Investigating General-Purpose Large Language Models for Patient Information Extraction: A Case Study on Real-World Cardiac MRI Reports

Sebin Sabu[1], Pavithra Rajendran[1], Ewart Jonny Sheldon[1], Alexandros Zenonos[2], Shiren Patel[1], Andrew Taylor[1], Rebecca Pope[2], Neil Sebire[1]

[1]Great Ormond Street Hospital for Children NHS Foundation Trust
[2]Roche Products Ltd.

Abstract

Electronic Patient Record (EPR) systems within healthcare systems contains a significant volume of free text written by clinicians in the form of unstructured data, meaning access to timely, potential pertinent data signals is precluded. For a clinician to analyse information for a cohort of patients for research, information extracted from unstructured data needs to be mapped with the routinely collected standard structured information and this can require lot of manual work and time. This paper studies the potential capabilities of general-purpose Large Language Models (LLMs) in the context of, (1) practical deployment using limited CPU computing resources, (2) usefulness in the context of extracting patient information within healthcare settings and (3) does not require fine-tuning or train models from scratch. In particular, we have investigated the utility of prompt-based zero-shot predictions by adapting these models in a question answering framework, which is deployed and run within a secure on-premise environment with CPU servers for extracting ten years of retrospective data containing 15,376 Cardiac MRI reports. Results are evaluated on a ground-truth dataset containing 400 randomly selected reports across the ten year period with the best performance having an averaged F1-score of 97.83%. Source code will be made available upon acceptance.

Introduction

Clinical decision making and research is often based on retrospective analysis of patient cohorts to guide prognostic or treatment decisions. Within the healthcare setting, a large volume of pertinent data is in the form of unstructured texts that hold extremely valuable clinical information that is not easily accessible for further study. As part of routine clinical care, by accessing information present within such unstructured texts, many documents and reports can be generated that contain important clinical information about a patient's condition, their presentation and phenotype, as well as conclusions drawn by medical professions which contribute to the overall diagnosis and prognosis. However, there are several challenges for current processes. Firstly, this requires extracting data elements from thousands of narrative reports which traditionally requires manual resource from the limited number of expert annotators with sufficient domain knowledge. This data must be linked to existing structured data and in order to do this, patient identifiable information present in reports are required to be extracted.

In order to address the above challenge, we have proposed an automated approach [1] that can help researchers and clinicians access structured information from large volumes of data in a timely format as described below:

- Modularised and reusable Question Answering pipeline for extracting the key information is built and deployed within our secure on-premise infrastructure. The pipeline is run in a scalable manner to process large volumes of data within the infrastructure where all required information governance protocols are adhered to (e.g., no patient data leaves the hospital).
- Automatically processing large volumes of unstructured texts for identifying patient key identifiers and mapping routinely collected data present within the unstructured texts with existing standard structured data used for clinical decision making and secondary research purposes.

In our approach, we follow a two-step automated process namely $\mathbf{S}_1$ and $\mathbf{S}_2$ as explained below:

- $\mathbf{S}_1$ is where documents $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_l\}$ are processed using simple, task-agnostic rule-based techniques for extracting tables and texts separately as an intermediate representation.
- $\mathbf{S}_2$ represents a prompt-based question answering pipeline with a set of predefined prompts $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_j\}$ and a set of NLP models $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_n\}$. The best performing model on a held-out validation data is then used within the pipeline for extracting patient identifiers from the documents. The outcomes are used to map the extracted data with our existing standard structured data.

Our proposed approach is used to extract relevant patient information from ten years of retrospective cardiac MRI reports (15,376). Figure 1 shows some example report structure and snippets of dummy data. From this, we randomly picked 400 reports across the time period. Manually annotated data are curated from this held-out dataset and used for evaluating results.

Our contributions in this paper are as follows:

[1]https://github.com/gosh-dre/AIMedHealth2025_CardiacMRIReports

- We explore the deployment of a prompt-based question answering pipeline that utilises existing "smaller" open-source LLMs within a real-world healthcare setting. We have evidenced the generalized and emergent behaviours of these LLMs using our investigation on automatically extracting entities within our clinical data, with the best performance on a groundtruth dataset having an averaged F1-score of 97.83%.
- Our results show the enhanced capabilities of general-purpose LLMs with the help of prompting templates in a zero-shot setting. Our experimental results and findings illustrate the potential advantage of existing general-purpose LLMs specifically **Flan-T5** (Longpre et al. 2023; Kanakarajan and Sankarasubbu 2023) within a real-world healthcare setting.
- Our approach makes use of prompting templates that are adaptable across various unstructured healthcare documents, which we have demonstrated on the cardiac MRI reports. This can enable automatic processing of large volumes of documents, in particular, similar types of reports by extracting relevant patient information with limited human effort.

