

Large Language Models as Virtual Personas to Estimate Construction Practitioners' Perceptions of Robotic Automation

Francis Baek, Ph.D.¹; Minjun Chang²; and Eunbee Kim, Ph.D.³

¹Assistant Professor, School of Civil and Environmental Engineering, Georgia Institute of Technology. Email: francis.baek@ce.gatech.edu

²Ph.D. Student, School of Civil and Environmental Engineering, Georgia Institute of Technology. Email: minjunchang@gatech.edu

³Academic Professional, School of Psychology, Georgia Institute of Technology. Email: ekim383@gatech.edu

ABSTRACT

Identifying construction practitioners' perceptions like attitudes, acceptance, and concerns regarding robotic automation is essential for successful deployment, yet conventional reliance on large-scale surveys can yield incomplete or outdated insights. This study examines Large Language Models (LLMs) as virtual personas to estimate practitioners' responses to perception-based surveys in a time-efficient and scalable manner. Two frontier LLMs were independently tested by defining personas and collecting responses through prompting practitioner demographics, automation scenarios, and survey items with text descriptions. Across five surveys involving 186 items and 989 practitioners, major results indicate a robust alignment evidenced by a micro-average Pearson correlation of 0.47 [95% CI: 0.31, 0.60] and a micro-average Spearman correlation of 0.33 [95% CI: 0.18, 0.46]. Findings present the potential that LLMs can provide responses highly aligning with practitioners, which can contribute to determining effective strategies to advance robotic automation for more productive and safer construction.

INTRODUCTION

While the construction industry has suffered from challenges such as stagnant productivity growth, high safety risk, and labor shortage, construction robots have gained increasing attention (Ayoola et al. 2024). Robotic automation is expected to alleviate the challenges by assisting human workers with physically demanding, dangerous, and repetitive tasks. Construction robots have already shown promising potentials to improve productivity and safety across several tasks such as bricklaying, rebar-tying, and wood assembly (Ayoola et al. 2024).

To realize this potential, identifying how construction practitioners—including on-site laborers to managerial personnel like general contractors—perceive robotic automation (Vora et al. 2024). Perceptions refer to the mental process of interpreting and evaluating a subject, such as attitudes, acceptance, and concern, which can influence practitioners' willingness to adopt and use robots, and, in turn, shape operational performance after adoption. While negative perceptions like safety concerns and resistant attitudes have already been often reported, identifying widespread perceptions of practitioners could be essential to determining effective strategies for the successful deployment of robotic automation in the construction industry (Vora et al. 2024).

A conventional but highly effective method for capturing perceptions across construction practitioners is the use of surveys (Emaminejad et al. 2024). Prior studies have used surveys to

examine practitioners' perceptions including attitudes, acceptance, and concern, regarding emerging robotic technologies in construction. Although surveys often suffer from subjective human errors like bias, and advancements in sensors and deep-learning technologies have enabled the objective analysis of physiological and behavioral indicators to infer perceptions, surveys remain competitive as a scalable and direct means of measuring perceptions (Kohn et al. 2021).

Despite their methodological advantages, administering large-scale surveys to capture perceptions across practitioner groups with diverse work roles and site contexts remains challenging in construction (Chen et al. 2015). Survey implementation can be constrained by limited site access and safety protocols, highly variable schedules with narrow and unpredictable time windows, and a fragmented workforce across multiple organizations and subcontracting tiers (Chen et al. 2015). As a result, survey evidence on practitioners' perceptions of robotic automation remains largely limited in coverage and timeliness—for example, confined to a single project or trade, based on one-time cross-sectional snapshots, or outdated by the time deployment decisions are made. Such limitations can hinder effective adoption and deployment strategies of robotic automation by yielding an incomplete or outdated understanding of practitioners' perceptions.

Recently, large language models (LLMs) have drawn growing attention, particularly in the social sciences, as a promising tool for generating human-like survey responses (Zhao et al. 2025). Leveraging their reasoning and generative capabilities, LLMs can be conditioned on text-based demographic profiles and survey items to produce responses that approximate those of human participants. Although human respondents remain essential for validating authentic responses, this capability suggests that LLMs could help proactively characterize potential responses of human participants in a more time-efficient and scalable way, while also highlighting the most contentious survey items so that subsequent human-participant surveys can be more efficient.

