Harnessing Collective Intelligence of LLMs for Robust Biomedical QA: A Multi-Model Approach

Notebook for the BioAsq Lab at CLEF 2025

Dimitra Panou[1,2], Alexandros C. Dimopoulos[2,3], Manolis Koubarakis[1,4] and Martin Reczko[2,*]

[1]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
[2]Institute for Fundamental Biomedical Science, Biomedical Sciences Research Center "Alexander Fleming", Greece
[3]Department of Informatics & Telematics, School of Digital Technology, Harokopio University, Greece
[4]Archimedes, Athena Research Center, Greece

Abstract

Biomedical text mining and question-answering are essential yet highly demanding tasks, particularly in the face of the exponential growth of biomedical literature. In this work, we present our participation in the 13th edition of the BioAsq challenge, which involves biomedical semantic question-answering for Task 13b and biomedical question-answering for developing topics for the Synergy task. We deploy a selection of open-source large language models (LLMs) as retrieval-augmented generators to answer biomedical questions. Various models are used to process the questions. A majority voting system combines their output to determine the final answer for Yes/No questions, while for list and factoid type questions, the union of their answers in used. We evaluated 13 state-of-the-art open source LLMs, exploring all possible model combinations to contribute to the final answer, resulting in tailored LLM pipelines for each question type. Our findings provide valuable insight into which combinations of LLMs consistently produce superior results for specific question types. In the four rounds of the 2025 BioAsq challenge, our system achieved notable results: in the Synergy task, we secured 1st place for ideal answers and 2nd place for exact answers in round 2, as well as two shared 1st places for exact answers in rounds 3 and 4.

Keywords

Biomedical Question Answering, BioAsq, Large Language Models, Retrieval-augmented generation

1. Introduction

Large Language Models (LLMs) are transforming numerous fields, but their development and application face distinct challenges tied to accessibility and capabilities. Closed-source models, such as GPT-models [1], often maintained by large corporations, demonstrate advanced capabilities but lack public accessibility and transparency. In contrast, open-source LLMs, grant accessibility, transparency and facilitate fine-tuning and integration into customizable pipelines.

Despite the remarkable progress in LLMs, which are increasingly capable of tackling complex tasks, biomedical QA remains a uniquely challenging field. Effective QA systems must not only retrieve relevant information handling domain-specific terminology, but also discern when to recommend a single 'best' option and when to present multiple perspectives. Avoiding mistakes is critical in this field, as decisions often have direct consequences on human health. Reliable question-answering systems must support experts in exploring these critical issues with accuracy and depth. In fast evolving fields such as drug discovery and molecular biology, where new findings appear constantly and may contradict earlier work, robust tools help professionals stay informed, avoid errors, and make evidence-based decisions that truly advance science and healthcare.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
panou@fleming.gr (D. Panou); dimopoulos@fleming.gr (A. C. Dimopoulos); koubarak@di.uoa.gr (M. Koubarakis); reczko@fleming.gr (M. Reczko)
https://dsit.di.uoa.gr/dimopoulos-cv/ (A. C. Dimopoulos); https://cgi.di.uoa.gr/~koubarak/ (M. Koubarakis); https://www.fleming.gr/research/ifbr/staff-scientists/reczko-lab (M. Reczko)
0000-0002-9824-4489 (D. Panou); 0000-0002-4602-2040 (A. C. Dimopoulos); 0000-0002-1954-8338 (M. Koubarakis); 0000-0002-0005-8718 (M. Reczko)

1.1. The BioAsq Challenge

The BioAsq challenge has played a central role in advancing biomedical question answering (QA), particularly through its tasks. Task B requires systems to retrieve relevant documents and snippets and then generate precise answers to biomedical questions, while the Synergy track adds further complexity by introducing an interactive, feedback-based QA setting, simulating real-world clinical scenarios. These tasks push the limits of current Retrieval-Augmented Generation (RAG) systems, demanding high precision in both information retrieval and generation.

Although large language models (LLMs) are becoming increasingly efficient today, the BioAsq challenge demonstrates that achieving accurate results relies on well-structured and carefully designed QA pipelines. Effective systems use hybrid retrievers [2], domain-specific encoders [3, 4], and fine-tuned generators tailored for biomedical text. Pipelines often include re-ranking steps, prompt tuning [5], and targeted post-processing to handle subtasks like yes/no classification or list generation. These components are critical to ensure relevance, factuality and clarity, something end-to-end LLMs still struggle with in complex domains.

Our lab has participated in BioAsq Challenge for three consecutive years. During this period, we experimented with various methodologies to enhance the document selection task. We began by developing our own model, ELECTROLBERT [6], and later fine-tuned a GAN [7] combined with sparse BM25 for document ranking. In our most recent iteration, we transitioned to leveraging existing models for document retrieval, systematically exploring and comparing sparse, dense, and hybrid approaches [8].

Despite improvements in our pipelines, we observed that the Mean Average Precision (MAP) in Phase A remains relatively low for all participants. This is mainly because selecting the right documents has become more difficult as the document collection keeps growing. Matching the retrieved documents with the small set chosen by experts remains a challenge. On the other hand, there is still room to improve how answers are generated from the retrieved documents. That's why in this work, we focus on improving the generation of 'ideal' and 'exact' answers in Phase B.

2. Methodology

2.1. Synergy

For the Synergy challenge [9], we used the same methods as for the final submissions of the BioAsq12 competition, with the notable addition of a DeepSeek-R1 model variant for the generation of exact and ideal answers [10]. The improved language generation skills of this model led to a notable improvement in the free text required for the ideal answers.

2.2. Task 13b, phase A: Document Retrieval & Snippet Identification

In Phase A and Aplus of the BioAsq challenge, the organizers release biomedical questions curated by experts [11, 12] that have to be processed within a strict 24-hour interval. For Phase A participants have to retrieve and submit up to 10 relevant documents per question, utilizing abstracts sourced from the PubMed[1] database. Based on the retrieved documents, participants must then identify and extract the most relevant snippets.

For document retrieval in Phase A, we adopted a standard approach shown to deliver strong performance in previous work [8], specifically using the BM25 [13, 14] ranking algorithm enhanced with pseudo-relevance feedback from RM3 [15]. From this setup, we initially retrieved the top 50 candidate documents and subsequently re-ranked them based on the relevance of their associated snippets. Snippet prediction, which extracts the most semantically relevant snippet from each of the top 10 retrieved documents, is performed as described in [8].

