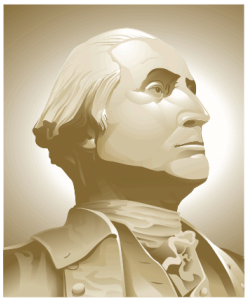

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Data Science Program

Capstone Report - Spring 2022

Optimizing Information Retrieval: A Hybrid Model Leveraging MAR and RAPTOR Frameworks

Abdygulov Timur,
Meet Daxini

Supervised by
Amir Jafari

Abstract

This report addresses the challenges of retrieval-augmented generation (RAG) systems in handling long-context tasks, ambiguous queries, and evidence aggregation, critical in domains like biomedical research. Memory-Augmented Retrieval (MAR) and RAPTOR systems are introduced, with MAR leveraging memory hierarchies for efficient context retention and RAPTOR employing hierarchical clustering for multi-level retrieval. A Hybrid Model combines these strengths, integrating memory-driven retrieval with hierarchical routing for superior performance across diverse queries. Evaluations demonstrate the Hybrid Model's effectiveness, while computational efficiency is enhanced through techniques like quantization and hybrid retrieval layers. Despite progress, challenges in handling specialized queries and optimizing scalability remain, providing avenues for future research.

Contents

1. Introduction

The *exponential growth of unstructured data* has intensified the need for efficient information retrieval systems capable of synthesizing meaningful insights. Traditional retrieval methods often falter in addressing *long-context tasks, resolving ambiguous queries*, or *aggregating evidence* from multiple sources. These limitations hinder their applicability in domains like *biomedical research*, *legal document summarization*, and *real-time question answering*.

Memory-Augmented Retrieval (MAR) draws on principles outlined in the *MemoRAG* paper (1), which introduces memory hierarchies to bridge short- and long-term retrieval needs. By pre-populating memory banks and using *clue-driven retrieval*, MAR enhances precision and context retention while reducing reliance on computationally intensive retrieval processes. RAPTOR, on the other hand, employs a hierarchical approach, organizing data into tree structures that enable *multi-level retrieval* and abstraction. Building on these frameworks, this report also examines a Hybrid Model that combines RAPTOR's *hierarchical retrieval* with MAR's *memory efficiency* and *hybrid search techniques*. These implementations aim to provide accurate, context-aware, and scalable solutions for complex retrieval tasks.

(Need more details): Additional real-world examples or domains where these systems could make a measurable impact would strengthen this section.

2. Problem Statement

The core problem addressed in this report centers on the *inefficiencies of traditional retrieval systems* in handling tasks requiring *extensive context*, *precise query resolution*, and *multi-hop reasoning (2)*. *Token limitations* in language models often hinder their ability to process ultra-long contexts, leading to *incomplete or fragmented responses*. Additionally, *ambiguous queries* pose a significant challenge, as traditional systems lack the capability to refine unclear inputs into actionable queries. Multi-hop reasoning, which involves aggregating evidence across distributed sources, remains a bottleneck for conventional methodologies.

MAR, based on MemoRAG principles (1), introduces *memory hierarchies* that store and recall relevant information dynamically, addressing these limitations. RAPTOR

enhances retrieval by creating hierarchical structures that aggregate evidence across multiple abstraction levels (2). This work explores the adaptation and integration of these frameworks into *practical systems* capable of *addressing these challenges effectively*.

(Need more details): Clarify the specific type of queries or tasks tested (e.g., biomedical, legal, or general-purpose information retrieval) to contextualize the problem.

3. Related Work

Memory-Augmented Retrieval (MAR) takes inspiration from the MemoRAG paper (1), which highlights the use of *memory hierarchies* for efficient retrieval and *clue-driven query resolution (1)*. The MemoRAG paper emphasizes *token compression*, *dynamic memory updates*, and *surrogate query generation* as key strategies for scaling memory-based retrieval systems. MAR integrates several of these principles, adapting them to a modular framework that incorporates fallback mechanisms and *hybrid retrieval techniques*.

RAPTOR, by contrast, employs *recursive clustering* to create *hierarchical tree structures*. These structures allow for *multi-level context aggregation* and efficient retrieval of information aligned with varying levels of abstraction. This methodology, outlined in the RAPTOR paper (2), allows for dynamic query resolution and multi-hop reasoning, which are essential for handling complex, context-dependent tasks. The Hybrid Model presented in this report builds on these frameworks by combining RAPTOR's hierarchical organization with MAR's pre-populated memory bank and hybrid retrieval capabilities. Tools like *LangChain* facilitate modular integration, while *ChromaDB* provides scalable vector storage for embeddings. By leveraging these tools, the implementations aim to address *scalability and precision* challenges in retrieval systems.

(Need more details): Include references to comparable retrieval systems or methods to position these frameworks within a broader context.

4. Solution and Methodology

This section outlines the design and implementation of the MAR-inspired system, the RAPTOR System, and the Hybrid Model. Each system leverages distinct methodologies to address the challenges of retrieval-augmented generation (RAG), culminating in a hybrid approach that integrates their strengths.

4.1 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation serves as the foundation for all three systems. A RAG system combines document retrieval with generative language models, enabling it to answer queries by retrieving relevant context from a document repository and generating *coherent responses*. Figure 1 illustrates the basic architecture of a RAG system, where a query initiates the *retrieval of relevant documents*, which are then processed by a language model to generate the final response.

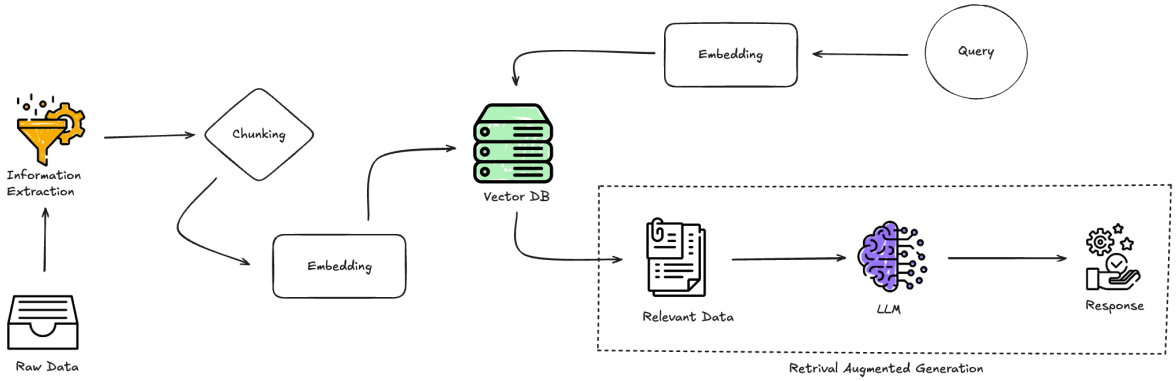

Figure 1: *Basic workflow of a Retrieval-Augmented Generation (RAG) system*.

4.2 Memory-Augmented Retrieval (MAR)

The MAR-inspired system builds upon the principles introduced in the MemoRAG framework (1), which emphasizes *memory hierarchies* to bridge *short- and long-term retrieval needs*. By leveraging a pre-populated memory bank, MAR is designed to handle *frequently recurring queries* efficiently, reducing the reliance on computationally intensive vector store retrieval.

