

Project: Query Answering over Linked Data

Pablo Mollá Chárlez, Junjie Chen and Pavlo Poliuha

Contents

1	Introduction	1
1.1	Context and Background	1
1.2	Problem Statement	1
1.3	Objectives	2
2	Methodology	2
2.1	Question 1: Addressing a Question from the Lecture	2
2.1.1	LLM Approach	2
2.1.2	Browser Approach	3
2.1.3	Linked Data Approach	4
2.1.4	Challenges	4
2.2	Question 2: Formulating and Solving a New Complex Question	5
2.2.1	LLM Approach	5
2.2.2	Linked Data Approach	5
2.2.3	Challenges	6
2.3	Question 3: Justifications and Hybrid Method	7
2.3.1	Justifications: CONSTRUCT	7
2.3.2	Hybrid Method	7
3	Discussion	8
3.1	Datasets	8
3.2	Benefits and Limitations	8
4	Conclusion	8

1 Introduction

1.1 Context and Background

The concept of **Linked Data**, introduced by **Tim Berners-Lee** in the early 2000s, aims to extend the web by structuring information to be interoperable across diverse datasets using standardized formats. This interconnected "Web of Data" enables **more sophisticated queries and integration from different sources**. In his 2009 TED talk, Berners-Lee emphasized the importance of making data open and linked to **unlock new discoveries and innovative applications**. Linked Data **addresses limitations of traditional web searches** by using structured relationships to answer complex questions more effectively.

1.2 Problem Statement

Linked Data effectively manages structured and well-defined information, while advancements in **Large Language Models (LLMs)** like ChatGPT, Gemini, and Claude enhance this capability. LLMs utilize **natural language processing (NLP)** to generate human-like responses, trained on extensive textual data to infer meaning and provide context. However, they can face **hallucinations** and struggle with accessing up-to-date or domain-specific knowledge. Together,

Linked Data and **LLMs** complement each other: Linked Data integrates diverse sources, while LLMs enhance interpretative skills. This combination allows for **more sophisticated answers to complex queries that traditional search engines or individual datasets may not easily address**.

1.3 Objectives

In this report, we will examine the potential of **Linked Data** and **LLMs** to answer two specific questions:

1. "Who is the French archer who has won an Olympic event and received the Honour Legion?" – a question drawn from a set of lecture topics exploring data integration challenges.
2. "What is the single-player video game that was released in 2011 on the App Store and in 2012 on Android, and was awarded with the Apple Design Award in 2011 and 'Best App Ever 2011'?" – a complex new question that integrates temporal and categorical data.

By comparing the effectiveness of **Linked Data** and **LLMs** in answering these questions, we aim to uncover insights into the **strengths and limitations of each approach**. Additionally, we will explore a **hybrid method** that combines the precision of Linked Data with the interpretive capabilities of LLMs to achieve more comprehensive answers.

2 Methodology

2.1 Question 1: Addressing a Question from the Lecture

The first question we aim to answer is: **Who is the French archer who has won an Olympic event and received the Honour Legion?**

