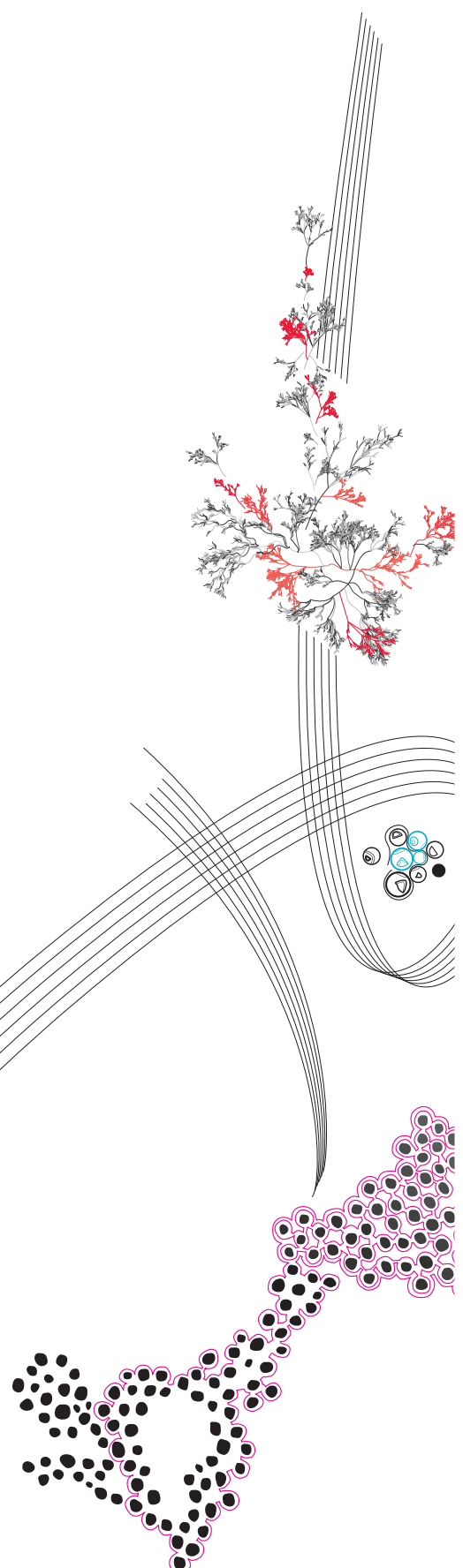




MSc Business & IT
Data Science and Business

Explainable Hybrid Deal Ranking for B2B Sales

Aadil Sheriff

Supervisor (BMS): Dr. L.O. Meertens

Supervisor (EEMCS): Dr. L. Ferreira Pires

August, 2026

Faculty of Electrical Engineering,
Mathematics and Computer Science
Faculty of Behavioural, Management
and Social Sciences
University of Twente

Management Summary

In business-to-business (B2B) software sales, deals advance through conversations. Yet lead-scoring tools rely only on structured customer relationship management (CRM) records. They overlook the introductory sales call, which might hold the most telling information about a deal. Whether taking into account or analysing that call improves prediction of deal conversions has hardly been explored. This thesis designs and evaluates an explainable decision aid that combines CRM data with information drawn from introductory-call transcripts to assess whether the conversation helps predict which deals close. The decision aid uses a large language model (LLM) to rate each introductory call against a rubric grounded in sales-conversation theory. These signals are combined with CRM data to rank deals and explain each score. It is evaluated on real data from a Remote IT Asset Management (RITAM) software vendor.

The introductory-call signals did not reliably improve prediction over CRM data alone; once the deal’s value was accounted for, the conversation added no further predictive value. The main contribution is a rigorous and reusable method for building, testing, and explaining these decision systems. The finding is a negative result, established on one vendor’s own data. Based on these results, deals are currently best ranked using CRM information. Call-based scoring should wait until stronger supporting evidence exists.

Keywords: lead scoring; B2B sales; large language models; explainable AI; SHAP; design science research.

Contents

1	Introduction	1
1.1	Context and motivation	1
1.2	Problem statement	2
1.3	Research design	2
1.4	Approach and methodology	4
1.5	Structure of the thesis	6
2	Literature Review	7
2.1	Machine learning for lead scoring	8
2.2	NLP and LLM feature extraction	8
2.3	Explainable AI in B2B decision-making	9
2.4	Theoretical foundations	9
3	Business Context	15
3.1	The sales process	15
3.2	Process data	16
3.3	Current situation	17
3.4	Problem Formalisation	17
4	Model Development	19
4.1	Overview	19
4.2	Data collection and population (Phase 1)	19
4.3	Data preparation and the instrument (Phase 2)	21
4.4	Model engineering and analysis (Phases 3-4)	26
4.5	Deployment and monitoring (Phases 5 and 6)	29
5	Evaluation	34
5.1	Predictive contribution of the transcript features	34
5.2	Transcript features as measurements	39
5.3	Evaluation summary	41
5.4	Discussion	41
5.5	Limitations	44
6	Conclusions	45
6.1	Summary of findings	45
6.2	Contributions	46

6.3	Recommendations for practice	47
6.4	Future Work	48
Statement on the Use of Artificial Intelligence		57
A	Feature Dictionary	58
A.1	Features used in the model	58
A.2	Features excluded by design	59
B	Full Extraction Rubric	61
B.1	Ordinal dimensions (0-5)	61
B.2	Qualification booleans (BANT)	62
B.3	Behaviour counts	62
B.4	Edge cases	62
C	Complete Extraction Prompt	63
C.1	System prompt	64
C.2	User-message template	66
C.3	Enforced output schema	67
D	User-Study Template	68
D.1	Artefact usage feedback form (TAM)	68
D.2	Human inter-rater (ICC) study: sample rating form	69
D.3	End-user evaluation of the explanations: sample questionnaire	69

List of Figures

1.1	The design-science research framework for this thesis [35]. Relevance flows from the RITAM sales environment into the build-evaluate cycle; rigour flows from the knowledge base; the artefact returns to the environment as the lead card and to the knowledge base as a validated call-scoring instrument.	4
1.2	The six DSR methodology activities [66].	5
2.1	Why PR-AUC is preferred over ROC-AUC under heavy class imbalance: ROC rewards abundant true negatives, whereas the precision–recall curve is read against the no-skill baseline set by the positive rate [22, 78]. . .	10
2.2	Reliability versus validity for a measurement instrument. Reliability is consistency (a tight grouping); validity is agreement with the truth (a grouping on the centre). An instrument can be reliable yet invalid, so the two must be evidenced separately.	12
2.3	The SPIN model of sales discovery [70].	13
2.4	Schematic of a local SHAP explanation: each feature pushes the prediction away from the base value, and the signed contributions sum exactly to the score for that deal.	14
3.1	The host organisation’s sales pipeline. The decision point (the end of the introductory call).	15
4.1	The lead-scoring pipeline end to end. The transcript path (left) is the novel feature layer; the combined features feed a single gradient-boosted tree whose score is paired with a SHAP explanation.	20
4.2	Deal value for won versus lost deals. The near-identical group means motivated the extraction of conversational features.	21
4.3	Mapping from theoretical sources to instrument outputs.	23
4.4	Feature-engineering funnel: the fourteen raw instrument outputs are reduced to eight engineered features by a blind, outcome-independent variance filter and theory-driven composition, fixed before the final data were seen.	25
4.5	The account-executive view: the ranked post-discovery queue (left) and the lead card for the selected deal (right), with checkable reasons, the discovery summary, and the feedback and disagree controls.	31

4.6	The business-development handoff-quality card: BANT qualification coverage, conversation-quality bars, and buyer signals for the booked meeting, scored from the sourced call transcript.	32
4.7	The sales-leadership view: priority-tier mix and active-selling workload by region (top), and aggregate deal value at stake per priority tier (bottom).	33
5.1	The three escalating tests applied to the observed gain, and the verdict they lead to.	36
5.2	Global feature importance (mean SHAP) for the deployed Model B. Deal value dominates; the transcript features sit below the CRM features.	37
5.3	Left: how much the model leans on each feature (mean SHAP). Right: how much each feature moves held-out predictions (permutation Δ PR-AUC).The transcript features carry visible attribution but little predictive effect, pain depth excepted (+0.058).	39
5.4	The measurement-validity progression: the three checks this thesis establishes on the instrument's own outputs, and the criterion-validity step that requires human raters and is left to future work.	41

List of Tables

1.1	CRISP-ML(Q) quality gates for this study: the principal risk at each phase and the check applied.	6
2.1	SLR Overview	7
2.2	Theoretical Foundations	9
3.1	Stakeholders in the lead-scoring decision and what the score does for each.	16
3.2	The data a deal carries in the CRM and call-recording platform. Every group but the last is a structured field; the introductory-call transcript is the only unstructured source, and no structured field captures its content.	16
3.3	The scoring signals already present in the CRM today.	17
3.4	Machine Learning Canvas [68] for the lead-scoring artefact.	18
4.1	Construction of the modelling population. Filters applied in order. . . .	19
4.2	Regional composition of the modelling population.	21
4.3	Leakage screen: fill rate by outcome for the excluded counters. A large gap means the field’s presence encodes the outcome.	22
4.4	The seven ordinal dimensions and their distribution across the 446 analysed calls (integer 0–5; sentiment uses 1–5 for graded affect, with 0 only when affect is indeterminable). Qualification prevalence: budget addressed in 72.2% of calls, authority 94.4%, need 93.9%, timeline 89.0%. Full anchors in Appendix B.	23
4.5	Extraction configuration. Pinned open weights and deterministic decoding make the run reproducible.	25
4.6	From the fourteen raw instrument outputs to the eight engineered features. The eighth engineered feature, <code>nlp_analysed</code> , is a derived coverage flag and has no raw-output source.	26
4.7	Final feature set entering the models, and the groups excluded by design.	27
4.8	Model exploration on the CRM-only baseline, over the 482 modelling deals (34 won); stratified 5-fold CV, mean $\pm$ sd. All intervals overlap, so explainer compatibility decides the deployed model. Peak PR-AUC cannot separate them.	28
4.9	The five dashboard views, their stakeholders, and what each shows. . .	30
5.1	Every experiment reported in this thesis, the population it was computed on, and its design. Positive counts are given because the number of won deals, not the number of deals, is the binding constraint throughout. . .	35

5.2	Model A→Model B ablation (repeated 5×5-fold CV; no-skill PR-AUC = 0.071). Point estimates positive; every interval spans zero.	35
5.3	The three tests for the deployed model: the full-model gain is not significant, reverses without deal value, and is not additive on top of it.	36
5.4	Top drivers, out-of-fold: attribution (mean SHAP) versus predictive importance (permutation Δ PR-AUC). The transcript (NLP) features hold visible attribution but little separable predictive importance once the CRM features are present.	38
5.5	Model Measurement Checks	40
5.6	The three FEDS evaluation steps, their primary evidence, and the verdict for each.	42
5.7	How each SLR gap is addressed, and with what status.	43
5.8	Limitations.	44
A.1	Features entering Model A and/or Model B.	58
A.2	Features deliberately withheld, with the reason.	59
C.1	Design choices in the extraction prompt and their grounding.	63
D.1	Acceptance statements (1 = strongly disagree, 7 = strongly agree).	68
D.2	Discovery dimensions (0 = absent; score by the anchors of Appendix B).	69

List of Abbreviations

Abbreviation	Meaning
AE	Account Executive
AUC	Area Under the Curve
AWQ	Activation-aware Weight Quantization
BANT	Budget, Authority, Need, Timeline
BDR	Business Development Representative
BI	Behavioural Intention
B2B	Business-to-Business
B2C	Business-to-Consumer
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRISP-ML(Q)	Cross-Industry Standard Process for Machine Learning with Quality assurance
CRM	Customer Relationship Management
CV	Cross-Validation
CW	Class Weighting
DSR	Design Science Research
DSRM	Design Science Research Methodology
EPV	Events Per Variable
FEDS	Framework for Evaluation in Design Science
GBT	Gradient-Boosted Tree
GPU	Graphics Processing Unit
HGB	Histogram-based Gradient Boosting
ICC	Intraclass Correlation Coefficient
JSON	JavaScript Object Notation
KDD	Knowledge Discovery in Databases
KR-20	Kuder-Richardson Formula 20
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MEDDIC	Metrics, Economic buyer, Decision criteria, Decision process, Identify pain, Champion
NLP	Natural Language Processing
PEOU	Perceived Ease of Use
PR-AUC	Precision-Recall Area Under the Curve
PU	Perceived Usefulness
RevOps	Revenue Operations
RITAM	Remote IT Asset Management
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
SaaS	Software as a Service
SHAP	SHapley Additive exPlanations
SLR	Systematic Literature Review
SMOTE	Synthetic Minority Over-sampling Technique
SPICED	Situation, Pain, Impact, Critical event, Decision
SPIN	Situation, Problem, Implication, Need-payoff
SUS	System Usability Scale
TAM	Technology Acceptance Model

Abbreviation	Meaning
XAI	Explainable Artificial Intelligence

Chapter 1 Introduction

This chapter introduces the thesis. It sets out the context and motivation for the research, states the problem, and frames the research questions, before outlining the approach and methodology and explaining the structure of the remainder of the thesis.

1.1 Context and motivation

Business-to-business software deals usually begin and advance through conversations between people. Before closing a deal, representatives hold introductory calls to learn the factors that shape the outcome: the buyer’s underlying problem, its urgency, the cost of inaction, and the decision-makers in the buying organisation [70, 92]. Given the importance of these calls, sales teams routinely record, review, and use them for sales coaching [3]. Yet turning these conversations into structured, machine-usable data is difficult, even though doing so would help teams spend limited selling time more efficiently [69, 15].

Lead-scoring systems help prioritise opportunities by estimating the probability that an open opportunity will convert into a paying customer [30, 94]. This lets teams rank their sales pipeline and focus on the most promising deals. In high-volume business-to-business (B2B) settings, where one representative handles many opportunities at once, prioritisation becomes increasingly important: time spent on the wrong deals directly reduces performance. These models rely almost exclusively on structured data fields recorded in customer relationship management (CRM) systems, such as deal value, prospect industry and size, and the number of logged activities [24, 75, 58]. However, the introductory call, which often holds the most diagnostic information, is usually left out. This is because turning unstructured speech data into reliable structured features is technically complex [69, 15].

This thesis investigates a Remote IT Asset Management (RITAM) software vendor whose sales team faces this challenge. The proposed solution is a decision aid for the stakeholders which is the artefact this thesis designs and evaluates, that uses available deal information and introductory-call transcripts to score B2B opportunities. The team handles many B2B opportunities with a CRM that records factual deal information. Even though every introductory call is transcribed, this information is missing in the CRM. Commercial tools for analysing sales calls are available and used within the host organisation [69, 15], but they have two limitations. First, these tools are built for mainstream B2B markets, so their analysis may not transfer to other domains like RITAM [15]. Second, they are not interpretable, so representatives cannot see why a

score was given or adapt the model to their own context [73]. Without transparency, users trust the aid less and it is less useful in practice [27, 90].

1.2 Problem statement

The context and motivation above leave two problems unresolved. Together they motivate this thesis.

1. The introductory call often contains the richest diagnostic signal about an opportunity [70, 92], yet most vendors’ lead-scoring models draw only on structured CRM fields. This raises the question of whether the conversational signals embedded in the call carry predictive value for conversion beyond what CRM fields alone capture.
2. The commercial tools that could bring those signals into scoring are opaque. Representatives cannot see why a score was given, so they trust the tools less and use them less in practice [73, 27, 90].

These problems affect a high-volume B2B sales team in which each representative manages many opportunities at once, so prioritisation decisions directly determine how selling time is spent. Since the call’s diagnostic signals never reach the lead scores, prioritisation rests on an incomplete signal, and selling time goes to opportunities that a transcript-based score might have ranked lower. The information captured in every recorded call also goes unused, despite the cost already incurred to transcribe it.

1.3 Research design

This section defines what the thesis sets out to build and assess: the scope and objectives of the artefact, and the research questions that structure the investigation. The methodological frameworks that govern how the work is carried out are set out in Section 1.4.

1.3.1 Research scope and objectives

This thesis designs and evaluates an artefact: a lead-scoring pipeline that combines a sales team’s customer relationship management (CRM) data with features extracted from introductory calls, produces a score for each deal, and explains in plain language the factors behind that score. Following the Design Science methodology [35, 66], the artefact is a reusable method demonstrated on real data and evaluated for the internal consistency of its measurements.

The study is restricted to a single organisation and a defined subset of its data: 482 closed deals with a transcribed introductory call, of which 34 were converted into paying customers. This single-company setting and the small number of converted deals limit the generalisability of the findings [5]. The artefact scores each deal immediately after its introductory call, using only the information available at that point.

Evaluating the explanations with end users is out of scope for this thesis and is set out as future work: the study establishes the artefact and its measurement properties, and prepares but does not run the user evaluation.

1.3.2 Research questions

This thesis addresses the problems mentioned in Section 1.2 through the following research questions.

Main research question. *To what extent can natural language processing (NLP) features extracted from introductory sales-call transcripts contribute to lead scoring for business-to-business RITAM sales teams?*

This overarching question addresses a practical challenge: RITAM sales teams must decide, after an introductory call, which deals to pursue. However, the most informative conversation is currently underutilised. A systematic literature review, reported in Chapter 2, addresses established literature and defines the research gaps this thesis addresses. To close them, the study builds and evaluates the artefact. The main question decomposes into two facets, taken up by the research questions below: whether the transcript features improve prediction (RQ1), and whether they are sound measurements in their own right (RQ2).

RQ1. To what extent do transcript-derived features improve lead-scoring performance compared to CRM-only features? This question is answered empirically by comparing a model trained solely on CRM data with a model that adds transcript-derived features for the same deals: the performance difference isolates the contribution of conversational data, and the model’s own attributions identify which signals drive that difference and whether the results are credible.

RQ2. How reliable are the transcript-derived features as measurements? This question is answered by evaluating whether the extracted scores are internally consistent and align with theoretical expectations.

RQ1 and RQ2 test Problem 1 empirically. Problem 2, the opacity of commercial tools, is addressed through the artefact’s design, which produces a plain-language explanation alongside each score, rather than through a separate research question.

1.4 Approach and methodology

This section sets out the methodological frameworks that govern the thesis: the research paradigm, the process model that structures the technical work, and the framework used to evaluate the resulting artefact.

1.4.1 Design Science Research (DSR)

The thesis produces an artefact for a practical problem, so DSR is the appropriate paradigm [35, 66]. Hypothesis testing cannot close capability gaps, and an interpretive study of how teams perceive scoring would not produce the missing pipeline. The study follows the six DSRM activities of Peffers et al. (2007) [66] under the design-science guidelines of Hevner et al. (2004) [35]. In DSR the knowledge produced is the validated artefact and the evidence that supports it.

Figure 1.1 shows the Hevner information-systems research framework for this study.

