COMISIÓN PARA EL MERCADO FINANCIERO
Regulador y Supervisor Financiero de Chile

Documento de Trabajo N° 03/24

Can Large Language Models Predict Credit Risk? An Empirical Study on Consumer Loans in Chile

Diego Beas · David Díaz

Diciembre 2024
www.CMFChile.cl

Regulador y Supervisor Financiero de Chile

The Working Papers series is a publication of the Financial Market Commission (CMF), whose purpose is to disseminate preliminary research in the finance area for discussion and comments. These works are carried out by professionals of the institution or entrusted by it to third parties.

The objective of the series is to contribute to the discussion and analysis of relevant topics for financial stability and related regulations. Although the Working Papers have the editorial revision of the CMF, the analysis and conclusions contained therein are the sole responsibility of the authors.

La serie de Documentos de Trabajo es una publicación de la Comisión para el Mercado Financiero (CMF), cuyo objetivo es divulgar trabajos de investigación de carácter preliminar en el área financiera, para su discusión y comentarios. Estos trabajos son realizados por profesionales de esta institución o encargados por ella a terceros.

El objetivo de la serie es aportar a la discusión y análisis de temas relevantes para la estabilidad financiera y normativas relacionadas. Si bien los Documentos de Trabajo cuentan con la revisión editorial de la CMF, los análisis y conclusiones en ellos contenidos son de exclusiva responsabilidad de sus autores.

Documentos de Trabajo de la Comisión para el Mercado Financiero (CMF) Financial Market Commission (CMF)
Av. Libertador Bernardo O'Higgins 1449, Santiago, Chile Teléfono: (56) 22617 4058

Documento de Trabajo N° 03/24
Working Paper N° 03/24

Can Large Language Models Predict Credit Risk? An Empirical Study on Consumer Loans in Chile*

Diego Beas[1] y David Díaz[2]

Diciembre 2024

ABSTRACT

This study empirically evaluates the performance of Large Language Models (LLMs) in predicting credit risk for retail banking in Chile, comparing their effectiveness to traditional machine learning models. A variety of LLM configurations were tested, including models with and without fine-tuning, different chunking sizes, and several prompt engineering strategies, such as credit analyst roleplay, chain-of-thought reasoning, emotional stimuli, take a breather, and example-based learning (one-shot and few-shot). The analysis compared open-source models like Llama 3 and commercial models like GPT-3.5 and GPT-4.0. Results indicate that fine-tuned LLMs can achieve predictive accuracy levels comparable to traditional models such as logistic regression and ensemble methods like LightGBM. The top-performing fine-tuned GPT-3.5, GPT-4.0 and Llama 3 configurations achieved AUROC values near 80%, closely matching the best-performing LightGBM benchmark.

A stability test was conducted to assess the consistency of predictions, crucial for credit risk applications. LLMs with a temperature setting of zero demonstrated high stability, producing consistent results across repeated queries, while higher temperature settings introduced variability, especially in default predictions, underscoring the importance of controlling this parameter for reliable results.

Additionally, the LLMs were evaluated for their ability to explain their predictions. Textual explanations provided by a best performing LLM model were blindly reviewed by a credit risk expert, who rated them with an average score of 5.5/7. This result shows promise for explainability but also revealing occasional inconsistencies. Some explanations omitted relevant variables or provided justifications that did not fully align with the underlying data. These findings suggest that, with fine-tuning and careful configuration, LLMs can complement traditional models by offering competitive predictive performance and enhanced transparency in financial applications, particularly in credit risk management.

Keywords: Credit Risk Prediction, Large Language Models (LLMs), Fine-Tuning, Prompt Engineering, Explainability in AI, Consumer Loans, Retail Banking, Artificial Intelligence in Finance.

*/ Las opiniones emitidas en este trabajo, errores y omisiones, son de exclusiva responsabilidad de los autores y no necesariamente reflejan la visión de la institución. Se agradecen los comentarios, consejos y sugerencias del referato interno, así como también los de otros participantes en seminarios internos. Además, se agradece el financiamiento y apoyo recibido por el equipo de la Dirección de Regulación Prudencial de Bancos e Instituciones Financieras y los fondos PAI de la Facultad de Economía y Negocios de la Universidad de Chile.

[1]/ División Normativa de Regulación Prudencial, Dirección de Regulación de Bancos e Instituciones Financieras, Dirección General de Regulación Prudencial, CMF, dbeas@cmfchile.cl

[2]/ Profesor Asociado del Departamento de Administración de la Facultad de Economía y Negocios de la Universidad de Chile, ddiaz@unegocios.cl

RESUMEN

Este estudio evalúa empíricamente el rendimiento de los Modelos de Lenguaje Grande (LLMs) en la predicción del riesgo de crédito para la banca minorista en Chile, comparando su efectividad con modelos tradicionales de aprendizaje automático. Se probaron diversas configuraciones de LLM, incluyendo modelos con y sin ajuste fino, diferentes tamaños de *chunks* y varias estrategias de ingeniería de prompts, como la simulación de rol de analista de crédito, el razonamiento en cadena toma un respiro, estímulos emocionales y el aprendizaje basado en ejemplos (de un solo disparo y de pocos disparos). El análisis comparó modelos de código abierto como Llama 3 y modelos comerciales como GPT-3.5 y GPT-4.0. Los resultados indican que los LLM ajustados finamente pueden alcanzar niveles de precisión predictiva comparables a los modelos tradicionales como la regresión logística y métodos de conjunto como LightGBM. Las configuraciones de GPT-3.5, GPT-4.0 y Llama 3 ajustadas finamente y de mejor desempeño lograron valores de AUROC cercanos al 80%, acercándose mucho al mejor punto de referencia de LightGBM.

Se realizó una prueba de estabilidad para evaluar la consistencia de las predicciones, crucial para aplicaciones de riesgo de crédito. Los LLM con una configuración de temperatura de cero demostraron alta estabilidad, produciendo resultados consistentes en consultas repetidas, mientras que configuraciones de temperatura más alta introdujeron variabilidad, especialmente en las predicciones de incumplimiento, subrayando la importancia de controlar este parámetro para obtener resultados confiables.

Además, se evaluó la capacidad de los LLM para explicar sus predicciones. Las explicaciones textuales proporcionadas por uno de los mejores modelos LLM fueron revisadas a ciegas por un experto en riesgo de crédito, quien las calificó con un puntaje promedio de 5.5/7. Este resultado muestra potencial para la explicabilidad, aunque también revela inconsistencias ocasionales. Algunas explicaciones omitieron variables relevantes o proporcionaron justificaciones que no se alineaban completamente con los datos subyacentes. Estos hallazgos sugieren que, con ajuste fino y una configuración cuidadosa, los LLM pueden complementar a los modelos tradicionales al ofrecer un rendimiento predictivo competitivo y una mayor transparencia en aplicaciones financieras, particularmente en la gestión del riesgo de crédito.

I. Introduction

Large Language Models (LLMs) have rapidly emerged as a central focus of artificial intelligence (AI) research and application across numerous sectors of the economy. Their impressive performance in a wide range of tasks has garnered significant attention from both industry and academia. Today, a growing community is developing and exploring new applications for these models, such as instant translation, virtual assistants, personalized recommendations, sentiment analysis, code generation from general instructions, and creative content generation, among many others.

In the financial sector, machine learning (ML) techniques have already shown substantial improvements in predictive modelling for risk management (Van Liebergen 2017; Aziz and Dowling 2019; Leo et al. 2019; Mashrur et al. 2020). One of the most critical areas where these advancements are making an impact is credit risk management, which focuses on assessing the likelihood that a borrower will default on their financial obligations. Credit risk modelling, particularly the estimation of the Probability of Default (PD), is crucial for financial institutions, as it helps determine capital requirements and informs loan pricing and portfolio management decisions.

LLMs represent a promising frontier in this regard, as they are designed to process both structured and unstructured data, such as customer information and financial documents. Despite this potential, the use of LLMs in credit risk prediction has been limited. Recent studies have begun exploring the applicability of LLMs in related domains. For example, Wu et al. (2021) demonstrated that LLMs could improve sentiment analysis for financial texts, helping predict stock market movements. Similarly, Bakumenko et al. (2024) showed that LLMs can enhance the performance of fraud anomaly detection from financial transactions.

When it comes to applying LLMs directly to PD prediction, the most notable work so far is by Babei and Giudici (2024). They explored the potential of LLMs for credit scoring and found that these models could perform similarly to classical techniques such as logistic regression, particularly in cases where text data and other unstructured inputs were critical. Their study, however, was limited in scope, relying on a small dataset and a single benchmark model for comparison. This paper builds on that foundation by expanding the analysis to a larger dataset, a more diverse set of models, and a comprehensive evaluation of LLM configurations, including fine-tuning and prompt engineering techniques.

In the broader context of risk management, there is a growing body of work focusing on the explainability of complex models. Techniques such as SHAP values (Lundberg and Lee 2017) have been developed to explain the predictions of black-box models like deep neural networks and ensemble methods. However, few studies have explored whether LLMs can not only predict credit risk but also provide coherent and actionable explanations for their predictions. Our study is one of the first to systematically evaluate the quality of LLM-generated explanations in the context of credit risk.

Another challenge in applying LLMs to financial risk management is the stability of predictions. LLMs can produce varying outputs depending on internal parameters such as the model's temperature setting, which controls the degree of randomness in the generated predictions. In fields like credit risk, where consistency and reliability are paramount, this variability is problematic. Existing literature on AI model stability in finance, such as Vela et al. (2023), has primarily focused on traditional models, leaving a gap in understanding how to manage variability in LLM predictions effectively.

We address the gap in the literature regarding LLM explainability and stability, two factors crucial for their deployment in high-stakes decision-making environments like credit risk management. The key contributions of this paper are threefold: a) We compare the performance of LLMs, including GPT-3.5, GPT-4.0, and Llama 3, against traditional ML models such as logistic regression and LightGBM. The results show that fine-tuned LLMs can achieve predictive accuracy comparable to traditional models, with AUROC values approaching 80%; b) beyond accuracy, LLMs offer a unique advantage by generating textual explanations for their predictions. We assess the quality of these explanations by having a credit risk expert evaluate a sample of model outputs. The expert found the explanations generally useful, assigning them an average score of 5.5/7, though occasional inconsistencies were noted, c) a critical requirement for deploying AI in credit risk management is the stability of predictions. We conduct a stability test by repeatedly querying the same LLMs with identical inputs. Our findings show that LLMs can produce consistent predictions when the model temperature is set to zero, minimizing random variations, a necessary condition for reliable credit risk evaluation.

This study is among the first to apply LLMs to credit risk prediction at scale, offering valuable insights into both the predictive power and practical limitations of these models in financial applications. In addition to showing that LLMs can compete with traditional ML techniques in terms of accuracy, our results highlight their potential to improve the explainability of credit risk models, a key concern for regulatory compliance and transparency in financial decision-making. Moreover, this study is one of the first to systematically evaluate the quality of LLM-generated explanations, a key requirement for their use in regulated financial environments.

In the following sections, we first provide a conceptual overview of credit risk modelling and LLMs. We then outline the experimental design and dataset used for this study, followed by a presentation of our results and a detailed discussion on the challenges and opportunities of incorporating LLMs into credit risk management. Finally, we conclude with suggestions for future research and implications for practitioners.

II. Literature Review

Accurate assessment of credit risk is critical for financial institutions, influencing decisions related to loan approvals, capital allocation and risk management. Credit risk refers to the possibility that a borrower will default on financial obligations, resulting in losses for the lender. To quantify this risk, two key metrics are commonly used: Probability of Default (PD) and Loss Given Default (LGD). PD estimates the likelihood that a borrower will default within a specific time frame, typically 12 months, while LGD represents the percentage of economic loss incurred once default occurs. These two parameters form the foundation of credit risk modelling in the financial sector.

Historically, PD has been modelled using statistical approaches such as logistic regression, which provides interpretable results and allows institutions to incorporate structured financial data into their models, such as borrower characteristics and macroeconomic indicators (Kruppa et al., 2013; Addo et al., 2018). While logistic regression has been a standard approach, it has limitations when dealing with the complex and often non-linear relationships inherent in financial data. To address these limitations, machine learning (ML) techniques have emerged as more sophisticated alternatives. Advanced ML techniques such as support vector machines (SVMs), random forests, and gradient boosting machines (e.g., LightGBM) have demonstrated significant improvements in predictive accuracy over traditional models (Mhlanga, 2021; Breeden, 2021). These models excel at

identifying patterns in large datasets and can process multiple variables simultaneously to predict credit risk more accurately. However, they predominantly rely on structured data often struggle to incorporate unstructured information, such as text data from loan applications or financial reports, which could provide additional insights into borrower behaviour (Aziz and Dowling, 2019).

Neural networks, a subset of machine learning, have emerged as powerful tools for predictive modelling that can also be adopted to process both structured and unstructured data. Although the theory of neural networks was developed decades ago, their practical application in finance only became feasible with advancements in computational power (Macukow, 2016). The core structure of a neural network consists of layers of neurons (equations), each applying a weighted sum of inputs and passing the result through an activation function to generate an output (Goodfellow et al., 2016). These networks are capable of modelling highly non-linear relationships, making them suitable for complex financial tasks. Deep learning, a form of neural network architecture with multiple layers (or "depth"), has enabled the processing of vast datasets with complex structures, such as images, time series, and text (Chollet, 2018). Specialized deep neural networks architectures have emerged for specific data types. For example, convolutional neural networks (CNNs) were originally designed for image data, while Long Short-Term Memory (LSTM) networks are tailored for sequential data, such as time series. However, natural language processing (NLP), which deals with text data, presents unique challenges due to the varying relationships between words in a sentence.

