LEVERAGING LARGE LANGUAGE MODELS (LLMS) TO AUTOMATE THE ANALYTIC HIERARCHY PROCESS (AHP) FOR SAAS PROCUREMENT

Author 1: Cha Hee Park[1]
Author 2. Keith Saniga[2]

Highlights

- Large Language Models (LLMs) can be used to partially automate the Analytic Hierarchy Process (AHP) for SaaS supplier evaluation.
- Evaluation framework incorporates five key criteria: Total Cost Ownership, Implementation Timeline, Deployment Type, Integration, and Vendor Reliability and Support.
- Repeated trials demonstrate the consistency and reliability of LLM-based assessments of vendor proposals.
- LLMs significantly reduce time and resources required for proposal evaluations.
- Findings support the potential of LLMs to automate and improve decision-making in procurement processes.

ABSTRACT

This study explores the use of Large Language Models (LLMs) to enhance an Analytic Hierarchy Process (AHP) for evaluating SaaS supplier proposals. Traditional assessments are often slow and biased, but LLMs offer a faster, objective alternative. The evaluation framework includes five criteria: Total Cost of Ownership, Implementation Time, Deployment Type, Integration with Existing Systems, and Supplier Reliability and Support, with weights refined by a procurement specialist. An LLM was used to assess synthetic vendor proposals for their quality along these criteria. Consistency of the LLM's responses was tested through repeated trials to calculate a reliable "true mean" of LLM assessments. Results from testing on 10 synthetic proposals indicate that LLMs can provide efficient, consistent, and unbiased proposal evaluations, supporting their potential in automating decision-making processes.

Keywords: AHP, LLMs, Procurement Process, Proposal Assessment, Decision Making.

1. Introduction

[1] Cha Hee Park, MS, University of Pittsburgh, PA 15219, USA, e-mail: chahee.p@outlook.com .

[2] Keith Saniga, MBA, University of Pittsburgh, PA 15219, USA, e-mail: sanigak@outlook.com .

The rapid expansion of software solutions and SaaS offerings has made evaluating supplier proposals a critical yet increasingly challenging task for organizations. The traditional approach of manually assessing proposals is often time-consuming, subjective, and prone to inconsistencies, particularly when dealing with dozens or even hundreds of responses to an open Request for Proposals (RFP). This research seeks to address these challenges by investigating the use of Large Language Models (LLMs) as a tool for streamlining and improving the proposal evaluation process.

The motivation behind this study is rooted in the potential of LLMs to provide efficient, scalable, and objective assessments. By leveraging an Analytic Hierarchy Process (AHP) framework, this research evaluates how LLMs can be used to systematically review and score supplier proposals based on key criteria, offering a more consistent and unbiased approach than human evaluators. The main research question driving this study is whether LLMs can reliably assist in the automated analysis of proposals while maintaining the rigor and fairness necessary for informed decision-making.

This work is significant as it addresses the dual challenges of scale and objectivity in the supplier selection process. By demonstrating the ability of LLMs to handle high volumes of proposals quickly and without inherent bias, this research aims to pave the way for more efficient and equitable software procurement practices, allowing organizations to focus their efforts on strategic decision-making rather than repetitive and resource-intensive evaluation processes.

2. Literature Review

Research on combining Analytic Hierarchy Process (AHP) with Large Language Models (LLMs) highlights various benefits and challenges in decision-making applications.

Lu et al. demonstrated that LLMs, such as GPT-4, can be effectively integrated with AHP to evaluate open-ended responses across multiple criteria, aligning with human judgments more closely than baseline models. This method, which involved criteria generation and evaluation phases, improved the objectivity and comprehensiveness of the assessments by incorporating diverse evaluation perspectives (Lu et al., 2024).

Zhao et al. investigated the reliability of LLMs in ordinal preference formation and uncovered significant issues with bias and consistency. Their analysis, focused on transitivity and independence from irrelevant alternatives (IIA), found that LLMs often struggle with maintaining coherent rankings when new options are introduced. This indicates a need for further refinement in their use for multi-criteria decision-making frameworks such as AHP (Zhao et al., 2024).

