Semantics Delivery Network

Rethinking Web Retrieval Infrastructure for LLM Agents

Peichun Hua and Yunming Xiao

School of Data Science,
The Chinese University of Hong Kong, Shenzhen



HotNets '26 · Salt Lake City

Paper: arXiv:2609.22486 · Code: github.com/sarendis56/SemDN_HotNets26

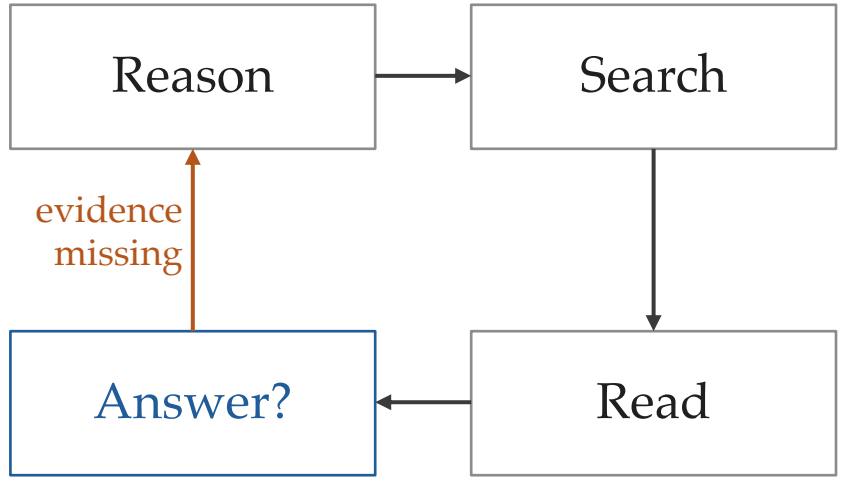
LLM agents now read the live web

- **RAG:** grounds answers in fetched evidence
- **Agentic search:** search, read, search again
- **AI fetcher traffic** is rising sharply at CDNs



Chunk

a passage or a table block
≈256-1024 tokens *(typically!)*
what the model actually reads

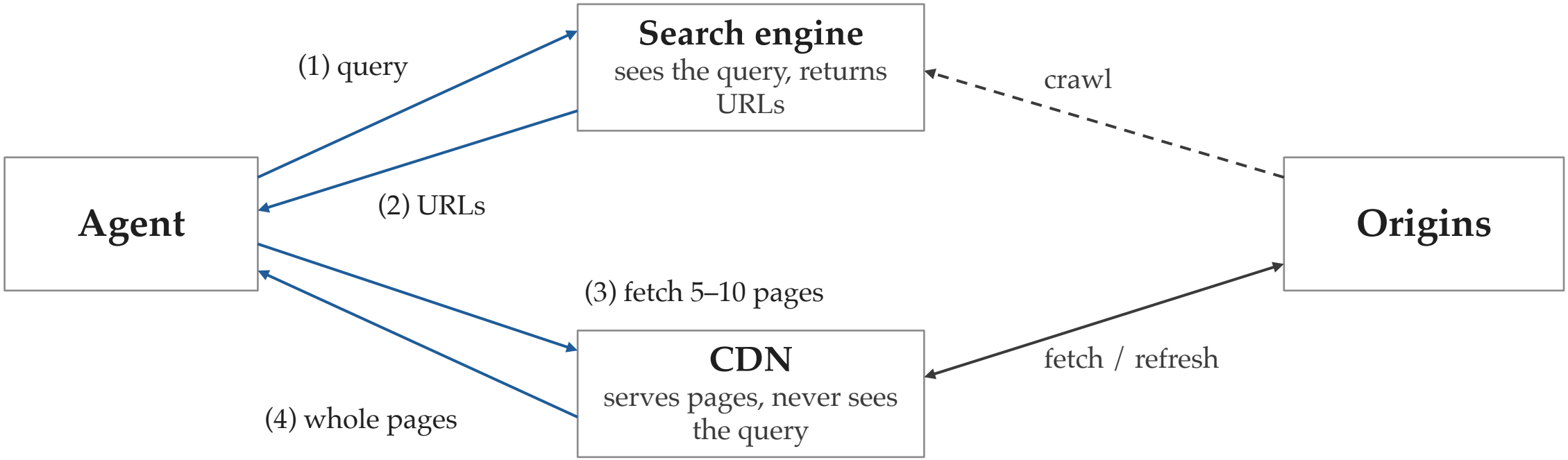


One question becomes a sequence of related sub-queries within a task.