Related Work

Clinical Data

A large volume of data within the healthcare setting (Sedlakova et al. 2023) is present as unstructured data, creating a barrier to access of such data by researchers, which in turn, reduces the progress within the research community. Given that much of the unstructured data contains valuable clinical information, the lack of availability of this data (Hemingway et al. 2018) can cause biases and hinder validation studies since most clinical information remains inaccessible at scale.

Document Understanding

Much of the reports stored within the Electronic Patient Record (EPR) systems are in the form of PDF reports, including tabular content, making it more challenging for extracting structured data based on rows and cells (Yang et al. 2021). Recent work on document understanding make use of transformer-based models for extracting information from documents that uses visual and spatial features and requires a large volume annotated data (Appalaraju et al. 2021; Xu et al. 2020). In this study, we do not focus on training from scratch or fine-tuning a transformer-based model on a large volume of data. Instead, we explored relatively simpler approaches for extracting information from texts.

Question Answering and Prompt Adaptation

One challenge in developing an entity recognition system requires the manual effort of annotating large volumes of data within the clinical context, which requires domain experts. Recent work has demonstrated capability of LLMs to work with language sequences of varying length to provide good text responses in a zero-shot setting (Dagdelen et al. 2024). (Singhal et al. 2023) showed a strong performance improvement with scaling of the parameters. Prior work (Tirskikh and Konovalov 2023; Liu et al. 2022) have also shown the potential capabilities on using pretrained extractive question answering models for detecting entities where, for a given question, the answer is predicted from a provided context. In this work, we investigate on zero-shot predictions in a question answering setting using both pretrained extractive question answering models and instruction-tuned text generation models.

Data

The study includes Cardiac MRI reports in PDF format for patients tested within our hospital over a period of ten years and each patient has multiple reports across the span of timelines. In total, we extracted 15,376 reports. From this, 400 reports were held-out for manually annotating the key information for validation purpose. The 400 reports were collected randomly across different timelines to ensure that we have a consistent performance.

Our Proposed Approach

In our proposed approach, Step 1 is used to extract an intermediate representation of content within documents for further processing. This is then fed into the pipelines present within Step 2. In Step 2, the raw text content present within the documents is used for processing the information and detecting the patient identifier information (see Table 1 *Data Representation* column). The pipelines developed for performing the different steps are explained in detail below.

S_1: DocProcessor

Given a PDF document $\mathbf{d}_i \in \mathcal{D}$, such that $\mathcal{D}$ represents the set of all reports, our custom developed DocProcessor pipeline uses an open-source Python package [2] for extracting the texts and tables separately and outputted as an intermediate representation $\mathbf{d}_i^{Inter}$ and used for further work.

S_{2a}: QnA Prompt based Entity Detection

Given an intermediate representation $\mathbf{d}_i^{Inter}$ for a document $\mathbf{d}_i$ and different existing open-source LLMs $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_n\}$, our custom pipeline is developed to perform an extractive question answering task. The extractive question answering task is where a question is provided with some context and key patient identifiers are extracted from the context as potential answers in a zero-shot setting. From the intermediate representation $\mathbf{d}_i^{Inter}$, the text content is further divided into chunks represented as $\mathbf{d}_{iv}^{Inter}$ such that v represents the chronological order of the chunks. For a given input $\mathcal{I}(\mathbf{d}_i^{Inter}) \equiv \mathcal{I}(\mathbf{Q}, \mathbf{d}_{i0}^{Inter}, \mathcal{P})$ where $\mathbf{Q}$ represents a predefined question template, $\mathbf{d}_{i0}^{Inter}$ represents the first chunk of text from the intermediate representation that is used as context and $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_j\}$ are predefined prompt templates that are appended to the context from which the answer is retrieved.

