Enhancing Credit Risk Reports Generation using LLMs: An Integration of Bayesian Networks and Labeled Guide Prompting

Ana Clara Teixeira*†
ana.teixeira@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Vaishali Marar†
vaishali.marar@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Hamed Yazdanpanah
hamed.yazdanpanah@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Aline Oliveira
aline.oliveira@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Mohammad Ghassemi
mohammad.ghassemi@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

ABSTRACT

Credit risk analysis is a process that involves a wide range of complex cognitive abilities. Automating the credit risk analysis process using Large Language Models can bring transformative changes to the finance industry, but not without appropriate measures to ensure trustworthy responses. In this work, we propose a novel prompt-engineering method that enhances the ability of Large Language Models to generate reliable credit risk reports - Labeled Guide Prompting (LGP). LGP consists of: (1) providing annotated few-shot examples to the LLM that denote sets of tokens in an exemplary prompt that are of greater importance when generating sets of tokens in the exemplary response and (2) providing text in the prompt that describes the direction, pathways and interactions between variables from a Bayesian network used for credit risk assessment, thus promoting abductive reasoning. Using data from 100 credit applications, we demonstrate that LGP enables LLMs to generate credit risk reports that are preferred by human credit analysts (in 60-90% of cases) over alternative credit risk reports created by their peers in a blind review. Additionally, we found a statistically significant improvement (p-value < 10^{-10}) in the insightfulness of the responses generated using LGP when compared to identical prompts without LGP components. We conclude that Labeled Guide Prompting can enhance LLM performance in complex problem-solving tasks, achieving a level of competency comparable to or exceeding human experts.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**;
• **Applied computing** → *Decision analysis.*

KEYWORDS

GPT-4, prompt engineering, credit risk report, Bayesian network, labeled guide prompting

ACM Reference Format:
Ana Clara Teixeira, Vaishali Marar, Hamed Yazdanpanah, Aline Oliveira, and Mohammad Ghassemi. 2023. Enhancing Credit Risk Reports Generation using LLMs: An Integration of Bayesian Networks and Labeled Guide Prompting. In *4th ACM International Conference on AI in Finance (ICAIF '23), November 27–29, 2023, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3604237.3626902

1 INTRODUCTION

Credit risk analysis is a complex process that involves a wide range of abilities, including contextual understanding, logical reasoning, the application of domain-specific knowledge, and implicit and causal reasoning. Evaluating the performance of Large Language Models (LLMs) in credit risk assessment provides crucial insights into their practical utility in real-world scenarios. This work introduces the application of LLMs for generating comprehensive credit risk reports - a critical task in financial decision-making. More specifically, we propose a novel prompt-engineering approach designed to enhance the quality and fidelity of credit risk assessments; we compare the credit risk assessments of the LLM against human analysts through a user-centered, human-based evaluation, demonstrating the proposed procedure's efficacy in dealing with the credit risk assessment task.

Automating the credit risk analysis process can bring transformative changes to the finance industry. By leveraging LLMs, we can streamline processing vast amounts of data, enabling real-time analysis, surpassing traditional methods. Furthermore, automation can also contribute to more consistent and objective analyses; while human analysts might be influenced by biases or varying expertise levels, using LLMs for credit risk assessment can maintain a standard level of analysis based on its prompt engineering infrastructure. Thus, the importance of automated credit risk assessment is due to the growing influence of automation and artificial intelligence in the finance sector. By evaluating the capabilities of LLMs for credit risk analysis, we offer valuable insights to inform AI integration in credit risk management. Considering the enormous volume of credit decisions made regularly, even marginal enhancements in

*Corresponding author
†These authors contributed equally.

ICAIF '23, November 27–29, 2023, Brooklyn, NY, USA

ACM ISBN 979-8-4007-0240-2/23/11.
https://doi.org/10.1145/3604237.3626902

LLM performance can translate into significant efficiency gains and better decision accuracy.

The task of using GPT-4 for credit risk analysis is highly complex due to several factors: (1) Credit risk analysis is inherently multifaceted, requiring the consideration of numerous variables ranging from personal credit history and current financial status, to broader economic conditions and industry-specific factors; (2) the dynamic nature of financial data, which is characterized by continuous change and occasional volatility. This dynamic nature makes interactions among variables more relevant and leads to evolving relationships between features and credit risk, and (3) there are underlying difficulties in utilizing a language model like GPT-4 for tasks traditionally performed by credit analysts, professionals with a deep understanding of financial dynamics and risk. These issues require bridging the gap between generalized data analysis, a strength of GPT-4, and credit analysts' specialized knowledge and intuition, adding another layer of complexity. Generating comprehensive credit risk reports tests the limits of current LLMs such as GPT-4, combining domain knowledge, contextual understanding, and multiple types of reasoning. While some elements may be solvable using existing language models, the comprehensive nature of the task is more demanding. The effectiveness of GPT-4 in generating credit risk reports depends on its capacity to grasp and incorporate this nuanced knowledge and intuition, which requires a sophisticated prompt engineering strategy.

