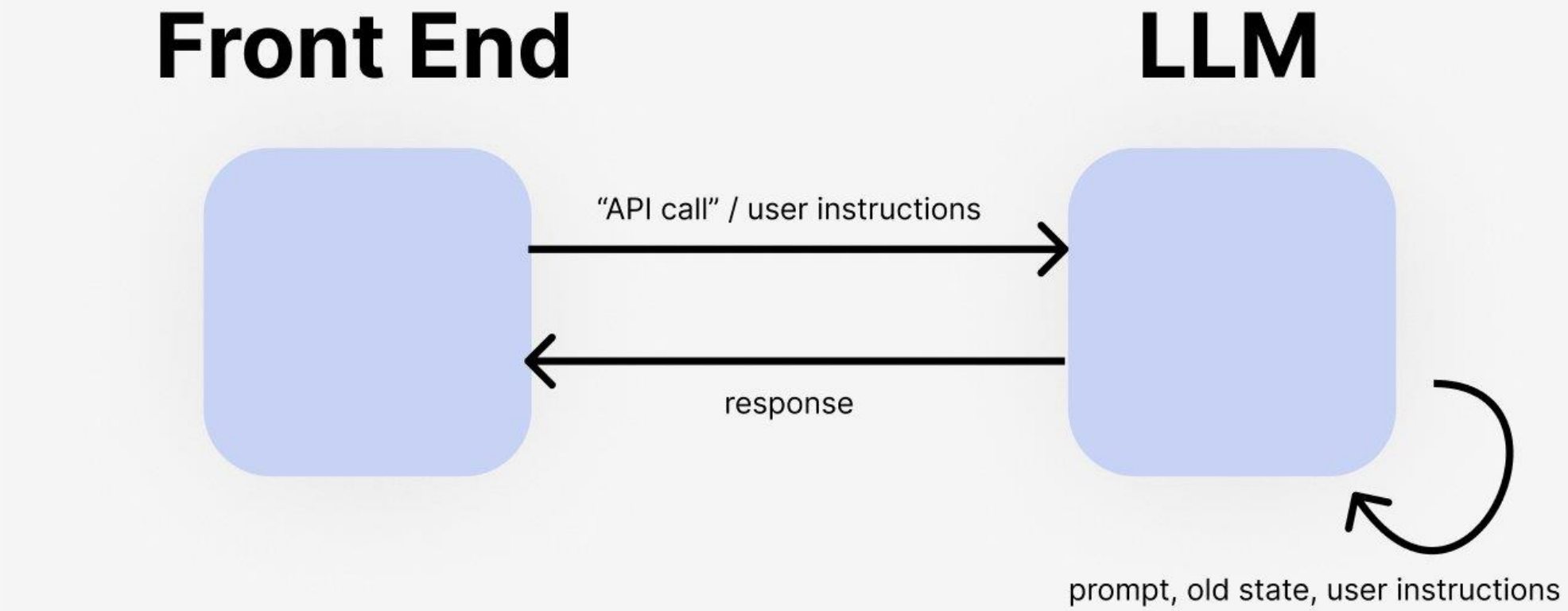
Prompt Caching

Saving our money :D

What is prompt caching?



Gemini Developer API (Per Million Tokens)