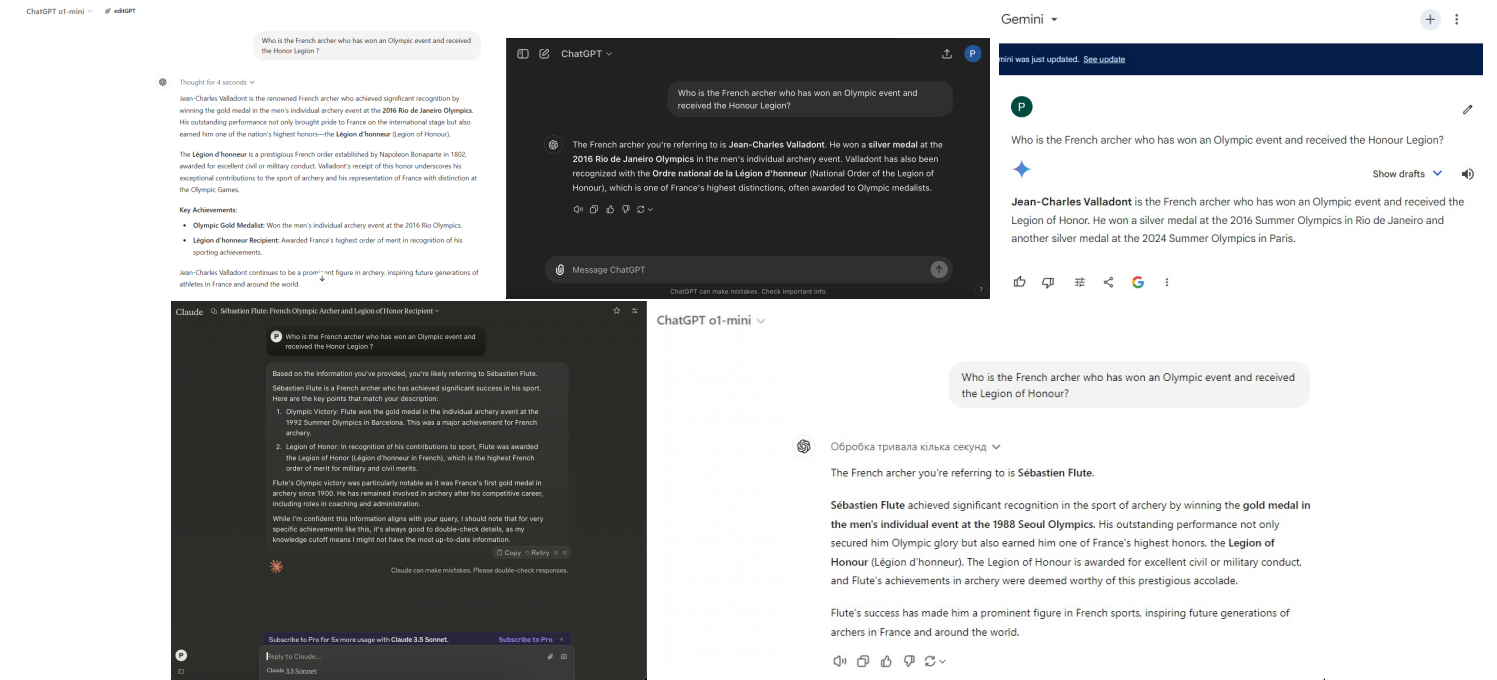
This question requires linking information about an individual's nationality (French), profession (archer), achievement (winning an Olympic event), and award (Honour Legion). The keywords associated with the query are: **French**, **archer**, **Olympic gold medal**, and **Honour Legion**. The Honour Legion is a notable French distinction awarded to individuals for outstanding service or achievements, which adds another layer of specificity to the question. The actual and correct answer is the french athlete **Sébastien Flute**, a French archer who won the gold medal in the 1992 Olympics and was awarded the Legion of Honour.

2.1.1 LLM Approach

To tackle this question, we will first utilize various versions of LLMs to see how they respond and what information they can retrieve. The LLMs used will be **ChatGPT o1-mini**, **ChatGPT o4**, **Claude 3.5**, and **Gemini**, as each has different training data and capabilities that might affect their ability to provide a correct answer. The observations for each model are as follows:

- **ChatGPT o4** and **ChatGPT o1-mini** Both versions provided **Jean-Charles Valladont** as the answer, a French archer who competed in the Olympics. Although Valladont is a well-known archer, he did not meet all criteria, such as winning an Olympic gold medal or receiving the Honour Legion. Unexpectedly, when we slightly modify the original question to "Who is the French archer who has won an Olympic event and received the **Legion of Honour**?" The model **ChatGPT o1-mini** is able to determine the right answer although fails providing the correct date, which is in 1992 instead of 1988. We believe that ChatGPT retrieved the most recent archer.
- **Claude 3.5** Claude correctly identified **Sébastien Flute**. This response directly answered the question and matched all criteria.
- **Gemini** Gemini, on the other hand, gave Jean-Charles Valladont as the answer, which was incorrect and resulted in the same answer given by **ChatGPT**'s models.

In the previous images we can notice how both LLMs, **Claude 3.5** and **ChatGPT o1-mini**, retrieve the correct athlete, being **Claude 3.5** the most correct response.



2.1.2 Browser Approach

We then examined how traditional search engines would perform when trying to retrieve the solution. For this, we tested three popular search engines: **Google**, **Bing**, and **Yahoo**. Here are the key observations:

- Google** Google provided results related to French archers, but most references were related to recent Olympic events, such as the 2024 Games, and included Jean-Charles Valladont as a prominent name. This was an improvement over the other engines but still did not fulfill the requirements of the question completely.
- Bing and Yahoo** Both search engines struggled to find relevant information that matched all criteria. The results often included general information about the Olympics or the Legion of Honour but did not successfully link the achievement of winning an Olympic gold medal as an archer with receiving the Honour Legion.

