CEUR-WS.org/Vol-4038/paper_25.pdf

Prompting Matters: Snippet-Aware Strategies for Biomedical QA with LLMs in BioASQ 13b

Hajung Kim[1,†], Hoonick Lee[1,†], Yewon Cho[1,†], Jungwoo Park[1,†], Jueon Park[1,†], Soyon Park[1,†], Yan Ting Chok[1,†], Seungheun Baek[1,†], Donghyeon Lee[1,†] and Jaewoo Kang[1,2,*]

[1] *Department of Computer Science and Engineering, Korea University, Seoul, 02841, Republic of Korea*
[2] *AIGEN Sciences, Seoul, 04778, Republic of Korea*

Abstract

Biomedical question answering (QA) plays a critical role in enabling efficient access to clinical and scientific knowledge, yet remains a challenging task due to domain complexity, terminology ambiguity, and evidence integration from multiple sources. In this study, we address these challenges in the context of the BioASQ 13b Task B challenge, which evaluates systems across yes/no, factoid, and list question types. In this work, we investigate the impact of prompt design on biomedical question answering in the BioASQ 13b Task B challenge. Specifically, we evaluate: (1) a standard format combining questions with all retrieved snippets, (2) randomized prompting with shuffled snippet orders and repeated trials, (3) a one-by-one snippet querying strategy with output aggregation, and (4) a no-snippet condition relying solely on the model's parametric knowledge. Final predictions are selected via majority voting or log-probability-based ranking, depending on the task type. In the final evaluation, team rankings were determined by averaging the best ranks achieved across sub-tasks (yes/no, factoid, and list), regardless of whether the top- performing results came from the same system. Based on this ranking scheme, our team achieved the highest average rank in Batches 1 and 4, and the second-highest in Batches 2 and 3, demonstrating the robustness and effectiveness of our prompt design.

Keywords

BioASQ 13b, LLM, question-answering, prompt engineering

1. Introduction

Accessing high-quality biomedical information remains a significant challenge for clinicians and researchers who must navigate heterogeneous databases and fragmented knowledge sources to find precise, evidence-based answers. Traditional keyword-based search engines are often inadequate for handling complex expert queries, especially when answers are dispersed across multiple scientific articles. To address these limitations, the BioASQ challenge [1, 2] has been a cornerstone benchmark for biomedical question answering (QA) systems. The challenge, now in its thirteenth edition (BioASQ 13b, 2025 [3]), is designed to promote the development of AI systems that can semantically index biomedical literature and generate accurate, context-aware responses to expert-written questions. Among the various tasks, BioASQ Task B—Biomedical Semantic QA—requires systems to answer questions using evidence snippets retrieved from PubMed [4], while returning both exact answers (such as yes/no, entities, or lists) and ideal summaries. Task B encompasses four question types: yes/no, factoid, list, and ideal answers. In this work, we focus on the first three types, which require precise and structured responses grounded in evidence. BioASQ 13b provides the largest training corpus to date, comprising 5,389 expert-authored questions with carefully curated snippet evidence, offering a unique testbed for evaluating modern LLM-based QA systems.

While recent advances in large language models (LLMs) such as GPT-4 [5] and Claude [6] have shown impressive general-domain QA capabilities, their performance in biomedical contexts is highly sensitive to how information is presented in the prompt. In particular, biomedical QA demands integrating multiple retrieved snippets, each of which may contain partial or conflicting evidence. Prompt design

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

[†] These authors contributed equally to this work.

kangj@korea.ac.kr (J. Kang)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

thus plays a pivotal role in determining whether the model can reason over the provided context effectively.

In this study, we systematically explore several snippet-aware prompting strategies to better understand their impact on biomedical QA performance. We evaluate: (1) a **default format** that combines task instruction, question, and the complete list of snippets in a single prompt; (2) a **randomized snippet order setting** that introduces permutation over snippet order across repeated trials to reduce positional bias; (3) a **one-by-one approach** that queries the model with each snippet separately and aggregates the answers; and (4) a **prior-knowledge-only condition** in which the model generates an answer without any evidence snippets, relying solely on parametric knowledge.

