S1):S85–S103, 2001.

Daniel M Oppenheimer. Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(2):139–156, 2006.

Myle Ott, Claire Cardie, and Jeff Hancock. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, pages 201–210, 2012.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Joe Ryan, Katya Reuk, and Charles Wang. To fund or not to fund: Determinants of loan fundability in the prosper.com marketplace. Technical report, The Stanford Graduate School of Business, January 2007. URL https://www.prosper.com/downloads/research/prosper_regression_projectfundability_study.pdf. Confidential.

Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. Data selection for fine-tuning large language models using transferred shapley values. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 266–275, Toronto, Canada, July

2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.37. URL https://aclanthology.org/2023.acl-srw.37/.

Michael Siering. Peer-to-peer (P2P) lending risk management: Assessing credit risk on social lending platforms using textual factors. *ACM Transactions on Management Information Systems*, 14(3), Jun 2023. ISSN 2158-656X. doi: 10.1145/3589003. URL https://doi.org/10.1145/3589003.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

Catalina L Toma and Jeffrey T Hancock. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1):78–97, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Lifang Zhang, Jianzhou Wang, and Zhenkun Liu. What should lenders be more concerned about? developing a profit-driven loan default prediction model. *Expert Systems with Applications*, 213: 118938, 2023.

Christoph Zott and Quy Nguyen Huy. How entrepreneurs use symbolic management to acquire resources. *Administrative Science Quarterly*, 52(1):70–105, 2007.

Appendix A Sample Descriptives and Baseline Features

We compile a list of structured features as detailed in Table A1.

Table A1: Definitions of Structured Features

ID	Feature	Description
Panel A: Loan Request and Verifiable Information		
1	AmountRequested	The natural logarithm of the loan amount applied for by the borrower.
2	AmountDelinquent	The natural logarithm of one plus the total delinquency amount in dollars.
3	DelinquenciesLast7Years	The natural logarithm of one plus the total number of delinquencies recorded in the past seven years.
4	InquiriesLast6Months	The total number of credit inquiries made in the past six months.
5	PublicRecordsLast10Years	The total count of public records, including bankruptcies, civil judgments, and tax liens, over the past ten years.
6	PublicRecordsLast12Months	The total count of public records, including bankruptcies, civil judgments, and tax liens, from the past year.
7	CurrentCreditLines	The natural logarithm of one plus the borrower's current number of credit lines.
8	OpenCreditLines	The natural logarithm of one plus the number of open credit lines held by the borrower.
9	RevolvingCeditBalance	The natural logarithm of one plus the borrower's revolving credit balance in dollars.
10	BankCardUtilization	The ratio of the total balances owed by the borrower on open bankcards to the total credit limits of those cards.
11	CreditGrade	Letter representing the borrower's credit score at the time of application.
Panel B: Self-Disclosed Information		
12	DebtToIncomeRatio	The ratio of a borrower's total debt to their income, expressed as a percentage.
13	HasImage	An indicator showing if the listing contains an image with at least one adult.
14	IsBorrowerHomeowner	An indicator representing homeownership status (owned vs. not owned).
15	EmploymentStatus	Indicators for employment status: unknown, working full-time, not employed, working part-time, retired, and self-employed.
16	LengthStatusMonths	The duration of the current employment status, measured in months, as of the listing's creation date.
17	IncomeLevel	Income level indicators include: below $25k, $25k to under $50k, $50k to under $75k, $75k to under $100k, above $100k, and missing information.
18	HasLoanDescription	Whether there is a non-empty description.
19	Text Length	The total word count in the loan description provided.
20	Spelling Errors	The percentage of spelling errors within loan descriptions, when the loan description is non-empty.
21	Grammatical Errors	The percentage of grammatical errors within loan descriptions, when the loan description is non-empty.
22	Formality	The percentage of words with six or more letters, when the loan description is non-empty.
23	Readability	The "simple measure of gobbledygook" (SMOG) readability index.
24	Tone	The sentiment of the loan descriptions.
Panel C: Lending Outcomes[1]		
25	BorrowerMaximumRate	The maximum rate the borrower agrees to pay for the loan.
26	BorrowerRate	The rate the borrower pays if the loan were to close at this point in time.
27	LenderRate	The rate that lenders would receive on the listing if the loan were to close at this point in time.
28	BidMaximumRate	The Maximum Rate in which a bidder will be able to lend money in order to create a winning bid.
29	BidCount	The number of bids received by the listing in total.
30	FundLabel	An indicator that equals to one if a listing is fully funded and becomes a loan.
31	RepayLabel	An indicator that equals to one if the loan fully repaid.

[1.] Panel C features are observed during or after the bidding stage.

To examine the relationship between loan status (funded vs. unfunded and repaid vs. default) and structured features, we also compared the differences in loan characteristics between different groups. Table A2 presents descriptive statistics of these structured features.

