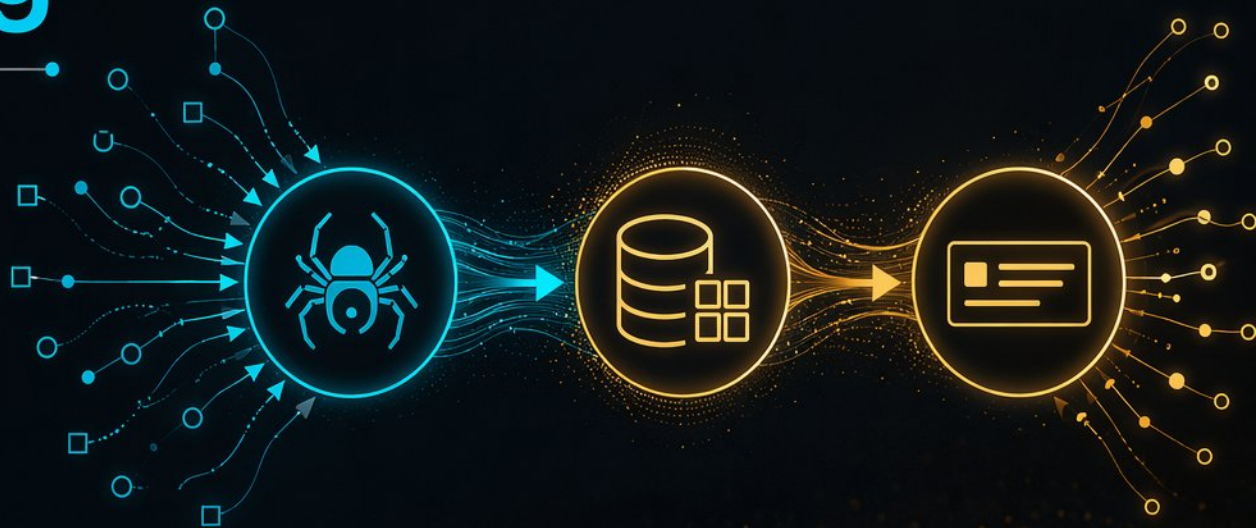


How Search Engines Work: Crawling, Indexing, and Serving

- Three core pillars: crawling, indexing, and ranking/serving
- A technical, systems-level view built for technical creators
- Follow a URL end-to-end: from discovery to the search result
- Understand the infrastructure, not the marketing myths



The Scale Problem:

Web Size, Freshness, and Crawl Budget



- Trillions of pages exist; exhaustive crawling is impossible.



- Crawl budget = URLs Google can and wants to crawl.



- Crawl rate limit: server capacity ceiling (TTFB, 5xx).



