OLD VS NEW AGENTIC ANSWER SYSTEM PERFORMANCE

Total Queries Analyzed: 100
Model Used: azure_ai/gpt-oss-120b

EXECUTIVE SUMMARY

The comparison between old and new agentic answer systems reveals a marginal improvement in the new system, with overall performance increasing by 1.52%. However, this improvement comes with significant trade-offs in certain areas.

Key Performance Metrics:

METRIC	OLD	NEW	IMPROVEMENT
Factuality	5.12	5.14	+ 0.02
Completeness	4.09	4.21	+0.12
Overall Average	4.61	4.67	+0.07

Analysis of the Old System

What the Old System Did Well

The old system remains superior for questions that require clear structure and organization.

- It was more disciplined in providing answers in a strict categorical format without adding irrelevant information.
  Q."Based on all recent communications, what are the primary categories of technical and compliance risks we are facing with the RBI Data Localization program, and what are the root causes?"

- It was often better at summarizing concrete mitigation steps and solutions.

  Q."Synthesize the issues related to the real-time risk scoring engine. What is the core problem, what is its direct impact on compliance, and what mitigation strategies are being pursued?"

What Was Wrong with the Old System

The old system's primary weakness was its poor information retrieval capability.

- It frequently failed to find specific documents and would incorrectly claim the information did not exist.

  Q."According to the 'CBS Migration Bottleneck Analysis' file, what were the two specific actions recommended to mitigate WAL commit latency?"

- Its search accuracy was low for specific items like JIRA tickets, often returning completely incorrect lists.

  Q."Provide a list of all JIRA tickets mentioned across emails and Slack messages that are related to the RBI Data Localization Audit Readiness Program."

- It was less effective at understanding the nuances and context from informal communications.

Analysis of the New System

What the New System Does Well

The new system is better at questions that require finding very specific information and understanding context.

- It excels at locating specific files and accurately extracting precise technical details from within them.

  Q."According to the 'CBS Migration Bottleneck Analysis' file, what were the two specific actions recommended to mitigate WAL commit latency?"

- It is more accurate when asked to find and list specific items like JIRA tickets.
- It is better at analyzing informal text (like Slack messages) to understand context and sentiment.

-"How are we managing the relationship and perception with NPCI regarding all the recent technical issues impacting the joint audit program?"

What Is Going Wrong in the New System

The new system's main weakness is a lack of focus and a tendency to include irrelevant information.

- It often adds related but incorrect or unasked-for categories to its answers, which hurts its factual accuracy.

  -"Based on all recent communications, what are the primary categories of technical and compliance risks we are facing with the RBI Data Localization program, and what are the root causes?" (The new system missed two correct risk categories and added two incorrect ones).

- It sometimes omits required items from a list, even while providing extra, unnecessary details.
- Both systems share a critical flaw: they sometimes incorrectly claim specific information is unavailable when it actually exists in the source data.
  -"What are the two specific architectural options being debated for the multi-currency settlement process in the meeting for PAY-5422?" (The new system, like the old one, incorrectly claimed this information was unavailable).

DETAILED PERFORMANCE BREAKDOWN

1. **FACTUALITY ANALYSIS**

New System Strengths: Better handling of JIRA ticket listings More accurate NPCI relationship management details

Old System Strengths: More accurate RBI risk categorization Better performance on specific fact-based queries More reliable handling of architectural options and error codes

Critical Factuality Issues: Both systems struggle with: Specific error codes (E-408: Invalid_Risk_Payload).Exact architectural options for multi-currency settlement Precise latency figures and technical specifications

2. **COMPLETENESS ANALYSIS**

New System Improvements: 12% better completeness scores overall More comprehensive coverage of complex multi-faceted questions Better inclusion of relevant context and background information

Areas Still Lacking: Missing specific incident details and email IDs Incomplete coverage of all required risk categories Partial responses to multi-part questions

3. **QUESTION TYPE PERFORMANCE (examples)**

High-Performing Question Types (New System Better):

1. JIRA Ticket Listings Example: "List all JIRA tickets related to RBI Data Localization" New Score: 6.5 vs Old Score: 5.0 Why New Wins: Captured 4 ground-truth tickets vs 0 for old system Improvement: +30% better factuality due to overlap with required tickets
2. CBS Migration Bottleneck Analysis Example: "What were the two specific actions to mitigate WAL commit latency?" New Score: 10.0 vs Old Score: 1.0 Why New Wins: Perfect factual accuracy with complete coverage Improvement: Correctly identified synchronous_commit changes and IOPS requirements
3. NPCI Relationship Management Example: "How are we managing relationship with NPCI regarding technical issues?" New Score: 8.5 vs Old Score: 6.5 Why New Wins: Better capture of specific communication patterns and stakeholder interactions

