Benchmarking Reasoning Models

This document compares three reasoning models from different providers:

Model Providers

OpenAI GPT

We selected `o3-mini-2025-01-31` as it is currently OpenAI's most cost-effective reasoning model. For detailed pricing information, please visit OpenAI Pricing. ![[Pasted image 20250301161216.png]]

The model accepts a `reasoning_effort` parameter with three possible values:

- `low`
- `medium`
- `high`

Note: While reasoning tokens are not visible via the API, they still occupy space in the model's context window and are billed as output tokens.

Anthropic Claude

For this comparison, we're using `claude-3-7-sonnet-20250219`, Anthropic's dedicated reasoning model. While it's the most expensive model in our comparison, it offers unique capabilities that justify the cost.

Token Configuration

By default, the model has a maximum token limit of 1024.

![[Pasted image 20250301162719.png]]

To prevent incomplete responses, we increased this limit to 4096.

Operation Modes

The `claude-3-7-sonnet-20250219` model offers two distinct operation modes:

1. **Standard Mode**
    - Functions similarly to previous Claude models
    - Provides direct responses without exposing internal reasoning
2. **Extended Thinking Mode**
    - Reveals Claude's reasoning process before delivering the final answer

API Example:

```
import Anthropic from "@anthropic-ai/sdk";
```

```
const client = new Anthropic();

const response = await client.messages.create({
  model: "claude-3-7-sonnet-20250219",
  max_tokens: 20000,
  thinking: {
    type: "enabled",
    budget_tokens: 16000,
  },
  messages: [
    {
      role: "user",
      content: "Explain quantum entanglement",
    },
  ],
});
```

Token Parameters

When using extended thinking mode, configure these key parameters:

budget_tokens

- Defines the maximum tokens allowed for internal reasoning processes
- Must be smaller than `max_tokens`
- Larger budgets enable more thorough analysis of complex problems
- The model may not consume the entire budget, especially above 32K tokens

max_tokens

- Defines the maximum total tokens for the entire response
- Includes `budget_tokens` as a subset
- For Claude 3.7 Sonnet: a validation error occurs if prompt tokens + `max_tokens` exceeds the context window

The extended thinking mode provides visibility into the reasoning process. For detailed examples, please refer to the attached Google Colab notebook.

DeepSeek R1

We're using `deepseek-reasoner` (also known as DeepSeek R1), which offers the most competitive pricing in our comparison.

DeepSeek provides full OpenAI API compatibility. Their documentation recommends using the OpenAI SDK, which enables seamless integration with LangChain-OpenAI as well.

Similar to Anthropic Claude, DeepSeek R1 exposes its reasoning process in the response body. Access both the reasoning and final content through the API:

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

Note: The reasoning process tokens are included in the total output token count for billing purposes.


```
{
    "completion_tokens": 1312,
    "prompt_tokens": 21,
    "total_tokens": 1333,
    "completion_tokens_details": {
        "accepted_prediction_tokens": None,
        "audio_tokens": None,
        "reasoning_tokens": 334,
        "rejected_prediction_tokens": None
    },
    "prompt_tokens_details": {
        "audio_tokens": None,
        "cached_tokens": 0
    },
    "prompt_cache_hit_tokens": 0,
    "prompt_cache_miss_tokens": 21
}
```


Key Findings

- All reasoning and thinking processes are billed as output tokens
- Only Claude and DeepSeek expose their reasoning processes in the response
- OpenAI processes reasoning internally within their system

Pricing Comparison (per million tokens)

Model	Price	Notes
Anthropic Claude	$15.00	Most expensive option
OpenAI o3-mini	$4.40	
OpenAI o1	$60.00	Premium pricing
DeepSeek R1	$2.19	Most cost-effective
DeepSeek R1 (discount)	$0.55	75% off (UTC 16:30-00:30)