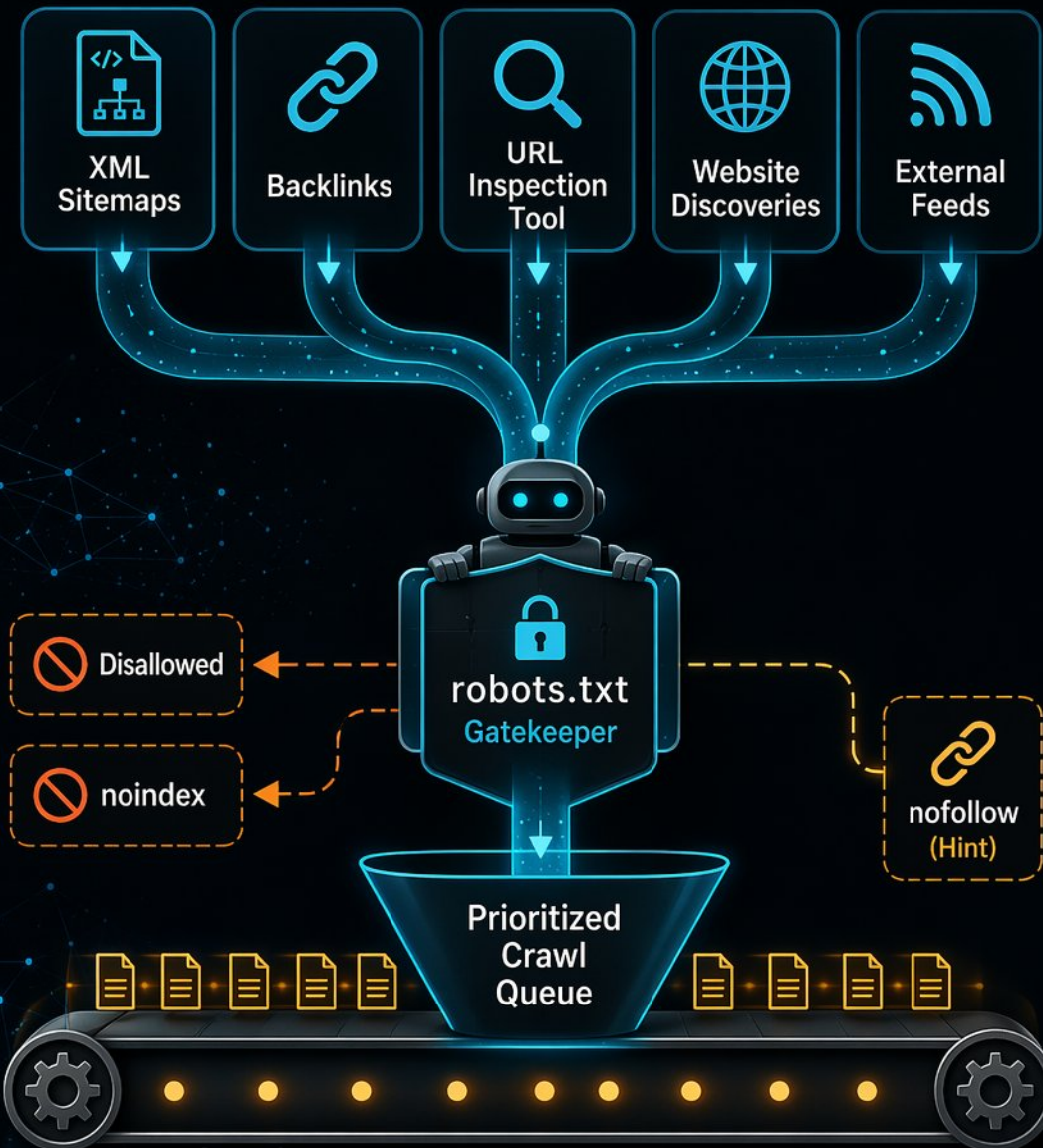
- Crawl demand: popularity, freshness, and perceived inventory.



- Wasted crawl budget directly limits AI visibility today.



Discovery: URL Sources and the Crawl Queue



- XML sitemaps, backlinks, and URL Inspection tool are the top discovery sources.
- Crawl queue priority is based on authority, freshness signals, and server performance.
- URL popularity and staleness drive crawl demand; server health sets the rate limit.
- Robots.txt and `noindex` control access; `nofollow` is a hint, not a crawl block.

Inside the Crawler: Fetching at Scale

- Googlebot fetches up to **2MB** per URL; bytes beyond are ignored entirely.
- **TTFB** directly governs crawl rate: fast servers get more concurrent connections.
- **5xx** errors and timeouts throttle host crawl activity, protecting your infrastructure.
- Paywalls and login walls block rendering; **WRS** operates statelessly without cookies.



