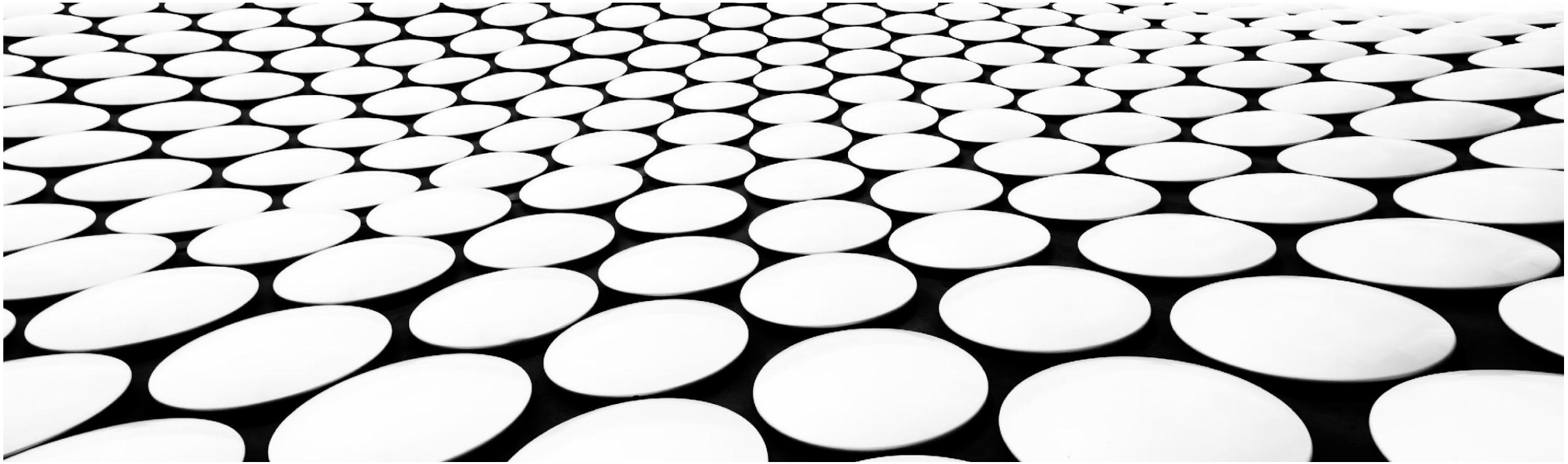
A CASE STUDY OF FINANCIAL CRIME PREVENTION THROUGH GENAI



DISCLAIMER

- All the views in this presentation by me are my personal and do not represent the views of the organization where I work.

FROM PREDICTIVE ML TO GEN-AI

Financial Crimes and Compliance risks keep evolving

Predictive Machine Learning learns patterns from historical data

GenAI has **potential** to assist in building clarity to complex alerts and regulations

Both must work together for stronger defenses

Validation ensures accuracy, explainability, and trust

GenAI's potential is real, but guardrails are critical

Like giving your guard dog night-vision goggles

THE TEMPTING POTENTIAL OF GENAI

GenAI is moving fast into pilot programs across the industry.

Use cases: chatboxes, onboarding, compliance reviews.

The potential is too tempting for banks and regulators to ignore.

But early pilots expose fragile points: hallucinations, retrieval gaps, prompt drift.

Without validation guardrails, exploration can turn into exposure.

It’s like hot cake, too tempting not to try, but check the ingredients first.

CASE STUDY: AML COMPLIANCE ASSISTANT

How such GenAI systems are built in practice

We will discuss an internal assistant for AML and compliance teams

It guides analysts on regulations, policies, memos, and procedures

Acts as a first-line explainer, not a decision-maker

Designed to improve efficiency and reduce interpretation risk

Like GPS for compliance, shows the route, but you are still driving!

BUSINESS PROBLEM

Analysts face an overwhelming flood of policies, memos, and regulations

Knowledge is scattered across systems, files, and teams

Institutional memory often lives in silos

The result: wasted time and inconsistent answers

Need: One assistant to answer, "*What does policy X require in scenario Y?*"

It's like looking for one sock in five different laundry baskets!

