**A screenshot of a computer

AI-generated content may be incorrect.**

**Task 1: Defining your Problem and Audience**

**You are an AI Solutions Engineer**. **What** problem do you want to solve? **Who** is it a problem for?

<aside> 📝

Task 1: Articulate the problem and the user of your application

*Hints:*

* *Create a list of potential questions that your user is likely to ask!*

*What is the user’s job title, and what is the part of their job function that you’re trying to automate?* </aside>

**✅ Deliverables**

1. Write a succinct 1-sentence description of the problem
2. Write 1-2 paragraphs on why this is a problem for your specific user

The problem: There is shortage of expertise in entry-level PTs . Physical therapy licensure candidates need to pass their exams the first time, and be able to excel from the start of their career. Up to today there is no personalized assistance to help them overcome their weaknesses and 20% of them have to repeat their expensive exams with expensive exam-help providers and have to wait another year. There is shortage of expert PTs in several cities and towns in USA.  
  
PT student complained that prep providers’ materials although included questions that they knew how to answer they did not address all their weaknesses. That would cause them fail their licensure exams, and to try other prep-providers or go back to school which all added to cost of both money and time. In the mean time the PT facilities are looking for more knowledgeable entry level therapists.

**Task 2: Propose a Solution**

Now that you’ve defined a problem and a user, *there are many possible solutions*.

Choose one, and articulate it.

<aside> 📝

Task 2: Articulate your proposed solution

*Hint:*

* *Paint a picture of the “better world” that your user will live in. How will they save time, make money, or produce higher-quality output?*
* *Recall the* [*LLM Application stack*](https://a16z.com/emerging-architectures-for-llm-applications/) *we’ve discussed at length* </aside>

**✅ Deliverables**

1. Write 1-2 paragraphs on your proposed solution. How will it look and feel to the user?
2. Describe the tools you plan to use in each part of your stack. Write one sentence on why you made each tooling choice.
   1. LLM opus 4 for reasoning gpt4.1-mini for SDG
   2. Embedding Model
   3. Orchestration AI
   4. Vector Database Qdrant
   5. Monitoring Arize Phoenix
   6. Evaluation
   7. User Interface
   8. (Optional) Serving & Inference
3. Where will you use an agent or agents? What will you use “agentic reasoning” for in your app?

Resources LLM

Below are solid, current picks (as of **July 30, 2025**) for each part of your PT‑licensure study app. I list **model → price** next to each, using provider list prices (USD, per **1M tokens** unless noted).

**1) RAG on your documents**

**Embeddings**

* **OpenAI text-embedding-3-large → $0.13/M** (best raw quality; can truncate dims to save storage). ([OpenAI](https://openai.com/index/new-embedding-models-and-api-updates/))
* **OpenAI text-embedding-3-small → $0.02/M** (cost‑efficient baseline). ([OpenAI](https://openai.com/index/new-embedding-models-and-api-updates/))
* **Voyage voyage‑3.5 → $0.06/M**; **voyage‑3.5‑lite → $0.02/M** (strong price‑performance; long‑context). ([Voyage AI](https://blog.voyageai.com/2025/05/20/voyage-3-5/?utm_source=chatgpt.com))
* **BAAI bge‑m3 (open‑source) → $0 model license** (MIT; self‑host infra cost only). ([Hugging Face](https://huggingface.co/BAAI/bge-m3/tree/main?utm_source=chatgpt.com), [GitHub](https://github.com/FlagOpen/FlagEmbedding?utm_source=chatgpt.com))

**Rerankers (for top‑k re‑ordering)**

* **Cohere Rerank v3.5 → $2.00 / 1k requests (Pinecone‑hosted)**; official Cohere v2 deprecation listed $1.00/1k searches, v3.5 is the current line. ([docs.pinecone.io](https://docs.pinecone.io/models/cohere-rerank-3.5?utm_source=chatgpt.com), [Cohere Documentation](https://docs.cohere.com/docs/deprecations?utm_source=chatgpt.com))
* **Voyage rerank‑2.5 → $0.05/M tokens** (first **200M tokens free**). ([Voyage AI](https://docs.voyageai.com/docs/pricing?utm_source=chatgpt.com))
* **Jina Reranker v2 → token top‑ups (first 10M tokens free), then paid packages** (pricing is token‑based, not per‑request). ([jina.ai](https://jina.ai/reranker/))

