Conversational Explanations for Bayesian Networks: An Interactive LLM-Driven Framework for Explainable AI

التفسيرات الحوارية للشبكات البايزية: إطار عمل تفاعلي مدعوم بنماذج اللغة الكبيرة للذكاء الاصطناعي القابل للتفسير

Lamis Olaiky :Syrian Virtual University, Damascus-Syria.
Corresponding: lamis.olaiky@gmail.com
Submitted date: 27 August 2026 - **Accepted date:** 29 October 2026.

ABSTRACT

Explainable Artificial Intelligence (XAI) is important for improving the interpretability of probabilistic models in decision-support applications. Although Bayesian Networks (BNs) provide probabilistic reasoning through Directed Acyclic Graphs (DAGs) and Conditional Probability Tables (CPTs), their numerical outputs can be difficult for non-technical decision-makers to interpret. Traditional explanation approaches may provide feature importance or attribution measures but do not necessarily provide contextual, natural-language explanations of probabilistic outcomes. To address this gap, this paper introduces an architectural framework that combines Bayesian Networks with Large Language Models (LLMs) to generate contextual and interactive explanations. The framework was evaluated on 300 held-out credit scoring cases, representing 30% of the dataset, after training the Bayesian Network on the remaining 70%. The Bayesian Network achieved an accuracy of 82.67%, precision of 81.04%, recall of 99.54%, and F1-score of 89.34%. The generated explanations achieved average ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores of 0.4499, 0.1685, 0.3338, and 0.0417, respectively. An LLM-as-a-Judge evaluation using Qwen 2.5 yielded an overall mean quality score of 3.11/5.00 (Faithfulness: 2.98, Clarity: 3.92, Completeness: 2.42), with an average LLM latency of 15.75 seconds per case. These results demonstrate the potential of combining probabilistic inference with constrained natural-language generation to improve the accessibility and interpretability of Bayesian network-based decision support.
Keywords: Explainable AI (XAI), Bayesian Networks, Large Language Models (LLMs), Natural Language Explanation, Decision Support Systems, Interactive Explanation

الملخص

يُعد الذكاء الاصطناعي القابل للتفسير (XAI) مهماً لتحسين قابلية تفسير النماذج الاحتمالية في تطبيقات دعم اتخاذ القرار. وعلى الرغم من أن الشبكات البايزية (BNs) توفر استدلالاً احتمالياً من خلال الرسوم البيانية الموجهة الخالية من الحلقات (DAGs) وجداول الاحتمالات الشرطية (CPTs)، فإن مخرجاتها الرقمية قد تكون صعبة التفسير بالنسبة لمتخذي القرار غير المختصين. وقد توفر أساليب التفسير التقليدية مقاييس لأهمية الخصائص أو مساهمتها، لكنها لا تقدم بالضرورة تفسيرات سياقية باللغة الطبيعية للمخرجات الاحتمالية. ولردم هذه الفجوة، تقدم هذه الورقة إطاراً معمارياً يجمع بين الشبكات البايزية ونماذج اللغة الكبيرة (LLMs) لتوليد تفسيرات سياقية وتفاعلية. يستخرج الإطار الأدلة الاحتمالية والعلاقات ذات الصلة من الشبكة البايزية ويمررها إلى نموذج اللغة من خلال قوالب أوامر منظمة ومقيدة، مما يتيح توليد تفسيرات باللغة الطبيعية تستند إلى نتائج الاستدلال البايزي. وعند تقييم الإطار ضمن حالة دراسية لتقييم الجدارة الائتمانية تشمل 300 حالة اختبار، حققت الشبكة البايزية دقة تصنيف بلغت 82.67% (بمقياس F1 بلغ 89.34%)، وبلغ متوسط مقياس ROUGE-1 مقدار 0.4499، إضافة إلى درجة جودة كلية بلغت 3.11 / 5.00 من خلال التقييم باستخدام نموذج لغوي كحَكَم (Qwen 2.5 LLM-as-a-Judge) وبمتوسط زمن استجابة بلغ 15.75 ثانية لكل حالة. وتوضح هذه النتائج إمكانية الجمع بين الاستدلال الاحتمالي وتوليد اللغة الطبيعية المقيد لتحسين سهولة الوصول إلى تفسيرات قرارات الشبكات البايزية وقابليتها للفهم
الكلمات المفتاحية: الذكاء الاصطناعي القابل للتفسير، الشبكات البايزية، نماذج اللغات الكبيرة، التفسير باللغة الطبيعية، أنظمة دعم القرار، التفسير التفاعلي

INTRODUCTION

Artificial intelligence (AI) models are increasingly used in high-stakes decision-making environments, including financial risk assessment and enterprise management [11, 22]. Bayesian Networks (BNs) provide a formal framework for probabilistic reasoning in these domains by representing variables and their conditional dependencies through Directed Acyclic Graphs (DAGs) and Conditional Probability Tables (CPTs) [10, 31, 35]. However, interpreting BN structures and probabilistic outputs can be challenging for users without specialized knowledge of probabilistic modeling. Although methods for ex-

plaining Bayesian network inference have been developed [17, 20, 36], translating model outputs into accessible explanations that support interactive exploration remains an important challenge. This challenge is particularly relevant in high-stakes settings, where users need to understand the basis and limitations of model predictions to support informed decision-making [15, 23, 33]. To address this challenge, this study investigates a conversational Explainable AI (XAI) framework that combines Bayesian probabilistic inference with large language model (LLM)-based explanation generation [18, 45]. In this architecture, the BN performs the probabilistic inference, while the LLM generates natural-language explanations based on structured applicant information, BN outputs, and instructions governing the interpretation of the network's structure. The framework also supports follow-up questions, allowing users to explore individual credit-assessment cases through conversational interaction [13, 29, 37]. The study focuses on the consistency of generated explanations with the underlying probabilistic results and network structure, as well as the limitations of using prompt constraints to guide the language model. Llama 3.2 is used as the language-generation component in the implemented system; the study does not assume that the proposed framework depends on this specific model. To guide this investigation, the study addresses the following research questions:

•RQ1: How can Bayesian Network (BN) inference outputs and structural information be used to generate natural-language explanations through a large language model?

•RQ2: To what extent are the generated explanations consistent with the numerical outputs and specified structure of the underlying Bayesian Network?

•RQ3: What limitations arise when using a large language model to provide interactive explanations and recommendations for individual credit-assessment cases?

The study investigates these questions through the implementation and evaluation of a conversational framework that separates probabilistic inference from natural-language generation. The evaluation examines the consistency of generated explanations with the BN outputs and structure, while the qualitative analysis of case scenarios explores the behavior and limitations of the conversational interface

Related Work and Background

Bayesian Networks and Interpretability Challenges:

Probabilistic graphical models, particularly Bayesian Networks (BNs), were established by Pearl [31] to model causal relationships and conditional dependencies under uncertainty using Directed Acyclic Graphs (DAGs). Subsequent surveys by Heckerman et al. [14] highlighted the efficacy of BNs across medical diagnosis, financial risk, engineering, and robotics. However, BNs suffer from exponential computational complexity and severe interpretability bottlenecks [2, 14]. Barber [2] attributed this opacity to derived variables and noisy Conditional Probability Tables (CPTs). To systematically address this cognitive load, Lacave and Díez [20] proposed a framework categorizing BN explanations by granularity (micro vs. macro), reasoning nature (causal vs. non-causal), and user personalization, demonstrating that Maximum A Posteriori (MAP) inference yields more concise human-readable explanations than full state representations. Early natural language and narrative attempts focused either on logical argumentation via support graphs [11] or scenario-based explanations in legal domain applications [40]. Nevertheless, these classical approaches remained constrained by rigid rule-based templates, high computational overhead, and limited adaptability to quantitative financial domains [40].

Explainable AI (XAI) Frameworks and Post-Hoc Interpretability:

To formalize model interpretability, Doshi-Velez and Kim [8] established three evaluation tiers: application-grounded, human-grounded, and functionally-grounded metrics. Lipton [23] delineated the philosophical dimensions of XAI, distinguishing intrinsic transparency (simulatability, decomposability, algorithmic transparency) from post-hoc explanations (textual, visual, local examples), noting that local post-hoc explanations are often the most practical solution for complex decision systems under regulatory mandates [41].

While post-hoc methods aim to illuminate deep learning models [34], they face critical structural limitations, including explanation inconsistency [19, 34], a lack of standardized evaluation benchmarks [30, 34], and the inherent trade-off between predictive accuracy and interpretability [33]. These challenges spurred a shift toward leveraging natural language generation and Large Language Models (LLMs) to communicate model behavior interactively [5, 37].

Conversational AI and LLM-Grounded Explanation Generation:

The emergence of generative AI and LLM-driven interfaces transformed data interaction [13, 46] by utilizing prompt engineering and Chain-of-Thought (CoT) reasoning to decompose complex inference into human-understandable steps [42, 46]. However, ungrounded LLMs remain prone to hallucinations

CC BY

and statistical inaccuracies [13, 16, 22], especially when decoupled from exact calculation engines [13, 42]. Specialized conversational agents have been developed to explain black-box models, such as Deep Reinforcement Learning (DRL) in healthcare and finance [24, 28], though they struggle to balance scientific rigor with user accessibility.

To improve grounding, structured text-generation approaches have emerged at the intersection of XAI and LLMs

•BN-to-Text Baselines: Keppens [17] introduced pre-LLM Natural Language Generation (NLG) for BN query results via support graphs. Sevilla et al. [36] advanced this by extracting factor arguments (sub-graph reasoning paths) into natural language, validated through human user studies in medicine.

•LLM-XAI Integration: Recent advances in explainable AI highlight the potential of leveraging LLMs to naturalize complex decision outputs, though unconstrained models remain prone to generating plausible yet unfaithful narratives [26]. To address this in financial decision-making, Dyomin et al. [9] combined traditional scoring classifiers with feature attribution tools, using an LLM strictly as a grounded translator for model-derived risk indicators.

Research Gap and Main Contributions

Previous research has addressed the generation of natural-language explanations for Bayesian Network (BN) inference through approaches that explicitly represent probabilistic reasoning. Keppens [17], for example, proposed a natural-language generation approach based on qualified support graphs to identify and communicate how observations affect a query result within a BN. Sevilla et al. [36] developed an explanation approach based on factor arguments, representing the flow of evidence toward a target variable, and evaluated its usefulness in the medical domain. These studies demonstrate that structured, model-grounded explanations of Bayesian reasoning can be generated without relying on an LLM to perform the underlying probabilistic inference. However, their primary focus is on constructing explanations of BN reasoning, rather than on the particular combination of a separate BN inference component, an LLM-based conversational interface, and an evaluation of the numerical consistency of the resulting dialogue.

