LLM Fine-Tuning With Biomedical Open-Source Data

Christopher Anaya[1], Maria Fernandes[2,1] and Francisco M Couto[1]

[1]*LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2]*Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark*

Abstract

In BioASQ Task 12b, we explored the potential of enhancing Large Language Models (LLMs) with external biomedical data. We fine-tuned Mistral-7B-Instruct v0.1 using open-source data and efficient techniques like QLoRA. To further enrich the model's knowledge, we incorporated manually curated biomedical data alongside open-source resources. During the competition, our model tackled three question types: yes/no, factoid, and summary. While the results weren't competitive, the process identified key areas for improvement, including data augmentation, hyperparameter tuning, and automation—aspects we intend to address in future iterations. The data is available at our group's GitHub: https://github.com/lasigeBioTM.

Keywords

BioAQ challenge Task 12b, Question answering, LLMs, Fine-tuning

1. Introduction

BioASQ [1] is an annual competition that releases a series of challenges in the area of Biomedical Semantic Indexing and Question Answering (QA). Specifically, Task 12b: Biomedical Semantic QA Phase b consists of responding to test questions in the biomedical domain with relevant answers [2]. The challenge is organized into a series of four question batches, usually, taking place between March and May. The competition provides a comprehensive training set, composed of the previous years batch releases with the corresponding answers. The participants then use this training set as guidance to develop their Question Answering systems. The training set is a manually curated corpus by a committee of experts in the field [3].

Recent developments in state-of-the-art QA models used in task b have largely mirrored advancements in transformer models in the wider Natural Language Processing community. Before the 11th edition of BioASQ in 2023, a majority of competitive models were built atop BERT based models that had been fine-tuned for a biomedical QA downstream task. Task 11b saw the emergence of GPT models yielding competitive results. Oftentimes these language models were modified minimally or not at all, relying on a prompt-learning strategy to improve results and appropriately format the responses.

In parallel, open-source Large Language Models (LLMs), such as LLaMA and Mistral 7B have removed barriers for end users to develop and implement language models. This has been carried out both through the publication of open-source models themselves, as well as through the development of computationally-efficient models that can run on consumer grade hardware.

For this work, we set out with the objective of developing a QA model from an available open-source LLM. Our goal was to test whether these models could be significantly improved by

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

incorporating additional specialized curated data and integrating specialized external knowledge bases. By participating in BioASQ 12b [2], we aimed to benchmark our model's performance against similar models and evaluate the impact of these enhancements.

2. Previous Work

The arrival of the Transformer [4] revolutionized NLP, with BERT [5] excelling in extracting knowledge from text and being adapted for biomedical QA through models like BioBERT [6]. Concurrently developed GPT [7] models have recently outperformed encoder models, generating responses autoregressively using large, diverse datasets and a next-word prediction objective. OpenAI's ChatGPT, starting from version 3.0, showcases these advanced, though not open-source, GPT models, driven by significant investments in computing infrastructure.

In recent time, efforts have been made to develop open-source LLMs that aim to have comparable performance to the popular GPT models. Many of these models have had an additional objective of increasing computational efficiency of the model prediction function so that end users without large computational resources can implement these LLMs in their own systems. Mistral 7B [8] exemplifies the use of Sliding Window Attention, which reduces computational load by limiting attention to a sliding window of nearby elements. Additionally, Parameter Efficient Fine-Tuning techniques, such as QLoRA (Quantized Low-Rank Adaptors), enhance LLM fine-tuning efficiency. QLoRA reduces memory usage for weight storage (quantization) and approximates weight updates with low-rank matrix products. Together, these methods enable computationally efficient fine-tuning of large language models.

3. Methods

This section first introduces the selected data sources used to fine-tune the model. Subsequently, it describes the model selection and training process used to create our QA system for participation in BioASQ Task 12b [2].

3.1. Data Sources

3.1.1. BioASQ Training Data 12b

The first dataset used for fine-tuning our model was the manually curated training data provided by the BioASQ challenge organizers. This dataset, comprising questions, text snippet and answers, serves as the gold standard for expected answers and aligns with the format required for BioASQ Task 12b [2]. It includes four question types: yes/no, factoid, list, and summary, and two answer types: exact answers (keywords or phrases) and ideal answers (a few sentences).