**Generators used inside RAG (to answer with retrieved context)**

* **OpenAI gpt‑4o → $2.50/M in, $10/M out** (general‑purpose, strong factuality). ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))
* **OpenAI gpt‑4o‑mini → $0.15/M in, $0.60/M out** (cheap & fast for most turns). ([OpenAI Platform](https://platform.openai.com/docs/models/compare?model=gpt-4o-mini&utm_source=chatgpt.com), [OpenAI](https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/?utm_source=chatgpt.com))
* **Anthropic Claude 3.7 Sonnet → $3/M in, $15/M out** (reasoning‑optimized; good judge/generator). ([Reuters](https://www.reuters.com/technology/artificial-intelligence/anthropic-launches-advanced-ai-hybrid-reasoning-model-2025-02-24/?utm_source=chatgpt.com))
* **Together AI: Llama 3.1 70B (Turbo) → $0.88/M** (open‑weights via API; attractive cost). ([Together AI](https://www.together.ai/pricing))
* **Fireworks: Qwen 2.5 72B Instruct → $0.90/M** (budget alternative). ([Fireworks AI](https://fireworks.ai/models/fireworks/qwen2p5-72b-instruct?utm_source=chatgpt.com))

**2) Web knowledge via Tavily**

Use Tavily for retrieval, then a **small model for query‑planning** and a **bigger model for synthesis**.

* **Tavily API plans**: **Free 1,000 credits/mo**; **Pay‑as‑you‑go $0.008/credit** (credits ≈ searches/tools). ([tavily.com](https://tavily.com/?utm_source=chatgpt.com))
* **Planner/Summarizer models**:
  + **OpenAI gpt‑4o‑mini → $0.15/M in, $0.60/M out** (great for query planning, snippet summarization). ([OpenAI Platform](https://platform.openai.com/docs/models/compare?model=gpt-4o-mini&utm_source=chatgpt.com))
  + **OpenAI o4‑mini (reasoning‑lean) → $1.10/M in, $4.40/M out** (helpful for tricky search decomposition). ([GPT for Work](https://gptforwork.com/tools/openai-chatgpt-api-pricing-calculator?utm_source=chatgpt.com))
  + **Claude 3.5 Haiku → $0.80/M in, $4/M out** (fast & inexpensive). ([Anthropic](https://docs.anthropic.com/en/docs/about-claude/models/overview?utm_source=chatgpt.com))

**3) Evaluation (scoring retrievers & answers)**

Use **Ragas** for metrics like *context\_precision* and *context\_recall* (open‑source; no license fee). Pair it with a modest judge model:

* **Ragas framework → $0** (library); you only pay for the judge LLM you call. ([Ragas](https://docs.ragas.io/en/stable/concepts/metrics/?utm_source=chatgpt.com))
* **Judge LLMs** (pick based on budget):
  + **OpenAI gpt‑4o‑mini → $0.15/M in, $0.60/M out** (good default). ([OpenAI Platform](https://platform.openai.com/docs/models/compare?model=gpt-4o-mini&utm_source=chatgpt.com))
  + **OpenAI gpt‑4o → $2.50/M in, $10/M out** (higher‑quality judging). ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))
  + **Claude 3.5 Haiku → $0.80/M in, $4/M out** (very cost‑effective for eval prompts). ([Anthropic](https://docs.anthropic.com/en/docs/about-claude/models/overview?utm_source=chatgpt.com))

(If you specifically need a *non‑LLM* retrieval eval, you can combine Ragas’ non‑LLM *context\_recall* with a reranker baseline; still $0 from Ragas itself.) ([Ragas](https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall/?utm_source=chatgpt.com))

