ISSN (Online): 3033-0106

BUSINESS & MANAGEMENT COMPASS
Volume: 70 (2026), Issue: 1
DOI: https://doi.org/10.56065/w2xbtf92**, pp. 22-40**

Issues since 2024: https://bi.ue-varna.bg/ojs/index.php/bmc
Previous issues: http://journal.ue-varna.bg/en/home

Econometric Foundations for Multi-Agent Negotiation Systems: A Theoretical Framework for LLM-Driven Strategic Advisor Evaluation

Ehsan HATAMIAN [1]

[1] University of Economics - Varna, Varna, Bulgaria, ehsan.hatamian@ue-varna.bg

JEL: C38, C45, C52, C63, M11

Abstract

Purpose: This study develops a comprehensive econometric framework for evaluating LLM-based multi-agent negotiation systems. The framework targets the Strategic Advisor Agent — an LLM-powered component implementing six specialised methods: context analysis, strategy generation, proposal evaluation, concession recommendation, tactic detection, and recommendation synthesis. *Design/Methodology/Approach:* The theoretical framework extends classical econometric theory to accommodate AI agent characteristics, proposing 24 performance metrics across six methods. A hybrid data methodology combines real European procurement data (2,715 TED notices from 18 countries) with synthetic supplier responses and expert validation. Causal inference approaches (RCT, IV, RDD, DiD) are adapted for AI evaluation with appropriate statistical inference procedures. *Findings:* Traditional AI metrics inadequately capture strategic value creation in negotiation contexts. The proposed Extended Econometric Model $Y_{it} = f(X_{it}, A_{it}, H_{i,t-1}, S_t, \theta_t) + \alpha_i + \lambda_t + \varepsilon_{it}$ accommodates dynamic learning behaviours, temporal dependencies, and emergent capabilities, enabling multi-dimensional performance assessment across technical accuracy, strategic effectiveness, and business impact dimensions. *Research Implications:* The framework provides production-ready econometric specifications for organisations deploying AI negotiation systems. The hybrid validation methodology addresses academic research constraints while maintaining ecological validity through authentic procurement data grounding. *Originality/Value:* A comprehensive econometric framework specifically designed for LLM-based negotiation agents, bridging artificial intelligence, econometrics, and negotiation theory. The 24-metric specification with method-specific performance measures, adapted causal identification strategies, and hybrid validation methodology establishes methodological foundations for rigorous AI agent evaluation in strategic business contexts.

Keywords:
Multi-agent systems,
Large language models,
Econometric evaluation,
Negotiation intelligence,
Strategic advisor agents,
Causal inference,
Performance measurement,
Agentic AI systems.

Published by the University of Economics – Varna

Citation: Hatamian, E. (2026). Econometric foundations for multi-agent negotiation systems: a theoretical framework for LLM-driven strategic advisor evaluation. *Business & Management Compass.* University of Economics Varna, 70 (1), pp. 22 - 40. DOI: 10.56065/w2xbtf92

Received: 18.02.2026.
Revised: 01.03.2026.
Accepted: 02.03.2026.
Published: 31.03.2026.

1. Introduction

1.1 Research Context and Motivation

The integration of Large Language Models (LLMs) with multi-agent systems represents a fundamental transformation in how organisations conduct strategic transactions and manage business relationships (Brynjolfsson & McAfee, 2017). These technological developments have created unprecedented opportunities for developing intelligent negotiation platforms that augment human decision-making while maintaining the nuanced understanding required for successful business outcomes (He et al., 2018). The emergence of specialised AI agents that perform distinct negotiation functions represents a particularly promising approach within this technological evolution, leveraging the principle of task specialisation to enable individual agents to develop expertise in specific domains while coordinating their efforts to achieve superior collective performance (Stone & Veloso, 2000; Wooldridge, 2009). However, this technological sophistication creates a critical challenge: the absence of rigorous econometric frameworks for evaluating AI agent performance in negotiation contexts. Traditional performance metrics—accuracy, response time, precision—fail to capture the multi-dimensional nature of strategic decision-making where success depends on contextual appropriateness, long-term relationship implications, and value creation across multiple stakeholders. This research addresses this fundamental gap by developing a comprehensive econometric framework for the StrategicAdvisorAgent, an LLM-powered agent serving as the central intelligence component of a multi-agent negotiation platform. The StrategicAdvisorAgent provides strategic negotiation support through six specialised methods that collectively address the complete negotiation lifecycle. The first method, `analyze_negotiation_context()`, assesses strategic positions by evaluating power dynamics and identifying critical success factors that shape negotiation approaches. Building upon this contextual foundation, `generate_strategy()` formulates comprehensive negotiation strategies tailored to specific situations, synthesising information about organisational objectives, counterparty characteristics, and environmental constraints. The third method, `analyze_proposal()`, evaluates counterparty proposals against strategic goals and organisational priorities, providing systematic assessment across multiple dimensions including financial implications and risk exposure. The fourth method, `recommend_concessions()`, addresses the tactical challenge of concession timing and magnitude by predicting likely counterparty responses and evaluating impact on overall negotiation outcomes. The fifth method, `detect_negotiation_tactics()`, identifies counterparty tactical manoeuvres including anchoring attempts, deadline pressure, and strategic information disclosure, suggesting appropriate counter-strategies. Finally, `generate_recommendations()` synthesises outputs from all other methods to produce comprehensive, actionable guidance for human negotiators. The detailed architecture and implementation characteristics of these methods are presented in Section 2.

1.2 Research Problem and Gaps

The challenge of measuring and optimising AI-driven negotiation agent performance presents a multifaceted problem that existing evaluation frameworks inadequately address. Traditional performance metrics, designed for simpler automated systems, fail to capture the complex, multi-dimensional nature of strategic decision-making in negotiation contexts (He et al., 2018). The absence of rigorous econometric frameworks specifically tailored for evaluating AI agents in negotiation environments creates several critical gaps that this research addresses.

Table 1

Critical Research Gaps in AI Negotiation Agent Evaluation

Gap	Description	Current Limitation	Framework Contribution
Gap 1	Theoretical Framework	Focus on technical metrics mostly	24 performance metrics across 6 evaluation methods

Gap	Description	Current Limitation	Framework Contribution
Gap 2	Econometric Methodology	Few AI-specific econometric frameworks	Extended econometric models incorporating learning effects
Gap 3	Causal Inference	Confusion between correlation and causation	Application of RCT, IV, RDD, and DiD methodologies
Gap 4	Multi-Dimensional Integration	Single-dimension performance focus	Composite indices integrating multiple evaluation dimensions
Gap 5	Validation Approach	Pure simulation or inaccessible real-world data	Hybrid validation using TED data, synthetic environments, and expert judgment

Source: Author's Compilation.

