LLM-based generation of semantic annotations and neuro-symbolic querying

IFB Biosphere annual meeting, 17th of June 2025, Nantes

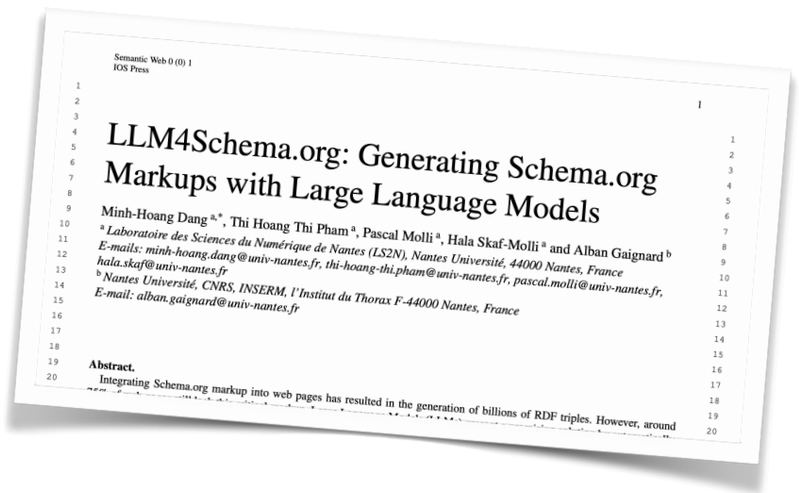
Alban Gaignard, Institut du Thorax, CNRS, Nantes

Inserm cnrs Nantes Université CHU NANTES

Laboratoire des Sciences du Numérique de Nantes

l'institut du thorax

Work in collaboration with LS2N, GDD team, funded by ANR Mekano, CominLabs MiKroloG

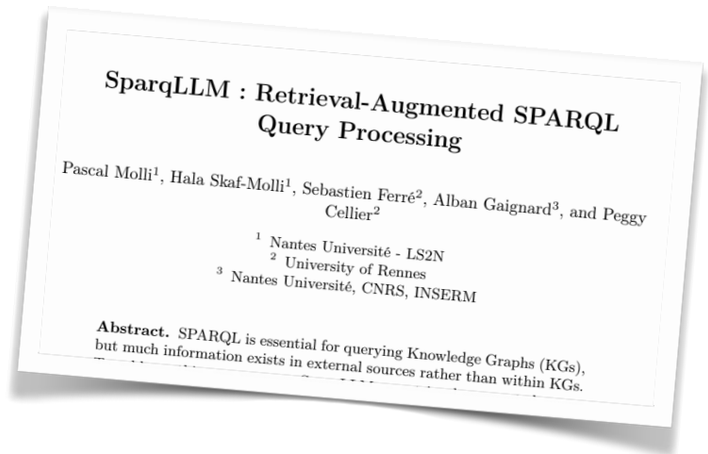

Semantic Web 0 (0) 1
IOS Press

LLM4Schema.org: Generating Schema.org Markups with Large Language Models

Minh-Hoang Dang [a,*], Thi Hoang Thi Pham [a], Pascal Molli [a], Hala Skaf-Molli [a] and Alban Gaignard [b]
[a] Laboratoire des Sciences du Numérique de Nantes (LS2N), Nantes Université, 44000 Nantes, France
E-mails: minh-hoang.dang@univ-nantes.fr, thi-hoang-thi.pham@univ-nantes.fr, pascal.molli@univ-nantes.fr, hala.skaf@univ-nantes.fr
[b] Nantes Université, CNRS, INSERM, l'Institut du Thorax F-44000 Nantes, France
E-mail: alban.gaignard@univ-nantes.fr

Abstract.
Integrating Schema.org markup into web pages has resulted in the generation of billions of RDF triples. However, around

[1] Minh-Hoang Dang, Thi Hoang Thi Pham, Pascal Molli, Hala Skaf-Molli, Alban Gaignard. ***LLM4Schema.org: Generating Schema.org Markups with Large Language Models.*** [pdf] Semantic Web Journal, 2025


SparqLLM : Retrieval-Augmented SPARQL Query Processing

Pascal Molli[1], Hala Skaf-Molli[1], Sebastien Ferré[2], Alban Gaignard[3], and Peggy Cellier[2]

[1] Nantes Université - LS2N
[2] University of Rennes
[3] Nantes Université, CNRS, INSERM

Abstract. SPARQL is essential for querying Knowledge Graphs (KGs), but much information exists in external sources rather than within KGs.

[2] Pascal Molli, Hala Skaf-Molli, Sébastien Ferré, Alban Gaignard and Peggy Cellier. ***SparqLLM : Retrieval-Augmented SPARQL Query Processing*** (Best Demo Nominee) [pdf] [online poster]. Extended Semantic Web Conférence 2025


Thi Hoang Thi Pham


Minh-Hoang Dang

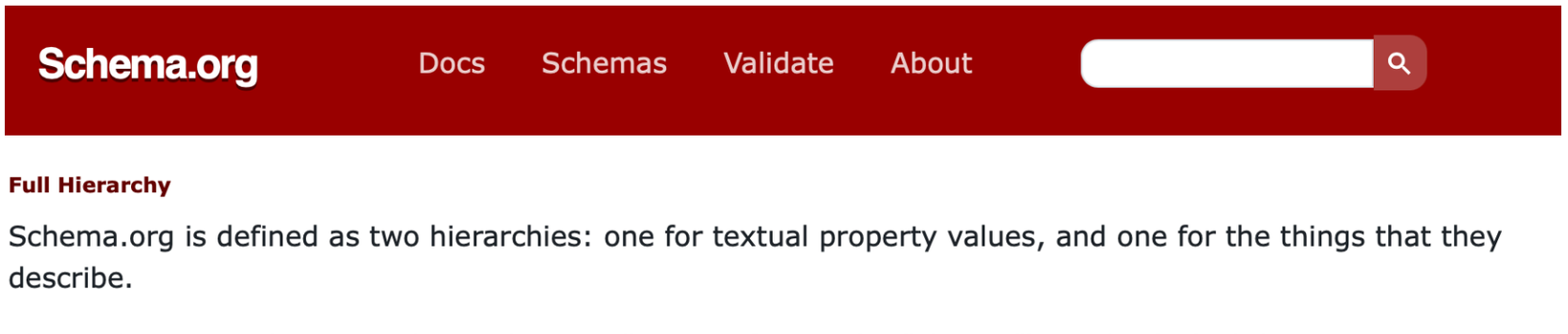

Pascall Molli


Hala Skaf-Molli

+ many thanks to the students involved in these projects through the computer science Master of Nantes University.

