Advanced AI applications in Healthcare

Krishna Gadiraju

Doctoral Candidate, Department of Electrical and Computer Engineering, Louisiana State University

Outline

Healthcare LLMs

Patient graph model for Identifying Infectious Hotspots in a city

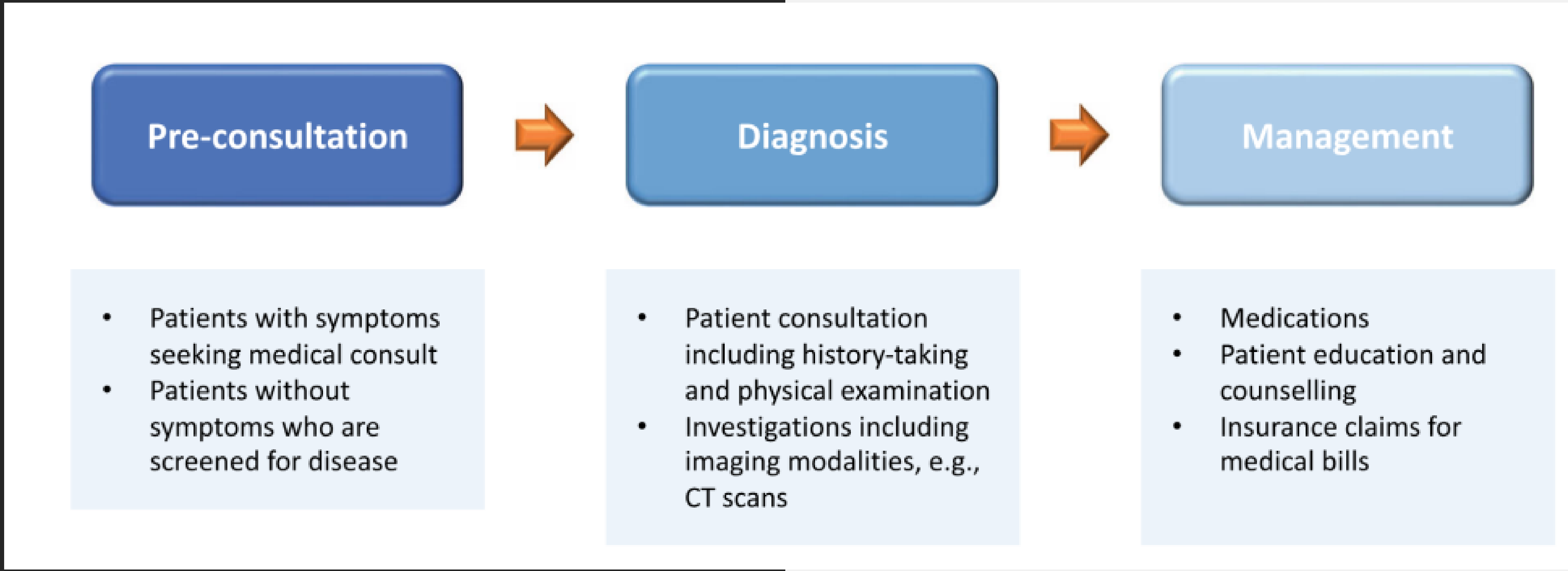
Graph theoretical framework to optimize the performance of SLMs

Introduction

- Large language models (**LLMs**) have garnered **significant attention** and widespread adoption across many fields, including **healthcare** [1].
- Within healthcare, LLMs may be classified into

  LLMs for the **biomedical domain** and

  LLMs for the **clinical domain** based on the corpora used for pre-training.
- In the last 3 years, these **domain-specific LLMs** have demonstrated **exceptional performance** on multiple natural language processing tasks, **surpassing** the performance of **general LLMs** as well [1].
- This not only emphasizes the significance of developing **domain-specific LLMs**, but also increases expectations for their applications in healthcare settings [2-4].
- **LLMs** maybe used widely in **pre-consultation**, **diagnosis**, and **management**, with appropriate **development** and **supervision**. [5-7]
- Additionally, **LLMs** hold tremendous promise in assisting with **medical education**, **medical writing** and other related applications. [8-10]



LLM applications in Patient care

Figure: Potential touch points along a patient's care journey for the application of large language models (LLMs) [1]

Patient-Graph Model for Identifying Infectious Hot-spots in an Urban Environments

Introduction

Infectious diseases pose a serious **threat** to **public health** and **well-being**, especially in **densely populated** urban areas.

Traditional methods of identifying and preventing infectious outbreaks **rely** on **reactive measures**, such as testing, tracing, and isolating [11].

However, **these methods** are often **insufficient**, **costly**, and **time-consuming**, resulting in **delayed responses** and **uncontrolled spread** of infections.

Therefore, there is a need for a **proactive approach** that can leverage **data-driven** techniques to **predict** and **prevent** infectious **hot-spots** in urban environments.

Synthea: Synthetic Patient Data Generation Tool

- ❑ **Synthea** is an open-source tool developed by The **MITRE Corporation** for generating **synthetic patient data**. This data is not based on real individuals, but rather **simulates** realistic **medical histories** and associated **health records**.

What it does:

- ❑ **Generates** extensive **patient data** covering demographics, diagnoses, procedures, medications, allergies, immunizations, social determinants of health, and more.
- ❑ **Offers** various output formats, including **FHIR** (Fast Healthcare Interoperability Resources), **C-CDA** (Continuity of Care Document), and even **DICOM images** for simulated medical scans.
- ❑ **Provides** configurable **population parameters** like city, state, age range, and desired level of detail, allowing customization based on research needs.

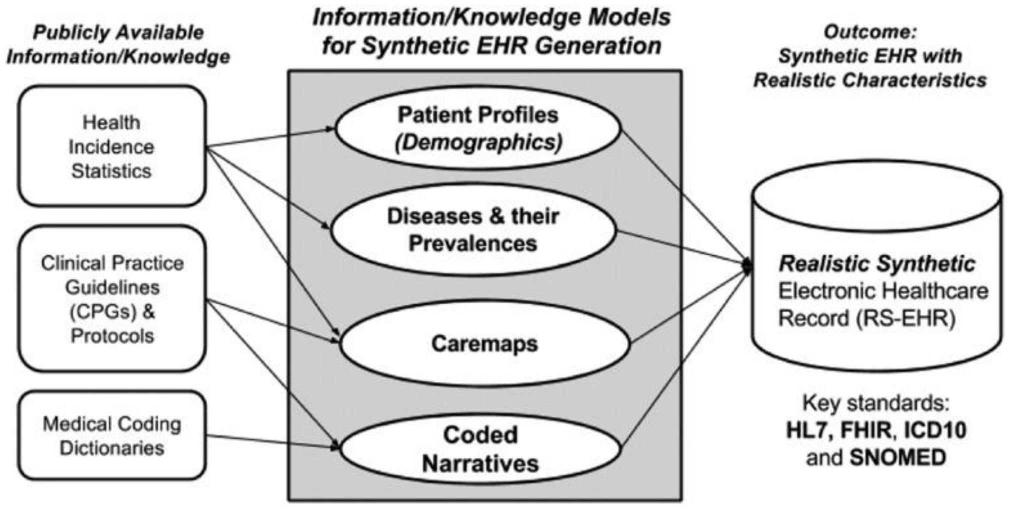
Benefits:

- **Privacy-friendly:** No real patient data is involved, reducing privacy concerns and regulatory hurdles.
- **Large-scale data access:** Enables research using large synthetic populations, overcoming limitations of real-world datasets.
- **Customization:** Tailor data generation to specific research questions by adjusting population characteristics and health trends.
- **Free and open-source:** Accessible to everyone, fostering research collaboration and transparency.

Use cases:

- Testing and development of healthcare IT systems and machine learning models.
- Research on population health, disease modeling, and healthcare interventions.
- Training healthcare professionals in data analysis and clinical decision-making.

Conceptual framework for synthetic EHR generation [12]

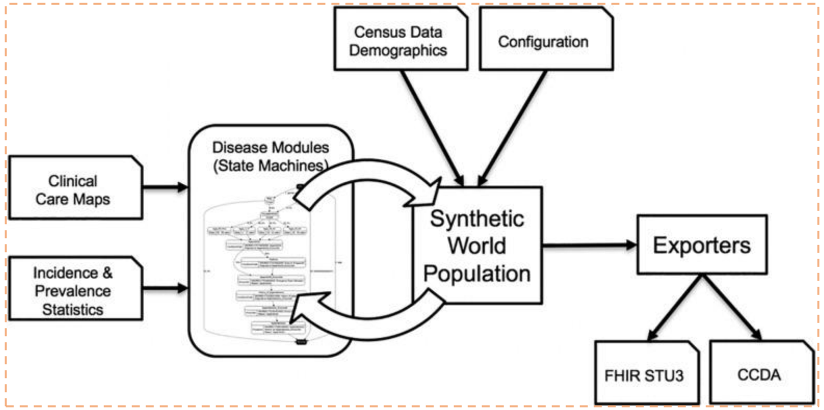


- **Public Data Approach:**
  - **Leverages** publicly available **health statistics**, avoiding need for real EHR access.
  - **Privacy focused:** uses aggregate data, clinical guidelines, and medical coding dictionaries.
- **Realistic Synthetic EHRs:**
  - **Care maps** guide **patient journey** based on **clinician** input and **clinical guidelines**.
  - Regional data, clinician expertise, and guidelines improve realism.
  - Resulting synthetic **EHRs (RS-EHRs)** suitable for many secondary uses (e.g., population studies).
- **Synthea and GRiSER Methods:**
  - **Synthea: top-down approach** generating **skeletal EHRs** with **FHIR** standard codes.
  - **GRiSER: bottom-up approach** generating **detailed entries** for specific health problems.
  - Both methods contribute to a future comprehensive **RS-EHR** generation system.