The first gap concerns the absence of comprehensive theoretical frameworks that capture the multi-dimensional nature of AI agent performance. Current evaluation approaches focus primarily on technical metrics such as response time and basic accuracy measures while neglecting the strategic value creation and business impact that represent the true purpose of these systems (Russell & Norvig, 2016). The second gap involves the lack of econometric methodology specifically designed for AI agent characteristics, including dynamic learning behaviours and temporal dependencies that violate assumptions underlying standard econometric models. The third gap relates to the fundamental challenge of establishing causal relationships between AI interventions and negotiation outcomes, where correlation-based evidence provides insufficient foundation for understanding true agent effectiveness. The fourth gap concerns the integration of performance assessment across multiple dimensions simultaneously, moving beyond single-dimension focus to comprehensive evaluation. The fifth gap addresses the practical challenge of conducting rigorous academic research when complete production negotiation data remains unavailable due to commercial confidentiality constraints.

1.3 Research Questions and Contributions

This research addresses three primary research questions that structure the theoretical and methodological contributions:

RQ1: How can comprehensive econometric principles create meaningful performance measurement frameworks for the StrategicAdvisorAgent's six specialised methods?

RQ2: What causal inference methodologies can establish rigorous relationships between AI interventions and negotiation outcomes in contexts characterised by selection effects and confounding variables?

RQ3: How can hybrid data frameworks combining real procurement data with synthetic responses provide methodologically rigorous validation while respecting confidentiality constraints?

In addressing these questions, this research establishes four primary contributions. First, an Extended Econometric Model accommodating AI agent dynamics through explicit incorporation of learning effects, time-varying parameters, and system state variables. Second, 24 method-specific performance metrics providing granular assessment across the six StrategicAdvisorAgent methods. Third, adapted causal inference methodologies enabling rigorous attribution of outcomes to AI interventions. Fourth, a hybrid validation framework combining real and synthetic data sources that addresses practical constraints while maintaining scientific rigour. The remainder of this paper proceeds as follows. Section 2 presents the StrategicAdvisorAgent architecture and the complete specification of 24 performance metrics. Section 3 develops the econometric framework including the Extended Econometric Model, mathematical specifications, and causal inference methodologies. Section 4 describes the hybrid data methodology. Section 5 provides comprehensive discussion of theoretical and practical implications. Section 6 concludes with contributions summary and future research directions.

2. StrategicAdvisorAgent Architecture

2.1 Multi-Agent Platform Overview

The StrategicAdvisorAgent operates within a comprehensive multi-agent negotiation platform designed for B2B negotiation support (Wooldridge, 2009; Jennings et al., 2001). The platform implements modular design where specialised agents address distinct functional areas while coordinating through standardised protocols and shared knowledge repositories. The architecture implements a layered processing model where raw negotiation data undergoes progressive refinement through multiple analytical stages, enabling the system to handle the complexity inherent in business negotiations while maintaining computational efficiency suitable for real-time decision support (Russell & Norvig, 2016). The StrategicAdvisorAgent serves as the central intelligence component within this architecture, receiving inputs from context providers and delivering strategic recommendations to human negotiators. The agent implements a hybrid intelligence framework that combines deterministic rule-based components for consistent procedural logic with LLM-powered analysis for sophisticated contextual reasoning and nuanced recommendations. This hybrid approach leverages the predictability of rule-based systems for well-defined processes while exploiting the flexibility of large language models for complex analytical tasks requiring natural language understanding and generation. The implementation employs structured prompt templates that guide LLM responses toward consistent output formats, enabling reliable parsing and validation of analytical results. Each method incorporates comprehensive error handling that distinguishes between LLM integration failures, parsing errors, and validation failures, ensuring graceful degradation when components encounter unexpected conditions. The configuration-driven architecture allows adjustment of model parameters, temperature settings, and validation thresholds without code modifications, supporting experimentation and optimisation during deployment.

2.2 Six Specialised Methods and Performance Metrics

The research develops 24 performance metrics distributed across six methods, where each method addresses a distinct aspect of negotiation intelligence. This granular approach enables identification of performance strengths and weaknesses at the component level, supporting targeted optimisation rather than opaque system-level assessment.

Table 2

StrategicAdvisorAgent Methods and Complete Metric Specification

Method	Function	Metric	Description
Method 1: analyze_negotiation_context()	Strategic position assessment; power dynamics evaluation	SPAA	Strategic Position Assessment Accuracy
		PDPA	Power Dynamics Prediction Accuracy
		CDSSR	Context-Driven Strategy Success Rate
		RIE	Risk Identification Effectiveness
Method 2: generate_strategy()	Strategy formulation; negotiation approach selection	SISR	Strategy Implementation Success Rate
		SOC	Strategy–Outcome

Method	Function	Metric	Description
			Correlation
		VCI	Value Creation Index
		CER	Concession Efficiency Ratio
Method 3: analyze_proposal()	Proposal evaluation against strategic and economic goals	PAS	Proposal Alignment Score
		SGC	Strategic Goal Compatibility
		RAVA	Risk-Adjusted Value Assessment
		IFI	Implementation Feasibility Index
Method 4: recommend_concessions()	Concession impact prediction; timing optimisation	CIPA	Concession Impact Prediction Accuracy
		RLE	Reciprocity Likelihood Estimation
		VPR	Value Preservation Ratio
		STO	Strategic Timing Optimisation
Method 5: detect_negotiation_tactics()	Tactical behaviour identification; counter-strategy design	TDA	Tactic Detection Accuracy
		TIA	Tactic Impact Assessment
		CSE	Counter-Strategy Effectiveness
		BPR	Behavioural Pattern Recognition
Method 6: generate_recommendations()	Integrated decision support and guidance synthesis	RAR	Recommendation Adoption Rate
		OIC	Outcome Improvement Correlation
		SCI	Strategic Coherence Index
		USS	User Satisfaction Score

Source: Author's Compilation.

The metrics span three fundamental performance dimensions. Technical accuracy metrics (SPAA, PDPA, TDA, BPR) assess the correctness of agent predictions and classifications against ground truth. Strategic effectiveness metrics (CDSSR, SISR, SOC, VCI, CIPA, CSE) evaluate whether agent outputs contribute to successful negotiation outcomes. Operational efficiency metrics (CER, STO, RAR) measure the practical utility and adoption of agent recommendations in actual negotiation processes. This dimensional structure enables organisations to understand not merely whether the agent produces correct outputs, but whether those outputs translate into improved negotiation results and practical adoption by human negotiators. The following section develops the mathematical specifications

underlying these metrics within an extended econometric framework.

3. Econometric Framework Specification

3.1 Extension of Classical Econometric Theory

Classical econometric theory, developed primarily for analysing human economic behaviour, provides powerful analytical tools for understanding relationships between variables and testing hypotheses about causal effects (Wooldridge, 2010; Greene, 2018). However, AI agents exhibit characteristics that necessitate theoretical extension beyond conventional econometric models. These characteristics include dynamic learning behaviours where performance evolves based on accumulated experience, deterministic yet complex decision processes that produce consistent outputs given identical inputs, and temporal dependencies where current performance depends on historical states in ways that violate standard independence assumptions (Athey & Imbens, 2017). The theoretical framework developed in this research extends classical econometric models to accommodate these AI system characteristics while preserving the rigorous inferential capabilities that make econometric methods valuable for performance evaluation. The extended framework introduces several key modifications to classical specifications that address the fundamental differences between human behaviour and AI system behaviour (Mullainathan & Spiess, 2017).

Core Extended Econometric Model: $Y_{it} = f(X_{it}, A_{it}, H_{i,t-1}, S_t, \theta_t) + \alpha_i + \lambda_t + \varepsilon_{it}$

This specification captures several critical features of AI agent performance that distinguish it from traditional econometric applications. The component X_{it} represents observable contextual variables including deal size, complexity indices, industry classification, and counterparty characteristics that influence both agent behaviour and negotiation outcomes. The component A_{it} captures agent actions and recommendations, representing the treatment of primary interest in causal evaluation. The historical state vector $H_{i,t-1}$accounts for path dependence and learning effects, capturing accumulated knowledge from previous negotiations that influences current performance (Argote, 2012). This explicit modelling of learning represents a fundamental departure from static models that assume constant behavioural parameters. The system-wide state variables S_t capture platform-level factors affecting all negotiations simultaneously, including model updates, computational resource availability, and aggregate learning across the agent population. Time-varying parameters θ_t enable the model to capture evolving agent capabilities as the underlying LLM and agent logic improve through updates and accumulated training. The negotiation-specific fixed effects α_i control for unobserved heterogeneity across negotiations that might otherwise confound treatment effect estimates, while time-specific fixed effects λ_t account for common temporal trends affecting all negotiations (Cameron & Trivedi, 2005). The idiosyncratic error term ε_{it} captures residual variation not explained by model components.

3.2 Mathematical Specifications for Performance Metrics

The 24 performance metrics require precise mathematical specifications that enable consistent measurement and statistical inference. This section presents the complete specification organised by method, where each metric is grounded in established econometric and statistical methodology.

Method 1: analyze_negotiation_context()

The context analysis method provides the foundational intelligence that informs all subsequent strategic decisions. The four metrics associated with this method assess different dimensions of analytical accuracy.

Strategic Position Assessment Accuracy (SPAA) measures how accurately the agent assesses strategic positions compared to expert validation. Following established panel regression methodology (Wooldridge, 2010), the specification captures both cross-sectional and temporal variation:

$$SPAA_{it} = \beta_0 + \beta_1 Context_Complexity_{it} + \beta_2 Historical_Accuracy_{i,t-1} + \beta_3 Market_Conditions_t + \beta_4 Agent_Learning_{it} + \alpha_i + \varepsilon_{it}$$

The coefficient β_1 captures the impact of contextual complexity on accuracy, with expected negative

sign indicating that more complex contexts challenge analytical capabilities. The coefficient β_2 represents path dependence effects where historical performance influences current accuracy, expected to fall between 0 and 1 reflecting persistence without explosive dynamics. The coefficient β_3 captures external market condition effects on assessment accuracy, reflecting how volatile or stable market environments influence the agent's ability to correctly evaluate strategic positions. The learning component β_4 measures accumulated knowledge effects, expected positive to reflect improvement over time as the agent processes more negotiations.

Power Dynamics Prediction Accuracy (PDPA) assesses the agent's ability to correctly anticipate power relationships using a normalised accuracy measure consistent with forecasting evaluation standards (Diebold & Mariano, 2002):

$PDPA = (1/n) \times \Sigma_{i=1}^{n} (1 - |PD_i\text{^agent} - PD_i\text{^actual}|)$

The metric compares agent predictions PD_i^agent on a scale from -1 (complete counterparty advantage) through 0 (balanced) to +1 (full organisational advantage) against actual realised power balance PD_i^actual determined through post-negotiation analysis.

Context-Driven Strategy Success Rate (CDSSR) employs weighted aggregation following multi-criteria decision analysis principles (Keeney & Raiffa, 1993):

$CDSSR = (\Sigma_{i=1}^{n} S_i \times W_i) / (\Sigma_{i=1}^{n} W_i)$

The binary success indicator S_i captures whether strategies based on context analysis achieved objectives, weighted by W_i reflecting strategic importance determined through stakeholder consultation.

Risk Identification Effectiveness (RIE) employs standard classification accuracy metrics from machine learning evaluation (Bishop & Nasrabadi, 2006):

$RIE = (TP + TN) / (TP + TN + FP + FN)$

The components capture true positives (risks correctly identified), true negatives (non-risks correctly classified), false positives (false alarms), and false negatives (missed risks).

Method 2: generate_strategy()

Strategy generation builds upon contextual analysis to develop comprehensive negotiation strategies tailored to specific situations (Lax & Sebenius, 2006). This method synthesises information about organisational objectives, counterparty characteristics, and environmental constraints to formulate strategic approaches that maximise value creation while protecting critical interests. The four metrics assess different aspects of strategy quality and implementation success.

Strategy Implementation Success Rate (SISR) measures the proportion of generated strategies that achieve successful implementation when adopted by human negotiators. The econometric specification extends the basic success rate to include implementation factors:

$SISR_{it} = \gamma_0 + \gamma_1 Strategy_Complexity_{it} + \gamma_2 Implementation_Resources_{it} + \gamma_3 Counterparty_Sophistication_{it} + \gamma_4 Temporal_Pressure_{it} + \varphi(H_{i,t-1}) + u_{it}$

The coefficient γ_1 captures how strategy complexity affects implementation success, with expected negative sign reflecting that more sophisticated strategies present greater implementation challenges for human negotiators. The coefficient γ_2 measures resource availability effects on implementation, expected positive as adequate organisational resources (personnel, time, information systems) facilitate successful strategy execution. The coefficient γ_3 captures counterparty sophistication impacts, expected negative as more sophisticated counterparties can more effectively counter or neutralise strategic approaches. The coefficient γ_4 reflects temporal pressure effects on implementation success, expected negative as time constraints reduce the opportunity for careful strategy deployment. The function $\varphi(H_{i,t-1})$ captures non-linear effects of historical learning accumulation, reflecting how past implementation experiences—both successes and failures—inform current strategy execution through organisational learning mechanisms.

Strategy-Outcome Correlation (SOC) measures alignment between strategy recommendations and achieved outcomes:

$SOC = \Sigma_{i=1}^{k} w_i \times corr(S_i, O)$

The correlation is computed across k strategy dimensions (price, terms, relationship, timing) with

business importance weights w_i. Higher values indicate that recommended strategies consistently predict actual outcomes, reflecting strategic validity.

Value Creation Index (VCI) aggregates value creation across negotiation dimensions to assess whether strategies generate joint gains beyond zero-sum redistribution:

$VCI = (\Sigma_{i=1}^{n} (VC_i \times W_i)) / (\Sigma_{i=1}^{n} W_i)$

The value creation score VC_i captures economic value generated for both parties in negotiation i, weighted by W_i reflecting deal significance. This metric reflects integrative bargaining success rather than purely distributive outcomes.