The Two-Wave Indexing Process: Crawl, Then Render

PHASE 1

IMMEDIATE HTML CRAWL

Extracts raw content, meta tags, and links within a 2MB limit



PHASE 2

DEFERRED WEB RENDERING SERVICE (WRS)

Queue executes JavaScript in headless Chromium



- Phase 1: Immediate HTML crawl extracts raw content, meta tags, and links within a 2MB limit



- Phase 2: Deferred Web Rendering Service (WRS) queue executes JavaScript in headless Chromium



- WRS operates statelessly—no cookies, localStorage, or Service Workers persist between renders



- Rendering delay ranges from hours to weeks, making SSR/SSG critical for indexing in 2026



- AI crawlers (GPTBot, ClaudeBot) do not render JavaScript; initial HTML must be complete

Parsing and Extraction: From HTML to Meaningful Tokens



- HTML parsing, DOM tree construction, and content extraction for text and links



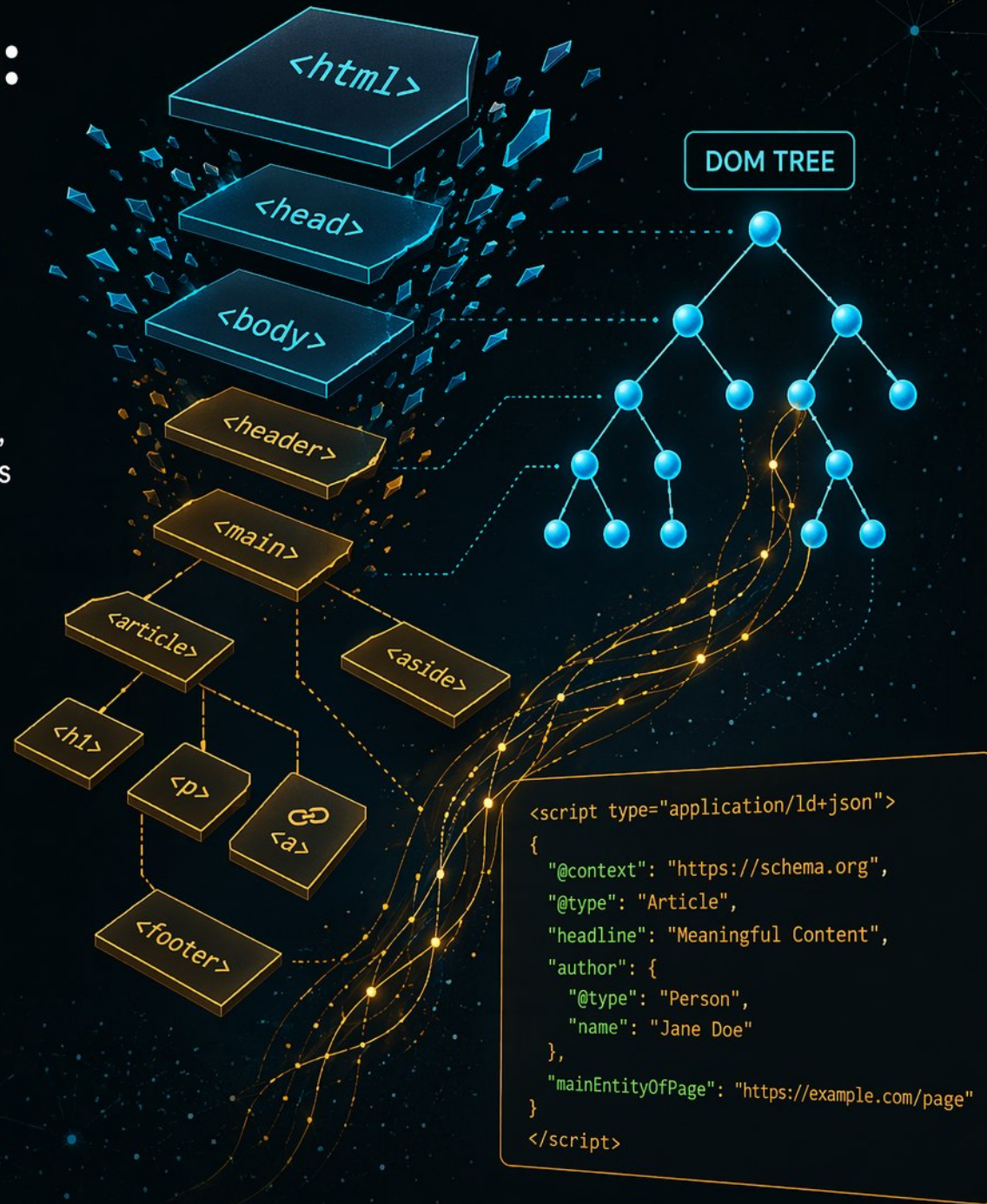
- Handles JavaScript-injected content, iframes, and shadow DOM during rendering



