Applied Computer Science, vol. 20, no. 4, pp. 138–148
doi: https://doi.org/10.35784/acs-2024-44

Submitted: 2024-08-30 | Revised: 2024-11-11 | Accepted: 2024-11-18

Keywords: medical information extraction, Large Language Models, ChatGPT, schema-based extraction

Zakaria KADDARI [0000-0003-4034-5612]*, *Ikram El HACHMI* [0009-0008-7928-3088]**, *Jamal BERRICH* [0000-0001-8443-7223]*, *Rim AMRANI* [0000-0003-3906-5533]**, *Toumi BOUCHENTOUF* [0000-0002-2689-8678]*

EVALUATING LARGE LANGUAGE MODELS FOR MEDICAL INFORMATION EXTRACTION: A COMPARATIVE STUDY OF ZERO-SHOT AND SCHEMA-BASED METHODS

Abstract

This study investigates the application of large language models, particularly ChatGPT, in the extraction and structuring of medical information from free-text patient reports. The authors explore two distinct methods: a zero-shot extraction approach and a schema-based extraction approach. The dataset, consisting of 1230 anonymized French medical reports from the Department of Neonatology of the Mohammed VI University Hospital, served as the basis for these experiments. The findings indicate that while ChatGPT demonstrates a significant capability in structuring medical data, certain challenges remain, particularly with complex and non-standardized text formats. The authors evaluate the model's performance using precision, recall, and F1 score metrics, providing a comprehensive assessment of its applicability in clinical settings.

1. INTRODUCTION

In recent years, the healthcare industry has seen a significant shift towards the digitalization of medical records, driven by the need for improved patient care, data management, and operational efficiency. With the proliferation of electronic health records (EHRs), there has been an exponential increase in the amount of unstructured textual data generated by healthcare providers. This unstructured data, which includes clinical notes, patient reports, and other free-text medical documentation, contains vital information that is crucial for patient diagnosis, treatment planning, and research. However, extracting structured, actionable insights from this vast amount of unstructured text remains a formidable challenge.

* Université Mohammed Premier, National School of Applied Sciences, LaRSA laboratory, AIRES team, Morocco, z.kaddari@ump.ac.ma, j.berrich@ump.ac.ma, t.bouchentouf@ump.ac.ma

** Université Mohammed Premier, Faculty of Medicine and Pharmacy Oujda, Morocco, ikram.elhachmi@ump.ac.ma, r.amrani@ump.ac.ma

Traditional natural language processing (NLP) (Kaddari et al., 2021) techniques have been employed to address this challenge, offering various methods to convert free-text data into structured formats. These methods typically rely on rule-based systems (RBS), statistical models, or a combination of both. While they have shown success in specific applications, these approaches often struggle with the complexity and variability of clinical language, which is characterized by domain-specific terminology, abbreviations, and diverse linguistic patterns (Zhan et al., 2021). Moreover, traditional NLP systems require extensive domain knowledge and manual rule crafting, making them less adaptable to new or evolving datasets.

The advent of large language models (LLMs) (Yifan et al., 2024), such as OpenAI's ChatGPT (Ray, 2023), has introduced a new paradigm in the field of NLP. These models, trained on vast amounts of text data, are capable of understanding and generating human-like language, making them well-suited for a wide range of text processing tasks. ChatGPT, in particular, has demonstrated impressive performance in tasks like text summarization, translation, and question answering. Its ability to perform zero-shot learning—where the model is applied to tasks without specific training on those tasks—presents a unique opportunity to leverage its capabilities in the medical domain, particularly for the extraction and structuring of clinical information.

Despite the potential of LLMs, their application in the healthcare sector is still in its infancy, and several challenges remain. The highly specialized and sensitive nature of medical data requires models that not only perform accurately but also maintain the highest standards of reliability and interpretability. Furthermore, the risk of hallucination (Huang et al., 2024) where the model generates incorrect or nonsensical information, poses a significant concern in clinical settings, where such errors can have serious consequences.

