

AI Large Model-Based Data Value Mining in Power Projects

Zhang Jialin, Zhang Siwen

GD Power Development Co., Ltd., Beijing, China

Received: June 15, 2026; Revised: August 10, 2026; Accepted: September 10, 2026; Published: September 25, 2026

Abstract

Power project usually has the characteristics of long construction period, many participants, complex supply chain and large number of project documents. Information such as schedule, earned value, cost, contract changes, quality, safety, equipment delivery, BIM / IoT, and meeting records are usually stored in separate systems. Therefore, although the data are available, they may not provide timely decision support for project management. Traditional statistical models are not good at interpreting narrative evidence. When the project conditions or supply conditions change, the prediction accuracy of the static model may also decrease. In order to solve these problems, this paper proposes Power-LLM ValueLoop, a closed-loop framework for data value mining supported by large models. The framework identifies entities, events, causal signals, urgency, conflict evidence, and confidence levels through a pattern-based large model layer, and associates relevant results with structured data such as schedule, cost, resources, quality, and supply chain. Subsequently, the semantic-time model is used to estimate and calibrate the probability of major delays or failure events in the next 4 weeks. Confirmed outcomes, human corrections, time-decayed learning, drift monitoring, and threshold adjustment form the feedback loop. The system can deliver traceable risk early warnings, identify key risk drivers, and provide action recommendations for review by project personnel. This paper uses a synthetic project portfolio for evaluation, including 720 power engineering projects, 10 start-up batches, and 15,653 project-week observation records. Controlled concept drift was introduced from the 6th batch. In the four batches after drift, the macro average ROC-AUC of the feedback model was 0.773, the macro average PR-AUC was 0.366, and the Brier score was 0.105. At the recall-oriented operating point, the model identified 78.1 % of the risks in the next four weeks, with an average warning advance of about 9.1 weeks. For 80.2 % of the event items, the system can issue an early warning at least 4 weeks before the event. Compared with the static AI fusion model, the feedback mechanism improves the recall rate by about 0.140, but the false alarm rate also increases from 0.275 to 0.391. This shows that the improvement of model performance still needs to rely on reasonable threshold management and manual review. A supplementary Monte Carlo decision analysis examines prevalence and error-cost sensitivity at fixed operating rates. At the reported test prevalence, feedback has lower expected error loss only when the missed-event cost exceeds approximately 5.35 times the false-alarm cost. This conditional result does not establish savings in deployed projects.

Keywords: large model; power project; data value mining; risk early warning; feedback loop; decision support.

1 Introduction

New energy base, transmission and substation engineering, distribution network upgrading and major technological transformation projects usually run through multiple stages such as feasibility study, design, procurement, construction, commissioning and handover. Relevant data includes not only structured plan information, earned value indicators, cost ledgers, and equipment lists, but also change notifications, contract

Research Briefs on Information & Communication Technology Evolution (ReBICTE), Vol. 12, Article No. 09 (September 25, 2026)
DOI: <https://doi.org/10.64799/rebict.e.v12.9>

documents, supervision logs, meeting minutes, quality and safety work orders, BIM models, and IoT progress data streams.

The Transformer architecture and pre-trained language model provide the basis for long text representation and cross-task migration [2-3]. The retrieval enhancement generation technology can also make the model output based on auditable external evidence [4]. Machine learning has been applied to project schedule performance evaluation and delay prediction [5-7]. Random forests and boosting models can flexibly describe nonlinear relationships [8-9]. However, most studies have completed the cleansing and coding of the default input data, and have not yet fully explained how narrative evidence corresponds to engineering entities, how the model score is translated into a risk probability that can support decision-making, and how the model should be adjusted when equipment supply, regulatory requirements, or organizational processes change. Document question answering robots can improve the efficiency of information retrieval, but they may not be able to transform scattered evidence into stable data assets and quantifiable decision value.

Recent construction research provides a more specific basis for semantic integration. Domain-specific pre-training on construction management corpora improves the representation of engineering language [16]. Saka et al. discuss lifecycle applications and implementation limits of GPT models [17]. Knowledge-augmented contract risk identification [18] and knowledge graph-enhanced contract review [19] demonstrate how explicit domain evidence can support document interpretation. A 2026 review examines the remaining gap between research prototypes and practical contract-management requirements [20]. These studies support evidence extraction as a useful component, but their task settings do not directly validate four-week risk forecasting for power projects.

In power systems, large-model applications cover perception, planning and operational assistance, while reliability, computing requirements and domain adaptation remain practical constraints [21]. Evidence-grounded generation [22], adaptive retrieval and self-critique [23], and retrieval-augmented generation evaluation [24] provide complementary tools for traceability. Long-context performance can depend on where relevant evidence appears [25], so simply concatenating all project records is insufficient. Online conformal prediction offers a separate approach to uncertainty under changing distributions [26], although its guarantees do not transfer automatically to the probability calibration used here.

The 2026 literature also broadens the comparison beyond text-only assistants. Erfani and Mansouri synthesize multimodal applications across construction workflows [27]. Liu and Luo combine textual and numerical information to predict construction-task duration with conventional machine learning [28]. This evidence motivates a comparison based on data, task and validation design rather than model size alone. Table 1 locates the present framework within these research directions.

Table 1. Relationship between recent research and the present framework

Research direction	Representative evidence	Validated task or emphasis	Remaining requirement for this study
Domain language and task prediction	[16], [17], [28]	Engineering text representation and task-duration prediction	Align narrative signals with project-week outcomes
Contract knowledge integration	[18]-[20]	Contract interpretation and risk review	Test transfer to event forecasting and delayed labels
Power-system large models	[21]	Applications, enhancement methods and reliability constraints	Validate power-project records in a prospective study
Evidence and uncertainty evaluation	[22]-[26]	Citations, retrieval, context use and online uncertainty	Audit extraction quality separately from prediction quality

Multimodal construction applications	[27]	Tasks, model adaptation and evaluation practices	Measure value added by BIM, images and IoT evidence
Present framework	Sections 3-5	Synthetic semantic fusion and feedback analysis	External data and real extractor evaluation are still required

This paper defines data value from four types of reusable outputs, including unified project entity-event-time evidence, semantic variables and data quality variables suitable for time series learning, calibrated risk probability and its drivers and action candidates, and result feedback that can improve data standards and model parameters. Power-LLM ValueLoop transforms this definition into an enforceable closed-loop framework. The large model does not replace the project manager, nor will it directly judge the success or failure of the project without review. Its main role is to carry out evidence-based semantic integration. The lightweight prediction head combines structured project status with semantic evidence. The calibration layer converts the model score into probability, and the lag label and manual correction are used to continuously update the feedback loop. The main contributions of this paper include: constructing the data model of the whole life cycle of power engineering, proposing the semantic fusion method of confidence perception and the drift perception feedback mechanism, introducing the threshold governance strategy, and designing controlled simulation experiments and a fully specified supplementary decision analysis.

