Reranking

- How it works
- Configuration
  - Considerations

Reranking can help improve the quality of AI Search results by reordering retrieved documents based on semantic relevance to the user's query. It applies a secondary model after retrieval to rerank the top results before they are returned.

How it works

By default, reranking is **disabled** for all AI Search instances. You can enable it during creation or later from the settings page.

When enabled, AI Search will:

1. Retrieve a set of relevant results from your index, constrained by your `max_num_results` and `score_threshold` parameters.
2. Pass those results through a reranking model.
3. Return the reranked results, which the text generation model can use for answer generation.

Reranking helps improve accuracy, especially for large or noisy datasets where vector similarity alone may not produce the optimal ordering.

Workers AI reranking usage is included in AI Search usage. It does not appear in your AI Gateway logs or analytics and is not billed separately as Workers AI usage.

Configuration

When you make a `/search` or `/chat/completions` request using the Workers binding or REST API, you can enable or disable reranking per request and specify the reranking model.

```
const instance = env.AI_SEARCH.get("my-instance");

const results = await instance.search({
  messages: [{ role: "user", content: "What is Cloudflare?" }],
  ai_search_options: {
    reranking: {
      enabled: true,
      model: "@cf/baai/bge-reranker-base",
    },
  },
});
```

Considerations

Adding reranking will include an additional step to the query request. As a result, there may be an increase in the latency of the request.

□□□ #2
□ □□□ □□□ 2026-10-31 23:48:24 UTC
□ □□□ □□□ 2026-10-31 23:48:25 UTC