- Extracts critical metadata: canonical URLs, hreflang, and structured data (JSON-LD)



- JSON-LD syntax errors (trailing commas, missing quotes) block rich result eligibility



The Inverted Index: How Documents Become Searchable

- Inverted index: maps tokens to document IDs, not documents to words
- Tokenization, stemming, and normalization for English and multilingual text
- Modern indexes integrate vector embeddings for semantic understanding
- Hybrid search combines lexical (BM25) and semantic retrieval in one query
- Enables AI features like RAG, AI Overviews, and intent-driven retrieval



Query Processing: Intent, Entities, and the Knowledge Graph



- Query rewriting, spelling correction, and synonym expansion refine what you actually meant.



- Named Entity Recognition (NER) classifies tokens as Person, Organization, or Location.



- The Knowledge Graph disambiguates "Apple" the company from "apple" the fruit.

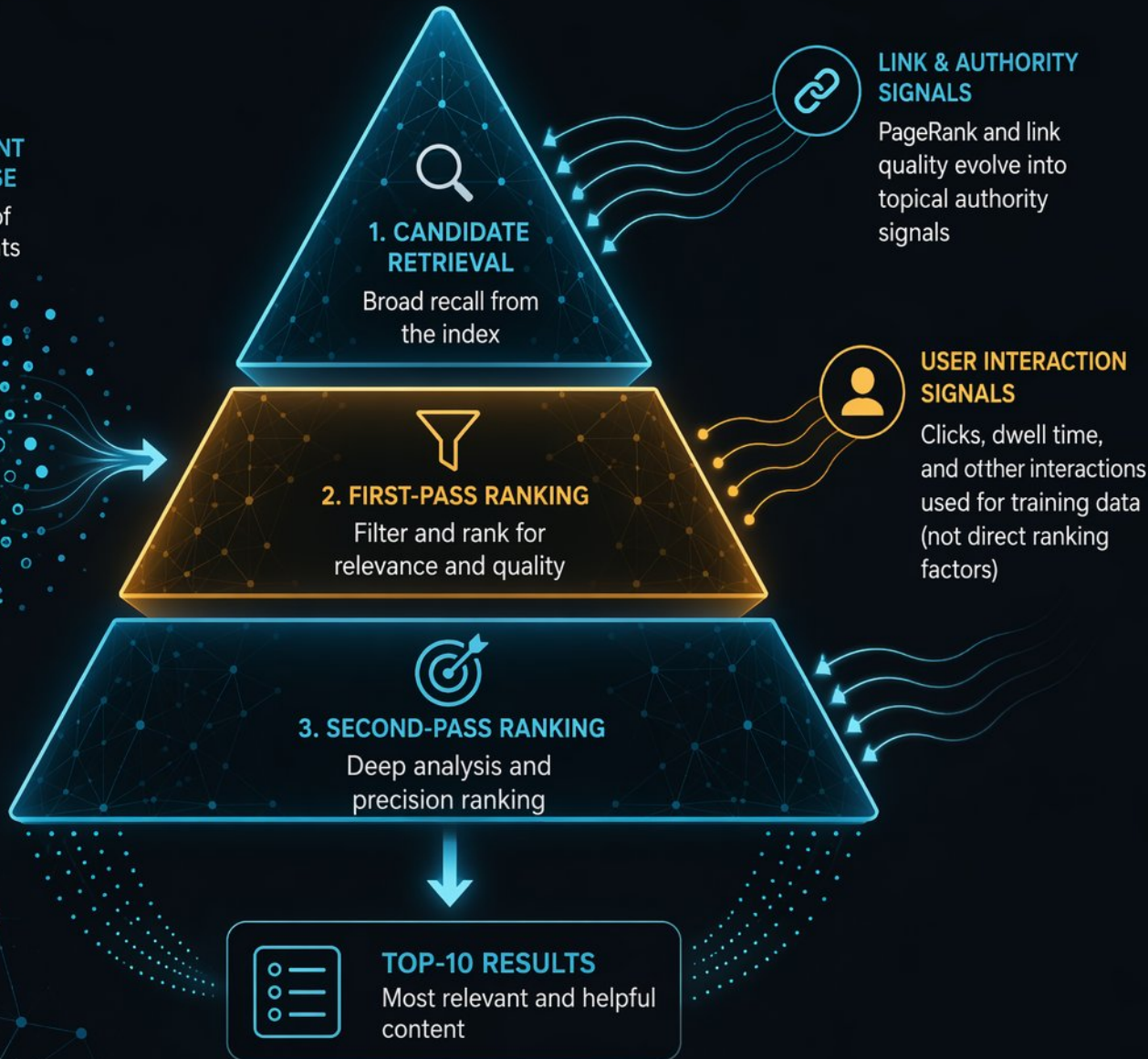