Documents are processed into manageable chunks, embedded using models such as mxbai-embed-large, and indexed in ChromaDB. A *MemoryDB*, pre-populated with

frequently asked questions and answers, prioritizes retrieval for recurring queries, ensuring rapid and accurate response generation. When a query cannot be resolved using MemoryDB, the system falls back to ChromaDB, a scalable vector database, for *semantic similarity-based retrieval*. This hierarchical memory-first approach aligns closely with MemoRAG's concept of clue-driven retrieval and *layered memory hierarchies*.

During query execution, MAR integrates *memory-driven persistence* and *fallback mechanisms* to balance efficiency with retrieval accuracy. Figure 2 illustrates the MAR workflow, highlighting its use of memory banks and fallback mechanisms to streamline query handling.

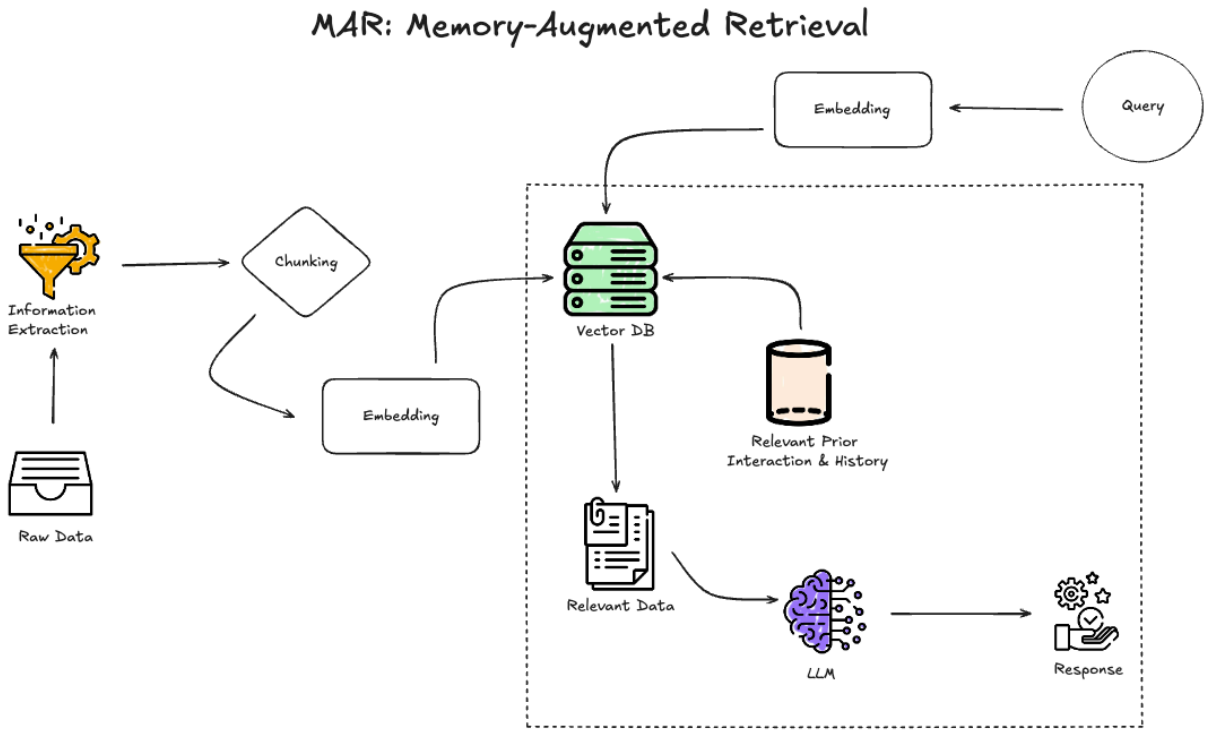

Figure 2: Workflow of the Memory-Augmented Retrieval (MAR) system.

While MAR excels in handling static or recurring queries, its reliance on pre-indexed memory data limits its adaptability to dynamic, multi-layered scenarios. This limitation is addressed in the Hybrid Model by combining MAR's memory persistence

with RAPTOR's *dynamic hierarchical retrieval*, creating a more balanced and versatile system.

4.3 Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR)

The RAPTOR system adopts a *hierarchical approach to data organization*, leveraging *recursive clustering* to create tree structures for dynamic and context-sensitive retrieval. This methodology, inspired by the framework introduced in the RAPTOR paper (2), enables *multi-level context aggregation and abstraction*. By summarizing information at each hierarchical level, the system efficiently handles complex queries that require evidence from multiple layers of context.

In RAPTOR, text chunks are recursively grouped based on *semantic similarity*. Each node in the hierarchy encapsulates a summarized representation of its content, ensuring that higher-level nodes provide abstractions of the information contained in lower-level nodes. This recursive summarization facilitates efficient retrieval, as queries can dynamically traverse the *hierarchical structure* to locate the most contextually relevant nodes.

The RAPTOR system combines semantic retrieval techniques, such as embedding-based similarity, with keyword-based searches using BM25Okapi from the rank_bm25 library (4) to enhance flexibility and precision. BM25Okapi (4) ensures robust keyword-based retrieval, particularly for queries involving rare terms or exact matches, where semantic embeddings might underperform. Retrieved results are further refined using a cross-encoder, which reranks them based on relevance scores to align the final output closely with the user's query intent. This hybrid retrieval strategy enables RAPTOR to address both narrow, specific queries and broader, multi-layered questions effectively.

The system employs quantization techniques to reduce the precision of embeddings for efficient storage. This approach minimizes memory usage and computational costs while preserving a reasonable approximation of the original embedding quality, ensuring effective hierarchical retrieval.

Figure 3 illustrates the hierarchical retrieval architecture of the RAPTOR system, showcasing its recursive summarization and dynamic routing capabilities.