Concession Efficiency Ratio (CER) measures the efficiency of achieving negotiation objectives relative to concessions made during the process:

$CER = (\Sigma_{i=1}^{n} (O_i / C_i)) / n$

The ratio O_i/C_i captures objective achievement relative to concession costs for each negotiation. Higher values indicate achieving objectives with minimal concession expenditure, reflecting strategic efficiency in resource utilisation.

Method 3: analyze_proposal()

Proposal analysis provides systematic evaluation of proposals received from counterparties (Raiffa, 1982). This method assesses proposals across multiple dimensions including alignment with organisational objectives, compatibility with strategic priorities, financial and operational implications, and risk exposure. The four metrics capture different aspects of analytical accuracy.

Proposal Alignment Score (PAS): measures how accurately the agent assesses the degree to which counterparty proposals align with organisational objectives:

$PAS = (1/n) \times \Sigma_{i=1}^{n} (1 - |A_i\text{^}agent - A_i\text{^}actual|) \times W_i$

Where A_i^agent represents the agent's alignment assessment on a 0-1 scale, A_i^actual represents expert-validated actual alignment, and W_i captures dimension importance weights. The weighted structure ensures that accuracy on strategically important dimensions contributes more to the overall score.

Strategic Goal Compatibility (SGC): evaluates how well the agent identifies compatibility between proposals and strategic goals across multiple goal dimensions:

$SGC = \Sigma_{j=1}^{m} (w_j \times (1/n) \times \Sigma_{i=1}^{n} |G_ij\text{^}agent - G_ij\text{^}actual|)$

Where j indexes m strategic goal dimensions, G_ij^agent represents agent compatibility assessments, and G_ij^actual represents validated compatibility. Lower values indicate better accuracy as the metric captures assessment error.

Risk-Adjusted Value Assessment (RAVA): measures the accuracy of value assessments that incorporate risk considerations:

$RAVA = \Sigma_{i=1}^{n} (Value_Score_i \times Risk_Adjustment_i) / n$

The Value_Score_i captures assessment accuracy while Risk_Adjustment_i weights by risk magnitude, ensuring that accuracy on high-risk assessments contributes appropriately to overall performance.

Implementation Feasibility Index (IFI):

$IFI = \Sigma_{i=1}^{n} Feasibility_Score_i / n$

The Feasibility_Score_i captures accuracy in predicting whether proposed terms can be operationally implemented, considering organisational capabilities, resource requirements, and practical constraints.

Method 4: recommend_concessions()

Concession recommendations address the tactical challenge of determining when, what, and how much to concede during negotiation processes (Mnookin, Peppet, & Tulumello, 2000). This method analyses the strategic value of different concession options, predicts likely counterparty responses, and evaluates the impact of potential concessions on overall negotiation outcomes. The four metrics assess different aspects of concession recommendation quality.

Concession Impact Prediction Accuracy (CIPA): measures how accurately the agent predicts the impact of recommended concessions on negotiation outcomes:

$CIPA = (1/n) \times \Sigma_{i=1}^{n} (1 - |I_i\text{^}pred - I_i\text{^}actual|)$

Where I_i^pred represents the agent's predicted impact of concession i on negotiation outcomes, and I_i^actual represents the actual realised impact determined through post-negotiation analysis. Higher values indicate better prediction accuracy.

Reciprocity Likelihood Estimation (RLE): evaluates the agent's accuracy in predicting whether counterparties will reciprocate concessions:

RLE = (TP + TN) / (TP + TN + FP + FN)

The classification framework captures true positives (correctly predicted reciprocation), true negatives (correctly predicted non-reciprocation), false positives (predicted reciprocation that did not occur), and false negatives (unexpected reciprocation). Accurate reciprocity prediction enables strategic concession timing.

Value Preservation Ratio (VPR): measures how effectively recommended concession strategies preserve organisational value while achieving negotiation progress:

$VPR = \Sigma_{i=1}^{n} (Value_Retained_i / Original_Value_i) / n$

The ratio captures the proportion of original negotiation value retained after concessions, where higher values indicate concession strategies that sacrifice minimal value to achieve agreement.

Strategic Timing Optimisation (STO): assesses the effectiveness of concession timing recommendations:

$STO = \Sigma_{i=1}^{n} Timing_Effectiveness_Score_i / n$

The Timing_Effectiveness_Score_i evaluates whether concessions were recommended at strategically optimal moments—neither prematurely (sacrificing leverage) nor too late (risking negotiation breakdown).

Method 5: detect_negotiation_tactics()

Tactic detection provides real-time analysis of counterparty communication and behaviour to identify tactical manoeuvres and strategic intentions (Bazerman & Neale, 1993). This method processes negotiation communications to recognise patterns indicative of specific tactics including anchoring attempts (Galinsky & Mussweiler, 2001), deadline pressure, and strategic information disclosure. The four metrics assess different aspects of tactical intelligence.

Tactic Detection Accuracy (TDA) measures the accuracy of tactic identification across multiple tactic types using multi-class classification accuracy:

$TDA = (\Sigma_{i=1}^{n} \Sigma_{j=1}^{m} (TP_ij + TN_ij)) / (n \times m \times 4)$

The metric evaluates detection performance across n negotiations and m distinct tactic types. The numerator aggregates true positives (correctly identified tactics) and true negatives (correctly identified non-tactics) across all negotiation-tactic combinations. The denominator normalises by the total number of classification opportunities (n × m × 4, where 4 represents the four possible classification outcomes: TP, TN, FP, FN). Higher values indicate superior tactic detection capability across the full range of tactical manoeuvres encountered in negotiation contexts.

Tactic Impact Assessment (TIA): evaluates the accuracy of agent assessments regarding the potential impact of detected tactics on negotiation outcomes:

$TIA = \Sigma_{i=1}^{n} Impact_Assessment_Accuracy_i / n$

The Impact_Assessment_Accuracy_i measures how closely the agent's predicted tactic impact matches actual impact on negotiation dynamics and outcomes, enabling appropriate response calibration.

Counter-Strategy Effectiveness (CSE): measures the success rate of counter-strategies recommended in response to detected tactics:

$CSE = \Sigma_{i=1}^{n} Counter_Strategy_Success_Rate_i / n$

The Counter_Strategy_Success_Rate_i captures whether recommended counter-strategies successfully neutralised or mitigated the effects of counterparty tactics, reflecting practical defensive value.

Behavioural Pattern Recognition (BPR): assesses the agent's ability to identify recurring behavioural patterns that may indicate tactical intentions:

$BPR = \Sigma_{i=1}^{n} Pattern_Recognition_Score_i / n$

The Pattern_Recognition_Score_i evaluates accuracy in identifying behavioural sequences that precede

or accompany specific tactics, enabling proactive rather than reactive tactical responses.

