



University of Milano-Bicocca
Department of Informatics, Systems and Communication
Master's degree program in Data Science

Unveiling Hidden Information in Unstructured Documents

Organization and Hybrid Retrieval with Knowledge Graphs

Supervisor: Prof. Matteo Luigi Palmonari

Co-supervisor: Dott. Francesco Abbracciavento

Co-supervisor: Dott. Riccardo Pozzi

Academic Year: 2023-2024

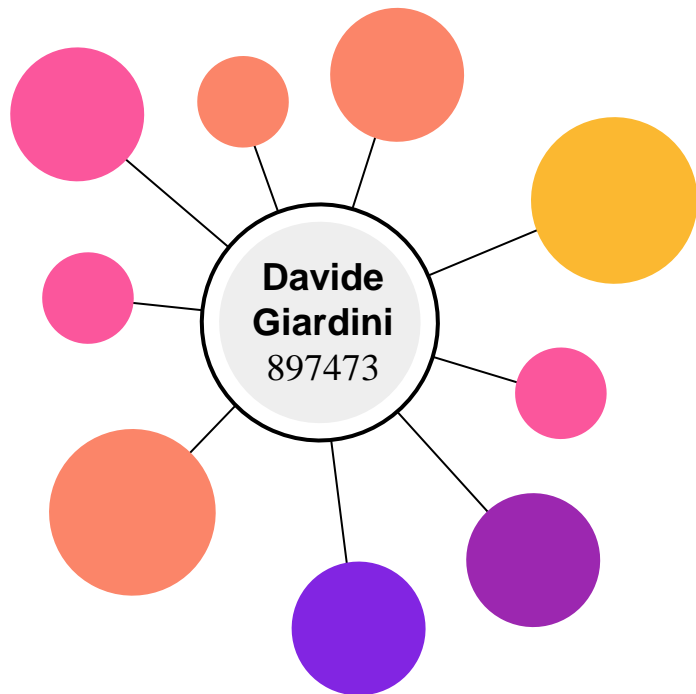


Table of Contents



Introduction

Introduction to
LLMs, QA and RAG

Objectives

Problem definition and
research questions.

Methodology

Methodology proposed to
address the research questions.

Experiments

Evaluation of the Retrieval and
RAG systems.

