An LLM-based Behavioral Intelligence Framework for Assessing Internal Team Conflict and Predicting Project Performance Outcomes

Santosh Reddy [1], Ashwini R Malipatil [2], Srinivasa Suresh Sikhakolli [3], Manoj Meghrajani [4], Samar Mansour Hassen [5], Vaibhav Sharma [6], Mohammed Saleh Al Ansari [7,*], and Mohd Aarif [8]

[1] Department of Computer Science & Technology, Dayananda Sagar University, Bangalore, India
[2] Department of Computer Science & Engineering, BNM Institute of Technology, Bangalore, India
[3] Department of Business Analytics, Kirloskar Institute of Management, Pune, India
[4] Department of Business Administration, Ramchandran International Institute of Management (RIIM), Pune, India
[5] Department of Management Information Systems, College of Business Administration, Jazan University, Jazan, Saudi Arabia
[6] School of Engineering & Technology, Shri Guru Ram Rai University, Dehradun, India
[7] Department of Chemical Engineering, College of Engineering, University of Bahrain, Zallak, Bahrain
[8] Department of Commerce, Aligarh Muslim University, Aligarh, India
Email: dr.santoshreddy-ct@dsu.edu.in (S.R.); ashwinim@bnmit.in (A.R.M.); sssuresh74@gmail.com (S.S.S.); manoj0708@yahoo.co.in (M.M.); shassen@jazanu.edu.sa (S.M.H.); vsdeveloper10@gmail.com (V.S.); malansari.uob@gmail.com (A.S.A.A.); drmohd03@gmail.com (M.A.)
*Corresponding author

***Abstract*—Team conflict within organizations is a vital predictor of project outcomes in today's organizations, especially in digitally mediated settings where interaction mostly happens through text-based communication. Determining the conflict state within projects and predicting project outcomes based on these interactions is a difficult task. Currently, there exist no methods that use automated analyses, dynamic linguistic features, and multivariate prediction models to understand behavioral dynamics, temporal relations, and cross-variables associations between conflict states and project outcomes. Moreover, many available datasets are not annotated for conflict and performance states and cannot be used for supervised learning methods. In order to tackle these issues, this paper suggests a new behavioral intelligence system based on a Large Language Model (LLM) to predict team conflicts and project outcomes from meetings' transcripts. This paper proposes an approach that incorporates an easily interpretable proxy labeling scheme for the generation of conflict and project outcome labels based on measurable signs of conversation, including sentiment polarity changes, interruptions, speaker dominance, and agenda management actions. Linguistic, interactional, and contextual cues are combined with transformer contextual embedding produced by a shared DeBERTa encoder. Multi-task learning is used to benefit from the common representation of the behavior across the two prediction tasks. Experiments show that the model reaches high levels of predictive power, with macro-average precision, recall, and F1-Scores of 91%, 89%, and 90%, respectively, on predicting conflict intensity and accuracy of 93% with a macroROC-AUC of 95% on predicting project outcomes. This approach is implemented using Python and PyTorch.**

***Keywords*—team conflict detection, project outcome prediction, behavioral intelligence, large language models, multi-task learning**

I. Introduction

The digital transformation has radically changed the manner in which project teams work together and most of the face-to-face interactions have been substituted by text-based communications in form of emails, issue-tracking systems and collaborative systems [1, 2]. Although these tools enhance coordination and scalability, they also lead to a loss of interpersonal dynamism which affects team effectiveness [3]. To a great extent, inner conflicts, emotional tension, and the unequal distribution of power are often hidden in the choice of words, the course of conversation, and patterns of interaction than in open words [4, 5]. The latest developments in Large Language Models (LLMs) have contributed to the computational aspects of natural language in a major way by representing the contextual semantics, discourse relations, and the intent to communicate [6, 7]. Transformer-based models like DeBERTa-v3 are more effective at improving the attention mechanisms and decoupling syntactic and semantic representations than previous language models. Similar studies in the field of organizational behavior have always shown that unresolved internal conflict leads to lack of trust, cognitive overload, and poor project results [8]. Such insights notwithstanding, most project analytics systems are still preoccupied with task completion indicators,

Manuscript received January 26, 2026; revised May 13, 2026; accepted June 9, 2026; published September 24, 2026.

schedules, and resource utilization without considering the behavioral indicators that are captured by the day-to-day communication [9]. The coming together of powerful language modeling and behavioral analytics forms a timely moment to study the interaction between teams as dynamic and data-driven processes. Nonetheless, to make this potential practical, it is necessary that methodological frameworks exist that will enable the conversion of crude textual exchanges to indicators of conflict and performance risk that are meaningful.

Despite the fact that sentiment analysis and simple text mining methods have been used to evaluate workplace communication, the methods merely give rough emotional classification and cannot determine the underlying intent in behaviors [10]. Models of polarity cannot distinguish constructive disagreement and damaging conflict or identify patterns of escalation that occur over time over conversations [11]. In addition, the current applications of LLM in organizational settings tend to focus on accuracy of predicting without consideration of interpretability and thus, cannot be accepted by managers when making decisions. The existing literature also prefers to discuss such analytical issues as conflict detection and project performance assessment despite the fact that they are not independent of each other in the actual team context [12]. This disjointed treatment limits the establishment of proactive mechanisms that can predict the risk of performance on the basis of emerging behavioral indicators. Recent progress in multi-task learning provides a promising future as it allows to achieve shared representations in similar prediction goals, leading to better generalization and stability. Together with explanation-capable behavioral signals, including stress cues, conflict signals, and tone of interaction, LLMs can be used to deliver predictive and explanatory capabilities. Nevertheless, it does not have an integration of systems that use semantic embeddings along with behavioral intelligence to understand the conflict between the team and project outcomes simultaneously. This problem needs to be overcome because it is due to this reason that the project has been working on the development of an LLM-Based Behavioral Intelligence Framework using communication data strategically for project outcomes prediction in conflict awareness.

A. Problem Statement

The current state of team communication analytics research suffers from several major drawbacks, which reduce its applicability [13]. Firstly, the majority of researchers base their work on surveys or post-project evaluations, which collect data retrospectively and do not allow for the identification of conflict as it arises at the real time [14]. Second, machine-based approaches tend to use sentiment or emotion-focused approaches to classification, which simplify interpersonal interactions (without contextual sensitivity or intentionality). These methods cannot be satisfactory to find implicit conflict indicators within linguistic patterns, conversation patterns and stress related manifestations [15]. Third, former studies commonly separate conflict detection and project performance analysis as independent processes instead of processes that are not independent. This distance restricts the capacity to bridge the analysis results into predictive action on the risk or success of the project [16]. Moreover, often advanced language models are used as black-box classifiers, which offer little transparency on the factors of behavior in which predictions are dependent. This inability to interpret diminishes the level of managerial trust and obstructs intervention planning. Lastly, no single learning architectures are available to apply semantic understanding, behavioral indicators and prediction of outcome within a single framework. These loopholes underscore the importance of a profound, interpretable and results-based strategy that can use team communication information to provide conflict aware project management.

B. Research Significance

The research contributes to the field of organizational analytics by proposing a unified behavioral intelligence framework that links the outcomes of team communication within the organization with the results of project performance. The combination of DeBERTa-v3 semantic embeddings with interpretable behavior indicators and multi-task learning contributes to the improvement of predictive accuracy and explanatory value of the research. The suggested methodological novelty has both methodological and practical significance to both proactive conflict management and data-driven project governance.

C. Research Motivation

The rationale behind this study is the increasing divide between the amount of digital communication produced by project-based teams and the analytical paucity in the available tools to discern its behavioral implications [17]. The internal conflict usually builds up over a period of time through minor verbal signs that are unidentified until a time when there is the deterioration in performance [18]. The current surveillance strategies do not have the capabilities of analyzing such signals in real time and within their context. The progress in LLM opens a possibility to convert a normal team-to-team communication into a behavioral intelligence source. This study aims to facilitate more and skillful organizational behavior by providing proactive conflict perception and evidence-based project management in order to make organizations more resilient and adaptable.

D. Key Contributions

- Proposed a unified LLM-based framework for predicting team conflict and project outcomes from workplace conversations.
- Introduced a transparent proxy labeling mechanism to enable supervision without manual annotations.
- Multi-task learning using a common transformer encoder that helps in identifying correlated behavior patterns.
- Behavior-aware representation learning using linguistic, interactional, and contextual features.
- Explainability to focus on important dialogues that influence the model's decision.

- Outperforms other models with a macro F1-Score of 90% for conflict prediction and 93% accuracy with 95% ROC-AUC for project outcome prediction.

E. Research Organization

The rest of this study is structured as follows. Section II does a review of related work, Section III contains the problem statement, Section IV outlines the proposed framework, Section V contains results and discussion and Section VI draws a conclusion leaving future directions.

II. Related Works

Davidaviciene *et al.* [19] focused on the early detection of social conflict drivers in civil infrastructure construction, which includes lack of appropriate monitoring mechanisms .The main aim of the study was to go beyond manual literature review and case-study-based analyses and propose a data-driven framework that is automated. The authors suggested a Natural Language Processing (NLP)-based pipeline with the ChatGPT with the purpose of identifying the conflict-relevant key phrases in a large set of news articles and categorizing them into the established categories of conflict drivers. The dataset used to conduct the analysis was obtained by web crawling in search of news items concerning infrastructure projects in South Korea. The research had a high level of performance with a micro-average F1-Score of 85.7 and exemplified the method by case studies. The findings indicated that it was possible to monitor conflicts in a scaled manner with the use of large language models. Nevertheless, the main weakness is the use of predetermined conflict types and news data specific to regions, which may limit the extrapolation to other countries and areas of the project.

