Can Language Models Critique Themselves? Investigating Self-Feedback for Retrieval Augmented Generation at BioASQ 2025

Notebook for the BioASQ Lab at CLEF 2025

Samy Ateia[1], Udo Kruschwitz[1]

[1]Information Science, University of Regensburg, Universitätsstraße 31, 93053, Regensburg, Germany

Abstract

Agentic Retrieval Augmented Generation (RAG) and 'deep research' systems aim to enable autonomous search processes where Large Language Models (LLMs) iteratively refine outputs. However, applying these systems to domain-specific professional search, such as biomedical research, presents challenges, as automated systems may reduce user involvement and misalign with expert information needs. Professional search tasks often demand high levels of user expertise and transparency. The BioASQ CLEF 2025 challenge, using expert-formulated questions, can serve as a platform to study these issues. We explored the performance of current reasoning and nonreasoning LLMs like Gemini-Flash 2.0, o3-mini, o4-mini and DeepSeek-R1. A key aspect of our methodology was a self-feedback mechanism where LLMs generated, evaluated, and then refined their outputs for query expansion and for multiple answer types (yes/no, factoid, list, ideal). We investigated whether this iterative self-correction improves performance and if reasoning models are more capable of generating useful feedback. Preliminary results indicate varied performance for the self-feedback strategy across models and tasks. This work offers insights into LLM self-correction and informs future work on comparing the effectiveness of LLM-generated feedback with direct human expert input in these search systems.

Keywords

Retrieval Augmented Generation, Large Language Models, Biomedical Question Answering, Professional Search, Self-Feedback Mechanisms, Query Expansion, BioASQ,

1. Introduction

Large Language Models (LLMs) are increasingly deployed in generative search engines that are embedded and offered through AI services such as ChatGPT or Microsoft Copilot. These systems are advertised and used to solve complex, often work-related tasks [1, 2]. Their newly introduced "deep research"[1][2] modes extend generative search by automating iterative, multistep information discovery to produce extensive reports. Tools like Elicit[3] demonstrate how similar approaches can be tailored specifically for professional search tasks, including technology-assisted systematic literature reviews [3]. However, these professional search scenarios typically require significant user expertise and system transparency [4, 5]. But current LLM limitations, such as the potential for generating unsupported statements [6], present challenges, especially when direct expert oversight is reduced through the promise of automation [7].

In this study, we evaluate the performance of current reasoning and non-reasoning LLMs in retrieving relevant biomedical information and answering expert-formulated questions. Specifically, we investigate whether these models can leverage iterative self-feedback mechanisms to enhance their query expansion and response quality.

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

Samy.Ateia@stud.uni-regensburg.de (S. Ateia); udo.kruschwitz@ur.de (U. Kruschwitz)

0009-0000-2622-9194 (S. Ateia); 0000-0002-5503-0341 (U. Kruschwitz)

[1]https://web.archive.org/web/20250516004701/https://openai.com/index/introducing-deep-research/

[2]https://web.archive.org/web/20250519025641/https://gemini.google/overview/deep-research/

[3]https://web.archive.org/web/20250502093323/https://elicit.com/solutions/systematic-reviews

1.1. BioASQ Challenge

The BioASQ challenge provides a long-running platform for evaluating systems on large-scale biomedical semantic indexing and question answering [8]. Participants are tasked with retrieving relevant documents and snippets from biomedical literature (PubMed[4]) and generating precise answers to expert-formulated questions, which can be in yes/no, factoid, list, or ideal summary formats. The structured, domain-specific nature of the BioASQ challenge makes it especially suitable for assessing advanced RAG methods for expert information needs.

1.2. Our Contribution

Our team has participated in previous iterations of the BioASQ challenge, examining the performance of various commercial and open-source LLMs, the impact of few-shot learning, and the effects of additional context from knowledge bases [9, 10]. In this year's challenge (CLEF 2025), we continued our participation across Task A (document and snippet retrieval), Task A+ (Q&A with own retrieved documents), and Task B (Q&A with retrieved and gold documents). Our primary investigation centered on the effectiveness of a self-feedback loop implemented with current LLMs, including Gemini-Flash 2.0, o3-mini, o4-mini, and DeepSeek Reasoner, to evaluate if models can improve their own generated query expansions and answers through self-critique.

2. Related Work

This work builds upon recent advancements in Large Language Models (LLMs), few-shot and zero-shot learning, Retrieval Augmented Generation (RAG), and their applications to professional search.

2.1. Large Language Models

The field of Natural Language Processing (NLP) has been significantly advanced by Large Language Models, mostly based on the transformer architecture [11]. Early influential models like BERT (Bidirectional Encoder Representations from Transformers) [12] demonstrated the power of pre-training on large text corpora. Parallel developments led to autoregressive models such as the GPT (Generative Pre-trained Transformer) series [13, 14]. The capabilities of these models were further improved through techniques like Reinforcement Learning from Human Feedback (RLHF), which helps align LLM outputs with human preferences and instructions, making them better at following prompts [15].

Recent months have seen the emergence of numerous so-called reasoning models from various developers, including Google's Gemini 2.5[5], OpenAI's o1[6] to o4-mini model series and models like DeepSeek R1 [16]. These models build on the idea of Chain of Thought (CoT) prompting [17] that showed that models perform better when they are prompted to generate additional tokens in their output that mimic reasoning or thinking steps. Fine-tuning models for this reasoning process and therefore enabling variable scaling of test-time compute [18] enabled further advances in model performance on popular benchmarks. Reinforcement learning on math and coding related datasets called Reinforcement Learning with Verifiable Reward (RLVR) [19] seems to be a current approach to enable models to find useful reasoning strategies.

Our work uses several of these current reasoning and non-reasoning models to compare their performance in a biomedical RAG setting and to see if these reasoning models are better at generating self-feedback.

[4] https://pubmed.ncbi.nlm.nih.gov/download/#annual-baseline

[5] https://web.archive.org/web/20250518193243/https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/

[6] https://web.archive.org/web/20250518101415/https://openai.com/index/learning-to-reason-with-llms/

2.2. Few and Zero-Shot Learning

A key characteristic of modern LLMs is their ability to perform tasks with minimal or no task-specific training data, often referred to as In-Context Learning (ICL). **Few-shot learning** allows LLMs to learn a new task by conditioning on a few input-output examples provided directly in the prompt. This approach removes the need for extensive, curated training datasets, a concept popularized by models like GPT-3 [14]. **Zero-shot learning** takes this further, enabling LLMs to perform tasks based solely on a natural language description or a direct question, without any preceding examples.

Our previous work has demonstrated the competitive performance of both zero-shot and few-shot approaches in the BioASQ challenge [9, 10]. These techniques are fundamental to the prompting strategies used in our current experiments, forming the basis for initial query/answer generation before any self-feedback loop.

2.3. Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) combines the generative capabilities of LLMs with information retrieved from external knowledge sources [20]. This approach aims to ground LLM responses in factual data, thereby reducing the likelihood of hallucinations and improving the reliability and verifiability of generated content [21]. A typical RAG pipeline involves a retriever that fetches relevant documents or snippets, and a generator LLM that synthesizes an answer based on the prompt and the retrieved context. The BioASQ challenge itself can be considered an example of a RAG setup in a specialized domain.

The RAG concept is evolving towards more dynamic and autonomous systems, sometimes termed Agentic RAG or 'deep research' systems [22]. One of the first of such systems was WebGPT a fine-tuned version of GPT-3 published by a team at OpenAI in 2021 [23]. It took OpenAI another 3 years[7] to roll out a similar system to their ChatGPT user base [8]. Their newest models, o3 and o4-mini are trained via reinforcement learning to decide autonomously when and how long to search among using other tools [9]. These advanced systems may involve LLM-powered agents performing multistep retrieval, reasoning over the retrieved information, and iteratively refining their outputs or search strategies. The deep research modes offered by both OpenAI and Google take these concepts even further and let the models search for over 5 minutes through up to hundreds of websites before synthesizing a multipage report.

Our test of a self-feedback mechanism, where an LLM critiques and revises its own generated queries and answers, is intended to analyze the abilities of off the shelf LLMs on such tasks. In future work, we plan to switch out the LLM generated feedback with feedback from human experts to compare the effectiveness of human and AI guided search processes.

2.4. Professional Search

Professional search refers to information seeking conducted in a work-related context, often by specialists who require high precision, control, and the ability to formulate complex queries [4, 24]. Domains such as biomedical research demand robust evidence-based answers, making transparency and the ability to trace information back to source documents crucial [25]. LLMs are increasingly being explored for professional search applications, offering potential benefits like advanced query understanding and generation of evidence-based summaries [3]. However, challenges such as LLM hallucinations and the need to align with expert workflows remain significant.

Our previous work, the BioRAGent system, has focused on making LLM-driven RAG accessible and transparent for biomedical question answering, enabling users to review and customize generated boolean queries in the search process [26]. This study builds upon this work by exploring the impact of generated critical feedback on query generation and answer generation, which will be compared against human feedback in future work.

[7] https://web.archive.org/web/20250516083609/https://openai.com/index/webgpt/
[8] https://web.archive.org/web/20250511211101/https://openai.com/index/introducing-chatgpt-search/
[9] https://web.archive.org/web/20250514114152/https://openai.com/index/introducing-o3-and-o4-mini/

3. Methodology

We evaluated several Large Language Models (LLMs) in the context of the BioASQ CLEF 2025 Challenge, specifically in Task 12 B, which is structured into Phase A (retrieval), Phase A+ (Q&A based on retrieved snippets), and Phase B (Q&A based on additional gold-standard snippets).

3.1. Models

The models used were grouped into two categories:

- **Non-reasoning models:**
  - Gemini Flash 2.0
  - Gemini 2.5 Flash (used without explicit reasoning mode)
- **Reasoning models:**
  - o3-mini
  - o4-mini (introduced mid-challenge and used in later batches)
  - DeepSeek Reasoner (initially used but replaced due to slow API)

3.2. Task 12 B Experimental Setup

We participated in all four batches of Task 13 B and submitted several systems under different configurations. Each batch comprised five runs, covering combinations of baseline prompting, feedback-augmented prompting, and few-shot learning (10-shot).

3.2.1. Phase A: Document and Snippet Retrieval

Each model configuration involved one of the following strategies:

- **Baseline**: Direct prompt-based query generation without iteration.
- **Feedback (FB)**: Prompt refinement using self-generated feedback.
- **Few-shot**: Prompting the model with 10 examples of successful queries.

UR-IW-1 and UR-IW-3 are paired for comparison, both using the same non-reasoning model (Gemini) with and without feedback, respectively. Similarly, UR-IW-2 and UR-IW-4 form a second pair using a reasoning model with and without feedback. While UR-IW-5 is always configured as a non-reasoning few-shot baseline.

The following table summarizes the configurations for Phase A across all four batches:

Table 1
Overview of Phase A configurations. UR-IW-1/3 compare non-reasoning models with and without feedback, UR-IW-2/4 compare reasoning models. FB = Feedback. UR-IW-5 10-shot baseline.

Batch	UR-IW-1	UR-IW-2	UR-IW-3	UR-IW-4	UR-IW-5
1	Gemini 2.0	DeepSeek	Gemini 2.0 + FB	DeepSeek + FB	Gemini 2.0 + 10-shot
2	Gemini 2.0	o3-mini	Gemini 2.0 + FB	o3-mini + FB	Gemini 2.0 + 10-shot
3	Gemini 2.0	o4-mini	Gemini 2.0 + FB	o4-mini + FB	Gemini 2.5 + 10-shot
4	Gemini 2.0	o4-mini	Gemini 2.0 + FB	o4-mini + FB	Gemini 2.5 + 10-shot

The non-feedback and few-shot approaches were mostly identical to our last years' participation, feedback in phase-A was only used for the query generation and refinement step. The top 10 results from the initial query were passed on to the feedback generating model as additional context. For snippet extraction and reranking no feedback was used.

3.2.2. Phase A+ and Phase B: Answer Generation

The system configurations used for Phase A+ and Phase B were similar to those of Phase A. However, the source of contextual snippets to ground the answer generation differed:

- **Phase A+**: Used the top-20 snippets retrieved by the corresponding model in Phase A.
- **Phase B**: Used a merged set combining the top-20 retrieved snippets from Phase A and the gold-standard snippets provided by the organizers.

As in Phase A, UR-IW-1/3 and UR-IW-2/4 are grouped to compare feedback vs. non-feedback performance for non-reasoning and reasoning models, respectively. UR-IW-5 serves as a consistent few-shot baseline using non-reasoning models.

Table 2
Phase A+ and Phase B configurations. Identical setups used across both phases, with differing snippet inputs. UR-IW-1/3 compare non-reasoning models with and without feedback, UR-IW-2/4 compare reasoning models, and UR-IW-5 is a non-reasoning 10-shot baseline.

Batch	UR-IW-1	UR-IW-2	UR-IW-3	UR-IW-4	UR-IW-5
1	Gemini 2.0	o3-mini	Gemini 2.0 + FB	o3-mini + FB	Gemini 2.0 + 10-shot
2	Gemini 2.0	o3-mini	Gemini 2.0 + FB	o3-mini + FB	Gemini 2.0 + 10-shot
3	Gemini 2.0	o4-mini	Gemini 2.0 + FB	o4-mini + FB	Gemini 2.0 + 10-shot
4	Gemini 2.0	o4-mini	Gemini 2.0 + FB	o4-mini + FB	Gemini 2.0 + 10-shot

The non-feedback and few-shot approaches were again mostly identical to our previous year's participation. For feedback-enhanced runs (UR-IW-3 and UR-IW-4), an additional feedback and refinement step was introduced between draft and final answers. This mechanism relied on task-specific feedback prompts followed by a final revision prompt.