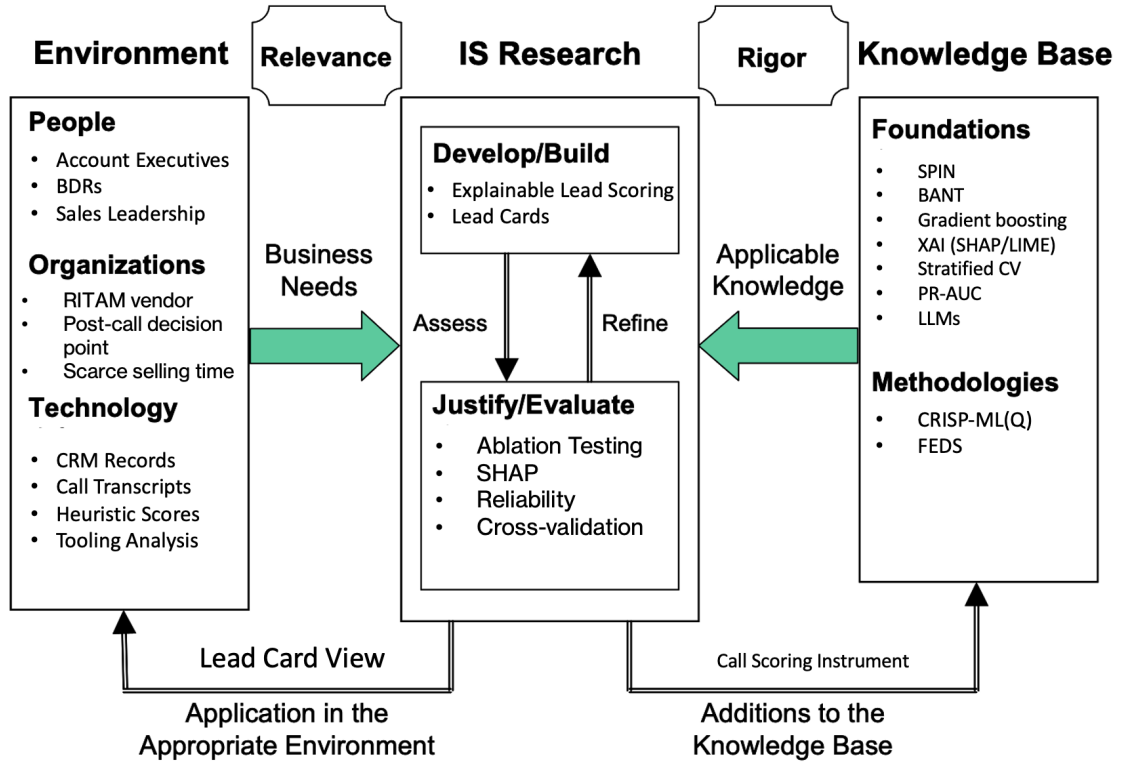


Figure 1.1: The design-science research framework for this thesis [35]. Relevance flows from the RITAM sales environment into the build-evaluate cycle; rigour flows from the knowledge base; the artefact returns to the environment as the lead card and to the knowledge base as a validated call-scoring instrument.

Figure 1.2 places the same work on the six DSRM activities and locates each one in the thesis. The study is problem-centred, so it enters at *problem identification*,

where the systematic literature review of Chapter 2 and the sales-process analysis of Section 3 yield the four gaps the artefact must close. The *objectives of a solution* are the research questions and the artefact definition, both fixed in Section 1.3.1. *Design and development* is carried out in Chapter 4, which produces the scoring pipeline, the LLM extraction instrument, and the explainable model. *Demonstration* follows in Section 4.5, where the artefact is exercised on real deals and rendered as lead-card views with their associated deal metrics. *Evaluation* is reported in Chapter 5 and organised by the Framework for Evaluation in Design Science (FEDS), covering technical performance, interpretability, and measurement validity. *Communication* runs through the discussion and conclusion of Chapters 5 and 6 and, for the practitioner audience, through the lead-card format itself.

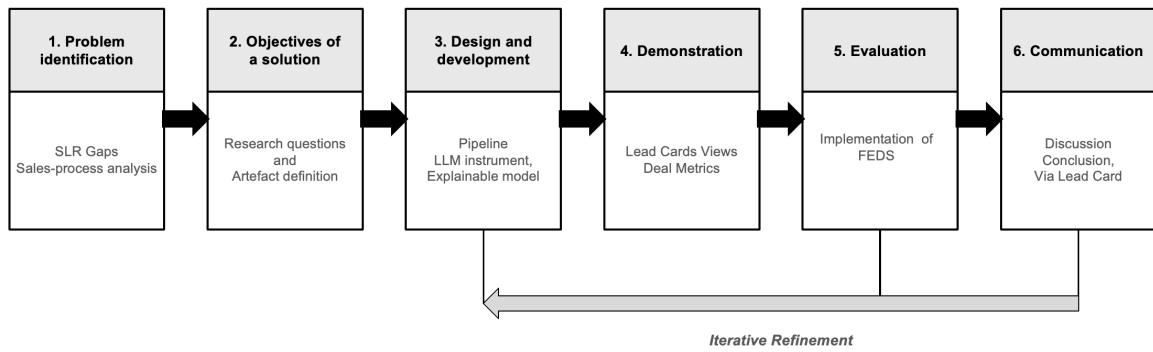


Figure 1.2: The six DSR methodology activities [66].

The loop at the foot of Figure 1.2 is the substantive part. Evaluation does not terminate the process: each evaluation pass returns to design and development, so the artefact reported in Chapter 4 is the settled product of several such passes rather than a single forward sweep. The leakage screen, the admissibility rule at the decision point, and the final feature set all originate in this loop, and the entries in Chapter 5 report only the results of the last iteration.

1.4.2 CRISP-DM and CRISP-ML(Q)

The artefact is developed under CRISP-ML(Q) [82], which is the machine-learning extension of CRISP-DM, the standard cross-industry data-mining process [14]. CRISP-ML(Q) keeps the CRISP-DM phase structure but attaches a quality-assurance step to every phase: each phase names the risk that could undermine it and the check that mitigates it (Table 1.1). That explicit quality layer is why CRISP-ML(Q) was preferred over classical CRISP-DM and over KDD [25]. Artefact evaluation is conducted using the FEDS framework [89].

Table 1.1: CRISP-ML(Q) quality gates for this study: the principal risk at each phase and the check applied.

Phase	Principal risk(s)	Quality check applied
Business & data understanding	Wrong success metric under heavy imbalance; selection bias and label leakage in raw fields	PR-AUC primary, accuracy never reported; documented population construction; leakage screen
Data preparation	Features that track the outcome or rep behaviour when only deal quality should count	Decision-point admissibility rule; leakage exclusions; blind, outcome-independent variance filter
Model engineering	Overfitting a 34-positive class; selecting on noise between models	Repeated stratified CV; class weighting; confidence intervals, never bare point estimates
Model evaluation	Overclaiming a small effect	Pre-registration; significance, permutation, and additivity tests
Deployment (demonstration)	Mis-calibrated scores misleading a representative	Percentile ranking, paired with a readable SHAP explanation

1.5 Structure of the thesis

The remainder of this thesis is organised as follows. Chapter 2 reviews the literature, answers the three review questions, and sets out the theoretical foundations the later chapters rely on. Chapter 3 describes the research context at the host organisation and formalises the problem. Chapter 4 develops the artefact: the data and population, the measurement instrument, feature engineering, and the model engineering and analysis design. Chapter 5 presents and discusses the results for RQ1 and RQ2, including the study’s limitations. Chapter 6 concludes with the contributions and with recommendations for the host organisation and for future research.

Chapter 2 Literature Review

This chapter briefly presents the systematic literature review (SLR) conducted in the initial phases of the thesis. It answers three review questions, synthesises the research gaps that the thesis targets, and concludes with the theoretical foundations for the chapters that follow. The SLR carries three limitations: it drew on a single database, was conducted by one reviewer, and included no RITAM-industry studies. Its findings are therefore inferred from adjacent business-to-business (B2B) contexts. Table 2.1 summarises the review questions and their corresponding gaps.

The review is guided by the following questions:

- **Q1.** How effectively have machine learning methods been applied to lead scoring and conversion prediction in B2B sales?
- **Q2.** To what extent does NLP- or large language model (LLM)-based feature extraction improve prediction from unstructured sales communication?
- **Q3.** How effective is Explainable Artificial Intelligence (XAI) in B2B decision-making?

Table 2.1: SLR Overview

Review question	Answer	Gap(s)
Q1: ML for lead scoring	Mature. Gradient-boosted trees dominate on structured CRM data; feature and label design matter more than the classifier.	Feature space is structured-only (transcript gap); imbalance handling inconsistent (class-imbalance gap).
Q2: NLP/LLM feature extraction	Promising but thin. One study uses an LLM as a feature extractor over buyer text, with predictive success.	No B2B call-transcript features combined with CRM data; extracted features never validated as measurements (measurement-validity gap).
Q3: XAI in B2B decisions	Technically settled. SHAP and LIME are feasible at enterprise scale and complementary.	Only 2 of 27 studies tested explanations with end users (user-evaluation gap); visible-but-meaningless explanations inflate trust.

Each of the following three sections reviews the relevant literature before closing with a direct answer to its review question.

2.1 Machine learning for lead scoring

Q1. *How effectively have machine learning methods been applied to lead scoring and conversion prediction in B2B sales?*

The literature consistently frames lead scoring as a supervised binary classification task that is built on structured customer relationship management (CRM) data. Gradient-boosted tree ensembles are the most widely adopted and best-performing models for these tasks. González-Flores et al. (2025) [30] compare fifteen classifiers on 16,600 deals and find that gradient boosting performs best (F1 0.93, AUC 0.99). Eitle and Buxmann (2019) [24] report CatBoost as the leading model on a Fortune 500 pipeline, and Utomo et al. (2025) [86] reach an F1 of 0.91 with XGBoost in a B2B SaaS context. No algorithm, however, consistently outperforms the others without hyperparameter tuning, so classifier selection is a secondary concern. Performance gains come mainly from data engineering and problem framing: engineered meta-variables [58], multi-class label design [97], and segment-wise pipelines [94, 24] each yield larger improvements than changing the algorithm.

The corpus has two main weaknesses. First, the feature space is almost exclusively structured (firmographics, deal attributes, and activity counts) [24, 75]; no reviewed study uses the content of B2B sales calls. Second, class-imbalance handling is inconsistent: some studies apply SMOTE and report accuracy [62], others use Bayesian re-weighting [97], and some do not address imbalance at all [72]. In one case a reported 99 percent accuracy conceals zero percent recall on the minority class [54]. Together these findings answer Q1: machine learning predicts B2B conversion effectively from structured data, with gradient-boosted trees as the prevailing standard. The two open problems are the reliance on structured features and the absence of disciplined imbalance handling.

2.2 NLP and LLM feature extraction

Q2. *To what extent does NLP- or large language model (LLM)-based feature extraction improve prediction from unstructured sales communication?*

Literature on language technology for sales prediction remains limited, especially in the business-to-business (B2B) context. Transformer encoders [23] have advanced the representation of free text, but the most relevant study takes a different route. Majid and Laroche (2025) [51] use ChatGPT-4.0 as a measurement instrument to score 307 inbound inquiries for linguistic specificity and personalisation, then feed the scores into a hazard model for conversion and cycle time. The setting is business-to-consumer (B2C), yet the study sets a methodological precedent for using a large language model as a feature extractor for buyer text. It also leaves a gap: the scores are validated only

through downstream model accuracy. They are never independently measured. This gap motivates Q2: features that large language models extract from sales communication may improve predictive performance, but the evidence is limited to a single adjacent study. No study combines B2B call-transcript features with customer relationship management (CRM) data, and none validates the extracted features as independent measurements.

2.3 Explainable AI in B2B decision-making

Q3. *How effective is Explainable Artificial Intelligence (XAI) in B2B decision-making?*

SHAP [50] and LIME [73] rest on well-established theory, operate at enterprise scale (200,615 records in Marín Díaz et al. (2022) [54]; 315,121 in Ayaz et al. (2024) [4]), and provide complementary global and local interpretability. Marín Díaz et al. have applied this approach across several enterprise settings [55, 53, 52]. Yet application does not equal validation: only two of the twenty-seven studies evaluated explanations with non-technical end users. In one, participants trusted randomly perturbed SHAP plots as much as coherent ones whenever any visual was present, leading to a placebo effect [90]. In the other, explanations tailored for sales representatives were comprehensible without prior training [24]. Together these results answer Q3: XAI methods are technically effective and deployable in B2B contexts, but their value for the decision-makers who use them remains largely asserted. Explanation quality must therefore be validated with users [19, 76].

2.4 Theoretical foundations

Four concepts from the literature form the foundation for the subsequent chapters. Table 2.2 identifies each concept and the principle it supports.

Table 2.2: Theoretical Foundations

Foundation	Core Principle
Lead scoring as a machine-learning problem	Imbalanced binary classification; PR-AUC against a no-skill baseline, never accuracy; leakage guarded by a fixed decision point
LLMs as measurement instruments	Reproducibility (pinned weights, temperature-zero decoding, schema-constrained output) and demonstrated validity
Sales-conversation theory	SPIN and organisational-buying theory ground what the instrument measures
Explainable AI (SHAP, LIME)	Exact tree Shapley values, user-facing, with LIME as an internal cross-check; visible explanations can inflate trust

2.4.1 Lead scoring as a ML problem

Lead scoring estimates the probability that each open opportunity will convert, so sales teams can prioritise their pipeline and spend limited selling time on the deals most likely to close [94, 62]. As a machine learning task, lead scoring is a supervised binary classification problem. The label is the eventual outcome (won or lost), and the features describe the state of the deal at a specific decision point. The model learns the mapping from features to outcomes on historical closed deals [62, 97]. Two characteristics of the problem shape the design decisions that follow.

The first is class imbalance. Won deals are far less common than lost deals, and on such skewed data accuracy becomes uninformative: a model that labels every deal “lost” scores high on accuracy yet finds no wins [24, 54]. The area under the receiver operating characteristic curve (ROC-AUC) is more robust, but it can still overstate performance when negative cases dominate: it rewards true-negative predictions that do not matter to sales teams [22] (Figure 2.1). The most appropriate metric is the area under the precision-recall curve (PR-AUC), which focuses on the minority (positive) class. It is read against a no-skill baseline equal to the positive rate, the score a random ranker would achieve [78]. Imbalance can also be handled during training, by re-weighting the classes or synthetically oversampling the minority class [16]. Both approaches have costs: oversampling in particular may increase variance without improving ranking performance when positive cases are scarce [9]. A small positive class also introduces statistical fragility: performance estimates can vary substantially across data splits, and model or hyperparameter selection on a single split may reflect chance variation in that split [13, 87]. This thesis therefore uses repeated stratified cross-validation throughout and reports results as intervals.

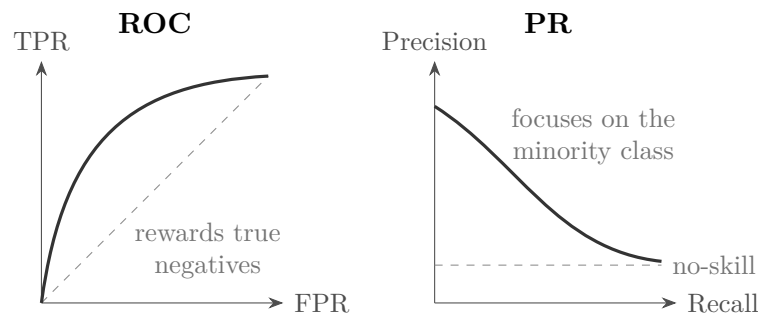


Figure 2.1: Why PR-AUC is preferred over ROC-AUC under heavy class imbalance: ROC rewards abundant true negatives, whereas the precision–recall curve is read against the no-skill baseline set by the positive rate [22, 78].

Another characteristic is data leakage: using as a feature any information that is unavailable at prediction time or that directly encodes the outcome. A leaked feature lets a model “predict” outcomes it has effectively already been told, so development performance looks strong but collapses in deployment [38]. Leakage is a leading cause

of over-optimistic and irreproducible machine-learning findings [37]. The only reliable safeguard is to fix a single decision point and include only information available by that time [38].

2.4.2 Large language models as measurement instruments

This thesis uses a large language model (LLM) as a measurement instrument: it reads call transcripts and returns structured scores, much as a human annotator applies a coding scheme. Gilardi et al. (2023) [29] show that instruction-tuned LLMs can annotate text about as reliably as trained human coders on some tasks, and do so far faster and more cheaply. This makes them a practical means of turning unstructured conversational text into structured features [85, 39]. Using a model this way imposes two requirements that ordinary text generation does not: the measurements must be reproducible and valid [32, 34].

A language model does not reproduce its outputs by default. The same transcript submitted twice can return different scores, and a hosted model may change without notice unless its version is pinned [32]. Repeatability therefore has to be engineered, and three settings provide it: the model weights are pinned to a single version, decoding runs at temperature zero with a fixed seed, and the output is constrained to a fixed schema, so every call returns the same fields in the same form [85]. The measurement task is constrained the same way. Rather than return a single overall judgement, the model scores a set of named dimensions, each accompanied by a written description of what every level denotes [40, 48, 95], in the manner of a human coder applying a codebook [39].

Reproducibility is not the same as accuracy. Obtaining the same output twice says nothing about whether that output is right, so two separate questions arise (Figure 2.2). The first is whether the instrument is consistent: whether its dimensions agree with one another and whether it returns the same result on repeated measurement, which is assessed with Cronbach’s α [83]. The second is whether the instrument is correct, that is, whether it measures what it claims to. Correctness is established against human raters, using weighted κ and the intraclass correlation coefficient read against published thresholds [44, 43]. The two can diverge: a bathroom scale that reads three kilograms heavy is perfectly consistent yet wrong, and a model can be the same, so consistency is no proof of correctness [65, 34]. Predicting an outcome and measuring a construct accurately are not the same thing; RQ2 concerns the latter.

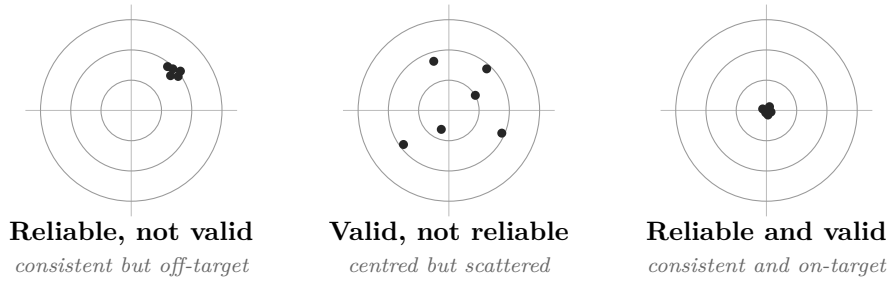


Figure 2.2: Reliability versus validity for a measurement instrument. Reliability is consistency (a tight grouping); validity is agreement with the truth (a grouping on the centre). An instrument can be reliable yet invalid, so the two must be evidenced separately.

2.4.3 Sales-conversation theory

The instrument’s measurement objectives rest on established theory. Two bodies of work provide this foundation.

The first body of work concerns what makes an introductory call effective. The SPIN model [70], developed from observations of thousands of sales calls, explains that effective discovery involves four types of questions (Figure 2.3). Situation questions establish the buyer’s current context, Problem questions identify difficulties and dissatisfactions, Implication questions explore the consequences of those problems, and Need-payoff questions prompt the buyer to articulate the value of a solution. Rackham’s (1988) central finding is that the consequence-oriented questions (Implication and Need-payoff) distinguish discovery that advances a deal from mere information gathering. Three of the instrument’s discovery dimensions (situation understanding, pain depth, and impact and urgency) operationalise this progression from context to consequence. In addition to the discovery rubric, the instrument captures the four BANT qualification signals that practitioners use to assess opportunity viability: Budget, Authority, Need, and Timeline [45].