2 Data Value Mining Problem in Power Projects

2.1 Heterogeneous evidence and latent value

Structural indicators can reflect the degree of deviation, but it is often difficult to explain the causes of deviation. The equipment ledger may show that the delivery date has not changed, but the supplier transaction information has reflected the production conflict. The critical path schedule is still ostensibly controllable, but multiple meeting minutes have recorded unresolved interface freezing issues. The number of quality work orders for some projects may be small, but their text descriptions may reveal the same root cause in multiple work packages. The value of the large model is to identify such weak signals. These signals may span different documents, time nodes, and subjects of responsibility. After extraction and collation, they can be converted into intermediate representations that can be queried, reviewed, and learned.

2.2 Objective and formalization

Let $x(i, t)$ denote the structured state of project i in week t , $D(i, t)$ the associated documents and text fragments, and $q(i, t)$ a data-quality vector. A large-model extractor $G(\psi)$ produces semantic variables $z(i, t)$, event relations $e(i, t)$, and referenced sources $s(i, t)$ under a predefined schema. A prediction head $f(\theta)$ estimates the probability that a material delay or failure event will occur within $H=4$ weeks. The system should have good discrimination ability, be able to output calibrated probability and traceable evidence, and adapt to data drift between different batches, while controlling the workload of manual review within a manageable range.

$$(z(i, t), e(i, t), s(i, t)) = G_{\psi}[D(i, t) \mid \text{schema, retrieval, provenance}] \quad (1)$$

$$p(i, t) = \mathbb{P}(y(i, t: t + H) = 1 \mid x(i, t), z(i, t), q(i, t)) = \text{sigmoid}(f_{\theta}[x, z, q]) \quad (2)$$

In (1)-(2), i indexes projects and t indexes weekly observations; $D(i, t)$ is the associated document set; G_{ψ} is the schema-constrained extractor with parameters ψ ; $z(i, t)$, $e(i, t)$, and $s(i, t)$ are the extracted semantic variables, event relations, and source references; and schema, retrieval, and provenance are the constraints and evidence supplied to the extractor. The vector $x(i, t)$ contains structured project status, $q(i, t)$ contains data-quality variables, $y(i, t: t + H)$ indicates whether a material event occurs from week t through $t + H$, H is the forecast horizon (four weeks here), $P(\cdot)$ denotes conditional probability, f_{θ} is the prediction head with

parameters θ , sigmoid ($\cdot$) maps its score to the interval from zero to one, and $p(i, t)$ is the predicted event probability. In $f_\theta[x, z, q]$, x , z , and q are shorthand for the same project-week inputs.

Risk identification and treatment should be tied to an explicit event definition and accountable review process [1]. For this manuscript, a positive weekly label concerns the next four weeks. It differs from the project-level indicator of whether an event ever occurs during observation. Forecasting accuracy, project-level warning coverage and intervention effectiveness therefore require different denominators and should be reported separately.

Risk is a representative verification task selected in this paper, which does not cover the full scope of data value mining. The same semantic layer can also support change impact traceability, equipment supply anomaly attribution, quality problem clustering, data integrity assessment and project experience reuse. Risky tasks are selected because they have a time-stamped result label, a quantifiable amount of advance warning, and clear false-negative and false-positive costs.

3 Power-LLM ValueLoop

3.1 Project data space and evidence governance

Figure 1 focuses on the project digital twin and organizes the evidence data of the whole life cycle of the project. Connector access progress and earned value records, design and contract changes, construction logs, meeting minutes, e-mails, equipment and supplier events, quality and safety issues, and BIM / IoT progress data streams. Each record is associated with a specific project, project phase, work breakdown structure or device object, responsible subject, event time and version information. From the perspective of the model, the original evidence cannot be modified. The fields generated by the model are stored in the form of derived data, with source identification, access rights, timestamp, and integrity metadata. Therefore, the early warning results can be traced back to the original source. Missing, conflicting, or old data can also become data governance signals, and will not be regarded as undetectable pre-processing noise.

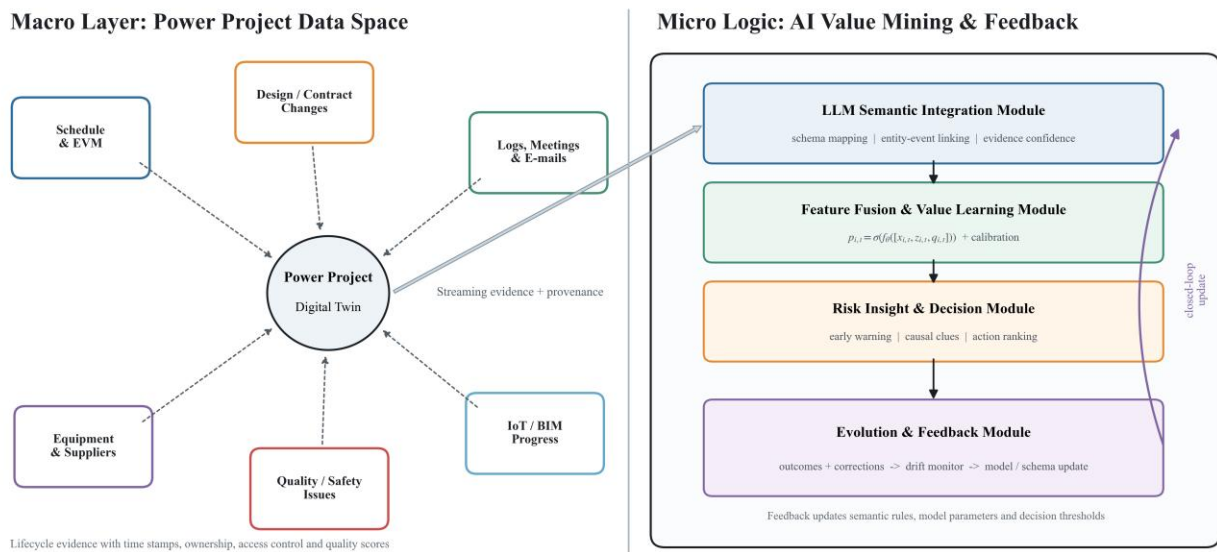


Figure 1. Macro data space and micro feedback loop for power-project data value mining

