Hyperflow Customer Support Knowledge Base

RAG audit report

Three-pillar diagnostics: document quality · retrieval performance · answer forensics

Prepared by
BlindspotLabs

Date
May 2026

Audit ID
BSL-2026-047

Version
1.0

Diagnostic scorecard

PILLAR 1	PILLAR 2	PILLAR 3
Document quality	Retrieval	Answer quality
54 / 100	74 / 100	69 / 100
C	B	C
c · target: 75+	b · target: 80+	c · target: 80+

OVERALL RAG READINESS **66 / 100 · C- · Significant improvement potential**

Scale guide: <50 D · critical | 50–69 C · needs improvement | 70–84 B · production-ready | 85+ A · excellent · Minimum B (75+) recommended for production chatbo

Executive summary

We assessed the performance of the chatbot running on the Hyperflow Customer Support Knowledge Base across three independent pillars. The audit findings highlight the following four key observations:

1. **Document quality is the primary bottleneck — Topic mix and Q-without-answer are critical.**
   The knowledge base scores 54/100 on document quality (C-range). Two metrics are in critical territory: topic mix risk (0.61, target ≤0.30) and Q-without-answer ratio (38.2%, target ≤5%). These directly cause retrieval_miss and wrong_doc_retrieved failures in Pillar 3.

2. **Keyword search unexpectedly outperforms Hybrid and Vector on this corpus.**
   All 10 Azure configurations were tested; Keyword k=3 Hard-grounded leads with 93.5% answer correctness and 3.2% severe fail rate. Vector retrieval performs worst (12.9% severe fail on k=3), likely because the knowledge base is terminology-heavy and benefits from exact-match retrieval.

3. **The real answer failure rate is 16.1% — significantly better than RAGAS suggests.**
   RAGAS flagged 16 of 31 runs (51.6%) as low-scoring. Our Answer Diagnostics module validated these: only 5 are real failures; the rest are false negatives (paraphrase mismatch, grounded expansion). The severe failure rate is 6.5% — low and manageable.

4. **60% of failures are retrieval-level, 40% are prompting-level — both addressable.**
   Retrieval failures (wrong_doc_retrieved, retrieval_miss) will decrease as Pillar 1 recommendations are implemented. Prompting failures (missing_relevant_detail, over_strict_grounding) are calibratable within a 1–2 week engineering cycle.

PILLAR 1

Document quality

54 / 100

Five metrics selected for their direct, proven impact on retrieval quality and answer correctness.
Each metric includes its causal chain: how the specific issue propagates from document structure to retrieval failure to wrong answer.

Document readiness metrics

Score: 54/100 overall · 642 articles in corpus · 5 metrics evaluated

METRIC + WHY IT MATTERS + IMPACT CHAIN	VALUE	TARGET	WORST DOCUMENT	ACTION
Topic mix risk score How many distinct topics are blended within a single article or chunk. High score → embedding vector blurs across topics → wrong chunk retrieved for related queries → answer drift / hallucination	0.61	≤ 0.30	Getting Started Guide (0.91)	CRITICAL
Q without answer ratio Questions posed in the text whose answers live in a structurally separate block. Chunking splits Q from A → retrieval returns Q-chunk only → LLM sees the question, not the answer → forced hallucination	38.2%	≤ 5%	Billing FAQ (100%)	CRITICAL
Title – question alignment Whether article titles match the vocabulary and format of real user queries. Title mismatch → weak embedding overlap with user query → correct article systematically underranked → retrieval miss	0.44	≥ 0.60	Platform Overview (0.11)	MEDIUM
Flesch-Kincaid grade level Text complexity in US school grade equivalents. Grade 19+ = above university level. Complex sentences → LLM loses specific details during synthesis → key numbers, thresholds and conditions dropped → incomplete answer	24.3	≤ 8	Enterprise SLA Terms (38.1)	HIGH
Context self-sufficiency Whether a retrieved chunk can be understood without reading surrounding text. Score above target → most chunks are self-contained → retrieval can succeed at chunk-level → positive signal for answer quality	0.71	≥ 0.60	API Auth Step 2 (0.03)	GOOD

How document quality connects to Pillar 2 and Pillar 3

- **Topic mix (0.61) → Pillar 3 wrong_doc_retrieved failures:**
  "Getting Started Guide" scores 0.91 topic mix — it covers account setup, product configuration, billing, and shipping in a single article. When users ask about billing, the mixed chunk is retrieved and ranked highly. The LLM generates from blended context, producing answers that mix unrelated product features. Splitting this article into focused units is expected to eliminate the wrong_doc_retrieved failure type.

