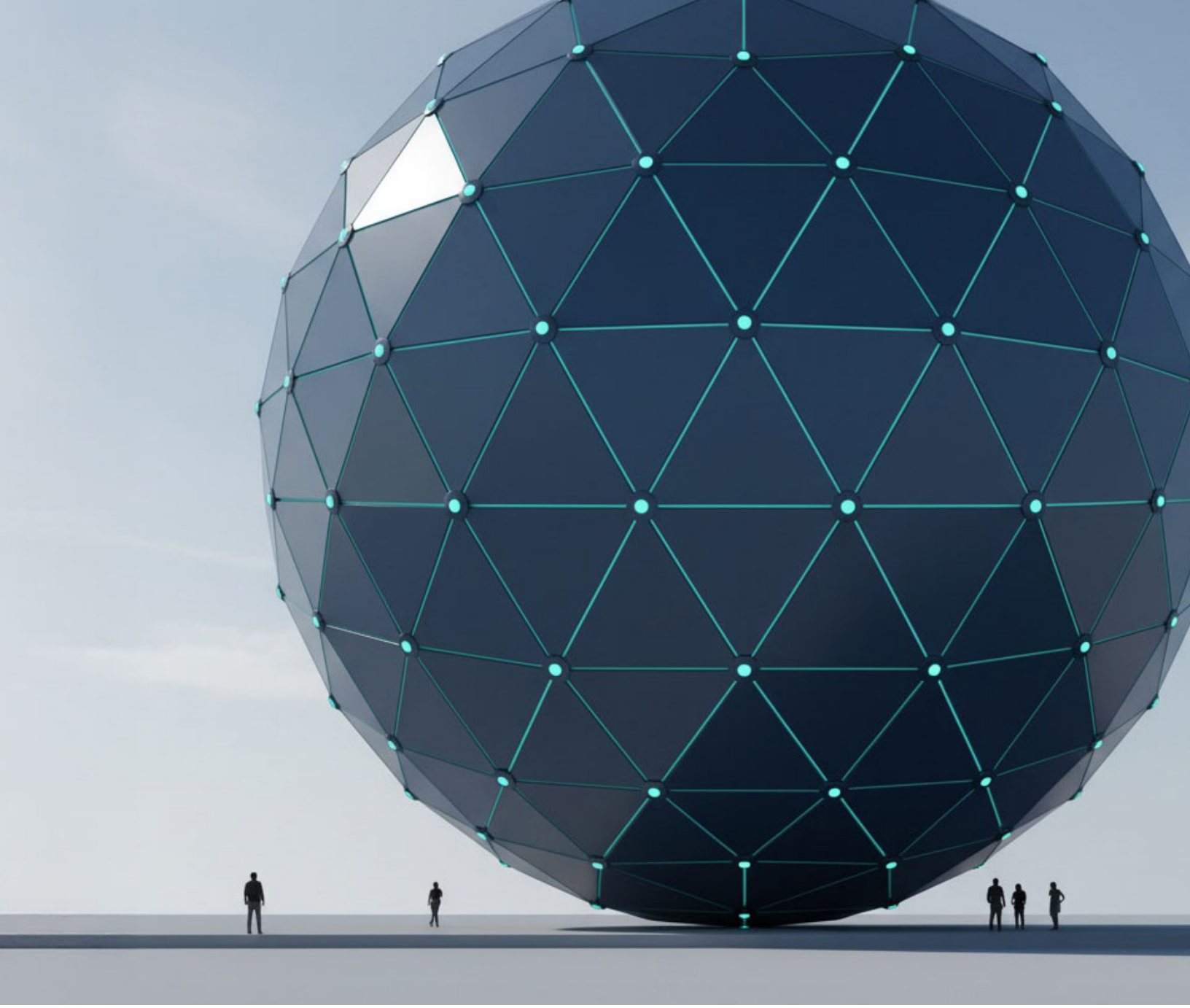


elastic

Erweiterte Vektorsuche für luftdichte Netzwerke

Der öffentliche Sektor befindet sich in einem kritischen Moment der Einführung von generativer KI (GenAI). Organisationen werden dazu gedrängt, größere KI-Initiativen zu starten, angetrieben durch landesweite Strategien sowie die Erwartungen der Stakeholder. Gleichzeitig sind GenAI und Suche miteinander verflochten. Nutzer wenden sich an interaktive GenAI, um schnelle Antworten auf ihre Anfragen zu erhalten, anstatt eine Suchmaschine zu nutzen, die eine Liste von Links liefert, die sie durchforsten müssen.

