Key Verbatim Extraction from Clinical Notes: A Hierarchical Multimodal Cross-Attention Approach

Anonymous ACL submission

Abstract

Clinical notes are essential for physicians to accurately assess patient conditions, particularly in oncology where records are extensive. Efficient and effective information extraction from these notes is crucial for effective treatment. This is not a trivial task due to the lengthy and specialized content in the notes. Current methods that capture token-level or sentence-level relations, which are context-dependent, are sometimes insufficient for knowledge-intensive tasks such as information extraction from EHR that require external knowledge. To address this, we introduce a knowledge-enhanced hierarchical multimodal cross-attention approach. This method employs a cross-attention mechanism to integrate textual knowledge with patient network knowledge, aiming to synthesize information across multiple data levels, including word, sentence, note, and patient levels. This approach can efficiently highlight key sentences in clinical notes. We validate our method using extensive experiments on a large real-world dataset. The results demonstrate that our proposed model outperforms baseline models by up to 4.17% and 2.79% regarding F1 and accuracy.

1 Introduction

Electronic Health Records (EHRs) play a crucial role in enabling physicians to assess a patient's condition precisely (Weed et al., 1968). However, in cases of severe illness, these records, along with associated textual materials, often become extensive and complex. This complexity poses a challenge for healthcare professionals to quickly extract essential information. Although the primary purpose of EHRs is to manage patients' health-related information, they are increasingly used for secondary purposes, such as addressing the above-mentioned challenges and improving healthcare practices (Sarwar et al., 2022). EHRs contain diverse data, including demographics, medical history, medications, lab results, and diagnoses, making them valuable for data mining and analytics (Yadav et al., 2018). These techniques have been used to study groups of patients, identify characteristics, provide personalized treatments, evaluate medical interventions, predict diseases, detect health conditions, and track disease progression (Yadav et al., 2018; Luque et al., 2019; Zeng et al., 2018; Karimi et al., 2015; Stiglic et al., 2020).

Summarising key information in EHRs holds substantial clinical significance, as it has the potential to expedite departmental workflows, diminish redundant human labor, and enhance clinical communication (Jin et al., 2024; Kahn Jr et al., 2009). Key verbatim is the exact, specific words, phrases or sentences extracted from the longer text (Siddiqi and Sharan, 2015). It is important in understanding and representing the longer text. This becomes more profound in oncology clinics, where patient records can span hundreds of pages due to frequent visits. Therefore, the efficient understanding of clinical notes and extraction of key verbatim from these EHRs are paramount for delivering timely and effective treatment.

With the advancement of natural language understanding techniques, language models like bidirectional encoder representations from transformers (i.e., BERT; (Devlin et al., 2018)) have been increasingly applied to tasks such as text extraction and classification. However, clinical notes present unique challenges due to their length and the specialized context, often containing terminology not found in standard datasets used for pretraining these models. While specialized algorithms like ClinicalBERT (Huang et al., 2019) have been developed to improve the accuracy of processing healthcare texts, and models adapted for longer texts are available, gaps remain in leveraging the potential useful information among sentences and across different notes in EHRs mining domain. Besides aiming to maximize the usefulness of textual information, approaches like DeepNote-GNN (Golmaei and Luo, 2021) are developed to extract relationships in EHRs. However, the potential for the fusion of different techniques still exists. Will exploring the relationships among different health-related entities (e.g., patients, drugs, physicians, treatments) help in understanding textual EHRs? We developed a novel multimodal approach to optimize the solution.

Overall, this study makes several contributions to the field of EHR analysis. First, we propose a hierarchical multimodal cross-attention approach for identifying key sentences associated with critical information in clinical notes. We employ ClinicalBERT for the textual representation of sentences, while capturing word-level details. Second, we leverage the external knowledge and textual notes to build a heterogeneous network and leverage Graph Attention Transformers (GATs) to learn implicit relations among patients and drugs, such as having shared illnesses or using similar treatments from the same physician. Third, we design a cross-attention mechanism that can bridge the intrinsic connection between information learned from text and the knowledge embedded in the patient network. We adopt Bi-LSTM to represent the combined textual and network knowledge at the sentence level within the context of individual notes. By aggregating information across word, sentence, note, and patient levels into the binary classification framework, our model is able to incorporate all relevant information in one unified framework. Finally, we empirically demonstrate that our approach facilitates the efficient extraction, highlighting key sentences in clinical notes. Our model is trained and evaluated using a dataset from an oncology clinic, where key sentences essential for diagnosis and other critical information have been labeled and verified by professional oncological physicians. The overall framework is shown in Figure 1.

2 Related Work

2.1 Extractive Summarization

Current text summarization methods originated from extractive algorithms. Following the initial use of rule-based extraction (Tas and Kiyani, 2007), deep learning language models have demonstrated superior performance, exemplified by fine-tuned BERT models (Liu, 2019). With the emergence of language generation models, such as Llama and ChatGPT (Touvron et al., 2023; OpenAI, 2024), abstractive summarization has developed and adapted to meet the varying requirements of different tasks (Mehta, 2016). Although abstractive summarization can outperform extractive methods in certain areas, such as statistical machine translation, extractive techniques remain crucial, in contexts where recognizing key information in lengthy texts is necessary and maintaining the originality of the output information is essential (Shi et al., 2021; Cho et al., 2014; Villanueva Jr and Simske, 2023; Mutlu et al., 2020). Compared to abstraction summarization, the key point of extraction summarization is to find the important paragraph or sentence in the texts (Moratanch and Chitrakala, 2017).

Earlier text extraction approaches include rule-based, statistical, machine learning approaches and domain-specific techniques (Siddiqi and Sharan, 2015; Moratanch and Chitrakala, 2017). Yang et al. (2022) highlighting the advancements and efficacy of deep learning in automatically understanding and processing large volumes of information. Jin et al. (2024) reviews Automatic Text Summarization(ATS) techniques, emphasizing practical implementations and the impact of Large Language Models (LLMs). Although LLM-based ATS achieves better performance in terms of consistency and relevance than human summarization and can handle tasks across a wide range of domains, which is superior to task-specific deep learning methods (Zhao et al., 2023; Tang et al., 2023; Basyal and Sanghvi, 2023; Zhang et al., 2024), the issues of prompt sensitivity and high resource requirements still dominate in real-world applications (Narayan et al., 2021; Liu et al., 2023).

2.2 Electronic Health Record Mining

Extraction summarization algorithms have found applications across diverse domains such as news, academia, law, and business (Venkatachalam et al., 2020; Mutlu et al., 2020; Jackson et al., 2003; Kitamori et al., 2017), where they enhance efficiency by condensing extensive texts into digestible summaries. This is particularly evident in the healthcare sector (Gao et al., 2017; Malmasi et al., 2017; Jackson et al., 2017; Wenzina and Kaiser, 2013). By distilling critical information from vast amounts of data, these algorithms support healthcare professionals in making informed decisions efficiently.

Given the volume and complexity of medical records, Wang et al. (2018) summarised clinical information extraction applications focusing on extracting key information from clinical texts. Commonly used health-related key information extraction tools, such as cTAKES (Savova et al., 2010), MetaMap (Aronson, 2001), and MedLEE (Friedman et al., 1994), are designed to extract information from unstructured, narrative, and redundant text data in EHRs. However, these tools are considered outdated due to their reliance on rule-based or heuristic methods, especially in light of the advancements in deep learning and LLMs.

Han et al. (2022) demonstrated that deep learning models, including CNN (LeCun et al., 1998), LSTM (Hochreiter and Schmidhuber, 1997), and BERT (Devlin et al., 2018), significantly outperformed traditional cTAKES in predicting social determinants of health from clinical notes. Similarly, Sarrouti et al. (2022) found that the fine-tuned encoder-decoder model T5 (Raffel et al., 2020) surpassed baseline models in biomedical text information extraction. Additionally, generative models such as BART (Lewis et al., 2019) also have been adopted in EHR mining. However, LLMs are not without limitations, including issues of inconsistency, lack of domain-specific knowledge, biases, hallucinations, high resource intensity, and limited handling of long documents (Reese et al., 2023; Kasneci et al., 2023; Chang et al., 2024).

2.3 Graph Neural Network in NLP

Graph Neural Networks (GNNs) have been extensively developed for graph data analysis, with popular models including GCN (Kipf and Welling, 2016), GraphSage (Hamilton et al., 2017), and GAT (Velickovic et al., 2017), among others. Recent research has witnessed a surge in interest in applying and developing various GNN variants for many NLP tasks, such as sentence classification(Huang and Carley, 2019; Lu et al., 2020), relation extraction (Qu et al., 2020; Sahu et al., 2019), and summarization(Fernandes et al., 2018; Yasunaga et al., 2017). In these studies, GNNs often serve as a rear-mounted module(Yang et al., 2021), further aggregating textual features modeled by pre-trained LLMs.

Another line of research employs GNNs as encoders of graph data for tasks such as retrieval augmentation (Abaho and Alfaifi, 2023), reasoning (Perozzi et al., 2024), and classification (Ostendorff et al., 2019; Chen et al., 2024). Despite these advancements, the potential of GNNs as knowledge enhancers for LLMs in extractive summarization remains under-explored. To address this gap, we propose a novel methodology that leverages GNNs for enhancing LLM-based extractive summarization.

