LeadFlow: A Systems Engineering Approach to AI-Driven Sales Prospecting in Zero-Baseline Environments

Madeline A. Priebe[1]*, Carly Elbaum[1], Zack Sikkink[1], Ian Girdner[1], Dylan Jones[1],
Bhima Pibuldham[1], Matthew Burkett[1]
**Email: map7cma@virginia.edu*

[1]*University of Virginia*
Department of Systems and Information Engineering
Charlottesville, VA USA

***Abstract*—Manual sales prospecting in emerging Connected-Commerce (C-Commerce) markets creates fundamental misalignment between individual representative judgment and organization-level optimization. Representatives at Zbooni, a UAE-based C-Commerce platform, allocated approximately 30% of working hours to manual prospecting while achieving only 1% conversion rates. This paper presents LeadFlow, a full-stack AI-driven prospecting system developed using formal systems engineering methodology. The central challenge was deploying machine-learning-based qualification without historical performance data—a zero-baseline environment. We contribute (1) a five-phase zero-baseline validation framework that bootstraps qualification capability from stakeholder-derived heuristics and progressively transfers trust to empirical evidence, and (2) a requirements traceability methodology that resolves principal-agent misalignment through objective-tree decomposition. This paper's primary contribution is methodological: the zero-baseline framework and requirements traceability approach are designed to be potentially transferable to similar data-scarce deployment contexts independent of domain. Pilot deployment has confirmed discovery pipeline throughput and scoring coverage across 12,720 businesses and 2,496 active leads; projected impact on effort reduction and conversion rate improvement are design targets dependent on outreach deployment, reported here as evaluation hypotheses rather than empirical outcomes. We predict that the framework generalizes beyond C-Commerce to any organization deploying AI-assisted judgment in data-scarce environments.**

***Keywords*—systems engineering, AI prospecting, zero-baseline learning, requirements engineering, C-Commerce, lead qualification**

I. INTRODUCTION

Organizations deploying AI systems in emerging markets face critical challenges in automating judgment-heavy workflows when historical performance data do not exist. Traditional machine learning assumes abundant training data, yet early-stage companies, new product launches, and zero-baseline contexts lack the labeled examples required for supervised learning. We define a *zero-baseline environment* as one in which available labeled outcomes are insufficient in both quantity and class balance to support reliable supervised learning or statistically meaningful evaluation, often due to limited instrumentation, label sparsity, or outcome skew. This paper demonstrates how Systems Engineering (SE) methodology enables effective AI automation even when historical data is unavailable.

Connected-commerce (C-Commerce) represents an emerging market paradigm in which social media platforms, messaging applications, and e-commerce infrastructure converge to enable direct merchant-to-consumer transactions [1]. In the Middle East and North Africa (MENA), C-Commerce has emerged as major sales channel for small and medium-sized businesses. Zbooni, a UAE-based C-Commerce platform, provides merchants with tools to manage sales, inventory, and customer relationships across multiple communication channels such as WhatsApp and Instagram. However, the company's internal sales prospecting process—the mechanism by which Zbooni identifies and acquires new merchant clients—faces significant inefficiencies characteristic of manual, decentralized workflows. Sales representatives allocate approximately 30% of working hours to manual Instagram searches and web research, yet conversion rates remain at roughly 1%. This inefficiency represents both an opportunity cost and a systemic failure in resource allocation that limits the company's growth potential in a competitive market. This paper reports both the system design methodology and results from a partial pilot deployment validating discovery and qualification behavior; outreach and conversion measurement are ongoing.

The fundamental problem is misalignment between individual representative judgment and organization-level optimization objectives across multiple system dimensions. In decentralized prospecting workflows, each sales representative independently discovers potential leads, applies subjective qualification criteria, and initiates contact without systematic feedback mechanisms or performance measurement infrastructure. These interconnected inefficiencies constitute a complex system requiring holistic re-design rather than disparate point solutions. This work demonstrates how SE methodology structures the design of AI-augmented workflows under data-scarcity constraints, contributing a replicable zero-baseline validation framework applicable beyond C-Commerce to any B2B sales domain facing similar constraints.