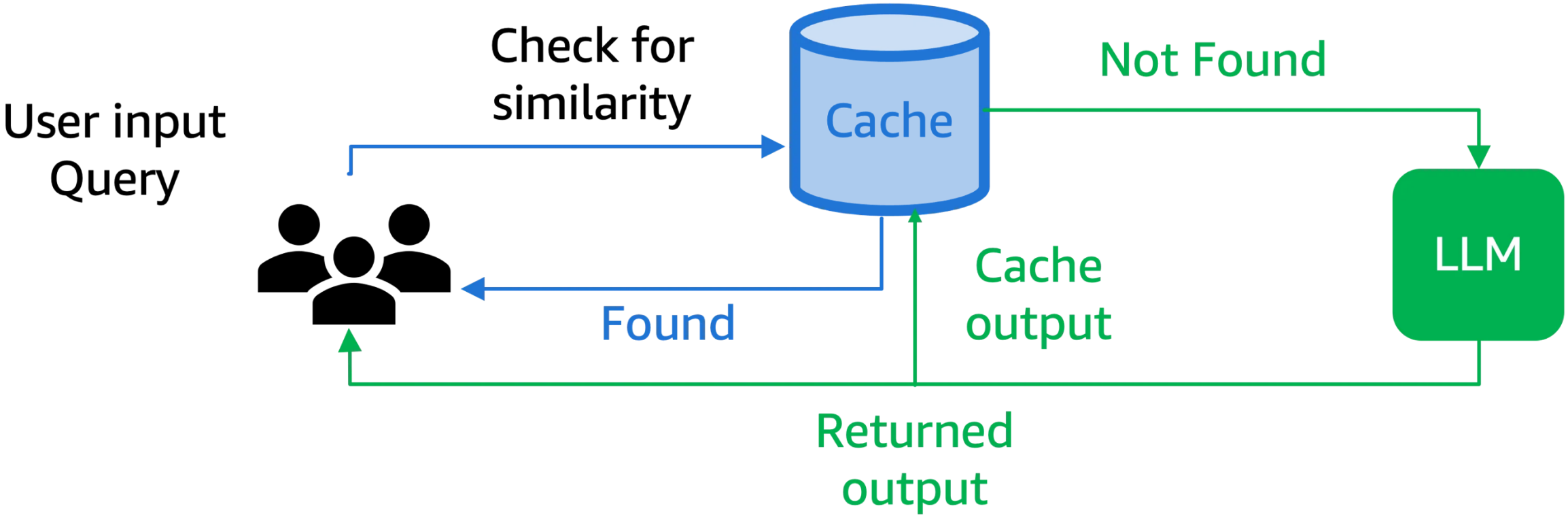
MODEL	TEXT/IMAGE/VIDEO INPUTS	AUDIO INPUTS	TEXT OUTPUTS	CONTEXT CACHING*
Gemini 2.0 Flash	$0.10	$0.70**	$0.40	*Text/image/video* $0.025 *Audio* $0.175
Gemini 2.0 Flash-Lite	$0.075	$0.075	$0.30	$0.01875
Gemini 1.5 Flash *(Provided for reference)*	$0.075 *(Prompts <= 128k)*	$0.075 *(Prompts <= 128k)*	$0.30 *(Prompts <= 128k)*	$0.01875 *(Prompts <= 128k)*
	$0.15 *(Prompts > 128k)*	$0.15 *(Prompts > 128k)*	$0.60 *(Prompts > 128k)*	$0.0375 *(Prompts > 128k)*

* Caching coming soon.

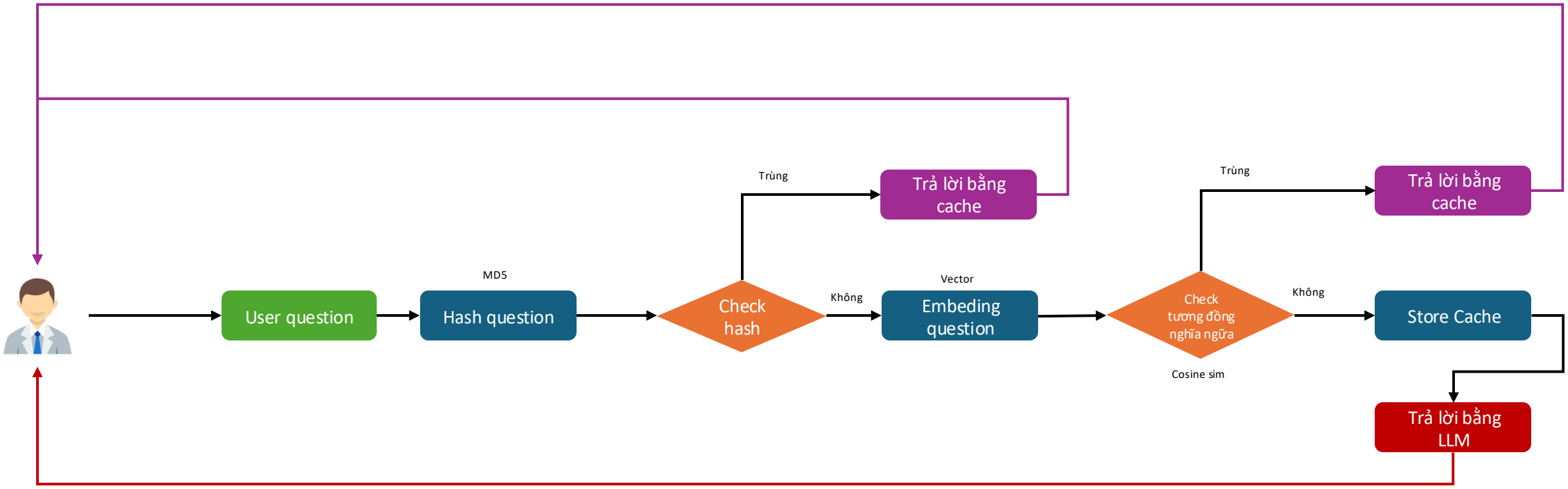
** The audio input pricing shown above will become effective February 20, 2025. Audio tokens will be charged the same as other modalities until then.

Model	Pricing	Pricing with Batch API*
gpt-4o	$5.00 / 1M input tokens	$2.50 / 1M input tokens
	$15.00 / 1M output tokens	$7.50 / 1M output tokens
gpt-4o-2024-08-06	$2.50 / 1M input tokens	$1.25 / 1M input tokens
	$10.00 / 1M output tokens	$5.00 / 1M output tokens
gpt-4o-2024-05-13	$5.00 / 1M input tokens	$2.50 / 1M input tokens
	$15.00 / 1M output tokens	$7.50 / 1M output tokens

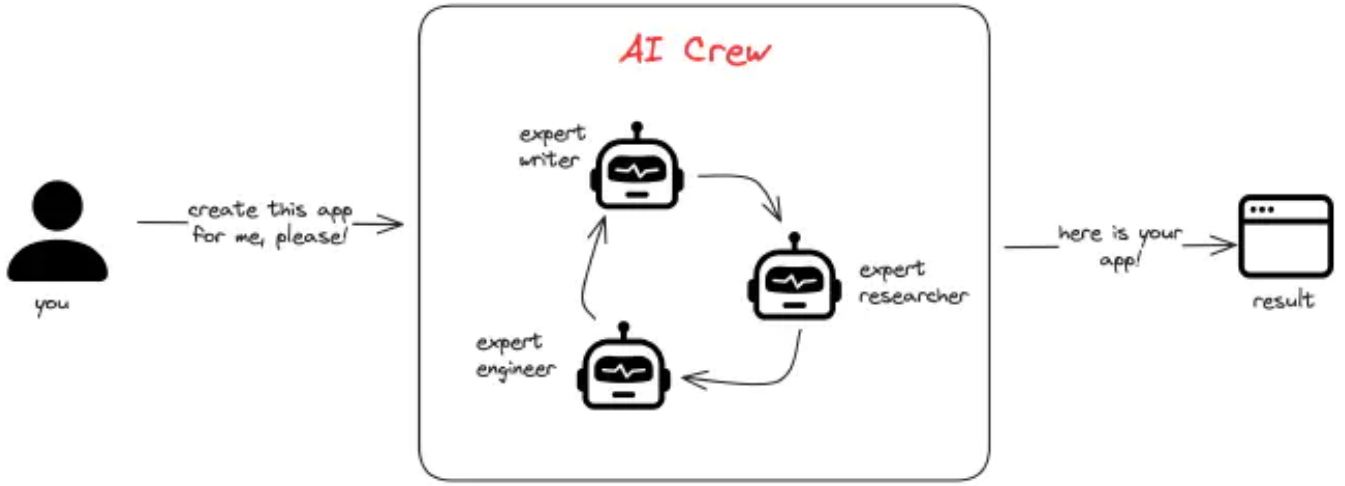
What is prompt caching?



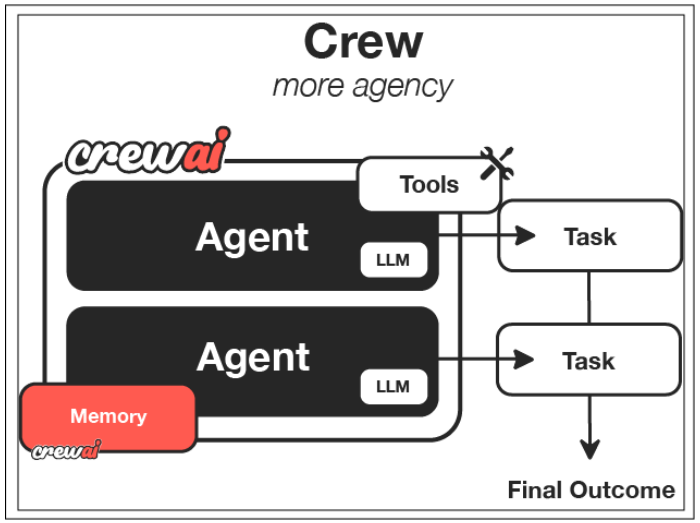
Similarity?


