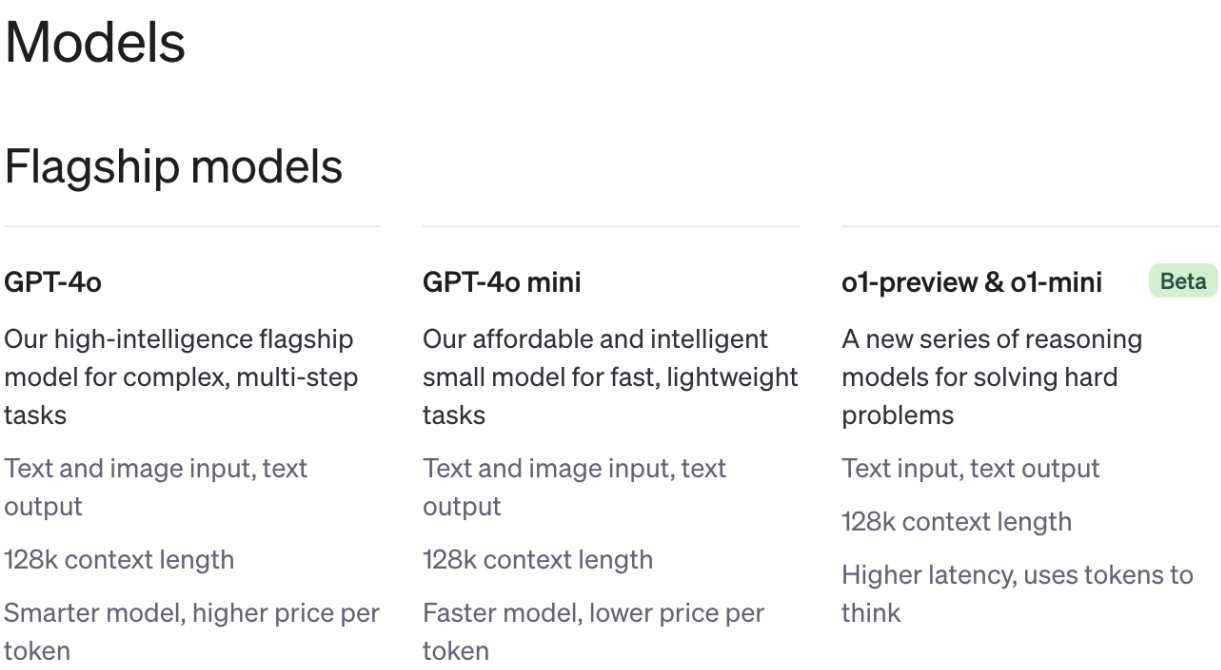
- Context Wiondow: 모델이 한 번에 처리할 수 있는 최대 입출력 토큰 수를 말합니다.
- max_tokens: 모델이 답변으로 생성할 수 있는 최대 토큰 수를 말합니다.



Context Window

- **Context Window**는 AI 모델이 한 번에 처리할 수 있는 최대 토큰 수를 의미합니다.
- 이 값은 주어진 텍스트 길이를 제한하며, 이 창 안에 들어가는 모든 입력과 출력의 합이 해당 창 크기를 넘지 않아야 합니다.
- 예를 들어, GPT-4의 경우 컨텍스트 윈도우가 8,192 토큰 또는 32,768 토큰 등 다양한 크기로 제공됩니다. 이 창이 클수록 긴 문장이나 문단을 더 잘 이해하고 처리할 수 있습니다.
- 더 큰 Context Window는 긴 대화나 복잡한 명령어를 처리할 때 유리하지만, 일반적으로 모델의 메모리 사용량도 증가하게 됩니다.