The feedback prompts varied by answer type and were designed to elicit critical evaluation from the model:

- **Yes/No questions**: *"Evaluate the draft answer ('yes' or 'no') against the provided snippets and the question. Indicate explicitly if it should change, with brief reasoning."*
- **Factoid questions**: *"Evaluate the draft JSON entity list answer against the provided snippets and the question. Clearly suggest corrections, removals, or additions."*
- **List questions**: Same as factoid prompt.
- **Ideal answer (summary)**: *"Evaluate the provided summary answer for accuracy, clarity, and completeness against the provided snippets and the question. Clearly suggest improvements."*

The generated feedback was then injected into a fixed refinement prompt to guide the model toward a final improved answer:

```
Expert Feedback: {feedback_response}
Revise and provide the final improved answer strictly following the
original instructions.
```

This two-step feedback-refinement process aimed to simulate expert review and enforce more robust quality control over generated answers.

3.3. Technical Implementation

All pipelines were implemented using Python notebooks and the OpenAI, Google and DeepSeek APIs. Query expansion used the `query_string` syntax of Elasticsearch. The PubMed annual baseline of 2024 was indexed (title, abstract only) on an Elasticsearch index using the standard English analyzer. Snippet extraction and reranking were performed via LLM prompts. Code and notebooks are publicly available on GitHub[10] to ensure full reproducibility.

[10] https://github.com/SamyAteia/bioasq2025

4. Results

As the final results of the BioASQ 2025 Challenge are still being rated by experts and won't be released before September, we can only report on the preliminary results published on the BioASQ website. We participated in Task A (document and snippet retrieval), Task A+ (question answering with own retrieved documents), and Task B (question answering with gold standard documents). The experiments were designed to evaluate the efficacy of different large language models (LLMs) and the impact of self-generated feedback. All results are preliminary and subject to change following manual expert evaluation.

4.1. Model Selection

We tested multiple of the current available models with different settings on a small subset of the BioASQ training set [27] from last year, specifically the fourth batch of BioASQ 12 Task B Phase B. These models included:

- deepseek-reasoner
- deepseek-chat
- gemini-2.5-pro-exp-02-05
- gemini-2.0-flash-thinking-exp-01-21
- gemini-2.0-pro-exp-02-05
- gemini-2.0-flash-lite
- gemini-2.0-flash
- claude-3-5-haiku-2024102
- claude-3-7-sonnet-20250219
- gpt-4.5-preview-2025-02-27
- o3-mini-2025-01-31
- gpt-4o-mini-2024-07-18

Table 3
Preliminary LLM Test Results Summary (best in **bold**)

Model Name	Yes/No Macro F1	Factoid MRR	List F-measure	Ideal R2 F1
gpt-4o-mini-2024-07-18	0.911	0.544	0.531	0.308
gpt-4o-mini-self-feedback	0.833	0.544	0.530	0.307
gpt-4o-mini-o3-feedback	0.833	0.526	0.550	0.308
o3-mini-2025-01-31	0.917	0.605	0.541	0.278
gpt-4.5-preview-2025-02-27	**0.957**	0.675	0.546	0.398
claude-3-7-sonnet-20250219	0.878	0.658	0.540	0.277
claude-3-5-haiku-20241022	**0.957**	0.658	0.482	0.240
gemini-2.0-flash	0.954	**0.684**	0.539	0.471
gemini-2.0-flash 10-shot	0.917	0.632	0.572	**0.566**
gemini-2.0-flash-lite	0.911	0.632	**0.588**	0.480
gemini-2.0-pro-exp-02-05	0.841	0.605	0.360	0.294
gemini-2.0-flash-thinking-exp-01-21	0.871	0.675	0.584	0.361
gemini-2.5-pro-exp-02-05	0.841	0.605	0.360	0.294
deepseek-chat	0.862	0.596	0.559	0.329
deepseek-chat-new	0.911	0.632	0.544	0.282
deepseek-reasoner	**0.957**	0.553	0.544	0.208

Key observations from these preliminary tests include:

gemini-2.0-flash: Demonstrated strong performance across multiple metrics, particularly in yesno_macro_f1 (0.954), and factoid_mrr (0.684), while being competitively priced.

deepseek-reasoner: Achieved high yesno_accuracy (0.962963) and yesno_macro_f1 (0.957075), comparable to gemini-2.0-flash, though with slightly lower performance in factoid and list question types in these preliminary tests.

We decided to choose gemini-2.0-flash as the non-reasoning LLM and also for our 10-shot baseline, as it was both competitive, fast and cheap. For the reasoning model we chose deepseek-reasoner because it is an open-weight model, cheaper to use via the official API and competitive with the other alternative reasoning models (o3-mini, gemini-2.0-flash-thinking).

4.2. Task A: Document and Snippet Retrieval

In Task A, systems were evaluated on their ability to retrieve relevant documents and snippets for given biomedical questions. Our systems were compared against other participating systems, with the "Top Competitor" representing the leading system in each batch.

Detailed preliminary result tables are available in Appendix A.

Document Retrieval: Across the four test batches, our systems demonstrated varied performance.

- **Batch 1**: UR-IW-5 (`gemini flash 2.0 + 10-shot`) was our top performer, ranking 22nd with a **MAP** of 0.2865, compared to the Top Competitor's **MAP** of 0.4246. Our other systems followed, with UR-IW-4 (`deepseek-reasoner + feedback`) having the lowest **MAP** (0.1739) among our submissions in this batch.
- **Batch 2**: UR-IW-5 again led our systems (25th, **MAP** 0.2634), with UR-IW-4 (`o3-mini + feedback`) closely following (26th, **MAP** 0.2601). The Top Competitor achieved a **MAP** of 0.4425.
- **Batch 3**: UR-IW-5 (`gemini-2.5-flash-preview + 10-shot`) was our best system (24th, **MAP** 0.1834). The Top Competitor's **MAP** was 0.3236.
- **Batch 4**: UR-IW-5 (`gemini-2.5-flash-preview + 10-shot`) ranked 27th with a **MAP** of 0.0794, while the Top Competitor had a **MAP** of 0.1801.

Snippet Retrieval: Similar trends were observed in snippet retrieval performance.

- **Batch 1**: UR-IW-5 (`gemini flash 2.0 + 10-shot`) performed best among our systems (8th, **MAP** 0.2768), with the Top Competitor achieving a **MAP** of 0.4535.
- **Batch 2**: UR-IW-5 again led our entries (12th, **MAP** 0.3080). The Top Competitor's **MAP** was 0.5522.
- **Batch 3**: UR-IW-1 (`gemini flash 2.0`) and UR-IW-5 (`gemini-2.5-flash-preview + 10-shot`) were our strongest performers, ranking 15th (**MAP** 0.1534) and 18th (**MAP** 0.1488) respectively. The Top Competitor had a **MAP** of 0.4322.
- **Batch 4**: UR-IW-5 (`gemini-2.5-flash-preview + 10-shot`) was our top system (18th, **MAP** 0.0511). The Top Competitor achieved a **MAP** of 0.1634.

