Models of Note w/ Pros & Cons

Gpt 3.5 Turbo Instruct

* Designed for Natural Language Processing with reduced bias
* Designed to ingest instructions and ingest feedback based on output (Rather than just answering questions or other tasks GPT may perform).
* Fast
* $0.0015/1k tokens input
* $0.002/1k tokens output

GPT 4

* Designed for many different fields, natural language processing being one of them. Demonstrates better understanding of prompts ingested, but can be shallow in application because of the generalization required for it to function on many fields.
* Notably slow
* $0.0025/1k tokens input
* $0.01/1k tokens output

GPT-4 Mini

* Less powerful than GPT 4, but more powerful than GPT 3.5 turbo for text processing
* $0.00015/1k tokens input
* $0.0006/1k tokens output

O1

* Wrapper for GPT-4
* Uses chain-of-thought processing and breaks down input tokens into what is needed to answer the question. Then feeds it to GPT 4 before producing an output
* Notably slow
* $0.003/1k tokens input
* $0.012/1k tokens output

Mini models

* Lower performance & lower cost than their adult counterparts (by what degree is unspecified)

Text Embedding 3 Small

* Able to determine similarities between prompts/tokens inputted on a given labeled/unlabeled vector of set size.
* $0.00002/1k tokens input
* $0.00001/1k tokens output