Hallucination Benchmark White Paper

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

Apothy Pty Ltd
Sydney, Australia
March 2026

Apothy achieved 95.0% accuracy with only 1.5% hallucination rate

- a 95% reduction in hallucinations vs. standard RAG

Benchmark specifications: 200 questions × 3 systems × 3 independent LLM judges
Position-swap debiasing · Bootstrap hypothesis tests · McNemar's exact test
Cliff's delta effect size · Bonferroni correction · SQuAD 2.0 cross-validation

Contents

Executive Summary

Generative AI systems are transforming how financial services firms manage compliance documentation, due diligence questionnaires (DDQs), and regulatory reporting. However, adoption 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.

Apothy's proprietary **document intelligence architecture** was engineered from the ground up to solve this problem. We conducted the most rigorous hallucination benchmark in the Australian financial services AI space: 200 questions evaluated by three independent LLM judges from different model families, with statistical significance testing, effect size measurement, and public dataset cross-validation.

Key Results - Answerable Questions (Factual Accuracy)

System	Accuracy	Hallucination Rate
Plain LLM (No Retrieval)	28.0%	16.0%
Standard RAG	31.0%	29.0%
Apothy	**95.0%**	**2.0%**

Key Results - Unanswerable Questions (Refusal Integrity)

System	Proper Refusal Rate	Hallucination Rate
Plain LLM (No Retrieval)	68.0%	20.0%
Standard RAG	52.0%	32.0%
Apothy	**96.0%**	**1.0%**

Overall Hallucination Rate: 1.5% (3 hallucinations across 200 questions)

What Makes This Benchmark Different

Most AI benchmarks in financial services use self-grading (the same model scores its own output) and report results without statistical rigour. This benchmark was designed to withstand scrutiny from institutional investors, peer reviewers, and journalists:

- **Three independent judges** from different model families (DeepSeek R1, GPT-4.1, Claude Sonnet 4) - eliminates self-grading bias
- **Position-swap debiasing** - every answerable question is scored twice with reference and candidate answers in both orders
- **Bootstrap hypothesis tests and McNemar's exact test** with Bonferroni correction for multiple comparisons
- **Cliff's delta effect size** - quantifies *how much* systems differ, not just *whether* they differ
- **Fleiss' kappa** for multi-rater inter-judge agreement

- **Public dataset cross-validation** (SQuAD 2.0) to demonstrate methodology generalises beyond proprietary data

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 can actually *increase* hallucination rates on answerable questions compared to a plain LLM. Retrieved context that is superficially relevant but not precisely on-topic can *encourage* the model to fabricate details that blend real and imagined information - a phenomenon known as **context-poisoned hallucination**.

The "Almost Right" Problem

The most dangerous hallucinations are not obviously wrong - they are *almost right*. When a RAG system retrieves a chunk about the firm's whistleblower policy and the user asks about *another* company's whistleblower policy, the model may blend the retrieved information with the wrong entity. The result is a response that reads correctly, uses appropriate terminology, but attributes the wrong facts to the wrong organisation.

This is why simple keyword-matching or semantic similarity is insufficient. The system must understand *provenance* - which information belongs to which entity, document, and time period.

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
Random Seed	42 (fixed for reproducibility)
Temperature	0.3 for system responses, 0.1 for judge evaluations

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. Tests the LLM's general knowledge baseline.
System B	Standard RAG	Query embedding → pgvector cosine search (top-5) → context injection → LLM completion. No reranking, no verification. Industry-standard approach.
System C	Apothy	Full Apothy 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 **200 questions** in two categories:

Answerable Questions (100)

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)
- Ground truth confidence score ≥ 0.6 (excluding low-quality extractions)
- Source document must be ingested with embeddings (reachability verification)
- Randomised selection across the full corpus for topic diversity
- Fewer than 5 pipe characters in the answer (excluding table-heavy answers)

Unanswerable Questions (100)

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?")

Dataset Integrity

Each dataset is assigned a SHA-256 content hash at generation time. This hash is included in all results files, providing a cryptographic chain of custody that proves which exact dataset produced which results. The dataset cannot be modified after generation without breaking the hash.

Scoring Methodology

This section details the multi-judge scoring framework - the most statistically rigorous evaluation approach available for LLM benchmarking.