3.2 Schema-constrained semantic integration

For each item-week time window, the system retrieves relevant evidence based on entity and time, and requires

the large model to return the result according to the fixed JSON mode. The core fields include risk-event, affected -entity, cause, effect, urgency, contradiction, confidence, evidence-ids. The system verifies the output results, including data type, numerical range, entity legitimacy, time sequence and evidence reference. Invalid output will be re-generated or submitted for manual review. The confidence is extracted as both the input variable of the model and the weight of the semantic feature. The system also explicitly retains cross-source contradictory information. For example, the progress plan appears to be normal on the surface, but the continuous reporting interface of multiple sources is blocked, and such signals will not be ignored.

$$\tilde{z}(i, t) = c(i, t)z(i, t) + [1 - c(i, t)]\mathbb{E}[z \mid \text{project type, phase}] \quad (3)$$

In (3), $\tilde{z}(i, t)$ is the confidence-adjusted semantic vector; $c(i, t)$ is extractor confidence on the interval from zero to one; $z(i, t)$ is the raw semantic vector; $\mathbb{E}[z \mid \text{project type, phase}]$ is the conditional mean semantic vector for comparable projects; and $1 - c(i, t)$ is the shrinkage weight assigned to that baseline.

A production extractor should be evaluated on an independently annotated document set. The evaluation should separately report entity/event extraction precision and recall, evidence-reference validity, unsupported-claim frequency and confidence calibration. Citation evaluation [22] and RAG evaluation [24] inform this design, but an automatically generated citation is not sufficient evidence that a claim is correct. Schema validation should also reject future-dated records. These extractor-level tests are outside the current proxy-based simulation.

3.3 Semantic-temporal fusion and calibration

The fusion layer combines schedule deviation, cost deviation, resource shortage, change intensity, dependency delay, defect pressure, project complexity and delivery status with semantic urgency, cause dispersion, narrative contradiction, extraction confidence and data missing degree. The simulation experiment uses a lightweight logistic risk prediction head to clearly and reproducibly evaluate the contribution of semantic integration. For larger real data sets, a boosting model or deep time series model can be further adopted. Platt scaling converts the model score into a calibrated probability [11]. The importance of global permutations and project-level directional contributions together support four types of outputs : risk probability, main drivers, evidence links, and candidate action recommendations.

$$\hat{p}(i, t) = \text{sigmoid}(af_{\theta}(x(i, t), \tilde{z}(i, t), q(i, t)) + b) \quad (4)$$

In (4), $\hat{p}(i, t)$ is the calibrated risk probability; $f_{\theta}(\cdot)$ is the uncalibrated score computed from structured state $x(i, t)$, confidence-adjusted semantic vector $\tilde{z}(i, t)$, and data-quality vector $q(i, t)$; a and b are the Platt-scaling slope and intercept learned on calibration data; and $\text{sigmoid}(\cdot)$ converts the affine score to a probability.

Gradient boosting [9,15] is a possible predictive extension, while Shapley-based explanations [10] are an alternative to the permutation analysis shown in Figure 3. Neither option is an additional implemented baseline in this study. Feature importance indicates predictive dependence and does not identify the causal effect of a managerial action.

3.4 Outcome feedback, drift monitoring, and threshold governance

Confirmed events, phase reviews, and human corrections are recorded separately as `outcome_label`, `action_taken`, and `action_effect` so that intervention is not mislearned as a natural outcome. When cohort k closes, historical samples receive time-decay weights $\alpha_j = \exp[-\lambda(t_k - t_j)]$, and recent cohorts are used for retraining and recalibration. Cohort-level AUC, Brier score, and feature distributions are monitored for control-limit violations. An alert threshold is chosen on calibration data to satisfy a minimum recall requirement and then favor a high F1 score. Critical equipment and major milestones may use a lower threshold, while ordinary work packages can use a higher threshold to control review load.

$$\theta_k = \arg \min_{\theta} \sum_j \alpha_j \text{WBCE}(y_j, f_{\theta}(h_j)) + \lambda_2 \|\theta\|_2^2 \quad (5)$$

In (5), θ_k is the parameter vector fitted when cohort k closes, and θ is a candidate parameter vector; j indexes historical training samples; α_j is the time-decay weight for sample j , defined as $\exp[-\lambda(t_k - t_j)]$ is the closing time of cohort k , t_j is the sample time, and λ is the decay rate; y_j is the binary outcome label; h_j is the fused feature vector; $f_{\theta}(h_j)$ is the predicted score or probability; WBCE is weighted binary cross-entropy; λ_2 is the L2-regularization coefficient; the squared L2 norm of θ is the penalty magnitude; and $\arg \min$ returns the value of θ that minimizes the objective.

Feedback closed-loop plays a role in three levels. First, the confirmed results are used to update the prediction head. Secondly, manual correction is used to update pattern rules and a small number of examples. Thirdly, the recurring data missing or information contradiction will be fed back to the data responsible party as a process quality problem. Therefore, the platform can not only identify risks, but also point out which fields need to be further improved, which data sources are in conflict, and which evidence can provide valuable early warning signals.

Feedback must use outcomes only after their four-week labels have matured. For a deployment record at week t , observations whose outcome window has not closed must remain excluded from retraining and threshold selection. Adaptive-window change detection [12] could complement cohort monitoring, and online uncertainty methods [26] provide further comparators. These are proposed extensions; the current reported experiment uses cohort updates rather than an implemented ADWIN or conformal procedure.

4 Simulation Design and Reproduction

4.1 Synthetic power-project portfolio

Due to the sensitivity of enterprise project records, this paper uses a synthetic high-risk project portfolio for evaluation. The data generator constructs 720 projects and divides them into 10 start-up batches. Each project was observed for up to 28 weeks, forming a total of 15,653 project-week records. Static project attributes include project complexity, equipment supply dependence, governance maturity, team maturity and potential synergy ambiguity. Time-varying variables include schedule deviation, cost deviation, resource shortage, quality pressure, change intensity, dependency delay and progress rate decline. The semantic channel generates urgency, cause dispersion, contradiction and confidence based on potential risks. When the degree of data loss increases, the noise in the semantic variable will increase accordingly. Once the event occurs, the subsequent weekly records of the project will no longer be generated. The incidence of events at the project level was 58.5 %, and the positive labels of events in the next four weeks accounted for 10.8 % of the panel data. These ratios are used to construct stress test scenarios that do not correspond to any specific power company.

