4most
ANALYTICS CONSULTING

Getting the GenAI balance right

Varying approaches to Retrieval Augmented Generation (RAG) system validation depending on scale, complexity and sensitivity of your application

James Barlow, Lucy Worlsey, and Gordon Baggott

August 2025

Generative AI

4most
ANALYTICS CONSULTING

Use Cases: Is GenAI a good solution to my problem?

What's the problem landscape

- Start with the problem
- Does it need GenAI
- Assess fit
- Consider alternatives

Capability and users

- Assess internal capabilities
- Advanced GenAI development
- User integration

Tech, Compliance and Control

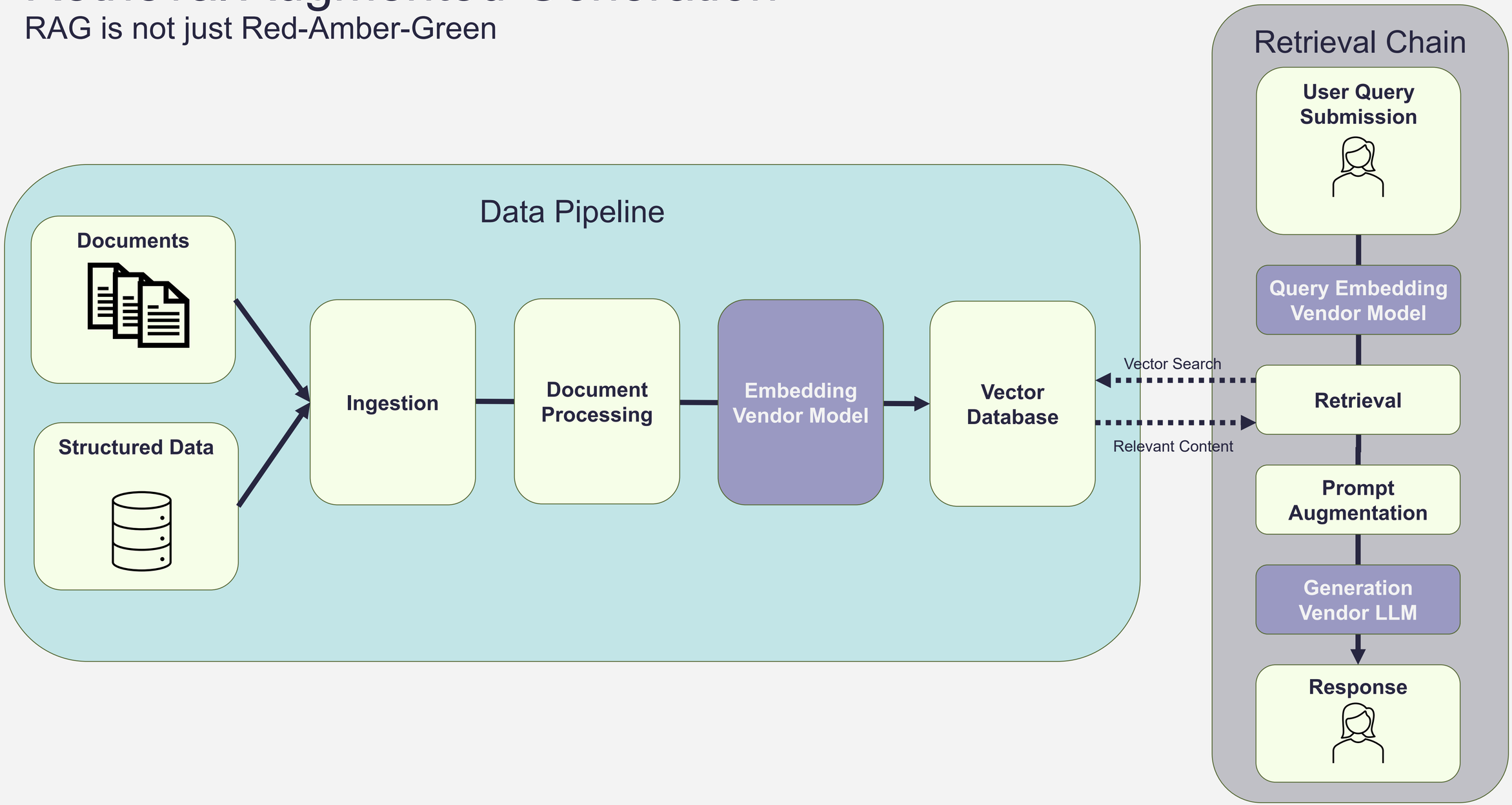
- Infrastructure & Governance
- Vendor & Model Assurance
- Regulatory & Policy Alignment
- Risk Controls & Guardrails