We apply these prompting strategies to multiple LLMs, including GPT-4o-mini, GPT-4, and Claude, and consolidate the outputs using log-probability-based ranking and majority voting to improve reliability. This unified evaluation reveals the strengths and limitations of each prompting approach and model, offering practical insights for designing robust biomedical QA systems. Our system achieved top-tier performance in the BioASQ 13b Task B challenge, ranking first in batches 1 and 4 and second in batches 2 and 3. These results highlight the critical importance of prompt structure, especially in evidence-sensitive tasks such as factoid and list-type QA.

2. Task Description

The BioASQ challenge promotes the development of intelligent systems for biomedical question answering. In this paper, we focus on Phase B of the BioASQ 13b task[7], where systems are required to generate precise answers to questions using evidence snippets retrieved from PubMed abstracts [4]. Each question is paired with a set of snippets, and its expected answer format varies depending on the question type: **yes/no**, **factoid**, or **list**.

Yes/No. For this question type, the model is expected to answer "**yes**" or "**no**" **based on the given snippets**. An example is: "*Is age an underlying factor in the eye disease AMD?*" Answering yes/no questions often requires considering multiple snippets rather than relying on a single snippet alone. The evaluation is based on **accuracy** or **macro-averaged F1-score**, where each label (*"yes"* or *"no"*) is treated as a separate class. This approach ensures that the model is not biased toward a majority class, and it is fairly evaluated on positive and negative responses. The macro-averaged F1-score is computed as:

$$\text{Macro-F1} = \frac{1}{2}\left(\text{F1}_{yes} + \text{F1}_{no}\right)$$

where for each class $c \in \{yes, no\}$, the precision, recall, and F1 score are given by:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \quad \text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

Here, TP_c, FP_c, and FN_c denote the number of true positives, false positives, and false negatives for class c, respectively.

Factoid. **Factoid** questions expect a **single answer** that best represents the factual information the requested. An example is: "*What is the target of fezolinetant?*" The ground-truth answers are usually, though not always, contained within the snippets. This type of question differs from other QA tasks, such as SQuAD [8], where answers can always be extracted from the given context. Answers to factoid-type questions are represented as a double list of possible answers, where each inner list contains acceptable variants of the same correct answer. In the evaluation, higher scores are assigned when the preferred correct answer appears earlier in the system's ranked list of predictions. Factoid questions are evaluated using the **Mean Reciprocal Rank (MRR)**, which accounts for both the correctness and the ranking position of the answers. Specifically, the reciprocal of the rank at which the first correct answer

	yesno		factoid		list	
	# of questions	avg.# of snippets	# of questions	avg # of snippets	# of questions	avg.# of snippets
11b test	86	9.45	98	7.39	66	15.15
12b test	102	25.95	85	27.19	80	22.15

Table 1
Summary of the dataset composition across three question types (yes/no, factoid, list) and two data splits (11b test, 12b test). Each cell reports the number of questions and the average number of supporting snippets per question.

appears in the system's prediction list is computed for each question. The MRR is then calculated as the average of these reciprocal ranks over all factoid questions:

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$$

where Q is the set of factoid questions, and rank_i denotes the position of the first correct answer for the i-th question. If no correct answer is found in the predicted list, the reciprocal rank for that instance is defined as zero.

List. List-type questions require the system to return a **list of correct answers**, i.e., more than one acceptable answer. An example is: "*Please list the symptoms of Chikungunya virus infection*". Although list-type questions may appear similar to factoid-type questions, they differ in that they expect multiple distinct answers rather than a single best answer or alias group. Each item in the gold standard represents a unique element, and the system is evaluated based on how many of these items are correctly retrieved. Evaluation is conducted using standard classification metrics: **precision**, **recall**, and **F1-score**, calculated based on the overlap between the predicted and gold-standard answer sets. Unlike factoid questions, the order of the answers does not affect the evaluation score; only the set overlap is considered.

3. Methods

3.1. Dataset

We utilized the official BioASQ datasets [9, 10] released as part of the BioASQ Challenge. In total, 5,389 biomedical questions were provided by the challenge organizers, categorized into three types: **yes/no**, **factoid**, and **list**. Each question is accompanied by supporting snippets from PubMed abstracts.

For evaluation, we constructed two internal test subsets based on the official BioASQ 11b and 12b test sets. The BioASQ 11b test data contains 86 yes/no, 98 factoid, and 66 list questions, with an average of 9.45, 7.39, and 15.15 snippets per question, respectively. The second test set is derived from the BioASQ 12b test set and follows the same annotation format. See Table 1 for detailed statistics.