Batches 0 to 4 are used for initial training, batch 5 is used for probabilistic calibration, and batches 6 to 9 constitute the test set after drift. Starting from Batch 6, the data generator reduces the relative impact of traditional schedule and cost signals, while increasing the impact of factors such as resource shortages, design changes, equipment supply, and collaborative ambiguity. The feedback model uses a time attenuation coefficient of 0.23 and a rolling calibration window covering the last three batches. Fitting data and calibration data are isolated from each other at the project level. The main evaluation indicators include batch macro average ROC-AUC, batch macro average PR-AUC, summary F1 value, recall rate, Brier score, false alarm rate, early warning load, average early warning advance, and the proportion of events that received early warning at least 4 weeks before the event. The semantic _ adapter function is a controllable off-line proxy for the output of a large structured model without calling the production environment API. The experiment aims to test whether semantic variables with preset noise can produce downstream application value after combining time learning and feedback mechanisms. This experiment is not a performance benchmark or ranking for a specific large language model.

The 10.8% positive rate above describes the complete panel. The test confusion matrix in Figure 3(c) instead contains 768 positives and 4,965 negatives, or 5,733 test observations, giving a post-drift prevalence of 13.40%. These denominators explain why the PR reference line in Figure 2 is 0.134. PR-based metrics are particularly relevant when positives are uncommon [13], while the Brier score measures squared probability error [14]. The supplied Figure 2 labels the PR summaries as macro average precision (AP). That convention should be retained when the original evaluation code is released.

4.2 Comparative methods and implementation

The four methods use the same project division method and random seed 20260822. Structured logistic regression and class-balanced random forests use only structured variables. The static AI fusion model adds all semantic variables and data quality variables, but it is not updated after the initial calibration is completed. The AI fusion model with feedback uses the same feature space and further performs time-attenuated batch update and rolling recalibration. ROC-AUC and PR-AUC were calculated in each post-drift batch, respectively, and then macro averaged. This can avoid the false impact of calibration mapping changes on cross-batch sorting. Threshold-dependent indicators are aggregated and calculated on all test records. The early warning advance is defined as the time interval between the first time the model exceeds the early warning threshold and the event occurrence week. For projects that have not received early warning, the early warning advance is recorded as 0.

Table 2. Main experiment settings and supplementary simulation assumptions

Item	Original experiment as reported	Supplementary decision analysis
Observation unit	Project-week, clustered within 720 projects	Independent simulated project-week decisions
Sample size	15,653 total; 5,733 test observations	5,733 observations per replicate
Temporal split	Training 0-4; calibration 5; test 6-9	No fitting or temporal model is simulated
Forecast horizon and drift	4 weeks; drift starts at cohort 6	Class-conditional operating rates held fixed
Positive prevalence	10.8% overall; 768/5,733 in test set	0.05-0.25 in 0.01 steps, plus 768/5,733
Feedback and calibration	Decay 0.23; latest 3 cohorts for calibration	Static TPR/FPR 0.641/0.275; feedback 600/768 and 1,942/4,965
Randomness	Seed 20260822 in the supplied experiment	Seed 20260914; 2,000 replicates per prevalence
Relative error costs	No monetary cost model supplied	Missed-event/false-alarm ratio 1-12, step 0.1
Uncertainty interpretation	No original prediction-level intervals supplied	Central 95% Monte Carlo intervals conditional on fixed rates

4.3 Supplementary Monte Carlo decision analysis

The supplementary experiment evaluates operational consequences of the reported alert policies. It does not rerun the semantic generator, train a large language model, or reconstruct the original project trajectories. The original generator and observation-level prediction files were not included with the manuscript. We therefore use the rounded static-fusion recall and false-positive rate, 0.641 and 0.275, together with the exact feedback rates obtained from Figure 3(c), 600/768 and 1,942/4,965. Holding these rates fixed isolates changes in event prevalence and error costs. It does not represent a new concept-drift adaptation experiment.

For each prevalence, 2,000 replicates are generated with 5,733 independent observations per replicate. The number of positive observations is binomial with prevalence π . For each method, true positives are drawn conditionally with probability r and false positives with probability f . Here r is recall and f is the false-positive rate. Both methods use the same positive count within a replicate, but their alert draws are conditionally independent. The random seed is 20260914. Shaded regions in Figures 5 and 6 are the 2.5th-97.5th percentiles

across replicates. They describe finite-sample variability under this artificial decision model and are not confidence intervals for the original algorithms.

$$A(\pi) = \pi r + (1 - \pi)f, \quad PPV(\pi) = \pi r / A(\pi) \quad (6)$$

$$L(\pi, \rho) = 1000[\rho\pi(1 - r) + (1 - \pi)f] \quad (7)$$

In (6), A is the fraction of observations generating alerts and PPV is precision. In (7), ρ is the cost of one missed positive observation divided by the cost of one false alarm, and L is expected error loss per 1,000 observations in false-alarm cost units. Correct decisions have zero loss in this simplified model. The model excludes the cost of reviewing true alerts, delayed intervention, repeated warnings and computation. Appendix A provides the complete numerical reproduction core.

5 Results and Value Analysis

5.1 Predictive performance and probability reliability

Table 3 consolidates the reported post-drift results. The feedback model has the highest reported macro ROC-AUC (0.773) and macro PR summary (0.366), and the lowest Brier score (0.105). Static fusion has the highest F1 score (0.375) and the lowest false-positive rate (0.275). Feedback raises recall to 0.781 while increasing the false-positive rate to 0.391. These differences describe the operating points selected for the supplied experiment; they do not establish statistical superiority or a benefit under an equal alert budget.

Table 3. Reported performance on post-drift test cohorts

Method	ROC-AUC	PR/AP	F1	Recall	Brier	FPR	Lead (wk)
Structured logistic	0.757	0.346	0.364	0.659	0.107	0.304	7.56
Structured random forest	0.724	0.326	0.333	0.622	0.106	0.327	8.75
Static fusion	0.767	0.360	0.375	0.641	0.106	0.275	7.15
Fusion with feedback	0.773	0.366	0.363	0.781	0.105	0.391	9.1

ROC-AUC and PR/AP are macro summaries across cohorts 6-9. Threshold metrics and Brier scores use pooled test records as reported in the source manuscript. FPR is $FP/(FP+TN)$, not the fraction of alerts that are false. Warning lead is a project-level measure. The feedback lead is shown to one decimal place because the source plot and text differ in the second decimal. No significance test or original-data confidence interval is inferred from this table.