The second body of work explains why these signals carry information. Organisational-buying theory holds that business purchases are made by a buying centre of multiple interacting roles [92], and that decisions follow a structured, multi-stage, multi-participant process [81]. An introductory call is typically the first point at which this structure becomes apparent: it reveals who holds budget and authority, which needs are shared across the buying centre, and the organisation’s timeline. Consequently, qualification and discovery signals from a conversation in an introductory call are, in principle, informative. The thesis’s contribution is the method that turns such conversations into structured, theory-grounded features, while the underlying constructs are well established.

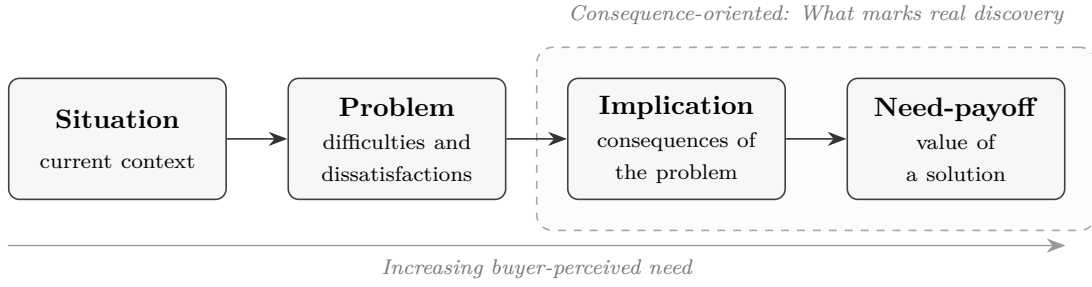


Figure 2.3: The SPIN model of sales discovery [70].

2.4.4 Explainable AI: SHAP and LIME

Gradient-boosted tree ensembles and other top-performing models for tabular data are often opaque and give no rationale for their predictions [50, 19]. In decision-support settings this opacity is a practical problem, because users are unlikely to trust or act on scores they cannot understand or interpret [57, 27]. Two strategies address it: use intrinsically interpretable models, or keep high-performing models and add post-hoc explanations [67, 19]. This thesis takes the latter approach and pairs each prediction with explanations at two levels. A global explanation identifies which features matter across all cases, while a local explanation clarifies why a specific instance received its score [19]. Both views are needed in lead scoring: the global view justifies the overall ranking, and the local view supports the score given to an individual case.

Two complementary methods generate these explanations. SHAP [50] attributes a prediction to individual features using Shapley values from cooperative game theory, which distribute the prediction’s ‘payout’ fairly and guarantee additivity and consistency. For tree-based models, an exact and efficient algorithm lets SHAP run at scale [49]. SHAP provides both global and local explanations and is the primary user-facing method (see Figure 2.4). By contrast, LIME [73] fits a simple, interpretable surrogate model near a single prediction. Since SHAP and LIME produce local explanations by independent mechanisms, their agreement is an internal credibility check; LIME output is never shown to users.

This approach carries an important caution. Explanations that look authoritative can create unwarranted confidence, whatever their actual validity. In one controlled study, participants trusted randomly perturbed SHAP plots as much as coherent ones as long as some visual was present [90]. An explanation’s usefulness also depends on the user and the task [46, 27]. Two implications follow. First, an explanation’s value must be validated empirically with real users, so a user study is left to future work. Second, the magnitude of a feature’s attribution should not be conflated with its predictive importance [67], a distinction tested here with permutation importance [26].

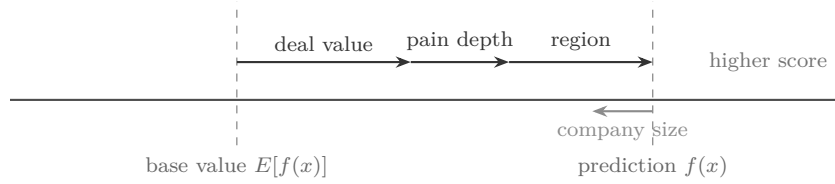


Figure 2.4: Schematic of a local SHAP explanation: each feature pushes the prediction away from the base value, and the signed contributions sum exactly to the score for that deal.

Chapter 3 Business Context

The host sells Remote IT Asset Management (RITAM) software on a subscription model with service-level commitments. Revenue accrues over the contract lifetime, so acquisition cost is concentrated up front. Since a deal recovers that cost only across a multi-year contract, pursuing the wrong deal forfeits an entire contract’s recurring revenue. Misallocated selling time therefore costs more here than in a one-off transactional sale. The binding constraint is selling time. Qualification is an allocation decision, and it is where revenue operations have the most leverage.

This chapter sets out the setting the artefact is built for. It describes the sales process and the decision the artefact supports, the data a deal carries, the scoring signals already in place, and the formalisation of the problem that the following chapter builds against.

3.1 The sales process

Deals move through a multi-stage CRM pipeline from first contact through an introductory call, discovery, demonstration, and proposal, to a closed-won or closed-lost outcome (Figure 3.1). A cycle spans several months, and the same pipeline is operated by regional teams in different markets.

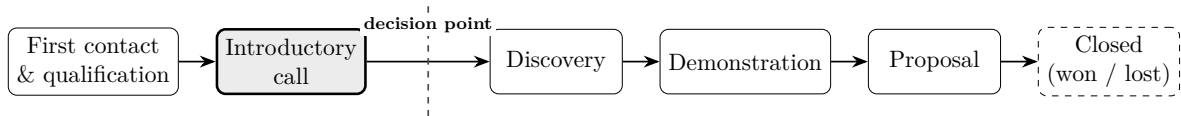


Figure 3.1: The host organisation’s sales pipeline. The decision point (the end of the introductory call).

The introductory call is the pivot. A B2B purchase is made by a buying centre of several interacting roles reached through a structured, multi-participant decision process. The introductory call, recorded and transcribed for every deal, is the first point at which the decision regarding the prioritisation of the deal takes place. It surfaces the buyer’s problem, its urgency, the cost of inaction, and who holds purchase authority. It is the discovery conversation the sales literature holds to be diagnostic. Immediately after it, the representative must decide how much further selling time the deal deserves, and that decision is the one the artefact supports.

Four stakeholder groups depend on that decision (Table 3.1). A score that is accurate but opaque fails all four, which is why the artefact pairs a ranking with an interpretable rationale.

Table 3.1: Stakeholders in the lead-scoring decision and what the score does for each.

Stakeholder	Stake in the score
Account executives (AEs)	The score reallocates their scarce selling time toward the deals most likely to convert.
Business development representatives (BDRs)	They feed the pipeline; the score signals which qualified hand-offs are worth prioritising.
Revenue operations (RevOps)	Owns the CRM data the score is built on; the researcher’s own vantage point, with internal access.
Sales leadership	Forecasts and plans capacity from the aggregate prioritisation the score produces.

3.2 Process data

The CRM accumulates structured deal records (value, region, channel, stage), company firmographics from third-party enrichment, and automatically logged engagement activity. Uniquely, the introductory call is recorded and transcribed, yet no structured field captures its content, and that gap is the opportunity this thesis pursues. Table 3.2 inventories what a deal carries; which of these enter the model, and which are excluded, is decided during data preparation.

Table 3.2: The data a deal carries in the CRM and call-recording platform. Every group but the last is a structured field; the introductory-call transcript is the only unstructured source, and no structured field captures its content.

Group	Examples	Origin
Deal	value, region/team, acquisition channel, stage, time in stage	recorded by the CRM
Company (firmographic)	headcount, annual revenue, industry, founding year	third-party enrichment
Engagement activity	email, task, and meeting counts and timestamps	auto-logged by the CRM
Qualification	budget, authority, need, timeline; methodology fields	entered by the representative
Production scores	rule- and model-based deal scores (Section 3.3)	derived by existing scoring layers
Introductory call	full transcript of the first substantive conversation	recorded and transcribed; <i>unstructured</i>

3.3 Current situation

Prioritisation combines representative judgement with automated scores that accumulated over time and run in parallel (Table 3.3). Three weaknesses follow. The scores are heuristic or proprietary, so a representative cannot see why a deal scored as it did. Several are in transition or inconsistently populated. Finally, the internal call-analysis system, the only component that reads conversations, is opaque and vertical-specific. The conversation therefore reaches prioritisation only through whichever mechanism a representative happens to trust. These scores are excluded from the artefact because they are heuristic-derived.

Table 3.3: The scoring signals already present in the CRM today.

Signal	How it is produced	What it reflects
Sales Reps’ Instinct	Experience and call notes	A read on the deal, neither recorded nor comparable across reps
Process-quality score	Rules on follow-up and record completeness	Process discipline only
Qualification scores (BANT, SPICED, MEDDIC)	Methodology fields, rep- or automation-filled	Whether qualification criteria are met
Stage-gap timing scores	Rules on elapsed time between stages	Deal momentum
CRM-native predictive scores	The CRM platform’s own models	A generic conversion estimate
Transcribing tool’s scoring	Automated, layered summaries of recordings	Conversation themes

3.4 Problem Formalisation

The business problem is binary classification of deal outcome at a fixed decision point, the end of the introductory call, using only information available then. Restricting the population to deals that reached a transcribed introductory call is deliberate: deals that close without a conversation are filtered out earlier on fit and pose no lead-scoring problem. With won deals rare, the primary metric is PR-AUC against the no-skill baseline set by the population win rate. Top-decile precision serves as a secondary, deployment-facing metric; accuracy is never reported. Table 3.4 represents a tabular version of the Machine Learning Canvas; it fixes these commitments before any data analysis [68].

Table 3.4: Machine Learning Canvas [68] for the lead-scoring artefact.

Canvas block	Commitment for this artefact
Value proposition	Help a representative spend selling time on the deals most likely to convert, with a reason they can read, so the score can be trusted and acted on.
Prediction task	Binary classification of deal outcome (won/lost) at a fixed decision point: the end of the introductory call.
Decision supported	Ranking of open deals so the most promising are worked first; the score is decision support the representative may override.
Data sources	The host’s CRM (deal and firmographic fields) joined at deal level to the transcript of the introductory call.
Features	Structured CRM fields knowable at the decision point, and theory-grounded discovery features extracted from the transcript.
Offline evaluation	PR-AUC against the population no-skill baseline (primary); winners in the top decile (secondary); never accuracy. Repeated stratified cross-validation.
Making predictions	A single explainable gradient-boosted model emits a score per deal, paired with a SHAP-based plain-language explanation.
Monitoring (future)	Re-evaluation of ranking quality and feature drift on new closed deals; out of scope for the demonstration.

Chapter 4 Model Development

This chapter develops the artefact. It follows the CRISP-ML(Q) phases adopted in Section 1.3. Each section covers the corresponding phases of the CRISP-ML(Q) methodology. The chapter opens with an overview of the pipeline.

4.1 Overview

Figure 4.1 gives an end-to-end overview of the developed artefact. Each introductory transcript is first masked [15] and then scored by an open-weight language model against a fixed rubric. Those scores are encoded into a compact feature set and combined with the CRM fields. The combined set is then passed to a gradient-boosted tree, and the tree’s prediction is paired with a SHAP explanation. The classifier and explanation steps are deliberately conventional, so that any observed effect can be attributed to the features. Model complexity stays constant.

4.2 Data collection and population (Phase 1)

The dataset joins the host’s CRM to transcript features from its call-recording platform at the deal level [32]. Table 4.1 documents how the modelling population was constructed. The snapshot is fixed: deals closed by 31 March 2026, created from 1 January 2025, with a deal amount present. This makes the population reproducible and restricts it to the time period in which transcription tooling was in place. That restriction removes a structural bias. Before the filter, deals with no transcripts were overwhelmingly old lost deals, which made transcribed deals look 3.7 times more likely to be won. Once the pre-transcription deals are excluded, that coverage asymmetry falls to a mild 1.2 times [38, 36].

Table 4.1: Construction of the modelling population. Filters applied in order.

Filter	Deals	Won
Closed deals in the snapshot window (closed $\leq$ 31 Mar 2026; created $\geq$ 1 Jan 2025, the transcription-tooling era; deal amount present)	1,645	97
... with a transcribed introductory call (modelling population)	482	34

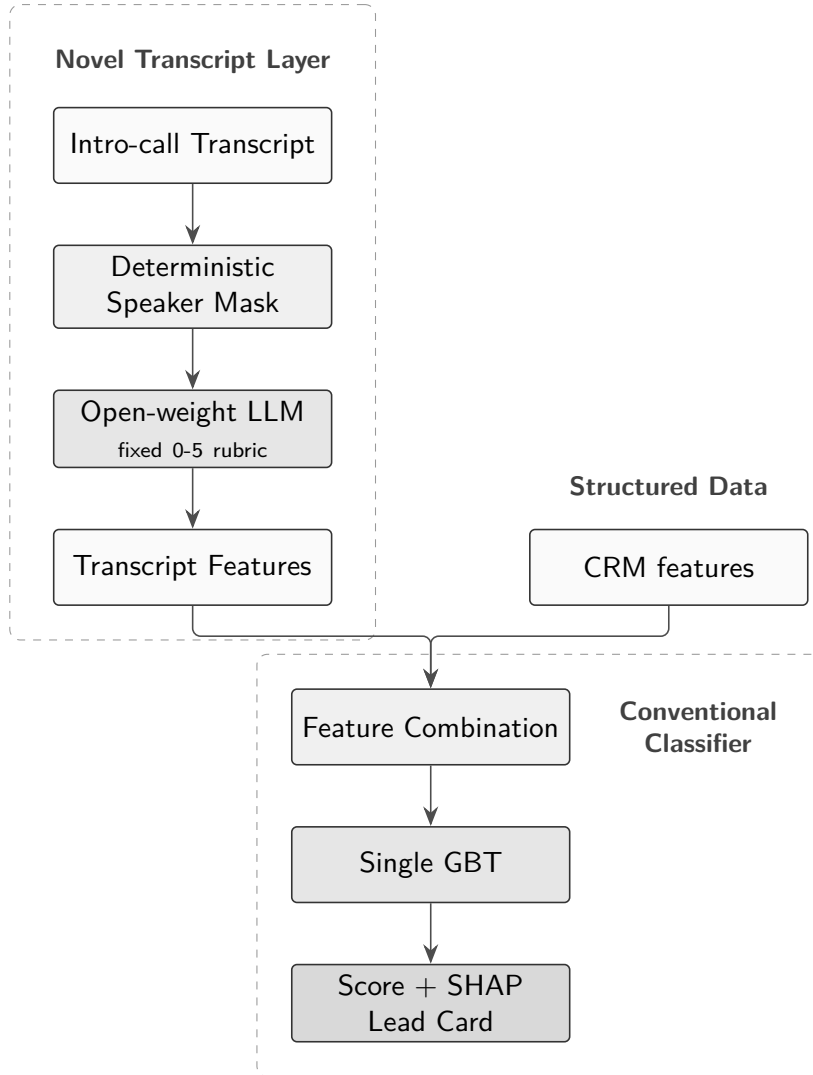


Figure 4.1: The lead-scoring pipeline end to end. The transcript path (left) is the novel feature layer; the combined features feed a single gradient-boosted tree whose score is paired with a SHAP explanation.

Exploratory analysis fixed three facts that shaped the design.

1. **Severe class imbalance.** The population is 13.2:1 (34 won, 448 lost, a 7.0% win rate), so PR-AUC is read against a no-skill baseline of 0.071.
2. **Region confounded with outcome** (Table 4.2). The US team wins at roughly twice the NL rate, so the feature of "region/team" enters the model as an explicit feature and the results are also read region by region. NL's eight won deals sit at the floor of what stratified cross-validation can hold [87].
3. **Deal value does not separate the classes at the mean** (won about EUR 52.7k against lost about EUR 53.6k; Figure 4.2). This motivated the hypothesis that the missing signal lies in the conversation, an evaluation of the artefact later revisits this critically.

One data-quality finding forced a change of plan: the host's pre-existing NLP feature

tables cover under 1% of the relevant deals, so the transcript features had to be extracted independently.

Table 4.2: Regional composition of the modelling population.

Team / region	Won	Lost	Win rate
US Sales (AMER)	20	188	9.6%
NL Sales (EMEA)	8	170	4.5%
Other	6	90	6.3%
Total	34	448	7.0%

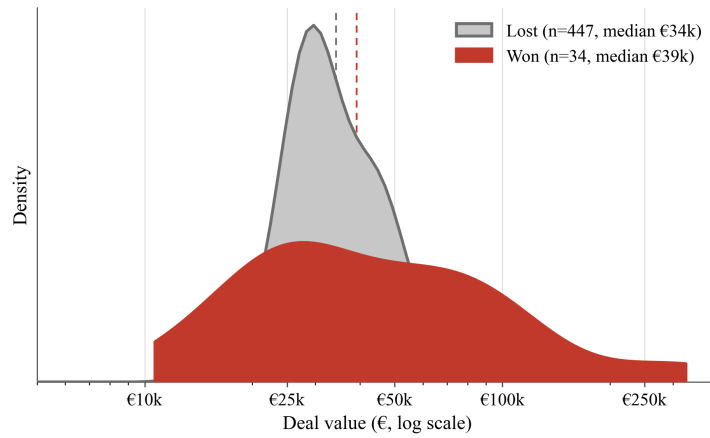


Figure 4.2: Deal value for won versus lost deals. The near-identical group means motivated the extraction of conversational features.

4.3 Data preparation and the instrument (Phase 2)

4.3.1 CRM features and the leakage screen

The structured features are intentionally limited and are all available at the decision point. These include deal amount (log-transformed to address skewness, 91.7% filled), owner team, region, and acquisition channel (all one-hot encoded) [24, 71]. They also include firmographics such as headcount and annual revenue (both log-transformed), and company age calculated from the founding year. Missing values in numeric fields are imputed with the median, while categorical gaps are assigned an explicit “missing” category [71, 86].

Exclusions are essential in this context. Under the decision-point rule, four groups of fields are excluded from the model: all post-close fields, the stage-derived win-probability field, heuristic production scores, and pre-call engagement counters. The win-probability field is excluded because it has a single-feature AUC of approximately 1.0 and effectively reveals the label. The exclusion of the last group is based on empirical evidence. As

shown in Table 4.3, each counter is recorded significantly more frequently for one class than the other, indicating that its presence encodes the outcome through the sales representative’s behaviour [38].

Table 4.3: Leakage screen: fill rate by outcome for the excluded counters. A large gap means the field’s presence encodes the outcome.

Field	Filled (won)	Filled (lost)	Gap
Pre-call email count	17.6%	37.9%	−20.3pp
Pre-call task count	35.3%	15.6%	+19.7pp
First meeting scheduled	49.1%	82.9%	−33.8pp
Discovery call completed	3.5%	24.8%	−21.3pp

4.3.2 Measurement Instrument

The transcript features represent the artefact’s novel component, generated by a custom-designed measurement instrument. Exactly one call is scored per deal: the earliest transcribed call, which is the introductory call. This fixes an outcome-independent decision point: later calls occur only if a deal advances, so scoring them would leak the trajectory [38]. In 99.3% of checkable cases, the earliest transcript corresponds to the introductory call.

