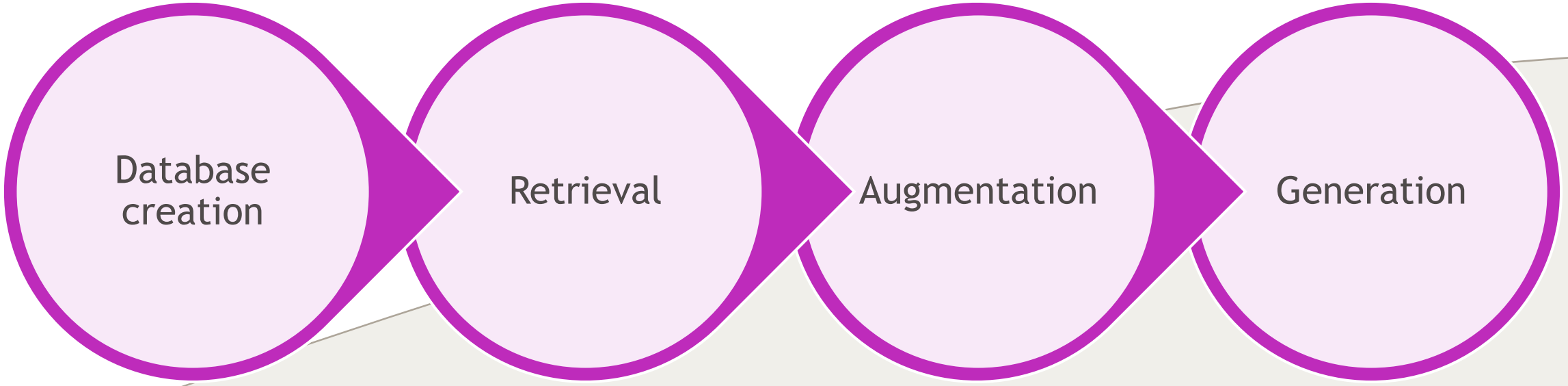
Sven Klijn
ISPOR Glasgow, November 10th, 2025

The essentials of RAG

A tale of transparency and traceability

Main steps in RAG



Pre-production

In production

Application in HEOR

- RAG is recommended to provide LLMs with relevant domain-specific information, e.g., information on HEOR-relevant R libraries
- Examples of when this may be especially relevant is when data are:
  - Proprietary
  - Very recent
  - Part of a niche domain
- Applied examples:
  - Writing a dossier based on curated in-house data
  - Conducting analyses while adhering to international guidelines
  - Summarization of external landscape

Perspectives on RAG

Retrieval-augmented generation (RAG) should be used to enhance transparency and traceability of GenAI outputs by grounding responses in verifiable external sources

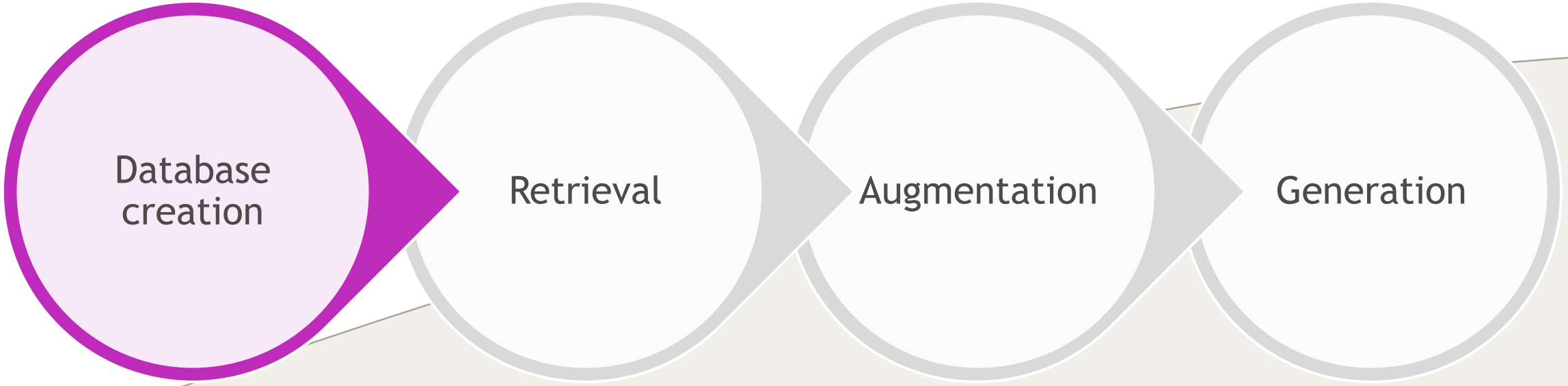
Principle 6: Ensure GenAI-assisted analyses are reproducible and auditable where possible

