**TL;DR**

* **Your biggest cost drivers** won’t be the graph database alone—they’ll be (1) **LLM tokens** and (2) **retrieval/query capacity**, followed by (3) **OCR/ingestion** and (4) **private networking/bandwidth**. Azure prices each of those transparently; Neo4j prices the database capacity by **GB RAM-month**, which is simpler but can get pricey at high memory footprints.
* **Azure-native** shines for **in-country operations** (private endpoints, RBAC, predictable RU-based scaling, semantic rerank pricing, reserved capacity discounts). It’s easier to forecast costs in procurement. Vector search has **no extra charge** beyond your Search units.
* **Neo4j-centric** is compelling for **fast time-to-demo** and a strong GraphRAG developer story. Managed AuraDB is priced per **GB RAM per month** (e.g., **$65/GB-mo** for Professional, **$146/GB-mo** for Business Critical), which is easy to explain—but you’ll still need to budget for LLM, OCR, and (optionally) Azure AI Search if you want hybrid/semantic features.

**Reference pricing you can rely on (the “knobs”)**

**Azure Cosmos DB (Gremlin API)**

* Throughput: **$0.008 per 100 RU/s per hour** (single-write region). Autoscale is ~1.5×, multi-region writes ~2×. Storage ~**$0.25/GB-mo** in Microsoft’s example. Reserved capacity offers discounts.

**Azure AI Search**

* Billed **per Search Unit (SU) per hour**; scale with replicas/partitions. **Semantic ranker**: first **1,000 queries/month free**, then **$1 per 1,000**. **Vector search has no extra fee** beyond the SU. (Exact SU $/h varies by tier/region; S1/S2/S3 are common.)

**Azure OpenAI (models)**

* Token-metered. As a concrete anchor, the Microsoft announcement for GPT-4o (2024-08-06) shows **$2.50 / 1M input tokens** and **$10 / 1M output tokens** (model/rates evolve; always check your region).

**Azure AI Document Intelligence (OCR)**

* Commitment tier example: **$720 per 100,000 pages** for prebuilt Read/OCR; volume tiers exist. Containers are also available.

**Private networking / bandwidth**

* **Private Endpoint** ~**$0.01/hour** each; **data processed** via Private Link ~**$0.01/GB** in/out. Inter-region egress within Middle East & Africa list is **$0.08/GB** (if you cross regions).

**Neo4j Aura (managed)**

* **AuraDB Professional:** **$65/GB-month** (min 1 GB).
* **AuraDB Business Critical:** **$146/GB-month** (min 2 GB).
* Marketplaces also show **$0.09/hour per GB** (Pro) as a PAYG view. **Bloom** full features require Enterprise; pricing is not public. AuraDS (managed GDS) appears on marketplaces with separate hourly line items.

**What actually dominates the bill?**

1. **LLM tokens** (plan/route/rerank/summarize): often the #1 line item once you hit scale. Use small models where possible and strict grounding.
2. **Retrieval capacity**:
   * Azure path = **RU/s** (Cosmos) + **SU/h** (Search) with optional **$1/1k** semantic.
   * Neo4j path = **GB RAM-month** (AuraDB) and possibly **AuraDS** for analytics.
3. **OCR ingestion** (Document Intelligence).
4. **Private Link + bandwidth** if you isolate all services.

**Worked monthly scenarios (order-of-magnitude)**

These are not commitments—just **transparent math** you can adjust. I’m using 720 h/month for simplicity.

**A) “Pilot” (single region)**

* ~**30k RU/s** graph reads/writes; **1× S1** AI Search; **200k semantic queries**/mo; **150M model tokens**/mo; **50k OCR pages**/mo; **6 private endpoints**; ~**2 TB** privatelink data.

**Azure-native**

* Cosmos: 30,000 RU/s ⇒ 30,000/100×$0.008 = **$2.40/h** ⇒ **$1,728/mo** (+storage e.g., 200 GB ≈ **$50**).
* AI Search SU: S1 example **≈$0.336/h ⇒ $242/mo** (representative third-party listing; use your region’s calculator).
* Semantic ranker: (200k−1k free) ≈ **$199/mo**.
* Azure OpenAI: 100M input × $2.50 + 50M output × $10 = **$750/mo**.
* OCR: 50k pages ≈ **$360/mo** (half of the 100k pack).
* Private Link: 6 endpoints × $0.01/h × 720 ≈ **$43/mo**; data via Private Link ~4 TB total I/O × $0.01/GB ⇒ **$40–80/mo**.

**Pilot subtotal (Azure-native): ~USD 3.3–3.5k/mo.**

(1.73k Cosmos + 0.24k Search + 0.20k Semantic + 0.75k LLM + 0.36k OCR + ~0.12k networking)

**Neo4j-centric (replacing Cosmos with AuraDB; two options):**

* **Option 1 (no Azure AI Search; use Neo4j FTS + vectors):**

AuraDB size for pilot, say **16 GB**: 16 × $65 = **$1,040/mo**. LLM **$750**, OCR **$360**.

**Subtotal ~USD 2.1–2.3k/mo** (plus any egress/peering if services are split).

* **Option 2 (keep Azure AI Search for hybrid + semantic):**

Add the same Search line items (**$242** + **$199**) and Private Link (~**$80–120**).