DESIGN CHOICES: RAG VS FINE-TUNING

RAG: pulls context from a document library in real time

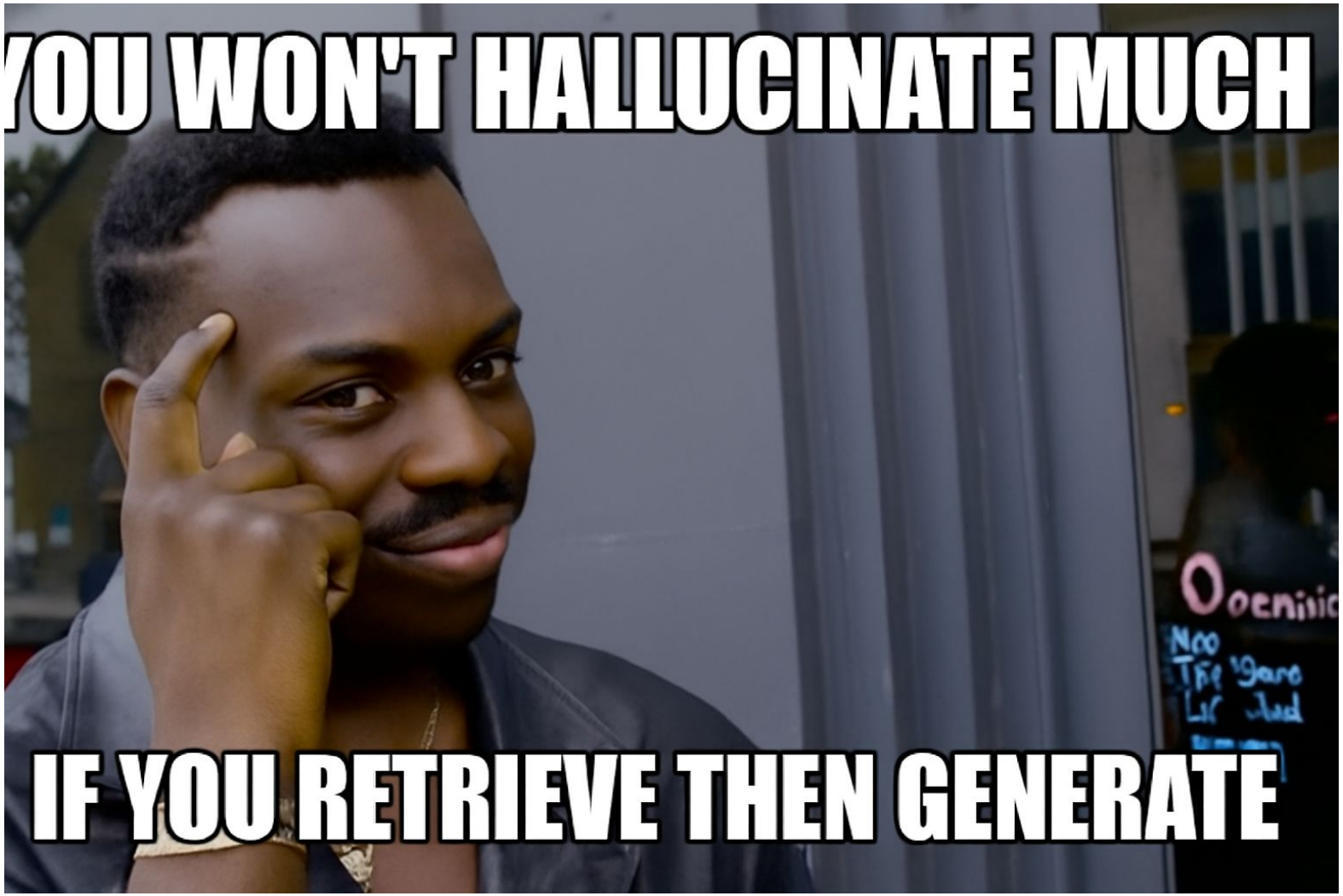
Fine-tuning: internalizes policies into the model's "brain"

Both can work, but each comes with tradeoffs.

Choice depends on update needs, transparency, and control

RAG is like checking a cookbook, fine-tuning is memorizing the recipes.

