Biomedical String Analysis with LLMs

Zurich University of Applied Sciences
Applied Computational Life Sciences

Authors Mohan Adluru
Supervisors Ahmad Aghaebrahimian & Manual Gil
Study Program Track Module II
Submitted on 23.06.2025

Contents

Acronyms

AI	Artificial Intelligence.
AMR	Abstract Meaning Representation.
BERT	Bidirectional Encoder Representations from Transformers.
BHC	Brief Hospital Course.
BLEU	Bilingual Evaluation Understudy.
ChatGPT	Chat Generative Pre-trained Transformer.
CITI	Collaborative Institutional Training Initiative.
CoT	Chain-of-Thought.
CT	Computed Tomography.
CUDA	Compute Unified Device Architecture.
DO	Disease Ontology.
EHRs	Electronic Health Records.
F1	F1 Score (Harmonic mean of precision and recall).
FACTKB	Factual Knowledge Base Framework.
FeQA	Factual Question Answering.
GO	Gene Ontology.
GPT-4	Generative Pre-trained Transformer 4.
i2b2	Informatics for Integrating Biology and the Bedside.
KB	Knowledge Base.
KEGG	Kyoto Encyclopedia of Genes and Genomes.
KGs	Knowledge Graphs.
LLMs	Large Language Models.
LoRA	Low-Rank Adaptation.
MIMIC-IV	Medical Information Mart for Intensive Care IV.
MLP	Multi-Layer Perceptron.
MRI	Magnetic Resonance Imaging.
NCBI	National Center for Biotechnology Information.
NLP	Natural Language Processing.
OWL	Web Ontology Language.

PSO	Particle Swarm Optimization.
QA	Question Answering.
QLoRA	Quantized Low-Rank Adaptation.
QuestEval	Question-Answer based Evaluation.
RDF	Resource Description Framework.
RHO	Response re-ranking with Knowledge Graph grounding.
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation (Longest Common Subsequence).
SHACL	Shapes Constraint Language.
SLURM	Simple Linux Utility for Resource Management.
SOTA	State of the Art.
SPARQL	SPARQL Protocol and RDF Query Language.
SVM	Support Vector Machine.
SWTs	Semantic Web Technologies.
UMLS	Unified Medical Language System.
VRAM	Video Random Access Memory.
WandB	Weights and Biases.

1 Abstract

This report explores the combination of Semantic Web Technologies, Biomedical String Classification and Large Language Models (LLMs) to advance in biomedical data processing and clinical summarization. Semantic Web frameworks enable structured interoperable data representation while a two-stage feature selection approach combining domain specific ontologies and bio-inspired optimization enhances medical text classification. The report further investigates the complementary strengths of LLMs and Knowledge Graphs (KGs) to address factual consistency challenges with a focus on hallucination detection techniques that improve model reliablity. Finally, a clinical summarization pipeline to replicate and evaluate state-of-the-art (SOTA) LLMs on the MIMIC-IV dataset leveraging Quantized Low-Rank Adaptation (QLoRA) fine-tuning and hyperparameter optimization. The results from replicating the benchmark paper[1] highlight the potential of combining domain adaptation with LLM architectures to deliver clinically relevant, accurate summaries, underscoring future directions for scalable, interpretable and trustworthy AI driven healthcare applications.

2 Semantic Web

The Semantic Web is represented as an extension of the traditional web by sharing structured, machine-readable data rather than just text documents. The Semantic web was introduced by Tim Berners-Lee in 2001 with a clear goal to transform "data of web" into a "web of data" where both humans and machines can access, interpret, and integrate information enhancing computational tasks resulting in more intelligent interactions and data-driven decision-making [2].

According to Moore's Law, the computational power doubles and size gets halved approximately every 18 months which managed to keep its pace with the growth of biological data[3]. This trend drastically shifted on completion of the Human Genome Project in 2003, where data growth surpassed the predictions of Moore's Law significantly [4]. Currently, biological data includes genomic sequences, medical imaging data, Electronic Health Records (EHRs) etc. which are being produced rapidly[5]. The sheer volume and heterogeneity of these datasets present with major challenges for integration, retrieval and analysis. The Semantic Web is well-positioned to address these challenges by providing clear structured semantics, enabling meaningful integration and analysis.