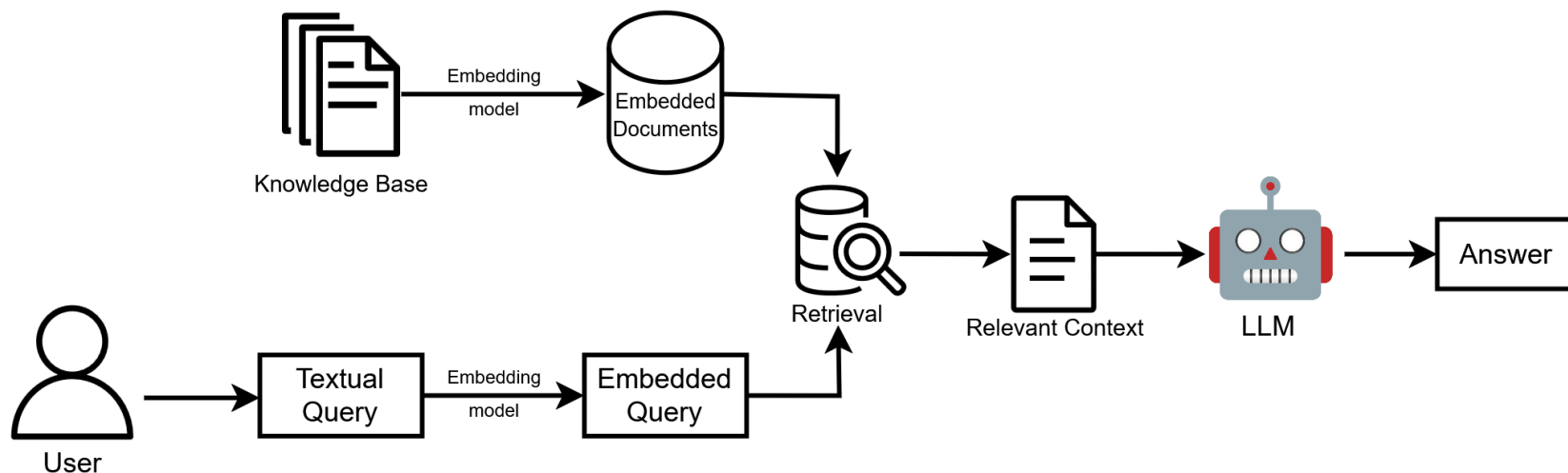
Conclusions

Conclusion and Future
Developments



Introduction

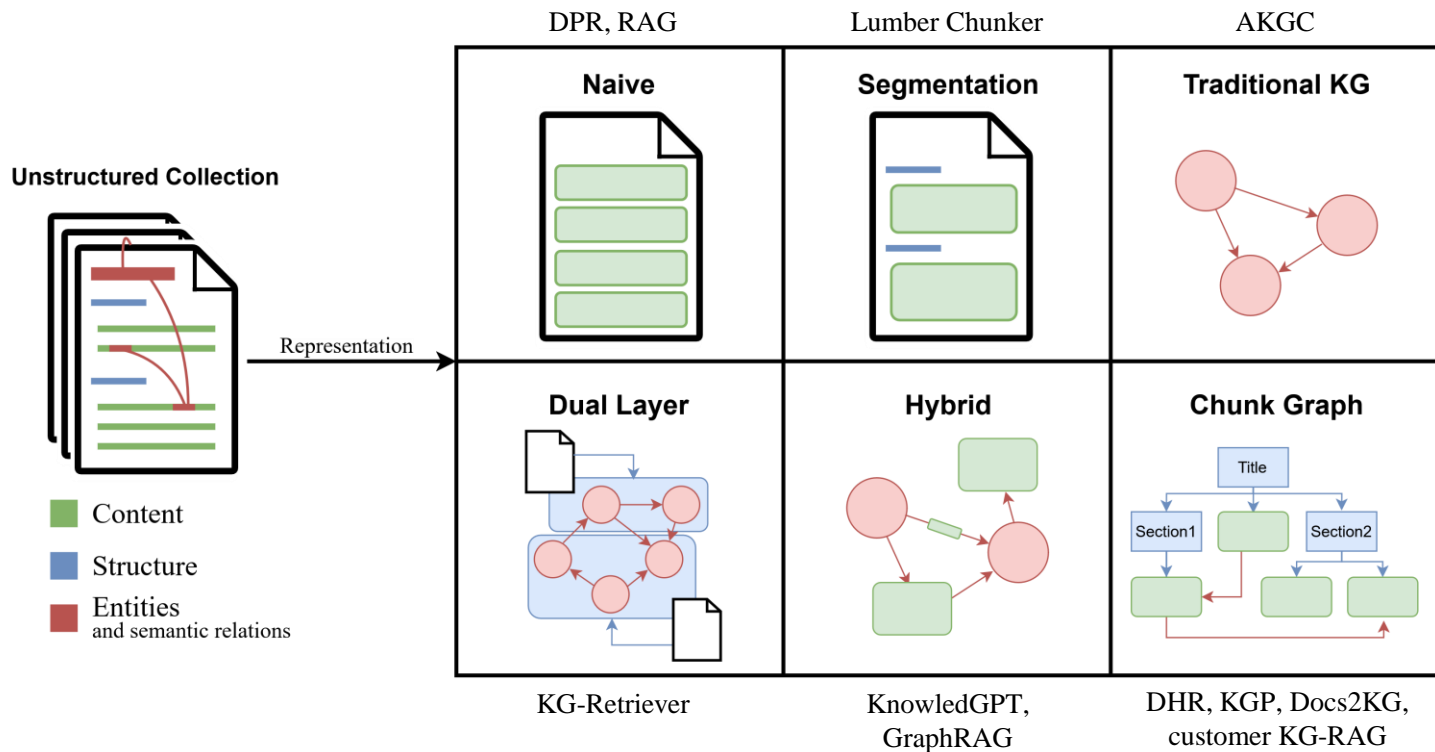
Retrieval Augmented Generation for QA





Objectives

Limitation: Unstructured Documents

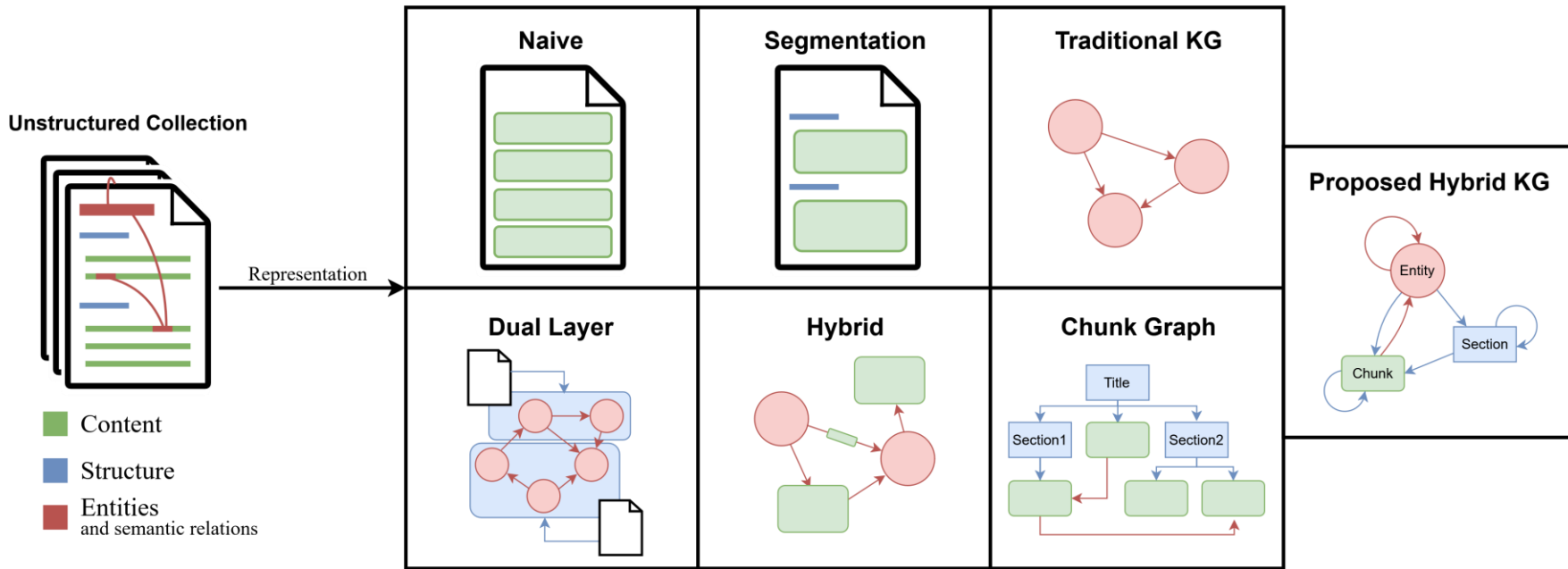




Objectives

Objective 1: Document Structuring
















Creating a KG structure that can effectively represent the initial set of unstructured documents in a new form, focusing on retaining the textual content while unveiling the more implicit information.





Objectives

Limitation: Hybrid Retrieval

	Structure	Entities	Content
Entity Based Question: Where was the song "New Gold", in which Tame Impala is featured, performed live for the first time?	<p>Tame Impala#Tours </p> <p>InnerSpeaker Tour (2010-2011) Lonerism Tour (2012-2014)</p> <p>New Gold (song)#Background </p> <p>It was first played live during the Gorillaz World Tour 2022 on 19 August 2022.</p>	<p>Gorillaz#2022-present </p> <p>the band performed the new song "New Gold" (featuring Tame Impala) at All Points East in London</p> <p>New Gold (song) </p> <p>New Gold is a song by British band Gorillaz, featuring Australian music project Tame Impala, and...</p>	<p>Tame Impala#2018 </p> <p>...2018 Mad Cool Festival in Spain, the first live music show the band agreed to play in 2018.</p> <p>the band performed the new song "New Gold" (featuring Tame Impala) at All Points East in London </p> <p>New Gold is a song by British band Gorillaz, featuring Australian music project Tame Impala, and... </p>
Broad Question: What is the main theme of Harry Potter?	<p>Harry Potter#Themes#1 </p> <p>Harry Potter's overarching theme is death. In the first book, when Harry looks in the Mirror of Erised, he...</p> <p>Harry Potter#Themes#2 </p> <p>Love distinguishes Harry and Voldemort. Harry is a hero because he loves others, even willing to...</p>	<p>Harry Potter </p> <p>Harry Potter is a series of seven fantasy novels written by the British author J.K. Rowling.</p> <p>J.K. Rowling </p> <p>...she is the author of Harry Potter, a seven-volume fantasy novel series published from 1997 to...</p>	<p>Harry Potter#Plot </p> <p>The series follows the life of a boy named Harry Potter. in the first book, Harry Potter and the... </p> <p>Harry Potter is a series of seven fantasy novels written by the British author J.K. Rowling. </p> <p>Harry Potter's overarching theme is death. In the first book, when Harry looks in the Mirror of Erised, he... </p>



Objectives

Objective 2: Hybrid Retrieval

Designing a retrieval system capable of leveraging this enriched structure, and therefore taking full advantage of all of the document's informational dimensions.

Secondary Research Question

Understanding which retrieval methods play a more crucial role in identifying the most relevant passages, and subsequently design a retrieval approach that weights different retrieval methods proportionally to their actual importance.

User's query

Dimension-specific
relevance scores

Combination

Comprehensive
relevance score

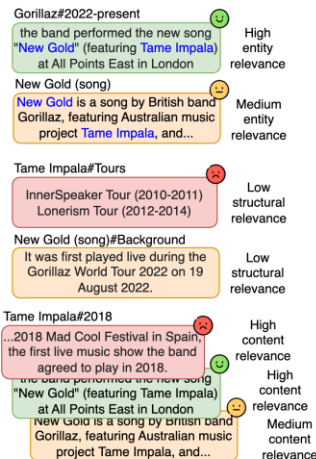
Entity Based Question:

Where was the song
"New Gold", in which
Tame Impala is
featured, performed
live for the first time?

