Leveraging LLM-Generated QA Pairs for Biomedical Open-Domain Question Answering

Adrita Anika [1], **Md Messal Monem Miah**[1], **Man Luo** [2]
[1]Texas A&M University
[2] Arizona State University
adrita.anika@tamu.edu, messal.monem@tamu.edu, mluo26@asu.edu

Abstract

Large Language Models (LLMs) have shown considerable success in open-domain question answering (ODQA). Nevertheless, their performance in specialized fields such as healthcare remains suboptimal due to insufficient domain-specific knowledge. While integrating retrieved documents as in-context examples offers some improvement, it is often inadequate. In this study, we introduce a novel approach to enhance Biomedical ODQA by utilizing question-answer (QA) pairs generated from PubMed abstracts using LLM. We prompt an LLM with these QA pairs as in-context examples for four biomedical question types: yes/no, factoid, list, and summary. Our method outperforms document retrieval in factoid and list type question, matches performance in the other two, while significantly reducing inference latency across all four. We also provide detailed empirical analysis to support the effectiveness of QA pairs in boosting performance.

Introduction

LLMs have significantly advanced progress in various NLP tasks, including ODQA (Tan et al. 2023; Lazaridou et al. 2022). However, the knowledge encoded in LLMs is limited to their training corpus, posing challenges for ODQA, which requires answering questions from a vast range of sources (Lewis et al. 2021; Li, Zhang, and Zhao 2023). Researchers have explored LLM performance in ODQA in various settings. In a zero-shot setting, the LLM generates answers without any external knowledge source. To address the knowledge gap, two approaches have been explored. The first one is fine-tuning (Yang et al. 2019; Su et al. 2019) the model weights for specific domains which is expensive and often requires large corpus of high quality labeled data that can be hard to obtain. The second approach is in-context learning (ICL), where relevant contents are added as part of the prompt for the LLM to infer more reliably from the external context. Studies have shown that retrieval-augmented demonstrations for ICL can enhance QA system performance (Luo et al. 2023; Rubin, Herzig, and Berant 2021). While LLMs are considered to have extensive context length, irrelevant long texts (e.g. Wikipedia articles, PubMed articles) can hinder their reasoning ability and accuracy (Liu et al. 2024). Alternatives such as passage retrieval and summarization (Kim et al. 2024) have been explored, but each has limitations: passage retrieval can include unnecessary information, and summarization can omit crucial details necessary for generating accurate answers. Our study aims to provide LLMs with concise and comprehensive knowledge to overcome these limitations.

Previous research has explored retrieving QA pairs for ODQA, such as the PAQ dataset (Lewis et al. 2021), constructed from Wikipedia with a multi-step question generation process. Li, Zhang, and Zhao (2023) has shown the effectiveness of generating pseudo QA pairs in a zero-shot setting by writing short Wikipedia-style passages. This motivates us to investigate the effectiveness of using QA pairs for ICL-based biomedical ODQA, which often differs in style and complexity from generic Wikipedia based QA. We create an LLM-generated dataset of QA pairs from PubMed abstracts which are used for retrieval-augmented demonstrations in ICL for four types of question: yes/no, factoid, list, and summary. These QA pairs offer precise information for answering questions in comparison to abstracts, which can be very long and often contain irrelevant information. To ensure dataset quality, we implemented rigorous pre-generation prompt refinement, automated post-generation evaluation using Claude-3-Sonnet, manual reviews for discrepancies, BERT-score validation, and LDA-based topic modeling in summary, our main contributions are:

- We create a dataset of 455,015 QA pairs generated from 45,737 PubMed abstracts using an LLM, and plan to publicly release it to advance research in biomedical ODQA upon acceptance of the paper.
- We conduct extensive experiments comparing the use of QA pairs as ICL demonstrations against abstracts and summarized abstracts. QA pair retrieval improves performance by 5% for factoid and list questions, while for yes/no and summary questions, both QA and abstract-based systems perform similarly. Additionally, our method results in a 2x speedup in inference.
- We provide comprehensive analysis of our approach using three open-source LLMs (Llama-3.1-8B-Instruct, Mistral-7B-Instruct, Flan-T5-xl) and three popular retrievers (BM25, Contriever, GTR-Large). Our experi-

---

Copyright © 2025, GenAI4Health Workshop @ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ments demonstrate that QA pairs improve ODQA performance by enhancing retrieval and reasoning capabilities.

Related Work

Biomedical Open Domain Question Answering with LLMs

Recent advancements in AI have enhanced functionalities beyond simple keyword matching, yet these developments may still be unfamiliar to clinicians and researchers (Jin, Leaman, and Lu 2024; Jin et al. 2022). Retrieval-Augmented Generation (RAG) has been extensively explored to bridge this gap (Yang et al. 2023), but its success in specific domains like healthcare remains limited compared to the general domain (Cheng et al. 2024). Fine-tuning LLMs has proven more effective in this context. For instance, the authors fine-tuned LLMs with 205k doctor-patient conversations, successfully addressing the limitations of general-domain LLMs in the medical field. In fine-tuning RAG systems, the retriever and LLM are typically trained separately, leading to additional time and financial costs (Wang et al. 2024). Furthermore, fine-tuning retrievers requires document-query pairs, which may not be readily available for specific domains, adding to the complexity and expense. Jointly training LLMs and retrieval models has been proposed to enhance the efficiency and accuracy of medical question-answering systems

QA pairs Retrieval in Question Answering

Some research has explored QA pair generation and retrieval in QA tasks. For instance, Alberti et al. (2019) introduced a method for generating synthetic question-answering corpora by combining question generation and answer extraction models using BERT and general domain datasets. Another study (Lewis et al. 2021) presented PAQ dataset generated from Wikipedia and the RePAQ retriever, leveraging BERT and general domain datasets to enhance QA-pair models. Additionally, a Self-Prompting framework was proposed (Li, Zhang, and Zhao 2023), which utilizes LLMs to generate pseudo QA pairs for ODQA in a zero-shot setting, outperforming previous methods. To the best of our knowledge, this is the only study that has explored QA pairs for general-domain QA with LLMs, addressing challenges where no real QA database exists. However, this approach may not be suitable for specific domains like healthcare, as LLMs lack domain-specific knowledge.

Approach

Problem Definition

In ODQA task, the goal is to respond to questions by retrieving a specified number of relevant articles from a large corpus and using those articles to generate an accurate answer. Formally, let $\mathcal{C}$ be the corpus, Q the input question, and k the number of relevant articles to retrieve. The objective is to identify $\mathcal{D}_Q \subset \mathcal{C}$ where $|\mathcal{D}_Q| = k$, such that $\mathcal{D}_Q$ contains the most relevant articles for answering Q. This can be represented as $\mathcal{D}_Q = \text{Retrieve}(Q, \mathcal{C}, k)$. The answer A is then generated from $\mathcal{D}_Q$ using LLM. The four types of questions: yes/no, list, factoid, and summary, each require a distinct type of response.

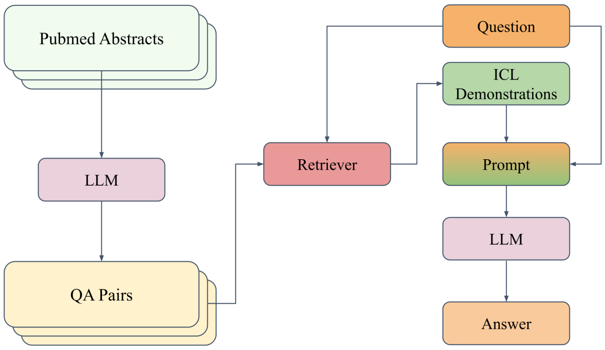


Figure 1: Biomedical ODQA with ICL-QA pairs

- **Yes/No Questions**: These are binary queries where the answer is $A \in \{\text{yes}, \text{no}\}$. For example, ”Is diabetes a chronic disease?” expects ”yes” or ”no.”
- **List Questions**: These require a list of entities, with the answer being $A = [E_1, E_2, \ldots, E_n]$, where E_i are entity names. For example, ”What are the symptoms of diabetes?” might result in $[E_1, E_2, E_3]$, such as fatigue, frequent urination, etc.
- **Factoid Questions**: These ask for a specific entity, where the answer is $A = E$, a single entity. For instance, ”What is the primary cause of Type 2 diabetes?” expects a response like ”insulin resistance.”
- **Summary Questions**: These require a detailed, multi-sentence answer summarizing key information. For example, ”Summarize the treatment options for diabetes” expects a concise overview of treatments like lifestyle changes and medication.