[2] https://pypi.org/project/pdfplumber/

1. CARDIOVASCULAR MRI REPORT\nDr Mickey Mouse MD (Res),\nDr John Doe MRCPCH DPhil MRCP FRACP \nDr John Smith MRCP, FRCR\nProf Katie Golden MRCPCH FRACP \nEnquiries: 121 121 121\nName: Alex Macdonald Date Of Birth: 21/06/1990\nGender: F MICKEY Hospital:2323232\nHosp No: 345 567 456 HH No. : 234351234 Date of Scan: 25/06/2013\nRequesting Consultant: Dr Adam Smith, Dr Lane Parrish\nDiagnosis … \nFindings

2. CARDIOVASCULAR MRI REPORT\nDr Mickey Mouse MD (Res),\nDr John Doe MRCPCH DPhil MRCP FRACP \nDr John Smith MRCP, FRCR\nProf Katie Golden MRCPCH FRACP nEnquiries: 121 121 121\nName: Alex Macdonald Date Of Birth: 21/06/1990 MICKEY Hospital:2323232\nHosp No: 345 567 456 HH No. : 234351234 Date of Scan: 25/06/2013\nRequesting Consultant: Dr Adam Smith and Dr Lane Parrish (MAP Hospital) \nDiagnosis

3. CARDIOVASCULAR MRI REPORT\nDr Mickey Mouse MD (Res),\nDr John Doe MRCPCH DPhil MRCP FRACP \nDr John Smith MRCP, FRCR\nProf Katie Golden MRCPCH FRACP nEnquiries: 121 121 121\nName: Alex Macdonald Date Of Birth: 21/06/1990\nGender: F MICKEY Hospital:2323232\nHosp No: 345 567 456 MAP Hosp No. : 234351234 Date of Scan: 25/06/2013\nRequested By: Dr Adam Smith, Dr Lane Parrish

Figure 1: Examples of dummy Cardiac MRI report snippets

Data Representation	BERT		RoBERTa-Large		Flan-T5-Large	
	F1	**Avg.CS**	**F1**	**Avg.CS**	**F1**	**Avg.CS**
Name	0.50	0.75	0.71	0.90	0.87	0.96
Date of Birth	1.0	1.0	1.0	1.0	1.0	1.0
Scan Date	1.0	1.0	1.0	1.0	1.0	1.0
Hospital Unique Identifier	0.97	0.98	0.97	0.98	0.99	0.99
Country Unique Identifier	0.90	0.92	0.94	0.94	0.94	0.94
Referrer Name	0.82	0.88	0.85	0.89	0.83	0.88

Table 1: Results on the held-out validation data containing 400 documents when no prompt is added to the context and using different models for the QA pipeline. Here *F1* represents the F1-score for each data representation and *Avg.CS* represents the average cosine similarity across the data based on an automatic comparison that compares the predicted data against the ground-truth data using the cosine similarity measure.

Data Representation	BERT		RoBERTa-Large		Flan-T5-Large	
	F1	**Avg.CS**	**F1**	**Avg.CS**	**F1**	**Avg.CS**
Name	0.67	0.81	0.81	0.97	0.99	1.0
Date of Birth	1.0	1.0	1.0	1.0	1.0	1.0
Scan Date	1.0	1.0	1.0	1.0	1.0	1.0
Hospital Unique Identifier	0.54	0.54	0.99	0.98	0.98	0.99
Country Unique Identifier	0.31	0.32	0.94	0.95	0.92	0.99
Referrer Name	0.82	0.88	0.88	0.88	0.98	0.99

Table 2: Results on the held-out validation data containing 400 documents when the best prompt is added to context and using different models for the QA pipeline. Here *F1* represents the F1-score for each data representation and *Avg.CS* represents the average cosine similarity across the data based on an automatic comparison that compares the predicted data against the ground-truth data using the cosine similarity measure.

Experimental Settings The purpose of using this pipeline is to extract major key patient identifiers namely *name of patient*, *date of birth*, *date of scan*, *hospital unique ID* and *country unique ID* respectively from our data. For the predefined question template **Q**, we use the phrase *"What is"* appended to the identifiers. For example, *name of patient* would be represented as *What is name of patient ?*. For the same example, let's assume we append $\mathbf{p}_1$ to the context. This would then be represented as below.