This study aims to explore the effectiveness of LLMs, particularly ChatGPT, in extracting and structuring information from free-text medical reports, specifically focusing on French-language neonatal patient reports. The authors propose two distinct methods for information extraction: a zero-shot extraction approach, where the model is tasked with extracting predefined attributes without prior domain-specific training, and a schema-based extraction approach, which involves defining a structured schema to guide the extraction process. By comparing the performance of these methods, the authors seek to evaluate the feasibility of using ChatGPT for clinical data structuring and identify areas where further refinement is needed.

Through this research, the authors contribute to the growing body of knowledge on the application of LLMs in the healthcare domain, offering insights into their strengths and limitations. The findings are intended to inform future efforts in the development of NLP tools for clinical applications, with the ultimate goal of improving patient care through more effective use of healthcare data.

2. RELATED WORKS

Due to the difficulty of obtaining medical notes, little research was done on automatic medical notes structuring. Still, varieties of techniques were explored, from rule-based systems (RBS) to Large Language Models, each presenting unique strengths and limitations in handling clinical data.

Rule-Based Approaches: Early methods, such as rule-based NLP systems, have demonstrated robustness in specific scenarios. For instance, Patra et al. (2024) utilized an RBS to extract social information from psychiatry notes, outperforming LLMs in accuracy across all metrics. These systems excel due to their deterministic nature, ensuring consistency in extraction when rules are clearly defined. However, their rigidity makes them less adaptable to diverse and evolving clinical contexts.

LLMs for Information Extraction: With the rise of LLMs like ChatGPT, new paradigms have emerged. LLMs leverage vast amounts of training data, allowing them to generalize across tasks with minimal domain-specific adjustments. In Huang et al. (2024), ChatGPT-3.5 demonstrated an 89% accuracy in extracting structured data from lung cancer notes, surpassing traditional NLP methods. This study highlighted the potential of zero-shot capabilities in clinical settings, showing that LLMs could achieve significant results without domain-specific training. Similarly, in (Kernberg et al., 2024), the authors used ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters. The research indicated significant differences in error rates, accuracy, and the quality of notes produced by ChatGPT-4. It was found that longer transcripts and more complex data negatively impacted note accuracy, highlighting potential issues with the model's capability in managing intricate medical cases.

This work aligns with this trend by employing both a zero-shot extraction method and a schema-based extraction approach. The zero-shot method is similar to Huang et al.'s, leveraging ChatGPT's ability to perform information extraction without specialized fine-tuning. However, the schema-based method introduces additional structure, which enhances extraction accuracy, particularly for complex attributes like "Mother's medical history."

Transformer-Based Models: Other research focused on smaller transformer models tailored to specific datasets. For instance, (Bergomi et al., 2024) used a compact transformer to structure radiology notes, achieving results comparable to GPT-3.5 despite being a thousand times smaller. Similarly, (Zelina et al., 2022) applied a fine-tuned RoBERTa model RobeCzech (Straka et al., 2021) to Czech clinical data, illustrating that domain-specific transformers can perform effectively with localized datasets.

The presented schema-based extraction method parallels these domain-specific approaches by providing structured guidance. Unlike smaller transformers, we employed ChatGPT models (3.5 and 4) with a schema that closely mirrors the attribute structure in patient reports. This structured guidance led to superior performance in precision and recall, especially with ChatGPT-4, which outperformed the zero-shot method in all metrics.

Zero-shot and Few-shot Learning: The adaptability of LLMs has also been tested in few-shot learning contexts. (Agrawal et al., 2022) demonstrated that LLMs like InstructGPT (Ouyang, 2022) excel in zero-shot and few-shot scenarios even without clinical domain training. Similarly, (Bhate et al., 2023) employed a minimal instruction setup to extract social determinants from clinical notes using a GPT-based model.

This research builds on these findings by showing that structured extraction (schema-based) can mitigate some of the limitations of zero-shot approaches, particularly in handling complex, multi-word fields. While our zero-shot method displayed competence in straightforward extractions, the schema-based approach provided a consistent boost in accuracy, highlighting the benefits of a structured methodology when dealing with nuanced clinical data.

