

Semantic ranking in Azure AI Search

Article • 06/12/2024

In Azure AI Search, *semantic ranking* is a feature that measurably improves search relevance by using Microsoft's language understanding models to rerank search results. This article is a high-level introduction. The section at the end covers [availability and pricing](#).

Semantic ranker is a premium feature, billed by usage. We recommend this article for background, but if you'd rather get started, follow these steps:

- ✓ [Check regional availability](#)
- ✓ [Sign in to Azure portal](#) to verify your search service is Basic or higher
- ✓ [Enable semantic ranking and choose a pricing plan](#)
- ✓ [Set up a semantic configuration in a search index](#)
- ✓ [Set up queries to return semantic captions and highlights](#)
- ✓ [Optionally, return semantic answers](#)

ⓘ Note


Semantic ranking doesn't use generative AI or vectors. If you're looking for vector support and similarity search? See [Vector search in Azure AI Search](#) for details.

What is semantic ranking?

Semantic ranker is a collection of query-side capabilities that improve the quality of an initial [BM25-ranked](#) or [RRF-ranked](#) search result for text-based queries. When you enable it on your search service, semantic ranking extends the query execution pipeline in two ways:

- First, it adds secondary ranking over an initial result set that was scored using BM25 or RRF. This secondary ranking uses multi-lingual, deep learning models adapted from Microsoft Bing to promote the most semantically relevant results.
- Second, it extracts and returns captions and answers in the response, which you can render on a search page to improve the user's search experience.

Here are the capabilities of the semantic reranker.

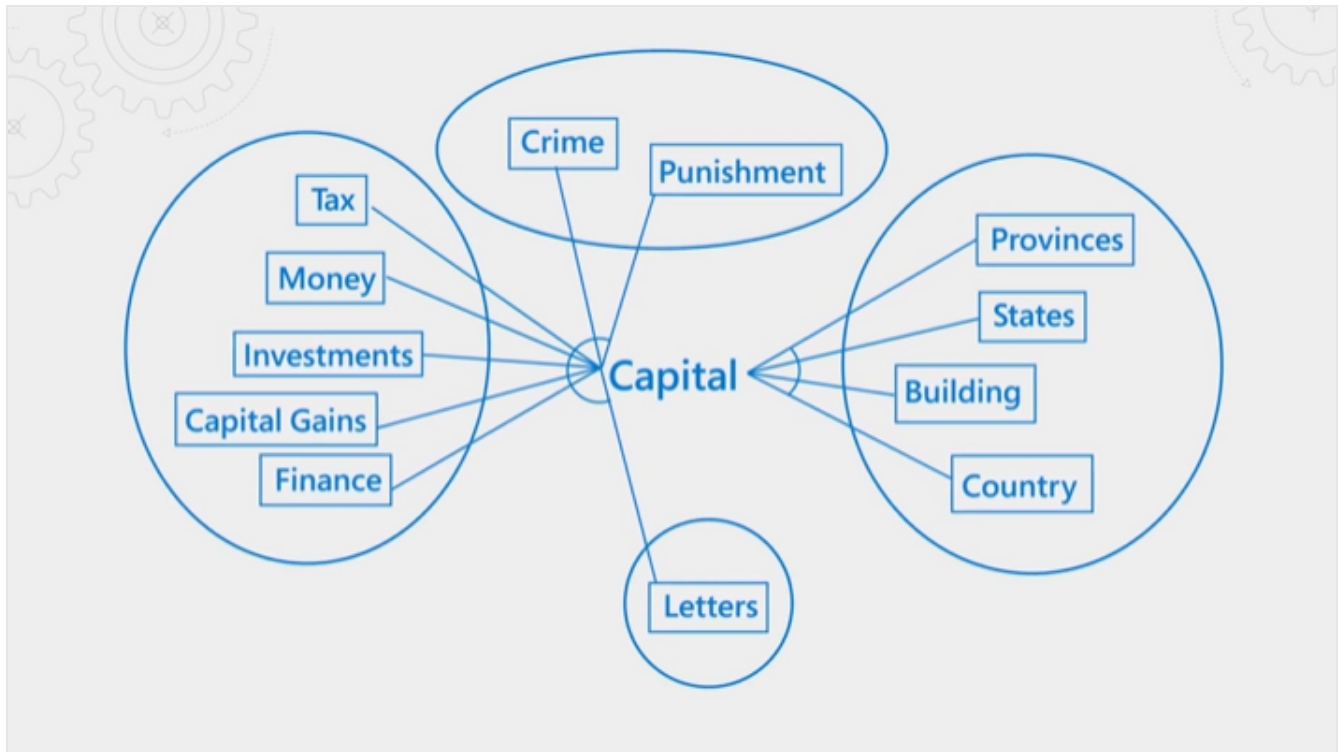
 Expand table

Feature	Description
Semantic ranking	Uses the context or semantic meaning of a query to compute a new relevance score over preranked results.
Semantic captions and highlights	Extracts verbatim sentences and phrases from a document that best summarize the content, with highlights over key passages for easy scanning. Captions that summarize a result are useful when individual content fields are too dense for the search results page. Highlighted text elevates the most relevant terms and phrases so that users can quickly determine why a match was considered relevant.
Semantic answers	An optional and extra substructure returned from a semantic query. It provides a direct answer to a query that looks like a question. It requires that a document has text with the characteristics of an answer.