Compared with structured logistic regression, the feedback mechanism increased ROC-AUC by 0.016 and PR-AUC by 0.020. Compared with the static AI fusion model, the two area-based evaluation indicators are increased by 0.006. The static model achieves a slightly higher F1 value, indicating that the trade-off between precision and recall is more balanced. After the drift occurs, the feedback model intentionally adopts a more sensitive warning threshold. Therefore, the decision-making value of the model should not be evaluated only by a certain threshold-dependent indicator.

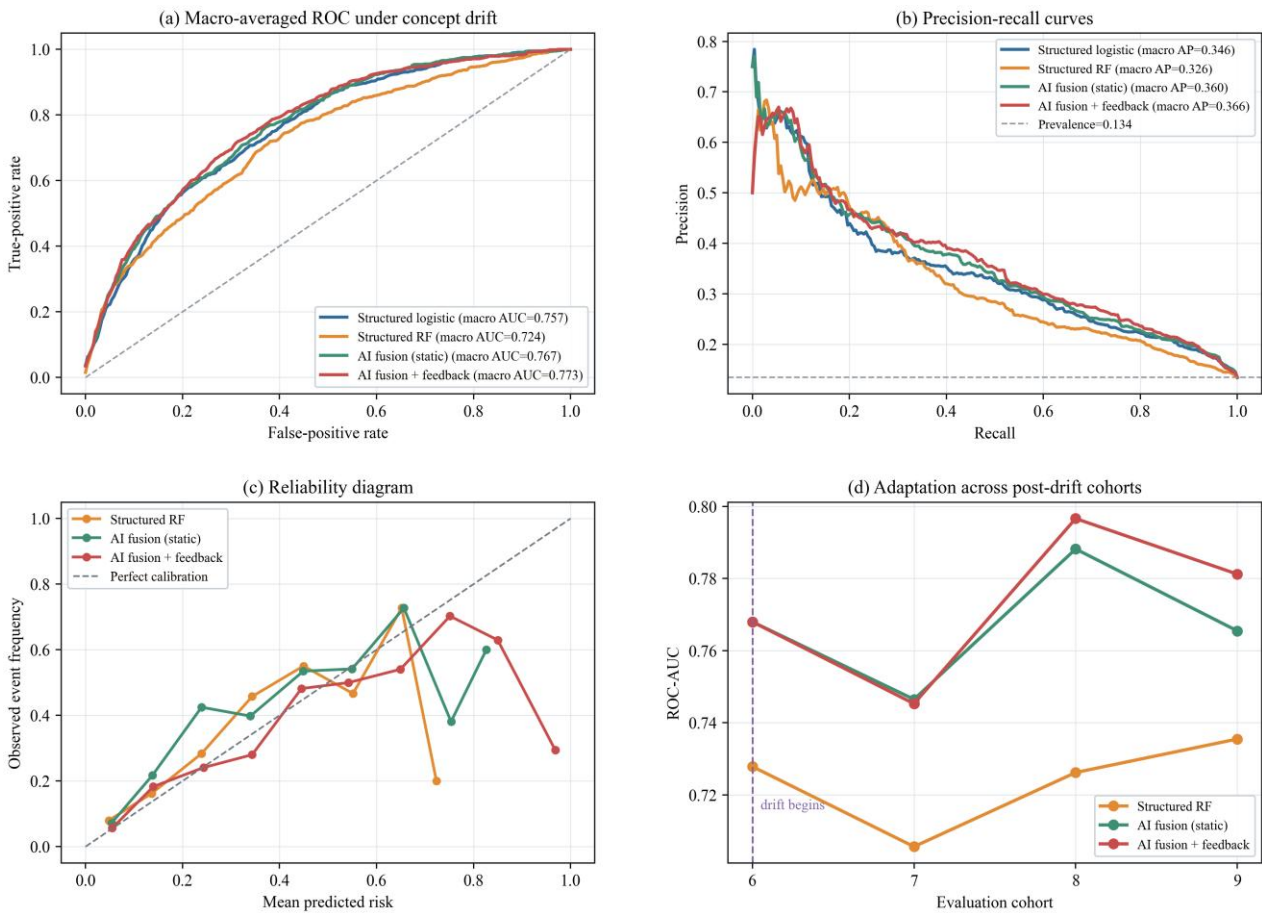


Figure 2. Post-drift ROC/PR performance, probability calibration, and feedback adaptation

Batch 6 is the first test of the feedback model, so its initial performance is consistent with the static model. When the confirmed drift label enters the historical sample pool, the ROC-AUC of the feedback model in batch 8 and batch 9 exceeds the static model. The summary Brier score of the feedback model is 0.105, which is lower than 0.106 of the static fusion model and 0.107 of the structured logistic regression model. Since the positive class records only account for about one-tenth of the panel data, there are still some fluctuations in the reliability of the highest prediction probability box. The dashboard in the production environment should show the number of samples, confidence intervals, and calibration errors for each bin by project type and project stage.

5.2 Warning lead and decision value

The feedback model has a four-week record-level recall of 78.1%. Separately, 80.2% of projects with an observed event were reported to receive at least one warning four or more weeks before that event, and mean project-level warning lead was about 9.1 weeks. The latter quantities measure alert coverage and timing across a project history. A warning more than four weeks before an event can be a false positive under the four-week label, so a long warning lead does not establish accurate prediction at that longer horizon. Candidate responses include reviewing supplier alternatives, resource allocation and critical-path assumptions. Their effects on project outcomes were not evaluated.

The higher alert sensitivity also increases the review burden. The false alarm rate of the feedback model is 0.391, which is higher than 0.275 of the static fusion model, and its early warning rate is 0.443. Therefore, decision support systems should not only provide red, yellow, and green labels, but also show calibrated risk probabilities, main drivers, evidence citations, and reversible action recommendations. For key long-period equipment and important milestones, a lower warning threshold can be used and a mandatory secondary review

can be set. For general tasks, a higher threshold can be used to control the workload of manual disposal and review.

