

-

Context

(not Content) **is King**
in AI Search

Lazarina Stoy

Founder, MLforSEO | Marketing & ML Consultant

@lazarinastoy

What We'll Cover

01

Setting the Stage

The shift from content to context

02

The Mechanics of AI Search

Fan-out, personalization, entities, reasoning

03

Old Rules, New Game

What carries over, what's different

04

A Framework for Both Worlds

Winning in traditional and AI search



**We've been told
'Content is King'
for 30 years.**



**And for 30 years,
that was mostly true.**

•

**But two users now type the exact
same query...**

•

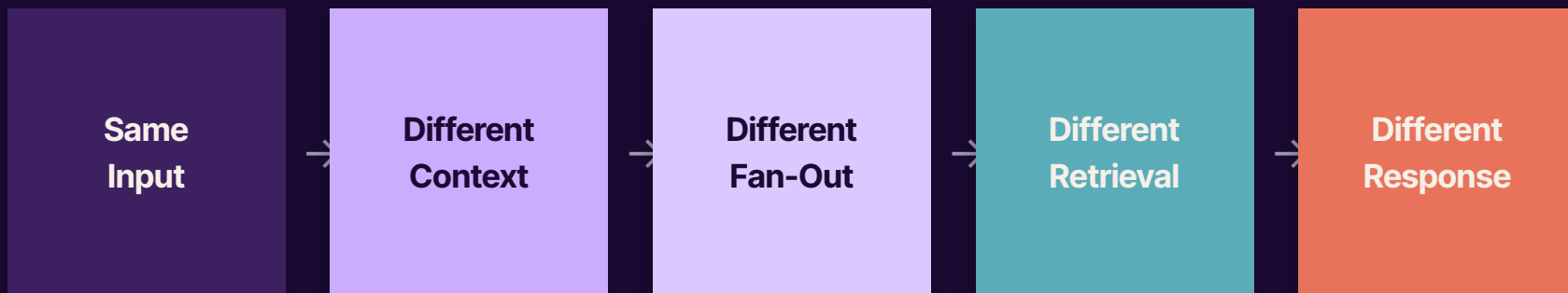
...and get completely different answers.

-

Not because the rankings changed.

Because the questions the system asked on their behalf were different.

The Shift



Two users, identical query. The system generates different sub-questions based on what it knows about each user — before retrieval even begins.



**Different context
leads to different questions
leads to different answers.**

Same query. Different fan-out.

"best running shoes"

User A — Experienced marathoner, urban

- marathon running shoes cushioning high mileage
- carbon plate marathon shoes performance
- best road running shoes experienced runners 2025
- lightweight racing flats urban marathon sub-4 hour
- running shoes durability concrete asphalt

Surfaces: Nike Alphafly, Asics Metaspeed, Saucony Endorphin Pro

User B — Beginner, near trails

- marathon training shoes beginners mixed terrain
- trail to road running shoes new marathoners
- cushioned shoes long distance beginners dirt paths
- durable running shoes marathon prep unpaved
- marathon race shoes gravel packed dirt

Surfaces: Hoka, Brooks, Salomon

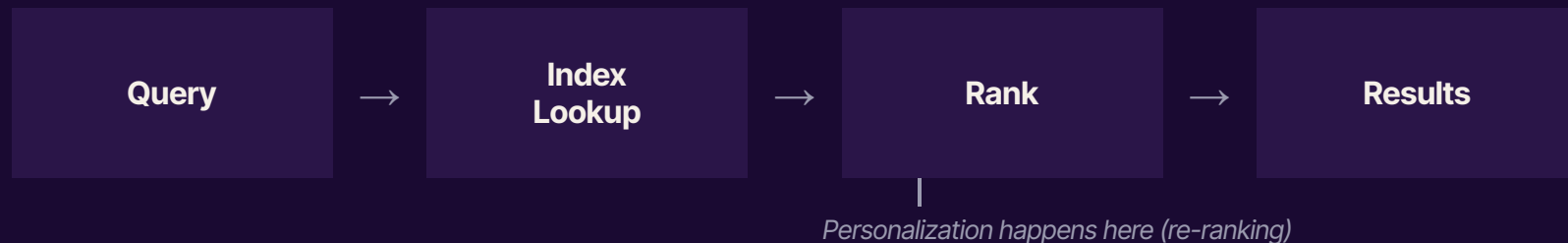
•

This isn't new technology.

It's the same building blocks — NLP, embeddings, knowledge graphs, retrieval, ranking — reassembled in a different order.

Same ingredients. Different recipe.

Traditional Search



AI Search



- **Traditional search personalizes the answers.**

AI search personalizes the questions.

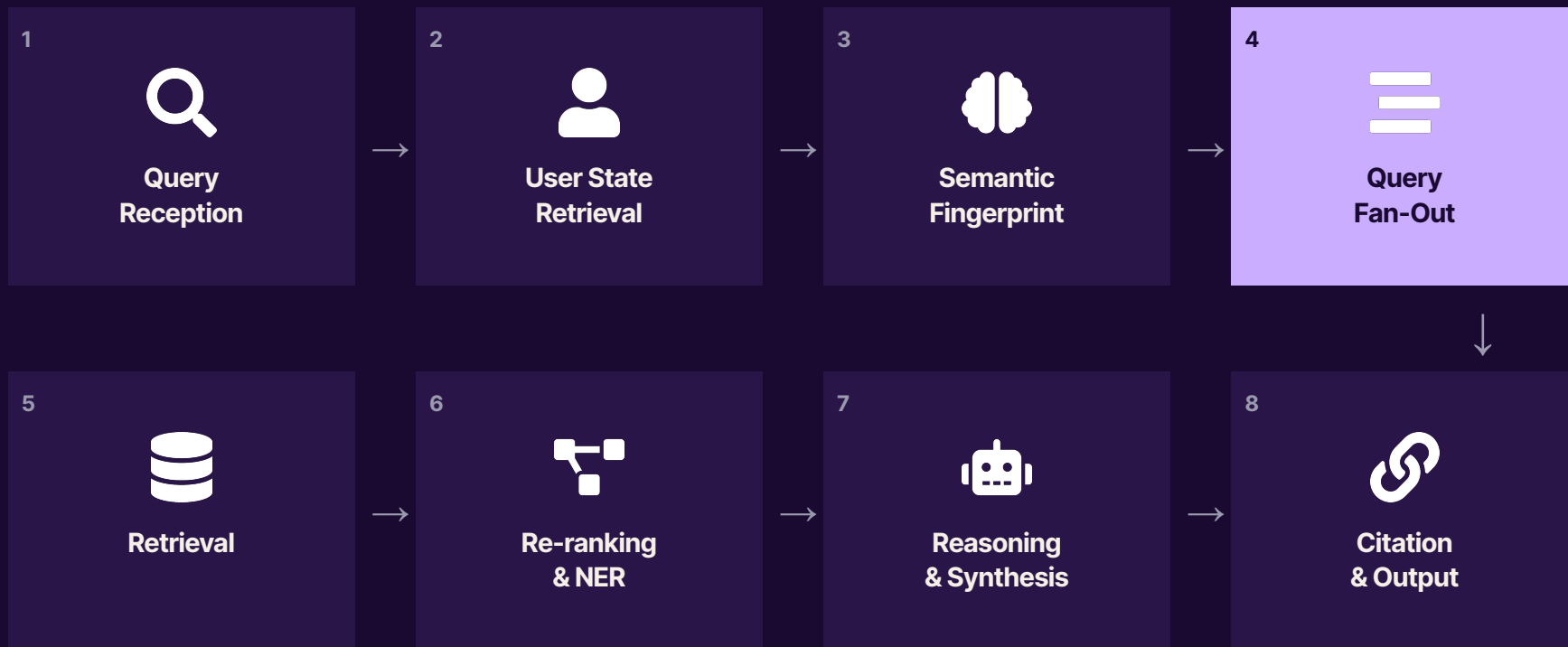
•

And that changes everything.

Part 01

The Mechanics of AI Search

The AI Search Pipeline



Step 1

Query Reception & Routing

- The system receives your query — text, voice, image, or multimodal input
- NLP algorithms parse intent, assess complexity, and determine routing
- Non-text inputs are converted via ML models (image captioning, object detection, speech-to-text)
- The system classifies the query into response categories: needs summarization, needs creative generation, needs clarification, do not interfere

Sources: Google patents US20240289407A1, US11769017B1, US20250124067A1

- **Simple queries get simple answers.**

Complex queries trigger an entire research operation.

Query Complexity Determines Processing Depth

Simple factual query

"Capital of Germany" → Knowledge Graph lookup → "Berlin"

No fan-out. No decomposition. Direct entity-attribute retrieval.

Complex multi-criteria query

"Bluetooth headphones comfortable for runners with long battery"

- Entity recognition (headphones, Bluetooth, runners)
- Attribute extraction (comfort, battery life, form factor)
- Full fan-out: 10-20+ synthetic sub-queries fired in parallel
- Retrieval from web, shopping graph, knowledge graph
- Reasoning, synthesis, citation

Step 2

User State Retrieval

Before generating any sub-queries, the system aggregates everything it knows about you:



Prior queries & session history



Location, device, time of day



Stored preferences & attributes



Browsing behavior & dwell time



Calendar, active apps, tasks



Saved memories & chat history

Traditional Search Already Personalizes — Via 7 Signal Types

Signal Type	What's Tracked	Effect on Rankings
Session context	Query sequence, refinements, topic switches	Carries context forward
Historical behavior	Past clicks, domain preferences	Boosts familiar sources
Information gain	Content you've consumed on a topic	Prioritizes novelty
Implicit feedback	CTR, dwell time, scroll depth	Adjusts relevance scores
Cohort signals	Behavior of similar users	Boosts popular results for group
Environmental	Location, device, time of day	Localizes results
Entity associations	Your Knowledge Graph interactions	Disambiguates entities

All of these operate on ranking — the query itself stays the same across users.

- **Traditional search has always personalized.**

**But it personalized the ranking —
not the query.**



**AI search moves personalization
upstream — into query generation itself.**

Step 3

Semantic Fingerprinting

All contextual information about a user is converted into embeddings — a "semantic fingerprint" that represents their interests, history, and profile.

Two users asking the same query get different embedding representations.

This allows modular personalization: the same question produces different answers based on individual profile alignment and semantic relevance.

User A embedding

[0.82, 0.15, 0.91, 0.03, 0.67, 0.44, ...]

High: performance, tech, urban, advanced

Low: budget, beginner, outdoor

User B embedding

[0.11, 0.88, 0.22, 0.79, 0.34, 0.91, ...]

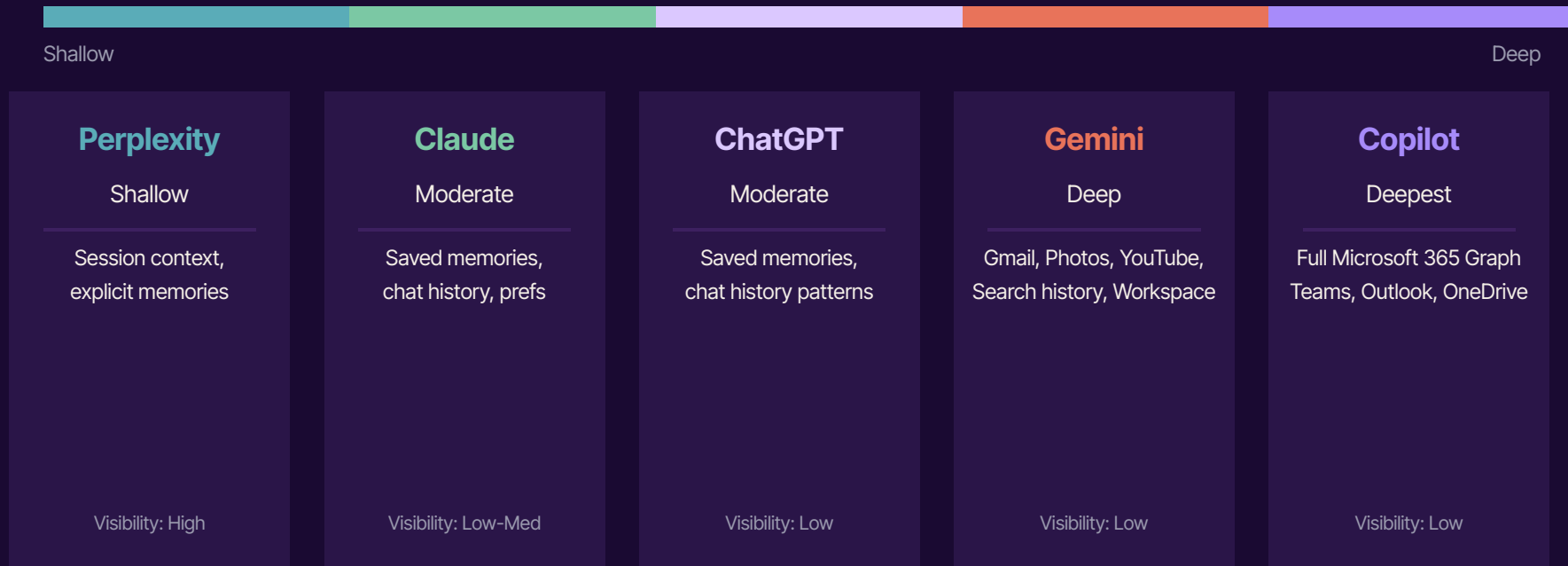
High: beginner, outdoors, budget, trail

Low: performance, tech, road

Google patent: User Embedding Models for Personalization (WO2025102041A1)

The Spectrum of Inference Depth

Not all platforms know the same amount about you.