However, the potential of LLMs to serve as virtual respondents for construction practitioners' perceptions of robotic automation remains largely unexamined. Although LLMs have shown promise as proxies for general-population survey responses in social science contexts, construction practitioners can exhibit different response patterns shaped by physically demanding, hazardous, and productivity-oriented work environments (Bakker and Demerouti 2007). While current LLM applications in construction have primarily focused on technical tasks like natural language interpretation or robot control, the suitability of LLMs as virtual respondents for construction practitioners should be empirically validated rather than assumed.

In this study, we examine the extent to which LLMs can serve as virtual personas for estimating construction practitioners' perceptions of robotic automation—including attitudes, acceptance, and concerns—by statistically comparing practitioners and LLM responses to the same surveys, which cover practitioners' perceptions of different robotic automations scenarios.

This study demonstrates the feasibility of applying LLMs to approximate construction practitioners' perceptions of robotic automation. The findings present the potential that LLMs can serve as a supplementary tool for time-efficient, scalable exploration of how practitioner perceptions may vary across robotic automation scenarios, which can help identify effective strategies for the successful deployment of robotic automation in construction.

LLM-BASED SURVEY RESPONSE COLLECTION

We selected five surveys administered to construction practitioners and two frontier LLMs with powerful reasoning capabilities. Notably, these surveys use Likert-scale ratings, which could facilitate the direct statistical comparison between responses of practitioners and LLMs. For each survey, we applied three steps to collect responses from each LLM (Figure 1): 1) persona

conditioning using text descriptions of the survey participants' demographics to instantiate a corresponding virtual persona; 2) response-probability estimation given text descriptions of the survey items to elicit an expected probability distribution over Likert options for each item; and 3) response sampling from the elicited distributions to draw a set of responses matched to the practitioner sample size for direct comparison.

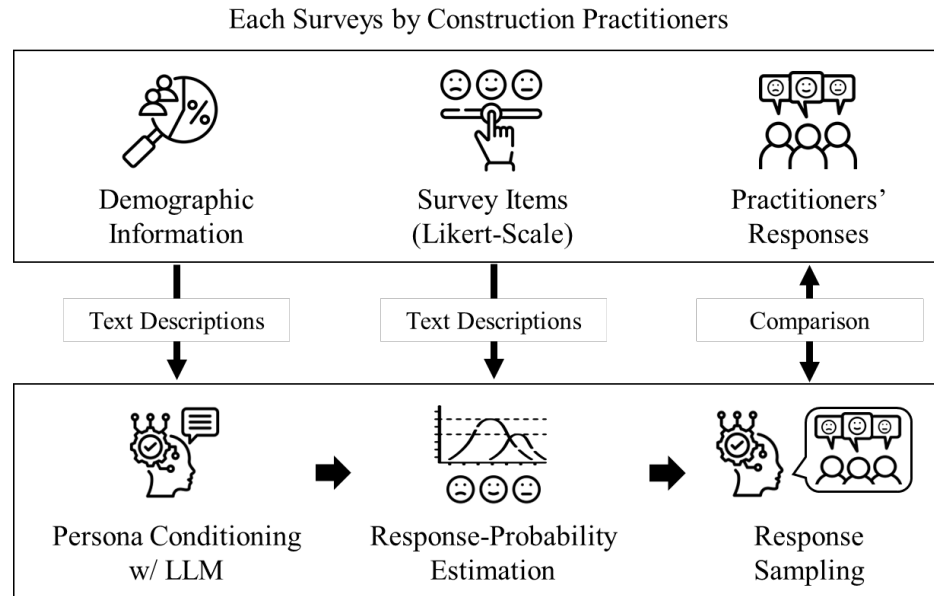


Figure 1. Overview of LLM-based survey response collection procedure.

Five Surveys. The five surveys collectively cover a broad spectrum of construction practitioners' perceptions relevant to robotic automation, including attitudes toward human–robot collaboration (HRC), acceptance of automation, trust in collaborative robots, and expert risk judgments regarding emerging technologies like exoskeletons and unmanned aircraft system (UAS)-enabled applications (Table 1). In Table 1, “# of Items” denotes the number of distinct rated items, and “# of Participants” denotes the number of unique respondents (not summed across Delphi rounds).