Die Erfüllung dieser Erwartungen im öffentlichen Sektor ist mit einigen Herausforderungen verbunden:

1. Es gibt Optionen für die Nutzung von GenAI, aber es gibt nicht genügend Entwickler. Organisationen müssen Ressourcen strategisch investieren, um die effizientesten Ergebnisse zu erzielen.

2. Organisationen des öffentlichen Sektors haben mit riesigen Datenmengen zu kämpfen. Sie benötigen eine skalierbare, schnelle und effiziente Wissensbasis, die als Grundlage für KI dient.

3. Daten sind nicht immer bereit für KI. Wichtige Informationen werden häufig in Formaten wie PDFs, Diagrammen und Videos gespeichert, aus denen sich relevante Informationen nur schwer extrahieren lassen.

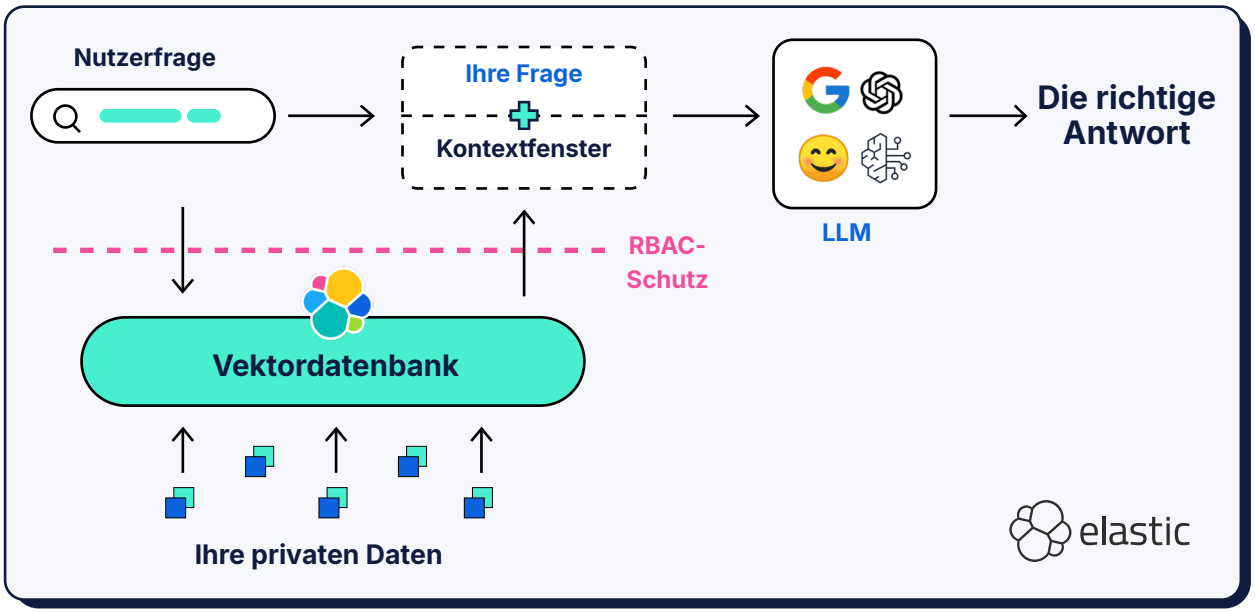
Verteidigungs- und Nachrichtendienste stehen vor der zusätzlichen Herausforderung, in einem luftdichten Netzwerk oder sogar mit einem technischen Kit zu arbeiten.

RAG: Ein effizientes Datenabrufmodell für KI

Viele Organisationen des öffentlichen Sektors finden es schwierig, große Sprachmodelle (LLMs) für die Art von Abfragen zu trainieren oder auch nur fein abzustimmen, die ihre Interessengruppen durchführen. Sie möchten ihre proprietären Daten sofort mit KI nutzen – auch wenn diese Daten komplex sind.

 | © 2025 Elasticsearch B.V. Alle Rechte vorbehalten.