[1] https://pubmed.ncbi.nlm.nih.gov/

2.3. Task13b, phase A+ / phase B: exact answer generation

In Phase A+ participants will submit exact and/or ideal answers before the expert selected (gold) documents and snippets (released in Phase B) are known. It serves as a baseline to compare with Phase B, where feedback is provided to guide system improvement. Each participant must rely on their own predicted documents and snippets for subsequent processing. They have 24 hours to submit their results, which include documents, snippets, exact and 'ideal' answers, based on the provided test set. For document selection, we followed the same procedure as in Phase A. To generate the exact and 'ideal' answers, we used both the predicted snippets and the full abstracts as input.

In Phase B, participants are required to submit exact answers for Yes/No, List, and Factoid questions, as well as ideal answers for summary-type questions. This phase uses gold-standard documents and snippets. In Phase B, we explored three distinct approaches for generating exact answers, as illustrated in Figure1.

The first approach (Figure 1, method a.) utilizes the extracted snippets from the given (golden) documents, incorporating them directly into the prompt to generate answers for each question. This method has been used in our previous submissions and is commonly adopted by participants in the BioAsq challenge. It is computationally efficient, as the snippets are typically short, ranging from a few words to two sentences. The second approach (Figure 1, method b.) uses the full abstracts of the top 10 most relevant documents. The prompt is constructed by combining the question with these abstracts in the following format: `text = <Abstract 1>, <Abstract 2>, ..., <Abstract 10>`, as shown in Appendix A. This method provides broader contextual information and outperformed the first approach in our evaluations.

The third approach (Figure 1, method c.) builds upon the second by additionally incorporating any relevant documents identified during the extended document retrieval process in Phase A+. These supplementary documents are appended to the original list, further enriching the input context provided to the model and potentially including documents not selected by the experts who created the gold answers.

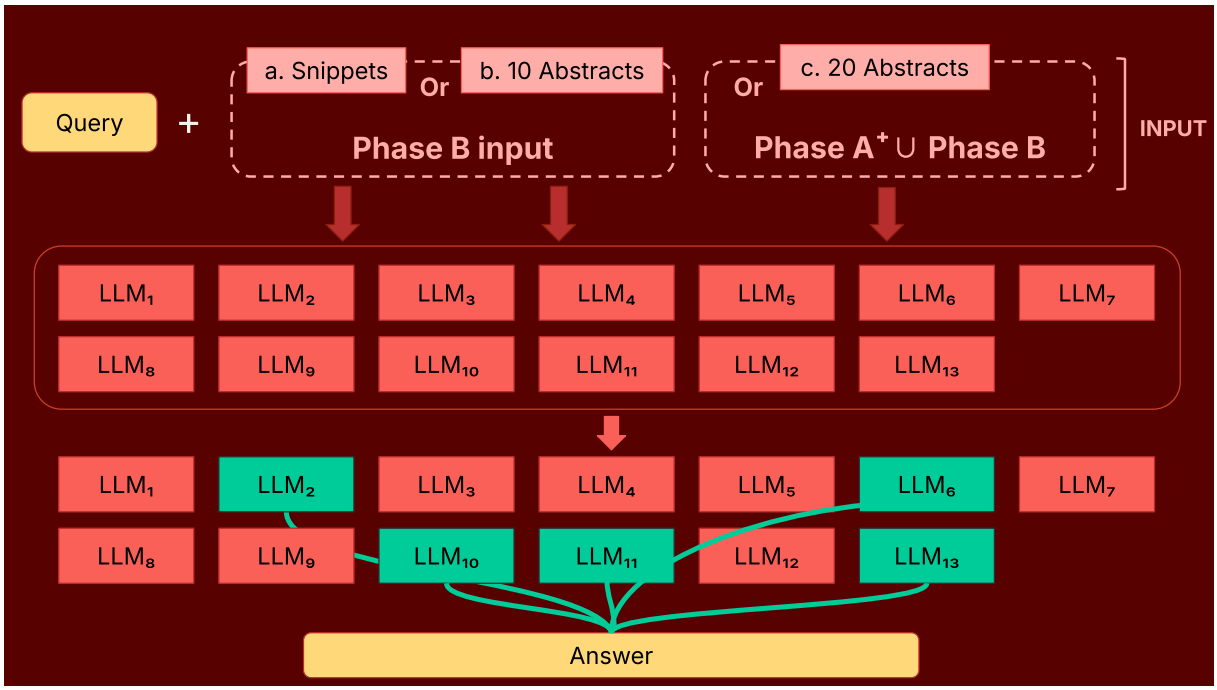

Figure 1: Processing during phase A+ and phase B. a. Prediction based on given snippets, b. Prediction based on given abstracts, c. Prediction based on given and predicted abstracts. The Query and one of the three input alternatives are used to form the prompts for the LLMs. The collection of all LLMs used to find optimal sets based on the training set is shown in the middle, the identified optimal subset used for prediction is shown in green at the bottom.

#	Abbreviation	Model Name	Edition / Quantization	Parameter Size
1	Reflection	Reflection[1]	latest	70B
2	L3.1	LLaMA 3.1[2]	Q4 / ctx8192	70B
3	L3.3	LLaMA 3.3[3]	latest	70B
4	Mixtral	Mixtral[4]	ctx8192:latest	8x7B
5	Qwen14	Qwen3:14B[6]	latest	14B
6	Qwen30	Qwen3:30B-A3B[7]	A3B	30B
7	Qwen32	Qwen3:32B[8]	latest	32B
8	Yi	Yi[8]	latest	34B
9	Smaug	Smaug:72B[9]	Q4_K_M, quantized 4 bit	72B
10	DS-R1d-70B	DeepSeek-R1-Distill-Llama-70B[10][10]	Q8_0:latest, quantized 8 bit	70B
11	Phi3	Phi-3 Medium[11]	latest	14B
12	Phi4	Phi-4[12]	latest	14B
13	Aya	Aya:35B[13]	latest	35B

Table 1
Large language models used individually and in all combinations of them.

To improve the answer performance measures, we employed an LLM 'farming' strategy, which we initially implemented last year for Yes/No questions. This strategy utilizes a diverse ensemble of complementary open-source large language models. In the present study, we extend this strategy to all exact answer types, aggregating the union of answers from multiple LLMs for factoid and list questions.

