

EVALUATING LOCAL SMALL LANGUAGE MODELS FOR ARCHITECTURAL MEETING TRANSCRIPTION SUMMARIZATION AND ACTION EXTRACTION

Sean Bayly^{1,2}, Sneha Verma¹, Mohamad Kassem¹ and Huizhi Liang¹

¹Newcastle University, ²Ryder Architecture

Abstract

This paper presents a locally hosted solution, for LLM driven summarization and action extraction for project-based meetings, that addresses the privacy, confidentiality, and cost concerns of externally hosted solutions. Developed for architectural project coordination, this system reduces the administrative burden of documenting multidisciplinary design meetings, tracking actions, and maintaining project records across distributed teams. Qwen3.6 27B was used as a judge to evaluate meeting transcript summarization and action extraction by Qwen3.5 4B and 9B SLMs in BF16 and AWQ INT4 formats. The 9B models achieved the best quality-speed tradeoff, producing accurate actions and concise, grounded summaries with lower wall-clock times.

Introduction and Related Research

Coordination between teams in the built environment involves frequent meetings with manual tracking of actions being onerous and prone to error (Ghimire et al., 2024). Many existing summarization workflows rely on external AI providers (e.g., Fireflies.ai and Otter.AI), requiring meeting transcripts and associated project discussions to be transmitted beyond the organisation's primary collaboration platform. In architectural practice, where meetings may contain commercially sensitive, contractual, or personally identifiable information, this can introduce additional privacy, data governance, and compliance concerns (Yan et al., 2024).

Table 1: Comparative Analysis of Meeting Intelligence Systems

Systems	Automatic Capture	Multi-Speaker Structuring	Project-Level Integration	Role-Based Access Control
Otter.ai	Yes	Strong	Limited	Platform-level
Fileflies.ai	Yes	Strong	Limited	Platform-level
Meeting Intelligence System	Yes (Graph Webhooks)	Structured Transcript Chunk	Deep relational embedding	Azure AD hierarchical role

This risk can be mitigated through the local deployment of LLM based models, which limits the egress of transcript data to third-party services during inference. The primary research challenge, therefore, is whether locally hosted SLMs can generate meeting summaries and action items of sufficient quality for use in architectural project coordination. In this paper, we investigated 4B and 9B parameter models from the Qwen3.5 family (Qwen Team, 2026). The impact of AWQ quantization on output quality is also evaluated.

Prototype: Development and Testing

Developed solution utilises three pipelines (Figure 1) to ingest and transform data atomically, from discovery through transcript ingestion and eventual summarisation.

30 transcripts were sampled, from project related

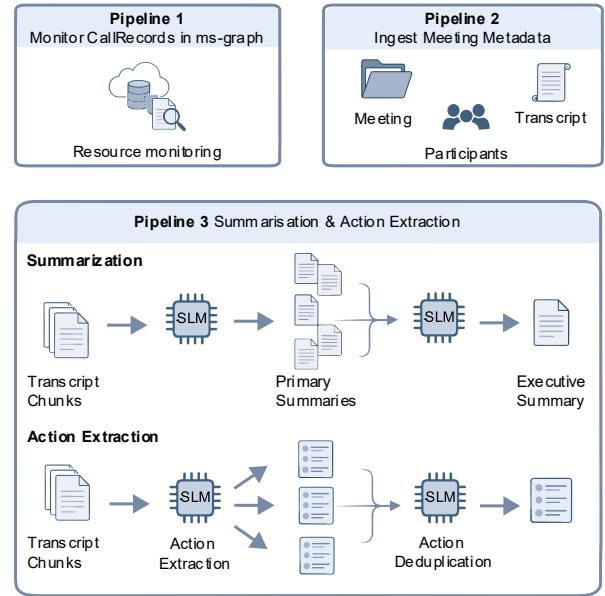


Figure 1: Architectural overview. The first two stages are concerned with data discovery and ingestion, with the third focused on processing transcripts into actions and summaries meetings in an architectural practice, to form the evaluation dataset. They vary from 2000-30000 tokens and are stratified into three groups of 10 by token length at boundaries 8000 and 15000 tokens. The summarisation process is evaluated via LLM as a Judge (Zheng et al., 2023) in which a Qwen3.6-27B (Qwen Team et al. 2026) is tasked with scoring the summarisations on a scale of 0-5 (Li et al., 2026) on two categories: Faithfulness, and Completeness (Zheng et al., 2023). These metrics correspond to tendency to hallucinate, and inclusion of all available facts, respectively. For action extraction, the judge model produces its own actions and compares them to the actions of the candidate models. Finally, F1 score was calculated to represent agreement between the candidate and the judge model.