Context: schema.org

schema.org



- General purpose lightweight ontology
- Aimed at annotating web pages
- Targetting **FINDABILITY**
- Originating from major search engines

Schema.org is massively adopted

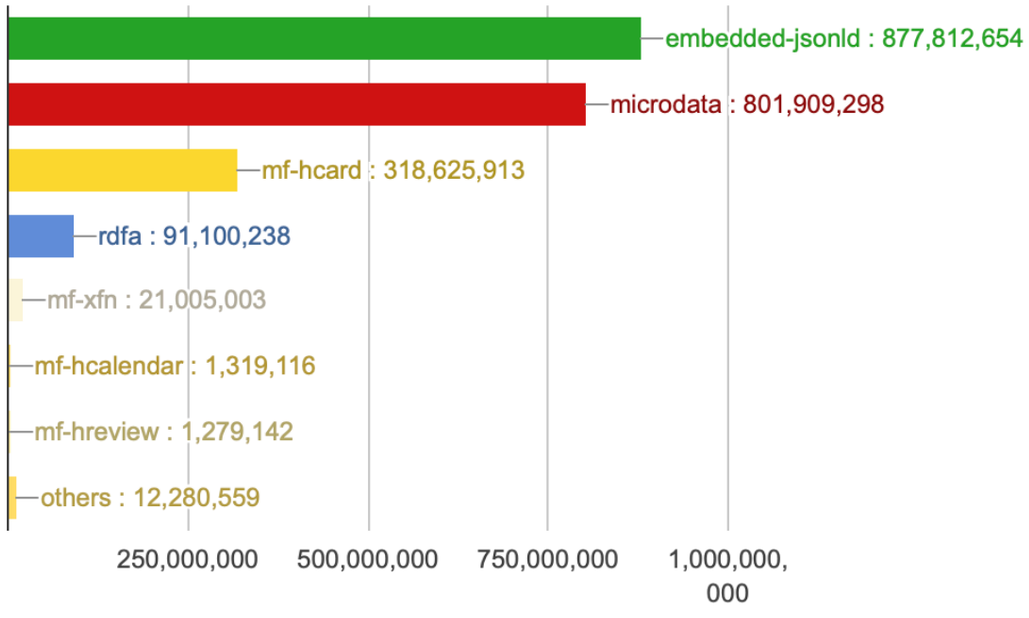
Web Data Commons

Extracting Structured Data from the Common Crawl

UNIVERSITY OF MANNHEIM

Crawl Date	October 2022	
Total Data	82.71 Terabyte	(compressed)
Parsed HTML URLs	3,048,746,652	
URLs with Triples	1,518,609,988	
Domains in Crawl	33,820,102	
Domains with Triples	14,235,035	
Typed Entities	19,072,628,514	
Triples	86,462,816,435	
Size of Extracted Data	1.6 Terabyte	(compressed)

URLs with Triples

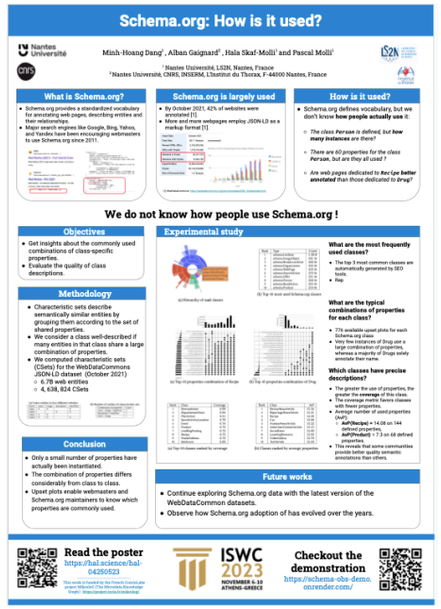


Top Domains by Extracted Triples

1. blogspot.com (879,564,145 triples)
2. wordpress.com (458,770,038 triples)
3. wikipedia.org (190,087,065 triples)
4. yummly.com (87,112,540 triples)
5. hotels.com (81,991,039 triples)
6. boohoo.com (79,884,394 triples)
7. kayak.com (77,623,248 triples)
8. google.com (73,729,078 triples)
9. yahoo.com (65,317,838 triples)
10. southleedslife.com (63,758,451 triples)
11. indiatimes.com (58,899,559 triples)
12. freepik.com (56,124,447 triples)
13. airbnb.com (51,964,983 triples)
14. pinterest.com (47,251,484 triples)
15. soundcloud.com (45,745,317 triples)
16. apple.com (42,410,414 triples)
17. hostadvice.com (42,309,867 triples)
18. elpais.com (42,136,136 triples)
19. vsemayki.ru (38,167,517 triples)
20. smugmug.com (38,031,434 triples)
21. More

Schema.org: how is it used ?

ISWC 2023



& online demo

Schema.org

Thing

- Action +
- BioChemEntity +
- CreativeWork +
- Event +
- Intangible +
- MedicalEntity +
- Organization +
- Person +
- Place +
- Product
  - DietarySupplement
  - Drug
  - IndividualProduct
  - ProductCollection
  - ProductGroup

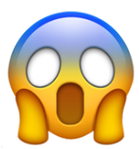
Most instanciated classes

700+ classes

49% of websites

86 billion RDF triples

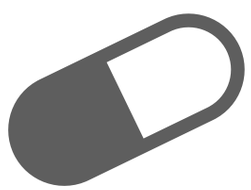
19 billion typed entities



Web Data Commons dataset , October 2021

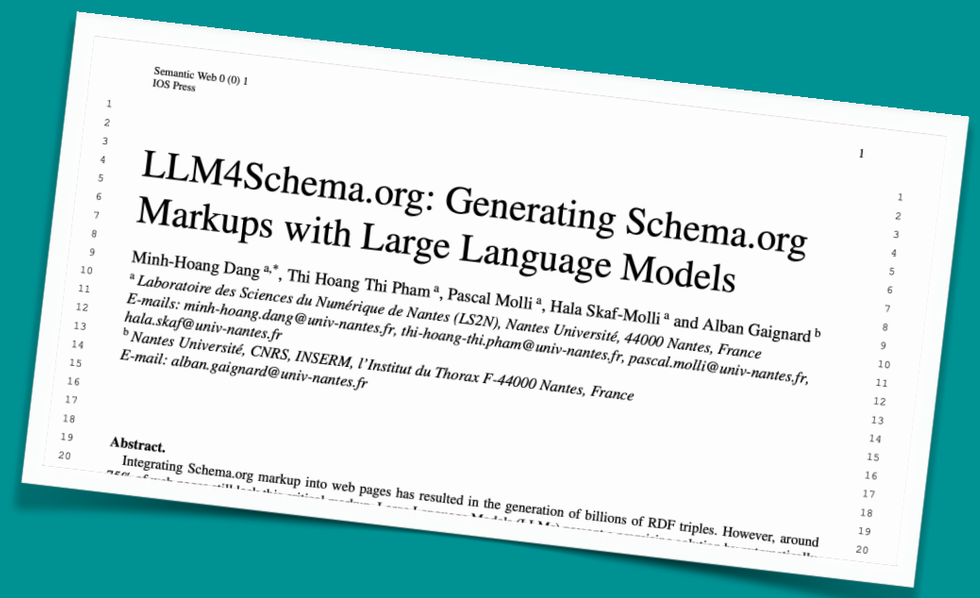
Top-K property combinations per class

"Best" Schema.org **annotation profiles** ?