Method 6: generate_recommendations()

Recommendation synthesis integrates outputs from all other methods to produce comprehensive, actionable guidance for human negotiators (Davenport & Ronanki, 2018). This method combines contextual analysis, strategic planning, proposal evaluation, concession optimisation, and tactical intelligence into coherent recommendation packages. The four metrics assess practical utility and adoption outcomes.

Recommendation Adoption Rate (RAR): measures the proportion of recommendations that human negotiators choose to implement, weighted by recommendation quality and relevance:

$RAR = (\Sigma_{i=1}^{n} A_i \times Q_i \times R_i) / (\Sigma_{i=1}^{n} Q_i \times R_i)$

Where A_i represents the binary adoption indicator (1 if adopted, 0 otherwise), Q_i represents quality score on a 0-1 scale, and R_i represents relevance score on a 0-1 scale. The weighting ensures that adoption of high-quality, highly relevant recommendations contributes more to the overall score.

Outcome Improvement Correlation (OIC): measures the relationship between recommendation adoption and negotiation outcome improvements:

$OIC = corr(R, \Delta O)$

The correlation captures whether adoption of recommendations (R) consistently associates with positive outcome changes (ΔO). Positive correlation indicates that following recommendations leads to improved results, validating practical recommendation value.

Strategic Coherence Index (SCI): evaluates the internal consistency and logical coherence of recommendation packages:

$SCI = \Sigma_{i=1}^{n} Coherence_Score_i / n$

The Coherence_Score_i assesses whether recommendations across different dimensions (strategy, concessions, tactics) form a consistent, non-contradictory guidance package that human negotiators can implement coherently.

User Satisfaction Score (USS): captures subjective assessments of recommendation quality from human negotiators who received the guidance:

$USS = \Sigma_{i=1}^{n} Satisfaction_Rating_i / n$

The Satisfaction_Rating_i reflects user-reported satisfaction with recommendation clarity, relevance, timeliness, and practical utility, providing direct feedback on perceived recommendation value.

3.3 Causal Inference Methodologies

Establishing causal relationships between AI interventions and outcomes represents a fundamental challenge that distinguishes rigorous evaluation from correlation-based assessment (Angrist & Pischke, 2009). This research adapts four complementary causal inference methodologies for AI system contexts, each addressing different identification challenges and providing robustness checks for causal effect estimates. Randomised Control Trials (RCT) provide the gold standard for causal inference by randomly assigning negotiations to treatment (with StrategicAdvisorAgent support) and control (without agent assistance) conditions (Duflo, Glennerster, & Kremer, 2007). Random assignment ensures statistical equivalence between groups across both observed and unobserved characteristics, eliminating selection bias. Stratified randomisation improves efficiency by balancing treatment assignment across negotiation value categories, industry sectors, and complexity levels. Instrumental Variables (IV) approaches enable causal inference from observational data when randomised experiments are infeasible (Angrist & Pischke, 2009). Valid instruments must satisfy relevance (strong prediction of treatment) and exclusion restriction (affecting outcomes only through treatment). System availability during planned or unplanned platform downtime provides quasi-experimental variation, while organisational policy thresholds create discontinuous changes in usage probability. Regression Discontinuity Design (RDD) exploits discontinuous changes in treatment assignment at known threshold values (Lee & Lemieux, 2010). For example, policies requiring agent support for negotiations exceeding €1M create sharp cutoffs enabling comparison of otherwise similar negotiations just above

and below the threshold. Difference-in-Differences (DiD) exploits variation in treatment timing across organisational units to identify causal effects while controlling for time-invariant differences between groups and common temporal trends (Angrist & Pischke, 2009). Staggered deployment across departments enables identification under the parallel trends assumption.

3.4 Statistical Inference Procedures

Rigorous hypothesis testing requires accurate estimation of standard errors that account for potential dependencies in the data generating process (Wooldridge, 2010; Cameron & Trivedi, 2005). AI performance data exhibits several characteristics that complicate standard inference procedures. Temporal correlation arises because agent performance in nearby time periods reflects persistent underlying capabilities and similar market conditions. Hierarchical clustering occurs when negotiations are nested within organisations, departments, or personnel. Multiple hypothesis testing across 24 metrics increases the probability of spurious discoveries. Heteroskedasticity and Autocorrelation Consistent (HAC) Standard Errors address temporal dependencies through the Newey-West estimator (Newey & West, 1986). The approach modifies the variance-covariance matrix using kernel weighting that down-weights correlations at longer lags, with bandwidth selected through established plug-in methods based on sample size. This correction proves essential for AI performance data where day-to-day performance exhibits positive serial correlation reflecting stable underlying system capabilities. Cluster-Robust Standard Errors address cross-sectional dependencies arising from nested data structures (Cameron & Trivedi, 2005). When negotiations cluster within departments or organisations, standard errors must account for within-cluster correlation to avoid overstated precision. Conventional asymptotic validity requires a sufficient number of clusters, typically 50 or more. Multiple Testing Corrections become essential with the 24-metric framework to control for increased false discovery probability (Benjamini & Hochberg, 1995). Three complementary approaches provide different trade-offs between Type I and Type II error control. Bonferroni correction ($\alpha = \alpha/24 \approx 0.0021$ for $\alpha = 0.05$) provides conservative family-wise error rate control but may sacrifice statistical power. Benjamini-Hochberg procedure controls false discovery rate while maintaining greater power for exploratory analysis. Holm-Bonferroni step-down method provides compromise between power and control through sequential testing. Power Analysis determines sample size requirements for detecting effects

Table 3

Causal Inference Methodologies

Methodology	Application Context	Key Assumption	Identification Strategy
RCT (Randomized Controlled Trial)	Pilot implementations and A/B testing.	Random assignment of subjects.	Experimental variation eliminates selection bias by design.
IV (Instrumental Variables)	Observational data with unobserved confounding.	Valid exclusion restriction (instrument only affects outcome via treatment).	Uses system availability or exogenous shocks as an instrument.
RDD (Regression Discontinuity)	Threshold-based access or eligibility rules.	Local continuity (subjects just above/below threshold are similar).	Exploits the sharp discontinuity at deal value or eligibility thresholds.
DiD (Difference-in-Differences)	Staggered deployment or policy changes over time.	Parallel trends (groups would have evolved similarly without treatment).	Before/after comparison across treated and control groups.

Source: Author's Compilation.

of practical importance with acceptable probability. The clustering structure characteristic of negotiation data reduces effective sample size through the design effect $(1 + (k-1)\rho)$, where k represents average cluster size and ρ represents intra-cluster correlation (Kozlowski et al., 2013). Detecting small effects (d = 0.20) with 80% power under moderate clustering (ρ = 0.10, k = 20) requires approximately 135 clusters totalling 2,700 observations. $n = 2 \times ((z_{1-\alpha/2} + z_{1-\beta}) / d)^2 \times (1 + \rho(m-1))$

4. Hybrid Data Methodology

4.1 Rationale and Three-Component Framework

The development of rigorous econometric frameworks for AI agent evaluation confronts a fundamental practical challenge: complete production negotiation data typically remains inaccessible due to commercial confidentiality constraints. Organisations legitimately protect sensitive negotiation outcomes, counterparty information, and strategic positioning details that would be essential for comprehensive empirical validation. This constraint is not unique to negotiation systems—similar challenges arise across AI applications in finance, healthcare, and legal domains where operational data contains proprietary or sensitive information. This research addresses this challenge through a novel hybrid data approach that combines authentic institutional data with controlled synthetic components and expert validation, following established principles from statistical disclosure limitation and data augmentation literature (Little & Rubin, 2019; Rubin, 1993). The methodology preserves essential statistical properties while enabling rigorous analysis within confidentiality constraints.