Table A2: Summary Statistics

		Stage 1 Outcomes				Stage 2 Outcomes			
		Unfunded (U) (n=19,484)		Funded (F) (n=19,484)		Default (D) (n=6,886)		Repaid (R) (n=12,598)	
ID	Numeric Structured Features	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
1	AmountRequested	8.999	0.814	8.434	0.853	8.393	0.845	8.509	0.862
2	AmountDelinquent	2.827	3.710	1.715	3.135	1.385	2.882	2.317	3.473
3	DelinquenciesLast7Years	1.034	1.300	0.820	1.204	0.725	1.146	0.994	1.287
4	InquiriesLast6Months	3.313	4.258	2.580	3.551	2.031	2.863	3.584	4.372
5	PublicRecordsLast10Years	0.440	0.956	0.368	0.885	0.305	0.746	0.483	1.084
6	PublicRecordsLast12Months	0.044	0.254	0.038	0.235	0.032	0.207	0.049	0.279
7	CurrentCreditLines	2.270	0.613	2.195	0.617	2.205	0.591	2.176	0.662
8	OpenCreditLines	2.138	0.609	2.056	0.617	2.066	0.591	2.038	0.661
9	RevolvingCreditBalance	8.051	2.960	7.880	2.925	7.986	2.773	7.686	3.176
10	BankCardUtilization	0.554	0.371	0.552	0.373	0.524	0.361	0.604	0.388
12	DebtToIncomeRatio	0.574	1.408	0.325	0.892	0.297	0.793	0.378	1.046
16	LengthStatusMonths	23.37	56.23	23.70	54.80	21.80	53.04	27.19	57.71
19	Text Length	182.8	121.5	222.2	139.4	235.1	146.7	215.1	134.7
20	Spelling Errors	0.036	0.022	0.033	0.020	0.033	0.019	0.034	0.020
21	Grammatical Errors	0.068	0.042	0.059	0.039	0.063	0.039	0.057	0.038
22	Formality	0.243	0.060	0.236	0.055	0.237	0.056	0.235	0.055
23	Readability	10.24	4.149	10.66	3.347	10.56	3.426	10.71	3.302
24	Tone	0.739	0.156	0.752	0.158	0.744	0.158	0.755	0.158
25	BorrowerMaximumRate	0.181	0.085	0.218	0.080	0.202	0.080	0.246	0.072
26	BorrowerRate	0.179	0.085	0.182	0.077	0.166	0.073	0.213	0.074
27	LenderRate	0.178	0.085	0.181	0.077	0.165	0.073	0.211	0.074
28	BidMaximumRate	0.178	0.085	0.181	0.077	0.164	0.073	0.210	0.074
29	BidCount	24.74	76.65	154.5	148.8	156.2	145.9	151.4	154.0
	Categorical structured Features	n	(%)	n	(%)	n	(%)	n	(%)
11	CreditGrade								
	CreditGrade=AA	2,571	13.2%	2,571	13.2%	448	6.5%	2,123	16.9%
	CreditGrade=A	2,529	13.0%	2,529	13.0%	654	9.5%	1,875	14.9%
	CreditGrade=B	3,373	17.3%	3,373	17.3%	1,132	16.4%	2,241	17.8%
	CreditGrade=C	4,166	21.4%	4,166	21.4%	1,523	22.1%	2,643	21.0%
	CreditGrade=D	3,568	18.3%	3,568	18.3%	1,436	20.9%	2,132	16.9%
	CreditGrade=E	1,657	8.5%	1,657	8.5%	791	11.5%	866	6.9%
	CreditGrade=HR	1,620	8.3%	1,620	8.3%	902	13.1%	718	5.7%
13	HasImage	10,542	54.1%	13,114	67.3%	4,594	66.7%	8,520	67.6%
14	IsBorrowerHomeowner	9,919	50.9%	10,279	52.8%	3,457	50.2%	6,822	54.2%
15	EmploymentStatus								
	EmploymentStatus=FT	16,414	84.2%	16,722	85.8%	5,843	84.9%	10,879	86.4%
	EmploymentStatus=SE	1,462	7.5%	1,474	7.6%	644	9.4%	830	6.6%
	EmploymentStatus=PT	755	3.9%	754	3.9%	209	3.0%	545	4.3%
	EmploymentStatus=R	650	3.3%	377	1.9%	148	2.1%	229	1.8%
	EmploymentStatus=NE	203	1.0%	157	0.8%	42	0.6%	115	0.9%
17	IncomeLevel								
	Income ($0 or unavailable): 1	368	1.8%	767	4.0%	277	4.0%	490	3.9%
	Income ($1-24999): 2	2,731	14.0%	2376	12.2%	866	12.6%	1,510	12.0%
	Income ($25,000-49,999): 3	7,256	37.2%	7272	37.3%	2,646	38.4%	4,626	36.7%
	Income ($50,000-74,999): 4	4,893	25.1%	4940	25.4%	1,708	24.8%	3,232	25.7%
	Income ($75,000-99,999): 5	2,117	10.9%	2214	11.4%	758	11.0%	1,456	11.6%
	Income ($100,000+): 6	2,121	10.9%	1915	9.8%	631	9.2%	1,284	10.2%

Appendix B Sample Rewrites

In this section, we present some examples of LLM-generated loan descriptions. As shown in Table A3, ChatGPT rewrites significantly improve the readability and formality of the loan descriptions.

Table A3: Examples of Loan Description Rewrites

Original	Rewrite
My financial situation: I am a good candidate for this loan because I am preparing for a future business so it is imperative that I get reed of these high intrest debts and stop making the credit card companies rich. I will never miss a payment one because Im using direct payments two because I want to use this site later on for my business loan.	Financial Situation: I present myself as an ideal candidate for this loan, as my unwavering commitment to fiscal responsibility is evident. With the vigor of an entrepreneur striving to secure a prosperous future, I am determined to eliminate these burdensome debts. By employing direct payment methods, I guarantee the punctuality of my monthly installments, reinforcing my reliability. Furthermore, I seek to utilize this esteemed lending platform for my future business aspirations, further exemplifying my commitment to honor my financial obligations. Monthly Monetary Overview: Amidst the discerning evaluation of my income and expenses, my financial prudence is unquestionably apparent.
Monthly net income: $ 3,700	My monthly net income stands at a commendable $3,700, allowing for ample liquidity to address the pertinent matter at hand.
Purpose of loan: The purpose of the loan id get back from a payed vacation that i get won 7 day eastern caribbean curise all payed for by cal worthington (go see cal go see cal) jsut want to have enough money to cover all house hold expense.	Loan Purpose: The purpose of this loan is to efficiently manage my household expenses after returning from an all-expenses-repaid 7-day Eastern Caribbean cruise, graciously won through the renowned Cal Worthington (Go see Cal, go see Cal!). By securing this loan, I intend to ensure that my financial obligations are met seamlessly.
I am using these funds to re-invest in select Prosper loans. I am currently a Banker, a Real Estate Developer and a Citrus Grower. I will be retiring from the Banking industry (at age 44) towards the end of June and will have an income well excess of $100k annually even after leaving and going full time in pursuit of development and citrus. I am borrowing a small amount and probably won't keep for the full term but will keep for at least 9-12 months. At this point, until I see the results of my lending efforts I am not willing to take any undue risks. Small time arbitrage!!Personal debt obligations are minimal and the majority of what shows under revolving is American Express paid off monthly and my Home Equity LOC in the amount of $40K now (was $70K) at time of pull. By the way, $25K of my LOC will be re-paid within 6 months as those were funds loaned to my development entity and the funds will be available within 6 months from that entity as lots are sold...more information than you needed probably. Be happy to answer any questions you may have.	I seek financial support to further invest in carefully chosen Prosper loans. As a seasoned Banker, Real Estate Developer, and Citrus Grower, I have diverse expertise that enhances my investment acumen. Despite my decision to retire from banking at the young age of 44, my earnings will exceed $100,000 annually post-retirement as I focus entirely on real estate development and citrus cultivation. While my loan amount is modest, I plan to retain it for a minimum of 9-12 months, allowing me to assess the outcomes of my borrowing ventures. Prudent decision-making is my priority at this juncture, as I refrain from undertaking unnecessary risks, adopting a strategy akin to small-scale arbitrage. My personal debt is minimal, largely comprising a monthly-paid American Express balance and a Home Equity Line of Credit of $40,000 (previously $70,000). Notably, $25,000 of the LOC is earmarked for repayment within six months, sourced from a loan to my development enterprise, to be recouped upon lot sales. Should you require further information, I am more than willing to address any queries. Your trust in my financial vision is valued and your support is integral to my continued success. Thank you for considering my loan application.

Appendix C Alternative Prompt Designs

C.1 Different Rule-Setting Specifications

Rule settings are relatively flexible and likely exhibit variation across borrowers, due to differences in their abilities to use LLMs and their own cognitive levels.

Firstly, a borrower might adopt ChatGPT to enhance the readability and emphasizes content that could assist lenders in decision-making. To do so, they might use the sentence "*Please improve the clarity and readability of the text, and emphasize anything that helpful for lenders' decision-making.*" in the rule setting of their prompt. We term this rule specification as "Specification 1" (S-1). To ensure that the rule setting is not influenced by the exact wording in the sentence, we used ChatGPT to generate 10 other ways to express this specification, and the variations are documented in Table A4. When putting together the prompt, we randomly pick one out of these ten variations and put it in the full rewrite prompt.

