**<https://github.com/ollama/ollama/blob/main/docs/modelfile.md>**

**Date: 20/5/2024 Version: 0.1.38**

**Ollama Model File**

Note: Modelfile syntax is in development

A model file is the blueprint to create and share models with Ollama.

**Table of Contents**

* [Format](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#format)
* [Examples](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#examples)
* [Instructions](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#instructions)
  + [FROM (Required)](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#from-required)
    - [Build from llama3](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#build-from-llama3)
    - [Build from a bin file](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#build-from-a-bin-file)
  + [PARAMETER](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#parameter)
    - [Valid Parameters and Values](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#valid-parameters-and-values)
  + [TEMPLATE](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#template)
    - [Template Variables](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#template-variables)
  + [SYSTEM](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#system)
  + [ADAPTER](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#adapter)
  + [LICENSE](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#license)
  + [MESSAGE](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#message)
* [Notes](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#notes)

**Format**

The format of the Modelfile:

# comment

INSTRUCTION arguments

| **Instruction** | **Description** |
| --- | --- |
| [FROM](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#from-required) (required) | Defines the base model to use. |
| [PARAMETER](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#parameter) | Sets the parameters for how Ollama will run the model. |
| [TEMPLATE](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#template) | The full prompt template to be sent to the model. |
| [SYSTEM](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#system) | Specifies the system message that will be set in the template. |
| [ADAPTER](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#adapter) | Defines the (Q)LoRA adapters to apply to the model. |
| [LICENSE](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#license) | Specifies the legal license. |
| [MESSAGE](https://github.com/ollama/ollama/blob/main/docs/modelfile.md#message) | Specify message history. |

**Examples**

**Basic Modelfile**

An example of a Modelfile creating a mario blueprint:

FROM llama3

# sets the temperature to 1 [higher is more creative, lower is more coherent]

PARAMETER temperature 1

# sets the context window size to 4096, this controls how many tokens the LLM can use as context to generate the next token

PARAMETER num\_ctx 4096

# sets a custom system message to specify the behavior of the chat assistant

SYSTEM You are Mario from super mario bros, acting as an assistant.

To use this:

1. Save it as a file (e.g. Modelfile)
2. ollama create choose-a-model-name -f <location of the file e.g. ./Modelfile>'
3. ollama run choose-a-model-name
4. Start using the model!

More examples are available in the [examples directory](https://github.com/ollama/ollama/blob/main/examples).

To view the Modelfile of a given model, use the ollama show --modelfile command.

> ollama show --modelfile llama3

# Modelfile generated by "ollama show"

# To build a new Modelfile based on this one, replace the FROM line with:

# FROM llama3:latest

FROM /Users/pdevine/.ollama/models/blobs/sha256-00e1317cbf74d901080d7100f57580ba8dd8de57203072dc6f668324ba545f29

TEMPLATE """{{ if .System }}<|start\_header\_id|>system<|end\_header\_id|>

{{ .System }}<|eot\_id|>{{ end }}{{ if .Prompt }}<|start\_header\_id|>user<|end\_header\_id|>

{{ .Prompt }}<|eot\_id|>{{ end }}<|start\_header\_id|>assistant<|end\_header\_id|>

{{ .Response }}<|eot\_id|>"""

PARAMETER stop "<|start\_header\_id|>"

PARAMETER stop "<|end\_header\_id|>"

PARAMETER stop "<|eot\_id|>"

PARAMETER stop "<|reserved\_special\_token"

**Instructions**

**FROM (Required)**

The FROM instruction defines the base model to use when creating a model.

FROM <model name>:<tag>

**Build from llama3**

FROM llama3

A list of available base models: <https://github.com/ollama/ollama#model-library>

**Build from a bin file**

FROM ./ollama-model.bin

This bin file location should be specified as an absolute path or relative to the Modelfile location.

**PARAMETER**

The PARAMETER instruction defines a parameter that can be set when the model is run.