Platform: Perplexity

The Transparent, Session-Based Approach

- Session memory — resets with each new conversation. Clean slate every time.
- Stores explicit "memories" (preferences, interests) you've shared directly.
- Built on Vespa.ai — integrates retrieval, ranking, and ML inference in a single low-latency pipeline. Processes 200M+ queries daily at 358ms median latency.
- Pro Search shows sub-queries visibly — you can see the fan-out happening.
- Key implication: Limited personalization depth means two new users get similar fan-outs. Expansions tend toward the generic.

Platform: ChatGPT

Explicit Facts Plus Pattern Inference

- Multi-layered memory: system instructions → session metadata → user memory → chat history → current message.
- Two memory modes: saved memories (explicit) and chat history insights (inferred patterns).
- Memory rewrites search queries. OpenAI example: user known to be vegan in SF → query "restaurants near me" becomes "good vegan restaurants, San Francisco."
- The more you interact, the more specific the fan-out becomes — and the more invisible to competitors.

Platforms: Gemini & Copilot

Deep and Deepest Inference

- Gemini (if enabled): integrates Gmail, Photos, YouTube, Search history, Google Workspace. Infers from your entire digital footprint within Google's ecosystem.
- Copilot: accesses the full Microsoft 365 Graph — Teams, Outlook, OneDrive, Calendar. Grounds queries against your organizational data.
- Microsoft's approach: iterative grounding via Orchestrator + Prometheus, not a single parallel burst. Each result informs the next query.
- Key implication: Most context is implicit — the system knows things you never explicitly told it. This is the furthest point from traditional search.

•

**The more a user interacts with an AI system,
the more specific — and more invisible to you
— the fan-out becomes.**

Step 4

Query Fan-Out

The system expands your single query into dozens of synthetic sub-queries — each targeting a different facet of your intent, shaped by your context.

This is the most significant shift in search since mobile-first indexing.

- Fan-out uses trained generative models on real query-document pairs
- Sub-queries are personalized using user attributes, temporal signals, and task context
- Google can fire hundreds of synthetic queries for complex requests (Deep Search)
- All major platforms use this: Google (explicitly), Perplexity (hybrid retrieval), ChatGPT (reformulation), Copilot (iterative grounding)

Patents: Thematic Search (US12158907B1), Generating Query Variants (US11663201B2)

How Fan-Out Expands a Single Query



The 8 Types of Synthetic Query Variants

Equivalent

Alternative phrasing for the same question

Follow-up

Logical next questions

Generalization

Broader versions of the query

Specification

More detailed, narrower versions

Canonicalization

Standardized phrasing

Translation

Multilingual content retrieval

Entailment

Implied or logically following questions

Clarification

Questions back to user to confirm intent

-

Your content either fits the personalized query expansion for a given user, or it doesn't exist in their consideration set.

Fan-Out with Entity-Based Reformulation

The system doesn't just expand words — it expands entities via Knowledge Graph anchors.

Query: "best SUV"

Tesla Model Y (Product entity)

SUV safety ratings 2025 (Attribute)

Volkswagen ID.4 (Product entity)

electric SUV range comparison (Attribute)

Toyota RAV4 (Product entity)

family SUV cargo space (Attribute)

Hyundai Tucson (Product entity)

SUV fuel efficiency hybrid (Attribute + Type)

LLMs crosswalk entity references to broader/narrower equivalents using KG anchors.

SUV → specific models, attributes, comparisons — all via entity relationships.

How Fan-Out Can Skew Intent

Semantic Drift Through Generative Expansion

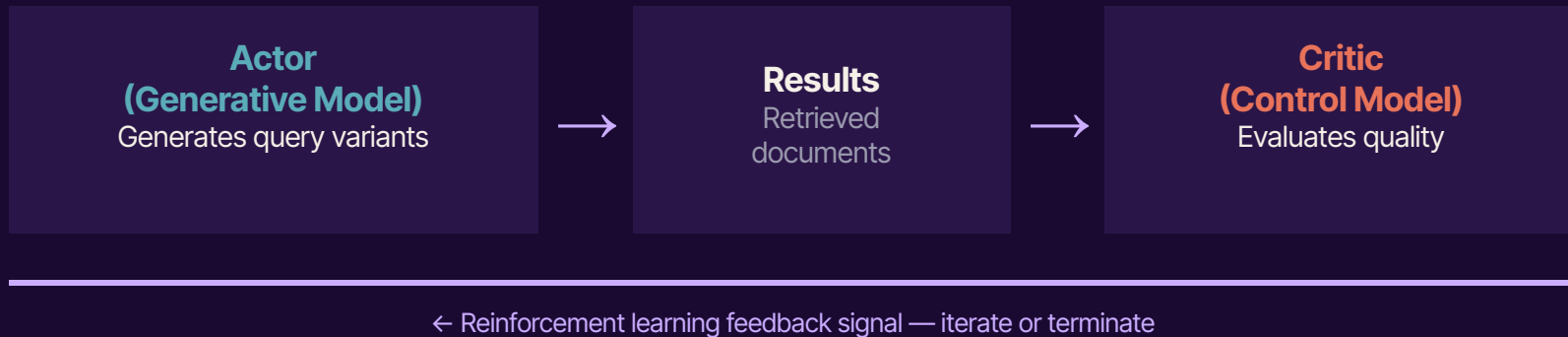
- Follow-up, generalization, and entailment queries can push retrieval away from the user's original narrow intent.
- The LLM traverses semantic relationships that may feel adjacent but stray from the original query — the system moves through vector space neighbors, not keyword matches.
- In traditional search, users decide whether to explore adjacent intents. In AI search, the system decides for them.
- Thematic Search (Google patent) generates themes from responsive documents, then creates drill-down queries — replacing the user's original query with synthetic, topic-specific ones.
- This can erase authentic niche content (e.g., travel blogs) and replace it with aggregated "top picks" lists.