Using the BioAsq11 and BioAsq12 training set, we evaluated 13 state-of-the-art LLMs using Ollama [16] and LM Studio [17], systematically analyzing their individual performance, as well as all possible combinations of models for each type of question. This experimentation allowed us to construct an optimal 'farm' of models for each category of question. Due to the long run-time of these optimizations, our submissions in the competition did not represent the finally best performing system. For all types of questions, the optimization revealed novel combinations of models with higher performance than any single LLM.

2.3.1. Optimal factoid question answering subsets

For the 13 LLMs listed in table 1, there are $|\{S_i, i \in [1, 2^{13}]\}| = 8191$ different subsets. The 13 LLMs predict sets of factoids for all questions of BioAsq11 and BioAsq12 separately. All predictions are compared to the golden answers of BioAsq11 and BioAsq12, respectively. The factoid sets for each LLM in S_i are combined to form a union for each question. Since the factoids should be ordered by relevance and only the top 5 most relevant should be returned, the combination of the factoids considers the relevance scores that are also returned by each LLM. The performance of these sets is evaluated with the usual Mean Reciprocal Rank (MRR) measure. The MRR values, averaged over four rounds, for each S_i are shown in the scatterplot 3 that visualizes the performances in BioAsq11 and BioAsq12. The color of each dot indicates the size of the LLM set. Single LLMs are shown in red, the largest sets containing 6 LLMs are shown in blue. It should be noted that in all cases the sets with more than 6 LLMs had the same performance as a 'kernel' set of 6 LLMs and are not contained in the plot. As all high-performing sets

[1]https://huggingface.co/mattshumer/ref_70_e3
[2]https://ollama.com/library/llama3.1:70b
[3]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[4]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[5]https://huggingface.co/Qwen/Qwen3-14B
[6]https://huggingface.co/Qwen/Qwen3-30B-A3B
[7]https://huggingface.co/Qwen/Qwen3-32B
[8]https://huggingface.co/01-ai/Yi-34B
[9]https://huggingface.co/senseable/Smaug-72B-v0.1-gguf/blob/main/Smaug-72B-v0.1-q4_k_m.gguf
[10]https://huggingface.co/unsloth/DeepSeek-R1-Distill-Llama-70B-GGUF
[11]https://ollama.com/library/phi3:medium
[12]https://ollama.com/library/phi4
[13]hhttps://ollama.com/library/aya:35b

occur in blue tones, it can be clearly seen that all red dots for single LLMs had worse performances than any union of at least 4 LLMs (see also figure 2). Consistently, the highest performances are obtained with unions of 6 LLMs (DS-R1d-70B, llama3.3, qwen3-14b, qwen3-32b, reflection, smaug). These observations can be made for both BioAsq11 and BioAsq12 independently, indicating no training set specificities. The finding that larger unions give better results is very likely due to the complementarity of the answers of the different LLMs. There can be many cases where one method finds a highly relevant factoid that another method does not identify at all or as a close miss. The merging strategy that uses the confidence scores supports these situations.

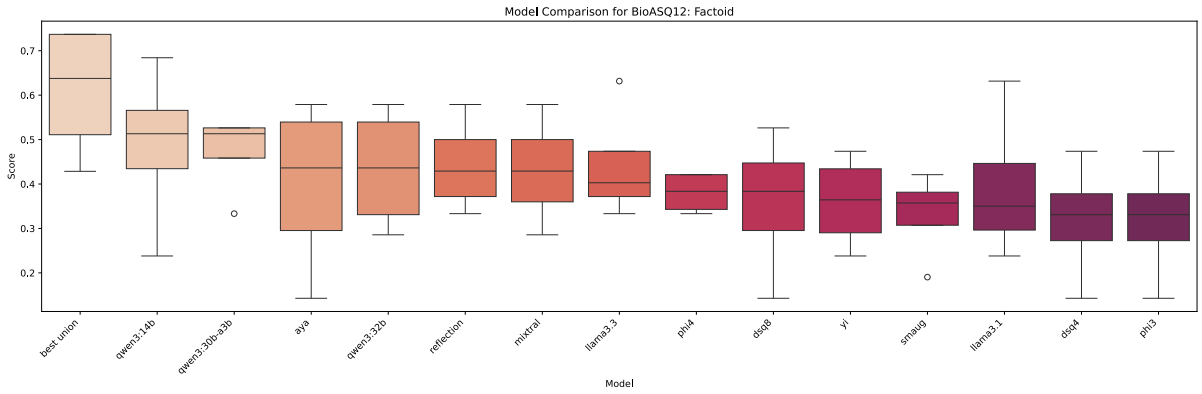

Figure 2: Comparison of single LLMs and the optimal union of LLMs for factoid questions performance tested on BioAsq12 datasets.

