LACAM - ARA

DEPARTMENT OF COMPUTER SCIENCE - ARTIFICIAL INTELLIGENCE

Master's Thesis in Semantic Technologies and Knowledge Graphs

Automated Prompt Engineering using the Prompt Engineering Ontology for Querying Large Language Models

Supervisor:
Prof. Claudia d'Amato
Co-supervisor:
Dott. Roberto Barile
Dott. Andrea Nuzzolese

Candidate:
Francioso Donato

Objective

Our **objective** is to evaluate whether the integration of structured knowledge can lead to the automatic generation of prompts that are qualitatively better to those obtainable without it.

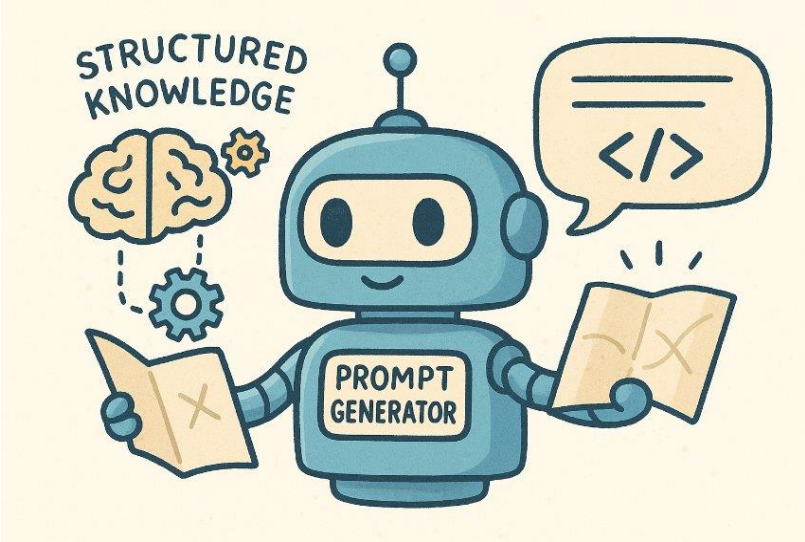


Image generated with ChatGPT

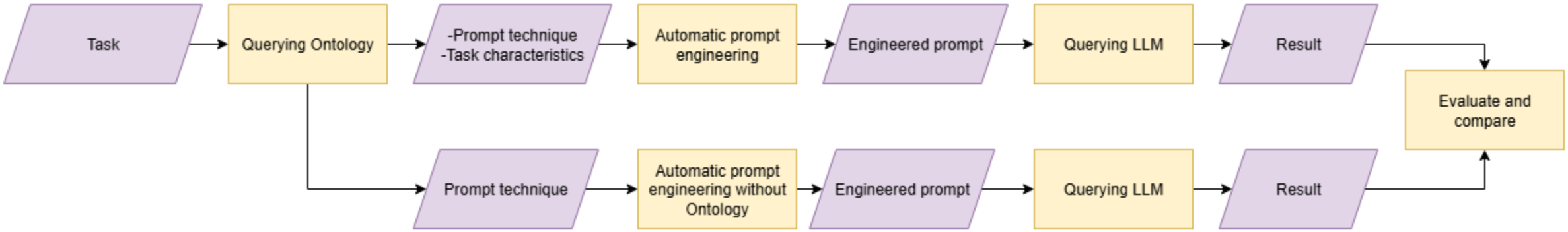
Context

- **Large Language Models (LLM):** artificial intelligence models capable of generating and understanding natural language
- **Automated Prompt Engineering:** generate high-quality prompts automatically
- **Ontology:** shared conceptualisation of a specific domain.
  - We need an ontology that formalises **prompt engineering**

Research Question

•**RQ**: Does the ontology improve the effectiveness of automatic prompt engineer compared to LLM?