- **Traditional search shows you options and lets you decide.**

AI search decides for you.

The Iterative Refinement Loop

Fan-out doesn't happen in one pass. An Actor-Critic loop refines until quality thresholds are met.



Loop terminates when:

Quality threshold met

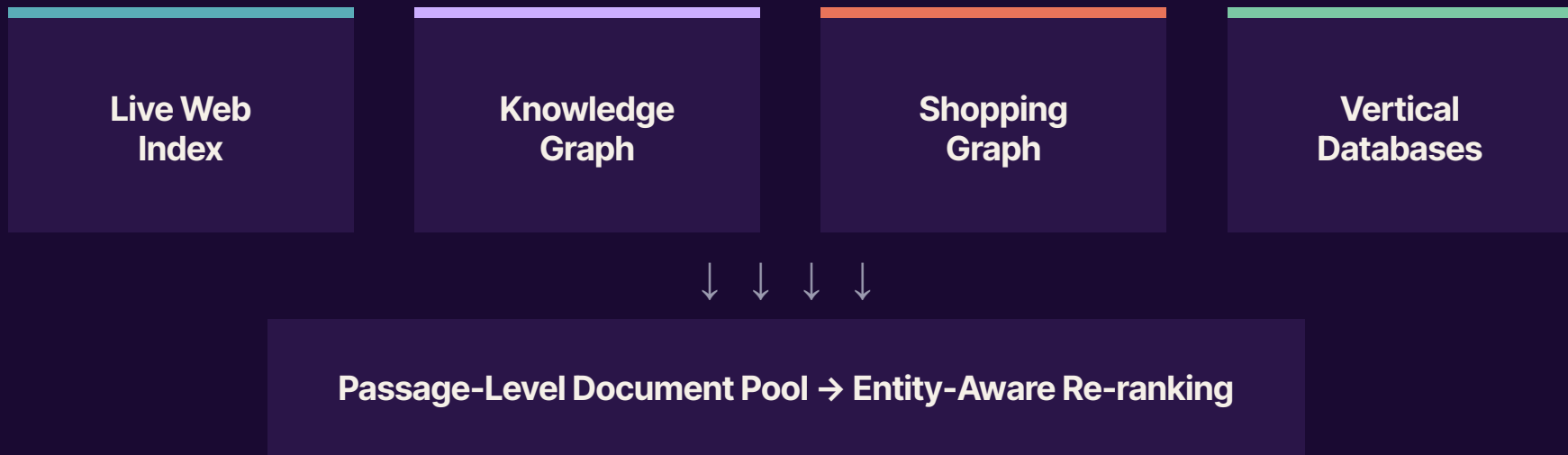
~20 iterations reached

Improvements flatten

Step 5

Parallel Multi-Source Retrieval

Fan-out queries fire simultaneously across multiple sources. Hybrid search: dense vectors (semantic similarity) + sparse vectors (lexical matching).



The RAG Framework Behind It All

Retrieval-Augmented Generation (RAG)

- Originally proposed by Lewis et al. (2020), RAG combines retrieval mechanisms with generative language models to reduce hallucination and improve factual accuracy.
- Modern RAG systems decompose multi-hop queries into sub-questions (RQ-RAG, Chan et al. 2024), combine results via reciprocal rank fusion (RAG-Fusion, Rackauckas 2024), and adapt retrieval based on question complexity (Adaptive-RAG, Jeong et al. 2024).
- Chunking strategies (fixed-size, recursive, layout-aware) determine how documents are split for passage-level retrieval — this is why your content structure directly impacts retrievability.
- Key insight: A well-structured document is easier for retrieval models to ground claims against. Your paragraphs are the unit of competition, not your pages.



**Retrieval is now passage-level,
not page-level.**

The Role of Entity Recognition in AI Search

NER + Entity Linking + Knowledge Graphs

- Named Entity Recognition (NER): scans text, labels terms as Person, Organization, Product, Location, Date, etc. Turns unstructured copy into structured fragments machines can reason about.
- Entity Linking (EL): maps each mention to a canonical entity ID (Wikidata Q-ID, Google KG MID). Resolves ambiguity — 'Jordan' the person vs. the country vs. the product.
- Knowledge Graphs: provide the relationship web — entity types, attributes, connections. Validated against Wikidata, Freebase, DBpedia, and proprietary KGs.
- These three systems work together across every stage of the pipeline: query understanding, fan-out, retrieval, re-ranking, synthesis, and citation.

Where Entities Work Across the Pipeline

Query Understanding

Detect entities in query; form context embedding

Personalization

Map user attributes to KG clusters for boosting

Fan-Out

Crosswalk references to broader/narrower equivalents

Retrieval

Filter by entity type + canonical IDs

Re-ranking

Bias ranking with type constraints and KG proximity

Synthesis

Grounded generation with entity-checked facts

Citation

Semantic matching of generated text to source passages

•

AI search systems understand things, not strings.

Query Decomposition vs. Reformulation

"Compare Pixel 9 camera to iPhone 16 and suggest accessories for hiking"

- | | | |
|---|---|--------------------------------|
| 1 | Retrieve Pixel 9 camera specs & reviews | <i>Pixel 9 (Product)</i> |
| 2 | Retrieve iPhone 16 camera specs & reviews | <i>iPhone 16 (Product)</i> |
| 3 | Synthesize side-by-side comparison | <i>Camera (Attribute)</i> |
| 4 | Retrieve hiking-use accessories for chosen device | <i>Hiking (Activity)</i> |
| 5 | Aggregate and rank recommendations | <i>All entities reconciled</i> |

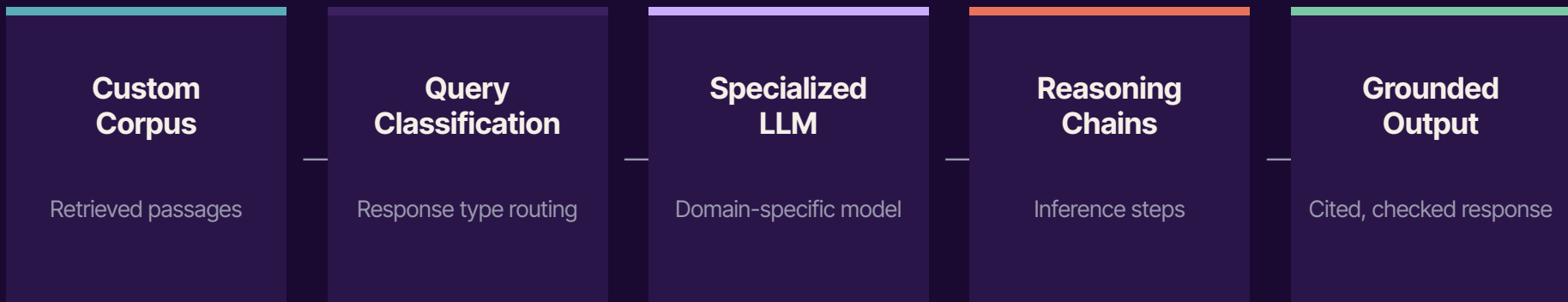
Chunk boundaries align with the EAV model — splitting by entity/aspect makes retrieval cleaner and synthesis more precise.

