GC-DAM: Graph and Contextual Embeddings for Heterogeneous Data Asset Matching

Maximilian Stäbler[1], Markus Lange[1], Samir Kipper[1], Chris Langdon[2] and Frank Köster[1]

[1] *German Aerospace Center (DLR) - Institute for AI Safety & Security, Germany*

[2] *Drucker School of Business, Claremont Graduate University, USA*

Abstract

Data assets—such as datasets, data services, APIs, algorithms, and analytical models—are valuable digital resources that organizations use to create value, support decision-making, and optimize business processes. Matching and integrating these assets, despite differences in semantic languages, ontologies, or schemas, is essential for building scalable and interoperable dataspaces. However, existing approaches often focus solely on semantic similarities, overlooking structurally similar assets from other domains that could be highly relevant. To address this gap, we present **G**raph and **C**ontextual Embeddings for Heterogeneous **D**ata **A**sset **M**atching (GC-DAM). GC-DAM employs two embedding strategies to match data assets based on both semantic and structural attributes. Structural (morphological) features are automatically incorporated into a knowledge graph, enabling the identification of assets that are structurally similar to a query but may originate from different domains, while metadata descriptions capture the semantic (contextual) features. This dual approach overcomes the limitations of methods that rely solely on semantic descriptions. We validate our approach against a custom dataset of 10,000 Kaggle data assets. Our multimodal embedding achieves 77% agreement on our custom dataset, demonstrating its ability to identify structurally similar assets across diverse domains, even when they are semantically different. The dataset and code are publicly available to the research community.

Keywords

Multi-Modal-Embedding, Heterogeneous dataspaces, Knowledge Graphs, Automated Interoperability, LLM

This paper addresses (Issue #2) of the W3C Dataspaces Community Group.

1. Introduction

Dataspaces have emerged as a pivotal concept in the European dataspace, driven by initiatives such as the European Data Governance Act (DGA) and the adoption of the FAIR Data Principles. These frameworks aim to foster trust, ensure data sovereignty, and promote seamless data sharing across diverse sectors, including research, business, and public services [1, 2]. Gaia-X (https://gaia-x.eu/) exemplifies the potential of dataspaces to enable collaborative value creation by integrating heterogeneous datasets under common standards. However, achieving semantic interoperability remains a critical challenge for their effective implementation.

Consider the following scenario: a mobility researcher needs to find traffic flow datasets to develop a congestion prediction model for a smart city project. In one dataspace, a municipality describes their dataset as "Urban Traffic Intensity Measurements" with metadata focusing on sensor locations and sampling frequency. In another, a similar dataset exists but is described as "City Vehicular Movement Analytics" emphasizing the analytical methods applied to the raw data. In a third, a dataset labeled "Metropolitan Transportation Metrics" documents comparable data but uses industry-specific terminology. Despite all three datasets containing structurally compatible traffic flow data that could benefit the researcher, purely semantic search methods would likely identify only one or two of these resources, missing potentially valuable data assets due to terminology discrepancies-a common challenge in the absence of globally established standards for data description.

The Third International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, June 01, 2025, Portorož, Slovenia

✉ maximilian.staebler@dlr.de (M. Stäbler)
0000-0003-1311-3568 (M. Stäbler); 0000-0001-7198-2871 (M. Lange)

The Semantic Web community has long contributed to addressing semantic heterogeneity through RDF-based solutions and ontologies. Yet, as dataspaces expand in scale and complexity, new methodologies are required to overcome challenges such as vocabulary mismatches, structural heterogeneity, and contextual divergence [3, 4]. These issues are particularly pronounced in federated environments where stakeholders operate under varying schemas and standards. Automating the discovery of similar concepts and data assets is essential for managing the growing complexity of dataspaces. Systems must identify and compare data assets accurately, even when expressed in different semantic languages, ontologies, or schemas, enabling seamless use across platforms. Despite progress, challenges remain, including the scalability of ontology-based solutions in dynamic ecosystems and the computational complexity of reasoning across conflicting ontologies. Nonetheless, ongoing research continues to advance semantic interoperability through refined terminological relationships (e.g., synonyms, hyponyms, hypernyms) and improved real-time context alignment, offering hope for resolving these issues in the near future.

To address these challenges, this paper introduces the GC-DAM (Graph and Contextual Embeddings for Heterogeneous Data Asset Matching), a framework designed to enhance semantic interoperability and facilitate efficient data asset discovery within dataspaces. The key research questions we address are:

1. How can we effectively match heterogeneous data assets across domains when semantic descriptions alone are insufficient? 2. Can the combination of structural and semantic embeddings overcome vocabulary mismatches and structural heterogeneity in dataspaces? 3. What are the optimal techniques for integrating graph-based structural representations with contextual semantic embeddings to maximize matching accuracy?

Our contributions include:

- A novel multimodal embedding approach that combines graph-based structural embeddings with contextual semantic embeddings to match heterogeneous data assets
- A framework that automatically extracts entities and relationships from metadata to build knowledge graphs that capture structural similarities
- An evaluation methodology demonstrating that our approach achieves superior matching performance compared to using either embedding type alone

The unique value of combining structural and semantic embeddings lies in their complementary strengths. While semantic embeddings excel at capturing thematic and contextual similarities based on textual descriptions, they often miss structurally compatible assets that use different terminology or conceptual frameworks. Structural embeddings, conversely, identify assets with similar organizational patterns, entity relationships, and morphological characteristics, regardless of domain-specific vocabulary. By leveraging both perspectives simultaneously, GC-DAM can discover relevant assets that would remain hidden to approaches relying solely on semantic matching, significantly expanding the pool of potential resources available to users in heterogeneous dataspaces.

GC-DAM directly addresses key challenges outlined in the W3C Dataspaces Community Group, particularly *Issue #2: Data Discovery*[1]. By leveraging multimodal embedding techniques and advanced semantic matching algorithms, GC-DAM enables the identification of structurally and semantically similar data assets across heterogeneous datasets. This capability is essential for discovering relevant resources within dynamic and distributed dataspaces. Furthermore, GC-DAM aligns especially with the "*F*" in the FAIR principles by ensuring that data assets are findable. To illustrate its potential applicability, we consider challenges that might arise in environments like the Gaia-X. In such scenarios, where diverse stakeholders share and access data assets, GC-DAM could address issues related to semantic alignment and integration. By enabling the accurate discovery of relevant data assets and ensuring seamless semantic interoperability, GC-DAM provides scalable solutions that help navigate the complexities of dataspaces. Additionally, its modular design promotes reusability and standardization, making it adaptable to a wide range of dataspace implementations.

[1] https://github.com/w3c-cg/dataspaces/issues/2

This work contributes to advancing the expressiveness and integration of semantic technologies in dataspace architectures. By fostering collaboration between researchers and practitioners at workshops such as Semantics in Dataspaces (SDS) 2025, we aim to drive innovation in this critical domain.

2. Related Work

To address our research questions on effective heterogeneous data asset matching through multimodal embeddings, we examine three key research areas: approaches for entity matching in heterogeneous environments, embedding techniques for structural and semantic representations, and methods for integrating multiple embedding modalities.

Heterogeneous Data Asset Matching

Traditional approaches to data integration, such as Extract-Transform-Load (ETL) systems and rule-based methods, often struggle with scalability and adaptability in dynamic federated environments [5, 1]. This is particularly evident in dataspaces with diverse formats ranging from structured databases to unstructured multimedia content [6]. Ontology-based matching approaches have long dominated the field, with systems like LogMap using lexical matching, graph structure matching, and logical reasoning to identify correspondences [7]. While these methods offer high semantic precision and formal consistency guarantees, they face several limitations that our approach addresses. First, they suffer from computational bottlenecks when applied to large-scale datasets [8]. As noted in [9], traditional ontology matching "severely lacks performance when dealing with large matching problems." Second, they struggle with understanding textual variations and nuanced semantics across domains, often relying on exact lexical matches or predefined synonyms.

Recent partition-based matching approaches like COMA++ and Falcon attempt to address scalability issues by dividing large ontologies into manageable partitions [7]. However, these methods have significant limitations that our multimodal embedding approach overcomes. COMA++ uses "relatively simple heuristic rules to partition the input schemas resulting often in too few or too many partitions" and relies on limited information about partitions (only root nodes) to determine similarity. In contrast, our GC-DAM framework leverages comprehensive structural information through graph embeddings, capturing deeper relationships between entities.

Semantic and Structural Embedding Methods

Recent advancements in embedding-based techniques demonstrate promise in addressing heterogeneous matching challenges. Large Language Models (LLMs) have emerged as powerful tools for generating semantic embeddings that capture contextual relationships between data entities [10, 7]. These semantic approaches excel at capturing thematic similarities and domain-specific vocabulary but often miss structurally compatible assets that use different terminology. For semantic embeddings, approaches range from basic word vector averaging to sophisticated contextual models. Word embedding methods have been applied to ontology alignment with moderate success, with hybrid approaches incorporating string-based similarity and semantic vector similarity showing improved performance [11]. However, as noted in [12], these methods primarily capture lexical similarity and struggle with cross-domain vocabulary differences-a limitation our multimodal approach explicitly addresses. Structural embedding methods, conversely, focus on graph-based representations of data assets. Geometric modeling approaches like JOIE transform RDFS ontologies into view graphs and model entities as vectors or shapes (e.g., Concept2Box representing concepts as boxes) [13]. While effective for capturing relationships within a domain, these methods typically operate independently of semantic embeddings, missing opportunities for complementary information. Our approach diverges by using Node2Vec's biased random walk strategy to capture both local and global graph structures while complementing this with rich semantic information.

Multimodal Embedding Integration

The integration of multiple embedding modalities for comprehensive data representation remains an underdeveloped area. Existing multimodal approaches primarily focus on integrating different data types (e.g., text and images) rather than different representation perspectives of the same data [14]. As highlighted in [15], multimodal embedding models facilitate "the integration of diverse data types into a unified vector space," enabling "seamless cross-modality vector similarity searches." Current multimodal models like CLIP, ImageBind, and visualBERT integrate visual and textual information but target cross-modal retrieval rather than complementary perspectives on structured data [16]. These approaches show "a generalized advantage of multimodal representations over language-only ones on concrete word pairs, but not on abstract ones," indicating domain-specific performance variations that must be considered in data asset matching.Unlike these approaches, our GC-DAM framework uniquely combines semantic embeddings (capturing contextual meaning) with structural embeddings (capturing morphological relationships) to provide a comprehensive representation of data assets. While methods like JOIE and EmbedS attempt to bridge knowledge graphs and ontologies, they typically use a single embedding strategy with different targets rather than truly multimodal representations [8].

In summary, while significant progress has been made in entity resolution, ontology alignment, and data discovery using various embedding techniques, existing approaches typically focus on either semantic or structural aspects in isolation. GC-DAM distinguishes itself by integrating both perspectives into a unified framework that captures comprehensive similarity across heterogeneous dataspaces. This novel integration enables identification of matches that would be missed by purely semantic or structural approaches alone, addressing a critical gap in current dataspace interoperability solutions.

3. Methodology

This section details our approach to matching and aligning heterogeneous data assets across diverse environments. By employing multimodal embeddings, we establish interoperability in complex dataspaces, enabling identification and integration of similar assets across domains.

We define three key concepts central to our approach:

- **Semantic Embeddings:** Dense vector representations capturing contextual meaning, thematic content, and semantic relationships within data asset descriptions and metadata.
- **Structural Embeddings:** Vector representations derived from knowledge graphs encoding morphological characteristics and entity relationships independent of specific terminology.
- **Multimodal Embeddings:** The integrated combination of both semantic and structural embedding spaces for comprehensive similarity assessment.

To illustrate, consider a traffic dataset: semantic embeddings capture that it contains "vehicle flow measurements on urban roads," while structural embeddings identify its temporal sequence format and geospatial reference structure. A dataset using different terminology (e.g., "metropolitan transit analytics") with similar structural patterns would be recognized through multimodal matching despite semantic differences.

Figure 1 provides an overview of our process in three steps: (a) heterogeneous data assets representation, (b) transformation into distinct structural and semantic embeddings, and (c) clustering matching datasets using cosine similarity measures. The output, termed a "complex information object" (CIO), represents query results by distinguishing between semantic and contextual representations [17].