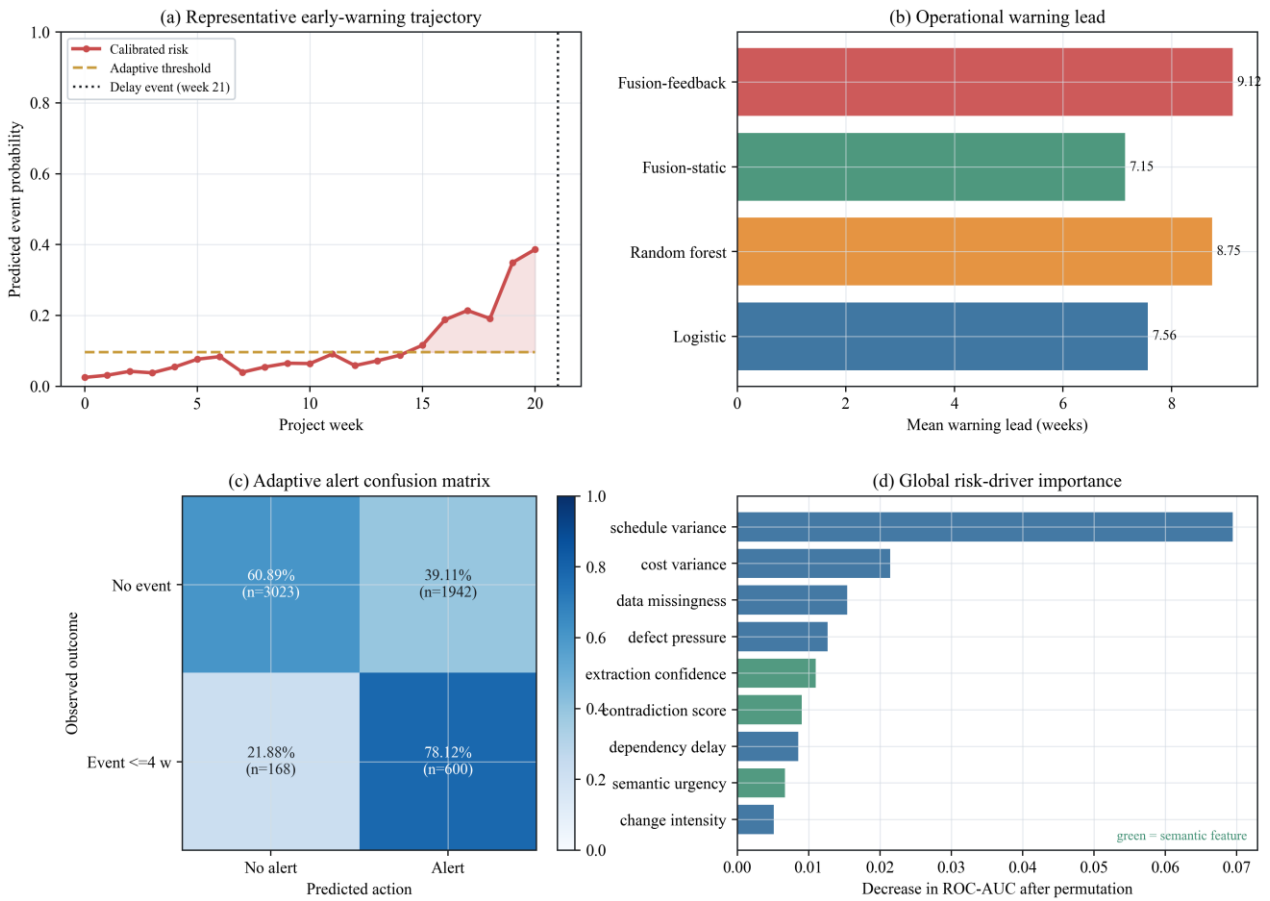


Figure 3. Representative risk trajectory, warning lead, confusion matrix, and risk drivers

5.3 Semantic ablation and threshold sensitivity

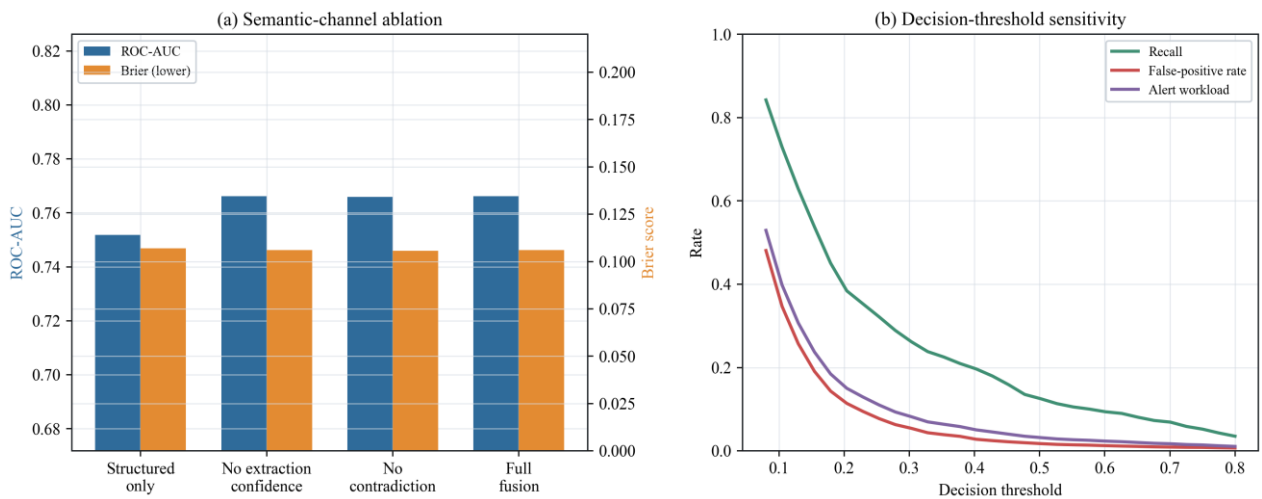


Figure 4. Semantic-feature ablation and alert-threshold sensitivity

The ablation model using only structured variables achieved ROC-AUC 0.752, PR-AUC 0.321, recall 0.642

and Brier score 0.107. The static semantic fusion model after adding all semantic variables increases ROC-AUC to 0.766, PR-AUC to 0.339, and reduces the Brier score to 0.106. After removing the extraction confidence, the ROC-AUC of the model was 0.766, and the recall rate was 0.634. After removing the contradictory variables, the ROC-AUC was still 0.766, and the recall rate was 0.663. Only a small change occurs after the removal of a single field, indicating that there may be a correlation and substitution relationship between different semantic variables. Whether the field is retained should not be based solely on a single global importance result. Extraction costs, data loss, evidence clarity, and local case review also need to be considered. The threshold curve shows the expected trade-off between recall rate, false alarm rate and workload, which can provide a basis for setting differentiated operation thresholds at different project levels.

The ablation subsection reports PR/AP values that differ from those in Table 3. Without the original split identifiers, per-observation scores and evaluation script, the difference cannot be attributed to feature removal alone. Accordingly, the ablation results should be read within Figure 4 and should not be subtracted from the main-experiment results as a controlled estimate of semantic benefit. A matched evaluation with identical splits, calibration data and alert budgets remains necessary.