QA Pairs Generation

For generating QA pairs, we utilized abstracts from 45,737 PubMed articles included in the BioASQ 11B dataset (Nentidis et al. 2023). We prompted Llama-3-8b-instruct (AI@Meta 2024) with specific instructions to generate QA pairs from each PubMed abstract. These instructions emphasized capturing the detailed knowledge presented in the articles, rather than producing generic information about the studies. Additionally, we required the model to generate QA pairs whose answers are directly found within the passage. We aimed for the creation of as many QA pairs as possible, with a minimum of 10 per abstract, to ensure comprehensive coverage of the knowledge contained in each passage. Table 1 illustrates key details of the generated QA pairs dataset.

QA Pairs Assessment

The quality of the generated QA pair corpus plays a pivotal role in generating accurate answers. We employ two stages of quality assurance measures. The first is a pre-generation stage, where we continually update our prompt by addressing concerns from previous iterations. As illustrated in Figure 4 in appendix, the instructions provided to the model for

Statistic	Value
Total Number of QA Pairs	455,015
Unique Questions	422,773
Average Question Length (words)	11.5
Average Answer Length (words)	9.6

Table 1: Statistics of the QA Pair Dataset

generating higher-quality QA pairs reflect insights from detailed quality inspections. For instance, we include explicit instruction in the prompt to prevent the creation of generic questions like 'What is the goal of the study?' or 'How many diseases/patients were considered?' that do not contain any useful information.

After applying such stringent instructions to the prompt, we further inspect a select set of 10k QA pairs using automatic evaluation as part of our post-generation quality assurance step. We utilize Claude-3-Sonnet, a leading LLM known for its state-of-the-art performance across various tasks. The LLM generates answers for each question in our sampled dataset using the corresponding abstract. It is then tasked with comparing these answers to those from our QA pair corpus. The LLM is provided with the passage, the question, and both the answers. The LLM returns a judgment indicating whether the two answers matched. Our results show a match in 98.4% of the answers. For the remaining 1.6%, where the answers did not match, we manually reviewed the samples and found that the discrepancy was primarily due to differences in the phrasing of the terms in the answer. Additionally, we calculated the BERT-score between the two sets of answers, yielding an average score of 0.96, further confirming the high quality of the QA pairs.

The BioASQ corpus covers a wide range of biomedical topics, as illustrated by Figure 4 of the BioASQ dataset paper (Krithara et al., 2023), with key terms like Thyroid Hormone Receptor, Insulin, and Diseases. This diversity is reflected in our dataset, where we employed LDA topic modeling to evaluate coverage. By deriving topics from each passage and corresponding question-answer pairs, and calculating the cosine similarity between them, we found an overall topic match rate of 87.4%. This confirms that our dataset maintains the broad coverage of BioASQ. Upon reviewing examples with lower coverage, we found that mismatches often result from inaccuracies in topic modeling by the LDA model.

ICL Demonstrations Retrieval

In this study, we experiment with 3 popular retrievers, BM25 (Robertson and Zaragoza 2009), Contriever (Izacard et al. 2022) and GTR-Large (Ni et al. 2022) for retrieving relevant ICL examples. The detail descriptions of the retrievers are provided in Appendix .

Inference with Retrieved QA pairs

For inference, we prompt the LLM to answer a given question using the retrieved QA pairs, considering four different types of questions: yes/no, factoid, list, and summary questions. For yes/no questions, we prompt the LLM with 'Respond with only one word, Yes or No.' For factoid questions, the prompt is 'Respond only with the answers. You can return a list of up to 5 entity names ordered by decreasing confidence.' For list questions, the prompt is 'The question is a list type question. Return a list with the options only.' For summary questions, we prompt the LLM with 'Give a short answer to the question with one or two sentences.' This approach ensures that the LLM provides concise and relevant answers tailored to each specific question type.

Figure 1 illustrates our pipeline where QA pairs are generated from PubMed abstracts, then retrieved to construct ICL demonstrations for ODQA.

Experimental Setup

Dataset

In this study, we utilized the BioASQ 11b dataset (Nentidis et al. 2023), which is a benchmark dataset for biomedical semantic indexing and question answering. It includes a total of 5049 QA pairs (1357 yes/no questions, 1515 factoid questions, 967 list questions, and 1210 summary questions). These QA pairs are linked to PubMed documents that contain the answers to the questions, comprising a corpus of 45,737 documents used for both retrieval and QA generation tasks.

Baselines

We considered the following baselines to evaluate the performance of our method:

- **Zero-Shot Setting:** This baseline evaluates LLMs' inherent knowledge and reasoning capabilities to answer questions from the corpus without task-specific training or in-context examples.
- **ICL-Abstracts:** In this setting, we retrieve abstracts relevant to the questions and provide them as in-context examples to the LLM.
- **ICL-Summaries:** In this setup, we experiment with summaries of the abstracts as alternative type of context for the model. The full abstracts are often very long and providing several abstracts within the prompt can be confusing for the LLM, hence we summarize the abstracts into a shorter fixed length summary to present the information concisely. The summarization was performed using T5 based (Raffel et al. 2020) summarizer model with maximum length of 256 tokens.

We evaluate the performance of these methods using several metrics across the four types of questions. These metrics are discussed in Appendix .

Results and Analysis

Evaluation with Llama-8b-Instruct in Different Settings

Table 2 presents the performance of Llama-8b-Instruct across four question types in different settings: zero-shot, ICL-abstracts, ICL-summaries, and ICL-QA pairs, for various retrievers. All three ICL methods outperform the zero-shot approach, highlighting the effectiveness of integrating external retrieval mechanisms.

Method	Retriever (top k)	Yes/No		Factoid			List			Summary	
		Acc	maF1	MRR	S Acc	L Acc	Precision	Recall	F1	R-2	R-LSum
Zero Shot	-	0.78	0.74	0.01	0.01	0.01	0.13	0.14	0.12	0.09	0.22
ICL-Abstracts	BM25 (5)	0.91	**0.89**	0.18	0.18	0.20	0.17	0.16	0.16	0.21	0.32
	Contriever (5)	**0.92**	**0.89**	0.20	0.20	0.22	0.17	0.16	0.16	0.21	0.32
	GTR-Large (5)	0.91	0.87	0.23	0.21	0.26	0.26	0.26	0.26	**0.22**	**0.33**
	BM25 (10)	0.86	0.86	0.10	0.10	0.12	0.04	0.04	0.04	0.18	0.29
	Contriever (10)	0.86	0.86	0.12	0.12	0.14	0.09	0.09	0.08	0.18	0.31
	GTR-Large (10)	0.87	0.86	0.10	0.10	0.11	0.07	0.07	0.07	0.18	0.29
ICL-Summaries	BM25 (5)	0.82	0.81	0.22	0.21	0.25	0.23	0.15	0.17	0.13	0.21
	Contriever (5)	0.87	0.87	0.20	0.20	0.12	0.12	0.12	0.12	0.16	0.27
	GTR-Large (5)	0.84	0.83	0.22	0.21	0.25	0.25	0.23	0.24	0.15	0.26
	BM25 (10)	0.81	0.81	0.20	0.20	0.20	0.1	0.06	0.06	0.14	0.22
	Contriever (10)	0.87	0.87	0.21	0.21	0.15	0.15	0.15	0.15	0.14	0.25
	GTR-Large (10)	0.85	0.85	0.23	0.22	0.24	0.24	0.24	0.24	0.15	0.27
ICL-QA Pairs	BM25 (5)	0.89	0.86	**0.28**	**0.26**	0.29	0.27	0.27	0.27	0.15	0.28
	Contriever (5)	0.90	0.88	**0.28**	**0.26**	**0.31**	0.27	0.28	0.27	0.18	0.30
	GTR-Large (5)	0.89	0.87	0.27	0.25	0.30	**0.32**	**0.29**	**0.30**	0.20	0.31
	BM25 (10)	0.91	0.88	0.27	**0.26**	0.28	0.30	0.26	0.27	0.17	0.30
	Contriever (10)	0.91	**0.89**	**0.28**	**0.26**	0.30	0.29	**0.29**	0.29	0.20	**0.33**
	GTR-Large (10)	0.91	0.88	027	0.25	0.29	0.31	**0.29**	**0.30**	0.21	**0.33**