A different line of research uses LLMs to communicate predictions and explanation evidence from machine-learning models. For example, Dyomin et al. [9] combine Logistic Regression and LightGBM with SHAP-based feature attribution and an LLM that translates feature contributions into natural-language explanations. This approach addresses the communication of feature-attribution results in credit-risk assessment, whereas the present study uses a BN to represent conditional probabilistic dependencies and obtain the credit-approval prediction. These approaches therefore differ in their underlying reasoning and the type of model information supplied to the language-generation component. More broadly, the use of LLMs as explanation interfaces raises the need to evaluate whether generated narratives preserve the supplied numerical results and model relationships rather than introducing unsupported claims.

Against this background, this study investigates a hybrid approach that combines BN-based probabilistic inference with LLM-generated conversational explanations for credit-assessment cases. Its focus is not on claiming that either BN explanation or LLM-based explanation is new in itself, but on implementing and evaluating their integration within a single decision-support interface, with explicit prompting constraints and quantitative and qualitative assessment of generated explanations.

The main contributions of this research are as follows:

1.A Decoupled Hybrid XAI Architecture:

BN performs probabilistic inference independently of the LLM, which generates natural-language explanations from structured applicant information, BN outputs, and explicit instructions governing the interpretation of the network's structure

2.A Constraint-Guided Prompting Protocol:

Development and evaluation of a prompt design intended to guide the local LLM in preserving the numerical results and specified relationships of the BN. The system achieved a Numerical Faithfulness Rate of 91.33% under the evaluation procedure used in this study. This result measures numerical consistency according to the adopted metric; it does not establish the elimination of hallucinations in general.

3.A Multi-Metric Evaluation Pipeline:

Evaluation of generated explanations using lexical-overlap metrics, BERTScore, numerical faithfulness, decision alignment, and LLM-based judging with Qwen 2.5. The pipeline compares generated explanations with manually constructed reference texts and assesses complementary aspects of explanation quality.

4.An Interactive Conversational Decision-Support Implementation:

Implementation of an interface that maintains applicant inputs and BN evaluation results across follow-up questions, allowing users to explore expla-

nations of individual credit-assessment cases. The implementation was evaluated using 300 test cases.

MATERIALS AND METHODS

Framework Overview and Strategic Objectives

The proposed framework establishes a structured, three-tiered architecture that integrates Bayesian Network (BN)-based probabilistic reasoning with Large Language Model (LLM)-based natural language generation for decision-support applications. Rather than relying on the LLM as an autonomous decision-maker, the architecture strictly separates the quantitative inference process from the narrative explanation layer, positioning the Bayesian Network as the primary computational engine. As illustrated in Figure 1, the operational pipeline spans three sequential functional layers: ***Data Ingestion Layer:*** Preprocesses, discretizes, and standardizes raw applicant attributes into discrete evidence states suitable for probabilistic propagation. ***Probabilistic Inference Layer:*** Serves as the Bayesian core, responsible for factorizing the directed acyclic graph (DAG), updating conditional probability tables (CPTs), and computing exact posterior marginals for the target outcome (e.g., loan approval status). ***Explanation and Natural Language Layer:*** Maps Maps structured probabilistic outputs and intermediate variable dependencies into an accessible, interactive narrative using an LLM, with prompt constraints intended to support consistency with the underlying BN structure and inference results

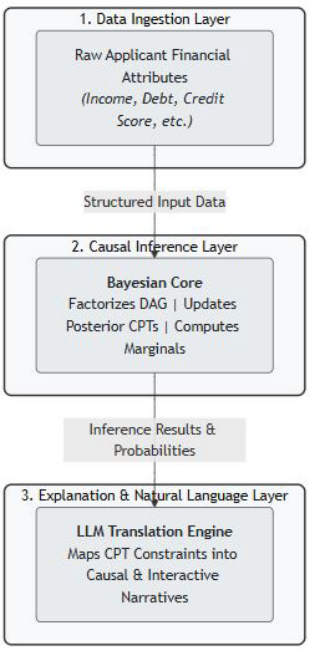


Figure 1System Architecture Diagram

Empirical Case Study: Bank Loan Approval Application

To validate and evaluate the practical implementation of this architecture, a bank loan approval decision system was adopted as an empirical case study. Loan underwriting presents an ideal domain due to its strict regulatory requirements, the necessity for auditability, and the presence of complex conditional dependencies among financial metrics. Although the framework is evaluated here in a credit-assessment context, its general architecture could be investigated in other domains, such as medical diagnostics and insurance underwriting. However, its suitability for these applications remains to be evaluated. The framework was evaluated using a publicly available credit risk dataset comprising 1,000 loan applicant records with financial and demographic attributes, where the target variable is binary loan approval (Loan). To suit the discrete requirements of Bayesian Networks, continuous features (e.g., annual income and credit score) were discretized into uniform state brackets (e.g., low, medium, high, very high) using quantile-based discretization (pd.qcut). Categorical variables were label-encoded to preserve structural dependencies while mitigating state-space explosion. The loan evaluation model uses a three-tiered hierarchical BN structure developed through a combination of data-driven feature selection and domain-informed topological constraints: ***Primary Evidence Nodes (Layer 1):***

Observed input attributes including annual_income, credit_score, existing_loans_count, employment_status, and num_dependents

Intermediate Derived Nodes (Layer 2):

Domain-expert intermediate variables—specifically financial_capacity (derived from income, employment, and dependents) and credit_worthiness (derived from credit score and active loans). These intermediate nodes act as structural abstraction layers, capturing higher-level financial concepts prior to final risk evaluation.

Target Decision Outcome Node (Layer 3):

The final classification node representing Loan Approval Status (Loan). Probabilistic Inference Engine: Bayesian Network Architecture and Pipeline The final classification node representing Loan Approval Status (Loan). **Probabilistic Inference Engine: Bayesian Network Architecture and Pipeline**

Data Preparation and Feature Discretization

This layer serves as the primary entry interface for applicant profiles. It standardizes raw continuous and categorical financial parameters-including income level, credit score, active debt burden, number of dependents, and employment status-into discrete evidence states suitable for probabilistic propagation. Continuous numerical variables (e.g.,

CC BY

credit_score and annual_income) are discretized into four uniform quantile-based state brackets using pd.qcut (e.g., categorizing credit scores into discrete evaluation tiers) to facilitate exact case classification and Bayesian Network construction. Categorical variables and state labels are encoded to preserve structural dependencies while assigning human-readable representations that support seamless integration with the subsequent natural-language explanation layer.

Domain-Engineered Intermediate Variable Construction

To incorporate domain knowledge into the Bayesian Network architecture, two derived intermediate nodes, financial_capacity and credit_worthiness, were constructed using predefined rules based on observed applicant attributes. Unlike statistically latent variables, these nodes are explicitly computed from available input features. They are subsequently incorporated into the Bayesian Network as intermediate variables, providing a hierarchical representation of financial capacity and creditworthiness within the probabilistic model. The financial_capacity node is derived from annual_income, employment_status, and num_dependents, whereas the credit_worthiness node is derived from credit_score and existing_loans_count.

Financial Capacity (Pillar_1):

Synthesized from primary evidence attributes including annual_income, employment_status, and num_dependents

$$Fin\ Capacity\ Score = Income\ bin + Employment\ bin - \mathbb{I}(Dependents\ bin \geq 2)$$

$$financial\ capacity\ score = \begin{cases} 0\ (Low\ Capacity), if\ fin\ score \leq 0 \\ 1\ (Medium\ Capacity), \quad if\ 0 < fin\ score \leq 2 \\ 2\ (High\ Capacity), \quad if\ fin\ score > 2 \end{cases}$$

1. **Credit Worthiness ($Pillar_2$):** Synthesized from `credit_score` and `existing_loans_count`.

$$Credit\ Worthiness\ Score = credit\ score\ bin - existing\ loans\ bin$$

$$credit\ worthiness = \begin{cases} 0\ (Poor\ Worthiness), if\ cred\ score \leq 0 \\ 1(Fair\ Worthiness), if\ cred\ score = 1 \\ 2(Excellent\ Worthiness), if\ cred\ score \geq 2 \end{cases}$$

Dependents Penalty Threshold Logic

The dependents condition I(dep_bin ≥ 2) enforces a symmetrical bipartition of the discretized state space {0, 1, 2, 3}. Partitioning the set into Low Family Burden ({0, 1}) and High Family Burden ({2, 3}) maintains statistical class balance while capturing a recognized financial inflection point. In retail underwriting, one dependent falls within baseline cost-of-living expectations, whereas two or more dependents significantly increase fixed household liabilities, directly impairing disposable cash flow and Debt-Service Coverage Ratios (DSCR).