3.1.2. GO Data

Gene Ontology [9, 10] is a structured data resource, which was developed with the aim of standardising the representation of our knowledge about genes and related concepts. It provides a hierarchy of terms and definitions related to genes. This data resource features a wide range of biological terms definitions that are of high relevance for the biomedical QA field.

The version of the GO dataset at the time of download was composed of 47,735 entries, which were all considered for the fine-tuning of our model. An entry in GO has a unique identifier, a GO ID or accession number, with the format GO then seven numerical digits. It also has a term name and a definition. From the term name we can generate a question ”What is x?” and we can use the definition as the answer. As an example, the term ”neuron recognition” with GO ID GO:0008038, has the following definition: The process in which a neuronal cell in a multicellular organism interprets its surroundings. From this data entry, we can generate a question ”What is neuron recognition?” with the definition as the answer.

3.1.3. DrugBank Data

DrugBank [11, 12] is a pharmaceutical knowledge database, a comprehensive repository with data about existing drugs, drug targets, and other drug-related information. This is an important resource for drug researchers and healthcare practitioners. We sought to use training data from DrugBank to integrate more drug-related knowledge in our model. The data from DrugBank is available for research upon access request and validation.

We used DrugBank data from release version 5.1.12., which encompasses more than 500,000 drugs and drug products. From the DrugBank data entries, we used the name to formulate a question and the description field to generate an answer. For example, the drug Aspirin with DrugBank ID DB00945 has the following entry for description: ”Aspirin, also known as acetylsalicylic acid (ASA), is a medication used to reduce pain, fever, or inflammation. Specific inflammatory conditions in which aspirin is used include Kawasaki disease, pericarditis, and rheumatic fever. It is also used long-term to help prevent heart attacks, strokes, and blood clots in people at high risk. Aspirin is an NSAID and works by inhibiting the enzyme cyclooxygenase.” From this data we can create a question ”What is Aspirin?” with the description as the answer.

3.1.4. BiQA Data

BiQA [13], consisting of mined data from forum-based social media platforms like Reddit, Stack Exchange, and Quora. The relevant data fields include the following: user-submitted questions and submissions of PubMed citations for articles that could answer the submitted question, provided by different users. We considered this dataset, as it presented a compiled biology related corpora (released in April 2020).

A particular challenge verified in the use of BiQA data was that the user-submitted answers were not directly saved in the dataset. Therefore, we manually annotated the answer for each question in the dateset, considering the cited articles as well as outside sources. Additionally, we excluded any question not relevant to the biomedical context. Upon processing the data, we noticed that the majority of the questions in the dataset would lie on the summary category.

As the curation and context validation was a full manual and time-consuming process, we compiled a total of 714 manually curated BiQA question-answer pairs. The following is an example of a generated data entry—question: ”Why does methylation not occur in viral DNA?” with answer: ”Overall, the absence of DNA methylation in viral DNA may be attributed to various factors, including viral replication strategies, genome size, evolutionary pressures, and interactions with host cell processes. While some viruses may encode proteins that can modulate host cell methylation machinery, the overall role of DNA methylation in viral replication and pathogenesis remains an active area of research."

3.2. Model Selection

LLMs have demonstrated comparable performance to BERT-like models for a variety of NLP tasks. [14] However, with the ease of training LLMs from unprocessed data and from the advances in computational efficiency, we assert that LLMs provide a more promising approach to base QA models.

Therefore, we focused on LLMs, and furthermore, our model selection relied on three main criteria: (i) model availability, (ii) performance, and (iii) fine-tuning capability. We considered open-source LLMs for increased usability. We searched for reported comparisons of existing open-source LLMs, i.e., Llama 2 and Mistral-7B [15]. At the time, January 2024, those were the main open-source high-performance LLMs. Regarding the fine-tuning, we looked for customization and training time, as our goal was to be able to integrate biomedical data and improve the QA task results. We chose to use Mistral-7B-Instruct v0.1 as our underlying language model, which is an already fine-tuned version of Mistral-7B model [8]. This was due to its strong performance among open source LLMs and its compatibility with computationally efficient fine tuning methods.