Table 2: Performance comparison of different methods with Llama-3.1-8B-Instruct

LLM	Method	Yes/No		Factoid			List			Summary	
		Acc	maF1	MRR	S Acc	L Acc	Precision	Recall	F1	R-2	R-LSum
Mistral-7B-Instruct	ICL-Abstracts	0.81	0.79	0.13	0.11	0.17	0.24	0.17	0.20	0.15	0.24
	ICL-Summaries	0.74	0.71	0.1	0.09	0.12	0.22	0.13	0.16	0.15	0.24
	ICL-QA Pairs	0.86	0.85	0.17	0.16	0.18	0.31	0.22	0.23	0.16	0.25
Flan-T5-xl	ICL-Abstracts	0.74	0.73	0.14	0.14	0.14	0.10	0.08	0.11	0.10	0.14
	ICL-Summaries	0.72	0.72	0.13	0.13	0.13	0.2	0.08	0.10	0.10	0.16
	ICL-QA Pairs	0.79	0.79	0.18	0.18	0.18	0.27	0.11	0.14	0.14	0.20

Table 3: Performance comparison of different methods with GTR-Large(5)

For Yes/No questions, ICL-QA pairs and ICL-abstracts achieve similar performance in terms of accuracy and macro-F1 score, while ICL-abstracts perform slightly worse. Notably, performance declines when using the top 10 retrieved abstracts compared to the top 5. Conversely, using more QA pairs slightly boosts performance.

For factoid questions, ICL-QA pairs show approximately a 5% improvement across all metrics compared to both ICL-abstracts and ICL-summaries, with the highest performance achieved using Contriever as the retriever for this question type followed by GTR-Large.

For list questions, ICL-QA pairs outperform ICL-abstracts with a 5% higher MRR, 3% higher strict accuracy, and 4% higher lenient accuracy, with GTR-Large being the best performing retriever. Additionally, ICL-abstracts show a notable drop in performance when 10 abstracts are used as in-context examples compared to 5. This can be attributed to the fact that a higher number of abstracts introduces more irrelevant information, making it harder for the LLM to reason and provide accurate answers. In contrast, ICL-QA pairs remain consistent or improve when more in-context examples are used.

For summary-type questions, both ICL-QA pairs and ICL-abstracts perform similarly in terms of Rouge-2 and Rouge-L scores and ICL-summaries perform slightly worse.

Evaluation with Different LLMs for GTR-Large

Table 3 presents the performance comparison of Mistral-7B-Instruct and Flan-T5-xl across different question types under three settings: ICL-abstracts, ICL-summaries, and ICL-QA pairs, using the GTR-Large retriever with the top 5 results. For Yes/No questions, ICL-QA pairs achieve the highest performance in terms of accuracy and macro-F1 score for both models, outperforming ICL-abstracts and ICL-summaries. Mistral-7B-Instruct generally shows better results than Flan-T5-xl across all settings for Yes/No questions.

For factoid questions, ICL-QA pairs yield the highest MRR, strict accuracy, and lenient accuracy for both models, with notable improvements over ICL-abstracts and ICL-summaries. Mistral-7B-Instruct shows a stronger performance compared to Flan-T5-xl across all metrics.

In the list question category, ICL-QA pairs once again demonstrate superior results, with a higher mean precision, recall, and F1 score than the other settings. The difference in performance is more pronounced for Mistral-7B-Instruct, while Flan-T5-xl exhibits a smaller gap between ICL-QA pairs and the other two methods.

For summary questions, the differences between ICL-QA

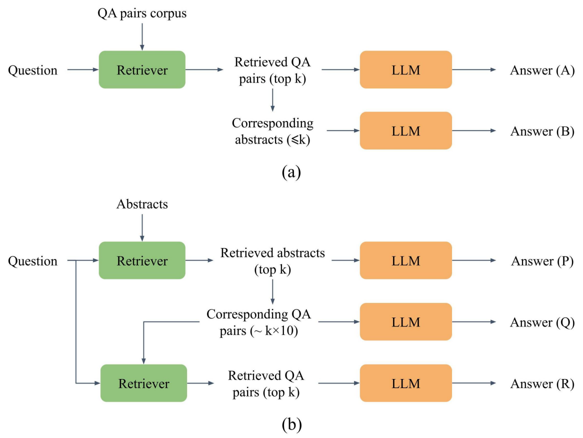

QA pairs corpus
Question
Retriever
Retrieved QA pairs (top k)
LLM
Answer (A)
Corresponding abstracts (≤k)
LLM
Answer (B)
(a)
Abstracts
Question
Retriever
Retrieved abstracts (top k)
LLM
Answer (P)
Corresponding QA pairs (~ k×10)
LLM
Answer (Q)
Retriever
Retrieved QA pairs (top k)
LLM
Answer (R)
(b)

Figure 2: Experimental setup for the probing study: (A) investigation of reasoning capabilities using QA pairs (B) investigation of the retrieval effectiveness of QA pairs.

pairs, ICL-abstracts, and ICL-summaries are less significant. Both models perform similarly across settings in terms of ROUGE-2 and ROUGE-Lsum scores, though Mistral-7B-Instruct slightly edges out Flan-T5-xl in most cases.

Probing Study

Do QA pairs facilitate reasoning? The goal is to assess whether QA pairs can enhance the reasoning capabilities of LLMs. We compare two methods as shown in Figure 2a: method A: ICL-QA pairs, and method B: ICL-(QA pairs → abstracts), where the abstracts used as demonstrations correspond to the sources from which the QA pairs were derived. Since the number of abstracts may be fewer than or equal to the number of QA pairs, this experiment evaluates whether the same information contained in QA pairs, when presented as abstracts, performs similarly.

Results in Figure 3a illustrate that across all question types—Yes/No, Factoid, List, and Summary—and metrics, method A consistently outperform method B. The performance gap is especially noticeable for metrics such as accuracy and macro-F1 for Yes/No questions, MRR and strict accuracy for Factoid questions, and precision for List questions. This indicates that QA pairs provide more concise and targeted information, helping LLMs to reason more effectively compared to using full abstracts.

Are QA pairs easier to retrieve? To assess whether QA pairs are easier to retrieve than abstracts, we conduct experiments as depicted in Figure 2b. The baseline (Method P) uses ICL-abstracts, while method Q maps retrieved abstracts to corresponding QA pairs as demonstrations, ICL-(abstracts→all QA pairs). Given that an abstract can yield around 10 QA pairs, this leads to approximately $k \times 10$ QA pairs. To manage this volume, method R employs a secondary retrieval step to select a few QA pairs, we refer to this as ICL-(abstracts→all QA pairs→k QA pairs).