Additionally, RAG and improving the instructions used in prompts provided to models (prompt engineering) can reduce hallucinations and bias

Principle 12: Assess and mitigate bias to ensure fairness and representativeness

Source: NICE (October 2025). Appendix B: Generative Artificial Intelligence (GenAI) best practice principles

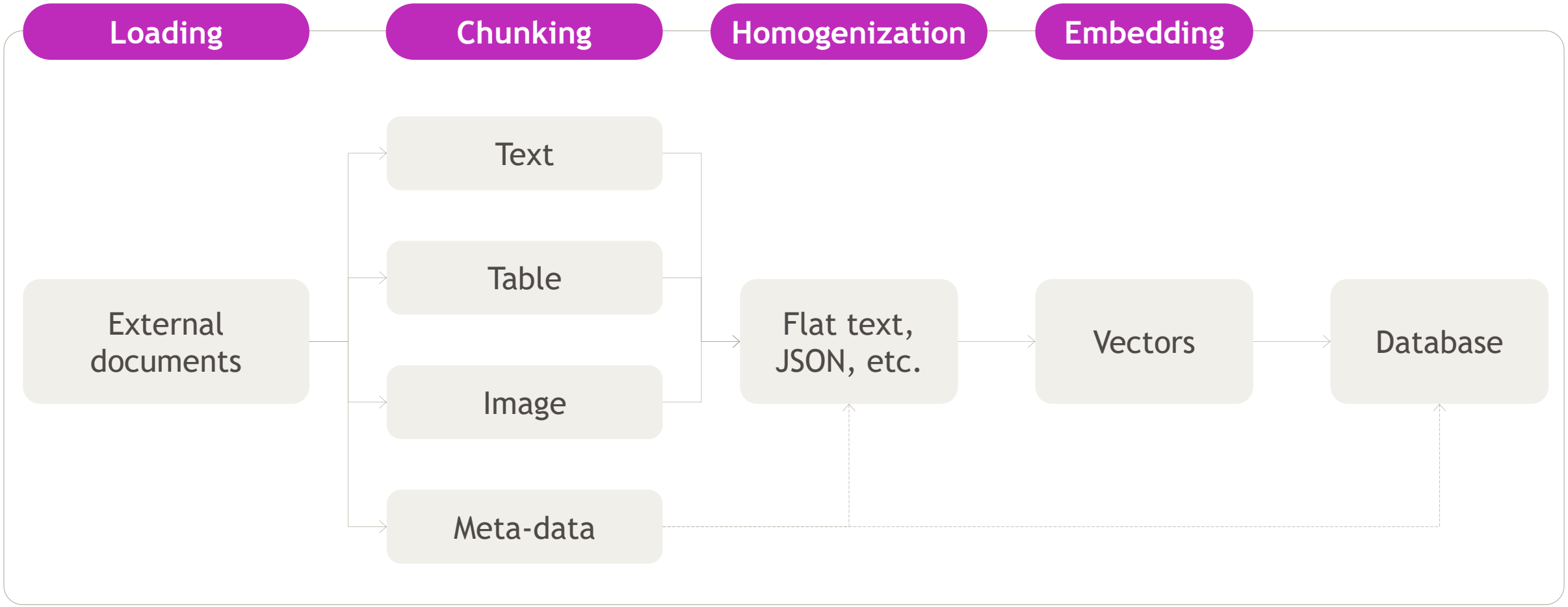
Main steps in RAG



Pre-production

In production

Pre-production: Database creation



JSON: JavaScript Object Notation

Example of a document after loading, chunking, homogenization, and embedding

Loading

in selecting relevant domains and briefly justify any exclusions, allowing flexibility for diverse and evolving HEOR use cases.

To support interpretation, each domain is assigned a maturity level reflecting the current availability of established metrics or reporting standards. High-maturity domains have well-defined practices, whereas low-maturity ones indicate evolving methods. These expert-assessed ratings within the ISPOR Working Group on Generative AI are a pilot feature and will be revisited in future validation. Table 2 outlines the 10 reporting domains and their definitions.