**4) Synthetic data generation (new exam questions targeting weaknesses)**

* **OpenAI gpt‑4o → $2.50/M in, $10/M out** (balanced quality for item‑writing with constraints/rationales). ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))
* **Claude 3.7 Sonnet → $3/M in, $15/M out** (excellent at rubric‑guided question crafting). ([Reuters](https://www.reuters.com/technology/artificial-intelligence/anthropic-launches-advanced-ai-hybrid-reasoning-model-2025-02-24/?utm_source=chatgpt.com))
* **Together: Llama 3.1 70B Turbo → $0.88/M** (low‑cost bulk generation; you can add a quality filter/judge pass). ([Together AI](https://www.together.ai/pricing))
* **DeepSeek‑R1 (Together) → $3/M in, $7/M out** (reasoning‑heavy variants; useful for multi‑step clinical vignettes). ([Together AI](https://www.together.ai/pricing))

**Quick “starter” stack that balances quality & cost**

* **Embeddings**: text-embedding-3-small for prod (cheap), keep ‑3-large for hard cases. ([OpenAI](https://openai.com/index/new-embedding-models-and-api-updates/))
* **Reranker**: Voyage rerank‑2.5 (cheap per‑token; generous free tier). ([Voyage AI](https://docs.voyageai.com/docs/pricing?utm_source=chatgpt.com))
* **Retriever judge**: gpt‑4o‑mini (Ragas). ([OpenAI Platform](https://platform.openai.com/docs/models/compare?model=gpt-4o-mini&utm_source=chatgpt.com))
* **Answering model**: gpt‑4o (fallback to Claude 3.7 Sonnet for borderline items). ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com), [Reuters](https://www.reuters.com/technology/artificial-intelligence/anthropic-launches-advanced-ai-hybrid-reasoning-model-2025-02-24/?utm_source=chatgpt.com))
* **Web search**: Tavily + gpt‑4o‑mini planner + gpt‑4o synthesizer. ([tavily.com](https://tavily.com/?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models/compare?model=gpt-4o-mini&utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))

**Notes:** Prices are list prices and can vary by region, batch/Prompt‑Caching discounts, or vendor (Azure/Bedrock/Together/Fireworks). Always check the provider’s pricing page before locking in. ([Together AI](https://www.together.ai/pricing), [Fireworks AI](https://fireworks.ai/models/fireworks/qwen2p5-72b-instruct?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models/compare?model=gpt-4o-mini&utm_source=chatgpt.com), [tavily.com](https://tavily.com/?utm_source=chatgpt.com))

Here’s a **quick, source‑backed comparison** of **GPT‑4o** vs **GPT‑4.1** as of **July 30, 2025**.

**Summary**

| **Model** | **Positioning** | **Reasoning behavior** | **Typical strengths** | **Price (per 1M tokens)** |
| --- | --- | --- | --- | --- |
| **GPT‑4o** | “Omni” GPT—multimodal, realtime | A GPT‑family model (not an “reasoning” series model); optimized for fast, multimodal interactions. ([OpenAI Platform](https://platform.openai.com/docs/guides/reasoning-best-practices?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models?utm_source=chatgpt.com)) | Real‑time voice/vision, multimodal IO, solid general text/code | **Input $2.50**, **Output $10.00**, **Cached input $1.25**. ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com)) |
| **GPT‑4.1** | Flagship GPT for complex tasks | Also a GPT‑family (not “reasoning” series) model; newer text/code flagship; larger max output and improved tool/diff behaviors. ([OpenAI Platform](https://platform.openai.com/docs/guides/reasoning-best-practices?utm_source=chatgpt.com), [OpenAI](https://openai.com/index/gpt-4-1/?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models?utm_source=chatgpt.com)) | Text/code tasks, tool use, long outputs | **Input $2.00**, **Output $8.00**, **Cached input $0.50**. ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4.1?utm_source=chatgpt.com)) |

