Token counting

Copy page

Token counting enables you to determine the number of tokens in a message before sending it to Claude, helping you make informed decisions about your prompts and usage. With token counting, you can

- Proactively manage rate limits and costs
- Make smart model routing decisions
- Optimize prompts to be a specific length

How to count message tokens

The token counting endpoint accepts the same structured list of inputs for creating a message, including support for system prompts, tools, images, and PDFs. The response contains the total number of input tokens.

> The token count should be considered an **estimate**. In some cases, the actual number of input tokens used when creating a message may differ by a small amount.
>
> Token counts may include tokens added automatically by Anthropic for system optimizations. **You are not billed for system-added tokens**. Billing reflects only your content.

Supported models

All active models support token counting.

Count tokens in basic messages

Claude Docs

```python
import anthropic

client = anthropic.Anthropic()

response = client.messages.count_tokens(
    model="claude-sonnet-4-5",
    system="You are a scientist",
    messages=[{
        "role": "user",
        "content": "Hello, Claude"
    }],
)

print(response.json())
```

JSON

```
{ "input_tokens": 14 }
```

Count tokens in messages with tools

Server tool token counts only apply to the first sampling call.

Python

Claude Docs

```python
client = anthropic.Anthropic()

response = client.messages.count_tokens(
    model="claude-sonnet-4-5",
    tools=[
        {
            "name": "get_weather",
            "description": "Get the current weather in a given location",
            "input_schema": {
                "type": "object",
                "properties": {
                    "location": {
                        "type": "string",
                        "description": "The city and state, e.g. San Francisco
                    }
                },
                "required": ["location"],
            },
        }
    ],
    messages=[{"role": "user", "content": "What's the weather like in San Fran
)

print(response.json())
```

JSON

```
{ "input_tokens": 403 }
```

Count tokens in messages with images

Shell

Claude Docs

```
IMAGE_URL="https://upload.wikimedia.org/wikipedia/commons/a/a7/Camponotus_flav
IMAGE_MEDIA_TYPE="image/jpeg"
IMAGE_BASE64=$(curl "$IMAGE_URL" | base64)

curl https://api.anthropic.com/v1/messages/count_tokens \
     --header "x-api-key: $ANTHROPIC_API_KEY" \
     --header "anthropic-version: 2023-06-01" \
     --header "content-type: application/json" \
     --data \
'{
    "model": "claude-sonnet-4-5",
    "messages": [
        {"role": "user", "content": [
            {"type": "image", "source": {
                "type": "base64",
                "media_type": "'$IMAGE_MEDIA_TYPE'",
                "data": "'$IMAGE_BASE64'"
            }},
            {"type": "text", "text": "Describe this image"}
        ]}
    ]
}'
```

JSON

```
{ "input_tokens": 1551 }
```

Count tokens in messages with extended thinking

See here for more details about how the context window is calculated with extended thinking

- Thinking blocks from **previous** assistant turns are ignored and **do not** count toward your input tokens
- **Current** assistant turn thinking **does** count toward your input tokens

Shell

Claude Docs

```
    --header "content-type: application/json" \
    --header "anthropic-version: 2023-06-01" \
    --data '{
      "model": "claude-sonnet-4-5",
      "thinking": {
        "type": "enabled",
        "budget_tokens": 16000
      },
      "messages": [
        {
          "role": "user",
          "content": "Are there an infinite number of prime numbers such that
        },
        {
          "role": "assistant",
          "content": [
            {
              "type": "thinking",
              "thinking": "This is a nice number theory question. Lets think a
              "signature": "EuYBCkQYAiJAgCs1le6/Pol5Z4/JMomVOouGrWdhYNsH3ukzUE
            },
            {
              "type": "text",
              "text": "Yes, there are infinitely many prime numbers p such tha
            }
          ]
        },
        {
          "role": "user",
          "content": "Can you write a formal proof?"
        }
      ]
    }'
```

JSON

```
{ "input_tokens": 88 }
```

Count tokens in messages with PDFs

Claude Docs

Shell

```shell
curl https://api.anthropic.com/v1/messages/count_tokens \
    --header "x-api-key: $ANTHROPIC_API_KEY" \
    --header "content-type: application/json" \
    --header "anthropic-version: 2023-06-01" \
    --data '{
      "model": "claude-sonnet-4-5",
      "messages": [{
        "role": "user",
        "content": [
          {
            "type": "document",
            "source": {
              "type": "base64",
              "media_type": "application/pdf",
              "data": "'$(base64 -i document.pdf)'"
            }
          },
          {
            "type": "text",
            "text": "Please summarize this document."
          }
        ]
      }]
    }'
```

JSON

```json
{ "input_tokens": 2188 }
```

Pricing and rate limits

Token counting is **free to use** but subject to requests per minute rate limits based on your usage tier. If you need higher limits, contact sales through the Claude Console.

Claude Docs

2	2,000
3	4,000
4	8,000

Token counting and message creation have separate and independent rate limits -- usage of one does not count against the limits of the other.

FAQ

Does token counting use prompt caching?

Claude Docs

Claude Docs

Company

Anthropic

Careers

Economic Futures

Research

News

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support

Discord

Terms and policies

Privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy