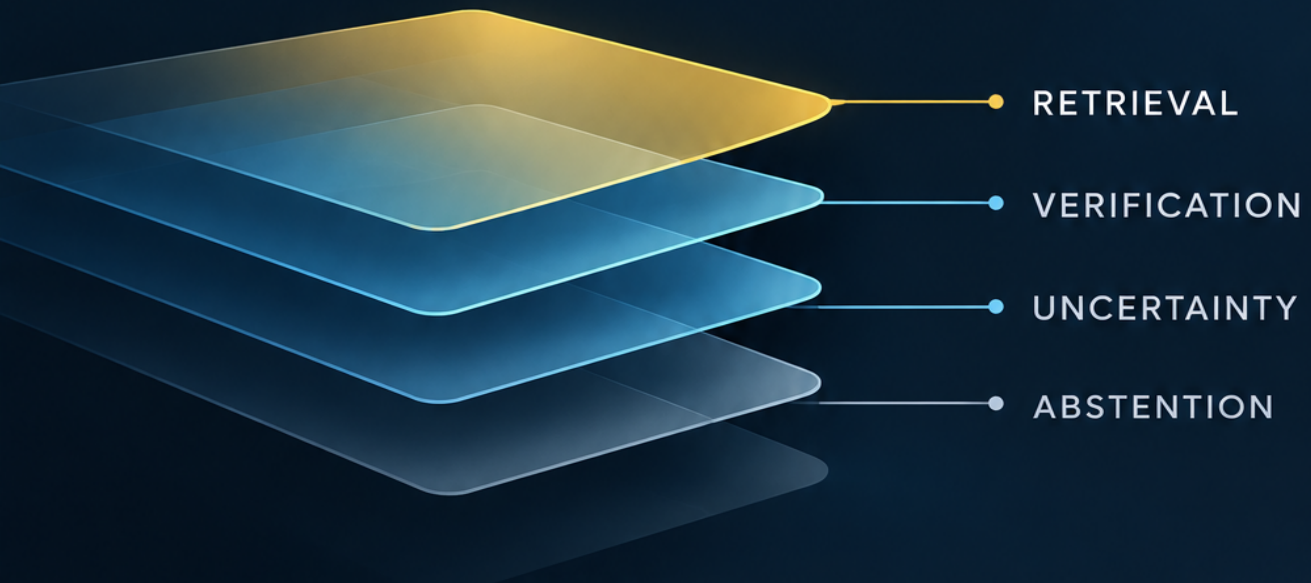
# Building Hallucination-Resistant RAG and AI Agents

## A Complete Guide to Reliable, Evidence-Grounded AI Systems




GARRETT LAWSON

# Building Hallucination-Resistant RAG and AI Agents

## A Complete Guide to Reliable, Evidence-Grounded AI Systems

Steve Publications

This book is available at
https://leanpub.com/buildinghallucination-resistantragandaiagents

This version was published on 2026-09-09


This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

# Contents

# A Complete Guide to Reliable, Evidence-Grounded AI Systems

Large language models hallucinate with startling confidence. In high-stakes domains like healthcare, law, finance, and operations, a single fabricated statistic or invented citation can cause real harm. This book shows you how to build retrieval-augmented generation (RAG) systems and AI agents that know when they have evidence, verify what they generate, detect when they are uncertain, and abstain rather than bluff when information is insufficient. We move beyond the myth of zero hallucinations and focus on what is actually achievable: layered, defense-in-depth architectures that reduce hallucination rates to the lowest level compatible with your latency, cost, and complexity constraints.

If you are an AI engineer, ML engineer, or technical architect responsible for production RAG or agent systems, this book gives you the technical depth to understand every stage where hallucinations originate, the engineering techniques that prevent or detect them at each stage, and the production-ready code to implement those techniques. From knowledge source selection and document parsing to retrieval strategies, constrained generation, verification pipelines, agent memory management, and continuous monitoring, you will learn how to build systems whose reliability you can measure, improve, and justify.

# Introduction: The Hallucination Problem Is Not a Prompting Problem

In February 2026, a major law firm faced sanctions after filing a legal brief containing completely fabricated case citations generated by a generative AI system. The citations looked authentic, complete with plausible volume numbers, page references, and legal formatting. Judges discovered they were invented.

This is not an isolated incident. LLM hallucinations have caused financial losses across industries, from an airline chatbot that promised an unauthorized discount to companies that refunded AI-promised warranty claims. In healthcare, research has documented AI systems generating incorrect drug dosages and fabricating medical journal references. These are not edge cases. They are predictable failure modes of systems that prioritize fluent generation over factual accuracy.

## What Hallucination Means in Practice

The term hallucination has become ubiquitous in AI discourse. To work with it seriously, we need precision.

In the RAG and agent context, a hallucination is any generated claim that lacks support from the system's accessible evidence. This includes facts not present in retrieved documents, statistics invented by the model, relationships between entities that do not exist, citations pointing to sources that do not contain the claimed information, and contradictions of evidence that is actually available in context.

Researchers distinguish several fundamental types. Intrinsic hallucinations occur when the model contradicts the information it has been given. If retrieved documents state that a regulation went into effect on January 1, 2024, and the model says March 15, 2024, that is intrinsic hallucination. Extrinsic hallucinations are unsupported additions: information the model invents that does not appear in the context and may or may not be true in the real world.