Lewis et al., "Retrieval-Augmented Generation", NeurIPS 2020 · Jin et al., "Search-R1", arXiv 2025 · Fastly, "Q2 2025 Threat Insights Report", 2025

Today the job is split across two systems

Request: *"Compare the API rate limits of Halden and Quillbase"*

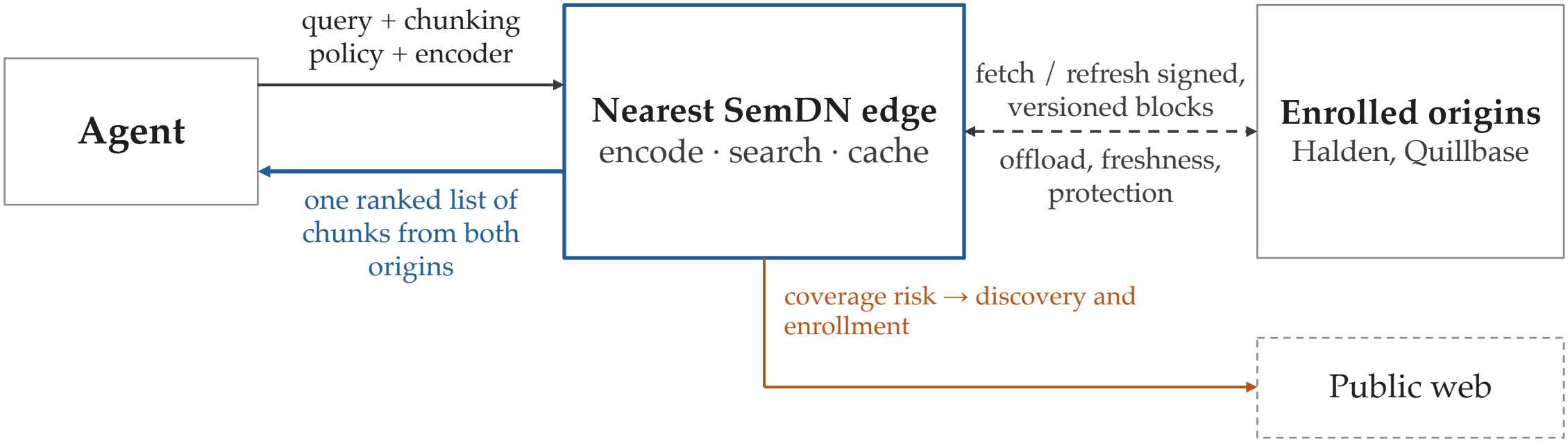


(5) clean, chunk, embed, rank
repeated by every operator; missing → search again

Search engine: knows the query, returns only URLs
CDN: delivers at the edge, never sees the query

Our proposal

SemDN: one query, one ranked list of chunks



Per-site AI indexes (Cloudflare, AISTRAL) rank one site at a time. **SemDN ranks across sites.**

Cloudflare, "An AI Index for All Our Customers", 2025 · Chen et al., "From Crawled to Published", arXiv 2026

