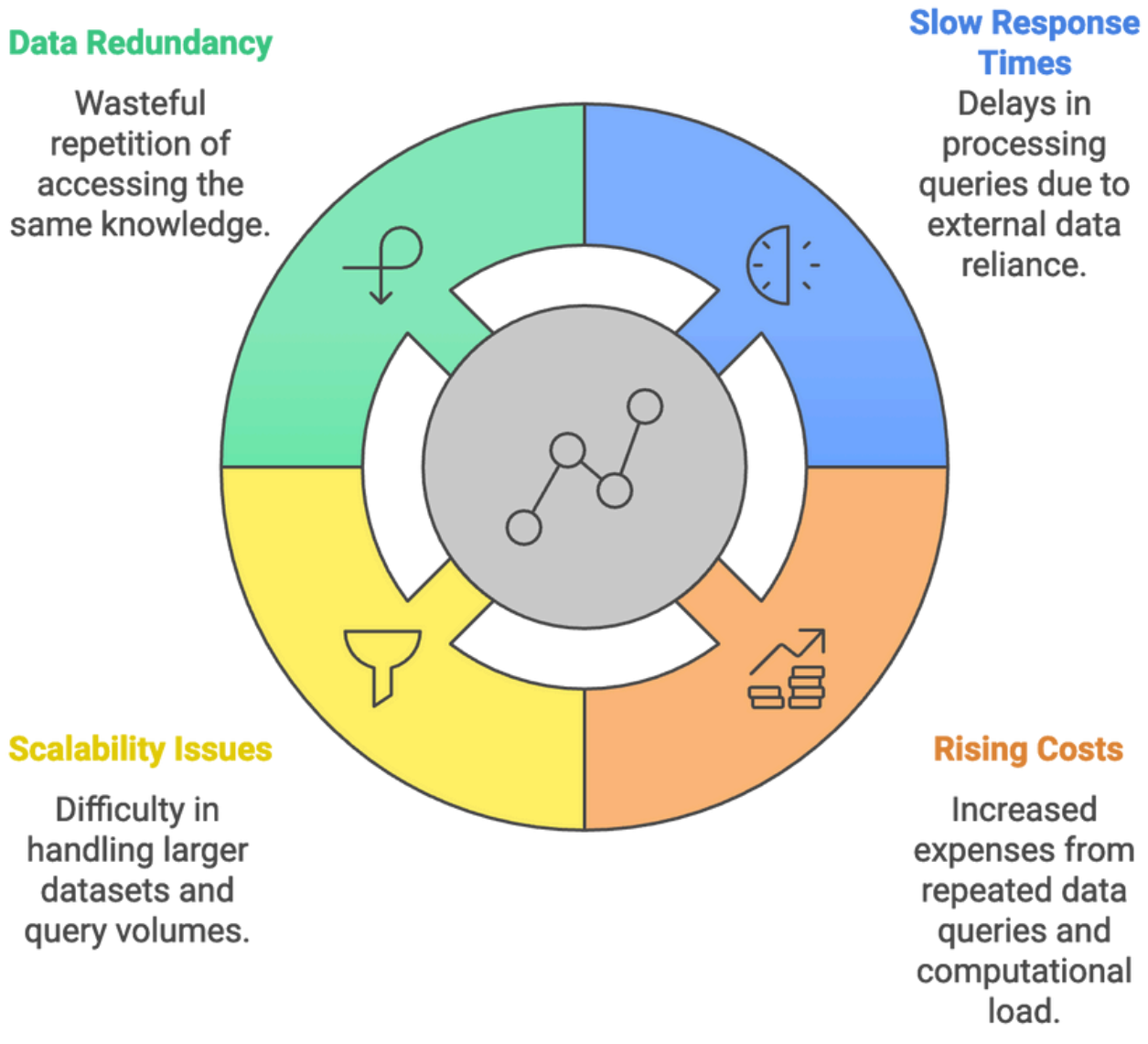
Trim cost of your RAG apps

Bhavishya Pandit

What is the challenge?