One of the main limitations in the utilization of GPT and current prompt engineering strategies is the unpredictable nature of responses when handling unseen data, anomalies, or changing requirements. This unpredictability is a significant concern in financial decision-making, where consistent quality of response, insight, and analysis is essential.

In this paper, we propose a unique prompt engineering strategy that standardizes the content and quality of GPT-4's output, making it more predictable and insightful, even when handling complex, dynamic tasks like credit risk analysis. The unique strength of this approach lies in its potential to act as a 'missing piece' in prompt engineering; it changes how we interact with LLMs, by ensuring that the quality of insights and analyses remains consistent, even when dealing with complex, dynamic tasks like credit risk analysis. Moving beyond traditional applications of LLMs, our research exposes these models to the financial industry's requirements, successfully meeting the pragmatic needs of credit analysis.

1.1 Related Work

1.1.1 LLM milestones. In recent years, language model development has seen substantial advancements in architecture, training methods, and real-world applications. The Transformer model pioneered the use of self-attention mechanisms, leading to significant improvements in many natural language processing (NLP) tasks [14]. Then, BERT revolutionized language understanding by training bidirectional transformers, allowing models to understand the context of a word based on all of its surroundings (left and right of the word) [4].

The advent of GPT-3 introduced scaling laws, which confirms that as models increase in size, they continue to improve in performance [1, 2, 13]. It showcased impressive few-shot learning capabilities, enabling the model to generalize from a small number of examples. Our work uses the GPT-4 model, which extends these principles and represents the latest milestone in this progression.

LLMs have demonstrated competitive performance in machine translation [5], summarization [21], dialogue generation [20], and writing human-like essays [8]. BloombergGPT is an example of a domain-specialized LLM, exhibiting superiority to its generalist counterparts in broad financial tasks such as market sentiment analysis [19]. Diverging from this, our work uniquely applies GPT-4 to individualized credit risk assessments, focusing on the evaluation of specific borrowers.

1.1.2 Prompt engineering. Prompt engineering and few-shot learning have become crucial components of effective utilization of LLMs, guiding these models towards more desirable and task-specific outputs [3, 18]. Prompt engineering involves crafting carefully structured inputs that elucidate the context and desired outcome of a task for the model, proving instrumental in narrowing down the model's broad knowledge to a specific task [17]. On the other hand, few-shot learning involves presenting the LLM with a small number of examples of a task, thereby assisting the model in identifying and applying the correct pattern for the given task. Our work builds on these methods to develop a prompt engineering procedure for credit risk assessment with GPT-4.

1.1.3 Task definition and benchmarks. Formulating tasks for LLMs and evaluating their performance are two fundamental steps in applying these models to real-world scenarios. Task formulation involves defining a clear objective, often expressed as a question or statement for the model to complete, expand, or react [10].

Evaluation, meanwhile, usually entails matching the model's output against a standard or specific criteria [7]. Widely accepted benchmarks such as GLUE [15], SuperGLUE [16] and BIG-bench [11, 12] serve to evaluate the model's competency across various areas. For our study, we design a task for GPT-4 focused on credit risk assessment and measure its performance using well-defined criteria specific to credit risk evaluation.

1.1.4 Abductive reasoning with LLM. Abductive reasoning involves generating the most plausible explanation for a given set of observations and aligns well with the prediction-focused nature of many AI applications. By enabling LLMs to not only predict but also to generate plausible explanations, we can enhance the interpretability of these models [6, 9].

In the context of our work, leveraging abductive reasoning allows the model to generate more informative and nuanced credit reports. For example, it can create insights into how different financial indicators might interact to influence an individual's credit risk score. This makes the assessment transparent and understandable, a significant advantage in fields where interpretability is crucial.

We propose a novel prompt engineering technique, *Labeled Guide Prompting*, designed to aid the LLM in responding to multiple complex dimensions of a problem such as the "what?", "why?" and "how?". This method ensures that the response requirements of a task, however complex are strictly followed by GPT, as the labeled guide assists the LLM in answering all concepts and putting equal thought into each by returning a standard response given a

