Entity Coverage vs Keyword Overlap

A portable field guide for deciding what each diagnostic reveals - and what neither can prove.

Executive rule: keyword overlap compares shared language; entity coverage compares required subjects; relationship coverage tests the connections; source support tests whether those connections are defensible.

Use keyword overlap when the task is a query-footprint or vocabulary comparison. Use entity coverage when the task is subject completeness. Add relationship coverage when the explanation matters, and source support when consequential connections need evidence.

Question	Use first	Do not conclude
Which queries or terms are shared?	Keyword overlap	Substantive similarity, intent, authority
Does the page represent the subjects required by the question?	Entity coverage	Ranking or citation
Are the required connections actually explained?	Relationship coverage	Source reliability
Are consequential connections backed by inspectable evidence?	Source support	Search or answer-engine selection

The five-layer model

1 Query set	2 Keyword overlap	3 Entity coverage	4 Relationship coverage	5 Source support
Language in scope	Shared normalized strings	Required subjects represented	Required connections stated	Consequential connections evidenced

A later layer cannot be inferred from an earlier one. Shared words do not prove shared entities; shared entities do not prove the same relationships; stated relationships do not prove source support.

Transparent calculations

Declare the normalization rule first. The web comparator and workbook use the same English-language rule: convert tabs, line breaks and non-breaking spaces to normal spaces; trim; collapse repeated spaces; and lowercase. Unicode compatibility folding, stemming, singularization, synonym merging and intent clustering are not applied silently. If you use any of them, apply the same transformation to both sets and record it.

Measure	Formula	Use	Limit
Jaccard similarity	|A intersection B| / |A union B|	Shared portion of the combined query set	Falls when either side has a large unique tail
Overlap coefficient	|A intersection B| / min(|A|, |B|)	Containment of the smaller set	Can look high when one set is very small
Entity coverage	required entities observed / required entities	Subject completeness for this scoped audit	Practitioner-defined diagnostic, not a disclosed ranking factor
Relationship coverage	required relationships observed / required relationships	Explanation completeness	Requires a predefined relationship model
Supported-relationship coverage	required relationships with acceptable evidence / required relationships	Defensibility	Depends on source-role rules

Four-state mismatch matrix

State	Diagnostic meaning	Next action
High keyword / high entity	Language and subject model broadly align.	Inspect relationships, sources and intent.
High keyword / low entity	Lexical false-positive risk.	Add missing subjects/relationships if the page should own the task.
Low keyword / high entity	Lexical false-negative risk.	Test query-language alignment before rewriting the subject model.
Low keyword / low entity	Both layers diverge.	Recheck whether the pages belong to the same task.

Do not use a universal high/low threshold unless the method you are following supplies and validates one. Set audit thresholds from the task and risk before looking at the result.

The seven-step method

1. Define the question and query universe
State task, locale, date, page set and what same task means.

2. Normalize the query sets
Apply the same declared rules; retain raw and normalized forms.

3. Measure lexical overlap
Calculate Jaccard and, when containment matters, the overlap coefficient. Show intersections and unique terms.

4. Build the required entity model
List entities needed to answer the question before reading the page; resolve aliases and ambiguity.

5. Observe entities and relationships
Record which required entities are represented and which required connections are stated.

6. Attach source support and diagnose the mismatch
Retain evidence locators/source roles, then interpret the four-state keyword/entity result.

7. Route the next action and define reopening conditions
Choose language, entity, relationship, source, intent or competitor follow-up; rerun when scope/evidence materially changes.

Operational refusal rules

- Keyword overlap does not establish commercial competition.
- High keyword overlap does not establish substantive similarity.
- Low keyword overlap does not establish substantive difference.
- Entity mention does not establish entity coverage.
- Entity coverage does not establish relationship coverage.
- Relationship coverage does not establish source support.
- Entity coverage does not establish ranking or AI citation.
- Schema does not equal entity coverage.
- Shared entities do not establish the same search intent.

Worked comparison: similar words, different knowledge

Composite editorial example only - not a live benchmark.

Question: Which page better explains how a dental CRM turns an eligible patient-reminder event into a booked appointment while respecting consent and practice workflow constraints?

Layer	Page A	Page B	Interpretation
Query set	dental CRM; patient reminders; recall automation; appointment reminders; dental software; patient reactivation	dental CRM; patient reminders; recall automation; appointment reminders; dental software; recall workflow	Five shared queries, one unique on each side
Overlap	Jaccard 5/7 = 0.714	Overlap coefficient 5/6 = 0.833	Lexical similarity is high for this bounded example
Entities	practice; patient; CRM; reminder	practice; patient; CRM; appointment; consent status; channel; reminder; treatment plan	Page A omits subjects required by the question
Relationships	CRM sends reminders	consent permits channel; appointment status triggers reminder; response creates booking; treatment plan changes follow-up	Page B explains the operating connections
Source support	generic category assertions	product documentation and workflow locators recorded	Support still needs authority/currency review

Interpretation: the lexical layer says the pages are close, but the entity and relationship layers show a material difference for the scoped question. The correct action is not to add more dental CRM synonyms to Page A. First decide whether Page A should answer this workflow question; if yes, add the missing subjects and connections with appropriate support.

Limit: the example does not establish which page would rank, be cited by an answer system, convert better or be treated as a commercial competitor.

Evidence, provenance and handoff

For each consequential relationship retain: subject, predicate, object, evidence locator, source role, observation date, evidence mode, reviewer status and any uncertainty.

Source role	Typical use
Primary technical authority	Platform or standards behavior within its documented scope
Original research	Empirical or methodological evidence from the research authors
Standard	Formal vocabulary, protocol or provenance model
Direct observation	Dated search, browser or answer-surface evidence
Competitor evidence	What a competitor page or product currently does
Secondary practitioner context	Ideas or terminology that require stronger support before consequential claims

Primary references

Google Search ranking systems guide
https://developers.google.com/search/docs/appearance/ranking-systems-guide

Google structured data introduction
https://developers.google.com/search/docs/appearance/structured-data/intro-structured-data

Google AI features and your website
https://developers.google.com/search/docs/appearance/ai-features

Google generative AI optimization guide
https://developers.google.com/search/docs/fundamentals/ai-optimization-guide

Schema.org sameAs
https://schema.org/sameAs

W3C PROV Data Model
https://www.w3.org/TR/prov-dm/

Entity-Relationship Search over the Web
https://arxiv.org/abs/1810.03235

Semantic Search at LinkedIn
https://arxiv.org/abs/2602.07309

Portable handoff checklist

- Save the raw and normalized query sets.
- Save the required entity and relationship model before interpreting coverage.
- Attach evidence locators to consequential relationships and ensure those locators exist in the provenance register.
- Record locale/date/surface for unstable observations.
- Keep ranking, citation, authority and commercial-competition conclusions outside the workbook unless separate evidence supports them.