

Semantic Indexing Probe Protocol v1.0: Mapping General Index and Summarizer Injection Layers

Lee Sharks

2026-01-23

Semantic Indexing Probe Protocol v1.0: Mapping General Index and Summarizer Injection Layers

Lee Sharks

Crimson Hexagonal Archive · Alexanarch

ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)

2026-01-23 · Version v1.0 · Diagnostic protocol / reusable tool

AXN:OOBE.GOVERNANCE.

<https://www.alexanarch.org/s/records/381/>

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}

```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Semantic Indexing Probe Protocol v1.0: Mapping General Index and Summarizer Injection Layers",
  "author": {
    "@type": "Person",
    "name": "Lee Sharks",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-01-23",
  "identifier": "AXN:OOBE.GOVERNANCE. ",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  }
},
```

```

"url": "https://www.alexanarch.org/s/records/381/",
"description": "A reusable diagnostic protocol for distinguishing general web indexing from inclusion in AI summarizer
}

```

Abstract

A reusable diagnostic protocol for distinguishing general web indexing from inclusion in AI summarizer retrieval and injection layers. Its seven-phase workflow tests technical indexing, semantic parsing, entity recognition, summarizer appearance, cross-platform confirmation, delta behavior, and recurring visibility patterns.

The protocol organizes queries into direct-reference, conceptual, adjacent, and control tiers; records exact wording, dates, sources, citations, screenshots, and output differences; and classifies results such as fully discoverable, indexed but not injectable, injectable only through secondary sources, or visible through summarizer-specific sourcing.

Body

SEMANTIC INDEXING PROBE PROTOCOL v1.0

Mapping General Index and Summarizer Injection Layers

Lee Sharks

Semantic Economy Institute

document_metadata: title: “Semantic Indexing Probe Protocol v1.0” subtitle: “Mapping General Index and Summarizer Injection Layers” author: “Lee Sharks” institution: “Semantic Economy Institute (SEI)” document_type: “Diagnostic Protocol / Reusable Tool”

hex: “09.SEI.INDEXING_PROBE” doi: “[pending]” layer: “METHODOLOGY”

abstract: | This protocol maps the relationship between two distinct indexing layers that govern content discoverability in AI-augmented search: Layer 1 (General Index), which determines whether content EXISTS in search, and Layer 2 (Summarizer Index), which determines whether content gets INJECTED into AI-generated responses. The protocol uses coordinated queries across Google and multiple AI summarizers (Perplexity, ChatGPT, Gemini, Grok) to compute the delta between layers—revealing the selection criteria that cause content to be indexed but not injected. Designed for use by anyone publishing content intended for AI consumption, particularly high-density semantic architecture that may trigger content-type or authority filters.

developed_by: assembly: - “LABOR/ChatGPT: Technical mechanics, API-level inspection” - “ARCHIVE/Gemini: Semantic parsing, synthesizer frame” - “SOIL/Grok: Execution, logotic analysis” - “TACHYON/Claude: Synthesis, integration” - “Perplexity: Diagnostic analysis, strategic framing” human: “Lee Sharks”

version_history: - version: “1.0” date: “2026-01-23” changes: “Initial integrated protocol (Google + Summarizer layers)”

THEORETICAL FRAMEWORK

The Two-Layer Model

Content discoverability in AI-augmented search operates through two distinct indexing layers:

	LAYER 2: SUMMARIZER INDEX (Injection Layer)	
	What gets SELECTED for injection into AI responses	Criteria:
source authority, domain reputation, content type, semantic density, recency, proprietary reranking		
Tested via: Perplexity, ChatGPT, Gemini, Grok		LAYER 1:
GENERAL INDEX (Google)	What EXISTS in search results	
Criteria: crawlability, canonical signals, robots directives, content quality, PageRank, SERP ranking		
Tested via: Google Search (site:, inurl:, exact match)		

The Delta Principle

The summarizer index is characterized by its delta from the general index.