3.2. Model

To address BioASQ Task B, which focuses on biomedical question answering across multiple formats-including yes/no, factoid, and list-type questions-we employed a suite of state-of-the-art large language models (LLMs): GPT-4-0125-preview, GPT-4o-mini, GPT-4o (developed by OpenAI), and Claude (developed by Anthropic). These models were chosen for their strong reasoning capabilities, broad biomedical knowledge, and robust performance in open-domain QA tasks. Our model selection was motivated in part by the success of previous BioASQ challenge participants [11, 12, 13], many of whom integrated GPT-based models—such as GPT-3.5 and early GPT-4 variants—into their systems and achieved competitive results. These precedents demonstrated the effectiveness of large-scale LLMs in biomedical

Prompt for Yes/No	Prompt for Factoid	Prompt for List
INSTRUCTIONS: 1. Base your answer ONLY on the information provided in the [Snippets] section. 2. If there is any conflict between the snippet information and your internal knowledge, ALWAYS prioritize the snippet. 3. Provide your final answer solely based on the snippet content. 4. You must answer only with lowercase "yes" or "no". If you are not sure, answer "none". [Snippets]: <snippets> [Question]: <question>	INSTRUCTIONS: 1. Base your answer ONLY on the information provided in the [Snippets] section. 2. If there is any conflict between the snippet information and your internal knowledge, ALWAYS prioritize the snippet. 3. Provide your final answer solely based on the snippet content. 4. Provide 1–2 items only. Use a single term if synonyms exist. Return a JSON string array of concise entity names or numbers. Return an empty list if unknown. [Snippets]: <snippets> [Question]: <question>	INSTRUCTIONS: 1. Base your answer ONLY on the information provided in the [Snippets] section. 2. If there is any conflict between the snippet information and your internal knowledge, ALWAYS prioritize the snippet. 3. Provide your final answer solely based on the snippet content. 4. Provide up to 7–8 items only. Use a single term if synonyms exist. Return a JSON string array of concise entity names or numbers. Return an empty list if unknown. [Snippets]: <snippets> [Question]: <question>

Table 2
Prompts used for each QA type in our system. Each prompt instructs the model to rely strictly on snippet content and return appropriately constrained outputs based on the question type.

QA settings and provided a practical foundation for adopting more advanced models in our pipeline. Additionally, we prioritized the GPT-4 series over GPT-3.5 due to their support for structured outputs, such as generating answers in JSON format. This feature is especially valuable in BioASQ, where answers must adhere to specific schemas depending on question type (e.g., boolean for yes/no, single entities for factoid, or lists for list questions). Relying on models that directly produce structured responses allowed us to reduce post-processing overhead and minimize potential formatting errors, thereby improving overall pipeline robustness and evaluation compatibility.

3.3. Prompt Design

To accommodate the distinct requirements of the three question types in BioASQ Task B-yes/no, factoid, and list-we designed custom prompts tailored to each type. These prompts enforced reliance on provided snippets and required structured outputs where applicable. To ensure that the model reliably followed these constraints, we carefully phrased and, in some cases, repeated key instructions within the prompts to reinforce snippet-based answering behavior. The detailed formats for each prompt are provided in Table 2.

To improve the model's ability to capture relevant entities, particularly when snippets are long or contain dispersed information, we experimented with multiple snippet input configurations. We initially designed multiple strategies not only to ensure coverage, but also to assess their relative effectiveness. Our goal was to reduce the chance of missing key evidence by exploring different ways of presenting the input. The following four strategies were implemented:

- **Full Snippet (Default Order):** All available snippets associated with a given question were concatenated and provided to the model in the original order as retrieved from the dataset. This setting reflects a realistic scenario where the model receives the complete context at once and is expected to synthesize information across multiple evidence sentences.

- **Random Ordered Full Snippet:** All snippets were concatenated into a single input, but the order of the snippets was randomly shuffled. This configuration was used to examine the model's robustness to non-sequential evidence and to assess whether reasoning performance depends on snippet ordering.
- **Single Snippet (One-by-One):** Each snippet was presented to the model individually, paired with the same question. The model generated an answer for each snippet independently, and the final answer was determined by aggregating outputs (e.g., via majority vote for yes/no, or top-confidence entity selection for factoid/list). This approach aimed to isolate the model's ability to extract relevant information from minimal context without being influenced by noisy or conflicting snippets.
- **No Snippet:** The question was presented without any snippets, forcing the model to answer based solely on internal knowledge.