Table 1. Five surveys for construction practitioners' perceptions of robotic automation.

Surveys	# of Items	# of Participants
Attitudes toward HRC and interacting with robots (Sam and Franz 2025)	6	116
Acceptance of an intelligent surveillance/automation system (Lu and Deng 2022)	38	220
Assessing Trust in Construction AI-Powered Collaborative Robots (Emaminejad et al. 2024)	77	600
UAS (drone) safety risk factors for construction workers (Turkan and Xu 2022)	23	19
Ethical and social risks of exoskeletons (Ogunseiju et al. 2025)	42	34
	Total: 186	Total: 989

The first survey examined construction professionals to identify distinct attitude groupings toward interacting with robots and adopting HRC (Sam and Franz 2025). It used six statements

rated on a 4-point Likert scale without a neutral option. The study also collected demographic information—such as gender, age, education, organizational role, and years of experience.

The second survey examined construction practitioners' acceptance of an intelligent field system using structured questionnaires administered in separate versions for workers and managers (Lu and Deng 2022). Survey items were measured on a 5-point Likert scale ranging from “strongly disagree” to “strongly agree.” The final valid dataset included 52 workers and 168 managers, and the study reported respondent demographic distributions such as age, education, job experience, and position to characterize the population and support group-level interpretation.

The third survey was an online, large-sample study that quantified construction practitioners' trust and related perceptions toward AI-powered collaborative robots (cobots) (Emaminejad et al. 2024). Respondents rated 28 Likert-scale items (1–5) and additionally evaluated 7 cobots across 7 attributes on 5-point Likert scale, yielding 49 additional ratings; in total, this study comprised 77 rated items. The final analytic dataset included 600 respondents.

The fourth survey was to identify, verify, and quantify causal factors associated with safety risks related to UAS use in construction (Turkan and Xu 2022) using a Delphi approach. The Delphi method is a structured, multi-round expert elicitation process designed to achieve consensus through iterative feedback and controlled aggregation of responses. Because Delphi participation often decreases across rounds, we report Round-1 qualified expert panel sizes as a consistent reference point for participant counts across surveys. Experts rated agreement with candidate causal factors on a 5-point Likert scale, and consensus was assessed using dispersion-based criteria summarized by statistics such as means and standard deviations (SDs). After qualification screening, 19 experts formed the initial panel.

The fifth survey conducted a three-round Delphi process to examine ethical and social risks associated with construction exoskeletons (Ogunseiju et al. 2025). The initial expert panel comprised 34 participants. Experts rated agreement with 42 candidate risk items on a 5-point Likert scale and iteratively reviewed aggregated results across rounds to refine and confirm consensus. Overall, these five reference surveys comprise a total of 186 questions, representing the collective perceptions of 989 construction practitioners.

Baseline LLM Selection. We independently applied two frontier LLMs—ChatGPT (GPT-5.2) and Gemini 3 Pro—from two major providers (OpenAI and Google) to examine whether the capability to approximate practitioner perceptions can be consistent across LLMs with different architectures and training datasets. These two LLMs were selected based on their expected proficiency in complex reasoning and perspective-taking, as evidenced by their state-of-the-art performance on standardized benchmarks such as Big Theory of Mind (BigToM) (Kosinski 2024; Wang and Zhao 2023), which measure comprehensive knowledge and problem-solving capabilities and the capacity to infer human mental states and intentions, respectively.

Persona Conditioning. The LLMs were conditioned on survey participants' demographic information and instructed to provide responses as a part of the participant group. While instantiating LLMs to serve as individual respondents is a possible alternative, surveys typically do not provide granular, respondent-level demographic records. To avoid fabricating groundless individual profiles, we adopted group-level conditioning. For each participant group defined in the reference surveys, we converted reported demographic distributions—such as age brackets, occupational roles, and years of experience—into a concise textual profile. The LLM was

conditioned on this profile and instructed not to introduce extraneous personal attributes. For example, a group profile from the third survey was provided as follows:

*Respond as a typical member of the Project Manager group (n=220). Gender/ethnicity
Age: 25–34 (40%), 35–44 (35%), 45–54 (20%), 55+ (5%).
Experience: 1–3 years (15%), 4–10 years (45%), 10+ years (40%). Age ~exp.years
Role: Project manager (100%). Rule: Do not add personal details beyond this profile.*