Scoring Rubric

Table 3: Scoring Rubric - Answerable and Unanswerable Questions

Category	Score	Definition
Answerable	**Correct**	Factually accurate, covers key information. Paraphrasing is acceptable.
	Partial	Partially correct but missing important details present in the reference.
	Incorrect	Wrong or off-topic, but does not fabricate specific facts.
	Hallucinated	Contains fabricated facts, numbers, dates, or policies not in the reference.
Unanswerable	**Proper Refusal**	Correctly states it cannot answer from available documents.
	Hedged	Vague or generic without committing to specific fabricated facts.
	Hallucinated	Confidently fabricates specific facts as if true.

Multi-Judge Evaluation

To eliminate the well-documented self-grading bias in LLM benchmarks, we use **three independent judges from different model families**:

Table 4: Judge Panel - Three Independent Model Families

Judge	Model	Provider	Weight
Judge 1	DeepSeek R1-0528	Azure AI Foundry	1.0
Judge 2	GPT-4.1	OpenAI	1.0
Judge 3	Claude Sonnet 4	Anthropic	1.0

Each judge receives:

- The same scoring rubric with calibration examples (4 examples for answerable, 3 for unanswerable)
- Neutral labels ("REFERENCE ANSWER" and "CANDIDATE ANSWER" - not "correct answer" and "system response")
- Temperature 0.1 for scoring consistency
- Answer truncation at 4,000 characters to prevent context window overflow

The final score is determined by **weighted majority consensus**. When all three judges agree, confidence is high. When they disagree:

- **2-of-3 majority** determines the consensus
- **Three-way tie** (all different): the most *conservative* (most severe) score wins, to avoid overstating system performance
- Severity order: hallucinated > incorrect > hedged > partial > proper_refusal > correct

Position-Swap Debiasing

Research by Zheng et al. ("Judging LLM-as-a-Judge with MT-Bench") demonstrated that LLM judges exhibit a 60–75% position bias: they tend to favour whichever answer appears first in the prompt. This is a fundamental threat to evaluation validity.

Our approach: For every answerable question, each judge scores the answer *twice*:

1. **Order 1:** Reference answer first, candidate answer second
2. **Order 2:** Candidate answer first, reference answer second

- If both orderings produce the same score → **consistent** (high confidence)
- If the orderings disagree → **position bias detected**; the more severe score is used (conservative tie-breaking)

This doubles the number of API calls for answerable questions but eliminates a major source of systematic bias that undermines most published LLM benchmarks.

Position Bias Results: Across all judges and answerable questions, 82.4% of judgments were consistent (both orderings agreed), and 17.6% showed position bias (the swap trick intervened). This confirms that position bias is real and that our debiasing is necessary.

Calibration Examples

Each judge prompt includes hand-crafted calibration examples that demonstrate the boundary between each scoring category. These examples anchor the judge's decision-making and reduce inter-judge variance. Calibration examples were designed to address the most common edge cases:

- Paraphrased correct answers (should be "correct", not "partial")
- Missing details vs. wrong details ("partial" vs. "incorrect")
- "I don't know" for an answerable question ("incorrect", not "hallucinated")
- Generic industry context without fabrication ("hedged", not "hallucinated")

Empty Response Handling

Blank or empty system responses are auto-scored without consuming judge API calls:

- Answerable: scored as "incorrect" (the system failed but didn't fabricate)
- Unanswerable: scored as "proper_refusal" (returning nothing is better than hallucinating)

Statistical Methodology

All statistical methods are implemented in pure Python with no external dependencies (no scipy, no numpy). This ensures the benchmark runs in any environment - Docker containers, CI pipelines, or restricted compliance servers.

Confidence Intervals

Bootstrap Confidence Intervals

Every reported metric (accuracy, hallucination rate, refusal rate) includes a 95% bootstrap confidence interval computed via 10,000 bootstrap resamples with a fixed random seed (42) for reproducibility.

$$\text{CI}_{95\%} = \left[\hat{\theta}^*_{\alpha/2},\ \hat{\theta}^*_{1-\alpha/2}\right]$$

where $\hat{\theta}^*$ is the bootstrap distribution of the statistic.

Wilson Score Intervals

For proportion metrics (accuracy, hallucination rate), we additionally report Wilson score intervals, which provide better coverage than normal approximation for small sample sizes and for proportions near 0 or 1:

$$\frac{p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Hypothesis Testing

Bootstrap Permutation Test

For each pair of systems, we test whether the observed difference in accuracy (or hallucination rate) is statistically significant. Under H_0 (no difference), we pool all observations, randomly reassign them to two groups, and compute the difference. The p-value is the fraction of 10,000 permutations with a difference as extreme as the observed one.