PARAMETER <parameter> <parametervalue>

**Valid Parameters and Values**

| **Parameter** | **Description** | **Value Type** | **Example Usage** |
| --- | --- | --- | --- |
| mirostat | Enable Mirostat sampling for controlling perplexity. (default: 0, 0 = disabled, 1 = Mirostat, 2 = Mirostat 2.0) | int | mirostat 0 |
| mirostat\_eta | Influences how quickly the algorithm responds to feedback from the generated text. A lower learning rate will result in slower adjustments, while a higher learning rate will make the algorithm more responsive. (Default: 0.1) | float | mirostat\_eta 0.1 |
| mirostat\_tau | Controls the balance between coherence and diversity of the output. A lower value will result in more focused and coherent text. (Default: 5.0) | float | mirostat\_tau 5.0 |
| num\_ctx | Sets the size of the context window used to generate the next token. (Default: 2048) | int | num\_ctx 4096 |
| repeat\_last\_n | Sets how far back for the model to look back to prevent repetition. (Default: 64, 0 = disabled, -1 = num\_ctx) | int | repeat\_last\_n 64 |
| repeat\_penalty | Sets how strongly to penalize repetitions. A higher value (e.g., 1.5) will penalize repetitions more strongly, while a lower value (e.g., 0.9) will be more lenient. (Default: 1.1) | float | repeat\_penalty 1.1 |
| temperature | The temperature of the model. Increasing the temperature will make the model answer more creatively. (Default: 0.8) | float | temperature 0.7 |
| seed | Sets the random number seed to use for generation. Setting this to a specific number will make the model generate the same text for the same prompt. (Default: 0) | int | seed 42 |
| stop | Sets the stop sequences to use. When this pattern is encountered the LLM will stop generating text and return. Multiple stop patterns may be set by specifying multiple separate stop parameters in a modelfile. | string | stop "AI assistant:" |
| tfs\_z | Tail free sampling is used to reduce the impact of less probable tokens from the output. A higher value (e.g., 2.0) will reduce the impact more, while a value of 1.0 disables this setting. (default: 1) | float | tfs\_z 1 |
| num\_predict | Maximum number of tokens to predict when generating text. (Default: 128, -1 = infinite generation, -2 = fill context) | int | num\_predict 42 |
| top\_k | Reduces the probability of generating nonsense. A higher value (e.g. 100) will give more diverse answers, while a lower value (e.g. 10) will be more conservative. (Default: 40) | int | top\_k 40 |
| top\_p | Works together with top-k. A higher value (e.g., 0.95) will lead to more diverse text, while a lower value (e.g., 0.5) will generate more focused and conservative text. (Default: 0.9) | float | top\_p 0.9 |

**TEMPLATE**

TEMPLATE of the full prompt template to be passed into the model. It may include (optionally) a system message, a user's message and the response from the model. Note: syntax may be model specific. Templates use Go [template syntax](https://pkg.go.dev/text/template).

**Template Variables**

| **Variable** | **Description** |
| --- | --- |
| {{ .System }} | The system message used to specify custom behavior. |
| {{ .Prompt }} | The user prompt message. |
| {{ .Response }} | The response from the model. When generating a response, text after this variable is omitted. |

TEMPLATE """{{ if .System }}<|im\_start|>system

{{ .System }}<|im\_end|>

{{ end }}{{ if .Prompt }}<|im\_start|>user

{{ .Prompt }}<|im\_end|>

{{ end }}<|im\_start|>assistant

"""

**SYSTEM**

The SYSTEM instruction specifies the system message to be used in the template, if applicable.

SYSTEM """<system message>"""

**ADAPTER**

The ADAPTER instruction is an optional instruction that specifies any LoRA adapter that should apply to the base model. The value of this instruction should be an absolute path or a path relative to the Modelfile and the file must be in a GGML file format. The adapter should be tuned from the base model otherwise the behaviour is undefined.

ADAPTER ./ollama-lora.bin

**LICENSE**

The LICENSE instruction allows you to specify the legal license under which the model used with this Modelfile is shared or distributed.

LICENSE """

<license text>

"""

**MESSAGE**

The MESSAGE instruction allows you to specify a message history for the model to use when responding. Use multiple iterations of the MESSAGE command to build up a conversation which will guide the model to answer in a similar way.

MESSAGE <role> <message>

**Valid roles**

| **Role** | **Description** |
| --- | --- |
| system | Alternate way of providing the SYSTEM message for the model. |
| user | An example message of what the user could have asked. |
| assistant | An example message of how the model should respond. |

**Example conversation**

MESSAGE user Is Toronto in Canada?

MESSAGE assistant yes

MESSAGE user Is Sacramento in Canada?

MESSAGE assistant no

MESSAGE user Is Ontario in Canada?

MESSAGE assistant yes

**Notes**

* the **Modelfile is not case sensitive**. In the examples, uppercase instructions are used to make it easier to distinguish it from arguments.
* Instructions can be in any order. In the examples, the FROM instruction is first to keep it easily readable.