Hallikainen and Kivijärvi explored the strategic importance and inherent risks of information systems (IS) procurement strategies using AHP. They emphasized the value of integrating both quantitative and qualitative evaluation criteria, including subjective measures, for a robust decision-making model. Their study's findings on the adaptability of AHP for complex IS investments align with the broader utility of AHP in assessing supplier proposals and ensuring consistency through sensitivity analyses (Hallikainen & Kivijärvi, 1999).

These studies collectively underscore the potential and limitations of combining AHP with LLMs. While LLMs can facilitate enhanced evaluations in open-ended and subjective decision-making scenarios, their susceptibility to inconsistency and positional bias necessitates ongoing refinement.

3. Hypotheses/Objectives

The integration of large language models (LLMs) for scoring SaaS proposals based on specific evaluation criteria presents a promising advancement in the procurement and decision-making process. The hypothesis driving this research is as follows:

The use of LLMs for scoring SaaS proposals along the criteria of Total Cost of Ownership, Implementation Time, Deployment Type, Integration with Existing Systems, and Vendor Reliability and Support will provide accurate and consistent assessments that are comparable to or better than human evaluations, while significantly reducing the time and resources required.

Objectives

1. To validate that LLMs can score SaaS proposals along the criteria of Total Cost of Ownership, Implementation Time, Deployment Type, Integration with Existing Systems, and Vendor Reliability and Support with accuracy and consistency comparable to human evaluators.
2. To determine the optimal number of trials required for LLM-based scoring to achieve reliable and stable mean results that minimize variability.
3. To demonstrate that the use of LLMs for proposal scoring reduces the time and effort involved in the evaluation process while maintaining or enhancing the quality of assessments.

4. Research Design/Methodology

The analysis focused on assessing the consistency of scoring for predefined criteria using measurements taken over multiple trails. Given the nature of using large language models (LLMs) for automated scoring, there are inherent limitations in consistency and precision due to potential model variability and context sensitivity. However, these limitations can be mitigated through a robust methodology involving repeated trials and the aggregation of results.

Process Overview

Data Collection: Measurements were collected across 50 trials, evaluating five distinct criteria: Total Cost of Ownership, Implementation Time, Deployment Type, Integration with Existing Systems, and Vendor Reliability and Support.

Criteria: These criteria were selected based on insights gained from recent engagements with business teams, reflecting their key priorities when selecting a CRM solution. Each

criterion addresses specific organizational needs, ensuring a comprehensive and balanced evaluation:

- C1 - Total Cost of Ownership (TCO):
  TCO provides a clear picture of both initial and ongoing costs, helping to assess the long-term financial impact. A thorough evaluation of setup, licensing, and maintenance costs ensures alignment with the project budget and helps avoid unexpected expenses.
- C2 - Implementation Timeline:
  The implementation timeline is critical for ensuring rapid deployment while minimizing disruptions. Supplier project plans must be reviewed to ensure they align with overall project timelines and avoid costly delays.
- C3 - Deployment Type (Cloud vs. On-Premises):
  Deployment type influences infrastructure needs, security, and scalability. Cloud-based solutions offer flexibility and potential cost savings, while on-premises deployments may be preferred for greater data control and compliance with regulations.
- C4 - Integration with Existing Systems:
  Integration is vital for ensuring that the CRM solution works seamlessly with existing business systems. Effective integration reduces data silos, enhances operational efficiency, and minimizes manual data entry through pre-built integrations or robust APIs.
- C5 - Vendor Reliability and Support:
  Vendor reliability and support are essential for long-term CRM stability. Evaluation of SLAs, customer service options, and supplier reputation ensures effective issue resolution and minimizes operational downtime.