User question
MD5
Hash question
Check hash
Trùng
Trả lời bằng cache
Không
Vector
Embeding question
Check tương đồng nghĩa ngữa
Cosine sim
Trùng
Trả lời bằng cache
Không
Store Cache
Trả lời bằng LLM

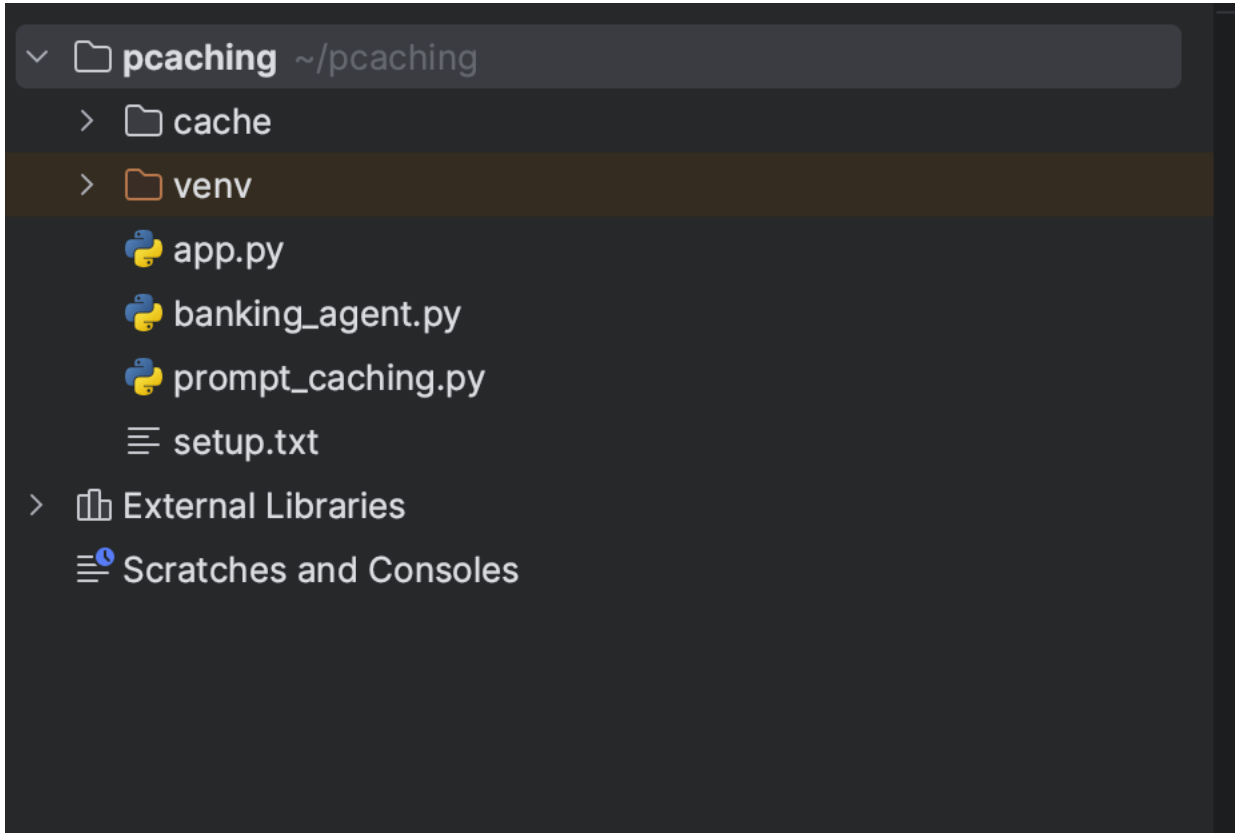
Hands-on



CrewAI is **an open-source Python framework for building and managing collaborative, multi-agent AI systems**. It allows users to create teams of AI agents, each with specific roles, goals, and tools, to tackle complex tasks collaboratively. This framework enhances AI by enabling autonomous decision-making and communication between agents, making them more effective at solving problems than individual agents working alone.



Hands-on



Need to change JSON to Redis or some DB