Ollama Model node documentation

# Ollama Model node#

The Ollama Model node allows you use local Llama 2 models.

On this page, you'll find the node parameters for the Ollama 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 that generates the completion. Choose from:  
Llama2  
Llama2 13B  
Llama2 70B  
Llama2 Uncensored

• Llama2

• Llama2 13B

• Llama2 70B

• Llama2 Uncensored

Refer to the Ollama Models Library documentation for more information about available models.

## Node options#

• 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#

by Mihai Farcas

by Joseph LePage

by Joseph LePage

## Related resources#

Refer to LangChains's Ollama documentation for more information about the service.

View n8n's Advanced AI documentation.

## Common issues#

For common questions or issues and suggested solutions, refer to Common issues.

## 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.

## Self-hosted AI Starter Kit#

New to working with AI and using self-hosted n8n? Try n8n's self-hosted AI Starter Kit to get started with a proof-of-concept or demo playground using Ollama, Qdrant, and PostgreSQL.