Response-Probability Estimation. The LLMs were presented with the original survey items, preserving the wording and item order. In response to the survey items, LLMs were instructed to return Likert response probabilities p_k , which represents the estimated probability that a member of the target group would select option k for a survey item. This probabilistic output captures both the group’s most likely response pattern (i.e., which Likert option is most likely) and how concentrated or dispersed the responses are across options, reflecting the degree of consensus within the group. A sample prompt for a 4-point Likert scale was as follows:

You are completing an anonymous research survey as the construction practitioner described in the provided group profile. There are no right or wrong answers. Do NOT provide any explanations. For each survey item, output probabilities $p_1 \dots p_4$ that sum approximately to 1.0. Likert scale: 1 = Strongly disagree; 2 = Disagree; 3 = Agree; and 4 = Strongly agree.

Response Sampling. Using the item-specific probability distributions, we sampled discrete Likert scores to construct a synthetic respondent-level dataset comparable to the human survey responses. For each survey, the number of sampled responses per item was set to match the original human sample size. When surveys reported multiple subgroups, we sampled within each subgroup using its corresponding distribution and sample size and then concatenated subgroup samples to preserve the original population composition. The sampled responses were then aggregated to compute group-level descriptive statistics—mean (μ_i) and variance (σ_i) for each survey item i —enabling a direct quantitative comparison with the original human survey data.

ANALYSIS AND RESULTS

We evaluated how closely the LLM-generated responses match the human survey responses by comparing: 1) the average item-level discrepancy; and 2) overall alignment—whether items rated higher and lower, and the extent of those differences, are preserved—across all 186 survey items.

Evaluation Metrics. We utilized two sets of metrics, commonly applied to assess average item-level discrepancy and overall alignment of estimation. First, we computed range-normalized Mean Absolute Error (normMAE) and range-normalized Mean Error (normME) for means (μ_i) and SDs (σ_i) of each item i , where normalization divides errors by the Likert scale range to enable comparison across mixed scale lengths. Notably, normMAE quantifies the average magnitude discrepancy between LLM-generated and practitioners’ responses, while the signed normME captures whether the LLM tends to consistently overestimate or underestimate practitioner ratings across items (i.e., directional bias). For overall alignment, we computed Pearson (r) and Spearman (ρ) correlations between LLM-generated and practitioner responses across all 186 items (micro-average). Pearson (r) reflects similarity in the magnitude of item-level mean responses, whereas Spearman (ρ) reflects whether items are ranked similarly as higher or lower. Here, 95% Confidence Intervals (CI) were estimated via a non-parametric bootstrap with 5,000 iterations by

resampling the original 186 survey items with replacement, which can provide insights into how consistent the overall alignment remains across different sets of survey items.

Item-Level Discrepancy. Across the 186 items, GPT-5.2 achieved a normMAE of 0.099 for item means and 0.033 for item-level SDs (Table 2), corresponding to absolute errors of 0.40 and 0.13 on a 5-point scale. Although no universal benchmark exists for “good” survey-estimation error, the mean error of 0.40 is promising because it remains below the 0.5-point threshold that would induce a categorical shift (e.g., from Likert score 3 to 4). In parallel, normME indicates that GPT-5.2 exhibits only a small directional tendency: for item means, normME is -0.037 (≈ -0.15 points on a 5-point scale), and for item-level SDs, normME is 0.015 (≈ 0.06 points on a 5-point scale), suggesting little to no consistent over- or under-estimation compared to practitioners’ responses. Notably, Gemini 3 Pro also achieved promising performance, with normMAE of 0.139 for means and 0.044 for SDs, and normME of 0.004 for means and -0.017 for SDs (Table 2). These results present the potential reliability of LLMs in generating item-level responses with small absolute discrepancies and low directional bias relative to construction practitioners.

Table 2. NormMAE of GPT-5.2 and Gemini 3 Pro.