**What that means in practice**

* **Reasoning:** Neither 4o nor 4.1 is an “o‑series reasoning model” (like o4‑mini/o3); they’re GPT models that don’t generate hidden chain‑of‑thought. For structured “think‑hard” tasks, OpenAI recommends the **o‑series**; for general complex tasks, **GPT‑4.1** is the newer flagship and typically stronger in text/code and tool use than 4o. ([OpenAI Platform](https://platform.openai.com/docs/guides/reasoning-best-practices?utm_source=chatgpt.com), [OpenAI](https://openai.com/index/gpt-4-1/?utm_source=chatgpt.com))
* **Multimodality & latency:** If you need **realtime voice/vision** and broad multimodal IO, **GPT‑4o** is designed for that “omni” use case. ([OpenAI Platform](https://platform.openai.com/docs/models?utm_source=chatgpt.com))
* **Cost:** **GPT‑4.1 is ~20% cheaper** than 4o on both input ($2.00 vs $2.50 per 1M) and output ($8.00 vs $10.00 per 1M). Cached‑input discounts are also better on 4.1 ($0.50 vs $1.25 per 1M). ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4.1?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))
* **Token/tool pricing footnote:** For specialized tool models (e.g., search/realtime), there can be extra per‑tool‑call fees; base token prices above still apply. Check the pricing page when enabling those tools. ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o-search-preview?utm_source=chatgpt.com))

**Recommendations**

* Choose **GPT‑4.1** when you want **best text/code quality**, **lower cost**, and **longer max output** (e.g., detailed rationales or diffs). ([OpenAI](https://openai.com/index/gpt-4-1/?utm_source=chatgpt.com))
* Choose **GPT‑4o** when your app needs **live voice, vision, or “omni” interactions** with strong general performance, and you’re OK with the higher price. ([OpenAI Platform](https://platform.openai.com/docs/models?utm_source=chatgpt.com), [OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))

If you share your usage profile (avg input/output tokens per call, % multimodal, concurrency), I can estimate monthly spend for each model and suggest caching/Batch API levers to reduce cost. ([OpenAI Platform](https://platform.openai.com/docs/models/gpt-4o?utm_source=chatgpt.com))

If you're looking for **one of the best tools to monitor an LLM application**, especially for production use monitoring cost, performance, data drift, prompt behavior, and output quality, an excellent choice is:

**🏆 Arize AI (specifically Arize Phoenix)**

**What it does:**

* Tracks **real-time latency**, **token usage**, **cost**, **error rates**, and **concept drift**.
* Traces **prompt chains** across multi-turn or RAG workflows with full context.
* Supports **evaluation scoring** (e.g. hallucination rates, accuracy) alongside real-time metrics.
* Offers filters, alerts, and root-cause analytics for prompt/model/version/regressor tracking. ([neptune.ai](https://neptune.ai/blog/llm-observability?utm_source=chatgpt.com), [LinkedIn](https://www.linkedin.com/pulse/edition-23-top-5-llm-observability-tools-2025-futureagi-ocnlc?utm_source=chatgpt.com), [edenai.co](https://www.edenai.co/post/top-5-paid-observability-platforms-for-llms-unlocking-advanced-monitoring-for-ai-systems?utm_source=chatgpt.com))

**Why it's standout:**

* Built specifically for LLM-driven systems.
* Designed to ensure **reliability, compliance**, and performance at scale.
* Strong in **troubleshooting production problems** and defining causality in prompt/model changes. ([voltagent.dev](https://voltagent.dev/blog/llm-observability-tools/?utm_source=chatgpt.com" \o "Top 5 LLM Observability Tools - VoltAgent))