Dwivedi *et al.* [20] explored how large language models can be used to aid teamwork conflict management training training in academic institutions. The aim of the study was to boost the confidence and competence of students with the management of interpersonal conflicts in software development project of collaboration. The authors suggested a framework on transformative learning that comprised three stages, including conceptual learning, simulated conflict practice with the use of an LLM, and organized reflection. The data was a collection of the answers of 56 undergraduate students who were in a course in systems development and their logs of interaction. The findings revealed that most students gained confidence in conflict management especially using collaborative and accommodative approaches. Some of the conflict management strategies that were also found to be used often included root-cause analysis and active listening, among others. Although the results show the pedagogical usefulness of simulation conducted with the help of LLM, the research faces limitations in terms of a small group of participants and the use of self-reported confidence levels instead of objective performance data.

Owolabi *et al.* [21] examined the connection between inter-team conflict, strength of ties and project success in Chinese megaprojects. The study was interested in exploring the impact of the characteristics of social networks on conflict and the project outcomes. A survey on the topic was carried out in the form of a questionnaire and the results of 306 respondents who participated in megaprojects were processed with the help of Structural Equation Modeling (SEM). The measured variables in the dataset that should be used include strong and weak ties, task conflict, relationship conflict and project success levels. The researchers discovered that strong ties had positive relationship to task-related conflict and negative relationship between strong ties and relationship-related conflict but no positive relationship was observed with weak ties on overall inter-team conflict. The task conflict was depicted as positively associated to project success, whereas relationship conflict was shown to have negative impact. The most important contribution is the emphasis on the conflict as a mediating variable between social connections and success. Causal inference and cross-cultural applicability are important consideration in this study, however, constrained by the use of self-reported survey data and the emphasis on only a single national situation.

Kalla *et al.* [22] focused on the factors that affect the decision-making process within virtual teams, with particular attention given to organizations that work in the Middle East region. The main aim of the research was to fill the gap in empirical research on virtual teams in the region because most of the previous studies were confined to the Western setting. The authors used a quantitative research design that included literature review, online survey study and structural equation modelling to examine the association between cultural intelligence, transformational leadership, and various forms of conflict. The survey was conducted using Google Forms with the respondents who were the employees of the IT-sector in the United Arab Emirates and were actively involved in virtual teams. The findings revealed that cultural intelligence, transformational leadership, and task-related conflict had a positive impact on the effectiveness of decision-making, and relationship-related conflict had a negative impact on the outcome. The research gives practical advice that managers can use to enhance the performance of virtual teams. Nevertheless, the study has shortcomings as it relies on self-reported information on surveys, cross-sectional design, and is limited to one industry and a geographical area which could restrict the generalizations.

Filice *et al.* [23] studied the incorporation of the LLM-driven cognitive agents into the Agile software project management or under the Scaled Agile Framework (SAFe). The authors of the study tried to evaluate the possibility of intelligent agents to model human project roles and enhance coordination, decision-making, and task execution. The authors used CogniSim simulation environment that was a synthetic data of real-life software development situations. Through the interaction of NLP the cognitive agents exhibited abilities in delegation of duties, communication amongst agents and adaptive planning. The findings revealed the fact that the time taken to complete a task, communicate, and quality of deliverables improved measurably. The research places emphasis on scalability and flexibility of the agents developed with LLM in complicated project environments. However, the primary

weakness is that it used simulated environments as opposed to real industrial data, which casts doubt on the reality of practice, ethical perspectives, and human-AI trust in living project ecosystems.

Baek *et al.* [24] described a detailed overview of the application of artificial intelligence and machine learning in project management, in which the authors focused on predictive analytics, resource allocation, risk assessment, and schedule optimization. The goal of the research was to show how the AI-based tools can promote accuracy in forecasts and management decision-making. The authors analyzed AI and ML models that have been trained with historical project data and provided their examples with real-life case studies. The research attained better forecasts of schedule delays, cost overruns and conflict of resources, thus with evident efficiency. Nevertheless, the paper also identified several major limitations such as reliance on good data, the impossibility of transparent algorithms, and difficulties in adopting it by the users. Given that the study is more of a review and a descriptive paper, it does not empirically benchmark competing AI models thus restricting direct performance comparison and reproducibility in a wide variety of project settings.

Aggrawal *et al.* [25] examined the use of the large language models to recognize dialogue acts during adaptive team training systems. The purpose of the study was to address the issue of team communication analysis on small datasets by exploiting the power of zero-shot and few-shot learning of LLMs. The performances of dialogue act classification were tested with a small team communication dataset. The authors not only compared GPT-4 based models to the traditional transformer-based models but also showed better accuracy and recall especially when domain transfer is involved. The findings proved that the timely engineering had a great effect on the classification efficiency particularly on uncertain acts of dialogue. This literature has a place in adaptive training since it allows real-time analysis of communication with a limited number of labeled data. Nonetheless, drawbacks are the reliance on proprietary LLMs and low interpretability of model decisions that can be inconvenient in safety-critical training circumstances.

DSouza *et al.* [26] suggested a multi-agent communication mining model to identify and predict silent resistance within a company. This paper dealt with the problem of detecting subtle resistance behaviors, which are not readily obtained through surveys or sentiment analysis. The model proposed regarded all the actors in the organization as agents and explored the frequency of interaction, semantic content and centrality of the network. Evaluation was done using a synthetic dataset that modeled organizational transformation situations. The framework was good in performance, and the F1-Score and Area Under Curve (AUC) 0.862 and 0.968, which are better than baseline sentiment-only models. The forecasting module was able to forecast resistance within a period of five weeks. Although the findings are very strong, the lack of external validity to the findings is caused by the use of synthetic data and lack of validation using actual organizational communication logs.

TABLE I. COMPARATIVE SUMMARY OF CONFLICT-RELATED STUDIES

Author	Methodology	Dataset Used	Advantage	Disadvantage
Davidaviciene *et al.* [19]	NLP pipeline using ChatGPT for keyphrase extraction and classification	Web-crawled news articles on infrastructure projects (South Korea)	Automated, scalable conflict detection with high accuracy	Region-specific data, predefined conflict categories limit generalization.
Dwivedi *et al.* [20]	LLM-based simulation with transformative learning framework	Student interaction logs and survey responses (56 students)	Improves conflict management confidence through experiential learning	Small sample size, reliance on self-reported measures
Owolabi *et al.* [21]	Questionnaire survey and Structural Equation Modeling (SEM)	Survey data from 306 Chinese megaproject professionals	Reveals mediating role of conflict in project success	Self-reported data, limited to one national context
Kalla *et al.* [22]	Literature review, online survey, and SEM	Online survey data from IT virtual teams (UAE)	Identifies leadership and cultural factors affecting decisions	Cross-sectional design, limited regional and industry scope
Filice *et al.* [23]	LLM-powered cognitive agents in Agile simulation	Synthetic data from CogniSim simulation environment	Enhances coordination and task efficiency in Agile settings	Lack of real-world industrial validation
Beak *et al.* [24]	Review of AI/ML models and case study analysis	Historical project datasets (varied sources)	Improves forecasting, risk assessment, and optimization	Data quality dependence, lack of empirical benchmarking
Aggrawal *et al.* [25]	Zero-shot and few-shot LLM-based dialogue act classification	Small team communication dataset	High accuracy with minimal labeled data	Limited interpretability, reliance on proprietary models
DSouza *et al.* [26]	Multi-agent communication mining and forecasting model	Synthetic organizational communication dataset	Early detection and prediction of silent resistance	Synthetic data limits real-world applicability

Table I provides a comparative summary of previous literature on conflict analysis, management, and choice-making in organizational and project-based settings. It compares the approaches that have been utilized, such as survey-based structural equation modeling to large language models-based simulations and multi-agent models, and also the datasets, from real-world surveys and web-crawled text to synthetic data. The strengths outline some of the significant contributions made, including automation and scalability, predictive capability and theoretical understanding of conflict dynamics. Such shortcomings are common to most of the existing researches, including regional specificity, self-reporting or simulation, small sample size, and lack of practical verification; hence, they motivate the creation of an integrated, evidence-based, and generalized approach to conflict identification.

III. LLM-Based Behavioral Intelligence Framework

The suggested research utilizes the dataset provided by the Microsoft Topic Conversation Relevance (TCR) (2024) to develop a behavioral intelligence system using an LLM for predicting the internal conflict of the team and the success of the project. Raw meeting transcripts are systematically preprocessed, such as text cleaning, tokenization, speaker turn division, and behavioral cues, such as sentiment polarity, interruptions, speaker dominance and turn length. A transformer-based LLM (DeBERTa) is used to create contextual embeddings, which are combined with behavioral features to create complete representations [27]. Proxies on conflict intensity (LOW, MEDIUM, HIGH) and project outcomes (SUCCESS, DELAY, FAILURE) are based on quantifiable conversational patterns. The hybrid capabilities are handled by an identical LLM encoder, and a multi-task prediction layer, which simultaneously predicts conflict and project performance. A module of explainability underlines influential utterances, which is indispensably interpretable. The framework is trained and tested by conventional classification metrics, and it has a sound predictive performance with transparency and reproducibility.