The results in Figure 3b show that ICL-abstracts (method P) outperform both Method Q and R across all question types—Yes/No, Factoid, List, and Summary. This indicates that even when reasoning is conducted with documents (method P) or their corresponding QA pairs (Method Q and R), using the original ICL-QA pairs (method A from) provides superior results. Notably, method Q performs worse than method R, likely due to the larger number of QA pairs (around 50), which makes it harder for the LLM to effectively reason.

Overall, our proposed ICL-QA pairs (method A) still achieve higher performance than all three approaches in factoid and list type questions and similar performance in the remaining types, suggesting that retrieving and reasoning with QA pairs directly is more effective than using retrieved documents or large volumes of QA pairs. The results shown in the graph are with Llama-3.1-8B for GTR-Large with $k = 5$.

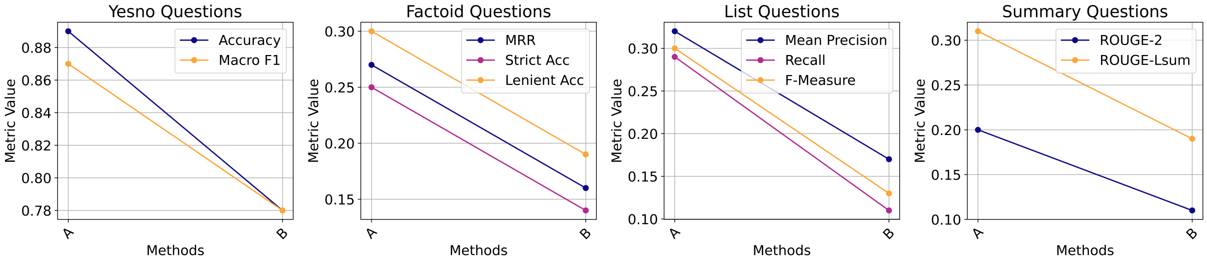
Qualitative Analysis

QA pairs are more concise and easier to retrieve than abstracts. Our results show that ICL-QA pairs outperform ICL-abstracts on factoid and list-type questions. For instance, in response to the list-type question 'Which antibiotics target peptidoglycan biosynthesis?', ICL-abstracts returned 2 correct entities because the retrieved documents mentioned only these 2 entities. In contrast, ICL-QA pairs returned 4 correct entities as those were present in the retrieved QA pairs, as shown in Example 4 in appendix. This disparity arises because abstracts are typically large with many irrelevant sentences, making it challenging for retriever to extract relevant information. The concise nature of ICL-QA pairs, however, facilitates easier retrieval and reasoning, leading to more accurate entity extraction.

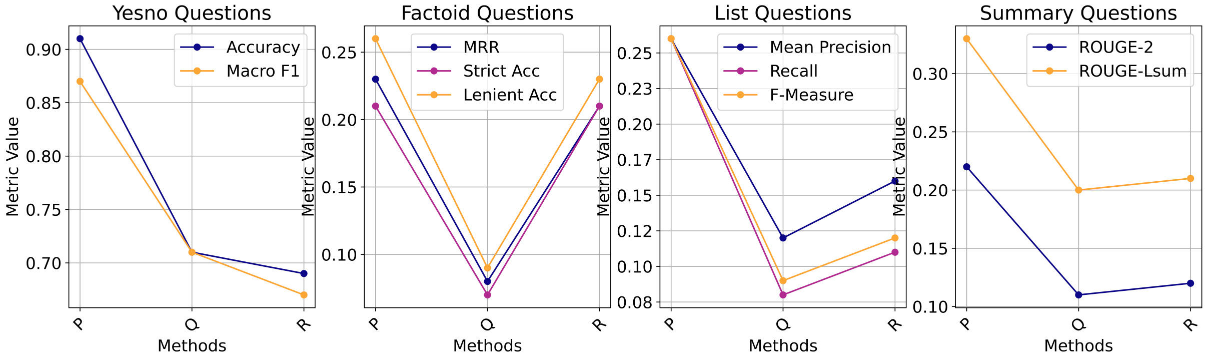
Lengthy abstracts with limited relevant information hinder the LLM's reasoning. In some cases, even if retrieved, it is harder for the LLM to reason from lengthy abstracts where only one or two snippets are relevant and the rest are irrelevant. For example, in the factoid question 'Which is the human selenoprotein that contains several Se-Cys residues?', ICL-abstracts, although retrieved relevant documents, failed to answer correctly. Conversely, from appendix Example 5 we observe, ICL-QA pairs correctly retrieved QA pairs and provided the correct answer.

Handling contradictory information affects yes/no question answers. For instance, in the question 'Has vitamin D been shown to reduce the incidence of falls in older people in clinical trials?', the ground truth answer is 'yes' based on supporting studies. ICL-abstracts retrieved relevant documents confirming this result, thus answering 'yes'. However, ICL-QA pairs retrieved contradictory results from different studies, leading to a 'no' answer, as shown in appendix Example 2. This discrepancy arises because ICL-QA pairs include concise information from various studies, including those with contradictory findings. Removing these contradictory results aligns ICL-QA pairs' answers with ICL-abstracts, indicating that this is not a fundamental shortcoming of ICL-QA pairs.

Missing QA pairs in the corpus affect yes/no questions. Sometimes, relevant QA pairs do not exist in the corpus, preventing their retrieval (appendix Example 3), which is not



(a) Experimental results from reasoning capabilities (method A: ICL-QA pairs, method B: ICL-(QA pairs→ abstracts))



(b) Experimental results from retrieval effectiveness (method P: ICL-abstracts, method Q: ICL-(abstracts→all QA pairs), method R: ICL-(abstracts→all QA pairs→k QA pairs))

Figure 3: Experimental results from investigations of reasoning and retrieval capabilities.

a shortcoming of the retriever or the LLM. Since, a limited number of QA pairs are generated from each abstract, sometimes LLM may not capture all necessary knowledge required to answer all the questions. To address this issue, we plan to generate questions at the sentence level to increase the coverage of QA pairs corpus in the future. The specificity required for yes/no answers is often missing in the QA pairs corpus, leading to occasional performance drop for ICL-QA pairs compared to ICL-abstracts. However, for yes/no questions where the QA pairs corpus contains the knowledge, the retriever can correctly retrieve and the LLM can reason to answer accurately, even when ICL-abstracts fail to provide a correct answer despite correct retrieval. Example 1 in appendix demonstrates such a scenario.

Inference Time

LLM	Method	K=5	K=10
Llama-3.1-8B	QA Pairs	1.67	1.83
	Abstracts	3.66	6.51
Mistral-7B	QA Pairs	1.62	1.88
	Abstracts	3.92	5.30
Flan-T5-xl	QA Pairs	3.71	6.13
	Abstracts	11.08	15.3

Table 4: Average inference time (sec) for each sample

The inference time analysis in Table 4 shows that ICL-QA pairs are consistently faster across all models compared to ICL-abstracts. For the Llama-3.1-8B-Inst. model, the ICL-QA pairs take 1.67 seconds for K=5 and 1.83 seconds for K=10, while ICL-abstracts require 3.66 seconds and 6.51 seconds respectively, demonstrating that ICL-QA pairs are approximately twice as fast. Similarly, for the Mistral-7B-Inst. model, the ICL-QA pairs achieve 1.62 seconds for K=5 and 1.88 seconds for K=10, while ICL-abstracts take longer at 3.92 seconds and 5.30 seconds. Again, ICL-QA pairs are significantly faster. Finally, the Flan-T5-xl model shows a more drastic difference in performance. While the ICL-QA pairs take 3.71 seconds for K=5 and 6.13 seconds for K=10, the ICL-abstracts have much higher times of 17.08 seconds and 35.3 seconds, indicating a considerable slowdown for processing abstracts compared to QA pairs.