Synthea Software Architecture: Example of a patient data

```
Golda945 O'Hara16
=================
Race:           White
Ethnicity:      Non-Hispanic
Gender:         F
Age:            45
Birth Date:     1971-10-04
Marital Status: M
--------------------------------------------------------------------------------
ALLERGIES: N/A
--------------------------------------------------------------------------------
MEDICATIONS:
2015-09-14 [CURRENT] : 3 ML liraglutide 6 MG/ML Pen Injector
2014-11-23 [STOPPED] : canagliflozin 100 MG Oral Tablet
2014-11-23 [STOPPED] : 3 ML liraglutide 6 MG/ML Pen Injector
2014-11-23 [CURRENT] : 24 HR Metformin hydrochloride 500 MG Extended Release Oral Tablet
2010-11-30 [STOPPED] : Amoxicillin 250 MG / Clavulanate 125 MG [Augmentin] for Viral sinusitis (disorder)
2007-07-05 [STOPPED] : Amoxicillin 250 MG / Clavulanate 125 MG [Augmentin] for Sinusitis (disorder)
--------------------------------------------------------------------------------
CONDITIONS:
2014-11-23 -            : Diabetes
2014-01-10 - 2014-02-05 : Viral sinusitis (disorder)
2010-11-22 - 2010-12-10 : Viral sinusitis (disorder)
2007-06-28 - 2007-07-22 : Sinusitis (disorder)
1998-04-22 -            : Prediabetes
1990-08-29 -            : Hypertension
--------------------------------------------------------------------------------
CARE PLANS:
1998-04-22 [CURRENT] : Diabetes self management plan
             Reason: Diabetes
             Activity: Diabetic diet
             Activity: Exercise therapy
--------------------------------------------------------------------------------
```

```
OBSERVATIONS:
2016-11-14 : Body Height                                       157.5 cm
2016-11-14 : Body Weight                                       104.3 kg
2016-11-14 : Body Mass Index                                    42.0 kg/m2
2016-11-14 : Systolic Blood Pressure                           198.0 mmHg
2016-11-14 : Diastolic Blood Pressure                          107.0 mmHg
2016-11-14 : Hemoglobin A1c/Hemoglobin.total in Blood 8.3 %
2016-11-14 : Glucose                                           133.0 mg/dL
2016-11-14 : Urea Nitrogen                                      13.0 mg/dL
2016-11-14 : Creatinine                                          1.0 mg/dL
2016-11-14 : Calcium                                             9.4 mg/dL
2016-11-14 : Sodium                                            136.0 mmol/L
2016-11-14 : Potassium                                           4.5 mmol/L
2016-11-14 : Chloride                                          102.0 mmol/L
2016-11-14 : Carbon Dioxide                                     27.0 mmol/L
2016-11-14 : Basic Metabolic Panel
2016-11-14 : Total Cholesterol                                 243.0 mg/dL
2016-11-14 : Triglycerides                                     340.0 mg/dL
2016-11-14 : Low Density Lipoprotein Cholesterol               145.0 mg/dL
2016-11-14 : High Density Lipoprotein Cholesterol               30.0 mg/dL
2016-11-14 : Lipid Panel
2016-11-14 : Microalbumin Creatine Ratio                         2.0 mg/g
2016-11-14 : Estimated Glomerular Filtration Rate               >60 mL/min/{1.73_m2}
--------------------------------------------------------------------------------
PROCEDURES:
2014-11-23 : Documentation of current medications
2011-01-02 : Documentation of current medications
2007-11-19 : Documentation of current medications
--------------------------------------------------------------------------------
ENCOUNTERS:
2016-11-14 : Outpatient Encounter
2015-09-14 : Outpatient Encounter
2015-03-23 : Outpatient Encounter
2014-11-23 : Outpatient Encounter
2014-01-15 : Encounter for Viral sinusitis (disorder)
2011-01-02 : Outpatient Encounter
2010-11-30 : Encounter for Viral sinusitis (disorder)
2007-11-19 : Outpatient Encounter
2007-07-05 : Encounter for Sinusitis (disorder)
```



Generic Module Framework:

- **Encodes** models of **disease progression** and **treatment** as state machines in JSON.
- Open and documented for **easy extension** and understanding.

Data Inputs:

- **Clinical care maps** and statistics **guide** patient journeys.
- **Census data** and **configuration options** populate the synthetic world.

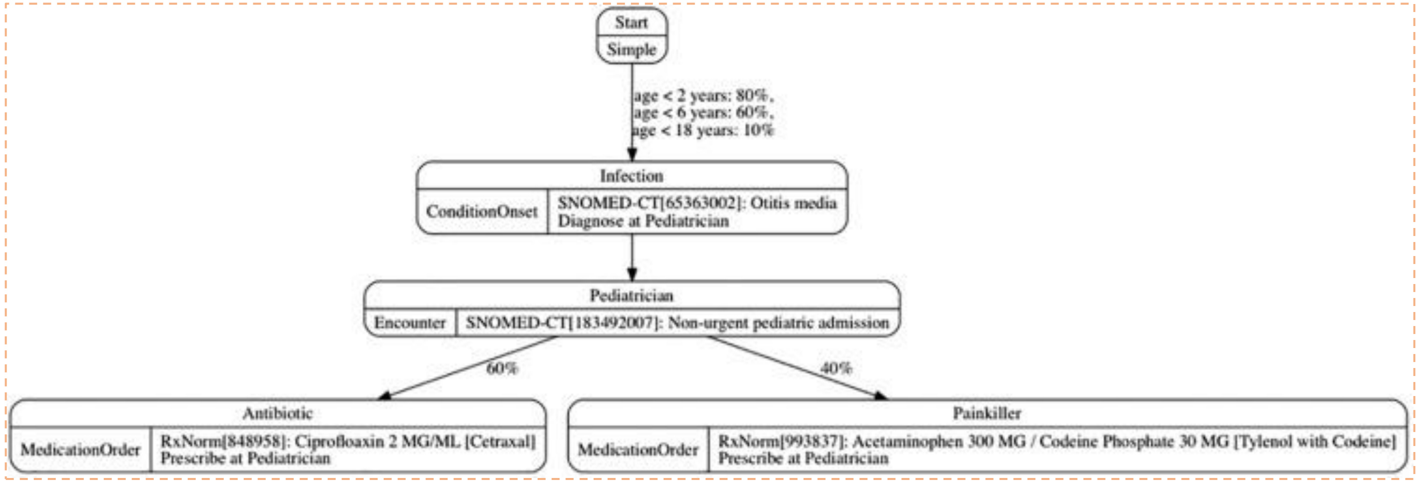
Processing:

- Modules **calculate state transitions** for each person in the synthetic world at each timestep (**default 7 days**).
- Events happening within a timestep are handled promptly.

Outputs:

- Transitions **trigger** various clinical events (**condition onsets, encounters, prescriptions, etc.**).

Example of childhood ear infections [12]




```
"Infection": {
 "type": "ConditionOnset",
 "target_encounter": "Pediatrician",
 "codes": [ { "system": "SNOMED-CT", "code": "65363002", "display": "Otitis media"} ],
 "direct_transition": "Pediatrician"
},
"Pediatrician": {
 "type": "Encounter",
 "encounter_class": "ambulatory",
 "codes": [ { "system": "SNOMED-CT", "code": "183492007",
        "display": "Non-urgent pediatric admission"} ],
 "distributed_transition": [
  { "distribution": 0.6, "transition": "Antibiotic" },
  { "distribution": 0.4, "transition": "Painkiller"} ]
}
```

Functionality:
- ❑ Simulates **ear infections** in children based on age.

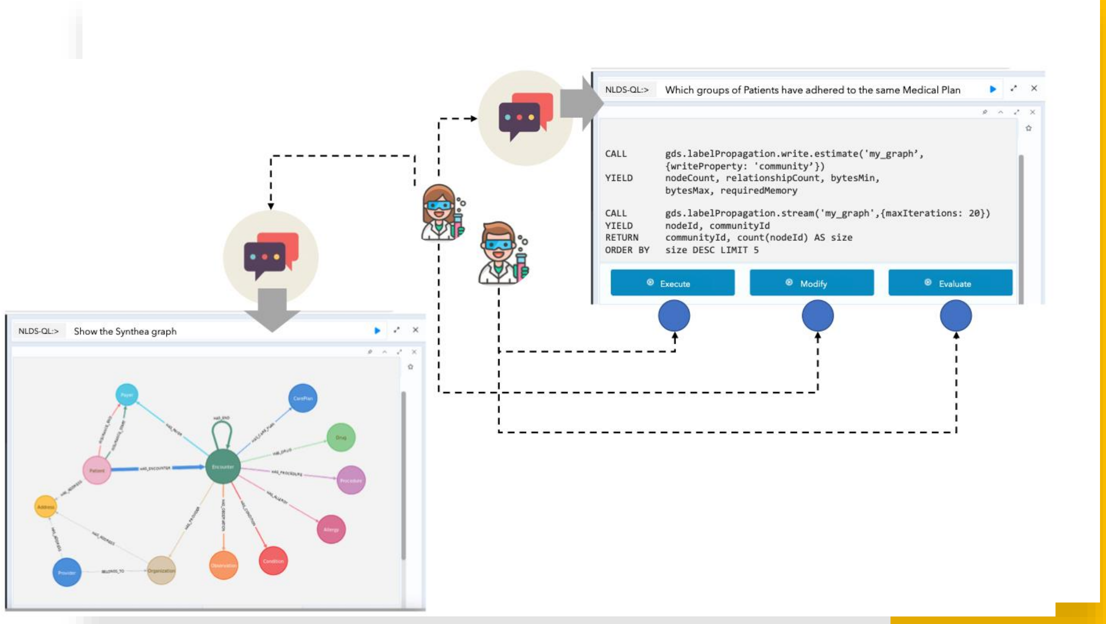
States:
- ❑ **Infection:** Child has an ear infection (**duration specified**).
- ❑ **Pediatrician:** Child visits a pediatrician.

