MedEx: Enhancing Medical Question-Answering with First-Order Logic based Reasoning and Knowledge Injection

Aizan Zafar[1*] **Kshitij Mishra**[1*] **Asif Ekbal**[2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]School of AI and Data Science, Indian Institute of Technology Jodhpur, India
aizanzafar@gmail.com, mishra.kshitij07@gmail.com, asif.ekbal@gmail.com

Abstract

In medical question-answering, traditional knowledge triples often fail due to superfluous data and their inability to capture complex relationships between symptoms and treatments across diseases. This limits models' ability to provide accurate, contextually relevant responses. To overcome this, we introduce **MedEx**, which employs First-Order Logic (FOL)-based reasoning to model intricate relationships between diseases and treatments. We construct FOL-based triplets that encode the interplay of symptoms, diseases, and treatments, capturing not only surface-level data but also the logical constraints of the medical domain. **MedEx** encodes the discourse (questions and context) using a transformer-based unit, enhancing context comprehension. These encodings are processed by a Knowledge Injection Cell that integrates knowledge graph triples via a Graph Attention Network. The Logic Fusion Cell then combines medical-specific logical rule triples (e.g., co-occurrence, causation, diagnosis) with knowledge triples and extracts answers through a feed-forward layer. Our analysis demonstrates **MedEx's** effectiveness and generalization across medical question-answering tasks. By merging logical reasoning with knowledge, MedEx provides precise medical answers and adapts its logical rules based on training data nuances. [1]

1 Introduction

The remarkable achievements of large pre-trained language models (PLM) (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Liu et al., 2019) have propelled extractive Question-Answering (QA) systems to unprecedented levels, matching and even surpassing human-level performance (Hermann et al., 2015; Rajpurkar et al., 2016; Lai et al., 2017). Nevertheless, these successes encounter a significant stumbling block when applied to the intricate domain of medical question-answering. This domain presents a unique challenge, demanding an understanding of the intricate dependencies between various medical symptoms and their corresponding treatments across different diseases. The limitations of the existing approaches are rooted in their reliance on conventional knowledge triples, which tend to introduce extraneous information and overlook the nuanced interplay of medical variables (Smith et al., 2020). For instance, in Figure 1, it can be seen that knowledge triples are not sufficient enough to convey the relationship between symptoms and solutions, whereas logical rules are efficiently able to extract the intricate relationship.

To address this challenge, QA systems must first recognize logical units, such as sentences, clauses, or other meaningful text spans, and subsequently discern the logical relationships between these units (Huang et al., 2021). Unfortunately, these logical structures often remain concealed within textual data and pose a considerable extraction challenge, aggravated by the fact that most datasets lack annotations for such structures. In recent years, there has been a growing interest in combining Pre-trained Language Models (PLMs) with logic reasoning, particularly in the context of addressing logical problems (Manhaeve et al., 2018; Dong et al., 2019).

Present knowledge triples fusion with PLMs fall short in providing the necessary reasoning capabilities, while symbolic reasoners face difficulties when applied to unstructured text. While Graph Neural Network (GNN)-based reasoning methods like DAGN (Huang et al., 2021) have shown promise, but they grapple with two notable limitations. First, despite its graph representation, it predominantly relies on neural techniques over discourse relations, raising questions about its ability to effectively approximate symbolic reasoning involving logical relations, such as implication and

[1]Code: https://github.com/aizanzafar/MedEx

Proceedings of the 31st International Conference on Computational Linguistics, pages 9701–9720
January 19–24, 2025.

Paragraph: You did your best to protect yourself from breaking a bone due to osteoporosis. Or maybe you didn't even know your bones were at risk. Either way, your fracture can heal, and you can work closely with your doctor to avoid it ever happening again. Fractures of the spine, hip, or wrist are the most common types in people... (truncated)

Question: What are treatment options for osteoporosis spine fractures?

Ground Truth: Treating a hip fracture depends on where your hip is broken, how severe the break is, and your overall health. Treatment options may include: Surgical repair with screws, nails, or plates A partial or total hip replacement Exercises so that you move better and build strength The best treatment depends on the location of the break.

KG Triples: ["treatment options", "co-occurs_with", "muscle spasms"], ["fractures", "co-occurs_with", "osteoporosis"], ["fractures","co-occurs_with", "broken hip"], ["fractures", "co-occurs_with", "side effects"], ["fractures", "co-occurs_with", "broken bone"], ["fractures", "co-occurs_with","falls"].....

Co-occurrence Impact Rule: ["osteoporosis", "affects", "compression fracture"], ["osteoporosis", "affects", "balance"], ["compression fracture", "affects", "side effects"], ["osteoporosis", "affects", "cavity"],.....

Diagnostic Correlation Rule: ["muscle spasms", "diagnoses", "fracture"], ["muscle spasms", "diagnoses", "trauma"], ["muscle spasms", "diagnoses", "fall"],....

Predicted Answer (KG): If your doctor recommends either procedure, talk with them about the risks, benefits, and recovery time. Treating a hip fracture depends on where your hip is broken, how severe the break is, and your overall health.

Predicted Answer (KG+FOL): Treatment options may include: Surgical repair with screws, nails, or plates A partial or total hip replacement Exercises so that you move better and build strength The best treatment depends on the location of the break.

Figure 1: Example of QA system based on KG triples alone and KG+FOL triples.

Rule	Condition	Implication
1	If X co-occurs with Y and Y affects Z	then X affects Z
2	If X prevents Y and Y causes Z	then X prevents Z
3	If X treats Y and Y is a type of Z	then X can be used to treat Z
4	If X is diagnosed with Y and X interacts with Z	then Z can be used for the diagnosis of Y
5	If X co-occurs with Y and X affects Z	then Y co-occurs with Z
6	If X prevents Y or Y causes Z	then X can either prevent or cause Z

Table 1: Proposed Logical Rules in the Medical Domain.

negation. Second, the resulting graph often exhibits loose connections and consists of long paths, potentially hindering the interaction between context and options, a critical aspect of answering multiple-choice questions. To confront this challenge, we propose **MedEx** (Medical Expertise) which leverages the First-Order Logic (FOL)-based reasoning to unravel intricate disease-solution dependencies. At its core, **MedEx** fuses FOL-based triples with traditional knowledge triples, to enrich the the surface-level data with the nuanced logical constraints defining the medical landscape. Employing six FOL-based rules *viz.* Co-occurrence, Prevention and Causation, Treatment and Classification, Diagnosis and Interaction, Conjunction and Disjunction, we first construct logic triples capturing intricate symptom-disease-treatment relationships using GNN. Second, contextual understanding is achieved by encoding context and questions through a PLM. Now we build our novel architecture employing two cells *viz. Knowledge Injection Cell* - integrating knowledge graph (KG) triples seamlessly through a Graph Attention Network, and *Logic Fusion Cell* - infusing domain-specific logical rules, encompassing co-occurrence, causation, treatment, diagnosis, and interaction. Lastly, answers are extracted from the context via a feed-forward layer. By uniting logical reasoning with a rich knowledge base, **MedEx** delivers precise responses while aligning with nuanced training data patterns. Our extensive experiments and empirical analysis demonstrate state-of-the-art results validating **MedEx**'s effectiveness and adaptability across diverse medical question-answering tasks. Our *key* contributions are summarized as follows:

1. Proposed a novel extractive medical question-answering system **MedEx** designing two cells *viz. Knowledge Injection Cell* and *Logic Fusion Cell* to facilitate logical reasoning with contextual reasoning.

2. Introduced six novel FOL based logical rules (Table 1) and constructed corresponding logic graphs resulting into logic triples which can be employed in Medical extractive Question Answers systems effectively.

3. Demonstrated the effectiveness of **MedEx** through extensive empirical analysis with state-of-the-art results across diverse medical question-answering tasks.