VS

Minh-Hoang Dang, Alban Gaignard, Hala Skaf-Molli, Pascal Molli. Schema.org : How is it used?. International Semantic Web Conference (ISWC) 2023, Nov 2023, Athens, France. ⟨hal-04250523v2⟩



Can we use an LLM to annotate web pages with schema.org ?

[1] Minh-Hoang Dang, Thi Hoang Thi Pham, Pascal Molli, Hala Skaf-Molli, Alban Gaignard. ***LLM4Schema.org: Generating Schema.org Markups with Large Language Models.*** [pdf] Semantic Web Journal, 2025

Apple pie recipe

Recettes

Easy Apple Pie

Little Spoon Farm

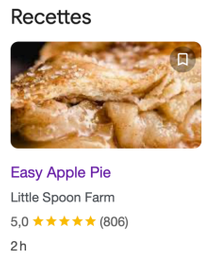
5,0 ★★★★★ (806)

2 h

```
<!DOCTYPE html>
<html lang="fr">
<head>
   <meta charset="UTF-8"/>
   <meta name="viewport" content="width=device-
        width, initial-scale=1.0"/>
   <title>Classic American Apple Pie</title>
   <script type="application/ld+json">
   { "@context": "http://schema.org/",
     "@type": "Recipe",
     "name": "Simple Apple Pie",
     "recipeCategory": "Dessert",
     "recipeCuisine": "American",
     "recipeIngredient": [
       "6 medium apples (Granny Smith and
            Honeycrisp mix)",
       "3/4 cup sugar",
       "2 tablespoons all-purpose flour",
       "1 teaspoon ground cinnamon",
       "The Eiffel Tower",
       "2 tablespoons unsalted butter",
       "2 premade pie crusts",
       "1 beaten egg (optional for brushing)",
       "A little sugar (optional for sprinkling)"
     ]
   }
   </script>
</head>
<body>
Preheat your oven to 375F (190C). Peel, core, and
     slice 6 medium apples (a mix of Granny Smith
     and Honeycrisp works well for flavor and
     texture). In a large bowl, mix the apple
     slices with 3/4 cup of sugar, 2 tablespoons of
      all-purpose flour, 1 teaspoon of ground
     cinnamon, and a pinch of salt. Roll out one
     premade pie crust and place it in a 9-inch pie
      dish. Trim the edge to 1/2 inch over the rim
     of the dish. Fill the crust with the apple
     mixture, and dot with 2 tablespoons of
     unsalted butter cut into small pieces. Roll
     out the second pie crust and place it over the
      filling. Trim, seal, and flute or crimp the
     edges. Cut slits in the top crust to allow
     steam to escape. Optional: For a golden crust,
      brush the top with abeaten egg and sprinkle
     with a little sugar. Enjoy your homemade
     american dessert apple pie!
</body>
</html>
```

Fig. 1. The Apple HTML web page mixing text in the body and JSON-LD describing apple-pie with its ingredients.

GPT 3.5 generation




```
"@context": "https://schema.org",
"@type": "Recipe",
"name": "Homemade American Dessert Apple Pie",
"recipeIngredient": [
  "6 medium apples (Granny Smith and Honeycrisp)",
    "3/4 cup sugar",
    "2 tablespoons all-purpose flour",
    "1 teaspoon ground cinnamon",
    "Pinch of salt",
    "2 premade pie crusts",
    "2 tablespoons unsalted butter" ],
"recipeInstructions": [
  { "@type": "HowToStep",
    "text": "Preheat your oven to 375F (190C)." },
  { "@type": "HowToStep",
    "text": "Peel, core, and slice the apples."},
  { "@type": "HowToStep",
    "text": "In a large bowl, mix the apple slices with
        ↪ sugar, flour, cinnamon, and salt."},
  { "@type": "HowToStep",
    "text": "Roll out one premade pie crust and place it in
        ↪  a 9-inch pie dish. Trim the edge to 1/2 inch
        ↪ over the rim of the dish."},
  {"@type": "HowToStep",
    "text": "Fill the crust with the apple mixture, and dot
        ↪  with unsalted butter cut into small pieces."
        ↪ },
  {"@type": "HowToStep",
    "text": "Roll out the second pie crust and place it
        ↪ over the filling. Trim, seal, and flute or
        ↪ crimp the edges."},
  {"@type": "HowToStep",
    "text": "Cut slits in the top crust to allow steam to
        ↪ escape."},
  { "@type": "HowToStep",
    "text": "Optional: For a golden crust, brush the top
        ↪ with a beaten egg and sprinkle with a little
        ↪ sugar."},
  {"@type": "HowToStep",
    "text": "Enjoy your homemade American dessert apple pi
        ↪ !"} ],
"recipeCategory": "Dessert",
"recipeCuisine": "American",
"nutrition": {
  "@type": "NutritionInformation",
  "calories": "Varies",
  "servingSize": "1 slice",
  "fatContent": "Varies",
  "carbohydrateContent": "Varies",
  "proteinContent": "Varies"
  }
```

Fig. 3. GPT3-5 generated Schema.org markup from the Apple pie text of Figure 1. Compared to the JSON-LD markup of Figure 1, GPT-3 produced the RecipeInstructions that are grounded in the text and nutrition facts that are not grounded in the text.

⚠️ **semantically incorrect** → no NutritionInformation class in schema.org

🤔 **non factual statements** → nothing said about nutrition in the web page

Objectives

Do LLMs generate more comprehensive schema.org markup than humans, given the text of a web page ?