Part 1 of 3

Chunks, not pages

Pages waste bytes and context; a similar passage is not always a useful one

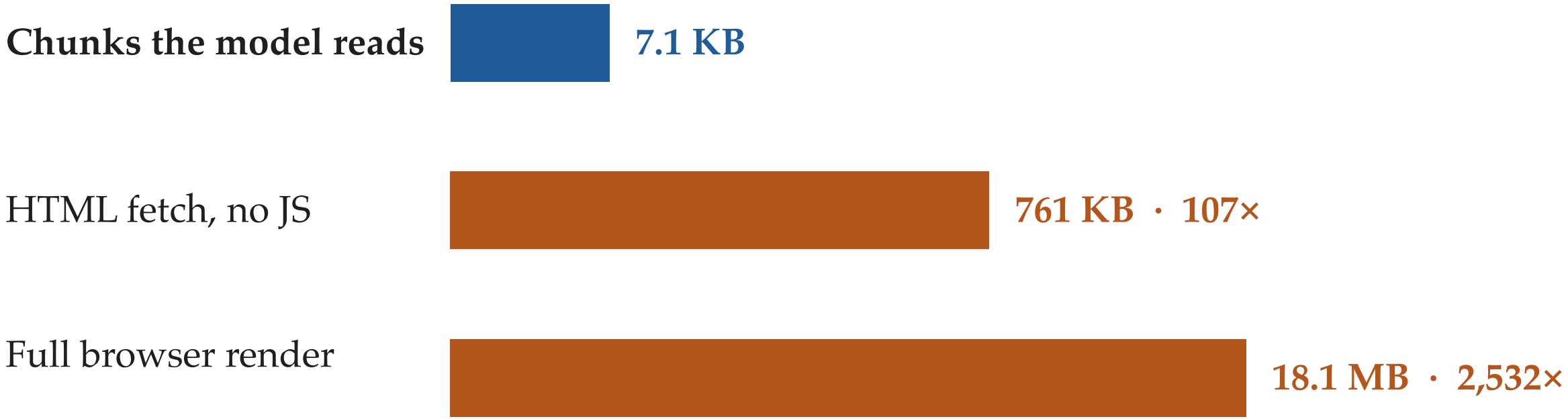
1 **Chunks, not pages**

2 Reuse at the edge

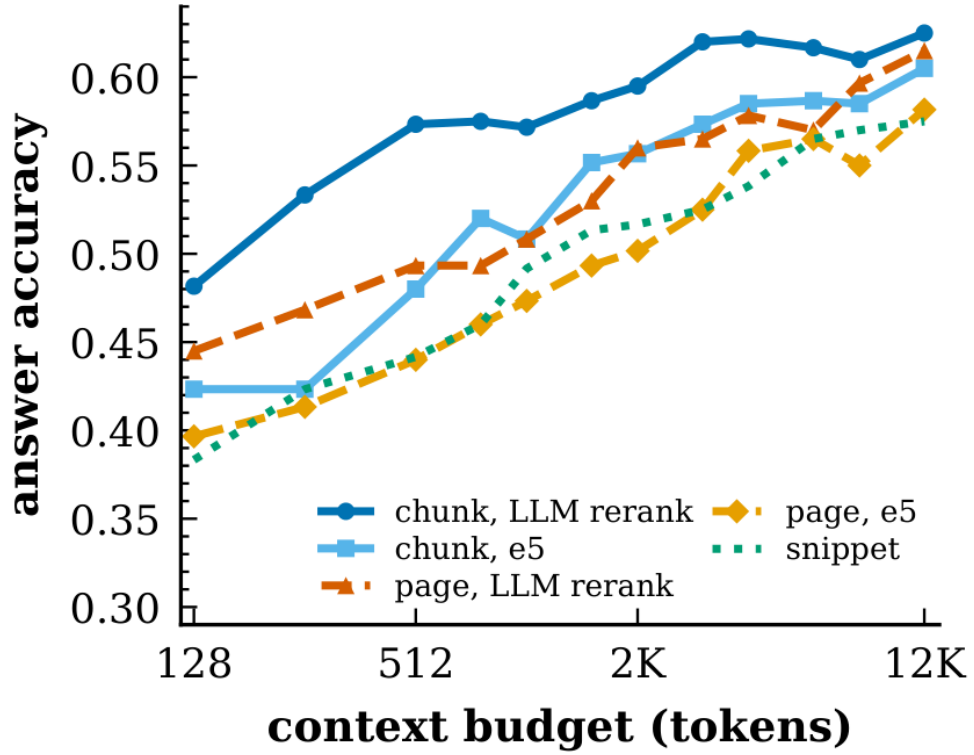
3 Open challenges

Pages move far more bytes than the model reads

Median bytes per query, top-5 search results, 100 live queries (bar length on a log scale)



Under a fixed context budget, chunks beat pages



CRAG Task 3, n = 600, Qwen3-32B reader (Fig. 3c)

To reach the same accuracy (≈57%)

	Context	≈ Text	1M questions*
Chunks	512 tokens	400 words	**≈ $2,000**
Pages	4K tokens	3,000 words	**≈ $16,000**

*Claude Opus 5.5 at $4 per million input tokens

"At the Oscars, who won best actor in 2000?"