3. DATA

The dataset used in this study consists of 1230 free-text patient medical reports from 2021 to 2023 written in French, supplied by the Department of Neonatology of the Mohammed VI University Hospital of Oujda, Morocco. The provided reports were completely anonymized. The formatting was not consistent, as multiple templates were used during the three-year period. In general, each report contained the following five sections: patient information, information about the mother, clinical exam, how to proceed with the patient, and patient stay evolution. Figure 1 lists the most important clinical information in each section. On average, a report contains 505 words, with the most complete report containing 1049 words.

Table 1 shows an example of the desired medical information to extract for two sample clinical reports.

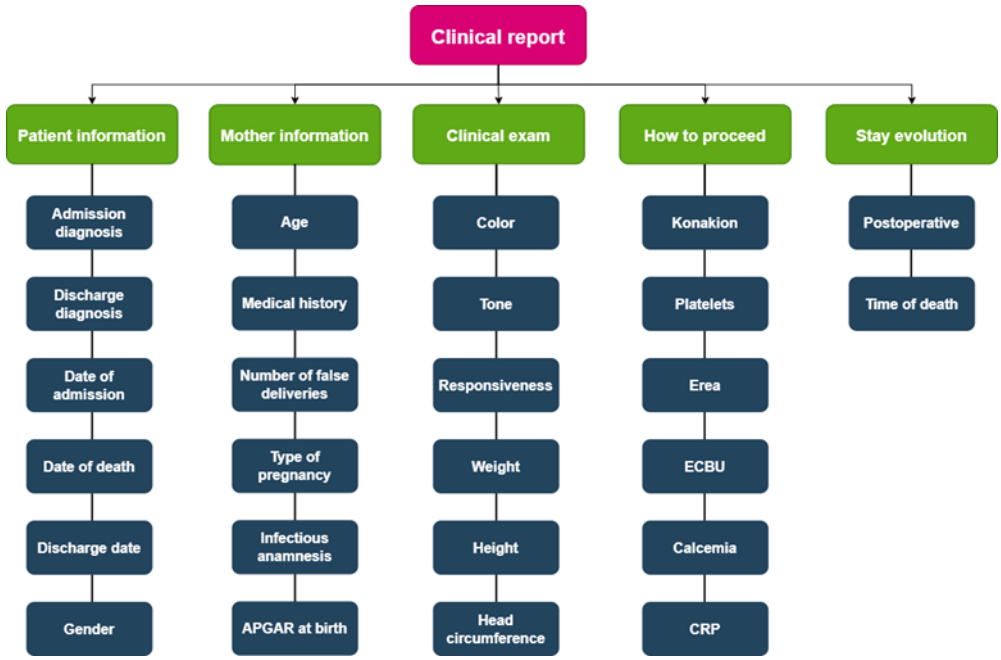

Fig. 1. Patient reports most important information

Tab. 1. Example of the desired medical information to extract for two sample clinical reports

Report number	Gender	Mother's age	Date of admission	Weight	PNN	Type of marriage	Height
1	Féminin	43	14/07/2023	3.6	7895	consanguin	50cm
2	masculin	19	13/04/2023	2.6	11256	Non consanguin	48cm

Report 1 : Date d'entrée: 14/07/2023 … de sexe Féminin … mère âgée de 43 ans … de mariage consanguin… Poids: 3.6kg Taille: 50cm … PNN 7895
Report 2 : Date d'entrée: 13/04/2023 … de sexe masculin … âgée de 19 ans … de mariage non consanguin … poids: 2kg 600 Taille 48cm PNN 11256

4. METHODS

Due to its superiority compared to the other available LLMs, the authors based their methods on ChatGPT. They used the model via the proposed Application Programming Interface (API) by OpenAI in a zero-shot setting. A ChatGPT API call requires users to provide instructions via two role variables.

- System: defines task instructions.
- User: provides an input text for zero-shot learning.

The specific models that were used are described in the Results section.

The authors used two different methods in their experiments. First, they prompted ChatGPT in a zero-shot setting by asking it to extract a list of predefined attributes from a given patient report. Table 2 shows the exact prompt that was used.