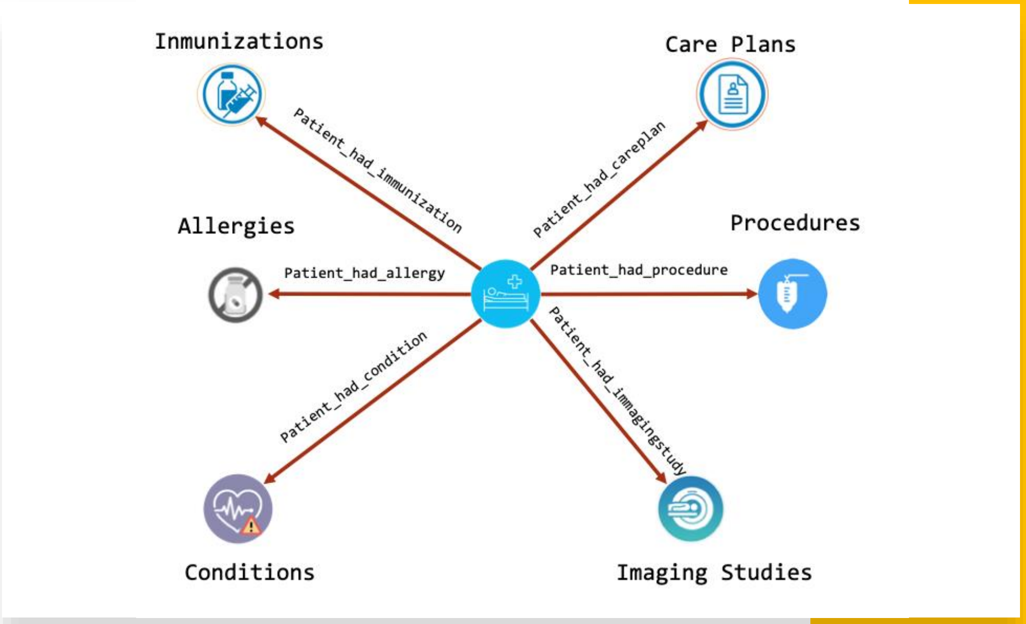
Transitions:
- ❑ Healthy child transitions to infection with **age-dependent probability**.
- ❑ Infected child transitions to **pediatrician** for **diagnosis**.
- ❑ **Pediatrician** visit leads to **treatment**: antibiotic or painkiller.

Listing 2: Details state definitions in JSON, including:
- ❑ State names and types.
- ❑ Attributes (e.g., medical codes for diagnosis).
- ❑ Transitions to other states with conditions and probabilities.

Application of Synthea in patient specific graph problem:

- **Start:** User interacts with **NLDS-QL** interface.
- **Ask Question:** User asks a question about the Synthea patient graph.
- **Generate Queries: NLDS-QL** generates one or more potential queries based on the user's question.
- **Refine & Execute:** User selects, **refines, and executes** one or more queries.
- **Evaluate:** User **evaluates** the results of the query execution with a satisfaction rating.
- **Explore More:** User continues **exploring** the graph by asking new questions or refining previous ones [13]





Data Statistics for our study:

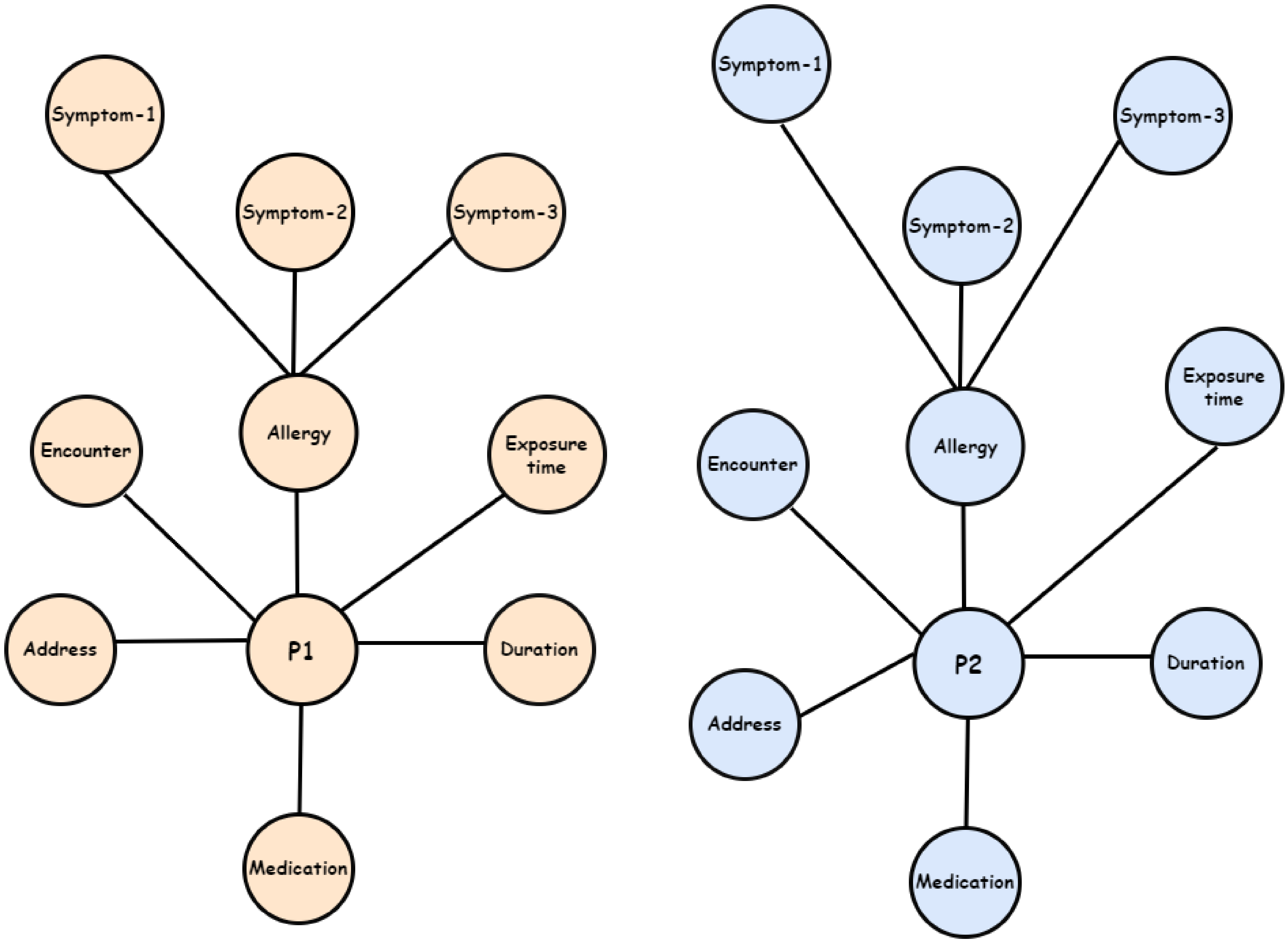
- **Selection Criteria:** A **subset** of patients is chosen from the **original graph** based on specific criteria, **reducing** the number of vertices from **800,000** to approximately **1000**.

- **Relationship Consideration:** The **original Synthea graph** likely includes **edges** representing various **relationships** or connections **between patients**, such as shared medical encounters, family relationships, or social interactions. When selecting a **subset of patients**, some of these **relationships** may be **preserved**, while **others** may be **omitted** based on the **simulation criteria** impacting the resulting graph's structure and reducing the number of edges from approximately **2,000,000** to around **2500**.