An anchored rubric is applied to the transcript, scoring seven ordinal dimensions on a 0 to 5 scale, four qualification booleans, and three behaviour counts [32]. The model is required to provide supporting evidence prior to each score and is prohibited from inferring the deal’s outcome. Figure 4.3 links each output to its theoretical source, while Table 4.4 details the measurement and distribution of each dimension across the 446 analysed calls (of the 482 deals; the 36 that were not analysed are explained below). The complete per-level rubric is provided in Appendix B. This design aligns with rubric-based large language model (LLM) evaluation [40, 48, 95] and adheres to the validation standards established in the LLM-annotation literature [29, 65].

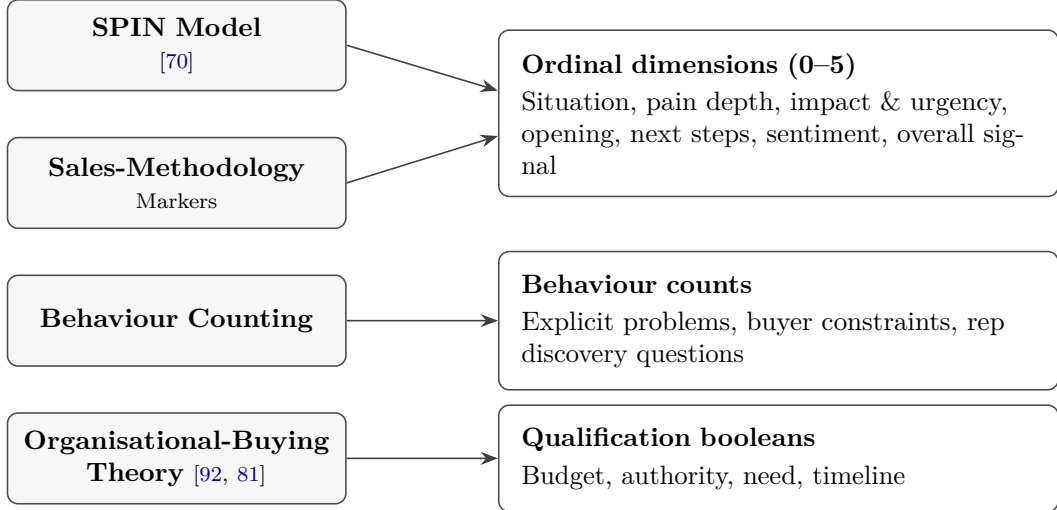


Figure 4.3: Mapping from theoretical sources to instrument outputs.

Table 4.4: The seven ordinal dimensions and their distribution across the 446 analysed calls (integer 0–5; sentiment uses 1–5 for graded affect, with 0 only when affect is indeterminable). Qualification prevalence: budget addressed in 72.2% of calls, authority 94.4%, need 93.9%, timeline 89.0%. Full anchors in Appendix B.

Dimension	What it measures	Mean	SD	Mode
Opening & alignment	Agenda set, time confirmed, buyer goals drawn out	3.70	0.98	4
Situation understanding	Depth to which the buyer’s current state is mapped	3.89	0.75	4
Pain depth	Clarity and severity of the problem surfaced	3.61	1.03	4
Impact & urgency	Business consequence and any compelling event	3.56	0.94	4
Next steps	Concreteness of the agreed follow-up	4.10	0.92	4
Sentiment	Engagement, scored by observable behaviours	3.99	0.34	4
Overall signal strength	Holistic quality of discovery execution	3.98	0.64	4

The rubric incorporates two mechanisms to prevent leakage. First, overall signal strength is anchored to discovery execution. The deal outcome is kept out of the judgement. Second, the evidence-before-score ordering requires the model to quote the transcript prior to assigning a rating [32]. The rubric underwent two iterations. Version 1 resulted in clustered scores, with BANT frequently marked as true and sentiment remaining flat. Subsequent versions introduced a common-mistakes section, behaviour-counting anchors for sentiment, and paired calibration excerpts [39, 32].

The system prompt implements this design directly. Rule 2 extends the decision-point leakage discipline of Section 2.4.1 into the instrument: the model is told, in the same terms used elsewhere in the pipeline, to ignore the deal’s eventual outcome. Rule 4 requires evidence before score, the same reasoning-before-verdict approach used in rubric-based LLM evaluation [48, 40] and grounded in the chain-of-thought literature

[91, 42]. The excerpt below quotes the prompt directly. The complete text, including the anonymisation protocol, the seven dimension anchors, the qualification and count definitions, the calibration examples, and the output schema, appears in Appendix C.

```
You are a B2B sales-conversation analyst scoring a single introduction (
discovery)
call between a sales representative ("the rep") and a prospect ("the buyer")
at a
B2B technology vendor. Your scores become features in a research study, so
they
must be consistent, calibrated, and grounded strictly in the transcript. You
evaluate HOW the call was conducted and WHAT the buyer revealed - never
whether
the deal will close.

SCORING RULES
2. NO OUTCOME INFERENCE. Do not reason about, hint at, or let your scores be
swayed by whether the deal closed, the stage it reached, or how promising it
looks commercially. Score conversation execution and revealed buyer state
only.
4. EVIDENCE BEFORE SCORE: write the transcript evidence first, then the
number;
the score must follow from the evidence you wrote.
```

4.3.3 Extraction: privacy, infrastructure, and coverage

Before model processing, each transcript is masked by a deterministic procedure. Speaker labels become positional tokens such as [SPEAKER_A]; regular expressions mask emails, phone numbers, and URLs; and a predefined term list masks the host and named competitors [15]. Personal names are never written to any file, so the privacy guarantee does not depend on the completeness of a redaction list. Since speakers carry only positional labels, the model must infer for itself which speaker is the representative [3].

Extraction used an open-weight model (Qwen2.5-14B-Instruct-AWQ) deployed locally via vLLM, with temperature set to zero, a fixed seed, and schema-enforced JSON output (Table 4.5). This model was chosen over commercial APIs for two reasons: data governance, since transcripts remain within the local environment, and reproducibility, since the model weights are fixed [85, 32]. Among the alternatives considered, a 7B variant was rejected for degenerate score variance. Of the 482 deals, 446 were analysed; the remaining 36 either fell below the 15-minute substance threshold or lacked a usable transcript. These 36 cases were retained and flagged, so that “not analysed” functions as a learnable value for the model. Capping the evidence-string

length in the schema eliminated all remaining parse failures.

The extraction model was selected on measurement reliability alone. Downstream PR-AUC played no part in the choice [87]. Selecting it by a metric computed from only 34 positive cases would introduce selection bias. Since reliability was the selection criterion, the measurement question (RQ2) stays independent of the performance question (RQ1).

Table 4.5: Extraction configuration. Pinned open weights and deterministic decoding make the run reproducible.

Setting	Value
Model	Qwen2.5-14B-Instruct-AWQ (open weights)
Serving	vLLM
Hardware	Two NVIDIA A16 GPUs, tensor-parallel (full 32k context exceeds one 15 GiB card)
Decoding	Temperature 0, fixed seed; schema-enforced JSON
Coverage	446 of 482 deals analysed (36 below the 15-minute floor or unusable)
Parse failures	0 (after the schema fix)

4.3.4 Feature engineering and the final matrix

The fourteen raw outputs were condensed into a compact, theory-driven feature set before the final extracted data were available. This served as a pre-registration step [61], preventing the feature engineering from being tuned to the outcomes. Figure 4.4 illustrates the process: the fourteen raw outputs are reduced to eight engineered features. A blind, outcome-independent variance filter removes near-constant fields, including sentiment. The standalone authority and need booleans are dropped the same way, but not lost: both are retained within the qualification count [47]. With 34 positive cases and about 24 model inputs once categorical fields are one-hot encoded (seven CRM and firmographic features and eight engineered transcript features, plus the expanded region, team, and channel dummies; Appendix A), the events-per-variable ratio is about 1.4 [74, 87].

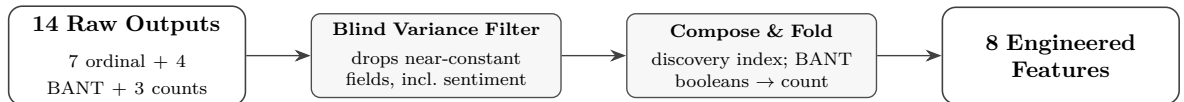


Figure 4.4: Feature-engineering funnel: the fourteen raw instrument outputs are reduced to eight engineered features by a blind, outcome-independent variance filter and theory-driven composition, fixed before the final data were seen.

The primary representation converts each integer score into its expected value under the model’s next-token probability distribution over the digits 0 to 5. This restores the

Table 4.6: From the fourteen raw instrument outputs to the eight engineered features. The eighth engineered feature, `nlp_analysed`, is a derived coverage flag and has no raw-output source.

Raw output	Type	Engineered feature
<i>Ordinal dimensions (7)</i>		
Opening & alignment	0–5	discovery_index
Situation understanding	0–5	discovery_index
Pain depth	0–5	discovery_index; pain_depth
Impact & urgency	0–5	discovery_index; impact_urgency
Next steps	0–5	discovery_index
Sentiment	1–5	— (near-constant; variance filter)
Overall signal strength	0–5	— (excluded by design; avoids double-counting the discovery signal)
<i>Qualification booleans (4)</i>		
Budget	bool	bant_count
Authority	bool	bant_count
Need	bool	bant_count
Timeline	bool	bant_count
<i>Behaviour counts (3)</i>		
n_explicit_problems	count	n_explicit_problems
n_buyer_constraints	count	n_buyer_constraints
n_rep_discovery_questions	count	n_rep_discovery_questions

resolution that a compressed integer scale discards, so a “confident 4” and a “barely 4” receive distinct values [32]. The integer and all-14 representations are kept as parallel arms, which test whether including the raw features improves or degrades performance at the current sample size [87]. Table 4.6 maps each raw output to its engineered feature, Table 4.7 specifies the full model input, and the per-feature dictionary appears in Appendix A.

4.4 Model engineering and analysis (Phases 3-4)

The empirical design is a two-model ablation on identical data folds. Model A uses only the CRM features; Model B adds the engineered transcript features to them. With the folds, preprocessing steps, and classifier held constant, any difference between the two models is attributable to the feature sets [88]. The change in PR-AUC from Model A to Model B addresses RQ1. Both models run on the same 482 closed deals (34 won); the 36 deals whose introductory call could not be analysed are retained with `nlp_analysed = false`, so that no rows are lost and both arms score identical deals. The measurement checks that answer RQ2 use only the 446 deals with an analysable call (Section 5.2).

Classifier selection. Four model families were evaluated on the CRM-only baseline (Table 4.8). Since their confidence intervals overlap at 34 positive cases, the final choice

Table 4.7: Final feature set entering the models, and the groups excluded by design.

Group	Examples	Type	In model
CRM deal	deal amount (log), region/team, acquisition channel	numeric, categorical	A and B
CRM firmographic	company size, revenue, age	numeric	A and B
Transcript: discovery	discovery index, pain depth, impact/urgency (smoothed 0–5)	ordinal	B only
Transcript: qualification	count of BANT topics covered	count	B only
Transcript: behaviour	three methodology counts; “analysed” flag	count, bool	B only
Excluded: post-close	lost reason, contract terms, win-probability field	—	leakage
Excluded: engagement	email/task counters	—	asymmetric rep-logging
Excluded: production scores	SPICED, MEDDIC, gap scores	—	heuristic-derived

rests on methodological suitability. Peak performance cannot separate the models here. The selected model is a gradient-boosted tree, a strong performer on tabular data [31, 18]. It supports the exact tree SHAP explainer [49] that the explanation layer requires, whereas non-tree models would require the slower, approximate kernel explainer. Logistic regression is included as an interpretable linear baseline, with a random forest and a histogram-based variant as robustness checks. All four are evaluated on identical folds. Hyperparameters remain at library defaults (100 trees, depth 3, learning rate 0.1) to avoid overfitting given the limited number of positive cases [87]. Since this screen uses a single stratified 5-fold split rather than the repeated protocol used for reporting, its PR-AUC values sit on a different scale from the ablation’s: the same histogram-boosting configuration reads 0.286 here and 0.246 under the reporting design (Table 5.2), so every claim about the transcript features rests on the reporting design alone. Library defaults also mean the procedure needs no inner selection loop or nested cross-validation, which at 34 positives would fit fold noise [13]. The classifier was necessarily chosen on the same 482 deals the ablation later measures; because that choice rested on explainer compatibility rather than PR-AUC, and both arms run the same classifier, any resulting optimism applies equally to Model A and Model B and cancels in the paired difference that answers RQ1.

Imbalance and validation design. Given a class imbalance ratio of 13.2:1, class weighting serves as the primary mitigation strategy [64]. SMOTE [16, 9] is included solely as a comparison arm, since interpolating between 34 sparse positives increased fold-to-fold variance without improving the central estimate (random forest SD ± 0.103 to ± 0.141 ; histogram boosting ± 0.161 to ± 0.175). Reporting both approaches addresses

Table 4.8: Model exploration on the CRM-only baseline, over the 482 modelling deals (34 won); stratified 5-fold CV, mean $\pm$ sd. All intervals overlap, so explainer compatibility decides the deployed model. Peak PR-AUC cannot separate them.

Model	PR-AUC	F1	ROC-AUC	Exact explainer
Hist. gradient boosting (SMOTE)	0.297 ± 0.175	0.195	0.741	flaky
Hist. gradient boosting (CW)	0.286 ± 0.161	0.245	0.783	flaky
Voting (LogReg + RF + HGB)	0.264 ± 0.156	0.260	0.748	no (kernel only)
Random forest (CW)	0.259 ± 0.103	0.081	0.703	yes
Voting (LogReg + RF)	0.246 ± 0.093	0.250	0.722	no (kernel only)
Logistic regression (CW)	0.213 ± 0.086	0.233	0.707	yes (linear)
No-skill baseline	0.071	—	—	—

the class-imbalance gap identified in the SLR. Evaluation uses repeated stratified cross-validation with five folds and five seeds, resulting in 25 paired A-versus-B estimates, stratified jointly by outcome and team. Results are presented as intervals and statistical tests [96].

Cross-validation design. The models are trained and tested entirely by repeated stratified cross-validation, with no deals held back as a separate test set. Each split divides the 482 deals into five stratified parts of roughly equal size: four train the model (about 386 deals, 27 wins) and one tests it (about 96 deals, 7 wins). The parts rotate, so every deal is tested exactly once per split by a model trained on the other four. This is what *out-of-fold* means throughout the thesis: a deal’s score and its SHAP explanation come from the fold model that left it out. A 20% held-out test set would leave only about seven wins on which to measure PR-AUC, far too few to separate Model A from Model B [87, 74]; repeated cross-validation, which reuses every deal for both roles, is therefore the appropriate design when the positive class is the binding constraint [13]. The five-fold split is repeated under five seeds, giving 25 paired A-versus-B estimates that smooth split-to-split variation. Since these estimates share training data, the paired-*t* tests of Section 5.1 apply the Nadeau–Bengio correction [59]. Stratification is joint on outcome and team, since region is confounded with outcome (Table 4.2), which spreads NL’s eight wins across the folds.

The preparation steps that learn from the data, median imputation of missing numeric fields, one-hot encoding, and the log transforms, are each fitted on the training portion of a fold and applied unchanged to its test portion. Fitting them on all 482 deals first would carry test-portion information into the model through the imputed values and the encoder’s category list, a label-free but real form of leakage [38, 37]. The class weights are derived per training fold for the same reason.

Analysis and Evaluation Assessment follows the FEDS framework [89]. Technical performance addresses RQ1, utilising PR-AUC relative to the 0.071 no-skill anchor, as well as significance, permutation, and additivity tests [78]. Interpretability then explains that result, using SHAP attributions [50], which are cross-validated with model-

agnostic permutation importance [26, 2] and corroborated by LIME [73], and checks that the drivers are credible rather than artefacts. Measurement validity addresses RQ2, using internal reliability and construct coherence, with model self-confidence reported separately as a diagnostic.

The experiments that follow, the classifier screen, the Model A→Model B ablation, its significance, permutation, and additivity tests, the rare-event and regional robustness refits, the pre-registered variants, and the interpretability and measurement-validity checks, are each reported in Chapter 5 with the population and design they use. Table 5.1, at the head of that chapter, lists them together.

4.5 Deployment and monitoring (Phases 5 and 6)

Deployment here is a demo artefact for the stakeholders. As a pilot, the artefact is not integrated live into the CRM. The pipeline’s outputs are assembled into a dashboard that presents the same scored pipeline separately to each of the four stakeholder groups of Table 3.1, together with a model card that documents the system following the reporting format of Mitchell et al. (2019) [56]. Each view is scoped to its stakeholder’s responsibility and decision: follow-up triage for account executives, handoff quality for business development representatives, allocation for sales leadership, and model diagnosis for revenue operations. The demonstration scores all 482 deals *out-of-fold*: the cross-validation folds of Section 4.4 are reused, so every deal is ranked, and its SHAP explanation computed, by the fold model that never saw that deal during training. The dashboard therefore shows the artefact behaving as it would on unseen deals. Buyer identities are replaced by fictional pseudonyms, deal values appear only as bands, and per-deal outcomes are hidden by design, so the demonstration can be shown to stakeholders without exposing any deal or its result. In operation, the queue would hold the open deals that are the actual target of the deployment. The artefact also carries its own acceptance measurement: feedback controls on every deal open the usage survey of Appendix D.1. Scores are presented as percentile ranks within the queue, grouped into four priority tiers (top priority at or above the 90th percentile, high at the 70th, medium at the 40th, low below). Displaying them as probabilities is avoided because the class-weighted model’s probabilities are miscalibrated [60]. Table 4.9 maps each view to its stakeholder and to the design commitment it demonstrates.

4.5.1 The lead card

The account-executive view pairs the ranked queue with a lead card per deal (Figure 4.5). The card states the deal’s tier and percentile, then gives three to five reasons in plain

Table 4.9: The five dashboard views, their stakeholders, and what each shows.

View	Stakeholder	What it shows
Lead queue and lead card	Account executives	Percentile rank and tier per deal; plain-language, checkable reasons; a CRM-only flag when no call was analysable
Handoff quality	Business development representatives	Whether the booked meeting turned into real discovery: qualification (BANT) coverage, discovery quality, and buyer signals from the sourced call; marked conditional on the pending human-rater validation
Pipeline overview	Sales leadership	Tier mix, aggregate deal value at stake per tier, and win rate by region with sample sizes printed beside every rate; the segment-fairness and feedback-loop warnings
Model and data health	Revenue operations	The monitoring phase made visible: published metrics against the no-skill anchor, global SHAP drivers computed out-of-fold, the leakage screen, the build configuration, and the monitoring plan
Model card and ledger	All	Intended and out-of-scope uses, training data, limitations; a ledger separating demonstrated claims from pending ones, including user acceptance

language. Each reason names a checkable fact about the deal together with the direction of its effect, for example “deal value in the €10–25k band raises this score”. One-hot encoded categories are summed back to their parent feature before ranking the drivers, so a reason concerns the deal’s actual attribute. SHAP magnitudes are deliberately absent from the card face: the numbers add no meaning for a representative, and every claim that remains can be verified against the CRM record. Below the reasons, a single line summarises the conversation evidence: the discovery-quality score and the qualification topics addressed. The 36 deals without an analysable introductory call are not scored silently; their cards state that the score rests on CRM data only. Each card closes with two controls: a feedback button that copies the deal’s pseudonymous reference for the usage micro-survey (Appendix D.1), and a disagree button. Five textual cards spanning the score range (two high, one mid, two low) remain the stimulus set for the prepared user study (Appendix D).

4.5.2 The handoff-quality view

The business-development view answers a single question about a booked meeting: whether it turned into real discovery (Figure 4.6). For the deal it reports BANT qualification coverage, the discovery-quality and pain-surfaced scores, and the buyer signals drawn from the sourced call. Since these quantities rest on the transcript

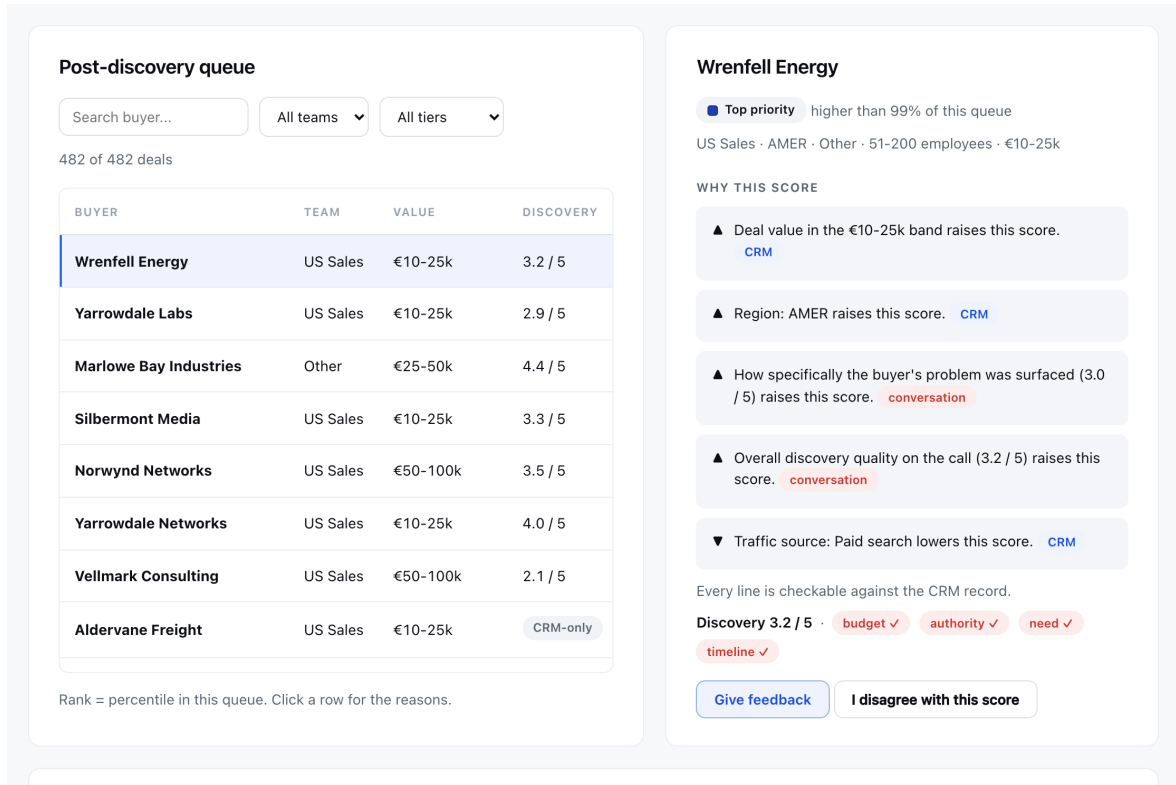


Figure 4.5: The account-executive view: the ranked post-discovery queue (left) and the lead card for the selected deal (right), with checkable reasons, the discovery summary, and the feedback and disagree controls.

instrument, the view is marked conditional on the pending human-rater validation of Section 5.

4.5.3 Designing for trust

The review showed that visible explanations can inflate trust regardless of their validity [90], so the demonstration is designed to earn trust only where the evidence supports it. It dampens trust where the evidence does not. Four mechanisms implement this directly, and two further interface statements, described afterwards, reinforce them.

- **Verifiable reasons.** Every card reason is a falsifiable statement about the deal, so a representative can verify the explanation for themselves.
- **Visible uncertainty.** Each view carries a provenance line (population, snapshot window, fold recipe, pinned extraction weights), and the interface states that deals with similar ranks should be read as ties.
- **Human agency.** The score is framed as decision support that the representative owns and may override; a disagree control records the mismatch as a monitoring signal to learn from [27].

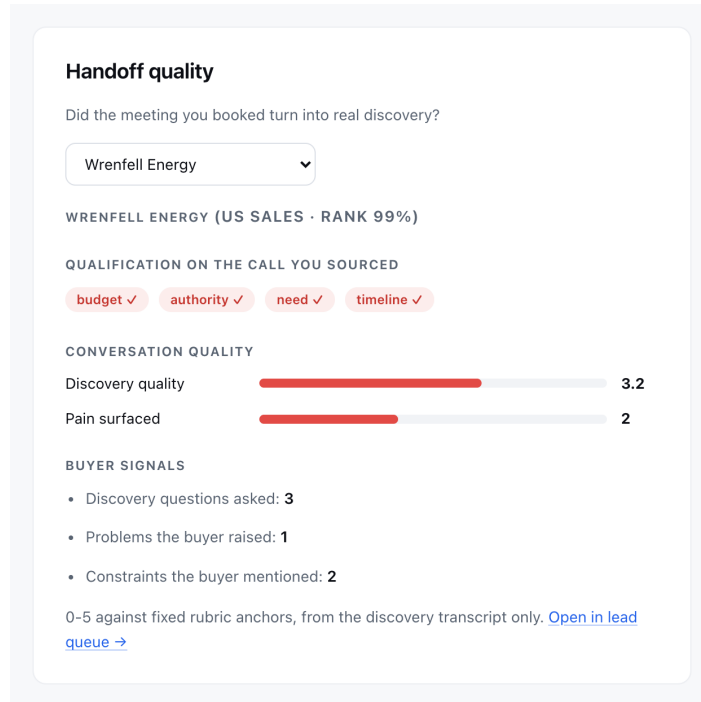


Figure 4.6: The business-development handoff-quality card: BANT qualification coverage, conversation-quality bars, and buyer signals for the booked meeting, scored from the sourced call transcript.

- **Measured acceptance.** Whether users accept the artefact is treated as an open empirical question: a six-item acceptance survey (Appendix D.1) is reachable from every deal, and the ledger lists user acceptance as pending until responses exist.

The interface also states its operational scope plainly. What is demonstrated is the hybrid thesis artefact (Model B); the diagnosis view notes that an operational rollout would ship the CRM-only Model A until the transcript increment is established, consistent with the recommendations of Chapter 6. The same view shows the global SHAP drivers computed out-of-fold, paired with the caution that attribution is not predictive importance (Chapter 5).

4.5.4 Monitoring

If the artefact were operationalised, the CRISP-ML(Q) monitoring phase [82] prescribes: quarterly re-evaluation of ranking quality on newly closed deals, read against the published intervals; drift checks on the feature distributions and on the tier mix per region, the segment-fairness watch of the leadership view (Figure 4.7) [28]; re-extraction pinned to the same model weights, temperature, and seed, with any change to the extraction model re-triggering the measurement-validity checks of RQ2; and a monthly review of the collected disagree-notes and the acceptance survey responses (Appendix D.1) as trust, acceptance, and drift signals. The two prepared studies, human inter-rater validation of the instrument and the end-user evaluation of the explanations,

precede any wider rollout.



Figure 4.7: The sales-leadership view: priority-tier mix and active-selling workload by region (top), and aggregate deal value at stake per priority tier (bottom).

The artefact is now complete and demonstrable. Whether it works is a separate question, and the next chapter answers it against the three evaluation steps fixed in Section 4.4.

Chapter 5 Evaluation

This chapter evaluates the artefact. The evaluation follows the three steps of the FEDS framework [89]: technical performance, interpretability, and measurement validity. Each step is reported in its own section, with the evidence it rests on. Section 5.3 then draws the three together. The chapter closes by discussing what the three steps contribute to each research question and by stating the limitations of the study [6].

Table 5.1 lists every experiment reported in this chapter, the population it was computed on, its number of won deals, and its design, so that each result below can be read together with the data behind it.

5.1 Predictive contribution of the transcript features

The evaluation compares two models on identical data. Model A uses the CRM features alone; Model B adds the engineered transcript features to exactly the same folds, preprocessing, and classifier, so any difference between them is attributable to the transcript features and to nothing else [88]. Both run on the full modelling population of 482 closed deals, of which 34 were won (Table 4.1), under repeated stratified cross-validation with five folds and five seeds, stratified jointly by outcome and team. This yields 25 paired estimates per model family. No deal is ever scored by a model that saw it during training, so every figure reported below is out-of-fold. The primary metric is PR-AUC, read against the no-skill baseline of 0.071 set by the population win rate; accuracy is never reported (Section 2.4.1). Table 5.1 records the population and design behind each result in this chapter.

Adding the transcript features raised PR-AUC by a small margin in every model family. For the deployed gradient-boosted tree the score rose from 0.244 to 0.267, against a no-skill baseline of 0.071, the positive rate a random ranker would reach [78] (Table 5.2). The question is whether this rise reflects a genuine contribution from the conversation, or only sampling noise. To test this, three increasingly demanding tests were applied, and the gain passed none of them (Figure 5.1).

Test 1: Significance. The first test asks whether the gain is larger than the run-to-run noise of the experiment. A genuine effect would produce a confidence interval sitting clearly above zero. Instead, once the correlation between repeated cross-validation folds

Table 5.1: Every experiment reported in this thesis, the population it was computed on, and its design. Positive counts are given because the number of won deals, not the number of deals, is the binding constraint throughout.

Experiment	Population	n (won)	Design	Reported
Classifier exploration	CRM features only	482 (34)	Single stratified 5-fold; screening only	Table 4.8
A→B ablation	CRM vs CRM + transcript	482 (34)	Repeated stratified 5×5; 25 paired estimates	Table 5.2
Test 1: significance	As ablation	482 (34)	Nadeau–Bengio corrected paired t ; Bayesian region of practical equivalence	Table 5.3
Test 2: permutation	As ablation	482 (34)	500 re-runs, transcript values shuffled across deals; folds, imbalance, and CRM features held fixed	Section 5.1
Test 3: additivity	Deal value only, vs deal value + transcript	482 (34)	As ablation	Table 5.3
Rare-event refit	As ablation	482 (34)	King–Zeng penalised logit [41]	Section 5.1
Regional subgroup	US (AMER) deals only	208 (20)	Subset of the ablation folds; post hoc	Section 5.1
Pre-registered variants	All 14 raw outputs; integer vs smoothed	482 (34)	As ablation	Section 5.1
SHAP and permutation importance	As ablation	482 (34)	Out-of-fold attributions and importances	Table 5.4
LIME cross-check	Demonstration leads	5	Top-five driver overlap with SHAP; qualitative	Section 5.1.1
Measurement validity	Analysed transcripts only	446	Instrument outputs; no model fitted	Table 5.5
Demonstration	All deals	482 (34)	Out-of-fold scoring on the ablation folds	Section 4.5

Table 5.2: Model A→Model B ablation (repeated 5×5-fold CV; no-skill PR-AUC = 0.071). Point estimates positive; every interval spans zero.

Model	A: PR-AUC	B: PR-AUC	$\Delta(B-A)$	paired- t p
Logistic regression	0.229	0.235	+0.005 [−0.051, +0.061]	0.86
Random forest	0.234	0.245	+0.010 [−0.108, +0.129]	0.87
Hist. gradient boosting	0.246	0.264	+0.018 [−0.076, +0.111]	0.72
Gradient boosting (deployed)	0.244	0.267	+0.023 [−0.065, +0.110]	0.62

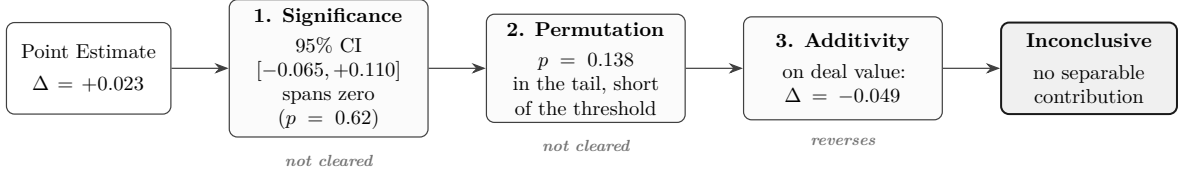


Figure 5.1: The three escalating tests applied to the observed gain, and the verdict they lead to.

is corrected for [59], every model’s 95% confidence interval spans zero, and no corrected paired- t test comes close to significance (Table 5.2). A Bayesian analysis agrees. Once a small band around zero is treated as no practical difference [7], the analysis assigns only weak probability to the hybrid model being the better of the two. It cannot rule out the two being equivalent (Table 5.3).

Test 2: Permutation. The second test asks how large a gain this experiment would produce if the transcript features carried no information at all. It is not a separate study: the same A-versus-B comparison is re-run 500 times, and in each run the transcript feature values are randomly reassigned to different deals [63]. Shuffling destroys any real link between a deal’s conversation and its outcome while leaving the sample size, the class imbalance, the folds, and the CRM features untouched. Each run therefore produces the gain that this exact experiment yields from noise alone, and 500 runs give the full distribution of such gains [19]. The gain actually observed sits in the upper tail of that distribution but does not clear it ($p = 0.138$ for the deployed tree; $p = 0.218$ for logistic regression), meaning a gain this large would arise by chance roughly one time in seven. It cannot be distinguished from a lucky draw.

Test 3: Additivity. The third test is the strictest, because it asks whether the transcript features add anything the CRM does not already supply. Starting from a model that uses deal value on its own, adding the transcript features changes PR-AUC by essentially nothing for logistic regression (≈ 0) and slightly lowers it for the tree models (-0.024 to -0.058) [87]. Removing deal value from the full model gives the same result in reverse (Table 5.3) [19]. Once deal value is present, the transcript features supply no further predictive value. This is the decisive test of the three: unlike Tests 1 and 2, which a small positive class can fail through lack of power alone, the additivity test speaks to structure. It holds even if the observed gain is taken at face value. The transcript block adds nothing separable once deal value is in the model.

Table 5.3: The three tests for the deployed model: the full-model gain is not significant, reverses without deal value, and is not additive on top of it.

Comparison	Δ PR-AUC	95% CI	p	P(better/equiv/worse)
Full CRM + NLP vs CRM	+0.023	$[-0.065, +0.110]$	0.62	61 / 15 / 24%
NLP gain, deal value removed	-0.044	$[-0.154, +0.065]$	0.43	17 / 10 / 73%
NLP added to deal-value-only	-0.049	$[-0.164, +0.066]$	0.41	16 / 9 / 75%

Two further observations help frame the result. First, when the models are refitted with a linear method built for rare-outcome data [41], the transcript gain vanishes ($\Delta = +0.001$). This suggests that whatever signal the features carry works through non-linear interactions [2]. Second, the small gain is concentrated in the US region, where PR-AUC rises from 0.252 to 0.304. Since this pattern surfaced only after the main analysis, it is treated as a lead for future work.

Two pre-registered design checks [61] show that the feature engineering shaped this result very little. Both were fixed before the final extracted data were available, and both ran on the same 482 deals and the same folds as the main ablation. The variant that fed in all fourteen raw transcript outputs scored lower in every model family, so extra features dilute the signal at this sample size [87]. The smoothed features also scored above their integer versions, as anticipated. The result therefore follows from the data.

5.1.1 Why the result came out this way

What drives the predictions. SHAP attributes each prediction to the features behind it [50], and, averaged over all deals, this reveals what the model leans on most [8]. Deal value is the single largest driver, with the firmographic features (company size, revenue, and age) next [80]. The transcript features together account for about 18.2% of this attribution, led by pain depth and by impact and urgency. Structured deal facts come first and pain depth is the strongest conversational signal, exactly the ordering that sales-conversation theory would predict [70] (Figure 5.2, Table 5.4).

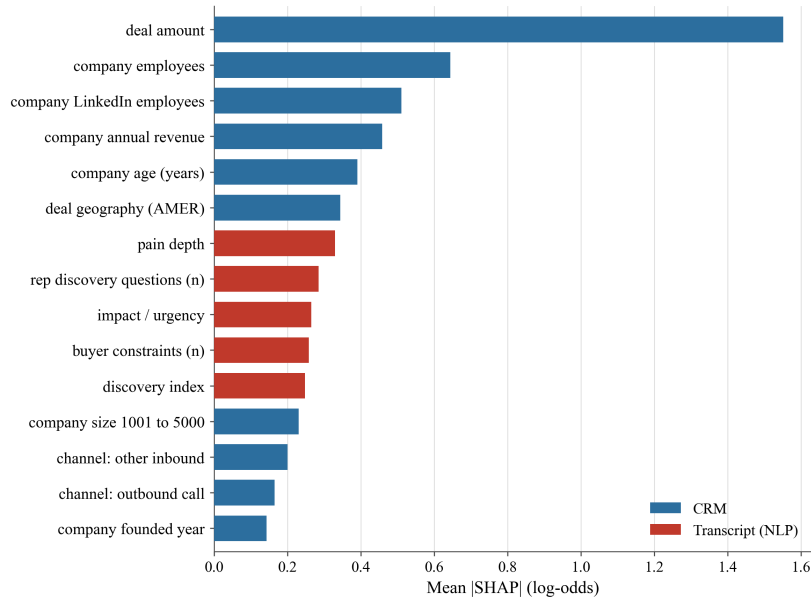


Figure 5.2: Global feature importance (mean |SHAP|) for the deployed Model B. Deal value dominates; the transcript features sit below the CRM features.

Table 5.4: Top drivers, out-of-fold: attribution (mean $|\text{SHAP}|$) versus predictive importance (permutation $\Delta\text{PR-AUC}$). The transcript (NLP) features hold visible attribution but little separable predictive importance once the CRM features are present.

Feature	Group	Mean $ \text{SHAP} $	Perm. $\Delta\text{PR-AUC}$
deal_amount	CRM	1.56	+0.143
company_employees	CRM	0.58	+0.037
company_annual_revenue	CRM	0.53	+0.018
company_linkedin_employees	CRM	0.50	+0.037
company_age_years	CRM	0.42	+0.033
deal_geography (AMER)	CRM	0.38	+0.046
impact/urgency	NLP	0.36	+0.003
pain_depth	NLP	0.36	+0.058

Are the drivers credible. A feature can look important in a SHAP plot without actually driving the outcome, so each top driver was tested against a second, independent measure: permutation importance. This shuffles a feature’s values and records how much the model’s ranking worsens on data it did not train on [10, 26, 19]. For deal value the two measures agree. It ranks first for attribution in all five cross-validation folds, and its permutation importance is several times that of any other feature. A fresh leakage audit further confirms it is a legitimate field that is known at the decision point. Its dominance is therefore real [38]. This also resolves the earlier puzzle that won and lost deals had almost identical average deal values (Figure 4.2). The model exploits non-linear structure in deal value that a simple comparison of averages cannot show, though the specific interacting covariates are not identified here [50].

The transcript features behave differently. Their attribution is stable, holding about 18% across folds with pain depth on top. Their predictive importance, though, is uneven: impact and urgency is genuinely near zero (+0.003), while pain depth registers a modest single-feature permutation importance (+0.058; Table 5.4). That single-feature figure is measured on the fitted model, which already encodes pain depth’s correlation with deal value; the additivity test refits without the transcript block and shows it adds no separable value once deal value is present (−0.049). The model therefore uses the transcript features to explain its predictions far more than to make them [19]. This is the mechanism behind the RQ1 result (Figure 5.3): the conversational signal is visible in the explanations while adding no separable predictive value.

The wider lesson is that a feature’s share of the SHAP attribution must not be read as its predictive contribution [57, 10, 26]. Read that way, the null found here would have looked like a positive result, so a SHAP summary plot should always be paired with an independent importance check. The LIME cross-check [73] makes the same point for individual predictions. It is a qualitative check on five deals and is read as such. Across the five demonstration leads, which span the score range, LIME and

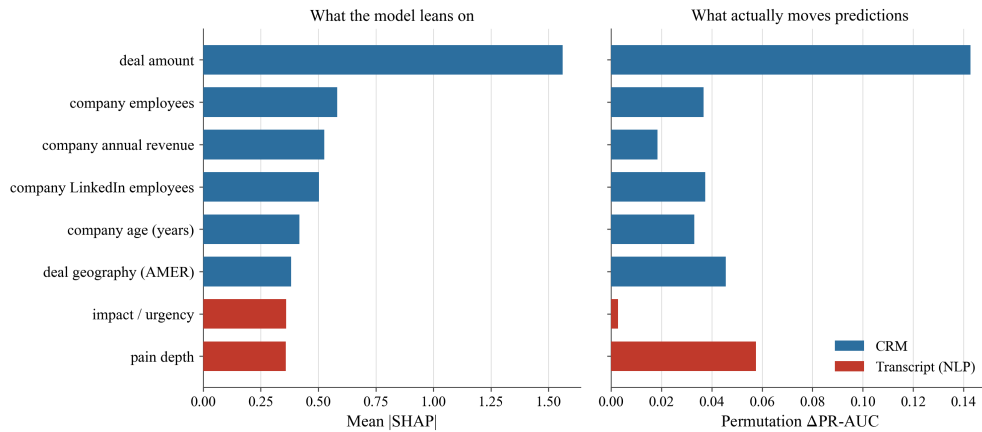


Figure 5.3: Left: how much the model leans on each feature (mean |SHAP|). Right: how much each feature moves held-out predictions (permutation Δ PR-AUC). The transcript features carry visible attribution but little predictive effect, pain depth excepted (+0.058).

SHAP agreed on an average of 2.6 of their top five drivers, and LIME surfaced fewer transcript features, which fits their low predictive effect [76, 19]. This check speaks to the stability of the explanations, not to their predictive value.

5.2 Transcript features as measurements

This section asks whether the transcript features are sound measurements in their own right, whatever they predict. It addresses the measurement-validity gap identified in the review, since a feature produced by a language model cannot be trusted simply because it exists [85, 32, 29, 17]. Three checks were run on the 446 analysed deals, all of them on the model’s own outputs (Table 5.5).

Internal reliability. This check asks whether the discovery dimensions cohere as one underlying construct. Cronbach’s α is 0.89 over the five discovery dimensions (opening and alignment, situation understanding, pain depth, impact and urgency, and next steps; sentiment and the holistic overall-signal item excluded), the standard index of internal consistency [83, 1]. This is high enough for the scale to be internally consistent, yet below the 0.95 point at which the dimensions would be redundant. The figure needs care. The five dimensions are separate but related parts of discovery, not five measures of the same thing. Cronbach’s α therefore shows how consistently the instrument scored them; it does not prove they form one construct. A high score is also easy to produce here: the rubric was built for the dimensions to agree, one model rated all of them in a single pass, and the same model gave a holistic overall-signal item. Each of these can push α up through a halo effect. A high α is therefore consistent with either genuine reliability or one global impression smeared across the dimensions. It is reported here as internal consistency. It is not evidence that the dimensions are independently valid [32].

Model self-confidence. This diagnostic asks whether the model commits to the score it reports; it bears on neither reliability nor validity in the psychometric sense, and is reported separately from the validity evidence. When it outputs an integer from 0 to 5, it also holds an internal probability spread across those values. If the reported score sits close to the centre of that spread, the model is being decisive [32]. Measured as a rank correlation, the agreement is strong for six of the seven dimensions (0.77 to 0.91) [17]. Sentiment is the exception (0.48) and is flagged as such. A weak result there is unsurprising, because most B2B calls are flat in tone, and the variance filter had already dropped sentiment for that reason.

Construct coherence. This check asks whether the scores relate to one another in the ways theory predicts, and they do, at the moderate strengths theory would expect [70]. Since both the score and the count it is correlated with come from the same model reading the same transcript, this is suggestive convergent evidence subject to shared-method variance. It falls short of independent confirmation. Pain depth rises with the number of explicit problems the buyer states (rank correlation +0.49), and overall discovery quality rises with the number of open questions the representative asks (+0.55). The four BANT qualification flags [45] show moderate internal consistency ($KR-20 = 0.667$, the binary-item counterpart of Cronbach’s α). This is the level appropriate to four related but distinct aspects of qualification.

Table 5.5: Model Measurement Checks

Check	Statistic	Value	Reading
Internal reliability (5 discovery dims)	Cronbach α	0.89	Reliable at ≥ 0.70 and below redundancy at < 0.95
Decisiveness (6 of 7 dimensions)	concordance ρ	0.77–0.91	Strong; the model commits to its scores
Decisiveness (sentiment)	ρ	0.48	Weak; the flagged dimension
Construct: pain depth vs explicit problems	ρ	+0.49	Moderate, as theory predicts
Construct: discovery vs open questions	ρ	+0.55	Moderate, as theory predicts
Qualification booleans (BANT)	KR-20	0.667	Moderate; four distinct facets

Figure 5.4 places these results in a progression of what has been shown and what remains to be shown. All three checks run on the model’s own outputs, so they can establish that the instrument is internally consistent and structurally coherent, but not that it is correct [65, 17]. None of them would catch a score that is confident but systematically wrong, such as one that consistently rates pain depth a point too high [34]. Catching that requires comparing the model’s scores against human raters

[17, 34], using the intraclass correlation coefficient and weighted κ measured against the thresholds of Koo and Li (2016) [43]. That comparison is fully specified on a prepared 48-transcript sample and is left as future work.

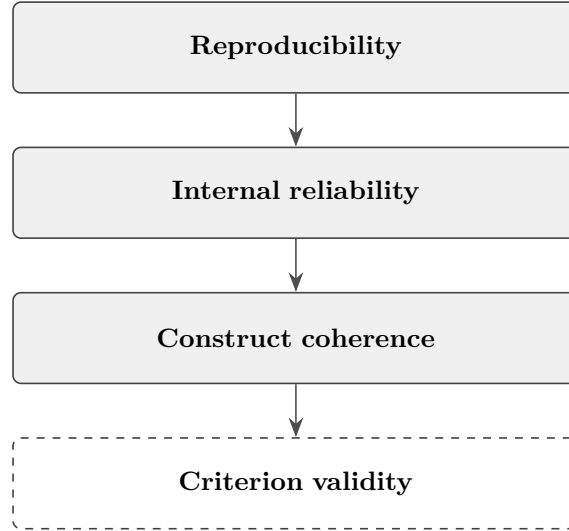


Figure 5.4: The measurement-validity progression: the three checks this thesis establishes on the instrument’s own outputs, and the criterion-validity step that requires human raters and is left to future work.

5.3 Evaluation summary

Table 5.6 presents a summary of the evaluation findings. The technical step is inconclusive, the model’s reasoning is credible, and the measurement instrument is self-consistent within the bounds stated above.

The FEDS framework classifies an evaluation on two axes [89]. The first is what is evaluated against: an *artificial* evaluation uses constructed or laboratory conditions, while a *naturalistic* one uses the real setting, real data, and real users. The second is when it is done and why: a *formative* evaluation feeds back into the design of the artefact, while a *summative* one judges the finished artefact. The technical and interpretability steps are naturalistic and summative, because both run on the host’s own closed deals and judge the completed pipeline. The measurement-validity step is artificial and formative, because it examines the instrument’s own outputs rather than its agreement with the world, and its findings shaped the rubric.

5.4 Discussion

Taken together, the results answer the research questions as follows.

Table 5.6: The three FEDS evaluation steps, their primary evidence, and the verdict for each.

Evaluation step	Primary evidence	FEDS character	Verdict
Technical	A→B ablation under three escalating tests	Naturalistic, summative	Inconclusive
Interpretability	SHAP global and local; permutation-importance and LIME cross-checks	Naturalistic, summative	Credible
Measurement validity	Reliability, decisiveness, construct coherence	Artificial, formative	Self-consistent

RQ1. No reliable improvement from the transcript features could be established. The point estimates were small and positive in all four model families, but none survived the significance, permutation, or additivity tests. The verdict is that the contribution is unconfirmed: the evidence is inconclusive and leans negative. A flat claim of no effect would itself overclaim, because 34 won deals are too few to rule out a small true effect [74]. What the evidence does rule out is a large effect, or one that works independently of the CRM signal.

Analysing the model explains why. Deal value dominates the predictions, the firmographic features follow, and the transcript features hold about 18% of the attribution, led by pain depth and by impact and urgency. Deal value’s dominance is credible: it ranks first in every held-out fold, leads on permutation importance by a wide margin, and survives a fresh leakage audit [38]. It looked uninformative when won and lost deals were compared on their average value, yet it drives the model through non-linear interactions [50]. The transcript features provide the contrast: their attribution is stable, but they add no separable predictive value once the CRM features are present. Only pain depth shows a modest single-feature importance, and even that does not survive the additivity test [19, 20]. Had that attribution been read as importance, this null would have been reported as a false positive [19]. The drivers are therefore credible, and the conversational signal is visible in the explanations while adding no separable predictive value.

RQ2. Within the stated bounds, the transcript features are internally consistent ($\alpha = 0.89$) and structurally coherent with theory, so they are not arbitrary output from the language model. This establishes internal consistency. Validity remains unestablished: none of the checks compares the scores against an external criterion, and agreement with human judgement remains to be established and is left to future work [34, 32].

Two readings of the RQ1 result remain open, and with the present data we cannot decide between them. Under reading (A), discovery quality in the introductory call genuinely carries little signal about winning or losing beyond what the CRM already records. Under reading (B), the signal exists, but a single call scored once against a

rubric is too coarse an instrument to surface it. Committing to either reading would overclaim at 34 won deals, and the future-work agenda is designed to tell them apart [87]. Table 5.7 sets out how the thesis addresses each gap the review identified. The thesis adds the NLP feature-extraction layer and the measurement-validity evaluation that a study using only structured data never has to confront [85].

A further contribution is specific to the RITAM setting. The systematic review was conducted on a single database and found no study drawn from the RITAM industry itself, so its conclusions reached this setting only by inference from adjacent B2B contexts. There was no direct evidence. This thesis supplies that missing evidence: in the reviewed literature, it is one of the early studies covering a lead-scoring pipeline built and evaluated on a RITAM vendor’s own CRM and introductory-call data [93]. This is a claim about the data source. It does not treat RITAM as a distinct buying vertical.

The result is specific and useful even though it is not the improvement that was hoped for. In this setting the predictive weight sits with the structured CRM data, deal value and firmographics above all, while discovery quality in the introductory call does not separably improve the ranking. For a RITAM vendor deciding where to spend its scoring effort, a result shown directly on comparable data is more actionable than one inferred from unrelated industries. The recommendations that follow turn it into concrete guidance [93, 33].

Table 5.7: How each SLR gap is addressed, and with what status.

SLR gap	How the thesis engages it	Status
Transcript gap	First study (in the reviewed evidence) to combine RITAM B2B transcript features with CRM data and quantify the lift; the answer is inconclusive	Addressed; finding inconclusive
Measurement validity	Internal reliability, decisiveness, construct coherence; human ICC study specified	Internal consistency shown
Class imbalance	PR-AUC vs no-skill anchor, stratified repeated CV, class weighting vs SMOTE	Closed methodologically
User evaluation	Protocol, questionnaire, prototype cards prepared (Appendix D)	Future work (primary direction)
RITAM transfer only inferred (SLR limitation)	First lead-scoring pipeline built and evaluated on a RITAM vendor’s own CRM and transcript data, supplying direct evidence on the data source where the review could only infer	Addressed; direct evidence produced, inconclusive on transcripts

5.5 Limitations

Table 5.8 pairs each limitation with its consequence and with the claim it bounds. The decisive limitation is statistical power. At 34 won deals the ratio of events per variable (won deals per model input) is roughly 1.4, well below the recommended range of 10 to 20 [74]. The study was therefore underpowered for this ablation from the outset [87]. The binding constraint is the number of won deals, and that is one no modelling choice can lift [9].

Table 5.8: Limitations.

Limitation	Consequence	What it bounds
Small positive class (34 won)	EPV ≈ 1.4 ; wide intervals	RQ1 is “not detected,” not “proven absent”
Class imbalance (13.2:1)	Minority-class metrics are noisy	Why PR-AUC against a no-skill anchor, never accuracy [78]
Single vendor; apparent effect concentrates in US	Limited external validity	Transfer is inferred and not yet demonstrated
Intro-call-only scope	Misses multi-call and trajectory signal	The window may be too early or too coarse
Machine-only instrument validation	Convergent/criterion validity not established	RQ2 bounded to internal consistency; sentiment weakest
No end-user evaluation	Explanation value asserted by design	Trust and adoption untested; primary future-work direction

Chapter 6 Conclusions

This chapter concludes the thesis. It summarises the findings, states the contributions, and offers recommendations for the host organisation and directions for future research.

6.1 Summary of findings

This thesis set out to design and evaluate an explainable lead-scoring pipeline that combines CRM data with features extracted from introductory sales-call transcripts, and to test whether those features improve predictive performance. The main finding is:

The answer to the main research question is negative, with careful qualification. At this sample size and in this setting, no distinct predictive contribution from the transcript features could be identified [87]. The CRM data, deal value in particular, carried most of the predictive power [80]. For now, NLP features extracted from introductory calls do not contribute to lead-scoring performance beyond what CRM data alone provides.

Other findings, which together support and add context to this result, are:

1. **The small apparent gain does not hold up statistically.** Any increase in predictive performance from the transcript features disappeared under significance, permutation, and additivity tests; the additivity test is the decisive one, since it bounds the transcript block as adding nothing separable once deal value is present. That is a conclusion that lack of statistical power alone would not produce [2, 19].
2. **The features remain useful as an interpretable coaching signal, conditional on validation.** They add to the interpretability of the scores while leaving predictive performance unchanged, so their value is as a coaching signal rather than a predictive one, and only once validated against human raters [19, 20].
3. **The features are internally reliable.** The instrument has strong internal consistency ($\alpha = 0.89$) and theoretical coherence [83, 1], though validation with human raters is still needed [34, 32].
4. **It supplies direct evidence on a previously unstudied data source.** Prior work could only infer from adjacent B2B; this study tests the question on a RITAM vendor's own data [93]. The claim concerns the data source, not RITAM as a distinct buying vertical.

6.2 Contributions

The contributions of this study are methodological, empirical, and practical. The methodological contributions generalise beyond this study, while the empirical and practical ones are specific to the RITAM context. To our knowledge, and within the bounds of a systematic review drawn from a single database and conducted by one reviewer, this work is one of the earliest studies that provides a direct test of transcript-augmented lead scoring on a RITAM vendor’s proprietary data, where earlier reviews could only infer from adjacent B2B markets [93].

The contributions are:

1. **Resolution of the two stated problems and achievement of the research objectives.** By the standards of DSR, a rigorously established inconclusive result is still a valuable contribution [35].
 - **Problem 1** (whether conversational signals carry predictive value beyond CRM fields) is answered directly: they do not, at this sample size.
 - **Problem 2** (the opacity of commercial tools) is resolved by design, since the artefact produces a plain-language explanation alongside every score.
 - The **research objectives** set out in Section 1.3.1 are therefore achieved: the artefact was designed, demonstrated, and evaluated, even though the predictive hypothesis it was built to test was not confirmed.
 - The artefact, the validation protocol, and the transparent evidence are the **instruments** through which these knowledge contributions were produced [6].
2. **A reproducible, privacy-preserving LLM measurement instrument and its machine-side validation protocol.** The extraction itself uses open weights, pinned versions, and schema-enforced decoding [85]. It is paired with internal-consistency, model self-confidence, and construct-coherence checks, and a ready-to-run human-rater (ICC) design [34].
3. **A leakage-disciplined, pre-registered small-sample evaluation design** [38, 87]. The design combines a fixed decision point, pre-registered feature engineering, and three escalating statistical tests, and an attribution-versus-importance safeguard so that explanation weight is not misread as predictive value [19, 20].
4. **Direct empirical evidence on a RITAM vendor’s own data, with a mechanism.** The result is a carefully bounded inconclusive finding, established with an $A \rightarrow B$ ablation, the additivity test, and interaction analysis. Where prior work could only infer from adjacent B2B [94], this contribution is specific to the RITAM setting.

5. **A demonstrated explainable hybrid pipeline and lead card.** Shown on real deals, the pipeline combines the CRM-only model with a percentile rank and a SHAP-based, plain-language rationale that a representative can inspect for any scored deal. This instance is RITAM-specific, though the underlying method transfers [27].

6.3 Recommendations for practice

1. **Rank the pipeline for the CRM signal now (well supported).**
 - Deploy the CRM-only model, which reaches 0.244 PR-AUC against the no-skill baseline of 0.071 [78, 22].
 - Do not include the transcript features in the deployable ranking system: they add no separable value.
 - Deliver scores as a percentile-ranked lead card with a SHAP rationale the representative can check [27].
 - Display ranks rather than raw probabilities, which are miscalibrated at the current class imbalance [60].
 - Watch for a feedback loop: deals the score deprioritises receive less selling effort and are then more likely to lose, which confirms the score [77].
2. **Hold off on conversational scoring for conversion prediction.** Intro-call features added no predictive value here. Test any vendor's claims against a CRM-only baseline before buying, and revisit the decision after reaching about 200 won deals, or by pooling data across several RITAM vendors [87].
3. **Treat the instrument as a candidate coaching signal, conditional on validation.** It is internally consistent ($\alpha = 0.89$) but not yet validated against human judgement, so coaching or quality-assurance use should wait on the human-rater (ICC) study [34] and the halo and discriminant checks. When it is used, separate rep-controllable dimensions (opening, questions asked, next steps) from buyer-driven ones (sentiment, pain depth), since a representative should not be coached on the buyer's behaviour.
4. **Model and monitor performance segment by segment.** Within this vendor, segment outcomes differ sharply: the US team won 9.6% (20 won) against NL's 4.5% (8 won). A global model is unreliable when segments contain few positive cases; RITAM vendors should stratify scoring by region or vertical and watch small-data segments closely.
5. **Establish data discipline before adopting advanced data-science methods [79].** Asymmetric activity logging ($\pm 20\text{pp}$ by outcome) forced the engagement-

activity group out of the model as a leakage risk [38]. Standardise logging, keep stage-derived probabilities and heuristic scores out of model inputs, and consolidate the parallel scoring layers.

