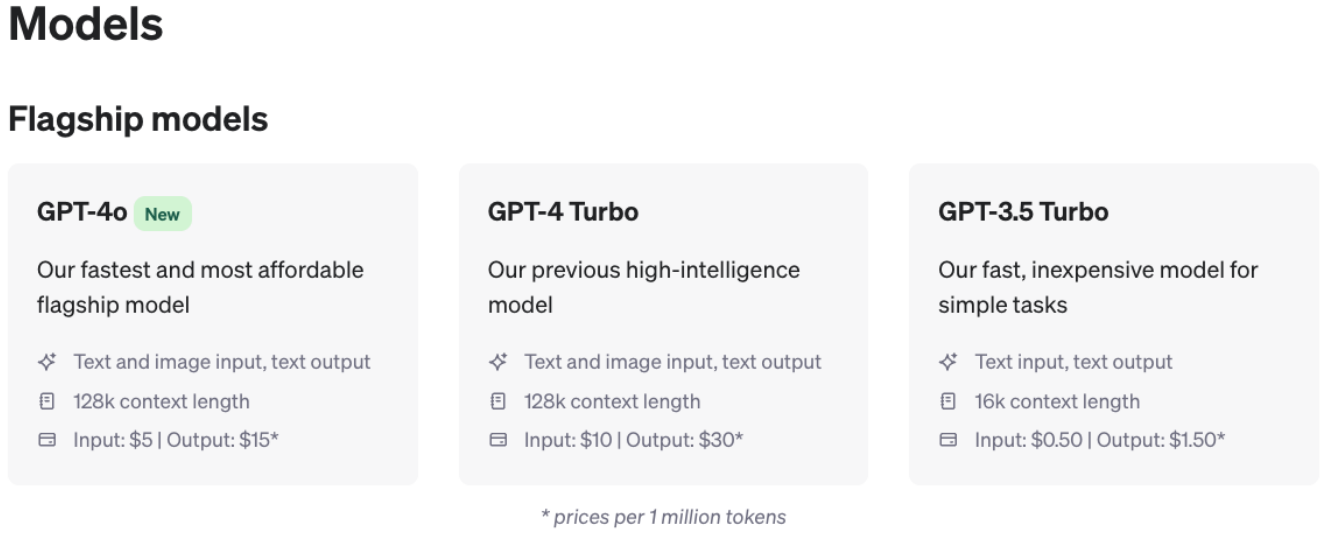
컨텍스트 윈도우(Context Window), 출력 토큰

Fast campus

- **Context Window**: 모델이 한 번에 처리할 수 있는 최대 입출력 토큰 수입니다.
- **max_tokens**: 모델이 답변으로 생성할 수 있는 최대 출력 토큰 수입니다.



참고: https://platform.openai.com/docs/models

Context Window (입력과 출력을 처리할 수 있는 최대 토큰 길이)

Context Length는 LLM이 한 번에 처리할 수 있는 최대 토큰 수를 의미합니다. GPT-3.5와 GPT-4 모델은 긴 텍스트를 다룰 수 있도록 설계되었으며, 일반적으로 수천 개의 토큰을 한 번에 처리할 수 있습니다.

- **GPT-3.5의 Context Length**: 약 16K 토큰
- **GPT-4의 Context Length**: 모델의 버전에 따라 다르지만, 기본 버전은 8,192 토큰, 확장 버전은 최대 32,768 토큰까지 처리할 수 있습니다.

max_tokens (답변에 대한 최대 출력 토큰수)

max_tokens는 모델이 생성할 수 있는 최대 출력 토큰 수를 지정하는 매개변수입니다. 이는 모델이 응답으로 생성할 수 있는 최대 텍스트 길이를 결정합니다. 사용자가 max_tokens 값을 설정하면, 모델은 지정된 수의 토큰까지 출력을 생성하게 됩니다.

- 예를 들어, max_tokens를 100으로 설정하면, 모델은 최대 100개의 토큰까지 출력합니다.
- 이 값은 모델의 context length를 초과하지 않아야 합니다.

