Reranking document passages

Use the reranker model and text reranker API that are available from watsonx.ai to rank a set of passages from most-to-least likely to answer a specified query.

Ways to develop

You can extract text from documents by using these programming methods:

- REST API
- Python

Overview

Ranking a set of passages in order of most-to-least relevant to a query is useful for search and retrieval tasks and chat workflows, especially in scenarios where grounded, factual answers are important.

Supported foundation models

To get a list of the reranker models that are available for use, you can use the *List the available foundation models* method in the watsonx.ai as a service API. Specify the **`filters=function_rerank`** parameter to return only the available reranker models.

For example:

```
curl -X GET \
  'https://{region}.ml.cloud.ibm.com/ml/v1/foundation_model_specs?version=2024-10-18&filters=function_rerank'
```

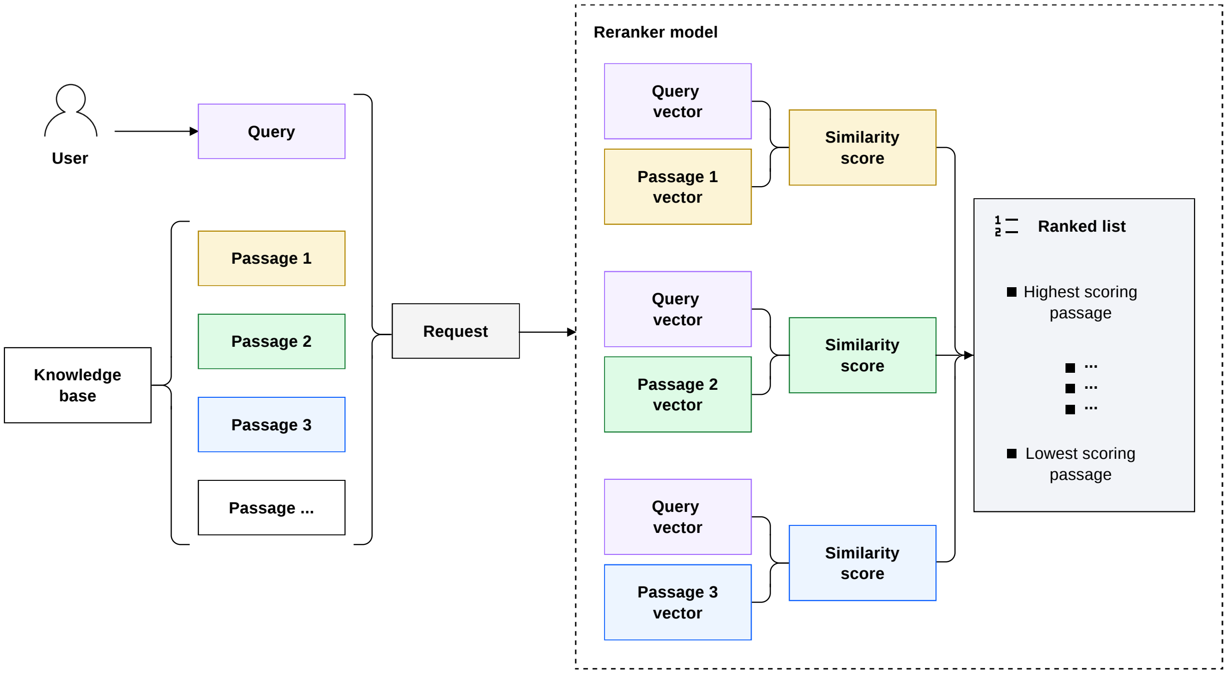
Note: Many of the embedding models also indicate that they support the rerank function. However, the embedding models support semantic reranking, which is a less accurate ranking method. When the embedding models rerank passages, they base the ranking on similarity scores that are derived from embedding vector values. Cross-encoder models are more effective at reranking because they explicitly compare each passage to the query and generate ranking scores per pairing.

Python

See the Reranker class in the watsonx.ai Python library.