**✅ Comparison vs. Other Tools**

| **Tool** | **Focus Area** | **Why Arize Phoenix Shines** |
| --- | --- | --- |
| **Arize Phoenix** | Full LLM observability + evaluation | Real-time tracing, cost/drift metrics, audit tools |
| **Langfuse** | Developer-centric Agent & prompt flow tracking | Great open-source SDK & dataset integration ([posthog.com](https://posthog.com/blog/best-open-source-llm-observability-tools?utm_source=chatgpt.com)) |
| **PostHog** | Built-in free LLM telemetry + product analytics | Best free-tier coverage w/ SQL dashboards ([posthog.com](https://posthog.com/blog/best-open-source-llm-observability-tools?utm_source=chatgpt.com)) |
| **Confident AI** | Evaluation-first workflow (metrics + dataset uplift) | Superior evaluation tooling, limited real-time tracing ([confident-ai.com](https://www.confident-ai.com/blog/greatest-llm-evaluation-tools-in-2025?utm_source=chatgpt.com)) |

**✅ Final Recommendation**

Go with **Arize Phoenix** if your priority is **robust LLM observability**, covering real-time monitoring, drift detection, cost tracking, and per-interaction traceability in production environments.

Let me know if you prefer open-source, free-hosted platforms like **Langfuse** or **PostHog**, and I can explain how to integrate them seamlessly.

**Here’s a concise comparison of OpenAI’s reasoning models—o3-mini, o4-mini, and (if applicable) full o3—versus Anthropic’s Claude reasoning models as of July 30, 2025. Focus areas: reasoning ability and price per 1M tokens.**

**🧠 Reasoning Performance**

**OpenAI Models**

* **o3-mini (high effort):**
  + **Substantial gains on math/science & STEM benchmarks: up to ~87% on AIME, ~79.7% on GPQA Diamond, ~2130 Elo on Codeforces, ~49.3 on SWE-bench Verified.**
  + **Test-time chain-of-thought reasoning with efficient internal strategy. ([arXiv](https://arxiv.org/abs/2502.15631?utm_source=chatgpt.com" \o "The Relationship Between Reasoning and Performance in Large Language Models -- o3 (mini) Thinks Harder, Not Longer),** [**Wikipedia**](https://en.wikipedia.org/wiki/Reasoning_language_model?utm_source=chatgpt.com)**,** [**Wikipedia**](https://en.wikipedia.org/wiki/OpenAI_o3?utm_source=chatgpt.com)**)**
* **o4-mini:**
  + **Released April 2025. Strong performer on math (AIME 2025: ~99.5% pass@1 with Python tool). Better efficiency and multimodal support. (**[**OpenAI**](https://openai.com/index/introducing-o3-and-o4-mini/?utm_source=chatgpt.com)**)**

**Claude Models**

* **Claude 3.5 Haiku:**
  + **Good baseline reasoning; lags both Sonnet and GPT-4o-mini on live reasoning benchmarks. On MMLU and coding it performs well, but weaker than Claude Sonnet. (**[**Reddit**](https://www.reddit.com/r/singularity/comments/1gk78gu/claude_35_haiku_scores_the_exact_same_as_3_haiku/?utm_source=chatgpt.com)**)**
* **Claude 3.5 Sonnet:**
  + **Competitive: 86.8% MMLU, 95% GSM8K, 60.1% MATH; strong graduate‑level, multilingual math (~90.7%), DROP F1 ~83.1%, BIG‑Bench‑Hard ~86.8%. Top-tier performance. ([PromptLayer](https://blog.promptlayer.com/big-differences-claude-3-5-vs-gpt-4o/?utm_source=chatgpt.com" \o "Claude 3.5 vs GPT 4o: Which LLM Reigns Supreme? - PromptLayer))**
* **Claude 3.7 Sonnet (hybrid):**
  + **Allows choice of fast vs extended reasoning dynamically. Improved accuracy on multi-step/complex tasks. (**[**Reuters**](https://www.reuters.com/technology/artificial-intelligence/anthropic-launches-advanced-ai-hybrid-reasoning-model-2025-02-24/?utm_source=chatgpt.com)**,** [**Tom's Guide**](https://www.tomsguide.com/ai/what-is-claude-everything-you-need-to-know-about-anthropics-ai-powerhouse?utm_source=chatgpt.com)**)**
* **Claude 4 Sonnet / Opus (2025 family):**
  + **Highest-level hybrid reasoning models. Opus 4 is flagship, deep multi-step, long-context support (200k tokens), best coding/reasoning among benchmarks. Sonnet 4 is slightly more efficient. (**[**Wikipedia**](https://en.wikipedia.org/wiki/Claude_%28language_model%29?utm_source=chatgpt.com)**)**