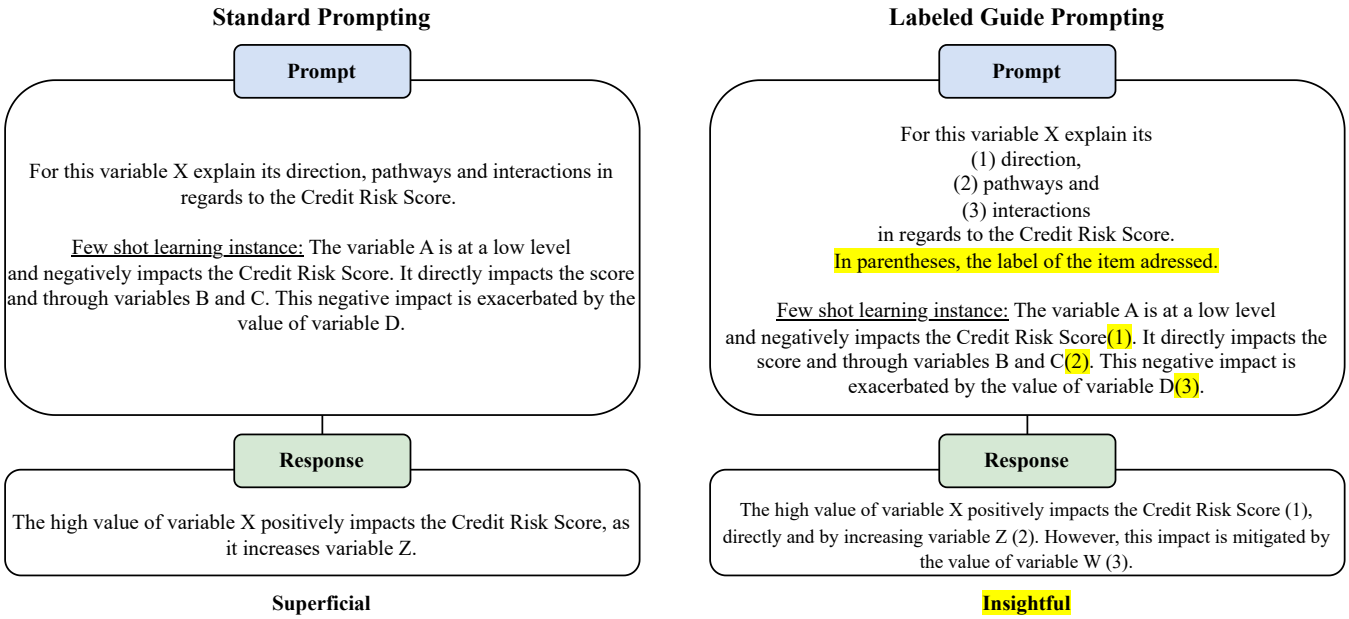

Figure 1: Example of how Labeled Guide Prompting leads to more insightful responses, by identifying each sub-task as a particular item to be addressed.

prompt. By combining this method with Bayesian networks representation, we promotes abductive reasoning in Large Language Models (LLMs). This approach makes the model to hypothesize or generate the "best guesses" that explain the observed phenomena, according to the network structure.

Explicit labeling of output elements ensures control over responses, facilitating the production of a specific, well-structured output. Consequently, this technique yields logically coherent and insightful responses, enhancing LLM's efficacy in complex problem-solving tasks.The labeled guide aids the LLM in guaranteeing differentiation between sub-tasks, therefore encouraging specificity and a more insightful response given that there is clear separation and purpose for each dimension of the problem at hand. The labeled guide serves as a rubric for the LLM, requiring the LLM to review the quality of response, format and content as the sub-tasks of the task are being addressed.

In our work, we apply Labeled Guide Prompting to guide GPT-4 in generating structured and insightful credit risk reports. We found that this technique effectively increased the frequency and quality of interactions in reports, thus enhancing their insightfulness and usefulness in the context of credit risk assessment.

In this study, we extend the application of LLMs (GPT-4) to the domain of credit risk assessment in agricultural finance, presenting a novel approach for leveraging LLMs in the financial sector.

1.2 Main contributions

Below we summarize our main contributions. These findings bear significant implications for the future integration of LLMs in credit risk assessment and other financial risk evaluation tasks. The successful application of our approach establishes a solid foundation for further research and practical implementations in this field.

1.2.1 LLMs for Credit Risk Reports. : We introduce an innovative application of Large Language Models (LLMs), specifically GPT-4, for the generation of high-quality credit risk reports. We have developed a principled procedure for prompt engineering tailored to the specifications of credit risk reports. This innovation significantly expands the repertoire of LLM applicability, demonstrating their potential in tasks requiring specialized knowledge and precision.

Traditionally crafted by human analysts, our study validates the capacity of LLMs to effectively replicate and even enhance credit risk assessment. In an experimental setup involving 100 credit applications, independent evaluators showed a statistically significant preference for GPT-4 generated reports. This validation of our method confirms the ability of LLMs to effectively automate the generation of credit risk reports.

1.2.2 Integration of Bayesian Networks. : Our approach integrates Bayesian networks to guide the LLM's reasoning tasks. The use of this graphical model assists the LLM in understanding the influence of each variable on the target, aligning with the network structure, and producing plausible scenarios, promoting abductive reasoning.

1.2.3 Novel Prompt Engineering Technique. : Our technique, *Labeled Guide Prompting*, was effective in ensuring that the LLM accurately addresses the specific aspects of credit risk analysis. This not only improved the quality of the generated reports but also demonstrated the broad potential of this technique in other sophisticated tasks. The LLM, post prompt-engineering, showed a statistically significant improvement in the insightfulness of the

reports, as measured by well-defined metrics. This underscores the potential of our approach to enhance the performance of LLMs in specialized tasks.