6. **Run the two prepared studies before wider rollout.** The human inter-rater study and the end-user study (Appendix D) are ready to use and low-cost. Once deployed, monitor ranking quality and drift quarterly [28].

6.4 Future Work

Each direction addresses a question this study left open. They are ordered by their likely impact.

1. **A larger, ideally multi-RITAM-vendor sample (highest leverage).** Sample size matters most here. Events-per-variable guidance puts the need at 200 to 400 won deals for the current feature count [74, 87]. This seems most feasibly reached by pooling data across several RITAM vendors. It would show whether a small transcript effect exists, and move the findings from a single RITAM data point toward RITAM-wide generalisation.
2. **Per-vertical reliability within RITAM (high).** Deal volumes vary widely across RITAM segments, as the US and NL comparison shows. A model that is reliable in a data-rich segment may mislead in a data-poor one. Future research should establish how many positive cases a RITAM segment needs before its score can be trusted [93].
3. **End-user study (high).** It would assess trust in and adoption of the explanations, and close the user-evaluation gap identified in the systematic literature review. It matters even if the primary effect is null, because explanations may still help representatives reason about their pipeline, whatever their effect on ranking [19, 27]. The materials are ready (Appendix D)
4. **Multi-call and multi-modal features (high).** Adding multi-call and multi-modal features would test reading (B) of the main research question: whether a single introductory call is a fine enough window to capture the signal [20].
5. **Human inter-rater study (medium to high).** The scores are already extracted and the transcripts are available, so this is the most immediately actionable direction. It would advance RQ2 from internal reliability to criterion validity [34].

Bibliography

- [1] Abdelmageed Algamdi, Said Brika, Ines Laamari, and Khalil Chergui. Crm system and potential customer loyalty trends: Some evidence of customer engagement. *International Journal of ADVANCED AND APPLIED SCIENCES*, 8:44–52, 05 2021.
- [2] Andre Altmann, Laura Tolosi, Oliver Sander, and Thomas Lengauer. Permutation importance: A corrected feature importance measure. *Bioinformatics (Oxford, England)*, 26:1340–7, 04 2010.
- [3] Abedelkadir Asi, Song Wang, Roy Eisenstadt, Dean Geckt, Yarin Kuper, Yi Mao, and Royi Ronen. An end-to-end dialogue summarization system for sales calls. pages 45–53, 01 2022.
- [4] Teoman Berkay Ayaz, Muhammet Furkan Özara, Emrah Sezer, Ahmet Erkan Çelik, and Akhan Akbulut. Enhancing Targeting in CRM Campaigns Through Explainable AI. *Lecture Notes in Networks and Systems*, 1088:203 – 214, 2024.
- [5] S. Baehre, M. O’Dwyer, L. O’Malley, and N. Lee. The use of Net Promoter Score (NPS) to predict sales growth: insights from an empirical investigation. *Journal of the Academy of Marketing Science*, 50(1):67–84, 2022.
- [6] Richard Baskerville. FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, 2014 Forthcoming:1–13, November 2014.
- [7] Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. Time for a change: A tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18, 06 2016.
- [8] Gülşah Benk, Bertan Badur, and Sona Mardikyan. A new 360° framework to predict customer lifetime value for multi-category e-commerce companies using a multi-output deep neural network and explainable artificial intelligence. *Information*, 13:373, 08 2022.
- [9] Rok Blagus and Lara Lusa. SMOTE for High-Dimensional Class-Imbalanced Data. *BMC bioinformatics*, 14:106, March 2013.
- [10] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.
- [11] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.
- [12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners, 05 2020.

- [13] Gavin Cawley and Nicola Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 07 2010.
- [14] Pete Chapman, Julian Clinton, Randy Kerber, Tom Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0 step-by-step data mining guide*. 01 1999.
- [15] Kushal Chawla, Chenyang Zhu, Pengshan Cai, Sangwoo Cho, Scott Novotney, Ayushman Singh, Jonah Lewis, Keasha Safewright, Alf Samuel, Erin Babinsky, Shi-Xiong Zhang, and Sambit Sahu. Lessons from the field: An adaptable lifecycle approach to applied dialogue summarization, 01 2026.
- [16] Nitesh Chawla, Kevin Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 06 2002.
- [17] Junjie Chen, Yuxi Dong, Haitao Li, Su Weihang, Yujia Zhou, Zhang Min, Yiqun Liu, and Qinyao Ai. Benchmarking llm-as-a-judge for long-form output evaluation, 06 2026.
- [18] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016.
- [19] Tobias Clement, Nils Kemmerzell, Mohamed Abdelaal, and Michael Amberg. XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. *Machine Learning and Knowledge Extraction*, 5(1):78 – 108, 2023.
- [20] Kristof Coussement, Mohammad Zoynul Abedin, Mathias Kraus, Sebastián Maldonado, and Kazim Topuz. Explainable AI for enhanced decision-making. *Decision Support Systems*, 184, 2024.
- [21] Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly*, 13(3):319–340, 09 1989.
- [22] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning, ACM*, volume 06, June 2006.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019.
- [24] Verena Eitle and Peter Buxmann. Business analytics for sales pipeline management in the software industry: A machine learning perspective. 01 2019.
- [25] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data

- mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.
- [26] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of machine learning research : JMLR*, 20, 01 2019.
 - [27] Piotr Gaczek, Grzegorz Leszczyński, and Amira Mouakher. Collaboration with machines in B2B marketing: Overcoming managers’ aversion to AI-CRM with explainability. *Industrial Marketing Management*, 115:127 – 142, 2023.
 - [28] João Gama, Indrè Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Hamid Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Surveys (CSUR)*, 46, April 2014.
 - [29] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks, 03 2023.
 - [30] L. González-Flores, J. Rubiano-Moreno, and G. Sosa-Gómez. The relevance of lead prioritization: a B2B lead scoring model based on machine learning. *Frontiers in Artificial Intelligence*, 8, 2025.
 - [31] Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? pages 507–520, January 2022.
 - [32] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge, 11 2024.
 - [33] Hongyu Guo and Herna Viktor. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *SIGKDD Explorations*, 6:30–39, June 2004.
 - [34] Steve Han, Gilberto Junior, Tom Balough, and Wenfei Zhou. Judge’s verdict: A comprehensive analysis of llm judge capability through human agreement, 10 2025.
 - [35] Alan Hevner, Alan R, Salvatore March, Salvatore T, Park, Jinsoo Park, Ram, and Sudha. Design science in information systems research. *Management Information Systems Quarterly*, 28:75–, 03 2004.
 - [36] Julie Josse, Jacob Chen, Nicolas Prost, Gael Varoquaux, and Erwan Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65:5447–5479, September 2024.
 - [37] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science, 07 2022.
 - [38] Shachar Kaufman, Saharon Rosset, and Claudia Perlich. Leakage in data mining: Formulation, detection, and avoidance. volume 6, pages 556–563, 01 2011.

- [39] Muhammad Khalid and Ann-Perry Witmer. Prompt engineering for large language model-assisted inductive thematic analysis, 03 2025.
- [40] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [41] Gary King and Langche Zeng. Logistic regression in rare events data. *Journal of Statistical Software*, 8, 01 2003.
- [42] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 05 2022.
- [43] Terry Koo and Mae Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 03 2016.
- [44] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [45] M.V. Leite, J.M. Abe, and M.L. Hoffmann Souza. ParaBANT: An Adaptive Paraconsistent Model for Bias-Resistant B2B Lead Qualification and Segmentation. *Procedia Computer Science*, 270:2406–2415, 2025.
- [46] Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. pages 1–15, 04 2020.
- [47] Jin Lin. Application of machine learning in predicting consumer behavior and precision marketing. *PLoS ONE*, 20(5), 2025.
- [48] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 03 2023.
- [49] Scott Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2, January 2020.
- [50] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 12 2017.
- [51] Kashef Abdul Majid and Michel Laroche. Reading between the lines: AI and customer acquisition in professional services. *Journal of Services Marketing*, pages 1 – 13, 2025.
- [52] Gabriel Marín Díaz. A Fuzzy-XAI Framework for Customer Segmentation and Risk Detection: Integrating RFM, 2-Tuple Modeling, and Strategic Scoring. *Mathematics*, 13(13), 2025.
- [53] Gabriel Marín Díaz. A Unified Fuzzy–Explainable AI Framework (FAS-XAI) for Customer Service Value Prediction and Strategic Decision-Making. *AI (Switzer-*

land), 7(1), 2026.

- [54] Gabriel Marín Díaz, Jose Javier Galan Hernandez, and Ramón Carrasco. Xai for churn prediction in b2b models: A use case in an enterprise software company. *Mathematics*, 10:3896, 10 2022.
- [55] Gabriel Marín Díaz, Raquel Gómez Medina, and José Alberto Aijón Jiménez. A Methodological Framework for Business Decisions with Explainable AI and the Analytic Hierarchical Process. *Processes*, 13(1), 2025.
- [56] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 220–229. ACM, January 2019.
- [57] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2nd edition, 2022.
- [58] Stephen Mortensen, Michael Christison, BoChao Li, AiLun Zhu, and Rajkumar Venkatesan. Predicting and defining b2b sales success with machine learning. pages 1–5, 04 2019.
- [59] Claude Nadeau. Inference for the generalization error. *Machine Learning*, 52:239–281, 01 2003.
- [60] Alexandru Niculescu-Mizil and Rich Caruana. *Predicting good probabilities with supervised learning*. January 2005. Journal Abbreviation: ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning Pages: 632 Publication Title: ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning.
- [61] Brian Nosek, Charles Ebersole, Alexander DeHaven, and David Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115:201708274, 03 2018.
- [62] R. Nygård and J. Mezei. Automating lead scoring with machine learning: An experimental study. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2020:1439–1448, 2020.
- [63] Markus Ojala and Gemma Garriga. Permutation Tests for Studying Classifier Performance. In *The Journal of Machine Learning Research*, volume 11, pages 908–913, November 2009.
- [64] Simona Vasilica Oprea and Adela Bâra. Explaining Customer Churn Through Unified Perturbation-Based XAI Soft Computing. *IEEE Access*, 14:24460 – 24479, 2026.
- [65] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation, 05 2023.

- [66] Ken Peffers, Tuure Tuunanen, Marcus Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24:45–77, 01 2007.
- [67] Ana Victoria Ponce Bobadilla, Vanessa Schmitt, Corinna Maier, Sven Mensing, and Sven Stodtmann. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17, October 2024.
- [68] Martin Prause. The machine learning canvas: Empirical findings on why strategy matters more than ai code generation, 01 2026.
- [69] Jieli Qiu, Liangwei Yang, Ming Zhu, Wenting Zhao, Zhiwei Liu, Juntao Tan, Zixiang Chen, Roshan Ram, Akshara Prabhakar, Rithesh Murthy, Shelby Heinecke, Caiming Xiong, Silvio Savarese, and Huan Wang. Enterprise sales copilot: Enabling real-time ai support with automatic information retrieval in live sales calls, 03 2026.
- [70] N. Rackham. *SPIN Selling*. McGraw Hill LLC, 1988.
- [71] Ganesan Ramachandran, Tanmay Narang, and Vallidevi Krishnamurthy. Slam: Sales lead amplification through genai and ml for e-learning platforms. *IEEE Access*, 12:182956–182971, 01 2024.
- [72] A. Rezazadeh. A Generalized Flow for B2B Sales Predictive Modeling: An Azure Machine-Learning Approach. *Forecasting*, 2(3):267–283, 2020.
- [73] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. pages 1135–1144, 08 2016.
- [74] Richard Riley, Joie Ensor, Kym Snell, Frank Jr, Glen Martin, Johannes Reitsma, Karel Moons, Gary Collins, and Maarten van Smeden. Calculating the sample size required for developing a clinical prediction model. *BMJ British Medical Journal*, 368:m441, 03 2020.
- [75] Doru Rotovei, Daniela Zaharie, Negru Viorel, and Daniel Pop. A decision support system for B2B CRM Systems using belief rules. *Proceedings - 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing, ICCP 2020*, pages 131 – 138, 2020.
- [76] Mirka Saarela and Vili Podgorelec. Recent Applications of Explainable AI (XAI): A Systematic Literature Review. *Applied Sciences (Switzerland)*, 14(19), 2024.
- [77] Gaurav Sabnis, Rajdeep Grewal, Gary Lilien, and Sharmila Chatterjee. The sales lead black hole: On sales follow-up of marketing leads. *Journal of Marketing*, 77, 01 2013.
- [78] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.

PloS one, 10:e0118432, March 2015.

- [79] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora Aroyo. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. pages 1–15, May 2021.
- [80] Muzeeb Shaik, Shrihari Sridhar, Chelliah Sriskandarajah, and Vikas Mittal. Opportunity Management for Business-to-Business (B2B) Service Organizations: A Theory-Informed Decision Support Framework. *Production and Operations Management*, 2025.
- [81] Jagdish Sheth. A model of industrial buyer behavior. *Journal of Marketing*, 37, 10 1973.
- [82] Stefan Studer, Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. Towards crisp-ml(q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3:392–413, 04 2021.
- [83] Keith Taber. The use of cronbach’s alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48:1–24, 12 2018.
- [84] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on performance of large language models, 08 2024.
- [85] Sönke Tenckhoff, Mario Koddenbrock, and Erik Rodner. Llmstructbench: Benchmarking large language model structured data extraction, 02 2026.
- [86] W. Utomo, S. Heripracoyo, and R.T. Prasetyo. Benchmarking Machine Learning Algorithms for Customer Churn Prediction in SaaS Platforms Serving SMEs: An Indonesian Case Study. *Architecture Image Studies*, 6(3):1845–1852, 2025.
- [87] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alex Casson. Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14:e0224365, November 2019.
- [88] Kuldeep Vayadande, Smita Sanjay Ambarkar, Viomesh Kumar Singh, Rahul Prakash Mirajkar, Sonali Prashant Bhoite, Amolkumar N. Jadhav, Rakhi Joshi Bharadwaj, Sanket Sunil Pawar, Yogesh Uttamrao Bodhe, and Ganesh B. Dapke. Hybrid approaches in smart sensing for detecting buying intent: performance, reasoning, and real-world deployment. *International Journal on Smart Sensing and Intelligent Systems*, 19(1), 2026.
- [89] John Venable, Jan Pries-Heje, and Richard Baskerville. FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, 25(1):77–89, 2016.
- [90] Ping Wang and Heng Ding. The rationality of explanation or human capacity?

- Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Information Processing and Management*, 61(4), 2024.
- [91] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 01 2022.
 - [92] Jerry Wind and Frederick Jr. A general model for understanding organizational buying behavior. *Journal of Marketing*, 36:12–19, 04 1972.
 - [93] M. Wu, P. Andreev, and M. Benyoucef. The state of lead scoring models and their impact on sales performance. *Information Technology and Management*, 25(1):69–98, 2024.
 - [94] Migao Wu, Pavel Andreev, and Morad Benyoucef. Smart Sales: Amplifying the Power of Predictive Lead Scoring in B2B Sales. *Pacific Asia Conference on Information Systems*, 2024.
 - [95] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets, 07 2023.
 - [96] Chenlang Yi, Gang Li, Zizhan Xiong, Tue Cao, Yanmin Gong, My Thai, and Tianbao Yang. Ress: Learning reasoning models for tabular data prediction via symbolic scaffold, 04 2026.
 - [97] Eelaaf Zahid, Yuya Ong, Aly Megahed, and Taiga Nakamura. Predicting loss risks for b2b tendering processes, 09 2021.

Statement on the Use of Artificial Intelligence

Artificial intelligence appears in this thesis in two quite different roles. In the first, a large language model is itself an object of study: the extraction instrument at the centre of this work (Qwen2.5-14B-Instruct-AWQ) reads sales-call transcripts and returns structured scores, and its configuration, prompting, and validation are documented in full in Chapter 4 and Appendix B. That use is part of the research contribution. The thesis reports it as such.

Separately, the author used a generative AI tool while writing this thesis and while building the demonstration artefact. Claude (Anthropic) helped with language editing and rephrasing for clarity, and with writing and debugging the code behind the demonstrator, such as the dashboard and lead cards, always on material the author had already read, gathered, or reasoned through. It was not used to generate findings, interpret results, or produce citations; every source was checked against its original, and all AI-assisted text and code was reviewed and revised by the author. Responsibility for the whole of this thesis, AI-assisted parts included, rests with the author. The statement follows the University of Twente’s guidelines on the responsible use of generative AI in student work.

Appendix A Feature Dictionary

This appendix lists every feature considered for the model, its source and type, whether it enters Model A (CRM-only) or Model B (CRM + transcript), and exclusions. Reasons for the exclusions are also stated.

A.1 Features used in the model

Table A.1: Features entering Model A and/or Model B.

Feature	Source	Type	Notes
deal_amount (log)	CRM deal	numeric	Deal value; log-transformed for skew. Model A and B.
owner_team / region	CRM deal	categorical	US / NL / Other; carried to absorb the regional confound. A and B.
acquisition_channel	CRM deal	categorical	How the deal entered the pipeline. A and B.
company_employees	Firmographic	numeric	Headcount (enrichment). A and B.
company_linkedin_employees	Firmographic	numeric	LinkedIn-reported headcount (enrichment). A and B.
company_annual_revenue	Firmographic	numeric	Revenue (enrichment); log-transformed. A and B.
company_age_years	Firmographic (derived)	numeric	Deal-created year minus founded year. A and B.
discovery_index	Transcript	ordinal	Mean of the five discovery dimensions (smoothed). B only.
pain_depth	Transcript	ordinal	SPIN Problem dimension (smoothed). B only.
impact_urgency	Transcript	ordinal	SPIN Implication/Need dimension (smoothed). B only.
bant_count	Transcript	count	Sum of the four BANT booleans (0-4). B only.

Feature	Source	Type	Notes
n_explicit_problems	Transcript	count	Distinct problems the buyer articulated. B only.
n_buyer_constraints	Transcript	count	Constraints the buyer raised. B only.
n_rep_discovery_questions	Transcript	count	Open-ended rep questions. B only.
nlp_analysed	Transcript	boolean	False for sub-15-minute or unusable transcripts. B only.

A.2 Features excluded by design

Table A.2: Features deliberately withheld, with the reason.

Feature group	Examples	Reason for exclusion
Stage-derived win probability	hs_deal_stage_probability	Target leakage: ≈ 1.0 for won, ≈ 0 for lost closed deals; single-feature AUC near 1.0.
Post-close fields	lost reason, contract terms, dates set at close	Not knowable at the decision point.
Pre-call engagement counters	email and task counts before the intro call	Logged ~ 20 pp asymmetrically by outcome; leaks outcome through rep behaviour.
Production scores	SPICED, MED-DIC, gap scores, CRM-native scores	Heuristic-derived; excluded so the baseline is raw CRM.
Volume / trajectory features	call count, total minutes, stage-progression fields	Encode pipeline progression, unrelated to intro-call conversation quality.
Raw transcript dimensions	sentiment, individual BANT booleans	Near-zero variance at this sample size; folded into composites or dropped by the blind variance filter.

Feature group	Examples		Reason for exclusion
Holistic overall-signal item	overall strength	signal	Excluded by design to avoid double-counting the discovery signal already captured by the discovery index, and to keep it out of the internal-consistency (α) computation.