2 Related Work

The achievements of large pre-trained language models (PLM) (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Liu et al., 2019) have

led to remarkable progress in extractive Question-Answering (QA) systems, often reaching or exceeding human-level performance (Hermann et al., 2015; Rajpurkar et al., 2016; Lai et al., 2017). However, the domain of medical question-answering poses unique challenges, requiring a deep understanding of intricate relationships between medical symptoms and treatments across diseases (Smith et al., 2020).

Knowledge-Based Medical QA Systems: Early approaches in medical question-answering relied on curated knowledge bases and rule-based systems. Notable systems like AskHERMES (Cao et al., 2017) and EON (Abacha et al., 2015) demonstrated the effectiveness of leveraging structured medical knowledge for answering questions. While informative, these systems often struggled with handling nuanced and context-specific queries.

Semantic Graphs and Medical Ontologies: The use of semantic graphs and medical ontologies, such as SNOMED CT (Donnelly, 2006) and UMLS (Bodenreider, 2004), has been explored to enhance medical QA systems. These approaches excel at capturing relationships between medical concepts but may face challenges in scaling to handle complex and diverse medical queries.

Deep Learning-Based QA Systems: Recent advancements in deep learning have led to the development of neural network-based medical QA systems. Models like BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) leverage pre-trained language models to understand medical text better. While these models offer improved context understanding, they often lack explicit logic reasoning capabilities.

Logic-Based Reasoning in QA: Logic-based reasoning has gained traction in the broader field of question-answering. Systems like Prolog (Colmerauer and Roussel, 1991) and Neural-Symbolic Integration (Garcez et al., 2015) have demonstrated the power of logic in answering complex questions. However, their application in medical QA has been limited.

Graph Neural Networks (GNNs): Graph-based approaches, particularly GNNs, have shown promise in combining structured knowledge with unstructured text. Methods like MedGNN (Zhang et al., 2020) have leveraged GNNs to navigate medical knowledge graphs effectively. **MedEx** builds upon this idea, integrating GNNs for knowledge injection, thereby enhancing the model's understanding of medical relationships.

Hybrid Models: Hybrid models that combine deep learning with logical reasoning have recently emerged. Systems like MedQA (authors , not specified in the provided context) aim to bridge the gap between neural networks and logical inference, showcasing the potential for more context-aware medical QA.

Question Decomposition and Answer Extraction: Another approach involves breaking down complex medical queries into subtasks. The use of question decomposition (Shrivastava et al., 2021) and answer extraction (Jin et al., 2020) techniques has demonstrated improvements in handling intricate medical questions.

Present approaches combining PLMs with knowledge triples have limitations, and symbolic reasoners struggle with unstructured text. While GNN-based reasoning methods, like DAGN (Huang et al., 2021), show promise, they may not fully address logical relations involving implication and negation. Additionally, GNNs can create loosely connected graphs, hindering interaction between context and options. **MedEx** stands at the intersection of these research directions, incorporating FOL-based reasoning to navigate the complexities of medical knowledge while harnessing the power of deep learning for context-awareness and knowledge triples to capture relationship between medical variables. By employing six FOL-based rules, we construct logic triples, capturing symptom-disease-treatment relationships, and integrate it with knowledge graph triples, resulting in improved accuracy and adaptability across diverse medical question-answering tasks.

3 Methodology

Our proposed model, **MedEx** consists of three components, namely a *PLM-encoding*, *Knowledge Injection Cell* and *Logic Fusion Cell*. **MedEx** takes as input: *question* q - the query posed by the user, *context* cxt - supporting paragraphs/documents containing the information needed to extract out the answer a - the output. *PLM-encoding* is a pre-trained LLM which encodes the given input. *knowledge injection cell* employs Graph Attention Networks to generate knowledge triples (h, r, t), enabling **MedEx** to focus on pertinent entities and their relationships. *Logic Fusion Cell* employs FOL based rules to steer the MedEX in conducting advanced reasoning to grasp the intricate relationships within medical data. This helps ***MedEx*** to tackle

the complex QA tasks within the medical domain.

Integrating external knowledge is a crucial component of **MedEx**, facilitated through the utilization of knowledge graphs (KG) (Wang et al., 2014) to establish links among each question, answer, and supporting paragraph triplet $< Q, A, P >$. A KG is defined as a multi-relational graph $G = (V, E)$, where V denotes the entity nodes in the KG, and $E \subseteq V \times R \times V$ represents the set of edges (triples) connecting nodes in V, with R representing the set of relation types. The structured information within the KG is depicted as triples $(\tau_1, \tau_2, ..., \tau_{N_g})$, each consisting of a head (h), a relation (r), and a tail entity (t), where N_g indicates the total number of triples. In this context, we employed the Unified Medical Language System (UMLS) (Bodenreider, 2004) to construct the KG. The primary aim of **MedEx** is to seamlessly merge unstructured information from the context with structured knowledge from the KG to extract accurate answers.

3.1 PLM Encoding

In constructing our architecture, we carefully select the pre-trained language models tailored to the specific datasets employed. We utilize RoBERTa (Liu et al., 2019) for MASH-QA (Zhu et al., 2020) and COVID-QA (Möller et al., 2020) datasets, and BioLinkBERT (Yasunaga et al., 2022b) for BioASQ (Nentidis et al., 2022) and MedQA-USMLE (Jin et al., 2021) datasets. These pre-trained models serve as the foundational components of our architecture. We employ the selected PLM to encode the concatenated context cxt and question q.

$$T_{enc} = \text{LM}(cxt; q) \tag{1}$$

where $T_{\text{enc}} \in \mathbb{R}^{n \times d}$ is the encoded output. d is the PLM's hidden representation and n is the total number of words. For more details about the selection of pre-trained language models, refer to Appendix A.

3.2 Knowledge Injection Cell

Medical Knowledge Graph Creation: We construct the Medical Knowledge Graph (MKG) using the Unified Medical Language System (UMLS) to overcome the limitations of medical datasets lacking comprehensive information about medical entities. Following the methodology outlined by (Zafar et al., 2024), which involves knowledge-infused abstractive question answering, we employ QuickUMLS to extract medical entities and relationships from the Metathesaurus and the Semantic Network. Our approach involves two strategies: merging all supporting paragraphs into a single document or processing each paragraph individually to create a smaller, more relevant KG. We focus on extracting meaningful triples by filtering medical entities and relationships, resulting in a concise and contextually relevant MKG. This process reduces overhead and enhances the efficiency of MKG creation, facilitating accurate medical question answering within the MedEx system. For more details, refer to Appendix B.

Knowledge Injection Cell (KI Cell): The Knowledge Injection Cell (KI Cell) enriches the system's understanding of medical concepts and relationships by injecting relevant medical knowledge. We use Graph Attention Networks (GATs) (Velivckovic et al., 2017) within this cell to analyze the knowledge graph, which consists of interconnected nodes representing various medical entities and their relationships.

Within the KI Cell, GATs allow the model to learn the significance of different medical concepts and their interconnections. By focusing on relevant nodes in the knowledge graph and considering their links, the model gains a deeper understanding of the medical question's context, leading to more informed and accurate answers.

The output of this cell is an attention mechanism, i.e., Att_{KI}, which plays a crucial role in determining the relevance and importance of the encoded output T_{enc} (as described in § 3.1), allowing the model to focus on specific nodes in the knowledge graph that are most relevant to the current medical query. The output is $Att_{KI} \in \mathbb{R}^{n \times d}$:

$$Att_{KI} = \text{Attention}(T_{\text{enc}}, \text{GAT\_enc}) \tag{2}$$

The Att_{KI} in the Knowledge Injection Cell of MedEx regulates how domain-specific knowledge is integrated into the encoded output T_{enc}, ensuring that the injected information complements the existing contextual representation effectively.