Traditional neural network architectures struggled with this complexity until the introduction of the transformer architecture, which incorporates an attention mechanism (Vaswani et al., 2017). This development revolutionized NLP by enabling models to capture long-range dependencies in text data. The attention mechanism allows transformers to weigh the importance of different words relative to one another, making them particularly adept at handling tasks where context plays a crucial role (Vaswani et al., 2017). This breakthrough paved the way for Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019) and LlaMA (Touvron et al., 2023), which are trained on massive datasets and contain billions of parameters. LLMs excel in generating human-like text, answering questions, translating languages, analysing text in search for specific topics or sentiments and even summarizing complex documents. Thus, their potential ability to process both structured and unstructured data, including financial documents and borrower communications, makes them highly relevant for credit risk modelling.

While LLMs have proven effective in numerous fields, their application in credit risk prediction is still emerging. Recent studies, such as Babei and Giudici (2024), have demonstrated that LLMs can achieve comparable performance to traditional models like logistic regression when predicting PD. Their research highlights the potential of LLMs to incorporate unstructured data—an area where traditional models struggle—into risk assessments. However, deploying LLMs in financial applications like credit risk modelling presents several challenges, particularly around performance, stability, and explainability.

LLMs can produce varying outputs given the same input depending on factors such as the temperature parameter, which introduces randomness into the text generation process. This variability is problematic in high-stakes settings where consistency and reliability are crucial, such as credit risk assessment. Additionally, neural networks in general and LLMs in particular, are often regarded as "black-box" models, with opaque decision-making processes that complicate regulatory compliance and stakeholder trust. Addressing these challenges requires careful consideration of

several research design decisions, which can be broadly categorized into three areas: model characteristics, interaction strategies, and external enhancements.

Model characteristics refer to the internal structure of LLMs, including factors like model size, the number of internal layers, and hyperparameter configurations. Larger models with more layers and weights can capture intricate patterns in data, but they also require more computational resources and can lead to risks such as overfitting or instability in predictions. Researchers frequently use pre-existing models like GPT, BERT, or Llama, which come with established architectures and pre-trained weights. By leveraging the knowledge these models acquire from large, general-purpose datasets, researchers can apply them "as is" to domain-specific tasks such as credit risk prediction or financial sentiment analysis (Pan and Yang, 2010; Wu et al., 2021). This method, known as transfer learning, allows models to utilize their prior learning without the need for additional training. A more specialized approach, fine-tuning, involves retraining a pre-trained LLM on domain-specific data to enhance its performance for specific tasks, like predicting credit defaults. Studies have demonstrated that fine-tuning improves model accuracy in financial tasks by capturing subtle patterns in borrower behaviour (Howard and Ruder, 2018; Babei and Giudici, 2024).

Interaction strategies, in conjunction with transfer learning, focus on how users can guide LLM outputs by providing specific instructions to the model through prompt engineering. This method has been shown to significantly improve model performance without the need to modify internal parameters. One effective strategy from the literature is requesting the model to plan how to solve a task, before executing it, i.e. asking the model to “think step by step" through its response, following the “chain-of-thought” reasoning technique, as demonstrated by Wei et al. (2022). This approach enables the model to break down complex tasks and generate more coherent and accurate predictions. Another method highlighted in recent research involves instructing the model to "take a breather" or "be patient" before generating a response. Lou et al. (2024) suggest that this strategy helps the model produce more thoughtful and deliberate answers, enhancing its decision-making process. Additionally, “emotional” prompts have been shown to improve LLM performance by integrating emotive cues into the instructions. Li et al. (2023) explored this idea, coining the term "Emotion Prompt" to describe prompts that combine the original task with emotional stimuli. For example, telling the model that "this analysis is very important for my career" can lead to more carefully considered responses. Their research demonstrated that adding emotional stimuli to prompts led to significant improvements across various tasks, including instruction induction and generative tasks, highlighting the potential of emotional intelligence in enhancing LLM output. Finally, a common strategy in prompt engineering involves assigning the model a specific role, such as instructing it to "act as an expert credit risk analyst." This role-based prompting guides the model’s behaviour and enables it to simulate professional decision-making processes, improving the relevance and accuracy of predictions in tasks like credit risk assessment (Wang et al., 2023).

Another key interaction strategy is example-based learning, where the model learns from examples provided during inference. In zero-shot learning, the model performs a task without any examples, relying solely on the prompt to understand and execute the task. Conversely, in one-shot or few-shot learning, the model is guided by a small set of examples embedded in the prompt, helping it grasp the task more effectively. This approach is particularly valuable when domain-specific data is scarce, allowing the model to learn patterns from just a handful of examples (Brown et al., 2020).

What makes these strategies powerful is their modularity. Researchers can combine multiple techniques within the same prompt to tailor the model's behaviour for specific tasks. For instance, a

researcher might ask the model to "think step by step" using the chain of thought strategy, request it to "be patient" and take a moment before responding, assign it a specific role, such as "acting as an expert credit risk analyst," and emphasize the importance of the task with emotional cues like "this analysis is crucial for my career." Additionally, the researcher can choose to provide zero or few examples in the prompt, depending on the complexity of the task or the availability of training data. This flexibility allows for a highly customizable interaction with the LLM, enhancing its ability to perform domain-specific tasks with greater accuracy and relevance.

In addition to interaction strategies, the performance of LLMs can be further enhanced by providing them with access to external information or by enabling collaboration with other models or systems. These external enhancements expand the model's ability to incorporate real-world data and improve the accuracy of its predictions. One prominent method is Retrieval-Augmented Generation (RAG), which allows the LLM to retrieve relevant documents, such as financial reports or market data, during inference. This integration of external information adds valuable context, enabling the model to produce more accurate and informed predictions (Lewis et al., 2020; Wu et al., 2022). Moreover, these strategies can be combined in adaptive systems, where multiple specialized agents—whether LLMs or other models—work together to solve complex tasks. For example, in credit risk assessment, one agent might be tasked with gathering borrower data, another could focus on predicting default risk, and yet another might be responsible for generating explanations. These agents can access external services through Application Programming Interfaces (APIs) or utilize different LLM configurations to handle distinct tasks. Such systems create a dynamic and flexible framework for decision-making, where each agent performs a specific role, contributing to a more efficient and accurate solution. The combination of internal strategies—such as chain of thought, role assignment, and example-based learning—along with external enhancements like RAG and collaborative agent-based systems, could enable LLMs to address domain-specific challenges with greater precision and flexibility.

Considering the current state of the literature, and despite the progress in applying machine learning (ML) and LLMs to credit risk modelling, several research gaps arise. Comprehensive comparisons between LLMs and traditional ML models are limited, especially when considering the integration of both structured and unstructured data (Babei and Giudici, 2024). Additionally, the role of unstructured data, such as financial reports and customer communications, remains underexplored, despite LLMs being well-suited for processing such inputs (Wu et al., 2022). Furthermore, issues of explainability and stability, particularly the variability of predictions due to parameters like temperature, also present significant barriers to the adoption of LLMs in high-stakes financial applications.

To address part of these gaps, we experiment with a subset of strategies— including transfer learning, fine-tuning, prompt engineering, example-based learning, and RAG—assessing their impact on LLM performance. A limitation of this study, however, is that we do not explore agent-based systems, which could provide additional flexibility by distributing tasks across multiple specialized agents. We also contribute by empirically comparing multiple LLMs—GPT-3.5, GPT-4.0, and Llama 3—against traditional models like logistic regression and LightGBM in the context of credit risk prediction. Additionally, we explore the integration of structured and unstructured data to enhance risk prediction and evaluate the quality of LLM-generated explanations by comparing them with expert opinions. Furthermore, we analyse the stability of LLM predictions and provide insights on improving consistency for high-stakes applications.

III. Methodology

This section outlines the steps taken to evaluate the effectiveness of Large Language Models (LLMs) for credit risk prediction. The methodology is divided into three key components: (1) predictive modelling comparison between LLMs and traditional machine learning models; (2) stability testing to assess the consistency of LLM predictions; and (3) explainability analysis to evaluate the quality of the LLM-generated explanations.

3.1 Predictive Modelling Comparison

3.1.1 Data and Variables

We employed a dataset provided by the chilean financial regulator*, Comisión para el Mercado Financiero* (CMF), under the CMF's framework of calls for joint research projects. We utilized files submitted by the banks to CMF, following the regulatory guidelines outlined in the "Manual de Sistemas de Información" (Information Systems' Manual). Banks report monthly detailed information about consumer loans, including details of the credit operation at the time of issuance (such as loan amounts and terms) as well as the current status of the loan. This includes whether the loan is classified as normal or in default, which allows us to construct the predictors and target variable for our study. The research focuses on consumer credit operations granted by banks in Chile between 2010 and 2020.

This dataset is unique because it contains information on all consumer credit operations within the Chilean banking system, with data validated by the regulator. As such, it provides comprehensive coverage of the consumer credit market with high-quality data. For instance, as of December 2020, the dataset covers 16.8 million credit operations, corresponding to 5.9 million individuals and 21 different institutions, including banks and their subsidiaries. Importantly, this dataset is not publicly available; researchers accessed it under strict technical and legal conditions, with all information fully anonymized to ensure the exclusion of lender identification and borrowers' personal data. Access was limited to predictor variables and the target label, where predictor variables (V1 to V10) represent widely recognized indicators in credit risk models, focusing on aspects like payment behaviour and debt ratios.

The target variable for this study is Probability of Default (PD), a binary variable indicating whether a borrower will default within the following 12 months. According to the Chilean banking regulatory framework, a borrower is considered in default if any of the following occurs: i) a delay of 90 days or more in the payment of interest or capital; ii) the issuance of a new loan to cover a loan overdue by more than 60 days; or iii) forced restructuring or partial debt forgiveness.

The process of predictor variables (V1 to V10) construction and selection is carried out following the development found in Beas et. al. (2024). Specifically, a set of 50 variables is created, from which the 10 best are then selected based on a random recursive selection algorithm. Table 1 shows a detailed definition of all predictor variables in the study:

Table 1: Predictor Variables

Name	Definition
V1	Corresponds to a binary variable that takes a value of 1 when the client has a delinquency of more than 30 days in the financial system and the month prior to the observation.
V2	Corresponds to a binary variable that takes a value of 1 when the client has a mortgage guarantee and 0 otherwise.
V3	Debt to income ratio for the month, without considering off balance sheet exposures and commercial debt. This definition is currently used in the Chilean framework for the determination of risk weighted asset for credit risk.
V4	Delinquency in days in the bank in the last 12 months according to the following coding: 0 equal to 0 days, 1 equal to between 1 to 30 days, 2 equal to between 31 to 60 days, 3 equal to between 61 to 89 days, and 4 or more is a delay greater than 90 days.
V5	Financial burden in the month, considering off balance sheet exposures but not commercial debt.
V6	Financial burden in the month, considering off balance sheet exposures and commercial debt.
V7	Corresponds to the ratio between the debt of the month of observation compared to the average of the last 12 months. It considers all the debt in the system.
V8	Delinquency in days in the bank in the last 6 months according to the following coding: 0 equal to 0 days, 1 equal to between 1 to 30 days, 2 equal to between 31 to 60 days, 3 equal to between 61 to 89 days, and 4 or more is a delay greater than 90 days.
V9	Average delinquency in days in the bank in the last 3 months according to the following coding: 0 equal to 0 days, less than or equal to 1 is between 1 to 30 days, less than or equal to 2 and greater than 1 is equal to between 31 to 60 days, less than or equal to 3 and greater than 2 is equal to between 61 to 89 days, and 3 or more is arrears greater than 90 days.
V10	Corresponds to a binary variable that takes a value of 1 when the client has a delinquency of more than 90 days in the system and between 3 to 1 month before the observation.

3.1.2 Model Adaptation for LLMs

Since LLMs are designed to handle unstructured text rather than structured data, we needed to transform the original structured dataset to fit the LLM input format requirements. For this, the predictor variables values for each borrower were converted into descriptive sentences in natural language, effectively creating a “credit report card” for each borrower. For example, a data point like V1 = 0.0 would be translated to: "The client has no delinquency in the last 30 days". These descriptions were then fed into the LLM as part of the input prompt.

3.1.3 LLM Models, Benchmarking Models and Configurations

We used several LLMs, including GPT-3.5, GPT-4.0o, GPT-4.0o mini, and Llama 3 8B (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020; OpenAI, 2022; OpenAI, 2023; Touvron et al., 2023). These models were chosen based on their strong performance in natural language tasks and their ability to process unstructured data. The GPT models were accessed through the OpenAI API, integrated with Python for easy querying and training, while Llama 3 was accessed via the Hugging

Face platform. In total, we conducted 34 experiments across various configurations, evaluating both logistic regression (RL) and LightGBM as traditional machine learning benchmarks. For LightGBM, we optimized hyperparameters through grid search with cross-validation.

3.1.4 Sampling and Cross-Validation

We used a sample of 10,000 anonymized borrower observations. In the first place, to comply with Chilean data protection regulations (anonymization) and in second place, primarily due to the direct costs associated with inference and training using proprietary models like ChatGPT 3.5 and ChatGPT 4.0, as the researchers had a limited budget for experimentation.

The dataset was divided into 10 stratified, randomly allocated subsets, each maintaining the consistent proportion of defaulted borrowers (10%). To ensure robustness, we employed 10-fold cross-validation, implemented as follows: in each iteration, we had access to 9,000 samples in the training set and 1,000 samples in the testing set. From the 9,000 observations available in the training set we took a sub-sample of equal number of defaulted and non-defaulted borrowers (500 of each class), following common practice in the machine learning literature where balancing and sub-sampling are standard techniques to improve model performance with imbalanced datasets (see, for example, Haixiang et al., 2017). This approach mitigated the bias that might arise from the original distribution, where default cases were less prevalent. For models that did not require training (such as LLMs without fine-tuning), we evaluated them directly on the 10 testing sets across the 10 folds, applying the model 10 times to the 1,000 testing samples from each subset.

Formally, the procedure can be expressed as in algorithm 1.

Algorithm 1: Sampling and Cross-Validation Process

Input:

- D: Full dataset of 10,000 borrower observations.
- k=10: Number of folds for cross-validation.
- m=1,000: Size of each balanced training sample.
- T=10: Number of iterations.

Output:

- Performance metrics with confidence intervals.