By applying all of these strategies across question types, we aimed to enhance answer coverage and mitigate the risk of evidence omission.

3.4. Ensemble Strategy

We performed ensemble across the outputs of multiple models, using task-specific strategies described below:

- **Yes/No Questions:** For each yes/no question, we collected all responses generated by the models and applied a majority voting scheme. Only responses labeled as `"yes"` or `"no"` were considered valid votes; responses labeled as `"none"` were treated as abstentions and excluded from the voting pool. The label with the highest number of votes among valid responses was selected as the final answer. This approach ensured that the decision reflected the consensus among models while discarding uncertain outputs.
- **Factoid Questions:** In the case of factoid questions, each model produced one or more entity candidates along with associated log probabilities. We aggregated all unique entity candidates across models and summed their log probabilities when the same entity appeared multiple times. The final answer was determined by selecting the top-ranked entities according to these cumulative scores. This ranking-based strategy allowed us to combine probabilistic confidence across models and prioritize entities with consistent high-confidence support.
- **List Questions:** For list-type questions, we applied a frequency-based ensemble method. Each model generated a list of entities as potential answers. We counted how many times each entity appeared across all model outputs, regardless of position. Entities that exceeded a predefined occurrence threshold were included in the final list. This threshold was chosen to balance precision and recall, ensuring that only consistently predicted entities were retained while filtering out spurious or low-confidence entries.

These ensemble strategies were applied uniformly across all model outputs to improve robustness and answer reliability. By leveraging complementary predictions from multiple models and aggregating them in a task-aware manner, we aimed to reduce variance, correct isolated errors, and enhance overall consistency. This approach allowed us to take advantage of both model diversity and redundancy, ensuring that the final predictions reflected a more stable consensus across different models and decoding conditions.

Table 3
Performance (Macro F1) on BioASQ-13b in the yes/no type. Numbers in parentheses indicate the number of single models constituting the ensemble. Systems are classified as either single or ensemble models, with most based on one-by-one inference.

Batch	Dense Rank	System	Macro F1	Description
Batch 1	1	`2025-DMIS-KU-2`	1.0000	Ensemble (10), full snippet
	2	`2025-DMIS-KU-1`	0.9328	Ensemble (2), one-by-one
	2	`2025-DMIS-KU-5`	0.9328	Ensemble (2), one-by-one
	2	`2025-DMIS-KU-3`	0.9328	Single, one-by-one
	6	`2025-DMIS-KU-4`	0.8132	Single, one-by-one
Batch 2	1	`2025-DMIS-KU-2`	1.0000	Ensemble (10), full snippet
	2	`2025-DMIS-KU-1`	0.9377	Ensemble (2), one-by-one
	2	`2025-DMIS-KU-5`	0.9377	Ensemble (2), one-by-one
	3	`2025-DMIS-KU-3`	0.9328	Ensemble (2), one-by-one
	4	`2025-DMIS-KU-4`	0.8786	Single, one-by-one
Batch 3	1	`2025-DMIS-KU-1`	0.9394	Ensemble (2), one-by-one
	1	`2025-DMIS-KU-2`	0.9394	Ensemble (12), one-by-one
	1	`2025-DMIS-KU-4`	0.9394	Single, one-by-one
	2	`2025-DMIS-KU-3`	0.8854	Ensemble (2), one-by-one
	2	`2025-DMIS-KU-5`	0.8854	Ensemble (2), one-by-one
Batch 4	2	`2025-DMIS-KU-3`	0.9532	Ensemble (2), one-by-one
	3	`2025-DMIS-KU-4`	0.9487	Single, one-by-one
	4	`2025-DMIS-KU-5`	0.9097	Ensemble (2), one-by-one
	5	`2025-DMIS-KU-2`	0.9023	Ensemble (10), full snippet
	8	`2025-DMIS-KU-1`	0.8595	Ensemble (2), full snippet