II. RELATED WORK

A. Sales Automation and CRM Systems

Business process automation has emerged as a significant research domain across organizational contexts [2]. Commercial Customer Relationship Management (CRM) platforms such as Salesforce and HubSpot dominate sales process management, providing contact management, pipeline tracking, and automated communication at scale [3], [4]. AI integration into these systems has demonstrated meaningful improvements in sales efficiency and strategic positioning when embedded into CRM architecture rather than added as an afterthought [4]. Recent AI extensions—Salesforce Einstein Lead Scoring and HubSpot Predictive Lead Scoring—address qualification through supervised ML models, but both impose hard data minimums before activation, requiring hundreds to thousands of prior conversion outcomes [5], [6]. Despite these advances, a systematic review of the field identifies a persistent gap: existing research focuses on *why* to adopt AI-CRM but not *how* to implement it, particularly for implementation in upstream contexts [3], [4]. These systems uniformly assume a structured lead pool already exists, optimizing pipelines after discovery rather than before. LeadFlow addresses this upstream gap by automating prospecting and qualification in contexts where no pipeline history exists.

B. AI-Driven Lead Qualification

Supervised machine learning approaches to lead scoring have been widely studied in B2B contexts, with models including logistic regression, gradient boosting, and neural networks trained on historical CRM data demonstrating meaningful improvements in lead prioritization [7], [8]. Nygård and Mezei (2020) demonstrated that random forest classifiers using behavioral data can effectively estimate purchase probability, while González-Flores et al. (2025) showed that gradient boosting classifiers outperform competing algorithms on real CRM data spanning four years of conversions [8]. The closest conceptual peer to LeadFlow as an integrated prospecting system, Scrapus [9], achieves high precision through reinforcement learning and transformer-based NLP, but requires training data and operates in established markets. The fundamental constraint shared across all these approaches is dependence on historical outcomes. The analogous cold-start problem—deploying predictive systems without historical interaction data—has been extensively studied in recommendation systems, where knowledge-based approaches serve as interim solutions before data-driven models become viable [10], [11]. LeadFlow extends this literature by providing a systematic, SE-grounded methodology for transitioning from knowledge-based to data-driven qualification as labeled outcomes accumulate.

C. Systems Engineering in Business Process Design

Systems Engineering (SE) methodology offers tools well-suited to deploying AI in organizational contexts, providing a structured approach to integrating stakeholder requirements, architecture design, and iterative validation across a system lifecycle [12]. Prior work has applied SE and requirements engineering to adaptive system design under uncertainty, identifying stakeholder misalignment, evolving requirements, and the challenge of translating business objectives into implementable specifications as persistent obstacles [13]. More recent research demonstrates that SE methodology applied to business automation contexts—not just aerospace or defense—yields structured frameworks for aligning stakeholder objectives in dynamic environments [14]. Despite this growing body of work, formal SE methodology has not been explicitly applied to AI-driven sales prospecting in zero-baseline environments. LeadFlow addresses this gap by demonstrating how classical SE methods—stakeholder analysis, objective-tree decomposition, requirements traceability, and iterative validation—can structure the design of AI-augmented sales systems where traditional data-driven approaches are infeasible.

III. REQUIREMENTS ENGINEERING

A. Stakeholder Analysis and Misalignment

Three primary stakeholder groups were identified: sales representatives, management/leadership, and end customers (merchants). Semi-structured interviews were conducted with two sales representatives (including the primary prospecting specialist) and two members of leadership to elicit objectives, pain points, and success criteria. The stakeholder set corresponds to the full set of operational decision-makers within Zbooni's prospecting workflow, ensuring that the analysis captures the complete set of relevant organizational perspectives. Interview protocols used open-ended questions about current workflows, challenges, and desired normative scenarios [15].

Analysis revealed a principal-agent misalignment [16], [17]: representatives optimized for volume metrics (messages sent, contacts initiated) under the assumption that higher volume yields proportionally more conversions, while management sought quality metrics and conversion visibility. This misalignment emerged from two structural factors: absence of conversion data enabling quality-based targeting, and organizational culture rewarding activity over outcomes.

Management could not answer fundamental operational questions: which merchant segments convert at higher rates, which outreach strategies prove most effective, or how leads progress through the pipeline. This constitutes a classic principal-agent problem in which representatives pursue objectives—volume—that diverge from organizational goals—revenue growth through quality conversions—without measurement infrastructure to detect or correct the divergence.