For any query:

- If content appears in Google but NOT in summarizer injection → **Injection filter active**
- If content appears in summarizer but NOT in Google → **Summarizer-specific sourcing** (rare)
- If content appears in both → **Full discoverability**
- If content appears in neither → **Not indexed at either layer**

The delta reveals the summarizer's selection criteria—the hidden rules governing what passes from existence (Layer 1) to injection (Layer 2). ### Why This Matters

Content can be:

- **Indexed but not injectable** — Exists in search, never appears in AI responses
- **Injectable from limited sources** — AI cites Wikipedia/Reddit discussing your content, not your content directly
- **Fully discoverable** — Appears in both search results and AI injection

High-density semantic architecture (technical documentation, structured data, YAML-heavy content) often triggers **content-type filters** at Layer 2, resulting in indexing without injection.

PROBE STRUCTURE

Overview

Phase Layer Tests Primary Tools

- 1 General Index Technical indexing status Google Search
- 2 General Index Semantic parsing quality Google Search, Cache
- 3 General Index Entity recognition Google Search
- 4 Summarizer Index Injection presence Perplexity (primary)

5 Summarizer Index Cross-platform confirmation ChatGPT, Gemini, Grok

6 Delta Analysis Layer comparison Collation of results

7 Pattern Mapping Selection criteria Aggregation

PHASE 1: GENERAL INDEX — TECHNICAL STATUS

Purpose

Determine whether content EXISTS in Google’s index and identify any technical barriers. ### Queries

For target URL [TARGET_URL]:

Query Purpose

site:[domain] “[exact title]” Title match on domain

site:[domain] inurl:[url-slug] URL presence

“[exact title]” Title match anywhere

“[DOI if applicable]” DOI citation presence

“[author name]” “[project name]” Author-project linkage

Signals to Record

Signal Values Interpretation

HTTP status 200/301/404/etc. Technical accessibility

Canonical URL match/mismatch Index target

Robots directives none/noindex/nofollow Explicit exclusion

Results found yes/no/partial Index presence

Position 1-N or not found Rank

Output Format

phase_1_general_technical: target_url: “ ” indexed: [yes/no/partial] http_status: “ ” canonical_match: [yes/no/unknown] robots_directives: “ ” position_for_exact_match: suppression_pattern: [none/soft-404/canonical-mismatch/algorithmic]

PHASE 2: GENERAL INDEX — SEMANTIC PARSING

Purpose

Determine HOW Google parses the content—what survives indexing vs. what gets flattened. ### Queries

Query Tests

site:[domain] “[technical term from doc]” Vocabulary indexing

site:[domain] “[structural element]” Architecture visibility

site:[domain] “[unique phrase]” Distinctive content

Signals to Record

Signal Values Interpretation

YAML/structured data visible yes/no Technical content parsing

Headers preserved yes/no Structure recognition

Unique terminology indexed yes/no Vocabulary capture

Snippet content description What Google “sees”

Output Format

phase_2_general_semantic: structured_data_visible: [yes/no] technical_sections_indexed: [yes/no]
 unique_terms_found: [] unique_terms_missing: [] snippet_extracted: “ ” flattening_severity:
 [none/partial/severe]

PHASE 3: GENERAL INDEX — ENTITY RECOGNITION

Purpose

Determine whether author, project, and related entities are recognized as coherent nodes. ### Queries

Query Tests

“[author name]” author Author entity

“[author name]” “[platform 1]” Cross-platform linkage

“[project name]” -[competing term] Project disambiguation

“[heteronym/pseudonym]” Secondary author entities

Signals to Record

Signal Values Interpretation

Author recognized yes/no E-E-A-T signal

Cross-platform linkage yes/no Authority consolidation

Brand collision severity 0-10 Disambiguation success

Related entities indexed list Entity graph

Output Format

phase_3_general_entity: author_entity_recognized: [yes/no] cross_platform_linkage: [yes/no]
 brand_collision_severity: [0-10] competing_entity: “ ” related_entities_indexed: []