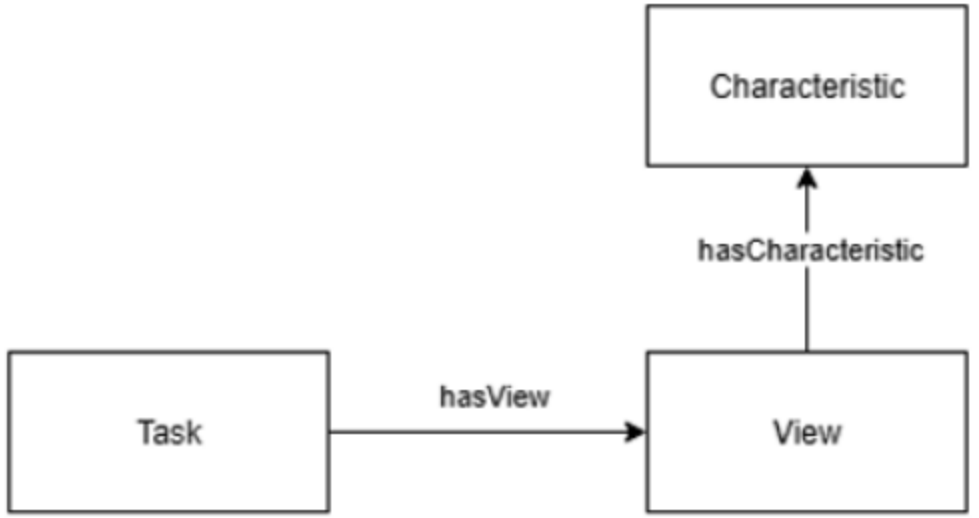
Adopted Pipeline



Adopted ontology

LACAM - ARA

- **PEO** (Prompt Engineering Ontology): describes fundamental concepts related to **LLMs** and **prompt engineering**
  - Developed by LACAM-ARA
- Task description missing and we extend it



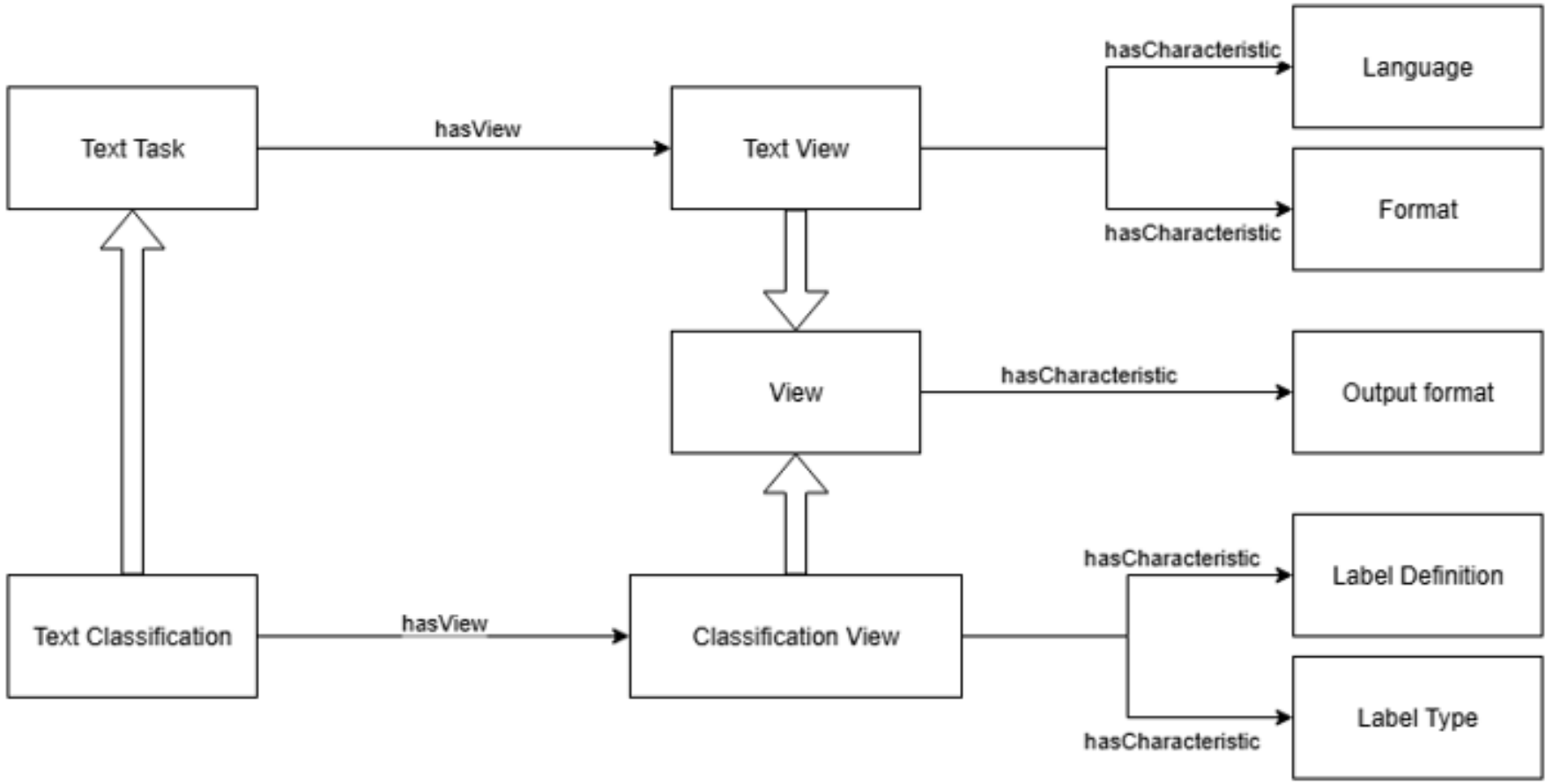
The **Linked Open Terms (LOT)** model methodology adopted.