How semantic ranker works

Semantic ranking feeds a query and results to language understanding models hosted by Microsoft and scans for better matches.

The following illustration explains the concept. Consider the term "capital". It has different meanings depending on whether the context is finance, law, geography, or grammar. Through language understanding, the semantic ranker can detect context and promote results that fit query intent.



Semantic ranking is both resource and time intensive. In order to complete processing within the expected latency of a query operation, inputs to the semantic ranker are consolidated and reduced so that the reranking step can be completed as quickly as possible.

There are two steps to semantic ranking: summarization and scoring. Outputs consist of rescored results, captions, and answers.


How inputs are collected and summarized

In semantic ranking, the query subsystem passes search results as an input to summarization and ranking models. Because the ranking models have input size constraints and are processing intensive, search results must be sized and structured (summarized) for efficient handling.

1. Semantic ranking starts with a [BM25-ranked result](#) from a text query or an [RRF-ranked result](#) from a hybrid query. Only text fields are used in the reranking exercise, and only the top 50 results progress to semantic ranking, even if results include more than 50. Typically, fields used in semantic ranking are informational

and descriptive.

- 2. For each document in the search result, the summarization model accepts up to 2,000 tokens, where a token is approximately 10 characters. Inputs are assembled from the "title", "keyword", and "content" fields listed in the [semantic configuration](#).
- 3. Excessively long strings are trimmed to ensure the overall length meets the input requirements of the summarization step. This trimming exercise is why it's important to add fields to your semantic configuration in priority order. If you have very large documents with text-heavy fields, anything after the maximum limit is ignored.

 Expand table

Semantic field	Token limit
"title"	128 tokens
"keywords"	128 tokens
"content"	remaining tokens

- 4. Summarization output is a summary string for each document, composed of the most relevant information from each field. Summary strings are sent to the ranker for scoring, and to machine reading comprehension models for captions and answers.

The maximum length of each generated summary string passed to the semantic ranker is 256 tokens.

Outputs of semantic ranker

From each summary string, the machine reading comprehension models find passages that are the most representative.

Outputs are:

- A [semantic caption](#) for the document. Each caption is available in a plain text version and a highlight version, and is frequently fewer than 200 words per document.
- An optional [semantic answer](#), assuming you specified the `answers` parameter, the query was posed as a question, and a passage is found in the long string that provides a likely answer to the question.

Captions and answers are always verbatim text from your index. There's no generative AI model in this workflow that creates or composes new content.

How summaries are scored

Scoring is done over the caption, and any other content from the summary string that fills out the 256 token length.

1. Captions are evaluated for conceptual and semantic relevance, relative to the query provided.
2. A **@search.rerankerScore** is assigned to each document based on the semantic relevance of the document for the given query. Scores range from 4 to 0 (high to low), where a higher score indicates higher relevance.
3. Matches are listed in descending order by score and included in the query response payload. The payload includes answers, plain text and highlighted captions, and any fields that you marked as retrievable or specified in a select clause.

ⓘ Note

For any given query, the distributions of **@search.rerankerScore** can exhibit slight variations due to conditions at the infrastructure level. Ranking model updates have also been known to affect the distribution. For these reasons, if you're writing custom code for minimum thresholds, or [setting the threshold property](#) for vector and hybrid queries, don't make the limits too granular.

Semantic capabilities and limitations

Semantic ranker is a newer technology so it's important to set expectations about what it can and can't do. What it *can* do:

- Promote matches that are semantically closer to the intent of original query.
- Find strings to use as captions and answers. Captions and answers are returned in the response and can be rendered on a search results page.

What semantic ranking *can't* do is rerun the query over the entire corpus to find semantically relevant results. Semantic ranking reranks the existing result set, consisting of the top 50 results as scored by the default ranking algorithm. Furthermore, semantic ranking can't create new information or strings. Captions and answers are extracted verbatim from your content so if the results don't include answer-like text, the language models won't produce one.

Although semantic ranking isn't beneficial in every scenario, certain content can benefit significantly from its capabilities. The language models in semantic ranking work best on searchable content that is information-rich and structured as prose. A knowledge base, online documentation, or documents that contain descriptive content see the most gains from semantic ranking capabilities.

The underlying technology is from Bing and Microsoft Research, and integrated into the Azure AI Search infrastructure as an add-on feature. For more information about the research and AI investments backing semantic ranking, see [How AI from Bing is powering Azure AI Search \(Microsoft Research Blog\)](#) .

The following video provides an overview of the capabilities.

https://www.youtube-nocookie.com/embed/yOf0WfVd_V0

Availability and pricing

Semantic ranker is available on search services at the Basic and higher tiers, subject to

[regional availability](#) .

When you enable semantic ranker, choose a pricing plan for the feature:

- At lower query volumes (under 1000 monthly), semantic ranking is free.
- At higher query volumes, choose the standard pricing plan.

The [Azure AI Search pricing page](#) shows you the billing rate for different currencies and intervals.

Charges for semantic ranking are levied when query requests include `queryType=semantic` and the search string isn't empty (for example, `search=pet friendly hotels in New York`). If your search string is empty (`search=*`), you aren't charged, even if the `queryType` is set to `semantic`.

See also

- [Enable semantic ranking](#)
- [Configure semantic ranking](#)
- [Blog: Outperforming vector search with hybrid retrieval and ranking capabilities](#)

Feedback

Was this page helpful?

 Yes

 No

[Provide product feedback](#) | [Get help at Microsoft Q&A](#)