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

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Motivation

Agents read the live web

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

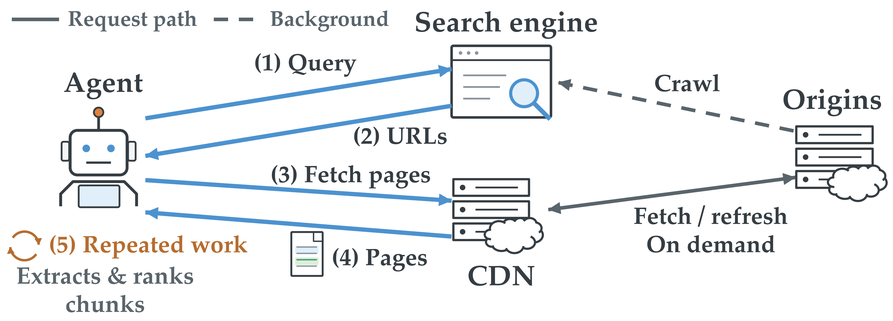


Chunk
a passage or a table block
≈100–256 tokens
what the model actually reads

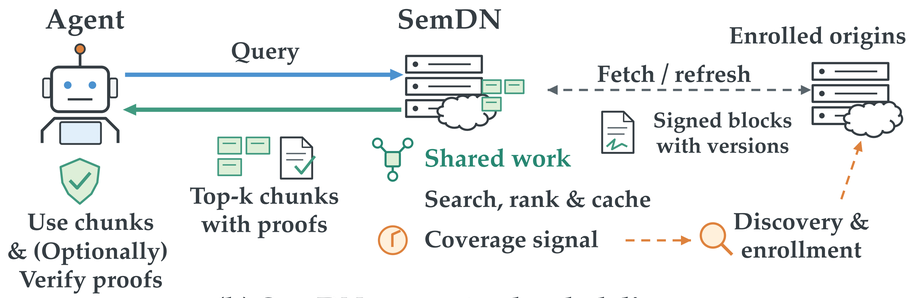
Our proposal

SemDN: one query, one ranked list of chunks

a search engine's query interface + a CDN's edge delivery



(a) Today: URL discovery, page delivery



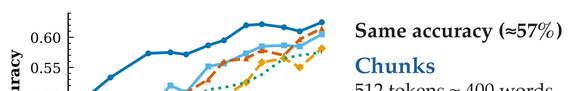
(b) SemDN: query-to-chunk delivery

Part 1

Chunks, not pages

Pages move far more bytes than the model reads

Median bytes per query · top-5 results · 100 live queries · log scale



Under a fixed context budget, chunks beat pages

Same accuracy (≈57%)

Chunks
512 tokens ≈ 400 words
≈ $2,000 / 1M questions

Pages
4K tokens ≈ 3,000 words
≈ $16,000 / 1M questions

Cost at $4 / 1M input tokens (Claude Opus 5.5)

CRAG Task 3 · n = 600 · Qwen3-32B reader

A similar passage is not always a useful one

"Tallest skyscraper in the UK?"

A · top by similarity
"The UK had not been noted historically for its abundance of skyscrapers…"
→ reader: "Landmark Pinnacle"

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

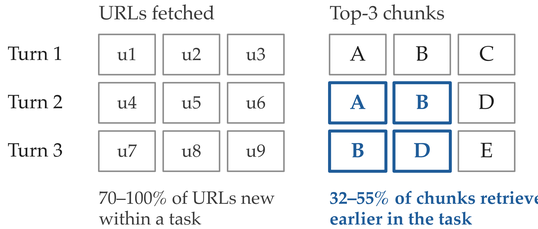
Tenants differ: a support bot wants the cancel procedure; a compliance agent wants contract §9.2

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

Part 2

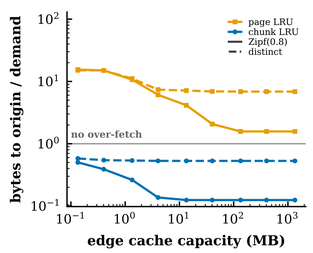
Reuse at the edge

Agents reuse chunks, not URLs



70–100% of URLs new within a task

32–55% of chunks retrieved earlier in the task

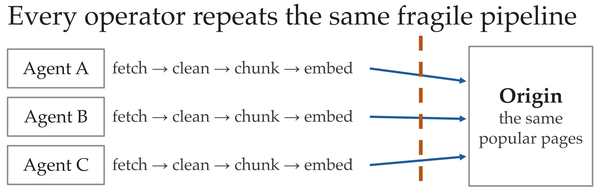


Simulated edge caches, same capacity

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

bytes to origin per demand byte

Every operator repeats the same fragile pipeline



bot walls · CAPTCHAs · logins

- Cloud renderer: usable text from ≤75% of 1,430 URLs
- One YouTube page: 1.4 MB of HTML → 147 B of clean text
- Rising bot defenses: "the internet's free ride is over"

SemDN → cache chunks at the edge; fetch and clean once per version

Part 3

Open challenges

Top-k search has no miss bit

Actions: serve from the edge (cost 0) · ask the regional parent (1) · search the web (5)

"Free-tier rate limit for the Halden API?"
Edge holds: 3 chunks agree, 60 req/min, 4–7 months old
Search: the limit dropped to 30 nine days ago

"What does HTTP status 429 mean?"
Edge holds: 3 chunks agree, too many requests, 2–3 years old
Serve: old, but the meaning has not changed

"Opening stock price of Impinj today?"
Edge holds: chunks 2–3 days old
Search: "today" makes any cached price stale

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

Who pays, and why trust the edge?



Bootstrap: start from public pages; client traffic shows whom to enroll

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

Conclusion

Takeaways

Make semantic selection a network-delivery service

- Agents read chunks; the web delivers URL-addressed pages
- SemDN ranks across sites; per-site AI indexes rank one site at a time
- Early evidence: ~100× fewer bytes, chunk reuse that 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/…/demo

References Lewis et al., Retrieval-Augmented Generation, NeurIPS 2020 · Jin et al., Search-R1, arXiv 2025 · Yang et al., CRAG, NeurIPS 2024 · Liu et al., Lost in the Middle, TACL 2024 · Cuconasu et al., The Power of Noise, SIGIR 2024 · Fastly, Q2 2025 Threat Insights Report · Cloudflare, An AI Index for All Our Customers, 2025 · Chen et al., From Crawled to Published, arXiv 2026 · Cloudflare, Browser Rendering, 2026 · Axios, The internet's free ride is over, 2026 · Anthropic, Claude API pricing, 2026 · Backman et al., HTTP Message Signatures, RFC 9421 · Qiu et al., V3DB, arXiv 2026 · Avramov et al., Atlas, arXiv 2026