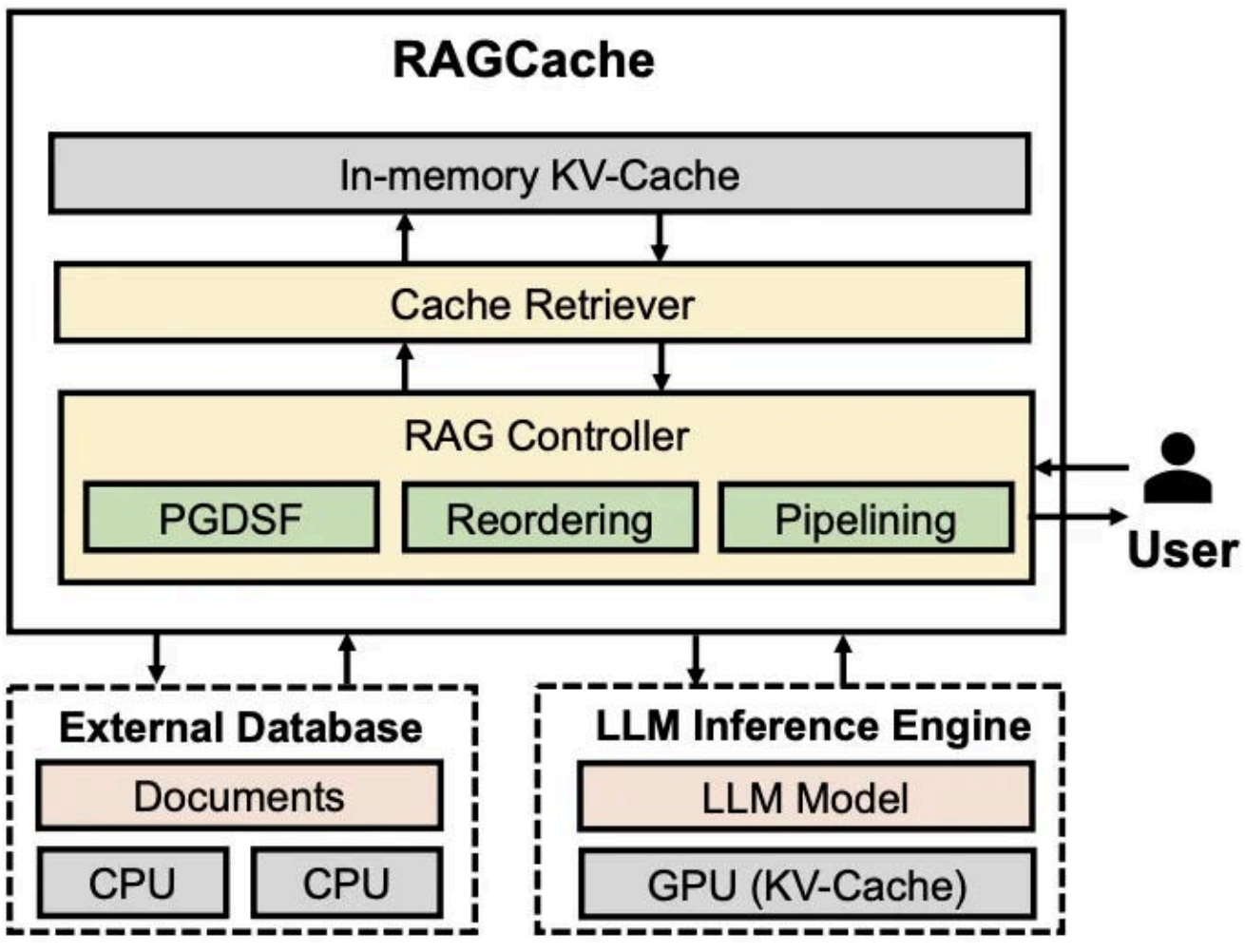
RAG systems have the power of integrating the power of knowledge retrieval with generative models. However they come with the following significant limitations:



- **Slow response times**: RAG systems rely heavily on external knowledge sources to fetch relevant information during queries. This dependence creates a lag, leading to frustratingly slow response times.
- **Rising costs**: Repeatedly querying the same data comes at a cost. The computational load increases drastically, leading to ballooning operational expenses.
- **Scalability issues**: As businesses scale, RAG systems struggle to process larger datasets and increased query volumes without compromising on speed or performance.
- **Data redundancy**: Frequently accessing the same knowledge without caching leads to unnecessary repetition, wasting resources and time.