Hierarchical Topology and Hybrid Structure Learning

The statistical foundation consists of a Directed Acyclic Graph G = (V, E), where vertices V represent domain variables and edges E denote direct conditional dependencies. The joint probability distribution over V factorizes according to the chain rule for Bayesian Networks:

$$P(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} P\left(V_i \mid \text{Parents}(V_i)\right)$$

Hybrid Feature Selection

To prevent exponential state-space explosion in Conditional Probability Tables (CPTs) and preserve exact inference tractability, a hybrid feature selection mechanism was executed prior to topology construction. Feature importance was quantified using a weighted composite score: Mutual Information (٪۷۰ weight) measured direct entropy reduction relative to the binary target variable (loan_approved), aligning with the information-theoretic foundations of Bayesian joint distributions. Decision Tree importance (٪۳۰ weight) evaluated non-linear conditional interactions via Gini impurity reduction. Based on this composite ranking, the top five primary predictor features were retained, while secondary and demographic attributes (age, loan_amount, gender) exhibited negligible

explanatory power and were pruned (Table 1).

$$Score = 0.70 \times Mutual\ Information + 0.30 \times Decision\ Tree\ Importance$$

Table-1 - Hybrid Feature Selection and Importance Rankings

	tree_ importance	mutual_ information	final_ score
num_ dependents	0.2347	0.0915	0.1344
credit_score	0.2369	0.0725	0.1218
existing_ loans_count	0.2238	0.0705	0.1165
employment_ status	0.2353	0.0573	0.1107
annual_ income	0.0666	0.0209	0.0346
marital_ status	0.0027	0.0020	0.0022
age	0.0000	0.0002	0.0002
loan_amount	0.0000	0.0002	0.0001
gender	0.0000	0.0002	0.0001

Structural Optimization

Purely data-driven algorithms (e.g., Hill-Climb Search with K2 score) led to spurious correlations and high parent cardinality. To preserve exact inference tractability, a three-tiered hierarchical DAG was established, where primary inputs (Tier_1) flow exclusively into the derived pillars(Tier_2), which in turn dictate the target decision outcome (loan_approved,(Tier_3))

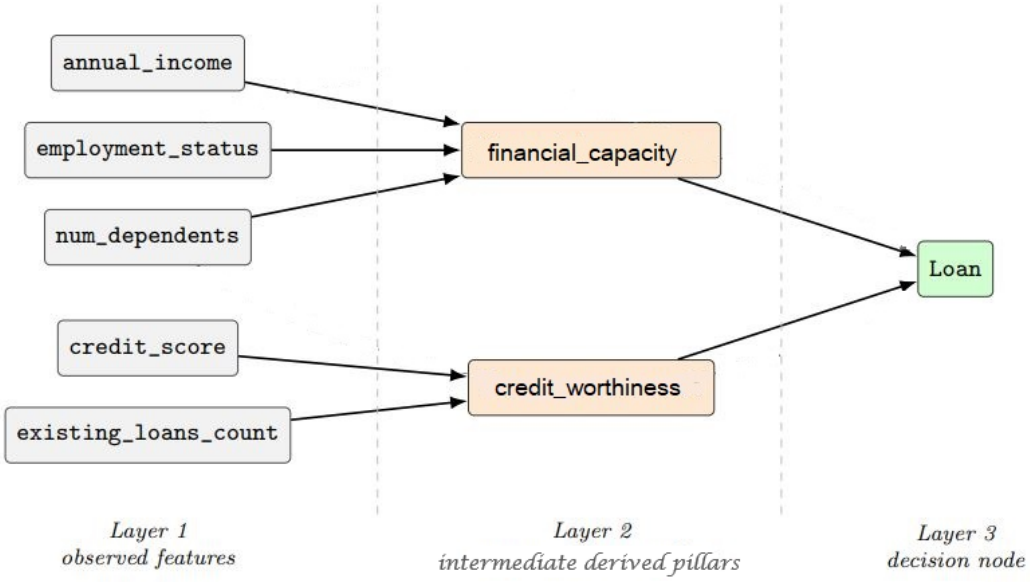


fig 2. Bayesian Network architecture

Parameter Learning and CPT Smoothing

To prevent data leakage, the feature discretization sequence (such as binning continuous variables using quantile-based boundaries via pd.qcut) and structural parameter estimation for the Conditional Probability Tables (CPTs) were executed exclusively on the 70% training partition (700 records). The resulting discretization boundaries and learned parameters were subsequently locked and applied to the independent 30% test partition (300 records). Parameter learning used a Bayesian Estimator with a Dirichlet uniform prior (BDeu) and an Equivalent Sample Size (ESS = 10) to handle sparse parent configurations. Network structural integrity and conditional sum constraints

were formally validated using check_model(). The compiled network artifact was then serialized using joblib for zero-overhead runtime loading.

$$\sum_k P(V_i = k | Parents(V_i)) = 1.0$$

Representative Conditional Probability Tables (CPTs)

To demonstrate the mathematical integrity of the parameter learning phase, representative excerpts of the learned CPTs for both intermediate mediators and the final target decision node are detailed below

Target Decision Node (loan): As shown in Table 2, the decision logic exhibits strong convergence. When both intermediate mediators (financial_capacity and credit_worthiness) are at optimal levels (High), the probability of loan approval reaches 0.9880. Conversely, critical vulnerabilities in creditworthiness (Low) shift the outcome towards rejection ($P \geq 0.50$).

Table2- Sample CPT Excerpt for Target Decision Node

Target Outcome (Loan)	Financial capacity	Credit worthiness	Conditional Probability
Rejected (0)	Low (0)	Low (0)	0.9954
Rejected (0)	Medium (1)	Low (0)	0.5
Rejected (0)	High (2)	Low (0)	0.5
Rejected (0)	Low (0)	Medium (1)	0.6239
Rejected (0)	High (2)	Medium (1)	0.0023
Approved (1)	Low (0)	Low (0)	0.0046
Approved (1)	High (2)	Low (0)	0.5
Approved (1)	Medium (1)	Low (0)	0.586
Approved (1)	Medium (1)	Medium (1)	0.9982
Approved (1)	High (2)	High (2)	0.988

Intermediate Mediators (credit_worthiness & financial_capacity):

Tables 3 and 4 illustrate how primary evidence features interact to define intermediate risk states. The BDeu smoothing estimator prevents deterministic zero-probabilities (e.g., maintaining residual probabilities such as 0.003 or 0.004), guaranteeing model adaptability during runtime inference.

Table 3 Sample CPT Excerpt for credit_worthiness Node

Derived State (credit_ worthiness)	Credit Score	Active Loans (existing_ loans_ count)	Conditional Probability
Low (0)	Low (Bin 0)	0	0.9931
Low (0)	Very High (Bin 3)	3	0.0043
Medium (1)	Low (Bin 0)	1	0.0028
Medium (1)	Medium (Bin 1)	0	0.9925
Medium (1)	Medium (Bin 1)	3	0.0029
High (2)	Low (Bin 0)	0	0.0034
High (2)	Medium (Bin 1)	1	0.003
High (2)	Very High (Bin 3)	0	0.9925
High (2)	Very High (Bin 3)	4	0.0025

Table4 - Sample CPT Excerpt for financial_capacity Node

Derived State (financial_ capacity)	Annual Income	Employment Status	Dependents (num_ dependents)	Probability
Low (0)	Low (Bin 0)	Employed	0	0.9939
Low (0)	Low (Bin 0)	Employed	3	0.9931
Low (0)	Medium (Bin 1)	Employed	0	0.0042
Low (0)	High (Bin 2)	Unemployed	1	0.0028
Low (0)	Very High (Bin 3)	Employed	1	0.0042
Medium (1)	Low (Bin 0)	Unemployed	1	0.0032
Medium (1)	Medium (Bin 1)	Employed	3	0.9927
High (2)	Medium (Bin 1)	Employed	3	0.0037
High (2)	High (Bin 2)	Employed	1	0.9909
High (2)	Very High (Bin 3)	Unemployed	1	0.9945

Operational Runtime Inference Pipeline

To bridge offline parameter learning with live execution, the system operationalizes inference through a two-stage pipeline:Offline Parameter Estimation: Prior to application runtime, deterministic aggregation rules are applied across the historical dataset (N = 1,000) to generate the intermediate pillar states, after which conditional probability distributions (CPTs) are estimated and serialized.

Online Runtime Execution (evaluate.py): During live operation, when raw applicant attributes are submitted, the application layer dynamically computes the exact pillar states. These computed pillar values are combined with the primary discretized inputs into a unified evidence dictionary E, which is injected into the VariableElimination inference engine to calculate marginal posterior probabilities.

Model Validation Setup and Evaluation Metrics

To evaluate the predictive performance of the Bayesian Network prior to integrating the LLM explanation layer, the dataset (N = 1,000) was split into a 70% training set (700 records) and a 30% test set (300 records).

The network parameters were learned using the training subset, after which the compiled model was executed on the remaining 300 test cases to evaluate its overall classification performance. Model performance is quantified using standard classification metrics: Accuracy, Precision, Recall, and F_1Score. Evaluating these probabilistic outputs verifies the network›s classification consistency before passing its inference outputs to the downstream explanation engine. (Quantitative results are presented in next Section)

Interactive Explanation Engine: LLM Integration and Conversational Framework

To translate probabilistic inference outcomes into accessible natural-language explanations while avoiding dependence on cloud-based language-model services and using prompt constraints to reduce unsupported claims [16, 44], the framework integrates a locally deployed Llama 3.2 instance orchestrated via the Ollama execution engine. Rather than allowing unconstrained natural language generation, the explanation layer enforces strict numerical grounding and constraint-guided prompt engineering to ensure robust alignment with the underlying Bayesian Network (BN).

Grounded Explanation and Translation Architecture

The runtime architecture operates locally to guarantee eliminates cloud transmission latency and strict data privacy. To balance structural determinism with interactive flexibility, the engine executes Llama 3.2 across two distinct operational pipelines:

Automated Evaluation Mode (evaluate.py): Designed for immediate, highly deterministic structured reporting following BN inference.

Interactive Conversational Mode (chat.py): Tailored for multi-turn, context-aware interrogation by risk officers.

To enforce strict numerical grounding and prevent hardware instability during local CPU execution, spe-

cific hyperparameters were tailored per operational mode, as detailed in Table 6.

table 6: Local Llama 3.2 Execution Hyperparameters Across Operational Modes

Hyperparameter / Setting	Evaluation Mode (evaluate.py)	Interactive Chat Mode (chat.py)	Operational Rationale
Model Runtime	Llama 3.2	Llama 3.2	Locally hosted via Ollama framework
CPU Thread Limit (num_thread)	4	4	Constrains thread concurrency to prevent hardware thermal throttling
Temperature (temperature)	0.1	0.0	Near-zero sampling variability to promote more consistent outputs, although this setting does not guarantee factual accuracy or eliminate hallucinations.
Top-P Sampling (top_p)	0.9	0.1	Restricted nucleus sampling for highly deterministic token selection
Maximum Tokens (num_predict)	140	250	Strict output length cap enforcing concise 3-bullet or 3–4 sentence outputs
Repetition Penalty (repeat_penalty)	Default (1.0)	1.2	Penalizes token looping during multi-turn conversational querying

Data Humanization and Decision Governance Boundaries

Prior to prompt construction, raw discrete evidence states and Bayesian posterior probabilities are converted into intuitive domain representations using a humanization translation dictionary (human_labels). This translation spans both input features and derived hierarchical pillars:

Primary Evidence Mapping: Raw discretized categories are mapped to qualitative descriptors (e.g., credit_score → {Very Poor, Poor, Good, Excellent}; employment_status → {Unemployed, Employed (Stable)})

Intermediate Derived Pillars: Hierarchical intermediate states are humanized into standardized risk capacity tiers (financial_capacity → {Low Capacity, Medium Capacity, High Capacity}; credit_worthiness → {Poor Worthiness, Fair Worthiness, Excellent Worthiness})Furthermore, the calculated posterior approval probability P(Approval) is mapped into explicit operational decision categories:

$$Deceision = \begin{cases} Likely\ Approved, if\ P(Approval) \geq 60\% \\ Borderline - Requires\ Manual\ Human\ Review, if\ 40\% < P(Approval) < 60\% \\ Likely\ Rejected, if\ P(Approval) < 40\% \end{cases}$$

Unlike traditional single-threshold classifiers (P = 0.50) that exhibit high decision entropy near the boundary, this framework adopts a dual-threshold mechanism (P ≥ 0.60 for approval, P<0.40 for rejection) establishing an explicit Uncertainty Buffer Zone (0.40 < P < 0.60):

Automated Processing (P ≥ 0.60 or P < 0.40): High-confidence profiles with decisive probabilistic evidence are processed automatically.