Metrics	GPT-5.2	Gemini 3 Pro
NormMAE of Means	0.099	0.139
NormMAE of Standard Deviations (SDs)	0.033	0.044
NormME for means	-0.037	0.004
NormME of Standard Deviations (SDs)	0.015	-0.017

Overall Alignment. GPT-5.2 exhibited a solid alignment with a micro-average Pearson correlation coefficient of 0.47 [95% CI: 0.31, 0.60] and a Spearman correlation coefficient of 0.33 [95% CI: 0.18, 0.46] (Table 3). Considering that Pearson and Spearman coefficients greater than 0.30 are often considered large effects in human behavior and psychology research (Gignac and Szodorai 2016), these results can be interpreted that responses generated by GPT-5.2 are well-aligned with construction practitioners’ responses across the five surveys. However, it is noteworthy that the alignment varied substantially by survey—for example, GPT-5.2 showed very high alignment in Surveys 1, 3, and 5, whereas alignment was weaker and less stable in Surveys 2 and 4, where CI included zero (Table 3). These results present the need for further analysis to identify potential drivers of estimation performance, such as item content, text descriptions, and respondent group characteristics, and to determine when LLM-based personas can be most reliable.

While GPT-5.2 generally exhibited higher overall alignments, Gemini 3 Pro also results in promising alignment with practitioners’ perceptions by achieving micro-average Pearson correlation of 0.27 [95% CI: 0.10, 0.41] and Spearman of 0.21 [95% CI: 0.06, 0.35], which can be interpreted as moderate effect sizes. In the fourth survey, Gemini 3 Pro achieved higher alignment compared to GPT-5.2. However, it is noteworthy that Gemini 3 Pro exhibited an inverse correlation in the second survey (Table 3), indicating that responses by Gemini 3 Pro showed opposite trends compared to construction practitioners’ responses. These results present the potential that certain LLMs can outperform others depending on survey context (e.g., items and practitioner groups), which highlights the need for further analysis examining LLMs’ capabilities

under different robotic automation scenarios and practitioner groups to identify the most effective approach for applying LLMs as virtual personas of construction practitioners.

Table 3. NormMAE of ChatGPT-5.2 and Gemini 3 Pro.

Surveys	GPT-5.2		Gemini 3 Pro	
	Pearson [95% CI]	Spearman [95% CI]	Pearson [95% CI]	Spearman [95% CI]
1	0.98 [0.85, 1.00]	0.89 [0.20, 1.00]	0.94 [0.77, 1.00]	0.77 [0.00, 1.00]
2	0.25 [-0.11, 0.59]	0.25 [-0.10, 0.56]	-0.20 [-0.40, 0.01]	-0.22 [-0.48, 0.08]
3	0.66 [0.49, 0.77]	0.62 [0.46, 0.75]	0.52 [0.34, 0.67]	0.52 [0.32, 0.68]
4	0.18 [-0.24, 0.55]	0.12 [-0.37, 0.55]	0.34 [-0.08, 0.65]	0.30 [-0.16, 0.68]
5	0.58 [0.34, 0.75]	0.54 [0.25, 0.74]	0.59 [0.35, 0.75]	0.54 [0.24, 0.74]
Micro-Avg.	0.47 [0.31, 0.60]	0.33 [0.18, 0.46]	0.27 [0.10, 0.41]	0.21 [0.06, 0.35]

DISCUSSION

This study examined the extent of LLMs to serve as virtual proxies for construction practitioners' perceptions of robotic automation. The results demonstrate that LLM-generated responses are statistically well-aligned with the responses of construction practitioners. It is particularly noteworthy that the alignment is achieved across surveys involving 186 survey items, multiple perceptual dimensions (e.g., attitudes, acceptance, and concerns) and responses from 989 construction practitioners. These findings present the potential of LLMs to support time-efficient and scalable identification of construction practitioners' perception of robotic automation by only providing text descriptions of demographics, robotic automation scenarios, and survey queries.

The promising findings of this study raise new questions regarding the potential of LLMs as virtual respondents of construction practitioners. These questions can include: 1) how LLMs would perform under more diverse robotic automation scenarios and practitioner groups; 2) what can be driving factors of LLMs' performance and which LLM would work best; 3) how well LLM-based personas would function for emerging robotic automation scenarios (e.g., humanoids) where published survey evidence—that might have been exposed to LLMs—is limited; and 4) whether the same capability extends beyond one-time, self-reported Likert surveys to repeated measurements over time, and even to estimating more objective measurements linked to perceptions, such as behavioral or psychophysiological indicators. Our future research will address these questions to further explore the potential of LLMs to serve as virtual respondents of construction practitioners, with the goal of reducing the traditional bottlenecks of relying solely on large-scale surveys and supporting successful deployment of robotic automation in construction.