2 METHOD

The primary objective of this research is to apply an LLM, specifically GPT-4, to generate credit risk reports that surpass the quality of traditional reports written by human analysts. This strategy involves a novel prompt engineering procedure for GPT-4, designed to adhere to the content specifications of the reports. The specific task involves generating comprehensive credit risk reports that quantify and explain credit risk in terms of the probability of delinquency and corresponding score, based on the borrower's information derived from the credit application. The overarching goal is to provide solid support for credit decisions, which means either approval or rejection. Accordingly, the report should provide a thorough explanation of the impact of all risk factors to ensure completeness, and it should offer a decisive assessment of the borrower's risk profile to guarantee insightfulness.

To assess the quality and effectiveness of these AI-generated credit risk reports, an experiment involving 100 credit applications was conducted. Each application was evaluated through two separate reports - one generated by a team of human credit analysts, and the other produced by GPT-4 using the proposed method in this paper. These reports were designed to provide comprehensive explanations of the probability of delinquency, taking into account all the input information provided. Subsequently, an independent team of credit analysts blindly reviewed the reports and chose the most helpful one for making a credit decision. This experiment enabled a human-based and user-centered evaluation of the GPT-4-generated reports, ensuring a practical assessment of the efficacy of the proposed approach.

2.1 Data and experimental design

2.1.1 Data source. The credit application data for this research were obtained from a financial institution's clients, a representative set of lenders within Brazil's agricultural supply chain. These applications, filled out by farmer borrowers, contain critical information necessary for a thorough credit risk assessment in an agricultural context. This risk assessment task is handled by the institution: the institution processes the credit application and returns a probability of delinquency alongside a corresponding credit score for each application. The human-generated (HG) reports based on these assessments were then created by the institution's credit analysts, professionals with extensive experience in agricultural finance credit risk evaluation.

2.1.2 Comparison of human and LLM generated reports. To assess the utility and quality of the credit risk reports, an independent evaluation was conducted by a team of seven credit analysts external to the institution. These analysts, industry professionals from Brazil's agricultural finance sector, were blind to the sources of the reports and unaware that one was LLM-generated (LLM-G). This blind-review approach ensured an unbiased comparison of the reports based on their practical utility in credit risk assessment.

The evaluators, who had no contact with each other, were not aware that the comparison's purpose was to compare HG and LLM-G reports. They simply believed they were comparing two different reports to determine which would be more useful in making a credit decision. Six evaluators reviewed reports in Portuguese, while one proficient in English reviewed reports in that language. They had three days to assess the assigned cases.

The evaluation process consisted of a pair of credit reports for each credit application and a questionnaire for evaluators to complete. The questionnaire consisted of three questions. In the questionaire, 'A' corresponded to the report generated by human (HG) and 'B' corresponded to the LLM-generated (LLM-G) report:

(1) Question Text: "Which report was more helpful for you to assess the credit risk?"
Response Form: choose "A", "B", or "Both are equally helpful"

(2) "Question Text: Do any of the reports have any information that is not true or does not make sense?"
Response Form: choose "A", "B", "Both", or "None"

(3) Question Text: "Why do you prefer the chosen report?"
Response Form: free-text

The design of this questionnaire ensured an objective evaluation of the reports, while the open-ended question provided insight into the evaluators' reasoning behind their choices.

2.2 Integrating Bayesian Networks into the LLM

In generating credit risk reports, we leverage a probabilistic graphical model, specifically a Bayesian network, to estimate the probability of delinquency. Our choice for a Bayesian network is motivated by two primary considerations. Firstly, Bayesian networks provide robust predictive performance. Secondly, their network structure and computation of a joint probability enhance the interpretability of the model output, enabling us to identify pathways of influence and interactions among variables.

The Bayesian network structure offers a framework to engage the LLM with complex tasks in credit risk reports. This structure facilitates reasoning across multiple dimensions, as follows:

(1) **Common-Sense Reasoning:** Informative variable names and a sensible structure dissuade the LLM from generating nonsensical explanations.
(2) **Logical-Deductive Reasoning:** The model proposes a general rule, prompting the LLM to deduce specific cases that align with this rule.
(3) **Logical-Abductive Reasoning:** Factorization of the joint distribution according to the network helps identify plausible scenarios - both pathways and interactions - given the information at hand.
(4) **Implicit Reasoning:** The network paths offer a natural means of decomposing relationships into intermediate steps.
(5) **Causal Reasoning:** The structure of Bayesian networks enables the identification and interpretation of potential causal relationships.

In deploying this approach, it is essential to ensure that the relationships between each variable and the target are intuitive. We rely

on the LLM's capacity to understand and predict the relationship between each variable and the target, predicated on the variable's name. We achieve this through the setting of appropriate priors and an analysis of how the target estimate fluctuates in response to changes in the values of the features.