Generally, the 10-shot run with `Gemini Flash 2.0 or 2.5` (UR-IW-5) tended to perform better in document and snippet retrieval tasks compared to our other configurations. The impact of feedback on retrieval tasks (UR-IW-3 and UR-IW-4) varied across batches and didn't consistently outperform the base models or the 10-shot variants in **MAP** scores.

4.3. Task A+: Question Answering (Own Retrieved Documents)

Task A+ required systems to answer questions based on the documents and snippets they retrieved in Phase A.

Yes/No Questions:

- UR-IW-1 (`gemini flash 2.0`) and UR-IW-2 (`o3-mini`) often performed strongly. In Batch 1, UR-IW-1, UR-IW-2 and UR-IW-4 (`o3-mini + feedback`) achieved perfect accuracy and **Macro F1** scores.
- UR-IW-5 (`gemini flash 2.0 + 10-shot`) achieved a perfect score in Batch 2.
- In Batch 4, UR-IW-2 (`o4-mini`) was our top performer (2nd, **Macro F1** 0.9097).
- The feedback mechanism (UR-IW-3, UR-IW-4) showed mixed results, sometimes improving (e.g., UR-IW-4 in Batch 3) and sometimes underperforming compared to non-feedback versions.

Factoid Questions:

- In Batch 1, UR-IW-2 (`o3-mini`) and UR-IW-5 (`gemini flash 2.0 + 10-shot`) were our best systems (7th and 8th, **MRR** 0.3782 and 0.3750 respectively).
- UR-IW-4 (`o3-mini + feedback`) performed well in Batch 2 (2nd, **MRR** 0.5370).
- UR-IW-5 (`gemini flash 2.0 + 10-shot`) took the top position in Batch 4 with an **MRR** of 0.5606.
- Feedback versions (UR-IW-3, UR-IW-4) had variable performance. For example, UR-IW-3 (`gemini flash 2.0 + feedback`) ranked 7th in Batch 3 (**MRR** 0.3100).

List Questions:

- Our systems achieved several top rankings in this category. In Batch 1, UR-IW-2 (`o3-mini`), UR-IW-1 (`gemini flash 2.0`), UR-IW-5 (`gemini flash 2.0 + 10-shot`), and UR-IW-4 (`o3-mini + feedback`) secured the top 4 positions with **F-Measures** of 0.2567, 0.2411, 0.2395, and 0.2357 respectively.
- In Batch 2, UR-IW-2 (`o3-mini`) was again a strong performer (2nd, **F-Measure** 0.3805).
- The effect of feedback and few-shot learning varied. For instance, in Batch 3, UR-IW-5 (`gemini flash 2.0 + 10-shot`) ranked 13th (**F-Measure** 0.3618).

4.4. Task B: Question Answering (Gold Standard Documents)

Task B involved answering questions using additional gold standard documents and snippets.

Yes/No Questions:

- UR-IW-1 (`gemini flash 2.0`) and UR-IW-5 (`gemini flash 2.0 + 10-shot`) achieved perfect scores in Batch 1.
- UR-IW-5 also achieved a perfect score in Batch 2.
- The feedback system UR-IW-4 (`o3-mini + feedback` or `o4-mini + feedback`) performed well, often outperforming its non-feedback counterpart in later batches (e.g., UR-IW-4 in Batch 3 and Batch 4 with **Macro F1** of 0.8706 and 0.9097 respectively).

Factoid Questions:

- UR-IW-3 (`gemini flash 2.0 + feedback`) was our best system in Batch 1 (17th, **MRR** 0.4821).
- In Batch 2, UR-IW-1 (`gemini flash 2.0`) performed strongly (11th, **MRR** 0.5926).
- UR-IW-4 (`o4-mini + feedback`) was our top performer in Batch 4 (6th, **MRR** 0.5909).
- The systems with feedback often showed competitive **MRR** scores, but overall the results were mixed.

List Questions:

- UR-IW-4 (`o3-mini + feedback` or `o4-mini + feedback`) consistently performed well, ranking 28th in Batch 1 (**F-Measure** 0.5069) and 28th in Batch 2 (**F-Measure** 0.5188).
- In Batch 3, UR-IW-5 (`gemini flash 2.0 + 10-shot`) and UR-IW-3 (`gemini flash 2.0 + feedback`) were our leading systems.
- The results suggest that both few-shot prompting and feedback mechanisms can be beneficial, though their relative effectiveness varied across batches.

5. Discussion and Future Work

Model Performance: Based on the initial model selection tests and the BioASQ task 13B results, `gemini-2.0-flash` and its variants showed strong and consistent performance, particularly the 10-shot version (UR-IW-5) in retrieval tasks and Yes/No questions. `o3-mini` and `o4-mini` (UR-IW-2 and UR-IW-4 configurations) also proved to be competitive, especially in question answering tasks. `deepseek-reasoner` was competitive, particularly in Task A, Batch 1. But due to the slow API we were unable to complete runs with it in later batches, therefore opting for a proprietary replacement (o3-mini, o4-mini).

Impact of Self-Generated Feedback: The motivation to explore self-generated feedback stemmed from our ongoing research into comparing the impact of human expert feedback to LLM generated feedback. In these BioASQ preliminary results, the impact of adding a feedback step (UR-IW-3 and UR-IW-4 configurations) was mixed across all tasks and batches. For **Task A (Retrieval)**, feedback configurations did not consistently outperform the base models or 10-shot configurations in terms of **MAP** scores. For **Task A+ and Task B (Question Answering)**, feedback sometimes led to improvements. For instance, in Task B Yes/No questions, UR-IW-4 (with feedback) often surpassed UR-IW-2 (without feedback) in later batches. Similarly, in Task B Factoid questions, feedback systems showed competitive **MRR** scores. However, there were also instances where feedback did not lead to better or even resulted in worse performance compared to the base model or the few-shot model. The preliminary tests on model selection also hinted that self-feedback might not always enhance performance for some base models.

Few-Shot Learning vs. Feedback: The UR-IW-5 configurations, typically employing `gemini flash 2.0 + 10-shot` or `gemini-2.5-flash-preview + 10-shot`, frequently emerged as strong performers, especially in retrieval (Task A) and some question-answering sub-tasks (e.g., Task A+ Factoid Batch 4, Task B Yes/No Batch 1). This suggests that providing a few examples is still a successful way to guide these LLMs. When comparing `Gemini Flash 2.0` base (UR-IW-1) with its feedback version (UR-IW-3) and its 10-shot version (UR-IW-5), the 10-shot approach often had an edge, particularly in retrieval.

Best Suited Models and Approaches:

- For **retrieval tasks (Task A)**, `gemini flash 2.0 + 10-shot` (UR-IW-5) appeared to be the most promising approach among our submissions.
- For **Yes/No questions (Task A+ & B)**, `gemini flash 2.0` (base and 10-shot) and `o3-mini/o4-mini` (with and without feedback) all showed the ability to achieve high or perfect scores.
- For **Factoid questions (Task A+ & B)**, performance was more varied. `o3-mini/o4-mini` with feedback (UR-IW-4) and `gemini flash 2.0 + 10-shot` (UR-IW-5) had good performances in certain batches.