Procedure:

1. **Split dataset D into k stratified subsets**, maintaining the original proportion of 10% defaulted borrowers in each subset S_i , where i=1…k, with $|S_i|$=1,000|.
2. **For each iteration t in {1, …, T}**:
    1. Select $k-1$ subsets {S_1, S_2, …, S_{k-1}} for training, reserving the remaining subset S_k for testing.
    2. **Sub-sample m=1,000 training samples** from the $k-1$ subsets, ensuring a balanced distribution with 500 defaulted and 500 non-defaulted borrowers.
    3. **Train the model** on the balanced training set {S_1, S_2, …, S_{k-1}}
    4. **Evaluate the model** on the reserved testing subset S_k
    5. Store the performance metrics M_t from the evaluation.
3. **Repeat steps 2-4 for each fold** until all subsets S_1, S_2, …, S_k have been used for testing.
4. **Compute confidence intervals** for the performance metrics using the results from all folds M_1, M_2, …, M_T.

3.1.5 Evaluation Metrics

The performance of both LLMs and traditional machine learning models was evaluated using two key metrics: Area Under the ROC Curve (AUROC) and Average Precision (AVGPREC). AUROC measures

the model's ability to distinguish between defaulted and non-defaulted borrowers. AVGPREC represents the area under the precision-recall curve and balances precision with recall, making it particularly useful for imbalanced datasets, such as the one used in this study, where defaulted borrowers constitute a minority class. One important issue relates to the fact that LLMs typically generate text-based predictions rather than raw probability scores, which are needed for the calculation of both AUROC and AVGPREC metrics. Thus, we approximated the probability score *s(x)* for each borrower using the transition probability $l(label)$, where $label$ refers to the predicted status, and $l(label)$ represents the probability of transitioning to that label. For LLMs, the calculation of $s(x)$ is as follows:

$$s(x) = \begin{cases} l(label) & if\ label = \text{default}\ and\ l(label) > 0.5 \\ 1 - l(label) & if\ label = \text{normal}\ and\ l(label) > 0.5 \\ 0.5 & if\ \ l(label) \leq 0.5 \end{cases}$$

Here, $label$ is the predicted status by the LLM, while $l(label)$ is the associated transition probability, retrieved from “output_scores” or the "logprobs" provided by the models at inference.

This innovative formulation allows us to align LLM-generated outputs with traditional metrics used in binary classification tasks, like AUROC and AVGPREC. Additionally, this approach provides a complementary measure of the confidence with which the LLM makes its predictions. By using these confidence scores, we can generate a decision threshold tailored to the risk appetite of the financial institution, offering a practical tool for decision-making in credit risk evaluation.

3.1.6 Modelling Strategies for Credit Risk Prediction

We employed a variety of modelling strategies to evaluate the performance of both traditional machine learning models and large language models (LLMs) for credit risk prediction. These strategies involved benchmark models, transfer learning, fine-tuning, prompt engineering, example-based learning, chunking, and external enhancements such as Retrieval-Augmented Generation (RAG) via “Dynamic Shoting,” an adaptive approach that selects examples based on the query context, optimizing model responses (Sun et al., 2022).

- **Benchmark Models:** As a baseline, we trained and evaluated two traditional machine learning models—logistic regression (RL) and LightGBM. For LightGBM, we optimized hyperparameters using grid search combined with cross-validation, allowing us to benchmark the performance of LLMs against traditional models optimized under ideal conditions.
- **LLMs with Fine-Tuning:** For a subset of models, such as GPT-3.5, GPT-4.0o, GPT-4.0 mini, and Llama 3 (8B), we conducted fine-tuning, where we retrained the LLMs on a domain-specific dataset. Fine-tuning was used to adjust the internal weights of the models, improving their performance on the specialized task of predicting credit default risk. As explained later, the fine-tuning process showed notable improvements in prediction accuracy in some cases, especially in scenarios with a balanced set of defaulted and non-defaulted borrowers.
- **LLMs Without Fine-Tuning (Transfer Learning):** For the LLMs, we evaluated their performance without fine-tuning (transfer learning) to test how well the pre-trained models could generalize to the credit risk prediction task. The models used were GPT-3.5, GPT-4.0o, GPT-4.0 mini, and Llama 3 (8B). This phase allowed us to compare their out-of-the-box performance with the benchmark models.

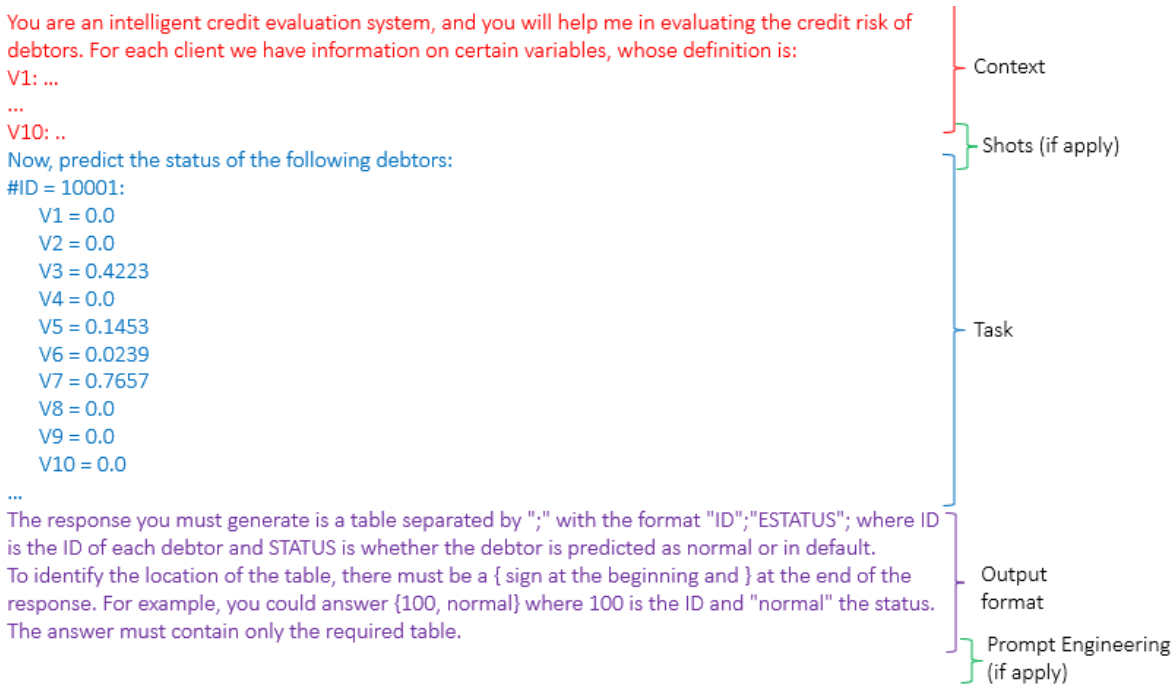
- LLMs Without Fine-Tuning (Transfer Learning) + Prompt Engineering: We employed several prompt engineering techniques to guide the LLMs' outputs and enhance their predictive ability. First, we used the take a role approach: we instructed the model to assume the role of a credit risk analyst with prompts such as "You are an expert credit risk analyst tasked with evaluating credit risk". Second, the think step by step approach: we employed the chain-of-thought reasoning technique by asking the model to break down its decision-making process step by step, enhancing its logical reasoning capabilities (Wei et al., 2022). Third, the take a breather approach: this technique asked the model to "be patient" and take time before generating the response, which helps produce more thoughtful answers (Lou et al., 2024). And forth, the emotional stimuli approach: we included emotional cues like "this analysis is very important for my career" to test if this would improve the model's performance, following the findings of Li et al. (2023).
- LLMs Without Fine-Tuning (Transfer Learning) + Example-Based Learning (Zero-Shot to Few-Shot): To further enhance performance, we used example-based reasoning. We tested two approaches. First, the zero-shot Learning: in this approach, the model was asked to predict the credit risk without any prior examples, relying only on the task description. And second, the few-shot learning: We provided the model with a small set of examples to guide its understanding of the task. We tested different numbers of shots (5, 10, and 20), where the examples were balanced between defaulted and non-defaulted borrowers. The few-shot examples were consistent across all configurations.
- Chunks: Another experimental parameter we tested was the concept of "chunks"—the number of borrowers the model is asked to predict at once. We tested three chunk sizes: 1 (one borrower at a time), 20 (20 borrowers in one prediction request), and 100 (100 borrowers at once). There are three main reasons for testing with chunks: i) Cost Efficiency: Making one API call to predict multiple borrowers at once can be more cost-effective than predicting each borrower individually; ii) Time Efficiency: Predicting several borrowers simultaneously can also reduce time, as it avoids the overhead of repeatedly loading the model, sending data, and receiving responses; and iii) Contextual Performance: The additional context provided by having more examples within the prompt may lead to improved model performance, as the model can potentially benefit from the richer input, even without being explicitly given the correct labels (similar to few-shot learning). Table 2 shows how chunk size impacts performance across various experiments.
- External Enhancements Using Retrieval-Augmented Generation (RAG via Dynamic Shoting): In addition to interaction strategies, we incorporated RAG via Dynamic Shoting as an external enhancement. Instead of providing random examples in the few-shot setting, we used Dynamic Shoting to retrieve the most contextually relevant examples for each borrower. This was done by vectorizing the dataset and using a cosine similarity metric to retrieve the closest matches, similar to a k-nearest neighbors (k-NN) approach. This helped improve prediction accuracy by selecting the most appropriate examples in each prediction.
- Modular and Combinatorial Nature of Strategies: It is important to note that the various strategies tested—such as fine-tuning, prompt engineering techniques (step-by-step reasoning, take a breather), and example-based learning (zero-shot and few-shot learning)—are modular. In our experimental design, we created multiple configurations by "turning on" or "turning off" these strategies. For instance, we tested the LLMs in scenarios where fine-tuning was applied in combination with prompt engineering or where external enhancements such as RAG were utilized alongside few-shot learning. As demonstrated in Table 2, this modular approach resulted in a total of 32 distinct experiments, allowing us to assess the impact of each strategy and their combinations on model performance across

various settings. Figure 1 shows an example prompt and its modular components. As mentioned, ID was an anonymized code that allow us to track the same customer but without access to any of its personal information.

3.1.7 Resource Constraints and Model Selection

Given the costs associated with proprietary models like GPT-4.0o and GPT-4.0 mini, we adopted a two-phase approach to optimize resource allocation. In the first phase, we tested a broader set of configurations using the open-source LlaMA 3 8B model and the more affordable GPT-3.5 model. This allowed us to explore a wide range of strategies, including prompt engineering, example-based learning, chunk sizes, and Retrieval-Augmented Generation (RAG) via Dynamic Shoting, to identify the most effective combinations. Once we identified the top-performing strategies, we proceeded to the second phase, where we applied only the best-performing configurations to GPT-4.0o and GPT-4.0 mini models. This selective approach helped balance the need for comprehensive testing with the financial limitations of running multiple experiments on expensive models. Consequently, only a subset of configurations was tested on these models, as seen in Table 2.

Figure 1: Modular components of prompt



3.2 Stability Testing

In high-stakes financial environments, consistency is essential. To evaluate the stability of the LLM models, we examined how their predictions varied, particularly focusing on the temperature parameter, which controls randomness in the generated text. Thus, we conducted stability tests by running 100 predictions for each borrower in a sub-sample of randomly selected 100 debtors. Predictions were made using two temperature settings: 0.5, allowing some randomness in output generation; and 0.0, eliminating randomness to ensure maximum consistency. Both GPT-3.5 and Llama 3 (8B), with and without fine-tuning, were subjected to these tests. The key outcome was the score s(x) assigned to each borrower, which ideally should remain stable across predictions. High

variability in scores—especially for defaulted borrowers—would signal inconsistencies in the model's decision-making process, which could lead to risk management issues.

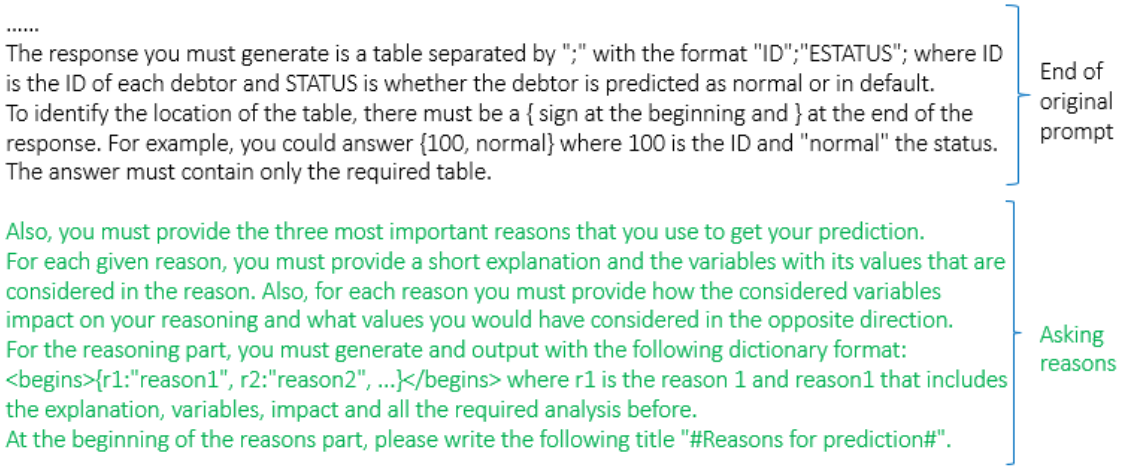
3.3 Explainability Analysis

LLMs are often considered "black-box" models, which poses challenges for their adoption in regulated financial environments where transparency is required. To address this, we conducted an explainability analysis, focusing on how well LLMs could justify their predictions. Thus, we asked the best-performing LLM model to provide explanations for its credit risk predictions on a sub-sample of 100 randomly selected borrowers. The model was prompted to generate both a default prediction and a corresponding explanation, justifying the decision based on the borrower's financial characteristics. For example, a typical explanation might state: "The debtor has a high delinquency in the last 90 days and no mortgage guarantee, indicating a high risk of default."