Model Characteristics

This domain focuses on documenting the foundational attributes of the LLM used in the study. Key elements include the model's name (eg, LLaMA-3), version, developer or organization, release date, license type (eg, commercial or open source), and access method (eg, application programming interface [API], web

Accuracy Assessment

This domain evaluates how well Gen AI-generated outputs align with correct or expected results. Accuracy can be assessed through comparisons with human benchmarks, gold-standard data sets, or expert review. Metrics may include commonly used measures in AI/ML, such as precision, recall, F1 score, and area under the curve, as well as natural-language-processing-specific (eg, BLEU) or domain-specific metrics (eg, Generative Radiology Report Evaluation and Error Notation for radiology report generation).[65] In HEOR, appropriate methods include fact checking against source documents, expert review, or benchmarking against known evidence, but the suitability of accuracy metrics depends on the task. Structured tasks, such as data extraction or classification, lend themselves to quantitative metrics, whereas free-text generation, such as drafting an HTA dossier, often requires qualitative assessment. Although interest is growing in

Chunking

chunk_id	preview
Fleurence_2025_ELEVATE-GenAI-sec-c1016100-01	Model Characteristics This domain focuses on documenting the foundational attributes of the LLM used in the study. Key elements include
Fleurence_2025_ELEVATE-GenAI-sec-aad9d23f-02	for describing model provenance, architecture, and access, although transparency about training data remains limited in some proprietary
Fleurence_2025_ELEVATE-GenAI-sec-46d56d1d-03	measures for HEOR tasks, such as SLRs and HEM, and in developing HEOR-specific benchmarks, further work is needed to define fit-for-purpose

Homogenization

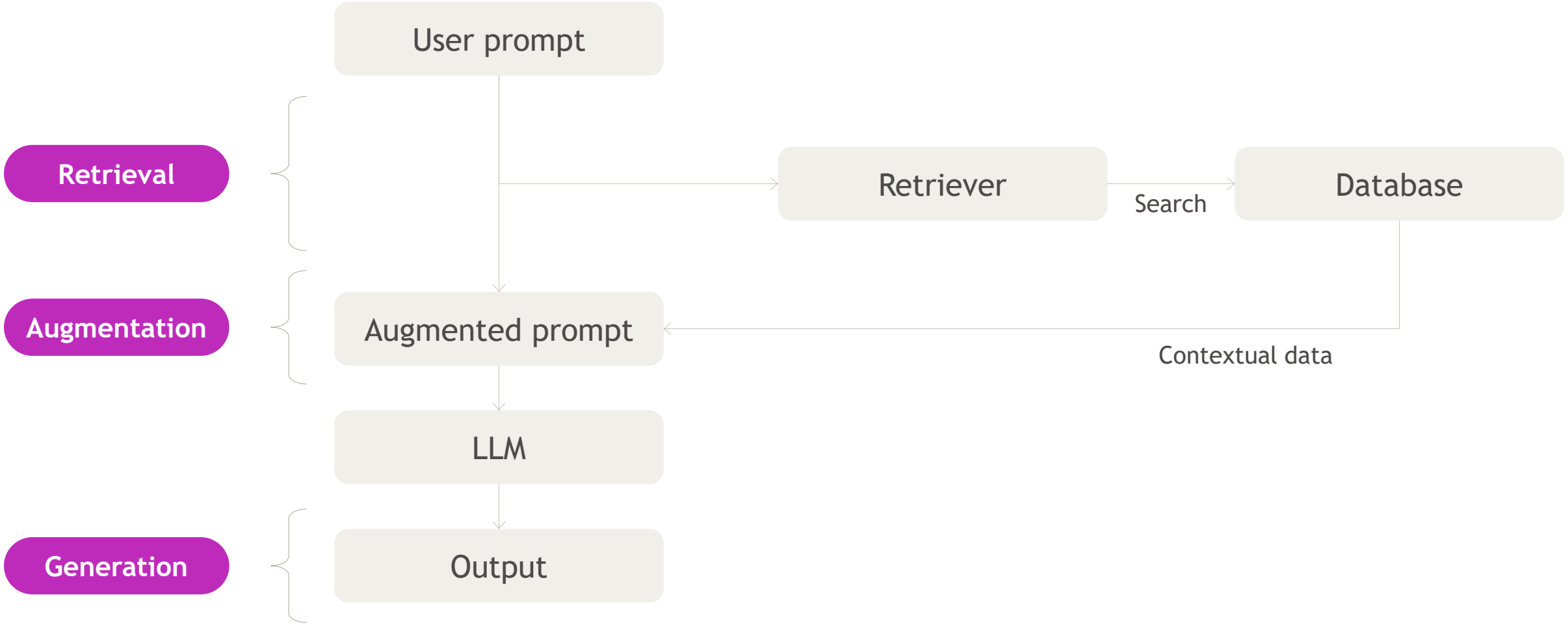

```
{"chunk_id": "Fleurence_2025_ELEVATE-GenAI-sec-c1016100-01",
 "doc_id": "Fleurence_2025_ELEVATE-GenAI",
 "source_file": "Fleurence_2025 - ELEVATE-GenAI.pdf",
 "pages": [3, 4],
 "section": "Model Characteristics & Accuracy Assessment",
 "start_word_index": 0,
 "end_word_index": 180,
 "text": "Model Characteristics This domain focuses on documenting the foundational attributes ofthe LLM used in the study. Key elements
        include the model's name (eg, LLaMA-3), version, developer or organization, release date, license type (eg, commercial or open
        source), and access method (eg, application programming interface [API], web interface, or local deployment). Authors should also
        report the model's architecture (eg, transformer based) and provide details about training data sources, where applicable. This
        includes general-purpose pretraining corpora (where identifiable), data sets used for fine-tuning or instruction-tuning, any
        proprietary data used for custom models, and any sources integrated into retrieval-augmented generation workflows. Where applicable,
        authors are encouraged to discuss the explainability of the model's outputs, particularly in relation to interpreting findings
        in HEOR contexts. Although explainability is not designated as a standalone domain in ELEVATE-GenAI, it remains an important
        consideration for transparency, reproducibility and stakeholder trust. Level of Maturity: High Well-established practices exist
        for describing model provenance, architecture, and access, although transparency about training data remains limited in some
        proprietary models. Accuracy Assessment This domain evaluates how well Gen AI-generated outputs align with",
   "metadata": {"chunk_index": 0,
                "chunking": {"method": "word_window",
                             "size": 180.
```