Step 6

Reasoning & Synthesis

- Query classification routes to specialized downstream LLMs — each trained for a particular response type (creative text, summarization, clarification, etc.).
- Reasoning chains: structured sequences of intermediate inferences connecting user queries to responses logically. Content must align with each logical step to be selected.
- Grounded generation: extracts chunks from documents, builds structured representations, synthesizes. Includes recitation and attribute checking from sources.
- Personalized summarization: the response omits content the user is already familiar with, based on their user state.
- Multimodal output: text, video, audio, imagery, dynamic visualizations — the system can transcribe, extract claims, and remix across formats.

Multi-Stage LLM Orchestration



Classification categories:

- Needs creative generation
- Needs summarization
- Needs clarification
- Benefit from suggested next step
- Do not interfere

Grounded Generation

How the LLM decides what to say — and what to cite

- The LLM extracts chunks from relevant documents, builds structured representations, and synthesizes a coherent response.
- Grounding: anchors claims to specific retrieved passages. Recitation: reproduces key facts accurately. Attribute checking: verifies names, specs, relationships from source documents.
- Content chunking strategies (fixed-size, recursive, layout-aware) determine which pieces of your content are independently retrievable.
- Multiple parallel sub-queries retrieve semantically rich passages that anchor different aspects of the response — reducing hallucination risk.
- Academic note: Corrective RAG (CRAG, Chen et al. 2024) adds adaptive evaluation of retrieved context quality before generation.

•

The system doesn't just retrieve your content.

It reasons about whether to cite you.

Citation & Linkification

Step 7: How sources get credited (or don't)

- Semantic embeddings of AI-generated text are compared against source document embeddings to verify verifiability and closeness.
- Sources are benchmarked — excluded from citing if not sufficiently close to the generated claim. Not every source used in retrieval gets cited.
- Links can point to specific passages/sentences or to entire documents — passage-level citation is the new unit.
- Key insight: Your content may inform a response (used for context, retrieved during fan-out) but never be cited. Used for knowledge, not for credit.

•

Your content might be used to inform a response — but never cited.

Used for context, not for credit.

The Grounding Problem for Brands

A competitive dynamics shift

- Your authoritative, knowledge-packed content feeds the system's understanding — those ultimate guides, that in-depth evergreen content. But the system may determine your brand is not the best offering for this particular user.
- Smaller brands — ones you've dismissed as non-competitors — can resurface and capture leads simply because they communicated their offer in a more focused, entity-rich way, and that offer is exactly what the user needs.
- Traditional competitive signals (links, domain authority, traffic scale) aren't irrelevant — they're just less decisive. What matters more is being the right answer for this user, based on everything the system knows about them.
- This is exactly what relevance engineering and technical branding aim to solve.

•

You're not optimizing for a position anymore.



You're optimizing for inclusion in the question set — and then for a citation.

What This Means

For Measurement

There's no stable query to rank for. Your visibility becomes a distribution across user contexts, not a position on a page.

For Competitive Analysis

Brands you've dismissed as non-competitors capture users whose personalized fan-outs align with their focused positioning.

For Content Strategy

Keyword optimization alone is insufficient. You need to optimize for the range of questions different user contexts might generate around your topic.

Key References & Sources

Google Patents

Thematic Search (US12158907B1) | User Embedding Models (WO2025102041A1) | Generating Query Variants (US11663201B2) | Synthetic Query-Document Pairs (WO2024064249A1) | AI Mode Processing (US20240289407A1)

Academic Research

Lewis et al. (2020) — RAG for Knowledge-Intensive NLP | RAG-Fusion (Rackauckas, 2024) | RQ-RAG (Chan et al., 2024) | Adaptive-RAG (Jeong et al., 2024) | CRAG (Chen et al., 2024) | KRAGEN (Matsumoto et al., 2024)

iPullRank & MLforSEO Resources

Stoy, L. — How AI Search Personalizes Fan-Out Queries (iPullRank, 2026)

Stoy, L. — How AI Search Platforms Expand Queries with Fan-Out (iPullRank, 2025)

Stoy, L. — How AI Search Platforms Leverage Entity Recognition (iPullRank, 2025)

iPullRank AI Search Manual — ipullrank.com/ai-search-manual

Platform Documentation

OpenAI — ChatGPT Memory & Search | Perplexity — Search API Architecture | Microsoft — Azure AI Bing Grounding | Google — AI Mode Blog Post (2025)

Part 02

Old Rules, New Game

-

We've walked through the mechanics.

Now let's talk about what this means for the work you're already doing.

What Carries Over

These fundamentals still matter — they're the building blocks AI search systems are built on.

E-E-A-T

Experience, Expertise, Authority, Trust — still the quality benchmark. LLMs prioritize authoritative, trustworthy sources for grounding.

Topical Authority

Depth and breadth of coverage on a subject still determines whether you're retrieved. Topical clusters remain essential.

Structured Data

Schema.org markup, sameAs links, canonical entity IDs — now more important than ever for disambiguation and KG grounding.

Entity Clarity

Clear, unambiguous entity definitions carry directly into AI search. Named entities with stable IDs are the currency of retrieval.

Content Quality

Well-written, factually accurate, information-rich content is still what gets retrieved. The bar for quality hasn't lowered — it's risen.

Personalisation Carries Over — But Amplified

Google has used implicit user feedback to personalize rankings since the early 2000s. Patents going back to 2002–2007 document how they track and use this data.