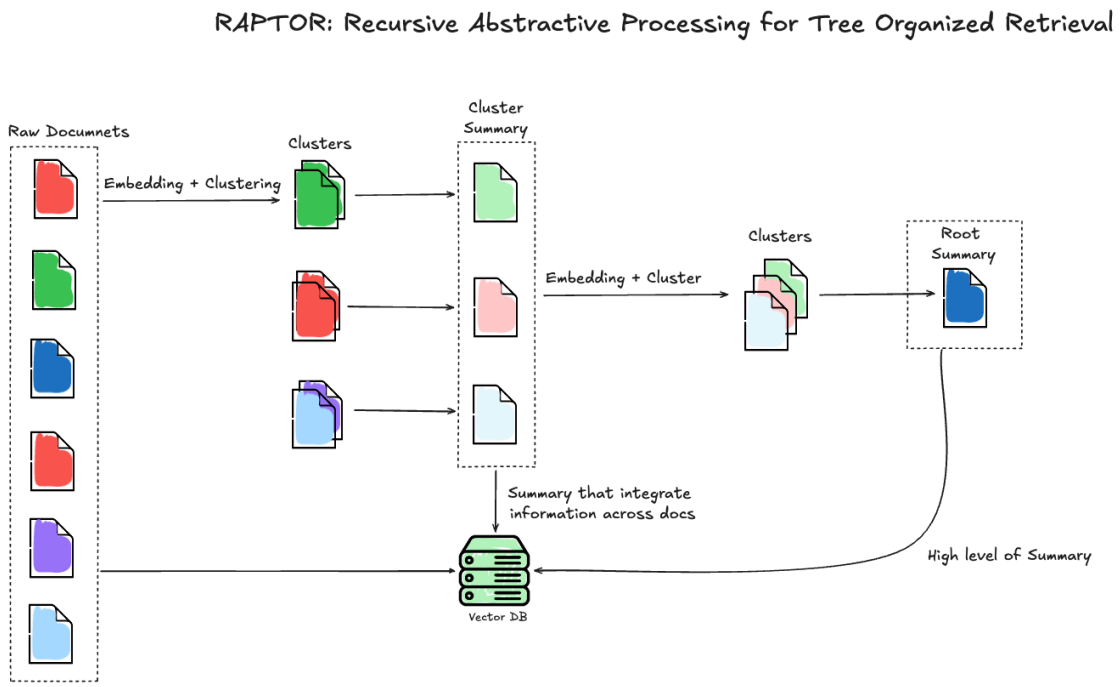

Figure 3: Hierarchical retrieval architecture of the RAPTOR system.

The RAPTOR methodology's reliance on *hierarchical abstraction* and dynamic query resolution makes it particularly suited for handling complex, *context-dependent tasks*. However, its lack of memory persistence can limit its ability to recall infrequently accessed or historical information. This limitation is addressed in the Hybrid Model, which integrates RAPTOR's strengths with MAR's memory-driven persistence for enhanced scalability and adaptability.

4.4 The Hybrid Model

The Hybrid Model integrates the strengths of MAR and RAPTOR to create a more robust retrieval-augmented generation system. It leverages MAR's *memory banks* to provide *rapid responses to frequently asked questions* while incorporating RAPTOR's hierarchical retrieval to handle complex, multi-layered queries. This integration enables the Hybrid Model to address both static and dynamic queries effectively.

The Hybrid Model uses *dynamic thresholding* to optimize query routing through the RAPTOR Tree. It aggregates context from memory banks, vector stores, and hierarchical retrieval layers, ensuring *comprehensive and precise response generation*. By balancing the *memory-driven persistence* of MAR with the dynamic adaptability of RAPTOR, the Hybrid Model achieves *superior scalability and accuracy*. Figure 4 illustrates the architecture of the Hybrid Model, showcasing its integration of MAR and RAPTOR components.

To enhance scalability, the Hybrid Model applies quantization to embeddings, reducing their precision for efficient storage. This process balances computational efficiency and retrieval accuracy, enabling the system to handle large-scale datasets with reduced resource requirements.

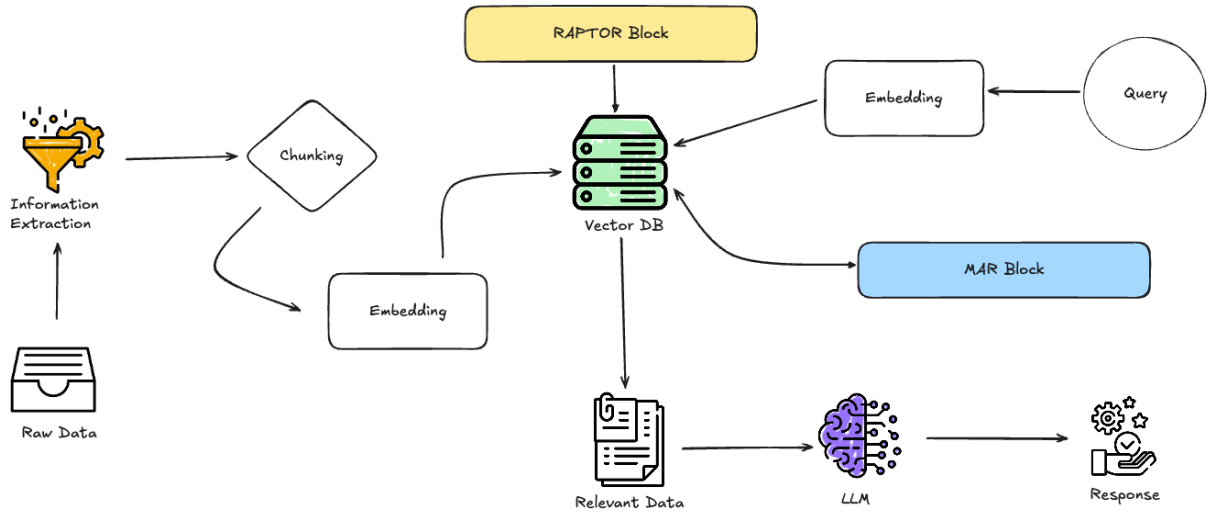

Figure 4: Architecture of the Hybrid Model, combining MAR and RAPTOR systems. (Insert Hybrid Model Diagram Here)

4.5 Integration of Components

The architecture of the Hybrid Model includes MAR's pre-populated *memory banks for rapid responses* to frequently asked questions. This reduces the computational burden on vector stores, which are otherwise required to process recurring queries. Simultaneously, RAPTOR's *hierarchical structure* organizes data into *semantic clusters*,

enabling dynamic *multi-level retrieval*. This allows the system to address more complex queries by routing them through contextually relevant nodes in the RAPTOR Tree.

To achieve seamless integration, the Hybrid Model employs *dynamic thresholding* to optimize query routing. During execution, a query is first directed to the MemoryDB for retrieval. If the memory lacks relevant data, the query is passed to RAPTOR's *hierarchical structure*, where semantic embeddings guide its traversal across different abstraction levels. This *layered approach* ensures comprehensive and precise response generation.

BM25Okapi is employed alongside semantic embeddings to form the hybrid retrieval layer. Its inclusion enhances the system's ability to handle queries requiring precise keyword matches, ensuring relevance when semantic models fall short for rare terms or domain-specific jargon. This dual-layered approach strengthens both the MAR and RAPTOR components of the Hybrid Model.

The integration of MAR and RAPTOR components into the Hybrid Model, as illustrated in Figure 4 (see 4.4 The Hybrid Model), demonstrates how memory-driven persistence and hierarchical retrieval are combined for *comprehensive query handling*.

5. Results and Discussion

This section evaluates the performance of MAR, RAPTOR, and the Hybrid Model across a diverse set of biomedical queries, as well as the influence of embedding models on retrieval accuracy and efficiency. The findings highlight both the strengths and limitations of each system while shedding light on the pivotal role of embeddings in optimizing retrieval-augmented generation (RAG) systems.

The biomedical data used in this study was derived from the BioASQ dataset (3), which contains essential features such as the *ideal answer, question, and links to related articles* from which PDFs were downloaded for processing.

