

An LLM-Augmented Framework for Risk-Informed Schedule Planning in Airside Construction: Integrating CPM and Probabilistic Simulation

Hossein Taheri¹, Mana Samadi²

¹Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX, USA

²Independent Researcher, Dallas, TX, USA

Email: Hxt7099@mavs.uta.edu, Mah.smd.18@gmail.com

How to cite this paper: Taheri, H. and Samadi, M. (2026) An LLM-Augmented Framework for Risk-Informed Schedule Planning in Airside Construction: Integrating CPM and Probabilistic Simulation. *World Journal of Engineering and Technology*, **14**, 947-972.
<https://doi.org/10.4236/wjet.2026.144057>

Received: October 6, 2026

Accepted: October 27, 2026

Published: October 30, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Construction scheduling in active airport environments is complicated by restricted work windows, operational access requirements, and dependencies that may not be fully reflected in conventional critical-path analysis. This study examines an LLM-augmented framework for identifying schedule vulnerabilities by combining Critical Path Method (CPM) indicators, a heuristic Baseline Risk Index (BRI), large language model (LLM) interpretation, and probabilistic duration analysis. The framework was examined through an anonymized airside rehabilitation project at a major U.S. airport, using 1591 baseline schedule records and nine retrospective validation cases. Baseline schedule characteristics were analyzed separately from subsequent project outcomes to reduce hindsight bias. Monte Carlo simulation was applied at the activity level using provisional duration distributions and risk-event assumptions. The results revealed substantial differences in schedule flexibility across work packages. Work Package E had approximately 67% of its eligible activities within one to five days of total float and subsequently experienced a 32-day schedule variance. Work Package D recorded a 70-day variance, despite having lower low-float exposure. Meanwhile, the project's substantial-completion milestone was achieved on its baseline target date. The LLM helped consolidate schedule-risk indicators into structured observations for planner review, although no independent improvement over conventional CPM and BRI analysis was demonstrated. The probabilistic results provided comparative activity-duration exposure rather than project completion forecasts. These findings suggest that combining conventional schedule measures with contextual risk screening and structured interpretation can support earlier recognition of localized

vulnerabilities, while further validation is needed to establish measurable improvements in planning performance.

Keywords

Large Language Models, Construction Scheduling, Airfield Construction, Schedule Risk, Critical Path Method, Monte Carlo Simulation, Probabilistic Scheduling, Risk-Informed Planning, Artificial Intelligence

1. Introduction

Construction scheduling requires more than assigning dates to activities. A reliable baseline schedule must represent work sequence, activity relationships, access conditions, contractual milestones, and the practical constraints under which the work will be performed. The Critical Path Method (CPM) provides a structured basis for this process through activity logic, total float, and schedule-driving paths. However, CPM remains deterministic: once durations, calendars, logic, and constraints are defined, the resulting dates are treated as fixed even though actual construction conditions are uncertain.

This limitation is especially important in airfield construction. Work in an active airport is often controlled by restricted access, limited closure windows, aircraft operations, safety requirements, and coordination among multiple organizations. These conditions can affect when a work package can begin, continue, or be completed, even when the underlying construction activity is technically straightforward. Previous research on airfield construction has emphasized the importance of early coordination, shared situational awareness, and proactive risk management in operationally constrained environments [1]-[3].

Artificial intelligence has created new opportunities for improving construction planning and schedule management. Recent studies have examined the use of natural language processing, machine learning, Building Information Modeling, and large language models for schedule generation, schedule interpretation, and dynamic project control. Prieto *et al.*, Jung *et al.*, and Naser [4]-[6] demonstrated that ChatGPT could generate a coherent preliminary construction schedule from natural-language input, while Chong *et al.* [7] showed how BIM and AI can support dynamic schedule management and risk identification. Donia *et al.* [8] further demonstrated that NLP-based methods can generate activities, durations, and logical relationships compatible with Primavera P6. These studies show the potential of AI-assisted scheduling, but they focus primarily on schedule generation, automation, or execution-stage adjustment.

A different opportunity is to use large language models as a reasoning layer over an existing baseline schedule. In this role, the LLM does not replace CPM calculations or professional scheduling judgment. Instead, it helps interpret combinations of baseline characteristics that may indicate schedule vulnerability [9] [10].

These characteristics can include low or zero float, driving-path exposure, long-duration activities, access restrictions, external approvals, third-party interfaces, procurement exposure, and other operational constraints. The present study represents these characteristics through a transparent Baseline Risk Index (BRI) and then uses the LLM to interpret the resulting risk patterns and identify areas that warrant planner attention.

Probabilistic simulation provides a complementary analytical layer. CPM establishes the logical structure of the baseline, while Monte Carlo simulation represents uncertainty in activity durations and selected risk events. The two approaches therefore serve different purposes: CPM describes schedule structure, the BRI summarizes baseline exposure, the LLM supports interpretation, and probabilistic simulation characterizes uncertainty. Historical project outcomes are kept separate from the baseline analysis and are introduced only after the initial assessment to reduce hindsight bias.

This study applies the framework to an anonymized airside rehabilitation project at a major U.S. airport. The case is useful because the project cannot be described simply as delayed or on time. The substantial-completion milestone was achieved on its baseline target date, while several intermediate work packages experienced later openings, sequencing changes, deferred mobilization, access or interface constraints, or scope removal. This creates an appropriate setting for examining whether localized schedule vulnerability could have been recognized from baseline information before the later outcomes were known.

Accordingly, this study proposes an LLM-augmented framework for risk-informed schedule planning in airside construction. The framework combines baseline CPM information, risk-feature extraction, the Baseline Risk Index, LLM-assisted interpretation, probabilistic simulation, planner review, and historical validation. The study has three objectives: first, to identify schedule vulnerability using baseline information alone; second, to evaluate the role of an LLM as a decision-support layer for interpreting concentrations of schedule risk; and third, to compare baseline-derived risk signals and probabilistic exposure with later observed project outcomes. The broader aim is to examine how AI can augment conventional scheduling practice without replacing established CPM methods or professional judgment.

2. Literature Review

2.1. AI and Natural Language Processing in Construction Scheduling

Construction scheduling has traditionally depended on structured planning tools and the experience of schedulers to define activity breakdowns, durations, relationships, and constraints. More recently, artificial intelligence has been introduced to automate portions of this process and to improve the way project teams interact with schedule information. Early applications largely focused on machine learning, BIM integration, optimization, and automated extraction of project data.

Natural language processing has expanded this direction by allowing construction information written for human use to be translated into structured data that can support planning and project control [11].

Prieto *et al.* [4] provided one of the early demonstrations of a conversational large language model applied directly to construction scheduling. Their study showed that ChatGPT could generate a reasonable activity sequence from a natural-language project description and respond to changes introduced by users. The results suggested that LLMs could reduce some of the effort associated with preliminary schedule development, although reliability, consistency, and professional verification remained important limitations.

More recent studies have moved beyond basic schedule generation. Chong *et al.* [7] integrated BIM and AI within a dynamic schedule-management framework that combined project information, predictive analysis, resource allocation, and progress visualization. Their results illustrate a broader role for AI in schedule management, particularly where the system is connected to structured project data rather than used as an independent prediction tool.

Donia *et al.* [8] developed this idea further through an integrated BIM-NLP framework for automated schedule generation. Their approach used semantic matching and rule-based reasoning to generate activities, estimate durations, establish relationships, and create schedule information compatible with Primavera P6. The work demonstrates that AI can participate directly in scheduling workflows while preserving links to conventional project-management software. At the same time, the authors noted that uncertainty associated with productivity, weather, and implementation risk was outside the scope of their framework. This limitation is particularly relevant when the objective moves from schedule generation toward schedule-risk assessment.

2.2. Large Language Models as a Planning and Decision-Support Layer

The growing capability of large language models creates an opportunity that extends beyond automatically writing an activity list. LLMs can interpret textual and structured information together, recognize relationships within large datasets, and provide an interface through which users can investigate complex project information.