3.3 Logic Fusion Cell

The Logic Fusion Cell is a crucial component in our architecture, responsible for integrating logical rules into the model's decision-making process. This cell enhances the model's reasoning capabilities, enabling it to comprehend complex relationships and dependencies within the data. It leverages

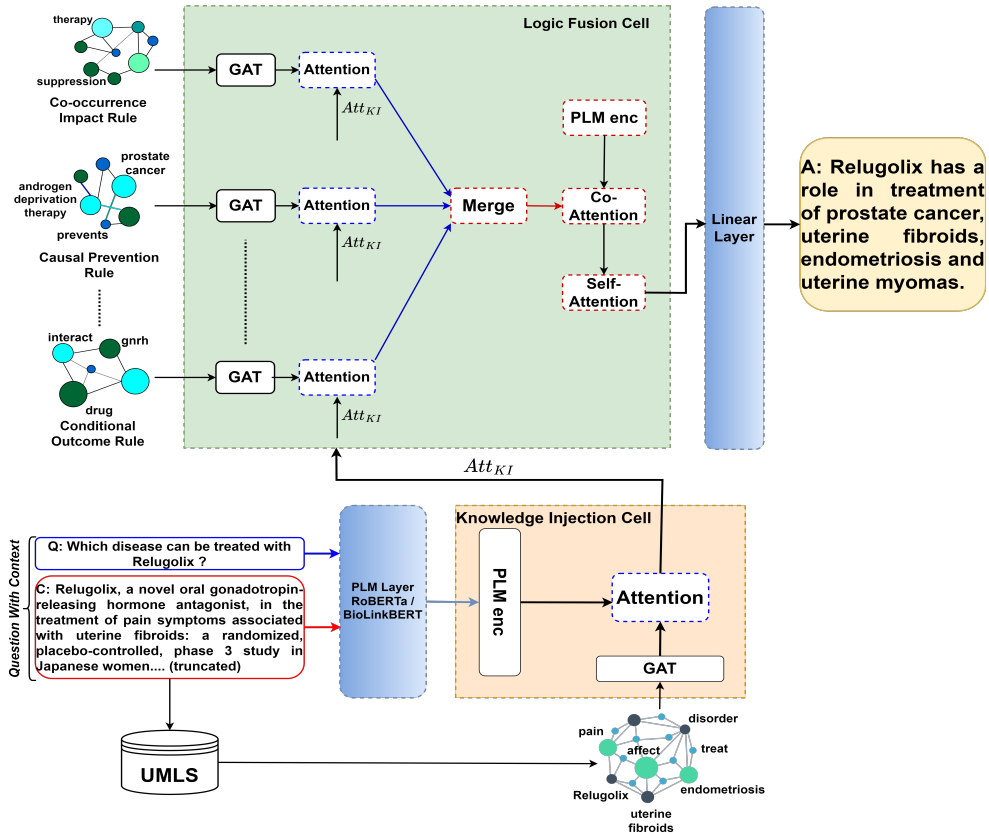

Figure 2: Overall architecture of the proposed system **MedEx**.

logical inference based on predefined rules specifically tailored for the medical domain, as illustrated in Table 1. These logical rules are seamlessly incorporated into the MedEx decision-making process, empowering it to provide informed and context-aware responses.

After evaluating multiple FOL-based rules, we finalized six that were most relevant and effective in enhancing the model's reasoning and inference capabilities. These rules, which serve as essential knowledge for the model, are as follows:

1. **Co-occurrence Impact Rule:** Captures the relationship where X co-occurs with Y, and Y affects Z, thus X also affects Z.

$$\text{co\_occurs\_with}(X, Y) \wedge \text{affects}(Y, Z) \Rightarrow \text{affects}(X, Z) \quad (3)$$

2. **Causal Prevention Rule:** If intervention X prevents event Y, and Y causes event Z, then X also prevents Z.

$$\text{prevent}(X, Y) \wedge \text{causes}(Y, Z) \Rightarrow \text{prevent}(X, Z) \quad (4)$$

3. **Treatment Inheritance Rule:** If X treats Y and Y is a type of Z, then X can be used to treat Z too.

$$\text{treat}(X, Y) \wedge \text{is\_a}(Y, Z) \Rightarrow \text{treat}(X, Z) \quad (5)$$

4. **Diagnostic Correlation Rule:** If X is diagnosed with Y and interacts with Z, then Z can assist in diagnosing Y.

$$\text{diagnosis}(X, Y) \wedge \text{interacts\_with}(X, Z) \Rightarrow \text{diagnosis}(Z, Y) \quad (6)$$

5. **Co-occurrence Propagation Rule:** If X co-occurs with Y and affects Z, then Y also co-occurs with Z.

$$\text{co\_occurs\_with}(X, Y) \wedge \text{affects}(X, Z) \Rightarrow \text{co\_occurs\_with}(Y, Z) \quad (7)$$

6. **Conditional Outcome Rule:** If X prevents Y or Y causes Z, then X either prevents or causes Z.

$$\text{prevent}(X, Y) \vee \text{causes}(Y, Z) \Rightarrow (\text{prevent}(X, Z) \vee \text{causes}(X, Z)) \quad (8)$$

These rules are integrated with knowledge graph (KG) triples extracted from the UMLS. Each rule generates a new set of KG triples, which are then compared pairwise using standard tri-linear attention mechanisms (Seo et al., 2016) with Att_{KI} (as described in § 3.2). This fusion process ensures the effective incorporation of logical rules into MedEx's reasoning framework. For more details on the derivation of these logical rules, please refer to Appendix C.

Further, Tri-linear attention is a critical mechanism that enables the model to capture interactions between the encoded context and the fused logical rules. It identifies the most relevant parts of the context and rules, allowing for better integration of logical rules into the reasoning process.

$$Att_{R_i}(u, v) = \text{W1} \cdot u + \text{W2} \cdot v + (\text{W3} \odot v) \cdot u \tag{9}$$

where, $W1, W2, W3 \in \mathbb{R}^d$ are trainable weights and $u \in \mathbb{R}^{x \times d}$, $v \in \mathbb{R}^{n \times d}$ are input matrices; x and n here are generic placeholder for input lengths; matrix multiplication and element-wise multiplication are denoted by $(\cdot)$ and $(\odot)$, respectively.

The tri-linear attention outputs, Att_{R_i}, for each rule are combined into a fused representation, R_{fuse}. This merging process aggregates the relevance scores of each rule with respect to the encoded context, ensuring that the model considers the combined influence of all rules during reasoning.

$$R_{fuse} = MLP(concat(Att_{R_i})_{i=0}^{6}) \tag{10}$$

The concatenated output is then passed through a feed-forward layer with d neurons to learn complex representations and capture intricate patterns, mapping the output to a d-dimensional space for further processing.

After merging, Co-attention (Xiong et al., 2016) mechanisms are subsequently applied to refine the hidden representation T_{enc} using the fused rule-correlation features. Specifically, co-attention mechanisms A_{tr} and A_{rt} are computed between and T_{enc} and R_{fuse}, allowing the model to recalibrate T_{enc} based on the weighted sum of R_{fuse}. This ensures the model focuses on the most relevant aspects of both the context and the fused rule representations.

$$A_{tr} = \text{Att}(T_{enc}, R_{fuse}) \tag{11}$$

$$A_{rt} = \text{Att}(R_{fuse}, T_{enc}) \tag{12}$$

$$R_t = A_{rt} \cdot [T_{enc}; A_{tr} \cdot R_{fuse}] \tag{13}$$

$$\hat{R}_t = \text{ReLU}(Wp([R_t; R_{fuse}])) \tag{14}$$

The concatenated vector $[R_t; R_{fuse}]$ is passed through a feed-forward layer with parameters W_p of size $\mathbb{R}^{3d \times d}$, followed by a ReLU activation function. This process enables the model to capture complex relationships, enhancing its reasoning and decision-making capabilities.

Finally, self-attention (Wang et al., 2017) is applied to compute the final rule representation, $\hat{R}_f$. This step captures relationships within the rule representation $\hat{R}_f$, further refining the model's understanding of the injected logic.

$$R_s = \hat{R}_t \cdot \text{Att}(\hat{R}_t, \hat{R}_t) \tag{15}$$

$$R_f = [\hat{R}_t; R_s, \hat{R}_t - R_s, \hat{R}_t \circ R_s] \tag{16}$$

$$\hat{R}_f = \text{ReLU}(W_f \cdot R_f) \tag{17}$$

Here, R_s represents the self-attention output, and $(\hat{R}_f)$ is the final rule representation, which the model uses for more informed decision-making by capturing relationships and dependencies within the rules.

