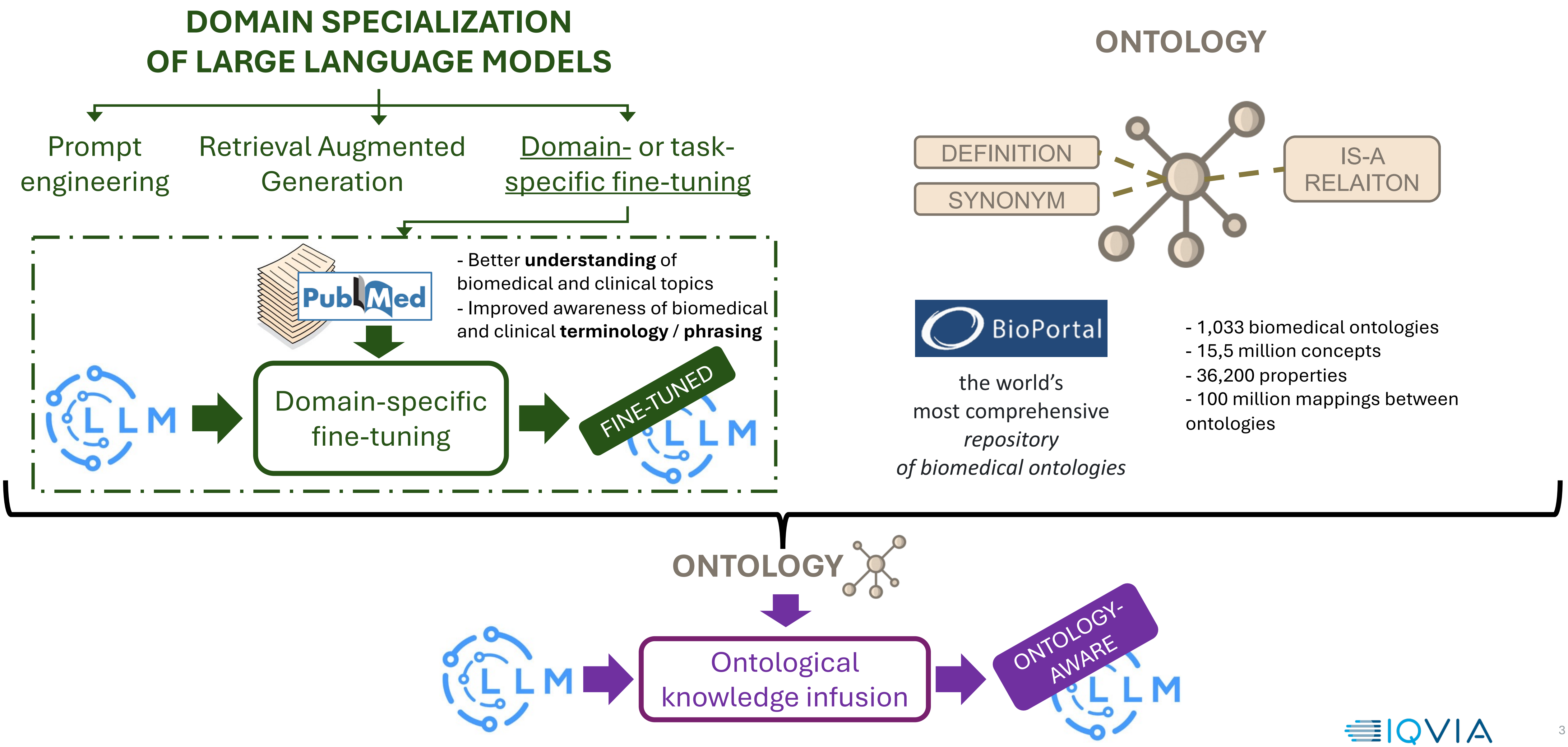
Towards Ontology-Enhanced Representation Learning for Large Language Models