max_tokens

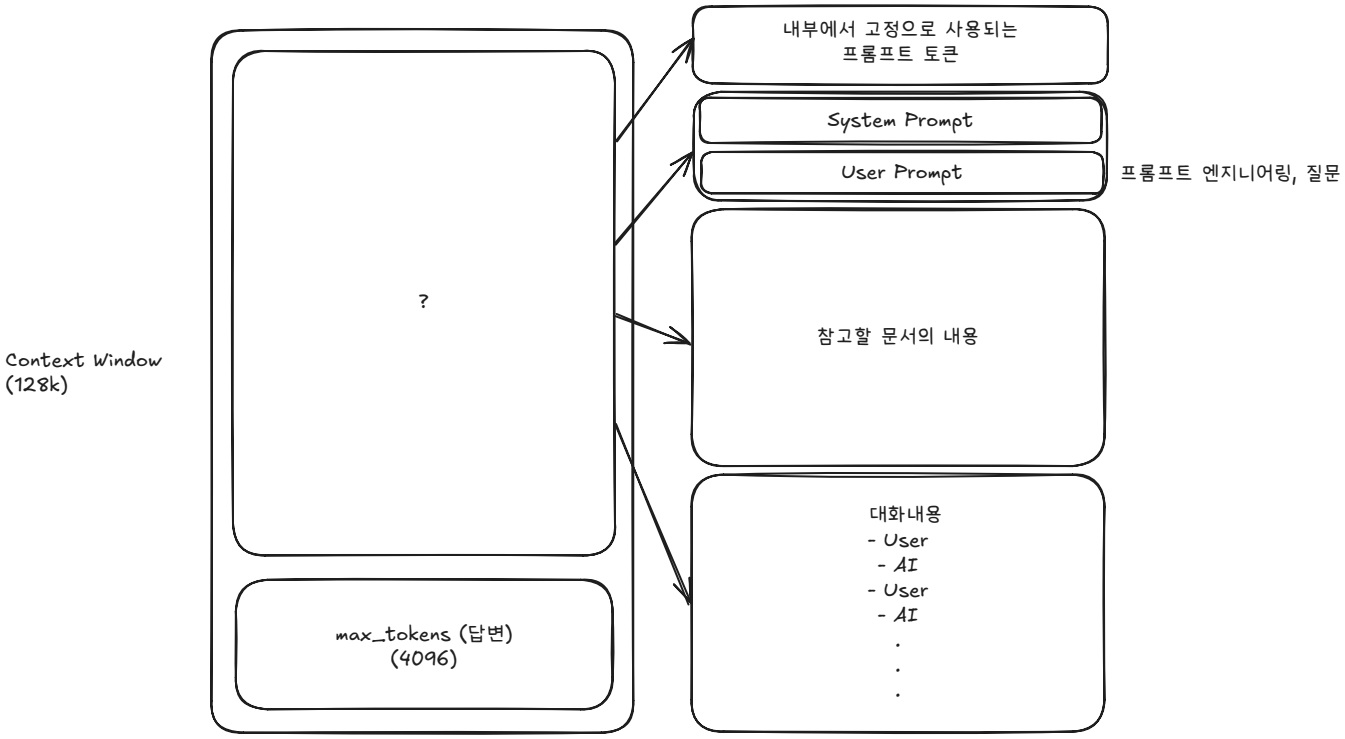
- **max_tokens**는 모델이 출력으로 생성할 수 있는 최대 토큰 수를 지정하는 설정입니다.
- 예를 들어, max_tokens를 100으로 설정하면 모델은 최대 100개의 토큰으로만 응답을 생성합니다.
- max_tokens와 Context Window는 상호작용합니다. 예를 들어, Context Window가 4,000 토큰이고 이미 3,000개의 토큰이 입력으로 사용되었다면, 최대 1,000개의 토큰만 출력으로 생성할 수 있습니다.

GPT-4o

GPT-4o ("o" for "omni") is our most advanced GPT model. It is multimodal (accepting text or image inputs and outputting text), and it has the same high intelligence as GPT-4 Turbo but is much more efficient—it generates text 2x faster and is 50% cheaper. Additionally, GPT-4o has the best vision and performance across non-English languages of any of our models. GPT-4o is available in the OpenAI API to paying customers. Learn how to use GPT-4o in our text generation guide.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS	TRAINING DATA
`gpt-4o` Our high-intelligence flagship model for complex, multi-step tasks. GPT-4o is cheaper and faster than GPT-4 Turbo. Currently points to `gpt-4o-2024-08-06`.	128,000 tokens	16,384 tokens	Up to Oct 2023
`gpt-4o-2024-08-06` Latest snapshot that supports Structured Outputs. `gpt-4o` currently points to this version.	128,000 tokens	16,384 tokens	Up to Oct 2023
`gpt-4o-2024-05-13` Original `gpt-4o` snapshot from May 13, 2024.	128,000 tokens	4,096 tokens	Up to Oct 2023
`chatgpt-4o-latest` The `chatgpt-4o-latest` model version continuously points to the version of GPT-4o used in ChatGPT, and is updated frequently, when there are significant changes.	128,000 tokens	16,384 tokens	Up to Oct 2023

구조



문서들을 참고해서 넣어줘야 하는 문서의 내용이 많으면 많을 수록 토큰의 비용이 크게 나올 것입니다. RAG 시스템을 만들때도 필요한 부분만 발췌해서 넣어주는 이유입니다.(비용과 성능적으로 유리)