The Review Of
DIABETIC
STUDIES
OPEN ACCESS

Enhancing Pharmaceutical Research Through Ontology-Driven Semantic Search And Real-World Evidence Integration

[1]Sandeep R Diddi, [2]Dr. Rajesh Sharma

[1]Alliance College of Engineering and Design, Alliance University, Bangalore, India
[1]Email: dsandeepPHD724@ced.alliance.edu.in
[2]Alliance College of Engineering and Design, Alliance University, Bangalore, India
[2]Email: rajeshsharma.r@alliance.edu.in

Abstract

The pharmaceutical research is highly affected by the lack of data sources that are fragmented, inconsistent metadata, and inefficient search systems based on keywords. To overcome these shortcomings, this research presents PharmaSeek+, an Ontology-Driven Semantic Search Framework aimed at transforming the way researchers' access, interpret and use real-world pharmaceutical evidence. PharmaSeek+ is designed using a systematic four step process: (1) metadata harmonization to harmonize the data of various clinical trials, drug databases, and pharmacovigilance systems by schema mapping and metadata alignment. (2) ontology development is applied by a pharmaceutical ontology that combines existing vocabularies (MeSH, SNOMED CT) with custom classes representing drug efficacy, interactions, molecular profiles, and adverse effects. (3) in semantic annotation via deep learning (DL) models, the biomedical documents are enhanced with the context-aware deep learning models such as BioBERT and contextual Natural Language Processing (NLP) in accordance with the ontology. (4) transformer-based sequence-to-query model handles the user queries which are converted into semantic queries and are SPARQL-based to support reasoning over the implicit knowledge and produce very relevant and contextual results. The PharmaSeek+ achieved 89% in search precision and 87% in recall compared to traditional search engines, and these results are confirmed using real-world pharmaceutical corpora. The findings validate the fact that PharmaSeek+ is a powerful tool that enhances the process of discovering and integrating important drug-related information. This smart structure allows quicker, more knowledgeable decision-making in pharmaceutical research and development, eventually closing the gap between fragmented biomedical information and researcher intention.

Keywords: - Ontology-Based Modeling, Pharmaceutical Data Integration, Real-World Evidence, SPARQL Querying, Natural Language Processing, Biomedical Ontologies.

1. Introduction

The pharmaceutical research is being confronted with a growing amount and complexity of data which are occasioned by a multiplicity of areas including clinical trials, genomics, drug interactions, and real-world clinical practice [1]. Nonetheless, such data sources are frequently very disparate, located in independent silos, and controlled by heterogeneous metadata requirements [2]. Such heterogeneity poses serious difficulties to data discovery, integration, and reuse, thus making the drug development lifecycle slow and scientific innovation difficult [3],[4]. The most problematic aspect of the modern pharmaceutical data systems is the use of the traditional, keyword-based search systems [5]. Such techniques do not usually capture contextual meaning, synonymy or hierarchical relationships between terms [6],[7]. Consequently, researchers fail to get pertinent information or access huge amounts of data that are not relevant, leading to inefficiencies in evidence collection, hypothesis formulation, and decision making [8].

In addition, the growing significance of real-world evidence (RWE) data obtained in real-life clinical practice, including electronic health records (EHRs), claims data, and patient registries, requires more advanced integration and interpretation approaches [9],[10]. Unless smart tools are available to semantically interconnect clinical and experimental data, useful patterns and relations are hidden in the noise [11]. Such increasing complexity explains why there should be a paradigm shift to intelligent, ontology-based search systems that are capable of identifying semantic relationships and adapting to different biomedical vocabularies [13],[14]. Ontologies structured representation of knowledge offers a means of integrating heterogeneous data sources by coordinating them on a common conceptualization, thus improving interoperability, data discoverability, and inferencing [15].

To overcome these difficulties, PharmaSeek+ is presented as a new semantic search system that are used in the pharmaceutical research field and which has an ontology-based structure. It uses domain-specific ontologies, natural language processing and knowledge graph technologies to offer context-aware searching, allowing the user to query biomedical datasets in a more natural and understandable fashion. PharmaSeek+ is able to combine structured ontological reasoning with real-world evidence sources and enable researchers to discover previously unknown relationships, simplify drug discovery processes, and enable precision medicine efforts. This paper discusses the design, functionality and the influence of PharmaSeek+ in promoting pharmaceutical research based on semantic technologies and integration of real-world data.

2. Literature Review

Stănescu & Oprea [16] discussed the use of ontologies and Semantic Web Technologies (SWT) in contemporary data management based on 10,037 scholarly articles published in 20192024. It employs bibliometric analysis, NLP, LDA, and BERT clustering to determine such main themes as ontology-driven systems, biomedical data integration, and ethical implications. The research identifies the lack of scalability, dynamic update, and semantic interoperability. It has a topic coherence score of 0.75 and perplexity of 48, and it has three significant research clusters. These results provide practical implications on how to enhance semantic search, data accessibility and automated knowledge discovery. Yao et al [17] introduced OntoPath, an ontology-aware hierarchical attention model to personalize prescription recommendation to chronic disease care in 2023. It uses the history of longitudinal diagnosis, hierarchical medical ontologies and side information to forecast the best course of treatment. OntoPath improves the profiling of patients and drug relevance modeling with domain knowledge and pre-training. It performs better than state-of-the-art baselines on a large depression cohort of more than 37,000 patients. The findings validate the accuracy, interpretability and clinical applicability of OntoPath.

Bakshi et al [18] in 2021 suggested semantic conflicts in data integration as a result of schema, terminology and domain interpretation differences, which lowers data interoperability and quality. In this paper we suggest ontology-based frameworks as a solution to such conflicts by formalizing domain knowledge into structured, machine-readable forms. The methodology is proved by case studies in the healthcare and e-commerce fields. The paper assesses the advantages and the drawbacks of these methods.

In 2022, Thirumahal et al [19] discussed the problem of semantically diverse biomedical data integration with an automatic ontology-based framework. The model works in three stages: local ontology generation, generation of unified global schema and querying heterogeneous sources to retrieve semantically aligned data. It is implemented on the patient records, chest X-rays, and COVID-19 questionnaires in SQL, MongoDB, and Excel databases. The system determines patients with moderate/high risk of severe COVID-19. This improves data unification, information retrieval and clinical decision making in healthcare.