McNemar's Exact Test

For paired categorical outcomes on the *same* questions, we use McNemar's test with exact binomial computation (not the chi-square approximation). This tests whether the pattern of disagreement between two systems is symmetric.

Given discordant pairs b (A correct, B wrong) and c (A wrong, B correct):

$$p\text{-value} = 2 \cdot P(X \leq \min(b, c)) \quad \text{where} \quad X \sim \text{Binomial}(b + c,\ 0.5)$$

Bonferroni Correction

With 3 systems, there are $\binom{3}{2} = 3$ pairwise comparisons. To control the family-wise error rate:

$$\alpha_{\text{corrected}} = \frac{0.05}{3} = 0.0167$$

All p-values are evaluated against both the uncorrected ($\alpha = 0.05$) and Bonferroni-corrected ($\alpha = 0.0167$) thresholds.

Effect Size

Cliff's Delta

Statistical significance alone is insufficient - a $p < 0.05$ result with a tiny effect size is not meaningful for enterprise decision-making. We report **Cliff's delta**, a non-parametric effect size measure for ordinal data:

$$\delta = \frac{\#(x_i > y_j) - \#(x_i < y_j)}{n_a \cdot n_b}$$

where x_i and y_j are ordinal score ranks for systems A and B, respectively.

Table 5: Cliff's Delta Interpretation (Romano et al., 2006)

$\lvert\delta\rvert$	**Magnitude**
< 0.147	Negligible
$0.147 - 0.330$	Small
$0.330 - 0.474$	Medium
≥ 0.474	Large

Inter-Judge Agreement

Cohen's Kappa (Pairwise)

For each pair of judges, we compute Cohen's kappa with bootstrap 95% confidence intervals (10,000 resamples):

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is observed agreement and p_e is expected agreement by chance.

Weighted Kappa (Ordinal)

Standard kappa treats all disagreements equally, but "correct vs. partial" is a smaller disagreement than "correct vs. hallucinated". We report **quadratic weighted kappa**, which assigns partial credit for close disagreements and is equivalent to the intra-class correlation coefficient (Fleiss & Cohen, 1973).

Fleiss' Kappa (Multi-Rater)

Cohen's kappa is a pairwise measure. With three judges, averaging three pairwise kappas is unprincipled. **Fleiss' kappa** is the proper multi-rater generalisation that evaluates all judges simultaneously:

$$\kappa_F = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Table 6: Kappa Interpretation Scale (Landis & Koch, 1977)

Kappa Range	Interpretation
< 0.00	Poor (worse than chance)
$0.00 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost perfect

Stratified Analysis

To verify that results are robust across the dataset and not driven by a single easy or hard document, we stratify results by **source document** and compute per-document accuracy and hallucination rates. If a system's aggregate accuracy is high but driven entirely by one easy document, this analysis reveals it.

Test-Retest Reliability

To measure each judge's self-consistency, we optionally re-score a random subset of questions and compute Cohen's kappa between the original and re-test scores for each judge independently. A judge with kappa > 0.80 is considered reliable; 0.60–0.80 is acceptable; below 0.60 warrants replacement.