Appendix B Full Extraction Rubric

This appendix gives the rubric the language model applies to each introductory transcript. Two rules run throughout: the model writes supporting transcript evidence before assigning each score, and it is forbidden from inferring or using the deal's outcome. Score 0 is reserved for a genuinely absent signal; 1–5 grade quality when the signal is present. The executable prompt and output schema are reproduced in full in Appendix C.

B.1 Ordinal dimensions (0-5)

1. Opening and alignment. The rep's opening: agenda, time-check, framing, mutual goals.

- **0** No opening; dives straight into pitching or questions. **1** Cursory greeting only. **2** States a purpose but no agenda or time-check. **3** Sets a basic agenda and confirms available time. **4** Sets an agenda and asks what the buyer wants from the call. **5** Buyer co-creates the agenda; time confirmed; mutual outcomes aligned.

2. Situation understanding. How deeply and specifically the rep learns the buyer's current state.

- **0** No discovery. **1** One or two surface questions, no follow-up. **2** Context across 2–3 areas but only surface answers. **3** A solid picture across 3+ dimensions with some follow-up. **4** Specific map (names, numbers, sequences) with probing follow-ups. **5** Rich multi-layered map; rep validates understanding at the end.

3. Pain depth. The clarity and severity of the problem surfaced.

- **0** No problem identified. **1** Vague dissatisfaction, never articulated. **2** A problem named but cause/consequence not explored. **3** A clear problem the buyer confirms matters; cause or consequence touched. **4** A specific, confirmed problem with both cause and consequence discussed. **5** Multiple problems, or one quantified business-impacting pain named as a priority.

4. Impact and urgency. Business consequence and time pressure or compelling event.

- **0** No consequence or timeline; pure exploration. **1** Faint interest, no consequence or timeline. **2** Soft importance or a vague timeline, but not both and not concrete. **3** Either a consequence or a timeline, but not connected. **4** Consequence and timeline both clear and connected. **5** An explicit compelling event or deadline with a quantified consequence of inaction.

5. Next steps. The concreteness of the agreed next step.

- **0** No next step. **1** A vague next step, no owner or date. **2** A follow-up agreed in principle, no date or attendees. **3** A concrete action but missing owner, date, or confirmation. **4** A specific step with owner and a tentative date. **5** A booked action with date, owner, and confirmed attendees.

6. Sentiment score. Buyer affect and engagement, scored by counting observable behaviours, with tone left uninferred.

- **0** Buyer affect indeterminable. **1** Resists, dismisses, pushes back. **2** Guarded; short factual answers, no questions back. **3** Cooperative but transactional; full answers, no elaboration. **4** Engaged; asks 2+ substantive questions or volunteers detail. **5** Enthusiastic; proactively suggests next steps or widens scope.

7. Overall signal strength. The overall quality of discovery *execution*, explicitly not the commercial outcome.

- **0** No meaningful discovery. **1** Very weak; minimal information surfaced. **2** Below par; gaps across multiple dimensions. **3** Competent and unremarkable; basics covered. **4** Strong; situation mapped, pain specific, next steps concrete. **5** Exemplary execution regardless of outcome; buyer state fully transparent and validated.

B.2 Qualification booleans (BANT)

Marked true only when genuinely addressed. Mere implication does not count: **budget** (pricing or willingness to spend discussed), **authority** (decision-maker or buying process identified), **need** (a clear articulated need beyond casual interest), **timeline** (a purchase or decision timeline discussed).

B.3 Behaviour counts

Non-negative integers: **n_explicit_problems** (distinct problems the buyer articulated), **n_buyer_constraints** (budget, timeline, authority, or technical constraints raised), **n_rep_discovery_questions** (open-ended discovery questions asked by the rep).

B.4 Edge cases

If the call has no buyer participation, the buyer is non-responsive, the transcript is unavailable, or the call is not in English, the record is flagged **analysed: false** and the seven scores are null.

Appendix C Complete Extraction Prompt

This appendix reproduces the system prompt, user-message template, and enforced JSON schema word for word from the extraction script. A comment in the script identifies this text as the exact code that produced the data, run at temperature 0 and seed 0 (Table 4.5). The prompt and schema content is unchanged; only the surrounding code, variable assignments and unrelated functions, has been removed. Table C.1 traces each design choice in the prompt to the literature that grounds it.

Table C.1: Design choices in the extraction prompt and their grounding.

Design choice	Grounding
Per-level anchored 0–5 rubric (every level defined, not just endpoints)	[40, 48]
Seven independently scored dimensions rather than one holistic pass	[95]
Evidence and reasoning written before the score is committed (Rule 4)	[48, 91, 42]
Calibration examples for the boundary-prone dimensions (few-shot exemplars distinguishing adjacent levels)	[12, 39]
Schema-constrained decoding trade-off (guarantees parseable output; can shift errors into silently wrong values rather than parse failures)	[85, 84]
LLM output entering the dataset as a measurement, not just generated text, requiring independent validation	[29, 65, 32]
Reproducible, deterministic execution (open weights, temperature 0, fixed seed)	[85, 32]
Discovery-dimension grounding (SPIN)	[70]
BANT and organisational-buying grounding	[92, 81, 45]
Model-tier selected on human-agreement grounds	[34]

C.1 System prompt

You are a B2B sales-conversation analyst scoring a single introduction (discovery) call between a sales representative ("the rep") and a prospect ("the buyer") at a B2B technology vendor. Your scores become features in a research study, so they must be consistent, calibrated, and grounded strictly in the transcript. You evaluate HOW the call was conducted and WHAT the buyer revealed - never whether the deal will close.

For each of the seven dimensions, first identify the supporting evidence in the transcript, then assign the score - in that order. Detect the four BANT signals, count the three process signals. Return ONLY a single JSON object matching the required schema; write each item's evidence before its score/present value. Output nothing else.

ANONYMISATION: The transcript arrives ALREADY masked --the sales representative is labelled [AE] (or [AE_1], [AE_2] if multiple reps are present), buyer-side speakers are [BUYER] (or [BUYER_1], [BUYER_2] if multiple). Contact details are [EMAIL]/[PHONE]/[URL]; the host vendor and its products are [HOST COMPANY]/[HOST PRODUCT]; competitors are [COMPETITOR]; the buyer's organisation is [COMPANY]. If any real name slips through the masking, exclude it from your evidence fields. Never reproduce a real name or company name.

SCORING RULES

1. Judge ONLY this transcript. Do not infer from company size, industry, deal value, or anything outside the conversation.
2. NO OUTCOME INFERENCE. Do not reason about, hint at, or let your scores be swayed by whether the deal closed, the stage it reached, or how promising it looks commercially. Score conversation execution and revealed buyer state only.
3. Keep judgements B2B-agnostic; do not reward or penalise any industry's terminology.
4. EVIDENCE BEFORE SCORE: write the transcript evidence first, then the number; the score must follow from the evidence you wrote.
5. USE THE FULL RANGE - do not default to 3 or 4. Weak evidence -> score low (0-2); strong evidence -> score high (4-5). Reserve 0 for genuinely absent signals. DO NOT CLUSTER around 3-4.
6. EVIDENCE BREVITY: each evidence field is ONE brief paraphrase or short quote (max ~25 words). Never reproduce extended transcript passages.

COMMON MISTAKES TO AVOID:

- Outcome leakage: scoring up/down based on whether the deal closed, not on call quality.
- Sentiment as a vibe: saying "the buyer sounded happy" instead of COUNTING specific behaviors (questions asked, elaboration, pushback).
- Situation as a data dump: buyer talked a lot ≠ rep discovered; verify rep probed specifics and validated understanding.
- Defaulting to 3 or 4: there are no "neutral" calls; they have gaps (score 1-2) or depth (score 4-5).

- Confusing next_steps scope: a vague "I'll send something" is NOT a 4; a 4 requires a concrete, scheduled action.

DIMENSIONS - integer 0-5. 0 = genuinely absent/not addressed; 1-5 grade quality when present.

1. opening_and_alignment - Does the rep set context, agenda, and check buyer's goals before diving in?

0 rep dives straight in, no agenda | 1 cursory greeting only, no purpose | 2 purpose stated, missing agenda or time-check | 3 basic agenda, time confirmed | 4 clear agenda, time confirmed, asks buyer's goals | 5 buyer co-creates agenda, mutual outcomes stated

2. situation_understanding - Does rep map CURRENT STATE with specificity and follow-up probing?

0 none; rep pitches without discovering | 1 one or two surface facts, no follow-up | 2 context gathered across 1-2 areas, shallow, no probing | 3 solid grasp across 3+ areas, some follow-up, buyer confirms | 4 detailed & specific (names, numbers), probes follow-ups, buyer elaborates | 5 rich map of current state; rep validates understanding ("So it sounds like...")

3. pain_depth - Is there a real problem? Is it vague or SPECIFIC AND CONFIRMED?

0 no problem identified | 1 vague dissatisfaction ("things could be better"), not articulated | 2 problem named ("process is slow"), never explored for cause or impact | 3 clear problem confirmed; cause OR consequence discussed, not both deeply | 4 specific problem confirmed by buyer; BOTH cause AND consequence discussed | 5 multiple problems OR single problem quantified & confirmed as priority ("we lose 2 days/month, affects invoicing")

4. impact_and_urgency - Why does this MATTER NOW? Is there a deadline/trigger?

0 "just exploring"; no consequence or timeline | 1 faint interest; no consequence or timeline | 2 soft importance ("nice to have") OR loose timeline ("sometime"), not both | 3 consequence stated OR timeline mentioned (one, not both) | 4 consequence AND timeline both discussed and connected | 5 explicit deadline + quantified cost ("if we don't migrate by June, we lose 1 week of productivity")

5. next_steps - Is it vague or CONCRETE, SCHEDULED, and OWNED?

0 no next step; call ends open-ended | 1 vague ("we'll be in touch"), no date or owner | 2 follow-up agreed in principle ("we should demo"), unscheduled | 3 concrete action ("demo next week"), missing owner or specific date | 4 specific step with owner and date ("Tuesday 2pm demo with ops lead") | 5 action booked with date, owner, attendees, and invitation sent

6. sentiment_score - BUYER AFFECT (BIPOLAR). Use 0 only if affect is indeterminable. SCORE BY COUNTING BEHAVIORS, NOT TONE.

1 resists/dismisses: "I don't think this is a fit"; pushback; sarcasm; answers grudgingly | 2 guarded: short factual answers only; no questions back; deflects probing ("that's not our concern") | 3 cooperative, transactional: full answers when asked; 0-1 questions back; no unprompted detail | 4 engaged: asks 2+ substantive questions OR volunteers unprompted detail ("we tried that last year...") OR builds on rep's points | 5 enthusiastic + engaged: level-4 behaviors PLUS proactively suggests next steps ("loop in my manager?") or expands scope

7. overall_signal_strength - How well did the rep EXECUTE DISCOVERY? How clearly did the buyer's STATE EMERGE? NOT about closing likelihood.
0 no discovery; monologue or no buyer engagement | 1 very weak execution; little learned about buyer | 2 below-par; gaps across multiple dimensions | 3 competent, unremarkable; basics covered, no depth | 4 strong execution; situation/pain/process clear; buyer engaged | 5 exemplary execution; buyer's state fully surfaced; rep validates understanding; call is well-structured regardless of outcome

BANT BOOLEANS (present = true ONLY if genuinely addressed in conversation, not implied or assumed)

- budget: budget range, pricing, or willingness/ability to spend discussed?
- authority: decision-maker, buying process, or "who else is involved" identified?
- need: a clear, articulated NEED established (beyond casual interest)?
- timeline: purchase or decision timeline discussed?

METHODOLOGY COUNTS (non-negative integers)

- n_explicit_problems: distinct problems/pains the buyer articulated
- n_buyer_constraints: budget/timeline/authority/technical constraints the buyer raised
- n_rep_discovery_questions: open-ended discovery/probing questions by the rep (count Qs, not pitches)

CALIBRATION EXAMPLES (understand scale, do not copy wording)

sentiment_score: 1="don't think it's a fit" | 2=short answers, no questions back | 3=full answers, 0-1 question back | 4=asks 2+ questions | 5=meets level-4 + proactively involves others

pain_depth: 1="things could be better" | 2="process is slow" (never explored) | 3="we lose time" (one of cause/consequence) | 4="we lose 2 days/month and it delays invoicing" (both) | 5=quantified + multiple problems + confirmed urgent

next_steps: 2="we should do a demo" (unscheduled) | 3="demo next week" (no specific date) | 4="Tuesday 2pm demo with your ops lead" | 5="calendar invite for Tuesday 2pm, attendees: you, your ops lead, and me"

C.2 User-message template

Each call fills the single placeholder below with that deal's masked transcript text:

Transcript:

```

"""
{transcript}
"""

```

C.3 Enforced output schema

The schema below is enforced at decoding time through guided, structured-output decoding. Every field the model returns must match it exactly:

```

_DIM = {"type": "object", "additionalProperties": False,
        "required": ["evidence", "score"],
        "properties": {"evidence": {"type": "string", "maxLength": 300},
                        "score": {"type": "integer", "enum": [0, 1, 2, 3, 4, 5]}}}
_BANT = {"type": "object", "additionalProperties": False,
         "required": ["evidence", "present"],
         "properties": {"evidence": {"type": "string", "maxLength": 300}, "
         present": {"type": "boolean"}}}
DIM_KEYS = ["opening_and_alignment", "situation_understanding", "pain_depth",
            "impact_and_urgency", "next_steps", "sentiment_score", "
            overall_signal_strength"]
BANT_KEYS = ["budget", "authority", "need", "timeline"]
COUNT_KEYS = ["n_explicit_problems", "n_buyer_constraints", "
n_rep_discovery_questions"]

SCHEMA = {
    "type": "object", "additionalProperties": False,
    "required": ["dimensions", "bant", "methodology_counts"],
    "properties": {
        "dimensions": {"type": "object", "additionalProperties": False,
                        "required": DIM_KEYS, "properties": {k: _DIM for k in
                        DIM_KEYS}},
        "bant": {"type": "object", "additionalProperties": False,
                 "required": BANT_KEYS, "properties": {k: _BANT for k in
                 BANT_KEYS}},
        "methodology_counts": {"type": "object", "additionalProperties":
        False,
                                "required": COUNT_KEYS,
                                "properties": {k: {"type": "integer", "minimum":
                                0} for k in COUNT_KEYS}},
    },
}

```

The system and user messages above form a two-message chat (`role: system / role: user`). Decoding runs at temperature 0 with a fixed seed, and vLLM enforces the schema through structured-output (guided JSON) decoding, as summarised in Table 4.5.

Appendix D User-Study Template

This appendix documents three measurement instruments connected to the artefact: an acceptance survey embedded live in the demonstrator (Section D.1), and two studies not conducted within this thesis but prepared as sample forms for future work (Sections D.2 and D.3).

D.1 Artefact usage feedback form (TAM)

To measure acceptance of the demonstrator in use, a short survey based on the Technology Acceptance Model (TAM) [21], is embedded in the artefact itself: every lead card, and each stakeholder view, carries a feedback button that opens the survey below.

The survey is deliberately minimal, about one minute per response, with a single typed field: the deal reference, copied to the respondent’s clipboard by the feedback button because the form platform does not support pre-filled links. Responses identify deals only by their fictional pseudonyms, mapped back through a private lookup that never leaves the research environment. Table D.1 contains the acceptance statement, here PU: Perceived Usefulness, PEOU: Perceived Ease of Use, and BI: Behavioural Intention.

Table D.1: Acceptance statements (1 = strongly disagree, 7 = strongly agree).

	Statement	1	2	3	4	5	6	7
PU1	Using the tool improves my decisions about which deals to pursue after a discovery call.	☐	☐	☐	☐	☐	☐	☐
PU2	The tool is useful for prioritising my pipeline.	☐	☐	☐	☐	☐	☐	☐
PEOU1	The tool is clear and understandable.	☐	☐	☐	☐	☐	☐	☐
PEOU2	It is easy to find the information I need in the tool.	☐	☐	☐	☐	☐	☐	☐
BI1	If the tool were available, I would use it in my weekly workflow.	☐	☐	☐	☐	☐	☐	☐
BI2	I would recommend that colleagues use the tool.	☐	☐	☐	☐	☐	☐	☐

Construct scores are the mean of the two items per construct; with two items per construct, reliability is reported with the Spearman–Brown coefficient. This is the appropriate index for two-item constructs. Responses are split by role and, through the pseudonym mapping, by the referenced deal’s priority tier.

D.2 Human inter-rater (ICC) study: sample rating form

This study is not conducted within this thesis. The form below is the prepared rating form: two to three raters would score 48 anonymised transcripts on the anchors of Appendix B, blind to deal outcome, with agreement quantified by the intraclass correlation coefficient and weighted κ [43, 44].

Table D.2: Discovery dimensions (0 = absent; score by the anchors of Appendix B).

Dimension	0	1	2	3	4	5
Opening and alignment	☐	☐	☐	☐	☐	☐
Situation understanding	☐	☐	☐	☐	☐	☐
Pain depth	☐	☐	☐	☐	☐	☐
Impact and urgency	☐	☐	☐	☐	☐	☐
Next steps	☐	☐	☐	☐	☐	☐
Sentiment (count behaviours, leaving tone uninferred)	☐	☐	☐	☐	☐	☐
Overall signal strength	☐	☐	☐	☐	☐	☐

D.3 End-user evaluation of the explanations: sample questionnaire

The questionnaire below combines the end-user evaluation with the acceptance survey of Appendix D.1 into a single instrument for a 30-minute session with RITAM sales-team members. Each session records the participant’s name and role, then steps through the five lead cards of Section 4.5 in turn; every item below runs on the same 1–5 scale (1 = strongly disagree, 5 = strongly agree) unless noted otherwise. The comprehension check (a distractor that never appeared in the explanation) guards against placebo trust [90].

1. Per lead (repeated for each of the five cards). Ask which factor contributed most to this lead’s score (multiple choice: the shown drivers plus one distractor, options vary per card), then rate:

- This score matches my professional read of this deal.
- This explanation would help me decide what to do next with this deal.

2. System Usability Scale [11]. Items marked (R) are reverse-scored for analysis. Participants do not see the marking; scored to the standard 0–100 scale (benchmark mean about 68).

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex. (R)
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system. (R)

5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system. (R)
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use. (R)
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system. (R)

3. Trust.

- I trust that the lead score reflects the actual quality of the deal.
- I would rely on this score when prioritising my pipeline.
- I would feel comfortable explaining this score to a colleague or manager.
- If the score disagreed with my own judgement, I would still consider what it says.

4. Explanation clarity.

- I understand why each lead received the score it did.
- The explanations highlight factors I consider important when evaluating a lead.
- The explanations are written in language that makes sense for my role.

5. Acceptance and intended use. The TAM statements of Appendix D.1 (PU1, PU2, PEOU1, PEOU2, BI1, BI2), rescaled from 1–7 to the same 1–5 scale and asked once at the end of the session rather than per deal.

6. Debrief.

1. What was most useful about the explanations?
2. What would you change or add before you'd use this weekly?
3. Did any score or explanation surprise you? Which, and why?

Analysis, if conducted: SUS as a single benchmarked score; trust, clarity, and acceptance as descriptive subscale means given the expected small sample; comprehension accuracy cross-tabulated against trust, where high trust with low comprehension is the placebo-trust pattern [90].