Seo and Choi [12] proposed an LLM-based agent system for large-scale project scheduling in which the language model was connected to a schedule knowledge graph and specialized analysis, modification, and optimization tools. Their findings are important for the present study because performance improved when the LLM was supported by external computational tools rather than being expected to perform the entire scheduling task internally. The authors also emphasized a human-in-the-loop arrangement in which the system can propose or analyze changes while the project manager retains responsibility for decisions.

A similar principle is appearing in other construction applications. Abdelsalam

et al. [13] for example, connected an LLM with BIM and external cost databases to automate a multi-step construction estimating workflow. Although cost estimation differs from scheduling, the underlying architecture illustrates an important direction in engineering AI: the language model acts as an orchestration and reasoning layer while specialized systems remain responsible for domain-specific data and calculations.

This distinction is central to the framework proposed in this study. CPM calculations, schedule logic, and probabilistic simulation do not need to be replaced by an LLM. Instead, the LLM can sit above these analytical tools and help the planner interpret patterns that may otherwise require substantial manual review. In schedule-risk analysis, these patterns may include combinations of low float, driving-path exposure, restrictive access, long durations, approval requirements, third-party interfaces, or other project-specific constraints. The value of the language model therefore lies less in producing a finish date and more in helping determine where the schedule deserves additional attention and why.

2.3. Probabilistic Analysis and Schedule Uncertainty

Deterministic schedules provide a necessary baseline for project execution, but they represent durations and relationships as defined planning assumptions. Actual construction conditions are less certain. Duration variability, access limitations, external dependencies, resource constraints, and operational changes can alter the way activities are ultimately executed.

Probabilistic simulation provides a way to represent this uncertainty explicitly. Rather than assigning only one duration to an uncertain activity, the input can be represented as a distribution and evaluated repeatedly. Jafari *et al.* [14] although addressing construction reporting requirements rather than construction scheduling directly, demonstrated this combined approach by linking NLP-based information extraction with stochastic simulation. Their model used Monte Carlo simulation to represent uncertainty in time and resource requirements and allowed triangular distributions based on minimum, most-likely, and maximum estimates when sufficient historical observations were unavailable.

For schedule planning, the same principle can complement CPM. CPM describes the network structure and identifies where activities have limited schedule flexibility. Probabilistic simulation can then examine the uncertainty associated with selected activities or risk conditions. These two functions should remain distinct. A high-risk interpretation from an LLM is not a probability of delay, and an activity-level Monte Carlo result is not automatically a project completion-date probability. A defensible framework therefore requires a clear separation between schedule logic, risk interpretation, and probabilistic calculation.

This distinction is also important for the present case study. The preliminary Monte Carlo component is used to develop and test the analytical workflow at the activity and schedule-grouping levels. Full propagation of uncertainty through the CPM network requires calibrated probability distributions and is treated as a sub-

sequent stage of model development rather than assumed from the exploratory simulation.

2.4. Airfield Construction and the Research Gap

Airfield construction provides a particularly demanding setting for risk-informed scheduling. The schedule must account not only for construction productivity and technical logic but also for an operating airport. Access to work areas, aircraft movements, temporary closures, safety requirements, operational coordination, and interfaces among multiple organizations can influence whether planned work can proceed when scheduled.

Taheri [1] discussed these conditions in the context of airfield construction safety management and highlighted the importance of timely information sharing, stakeholder coordination, and early recognition of operational risks. Taheri *et al.* [2] similarly emphasized the value of combining prescriptive controls with proactive risk identification and continuous risk management in airport environments. Together, these studies support the broader premise that airport decision-making benefits from approaches that identify emerging risk before it becomes an operational problem.

The scheduling literature reviewed above shows substantial progress in AI-assisted schedule generation, NLP-based interpretation, BIM integration, dynamic schedule management, and LLM-enabled interaction with project data. However, the reviewed studies do not directly address the combination examined here: using an LLM to interpret risk signals from an existing baseline CPM schedule, linking those signals to probabilistic uncertainty analysis, and then testing the baseline-derived assessment against later project outcomes that were withheld from the initial analysis.

The present study addresses that gap through an anonymized airside construction case study. Rather than asking AI to create a schedule from scratch, the proposed framework begins with an established baseline schedule and asks a different question: can AI help the planner recognize where the baseline is vulnerable before the subsequent project outcome is known? CPM provides the schedule structure, the LLM supports interpretation of risk characteristics, probabilistic simulation represents uncertainty, and historical project records provide an independent basis for evaluating the resulting risk signals. This positioning places the proposed framework between conventional schedule analysis and fully automated AI scheduling, with the objective of supporting professional judgment rather than replacing it.

3. Materials and Methods

3.1. Research Design

This study used a retrospective case-study design to examine whether schedule risks could have been identified from information available at the baseline stage, before later project outcomes were known. The method was intentionally struc-

tured to separate prediction from validation. Baseline schedule information was analyzed first, while subsequent project records were withheld and used only after the risk-identification process had been completed.

The proposed framework combines three analytical components: CPM-based schedule information, LLM-assisted interpretation of schedule risk, and probabilistic simulation. The overall workflow is:

Baseline schedule → risk-feature extraction → LLM-assisted risk interpretation → probabilistic simulation → risk review → historical validation.

The overall analytical workflow is summarized in **Figure 1**. The LLM was not used to calculate CPM dates or independently generate probabilistic results. CPM calculations were retained as structured schedule data, and Monte Carlo simulation was performed computationally. The LLM was used as an interpretive layer to organize baseline risk signals and identify activities or work packages that warranted closer planning attention. This approach follows the broader direction of tool-supported LLM applications in engineering, where the language model assists with reasoning while specialized analytical tools perform deterministic or numerical calculations [12].

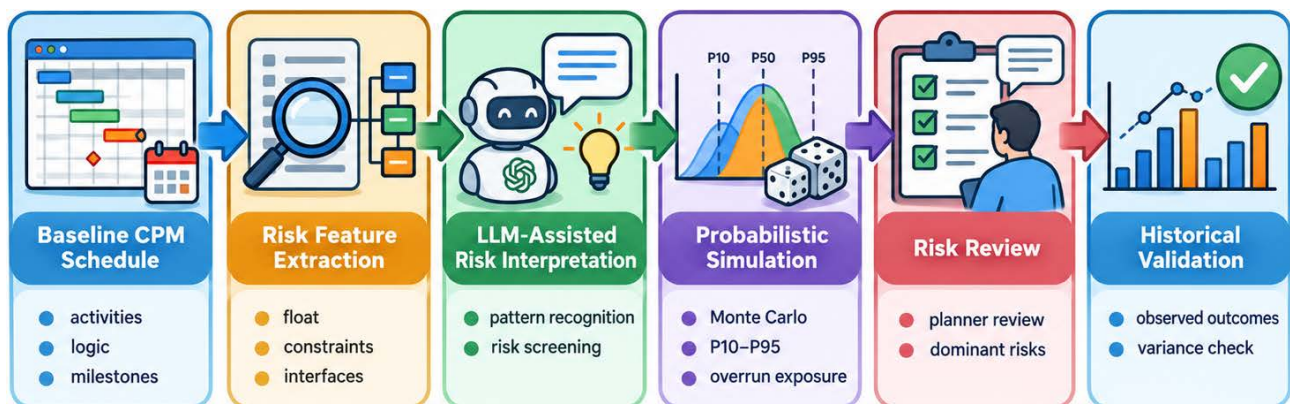


Figure 1. Proposed LLM-augmented workflow for baseline schedule-risk screening, probabilistic analysis, planner review, and historical validation. Historical outcomes are withheld from the baseline analysis and introduced only during the validation stage.