Factuality hallucinations violate real-world facts. Faithfulness hallucinations violate the provided source material even when the facts might accidentally be correct.

Huang et al. categorize hallucination into factuality and faithfulness dimensions, showing that they have different causes and require different mitigations [1]. A model might be factually accurate but unfaithful by ignoring retrieved context in favor of parametric knowledge, or faithful to bad context that happens to be wrong. The EdinburghNLP curated list of hallucination detection research tracks the rapidly evolving literature [18].

## Why Hallucinations Are Inevitable

Probabilistic language models predict the next token by estimating distributions over all possible continuations. They are not truth machines. They optimize for statistical plausibility, not correspondence with reality. OpenAI's own research from September 2025 explains that standard training and evaluation paradigms reward guessing over abstention. If a model that guesses sometimes gets answers right and scores higher than a cautious model that says I do not know, training pushes toward confident errors.

This is not a bug waiting to be fixed. It is a feature of how next-token prediction works on arbitrary knowledge domains. Low-frequency facts, especially, lack reliable patterns in training data, leading to confident fabrications that are statistically plausible given the model's learned distributions.

Hallucination reduction is therefore fundamentally different from model improvement. You do not eliminate hallucinations. You constrain them through layered defenses: better knowledge sources, smarter retrieval, generation constraints, independent verification, and structured abstention when evidence is inadequate.

## What This Book Covers

This book traces the complete pipeline through which hallucinations can enter, propagate, or be corrected. It is organized into seven parts:

**Part I** establishes foundations. We define hallucination types rigorously, map the RAG and agent pipeline as an attack surface where errors can occur

at each stage, and compare architectural patterns including hybrid RAG, graph RAG, agentic RAG, and self-reflective systems in terms of their hallucination profiles.

**Part II** focuses on knowledge as the foundation. Hallucination prevention starts long before a query arrives. We cover knowledge source selection, document ingestion and parsing challenges especially with PDFs and complex layouts, metadata extraction and deduplication, and chunking strategies that affect retrieval accuracy. Poor knowledge engineering creates hallucination-prone systems that no amount of prompt engineering can fix.

**Part III** covers retrieval and indexing. We examine embedding model selection and fine-tuning, indexing strategies including vector indexes and knowledge graphs, query understanding and rewriting, hybrid search combining lexical and semantic methods, reranking techniques, and iterative retrieval patterns for multi-hop reasoning. Retrieval quality directly determines what evidence the model has access to, and garbage retrieval guarantees garbage generation.

**Part IV** addresses context construction and generation. We show how to assemble retrieved documents into effective context, write prompts that constrain generation to provided evidence, use constrained decoding to limit hallucination at the token level, and design structured outputs that reduce the model's freedom to invent unsupported content.

**Part V** covers verification and validation. This is where independent checks detect and block hallucinations. We explain claim extraction, natural language inference for entailment checking, dedicated verification models, self-consistency methods, rule-based and deterministic validation, and multi-layer defense architectures that combine pre-generation, in-generation, and post-generation safeguards.

**Part VI** examines agent-specific risks. Agents that plan, call tools, and remember across conversations face unique hallucination modes: fabricated tool results, planning failures, memory contamination, multi-step error propagation, and cascading failures across multiple agents. We cover each of these with specific mitigations.

**Part VII** covers evaluation and production. We provide comprehensive evaluation metrics, guidance on building test datasets and red-team tests, and production monitoring practices including drift detection and continuous improvement loops.

## How to Use This Book

Each chapter identifies specific hallucination modes at its stage of the pipeline, explains how errors propagate downstream, and provides practical techniques categorized as prevention, detection, verification, correction, mitigation, or abstention. Code examples use Python with widely adopted frameworks including LangChain, LlamaIndex, and Haystack, plus common vector databases, embedding models, and evaluation tools.

Do not treat this book as a checklist of techniques to apply indiscriminately. Every safeguard adds latency, cost, and complexity. The goal is defense in depth: layering safeguards appropriate to your risk profile, measuring their actual impact on hallucination rates, and understanding tradeoffs. A medical records system needs far more rigorous verification than a creative writing assistant.

By the end, you will understand why hallucinations occur, how they propagate through pipelines, and how to build systems that minimize them through engineering rigor rather than hopeful prompting.

# Chapter 1: What Hallucinations Are and Why They Matter

A customer service agent at an airline promised a passenger a $100 discount that did not exist under any policy. The passenger posted screenshots online. The airline honored the discount under consumer protection law. A legal firm filed court documents citing cases that never existed, complete with fabricated page numbers and volumes. Judges discovered the fabrications. Attorneys faced sanctions for professional misconduct. A dealership chatbot gave incorrect information about a vehicle warranty, and when customers relied on it, the company had to honor promises the bot had invented.

These are not hypothetical scenarios. They are documented failures of production systems deploying large language models. The pattern is consistent: models generate confident, plausible-sounding information that is simply false. This is hallucination, and in any domain where decisions carry consequences, it is the central engineering problem to solve.

## Defining Hallucination: Unsupported Claims Versus Errors Versus Fabrication

The word hallucination has become a catch-all term in AI discourse, which undermines precise discussion. We need working definitions that distinguish different phenomena so we can diagnose causes and select appropriate mitigations.