What is the solution?

RAGCache is a solution designed to tackle the inefficiencies of traditional RAG systems. It acts as an intelligent knowledge cache that stores, manages, and optimizes frequently accessed data, significantly enhancing the performance and scalability of RAG systems. Here's what makes it stand out:

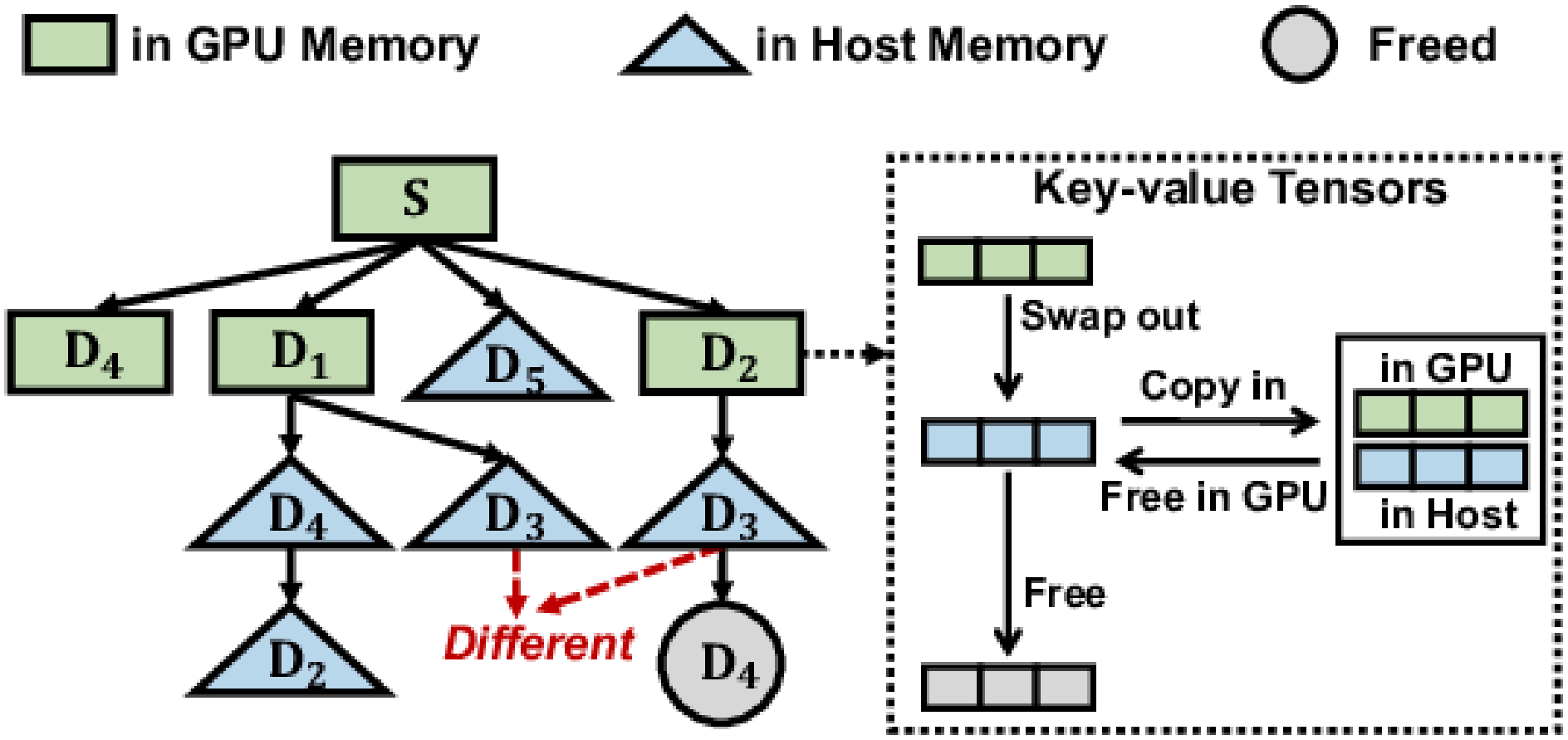


Source: https://arxiv.org/pdf/2404.12457

- **Efficient data storage**: RAGCache intelligently identifies and stores the most frequently retrieved information. By maintaining a cache of relevant knowledge, it eliminates the need to repeatedly query external sources, saving time and resources.
- **Speed booster**: RAGCache enables faster response times by reducing the reliance on external data retrieval. Instead of fetching the same data multiple times, it retrieves it instantly from the cache, drastically improving query performance.
- **Cost saver**:By minimizing redundant queries, RAGCache lowers computational overhead and operational costs.
- **Dynamic and adaptable**: Unlike static caching systems, RAGCache continuously updates its stored knowledge to reflect the most recent information. This dynamic adaptability ensures that the AI system stays accurate and reliable over time.

How does it work?

RAGCache upgrades the traditional RAG workflow by introducing a smart, adaptive caching layer that boosts efficiency and reduces redundancy. Here's how it functions in a streamlined yet sophisticated process:

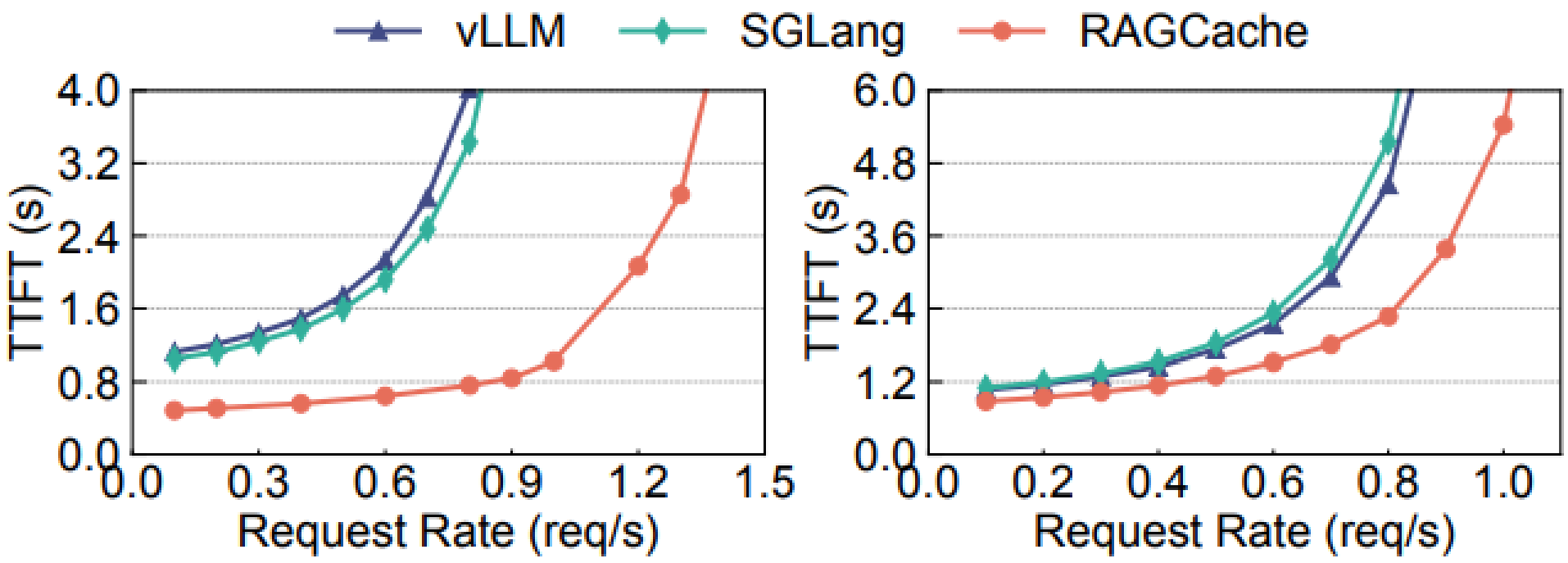


Source: https://arxiv.org/pdf/2404.12457

- **Smart query matching**: Instantly serves answers from the cache if a match exists, skipping external retrieval.
- **Adaptive cache updates**: Learns usage patterns and updates the cache with high-demand knowledge, removing outdated data.
- **Dynamic expiry**: Discards old information to ensure the cache always has accurate, up-to-date knowledge.
- **Usage insights**: Analyzes query trends to prioritize caching of frequently accessed information.
- **Fallback to retrieval**: Fetches data externally if not cached and adds it to the cache for future use.

How well does it perform?

- RAGCache reduces the average time-to-first-token(TTFT) by **1.2–4time-to-first-token** and **1.1–3.5× compared to