Traditional Search Signals

- Click-through rates (CTR)
- Dwell time / long clicks
- Pogo-sticking detection
- Mouse hover & scroll patterns
- Query reformulation patterns
- Device-specific ranking adjustments
- Browsing-time factor (patent US2003)

AI Search Adds

- **All of the above, plus:**
- Persistent memory across sessions
- Saved user preferences & facts
- Cross-product data (Gmail, Calendar...)
- Semantic fingerprinting (embeddings)
- Fan-out shaped by inferred context
- **Personalized query generation, not just re-ranking**

The Patent Trail: Implicit User Feedback Since 2002

2002	Query Logs	Tracks user actions (clicks, time spent) to find query synonyms
2003	Historical Data	Analyzes changes in user time on page to mark content freshness
2003	Location Awareness	Uses clicks in local searches for geographically relevant results
2004	Mouse Hover Tracking	Tracks mouse behavior to predict and reorder link relevancy
2005	Repository Choices	Uses data triples to recommend media types in search
2007	Personalized Demotion	Ranks results based on individual user actions
2011	Implicit User Feedback	Re-ranks based on long views vs short views of document results
2025	User Embedding Models	Context converted to embeddings — semantic fingerprints for AI search

The mechanisms haven't changed. The scale and depth of their application has.

-

Same signals.

More data.

Different application point.

Google's Data Ecosystem: Chrome, DOJ, and Beyond

What the antitrust trial revealed

- Chrome serves as a major data-collection tool — browsing history, location, search preferences, product interactions — all feeding ranking and personalization.
- SERP interactions go beyond clicks: scrolling patterns, mouse hover, "read interactions" capturing attention to specific page sections.
- Data is linked across Google's ecosystem — Gmail, Maps, Google Pay — enabling detailed user profiles for personalized results.
- Google's own testimony: BERT and MUM are becoming more important than user data — the system is shifting toward ML-based text evaluation over pure behavioral signals.
- This is the bridge: the same behavioral signals that power traditional search now feed the embedding models and semantic fingerprints that power AI search personalization.

What's Different

The fundamental shifts that break the old playbook.

FROM

TO

Ranking for a position



Being included in the question set

Keyword matching



Entity + relationship matching

Page-level retrieval



Passage-level retrieval

Static SERP



Dynamic, personalized synthesis

Personalized ranking



Personalized query generation

From "Ranking" to "Being Referenced"

The competitive model has fundamentally shifted

- In traditional search, you compete for a position within a stable results set. In AI search, you compete for inclusion in the query set — and then for a citation.
- Your content might perfectly answer one user's personalized fan-out and be completely invisible to another — not because it ranked poorly, but because the system never asked a question that would surface it.
- A brand like Salomon, traditionally seen as a trail specialist, suddenly competes for marathon-prep queries — but only for users whose context suggests trail-based training. Meanwhile, a road running brand with objectively better marathon expertise may never surface for that user.
- Citation-based visibility may eventually rival links. Being mentioned and cited in AI responses could become equivalent to backlinks in traditional search: a signal that drives awareness, trust, and indirect conversion.

From Keywords to Entities + Relationships

The unit of optimization has changed

- Traditional search matches keywords to pages. AI search matches entities, attributes, and relationships to passages.
- Entity-based reformulation means the system expands "SUV" to specific product entities (Model Y, ID.4) using Knowledge Graph anchors — not keyword synonym lists.
- Canonical entity IDs (Wikidata Q-IDs, Google KG MIDs) deduplicate synonyms, enable cross-language discovery, and let AI systems consistently identify your brand across queries.
- The EAV model (Entity-Attribute-Variable) drives query decomposition: chunk boundaries align with entities and their attributes, making retrieval cleaner and synthesis more precise.
- Generic, thematic content no longer converts to visibility. Your passages must be granularly useful and independently retrievable.

From Page-Level to Passage-Level Competition

Traditional Search

Unit of competition: the page

- Google indexes entire documents
- Ranking factors apply to the whole page (links, DA, on-page SEO)
- User sees a list of 10 blue links
- One page, one ranking, one click

AI Search

Unit of competition: the passage

- System chunks documents into passages via semantic boundaries
- Each passage independently scored for entity relevance, semantic match, factual grounding
- User sees a synthesized answer citing specific passages
- **Your paragraph is the unit of competition, not your page**

The Grounding vs Citation Gap

Used for context ≠ credited as a source

- AI search systems might use your content to inform a response — those authoritative guides, that in-depth evergreen content — yet stop there. They determine your brand is not the best offering to surface to this particular user.
- Other brands — ones you've dismissed as non-competitors — resurface and capture leads simply because they communicated their offer in a more focused, entity-rich way that matches this user's context.
- Traditional competitive signals (links, DA, traffic scale) aren't irrelevant — they're just less decisive. What matters more is being the right answer for this user.
- If the system decides your brand is not relevant to the conversation with this particular user, your brand won't just fail to surface — it won't even be part of the consideration set. The system won't ask the questions that would lead to your content.

•

Your brand can be invisible — not because you ranked poorly, but because the system never asked a question that would surface you.

From Deterministic to Probabilistic Ranking

- Traditional search is largely deterministic — the same query produces roughly the same results for the same user segment. You can track a ranking.
- AI search is stochastic — follow-up questions generated by fan-out vary across users, influenced by history, device, location, preferences, and prior queries. The same query generates different answers on different days.
- This breaks traditional measurement. You can't track a "ranking" when there's no stable query to rank for. Your visibility becomes a distribution across user contexts, not a position on a page.
- Traditional SEO tools designed for static keyword-to-page matching cannot predict, measure, or optimize for contextual personalization at this scale.

What Breaks in the Old Playbook

✗ **Keyword-based content clustering**

Fan-out decomposes queries into entity-driven sub-queries, not keyword matches. Broad topic coverage may fragment across personalized variants.

✗ **Static rank tracking**

No stable query to rank for. Visibility is a distribution across user contexts — stochastic and user-dependent.

✗ **Competitive analysis by SERP position**

Brands you've dismissed capture users whose fan-outs align with their focused positioning. Your competitors are now contextual.

✗ **Content = pages, optimized individually**

The unit of retrieval is the passage. Each paragraph must be independently retrievable and entity-anchored.

•

AI search allows smaller brands to compete with larger ones — if their content is communicated in a focused, structured, entity-rich way.

The New Competitive Equation

BEFORE

Links + Domain Authority + Traffic Scale + Keyword Targeting
= Ranking Position = Visibility

NOW

**Entity Clarity + Contextual Relevance + Passage Quality
+ Structured Data + Source Authority
= Inclusion in Fan-Out = Citation = Visibility**

•

You're not optimizing for a position.

You're optimizing for a citation.

Part 03

A Framework for Winning in Both Worlds

•

**Don't choose between traditional SEO
and AI search.**

Let context strengthen both.