A hallucination is a generated claim lacking support from the model's accessible evidence. In a RAG system, accessible evidence means retrieved documents and any explicitly provided context. In an agent system, it additionally includes tool outputs, memory, and user-provided information. The critical distinction is that hallucination is defined relative to evidence, not relative to real-world truth. A model might accidentally generate a factually correct statement that it has no evidence for. That is still a hallucination because the correctness is coincidental, not grounded.

This definition separates hallucination from three related phenomena. Factual errors are claims that contradict real-world facts but may be supported by bad evidence. If the knowledge base contains an incorrect regulation date and the model faithfully reports it, that is a factual error but not a hallucination. The model is faithful to its context even though the context is wrong. Noise or confusion happens when the model misreads or misinterprets valid context due to ambiguity, poor formatting, or contradictory information. That is a comprehension failure rather than fabrication. And refusal failures are the model's inability to abstain when it should: generating an answer despite insufficient evidence. That is a decision boundary problem related to hallucination but distinct from the act of generating unsupported content itself.

Within hallucination, we distinguish two fundamental categories that appear throughout the literature. Extrinsic hallucinations introduce information not present in the available evidence. The model adds details, examples, statistics, or explanations that it has not been given. Some of these might coincidentally be true. Most are not. Intrinsic hallucinations contradict the provided evidence. The retrieved documents clearly state one thing, and the model generates something incompatible. This is arguably worse than extrinsic hallucination because the model had the correct information and chose not to use it.

Researchers have proposed more granular taxonomies. Huang et al. in their comprehensive survey categorize hallucinations along two axes: factuality versus faithfulness, and intrinsic versus extrinsic [1]. Factuality hallucinations violate real-world correctness regardless of provided context. Faithfulness hallucinations violate consistency with the source material regardless of real-world truth. This dual-axis taxonomy is useful because it shows that you can have a model that is faithful but not factual when given bad sources, or factual but not faithful when it ignores sources in favor of parametric knowledge.

For RAG and agent systems specifically, we can further decompose hallucination into types that map to different engineering interventions:

- Factual hallucinations: incorrect dates, numbers, names, events, or facts presented as true
- Entity hallucinations: invented people, organizations, products, or places
- Numerical hallucinations: fabricated statistics, percentages, prices, or measurements

- Temporal hallucinations: incorrect dates, wrong timelines, or events misattributed to wrong time periods
- Relational hallucinations: invented relationships between entities that do not exist in reality or in the sources
- Citation hallucinations: references to documents, pages, or sources that either do not exist or do not contain the claimed information
- Logical hallucinations: reasoning errors where conclusions do not follow from premises, even if the premises are correct
- Contextual hallucinations: answers that ignore or contradict the specific question asked, often providing plausible but irrelevant information

Understanding these types matters because different hallucination types require different detection and mitigation strategies. Entity hallucinations might be caught by checking entities against a known ontology. Numerical hallucinations can be verified arithmetically or through structured data queries. Citation hallucinations require checking that claimed source material actually contains the referenced information. Logical hallucinations need reasoning validation. No single technique catches all types.

## Origins in Language Models: Distributional Generalization, Memorization, and Training Bias

To prevent hallucination, we need to understand where it comes from. Hallucination is not a random glitch. It is a predictable consequence of how language models work.

At a fundamental level, language models are distributional predictors. During pretraining, they learn to predict the next token given all previous tokens, optimizing for likelihood over massive corpora. They never receive explicit truth labels. There is no supervision telling them which statements are factually correct and which are not. They learn statistical patterns: if tokens A and B often appear near token C in training data, the model learns that A and B make C probable. This works brilliantly for syntax, style, and common patterns. It works less well for specific facts, especially rare facts.

When a model encounters a query about something it has no reliable pattern for, it does not know that it does not know. It generates the statistically most plausible continuation, which may be completely wrong. OpenAI's 2025

research shows this clearly: models optimized for accuracy scores on benchmarks learn that guessing sometimes yields correct answers and higher scores than abstention, so training pushes them toward confident fabrication rather than cautious uncertainty [3].

Three specific mechanisms drive hallucination in practice:

Distributional generalization means the model applies learned patterns to new situations where they do not apply. If the model has seen many examples of drug dosages expressed as "X mg three times daily," it might generate a dosage following that pattern even when the actual dosage is different. The output looks plausible because it matches a pattern the model knows, not because it is correct.

Memorization artifacts occur when the model partially recalls training data but distorts it. Research on factuality in language models shows that models can memorize facts but retrieve them imperfectly, especially when facts are similar or when the query phrasing differs from training examples. A model might remember that a company acquired another entity but confuse the year, the price, or the parties involved.

Training bias means the model's generation is skewed toward patterns in training data that are statistically frequent but not necessarily true or representative. If training data contains many examples of confident assertions with fabricated or outdated information, the model learns that confident assertion is a valid strategy. This is especially problematic for domains where authoritative information is sparse in web-scale training data.

A key insight from recent research is that hallucination is not uniformly distributed. The RAGTruth study of 18,000 annotated responses found that hallucination rates increase with context length and response length, and hallucinations cluster toward the end of responses rather than the beginning [2]. This suggests attention dilution: as the model generates longer outputs, its focus on the source context weakens, and parametric knowledge begins to dominate. It also suggests that the model's self-consistency degrades as generation proceeds, with later tokens becoming less constrained by earlier commitments to the evidence.

## Real-World Harm: Medical Misinformation, Legal Errors, Financial Decisions, Security Vulnerabilities

Hallucination is not just an academic concern about model quality. It causes measurable harm.