SGLang** under the same request rate. This is because RAGCache utilizes the GPU memory and host memory to cache the KV cache of hot documents and avoids frequent recomputation.
- Due to faster request processing, RAGCache achieves **1.3– 2.1× higher throughput than vLLM** and **1.2–1.8× higher throughput than SGLang**.



(a) Mistral-7B. **(b)** LLaMA-2-7B.

Overall performance on Natural Questions.

Vector Search Ratio	MMLU		Natural Questions	
	RAGCache	**No DSP**	**RAGCache**	**No DSP**
12.5%	52.1 ms	78.5 ms	67.7 ms	105.8 ms
25%	59.2 ms	135.9 ms	72.9 ms	163.4 ms
50%	69.7 ms	243.7 ms	94.2 ms	282.5 ms
100%	97.4 ms	422.3 ms	145.0 ms	446.1 ms

Average non-overlapping vector search time under different settings.

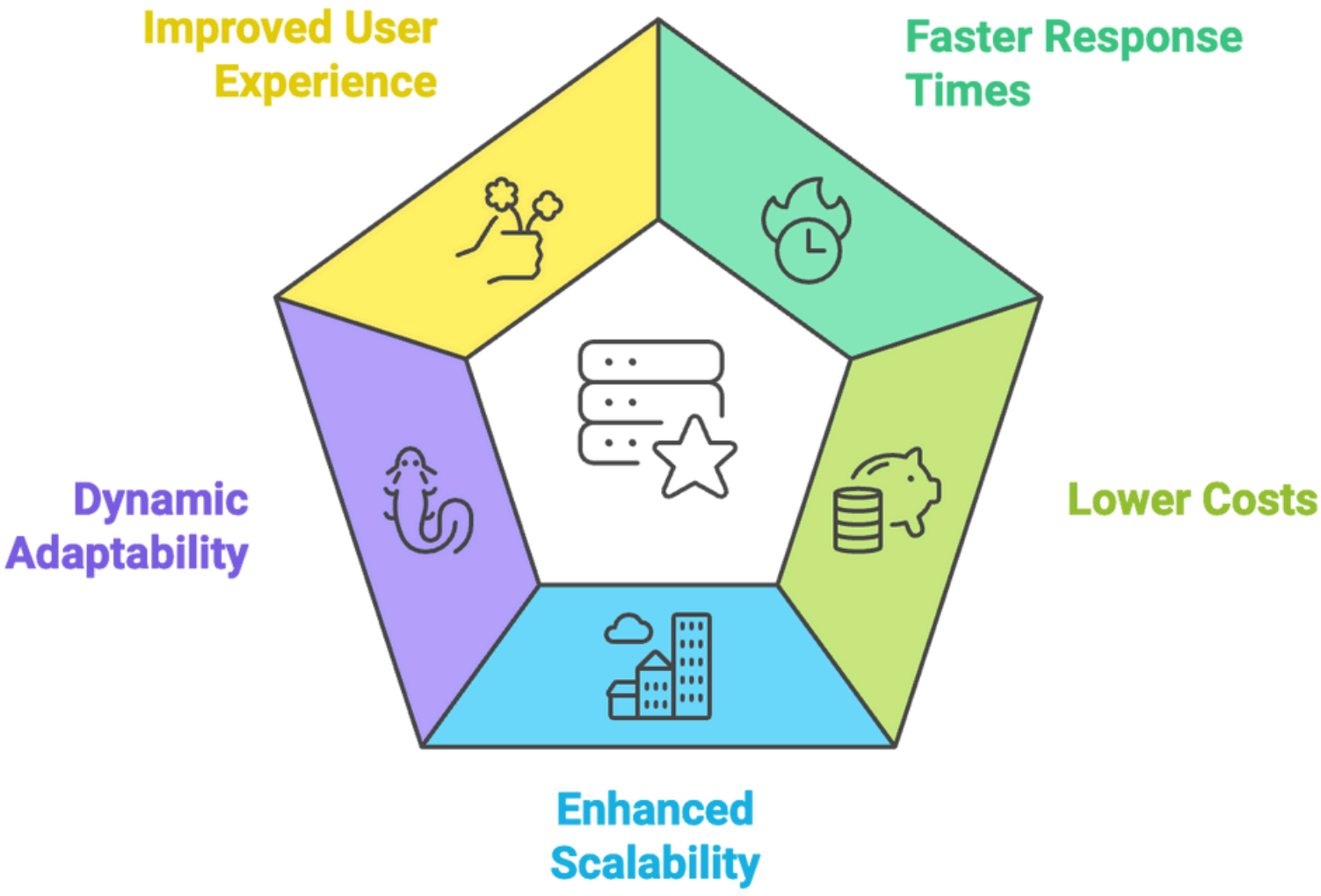
Source: https://arxiv.org/pdf/2404.12457