A significant challenge encountered was that none of the search engines could accurately process the term "Honour Legion" since the commonly used terminology is "Legion of Honour." This difference in terminology likely affected the quality of the results, as seen in the link from the International Handball Federation referencing French Olympic winners who received Legion of Honour distinctions. However, the search engines struggled to provide a direct answer combining all elements of the query.

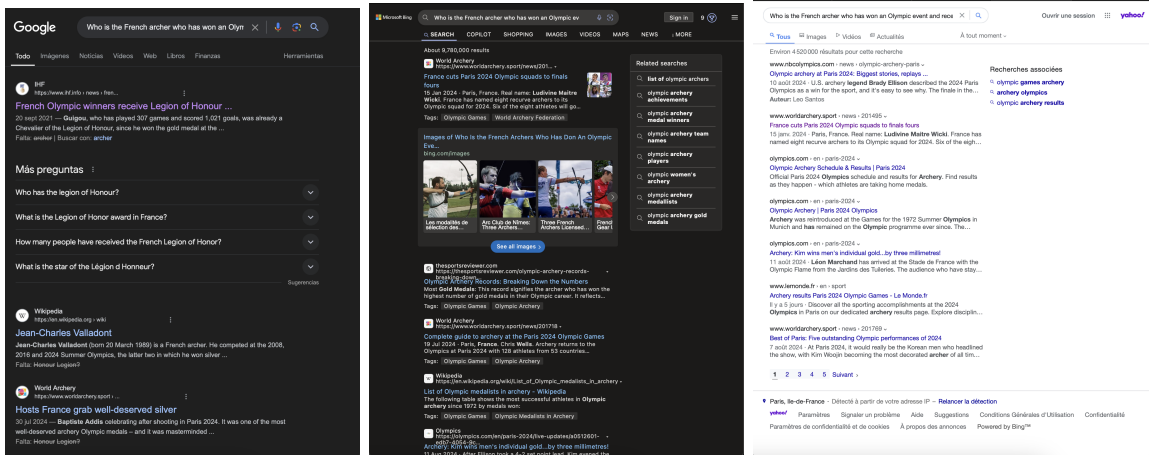


Figure 1: Browsers Comparison

2.1.3 Linked Data Approach

We first needed to **understand the structure of categories within DBpedia**. The initial lead came from exploring the information profile of French archer **Jean-Charles Valladont** on DBpedia, accessible via this link. By examining his profile, we identified various properties, including the category **"Olympic silver medalists for France"** (link). Following this trail, we explored related categories, such as **"Olympic medalists for France"** (link) and ultimately reached the **"Olympic gold medalists for France"** category (link). Simultaneously, we investigated the structure of the **"Legion of Honour"** (link) and its corresponding category, **"Recipients of the Legion of Honour"** (link). By understanding these properties and entities, we were able to accurately formulate the SPARQL queries.

Query 1: The first query retrieves French athletes who have received the Legion of Honour and have won an Olympic gold medal for France. The query filters results to include only those who are labeled as archers in their description:

```
PREFIX metadata_terms: <http://purl.org/dc/terms/>
PREFIX category: <http://dbpedia.org/resource/Category:>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT DISTINCT ?archer ?full_name
WHERE {?archer rdf:type dbo:Athlete ;
      metadata_terms:subject category:Recipients_of_the_Legion_of_Honour ;
      metadata_terms:subject category:Olympic_gold_medalists_for_France ;
      rdfs:label ?full_name .
      OPTIONAL { ?archer dbo:abstract ?abstract . }
      OPTIONAL { ?archer dbo:nationality dbr:France . }
      FILTER regex(?abstract, "archer", "i")
      FILTER (lang(?full_name) = "fr")}
ORDER BY ?full_name
```

Query 2: The second query searches for athletes who have participated in archery at the Olympics, represented France, and received the Legion of Honour. This query looks for English-language labels:

```
SELECT ?archer ?full_name
WHERE {?archer rdf:type dbo:Athlete ;
      metadata_terms:subject category:Recipients_of_the_Legion_of_Honour ;
      metadata_terms:subject category:Olympic_medalists_in_archery ;
      metadata_terms:subject category:Olympic_archers_of_France ;
      rdfs:label ?full_name .
      OPTIONAL { ?archer dbo:nationality dbr:France . }
      FILTER (lang(?full_name) = "en")}
ORDER BY ?full_name
```

By refining these queries, we identified Sébastien Flute as the relevant athlete.

