OpenRouter Chat Model node documentation

# OpenRouter Chat Model node#

Use the OpenRouter Chat Model node to use OpenRouter's chat models with conversational agents.

On this page, you'll find the node parameters for the OpenRouter Chat 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.

n8n dynamically loads models from OpenRouter and you'll only see the models available to your account.

## Node options#

Use these options to further refine the node's behavior.

### Base URL#

Enter a URL here to override the default URL for the API.

### 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.

### Response Format#

Choose Text or JSON. JSON ensures the model returns valid JSON.

### 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.

### Timeout#

Enter the maximum request time in milliseconds.

### Max Retries#

Enter the maximum number of times to retry a request.

### 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#

by Davide

by Davide

by Davide

## Related resources#

As OpenRouter is API-compatible with OpenAI, you can refer to LangChains's OpenAI 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.