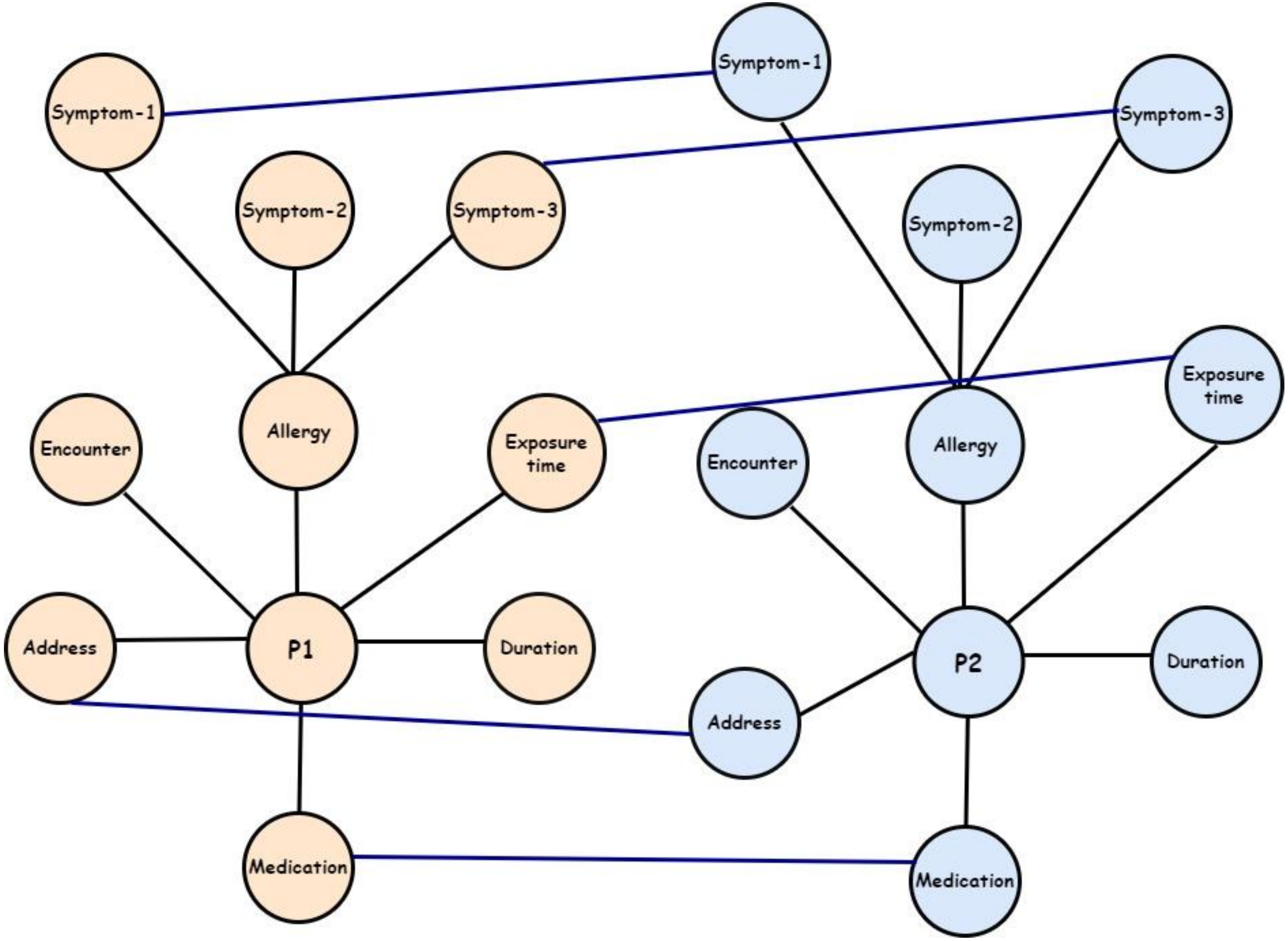
- **Scaling Effect:** Applying a **linear scaling** approach provides an estimate, with the number of vertices for the **1000 patients** being approximately **1000 times smaller** than the **original**, and the number of **edges** being roughly **1000 times smaller** as well.

- **Graph Connectivity:** Changes in the number of vertices may affect the graph's overall **connectivity** and **edge density**, influencing its structure and the number of connections between patients.

Data Structure:



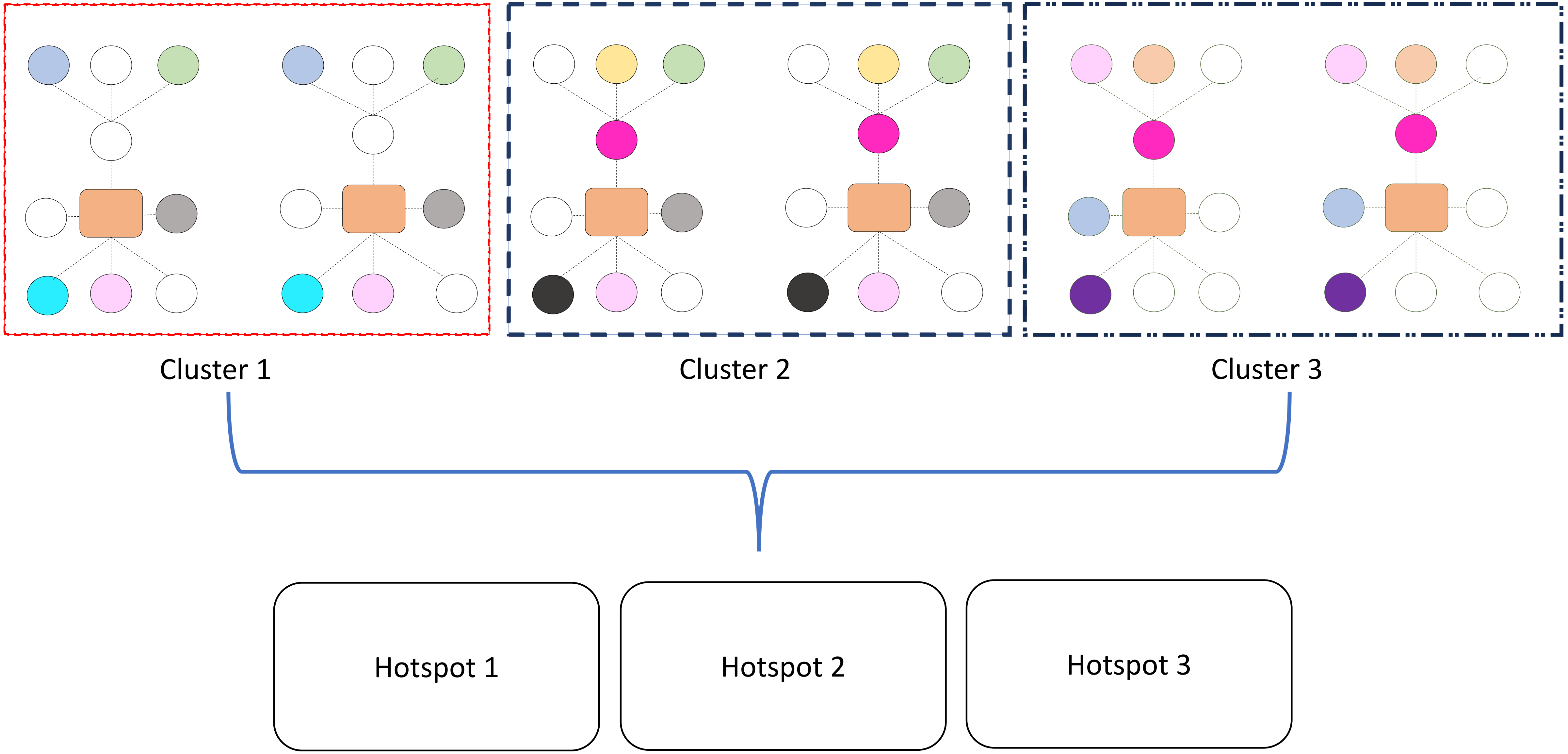
Cohort Attributes



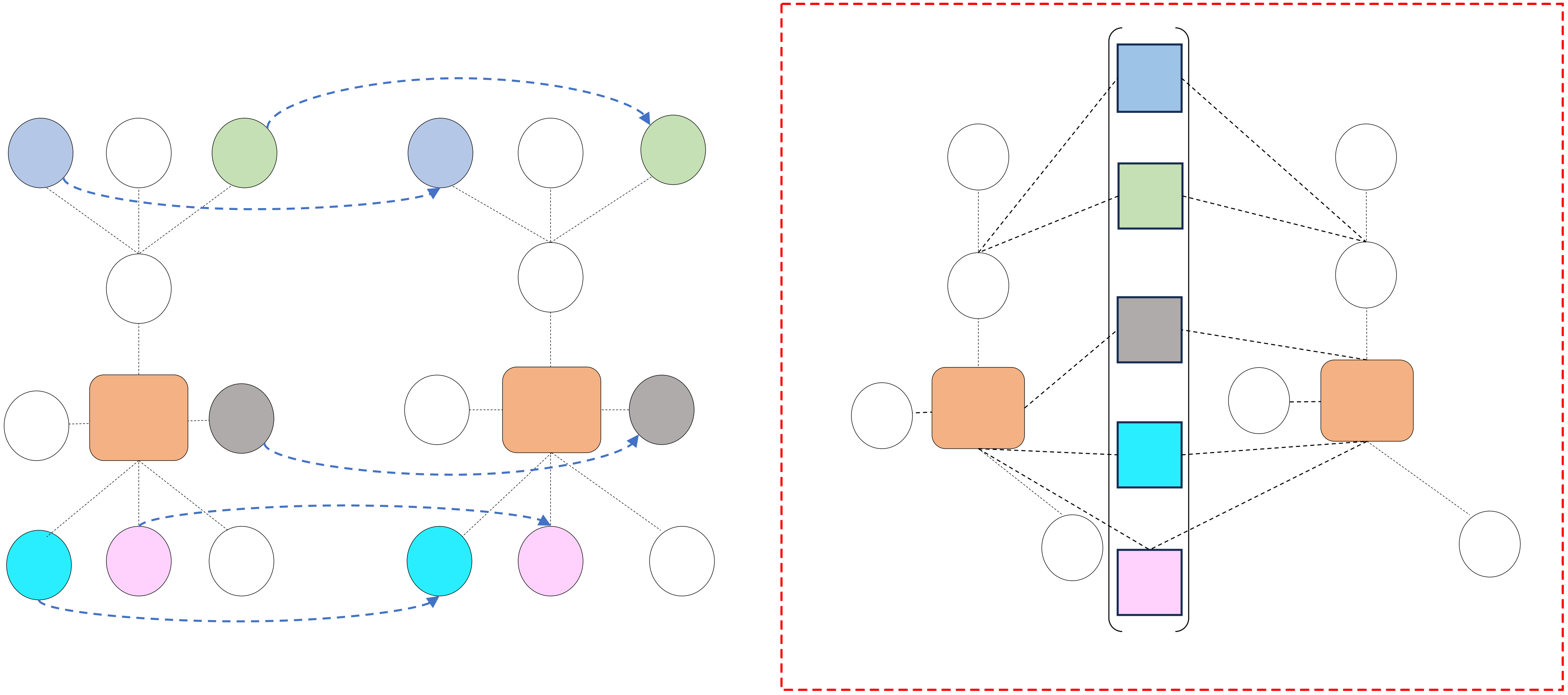
Problem Statement

Given a patient graph, **identify cohorts** with similar disease thresholds (symptoms) such that **infectious hot-spots** can be identified prematurely and **risk of infection spread** in given urban setting can be **mitigated**.

Approach1: Graph Clustering and Hotspot Identification (Unpublished work)



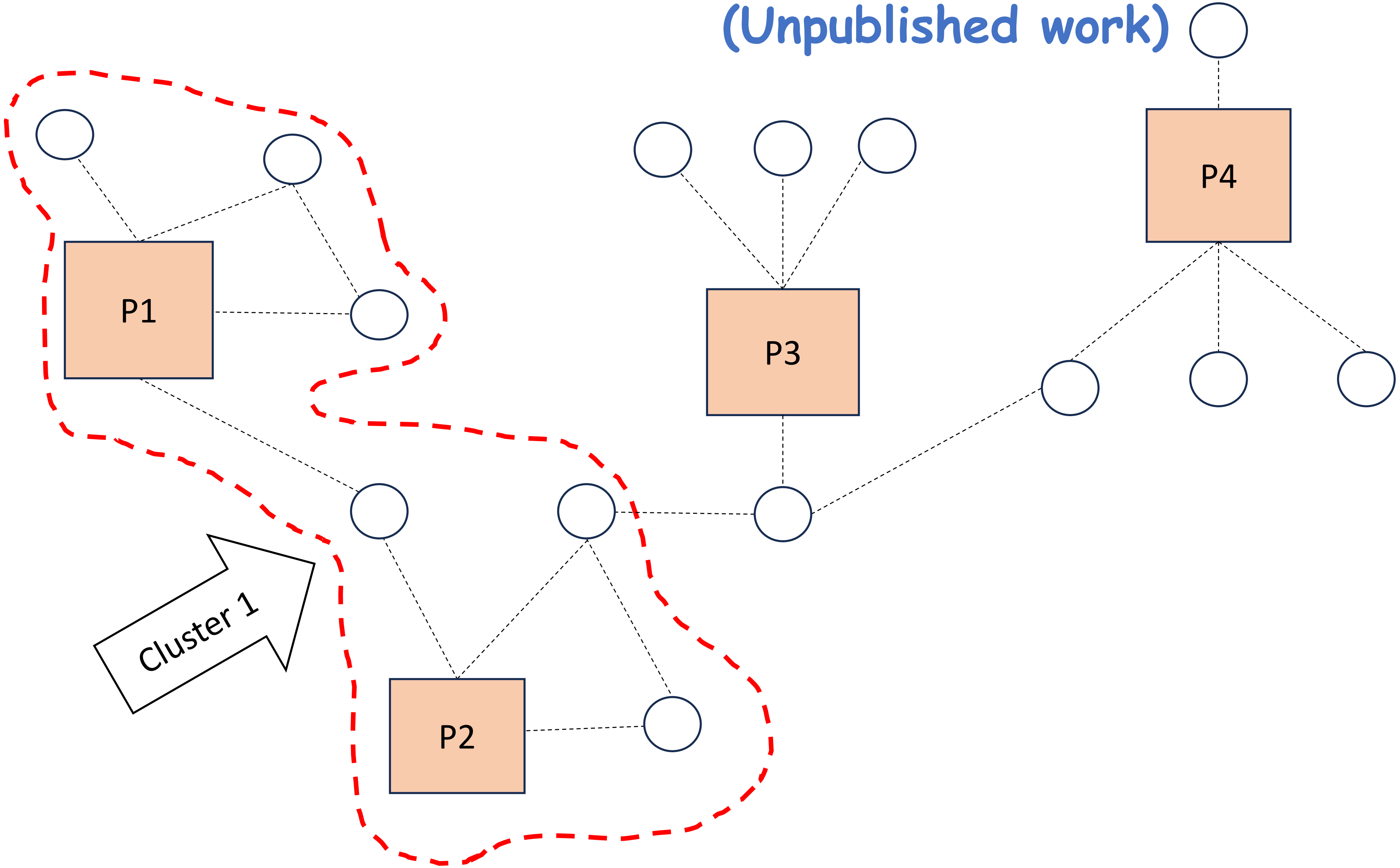
Approach 2: Super node clustering and Hotspot Identification using Edge Contraction (Unpublished work)



Cluster 1

Approach 3: Graph Topological Clustering and Hotspot Identification

(Unpublished work)



Development of LLMs in Healthcare

- Although **LLMs** have shown **impressive performance** across a range of **NLP tasks**, their **efficacy** in specialized tasks is **limited** [19].
- Moreover, there are **significant differences** between **general corpora** and **professional corpora**, which further **hinder** the **ability of LLMs** to perform well in **biomedical** or **clinical** settings [20].
- To **improve** domain-specific **performance** by addressing these weaknesses, **domain-specialized LLMs** have been developed.
  - BioMistral [14]
  - ClinicalBERT [3]
  - BioBERT [3]
  - GatorTron [15]
  - Med-PaLM[16] and Med-PaLM 2 [17]
  - ChatDoctor[18]

Key takeaways [1]:

Rather than training domain-specific models from the ground up, further research may seek to **fine-tune** or **prompt-tune** these **general LLMs** to optimize performance in **domain-specific** clinical settings.

Using larger **open-source base models** and newer **interactive LLMs** could further **improve** the **capabilities** of decentralized researchers around the world, who could then **fine-tune LLMs** to optimize performance for **clinical tasks**.

Through **fine-tuning**, domain-specific LLMs may be produced to serve narrowly defined, well-specified tasks—**minimizing error** and **maximizing clinical utility**.

Whether developed from scratch or fine-tuned using existing models, **LLM applications** will become more sophisticated and begin to **impact patients** and **practitioners** at scale.

Graph-Theoretical Framework to Optimize the Performance of SLMs

LLMs -> SLMs

- In recent years, large language models (**LLMs**) have been widely applied in artificial intelligence (AI) driven **prompt engineering** such as **question-answering** and **text summarization** functionalities [21].
- There is a growing interest in small language models (**SLMs**) for **resource-constrained application-specific data mining** [22]
- Small Language Models (**SLMs**) involves **much fewer** parameters than **LLMs**, offer **advantages** in terms of **reduced carbon emission**, **short training times**, and **low computational complexity** [23]. **SLMs** provide **quick** inference and responses and thus are **preferred often** in practice.
- When an **SLM** is **fine-tuned** for a specific domain or task, it can provide **accurate** and **contextually-relevant** answers to user queries [24].
- The capabilities of **SLMs** can be significantly **improved** by **incorporating knowledge** from **LLMs**. **Pre-trained LLMs**, which have learned **high-fidelity** information from **big data**, can **transfer** valuable **digested information** to SLMs through "**fine-tuning**".

Drawbacks of Conventional Finetuning

To **improve** the **response quality** provided by **SLMs**, the **conventional model-training** procedures often rely on **enormous training data**.

However, the obvious **drawbacks** can be found as follows:

- ❑ **Training** on **big datasets** demands **substantial computational power** often beyond the capability of any resource-constrained system.
- ❑ The **tremendous computing resource** required by training big data implies **high operational costs**.
- ❑ The incurred **extraordinary computational burden** turns out to be **huge carbon emission** against the globally demanded green computing agenda.
- ❑ Training **large datasets** usually requires a very long time, thus **hindering** the timeliness of any **model deployment** for **time-sensitive applications**.

Fine-tuning under resource-constrained scenarios

There exist **three primary approaches** for **fine-tuning SLMs** subject to **computational resource constraints**, namely:

- ❑ **Transfer learning:** adopting the **pre-trained LLMs or SLMs** and adapting them to specific tasks subject to **minimum additional training** [25–27],
- ❑ **Knowledge distillation: transferring knowledge** from a large teacher model (a **pretrained LLM**) to a small student model (an **SLM**) by preserving the essential information efficiently [28– 30], and
- ❑ **Prompt Engineering: crafting** specialized users' prompts to guide the **responses** of an **SLM** and enabling **targeted-performance** improvements [31–33].

Unfortunately, these three approaches suffer from **domain mismatch**, **high training complexity**, and **limited application-specific knowledge** [34].

Training Data Reduction – Literature attempts

A **possible strategy** to **combat** the aforementioned **drawbacks** of the existing approaches is to **extract** the **subset** of the **tremendous training data**, which encompasses the essential characteristics of the entire dataset. This idea is called **training data reduction (TDR)**.

- ❑The **graph-based heuristic method** has been proposed to **partition** a **big dataset** and select one or a **few subsets** for scalable supervised training to **reduce** the **computation time** and **enhance** the overall **accuracy** across various **classification** algorithms [35].
- ❑The **TDR scheme** has been applied to **fine-tune** multilingual **BERT** models for spoken language understanding [36].
- ❑A **data-efficient** learning algorithm was introduced, which **compressed** large vision language datasets into a **small**, **high-quality subset** by selecting the **representative samples** and **generating** the **new captions** [37].

Training Data Reduction – Literature attempts

A strategy for **reducing large datasets** for machine learning model training, which involved the **discretization of data** through **multidimensional histograms** and the **reduction** of the **sample size** within **each bin** [38].

A minimum **data augmentation framework** for few-shot question-answering was proposed using a graph algorithm and an unsupervised question generation mechanism to **synthesize** the **most informative** training samples from the **raw text** [39].

However, the **above** stated **TDR schemes** cannot be directly applied to **personalized prompt datasets** as the **domain-relevance information** among prompts **cannot be captured** to provide correct responses.

Problem Statement

Given a **prompt dataset**, consisting of individual prompts, what is the **optimum subset** of prompts, one can select to **train** an **SLM** so as to **reduce** the **training time** and **simultaneously** achieving a **satisfactory data-mining performance** not much worse than that resulting from a prominent LLM.

Fine-tuning optimization

The primary contributions made so far can be summarized as follows:

- ❑ A **graph-theoretical approach** to extract the **semantic, contextual, and domain-relevance relationships** among users' prompts is developed. This approach can be applied to any large prompt datasets of multiple domains.
- ❑The **conventional clique-finding paradigm** is extended for **TDR** and the proposed scheme is evaluated for the **GPT-2 model** (an SLM) involving **117 million parameters** trained by **three artificial prompt datasets** crafted for domain experts such as **clinicians**, **bio-informatics scientists**, **AI/ML engineers**, and **data scientists**.
- ❑The **time-complexity** analysis is studied for the proposed **TDR scheme**.
- ❑The **conventional paradigm** trained by at least **70%** of the training data is **compared** with the **proposed TDR approach**. The **proposed approach** shows the **on-par** and **better performance** than the **conventional method** in terms of **BERTScore** [40].

Preliminaries

Definition 1: Prompt Semantic Measure Ψ(A, B): The prompt semantic measure Ψ(A, B) is defined by the degree of **similarity or relatedness** in meaning between two prompts $\mathbf{P_A}$ and $\mathbf{P_B}$ based on their respective **semantic embeddings** [41].

$$\Psi(A,B) \stackrel{\text{def}}{=} \frac{\langle \mathfrak{E}^{S}_{A}, \mathfrak{E}^{S}_{B} \rangle}{\|\mathfrak{E}^{S}_{A}\| \, \|\mathfrak{E}^{S}_{B}\|},$$

Where,

"$\langle\ \rangle$" - denotes the inner-product

"$\|\ \|$" - denotes the vector norm

$\mathfrak{E}^{S}_{A}$ and $\mathfrak{E}^{S}_{B}$ - Represents semantic word embeddings of Prompts $\mathbf{P_A}$ and $\mathbf{P_B}$

Preliminaries

Definition 2: Prompt Contextual Measure Δ(A, B): The prompt contextual measure Δ(A, B)) is defined by the **degree of similarity or relatedness** between two prompts $\mathbf{P_A}$ and $\mathbf{P_B}$ based on their **contextual embeddings** [42].

Where,

$$\Delta(A, B) \stackrel{\text{def}}{=} \frac{\langle \mathfrak{E}^{\mathfrak{c}}_A, \mathfrak{E}^{\mathfrak{c}}_B \rangle}{\|\mathfrak{E}^{\mathfrak{c}}_A\| \, \|\mathfrak{E}^{\mathfrak{c}}_B\|}.$$

"$\langle\ \rangle$" - denotes the inner-product

"$\|\ \|$" - denotes the vector norm

$\mathfrak{E}^{\mathfrak{c}}_A$ and $\mathfrak{E}^{\mathfrak{c}}_B$ - Represents contextual embeddings of Prompts $\mathbf{P_A}$ and $\mathbf{P_B}$

Preliminaries

Definition 3: (Prompt Graph $G_P(\eta, \rho)$): A **prompt dataset** can be transformed into the corresponding **prompt graph**, say **$G_P(\eta, \rho)$ = (V, $E_{\eta,\rho}$)**, where the vertex set **V** consists of all prompts in **P**, i.e., **V = P**,

while there exists an edge between **P_i** and **P_j** **(P_i , $P_j \in$ V)**

If:

- the respective **prompt semantic measure $\Psi(i, j) \geq \eta$**,
- the respective **prompt contextual measure $\Delta(i, j) \geq \rho$**, and
- **P_i** and **P_j** belong to the **same domain** or subject area, i.e., **$P_i \leftrightarrow P_j$** .

Note that **η** and **ρ** here are called the **semantic** and **contextual** relevance thresholds, respectively.

Preliminaries

Definition 4: (Maximal Clique and Maximum Clique): A **maximal clique**, of $\mathbf{G_P(\eta, \rho)}$ represents a **clique** from which **no further extension of node(s)** is possible to form a **bigger clique** containing **extra node(s)**. Furthermore, a **maximum clique** is one of the maximal cliques of $\mathbf{G_P(\eta, \rho)}$, which has the **largest number of vertices (graph order)**.

Definition 5: (Union of Maximum Cliques' Vertices (UMCV) $\mathbf{V_u(P:\Theta)}$) Given a **prompt dataset P**, one can form **various prompt graphs** as subject to **Q** pairs of $\boldsymbol{\eta_q}$ **and** $\boldsymbol{\rho_q}$ for **q=1, 2, . . ., Q** according to **Definition 3**. Furthermore, one can find the respective **maximum cliques** for **q=1, 4 2, . . ., Q** according to **Definition 4**. The **union of maximum cliques' vertices UMCV $\mathbf{V_u(P)}$** is thus defined by

$$\mathcal{V}_{\mathrm{u}}(\mathbb{P}:\Theta) \stackrel{\mathrm{def}}{=} \bigcup_{q=1}^{Q} \mathcal{V}'_{\mathrm{ma}}\left(\mathcal{G}_{\mathbb{P}}(\eta_q, \rho_q)\right)$$

Where,

$$\Theta \stackrel{\mathrm{def}}{=} \{(\eta_q, \rho_q), q = 1, 2, \ldots, Q\}$$

Proposed Framework

Proposed graph-theoretical **framework** for prompt dataset reduction, includes **three** key **mechanisms**:

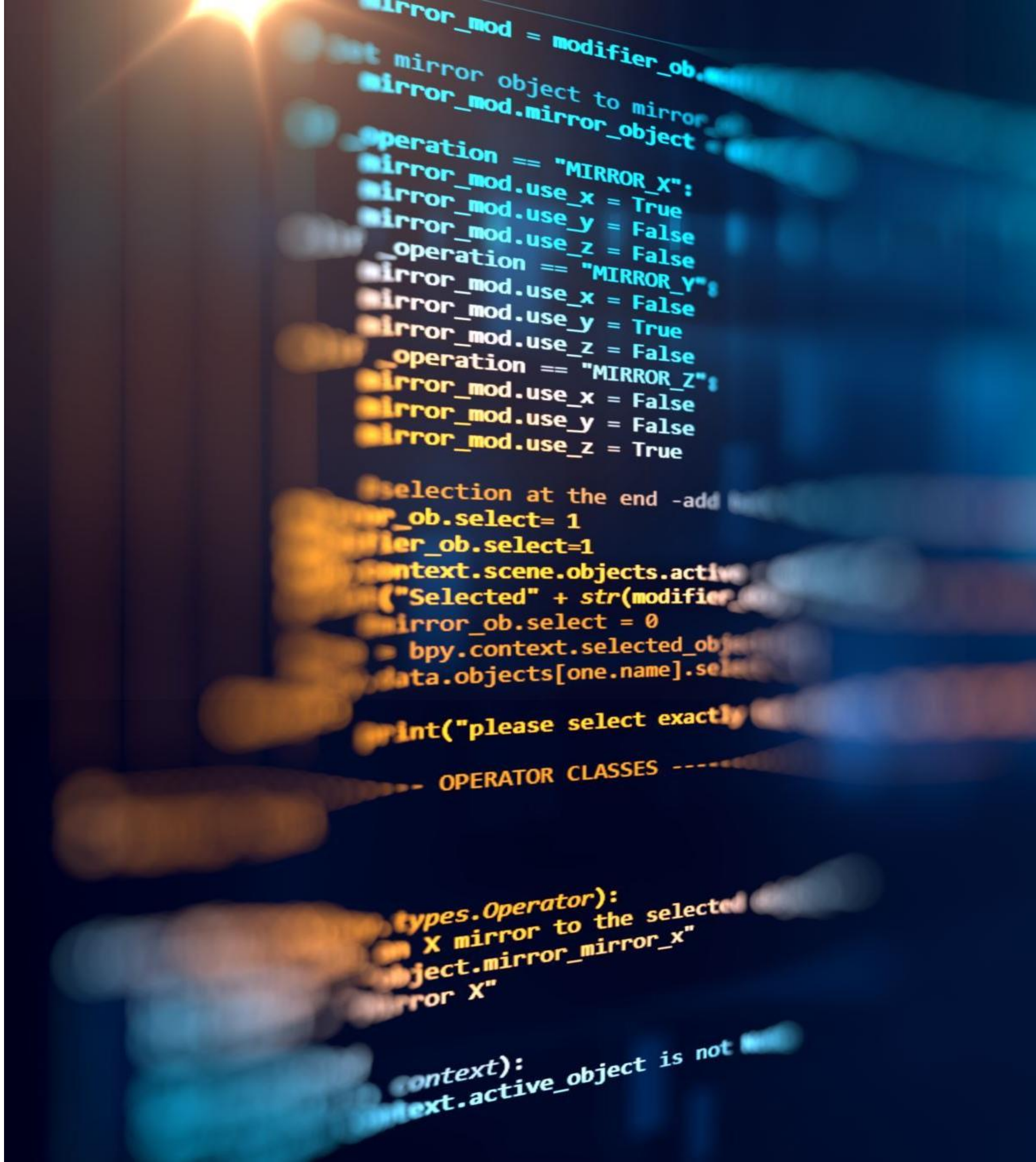
- ❑**Relevance thresholds determination**,
- ❑**Prompt graph construction**, and
- ❑**Graph-theoretical TDR scheme**.

Proposed Framework - Relevance Thresholds Determination

For a given **prompt dataset P**:

Step:1 Obtain **prompt semantic measure** (according to **Definition 1**) and **prompt contextual measure** (according to **Definition 2**) for **all pairs of prompts**

Step:2 Then **compute** the **mean, first quartile (Q1), second quartile (Q2), and third quartile (Q3)** values of **prompt semantic measure** and **prompt contextual measure** for the entire **prompt dataset P**. These values form the **set of relevance thresholds Θ**.



Proposed Framework- Prompt Graph Construction

For a given **prompt dataset P** and the **set of relevance thresholds Θ** :

Step:1 Treat **each prompt** in **P** as a vertex **V**

Step:2 Form **edge set** such that $\mathbf{E_{\eta,\rho}}$ **for any two** distinct vertices (**prompts**) in **V**, the corresponding **edge weight** is set to be **1** if **prompt semantic measure Ψ(i, j)≥η** and **prompt contextual measure Δ(i, j)≥ρ** and prompts $\mathbf{P}_i$ and $\mathbf{P}_j$ belong to the **same domain or subject area**, i.e., $\mathbf{P}_i \leftrightarrow \mathbf{P}_j$

Likewise, we obtain **four prompt graphs**.

Proposed Framework- Graph- Theoretical TDR Scheme

For each **prompt graph**,

Step:1 Obtain **maximum clique** using **Bron-Kerbosch algorithm** [43] or the **approximate maximum-clique finding algorithm** (for a **large graph order**) [44].

Step:2 Obtain **UMCV** $\mathbf{V_u(P:\Theta)}$ (according to **Definition 5**).

Step:3 The **optimal set of prompts** are nothing but $\mathbf{V_u(P:\Theta)}$.

Simulation – Data Acquisition

- **Proposed TDR approach** is **evaluated** on **fine-tuning GPT-2** [45] language model involving **117 million** parameters with three artificial prompt datasets.
- ChatGPT was used to generate **three batches** of artificial **question-answering** prompt data (approximately **uniformly distributed** user-persona-specific **prompts** over **four** different categories) of size **100**, **500,** and **1000** prompts crafted for four domain experts: **clinicians, bio-informatics scientists, AI/ML engineers, and data scientists**.

Simulation – Application of proposed TDR approach

MISTRAL 7B model [46] was used to **infer** which prompts $\mathbf{P}_i$ and $\mathbf{P}_j$ in a prompt dataset **P** belong to the **same domain or subject area**, i.e., $\mathbf{P}_i \leftrightarrow \mathbf{P}_j$

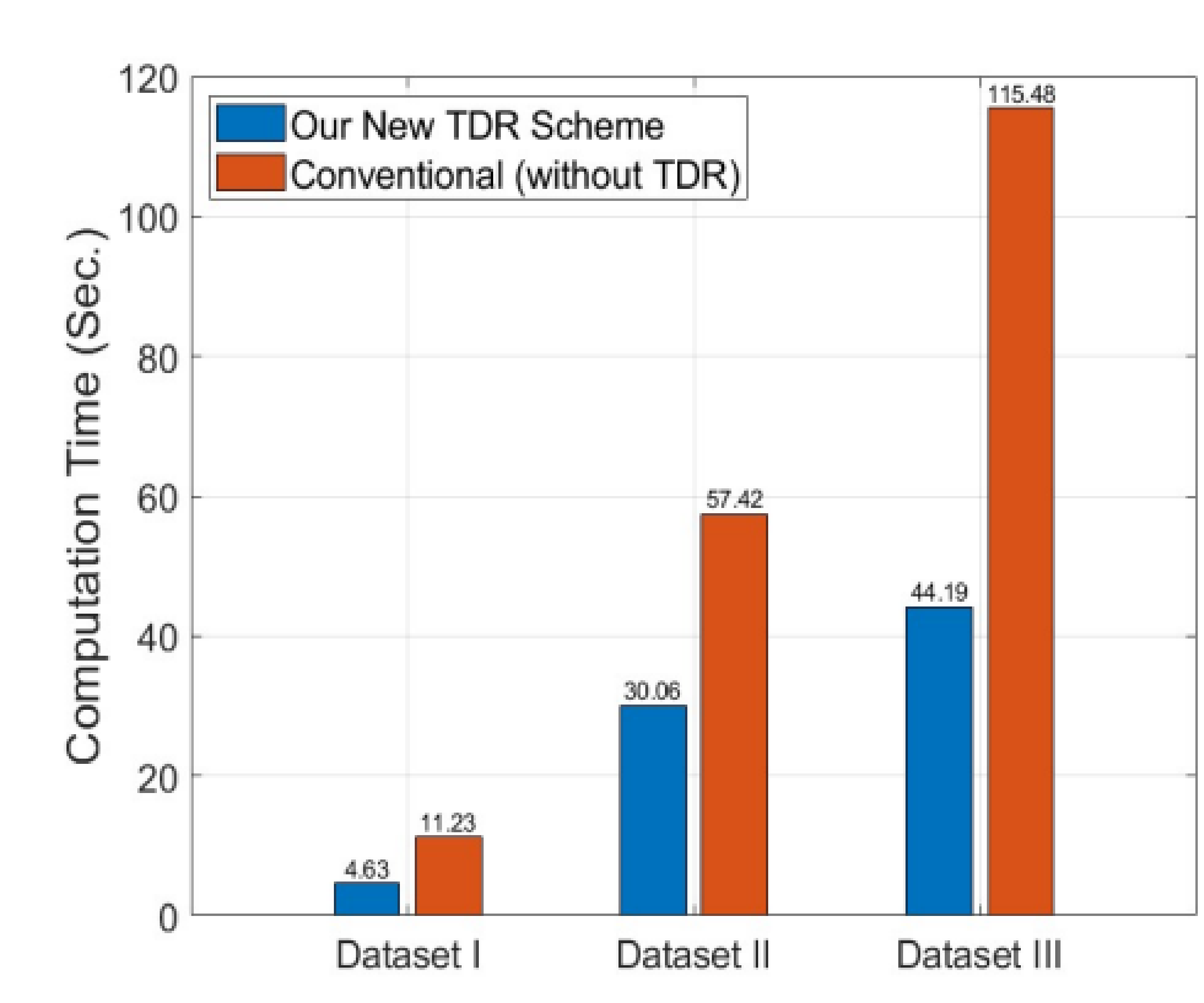
Then the **same model** was used to generate the "**ground truth**" for the **BERTScore** evaluation of the **question-answering** task.

Then, the **set of relevance thresholds Θ** is obtained using key mechanism (**Relevance thresholds determination**).

Using the above information **four prompt graphs** are obtained by implementing key mechanism (**Prompt graph construction**).

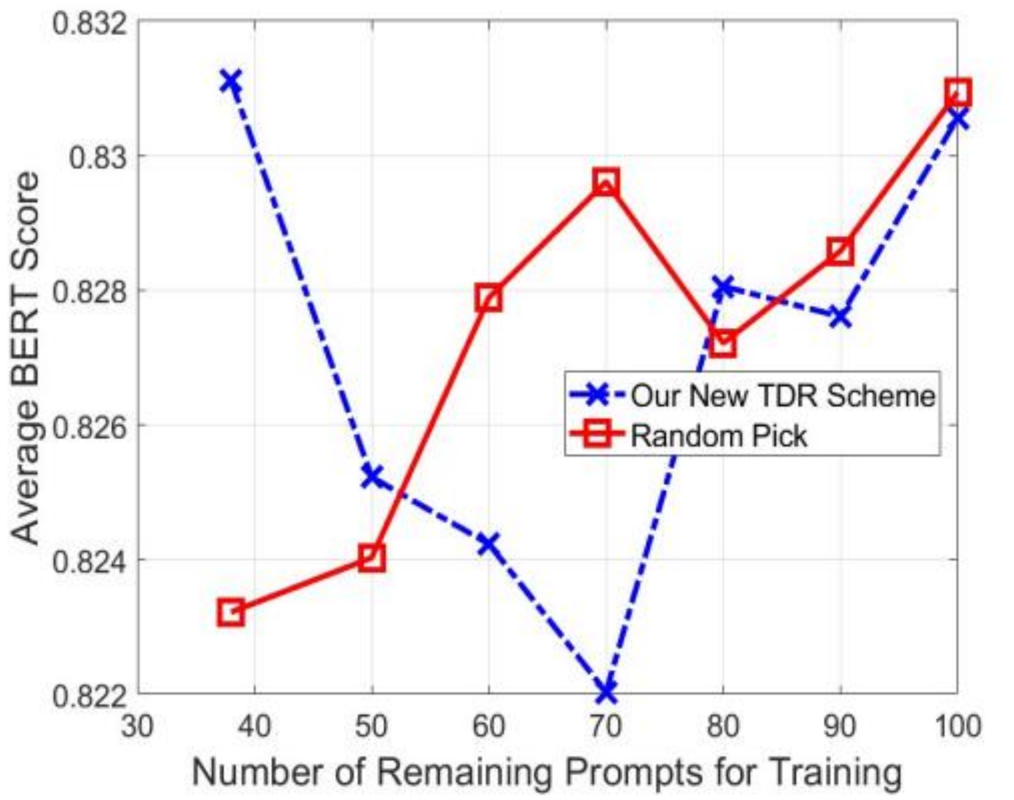
Finally, **maximum cliques** are computed and the **optimal set of prompts UMCV Vu(P:Θ)** is obtained using key mechanism (**Graph-theoretical TDR scheme**).

Results and Discussion- Actual Run-time comparison



Fine-Tuning Optimization of Small Language Models: A Novel Graph-Theoretical Approach for Efficient Prompt Engineering (submitted)

Results and Discussion- BERT Score Performance Evaluation



	Conv. Random-Pick Method	Our Proposed New Approach
Dataset I ($\lvert\mathbb{V}^{opt}\rvert$=38)	0.8159	0.8236
Dataset II ($\lvert\mathbb{V}^{opt}\rvert$=161)	0.8238	0.8262
Dataset III ($\lvert\mathbb{V}^{opt}\rvert$=357)	0.8278	0.8287

Fine-Tuning Optimization of Small Language Models: A Novel Graph-Theoretical Approach for Efficient Prompt Engineering (submitted)

Future work

- ❑Designing a **dynamic edge contraction TDR scheme** to further **reduce** the **run-time** of the proposed framework.
- ❑Develop a **Graph topological compression TDR scheme** using Topological GNNs [] to facilitate the **reduction** of **large-scale corpus knowledge graphs**.
- ❑Explore **computational-geometry** approaches such as **Voronoi partition, Delaunay triangulation** to **pre-partition** the **large-scale graphs** and design a novel graph topological compression **TDR mechanisms.**

References

[1]Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, *2*(4), 255-263.

