API Guides Reasoning Model (deepseek-reasoner)

Reasoning Model (`deepseek-reasoner`)

`deepseek-reasoner` is a reasoning model developed by DeepSeek. Before delivering the final answer, the model first generates a Chain of Thought (CoT) to enhance the accuracy of its responses. Our API provides users with access to the CoT content generated by `deepseek-reasoner`, enabling them to view, display, and distill it.

When using `deepseek-reasoner`, please upgrade the OpenAI SDK first to support the new parameters.

```
pip3 install -U openai
```

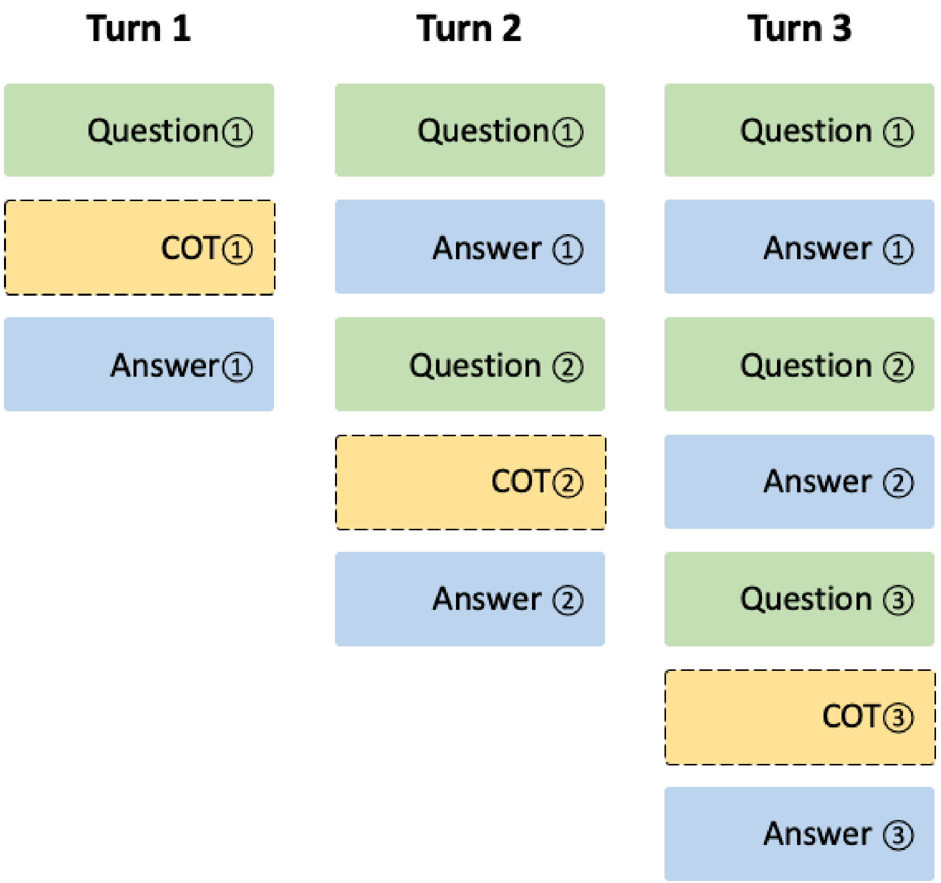
API Parameters

- **Input**：
  - `max_tokens`： The maximum length of the final response after the CoT output is completed, defaulting to 4K, with a maximum of 8K. Note that the CoT output can reach up to 32K tokens, and the parameter to control the CoT length (`reasoning_effort`) will be available soon.
- **Output**：
  - `reasoning_content`： The content of the CoT， which is at the same level as `content` in the output structure. See API Example for details
  - `content`The content of the final answer
- **Context Length**： The API supports a maximum context length of 64K, and the length of the output `reasoning_content` is not counted within the 64K context length.
- **Supported Features**： Chat Completion、Chat Prefix Completion (Beta)
- **Not Supported Features**： Function Call、Json Output、FIM (Beta)

- **Not Supported Parameters**：`temperature`、`top_p`、`presence_penalty`、`frequency_penalty`、`logprobs`、`top_logprobs`. Please note that to ensure compatibility with existing software, setting `temperature`、`top_p`、`presence_penalty`、`frequency_penalty` will not trigger an error but will also have no effect. Setting `logprobs`、`top_logprobs` will trigger an error.

Multi-round Conversation

In each round of the conversation, the model outputs the CoT (`reasoning_content`) and the final answer (`content`). In the next round of the conversation, the CoT from previous rounds is not concatenated into the context, as illustrated in the following diagram:



Please note that if the `reasoning_content` field is included in the sequence of input messages, the API will return a `400` error. Therefore, you should remove the `reasoning_content` field from the API response before making the API request, as demonstrated in the API example.

API Example

The following code, using Python as an example, demonstrates how to access the CoT and the final answer, as well as how to conduct multi-round conversations:

NoStreaming Streaming

```python
from openai import OpenAI
client = OpenAI(api_key="<DeepSeek API Key>", base_url="https://api.deepseek.com")

# Round 1
messages = [{"role": "user", "content": "9.11 and 9.8, which is greater?"}]
response = client.chat.completions.create(
    model="deepseek-reasoner",
    messages=messages
)

reasoning_content = response.choices[0].message.reasoning_content
content = response.choices[0].message.content

# Round 2
messages.append({'role': 'assistant', 'content': content})
messages.append({'role': 'user', 'content': "How many Rs are there in the word 'strawberry'?"})
response = client.chat.completions.create(
    model="deepseek-reasoner",
    messages=messages
)
# ...
```