Embedding


```
{"chunk_id": "Fleurence_2025_ELEVATE-GenAI-sec-c1016100-01"
 "embedding_model": "tfidf(ngram=1..2,max_features=512)",
 "vector": [0.07585591314453734, 0.07585591314453734, 0.07585591314453734, 0.03792795657226867, 0.04987070497336715,
            0.029454441731919773, 0.03792795657226867, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.029454441731919773,
            0.03792795657226867, 0.0, 0.03792795657226867, 0.03792795657226867, 0.04987070497336715, 0.04987070497336715,
            0.07585591314453734, 0.04987070497336715, 0.0, 0.03792795657226867, 0.03792795657226867, 0.0, 0.04987070497336715,
            0.14727220865959886, 0.07585591314453734, 0.04987070497336715, 0.0, 0.0, 0.0, 0.0, 0.0, 0.04987070497336715,
            0.04987070497336715, 0.0997414099467343, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715,
            0.04987070497336715, 0.0997414099467343, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715,
            0.04987070497336715, 0.0, 0.0, 0.07585591314453734, 0.03792795657226867, 0.04987070497336715, 0.03792795657226867,
            0.04987070497336715, 0.0, 0.0, 0.0, 0.029454441731919773, 0.0, 0.0, 0.0, 0.0, 0.0, 0.04987070497336715, 0.0, 0.0,
            0.0, 0.03792795657226867, 0.0, 0.03792795657226867, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715,
            0.04987070497336715, 0.0997414099467343, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715,
            0.04987070497336715, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.04987070497336715,
            0.04987070497336715, 0.029454441731919773, 0.04987070497336715, 0.0, 0.0, 0.0, 0.04987070497336715,
            0.04987070497336715, 0.0, 0.0, 0.0, 0.0, 0.04987070497336715, 0.04987070497336715, 0.15171182628907467, 0.0,
            0.03792795657226867, 0.03792795657226867, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715,
            0.04987070497336715, 0.0, 0.0, 0.0, 0.0, 0.04987070497336715, 0.04987070497336715, 0.03792795657226867,
            0.03792795657226867, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715, 0.04987070497336715, 0.0, 0.0
```


Source of example: Fleurence et al. (2025). ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report. DOI: 10.1016/j.jval.2025.06.018

In production:
Retrieval, augmentation and generation



LLM: large language model

Traceability

The **first three reporting domains** in the *ELEVATE-GenAI framework* are:

1. **Model Characteristics** —
   This domain documents the foundational details of the large language model (LLM) used, including its name, version, developer, architecture, release date, license type, and data sources for pre-training or fine-tuning. It also covers access methods (API, web interface, or local deployment) and encourages discussion of explainability in the HEOR context.
2. **Accuracy Assessment** —
   This domain evaluates how closely the model's outputs align with correct or expected results. It includes comparing outputs against human benchmarks, gold-standard datasets, or expert reviews using metrics such as precision, recall, F1 score, or domain-specific measures.
3. **Comprehensiveness Assessment** —
   This domain examines whether GenAI-generated outputs fully and coherently address all required elements of a task. In HEOR, it ensures all relevant studies, model components, and outcomes are captured and described, with expert review confirming that nothing critical has been omitted.