[2] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: apre-trained biomedical language representation model forbiomedical text mining. Bioinformatics. 2020;36(4):1234–40

[3] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D,Naumann T, et al. Publicly available clinical BERT embed-dings. 2019

[4] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language modelfor scientific text. Proceedings of the 2019 conference on empiricalmethods in natural language processing and the 9th InternationalJoint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019

[5] He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing diseaseknowledge into BERT for health question answering, medicalinference and disease name recognition. 2020

[6] Li C, Zhang Y, Weng Y, Wang B, Li Z. Natural languageprocessing applications for Computer-Aided diagnosis inoncology. Diagnostics. 2023;13(2):286

[7] Omoregbe NAI, Ndaman IO, Misra S, Abayomi-Alli OO,Damaševičius R. Text Messaging-Based medical diagnosisusing natural language processing and fuzzy logic. J HealthcEng. 2020;2020(4):1–14

References

[8] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L,Elepaño C, et al. Performance of ChatGPT on USMLE:potential for AI-assisted medical education using largelanguage models. PLOS Digital Health. 2023;2(2):e0000198.

[9] Gilson A, Safranek CW, Huang T, Socrates V, Chi L,Taylor RA, et al. How does ChatGPT perform on the UnitedStates medical licensing examination? The implications oflarge language models for medical education and knowledgeassessment. JMIR Med Educ. 2023;9:e45312.

[10] Kitamura FC. ChatGPT is shaping the future of medicalwriting but still requires human judgment. Radiology.2023;307(2):230171

11] Houlihan, C. F., & Whitworth, J. A. (2019). Outbreak science: recent progress in the detection and response to outbreaks of infectious diseases. Clinical Medicine, 19(2), 140.

[12] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238,

[13] Vargas-Solar, G., Dao, K., & Alves, M. H. F. (2022). NLDS-QL: From natural language data science questions to queries on graphs: analysing patients conditions & treatments. arXiv preprint arXiv:2208.10415.

[14] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P. A., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv preprint arXiv:2402.10373.

References

[15] YangX,ChenA,PourNejatianN,ShinHC,SmithKE,Parisien C, et al. A large language model for electronichealth records. npj digital Medicine. 2022;5(1):1–9.

[16] Med-PaLM. Med-PaLM [Internet]. Available from:https://sites.research.google/med-palm/27.

[17] Matias Y. Our latest health AI research updates. Google[Internet]. Available from:https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/28.

[18] Li Y, Li Z, Zhang K, Dan R, Zhang Y. ChatDoctor: a medicalchat model fine-tuned on LLaMA model using medicaldomain knowledge. 2023

[19] Thirunavukarasu A, Hassan R, Mahmood S, Sanghera R,Barzangi K, El Mukashfi M, et al. Trialling a large languagemodel (ChatGPT) with Applied Knowledge Test questions:what are the opportunities and limitations of artificialintelligence chatbots in primary care? (Preprint). 2023

[20] Lei L, Liu D. A new medical academic word list: a corpus-based study with enhanced methodology. J English Acad Purp. 2016;22:42–53

[21] Z. Shao, Z. Yu, M. Wang, and J. Yu, “Prompting large language models with answer heuristics for knowledge-based visual question answering,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023

[22] V. Haswani and P. Mohankumar, “Methods to Optimize Wav2Vec with Language Model for Automatic Speech Recognition in ResourceConstrained Environment,” in Proceedings of the 19th International Conference on Natural Language Processing (ICON), 2022.

[23] T. Schick and H. Schutze, “It’s not just size that matters: Small language ¨ models are also few-shot learners,” arXiv preprint arXiv:2009.07118, 2020.

[24] J. Oza and H. Yadav, “Enhancing Question Prediction with flan t5-a context-aware language model approach,” Authorea Preprints, 2023.

References

[25] R. Wang, J. Du, and T. Gao, “Quantum Transfer Learning Using the Large-Scale Unsupervised Pre-Trained Model Wavlm-Large for Synthetic Speech Detection,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

[26] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mosner, L. Burget, and ˇ J. Cernock ˇ y, “Parameter-efficient transfer learning of pre-trained Trans- ` former models for speaker verification using adapters,” in ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

[27] Y. Li, A. Mehrish, R. Bhardwaj, N. Majumder, B. Cheng, S. Zhao, A. Zadeh, R. Mihalcea, and S. Poria, “Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

[28] L. Zhang, R. Dong, H.-S. Tai, and K. Ma, “Pointdistiller: Structured knowledge distillation towards efficient and compact 3D detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023.

[29] Z. Li, P. Xu, X. Chang, L. Yang, Y. Zhang, L. Yao, and X. Chen, “When object detection meets knowledge distillation: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[30] J. Gou, L. Sun, B. Yu, S. Wan, and D. Tao, “Hierarchical multi-attention transfer for knowledge distillation,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 20, no. 2, pp. 1–20, 2023.

[31] M. Wang, M. Wang, X. Xu, L. Yang, D. Cai, and M. Yin, “Unleashing ChatGPT’s Power: A Case Study on Optimizing Information Retrieval in Flipped Classrooms via Prompt Engineering,” IEEE Transactions on Learning Technologies, 2023.

[32] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush, “Interactive and visual prompt engineering for ad-hoc task adaptation with large language models,” IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 1, pp. 1146–1156, 2022.

References

[33] C. Clemmer, J. Ding, and Y. Feng, “PreciseDebias: An Automatic Prompt Engineering Approach for Generative AI To Mitigate Image Demographic Biases,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024

[34] J. Rao, X. Meng, L. Ding, S. Qi, X. Liu, M. Zhang, and D. Tao, “Parameter-efficient and student-friendly knowledge distillation,” IEEE Transactions on Multimedia, 2023.

[35] S. Yadav and M. Bode, “A graphical heuristic for reduction and partitioning of large datasets for scalable supervised training,” Journal of Big Data, vol. 6, no. 1, p. 96, 2019.

[36] A. Bansal, A. Shenoy, K. C. Pappu, K. Rottmann, and A. Dwarakanath, “Training data reduction for multilingual Spoken Language Understanding systems,” in Proceedings of the 18th International Conference on Natural Language Processing (ICON), December 2021.

[37] A. J. Wang, K. Q. Lin, D. J. Zhang, S. W. Lei, and M. Z. Shou, “Too large; Data Reduction for Vision-Language Pre-Training,” arXiv preprint arXiv:2305.20087, 2023.

[38] J. Wibbeke, P. Teimourzadeh Baboli, and S. Rohjans, “Optimal data reduction of training data in machine learning-based modelling: a multidimensional bin packing approach,” Energies, vol. 15, no. 9, p. 3092, 2022.

[39] X. Chen, J.-Y. Jiang, W.-C. Chang, C.-J. Hsieh, H.-F. Yu, and W. Wang, “MinPrompt: Graph-based Minimal Prompt Data Augmentation for Few-shot Question Answering,” arXiv preprint arXiv:2310.05007, 2023.

[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” CoRR, vol. abs/1904.09675, 2019.

References

[41] L. K. Senel, I. Utlu, V. Yucesoy, A. Koc, and T. Cukur, “Semantic ¨ structure and interpretability of word embeddings,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1769–1779, 2018.

[42] S. Arora, A. May, J. Zhang, and C. Re, “Contextual embeddings: When ´ are they worth it?” arXiv preprint arXiv:2005.09117, 2020.

[43] E. A. Akkoyunlu, “The enumeration of maximal cliques of large graphs,” SIAM Journal on Computing, vol. 2, no. 1, pp. 1–6, March 1973.

[44] R. Boppana and M. M. Halldorsson, “Approximating maximum indepen- ´ dent sets by excluding subgraphs,” BIT Numerical Mathematics, vol. 32, no. 2, pp. 180–196, 1992

[45] P. Budzianowski and I. Vulic, “Hello, it’s gpt-2–how can i help you? to- ´ wards the use of pre-trained language models for task-oriented dialogue systems,” arXiv preprint arXiv:1907.05774, 2019

[46] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier et al., “Mistral 7b,” arXiv preprint arXiv:2310.06825, 2023

Under the esteemed guidance of my PI:

Dr. Hsiao-Chun Wu

Professor,

Division of Electrical and Computer Engineering

School of Electrical Engineering

and Computer Science

Louisiana State University

Thanks for your valuable support !!!

Dr. Manali Singha

IGM Bioinformatics Scientist

Nationwide Children's Hospital

Columbus OH

Thanks for research infrastructure support !!!

Hao-yu-Tsai

Ph.D Student

Department of Electrical Engineering

National Tsing Hua University (NTHU)

Hsinchu 30013, Taiwan

Dr. Scott Huang

Professor

Department of Electrical Engineering and Institute of Communication Engineering

National Tsing Hua University (NTHU)

Hsinchu 30013, Taiwan

Thanks to co-authors !!!

Dr. Guannan Liu

Assistant Professor

Department of Applied Data Science

San Jose State University, San Jose, USA

Dr. Shih Yu Chang

Assistant Professor

Department of Applied Data Science

San Jose State University, San Jose, USA

Dr. Yiyan Wu

Principal Research Scientist

Communications Research Centre

Ottawa, Canada

Q&A session

Please post your Questions in the chat


Thank
You