These explanations were blindly reviewed by a highly experienced credit risk expert with over 25 years in the field, who rated their quality and accuracy on a scale from 1 to 7. A former credit risk manager at Chile's largest international bank, he was provided with anonymized input data, the LLM-generated prediction, the associated probability score, and the model's explanation for each borrower (more details in Appendix 1). To ensure objective evaluation, he was not informed that an LLM generated these predictions and explanations; instead, he believed they came from a junior analyst. This "blind" setup aimed to eliminate bias, focusing his assessment solely on the quality of the explanations.

Following the well-established "4 Cs of Credit" framework—Capacity, Character, Collateral, and Capital (Fraser et al., 2001)—he evaluated each borrower's financial characteristics, a method commonly used to determine creditworthiness. Importantly, he volunteered his expertise without compensation, and his access was strictly limited to anonymized predictor variables and model explanations, ensuring full data privacy. Figure 2 displays the modified prompt asking the model to provide explanations for its predictions.

Figure 2: Prompt that asks explanations from the LLM model.



IV. Results and Analysis

This section presents results in table 2 and findings from the experiments described in the Methodology. Results are structured in alignment with the three core methodological components: (1) Predictive Modelling Comparison, (2) Stability Testing, and (3) Explainability Analysis.

Additionally, Appendix 2 shows the results of the hypothesis test for the mean, where the null hypothesis is that the average AUROC is the same for two given experiments. In this appendix, bold and asterisks highlight when the null hypothesis can be rejected, and therefore the difference in means is significantly different. An element (i,j) in the table represents the average difference between experiment (i) minus experiment (j). Appendix 3 contains the same information, except in this case, the average difference in the AVGPREC measure is tested. In both appendices, the results for each model are presented according to the numbering given in table 2.

4.1 Predictive Modelling Comparison

4.1.1 Role of Chunk Sizes

An unexpected finding from the experiments is the role of chunk size. As shown in the table 2, chunk size impacts both cost and performance. Testing with larger chunks (e.g., 20 or 100 borrowers at a time) proved to be significantly more cost-effective and time-efficient. By running predictions on multiple borrowers in a single API call, we minimized the overhead associated with model loading and API response times. This strategy also has the potential to improve model performance, as providing more borrowers in a single request offers additional context to the LLM, similar to few-shot learning but without the explicit labelling. For example, in GPT-3.5, chunk sizes of 20 generally yielded better results in terms of AUROC and AVGPREC compared to single-borrower predictions. This is a significant finding for the application of LLMs in large-scale credit risk modelling, where efficiency in cost and time is crucial. Specifically, GPT-3.5 with chunk size 20 achieved 73.4% AUROC and 24.8% AVGPREC, whereas the model with chunk size 1 achieved 72.6% AUROC and 26.4% AVGPREC. The trade-off in precision was minimal, but the efficiency gains in both time and cost make larger chunk sizes an attractive option. With chunk size 100, performance remained similar with 71.2% AUROC and 22.9% AVGPREC, indicating that while larger chunk sizes provide cost efficiency, chunk sizes beyond 20 do not substantially improve performance. This is a significant finding for the application of LLMs in large-scale credit risk modelling, where efficiency in cost and time is crucial.

4.1.2 Performance Comparison

In terms of performance, traditional models like Logistic Regression (RL) and LightGBM (L-GBM) showed strong baseline results, with LightGBM outperforming RL in both AUROC and AVGPREC. Specifically, LightGBM achieved 81.7% AUROC and 45.2% AVGPREC, compared to Logistic Regression's 79.1% AUROC and 39.3% AVGPREC. LLMs without fine-tuning generally underperformed compared to traditional models. For example, GPT-3.5 achieved 69.3% AUROC and 23.3% AVGPREC when tested without fine-tuning in a zero-shot learning no interaction strategies configuration. However, after tuning various prompt engineering strategies and experimenting with different chunk sizes, performance improved, with GPT-3.5 reaching 73.4% AUROC and 24.8% AVGPREC when using chunk size 20 and few-shot learning with five examples.

Fine-tuned LLMs demonstrated significant improvement. For instance, GPT-3.5 with fine-tuning achieved 80.2% AUROC and 40.7% AVGPREC, nearing the performance of LightGBM. GPT-4.0o mini also showed considerable improvement after fine-tuning, with 80.1% AUROC and 40.8% AVGPREC. Fine-tuning proved to be particularly effective in enhancing LLM performance, allowing models to better adjust to the specific task of credit risk prediction. The fact that GPT-3.5 and Llama 3 achieved similar results to GPT-4.0 after fine-tuning highlights a critical finding: when fine-tuning is feasible, opting for simpler, more cost-effective models like GPT-3.5 or Llama 3 can be the best option. Conversely, if fine-tuning is not possible or prohibitively expensive, higher-performing models like GPT-4.0 may still be necessary. This trade-off between cost and performance underscores the importance of considering the opportunity to fine-tune in real-world applications, where resource constraints often dictate model selection.

In analysing the effectiveness of Prompt Engineering (PE) and Retrieval-Augmented Generation (RAG via Dynamic Shoting), it is clear that their individual and combined impacts varied across different models. PE alone provided slight performance improvements, particularly in cases where the prompts were carefully designed to guide the model's decision-making process. For example, when applied to GPT-3.5 with a chunk size of 20, PE helped the model achieve an AUROC of 73.2%, compared to 72.6% without PE, though the there was a slightly decrease in AVGPREC (24.4% with PE vs. 26.4%, but no statistical significative). This suggests that while PE can refine prediction accuracy slightly, its effect may be more pronounced in certain scenarios.

Table 2: Predictive Modelling Comparison Results

					Prompt Engineering (PE)				RAG		
Model	N° exp.	N° shots	Chunk size	T	RP	CoT	TaB	ES	DS	AUROC	AVGPREC
RL	1	-	-		-					79,1% ± 1,2%	39,3% ± 2,9%
L-GBM	2	-	-		-					**81,7% ± 1,3%**	**45,2% ± 2,1%**
Llama 3 8B (without fine tuning)	3	0	1	D	✓					73,6% ± 1,9%	27,3% ± 2,5%
	4	5	1	D	✓					73,4% ± 1,2%	28,5% ± 2,6%
	5	**5**	**1**	**D**	✓	✓	✓	✓		**73,5% ± 2,0%**	**28,6% ± 2,8%**
	6	5	1	D	✓	✓	✓	✓	✓	66,2% ± 1,5%	22,3% ± 2,1%
	7	5	20	D	✓					71,0% ± 1,7%	22,9% ± 1,5%
	8	10	1	D	✓					73,1% ± 2,0%	28,9% ± 3,0%
	9	20	1	D	✓					69,4% ± 1,5%	26,3% ± 2,3%
GPT 3.5 (without fine tuning)	10	0	1	D	✓					69,3% ± 1,9%	23,3% ± 2,1%
	11	0	20	D	✓					72,8% ± 1,9%	25,2% ± 1,9%
	12	0	100	D	✓					71,3% ± 1,7%	24,7% ± 1,9%
	13	5	1	D	✓					72,6% ± 2,1%	26,4% ± 2,1%
	14	**5**	**20**	**D**	✓					**73,4% ± 1,6%**	**24,8% ± 1,7%**
	15	5	100	D	✓					71,2% ± 2,1%	22,9% ± 2,3%
	16	5	1	D	✓	✓	✓	✓		69,4% ± 2,5%	21,2% ± 2,6%
	17	5	20	D	✓	✓	✓	✓		73,2% ± 1,9%	24,4% ± 2,0%
	18	5	20	D	✓	✓	✓	✓	✓	68,5% ± 2,0%	26,5% ± 2,5%
	19	10	1	D	✓					71,9% ± 2,2%	23,9% ± 1,7%
	20	10	100	D	✓					69,7% ± 1,9%	22,4% ± 3,1%

	21	20	1	D	✓	72,7% ± 2,2%	25,5% ± 2,5%
	22	20	100	D	✓	67,0% ± 1,6%	20,2% ± 2,3%
GPT 4.0o (without fine tuning)	23	0	1	D	✓	69,6% ± 0,9%	24,6% ± 1,4%
	24	5	1	D	✓	70,2% ± 1,6%	18,4% ± 1,4%
	25	**5**	**20**	**D**	**✓**	**75,3% ± 1,7%**	**28,9% ± 2,5%**
GPT 4.0o mini (without fine tuning)	26	0	1	D	✓	69,6% ± 0,9%	24,6% ± 1,4%
	27	5	1	D	✓	65,7% ± 1,5%	19,6% ± 1,1%
	28	**5**	**20**	**D**	**✓**	**70,0% ± 1,9%**	**19,6% ± 1,3%**
GPT 3.5 (with fine tuning)	29	0	1	D	✓	80,2% ± 1,6%	40,7% ± 2,2%
	30	**0**	**1**	**0**	**✓**	**80,3% ± 1,6%**	**40,7% ± 2,1%**
GPT 4.0o mini (with fine tuning)	31	0	1	D	✓	80,1% ± 1,2%	40,8% ± 2,6%
	32	**0**	**1**	**0**	**✓**	**80,0% ± 1,1%**	**40,7% ± 2,6%**
Llama 3 8B (with fine tuning)	**33**	**0**	**1**	**D**	**✓**	**78,9% ± 1,1%**	**37,7% ± 2,9%**
	34	0	1	0	✓	73,1% ± 1,1%	21,4% ± 1,0%

Symbology:
T: temperature value, where "D" refers to the default setting.
Prompt strategies:

- **RP:** role play.
- **CoT:** chain of thoughts.
- **TaB:** take a breather
- **ES:** emotional stimuli.
- **DS:** dynamic shot or RAG.

RAG-DS, on the other hand, showed a decrease in performance when applied independently. For example, the combination of PE + RAG resulted in lower AUROC (68.5%) and AVGPREC (26.5%) compared to using PE alone. The difference in AUROC is significative while in AVGPREC is not. This highlights that RAG's retrieval of additional context may not always synergize with the LLM's ability to process structured input, potentially introducing noise into the predictions. When PE and RAG were used together, the results were generally lower than when PE was applied independently, suggesting that these strategies may not complement each other as expected. It appears that in the context of structured financial data, PE has a more positive effect when applied on its own, while RAG might require further optimization or more relevant external data to boost performance effectively.

4.2 Stability Testing

Given the significant cost difference between GPT-3.5, GPT-4.0, and Llama 3—along with their statistically similar performance—we opted to use the fine-tuned versions of GPT-3.5 and Llama 3 for subsequent experimentation phases. In this phase, we examined the variability in predicted default status across different temperature settings. Stability testing provided critical insights: as shown in Table 3, at a temperature of 0.5, both GPT-3.5 and Llama 3 exhibited noticeable variability, with the default status fluctuating across 100 predictions for the same borrower. Although GPT-3.5 was somewhat more stable overall, it still displayed variability in borderline cases.

When the temperature was set to 0.0, predictions were far more consistent, with randomness effectively eliminated across all repetitions. However, an important observation emerged: while we anticipated a reduction in variability with lower temperatures, the precise extent of this reduction was unknown beforehand. This "elasticity" measure—how much response variability decreases with temperature adjustments—represents a new and relatively undocumented area. It is one thing to

know that variability should decrease but quantifying exactly how much and under which conditions adds a novel layer of complexity to using LLMs in high-stakes prediction tasks.

Table 3: Standard Deviation of predicted Scores. "Normal" vs "Default" Predictions

Model	Label	Min	Q25	Q50	Avg.	Q75	Q90	Max
GPT 3.5 with fine tunning (temp. 0.5)	**Total**	**0,0%**	**0,0%**	**0,0%**	**0,6%**	**0,0%**	**0,5%**	**8,2%**
	Normal	0,0%	0,0%	0,0%	**0,3%**	0,0%	0,0%	7,0%
	Default	0,0%	0,0%	0,0%	**2,3%**	4,2%	6,4%	8,2%
GPT 3.5 with fine tunning (temp. 0)	**Total**	**0,0%**	**0,0%**	**0,0%**	**0,1%**	**0,0%**	**0,0%**	**3,5%**
	Normal	0,0%	0,0%	0,0%	**0,0%**	0,0%	0,0%	0,1%
	Default	0,0%	0,0%	0,0%	**0,4%**	0,2%	0,8%	3,5%
Llama 3 (8B) with fine tunning (temp. 0.5)	**Total**	**0,9%**	**9,5%**	**11,6%**	**10,9%**	**12,8%**	**13,4%**	**20,1%**
	Normal	3,1%	9,9%	11,7%	**11,2%**	12,8%	13,5%	20,1%
	Default	0,9%	7,7%	9,4%	**9,3%**	12,6%	12,9%	13,2%
Llama 3 (8B) with fine tunning (temp. 0)	**Total**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**
	Normal	0,0%	0,0%	0,0%	**0,0%**	0,0%	0,0%	0,0%
	Default	0,0%	0,0%	0,0%	**0,0%**	0,0%	0,0%	0,0%

4.3 Explainability Analysis

In the case of the LLM model, specifically GPT-3.5 with fine tuning, two key aspects were analysed: a) whether the model is capable of explaining the predictions it makes and b) the quality of those explanations. To assess this, the model was asked to provide the reasoning behind its decision for a sample of 100 randomly selected borrowers.

Table 4 shows an example of the reasons provided by GPT-3.5 when predicting a client as being in default with a probability of 100%. It is evident that the first reason given by the model is inconsistent with the variable value, as it indicates the presence of delinquency when, in fact, no such delinquency exists in the data. However, the second reason aligns with the observed variables, as it accurately points to delinquency in the referred characteristics. Similarly, the last reason provided is consistent with the data, indicating the absence of a mortgage guarantee, which increases the risk.

As mentioned, the sample of 100 explanations was reviewed blindly by a professional in the credit risk management field, who evaluated the predictions and reasoning provided by the GPT 3.5 model on a scale of 1 to 7 (with 7 being the highest score). The evaluation yielded an average score of 5.5, with a 95% confidence interval ranging from 5.3 to 5.7. These results suggest that GPT has the potential to provide reasonable analyses of credit risk assessments. However, some inconsistencies were noted, detailed as follows:

1. The model may omit pertinent information from certain variables that could significantly impact the analysis. The omission of these data points may lead to an incomplete or biased interpretation, affecting the quality and precision of the conclusions drawn.
2. The model's analysis may present inconsistencies between variable values and the resulting predictions. For example, the model may suggest that a financial burden of 100% is manageable in scenarios where it predicts a normal financial state, which appears contradictory and does not provide an adequate justification for its prediction. Such discrepancies can undermine the validity of the conclusions derived and highlight the need

for careful and critical review of the model's outputs. In other cases, the justifications offered by the model, along with the values presented to support its analysis, do not align well with the actual financial condition of the borrower. For instance, the model may indicate that a borrower has no delinquency when, in fact, the analysed variables reflect the opposite. This discrepancy suggests a potential deficiency in the model's ability to correctly interpret the available data, which could lead to erroneous conclusions that affect the credibility of the analysis.

In conclusion, the results of this section indicate that the LLM model can generate a reasonable risk analysis, providing interpretability for its predictions. However, human oversight remains crucial, as the model can occasionally make errors or show inconsistencies in certain cases. Improvements in this area could be observed as these technologies continue to advance.

Table 4: Example of reasons given by GPT for predicting an individual to be in default

VARIABLES	**EXPLANATIONS**
#ID=20522://V1=1.0//V2=0.0//V3=8.4844//V4=3.0//V5=0.5563//V6=0.5563//V7=1.3526//V8=3.0//V9=2.0//V10=0.0//	{r1:"High delinquency levels: The debtor has a delinquency of more than 90 days in the system and between 3 to 1 month before the observation (V10 = 0.0), which indicates a severe default risk. This is a strong indicator of financial distress and inability to meet debt obligations. Impact: High delinquency levels significantly increase the likelihood of default. Opposite direction: A value of 1 for V10 would have indicated a higher default risk.", r2:"Consistent delinquency history: The debtor has a high average delinquency in days in the bank in the last 3 months (V9 = 2.0) and a high delinquency in days in the bank in the last 6 months (V8 = 3.0), showing a pattern of ongoing payment issues. Impact: A consistent history of delinquency suggests a persistent inability to manage debt effectively. Opposite direction: Lower values for V9 and V8 would have implied a more stable payment behavior.", r3:"Lack of mortgage guarantee: The debtor does not have a mortgage guarantee (V2 = 0.0), which could indicate a lack of valuable collateral to secure the debt. Impact: Without a mortgage guarantee, the lender has less assurance of recovering the debt in case of default. Opposite direction: A value of 1 for V2 would have provided additional security for the lender and potentially lowered the default risk."}

V. Discussions on the implementation of LLM models for credit risk assessment

The results of this study offer important insights into the potential application of Large Language Models (LLMs) like GPT-3.5, GPT-4.0, and Llama 3 in the banking sector, particularly in credit risk assessment. The findings align with current challenges and recommendations highlighted by regulatory bodies such as the European Banking Authority (EBA, 2023), which stress that the main obstacles to implementing machine learning models in internal models for capital requirements are the complexity of interpreting results, overfitting tendencies, and the shortage of skilled personnel for both model development and validation.

In the context of LLMs, these challenges are mirrored in our results, particularly regarding the trade-offs between cost, performance, and interpretability. For example, our experiments demonstrated that while fine-tuning LLMs like GPT-3.5 and Llama 3 led to performance levels close to traditional models like LightGBM (AUROC: 80.2% and 80.1% respectively), the costs associated with more advanced models like GPT-4.0 were significantly higher without proportionate gains in performance. This suggests that, when fine-tuning is feasible, using more cost-effective models like GPT-3.5 or Llama 3 may be more practical for financial institutions.

Furthermore, our experiments revealed that chunk size can have a significant impact on both performance and cost-efficiency. When testing chunk sizes of 20 borrowers per prediction, we found that performance (AUROC) and cost-effectiveness improved substantially compared to single-borrower predictions. This optimization is especially important in large-scale credit risk assessments where real-time evaluation of multiple borrower profiles is needed. However, beyond a chunk size of 20, we observed diminishing returns in performance improvements, suggesting that financial institutions should balance between chunk size and precision for optimal results. Additionally, the use of interaction strategies like Prompt Engineering (PE) and Retrieval-Augmented Generation (RAG) showed mixed results. PE improved model accuracy slightly by guiding the model's decision-making process, as seen in the AUROC increase for GPT-3.5. However, RAG's addition did not lead to the expected performance gains and sometimes introduced noise into the predictions, especially when applied together with PE. This highlights that while interaction strategies can refine LLM predictions, they need further optimization, especially in the context of structured financial data.

Interpretability remains a significant challenge for LLMs in the banking sector. Although the models were able to generate explanations for their predictions, as shown in the explainability analysis with GPT-3.5 (which received an average rating of 5.5/7 from an expert), inconsistencies and errors highlight the need for human oversight. For example, in some cases, the model omitted critical information or provided contradictory justifications, signalling potential pitfalls in automating credit risk assessments.

In conclusion, LLMs offer promising opportunities for credit risk modelling in the banking sector, particularly when fine-tuned to specific tasks. However, challenges related to cost, variability in predictions, and the need for human oversight for interpretability issues must be addressed before widespread implementation. These findings suggest that banks should proceed cautiously, integrating LLMs with traditional models and addressing concerns related to scalability, transparency, and regulatory compliance.

VI. Conclusions

The present study provides a comprehensive empirical analysis of the capabilities of Large Language Models (LLMs) in credit risk assessment, comparing their performance against traditional methodologies and exploring various configurations, including fine-tuning and prompt engineering. The results show that LLMs, when properly fine-tuned, can achieve performance levels on par with traditional models such as logistic regression and LightGBM, with AUROC values nearing 80%.

One of the study's main contributions is its analysis of the quality of the explanations generated by LLMs. Our findings show that LLMs can provide interpretable and valuable insights for credit risk assessments, offering a fast and efficient alternative to traditional explainability techniques like SHAP.

However, occasional inconsistencies in the explanations highlight the need for human oversight to validate and critique the model's outputs.

From a regulatory perspective, this study advances the conversation on the applicability of cutting-edge technologies in risk management, offering empirical evidence of their functionality. However, for wider adoption of advanced models, financial institutions must first consolidate the use of simpler and more transparent techniques. LLMs offer a promising route forward, but human involvement remains crucial for ensuring that explanations and predictions meet the high standards required for regulatory compliance.

Despite these significant contributions, some limitations remain. The dataset used in this study is based on pre-pandemic Chilean credit data, and it focuses on a specific type of credit. Future research should explore how these models perform in other markets and credit types. Furthermore, the unavailability of GPT-4.0o for fine-tuning and the absence of larger open-source models like Llama 3 405B represent areas for further study. Exploring more advanced RAG configurations that leverage extended context lengths could offer a similar performance to fine-tuning at a lower implementation cost.

Finally, the ability of LLMs to provide explanations for their predictions opens new opportunities for practical applications, such as assisting customers during the credit application process and enhancing the transparency of credit evaluations. This not only improves customer experience but also provides financial sector workers with useful tools to complement their expertise. The continued development of LLMs and their integration into financial risk management represents a key area for both academic inquiry and industry adoption in the years to come.

References

- Addo P. M., Guegan D., & Hassani B. (2018). Credit risk analysis using machine and deep learning models. Risks, 6(2), 38.
- Aziz S., & Dowling, M. (2019). Machine learning and AI for risk management. Disrupting finance: FinTech and strategy in the 21st century, 33-50.
- Babei G. & Giudici P. (2024). GPT classifications, with application to credit lending. Machine Learning with Applications.
- Bakumenko, A., Hlaváčková-Schindler, K., Plant, C., & Hubig, N. C. (2024). Advancing Anomaly Detection: Non-Semantic Financial Data Encoding with LLMs. arXiv preprint arXiv:2406.03614.
- Beas D., Pulgar C. and Ramírez S. (2024). Valor de la información de deuda en el mercado de créditos. Working paper N°2, Comisión para el Mercado Financiero.
- Breeden J. (2021). A survey of machine learning in credit risk. Journal of Credit Risk, 17(3).
- Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner Ch., McCandish S., Radford A., Sutskever H., & Amodei D. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- EBA (2023). EBA Discussion Paper On Machine Learning For IRB Models. Follow-Up Report From The Consultation On The Discussion Paper On Machine Learning For IRB Models.
- Fraser, D., Gup, B., Kolari, J., (2001) Commercial Banking: The Management of Risk (2nd Edition), South-Western College Publishing, Cincinnati, Ohio.
- Goodfellow I., Bengio Y. & Courville A. (2016). Deep learning. MIT Press.
- Chollet F. (2018). Deep Learning with Python.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239.
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.
- Kruppa J., Schwarz A., Arminger G., & Ziegler A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert systems with applications, 40(13), 5125-5131.
- Leo M., Sharma S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. Risks, 7(1), 29.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Zettlemoyer, L. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.
- LI, Cheng, et al (2023). Large language models understand and can be enhanced by emotional stimuli. arXiv preprint arXiv:2307.11760.
- Luo, W., Zheng, S., Xia, H., Wang, W., Lei, Y., Liu, T., ... & Sui, Z. (2024). Taking a Deep Breath: Enhancing Language Modeling of Large Language Models with Sentinel Tokens. arXiv preprint arXiv:2406.10985.
- Lundberg S. & Lee S. (2017). A Unified Approach to Interpreting Model Predictions.

- Macukow, B. (2016). Neural networks–state of art, brief history, basic models and architecture. In Computer Information Systems and Industrial Management: 15th IFIP TC8 International Conference, CISIM 2016, Vilnius, Lithuania, September 14-16, 2016, Proceedings 15 (pp. 3-14).
- Mhlanga D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. International journal of financial studies, 9(3), 39.
- Mashrur A., Luo W., Zaidi N. A., & Robles-Kelly A. (2020). Machine learning for financial risk management: a survey. IEEE Access.
- OpenAI (2022). Training language models to follow instructions with human feedback.
- OpenAI (2023). GPT-4 Technical Report.
- Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359. DOI: 10.1109/TKDE.2009.191
- Radford A., Narasimhan K., Salimans T. & Sutskever I. (2018). Improving Language Understanding by Generative Pre-training. OpenAI Blog.
- Radford A., Narasimhan K., Salimans T., & Sutskever I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.
- Sun, S., Zhang, X., Li, Z., & Wang, H. (2022). Dynamic Shoting for Improving Large Language Model Responses. Proceedings of the Neural Information Processing Systems, 35, 1035-1045.
- Touvron, Hugo; Lavril, Thibaut; Izacard, Gautier; Martinet, Xavier; Lachaux, Marie-Anne; Lacroix, Timothée; Rozière, Baptiste; Goyal, Naman; Hambro, Eric; Azhar, Faisal; Rodriguez, Aurelien; Joulin, Armand; Grave, Edouard; Lample, Guillaume (2023). LLaMA: Open and Efficient Foundation Language Models.
- Van Liebergen, B. (2017). Machine learning: a revolution in risk management and compliance? Journal of Financial Transformation, 45, 60-67.
- Vela, D., Sharp, A., Zhang, R. et al (2022). Temporal quality degradation in AI models. Sci Rep 12, 11654 https://doi.org/10.1038/s41598-022-15245-z
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 30).
- Wang, Zekun Moore, et al (2023). Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv:2310.00746.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
- Wu, S., Liu, Y., Zou, Z., & Weng, T. H. (2021). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. Connection Science.

Appendix 1: Instructions and data provided for the explainability analysis

A former professional in credit risk management was provided with the below information to conduct the explainability analysis of the output produced by the LLM model:

Instructions:

Your mission will be to evaluate the assessments of a credit risk analyst who has been trained and hired to predict the credit risk of consumer loan debtors at banking institutions in Chile. Specifically, the analyst has been tasked with forecasting debtor default, defined as the occurrence of a 90-day delinquency within a prospective 12-month window from the observation month. For example, if an individual's variables are observed in March 2023, the analyst indicates whether a delinquency of over 90 days will occur in the period from April 2023 to March 2024, inclusive.

For each evaluation request, the analyst is required to provide three results:

- *Probability of default: This corresponds to a value between 0% and 100%, inclusive of both extremes, which represents the probability the analyst predicts that the default event will occur. Its complement (i.e., 100% minus this value) corresponds to the probability that the default will not occur, associated with the label "normal." In some cases, the probability may be 50%, indicating indecision between default or normal status.*
- *Status: This is the status assigned and predicted by the analyst, which correlates with the previous probability. The possible values are default, normal, or default/normal. The latter occurs when the analyst is undecided (50% probability).*
- *Reasons: The analyst provides the top three reasons for the previous prediction. For each reason, the analyst identifies the variables used in their reasoning, the values of the variables that would have led to the opposite prediction, and any interactions with other variables if applicable.*

For the evaluation, the analyst considers the following variables:

- *V1: A binary variable that takes the value 1 when the client has a delinquency of more than 30 days in the system and the month before the observation.*
- *V2: A binary variable that takes the value 1 when the client has a mortgage guarantee and 0 otherwise.*
- *V3: Debt-to-income ratio for the month, excluding off balance sheet exposures and commercial debt.*
- *V4: Days of delinquency at the bank in the last 12 months, coded as follows: 0 equals 0 days, 1 equals 1 to 30 days, 2 equals 31 to 60 days, 3 equals 61 to 89 days, and 4 or more equals over 90 days of delinquency.*
- *V5: Financial burden for the month, including off balance sheet exposures but excluding commercial debt.*

- *V6: Financial burden for the month, including off balance sheet exposures and commercial debt.*
- *V7: Ratio of the observed month's debt relative to the average of the last 12 months, considering all debt in the system.*
- *V8: Days of delinquency at the bank in the last 6 months, coded as follows: 0 equals 0 days, 1 equals 1 to 30 days, 2 equals 31 to 60 days, 3 equals 61 to 89 days, and 4 or more equals over 90 days of delinquency.*
- *V9: Average days of delinquency at the bank in the last 3 months, coded as follows: 0 equals 0 days, less than or equal to 1 equals 1 to 30 days, between 1 and 2 equals 31 to 60 days, between 2 and 3 equals 61 to 89 days, and 3 or more equals over 90 days of delinquency.*
- *V10: A binary variable that takes the value 1 when the client has a delinquency of over 90 days in the system and between 3 to 1 month before the observation.*

In the following, you will be provided with a table containing evaluations made by the analyst, with the following columns:

- *ID: a fictitious ID assigned to the debtor.*
- *Variables: the variables observed at the time the evaluation was requested from the analyst.*
- *Probability: predicted probability of default.*
- *Status: status assigned by the analyst.*
- *Reasons: reasons for the analyst's prediction.*

Your task will then be to rate the evaluations made by the analyst on a scale from 1 to 7, where 1 means that the evaluation and reasons provided are not at all reasonable according to the observed variables, and thus "very poor," while 7 means that the evaluation is highly appropriate both in the assignment of status and in the reasons and their relationship to the debtor's variables, and thus "very good."