Retrieval-Augmented Generation (RAG) ist eine skalierbare Lösung, die größere Herausforderungen im Zusammenhang mit begrenzter Zeit und Ressourcen sowie der Skalierung großer Mengen unterschiedlicher Daten bewältigt. Im RAG-Modell von Elastic wird vor der Übermittlung einer Nutzeranfrage an ein LLM zunächst unsere Vektordatenbank abgefragt, die über eine Vektorsuche relevanten Kontext aus unternehmenseigenen Daten ermitteln kann. Dieser Schritt bzw. diese Kontextebene reduziert das Datenuniversum, auf das sich ein LLM für seine Ausgabe stützt, erheblich. Agentische KI verändert die Natur dieses Modells noch weiter, da eine Abfrage mit KI-Agenten interagiert, die zwischen Nutzer und RAG sitzen.

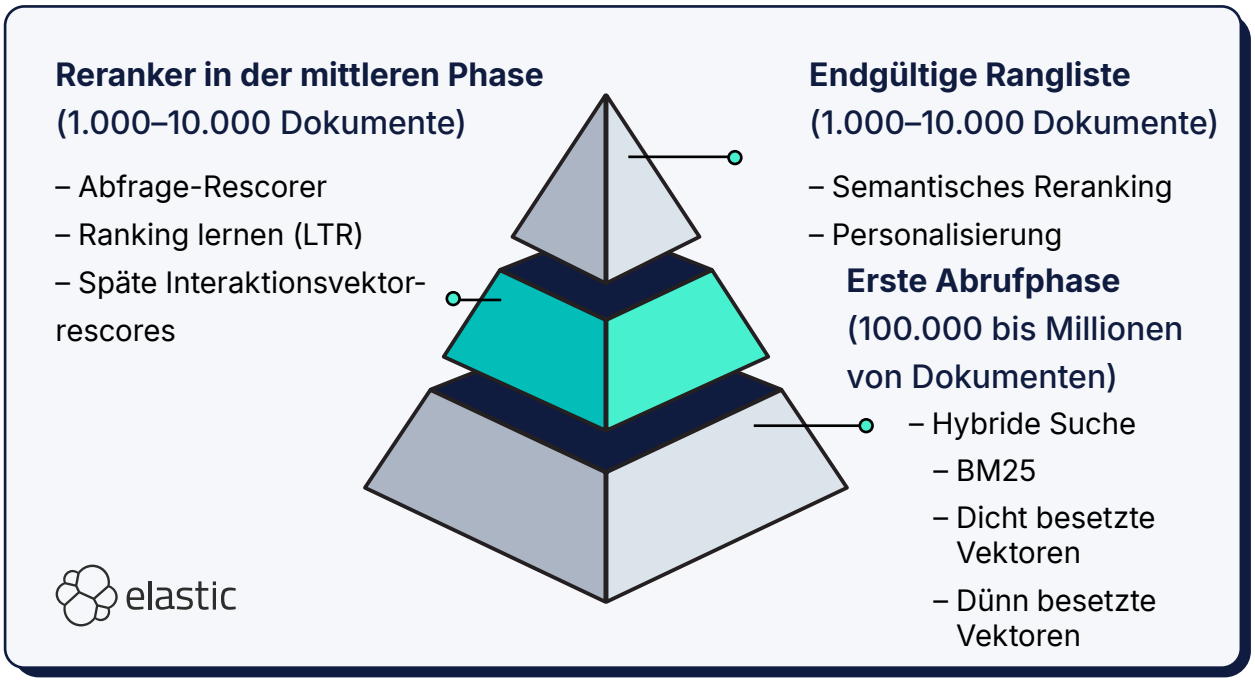


Mehrstufige Abfrage und multimodale Suche für komplexe Datentypen

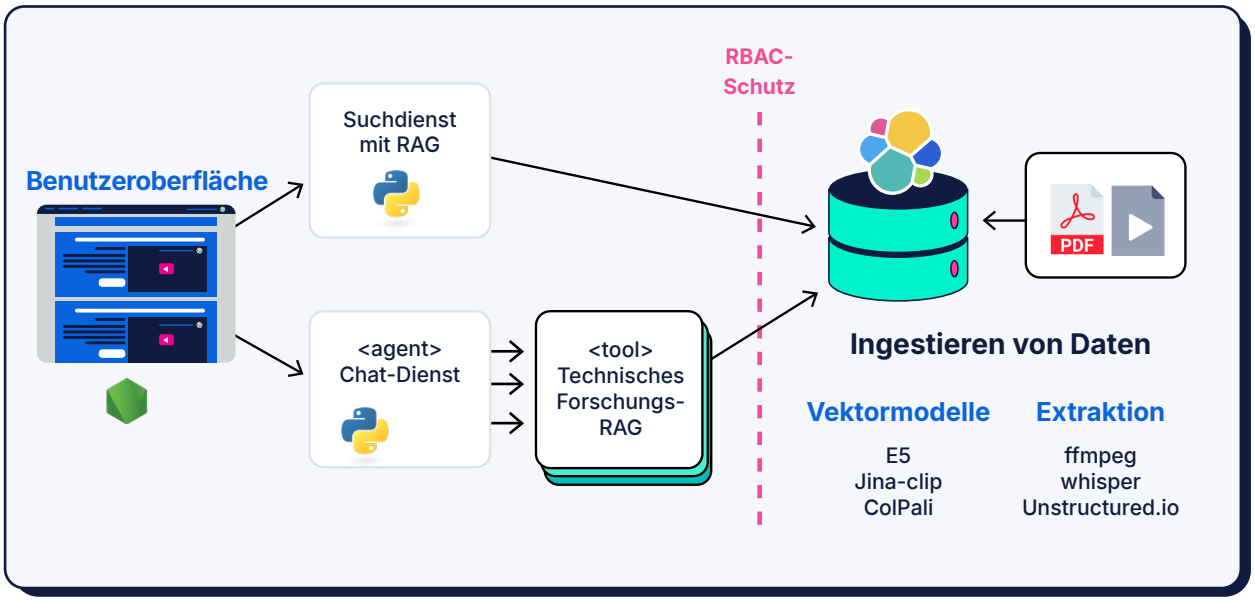
Um Organisationen des öffentlichen Sektors dabei zu helfen, ihre komplexen Daten für GenAI im Rahmen des RAG-Ansatzes nutzbar zu machen, nutzt Elastic **mehrstufige Abfrage** und **multimodale Suche**.

Die mehrstufige Abfrage zerlegt komplexe Anfragen im RAG-Prozess in kleinere Aufgaben, um eine endgültige Antwort zu liefern. Dieser Ansatz basiert darauf, dass für ungefähre Ergebnisse eine andere Technologie erforderlich ist als für spezifische Ergebnisse. Bemerkenswerterweise ist Elastic die erste Vektordatenbank, die eine vollständig unterstützte Air-Gapped-Option für semantisches Reranking bietet.

 | © 2025 Elasticsearch B.V. Alle Rechte vorbehalten.



Ebenso gewinnen multimodale Anwendungsfälle schnell an Bedeutung, da Organisationen nach intuitiveren und leistungsfähigeren Möglichkeiten suchen, Informationen über verschiedene Daten hinweg abzurufen, wie etwa PDF-Dateien mit Tabellen oder Videos mit Audio- und visuellen Komponenten. Die multimodale Suche kombiniert Text-, Bild-, Audio- und Videoeingaben und ermöglicht es KI-Anwendungen, Fragen schneller und mit dem richtigen Kontext zu beantworten.