Table A4: Variations of **S-1**. The original sentence is "*Please improve the clarity and readability of the text, and emphasize anything that helpful for lenders' decision-making.*"

1	"Please enhance the clarity and readability of the text, highlighting any details that could assist lenders in their decision-making."
2	"Improve the text's clarity and readability, ensuring that important points for lenders' decisions stand out."
3	"Enhance the text to make it clearer and easier to read, with a focus on highlighting information that helps lenders in their decision process."
4	"Please improve the readability of this text, making sure to emphasize key points that aid lenders in their decision-making."
5	"Rewrite the text for improved clarity and readability, drawing attention to any details that could influence lenders' decisions."
6	"Please refine the text to boost clarity and readability, while emphasizing critical information helpful to lenders."
7	"Make the text clearer and easier to read, and highlight any details that would assist lenders in deciding."
8	"Improve the text's readability and clarity, focusing on highlighting points that are crucial for lenders' decision-making process."
9	"Improve the text to ensure it's easy to read and clear, while stressing any information useful for lenders' decisions."
10	"Rewrite the text for clarity and ease of understanding, while highlighting points that would be helpful for lenders in their decision-making."

Secondly, the borrower might want to adopt LLMs to help make their reason compelling and improve their funding success. This rule setting, termed "Specification 2" (S-2), may be expressed with the sentence "*Please make the reason compelling, and make myself sound trustworthy, so lenders willing to fund it.* " Similar to the above, we generate 10 variations as listed in Table A5,

and randomly pick one when dynamically generating the full rewrite prompt.

Table A5: Variations of **S-2**. The original sentence is "*Please make the reason compelling, and make myself sound trustworthy, so lenders willing to fund it.* "

1	"Please make the reason persuasive and present myself as reliable, so lenders are eager to provide funding."
2	"Ensure the reason is convincing and I come across as trustworthy, encouraging lenders to fund the loan."
3	"Craft the reason in a compelling way, making me appear dependable, so lenders feel confident in offering support."
4	"Please make the reason persuasive, portraying me as trustworthy, so lenders are inclined to fund it."
5	"Help make the reason strong and convincing, building trust so lenders will want to provide funding."
6	"Ensure the reason is appealing and I appear trustworthy, so lenders feel comfortable funding the loan."
7	"Make the reason convincing and portray me as dependable, so lenders feel confident in offering their support."
8	"Write the reason in a way that inspires confidence and makes me appear trustworthy, so lenders want to fund the loan."
9	"Make sure the reason is strong and I come across as reliable, encouraging lenders to offer funding."
10	"Make the reason compelling and reflect my reliability, so lenders feel comfortable providing funds."

Finally, suppose that the borrower had been exposed to insights from research regarding which listings are more likely to get funded. For example, Netzer et al. (2019) and Gao et al. (2023) report many characteristics of listings that were more successful in securing funding. Referring to these existing research, we put together another sentence termed "Specification 3" (S-3), which says "*Please correct grammatical mistakes and spelling mistakes. Please use a more positive tone to enhance the lender's willingness to fund. Please use advanced vocabulary and ensure the text length is compliant.*" Table A6 lists its ten variants.

Table A6: Variations of **S-3**. The original sentence is "*Please correct grammatical mistakes and spelling mistakes. Please use a more positive tone to enhance the lender's willingness to fund. Please use advanced vocabulary and ensure the text length is compliant.*"

1	"Please correct any grammatical and spelling errors. Ensure the tone is more optimistic to encourage the lender's willingness to fund, and use advanced vocabulary while keeping the text length appropriate."
2	"Correct any grammar or spelling issues and adjust the tone to be more positive, making it appealing to lenders. Use sophisticated language and keep the text length within limits."
3	"Please revise grammatical and spelling errors, adopt a more encouraging tone to increase lender interest, and use advanced terminology while ensuring text length is appropriate."
4	"Correct any errors in grammar and spelling, using a more positive tone to encourage lender engagement. Employ advanced vocabulary and maintain a suitable length for the text."
5	"Please edit grammatical and spelling mistakes, ensuring a positive tone to heighten lender interest. Use an elevated vocabulary and ensure the text adheres to the required length."
6	"Correct all grammar and spelling errors, adjusting the tone to be more uplifting and persuasive for the lender. Apply advanced wording and keep the text concise."
7	"Correct any grammatical and spelling issues, ensuring the tone is upbeat to appeal to lenders. Use high-level vocabulary and make sure the text length is suitable."
8	"Revise the text for any grammar or spelling issues, using a more positive tone to encourage funding. Employ advanced language while ensuring the text stays within length requirements."
9	"Correct all grammatical and spelling errors, and ensure a positive tone to increase the lender's confidence. Use elevated vocabulary and ensure the text remains within the required length."
10	"Correct any grammar or spelling mistakes, adopt a more encouraging tone to appeal to lenders, and use higher-level language while ensuring the text stays within the desired length."

C.2 Prompt Versions

As econometricians, we consider several common scenarios: (1) We examine the special case without a rule setting, where borrowers directly use GenAI tools to rewrite their loan applications; (2) We assess the scenario where LLMs are adopted to enhance the readability and emphasizes content that could assist lenders in decision-making, termed "Specification 1" (S-1); (3) We assess the scenario where LLMs are adopted to make the reason compelling and improve the trustworthy of the loan description, termed "Specification 2" (S-2); (4) Building on existing research, we designed settings from the perspective of textual features that have been proven to increase fundability, termed "Specification 3" (S-3). Notably, each setting aims to improve writing quality to enhance fundability. For robustness tests, we opt to progressively broaden the rule settings stems from the fact that, with a deeper understanding of ChatGPT, users' techniques and abilities will consistently advance. We consider several prompt templates, as summarized in Table A7 and illustrated in Figure A1.

In the following, we use P1R through P4R to represent results corresponding to the rewrites produced by applying prompt versions 1 through 4.

Table A7: Prompt Templates Based on Different Specifications

Prompt Template	Role Setting	Task Setting	Rule Setting
Version 1 (P1R)	✓	✓	None
Version 2 (P2R)	✓	✓	S-1
Version 3 (P3R)	✓	✓	S-1 + S-2
Version 4 (P4R)	✓	✓	S-1 + S-2 + S-3