3.2. Case Study and Data Preparation

The framework was evaluated using an anonymized airside rehabilitation project at a major U.S. airport. The original project schedule was developed in Primavera P6 and contained the activity structure, work breakdown structure, activity durations, calendars, logical relationships, constraints, milestones, and float information required to reconstruct the baseline planning condition.

To protect project-specific and operationally sensitive information, identifiable project names, organizations, activity locations, runway or taxiway identifiers, and other potentially sensitive descriptions were removed or generalized. Dates were converted to relative Project Day values rather than reported as calendar dates. The analysis therefore preserves the temporal relationships required for schedule evaluation without disclosing the original project identity.

The research dataset was organized into separate analytical layers. Baseline activities and CPM attributes formed the primary schedule dataset. Predecessor and successor relationships were retained separately to preserve network structure, while baseline milestones and constraints were maintained as an additional reference layer. Engineered risk features were then derived from these baseline datasets.

Historical project outcomes were stored separately from the baseline information. These records included later-reported milestone dates, changes in work-package completion, scope adjustments, access limitations, and other documented schedule outcomes. Nine anonymized validation cases were identified for comparison with the baseline analysis. Importantly, these historical records were not supplied to the LLM during the primary risk-identification stage.

This separation was used to reduce hindsight leakage. The intent was to reproduce, as closely as possible, the information environment that would have existed when the baseline schedule was originally available.

3.3. Baseline Schedule-Risk Features

The baseline schedule was evaluated using a combination of conventional CPM indicators and contextual risk characteristics. Conventional indicators included total float, low-float exposure, zero-float exposure, activity duration, driving-path status, constraints, and network connectivity.

Additional features were introduced to represent conditions that may be particularly important in operationally constrained construction environments. These included:

- restricted-access exposure;
- third-party or external interfaces;
- approval dependencies;
- procurement exposure;
- weather sensitivity;
- long-duration activities;
- cost exposure; and
- dependencies associated with operational sequencing.

The contextual risk indicators were developed through provisional baseline-only screening of anonymized activity descriptions and schedule classifications. External approval (EA) represented activities involving submittals, permitting, reviews, or approval requirements. Procurement exposure (PR) represented material procurement, submittal, review, and delivery dependencies. Restricted access (RA) represented activities associated with gates, fencing, barricades, controlled access points, or restricted work areas. Third-party interface (TP) represented dependencies involving suppliers, utility providers, external approvals, or other stakeholders. These indicators were treated as heuristic screening flags rather than independently verified records of actual risk events. Their activity-level assignments were used to calculate the Baseline Risk Index (BRI) for each activity.

The activity-level risk-feature classifications were recorded as part of the baseline screening process. Among the 1591 baseline records, 47 were flagged for external approval, 73 for procurement exposure, 105 for restricted access, and 79 for third-party interfaces. The binary classifications and associated BRI values were preserved independently of the historical validation outcomes.

The Baseline Risk Index (BRI) for activity (i) was calculated as:

$$\text{BRI}_i = 20\text{ZF}_i + 12\text{LF}_i + 10\text{DP}_i + 8\text{C}_i + 5\text{M}_i + 5\text{LD}_i + 8\text{EA}_i + 8\text{PR}_i + 10\text{RA}_i + 8\text{TP}_i + 5\text{W}_i + 6\text{HC}_i + 5\text{CW}_i + 5\text{CE}_i$$

where ZF denotes zero float, LF denotes 1 - 5 days of total float, DP denotes driving-path status, C denotes a schedule constraint, M denotes milestone status, LD denotes long duration (≥ 20 days), EA denotes external approval, PR denotes procurement exposure, RA denotes restricted access, TP denotes third-party interface, W denotes weather sensitivity, HC equals 1 when the combined number of direct predecessor and successor relationships is five or more, and 0 otherwise. CW equals 1 when at least one predecessor belongs to a different work breakdown structure (WBS) node, and 0 otherwise. CE equals 1 when the activity's budgeted cost share exceeds 0.5% of the total baseline project cost, and 0 otherwise. All terms are binary indicators. The index is a heuristic screening measure and is not a probability of delay.

The weather-sensitivity indicator was assigned a value of 1 for all 1591 baseline records. As a result, it contributes five points to every BRI score without affecting the relative ranking of activities. Although weather conditions can influence airside construction, the available data did not support assigning different levels of weather exposure to individual activities. The indicator was therefore retained for consistency with the original calculations but was not used to distinguish weather-related risk among activities.

Low-float exposure (LFE) and zero-float exposure (ZFE) were calculated at the schedule-grouping level as:

$$\text{LFE}_j = \left[\sum I(1 \leq \text{TF}_i \leq 5) / N_j \right] \times 100$$

$$\text{ZFE}_j = \left[\sum I(\text{TF}_i = 0) / N_j \right] \times 100$$

where TF_i is the total float of activity (i), I is an indicator function equal to 1 when the stated condition is satisfied and 0 otherwise, and N_j is the number of eligible activities included in schedule grouping (j). LFE represents the percentage of activities with 1 - 5 days of total float, while ZFE represents the percentage of activities with zero total float.

Eligible activities were defined as task-type schedule records, excluding 26 start milestones, 26 finish milestones, and 26 level-of-effort activities, leaving 1513 eligible activities from the 1591 baseline records. Schedule-grouping membership was determined using the hierarchical WBS paths. The major airside closure grouping included records within Phase 2, while intermediate Work Packages B, D, E, F, G, and H corresponded to their respective Phase 2B, 2D, 2E, 2F, 2G, and 2H WBS branches. Both project-level completion groupings included all 1513 el-

igible activities. The resulting denominators (N_j) were 839 for Phase 2 overall, 103 for B, 87 for D, 43 for E, 5 for F, 20 for G, and 50 for H. The hierarchical groupings were not mutually exclusive.

The BRI weights were assigned as heuristic values to reflect the relative importance of the selected schedule-risk indicators. Greater weights were given to zero float, low float, and restricted-access conditions, while the remaining indicators received lower weights according to their intended contribution to baseline risk screening. The weights were not statistically calibrated, and the historical validation outcomes were not used to adjust them. Therefore, the BRI was used only to compare relative risk exposure and was not treated as a predictor of actual schedule delay.

The purpose of the feature set was not to declare that an activity would be delayed. Rather, the features were used to characterize the degree to which an activity or schedule grouping was exposed to conditions that could reduce schedule flexibility.

The resulting BRI was used to support comparative analysis among activities and schedule groupings. This distinction is important because a score of 60, for example, should not be interpreted as a 60% probability that an activity will finish late.

At the schedule-grouping level, additional measures were calculated, including the percentage of activities with zero float and the percentage with one to five days of total float. These measures were used to identify concentrations of limited schedule flexibility rather than relying only on individual critical activities.

3.4. LLM-Assisted Risk Interpretation

The LLM was provided with structured baseline information derived from the schedule and engineered risk features. Its role was to interpret combinations of schedule characteristics and identify potential concentrations of risk that might warrant review by a planner.

The LLM received only baseline-derived schedule information, including activity and WBS identifiers, original duration, total float, driving-path and constraint indicators, and the contextual risk features defined above. Historical validation outcomes were excluded from the input. A fixed analysis prompt instructed the model to identify concentrations of baseline vulnerability, explain the contributing schedule characteristics, and flag areas requiring planner review without recalculating CPM dates or predicting specific delay durations. Historical outcomes were introduced only after completion of the baseline analysis. The LLM-assisted baseline risk analysis was conducted using ChatGPT with a session-identified GPT-6 model (OpenAI). The analysis incorporated 1591 anonymized baseline schedule records, 2750 predecessor-successor relationships, and their corresponding baseline risk features. Historical project outcomes were excluded from the model inputs. The analysis produced five evidence-based observations identifying potential schedule vulnerabilities, supporting activity characteristics, and areas re-

quiring professional planning review. The analysis focused on questions such as whether low-float activities were concentrated within a work package, whether critical or driving activities also depended on restricted access or external interfaces, and whether multiple risk characteristics occurred simultaneously in the same portion of the schedule.

This design differs from studies that use an LLM primarily to generate a construction schedule from natural-language input [4]. In the present framework, the baseline schedule already exists. The purpose of the language model is therefore not to replace schedule development, but to assist with interpretation of an established CPM network.

Human judgment remains part of the process. LLM-generated observations are treated as decision-support information that a scheduler or project manager can review before considering mitigation measures. This human-in-the-loop arrangement is consistent with recent tool-augmented scheduling research in which analytical recommendations are generated computationally but managerial responsibility remains with the project team [12].

The LLM output was used as a qualitative screening and interpretation layer rather than as an independently calibrated predictive model. Its observations were reviewed in conjunction with the structured baseline indicators, including the Baseline Risk Index (BRI), float exposure, access constraints, and interface conditions. The study therefore does not assign a separate prediction-accuracy statistic to the LLM. Instead, the LLM contribution is evaluated as part of the integrated framework by examining whether baseline-only risk interpretations were consistent with areas that later warranted schedule attention.

The baseline risk assessment was carried out through the ChatGPT interface using a session-identified GPT-6 model. The model was provided with anonymized activity records, predecessor-successor relationships, and provisional baseline risk features. It was instructed to identify three to five schedule-risk observations supported by activity identifiers, total float, network relationships, and relevant contextual indicators. The instructions also required the model to explain the scheduling implications of each observation and recommend areas for planner review without assuming that future delays or risk events had occurred. Historical validation records and Monte Carlo results were not included in the model inputs. Because the analysis was performed through the ChatGPT interface, generation settings such as temperature and random seed were not available for independent verification.

3.5. Probabilistic Simulation

Monte Carlo simulation was used to introduce duration uncertainty into the analysis. For each simulated activity, a triangular distribution was defined using minimum, most-likely, and maximum duration assumptions. Triangular distributions are commonly used when complete historical distributions are unavailable but reasonable lower, most-likely, and upper estimates can be established from

available information or expert judgment. A similar approach has been used in construction simulation research to represent uncertain duration and resource inputs [14].

For the exploratory Monte Carlo analysis, minimum, most-likely, and maximum activity durations were estimated using baseline durations and the corresponding BRI categories. Risk-event probabilities and duration impacts were assigned according to the same risk categories and the relevant activity-level exposure indicators. These assumptions were kept separate from the withheld historical validation outcomes and were not calibrated using the later observed schedule results. The provisional triangular duration bounds were assigned using the baseline activity duration and three BRI categories: Low (BRI below 25), Medium (BRI 25 - 44), and High (BRI 45 or greater). The optimistic duration was set at 85%, 90%, and 95% of the baseline duration for High-, Medium-, and Low-risk activities, respectively, while the most-likely duration equaled the baseline duration. The corresponding pessimistic multipliers were 1.50, 1.30, and 1.15. For activities flagged for external approval or procurement, the pessimistic multiplier was increased to 1.60, or 1.70 when restricted-access exposure was also identified. Discrete risk-event probabilities were assigned only to activities flagged for external approval, procurement, restricted access, or third-party interfaces, using provisional probabilities of 0.05, 0.15, and 0.30 for Low-, Medium-, and High-risk activities, respectively. The associated event-impact intervals were 1 - 5, 2 - 10, and 3 - 21 days. The duration multipliers, event probabilities, and impact ranges were selected as provisional assumptions to examine how different levels of baseline risk could influence activity-duration uncertainty. These values were not derived from historical frequency data or calibrated against observed project outcomes. Accordingly, the simulation results were used to compare relative duration exposure under the assumed conditions rather than to estimate the likelihood of actual schedule delays. Because the input distributions remain provisional, the simulation is used to compare relative duration exposure rather than to produce a calibrated forecast of package or project completion.

The simulation generated repeated duration samples and produced activity-level statistics including mean simulated duration, P10, P50, P80, P90, and P95 durations, expected overrun relative to the baseline, and probability of exceeding the baseline duration.

The provisional activity-level Monte Carlo analysis was reproduced using Python 3.13.5 and NumPy 2.3.5, with 20,000 iterations per simulated activity and a fixed random seed of 20261012 using the PCG64 random-number generator. For each iteration, activity duration was sampled from the corresponding triangular distribution defined using the duration assumptions described above. Discrete risk events were modeled through independent Bernoulli trials using the assigned event probabilities; when triggered, an additional duration impact was sampled from a uniform distribution bounded by the specified minimum and maximum impact values. The resulting activity-duration samples were used to calculate

mean duration, empirical percentiles, and the proportion exceeding baseline duration. Activity uncertainties were sampled independently, without CPM-network propagation or cross-activity correlation. The fixed random seed and documented simulation settings support consistency across computational runs. However, independent reproduction of the activity-level results requires access to the complete input parameters and simulation code.

At the schedule-grouping level, activity results were aggregated to support comparative visualization. Particular attention was given to average P80 duration-overrun exposure and expected overrun measures. Of the 1591 baseline schedule records, 1513 activities were included in the activity-level simulation. Seventy-eight records were excluded, consisting of 26 start milestones, 26 finish milestones, and 26 level-of-effort activities.

These results require an important qualification. The simulation used in the present exploratory stage is activity-based. Uncertainty was not propagated through a complete CPM network during every Monte Carlo iteration. Consequently, the schedule-grouping-level P80 measures presented in this study should not be interpreted as probabilities of project or package completion dates.

The probability distributions and event assumptions also remain subject to calibration. The Monte Carlo results are therefore used to demonstrate the analytical framework and compare relative exposure among schedule components, rather than to claim a final probabilistic completion forecast.

3.6. Historical Validation

After the baseline-only analysis had been completed, historical project outcomes were introduced for validation. Nine anonymized validation cases were used. These included major project milestones and intermediate work packages that experienced different types of schedule outcomes.

For cases with both a baseline target and a later-reported completion date, schedule variance was calculated as:

Schedule Variance = Later Reported Project Day – Baseline Target Project Day

Positive values indicate completion later than the baseline target, negative values indicate earlier completion, and zero indicates completion on the baseline target.

Not every historical outcome could reasonably be expressed as a numerical delay. Cases involving scope removal, deferred mobilization, or unresolved outcomes were therefore retained as categorical observations rather than assigned an artificial number of delay days.

This validation strategy is important because the case study did not experience a simple project-wide delay. The substantial-completion milestone was achieved on its baseline target date even though several intermediate work packages experienced schedule changes. The evaluation therefore focused on whether the proposed framework identified localized schedule vulnerability, rather than whether it simply predicted the final project completion date correctly.

3.7. Evaluation and Visualization

The final analytical stage compared baseline-derived risk indicators with the historical validation cases. Three forms of evidence were considered together:

1. conventional CPM exposure, particularly zero- and low-float activity concentration;
2. baseline-derived risk indicators, including the BRI, together with LLM-assisted qualitative interpretations; and
3. exploratory Monte Carlo duration-overrun exposure.

The comparisons were visualized at the schedule-grouping level using scatter plots and bar charts. Observed schedule variance was then compared with baseline risk characteristics to determine whether areas that later experienced schedule disruption had shown recognizable warning signals during baseline planning.

The evaluation was conducted as a retrospective comparison rather than a formal test of predictive accuracy. The nine validation cases included five outcomes with measurable schedule variance, three categorical outcomes involving scope changes or deferred mobilization, and one pending milestone. Given the limited number of cases and differences in outcome types, the analysis focused on whether baseline risk indicators were consistent with the later project records. The findings were interpreted at the case-study level and were not used to establish statistical relationships or predictive performance beyond the project examined.

4. Results

4.1. Baseline Schedule-Risk Exposure

The baseline analysis showed that schedule flexibility was not distributed evenly across the project. Some work packages contained a much larger concentration of activities with limited total float than others. This was particularly evident for Intermediate Airside Work Packages D and E.