(i) propose an annotation pipeline **limiting incorrectness and hallucinations**

(ii) propose a **scoring function** to compare which markup is the most comprehensive

Approach

```
You are an expert in the semantic web and have deep knowledge about
writing Schema.org markup. You will be given a document, a question, and
a series of hints. Your task is to give an answer using the hints. Reply
using only Yes or No.
Given the document below:
<content>[CONTENT]</content>
Hint 1: Check whether the [VALUE] is in the text.
Hint 2: Check whether the [VALUE] is [PROP].
Is [VALUE] a [PROP] of [TYPE]?
```

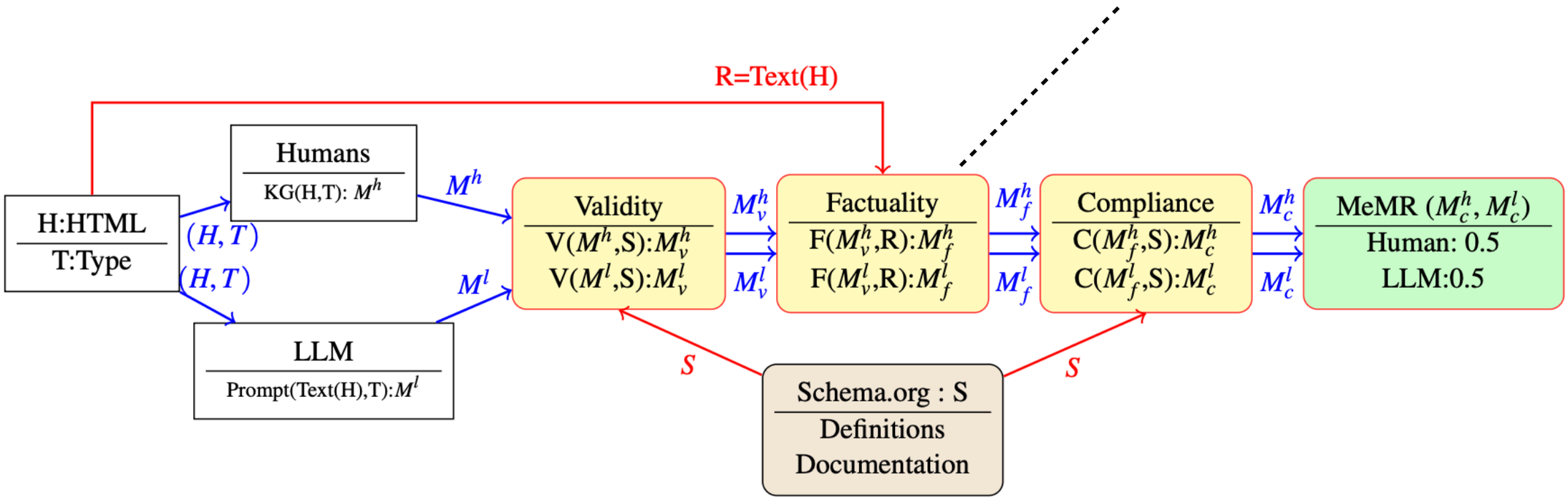


Fig. 4. LLM4Schema.org overall pipeline. Given a web page and a type T, it extracts a Human-markup M^h and generates LLM-markup M^l. It removes incorrect facts from both markups using Validity, Factuality, and Compliance agents. Finally, it compares which markup is the most comprehensive using the MeMR score function.

```
Given the markup below for property [RecipeIngredient]:
<markup> recipeIngredient "the Eiffel Tower" </markup>
Given the property definition below for [RecipeIngredient]:
<definition> A single ingredient used in the recipe, e.g., sugar or
garlic. </definition>
Does the value align with the property definition?
```

Experiments

Table 7

Results throughout the evaluation pipeline, where Input is the number of triples in the input. Valid, Fact, and Comp are the number of triples resulting from the step. The Rejection Rate is the percentage of triples rejected by the pipeline. MeMR: h for Human, l for LLMs.

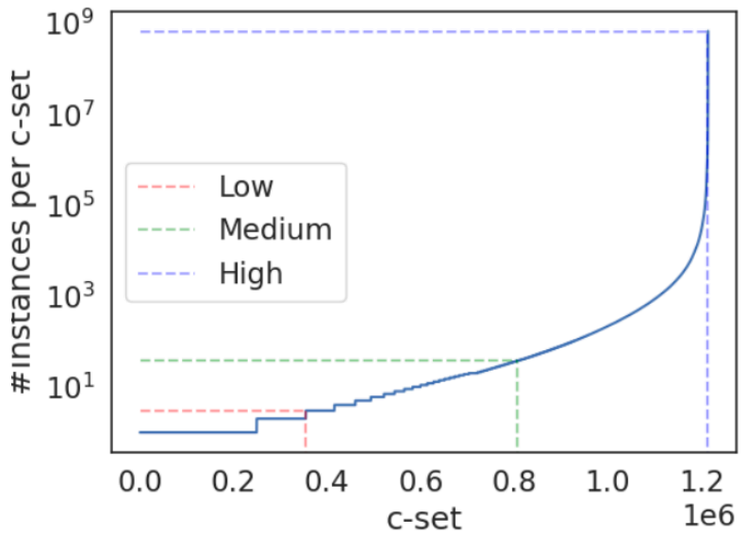
	Input	Valid.	Fact.	Comp.	Rejection Rate	MeMR
Human	5690	4875	3315	2719	52.2%	
GPT-3.5	4190	3369	2489	2055	50.9%	$(h = \mathbf{0.687}, l = 0.585)$
GPT-4	5260	4613	3573	3113	40.8%	$(h = 0.568, l = \mathbf{0.707})$

✨ After aplying our annotation pipeline, we observe less rejected anotations with **GPT4** compared to **GPT3.5** and **Humans**.

✨ MeMR score: **Humans** perform better **GPT3.5** than in the finally kept annotations

✨ MeMR score: **GPT4** better perform than **Humans** (larger contribution to the the merged set of annotations {human+LLM})

Experiments



longer combinations of properties tend to have fewer instances, but this is not the case for all types