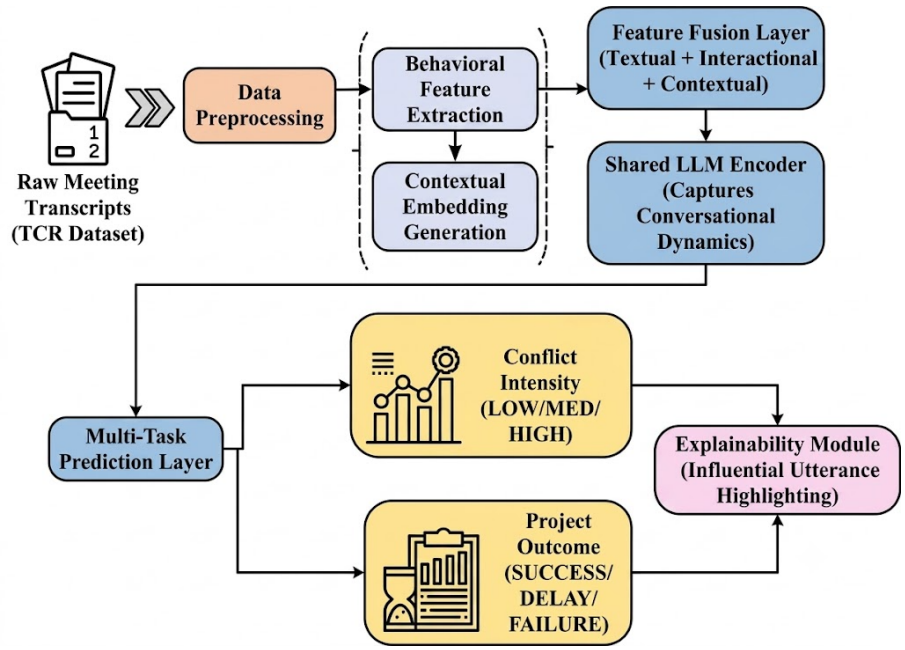

Fig. 1. Block diagram of the proposed LLM-based framework.

The proposed structure in Fig. 1 works with TMC meeting transcripts and preprocesses them to retrieve behavioral features (sentiment, interruptions, speaker dominance) and contextual embeddings with the help of a transformer-based LLM. They are combined and stored within a common LLM, and a multi-task prediction layer at the same time predicts the intensities of conflicts and project results. An explainability module shows influential utterances, which makes the decision-making process transparent and interpretable.

A. Data Collection

For this study, the Microsoft Topic Conversation Relevance (TCR) Dataset (2024) [28] was utilized , which is a publicly available dataset of real-life interactions during meetings in the workplace. The data set is 1500 multi-participant meeting transcripts, which include more than 22 million words in various organizational conversations. All transcripts are represented in JSON format, with speaker codes, dialogue turns, and agenda topics related to it, representing the dynamics of team communication in digital form in detail.

Microsoft used the methods of data collection based on recording and transcribing of organizational meetings, which guaranteed a natural flow of conversation without artificial interventions [29]. In spite of the fact that the dataset will include in-depth dialogue and topic development, conflict labels and indicators of project outcomes are not inherent to it. To overcome this shortcoming, this paper obtains proxy labels using quantifiable conversational properties, including sentiment changes, interruption and topic-resolution patterns. The TCR data therefore provides a true and holistic basis of modeling internal team behavior and the project performance outcomes in contemporary organizational environment [30].

B. Data Preprocessing

The raw transcripts on the Microsoft TCR data are preprocessed (in a systematic preprocessing pipeline) to make them appropriate to the modeling. The first step that is performed is cleaning and normalization of all dialogue texts, such as the removals of irrelevant characters, additional spaces, and aligning of text casings [31]. The speaker turns divide each conversation and maintain the

temporal flow and context of interactions. Linguistic features of tokenization, lemmatization, and removal of stop-words are then implemented to improve the representation of the text. The dialogues are then extracted to identify behavioral patterns like sentiment polarity, speaker dominance, and interruption. Furthermore, the semantic relations between phrases are found using contextual embeddings produced by transformers (e.g., DeBERTa) [32]. This process ensures that the data set contains linguistic and behavioral information, which forms an excellent basis for predicting conflicts and project outcomes.

1) Text cleaning and normalization

Raw data transcripts should be cleaned to remove any noise such as special characters, additional space, and other symbols by Eq. (1).

$$T_c = f_{clean}(T_r) \tag{1}$$

T_r is the raw transcript that undergoes the process of cleaning through function f_{clean} that eliminates unwanted characters. This generates the clean text transcript T_c.

2) Tokenization and lemmatization

Break down the cleaned text into tokens and reduce each word to its base form to standardize the text through Eq. (2).

$$W_i = lemma\big(token(T_c)\big) \tag{2}$$

For each cleaned transcript T_c, the process begins with tokenizing the text into words followed by lemmatization to reduce each token to its base form. This results in the production of W_i, a collection of lemmatized tokens.

3) Speaker turn segmentation

One on one conversation between individual speaker in order to keep conversation flow

$$S_j = \{\, T_{c,k} \mid speaker = j \,\} \tag{3}$$

In the case of Eq. (3), the utterance tokens are grouped per speaker in order to preserve the conversation flow. For every participant j, all turns $T_{c,k}$ belonging to the participant will be collected to form S_j

4) Sentiment and behavioral feature extraction

Determine the polarity and behavior of every utterance.

$$Pol_k = f_{sent}(T_{c,k}), I_k = f_{int}(S_j, k) \tag{4}$$

In order to determine the polarity of every utterance $T_{c,k}$ given in Eq. (4), the sentiment analysis function f_{sent} is used, which represents the emotional content of the conversation. At the same time, the value of the interruption or behavior metric I_k is calculated by the use of f_{int}, which analyzes the sequence of speaker S_j and identifies interruptions, overlaps, and other conversational elements.

C. Proxy Labeling Strategy

In the case of the Microsoft TCR dataset, there are no annotations available for the presence of conflict within teams and the success of the project. As an approach to solve this issue, proxy labeling is employed through which measures are obtained from the conversation transcript. The intensity of conflict is computed on the basis of linguistic and behavioral cues such as change in polarity of sentiment, interruption, and disagreement among the participants. To calculate the Conflict Intensity Score (CIS) of $T_{c,k}$, CIS is calculated as follows:

$$CIS = \alpha \times \bar{Pol}_k + \beta \times I_k + \gamma \times D_k \tag{5}$$

For Eq. (5), Pol_k will be the average value of negative sentiment within the utterances, I_k will be the frequency of interruption, while D_k will be the number of disagreement occurrences in the conversation, with α, β, γ being the empirical weighting values for these features. In accordance with the overall CIS score, the severity of conflicts was categorized as either LOW, MEDIUM, or HIGH based on the meeting. The empirical weights α, β, and γ were selected to reflect the impact of sentiment, interruptions, and disagreements on conflict intensity. We found that $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$ provided the best balance between LOW, MEDIUM, and HIGH conflict severities after analyzing a validation subset of conversations. Thus, the empirically tuned weights ensure the resulting proxy labels' accuracy by reflecting actual behavior patterns, and, at the same time, they remain methodologically transparent.

The proxies of project outcomes are based on the indicators of discussion resolution and adherence to agenda. Project outcome Score (POS) is defined as:

$$POS = \frac{R_t}{T_t} - \delta \times TD \tag{6}$$

where in Eq. (6), R_t is the number of resolved agenda topics, T_t is the total topics discussed, TD represents the topic drift or deviation from the agenda, and δ is a scaling factor. Based on POS thresholds, meetings are labeled as SUCCESS, DELAY, or FAILURE.

D. Proxy Labeling Strategy

After preprocessing and proxy labeling, several features are derived out of the TCR transcripts to give deep inputs into the behavioral intelligence framework that is based on LLM. The representations of features include textual, interactional, and contextual features.

1) Textual feature representation

Text characteristics elicit the semantic and sentiment context of every utterance. Sentiment polarity on the k-th utterance $T_{c,k}$ is calculated as follows Eq. (7)

$$Pol_k = f_{\text{sent}}(T_{c,k}) \tag{7}$$

where, Pol_k represents the polarity score and f_{sent} is the sentiment analysis function (e.g., transformer-based or

VADER). The aggregated sentiment for a meeting is calculated as given in Eq. (8):

$$\bar{Pol} = \frac{1}{N}\sum_{k=1}^{N} Pol_k \quad (8)$$

where, N is the total number of utterances. Additionally, psycholinguistic indicators such as negation counts, modal verbs, and intensifiers are computed to quantify disagreement and assertiveness patterns.

2) Interactional feature representation

Interactional features measure the dynamics among the members of the team. The main characteristics are the interruption frequency *i, k* and speaker dominance SD_j and the length of turns N_i_i, which are defined as follows Eq. (9):

$$SD_j = \frac{\sum_{k=1}^{N_j} 1}{\sum_{i=1}^{M} N_i} \quad (9)$$

where, N_j is the number of turns by speaker j and M is the total number of speakers. This set of features encapsulates engagement, dominance, and conversation flow during meetings.

3) Contextual embedding representation

This set of features encapsulates engagement, dominance, and conversation flow during meetings $Tc_{,k}$ is mapped to a dense vector representation as follows:

$$E_k = f_{embed}(T_{c,k}) \quad (10)$$

where in Eq. (10), $E_k \in \mathbb{R}^d$ is the embedding vector of dimension d. These embeddings are combined with other features like textual and interaction features to create a complete feature vector for each utterance as illustrated in Eq. (11):

$$F_k = [Pol_k, I_k, TL_k, SD_j, E_k] \quad (11)$$

where, F_k is the enriched feature vector used as the input for the LLM model. This method guarantees that both behavioral quantifiers and contextual semantics are considered, improving the predictive power and interpretability of the results.