Work Package E had the highest low-float exposure, with approximately 67% of its activities having between one and five days of total float. Work Package D followed at approximately 28%. By comparison, Work Package B had only about 5% of its activities within this range. These differences suggest that the packages were entering execution with substantially different levels of schedule flexibility.

Among the intermediate work packages, Work Package B had the highest zero-float exposure at approximately 11.7%, followed by Work Package D at approximately 5.8%. The project-to-substantial-completion and project-to-final-completion groupings each showed approximately 6.9% zero-float exposure, while the major airside closure grouping showed approximately 5.4%. Work Packages E, F, G, and H contained no zero-float activities under the adopted analytical definition. The comparison illustrates why a single CPM indicator is insufficient for risk screening. Work Package E, for example, had no zero-float activities but had a very large concentration of activities with only one to five days of float. Its baseline

was therefore not technically dominated by zero-float activities, but it nevertheless contained relatively little schedule flexibility across a substantial portion of the package. **Figure 2** summarizes low-float exposure across the analyzed schedule groupings.

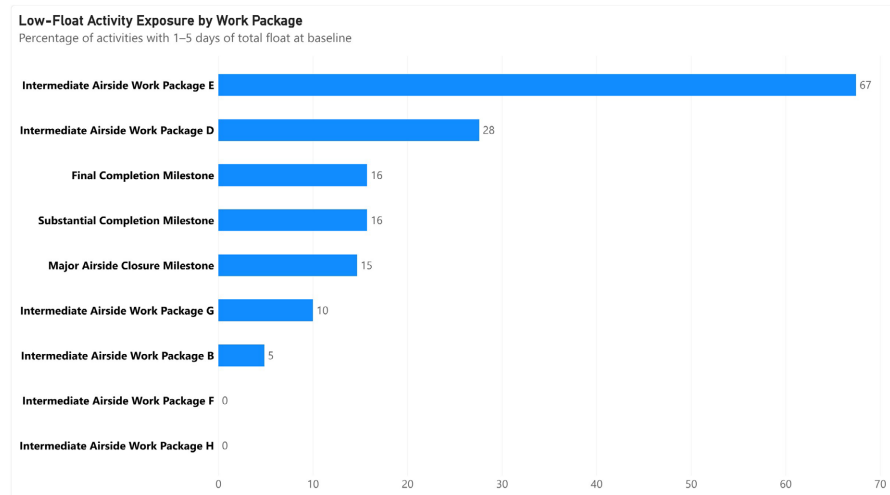


Figure 2. Low-float activity exposure across the analyzed schedule groupings, including intermediate work packages and milestone groupings. Values represent the percentage of eligible baseline activities with one to five days of total float.

4.2. Historical Schedule Outcomes

The historical validation records showed that project performance cannot be described adequately using the final milestone alone. The substantial-completion milestone was achieved on its baseline target date, and the major airside closure milestone occurred seven days earlier than its baseline target. At the same time, several intermediate work packages experienced substantial deviations from their original baseline dates.

Among the cases for which directly comparable dates were available, Work Package D recorded the largest variance, with its later-reported date occurring 70 days after the baseline target. Work Package E was 32 days later than its baseline package finish, while Work Package B experienced an 11-day delay in opening.

Other outcomes were not appropriately represented by a numerical delay value. Work Packages F and G were removed from the relevant scope, while Work Package H experienced deferred mobilization associated with access or operational conditions. These cases were retained as categorical outcomes because assigning an artificial delay duration would misrepresent what occurred.

The difference between intermediate work-package outcomes and the on-time substantial-completion milestone suggests that localized schedule changes did not ultimately prevent the project from meeting its principal contractual milestone. These results may reflect schedule adjustments, resequencing, or other management actions during construction, although the available records do not establish the specific contribution of each measure. **Figure 3** illustrates the observed sched-

ule variances for the five historical validation cases with comparable baseline and actual completion dates.

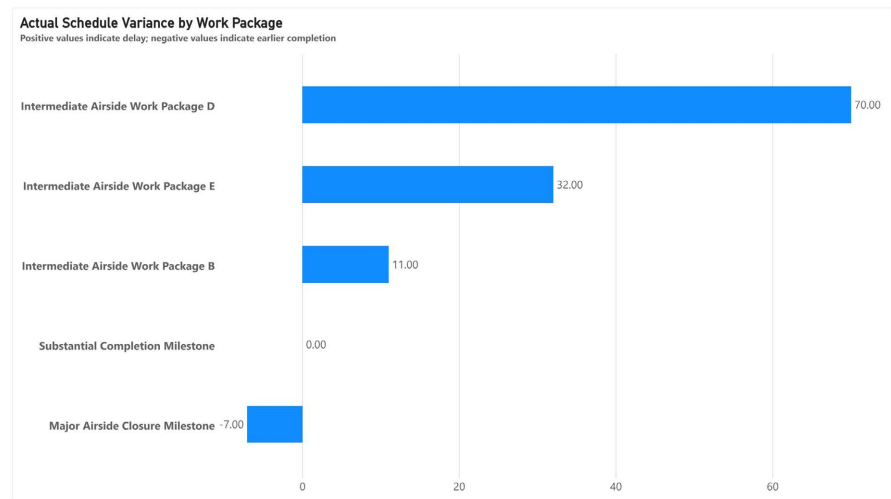


Figure 3. Actual schedule variance for historical validation cases with comparable dates. Positive values indicate later completion relative to baseline, negative values indicate earlier completion, and zero indicates completion on the baseline target.

4.3. Baseline Risk Index and Observed Variance

The schedule-grouping-level comparison between average Baseline Risk Index (BRI) values and later schedule variance produced a mixed but informative pattern. Work Packages D and E, both of which later experienced substantial schedule variance, showed elevated baseline risk exposure relative to some other intermediate packages. Work Package E had the highest average BRI among the three delayed packages included in the numerical comparison, while Work Package D had a lower average BRI but ultimately experienced the largest observed variance. The relationship was therefore not monotonic. A higher average Baseline Risk Index (BRI) did not translate directly into a greater number of delay days. This is an important result because the Baseline Risk Index (BRI) was designed as a relative indicator of schedule exposure, not as a calibrated prediction of delay magnitude.

The milestone cases reinforce this distinction. The substantial-completion milestone was achieved on target, while the major closure milestone occurred earlier than baseline, even though the associated schedule groupings still contained measurable baseline risk exposure. Risk identification should therefore be interpreted as an indication of vulnerability or management attention, rather than a deterministic forecast that the corresponding milestone will be missed.

The comparison suggests that the primary value of the LLM-assisted interpretation layer may lie in identifying where additional review is warranted, rather than estimating exactly how many days of variance will occur. This is consistent with the intended decision-support role of the framework. **Figure 4**

compares the average Baseline Risk Index (BRI) with observed schedule variance, with bubble size representing exploratory Monte Carlo P80 duration-overrun exposure.

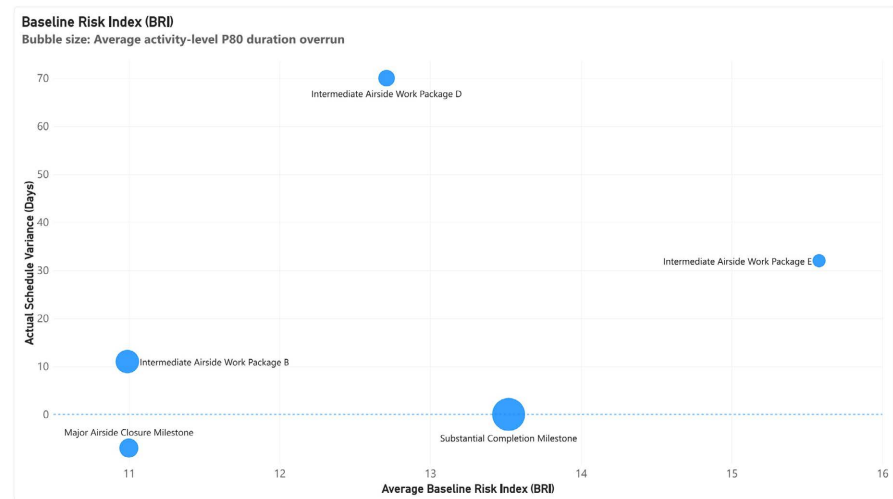


Figure 4. Average Baseline Risk Index (BRI) versus observed schedule variance for validation cases with comparable dates. Bubble size represents the average activity-level P80 duration overrun from the exploratory Monte Carlo simulation. The simulated values reflect duration uncertainty under provisional assumptions and do not represent predicted package completion delays.