3.1. Semantic Embedding – Contextual Metadata

For semantic representation, we incorporate rich contextual metadata capturing domain-specific insights critical for accurate data interpretation. This includes functional descriptions, column annotations,

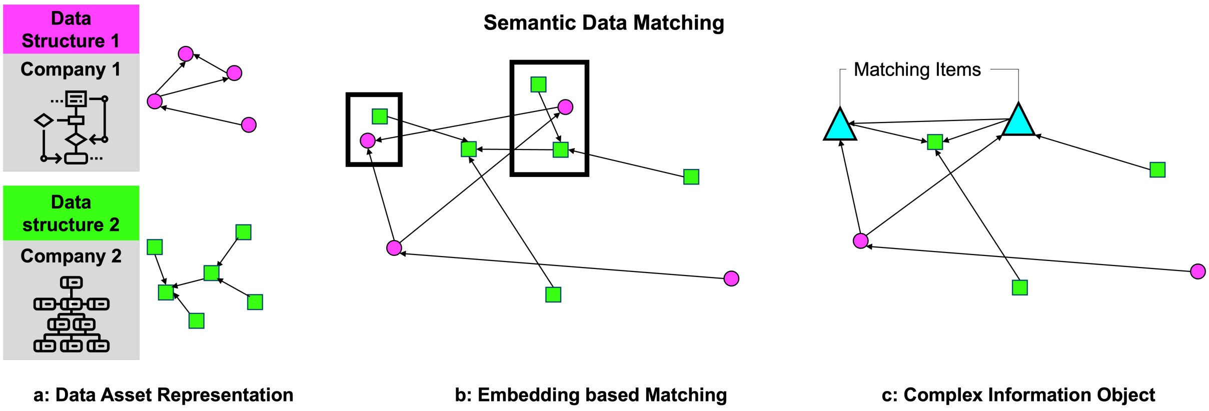


Figure 1: A multimodal embedding-based semantic matching process illustrated in three stages. (**a**) Data Asset Representation: Two distinct data structures from different organizations are visualized as graphs, where nodes represent entities and edges denote relationships. (**b**) Embedding-Based Matching: Semantic and structural embeddings are generated for each graph, enabling the identification of similarities between data assets across heterogeneous sources. (**c**) Complex Information Object: Matched items are aggregated into a unified representation, capturing both semantic coherence and structural alignment to facilitate integration of datasets, data services, or interconnected data assets.

and API interface details transformed into high-dimensional embeddings using the *stella_en_1.5B_v5*[2] (S_EN) model. Built on Alibaba's GTE architecture [18], S_EN is optimized for semantic similarity tasks with several key adaptations:

- **Embedding Dimensionality:** 1536-dimensional output space balancing representational capacity and efficiency
- **Specialized Loss Function:** Enhanced contrastive learning with hard negative mining
- **Domain Adaptation:** Transfer learning capabilities for dataspace-specific terminology

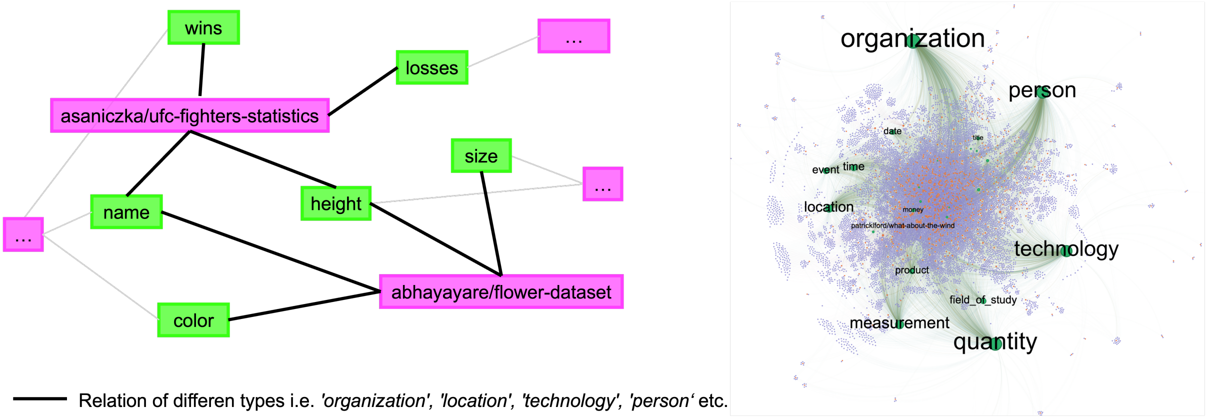
Metadata Integration and Preprocessing We systematically integrate five metadata categories: (1) Descriptive (titles, descriptions, keywords), (2) Structural (schemas, data types, relationships), (3) Administrative (authorship, version history), (4) Technical (quality metrics, update frequency), and (5) Domain-Specific annotations (sensor specifications for IoT data, geographic coordinates for spatial data). These elements undergo preprocessing through normalization (standardizing formats, resolving acronyms), noise reduction (removing redundant fields), domain-aware tokenization, and entity linking to standard ontologies where available. This preprocessing is adaptive to specific domains while maintaining cross-domain applicability.

Embedding Generation Process The S_EN model transforms preprocessed metadata into dense vector embeddings that capture semantic relationships beyond simple keyword matching, including conceptual associations, hierarchical relationships, functional similarities, and implicit contextual connections. These embeddings enable precise entity matching across diverse datasets through similarity computation (using cosine distance), cross-domain alignment, and contextual disambiguation. This approach enhances precision, reduces ambiguity, and enables integration of datasets that differ structurally but share similar contextual meanings.

3.2. Structural Embedding – Structure and Dependencies

To capture structural characteristics, we convert data assets into knowledge graphs (KGs). Nodes represent real-world entities (organizations, people, products) while edges denote relationships between

[2] https://huggingface.co/dunzhang/stella_en_1.5B_v5


wins
...
losses
asaniczka/ufc-fighters-statistics
size
...
name
height
...
abhayayare/flower-dataset
color
Relation of differen types i.e. 'organization', 'location', 'technology', 'person' etc.
organization
person
date
event time
location
money
patrickford/what-about-the-wind
technology
product
field_of_study
measurement
quantity

a: KG Construction

b: Resulting KG

Figure 2: a describe the resulting KG structure. Both the nodes and the edges have different types. *asaniczka/ufc-fighters-statistics* and *abhayayare/flower-dataset* are two data sets. The entities extracted from the metadata are shown in green. Edges model the relationships between the data sets and the entities. An exact overview is described in Section 4. **b** shows the resulting knowledge graph visualized in *Gephi*. The labels are shown in green, the data assets in orange, and the extracted entities in purple. The size of the nodes is determined by the degree. The edges are weighted by the number of occurrences.

them, capturing data element interconnections.

The process begins with entity extraction using the *GLiNER* model [19], a Named Entity Recognition system that frames NER as a matching problem in a shared latent space. We adapted GLiNER with:

- Custom entity types for dataspace-specific elements
- Improved span representation techniques using contextualized attention
- Specialized entity type embeddings aligned with common dataspace ontologies

GLiNER employs a DeBERTa-v3-large encoder to generate contextualized token embeddings aggregated into span representations. The span representation is computed as:

$$\mathbf{s}_{i,j} = \text{SpanAttn}(\mathbf{h}_i, \mathbf{h}_{i+1}, \ldots, \mathbf{h}_j) \tag{1}$$

where $\mathbf{h}_i$ through $\mathbf{h}_j$ are token embeddings and SpanAttn weights tokens based on contextual importance. After entity extraction, relationships are identified through dependency parsing and domain-specific rules. The resulting KG is transformed into embeddings using Node2Vec with the following configuration: The embedding was generated using the following parameters: a walk length of 80 steps per walk, 10 walks per node, a return parameter (p) of 1, an in-out parameter (q) of 0.5, a window size of 10, and 128 embedding dimensions. Before embedding generation, we include an Entity Resolution step to identify and merge duplicate entities, ensuring accurate and non-redundant structural representations.

Figure 2 shows an example KG structure and visualization where nodes represent entities color-coded by type, and edges indicate relationships. Node size reflects connection degree, highlighting central entities. These structural embeddings enable matching of data assets with similar organizational patterns regardless of terminology.

3.3. GC-DAM Framework

Figure 3 illustrates our procedure for generating and utilizing semantic and structural embeddings for data asset matching. While semantic embeddings (1536-dimensional) capture rich contextual information, structural embeddings (128-dimensional) capture morphological relationships like entity connectivity.

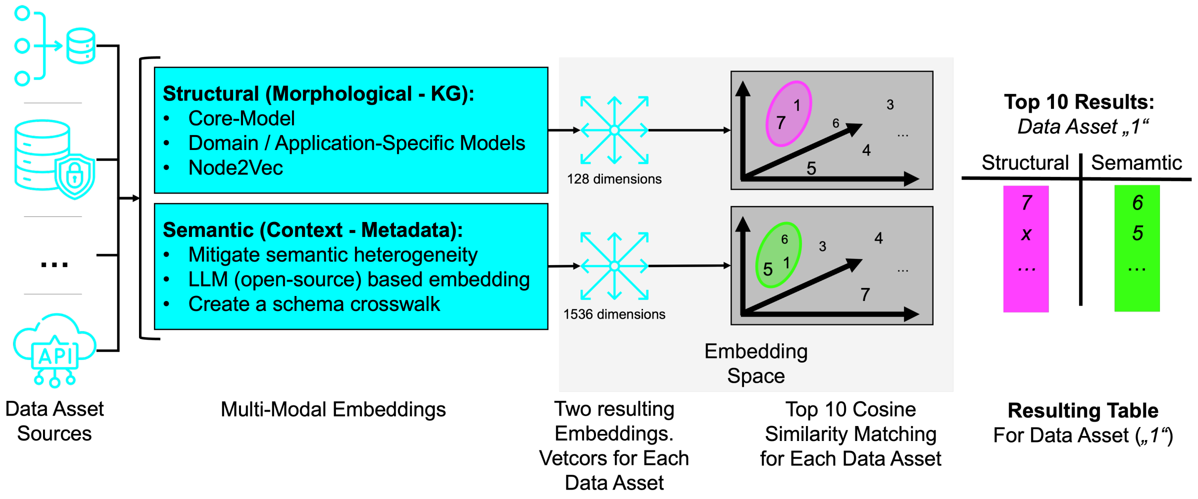

Structural (Morphological - KG):
• Core-Model
• Domain / Application-Specific Models
• Node2Vec
Semantic (Context - Metadata):
• Mitigate semantic heterogeneity
• LLM (open-source) based embedding
• Create a schema crosswalk
...
128 dimensions
1536 dimensions
1
7
6
3
4
5
...
6
5
1
3
4
7
...
Embedding Space
Top 10 Results:
Data Asset „1“
Structural
Semamtic
7
x
...
6
5
...
Data Asset Sources
Multi-Modal Embeddings
Two resulting Embeddings. Vetcors for Each Data Asset
Top 10 Cosine Similarity Matching for Each Data Asset
Resulting Table
For Data Asset („1“)

Figure 3: Visualization of the GC-DAM embedding generation and matching process. Structural embeddings are derived using the *Node2Vec* algorithm to capture morphological relationships in a 128-dimensional space, while semantic embeddings utilize the S_EN model to encode contextual metadata into a 1536-dimensional space. Cosine similarity is applied within each embedding space to identify the top 10 most similar data assets for a given query asset. The results are consolidated into a final table for comprehensive multimodal matching.

For each embedding space, we generate cosine similarity matrices representing pairwise distances between all data assets. For each query asset, we identify the top 10 most similar assets based on these scores, then consolidate results into a final table combining both similarity types. This ensures matches are both structurally aligned and semantically coherent.

Formally, let $\mathcal{D} = \{D_1, \ldots, D_n\}$ be a dataspace of data assets. Let $E_{\text{Descr}} : \mathcal{D} \to \mathbb{R}^{1536}$ denote the description embedding using the S_EN model processing titles, subtitles and descriptions. Let $E_{\{0,1\},\mathcal{M}} : \mathcal{D} \to G_{\theta,\mathcal{M}}$ be the graph embedding with boolean parameter θ and embedding model $\mathcal{M}$, where $G_{\theta,\mathcal{M}} = (V, E)_{\theta,\mathcal{M}}$ is the knowledge graph with vertices V and edges E. The structural embedding $E^{\theta,\mathcal{M}}_{\text{Struc}} : \mathcal{D} \to \mathbb{R}^{128}$ maps data assets to Node2Vec embeddings.