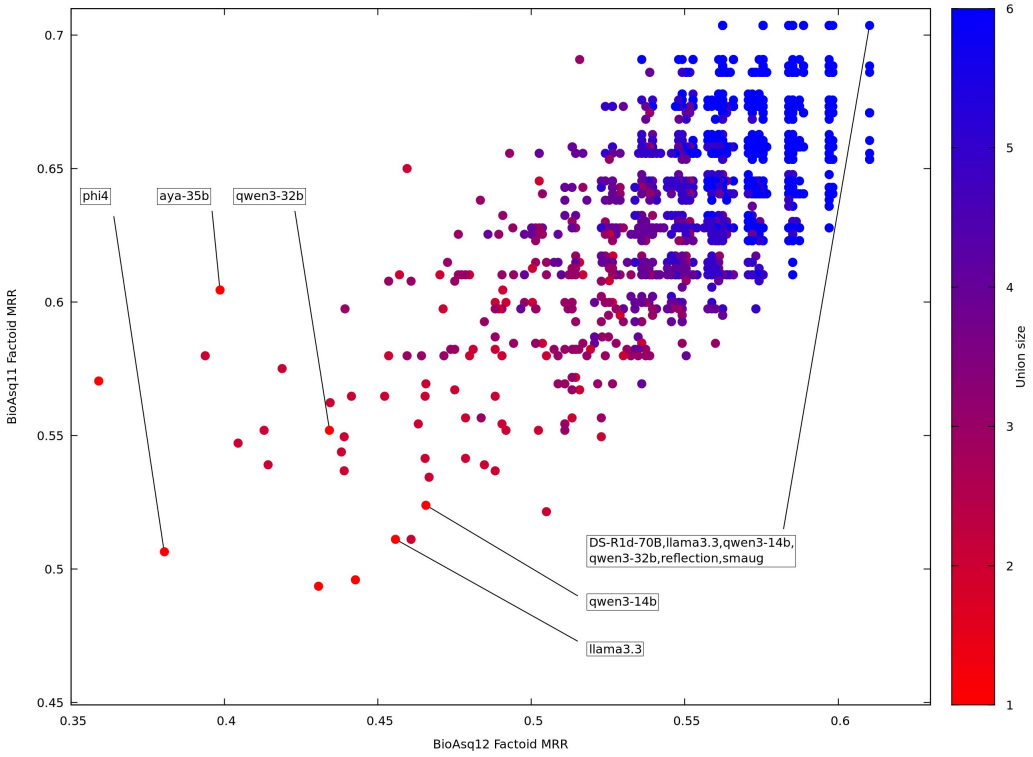

Figure 3: Model unions for Factoid questions performance tested on BioAsq11 & BioAsq12 datasets

Factoid deduplication As forming the union might introduce multiple occurrences of exactly the same factoid phrase or semantically similar phrases, we investigated a simple deduplication procedure. Each factoid phrase is embedded with a standard transformer (all-MiniLM-L6-v2) and the cosine

similarity of the embedding between all factoids is measured. With different thresholds for the cosine similarity, semantically similar phrases can be removed from the set. The MRR performance with different thresholds for the LLM subset with the best performance on BioAsq12 is shown in figure 4. It can be observed that deduplication does not improve MRR performance, consistent with our observation that larger subsets in general have higher MRR performances.

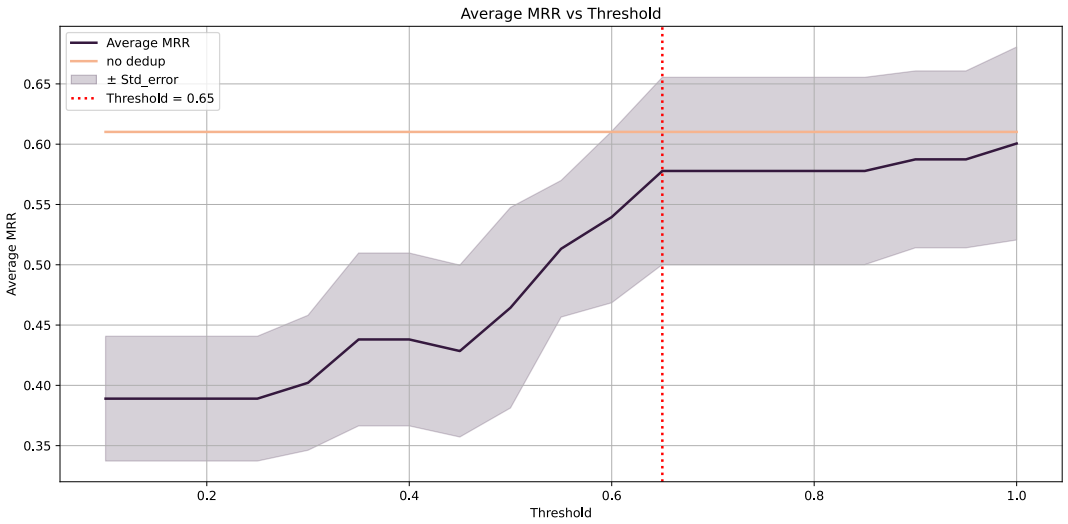

Figure 4: Deduplication performance for factoid questions

2.3.2. Optimal list question answering subsets

A procedure similar to the processing of the factoid questions is used for list type questions. As no relevance order and no limit is required for the list items in the answer, the set of list items is the simple union of the list items predicted by each LLM in the subset S_i. The usual performance measure for list type questions is the F-Measure, which is the harmonic mean of precision and recall. With a growing size of the list items in the union for each additional LLM in the subset S_i, the chance of false positive items increases and precision decreases. In the scatterplot showing the F-Measure scores (averaged over four rounds) for BioAsq11 and BioAsq12 in figure 6, it can be clearly seen that the large subsets with more than 7 LLMs have significantly lower performance than the smaller subsets, independently of the BioAsq dataset used. However, as detailed in figure 5, several specific combinations, such as the set [DS-R1d-70B, L3.3, Qwen14] have better performance than any of the single LLMs.

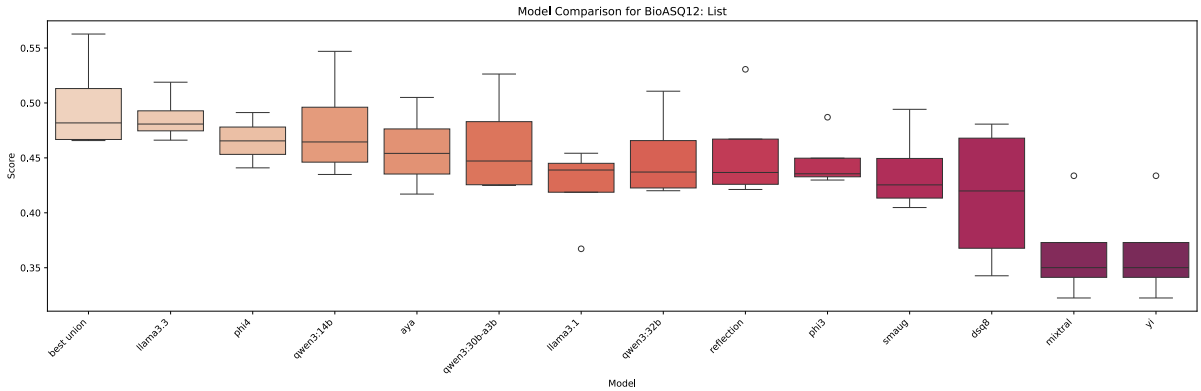

Figure 5: Comparison of single LLMs and the optimal union of LLMs for list questions performance tested on BioAsq12 datasets.