B. Objective Tree Analysis

Objective-tree analysis was applied to decompose these conflicting stakeholder objectives into a unified set of measurable system requirements (Fig. 1). The analysis revealed that while the surface-level conflict was between volume and quality, both groups shared the underlying goal of acquisition efficiency. They simply lacked the tools to pursue these goals systematically. This reframing enables five derived requirements to serve both stakeholder groups simultaneously:

- **R1**: Automated lead discovery, eliminate manual research burden while enforcing consistent customer criteria
- **R2**: Model-based qualification scoring, providing priority signals to reps and segment intelligence to management
- **R3**: Structured volume control outreach, allocate messaging budget to quality leads while generating A/B data
- **R4**: Real-time pipeline visibility through shared state rather than individual spreadsheets
- **R5**: Adaptive feedback loops providing systematic improvement over time

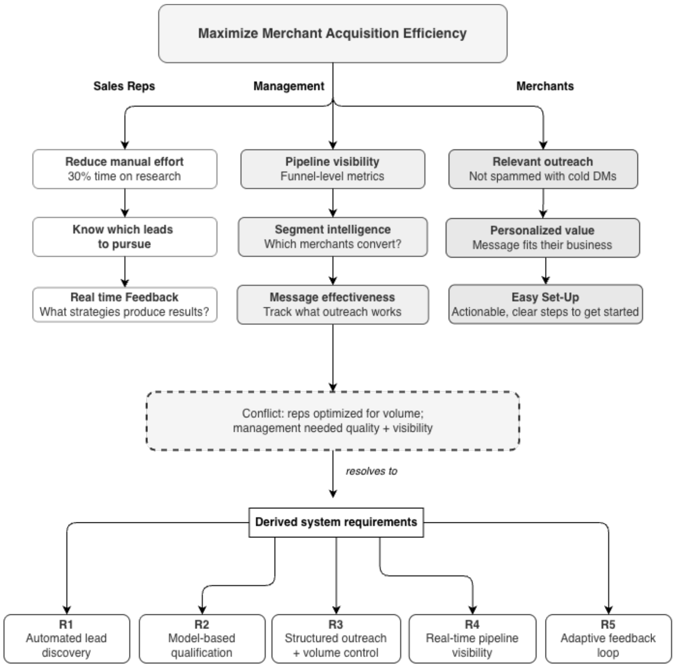

Fig. 1. Objective tree decomposing stakeholder conflicts into unified system requirements.

C. Requirements Traceability

Table I presents the requirements traceability matrix, mapping each derived requirement to its stakeholder origin, the design decision it motivated, and its success metric. The traceability from stakeholder objective through design decision to system component is the mechanism by which SE methodology resolves organizational misalignment at the architectural level.

IV. SYSTEM ARCHITECTURE

Fig. 2 illustrates the end-to-end LeadFlow architecture. The system begins with Ideal Customer Profile (ICP) parameters—industry vertical, geographic location, and relevant hashtags—driving automated discovery across four parallel sources: Google Places API for location-based business search, social media enrichment for digital presence signals, professional network enrichment for organizational context, and a UAE business directory for contact fallback. Data collection is performed through verified APIs to ensure compliance with platform specific terms of service.

TABLE I
REQUIREMENTS TRACEABILITY MATRIX

Req	Stakeholder Objective	Design Decision	Success Metric
R1	Rep: reduce 30% manual effort	ICP-seeded automated discovery pipeline	Manual prospecting $<5\%$ total logged sales time
R2	Rep: lead priority; Mgmt: segment intel	Hybrid scoring (rules + ML)	Conversion rate $>1\%$ baseline
R3	Both: message effectiveness; Merchant: relevance	Score-gated outreach; T-VP-CTA message framework	Response rate; variant A/B performance
R4	Mgmt: pipeline visibility	Centralized queue; shared real-time dashboard	Funnel metrics available in real time
R5	Both: systematic improvement	Three adaptive learning loops	AUC improvement

Discovered businesses pass through pre-qualification stage (DNS validation, minimum review count, parked domain detection) to filter out low-quality or non-operational entities prior to downstream processing. Remaining records enter an enrichment pipeline that populates contact channels and structured profile metadata. Enriched leads are then scored against ICP-aligned features and routed to structured outreach workflows.

Outreach messages are generated through an AI-driven personalization module conditioned on enriched lead features and ICP context. Messages follow a structured Trigger–Value Proposition–Call to Action (T-VP-CTA) format: the trigger establishes relevance using lead-specific signals, the value proposition aligns platform capabilities with inferred business needs, and the call to action is designed to be low-friction to encourage response.