- For **List questions (Task A+ & B)**, `o3-mini` (UR-IW-2) had particularly strong showings in Task A+, Batch 1 and 2. In Task B, `o3-mini/o4-mini` with feedback (UR-IW-4) also performed well.

The choice of "best" model and approach appears to be task-dependent. Few-shot learning with `gemini-2.0-flash` seems broadly effective. The feedback mechanism shows potential but requires further refinement to ensure consistent improvements across diverse tasks and models. The preliminary test data indicated that `gemini-2.0-flash` had strong baseline factoid performance, which was reflected in some of the task results.

These are preliminary observations, and a more in-depth analysis will be conducted once the final, manually evaluated results are available. Future work will involve a more granular analysis of the generated answers and the types of errors made by different models and approaches to refine our strategies for future BioASQ challenges. The example code used for feedback and few-shot prompting can be found online[11].

6. Ethical Considerations

Even if the accuracy and reliability of LLM generated answers in RAG improve, they still tend to make subtle errors or hallucinate information that is not supported by the source documents. These errors can be especially difficult to catch when expert information needs such as the questions posed in the BioASQ challenge are answered. The output of these systems should therefore not be used to inform clinical decision-making without thorough expert oversight.

Another ethical issue is the environmental costs of complex multistep RAG systems. As each LLM call is processed on GPU clusters with SOTA models having billions of parameters distributed over these GPUs, every call produces considerably more co2 than a simple TF_IDF based search result ranking.

7. Conclusion

Overall, our feedback-based approach returned mixed results. There was no clear improvement over the zero-shot baselines with the same models. The few-shot approach from last year's participation that we reused as a baseline this year, was, according to the preliminary results, still the most competitive approach from our runs. It was also interesting to see that in our model selection test, the presumably cheaper and smaller distilled models (Gemini flash) were achieving better results than their pricier and presumably bigger counterparts (Gemini Pro) or the reasoning models (o3-mini, DeepSeek R1).

We will build on the introduced feedback approach in future work, comparing the impact of human and LLM generated feedback on overall task performance in professional search [28]. We believe this will be a valuable contribution to assess the performance of systems that foster human engagement vs. systems that promise full automation.

Acknowledgments

We thank the organizers of the BioASQ challenge for their continued support and quick response time. This work is supported by the German Research Foundation (DFG) as part of the NFDIxCS consortium (Grant number: 501930651).

Declaration on Generative AI

The authors used the following generative-AI tools while preparing this paper[12]:

[11] https://github.com/SamyAteia/bioasq2025
[12] https://ceur-ws.org/GenAI/Policy.html

- OpenAI ChatGPT (o3, 4o, 4.5 preview) (May 2025) - drafting content, latex formatting, paraphrase and reword.
- Google Gemini 2.5 Pro (May 2025) - drafting content, latex formatting, paraphrase and reword.
- LanguageTool - spellchecking, paraphrase and reword.

All AI-generated material was critically reviewed, revised and verified by the human authors. The authors accept full responsibility for the integrity and accuracy of the final manuscript.

References

[1] S. Suri, S. Counts, L. Wang, C. Chen, M. Wan, T. Safavi, J. Neville, C. Shah, R. W. White, R. Andersen, G. Buscher, S. Manivannan, N. Rangan, L. Yang, The Use of Generative Search Engines for Knowledge Work and Complex Tasks, 2024. URL: https://arxiv.org/abs/2404.04268. arXiv:2404.04268.

[2] B. G. Edelman, D. Ngwe, S. Peng, Measuring the impact of AI on information worker productivity, Available at SSRN 4648686 (2023).

[3] M. P. Bron, B. Greijn, B. M. Coimbra, R. van de Schoot, A. Bagheri, Combining large language model classifications and active learning for improved technology-assisted review, in: Proceedings of the International Workshop on Interactive Adaptive Learning (IAL@PKDD/ECML 2024), volume 3770 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 77–95. URL: https://ceur-ws.org/Vol-3770/paper6.pdf.

[4] S. Verberne, J. He, U. Kruschwitz, G. Wiggers, B. Larsen, T. Russell-Rose, A. P. de Vries, First International Workshop on Professional Search, SIGIR Forum 52 (2019) 153–162. URL: https://doi.org/10.1145/3308774.3308799. doi:10.1145/3308774.3308799.

[5] S. Verberne, Professional Search, 1 ed., Association for Computing Machinery, New York, NY, USA, 2024, p. 501–514. URL: https://doi.org/10.1145/3674127.3674141.

[6] N. F. Liu, T. Zhang, P. Liang, Evaluating verifiability in generative search engines, arXiv preprint arXiv:2304.09848 (2023).

[7] S. E. Spatharioti, D. Rothschild, D. G. Goldstein, J. M. Hofman, Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3714082. doi:10.1145/3706598.3714082.

[8] 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, G. Maria Di Nunzio, N. Ferro, 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: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[9] S. Ateia, U. Kruschwitz, Is chatgpt a biomedical expert?, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 73–90. URL: https://ceur-ws.org/Vol-3497/paper-006.pdf.

[10] S. Ateia, U. Kruschwitz, Can open-source llms compete with commercial models? exploring the few-shot performance of current GPT models in biomedical tasks, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 78–98. URL: https://ceur-ws.org/Vol-3740/paper-07.pdf.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is All You Need, in: Proceedings of the 31st International Conference on Neural

Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, preprint, 2018. URL: https://web.archive.org/web/20240522131718/https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[16] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, Z. Zhang, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, F. Xia, Q. Le, D. Zhou, Chain of Thought Prompting Elicits Reasoning in Large Language Models, ArXiv abs/2201.11903 (2022).

[18] C. Snell, J. Lee, K. Xu, A. Kumar, Scaling llm test-time compute optimally can be more effective than scaling model parameters, arXiv preprint arXiv:2408.03314 (2024).

[19] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, H. Hajishirzi, Tulu 3: Pushing Frontiers in Open Language Model Post-Training, 2025. URL: https://arxiv.org/abs/2411.15124. arXiv:2411.15124.

[20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[21] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguis-

tics, Punta Cana, Dominican Republic, 2021, pp. 3784–3803. URL: https://aclanthology.org/2021.findings-emnlp.320/. doi:10.18653/v1/2021.findings-emnlp.320.
[22] OpenAI, Deep Research System Card, https://cdn.openai.com/deep-research-system-card.pdf, 2025. System card, accessed 8 Jul 2025.
[23] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, J. Schulman, Webgpt: Browser-assisted question-answering with human feedback, 2022. URL: https://arxiv.org/abs/2112.09332. arXiv:2112.09332.
[24] T. Russell-Rose, P. Gooch, U. Kruschwitz, Interactive query expansion for professional search applications, Business Information Review 38 (2021) 127–137. URL: https://doi.org/10.1177/02663821211034079. doi:10.1177/02663821211034079. arXiv:https://doi.org/10.1177/02663821211034079.
[25] J. Higgins, Cochrane handbook for systematic reviews of interventions, Cochrane Collaboration and John Wiley & Sons Ltd (2008).
[26] S. Ateia, U. Kruschwitz, BioRAGent: A Retrieval-Augmented Generation System for Showcasing Generative Query Expansion and Domain-Specific Search for Scientific Q&A, in: European Conference on Information Retrieval, 2025.
[27] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.
[28] S. Ateia, From professional search to generative deep research systems: How can expert oversight improve search outcomes?, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (Doctoral Consortium), 2025. Doctoral Consortium, to appear.

