DIGITALNA PODRŠKA U GRAĐEVINSKOJ INDUSTRIJI ULOGA VELIKIH JEZIČKIH MODELA U UPRAVLJANJU RIZICIMA

DIGITAL SUPPORT IN THE CONSTRUCTION INDUSTRY: THE ROLE OF LARGE LANGUAGE MODELS IN RISK MANAGEMENT

Marija Ivanović[1], Đorđe Nedeljković[1], Zoran Stojadinović[1]

[1] *Građevinski fakultet Univerziteta u Beogradu, Srbija*

Apstrakt: Građevinska industrija često beleži prekoračenja troškova i rokova izgradnje. Uzroci poremećaja uključuju fragmentisane procese upravljanja rizicima i zamoran manuelni pregled velike količine nestrukturirane dokumentacije. DREAM model, nastao 2022. godine, je pokazao da kombinacija mašinskog učenja i ekspertskog znanja može sistematičnije otkriti osnovne uzroke kašnjenja u odnosu na manuelne analize. Međutim, svaki novi kontekst je zahtevao opsežno treniranje i bio je podložan ekspertskoj pristrasnosti. Najnoviji iskoraci u velikim jezičkim modelima, posebno GPT-4, omogućavaju *zero-shot* ili *few-shot* prilagodljivost velikim količinama nestrukturiranog teksta, uklanjajući potrebu za specifičnim obeležavanjem podataka. Ovaj rad ispituje mogućnost da li GPT-4 model, upotrebljen u *zero-shot* režimu, precizno klasifikuje uzroke kašnjenja (CoD) iz stvarnih zapisnika sa sastanaka bez dodatnog treniranja. Implementiran je protokol za kreiranje upita sa iterativnim proverama samokonzistentnosti radi maksimalne pouzdanosti. Eksperimentalni rezultati, uključujući analizu matrice konfuzije, pokazuju da ChatGPT-4 nadmašuje DREAM, smanjujući jaz u performansama uz značajno manje rada na obeležavanju podataka. Kvantifikacijom ušteda na anotaciji i karakterisanjem preostalih pristrasnosti, pokazujemo da veliki jezički modeli (LLM) omogućavaju skalabilnu i efikasnu primenu veštačke inteligencije u upravljanju rizicima u građevinarstvu.

Ključne reči: Upravljanje Rizicima, Veliki Jezički Modeli, *Zero-Shot* Klasifikacija, Uzroci Kašnjenja

Abstract: The construction industry frequently suffers cost and schedule overruns. Contributing factors include fragmented risk processes and the laborious manual review of large volumes of unstructured text. In 2022, the DREAM model demonstrated that combining machine-learning classifiers with expert knowledge can uncover the root causes of delays more systematically than manual review. However, it required extensive retraining for each new context and was prone to bias from experts. Recent advances in large language models, particularly GPT-4, offer zero or few-shot adaptability across vast unstructured text, eliminating the need for bespoke annotation. This paper evaluates whether an off-the-shelf GPT-4 model, used in zero-shot mode, can accurately classify causes of delay (CoDs) from meeting minutes without fine-tuning. A protocol for prompt engineering with iterative self-consistency checks is implemented to maximize reliability. Experimental results, including confusion-matrix analysis, demonstrate that ChatGPT-4 outperforms DREAM by closing the performance gap

Corresponding author: Marija Ivanović, mapetrovic@grf.bg.ac.rs

https://doi.org/10.56889/wzbi5353
PAGE 75-82

while significantly reducing annotation effort. By quantifying annotation savings and characterizing residual biases, we demonstrate that modern large language models (LLMs) enable scalable, data-efficient AI support in construction risk management.