The purpose of applying self-attention is to allow the model to focus on different aspects of the rule representation $\hat{R}_t$ and learn complex dependencies within the rules. By capturing relationships and interactions within the rule representation, the final rule representation $\hat{R}_f$ enhances the model's understanding of the injected logic. It facilitates more informed decision-making during the reasoning process.

4 Dataset

To check the robustness of the proposed model, we train and test **MedEx** on four diverse medical question-answering datasets: **(i.)** MASH-QA (Zhu et al., 2020), encompasses consumer healthcare questions from WebMD, with a broad spectrum of health topics and approximately 25K question-answer pairs, (ii.) COVID-QA (Möller et al., 2020), focuses on COVID-19 with 2,019 curated question-answer pairs related to the pandemic, (iii.) BioASQ Task 10b Phase B (QA Task) (Nentidis et al., 2022) is designed as part of the BioASQ challenge, emphasizing biomedical question-answering in the life

sciences domain, (iv.) MedQA-USMLE (Jin et al., 2021) comprises 4-way multiple-choice questions from USMLE practice tests, providing a rigorous assessment of the model's medical knowledge with a total of 12,723 questions. Detailed dataset details and statistics, please refer to the Appendix D.

5 Experiments

5.1 Baselines

We evaluated the **MedEx** model's performance against eleven strong baseline models (fine-tuned on all considered datasets): BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2020), XLNet (Yang et al., 2019), MultiCo (Zhu et al., 2020), RoBERTa (Liu et al., 2019), BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019), BioLinkBERT (Yasunaga et al., 2022b), QA-GNN (Yasunaga et al., 2021), GreaseLM (Zhang et al., 2021), and DRAGON (Yasunaga et al., 2022a). These baselines were selected for their diverse representation capabilities and relevance to the medical domain. Each model offers unique strengths, such as capturing contextual relationships, emphasizing span representations, using autoregressive techniques for bidirectional context, or specializing in biomedical and clinical text analysis. Detailed descriptions of these baselines are provided in section E of the Appendix.

5.2 Implementation Details

We implement all the models on a train:test split of 80:20. For all the models, we used random_seed=40, learning rate = 1e-5, dropout = 0.2, Adam optimizer (Kingma and Ba, 2014), and n_epochs = 15. The implementation utilized the A100-PCIE-40GB with CUDA version 11.2 for GPU acceleration. Each training epoch lasted approximately 1.5 hours. Additional information about hyperparameters can be found in Appendix F.

5.3 Evaluation Metrics

For MASH-QA and COVID-QA datasets, all models are assessed employing two evaluation metrics: Exact Match (**EM**)

$$EM = \mathbb{E}_{[cxt+q],a} \mathbb{1}\{\sum p_\theta(y|[cxt+q]) = a)\} \quad (18)$$

and F1-score (F1).

For BioASQ and MedQA-USMLE datasets, we evaluated all models using accuracy ACC measure.

6 Results and Analysis

We perform both quantitative analysis and qualitative analysis to assess the the performance of proposed model on fours datasets *viz.* MASH-QA, COVID-QA, BioASQ and MedQA-USMLE.

6.1 Quantitative Analysis

Table 2 presents the performance of **MedEx** on the MASH-QA and COVID-QA datasets, compared to eleven baseline models. **MedEx** achieved the highest scores, with an EM of 39.52 and F1 of 69.17 on MASH-QA, and an EM of 32.72 and F1 of 61.98 on COVID-QA, outperforming all baselines, including the state-of-the-art DRAGON model. Specifically, **MedEx** surpassed DRAGON by 3.63 EM and 2.19 F1 on MASH-QA, and by 2.15 EM and 4.26 F1 on COVID-QA.

The superior performance of **MedEx** can be attributed to its innovative integration of First-Order Logic (FOL) with domain-specific knowledge, enhancing its ability to capture intricate medical relationships. Notably, using RoBERTa-large as the PLM encoder in **MedEx** contributed to these improved results.

Further results in Table 3 on the BioASQ and MedQA-USMLE datasets show that **MedEx** maintained its lead, scoring 96.6 on BioASQ and 48.3 on MedQA-USMLE. On BioASQ, **MedEx** edged out DRAGON by 0.2, while on MedQA-USMLE, it outperformed all baselines, including the specialized models like BioLinkBERT and GreaseLM.

Statistical significance testing using the t-test confirmed that the performance improvements of **MedEx** over DRAGON on MASH-QA and COVID-QA were statistically significant with 95% confidence ($p < 0.05$). This indicates that the observed gains were not due to chance, underscoring the effectiveness of **MedEx** in handling both general and specialized biomedical queries.

6.2 Qualitative Analysis

6.2.1 Impact of Each Module in MedEx:

The impact of each module within the MedEx framework is evident from the ablation study results in Table 4. The baseline model, PLM-enc (RoBERTa for MASH-QA and COVID-QA, BioLinkBERT for BioASQ and MedQA-USMLE), shows moderate performance across all metrics.

Incorporating the Knowledge Injection (KI) module, denoted as PLM-enc+KI, leads to improved scores across all datasets, enhancing the

Model	MASH-QA		COVID-QA	
	EM	F1	EM	F1
BERT	3.83	27.93	20.93	40.44
SpanBERT	5.62	31.47	23.01	44.14
XLNet	22.70	56.46	28.54	55.45
MultiCo	29.34	64.94	na	na
RoBERTa	34.77	66.08	25.9	59.53
BioBERT	27.81	56.73	23.41	44.29
ClinicalBERT	24.90	53.85	23.10	44.33
BioLinkBERT	32.53	64.52	24.83	51.52
QA-GNN	33.87	65.23	25.41	52.51
GreaseLM	34.37	66.03	26.11	54.09
DRAGON	35.89	66.98	30.57	57.72
MedEx	**39.52**	**69.17**	**32.72**	**61.98**

Table 2: Result on MASH-QA and COVID-QA dataset

Model	BioASQ	MedQA-USMLE
BioBERT	84.1	36.7
PubmedBERT	87.5	38.1
BioLinkBERT	94.2	44.6
QA-GNN	95.0	45.0
GreaseLM	94.9	45.1
DRAGON	96.4	47.5
MedEx	**96.6**	**48.3**

Table 3: Result on BioASQ and MedQA-USMLE dataset

model's performance compared to the baseline. Further integration of the Logic Fusion (FOL) module, as seen in the PLM-enc+FOL configuration, results in additional performance gains, surpassing the PLM-enc+KI configuration.

The combined use of both KI and FOL in the PLM-enc+KI+FOL configuration achieves the highest performance across all metrics, highlighting the synergy between knowledge injection and logical fusion in enhancing model accuracy and effectiveness for complex medical queries. A detailed analysis of the impact of each logical rule on model performance is provided in section H of the appendix due to space constraints.

6.3 Error Analysis

To assess MedEx's strengths and limitations, we analyzed cases where the model underperformed across different datasets. In the MASH-QA and COVID-QA datasets, we examined 200 incorrect predictions based on EM scores, categorizing the errors into two types:

1. **Ambiguously Defined Answer Spans (50%)**: These errors arose when ground truth spans were overly extended, creating ambiguity. For example, in Table **??**, the model generated a concise and accurate answer, while the ground truth was lengthy and redundant. Despite a low EM score, the model's response was logically sound and more relevant. 2. **Semantic Inadequacies (28%)**: Here, the model's answers were semantically better than the ground truth. For instance, when asked, "How can you recover from delusional disorder?" the model provided a recovery-focused response, while the ground truth described the disorder itself, highlighting challenges with medical terminology interpretation.

For BioASQ and MedQA-USMLE, we reviewed 50 and 100 incorrect predictions, respectively. A key issue was incomplete KG triples. For instance, in BioASQ, missing details in KG triples prevented the model from answering, "Is the protein Papilin secreted?" accurately. In MedQA-USMLE, missing or inaccurately labeled triples similarly impacted performance. This analysis pinpoints areas where MedEx struggles, providing insights for targeted improvements in medical QA tasks. Detailed comparison studies are provided in Appendix G.

7 Conclusion

In this study, we proposed **MedEx**, a medical question-answering system that effectively addresses complex medical queries by leveraging FOL-based reasoning, knowledge injection, and logic infusion cells. We assessed its performance across multiple datasets, including MASH-QA, COVID-QA, BioASQ, and MedQA-USMLE, comparing it against 11 baselines. **MedEx** demonstrated remarkable proficiency in addressing biomedical and healthcare-related queries. On MASH-QA and COVID-QA datasets, it surpassed state-of-the-art baselines, achieving EM scores of 39.52 and 32.72, and F1 scores of 69.17 and 61.98, respectively. For BioASQ and MedQA-USMLE datasets as well, it outperformed all baselines, achieving an impressive EM score of 96.6 and 48.3, respectively. These results underscore MedEx's effectiveness in handling both general biomedical questions and specialized medical queries. The confluence of knowledge and FOL-based reasoning exhibited a synergistic effect, enhancing **MedEx**'s ability to provide precise, context-aware answers. In future, we aim to refine logical rules and address ambiguous questions while exploring a *transitivity relevance* measure to classify inference requirements, enhancing MedEx's reasoning evaluation and understanding transitivity's impact on performance.

Model	MASH-QA		COVID-QA		BioASQ	MedQA
	EM	F1	EM	F1	Acc	Acc
PLM-enc	34.77	66.08	25.90	59.53	94.2	44.6
PLM-enc+KI	36.75	66.82	27.35	60.41	94.9	45.8
PLM-enc+FOL	37.34	67.53	28.22	61.01	95.6	47.1
PLM-enc+KI+FOL	39.52	69.17	32.72	61.98	96.6	48.3

Table 4: Ablation study results for **MedEx**, where 'PLM-enc' represents RoBERTa for MASH-QA and COVID-QA, and BioLinkBERT for BioASQ and MedQA-USMLE.

Limitations

Despite its impressive performance in medical question-answering, **MedEx** comes with it's limitations. As it relies on the knowledge graphs, hence, it is susceptible to inaccuracies when external knowledge is missing or inadequately complements contextual information, potentially impacting response accuracy. Additionally, due to computational constraints, the model primarily accesses information within two-hop neighborhoods of the knowledge graph, which can limit its responses to queries requiring data beyond this range. Furthermore, addressing highly complex or multi-faceted queries remains a challenge. Improving the model's logical reasoning, especially for complex First-Order Logic (FOL) rules, is a potential area for future research. These limitations highlight the need for enhancing knowledge coverage, scalability, domain adaptability, knowledge graph accuracy, and the model's handling of complex queries to maximize **MedEx**'s utility and robustness in medical question-answering tasks. This constitutes towards future directions of our work.

Ethics Statement

We acknowledge the importance of data ethics and copyright compliance in our research endeavors. We strictly employ publicly available datasets while adhering to copyright laws and the policies defined by the dataset providers. We prioritize due diligence and compliance to ensure the lawful and ethical use of data, thereby avoiding any infringement of copyright issues. Furthermore, it is crucial to note that our research paper does not involve studies with human participants conducted by any of the authors. We uphold the principles of research ethics and human subject protection, ensuring that our work does not encompass any direct involvement with human subjects. We emphasize the importance of respecting ethical guidelines concerning studies involving humans and confirm our adherence to these standards. As our system relies on a LLM-based model extracted responses, there is a risk that users might overly depend on these responses, potentially diminishing critical thinking and clinical judgment.

Acknowledgement

Authors gratefully acknowledge the generous support for the project "Percuro-A Holistic Solution for Text Mining", sponsored by Wipro Ltd.

References

Abdel Fattah Abacha, Md Faisal Mahbub Chowdhury, Abeed Sarker, and Elke A Rundensteiner. 2015. Overview of the mediqa 2015 question answering for clinical decision support challenge. *arXiv preprint arXiv:1511.03417*.

MedQA authors (not specified in the provided context). 2019. Medqa: A large-scale medical question answering dataset. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Yu Cao, Shuai Wang, George Hripcsak, and Fei Xia. 2017. Askhermes: An online question answering system for complex clinical questions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Alain Colmerauer and Pierre Roussel. 1991. *Prolog and natural-language analysis*. Ablex Publishing Corporation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Bidirectional encoder representations from transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hu Dong, Yang Yang, and Mirella Lapata. 2019. Neural logic machines. *arXiv preprint arXiv:1904.11694*.

K Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.

Artur d'Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. 2015. Neural-symbolic learning and reasoning: contributions and challenges. In *2015 AAAI Spring Symposium Series*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*.

Kevin Huang, Jaan Altosaar, Rajesh Ranganath, and Dustin Tran. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5848–5855.

Di Jin, Gamal Crichton, Cody Buntain, Will Rodriguez, Devendra Sachan, and Dragomir Radev. 2020. Formalizing medical knowledge into a question answering dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4843–4853.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset for machine reading. *arXiv preprint arXiv:1704.04683*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*, 31.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2022. Overview of bioasq 2022: The tenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 337–361. Springer.

Alec Radford, Jiajun Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Pranav Rajpurkar, Jian Zhang, Dmitry Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Vineet Shrivastava, Shikhar Gupta, and Ankit Arya. 2021. Contextual decomposition for medical question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Joshua Smith, Michael R Jones, Ryen W White, and Stephen H Muggleton. 2020. Explainable and justifiable reasoning in ai question answering. *arXiv preprint arXiv:2001.06385*.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Petar Velivckovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1591–1601.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. In *International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022a. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546.

Aizan Zafar, Sovan Kumar Sahoo, Harsh Bhardawaj, Amitava Das, and Asif Ekbal. 2024. Ki-mag: A knowledge-infused abstractive question answering system in medical domain. *Neurocomputing*, 571:127141.

Jing Zhang, Xiaolong Wang, Liheng Xu, Xiang Wang, and Minlie Huang. 2020. Medgnn: Graph neural networks for medical entity linking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. Greaselm: Graph reasoning enhanced language models. In *International conference on learning representations*.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Selection of Pre-trained Language Models

The decision to utilize RoBERTa for the MASH-QA and COVID-QA datasets and BioLinkBERT for the BioASQ and MedQA-USMLE datasets is primarily guided by their performance and suitability for the respective tasks.

1. RoBERTa for MASH-QA and COVID-QA:
    - **General Language Understanding:** RoBERTa's extensive pre-training captures diverse linguistic patterns and context, suitable for varied topics in MASH-QA and COVID-QA.
    - **Fine-Tuning Capabilities:** Its architecture allows efficient adaptation to medical question-answering tasks.
    - **Large Model Capacity:** RoBERTa's capacity enables capturing complex relationships, vital for accurate responses.
2. BioLinkBERT for BioASQ and MedQA-USMLE:
    - **Biomedical Text Mining Specialization:** BioLinkBERT is tailored for biomedical tasks, ideal for the medical domain in BioASQ and MedQA-USMLE.
    - **Domain-Specific Vocabulary:** Trained on biomedical literature, it effectively understands medical terminology.
    - **Semantic Understanding:** BioLinkBERT's focus on biomedical text semantics ensures contextually relevant answers.

These characteristics make RoBERTa and BioLinkBERT well-suited for their respective datasets, providing the necessary capabilities to address the challenges inherent in medical question-answering tasks.

B Medical Knowledge Graph

To overcome the limitation of medical datasets lacking comprehensive information about medical entities, we leverage the Unified Medical Language System (UMLS) as a rich source of medical knowledge. Specifically, we construct a self-built knowledge graph using Quick-UMLS (Soldaini and Goharian, 2016), which is based on the UMLS (Bodenreider, 2004).