A. Detailed Preliminary Results

Table 4
Task 13 Phase A Document Retrieval

Batch	Position	System	Precision	Recall	F-Measure	**MAP**	GMAP
Test batch 1	1 of 51	Top Competitor	0.1047	0.5043	0.1605	**0.4246**	0.0104
Test batch 1	**22** of 51	UR-IW-5	0.1677	0.3471	0.2038	**0.2865**	0.0015
Test batch 1	**26** of 51	UR-IW-1	0.1415	0.3194	0.1776	**0.2527**	0.0010
Test batch 1	**29** of 51	UR-IW-2	0.1376	0.2941	0.1699	**0.2272**	0.0007
Test batch 1	**32** of 51	UR-IW-3	0.1344	0.2547	0.1557	**0.2064**	0.0005
Test batch 1	**34** of 51	UR-IW-4	0.0979	0.1892	0.1135	**0.1739**	0.0001
Test batch 2	1 of 41	Top Competitor	0.0976	0.5093	0.1546	**0.4425**	0.0096
Test batch 2	**25** of 41	UR-IW-5	0.1930	0.3237	0.2088	**0.2634**	0.0011
Test batch 2	**26** of 41	UR-IW-4	0.1575	0.3184	0.1820	**0.2601**	0.0009
Test batch 2	**27** of 41	UR-IW-1	0.1643	0.2890	0.1855	**0.2523**	0.0008
Test batch 2	**28** of 41	UR-IW-3	0.1742	0.3064	0.1996	**0.2443**	0.0008
Test batch 2	**31** of 41	UR-IW-2	0.1181	0.2399	0.1335	**0.1846**	0.0005
Test batch 3	1 of 47	Top Competitor	0.0941	0.4228	0.1445	**0.3236**	0.0059
Test batch 3	**24** of 47	UR-IW-5	0.1341	0.2507	0.1560	**0.1834**	0.0005
Test batch 3	**27** of 47	UR-IW-1	0.1114	0.2283	0.1273	**0.1615**	0.0004
Test batch 3	**30** of 47	UR-IW-3	0.0854	0.2086	0.1093	**0.1456**	0.0004
Test batch 3	**31** of 47	UR-IW-2	0.0703	0.1588	0.0871	**0.1187**	0.0002
Test batch 3	**36** of 47	UR-IW-4	0.0644	0.1490	0.0818	**0.1043**	0.0001
Test batch 4	1 of 79	Top Competitor	0.0600	0.2512	0.0927	**0.1801**	0.0008
Test batch 4	**27** of 79	UR-IW-5	0.0427	0.1391	0.0632	**0.0794**	0.0002
Test batch 4	**31** of 79	UR-IW-4	0.0451	0.1371	0.0622	**0.0713**	0.0001
Test batch 4	**38** of 79	UR-IW-2	0.0408	0.1227	0.0555	**0.0655**	0.0001
Test batch 4	**39** of 79	UR-IW-1	0.0418	0.1040	0.0522	**0.0627**	0.0001
Test batch 4	**45** of 79	UR-IW-3	0.0396	0.0900	0.0460	**0.0574**	0.0001

Table 5
Task 13 Phase A Snippet Retrieval

Batch	Position	System	Precision	Recall	F-Measure	**MAP**	GMAP
Test batch 1	1 of 51	Top Competitor	0.0803	0.3050	0.1186	**0.4535**	0.0014
Test batch 1	**8** of 51	UR-IW-5	0.1189	0.1928	0.1202	**0.2768**	0.0006
Test batch 1	**9** of 51	UR-IW-1	0.0978	0.1594	0.1071	**0.2762**	0.0005
Test batch 1	**10** of 51	UR-IW-2	0.1136	0.1633	0.1110	**0.2478**	0.0005
Test batch 1	**11** of 51	UR-IW-3	0.0863	0.1393	0.0912	**0.2447**	0.0003
Test batch 1	**17** of 51	UR-IW-4	0.0795	0.1035	0.0778	**0.1844**	0.0001
Test batch 2	1 of 41	Top Competitor	0.0941	0.3625	0.1421	**0.5522**	0.0035
Test batch 2	**12** of 41	UR-IW-5	0.1407	0.1885	0.1290	**0.3080**	0.0009
Test batch 2	**13** of 41	UR-IW-1	0.1233	0.1713	0.1200	**0.3023**	0.0005
Test batch 2	**14** of 41	UR-IW-3	0.1287	0.1715	0.1212	**0.2949**	0.0007
Test batch 2	**18** of 41	UR-IW-4	0.1397	0.1635	0.1149	**0.2543**	0.0007
Test batch 2	**24** of 41	UR-IW-2	0.0916	0.1287	0.0877	**0.1654**	0.0003
Test batch 3	1 of 47	Top Competitor	0.0749	0.2855	0.1098	**0.4322**	0.0015
Test batch 3	**15** of 47	UR-IW-1	0.0838	0.1160	0.0806	**0.1534**	0.0003
Test batch 3	**18** of 47	UR-IW-5	0.0961	0.1271	0.0938	**0.1488**	0.0003
Test batch 3	**19** of 47	UR-IW-3	0.0602	0.1130	0.0745	**0.1463**	0.0002
Test batch 3	**24** of 47	UR-IW-2	0.0680	0.0662	0.0595	**0.0968**	0.0001
Test batch 3	**25** of 47	UR-IW-4	0.0478	0.0623	0.0475	**0.0721**	0.0001
Test batch 4	1 of 79	Top Competitor	0.0411	0.1135	0.0560	**0.1634**	0.0001
Test batch 4	**18** of 79	UR-IW-5	0.0308	0.0633	0.0399	**0.0511**	0.0001
Test batch 4	**20** of 79	UR-IW-3	0.0247	0.0492	0.0297	**0.0459**	0.0000
Test batch 4	**21** of 79	UR-IW-4	0.0229	0.0564	0.0314	**0.0446**	0.0001
Test batch 4	**22** of 79	UR-IW-1	0.0254	0.0639	0.0323	**0.0434**	0.0000
Test batch 4	**23** of 79	UR-IW-2	0.0250	0.0700	0.0344	**0.0411**	0.0001