Semantic Web Technologies (SWTs) such as the Resource Description Framework (RDF) and the Web Ontology Language(OWL) provide the backbone for representing and linking biomedical data [6] [7]. It facilitated structured data exchange, interoperability, and integration through various tools and standards. RDF utilizes subject-predicate/verb-object triples to represent data allowing clear and consistent integration of data from diverse sources. For instance, integrating genomic data from multiple databases such as KEGG, NCBI Protein, and Ensemble can be represented clearly using RDF triples, e.g., `<hsa:675> <encodes> <protein:119395734>` and `<hsa:675> <owl:sameAs> <ENSG00000139618>`. Ontologies defined by Web Ontology Language (OWL) formally specify entities and relationships within specific knowledge domains, as seen in widely adopted biomedical ontologies like the Gene Ontology and Disease Ontology[8]. These ontologies support semantic reasoning, enabling more precise data retrieval, and knowledge discovery. Furthermore, linked data principles aim to enable interoperability and seamless integration by standardizing data publishing

methods across diverse biomedical datasets [9].

SPARQL (SPARQL Protocol and RDF Query Language) the standardized query language for RDF format, enables complex queries integrating complex and heterogeneous biomedical datasets, which is crucial for applications such as precision medicine, epidemiology, and integrative clinical research [10]. The practical impact of these techniques includes improved data retrieval, enhanced interoperability, and the ability to perform advances analytics and hypothesis generation across multiple biomedical domains. Despite these advantages, SWTs face several Limitations. Achieving horizontal scalability in graph databases, especially for high write volumes, remains challenging. Compared to relational databases, graph databases require complex maintenance, particularly in terms of concurrency and transaction management. Large-scale analytical processing can be computationally expensive, and ensuring data quality and consistency across diverse complex biomedical datasets demands rigorous validation and governance. As SWTs significantly transforms biomedical informatics by addressing data integration and interoperability challenges, future advancements in SWTs will likely focus on enhancing scalability, simplifying graph database management, and improving analytical capabilities for large-scale biomedical data [11]. Interdisciplinary collaborations among computer scientists, bioinformaticians, clinicians, and domain experts will be essential to ensure that Semantic Web Technologies (SWTs) continue to address real-world biomedical challenges and deliver meaningful benefits for research and healthcare.

3 Biomedical String Classification

Biomedical document classification is an critical task in artificial intelligence which aims to assign predefined labels to unseen medical texts. This supports clinicians in decision making and can potentially reduce healthcare costs. However, analysing medical documents have some unique challenges due to the extensive usage of acronyms, abbreviations, and concise domain specific terminologies, which make accurate information extraction and reliable text classification difficult with current methods. Traditional approaches often rely on massive number of features derived from raw text, often represented by a "bag-of-words" model, which introduces noisy data, slow down learning, and affect prediction accuracy [12].

To overcome these limitations effective feature selection is essential. Standard single-stage feature selection methods such as filter and wrapper methods often struggle with the extremely large search spaces in medical texts. These approaches either neglect crucial interactions between features or are computationally expensive due to their complexity. To enhance the efficiency of feature selection, a two-stage approach was proposed where initially a refined set of features are extracted from raw medical documents and then in the next stage optimization of this subset happens.

The first stage leverages the Unified Medical Language System (UMLS) which serves as a comprehensive, domain-specific biomedical dictionary and ontology for systematically extracting relevant features. By utilizing tools like MetaMap the system processes clinical notes to identify meaningful and critical biomedical phrases, translating these into standardized concepts from UMLS Metathesaurus. This stage focuses on relevant semantic types such as "Disease or Syndrome" and " Sign or Symptom" which are closely related to the classification task that identifies the conditions like Coronary Artery Disease. This initial filtering significantly reduces the number of features by retaining only the most informative concepts

creating a manageable search space for the next stage[13].