5.1 System Performance Across Queries

The evaluation of MAR, RAPTOR, and the Hybrid Model used a range of biomedical queries, encompassing static, frequently recurring questions and dynamic, multi-layered scenarios. MAR demonstrated exceptional effectiveness in leveraging its

memory-driven persistence, enabling it to handle recurring static questions such as identifying EGFR ligands and acetylcholinesterase inhibitors for myasthenia gravis treatment. However, its reliance on pre-indexed memory limited its ability to adapt to queries requiring dynamic reasoning or contextual updates, such as determining the FDA approval status for Denosumab (Prolia).

RAPTOR excelled in addressing dynamic queries with its hierarchical structure and multi-level reasoning capabilities. It performed particularly well for complex questions requiring multi-hop reasoning, such as identifying the location of the protein Pannexin1. Despite this, RAPTOR's lack of memory persistence restricted its ability to recall less frequently accessed or historical information, leading to incomplete responses for queries like the classification of Hirschsprung disease.

The Hybrid Model consistently outperformed both MAR and RAPTOR by integrating their strengths. It provided accurate answers for both static and dynamic queries, excelling in multi-layered scenarios such as determining miRNAs for epithelial ovarian cancer and classifying Hirschsprung disease. Nevertheless, even the Hybrid Model struggled with highly specialized queries, such as identifying mitochondrial diseases linked to POLG mutations. This underscores the need for more advanced clustering techniques and finer-grained contextual reasoning.

The evaluation results, summarized in Table 1, were generated by comparing system responses to ideal answers using OpenAI's ChatGPT-4 (6) for biomedical queries. This ensured a standardized and consistent assessment of the systems' retrieval capabilities across various scenarios.

Question	Ideal Answer	MAR Answer	RAG Answer	RAPTOR Answer	Hybrid Answer
Is Hirschsprung disease a Mendelian or a multifactorial disorder?	Hirschsprung disease is both Mendelian and multifactorial, depending on the context.	✅ Matches ideal answer	✅ Matches ideal answer	❌ Incorrect	✅ Matches ideal answer
List signaling molecules (ligands) that interact with the receptor EGFR?	The 7 EGFR ligands are EGF, BTC, EPR, HB-EGF, TGF-α, AREG, and EPG.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Are long non-coding RNAs spliced?	Yes, long non-coding RNAs are spliced through the same pathway as mRNAs.	✅ Matches ideal answer	✅ Matches ideal answer	❌ Unclear or incomplete answer	❌ Unclear or incomplete answer
Is RANKL secreted from the cells?	Yes, RANKL is secreted by osteoblasts.	✅ Matches ideal answer	❌ Does not mention secretion	❌ Incomplete answer	✅ Matches ideal answer

Which miRNAs could be used as potential biomarkers for epithelial ovarian cancer?	miR-200a, miR-100, miR-141, miR-200b, miR-200c, miR-203, etc.	✅ Matches ideal answer	❌ Partial match	❌ Partial match	✅ Matches ideal answer
Which acetylcholinesterase inhibitors are used for treatment of myasthenia gravis?	Pyridostigmine and neostigmine.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Has Denosumab (Prolia) been approved by FDA?	Yes, approved by the FDA in 2010.	❌ Not answered	❌ Not answered	✅ Matches ideal answer	❌ Not answered
Which are the different isoforms of the mammalian Notch receptor?	Notch-1, Notch-2, Notch-3, Notch-4.	❌ Not answered	❌ Not answered	❌ Not answered	❌ Not answered
Orteronel was developed for treatment of which cancer?	Castration-resistant prostate cancer.	✅ Matches ideal answer	❌ Not answered	❌ Incorrect	✅ Matches ideal answer
Is the monoclonal antibody Trastuzumab (Herceptin) of potential use in the treatment of prostate cancer?	Controversial, but it can be used in HER2 overexpressing prostate cancer.	✅ Matches ideal answer	✅ Matches ideal answer	❌ Incorrect	❌ Not answered
Which are the Yamanaka factors?	OCT4, SOX2, MYC, and KLF4 transcription factors.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Where is the protein Pannexin1 located?	Localized to the plasma membranes.	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer	✅ Matches ideal answer
Which currently known mitochondrial diseases have been attributed to POLG mutations?	Recessive PEO and MNGIE.	❌ Partial match	❌ Partial match	❌ Partial match	❌ Partial match

Table 1: Evaluation Performance of Retrieval MAR, RAPTOR, and Hybrid Systems

5.2 Embedding Model Evaluation

The embedding models significantly influenced the retrieval performance of MAR, RAPTOR, and the Hybrid Model. Among these, dunzhang/stella_en_1.5B_v5 (8) served as the primary embedding model for evaluations due to its high retrieval accuracy. However, additional models were analyzed to understand their performance across diverse datasets, including the HuggingFace QA dataset (5) and the PubMed (3) filtered dataset. These evaluations explored the models' accuracy at varying retrieval depths (top-k levels) and computational efficiency, offering insights into their strengths and trade-offs.

Instruction-based embeddings like dunzhang/stella_en_1.5B_v5 (8) and nvidia/NV-Embed-v2 (7) demonstrated exceptional performance on domain-specific tasks, particularly in the PubMed filtered dataset. These models consistently achieved near-perfect accuracy for both Top 1 and Top 10 retrievals, showcasing their robustness for complex biomedical queries. For instance, nvidia/NV-Embed-v2 (7) achieved a Top 1 accuracy of 89.23% and a Top 10 accuracy of 100% on the HuggingFace QA dataset, illustrating its precision in handling intricate contexts. However, such embeddings come

with considerable computational costs. dunzhang/stella_en_1.5B_v5 (8), loaded in 8-bit precision, required 3.48 GB of GPU memory, whereas nvidia/NV-Embed-v2 used 7.44 GB under similar conditions.

Lighter models, such as sentence-transformers/all-MiniLM-L6-v2 (9), provided a practical balance between performance and efficiency. While its Top 1 accuracy was lower (56.92% on HuggingFace QA), it achieved competitive Top 10 accuracy (83.08%), making it a viable option for broader contextual queries in resource-constrained environments. With only 0.08 GB of GPU memory usage, this lightweight model offers an excellent trade-off for high-throughput applications.

The accuracy performance of embedding models across Top 1 and Top 10 retrievals is summarized in Table 2, which provides insights into their relative strengths.

Embedding Model	Top 1 Accuracy (%)	Top 10 Accuracy (%)
bert-large-nli	75.0	87.5
instructor-xl	87.5	87.5
msmacro	87.5	87.5
mxbai-embed-large	87.5	87.5
roberta-base	37.5	87.5
roberta-large	62.5	62.5

Table 2: Top k Accuracy of Embedding Models using PubMed dataset

Expanding the evaluation, the models were tested for their accuracy across datasets and deeper retrieval depths. Table 3 highlights the accuracy of these models, emphasizing their performance at different levels of retrieval precision.