- Question: What is *name of patient* ?
- Context: $\mathbf{p}_1$ $\mathbf{d}_{i0}^{Inter}$

To understand whether these prompts can boost the performance of the pipeline, we investigated the same set of prompts across the different models $\mathcal{M}$. For our experiments, we carefully chose the following models due to our hospital infrastructure limitations: pretrained **BERT-base** (Devlin et al. 2018), **RoBERTa-Large** (Liu et al. 2019) models that were further fine-tuned on the SQUAD (Rajpurkar et al. 2016) dataset and, instruction-tuned text generation models **FLAN-T5-Small** and **FLAN-T5-Large** (Longpre et al. 2023) respectively.

Further, to understand the limitations of a rule-based approach, we chose the common entity keyphrases and obtained the results. Performance of a rule-based approach is compared with the best performing LLM as shown in Figure 2.

Experiments are conducted on a held-out validation data containing 400 documents. To automatically compare the extracted outcomes with the ground-truth data, we represent both the extracted outcomes and the ground-truth data as Bag-of-Words (BoW) vector representations. Each extracted outcome and corresponding ground-truth data is compared using the cosine similarity score between their vector representations. Here, a high threshold value of 0.95 is used such that the cosine similarity scores above the threshold were considered as correct.

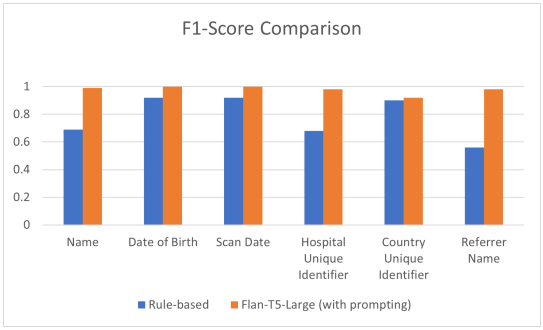


Figure 2: Results on the held-out validation data containing 400 documents with the best prompt when using **Flan-T5-Large** QA pipeline and the results obtained using a rule-based approach. Here F1-score for each data representation is compared.

Results and Evaluation

Different experiments were carried out using various prompts appended to the context. We present the results for predicting a specific set of patient identifiers for evaluating the performance when using the best prompt appended to the context vs without any prompt added to the context. The experiments were also carried out across a different set of models . In this experiment, we were more interested in understanding whether introducing prompts provide significant results rather than on engineering better prompts. The best prompt was *Following is an extracted unstructured text from a report of a patient in the XXXX hospital.* where XXXX is replaced by the hospital name. Results are provided in Table 1 and 2 respectively. For each data representation or key patient identifier, the F1-score and the average cosine similarity (as described in the section Experimental Settings) are reported across the different models used. Based on the results, we find that all the three models, with and without any prompting when provided to the context, can pick the *Date of Birth* and *Scan Date* correctly.

Our best prompt makes use of information provided about the *hospital* and the type of *document*. In general, the reports contain a lot of clinician names included which causes a lot of ambiguity in distinguishing between the patient name, clinician names and the referring clinician name. However, with the addition of the best prompt to the context, this shows a significant improvement across all the models when extracting names i.e. *Patient Name* and *Referrer Name*. However, **BERT-Base** model (Figure 3) did not provide any improvement with the addition of prompts to the context and adding more information worsened the performance of the model when extracting identifiers that included numbers (e.g., Hospital Unique Identifier).

Among the three models, **Flan-T5-Large** provides the best performance with or without prompts followed by **RoBERTa-Large**. This shows that instruction-tuned models such as **Flan-T5-Large** provide a better performance in extracting key entities when there is less information from their context.

Further, we also compared the results of **Flan-T5-Large** with **Flan-T5-Small** and it does show that scaling the parameters of the models significantly improves the performance as shown in prior work. In a zero-shot setting, **Flan-T5-Small** was unable to provide any significant extraction but was repeating the question template.