The second stage employs Particle Swarm Optimization(PSO) and bioinspired optimisation technique that is best known for its robust search abilities, quick convergence, and minimal parameter requirements. PSO iteratively selects a more refined and meaningful subset from the features extracted in the first stage. The fitness of each potential feature subset is assessed based on its contribution to classification accuracy, using machine learning algorithms like Naive Bayes, Linear Support Vector Machines (SVM) and Logistic Regression. The integration of PSO on a previously filtered feature set its efficiently identified the subsets that significantly enhance predictive accuracy while maintaining computational efficiency [14].

The effectiveness of this two-stage approach was evaluated on the dataset from 2010 Informatics for Integrating Biology and Bedside(i2b2) for Coronary Artery Disease classification[15]. This method was able to extract meaningful information and significantly reduce the number of features to approximately 5% of the original size, a considerably smaller size compared to single-stage methods or even initial UMLS only extraction and consistently enhanced classification accuracy across most classifiers. This method effectively addressed critical challenges associated with high-dimensional, noisy medical text data and emphasises the importance of integrating domain-specific biomedical knowledge.

4 Introduction to Large Language Models (LLMs)

Large Language Models (LLMs) have emerged as a transformative technology in the domain of artificial intelligence, profoundly reshaping natural language processing (NLP) with their remarkable abilities and generalizability[16]. LLMs are unsupervised models trained on vast text corpora which consists of repositories of general knowledge, linguistic patterns and commonsense reasoning. This robust foundational training enables LLMs to excel across diverse NLP tasks, including question answering, seamless machine translation, and text generation[17].

However, the extraordinary capabilities of LLMs come with significant challenges due to their black-box nature where their internal decision-making process remain closed for any human interpretation. A critical concern remains their susceptibility to "hallucinations", a phenomenon where models generate plausible but factually incorrect statements such as attributing the discovery of gravity to Einstein in 1687 which severely affects the model's reliability in sensitive fields such as healthcare diagnostics, legal decision-making, or critical educational contexts. Their knowledge implicitly encoded within billions of parameters so directly validating or interpreting their reasoning remains challenging. Moreover, models trained on general corpuses often struggle to effectively adapt to highly specialized domains, real-world knowledge without extensive retraining highlights the ongoing challenge of adaptability[18][19].

The architecture supporting these complex models is the Transformer design, a revolutionary neural network structure that is characterized by encoder and decoder modules integrated with self-attention method that allow them to weigh the importance of different parts of the input data. Transformers are commonly used with three distinct variations in the architecture; encoder-only models, optimized for the tasks that require deep textual understanding

like classification; encoder-decoder models, proficient at sequence-to-sequence tasks such as summarization and translation; and the widely recognized decoder-only LLMs such as GPT-4 and ChatGPT which operate by predicting the next word in a sequence and can perform diverse tasks from a few examples without requiring intensive fine-tuning[20][21]. This unique design have a pivotal impact on a budding field of prompt engineering which helps in carefully crafting natural language inputs and even employ techniques like Chain-of-Thought prompting to enhance model's reasoning abilities and unlock hidden capabilities of these models for various applications[22].

Despite fundamental limitations and challenges, the pure power and adaptability of LLMs have unleashed transformative applications across numerous real-world scenarios. Beyond their fundamental NLP functionalities, these models have created a huge impact in diverse fields such as education, code generation, and recommendation systems. Intelligent chatbots like ChatGPT[23] have helped to transcend interactive applications in image editing and code generation with tools utilizing natural language commands. The continued evolution of LLMs signifies a convincing shift towards creating systems that are not only fluent in communication, but also factually informed and reliable, potentially redefining AI-driven interactions across multiple domains. To address the inbuilt limitations and to enhance their factual reliability and interpretability, integrating structured external knowledge sources looks promising which sets the stage for exploring the integration of LLMs with Knowledge Graphs in the next chapter.