- **Q-without-answer (38.2%) → Pillar 3 retrieval_miss failures:**
  "Billing FAQ" is the most extreme case: every question in the article is answered in a separate policy document. When users ask a billing question, the FAQ chunk is retrieved (perfect semantic match to the question) but contains no answer — only more questions. The LLM is forced to hallucinate or refuse. This pattern explains 2 of the 3 retrieval-level failures in Pillar 3.

- **Context self-sufficiency (0.71) is the one strong signal:**
  Above-target self-sufficiency means most retrieved chunks are coherent in isolation — the chunking strategy preserves logical units well. This is a positive baseline: improvements to topic mix and Q-without-answer will build on a structurally sound foundation.

PILLAR 2

Retrieval performance

74 / 100

We tested 10 Azure AI Search configurations with identical evaluation parameters (31 representative user questions, Azure OpenAI GPT-4.1 mini, Hard
Measurement uses the RAGAS framework jointly evaluating retrieval and generation quality. Experiments ranked by Answer correctness and Severe fa

Experiment leaderboard

EXPERIMENT	MODE	K	GROUNDING	FAITHFUL.	CTX REC.	ANS.CORR.	SEV.FAIL	ROOT CAUSE
★ Keyword k=3 Hard-grounded	Keyword	3	Hard	0.83	0.94	93.5%	3.2%	retrieval
Keyword k=5 Hard-grounded	Keyword	5	Hard	0.81	0.91	90.3%	6.5%	retrieval
Vector k=3 Hard-grounded	Vector	3	Hard	0.74	0.68	80.6%	12.9%	reasoning
Vector k=5 Hard-grounded	Vector	5	Hard	0.76	0.72	83.9%	9.7%	reasoning
Hybrid k=5 Hard-grounded	Hybrid	5	Hard	0.79	0.81	87.1%	6.5%	retrieval
Hybrid k=8 Hard-grounded	Hybrid	8	Hard	0.77	0.83	85.5%	9.7%	retrieval
Keyword k=3 Standard	Keyword	3	Standard	0.69	0.92	80.6%	12.9%	prompting
Hybrid k=5 Standard	Hybrid	5	Standard	0.72	0.80	77.4%	16.1%	prompting
◆ Keyword k=3 + Reranker OPT	Keyword	3	Hard	0.86	0.97	96.8%	0.0%	retrieval
Hybrid k=5 + Metadata filter	Hybrid	5	Hard	0.80	0.83	87.1%	6.5%	retrieval

Key observations

- **★ Keyword k=3 Hard-grounded is the production candidate (93.5% correctness, 3.2% severe fail):**
  Keyword search outperforms Hybrid and Vector across all quality dimensions on this corpus. This is a significant and counter-intuitive finding: the knowledge base is terminology-heavy (product names, feature labels, SLA abbreviations), where exact-match keyword retrieval outperforms semantic similarity. Switching to Hybrid or Vector would increase cost with no quality benefit.

- **Vector retrieval performs worst — 12.9% severe fail on k=3 (vs 3.2% for Keyword):**
  Vector search creates the most failures on this corpus. The likely cause: the embedding model over-generalizes on short, keyword-dense chunks, ranking semantically adjacent but topically wrong articles. This reinforces the recommendation against vector-first retrieval for this knowledge base type.

- **Hard vs Standard grounding makes a 12.9pp difference in severe fail rate:**
  Keyword k=3 Standard produces 12.9% severe fail (vs 3.2% Hard). This shows the current Hard grounding policy is well-calibrated for precision. Relaxing it would significantly increase hallucination risk. The over_strict_grounding failure in Pillar 3 is an edge case, not a systemic grounding problem.

- **◆ Reranker adds 3.3pp answer correctness and eliminates severe failures:**
  Keyword k=3 + Semantic Reranker achieves 96.8% correctness and 0.0% severe fail rate. This is the highest-quality configuration tested. Recommended as a medium-term optimization after Pillar 1 quick wins are completed and the baseline has improved.

Production recommendation:
Keyword k=3 Hard-grounded on the current corpus. Once Pillar 1 issues are addressed, Top1 is expected to move from 0.76 into the 0.82+ range.

PILLAR 3

Answer failure forensics

69 / 100

Based on RAGAS signals, 16 of 31 runs qualified as low-scoring. The BlindspotLabs Answer Diagnostics module applied two-stage LLM-judge validatio
(1) distinguishes real failures from RAGAS false negatives, (2) classifies real failures across the 5-layer RAG root cause model.