4.4. Probabilistic Duration Exposure

The exploratory Monte Carlo simulation provided an additional view of uncertainty at the activity and schedule-grouping levels. Average P80 duration-overrun exposure varied across the analyzed schedule groupings. The project-level milestone groupings showed the largest average P80 exposure, at approximately 1.19 days per simulated activity. Among the intermediate work packages, Work Package B showed the highest average P80 exposure at approximately 0.57 days per activity. Work Packages F and D followed at approximately 0.34 and 0.25 days, respectively, while Work Packages E, G, and H showed lower average values. These values do not correspond directly with the observed schedule variance. For example, Work Package D experienced the largest historical variance even though its average activity-level P80 overrun was below that of Work Package B. Similarly, Work Package E experienced a 32-day historical variance despite having one of the smaller average simulated P80 overruns.

This divergence is expected given the present stage of the model. The Monte Carlo analysis was performed at the activity level using provisional duration distributions, and uncertainty was not propagated through the complete CPM network. The metric therefore describes relative activity-duration exposure rather than predicted package completion delay.

This result also highlights an important methodological lesson. Schedule risk is not determined by duration uncertainty alone. The consequence of a duration

change depends on where the affected activity sits within the network, the amount of available float, the sequence of downstream work, and the operational constraints surrounding that activity. A relatively small duration change on a strategically important activity may have a greater schedule consequence than a larger change on an activity with sufficient flexibility.

For that reason, the Monte Carlo results are treated here as one component of the integrated framework rather than as a standalone forecast.

4.5. Combined Interpretation of the Validation Cases

Taken together, the baseline CPM indicators, baseline-derived risk signals and LLM-assisted interpretations, probabilistic exposure measures, and historical records indicate that no single metric adequately explained the observed project outcomes. (Table 1)

Table 1. Summary of baseline schedule-risk observations identified through LLM-assisted analysis.

Risk observation	Supporting baseline evidence	Recommended planner review
Phase 0 - 1 approval interface	A0034: TF 0, BRI 71; A0073: TF 0, BRI 61; A0115: TF 0, five incoming cross-WBS relationships	Review approval sequencing and construction readiness requirements.
Paint-marking procurement	A0046: TF 4, BRI 51; A0085: TF 4, BRI 57; A0128: TF 4, BRI 49; A0205: TF 0, BRI 46	Review material lead times, submittal approvals, and installation readiness.
Phase 2B restricted-access preparation	A0294: TF 0, BRI 56, ten outgoing relationships, nine cross-WBS; A0292, A0321, A0339: TF 0	Review barricade sequencing, access release, and downstream work coordination.
Phase 2 - 3 handover	A1196: TF 0, BRI 59, 24 predecessors, 23 cross-WBS; linked activities A1195, A1198, A1197	Review milestone constraints and phase-transition dependencies.
Phase 4 closeout	A1534: 158-day original duration, TF 0, BRI 51; A1590: TF 0, BRI 59	Review closeout duration, punch-list requirements, and completion logic.

TF = total float (days); BRI = Baseline Risk Index; WBS = work breakdown structure. These are baseline-derived observations, not predictions of actual schedule delay.

The baseline-only LLM analysis identified a potential schedule vulnerability in Phase 2B involving activity A0294 (installation of lighted barricades, SW3P, and make-safe measures). This activity had zero total float, a provisional BRI of 56, and ten outgoing schedule relationships, nine of which crossed WBS boundaries. Its downstream activities included A0292, A0321, and A0339, each with zero total float. Beyond examining individual CPM float values or BRI scores, the LLM interpretation highlighted how restricted-access preparation could become a coordination point affecting multiple downstream work fronts. The planner's assessment indicated that this LLM observation primarily synthesized information already available through CPM and BRI analysis, rather than identifying an additional schedule vulnerability. Its contribution was therefore limited to organizing existing risk indicators into a structured planning-review recommendation. No independent improvement in risk identification or delay prediction was demonstrated.

Work Package E provides a useful example. It contained the greatest concen-

tration of low-float activities and subsequently finished 32 days later than its baseline package target. Work Package D combined meaningful low-float and zero-float exposure and later recorded the largest numerical variance among the validation cases. Work Package B had the highest zero-float exposure and experienced an 11-day delayed opening.

At the same time, the evidence does not support a simple rule in which greater baseline risk exposure inevitably produces greater delay. The substantial-completion milestone was achieved on target despite underlying schedule exposure, and the major closure milestone was achieved earlier than baseline. The results instead suggest that baseline vulnerability and realized project delay are related through the way risk is absorbed, mitigated, resequenced, or transferred through the schedule network.

The categorical validation cases provide additional context for the baseline risk assessment. Work Packages F and G were removed from the project scope, while Work Package H experienced deferred mobilization. These outcomes highlight project changes that cannot be adequately described by total float alone. Although the available records do not establish a direct relationship between the baseline risk indicators and these outcomes, they illustrate the relevance of considering scope, access, and interface conditions alongside conventional CPM measures.

Overall, the case study provides preliminary evidence that an LLM-supported framework can assist in organizing and interpreting multiple schedule-risk signals at the baseline stage. The present analysis does not establish the LLM as an independent predictor of schedule delay. Rather, its role is to support interpretation of quantitative and contextual baseline information and to direct planner attention toward areas requiring closer review. Likewise, the BRI should not be treated as a direct delay predictor, and the current Monte Carlo outputs should not be interpreted as completion-date forecasts. Together, these components provide a structured basis for identifying schedule vulnerability and examining where mitigation, resequencing, or additional contingency analysis may be appropriate.

5. Discussion

5.1. Interpreting Schedule Risk beyond the Final Milestone

One of the clearest findings from the case study is that project-level completion performance does not fully describe what occurred within the schedule. The substantial-completion milestone was achieved on its baseline target date, yet several intermediate work packages experienced delayed openings, sequencing changes, deferred mobilization, access limitations, or scope adjustments. A schedule can therefore appear successful when viewed only from the final milestone while still containing significant internal disruption.

This distinction matters for risk-informed planning. If schedule performance is evaluated only by the final completion date, many early warning signals can be overlooked. The present analysis indicates that low-float concentration, zero-float exposure, interface conditions, and restricted-access characteristics can reveal ar-

eas of schedule vulnerability even when the overall project milestone is ultimately protected.

The results also show that traditional CPM indicators should not be interpreted in isolation. Work Package E, for example, had no zero-float activities under the package-level definition used in this study, yet it had the highest concentration of activities with only one to five days of float and later experienced a substantial schedule variance. Conversely, the existence of zero-float activities did not automatically lead to a missed project milestone. The practical implication is that schedule risk is better understood as a combination of network position, available flexibility, operational constraints, and the ability of the project team to absorb or mitigate disruption.

5.2. Role of the LLM in Schedule Planning

Large language models may support schedule-risk interpretation by organizing multiple baseline indicators into coherent planning-review observations rather than replacing CPM calculations or professional scheduling judgment. Prior studies have demonstrated that LLMs can generate activities, interpret schedule information, and support interaction with complex project data [4] [12]. The present study extends this application toward baseline schedule-risk interpretation.