3 Methodology

Our proposed model is shown in Figure 1. First, we pre-train a graph encoder to derive low-dimensional patient representations. Subsequently, the hierarchical sentence embeddings, concatenated with the updated sentence embeddings, are propagated through a classification layer for inference. Finally, a cross-attention module is applied, as a fusion layer, to update the sentence embeddings obtained from the hierarchical language model with the patient representations. This architecture incorporates multi-modality from both the language model, capturing word-to-word and sentence-to-sentence relations, and the graph model, capturing the prior knowledge of patients. And this prior-knowledge-enhanced architecture thereby can facilitate a more precise extraction of key verbatim.

3.1 Graph Construction

Consider an undirected heterogeneous graph $G = (V, E)$, where V represents the set of nodes, and E represents the set of edges. In this healthcare context, our proposed graph consists of three types ($T = \{p, o, m\}$) of nodes: patient Nodes (V_p), oncology nodes (V_o), and medication nodes (V_m). Nodes $v_p \in V_p$, $v_o \in V_o$ and $v_m \in V_m$ represent a patient, a specific oncology diagnosis, and a specific medication prescribed to patients correspondingly.

The edges in the graph represent relationships between these nodes and are of two types: patient-oncology edges (E_{po}) and patient-medication edges (E_{pm}). An edge $e_{po} = (v_p, v_o) \in E_{po}$ indicates that patient v_p has been diagnosed with oncology condition v_o, and an edge $e_{pm} = (v_p, v_m) \in E_{pm}$ indicates that patient v_p has been prescribed medication v_m.

Formally the graph can be represented as:

$$G = (V_p \cup V_o \cup V_m, E_{po} \cup E_{pm})$$

3.2 Graph Encoder

We adopt a heterogeneous GAT to derive meaningful embeddings for the patient nodes that capture the complex relationships within the heterogeneous healthcare graph data.

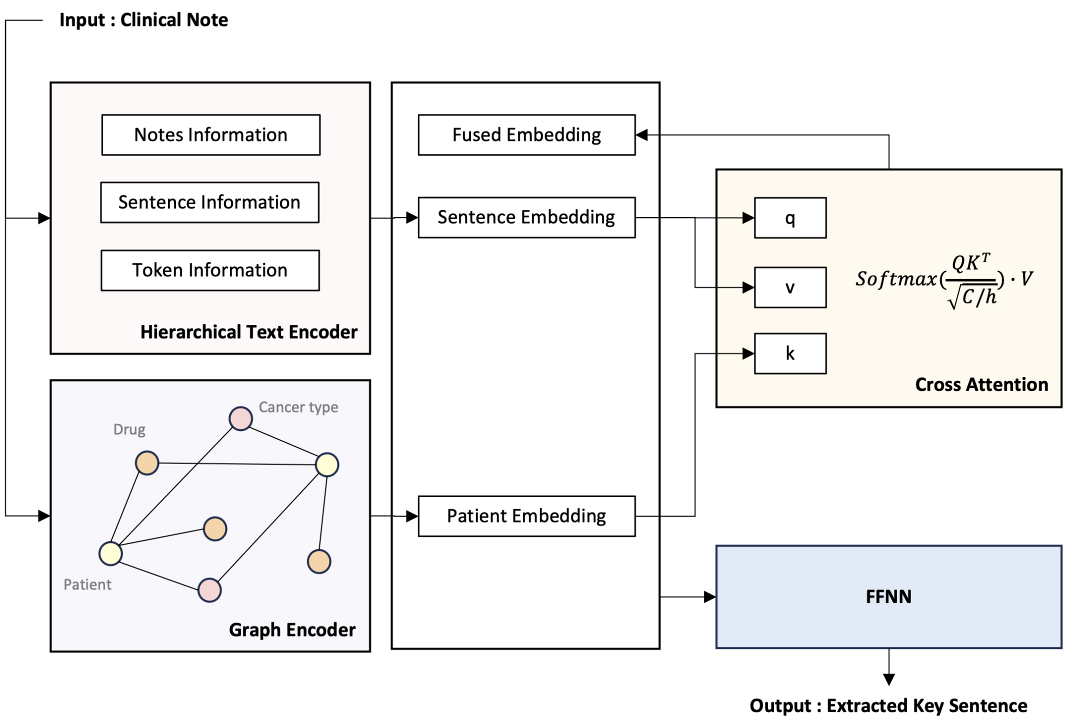

Figure 1: The overall framework of our proposed method.