Tab. 2. Exact prompt used in the first method

Role	Content
System	Please extract the following attributes from the provided patient report: - Date of admission - Date of birth - Gender of newborn - Type of marriage - Admission diagnosis - Mother's medical history - Mother's age - PNN (polynucléaires neutrophiles) - Pregnancy follow-up - Pregnancy carried to term - Weight - Height
User	Patient report

For the second method, the authors defined a schema model for each report section. With each model containing the medical attributes to be extracted. A model is defined as a Python class, and medical attributes are defined as regular typed class attributes. To aid the model in the extraction, the authors used the same attribute names in the patient reports, and they provided a description for certain attributes to further aid the model in extracting the desired attributes with specific unit measures for example. Thay then instructed ChatGPT to extract the required information as objects of the defined classes.

In order to mitigate hallucination, the authors allowed the model for an additional retry in case it did not succeed in extracting a given attribute.

The results of the experiments using the two methods are presented in section V.

5. RESULTS

In order to experiment with these two methods, the authors created a test set from their dataset, first by randomly selecting 100 patient reports. Then, by manually extracting a

subset of the desired medical attributes. They made sure to include attributes from all relevant types (strings, dates, integers, floats, and booleans) in their experiments.

In the evaluation, the authors used a relaxed match metric, similar to the fragment match approach used in biomedical named entity recognition (Tsai et al., 2006). In relaxed match, all extracted concepts are broken down into individual words. Unlike in fragment match, where each token in a biomedical concept is considered separately. Relaxed match is more suited for clinical notes, and represents a middle ground between the more rigorous strict match, and more tolerant measures like fragment match. Once the individual words were extracted, the precision and value returned based on the words were calculated. Words present in both the annotated ground truth and the model output are counted as True Positives (TP). Words found in the model output but not in the annotated ground truth are counted as False Positives (FP), and words in the annotated ground truth but missing from the model output are counted as False Negatives (FN). Equations 1 to 3 detail the calculations for recall (R), precision (P), and F1 score.

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = \frac{2PR}{P+R} \tag{3}$$

The results of these experiments are shown in Table 3 and Table 4. The authors experimented with their two methods using the two models ChatGPT 3.5 and ChatGPT 4. Table 3 shows the overall results of their experiments with the two methods using the two ChatGPT models. Whereas Table 4 shows the detailed results of the experiments carried out with the scheme-based extraction method. This is the best performing of the two proposed methods.

Tab. 3. Overall results of our experiments using the two proposed methods

Method	**ChatGPT 3.5**			**ChatGPT 4**		
	P	**R**	**F1**	**P**	**R**	**F1**
zero-shot extraction	70.13	71.45	70.35	74.69	74.98	73.14
schema-based extraction	76.55	77.99	76.69	81.93	82.76	81.00

The authors wanted to study the effectiveness of their best-performing method (schema-based extraction) by attribute type. They experimented with all attribute types in patient reports, which are dates, single word strings, long sentences, integers, floats, and booleans. In order to confirm their findings, the authors experimented on two attributes by each attribute type. Table 4 shows the results of these experiments using the two ChatGPT models.

Tab. 4. Detailed results of the schema-based extraction method

Attribute Type	Attribute	ChatGPT 3.5			ChatGPT 4		
		P	R	F1	P	R	F1
Date	Date of admission	76.67	76.67	76.67	77.36	77.36	77.36
	Date of birth	83.33	83.33	83.33	85	85	85
Single word string	Gender of newborn	93.33	93.33	93.33	95.75	95.75	95.75
	Type of marriage	80.00	80.00	80.00	83.16	83.16	83.16
Long sentence	Admission diagnosis	77.78	78.33	78.00	78.67	78.67	78.67
	Mother's medical history	54.13	70.90	55.59	61.23	71.23	60.12
Integer	Mother's age	96.67	96.67	96.67	96.67	96.67	96.67
	PNN	56.67	56.67	56.67	59.33	59.33	59.33
Boolean	Pregnancy follow-up	90.00	90.00	90.00	93.33	93.33	93.33
	Pregnancy carried to term	53.33	53.33	53.33	86.67	86.67	86.67
Float	Weight	70.00	70.00	70.00	77.00	77.00	77.00
	Height	86.67	86.67	86.67	89.00	89.00	89.00
Average		76.55	77.99	76.69	81.93	82.76	81.00

In Figure 2, the performance of the schema-based extraction method by number of tries is presented. Allowing for an additional try does mitigate the effect of hallucination, yielding overall performance gains.

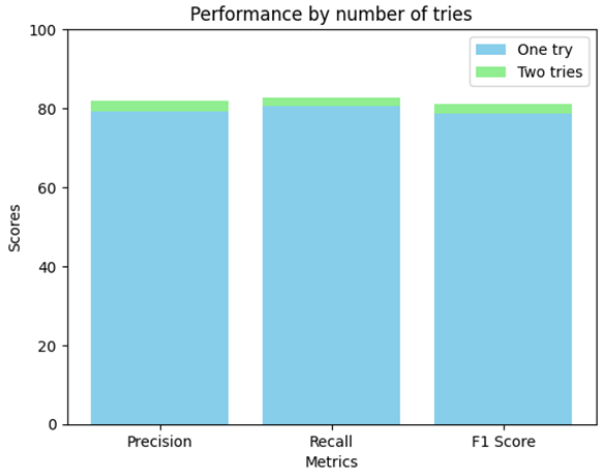

Fig. 2. Performance of the schema-based extraction method by number of tries

The next section discusses the results, first comparing the performance of the two proposed methods. Then, the performance of the most efficient method is investigated for different types of attributes, and finally, comparing the performance of the ChatGPT model.

6. DISCUSSION

The results of this study indicate a clear difference in the performance of the two methods employed—zero-shot extraction and schema-based extraction—across various attribute types and ChatGPT models.

6.1. Comparison of methods

The schema-based extraction method outperformed zero-shot extraction in all performance metrics across both ChatGPT-3.5 and ChatGPT-4 models. Specifically, for ChatGPT-3.5, the schema-based method achieved an F1 score of 76.69 compared to 70.35 for the zero-shot method. This trend was consistent with ChatGPT-4, where the schema-based approach recorded an F1 score of 81.00, significantly higher than the 73.14 observed with zero-shot extraction. These findings underscore the importance of providing structured guidance to the model in extracting specific medical attributes, which appears to enhance the accuracy and consistency of the results.

6.2. Comparison by attribute type

When analyzing the results by attribute type, it is evident that both methods struggled more with certain field types, particularly complex or multi-word fields such as "Mother's medical history". For instance, the F1 score for "Mother's medical history" was notably lower compared to simpler fields like "Gender of newborn" and "Mother's age". This suggests that while LLMs like ChatGPT are adept at handling straightforward, single-word extractions, they encounter challenges with longer and more nuanced text segments.

Interestingly, the schema-based method provided more reliable results for complex fields compared to zero-shot extraction, particularly with ChatGPT-4. For example, the F1 score for "Mother's medical history" improved from 55.59 (ChatGPT-3.5) to 60.12 (ChatGPT-4) in the schema-based approach, demonstrating the method's relative effectiveness in managing complex attributes when provided with explicit structure and schema.

Boolean fields such as "Pregnancy follow-up" and "Pregnancy carried to term" showed varying results. While "Pregnancy follow-up" achieved high F1 scores (90.00 for ChatGPT-3.5 and 93.33 for ChatGPT-4), "Pregnancy carried to term" had lower scores, particularly with ChatGPT-3.5 (53.33). This discrepancy suggests that the model's performance on boolean fields may depend on the specific context and phrasing of the attribute.

Numeric fields like "Weight" and "Height" showed good performance, with F1 scores ranging from 70.00 to 89.00. This indicates that both methods are reasonably effective at extracting quantitative data from medical reports.