**💰 Pricing (per million tokens)**

**OpenAI Pricing**

* **o3-mini:**
  + **Input: ~$0.40/M (recent price drop), Output: ~$1.60/M. (**[**OpenAI Community**](https://community.openai.com/t/is-the-api-pricing-for-gpt-4-1-mini-and-o3-really-identical-now/1286911?utm_source=chatgpt.com)**)**
* **o4-mini:**
  + **Input: $? Not explicitly in docs but consistent with GPT‑4o‑mini: Input $0.15/M, Output $0.60/M. ([PromptLayer](https://blog.promptlayer.com/big-differences-claude-3-5-vs-gpt-4o/?utm_source=chatgpt.com" \o "Claude 3.5 vs GPT 4o: Which LLM Reigns Supreme? - PromptLayer))**

**Claude Pricing**

* **Claude 3.5 Haiku:**
  + **Input: $0.80/M, Output: $4.00/M (cache write: $1, cache hit: $0.08). (**[**Anthropic**](https://docs.anthropic.com/en/docs/about-claude/pricing?utm_source=chatgpt.com)**)**
* **Claude 3.5 Sonnet / 3.7 Sonnet:**
  + **Input: $3/M, Output: $15/M (cache write/hit similar). (**[**Anthropic**](https://docs.anthropic.com/en/docs/about-claude/pricing?utm_source=chatgpt.com)**,** [**Zapier**](https://zapier.com/blog/claude-vs-chatgpt/?utm_source=chatgpt.com)**)**
* **Claude 4 Sonnet / Opus 4:**
  + **Input: $3/M (Sonnet) or $15/M (Opus), Output: $15/M (Sonnet) or $75/M (Opus). (**[**Zapier**](https://zapier.com/blog/claude-vs-chatgpt/?utm_source=chatgpt.com)**)**

**📊 Side‑by‑Side Comparison**

| **Model** | **Reasoning Strength** | **Input $/M** | **Output $/M** |
| --- | --- | --- | --- |
| **o3‑mini (high)** | **Very good STEM/coding reasoning** | **~$0.40** | **~$1.60** |
| **o4‑mini** | **Excellent math & coding; multimodal** | **~$0.15** | **~$0.60** |
| **Claude 3.5 Haiku** | **Moderate reasoning, low price** | **$0.80** | **$4.00** |
| **Claude 3.5 Sonnet / 3.7** | **Strong graduate‑level reasoning** | **$3.00** | **$15.00** |
| **Claude 4 Sonnet** | **Improved instruction following, coding** | **$3.00** | **$15.00** |
| **Claude 4 Opus** | **Top-tier multi-step reasoning & code** | **$15.00** | **$75.00** |

**✅ Recommendation Notes**

* **Best reasoning vs cost:**
  + **o4‑mini is the most cost-efficient with strong math/coding reasoning.**
  + **o3‑mini (high effort) offers high accuracy at modest cost too.**
* **For large context & deep logic tasks:**
  + **Claude 3.5 Sonnet / 3.7 is strong, and Claude 4 Sonnet / Opus if absolute reasoning edge matters and cost is acceptable.**
* **For very low cost pipelines:**
  + **Claude 3.5 Haiku is mid-tier reasoning at moderate price but weaker than o4‑mini for multi-step benchmarks.**

**Let me know your task focus (e.g. STEM, clinical reasoning, long‑form logic) and usage volume, and I can help choose the best fit or estimate your expected costs/performance!**

**Task 2: Propose a Solution**