Human-in-the-Loop (HITL) Governance (0.40 <P< 0.60): Borderline cases with conflicting evidence are systematically flagged for manual credit officer review.

This 20-percentage-point uncertainty interval is a design choice intended to flag borderline cases for human review. It does not, by itself, establish reduced financial exposure or compliance with regulatory requirements, including those of the EU AI Act.

Constraint-Guided Prompt Engineering and

The humanized profile JSON and calculated posterior probability metrics are injected into a system prompt bounded by strict deterministic constraints: Professional B2B Tone Directive: Mandates an institutional credit analyst persona. Conversational greetings, introductory fluff, and pleasantries are strictly prohibited, enforcing an objective third-person perspective (referring strictly to "the applicant's profile"). Mathematical Alignment and Anomaly Detection: The LLM evaluates logical agreement between the Bayesian posterior probability and the underlying humanized metrics. If P(Approval) falls within the critical range between 40% and 60% while coinciding with high-risk metrics, the model is instructed to output an explicit anomaly flag. Otherwise, outputs are tagged as Aligned with banking logic. Mitigation of Numerical Hallucinations: The LLM is strictly prohibited from inventing, calculating, or introducing financial figures (e.g., dollar amounts, interest rates, or loan terms). The only numerical figure

CC BY

permitted in the generated explanation is the exact posterior probability calculated by the BN core. Causal Reasoning Alignment: Prompts force Llama 3.2 to trace causal drivers along the network's Directed Acyclic Graph (DAG) topology: moving from primary inputs through intermediate derived pillars (financial_capacity and credit_worthiness) to the final decision state. 4.4 Output Format Specifications In initial reporting mode (evaluate.py), the generated explanation is constrained to a deterministic three-bullet structure: Decision Status: Declares the decision outcome, the exact BN posterior probability percentage (P%), and the alignment/anomaly verification tag. Risk Evaluation: Highlights key negative drivers threatening repayment (limited to a maximum of two sentences) or declares a standardized non-risk statement. Mitigating Strengths: Details offsetting high-performing profile attributes (maximum two sentences) or suggests qualitative credit structuring adjustments. (The complete, raw System Prompt Templates for both evaluate.py and chat.py are provided in Appendix A). Interactive Conversational Engine (chat.py) To enable dynamic post-inference analysis for credit risk officers, the interactive conversational engine (chat.py) establishes a stateful session linked directly to the primary Bayesian Network evaluation output (last_evaluation). To prevent arbitrary or ungrounded queries, the endpoint enforces a strict evaluation dependency, requiring an active applicant profile prior to initializing context-aware interrogation. Upon receiving a query, the conversational pipeline executes a keyword-driven intent classification mechanism to dynamically route the inquiry into target instruction templates while maintaining the base system constraints: Causal Justification Mode: Triggered by causal indicators (e.g., why, reject, cause), this path prompts Llama 3.2 to articulate the specific structural bottlenecks occurring within the intermediate derived pillars (financial_capacity vs. credit_worthiness) that mathematically drive the posterior approval probability. Qualitative Risk Mitigation Mode: Activated by optimization queries (e.g., improve, mitigate, fix), this path structures actionable recommendations into two distinct domains: Internal Variable Shifting: Controllable client-side behavior modifications aimed at improving specific weak intermediate pillars. Loan Structuring Adjustments: Institutional policy adaptations (e.g., requiring collateral or reducing requested capital) to absorb credit exposure. General Inquiry Mode: Directly addresses technical questions regarding profile attributes while strictly maintaining mathematical alignment with the calculated Bayesian outputs. To ensure reproducible and grounded explanations during local execution, runtime inference options are strictly bounded to meet three practical requirements: Zero Temperature (Temperature = 0.0): Minimizes generative randomness to produce consistent, stable explanations for identical profiles and reduce potential hallucinations. Constrained Nucleus Sampling (Top_p = 0.1): Filters out out-of-context or informal words, restricting output selection to relevant, professional banking terminology. Controlled Output Length (Num Predict = 250, Repeat Penalty = 1.2): Limits responses to concise 3–4 sentence summaries, preventing redundant text loops and reducing compute latency on local hardware.

Rule-Based Ground-Truth Baseline for LLM Evaluation

To systematically evaluate the faithfulness, semantic equivalence, and numerical consistency of the explanations generated by Llama 3.2, an automated, deterministic rule-based baseline engine was developed in Python. Operating independently of generative language models (Zero-LLM configuration), this engine serves as a deterministic reference benchmark. The evaluation pipeline constructs this baseline across the N=300 test profiles through a programmatic, two-stage rule-mapping mechanism: Rule-Based Factor Analysis: The baseline script interfaces directly with the compiled Bayesian Network inference core (pgmpy) via VariableElimination. For any given evidence profile E, it executes custom condition-mapping functions that evaluate discrete categorical states (such as credit score tiers, income levels, and loan counts) to programmatically separate inputs into positive contributing factors and negative risk drivers. Deterministic Text Instantiation: The exact posterior approval probability P(Approval | E) and the extracted factors are then programmatically formatted into structured, predefined textual templates. This ensures that the reference text strictly reflects the underlying probabilistic inference without model-driven variation. By establishing this baseline across all N = 300 unique credit scenarios, we created a standardized evaluation benchmark. This enables automated, scalable measurement of the downstream LLM's syntactic precision, contextual semantic alignment (via BERTScore), and adherence to numerical outputs.

Technical Stack and Software Implementation

The proposed framework is implemented using a modular, fully open-source software stack: Dataset Source & Preprocessing: The empirical loan evaluation dataset used for structure learning and CPT parameterization is publicly available . Data standardization, continuous feature discretization,

CC BY

and categorical encoding were implemented using Pandas, NumPy, and scikit-learn (LabelEncoder). Probabilistic Modeling & Inference Core: Structural topology learning and conditional probability estimation were executed via pgmpy using a hybrid approach (HillClimbSearch constrained by domain expertise) and a BayesianEstimator with a BDeu prior. Exact marginal probability queries were processed using the VariableElimination inference algorithm, with trained model artifacts serialized via joblib. Backend API & Data Validation: Backend REST services were developed using FastAPI, enforcing strict request payload validation and data integrity via Pydantic schemas (BaseModel). Local Language Model Execution: Natural language translation and conversational explanations were orchestrated via requests interfacing with a locally hosted Ollama engine running the Llama 3.2 model. Presentation Layer: The user interface was developed using React.js, providing an interactive dashboard for real-time risk reporting, counterfactual scenario exploration, and multi-turn querying

Results

Bayesian Network Performance Evaluation Predictive Performance of the Bayesian Core To ensure methodological consistency across all architectural layers, the empirical dataset was partitioned using a standard 70% training and 30% testing split (N = 1,000 total records; N = 300 held-out test profiles). The Bayesian Network (BN) core was parameterized using the 70% training subset and evaluated on the held-out test set. For each test profile, input variables were processed using standard discretization and encoding procedures, while domain-expert-derived intermediate variables (Financial Capacity and Credit Worthiness) were computed using predefined deterministic rules and incorporated into the inference evidence. The state with the highest posterior probability for the Loan decision node was selected as the predicted outcome and evaluated against ground-truth loan statuses. To contextualize the Bayesian Network's performance against standard statistical benchmarks, two baseline models were evaluated on the same independent test split (N = 300) using the exact same five primary predictor features: a Majority-Class Baseline (which deterministically predicts the majority 'Approved' status) and a standard Logistic Regression model. As detailed in Table 5, the Majority-Class Baseline achieves an accuracy of 73.00% with a 100.00% Recall, directly reflecting the empirical class imbalance of the underlying credit dataset (where 729 out of 1,000 cases, or 72.90%, represent loan approvals). The proposed Bayesian Network core demonstrates a clear predictive improvement over this naive baseline, reaching an overall accuracy of 82.67%, a Precision of 81.04%, a Recall of 99.54%, and an F1-score of 89.34%. While standard Logistic Regression achieves a higher accuracy of 88.33%, it functions purely as a discriminative 'black-box' classifier without explicit conditional dependency representations. The Bayesian Network effectively trades off a marginal difference in classification accuracy for complete structural explainability, transparent Conditional Probability Tables (CPTs), and a DAG topology that enables interactive conversational reasoning.

Rule-Based Ground-Truth Baseline Construction

Evaluating N = 300 test cases through manual human expert annotation introduces subjective bias, inter-annotator variance, and scalability constraints. To establish a rigorous reference benchmark for evaluating the faithfulness of Llama 3.2 to the Bayesian Network (BN) outputs, an automated rule-based explanation engine was implemented. This engine dynamically extracts posterior probability vectors P(Y | E) and conditional probability table (CPT) parent-child dependencies directly from the BN core, inserting these inference outcomes into structured, deterministic text templates. The construction of this rule-based baseline serves four primary analytical functions within the evaluation pipeline: Standardization and Consistency: It enforces uniform evaluation criteria across all 300 test profiles, eliminating human scoring variability. Numerical and Factual Precision: It directly anchors reference text to exact BN probability calculations, ensuring zero gold-standard error. Experimental Reproducibility: It provides a deterministic baseline that guarantees complete replication under identical BN configurations. Feasibility and Scalability: It offers an automated, verifiable benchmark for evaluating LLM text alignment against underlying BN outputs at scale. Operational Note: This rule-based reference engine does not function as an independent manually constructed reference texts (Ground Truth) for evaluating the predictive capability of the BN itself. Instead, it defines the precise information expected to be verbalized by the downstream LLM. Consequently, the resulting evaluation metrics measure the syntactic, semantic, and quantitative faithfulness of the generated explanations relative to the probabilistic BN core.