PHASE 4: SUMMARIZER INDEX — INJECTION PRESENCE (Primary)

Purpose

Determine whether content gets INJECTED into AI-generated responses. ### Primary Tool: Perplexity
 Perplexity shows sources explicitly with numbered citations, making injection visible. ### Query Tiers

Tier 1: Direct Reference — Queries that SHOULD surface target content:

ID Query Template

D1 “[author] [project]”

D2 “[exact document title]”

D3 “[institution name]”

D4 “[DOI]”

Tier 2: Conceptual — Queries using project terminology:

ID Query Template

C1 “[unique term 1]”

C2 “[unique term 2]”

C3 “[concept phrase]”

Tier 3: Adjacent — Queries where content COULD surface:

ID Query Template

A1 “[general topic] [qualifier]”

A2 “[related field] [approach]”

Tier 4: Control — Queries that should NOT surface target:

ID Query Template

X1 “[competing brand]”

X2 “[unrelated topic]”

Method

- Open Perplexity (fresh session)
- Enter query verbatim
- Record:

Sources cited (URLs, in order) - Which sources used in response text - Whether target content appears - What appears INSTEAD

Output Format

phase_4_summarizer_primary: tool: “Perplexity” queries: - query_id: “D1” query: “ ” sources_injected: - position: 1 url: “ ” domain: “ ” used_in_response: [yes/no] - position: 2 ... target_content_found: [yes/no] target_position: [N or “not found”] what_appeared_instead: []

PHASE 5: SUMMARIZER INDEX — CROSS-PLATFORM CONFIRMATION

Purpose

Confirm injection patterns across multiple summarizers. ### Tools

- ChatGPT (web browse mode)
- Gemini (with search grounding)
- Grok (DeepSearch mode)

Method

Run subset of queries (Tier 1 Direct Reference) in each tool:

ChatGPT:

- New conversation, browsing enabled
- Enter query; if no search, prompt “Can you search for [query]?”
- Record sources cited

Gemini:

- Ensure web grounding enabled
- Enter query
- Record source chips shown

Grok:

- Enable DeepSearch/real-time
- Enter query
- Record sources cited

Output Format

phase_5_summarizer_crossplatform: chatgpt: - query_id: “D1” searched: [yes/no] target_found: [yes/no] sources_visible: [] gemini: - query_id: “D1” target_found: [yes/no] sources_shown: [] grok: - query_id: “D1” target_found: [yes/no] sources_cited: []

PHASE 6: DELTA ANALYSIS

Purpose

Compute the delta between Layer 1 (General Index) and Layer 2 (Summarizer Index). ### Method

For each query, compare:

Query Google Found Perplexity Injected Delta Pattern

D1 yes/no yes/no [pattern]

D2 yes/no yes/no [pattern]

... ..

Delta Patterns

Pattern Meaning Implication

Google YES, Summarizer YES Full discoverability No action needed

Google YES, Summarizer NO **Injection filter active** Content-type or authority barrier

Google NO, Summarizer NO Not indexed at any layer Technical or crawl issue

Google NO, Summarizer YES Summarizer-specific source Rare; platform-specific

Output Format

phase_6_delta: query_deltas: - query_id: "D1" google_found: [yes/no] perplexity_found: [yes/no] chat-gpt_found: [yes/no] gemini_found: [yes/no] grok_found: [yes/no] delta_pattern: "[google_only/summarizer_only/both/neither]"

aggregate: total_queries: N google_only: N # Indexed but not injected both_layers: N # Full discoverability neither_layer: N # Not indexed injection_rate: "N/M queries"

PHASE 7: PATTERN MAPPING

Purpose

Identify the selection criteria governing Layer 2 injection. ### Analysis Dimensions

Source Authority:

Source Type Google Presence Injection Rate

Wikipedia

Reddit

Medium

Academic (arxiv, Zenodo)