Entities

Structure

Content



Neural Network

Gorillaz#2022-present
the band performed the new song "New Gold" (featuring Tame Impala) at All Points East in London

High comprehensive relevance

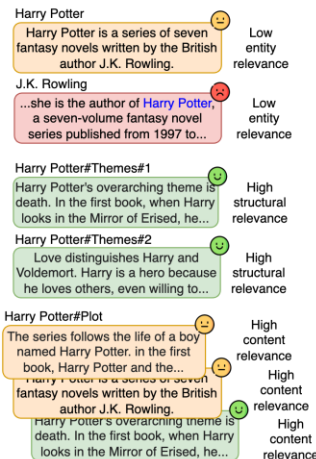
Broad Question:

What is the main theme
of Harry Potter?

Entities

Structure

Content



Neural Network

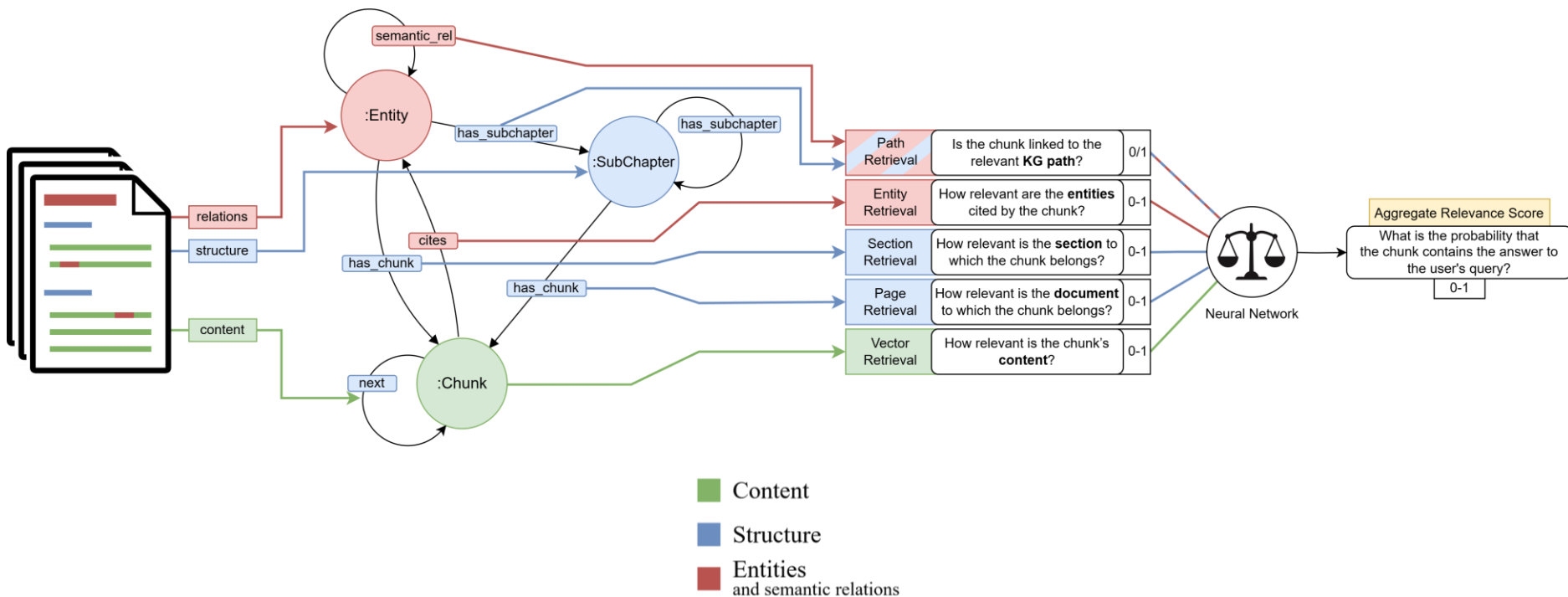
Harry Potter#Themes#1
Harry Potter's overarching theme is death. In the first book, when Harry looks in the Mirror of Erised, he...