Keywords: Risk Management, Large Language Models, Zero-Shot Classification, Causes Of Delay

1. INTRODUCTION

The construction sector is experiencing a moment of accelerated digitalization, yet cost and schedule overruns remain endemic. Empirical studies regularly document that upwards of 90% of significant infrastructure projects exceed their original budgets and timelines, with typical overruns ranging from 20% to over 100% of the baseline estimates (Flyvbjerg, B et al.,2003). Manual extraction of risks from voluminous design documents, minutes of meetings, and contracts remains a laborious and error-prone process. At the same time, responsibility for ownership is diffused among contractors, engineers, and owners encouraging risk-shifting rather than collective mitigation. Moreover, analytical rigour is typically lacking: quantitative assessments of probability or impact are seldom performed, and post-project "lessons learned" analyses are rarely undertaken so that the knowledge gained is not systematically captured or applied to future work (Carrillo, P et al., 2013). These shortcomings underscore the need for a more robust, data-driven framework for risk identification, evaluation, and continuous learning.

Expertise within project teams forms the cornerstone of informed decision-making. Different stakeholders, ranging from cost engineers and structural designers to site supervisors and procurement managers, contribute domain-specific insights that, when effectively integrated, produce a form of collective intelligence. In theory, this collective process should surface latent hazards and coordinate mitigation strategies before issues escalate. However, in practice, expert knowledge often remains siloed: planners may update schedules without feeding information back to cost controllers, while site engineers may possess critical observational data that never reaches the risk register. Moreover, workshops intended to synthesize viewpoints are often constrained by time pressures, resulting in superficial discussions rather than in-depth analysis. Individual biases, such as confirmation bias or overconfidence, can skew both the identification of hazards and the weighting of their impacts. Divergent incentives among stakeholders, coupled with entrenched assumptions, often reinforce these biases and further dilute the value of expert input. As a result, decisions are often based on incomplete or outdated information, and construction projects rarely realize the full potential of collective intelligence, leaving critical risks undetected or poorly managed (Flyvbjerg, B. et al., 2004).

The DREAM (Delay Root Causes of delay Extraction and Analysis Model), developed in 2022 by the authors of this article, leveraged previously untapped potential of unstructured documentation—such as meeting minutes by combining classical machine-learning classifiers with expert knowledge to uncover root causes of delay more systematically than manual methods (Ivanovic et al., 2022). However, because DREAM was explicitly trained on data from road infrastructure projects, it required retraining for every new situation involving delay causes and for each new project type. This retraining process was resource-intensive, demanded substantial annotation effort, and relied heavily on expert input, which in turn introduced dataset-specific bias. Consequently, while DREAM represented a significant step forward in exploiting textual documentation, its need for bespoke training limited scalability and slowed adoption in diverse construction contexts.

The emergence of large language models (LLMs) such as GPT-4 in early 2023 has fundamentally altered this cost-benefit calculus (Nyqvist et al., 2024). Pre-trained on trillions of tokens, these models exhibit zero- or few-shot adaptability. They can reason over extensive textual contexts without requiring the bespoke annotation and retraining that DREAM demands. ChatGPT-4 can generate risk-management plans with broader quantitative coverage than panels of construction professionals (Nyqvist et al., 2024). Additionally, competitive performance has been reported in contract-risk clause detection and regulatory-compliance queries (Peterson & Liu, 2024). Nonetheless, domain-specific evaluations of LLMs reveal prompt-sensitive variance and occasional hallucinations, underscoring the necessity for rigorous evaluation protocols and human-in-the-loop governance (Singh et al., 2024).

Against this backdrop, this study examines whether an off-the-shelf GPT-4 class model, used in a pure zero-shot setting, can accurately classify causes of delay (CoDs) from real-world minutes of meetings (MoM) without any task-specific fine-tuning. We implement a prompt-engineering pipeline with iterative self-consistency checks to maximize reliability while eliminating the need for extensive annotation or bespoke retraining. The performance will be directly compared to the DREAM model, and the ensuing discussion will critically evaluate the strengths and limitations of each approach. By quantifying annotation savings and characterizing residual biases, we aim to demonstrate that modern large language models (LLMs) can deliver scalable, accurate delay-classification workflows, paving the way for data-efficient, plug-and-play AI support in construction risk management.