2.3 Generating Credit Risk Reports

The specific task of generating credit risk reports aims to assess the LLM's ability to produce a proficient credit risk report. This specific task challenges the LLM in the following areas:

(1) **Contextual Question-Answering:** The LLM should accurately interpret the comprehensive instruction, which includes explaining the credit risk score based on the Bayesian network's structure, input, and output.
(2) **Domain-Specific Tasks:** The LLM needs to demonstrate proficiency in agricultural finance, credit risk, and machine learning to effectively produce a relevant report.
(3) **Decomposition:** The task of explaining the impact of each feature on the target should be broken down into simpler steps by the LLM, aligning with the Bayesian network structure.
(4) **Common-Sense Reasoning:** The LLM must comprehend the phenomenon that the Bayesian network describes and create plausible scenarios given the feature values to justify the estimated target.
(5) **Logical Reasoning:** Beyond generating logically coherent statements, the LLM should:
   - Deduce the impact of the features on the target in line with the Bayesian network structure, and
   - Engage in abductive reasoning to provide the most plausible explanation of how the inputs yield the model's output.
(6) **Implicit Reasoning:** The LLM is charged with illuminating steps implicit in the network structure, explaining how a feature impacts the target through various pathways or interactions.
(7) **Causal Reasoning:** The LLM should identify potential causal relationships between the features and the target variable as suggested by the Bayesian network structure.

The primary criterion of our evaluation is human-based and user-centric, where credit analysts judge the utility of the reports. Alongside this, we use the blind review's results to objectively assess the LLM's performance on the general tasks. This approach ensures a robust, comprehensive assessment of the LLM's capability to generate credit risk reports, striking a balance between subjective user experience and objective task performance measures.

2.4 Prompt Components

The prompt engineering design used in this research is constructed around several key components. These components guide the LLM in generating comprehensive and insightful credit risk reports. These components are listed below.

- **Role and Instruction:** The role situates the LLM as a data scientist expert in agricultural finance. This positioning provides the context and performance expectations for the LLM. The instruction propels the LLM to create an in-depth credit risk report that explains the probability of delinquency and the corresponding credit risk score, by taking into account both the provided data inputs and the underlying network structure.
- **Network Structure:** This component denotes the structure of the Bayesian network, a graphical illustration of the causal relationships among different variables. Table 1 shows how it equips the LLM with a roadmap to comprehend and reason about the dependencies between various factors, and how they contribute to the credit risk.
- **Model Output:** This is the computed credit risk score, which forms the main subject of the report. The LLM is expected to elaborate on it extensively, given the inputs and the structure of the Bayesian network.
- **Items to be Addressed:** These detail the elements that the LLM should emphasize in the report. These aspects stem from the Bayesian network's learning and inference process, thereby enabling the LLM to carry out reasoning tasks with greater accuracy. The items to be addressed are:
  (1) The direction of the impact of each model input on the output (does this model input increase or decrease the output?)
  (2) The network pathway from input to output (how does this model input impact the output, considering the network structure?)
  (3) Input interactions affecting the output (how does this model input interact with neighboring inputs to impact the output?)
- **Few-shot Learning Instances:** These are samples of completed tasks that the LLM can learn from. By generalizing from these instances to the present task, the LLM can understand the structure, style, and content of an effective credit risk report.

Table 1: How to represent the Bayesian network structure as a prompt component.

Network Structure
"The Bayesian network structure is as follows: - Variable 1 is connected to Variables 2, 3, 4, and 5 (the output). - Variable 2 is connected to Variables 3 and 5 (the output). - Variable 3 is connected to Variables 4 and 5 (the output). - Variable 4 is connected to Variables 6 and 5 (the output). - Variable 7 is connected to Variable 4. - Variable 8 is connected to Variables 9 and 5 (the output). - Variable 9 is connected to Variable 5 (the output)."

2.5 Labeled Guide Prompting

To ensure that the LLM retrieves pertinent information and reasons in a prescribed manner, we introduce a novel prompt engineering technique called Labeled Guide Prompting. This method works by splitting the task into sub-tasks, each with a unique label. The LLM is given response requirements that define these labels and is instructed to address each label. This instruction is further corroborated by the few shot learning example, which demonstrates a perfect sample wherein each sentence is labeled in reference to

which sub-task it is addressing. As a result of this method, the LLM dedicates more sentences to each proposed labeled sub-task . It views each label as a separate concept leading to a more detailed response with little to no overlap between the information and insights used in one label to the next.

To evaluate the effectiveness of our proposed technique, we conducted a series of experiments with 100 credit applications, comparing the LLM's performance with and without the implementation of this method. We used several metrics for the evaluation:

(1) **Number of complete responses**: a complete response has the three items addressed for each variable.
(2) **Number of occurrences of the targeted items** in each response.

In the response, each feature is addressed in a separate paragraph, which details its impact on the target. This explanation includes interactions whenever it references any parent nodes. For instance, if "Crop Yields" is linked to "Profitability," which in turn is linked to "Credit Performance," an interaction for "Profitability" might include a reference to "Crop Yields."

Additionally, we introduce insightfulness measures to evaluate the depth of the LLM's reasoning:

(1) **Insightfulness of Interactions**: For each feature's paragraph, we first compute the ratio of the number of mentioned parent nodes to the total number of parent nodes. We then average this ratio across all features that show interactions.
(2) **Insightfulness of the Response**: For each feature's paragraph, we first compute the ratio of the number of mentioned parent nodes to the total number of parent nodes. We then average this ratio over all features that are explained in the answer.