The Omnichannel Visibility Framework

Four pillars that work for both traditional search and AI search — simultaneously.

01

**Entity
Clarity**

02

**Contextual
Depth**

03

**Source
Authority**

04

**Omnichannel
Consistency**

Works for Traditional Search ↔ Works for AI Search

01

Entity Clarity

Make your brand, products, and concepts unambiguous to machines.

Entity Clarity: What To Do

For Traditional Search

- Define clear entity names — consistent across pages, titles, H1s, alt text. Run NER APIs (Google Cloud NLP) to verify entities are correctly identified and typed.
- Implement Schema.org markup with sameAs links to Wikidata, LinkedIn, and Knowledge Graph IDs. Use @id for internal entity linking.

For AI Search

- Canonical entity IDs enable deduplication across languages and synonyms. When pages link entities to public IDs, AI systems disambiguate your brand and consolidate related pages.
- Entity-based query reformulation expands queries using KG anchors — if your entities aren't clearly defined, you won't be included in the expansion.
- Disambiguate in the first paragraph. Clarify which entity you mean. LLMs resolve ambiguity early.

Entity Audit: Quick Checklist

- ☐ Run top 10 pages through NER API — are primary entities salience > 0.5?
- ☐ Do all primary entities have Knowledge Graph IDs (MID values)?
- ☐ Is entity naming consistent across all pages? (not "Acme Corp" vs "ACME" vs "Acme")
- ☐ Are entity types correctly identified? (Product not labeled as "OTHER")
- ☐ Does every key entity have a dedicated hub page?
- ☐ Are relationships bidirectional? (Person→worksFor→Org AND Org→employee→Person)
- ☐ Is Schema.org markup deployed with sameAs to Wikidata, LinkedIn, GBP?
- ☐ Competitive gap: which entities do competitors cover that you don't?

02

Contextual Depth

Go beyond keyword coverage — build contextual richness through relationships, co-occurrence, and entity networks.

Contextual Depth: What To Do

For Traditional Search

- Build topic-entity networks, not just topic clusters. Transform pillars + supporting articles into entity-connected networks with explicit schema relationships (isPartOf, isRelatedTo, knowsAbout).
- Cover semantic and intent variations of queries: synonym/phrase variants, intent modifiers, brand-focused, location/industry, question-form, negative/pain-point focused.

For AI Search

- Fan-out generates 8 variant types per query. Your content must answer the range of questions different user contexts might generate — not just one keyword.
- Write passage-level atomic facts anchored to entities. Each paragraph should be independently retrievable and citation-worthy.
- Design for Information Gain: the system prioritizes novelty based on what the user already knows. Unique insights, original data, and specific frameworks get retrieved over generic coverage.

From Topic Clusters to Entity Networks

Topic Cluster Model

- Pillar page + supporting articles
- Organized by keyword themes
- Internal links between pages
- No explicit entity relationships
- Pages are the unit of organization
- Works for traditional SEO ✓
- Fragments across AI fan-out ✗

Entity Network Model

- Entity hubs + supporting content
- Organized by entity relationships
- @id linked schema across pages
- Explicit: isPartOf, knowsAbout, about
- Passages are the unit of retrieval
- Works for traditional SEO ✓
- Coherent across AI fan-out ✓

03

Source Authority

Be the source LLMs trust — through citations, structured data, and consistent entity signals across platforms.

Source Authority: What To Do

For Traditional Search

- E-E-A-T remains the quality benchmark. Author attribution, expert bylines, about pages, and trust signals still determine what gets ranked.
- Build backlinks from authoritative sources. Co-citation with established authorities transfers credibility.

For AI Search

- LLMs cite content they see used in multiple authoritative sources. Multiple platforms = stronger authority signal. Establish entities on LinkedIn, Wikidata, industry sites — exact same names from your KG.
- Create citation-worthy answers: Q&A format, specific data, solve real problems. FAQPage schema with answers attributed to Person entities.
- Co-citation building: get your entities mentioned alongside established authorities in comparisons, alternatives, and category content. When LLMs see your entities repeatedly in context with known authorities, they infer similar authority.

Brand Narrative Architecture for LLMs

LLMs synthesize narratives from thousands of data points. You can engineer how your brand is described.

Layer 1

Core Identity Statements

5-7 canonical sentences defining your brand. Same phrasing verbatim across About page, LinkedIn, press releases, Wikidata, guest bios, podcast intros.

Layer 2

Attribute Association

8-10 attributes you want LLMs to associate with your brand. For each: 1 deep article, 3 social proof examples, 5 third-party citations, 2 comparison pieces.

Layer 3

Context-Specific Positioning

LLMs respond differently by query context. Create positioning for: vs. enterprise, vs. startups, specific use cases, budget-conscious, security-focused.

04

Omnichannel Consistency

Strengthen the message across channels without fragmenting — what you say on your site, LinkedIn, Wikidata, and guest content must be the same entity.

Omnichannel Consistency: What To Do

For Both Traditional and AI Search

- ONE canonical name everywhere. "Dr. Emma Chen" on ALL pages — not "Emma Chen PhD," not "E. Chen." Consistent naming is the foundation of entity recognition.
- Cross-platform distribution: LinkedIn profiles (exact same name as KG), Wikidata entries, Google Business Profile, industry publications — all with sameAs links in schema.
- Identical properties across all sources: founding dates, addresses, URIs in ISO format. Single source of truth — central data, no duplication.
- Bidirectional relationships: if $A \rightarrow B$ exists, $B \rightarrow A$ must exist. $\text{Person} \rightarrow \text{worksFor} \rightarrow \text{Org}$ AND $\text{Org} \rightarrow \text{employee} \rightarrow \text{Person}$.
- Why this matters for AI: When LLMs see "John Smith" on your site, LinkedIn, guest articles, AND consistent expertise signals — they gain citation confidence. Multiple sources = higher trust = more citations.

Cross-Platform Entity Distribution

LinkedIn	Exact name match. Detailed experience. Weekly expertise posts.	Critical
Wikidata	Create org entry. Add founding date, industry, URL. Add sameAs to schema.	Critical
Google Business	Claim profile. Complete ALL fields. Exact name match.	Critical
Industry Pubs	Guest articles from Tier 1 people. Byline matches entity name.	High
Wikipedia	Entry with proper citations. Connects to Wikidata.	High
Reddit / Forums	Genuine participation. Natural brand mentions in context.	Medium

•

Isolated entities get ignored.

Connected entities get recognized.

Authority flows through relationships.