Li et al [20] proposed an Ontology-Driven Medication Query (ODMQ) optimization scheme to enhance the accuracy of medication information retrieval in EHRs in 2025. Through the OMOP Common Data Model, ODMQ broadens queries semantically on drug names, codes, and generics, which increases the accuracy and completeness. When applied to the real-world COVID-19 EHR data, it demonstrates better performance and less manual work. It has a friendly interface that facilitates query execution and viewing of patient history. The relevance of expanded search terms to clinical needs is confirmed with the help of manual review.

Fareedi et al [21] suggested a hybrid Ontology-Based Design Science Research Engineering (ODSRE) approach to improve data integration and semantic interoperability in healthcare in 2025. It uses the Ontop virtual paradigm to propose a Federated Virtual Knowledge Graph (FVKG) system to effectively and in real-time access various data sources. FVKG reduces data migration, has a low latency, and semantic coherence across systems. The model employs ontology-based data access (OBDA) and schema mapping in order to match semantic artifacts. This provides a scalable federated information systems (FIS) solution in patient-centric healthcare settings.

In 2029, Kamdar et al [22] proposed the problem of combining fragmented biomedical data and highlighted the opportunities of Semantic Web technologies and Linked Data principles. It deals with the Life Sciences Linked Open Data (LSLOD) cloud as a means of harmonizing data in pharmacology, cancer research, and infectious diseases. Although promising, LSLOD has adoption obstacles that are related to technical and usability challenges. The paper suggests the possible ways to improve LSLOD usability and accessibility. Finally, LSLOD was able to facilitate AI-powered biomedical research and enhance clinical outcomes.

In 2025, Abraham et al [23] introduced a digital-twin framework of precision oncology that combines machine learning and semantic models to personalize treatment. It identifies subtypes of brain cancer based on clinical and molecular data, in the form of ontologies with diagnostic rules. The potential therapies are prioritized in preclinical models that are semantically aligned with patient subtypes. The method allows cross-domain reasoning to assist in personalized treatment decisions.

In 2023, Lazarova et al [24] proposed AD-DPC, an ontology to facilitate interdisciplinary cooperation in the study of Alzheimer disease. It organizes knowledge in six major conceptual categories, such as pathology, diagnosis, and clinical findings. AD-DPC is marked with definitions, synonyms, and resources to assist users of different levels. Usability testing indicates that non-medical users were able to use AD-DPC to learn and share concepts related to Alzheimer with good results. The ontology promotes better knowledge sharing and interdisciplinary participation in the complex medical research.

Kawas et al [25] (2023) proposed a sophisticated mechanism of ontology integration that was applied to medical text, which is based on the shortcomings of the current methods, including semantic precision and adaptability. It integrates ontological, lexical, logical and machine learning methods to align disparate sources of data through a shared upper ontology and transformation rules. A supervised model foresees concept mappings between various ontologies. The method improves the semantic integration, which promotes data-driven decision-making in clinical and research environments.

2.1. Problem Statement

Although there is increasing use of Electronic Health Records (EHRs) and biomedical data systems, the healthcare sector continues to experience serious problems in integrating, querying, and interpreting semantically heterogeneous and fragmented data. The conventional solutions have problems with schema inconsistencies, terminology inconsistencies, and domain interpretation inconsistencies, which result in decreased data interoperability, ineffective information retrieval, and inferior clinical decision-making. Such constraints impede the scalability, accuracy and applicability in real-time of data-driven solutions in healthcare. Ontology-based frameworks and semantic web technologies have become indispensable to integrate heterogeneous data, improve semantic search, and to provide personalized care as well as intelligent decision support systems in various medical fields. The summary of some of the existing works is depicted in Table 1.

Table 1: Summary on Several Existing Works

Author, Year [Citation]	**Technique Used**	**Application in Healthcare**	**Advantages**	**Limitations**

Stănescu & Oprea, 2025 [16]	Bibliometric analysis, NLP, LDA, BERT clustering	Biomedical data integration, ontology analysis	Identifies research trends, ethical gaps, and ontology-driven systems; coherence score: 0.75	Scalability, dynamic update challenges, and semantic interoperability issues
Yao et al., 2023 [17]	Ontology-aware hierarchical attention model (OntoPath), longitudinal diagnosis, pre-training	Personalized prescription for chronic disease	Improves patient profiling and prescription accuracy; interpretable and clinically useful	Condition was specific; scalability across diseases not evaluated
Bakshi et al., 2021 [18]	Ontology-based data integration, schema alignment, case study validation	Healthcare and e-commerce	Resolves semantic conflicts; structured, machine-readable knowledge representation	Complexity in ontology modeling and domain adaptation
Thirumahal et al., 2022 [19]	Automatic ontology generation, global schema synthesis, heterogeneous query system	COVID-19 patient risk prediction using multi-source biomedical data	Enhances data unification and retrieval; supports diverse data formats (SQL, Excel, MongoDB)	Limited scope to specific disease domains; reliance on structured data
Li et al., 2025 [20]	Ontology-Driven Medication Query (ODMQ), OMOP CDM, semantic query expansion	Medication retrieval from COVID-19 EHRs	Improves query completeness and accuracy; reduces manual effort; supports query UI and visualization	Generalizability to non-COVID datasets needs validation
Fareedi et al., 2025 [21]	ODSRE methodology, Federated Virtual Knowledge Graph (FVKG), Ontop, OBDA	Semantic interoperability in patient-centric systems	Enables real-time, federated access; low latency; minimal data migration	Requires robust semantic schema mapping; technical complexity
Kamdar et al., 2029 [22]	Linked Data principles, LSLOD Cloud, Semantic Web technologies	Biomedical data integration	Promotes unified data access; potential for AI-driven research	Usability and adoption barriers in LSLOD
Abraham et al., 2025 [23]	Digital twin framework, ontology rules, ML integration, cross-domain reasoning	Brain cancer subtype discovery and precision oncology	Supports personalized treatment; semantic alignment with preclinical models	Complexity in semantic modeling and interpretability across domains