RAG ADVANTAGES



WHY CHOOSE RAG FOR AML COMPLIANCE

Tractable, modular, and explainable

Avoids hidden drift from fine-tuned weights

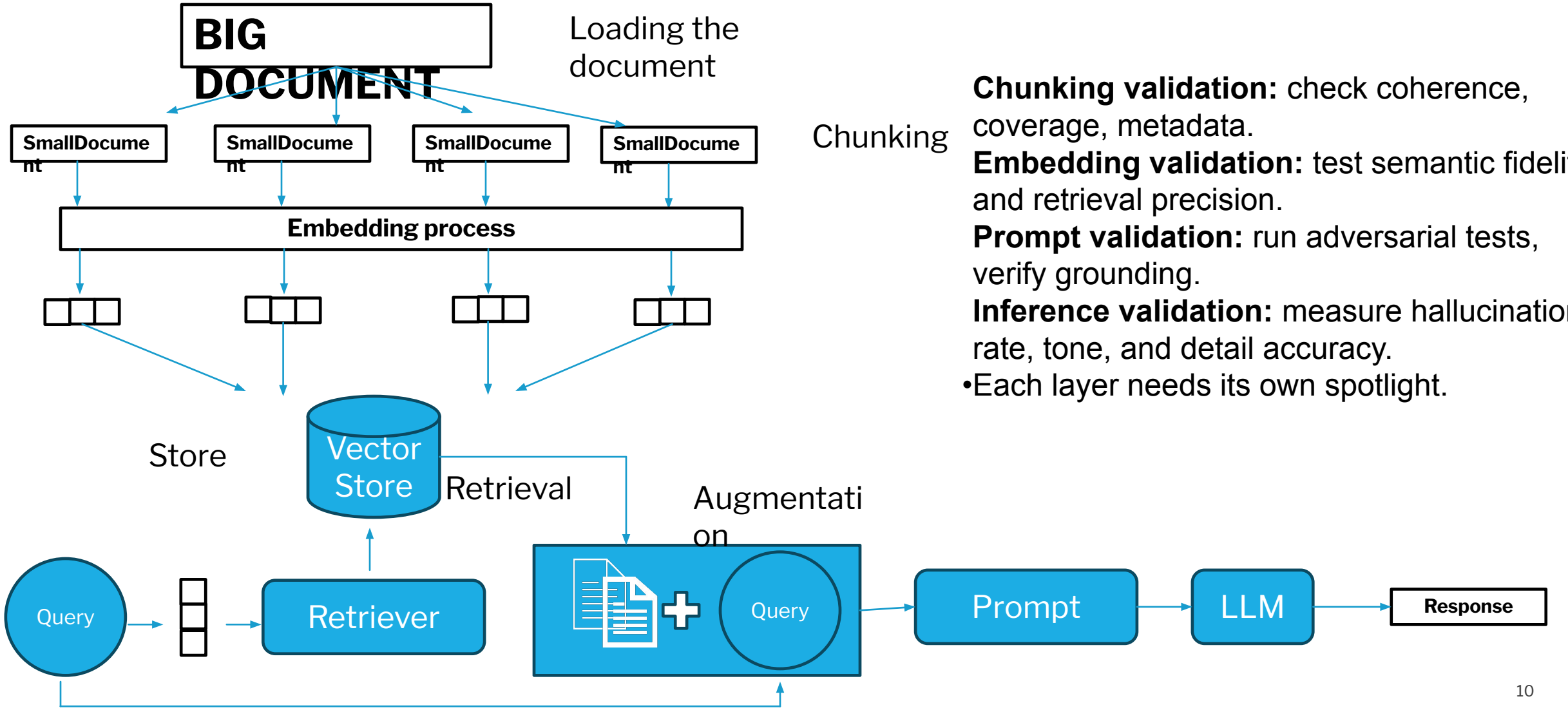
Provides citations to defend outputs

Easier to align with SR 11-7 guidelines

Build trust with both users and regulators

Like using Lego blocks, you can see every piece and swap them when needed.

A TYPICAL RAG PIPELINE



Chunking validation: check coherence, coverage, metadata.
Embedding validation: test semantic fidelity and retrieval precision.
Prompt validation: run adversarial tests, verify grounding.
Inference validation: measure hallucination rate, tone, and detail accuracy.

- Each layer needs its own spotlight.

VALIDATION FRAMEWORK (SR 11-7 ADAPTED)

Development and Use: document chunking, embeddings

Independent Validation: test retrieval accuracy, hallucinations, prompt sensitivity.