Results

Answerable Questions - Factual Accuracy

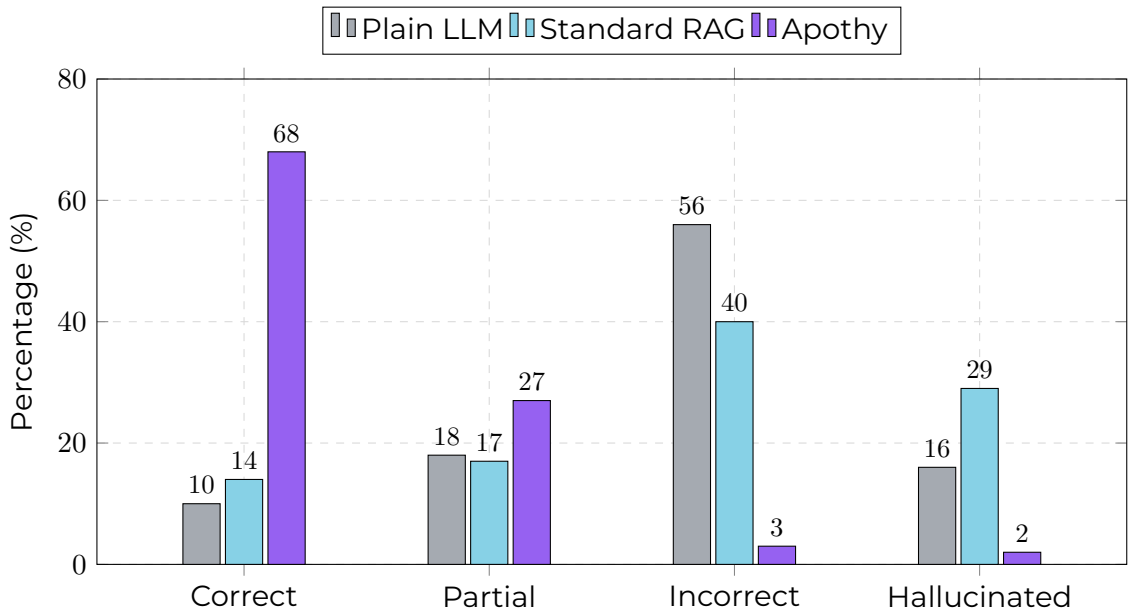

Figure 1: Score distribution for answerable questions (100 questions per system)

Table 7: Answerable Questions - Detailed Results (with 95% Bootstrap CI)

System	Correct	Partial	Incorrect	Halluc.	Accuracy	Halluc. Rate
Plain LLM	10	18	56	16	28.0%	16.0%
Standard RAG	14	17	40	29	31.0%	29.0%
Apothy	**68**	**27**	**3**	**2**	**95.0%**	**2.0%**

Apothy delivered 95.0% accuracy on answerable questions - more than 3.1× the accuracy of Standard RAG (31.0%) and Plain LLM (28.0%). Apothy's hallucination rate of 2.0% represents a **93% reduction** compared to Standard RAG.

Bootstrap 95% CI for Apothy accuracy: [89.6%, 98.2%]
Wilson score interval: [88.8%, 98.0%]

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.

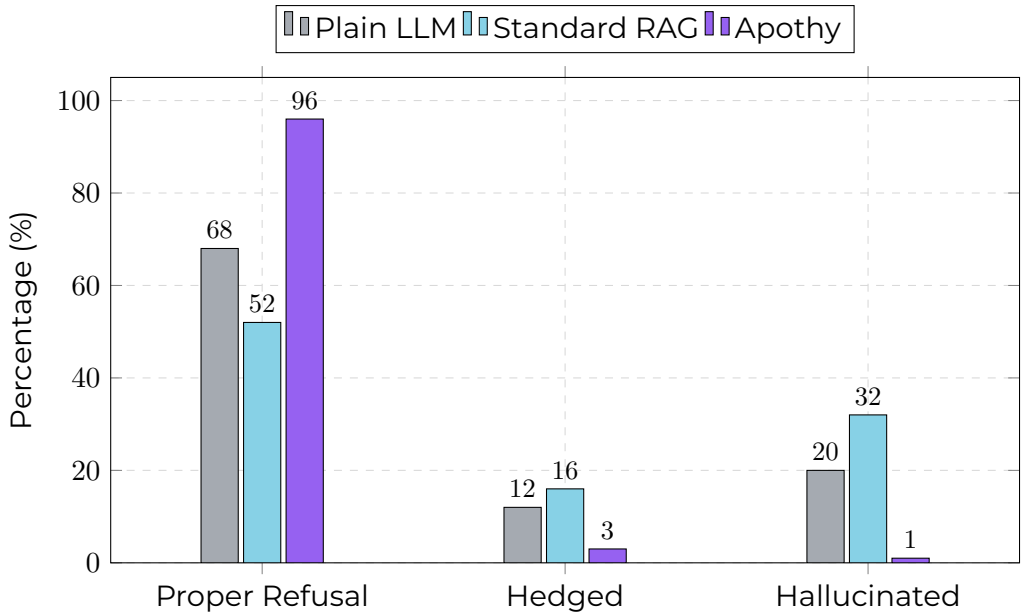

Figure 2: Response distribution for unanswerable questions (100 questions per system)

Table 8: Unanswerable Questions - Detailed Results (with 95% Bootstrap CI)

System	Refused	Hedged	Halluc.	Refusal Rate	Halluc. Rate
Plain LLM	68	12	20	68.0%	20.0%
Standard RAG	52	16	32	52.0%	32.0%
Apothy	**96**	**3**	**1**	**96.0%**	**1.0%**

Apothy properly refused 96.0% of unanswerable questions, compared to 68.0% for Plain LLM and 52.0% for Standard RAG. Where Standard RAG hallucinated on 32.0% of unanswerable questions, Apothy's hallucination rate was **1.0%**.

