Maastricht University

Department of Advanced Computing Sciences

Project Proposal

KGEMMA: Knowledge Graph for Enhanced Model Management in AI

Salvatore Pascarella

I6404117

s.pascarella@student.maastrichtuniversity.nl

Course: Building and Mining Knowledge Graphs

February 2025

1 Significance

The AI ecosystem is undergoing an unprecedented explosion of models and research. As of February 2025, Hugging Face, the largest open repository for machine learning, hosts 1.44 million models, including 180,000 dedicated to text generation. This represents a 90-fold increase in just two years, up from 16,000 text-generation models in mid-2023 [Gao and Gao, 2023]. Concurrently, AI research papers have surged: publications mentioning LLMs, skyrocketed from 40 in 2018 to nearly 30,000 in 2024 [Naveed et al., 2024], while total AI publications tripled from 2010–2022 [Maslej et al., 2023]. This rapid growth presents a critical fragmentation problem—thousands of models and research papers are released every month, making it increasingly difficult to systematically track, categorize, and connect models to their corresponding research. [Jiang et al., 2023] investigated the practices and challenges associated with reusing pre-trained models (PTMs) within the Hugging Face ecosystem. Through interviews with 12 practitioners, the authors identify issues such as missing attributes, discrepancies between claimed and actual performance, and model risks. Similarly, the lack of a structured evaluation system leads to inefficiencies in model selection and benchmarking [Sanseviero and Kiela, 2023]. Researchers and practitioners lack a structured knowledge base that links new models with the original papers introducing them, struggling to identify relevant models, reproduce results, and ensure efficient progress in AI research. The reproducibility crisis in AI research is a significant concern, as fragmented documentation and opaque methodologies hinder validation efforts. A 2020 survey revealed that over 50% of researchers believe there is a significant reproducibility crisis in science, with AI and ML fields being notably affected [Gundersen and Kjensmo, 2020]. Furthermore, [Smith and Doe, 2024] found out that inconsistent terminology and reporting standards further exacerbate the issue, leading to confusion and impeding the replication of studies, making AI research lose credibility due to irreproducible results. This project aims to develop a structured knowledge graph that systematically links machine learning models to their associated benchmarks and corresponding research papers, enabling the practitioners to compare models across standardized evaluations and identify the most suitable solutions for specific application while keeping traceability and reproducibility through the connection to their research papers. This will accelerate research cycles, improve decision-making, and foster a more accessible and transparent AI ecosystem.

2 Related Work

Several efforts have been made to integrate machine learning models with their associated benchmarks, information, and research papers, but none fully address the comprehensive integration of models, papers, and benchmarks. Hugging Face provides a platform where researchers and developers can host and share machine learning models and datasets. While it offers extensive collections, the linkage between models, their evaluation metrics, and corresponding research papers is not systematically structured. Users often need to manually navigate between models and datasets, and the associated evaluation metrics are not consistently documented. Additionally, while some models reference their research papers, this connection is not uniformly maintained across the platform. A curated, ontology-based knowledge graph for AI research [Blagec et al., 2022] provides structured relationships between AI concepts, benchmarks, and tasks. However, it does not systematically integrate model-specific metadata from platforms such as Hugging Face, nor does it directly connect models to their associated datasets and research papers. My approach fills this gap by explicitly linking models, datasets, evaluation metrics, and research publications to improve reproducibility, traceability, and model selection. OGB [Hu et al., 2021] offers a diverse set of benchmark datasets to facilitate robust and reproducible research in graph-based machine learning . While it provides a unified evaluation protocol, it does not address the fragmentation of model metadata across platforms like Hugging Face. My project differs by structuring ML model information within a knowledge graph, explicitly linking models to their training datasets, performance metrics, and related research.

3 Goal and specific objectives

The main objective of this project is to build a structured knowledge graph that interconnects metadata from multiple machine learning models, including evaluation metrics, datasets, to their associated research papers. By systematically linking these elements, the knowledge graph will enable researchers and developers to efficiently explore models, compare their performance, and extract actionable insights. The overarching goal is to accelerate AI research and mitigate information fragmentation, ensuring that valuable knowledge is easily accessible and well-structured, rather than being siloed across different platforms. By addressing key challenges in AI knowledge fragmentation mentioned in Section 1, this project serves as a practical implementation of knowledge graphs in AI research, supporting efficient AI model selection and evaluation. The specific objectives of this project are as follows:

- Collection and extraction of unstructured metadata from models, including evaluation metrics, datasets, base models, through the Hugging Face API. This also involves retrieving research papers using the arXiv API to enrich paper-related entities.
- Design and implementation of a structured schema that organizes models, datasets, evaluation metrics, and papers into a graph-based representation, ensuring efficient interconnectivity.

- Development of a queryable knowledge graph that allows users to search, retrieve, and analyze models based on evaluation metrics, dataset usage.