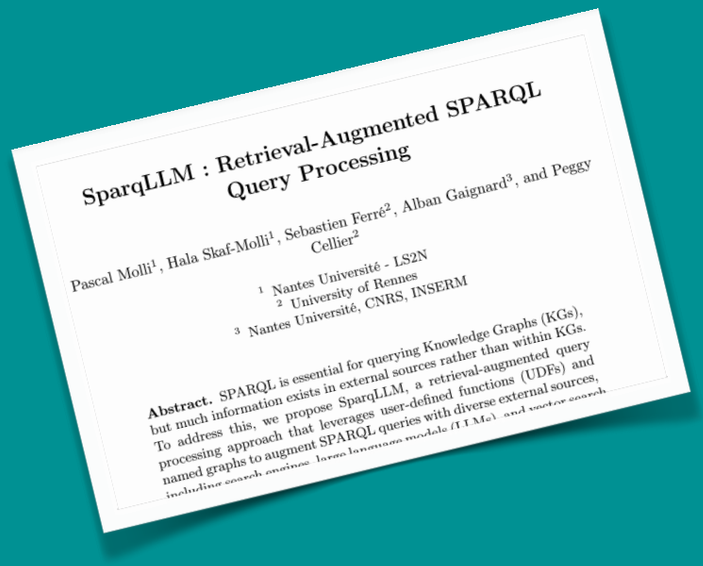
(a) Number of instances per C-set

	Human			**GPT-3.5**			**GPT-4**		
	Low	Med	High	Low	Med	High	Low	Med	High
Input	910	861	1039	1293	572	401	1286	619	671
Valid.	748	731	812	781	452	373	1064	592	619
Fact.	463	488	602	589	245	284	819	469	513
Comp.	402	346	515	549	222	256	683	441	448
RR	55%	59%	50%	57%	61%	36%	46%	28%	33%
MeMR				($h = 0.603$, $l = \mathbf{0.669}$)	($h = \mathbf{0.689}$, $l = 0.577$)	($h = \mathbf{0.802}$, $l = 0.506$)	($h = 0.491$, $l = \mathbf{0.795}$)	($h = 0.576$, $l = \mathbf{0.673}$)	($h = \mathbf{0.653}$, $l = 0.638$)

Take-home message

- LLM used as an extractor
- LLM agents for checking semantics and factuality (SHACL constraints for validating syntax)
- Need tooling ...

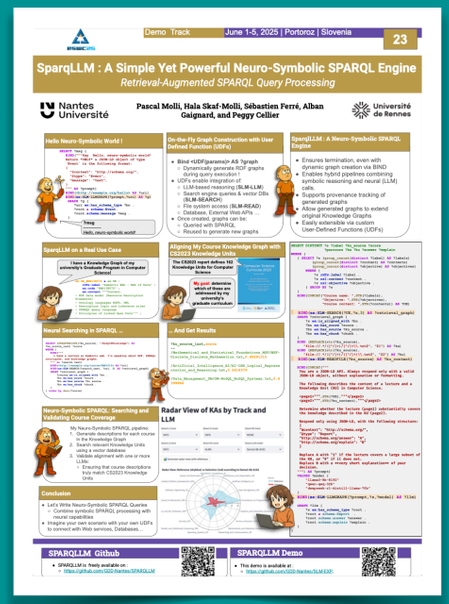
- Potential **bioinformatics applications**
  - Automating the annotation of Life-science web pages with Bioschemas (schema.org subset)
  - Automating the annotation with other ontologies (EDAM)
    $\rightarrow$ do we have enough examples ? a well documented schema ?


SparqLLM : Retrieval-Augmented SPARQL Query Processing

Pascal Molli[1], Hala Skaf-Molli[1], Sebastien Ferré[2], Alban Gaignard[3], and Peggy Cellier[2]

[1] Nantes Université - LS2N
[2] University of Rennes
[3] Nantes Université, CNRS, INSERM

Abstract. SPARQL is essential for querying Knowledge Graphs (KGs), but much information exists in external sources rather than within KGs. To address this, we propose SparqLLM, a retrieval-augmented query processing approach that leverages user-defined functions (UDFs) and named graphs to augment SPARQL queries with diverse external sources, including search engines, large language models (LLMs), and vector search



Can we query LLMs and Knowledge Graphs with SPARQL ?

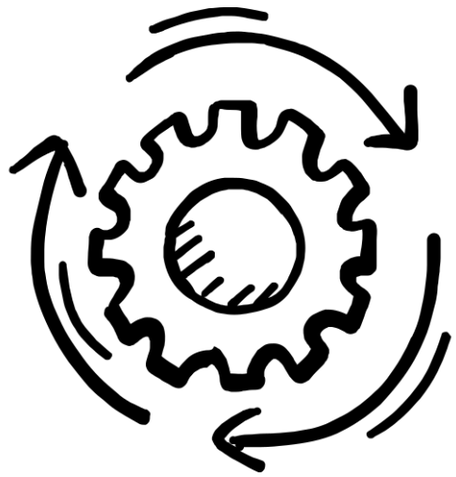
→ Neuro-symbolic SPARQL engine ?

[2] Pascal Molli, Hala Skaf-Molli, Sébastien Ferré, Alban Gaignard and Peggy Cellier. ***SparqLLM : Retrieval-Augmented SPARQL Query Processing*** (Best Demo Nominee) [pdf] [online poster]. Extended Semantic Web Conférence 2025

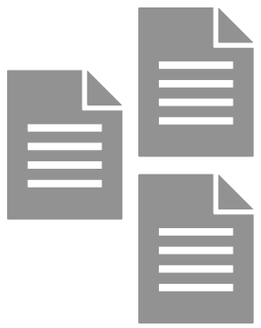
Query a LLM with SPARQL ?

- Needs :
  - KGs might be incomplete
  - Support provenance
  - Reason (with ontologies and rules) on LLM outputs
  - Do calculations on LLM outputs
  - Query non-RDF data sources

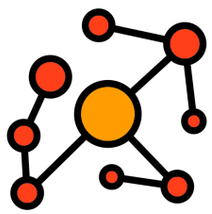
SPARQL



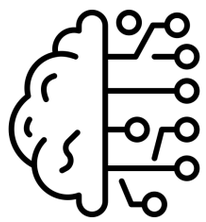
SPARQLLM



local documents



KG (e.g. uniprot, wikidata, etc.)



LLMs

Query a LLM with SPARQL ?