3.3. BiQA manual curation

Starting from the BiQA dataset, we manually evaluated the relevance of each question for Biomedical Research. As the questions were posed by users from the general public, some were not well posed or were not within a biomedical context. We then accessed the PubMed API and retrieved the abstracts of the cited articles. For relevant questions, we classified each question by type, as defined by BioASQ. We then used the article abstracts as well as other open source information to annotate an answer to the question. The question-answer pairs were then formatted in batches similar to what BioASQ uses.

3.4. Fine-Tuning

The ”out-of-the-box” Mistral-7B-Instruct v0.1 model is a general purpose fine-tuned generative text model. From this base model, we further fine-tuned this model to improve performance on biomedical QA task. Due to the model size (7 billion parameters) of Mistral-7B-Instruct, we needed to use a Parameter Efficient Fine-Tuning method; if we had not used a PEFT method, the computation time would be prohibitively long. We chose QLoRA, which combines datatype quantization with Low Rank Adaptors (LoRA), as our PEFT method, as it works well with our chosen language model, Mistral 7B. The fine-tuning was made through prompt-learning, a technique where the user provides instructions to the model in order to integrate further knowledge in it. This required a pre-processing step, where we integrate the question with the instructions on how the model should answer the question, and provide the answer. Further pre-processing details are described for each dataset, alongside the corresponding training

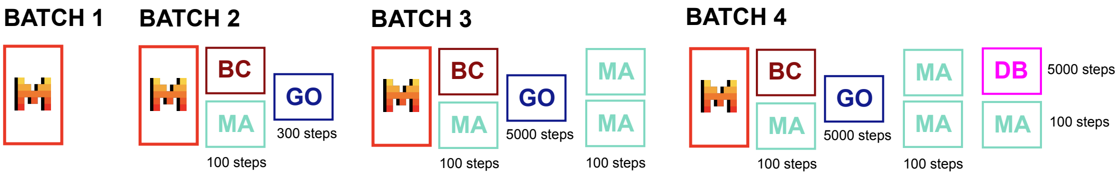

Figure 1: Overview of the fine-tuning procedure and its alignment with the competition batches. M is the base Mistral 7B instruct model. BC denotes the BioASQ provided training dataset. MA are the manual annotations we generated from BiQA. GO refers to the Gene Ontology dataset. DB is the DrugBank dataset.

parameters, which were adapted to the different dataset sizes. Therefore, under the assumption that the model should not repeat questions to not overfit we adjusted the maximum number, taking into account the batch size per GPU. The learning rate was kept constant through all fine-tuning steps (0.00025), and per fine-tuning round 3 evaluation steps.

Regarding input format, the model accepts JSONL as input and the answers are saved into a text file (.txt), which is then post-processed back to JSON format for the challenge answers submissions.

To optimize our process and ensure the robustness of our results, we initially conducted a brief fine-tuning of GO terms for batch 2, limiting the steps to 300 due to time constraints. For batch 3, however, we extended the fine-tuning to 5000 steps. This adjustment aimed to enhance the quality of our findings while managing the risk of over-fitting.

3.4.1. BioASQ Training Data 12b

BioASQ training data featured our first fine-tuning round and it was provided in JSON format. It shares the same format as BioASQ test data, with provided question, question type, exact and ideal answers, text snippets among other data fields. For our fine-tuning, we only used these data fields in the model input. An example of a data entry is the following—body: ”Is Hirschsprung disease a mendelian or a multifactorial disorder?”, type: summary, ideal answer: ”Coding sequence mutations in RET, GDNF, EDNRB, EDN3, and SOX10 are involved in the development of Hirschsprung disease. The majority of these genes was shown to be related to Mendelian syndromic forms of Hirschsprung's disease, whereas the non-Mendelian inheritance of sporadic non-syndromic Hirschsprung disease proved to be complex; involvement of multiple loci was demonstrated in a multiplicative model.”. As this example belongs is for a 'summary' question there is no exact answer, as defined by BioASQ.

Batch size is the number of training examples utilized in one iteration and is described in each data type fine-tuning section.

3.4.2. GO terms fine-tuning

For the fine-tuning with GO terms data, the training was conducted for a maximum of 100 steps (iterations). Additionally, in each step, a batch size of 6 was used, meaning that 6 data samples were processed together in each iteration.

