Chat Completion API

Tap into Galadriel’s LLM inference network. Follows the exact schema as OpenAI’s chat completion API.

POST

/ v1 / chat / completions Send Authorization Authorization string * Bearer

Authorization Required string Bearer authentication header.

Bearer Galadriel-API-key

Get API key fromGaladriel dashboard . Body object * Required object Add Example Value messages array * messages Required array A list of messages comprising the conversation so far. Add Example Value model string * model Required string ID of the model to use. Get ID for availablemodels . Add Example Value frequency_penalty number

number null frequency_penalty number Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim. logit_bias object

object null logit_bias object Modify the likelihood of specified tokens appearing in the completion. logprobs boolean

boolean null Select option logprobs boolean Whether to return log probabilities of the output tokens or not. If true, returns the log probabilities of each output token returned in thecontent ofmessage . top_logprobs integer

integer null top_logprobs integer An integer between 0 and 20 specifying the number of most likely tokens to return at each token position, each with an associated log probability.logprobs must be set totrue if this parameter is used. max_tokens integer

integer null max_tokens integer The maximum number of tokens to generate in the chat completion. n integer

integer null n integer How many chat completion choices to generate for each input message. presence_penalty number

number array null presence_penalty number Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics. response_format object

object null response_format object An object specifying the format that the model must output. Compatible with GPT-4o, GPT-4o mini, GPT-4 Turbo and all GPT-3.5 Turbo models newer than gpt-3.5-turbo-1106. type enum * Select option type Required enum The type of response format being defined:text json_schema object

object null json_schema object description string

string null description string A description of what the response format is for, used by the model to determine how to respond in the format. name string name string The name of the response format. Must be a-z, A-Z, 0-9, or contain underscores and dashes, with a maximum length of 64. schema object

object null schema object The schema for the response format, described as a JSON Schema object. strict boolean

boolean null Select option strict boolean Whether to enable strict schema adherence when generating the output. If set to true, the model will always follow the exact schema defined in the schema field. Only a subset of JSON Schema is supported whenstrict istrue . seed integer

integer null seed integer This feature is in Beta. If specified, our system will make a best effort to sample deterministically, such that repeated requests with the sameseed and parameters should return the same result. Determinism is not guaranteed, and you should refer to thesystem_fingerprint response parameter to monitor changes in the backend. stop string

string array null stop string Up to 4 sequences where the API will stop generating further tokens. stream boolean

boolean null Select option stream boolean If set, partial message deltas will be sent, like in ChatGPT. stream_options object

object null stream_options object Options for streaming response. Only set this when you setstream: true . include_usage boolean * Select option include_usage Required boolean If set, an additional chunk will be streamed before thedata: [DONE] message. The usage field on this chunk shows the token usage statistics for the entire request, and the choices field will always be an empty array. All other chunks will also include a usage field, but with a null value. temperature number

number null temperature number What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic. top_p number

number null top_p number An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered. tools array

array null tools array Currently the 8b model does not support tools. A list of tools the model may call. Currently, only functions are supported as a tool. Use this to provide a list of functions the model may generate JSON inputs for. tool_choice string

string object null tool_choice string Controls which (if any) tool is called by the model. none means the model will not call any tool and instead generates a message. auto means the model can pick between generating a message or calling one or more tools. required means the model must call one or more tools. Specifying a particular tool via{"type": "function", "function": {"name": "my_function"}} forces the model to call that tool.

Authorizations

Authorization string header required Bearer authentication header.

Bearer Galadriel-API-key

Get API key from[Galadriel dashboard](#) .

Body

application/json messages object[] required A list of messages comprising the conversation so far. Showchild attributes model string required ID of the model to use. Get ID for available[models](#) . frequency_penalty number | null Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim. logit_bias object | null Modify the likelihood of specified tokens appearing in the completion. logprobs boolean | null default:false Whether to return log probabilities of the output tokens or not. If true, returns the log probabilities of each output token returned in thecontent ofmessage . top_logprobs integer | null An integer between 0 and 20 specifying the number of most likely tokens to return at each token position, each with an associated log probability.logprobs must be set totrue if this parameter is used. max_tokens integer | null The maximum number of tokens to generate in the chat completion. n integer | null How many chat completion choices to generate for each input message. presence_penalty number | null any[] | null default:0 Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics. response_format object | null An object specifying the format that the model must output. Compatible with GPT-4o, GPT-4o mini, GPT-4 Turbo and all GPT-3.5 Turbo models newer than gpt-3.5-turbo-1106. Showchild attributes seed integer | null This feature is in Beta. If specified, our system will make a best effort to sample deterministically, such that repeated requests with the sameseed and parameters should return the same result. Determinism is not guaranteed, and you should refer to thesystem_fingerprint response parameter to monitor changes in the backend. stop string | null any[] | null Up to 4 sequences where the API will stop generating further tokens. stream boolean | null default:false If set, partial message deltas will be sent, like in ChatGPT. stream_options object | null Options for streaming response. Only set this when you setstream: true . Showchild attributes temperature number | null default:1 What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic. top_p number | null default:1 An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered. tools object[] | null Currently the 8b model does not support tools. A list of tools the model may call. Currently, only functions are supported as a tool. Use this to provide a list of functions the model may generate JSON inputs for. Showchild attributes tool_choice string | null object | null Controls which (if any) tool is called by the model. none means the model will not call any tool and instead generates a message. auto means the model can pick between generating a message or calling one or more tools. required means the model must call one or more tools. Specifying a particular tool via{"type": "function", "function": {"name": "my_function"}} forces the model to call that tool.

Response

200 - application/json id string required choices object[] required Showchild attributes created integer required model string required object enum required Available options:chat.completion service_tier enum | null Available options:scale , default system_fingerprint string | null usage object | null Showchild attributes