Lazarova et al., 2023 [24]	AD-DPC ontology design, usability studies, conceptual structuring	Alzheimer's disease research collaboration	Facilitates interdisciplinary communication; usable by non-medical professionals	Domain-specific ontology; not evaluated for clinical decision-making
Kawas et al., 2023 [25]	Ontology integration via ontological, lexical, logical, ML techniques; supervised concept mapping	Medical text integration	Accurate semantic alignment; supports data-driven decisions; superior to baseline methods	Requires training data for supervised model; transformation rule complexity

3. Methodology

The suggested approach is aimed at ensuring a smooth harmonization of metadata and semantic integration of diverse pharmaceutical data by means of a multi-layered approach. First, schema mapping methods are used to harmonize heterogeneous data structures, so that metadata is aligned across clinical trials, drug databases, and pharmacovigilance systems, in particular with regard to terminology standardization. It is then succeeded by creation of a rich pharmaceutical ontology that combines standard vocabularies such as MeSH and SNOMED CT, with bespoke domain-specific classes to model drug efficacy, molecular properties, interactions and adverse effects. Contextual NLP and concept tagging are used to carry out semantic annotation of biomedical texts so as to guarantee ontology alignment. Lastly, intelligent semantic query processing uses SPARQL and description logic to map user queries into semantically augmented forms, and applies reasoning on implicit relations, and assists context-sensitive search of pertinent biomedical data. Figure 1 illustrates the overall architecture of the proposed PharmaSeek+ model.

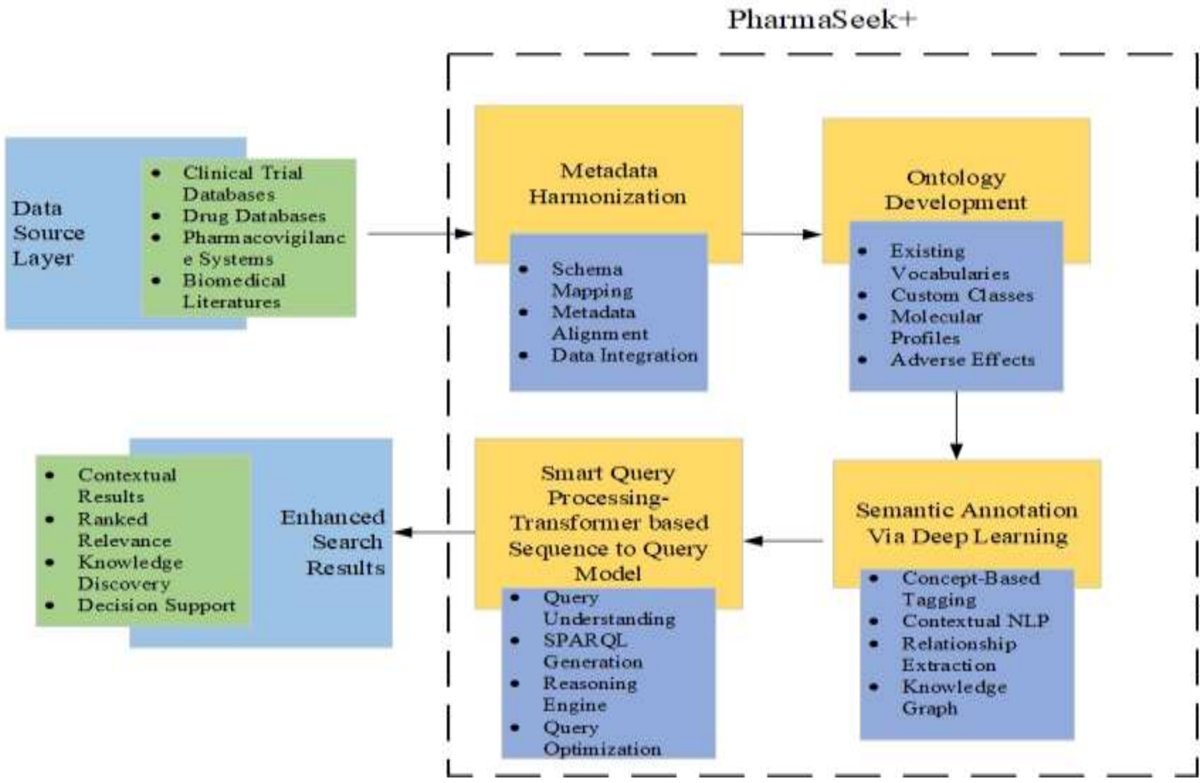


Figure 1: Overall Architecture of Proposed PharmaSeek+ Model.

3.1. Metadata Harmonization and Data Integration

Within pharmaceutical informatics, the process of schema mapping, terminology normalization and metadata alignment are a formal process of integrating various datasets like clinical trials, drug databases, pharmacovigilance systems. The sources of pharmaceutical data are heterogeneous and include such data as clinical trials, EHRs, drug registries, and pharmacovigilance reports, each having different formats and semantics. This diversity hinders easy integration, comparison and reuse of data. The consolidation of such datasets guarantees the uniformity of interpretation and facilitate the thorough analytics. This also facilitates interoperability for research, regulatory, along with clinical applications. Let there be a set of data sources is as shown in Eq. (1), where each D_i represents a dataset from a different source like D_1: clinical trials and D_2: EHR.

$$D = \{D_1, D_2, \dots, D_n\} \qquad (1)$$

Each dataset D_i has its own schema S_i, and it was expressed in Eq. (2), where a_j^i represents attributes in that schema.

$$S_i = \left\{a_1^i, a_2^i, \dots a_{m_i}^i\right\} \qquad (2)$$

The goal of harmonization is to transform all S_i into a unified global schema S^* is represented by Eq. (3),

$$\forall i, \exists f_i : S_i \to S^* \qquad (3)$$

Schema mapping is the process of matching heterogeneous sources based on fields that have similar semantics. Methods involve rule-based methods, ontology-based mappings and machine learning algorithms to identify correspondences with drug_name ↔ medication. This allows querying and accessing heterogeneous structural data. The schema mapping function f_i is defined by similarity measures. For attributes $a \in S_i, b \in S^*$, this has been computed by Eq. (4), where, $lexsim$ represents the lexical similarity $typesim$ indicates as datatype similarity,$contextsim$ represents as co-occurrence/context similarity, in which $lexsim(a,b), typesim(a,b), contextsim(a,b) \in [0,1]$ $also; \alpha, \beta, \gamma \ \in [0,1] and \ \ \alpha + \ \beta + \ \gamma = 1$. A match is confirmed if: $sim(a,b) \geq \theta$, where $\theta \in [0,1]$ is a predefined threshold.