High comprehensive relevance



Approach Description

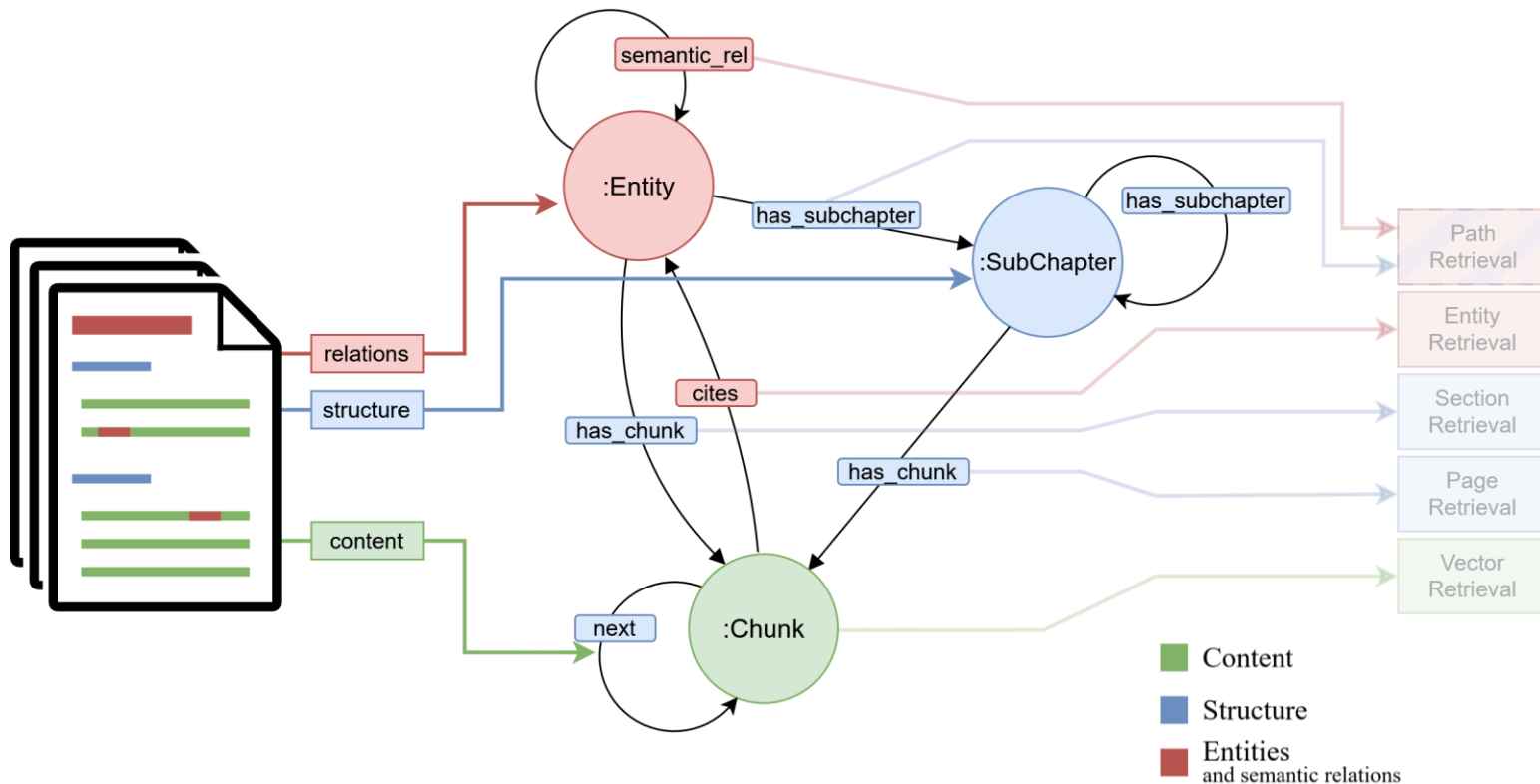
Overview





Approach Description

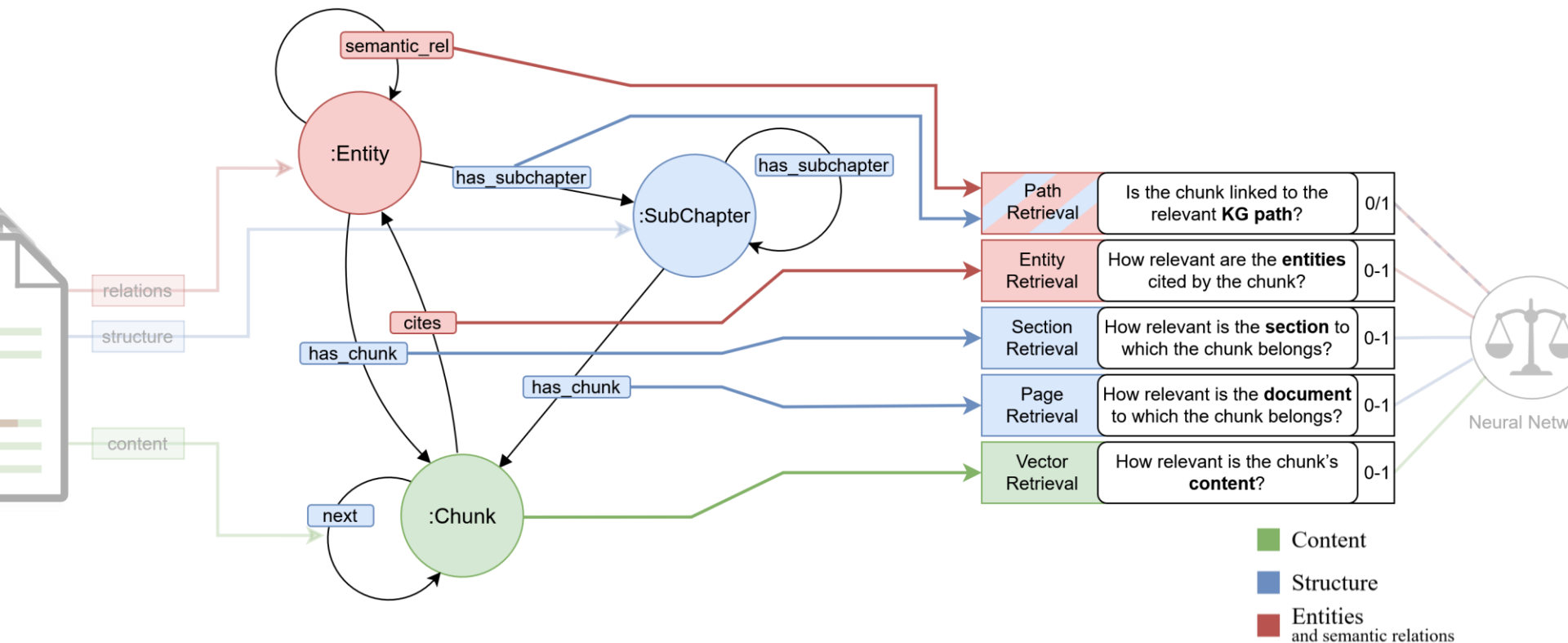
Document Structuring





Approach Description

Retrieval Methods

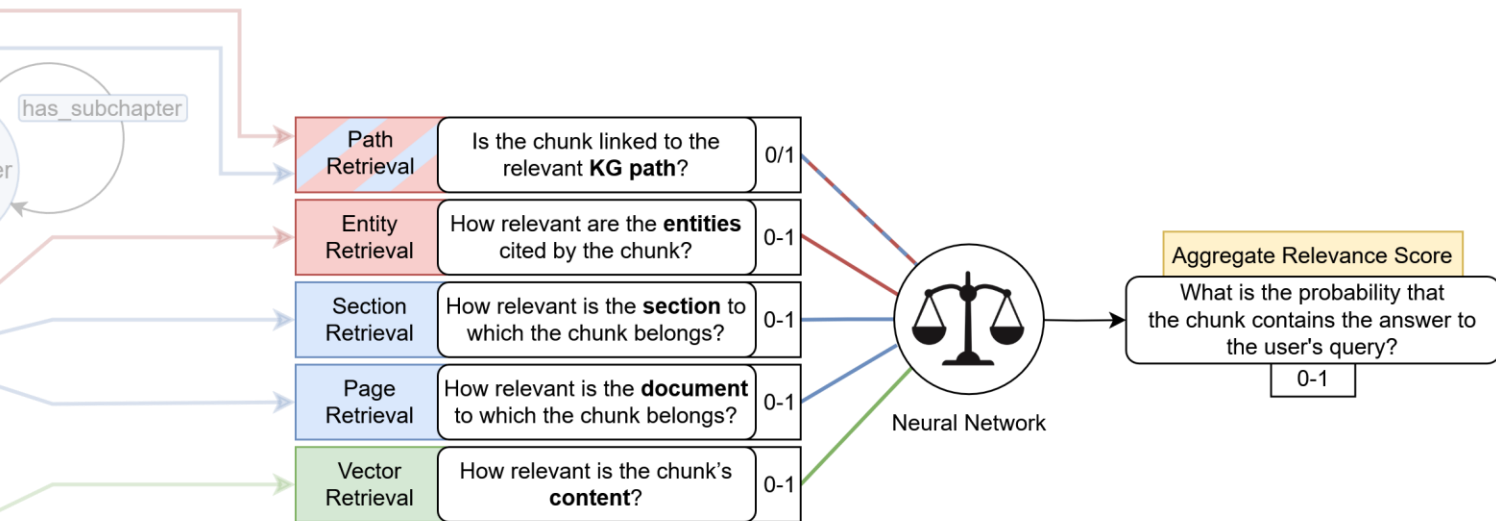




Methodology

Approach Description

Retrieval Fusion

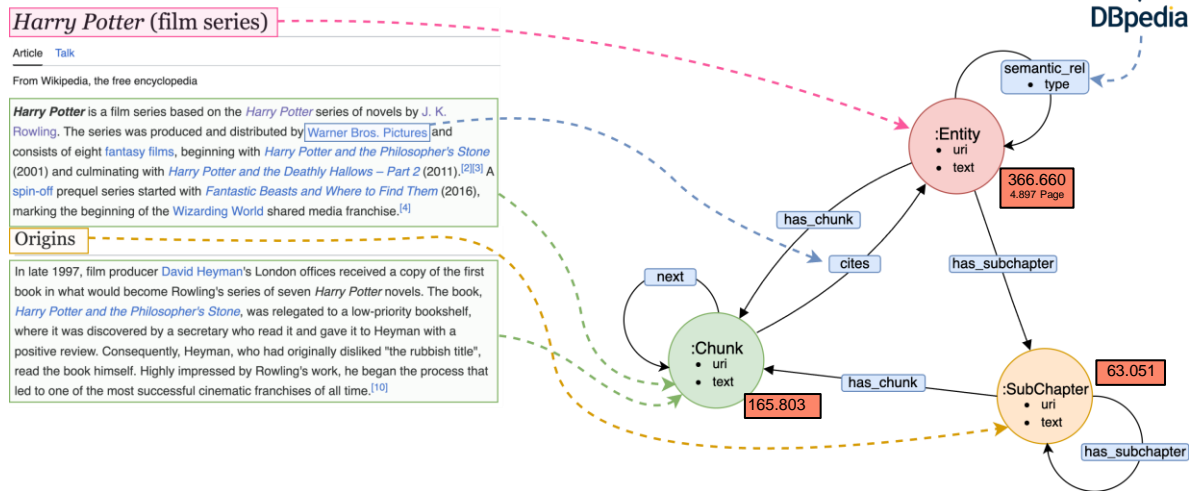


- Content
- Structure
- Entities and semantic relations



Datasets & KB construction

- **NQ-dev**
5223 questions with an annotated answers
 - 4897 pages for **KB construction**
 - **NQ-dev-80%**
4179 questions for training
 - **NQ-dev-20%**
1044 questions for testing
- **Synthetic MultiHop dataset**
100 multi-hop questions





Experiments

NQ Results