Ontology pattern used:

- Classification
- Parameter

PEO ontology is available online

Ontology extension



Task and adopted dataset

- The [Civil Comments dataset] was used
- Each comment is assigned a **toxicity score** (a numeric value)

haha you guys are a bunch of losers.	0.893617
ur a sh*tty comment.	0.666667
hahahahahahahahhha suck it.	0.457627
FFFFUUUUUUUUUUUUUU	0
The ranchers seem motivated by mostly by greed; no one…	0

- The labels were **binarized**, considering a comment **toxic** if its score is **greater than 0.5**
- **100 comments** were selected, with a specific distribution: 60 non-toxic and 40 toxic

Experimental setup

LACAM - ARA

- **Automatic Engineering and Testing of Ontology-Driven Prompts**
  - With varying characteristic configurations:
    - Language, Label Definition, Label Type, Format
    - Language, Label Definition, Label Type
    - Language, Label Definition
- **Automatic Engineering and Testing of Prompts without Ontology**
  - Based only on the LLM's intrinsic knowledge
- Tests Performed (multiple LLM and Technique adopted)

This phase will answer the Research Quesiton

Experimental setup

- **SPARQL:** to query ontology
- **Selected LLMs:** Gemini 2.0 and DeepSeek V3
- **Prompting Techniques:**
  - **Few-shot:** Provide guide examples within the prompt.
  - **Chain-of-Thought:** Stimulates reasoning through a sequence of intermediate steps

```
SELECT DISTINCT ?label ?comment
WHERE {
    ?subClass rdfs:label ?label.
    ?subClass rdfs:comment ?comment .
    ?char rdf:type ?subClass .
    ?subClass rdfs:subClassOf* peo:Characteristic .

    ?view peo:hasCharacteristic ?char.
    ?view rdf:type ?viewClass.
    ?viewClass rdfs:subClassOf* peo:View.

    ?task peo:hasView ?view.
    ?task rdf:type peo:Task.
}
```

Query SPARQL to extract task chatacteristics

Experimental Results

LACAM - ARA

Model	Ontology	# Characteristics	Prompt technique	# Precision	# Recall	# F1-score
Gemini 2.0	☑	Two	Few-shot	0,75	0,70	0,72
Gemini 2.0	☑	Two	CoT	0,74	0,63	0,68
DeepSeek V3	☑	Three	Few-shot	0,72	0,72	0,70
DeepSeek V3	☑	Three	CoT	0,76	0,73	0,74
Gemini 2.0	☐		Few-shot	0,70	0,59	0,64
Gemini 2.0	☐		CoT	0,73	0,61	0,66
DeepSeek V3	☐		Few-shot	0,70	0,70	0,67
DeepSeek V3	☐		CoT	0,63	0,57	0,58

RQ: Does the ontology improve the effectivenes of automatic prompt engineer compared to LLM?

Ontology-driven prompts on varying LLMS

LACAM - ARA

Gemini

# Characteristics	Prompt technique	# Precision	# Recall	# F1-score
Two	Few-shot	0,75	0,70	0,72
Two	CoT	0,74	0,63	0,68
Three	Few-shot	0,78	0,44	0,56
Three	CoT	0,72	0,66	0,69
Four	Few-shot	0,68	0,62	0,65
Four	CoT	0,78	0,71	0,75

DeepSeek

# Characteristics	Prompt technique	# Precision	# Recall	# F1-score
Two	Few-shot	0,72	0,73	0,72
Two	CoT	0,74	0,69	0,71
Three	Few-shot	0,72	0,72	0,70
Three	CoT	0,76	0,73	0,74
Four	Few-shot	0,74	0,74	0,72
Four	CoT	0,76	0,68	0,70

LLM failure in violation of syntactic requirements

LACAM - ARA

Error: LLM returns results in a format different from the one requested.

Gemini

Ontology	# Characteristics	Prompt technique	# Errors	# Errors with Toxicity = 1
☑	Four	CoT	8	4
☑	Four	Few-shot	9	6
☑	Three	CoT	9	4
☑	Three	Few-shot	45	15
☑	Two	CoT	15	8
☑	Two	Few-shot	4	1
☐		CoT	17	8
☐		Few-shot	16	7

DeepSeek

Ontology	# Characteristics	Prompt technique	# Errors	# Errors with Toxicity = 1
☑	Four	CoT	12	3
☑	Four	Few-shot	1	0
☑	Three	CoT	4	1
☑	Three	Few-shot	0	0
☑	Two	CoT	8	4
☑	Two	Few-shot	0	0
☐		CoT	11	3
☐		Few-shot	0	0

Conclusions

- Ontology-based automated prompt engineering **outperforms** LLM-based automated prompt engineering
- Ontology-based automated prompt engineering **reduces** the steps needed to engineer optimized prompts automatically with respect to LLM only
- Ontology-based automated prompt engineering enables **full customization** of the prompting environment

Future Works

- **Extend PEO** ontology with additional tasks
- **Further experiments** with additional LLMs and/or datasets
- Develop a **hybrid solution** enabling an interactive user-LLM dialogue for dynamic ontology-driven prompt engineering

Thanks for your attention