3.4. Similarity Measurement and Matching

We use cosine similarity to compare embedding vectors. For vectors $v, w \in \mathbb{R}^n$, the cosine similarity is:

$$S_C(v, w) = \frac{v \cdot w}{\|v\|\|w\|}$$

Given a search query q_{search}, we identify the ten most similar data assets in both embedding spaces:

$$\begin{aligned}
\mathfrak{D}_{\text{Descr}} = \{&\{D_1, \ldots D_{10}\} \subseteq \mathcal{D} : \\
&S_C(E_{\text{Descr}}(D_i), E_{\text{Descr}}(q_{\text{search}})) \geq S_C(E_{\text{Descr}}(D), E_{\text{Descr}}(q_{\text{search}})) \\
&\text{for all } D \in \mathcal{D} \setminus \{D_1, \ldots, D_{10}\}\} \\
\mathfrak{D}_{\text{Struc}} = \{&\{D_1, \ldots D_{10}\} \subseteq \mathcal{D} : \\
&S_C(E^{\theta,\mathcal{M}}_{\text{Struc}}(D_i), E^{\theta,\mathcal{M}}_{\text{Struc}}(q_{\text{search}})) \geq S_C(E^{\theta,\mathcal{M}}_{\text{Struc}}(D), E^{\theta,\mathcal{M}}_{\text{Struc}}(q_{\text{search}})) \\
&\text{for all } D \in \mathcal{D} \setminus \{D_1, \ldots, D_{10}\}\}
\end{aligned}$$

We then compare the entries of $\mathfrak{D}_{\text{Descr}}$ and $\mathfrak{D}_{\text{Struc}}$ to determine parameters for the structural embedding that yield the highest number of matching entries. This achieves comparable similarity between found data assets with respect to both semantic and structural representations. The structural approach captures similarity influenced by morphological aspects arising from the particular graph model, complementing the semantic approach's focus on contextual meaning.

The combined multimodal embedding approach enables identification of matches that would be missed by approaches relying on either embedding type alone, providing a comprehensive framework for heterogeneous data asset matching across diverse dataspaces.

4. Experiments

This section outlines the implementation and experiments conducted to evaluate the proposed method. We present the frameworks, algorithms, datasets, experimental scenarios, and performance metrics used. The goal is to demonstrate the method's effectiveness across different data environments and benchmark it against existing approaches. All implementations are available on GitHub[3]. The Kaggle-dataset can be downloaded here[4]. The Section covers the creation of our heterogeneous dataset for evaluating GC-DAM, followed by the implementation details, concluding with experiments on our Kaggle dataset.

4.1. Kaggle Dataset Evaluation

The evaluation of the Kaggle dataset focuses on assessing whether GC-DAM can effectively identify suitable matches within a highly heterogeneous and cross-domain dataset. These matches are not necessarily semantically similar but align with the query in terms of structure and context. The evaluation specifically aims to quantify the relevance of the top 10 matching results returned by GC-DAM for a given query dataset. The dataset is organized as a structured dataframe with multiple attributes describing each dataset. Key columns include:

- **id**: A unique identifier for each dataset, formatted as `username/dataset-slug`.
- **titleNullable**: The title of the dataset.
- **subtitle**: A descriptive subtitle providing additional context about the dataset.
- **description**: A detailed explanation of the dataset's contents and purpose.
- **usabilityRatingNullable**: A numerical rating (e.g., 0.941176) representing the dataset's usability.
- **keywords**: A list of tags or keywords associated with the dataset.
- **domains**: Categorical classifications indicating the dataset's domain (e.g., "CROSS_SECTOR", "EDUCATION").
- **licenses**: Information about the licensing terms for each dataset.
- **isPrivate**: A boolean value indicating whether the dataset is private or publicly accessible.

This structured metadata enables a comprehensive analysis of datasets, including their usability, domain classification, and descriptive content. The inclusion of categorical and textual attributes supports diverse applications such as trend analysis, domain-specific studies, and metadata-driven recommendations.

Figure 4 provides an overview of the distribution of datasets within the Kaggle dataset based on domain classification and keyword frequency. The top chart reveals that "CROSS-SECTOR" datasets dominate, indicating their broad applicability across multiple fields, while other domains like "SPORT," "HEALTH," and "FINANCE" also feature prominently. The logarithmic scale highlights the long-tail distribution of domain-specific datasets. The bottom chart focuses on keyword frequencies, showcasing common tags such as "business," "tabular," and "data visualization," which reflect popular themes in data science applications. Keywords like "classification," "image," and "computer science" further emphasize the diversity of dataset topics available on Kaggle. Together, these visualizations underscore the versatility and thematic richness of the platform's dataset collection. Table 1 presents the top 10 GC-DAM matching results based on semantic and structural embeddings of the *Sample Sales Data dataset* (part of the Kaggle dataset).