REST API

As shown in the following diagram, the reranker method of the watsonx.ai API accepts a REST API request with a query and a list of document passages. The API submits these text strings to a reranker model. The reranker model pairs the query with each passage in turn, converts the text to embedding vectors, and then compares the vectors in each pair and scores their similarity to one another. The model then reranks the passages based on the generated similarity scores. Although the model converts text to text embeddings, the embeddings are not returned. The text embeddings are used to generate a similarity score, and only the score is returned.



API reference

For details, see the API reference documentation.

REST API example

The code in the example uses the `ms-marco-minilm-l-12-v2` model to rerank the following passages of text such that the passage that has information that is most likely to answer the query is listed first.

Query

What is an agent-driven AI workflow?

Passages

The input passages are definitions from the watsonx.ai glossary:

- Item 0. *A foundation model is a large-scale generative AI model that can be adapted to a wide range of downstream tasks.*
- Item 1. *Generative AI is a class of AI algorithms that can produce various types of content including text, source code, imagery, audio, and synthetic data.*
- Item 2. *Agentic AI is a generative AI flow that can decompose a prompt into multiple tasks, assign tasks to appropriate gen AI agents, and synthesize an answer without human intervention.*
- Item 3. *AI ethics is a multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes. Examples of AI ethics issues are data responsibility and privacy, fairness, explainability, robustness, transparency, environmental sustainability, inclusion, moral agency, value alignment, accountability, trust, and technology misuse.*
- Item 4. *AI governance is an organization's act of governing, through its corporate instructions, staff, processes and systems to direct, evaluate, monitor, and take corrective action throughout the AI lifecycle, to provide assurance that the AI system is operating as the organization intends, as its stakeholders expect, and as required by relevant regulation.*

REST API request example

In this example, only 5 passages are submitted for reranking. You can specify up to 1,000 inputs. However, the more passages you specify, the longer the reranking process takes because a cross-encoder model processes each passage together with the query one after the other.

Each input that you submit must conform to the maximum input token limit that is defined by the reranker model.

To address cases where a passage might have more tokens than the model allows, the **`truncate_input_tokens`** parameter is specified in this example to force the line to be truncated. Otherwise, the request might fail.

The **`inputs`** return options parameter is included in this example so that text from the original passage is included in the response, making it easier to understand how the original passages are reranked. In a real workflow, you might not need to include the **`inputs`** parameter.

You specify the reranker model that you want to use as the **`model_id`** in the payload for the text reranker method.

In the following example, replace the '{url}**`variable with the right value for your instance, such as`**us-south.ml.cloud.ibm.com` . Add your own bearer token and project ID.

```
curl -X POST \
  'https://{url}/ml/v1/text/rerank?version=2024-10-17' \
  --header 'Accept: application/json' \
  --header 'Content-Type: application/json' \
  --header 'Authorization: Bearer eyJraWQiOi...' \
  --data '{
      "inputs": [
        {
          "text": "A foundation model is a large-scale generative AI model that can be adapted 
to a wide range of downstream tasks."
        },
        {
          "text": "Generative AI is a class of AI algorithms that can produce various types of 
content including text, source code, imagery, audio, and synthetic data."
        },
        {
          "text": "Agentic AI is a generative AI flow that can decompose a prompt into multiple 
tasks, assign tasks to appropriate gen AI agents, and synthesize an answer without human 
intervention."
        },
        {
          "text": "AI ethics is a multidisciplinary field that studies how to optimize AI'\''s 
beneficial impact while reducing risks and adverse outcomes. Examples of AI ethics issues are 
data responsibility and privacy, fairness, explainability, robustness, transparency, 
environmental sustainability, inclusion, moral agency, value alignment, accountability, trust, 
and technology misuse."
        },
        {
          "text": "AI governance is an organization'\''s act of governing, through its corporate 
instructions, staff, processes and systems to direct, evaluate, monitor, and take corrective 
action throughout the AI lifecycle, to provide assurance that the AI system is operating as the 
organization intends, as its stakeholders expect, and as required by relevant regulation."
        }
      ],
      "query": "What is an agent-driven AI workflow?",
      "parameters":{
        "truncate_input_tokens": 512,
        "return_options":{
          "inputs":true
        }
      },
      "model_id": "cross-encoder/ms-marco-minilm-l-12-v2",
      "project_id": "51f3a990-4372-4ac3-9ddb-ed99d9b50840"
    }'
```

If you want to return only the highest-ranking passages, you can specify the **`top_n`** parameter. The following sample shows how to indicate that you want the model to return only one passage, the passage that is most related to the query:

```
"parameters":{
    "truncate_input_tokens": 512,
    "return_options":{
      "inputs":true,
      "top_n": 1
    }
}
```

REST API response example

The following response is returned for the example request. A few things to notice:

- The passages are reordered to show the passages with the highest scores first.

  The model accurately returns the definition for *Agentic AI* at the top of the list because it is most closely related to the question about agent-driven workflows.

- The index values, which range from 0–4, represent the original positions of the passages in the **`inputs`** array that was submitted in the request.

  For example, the passage about Agentic AI was the third passage in the list, and therefore has the index value of **`2`**.

- The **`input_token_count`** field shows the total number of tokens that were processed by the request.

```
{
  "model_id": "cross-encoder/ms-marco-minilm-l-12-v2",
  "created_at": "2024-10-17T14:01:47.322Z",
  "results": [
    {
      "index": 2,
      "score": 5.063366413116455,
      "input": {
        "text": "Agentic AI is a generative AI flow that can decompose a prompt into multiple tasks, assign tasks to appropriate gen AI agents, and synthesize an answer without human intervention."
      }
    },
    {
      "index": 4,
      "score": -0.992393970489502,
      "input": {
        "text": "AI governance is an organization's act of governing, through its corporate instructions, staff, processes and systems to direct, evaluate, monitor, and take corrective action throughout the AI lifecycle, to provide assurance that the AI system is operating as the organization intends, as its stakeholders expect, and as required by relevant regulation."
      }
    },
    {
      "index": 1,
      "score": -3.5372314453125,
      "input": {
        "text": "Generative AI is a class of AI algorithms that can produce various types of content including text, source code, imagery, audio, and synthetic data."
      }
    },
    {
      "index": 0,
      "score": -4.646212100982666,
      "input": {
        "text": "A foundation model is a large-scale generative AI model that can be adapted to a wide range of downstream tasks."
      }
    },
    {
      "index": 3,
      "score": -4.926990032196045,
      "input": {
        "text": "AI ethics is a multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes. Examples of AI ethics issues are data responsibility and privacy, fairness, explainability, robustness, transparency, environmental sustainability, inclusion, moral agency, value alignment, accountability, trust, and technology misuse."
      }
    }
  ],
  "input_token_count": 292,
  "system": {
    "warnings": [
      {
        "message": "This model is a Non-IBM Product governed by a third-party license that may
impose use restrictions and other obligations. By using this model you agree to its terms as
identified in the following URL.",
        "id": "disclaimer_warning",
        "more_info": "https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-
models.html?context=wx"
      }
    ]
  }
}
```

If you include `"top_n": 1` in the parameters object as shown in the earlier sample, the response looks as follows:

```
{
  "model_id": "cross-encoder/ms-marco-minilm-l-12-v2",
  "created_at": "2024-10-17T20:18:50.867Z",
  "results": [
    {
      "index": 2,
      "score": 5.063366413116455,
      "input": {
        "text": "Agentic AI is a generative AI flow that can decompose a prompt into multiple
tasks, assign tasks to appropriate gen AI agents, and synthesize an answer without human
intervention."
      }
    }
  ],
  "input_token_count": 292,
  "system": {
    "warnings": [
      {
        "message": "This model is a Non-IBM Product governed by a third-party license that may
impose use restrictions and other obligations. By using this model you agree to its terms as
identified in the following URL.",
        "id": "disclaimer_warning",
        "more_info": "https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-
models.html?context=wx"
      }
    ]
  }
}
```