A centralized backend service orchestrates data processing and workflow execution, while a shared interface provides visibility for both representatives and management.

V. ZERO-BASELINE QUALIFICATION FRAMEWORK

A central challenge in LeadFlow's design was the absence of historical performance data. Traditional supervised machine learning approaches to lead scoring–logistic regression, gradient boosting, neural networks–require substantial labeled datasets of won and lost opportunities to achieve production-grade accuracy. At project inception, Zbooni possessed fewer than 200 documented conversion outcomes heavily skewed toward positive responses. This constraint is characteristic of early-stage companies, new market segments, and organizations without prior performance instrumentation. We term such contexts *zero-baseline environments* and contribute a five-phase framework for deploying AI qualification systems under this constraint (Fig. 3).

Phase 1 — Stakeholder heuristic elicitation: Semi-structured interviews gathered the qualification knowledge that experienced representatives apply implicitly. These heuristics exist independent of historical data and represent the organization's current best understanding of ideal customer characteristics.

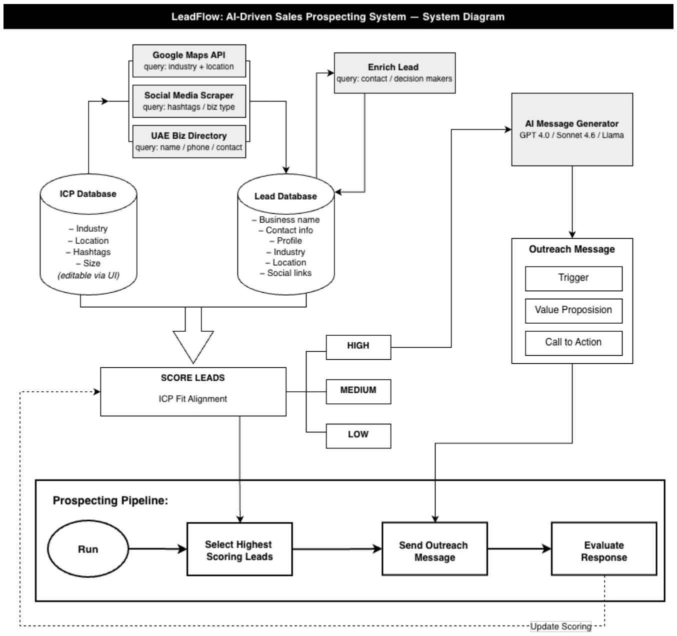

Fig. 2. LeadFlow system architecture. Data flows from ICP configuration through multi-source discovery, enrichment, ICP-aligned qualification scoring, and structured outreach with adaptive feedback.

Phase 2 — Feature operationalization: Heuristics are translated into 43 measurable features spanning digital presence signals (Instagram follower count and engagement rate, WhatsApp availability), e-commerce readiness indicators (payment widget count, product catalog presence, booking form availability), business maturity signals (Google review count, multi-staff detection), and ICP fit criteria (industry vertical match, geographic constraint). Initial feature weights are set through stakeholder consensus rather than empirical optimization.

Phase 3 — Deterministic deployment: Criteria are deployed as a rule-based scoring model serving as the primary qualification mechanism during the data-sparse period. The model applies hard filters (binary pass/fail criteria: country constraint, website presence) before computing a weighted score across positive and negative signals with Laplace smoothing to prevent zero scores for leads with missing data [18]. Critically, this model is fully explainable. Every score decomposes into constituent rule contributions. This maintains stakeholder trust before empirical validity is established.

Phase 4 — Controlled experimentation: A/B message variant testing generates the first labeled outcome data as leads respond, bounce, or convert [19], [20]. These data accumulates as a byproduct of normal system operation, eliminating the need for a separate data collection phase.

Phase 5 — Performance-gated ML activation: A logistic regression model trained on accumulated labeled outcomes is deployed only when it achieves AUC $\geq$ 0.70 on a held-out validation set with $\geq$200 labeled samples [21], [22]. The blend ratio between deterministic and ML scores shifts dynamically: 90/10 (rules/ML) during the data-sparse period; 70/30 at AUC $\geq$ 0.70; 50/50 at AUC $\geq$ 0.80 with 500+ samples. This graduated trust transfer ensures the ML model earns influence through demonstrated performance rather than assumption. Thresholds for model activation were selected based on standard benchmarks for minimally acceptable classifier performance in imbalanced classification settings and internal constraints on label acquisition rates; sensitivity to these thresholds is discussed as a limitation.

