External Knowledge Integration Architecture for Autonomous Agents

Date: January 2026 (Reflecting latest 2025/26 ecosystem changes)
Context: Technical architecture for "ATLAS" agent (<$20/mo, <500ms latency)

1. Executive Summary

The "Agentic Retrieval" landscape has bifurcated in 2025. While high-end semantic search (Exa.ai) has become the gold standard for accuracy, its cost structure ($5–$25 per 1k requests) makes it prohibitive as a "default" search for budget-constrained agents. Conversely, the "raw" web has become increasingly hostile, with over 70% of top publishers now blocking AI scrapers via robots.txt.[1]

To meet the ATLAS budget (<$20/mo) and latency (<500ms) targets, we recommend a **Tiered Retrieval Architecture**:

1. **Ground Truth Layer (Free):** Use **deps.dev** and **GitHub Core API** for deterministic package/code lookups.
2. **Reading Layer (High Efficiency):** Use **Jina Reader** (10M free tokens) for URL-to-Markdown conversion.
3. **Search Layer (Hybrid):** Use **Tavily** (Free tier) for general queries and **Exa.ai** (Paid) strictly for complex code discovery.

---

2. Semantic Search & Code Discovery

2.1 Exa.ai (Neural Search)

Exa remains the most capable engine for "code-aware" search but represents the highest budget risk.

- **Pricing:**
  - **Fast / Auto Mode: $5.00 per 1,000 requests** (1–25 results).[2]
  - **Deep Mode:** $15.00 per 1,000 requests.[2]
  - **Deep Search + High Volume:** Jumps to **$25.00/1k** if requesting 26–100 results.[2]
- **Rate Limits:** 5 QPS (Queries Per Second) for Search; 50 QPS for Contents.
- **Latency:** <350ms P50 latency for the "Fast" endpoint.[3]
- **Recommendation:** Strictly cap usage. Do not use "Deep" mode for routine loops. Hard-limit results to <25 to stay in the $5 tier.

2.2 Tavily (Fallback & General)

Tavily is less semantically precise for code but offers a critical financial buffer.

- **Pricing: Free tier** includes 1,000 credits/month.[4] Overages are ~$8/1k.
- **Latency:** Generally >1,000ms as it aggregates sources in real-time, making it slower than Exa.[4]
- **Strategy:** Use Tavily for general knowledge (e.g., "What is the latest version of Next.js?") to conserve budget, and Exa for semantic code questions (e.g., "Implement A* search in Rust").

---

3. Documentation Retrieval (The "Reading" Layer)

Reading documentation pages is expensive due to token bloat. Two primary tools compete here: Ref.tools and Jina Reader.

3.1 Jina Reader (r.jina.ai)

Winner for Budget.
Jina converts URLs to LLM-friendly Markdown.

- **Pricing: 10 Million tokens FREE** for non-commercial/hobby use.[5]
- **Paid:** ~$0.05 per 1M tokens after the free tier.[5]
- **Capabilities:** Excellent "URL to Markdown" conversion. Can essentially read the entire documentation web for free for a personal agent.

3.2 Ref.tools

Winner for Precision.
Ref.tools is an MCP-native server that indexes documentation to serve only relevant snippets, saving inference tokens.

- **Pricing:** The server is open-source, but the hosted API (required for the index) is a paid service. User reports indicate a **$7/month** subscription plan.[6]
- **Token Savings:** Claims to reduce token usage by up to 95% compared to raw scraping, potentially saving ~$0.09 per step in inference costs.[7]
- **Verdict:** While Ref.tools is superior for *inference* cost, its $7/mo subscription consumes 35% of the total ATLAS budget. **Jina Reader is recommended** for the initial MVP to keep fixed costs at $0.

---

4. Package Registries (Zero-Latency / Zero-Cost)

Querying pypi.org or npmjs.com directly is inefficient due to disparate API schemas and rate limits.

4.1 The Unified Solution: deps.dev

Google's **deps.dev API** is the hidden gem for 2025/26 agents.

- **Coverage:** PyPI, npm, Maven, Cargo, Go, NuGet.[8]
- **Pricing: Free**.
- **Rate Limits:** No published hard limit; standard "polite" throttling applies. Supports **Batch Requests** (up to 5,000 items) [9], allowing ATLAS to check an entire requirements.txt in a single HTTP call (latency <300ms).