16 / 31	5 / 16	6.5%
RAGAS-flagged low-scoring runs (51.6%)	Real failures out of flagged 31.3% — rest filtered by validation	Severe failure rate 2 cases out of 31 — low

Real failure root cause distribution

Ordered by frequency — Keyword k=3 Hard-grounded strategy. N = 5 real failures.

ROOT CAUSE LAYER	COUNT	%	DOMINANT FAILURE TYPE	AVG SEVERITY
Retrieval	3	60%	retrieval_miss (2) · wrong_doc_retrieved (1)	0.61
Prompting	2	40%	missing_relevant_detail (1) · over_strict_grounding (1)	0.46
Context assembly	0	0%	—	—
Model reasoning	0	0%	—	—
Content gap	0	0%	—	—

Severity distribution

Minor 2 cases (40%) · small gap, answer still usable — low business risk

Moderate 1 case (20%) · partially correct, potentially misleading downstream

Severe 2 cases (40%) · completely wrong or answer refused — high business risk

Average severity score: 0.54 · Industry average: 0.45–0.55 range · Within normal range

Representative diagnoses

REAL FAILURE · MODERATE

Question: "How do I downgrade my subscription plan?"

Diagnosis: The answer correctly describes the subscription management flow but omits the critical 30-day advance notice requirement for plan downgrades. This detail is documented in the "Subscription Terms" article. A customer following the answer would attempt same-day downgrade and face unexpected billing — a direct business impact. Root cause: the system prompt did not instruct the model to surface edge cases and exceptions when describing processes.

Root cause: prompting / missing_relevant_detail · severity 0.54

REAL FAILURE · SEVERE

Question: "What payment methods do you accept for enterprise invoicing?"

Diagnosis: The corpus contains a dedicated "Enterprise Payment Methods" article listing bank transfer, SEPA direct debit, and purchase orders. The retrieval pipeline returned the generic "Payment FAQ" article instead. The model generated a list of consumer payment methods (credit card, PayPal) — completely incorrect for enterprise invoicing. The article title mismatch ("Payment FAQ" vs "Enterprise Payment Methods") is the root cause of this retrieval failure, directly connected to the 0.44 title-question alignment score in Pillar 1.

Root cause: retrieval / wrong_doc_retrieved · severity 0.76

REAL FAILURE · SEVERE

Question: "What is the SLA response time for critical priority support tickets?"

Diagnosis: The SLA table is documented in the corpus (4-hour response for critical tickets). The retrieval pipeline returned the correct article ("Enterprise SLA Terms") but the chunk contained only the table header row — the data rows were split into a separate chunk by the chunking pipeline. With the table header but no data, the model could not construct an answer. Under Hard grounding, it refused to respond. The user experienced total chatbot failure on a high-stakes question. Fixing: merge table header and data rows into a single chunk.

Root cause: prompting / over_strict_grounding · severity 0.71

CROSS-CUTTING ANALYSIS

Cost & latency

Retrieval configuration determines not only answer quality but also operational cost and response time.
The directions are reliable; precise absolute values require endpoint instrumentation.

CONFIGURATION FACTOR	COST	LATENCY	MECHANISM
Keyword / BM25 index	Low	Low	No query embedding; inverted-index lookup
Vector / dense index	Medium	Medium	Query embedding + ANN search
Hybrid index (vector + BM25)	High	High	Two retrievers in parallel + RRF fusion
Semantic reranker (add-on)	Medium+	Medium+	Cross-encoder re-scoring of retrieved top-k
Higher k (e.g. k=5 vs k=3)	↑	↑	More context tokens entering the LLM input
Metadata filter (add-on)	≈	≈	Pre-filters index; negligible overhead at scale
Grounding: Hard vs Standard	≈	≈	Prompt-only constraint; no retrieval cost impact

Key insight: on this corpus, the cheapest configuration is also the best

Keyword k=3 (BM25) is the lowest-cost, lowest-latency configuration tested — and it delivers the highest answer quality (93.5%). Hybrid retrieval, which is more expensive and slower, performs 6.4pp worse in answer correctness on this corpus. This is a double win: no quality trade-off is required to stay at the low end of the cost scale. The reranker (+3.3pp quality, Medium+ cost) is the only configuration that meaningfully improves on Keyword k=3 and represents the only cost increase worth considering.

Hyperflow Commerce context:

The recommended Keyword k=3 Hard-grounded configuration sits at the low end of both cost and latency axes. Pillar 1 improvements (title rewriting, Q-without-answer resolution) can raise answer quality from 93.5% toward 96%+ without additional operational cost — same configuration, better corpus. The reranker (+3.3pp, Medium+ cost) is the only configuration worth adding cost for, and is recommended as a medium-term step after the Pillar 1 gains are measured.