E. Proposed LLM-Based Behavioral Intelligence Model

The proposed model would be a behavioral intelligence system that employs Large Language Models (LLM) in predicting the outcomes of the conflict within the team as well as the performance of the project. The model uses the text, interaction, and context features extracted from the TCR dataset and applies multi-task learning to capture the shared behaviors. The design of the model ensures that it is highly accurate while at the same time being easily understandable as it has an explanation module indicating the most significant parts of conversations.

1) Model architecture

The architecture consists of three components, that include input representation, shared LLM encoder, and multi-task prediction heads. The input representation of each utterance is comprised of sentiment polarity, interruption rate, utterance duration, and speaker dominance, along with contextual embedding generated during feature engineering [33]. These input representations are fed into a transformer based LLM encoder (DeBERTa), which helps in capturing long-range dependencies as well as complex conversations. The utterances are encoded to generate hidden representations, which are further fed to two separate prediction heads for generating predictions regarding project success (SUCCESS/DELAY/FAILURE) and conflict severity (LOW/MEDIUM/HIGH) [34]. Using the correlations among conflict patterns and project outcomes, the shared encoder architecture improves the performance of the model. Hidden representations of utterances H_k are generated using the following Eq. (12):

$$H_k = LLM(F_k) \quad (12)$$

where, F_k represents the concatenated feature vector for the utterance.

2) Multi-task training

Joint loss function approach is adopted for training the model through the trade-off between conflict intensity and project outcome prediction. In doing so, the shared LLM encoder is able to capture behavioral clues that have an influence on both the tasks, resulting in better generalization. The total loss $\mathcal{L}_{total}$ which can be considered as a weighted sum of conflict $\mathcal{L}_{conflict}$ and project outcome $\mathcal{L}_{outcome}$ is as follows Eq. (13):

$$\mathcal{L}_{total} = \lambda_1 \times \mathcal{L}_{conflict} + \lambda_2 \times \mathcal{L}_{outcome} \quad (13)$$

This is where λ_1 and λ_2 are weight parameters for the importance of the tasks. Both loss functions use the categorical cross-entropy method which is suitable since both label sets are discrete. This allows for learning of the relationship between conflict behavior and project success rate.

3) Explainability module

This is another component that is included to ensure transparency and confidence on the part of the reviewer. It extracts the important features in predicting conflicts and project outcomes, since it helps in identifying the relevant parts of the conversation, and qualitative evaluation of the decisions made by the model is possible. First paragraph: Explainability works based on attention weights and gradient methods such as Grad-CAM for transformers. The contribution of each utterance C_k is measured quantitatively as follows:

$$C_k = \frac{\partial y}{\partial H_k} \quad (14)$$

where, y refers to the predicted probability of the class, while H_k refers to the hidden representation of the utterance. The high contribution scores denote the important turn or interaction which impacted the prediction of the model. This layer of explainability helps in making sure that the predictions made are interpretable and aligned with team behavior.

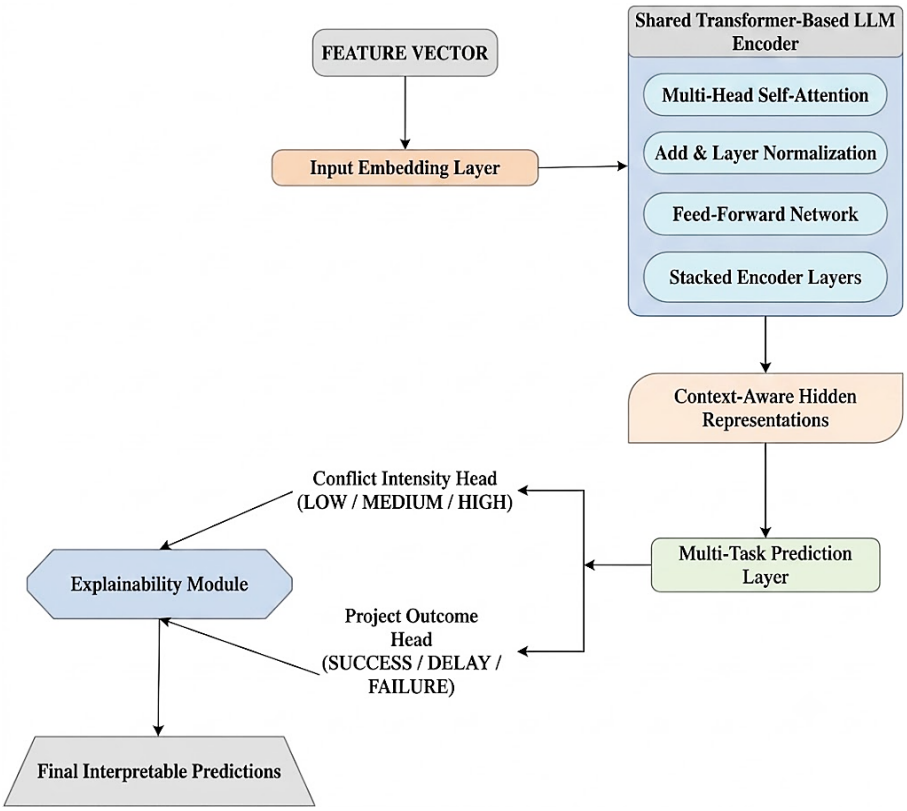

Fig. 2. Proposed behavioral intelligence architecture.

The Fig. 2 shows the proposed structure of the evaluation of the internal team conflict and predicting the outcomes of the project performance on the basis of the approach based on the application of the Large Language Model (LLM). Precomputed transcripts of the Microsoft TCR are then sent through a feature engineering module which generates textual (sentiment polarity, psycholinguistic indicators) and interactional (interruption, speaker dominance, turn length) and contextual embeddings using a transformer-based language model (DeBERTa). These enriched feature vectors are fed in to a common encoder of LLM which learns long-term dependencies and subtle patterns of conversation. Two task prediction heads are fed by the encoder, one of which is the classification of conflict intensity (LOW, MEDIUM, HIGH), and another one is predicting project outcome (SUCCESS, DELAY, FAILURE). Lastly, an explainability module brings out key contextual areas of conversation, which offer insight and clarity to the choices of the model. This architecture can support high predictive accuracy and interpretability and learn multiple tasks at the same time.

F. Evaluation Metrics and Experimental Setup

In order to measure the effectiveness of the proposed behavioral intelligence system based on the use of the LLM, a series of conventional classification indicators and a strict experimental design is utilized. The test is aimed at conflict intensity prediction and project outcome prediction.

To assess the model, it is analyzed in terms of accuracy, precision, recall, and F1-Score that gives integral information on the performance of classification. In case of multi-class issues like intensity of conflict (LOW, MEDIUM, HIGH) and outcomes of projects (SUCCESS, DELAY, FAILURE), the macro-averaged F1-Scores are calculated to balance the assessment of all classes. As well, the confusion matrices are examined to determine certain misclassification trends. To obtain additional confirmation, area under ROC curve (AUC) is calculated on each of the classes to analyze discriminative ability.

The obtained TCR transcripts (postprocessed and in the form of proxy labels) are further divided into training (70%), validation (15%), and test (15%) sets. The computed multi task loss is directly computed using the derived proxy labels in model training. In order to verify the reliability of the labels, a subset of meetings was checked manually, and it was established that the labels of the conflict intensity derived with the help of proxies and the project outcome outcomes correspond to the observable team behaviors. This validation measure will ensure that the labels have meaningful supervisory information about both tasks as well as improve the overall methodological rigor of the study. The model is coded in Python and

PyTorch, where the transformer based DeBERTa encoder shall be used as the embedding layer. The heads of multi-task prediction are simultaneously trained with the joint loss function.

Other hyperparameters are also chosen to maximize the performance of the model and guarantee reproducibility. The transformer encoder has 12 layers which are dropped by 0.1 to avoid overfitting. It is trained with 20 epochs with 12 attention heads per layer. These hyperparameters were chosen after a combination of both empirical tuning and a reliance to best practices in training transformer-based LLM, so as to have a stable and robust learning process. The experiments are all implemented on an NVIDIA GPU based workstation, so that the training is efficient, reproducible, and results consistent.