Each node $v_i^s \in V_s$ is associated with a feature vector $\mathbf{h}_i^{s,l}$ at layer l and $s \in \{p, o, m\}$. We compute the attention coefficients α_{ij}^l that quantify the importance of node features of node $v_j^{q,l}$ ($q \in \{p, o, m\}$) to node v_i^s:

$$\alpha_{ij}^{l} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}_{sq}^{T}[\mathbf{W}^{s}\mathbf{h}_i^{s,l}\|\mathbf{W}^{q}\mathbf{h}_j^{q,l}]\right)\right)}{\sum_{t\in T}\sum_{k\in\mathcal{N}(i)\in V_t}\exp\left(\text{LeakyReLU}\left(\mathbf{a}_{st}^{T}[\mathbf{W}^{s}\mathbf{h}_i^{s,l}\|\mathbf{W}^{t}\mathbf{h}_k^{t,l}]\right)\right)}$$

$\mathbf{W}^s, \mathbf{W}^q, \mathbf{W}^t$ is node-type-specific learnable weight matrix and $t \in \{p, o, m\}$, $\mathbf{a}_{sq}$ and $\mathbf{a}_{st}$ is a learnable attention vector, $\|$ denotes concatenation, and $\mathcal{N}(i)$ denotes the neighborhood of node v_i^s. The embedding of node v_i^s is updated by aggregating the features of its neighbors of different types, weighted by the attention coefficients:

$$\mathbf{h}_i^{s,l+1} = \Big\|_{k=1}^{K} \sigma\left(\sum_{t\in T}\sum_{j\in\mathcal{N}(i)\in V_t}\alpha_{ij}^{l}\mathbf{W}^t\mathbf{h}_j^{t,l}\right)$$

$\mathbf{h}_j^{t,l}$ is the feature vector of node v_j^t at layer l and σ is a non-linear activation function, ReLU. In order to stabilize the learning process, we use K independent attention heads and their outputs are concatenated.

The training objective is to optimize the embeddings for the link prediction task. Specifically, we aim to predict the existence of edges between nodes in the graph. For this purpose, we employ a binary cross-entropy loss function over the observed and non-observed edges:

$$\mathcal{L} = -\sum_{(u,v)\in E}\log\sigma(\mathbf{h}_u^T\mathbf{h}_v) - \sum_{(u,v)\notin E}\log(1-\sigma(\mathbf{h}_u^T\mathbf{h}_v))$$

where σ is the sigmoid function, and (u, v) represents a node pair, with $(u, v) \in E$ indicating an existing edge and $(u, v) \notin E$ indicating a non-existent edge.

We pre-train a GAT on the constructed heterogeneous healthcare graph and obtain patient embeddings, which will be used in the following steps for better extract key informaiton from the healthcare documents, that capture the intricate relationships between patients, their oncological diagnoses, and prescribed medications.

3.3 Hierarchical Sentence Encoder

As shown in Figure 2, we adopt ClinicalBERT to obtain the textual representation of each sentence. ClinicalBERT, a transformer-based model pre-trained on clinical text, is capable of capturing token-level details effectively (Huang et al., 2019). For each sentence S consisting of n tokens $[t_1, t_2, \ldots, t_n]$, the initial token embeddings $\mathbf{E}_t$ are

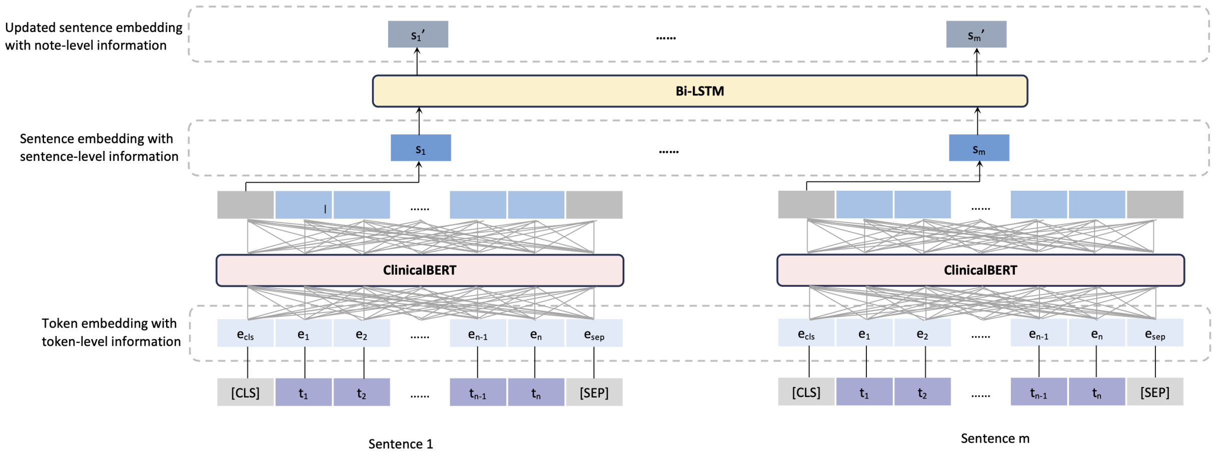

Figure 2: The framework of hierarchical sentence encoder.

passed through ClinicalBERT to generate contextualized embeddings.