The UMLS comprises three primary knowledge sources: the Metathesaurus, the Semantic Network, and the Specialist Lexicon and Lexical Tools. For our purposes, we focus on two of these sources: the Metathesaurus and the Semantic Network. The Metathesaurus contains an extensive collection of biomedical concepts and their relationships, with each concept associated with one or more Semantic Types. The Semantic Network provides detailed information about semantic types and relationships between them, including examples such as ''*disease*', '*symptom*', and '*laboratory test*', as well as relationships like '*is-a*', '*part-of*', and '*affects*'

B.1 Relevant KG Construction

There are two strategies for constructing knowledge graph (KG) triples. The first method involves merging all supporting paragraphs into a single document, while the second method processes each supporting paragraph individually.

In the first approach, all supporting paragraphs, such as those in the COVID-QA dataset with 147 contexts, are merged into a unified document. This document is then processed through the UMLS (Bodenreider, 2004) to generate a large KG. However, this method has drawbacks, including significant computing time and the risk of introducing irrelevant or unwanted triples into the KG.

Conversely, the second method processes each supporting paragraph separately through the UMLS (Bodenreider, 2004) to generate a smaller and more pertinent KG. For example, the head entity "Relugolix" may generate approximately 500-600 tail entities using this method. By focusing on individual contexts, this approach ensures that the KG contains only meaningful triples relevant to the specific question and context at hand.

Our approach for constructing the Medical Knowledge Graph involves the following steps:

1. **Medical Entity Extraction:** Identifying medical entities from each context using the Metathesaurus. Each distinct concept found in the UMLS is represented as a node in our knowledge graph.

2. **Relation Extraction:** Sourcing relations from both the Metathesaurus and the Semantic Network of UMLS.

3. **Graph Construction:** Using the extracted relations, we establish connections between the filtered medical concepts retrieved from UMLS. This process results in the construction of a smaller Knowledge Graph, which functions as a pertinent subgraph of UMLS.

By leveraging UMLS and following these steps, we create a Medical Knowledge Graph that enriches our understanding of medical concepts and their relationships. This facilitates more accurate and context-aware medical question answering within the *MedEx* system.

The decision to utilize the second method, which involves processing each supporting paragraph individually, was guided by several key considerations:

- **Precision and Relevance:** Ensuring that the KG contains only relevant information directly related to the specific context of each medical question.

- **Efficiency and Scalability:** Enhancing computational efficiency and scalability by processing each supporting paragraph separately.

- **Risk Mitigation:** Reducing the risk of introducing irrelevant or unwanted triples into the KG, thereby improving its quality and reliability.

- **Contextual Relevance and Tailoring:** Capturing the unique context of each medical question to tailor the KG construction process accordingly.

- **Enhanced Understanding:** Creating a KG that enriches our understanding of medical concepts and their relationships, facilitating more accurate and context-aware medical question answering.

C Derivation of Logical Rules

The derivation of the six logical rules was a multi-step process designed to ensure their effectiveness across various medical knowledge domains. We began by constructing a medical knowledge graph (KG) using the Unified Medical Language System (UMLS), which includes a vast array of semantic relations between medical entities. From the 54 available semantic relations in UMLS, we carefully selected seven key relations— "co-occurs-with," "prevent," "treat," "diagnosis," "interacts-with," "affects," and "causes."

This selection was guided by the need to balance computational efficiency with relevance, as incorporating too many relations could overburden computing resources and introduce unnecessary information into the KG. The KG was then built using this curated set of relationships, drawing on contextual information from medical documents.

For example, when analyzing content related to the treatment of uterine fibroids with Relugolix, the KG included triples like ["androgen_deprivation_therapy", "affects", "uterine_fibroids"] and ["heavy_menstrual_bleeding", "co-occurs-with", "uterine_fibroids"]. After analyzing these triples, we identified patterns that suggested logical relationships, such as a rule "co-occurs-with(X, Y) $\wedge$ affects(Y, Z) leads to affects (X, Z)." Medical experts validated the initially derived rules to ensure accuracy and domain relevance. Out of an initial set of twelve candidate rules, six were ultimately retained, as the others did not produce valid logical triples. Through this thorough cross-verification with medical professionals, the final six rules provided an optimal balance between interpretability and practical relevance for medical QA tasks, achieving accuracy without exhaustive manual intervention. This framework thus ensures reliable, end-to-end question-answering performance with minimal manual oversight.

D Datasets

To check the robustness of the proposed model, we train and test **MedEx** on four diverse medical question-answering datasets:

1. **MASH-QA (Zhu et al., 2020):** It comprises of consumer healthcare questions collected from the well-known health website WebMD, which includes content from a wide range of consumer healthcare sectors like general health, mental health, and nutrition, among others. These sections cover questions regarding frequent healthcare difficulties that people confront, such as symptoms, treatment options, and general health advice. It is the largest available dataset with approximately 25K question-answer pairs. The answers to the questions are generally non-factoid in nature, with longer answer spans. Statistics of MASH-QA are shown in Table 5

	MASH-QA
Contexts	5,574
QA pairs	34,808
Train QA	27,728
Dev QA	3,493
Test QA	3,587

Table 5: MASH-QA Dataset statistics

2. **COVID-QA (Möller et al., 2020):** It consists of 2,019 question-answer pairs related to COVID-19, curated by volunteer biomedical experts. This dataset is unique in its focus on COVID-19 and differs from typical open-domain Machine Reading Comprehension (MRC) datasets. Statistics of COVID-QA are shown in Table 6.

	COVID-QA
Contexts	147
QA pairs	2019
Train QA	1414
Dev QA	404
Test QA	201

Table 6: COVID-QA Dataset statistics

3. **BioASQ Task 10b Phase B (QA Task)** (Nentidis et al., 2022)**:** It had been explicitly designed as part of the BioASQ challenge, an initiative dedicated to advancing the state of the art in biomedical semantic indexing and question-answering systems. Task 10b Phase B focuses on the question-answering aspect and is an integral component of BioASQ's multifaceted challenges in the biomedical and life sciences domain. Statistics of BioASQ are shown in Table 7.

	BioASQ
Contexts	1148
QA pairs	1148
Train QA	861
Dev QA	172
Test QA	115

Table 7: BioASQ Dataset statistics

4. **MedQA-USMLE (Jin et al., 2021):**The MedQA-USMLE dataset serves as a specialized benchmark for evaluating medical question-answering systems, particularly in the context of preparing for the United States Medical Licensing Examination (USMLE). It encompasses a meticulously curated assortment of medical questions and corresponding answers designed to gauge candidates' understanding of medical concepts pertinent to the USMLE. These questions cover a wide spectrum of medical domains, mirroring real-world clinical scenarios and requiring candidates to apply medical reasoning and problem-solving skills. It consists of 4-way multiple-choice questions that require deep biomedical and clinical knowledge. Its specialized nature and focus on USMLE-style questions make it an invaluable resource for benchmarking the performance of medical question-answering systems. It comprises of 12,723 questions. Statistics of MedQA are shown in Table 8

	MedQA
QA pairs	12,743
Train QA	10,195
Dev QA	1,274
Test QA	1,274

Table 8: MedQA-USMLE Dataset statistics

Examples from different datasets are shown in Tables 9 and 10.