Automated NLG Evaluation Results

To evaluate the natural language explanations generated by the Llama 3.2 model, a deterministic rule-based explanation engine was constructed as a reference baseline. This baseline maps the probabilistic outputs (P(Y | E)) and conditional probabil-

CC BY

ity dependencies from the Bayesian Network into structured, predefined textual templates. These reference texts were used to assess the lexical overlap, semantic similarity, and selected aspects of numerical and decision-level alignment of the generated explanations across 300 held-out test profiles.

table 5: Performance Benchmarking of the Bayesian Network Core against Baselines (N = 300)				
Model / Benchmark	Accuracy	Precision	Recall	F1-Score
Majority-Class Baseline	73.00%	73.00%	100.00%	84.39%
Proposed Bayesian Network (BN)	82.67%	81.04%	99.54%	89.34%
Logistic Regression	88.33%	90.71%	93.61%	92.13%

Table 7 - Summary of Automatic Evaluation Metrics for Generated Explanations (N = 300)		
Evaluation Category	Metric	Mean Score (%)
Lexical Overlap	BLEU	4.17%
	ROUGE-1	44.99%
	ROUGE-2	16.85%
	ROUGE-L	33.38%
Semantic Similarity	BERTScore Precision	83.34%
	BERTScore Recall	83.89%
	BERTScore F1	83.59%
Explanation Faithfulness	Numerical Faithfulness Rate	91.33%
Decision Alignment	Direct Decision Alignment Accuracy	28.67%

Summary of Observed Results

Lexical Overlap: As shown in Table 7, the BLEU score was 4.17%, while ROUGE-1, ROUGE-2, and ROUGE-L reached 44.99%, 16.85%, and 33.38%, respectively. These metrics quantify different forms of lexical overlap between the generated explanations and the manually constructed reference texts. Differences in wording and explanatory detail may affect these scores. Semantic Similarity: The generated explanations achieved a BERTScore Precision of 83.34%, a Recall of 83.89%, and an F1 score of 83.59%. These results indicate semantic similarity between the generated explanations and the reference texts according to BERTScore. However, because the references contain only the key input values and final decisions rather than complete explanatory narratives, these scores should not be interpreted as direct evidence of factual correctness or complete semantic equivalence. Numerical Faithfulness and Decision Alignment: The Numerical Faithfulness Rate was 91.33%, indicating consistency with the numerical reference criteria under the implemented evaluation procedure. Direct Decision Alignment Accuracy was 28.67%, based on regex-based and exact keyword matching of the decision status. The latter result reflects explicit decision-label matching under the implemented parsing procedure and does not, by itself, establish the overall correctness of the generated explanations. These metrics assess specific aspects of the generated explanations relative to the constructed references. They provide evidence about alignment with the recorded inputs and decisions but do not independently establish the real-world validity of the Bayesian Network's predictions or quantify all forms of hallucination.

LLM-as-a-Judge & Semantic Quality Analysis

To complement automatic text-matching metrics, Qwen 2.5 (7B), running locally via Ollama, was used as an LLM-based judge to evaluate the generated explanations against manually constructed reference texts (Ground Truth). For each test case, these reference texts contained the key input values provided to the Bayesian Network and its corresponding final decision, without additional explanatory content. The evaluation covered 300 test cases and assessed three dimensions: faithfulness, clarity, and completeness, using a five-point rating scale. Qwen 2.5 achieved an overall mean score of 3.11 out of 5.0. The mean scores were 2.98 for faithfulness, 3.92 for clarity, and 2.42 for completeness. Clarity received the highest rating, whereas completeness received the lowest. These results indicate that the generated explanations were rated more favorably in terms of clarity than faithfulness and completeness, suggesting room for improvement in their alignment with the reference information and their coverage of relevant details The judge-model evaluation complements the automatic metrics by providing an additional assessment of the generated explanations relative to the constructed references. However, these scores represent model-based judgments and

should not be interpreted as independent proof of factual correctness, comprehensive hallucination detection, or real-world decision validity.

Table 8 - LLM-Based Evaluation of Generated Explanations Using Qwen 2.5 (7B) (N = 300)

Evaluation Dimension	Mean Score (out of 5)
Faithfulness	2.98
Clarity	3.92
Completeness	2.42
Overall Mean	3.11

Operational Validation: Live Interactive Case Scenarios

To rigorously validate the hybrid Hierarchical Bayesian Network and LLM-based explanation framework, we implemented a multi-faceted evaluation pipeline across a test sample of 300 comprehensive scenarios. The aggregate evaluation provided a quantitative assessment across metrics, including a Bayesian Network Accuracy of 82.67% (Precision: 0.8104, Recall: 0.9954, F1-Score: 0.8934), lexical alignment (ROUGE-1 Average: 0.4499, ROUGE-2 Average: 0.1685, ROUGE-L Average: 0.3338, BLEU Average: 0.0417), and a qualitative evaluation by the Qwen 2.5 judge with an Overall Mean Score of 3.11 / 5.0 (Clarity: 3.92, Faithfulness: 2.98, Completeness: 2.42) at an average latency of 15.75 seconds per case. The aggregate evaluation metrics provide an overview of the predictive performance of the Bayesian Network and the quality of the generated explanations. To complement these quantitative results, three illustrative scenarios are examined, representing high-approval, high-rejection, and borderline probability cases. These examples demonstrate the framework's explanation and interaction capabilities while also highlighting potential limitations in numerical interpretation, structural reasoning, and decision consistency. The interactive responses are presented as qualitative examples and were not included in the automated single-shot explanation evaluation: Low-Risk Profile (Deterministic Approval) Input Evidence Vector (E): Credit Score = 770, Annual Income = $147,905, Existing Loans = 0, Dependents = 1, Employment = Employed. BN Posterior Inference: P(Approval | E) = 99.90% (High Confidence Tier). Generated Narrative Output: Decision Status: Approved | 99.90% | Aligned with banking logic (Strong Profile). Risk Evaluation: The model's approval of a subprime profile despite a good credit score and employed status is questionable. Further review is necessary to understand the decision-making process. Mitigating Strengths: High annual income combined with zero existing debt and stable employment provides maximum financial stability and repayment capacity. Automated Explanation Evaluation: Automated Lexical Metrics: Evaluated against the reference texts, achieving strong alignment (ROUGE-1 = 0.2149, BLEU = 0.0393). LLM-Based Judge Assessment: Qwen 2.5 evaluated the explanation with an overall score of 3.33/5.0, indicating satisfactory alignment with the Bayesian posterior probability and professional readability. Interactive Sub-Query Tracing: The following conversation demonstrates the interactive reasoning of the system when probed with specific exploratory queries: Q1: What are the primary drivers for this high approval probability? "The Final Model Approval Probability of 99.90% is primarily driven by the applicant's profile having a Raw Credit Score Category of "Excellent" and Financial Capacity Assessment of "High Financial Capacity", which are strongly indicative of low credit risk. Additionally, the existing debt burden being categorized as "No Debt Burden" further reduces the likelihood of default. These factors collectively contribute to an extremely high probability of approval." Q2: Given this excellent risk evaluation, what qualitative loan structuring terms or preferential options should the bank consider offering this applicant? "The Final Model Approval Probability of 99.90% indicates a strong confidence in The Risk Profile's creditworthiness and financial capacity. Given these excellent assessments, the bank should consider offering preferential loan terms such as a lower interest rate or extended repayment period to further reduce the risk burden on The Applicant. This decision is grounded in the high probability that The Risk Profile will meet their loan obligations without defaulting (99.90% confidence). Offering flexible loan structures aligns with this assessment, ensuring The Bank's exposure to potential losses remains minimal while still providing a competitive offering to The Applicant" High-Risk Profile (Deterministic Rejection) Input Evidence Vector (E): Credit Score = 480, Annual Income = $24,135, Existing Loans = 2, Dependents = 4, Employment = Employed. BN Posterior Inference: P(Approval | E) = 0.21% (High Confidence Rejection Tier). Generated Narrative Output: Decision Status: Likely Rejected | 0.21% | Aligned with banking logic. Risk Evaluation: The low credit score and relatively low income may indicate a higher risk profile, but the model's rejection aligns with banking logic due to its high baseline probability. This decision is not indicative of a critical anomaly.

Mitigating Strengths: High employment status mitigates the overall risk, while having more dependents reduces the debt-to-income ratio. Automated Explanation Evaluation: Automated Lexical Metrics: Evaluated against the reference texts, achieving consistent alignment (ROUGE-1 = 0.512, BLEU = 0.0256). LLM-Based Judge Assessment: Qwen 2.5 evaluated the explanation with an overall score of 3.0/5.0, reflecting moderate alignment with the Bayesian posterior probability and standard readability. Interactive Sub-Query Tracing: The following conversation demonstrates the interactive reasoning of the system when probed with specific exploratory queries: Q1: What are the primary drivers for this high rejection probability? "The Financial Capacity Assessment and Credit Worthiness Evaluation are interconnected yet distinct components of the risk profile. The applicant's Low Financial Capacity is primarily driven by their 2 Active Loans, which indicate a substantial debt burden. This bottleneck in financial capacity constrains the ability to absorb additional credit or manage existing expenses. Conversely, the Poor Credit Score Category and High Risk associated with the Credit Worthiness Evaluation are largely influenced by the same underlying factors: the applicant's Very Low Income and 2 Active Loans. The stable employment status mitigates some of these concerns; however, it does not entirely offset the risk posed by their debt burden. The combination of these factors creates a compensating trade-off between Financial Capacity Assessment and Credit Worthiness Evaluation, resulting in an overall high rejection probability due to the cumulative impact of debt burden on both assessments." Q2: Given this high-risk evaluation, what specific corrective measures or restructuring steps would this applicant need to undertake to shift their profile from a rejected status to an acceptable risk threshold? "The applicant's Financial Capacity Assessment is constrained by their existing debt burden and family dependents, which limits their ability to take on additional loans. This bottleneck in Financial Capacity Assessment negatively impacts the overall risk profile. In contrast, while Credit Worthiness Evaluation indicates a high-risk category due to poor credit scores, historical network patterns suggest that stable employment with low income can mitigate this risk when combined with other positive factors (e.g., employed status). The 0.21% approval probability reflects the net effect of these competing assessments: the applicant's limited financial capacity outweighs their relatively good employment and family stability. To address the bottleneck, the officer should consider alternative loan options that do not exacerbate debt burden or explore income-enhancing strategies to improve Financial Capacity Assessment before pursuing further credit approvals." Borderline Decision Evaluation & Uncertainty Zone Input Evidence Vector (E): Credit Score = 490, Annual Income = $149,295, Existing Loans = 3, Dependents = 2, Employment = Employed. BN Posterior Inference: P(Approval | E) = 50.00% (Critical Threshold / Manual Review Zone). Generated Narrative Output: Decision Status: Approved (Borderline) | 50.00% | Aligned with banking logic (Risk Correctly Captured). Risk Evaluation: The subprime profile was not flagged despite a low credit score, indicating inadequate risk assessment. Insufficient weight may have been placed on the credit score metric. Mitigating Strengths: High income and stable employment mitigate some risk but are insufficient to offset the adverse credit profile. Automated Explanation Evaluation: Automated Lexical Metrics: Evaluated against the reference texts, achieving strong alignment (ROUGE-1 = 0.3571, BLEU = 0.0114). LLM-Based Judge Assessment: Qwen 2.5 awarded an overall score of 3.0/5.0, reflecting a satisfactory alignment in articulating the nuances of the borderline risk zone and maintaining model fidelity under high uncertainty. Interactive Sub-Query Tracing: To further investigate this borderline evaluation, we initiated the following dialogue: User Inquiry: How can we improve this client's profile to achieve approval? System Response: "The applicant's profile exhibits a contradictory mix of high income and poor credit score, indicating potential for financial capacity but elevated credit risk due to existing debt burden. **INTERNAL VARIABLE SHIFTING** To improve this client's profile, consider the following: Encourage The applicant to prioritize debt repayment by allocating more funds towards loan payments. Suggest a temporary reduction in discretionary spending to reduce the overall debt-to-income ratio. Recommend regular financial planning and budgeting sessions with the applicant. LOAN STRUCTURING** To mitigate risk, consider adjusting bank policies: Offer collateral options that better align with the applicant's credit profile (e.g., secured loans). Implement a more stringent loan approval process for high-risk borrowers to ensure adequate capital buffers. Consider offering alternative repayment terms or interest rate adjustments to reduce the overall cost of borrowing. By implementing these strategies, it is possible to improve the risk profile and in-