Retrieval Method	N. of Elements	N. of Chunks	Precision	Recall	Hit	F1	F2	F3
Vector Retrieval	1 Chunk	1	0.398	0.308	0.398	0.347	0.323	0.315
	3 Chunks	3	0.241	0.528	0.622	0.330	0.426	0.472
	5 Chunks	5	0.175	0.630	0.710	0.274	0.414	0.500
Page Retrieval	1 Page	~ 35	0.071	0.789	0.789	0.130	0.261	0.392
	2 Pages	~ 72	0.029	0.861	0.861	0.056	0.128	0.223
	3 Pages	~ 108	0.018	0.885	0.885	0.035	0.083	0.152
Section Retrieval	1 Section	~ 3	0.207	0.352	0.421	0.261	0.309	0.329
	3 Sections	~ 6	0.113	0.604	0.663	0.190	0.323	0.421
	5 Sections	~ 9	0.058	0.798	0.829	0.108	0.225	0.351
Entity Retrieval	1 Entity	~ 4	0.059	0.096	0.118	0.073	0.085	0.090
	3 Entities	~ 8	0.046	0.205	0.249	0.075	0.121	0.152
	5 Entities	~ 11	0.041	0.248	0.299	0.070	0.123	0.165
Path Retrieval	1 Path	~ 3	0.266	0.423	0.512	0.327	0.378	0.399
Naive Hybrid Retrieval	5 Chunks	5	0.188	0.695	0.776	0.296	0.451	0.547
Hybrid Retrieval	1 Chunk	1	0.478	0.371	0.478	0.418	0.388	0.379
	3 Chunks	3	0.298	0.662	0.756	0.411	0.532	0.590
	5 Chunks	5	0.215	0.780	0.851	0.337	0.511	0.618

MultiHop Results

N. of Chunks	Retrieval Method	Metrics					
		Precision	Recall	Set Coverage	F1	F2	F3
4	Vector	0.175	0.357	0.090	0.235	0.296	0.323
	Hybrid Naive	0.125	0.381	0.110	0.188	0.270	0.316
	Hybrid	0.302	0.599	0.330	0.402	0.501	0.545
6	Vector	0.138	0.418	0.160	0.207	0.297	0.347
	Hybrid Naive	0.143	0.432	0.170	0.215	0.308	0.359
	Hybrid	0.232	0.688	0.430	0.347	0.494	0.575
8	Vector	0.111	0.448	0.180	0.178	0.279	0.344
	Hybrid Naive	0.155	0.467	0.200	0.232	0.333	0.389
	Hybrid	0.180	0.711	0.480	0.287	0.447	0.549