Dataset	Example
MASH-QA	Q: What should I do if I smoke and have prediabetes? Context: When your doctor tells you that you have prediabetes, you might think there's no reason to take action just yet. Or you might assume that you're definitely going to get diabetes. Not so! You do need to take prediabetes seriously, but there's still time to turn things around – if you start now. The goal is to get your blood sugar level out of the prediabetes range, and keep it that way. What you do every day makes a big difference. Making lifestyle changes may be even more powerful than just taking medication.. (truncated) A: Smoking is strongly linked to diabetes: People who smoke are 30% to 40% more likely to develop type 2 diabetes than those who don't. And people with diabetes who continue smoking are more likely to develop complications such as heart disease and blindness. So the sooner you ditch the cigarettes, the better.
COVID-QA	Q: What cells are infected by the PED virus? Context: Mucosal immune responses induced by oral administration recombinant Bacillus subtilis expressing the COE antigen of PEDV in newborn piglets. Abstract: Porcine epidemic diarrhea (PED) is a highly contagious disease in newborn piglets and causes substantial economic losses in the world. PED virus (PEDV) spreads by fecal oral contact and can be prevented by oral immunization. Therefore, it is necessary to develop an effective oral vaccine against PEDV infection. Currently, Bacillus subtilis as recombinant vaccine carrier has been used for antigen delivery and proved well in immune effect and safety. The present study evaluated the immunogenicity of recombinant A: intestine epithelial cells

Table 9: Examples from Different Datasets

E Baseline

We compared **MedEx**'s performance with eleven strong baselines due to their recognition within the NLP community, their capacity to represent a wide range of approaches and their relevance to the medical domain.

1. **BERT** (Devlin et al., 2019): It's an encoder based transformer employing attention mechanism to capture contextual relationships between words in a text. For extractive question-answering tasks, we utilize a BERT model with a span classification head, which includes a linear layer for predicting span start and end logits.

2. **SpanBERT** (Joshi et al., 2020): It's a variant of the BERT model that emphasizes span representations. It is trained by masking contiguous token spans and optimizing two objectives: masked

Dataset	Example
MedQA-USMLE	A 57-year-old man presents to his primary care physician with a 2-month history of right upper and lower extremity weakness. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had increasing difficulty with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his family have had musculoskeletal problems. His right upper extremity shows forearm atrophy and depressed reflexes while his right lower extremity is hypertonic with a positive Babinski sign. Which of the following is most likely associated with the cause of this patients symptoms? (A) HLA-B8 haplotype (B) HLA-DR2 haplotype (C) A: Mutation in SOD1 (D) Mutation in SMN1
BioASQ	Q: Has FTY720 been considered for the treatment of stroke? Context: FTY720 (Fingolimod) Ameliorates Brain Injury through Multiple Mechanisms and is a Strong Candidate for Stroke Treatment Many researchers have recognized the positive effects of FTY720 and launched basic and clinical experiments to test the use of this agent against stroke. Although the mechanism of FTY720 has not been fully elucidated, its efficacy against cerebral stroke is becoming clear, not only in animal models, but also in ischemic stroke patients through clinical trials. In this article, we review the data obtained from laboratory findings and preliminary clinical trials using FTY720 for stroke treatment. (A) A: yes (B) no

Table 10: Examples from Different Datasets

language modeling to predict its own vector representation, and the span boundary objective, which predicts each masked token based on representations from unmasked tokens at the span's start and end.

3. **XLNet** (Yang et al., 2019): It's a pre-trained version of the Transformer-XL model, leveraging an autoregressive technique to maximize expected likelihood across all permutations of the input sequence factorization order, enabling the learning of bidirectional contexts.

4. **MultiCo** (Zhu et al., 2020): It addresses the extended context challenge by identifying and composing phrases that span the document. It combines a query-based contextualized phrase selection technique with a sparse self-attention mechanism.

5. **RoBERTa** (Liu et al., 2019): It shares the same model architecture as BERT but varies in tokenization, pre-training data, and training techniques. It adjusts critical hyperparameters, including the removal of the next-sentence pre-training objective and training with significantly larger mini-batches and learning rates. It had been pretrained on five English-language corpora[2].

6. **BioBert** (Lee et al., 2020): It's based on BERT architecture pre-trained on a large biomedical corpus.

7. **ClinicalBert** (Huang et al., 2019): This is also based on BERT architecture, designed to enhance performance in biological and clinical NLP tasks. Basically, it had been tailored for clinical text mining applications.

8. **BioLinkBERT**(Yasunaga et al., 2022b): It is a specialized pre-trained model designed for biomedical NLP tasks, and is notable for its significant contributions to biomedical question answering.

9. **QA-GNN** (Yasunaga et al., 2021): This is an end-to-end question answering model that leverages relevance scoring for graphs and performs joint reasoning over the QA context and graph to harness the representation capabilities of language models and knowledge graph data.

10. **GreaseLM** (Zhang et al., 2021): GreaseLM utilizes multiple layers of modality interaction operations to combine pretrained language model representations with GNN-based information.

11. **DRAGON** (Yasunaga et al., 2022a): It's a self-supervised model, trained on a deep fusion of text and KG (Knowledge Graph) data, enhancing the potential for joint representations and reasoning in downstream NLP tasks.

[2]BookCorpus (Zhu et al., 2015), CC-NEWS, OpenWebText, Stories (Trinh and Le, 2018) and English Wikipedia

F Implementation Details

We employed the RoBERTa-large (24 layers, 1024 hidden units, 16 attention heads, 355M parameters) (Radford et al., 2019) for PLM-encoding and fine-tuned it on MASH-QA and COVID-QA datasets. After thorough experimentation with learning rates ranging from [1e-5, 2e-5, 3e-5, 4e-5], we opted for the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2e-5. This choice was made based on its superior performance in terms of both convergence speed and training stability. The maximum token limit for input was set to 512, encompassing 384 tokens for context and 128 tokens for the query. For BioASQ and MedQA-USMLE dataset, we used BioLinkBERT-Large(Yasunaga et al., 2022b) (24 layers, 512 hidden units, 16 attention heads, 340M parameters) for PLM-encoding and fine-tuning.

For MASH-QA and COVID-QA datasets, we adopted RoBERTa-large with 24 layers, 1024 hidden units, 16 attention heads, and a total of 355 million parameters, as detailed in (Radford et al., 2019). Similarly, for BioASQ and MedQA-USMLE datasets, we adopted BioLinkBERT-Large(Yasunaga et al., 2022b) with 24 layers, 512 hidden units, 16 attention heads, and a total of 340M parameters. We used baselines BERT-large uncased (24-layer, 1024 hidden dimension, 16 attention heads and 336M parameters), SpanBERT (24-layer, 1024 hidden dimension, 16 attention heads and 340M parameters), XLNet (24-layer, 1024 hidden dimension, 16 attention heads and 340M parameters), MultiCo (24-layer, 1024 hidden dimension, 16 attention heads and 340M parameters), BioBERT (24-layer, 1024 hidden dimension, 16 attention heads and 345M parameters), ClinicalBERT (24-layer, 1024 hidden dimension, 16 attention heads and 345M parameters), QA-GNN (24 layers, 1024 hidden units, 16 attention heads, and 360M parameters), and DRAGON (24 layers, 1024 hidden units, 16 attention heads, and 360M parameters).

To fine-tune *MedEx* and all baselines, we utilized the *Adam* optimizer (Kingma and Ba, 2014) with a learning rate selected from the range [1e-5, 2e-5, 3e-5, 4e-5]. The model was trained over various epochs, including 10, 20, 30, 40, and 50, to optimize performance. The input token limit was set to 512 tokens, allocating 384 tokens for context and 128 tokens for the query. Additionally, we utilized a training batch size of 16 with gradient accumulation steps set to 8 to ensure efficient training.

The computational infrastructure used was the A100-PCIE-40GB with CUDA version 11.2. Each training epoch had a duration of approximately 1.5 hours. To ensure robustness, the experiment was repeated ten times, and the final result was determined by computing the average score across these runs.

G Comparision of MedEx

MedEx represents a significant advancement in medical question-answering systems compared to existing models. Unlike traditional QA systems relying solely on pre-trained language models and graph neural networks, MedEx introduces a novel approach integrating knowledge injection and logic fusion cells. This integration allows MedEx to leverage domain-specific knowledge and logical rules, resulting in more accurate and contextually relevant answers. Furthermore, our ablation study reveals that MedEx outperforms baseline models across multiple datasets, including MASH-QA, COVID-QA, BioASQ, and MedQA-USMLE. Specifically, MedEx achieves higher scores in terms of Exact Match (EM), F1-score, and accuracy (Acc) compared to models relying solely on pre-trained language models. The comprehensive model configuration of MedEx, incorporating both knowledge injection and logical rules, demonstrates its efficacy in delivering precise and context-aware responses in the challenging domain of medical question-answering.