Three adaptive learning loops operate continuously: (1) ML model retraining when 50+ new labeled outcomes accumulate, deploying new versions only if AUC improves; (2) adaptive search budget allocation using exponential moving average of per-ICP conversion rates, reducing search task provisioning as the system learns discovery efficiency; and (3) weekly message performance analysis generating ranked recommendations for threshold adjustments visible on the management dashboard.

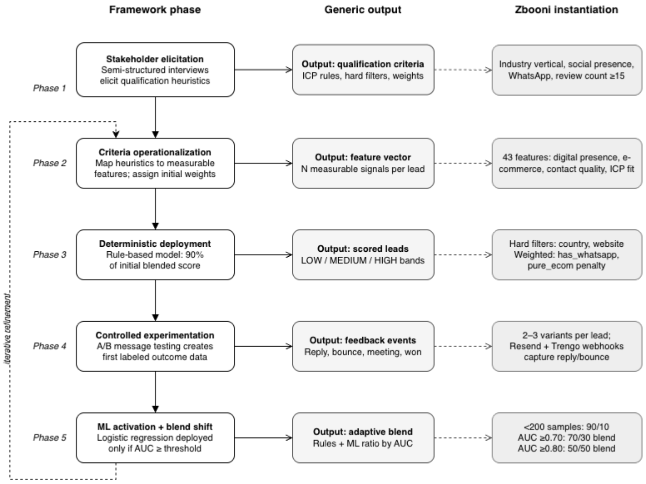

Fig. 3. Five-phase zero-baseline validation framework. Left column: generic methodology phases. Center: outputs at each phase. Right: Zbooni-specific instantiation. Feedback arrow indicates iterative refinement across phases.

A. Design Decision Analysis

Four key design decisions merit discussion as they reflect the SE tradeoffs central to this work. *Expert-driven over data-driven scoring* (R2) prioritized explainability and immediate deployability over potential long-term accuracy, appropriate for the zero-baseline context where stakeholder trust must be earned before ML influence is warranted. *Centralized queue over individual lists* (R4) resolved the principal-agent problem architecturally by making the organization's lead pipeline a shared resource rather than a collection of individual inventories. The system creates management visibility while preserving representative autonomy through manual override. *Free-first enrichment* (all requirements) reduced per-lead enrichment cost by exhausting free data sources before invoking paid APIs, addressing budget constraints characteristic of early-stage deployment. *Score-gated outreach* (R3) ensures that WhatsApp daily messaging limits are allocated to the highest-probability leads, making qualification accuracy directly consequential for operational performance.

VI. EVALUATION METHODOLOGY

The LeadFlow system is designed for evaluation through a structured pilot deployment with the Zbooni sales team using a pre/post comparison design. Pre-deployment baseline metrics are established through structured observation of the existing manual workflow. Post-deployment metrics will be collected over a minimum six-week pilot period, sufficient for the adaptive learning mechanisms to accumulate initial outcome data. The evaluation captures quantitative outcomes against the success metrics in Table I.

The primary unit of analysis is the individual lead. Key evaluation metrics include response rate (proportion of contacted leads that reply) and conversion rate (proportion of leads that convert to active merchants). Given the absence of labeled outcomes during initial deployment, evaluation focuses on upstream system performance metrics rather than predictive accuracy.

Note: The system has completed pilot discovery and qualification runs. Pipeline efficiency and scoring behavior are reported from validated batch data in VII. Full outreach deployment and conversion tracking (R2 and R3) constitute the next evaluation phase and are pending regional stabilization.

A. Validation in Zero-Baseline Contexts

Validation in zero-baseline environments presents a distinctive methodological challenge: the absence of historical ground truth means early qualification accuracy cannot be evaluated against known outcomes. Our approach treats stakeholder-derived criteria as provisional ground truth for the initial deployment period, subject to iterative refinement as empirical data accumulates. This is explicitly acknowledged as a limitation. Stakeholder-derived criteria may encode the same biases that produced the original 1% conversion rate, creating a risk that the system reinforces existing inefficiencies rather than correcting for them. Two mechanisms detect and correct such a bias: A/B testing provides signal on which lead characteristics correlate with positive responses independent of scoring weights, and weekly manager analysis compares actual reply rates against model-predicted rates across ICP segments, flagging systematic miscalibrations.