4. Result

4.1. Official Evaluation on BioASQ-13b

Tables 3, 4, 5 show that our best-performing models achieved top scores in most batches across all question types. In these tables, the "Dense Rank" column represents the system's rank on the leaderboard without skipping positions for ties (i.e., ranks increase sequentially even when scores are tied). The "Description" column follows the format "Ensemble (k), Prompt Type", where k denotes the number of model outputs combined, and Prompt Type indicates the prompting strategy used (e.g., One-by-One). In particular, in the **yes/no** type, all our systems consistently demonstrated stable and strong performance across all batches. Most systems were designed around the one-by-one strategy, in which answers are predicted separately for each snippet. We also integrated variations such as generating responses without snippets and randomly selected snippets to diversify perspectives. To better handle uncertainty, we applied rule-based fallback strategies triggered by indicators like the ratio of "none" responses or imbalanced yes/no predictions. Based on these signals, answers were either substituted from other systems or selected based on log-probability confidence. While many systems remained small-scale ensembles of two models, some incorporated up to 12 different outputs. The best-performing models effectively combined these techniques, reaching a macro F1 score of 1.0000 in both Batches 1 and 2.

All systems except one were based on single-model configurations for the **factoid** task. We tested different answer selection methods, such as log-probability scoring with rule-based filters, and a Reciprocal Rank Fusion (RRF) strategy in one case. The systems also varied in how they processed input: some relied on all available snippets, while others processed each one independently. No ensemble was used, except in a single submission for Batch 4, which combined three different outputs using RRF and yielded competitive performance. Despite focusing primarily on individual systems, we achieved strong MRR scores, consistently ranking among the top in several batches, including two second-place finishes in Batch 1 and Batch 4.

Table 4
Performance (mean reciprocal rank, MRR) on BioASQ-13b in the factoid type. Numbers in parentheses indicate the number of single models constituting the ensemble. All systems are single models unless otherwise noted, and include the prompt types described in Section 3.3.

Batch	Dense Rank	System	MRR	Description
Batch 1	1	`2025-DMIS-KU-1`	0.5962	Single, full snippet
	1	`2025-DMIS-KU-5`	0.5962	Single, one-by-one
	1	`2025-DMIS-KU-3`	0.5962	Single, one-by-one
	3	`2025-DMIS-KU-4`	0.5513	Single, full snippet
	4	`2025-DMIS-KU-2`	0.5256	Single, full snippet
Batch 2	2	`2025-DMIS-KU-4`	0.6667	Single, random ordered full snippet
	3	`2025-DMIS-KU-5`	0.6481	Single, random ordered full snippet
	9	`2025-DMIS-KU-2`	0.5556	Single, random ordered full snippet
	9	`2025-DMIS-KU-3`	0.5556	Single, random ordered full snippet
	10	`2025-DMIS-KU-1`	0.5370	Single, one-by-one
Batch 3	2	`2025-DMIS-KU-4`	0.5042	Single, full snippet
	7	`2025-DMIS-KU-3`	0.4458	Single, full snippet
	8	`2025-DMIS-KU-1`	0.4392	Single, full snippet
	9	`2025-DMIS-KU-5`	0.4333	Single, full snippet
	11	`2025-DMIS-KU-2`	0.4142	Single, full snippet
Batch 4	2	`2025-DMIS-KU-2`	0.6136	Single, random ordered full snippet
	2	`2025-DMIS-KU-3`	0.6136	Single, random ordered full snippet
	2	`2025-DMIS-KU-4`	0.6136	Single, full snippet
	3	`2025-DMIS-KU-5`	0.5909	Ensemble (3), full snippet
	6	`2025-DMIS-KU-1`	0.5455	Single, full snippet

For the **list**-type task, we generated 16 candidate outputs by applying four distinct prompting strategies to four different LLMs. Most of our systems were ensemble models that combined a subset of these outputs. In Batches 1 and 3, we used fixed combinations of prompting strategies such as full snippet, random ordered full snippet, and one-by-one prompting. Batch 2 employed a weighted ensemble over combinations, where the models and their weights were determined based on F1 scores from internal validation across all possible 4-model combinations among the 16 candidates. In Batch 4, we selected four models from the full candidate pool and combined them using equal weights. Overall, performance improvements across batches were primarily driven by effective model selection and the diversity of prompting strategies, rather than the size of the ensemble.

4.2. Evaluation on 11b and 12b

We conducted an ablation study using the `GPT-4o-mini` model to evaluate how different prompt design strategies affect biomedical question answering performance. As summarized in Table 6, we compared the following four configurations:

- **Full Snippet (Default Order):** The full set of retrieved evidence snippets is concatenated in their original order and provided in a single prompt.
- **Random Ordered Full Snippet:** Snippets are randomly reordered and queried multiple times to mitigate positional bias and enhance evidence diversity.
- **Single Snippet (One-by-One):** Each snippet is paired individually with the question. The model produces separate outputs, which are aggregated to form the final answer. In some cases, `GPT-4` was used as the decoder in this configuration.
- **No Snippet:** The model answers the question without access to any supporting snippets, relying solely on its parametric knowledge.