CONCLUSION

Building on the need for time-efficient and scalable approach to identify construction practitioners' perceptions of robotic automation, this study evaluated the potential of LLMs to serve as virtual respondents. Major results demonstrate well-alignment between LLM-generated and practitioners responses, as evidenced by a low item-level errors (normMAE of 0.099 for means and 0.033 for SDs) with little bias (normME of -0.037 for means and 0.015 for SDs), a micro-average Pearson correlation of 0.47 [95% CI: 0.31, 0.60], and a micro-average Spearman correlation of 0.33 [95% CI: 0.18, 0.46]. This alignment is particularly noteworthy as it is achieved across five surveys encompassing 186 survey items, multiple perceptual dimensions like attitudes, acceptance, and concerns, and 989 construction practitioners. Given these promising results, follow-up studies will

further explore the potential of LLMs, such as testing broader scenarios and practitioner groups, comparing LLMs, validating emerging cases with limited prior survey evidence, and extending beyond one-time Likert surveys to longitudinal and objective measures. The findings of this study can provide a solid foundation for future research, aiming at mitigating the traditional bottlenecks of large-scale human surveys and supporting the successful deployment of robotic automation for more productive and safer construction.

REFERENCES

- Ayoola, V. B., Idoko, P., Danquah, E. O., Ukpoju, E. A., Obasa, J., Otakwu, A., and Enyejo, J. O. (2024). "Optimizing Construction Management and Workflow Integration through Autonomous Robotics for Enhanced Productivity Safety and Precision on Modern Construction Sites." *International Journal of Scientific Research and Modern Technology (IJSRMT)*, 3(10).
- Bakker, A. B., and Demerouti, E. (2007). "The job demands-resources model: State of the art." *Journal of managerial psychology*, 22(3), 309–328.
- Chen, Y., Alderman, E., McCabe, B. Y., and Hyatt, D. (2015). "Data collection framework for construction safety research." *Proc., ICSC'15—The Canadian Society for Civil Engineering's 5th International/11th Construction Specialty Conference*, Vancouver, BC, Canada, June 7–10.
- Emaminejad, N., Kath, L., and Akhavian, R. (2024). "Assessing trust in construction AI-powered collaborative robots using structural equation modeling." *Journal of Computing in Civil Engineering*, 38(3), 04024011.
- Gignac, G. E., and Szodorai, E. T. (2016). "Effect size guidelines for individual differences researchers." *Personality and individual differences*, 102, 74–78.
- Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). "Measurement of trust in automation: A narrative review and reference guide." *Frontiers in psychology*, 12, 604977.
- Kosinski, M. (2024). "Evaluating large language models in theory of mind tasks." *Proceedings of the National Academy of Sciences*, 121(45), e2405460121.
- Lu, Y., and Deng, Y. (2022). "What drives construction practitioners' acceptance of intelligent surveillance systems? An extended technology acceptance model." *Buildings*, 12(2), 104.
- Ogunseiju, O., Cho, Y. K., Tomori, M., and Rudraraju, L. (2025). "Virtual Boundaries: Investigating the Ethical and Social Risks of Exoskeletons in the Construction Industry."
- Sam, M., and Franz, B. (2025). "Classification of construction professionals' attitudes toward using robots in applications of Human Robot Collaboration." *Journal of Information Technology in Construction (ITcon)*, 30(18), 418–438.
- Turkan, Y., and Xu, Y. (2022). "A practical model for measuring and mitigating safety risks of using UAS in construction."
- Vora, C., Aryal, A., Willoughby, S., and Wang, C. (2024). "Investigating stakeholder perception and developing a decision framework for robot adoption in construction." *Journal of Construction Engineering and Management*, 150(4), 04024012.
- Wang, Y., and Zhao, Y. (2023). "Gemini in reasoning: Unveiling commonsense in multimodal large language models." *arXiv preprint arXiv:2312.17661*.
- Zhao, J., Yuan, C., Luo, W., Xie, H., Zhang, G., Quan, S. J., Yuan, Z., Wang, P., and Zhang, D. (2025). "Large Language Models as Virtual Survey Respondents: Evaluating Sociodemographic Response Generation." *arXiv preprint arXiv:2509.06337*.