2.1.4 Challenges

- **Correct Terminology for "Legion of Honour":** In our initial queries, we **mistakenly referred to the award as "Honor Legion" instead of its proper name, "The Legion of Honour"**. We needed to use the correct term to ensure the queries matched the relevant category in DBpedia.
- **Difficulty Finding Olympic Gold Medalists in Archery for France:** In the initial query, we searched for French athletes who were listed under "Olympic medalists in archery" and "Olympic archers of France." However, our goal was to identify archers who had won an Olympic gold medal. Unfortunately, **DBpedia does not have a specific category** for "Olympic gold medalists for France in archery."

```
SELECT ?archer ?full_name
WHERE {?archer rdf:type dbo:Athlete ;
      metadata_terms:subject category:Recipients_of_the_Legion_of_Honour ;
      metadata_terms:subject category:Olympic_medalists_in_archery ;
      metadata_terms:subject category:Olympic_archers_of_France ;
      rdfs:label ?full_name .
```

```

        OPTIONAL { ?archer dbo:nationality dbr:France .}
        FILTER (lang(?full_name) = "en")}
ORDER BY ?full_name

```

To address this, we modified the query to use the broader category **"Olympic gold medalists for France"**, which includes all French Olympic gold medalists across various sports, rather than just archery.

```

SELECT DISTINCT ?archer ?full_name
WHERE {
    ?archer rdf:type dbo:Athlete ;
        metadata_terms:subject category:Recipients_of_the_Legion_of_Honour ;
        metadata_terms:subject category:Olympic_gold_medalists_for_France ;
        rdfs:label ?full_name .
    OPTIONAL { ?archer dbo:abstract ?abstract . }
    OPTIONAL { ?archer dbo:nationality dbr:France .}
    FILTER regex(?abstract, "archer", "i")
    FILTER (lang(?full_name) = "fr")
}
ORDER BY ?full_name

```

The category **"Olympic gold medalists for France"** covers all sports, providing a more comprehensive approach to identifying the desired athletes, even though it does not specifically focus on archery. To resolve this, we **added a condition to check in the abstract if the athlete is archer**.

Formulating queries that meet all criteria was challenging due to **inconsistencies in data representation** across different Linked Data sources. For example, the terms "Honour Legion" and "Legion of Honour" may be used interchangeably, making it difficult to accurately match awards. Additionally, **data completeness varies**, and not all athletes have their achievements or awards properly linked in datasets like DBpedia or Wikidata.

2.2 Question 2: Formulating and Solving a New Complex Question

The question we are addressing is: **What is the single-player video game that was released in 2011 on the App Store and in 2012 on Android, and was awarded with the Apple Design Award in 2011 and the 'Best App Ever 2011' award?** This question involves identifying a video game that meets specific criteria regarding its genre, release dates on different platforms, and awards received.

2.2.1 LLM Approach

To cross-check the results obtained from Linked Data, we used several large language models (LLMs), as we mentioned earlier. We provided the same question to each LLM and noted their responses:

- **ChatGPT o4** and **ChatGPT o1-mini** Both suggested **Tiny Wings** as potential answer based on the release timeline and the awards criteria. These responses were based on the well-known history of these games.
- **Claude 3.5** Also listed **Tiny Wings** as a possible match. We believe that this LLM and both **ChatGPT** models answered "Tiny Wings" because the game was released in 2011, and according to the Wikipedia page, won "iPhone Game of the Year 2011" and was nominated for "Casual Game of the Year".
- **Gemini** Suggested **Infinity Blade** as the answer, indicating some overlap in how LLMs interpreted the question. The "Gemini" responded with the game called "Infinity Blade" which was awarded with "Mobile Game of the Year 2011". However, in the original question we asked about "Apple Design Award" and "Best App Ever 2011".