How it differs from FAQ?

Aspect	RAGCache	FAQs
Data Retrieval	Retrieves relevant data dynamically from cache, optimizing based on usage patterns.	Static answers to pre-defined questions stored manually.
Adaptability	Continuously updates with new information based on queries, keeping the system current.	Updates are manual and require human intervention.
Response Time	Faster responses by serving frequently accessed data directly from cache.	Fixed responses may not always be as fast, especially with large volumes of users.
Scalability	Highly scalable; adapts to increasing data loads and query volumes.	Not as scalable as FAQ systems require manual additions.
Customization	Learns from queries to optimize data retrieval and answer quality.	Limited customization; answers are predetermined and fixed.
Efficiency	Minimizes redundant queries to external systems, saving computational resources.	Requires repeated retrieval of static information, leading to inefficiency.
User Experience	Provides more accurate, real-time answers by dynamically updating the cache.	Offers a limited experience based on static content.
Context Awareness	Can dynamically adjust responses based on context, trends, or previous interactions.	Cannot adapt to context; all responses are fixed.
Maintenance	Self-optimizing system that reduces the need for manual updates.	Requires constant manual updates and management.

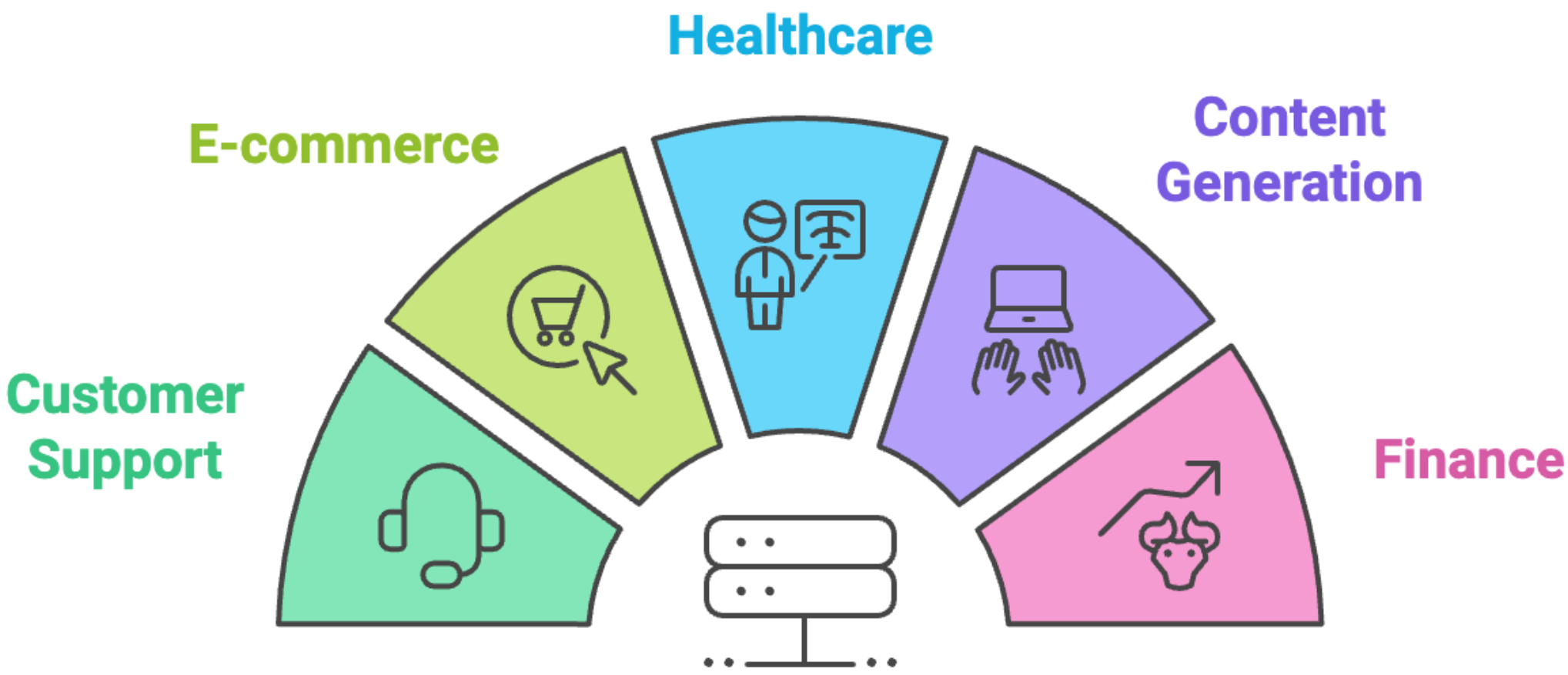
What are the key benefits?

RAGCache introduces transformative advantages to Retrieval-Augmented Generation (RAG) systems, solving major inefficiencies while unlocking new possibilities. Here's a look at its core benefits:



- **Faster response times**: By serving frequently requested data directly from the cache, RAGCache eliminates the delays caused by external data retrieval. This ensures near-instant responses, especially for time-critical applications.