$$sim(a,b) = \alpha \cdot lexsim(a,b) + \beta \cdot typesim(a,b) + \gamma \cdot contextsim(a,b) \qquad (4)$$

Metadata alignment provides consistency of the representation of data attributes such as units, formats, and definitions across datasets like clinical trials, drug registries, and pharmacovigilance systems. It entails harmonizing data differences in types, labels and standards of measurement. This consistency enhances data integration, interoperability and proper cross-dataset analysis. Let metadata is termed as a tuple is shown in Eq. (5), where; T represents as Data type, U indicates as Unit, D indicates as Definition or description.

$$M = (T, U, D) \qquad (5)$$

Let two datasets D_i, D_j have metadata for the same concept is as shown in Eq. (6), where, U_i and U_j are unit-convert which evaluates whether units U_i and U_j are equivalent and convertible using a deterministic transformation rule, such as 1 g=1000 mg, 1 hr=60 min, Temperature, concentration, and currency transformations

$$M_i = (T_i, U_i, D_i), M_j = \left(T_j, U_j, D_j\right) \qquad (6)$$

Metadata alignment ensures which are displayed by Eq. (7), where δ is a semantic similarity threshold.

$$T_i \equiv T_j, unit - convert(U_i, U_j) = true, sim(D_i, D_j) \geq \delta \qquad (7)$$

Terminology standardization and normalization is the process of mapping diverse medical or pharmaceutical terminologies to standard and controlled vocabularies such as RxNorm, MeSH, or SNOMED CT. This makes it so that synonyms such as acetaminophen and paracetamol are considered to be the same concept. It minimizes ambiguity and enhances the retrieval, integration, and semantic reasoning of data. Finally, it allows a meaningful comparison and aggregation of various healthcare data sets. Each term $t \in T$ from the dataset is mapped to a standard vocabulary V is as given by Eq. (8), with an example: $\mu("Advil") = RxNorm - ID{:}\,5640$.

$$\mu = T \rightarrow V \qquad (8)$$

This allows semantic equivalence is shown by Eq. (9), even if $t_1 = "Advil"$ and $t_2 = "Ibuprofen"$

$$\mu(t_1) = \mu(t_2) \Rightarrow t_1 \equiv t_2 \qquad (9)$$

The terminology normalization process, computed semantic similarity between two terms t_1 and t_2 using BERT embeddings. The cosine similarity is calculated by Eq. (10), where; v_{t_1} and $v_{t_2} \in \mathbb{R}^d$ are the vector embeddings of terms t_1 and t_2 generated by a contextual model such as BioBERT, . denotes the dot product, $\|v_{t_i}\|$ is the Euclidean norm of the vector.

$$sim_{BERT}(t_1, t_2) = cos(v_{t_1}, v_{t_2}) = \frac{v_{t_1} . v_{t_2}}{\|v_{t_1}\| . \|v_{t_2}\|} \qquad (10)$$

3.2. Ontology Development and Semantic Annotation

Ontology Development and Semantic Annotation is the development of structured representations of knowledge in pharmaceuticals by combining existing ontologies such as MeSH and SNOMED CT with domain-specific concepts. It allows formal encoding of drug related entities such as efficacy, interactions in formal logic of OWL. NLP is used in semantic annotation to create connections between terms in unstructured text and these ontology concepts to analyze them in more detail.

A) Construction of a pharmaceutical ontology

MeSH, SNOMED CT and custom classes integration unites standardized biomedical vocabularies with concepts in the domain of drugs. This forms a coherent ontology of high-quality and complete semantic representation. Let $\mathcal{O}_{MeSH}, \mathcal{O}_{SNOMED}$ and $\mathcal{O}_{Custom}$ be the ontologies for MeSH, SNOMED CT, and custom domain-specific concepts accordingly. The integrated pharmaceutical ontology$\mathcal{O}_{Pharma}$ is defined as in Eq. (11),

$$\mathcal{O}_{Pharma} = \mathcal{O}_{MeSH} \cup \mathcal{O}_{SNOMED} \cup \mathcal{O}_{Custom} \qquad (11)$$

The ontological triples in the form ⟨subject, predicate, object⟩ like ⟨DrugA, inhibits, EnzymeX⟩ are used to model drug efficacy, molecular profiles, interactions and side effects. OWL or RDF is used to encode these relationships in order to support machine-readable reasoning. This makes it possible to query and infer the drug behaviors and risks automatically. They are represented formally as a description logic (DL) in OWL or RDF triples. Let drug D which reduces blood pressure and targets receptor R is shown in Eq. (12),

$$Drug(D) \wedge has\ Effect(D, Lower\ Blood\ Pressure) \wedge targets(D, R) \qquad (12)$$

3.3. Semantic Annotation of Biomedical Texts

By use of semantic annotation one can convert the unstructured biomedical text into structured and ontology aligned knowledge. This is done by accurately identifying drug related entities and relationships and associating them with standardized biomedical ontologies so they can be used downstream to reason and to make semantic queries.

A) Contextual NLP with Deep Learning (BioBERT)

In contrast to universal models, such as BERT, BioBERT is a task-specific neural network -deep learning model which is pre-trained over massive biomedical data like PubMed abstracts and PMC full-text articles. It has been optimized on Named Entity Recognition (NER) and Relation Extraction (RE) on clinical and pharmaceutical text. Let T be an unstructured biomedical text. A contextual NLP model such as BioBERT processes this text to extract a set of named biomedical entities through token-level classification and contextual embedding. This process is represented in Eq. (13), where; T denoted as unstructured biomedical text, $f_{BioBERT}$ be the deep learning function based on the BioBERT model that does NER over T, e_i are named, entities extracted from T each entity e_i typically includes token span in the input text, entity type label ,confidence score from the model output

$$f_{BioBERT}(T) \rightarrow \{e_1, e_2, \dots, e_n\} \qquad (13)$$

B) Ontology-Aligned Concept Tagging

Once entity vectors $\bar{v}_{e_i}$ are generated, semantic annotation maps each e_i to a concept c_j in a pharmaceutical ontology O_{Pharma} using similarity matching is as expressed by Eq. (14), Where: $\bar{v}_{e_i}$: BioBERT embedding of entity e_i, $\bar{v}_{c_j}$: Pre-computed or fine-tuned embedding of ontology concept c_j, $sim(.)$ represents as Similarity function

$$Align(e_i) = arg_{c_j \in O_{Pharma}}^{max} sim\left(\bar{v}_{e_i}, \bar{v}_{c_j}\right) \qquad (14)$$