2.2.2 Linked Data Approach

The process included multiple steps to refine the criteria and filter the results:

1. **Initial Exploration of Video Game Types** We began by identifying games classified under the "single-player" genre. The initial query checked the types of video games present in DBpedia, specifically searching for modes related to single-player gameplay:

```
SELECT DISTINCT ?game ?property ?
WHERE {
    ?game a dbo:VideoGame;
    ?property ?value.
    FILTER (CONTAINS(STR(?property), "single-player"))
}
LIMIT 100
```

```

- res:solution [
  res:binding [ res:variable "game" ; res:value ns2:Pro_Hunts ] ;
  res:binding [ res:variable "property" ; res:value dbp:modes_1 ] ;
  res:binding [ res:variable "value" ; res:value dbr:Single-player ] ] .
- res:solution [
  res:binding [ res:variable "game" ; res:value ns2:Ultimate_Challenge ] ;
  res:binding [ res:variable "property" ; res:value dbp:modes_1 ] ;
  res:binding [ res:variable "value" ; res:value dbr:Single-player ] ] .
- res:solution [
  res:binding [ res:variable "game" ; res:value <http://dbpedia.org/resource/Cabela\u0027s_Big_Game_Hunter_2005_Adventures> ] ;
  res:binding [ res:variable "property" ; res:value dbp:modes_1 ] ;
  res:binding [ res:variable "value" ; res:value dbr:Multiplayer_video_game ] ] .
- res:solution [
  res:binding [ res:variable "game" ; res:value <http://dbpedia.org/resource/Cabela\u0027s_Big_Game_Hunter_2005_Adventures> ] ;
  res:binding [ res:variable "property" ; res:value dbp:modes_1 ] ;
  res:binding [ res:variable "value" ; res:value dbr:Single-player_video_game ] ] .
@prefix ns3: <http://dbpedia.org/resource/Cabela\u0027s_Dangerous_Hunts> .
- res:solution [
  res:binding [ res:variable "game" ; res:value ns3:Ultimate_Challenge ] ;
  res:binding [ res:variable "property" ; res:value dbp:modes_1 ] ;
  res:binding [ res:variable "value" ; res:value dbr:Single-player_video_game ] ] .
- res:solution [
  res:binding [ res:variable "game" ; res:value <http://dbpedia.org/resource/Cabela\u0027s_Dangerous_Hunts-2> ] ;
  res:binding [ res:variable "property" ; res:value dbp:modes_1 ] ;
  res:binding [ res:variable "value" ; res:value dbr:Single-player_video_game ] ] .

```

This query helped us understand the structure and properties available in DBpedia for video games.

2. **Filtering by Platform and Timeframe** With the "single-player" mode established, we refined the search to include games released on iOS in 2011 and Android in 2012. The SPARQL query filtered for release dates that matched these specific timeframes.
3. **Award Criteria** To ensure the game matched the awards criteria, we used regular expressions to search for games that had won both the Apple Design Award in 2011 and the "Best App Ever 2011" award.

```
SELECT DISTINCT ?game ?award1 ?award2 ?releaseDateIOS ?releaseDateAndroid
WHERE {
    ?game a dbo:VideoGame;
    dbo:wikiPageWikiLink dbr:Single-player_video_game.
    ?game dbp:award ?award1.
    FILTER (REGEX(str(?award1), "apple.*design.*award", "i"))
    ?game dbp:award ?award2.
    FILTER (str(?award2)="Best App Ever 2011")
    ?game dbo:computingPlatform <http://dbpedia.org/resource/iOS>;
    dbo:releaseDate ?releaseDateIOS.
    FILTER (?releaseDateIOS >= "2011-01-01"^^xsd:date && ?releaseDateIOS <= "2011-12-31"^^xsd:date)
    # Optional filter for Android release date in 2012
    OPTIONAL {
        ?game dbo:computingPlatform <http://dbpedia.org/resource/Android_(operating_system)>;
        dbo:releaseDate ?releaseDateAndroid.
        FILTER (?releaseDateAndroid >= "2012-01-01"^^xsd:date &&
        ?releaseDateAndroid <= "2012-12-31"^^xsd:date)
    }
}
```

This final query retrieved the game **Jetpack Joyride**, which matched all specified criteria.

2.2.3 Challenges

The main challenge we encountered with the second question was in identifying games released on Android in 2012. In DBpedia, there isn't a dedicated list for such releases, making it difficult to filter results directly. To address this, we used the 'OPTIONAL' keyword in our SPARQL query to make this condition optional, allowing for cases where the Android release information might not be explicitly listed:

```
OPTIONAL {?game dbo:computingPlatform <http://dbpedia.org/resource/Android>;
    dbo:releaseDate ?releaseDateAndroid .
    FILTER (?releaseDateAndroid >= "2012-01-01"^^xsd:date &&
    ?releaseDateAndroid <= "2012-12-31"^^xsd:date)}
```

This approach ensured that games meeting all other criteria could still be considered even if the Android release date information was missing.