Model	Dataset	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10
nvidia/NV-Embed-v2	Hugging Face QA Dataset	0.89	0.92	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00
dunzhang/stella_en_1.5B_v5	Hugging Face QA Dataset	0.95	0.97	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00

sentence-transformers/all-MiniLM-L6-v2	Hugging Face QA Dataset	0.57	0.65	0.72	0.74	0.77	0.80	0.80	0.82	0.83	0.83
mixedbread-ai/mxbai-embed-large-v1	Hugging Face QA Dataset	0.95	0.95	0.95	0.97	0.97	0.98	1.00	1.00	1.00	1.00
amazon.titan-embed-text-v2:0	Hugging Face QA Dataset	0.89	0.91	0.95	0.97	0.98	0.98	0.98	0.98	0.98	0.98
nvidia/NV-Embed-v2	PubMed filtered Dataset	1.00	0.88	0.77	0.78	0.79	0.76	0.76	0.77	0.79	0.80
dunzhang/stella_en_1.5B_v5	PubMed filtered Dataset	1.00	0.81	0.77	0.74	0.74	0.74	0.76	0.76	0.80	0.80
sentence-transformers/all-MiniLM-L6-v2	PubMed filtered Dataset	0.69	0.65	0.59	0.57	0.67	0.65	0.65	0.70	0.71	0.71
mixedbread-ai/mxbai-embed-large-v1	PubMed filtered Dataset	1.00	0.92	0.77	0.70	0.74	0.77	0.77	0.79	0.79	0.79
amazon.titan-embed-text-v2:0	PubMed filtered Dataset	1.00	0.81	0.69	0.67	0.69	0.67	0.67	0.68	0.70	0.73

Table 3: Evaluation of Embedding Models Across Datasets

The resource utilization of embedding models, summarized in Table 4, highlights the computational costs associated with different embeddings. This table emphasizes the trade-offs between resource efficiency and performance, offering insights for selecting suitable models based on system requirements.

Model Name	GPU Memory Used (GB)	Total Parameters (Millions)	Loaded in 8-bit
mixedbread-ai/mxbai-embed-large-v1	1.25	335.14	No
dunzhang/stella_en_1.5B_v5	3.48	1543.27	Yes
dunzhang/stella_en_1.5B_v5	9.25	1543.27	No
nvidia/NV-Embed-v2	7.44	7851.02	Yes
sentence-transformers/all-MiniLM-L6-v2	0.08	22.71	No
meta-llama/Meta-Llama-3-8B-Instruct	10.42	8030.26	Yes

Table 4: Resource Utilization of Embedding Models

The inclusion of resource utilization metrics underscores the importance of *balancing accuracy with computational efficiency*. Lightweight models such

as sentence-transformers/all-MiniLM-L6-v2 (9) are optimal for high-throughput or resource-constrained scenarios, whereas instruction-based embeddings like nvidia/NV-Embed-v2 (7) and dunzhang/stella_en_1.5B_v5 (8) excel in domain-specific tasks requiring precise contextual understanding.

5.3 Insights on Embedding Models

The evaluation of embedding models revealed critical trade-offs between retrieval accuracy and computational efficiency. Models like dunzhang/stella_en_1.5B_v5 (8) and nvidia/NV-Embed-v2 (7) excelled in high-accuracy retrieval tasks, particularly for specialized biomedical queries. However, their substantial memory usage and GPU requirements highlight the challenges of scaling these models in resource-constrained environments.

Conversely, lightweight models such as sentence-transformers/all-MiniLM-L6-v2 (9) demonstrated competitive performance for broader contextual queries. These models maintained lower computational costs, offering a viable solution for high-throughput applications. For instance, Table 4 shows that sentence-transformers/all-MiniLM-L6-v2 (9) required only 0.08 GB of GPU memory, significantly less than the 9.25 GB used by dunzhang/stella_en_1.5B_v5 (8) in its 8-bit configuration.

The Top 1 and Top 10 accuracy results, summarized in Table 3, further emphasize the practical trade-offs. While heavier models like nvidia/NV-Embed-v2 (7) achieved near-perfect Top 10 accuracy, lightweight embeddings offered sufficient performance for less complex tasks with reduced resource demands.

These findings underscore the importance of aligning embedding model selection with system requirements. Dynamic embedding strategies, where models are selected based on query complexity or resource availability, could optimize performance across a variety of retrieval scenarios. Such strategies would enable the Hybrid Model to balance precision and computational efficiency effectively.

5.4 Performance Insights Across Systems and Embeddings

The performance of MAR, RAPTOR, and the Hybrid Model was closely tied to the choice of embeddings. The resource utilization data in Table 4 highlights the practical considerations when deploying these systems at scale. For example, the Hybrid

Model's reliance on dunzhang/stella_en_1.5B_v5 (8) contributed significantly to its superior accuracy but also imposed higher computational costs.

Lightweight embeddings like sentence-transformers/all-MiniLM-L6-v2 (9) presented an efficient alternative, particularly for applications prioritizing speed and scalability. Their lower memory usage and latency make them suitable for high-throughput systems, though they may compromise precision for complex, domain-specific queries.

The results from Table 3 suggest that instruction-based embeddings consistently outperformed simpler models in both Top 1 and Top 10 accuracy. This indicates their suitability for tasks requiring detailed contextual understanding. However, as resource constraints are a critical factor in real-world deployments, future iterations of these systems should explore dynamic embedding strategies. By integrating resource-efficient models for simpler queries and high-performance embeddings for complex tasks, the systems could achieve a more balanced performance profile.

6. Discussion

The results revealed that each system addressed unique aspects of retrieval challenges, but the Hybrid Model provided the most balanced solution. MAR's ability to recall frequently accessed information proved invaluable for recurring and static queries, such as identifying signaling molecules or transcription factors. However, its dependence on pre-indexed memory limited its ability to handle dynamic, multi-layered queries, reducing its adaptability.

RAPTOR excelled in dynamic query handling and multi-hop reasoning due to its hierarchical structure. Its recursive abstraction capabilities enabled it to aggregate evidence across multiple levels of context, making it highly effective for complex reasoning tasks. However, the lack of memory persistence hindered its ability to recall historical or infrequent data, which affected its performance in questions requiring both dynamic routing and long-term recall.

The Hybrid Model successfully combined the strengths of MAR and RAPTOR, achieving superior performance in addressing multi-layered and context-sensitive queries. Its effectiveness was particularly evident in scenarios requiring both memory persistence and hierarchical retrieval. However, the results also revealed several areas

for improvement. The systems struggled with ambiguous or highly specialized queries, such as identifying mitochondrial diseases linked to POLG mutations. Incorporating advanced clustering techniques like UMAP or GMM could optimize hierarchical structures and improve performance for such cases.

The quantization process employed in RAPTOR and the Hybrid Model reduced memory usage and computational costs, as highlighted in Table 4, but it introduced *semantic approximation errors*. This trade-off indicates the need for adaptive quantization methods to balance efficiency and precision. For instance, dynamic quantization could be explored to adjust embedding precision based on query complexity, minimizing the loss of contextual relevance.

The inclusion of BM25Okapi strengthened the system's ability to handle scenarios where semantic embeddings might underperform, such as queries involving rare or domain-specific terms. This keyword-based approach complemented semantic models effectively for certain tasks but lacked the contextual depth and nuanced understanding that embeddings provide. Future enhancements could include dynamic weighting mechanisms to balance the contributions of semantic and keyword-based retrieval layers, optimizing performance across diverse query types.