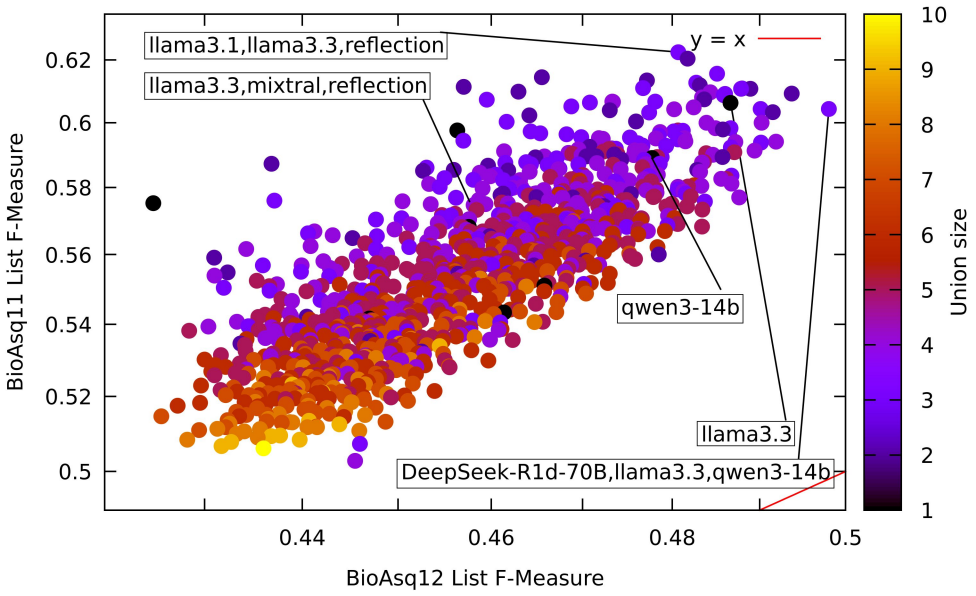

Figure 6: Model unions for List questions performance tested on BioAsq11 & BioAsq12 datasets

List deduplication The same deduplication procedure used for factoids is also evaluated for the union of list items. For the subset with the best performance on the BioAsq12 set, the F-Measure performance for different thresholds for the cosine similarity is shown in figure 7. It can be observed that two levels of deduplication achieve higher performance than without deduplication, with an optimal F-Measure values at a threshold of 0.76. A threshold of 0.7 was used for the list type submission in the BioAsq13 competition. Comparing to the deduplication results for factoid questions, we confirm the general effect that deduplication improves the F-measure (used to evaluate list questions) by increasing precision without harming recall, but it does not change the Mean Reciprocal Rank (MRR, used to evaluate factoid questions) because the position of the first correct result usually remains unchanged.

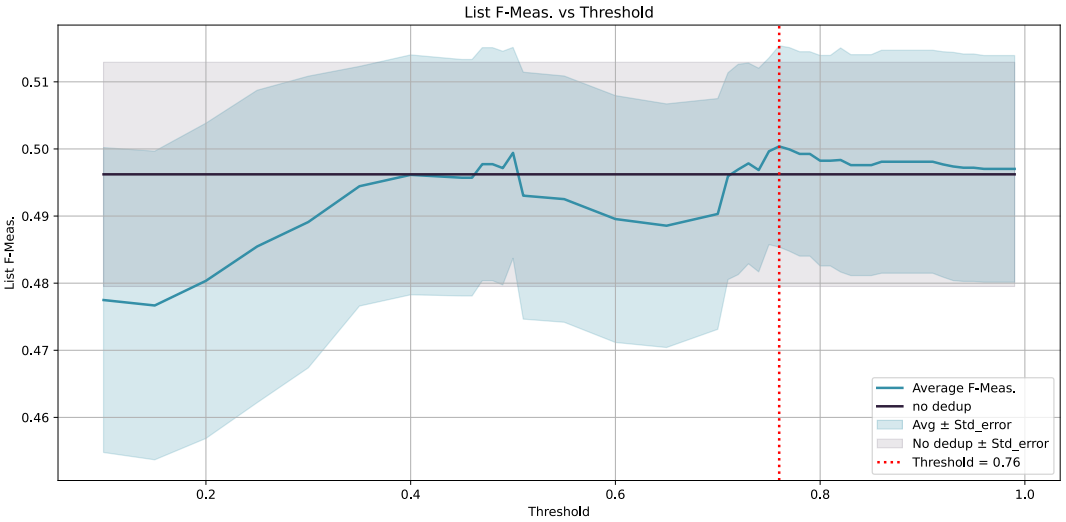

Figure 7: Deduplication performance for list questions

2.3.3. Optimal jury sets answering Yes/No questions

The concept of using a jury (or 'farm') of LLMs was introduced by our lab for BioAsq12 [8]. Here we further optimize this by evaluating all possible combinations of LLMs and adding more recent LLMs. A subset S_i of LLMs generates an answer by counting the number of 'Yes' and 'No' outcomes for each participating LLM. The final answer will be 'Yes' if there are a higher or equal number of 'Yes' outcomes than 'No' outcomes. The performances of the different subsets with the usual macroF1 measure (averaged over four rounds) is shown for BioAsq11 and BioAsq12 in figure 9. The discrete nature of this question type leads to more discrete performance levels that are visualized in the plot by applying a small jitter. As in the case of the list type questions, it can be seen that there are several combinations of a few LLMS like [Aya, Qwen32, Smaug] that outperform any of the individual LLM alone.

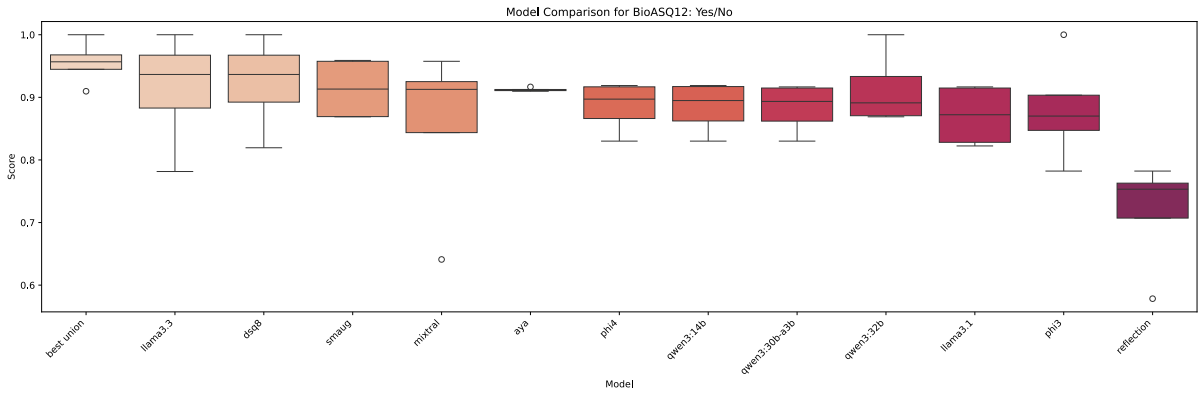

Figure 8: Comparison of single LLMs and the optimal union of LLMs for Yes/No questions performance tested on BioAsq12 datasets.