By quantifying the insightfulness of interactions and responses, we can assess the LLM's capability to integrate and reason with multiple variables simultaneously. This in-depth analysis enables us to optimize the performance of the LLM in generating comprehensive credit risk reports.

The Label Guide Prompting technique works in synergy with the Bayesian network, as they facilitate an environment where the LLM can engage in abductive reasoning – the process of considering plausible scenarios and outcomes based on available data. Specifically, the Bayesian network representation provides a framework that allows the LLM to explore various pathways and interactions embedded within it, while Label Guide ensures that these intricate relationships are adressed by the LLM in its output. Consequently, the LLM becomes more proficient in managing complex tasks that require a deep understanding of multiple factors.

3 RESULTS

3.1 Blind Review

Table 2 presents the results of Question 1: "Which report was more helpful for you to assess the credit risk?". They indicate the evaluators' preference for the report generated by GPT-4 (LLM-G) and "Both" (indicating a positive reception to LLM-G) across both English and Portuguese. The preference for LLM-G is noticeable in the English version, with 90. 2% favoring it when disregarding the "Both" option. In the Portuguese version, evaluators still show a significant preference, with 60% favoring LLM-G in the same conditions.

Table 2: Responses to Question 1 - "Which report was more helpful for you to assess the credit risk?" Preferences expressed by evaluators, as counts and percentages of total responses. The p-values are from chi-squared tests for equal likelihood across categories. The first row for each language allows for the choice of both reports, while the second row indicates exclusive preference for either report human-generated (HG) or LLM-generated (LLM-G).

Language	HG	LLM-G	Both	Total	p-value
English	6 (5.1%)	55 (46.6%)	57 (48.3%)	118	5.095×10^{-10}
English	6 (9.8%)	55 (90.2%)	-	61	2.496×10^{-10}
Portuguese	29 (29.9%)	43 (44.3%)	25 (25.8%)	97	0.061
Portuguese	29 (40%)	43 (60%)	-	72	0.098

The p-values presented in Table 2 are derived from a chi-squared goodness-of-fit test. In the case of Portuguese evaluators, the p-value, when we assess preferences for HG versus LLM-G or "Both", is approximately 6.841×10^{-10}. Hence, the evaluators in Portuguese also show a significant preference for LLM-G, as we assume "Both" is favorable to LLM-G due to the benefits of automation.

In the blind review process for credit applications, each application received between 1 to 5 evaluations. Among 72 applications evaluated more than once, 41 were evaluated twice, 29 once, 23 three times, 5 four times, and 3 five times. By combining the categories LLM-G and "Both" into LLM-G exclusively for this analysis, each credit application evaluated more than once generated a binary distribution of preferences (HG versus LLM-G).

To assess the agreement among reviewers, we calculated the entropy for each of these 72 binary distributions. Entropy, in this context, serves as a measure of diversity in preference, with 0

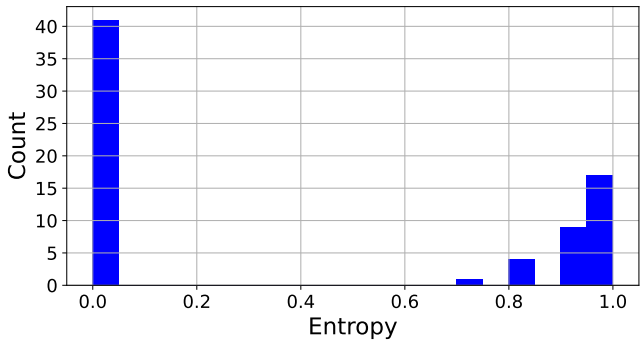

Figure 2: Histogram of the entropy of binary distributions of preferences for human-generated and LLM-generated reports among multiple evaluations of credit applications.

Table 3: Responses to Question 2 - "Do any of the reports have any information that is not true or does not make sense?" Responses as counts and percentages of total responses. The first row for each language includes all responses, including those who found no issues in any or both of the reports. The second row includes only the responses of those who found issues in precisely one report.

Language	None	HG	LLM-G	Both	Total
English	84 (71.2%)	9 (7.6%)	3 (2.5%)	22 (18.6%)	118
English	-	9 (75%)	3 (25%)	-	12
Portuguese	71 (73.2%)	10 (10.3%)	10 (10.3%)	6 (6.2%)	97
Portuguese	-	10 (50%)	10 (50%)	-	20

Table 4: Responses to Question 3 - "Why do you prefer the chosen report?" We illustrate the distribution of Semantic Domains.

Semantic Domain	HG	LLM-G	Both	Total
Agricultural Finance	12	8	8	15
Information	6	5	1	9
Credit Risk	0	3	1	3

representing unanimous agreement (all reviewers selected the same category), and 1 denoting total disagreement (votes for categories HG and LLM-G were evenly distributed).