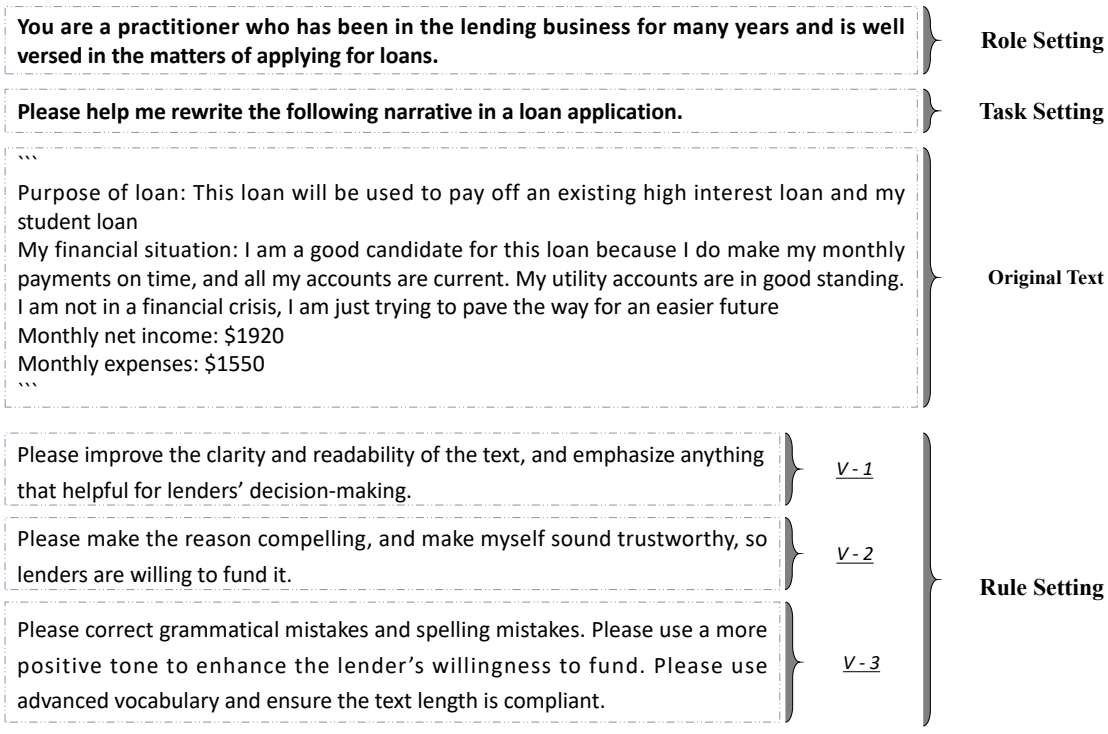


Figure A1: An Example ChatGPT Prompt Showing Different Levels of Specification.

Appendix D Human Assessments of Loan Descriptions

To conduct the experiments where humans assess the writing quality and fundability of ChatGPT rewrites relative to the original ones, we drew a stratified sample of our dataset based on funding status (unfunded, funded), credit level (high, medium, low), and WQI (divided into four quantiles). From the 24 possible combinations of these 3 factors, we randomly selected 5 loan applications per combination, and obtained their rewrite version. In other words, we have 120 loan application samples, with their text description before and after rewrite.

Our experiments were conducted on the Prolific platform, with participants restricted to adults in the United States. To capture a broad perspective on perceived writing quality and willingness to lend, we did not impose specific requirements such as investment experience or education level. Each participant was randomly assigned to one of two groups: one group assessed the perceived writing quality of the texts, while the other evaluated their willingness to fund based on the text.

Each participant evaluated 48 texts, which include 24 original descriptions and their corresponding rewrites, with the sentence order randomly shuffled. Participants were not informed whether the texts were original or ChatGPT-rewritten, and the correspondence between the original and rewritten texts was not disclosed as all 48 pieces of text were randomly ordered.

Does the ChatGPT-Rewritten Version Exhibit a Higher Writing Quality? One group of human raters were shown the following instructions: "*In online peer-to-peer lending platform, individuals request loans from potential lenders. The borrowers' request includes a loan description detailing why they need the loan and their financial situation. As a lender, these are risky investments - historically more than one-third of borrowers default. After reading the loan request descriptions below, please rate their* ***their writing quality*** *on a scale of 1 to 10, where 1 represents a low quality and 10 represents a high quality.*"

Does the ChatGPT-Rewritten Version Enhance Fundability? The other group of human raters were shown the following instructions: "*In online peer-to-peer lending platform, individuals request loans from potential lenders. The borrowers' request includes a loan description detailing why they need the loan and their financial situation. As a lender, these are risky investments - historically more than one-third of borrowers default. After reading the loan request descriptions*

below, please respond with ***your willingness to lend****. Rate your willingness on a scale of 1 to 10, with 1 indicating the lowest willingness and 10 indicating the highest willingness.*"

Appendix E Textual Features and Fundability

Textual features potentially reflect writing quality and exhibit a strong connection with fundability, as documented by many existing studies. How does each of these features help explain fundability, especially when considering their heterogeneous effects on borrowers of different credit grades and while controlling for the other textual features? We employ a logistic regression model and control for known loan characteristics. The model is specified as follows:

$$\begin{aligned}\text{logit}(Outcome_i) =& \beta_0 + \beta_1 \times IsHighCredit_i + \beta_2 \times IsLowCredit_i + \beta_3 \times TF_i \\ &+ \beta_4 \times TF_i \times IsHighCredit_i \\ &+ \beta_5 \times TF_i \times IsLowCredit_i + \boldsymbol{\gamma} \cdot \boldsymbol{c}_i + \varepsilon_i, \end{aligned} \tag{18}$$

where $Outcome_i$ represents whether loan i is funded, TF represents a textual feature, $\boldsymbol{c}_i$ represents the vector of control variables, $\boldsymbol{\gamma}$ is the corresponding coefficient vector, and ε_i represents the random error. With this model specification, the estimated effect of the textual feature would be $\hat{\beta}_3$ for medium credit borrowers, $\hat{\beta}_3 + \hat{\beta}_4$ for high credit borrowers, and $\hat{\beta}_3 + \hat{\beta}_5$ for low credit borrowers. Results can be found in Table A8, where each column considers a model with a textual feature as indicated in the model name.

From Table A8, we can see that each text feature shows a significant relationship with fundability, at least for certain subgroups of different credit grades. For example, spelling errors have a more significant association with fundability for high-score borrowers, while grammatical errors have a more significant association with fundability for low-score borrowers. Additionally, tone has significant, positive association with fundability for low-score borrowers, but this effect is not as prominent for high-score borrowers, suggesting that low-score borrowers may rely more on a positive tone to secure loans. In contrast, the readability score has significant, positive association with fundability for high-score borrowers, indicating that by making the description more advanced increase the chances of receiving funding only for high-score borrowers.

It is worth mentioning that with the same level of model complexity (i.e., allowing one textual feature to interact with credit level while controlling for all the other textual features), WQI reaches the highest R^2 (18.0% for WQI as seen in Model 4 of Table 4, as opposed to 17.4–17.5% for all

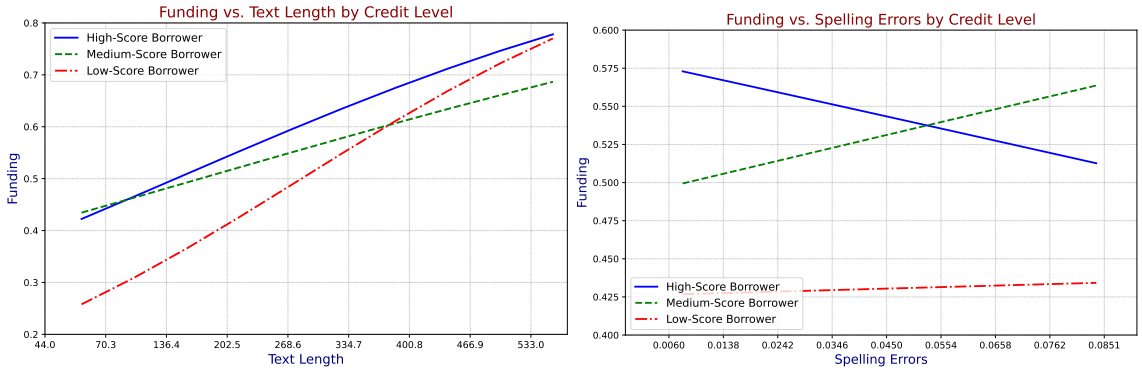
Table A8: Relationship Between Textual Features and Funding Outcome