The baseline-only LLM analysis generated five observations concerning approval dependencies, procurement interfaces, restricted-access activities, phase transitions, and completion-related dependencies. These observations combined activity float, network relationships, and contextual risk indicators to identify areas potentially requiring closer review.

However, the planner's assessment of the representative Phase 2B observation indicated that the LLM primarily synthesized information already available through CPM, BRI, and schedule-logic records. Although the resulting explanation provided a structured summary of potential access and sequencing concerns, it did not establish an additional risk beyond conventional schedule analysis. Furthermore, the study did not include a controlled CPM/BRI-only comparison or quantitative measures of improvement in risk identification or planning efficiency.

The use of structured schedule data alongside LLM interpretation is consistent with previous research on AI-supported engineering applications [12] [13]. In this study, the LLM helped organize existing schedule-risk information into observations that could be reviewed by a planner. However, the findings do not demonstrate that the LLM identified risks beyond those available through conventional CPM and BRI analysis or improved schedule prediction. Further research comparing LLM-assisted and conventional planning reviews would be needed to determine whether this approach offers measurable benefits in risk identification, review time, or planning decisions.

5.3. Relationship between the Baseline Risk Index and Probabilistic Analysis

The results also demonstrate why LLM-assisted risk interpretation and Monte

Carlo simulation should remain separate but complementary components. The Baseline Risk Index (BRI) was intended to represent relative exposure based on baseline schedule characteristics. It was not designed as a probability of delay. Likewise, the exploratory Monte Carlo outputs represented activity-duration uncertainty and not the probability that an entire work package or contractual milestone would finish late. The lack of a simple one-to-one relationship between the Baseline Risk Index (BRI), P80 exposure, and observed delay should therefore not be interpreted as a failure of the framework. Rather, it shows that these measures describe different aspects of schedule risk. The LLM-assisted interpretation layer helps identify combinations of structural and contextual vulnerability, while probabilistic simulation represents uncertainty in activity duration. The realized project outcome is then influenced by CPM logic, available float, mitigation actions, resequencing, scope decisions, and field conditions.

This is particularly visible in Work Packages D and E. Both showed meaningful baseline warning signals and later experienced schedule variance, but the magnitude of the realized delay did not follow the Baseline Risk Index (BRI) or the activity-level Monte Carlo result proportionally. This suggests that the next stage of the research should focus less on creating a single combined risk number and more on understanding how multiple indicators interact within the CPM network.

5.4. Implications for Airfield Construction Planning

Airfield construction is particularly suited to this type of layered analysis because operational constraints can influence schedule performance in ways that are not fully represented by activity duration alone. Access to a work area may be available only during specific operating windows. Work may depend on another contractor, airport operations, inspection, clearance, or temporary closure. These conditions can create schedule exposure even when the underlying construction activity is technically straightforward.

The categorical validation cases in this study are especially important in this regard. Scope removal, deferred mobilization, and interface-driven changes are difficult to describe using a simple delay-duration model [15] [16]. Yet they are highly relevant to the quality of the original schedule plan. A risk-informed planning framework should therefore identify not only activities that may take longer than expected but also activities whose execution depends on conditions outside the direct control of the construction team.

For practitioners, this suggests a useful planning workflow. A conventional CPM schedule can first establish the logical baseline. LLM-assisted interpretation can then help screen the schedule for concentrations of operational and interface risk identified through the baseline indicators and BRI. Probabilistic simulation can be used to examine uncertainty in selected activities. The planner can then evaluate whether additional float protection, alternative sequencing, earlier approvals, access coordination, or contingency planning is warranted before work begins.

This approach is more practical than treating AI as a replacement for Primavera P6 or professional scheduling judgment. The value is in improving the quality of the review performed around the schedule [17]-[19].

5.5. Study Limitations

Several limitations should be recognized when interpreting the findings.

First, the research is based on a single anonymized airfield rehabilitation project. The case provides a realistic and operationally complex environment, but the results should not be assumed to generalize directly to all airport or construction projects.

Second, the Baseline Risk Index (BRI) used in the current analysis is a provisional research construct. It provides a structured method for comparing baseline risk characteristics, but it has not yet been calibrated as a predictive model using a large population of completed projects.

Third, the Monte Carlo analysis presented in this stage is activity-based. Duration uncertainty was simulated at the activity level, but the sampled durations were not propagated through the complete CPM network during each simulation iteration. As a result, the current P80 results should not be interpreted as project or package completion-date distributions.

Fourth, the probability distributions used in the exploratory simulation require further calibration. Triangular distributions provide a practical starting point where detailed historical data are limited, but final probabilistic analysis should be supported by project-specific historical data, expert elicitation, or a documented calibration procedure.

Finally, the study was limited to nine historical validation cases from a single project, including five numerical outcomes, three categorical outcomes, and one pending milestone. The small sample and differences in outcome types limit the ability to evaluate predictive accuracy. In addition, the LLM analysis was conducted through the ChatGPT interface, where some model settings were unavailable, and the risk-feature classifications were based partly on provisional screening rules. These limitations affect reproducibility and should be considered when interpreting the findings. The results are therefore presented as an exploratory case-study assessment rather than evidence of a validated predictive model.

5.6. Practical and Research Significance

Despite these limitations, the study demonstrates a practical way to position AI within established project-controls practice. The framework does not require replacing CPM, existing scheduling software, or the scheduler. Instead, it adds a reasoning layer that can help identify where additional planning attention may be justified.

The main contribution of this study is a structured approach to reviewing schedule risk using information already available during baseline planning. By combining CPM indicators, contextual risk features, LLM-assisted interpretation,

and exploratory probabilistic analysis, the framework brings different sources of schedule information into a common review process. The retrospective comparison with historical project outcomes provides an initial assessment of this approach, although further testing is needed to establish its effectiveness in identifying schedule risks before construction begins. For airfield construction, where the consequences of poor sequencing or missed coordination can extend beyond construction performance into airport operations, this type of early risk screening may be particularly valuable [20].

The methodological contribution of this study lies in bringing conventional CPM indicators, contextual risk features, LLM-assisted interpretation, and exploratory probabilistic analysis into a single baseline risk-review process. Rather than introducing a new delay-prediction model, the framework uses the LLM to organize existing schedule information into traceable observations and practical recommendations for planner review. The retrospective comparison with actual project outcomes provides an initial examination of how this approach can be applied to identify localized schedule vulnerabilities. Although the study does not establish that LLM-assisted analysis performs better than conventional methods, it demonstrates a structured way to incorporate language-model interpretation into established scheduling practice without replacing CPM calculations or professional judgment.

6. Conclusions

This study developed an LLM-augmented framework for risk-informed schedule planning in airfield construction by combining CPM information, baseline risk features, the Baseline Risk Index, LLM-assisted interpretation, probabilistic simulation, and historical validation. The purpose was not to use an LLM as a replacement for the scheduler or for established scheduling software. Instead, the language model was positioned as an additional reasoning layer that can help planners interpret combinations of schedule characteristics and recognize areas that may deserve closer review before construction progresses.

The case study showed that schedule vulnerability can exist below the level of the final contractual milestone. Although substantial completion was achieved on its baseline target date, several intermediate work packages experienced later openings, revised sequencing, deferred mobilization, access or interface constraints, and scope changes. Baseline indicators also varied considerably among these packages. In particular, concentrations of low- and zero-float activities provided useful evidence of limited schedule flexibility, while contextual features such as restricted access and third-party interfaces helped explain risks that could not be described by CPM measures alone.

The comparison between baseline-derived risk signals and LLM-assisted interpretations, exploratory Monte Carlo results, and historical outcomes also showed that no single indicator was sufficient to explain schedule performance. A higher BRI did not translate directly into a specific number of delay days, and activity-

level probabilistic exposure did not necessarily correspond to the magnitude of later package variance. These findings support the use of the framework as a decision-support method rather than a deterministic delay-prediction model. Its practical value lies in directing attention toward vulnerable portions of the schedule so that planners can examine mitigation, alternative sequencing, access coordination, or additional contingency before those risks develop into field problems.