SYNTHESIS

Prioritized recommendations

Based on findings from all three pillars, ordered by priority and feasibility. Each item is tied to a specific metric or failure pattern from the audit.

A. Quick wins

1–2 weeks · measured impact

1. **Rewrite the top 50 article titles to FAQ-style. (Title–question alignment: 0.44 → target 0.60+)**
   Focus on articles with alignment score below 0.40: Platform Overview (0.11), Getting Started Guide (0.18), Subscription Management (0.22), and 47 others. Simple rewrites like "Platform Overview" → "What does Hyperflow Commerce do?" directly improve embedding match. Expected: +0.10–0.15 alignment score, visible answer relevancy improvement in next benchmark.

2. **Resolve the Q-without-answer structure in 38.2% of articles — starting with "Billing FAQ". (Impact: eliminates 2 retrieval_miss failures**
   The "Billing FAQ" article poses 12 billing questions with no answers (100% Q-without-answer). Either embed the answers directly in the FAQ, or add cross-references that land in the same chunk. Second priority: the 8 other articles at 80%+ Q-without-answer ratio. Estimated 3–4 hours per article for a content editor.

3. **Split the "Getting Started Guide" into 4 focused articles. (Topic mix risk: 0.91 → eliminate worst offender)**
   This single article covers account setup, product configuration, billing overview, and shipping settings. Splitting into "Account Setup Guide", "Product Configuration Guide", "Billing Overview", "Shipping Setup" will directly address the wrong_doc_retrieved failure type and reduce the overall topic mix score from 0.61 toward the 0.30 target.

B. Medium-term

1–2 months

1. **Fix the SLA table chunking issue and revise the grounding policy. (Eliminates over_strict_grounding failure)**
   The SLA table in "Enterprise SLA Terms" must be kept as a single chunk — header + data rows together. Additionally, a targeted grounding policy revision: allow the model to answer from partial table context by adding an explicit instruction to interpolate from available table structure. This eliminates the severe failure that causes total chatbot failure on SLA queries.

2. **Add explicit exception/edge-case instructions to the system prompt. (Reduces missing_relevant_detail failures)**
   The current prompt focuses the model on answering the main question but does not instruct it to surface conditions, thresholds, and exceptions. Adding a prompt instruction like "always note limitations, conditions, and exceptions when they are documented in the context" is expected to reduce the missing_relevant_detail failure type significantly.

3. **Introduce a semantic reranker. (Expected: 93.5% → 96%+ answer correctness)**
   Experiment #9 shows Keyword k=3 + Reranker achieves 96.8% correctness and 0.0% severe fail. Implement after completing the A. package so the baseline improvement and the reranker contribution can be measured separately. Azure AI Search Semantic Reranker is the recommended implementation.

C. Strategic

3–6 months

1. **Simplify "Enterprise SLA Terms" and "Billing Terms" using LLM-assisted rewriting. (Flesch-Kincaid: 24.3 → target ≤8)**
   These two articles alone account for 60% of the Flesch-Kincaid overshoot. LLM-assisted simplification: feed the original text to GPT-4 with the instruction to rewrite at Grade 8 level, then have the content team review for accuracy. Target: sentence length ≤25 words, complex word ratio ≤12%.

2. **Continuous audit benchmark in the deployment pipeline.**
   Automate the 31-question regression suite as a CI/CD gate on knowledge base changes. Prevents well-intentioned content edits from silently degrading the answer correctness gains achieved here. BlindspotLabs can supply the benchmark tooling and threshold alerting.

What not to change right now

Do not switch from Keyword to Hybrid or Vector retrieval — the data clearly shows Keyword outperforms both on this corpus. Do not invest in new embedding models or index rebuilds before completing the A. package. The primary leverage is in the document and prompting layers. Retrieval improvements are constrained by document quality; fix the foundation first.

Proposed next step:

A 30-minute working session to align on the A. package priorities, followed by delivery within 1–2 weeks. The current composite score of 66/100 can realistically be raised to 78+ within 1–2 months by completing the B. package — entering the B (production-ready) range.

Methodology:

Audit: n=31 questions · k=3 · Azure OpenAI GPT-4.1 mini · Hard grounding · 642 articles in Azure AI Search. Document quality: BlindspotLabs Advanced Audit, 5 key metrics from a 50-metric evaluation framework. Retrieval: RAGAS framework. Answer Diagnostics: Claude Opus judge, 5-layer RAG taxonomy, two-stage false-negative filtering. Full dataset and failure log available on request.