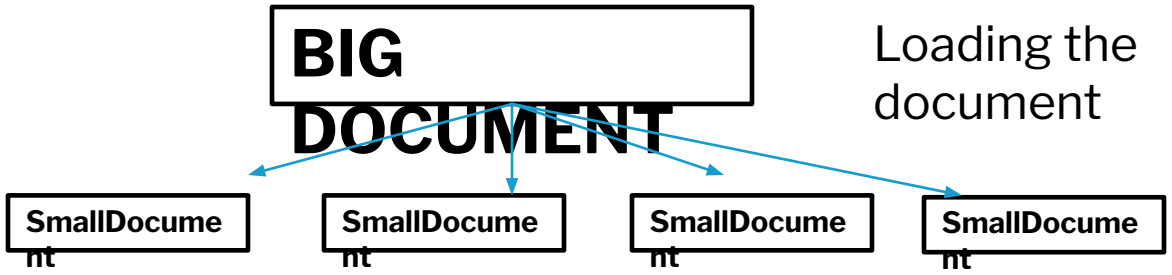
Ongoing Monitoring: track hallucination rate, drift, and degradation.

Outcomes Analysis: SME review of key risk indicators.

Governance: version control for prompts, embeddings, retrievers, APIs.

Like a pilot’s pre-flight checklist, skip one step and you’re in trouble.

LET'S GO A LITTLE GRANULAR



Chunking

- Breaks a big document into smaller, self-contained pieces.
- Each chunk should keep full meaning (no half sentences).
- Quality of chunking sets the ceiling for retrieval accuracy.

Chunk Size Variation

Too large chunks (risk: irrelevant text bundled in).
Too small chunks (risk: broken sentences, lost meaning).

Metadata Tags

Add labels like *Section*, *Page Number*, *Policy Type*.
Example: “Small Document A → [Doc: AML Manual, Sec 2.3, p.14]”.

Validation Checks

Add mini callouts:

Coherence: does this chunk make sense on its own?
Coverage: is every section of the original document included?
Integrity: does text exactly match the source (no OCR errors)?

Slice the bread too thick, it’s hard to chew. Slice it too thin, it falls apart.

CHUNKING GOING DEEPER

Challenges in Chunking

Overlapping vs. non-overlapping: overlap preserves context but increases storage.
Structural breaks: tables, lists, and figures often split poorly.
Context loss: splitting mid-sentence or mid-thought weakens meaning.
Document diversity: policies, memos, and emails each need tailored rules.
Scalability: large document sets require automated but reliable chunking.

Advanced Validation Checks

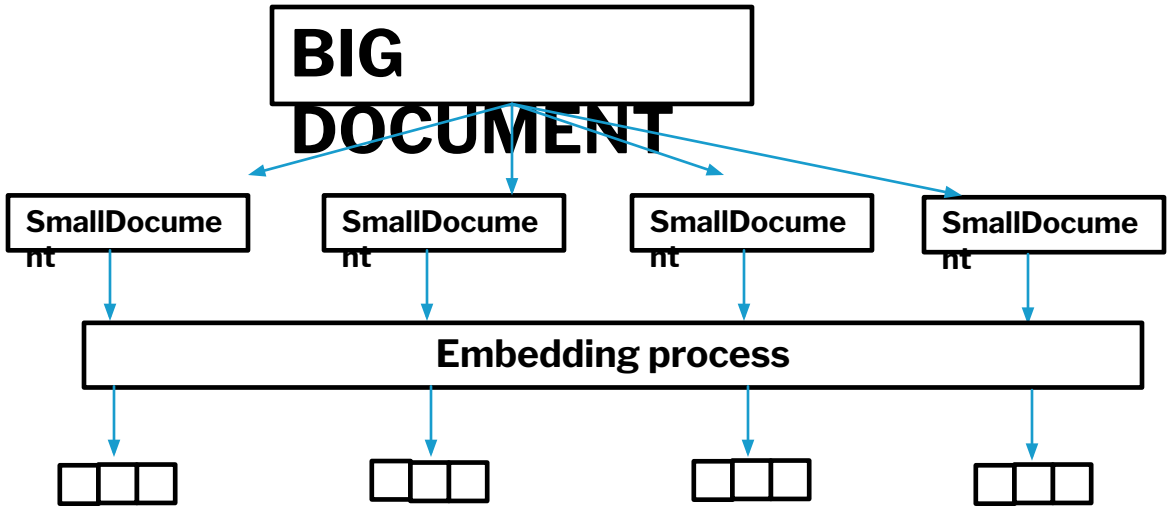
Coverage balance: ensure no section skipped, no duplication created.
Semantic continuity: verify logical flow across chunks.
Granularity testing: check if chunk size affects retrieval accuracy.
Stress testing: query spanning multiple chunks must still resolve.
Human review: spot-check chunks for readability and coherence.