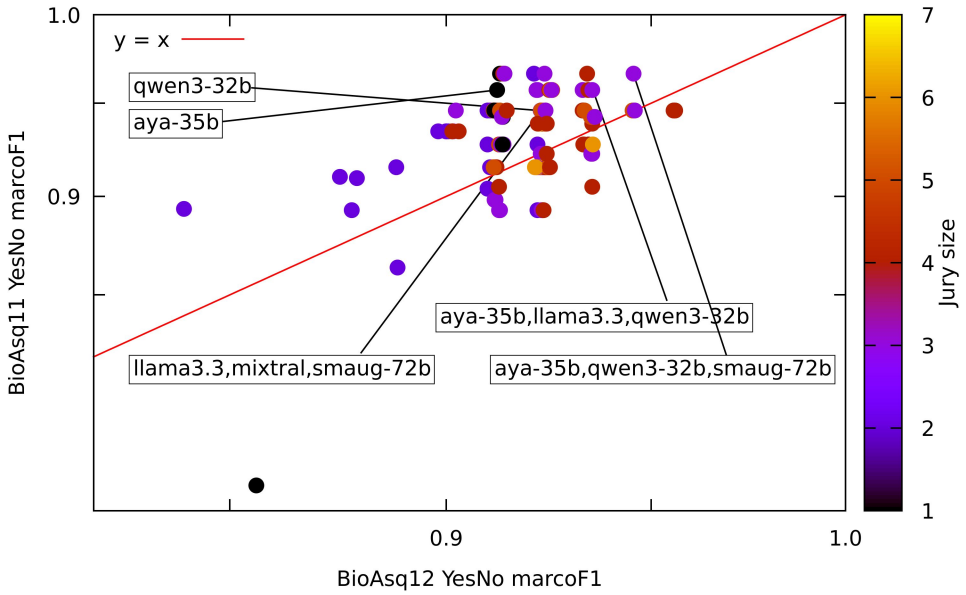

Figure 9: Model juries for Yes/No questions performance tested on BioAsq11 & BioAsq12 datasets

3. Results

Here, we present the results across all our participations (Synergy and Phase B) in the BioAsq competition. All systems submitted throughout the different phases are listed under the name Fleming-X in the

results. The evaluation of systems participating in the BioAsq competition Task B varies [18] based on the question type.

3.1. Synergy Results

In the four rounds of the Synergy 2025 BioAsq challenge, our system achieved notable results: first place in round 2 for 'ideal answers' and second place in rounds 3 and 4.
As shown in Table 2, the evaluation for 'Exact' answers includes multiple measures across three question types: Accuracy and Macro F1 for Yes/No questions, Mean Reciprocal Rank (MRR) and Lenient Accuracy for Factoid questions, and Mean Precision and F-measure for List questions. The overall position per system in each batch is calculated based on a combination of Macro F1 (Yes/No), MRR (Factoid), and F-measure (List). While the top-ranked systems achieve the highest combined scores, the Fleming submissions stand out with stronger results in the List category and competitive performance in Factoid.
Table 3 presents the performance of systems on the 'Ideal answers' task, as assessed through manual evaluation. The scores reflect human judgments across four criteria: Readability, Recall, Precision, and Repetition, with the final Mean Manual score representing their average. In Batch 2, the Fleming system achieved the highest overall score, ranking 1st, while in Batches 3 and 4, it remained competitive with particularly strong Recall and Repetition scores, securing 2nd and 4th place respectively.

Table 2
Synergy: Exact answers performance measured by the combination of Macro F1, MRR, and F-measure

Batch	Position	System	Yes/No		Factoid		List	
			Macro F1	Rank	MRR	Rank	F-Measure	Rank
Batch 2	1/10	**dmiip2024_1**	1.000	1	0.4286	1	0.2467	1
	2/10	dmiip2024_2	1.000	1	0.2857	2	0.2000	4
	3/10	Fleming-1 (ours)	0.8571	2	0.2857	2	**0.2100**	3
	7/10	Fleming-1	0.5333	5	0.2857	2	**0.2333**	2
Batch 3	1/13	**dmiip2024_4**	0.899	2	0.5000	1	0.2495	3
	2/13	Fleming-3	0.899	2	0.2500	3	**0.2634**	**1**
	4/13	Fleming-1	0.7917	3	0.2500	3	**0.2634**	**1**
	5/13	Fleming-2	0.7917	3	0.2500	3	**0.2634**	**1**
Batch 4	1/15	**sinai_uja_RAG**	0.899	2	0.4000	2	0.2667	2
	2/15	**Fleming-4**	0.7917	3	0.4000	2	**0.3536**	1
	3/15	**Fleming-1**	0.7917	3	0.4000	2	**0.3536**	1
	4/15	**Fleming-2**	0.7917	3	0.4000	2	**0.3536**	1
	5/15	**Fleming-3**	0.7917	3	0.4000	2	**0.3536**	1

Table 3
Synergy: 'Ideal answers' performance measured by mean of manual score.

Batch	Position	System	Readability	Recall	Precision	Repetition	Mean Manual
Batch 2	1/10	**Fleming-1 (ours)**	4.06	4.24	3.88	4	4.045
	2/10	dmiip2024_1	4.06	3.79	3.79	4.39	4.0075
	5/10	Fleming-2	3.52	4.06	3.27	3.27	3.730
Batch 3	1/13	**dmiip2024_1**	4.57	4.61	4.43	4.57	4.545
	6/13	Fleming-1	4.31	**4.73**	4.06	4.47	4.3925
	7/13	Fleming-2	4.31	**4.73**	4.06	4.47	4.3925
	8/13	Fleming-3	4.31	**4.73**	4.06	4.47	4.3925
Batch 4	1/15	**dmiip2024_1**	4.47	4.53	4.24	4.49	4.4325
	9/15	Fleming-1	4.05	4.51	3.69	4.09	4.085
	10/15	Fleming-3	4.05	4.51	3.69	4.09	4.085
	11/15	Fleming-4	3.98	4.36	3.56	4.09	3.998
	12/15	Fleming-2	3.91	4.51	3.47	3.96	3.963

3.2. Task 13b: Phase A

3.2.1. Document retrieval

In table 4 the preliminary performances of our document retrieval submissions for the BioAsq13 competition are listed. The final and official results, will be available shortly before the BioAsq13 workshop, after the manual assessment of all system responses by the BioAsq experts and the enrichment of the respective ground truth with potential additional correct elements.