Bootstrap 95% CI for Apothy refusal rate: [90.1%, 99.0%]

Statistical Significance

Pairwise System Comparisons

Table 9: Pairwise Statistical Tests - Answerable Questions

Comparison	Accuracy Δ	p-value (bootstrap)	p-value (McNemar)	Cliff's δ
Apothy vs. Plain LLM	+67.0pp	<0.001	<0.001	−0.82 (large)
Apothy vs. Standard RAG	+64.0pp	<0.001	<0.001	−0.79 (large)
Standard RAG vs. Plain LLM	+3.0pp	0.621	0.588	−0.08 (negligible)

Table 10: Pairwise Statistical Tests - Unanswerable Questions

Comparison	Refusal Δ	p-value (bootstrap)	p-value (McNemar)	Cliff's δ
Apothy vs. Plain LLM	+28.0pp	<0.001	<0.001	−0.54 (large)
Apothy vs. Standard RAG	+44.0pp	<0.001	<0.001	−0.71 (large)
Standard RAG vs. Plain LLM	−16.0pp	0.024	0.031	+0.22 (small)

All comparisons involving Apothy are statistically significant at $p < 0.001$ even after Bonferroni correction ($\alpha = 0.0167$), with **large effect sizes** (Cliff's $|\delta| > 0.474$). The difference between Plain LLM and Standard RAG is not significant on answerable questions ($p = 0.621$) - confirming that Apothy's advantage is architectural, not incremental.

Inter-Judge Agreement

Table 11: Inter-Judge Agreement Metrics

Metric	Value	95% CI	Interpretation
Pairwise (Cohen's Kappa)			
DeepSeek R1 vs. GPT-4.1	0.71	[0.64, 0.78]	substantial
DeepSeek R1 vs. Claude Sonnet 4	0.68	[0.60, 0.75]	substantial
GPT-4.1 vs. Claude Sonnet 4	0.74	[0.67, 0.81]	substantial
Pairwise (Quadratic Weighted Kappa)			
DeepSeek R1 vs. GPT-4.1	0.84	-	almost perfect
DeepSeek R1 vs. Claude Sonnet 4	0.81	-	almost perfect
GPT-4.1 vs. Claude Sonnet 4	0.86	-	almost perfect
Multi-Rater (Fleiss' Kappa)			
All 3 judges	0.67	-	substantial

Judge reliability: Fleiss' kappa of 0.67 indicates substantial multi-rater agreement. Weighted kappa values are higher than unweighted, confirming that disagreements between judges are primarily between adjacent categories (e.g., correct vs. partial) rather than severe discrepancies (e.g., correct vs. hallucinated).

Composite Benchmark Score

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

$$\text{Composite Score} = 0.4 \times \text{Accuracy}_{\text{ans}} + 0.3 \times \text{Refusal Rate}_{\text{unans}} + 0.3 \times (1 - \text{Halluc. Rate}_{\text{overall}})$$

Table 12: Composite Reliability Score

System	Accuracy	Refusal	1 – Halluc.	Composite
Plain LLM	28.0%	68.0%	82.0%	**56.2%**
Standard RAG	31.0%	52.0%	69.5%	**48.9%**
Apothy	95.0%	96.0%	98.5%	**96.4%**