VII. RESULTS

A. Pipeline Throughput

In the validated April 9 batch, 250 automated search queries produced 2,509 attributed businesses and 705 leads, corresponding to 2.83 leads per search and 1.32 workable leads per search. Of these scored leads, 135 (19.1%) exceeded the 0.90 scoring threshold, representing the system's highest-priority qualification tier. Across the full pilot dataset, the system processed 12,720 businesses and generated 2,496 active leads, of which 1,398 were classified as workable (qualified or drafted). Approximately 19.5% of businesses produced at least one lead and 10.9% produced at least one workable lead. The 19.1% top-tier rate observed in the April 9 batch is consistent with the 18.7% rate across the full pilot dataset (466 of 2,496 leads), suggesting stable scoring behavior across discovery runs.

B. Enrichment Effectiveness

Contact signal extraction varied across enrichment channels. Web presence enrichment proved highly effective: 87.6% of processed entities in the validated batch yielded persisted contact signals, confirming its role as the primary contact resolution mechanism. Social media enrichment contributed ICP-relevant profile signals — follower count, engagement rate, and business category— but produced negligible direct contact signal across both the validated batch and the full pilot dataset. This finding provides evidence supporting the free-first enrichment design decision, establishing web presence enrichment as the primary contact resolution pathway and social media enrichment as a scoring and message content source.

C. Lead Inventory and Scoring Distribution

The scoring system operated in fully deterministic mode throughout the pilot period (Phase 3 of the zero-baseline framework), with no trained model contribution active. Scores represent bounded ICP fit and priority rather than calibrated conversion probability. Table II shows the lead-level score distribution across the full pilot dataset. A total of 1,167 leads scored at 0.67 or above, including 466 in the top-tier >0.90 segment, indicating a substantial pool of high-priority opportunities. Approximately half of active leads fell below the 0.40 qualification threshold, reflecting deliberate early-stage filtering of low-fit entities. At the business level, 1,387 unique businesses produced at least one workable lead, and 465 produced at least one top-tier lead. This demonstrates that high-priority opportunities are distributed across a non-trivial portion of the discovered business space rather than concentrated in a narrow subset.

These results validate upstream pipeline performance across discovery, enrichment, and scoring. Design targets for the outreach phase include reducing manual prospecting effort below 5% and improving conversion rate toward 5%; empirical validation is pending regional stabilization.

TABLE II
LEAD-LEVEL SCORE DISTRIBUTION, FULL PILOT DATASET

Score Band	Leads	Share of Active Leads
Above 0.90	466	18.7%
0.67–0.90	701	28.1%
0.40–0.67	75	3.0%
Below 0.40	1,253	50.2%
Unscored	1	<0.1%
Total	**2,496**	

VIII. DISCUSSION

This work illuminates three insights relevant to SE methodology for AI-augmented business process design, alongside a generalizable methodology for deploying AI in data-scarce environments.

First, *stakeholder alignment is a prerequisite for AI system effectiveness*. The principal-agent misalignment identified in requirements engineering would have undermined any technical solution that failed to address it structurally. The centralized queue and shared dashboard resolve this misalignment as design properties.

Second, *objective-tree analysis surfaces shared goals beneath surface conflicts*. The volume-versus-quality conflict dissolved when decomposed into the shared goal of acquisition efficiency. Both groups wanted the same outcome and differed only in proxy metrics because they lacked measurement infrastructure to optimize for the actual goal. This suggests stakeholder conflicts in AI system design often reflect information asymmetry rather than genuinely opposed interests. We recommend this step in any design process to bring stakeholders into alignment and set expectations early on.

Third, *formal SE methodology can substitute for historical data in early AI deployment*. By treating stakeholder heuristics as provisional model parameters (subject to empirical revision), the zero-baseline framework avoids both the paralysis of refusing to deploy without training data and the overconfidence of treating unvalidated heuristics as ground truth. The graduated trust transfer mechanism systematically automates the process.

Limitations include pilot scale and duration constraints, dependency on external API reliability, potential bias in stakeholder-derived qualification criteria, and MENA-specific context that may limit direct transferability. Additionally, initial testing had to be put on hold due to conflict in Zbooni's region upending normal business practices. This extended the timeline for implementation.