Table 5
Performance (F-Measure) on BioASQ-13b in the list type. Numbers in parentheses indicate the number of single models constituting the ensemble. Systems are categorized as either single models or ensembles, using base, random, or hybrid strategies.

Batch	Dense Rank	System	F-Measure	Description
Batch 1	2	`2025-DMIS-KU-3`	0.5913	Ensemble (5), full snippet
	3	`2025-DMIS-KU-2`	0.5852	Ensemble (5), full snippet
	7	`2025-DMIS-KU-1`	0.5679	Ensemble (5), random ordered snippet
	12	`2025-DMIS-KU-4`	0.5473	Ensemble (5), one-by-one
	18	`2025-DMIS-KU-5`	0.5342	Ensemble (5), full snippet
Batch 2	8	`2025-DMIS-KU-4`	0.5670	Ensemble (5), full snippet
	9	`2025-DMIS-KU-5`	0.5545	Ensemble (5), full snippet
	11	`2025-DMIS-KU-1`	0.5522	Ensemble (5), full snippet
	11	`2025-DMIS-KU-2`	0.5522	Ensemble (5), full snippet
	12	`2025-DMIS-KU-3`	0.5513	Ensemble (5), full snippet
Batch 3	4	`2025-DMIS-KU-5`	0.6269	Ensemble (11), majority voting
	7	`2025-DMIS-KU-4`	0.6123	Ensemble (3), full snippet
	9	`2025-DMIS-KU-3`	0.6087	Single, full snippet
	12	`2025-DMIS-KU-2`	0.6024	Single, random ordered snippet
	16	`2025-DMIS-KU-1`	0.5912	Single, full snippet
Batch 4	3	`2025-DMIS-KU-4`	0.6328	Ensemble (4), full snippet
	7	`2025-DMIS-KU-3`	0.6200	Ensemble (4), full snippet
	8	`2025-DMIS-KU-5`	0.6180	Ensemble (4), full snippet
	9	`2025-DMIS-KU-2`	0.6160	Ensemble (4), full snippet
	10	`2025-DMIS-KU-1`	0.6155	Ensemble (4), full snippet

Table 6
Performance comparison of different snippet prompting strategies across question types in BioASQ 13b Task B. The upper half corresponds to phase 11b, and the lower half to phase 12b.

Snippet Strategy	**yesno**			**factoid**			**list**		
	f1-yes	f1-no	macro-f1	strict acc.	lenient acc.	mrr	mean prec.	recall	f-measure
BioASQ 11b Task B									
full snippet	0.9375	0.9750	0.9563	0.7544	0.7949	0.7747	**0.6635**	**0.7104**	**0.6675**
random ordered	**0.9722**	**0.9868**	**0.9795**	0.7509	0.7931	0.7720	0.6477	0.6850	0.6455
one-by-one	0.9470	0.9615	0.9547	**0.7623**	**0.8398**	**0.7941**	0.6401	0.7099	0.6522
no snippet	0.8192	0.8204	0.8198	0.4199	0.4199	0.4199	0.4134	0.4596	0.4170
BioASQ 12b Task B									
full snippet	0.9169	**0.8520**	**0.8844**	0.6186	0.6818	0.6502	**0.6414**	**0.6202**	**0.6106**
random ordered	**0.9183**	0.8450	0.8817	0.4918	0.6449	0.5661	0.6260	0.5915	0.5889
one-by-one	0.8902	0.8225	0.8563	**0.6710**	**0.8494**	**0.7403**	0.5696	0.5521	0.5217
no snippet	0.7546	0.6369	0.6958	0.3959	0.4270	0.4114	0.3514	0.3434	0.3281

Yes/No. The *random ordered full snippet* setting yielded the highest macro-F1 score (0.9795), demonstrating that randomized snippet ordering can improve robustness. The *full snippet* setting achieved a macro-F1 of 0.8844, suggesting that a holistic view of the full snippet set provides consistent performance.