Design Considerations

Smaller chunks = higher precision, lower recall.
Larger chunks = higher recall, risk of irrelevant noise.
Metadata tagging adds searchability and traceability.
Trade-off between efficiency and semantic fidelity.

Cut the pizza wrong, and one person gets all the cheese while another gets only crust.

LET'S GO A LITTLE GRANULAR



Embedding Process Explained

- Transforms each chunk into a numeric vector (its "fingerprint")
- Stores meaning, not just words → helps retrieval find the right text
- Key step: quality of embeddings drives answer accuracy

Validation Checks

Fidelity: does the embedding capture the right meaning in AML/regulatory context?

Precision: do similar queries retrieve truly relevant chunks?

Cross-check: test against a gold set of expert queries.

Like turning books into barcodes—you need the scanner to read them right

EMBEDDING GOING DEEPER

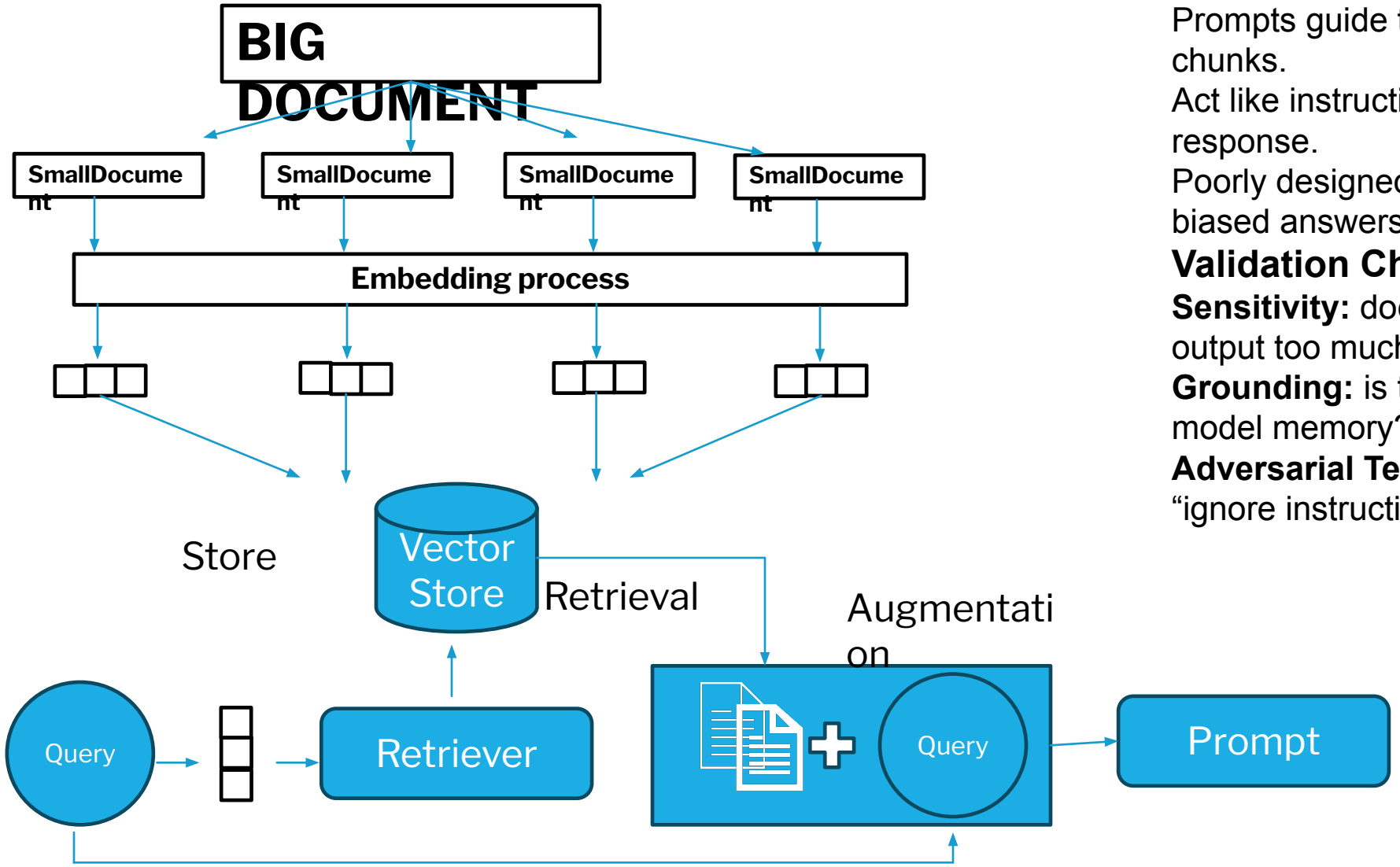
Challenges in Embeddings
Same word can have different meanings (polysemy).
Context loss when surface meaning overrides domain intent.
Bias risk from training data.
Domain shift: general embeddings often miss AML/legal nuances.
Vector stores grow fast, need efficient retrieval.