- BERT parses context bidirectionally; MUM handles multimodal, cross-language queries.



Ranking: The Multi-Stage Architecture

**DOCUMENT
UNIVERSE**
Billions of
Documents



- Candidate retrieval, then first-pass, then second-pass ranking layers
- PageRank and link quality evolve into topical authority signals
- User interactions used as training data, not direct ranking factors
- Information Gain score rewards content that teaches something new
- Site-level classifiers: thin pages can drag down strong pages

E-E-A-T: The Core Ranking Signals in 2026

- - E-E-A-T: Experience, Expertise, Authoritativeness, Trustworthiness
- - Helpful Content System: a site-wide classifier, not a page-level penalty
- - Information Gain: rewards content that teaches something new
- - Topical authority & original insights outweigh publishing volume



Serving: Assembling the Modern SERP



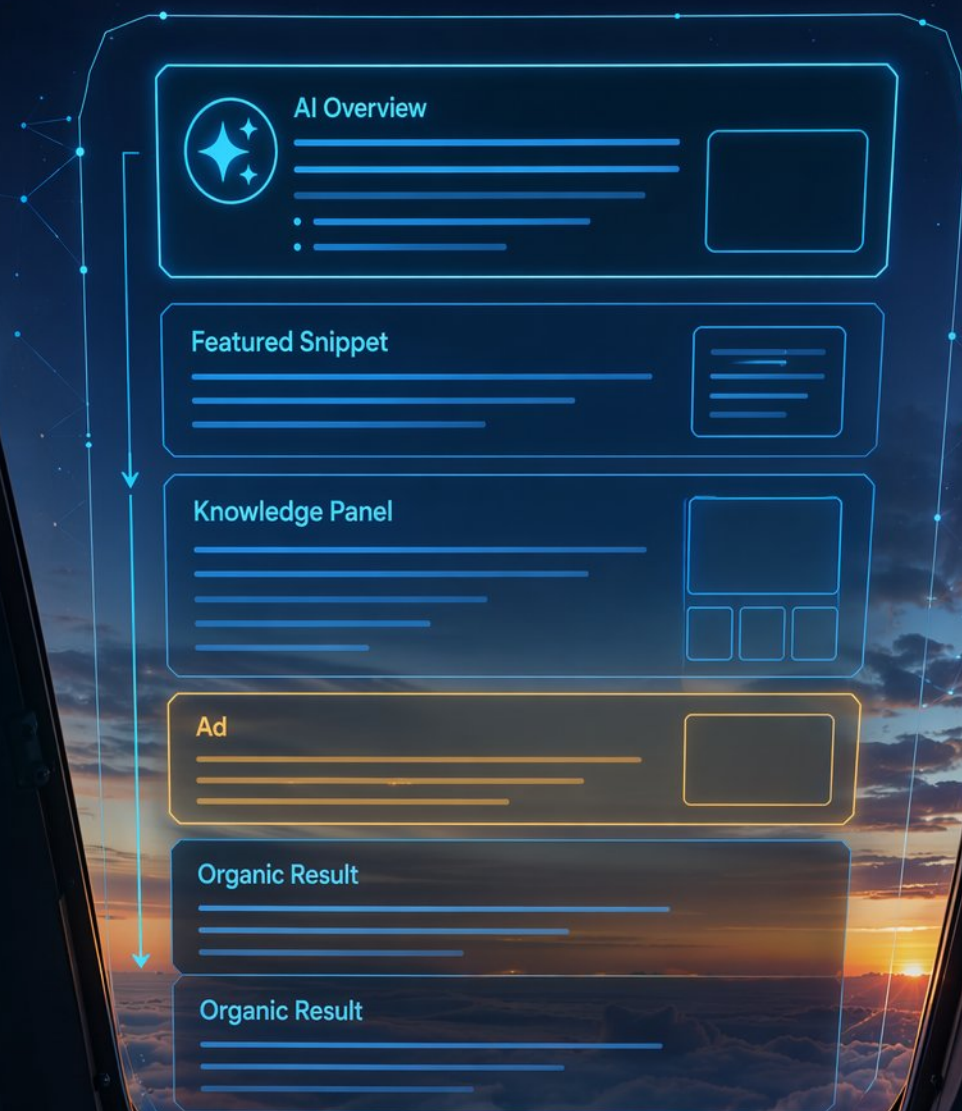
- SERP features—AI Overviews, snippets, panels—appear in 65%+ of queries, dominating above-the-fold space