The present probabilistic analysis should be viewed as an intermediate stage of the research. The Monte Carlo simulation was performed at the activity level using provisional duration and risk-event assumptions. A stronger next step is to calibrate these assumptions using historical project data and expert elicitation and then propagate uncertainty through the complete CPM network. This would allow the analysis to move from relative activity-level exposure toward defensible distributions for work-package and project milestone completion.

Future research should also test the framework on additional airfield and infrastructure projects. A larger sample would allow the baseline risk features and BRI formulation to be calibrated more rigorously [21]. Another important extension is to evaluate the planning-enhancement stage directly by comparing an original baseline with an AI-assisted alternative schedule under the same project constraints. The alternative could include changes in sequencing, float protection, access planning, or interface management and then be evaluated through the same probabilistic model. Such an experiment would provide stronger evidence of whether LLM-assisted reasoning can improve schedule resilience rather than simply identify areas of risk.

Overall, this study shows how an LLM can be incorporated into an existing CPM-based scheduling process to support the review of baseline risk conditions. The LLM helped organize schedule information and identify issues for further examination, although its added value over conventional CPM and BRI analysis was not independently demonstrated. The findings provide a basis for further development of the framework, particularly through calibrated probabilistic simulation, network-level schedule analysis, and evaluation across additional construction projects.

Author Contributions

Hossein Taheri: Conceptualization, methodology, data collection, data analysis, investigation, visualization, and original draft writing. Mana Samadi: Supported methodology, interpreted data, validated results, and wrote, reviewed, and edited the manuscript. Both authors reviewed and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Taheri, H. (2025) A KPI-Driven Dashboard Framework for Enhanced Safety Man-

- agement in Airfield Construction. <https://doi.org/10.31224/5432>
- [2] Taheri, H.J., Rousta, S. and Zare, S. (2026) A Hybrid Framework for Airport Safety Oversight: Integrating FAA Part 139 and ICAO SMS. *Journal of Transportation Technologies*, **16**, 342-360. <https://doi.org/10.4236/jtts.2026.163019>
 - [3] Umoru, C. (2025) AI and Construction Project Schedules Efficiency: A Review. *International Journal of Scientific Research and Modern Technology*, **4**, 89-94. <https://doi.org/10.38124/ijsrmt.v4i6.761>
 - [4] Prieto, S.A., Mengiste, E.T. and García de Soto, B. (2023) Investigating the Use of ChatGPT for the Scheduling of Construction Projects. *Buildings*, **13**, Article 857. <https://doi.org/10.3390/buildings13040857>
 - [5] Naser, M.Z. (2025) A Step-by-Step Tutorial on Machine Learning for Engineers Unfamiliar with Programming. *AI in Civil Engineering*, **4**, Article No. 10. <https://doi.org/10.1007/s43503-025-00053-x>
 - [6] Jung, Y. and Golparvar-Fard, M. (2026) Expertplanner: A Mixture-of-Experts Transformer Language Model for Decomposing Construction Look-Ahead Plan Tasks from Long-Term Master Schedules. *Advanced Engineering Informatics*, **71**, Article 104439. <https://doi.org/10.1016/j.aei.2026.104439>
 - [7] Chong, H., Yang, X., Goh, C.S. and Luo, Y. (2025) BIM and AI Integration for Dynamic Schedule Management: A Practical Framework and Case Study. *Buildings*, **15**, Article 2451. <https://doi.org/10.3390/buildings15142451>
 - [8] Donia, M., Elbeltagi, E., Elhakeem, A. and Wefki, H. (2026) An Integrated BIM-NLP Framework for Design-Informed Automated Construction Schedule Generation. *Designs*, **10**, Article 43. <https://doi.org/10.3390/designs10020043>
 - [9] Kim, J., Park, S., Lim, S., Lee, H. and Mun, D. (2026) LLM-Based Design Document Understanding for Semantic Knowledge Extraction and Design Verification. *Journal of Computational Design and Engineering*, **13**, 29-51. <https://doi.org/10.1093/jcde/qwag074>
 - [10] Mandal, S., Shiuly, A., Sau, D., Mondal, A.K. and Sarkar, K. (2024) Study on the Use of Different Machine Learning Techniques for Prediction of Concrete Properties from Their Mixture Proportions with Their Deterministic and Robust Optimisation. *AI in Civil Engineering*, **3**, Article No. 7. <https://doi.org/10.1007/s43503-024-00024-8>
 - [11] Amer, F., Jung, Y. and Golparvar-Fard, M. (2023) Construction Schedule Augmentation with Implicit Dependency Constraints and Automated Generation of Lookahead Plan Revisions. *Automation in Construction*, **152**, Article 104896. <https://doi.org/10.1016/j.autcon.2023.104896>
 - [12] Seo, H. and Choi, M. (2026) A Large Language Model-Based Agent System for Interactive Large-Scale Project Scheduling. *Journal of Computational Design and Engineering*, **13**, 183-203. <https://doi.org/10.1093/jcde/qwag082>
 - [13] Abdelsalam, M., Ashmawi, A. and Nguyen, P.H.D. (2026) AI-Driven Automation of Construction Cost Estimation: Integrating BIM with Large Language Models. *Buildings*, **16**, Article 485. <https://doi.org/10.3390/buildings16030485>
 - [14] Jafari, P., Al Hattab, M., Mohamed, E. and AbouRizk, S. (2021) Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation. *Applied Sciences*, **11**, Article 6188. <https://doi.org/10.3390/app11136188>
 - [15] Biswas, P., Rashid, A., Biswas, A., Nasim, M.A.A., Chakraborty, S., Gupta, K.D., *et al.* (2024) AI-Driven Approaches for Optimizing Power Consumption: A Comprehensive

- sive Survey. *Discover Artificial Intelligence*, **4**, Article No. 116. <https://doi.org/10.1007/s44163-024-00211-7>
- [16] Cao, L., Huo, T., Li, S., Zhang, X., Chen, Y., Lin, G., *et al.* (2024) Cost Optimization in Edge Computing: A Survey. *Artificial Intelligence Review*, **57**, Article No. 312. <https://doi.org/10.1007/s10462-024-10947-4>
- [17] Aritra Pal, A.P., Danny Murguia, D.M. and Campbell Middleton, C.M. (2025) AI-Powered Application for Construction Schedule Management Using Natural Language. *Computing in Construction*, Porto, 14-17 July 2025, 1-8. <https://doi.org/10.35490/ec3.2025.407>
- [18] Doneva, S.E., Qin, S., Sick, B., Ellendorff, T., Goldman, J., Schneider, G., *et al.* (2024) Large Language Models to Process, Analyze, and Synthesize Biomedical Texts: A Scoping Review. *Discover Artificial Intelligence*, **4**, Article No. 107. <https://doi.org/10.1007/s44163-024-00197-2>
- [19] Memiş, M.A., Keskin, I., Demir, S. and Memiş, Ş.U. (2025) Machine Learning-Based Prediction of Permafrost Degradation and Its Implications on Geotechnical Infrastructure: A Comprehensive Review. *AI in Civil Engineering*, **4**, Article No. 28. <https://doi.org/10.1007/s43503-025-00080-8>
- [20] Mirzamohammadi, M., Haghighatjoo, M., Taheri, H.J. and Varzeghani, N.M. (2026) Physics-Informed and Transferable Explainable AI Framework for Reliable Prediction of Tunnel-Induced Settlement in Coastal Environments. *Discover Artificial Intelligence*, **6**, Article No. 558. <https://doi.org/10.1007/s44163-026-01108-3>
- [21] Tashkinov, A.G. (2024) The Implementation of Lean and Digital Management Techniques Using Artificial Intelligence in Industrial Settings. *Discover Artificial Intelligence*, **4**, Article No. 94. <https://doi.org/10.1007/s44163-024-00186-5>