End 2 End Value

- Upfront & Ongoing Costs
- Scaling with Usage
- Cost Dynamics

Retrieval Augmented Generation

RAG is not just Red-Amber-Green



Validation principles are qualitative and quantitative

Visibility

Invisible and unknown model risk

Is anything known about performance?
How are organisational decisions made to onboard?

Relevance

Task misalignment

Is the system design aligned with the use case?
How are prompts engineered and controlled?

Reliability

System failure

How robust is the implementation?
Is the system operationally suitable and backed up?

Security

Data and information loss

How is access approved for usage and document store?
What are the controls on internal information loss?

Performance

Compromised outputs

Does the system respond with an accurate response?
Does the system respond using the correct set of facts?

Trust

Tool/vendor/assumption error

Has the vendor disclosed a suitable level of detail?
Why can the system be trusted?

Qualitative validation aspects

4most ANALYTICS CONSULTING

Language
Could inappropriate outputs arise?

Cost
Are the benefits justified?

Ethics
Are users treated fairly?

Latency
Is the process efficient?

Effective Challenge

Compliance
Full compliance evaluation?

Usefulness
Outputs having positive results?

SME Knowledge
Stakeholders best placed?

User design
Tailored to users?

Driving effectiveness in validation techniques

Evaluation Approaches

- **Standardised Q&A Benchmark**
- **Bespoke Q&A Benchmark**
- **LLM Evaluation**
- **Human Evaluation**

Evaluation Metrics and Remediation

Retrieval and Prompt Tuning

Retrieval Accuracy
- Use of well-known metrics

Ranking and Relevance
- Assess semantic relevance of retrieved chunks

Structure and Robustness
- Ensure prompts are clear
- Guard against prompt injection (user manipulation of prompts)

Chunking and Indexing

Chunking & Embedding
- Use fit-for-purpose chunking aligned with content.
- Validate embeddings

Indexing
- Minimise latency
- Audit vector store.

Data Quality & Ground Truth Improvement

Metadata Filtering
- Precise metadata filtering is key to ensuring accurate input.

Ground Truth Quality
- Ground truths must be accurate and consistently meet high standards during development.

Improvements

Metrics
- Advanced metrics can enhance evaluation.

Ground Truth
- Expanding ground truth to include edge cases.

Humans
- Human-in-the-loop feedback is crucial for aligning the system with real-world use.

Summary and Conclusions

A structured four-stage assessment helps determine the appropriateness of using GenAI	• Assessing the problem landscape • Capabilities and users • Tech, compliance and control • End 2 End value
Core validation principles remain unchanged for GenAI systems	• Understand the input data, model architecture, and alignment with intended business use.
System complexity does not necessarily require complex validation	• Large-scale, low-materiality applications may not warrant the same level of scrutiny as high-materiality use cases.
Validation should be proportionate to risk	• Higher materiality and sensitivity require more robust validation. • Validation complexity should reflect the model's risk profile, not merely its technical complexity.
Evaluation can use traditional or contemporary methods	• Traditional metrics can be used to for evaluation of a RAG system • More advanced metrics can be introduced to improve use case specific validation

Thank you.

Lucy Worsley
Associate Partner

Linkedin:

James Barlow
Principal Consultant

Linkedin: