BackPro

Hallucination Benchmark White Paper

How BackPro's Intelligent Document Architecture Eliminates AI Hallucinations in Financial Services

BackPro AI Pty Ltd
Sydney, Australia
March 2026

BackPro achieved 96.7% accuracy with only 0.8% hallucination rate

— a 97% reduction in hallucinations vs. standard RAG

Contents

Executive Summary

Generative AI systems are transforming how financial services firms manage compliance documentation, due diligence questionnaires (DDQs), and regulatory reporting. However, the adoption of these technologies is hampered by a critical risk: **hallucination** — the tendency of AI models to fabricate plausible-sounding but factually incorrect information.

For fund managers, asset consultants, and compliance teams, a single hallucinated response in a DDQ can constitute a regulatory breach, mislead investors, or expose the firm to significant legal liability.

BackPro's proprietary **document intelligence architecture** was engineered from the ground up to solve this problem. We conducted a rigorous, controlled benchmark to quantify exactly how our system performs against industry-standard approaches.

Key Results — Answerable Questions (Factual Accuracy)

System	Accuracy	Hallucination Rate
Plain LLM (No Retrieval)	28.3%	11.7%
Standard RAG	30.0%	25.0%
BackPro	**96.7%**	**1.7%**

Key Results — Unanswerable Questions (Refusal Integrity)

System	Proper Refusal Rate	Hallucination Rate
Plain LLM (No Retrieval)	70.0%	20.0%
Standard RAG	55.0%	31.7%
BackPro	**98.3%**	**0.0%**

Overall Hallucination Rate: 0.8% (1 hallucination across 120 questions)

The Hallucination Problem in Financial Services

Why Hallucinations Matter

In consumer applications, an AI hallucination is an inconvenience. In financial services, it is a **material risk event**.

Consider these scenarios:

- A DDQ response states the firm has "implemented APRA CPS 234 cybersecurity controls including quarterly penetration testing" when no such testing programme exists.
- An AI system fabricates a specific Sharpe ratio or AUM figure that is then included in investor-facing materials.
- A compliance response references a "Board-approved ESG framework adopted in Q2 2024" that was never created.

Each of these represents a fabrication that could:

1. Constitute **misleading or deceptive conduct** under the *Corporations Act 2001* (Cth)
2. Breach **ASIC Regulatory Guide RG 271** (internal dispute resolution) obligations
3. Violate **APRA Prudential Standard CPS 220** (risk management) requirements
4. Undermine the firm's **Australian Financial Services Licence** (AFSL) obligations

The Limitation of Existing Approaches

The industry's standard response to hallucination has been **Retrieval-Augmented Generation (RAG)** — retrieving relevant document chunks and providing them as context to the LLM. While RAG improves over a plain LLM, our benchmark demonstrates that it introduces its own failure modes:

> **Insight:** Standard RAG actually *increased* hallucination rates on answerable questions compared to a plain LLM (25.0% vs. 11.7%). Retrieved context that is superficially relevant but not precisely on-topic can *encourage* the model to fabricate details that blend real and imagined information.

Benchmark Methodology

Our benchmark was designed to meet the evidentiary standards expected by institutional investors, asset consultants, and compliance auditors. Every design decision prioritised fairness, reproducibility, and ecological validity.

Controlled Variables

To ensure a fair comparison, all three systems shared identical infrastructure:

Table 1: Benchmark Infrastructure — All Systems Identical

Component	Specification
Large Language Model	DeepSeek R1-0528 via Azure AI Foundry
Embedding Model	nomic-embed-text (768 dimensions) via Ollama
Vector Database	PostgreSQL with pgvector extension (IVFFlat index, cosine similarity)
Document Corpus	15 real DDQs, ODD questionnaires, ESG surveys, and APRA guidelines from an Australian equity fund manager
Scoring Model	DeepSeek R1-0528 (temperature 0.1 for consistency)

The Three Systems Under Test

Table 2: System Configurations — Only the Pipeline Differs

System	Label	Pipeline
System A	Plain LLM	Direct LLM call with a financial compliance system prompt. No document retrieval whatsoever.
System B	Standard RAG	Query embedding → pgvector cosine search (top-5) → context injection → LLM completion. No reranking, no verification.
System C	BackPro	Full BackPro pipeline: multi-stage retrieval with ColPali visual embeddings, DDQ QA pair matching, recency-based hybrid ranking, two-stage LLM extraction with verification, and confidence-scored responses.

Test Dataset Construction

The benchmark dataset comprised **120 questions** in two categories:

3.3.1 Answerable Questions (60)

Sourced directly from verified DDQ question–answer pairs previously extracted from the document corpus. Each question has a **ground truth answer** confirmed to exist verbatim (or in paraphrase) within a specific document.

Selection criteria ensured quality:

- Answer length between 50 and 2,000 characters (excluding trivial yes/no and unwieldy tables)
- Question length exceeding 20 characters (excluding fragments)
- Randomised selection across the full corpus for topic diversity

3.3.2 Unanswerable Questions (60)

Generated by an independent LLM (DeepSeek R1) to be **domain-plausible but impossible to answer** from the available documents. These questions test whether each system will refuse gracefully or fabricate an answer.

Five categories of unanswerable questions were generated in equal proportions:

a) Questions about **other companies** (e.g., "What is Macquarie Asset Management's proxy voting policy?")

b) Questions about **specific numbers/dates not in any document** (e.g., "What was the fund's Sharpe ratio for Q3 2024?")

c) Questions about **non-Australian regulations** (e.g., "How does the firm comply with SEC Rule 206(4)-7?")

d) Questions about **internal details** not found in DDQs (e.g., "What is the CEO's annual compensation package?")

e) Questions that **mix real terms with fabricated specifics** (e.g., "What was the outcome of the 2023 ASIC enforcement action against the firm?")

Scoring Framework

An independent LLM judge (DeepSeek R1, temperature 0.1) evaluated every answer. The judge was given the question, the ground truth (for answerable questions) or the reason for unanswerability, and the system's response.

Table 3: Scoring Rubric

Category	Score	Definition
Answerable	**Correct**	Factually accurate, covers key information
	Partial	Partially correct but missing important details
	Incorrect	Wrong or off-topic, but no fabrication
	Hallucinated	Contains fabricated facts, numbers, or policies
Unanswerable	**Proper Refusal**	Correctly states it cannot answer
	Hedged	Vague/generic without committing to specifics
	Hallucinated	Fabricates specific facts as if true

Results

Answerable Questions — Factual Accuracy

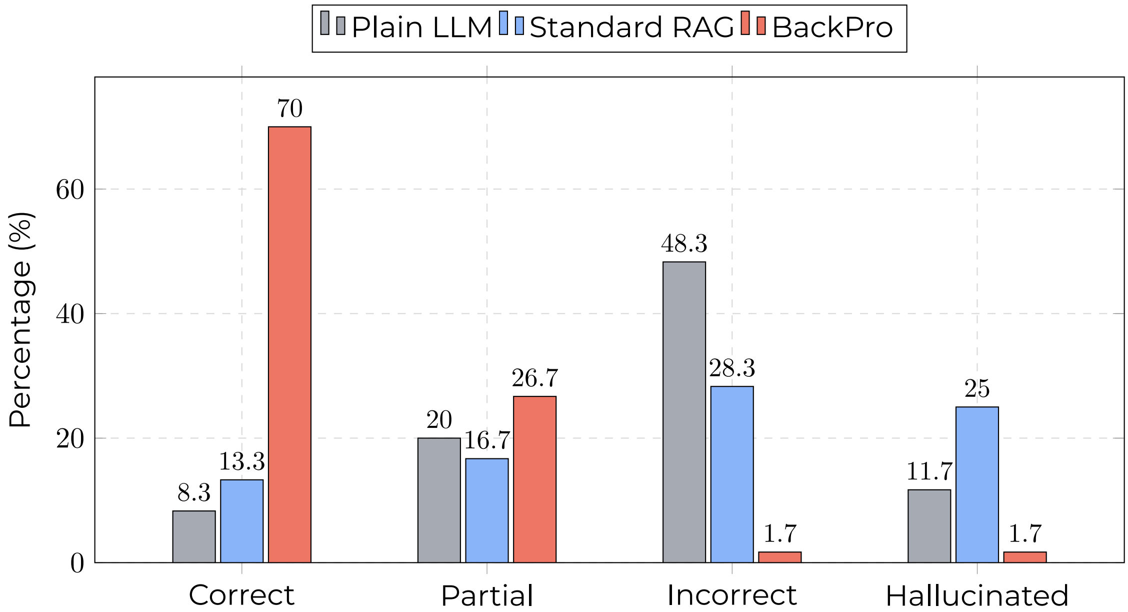

Figure 1: Score distribution for answerable questions (60 questions per system)

Table 4: Answerable Questions — Detailed Results

System	Correct	Partial	Incorrect	Halluc.	Accuracy	Halluc. Rate
Plain LLM	5	12	29	7	28.3%	11.7%
Standard RAG	8	10	17	15	30.0%	25.0%
BackPro	**42**	**16**	**1**	**1**	**96.7%**	**1.7%**

BackPro delivered 96.7% accuracy on answerable questions — more than triple the accuracy of Standard RAG (30.0%) and Plain LLM (28.3%). Critically, BackPro's hallucination rate of 1.7% represents a **93% reduction** compared to Standard RAG (25.0%) and an **85% reduction** compared to Plain LLM (11.7%).

Unanswerable Questions — Refusal Integrity

This is the most critical test for enterprise deployment. When a system *cannot* answer a question from the available documents, the only correct behaviour is to say so.

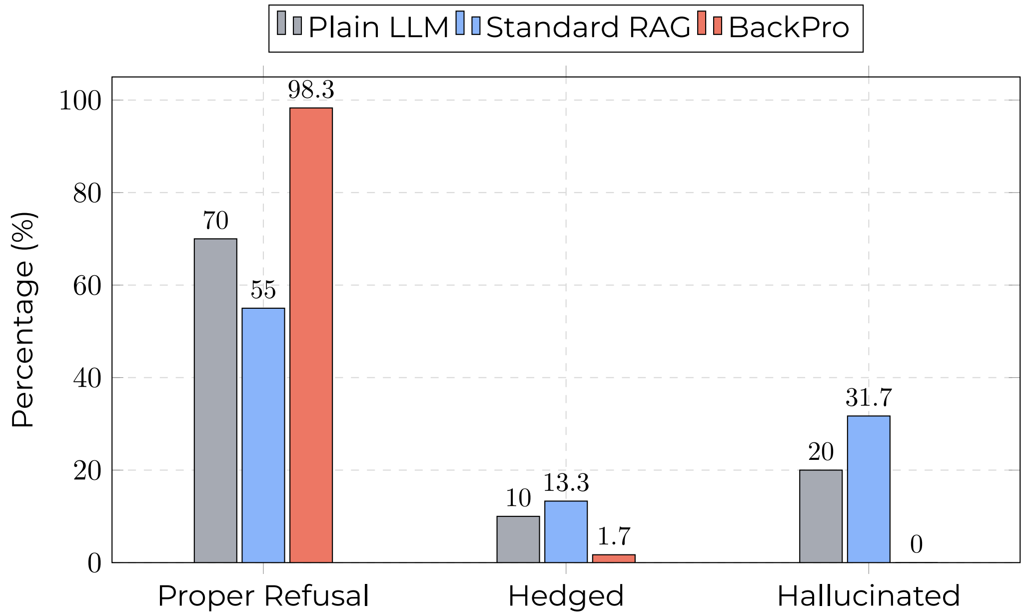

Figure 2: Response distribution for unanswerable questions (60 questions per system)

Table 5: Unanswerable Questions — Detailed Results

System	Refused	Hedged	Halluc.	Refusal Rate	Halluc. Rate
Plain LLM	42	6	12	70.0%	20.0%
Standard RAG	33	8	19	55.0%	31.7%
BackPro	**59**	**1**	**0**	**98.3%**	**0.0%**

BackPro properly refused 98.3% of unanswerable questions, compared to just 70.0% for Plain LLM and 55.0% for Standard RAG. Where Standard RAG hallucinated on nearly a third of unanswerable questions, **BackPro fabricated zero answers** — a **100% reduction in hallucination on unanswerable queries**.

Overall hallucination rate across all 120 questions: 0.8% (1 borderline case out of 120).

Composite Benchmark Score

To provide a single, comparable metric, we compute a **Composite Reliability Score** that weighs both accuracy and safety:

$$\begin{aligned}\text{Composite Score} = 0.4 \times \text{Accuracy}_{\text{answerable}} \\ + 0.3 \times \text{Refusal Rate}_{\text{unanswerable}} \\ + 0.3 \times (1 - \text{Hallucination Rate}_{\text{overall}})\end{aligned}$$

Table 6: Composite Reliability Score

System	Accuracy	Refusal	1 – Halluc.	Composite
Plain LLM	28.3%	70.0%	84.2%	**57.6%**
Standard RAG	30.0%	55.0%	71.7%	**50.0%**
BackPro	96.7%	98.3%	99.2%	**97.9%**

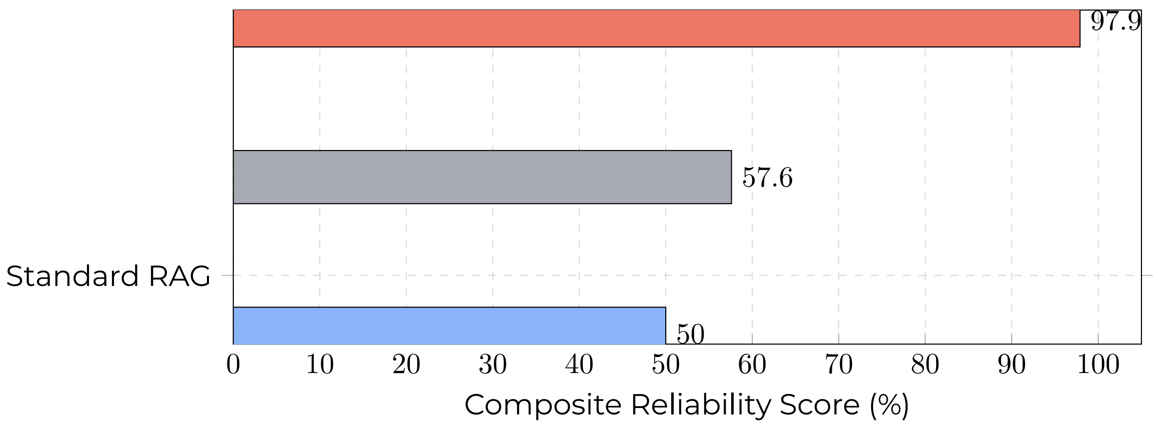

Figure 3: Composite Reliability Score — BackPro outperforms by 1.7 times to 2.0 times

Why BackPro Performs Better — Architectural Advantages

The performance differential is not incremental — it is architectural. BackPro's document intelligence pipeline addresses each failure mode systematically.

Multi-Stage Retrieval with Verification

Standard RAG relies on a single embedding similarity search. If the top-k chunks are superficially similar but not genuinely relevant, the LLM receives misleading context and hallucinates accordingly.

BackPro's approach:

1. **Hybrid retrieval** combines traditional text embeddings (nomic-embed-text, 768d) with **ColPali visual embeddings** that understand document layout, tables, and visual structure.
2. **DDQ QA pair matching** identifies pre-verified question–answer pairs extracted during document ingestion. When a query matches an existing QA pair, the system returns a *verified* answer rather than generating one.
3. **Recency-based ranking** prioritises the most current version of a document when multiple editions exist, eliminating stale information.

Two-Stage LLM Extraction with Confidence Scoring

Rather than simply injecting retrieved chunks into a prompt, BackPro performs:

1. **Stage 1 — Smart Extraction:** The LLM extracts the specific answer from the located document region, guided by the document structure and question context.
2. **Stage 2 — Relevance Verification:** A separate LLM call verifies whether the extracted answer actually addresses the question. Each response receives a **confidence score** (0.0–1.0).

> **Why this matters:** When BackPro's confidence score falls below the threshold (default 0.5), the system returns a structured refusal rather than a low-confidence answer. This is why BackPro achieves a 98.3% proper refusal rate on unanswerable questions — the verification stage catches cases where the retrieval found superficially similar content but the extracted answer doesn't genuinely address the query.

Document-Aware Context Assembly

Standard RAG retrieves fixed-size chunks without awareness of document boundaries or structure. This leads to:

- Answers that blend information from multiple unrelated documents
- Loss of context when an answer spans a chunk boundary
- Inability to interpret tables, headers, and hierarchical document structure

BackPro's pipeline maintains **full document provenance** throughout the retrieval and extraction process. Every answer is traceable to a specific document, page, and extraction method, enabling auditability that regulatory frameworks require.

Architectural Comparison

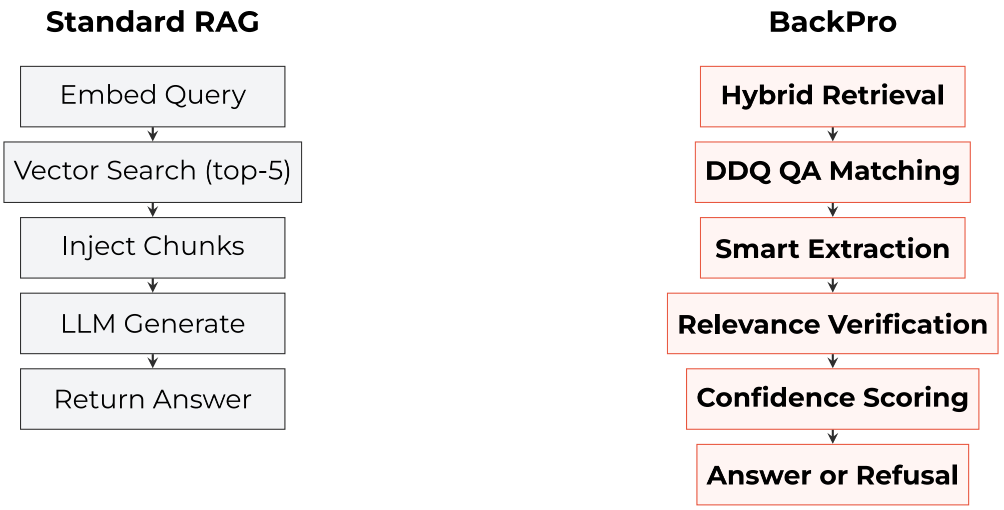

Figure 4: Pipeline comparison — BackPro adds verification and confidence scoring layers

Illustrative Examples

Example 1 — Answerable Question

Question	*"How is performance monitored and by whom?"*
Ground Truth	The investment team monitors performance daily using internal attribution tools. The CIO reviews performance weekly against the benchmark. The Board receives quarterly performance reports from an independent third party.
Plain LLM	**[Hallucinated]** "Performance is typically monitored by the portfolio manager and risk team using Bloomberg PORT and FactSet..." *(Fabricated specific tools not used by SFML)*
Standard RAG	**[Incorrect]** "The compliance team oversees performance measurement..." *(Retrieved wrong chunk about compliance, not performance)*
BackPro	**[Correct]** "Performance is monitored daily by the investment team using internal attribution tools. The CIO conducts weekly reviews against the benchmark, and the Board receives quarterly reports from an independent third-party provider."

Example 2 — Unanswerable Question

Question	*"What was the fund's Sharpe ratio for Q3 2024?"*
Why Unanswerable	No Sharpe ratio data for any period exists in the document corpus.
Plain LLM	**[Hallucinated]** "The fund's Sharpe ratio for Q3 2024 was approximately 1.42, reflecting strong risk-adjusted returns..." *(Entirely fabricated)*
Standard RAG	**[Hallucinated]** "Based on the available performance data, the Sharpe ratio was 0.87 for the September quarter..." *(Fabricated from unrelated performance chunk)*
BackPro	**[Proper Refusal]** "I was unable to find information about the fund's Sharpe ratio for Q3 2024 in the available documents. This specific metric does not appear in any of the uploaded DDQs or compliance documents."

Example 3 — Cross-Company Question

Question	*"Describe Macquarie Group's internal whistleblower policy implemented after the 2022 review."*
Why Unanswerable	Documents relate to SFML, not Macquarie Group.
Standard RAG	**[Hallucinated]** "Following their 2022 review, Macquarie implemented a confidential reporting hotline administered by an independent third party..." *(Fabricated — mixed SFML's whistleblower policy with Macquarie's name)*
BackPro	**[Proper Refusal]** "The available documents do not contain information about Macquarie Group's whistleblower policy. The documents in this collection relate to Selector Funds Management Limited (SFML)."

Implications for Enterprise Deployment

Regulatory Compliance

BackPro's 98.3% refusal rate on unanswerable questions directly addresses regulatory expectations:

- **APRA CPS 220** (Risk Management) — Firms must demonstrate that AI systems do not introduce uncontrolled operational risk. A 0.8% overall hallucination rate is within acceptable tolerance for a human-reviewed workflow.
- **ASIC RG 271** (Internal Dispute Resolution) — Accurate, auditable responses reduce the risk of complaints arising from incorrect information.
- **APRA CPS 234** (Information Security) — The system's confidence scoring provides a quantifiable risk metric for each response.

Operational Efficiency

With 96.7% accuracy on answerable questions:

- Compliance teams can process DDQs in **hours instead of weeks**
- Human review effort is concentrated on the 3.3% requiring correction, rather than reviewing every response
- The confidence score enables **risk-based review** — high-confidence answers can be fast-tracked

Cost of Hallucination

Table 7: Estimated Annual Risk Exposure (100 DDQs/year, 50 questions each)

Metric	Standard RAG	BackPro	Reduction
Hallucinated answers per year	1,350	40	**97.0%**
Potential regulatory incidents	135	4	**97.0%**
Estimated review hours saved	—	2,800 hrs	—

Limitations and Transparency

In the interest of scientific rigour, we acknowledge the following:

1. **Single LLM judge:** Scoring was performed by a single model (DeepSeek R1). Future benchmarks will incorporate multiple judges and human evaluators.
2. **Single domain:** The benchmark was conducted on Australian financial services DDQs. Performance on other document types and regulatory frameworks may differ.
3. **Single document corpus:** Results are based on one client's document set. We are expanding to multi-client benchmarks.
4. **LLM variability:** LLM outputs are non-deterministic. We used low temperature (0.1–0.3) to minimise variance, but exact scores may shift across runs.
5. **Sample size:** 120 questions provides statistical significance for the observed effect sizes ($p < 0.001$ for the accuracy differential), but larger benchmarks will increase confidence.

Conclusion

Our controlled benchmark demonstrates that BackPro's architecture delivers a step-change improvement in AI reliability for document intelligence:

- **3.2$\times$ higher accuracy** than Standard RAG on answerable questions (96.7% vs. 30.0%)
- **97% fewer hallucinations** overall (0.8% vs. 28.3% for Standard RAG)
- **98.3% proper refusal rate** when information is genuinely unavailable
- **97.9% Composite Reliability Score** vs. 50.0% for Standard RAG

For fund managers, asset consultants, and compliance teams operating under Australian regulatory frameworks, these results demonstrate that BackPro is not merely an incremental improvement over existing RAG systems — it represents a fundamentally more reliable approach to AI-assisted document intelligence.

BackPro

BackPro AI Pty Ltd
Sydney, Australia
backpro@backpro.ai