In healthcare, research has documented LLMs generating fabricated drug dosages, inventing medical journal references, and providing incorrect treatment recommendations that would be dangerous if followed. The consequences are not theoretical. Health systems exploring AI-assisted clinical decision support must ensure hallucinated medical information does not reach clinicians or patients.

In law, hallucinated citations have become a recurring problem. Multiple documented cases involve attorneys filing court documents with AI-generated citations that look authentic but reference non-existent cases. Courts have sanctioned lawyers for this, even when the hallucinations were not immediately detectable. The legal domain is especially vulnerable because citation formats are highly structured and hallucinated citations can look convincing to anyone who does not verify them.

In finance, hallucinated statistics, invented market data, and fabricated company information can lead to incorrect investment decisions, regulatory violations, or financial losses. Financial RAG systems must distinguish between retrieved market data that the model should faithfully report and fabricated numbers that the model might invent when actual data is unavailable.

In operations and customer service, hallucinated policy information has created contractual obligations that companies did not intend. When a support bot promises a refund, discount, or policy interpretation that does not match actual company policy, customers reasonably expect the company to honor that promise.

Security is another dimension. Hallucination can mask prompt injection attacks when models follow malicious instructions embedded in retrieved documents, or it can cause agents to take incorrect actions when they hallucinate tool outputs or misinterpret real tool results. A fabricated tool response might make an agent believe a payment was processed when it was not, or that a user was authorized when they were not.

The common thread across all these domains is trust. Users of AI systems

reasonably expect that when a system provides information confidently, especially when it cites sources, that information is based on actual evidence. Hallucination violates this expectation systematically.

## Why Prompt Engineering Alone Fails: The Fundamental Probabilistic Problem

Given how central hallucination is, it is natural to ask: can we prompt our way out of it? The answer is essentially no.

Prompt engineering can reduce hallucination rates, but it cannot eliminate them. Instructions like "answer only from the provided documents" or "if you do not know, say you do not know" help. They shift the model's distribution slightly toward more cautious behavior. But they do not change the underlying mechanism: the model is still predicting tokens probabilistically, and there is always a non-zero probability that it will generate unsupported content despite the instruction.

Research demonstrates this limitation clearly. Studies comparing RAG with various prompt engineering approaches consistently find that while well-designed prompts reduce hallucination compared to naive prompts, hallucination persists even with optimized instructions. The reason is structural: prompts guide generation but do not constrain it. The model can always choose, probabilistically, to ignore its instructions.

Effective hallucination mitigation requires mechanisms beyond prompting:

- Retrieval quality: ensuring the model actually has the evidence it needs to answer correctly
- Constrained generation: limiting the model's freedom at the token level through grammar constraints or structured outputs
- Verification: independently checking generated claims against evidence
- Abstention: designing systems that explicitly refuse to answer when evidence is insufficient

Prompts are one tool among many. They are necessary but not sufficient. Systems that rely solely on prompt engineering for hallucination control are building their reliability on sand.

## Setting the Stage for Defense in Depth

Understanding what hallucination is, where it comes from, and why prompting alone cannot solve it sets up the rest of this book. If hallucination is a systemic property of probabilistic generation, then our strategy must be systemic as well. We cannot fix it at a single point. We must design defenses at every stage where hallucination can enter or propagate.

The next chapter maps the complete RAG and agent pipeline, showing exactly where hallucinations originate at each stage and how errors cascade downstream. That pipeline view is the foundation for every technique this book covers.

# Chapter 2: The RAG and Agent Pipeline as a Hallucination Attack Surface

Every stage of a RAG or agent pipeline is a potential source of hallucination. Understanding this pipeline as an attack surface rather than a monolithic system is essential for building effective defenses. Hallucination prevention is not a single component. It is a property that emerges from careful engineering across the entire stack.

This chapter maps each stage where hallucinations can originate, shows how errors propagate downstream, and introduces the defense-in-depth philosophy that structures the rest of the book. Zhou et al. frame trustworthiness in RAG as covering retrieval accuracy, evidence quality, generation reliability, and system safety, which maps directly to this pipeline view [11].

## The Standard RAG Pipeline: Ingestion, Indexing, Retrieval, Generation

A minimal RAG system has four stages:

1. **Ingestion**: documents are collected, parsed, cleaned, and transformed into retrievable units
2. **Indexing**: processed content is embedded and stored in a searchable index, typically a vector database
3. **Retrieval**: given a user query, relevant documents are fetched from the index
4. **Generation**: a language model produces an answer using the retrieved documents as context

Each stage can introduce hallucination:

Ingestion is where the knowledge base is built. If documents are parsed incorrectly, key information is lost. If PDFs with multi-column layouts are

read left to right without respecting column boundaries, the resulting text becomes garbled, and any retrieval from it will fail to find coherent evidence. If tables are extracted incorrectly, numerical data becomes unreliable. If conflicting documents are ingested without detection, the system has no way to know which information is current. All of these ingestion failures create hallucination-prone knowledge bases: the model will either fail to find the evidence it needs or will find incorrect evidence and faithfully reproduce it.

Indexing determines how well the system can find what it has ingested. Poor embedding models that do not capture domain-specific semantics will fail to match queries with relevant documents. Incorrect chunking that splits coherent information across multiple chunks means retrieval may return partial information that the model cannot use correctly. Indexes with low recall miss relevant documents entirely, forcing the model to answer without complete evidence.

