Hugging Face Inference Model node documentation

# Hugging Face Inference Model node#

Use the Hugging Face Inference Model node to use Hugging Face's models.

On this page, you'll find the node parameters for the Hugging Face Inference Model node, and links to more resources.

Credentials

You can find authentication information for this node here.

Parameter resolution in sub-nodes

Sub-nodes behave differently to other nodes when processing multiple items using an expression.

Most nodes, including root nodes, take any number of items as input, process these items, and output the results. You can use expressions to refer to input items, and the node resolves the expression for each item in turn. For example, given an input of five name values, the expression {{ $json.name }} resolves to each name in turn.

name

{{ $json.name }}

In sub-nodes, the expression always resolves to the first item. For example, given an input of five name values, the expression {{ $json.name }} always resolves to the first name.

name

{{ $json.name }}

## Node parameters#

• Model: Select the model to use to generate the completion.

## Node options#

• Custom Inference Endpoint: Enter a custom inference endpoint URL.

• Frequency Penalty: Use this option to control the chances of the model repeating itself. Higher values reduce the chance of the model repeating itself.

• Maximum Number of Tokens: Enter the maximum number of tokens used, which sets the completion length.

• Presence Penalty: Use this option to control the chances of the model talking about new topics. Higher values increase the chance of the model talking about new topics.

• Sampling Temperature: Use this option to control the randomness of the sampling process. A higher temperature creates more diverse sampling, but increases the risk of hallucinations.

• Top K: Enter the number of token choices the model uses to generate the next token.

• Top P: Use this option to set the probability the completion should use. Use a lower value to ignore less probable options.

## Templates and examples#

Browse Hugging Face Inference Model integration templates, or search all templates

## Related resources#

Refer to LangChains's Hugging Face Inference Model documentation for more information about the service.

View n8n's Advanced AI documentation.

## AI glossary#

• completion: Completions are the responses generated by a model like GPT.

• hallucinations: Hallucination in AI is when an LLM (large language model) mistakenly perceives patterns or objects that don't exist.

• vector database: A vector database stores mathematical representations of information. Use with embeddings and retrievers to create a database that your AI can access when answering questions.

• vector store: A vector store, or vector database, stores mathematical representations of information. Use with embeddings and retrievers to create a database that your AI can access when answering questions.