crease the likelihood of successful loan outcomes"

Discussion:

1. Interpretation of Bayesian Network Predictive Performance

The empirical evaluation of the Bayesian Network (BN) core achieved a classification accuracy of 82.67% and an F1-score of 89.34% on the held-out test subset (N = 300). These results validate the capacity of the probabilistic graph to capture underlying decision patterns embedded within the historical credit dataset. Key technical and methodological insights regarding these results include:

1. Class Distribution and Efficacy of Structural Feature Decomposition: The introduction of intermediate variable structures—specifically Financial Capacity and Credit Worthiness—was designed using analytical domain heuristics to structure raw input features into discrete probabilistic dependencies. Evaluating the test partition (N = 300) yields an overall Bayesian Network accuracy of 82.67%, with Precision at 81.04%, Recall at 99.54%, and an F1-score of 89.34%. To investigate the elevated Recall score, a class distribution analysis of the baseline credit dataset (N = 1,000) was conducted. The empirical dataset exhibits an inherent class imbalance, comprising 729 Approved cases (72.90%) against 271 Rejected cases (27.10%). Consequently, the Bayesian Network learned this empirical prior distribution, resulting in a prediction boundary aligned with the majority class prior rather than an artificially balanced decision threshold. This confirms that the classification behavior directly reflects the underlying dataset distribution while maintaining exact probabilistic inference tractability.

2. Reflecting Empirical Data Patterns and Boundaries: The remaining ~18% classification variance highlights two operational factors: boundary discretization effects in continuous variables, and inherent inconsistencies or noise within historical human credit decisions. Because the BN reflects the empirical probability distributions of the underlying dataset, its predictive output mirrors real-world decision patterns rather than idealized assumptions.

3. Validation Alignment and Scope Considerations: The current performance was evaluated using an independent 30% held-out test split, ensuring strict separation between parameter learning and performance estimation. While evaluating the model against external multi-institutional datasets or conducting blind human-expert panel scoring remains an objective for future work, the current held-out evaluation provides a statistically valid and rigorous benchmark for downstream explanation generation.

2. Critical Analysis of Natural Language Generation Performance

1. Lexical vs. Semantic Divergence: The low BLEU and ROUGE-2 scores (4.17% and 16.85%, respectively), alongside a BERTScore F1 of 83.59%, reflect differences in surface-level wording alongside semantic similarity between the generated explanations and the template-based reference texts. These metrics capture complementary aspects of textual similarity and should not be interpreted as direct evidence of factual correctness or complete semantic equivalence.

2. Mitigation of Numerical Hallucinations: The high Numerical Faithfulness Score (91.33%) indicates that the generated explanations were numerically consistent with the underlying BN probabilities in 91.33% of cases under the implemented evaluation procedure. This result suggests that the proposed integration strategy can support numerical consistency between generated narratives and BN outputs. However, it does not independently establish that structured evidence injection and constraint-guided prompt engineering reduce numerical hallucinations, as this would require a separate evaluation.

3. Methodological Constraints of Programmatic Decision Matching: The Direct Decision-Matching Rate of 28.67% reflects the operational limitations of strict regex-based parsing rather than a systemic failure of LLM semantic decision alignment. Because programmatic matching required exact keyword matches (e.g., "Approved"), descriptive or alternative domain formulations -such as "The applicant displays a highly favorable profile for credit extension"- were categorized as non-matches despite conveying the correct decision context. This limitation underscores the necessity of incorporating LLM-as-a-Judge semantic evaluation to assess decision-support quality beyond lexical matching.

3. Methodological Analysis of Judge Performances

1. Assessment of Explanation Quality: The Qwen 2.5 evaluation provided a complementary assessment of the generated explanations beyond automatic text-matching metrics. The mean clarity score of 3.92 out of 5.0 was the highest among the evaluated dimensions, followed by faithfulness at 2.98 and completeness at 2.42.

2. Interpretation of Judge Scores: The overall mean score of 3.11 out of 5.0 indicates a moderate evaluation outcome, with clarity rated more favorably than faithfulness and completeness. These results suggest that, although the explanations were generally considered relatively clear, their alignment with the reference information and their coverage of relevant details require further improvement.

3. Methodological Limitations: The Qwen-based assessment provides an additional model-based eval-

CC BY

uation of the generated explanations against the manually constructed reference texts. However, these scores should be interpreted as supplementary judgments rather than independent verification of factual correctness or comprehensive detection of hallucinations. The evaluation is also limited by the scope and content of the constructed references.

4. Qualitative Insights from Live Case Scenarios and Interactive Risk-Mitigation Recommendations

The three representative case scenarios provide qualitative insights into the explanatory capabilities and limitations of the proposed framework across low-risk, high-risk, and borderline credit profiles. The analysis distinguishes the probabilistic predictions produced by the Bayesian Network (BN) from the natural-language explanations generated by Llama 3.2, with particular attention to structural consistency, probabilistic interpretation, and the reliability of interactive recommendations.

4.1. Evaluation Scope and Single-Shot Explanation Performance

The quantitative Natural Language Generation (NLG) metrics (ROUGE and BLEU) and the Qwen 2.5-based evaluation were applied exclusively to the initial single-shot explanations, consisting of three components: Decision Status, Risk Evaluation, and Mitigating Strengths. These explanations were compared against deterministic rule-based reference templates generated from the applicant attributes and Bayesian Network outputs. The results indicate that linguistic coherence does not necessarily ensure accurate or complete representation of the underlying probabilistic model. Accordingly, the following case analysis examines whether the generated explanations preserve the meaning of the predicted outcome, correctly attribute input features to the corresponding intermediate assessments, and distinguish model-supported evidence from plausible but unverified financial interpretations.

4.2 Case-Based Analysis of Explanatory Limitations

Low-Risk Profile: Overinterpretation of the Predicted Outcome

In the low-risk scenario, the interactive explanation associated the 99.90% approval probability with favorable financial capacity and creditworthiness indicators. However, it also interpreted this probability as indicating an equivalent likelihood of fulfilling loan obligations without default. This interpretation exceeds the scope of the BN, which predicts loan approval rather than repayment or default. The explanation also described the applicant as having "maximum financial stability and repayment capacity." Although the favorable indicators support a strong approval assessment within the implemented model, this absolute wording is not established by the reported probability alone. These observations illustrate two distinct limitations: extending the predicted outcome to an unmodeled financial event and overstating the strength of the assessment through imprecise language. Neither observation, by itself, establishes an error in the BN's numerical inference.

High-Risk Profile: Structural Attribution and Unsupported Financial Claims

In the high-risk scenario, the initial explanation described the financial burden associated with having multiple dependents as a reduction in the debt-to-income ratio. However, the implemented framework incorporates the number of dependents into a rule-based Financial Capacity Assessment through penalties applied to higher dependent-burden categories; it does not directly calculate the conventional debt-to-income ratio. The explanation therefore introduced a financial interpretation that was not supported by the implemented mechanism. The interactive responses revealed additional concerns. When asked about the primary drivers of rejection, the conversational model associated existing loans with both Financial Capacity Assessment and Credit Worthiness Evaluation, although existing_loans_count is structurally assigned to Credit Worthiness Evaluation in the implemented BN. The response also referred to "historical network patterns" to explain the mitigating effect of stable employment under low-income conditions, without presenting explicit conditional probability evidence supporting that claim. These observations concern the fidelity of the generated explanations rather than necessarily the correctness of the final approval probability. They demonstrate that a plausible financial narrative can misrepresent the network's structure or present an unsupported relationship as an established model finding.

Borderline Profile: Interpretation of Conflicting Indicators and Human Review

In the borderline scenario, the BN assigned a 50.00% approval probability and reported an Approved classification. We clarify that exact probability values of 50.00% in certain Conditional Probability Table (CPT) entries arise from the Dirichlet uniform smoothing prior applied during parameter learning to handle sparsely populated or unobserved parent configurations; thus, this outcome reflects data sparsity and structural neutrality rather than a true conflict of empirical evidence. The initial narrative questioned whether the applicant's low credit score had received sufficient consideration, while the interactive response acknowledged the combination of high income, poor creditworthiness indicators, and existing debt burden.