GO, as the name indicates, is an ontology [1] We used the defined GO term and the corresponding definition. Due to limited time, this data was only considered for summary questions fine-tuning. GO terms training set was built from including each GO term into ”What is ...?” question structure and provide as reply the corresponding description. From the example given in the data sources section, this corresponds to ”What is neuron recognition?” In case of multiple paragraphs, the answer/description was restricted to the first paragraph. We opt for this approach as BioASQ ideal answer was limited to 200 words.

For the fine-tuning with GO terms data we used a maximum number of steps of 5000, and batch size of 6.

3.4.3. DrugBank fine-tuning

For our approach we used only the description field, truncated at the first paragraph due to answers length limitation. Similar to the GO terms setting, here we also used DrugBank only for summary questions fine-tuning. Drawing from our previous example, for the drug entry ”Aspirin”, the given question is ”What is Aspirin?” and the corresponding answer will be the content of the Description section in the database.

For the fine-tuning with DrugBank data we used a maximum number of steps of 2300, and batch size of 4.

3.4.4. BiQA data fine-tuning

Depending on the annotated questions dataset size, we adjusted the number of steps, which was 100, and fixed the batch size to 4, for all the BiQA fine-tuning rounds.

3.5. Metrics

The BioASQ competition results for Task 12b [2] are evaluated using different methods depending on question type and answer type. [16] For yes/no questions, the official metric is the ”macro F1 score” between the yesses and nos. For the factoid questions, the official metric is Mean Reciprocal Rank. Here, reciprocal rank refers to the inverse of the rank or position of the entry containing the correct entity. For the list questions, the metric used by the competition is mean average F-measure, which is a ”micro F1 score” among all correct answer classes given in the golden entity list.

For the ideal answers, BioASQ conducts a manual scoring of submitted answers. However, a set of automatic metrics, ROUGE scores, [17] are also calculated. These scores use n-grams and skip-grams to characterize the similarity between the ideal answers generated by experts and submitted by participant systems.

[1] Ontology – is a structured representation of a set of concepts and categories in a domain, which also describes their properties and relations between them.

4. Results

Our QA model participated in BioASQ Task 12b [2], in all four test batches, that is, rounds of competition, for yes/no and summary questions, and in factoid questions for batches three and four. The different evaluation metrics are summarized in Tables 1 and 2.

Table 1
Evaluation results from BioASQ Task 12b - yes/no and factoid questions

Batch	yes/no				factoid		
	Accuracy	F1 Yes	F1 No	Macro F1	Strict Acc.	Lenient Acc.	MRR
Batch 1	0.6400	0.7568	0.3077	0.5322	-	-	-
Batch 2	0.6154	0.7222	0.3750	0.5486	-	-	-
Batch 3	0.5000	0.2500	0.6250	0.4375	0.0769	0.0769	0.0769
Batch 4	0.4074	0.5294	0.2000	0.3647	0.1579	0.1579	0.1579

Table 2
Evaluation results from BioASQ Task 12b - summary questions

Batch	summary			
	R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
Batch 1	0.0516	0.0331	0.0808	0.0432
Batch 2	0.0416	0.0227	0.0783	0.0404
Batch 3	0.1122	0.0474	0.1436	0.0607
Batch 4	0.0845	0.0543	0.1172	0.0719

As the competition rounds proceeded, we observed different trends depending on question type. For yes/no questions, we saw a decrease in performance as measured by Accuracy and Macro-F1 over the competition, with a Batch 1 accuracy of 0.64 and an Batch 4 accuracy of 0.4074. For the automated summary scores, there was a slight improvement over the course of the competition over all of the ROUGE metrics. The factoid scores, both accuracy metrics and the MRR score, also improved between Batches 3 and 4, the two in which we competed.

In terms of peer systems, the results obtained were within the middle part of the competition results. As an illustrative example, for Batch 1 yes/no performance, the accuracy scores ranged from 0.04 and 0.96, with our model performing at 0.64. However, we started with a limited participation in yes/no and summary questions, and expanded our model to provide results for factoid question.

5. Discussion and Conclusion

Our model did not achieve competitive results compared to peer systems participating in the BioASQ challenge Task 12B.

The adopted strategies for fine-tuning with open-source biomedical data allowed the improvement of QA, with some fluctuations. We verified improvement when fine-tuning with large datasets, such as GO terms and DrugBank data for the factoid MRR scores as well and the various ROUGE scores, while when using manually curated data, as is the case of BiQA data, the model often dropped its performance. One possible reason behind the decreased performance is the size of the dataset, where manual annotation necessarily is time-intensive and limited our capacity to generate data in large volume.