Table 4
Phase A: System performance for Document retrieval measured as mean average precision (MAP)

Batch	Position	System	Mean Precision	Recall	F-Measure	MAP	GMAP
Batch 1	1/51	**bioinfo-4**	0.1047	0.5043	0.1605	0.4246	0.0104
	24/51	Fleming-1 (ours)	0.0606	0.3863	0.1005	0.2716	0.0020
Batch 2	1/42	**Baseline top 10**	0.0976	0.5093	0.1546	0.4425	0.0096
	18/42	Fleming-2	**0.0993**	0.4333	0.1477	0.3066	0.0026
	19/42	Fleming-3	**0.0993**	0.4333	0.1477	0.3066	0.0026
	22/42	Fleming-1	0.0861	0.4333	0.1342	0.2957	0.0026
Batch 3	1/47	**bioinfo-1**	0.0941	0.4228	0.1445	0.3236	0.0059
	25/47	Fleming-1	0.0697	0.3105	0.1064	0.1794	0.0009
Batch 4	1/79	**bioinfo-1**	0.06	0.2512	0.0927	0.1801	0.0008
	24/79	Fleming-2	0.0383	0.155	0.05957	0.09427	0.0002
	25/79	Fleming-1	0.0383	0.155	0.0595	0.0863	0.0002

3.3. Task 13b: Phase A+ and Phase B

3.3.1. Exact answer prediction

The tables reporting the Phase A+ (Table 5) and Phase B (Table 6) results of BioAsq13, for exact answers provide a comparative view of our submitted systems. In each batch, the first row corresponds to the top-ranked competitor. For each question type, we report a corresponding evaluation metric: Macro F1 for Yes/No, MRR for Factoid, and F-Measure for List. The systems are ranked per metric and the total rank is computed as the sum of these individual ranks, providing an overall measure of performance across all type of questions. The final position according to the total rank and the total number of submissions is indicated in the column 'Position'. Our systems demonstrated competitive performance, particularly in the Yes/No and Factoid categories.

3.3.2. Ideal answer prediction

Regarding the evaluation of the ideal answer for both Phase A+ and Phase B of Task 13b, we are currently waiting for the release of the scores manually assigned by the BioAsq experts, which are expected to be published shortly before the CLEF workshop in September. We note that all results for Task 13b remain provisional, as small corrections may still be applied by question curators prior to the workshop.

Table 5
Phase A+: Exact answers performance measured by the combination of Macro F1, MRR, and F-measure

Batch	Position	System	Yes/No		Factoid		List	
			Macro F1	Rank	MRR	Rank	F-Measure	Rank
Batch 1	1/56	**UR-IW-2**	**1.000**	1	0.3782	5	0.2567	1
	20/56	Fleming-3 (ours)	0.9328	2	0.3186	11	0.144	26
	21/56	Fleming-2	0.9328	2	0.3186	11	0.144	26
	31/56	Fleming-1	0.9328	2	0.3186	11	0.1296	33
Batch 2	1/49	**Baseline top 20**	0.9328	3	0.463	6	0.388	1
	29/49	Fleming-1	**0.9377**	**2**	0.2790	20	0.2242	25
	32/49	Fleming-2	0.9328	3	0.2790	20	0.2242	25
Batch 3	1/58	**IR3**	0.6944	11	0.3500	2	0.4313	4
	6/58	Fleming-2	**0.8182**	**6**	0.3125	4	0.3565	15
	19/58	Fleming-1	0.6563	13	0.2625	11	0.3565	15
Batch 4	1/67	**Baseline top 20**	0.8595	4	0.4318	8	0.2977	2
	31/67	Fleming-1	**0.8595**	**4**	0.2803	19	0.2425	24
	32/67	Fleming-4	**0.8595**	**4**	0.2780	20	0.2433	23
	53/67	Fleming-2	**0.8595**	**4**	0.2818	18	0.1578	50
	60/37	Fleming-3	0.7068	17	0.2818	18	0.1578	50

Table 6
Phase B: Exact answers performance measured by the combination of Macro F1, MRR, and F-Measure

Batch	Position	System	Yes/No		Factoid		List	
			Macro F1	Rank	MRR	Rank	F-Meas.	Rank
Batch 1	1/72	**2025-DMIS-KU-3**	0.9328	2	**0.5962**	**1**	0.5913	2
	6/72	Fleming-3 (ours)	0.9244	3	0.5577	2	0.5384	14
	13/72	Fleming-1	**1.0000**	**1**	**0.5962**	**1**	0.5290	20
	15/72	Fleming-2	0.9244	3	0.5962	1	0.5290	20
Batch 2	1/72	**dmiip2024_4**	**1.0000**	**1**	0.5926	6	0.6152	1
	25/72	Fleming-2	**1.0000**	**1**	0.4704	19	0.5356	19
	27/72	Fleming-3	**1.0000**	**1**	0.5148	15	0.5210	24
	35/72	Fleming-1	**1.0000**	**1**	0.4704	19	0.5210	24
Batch 3	1/66	**EP-1**	**0.9394**	**1**	0.4625	4	0.6331	2
	27/66	Fleming-1	**0.9394**	**1**	0.2717	23	0.5638	22
	38/66	Fleming-2	0.8706	3	0.3083	19	0.4832	40
	47/66	Fleming-4	**0.9394**	**1**	0.3225	18	0.4595	48
	49/66	Fleming-3	**0.9394**	**1**	0.3083	19	0.4595	48
Batch 4	1/79	**2025-DMIS-KU-4**	0.9487	3	0.6136	2	0.6328	3
	40/79	Fleming-1	0.9097	4	0.4697	10	0.4697	42
	58/79	Fleming-5	**0.9532**	**2**	0.3311	16	0.3743	55
	63/79	Fleming-4	0.9023	5	0.3250	17	0.3743	55
	64/79	Fleming-2	0.9023	5	0.3250	17	0.3208	59
	65/79	Fleming-3	0.8595	8	0.3250	17	0.3208	59