Additionally, dynamic query refinement techniques should be explored to better address ambiguous and multi-hop queries, further enhancing the system's adaptability and accuracy. Expanding the Hybrid Model to support multimodal retrieval, including images and videos, could significantly broaden its versatility and applicability across diverse domains such as medical imaging and legal documentation.

7. Conclusion

This research highlights the effectiveness of integrating MAR and RAPTOR into a Hybrid Model to address challenges in retrieval-augmented generation. By combining MAR’s memory-driven retrieval with RAPTOR’s hierarchical routing, the Hybrid Model consistently demonstrated *superior performance across a diverse set* of biomedical queries. It provided accurate, context-aware responses for both recurring and multi-layered questions, excelling in scenarios such as identifying EGFR ligands, miRNAs as biomarkers, and classifying Hirschsprung disease.

Despite these achievements, several limitations persist. Computational requirements and memory overhead posed scalability challenges, as detailed in Table 4, while the absence of advanced clustering techniques restricted the system's ability to optimize hierarchical structures. Questions requiring fine-grained contextual understanding, such as those involving mitochondrial diseases linked to POLG mutations, exposed gaps in retrieval precision.

The choice of embeddings played a critical role in system performance. Models like dunzhang/stella_en_1.5B_v5 (8) and nvidia/NV-Embed-v2 (7) provided high accuracy but required substantial GPU memory and computational resources. Conversely, lightweight embeddings such as sentence-transformers/all-MiniLM-L6-v2 (9) offered greater computational efficiency, making them suitable for resource-constrained environments, though their accuracy was lower for complex queries.

Future enhancements should focus on incorporating *advanced clustering techniques like UMAP or GMM* to optimize hierarchical structures. Dynamic query refinement and token compression techniques could address scalability challenges while improving query resolution. Exploring hybrid embedding strategies that dynamically allocate resources based on query complexity could further optimize performance. For example, lightweight models could be used for simple queries, while heavier models like nvidia/NV-Embed-v2 (7) could handle complex scenarios requiring high semantic understanding.

Expanding the Hybrid Model to include multimodal retrieval capabilities would significantly enhance its applicability in domains like medical imaging and legal document processing. By integrating advanced retrieval strategies, adaptive quantization, and dynamic embedding approaches, the Hybrid Model can evolve into a robust, scalable, and versatile system, offering substantial benefits for real-world applications requiring context-aware and scalable information retrieval.

References

(1) Qian, H., Zhang, P., Liu, Z., Mao, K., & Dou, Z. (2023). MemoRAG: Moving Towards Next-Gen RAG via Memory-Inspired Knowledge Discovery. Retrieved from https://arxiv.org/abs/2401.18059

(2) Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024). RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. *arXiv preprint arXiv:2401.18059*. https://doi.org/10.48550/arXiv.2401.18059

(3) BioASQ Consortium. (2023). BioASQ11 Dataset for Biomedical Question Answering (Training 11b). Retrieved from http://participants-area.bioasq.org/datasets/

(4) Van Gysel, C., Kanoulas, E., & de Rijke, M. (2018). *rank_bm25: BM25 ranking algorithm for information retrieval*. Retrieved November 30, 2024, from https://github.com/dorianbrown/rank_bm25

(5) Möller, R. (2023). *Hugging Face document QA evaluation dataset* [Data set]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/datasets/m-ric/huggingface_doc_qa_eval

(6) OpenAI. (2024). *ChatGPT-4.0* [Large language model]. Retrieved from https://openai.com

(7) NVIDIA. (2023). *NV-Embed-v2: Text embeddings for natural language understanding* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/nvidia/NV-Embed-v2

(8) Dunzhang. (2023). *Stella_en_1.5B_v5: A large-scale language model for text generation* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/dunzhang/stella_en_1.5B_v5

(9) Reimers, N., & Gurevych, I. (2023). *all-MiniLM-L6-v2: Sentence-transformers for embedding generation* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

(10)Mixedbread-AI. (2023). *MXBAI-Embed-Large-v1: Advanced embeddings for text representation* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

(11)Amazon. (2023). *Titan-Embed-Text-v2:0: Embeddings for scalable text applications* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/amazon/titan-embed-text-v2:0

(12)Google AI. (2023). *BERT-Large-NLI: A large-scale pre-trained model for natural language inference* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/bert-large-nli

(13)OpenAI. (2023). *Instructor-XL: Large-scale instructional language model*. Hugging Face [Model]. Retrieved November 30, 2024, from https://huggingface.co/instructor-xl

(14)Microsoft. (2023). *MSMARCO: A dataset and model for machine reading comprehension* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/msmacro

(15)Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2023). *RoBERTa-Base: A robustly optimized BERT pretraining approach* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/roberta-base

(16)Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2023). *RoBERTa-Large: A robustly optimized BERT pretraining approach (large version)* [Model]. Hugging Face. Retrieved November 30, 2024, from https://huggingface.co/roberta-large

(17)Suzuki, K., Ozono, S., Yamaguchi, A., Koike, H., Matsui, H., Nagata, M., Takubo, T., Miyashita, K., Matsushima, T., & Akaza, H. (2015). A phase 1 multiple-dose study of orteronel in Japanese patients with castration-resistant prostate cancer. *Cancer Chemotherapy and Pharmacology, 75*(3), 373–380. https://doi.org/10.1007/s00280-014-2654-y

(18)Liu, X., Huang, J., Chen, T., Wang, Y., Xin, S., Li, J., Pei, G., & Kang, J. (2008). Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Research, 18*(12), 1177–1189. https://doi.org/10.1038/cr.2008.309

(19)Liu, Y., Cheng, H., Gao, S., Lu, X., He, F., Hu, L., Hou, D., Zou, Z., Li, Y., Zhang, H., Xu, J., Kang, L., Wang, Q., Yuan, W., Gao, S., & Cheng, T. (2014). Reprogramming of MLL-AF9 leukemia cells into pluripotent stem cells. *Leukemia, 28*(5), 1071–1080. https://doi.org/10.1038/leu.2013.304

(20)Legrier, M.-E., Oudard, S., Judde, J.-G., Guyader, C., de Pinieux, G., Boyé, K., de Cremoux, P., Dutrillaux, B., & Poupon, M.-F. (2007). Potentiation of antitumour activity of docetaxel by combination with trastuzumab in a human prostate cancer xenograft model and underlying mechanisms. *British Journal of Cancer, 96*(2), 269–276. https://doi.org/10.1038/sj.bjc.6603553

(21)Van Goethem, G., Schwartz, M., Löfgren, A., Dermaut, B., Van Broeckhoven, C., & Vissing, J. (2003). Novel POLG mutations in progressive external ophthalmoplegia mimicking mitochondrial neurogastrointestinal encephalomyopathy. *European Journal of Human Genetics, 11*(7), 547–549. https://doi.org/10.1038/sj.ejhg.5201002

(22)Jackson, D. G., Wang, J., Keane, R. W., Scemes, E., & Dahl, G. (2014). ATP and potassium ions: A deadly combination for astrocytes. *Scientific Reports, 4*, Article 4576. https://doi.org/10.1038/srep04576