Error Analysis: An example We performed error analysis by investigating on a subset of the validation dataset to understand the performance of the individual models and the impact of introducing prompting templates. To illustrate this, we will consider the following example: "Question: What is *name of patient* ? Context: *Following is an extracted unstructured text from a report of a patient in the XXXX hospital. CMR Report Dr. John Brown Name: John Doe Master Mickey Hospital number: 1234567 Date of Birth: 01/01/1990 [SEP] Gender: M XXXX [SEP]*"

Our error analysis on the errors present in the outputs from the different models showed us that **BERT-Base** model was not able to distinguish *clinician names* from *patient names* even when the best prompt was appended to the context. On the other hand, **RoBERTa-Large** was able to extract partial answers with no prompt and additional texts along with the correct answers when prompting was provided. For the given example, Figure 3 shows the attention vectors across the different layers of the model and we found that introducing a prompt did not provide any significant attention to the key entities. For the same example, Figure 4, 5, 7 and 6 shows the word importance and attribution scores obtained using layered integrated gradients [3] for each token in the prediction output for both **BERT-Base** and **RoBERTa-Large** fine-tuned on SQUAD dataset respectively. The more positive the score, it represents the feature agreement with the model's prediction. While the prompting template did not have any direct feature agreement, it helped **RoBERTa-Large** model to capture the correct prediction whereas, the **BERT-base** failed to capture it.

Without any prompt, **Flan-T5-Large** appended additional texts to the correct answers for some documents similar to **RoBERTa-Large**. Figure 8 provides a target saliency heatmap with the attribute score computed using integrated gradients and the visualisation is produced using the interpretability toolkit Inseq (Sarti et al. 2023). Our overall analysis, results and experiments showed us that **Flan-T5-Large** was able to outperform the other models and provide correct answers. Figure. 9 provides a quantitative overview of the fraction of times where the models made an error.

Limitations

There are some limitations and constraints based on our infrastructure capabilities and information governance policies. Our work is carried out within a hospital trust in England which means using real-world data involves strict information governance compliances and subsequently on-premise infrastructure that meets these compliances.

Processing of patient-identifiable information is restricted to infrastructure that allows no access to the Internet and has been approved by the Information Governance team. This

[3] https://github.com/cdpierse/transformers-interpret

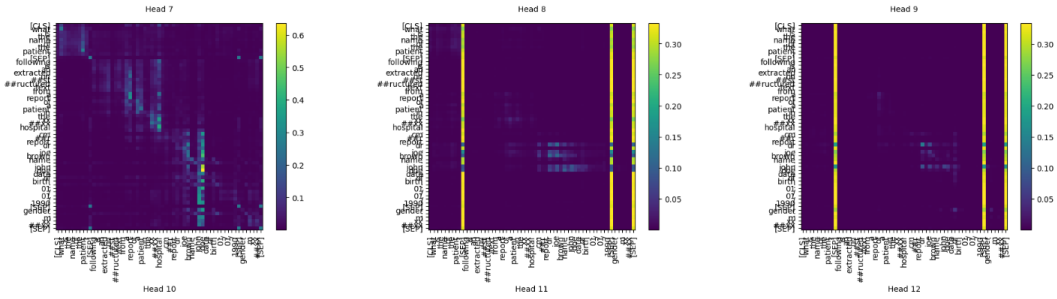


Figure 3: Visualisation of attention matrices between tokens present in a prompt, a context and the question template for the last layer of **BERT-Base** model (Vig 2019).

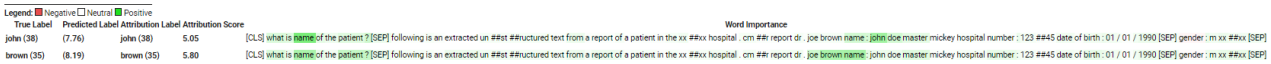


Figure 4: Word importance and attribution score using **BERT-Base** with prompt added to the context.

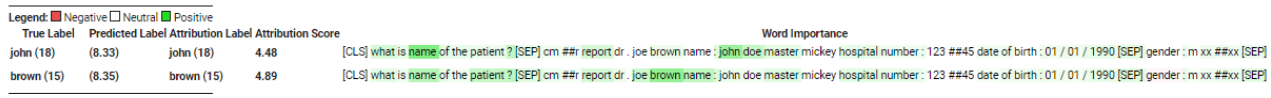


Figure 5: Word importance and attribution score using **BERT-Base** with no prompt added to the context.

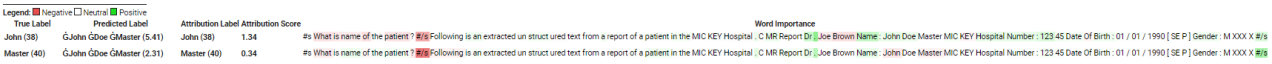


Figure 6: Word importance and attribution score using **RoBERTa-Large** with prompt added to the context.