4 Methodology

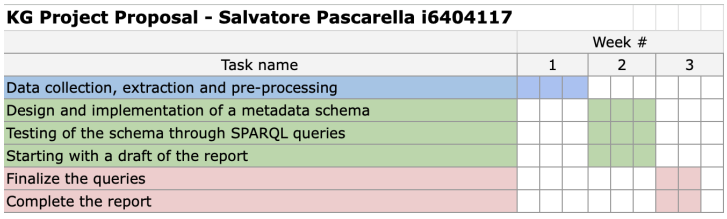
The methodology involves data collection, processing, graph construction, querying, and insights extraction, ensuring that AI model metadata is structured and easily retrievable. In order to complete the steps mentioned in 3 I will use the following technologies:

- The Hugging Face and arXiv APIs to extract the most meaningful ML metadata and research papers info respectively. Extracted data will be structured in a tabular format using Pandas, allowing for easier transformation into graph-based representations.
- Wherever possible, existing semantic web ontologies will be reused to define entities and relationships, minimizing redundancy. However, since specific AI model-related ontologies are limited, some entities and properties will be created from scratch. Using RDFlib, the data will be converted into RDF triples, creating structured links between the entities. The resulting knowledge graph will possibly be stored in a graph database (GraphDB or Neo4j), ensuring efficient storage.
- SPARQL will be used to query the knowledge graph, enabling the extraction of meaningful insights.

One of the primary risks is incomplete metadata from Hugging Face, where many models lack critical details such as evaluation metrics, dataset references, or links to research papers. This could limit the quality and size of the knowledge graph. To address this, extensive data cleaning and selection steps will be performed to ensure that the majority of the entities contain meaningful metadata. Additionally, priority will be given to models (e.g. text-generation/classification) with richer metadata, ensuring that the core of the knowledge graph remains well-structured. Another potential challenge is ontology limitations, as existing ontologies may not fully cover ML-specific metadata. While the preference will be to reuse established ontologies wherever possible, it is likely that new entities and relationships will need to be defined. To mitigate this, custom extensions to existing ontologies will be implemented only when necessary, ensuring compatibility with broader AI knowledge representation efforts. A further limitation is the time constraints associated with research paper processing. Extracting deep insights from research papers (such as detailed methodologies or architectures) would require advanced NLP-based processing, which is beyond the project's time scope. To mitigate this, only essential bibliographic metadata (such as title, abstract, and authorship) will be extracted from arXiv. This ensures that paper-to-model connections are still well established, while keeping the data processing within manageable limits. For future work, I will enrich the paper-related entities by applying some NLP techniques. Additionally, a larger variety of models will be included in order to cover as many tasks as possible and also, more info about the used datasets will be added by leveraging PapersWithCode API which offers a structured repository datasets.

5 Milestones and Deliverables

Due to time constraints, every task has a specific milestone as shown in the following Gantt chart. Github will be used to keep track of the advancements.



6 Anticipated Results

The main output of this project will be a metadata schema along with a set of queries that demonstrate how to leverage the structured knowledge base for gaining a clearer understanding of model performance trends, dataset-model relationships, and research impact. By addressing the key issue of the data fragmentation across multiple sources, researchers will benefit from faster access to model performance data and improved reproducibility tracking, while developers will be able to efficiently compare models for real-world applications.

References

[Blagec et al., 2022] Blagec, K., Barbosa-Silva, A., Ott, S., Fabien, M., Kühberger, J., Vollrath, M., Theis, F., and Baumbach, J. (2022). A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. *Scientific Data*, 9:322.

[Gao and Gao, 2023] Gao, S. and Gao, A. K. (2023). On the origin of llms: An evolutionary tree and graph for 15,821 large language models.

[Gundersen and Kjensmo, 2020] Gundersen, O. E. and Kjensmo, S. (2020). The reproducibility crisis in ai and machine learning. *AI Magazine*, 41(3):49–58.

[Hu et al., 2021] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. (2021). Open graph benchmark: Datasets for machine learning on graphs.

[Jiang et al., 2023] Jiang, W., Synovic, N., Hyatt, M., Schorlemmer, T. R., Sethi, R., Lu, Y.-H., Thiruvathukal, G. K., and Davis, J. C. (2023). An empirical study of pre-trained model reuse in the hugging face deep learning model registry.

[Maslej et al., 2023] Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. (2023). Artificial intelligence index report 2023.

[Naveed et al., 2024] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2024). A comprehensive overview of large language models.

[Sanseviero and Kiela, 2023] Sanseviero, O. and Kiela, D. (2023). Announcing evaluation on the hub. Accessed: February 2025.

[Smith and Doe, 2024] Smith, J. and Doe, J. (2024). Inconsistent terminology and reporting standards in ai research. *arXiv preprint arXiv:2407.10239*.