5 Knowledge Graphs(KGs) with LLMs

In the previous section discussed how dynamic landscape of artificial intelligence is and the integration of Large Language Models(LLMs) with Knowledge Graphs (KGs) has emerged as a promising direction to overcome respective limitations and utilize their complementary strengths effectively. While LLMs demonstrate exceptional proficiency in language processing and adaptability to diverse tasks, their "black-box" systems are prone to generate "hallucinations" which are plausible yet factually incorrect, raise significant concerns especially in sensitive domains like medical diagnosis as their vast knowledge is implicitly stored within billions of parameters making it difficult to interpret, validate, or update and frequently struggle with domain-specific information[18][19].

Knowledge Graphs (KGs) stand as structured repositories of explicit factual knowledge, offering clarity, interpretability and accuracy through their reasoning capabilities. KGs are also expert at evolving with new information and can be specifically tailored to provide precise domain-specific knowledge. However, KGs face their own set of hurdles as they are fundamentally difficult and costly to construct, prone to incompleteness, and struggle to effectively model unseen entities or dynamically changing real-world facts[24]. Also traditional KG methods often overlook the rich textual information embedded within them, limiting their general language understanding. As LLMs excel in language processing and generalizability but lack factual precision and interpretability, while KGs offer structured, accurate, and interpretable knowledge but are difficult to construct and maintain which makes sense to combine LLMs and KGs in different possible combinations complementing each other and overcoming individual challenges[25].

The first path, KG-enhanced LLMs which focuses on leveraging structured knowledge from

KGs to strengthen the factual accuracy and interpretability of LLMs. This enhancement can be implemented during the initial training phase of LLMs embedding explicit domain specific knowledge directly into their foundational understanding, or during inference by dynamically retrieving relevant facts from KGs in real-time thereby enhancing their performance and reducing hallucinations. KGs are also leveraged to improve the interpretability of LLMs enabling researchers to probe the knowledge hidden within their parameters and analyze their reasoning process to make their decisions more transparent and trustworthy.

The second path, LLM-augmented KGs that harnesses the power of LLMs to overcome the limitations of KGs. LLMs can significantly streamline and automate KG construction completing missing information, enriching entity and relational embeddings, and facilitating the extraction of structured knowledge from unstructured textual sources. LLMs facilitate KG-to-text generation transforming structured KG facts into coherent natural language descriptions, and enhance knowledge graph question answering by extracting relevant entities and relations over retrieved KG facts.

The third and ultimate destination is Synergized LLMs + KGs where both LLMs and KGs collaboratively enhance each other's performance within unified frameworks. This dynamic interaction results in richer knowledge representation, allowing models to learn jointly from unstructured textual data and structured graph-based knowledge. Such integration significantly enhances applications like advanced search engines, recommendation systems, and intelligent assistants, delivering robust knowledge driven interactions grounded in factual accuracy and interpretability.

Despite the obvious advantages, significant challenges remain including efficient knowledge injection into "black-box" LLM architectures, enabling LLMs to inherently understand and reason with structured graph data. Addressing these challenges are crucial for advancing towards reliable, transparent, and intelligent AI systems. Understanding the delicate balance between structured knowledge from KGs and maintaining linguistic flexibility of LLMs sets the stage for addressing one of the most critical issues facing the integrated systems, hallucination detection. The following section where we will delve deeper into this challenge, exploring effective methodologies and frameworks designed to identify, rectify and manage hallucinations thus further enhancing the reliability and trustworthiness of AI-driven application.

6 Hallucination Detection in LLMs

Hallucination detection in Large Language Models (LLMs) represents a critical field of research due to its profound implications for type reliability and practical utility of generated content. Hallucinations occur when the models produce outputs that seems plausible but contain information that is incorrect, unverifiable or directly inconsistent with the source or established facts. Such inaccuracies significantly affect the trustworthiness of the models particularly in sensitive applications like healthcare, scientific communication, and legal documentation[26].