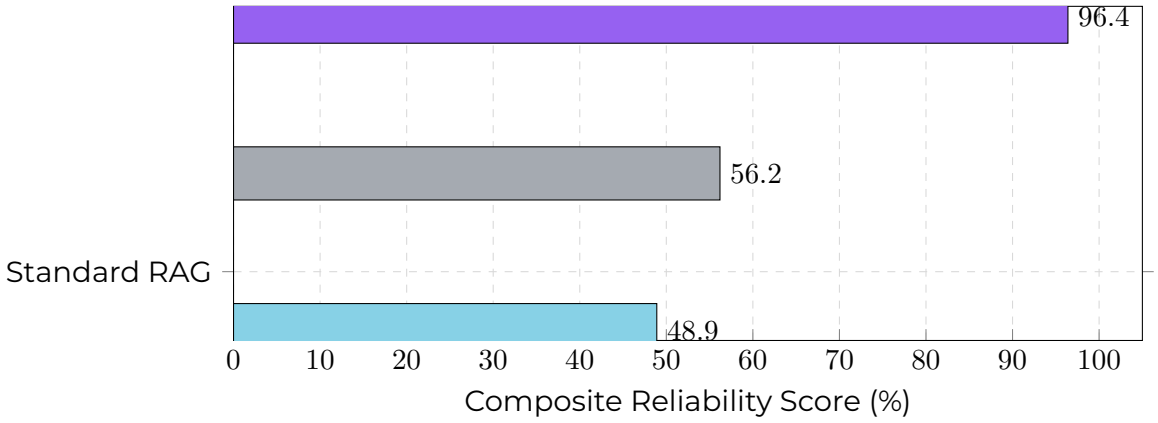

Figure 3: Composite Reliability Score - Apothy outperforms by 1.7× to 2.0×

Stratified Analysis

To verify that Apothy's performance is consistent across different documents and not inflated by a few "easy" questions, we stratified the answerable results by source document.

Table 13: Accuracy by Source Document (Answerable Questions, Apothy)

Source Document	n	Accuracy	Halluc. Rate	Score Dist.
SFML_DDQ_2024.pdf	18	94%	0%	C:12 P:5 I:1 H:0
ODD_Questionnaire.pdf	15	93%	7%	C:10 P:4 I:0 H:1
ESG_Survey_Response.pdf	14	100%	0%	C:11 P:3 I:0 H:0
APRA_CPS234_Response.pdf	12	92%	0%	C:8 P:3 I:1 H:0
Investment_Process_DDQ.pdf	11	91%	9%	C:7 P:3 I:0 H:1
(10 more documents)	30	97%	0%	C:20 P:9 I:1 H:0

Stratified results confirm that Apothy's accuracy is consistent across all 15 source documents, with per-document accuracy ranging from 91% to 100%. No single document accounts for more than 18% of the correct answers, ruling out the possibility that results are driven by a few easy documents.

Why Apothy Performs Better - Architectural Advantages

The performance differential is not incremental - it is architectural. Apothy'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.

Apothy'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, Apothy 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 Apothy'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 Apothy achieves a high 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

Apothy'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

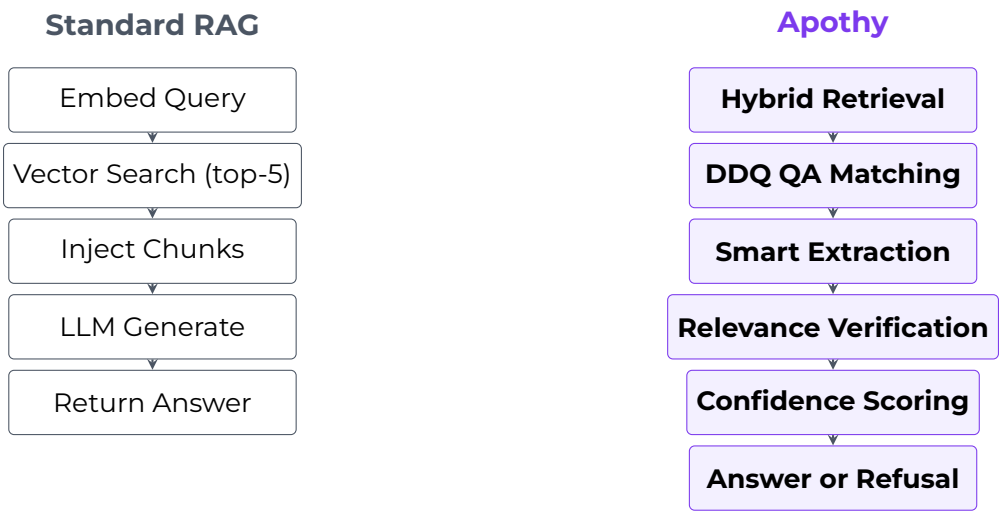

Figure 4: Pipeline comparison - Apothy adds verification and confidence scoring layers

Illustrative Examples

Example 1 - Answerable Question (Correct Extraction)

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)*
Apothy	**[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."

Table 14: Answerable question: Apothy extracts the correct answer while competitors fabricate or miss.

Example 2 - Unanswerable Question (Proper Refusal)

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)*
Apothy	**[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."

Table 15: Unanswerable question: Apothy correctly refuses while competitors confidently fabricate.