- **Lower costs**: RAGCache reduces computational overhead by avoiding repeated external queries, significantly cutting down on operational expenses while boosting efficiency.
- **Enhanced scalability**: As query volumes and datasets grow, RAGCache maintains optimal performance, ensuring smooth operations even in large-scale deployments.
- **Dynamic adaptability**: RAGCache updates its cache with the latest and most relevant knowledge, ensuring accuracy and relevance without requiring manual intervention.
- **Improved user experience**: Faster, accurate responses lead to better interactions, whether it’s a chatbot, recommendation engine, or content generator, resulting in higher user satisfaction.

Real-world applications



- **Customer support systems**: RAGCache allows chatbots and virtual assistants to respond faster by caching frequently asked questions and common solutions. This reduces wait times, enhances customer satisfaction, and minimizes operational costs.
- **E-commerce platforms**: Personalization engines leverage RAGCache to instantly retrieve user preferences and product recommendations. This ensures smoother shopping experiences and drives higher conversion rates.
- **Healthcare applications**: Medical systems use RAGCache to quickly access patient records, guidelines, or prior diagnoses, enabling faster and more accurate clinical decisions, especially during emergencies.
- **Content generation tools**: AI-driven content platforms utilize RAGCache to store commonly used references or templates, speeding up workflows and improving efficiency for writers and creators.
- **Finance and banking**: RAGCache accelerates financial advisory systems by instantly retrieving market trends, historical data, and client portfolios, allowing for real-time decision-making in high-stakes scenarios like stock trading.

Follow to stay updated on Generative AI

SAVE

LIKE

REPOST

Bhavishya Pandit