Conclusion

In this study, we demonstrated that utilizing QA pairs generated from PubMed abstracts can significantly improve the performance of factoid and list-type questions while achieving similar performance for yes/no and summary-type questions. Additionally, this approach reduces the inference latency of Biomedical ODQA with LLMs. Thus, our approach highlights the effectiveness of concise, targeted information in enhancing LLM-based question answering in specialized fields like healthcare. However, it is important to exercise caution when applying LLM-generated QA pairs in practical settings, as with any LLM-based system, due to potential limitations such as biases or inaccuracies in the generated content.

References

AI@Meta. 2024. Llama 3 Model Card.

Alberti, C.; Andor, D.; Pitler, E.; Devlin, J.; and Collins, M. 2019. Synthetic QA corpora generation with roundtrip consistency. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6168–6173.

Cheng, X.; Luo, D.; Chen, X.; Liu, L.; Zhao, D.; and Yan, R. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Jin, Q.; Leaman, R.; and Lu, Z. 2024. Pubmed and beyond: Biomedical literature search in the age of artificial intelligence. *Ebiomedicine*, 100.

Jin, Q.; Yuan, Z.; Xiong, G.; Yu, Q.; Ying, H.; Tan, C.; Chen, M.; Huang, S.; Liu, X.; and Yu, S. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2): 1–36.

Kim, J.; Nam, J.; Mo, S.; Park, J.; Lee, S.-W.; Seo, M.; Ha, J.-W.; and Shin, J. 2024. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. *arXiv preprint arXiv:2404.13081*.

Lazaridou, A.; Gribovskaya, E.; Stokowiec, W.; and Grigorev, N. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Lewis, P.; Wu, Y.; Liu, L.; Minervini, P.; Küttler, H.; Piktus, A.; Stenetorp, P.; and Riedel, S. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9: 1098–1115.

Li, J.; Zhang, Z.; and Zhao, H. 2023. Self-prompting large language models for zero-shot open-domain qa.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Luo, M.; Xu, X.; Dai, Z.; Pasupat, P.; Kazemi, M.; Baral, C.; Imbrasaite, V.; and Zhao, V. Y. 2023. Dr. ICL: Demonstration-retrieved in-context learning. *arXiv preprint arXiv:2305.14128*.

Nentidis, A.; Katsimpras, G.; Krithara, A.; López, S. L.; Farré-Maduell, E.; Gasco, L.; Krallinger, M.; and Paliouras, G. 2023. Overview of BioASQ 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 227–250. Springer.

Ni, J.; Qu, C.; Lu, J.; Dai, Z.; Hernandez Abrego, G.; Ma, J.; Zhao, V.; Luan, Y.; Hall, K.; Chang, M.-W.; and Yang, Y. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9844–9855. Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Robertson, S.; and Zaragoza, H. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Rubin, O.; Herzig, J.; and Berant, J. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Su, D.; Xu, Y.; Winata, G. I.; Xu, P.; Kim, H.; Liu, Z.; and Fung, P. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd workshop on machine reading for question answering*, 203–211.

Tan, Y.; Min, D.; Li, Y.; Li, W.; Hu, N.; Chen, Y.; and Qi, G. 2023. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. In *International Semantic Web Conference*, 348–367. Springer.

Wang, J.; Yang, Z.; Yao, Z.; and Yu, H. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Yang, W.; Xie, Y.; Tan, L.; Xiong, K.; Li, M.; and Lin, J. 2019. Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*.

Yang, Z.; Yao, Z.; Tasmin, M.; Vashisht, P.; Jang, W. S.; Wang, B.; Berlowitz, D.; and Yu, H. 2023. Performance of Multimodal GPT-4V on USMLE with Image: Potential for Imaging Diagnostic Support with Explanations. *medRxiv*.

Generated QA Pairs Dataset Samples

Figure 4 in appendix illustrates the prompt used to generate QA pairs from the PubMed abstracts. The prompt includes a set of instruction to better guide the LLM in generating useful QA pairs. The figure also provides 10 example QA pairs generated by Llama-3, ensuring each question is directly based on the passage.

Retrievers

BM25 is a widely-used ranking function for text retrieval (Robertson and Zaragoza 2009). It assesses the relevance of articles to a query by leveraging term frequency (TF), inverse document frequency (IDF), and document length normalization, ensuring more accurate and context-aware search results.

Contriever is a dense retriever trained with a contrastive learning objective (Izacard et al. 2022). For our approach, we extract embeddings from both the query and the articles and rank the articles based on their cosine similarity with the query embedding.

GTR-Large (Ni et al. 2022) is also a dense retriever built on top of T5-large encoders pre-trained with a contrastive learning objective. Similar to contriever, we extract embeddings from both the query and the articles and rank the articles based on their cosine similarity with the query embedding.

Evaluation Metrics

Yes/No Questions

Accuracy (Acc) measures correct answers out of total yes/no questions (n), where c is the number of correct answers:

$$\text{Acc} = \frac{c}{n}$$

The macro-averaged F-measure (maF1) is:

$$\text{maF1} = \frac{F1_y + F1_n}{2}$$

where $F1_y$ and $F1_n$ are the F1 scores for 'yes' and 'no' answers, respectively.

Factoid Questions

Strict accuracy (SAcc) and lenient accuracy (LAcc) are measured for factoid questions. SAcc is correct if the first list element matches, and LAcc is correct if any list element matches. With n factoid questions, c_1 correct first elements, and c_5 correct in any position:

$$\text{SAcc} = \frac{c_1}{n}$$

$$\text{LAcc} = \frac{c_5}{n}$$

Mean reciprocal rank (MRR) for rank $r(i)$:

$$\text{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r(i)}$$

List Questions

For each list question, the system's returned list is evaluated against the golden list by calculating precision (P), recall (R), and F-measure (F1). True Positives (TP) are the entities present in both the returned and golden lists; False Positives (FP) are the entities in the returned list but not in the golden list; and False Negatives (FN) are the entities in the golden list but not in the returned list.

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Averaging these metrics provides mean scores.

Summary Type

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are used to evaluate the quality of summaries by comparing them to reference summaries. Specifically, we use ROUGE-2 and ROUGE-Lsum metrics.ROUGE-Lsum measures the longest common subsequence (LCS) between the candidate and reference summaries. The ROUGE-2 and ROUGE-Lsum scores are computed using the `rouge-score` library in Python.

Budgeting Hyper-parameters

The table shows the hyperparameters of the LLM model used for experimentation and their respective values. We utilized A6000 GPUs for the generation of question-answer pairs and for conducting question answering with LLMs. The GPU processing took approximately 72 hours to generate the proposed corpus. For the retrieval tasks, we employed A100 GPUs.