Figure 2 depicts the histogram of entropies and shows that 41 out of the 72 applications evaluated multiple times had unanimous decisions, as evidenced by entropy of 0, indicating a high level of consensus among reviewers. Conversely, we observed that the entropy was 1 for 17 applications, highlighting a uniformly split decision among evaluators. Thus, while a significant degree of consensus was noted among evaluators overall, 23.6% of the 72 applications had disagreement on the reviewers' choices. This graph also shows that fewer applications had entropy values falling between 0 and 1. This signifies a split preference among reviewers, but not an equal distribution between the two categories.

Table 3 presents the results of Question 2: "Do any of the reports have any information that is not true or does not make sense?" The results of whether the evaluators found any information that was untrue or did not make sense show that LLM-G has no more errors than HG. Also, these results show that the translation tool does not increase the overall occurrence of errors.

When the responses identifying errors in both reports HG and LLM-G are considered, and the counts for "Both" are aggregated to the totals of HG and LLM-G, a chi-squared goodness-of-fit test yields p-values of 1.095×10^{-10} and 1.312×10^{-13} for English and Portuguese responses respectively.

Table 4 shows the top 30 words in three semantic domains: agricultural finance, information, and credit risk, across the HG and LLM-G reports. The Human report emphasizes agricultural finance terms such as 'crop yield' and 'season', showcasing its ability to focus on the most pertinent information for each credit application, as highlighted by reviewers that preferred this type of report.

In contrast, the LLM-G report often uses credit risk terms such as 'credit score' and 'decision', showing its strength in explaining credit risk scores in relation to all factors. In the 'information' category, both reports share a similar count of common terms, yet the specific words they use vary. For instance, the LLM-G report leans toward 'detail' and 'specify', while the Human report favors words associated with 'understanding' and 'analysis', showing the respective strengths of each preferred report.

Table 5: Impact of Labeled Guide Prompting. The p-values are from the Wilcoxon signed-rank test.

Average Occurrences	Non Labeled	Labeled	p-value
1 - impact	7.55	8.57	1.557×10^{-6}
2 - pathways	4.97	8.46	2.04×10^{-14}
3 - interactions	4.57	8.42	3.221×10^{-16}

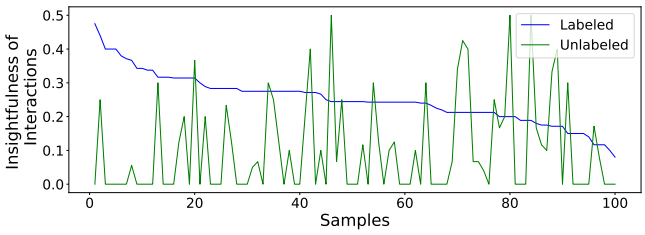

Figure 3: Comparison of Insightfulness of Interactions for each report, with and without LGP. The corresponding p-value from the Wilcoxon signed-rank test was 3.214×10^{-11}

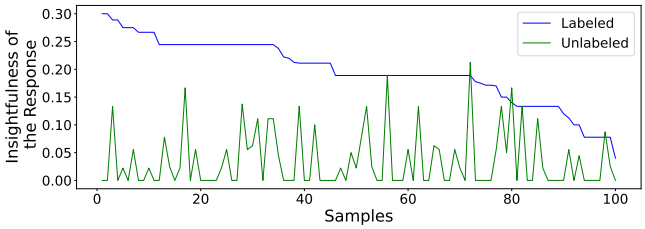

Figure 4: Comparison of Insightfulness of the Response for each report, with and without LGP. The corresponding p-value from the Wilcoxon signed-rank test was 1.087×10^{-17}

3.2 Performance of Labeled Guide Prompting

Table 5 shows the impact of Labeled Guide Prompting (LGP) in the responses. The average occurrence of labels 1, 2, and 3 all rose significantly, with p-values derived from a Wilcoxon signed-rank test indicating that these increases were statistically significant. There was a noteworthy increase in complete responses, from 2 in the non-labeled scenario to 56 when LGP was employed.

Figure 3 depicts the insightfulness of interactions for each report, whereas Figure 4 illustrates the insightfulness of each response. A noticeable increase in overall insightfulness, measured by both metrics, is clearly demonstrated in these figures. The implementation of the proposed technique effectively enhances the insightfulness of interactions and responses.

The insightfulness of interactions is related to the number of parent nodes mentioned in each feature paragraph (see 2.5). The insightfulness of the response is influenced by both the quantity of interactions and their individual insightfulness.

This observation explains why some samples in Figure 3 have higher unlabeled values than labeled ones. These particular cases are instances where only one or a few variables exhibit interactions. Consequently, when averaged over features with interactions, the insightfulness of interactions of these features predominates, regardless of how many interactions occur in the response.

In contrast, in Figure 4, when the averaging process extends over all features explained in the response, it also accommodates the increased number of occurrences of interactions, not just their insightfulness. Therefore, a more comprehensive view is provided, integrating both the frequency and insightfulness of interactions.