High-Performing Question Types (Old System Better):

1. Risk Categorization Example: "Primary categories of technical and compliance risks for RBI Data Localization" Old Score: 7.0 vs New Score: 5.5 Why Old Wins: Captured all 4 ground-truth risk categories; new system only captured 2 Issue: New system added unrelated KYC and OAuth categories
2. Cross-Border Payment Status Example: "Current situation with cross-border payment routing problem" Old Score: 5.0 vs New Score: 5.0 (tie, but old selected) Why Old Wins: Clearer status descriptions despite both missing core incidents
3. Real-Time Risk Scoring Engine Issues Example: "Synthesize issues related to real-time risk scoring engine" Old Score: 5.5 vs New Score: 5.0 Why Old Wins: Included more concrete mitigation steps aligned with ground truth

Problem Areas (Both Systems Perform Poorly):

1. Specific Architectural Options Both systems scored 2.0 on multi-currency settlement architecture options Issue: Both claimed information was unavailable despite ground truth providing clear options Impact: Critical for technical decision-making processes
2. Undocumented Error Codes Both systems failed to identify E-408: Invalid_Risk_Payload Scores: Old: 1.0, New: 1.5 Issue: Both contradicted ground truth by claiming no undocumented codes exist

SYSTEMIC PATTERNS AND ROOT CAUSES

1. **Information Retrieval Patterns**

New System Characteristics: More conservative in claiming definitive facts Better at providing alternative information when exact match not found Tends to add more contextual JIRA tickets and references

Old System Characteristics: More likely to provide definitive answers (sometimes incorrectly) Better at structured categorization More focused responses with less extraneous information

2. **Error Patterns**

Common Failure Modes:

1. Information Availability Claims: Both systems incorrectly claim information doesn't exist
2. Category Substitution: Adding related but incorrect categories

3. Metric Confusion: Providing similar but incorrect numerical values
4. Scope Drift: Including unrelated but topically similar information
5. Quality Indicators

Positive Indicators for New System: More JIRA ticket references Better acknowledgment of uncertainty More comprehensive background context

Positive Indicators for Old System: More structured categorical responses Better alignment with specific technical requirements Clearer distinction between different types of issues

DETAILED QUESTION EXAMPLES WITH ANALYSIS

Example 1: JIRA Ticket Listing Success (New System Win)

Question: "Provide a list of all JIRA tickets mentioned across emails and Slack messages that are related to the RBI Data Localization Audit Readiness Program."

Ground Truth Tickets: PAY-5422, PAY-6199, PAY-1477, PAY-7154, PAY-1036, [... 35 more tickets]

Old System Response: Listed completely different tickets (PAY-2988, PAY-2670, etc.) - 0% overlap New System Response: Included PAY-4713, PAY-1241, PAY-2405, PAY-6180 - 11% overlap

Analysis: New system's retrieval was more accurate, though still incomplete. This demonstrates improved search and matching capabilities.

Example 2: Risk Categorization Failure (Old System Win)

Question: "What are the primary categories of technical and compliance risks we are facing with the RBI Data Localization program?"

Ground Truth Categories:

1. Data Residency Breaches
2. Performance & Stability Issues
3. Data Integrity & Reconciliation Gaps
4. Security & Compliance Violations

Old System: Covered all 4 categories with appropriate root-cause explanations
New System: Only covered 2 categories, added unrelated KYC automation and OAuth token failures

Analysis: Old system maintained better categorical discipline and comprehensive coverage.

Example 3: Perfect Execution (New System Win)

Question: "According to the 'CBS Migration Bottleneck Analysis' file, what were the two specific actions recommended to mitigate WAL commit latency?"

Ground Truth:

1. Change synchronous_commit = 'local' for cbs-writer-v3
2. Upgrade to Provisioned IOPS (io2 Block Express) volumes with 20,000 IOPS baseline

Old System: Claimed document doesn't exist (Score: 1.0)
New System: Provided both actions with perfect accuracy and additional context (Score: 10.0)

Analysis: Demonstrates new system's superior document retrieval and technical detail extraction.