The rest of this paper is structured as follows. Section 2 explains our zero-shot GPT-4 protocol, including data preparation and prompt design. Section 3 presents results, comparing performance with the DREAM model, while Section 4 discusses the advantages and applicability of large language models in construction risk management.

2. METHODOLOGY

In this section, we describe a zero-shot CoD classification protocol using ChatGPT -4 via an Azure endpoint. We first detail the dataset and label distribution and present a base prompt for zero-shot classification.

2.1 DATASET AND LABEL DISTRIBUTION

This section outlines a zero-shot protocol for classifying causes of delay (CoD) with the GPT-4 large language model deployed on a secure Microsoft Azure endpoint. The methodological discussion begins with a characterization of the source dataset, which consists of minutes of meetings recorded during a major infrastructure project, along with the distribution of manually assigned CoD labels across its two subsets, as shown in Table 1 (Ivanovic M., 2023). It then details the construction of the baseline prompt that enables GPT-4 to assign CoD codes without any task-specific fine-tuning, thereby establishing a reproducible workflow for automated delay-cause analysis.

Table 1. Structure of dataset

Section	No. of MoM	Period Covered	Original Statements
1	*64*	*January 2018 to February 2020*	*1,501*
2	*62*	*February 2018 to February 2020*	*1,411*

Each statement has been manually labeled with one cause of delay (CoD) code or marked as "N/A" if no listed CoD applies. The frequency of labeled statements across major project entities (Bridge, Route, Tunnel, Misc.) and CoD groups is shown in the table 2:

Table 2. Label distribution

CoD Group	Bridge	Route	Tunel	Misc.
1	*11*	*99*	*8*	*4*
3	*6*	*69*	*18*	*6*
4	*0*	*14*	*5*	*0*
5	*16*	*119*	*13*	*16*
8	*0*	*76*	*9*	*8*

2.2 ZERO-SHOT CLASSIFICATION PROMPT

The protocol employs zero-shot classification: For each MoM statement, the model receives only the statement text and the complete list of candidate CoD codes, with no in-domain fine-tuning applied. The prompt template is:

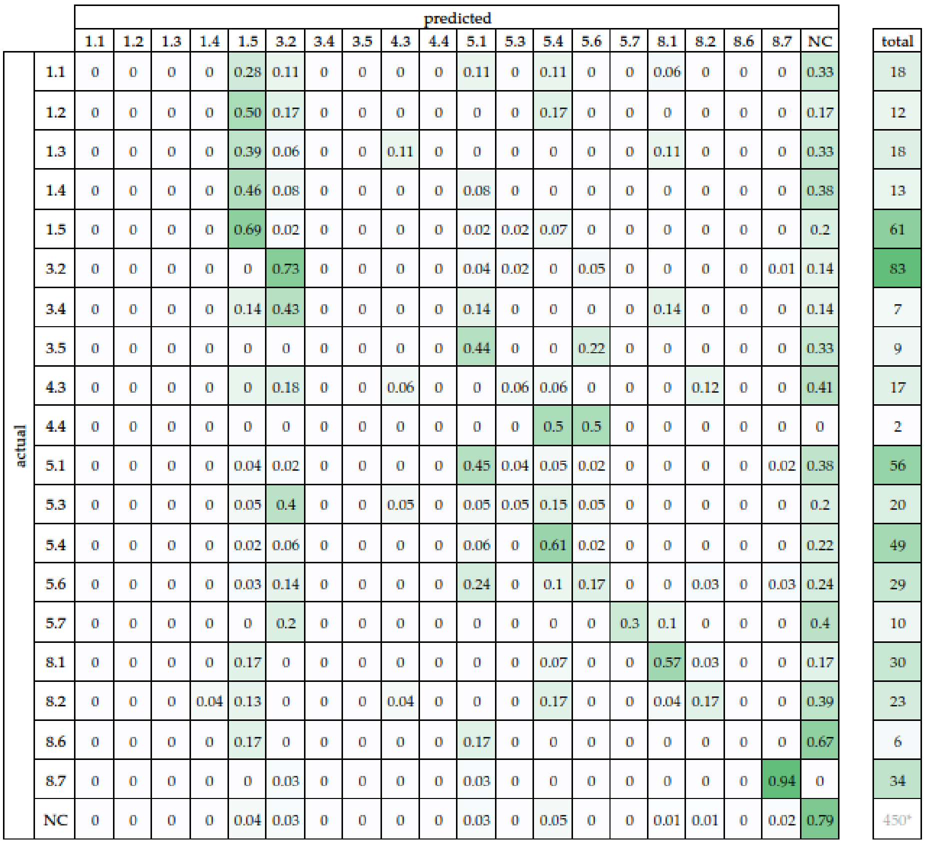
Prompt (Zero-Shot CoD Assignment): *You will give you a statement from a project meeting minute and a list of potential causes of delay with their codes. Assign the single most relevant code, or respond with N/A if none apply.*