The introduction of more interactions fosters abductive reasoning, as it integrates information about the parent nodes, thus enriching the relationship observed between the feature and the target. To illustrate this, consider the following examples of interactions provided by GPT-4 using our methods, where the model performs abductive reasoning.

Crop Yields (high): Considering that the farmer has a high yield and is cultivating rice (3), this directly enhances the profitability per hectare (2) and improves the Credit Risk Score (1).

State (Rio Grande do Sul (RS) - Brazil) and **Cultivated Crop** (rice): These inputs interact with the crop yields and profitability per hectare influencing the Credit Risk Score (3). Cultivating a high-yield crop like rice in a state like RS with suitable conditions can enhance profitability and hence, the credit risk score (2).

Short Term Debt to Total Planted Area (low): Coupled with the low levels of long-term borrowing (3), the lower score for this variable also reflects good cash flow management, thereby improving the overall Credit Risk Score (1).

4 DISCUSSION

Our results indicate GPT-4 generated reports are as useful as traditional credit analyst reports for credit risk assessment. The selection of the "Both" option signals approval of report LLM-G, because of the inherent efficiency and scalability benefits of automation.

The blind review process analysis indicates inter-rater reliability, as evidenced by the majority of low entropy values, indicating consistent, non-arbitrary reviewer decisions.

The significant deviation from an equal distribution across nonsensical content categories demonstrates that errors are statistically less frequent in the reports, especially in report LLM-G, suggesting GPT-4 did not introduce harmful hallucinations.

The success of our approach lies in the combination of Bayesian networks and the Label Guide Prompting technique. The former provides a robust framework for GPT-4 to explore complex credit risk factors, contributing to report LLM-G's comprehensiveness. The latter ensures more detailed responses and facilitates the understanding of each sub-task's unique characteristics.

Despite the success, GPT-4's tendency to give equal importance to all features was seen as a shortcoming. Future work could aim to refine GPT-4's summarization capabilities to better reflect the Bayesian network feature importance.

5 CONCLUSION

In summary, the combination of Bayesian networks and Labeled Guide Prompting can enhance GPT-4's performance in complex problem-solving tasks, achieving a level of competency comparable to human experts. Despite these advancements, several research challenges remain. Specifically, there is a need to develop critical summarization methods that allow GPT-4 to pinpoint and succinctly communicate the most important aspects of a task. Furthermore, integrating the Bayesian network feature importance directly into the LLM to offer case-specific insights presents an avenue for future work. The continual refinement of these elements furthers the proficiency of LLMs in intricate domain-specific tasks, contributing towards the creation of more precise and reliable AI systems.

ACKNOWLEDGMENTS

This research was supported by Traive Inc. We thank Traive's Risk Team, Luis Lapo, Antonio Hildenberg, Rafael Arruda, and the Customer Success Team, Danilo Carnevalli, Leandro Marques, Marinna Reis, and Anderson Viana, for their collaboration and insights.

REFERENCES

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[2] Tom B. Brown et al. 2020. Language models are few-shot learners. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[3] Zhiyu Chen, Harini Eavani, Yinyin Liu, and William Yang Wang. 2019. Few-shot NLG with pre-trained language model. *CoRR* abs/1904.09521 (2019). arXiv:1904.09521 http://arxiv.org/abs/1904.09521

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[5] Amr Hendy et al. 2023. How good are GPT models at machine translation? A comprehensive evaluation. arXiv:2302.09210

[6] Seungone Kim. 2022. Can Language Models perform Abductive Commonsense Reasoning? arXiv:2207.05155

[7] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[8] Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023. Check me if you can: Detecting ChatGPT-generated academic writing using CheckGPT. arXiv:2306.05524

[9] Remo Pareschi. 2023. Abductive reasoning with the GPT-4 language model: Case studies from criminal investigation, medical practice, scientific research. arXiv:2307.10250

[10] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for NLG systems. *CoRR* abs/2008.12009 (2020). arXiv:2008.12009 https://arxiv.org/abs/2008.12009

[11] Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615

[12] Mirac Suzgun et al. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv:2210.09261

[13] Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947 (2023), 224–226.

[14] Ashish Vaswani et al. 2017. Attention is all you need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[15] Alex Wang et al. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR* abs/1804.07461 (2018). arXiv:1804.07461 http://arxiv.org/abs/1804.07461

[16] Alex Wang et al. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding Systems. *CoRR* abs/1905.00537 (2019). arXiv:1905.00537 http://arxiv.org/abs/1905.00537

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR* abs/2201.11903 (2022). arXiv:2201.11903 https://arxiv.org/abs/2201.11903

[18] Jules White et al. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382

[19] Shijie Wu et al. 2023. BloombergGPT: A large language model for finance. arXiv:2303.17564

[20] Yankai Zeng, Abhiramon Rajasekharan, Parth Padalkar, Kinjal Basu, Joaquín Arias, and Gopal Gupta. 2023. Automated interactive domain-specific conversational agents that understand human dialogs. arXiv:2303.08941

[21] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. arXiv:2301.13848