The similarity function is typically expressed by Eq. (15),

$$sim\left(\bar{v}_{e_i}, \bar{v}_{c_j}\right) = \frac{\bar{v}_{e_i} \cdot \bar{v}_{c_j}}{\left\|\bar{v}_{e_i}\right\| \left\|\bar{v}_{c_j}\right\|} \qquad (15)$$

The use of a BioBERT, a fine-tuned deep learning model that achieved state-of-the-art performance on multiple biomedical tasks because it was trained on a large set of biomedical corpora, makes it accurate when detecting drug-related entities. In contrast to the use of the traditional keyword matching, it resorts to ontology-aware embedding alignment, thus allowing a close disambiguation of the biomedical terms used in context. It uses the semantics more deeply in that it computes similarity between the entity and ontology embeddings through this method. Moreover, it is scalable and interoperable since it allows mapping of extracted concepts to vocabularies, such as MeSH, SNOMED CT, and RxNorm. This acts as an easy synchronization of systems, enabling the applications semantic search, clinical decision support, and pharmacovigilance.

3.4. Smart Semantic Query Processing and Reasoning

Smart Semantic Query Processing and Reasoning allows systems to understand natural language queries and output into semantic query languages such as SPARQL that go be understood by machines. It extends keyword-based retrieval, allowing reasoning both on ontologies using Description Logic (DL) and Transformer-based models, and extracting both explicit and implicit and indirect biomedical knowledge. The combination of context-awareness will deliver quite contextual and personal reaction in clinical and pharmaceuticals applications.

A) Transformation of user input into semantic queries

In healthcare, user queries are usually formulated using natural languages that are incompatible with the machine. These queries are formally encoded in formal semantic formats such as SPARQL before they are used to access related data in well-structured knowledge bases such as RDF graphs. This consists in determining some important things such as drugs, hypertension and matching them with concepts in the ontology. The resulting SPARQL query is able to search the ontology to derive meaningful results. It is a way in which human language and machine-readable knowledge become bridged so that information could be retrieved accurately. A natural language query is expressed in Eq. (15),

$$\text{"}Show\ drugs\ that\ treats\ hypertension\text{"} \qquad (15)$$

B) Intelligent Query Processing with Transformer-Based Models

These are embedded and semantically further interpreted with the help of Transformer-based sequence-to-query models, like T5, BART, or GPT, which are fine-tuned on biomedical QA-related data like BioASQ. These models learn to translate a user query Q_{nl} into a formal SPARQL query Q_{sparql}. Let Q_{nl} be the natural language query, $M_{Transformer}$ represents as Transformer-based model trained to generate queries, Q_{sparql} be the Output SPARQL query. The transformation is given by Eq. (16),

$$Q_{sparql} = M_{Transformer}(Q_{nl}) \qquad (16)$$

C) Use of SPARQL and Description Logic for Query Execution

SPARQL is a query language of data in the RDF triple format of subject predicate object. The formal semantics of ontologies are offered by the DL, which allows one to make logical inferences. Collectively, they used to query and reason over biomedical knowledge graphs in a very specific way. Let $Treats$ be an object property, $Condition\ X \subseteq Cardiovascular\ Disease$, $Treats(DrugA, Condition\ X)$ be an asserted fact. Then the reasoner infer is shown by Eq. (17),

$$Condition\ X \subseteq Cardiovascular\ Disease,\ Treats(DrugA, Condition\ X) \Rightarrow Treats(DrugA, Cardiovascular\ Disease) \qquad (17)$$

D) Reasoning Over Implicit and Indirect Relationships

Having to reason over things that are implicit and indirect relationships means that logical inference is made to infer new knowledge based on existing ontology structures. Such methods as subsumption with in case of "Hypertension" being a subclass of "cardiovascular disease" and transitivity with in case A treats B, and B causes C, A have an effect on C are used. The deduction is done by ontology reasoners such as Pellet or HermiT. This allows systems to find more deep and intuitive knowledge in the fields of biomedicine. Let $HasCondition\ (Patient\ , A)$ be an object property assertion, $A \subseteq B$ where A and B are disease concepts in an ontology. Then the system infers is given by Eq. (18),

$$HasCondition\ (Patient\ , A), A \subseteq B \Rightarrow HasCondition\ (Patient\ , B) \qquad (18)$$

E) Context-Aware Information Retrieval

Context-sensitive information search is the one that modifies the outcome of a query or search by a consideration of several parameters about the user such as age/sex, medical history, presence of comorbidities. A drug that is fit in adults are not be fit in the case of the pediatric patient, i.e., in terms of dosage or in the profile of side effects. Formalization of context models' Semantic rules or filters applied in executing the query are common forms of formalizing context models. This method results in more pertinent and secure and individualistic biomedical ideas. Let $Q(u,c)$ be the query function for user u under context c , $r \in R$ be the result from the candidate set R, $Score\ (r,u,c)$ be the context-relevance score. Then the system returns are expressed in Eq. (19), Figure 2 illustrates the framework for transformer-based query translation.

