Documentation

Reasoning

Reasoning

Reasoning models excel at complex problem-solving tasks that require step-by-step analysis, logical deduction, and structured thinking and solution validation. With Groq inference speed, these types of models can deliver instant reasoning capabilities critical for real-time applications.

Why Speed Matters for Reasoning

Reasoning models are capable of complex decision making with explicit reasoning chains that are part of the token output and used for decision-making, which make low-latency and fast inference essential. Complex problems often require multiple chains of reasoning tokens where each step build on previous results. Low latency compounds benefits across reasoning chains and shaves off minutes of reasoning to a response in seconds.

Supported Model

MODEL ID	MODEL
`deepseek-r1-distill-llama-70b`	DeepSeek R1 (Distil-Llama 70B)

Reasoning Format

Groq API supports explicit reasoning formats through the `reasoning_format` parameter, giving you fine-grained control over how the model's reasoning process is presented. This is particularly valuable for valid JSON outputs, debugging, and understanding the model's decision-making process.
Note: The format defaults to `raw` or `parsed` when JSON mode or tool use are enabled as those modes do not support `raw`. If reasoning is explicitly set to `raw` with JSON mode or tool use enabled, we will return a 400 error.

Options for Reasoning Format

REASONING_FORMAT OPTIONS	DESCRIPTION
`parsed`	Separates reasoning into a dedicated field while keeping the response concise.
`raw`	Includes reasoning within think tags in the content.

REASONING_FORMAT OPTIONS	DESCRIPTION
hidden	Returns only the final answer for maximum efficiency.

Quick Start

Python JavaScript curl

```python
from groq import Groq

client = Groq()
completion = client.chat.completions.create(
    model="deepseek-r1-distill-llama-70b",
    messages=[
        {
            "role": "user",
            "content": "How many r's are in the word strawberry?"
        }
    ],
    temperature=0.6,
    max_completion_tokens=1024,
    top_p=0.95,
    stream=True,
    reasoning_format="raw"
)

for chunk in completion:
    print(chunk.choices[0].delta.content or "", end="")
```

Quick Start with Tool use

```curl
curl https://api.groq.com//openai/v1/chat/completions -s \
  -H "authorization: bearer $GROQ_API_KEY" \
  -d '{
    "model": "deepseek-r1-distill-llama-70b",
    "messages": [
        {
            "role": "user",
            "content": "What is the weather like in Paris today?"
        }
    ],
    "tools": [
        {
```

```
            "type": "function",
            "function": {
                "name": "get_weather",
                "description": "Get current temperature for a given location.",
                "parameters": {
                    "type": "object",
                    "properties": {
                        "location": {
                            "type": "string",
                            "description": "City and country e.g. Bogotá, Colombia"
                        }
                    },
                    "required": [
                        "location"
                    ],
                    "additionalProperties": false
                },
                "strict": true
            }
        }
    ]}'
```

Recommended Configuration Parameters

PARAMETER	DEFAULT	RANGE	DESCRIPTION
messages			Array of message objects. Important: Avoid system prompts - include all instructions in the user message!
temperature	0.6	0.0 - 2.0	Controls randomness in responses. Lower values make responses more deterministic. Recommended range: 0.5-0.7 to prevent repetitions or incoherent outputs
max_completion_tokens	1024	-	Maximum length of model's response. Default may be too low for complex reasoning - consider increasing for detailed step-by-step solutions
top_p	0.95	0.0 - 1.0	Controls diversity of token selection
stream	false	boolean	Enables response streaming. Recommended for interactive reasoning tasks
stop	null	string/array	Custom stop sequences
seed	null	integer	Set for reproducible results. Important for benchmarking - run multiple tests with different seeds
json_mode	-	boolean	Set to enable JSON mode for structured output.
reasoning_format	raw	"parsed", "raw", "hidden"	Controls how model reasoning is presented in the response. Must be set to either `parsed` or `hidden` when using tool calling or JSON mode.

Optimizing Performance

Temperature and Token Management

The model performs best with temperature settings between 0.5-0.7, with lower values (closer to 0.5) producing more consistent mathematical proofs and higher values allowing for more creative problem-solving approaches. Monitor and adjust your token usage based on the complexity of your reasoning tasks - while the default max_completion_tokens is 1024, complex proofs may require higher limits.

Prompt Engineering

To ensure accurate, step-by-step reasoning while maintaining high performance:

- DeepSeek-R1 works best when all instructions are included directly in user messages rather than system prompts.
- Structure your prompts to request explicit validation steps and intermediate calculations.
- Avoid few-shot prompting and go for zero-shot prompting only.