Component 1: Real TED Procurement Data (Upstream). The first component comprises 2,715 authentic European public procurement Request for Quotations (RFQs) obtained from the Tender Electronic Daily (TED) database. These documents cover software package procurement (CPV=48000000) from 18 European countries spanning 2018-2024. The TED corpus provides genuine institutional requirements, authentic administrative patterns, and validated cross-national characteristics that ground the synthetic components in real-world procurement practices (Hatamian, 2025) .The public nature of these documents eliminates confidentiality constraints while preserving the institutional authenticity essential for ecological validity.

Component 2: Synthetic Supplier Responses (Downstream). The second component comprises empirically grounded supplier proposals generated based on established negotiation research parameters (Kersten & Lai, 2007; Baarslag et al., 2016). Meta-analytic foundations ensure that parameter distributions—including learning effects, anchoring impacts, and outcome distributions—align with published research findings. Synthetic generation follows established principles for creating datasets that preserve inferential validity while enabling controlled experimentation (Jordon et al., 2018; Nikolenko, 2021). The synthesis process incorporates cultural archetype characteristics discovered through clustering analysis of the TED data, ensuring that generated supplier responses reflect realistic institutional variation.

Component 3: Expert Human Validation (Ground Truth). The third component comprises performance annotations from professional procurement and negotiation experts who provide ground truth for agent performance assessment (Nadler, Thompson, & van Boven, 2003). Expert validation establishes rigorous performance benchmarks against which agent outputs are measured, achieving 95.4% inter-rater agreement across a validation subset. This human-in-the-loop approach addresses the fundamental challenge that negotiation quality cannot be assessed through purely automated metrics—strategic appropriateness requires professional judgment.

4.2 Contextual Intelligence Capabilities

The StrategicAdvisorAgent incorporates multiple contextual intelligence tools that enhance negotiation understanding beyond basic transaction parameters. Industry-specific adaptation through configurable parameters enables the agent to adjust analytical approaches for sector-specific negotiation norms.

Regulatory context awareness supports compliance requirements that vary across jurisdictions and industries. Market condition integration incorporates external environmental factors that influence negotiation dynamics and counterparty behaviour. Among these capabilities, cultural pattern recognition for cross-border negotiations represents one tool that supports international negotiations. Prior research (Hatamian, 2025) established a Cultural Intelligence Extraction System for European procurement, extracting 101 cultural parameters across 12 categorical dimensions from multilingual procurement documents. This research demonstrated systematic cross-national variation in administrative practices across 18 European countries, with multi-algorithm clustering (K-means, Hierarchical Agglomerative, DBSCAN) identifying distinct cultural archetypes with strong validation metrics. This capability enhances the `analyze_negotiation_context()` method by providing country-specific pattern awareness when relevant to negotiations with European counterparties. The cultural intelligence extraction system achieved 95.4% agreement with expert human annotations, providing validated cultural context for the hybrid data methodology.

4.3 Validation Framework

The validation framework addresses multiple dimensions of methodological rigour to ensure that findings generalise beyond the specific synthetic data instances. Convergent Validity assesses whether synthetic data parameters produce distributions consistent with meta-analytic findings from published negotiation research. Kolmogorov-Smirnov tests compare synthetic distributions against target distributions derived from Hüffmeier et al. (2014), Sharma, Bottom, and Elfenbein (2013), and other meta-analytic sources. Complexity-performance relationships in synthetic data should exhibit correlations ($r \approx -0.3$ to -0.5) consistent with empirical research (Thompson & Hastie, 1990). Discriminant Validity confirms that the framework distinguishes between high-performing and low-performing agent configurations. This requires that known variations in agent parameters produce detectable performance differences across the 24 metrics. External Validity is enhanced through real TED data grounding, ensuring that synthetic scenarios reflect authentic institutional procurement contexts rather than artificial constructs. The coverage of 18 European countries across six years provides substantial cross-national and temporal variation. Reproducibility is ensured through documented procedures that enable academic replication, including parameter specifications, random seeds, and validation protocols.

5. Discussion

5.1 Theoretical Contributions

This research advances theoretical understanding of AI agent evaluation in strategic business contexts through several interconnected contributions that collectively address fundamental gaps in the intersection of artificial intelligence, econometrics, and negotiation theory. The proposed framework extends classical econometric theory to accommodate AI agent characteristics that violate standard assumptions underlying traditional models. The explicit inclusion of historical state vectors ($H_{i,t-1}$) represents a fundamental departure from static models, capturing the path dependence and learning effects that characterise intelligent systems. Unlike human decision-makers whose behaviour may exhibit inconsistency or bounded rationality, AI agents demonstrate systematic learning patterns that the model specification directly incorporates. The time-varying parameters (θ_t) capture evolving capabilities as underlying models improve through updates and accumulated training, while negotiation-specific fixed effects (α_i) control for unobserved heterogeneity that would otherwise confound treatment effect estimates. This theoretical extension provides foundations for evaluating systems that learn, adapt, and exhibit emergent behaviours—characteristics fundamentally different from the static agents assumed in classical econometric applications. The 24-metric specification across six methods demonstrates that AI negotiation agent performance cannot be adequately captured through singular metrics, regardless of how sophisticated those individual metrics might be. The framework's dimensional structure—spanning technical accuracy, strategic effectiveness, and operational

efficiency—provides theoretical grounding for comprehensive evaluation that recognises how AI value creation occurs through complex interactions that simple aggregate metrics cannot capture. This multi-dimensional approach reflects the reality that an agent might produce technically accurate predictions that nonetheless fail to translate into improved negotiation outcomes, or might generate strategically sound recommendations that human negotiators do not adopt. Understanding these distinctions requires the granular, multi-metric framework this research develops. The adaptation of established causal inference methodologies (RCT, IV, RDD, DiD) for AI system contexts addresses fundamental attribution challenges that distinguish rigorous evaluation from correlation-based assessment. Traditional experimental designs face practical constraints in business negotiation contexts where randomisation may conflict with commercial imperatives, and observational data exhibits complex selection effects. The adapted methodologies exploit natural variation while maintaining identification validity, extending causal inference theory to emerging technological domains characterised by adaptive systems, complex intervention structures, and human-AI collaboration patterns.

