Web3 Journal: ML in Health Science
ISSN 2942-8726
2025. Vol. 2(1)
Short Report
DOI: 10.62487/5a1m5ara

Architecture and Workload as Primary Sources of Error in RAG and Agentic AI Systems: Summary of Two Years of SAIMSARA Development

SAIMSARA core team [1]

Keywords: RAG, Agentic AI, LLM, AI Engineering, SAIMSARA, AI Architecture

Most Retrieval-Augmented Generation (RAG) and AI agent errors are not Large Language Model (LLM) model failures — they are architecture and workload failures.

This editorial short report summarizes the most common pitfalls and outlines practical mitigation strategies based on two years of developing SAIMSARA, a Systematic, AI-powered Medical Scientific Article Review Agent (saimsara.com).

Common Architectural Pitfalls

1. **Prompt Overload**: Prompts that combine many rules, constraints, and strict formatting requirements consume a disproportionate share of the model's attention, leaving insufficient capacity for reliable data processing.
2. **Excessive Batch Size**: Large batches amplify error probability through cumulative effects, including skipped items, cross-item interference, and degradation toward the end of long sequences.
3. **Oversized Input Items**: Long text passages, dense token sequences, or high-resolution images increase per-item processing cost and reduce overall system stability.
4. **Speed-Optimized Model Selection**: LLM optimized primarily for throughput often lack the step-by-step discipline required for multi-stage reasoning tasks, leading to skipped reasoning steps and structural output errors.

Practical Mitigation Strategies

Error rates can be significantly reduced by aligning system workload with model capacity:

- **Use a stronger or more deliberative LLM** for complex, multi-step tasks
- **Simplify prompts** by reducing rules, exceptions, and implicit assumptions
- **Reduce batch size** to limit cumulative cognitive LLM load
- **Decrease item size** (characters, tokens, pixels) to improve per-item reliability

Editorial Conclusion

These interventions do not eliminate errors entirely, but they move agentic AI systems into a stable operating regime. The key insight is that robustness in RAG and agentic AI is primarily a systems engineering challenge, not a model selection problem.

As these systems scale, architectural discipline—rather than marginal gains in model performance—will determine reliability and reproducibility in real-world applications.

No Conflict of Interest

[1]SAIMSARA core team
Email: admin@saimsara.com