[3] https://github.com/maxistaebler/GC-DAM

[4] https://tinyurl.com/ty8xvzte

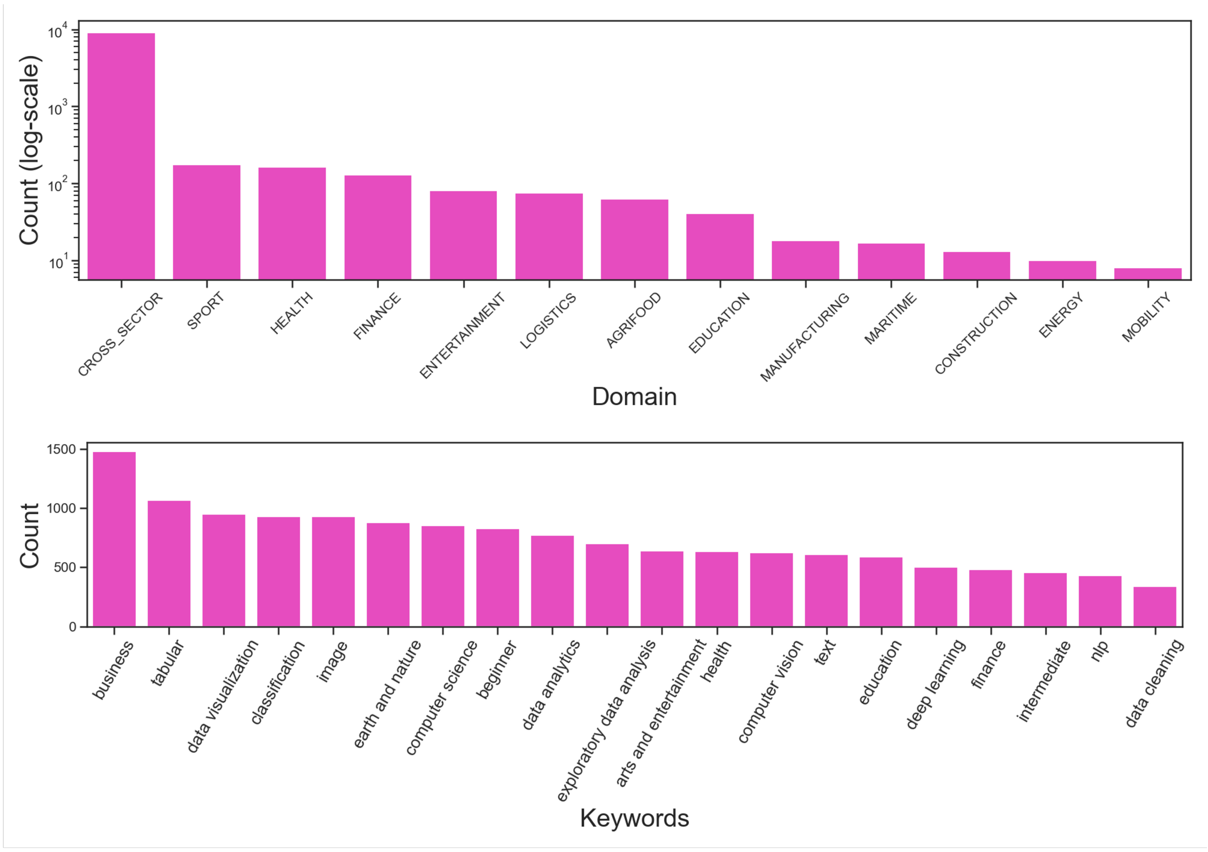


Figure 4: Distribution of Kaggle datasets by domain and keyword frequency. The top chart shows the logarithmic-scale count of datasets across various domains. The bottom chart illustrates the frequency of keywords associated with datasets.

Table 1
Top-10 Semantic and Contextual Matching Results for *Sample Sales Data*

ID	Semantic	Contextual
0	blastchar/telco-customer-churn	armitaraz/google-war-news
1	nnthanh101/aws-saas-sales	jacopoferretti/superstore-dataset
2	shashwatwork/dataco-smart-supply-chain-for-big...	aslanahmedov/walmart-sales-forecast
3	samuelcortinhas/2d-clustering-data	kapatsa/modelled-time-series
4	terencicp/e-commerce-dataset-by-olist-as-an-sq...	sanaaafrine/covid-19-dataset
5	dansbecker/powerlifting-database	winson13/heart-disease-dataset
6	ylchang/coffee-shop-sample-data-1113	moradrawashdeh/attack-simulation-lab
7	dillonmyrick/bike-store-sample-database	rutuspatel/walmart-dataset-retail
8	kyanyoga/sample-sales-data	kyanyoga/sample-sales-data
9	laibaanwer/superstore-sales-dataset	iqmansingh/data-preprocessing-dataset

The *Sample Sales Data dataset*[5], as described by its authors, is a valuable resource for segmentation, customer analytics, and clustering. It contains anonymized sales information, including order details, customer data, sales figures, and shipping details. This dataset serves as an illustrative example of the challenges posed by heterogeneous dataspaces, where datasets differ significantly in structure, content, and application domains. The experiments conducted on this dataset demonstrate the effectiveness of GC-DAM in enabling robust and accurate search functionality within such complex environments.

The results highlight the complementary strengths of semantic and structural embeddings. Semantic

[5] https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

embeddings primarily rely on textual metadata, such as names and content descriptions, to identify matches. Manual inspection reveals that matched datasets often share similar thematic content or application domains. However, relying solely on semantic embeddings can overlook datasets with structural relevance but limited textual similarity.

For example, datasets like *armitaraz/google-war-news*[6] or *kapatsa/modelled-time-series*[7] may not appear semantically related to the query but are highly relevant in terms of structural alignment and contextual relevance. Structural embeddings address this limitation by capturing graph-based relationships and dependencies within datasets. These embeddings excel at identifying matches based on structural patterns, such as schema similarities or shared data formats. For instance, the *google-war-news* dataset discusses economic impacts of war, aligning with the query's context when viewed through its structural focus on economic indicators. Similarly, the *kapatsa/modelled-time-series* dataset provides yearly US GDP values corrected for inflation, resonating with the structural characteristics of sales data.

The GC-DAM approach combines these two embedding spaces to leverage their complementary strengths. By integrating semantic and structural perspectives, GC-DAM identifies datasets that are both contextually and structurally relevant to the query. This multimodal approach ensures that matches are not only thematically coherent but also aligned in terms of schema and data organization. Compared to using only one type of embedding, GC-DAM significantly improves recall and precision in heterogeneous dataspaces. Currently, both semantic and structural embeddings are treated equally when selecting the top 10 matching results for a given query. However, future applications or domain-specific use cases may require weighting these embeddings differently based on their relative importance for the task at hand. Such adjustments would need to be carefully evaluated for each specific use case to ensure optimal performance. This example with the *Sample Sales Data dataset* is one of many used for illustration purposes; readers are encouraged to explore additional examples and further analyses by accessing the accompanying code or the dataset itself.

To systematically evaluate the relevance of the top 10 matching results, we employed an "LLM-as-a-judge" approach using *ChatGPT-4o.* This method leverages the advanced capabilities of Large Language Models to act as evaluators for assessing the suitability of matches based on specific criteria derived from the query. The primary criterion for judgment was whether the retrieved dataset's inferred content was comparable to the reference dataset, based solely on the dataset names provided to the LLM.

Acknowledging the potential limitations and biases inherent in LLM-based evaluation, we incorporated a Human-in-the-Loop validation step to ensure the reliability of the LLM's judgments. Two experienced domain scientists independently reviewed a random sample of 1000 judgments made by the LLM across various queries and embedding types. This human validation served as a crucial benchmark for the LLM's performance on the relevance assessment task. The results showed a high level of agreement between the human evaluators and the LLM, with a concordance rate of 87.4%. This substantial agreement indicates that, for the specific task of judging dataset relevance based on names, the LLM-as-a-judge approach provided sufficiently reliable and consistent evaluations, validating its use for large-scale assessment in this study.