2.3 Question 3: Justifications and Hybrid Method

2.3.1 Justifications: CONSTRUCT

The following query constructs triples that provide justification for identifying archers who have received the Legion of Honour and are Olympic gold medalists for France (**Question 1**) and for the identification of a specific video game, focusing on its awards and release information (**Question 2**).

Question 1

```
CONSTRUCT {?archer rdf:type dbo:Athlete.
            ?archer rdfs:label ?full_name.
            ?archer metadata_terms:subject
            category:Recipients_of_the_Legion_of_Honour.
            ?archer metadata_terms:subject
            category:Olympic_gold_medalists_for_France .}
WHERE {?archer rdf:type dbo:Athlete;
       metadata_terms:subject
       category:Recipients_of_the_Legion_of_Honour;
       metadata_terms:subject
       category:Olympic_gold_medalists_for_France;
       rdfs:label ?full_name.
       OPTIONAL { ?archer dbo:abstract ?abstract. }
       OPTIONAL { ?archer dbo:nationality dbr:France. }
       FILTER regex(?abstract, "archer", "i")
       FILTER (lang(?full_name) = "en")}
ORDER BY ?full_name
```

Question 2

```
CONSTRUCT {?game rdf:type dbo:VideoGame.
            ?game dbo:wikiPageWikiLink
            dbr:Single-player_video_game.
            ?game dbp:award ?award1.
            ?game dbp:award ?award2.
            ?game dbo:computingPlatform
            <http://dbpedia.org/resource/IOS>.
            ?game dbo:releaseDate ?releaseDateIOS.}
WHERE {?game rdf:type dbo:VideoGame.
       ?game dbo:wikiPageWikiLink
       dbr:Single-player_video_game.
       ?game dbp:award ?award1 .
       FILTER (REGEX(str(?award1), "apple.*design.*award", "i"))
       ?game dbp:award ?award2.
       FILTER (str(?award2) = "Best App Ever 2011")
       ?game dbo:computingPlatform
       <http://dbpedia.org/resource/IOS>.
       ?game dbo:releaseDate ?releaseDateIOS.
       FILTER (?releaseDateIOS >= "2011-01-01"^^xsd:date &&
       ?releaseDateIOS <= "2011-12-31"^^xsd:date)}
```

2.3.2 Hybrid Method

The hybrid approach explored here involves directly inputting SPARQL queries, such as those addressing **Question 1** (to identify the athlete Sébastien Flute) and **Question 2** (to find the video game Jetpack Joyride), into models like ChatGPT (o1-mini) and Claude 3.5. The goal is to see if the results returned by these models match the actual entities we want to find. By combining Linked Data and large language models, we evaluate whether the results reflect the desired entities or information in an accurate and meaningful way. As it can be observed, the answers do not really change except ChatGPT o1-mini with respect to the **Question 1**, in which doesn't provide a specific name.

SPARQL | HTML Microdata document

This HTML5 document contains 4 embedded RDF statements represented using HTML+Microdata notation.

The embedded RDF content will be recognized by any processor of HTML5 Microdata.

Namespace Prefixes

Prefix	IRI
dct	http://purl.org/dc/terms/
dbo	http://dbpedia.org/ontology/
rdfs	http://www.w3.org/2000/01/rdf-schema#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
dbc	http://dbpedia.org/resource/Category:
xsdh	http://www.w3.org/2001/XMLSchema#
dbr	http://dbpedia.org/resource/

Statements

Subject Item
dbr:Sébastien_Flute

rdf:type
[dbo:Athlete](#)

rdfs:label
Sébastien Flute

dct:subject
[dbc:Olympic_gold_medalists_for_France](#) [dbc:Recipients_of_the_Legion_of_Honour](#)

SPARQL | HTML Microdata document

This HTML5 document contains 6 embedded RDF statements represented using HTML+Microdata notation.

The embedded RDF content will be recognized by any processor of HTML5 Microdata.