Francesco Ronzano and Jay Nanavati
IQVIA Advanced NLP Team

Agenda

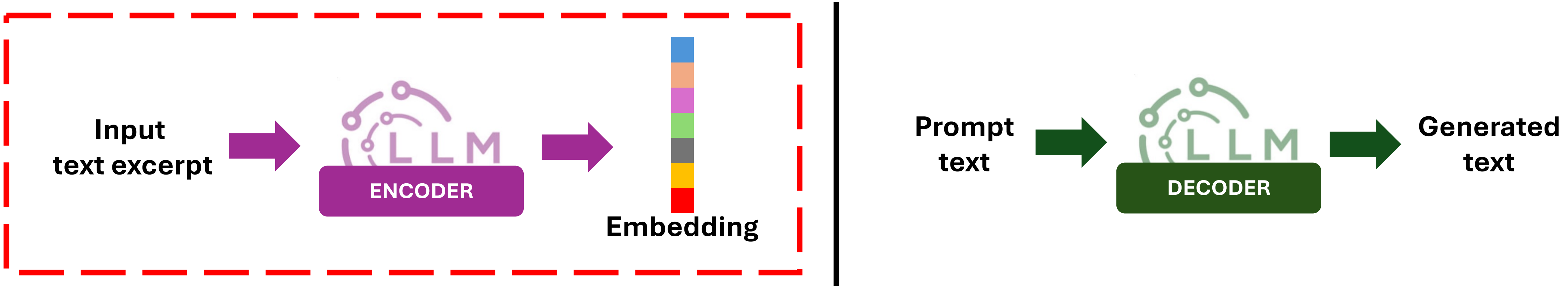
+ Exploiting ontologies to teach Large Language Models
+ The ontological knowledge infusion approach
    - Choosing the *source ontology* and the *target embedding-Large Language Model*
    - Generating *synthetic definitions of concepts*
    - Selecting *positive and negative pairs of concept definitions*, driven by the ontology
    - Fine-tuning the embedding-Large Language Model by *contrastive representation learning*
+ Evaluation: do embedding-Large Language Models better understand diseases, after infusing the MONDO disease ontology?
+ Key learnings
+ Next steps

Exploiting ontologies to teach Large Language Models



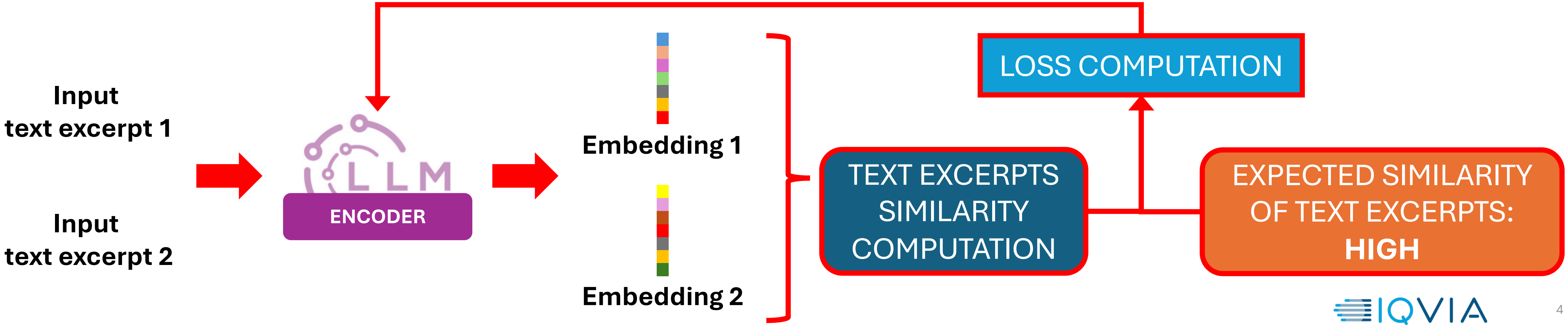
The ontological knowledge infusion approach

Which Large Language Model architecture do we target to infuse ontological knowledge?



Which fine-tuning framework is exploited to support ontological knowledge infusion?

Contrastive learning framework



Text A.1: A car is a four-wheeled road vehicle.
Text A.2: A car is a mean of transport moving on wheels.

The ontological knowledge infusion approach

Which fine-tuning framework is exploited to support ontological knowledge infusion? (cont.)

Contrastive training objective: **InfoNCE**

- Text A.1, Text A.2
- Text B.1, Text B.2
- Text C.1, Text C.2
- Text D.1, Text D.2

Both texts of each pair share the same meaning

LLM ENCODER

Text similarity matrix

	A.2	B.2	C.2	D.2
A.1	**sim**(A.1, A.2)	**sim**(A.1, B.2)	**sim**(A.1, C.2)	**sim**(A.1, D.2)
B.1	**sim**(B.1, A.2)	**sim**(B.1, B.2)	**sim**(B.1, C.2)	**sim**(B.1, D.2)
C.1	**sim**(C.1, A.2)	**sim**(C.1, B.2)	**sim**(C.1, C.2)	**sim**(C.1, D.2)
D.1	**sim**(D.1, A.2)	**sim**(D.1, B.1)	**sim**(D.1, C.2)	**sim**(D.1, D.2)

sim: similarity function among text embeddings

COMPUTATION OF **CATEGORICAL CROSS-ENTROPY LOSS**

Predicted probability distribution

softmax(	**sim**(A.1, A.2)	**sim**(A.1, B.2)	**sim**(A.1, C.2)	**sim**(A.1, D.2))
softmax(	**sim**(B.1, A.2)	**sim**(B.1, B.2)	**sim**(B.1, C.2)	**sim**(B.1, D.2))
softmax(	**sim**(C.1, A.2)	**sim**(C.1, B.2)	**sim**(C.1, C.2)	**sim**(C.1, D.2))
softmax(	**sim**(D.1, A.2)	**sim**(D.1, B.2)	**sim**(D.1, C.2)	**sim**(D.1, D.2))