Let $\mathbf{E}_t = [\mathbf{e}_{cls}, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$ be the initial embeddings of the tokens in sentence S. These embeddings are processed by ClinicalBERT to produce updated embeddings $\mathbf{H}_t = [\mathbf{h}_{cls}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n]$ The embedding of the initial token [CLS], $\mathbf{h}_{\text{cls}}$, represents the entire sentence embedding.

$$\mathbf{H} = \text{ClinicalBERT}(\mathbf{E}_t) \qquad \mathbf{s}_{BERT} = \mathbf{h}_{\text{cls}}$$

where $\mathbf{s}_{BERT}$ denotes the sentence embedding derived from the [CLS] token.

ClinicalBERT only captures contextual information within a sentence. We adopt BiLSTM to capture inter-sentence information.To integrate note-level (between-sentence) context in the sentence embedding, we employ a Bidirectional Long Short-Term Memory (Bi-LSTM) layer. This layer processes the sequence of sentence embeddings obtained from ClinicalBERT, capturing dependencies and contextual information at the note level. Let $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ be the sequence of sentence embeddings for a clinical note containing m sentences. These embeddings are input into the Bi-LSTM layer to obtain updated sentence embeddings. The Bi-LSTM processes the sequence as follows:

$$\begin{aligned}
\overrightarrow{\mathbf{h}}_i &= \overrightarrow{\text{LSTM}}(\mathbf{s}_i, \overrightarrow{\mathbf{h}}_{i-1}) \\
\overleftarrow{\mathbf{h}}_i &= \overleftarrow{\text{LSTM}}(\mathbf{s}_i, \overleftarrow{\mathbf{h}}_{i+1}) \\
\mathbf{s}'_i &= [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]
\end{aligned}$$

where $\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_i}$ are the forward and backward hidden states of the Bi-LSTM at position i, and $\mathbf{h}_i$ is the concatenated hidden state representing the updated sentence embedding. This process yields the note-level-context-updated sentence embeddings $\{\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_m\}$

3.4 Cross-Attention for Multi-modality Fusion

To include the patient-level information from GAT and combine the sentence embeddings with patient embeddings, we employ a cross-attention mechanism. This mechanism allows the model to attend to relevant parts of both embeddings, resulting in a fused representation.

Let $\mathbf{P}$ be the patient embedding obtained from the GAT, and $\mathbf{s}_{BERT}$ be the sentence embedding obtained from ClinicalBERT. The cross-attention mechanism is formulated as follows:

$$\begin{aligned}
\mathbf{Q} &= \mathbf{W}_Q \mathbf{p} \\
\mathbf{K} &= \mathbf{W}_K \mathbf{s}_{\text{BERT}} \\
\mathbf{V} &= \mathbf{W}_V \mathbf{s}_{\text{BBRT}}
\end{aligned}$$

$\mathbf{W}_Q$,$\mathbf{W}_K$ and $\mathbf{W}_V$ are learned weight matrices that transform the patient embedding and the sentence embedding into the query, key, and value matrices, respectively.

The attention scores $\mathbf{A}$ are computed by taking the dot product of the query and key matrices, scaled by the square root of the dimension of the key vectors d_k followed by a softmax function to normalize the scores.

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$$

The final fused embedding $\mathbf{c}$ is obtained by multiplying the attention scores $\mathbf{A}$ with the value matrix $\mathbf{V}$. This embedding captures the combined information from both the patient knowledge graph and the sentence embeddings. The fused embedding $\mathbf{c}$ is then concatenated with the original sentence embedding $\mathbf{s}_{BERT}$ and the note-level updated sentence embedding $\mathbf{s}'_i$ to form the final representation for classification.

3.5 Final Classification

We concatenate the fused embedding $\mathbf{c}$, the note-level updated sentence embedding $\mathbf{s}'_i$ and the original sentence embedding $\mathbf{s}_{BERT}$ to form the final representation:

$$\mathbf{f}_i = [\mathbf{c}; \mathbf{s}_{\text{BERT}}; \mathbf{s}'_i] \qquad y_i = \text{FFNN}(\mathbf{f}_i)$$

where y_i is the predicted label for sentence S_i.