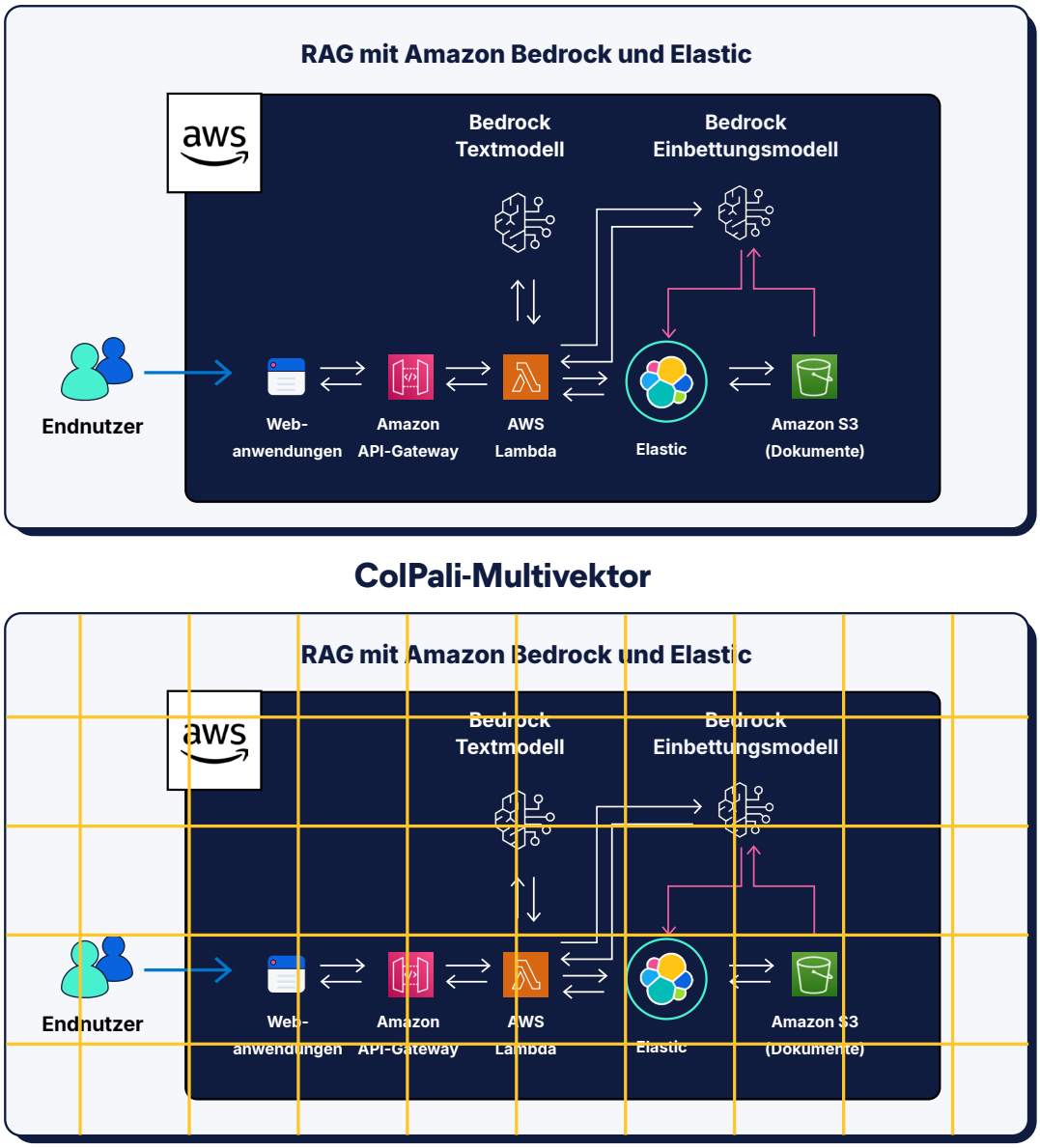


Mithilfe der obigen Struktur können Unternehmen visuelle Informationen finden, die tief in einer 120-minütigen .MOV-Datei eingebettet sind, oder einen Satz in einem Diagramm auf Seite 63 einer PDF-Datei von 1997. Diese Fähigkeit stellt sicher, dass komplexe, proprietäre Daten ganzheitlich durchsuchbar und für generative KI-Tools zugänglich

 | © 2025 Elasticsearch B.V. Alle Rechte vorbehalten.

sind, die sie für ihre Ausgaben nutzen können. Elastic ermöglicht außerdem die Suche nach Teilbildern, um sicherzustellen, dass selbst wenn ein Element nur einen relativ kleinen Teil eines Gesamtbilds oder Einzelbilds aus einem Video einnimmt, es aufgrund seiner Gesamtrelevanz für die Nutzerabfrage dennoch hoch in den Suchergebnissen eingestuft wird. Egal, ob etwas auf dem Audiotrack besprochen und nicht auf dem Bildschirm gezeigt wird oder umgekehrt, Elastic kann es trotzdem finden und hoch in den Suchergebnissen einordnen, wenn es für die Frage eines Nutzers relevant ist.