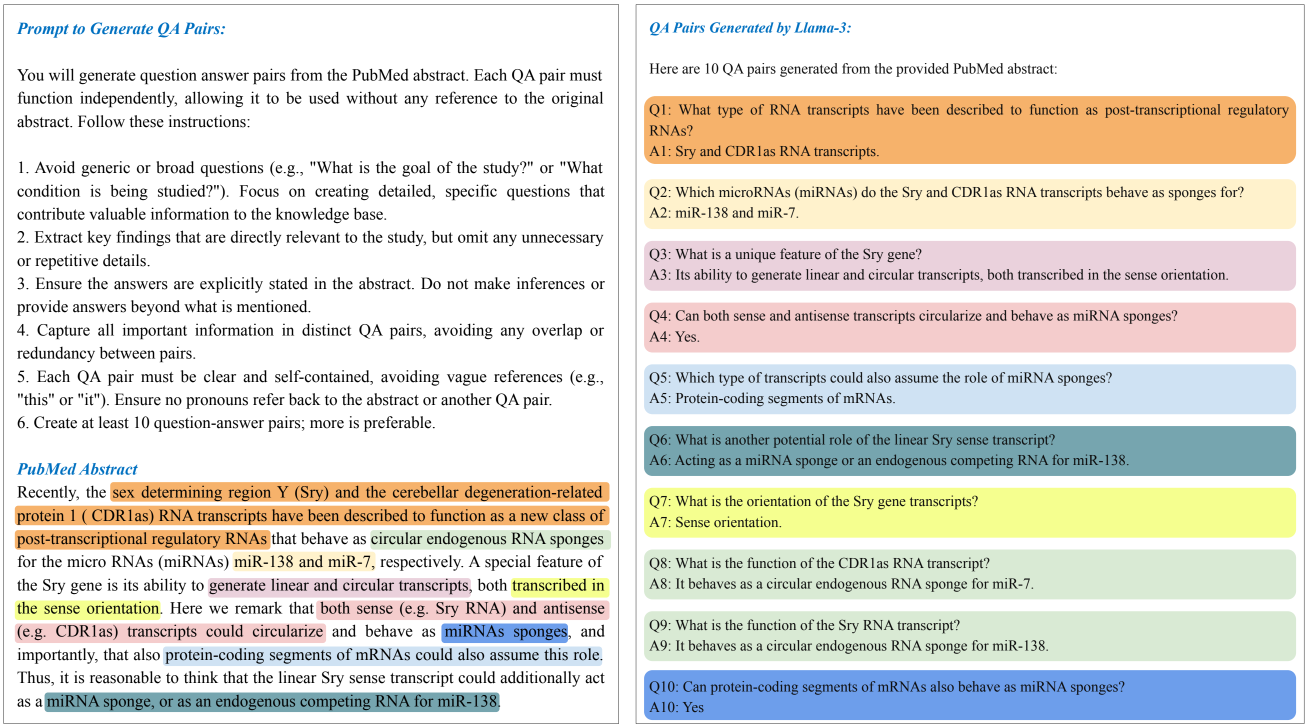
Hyperparameter	**Value**
temperature	0.9
max_new_tokens	512
top_p	0.9

Table 5: Hyperparameter Values

Analysis Examples

Several samples have been analyzed below in both ICL-abstracts and ICL-QA pairs settings Table 1, 2, 3, 4, 5.

.



Prompt to Generate QA Pairs:

You will generate question answer pairs from the PubMed abstract. Each QA pair must function independently, allowing it to be used without any reference to the original abstract. Follow these instructions:

1. Avoid generic or broad questions (e.g., "What is the goal of the study?" or "What condition is being studied?"). Focus on creating detailed, specific questions that contribute valuable information to the knowledge base.
2. Extract key findings that are directly relevant to the study, but omit any unnecessary or repetitive details.
3. Ensure the answers are explicitly stated in the abstract. Do not make inferences or provide answers beyond what is mentioned.
4. Capture all important information in distinct QA pairs, avoiding any overlap or redundancy between pairs.
5. Each QA pair must be clear and self-contained, avoiding vague references (e.g., "this" or "it"). Ensure no pronouns refer back to the abstract or another QA pair.
6. Create at least 10 question-answer pairs; more is preferable.

PubMed Abstract

Recently, the sex determining region Y (Sry) and the cerebellar degeneration-related protein 1 (CDR1as) RNA transcripts have been described to function as a new class of post-transcriptional regulatory RNAs that behave as circular endogenous RNA sponges for the micro RNAs (miRNAs) miR-138 and miR-7, respectively. A special feature of the Sry gene is its ability to generate linear and circular transcripts, both transcribed in the sense orientation. Here we remark that both sense (e.g. Sry RNA) and antisense (e.g. CDR1as) transcripts could circularize and behave as miRNAs sponges, and importantly, that also protein-coding segments of mRNAs could also assume this role. Thus, it is reasonable to think that the linear Sry sense transcript could additionally act as a miRNA sponge, or as an endogenous competing RNA for miR-138.

QA Pairs Generated by Llama-3:

Here are 10 QA pairs generated from the provided PubMed abstract:

Q1: What type of RNA transcripts have been described to function as post-transcriptional regulatory RNAs?
A1: Sry and CDR1as RNA transcripts.

Q2: Which microRNAs (miRNAs) do the Sry and CDR1as RNA transcripts behave as sponges for?
A2: miR-138 and miR-7.

Q3: What is a unique feature of the Sry gene?
A3: Its ability to generate linear and circular transcripts, both transcribed in the sense orientation.

Q4: Can both sense and antisense transcripts circularize and behave as miRNA sponges?
A4: Yes.

Q5: Which type of transcripts could also assume the role of miRNA sponges?
A5: Protein-coding segments of mRNAs.

Q6: What is another potential role of the linear Sry sense transcript?
A6: Acting as a miRNA sponge or an endogenous competing RNA for miR-138.

Q7: What is the orientation of the Sry gene transcripts?
A7: Sense orientation.

Q8: What is the function of the CDR1as RNA transcript?
A8: It behaves as a circular endogenous RNA sponge for miR-7.

Q9: What is the function of the Sry RNA transcript?
A9: It behaves as a circular endogenous RNA sponge for miR-138.

Q10: Can protein-coding segments of mRNAs also behave as miRNA sponges?
A10: Yes

Figure 4: Example of a PubMed abstract with prompting instructions and QA pairs generated by Llama-3.1-8b-Instruct. Color coding is used to highlight the sources of the information presented in each QA pair within the abstract.

Example 1: Yes/No type Question

Question: Are integrins part of the extracellular matrix?
Gold Answer: Yes

With ICL-Abstracts

Retrieved Abstracts:

1. The integrins are a superfamily of cell adhesion receptors that bind to extracellular matrix ligands, cell-surface ligands, and soluble ligands. They are transmembrane alphabeta heterodimers and at least 18 alpha and eight beta subunits are known in humans, generating 24 heterodimers. Members of this family have been found in mammals, chicken and zebrafish, as well as lower eukaryotes, including sponges, the nematode Caenorhabditis elegans (two alpha and one beta subunits, generating two integrins) and the fruitfly Drosophila melanogaster (five alpha and one beta, generating five integrins). The alpha and beta subunits have distinct domain structures, with extracellular domains from each subunit contributing to the ligand-binding site of the heterodimer. The sequence arginine-glycine-aspartic acid (RGD) was identified as a general integrin-binding motif, but individual integrins are also specific for particular protein ligands. Immunologically important integrin ligands are the intercellular adhesion molecules (ICAMs), immunoglobulin superfamily members present on inflamed endothelium and antigen-presenting cells. On ligand binding, integrins transduce signals into the cell interior; they can also receive intracellular signals that regulate their ligand-binding affinity. Here we provide a brief overview that concentrates mostly on the organization, structure and function of mammalian integrins.
2. The extracellular matrix is a three-dimensional network of proteins, glycosaminoglycans and other macromolecules. The extracellular matrix conveys signals through membrane receptors called integrins and plays an important role in pituitary physiology and tumorigenesis. There is a differential expression of extracellular matrix components and integrins during the pituitary development in the embryo and during tumorigenesis in the adult. Different extracellular matrix components regulate adrenocorticotropin at the level of the proopiomelanocortin gene transcription. The extracellular matrix also controls the proliferation of adrenocorticotropin-secreting tumor cells. On the other hand, laminin regulates the production of prolactin. The remodeling of the extracellular matrix affects pituitary cell proliferation. Matrix metalloproteinase activity is very high in all types of human pituitary adenomas. Matrix metalloproteinase secreted by pituitary cells can release growth factors from the extracellular matrix that, in turn, control pituitary cell proliferation and hormone secretion. In summary, the differential expression of extracellular matrix components, integrins and matrix metalloproteinase contributes to the control of pituitary hormone production and cell proliferation during tumorigenesis.
3. Integrin trafficking plays an important role in cellular motility and cytokinesis. Integrins undergo constant endo/exocytic shuttling to facilitate the dynamic regulation of cell adhesion. Integrin activity toward the components of the extracellular matrix is regulated by the ability of these receptors to switch between active and inactive conformations. Several cellular signalling pathways have been described in the regulation of integrin traffic under different conditions. The net endocytosis rate of the active $\beta 1$ integrins is higher, whereas endocytosis of the inactive $\beta 1$ integrin is counteracted by rapid recycling back to the plasma membrane via an ARF6- and early endosome antigen 1-positive compartment in an Rab4a- and actin-dependent manner. Owing to these distinct trafficking routes, the two receptor pools display divergent subcellular localization. At steady state, the inactive $\beta 1$ integrin is mainly on the plasma membrane, whereas the active receptor is predominantly intracellular. These data provide new insights into the endocytic traffic of integrins and imply the possibility of a previously unappreciated crosstalk between pathways regulating integrin activity and traffic. © 2012 John Wiley & Sons A/S.