Algorithm 1. LLM-Based Behavioral Intelligence Framework

Input:
  Meeting transcripts from the Microsoft TCR dataset
Output:
  Conflict intensity label and project outcome label for each meeting
Initialize transformer-based language model (DeBERTa)
Initialize multi-task prediction heads for conflict and project outcome
Load raw meeting transcripts
For each meeting transcript do
  Clean and normalize textual content
  Segment transcript into speaker-wise utterances
  Tokenize and lemmatize utterances
  Extract textual features including sentiment and psycholinguistic cues
  Extract interactional features including interruptions, speaker dominance, and turn length
  Generate contextual embeddings for each utterance using the transformer encoder
  Derive proxy conflict indicators from sentiment shifts, interruptions, and disagreements
  Assign conflict intensity label (LOW, MEDIUM, HIGH)
  Derive proxy project outcome indicators from agenda resolution and topic adherence
  Assign project outcome label (SUCCESS, DELAY, FAILURE)
  Fuse textual, interactional, and contextual features into a unified feature representation
End For
Split dataset into training, validation, and test sets
Train shared LLM encoder using multi-task learning
Optimize joint loss for conflict intensity and project outcome prediction
Apply early stopping based on validation performance
Evaluate trained model on test set
Compute accuracy, precision, recall, F1-Score, and macro-averaged metrics
Generate explainability outputs highlighting influential conversation segments
Return predicted conflict intensity and project outcome labels

The workflow of the proposed end-to-end process of the LLM-based behavioral intelligence framework in predicting internal team conflict and project performance outcomes is described in Algorithm 1. This is done by loading raw meeting transcripts of the Microsoft TCR dataset, which is then preprocessed systematically to clean the text, maintaining speaker-wise conversational structure. The linguistic content, dynamic behavior, and semantic context are then captured using textual, interactional and contextual features. Observable conversational indicators are used to derive proxy labels of the conflict intensity and project outcomes. The representations of the features are combined and inputted into a common transformer-based LLM encoder that is trained in a multi-task learning configuration. Lastly, an evaluation of the trained model is performed through conventional metrics of classification, and an explainability module identifies the parts of the conversation most strongly affecting the predictions to provide straightforward and interpretable ones.

II. Result and Discussion

In this section, the proposed Behavioral Intelligence Framework implementing LLM-based Conflict-Aware Project Outcome Prediction (LBIC-POP) is fully evaluated based on the Microsoft Topic Conversation Relevance (TCR) Dataset (2024). The experimental analysis is concerned with the capacity of the framework to analyze internal team conflict together and predict the outcomes of project performance based on the data of digital communication. Findings have been provided on three conflict detection, conflict severity estimation and project outcome prediction tasks. Effectiveness is shown by the performance compared to the traditional NLP and single-task learning baselines. Every experiment was made with the help of PyTorch, HuggingFace Transformers, scikit-learn, and the model training was held on the systems with the usage of the GPUs. The strength, accuracy and practical viability of the behavioral intelligence approach is validated by the quantitative measures, ablation and interpretability studies respectively.

A. Experimental Setup and Data Splits

The experiments were carried out based on Microsoft Topic Conversation Relevance (TCR) Dataset (2024), which contains multi-turn conversational interactions that were annotated in topic relevance and interaction dynamics. To exclude information leaks, speaker-independent splits were used in partitioning the dataset into 70% training, 15% validation and 15% testing sets. Raw conversational text was processed by removing noise, lowercasing, and getting rid of system generated artifacts. To maintain contextual dependencies, sentence segmentation and tokenization were done using the DeBERTa tokenizer. The conversational turns were systematically organized in a way that they preserved the flow of interaction and the behavioral cues, such as negativity degree, conflict signs, and stress signs, were desiccated on an utterance level and temporally correlated with semantic embedding before feature fusion.

Table II was trained on hardware that is accelerated by a GPU to aid in efficient fine-tuning of the DeBERTa-v3 model. End-to-end training was made possible through PyTorch and HuggingFace Transformers, and the optimization was made stable using Adam optimizer. The choice of hyperparameters has been to strike a balance

between predictive performance and computational efficiency of all the multi-task learning goals.

Fig. 3 shows the convergence pattern of training and validation loss in fine-tuning of the multi-task behavioral intelligence model that is based on DeBERTa-v3. Stable optimization and efficient learning of shared representations in conflict detection and project outcome prediction tasks are observed when there is a systematic decrease in both curves. Small divergence among losses is an indication of limited overfitting as well as good generalization ability.

TABLE II. SIMULATION PARAMETER AND HARDWARE SETUP

Parameter	Value
Framework	PyTorch, HuggingFace Transformers
LLM Backbone	DeBERTa-v3-base
Optimizer	Adam
Learning Rate	2e−5
Batch Size	16
Epochs	10
GPU	NVIDIA RTX 3090 (24 GB)
CPU	Intel Xeon
RAM	64 GB

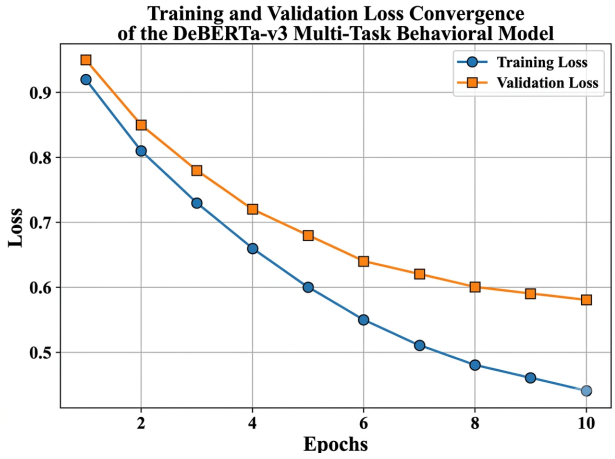

Fig. 3. Loss convergence of DeBERTa-v3 model.

B. Feature and Proxy Label Statistics

The descriptive analysis of the proxy labels will give an indication on how the inferred conflict intensity and project performance outcomes are distributed in the data. There is an equal representation of the different levels of conflict severity and outcomes categories into the statistics, which allows the achievement of multi-task learning. The use of these distributions proves the possibility of proxy-based supervision to predict the dynamics of behavior and future downstream project performance.

Fig. 4 shows the frequency of the proxy-based conflict severity categorical labels that fall in the low, medium and high categories. All the levels of severity with adequate sample numbers facilitate constructive supervised learning to detect the level of conflict and estimate its severity. The distribution is also realistic in terms of team communication dynamics that are observed within collaborative settings.

Fig. 5 presents the distribution of the proxy-labeled project outcomes such as successful, partially successful and unsuccessful. With a comparatively even distribution of outcome classes, it becomes possible to train a model and evenly assess. These proxy labels give a feasible mechanism of connecting the pattern of conversation behavior with the high levels of project performance results.

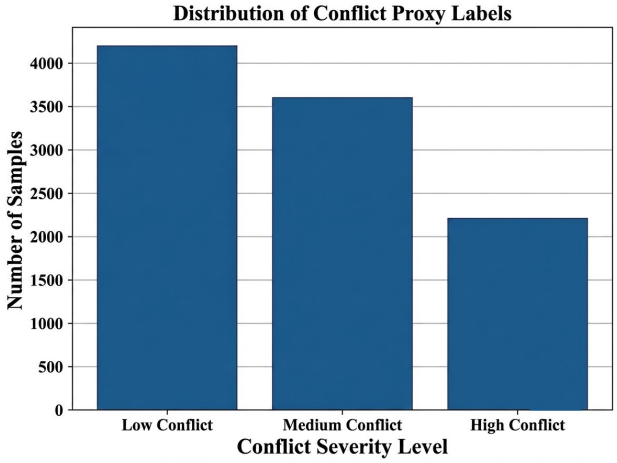

Fig. 4. Distribution of conflict severity levels.

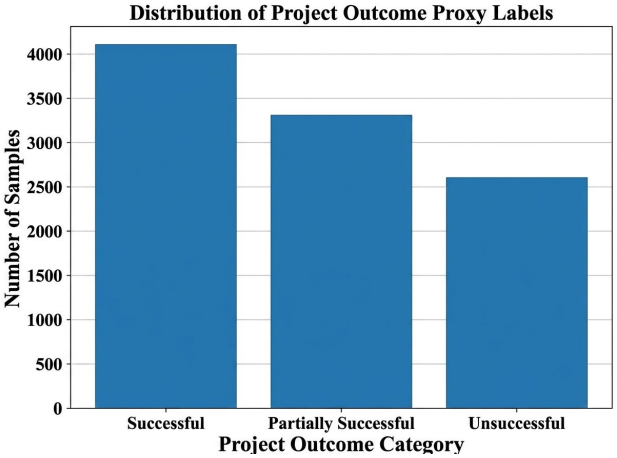

Fig. 5. Distribution of project outcome labels.

C. Experimental Setup and Data Splits

The results of the experiment prove that the developed behavioral intelligence framework of the form of LLCM is effective in the modeling of latent conflict dynamics in the team and their role in project performance. The coherent behavioral reasoning is possible through joint learning applied to the detection of conflict severity, estimation of outcomes, and conflict severity. Comparative and explainability studies show that integrating contextual semantics with interpretable behavioral information can result in consistent, context-sensitive predictions consistent with behavioral patterns of collaborative communication in the real world.

Fig. 6 shows the confusion matrix of binary conflict detection through TCR dataset. Good diagonal dominance implies proper discrimination of conflict prone and neutral interactions. The small off-diagonal scores illustrate that the model can be used to exploit context-based semantics and behavior patterns in identifying conflicts in team discourse in a reliable manner.

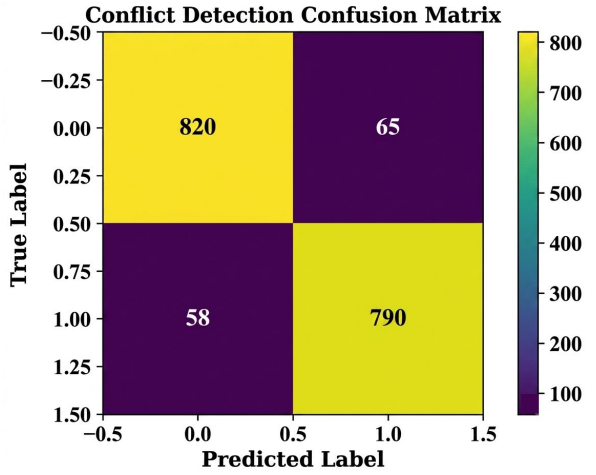

Fig. 6. Conflict detection confusion matrix.