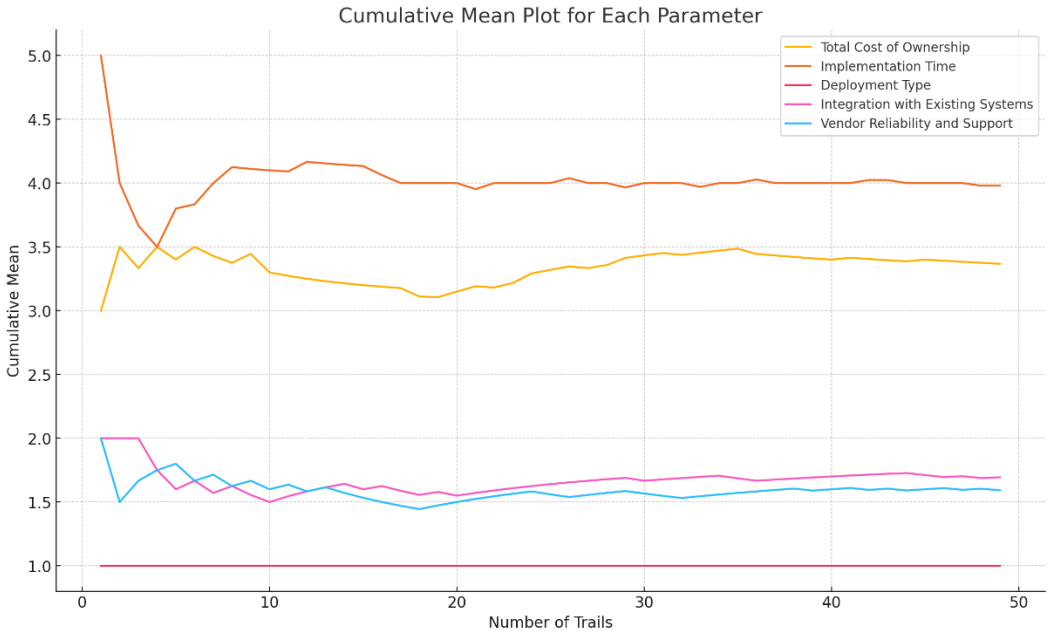
Initial Descriptive Analysis

Basic descriptive statistics, including the mean and standard deviation, were calculated to provide a comprehensive overview of the central tendency and spread of the data.
The Coefficient of Variation (CV) was used to measure the relative variability for each criterion, indicating the consistency of scores.

Convergence Analysis:

The cumulative mean for each parameter was plotted as the number of trails increased. This helped identify the point at which the mean stabilized, suggesting the minimum number of trails required for reliable estimation.
Bootstrapping for Confidence Intervals:

A bootstrapping approach was employed to resample the data and calculate the confidence intervals for the mean of each criterion. This allowed for an estimation of the mean's variability and provided insight into the precision of the average score.



Through analysis of cumulative mean plots and bootstrapping for confidence intervals, it was determined that 10 trials are adequate to achieve a reliable measurement for each criterion when using an LLM. The convergence analysis showed that beyond approximately 10 trials, the cumulative mean for most parameters stabilized with minimal fluctuation, indicating that further trials did not significantly alter the mean.

This threshold of 10 trials ensures that the influence of any individual outlier or model variability is minimized. The bootstrapping results further reinforced this finding, demonstrating that the confidence intervals for the means narrowed sufficiently after 10 trials to provide consistent and reliable estimates.

5. Results/Model Analysis

Weighting Approach

As outlined earlier, the weights reflect the relative importance of each criterion, determined through a combination of objective project requirements and subjective insights from the procurement process. The goal is to ensure that the final decision aligns with the project's primary objectives and constraints.

Rationale Behind Weight Assignment

The weights were derived using a pairwise comparison matrix, which quantifies the relative importance of each criterion. The results of these comparisons are summarized below:

	C1	C2	C3	C4	C5	Priority Value
C1	1	1	3	7	3	0.3968
C2	1	1	3	5	3	0.3425
C3	1/3	1/3	1	3	1	0.1290
C4	1/7	1/5	1/3	1	1/3	0.0587
C5	1/3	1/3	1	3	1	0.0730
Sum	2.8095	2.8667	8.3333	19.0000	8.3333	1.0000

The Total Cost of Ownership (TCO) (C1) was assigned the highest weight of 0.3968, reflecting its dominant role in the decision-making process. The pairwise comparisons indicate that TCO is significantly more important than the other criteria, particularly Integration with Existing Systems and Vendor Reliability and Support. This is consistent with the project's focus on minimizing costs, which often takes precedence in procurement decisions, especially when operating within a fixed budget.