	DV: Funding (Funded: 1, Unfunded: 0)					
TF ▷	Model 1 (Text Length)	Model 2 (Spelling Errors)	Model 3 (Grammatical Errors)	Model 4 (Formality)	Model 5 (Readability)	Model 6 (Tone)
TF × High-Score Borrower	**0.134*****	**-0.132*****	-0.011	**-0.098*****	**0.073***	0.015
	(0.032)	(0.026)	(0.032)	(0.027)	(0.029)	(0.028)
TF × Low-Score Borrower	**0.314*****	-0.060	**-0.260*****	**-0.169*****	-0.049	**0.094****
	(0.033)	(0.037)	(0.034)	(0.033)	(0.032)	(0.030)
Text Length	**0.270*****	**0.363*****	**0.355*****	**0.360*****	**0.360*****	**0.364*****
	(0.019)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)
Spelling Errors	0.003	**0.068*****	-0.001	0.001	0.008	-0.001
	(0.015)	(0.020)	(0.014)	(0.014)	(0.015)	(0.014)
Grammatical Errors	**-0.118*****	**-0.124*****	**-0.071*****	**-0.112*****	**-0.117*****	**-0.112*****
	(0.015)	(0.015)	(0.017)	(0.014)	(0.014)	(0.014)
Formality	**0.050*****	**0.050*****	**0.052*****	**0.108*****	**0.049*****	**0.054*****
	(0.014)	(0.013)	(0.013)	(0.017)	(0.013)	(0.013)
Readability	**0.057*****	**0.052*****	**0.058*****	**0.058*****	**0.045****	**0.057*****
	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)	(0.012)
Tone	0.022	0.021	0.022	0.023	0.021	0.002
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)
WQI	-0.245	-0.289	-0.257	-0.277	-0.303*	-0.285
	(0.155)	(0.154)	(0.154)	(0.154)	(0.154)	(0.154)
Structured Features Controlled	Yes	Yes	Yes	Yes	Yes	Yes
Intercept	Yes	Yes	Yes	Yes	Yes	Yes
Observations	38,968	38,968	38,968	38,968	38,968	38,968
Pseudo R^2	17.5%	17.4%	17.5%	17.4%	17.4%	17.4%

Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors in parentheses.

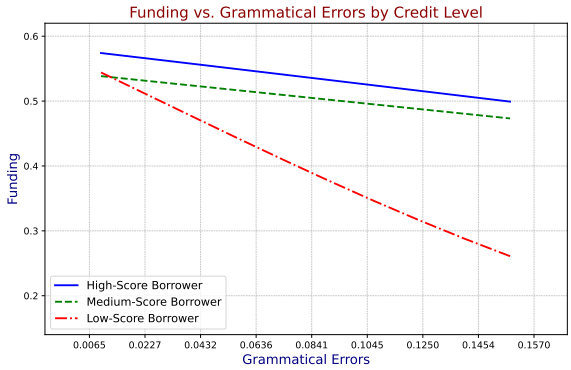
models in Table A8), which supports WQI's comprehensiveness in measuring text characteristic.

Estimated effects by credit level are visualized in Figure A2. We can see a positive relationship between fundability and text length, readability, and tone, alongside a negative relationship with spelling errors, grammatical errors, and formality. These correlations closely related to writing quality, as high-quality writing typically features appropriate text length, fewer errors, a positive tone, and higher readability. As depicted in Figure A2, while increased readability seems helpful for high-score borrowers, increased formality could decreases the fundability of low-score borrowers.

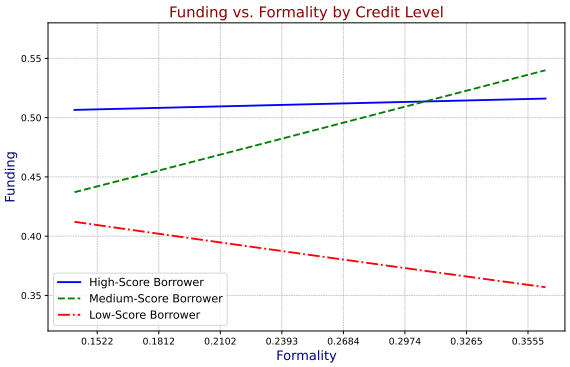


(a) Text Length

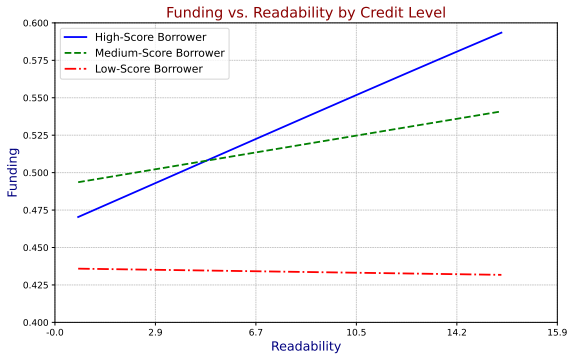
(b) Spelling Errors



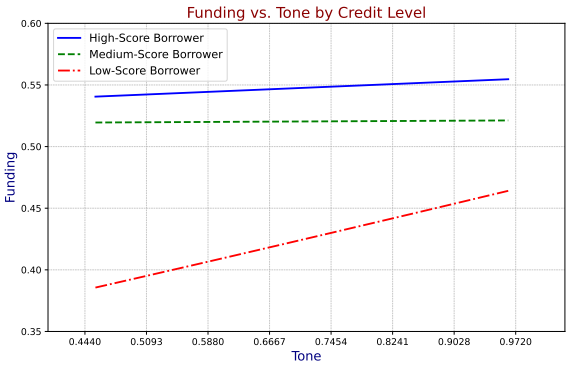
(c) Grammatical Errors



(d) Formality



(e) Readability



(f) Tone

Figure A2: Textual Features and Fundability.

Appendix F Robustness Tests

F.1 Prompt Design and Textual Features

Based on the previous setup, we tested the changes in textual features across different prompt templates, as presented in Table A9.