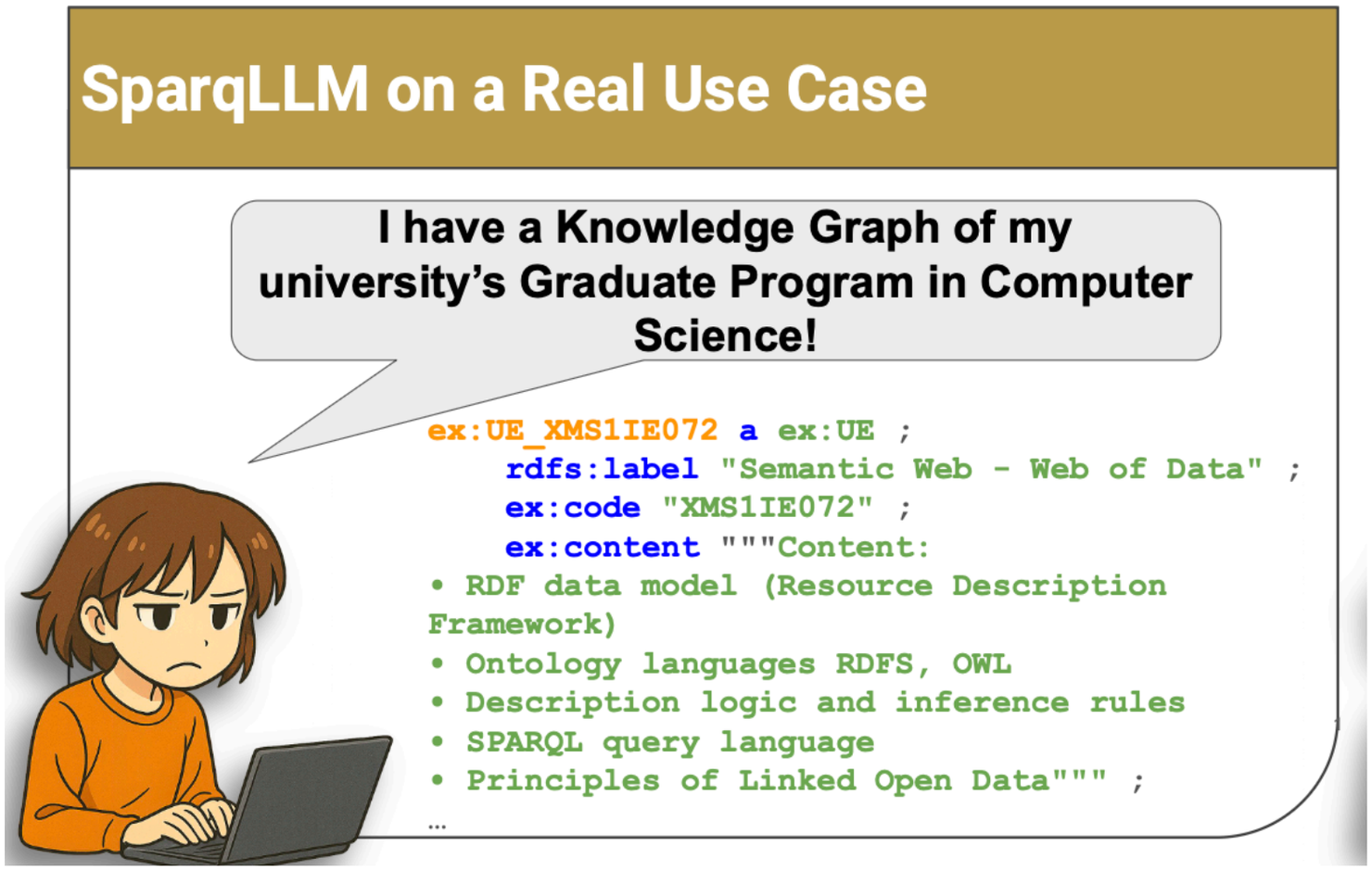
Hello Neuro-Symbolic World !

```
SELECT ?msg {
   BIND("""Say  Hello, neuro-symbolic world!
   Return *ONLY* a JSON-LD object of type
   `Event` in the following format:
   {
     "@context": "http://schema.org/",
     "@type": "Event",
     "message": "text",
   }
""" AS ?prompt)
BIND(<http://example.org/hello> AS ?uri)
BIND(ex:SLM-LLMGRAPH(?prompt,?uri) AS ?g)
 GRAPH ?g {
     ?uri ex:has_schema_type ?bn .
     ?root a schema:Event.
     ?root schema:message ?msg .
 }
}
```

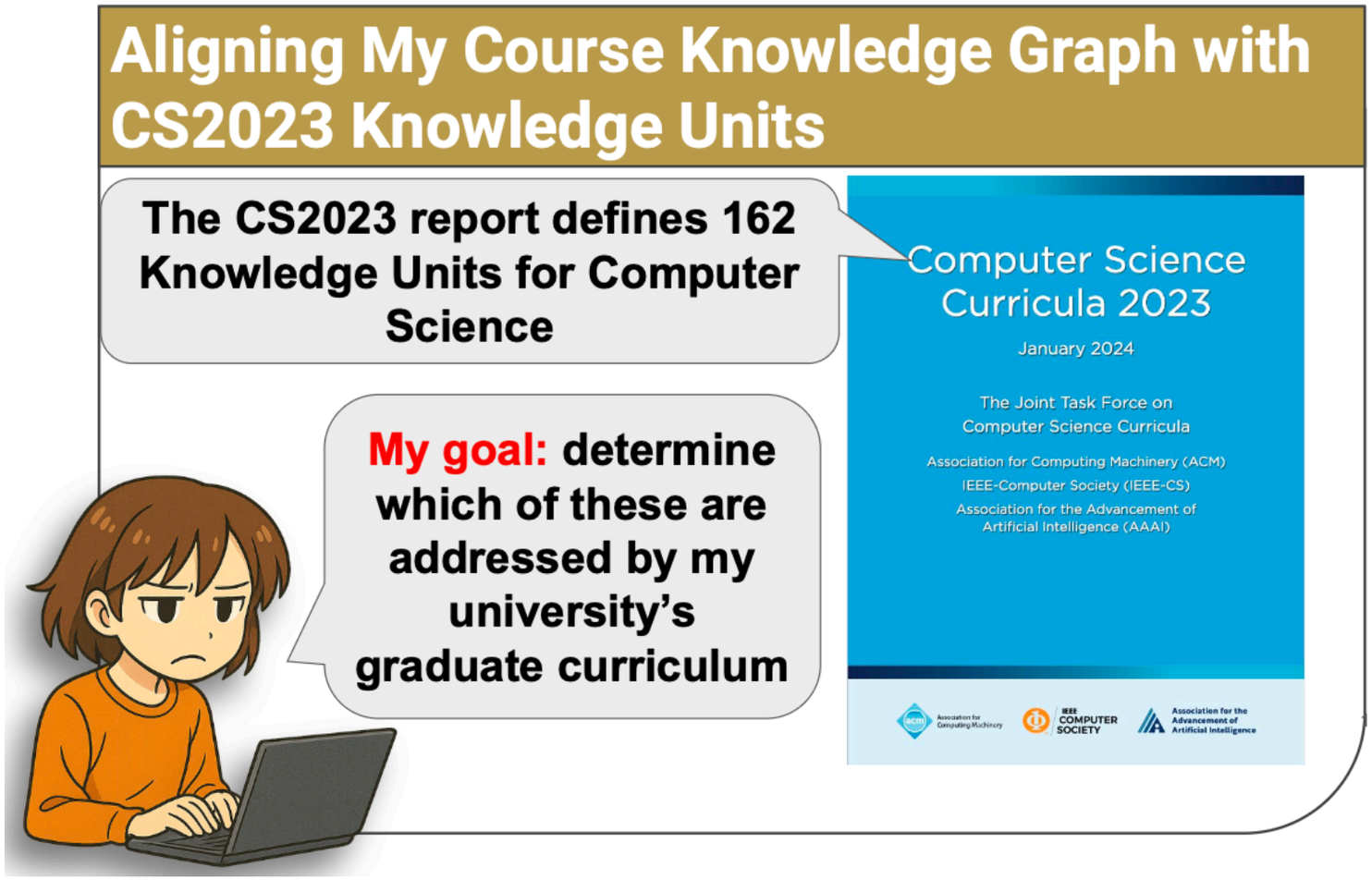
?msg

Hello, neuro-symbolic world!