The Implementation Timeline (C2) received the second-highest weight of 0.3425, signifying its importance in ensuring the solution is deployed within the desired timeframe. Timely implementation is essential to avoid delays that could result in operational disruptions or increased costs, making it a critical factor in supplier selection.

Deployment Type (C3) was assigned a moderate weight of 0.1290, as it is relevant to the project's long-term scalability and infrastructure needs but is secondary to cost and timeline considerations. Given the project's flexibility in adapting the deployment method, it was not viewed as a priority compared to TCO and Implementation Timeline.

Integration with Existing Systems (C4) received the lowest priority of 0.0587. While necessary for operational efficiency and reducing data silos, the project team had some flexibility in adapting systems to accommodate the solution, making it less critical compared to TCO, Implementation Timeline, and Vendor Reliability and Support.

Finally, Vendor Reliability and Support (C5) was assigned the lowest weight of 0.0730, reflecting its importance in ensuring the long-term stability of the solution. While supplier

relationships are important, this criterion was considered secondary to TCO and Implementation Timeline. The project team valued reliable support and service but prioritized other factors like cost control and timely implementation.

Summary
The weighted criteria clearly indicate that TCO and Implementation Timeline are the primary decision drivers, with Vendor Reliability and Support and Integration with Existing Systems as secondary considerations. Deployment Type was the least prioritized factor, given the project's flexibility in adapting the system as needed.

6. Conclusions

This study successfully demonstrated that Large Language Models (LLMs) can provide consistent, unbiased, and accurate assessments of SaaS proposals when evaluated against tailored criteria set by a procurement specialist. The LLM's ability to process and rank proposals efficiently addresses a critical need in the software procurement process, where organizations often face the daunting task of reviewing a large volume of responses to RFPs. By employing a structured approach and repeating trials to ensure consistency, the results indicate that LLMs are reliable tools for proposal assessment.

These findings highlight that LLMs can be integrated into an Analytic Hierarchy Process (AHP) framework, where weighted criteria evaluations guide quick and accurate decision-making. This approach supports procurement teams by automating the initial phases of assessment, enabling them to focus on strategic analysis and final decisions. Overall, LLMs show great promise for enhancing the speed, consistency, and objectivity of supplier evaluations, streamlining the procurement process in an increasingly complex software landscape.

7. Limitations

While this study demonstrates the potential of LLMs in enhancing the efficiency and objectivity of proposal evaluations, certain limitations must be acknowledged. One significant constraint was the inability to conduct comprehensive verifications of LLM-generated assessments against human evaluations by procurement specialists. Ideally, a thorough comparison would have involved a team of specialists to independently review and score each proposal for direct comparison with LLM-generated assessments. However, the cost associated with the time and expertise of procurement professionals rendered this approach impractical for our study.

Despite this limitation, partial validation was achieved through the involvement of one of the authors, Park, who is an experienced SaaS procurement expert. Park reviewed the LLM-generated results and provided a limited but meaningful degree of validation, confirming that the AI's assessments aligned well with professional standards in the field. While these findings are promising, future research with broader human verification would further strengthen the reliability and applicability of LLMs in this context.

8. Key References

Hallikainen, Petri, and Hannu Kivijärvi. "Determining Information Systems Procurement Strategy—AHP Approach." ISAHP 1999, Kobe, Japan, August 12-14, 1999. Helsinki School of Economics and Business Administration.

Lu, Xiaotian, Jiyi Li, Koh Takeuchi, and Hisashi Kashima. "AHP-Powered LLM Reasoning for Multi-Criteria Evaluation of Open-Ended Responses." Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, 2024.

Zhao, Xiutian, Ke Wang, and Wei Peng. "Measuring the Inconsistency of Large Language Models in Ordinal Preference Formation." Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), Association for Computational Linguistics, August 2024, pp. 171–176.