Expected probability distribution

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Text A.1: A car is a four-wheeled road vehicle.
Text A.2: A car is a mean of transport moving on wheels.
Text A.3n: Two over six wheels of that bus were damaged because of the accident with a car.

The ontological knowledge infusion approach

Which fine-tuning framework is exploited to support ontological knowledge infusion? (cont.)

Contrastive training objective: **InfoNCE** with (hard-)negative texts

- Text A.1, Text A.2 + Text A.3n
- Text B.1, Text B.2 + Text B.3n
- Text C.1, Text C.2 + Text C.3n
- Text D.1, Text D.2 + Text D.3n

Both texts of each pair share the same meaning

LLM ENCODER

Text similarity matrix

	A.2	B.2	C.2	D.2	A.3n	B.3n	C.3n	D.3n
A.1	**sim**(A.1, A.2)	**sim**(A.1, B.2)	**sim**(A.1, C.2)	**sim**(A.1, D.2)	**sim**(A.1, A.3n)	**sim**(A.1, B.3n)	**sim**(A.1, C.3n)	**sim**(A.1, D.3n)
B.1	**sim**(B.1, A.2)	**sim**(B.1, B.2)	**sim**(B.1, C.2)	**sim**(B.1, D.2)	**sim**(B.1, A.3n)	**sim**(B.1, B.3n)	**sim**(B.1, C.3n)	**sim**(B.1, D.3n)
C.1	**sim**(C.1, A.2)	**sim**(C.1, B.2)	**sim**(C.1, C.2)	**sim**(C.1, D.2)	**sim**(C.1, A.3n)	**sim**(C.1, B.3n)	**sim**(C.1, C.3n)	**sim**(C.1, D.3n)
D.1	**sim**(D.1, A.2)	**sim**(D.1, B.1)	**sim**(D.1, C.2)	**sim**(D.1, D.2)	**sim**(D.1, A.3n)	**sim**(D.1, B.3n)	**sim**(D.1, C.3n)	**sim**(D.1, D.3n)

sim: similarity function among text embeddings

COMPUTATION OF **CATEGORICAL CROSS-ENTROPY LOSS**

Predicted probability distribution

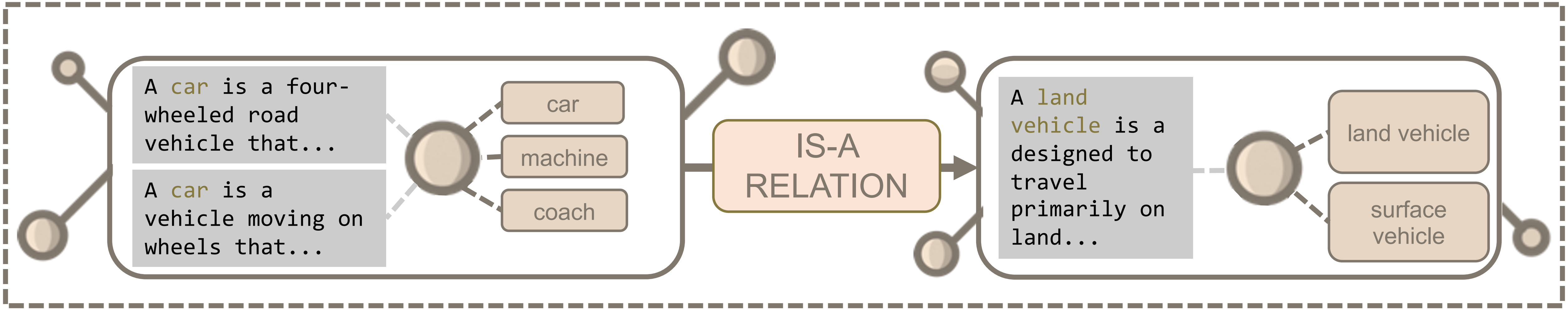
softmax(	**sim**(A.1, A.2)	**sim**(A.1, B.2)	**sim**(A.1, C.2)	**sim**(A.1, D.2)	**sim**(A.1, A.3n)	**sim**(A.1, B.3n)	**sim**(A.1, C.3n)	**sim**(A.1, D.3n))
softmax(	**sim**(B.1, A.2)	**sim**(B.1, B.2)	**sim**(B.1, C.2)	**sim**(B.1, D.2)	**sim**(B.1, A.3n)	**sim**(B.1, B.3n)	**sim**(B.1, C.3n)	**sim**(B.1, D.3n))
softmax(	**sim**(C.1, A.2)	**sim**(C.1, B.2)	**sim**(C.1, C.2)	**sim**(C.1, D.2)	**sim**(C.1, A.3n)	**sim**(C.1, B.3n)	**sim**(C.1, C.3n)	**sim**(C.1, D.3n))
softmax(	**sim**(D.1, A.2)	**sim**(D.1, B.1)	**sim**(D.1, C.2)	**sim**(D.1, D.2)	**sim**(D.1, A.3n)	**sim**(D.1, B.3n)	**sim**(D.1, C.3n)	**sim**(D.1, D.3n))