Fig. 7 represents the low, medium, and high conflict severity prediction based on performance in terms of classes. The similarity in the scores between classes means that there is no bias in favor of dominant categories or a different type of learning. This pattern of behavior proves that the framework can be helpful in capturing finer escalation patterns in team interactions.

Fig. 8 demonstrates the correspondence between the expected levels of conflict severity and labels obtained by proxies. Strong inter-sample correspondence reflects the fact that the model is able to learn severity gradients based on behavioral and semantic features, which confirms the accuracy of proxy supervision in the TCR dataset.

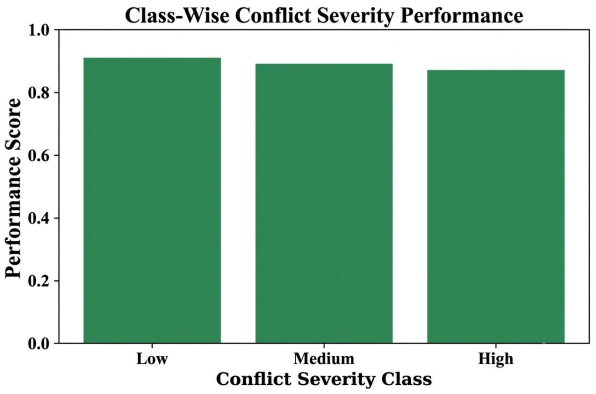

Fig. 7. Conflict severity class performance.

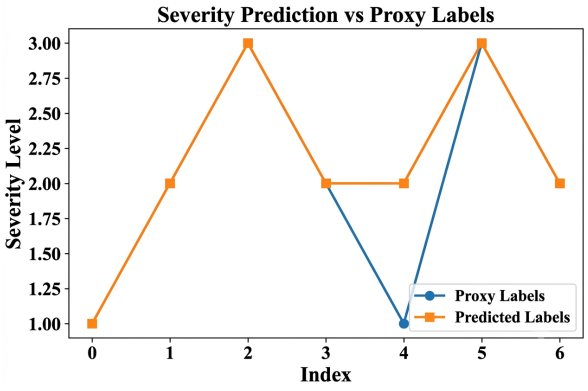

Fig. 8. Predicted versus proxy conflict severity.

Fig. 9 demonstrates how semantic, behavioral and fused feature settings contribute to predictive effectiveness. This high-performance of fused features proves that integrating interpretable behavioral features with contextual embeddings is important to improve the robustness of a model and contribute to more effective behavioral reasoning.

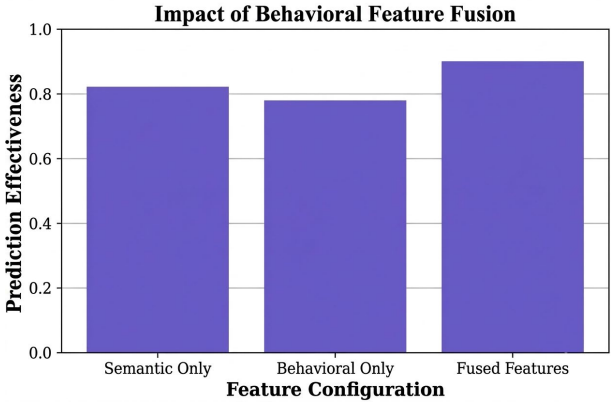

Fig. 9. Impact of behavioral feature fusion.

Table III presents the sensitivity analysis of the Conflict Intensity Score weighting parameters by varying the values of α, β, and γ. The experimental results show only minor fluctuations in precision, recall, and F1-Score across different weight configurations, indicating that the proposed framework is robust to moderate hyperparameter variations. The baseline configuration $(0.5, 0.3, 0.2)$ and nearby weight combinations achieved the best performance with a macro F1-Score of 90%.

TABLE III. SENSITIVITY ANALYSIS OF CIS WEIGHT PARAMETERS

α	β	γ	Precision (%)	Recall (%)	F1-Score (%)
0.5	0.3	0.2	91	89	90
0.4	0.4	0.2	90	88	89
0.6	0.2	0.2	90	87	88
0.5	0.2	0.3	89	88	88
0.45	0.35	0.2	91	89	90
0.55	0.25	0.2	90	89	89

Fig. 10 illustrates the effect of varying the CIS weighting coefficients on the macro F1-Score of the proposed behavioral intelligence framework. The graph shows only slight fluctuations in performance across different weight configurations, indicating that the model is stable and robust to moderate hyperparameter changes. The highest F1-Score is achieved with the baseline and nearby configurations, validating the effectiveness of the empirically selected weights.

Fig. 11 shows how intensity of conflict changes after consecutive conversational turns. The chronological development brings out the trend in which latent conflict grows or remains constant in the course of team interactions. Such a visualization helps the framework to reflect the dynamic pattern of behavior and not just the isolated linguistic cues, as is the case with team communication behavior in the real world.

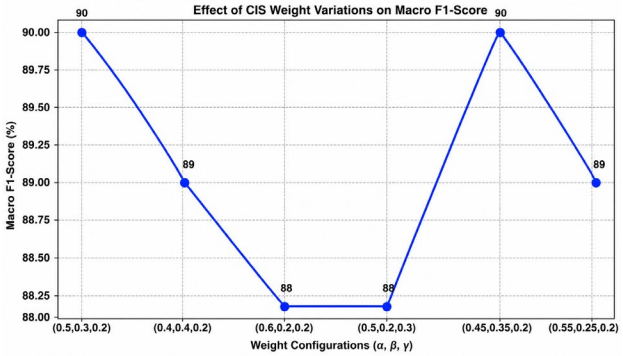

Fig. 10. Effect of CIS weight variations on macro F1-Score.

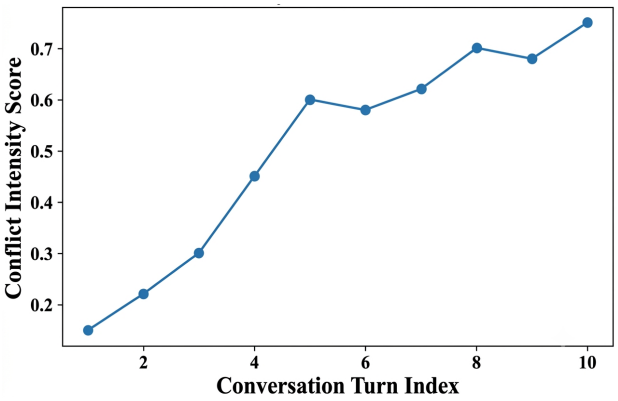

Fig. 11. Conflict intensity across conversation turns.

D. Performance Metrics

The performance indicators prove that the proposed behavioral intelligence framework based on the LLM allows modeling the dynamics of the internal conflict and the project performance results. The results of the conflict intensity prediction in macro-averaged show that there are no differences in performance in the high levels of severity but the learning of the classes of minority and majority conflict is balanced. To predict the outcome of the project, the strong weighted measures and high discriminating ability of the framework emphasize the strength of the framework in addressing the imbalance of outcomes. The findings are truly coherent to the effect that when contextual semantic representation is integrated with explanatory behavioral measures, reliable learning of multi-tasks can be achieved, which can make stable predictions at the same time involve predictability, and can be relevant to real life communication, and project dynamics in a team.

The summary of the macro-averged performance of the conflict intensity prediction task is given in Table IV. The reported measures represent the same level of classification based on low, medium, high levels of conflict and illustrate the capability of the framework to gain balanced displays of the severity of conflict in the absence of favoring dominant classes.

TABLE IV. PERFORMANCE OF CONFLICT INTENSITY PREDICTION TASK

Metric (Macro-Averaged)	Value (%)
Precision (Macro)	91
Recall (Macro)	89
F1-Score (Macro)	90

The assessment outcomes of project outcome prediction are given in Table V. Strong performance on the discriminative metrics and weighted metrics shows that learning under imbalance of classes is strong. These findings justify the ability of the framework to relate behavioral and semantic cues of team communication with high-level project performance outcomes.

TABLE V. PROJECT OUTCOME PREDICTION PERFORMANCE METRICS

Metric	Value (%)
Accuracy	93
Precision (Weighted)	92
Recall (Weighted)	93
F1-Score (Weighted)	92
ROC–AUC (Macro)	95

Fig. 12 shows the percentage-based project outcome performance prediction using the suggested behavioral intelligence framework. Large values of accuracy, precision, recall, F1-Score, and ROC-AUC denote the strong outcome-level reasoning. These findings attest to successful collaboration of semantic and behavioral characteristics to ensure sound project performance estimation.

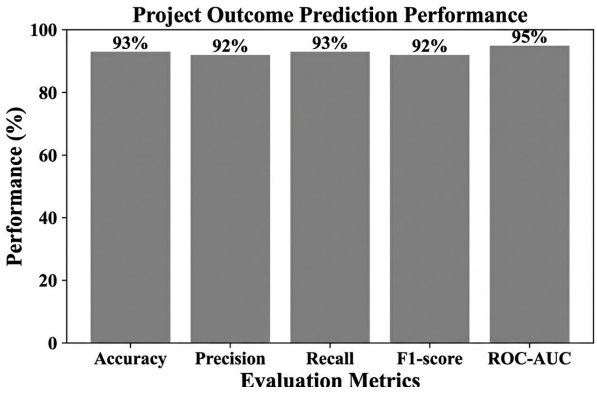

Fig. 12. Project outcome prediction performance metrics.