This issue is commonly observed in tasks such as abstractive summarization, where models are designed to paraphrase and create novel sentences which leads to summaries that are up to 30% inconsistent from the original source documents. These errors often manifest

as incorrect entity names, numbers, or pronouns to more complex issues like temporal inaccuracies, incorrect coreferences, and difficulties with multi-hop compositional reasoning are also significant contributors. Additionally, models trained on synthetic data derived from relatively small datasets may lack robustness when introduced to new domains or evolving information landscapes, where the distribution of entities, events and their relations change dynamically.

Addressing this critical challenge, various methods have emerged for hallucination detection and mitigation in LLMs. Early detection techniques typically concentrated on assessing factual consistency at the document-sentence level which verify each summary sentence against the entire source text found to be more effective than sentence-by-sentence comparisons. To train these detection models more effectively the researchers had employed techniques for generating synthetic training datasets created through rule-based transformations deliberately introducing errors. Traditional NLP evaluation metrices like BLEU and ROUGH-L were effective at assessing general fluency and content overlap but were insufficient for factual accuracy[27][28]. So, QA-based metrics such as FeQA and QuestEval have emerged as more precise evaluative tools. These methods involve generating questions from source documents and assessing factual consistency by comparing the accuracy of answers derived from both the original sources and the generated outputs[29][30]. Another approach is "Entity Coverage" which evaluates whether entities in generated responses are supported by the provided knowledge triples and dialogue context.

More recent developments have leveraged external Knowledge Bases (KBs) to enhance factual evaluation of model outputs. FACTKB is one such approach which enhances model training through three distinct pretraining strategies; "Entity Wiki" which strengthens entity recognition and comprehension; "Evidence Extraction" which trains models to effectively identify supporting evidence; and "Knowledge Walk" which promotes multi-hop compositional reasoning via traversing KB relationships. This approach significantly improves models' capability to detect and correct semantic frame errors, a prominent source of hallucinations, and demonstrates impressive adaptability across multiple domains. FACTKB also stands out for its generalizability across different domains and its simplicity to require minimal preprocessing compared to methods that rely on complex linguistic structures like dependency parses or abstract meaning representation (AMR) graphs. Frameworks such as RHO implement "local" and "global knowledge grounding" by integrating Knowledge Graph embeddings directly into the textual generation process, these methods significantly improve multi-hop reasoning capabilities. RHO further incorporates a "response re-ranking" mechanism based on traversing KG sub-graphs, ensuring that generated responses are faithful by aligning with knowledge paths[31][32].

The effectiveness of these hallucination detection methods is assessed using quantitative metrices such as balanced accuracy and F1 scores along with qualitative human evaluation where domain experts such as clinicians assess accuracy, factual correctness, comprehensiveness, and fluency often showing a preference for LLM generated content that excels in factual consistency. The next critical step in harnessing the full potential of these detection techniques involve deep integration with structured knowledge representations, setting the stage for continued refinement and enhanced trustworthiness in AI-driven applications in biomedical setting.

7 MIMIC-IV-Ext-BHC

To accurately replicate the original study on hospital course summarization using Large Language Models (LLMs), a structured approach was implemented. The clinical summarization pipeline systematically explored and optimized the use of Large Language Models (LLMs) Clinical-T5, LLaMA2-13B, and GPT-4 for generating concise and accurate summaries of the Brief Hospital Course (BHC) sections extracted from the MIMIC-IV dataset[33][34][21][1].

As the BHC dataset is confidential and available only to credentialed users upon completing 'CITI Data or Specimens Only Research' certification [35]. Once acquiring the dataset, data preprocessing involves regex-based BHC section extraction using multiple pattern matching strategies, clinical text normalization (removing whitespace, bracketed content, underscores, symbols) are performed along with column validation which ensured data integrity with source text > 20 characters and target text > 10 characters filtering also removing null and empty entries improves the data quality. The 95:5 train/test split was satisfied across all bins to evaluate the results efficiently. The bins are classified into short (≤ 1024 tokens), medium (1025–2048 tokens) and long (> 2048). Then using binned dataset a separate dataset for each context bin using HuggingFace dataset to convert pandas DataFrames to dataset objects was created and a validation is performed to ensure each bin has both training and test examples. This rigorous preprocessing ensure to produce high quality and standardized datasets segmentation into bins enables precise and targeted model training and evaluation using Unsloth framework for memory-efficient operations and FLAN-T5 tokenizer for context length determination[36].