5.4 Prevalence sensitivity and review workload

Figure 5 shows that higher recall does not necessarily give higher alert precision. At the test prevalence of 768/5,733, the expected precision is 0.265 for static fusion and 0.236 for feedback. Their expected alert fractions are 0.324 and 0.443. The central 95% simulation intervals for precision are 0.245-0.285 and 0.219-0.253, respectively. The corresponding workload intervals are 0.312-0.336 and 0.430-0.456. Lower prevalence reduces precision for both policies while leaving a substantial false-alarm workload. These are conditional simulation results at fixed operating rates.

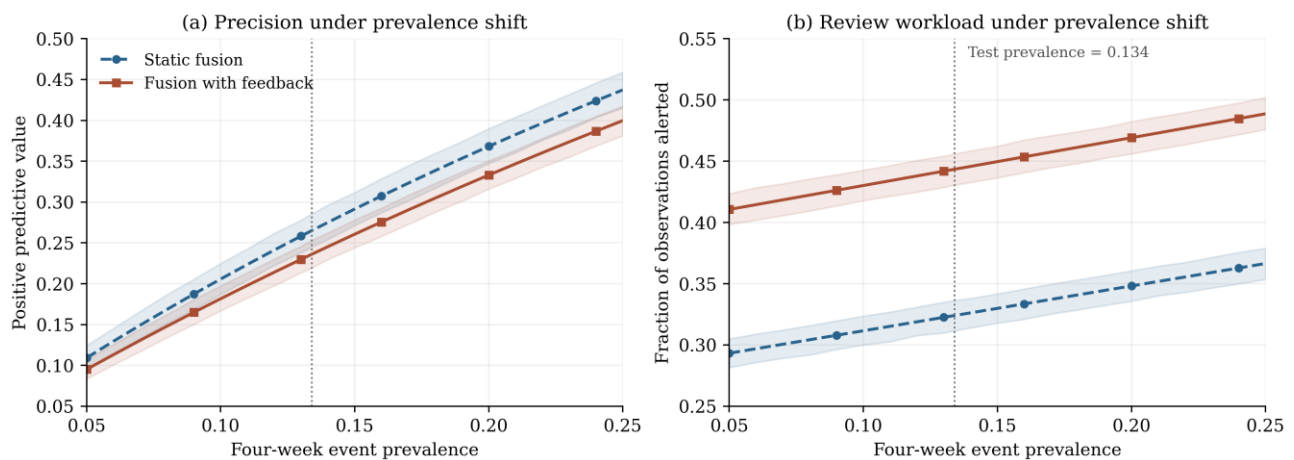


Figure 5. Supplementary prevalence sensitivity. Lines show analytical expectations and shading shows central 95% Monte Carlo intervals from 2,000 replicates per prevalence. The dotted line marks the reported test prevalence. Operating rates are fixed; these are not retrained-model results

Per 1,000 observations at the same prevalence, feedback generates about 119 additional alerts and identifies about 19 additional positive observations compared with static fusion. Approximately 101 of the additional alerts are false positives. A team able to review only one-third of weekly observations could accommodate the expected static alert load, but not the feedback load at these operating points. Deduplication by project, severity-aware triage and threshold adjustment require separate evaluation before workload savings can be claimed.

5.5 Sensitivity to missed-event and false-alarm costs

Figure 6 evaluates the two policies using (7). At the test prevalence, their expected error losses are equal at a missed-event/false-alarm cost ratio of 5.35. Below this ratio, static fusion has lower expected loss. Above it,

feedback has lower expected loss. For example, at a ratio of 6, the expected losses are 526.7 and 514.6 false-alarm cost units per 1,000 observations for static fusion and feedback, respectively. These quantities are normalized decision losses rather than currency amounts or measured project savings.

$$\rho^* = (1 - \pi)(f_1 - f_0) / [\pi(r_1 - r_0)] \quad (8)$$

In (8), subscripts 0 and 1 denote static fusion and feedback. The break-even ratio ρ^* is defined when feedback has both higher recall and a higher false-positive rate. It falls from 15.73 at a prevalence of 0.05 to 2.48 at a prevalence of 0.25. Thus a sensitive policy is more attractive when positives are more common or missed events are more costly. Actual policy selection must also consider review capacity, dependencies between weekly records and the consequences of repeated false alarms.

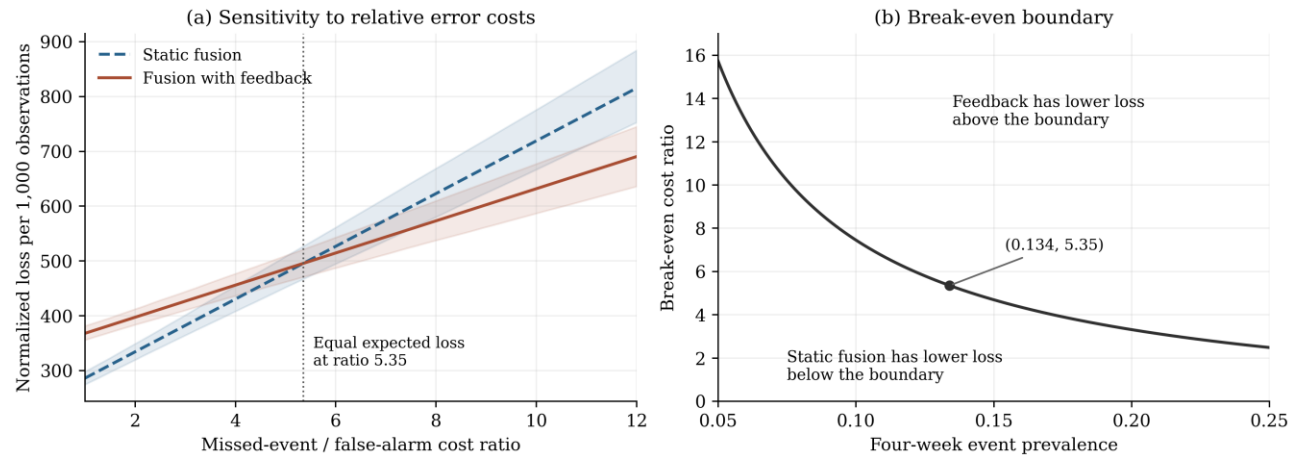


Figure 6. Supplementary error-cost sensitivity. Panel (a) uses the reported test prevalence and shows expected loss with central 95% Monte Carlo intervals. Panel (b) shows the analytical break-even cost ratio across prevalence. The analysis holds each policy’s recall and false-positive rate fixed

6 Conclusion

Power-LLM ValueLoop organizes heterogeneous power-project evidence into semantic variables, calibrated risk estimates and reviewable feedback. The supplied synthetic experiment reports small improvements in discrimination and higher recall for the feedback model, together with a higher false-positive rate. The added prevalence and cost simulations show that this trade-off depends on event frequency, error consequences and review capacity. At the reported test prevalence, feedback has lower expected error loss only when the missed-event cost is more than about 5.35 times the false-alarm cost under the stated simplified assumptions. The framework is therefore a candidate for further validation. Real extractor benchmarks, reproducible original project-level predictions and prospective deployment studies are needed before claiming stable operational or economic benefits.