This final representation $\mathbf{f}_i$ is fed into a Feed-Forward Neural Network (FFNN) to predict whether each sentence contains key information. The loss function is Binary Cross-Entropy Loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where N is the number of sentences in the dataset, y_i is the true label for the i-th sentence (1 if the sentence contains key information, 0 otherwise), $\hat{y}_i$ is the predicted probability that the i-th sentence contains key information, obtained from the FFNN. And parameters in ClinicalBERT, Bi-LSTM, Cross-attention module and FFNN are jointly optimized.

4 Experiments

4.1 Data

We collected 300,000 pages of clinical notes obtained from an oncology clinic, comprising clinical notes from roughly 2,000 patients. Each patient's documentation includes records from multiple visits, with lengths ranging from 90 to 700 pages. We engaged 20 physicians to annotate key sentences indicative of ten specific elements: clinical events, medical history, medication, family history, oncology events, oncology medication, procedures, oncology procedures, reproductive potential, and social history. We only keep the pages that containing key sentences. Then we refined the dataset to 16,000 pages containing positive samples.

We collected 300,000 pages of clinical notes from an oncology clinic, encompassing records from approximately 2,000 patients. Each patient's documentation includes records from multiple visits, with document lengths ranging from 90 to 700 pages. To annotate key sentences indicative of ten specific elements—clinical events, medical history, medication, family history, oncology events, oncology medication, procedures, oncology procedures, reproductive potential, and social history—we engaged 20 physicians.

After annotation, we filtered the dataset to retain only the pages containing key sentences, resulting in a refined dataset of 16,000 pages with positive samples. We adopt 8:1:1 split for train, test and validation datasets. This unique dataset, annotated by experts, provides a robust foundation for developing and evaluating our model.

4.2 Experimental Results

We selected a diverse set of baseline models to comprehensively evaluate the performance of our proposed GEHE (Graph-Enhanced Hierarchical Encoder) framework. The chosen baselines include state-of-the-art models for contextualized text representation and generative language modeling. BERT, a widely used transformer model, captures bidirectional context, while RoBERTa improves upon BERT with enhanced training procedures (Devlin et al., 2019; Liu et al., 2019). BART, a denoising autoencoder, integrates bidirectional and autoregressive transformers (Lewis et al., 2020). T5 frames NLP tasks as text-to-text problems, excelling across benchmarks (Raffel et al., 2020). These models are considered to have strong performance among language models. For generative models, we selected the latest LLMs that are designed to handle various text-generation tasks, including Llama3, GPT3.5 and GPT4 (Brown et al., 2020; OpenAI, 2023). These baselines provide a robust comparison for evaluating the effectiveness of our GEHE framework.

Our model is trained on a A10G Nvidia GPU. The models are evaluated using four standard metrics for information extraction, including Accuracy (ACC), F1 Score (F1), Precision (Prec), and Recall (Rec). The results of performance comparison with baselines are presented in Table 1 and Table 2.

Our model (GEHE) achieves a substantial improvement over the baselines. GEHE boosts the highest baseline accuracy by 3.77% (0.7711), demonstrating superior capability in correctly classifying sentences as containing key information or not. It also achieves the highest F1 Score at 0.8035, which is 2.3% higher than the best baseline, effectively balancing Precision and Recall. GEHE's Precision of 0.7358 is 4.11% higher than the best baseline, underscoring its strength in accurately

Table 1: Performance comparison with baseline models.

Model	ACC	F1	Prec	Rec
BERT	0.7317	0.7765	0.6900	0.8877
RoBERTa	0.7334	0.7756	0.6947	0.8779
BART	0.7320	0.7805	0.6847	0.9076
T5	0.7294	0.7801	0.6802	**0.9145**
GEHE (Ours)	**0.7711**	**0.8035**	**0.7358**	0.8915

identifying key information sentences. While T5's Recall is 2.3% higher than our model's, GEHE maintains a remarkable performance with a well-balanced Precision and Recall, ensuring accurate identification of most key sentences. Our model stands out as the best in the task of identifying important sentences in clinical notes.

Table 2: Performance for generative models. Note that we use a sample of 300 sentences.

Model	ACC	F1	Prec	Rec
Llama3	0.5799	0.5156	0.6751	0.4559
GPT3.5	0.5933	0.4404	0.7059	0.3200
GPT4	0.7233	0.7296	0.7133	0.7467
GEHE(Ours)	**0.7726**	**0.8046**	**0.7348**	**0.8890**

When compared to the aforementioned state-of-the-art closed-source and open-source LLMs, GEHE demonstrates superior performance across all metrics, particularly in F1 Score and Accuracy. Due to limited computing resources and data privacy issues, we evaluated the models on a seperated 300-sentence dataset. Despite the generative models' strength in language generation tasks, they fall short in the specific task of key verbatim extraction from clinical notes. GEHE's focused approach and its ability to integrate graph-based patient information with hierarchical textual representations contribute significantly to its superior performance.