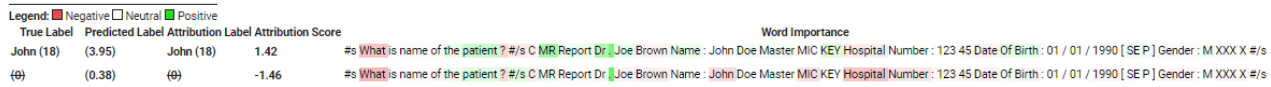


Figure 7: Word importance and attribution score using **RoBERTa-Large** with no prompt added to the context.

restricts us from using services like ChatGPT/ or other commercial external API-based systems. Our current infrastructure is a CPU based system with no GPUs and hence we explored using models which can give a reasonable performance during training on a CPU based system. This means that heavy resource-intensive models like Llama or Mistral have not been explored. Our decisions on choosing the models for evaluation are shaped by the constraints around infrastructure and information governance. We believe that this approach can be used as a pilot study in situations to show the potential capabilities of LLMs, and the results and evaluation using our existing infrastructure capabilities will be useful in other regulated domains like finance and in resource-constrained scenarios.

Deployment Details

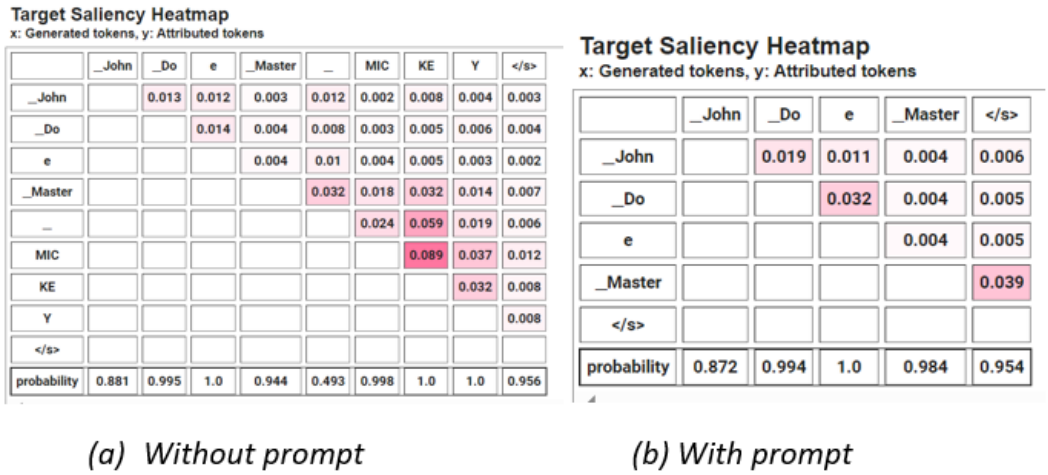
The pipeline was deployed within our secure on-premise environment. The source code was developed within one of our development servers and code pushed into our internal GitLab server. The source code is containerised and used within our Staging server, which has no access to internet and where we are allowed to process patient identifiable data. The infrastructure have 48 GB RAM,12 CPU cores and 500 GB storage with a linux based OS.

Conclusion

In this study, we explored the automatic extraction of patient information from unstructured data using a prompt-based question answering pipeline and utilising open-source LLMs. Our study demonstrates the potential capability of using existing open-source LLMs in a zero-shot setting for a real-world healthcare data with the use of effective prompt-based information in a question answering pipeline. The open-source instruction-tuned text generation model **Flan-T5-Large** (Longpre et al. 2023), which is significantly smaller than many current existing models and that is not fine-tuned or trained using our own data, performs significantly better than others.

Acknowledgements

We would like to thank Daniel Key and Alex Eze for their support with infrastructural development. This activity is part of a collaborative working agreement between Great Ormond Street Hospital NHS Foundation Trust and Roche Products Ltd. M-GB-00021188 (January 2025).