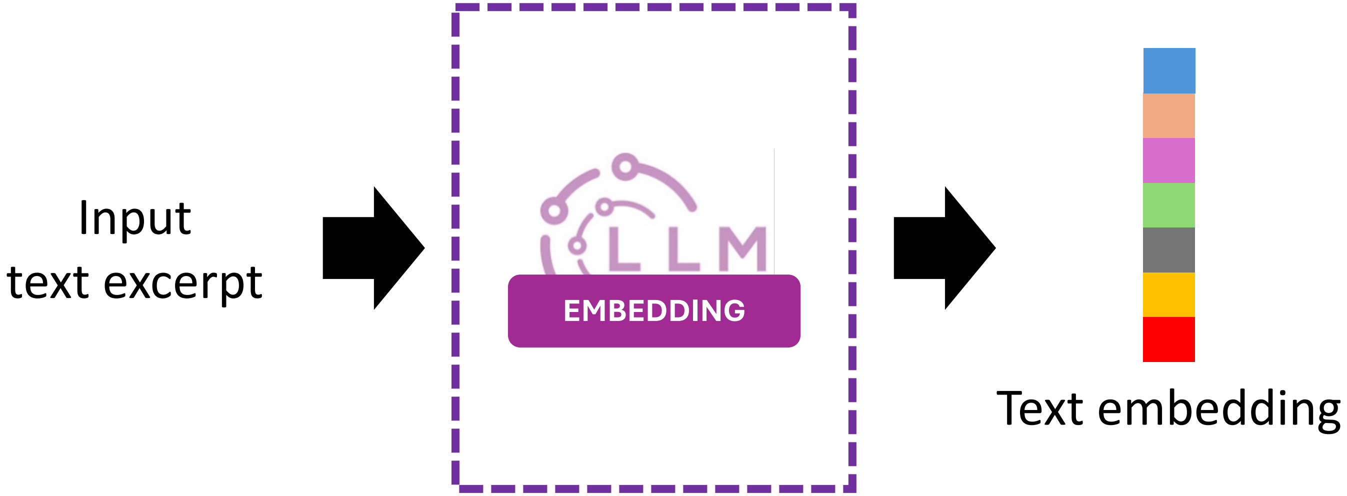
Expected probability distribution

1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0

The ontological knowledge infusion approach

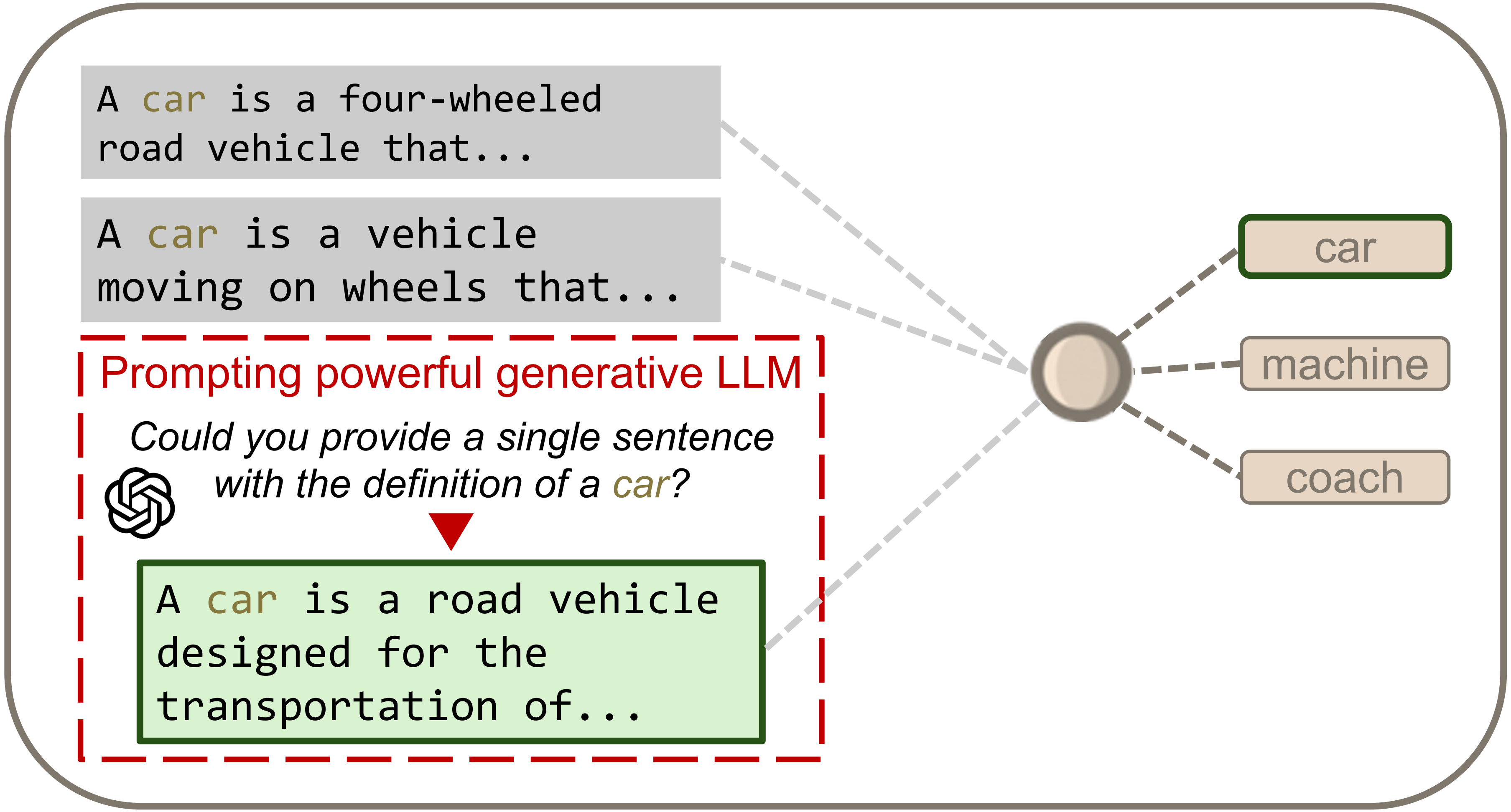