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-4-turbo`	The latest GPT-4 Turbo model with vision capabilities. Vision requests can now use JSON mode and function calling. Currently points to `gpt-4-turbo-2024-04-09`.	128,000 tokens	Up to Dec 2023
`gpt-4-turbo-2024-04-09`	GPT-4 Turbo with Vision model. Vision requests can now use JSON mode and function calling. `gpt-4-turbo` currently points to this version.	128,000 tokens	Up to Dec 2023
`gpt-4-turbo-preview`	GPT-4 Turbo preview model. Currently points to `gpt-4-0125-preview`.	128,000 tokens	Up to Dec 2023
`gpt-4-0125-preview`	GPT-4 Turbo preview model intended to reduce cases of "laziness" where the model doesn't complete a task. Returns a maximum of 4,096 output tokens. Learn more.	128,000 tokens	Up to Dec 2023
`gpt-4-1106-preview`	GPT-4 Turbo preview model featuring improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. This is a preview model. Learn more.	128,000 tokens	Up to Apr 2023

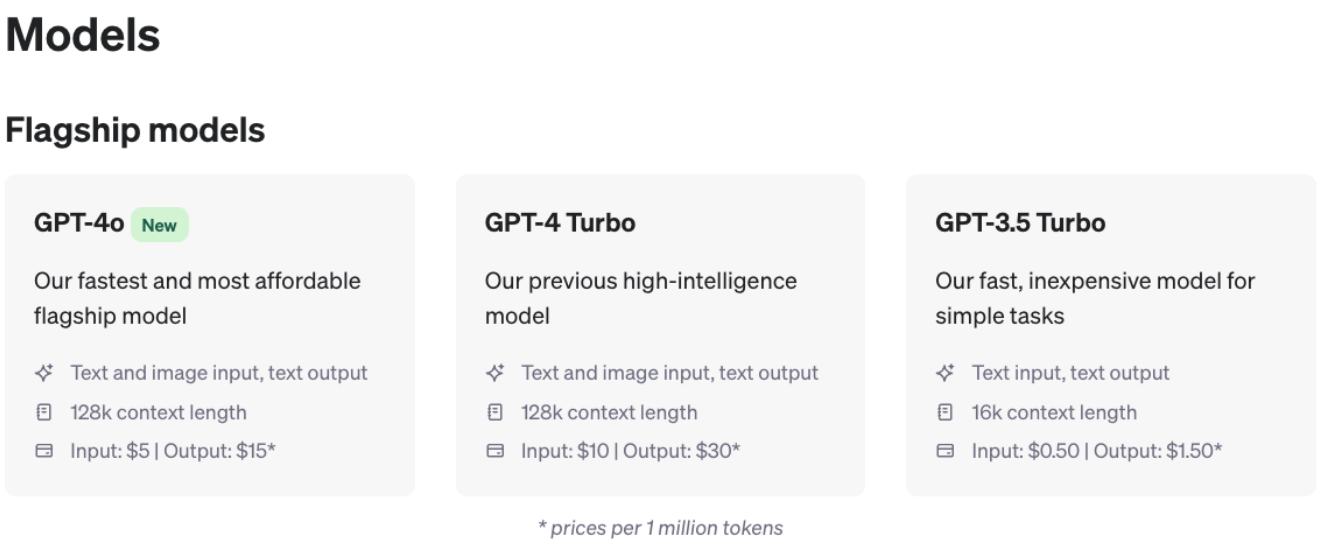
대략적인 비용

출처: https://invertedstone.com/calculators/openai-pricing/

참고
4096개의 토큰은 대략 **한글 1350자(글자수 기준, 단어 아님) 정도 되는 길이**이며, **워드 문서 1장에 500자 정도 기입한다고 계산하였을 때 약 3장** 조금 못되는 정도되는 분량입니다.

입력과 출력의 비용은 다르다!

링크: https://livechatai.com/gpt-4o-pricing-calculator



참고: https://platform.openai.com/docs/models

구조에 대한 이해가 중요!
(그렇지 않으면 불필요한 비용이 늘어날 수 있음)