4.3 Ablation Study

We compared our model with various ablation settings to isolate the impact of different components of our approach.

The baseline ClinicalBERT model achieves an accuracy of 0.7300 and an F1 score of 0.7759. Adding contextual information beyond individual sentences, the stacked ClinicalBERT setup reaches an accuracy of 0.7449 and precision of 0.7081 without significantly enhancing recall. Adding a Bi-LSTM layer to ClinicalBERT to capture note-level context achieves the highest recall, comparable to our model. Introducing graph-based patient embeddings and using them in the cross-attention mechanism (with values from the graph and queries and keys from the text) boosts precision to 0.7560 and overall accuracy to 0.7690, though recall drops to 0.8268. Our GEHE model significantly enhances the ability to extract key information, achieving an accuracy of 0.7711 and an F1 score of 0.8035.

The ablation study highlights the importance of each component in our GEHE framework. Adding Bi-LSTM to ClinicalBERT enhances note-level context, improving overall performance. Incorporating patient-specific information through graph embeddings in the cross-attention mechanism significantly boosts precision, and the cross-attention fusion balances precision and recall, crucial for minimizing false positives and negatives in clinical applications.

4.4 Discussion

The results suggest that incorporating patient-specific information through graph-based embeddings, combined with sentence embeddings derived from ClinicalBERT and contextualized via Bi-LSTM, significantly enhances the model's ability to accurately extract key information from clinical notes. The cross-attention mechanism effectively fuses these multimodal representations, leading to improved classification performance.

The ablation study results highlight the importance of each component in our GEHE framework:

1. Contextual Integration: Adding Bi-LSTM to ClinicalBERT demonstrates the value of note-level context, improving the model's performance across several metrics.

2. Graph-based Enhancements: Incorporating patient-specific information through graph embeddings in the cross-attention mechanism provides a substantial boost to precision, showing that patient context is crucial for accurate extraction of key sentences.

3. Cross-Attention Fusion: The cross-attention mechanism effectively combines multimodal

Table 3: Ablation study for model evaluation.

Model	ACC	F1	Prec	Rec
ClinicalBERT	0.7300	0.7759	0.6877	0.8899
ClinicalBERT-ClinicalBERT	0.7449	0.7827	0.7081	0.8748
ClinicalBERT-BiLSTM	0.7400	0.7827	0.6973	**0.8918**
Gragh-Enhanced ClinicalBERT-BiLSTM (only v from Gragh)	0.7690	0.7898	**0.7560**	0.8268
Gragh-Enhanced ClinicalBERT-BiLSTM (GEHE, ours)	**0.7711**	**0.8035**	0.7358	0.8915

information, leading to a balanced improvement in both precision and recall, which is critical for clinical applications where both false positives and false negatives carry significant consequences.

5 Conclusion

Our hierarchical multimodal cross-attention framework, GEHE, provides a novel and effective graph-knowledge-enhanced methods for Key Verbatim Extraction. The model's superior performance in terms of Accuracy, F1 Score, and Precision underscores the importance of integrating diverse sources of information and leveraging advanced attention mechanisms. This approach not only advances the state-of-the-art in clinical text analysis but also holds potential for broader applications in healthcare and other domains where accurate information extraction is critical.

The ablation study confirms that the hierarchical multimodal cross-attention approach in our GEHE model significantly enhances the performance of key verbatim extraction from clinical notes. Each component—Bi-LSTM for contextual note-level information, graph-based patient embeddings, and cross-attention fusion—contributes to the model's overall effectiveness, making it a robust solution for clinical text analysis.

This research effectively addresses the complexities inherent in clinical text analysis. Our approach is unique in its ability to combine word, sentence, note, and patient-level data, providing a comprehensive framework for understanding clinical narratives. Furthermore, by pretraining our model on datasets that include relational information between patients, we open new avenues for understanding how inter-patient relationships can be leveraged to improve information extraction in healthcare contexts. For practical implications, our model contributes to the efficiency and effectiveness of healthcare delivery. By facilitating the rapid identification of critical information in clinical texts, our approach can assist healthcare providers in making informed decisions more swiftly, leading to better patient outcomes. Our validation of the model using real-world oncology clinic reports, verified by professional oncological physicians, underscores the applicability and potential impact of our method in clinical settings.

Limitations

Our model's performance heavily depends on the quality and quantity of available clinical notes, and it may not perform optimally with sparse or poor-quality data. Future work should explore data augmentation techniques and improved preprocessing to enhance data quality and standardize clinical terminology. Additionally, our GEHE framework's reliance on network data and defined entity relationships limits its effectiveness for documents lacking these relationships, reducing the accuracy of graph-based embeddings and cross-attention mechanisms.