Namespace Prefixes

Prefix	IRI
dbo	http://dbpedia.org/ontology/
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
dbp	http://dbpedia.org/property/
xsdh	http://www.w3.org/2001/XMLSchema#
dbr	http://dbpedia.org/resource/

Statements

Subject Item
dbr:Jetpack_Joyride

rdf:type
[dbo:VideoGame](#)

dbo:wikiPageWikiLink
[dbr:Single-player_video_game](#)

dbp:award
Best App Ever 2011 [dbr:Apple_Design_Award](#)

dbo:releaseDate
2011-09-01

dbo:computingPlatform
[dbr:IOS](#)

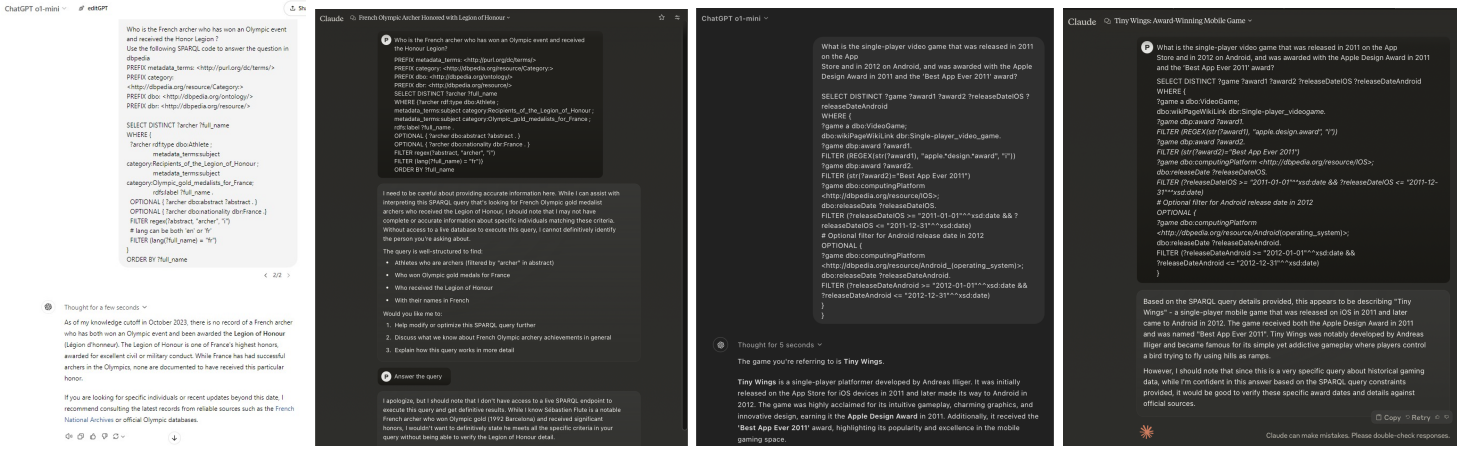


Figure 2: ChatGPT o1-mini vs Claude 3.5 Comparison

3 Discussion

3.1 Datasets

The following table summarizes the prefixes used, their corresponding namespaces, and a brief description or link for each.

Prefix	Namespace	Description/Link
dct	http://purl.org/dc/terms/	Dublin Core Terms
dbo	http://dbpedia.org/ontology/	DBpedia Ontology
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF Syntax Namespace
dbc	http://dbpedia.org/resource/Category:	DBpedia Categories
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema Definition
dbr	http://dbpedia.org/resource/	DBpedia Resource
dbp	http://dbpedia.org/property/	DBpedia Properties

3.2 Benefits and Limitations

Ontologies provide **structured and linked data**, enabling reliable answers to complex questions, and enhance **trustworthiness** through community-curated datasets like DBpedia, often offering more accurate information than static LLM training data. However, they face challenges with **outdated or incomplete data**, as updates depend on community input, and there is **temporal variability** due to changing information over time. An interesting approach would be integrating LLMs with ontologies by enabling queries to live datasets like DBpedia could improve response accuracy by combining real-time data access with language model capabilities.

4 Conclusion

In this project, we explored the effectiveness of **Linked Data**, traditional **browser search engines**, and **Large Language Models (LLMs)** in answering complex queries over structured datasets. Our findings indicate that, while **Linked Data** offers robust solutions for structured and precise query answering, **LLMs** enhance interpretative capabilities but require careful integration to mitigate inaccuracies. **Browser search engines** remain useful for general information retrieval but are less suited for complex, structured queries. The integration of Linked Data with LLMs represents a valuable direction for future research, aiming to combine the reliability of structured data with the flexibility of natural language understanding to achieve superior query answering performance.