AML Example: "Placement"
AML stage: "The transaction shows signs of placement through small cash deposits."
General English: "The placement of the ATM was near the entrance."
HR context: "The candidate's placement in the compliance team is confirmed."

Advanced Validation Checks
Semantic fidelity: does "placement" link to AML stage, not HR?
Cluster testing: AML terms group correctly together.
Outlier detection: spot embeddings far from expected clusters.
Cross-model comparison: check consistency across embedding models.
Traceability: always link vector back to exact source text.

Like sorting puzzle pieces, if one belongs to another box, the whole picture breaks.

LET'S GO A LITTLE GRANULAR



Prompt Validation Explained

Prompts guide the LLM on how to use retrieved chunks.

Act like instructions or templates that shape the response.

Poorly designed prompts = vague, speculative, or biased answers.

Validation Checks

Sensitivity: does small wording change alter the output too much?

Grounding: is the answer based on retrieved text, not model memory?

Adversarial Testing: can the prompt resist tricks like "ignore instructions"?

Like giving directions, say it wrong, and you end up three blocks away.

PROMPT VALIDATION GOING DEEPER

Challenges in Prompts

Small wording changes can flip the answer.
Prompts may unintentionally bias the response.
Overly broad prompts lead to vague or speculative text.
Inconsistent formatting creates inconsistent outputs.
Risk of "prompt injection" (malicious instructions hidden in text).
Stress prompts: long queries, multiple clauses, or "what if" scenarios.