The implementation leveraged cutting edge technical optimizations including 4-bit quantization (`load_in_4bit=True`) reducing the memory footprint by $\sim 75\%$, Unsloth framework integration for accelerated training, and comprehensive cache management across HuggingFace, WandB and Triton directories[37][38]. The system required NVIDIA GPU with 24GB+ VRAM, 32GB system RAM and SLURM compatible high performance computing infrastructure. Environment configuration included CUDA capability detection, mixed precision training and reproducible random seed management across NumPy, PyTorch and CUDA backends. The pipeline incorporated comprehensive logging and monitoring through WandB integration, SLURM batch script compatibility for cluster deployment and memory optimization techniques including checkpointing, 8-bit optimizer, and intelligent caching mechanisms.

Multiple adaptation strategies were applied to the models in the paper but for simplicity we have only used Quantized Low-Rank Adaptation (QLoRA) fine tuning for domain adapted performance, zero-shot baseline for raw model capability and prefix prompting for instruction guided performance[39][40][41]. The technical implementation utilized specific LoRA configurations targeting attention and MLP layers [`"q_proj"`, `"k_proj"`, `"v_proj"`, `"o_proj"`, `"gate_proj"`, `"up_proj"`, `"down_proj"`] with rank= 16, alpha= 16, and zero dropout for maximum learning capacity. Explored critical parameters such as learning rate, batch size, LoRA rank (`lora_r`), LoRA alpha (`lora_alpha`), dropout rates, warmup ratios, and weight decay to enhance model performance as part of hyperparameter tuning. Clinical-T5 performed best with a moderately high learning rate ($\sim 5 \times 10^{-4}$), a batch size of 8, LoRA rank and alpha at 32, dropout at 0.1, warmup ratio at 0.1, and weight decay at 0.01, whereas LLaMA2-13B required a lower learning rate ($\sim 2 \times 10^{-4}$), smaller batch size of 4, a higher LoRA rank (64) paired with a lower alpha (16), dropout at 0.1.

The complete pipeline execution required over 8+ hours(just for short context bin) of intensive computation including model loading, for context specific training across three bins (short/medium/long) took even longer as the dataset is huge across all three bins with comprehensive inference across multiple adaptation strategies, and quantitative evaluation using BERT score metrics[42]. This extended runtime reflects the computational complexity of training large language models with parameter efficient techniques while maintaining production quality standards.

Quantitative evaluation using BERT Score demonstrated substantial benefits from QLoRA fine-tuning with LLaMA2-13B achieving the highest scores among open-source models even for the contexts in the short bins. LLaMA2-13B's performance degraded with longer inputs revealing limitations in managing extensive clinical texts but GPT-4 consistently maintained robust performance across all context lengths without finetuning showcasing its generalization capabilities. According to the paper the clinicians strongly preferred GPT-4 for summaries over QLoRA LLaMA2-13B summaries and original clinician written summaries across all four criteria (comprehensiveness, conciseness, factual correctness, and fluency). The results achieved trying to replicate the paper as shown in the table below supports the clinicians' claim.

Table 7.1: BERT F1-Score Results for Clinical Note Summarization Models and Adaptation Strategies

Model	**Adaptation Strategy**	**BERT F1-Score**
Clinical-T5-Base	Zero-shot Prompting	0.584
	Prefix Prompting	0.601
	QLoRA Fine-tuning	0.647
LLaMA2-13B-Chat	Zero-shot Prompting	0.612
	Prefix Prompting	0.629
	QLoRA Fine-tuning	0.683
GPT-4	Zero-shot Prompting	0.667
	Prefix Prompting	0.673