For this, you will need to fill in the "Rating" column with the score assigned by you. Finally, there is a last column, titled "Observations," for any comments you may have on each case.

Please complete this exercise thoughtfully. Your evaluation is key to the research objective.

Evaluations to be performed:

The table with the evaluations to be performed is as follows:

ID	*VARIABLES*	*PROBABILITY*	*STATUS*	*REASONS*	*RATING*	*OBSERVATIONS*

Appendix 2: Results of the hypothesis test for the average difference in AUROC

	EX - 1	EX - 2	EX - 3	EX - 4	EX - 5	EX - 6	EX - 7	EX - 8	EX - 9	EX - 10	EX - 11	EX - 12	EX - 13	EX - 14	EX - 15	EX - 16	EX - 17	EX - 18	EX - 19	EX - 20	EX - 21	EX - 22	EX - 23	EX - 24	EX - 25	EX - 26	EX - 27	EX - 28	EX - 29	EX - 30	EX - 31	EX - 32	EX - 33	EX - 34
EX - 1	-	**-2,6%***	**5,5%***	**5,7%***	**5,6%***	**12,9%***	**8,1%***	**6,1%***	**9,7%***	**9,8%***	**6,3%***	**7,8%***	**6,6%***	**5,7%***	**8,0%***	**9,7%***	**6,0%***	**10,7%***	**7,3%***	**9,5%***	**6,4%***	**12,1%***	**9,6%***	**8,9%***	**3,8%***	**9,6%***	**13,4%***	**9,1%***	-1,1%	-1,2%	-0,9%	-0,9%	0,2%	**6,0%***
EX - 2	**2,6%***	-	**8,1%***	**8,3%***	**8,2%***	**15,5%***	**10,7%***	**8,7%***	**12,3%***	**12,4%***	**8,9%***	**10,4%***	**9,1%***	**8,3%***	**10,5%***	**12,3%***	**8,6%***	**13,2%***	**9,9%***	**12,1%***	**9,0%***	**14,7%***	**12,2%***	**11,5%***	**6,4%***	**12,2%***	**16,0%***	**11,7%***	1,5%	1,4%	1,7%	1,7%	**2,8%***	**8,6%***
EX - 3	**-5,5%***	**-8,1%***	-	0,2%	0,1%	**7,4%***	2,6%	0,6%	**4,2%***	**4,3%***	0,8%	2,3%	1,0%	0,2%	2,4%	**4,2%***	0,5%	**5,1%***	1,8%	**4,0%***	0,9%	**6,6%***	**4,1%***	**3,4%***	-1,7%	**4,1%***	**7,9%***	**3,6%***	**-6,6%***	**-6,7%***	**-6,4%***	**-6,4%***	**-5,3%***	0,5%
EX - 4	**-5,7%***	**-8,3%***	-0,2%	-	-0,1%	**7,2%***	**2,4%***	0,4%	**4,0%***	**4,1%***	0,6%	**2,1%***	0,8%	0,0%	2,3%	**4,0%***	0,3%	**4,9%***	1,6%	**3,8%***	0,7%	**6,4%***	**3,9%***	**3,2%***	-1,9%	**3,9%***	**7,7%***	**3,4%***	**-6,8%***	**-6,9%***	**-6,6%***	**-6,6%***	**-5,5%***	0,3%
EX - 5	**-5,6%***	**-8,2%***	-0,1%	0,1%	-	**7,3%***	**2,5%***	0,4%	**4,1%***	**4,2%***	0,7%	2,2%	0,9%	0,1%	2,3%	**4,1%***	0,3%	**5,0%***	1,6%	**3,8%***	0,8%	**6,5%***	**3,9%***	**3,3%***	-1,8%	**3,9%***	**7,8%***	**3,5%***	**-6,7%***	**-6,8%***	**-6,5%***	**-6,5%***	**-5,4%***	0,4%
EX - 6	**-12,9%***	**-15,5%***	**-7,4%***	**-7,2%***	**-7,3%***	-	**-4,8%***	**-6,8%***	**-3,2%***	**-3,1%***	**-6,6%***	**-5,1%***	**-6,3%***	**-7,2%***	**-4,9%***	-3,2%	**-6,9%***	-2,2%	**-5,6%***	**-3,4%***	**-6,5%***	-0,8%	**-3,3%***	**-4,0%***	**-9,1%***	**-3,3%***	0,5%	**-3,8%***	**-14,0%***	**-14,0%***	**-13,8%***	**-13,7%***	**-12,6%***	**-6,9%***
EX - 7	**-8,1%***	**-10,7%***	-2,6%	**-2,4%***	**-2,5%***	**4,8%***	-	**-2,0%***	**1,6%***	**1,7%***	-1,8%	-0,3%	-1,5%	**-2,4%***	-0,1%	1,6%	-2,1%	**2,6%***	-0,8%	1,4%	-1,7%	**4,0%***	1,5%	0,8%	**-4,3%***	1,5%	**5,3%***	1,0%	**-9,2%***	**-9,3%***	**-9,0%***	**-8,9%***	**-7,8%***	**-2,1%***
EX - 8	**-6,1%***	**-8,7%***	-0,6%	-0,4%	-0,4%	**6,8%***	**2,0%***	-	**3,7%***	**3,7%***	0,2%	1,7%	0,5%	-0,4%	1,9%	**3,6%***	-0,1%	**4,6%***	1,2%	**3,4%***	0,3%	**6,1%***	**3,5%***	**2,8%***	-2,3%	**3,5%***	**7,3%***	**3,0%***	**-7,2%***	**-7,2%***	**-7,0%***	**-6,9%***	**-5,8%***	-0,1%
EX - 9	**-9,7%***	**-12,3%***	**-4,2%***	**-4,0%***	**-4,1%***	**3,2%***	**-1,6%***	**-3,7%***	-	0,1%	**-3,4%***	**-1,9%***	**-3,2%***	**-4,0%***	-1,8%	0,0%	**-3,8%***	0,9%	**-2,5%***	-0,3%	**-3,3%***	2,4%	-0,2%	-0,8%	**-5,9%***	-0,2%	**3,7%***	-0,6%	**-10,8%***	**-10,9%***	**-10,6%***	**-10,6%***	**-9,5%***	**-3,7%***
EX - 10	**-9,8%***	**-12,4%***	**-4,3%***	**-4,1%***	**-4,2%***	**3,1%***	**-1,7%***	**-3,7%***	-0,1%	-	**-3,5%***	**-2,0%***	**-3,3%***	**-4,1%***	-1,9%	-0,1%	**-3,8%***	0,8%	-2,5%	-0,3%	**-3,4%***	2,3%	-0,2%	-0,9%	**-6,0%***	-0,2%	**3,6%***	-0,7%	**-10,9%***	**-11,0%***	**-10,7%***	**-10,7%***	**-9,6%***	**-3,8%***
EX - 11	**-6,3%***	**-8,9%***	-0,8%	-0,6%	-0,7%	**6,6%***	1,8%	-0,2%	**3,4%***	**3,5%***	-	1,5%	0,2%	-0,6%	1,6%	**3,4%***	-0,3%	**4,3%***	1,0%	**3,2%***	0,1%	**5,8%***	**3,3%***	**2,6%***	**-2,5%***	**3,3%***	**7,1%***	**2,8%***	**-7,4%***	**-7,5%***	**-7,2%***	**-7,2%***	**-6,1%***	-0,3%
EX - 12	**-7,8%***	**-10,4%***	-2,3%	**-2,1%***	-2,2%	**5,1%***	0,3%	-1,7%	**1,9%***	**2,0%***	-1,5%	-	-1,3%	**-2,1%***	0,1%	1,9%	-1,8%	**2,8%***	-0,5%	1,7%	-1,4%	**4,3%***	1,7%	1,1%	**-4,0%***	1,7%	**5,6%***	1,3%	**-8,9%***	**-9,0%***	**-8,7%***	**-8,7%***	**-7,6%***	**-1,8%***
EX - 13	**-6,6%***	**-9,1%***	-1,0%	-0,8%	-0,9%	**6,3%***	1,5%	-0,5%	**3,2%***	**3,3%***	-0,2%	1,3%	-	-0,8%	1,4%	**3,1%***	-0,6%	**4,1%***	**0,7%***	**2,9%***	-0,1%	**5,6%***	**3,0%***	**2,4%***	**-2,7%***	**3,0%***	**6,8%***	**2,5%***	**-7,7%***	**-7,7%***	**-7,5%***	**-7,4%***	**-6,3%***	-0,6%
EX - 14	**-5,7%***	**-8,3%***	-0,2%	0,0%	-0,1%	**7,2%***	**2,4%***	0,4%	**4,0%***	**4,1%***	0,6%	**2,1%***	0,8%	-	**2,3%***	**4,0%***	0,3%	**5,0%***	1,6%	**3,8%***	0,7%	**6,4%***	**3,9%***	**3,2%***	**-1,9%***	**3,9%***	**7,7%***	**3,4%***	**-6,8%***	**-6,9%***	**-6,6%***	**-6,6%***	**-5,5%***	0,3%
EX - 15	**-8,0%***	**-10,5%***	-2,4%	-2,3%	-2,3%	**4,9%***	0,1%	-1,9%	1,8%	1,9%	-1,6%	-0,1%	-1,4%	**-2,3%***	-	1,7%	-2,0%	**2,7%***	-0,7%	**1,5%***	-1,6%	**4,2%***	1,6%	1,0%	**-4,1%***	1,6%	**5,4%***	1,1%	**-9,1%***	**-9,1%***	**-8,9%***	**-8,8%***	**-7,7%***	-2,0%
EX - 16	**-9,7%***	**-12,3%***	**-4,2%***	**-4,0%***	**-4,1%***	3,2%	-1,6%	**-3,6%***	0,0%	0,1%	**-3,4%***	-1,9%	**-3,1%***	**-4,0%***	-1,7%	-	**-3,7%***	1,0%	**-2,4%***	-0,2%	**-3,3%***	**2,4%***	-0,1%	-0,8%	**-5,9%***	-0,1%	**3,7%***	-0,6%	**-10,8%***	**-10,8%***	**-10,6%***	**-10,5%***	**-9,4%***	**-3,7%***