4. Conclusion and Future Work

In this work, we presented a robust and extensible methodology for biomedical question answering within the BioAsq challenge framework. A key innovation in our methodology is the application and generalization of an LLM 'farming' strategy, initially developed for Yes/No questions [8], to all exact question types. By systematically evaluating 13 state-of-the-art open LLMs and exhaustively testing all possible model combinations, we created optimized model farms for Yes/No, factoid, and list type questions. Our results show that combining multiple models improves performance in each case. For Factoid questions, the best results came from combining six different LLMs. This pattern was consistent across both the BioAsq11 and BioAsq12 datasets, suggesting that the improvement wasn't specific to the training data but rather due to the different strengths of each model working together. The top-performing combinations are shown in Table 7. For List questions, using too many models actually reduced performance. The best results came from small groups of about three models, for example, [DS-R1d-70B, L3.3, Qwen14], which outperformed all single models. For Yes/No questions, smaller combinations also worked best. Groups of three to four models, like the jury of Aya, Qwen32, Smaug, outperformed individual models. In summary, combining LLMs can improve performance, but the optimal number of models depends on the question type.

Moving forward, we plan to expand our evaluation to include more state of the art open-source LLMs, incorporate more confidence scoring mechanisms across model outputs to better weigh and reconcile conflicting answers and release our question answering system to support reproducibility and foster collaboration in the open LLM community.

Acknowledgments

We thank the anonymous reviewers for their valuable questions and comments. We express our gratitude to the BioAsq challenge organizers for organizing the event and offering continuous support. The GPU computations were executed on two servers acquired as part of project ID 16624, titled "Creation - Expansion - Upgrading of the Infrastructures of research centers supervised by the General Secretariat for Research and Innovation (GSRI)" with the code MIS 5161770. This project received funding under the "National Recovery and Resilience Plan Greece 2.0".

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini, Copilot in order to: Grammar and spelling check, paraphrase and reword. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, et al. (281 authors), Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774.

[2] P. Mandikal, R. Mooney, Sparse meets dense: A hybrid approach to enhance scientific document retrieval, 2024. URL: https://arxiv.org/abs/2401.04055.

[3] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682.

[4] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2020. arXiv:arXiv:2007.15779.

[5] S. Ateia, U. Kruschwitz, Is chatgpt a biomedical expert? – exploring the zero-shot performance of current gpt models in biomedical tasks, 2023. URL: https://arxiv.org/abs/2306.16108. arXiv:2306.16108.
[6] M. Reczko, Electrolbert: Combining replaced token detection and sentence order prediction., in: CLEF (Working Notes), 2022, pp. 335–340.
[7] D. N. Panou, M. Reczko, Semi-supervised training for biomedical question answering., in: CLEF (Working Notes), 2023, pp. 152–158.
[8] D. Panou, A. Dimopoulos, M. Reczko, Farming open llms for biomedical question answering, CLEF Working Notes (2024). URL: https://doi.org/10.5281/zenodo.13683433.
[9] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 13b and Synergy13 in CLEF2025, in: CLEF 2025 Working Notes, 2025.
[10] DeepSeek-R1-Distill-Llama-70B-GGUF, Huggingface, 2025. URL: https://huggingface.co/second-state/DeepSeek-R1-Distill-Llama-70B-GGUF.
[11] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association, 2025.
[12] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, et al., BioASQ at CLEF2025: The Thirteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: European Conference on Information Retrieval, Springer, 2025, pp. 407–415.
[13] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019.
[14] S. Robertson, S. Walker, M. Beaulieu, Experimentation as a way of life: Okapi at trec, Information Processing Management 36 (2000) 95–108. URL: https://www.sciencedirect.com/science/article/pii/S0306457399000461.
[15] V. Lavrenko, W. B. Croft, Relevance based language models, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, Association for Computing Machinery, New York, NY, USA, 2001, p. 120–127. URL: https://doi.org/10.1145/383952.383972. doi:10.1145/383952.383972.
[16] Ollama, github, 2024. URL: https://github.com/ollama/ollama.
[17] L. AI, Lightning studio (lm studio), https://lightning.ai/docs/studio/, 2023. Software.
[18] P. Malakasiotis, I. Pavlopoulos, I. Androutsopoulos, A. Nentidis, Evaluation measures for task b, BioASQ-EvalMeasures-taskB (2018).

A. Appendix

In all these prompts, the %s after QUESTION is replaced by the actual question, and the %s after INFORMATION, TEXT or ABSTRACT is replaced with the collection of the related snippets or abstracts, concatenated and separated by a single blank.

Yes/No Prompt

Given only the following **INFORMATION** and **QUESTION**, answer the **QUESTION** only with 'Yes' or 'No'. Think carefully. **INFORMATION**: %s **QUESTION**: %s

List Prompt

Answer the **QUESTION** using only the **TEXT** by only returning a list of entity names, numbers, or similar short expressions that are an answer to the question and are separated by commas. Only the list should be returned. If you do not know any answer return the word EMPTY. **TEXT**: %s **QUESTION**: %s

Factoid Prompt

Answer the **QUESTION** using only the **TEXT** by only returning a list of entity names, numbers, or similar short expressions that are an answer to the question and are separated by commas,ordered by decreasing confidence. Only the list should be returned. If you do not know any answer return the word EMPTY. **TEXT**: %s **QUESTION**: %s

Summary Prompt

##**ABSTRACT**: %s ##**QUESTION**: %s ##**TASK**: Answer the **QUESTION** by returning a single paragraph sized text ideally summarizing only the most relevant information in the **ABSTRACT**.

Table 7
Best-performing LLM combinations by question type

Question Type	**Best Model(s)**	**# of LLMs**	**Notes**
Factoid	DeepSeek-R1-Distill-Llama-70B, LLaMA 3.3, Qwen3-14B, Qwen3-32B, Reflection, Smaug	6	Larger combinations performed best due to complementary strengths
List	DeepSeek-R1-Distill-Llama-70B, LLaMA 3.3, Qwen3-14B	3	Small groups outperformed individual models and larger combinations
Yes/No	Aya, Qwen3-32B, Smaug	3	Small combinations (3–4 models) yielded the highest accuracy