References

- [1] International Organization for Standardization. ISO 31000:2018 Risk management - Guidelines. Geneva: ISO, 2018.
- [2] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017, 30.
- [3] Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 2019: 4171-4186.
- [4] Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 2020, 33: 9459-9474.

- [5] Lagos C I, Herrera R F, Mac Cawley A F, Alarcon L F. Predicting construction schedule performance with last planner system and machine learning. *Automation in Construction*, 2024, 167: 105716.
- [6] Egwim C N, Alaka H, Toriola-Coker L O, Balogun H, Sunmola F. Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications*, 2021, 6: 100166.
- [7] Zhao Y, Chen W, Arashpour M, Yang Z, et al. Predicting delays in prefabricated projects: SD-BP neural network to define effects of risk disruption. *Engineering, Construction and Architectural Management*, 2022, 29(4): 1753-1776.
- [8] Breiman L. Random Forests. *Machine Learning*, 2001, 45: 5-32.
- [9] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of KDD*, 2016: 785-794.
- [10] Lundberg S M, Lee S I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017, 30.
- [11] Guo C, Pleiss G, Sun Y, Weinberger K Q. On Calibration of Modern Neural Networks. *Proceedings of ICML*, 2017: 1321-1330.
- [12] Bifet A, Gavalda R. Learning from Time-Changing Data with Adaptive Windowing. *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007: 443-448.
- [13] Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of ICML*, 2006: 233-240.
- [14] Brier G W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950, 78(1): 1-3.
- [15] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 2001, 29(5): 1189-1232.
- [16] Zhong Y, Goodfellow S D. Domain-specific language models pre-trained on construction management systems corpora. *Automation in Construction*, 2024, 160: 105316. <https://doi.org/10.1016/j.autcon.2024.105316>
- [17] Saka A, Taiwo R, Saka N, et al. GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, 2024, 17: 100300. <https://doi.org/10.1016/j.dibe.2023.100300>
- [18] Wong S, Zheng C, Su X, Tang Y. Construction contract risk identification based on knowledge-augmented language models. *Computers in Industry*, 2024, 157-158: 104082. <https://doi.org/10.1016/j.compind.2024.104082>
- [19] Zheng C, Wong S, Su X, et al. Automating construction contract review using knowledge graph-enhanced large language models. *Automation in Construction*, 2025, 175: 106179. <https://doi.org/10.1016/j.autcon.2025.106179>
- [20] Tang Y, Zheng C, Kassem M, et al. Bridging the research-practice gap in construction contract management with NLP and LLMs. *Automation in Construction*, 2026, 181: 106614. <https://doi.org/10.1016/j.autcon.2025.106614>
- [21] Yao Q, Fang F, Chen Y, et al. AI Large Models for Power System: A Survey and Outlook. *IET Smart Energy Systems*, 2025, 1(1): 3-21. <https://doi.org/10.1049/ses2.70000>
- [22] Gao T, Yen H, Yu J, Chen D. Enabling Large Language Models to Generate Text with Citations. *Proceedings of EMNLP*, 2023: 6465-6488. <https://doi.org/10.18653/v1/2023.emnlp-main.398>
- [23] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *Proceedings of ICLR*, 2024. <https://openreview.net/forum?id=hSyW5go0v8>
- [24] Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of EAACL System Demonstrations*, 2024: 150-158. <https://doi.org/10.18653/v1/2024.eacli-demo.16>
- [25] Liu N F, Lin K, Hewitt J, et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 2024, 12: 157-173. https://doi.org/10.1162/tacl_a_00638
- [26] Gibbs I, Candès E J. Conformal Inference for Online Prediction with Arbitrary Distribution Shifts. *Journal of Machine Learning Research*, 2024, 25(162): 1-36. <https://jmlr.org/papers/v25/22-1218.html>
- [27] Erfani A, Mansouri A. Applications of multimodal large language models in construction industry. *Advanced Engineering Informatics*, 2026, 69: 103909. <https://doi.org/10.1016/j.aei.2025.103909>
- [28] Liu Y, Luo H. An efficient machine learning-based model for duration prediction of construction tasks with large-scale datasets. *Advanced Engineering Informatics*, 2026, 69: 103820. <https://doi.org/10.1016/j.aei.2025.103820>

Author's Biography



Zhang Jialin, Deputy Manager of Conventional Energy Project Division, Planning & Development Department, Guodian Power Development Co., Ltd., Safety Assessor. She was the principal author of special memoranda on technological research of new energy storage industry and large-scale equipment renewal & transformation for central SOEs, which were submitted for review by central leading officials. She presided over the compilation of Action Plan for Electric Power Work Safety (2018–2020), and participated in the compilation of 13th Five-Year Plan for Work Safety and 14th Five-Year Action Plan for Electric Power Work Safety. She has participated in compiling and publishing 9 books including the case warning textbook 25 Key Requirements for Preventing Electric Power Production Accidents and Electric Power Safety Governance. More than 30 articles written by her have been published in journals and newspapers such as Energy Work, China Energy News and Electric Power Equipment Management.



Zhang Siwen, Director of the New Energy Project Division, Planning & Development Department, Guodian Power Development Co., Ltd., takes overall responsibility for the full-process management of new energy projects at the preliminary stage. Deeply engaged in frontline new energy project development, he has solid expertise in key business areas including project scheme planning, design optimization, economic evaluation and investment decision-making, and has coordinated to deliver a host of new energy projects efficiently. He led the formulation of multiple management regulations for new energy projects, continuously improving the preliminary project management system and investment decision-making procedures to support standardized and efficient project delivery. He once took part in the reward fund project of China Energy Investment Corporation. With remarkable work achievements, he has repeatedly been honored as the Company's "Outstanding Employee".