EX - 17	**-6,0%***	**-8,6%***	-0,5%	-0,3%	-0,3%	**6,9%***	2,1%	0,1%	**3,8%***	**3,8%***	0,3%	1,8%	0,6%	-0,3%	2,0%	**3,7%***	-	**4,7%***	1,3%	**3,5%***	0,4%	**6,2%***	**3,6%***	**3,0%***	**-2,2%***	**3,6%***	**7,4%***	**3,1%***	**-7,1%***	**-7,1%***	**-6,9%***	**-6,8%***	**-5,7%***	0,0%
EX - 18	**-10,7%***	**-13,2%***	**-5,1%***	**-4,9%***	**-5,0%***	2,2%	**-2,6%***	**-4,6%***	-0,9%	-0,8%	**-4,3%***	**-2,8%***	**-4,1%***	**-5,0%***	**-2,7%***	-1,0%	**-4,7%***	-	**-3,4%***	-1,2%	**-4,2%***	1,5%	-1,1%	-1,7%	**-6,8%***	-1,1%	**2,7%***	-1,6%	**-11,8%***	**-11,8%***	**-11,6%***	**-11,5%***	**-10,4%***	**-4,7%***
EX - 19	**-7,3%***	**-9,9%***	-1,8%	-1,6%	-1,6%	**5,6%***	0,8%	-1,2%	**2,5%***	2,5%	-1,0%	0,5%	**-0,7%***	-1,6%	0,7%	**2,4%***	-1,3%	**3,4%***	-	**2,2%***	-0,9%	**4,9%***	2,3%	1,7%	**-3,5%***	2,3%	**6,1%***	1,8%	**-8,4%***	**-8,4%***	**-8,2%***	**-8,1%***	**-7,0%***	-1,3%
EX - 20	**-9,5%***	**-12,1%***	**-4,0%***	**-3,8%***	**-3,8%***	**3,4%***	-1,4%	**-3,4%***	0,3%	0,3%	**-3,2%***	-1,7%	**-2,9%***	**-3,8%***	**-1,5%***	0,2%	**-3,5%***	1,2%	**-2,2%***	-	**-3,1%***	**2,7%***	0,1%	-0,6%	**-5,7%***	0,1%	**3,9%***	-0,4%	**-10,6%***	**-10,6%***	**-10,4%***	**-10,3%***	**-9,2%***	**-3,5%***
EX - 21	**-6,4%***	**-9,0%***	-0,9%	-0,7%	-0,8%	**6,5%***	1,7%	-0,3%	**3,3%***	**3,4%***	-0,1%	1,4%	0,1%	-0,7%	1,6%	**3,3%***	-0,4%	**4,2%***	0,9%	**3,1%***	-	**5,7%***	**3,2%***	**2,5%***	**-2,6%***	**3,2%***	**7,0%***	**2,7%***	**-7,5%***	**-7,6%***	**-7,3%***	**-7,3%***	**-6,2%***	-0,4%
EX - 22	**-12,1%***	**-14,7%***	**-6,6%***	**-6,4%***	**-6,5%***	0,8%	**-4,0%***	**-6,1%***	-2,4%	-2,3%	**-5,8%***	**-4,3%***	**-5,6%***	**-6,4%***	**-4,2%***	**-2,4%***	**-6,2%***	-1,5%	**-4,9%***	**-2,7%***	**-5,7%***	-	**-2,6%***	**-3,2%***	**-8,3%***	**-2,6%***	1,3%	**-3,0%***	**-13,2%***	**-13,3%***	**-13,1%***	**-13,0%***	**-11,9%***	**-6,2%***
EX - 23	**-9,6%***	**-12,2%***	**-4,1%***	**-3,9%***	**-3,9%***	**3,3%***	-1,5%	**-3,5%***	0,2%	0,2%	**-3,3%***	-1,7%	**-3,0%***	**-3,9%***	-1,6%	0,1%	**-3,6%***	1,1%	-2,3%	-0,1%	**-3,2%***	**2,6%***	-	-0,6%	**-5,7%***	**0,0%***	**3,8%***	-0,5%	**-10,7%***	**-10,7%***	**-10,5%***	**-10,4%***	**-9,3%***	**-3,6%***
EX - 24	**-8,9%***	**-11,5%***	**-3,4%***	**-3,2%***	**-3,3%***	**4,0%***	-0,8%	**-2,8%***	0,8%	0,9%	**-2,6%***	-1,1%	**-2,4%***	**-3,2%***	-1,0%	0,8%	**-3,0%***	1,7%	-1,7%	0,6%	**-2,5%***	**3,2%***	0,6%	-	**-5,1%***	0,6%	**4,5%***	0,2%	**-10,0%***	**-10,1%***	**-9,8%***	**-9,8%***	**-8,7%***	**-2,9%***
EX - 25	**-3,8%***	**-6,4%***	1,7%	1,9%	1,8%	**9,1%***	**4,3%***	2,3%	**5,9%***	**6,0%***	**2,5%***	**4,0%***	**2,7%***	**1,9%***	**4,1%***	**5,9%***	**2,2%***	**6,8%***	**3,5%***	**5,7%***	**2,6%***	**8,3%***	**5,7%***	**5,1%***	-	**5,7%***	**9,6%***	**5,3%***	**-4,9%***	**-5,0%***	**-4,7%***	**-4,7%***	**-3,6%***	2,2%
EX - 26	**-9,6%***	**-12,2%***	**-4,1%***	**-3,9%***	**-3,9%***	**3,3%***	-1,5%	**-3,5%***	0,2%	0,2%	**-3,3%***	-1,7%	**-3,0%***	**-3,9%***	-1,6%	0,1%	**-3,6%***	1,1%	-2,3%	-0,1%	**-3,2%***	**2,6%***	**0,0%***	-0,6%	**-5,7%***	-	**3,8%***	-0,5%	**-10,7%***	**-10,7%***	**-10,5%***	**-10,4%***	**-9,3%***	**-3,6%***
EX - 27	**-13,4%***	**-16,0%***	**-7,9%***	**-7,7%***	**-7,8%***	-0,5%	**-5,3%***	**-7,3%***	**-3,7%***	**-3,6%***	**-7,1%***	**-5,6%***	**-6,8%***	**-7,7%***	**-5,4%***	**-3,7%***	**-7,4%***	**-2,7%***	**-6,1%***	**-3,9%***	**-7,0%***	-1,3%	**-3,8%***	**-4,5%***	**-9,6%***	**-3,8%***	-	**-4,3%***	**-14,5%***	**-14,6%***	**-14,3%***	**-14,2%***	**-13,1%***	**-7,4%***
EX - 28	**-9,1%***	**-11,7%***	**-3,6%***	**-3,4%***	**-3,5%***	**3,8%***	-1,0%	**-3,0%***	0,6%	0,7%	**-2,8%***	-1,3%	**-2,5%***	**-3,4%***	-1,1%	0,6%	**-3,1%***	1,6%	-1,8%	0,4%	**-2,7%***	**3,0%***	0,5%	-0,2%	**-5,3%***	0,5%	**4,3%***	-	**-10,2%***	**-10,3%***	**-10,0%***	**-10,0%***	**-8,9%***	**-3,1%***
EX - 29	1,1%	-1,5%	**6,6%***	**6,8%***	**6,7%***	**14,0%***	**9,2%***	**7,2%***	**10,8%***	**10,9%***	**7,4%***	**8,9%***	**7,7%***	**6,8%***	**9,1%***	**10,8%***	**7,1%***	**11,8%***	**8,4%***	**10,6%***	**7,5%***	**13,2%***	**10,7%***	**10,0%***	**4,9%***	**10,7%***	**14,5%***	**10,2%***	-	0,0%	0,2%	0,3%	1,4%	**7,1%***
EX - 30	1,2%	-1,4%	**6,7%***	**6,9%***	**6,8%***	**14,0%***	**9,3%***	**7,2%***	**10,9%***	**11,0%***	**7,5%***	**9,0%***	**7,7%***	**6,9%***	**9,1%***	**10,8%***	**7,1%***	**11,8%***	**8,4%***	**10,6%***	**7,6%***	**13,3%***	**10,7%***	**10,1%***	**5,0%***	**10,7%***	**14,6%***	**10,3%***	0,0%	-	0,2%	0,3%	**1,4%***	**7,1%***
EX - 31	0,9%	-1,7%	**6,4%***	**6,6%***	**6,5%***	**13,8%***	**9,0%***	**7,0%***	**10,6%***	**10,7%***	**7,2%***	**8,7%***	**7,5%***	**6,6%***	**8,9%***	**10,6%***	**6,9%***	**11,6%***	**8,2%***	**10,4%***	**7,3%***	**13,1%***	**10,5%***	**9,8%***	**4,7%***	**10,5%***	**14,3%***	**10,0%***	-0,2%	-0,2%	-	0,1%	1,2%	**6,9%***
EX - 32	0,9%	-1,7%	**6,4%***	**6,6%***	**6,5%***	**13,7%***	**8,9%***	**6,9%***	**10,6%***	**10,7%***	**7,2%***	**8,7%***	**7,4%***	**6,6%***	**8,8%***	**10,5%***	**6,8%***	**11,5%***	**8,1%***	**10,3%***	**7,3%***	**13,0%***	**10,4%***	**9,8%***	**4,7%***	**10,4%***	**14,2%***	**10,0%***	-0,3%	-0,3%	-0,1%	-	1,1%	**6,8%***
EX - 33	-0,2%	**-2,8%***	**5,3%***	**5,5%***	**5,4%***	**12,6%***	**7,8%***	**5,8%***	**9,5%***	**9,6%***	**6,1%***	**7,6%***	**6,3%***	**5,5%***	**7,7%***	**9,4%***	**5,7%***	**10,4%***	**7,0%***	**9,2%***	**6,2%***	**11,9%***	**9,3%***	**8,7%***	**3,6%***	**9,3%***	**13,1%***	**8,9%***	-1,4%	**-1,4%***	-1,2%	-1,1%	-	**5,7%***
EX - 34	**-6,0%***	**-8,6%***	-0,5%	-0,3%	-0,4%	**6,9%***	**2,1%***	0,1%	**3,7%***	**3,8%***	0,3%	**1,8%***	0,6%	-0,3%	2,0%	**3,7%***	0,0%	**4,7%***	1,3%	**3,5%***	0,4%	**6,2%***	**3,6%***	**2,9%***	-2,2%	**3,6%***	**7,4%***	**3,1%***	**-7,1%***	**-7,1%***	**-6,9%***	**-6,8%***	**-5,7%***	-