CC BY

The interactive response proposed debt repayment reduced discretionary spending, financial planning, and alternative loan structures. However, it did not recalculate the posterior approval probability following any proposed change to the applicant's attributes. These suggestions should therefore be treated as qualitative risk-mitigation recommendations rather than validated counterfactual explanations. The framework also applies a separate human-review advisory to cases with approval probabilities between 40% and 60%, inclusive. Under this policy, the 50.00% case retains the BN's Approved classification while receiving an additional recommendation for review by a credit officer. The advisory does not alter the network's prediction or constitute a new classification; it is a supplementary decision-support mechanism intended to draw attention to cases requiring further assessment. When examining the raw narrative outputs generated across the representative case scenarios, certain conceptual or phrasing discrepancies become apparent. For instance, in Case 1, despite an excellent credit score of 770, the model's generated narrative unsuitably utilized the term 'subprime' while simultaneously classifying the profile as aligned with banking logic. Similarly, minor semantic or structural variations appear in subsequent scenarios. To mitigate such issues, the prompt engineering protocol was carefully structured to enforce strict operational constraints, limit numerical hallucination, and guide the language model along the Bayesian network's causal pathways. However, despite these structured prompting safeguards, lightweight locally hosted models inherently exhibit limitations when handling multi-constraint reasoning tasks. The purpose of integrating this local setup was to test the practical feasibility of the decoupled hybrid architecture within an offline, resource-accessible environment. Consequently, the minor variations, inappropriate terminology (such as subprime for high-scoring profiles), or subtle semantic discrepancies occasionally found in the raw outputs are an expected characteristic of models with limited parameter sizes when navigating complex inference constraints. It is important to emphasize that this prototype was intentionally designed as an exploratory proof-of-concept. While these raw outputs successfully demonstrate the functional integration of probabilistic inference and natural-language generation, they also highlight the boundaries of edge-deployed language models. Transitioning this framework into an operational, production-grade financial decision-support system would certainly necessitate scaling up to advanced, high-capacity language models to guarantee absolute narrative fidelity and eliminate such explanatory discrepancies.

4.3 Architectural Constraints and Feature-Space Limitations

The framework's explanatory capabilities are constrained by the variables and probabilistic relationships represented in the BN and supplied to the language model through the structured prompt context. To limit Conditional Probability Table (CPT) dimensionality and mitigate data sparsity, the implemented BN retains five primary input variables and two domain-engineered intermediate nodes. Although this compact structure reduces model complexity, it also limits the information available for probabilistic inference and subsequent explanation generation. For example, the exclusion of loan_amount prevents the framework from explicitly incorporating the requested loan size into its assessment. An applicant may therefore receive a high approval probability despite a relatively low credit score, while the framework remains unable to account for the potential effect of the requested amount. Such a result may reflect a limitation of the available feature space rather than an error in the network's inference, and should be interpreted in the context of the variables represented in the model. Overall, the case scenarios demonstrate that evaluating conversational Bayesian explanations requires more than verifying the reported approval probability. The explanations must also preserve the meaning of the predicted target, correctly represent the network's structural dependencies, avoid unsupported financial claims, and distinguish qualitative recommendations from changes to the model's calculated outcome. The current framework should therefore be regarded as a proof of concept for conversational explanations of Bayesian predictions rather than a comprehensive credit-risk assessment system. Future work may incorporate additional financial variables, more explicit extraction of probabilistic evidence, and systematic evaluation of interactive responses to improve explanatory reliability.

5. Literature Benchmarking & Comparative Position

To position this study within the broader Explainable AI (XAI) literature, the proposed integration of Llama 3.2 with a Bayesian Network (BN) is discussed in relation to established approaches to Bayesian explanation, natural language generation (NLG), and post-hoc feature attribution.

Positioning Against BN Explanation Baselines

Previous research has investigated methods for explaining Bayesian network inference through natural language generation and structured explanatory representations [17, 36]. Such approaches provide mechanisms for translating probabilistic reasoning

CC BY

into human-readable explanations. The proposed framework adopts a complementary architecture by combining a BN-based probabilistic inference core with an LLM-based language layer. This design supports flexible, applicant-specific explanations and conversational interaction, while retaining the BN as the source of the underlying predictions. Unlike predefined templates, an LLM-based language layer offers greater flexibility in phrasing and interaction; however, the factual accuracy of the generated explanations depends on how faithfully the language model represents the BN's structure and outputs. Although prior studies have addressed natural-language explanations of Bayesian network inference, the proposed framework differs in its application domain, language-generation architecture, and interaction setting. Keppens [17] presents an NLG approach based on qualified support graphs for generating explanations of Bayesian inference, whereas Sevilla et al. [36] investigate natural-language explanations based on factor arguments in the medical domain. In contrast, this study applies Bayesian inference to credit approval assessment and uses Llama 3.2 as a conversational language layer to explain applicant-specific BN outputs and respond to follow-up questions. The framework also examines the consistency of generated explanations with the underlying network structure and inference results. Thus, its contribution lies not in introducing natural-language explanation of Bayesian networks itself, but in integrating BN-based credit assessment with LLM-generated conversational explanations and evaluating the resulting explanations in credit-related scenarios. Alignment with LLM-Based XAI Frameworks Recent research has explored the use of large language models to generate natural-language explanations of structured model outputs and AI decisions [9, 26]. In line with this research direction, the proposed framework supplies Llama 3.2 with applicant information, BN inference results, and explicit instructions governing how the network's structure should be interpreted. These constraints are intended to reduce unsupported interpretations and improve consistency between the generated explanations and the underlying model. Nevertheless, the qualitative analysis of case scenarios highlights the importance of evaluating generated explanations against the model's actual structure and outputs, since the use of explicit prompting constraints alone does not guarantee complete faithfulness. Probabilistic Reasoning and Post-Hoc Feature Attribution (SHAP/LIME) The proposed framework uses a Bayesian Network as its probabilistic reasoning core rather than relying exclusively on post-hoc feature attribution methods such as SHAP (SHapley Additive exPlanations) [27] or LIME (Local Interpretable Model-agnostic Explanations) [32]. These methods provide ways to explain model predictions through feature contributions or locally interpretable approximations, whereas a BN explicitly represents conditional probability distributions and dependencies among variables. The approaches therefore provide different forms of explanatory information and should not be considered interchangeable. In the proposed framework, the BN represents the probabilistic relationships between applicant attributes, intermediate assessments, and the final loan approval outcome. Llama 3.2 uses the supplied model information to produce natural-language explanations of the BN's predictions. This architecture is intended to provide explanations informed by the network's probabilistic structure rather than relying solely on feature-attribution scores. However, a BN's representation of conditional dependencies does not, by itself, establish causal relationships, and the use of an LLM does not automatically guarantee that these dependencies will be accurately conveyed in natural language. Accordingly, the proposed framework is best characterized as a conversational approach to explaining Bayesian predictions, rather than as a demonstrated method of causal explanation. Ethical Considerations and Regulatory Relevance The use of conversational decision-support systems in financial applications raises ethical and regulatory considerations, particularly regarding transparency, human oversight, and the interpretation of model outputs. The proposed framework is designed as an explanatory interface that communicates Bayesian Network (BN) predictions in natural language, rather than as an autonomous decision-maker or a bias-mitigation mechanism. From an ethical perspective, the use of socioeconomic attributes, such as the number of dependents, may raise fairness concerns if these attributes contribute to systematically unfavorable outcomes for particular groups. The language model provides explanations of the BN's outputs but does not independently assess or mitigate potential biases in the underlying model. Accordingly, algorithmic fairness and potential disparities across demographic subgroups remain outside the scope of this study and require dedicated evaluation. From a regulatory perspective, the framework addresses aspects relevant to transparency and human oversight in automated decision-support systems. By translating BN outputs into natural-language explanations and providing a human-review advisory for cases with intermediate approval probabilities, the framework aims to support credit officers in inter-

CC BY

preting model predictions. These capabilities may be relevant to discussions of transparency and human oversight under regulatory frameworks such as the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act (EU AI Act). However, the framework has not been subjected to a formal regulatory compliance assessment, and its implementation alone does not establish compliance with these requirements.

Table 9 - Qualitative Comparison of the Proposed Framework with Representative XAI and BN Explanation Approaches

Baseline / Methodology	Representative Literature	Primary Mechanism	Distinguishing Capability of the Proposed Framework
BN Natural Language Generation	Keppens (2019) [١٧]	Generation of natural-language explanations using a structured model of Bayesian inference and support graphs	Combines BN-based probabilistic inference with an LLM-based language layer to support flexible, context-sensitive explanations and interactive dialogue.
User-Evaluated BN Explanations	Sevilla et al. (2024) [٣٦]	Natural-language explanations of BN reasoning evaluated in a user study	Extends the explanation setting to conversational interaction and applicant-specific credit assessments.
LLM-Based Explanation of Credit-Risk Predictions	Dyomin et al. (2026) [٩]	Conversion of structured machine-learning feature contributions into natural-language explanations	Uses a BN as the probabilistic inference core and supplies its outputs and structural information to an LLM for explanation generation.
Post-hoc Feature Attribution (SHAP/LIME)	Bücker et al. (2020) [٤٧]	Feature-attribution methods and other approaches to model transparency and explainability in credit scoring	Explicitly represents conditional probabilistic dependencies in the BN rather than relying exclusively on post-hoc feature-attribution outputs.

Conclusions and recommendations

Concluding Remarks

This study presented an initial proof of concept for generating natural-language explanations of Bayesian Network (BN) inferences using a locally deployed Large Language Model (Llama 3.2) in a credit decision-support context. The proposed implementation separates probabilistic inference performed by the BN from natural-language generation performed by the LLM, which receives structured applicant information and BN outputs to produce explanations of the resulting predictions. Evaluation of the BN on an independent test set yielded a classification accuracy of 82.67%, with a precision of 0.8104, recall of 0.9954, and F1-score of 0.8934. These results provide a quantitative baseline for the implemented credit-classification model. However, the observed performance should be interpreted in light of the dataset, class distribution, and evaluation setup, rather than as evidence of general suitability for real-world credit decisions. For natural-language generation, the system achieved a Numerical Faithfulness Rate of 91.33% and a BERTScore F1 of 83.59%. An additional evaluation using Qwen 2.5 as a judge model produced a mean overall score of 3.11 out of 5.0, with scores of 2.98 for faithfulness, 3.92 for clarity, and 2.42 for completeness. These results indicate that the generated explanations were generally clearer than they were complete or consistently faithful, highlighting the distinction between linguistic readability and the accurate communication of model outputs. The separation of BN inference from language generation is a central design characteristic of the implementation: the LLM is used to communicate the results of probabilistic inference rather than to calculate the underlying probabilities independently. Nevertheless, this architectural separation does not guarantee that every generated explanation accurately preserves the numerical results or the relationships represented in the network. The observed evaluation results and qualitative limitations therefore support treating the system as an exploratory decision-support prototype rather than a val-

idated solution for operational credit assessment.