Table 6
Task 13 B Phase A+ Yes/No questions

Batch	Position	System	Accuracy	F1 Yes	F1 No	**Macro F1**
Test batch 1	1 of 56	Top Competitor	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**1** of 56	UR-IW-1	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**1** of 56	UR-IW-2	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**1** of 56	UR-IW-4	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**21** of 56	UR-IW-3	0.9412	0.9565	0.9091	**0.9328**
Test batch 1	**42** of 56	UR-IW-5	0.8235	0.8696	0.7273	**0.7984**
Test batch 2	**1** of 49	UR-IW-5	1.0000	1.0000	1.0000	**1.0000**
Test batch 2	**10** of 49	UR-IW-1	0.9412	0.9565	0.9091	**0.9328**
Test batch 2	**13** of 49	UR-IW-3	0.8824	0.9000	0.8571	**0.8786**
Test batch 2	**14** of 49	UR-IW-2	0.8824	0.9091	0.8333	**0.8712**
Test batch 2	**37** of 49	UR-IW-4	0.7059	0.7368	0.6667	**0.7018**
Test batch 3	1 of 58	Top Competitor	0.9545	0.9697	0.9091	**0.9394**
Test batch 3	**5** of 58	UR-IW-5	0.9091	0.9412	0.8000	**0.8706**
Test batch 3	**10** of 58	UR-IW-4	0.8636	0.9091	0.7273	**0.8182**
Test batch 3	**19** of 58	UR-IW-1	0.8636	0.9143	0.6667	**0.7905**
Test batch 3	**26** of 58	UR-IW-2	0.8182	0.8824	0.6000	**0.7412**
Test batch 3	**42** of 58	UR-IW-3	0.6818	0.7742	0.4615	**0.6179**
Test batch 4	1 of 67	Top Competitor	0.9231	0.9444	0.8750	**0.9097**
Test batch 4	**2** of 67	UR-IW-2	0.9231	0.9444	0.8750	**0.9097**
Test batch 4	**34** of 67	UR-IW-3	0.8462	0.8889	0.7500	**0.8194**
Test batch 4	**37** of 67	UR-IW-1	0.8462	0.8947	0.7143	**0.8045**
Test batch 4	**39** of 67	UR-IW-4	0.8077	0.8485	0.7368	**0.7927**
Test batch 4	**43** of 67	UR-IW-5	0.8462	0.9000	0.6667	**0.7833**

Table 7
Task 13 B Phase A+ factoid questions

Batch	Position	System	Strict Acc.	Lenient Acc.	**MRR**
Test batch 1	1 of 56	Top Competitor	0.4231	0.5000	**0.4551**
Test batch 1	**7** of 56	UR-IW-2	0.3462	0.4231	**0.3782**
Test batch 1	**8** of 56	UR-IW-5	0.3462	0.4231	**0.3750**
Test batch 1	**24** of 56	UR-IW-1	0.2692	0.3462	**0.3077**
Test batch 1	**25** of 56	UR-IW-3	0.2692	0.3462	**0.3077**
Test batch 1	**29** of 56	UR-IW-4	0.2692	0.3077	**0.2885**
Test batch 2	1 of 49	Top Competitor	0.5926	0.5926	**0.5926**
Test batch 2	**2** of 49	UR-IW-4	0.5185	0.5556	**0.5370**
Test batch 2	**4** of 49	UR-IW-2	0.4815	0.5185	**0.5000**
Test batch 2	**14** of 49	UR-IW-1	0.4074	0.4815	**0.4383**
Test batch 2	**16** of 49	UR-IW-3	0.3704	0.4444	**0.3920**
Test batch 2	**29** of 49	UR-IW-5	0.2963	0.3333	**0.3086**
Test batch 3	1 of 58	Top Competitor	0.3500	0.4000	**0.3750**
Test batch 3	**7** of 58	UR-IW-3	0.2000	0.5000	**0.3100**
Test batch 3	**12** of 58	UR-IW-1	0.2000	0.4000	**0.2875**
Test batch 3	**18** of 58	UR-IW-2	0.2500	0.2500	**0.2500**
Test batch 3	**27** of 58	UR-IW-4	0.2000	0.2000	**0.2000**
Test batch 3	**29** of 58	UR-IW-5	0.1000	0.3000	**0.2000**
Test batch 4	**1** of 67	UR-IW-5	0.5455	0.5909	**0.5606**
Test batch 4	**4** of 67	UR-IW-1	0.5000	0.5455	**0.5152**
Test batch 4	**13** of 67	UR-IW-2	0.4545	0.4545	**0.4545**
Test batch 4	**14** of 67	UR-IW-3	0.4545	0.4545	**0.4545**
Test batch 4	**34** of 67	UR-IW-4	0.3636	0.4091	**0.3788**

Table 8
Task 13 B Phase A+ list questions

Batch	Position	System	Mean Prec.	Recall	**F-Measure**
Test batch 1	**1** of 56	UR-IW-2	0.2290	0.3056	**0.2567**
Test batch 1	**2** of 56	UR-IW-1	0.2070	0.3232	**0.2411**
Test batch 1	**3** of 56	UR-IW-5	0.2164	0.3003	**0.2395**
Test batch 1	**4** of 56	UR-IW-4	0.2134	0.2783	**0.2357**
Test batch 1	**15** of 56	UR-IW-3	0.1685	0.2804	**0.2004**
Test batch 2	1 of 49	Top Competitor	0.3785	0.4357	**0.3880**
Test batch 2	**2** of 49	UR-IW-2	0.3449	0.4626	**0.3805**
Test batch 2	**11** of 49	UR-IW-4	0.2859	0.3652	**0.3023**
Test batch 2	**17** of 49	UR-IW-1	0.2307	0.3536	**0.2682**
Test batch 2	**33** of 49	UR-IW-5	0.1796	0.3432	**0.2144**
Test batch 2	**34** of 49	UR-IW-3	0.1696	0.3213	**0.2118**
Test batch 3	1 of 58	Top Competitor	0.4674	0.4446	**0.4541**
Test batch 3	**13** of 58	UR-IW-5	0.3455	0.4111	**0.3618**
Test batch 3	**14** of 58	UR-IW-2	0.3656	0.3969	**0.3599**
Test batch 3	**20** of 58	UR-IW-1	0.3271	0.4040	**0.3482**
Test batch 3	**26** of 58	UR-IW-3	0.3114	0.3777	**0.3279**
Test batch 3	**37** of 58	UR-IW-4	0.2206	0.2896	**0.2371**
Test batch 4	1 of 67	Top Competitor	0.3217	0.2929	**0.3014**
Test batch 4	**9** of 67	UR-IW-5	0.2345	0.3680	**0.2742**
Test batch 4	**11** of 67	UR-IW-3	0.2640	0.3114	**0.2739**
Test batch 4	**34** of 67	UR-IW-4	0.2122	0.2936	**0.2270**
Test batch 4	**37** of 67	UR-IW-1	0.1819	0.3429	**0.2246**
Test batch 4	**42** of 67	UR-IW-2	0.1846	0.3349	**0.2172**