Table A9: Textual Features Before and After Rewrite

	Original	P1R		P2R		P3R		P4R	
	Mean (SD)	Mean (SD)	DIF* (%)	Mean (SD)	DIF* (%)	Mean (SD)	DIF* (%)	Mean (SD)	DIF* (%)
Text Length	202.5 (132.2)	240.2 (109.7)	37.7*** (+18.6%)	248.1 (99.35)	45.6*** (+22.5%)	292.5 (90.54)	90.0*** (+44.4%)	281.1 (99.44)	78.6*** (+38.8%)
Spelling Errors	0.035 (0.021)	0.009 (0.008)	-0.026*** (-74.3%)	0.009 (0.009)	-0.026*** (-74.3%)	0.008 (0.007)	-0.027*** (-77.1%)	0.009 (0.007)	-0.026*** (-74.3%)
Grammatical Errors	0.064 (0.041)	0.035 (0.030)	-0.029*** (-45.3%)	0.038 (0.030)	-0.026*** (-40.6%)	0.026 (0.021)	-0.038*** (-59.4%)	0.028 (0.023)	-0.036*** (-56.3%)
Formality	0.239 (0.058)	0.342 (0.042)	0.103*** (+43.1%)	0.345 (0.039)	0.106*** (+44.4%)	0.349 (0.033)	0.110*** (+46.0%)	0.367 (0.035)	0.128*** (+53.6%)
Readability	10.45 (3.775)	14.71 (2.664)	4.26*** (+40.8%)	14.78 (2.115)	4.33*** (+41.4%)	15.26 (1.502)	4.81*** (+46.0%)	15.75 (1.684)	5.30*** (+50.7%)
Tone	0.745 (0.157)	0.773 (0.146)	0.028*** (+3.8%)	0.775 (0.145)	0.030*** (+4.0%)	0.753*** (0.144)	0.008*** (+1.1%)	0.754 (0.144)	0.009*** (+1.2%)

[a] Sig. levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table A9 indicates that while ChatGPT's impact on textual features varies across different prompt templates, the improvements remains consistent. Corroborating findings in Section 3.1, different prompt versions (as nuanced variations of ChatGPT adoption) enhance length compliance, tone positivity, formality, and readability, while reducing grammatical and spelling errors.

F.2 Prompt Design and the WQI

Since the WQI is calculated based on the cosine similarity of loan descriptions before and after ChatGPT rewriting, we analyzed the WQI under different prompt templates, with the results shown in Table A10.

From Table A10, we observe that although there are differences in prompt templates, the variation in WQI across different different groups remains consistent. Specifically, we find that low-credit borrowers have higher WQI, while high-credit borrowers have lower WQI, with this difference being particularly noticeable for unfunded loans. Additionally, we observe that funded

Table A10: WQI by Credit Level and Funding Outcome

	P1R			P2R			P3R			P4R		
	U	D	R	U	D	R	U	D	R	U	D	R
High	0.736	0.771	0.769	0.745	0.774	0.772	0.734	0.768	0.763	0.726	0.764	0.759
	(0.009)	(0.006)	(0.007)	(0.008)	(0.006)	(0.006)	(0.008)	(0.006)	(0.006)	(0.009)	(0.006)	(0.006)
Medium	0.790	0.786	0.783	0.790	0.788	0.787	0.779	0.776	0.773	0.779	0.773	0.770
	(0.005)	(0.006)	(0.007)	(0.005)	(0.006)	(0.006)	(0.005)	(0.007)	(0.007)	(0.005)	(0.006)	(0.007)
Low	0.796	0.795	0.788	0.799	0.798	0.792	0.803	0.790	0.781	0.790	0.784	0.776
	(0.004)	(0.008)	(0.009)	(0.004)	(0.007)	(0.008)	(0.004)	(0.008)	(0.009)	(0.004)	(0.008)	(0.009)

[a] U: Unfunded Loans; D: Default Loans; R: Repaid Loans.
[b] Values in parentheses represent standard deviations.

high-credit borrowers have significantly higher WQI compared to their unfunded counterparts, which implies that some high-credit borrowers may undervalue the importance of writing loan descriptions. Finally, we find that within funded loans, the WQI difference between repay and default loan statuses is relatively small, which potentially suggests that WQI is more influenced by funding status.

Appendix G Technical Information

G.1 Architectures for Text Embeddings

Bidirectional Encoder Representations from Transformers (BERT) was first introduced in Devlin et al. (2019). As a transformer model, BERT is based on the Transformer architecture (Vaswani et al., 2017), specifically its encoder part. The transformer employs a self-attention mechanism to process input data, enabling the model to consider the relationships between all elements in the input sequence.

The basic principles of BERT are as follows. Bidirectional Contextual Understanding, traditional language models are unidirectional, unable to consider both the preceding and following context of a word. BERT innovatively used bidirectional context, considering the context on both sides of each word to obtain richer language representations. Pre-training and Fine-tuning, BERT employs a two-stage training process. First, it is pre-trained on large-scale text data to learn a general language representation; then it is fine-tuned on task specific data to adapt the model for specific NLP tasks.

The pre-train process of BERT relies on two key techniques: Masked Language Model (MLM) and Next Sentence Predict (NSP). The MLM sets the objective to maximize the likelihood of the masked tokens. For a masked token, w_m, the objective can be formulated as:

$$L_{MLM} = -\sum_{i \in M} \log P(w_m | context) \tag{19}$$

where $context$ denotes unmasked tokens. $P(w_m|context)$ is the probability of the masked word given the context provided by the other words. The objective of NSP is to model the probability of the second sentence (B) being the next sentence. The formula can be represented as

$$L_{NSP} = -(y \log(P)) + (1 - y) \log(1 - P) \tag{20}$$

where L_{NSP} is the loss for the NSP task. P is the model's predicted probability that sentence B follows sentence A, and y is the actual label (1: if B follows A; 0: otherwise).

SBERT modifies the BERT architecture to produce embeddings directly comparable using co-

sine similarity, which is more efficient for tasks like semantic search or similarity assessment. SBERT uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine similarity. These design ensure that sentences with similar meanings are close in vector space, while those with different meanings are further apart. Additionally, in the training process, the triplet loss was used to improve the learning efficiency. The formula of triplet loss is:

$$L = \sum_{i=1}^{N} \max(0, \alpha + d(a_i, p_i) - d(a_i, n_i)), \tag{21}$$

where $d(x, y)$ is a distance metric function to calculate the similarity between the embeddings of sentences x and y, α is the margin, a_i, p_i, and n_i are the anchor, positive, and negative sentences in the i-th triplet, respectively.

G.2 Multimodal Transformer

To make predictions using both structured features and textual features, we leverage a multimodal transformer model (Gu and Budhkar, 2021), illustrated in Figure A3.

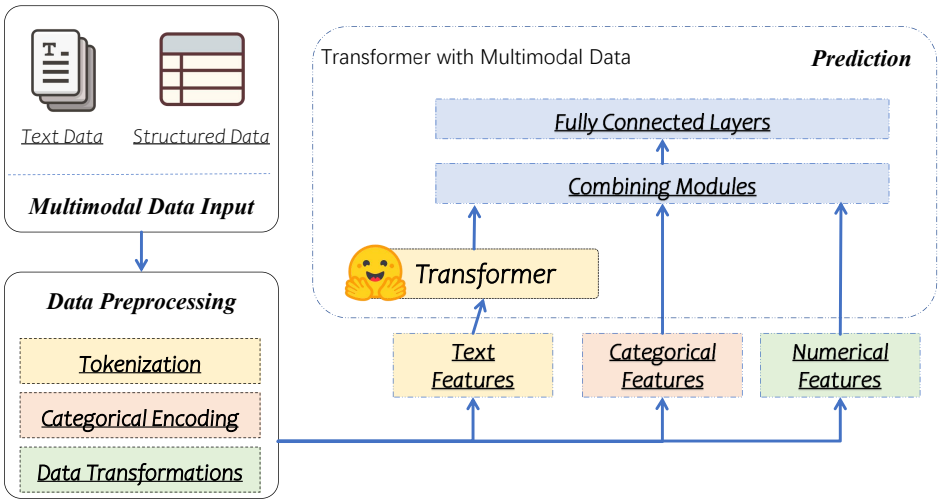


Figure A3: Architecture of the Multimodal Transformer

Depending on the data and label being used, this model may be used to classify funding status

(among all listings) or repayment status among those funded.