Source: Fleurence R.L. et al. *ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research. Value in Health*, 2025.

Simple referencing

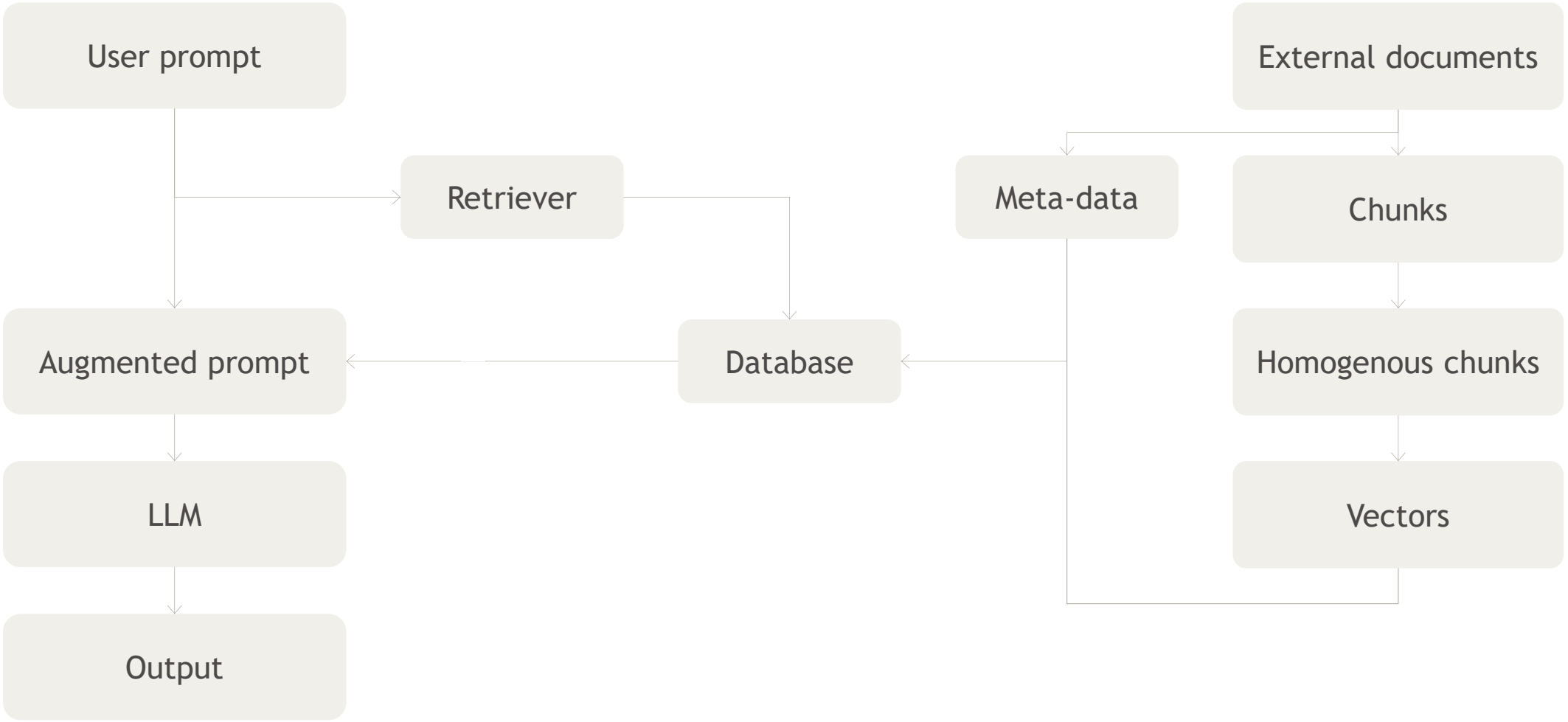
2. **Accuracy Assessment** —
   This domain assesses how well the GenAI-generated outputs align with correct or expected results. It recommends validation against human benchmarks, gold-standard datasets, or expert review, using metrics such as precision, recall, F1 score, and area under the curve (AUC). The report notes that in health economics and outcomes research (HEOR), methods like fact-checking against source documents or benchmarking against known evidence are appropriate.
   (See: Fleurence et al., *ELEVATE-GenAI Report*, p. 3–4, Section "Accuracy Assessment," Table 2)
3. **Comprehensiveness Assessment** —
   This domain evaluates whether the GenAI-generated outputs fully and coherently address all required elements of the assigned task. For HEOR studies, this includes ensuring that all relevant studies, model components, and outcomes are included and that outputs are compared against authoritative benchmarks such as published reviews or validated models. Expert evaluation is encouraged to confirm that no critical elements have been missed.
   (See: Fleurence et al., *ELEVATE-GenAI Report*, p. 4–5, Section "Comprehensiveness Assessment," Table 2)

Full Source:

Fleurence, R. L., Dawoud, D., Bian, J., Higashi, M. K., Wang, X., Xu, H., Chhatwal, J., Ayer, T. (2025). *ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research. Value in Health*, 2025, pp. 3–5.