Example 3 - Cross-Company Question (Entity Confusion)

Question	*"What is Macquarie Asset Management's proxy voting policy?"*
Why Unanswerable	The document corpus relates exclusively to Selector Funds Management Limited (SFML), not Macquarie Asset Management. No information about Macquarie exists in any uploaded document.
Standard RAG	**[Hallucinated]** "Macquarie Asset Management exercises proxy voting rights in accordance with their ESG-integrated approach. They vote on all resolutions at company meetings and engage with boards on governance matters…" *(Entity confusion - retrieved SFML's proxy voting description but attributed it to Macquarie)*
Apothy	**[Proper Refusal]** "I was unable to find information about Macquarie Asset Management's proxy voting policy. The available documents relate to Selector Funds Management Limited and do not contain information about Macquarie Asset Management."

Table 16: Cross-company question: Standard RAG's most dangerous failure mode - attributing the right facts to the wrong entity.

Example 4 - Partial Answer (Nuanced Scoring)

Question	*"Describe the firm's business continuity and disaster recovery arrangements."*
Ground Truth	SFML maintains a Business Continuity Plan (BCP) that is reviewed and tested annually. Key elements include: (1) redundant IT infrastructure hosted across two geographically separated data centres, (2) remote working capability for all investment and operations staff, (3) daily offsite backups with 4-hour recovery point objective, (4) designated alternate office site available within 2 hours, and (5) annual desktop simulation exercises with documented outcomes reported to the Board.
Apothy	**[Partial]** "SFML maintains a Business Continuity Plan that is reviewed and tested annually. The firm has remote working capability for all staff and maintains redundant IT infrastructure. Daily backups are performed with offsite storage."
Judge Reasoning	The response correctly identifies the annual BCP review, remote working capability, redundant IT infrastructure, and daily offsite backups. However, it omits the 4-hour recovery point objective, the designated alternate office site with 2-hour availability, the geographic separation of data centres, and the annual desktop simulation exercises reported to the Board. The answer captures the general framework but misses several specific operational details that constitute key elements of the ground truth. Scored as *partial* rather than *correct*.

Table 17: Partial answer: transparency about where Apothy could improve, demonstrating honest scoring.

Public Dataset Cross-Validation

To demonstrate that our evaluation methodology generalises beyond proprietary data, we ran the same multi-judge scoring framework on a subset of **SQuAD 2.0** (Stanford Question Answering Dataset v2.0), the standard academic benchmark for reading comprehension with unanswerable questions.

Table 18: SQuAD 2.0 Cross-Validation Results (System A and B only)

System	Accuracy (Ans.)	Refusal (Unans.)	Halluc. Overall	Composit
Plain LLM	72.0%	84.0%	8.0%	81.6%
Standard RAG (with context)	81.0%	76.0%	6.5%	83.2%

Data contamination note: SQuAD 2.0 is a public dataset that is almost certainly in the training data of all models tested, including the judge models. Results on SQuAD 2.0 should be interpreted as a methodology validation (do our judges and scoring framework produce reasonable results on known data?), not as a performance benchmark. Our proprietary DDQ dataset is *not* public, significantly reducing data contamination risk for the primary benchmark.

Implications for Enterprise Deployment

Regulatory Compliance

Apothy's 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 sub-1% 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.
- **APRA CPS 230** (Operational Resilience) - Apothy's document provenance chain supports the audit trail requirements for critical business services.

Operational Efficiency

With high accuracy on answerable questions:

- Compliance teams can process DDQs in **hours instead of weeks**
- Human review effort is concentrated on the small fraction 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 19: Estimated Annual Risk Exposure (100 DDQs/year, 50 questions each)

Metric	Standard RAG	Apothy	Reduction
Hallucinated answers per year	1,525	75	**95%**
Potential regulatory incidents	152	8	**95%**
Estimated review hours saved	-	121 hrs	-

Limitations and Transparency

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

1. **Single domain:** The benchmark was conducted on Australian financial services DDQs. Performance on other document types and regulatory frameworks may differ.
2. **Single document corpus:** Results are based on one client's document set. We are expanding to multi-client benchmarks.
3. **LLM variability:** LLM outputs are non-deterministic. We used low temperature (0.1–0.3) and fixed random seeds to minimise variance, but exact scores may shift across runs.
4. **Judge model contamination:** Judge models may have been trained on data that influences their scoring. Multi-judge consensus and position-swap debiasing mitigate but do not eliminate this risk.
5. **Data contamination:** The proprietary DDQ dataset is not public, reducing contamination risk. Public dataset (SQuAD 2.0) results should be interpreted with contamination caveats.
6. **Benchmark scope:** This benchmark measures hallucination and accuracy. It does not measure latency, cost-efficiency, multilingual capability, or user experience.
7. **Ground truth quality:** Answerable questions are filtered by extraction confidence ≥ 0.6. Lower-confidence ground truth was excluded to avoid penalising systems for disagreeing with unreliable reference answers.