$$Q(u,c) = arg_{r \in R}^{max} Score\ (r,u,c) \qquad (19)$$

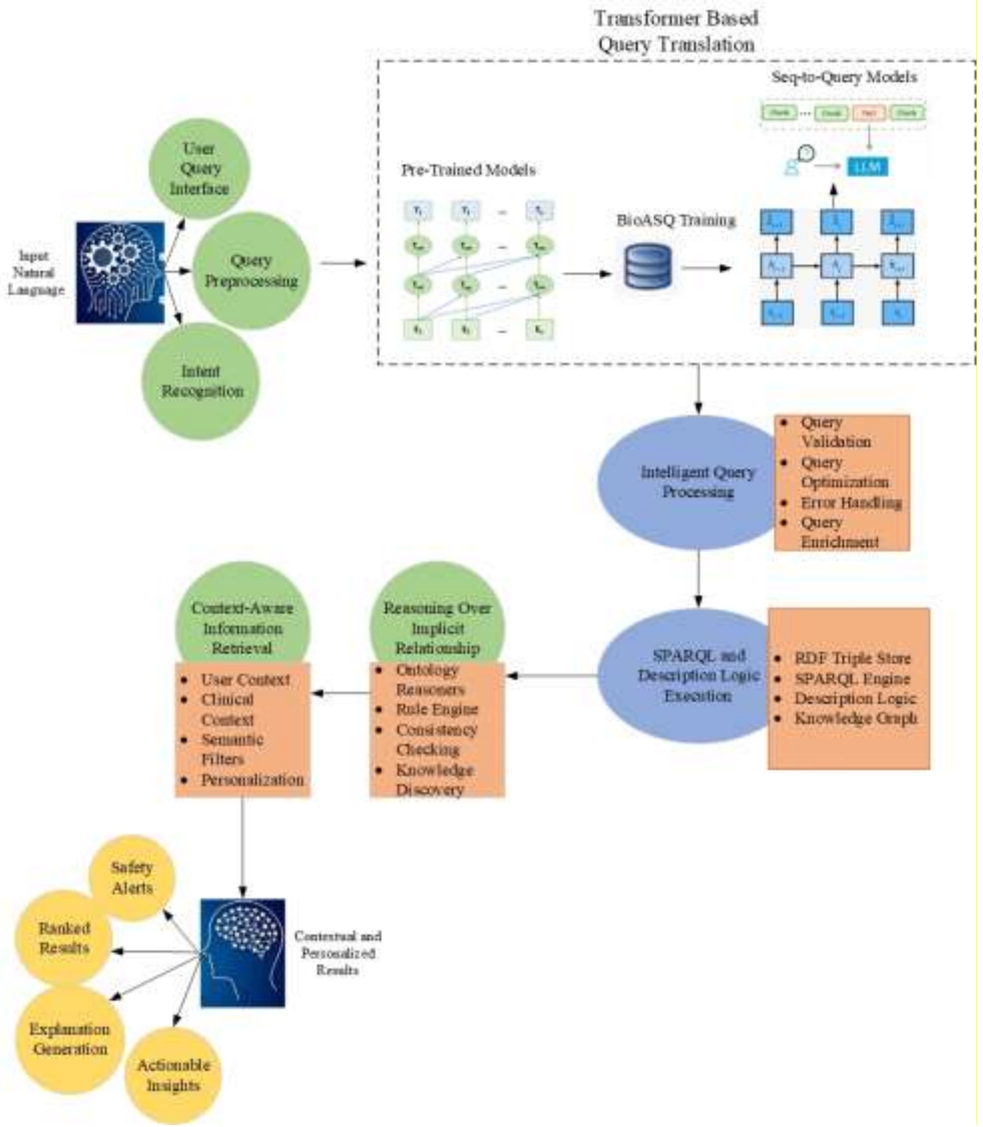


Figure 2: Transformer Based Query Translation Framework

4. Evaluation, Results, and Impact

The evaluation of performance metrics like precision, Recall, F1-score, MAP (Mean Average Precision), USS (User Satisfaction Score) and SQRT (SPARQL Query Response Time) with the proposed model has been compared with the existing models of CASBERT [26], Transformer-based language models [27], Transformer-based RoBERTa [28], and Transformer-based embedding model (TEM) [29]. Table 2 determines the formulas for performance metrics.

Table 2: Formula and Description of Performance Metrics

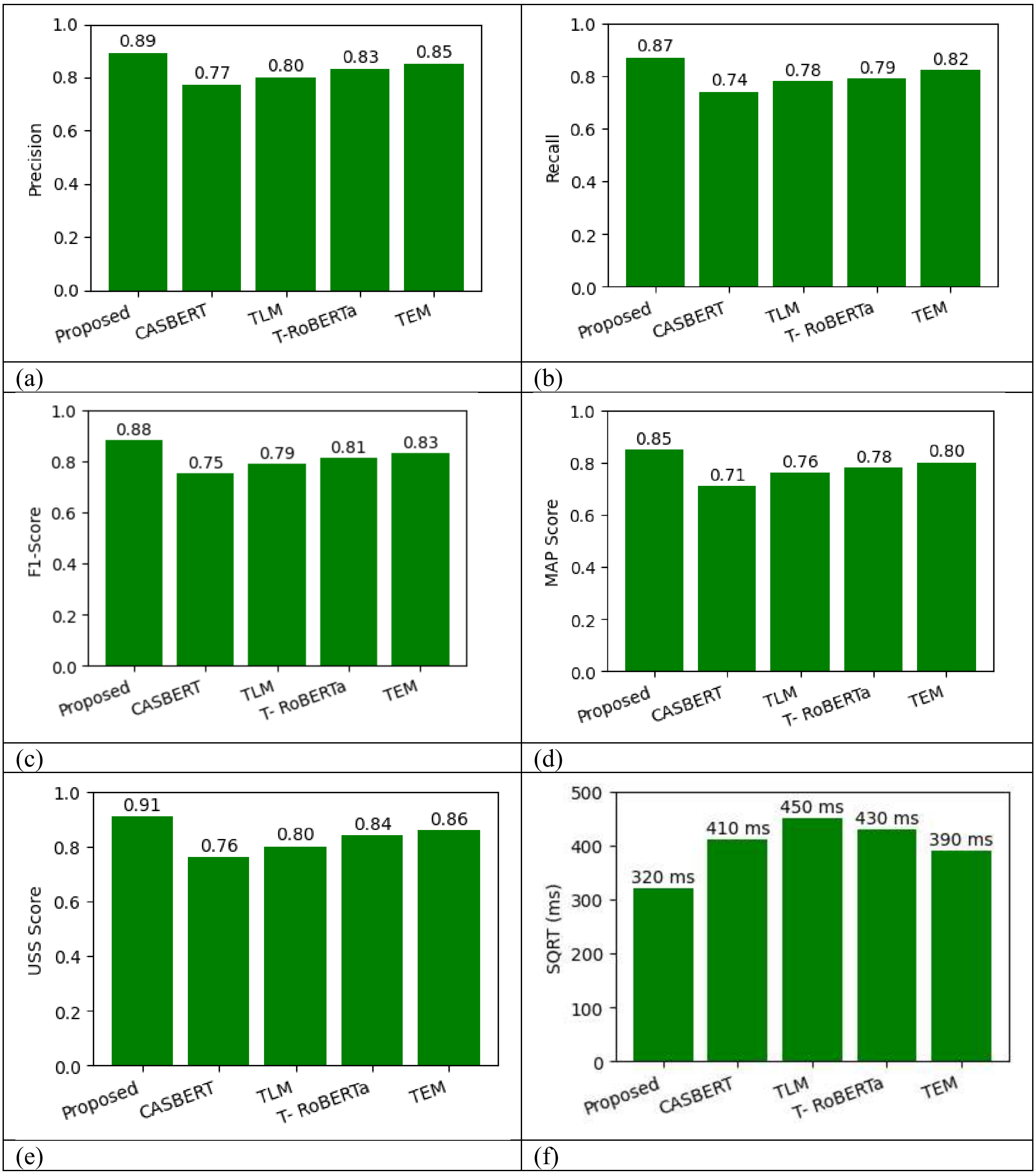
Metrics	Equation	Description
Precision	$Precision = \frac{TP}{TP+FP}$	Measures how many of the predicted positive cases are actually positive
Recall	$Recall = \frac{TP}{TP+FN}$	Measures the proportion of relevant documents that were retrieved.
F1-Score	$F_1 = 2 \times \frac{Precision \times Sensitivity}{Precision + \mathrm{S}ensitivity}$	Harmonic mean of precision and sensitivity
MAP	$\mathrm{MAP} = \frac{1}{\lvert Q \rvert}\sum_{q=1}^{\lvert Q \rvert}\left(\frac{1}{\lvert R_q \rvert}\sum_{k=1}^{\lvert R_q \rvert}\mathrm{Precision\ @k}\right)$	Evaluates precision across multiple queries by averaging precision at each relevant document.
USS	$\mathrm{USS} = \frac{1}{N}\sum_{i=1}^{N} S_i$	Provides a direct measure of real-world system effectiveness from the user's perspective, enabling user-centric evaluation.
SQRT	$\mathrm{SQRT} = T_{end} - T_{start}$	Measures how long it takes for the system to process and return a result for a SPARQL query.