- Real-time signals including location, search history, and device type personalize the full results page



- Zero-click searches rise when features answer queries directly; being cited can still drive qualified traffic



AI Overviews: The New Front Door to Search



- AI Overviews synthesize multi-source answers with citations, appearing above organic results for complex queries.



- Citation eligibility requires indexing and top-ranking positions; content outside top results is rarely cited.



- AI crawlers like GPTBot and ClaudeBot do not render JavaScript, so SSR/SSG is essential for AI visibility.



- Content structured for featured snippets also strengthens AI Overview citation potential.



Practical Audit: Crawlability and Indexability



- Check Crawl Stats for 90-day trends, host status, and 5xx/server errors.



- Cross-reference Page Indexing report to find crawled-but-not-indexed causes.



- Monitor server health: TTFB, DNS, and error rate above 0.1%.



- Use URL Inspection Tool for rendered HTML, canonical, and structured data.



- Validate JSON-LD with Rich Results Test; fix syntax errors like trailing commas.



Structuring Content for Modern Indexability



- Semantic HTML and clean IA give crawlers a clear content hierarchy for priority



- Strong internal linking distributes crawl budget across your most valuable pages



- JSON-LD must be syntactically perfect: trailing commas, smart quotes, or missing braces break parsing



- SSR/SSG as the foundation ensures all crawlers (Google, GPTBot, ClaudeBot) see complete content



- Client-side JavaScript only for interactivity—never for critical content or structured data



Client-Side JS

Interactivity
and dynamic
experiences



Internal Links

Distributes crawl
budget to your
most valuable
pages



SSR / SSG

The foundation
ensures all crawlers
see complete
content

Diagnosing and Recovering from Ranking Drops

- Cross-reference algorithm update timelines with site changes and traffic dips
- Identify site-wide quality issues vs. page-level technical regressions
- Prune thin content, improve engagement, and demonstrate real expertise
- The Helpful Content System evaluates sites holistically, not page-by-page
- Recovery requires sustained, site-wide quality improvement over months



Key Takeaways and Action Plan for Technical Creators



- Crawl budget governs entry; indexability determines eligibility; quality and E-E-A-T drive ranking.



- Fix server health, eliminate crawl waste, and build topical authority as a continuous process.



- Verify rendering: SSR/SSG as foundation, client-side JS only for interactivity.



- Optimize for entity understanding and Information Gain, not just keywords.

PRE-FLIGHT CHECKLIST



Server Health



Crawl Waste



Topical Authority

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/c368097d/public