STEP 1: Choosing the **source ontology** and the **target embedding-Large Language Model**





The ontological knowledge infusion approach

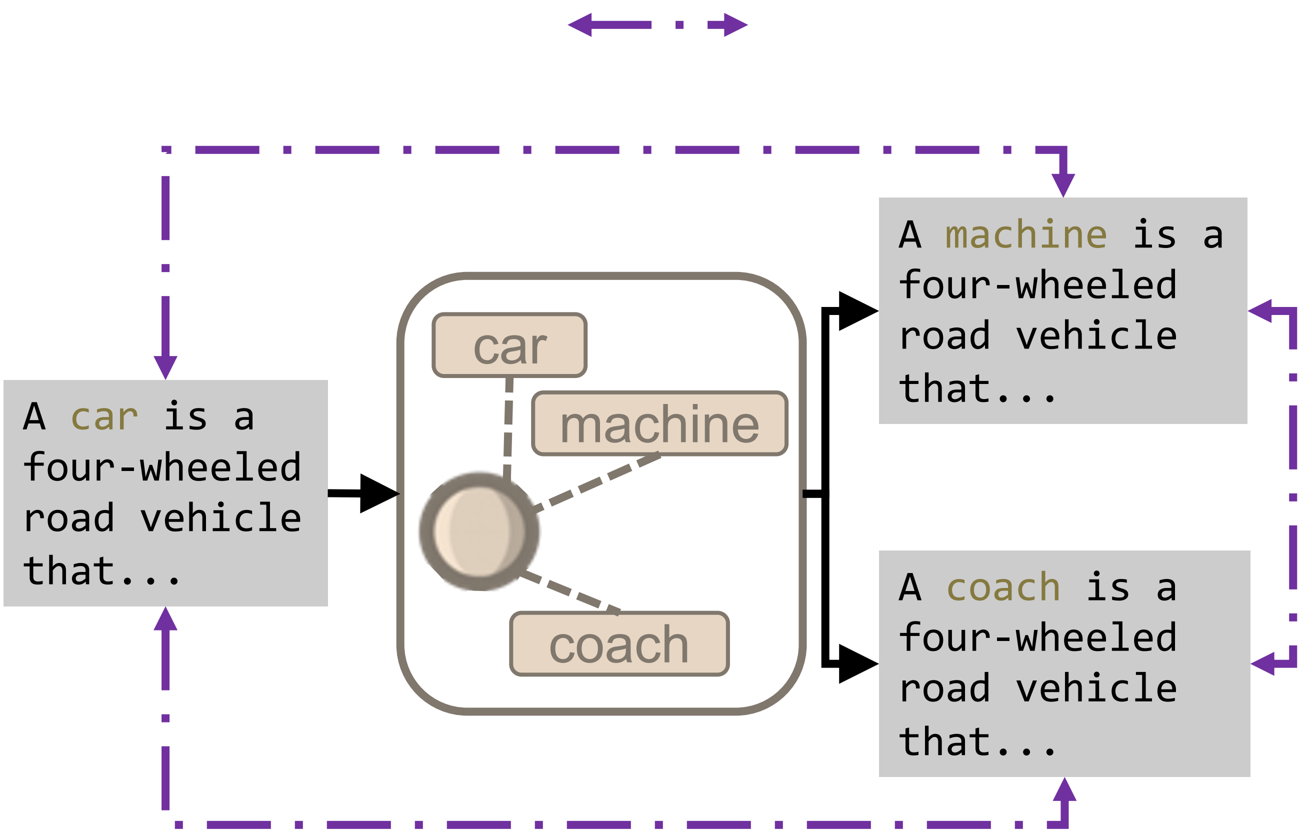
STEP 2: Generating **synthetic definitions of concepts**