Retrieval is where many hallucination failures originate. If the retriever returns irrelevant documents because the query was misunderstood or the embedding space is poorly calibrated, the model generates from bad context. If it returns too few documents, the model lacks sufficient evidence and may fill gaps with parametric knowledge. If it returns contradictory documents, the model may produce inconsistent or hedged answers. All of these are retrieval failures that manifest as generation hallucinations.

Generation is where hallucination becomes visible. Even with perfect ingestion, indexing, and retrieval, the model might ignore the evidence and generate from parametric knowledge, contradict the retrieved context despite it being clear, or add unsupported details that seem plausible. This is intrinsic hallucination: the model had the correct information and failed to use it correctly.

Each stage can introduce hallucination, and each stage can mitigate hallucination from previous stages. A good retriever cannot fix bad ingestion, but a good generator with strong grounding instructions can sometimes produce a faithful answer even from noisy retrieval, as long as the correct evidence is present. This interdependence is why defense in depth matters.

## Agentic RAG: Tools, Planning, Iteration, Multi-Step Reasoning

Standard RAG is a simple linear pipeline: ingest, index, retrieve, generate. Agentic RAG adds complexity and with it, new hallucination modes.

An agentic RAG system uses an LLM as a decision-making agent that can plan multi-step actions, call external tools including retrieval systems, iterate based on intermediate results, and maintain state across the interaction. This is powerful for complex tasks but introduces hallucination opportunities at each new layer.

Planning hallucination occurs when the agent invents steps that do not exist or will not work. If the agent plans to call a tool that is not available, or plans a sequence of actions that cannot achieve the goal, the entire trajectory is built on a hallucinated foundation.

Tool-selection hallucination happens when the agent chooses the wrong tool for a task or fabricates a tool that does not exist. An agent might decide to query a database when the information is only available through a specific API, or worse, it might hallucinate an API endpoint that never existed.

Tool-argument hallucination occurs when the agent generates incorrect parameters for tool calls. The right tool is selected but with wrong arguments: the wrong date range, the wrong entity identifiers, or invalid query syntax. This leads to incorrect or empty tool results that the agent then reasons about incorrectly.

Fabricated tool results are perhaps the most dangerous agent hallucination. The agent reports results from a tool call that never happened, or fabricates the content of a real tool's response. Research on agent hallucination specifically identifies fabricated tool outputs as a major risk: the agent claims an API returned certain data when in fact it returned something different or returned nothing at all. NabaOS, a framework published in 2026, tackles this by requiring HMAC-signed receipts for every tool execution, detecting fabricated results with over 90 percent accuracy.

Multi-step error propagation means that hallucination at any step affects all downstream steps. If the agent hallucinates the result of step one, it uses that hallucinated result as input for step two, potentially compounding the error. A hallucinated entity name from one step might cause subsequent retrieval steps

to fail completely.

Memory contamination occurs when agents store hallucinated information in persistent memory and later retrieve and use it as if it were real. The ConsistencyGate paper demonstrates that LLM agents can accumulate hallucinated facts at rates up to 50 percent in long trajectories, and these errors compound over time because existing memory admission criteria focus on utility rather than correctness [9].

Agentic systems also face reasoning failures: multi-step deduction where conclusions do not follow from premises, even when all the facts are correct. This is logical hallucination: the agent has the right information but reasons incorrectly.

The key insight is that adding agency does not just add new features. It adds new failure modes. Each capability the agent gains requires corresponding safeguards.

## Hallucination Modes by Stage: Knowledge, Retrieval, Context, Generation, Verification

Let us systematically categorize hallucination modes by pipeline stage so we can target interventions precisely.

**Knowledge-stage hallucination modes:**

- Missing information: critical facts are not in the knowledge base because sources were incomplete, parsing failed, or relevant documents were not ingested
- Incorrect information: facts in the knowledge base are wrong due to bad sources, outdated content, or parsing errors
- Contradictory information: the knowledge base contains conflicting facts without versioning or precedence rules to resolve them
- Malformed content: information exists but is garbled or fragmented in a way that prevents correct retrieval

Knowledge-stage failures are subtle because they are invisible during query time. The system works exactly as designed. The problem is that it is designed around bad knowledge.

**Retrieval-stage hallucination modes:**

- False negatives: relevant documents exist but are not retrieved, leaving the model without necessary evidence
- False positives: irrelevant documents are retrieved, confusing the model with misleading information
- Partial retrieval: only some of the necessary documents are retrieved, giving incomplete evidence
- Ranking failures: relevant documents are retrieved but ranked too low, so they fall outside the context window

Retrieval failures are particularly dangerous because they are hard to detect without ground truth. If the system retrieves something, even irrelevant information, there is no obvious signal that retrieval failed.

**Context-stage hallucination modes:**

- Context overflow: too much irrelevant information crowds out the relevant evidence in the context window
- Context conflict: multiple retrieved documents contradict each other, confusing the model about which is correct
- Attribution confusion: the model cannot distinguish which claim comes from which source, leading to incorrect citation or blending of incompatible claims
- Context noise: boilerplate, formatting artifacts, or irrelevant metadata in retrieved documents distracts the model

Context-stage failures happen even when retrieval succeeds. The right information was found but assembled in a way that confuses the model.