Table 3 shows the comparison of PharmaSeek+ with the available models considering six major performance metrics determined as Precision, Recall, F1-Score, Mean Average Precision (MAP), User Satisfaction Score (USS), and SPARQL Query Response Time (SQRT). Among any of the baseline models, the proposed PharmaSeek+ framework has the best values of Precision (0.89), Recall (0.87), and F1-Score (0.88) showing that it is the most accurate and balanced to find relevant results. It is also the leader in MAP (0.85), which proves its high-ranking efficiency, and is the highest USS (0.91), that proves it corresponds to user recommendations. Besides, PharmaSeek+ offers the most responsive time, i.e., 320 ms, proving its effectiveness in terms of semantics query answering, unlike CASBERT and other transformer-based approaches, which though competitive produce lower scores and increased execution times. It gives emphasis on how well PharmaSeek+ performs as well as is user-friendly in delivery.

Table 3: Comparison of Proposed Model with Performance Metrics

Models	**Precision**	**Recall**	**F1-Score**	**MAP**	**USS**	**SQRT (ms)**
Proposed PharmaSeek+	0.89	0.87	0.88	0.85	0.91	320
CASBERT	0.77	0.74	0.75	0.71	0.76	410
Transformer-based language models (TBLM)	0.8	0.78	0.79	0.76	0.8	450
Transformer-based RoBERTa (T-RoBERTa)	0.83	0.79	0.81	0.78	0.84	430
Transformer-based embedding model (TEM)	0.85	0.82	0.83	0.8	0.86	390

Figure 3 shows a comparative study of the proposed PharmaSeek+ model with some of the baseline models based on some important performance measures like Precision, Recall, F1-Score, Mean Average Precision (MAP), User Satisfaction Score (USS), and SPARQL Query Response Time (SQRT). The figure shows vividly that PharmaSeek+ is much better than any standard transformer-based model such as RoBERTa, CASBERT, and embedding-based architecture in most of the evaluation metrics. It should be noted that PharmaSeek+ performs best (0.89 precision and 0.87 recall) than other systems and thus poses the least possibility of retrieving irrelevant information about drugs but with fewer false positive and negative results. Then, it also shows the best USS (0.91) and this underlines that it is effective in addressing intent and expectations of the users. Moreover, it is efficient and precise since its lower SQRT (320 ms) shows that its query processing speed is optimized. These findings confirm that PharmaSeek+ is a powerful, intelligent semantic search environment to be applied to pharmaceutical research.



(a) (b) (c) (d) (e) (f)

Figure 3: Model Comparison with Proposed PharmaSeek+ model with (a) Precision, (b) Recall, (c) F1-Score, (d) MAP, (e) USS and (f) SQRT

5. Conclusion

In this study, an ontology-based semantic search framework, PharmaSeek+ is introduced, which substantially improves the process of pharmaceutical research through the integration of heterogeneous data resources and intelligent information retrieval. The framework provides strong metadata integration by means of schema mapping and alignment of clinical trials, drug databases, and pharmacovigilance systems. A pharmaceutical ontology is built on the basis of integrating MeSH, SNOMED CT, and custom classes to represent drug efficacy, molecular interactions, and adverse effects. Deep learning models such as BioBERT used to perform semantic annotation to make sure that the entity recognition and ontology-based tagging of biomedical texts is accurate. Moreover, the reasoning over implicit relationships is also possible due to the intelligent query processing with a transformer-based sequence-to-SPARQL model, which guarantees highly contextual and relevant results. According to the evaluation results, PharmaSeek+ has significantly outperformed the existing models, including CASBERT and RoBERTa, in all the key metrics, including precision (0.89), recall (0.87), and user satisfaction score (0.91) with the fastest query response time (320 ms). Such results confirm the efficacy of the framework and point to its potential future scalability, interoperability, and clinical awareness in pharmaceutical knowledge discovery, opening the door to applying it in the real world to drug development, decision support, and biomedical research.