Erweiterte Bildersuche mit RAG



Betrachten wir ein konkretes Beispiel dafür, wie ein RAG-Ansatz relevante Informationen aus einer Bilddatei bereitstellen kann. Nehmen wir an, wir möchten nach „Diagrammen der Elastic-Architektur unter Verwendung von Amazon Bedrock“ suchen. Um die Informationen von der ersten Folie oben zu erhalten, müssten wir

 © 2025 Elasticsearch B.V. Alle Rechte vorbehalten.

wahrscheinlich eine Texterkennung (OCR) und Textextraktion durchführen und einige Datenflüsse erstellen, um die Pfade, die die Daten nehmen, zu verfolgen. Das kann kompliziert werden, insbesondere da jedes Dokument anders sein wird. Und wenn man lange Regierungsdokumente im PDF-Format mit zahlreichen Spalten und Tabellen durchsucht, die sich über mehr als acht Seiten erstrecken, wird dies zu einer deutlich größeren Herausforderung. ColPali kann dabei helfen, da es ein Raster aus dichten Vektoren über dem Bild erzeugt.

Anhand der obigen Bilder können wir erkennen, dass ColPali bestimmte Flächen des Bildes identifiziert hat, in denen Amazon Bedrock und Elastic sowohl in Text- als auch in Logoform vorherrschen. Das Elastic-Logo und die Verweise sind eigentlich recht klein, aber unsere Abfrage ist dennoch in der Lage, diese Folie zu finden und uns anzuzeigen, da die gesamte Bilddichtevektorisierung mit der Neusortierung der Bilder mithilfe des ColPali-Rangvektors kombiniert wird. Elastic hat diese Funktionen in das Produkt integriert, um diesen Prozess zu vereinfachen und es Entwicklern zu erleichtern, Anwendungen zu erstellen, die den Bedürfnissen ihrer Organisationen gerecht werden.

Elastic ist Ihre Lösung für offene, skalierbare Vektorsuche und RAG im öffentlichen Sektor

- Interoperable, LLM-unabhängige Unterstützung
- Ein RAG-Ansatz, der die Notwendigkeit für LLM-Schulungen und Feinabstimmung überflüssig macht
- Eine umfassend dokumentierte, API-gesteuerte Architektur, die vollständige Anpassung und Integration mit Tools in Ihrem Anwendungs-Stack ermöglicht
- Integrierte rollen- und attributbasierte Sicherheitskontrollen (RBAC und ABAC)
- Die weltweit am häufigsten heruntergeladene Vektordatenbank
- Inferenz-APIs zur Registrierung von KI-Diensten in jeder Umgebung
- Zwei sofort einsatzbereite Vektormodelle – E5 für dichte Codierung und ELSER für weniger dichte Codierung – und die Möglichkeit, jedes andere Einbettungsmodell Ihrer Wahl zu verwenden
- Vektorsuchfunktionen, die für Rankingqualität, Rechenressourcen und die Reduzierung der Abfragelatenz optimiert sind
- Bessere binäre Quantisierung (BBQ) zum Komprimieren von Vektoren, wodurch sie schneller durchsucht und verarbeitet werden können
- Eine skalierbare Grundlage für agentenbasierte KI, die erweiterte Retrieval-Modelle und Vektorsuchfunktionen nutzt

 © 2025 Elasticsearch B.V. Alle Rechte vorbehalten.

elastic

Die nächsten Schritte

Kontaktieren Sie unser technisches Team für den öffentlichen Sektor, um eine Demo der Vektorsuche mit Ihren Daten zu sehen: elastic.co/de/contact/public-sector.