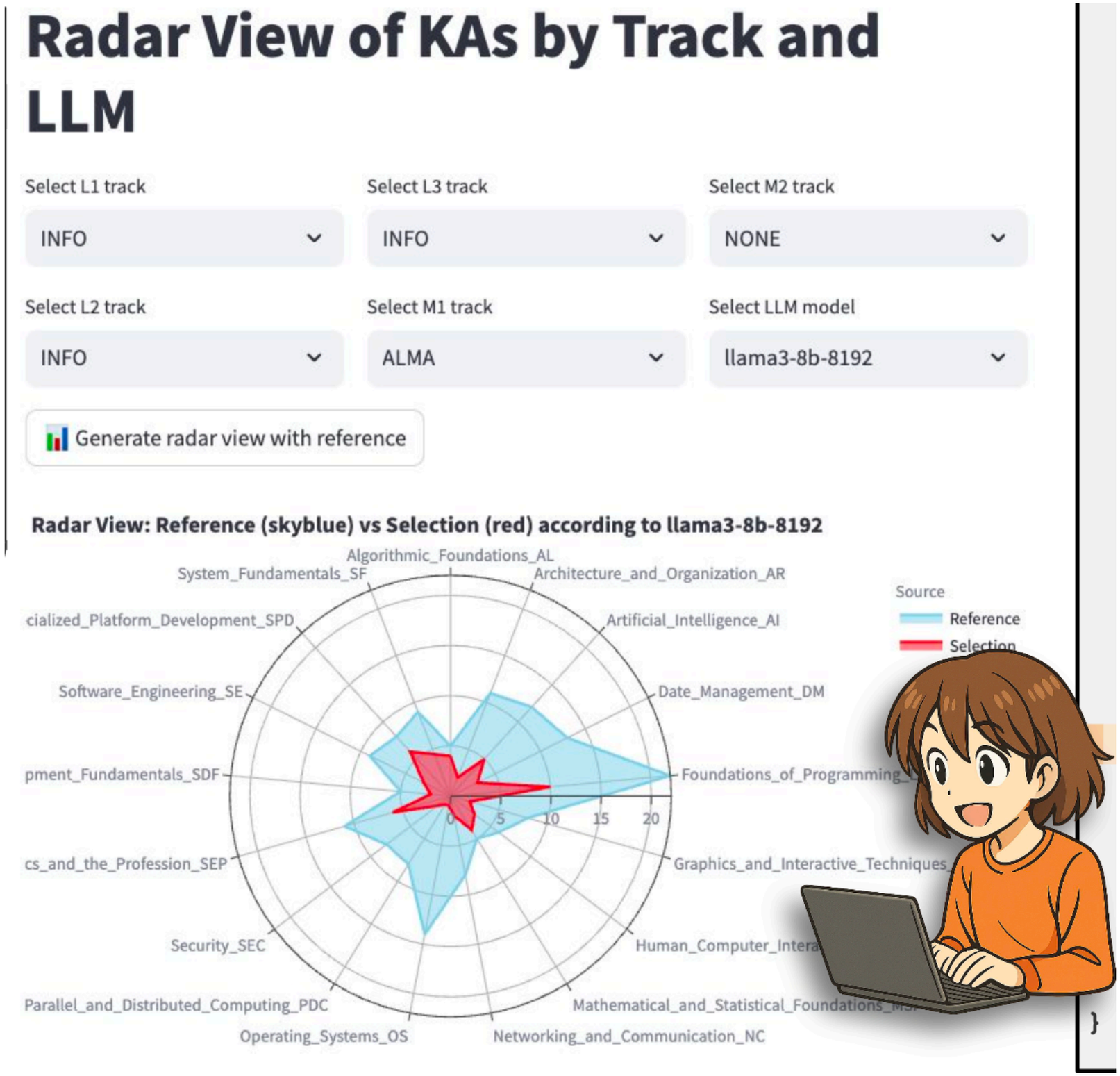
Query a LLM with SPARQL ?



Query a LLM with SPARQL ?



Expected results

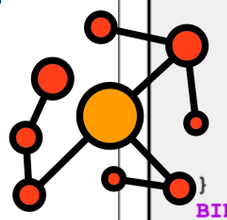


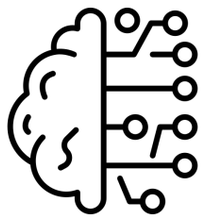
A SPARQL query to perform RAG

① Select course content and learning objectives

② Search matching textual document (vector database)