Page
starts with the 1929 winner; budget runs out before 2000

Chunk
the 2000 ceremony passage names Kevin Spacey

Yang et al., "CRAG – Comprehensive RAG Benchmark", NeurIPS 2024 · Liu et al., "Lost in the Middle", TACL 2024 · Anthropic, "Claude API pricing", 2026

A similar passage is not always a useful one

"Tallest skyscraper in the UK?" (CRAG)

A · top by embedding similarity
"The UK had not been noted historically for its abundance of skyscrapers…"

B · top by LLM reranker
"The Shard, with a height of 309.6 meters, is the tallest skyscraper in the UK."

With A the reader answers "Landmark Pinnacle".
With B it answers "The Shard".

"Can I cancel Halden mid-contract?"

0.88 · To cancel, open Billing and choose End plan.

0.86 · §9.2: either party may terminate for material breach.

0.60 · Many customers pause rather than cancel.

A support bot wants the procedure.
A compliance agent wants §9.2.

SemDN → deliver chunks; share content and candidates; rank per tenant

Yang et al., "CRAG – Comprehensive RAG Benchmark", NeurIPS 2024 · Cuconasu et al., "The Power of Noise", SIGIR 2024 · Jin et al., "Search-R1", arXiv 2025

Part 2 of 3

Reuse at the edge

Agents reuse chunks; every operator repeats the same fragile preprocessing

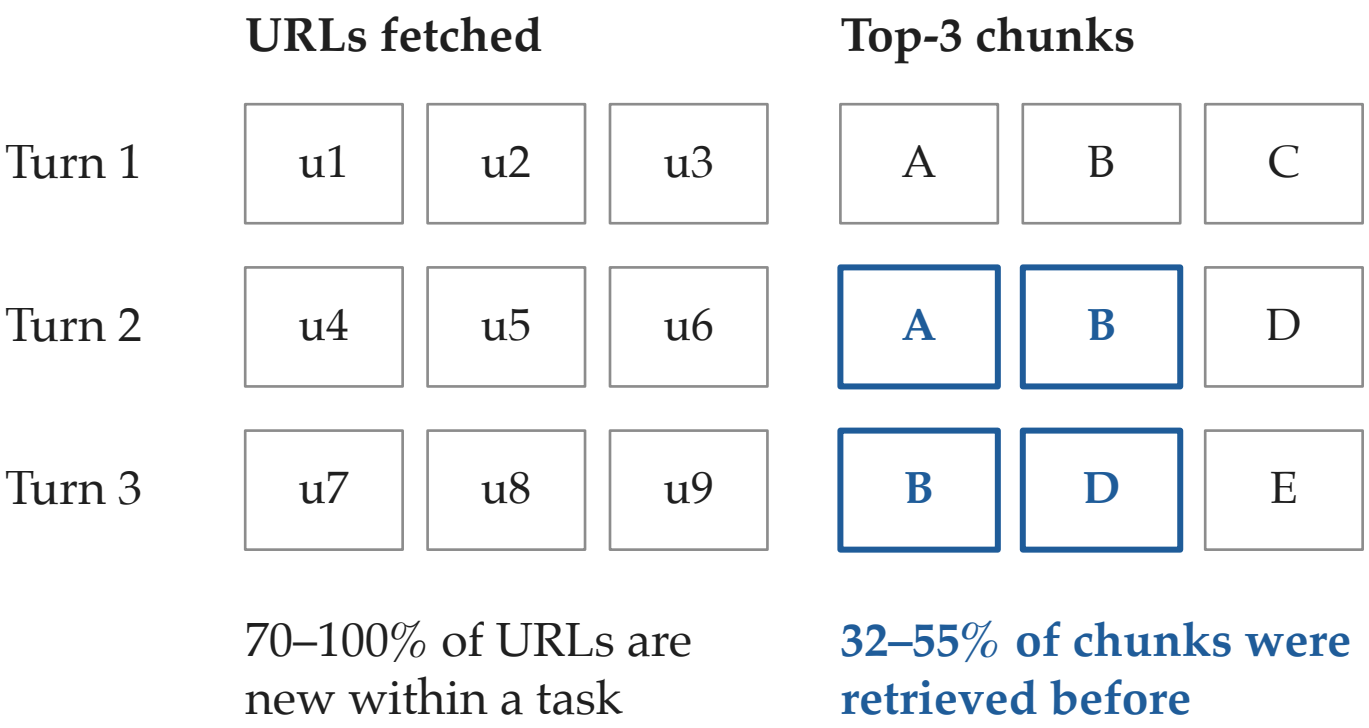
---

1 Chunks, not pages

2 Reuse at the edge

3 Open challenges

Agents reuse chunks, not URLs



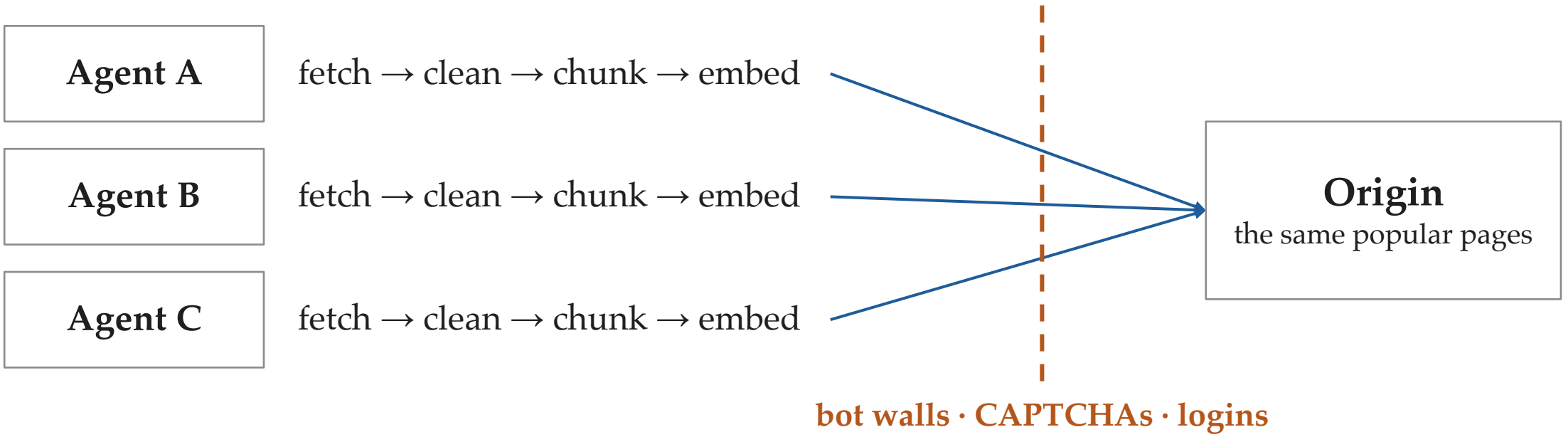
Simulated edge caches, same capacity

A chunk cache sends 13–41× fewer bytes to origin than a URL cache

SemDN → cache chunks at the edge; place and evict by the query stream

Jin et al., "Search-R1", arXiv 2025 · Yang et al., "CRAG – Comprehensive RAG Benchmark", NeurIPS 2024 · Zhang et al., "Rethinking Web Cache Design for the AI Era", SoCC 2025

Every operator repeats the same fragile pipeline



- Cloud renderer: usable text from at most 75% of 1,430 search-result URLs
- One YouTube page: 1.4 MB of HTML, 147 bytes of clean text

SemDN → fetch and clean once per content version; origins enroll for offload

Axios, "The internet's free ride is over", 2026 · Cloudflare, "Browser Rendering", 2026 · Fastly, "Q2 2025 Threat Insights Report", 2025