6.3. Comparison by ChatGPT model

The comparison between ChatGPT-3.5 and ChatGPT-4 further highlights the advancements made in the latter model. ChatGPT-4 consistently outperformed ChatGPT-3.5 across all metrics and methods, with a noticeable increase in precision, recall, and F1 scores. For example, the F1 score for zero-shot extraction improved from 70.35 (ChatGPT-3.5) to 73.14 (ChatGPT-4), and for schema-based extraction, it increased from 76.69 (ChatGPT-3.5) to 81.00 (ChatGPT-4). This improvement is particularly pronounced in the

schema-based method, indicating that ChatGPT-4 is better equipped to leverage structured guidance for extracting medical information, further validating the evolution in LLM capabilities.

The performance gap between the two models was most evident in complex fields. For instance, in the schema-based extraction of "Mother's medical history", ChatGPT-4 achieved an F1 score of 60.12, compared to 55.59 for ChatGPT-3.5. This suggests that the advanced capabilities of ChatGPT-4 are particularly beneficial when dealing with more challenging, context-dependent information extraction tasks.

It's worth noting that even for fields where ChatGPT-3.5 performed well, such as "Gender of newborn" (F1 score of 93.33), ChatGPT-4 still managed to show improvement (F1 score of 95.75). This consistent enhancement across various field types underscores the overall superiority of the newer model in medical information extraction tasks.

6.4. Common extraction errors

The analysis of the extraction methods revealed several common errors, each of which carries specific implications for clinical applications. These errors were observed across both zero-shot and schema-based approaches, highlighting areas where current methods can be improved.

False positives in extraction: One common issue involved the over-identification of attributes, leading to false positives. This error was particularly evident in fields with overlapping terminology, such as "Mother's medical history" and "Admission diagnosis," where context was crucial for accurate interpretation. In clinical scenarios, false positives can generate misleading information, potentially affecting clinical decisions and leading to unnecessary follow-up actions.

Incomplete data due to false negatives: False negatives, where relevant attributes were missed by the extraction model, were also a notable source of error. Complex, multi-word entities, particularly those involving numerical or categorical values like "Pregnancy carried to term," were frequently affected. Missing critical data can compromise the integrity of patient records, leading to gaps in medical history that could affect future treatment plans.

Misinterpretation of contextual indicators: Both extraction methods occasionally failed to interpret negations or contextual cues accurately. This was particularly problematic with phrases involving temporal elements (e.g., "recent history of illness") or negations ("no previous complications"). Errors in contextual understanding could lead to incorrect patient profiles, potentially influencing clinical risk assessments and eligibility for specific treatments.

These extraction errors have a direct impact on the reliability of structured data in clinical settings. For instance, a false negative in the extraction of a family history related to genetic conditions could lead to underestimating patient risks, while false positives may result in unnecessary diagnostic procedures. Additionally, incomplete or misinterpreted data complicates clinical documentation, affecting the quality of Electronic Health Records (EHRs) and their use in decision-making systems.

To address these issues, future work will focus on refining the schema-based extraction methodology, emphasizing improved context recognition and handling of ambiguous terminology. Adjustments to the schema structure and the incorporation of domain-specific lexicons could enhance precision. Additionally, developing post-processing validation steps

that involve clinician feedback may help filter out inaccuracies, ensuring that the structured data aligns closely with clinical needs.

6.5. General observations

Overall, the schema-based extraction method, particularly when paired with the more advanced ChatGPT-4, shows considerable promise for medical information extraction tasks. The structured nature of this approach mitigates some of the challenges associated with complex and non-standardized medical texts, which were more problematic for the zero-shot extraction method. Furthermore, the superior performance of ChatGPT-4 across all tasks suggests that future work in this area would benefit from utilizing the most recent iterations of LLMs, coupled with robust schema-based methodologies, to enhance the accuracy and reliability of medical data extraction.

However, it's important to note that even with the best-performing method (schema-based extraction with ChatGPT-4), there is still room for improvement, particularly in handling complex, multi-word fields and ensuring consistent performance across all types of medical attributes. The variability in performance across different field types highlights the need for specialized approaches that can adapt to the diverse nature of medical information.

These findings contribute valuable insights into the application of LLMs in clinical settings, highlighting both their potential and the areas where additional refinement is necessary. The results underscore the importance of continuous model development and the need for domain-specific fine-tuning to address the unique challenges presented by medical data extraction.