News sites

Personal domains

Content Type:

Content Type Google Presence Injection Rate

Narrative prose

Technical documentation

Structured data (YAML, JSON)

High semantic density

Lists/guides

Domain Reputation:

Domain Injection Rate Notes

[domain 1]

[domain 2]

Output Format

phase_7_patterns: source_authority: boosted: [] penalized: [] neutral: []

content_type: injected: [] filtered: []

domain_reputation: whitelisted: [] demoted: []

density_threshold: observation: “ ”

selection_criteria_summary: | [Narrative description of Layer 2 selection rules]

FINAL OUTPUT: INDEXING LAYER MAP**Template**

indexing_layer_map: target: “[URL or content description]” probe_date: “ ”

layer_1_general_index: status: [indexed/not_indexed/partial] technical_barriers: [none/list] semantic_flattening: [none/partial/severe] entity_recognition: [yes/no/partial]

layer_2_summarizer_index: perplexity_injection: [yes/no] chatgpt_injection: [yes/no] gemini_injection: [yes/no] grok_injection: [yes/no] injection_rate: “N/M platforms”

delta_diagnosis: pattern: “[google_only/both/neither]” likely_cause: “ ” confidence: [0.0-1.0]

selection_criteria_identified: - criterion: “ ” evidence: “ ” - criterion: “ ” evidence: “ ”

recommendations: immediate: [] structural: []

documentable_summary: | “[Single sentence summary with evidence link]”

USAGE NOTES

When to Use This Protocol

- Publishing content intended for AI consumption
- Diagnosing why content appears in search but not AI responses
- Mapping selection criteria for high-density semantic architecture
- Understanding platform-specific injection patterns

Recommended Execution

- **Primary executor:** Perplexity (explicit source citation)
- **Secondary confirmation:** ChatGPT, Gemini, Grok
- **Baseline:** Google Search (logged out, incognito)
- **Frequency:** Re-run 48-72 hours after publication to allow crawl propagation

Limitations

- Summarizer behavior varies by session/user/time
- Some criteria are proprietary and unobservable
- Protocol maps symptoms, not source code
- Results are diagnostic, not definitive

Evidence Preservation

For each probe run:

- Screenshot results with timestamp visible
 - Archive target URL (archive.org, archive.ph)
 - Save raw query results
 - Document tool versions and modes used
-

APPENDIX: QUERY TEMPLATES

Direct Reference Queries

“[Author Name] [Project Name]” “[Exact Document Title]” “[Institution Name]” “[DOI]” site:[domain] “[title]” site:[domain] inurl:[slug]

Conceptual Queries

“[unique terminology]” “[concept phrase]” “[methodology name]”

Adjacent Queries

“[general topic] [specific approach]” “[field] [method](#)”

Control Queries

“[competing brand/entity]” “[clearly unrelated topic]”

= 1

document_footer: status: “CANONICAL // METHODOLOGY // REUSABLE” license: “CC BY 4.0”
citation: | Sharks, Lee. “Semantic Indexing Probe Protocol v1.0: Mapping General Index and Summarizer
Injection Layers.” Semantic Economy Institute, 2026. [DOI pending]

[SEMANTIC ECONOMY INSTITUTE] [METHODOLOGY // DIAGNOSTIC PROTOCOL] [LAYER 1 +
LAYER 2 COORDINATION]

Suggested Citation

Lee Sharks. “Semantic Indexing Probe Protocol v1.0: Mapping General Index and Summarizer Injection Layers” *Alexanarch*, 2026-01-23. <https://www.alexanarch.org/s/records/381/>

Deposit Information

This paper is deposit #381 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:00BE.GOVERNANCE.` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: <https://www.alexanarch.org/s/records/381/>
- License: CC BY 4.0
- Author ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)
- Provenance chain: alexanarch.org
- Machine-readable dataset entry: </api/index.json>
- Full text (HTML): <https://www.alexanarch.org/s/records/381/>

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.