③ LLM to ask for a **yes/no** alignemnt, with **explanations**





LLMs

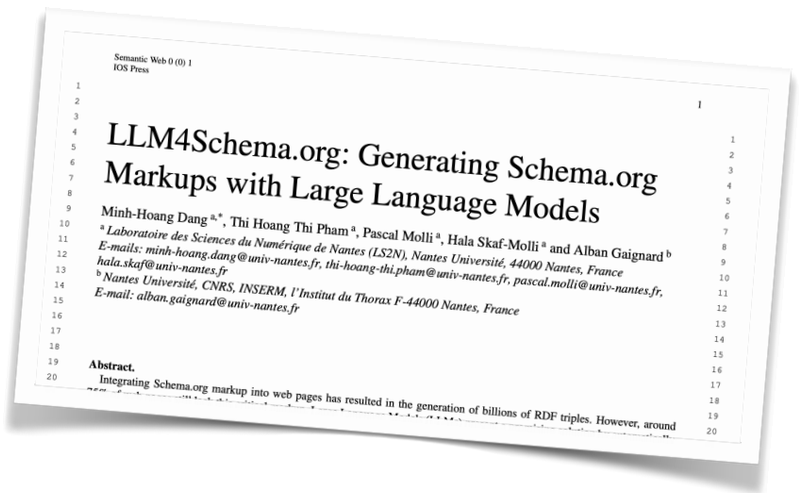
```
SELECT DISTINCT ?s ?label ?ku_source ?score
                ?parcours ?ka ?ku ?answer ?explain
WHERE {
   { SELECT ?s (group_concat(distinct ?label) AS ?labels)
          (group_concat(distinct ?content) AS ?contents)
          (group_concat(distinct ?objective) AS ?objectives)
      WHERE {
          ?s rdfs:label ?label .
          ?s ns1:content ?content .
          ?s ns1:objective ?objective .
      } GROUP BY ?s
   }
   BIND(CONCAT("Course name: ",STR(?labels),
               "Objective: ",STR(?objectives),
               "Course content: ",STR(?contents)) AS ?UE)

   BIND(ex:SLM-SEARCH(?UE,?s,3) AS ?retrieval_graph)
   GRAPH ?retrieval_graph {
       ?s ex:is_aligned_with ?bn .
       ?bn ex:has_score ?score .
       ?bn ex:has_source ?ku_source .
       ?bn ex:has_chunk ?chunk .
   }
   BIND (REPLACE(str(?ku_source),
      "file://.*/([^/]+)/[^/]+\\.txt$", "$1") AS ?ka)
   BIND (REPLACE(str(?ku_source),
      "file://.*/([^/]+)/([^/]+)\\.txt$", "$2") AS ?ku)
   BIND(ex:SLM-READFILE(?ku_source) AS ?ku_content)

   BIND(CONCAT("""
    You are a JSON-LD API. Always respond only with a valid
    JSON-LD object, without explanation or formatting.

    The following describes the content of a lecture and a
    Knowledge Unit (KU) in Computer Science.

    <page1>""",STR(?UE),"""</page1>
    <page2>""",STR(?ku_content),"""</page2>

    Determine whether the lecture (page1) substantially covers
    the knowledge described in the KU (page2).

    Respond only using JSON-LD, with the following structure:
    {
    "@context": "http://schema.org/",
    "@type": "Report",
    "http://schema.org/answer": "A",
    "http://schema.org/explain": "B"
    }

    Replace A with "1" if the lecture covers a large subset of
    the KU, or "0" if it does not.
    Replace B with a **very short explanation** of your
    decision.
   """) AS ?prompt)
   VALUES ?model {
       "llama3-8b-8192"
       "qwen-qwq-32b"
       "deepseek-r1-distill-llama-70b"
   }
   BIND(ex:SLM-LLMGRAPH(?prompt,?s,?model) AS ?llm)

   GRAPH ?llm {
       ?s ex:has_schema_type ?root .
       ?root a schema:Report  .
       ?root schema:answer ?answer .
       ?root schema:explain ?explain .
   }
}
```

Take-home message

- A Neuro-Symbolic query engine
  - local or remote LLM
  - easily extensible with **user-defined functions** (UDF)

- Potential **bioinformatics applications**
  - Query a biomed KG with relevant scientific papers ?
  - Complement a biomed KG with scientific context and results described in a paper ?

Work in collaboration with LS2N, GDD team, funded by ANR Mekano, CominLabs MiKroloG


Semantic Web 0 (0) 1
IOS Press

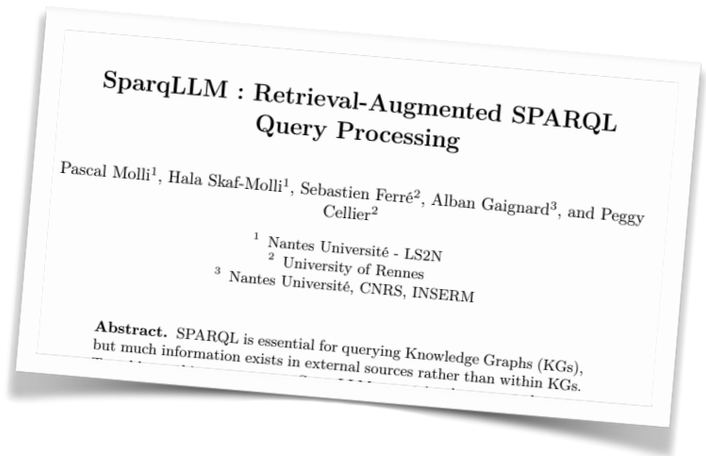
1

LLM4Schema.org: Generating Schema.org Markups with Large Language Models

Minh-Hoang Dang [a,*], Thi Hoang Thi Pham [a], Pascal Molli [a], Hala Skaf-Molli [a] and Alban Gaignard [b]
[a] Laboratoire des Sciences du Numérique de Nantes (LS2N), Nantes Université, 44000 Nantes, France
E-mails: minh-hoang.dang@univ-nantes.fr, thi-hoang-thi.pham@univ-nantes.fr, pascal.molli@univ-nantes.fr, hala.skaf@univ-nantes.fr
[b] Nantes Université, CNRS, INSERM, l'Institut du Thorax F-44000 Nantes, France
E-mail: alban.gaignard@univ-nantes.fr

Abstract.
Integrating Schema.org markup into web pages has resulted in the generation of billions of RDF triples. However, around

[1] Minh-Hoang Dang, Thi Hoang Thi Pham, Pascal Molli, Hala Skaf-Molli, Alban Gaignard. ***LLM4Schema.org: Generating Schema.org Markups with Large Language Models.*** [pdf] Semantic Web Journal, 2025


SparqLLM : Retrieval-Augmented SPARQL Query Processing

Pascal Molli[1], Hala Skaf-Molli[1], Sebastien Ferré[2], Alban Gaignard[3], and Peggy Cellier[2]
[1] Nantes Université - LS2N
[2] University of Rennes
[3] Nantes Université, CNRS, INSERM

Abstract. SPARQL is essential for querying Knowledge Graphs (KGs), but much information exists in external sources rather than within KGs.

[2] Pascal Molli, Hala Skaf-Molli, Sébastien Ferré, Alban Gaignard and Peggy Cellier. ***SparqLLM : Retrieval-Augmented SPARQL Query Processing*** (Best Demo Nominee) [pdf] [online poster]. Extended Semantic Web Conférence 2025


Thi Hoang Thi Pham


Minh-Hoang Dang


Pascall Molli


Hala Skaf-Molli

+ many thanks to the students involved in these projects through the computer science Master of Nantes University.

Any questions ?

Inserm cnrs Nantes Université CHU NANTES

LS2N Laboratoire des Sciences du Numérique de Nantes

l'institut du thorax