Table 9
Task 13 B Phase B Yes/No questions

Batch	Position	System	Accuracy	F1 Yes	F1 No	**Macro F1**
Test batch 1	1 of 72	Top Competitor	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**1** of 72	UR-IW-1	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**1** of 72	UR-IW-5	1.0000	1.0000	1.0000	**1.0000**
Test batch 1	**36** of 72	UR-IW-2	0.9412	0.9600	0.8889	**0.9244**
Test batch 1	**53** of 72	UR-IW-3	0.8824	0.9091	0.8333	**0.8712**
Test batch 1	**60** of 72	UR-IW-4	0.8235	0.8696	0.7273	**0.7984**
Test batch 2	**1** of 72	UR-IW-5	1.0000	1.0000	1.0000	**1.0000**
Test batch 2	**35** of 72	UR-IW-4	0.9412	0.9565	0.9091	**0.9328**
Test batch 2	**47** of 72	UR-IW-3	0.8824	0.9000	0.8571	**0.8786**
Test batch 2	**52** of 72	UR-IW-2	0.8824	0.9091	0.8333	**0.8712**
Test batch 2	**56** of 72	UR-IW-1	0.8824	0.9167	0.8000	**0.8583**
Test batch 3	1 of 66	Top Competitor	0.9545	0.9697	0.9091	**0.9394**
Test batch 3	**30** of 66	UR-IW-4	0.9091	0.9412	0.8000	**0.8706**
Test batch 3	**31** of 66	UR-IW-5	0.9091	0.9412	0.8000	**0.8706**
Test batch 3	**42** of 66	UR-IW-2	0.8636	0.9091	0.7273	**0.8182**
Test batch 3	**43** of 66	UR-IW-3	0.8636	0.9091	0.7273	**0.8182**
Test batch 3	**52** of 66	UR-IW-1	0.8636	0.9143	0.6667	**0.7905**
Test batch 4	1 of 79	Top Competitor	1.0000	1.0000	1.0000	**1.0000**
Test batch 4	**21** of 79	UR-IW-4	0.9231	0.9444	0.8750	**0.9097**
Test batch 4	**30** of 79	UR-IW-2	0.9231	0.9474	0.8571	**0.9023**
Test batch 4	**31** of 79	UR-IW-5	0.9231	0.9474	0.8571	**0.9023**
Test batch 4	**45** of 79	UR-IW-1	0.9231	0.9500	0.8333	**0.8917**
Test batch 4	**66** of 79	UR-IW-3	0.7692	0.8235	0.6667	**0.7451**

Table 10
Task 13 B Phase B factoid questions

Batch	Position	System	Strict Acc.	Lenient Acc.	**MRR**
Test batch 1	1 of 72	Top Competitor	0.5385	0.6538	**0.5962**
Test batch 1	**17** of 72	UR-IW-3	0.4231	0.5769	**0.4821**
Test batch 1	**26** of 72	UR-IW-4	0.4231	0.5000	**0.4615**
Test batch 1	**27** of 72	UR-IW-5	0.4231	0.5000	**0.4615**
Test batch 1	**33** of 72	UR-IW-2	0.4231	0.5000	**0.4551**
Test batch 1	**37** of 72	UR-IW-1	0.3846	0.5385	**0.4423**
Test batch 2	1 of 72	Top Competitor	0.7037	0.7037	**0.7037**
Test batch 2	**11** of 72	UR-IW-1	0.5556	0.6296	**0.5926**
Test batch 2	**21** of 72	UR-IW-2	0.5185	0.5926	**0.5556**
Test batch 2	**22** of 72	UR-IW-4	0.5185	0.5926	**0.5556**
Test batch 2	**30** of 72	UR-IW-5	0.5185	0.5556	**0.5370**
Test batch 2	**35** of 72	UR-IW-3	0.5185	0.5556	**0.5309**
Test batch 3	1 of 66	Top Competitor	0.4000	0.6500	**0.5100**
Test batch 3	**24** of 66	UR-IW-1	0.3500	0.4500	**0.3725**
Test batch 3	**31** of 66	UR-IW-4	0.3000	0.3500	**0.3250**
Test batch 3	**32** of 66	UR-IW-5	0.2500	0.4000	**0.3250**
Test batch 3	**38** of 66	UR-IW-3	0.2500	0.3500	**0.3000**
Test batch 3	**43** of 66	UR-IW-2	0.2500	0.3000	**0.2750**
Test batch 4	1 of 79	Top Competitor	0.6364	0.6364	**0.6364**
Test batch 4	**6** of 79	UR-IW-4	0.5455	0.6364	**0.5909**
Test batch 4	**17** of 79	UR-IW-1	0.5455	0.5909	**0.5606**
Test batch 4	**27** of 79	UR-IW-5	0.5000	0.5455	**0.5227**
Test batch 4	**30** of 79	UR-IW-2	0.5000	0.5000	**0.5000**
Test batch 4	**46** of 79	UR-IW-3	0.4545	0.4545	**0.4545**

Table 11
Task 13 B Phase B List questions

Batch	Position	System	Mean Prec.	Recall	**F-Measure**
Test batch 1	1 of 72	Top Competitor	0.5820	0.6224	**0.5959**
Test batch 1	**28** of 72	UR-IW-4	0.4817	0.5601	**0.5069**
Test batch 1	**50** of 72	UR-IW-2	0.3740	0.4944	**0.4042**
Test batch 1	**52** of 72	UR-IW-1	0.3361	0.5653	**0.3978**
Test batch 1	**57** of 72	UR-IW-3	0.3199	0.5419	**0.3769**
Test batch 1	**60** of 72	UR-IW-5	0.2877	0.5341	**0.3515**
Test batch 2	1 of 72	Top Competitor	0.6360	0.6315	**0.6152**
Test batch 2	**28** of 72	UR-IW-4	0.4610	0.6425	**0.5188**
Test batch 2	**31** of 72	UR-IW-3	0.4312	0.6771	**0.5010**
Test batch 2	**41** of 72	UR-IW-1	0.4088	0.6561	**0.4716**
Test batch 2	**47** of 72	UR-IW-5	0.3916	0.6048	**0.4463**
Test batch 2	**49** of 72	UR-IW-2	0.3833	0.5784	**0.4407**
Test batch 3	1 of 66	Top Competitor	0.6433	0.6429	**0.6337**
Test batch 3	**44** of 66	UR-IW-5	0.4465	0.5790	**0.4783**
Test batch 3	**45** of 66	UR-IW-3	0.4324	0.6102	**0.4774**
Test batch 3	**46** of 66	UR-IW-4	0.4472	0.5272	**0.4687**
Test batch 3	**47** of 66	UR-IW-1	0.4371	0.6368	**0.4684**
Test batch 3	**54** of 66	UR-IW-2	0.4009	0.5549	**0.4369**
Test batch 4	1 of 79	Top Competitor	0.7491	0.5980	**0.6492**
Test batch 4	**41** of 79	UR-IW-3	0.4140	0.6127	**0.4711**
Test batch 4	**46** of 79	UR-IW-4	0.4163	0.5536	**0.4479**
Test batch 4	**47** of 79	UR-IW-5	0.3671	0.5911	**0.4338**
Test batch 4	**55** of 79	UR-IW-1	0.3303	0.5425	**0.3872**
Test batch 4	**56** of 79	UR-IW-2	0.3285	0.5413	**0.3797**