Result and Analysis: Action extraction F1 score (Figure 2) across all conditions was consistently above 0.85, indicating that the SLMs and the Judge were mostly in agreement. Though, the quantized models show larger variation in standard error. Upon inspection, this was caused by the AWQ models occasionally failing completely and generating zero actions, upsetting the

mean and reported standard error. In practice, a system can detect and retry under such circumstances. Scores were broadly flat across meeting lengths, validating the small window, high overlap and dedupe approach.

Faithfulness (Figure 2) for immediate summarizations of transcript windows are consistently strong across models with quantization having a minimal impact. Executive summaries exhibit a steep decline, likely due to the high compression and propagation of hallucinations from tier

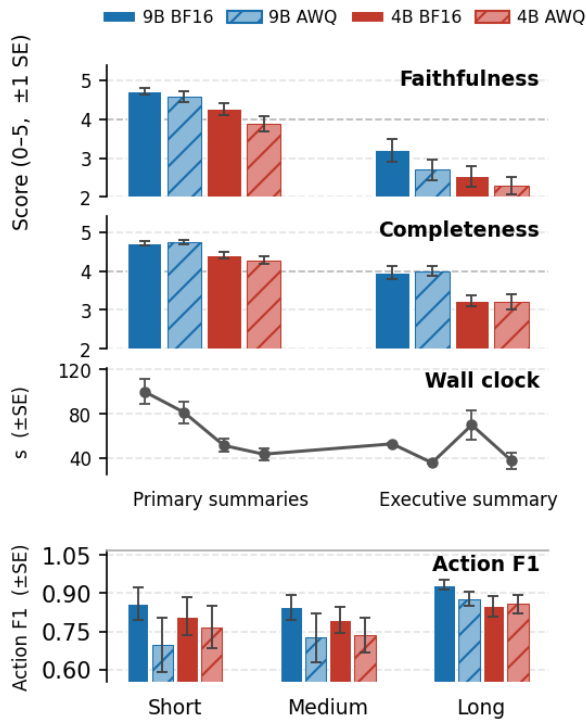


Figure 2: LLM Judge scores on Faithfulness and Completeness, Action F1 for 9B/4B variants of Qwen3.5 in BF16 and AWQ INT4. Wall clock time represents the mean average time taken for completion of a summary

one summarizations.

The 9B parameter models clearly outperform the 4B parameter models in Completeness (Figure 2) indicating that they tend to preserve more of the available information from the transcript.

Wall clock time (Figure 2) decreases steadily with reduction in model size, this is expected as LLM decode is memory bandwidth bound, smaller models are faster. However, for executive summaries, the 4B models break that trend, taking more time to produce summaries. This is because the smaller models are consistently producing longer outputs at that stage. Alongside a decline in Faithfulness, and a steeper decline in Completeness, this indicates that the outputs are not only longer, but less valuable.

Conclusion

This work presented a locally deployed LLM based summarisation system for architectural meetings, offering a privacy-conscious alternative to cloud-based

documentation tools. The 9B model in quantization (INT4 AWQ) was identified as the optimal configuration, delivering acceptable summarisation quality with significant decode speedup over the BF16 variant, while producing more concise and consistent outputs than the 4B model. These results demonstrate the feasibility of locally deployed LLMs for supporting privacy-conscious architectural workflows and reducing manual meeting documentation overhead. Although, Meetings organised by external tenants cannot be automatically accessed through MS Graph due to tenant ownership restrictions, and users within meeting rooms cannot be directly identified. Additionally, the asynchronous architecture introduces eventual consistency delays. Future work will explore broader evaluation strategies using both open-source and frontier judge models to reduce model-family bias, alongside extending experiments to next-generation SLM families such as Gemma4.

Database and Source Code Availability: All source code, datasets, and supplementary implementation details related to this work are publicly available at the following [repository](#).

Acknowledgments

Innovate UK (Grant 10113484) for facilitation of a Knowledge Transfer Partnership (KTP)

References Qwen Team. (2026) Qwen3.5: Towards Native Multimodal Agents, Available at: <https://qwen.ai/blog?id=qwen3.5>

Qwen Team. (2026) Qwen3.6-27B: Flagship-Level Coding in a {27B} Dense Model, Available at: <https://qwen.ai/blog?id=qwen3.6-27b>

Ghimire, P., Kim, K. & Acharya, M. (2024)

Opportunities and Challenges of Generative AI in Construction Industry: Focusing on Adoption of Text-Based Models. Buildings, 14(1), p.220.

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z. & Cheng, X. (2024) On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. arXiv preprint arXiv:2403.05156

Li, W., Zhao, M., Dong, W., Cai, J., Wei, Y. et al. (2026) Grading Scale Impact on LLM-as-a-Judge: Human-LLM Alignment Is Highest on 0-5 Grading Scale. arXiv preprint arXiv:2601.03444

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E. & Stoica, I. (2023) Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 3