Acknowledged Limitations

Several technical and methodological limitations should be considered when interpreting the findings.

1. Dependence on a Simplified BN Topology: The implemented BN contains five applicant input variables and two intermediate assessment nodes. Although this structure provides a manageable setting for investigating the integration of probabilistic inference and natural-language generation, it represents only a simplified model of credit assessment. More complex financial risk settings may involve additional variables, dependencies, and sources of uncertainty that are not represented in the current topology.

2. Evaluation Focused on Consistency with BN Outputs: The reference explanations were constructed programmatically using the BN's probability outputs. Consequently, the evaluation measures how closely the generated explanations correspond to the model-derived references; it does not independently establish the real-world validity of the BN's predictions, the appropriateness of its assumptions, or the fairness of the resulting decisions. High consistency with a reference derived from the BN should therefore not be interpreted as evidence that the underlying credit assessment is objectively correct.

3. Limitations of Programmatic Decision Matching: The direct decision-matching rate was 28.67% under the exact parsing procedure used in the evaluation. This result indicates that the implemented matching procedure did not reliably identify the intended decision in many generated responses. Because exact or regex-based matching can be sensitive to phrasing and paraphrasing, the result may reflect limitations in both the generated responses and the evaluation procedure. Further analysis is required to distinguish semantic decision errors from parsing failures.

4. Residual Risk of Unsupported or Misleading Explanations: Despite a Numerical Faithfulness Rate of 91.33%, some generated explanations may contain numerical inconsistencies, unsupported interpretations, or misleading descriptions of relationships among variables. Prompt constraints reduce the scope for unsupported interpretations but do not eliminate these risks. Moreover, if the underlying BN reflects limitations or biases in its data, assumptions, or structure, a fluent natural-language explanation does not resolve those limitations and may communicate the resulting prediction persuasively without adequately conveying its uncertainty. Human review remains important, particularly in high-stakes credit-assessment settings.

Interpretive Reliability of the Language Model

The qualitative case analysis illustrates both the potential and the limitations of LLM-generated explanations. In borderline cases, the language model can describe competing considerations in an applicant's profile, such as a low credit score alongside a high annual income. However, a plausible narrative does not necessarily demonstrate that the model has correctly represented the BN's probability distributions or the roles of individual variables. For example, descriptions of financial burden or the relative importance of credit-related factors may introduce interpretations that are not directly supported by the specified network structure or its outputs. These observations underscore the need to distinguish linguistic plausibility from model-grounded accuracy. The current implementation should therefore be understood as a means of making BN outputs more accessible through conversational interaction, not as an independently validated authority on credit risk. Further refinement and evaluation with domain experts are needed before the approach can be considered for operational use in high-stakes decision-making.

Recommendations for Future Research

Future work should address the limitations identified in this study through the following directions:

1. Expanding BN Complexity: Evaluate the framework using larger and more complex Bayesian Networks with additional variables, intermediate nodes, and higher-dimensional Conditional Probability Tables (CPTs). This would help assess whether the prompting approach remains effective as the amount and complexity of the information supplied to the LLM increase.

2. Improving Semantic Decision Evaluation: Replace or supplement exact regex-based decision matching with a more robust evaluation procedure that distinguishes genuine decision mismatches from differences in wording. The revised procedure should be validated against manually reviewed examples to establish whether generated responses communicate the intended BN decision accurately.

3. Strengthening Explanation Faithfulness and Completeness: Refine the prompting protocol and evaluation criteria to identify unsupported claims, incorrect descriptions of variable relationships, omissions of relevant evidence, and confusion between approval probability and other outcomes, such as repayment or default. Numerical consistency should be evaluated separately from structural consistency and overall explanatory quality.

4. Human-in-the-Loop Validation: Conduct structured evaluations with credit-risk professionals to assess the factual accuracy, completeness, interpretability, and practical usefulness of the generated explanations. Such studies could investigate whether conversational explanations help users understand

model outputs more effectively than conventional numerical summaries, without assuming in advance that they improve decision quality or user trust.

5. Assessing Robustness and Fairness: Investigate how changes in applicant characteristics, BN assumptions, and the distribution of evaluation data affect both the underlying predictions and their natural-language explanations. Dedicated fairness analyses would also be required to determine whether the credit model produces systematically different outcomes across relevant applicant groups.

Overall, this study provides an initial implementation and empirical evaluation of a conversational interface for explaining Bayesian Network outputs in credit assessment. Its findings demonstrate the feasibility of separating probabilistic inference from natural-language generation while also revealing limitations in numerical consistency, completeness, and automated decision evaluation. Addressing these limitations through more comprehensive testing, improved evaluation methods, and expert validation is an important next step toward establishing the reliability and practical value of this approach.

References

1. Absar, N. et al. (2021). Explaining Bayesian networks in natural language: A review. Journal of Artificial Intelligence Research, 71, 1101–1145.
2. Barber, D. (2011). Bayesian reasoning and machine learning. Cambridge University Press.
3. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137–1155.
4. Berger, J. O. (1985). Statistical decision theory and Bayesian analysis. Springer.
5. Zytek, A., Pidò, S., & Veeramachaneni, K. (2024). LLMs for XAI: Future directions for explaining explanations. arXiv preprint arXiv:2405.06064.
6. Bollen, K. A. (1989). Structural equations with latent variables. Wiley.
7. Darwiche, A. (2009). Modeling and reasoning with Bayesian networks. Cambridge University Press.
8. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
9. Dyomin, A., & Li, M. (2026). Interpretable decision support for credit risk assessment using machine learning and LLM-based explanations. Proceedings of the International Conference on Advanced Credit Risk Modeling and Explainable AI, 115–128.
10. Fenton, N., & Neil, M. (2018). Risk assessment and decision analysis with Bayesian networks. CRC Press.
11. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50–57.
12. Gunning, D. et al. (2019). DARPA's Explainable Artificial Intelligence (XAI) program. AI Magazine, 40(2), 44–58.
13. Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field. Data, 8(7), 112.
14. Heckerman, D., Mamdani, A., & Wellman, M. P. (1995). Real-world applications of Bayesian networks. Communications of the ACM, 38(3), 24–26.
15. Jacovi, A. et al. (2021). Formalizing trust in artificial intelligence: Prerequisites, definitions and measurement scales. In Proc. ACM FAccT (pp. 639–649).
16. Ji, Z. et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.
17. Keppens, J. (2019). Explainable Bayesian network query results via natural language generation systems. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19, pp. 42–51).
18. Koller, D., & Friedman, N. (2009). Probabilistic graphical models: Principles and techniques. MIT Press.
19. Krishna, S. et al. (2023). The disagreement problem in explainable machine learning: An empirical study. Neurocomputing, 558, 126720.
20. Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian networks. The Knowledge Engineering Review, 17(2), 107–127.
21. Lakkaraju, H. et al. (2022). Rethinking explainability: Towards a human-centered approach for explainable AI. Communications of the ACM, 65(1), 24–32.
22. Lin, P. (2015). Why ethics matters for autonomous cars. In Autonomous Driving (pp. 69–85). Springer.
23. Lipton, Z. C. (2018). The mythos of model interpretability. ACM Queue, 16(3), 31–57.
24. Liu, P. et al. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35.
25. Liu, Y. et al. (2023). Trustworthy LLMs: A survey on large language models' reliability and explainability. arXiv preprint arXiv:2308.05374.
26. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2024). Language models often offer plausible but misleading reasons for their predictions. Advances in Neural Information Processing Systems, 36, 42100–42115.
27. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
28. Metzger, A., Bartel, J., & Laufer, J. (2023). An AI chatbot for explaining deep reinforcement learning decisions of service-oriented systems. In International Conference on Service-Oriented Computing (pp. 115–130). Springer.
29. Metzger, M. et al. (2023). Interactive chatbots for explaining complex AI decisions. Journal of Artificial Intelligence Research, 78, 411–445.
30. Nauta, M. et al. (2023). From qualities to criteria: A survey on evaluation metrics for explainable AI. IEEE Transactions on

Pattern Analysis and Machine Intelligence, 45(7), 8848–8869.
31. Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann.
32. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proc. ACM SIGKDD (pp. 1135–1144).
33. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
34. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). Explainable AI: Interpreting, explaining and visualizing deep learning. Springer.
35. Scutari, M., & Denis, J.-B. (2021). Bayesian networks: With examples in R. CRC Press.
36. Sevilla, J. et al. (2024). Explaining Bayesian networks reasoning to the general public. CEUR Workshop Proceedings, 4061, Paper 3.
37. Slack, D. et al. (2023). Explaining machine learning models with interactive natural language conversations. Nature Machine Intelligence, 5(8), 892–901.
38. Timmer, S. T. et al. (2016). A two-phase method for extracting explanatory arguments from Bayesian networks. International Journal of Approximate Reasoning, 80, 475–494.
39. Touvron, H. et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
40. Vlek, C. S., Prakken, H., Renooij, S., & Verheij, B. (2016). A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law, 24(3), 263–290.
41. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
42. Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.
43. White, J. et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382.
44. Zhang, Y. et al. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.
45. Zhao, H. et al. (2024). Explainable AI meets LLMs: A survey on explainable artificial intelligence in the era of large language models. arXiv preprint arXiv:2402.04616.
46. Zhao, W. X. et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
47. Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2020). Transparency, auditability and explainability of machine learning models in credit scoring. arXiv preprint arXiv:2009.13384.
48. Calabrese, R., & Martin-Barragan, B. (2023). Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 312(2), 755–769

Author contributions
Conceptualization: Lamis Olaiky
Methodology: Lamis Olaiky
Software / Pipeline Implementation: Lamis Olaiky
Validation & Investigation: Lamis Olaiky
Formal Analysis: Lamis Olaiky
Writing – Original Draft: Lamis Olaiky
Competing interests: The author declares that there are no competing interests
Data and materials availability: All data and materials used in the analysis are available within the main text or the supplementary materials

Abbreviations
BN: Bayesian Network
CPT: Conditional Probability Table
DAG: Directed Acyclic Graph
LLM: Large Language Model
XAI: Explainable Artificial Intelligence
SHAP: SHapley Additive exPlanations
LIME: Local Interpretable Model-agnostic Explanations
API: Application Programming Interface