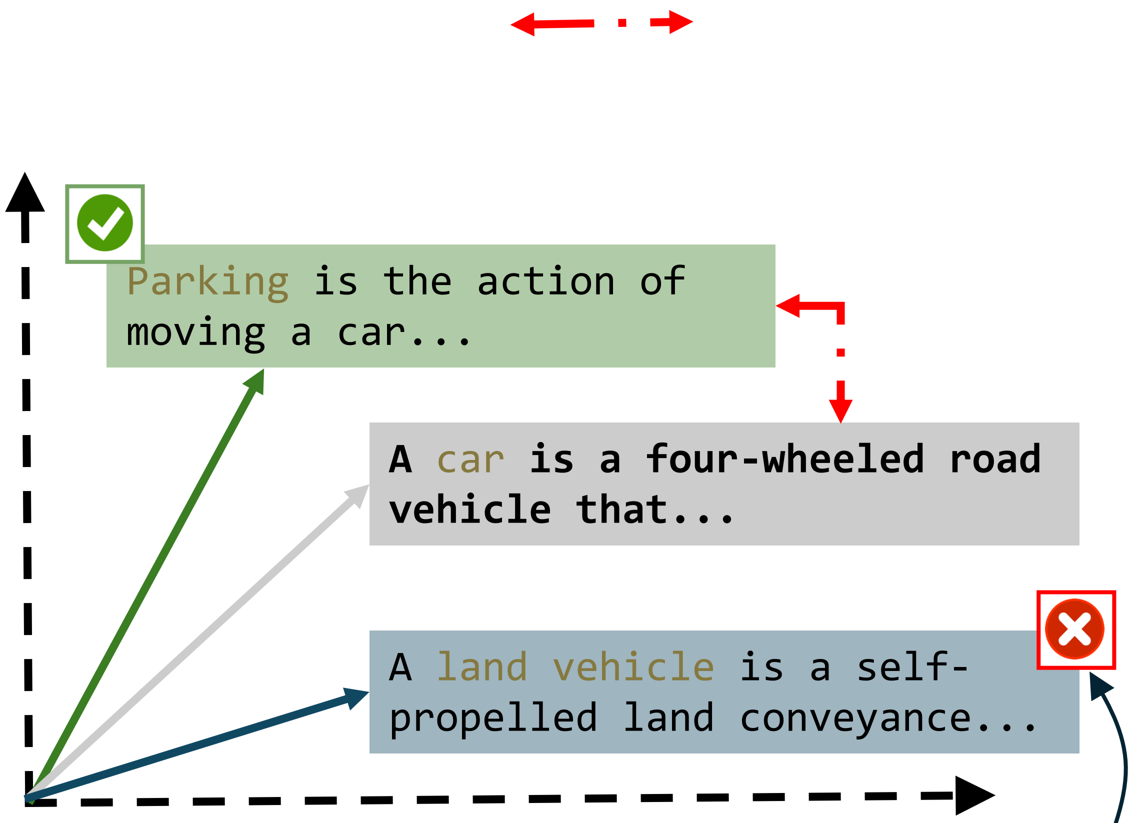
The ontological knowledge infusion approach