Table VI presents the mean and standard deviation of evaluation metrics across multiple random seeds. The low standard deviation values indicate that the proposed multi-task framework achieves stable and reproducible performance.

Table VII summarizes the paired t-test analysis between the single-task baseline and the proposed multi-task model. The obtained *p*-value confirms that the performance improvement achieved by the proposed framework is statistically significant.

TABLE VI. STATISTICAL STABILITY AND SIGNIFICANCE ANALYSIS

Model	Precision (%)	Recall (%)	F1-Score (%)
Single-task Baseline	86.4 ± 1.2	84.9 ± 1.5	85.3 ± 1.3
Proposed Multi-task Model	91.0 ± 0.5	89.0 ± 0.6	90.0 ± 0.4

TABLE VII. STATISTICAL SIGNIFICANCE ANALYSIS USING PAIRED T-TEST

Comparison	*p*-value	Significance
Single-task vs Multitask F1-Score	0.003	Significant ($p < 0.05$)

TABLE VIII. COMPARISON OF CONFLICT AND OUTCOME PREDICTION

Model	Task	Common Metrics Reported	Performance (%)
Proposed LBIC POP	Conflict Intensity Prediction	Precision (Macro)	91
		Recall (Macro)	89
		F1-Score (Macro)	90
NLP [19]	Conflict Detection	F1-Score (Micro)	85.7
		Macro F1-Score	83.6
		Weighted F1-Score	84.7
Proposed LBIC POP	Project Outcome Prediction	Accuracy	93
		Precision (Weighted)	92
		Recall (Weighted)	93
		F1-Score (Weighted)	92
		ROC–AUC (Macro)	95
Multi-agent communication mining framework [26]	Organizational Resistance Prediction	Precision	86.2
		F1-Score	86.2
		ROC–AUC	96.8
Mobile App Feedback Conflict Detection [28]	Conflict Detection	Precision	84.9
		Recall	75.2
		F1-Score	79.7

Table VIII contrasts the suggested LBIC-POP framework with the representative NLP-based and multi-agent models, in the conflict intensity, conflict detection, and prediction of the organizational outcome tasks. The metrics are reported only on the measures that are available in each study in order to compare fairly. The findings demonstrate the success of applying semantic language modeling and behavioral intelligence in modeling dynamic team interaction patterns.

The performance of the proposed LBIC-POP framework in terms of comparative performance in the F1-Score against the related conflict and organizational behavior prediction studies is presented in Fig. 13. The visualization also emphasizes that the similar improvement in performance has been attained by the proposed strategy, which indicates the efficiency of the application of contextual semantic representations as well as behavioral intelligence to the representation of conflict intensity and outcomes of the project.

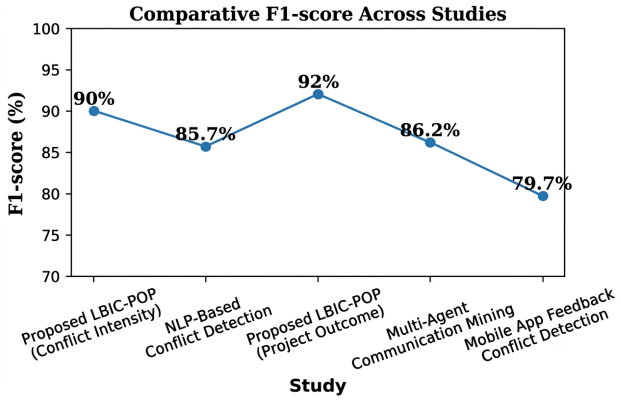

Fig. 13. Comparative F1-Score across related studies.

TABLE IX. COMPARISON BETWEEN FINE-TUNED ENCODERS AND GENERATIVE LLM PROMPTING

Model Type	Advantages	Limitations
Generative LLMs (Zero/ Few-shot Prompting)	Strong reasoning ability, flexible prompting, broad generalization	Prompt sensitivity, unstable outputs, high computational cost
Fine-tuned DeBERTa-v3 Encoder	Stable classification performance, efficient training, domain adaptation, reproducible predictions	Requires supervised fine-tuning and labeled data

Table IX compares generative LLM prompting approaches with the fine-tuned DeBERTa-v3 encoder used in the proposed framework. This comparison highlights the high reasoning capacity and prompt flexibility of generative LLMs compared to the stability and reproducibility of fine-tuned encoders in classification tasks. Based on these results, the choice of DeBERTa-v3 in behavioural conflict analysis and multiclass project outcomes prediction is justified.

Fig. 14 shows the comparative study between generative LLM prompting techniques and the fine-tuned DeBERTa-v3 encoder in terms of various behavioural NLP features. The findings indicate that while the reasoning ability is higher in case of generative LLMs, stability, adaptability, and efficiency in computations are higher for the fine-tuned DeBERTa-v3 encoder. This justifies the use of the fine-tuned encoder for multi-task behavioural conflict analysis and project outcome prediction.

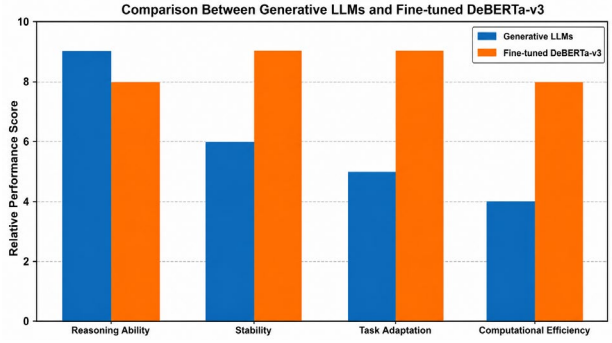

Fig. 14. Comparative study of generative LLMs and fine-tuned DeBERTa-v3.

Fig. 15 demonstrates the performance stability between the proposed multi-task learning framework and the single task baseline with multiple random initialization seeds. The proposed method is able to achieve a better mean F1-Score with significantly narrower error bars, implying less variability in the performance. These findings demonstrate that the multi-task behavioral intelligence framework yields reliable performance stability.

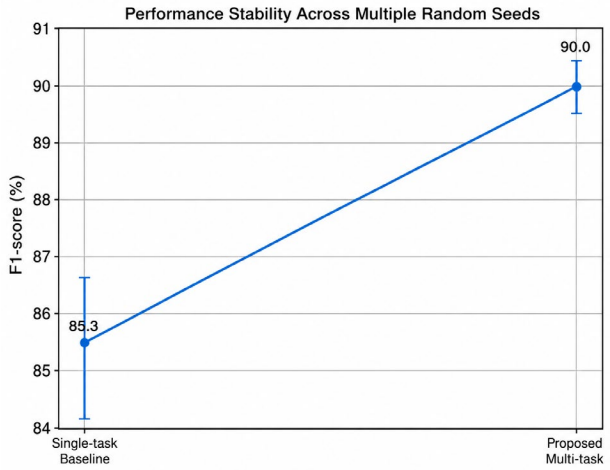

Fig. 15. Stability of performance with multiple random seeds.

E. Discussion

The results of the experiment show that the LBIC-POP framework suggested is effective in explaining the latent behavioral dynamics inherent in team communication and is able to connect them to the outcomes of conflict intensity and the project performance. The macro-averaged conflict intensity statistics show that the conflict intensity metrics are balanced in the low, medium, and high severity levels, which can be interpreted as the resistance to the dominance of classes and enhanced visibility to minor escalation patterns. The high weighted performance in project outcome prediction shows that the framework can be used to deal with outcome imbalance and at the same time have high discriminative power. The fact that the model reflects the dynamics of interaction, but not individual lexical clues is further supported by visual analyses, such as confusion matrices, and temporal conflict evolution plots. The hypothesis that contextual semantic embedding predictive stability and reasoning depth would be improved when integrating contextual semantic embedding with interpreted behavioral indicators is proven by the feature fusion experiments. All in all, it can be implied that joint multi-task learning allows performing coherent behavioral reasoning in objectives that are linked, which makes the framework a valid instrument of conflict-conscious project intelligence in collaborative digital spaces. The statistical stability analysis further demonstrates that the proposed framework maintains consistent predictive performance across different random initialization settings. The significance testing confirms that the improvements over single-task learning are reliable and not attributable to stochastic variation.

III. Conclusion and Future Work

This research paper has come up with an LLM based Behavioral Intelligence Framework for measuring internal team conflict and predicting project performance outcomes. This framework is named LBIC-POP and represents the state of the art in detecting conflicts and estimating conflict severity along with predicting project outcomes in a multi-task learning approach. Unlike conventional approaches, this framework makes use of similar backbone semantic representation using DeBERTa-v3 but includes behavioral features as well in order to better reason about interactions within teams. The empirical tests have produced highly accurate results for both conflict severity and results prediction even in case of class imbalance proving the effectiveness of using proxy-based supervision and feature fusion approaches. The three main elements of the novelty of the work include the combination of behavioral intelligence with large language models, making conflict and outcome prediction tasks interdependent, and using proxy labels to realize abstract organizational constructs based on conversational data. All of these contributions help to more realistically model collaborative environments, whose conflict and performance are necessarily interconnected. The convergence stability of the framework, its class-wise high performance, and overall high-performance show its suitability in the real-world implementation in the organizational analytics, project management systems, and digital collaboration platforms.