**Subtotal ~USD 2.8–3.0k/mo.**

Pilot takeaway: At this scale, either stack lands in the same ballpark. If your indexed corpus fits in a modest AuraDB memory footprint, Neo4j can be slightly cheaper; if you need elastic RU bursting and reserved capacity, Cosmos can be comparable and more predictable on spend.

**B) “National” (2 regions, much higher usage)**

* **150k RU/s** sustained (graph), **3× S2** AI Search, **~5M semantic queries/mo**, **2B input + 500M output tokens/mo**, **1M OCR pages/mo**, **12 endpoints**, **~20 TB** Private Link data.

**Azure-native**

* Cosmos: (150,000/100×$0.008) = **$12/h/region** × 2 regions ⇒ **$24/h** ⇒ **$17,280/mo** (storage add e.g., 2 TB total ≈ **$1,000**). Reserved capacity can reduce this.
* AI Search: 3× S2 at **≈$1.344/h** ⇒ **$2,903/mo** (representative).
* Semantic ranker: **5,000k − 1k free ≈ $5,000/mo**.
* Azure OpenAI: 2B×$2.50 + 0.5B×$10 = **$10,000/mo**.
* OCR: 1M pages ⇒ 10×100k pack ⇒ **$7,200/mo**.
* Private Link: 12×$0.01/h×720 ≈ **$86/mo**; 20 TB I/O ≈ **$200–$400/mo** depending on pattern.

**National subtotal (Azure-native): ~USD 43–45k/mo.**

(Cosmos ~17.3k + Search ~2.9k + Semantic 5k + LLM 10k + OCR 7.2k + networking ~0.3k + storage ~1k)

**Neo4j-centric (replace Cosmos with AuraDB):**

* Memory footprint for this size often lands in **128–256 GB** for low-latency traversals of a large legal KG.
  + **128 GB Business Critical:** 128×$146 ≈ **$18,688/mo**;
  + **256 GB Business Critical:** **$37,376/mo**.
* If you keep Azure AI Search (for hybrid + semantic): add **~$2.9k + $5k + networking**; if you skip it, expect more LLM reranking tokens (cost shifts to OpenAI).
* LLM and OCR costs **don’t change**; they dominate either way.

National takeaway: At high scale, the LLM + semantic reranking lines rival or exceed the graph cost. Cosmos RU/s with reserved capacity can be very competitive; AuraDB can be competitive if your workload stays within a tight memory envelope. The wrong memory tier in Aura can outpace Cosmos; the wrong RU target in Cosmos can outpace Aura.

**Where each path tends to win on *cost governance***

**Azure-native (Cosmos + AI Search):**

* **Predictable scaling**: RU/s and SU/h are linear knobs, with **reserved capacity** discounts and clear multi-region math. Good for procurement and budget control.
* **Semantic ranker** is pay-as-you-go (free 1k/mo, then **$1/1k**); **vector** has **no additional charge**. Easy to dial up/down.
* **Private networking** is standardized (**$0.01/h endpoint; $0.01/GB**), and **in-country residency** options exist across services.

**Neo4j-centric (AuraDB ± AuraDS):**

* **Simple database pricing** by **GB RAM-month**, nice for explainability. **Bloom** full features available under Enterprise.
* **But** total TCO still depends on **LLM tokens**, **OCR**, and whether you **also** run Azure AI Search (for hybrid/semantic). Net: you often end up paying both **Neo4j** and some **Azure AI** line items.

**Practical guidance for your proposal**

1. **Budget model**: make the LLM + retrieval costs first-class in the deck (they dominate at scale). Use the token rates for GPT-4o as the anchor (update per model/region at procurement).
2. **Pick a graph cost curve and show the “crossover”**:
   * If your graph hot set fits in **≤64–128 GB**, **AuraDB** looks attractive. Above that, or if you need multi-region HA, **Cosmos (RU/s + reserved capacity)** can edge out on $/throughput.
3. **Be explicit about semantic reranking**: great quality/price lever—**$1 per 1k** queries beyond the free 1k. If you turn it off, expect higher LLM token costs to compensate in answer quality.
4. **Lock down networking early**: Private endpoints are cheap but numerous; show the math (endpoints × hours + GB processed) so security doesn’t become a surprise line item.
5. **Ingestion**: OCR is real money—quote the **$720/100k pages** and your expected monthly pages.

**Bottom line for UAE legal GraphRAG**

* For the **demo and early pilot**, Neo4j (AuraDB Pro + Bloom) keeps graph costs simple and the developer workflow fast. Total monthly will likely be **dominated by LLM + OCR**, not the graph itself.
* For **production at national scale with sovereignty and multi-region**, the **Azure-native route** (Cosmos Gremlin + Azure AI Search + Azure OpenAI + Private Link) gives you **predictable scaling knobs**, **line-item cost control** (RU, SU, $/1k semantic), and **discount levers** (reserved capacity). In many “big” configurations it ends up **cost-comparable or cheaper** than sizing AuraDB into high memory bands—**especially** if you’re already paying for AI Search and semantic reranking.