STEP 3: Selecting **positive** and **negative** pairs of concept definitions, driven by the ontology

POSITIVE PAIR of semantically related texts



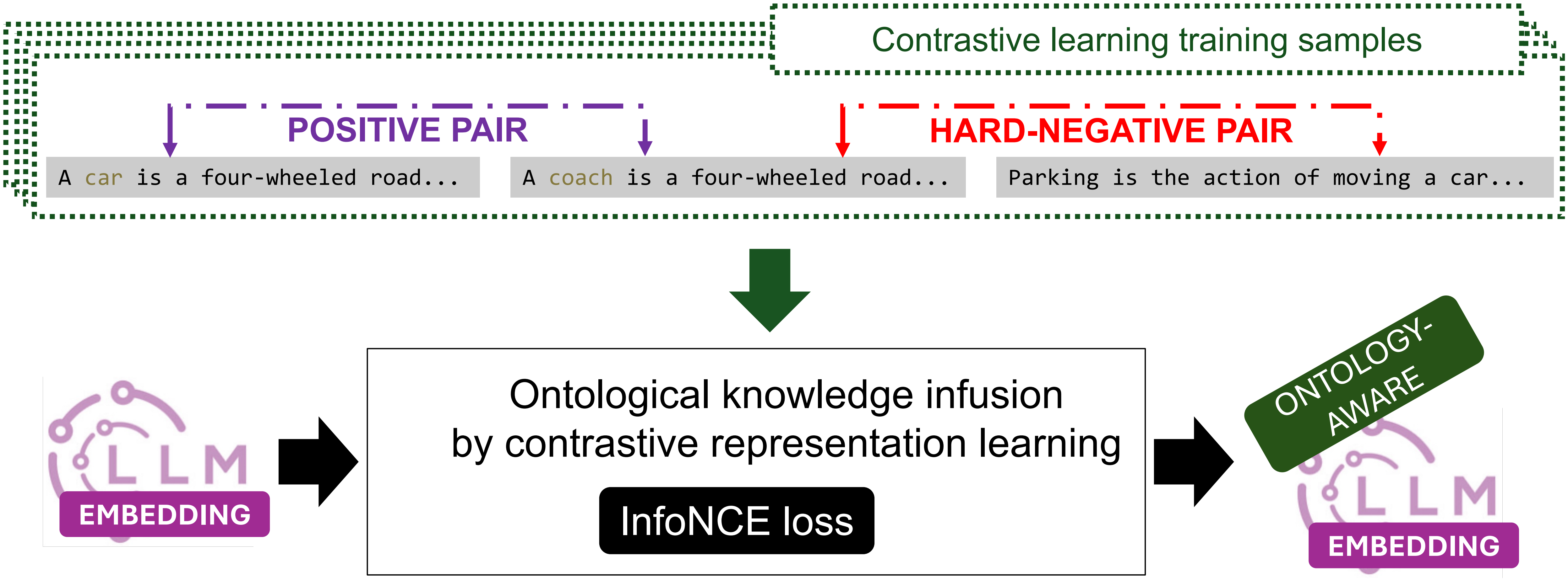
HARD-NEGATIVE PAIR of texts



Do not pair definitions of ancestor or descendant concepts

The ontological knowledge infusion approach

STEP 4: Fine-tuning the embedding-Large Language Model by **contrastive representation learning**



Implementation available on GitHub at https://github.com/iqvianlp/llm-onto-infuse/

Do embedding-Large Language Models better understand diseases, after infusing the MONDO disease ontology?

mondo
Disease Ontology

April-2024 version

- → 24,201 disease-related concepts
- → ~75k exact synonyms
- → definitions of ~70% of concepts
- → 36,459 is-a relations

LLM EMBEDDING → Ontological knowledge infusion by contrastive representation learning → ONTOLOGY-AWARE LLM EMBEDDING

EVALUATION

Do embedding-Large Language Models better understand diseases, after infusing the MONDO disease ontology?

PubMedBERT: 110M parameters, pre-trained from scratch using *abstracts from PubMed and full-text articles from PubMedCentral* with masked language modelling and next sentence prediction

SapBERT: 110M parameters, fine-tuned in a contrastive learning framework to increase the similarity of *pairs of synonyms of biomedical concepts*, from the UMLS meta-thesaurus

GTEbase: 110M parameters, fine-tuned by means of a *two-stages contrastive learning framework*: pre-training text pairs by weak supervision, a subsequent training on higher-quality annotated datasets

GIST: 100M parameters, *one of the best performing small embedding-LLMs in the MTEB leader-board*, fine-tuned by a contrastive objective relying on an dynamic selection strategy to identify in-batch negative samples