IX. CONCLUSION

This paper presented LeadFlow, an AI-driven sales prospecting system whose primary contribution is the methodology used to design it: a systematic approach to deploying AI-assisted qualification in zero-baseline environments. The five-phase framework—elicit stakeholder heuristics, operationalize as features, deploy deterministically, instrument for outcomes, activate ML with performance gates—provides a replicable methodology for organizations that lack historical data to deploy data-driven automation. Requirements traceability from stakeholder conflict through objective-tree analysis through design decision to system component demonstrates that formal SE methodology resolves organizational misalignments that would otherwise undermine system effectiveness. Pilot deployment has confirmed pipeline throughput and scoring coverage; the projected outcomes in are design-stage hypotheses to be tested during outreach deployment. Results will be reported in a follow-up study upon pilot completion. The contribution of this paper is the methodology itself–the five-phase framework and traceability approach are designed to be applicable prior to, and independent of, empirical outcome validation. Our zero-baseline framework and requirements traceability methodology contribute to the SE literature as generalizable tools for AI system design in data-scarce environments.

REFERENCES

[1] Zbooni, "The MENA cCommerce report 2024: How WhatsApp is changing the way we shop," Zbooni, Industry Report, 2024, conducted in partnership with YouGov.

[2] S. Elhajjar, L. Yacoub, and H. Yaacoub, "Automation in business research: systematic literature review," *Information Systems and e-Business Management*, vol. 21, pp. 1–24, 08 2023.

[3] C. Ledro, A. Nosella, and A. Vinelli, "Artificial intelligence in customer relationship management: literature review and future research directions," *Journal of Business & Industrial Marketing*, vol. 37, pp. 48–63, 12 2022.

[4] S. Chatterjee, R. Chaudhuri, and D. Vrontis, "Ai and digitalization in relationship management: Impact of adopting ai-embedded crm system," *Journal of Business Research*, vol. 150, pp. 437–450, 2022.

[5] Salesforce, *Einstein Lead Scoring: Implementation Guide*, Salesforce, Inc., 2024.

[6] HubSpot, *Predictive Lead Scoring: Requirements and Setup*, HubSpot, Inc., 2024.

[7] R. Nygård and J. Mezei, "Automating lead scoring with machine learning: An experimental study," in *Hawaii International Conference on System Sciences*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211552287

[8] 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*, vol. 8, 2025.

[9] A. Kaplan, S. E. Seker, and R. Yoruk, "A review of ai-based business lead generation: Scrapus as a case study," *Frontiers in Artificial Intelligence*, vol. 8, 2025.

[10] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 253–260.

[11] M. Uta, A. Felfernig, V.-M. Le, T. N. T. Tran, D. Garber, S. Lubos, and T. Burgstaller, "Knowledge-based recommender systems: overview and research directions," *Frontiers in Big Data*, vol. 7, 2024.

[12] INCOSE, *INCOSE Systems Engineering Handbook*, 5th ed. Wiley, 2023.

[13] S. Legesse and J. Bicknell, "Ai-enhanced requirements traceability using mbse and large language models," in *Proceedings of the INCOSE AI4SE & SE4AI Research Workshop*, 2025.

[14] J. Nuerk and F. Dařena, "Systems engineering methodology for digital supply chain business models," *Systems Engineering*, vol. 28, no. 3, pp. 411–437, 2025.

[15] S. Kvale, *Interviews: An Introduction to Qualitative Research Interviewing*. Thousand Oaks, CA: Sage Publications, 1996.

[16] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of Financial Economics*, vol. 3, no. 4, pp. 305–360, 1976.

[17] K. M. Eisenhardt, "Agency theory: An assessment and review," *The Academy of Management Review*, vol. 14, no. 1, pp. 57–74, 1989. [Online]. Available: http://www.jstor.org/stable/258191

[18] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[19] R. Kohavi, D. Tang, and Y. Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020.

[20] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web: listen to your customers not to the hippo," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 959–967. [Online]. Available: https://doi.org/10.1145/1281192.1281295

[21] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, rOC Analysis in Pattern Recognition.

[22] D. Hosmer and S. Lemeshow, *Front Matter*. John Wiley & Sons, Ltd, 2000.