Model	MASH-QA		COVID-QA	
	EM	F1	EM	F1
ChatGPT	NA	25.82	NA	37.29
Llama2-7b	NA	22.36	NA	34.34

Table 11: Automatic result comparision for ChatGPT and Llama2 with our proposed model Medex.

Model	Fluency	Adequacy	Medical Entity Relevance
ChatGPT	4.58	4.15	4.15
Llama2-7b	4.12	3.97	3.5
MedEx	4.46	4.48	4.42

Table 12: Human assessment result comparison of ChatGPT and Llama2 with our proposed model MedEX for 200 randomly selected QA pair.

G.1 Comparision with ChatGPT and Llama2-7b

Since ChatGPT [3] and Llama2-7b (Touvron et al., 2023) operate as a generative model, while our proposed model follows an extractive approach, the systems inherently involve different evaluation metrics, making direct comparisons challenging across all metrics. In our current experiments, we focused on comparing results using the F1-score. Additionally, we conducted human evaluations to assess adequacy and fluency. *Adequacy* gauges the acceptability of responses, while *fluency* measures grammatical correctness. The findings indicate that our model excels in terms of the F1-score and adequacy in comparison to ChatGPT and Llama2, as demonstrated in Table-11 and Table-12. Notably, our model's fluency is reflective of the context's fluency, given its extractive nature. Despite ChatGPT exhibiting superior fluency, our model outperforms it in terms of adequacy, particularly in capturing correct answers containing medical entities. This distinction arises from our model's ability to detect text spans from the context, ensuring the inclusion of crucial medical entities in responses. While we refrain from asserting overall superiority over ChatGPT, we emphasize the efficacy of a specialized model, like ours, within its trained domain, highlighting its enhanced capability over a general-purpose model.

H Logical Rules Analysis

To evaluate the influence of individual logical rules within the MedEx framework, we analyzed various model configurations as detailed in Table 13.

When integrating individual rules (PLM-enc+KI+R1 to PLM-enc+KI+R6), the model's performance varied across datasets, with some rules contributing more significantly to improvements in EM, F1, and accuracy. For instance, the rule R2 consistently led to better performance across most datasets.

Combining multiple rules further enhanced the model's performance. The configuration with all rules integrated (PLM-enc+KI+R1+R2+R3+R4+R5+R6) achieved the highest metrics, showcasing the cumulative benefit of incorporating diverse logical rules. This underscores the importance of leveraging multiple logical inferences to improve MedEx's capability to generate accurate responses in medical question-answering tasks.

Model	MASH-QA		COVID-QA		Bio ASQ	Med QA
	EM	F1	EM	F1	Acc	Acc
PLM-enc+KI+R1	36.85	66.86	27.97	58.03	94.4	44.6
PLM-enc+KI+R2	38.51	68.12	30.26	61.24	95.6	47.7
PLM-enc+KI+R3	37.11	67.27	28.23	59.66	93.1	46.1
PLM-enc+KI+R4	38.29	68.82	29.92	59.98	95.3	47.9
PLM-enc+KI+R5	37.89	67.10	28.74	59.41	93.3	46.5
PLM-enc+KI+R6	36.69	66.25	27.66	58.88	94.2	44.8
PLM-enc+KI+R1+R2	38.77	68.83	29.68	60.87	95.7	47.8
PLM-enc+KI+R1+R2+R3	37.34	67.53	28.22	61.01	95.6	47.1
PLM-enc+KI+R1+R2+R3+R4	39.01	68.98	32.14	61.28	96.0	47.9
PLM-enc+KI+R1+R2+R3+R4+R5	39.12	66.05	32.45	61.44	96.3	48.2
PLM-enc+KI+R1+R2+R3+R4+R5+R6	39.52	69.17	32.72	61.98	96.6	48.3

Table 13: Performance metrics of different models for **MedEx**, where 'PLM-enc' represents RoBERTa for MASH-QA and COVID-QA, and BioLinkBERT for BioASQ and MedQA-USMLE.

[3] https://openai.com/chatgpt

I Error Analysis

Error Type	Question	Ideal Answer	Predicted	Paragraph
Ambiguous Answer Spans	What are some signs of processing issues in people with autism spectrum disorder (ASD)?	That's because some researchers estimate that eight out of 10 children with autism have problems processing sensory input. For example, they can't filter out background noise. Other signs of processing issues include: Problems with balance, body position, and oversensitivity to touch and certain types of clothing, such as socks with seams. With autism, social, behavioral, or attention problems can be partly a result of these sensory challenges.	Other signs of processing issues include: Problems with balance, body position, and oversensitivity to touch and certain types of clothing, such as socks with seams. With autism, social, behavioral, or attention problems can be partly a result of these sensory challenges.	A person with autism spectrum disorder (ASD) often has trouble communicating and interacting with others; their interests, activities, and play skills may be limited. Occupational therapy may help them develop these skills at home and in school. Occupational therapists study human growth and development and a person's interaction with the environment through daily activities. They are experts in the social, emotional, and physiological effects of illness and injury. This knowledge helps them promote skills for independent living in people ...(truncated)
Semantic Inadequacies	How can you recover from delusional disorder?	Delusional disorder is typically a chronic condition, but with proper treatment, many find relief. Recovery varies depending on the person, disorder type, and life circumstances. Unfortunately, some with this disorder don't seek help.	It varies, depending on the person, disorder type, and life circumstances, including support and willingness to stick with treatment.	Delusional disorder, previously called paranoid disorder, is a serious mental illness. People with it can't distinguish reality from imagination ...(truncated)
Model Error	How does surgery treat a deviated septum?	Sometimes, rhinoplasty is combined with septoplasty for appearance improvement. Surgery, usually outpatient, takes one to one and a half hours. Internal splints or packing stabilize the septum. Swelling or bruising is normal after septorhinoplasty.	Septoplasty may also be combined with sinus surgery.	Most people have some sort of imbalance in the size of their breathing passages. Estimates indicate that 80% of people have some ...(truncated)

Table 14: Semantic Inadequacies and Model Error

Frequently Asked Questions (FAQ)

*** What is the primary goal of the methodology used in this study?**

The primary goal of the methodology is to develop and evaluate MedEx, a medical question-answering system that integrates knowledge graphs (KGs) and logical rules with pre-trained language models (PLMs). This integration aims to enhance the accuracy and contextual relevance of answers provided by the system in various medical domains.

*** How does MedEx leverage knowledge graphs and logical rules to improve medical question-answering?**

MedEx utilizes knowledge graphs to incorporate structured medical information and logical rules to apply this information effectively. This approach enables the system to understand complex medical queries by combining structured knowledge with reasoning capabilities, thus improving the accuracy and relevance of the answers.

*** What role do pre-trained language models (PLMs) play in MedEx, and how do they interact with knowledge graphs and logical rules?**

In MedEx, pre-trained language models (PLMs) are employed to process and generate natural language text. They are fine-tuned with medical datasets to enhance their ability to handle medical queries. PLMs work alongside knowledge graphs and logical rules to ensure that the responses are not only accurate but also contextually appropriate, leveraging both language understanding and structured knowledge.

*** Why were knowledge graphs and logical rules specifically chosen for MedEx, and how do they complement each other?**

Knowledge graphs and logical rules were chosen for MedEx because they provide complementary strengths. Knowledge graphs offer structured, comprehensive medical information, while logical rules facilitate reasoning and application of this information. Together, they allow MedEx to deliver more precise and contextually relevant answers than systems relying solely on either approach.

*** What are the advantages of MedEx over existing medical question-answering systems?**

MedEx distinguishes itself from existing systems by integrating knowledge graphs, logical rules, and pre-trained language models. This multi-faceted approach enhances the system's ability to generate accurate and contextually relevant answers, addressing the limitations of systems that depend on only one of these components. The integration ensures a more robust handling of medical queries and improves overall answer quality.