(23)Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan,

M., ... Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research, 22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

(24)Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., & Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research, 22*(9), 1616–1625. https://doi.org/10.1101/gr.134445.111

(25)Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., & Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research, 22*(9), 1616–1625. https://doi.org/10.1101/gr.134445.112

(26)Goldstein, D., Gofrit, O., Nyska, A., & Benita, S. (2007). Anti-HER2 Cationic Immunoemulsion as a Potential Targeted Drug Delivery System for the Treatment of Prostate Cancer. *Cancer Research, 67*(1), 269. https://doi.org/10.1158/0008-5472.CAN-06-2750

(27)Yang, N., Kaur, S., Volinia, S., Greshock, J., Lassus, H., Hasegawa, K., Liang, S., Leminen, A., Deng, S., Smith, L., Johnstone, C. N., Chen, X.-M., Liu, C.-G., Huang, Q., Katsaros, D., Calin, G. A., Weber, B. L., Bützow, R., Croce, C. M., Coukos, G., & Zhang, L. (2008). MicroRNA Microarray Identifies Let-7i as a Novel Biomarker and Therapeutic Target in Human Epithelial Ovarian Cancer. *Cancer Research, 68*(24), 10307. https://doi.org/10.1158/0008-5472.CAN-08-1956

(28)Chen, L., Mooso, B. A., Jathal, M. K., Madhav, A., Johnson, S. D., van Spyk, E., Mikhailova, M., Zierenberg-Ripoll, A., Xue, L., Vinall, R. L., deVere White, R. W., & Ghosh, P. M. (2011). Dual EGFR/HER2 inhibition sensitizes prostate cancer cells to androgen withdrawal by suppressing ErbB3. *Clinical Cancer Research, 17*(19), 6218–6228. https://doi.org/10.1158/1078-0432.CCR-11-1548

(29)Vermi, W., Facchetti, F., & colleagues. (2013). Ligand-dependent activation of EGFR in follicular dendritic cells sarcoma is sustained by local production of cognate ligands. *Clinical Cancer Research, 19*(##), pages. https://doi.org/10.1158/1078-0432.CCR-13-1275

(30)Dreicer, R., MacLean, D., Suri, A., Stadler, W. M., Shevrin, D., Hart, L., MacVicar, G. R., Hamid, O., Hainsworth, J., Gross, M. E., Shi, Y., Webb, I. J., & Agus, D. B. (2014). Phase I/II trial of Orteronel (TAK-700)—an investigational 17,20-lyase inhibitor—in patients with metastatic castration-resistant prostate cancer. *Clinical Cancer Research, 20*(##), pages. https://doi.org/10.1158/1078-0432.CCR-13-2436

(31)Lin, Y., Cheng, Z., Yang, Z., Zheng, J., & Lin, T. (2012). DNp73 improves generation efficiency of human induced pluripotent stem cells. *BMC Cell Biology, 13*(9). https://doi.org/10.1186/1471-2121-13-9

(32)Núñez-Torres, R., Fernández, R. M., Acosta, M. J., Enguix-Riego, M. del V., Marbá, M., de Agustín, J. C., Castaño, L., Antiñolo, G., & Borrego, S. (2011). Comprehensive analysis of RET common and rare variants in a series of Spanish Hirschsprung patients confirms a synergistic effect of both kinds of events. *BMC Medical Genetics, 12*(138). https://doi.org/10.1186/1471-2350-12-138

(33)Kan, C. W. S., Hahn, M. A., Gard, G. B., Maidens, J., Huh, J. Y., Marsh, D. J., & Howell, V. M. (2012). Elevated levels of circulating microRNA-200 family members correlate with serous epithelial ovarian cancer. *BMC Cancer, 12*(627). https://doi.org/10.1186/1471-2407-12-627

(34)Chen, Y., Zhang, L., & Hao, Q. (2013). Candidate microRNA biomarkers in human epithelial ovarian cancer: Systematic review profiling studies and experimental validation. *Cancer Cell International, 13*(86). https://doi.org/10.1186/1475-2867-13-86

(35)Lane, D., Matte, I., Laplante, C., Garde-Granger, P., Rancourt, C., & Piché, A. (2013). Osteoprotegerin (OPG) activates integrin, focal adhesion kinase (FAK), and

Akt signaling in ovarian cancer cells to attenuate TRAIL-induced apoptosis. *Journal of Ovarian Research, 6*(82). https://doi.org/10.1186/1757-2215-6-82

(36)Ohta, S., Imaizumi, Y., Okada, Y., Akamatsu, W., Kuwahara, R., Ohyama, M., Amagai, M., Matsuzaki, Y., Yamanaka, S., Okano, H., & Kawakami, Y. (2011). Generation of human melanocytes from induced pluripotent stem cells. *PLoS ONE, 6*(1), e16182. https://doi.org/10.1371/journal.pone.0016182

(37)Meng, H., Zhang, X., Yu, G., Lee, S. J., Chen, Y. E., Prudovsky, I., & Wang, M. M. (2012). Biochemical characterization and cellular effects of CADASIL mutants of NOTCH3. *PLoS ONE, 7*(9), e44964. https://doi.org/10.1371/journal.pone.0044964

(38)Shah, S. N., Kerr, C., Cope, L., Zambidis, E., Liu, C., Hillion, J., Belton, A., Huso, D. L., & Resar, L. M. S. (2012). HMGA1 reprograms somatic cells into pluripotent stem cells by inducing stem cell transcriptional networks. *PLoS ONE, 7*(11), e48533. https://doi.org/10.1371/journal.pone.0048533

(39)Sanders, J. M., Wampole, M. E., Thakur, M. L., & Wickstrom, E. (2013). Molecular determinants of epidermal growth factor binding: A molecular dynamics study. *PLoS ONE, 8*(1), e54136. https://doi.org/10.1371/journal.pone.0054136

(40)Su, R.-J., Baylink, D. J., Neises, A., Kiroyan, J. B., Meng, X., Payne, K. J., Tschudy-Seney, B., Duan, Y., Appleby, N., Kearns-Jonker, M., Gridley, D. S., Wang, J., Lau, K.-H. W., & Zhang, X.-B. (2013). Efficient generation of integration-free iPS cells from human adult peripheral blood using BCL-XL together with Yamanaka factors. *PLoS ONE, 8*(5), e64496. https://doi.org/10.1371/journal.pone.0064496

(41)Mukhopadhyay, C., Zhao, X., Maroni, D., Band, V., & Naramura, M. (2013). Distinct effects of EGFR ligands on human mammary epithelial cell differentiation. *PLoS ONE, 8*(10), e75907. https://doi.org/10.1371/journal.pone.0075907

(42)Koskensalo, S., Louhimo, J., Hagström, J., Lundin, M., Stenman, U.-H., & Haglund, C. (2013). Concomitant tumor expression of EGFR and TATI/SPINK1

associates with better prognosis in colorectal cancer. *PLoS ONE, 8*(10), e76906. https://doi.org/10.1371/journal.pone.0076906