Another limitation is that our model has only been validated on a medical dataset, raising concerns about its generalizability to other domains. The unique characteristics of medical data may not be present in other types, potentially limiting its applicability. Future work should test the model across various domains to ensure broader applicability and identify necessary adjustments.

Additionally, the use of patient-specific information, such as embeddings from a Graph Attention Network (GAT), raises concerns about privacy and data security. Ensuring strict privacy standards and data protection is essential but not fully addressed in this study. Future work should incorporate privacy-preserving techniques like differential privacy or federated learning to secure patient data and enable use across multiple institutions.

References

Micheal Abaho and Yousef H Alfaifi. 2023. Select and augment: Enhanced dense retrieval knowledge graph augmentation. *Journal of Artificial Intelligence Research*, 78:269–285.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Patrick Fernandes, Miltiadis Allamanis, and Marc Brockschmidt. 2018. Structured neural summarization. *arXiv preprint arXiv:1811.01824*.

Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Jun Gao, Ninghao Liu, Mark Lawley, and Xia Hu. 2017. An interpretable classification framework for information extraction from online healthcare forums. *Journal of healthcare engineering*, 2017(1):2460174.

Sara Nouri Golmaei and Xiao Luo. 2021. Deepnote-gnn: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Sifei Han, Robert F Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang R Tsui. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of biomedical informatics*, 127:103984.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Binxuan Huang and Kathleen M Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. *arXiv preprint arXiv:1909.02606*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. 2003. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150(1-2):239–290.

Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, and Robert Stewart. 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project. *BMJ open*, 7(1):e012012.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Shiori Kitamori, Hiroyuki Sakai, and Hiroki Sakaji. 2017. Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: augmenting bert with graph embedding for text classification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*, pages 369–382. Springer.

Carmen Luque, José M Luna, Maria Luque, and Sebastian Ventura. 2019. An advanced review on text mining in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1302.

Shervin Malmasi, Naoshi Hosomura, Lee-Shing Chang, C Justin Brown, Stephen Skentzos, and Alexander Turchin. 2017. Extracting healthcare quality information from unstructured data. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1243. American Medical Informatics Association.

Parth Mehta. 2016. From extractive to abstractive summarization: A journey. In *ACL (Student Research Workshop)*, pages 100–106. Springer.

N Moratanch and S Chitrakala. 2017. A survey on extractive text summarization. In *2017 international conference on computer, communication and signal processing (ICCCSP)*, pages 1–6. IEEE.

Begum Mutlu, Ebru A Sezer, and M Ali Akcayol. 2020. Candidate sentence selection for extractive text summarization. *Information Processing & Management*, 57(6):102359.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Chatgpt (june 13 version). `https://www.openai.com/chatgpt`.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification. *arXiv preprint arXiv:1909.08402*.

Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. 2024. Let your graph do the talking: Encoding structured data for llms. *arXiv preprint arXiv:2402.05862*.

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International conference on machine learning*, pages 7867–7876. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Justin T Reese, Daniel Danis, J Harry Caufield, Tudor Groza, Elena Casiraghi, Giorgio Valentini, Christopher J Mungall, and Peter N Robinson. 2023. On the limitations of large language models in clinical diagnosis. *medRxiv*.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. *arXiv preprint arXiv:1906.04684*.

Mourad Sarrouti, Carson Tao, and Yoann Mamy Randriamihaja. 2022. Comparing encoder-only and encoder-decoder transformers for relation extraction from biomedical texts: An empirical study on ten benchmark datasets. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 376–382.

Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. 2022. The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–40.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, 2(1):1–37.

Sifatullah Siddiqi and Aditi Sharan. 2015. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2).

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Oguzhan Tas and Farzad Kiyani. 2007. A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Swathilakshmi Venkatachalam, Lakshmana Pandian Subbiah, Regan Rajendiran, and Nithya Venkatachalam. 2020. An ontology-based information extraction and summarization of multiple news articles. *International Journal of Information Technology*, 12(2):547–557.

Arturo N Villanueva Jr and Steven J Simske. 2023. Algorithm parallelism for improved extractive summarization. In *Proceedings of the ACM Symposium on Document Engineering 2023*, pages 1–4.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Lawrence L Weed et al. 1968. Medical records that guide and teach. *N Engl J Med*, 278(11):593–600.

Reinhardt Wenzina and Katharina Kaiser. 2013. Identifying condition-action sentences using a heuristic-based information extraction method. In *International Workshop on Process-oriented Information Systems in Healthcare*, pages 26–38. Springer.

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. 2021. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681*.

Zexian Zeng, Yu Deng, Xiaoyu Li, Tristan Naumann, and Yuan Luo. 2018. Natural language processing for ehr-based computational phenotyping. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(1):139–153.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.