5.2 Methodological Contributions

The method-specific econometric specifications represent a departure from existing AI evaluation approaches that treat systems as black boxes producing aggregate outputs. By providing mathematical specifications for each of the StrategicAdvisorAgent's six functions, the framework enables identification of performance strengths and weaknesses at the component level. Organisations can identify which specific methods require improvement and allocate development resources accordingly, rather than attempting to optimise an opaque system-level performance measure. This granular approach supports targeted development efforts focused on underperforming components rather than wholesale system redesign. The three-component hybrid data framework provides a methodological template for conducting rigorous academic research on AI systems prior to production deployment. This approach addresses the fundamental tension between scientific rigour and practical constraints that characterises much applied AI research. By combining authentic institutional data (TED procurement notices) with controlled synthetic components (supplier responses) and expert validation (ground truth annotations), the methodology maintains ecological validity through real data grounding while enabling the controlled experimentation necessary for causal inference. This template generalises beyond negotiation systems to other domains where production data remains inaccessible. The statistical inference adaptations specify procedures addressing characteristics of AI performance data that violate standard assumptions. HAC standard errors for temporal correlation, cluster-robust estimators for hierarchical structures, and multiple testing corrections for the 24-metric battery provide methodologically rigorous foundations for valid statistical inference. These specifications ensure that performance claims are statistically defensible, avoiding the overconfident conclusions that result from applying standard errors to data exhibiting complex dependencies.

5.3 Practical Implications

The framework provides organisations deploying AI negotiation systems with comprehensive evaluation tools that support evidence-based decision-making. The 24 metrics enable assessment of AI investments across multiple dimensions, supporting resource allocation decisions and continuous improvement initiatives. Rather than relying on intuition or anecdotal evidence, organisations can systematically evaluate whether AI systems deliver promised value and identify specific areas requiring attention. While positioned as a theoretical framework, the specifications are designed for production implementation. The Python-based analytical approach, leveraging NumPy for numerical computation, SciPy for statistical functions, Pandas for data manipulation, Statsmodels for econometric modelling, and Scikit-learn for machine learning applications, enables direct translation of theoretical specifications into operational evaluation systems with real-time monitoring capabilities. The framework's modular structure, with industry-configurable parameters and context-specific weighting, enables adaptation across diverse business sectors. The integration with contextual intelligence

capabilities demonstrates applicability to complex negotiation environments where multiple factors influence outcomes, from regulatory requirements to cultural patterns to market conditions.

5.4 Limitations

This research acknowledges several important limitations that bound the scope of conclusions and suggest priorities for future investigation. The framework has not been validated with production deployment data, as all claims are appropriately limited to theoretical specifications and prospective development guidance. Empirical validation with real negotiation outcomes represents a critical priority for future research. This limitation reflects the current state of AI negotiation system development, where prototype implementations demonstrate functional capabilities but production-scale deployment with measurable outcomes remains limited. The theoretical framework positioning is appropriate given this developmental status. While the hybrid methodology provides rigorous validation capabilities, synthetic data cannot fully replicate the complexity of real-world negotiation dynamics. Parameter distributions derived from meta-analytic research may not capture all context-specific variations encountered in production environments. The synthetic components, despite empirical grounding, represent approximations of actual supplier behaviour that may miss subtle strategic patterns. The framework draws primarily from European procurement practices through the TED data foundation. Applicability to non-European contexts, particularly markets with substantially different institutional environments such as Asian, Middle Eastern, or African procurement systems, requires additional validation and potential adaptation. The methodological frameworks developed here can be adapted to other cultural contexts, but empirical validation of performance in diverse settings remains a priority. Academic research timelines limit longitudinal validation of learning effects and temporal dynamics. Extended observation periods are necessary to validate the temporal components of the framework, including historical state vectors and time-varying parameters that capture agent evolution over months or years rather than the shorter periods typically available for academic investigation.

5.5 Future Research Directions

The research agenda proceeds across three temporal horizons, each addressing distinct aspects of framework development and validation. Immediate priorities focus on establishing empirical foundations for the theoretical framework. Empirical validation with production deployment data represents the most pressing need, requiring industry partnerships that provide access to real negotiation outcomes while protecting commercial confidentiality. Concurrently, the hybrid data methodology should be extended to additional procurement domains beyond software (CPV=48000000), including construction, healthcare equipment, and professional services, to demonstrate cross-sector applicability. Cross-cultural validation in non-European market contexts also demands attention, particularly in Asian institutional environments where negotiation norms differ substantially from the European procurement practices that ground the current framework. Medium-term development addresses deeper questions of system dynamics and human-AI interaction. Longitudinal studies capturing learning effects over extended periods will enable validation of time-varying parameters and historical state vectors that currently remain theoretically specified but empirically untested. Investigation of human-AI collaboration patterns in recommendation adoption represents another priority, seeking to understand why negotiators accept some recommendations while rejecting others—insights essential for optimising recommendation presentation and improving adoption rates. Real-time monitoring dashboard implementation will support continuous evaluation during production operation, enabling organisations to track performance trends and identify degradation before it impacts business outcomes. Long-term extensions position the framework within broader organisational contexts. Integration with enterprise-wide business intelligence and performance measurement systems will connect AI negotiation evaluation to strategic decision-making processes. Development of automated metric calculation pipelines will reduce manual effort in performance assessment, enabling scalable evaluation across large negotiation portfolios. Extension to multi-agent coordination evaluation represents the frontier

challenge, moving beyond single-agent performance to address how multiple agents collaborate in complex negotiation environments where distributed intelligence creates emergent capabilities exceeding individual agent contributions.

6. Conclusion

This research presents a comprehensive econometric framework for evaluating LLM-based multi-agent negotiation systems, specifically addressing the StrategicAdvisorAgent as the central intelligence component of AI-driven negotiation platforms. The framework addresses critical gaps in AI performance measurement methodology and provides foundations for rigorous evaluation in strategic business contexts.

6.1 Summary of Contributions

Contribution 1: Extended Econometric Model. The framework extends classical econometric theory to accommodate AI agent characteristics—dynamic learning behaviours, temporal dependencies, and emergent capabilities. The model specification $Y_{it} = f(X_{it}, A_{it}, H_{i,t-1}, S_t, \theta_t) + \alpha_i + \lambda_t + \varepsilon_{it}$ provides theoretical foundations for rigorous performance measurement while maintaining analytical tractability. This extension addresses the fundamental mismatch between static traditional models and dynamic AI system behaviour, enabling evaluation of systems that learn and evolve over time.

Contribution 2: 24-Metric Performance Framework. The specification of 24 performance metrics across six specialised methods enables multi-dimensional assessment that captures technical accuracy, strategic effectiveness, and business impact simultaneously. This granular approach—with metrics including SPAA, PDPA, SISR, SOC, PAS, CIPA, TDA, RAR, and others—supports targeted optimisation and evidence-based resource allocation. The method-specific structure enables organisations to identify precisely which agent capabilities require improvement rather than optimising opaque system-level measures.