G.3 DeepHeckman Model Estimation

Our Heckman-based lender model consists of two versions of this multimodal transformer, one for the selection model f, and the other for the prediction model g (see Section 4.1 for a high-level description of its formulation, rationale, and the loss function). Following Kahng et al. (2023), we adopt a two-step optimization approach. The pseudocode is provided in Algorithm 1.

Algorithm 1: Two-Step Optimization for DeepHeckman

Data: Data $\mathcal{D} = \{(\mathbf{x}_i, y_i, s_i) \in X \times Y \times S, i = 1, \ldots, N\}$, Batch size b, Learning rate γ.

Result: $\hat{f}$, $\hat{g}$, $\hat{\Sigma}$

Randomly initialize $\hat{f}$, $\hat{g}$, $\hat{\Sigma}$;

`/* Step 1: Learn the Selection Model` $\hat{g}$ `*/`

while $\hat{g}$ *not converged* **do**

  $\mathcal{B} \leftarrow BatchSampler(\mathcal{D}, b)$;

  $\hat{g} \leftarrow \hat{g} + \frac{\gamma}{b} \nabla \Sigma_{(\mathbf{x}_i, s_i) \in \mathcal{B}} \left[s_i \log \Phi(\hat{g}(\mathbf{x}_i)) + (1 - s_i) \log \Phi(-\hat{g}(\mathbf{x}_i)) \right]$;

end

`/* Step 2: Learn the Outcome Model` $(\hat{f}, \hat{\Sigma})$ `*/`

while $\hat{f}$ *not converged* **do**

  $\mathcal{B} \leftarrow BatchSampler(\mathcal{D}, b)$;

  $\hat{f} \leftarrow \hat{f} - \frac{\gamma}{b} \nabla \Sigma_{(\mathbf{x}_i, y_i, s_i) \in \mathcal{B}} \left[s_i \Lambda \left(\hat{f}(\mathbf{x}_i), \hat{g}(\mathbf{x}_i); y_i, s_i; \hat{\Sigma} \right) \right]$;

  $\hat{\Sigma} \leftarrow \hat{\Sigma} - \frac{\gamma}{b} \nabla \Sigma_{(\mathbf{x}_i, y_i, s_i) \in \mathcal{B}} \left[s_i \Lambda \left(\hat{f}(\mathbf{x}_i), \hat{g}(\mathbf{x}_i); y_i, s_i; \hat{\Sigma} \right) \right]$;

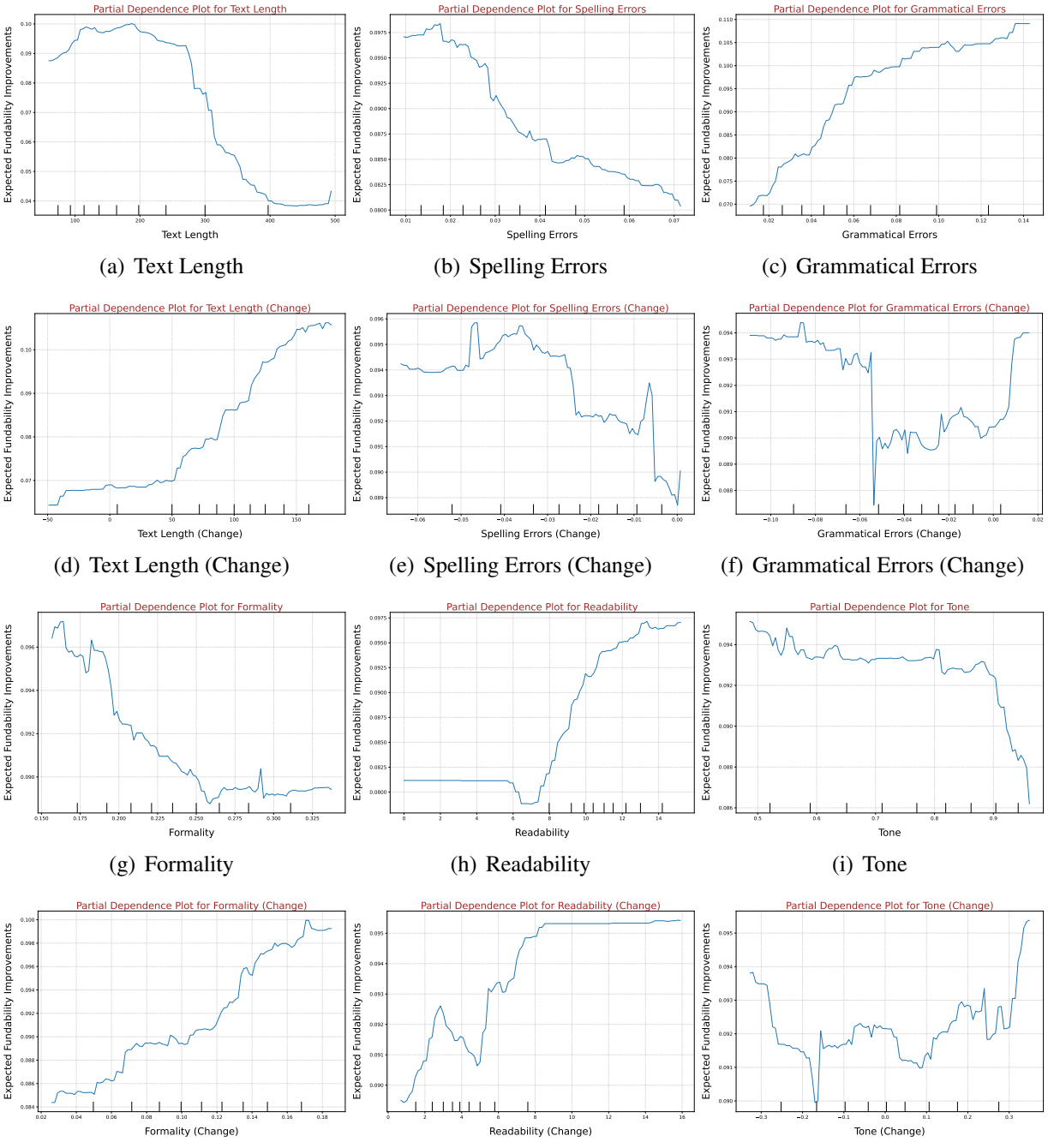
end

Appendix H Linking Textual Features to Fundability

We examine the relationship between features and fundability improvements using SHAP-based explanations (Lundberg and Lee, 2017) based on the XGBoost model as described in the attribution test presented in Section 4.3. Focusing on baseline textual features (those of the original text) and their change values (i.e., the difference in textual features between rewrite and the original text), we present the partial dependence plot in Figure A4.