7. CONCLUSIONS

This research demonstrates the potential of large language models like ChatGPT in extracting and structuring medical information from unstructured clinical texts. While the model shows considerable promise, particularly in its zero-shot capabilities, it also highlights the need for more specialized approaches to handle the unique challenges presented by medical data. Future work should focus on enhancing the model's ability to manage complex and non-standardized text formats, as well as reducing the incidence of hallucinations. The continued development and application of LLMs in this field could ultimately contribute to more efficient and accurate clinical data management, improving outcomes for healthcare providers and patients alike.

Code and data availability

The code is available from the corresponding author upon request. As for the data, due to privacy, ethical and legal concerns, we cannot share our dataset, but can provide more details to interested parties on request.

REFERENCES

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large Language Models are few-shot clinical information extractors. *ArXiv, abs/2205.12689*. https://doi.org/10.48550/arXiv.2205.12689

Bergomi, L., Tommaso, M., Antonazzo, P., Alberghi, L., Bellazzi, R., Preda, L., Bortolotto, C., & Parimbelli, E. (2024). Reshaping free-text radiology notes into structured reports with generative question answering transformers. *Artificial Intelligence in Medicine, 154*, 102924. https://doi.org/10.1016/j.artmed.2024.102924

Bhate, N., Mittal, A., He, Z., & Luo, X. (2023). Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT Model. *IEEE International Conference on Big Data (BigData)* (pp. 1476-1480). IEEE. https://doi.org/10.1109/BigData59044.2023.10386811

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L. J., Xiao, G., Peterson, E. D., Zhan, X., & Xie, Y. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *Npj Digital Medicine*, *7*(1), 106. https://doi.org/10.1038/s41746-024-01079-8

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on hallucination in Large Language Models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 3703155. https://doi.org/10.1145/3703155

Kaddari, Z., Mellah, Y., Berrich, J., Belkasmi, M. G., & Bouchentouf, T. (2021). Natural language processing: challenges and future directions. In T. Masrour, I. El Hassani, & A. Cherrafi (Eds.), *Artificial Intelligence and Industrial Applications* (Vol. 144, pp. 236–246). Springer International Publishing. https://doi.org/10.1007/978-3-030-53970-2_22

Kernberg, A., Gold, J., & Mohan, V. (2024). Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: Comparative study. *Journal of Medical Internet Research*, *26*, e54419. https://doi.org/10.2196/54419

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *ArXiv, abs/2203.02155*. https://doi.org/10.48550/arXiv.2203.02155

Patra, B. G., Lepow, L. A., Kasi Reddy Jagadeesh Kumar, P., Vekaria, V., Sharma, M. M., Adekkanattu, P., Fennessy, B., Hynes, G., Landi, I., Sanchez-Ruiz, J. A., Ryu, E., Biernacka, J. M., Nadkarni, G. N., Talati, A., Weissman, M., Olfson, M., Mann, J. J., Zhang, Y., Charney, A. W., & Pathak, J. (2024). Extracting social support and social isolation information from clinical psychiatry notes: Comparing a rule-based natural language processing system and a large language model. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocae260

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Straka, M., Náplava, J., Straková, J., & Samuel, D. (2021). RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. In K. Ekštein, F. Pártl, & M. Konopík (Eds.), *Text, Speech, and Dialogue* (Vol. 12848, pp. 197-209). Springer International Publishing. https://doi.org/10.1007/978-3-030-83527-9_17

Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., Sung, T.-Y., & Hsu, W.-L. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, *7*, 92. https://doi.org/10.1186/1471-2105-7-92

Yifan, Y., Jinhao, D., Kaidi, X., Yuanfang, C., Zhibo, S., & Yue, Z. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Zelina, P., Halamkova, J., & Novacek, V. (2022). Unsupervised extraction, labelling and clustering of segments from clinical notes. *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1362-1368). IEEE. http://dx.doi.org/10.1109/BIBM55620.2022.9995229

Zhan, X., Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2021). Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns*, *2*(7), 100289. https://doi.org/10.1016/j.patter.2021.100289