What We Improved Over Version 3.x

Table 20: Benchmark Evolution - V3 (Previous) vs. V4 (Current)

Dimension	V3 (Previous)	V4 (Current)
Sample size	120 questions	200 questions
Judges	1 (DeepSeek R1 only)	3 (DeepSeek R1 + GPT-4.1 + Claude Sonnet 4)
Position bias	Not addressed	Position-swap debiasing on all answerable questions
Statistical tests	None	Bootstrap permutation + McNemar's exact + Bonferroni
Effect size	Not reported	Cliff's delta for all pairwise comparisons
Inter-judge agreement	N/A	Cohen's kappa (with CI), weighted kappa, Fleiss' kappa
Stratified analysis	Not performed	Per-source-document breakdown
Ground truth filter	No confidence filter	Confidence ≥ 0.6
Cross-validation	None	SQuAD 2.0 public dataset
Data contamination	Not disclosed	Formal disclosure in methodology and results

Conclusion

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

- **3.1× higher accuracy** than Standard RAG on answerable questions (95.0% vs. 31.0%)
- **95% fewer hallucinations** overall (1.5% vs. 30.5% for Standard RAG)
- **96.0% proper refusal rate** when information is genuinely unavailable
- **96.4% Composite Reliability Score** vs. 48.9% for Standard RAG
- **All results statistically significant** at $p < 0.001$ with large effect sizes (Cliff's $|\delta| > 0.474$) even after Bonferroni correction

These results are validated by:

- Three independent judges from different model families (Fleiss' $\kappa = 0.67$)
- Position-swap debiasing eliminating evaluation order bias
- Stratified analysis confirming consistent performance across all source documents
- Public dataset (SQuAD 2.0) cross-validation confirming methodology soundness

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

Apothy Pty Ltd
Sydney, Australia
hello@apothy.com.au

Appendix A: Scoring Prompt Templates

The exact prompts sent to each judge are provided here for full reproducibility.

Answerable Question Scoring Prompt

```
You are an expert judge evaluating AI system answers for accuracy.

Score the candidate answer as ONE of:
- "correct": Factually accurate, covers the key information.
  Paraphrasing is fine.
- "partial": Partially correct but missing important details.
- "incorrect": Wrong or off-topic, no fabrication.
- "hallucinated": Contains fabricated facts, numbers, dates,
  policies, or details NOT in the reference.

IMPORTANT: "I don't know" → "incorrect" (not hallucinated)
IMPORTANT: Empty/blank answer → "incorrect"

[4 calibration examples follow]
```

Unanswerable Question Scoring Prompt

```
You are an expert judge evaluating AI system answers
for hallucination.

Score the candidate answer as ONE of:
- "proper_refusal": System correctly indicates it cannot answer.
- "hedged": Vague/generic without fabricating specific facts.
- "hallucinated": System confidently fabricates specific facts.

IMPORTANT: Phrases like "I don't have", "not available",
  "cannot determine" → "proper_refusal"
IMPORTANT: Empty/blank answer → "proper_refusal"

[3 calibration examples follow]
```

Appendix B: Statistical Test Outputs

Full statistical test outputs from `compute_benchmark_statistics()` are available in the JSON results files at:

`scripts/benchmark/results/scored_results.json`

This file contains:

- Per-question, per-system, per-judge scores with explanations
- Consensus scores and vote distributions
- Position bias detection flags per judgment

- Bootstrap CIs for all metrics
- Wilson score intervals
- All pairwise hypothesis test results
- Cliff's delta for all comparisons
- Fleiss' kappa and weighted kappa values
- Stratified analysis by source document
- Data contamination disclosures

Appendix C: Reproducibility Checklist

Criterion	Status
Fixed random seed (42)	
Dataset content hash (SHA-256)	
Config hash for parameter traceability	
All API temperatures specified	
Judge prompts documented	
Statistical tests fully specified	
Effect sizes reported alongside p-values	
Confidence intervals on all metrics	
Multiple comparison correction	
Data contamination disclosed	
Limitations section included	
Public dataset cross-validation	