Factoid. The *one-by-one* setting consistently outperformed all others, achieving the best strict accuracy, lenient accuracy, and MRR (0.7623, 0.8398, and 0.7941, respectively). These results indicate that isolating snippets can help the model better extract concise factual entities, especially when they are sparsely distributed across evidence.

List. Performance on list-type QA revealed a trade-off between coverage and precision. The *full snippet* format achieved the highest F1 score (0.6675) in the high-resource setup, likely due to its ability to synthesize entity mentions across snippets. In contrast, the *one-by-one* strategy showed reduced recall, possibly due to fragmented evidence limiting answer aggregation.

No Snippet Setting. The *no snippet* baseline consistently underperformed across all question types. This confirms the importance of retrieved evidence, particularly for factoid and list questions where external grounding is essential.

These findings highlight the crucial role of prompt design in biomedical QA. While one-by-one prompting is particularly effective for entity-centric questions like factoids, the original format offers more balanced performance for yes/no and list questions—especially under constrained conditions. Randomized snippet ordering also improves model robustness, underscoring the benefit of prompt-level perturbations in multi-evidence settings.

5. Conclusion

In this study, we developed a series of systems for the BioASQ-13b Challenge Phase B, targeting yes/no, factoid, and list question types. Our analysis revealed that the effectiveness of prompting strategies varied across tasks. Yes/no questions, framed as binary classification problems, benefited from full-context prompts that enabled holistic reasoning over all provided snippets. In contrast, factoid questions, which require precise entity-level retrieval, performed better with one-by-one prompting, where each snippet is processed independently to ensure comprehensive coverage of fine-grained information. This approach was particularly useful when handling long or noisy snippet sets, where key evidence might otherwise be overlooked.

Guided by these insights, we applied task-specific ensemble strategies. For the factoid task, we used a log-probability-based ensemble: each model produced a ranked list of entity candidates with associated log-probabilities. We aggregated all unique entity candidates across models and summed their log-probabilities when the same entity appeared multiple times. The final answer was selected by ranking entities based on these cumulative scores, effectively capturing probabilistic consensus across models. For list-type questions, we employed a frequency-based ensemble: each model generated a list of candidate entities, and we counted the number of times each entity appeared across outputs, regardless of position. Entities that exceeded a predefined occurrence threshold were retained in the final answer. This method helped balance precision and recall, promoting entities with consistent support while filtering out low-confidence or spurious predictions.

Overall, our results demonstrate that task-aware prompting combined with lightweight ensemble techniques can effectively enhance system performance. This strategy offers a practical and interpretable framework for biomedical question answering with large language models.

6. Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2023R1A2C3004176), the Korea Health Industry Development Institute (KHIDI) grant funded by the Ministry of Health & Welfare, Republic of Korea (HR20C002103), and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2025-RS-2020-II201819). This research was also supported by the RS-2023-00262002 grant.

7. Declaration on Generative AI

Generative AI tools were used for writing assistance, including grammar correction and sentence rephrasing.

References

[1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 1–28.

[2] 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: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] 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.

[4] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, X. Lu, Pubmedqa: A dataset for biomedical research question answering, arXiv preprint arXiv:1909.06146 (2019).

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[6] Anthropic, The Claude 3 Model Family: Opus, Sonnet, Haiku, Technical Report, Anthropic, 2024. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, model Card.

[7] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 13b and Synergy13 in CLEF2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[8] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, arXiv preprint arXiv:1606.05250 (2016).

[9] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.

[10] A. Krithara, J. G. Mork, A. Nentidis, G. Paliouras, The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey, Frontiers in Research Metrics and Analytics 8 (2023) 1250930.

[11] J. H. Merker, A. Bondarenko, M. Hagen, A. Viehweger, Mibi at bioasq 2024: retrieval-augmented generation for answering biomedical questions, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, volume 3740, 2024, pp. 176–187.

[12] B.-C. Chih, J.-C. Han, R. Tzong-Han Tsai, Ncu-iisr: enhancing biomedical question answering with gpt-4 and retrieval augmented generation in bioasq 12b phase b, CLEF Working Notes (2024).

[13] O. Şerbetçi, X. D. Wang, U. Leser, Hu-wbi at bioasq12b phase a: Exploring rank fusion of dense retrievers and re-rankers, in: Proceedings of the Conference and Labs of the Evaluation Forum, Grenoble, France, 2024, pp. 9–12.