The comprehensive clinical summarization pipeline successfully demonstrates the practical deployment of state-of-the-art (SOTA) LLMs in healthcare setting through advanced technical optimizations, memory-efficient training strategies, and production ready infrastructure. The 8+ hours (just for short context bin) computational investment resulted in valuable insights into parameter efficient fine-tuning effectiveness, context based model performance and showed the generalization capabilities of proprietary models like GPT-4. These findings emphasize the importance of combining precise hyperparameter tuning with qualitative evaluation methods to achieve clinically relevant NLP solutions. Future work should focus on further refining long-context performance and integrating ensemble strategies to leverage complementary strengths of multiple LLMs.

Bibliography

[1] A. Aali, D. Veen, Y. I. Arefeen, *et al.*, "A dataset and benchmark for hospital course summarization with adapted large language models," *Journal of the American Medical Informatics Association*, 2024. DOI: 10.1093/jamia/ocae312.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001. DOI: 10.1038/scientificamerican0501-34. [Online]. Available: https://doi.org/10.1038/scientificamerican0501-34.

[3] G. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998. DOI: 10.1109/jproc.1998.658762.

[4] National Human Genome Research Institute, *The human genome project*, https://www.genome.gov/human-genome-project, Accessed: 2025-06-17, 2020.

[5] T. Seymour, D. Frantsvog, and T. Graeber, "Electronic health records (ehr)," *American Journal of Health Sciences (AJHS)*, vol. 3, no. 3, p. 201, 2018. DOI: 10.19030/ajhs.v3i3.7139.

[6] W3.org, *Rdf primer*, https://www.w3.org/TR/rdf-primer/, Accessed: 2025-06-17, 2014.

[7] W3.org, *Owl web ontology language overview*, http://www.w3.org/TR/owl-features/, Accessed: 2025-06-17, 2025.

[8] H. Wu and A. Yamaguchi, "Semantic web technologies for the big data in life sciences," *BioScience Trends*, vol. 8, no. 4, pp. 192–201, 2014. DOI: 10.5582/bst.2014.01048. [Online]. Available: https://doi.org/10.5582/bst.2014.01048.

[9] LinkedDataBook.com, *Linked data: Evolving the web into a global data space*, http://linkeddatabook.com/editions/1.0/, Accessed: 2025-06-17, n.d.

[10] W3.org, *Sparql query language for rdf*, https://www.w3.org/TR/rdf-sparql-query/, Accessed: 2025-06-17, 2013.

[11] A. Patel and S. Jain, "Present and future of semantic web technologies: A research statement," *International Journal of Computers and Applications*, vol. 43, no. 5, pp. 1–10, 2019. DOI: 10.1080/1206212x.2019.1570666. [Online]. Available: https://doi.org/10.1080/1206212x.2019.1570666.

[12] O. Bodenreider, "The unified medical language system (umls): Integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 267D–270, 2004. DOI: 10.1093/nar/gkh061.

[13] U.S. National Library of Medicine, *Umls metathesaurus browser*, https://uts.nlm.nih.gov/uts/, Accessed: 2025-06-23, n.d.

[14] M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, and J. Li, "An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation," in *Open Publications Of UTS Scholars (University of Technology Sydney)*, 2019. DOI: 10.1109/cec.2019.8790259.

[15] I. Kohane, S. Churchill, and S. Murphy, "A translational engine at the national scale: Informatics for integrating biology and the bedside," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 181–185, 2012. DOI: 10.1136/amiajnl-2011-000492.

[16] S. Minaee, T. Mikolov, N. Nikzad, *et al.*, "Large language models: A survey," *arXiv*, 2024. DOI: 10.48550/arxiv.2402.06196. arXiv: 2402.06196 [cs.CL].

[17] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 1–20, 2024. DOI: 10.1109/tkde.2024.3352100.

[18] Arxiv.org, "Predicting the performance of black-box llms through self-queries," *arXiv*, 2022, Accessed: 2025-06-23. arXiv: 2501.01558v1 [cs.CL]. [Online]. Available: https://arxiv.org/html/2501.01558v1.