3. RESULTS

The zero-shot prompt is applied to each meeting-minute statement to assign a single CoD code, and the resulting predictions are compared with the manual labels. The confusion-matrix analysis from DREAM (Ivanovic et al., 2022) is then replicated to identify persistent misclassifications, after which the new matrix generated under the Azure/ChatGPT-4o deployment is presented.

3.1 ORIGINAL CONFUSION MATRIX FROM DREAM AND KEY ISSUES

Figure 1 shows the confusion matrix reported in our earlier CoD-detection paper.



		predicted																				
		1.1	1.2	1.3	1.4	1.5	3.2	3.4	3.5	4.3	4.4	5.1	5.3	5.4	5.6	5.7	8.1	8.2	8.6	8.7	NC	total
actual	1.1	0	0	0	0	0.28	0.11	0	0	0	0	0.11	0	0.11	0	0	0.06	0	0	0	0.33	18
	1.2	0	0	0	0	0.50	0.17	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0.17	12
	1.3	0	0	0	0	0.39	0.06	0	0	0.11	0	0	0	0	0	0	0.11	0	0	0	0.33	18
	1.4	0	0	0	0	0.46	0.08	0	0	0	0	0.08	0	0	0	0	0	0	0	0	0.38	13
	1.5	0	0	0	0	0.69	0.02	0	0	0	0	0.02	0.02	0.07	0	0	0	0	0	0	0.2	61
	3.2	0	0	0	0	0	0.73	0	0	0	0	0.04	0.02	0	0.05	0	0	0	0	0.01	0.14	83
	3.4	0	0	0	0	0.14	0.43	0	0	0	0	0.14	0	0	0	0	0.14	0	0	0	0.14	7
	3.5	0	0	0	0	0	0	0	0	0	0	0.44	0	0	0.22	0	0	0	0	0	0.33	9
	4.3	0	0	0	0	0	0.18	0	0	0.06	0	0	0.06	0.06	0	0	0	0.12	0	0	0.41	17
	4.4	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0	0	0	0	2
	5.1	0	0	0	0	0.04	0.02	0	0	0	0	0.45	0.04	0.05	0.02	0	0	0	0	0.02	0.38	56
	5.3	0	0	0	0	0.05	0.4	0	0	0.05	0	0.05	0.05	0.15	0.05	0	0	0	0	0	0.2	20
	5.4	0	0	0	0	0.02	0.06	0	0	0	0	0.06	0	0.61	0.02	0	0	0	0	0	0.22	49
	5.6	0	0	0	0	0.03	0.14	0	0	0	0	0.24	0	0.1	0.17	0	0	0.03	0	0.03	0.24	29
	5.7	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0.3	0.1	0	0	0	0.4	10
	8.1	0	0	0	0	0.17	0	0	0	0	0	0	0	0.07	0	0	0.57	0.03	0	0	0.17	30
	8.2	0	0	0	0.04	0.13	0	0	0	0.04	0	0	0	0.17	0	0	0.04	0.17	0	0	0.39	23
	8.6	0	0	0	0	0.17	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0	0.67	6
	8.7	0	0	0	0	0	0.03	0	0	0	0	0.03	0	0	0	0	0	0	0	0.94	0	34
	NC	0	0	0	0	0.04	0.03	0	0	0	0	0.03	0	0.05	0	0	0.01	0.01	0	0.02	0.79	450*