Figure 5 illustrates the distribution of positive matches among the top 10 results as judged by *ChatGPT-4o.* We designed a specific prompt for *ChatGPT-4o* to evaluate all top 10 matching results for both semantic and contextual embeddings. The prompt instructed the model to determine whether each result was relevant to the query based on inferred similarity in dataset content, explicitly stating that the judgment should be based on whether the name suggests comparable data, irrespective of domain or application. The use of LLMs in this context provides several advantages:

1. **Scalability:** LLMs like ChatGPT-4o can efficiently evaluate large datasets without requiring extensive manual effort[20].
2. **Contextual Understanding:** LLMs are capable of capturing nuanced relationships between datasets based on structure and context rather than relying solely on surface-level semantic

[6] https://www.kaggle.com/datasets/armitaraz/google-war-news

[7] https://www.kaggle.com/datasets/kapatsa/modelled-time-series

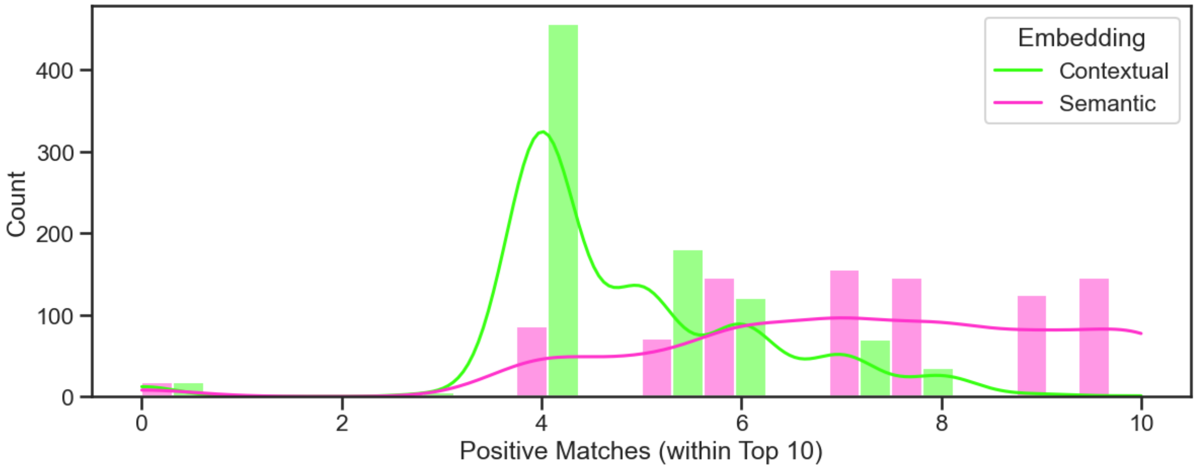


Figure 5: Results of the LLM-based evaluation of GC-DAM matching using *ChatGPT-4o*. The X-axis represents the number of matches (out of the top 10 results) deemed suitable by *ChatGPT-4o*, while the Y-axis shows their frequency.

similarity, particularly when considering dataset names and inferred content[20, 21].

3. **Consistency:** Unlike human evaluators who may introduce variability in judgment, LLMs can provide consistent evaluations across queries when presented with the same prompt and input[21].

The detailed prompt was as follows:

> The following is the name of the reference dataset: `{reference_id}`.
>
> Compare the names in the list below to see if they describe similar datasets to the reference dataset. By "similar," it is meant that you can infer from the name that the dataset contains comparable data. It does not need to be in the same domain or application—the key is whether the dataset content aligns with the reference dataset.
>
> Please return a list of True/False values for each dataset in the list, indicating whether there is a similarity or not.
>
> List to compare: `{dataset_values}`

However, it is crucial to critically reflect on the LLM-as-a-judge paradigm. LLMs can be sensitive to prompt wording, prone to hallucination, and may perpetuate biases present in their training data. Their decision-making process is often opaque, making it difficult to fully understand the rationale behind each judgment compared to explicit rule-based systems or human reasoning. Furthermore, relying solely on LLMs without validation can lead to unreliable results, especially in complex or subjective evaluation tasks. The high agreement observed in our human validation step underscores the necessity of such corroboration, demonstrating that while LLMs can be powerful evaluation tools, their judgments require verification for academic rigor, particularly when assessing nuanced concepts like dataset relevance across heterogeneous types.

The evaluation yielded average agreement percentages (indicating the percentage of relevant matches within the top 10) of 71.90% for semantic embeddings, 48.20% for contextual embeddings, and 77% for the combined GC-DAM approach. Figure 5 shows the distribution of the number of relevant matches within the top 10 for each query. Both embedding approaches achieved at least four positive matches for almost all queries. Semantic embeddings exhibited broader coverage with approval ratings ranging from five to nine positive matches, while contextual embeddings displayed a concentrated distribution around four positive matches. The combined GC-DAM results show a strong propensity for 6 to 8 relevant matches in the top 10, reflecting the benefit of integration. The "LLM-as-a-judge" paradigm, validated by human experts, proved effective in this scenario due to its ability to assess datasets holistically by considering inferred content and potential structural or contextual relevance suggested by the names. This approach aligns well with GC-DAM's goal of identifying suitable matches in heterogeneous environments while ensuring compliance with FAIR principles through accurate data discovery mechanisms.

In summary, leveraging LLMs as evaluators, critically validated by human experts, enhances our ability to systematically assess GC-DAM's performance across diverse datasets. This methodology not only streamlines evaluation processes but also provides actionable insights into improving embedding techniques for dataspace applications, demonstrating the utility of LLMs as a robust, albeit not standalone, evaluation tool.

5. Conclusion and Future Work

This paper presented GC-DAM, a multimodal embedding-based framework designed to address the challenges of data asset matching in heterogeneous and cross-domain dataspaces. By integrating structural and semantic embeddings, GC-DAM effectively identifies data assets that align with queries in both structure and context. This dual embedding strategy surpasses traditional approaches by capturing nuanced relationships between datasets, enabling robust matching across diverse domains. Our evaluation demonstrate GC-DAM's capacity to uncover meaningful connections between datasets that may not appear semantically similar at first glance but align contextually and structurally.