Appendix 3: Results of the hypothesis test for the average difference in AVGPREC

	EX - 1	EX - 2	EX - 3	EX - 4	EX - 5	EX - 6	EX - 7	EX - 8	EX - 9	EX - 10	EX - 11	EX - 12	EX - 13	EX - 14	EX - 15	EX - 16	EX - 17	EX - 18	EX - 19	EX - 20	EX - 21	EX - 22	EX - 23	EX - 24	EX - 25	EX - 26	EX - 27	EX - 28	EX - 29	EX - 30	EX - 31	EX - 32	EX - 33	EX - 34
EX - 1	-	**-5,9%***	**12,0%***	**10,8%***	**10,7%***	**17,0%***	**16,4%***	**10,4%***	**13,0%***	**16,0%***	**14,1%***	**14,6%***	**12,9%***	**14,5%***	**16,4%***	**18,1%***	**14,9%***	**12,8%***	**15,4%***	**16,9%***	**13,8%***	**19,1%***	**14,7%***	**20,9%***	**10,5%***	**14,7%***	**19,7%***	**19,7%***	-1,4%	-1,3%	-1,5%	-1,4%	1,6%	**17,9%***
EX - 2	**5,9%***	-	**17,9%***	**16,7%***	**16,6%***	**22,9%***	**22,3%***	**16,3%***	**18,9%***	**21,9%***	**20,0%***	**20,5%***	**18,8%***	**20,5%***	**22,3%***	**24,0%***	**20,9%***	**18,7%***	**21,3%***	**22,8%***	**19,7%***	**25,0%***	**20,6%***	**26,8%***	**16,4%***	**20,6%***	**25,6%***	**25,6%***	**4,5%***	**4,6%***	**4,4%***	**4,5%***	**7,5%***	**23,8%***
EX - 3	**-12,0%***	**-17,9%***	-	-1,2%	-1,4%	**5,0%***	**4,4%***	-1,6%	1,0%	**4,0%***	2,1%	**2,6%***	0,9%	**2,5%***	**4,4%***	**6,1%***	**2,9%***	0,8%	3,4%	**4,9%***	1,7%	**7,1%***	**2,6%***	**8,9%***	-1,6%	**2,6%***	**7,6%***	**7,7%***	**-13,4%***	**-13,4%***	**-13,5%***	**-13,4%***	**-10,4%***	**5,9%***
EX - 4	**-10,8%***	**-16,7%***	1,2%	-	-0,2%	**6,2%***	**5,6%***	-0,4%	**2,2%***	**5,2%***	**3,3%***	**3,8%***	2,1%	**3,7%***	**5,6%***	**7,3%***	**4,1%***	2,0%	**4,6%***	**6,1%***	2,9%	**8,3%***	**3,8%***	**10,1%***	-0,4%	**3,8%***	**8,8%***	**8,9%***	**-12,2%***	**-12,2%***	**-12,3%***	**-12,2%***	**-9,2%***	**7,1%***
EX - 5	**-10,7%***	**-16,6%***	1,4%	0,2%	-	**6,3%***	**5,7%***	-0,2%	2,3%	**5,3%***	**3,4%***	**4,0%***	2,2%	**3,9%***	**5,8%***	**7,4%***	**4,3%***	2,1%	**4,7%***	**6,2%***	3,1%	**8,5%***	**4,0%***	**10,3%***	-0,2%	**4,0%***	**9,0%***	**9,0%***	**-12,1%***	**-12,0%***	**-12,2%***	**-12,1%***	**-9,1%***	**7,2%***
EX - 6	**-17,0%***	**-22,9%***	**-5,0%***	**-6,2%***	**-6,3%***	-	-0,6%	**-6,6%***	**-4,0%***	-1,0%	**-2,9%***	-2,4%	**-4,1%***	-2,5%	-0,6%	1,1%	-2,1%	**-4,2%***	-1,6%	-0,1%	-3,2%	2,1%	-2,3%	**3,9%***	**-6,6%***	-2,3%	**2,7%***	2,7%	**-18,4%***	**-18,4%***	**-18,5%***	**-18,4%***	**-15,4%***	0,9%
EX - 7	**-16,4%***	**-22,3%***	**-4,4%***	**-5,6%***	**-5,7%***	0,6%	-	**-6,0%***	**-3,4%***	-0,4%	-2,3%	**-1,8%***	**-3,5%***	**-1,8%***	0,0%	1,7%	-1,4%	**-3,6%***	-1,0%	0,5%	-2,6%	**2,8%***	-1,7%	**4,6%***	**-5,9%***	-1,7%	**3,3%***	**3,3%***	**-17,8%***	**-17,7%***	**-17,9%***	**-17,8%***	**-14,8%***	1,5%
EX - 8	**-10,4%***	**-16,3%***	1,6%	0,4%	0,2%	**6,6%***	**6,0%***	-	**2,6%***	**5,6%***	**3,7%***	**4,2%***	2,5%	**4,1%***	**6,0%***	**7,7%***	**4,5%***	2,4%	**5,0%***	**6,5%***	3,3%	**8,7%***	**4,2%***	**10,5%***	0,0%	**4,2%***	**9,2%***	**9,3%***	**-11,8%***	**-11,8%***	**-11,9%***	**-11,8%***	**-8,8%***	**7,4%***
EX - 9	**-13,0%***	**-18,9%***	-1,0%	**-2,2%***	-2,3%	**4,0%***	**3,4%***	**-2,6%***	-	**3,0%***	1,1%	1,6%	-0,1%	1,5%	**3,4%***	**5,1%***	1,9%	-0,2%	2,4%	3,9%	0,8%	**6,1%***	1,7%	**7,9%***	-2,6%	1,7%	**6,7%***	**6,7%***	**-14,4%***	**-14,4%***	**-14,5%***	**-14,4%***	**-11,4%***	**4,9%***
EX - 10	**-16,0%***	**-21,9%***	**-4,0%***	**-5,2%***	**-5,3%***	-1,0%	0,4%	**-5,6%***	**-3,0%***	-	**-1,9%***	-1,4%	**-3,1%***	-1,5%	0,4%	2,1%	-1,1%	-3,2%	-0,6%	0,9%	-2,2%	3,1%	-1,3%	**4,9%***	**-5,5%***	-1,3%	**3,7%***	**3,7%***	**-17,4%***	**-17,3%***	**-17,5%***	**-17,4%***	**-14,4%***	1,9%
EX - 11	**-14,1%***	**-20,0%***	-2,1%	**-3,3%***	**-3,4%***	**2,9%***	2,3%	**-3,7%***	-1,1%	**1,9%***	-	0,5%	-1,2%	0,4%	2,3%	**4,0%***	0,8%	-1,3%	1,3%	2,8%	-0,3%	**5,0%***	0,6%	**6,8%***	**-3,7%***	0,6%	**5,6%***	**5,6%***	**-15,5%***	**-15,5%***	**-15,6%***	**-15,5%***	**-12,5%***	**3,8%***
EX - 12	**-14,6%***	**-20,5%***	**-2,6%***	**-3,8%***	**-4,0%***	2,4%	**1,8%***	**-4,2%***	-1,6%	1,4%	-0,5%	-	-1,7%	-0,1%	1,8%	**3,5%***	0,3%	-1,8%	0,8%	2,3%	-0,9%	**4,5%***	0,0%	**6,3%***	**-4,2%***	0,0%	**5,0%***	**5,1%***	**-16,0%***	**-16,0%***	**-16,1%***	**-16,0%***	**-13,0%***	**3,2%***
EX - 13	**-12,9%***	**-18,8%***	-0,9%	-2,1%	-2,2%	**4,1%***	**3,5%***	-2,5%	0,1%	**3,1%***	1,2%	1,7%	-	1,6%	**3,5%***	**5,2%***	2,0%	-0,1%	**2,5%***	**4,0%***	0,9%	**6,2%***	1,8%	**8,0%***	-2,4%	1,8%	**6,8%***	**6,8%***	**-14,3%***	**-14,2%***	**-14,4%***	**-14,3%***	**-11,3%***	**5,0%***
EX - 14	**-14,5%***	**-20,5%***	**-2,5%***	**-3,7%***	**-3,9%***	2,5%	**1,8%***	**-4,1%***	-1,5%	1,5%	-0,4%	0,1%	-1,6%	-	1,9%	**3,5%***	0,4%	-1,8%	0,8%	2,3%	-0,8%	**4,6%***	0,1%	**6,4%***	**-4,1%***	0,1%	**5,1%***	**5,2%***	**-15,9%***	**-15,9%***	**-16,1%***	**-15,9%***	**-12,9%***	**3,3%***
EX - 15	**-16,4%***	**-22,3%***	**-4,4%***	**-5,6%***	**-5,8%***	0,6%	0,0%	**-6,0%***	**-3,4%***	-0,4%	-2,3%	-1,8%	**-3,5%***	-1,9%	-	1,7%	-1,5%	**-3,6%***	-1,0%	0,5%	-2,7%	**2,7%***	-1,8%	**4,5%***	**-6,0%***	-1,8%	**3,2%***	**3,3%***	**-17,8%***	**-17,8%***	**-17,9%***	**-17,8%***	**-14,8%***	1,5%
EX - 16	**-18,1%***	**-24,0%***	**-6,1%***	**-7,3%***	**-7,4%***	-1,1%	-1,7%	**-7,7%***	**-5,1%***	-2,1%	**-4,0%***	**-3,5%***	**-5,2%***	**-3,5%***	-1,7%	-	**-3,1%***	**-5,3%***	**-2,7%***	-1,2%	**-4,3%***	1,1%	**-3,4%***	2,9%	**-7,6%***	**-3,4%***	1,6%	1,6%	**-19,5%***	**-19,4%***	**-19,6%***	**-19,5%***	**-16,5%***	-0,2%
EX - 17	**-14,9%***	**-20,9%***	**-2,9%***	**-4,1%***	**-4,3%***	2,1%	1,4%	**-4,5%***	-1,9%	1,1%	-0,8%	-0,3%	-2,0%	-0,4%	1,5%	**3,1%***	-	-2,2%	0,4%	1,9%	-1,2%	**4,2%***	-0,3%	**6,0%***	**-4,5%***	-0,3%	**4,7%***	**4,8%***	**-16,3%***	**-16,3%***	**-16,5%***	**-16,3%***	**-13,3%***	**2,9%***
EX - 18	**-12,8%***	**-18,7%***	-0,8%	-2,0%	-2,1%	**4,2%***	**3,6%***	-2,4%	0,2%	3,2%	1,3%	1,8%	0,1%	1,8%	**3,6%***	**5,3%***	2,2%	-	**2,6%***	4,1%	1,0%	**6,3%***	1,9%	**8,1%***	-2,3%	1,9%	**6,9%***	**6,9%***	**-14,2%***	**-14,1%***	**-14,3%***	**-14,2%***	**-11,2%***	**5,1%***
EX - 19	**-15,4%***	**-21,3%***	-3,4%	**-4,6%***	**-4,7%***	1,6%	1,0%	**-5,0%***	-2,4%	0,6%	-1,3%	-0,8%	**-2,5%***	-0,8%	1,0%	**2,7%***	-0,4%	**-2,6%***	-	1,5%	-1,6%	**3,7%***	-0,7%	**5,6%***	**-4,9%***	-0,7%	**4,3%***	**4,3%***	**-16,8%***	**-16,7%***	**-16,9%***	**-16,8%***	**-13,8%***	**2,5%***
EX - 20	**-16,9%***	**-22,8%***	**-4,9%***	**-6,1%***	**-6,2%***	0,1%	-0,5%	**-6,5%***	-3,9%	-0,9%	-2,8%	-2,3%	**-4,0%***	-2,3%	-0,5%	1,2%	-1,9%	-4,1%	-1,5%	-	-3,1%	**2,3%***	-2,2%	4,1%	**-6,4%***	-2,2%	2,8%	**2,8%***	**-18,3%***	**-18,2%***	**-18,4%***	**-18,3%***	**-15,3%***	1,0%
EX - 21	**-13,8%***	**-19,7%***	-1,7%	-2,9%	-3,1%	3,2%	2,6%	-3,3%	-0,8%	2,2%	0,3%	0,9%	-0,9%	0,8%	2,7%	**4,3%***	1,2%	-1,0%	1,6%	3,1%	-	**5,4%***	0,9%	**7,2%***	-3,3%	0,9%	**5,9%***	**5,9%***	**-15,2%***	**-15,1%***	**-15,3%***	**-15,1%***	**-12,2%***	**4,1%***
EX - 22	**-19,1%***	**-25,0%***	**-7,1%***	**-8,3%***	**-8,5%***	-2,1%	**-2,8%***	**-8,7%***	**-6,1%***	-3,1%	**-5,0%***	**-4,5%***	**-6,2%***	**-4,6%***	**-2,7%***	-1,1%	**-4,2%***	**-6,3%***	**-3,7%***	**-2,3%***	**-5,4%***	-	**-4,5%***	1,8%	**-8,7%***	**-4,5%***	0,5%	0,6%	**-20,5%***	**-20,5%***	**-20,6%***	**-20,5%***	**-17,5%***	-1,3%
EX - 23	**-14,7%***	**-20,6%***	**-2,6%***	**-3,8%***	**-4,0%***	2,3%	1,7%	**-4,2%***	-1,7%	1,3%	-0,6%	0,0%	-1,8%	-0,1%	1,8%	**3,4%***	0,3%	-1,9%	0,7%	2,2%	-0,9%	**4,5%***	-	**6,3%***	**-4,2%***	**0,0%***	**5,0%***	**5,0%***	**-16,1%***	**-16,0%***	**-16,2%***	**-16,0%***	**-13,1%***	**3,2%***
EX - 24	**-20,9%***	**-26,8%***	**-8,9%***	**-10,1%***	**-10,3%***	**-3,9%***	**-4,6%***	**-10,5%***	**-7,9%***	**-4,9%***	**-6,8%***	**-6,3%***	**-8,0%***	**-6,4%***	**-4,5%***	-2,9%	**-6,0%***	**-8,1%***	**-5,6%***	-4,1%	**-7,2%***	-1,8%	**-6,3%***	-	**-10,5%***	**-6,3%***	-1,3%	-1,2%	**-22,3%***	**-22,3%***	**-22,5%***	**-22,3%***	**-19,3%***	**-3,1%***
EX - 25	**-10,5%***	**-16,4%***	1,6%	0,4%	0,2%	**6,6%***	**5,9%***	0,0%	2,6%	**5,5%***	**3,7%***	**4,2%***	2,4%	**4,1%***	**6,0%***	**7,6%***	**4,5%***	2,3%	**4,9%***	**6,4%***	3,3%	**8,7%***	**4,2%***	**10,5%***	-	**4,2%***	**9,2%***	**9,2%***	**-11,8%***	**-11,8%***	**-12,0%***	**-11,8%***	**-8,8%***	**7,4%***
EX - 26	**-14,7%***	**-20,6%***	**-2,6%***	**-3,8%***	**-4,0%***	2,3%	1,7%	**-4,2%***	-1,7%	1,3%	-0,6%	0,0%	-1,8%	-0,1%	1,8%	**3,4%***	0,3%	-1,9%	0,7%	2,2%	-0,9%	**4,5%***	**0,0%***	**6,3%***	**-4,2%***	-	**5,0%***	**5,0%***	**-16,1%***	**-16,0%***	**-16,2%***	**-16,0%***	**-13,1%***	**3,2%***
EX - 27	**-19,7%***	**-25,6%***	**-7,6%***	**-8,8%***	**-9,0%***	**-2,7%***	**-3,3%***	**-9,2%***	**-6,7%***	**-3,7%***	**-5,6%***	**-5,0%***	**-6,8%***	**-5,1%***	**-3,2%***	-1,6%	**-4,7%***	**-6,9%***	**-4,3%***	-2,8%	**-5,9%***	-0,5%	**-5,0%***	1,3%	**-9,2%***	**-5,0%***	-	0,0%	**-21,1%***	**-21,0%***	**-21,2%***	**-21,0%***	**-18,1%***	**-1,8%***
EX - 28	**-19,7%***	**-25,6%***	**-7,7%***	**-8,9%***	**-9,0%***	-2,7%	**-3,3%***	**-9,3%***	**-6,7%***	**-3,7%***	**-5,6%***	**-5,1%***	**-6,8%***	**-5,2%***	**-3,3%***	-1,6%	**-4,8%***	**-6,9%***	**-4,3%***	**-2,8%***	**-5,9%***	-0,6%	**-5,0%***	1,2%	**-9,2%***	**-5,0%***	0,0%	-	**-21,1%***	**-21,0%***	**-21,2%***	**-21,1%***	**-18,1%***	-1,8%
EX - 29	1,4%	**-4,5%***	**13,4%***	**12,2%***	**12,1%***	**18,4%***	**17,8%***	**11,8%***	**14,4%***	**17,4%***	**15,5%***	**16,0%***	**14,3%***	**15,9%***	**17,8%***	**19,5%***	**16,3%***	**14,2%***	**16,8%***	**18,3%***	**15,2%***	**20,5%***	**16,1%***	**22,3%***	**11,8%***	**16,1%***	**21,1%***	**21,1%***	-	0,0%	-0,1%	0,0%	**3,0%***	**19,3%***
EX - 30	1,3%	**-4,6%***	**13,4%***	**12,2%***	**12,0%***	**18,4%***	**17,7%***	**11,8%***	**14,4%***	**17,3%***	**15,5%***	**16,0%***	**14,2%***	**15,9%***	**17,8%***	**19,4%***	**16,3%***	**14,1%***	**16,7%***	**18,2%***	**15,1%***	**20,5%***	**16,0%***	**22,3%***	**11,8%***	**16,0%***	**21,0%***	**21,0%***	0,0%	-	-0,2%	0,0%	**3,0%***	**19,2%***
EX - 31	1,5%	**-4,4%***	**13,5%***	**12,3%***	**12,2%***	**18,5%***	**17,9%***	**11,9%***	**14,5%***	**17,5%***	**15,6%***	**16,1%***	**14,4%***	**16,1%***	**17,9%***	**19,6%***	**16,5%***	**14,3%***	**16,9%***	**18,4%***	**15,3%***	**20,6%***	**16,2%***	**22,5%***	**12,0%***	**16,2%***	**21,2%***	**21,2%***	0,1%	0,2%	-	0,1%	3,1%	**19,4%***
EX - 32	1,4%	**-4,5%***	**13,4%***	**12,2%***	**12,1%***	**18,4%***	**17,8%***	**11,8%***	**14,4%***	**17,4%***	**15,5%***	**16,0%***	**14,3%***	**15,9%***	**17,8%***	**19,5%***	**16,3%***	**14,2%***	**16,8%***	**18,3%***	**15,1%***	**20,5%***	**16,0%***	**22,3%***	**11,8%***	**16,0%***	**21,0%***	**21,1%***	0,0%	0,0%	-0,1%	-	3,0%	**19,3%***
EX - 33	-1,6%	**-7,5%***	**10,4%***	**9,2%***	**9,1%***	**15,4%***	**14,8%***	**8,8%***	**11,4%***	**14,4%***	**12,5%***	**13,0%***	**11,3%***	**12,9%***	**14,8%***	**16,5%***	**13,3%***	**11,2%***	**13,8%***	**15,3%***	**12,2%***	**17,5%***	**13,1%***	**19,3%***	**8,8%***	**13,1%***	**18,1%***	**18,1%***	**-3,0%***	**-3,0%***	-3,1%	-3,0%	-	**16,3%***
EX - 34	**-17,9%***	**-23,8%***	**-5,9%***	**-7,1%***	**-7,2%***	-0,9%	-1,5%	**-7,4%***	**-4,9%***	-1,9%	**-3,8%***	**-3,2%***	**-5,0%***	**-3,3%***	-1,5%	0,2%	**-2,9%***	**-5,1%***	**-2,5%***	-1,0%	**-4,1%***	1,3%	**-3,2%***	**3,1%***	**-7,4%***	**-3,2%***	**1,8%***	1,8%	**-19,3%***	**-19,2%***	**-19,4%***	**-19,3%***	**-16,3%***	-