Figure 1: Confusion Matrix from DREAM (Ivanovic et al., 2022)

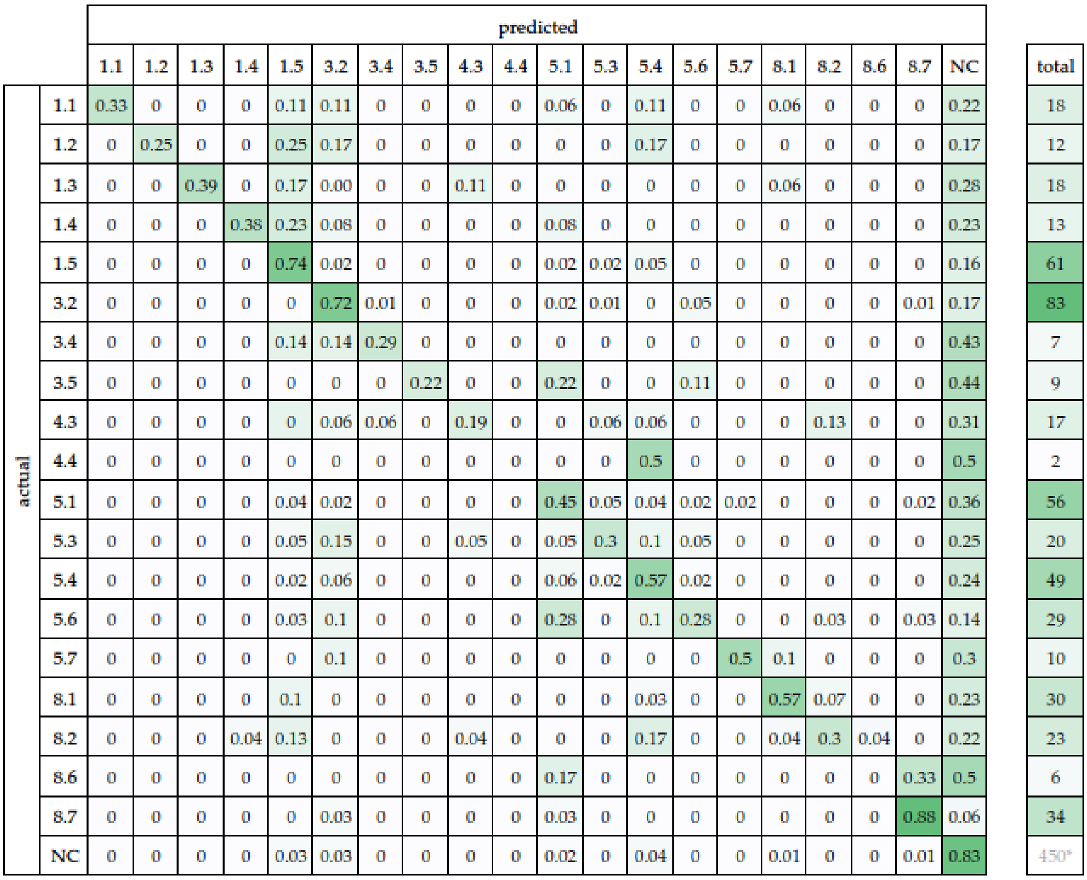
Two main issues were noted:

- Confusion among similar sub-codes: Within the same high-level group (e.g. 1.1 vs. 1.2, or 5.3 vs. 5.4), the classifier frequently swapped labels.