	_John	_Do	e	_Master	_	MIC	KE	Y	</s>
_John		0.013	0.012	0.003	0.012	0.002	0.008	0.004	0.003
_Do			0.014	0.004	0.008	0.003	0.005	0.006	0.004
e				0.004	0.01	0.004	0.005	0.003	0.002
_Master					0.032	0.018	0.032	0.014	0.007
_						0.024	0.059	0.019	0.006
MIC							0.089	0.037	0.012
KE								0.032	0.008
Y									0.008
</s>									
probability	0.881	0.995	1.0	0.944	0.493	0.998	1.0	1.0	0.956

	_John	_Do	e	_Master	</s>
_John		0.019	0.011	0.004	0.006
_Do			0.032	0.004	0.005
e				0.004	0.005
_Master					0.039
</s>					
probability	0.872	0.994	1.0	0.984	0.954

Figure 8: Target saliency heatmap with the attribute score computed using integrated gradients on the Flan-T5-Large output.

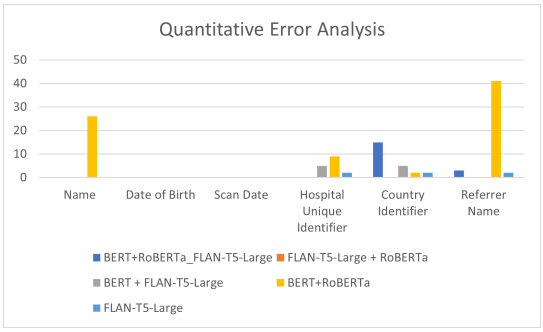


Figure 9: Quantitative error analysis based on the models that made errors while the models not included in their labels predicted the correct answer.

References

Appalaraju, S.; Jasani, B.; Kota, B. U.; Xie, Y.; and Manmatha, R. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 993–1003.

Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; and Jain, A. 2024. Structured information extraction from scientific text with large language models. *Nat. Commun.*, 15(1): 1–14.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hemingway, H.; Asselbergs, F. W.; Danesh, J.; Dobson, R.; Maniadakis, N.; Maggioni, A.; van Thiel, G. J. M.; Cronin, M.; Brobert, G.; Vardas, P.; Anker, S. D.; Grobbee, D. E.; Denaxas, S.; and Innovative Medicines Initiative 2nd programme, Big Data for Better Outcomes, BigData@Heart Consortium of 20 academic and industry partners including ESC. 2018. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur. Heart J.*, 39(16): 1481–1495.

Kanakarajan, K. R.; and Sankarasubbu, M. 2023. Saama AI Research at SemEval-2023 Task 7: Exploring the Capabilities of Flan-T5 for Multi-evidence Natural Language Inference in Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 995–1003.

Liu, A. T.; Xiao, W.; Zhu, H.; Zhang, D.; Li, S.-W.; and Arnold, A. 2022. Qaner: Prompting question answering models for few-shot named entity recognition. *arXiv preprint arXiv:2203.01543*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 22631–22648. PMLR.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Sarti, G.; Feldhus, N.; Sickert, L.; van der Wal, O.; Nissim, M.; and Bisazza, A. 2023. Inseq: An Interpretability Toolkit for Sequence Generation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 421–435. Toronto, Canada: Association for Computational Linguistics.

Sedlakova, J.; Daniore, P.; Horn Wintsch, A.; Wolf, M.; Stanikic, M.; Haag, C.; Sieber, C.; Schneider, G.; Staub, K.; Alois Ettlin, D.; Grübner, O.; Rinaldi, F.; von Wyl, V.; and University of Zurich Digital Society Initiative (UZH-DSI) Health Community. 2023. Challenges and best practices for digital unstructured data enrichment in health research:

A systematic narrative review. *PLOS Digit. Health*, 2(10): e0000347.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; Payne, P.; Seneviratne, M.; Gamble, P.; Kelly, C.; Babiker, A.; Schärli, N.; Chowdhery, A.; Mansfield, P.; Demner-Fushman, D.; Agüera y Arcas, B.; Webster, D.; Corrado, G. S.; Matias, Y.; Chou, K.; Gottweis, J.; Tomasev, N.; Liu, Y.; Rajkomar, A.; Barral, J.; Semturs, C.; Karthikesalingam, A.; and Natarajan, V. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Tirskikh, D.; and Konovalov, V. 2023. Zero-Shot NER via Extractive Question Answering. In *International Conference on Neuroinformatics*, 22–31. Springer.

Vig, J. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 37–42. Florence, Italy: Association for Computational Linguistics.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1192–1200.

Yang, Y.; Cao, J.; Wen, Y.; and Zhang, P. 2021. Table to text generation with accurate content copying. *Sci. Rep.*, 11(1): 22750.