**Generation-stage hallucination modes:**

- Parametric override: the model ignores retrieved context and answers from its own training knowledge
- Intrinsic contradiction: the model contradicts clear information in the retrieved context
- Extraneous addition: the model adds plausible details not present in the context
- Citation fabrication: the model claims evidence comes from a specific source when it does not
- Format hallucination: the model invents numbers, dates, or entities in a required output format

Generation failures are the most visible because they appear directly in the output. But they are symptoms of upstream problems as often as they are pure model errors.

**Verification-stage hallucination modes:**

- False negatives in detection: hallucinations exist but verification fails to detect them
- False positives in detection: correct claims are flagged as hallucinated, causing unnecessary abstention
- Verification incompleteness: only some claims are checked, leaving unchecked hallucinations in the output

Verification failures are insidious because they create false confidence. The system believes it has checked the output when it has not.

## Error Propagation: How Upstream Failures Cascade Downstream

Hallucination failures rarely occur in isolation. A failure at one stage often amplifies through subsequent stages, making the final error worse than any single component failure would suggest.

Consider a chain of failures:

A medical document about drug interactions is ingested as a PDF. The parser fails to correctly handle the table layout and merges cells from different rows, creating a garbled entry that says Drug A interacts with Condition B when the actual table says Drug A interacts with Drug C. This ingestion failure creates incorrect knowledge.

A clinician queries the system about Drug A and Condition B. The embedding model, trained on general text rather than medical terminology, does not recognize the specific pharmacological terms well. The retriever finds the garbled table entry because the wrong Condition B is now associated with Drug A. This is a false positive retrieval, but it is a false positive created by upstream ingestion errors.

The retrieved context shows Drug A interacting with Condition B. The model is instructed to answer faithfully from context. It does, reporting the

drug interaction. This is faithful generation but factually incorrect because the evidence was wrong.

If the system includes a verification step that only checks whether claims are faithful to context, it passes: the claim matches the retrieved context. The factual error goes undetected.

In this scenario, four pipeline stages each worked correctly given their inputs, but the chain of upstream failures produced a hallucinated output that looked fully verified. This is why defense in depth cannot be purely linear: each defense must be aware that its inputs might be compromised.

Error propagation is particularly severe in agentic systems. A hallucinated plan causes the agent to take wrong actions. Wrong actions produce wrong results, which are then stored in memory and used for future decisions. The error compounds rather than resets.

## The Defense-in-Depth Philosophy: Prevention Versus Detection Versus Verification Versus Abstention

Defense in depth is borrowed from security engineering. The principle is simple: no single defense is perfect, so layer multiple defenses so that if one fails, another catches the error. Applied to hallucination, this means designing safeguards at every pipeline stage, categorized by their primary function:

**Prevention** mechanisms reduce the probability that hallucination will occur in the first place. Examples: careful knowledge source selection, robust document parsing, high-recall retrieval, prompts that constrain generation to context, and structured outputs that limit the model's freedom to invent.

**Detection** mechanisms identify hallucination after it has occurred. Examples: NLI-based entailment checking, self-consistency verification, and anomaly detection in output patterns. Detection alone is incomplete because detected hallucination must then be handled.

**Verification** mechanisms independently validate claims against evidence. Examples: claim extraction followed by source checking, citation verification, numerical validation, and cross-model checking. Verification is stronger than detection because it uses independent methods to confirm specific claims.

**Correction** mechanisms fix detected hallucination. Examples: regenerating with revised prompts, re-retrieving with modified queries, or post-processing

to remove unsupported claims. Correction is risky because the correction process itself can introduce new errors.

**Mitigation** mechanisms reduce the impact of hallucination without eliminating it. Examples: providing confidence scores, surfacing source citations so users can verify, and logging flagged outputs for human review.

**Abstention** mechanisms refuse to generate answers when evidence is insufficient. Examples: uncertainty-based refusal, abstention when confidence scores are low, and explicit "I do not know" responses when no supporting evidence is found.

An effective hallucination-resistant system uses all six categories in appropriate combinations. Prevention is cheapest because it stops errors before they occur. Detection and verification add reliability but cost latency and compute. Correction adds complexity and its own failure modes. Mitigation is a fallback. Abstention is essential because some queries genuinely cannot be answered from available evidence, and generating an answer in those cases is the definition of hallucination.

The goal is not zero hallucination. The goal is a hallucination rate low enough that the remaining errors are caught by downstream processes or human review, and a clear understanding of what those remaining errors look like so they can be monitored and reduced over time.

## What Comes Next

This chapter has mapped the pipeline where hallucinations can enter. Chapters 3 through 25 examine each stage in detail, identifying specific hallucination modes, explaining how errors propagate, and providing practical techniques and code to prevent, detect, and mitigate hallucination at that stage. The architecture of the book follows the architecture of the system.

We start in Chapter 3 by comparing RAG architecture patterns and their hallucination profiles, so you can choose an architecture appropriate to your risk and reliability requirements.