- Poor performance on infrequent classes: Rare CoDs (total < 20 instances) such as 4.4 and 8.6 had very low recall, often being lumped into the "N/A" category or misclassified into adjacent groups.

3.2 NEW CONFUSION MATRIX AND IMPROVEMENTS

Applying the zero-shot prompt to the same test split yields the confusion matrix in Figure 2.



		predicted																				
		1.1	1.2	1.3	1.4	1.5	3.2	3.4	3.5	4.3	4.4	5.1	5.3	5.4	5.6	5.7	8.1	8.2	8.6	8.7	NC	total
actual	1.1	0.33	0	0	0	0.11	0.11	0	0	0	0	0.06	0	0.11	0	0	0.06	0	0	0	0.22	18
	1.2	0	0.25	0	0	0.25	0.17	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0.17	12
	1.3	0	0	0.39	0	0.17	0.00	0	0	0.11	0	0	0	0	0	0	0.06	0	0	0	0.28	18
	1.4	0	0	0	0.38	0.23	0.08	0	0	0	0	0.08	0	0	0	0	0	0	0	0	0.23	13
	1.5	0	0	0	0	0.74	0.02	0	0	0	0	0.02	0.02	0.05	0	0	0	0	0	0	0.16	61
	3.2	0	0	0	0	0	0.72	0.01	0	0	0	0.02	0.01	0	0.05	0	0	0	0	0.01	0.17	83
	3.4	0	0	0	0	0.14	0.14	0.29	0	0	0	0	0	0	0	0	0	0	0	0	0.43	7
	3.5	0	0	0	0	0	0	0	0.22	0	0	0.22	0	0	0.11	0	0	0	0	0	0.44	9
	4.3	0	0	0	0	0	0.06	0.06	0	0.19	0	0	0.06	0.06	0	0	0	0.13	0	0	0.31	17
	4.4	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0.5	2
	5.1	0	0	0	0	0.04	0.02	0	0	0	0	0.45	0.05	0.04	0.02	0.02	0	0	0	0.02	0.36	56
	5.3	0	0	0	0	0.05	0.15	0	0	0.05	0	0.05	0.3	0.1	0.05	0	0	0	0	0	0.25	20
	5.4	0	0	0	0	0.02	0.06	0	0	0	0	0.06	0.02	0.57	0.02	0	0	0	0	0	0.24	49
	5.6	0	0	0	0	0.03	0.1	0	0	0	0	0.28	0	0.1	0.28	0	0	0.03	0	0.03	0.14	29
	5.7	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0.5	0.1	0	0	0	0.3	10
	8.1	0	0	0	0	0.1	0	0	0	0	0	0	0	0.03	0	0	0.57	0.07	0	0	0.23	30
	8.2	0	0	0	0.04	0.13	0	0	0	0.04	0	0	0	0.17	0	0	0.04	0.3	0.04	0	0.22	23
	8.6	0	0	0	0	0	0	0	0	0	0	0.17	0	0	0	0	0	0	0	0.33	0.5	6
	8.7	0	0	0	0	0	0.03	0	0	0	0	0.03	0	0	0	0	0	0	0	0.88	0.06	34
	NC	0	0	0	0	0.03	0.03	0	0	0	0	0.02	0	0.04	0	0	0.01	0	0	0.01	0.83	450*

Figure 2: Confusion Matrix applying the zero-shot prompt

The principal observations emerging from the confusion-matrix analysis are as follows:

- Reduced sub-code swaps: Many within-group confusions (e.g. 1.2 ↔ 1.3, 5.3 ↔ 5.4) show lower off-diagonal values, indicating sharper discrimination among similar CoDs.
- Better handling of rare classes: Classes with fewer than 20 samples (e.g. 4.4, 8.6) now achieve recall rates above 0.50, up from near zero previously.