4.2 Registry-Specific Limits

- **Crates.io (Rust):** Strict limit of **1 request per second**.[10] Agents *must* throttle calls here or risk IP bans.
- **PyPI:** No official rate limit on JSON API, but backed by Fastly CDN. High volume is tolerated if User-Agent is set.[11]
- **npm:** Replicating to replicate.npmjs.com for bulk access; standard registry is robust.[12]

---

5. Code & Community Knowledge

5.1 GitHub API

- **Search API:** Extremely restrictive. **30 requests per minute**.[13]
  - *Workaround:* Do not use GitHub Search API for code discovery. Use Exa with site:github.com instead.
- **Core API (Reading Files):** Generous. **5,000 requests per hour** with a Personal Access Token (PAT).[14]
- **Cost:** Free.

5.2 Stack Overflow

- **API Limits: 10,000 requests per day** with an API Key (300/day without).
- **Policy Warning:** Terms of Service explicitly **prohibit** using Stack Overflow data to "train, test, or improve" generative AI models.
  - *Runtime Access:* Fetching answers at runtime (RAG) to solve a user's specific problem is generally considered distinct from "training," but commercial use is strictly gated. For a personal agent, this is likely acceptable, but "Commercial AI Agents" are being funneled to paid API tiers.

---

6. Final Architecture & Cost Estimate

Constraint: 100 queries/day (~3,000/month).

Component	Service	Tier/Cost	Monthly Est.	Notes
Search (Code)	Exa.ai	$5/1k req	**$7.50**	Assumes 50% of queries use Exa (1,500 reqs).
Search (General)	Tavily	Free (1k)	**$0.00**	Use for first 1k general queries.
Docs Reading	Jina Reader	Free (10M tok)	**$0.00**	10M tokens is ~20k pages.
Packages	deps.dev	Free	**$0.00**	Use Batch API for speed.
Code Fetch	GitHub API	Free	**$0.00**	Use PAT for 5k/hr limit.
Caching	SQLite	Local	**$0.00**	**Required** to stay under limits.
Buffer	-	-	**$12.50**	Remaining budget for LLM inference.

Total Est. Cost: $7.50 / month (leaving ~$12.50 for the LLM inference itself).

Implementation Note:
To handle the Crates.io 1 req/sec and GitHub Search 30 req/min limits, ATLAS must implement a client-side "Leaky Bucket" rate limiter. Failing to do so will result in immediate 429 errors during multi-step reasoning loops.

Works cited

1. Which News Sites Block AI Crawlers in 2025? [New Data] - BuzzStream, accessed on January 4, 2026, https://www.buzzstream.com/blog/publishers-block-ai-study/
2. Exa Pricing | AI Search Engine & Semantic Search Technology, accessed on

January 4, 2026, https://exa.ai/pricing
3. Introducing Exa 2.0 - Exa AI Research Blog | Semantic Search & Neural Network Search Engine, accessed on January 4, 2026, https://exa.ai/blog/exa-api-2-0
4. 5 Tavily Alternatives for Better Pricing, Performance, and Extraction Depth - Firecrawl, accessed on January 4, 2026, https://www.firecrawl.dev/blog/tavily-alternatives
5. Reader API - Jina AI, accessed on January 4, 2026, https://jina.ai/reader/
6. Is anyone actually making money selling MCP tools or servers? - Reddit, accessed on January 4, 2026, https://www.reddit.com/r/mcp/comments/1p2av9o/is_anyone_actually_making_money_selling_mcp_tools/
7. ref-tools/ref-tools-mcp: Helping coding agents never make mistakes working with public or private libraries without wasting the context window. - GitHub, accessed on January 4, 2026, https://github.com/ref-tools/ref-tools-mcp
8. Open Source Insights, accessed on January 4, 2026, https://deps.dev/
9. API | Open Source Insights, accessed on January 4, 2026, https://docs.deps.dev/api/v3alpha/
10. Data Access Policy - crates.io: Rust Package Registry, accessed on January 4, 2026, https://crates.io/data-access
11. Introduction - PyPI Docs, accessed on January 4, 2026, https://docs.pypi.org/api/
12. `npm replication API` changes and migration guide · community · Discussion #152515 - GitHub, accessed on January 4, 2026, https://github.com/orgs/community/discussions/152515
13. Rate Limit Error · community · Discussion #179480 - GitHub, accessed on January 4, 2026, https://github.com/orgs/community/discussions/179480
14. Rate limits for the REST API - GitHub Docs, accessed on January 4, 2026, https://docs.github.com/en/rest/using-the-rest-api/rate-limits-for-the-rest-api