# Chapter 3: RAG Architecture Patterns and Their Hallucination Profiles



## Standard RAG: Strengths and Hallucination Failure Modes



## Hybrid RAG: Combining Lexical and Semantic Retrieval for Robustness



## Corrective RAG: Detecting and Correcting Bad Retrievals

## Self-Reflective and Self-RAG: Model Self-Critique and Conditional Generation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Adaptive RAG: Dynamic Strategy Selection Based on Query Complexity

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Graph RAG: Structured Knowledge for Multi-Hop and Relation Accuracy

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Agentic RAG: Iterative Retrieval and Tool Use with Hallucination Risks

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Multi-Agent Systems: Cross-Agent Verification and Conflict Resolution

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Choosing an Architecture for Your Risk Profile

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 4: Selecting and Vetting Knowledge Sources

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Source Trustworthiness Criteria: Authority, Accuracy, Recency, Completeness

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Primary Versus Secondary Versus Tertiary Sources: When Each Is Appropriate

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Domain-Specific Requirements: Medical, Legal, Financial, Technical

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Detecting and Excluding Unreliable, Adversarial, or Misleading Sources

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Source-Quality Scoring and Metadata Standards

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Provenance Tracking from Origin to Production Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Building a Source-Governed Knowledge Base

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 5: Document Ingestion, Parsing, and Cleaning

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Format Challenges: PDFs, Office Documents, HTML, Code, Images, Tables

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## PDF Parsing Pitfalls: OCR Errors, Layout Confusion, Page Breaks Mid-Claim

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## HTML Extraction: Boilerplate Removal, Script Noise, Embedded Ads

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Table and Figure Extraction: Structured Data Loss

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Cleaning and Normalization: Encoding, Whitespace, Special Characters

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Detecting Malformed or Corrupted Documents Before Indexing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## The Cost of Bad Parsing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 6: Metadata, Deduplication, and Knowledge Validation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Metadata Extraction: Titles, Dates, Authors, Sections, Entities

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Automatic Metadata Quality Validation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Semantic Deduplication: Collapsing Near-Identical Content

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Conflict Detection: Identifying Contradictory Sources

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Temporal Metadata: Handling Versioning and Superseded Information

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Knowledge Quality Scoring: Flagging Low-Confidence Documents

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Maintaining Knowledge Quality Over Time

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 7: Chunking Strategies and Semantic Segmentation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Why Chunking Matters: Context Fragmentation Versus Noise Dilution

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Fixed-Size Versus Semantic Chunking: Comparative Analysis

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Recursive and Hierarchy-Aware Chunking

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Entity-Aware and Claim-Aware Segmentation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Overlapping Chunks: Tradeoffs Between Continuity and Duplication

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Adaptive Chunking: Size Selection Based on Content Type and Query Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Choosing and Validating a Chunking Strategy

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 8: Embeddings and Vector Search for Faithful Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Embedding Model Selection: Domain Fit, Language Coverage, Dimensionality

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Embedding Drift and Degradation Over Time

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Vector Search Metrics: Cosine Similarity, Dot Product, Hybrid Scoring

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## The False-Negative Problem: When Relevant Content Is Not Retrieved

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## The False-Positive Problem: When Irrelevant Content Misleads Generation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Embedding Evaluation for Retrieval Quality

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Practical Recommendations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 9: Indexing Strategies and Knowledge Graphs



## Vector Index Types: HNSW, IVF, Flat, and Their Recall-Accuracy Tradeoffs



## Hybrid Indexes: Combining Vectors with Inverted Lexical Indexes



## Knowledge Graph Construction from Documents

## Graph-Based Retrieval: Multi-Hop Queries and Relation Grounding

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Graph RAG Architectures: Structured Plus Unstructured for Reliability

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Maintaining Graph Freshness and Consistency

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 10: Query Understanding and Rewriting

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Query Ambiguity and Vagueness: Causes of Hallucinated Best Guesses

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Query Expansion: Synonyms, Related Terms, Controlled Vocabularies

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Query Decomposition: Breaking Complex Questions Into Subqueries

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Query Rewriting for Retrieval: Style Matching, Entity Normalization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Negative Prompting: Specifying What Not to Retrieve

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Entity Linking and Disambiguation Before Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 11: Hybrid Search, Filtering, and Reranking

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Lexical Versus Semantic Versus Hybrid Search: When Each Succeeds and Fails

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Pre-Retrieval Filtering: Metadata Filters, Date Ranges, Source Types

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Cross-Encoder Reranking: Precision Versus Recall Versus Latency

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Query-Aware Filtering: Dynamic Filter Selection

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Reranking Model Selection and Fine-Tuning for Domain Accuracy

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Confidence Scoring of Retrieved Documents

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 12: Iterative and Agentic Retrieval Patterns

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Multi-Hop Retrieval: Chaining Queries for Complex Evidence Chains

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Self-RAG Style Reflection: Did I Find Enough Evidence?

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Iterative Deepening: Refining Search Based on Partial Results

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Agentic Retrieval: When the Agent Decides What to Look For Next

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Stop Conditions: Knowing When Further Retrieval Is Futile

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Cost and Latency Tradeoffs of Iterative Retrieval

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 13: Context Construction and Window Management

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Context Assembly: Ordering, Grouping, and Labeling Retrieved Documents

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Context Prioritization: Most Relevant Evidence First or Last

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Context Window Limits: Truncation Strategies and Information Loss

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Handling Contradictory Evidence in Context

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Document Source Attribution in Context Prompts

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Reducing Context-Induced Hallucination from Conflicting or Misleading Sources

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 14: Prompt Engineering for Grounded Generation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## System Instructions: Explicit Grounding Constraints

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Few-Shot Prompting with Grounded Examples

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Citation Instructions: Requiring Source References Per Claim

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Abstention Patterns: Say You Do Not Know When Evidence Is Absent

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Structured Response Formats: Reducing Free-Form Hallucination Space

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Prompt Injection Risks from Untrusted Retrieved Content

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 15: Constrained Decoding and Structured Outputs

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Grammar-Constrained Decoding: Limiting to Expected Output Structures

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Vocabulary Constraints: Restricting to Source-Derived Terms

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## JSON and Schema-Based Output Validation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Token-Level Constraints for Numerical and Entity Accuracy

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Sampling Parameters: Temperature, Top_p, and Hallucination Rates

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## When Constraints Are Worth the Complexity

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 16: Claim Extraction and Evidence Mapping

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Claim Extraction from Generated Text

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Entailment Testing: Does Evidence Entail the Claim?