The further research directions will be attempting to expand the framework to multilingual and cross-cultural communication environments to determine the generalizability not only across one dataset. When used with multimodal features and signals (voice tone, response latency, interaction networks), it might further improve behavioral concepts. Besides, longitudinal assessment of entire lifecycle of project could help to better understand causal relations between conflict during initial stages and final results. The study of explainable AI processes based on behavioral indicators would also enhance transparency and managerial confidence. These extensions will also enhance the framework in terms of application as a scalable, conflict-sensitive decision support system, in a complex team-based environment.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

S.R. is responsible for Conceptualization, methodology, project management, supervision, and manuscript review. A.R.M.: Preprocessing, formal analysis, data curation, literature review, and manuscript drafting. S.S.S.: Critical evaluation, validation, framework creation, and methodology improvement. M.M.: Designing experiments, interpreting results, creating visuals, and editing manuscript. S.M.H.: Review, editing, quality control, validation, and investigation. V.S.: Model creation, experiments, software implementation, and visualization.

M.S.A.A.: Final approval, correspondence, resource coordination, and supervision. M.A. Final approval, proofreading, data interpretation, validation, and paper editing. All authors reviewed and approved the final manuscript.

REFERENCE

[1] B. Lal, Y. K. Dwivedi, and M. Haag, "Working from home during Covid-19: doing and managing technology-enabled social interaction with colleagues at a distance," *Information Systems Frontiers*, vol. 25, no. 4, pp. 1333–1350, 2023.

[2] N. Almarimi, A. Ouni, M. Chouchen, and M. W. Mkaouer, "Improving the detection of community smells through socio-technical and sentiment analysis," *Journal of Software: Evolution and Process*, vol. 35, no. 6, e2505, 2023.

[3] S. Berrezueta-Guzman, P. Bassner, S. Wagner, and S. Krusche, "Code collaborate: Dissecting team dynamics in first-semester programming students," in *Proc. 2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, IEEE, 2024, pp. 1–10.

[4] C. Dash, M. S. A. Ansari, C. Kaur *et al.,* "Cloud computing visualization for resources allocation in distribution systems," in *Proc. AIP Conference*, 2025, vol. 3137, no. 1, 020038.

[5] A. F. Ahmad, "The influence of interpersonal conflict, job stress, and work life balance on employee turnover intention," *International Journal of Humanities and Education Development (IJHED)*, vol. 4, no. 2, pp. 1–14, 2022.

[6] N. Looman, T. van Woezik, D. van Asselt, N. Scherpbier-de Haan, C. Fluit, and J. de Graaf, "Exploring power dynamics and their impact on intraprofessional learning," *Medical education*, vol. 56, no. 4, pp. 444–455, 2022.

[7] M. Aarif, A. Anjum, T. Sharma *et al.,* "Implementing fuzzy logic in cognitive sensor networks for environmental monitoring," in *Proc. 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2024, pp. 1–6.

[8] S. Aggrawal and A. J. Magana, "Teamwork conflict management training and conflict resolution practice via large language models," *Future Internet*, vol. 16, no. 5, 177, 2024.

[9] D. Pacella and D. Marocco, "Understanding negotiation: A text-mining and NLP approach to virtual interactions in a simulation game," *Applied Sciences*, vol. 12, no. 10, 5243, 2022.

[10] T. Gavrić and I. N. Braje, "Managing cultural diversity and conflict in family businesses: An organizational perspective," *Administrative sciences*, vol. 14, no. 1, 13, 2024.

[11] C. Kaur and A. Balobaid, "Artificial Intelligence in Robotics: Revolutionizing Industrial Automation and Beyond," in *Proc. 2025 International Conference on Frontier Technologies and Solutions (ICFTS)*, 2025, pp. 1–7.

[12] L. Poetz and J. Volmer, "What does leadership do to the leader? Using a pattern-oriented approach to investigate the association between daily leadership profiles and daily leader well-being," *Journal of Business and Psychology*, vol. 39, no. 6, pp. 1259–1281, 2024.

[13] C. Emexidis, P. Gkonis, and A. Liapakis, "Analyzing Employee Job Satisfaction Through Sentiment Analysis for Enhanced Workplace Improvement and Business Success," *Theoretical and Applied Ergonomics*, vol. 1, no. 2, 10, 2025.

[14] R. Martínez-España, J. Fernández-Pedauye, J. G.-P. De Lucía, J. M. Rojo-Martínez, K. Bakdid-Albane, and J. J. García-Escribano, "Methodology for measuring individual affective polarization using sentiment analysis in social networks," *IEEE Access*, vol. 12, pp. 102035–102049, 2024.

[15] H. Donelan and K. Kear, "Online group projects in higher education: persistent challenges and implications for practice," *Journal of computing in higher education*, vol. 36, no. 2, pp. 435–468, 2024.

[16] P. C. Behera, M. Aarif, S. H. I. Gardezi *et al.,* "Blockchain-Based Audit Mechanism for Verifiable and Transparent Ethical Hacking in Enterprise Penetration Testing Systems," in *Proc. 2026 1st International Conference on Emerging Trends in Advancements and Applications of Computational Intelligence Techniques (ETAACT)*, 2026.

[17] D. Siemon, "Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration," *Group Decision and Negotiation*, vol. 31, no. 5, pp. 871–912, 2022.

[18] I. O. Evans-Uzosike, C. G. Okatta, B. O. Otokiti, O. G. Ejike, and O. T. Kufile, "Modeling the Impact of Project Manager Emotional Intelligence on Conflict Resolution Efficiency Using Agent-Based Simulation in Agile Teams," *International Journal of Scientific Research in Civil Engineering*, vol. 8, no. 5, pp. 154–167, 2024.

[19] V. Davidaviciene and K. Al Majzoub, "The effect of cultural intelligence, conflict, and transformational leadership on decision-making processes in virtual teams," *Social Sciences*, vol. 11, no. 2, 64, 2022.

[20] J. Dwivedi, I. Muda, M. B. Alazzam *et al.,* "Natural Language Processing in Financial Reports: Analyzing Annual Reports Using AI for Investment Insights," in *Proc. 2025 IEEE International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, 2025, pp. 1–5.

[21] H. A. Owolabi *et al.,* "Big data innovation and implementation in projects teams: Towards a SEM approach to conflict prevention," *Information Technology & People*, vol. 38, no. 3, pp. 1363–1402, 2025.

[22] M. Kalla, M. Jerowsky, B. Howes, and A. Borda, "Expanding formal school curricula to foster action competence in sustainable development: A proposed free-choice project-based learning curriculum," *Sustainability*, vol. 14, no. 23, 16315, 2022.

[23] L. Filice and W. J. Weese, "Developing emotional intelligence," *Encyclopedia*, vol. 4, no. 1, pp. 583–599, 2024.

[24] S. Baek, D. Namgoong, J. Won, and S. H. Han, "Automated detection of social conflict drivers in civil infrastructure projects using natural language processing," *Applied Sciences*, vol. 13, no. 20, 11171, 2023.

[25] S. Aggrawal and A. J. Magana, "Teamwork conflict management training and conflict resolution practice via large language models," *Future Internet*, vol. 16, no. 5, 177, 2024.

[26] M. DSouza, R. Pradhan, C. Kaur *et al.,* "Quantum-Inspired Genetic Algorithms for Secure and Scalable Cloud-Based Decision Support Systems," in *Proc. 2025 2nd International Conference on Integration of Computational Intelligent System (ICICIS)*, 2025.

[27] H. Zeng, J. Cao, and Q. Fu, "Unpacking the 'black box': understanding the effect of strength of ties on inter-team conflict and project success in megaprojects," *Buildings*, vol. 12, no. 11, 1906, 2022.

[28] V. Davidaviciene and K. Al Majzoub, "The effect of cultural intelligence, conflict, and transformational leadership on decision-making processes in virtual teams," *Social sciences*, vol. 11, no. 2, 64, 2022.

[29] K. Cinkusz, J. A. Chudziak, and E. Niewiadomska-Szynkiewicz, "Cognitive agents powered by large language models for agile software project management," *Electronics*, vol. 14, no. 1, 87, 2024.

[30] D. K. D. Pal, S. Chitta, V. S. M. Bonam, P. Katari, and S. Thota, "AI-Assisted Project Management: Enhancing Decision-Making and Forecasting," *J. Artif. Intell. Res*, vol. 3, pp. 146–171, 2023.

[31] R. Spain, W. Min, V. Kumaran, J. Pande, J. Saville, and J. Lester, "Applying Large Language Models to Enhance Dialogue and Communication Analysis for Adaptive Team Training," *International Journal of Artificial Intelligence in Education*, pp. 1–35, 2025.

[32] M. A. Deif, A. Bounceur, S. Mouakket, M. A. Hafez, and M. Elhoseny, "A Multi-Agent Framework for Detecting and Forecasting Silent Resistance in Organizational Communication: Implications for Open Innovation Dynamics," *Journal of Open Innovation: Technology, Market, and Complexity*, 100646, 2025.

[33] Github. (September 2025). *microsoft/topic_conversation.* Python. Microsoft. [Online]. Available: https://github.com/microsoft/topic_conversation

[34] I. Gambo, R. Massenon, R. O. Ogundokun, S. Agarwal, and W. Pak, "Identifying and resolving conflict in mobile application features through contradictory feedback analysis," *Heliyon*, vol. 10, no. 17, e36729, 2024. doi: 10.1016/j.heliyon.2024.e36729