AI-Powered Document Intelligence for Insurance & Compliance

Reducing Document Review Time from Minutes to Seconds

---

1. Executive Summary

Insurance, cybersecurity, and compliance teams rely heavily on long, dense PDFs - coverage forms, SOC 2 reports, NIST frameworks, cyber endorsements, and audit evidence. These documents are difficult to search, inconsistent across carriers, and slow down decision-making.

To address this, BIS Advisors developed and deployed an **AI-powered document intelligence system** capable of reading complex policy and compliance documents and returning **grounded, citation-backed answers in a few seconds**.

Across evaluations on both internal materials and real insurance/compliance PDFs, the system consistently delivered:

- **80–90% reduction in lookup time**
- **Accuracy in the low-to-mid 90% range**
- **Zero unsupported claims** when strict grounding is enforced
- **Consistent answers between reviewers**
- **Hybrid deployment across on-prem RTX 4090 and GCP L4**
- **Predictable, low compute cost per question**

While AI does not replace actuarial, legal, or underwriting judgment, it dramatically reduces the manual work required to find information inside complex documents.

---

2. Industry Challenge

Insurance and compliance workflows depend on documents such as:

- 30–200+ page cyber, property, liability, and ESG policies
- SOC 2 Type I/II reports
- NIST 800-53 and 800-171 frameworks
- Cyber endorsements, exclusions, conditions
- Regulatory guidance
- Audit evidence and supporting documentation

These documents often contain:

- Dense technical language
- Repetitive cross-references
- Inconsistent formatting across carriers
- OCR-heavy or scanned pages
- Definitions located far from the clauses they influence

A single coverage or control question may require **20–40 minutes** of manual searching across multiple PDFs.

This leads to:

- Slower underwriting turnaround
- Inconsistent interpretation
- Higher compliance and audit risk
- Analyst fatigue and cognitive load

As one analyst put it:

> “I spend a third of my day just trying to find where things are in these PDFs.”

---

3. Solution Overview

BIS Advisors implemented a domain-optimized **Retrieval-Augmented Generation (RAG)** system designed specifically for insurance and compliance work.

System Capabilities

- Parses and indexes large, complex PDFs
- Retrieves the most relevant clauses, sections, and controls
- Generates answers grounded strictly in retrieved text
- Returns responses in **2–7 seconds**
- Provides citations to exact policy sections or SOC/NIST controls
- Handles multi-document libraries and cross-document questions
- Responds with **“Not present in the document.”** when appropriate

Technical Architecture

Component	Implementation
Language	Python
LLM	Llama 3.1 Instruct 8B (self-hosted)
Inference Engine	vLLM
Embeddings	Llama 3.1 Embeddings

Component	Implementation
Vector DB	FAISS
Deployment	GCP L4 + on-prem RTX 4090
Guardrails	Grounded-answer prompting + refusal when unsupported
Pipeline	Chunk → Embed → Retrieve → Rank → Answer

Engineering Observations (Generalized)

- A clean Linux environment provides the most stable on-prem inference experience.
- vLLM performs best when GPU memory utilization remains within approximately 80–85%.
- Retrieval accuracy depends heavily on PDF parsing quality (OCR, tables, headers).
- Optimal chunking and overlap vary by document type (policies vs SOC/NIST vs endorsements).

These real-world nuances are where production-ready RAG systems differentiate themselves.

---

4. Evaluation Methodology

To ensure defensible results, we implemented a structured evaluation strategy:

- ~120 total questions
- Roughly half insurance, half SOC/NIST
- Scored using **Correct / Partially Correct / Incorrect**
- All answers manually reviewed by a domain-knowledgeable evaluator
- Grounding checked against retrieved text
- Latency measured across repeated runs on **GCP L4** and **on-prem RTX 4090**

This produced realistic, operationally meaningful metrics - not lab-optimized benchmarks.

---

5. Performance Results

Latency (L4 and 4090)

Across all question types:

- **Average latency on L4:** 5–7 seconds
- **Range:** ~0.5 to 15 seconds depending on context depth and answer length
- **On-prem 4090:** ~3.5× faster than L4

Latency was influenced more by retrieval depth and answer length than by document type.