The implementation of GC-DAM aligns with key challenges outlined in the W3C Dataspaces Community Group, particularly *Challenge #2: Data Discovery*. The framework facilitates the discovery of structurally and semantically similar data assets, supporting FAIR-compliant data sharing practices by making datasets findable and interoperable. Moreover, GC-DAM's modular design ensures adaptability across various dataspace architectures, offering a scalable foundation for addressing interoperability issues in federated environments like Gaia-X. To enhance the evaluation of GC-DAM's performance, we employed an "LLM-as-a-judge" approach using state-of-the-art models such as *ChatGPT-4o*. This methodology provided consistent and scalable assessments of dataset matches by leveraging the contextual understanding and reasoning capabilities of LLMs. The use of LLMs as evaluators proved particularly effective in capturing subtle contextual alignments between datasets, which are often overlooked by traditional metrics. In conclusion, GC-DAM represents a significant step forward in advancing semantic interoperability within dataspaces. By aligning with the goals of workshops like SDS 2025 and addressing key W3C challenges, this work lays the groundwork for future innovations in dataspace architectures while fostering collaboration between researchers and practitioners.

Challenges Despite its promising results, several challenges remain. Scalability is a critical limitation when applying GC-DAM to large-scale or highly dynamic dataspaces. The computational complexity of generating and comparing multimodal embeddings can lead to increased overhead, particularly in real-time applications. Additionally, while the integration of structural and semantic embeddings provides a balanced perspective, further research is needed to refine this integration to prevent biases toward the one or the other.

Future Work Future work will focus on addressing these limitations by exploring efficient algorithms for embedding generation and comparison to improve scalability. Advanced techniques for balancing semantic and structural embeddings will be investigated to enhance robustness in cross-domain scenarios. Furthermore, we aim to integrate GC-DAM more closely with ongoing W3C Dataspaces initiatives by proposing new use cases and challenges, such as multimodal embedding techniques for data asset matching (e.g., Issue #6[8]). These efforts will ensure that GC-DAM continues to evolve as a foundational tool for enabling trusted and efficient data sharing within dataspaces.

[8] https://github.com/w3c-cg/dataspaces/issues/6

Declaration on Generative AI

During the writing of this paper, the author(s) used DeepL and GPT-4o in order to: Grammar, translation and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] B. Otto, M. Ten Hompel, S. Wrobel (Eds.), Designing Data Spaces: The Ecosystem Approach to Competitive Advantage, Springer International Publishing, Cham, 2022. doi:10.1007/978-3-030-93975-5.

[2] J. Theissen-Lipp, M. Kocher, C. Lange, S. Decker, A. Paulus, A. Pomp, E. Curry, Semantics in Dataspaces: Origin and Future Directions, in: Companion Proceedings of the ACM Web Conference 2023, ACM, Austin TX USA, 2023, pp. 1504–1507. doi:10.1145/3543873.3587689.

[3] W. Kim, J. Seo, Classifying schematic and data heterogeneity in multidatabase systems, Computer 24 (1991) 12–18. doi:10.1109/2.116884.

[4] A. Hutterer, B. Krumay, Integrating Heterogeneous Data in Dataspaces - A Systematic Mapping Study (2022).

[5] Z. Boukhers, C. Lange, O. Beyan, Enhancing Data Space Semantic Interoperability through Machine Learning: A Visionary Perspective, in: Companion Proceedings of the ACM Web Conference 2023, ACM, Austin TX USA, 2023, pp. 1462–1467. doi:10.1145/3543873.3587658.

[6] M. Ganzha, M. Paprzycki, W. Pawłowski, P. Szmeja, K. Wasielewska, Towards Semantic Interoperability Between Internet of Things Platforms, in: R. Gravina, C. E. Palau, M. Manso, A. Liotta, G. Fortino (Eds.), Integration, Interconnection, and Interoperability of IoT Systems, Springer International Publishing, Cham, 2018, pp. 103–127. doi:10.1007/978-3-319-61300-0_6.

[7] Z. Sun, Q. Zhang, W. Hu, C. Wang, M. Chen, F. Akrami, C. Li, A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs, Proceedings of the VLDB Endowment 13 (2020) 2326–2340. doi:10.14778/3407790.3407828. arXiv:2003.07743.

[8] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, J. Widom, Swoosh: A generic approach to entity resolution, The VLDB Journal 18 (2009) 255–276. doi:10.1007/s00778-008-0098-x.

[9] G. Papadakis, N. Kirielle, P. Christen, T. Palpanas, A Critical Re-evaluation of Benchmark Datasets for (Deep) Learning-Based Matching Algorithms, 2023. arXiv:2307.01231.

[10] X. Xie, Z. Li, X. Wang, Z. Xi, N. Zhang, LambdaKG: A Library for Pre-trained Language Model-Based Knowledge Graph Embeddings, 2023. arXiv:2210.00305.

[11] N. Barlaug, J. A. Gulla, Neural Networks for Entity Matching: A Survey, ACM Transactions on Knowledge Discovery from Data 15 (2021) 1–37. doi:10.1145/3442200.

[12] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, D. Zhao, Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019, pp. 5278–5284. doi:10.24963/ijcai.2019/733. arXiv:1908.08210.

[13] M. Chen, Y. Tian, M. Yang, C. Zaniolo, Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment, 2017. arXiv:1611.03954.

[14] N. Tang, C. Yang, J. Fan, L. Cao, Y. Luo, A. Halevy, VerifAI: Verified Generative AI (2024).

[15] R. Peeters, C. Bizer, Entity Matching using Large Language Models, 2024. arXiv:2310.11244.

[16] Y. Zhou, C. Cui, R. Rafailov, C. Finn, H. Yao, Aligning Modalities in Vision Large Language Models via Preference Fine-tuning, 2024. arXiv:2402.11411.

[17] P. Knowles, P. Page, R. Mitwicki, Decentralised semantics in distributed data ecosystems: Ensuring the structural, definitional, and contextual harmonisation and integrity of deterministic objects and objectual relationships (2022).

[18] Alibaba, GTE-Multilingual Series: A Key Model for Retrieval-Augmented Generation,

https://www.alibabacloud.com/blog/gte-multilingual-series-a-key-model-for-retrieval-augmented-generation_601776, 2024.

[19] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, 2023. doi:10.48550/arXiv.2311.08526. arXiv:2311.08526.

[20] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, H. Liu, From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge, 2025. doi:10.48550/arXiv.2411.16594. arXiv:2411.16594.

[21] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A Survey on LLM-as-a-Judge, 2025. doi:10.48550/arXiv.2411.15594. arXiv:2411.15594.