Of note, we observed a significant improvement in the ROUGE scores from batch 2 to batch 3. This is probably due to the re-run of GO terms fine-tuning with 5000 steps (batch 3) instead of the previous 300 steps (batch 2), which allows a better integration of the data features without much underfiting. This is only demonstrated for the summary questions evaluation metrics, as GO content was only used for fine-tuning in this category of questions.

The use of the snippets field was limited to the BioASQ training set, therefore we did not included in our model prompt-learning. We faced several challenges when including the large volume of text in the snippets in the prompt-learning methodology, namely model confusion an hallucination. We are aware of the importance of using snippets in our training and prediction procedures to develop QA systems that include context information in generating answers.

Developing a new QA system from a base language model has yielded many areas where we can improve in the future. We believe by adding more data sources, specifically from large open source repositories, and better curating the included data we can improve our system. Particularly, we will focus on generating question-answer data in question types other than summaries. Also, we can expand our hyperparameter space for our underlying language model and for the fine tuning process. The integration of snippets data will be another major improvement in our QA system, providing it with a better context for each question that can be integrated in the answer generation. Where possible, we would like to automate as many processes as possible to facilitate a greater volume of data processing. This would be tied to finding data sources that require minimal manual annotation.

Acknowledgments

This work was funded in part by FCT through LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020), and in part by the Novo Nordisk Foundation through NNF23SA0084103 and NNF18CC0034900".

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 12b and Synergy12 in CLEF2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de

Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[3] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, Bioasq-qa: A manually curated corpus for biomedical question answering, Scientific Data 10 (2023) 170. URL: https://doi.org/10.1038/s41597-023-02068-4. doi:10.1038/s41597-023-02068-4.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[5] 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, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[6] 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, CoRR abs/1901.08746 (2019). URL: http://arxiv.org/abs/1901.08746. arXiv:1901.08746.

[7] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[9] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, Gene ontology: Tool for the unification of biology, The Gene Ontology Consortium. Nat Genet 25 (2000) 25–29.

[10] T. G. O. Consortium, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J.

Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, M. Westerfield, The Gene Ontology knowledgebase in 2023, Genetics 224 (2023) iyad031. doi:10.1093/genetics/iyad031.

[11] D. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, Drugbank: A comprehensive resource for in silico drug discovery and exploration, Nucleic acids research 34 (2006) D668–72. doi:10.1093/nar/gkj067.

[12] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. L. Chin, S. Strawbridge, M. Garcia-Patino, R. Kruger, A. Sivakumaran, S. Sanford, R. Doshi, N. Khetarpal, O. Fatokun, D. Doucet, A. Zubkowski, D. Y. Rayat, H. Jackson, K. Harford, A. Anjum, M. Zakir, F. Wang, S. Tian, B. Lee, J. Liigand, H. Peters, R. Q. R. Wang, T. Nguyen, D. So, M. Sharp, R. da Silva, C. Gabriel, J. Scantlebury, M. Jasinski, D. Ackerman, T. Jewison, T. Sajed, V. Gautam, D. S. Wishart, DrugBank 6.0: the DrugBank Knowledgebase for 2024, Nucleic Acids Research 52 (2023) D1265–D1275. doi:10.1093/nar/gkad976.

[13] A. Lamurias, D. Sousa, F. M. Couto, Generating biomedical question answering corpora from q&a forums, IEEE Access 8 (2020) 161042–161051. doi:10.1109/ACCESS.2020.3020868.

[14] M. Luo, K. Hashimoto, S. Yavuz, Z. Liu, C. Baral, Y. Zhou, Choose your qa model wisely: A systematic study of generative and extractive readers for question answering, 2022. arXiv:2203.07522.

[15] Mistral 7b: The best 7b model to date, apache 2.0, https://mistral.ai/news/announcing-mistral-7b/, 2024. Accessed: 2024-05-29.

[16] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artiéres, A.-C. N. Ngomo, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (2015) 138. URL: https://doi.org/10.1186/s12859-015-0564-6. doi:10.1186/s12859-015-0564-6.

[17] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.