---

6. Deployment Considerations

GPU VRAM & Model Stability

Optimal settings for stable inference:

- Keep GPU utilization in the **80–85% range**
- Very long answers (>2–3k tokens) increase latency noticeably
- Clean Linux installations ensure consistent GPU behavior

On-Prem vs Cloud Options

On-Prem RTX 4090

- Fast, predictable performance
- Minimal incremental cost
- Full data control
- Ideal for daily workloads

GCP L4

- Easy to deploy and scale
- Handles 24GB models well
- Suitable for variable or distributed workloads
- Estimated business-hours cost: **$400–$500/month**

Hybrid Deployment

Most teams benefit from a hybrid setup:

- On-prem for routine use
- Cloud for burst traffic or after-hours support

---

7. Cost Analysis

Self-Hosted Compute Cost

Example (scheduled GCP L4 instance):

- ~$400/month for 8–10 hrs/day availability
- At ~40k–50k monthly questions:
  $0.00001–$0.00002 per question

On-prem 4090 amortizes to near-zero incremental cost.

(Implementation and integration work not included.)

API Model Comparisons

Model	Approx Cost per Question
GPT-4o mini	~$0.0002
Claude 3.7 Sonnet	~$0.005
GPT-4.1	~$0.007

Recommended:
Use local Llama 8B for 80–90% of requests, escalate to premium API models for complex or ambiguous tasks.

---

8. Business Impact

Before AI

- 20–40 minutes to answer many coverage or compliance questions
- Manual paging through multiple PDFs
- Re-reading exclusions, definitions, and endorsements
- High variance between reviewers

After AI

- **Seconds** to retrieve grounded, citation-backed answers
- Consistent interpretation across analysts
- Lower cognitive load
- **80–90% reduction** in lookup time

Value Across Teams

Underwriting

- Faster review of submissions
- Faster identification of conditions, deductibles, sublimits
- More consistent interpretation of ambiguous clauses

Claims

- Faster adjudication
- Improved consistency
- Stronger defensibility with citations

Compliance & Audit

- Rapid SOC 2 evidence lookup
- Reliable NIST control interpretation

- Better audit readiness and documentation

Known Limitations

- OCR-heavy PDFs reduce recall and require specialized handling
- Highly cross-referenced documents occasionally need manual confirmation
- Not intended to replace legal or actuarial review
- Requires tuning per document type

---

9. Change Management & Adoption

The technology performs well, but organizational adoption requires:

- Clear communication of what the system can and cannot answer
- Training on reading grounded answers
- Governance and usage policies
- Managerial support to encourage adoption
- Trust-building over the first weeks of use

In every deployment, **culture, not model performance, determines long-term success**.

---

10. Key Lessons Learned

Technical

- Domain-optimized chunking and embeddings matter more than model size
- Instruct models outperform base models for grounded QA
- Retrieval accuracy is heavily influenced by PDF parsing quality
- vLLM requires careful VRAM tuning
- WSL is not suitable for production GPU workloads

Operational

- Integrations drive the real business value
- Insurance and compliance content require domain-specific evaluation sets

Organizational

- Analysts must understand grounded answers
- Adoption requires deliberate change management
- Consistency improves dramatically with structured workflows

11. Conclusion

AI-powered document intelligence is no longer experimental.
It is already delivering measurable value across underwriting, claims, cyber, audit, and compliance.

With:

- **2–7 second grounded answers**
- **80–90% reduction in lookup time**
- **Accuracy in the low-to-mid 90% range**
- **Stable on-prem and cloud deployments**
- **Proven Python-based RAG architecture**

Organizations can reduce manual effort, improve decision-making speed, and increase operational consistency.

The combination of grounded retrieval, hybrid deployment, and domain-optimized tuning makes AI document intelligence a practical, repeatable tool for real insurance and compliance workflows.

About BIS Advisors

BIS Advisors helps insurance, finance, and other regulated industries deploy **practical, reliable AI systems** - with a strong focus on document intelligence, hybrid cloud/on-prem deployment, and cost-efficient self-hosted LLMs.

We build AI solutions that teams trust, adopt, and use every day.

https://bisadvisors.com
contact@bisadvisors.com