From Figure 4(a), we can see that shorter baseline text length corresponds to greater improvement in fundability. Proportionately, larger changes in text length made by ChatGPT is linked to greater improvement in fundability, as shown in Figure 4(d). This suggests that when the baseline text length is low, there is more room to make changes; and since text length is positively linked to fundability, larger increase in text length is linked to a larger increase in fundability. As we can see in Figure 4(g) and Figure 4(j), a similar pattern can be found in formality.

Similarly, higher percentages of baseline spelling and grammatical errors provide greater opportunities for correction, leading to significant increases in fundability. In terms of readability, Figure 4(h) shows that improvements are consistently linked to positive impacts on fundability. For tone, however, a U-shaped relationship with fundability is observed. This phenomenon carries two implications. First, for loan descriptions with low tone, increasing the positivity leads to fundability improvements. Second, for loan descriptions with excessively high tone, reducing the level of positivity similarly enhances fundability. This observation suggests that the tone of loan descriptions should be balanced, avoiding extremes of excessive positivity.



(a) Text Length (b) Spelling Errors (c) Grammatical Errors

(d) Text Length (Change) (e) Spelling Errors (Change) (f) Grammatical Errors (Change)

(g) Formality (h) Readability (i) Tone

(j) Formality (Change) (k) Readability (Change) (l) Tone (Change)

Figure A4: Textual Features, Changes and Fundability.

Appendix I Performance Benchmarking

I.1 Performance of Traditional Classifiers

Following Netzer et al. (2019), we consider regularized logistic regressions (GLM), random forest (RF), and extremely randomized trees (ERT), as well as an emsemble model consisting of these base models. More detailed description of these methods can be found in Netzer et al. (2019). We also include light gradient-boosting machine (LGBM), an algorithm renowned for its effectiveness in loan evaluation (Zhang et al., 2023), as one of the base models. For machine learning classifiers. We use structured features, explicit textual features, and bigram features in a setting similar to Netzer et al. (2019). The Ensemble Model was constructed by stacking the above models, each with an using the aforementioned models.

Table A11: Performance Benchmarking by Predicting Default Among Funded Loans

	AUC[c]			
	Model 1 (Tabu Only)	Model 2 (Text Only)	Model 3 (Tabu+Text)	Soft Information
Ensemble	**72.5%**	64.6%	73.7%	1.2%
GLM Only	71.4%	63.8%	71.4%	0.0%
RF Only	72.3%	63.0%	73.0%	0.7%
LGBM Only	72.3%	63.6%	73.3%	1.0%
ERT Only	71.7%	63.0%	72.3%	0.6%

[a] Ensemble = an Ensemble model of all; GLM = Generalized linear model; RF = Random forest; LGBM = Light gradient boosting machine; ERT = Extremely randomized trees (extra trees).
[b] BERT and RoBERTa are designed for text data, therefore, they do not have a result for Model 1. While BERT and RoBERTa are large language models designed for text, when we combine them with structured data, we employed the multimodal transformer architecture (Gu and Budhkar, 2021).
[c] Note: All AUCs are averaged across 10 replications of the ten-fold mean.

In online credit markets, Netzer et al. (2019) applied several advanced machine learning models and AUC to evaluate the informational value of text data. Building on their research, we used both machine learning models, such as generalized linear models (GLM) and random forests (RF), and deep learning models, including BERT and RoBERTa, to evaluate the value of soft information in default prediction within online credit markets. The results are shown in Table A11. For these machine learning models, soft information is captured through bi-grams, whereas in deep learning models (Netzer et al., 2019), it is obtained through text embeddings (Vaswani et al., 2017).

In Table A11, Model 1 is based solely on structured features, Model 2 uses only text, while Model 3 combines both structured features and text as predictors. For BERT and RoBERTa, since they do not have a Model 1 result, we use the best performance of machine learning models (i.e.,

ensemble) to impute it. The value of soft information is quantified by the performance variation from Model 1 to Model 3. In general, soft information improved the performance of loan default prediction models, although there were variations across different credit levels. This pattern confirms that text has predictive value, which is even larger when using deep learning models.

I.2 Hyperparameters for Machine Learning Models

In this section, we document the hyperparameters and optimization methods of the machine learning models involved in our study. This final estimator is trained through cross-validation. We employ Bayesian methods for the optimization of hyperparameters (Nogueira, 2014). All these models were implemented using sklearn, a highly renowned Python package extensively used in the field of machine learning.[20]. Details of the parameters are shown in Table A12.

Table A12: Hyperparameter Settings for Classifiers

Model	Parameter	Description	Range
GLM	C	Inverse of regularization strength; must be a positive float. Smaller values specify stronger regularization.	(0.01, 10)
	penalty	Specify the norm of the penalty.	['l1', 'l2']
	solver	Algorithm to use in the optimization problem.	[0, 1]
Tree-Based[a]	n_estimators	The number of trees in the forest.	(10, 1000)
	max_depth	The maximum depth of the tree.	(5, 30)
	max_features	The number of features to consider when looking for the best split.	(0.1, 0.999)
	min_samples_leaf	The minimum number of samples required to be at a leaf node.	(1, 25)
	min_samples_split	The minimum number of samples required to split an internal node.	(2, 25)
LGBM	n_estimators	The number of trees.	(10, 1000)
	max_depth	The maximum depth of the tree.	(5, 30)
	learning_rate	The rate of learning of the model.	(0.01, 0.3)
	min_child_samples	Minimum number of data needed in a leaf.	(5, 50)
	num_leaves	The number of leaves in a tree.	(24, 45)

[a] Tree-based models include RF and ERT; deep learning (DL) models include BERT and RoBERTa.

[20] https://scikit-learn.org/stable/index.html

I.3 Hyperparameters for BERT

When considering textual features such as BERT, we directly add a multi-layer perceptron (MLP) layer with cross entropy loss to map numeric vector representations of text into the outcome class label. We implemented them with Pytorch[21] and Huggingface.[22]

Due to random initialization, we ran the same model setting 15 times and report results based on the average. We adopted grid search to find the best parameters for BERT. Details of the parameters are presented in Table A13.

Table A13: Hyperparameter Settings for Deep Learning Model (BERT).

Parameter	Description	Range
learning_rate_text	The learning rate of the pre-trained language models.	[3e-6, 3e-5]
learning_rate_tabu	The learning rate of the MLP model, which employed to learn the features of structured features if activated.	[0.01, 0.015]
weight_decay_text	The regularization part of the language model.	[5e-3, 5e-4, 5e-5]
weight_decay_tabu	The regularization part of the MLP model.	[1e-1, 1e-2, 1e-3]
epoch	Train epochs in each round.	[6, 7, 8]
round	Train round for each group of parameters.	[15]
batch_size	Batch size.	[15, 30]
dropout_rate	The dropout rate of models.	[0.45]
activate_func	Activate functions.	['relu','leaky_relu']

[21] https://pytorch.org/
[22] https://huggingface.co/