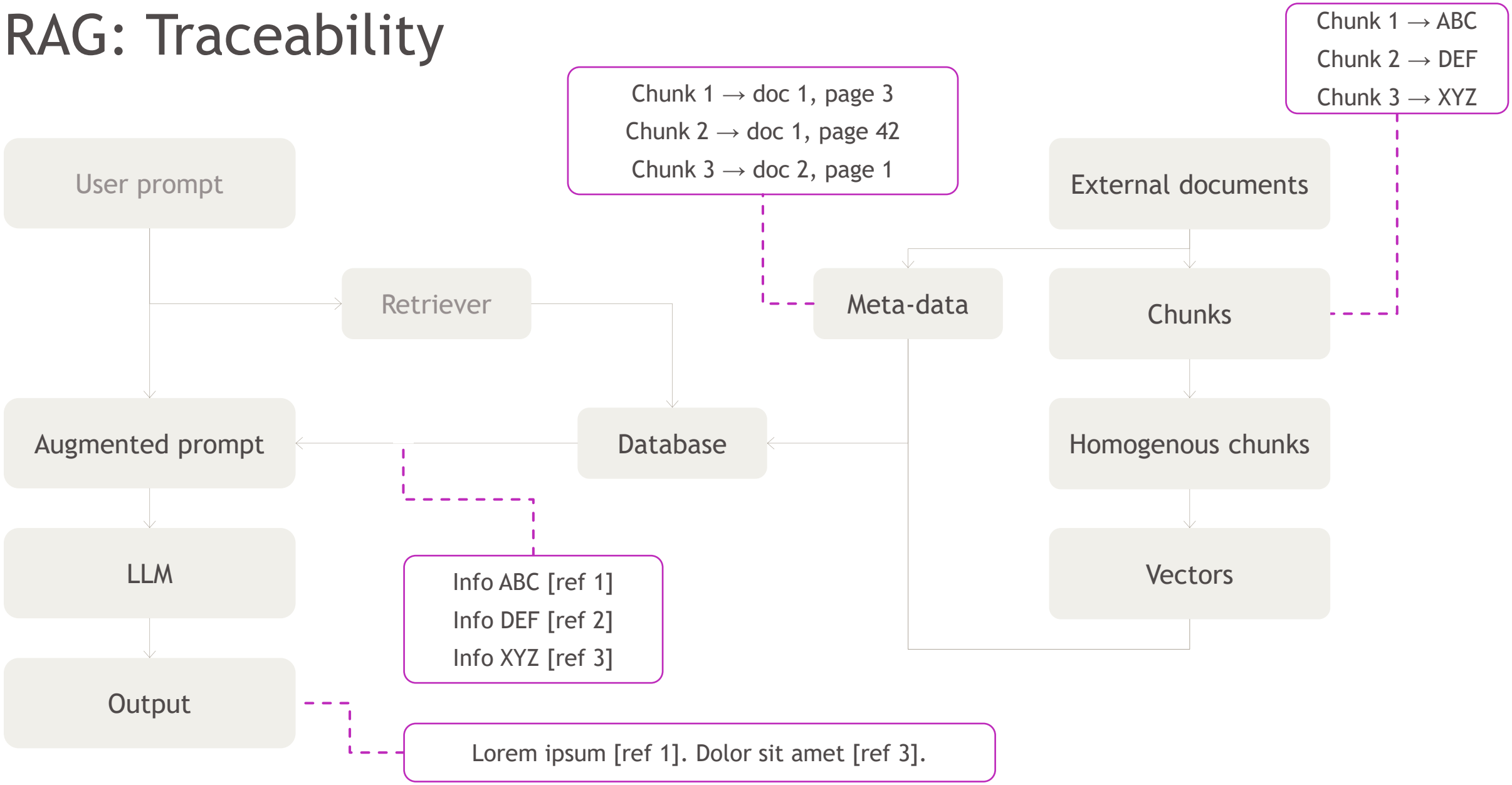
Detailed referencing

RAG: Full system



LLM: large language model

RAG: Traceability



LLM: large language model

Data curation

GenAI solution	Characteristics	Limitations
Vanilla LLM	Only includes information from its original training dataset	• Information may be incorrect or outdated • Limited to no traceability • Risk of hallucinations
Prompt + context	Insert relevant contextual information in the prompt	• Contextual information limited by token window • LLM performance with extremely long prompts may be suboptimal
Web-connected chatbot / Deep Research	LLM combined with internet connectivity and/or tools	• Limited to no control over which sources are consulted • Only works with publicly available sources
RAG	LLM combined with vector database	• Requires careful design and validation • Information may be outdated if vector database is not maintained

LLM: large language model

Validating a RAG: What questions should we ask?

Relevant metrics include

01 Retrieval metrics, such as precision and mean reciprocal rank

02 Response metrics, such as faithfulness and sufficiency

03 System metrics, such as latency and efficiency

A RAG is not a RAG is not a RAG

- Careful evaluation of all RAG steps is essential

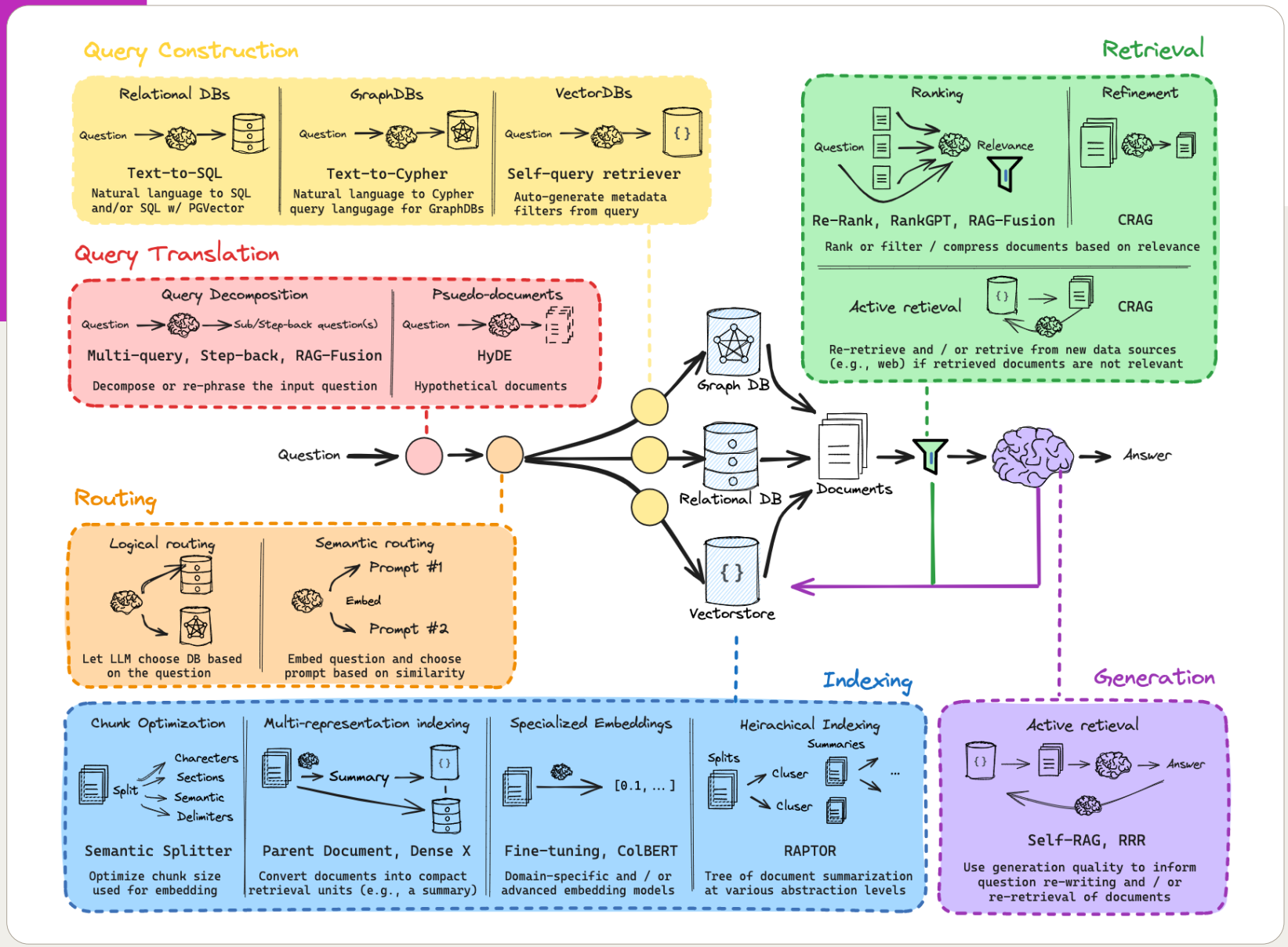
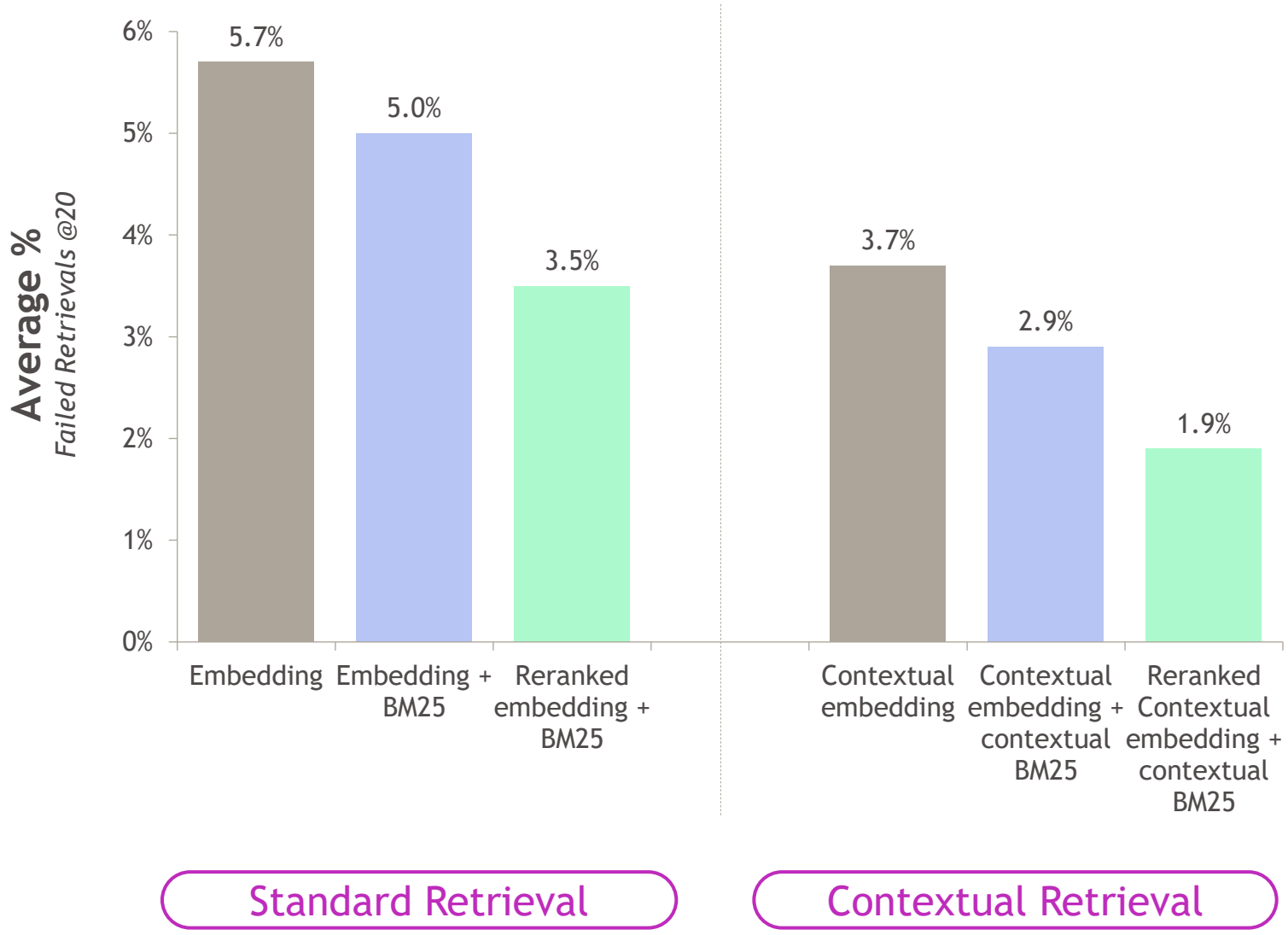


Image source: Ababou TH (2025), MIT License, https://github.com/bragai/brag-langchain

A RAG is not a RAG is not a RAG

- Careful evaluation of all RAG steps is essential
- The methods employed in each step of the RAG can have a large impact on performance[1]



Anthropic (2024). Introducing Contextual Retrieval. Accessed on March 20th, 2025, at: https://www.anthropic.com/news/contextual-retrieval