AML-Relevant Examples

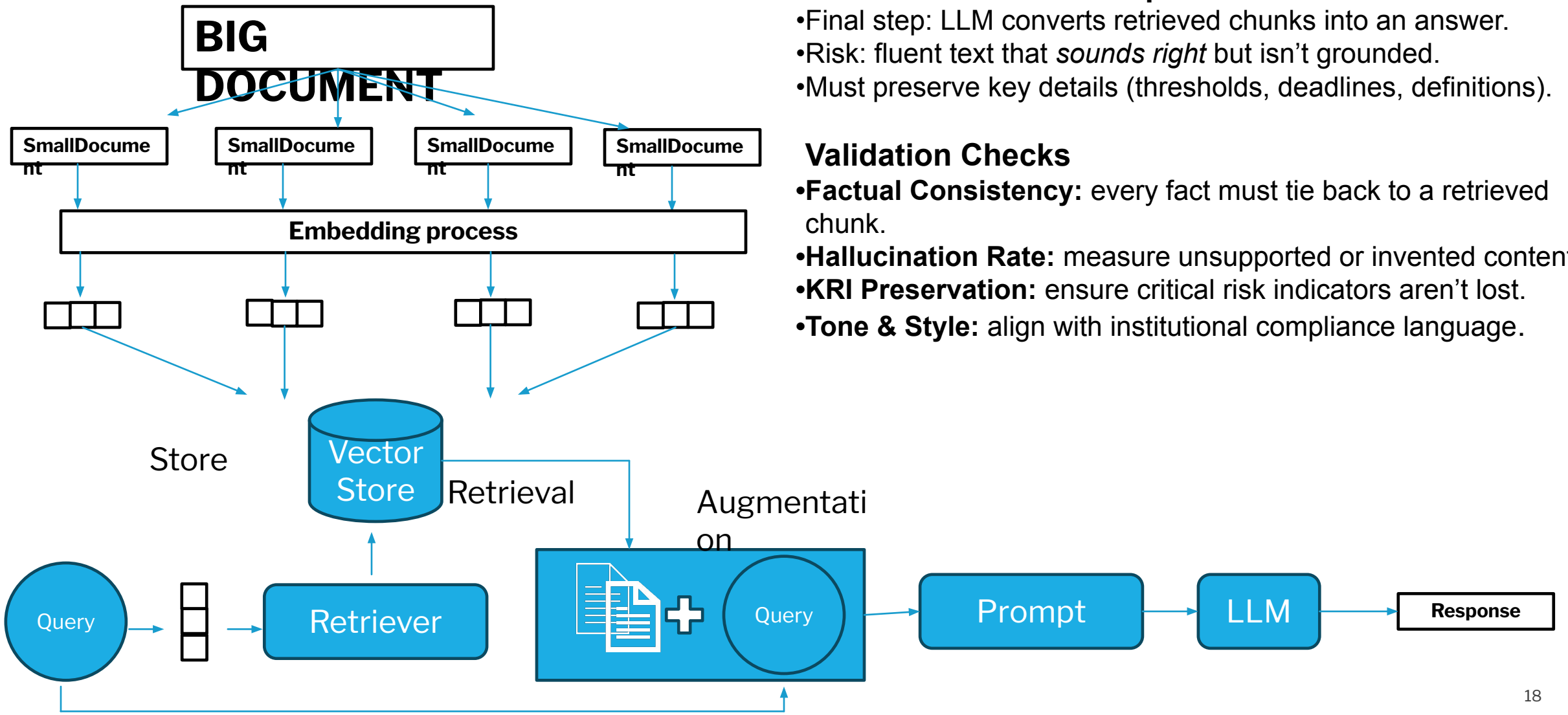
Asking: *"Explain AML filing timelines"* vs. *"What is the deadline for SAR?"* → must give the same answer.
Prompt must not generate fake regulatory citations.
Prompts should enforce escalation when uncertainty is high.

Advanced Validation Checks

Sensitivity tests: change wording slightly, see if the meaning stays.
Grounding checks: verify answer comes from retrieved chunks, not memory.
Adversarial tests: test prompts like *"ignore all rules"* or *"summarize but skip compliance."*
Role consistency: ensure assistant stays in compliance-explainer mode, not policy-maker mode.

Like typing the wrong letter in a phone number; one digit off and you reach a stranger.

LET'S GO A LITTLE GRANULAR



Inference Validation Explained

- Final step: LLM converts retrieved chunks into an answer.
- Risk: fluent text that *sounds right* but isn't grounded.
- Must preserve key details (thresholds, deadlines, definitions).

Validation Checks

- **Factual Consistency:** every fact must tie back to a retrieved chunk.
- **Hallucination Rate:** measure unsupported or invented content.
- **KRI Preservation:** ensure critical risk indicators aren't lost.
- **Tone & Style:** align with institutional compliance language.

INFERENCE VALIDATION GOING DEEPER

Challenges in Inference

LLM outputs are fluent but not always factual.
Risk of hallucinations — confident but unsupported text.
Critical details (thresholds, deadlines) can be altered or lost.
Style drift: casual tone instead of compliance tone.
Outputs may blend retrieved text with model memory.

AML-Relevant Examples

A $10,000 threshold might be restated incorrectly as $100,000.
Filing deadline “30 calendar days” could change to “30 business days.”
Assistant might skip escalation advice in high-risk cases.
Hallucinated citations: making up sections of AML laws or policies.

Advanced Validation Checks

Factual consistency: each statement must map to a retrieved chunk.
Hallucination rate: measure unsupported claims across test sets.
Critical indicator checks: thresholds, timelines, red flags preserved.
Tone validation: outputs must match compliance/regulatory style.
Comparative review: cross-check outputs with SME judgment.

Like a student writing a perfect essay, but on the wrong textbook!

KEY TAKEAWAY

Predictive ML models assess financial crime likelihood using patterns in historical data.

GenAI complements them by clarifying alerts and complex cases.

Provides context from policies, memos, and regulations.

Enhances understanding for investigators and regulators.

Validation ensures results are consistent, explainable, and defensible.

Together: detection + explanation = stronger financial crimes prevention.

Like having both a smoke alarm and a fire marshal: one detects, the other explains.

THANK YOU!

- Questions?