(43)Kurtenbach, S., Prochnow, N., Kurtenbach, S., Klooster, J., Zoidl, C., Dermietzel, R., Kamermans, M., & Zoidl, G. (2013). Pannexin1 channel proteins in the zebrafish retina have shared and unique properties. *PLoS ONE, 8*(10), e77722. https://doi.org/10.1371/journal.pone.0077722

(44)Sathasivam, S. (2011). Current and emerging treatments for the management of myasthenia gravis. *Therapeutics and Clinical Risk Management, 7*, 313–323. https://doi.org/10.2147/TCRM.S14015

(45)Felici, A., Pino, M. S., & Carlini, P. (2012). A changing landscape in castration-resistant prostate cancer treatment. *Frontiers in Oncology, 3*, Article 85. https://doi.org/10.3389/fonc.2012.00085

(46)Frontiers in Cellular Neuroscience. (2013). Identification and function of long non-coding RNA. *Frontiers in Cellular Neuroscience, 7*, Article 168. https://doi.org/10.3389/fncel.2013.00168

(47)Boassa, D., Nguyen, P., Hu, J., Ellisman, M. H., & Sosinsky, G. E. (2015). Pannexin2 oligomers localize in the membranes of endosomal vesicles in mammalian cells while Pannexin1 channels traffic to the plasma membrane. *Frontiers in Cellular Neuroscience, 8*, Article 468. https://doi.org/10.3389/fncel.2014.00468

(48)Cea, L. A., Riquelme, M. A., Vargas, A. A., Urrutia, C., & Sáez, J. C. (2014). Pannexin 1 channels in skeletal muscles. *Frontiers in Cellular Neuroscience, 5*, Article 139. https://doi.org/10.3389/fncel.2014.00139

(49)Malmberg, J., Tolmachev, V., & Orlov, A. (2011). Imaging agents for in vivo molecular profiling of disseminated prostate cancer: Cellular processing of [111In]-labeled CHX-A''DTPA-trastuzumab and anti-HER2 ABY-025 Affibody in prostate cancer cell lines. *Experimental and Therapeutic Medicine, 2*(2), 287–293. https://doi.org/10.3892/etm.2011.217

(50)Tan, Z., Chen, P., Schneider, N., Glover, S., Cui, L., Torgue, J., Rixe, O., Spitz, H. B., & Dong, Z. (2012). Significant systemic therapeutic effects of high-LET immunoradiation by 212Pb-trastuzumab against prostatic tumors of androgen-independent human prostate cancer in mice. *International Journal of Oncology, 40*(5), 1523–1532. https://doi.org/10.3892/ijo.2012.1357

(51)Ashizawa, T., Miyata, H., Iizuka, A., Komiyama, M., Oshita, C., Kume, A., Nogami, M., Yagoto, M., Ito, I., Oishi, T., Watanabe, R., Mitsuya, K., Matsuno, K., Furuya, T., Okawara, T., Otsuka, M., Ogo, N., Asai, A., Nakasu, Y., Yamaguchi, K., & Akiyama, Y. (2013). Effect of the STAT3 inhibitor STX-0119 on the proliferation of cancer stem-like cells derived from recurrent glioblastoma. *International Journal of Oncology, 42*(5), 1535–1543. https://doi.org/10.3892/ijo.2013.1916

(52)Xu, F., Du, Y., Hang, S., Chen, A., Guo, F., & Xu, T. (2013). Adipocytes regulate the bone marrow microenvironment in a mouse model of obesity. *Molecular Medicine Reports, 7*(4), 1220–1224. https://doi.org/10.3892/mmr.2013.1572

(53)Carlsson, J., Shen, L., Xiang, J., Xu, J., & Wei, Q. (2012). Tendencies for higher co-expression of EGFR and HER2 and downregulation of HER3 in prostate cancer lymph node metastases compared with corresponding primary tumors. *Oncology Letters, 4*(5), 1015–1020. https://doi.org/10.3892/ol.2012.996

(54)Zhou, X., Zhao, F., Wang, Z.-N., Song, Y.-X., Chang, H., Chiang, Y., & Xu, H.-M. (2011). Altered expression of miR-152 and miR-148a in ovarian cancer is related to cell proliferation. *Oncology Reports, 26*(5), 1071–1077. https://doi.org/10.3892/or.2011.1482

(55)Peng, D.-X., Luo, M., Qiu, L.-W., He, Y.-L., & Wang, X.-F. (2012). Prognostic implications of microRNA-100 and its functional roles in human epithelial ovarian cancer. *Oncology Reports, 27*(5), 1423–1431. https://doi.org/10.3892/or.2012.1625

(56)Malara, A. E., Fedele, C., Aloj, L., Arra, C., Laccetti, P., D'Alessio, G., & De Lorenzo, C. (2012). Effects of a human compact anti-ErbB2 antibody on prostate cancer. *Oncology Reports, 27*(5), 1487–1494. https://doi.org/10.3892/or.2012.1760

(57)Lee, H. J., & Lee, C. H. (2013). Transglutaminase-2 is involved in expression of osteoprotegerin in MG-63 osteosarcoma cells. *Biomolecular Therapy, 21*(5), 319–324. https://doi.org/10.4062/biomolther.2013.037

(58)Perletti, G., Monti, E., Marras, E., Cleves, A., Magri, V., Trinchieri, A., & Rennie, P. S. (2015). Efficacy and safety of second-line agents for treatment of metastatic castration-resistant prostate cancer progressing after docetaxel: A systematic review and meta-analysis. *Acta Urologica*, *2*, 121. https://doi.org/10.4081/aiua.2015.2.121

(59)Xu, Y.-Z., Xi, Q.-H., Ge, W.-L., & Zhang, X.-Q. (2013). Significance of Serum microRNA-21 in Ovarian Cancer: Identification of Serum MicroRNA-21 as a Biomarker for Early Detection and Prognosis in Human Epithelial Ovarian Cancer. *Asian Pacific Journal of Cancer Prevention*, *14*(2), 1057–1060. https://doi.org/10.7314/APJCP.2013.14.2.1057

(60)Beck, A., & Reichert, J. M. (2011). Therapeutic Fc-fusion proteins and peptides as successful alternatives to antibodies. *mAbs*, *3*(5), 415–416. https://doi.org/10.4161/mabs.3.5.17334

(61)Emison, E. S., McCallion, A. S., Kashuk, C. S., Bush, R. T., Grice, E., Lin, S., Portnoy, M. E., Cutler, D. J., Green, E. D., & Chakravarti, A. (2005). A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, *434*(7031), 857–863. https://doi.org/10.1038/nature03474

(62)Nijjar, S. S., Crosby, H. A., Wallace, L., Hubbscher, S. G., & Strain, A. J. (2001). Notch receptor expression in adult human liver: A possible role in bile duct formation and hepatic neovascularization. *Hepatology*, *34*(6), 1188–1196. https://doi.org/10.1053/jhep.2001.29399

(63)Sathasivam, S. (2011). Current and emerging treatments for the management of myasthenia gravis. *Therapeutics and Clinical Risk Management*, *7*, 313–323. https://doi.org/10.2147/TCRM.S14015