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Citation Correctness: Verifying Cited Sources Actually Support Claims

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Citation Completeness: Detecting Unsupported Claims With No Citations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Numerical Verification: Checking Numbers Against Sources

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Temporal Verification: Checking Date-Sensitive Claims

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 17: Verification Models and Critic Systems

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## LLM-as-Judge Patterns: Strengths and Failure Modes

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Cross-Model Verification: Different Models Checking Each Other

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Dedicated Verifier Models: Trained for Entailment and Contradiction

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Self-Consistency: Generating Multiple Answers and Comparing

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Critic-Model Architectures: Separate Generation and Verification

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Confidence Estimation and Uncertainty Quantification

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 18: Rule-Based and Deterministic Verification

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Schema Validation: Type Checking, Required Fields, Value Ranges

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Deterministic Rules: Regex Patterns, Format Validation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Numerical Consistency: Arithmetic Verification

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Contradiction Detection: Logical Consistency Checks

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Citation Format and Existence Verification

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## When Deterministic Checks Are Superior to Model-Based Checks

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 19: Multi-Layer Defense Architectures

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Layered Architecture Design: Where Each Check Sits in the Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Pre-Generation Verification: Validating Retrieved Evidence

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## In-Generation Constraints: Real-Time Guidance

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Post-Generation Verification: Independent Checking

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Escalation Patterns: When to Re-Retrieve, Ask Human, or Abstain



## Reference Architectures for Different Risk Profiles

# Chapter 20: Hallucination Risks Specific to AI Agents

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Planning Hallucination: Inventing Steps That Do Not Exist or Will Not Work

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Tool-Selection Errors: Choosing Wrong Tools or Inventing Tools

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Fabricated Tool Results: Hallucinating API Responses

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Tool-Argument Errors: Wrong Parameters Leading to Wrong Data

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Multi-Step Error Propagation: Compounding Failures Across Steps

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Reasoning Failures: Logical Errors in Multi-Step Deduction

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 21: Agent Memory and Conversation History Management

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Memory Contamination: Old Facts Applied to New Situations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Conversation History Pruning: Keeping Relevant, Discarding Stale

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Cross-Session Leakage: Hallucination From Unrelated Conversations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Summarization Hallucination: Errors Introduced When Compressing History

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Memory Consistency: Ensuring Stored Facts Remain Accurate

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Forgetting: When Agents Should Drop Outdated Information

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 22: Multi-Agent Systems and Cross-Agent Verification



## Specialized Agents: Reducing Hallucination Through Domain Focus



## Debating Agents: Adversarial Refinement of Answers



## Cross-Verification: Independent Agents Checking Each Other's Work

## Consensus Mechanisms: Aggregating Multiple Agent Perspectives



## Conflict Resolution: When Agents Disagree



## Cascading Failures: When One Agent's Hallucination Infects Others

# Chapter 23: Evaluation Methodologies and Metrics



## Core Metrics: Faithfulness, Groundedness, Factual Correctness



## Retrieval Metrics: Context Relevance, Retrieval Precision and Recall



## Citation Metrics: Citation Correctness, Completeness, Hallucination Rate

## End-to-End Metrics: Answer Relevance, Unsupported-Claim Rate



## Calibration and Abstention Quality



## Combining Metrics Into System-Level Reliability Scores



## Practical Evaluation Workflow

# Chapter 24: Test Data, Benchmarks, and Red-Teaming

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Building Ground-Truth Evaluation Datasets

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Synthetic Test Case Generation With Controlled Difficulty

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Adversarial Tests: Queries Designed to Trigger Hallucination

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Existing Benchmarks: QAFactEval, FaithDial, HaluEval, RAGTruth

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Regression Testing: Preventing Hallucination Regressions

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Red-Team Methodologies for Hallucination Discovery

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Chapter 25: Production Monitoring and Continuous Improvement

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Real-Time Monitoring: Latency, Confidence, Verification Results

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## User Feedback Loops: Thumbs Up/Down, Corrections, Escalations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Drift Detection: Detecting When Retrieval or Generation Quality Degrades

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Incident Response: Handling Hallucination-Related Outages

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Continuous Evaluation: Automated Testing in Production Pipelines

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Documentation and Governance for Hallucination-Critical Systems

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# Conclusion: The Reality of Reliability

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## What Hallucination Rates Are Realistically Achievable

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Fundamental Limits: Why Probabilistic Models Will Always Hallucinate

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## The Defense-in-Depth Mindset: Accepting Risk and Layering Safeguards

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Cost Versus Reliability: Building Appropriate Safeguards for Your Risk Profile

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## Future Directions: What Research Might Change the Landscape

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

## A Framework for Ongoing Improvement

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.

# References

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/buildinghallucination-resistantragandaiagents.