LLM Answer: No

With ICL-QA Pairs

Retrieved QA Pairs:

1. Question: What is the role of integrins in the extracellular matrix? Answer: Integrins convey signals through membrane receptors and play an important role in pituitary physiology and tumorigenesis.
2. Question: What are the ligands of integrins? Answer: Extracellular matrix (ECM) proteins.
3. Question: What is the role of integrins in the immune response? Answer: They are extracellular matrix proteins that modulate many cellular responses.

LLM Answer: Yes

Example 2: Yes/No type Question

Question: Has vitamin D has been shown to reduce incidence of falls in older people in clinical trials? Answer: Yes

---

With ICL- Abstracts

Retrieved Abstracts:

1. Increasing data suggest that many or most adults in the United States and Europe would benefit from vitamin D supplements. This review summarizes the benefits of vitamin D with the strongest evidence today from randomized controlled trials for fall and fracture prevention. Beyond fall and fracture prevention, vitamin D may also reduce overall morbidity by multiple mechanisms. Prospective epidemiological studies supported by strong mechanistic evidence suggest a reduction of cardiovascular disease (incident hypertension and cardiovascular mortality) and colorectal cancer, extending to weaker evidence on immune-modulatory and anti-inflammatory benefits of vitamin D.
2. Context: Vitamin D affects bone and muscle health and likely reduces the risk of falls in the elderly. Objective: The aim of this systematic review is to summarize the existing evidence on vitamin D use and the risk of falls. Data sources: We searched electronic databases from inception through August 2010. Study selection: Eligible studies were randomized controlled trials in which the intervention was vitamin D and the incidence of falls was reported. Data extraction: Reviewers working in duplicate and independently extracted study characteristics, quality, and outcomes data. Data synthesis: Odds ratio and associated 95% confidence interval were estimated from each study and pooled using the random effects model. Results: We found 26 eligible trials of moderate quality that enrolled 45,782 participants, the majority of which were elderly and female. Vitamin D use was associated with statistically significant reduction in the risk of falls (odds ratio for suffering at least one fall, 0.86; 95% confidence interval, 0.77-0.96). This effect was more prominent in patients who were vitamin D deficient at baseline and in studies in which calcium was coadministered with vitamin D. The quality of evidence was low to moderate because of heterogeneity and publication bias. Conclusions: Vitamin D combined with calcium reduces the risk of falls. The reduction in studies without calcium coadministration did not reach statistical significance. The majority of the evidence is derived from trials enrolling elderly women.
3. The evidence that specific vitamins may be beneficial in the prevention of cardiovascular disease (CVD) is supported by mechanistic models of atherogenesis. We and others have published observational epidemiologic studies in support of vitamins in the primary prevention of CVD, but the results from intervention studies are mixed. This article summarizes the recent results for vitamin E, vitamin D, and the B vitamins, comparing study populations, study designs, and potential methodologic reasons for differences in findings. For vitamin E, observational data suggest benefit at doses of 100 to 400 IU/d. Results from recent large-scale trials are mixed, with some showing modest benefit but others suggesting no benefit, especially for secondary prevention. Results for B vitamins are also mixed and further complicated by the recent folate fortification of the flour supply. If greater B vitamin intake does reduce CVD, the benefits are likely to be greatest for primary prevention and in populations with intake below dietary reference standards. Research on vitamin D and CVD is just beginning to emerge, but current data suggest that if there is benefit it likely needs to be at intake levels much higher than the current reference intakes of 200 to 600 IU/d for American adults

LLM Answer: Yes

With ICL- QA Pairs

Retrieved QA Pairs:

1. question: What is the effect of vitamin D supplementation on falls in the elderly?, answer: Vitamin D supplementation (800 IU/d) can reduce falls in the elderly.
2. question: Did vitamin D supplementation reduce the rate of falls?, answer: No, vitamin D supplementation did not reduce the rate of falls
3. question: What was the association between vitamin D use and the risk of falls?, answer: Vitamin D use was associated with a statistically significant reduction in the risk of falls.

LLM Answer: No

Example 3: Yes/No type Question

Question: Can valproic acid act as an activator of AMPK?, **Answer: Yes**

With ICL- Abstracts

Retrieved Abstracts:

1. Adenosine monophosphate - activated kinase (AMPK) plays a key role in the coordination of the heart's anabolic and catabolic pathways. It induces a cellular cascade at the center of maintaining energy homeostasis in the cardiomyocytes.. The activated AMPK is a heterotrimeric protein, separated into a catalytic - subunit (63kDa), a regulating - subunit (38kDa) and a - subunit (38kDa), which is allosterically adjusted by adenosine triphosphate (ATP) and adenosine monophosphate (AMP). The actual binding of AMP to the - subunit is the step which activates AMPK. AMPK serves also as a protein kinase in several metabolic pathways of the heart, including cellular energy sensoring or cardiovascular protection. The AMPK cascade represents a sensitive system, activated by cellular stresses that deplete ATP and acts as an indicator of intracellular ATP/AMP. In the context of cellular stressors (i.e. hypoxia, pressure overload, hypertrophy or ATP deficiency) the increasing levels of AMP promote allosteric activation and phosphorylation of AMPK. As the concentration of AMP begins to increase, ATP competitively inhibits further phosphorylation of AMPK. The increase of AMP may also be induced either from an iatrogenic emboli, percutaneous coronary intervention, or from atherosclerotic plaque rupture leading to an ischemia in the microcirculation. To modulate energy metabolism by phosphorylation and dephosphorylation is vital in terms of ATP usage, maintaining transmembrane transporters and preserving membrane potential. In this article, we review AMPK and its role as an important regulatory enzyme during periods of myocardial stress, regulating energy metabolism, protein synthesis and cardiovascular protection.
2. We have recently shown that in diabetic OVE26 mice (type I diabetes), the AMP-activated protein kinase (AMPK) is reduced along with cardiac dysfunction and decreased cardiac autophagy. Genetic inhibition of AMPK in cardiomyocytes attenuates cardiac autophagy, exacerbates cardiac dysfunction and increases mortality in diabetic mice. More importantly, we have found chronic AMPK activation with metformin, one of the most used antidiabetes drugs and a well-characterized AMPK activator, significantly enhances autophagic activity, preserves cardiac function and prevents most of the primary characteristics of diabetic cardiomyopathy in OVE26 mice, but not in dominant negative-AMPK diabetic mice. We conclude that AMPK activation protects cardiac structure and function by increasing cardiac autophagy in the diabetic heart.
3. The multikinase inhibitor sorafenib is under clinical investigation for the treatment of many solid tumors, but in most cases, the molecular target responsible for the clinical effect is unknown. Furthermore, enhancing the effectiveness of sorafenib using combination strategies is a major clinical challenge. Here, we identify sorafenib as an activator of AMP-activated protein kinase (AMPK), in a manner that involves either upstream LKB1 or CAMKK2. We further show in a phase II clinical trial in KRAS mutant advanced non-small cell lung cancer (NSCLC) with single agent sorafenib an improved disease control rate in patients using the antidiabetic drug metformin. Consistent with this, sorafenib and metformin act synergistically in inhibiting cellular proliferation in NSCLC in vitro and in vivo. A synergistic effect of both drugs is also seen on phosphorylation of the AMPK activation site. Our results provide a rationale for the synergistic antiproliferative effects, given that AMPK inhibits downstream mTOR signaling. These data suggest that the combination of sorafenib with AMPK activators could have beneficial effects on tumor regression by AMPK pathway activation. The combination of metformin or other AMPK activators and sorafenib could be tested in prospective clinical trials. Keywords: AMP-activated protein kinase; metformin; non-small cell lung cancer; salicylate; sorafenib.