Reference

1. Zhu, R., Vora, B., Menon, S., Younis, I., Dwivedi, G., Meng, Z., ... & International Consortium for Innovation and Quality in Pharmaceutical Development (IQ) Real-World Data Working Group. (2023). Clinical Pharmacology Applications of Real-World Data and Real-World Evidence in Drug Development and Approval–An Industry Perspective. Clinical Pharmacology & Therapeutics, 114(4), 751-767.
2. Chelazzi, F. (2023). From isolated data silos to an integrated and multi-proxy regional synthesis. Cyprus in the context of exploring ancient patterns of human-environment-climate interaction. Cahiers du Centre d'Études Chypriotes, (52-53), 153-178.
3. Mirakhori, F., & Niazi, S. K. (2025). Harnessing the AI/ML in Drug and Biological Products Discovery and Development: The Regulatory Perspective. Pharmaceuticals, 18(1), 47.
4. Beierle, J., Algorri, M., Cortés, M., Cauchon, N. S., Lennard, A., Kirwan, J. P., ... & Abernathy, M. J. (2023). Structured content and data management—enhancing acceleration in drug development through efficiency in data exchange. AAPS open, 9(1), 11.
5. Martenot, V., Masdeu, V., Cupe, J., Gehin, F., Blanchon, M., Dauriat, J., ... & Zucker, J. D. (2022). LiSA: an assisted literature search pipeline for detecting serious adverse drug events with deep learning. BMC medical informatics and decision making, 22(1), 338.
6. Alrasheed, H. (2021). Word synonym relationships for text analysis: A graph-based approach. Plos one, 16(7), e0255127.
7. Fei, H., Tan, S., & Li, P. (2019, July). Hierarchical multi-task word embedding learning for synonym prediction. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 834-842).
8. Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., ... & Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. Trials, 22, 1-15.
9. Knevel, R., & Liao, K. P. (2023). From real-world electronic health record data to real-world results using artificial intelligence. Annals of the Rheumatic Diseases, 82(3), 306-311.
10. Zou, K. H., & Berger, M. L. (2024). Real-World Data and Real-World Evidence in Healthcare in the United States and Europe Union. Bioengineering, 11(8), 784.
11. Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. Frontiers in cell and developmental biology, 8, 673.
12. Mannion, A. (2025). Demystifying machine learning applications in general medicine (Doctoral dissertation, Laboratoire d'Informatique de Grenoble).
13. Yu, H. Q., O'Neill, S., & Kermanizadeh, A. (2023). AIMS: An Automatic Semantic Machine Learning Microservice Framework to Support Biomedical and Bioengineering Research. Bioengineering, 10(10), 1134.

14. Bains, R., Huzard, D., McCutcheon, J., Boguszewski, P., Virag, D., Restivo, L., ... & Girard, B. (2025). Too big to lose-a FAIR repository for biomedical data derived from home-cage monitoring.
15. Bernasconi, A., Guizzardi, G., Pastor, O., & Storey, V. C. (2022). Semantic interoperability: ontological unpacking of a viral conceptual model. BMC bioinformatics, 23(Suppl 11), 491.
16. Stănescu, G., & Oprea, S. V. (2025). Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling. Electronics, 14(7), 1313.
17. Yao, Z., Liu, B., Wang, F., Sow, D., & Li, Y. (2023). Ontology-aware prescription recommendation in treatment pathways using multi-evidence healthcare data. ACM Transactions on Information Systems, 41(4), 1-29.
18. Bakshi, W. J., Aasim, S., Butt, M. A., & Qadri, M. H. Ontology-Driven Solutions for Resolving Semantic Conflicts in Data Integration.
19. Thirumahal, R., Sudha Sadasivam, G., & Shruti, P. (2022). Semantic integration of heterogeneous data sources using ontology-based domain knowledge modeling for early detection of COVID-19. SN Computer Science, 3(6), 428.
20. Li, X., Huang, Y., Cui, L., Tao, S., & Zhang, G. Q. (2025, May). Optimizing Medication Querying Using Ontology-Driven Approach with OMOP: with an application to a large-scale COVID-19 EHR dataset. In AMIA Annual Symposium Proceedings (Vol. 2024, p. 693).
21. Fareedi, A. A., Gagnon, S., Ghazawneh, A., & Valverde, R. (2025). Semantic Fusion of Health Data: Implementing a Federated Virtualized Knowledge Graph Framework Leveraging Ontop System. Future Internet, 17(6), 245.
22. Kamdar, M. R., Fernández, J. D., Polleres, A., Tudorache, T., & Musen, M. A. (2019). Enabling web-scale data integration in biomedicine through linked open data. NPJ digital medicine, 2(1), 90.
23. Abraham, J., Austin, M., Gilbert, M. R., & Celiku, O. (2025). Semantic Foundations for Precision Medicine. ACM Transactions on Computing for Healthcare.
24. Lazarova, S., Petrova-Antonova, D., & Kunchev, T. (2023). Ontology-Driven Knowledge Sharing in Alzheimer's Disease Research. Information, 14(3), 188.
25. Kawas, M., Alkhatib, B., & Dashash, M. (2023). Enhancing ontology integration in medical texts through advanced mechanisms.
26. Munarko, Y., Rampadarath, A., & Nickerson, D. (2023). Building a search tool for compositely annotated entities using Transformer-based approach: Case study in Biosimulation Model Search Engine (BMSE). F1000Research, 12, 162.
27. Coelho, J., Neto, A., Tavares, M., Coutinho, C., Oliveira, J., Ribeiro, R., & Batista, F. (2021). Transformer-based language models for semantic search and mobile applications retrieval. Transformer-based language models for semantic search and mobile applications retrieval, 225-232.
28. Kamil, M., & Çakır, D. (2025). Advances in Transformer-Based Semantic Search: Techniques, Benchmarks, and Future Directions. Turkish Journal of Mathematics and Computer Science, 17(1), 145-166.
29. Bi, K., Ai, Q., & Croft, W. B. (2020, July). A transformer-based embedding model for personalized product search. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1521-1524).