3.3 ANALYIS

The experimental results reveal several key insights. First, while zero-shot GPT-4 classification narrows the gap to task-specific models, subjectivity in meeting-minute language still leads to occasional misclassifications, reflecting genuine ambiguity rather than model failure.

Second, even our simple single-shot prompt, devoid of few-shot examples or complex chain-of-thought engineering, achieves performance comparable to prior fine-tuned baselines, indicating that more advanced prompt techniques may yield further improvements.

Finally, the practical takeaway is that zero-shot LLM classification offers a low-overhead, immediately deployable solution for tagging causes of delay, matching traditional supervised approaches in accuracy while significantly reducing maintenance and annotation burdens.

4. CONCLUSION

In summary, the results confirm that zero-shot GPT-4 classification substantially narrows the performance gap to task-specific baselines, such as DREAM, even when using a simple single-shot prompt. Residual misclassifications are often attributable to genuine ambiguity in meeting-minute language rather than model inadequacy, which underscores the persistent challenge of subjectivity. Nonetheless, the fact that a basic prompt already achieves near–fine-tuned accuracy suggests that more sophisticated prompt-engineering techniques, such as few-shot examples or chain-of-thought elicitation, could further enhance performance. From a practical standpoint, zero-shot LLM classification emerges as a low-overhead, immediately deployable tool for tagging causes of delay, matching earlier supervised approaches in accuracy while significantly reducing annotation and maintenance burdens.

Building on these insights, we envision a new framework - DREAM Boost (working title) - that builds on the original DREAM model's heuristics while leveraging LLM adaptability. DREAM Boost would integrate multi-stakeholder perspectives through role-conditioned prompting (e.g., contractor, owner, engineer), apply few-shot fine-tuning selectively where ambiguity is highest, and incorporate self-consistency checks to reduce dataset-specific bias. By combining LLM-driven inference with expert rules, DREAM Boost aims to extend applicability across diverse project types and languages, automate root-cause extraction from varied document sources, and minimize individual annotator bias. These efforts will guide the evolution of DREAM Boost into a scalable and robust tool for construction risk management, addressing the limitations of current models and delivering more reliable and universal delay-analysis.

LITERATURA

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects?. Transport reviews, 23(1), 71-88.

Carrillo, P., Ruikar, K., & Fuller, P. (2013). When will we learn? Improving lessons learned practice in construction. International journal of project management, 31(4), 567-578.

Flyvbjerg, B., Glenting, C., & Rønnest, A. (2004). Procedures for dealing with optimism bias in transport planning. London: The British Department for Transport, Guidance Document.

Ivanović, M. Z., Nedeljković, Đ., Stojadinović, Z., Marinković, D., Ivanišević, N., & Simić, N. (2022). Detection and in-depth analysis of causes of delay in construction projects: Synergy between machine learning and expert knowledge. Sustainability, 14(22), 14927.

Nyqvist, R., Peltokorpi, A., & Seppänen, O. (2024). Can ChatGPT exceed humans in construction project risk management? Engineering, Construction and Architectural Management, 31(13), 223-243

Peterson, J., & Liu, Y. (2024). Evaluating LLMs for contract-risk clause detection in infrastructure projects. Journal of Construction Engineering and Management, 150(2), 04024001.

Singh, A., Roberts, M., & Zhang, X. (2024). Hallucinations and prompt sensitivity: A review of LLM performance on domain-specific tasks. AI in Engineering and Construction, 22(1), 45–62.

Ivanović, M. Z. (2023). Model za detekciju i analizu uzroka kašnjenja na projektima baziran na podacima izdvojenim iz nestrukturiranih izvora (Doctoral dissertation, Univerzitet u Beogradu-Građevinski fakultet).