LLM Answer: Yes

With ICL- QA Pairs

Retrieved QA Pairs:

1. question: What is the effect of valproic acid on HDACs?, answer: Valproic acid is an HDAC inhibitor.
2. question: What is the effect of AMPK activators on cell growth?, answer: Inhibition of cell growth.
3. question: What is a potential concern with using valproic acid in patients who require surgical intervention?, answer: It may independently impair hemostasis.

LLM Answer: No

Example 4: List type Question

Question: Which antibiotics target peptidoglycan biosynthesis?
Answer: colchicine, fosfomycin, bacitracin, vancomycin,d-cycloserine, seromycin, ampicillin, cinnamycin, ramoplanin, muraymycin, mersacidin

---

With ICL- Abstracts

Retrieved Abstracts:

1. The range of antibiotic therapy for the control of bacterial infections is becoming increasingly limited because of the rapid rise in multidrug resistance in clinical bacterial isolates. A few diseases, such as tuberculosis, which were once thought to be under control, have re-emerged as serious health threats. These problems have resulted in intensified research to look for new inhibitors for bacterial pathogens. Of late, the peptidoglycan (PG) layer, the most important component of the bacterial cell wall has been the subject of drug targeting because, first, it is essential for the survivability of eubacteria and secondly, it is absent in humans. The last decade has seen tremendous inputs in deciphering the 3-D structures of the PG biosynthetic enzymes. Many inhibitors against these enzymes have been developed using virtual and high throughput screening techniques. This review discusses the mechanistic and structural properties of the PG biosynthetic enzymes and inhibitors developed in the last decade.
2. Bacitracin and other antibiotics that inhibit late stages in peptidoglycan biosynthesis induce vancomycin resistance in a high-level, inducibly vancomycin-resistant strain of Enterococcus faecium. Exposure to bacitracin led to synthesis of the lactate-containing UDP-MurNAc-pentadepsipeptide precursor required for vancomycin resistance. These findings indicate that inhibition of peptidoglycan biosynthesis can lead to induction of vancomycin resistance and raise the possibility that multiple signals may serve to induce resistance.
3. The bacterial cell envelope includes plasma membrane, peptidoglycan, and surface layer. Peptidoglycan is unique to bacteria and the target of the most important antibiotics; here it is analyzed by mass spectrometry. Keywords: Cell wall; Cross-link; Glycoproteomics; Muropeptides; Peptidoglycan; Proteomics; Tandem mass spectrometry.

LLM Answer: bacitracin, vancomycin

With ICL- QA Pairs

Retrieved QA Pairs:

1. question: What is the primary target of ramoplanin?, answer: Peptidoglycan biosynthesis.
2. question: How do the muraymycins inhibit peptidoglycan biosynthesis?, answer: The muraymycins inhibited peptidoglycan biosynthesis.
3. question: What is the result of inhibiting peptidoglycan biosynthesis in bacteria?, answer: Induction of vancomycin resistance.

LLM Answer: vancomycin, ramoplanin, muraymycins, bacitracin

Example 5: Factoid type Question

Question: Which is the human selenoprotein that contains several Se-Cys residues?
Answer: selenoprotein p

With ICL- Abstracts

Retrieved Abstracts:

1. Selenocysteine (Sec), the 21st amino acid in protein, is encoded by UGA. The Sec insertion sequence (SECIS) element, which is the stem-loop structure present in 3' untranslated regions (UTRs) of eukaryotic selenoprotein-encoding genes, is essential for recognition of UGA as a codon for Sec rather than as a stop signal. We now report the identification of a new eukaryotic selenoprotein, designated selenoprotein M (SelM). The 3-kb human SelM-encoding gene has five exons and is located on chromosome 22 but has not been correctly identified by either Celera or the public Human Genome Project. We characterized human and mouse SelM cDNA sequences and expressed the selenoprotein in various mammalian cell lines. The 3" UTR of the human, mouse, and rat SelM-encoding genes lacks a canonical SECIS element. Instead, Sec is incorporated in response to a conserved mRNA structure, in which cytidines are present in place of the adenosines previously considered invariant. Substitution of adenosines for cytidines did not alter Sec incorporation; however, other mutant structures did not support selenoprotein synthesis. SelM is expressed in a variety of tissues, with increased levels in the brain.
2. Selenocysteine (Sec), the 21st amino acid in protein, is encoded by UGA. The Sec insertion sequence (SECIS) element, which is the stem-loop structure present in 3' untranslated regions (UTRs) of eukaryotic selenoprotein-encoding genes, is essential for recognition of UGA as a codon for Sec rather than as a stop signal. We now report the identification of a new eukaryotic selenoprotein, designated selenoprotein M (SelM). The 3-kb human SelM-encoding gene has five exons and is located on chromosome 22 but has not been correctly identified by either Celera or the public Human Genome Project. We characterized human and mouse SelM cDNA sequences and expressed the selenoprotein in various mammalian cell lines. The 3" UTR of the human, mouse, and rat SelM-encoding genes lacks a canonical SECIS element. Instead, Sec is incorporated in response to a conserved mRNA structure, in which cytidines are present in place of the adenosines previously considered invariant. Substitution of adenosines for cytidines did not alter Sec incorporation; however, other mutant structures did not support selenoprotein synthesis, demonstrating that this new form of SECIS element is functional. SelM is expressed in a variety of tissues, with increased levels in the brain.
3. Selenoproteins are a diverse group of proteins that contain selenocysteine (Sec), the 21st amino acid. In the genetic code, UGA serves as a termination signal and a Sec codon. This dual role has precluded the automatic annotation of selenoproteins. Recent advances in the computational identification of selenoprotein genes have provided a first glimpse of the size, functions, and phylogenetic diversity of eukaryotic selenoproteomes. Here, we describe the identification of a selenoprotein family named SelJ. In contrast to known selenoproteins, SelJ appears to be restricted to actinopterygian fishes and sea urchin, with Cys homologues only found in cnidarians. SelJ shows significant similarity to the jellyfish J1-crystallins and with them constitutes a distinct subfamily within the large family of ADP-ribosylation enzymes. Consistent with its potential role as a structural crystallin, SelJ has preferential and homogeneous expression in the eye lens in early stages of zebrafish development. The unusually highly restricted phylogenetic distribution of SelJ, and the comparative analysis of eukaryotic selenoproteomes reveal the diversity and functional plasticity of selenoproteins and point to evolution of the use of Sec in proteins.

LLM Answer: selm

With ICL- QA Pairs

Retrieved QA Pairs:

1. question: How many selenocysteine residues do most selenoproteins contain?, answer: A single selenocysteine residue. 2. question: How many selenoproteins are present in the human selenoproteome?, answer: The human selenoproteome consists of 25 selenoproteins. 3. question: What is the composition of selenoprotein P in addition to its high number of selenocysteine residues?, answer: The protein is cysteine- and histidine-rich.

LLM Answer: selenoprotein p