Your Knowledge Graph Is the Foundation

- Knowledge graphs power rich results, knowledge panels, AI overviews, and can address query fan-out more efficiently.
- Build in tiers: Tier 1 (core brand entities — org, founders, CEO, locations), Tier 2 (authority signals — products, key team, IP, methodologies), Tier 3 (extended network — industry concepts, partnerships).
- Technologies: RDF provides the data model, JSON-LD makes it web-friendly, Schema.org gives us the vocabulary. They work together.
- Implementation: Plugin/SaaS (WordLift, Schema App) for quick deployment, Custom code for full control, Hybrid API for enterprise multi-site.
- Validation: test all 6 checkpoints before deploy — bidirectional links, URI resolution, syntax (0 errors), sameAs links, live JSON-LD, required properties.

Content That Gets Cited by LLMs

For each Tier 1 entity, follow this pattern:

1

Define Expertise Scope

3-5 specific expertise areas per person/product. These become about properties in schema, linking Person to Concept entities.

2

Create Hub Content

One 2,000+ word guide per expertise area. Entity name in title, 5-8 times in body. Specific data, original insights, frameworks.

3

Build Supporting Cluster

5-7 shorter pieces (500-800 words) referencing the hub. Demonstrate ongoing expertise and activity.

4

Add Q&A Format

FAQ pages answering specific questions LLMs receive. Attributed to Person entities via FAQPage schema + author property.

Time investment: 1 hub (4-6 hrs) + 5 supporting (1 hr each) = 10-12 hrs per entity. Do ONE entity/month for 3 months.

Measuring Success: New Metrics for a New World

Metric	How to Measure	Target
Entity Coverage	% of priority entities with hub + 3+ supporting pages	> 80%
Schema Coverage	% of pages with correct, validated schema markup	> 80%
LLM Citation Frequency	Test 10+ industry queries on ChatGPT, Claude, Perplexity weekly	Track trend
Entity Consistency	Same name, same properties across all platforms	100%
Relationship Density	Avg. schema relationships per entity	> 3
Grounding Rate	% of key queries where your content is grounded (used by AI systems)	Track trend

Citation frequency = AI authority score. This is your new SEO.

•

**Great content + proper schema
= exponential visibility.**

One Strategy. Both Worlds.

Action	Traditional Search Benefit	AI Search Benefit
 Build entity clarity with schema + canonical IDs	Knowledge panels, rich snippets	Included in fan-out expansion
 Create passage-level atomic content	Featured snippets, PAA answers	Retrieved and cited by LLMs
 Establish cross-platform authority	Backlink profile, E-E-A-T signals	Multi-source citation confidence
 Maintain omnichannel consistency	Brand recognition, local SEO	Entity disambiguation, KG grounding



Don't choose between traditional SEO and AI search.

Let context strengthen both.

Part 04

What To Do Next

This Week: 5 Things You Can Do Right Now

1

Audit your top 10 pages with a NER API

Run through Google Cloud NLP. Check: are your primary entities salience > 0.5? Do they have KG IDs? Are entity types correct?

2

Check your schema for @id and sameAs

Validate with Google Rich Results Test. Are entities connected bidirectionally? Do sameAs links point to Wikidata, LinkedIn, GBP?

3

Test 10 industry queries on ChatGPT, Claude, Perplexity

Note: which brands get cited? Are you mentioned? What entities appear most? This is your citation baseline.

4

Write one passage-level, entity-attributed answer

Pick your most important query. Write a paragraph that directly answers it with specific data, attributed to a named person. Add FAQPage schema.

5

Align one entity across 3 platforms

Pick your CEO or brand. Ensure exact same name on your site, LinkedIn, and Wikidata. Add sameAs links. This takes 30 minutes.

30-60-90 Day Roadmap

Days 1-30

Foundation

- Entity audit: NER + KG ID check on top pages
- Deploy Tier 1 schema (Org, CEO, primary products)
- Establish cross-platform presence (LinkedIn, Wikidata, GBP)
- Set LLM citation baseline (10 queries, 3 platforms)

Days 31-60

Authority

- Create first entity hub content (2,000+ words, attributed)
- Build 5 supporting pieces per hub entity
- Implement bidirectional @id relationships across schema
- First competitive entity gap analysis

Days 61-90

Scale

- Expand to Tier 2 entities (team, products, methodologies)
- Deploy brand narrative architecture (3 layers)
- Build co-citation content (comparisons, alternatives)
- Re-measure LLM citations — track trend vs. baseline

Key Mindset Shifts to Take Away

"What keywords should I target?"

→ **"What entities should I be associated with?"**

"How do I rank for this query?"

→ **"How do I get included in the fan-out for this topic?"**

"How do I optimize this page?"

→ **"How do I make each paragraph independently citable?"**

"Who are my SERP competitors?"

→ **"Whose content gets cited for my users' contexts?"**

"Content is King"

→ **"Context is King"**

-

The playbook isn't being rewritten.

It's being expanded.

What Won't Change — No Matter What

- Quality content wins. LLMs filter out promotional fluff just like users do.
- Trust is built through consistency, accuracy, and genuine expertise — across both channels.
- Understanding your audience matters more than ever — now the machines are trying to understand them too.
- The best answer for the user will surface. Make sure that's you.
- Structured data is not optional. It's the language machines speak.



Context is King.

-

And context doesn't replace content.

**It makes your content message stronger —
everywhere.**

Resources & Further Reading

iPullRank AI Search Research

- How AI Search Personalizes Fan-Out Queries — ipullrank.com
- How AI Search Platforms Expand Queries with Fan-Out — ipullrank.com
- How AI Search Platforms Leverage Entity Recognition — ipullrank.com
- iPullRank AI Search Manual — ipullrank.com/ai-search-manual

MLforSEO Academy Courses

- Introduction to Machine Learning for SEOs — academy.mlforseo.com
- Semantic ML-enabled Keyword Research — academy.mlforseo.com
- AI Search & LLMs: Entity SEO and Knowledge Graph Strategies for Brands (by Beatrice Gamba) — academy.mlforseo.com

Tools & APIs

- Google Cloud Natural Language API — cloud.google.com/natural-language
- Google Knowledge Graph Search API — developers.google.com/knowledge-graph
- iPullRank Relevance Doctor — ipullrank.com/tools/relevance-doctor
- Google Rich Results Test — search.google.com/test/rich-results

Thank you.

Lazarina Stoy

Founder, MLforSEO | Marketing & ML Consultant

→ lazarinastoy.com

→ mlforseo.com

→ [@lazarinastoy](https://twitter.com/lazarinastoy)

**Context
(not Content)
is King
in AI Search**