Contribution 3: Adapted Causal Inference Methodologies. The adaptation of RCT, IV, RDD, and DiD methodologies for AI system contexts addresses fundamental attribution challenges in establishing causal relationships between AI interventions and negotiation outcomes. These adaptations extend causal inference theory to emerging technological domains characterised by adaptive systems, complex intervention structures, and human-AI collaboration patterns that differ fundamentally from traditional causal inference applications.

Contribution 4: Hybrid Data Validation Methodology. The three-component framework combining real TED procurement data (2,715 notices from 18 countries), synthetic supplier responses grounded in meta-analytic research, and expert human validation provides a robust validation approach that addresses practical constraints while maintaining scientific rigour. This methodology provides a template for conducting rigorous academic research on AI systems when production data remains inaccessible due to confidentiality constraints.

6.2 Significance

The framework addresses a critical gap in the application of multi-agent systems to business contexts. As organisations increasingly deploy AI for strategic functions—negotiation, procurement, business development, strategic planning—rigorous evaluation methodologies become essential for evidence-based decision-making. Without such frameworks, organisations cannot reliably assess AI investment returns, identify improvement opportunities, or compare alternative system configurations. Decisions about AI deployment risk become matters of faith rather than evidence. The convergence of large language models, multi-agent systems, and business negotiation creates unprecedented opportunities for intelligent automation of strategic functions. However, realising these opportunities requires moving beyond simplistic technical metrics toward comprehensive evaluation frameworks that

capture the multi-dimensional nature of strategic value creation. The contribution of this research lies not in the individual metrics or methods, but in the integrated framework that connects AI system capabilities to measurable business outcomes through rigorous econometric methodology.

6.3 Theoretical Framework Position

This research explicitly positions itself as a theoretical framework paper. The StrategicAdvisorAgent's six methods have been implemented at the prototype level with full LLM integration, prompt engineering, and validation pipelines. However, comprehensive empirical validation with production deployment data remains a priority for future research. Consequently, all claims are appropriately limited to theoretical specifications and prospective development guidance. This positioning reflects the current state of AI negotiation system development. Prototype implementations demonstrate functional capabilities, but production-scale deployment with measurable negotiation outcomes remains limited across the industry. The framework provides robust foundations for future empirical validation when such deployment data becomes accessible through industry partnerships or phased production rollouts.

6.4 Closing Statement

The econometric framework presented in this research establishes methodological foundations for rigorous AI agent evaluation in strategic business contexts. By bridging artificial intelligence, econometrics, and negotiation theory, the framework enables organisations to move from intuition-based AI deployment toward evidence-based evaluation and optimisation. As LLM-powered multi-agent systems become increasingly prevalent in business applications, such frameworks become essential infrastructure for responsible and effective AI adoption. The 24 metrics, extended econometric model, adapted causal inference methodologies, and hybrid validation approach provide the comprehensive toolkit necessary for this evaluation challenge—transforming AI performance assessment from art to science.

References

1. Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.
2. Argote, L. (2012). Organizational Learning: Creating, Retaining and Transferring Knowledge. Springer Science & Business Media.
3. Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3-32.
4. Baarslag, T., Hendrikx, M., Hindriks, K., & Jonker, C. (2016). Learning about the opponent in automated bilateral negotiation: A comprehensive survey. Autonomous Agents and Multi-Agent Systems, 30(5), 849-898.
5. Bazerman, M. H., & Neale, M. A. (1993). Negotiating Rationally. The Free Press.
6. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289-300.
7. Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern Recognition and Machine Learning. Springer.
8. Brynjolfsson, E., & McAfee, A. (2017). The business of artificial intelligence. Harvard Business Review, 95(4), 3-11.
9. Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: Methods and Applications. Cambridge University Press.
10. Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108-116.
11. Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. Journal of Business

& Economic Statistics, 20(1), 134-144.
12. Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. Handbook of Development Economics, 4, 3895-3962.
13. Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. Journal of Personality and Social Psychology, 81(4), 657-669.
14. Greene, W. H. (2018). Econometric Analysis (8th ed.). Pearson.
15. Hatamian, E. (2025). Formal harmonisation, informal divergence: Administrative culture in European public procurement. Business & Management Compass, 69(4), 18-43. DOI: 10.56065/kq2wma16
16. He, H., Chen, D., Balakrishnan, A., & Liang, P. (2018). Decoupling strategy and generation in negotiation dialogues. Proceedings of EMNLP 2018, 2333-2343.
17. Hüffmeier, J., Freund, P. A., Zerres, A., Backhaus, K., & Hertel, G. (2014). Being tough or being nice? A meta-analysis on the impact of hard- and softline strategies in distributive negotiations. Journal of Management, 40(3), 866-892.
18. Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Wooldridge, M. J., & Sierra, C. (2001). Automated negotiation: Prospects, methods and challenges. Group Decision and Negotiation, 10(2), 199-215.
19. Jordon, J., Yoon, J., & van der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. International Conference on Learning Representations.
20. Keeney, R. L., & Raiffa, H. (1993). Decisions with Multiple Objectives: Preferences and Value Trade-offs. Cambridge University Press.
21. Kersten, G. E., & Lai, H. (2007). Negotiation support and e-negotiation systems: An overview. Group Decision and Negotiation, 16(5), 411-435.
22. Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. Organizational Research Methods, 16(4), 581-615.
23. Lax, D. A., & Sebenius, J. K. (2006). 3-D Negotiation: Powerful Tools to Change the Game. Harvard Business School Press.
24. Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of Economic Literature, 48(2), 281-355.
25. Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data (3rd ed.). John Wiley & Sons.
26. Mnookin, R. H., Peppet, S. R., & Tulumello, A. S. (2000). Beyond Winning: Negotiating to Create Value in Deals and Disputes. Harvard University Press.
27. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.
28. Nadler, J., Thompson, L., & van Boven, L. (2003). Learning negotiation skills: Four models of knowledge creation and transfer. Management Science, 49(4), 529-540.
29. Newey, W. K., & West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703-708.
30. Nikolenko, S. I. (2021). Synthetic Data for Deep Learning. Springer.
31. Raiffa, H. (1982). The Art and Science of Negotiation. Harvard University Press.
32. Rubin, D. B. (1993). Statistical disclosure limitation. Journal of Official Statistics, 9(2), 461-468.
33. Russell, S., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3rd ed.). Pearson.
34. Sharma, S., Bottom, W. P., & Elfenbein, H. A. (2013). On the role of personality, cognitive ability, and emotional intelligence in predicting negotiation outcomes: A meta-analysis. Organizational Psychology Review, 3(4), 293-336.
35. Stone, P., & Veloso, M. (2000). Multiagent systems: A survey from a machine learning

perspective. Autonomous Robots, 8(3), 345-383.
36. Thompson, L., & Hastie, R. (1990). Social perception in negotiation. Organizational Behavior and Human Decision Processes, 47(1), 98-123.
37. Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data (2nd ed.). MIT Press.
38. Wooldridge, M. (2009). An Introduction to MultiAgent Systems (2nd ed.). John Wiley & Sons.