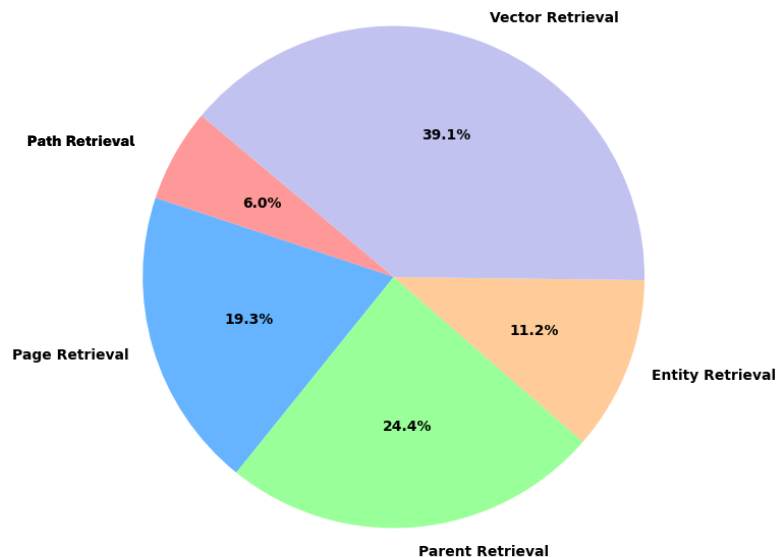


Additional Results

RAG quality

	Retrieval	EM	ROUGE	Cosine	FC
NQ-dev-20%-sa	Vector	0.355	0.461	0.891	0.636
	Hybrid	0.399	0.515	0.903	0.672
NQ-MH-100	Vector	0.040	0.394	0.843	0.424
	Hybrid	0.060	0.507	0.887	0.519

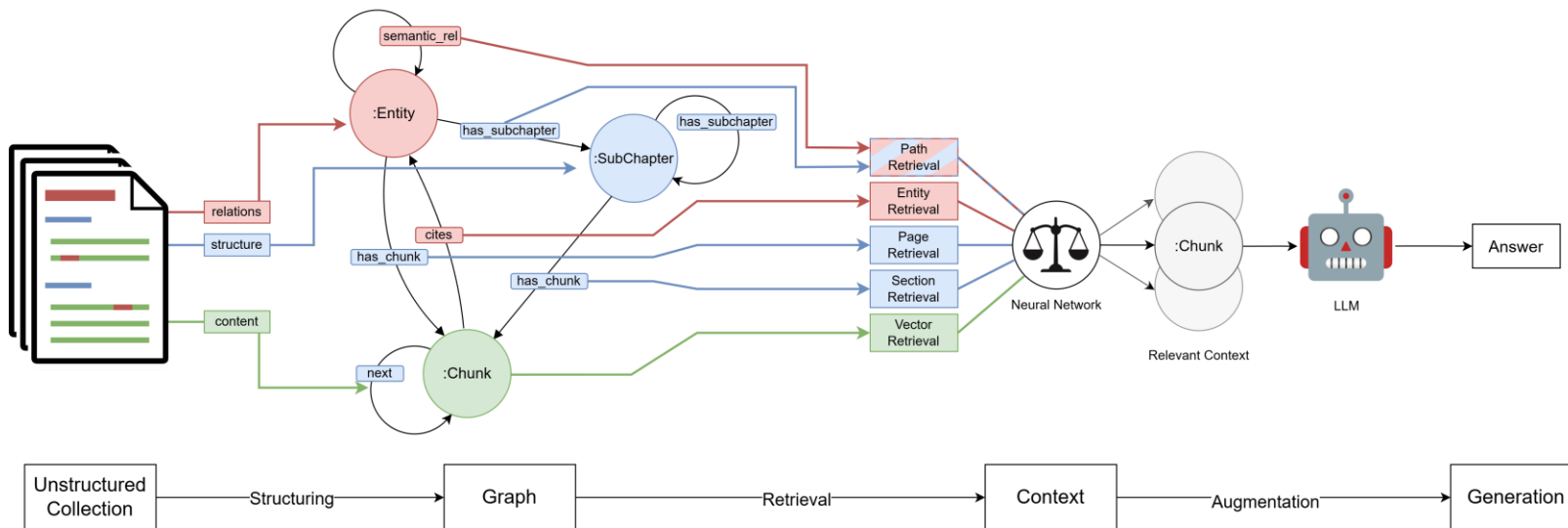
Features Importance in Retrieval





Conclusions

Conclusion and Future Developments



Limitations and Future Developments:

- Transfer Learning
- Extraction Sensitivity
- Hyperparameter Finetuning
- Scalability
- Adaptive Retrieval