[19] L. Huang, Y. Yang, W. Ma, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv*, 2023. DOI: 10.48550/arxiv.2311.05232. arXiv: 2311.05232 [cs.CL].

[20] K. Aitken, V. Ramasesh, Y. Cao, and N. Maheswaranathan, "Understanding how encoder-decoder architectures attend," *arXiv*, 2021. DOI: 10.48550/arXiv.2110.15253. arXiv: 2110.15253 [cs.CL].

[21] OpenAI, "Gpt-4 technical report," *arXiv*, 2023. DOI: 10.48550/arxiv.2303.08774. arXiv: 2303.08774 [cs.CL].

[22] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv*, 2023. arXiv: 2201.11903 [cs.CL]. [Online]. Available: https://arxiv.org/pdf/2201.11903.

[23] OpenAI, *Chatgpt*, https://chatgpt.com/, 2025.

[24] A. Hogan, E. Blomqvist, M. Cochez, *et al.*, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2022. DOI: 10.1145/3447772.

[25] Z. Zhao, Q. Jin, F. Chen, T. Peng, and S. Yu, "A large-scale dataset of patient summaries for retrieval-based clinical decision support systems," *Scientific Data*, vol. 10, no. 1, 2023. DOI: 10.1038/s41597-023-02814-8.

[26] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," *arXiv*, 2019. DOI: 10.48550/arxiv.1910.12840. arXiv: 1910.12840 [cs.CL].

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, 2001. DOI: 10.3115/1073083.1073135.

[28] K. Ganesan, "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks," *arXiv*, 2018. DOI: 10.48550/arxiv.1803.01937. arXiv: 1803.01937 [cs.CL].

[29] E. Durmus, H. He, and M. Diab, "Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of ACL 2020*, 2020. DOI: 10.18653/v1/2020.acl-main.454.

[30] T. Scialom, P.-A. Dray, P. Gallinari, *et al.*, "Questeval: Summarization asks for fact-based evaluation," *arXiv*, 2021. DOI: 10.48550/arxiv.2103.12693. arXiv: 2103.12693 [cs.CL].

[31] S. Feng, V. Balachandran, Y. Bai, and Y. Tsvetkov, "Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge," *arXiv*, 2023. DOI: 10.48550/arxiv.2305.08281. arXiv: 2305.08281 [cs.CL].

[32] Z. Ji, Z. Liu, N. Lee, *et al.*, "Rho: Reducing hallucination in open-domain dialogues with knowledge grounding," in *Findings of ACL 2023*, 2023. DOI: 10.18653/v1/2023.findings-acl.275.

[33] E. Lehman and A. Johnson, *Clinical-t5: Large language models built using mimic clinical text*, https://www.physionet.org/content/clinical-t5/1.0.0/, 2023.

[34] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv*, 2023. DOI: 10.48550/arxiv.2302.13971. arXiv: 2302.13971 [cs.CL].

[35] physionet.org, *Citi course instructions*, https://physionet.org/about/citi-course/, Accessed: 2024-06-23, n.d.

[36] Unsloth - Open source Fine-tuning for LLMs, *Unsloth - open source fine-tuning for llms*, https://unsloth.ai/, 2024.

[37] Hugging Face, *Hugging face – on a mission to solve nlp, one commit at a time*, https://huggingface.co/, 2024.

[38] wandb.ai, *Weights & biases – developer tools for ml*, https://wandb.ai/site.

[39] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv*, 2023. DOI: 10.48550/arxiv.2305.14314. arXiv: 2305.14314 [cs.CL].

[40] Y. Li, "A practical survey on zero-shot prompt design for in-context learning," in *Proceedings*, 2023. DOI: 10.26615/978-954-452-092-2_069.

[41] X. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv*, 2021. DOI: 10.48550/arxiv.2101.00190. arXiv: 2101.00190 [cs.CL].

[42] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv*, 2019. DOI: 10.48550/arxiv.1904.09675. arXiv: 1904.09675 [cs.CL].