Part 3 of 3

Open challenges

Misses, incentives, and trust

1 Chunks, not pages

2 Reuse at the edge

3 Open challenges

Top-k search has no miss bit

Three actions: serve what the edge holds (cost 0) · ask the regional parent (1) · search the web (5)

Query	What the edge holds	Right action
"Current free-tier rate limit for the Halden API?"	Three chunks agree: 60 requests per minute; 4–7 months old	**Search: the limit dropped to 30 nine days ago**
"What does HTTP status 429 mean?"	Three chunks agree: too many requests; 2–3 years old	**Serve: old, but the meaning has not changed**
"Opening stock price of Impinj today?"	Chunks are 2–3 days old	**Search: "today" makes any cached price stale**

Agreement among chunks and their age do not decide the miss. The query does.

Yang et al., "CRAG – Comprehensive RAG Benchmark", NeurIPS 2024

Who pays, and why trust the edge?

A two-sided service



- Neither side pays just because the other uses it
- Bootstrap: start from public pages; client traffic shows whom to enroll

Why trust the edge?

Is this what the site said?
→ origins sign versioned blocks

Did I get the best chunks?
→ verifiable search: 0.8–2 s per query today

Was any site left out?
→ a public log of enrolled origins

Backman et al., "HTTP Message Signatures", RFC 9421 · Qiu et al., "V3DB", arXiv 2026 · Avramov et al., "Atlas", arXiv 2026

Make semantic selection a network-delivery service

- Agents read chunks; today's web delivers URL-addressed pages
- **SemDN:** a search engine's query interface + a CDN's edge delivery
- **Early evidence:** ~100× fewer bytes, chunk reuse URL caches miss, more accuracy per token

Open questions

1. When is cached evidence good enough to serve, given timeliness and coverage?
2. What belongs in the network, and what in a client cache?
3. How do origins and clients join first, and what if an origin refuses?
4. Can clients verify they got the best chunks, without seeing the index?
5. How does one query find its sources across millions of sites?

Paper: arXiv:2609.22486 · Interactive demo: peichun.xyz/papers/semantics-delivery-network/demo

References

[1] Peichun Hua and Yunming Xiao. Semantics Delivery Network: Rethinking Web Retrieval Infrastructure for LLM Agents. HotNets 2026. arXiv:2609.22486.

[2] Patrick Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.

[3] Bowen Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arXiv:2503.09516, 2025.

[4] Fastly. Q2 2025 Threat Insights Report: The Rise of AI Crawlers and Fetchers. 2025.

[5] Cloudflare. An AI Index for All Our Customers. Blog, 2025.

[6] Yixi Chen et al. From Crawled to Published: A Publisher-Controlled Web Architecture for an AI-Native Internet. arXiv:2511.18354v2, 2026.

[7] Xiao Yang et al. CRAG – Comprehensive RAG Benchmark. NeurIPS 2024.

[8] Nelson F. Liu et al. Lost in the Middle: How Language Models Use Long Contexts. TACL 2024.

[9] Anthropic. Claude API pricing (Claude Opus 5.5: $4 per million input tokens). Accessed November 2026.

[10] Florin Cuconasu et al. The Power of Noise: Redefining Retrieval for RAG Systems. SIGIR 2024.

[11] Yazhuo Zhang et al. Rethinking Web Cache Design for the AI Era. ACM SoCC 2025.

[12] Axios. The internet's free ride is over. Media Trends, November 3, 2026.

[13] Cloudflare. Browser Rendering: Markdown endpoint. Developer documentation, 2026.

[14] Annabelle Backman, Justin Richer, and Manu Sporny. HTTP Message Signatures. RFC 9421, 2024.

[15] Zipeng Qiu et al. V3DB: Audit-on-Demand Zero-Knowledge Proofs for Verifiable Vector Search over Committed Snapshots. arXiv:2603.03065, 2026.

[16] Nikolay Avramov et al. Atlas: Efficient Verifiable Semantic Search. arXiv:2609.11841, 2026.