EMBEDDING

publicly available on

Hugging Face

Specialized to the biomedical domain

Not specialized to a specific knowledge domain

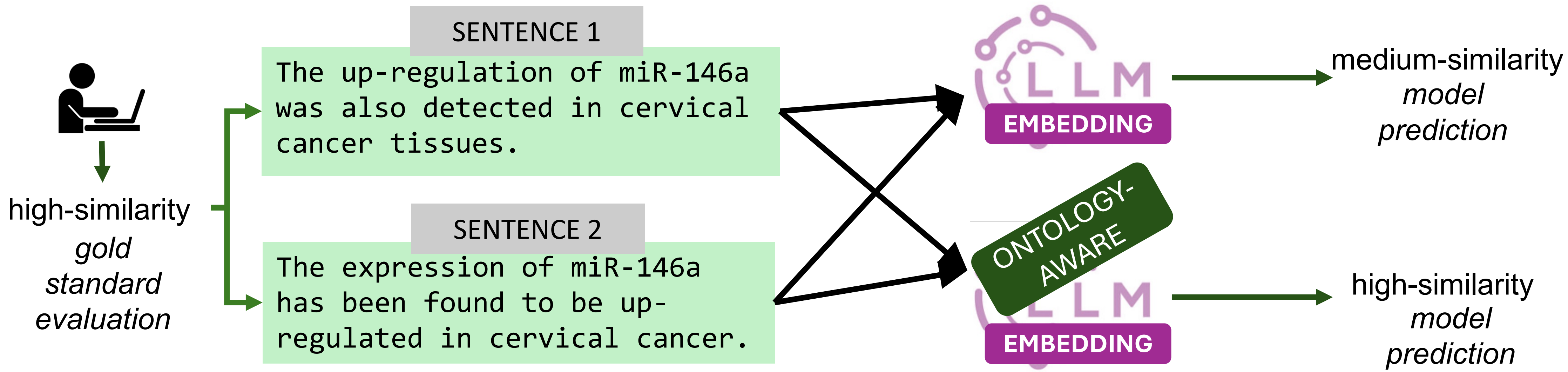
Pre-trained with standard objectives

Fine-tuned by contrastive objectives

EVALUATION

Do embedding-Large Language Models better understand diseases, after infusing the MONDO disease ontology?

TASK: Sentence Similarity



DATASETS: **BIOSSES**: 100 sentence pairs from biomedical publications

SEMEVAL: about 12k pairs of sentences dealing with several distinct knowledge domains

EVALUATION

Do embedding-Large Language Models better understand diseases, after infusing the MONDO disease ontology?

Embedding-Large Language Model		BIOSSES		STS 12		STS 13		STS 14		STS 15		STS 16	
		All	Dis	All	Dis	All	Dis	All	Dis	All	Dis	All	Dis
PubMedBERT	ORIGINAL	53.74	69.80	25.99	46.34	28.09	16.21	25.80	00.30	37.33	21.31	47.99	**80.33**
	ONTOLOGY-AWARE	**71.23**	**77.41**	**41.90**	**47.83**	**42.19**	**18.30**	**37.94**	**12.32**	**49.17**	**23.55**	**58.37**	72.78
SapBERT	ORIGINAL	81.86	83.21	70.89	68.84	79.23	35.73	70.37	47.64	77.85	56.99	76.71	89.73
	ONTOLOGY-AWARE	**85.45**	**84.79**	**72.31**	**79.99**	**80.66**	**46.04**	**72.44**	**52.07**	**79.79**	**64.05**	**77.58**	**92.86**
GTEbase	ORIGINAL	87.26	**90.30**	75.70	69.85	85.72	87.91	81.51	76.66	88.81	87.40	83.82	93.60
	ONTOLOGY-AWARE	**87.40**	89.62	**76.44**	**70.17**	**86.12**	**88.15**	**81.94**	**77.69**	**88.86**	**88.18**	**84.21**	**94.71**
GIST	ORIGINAL	87.96	89.66	76.15	63.88	87.85	88.64	83.39	74.52	89.43	**85.75**	85.35	**93.78**
	ONTOLOGY-AWARE	**88.86**	**92.05**	**76.69**	**65.94**	**87.99**	**89.26**	**83.64**	**75.45**	**89.56**	85.42	**85.69**	**93.78**

BIOMEDICAL SENTENCE SIMILARITY (in-domain)

NON-BIOMEDICAL SENTENCE SIMILARITY (out-of-domain)

All: all pairs of sentences considered

Dis: pairs of sentences mentioning diseases

IQVIA

To be presented at Data-centric Machine Learning Research Workshop

ICML International Conference On Machine Learning

arXiv > cs > arXiv:2405.20527

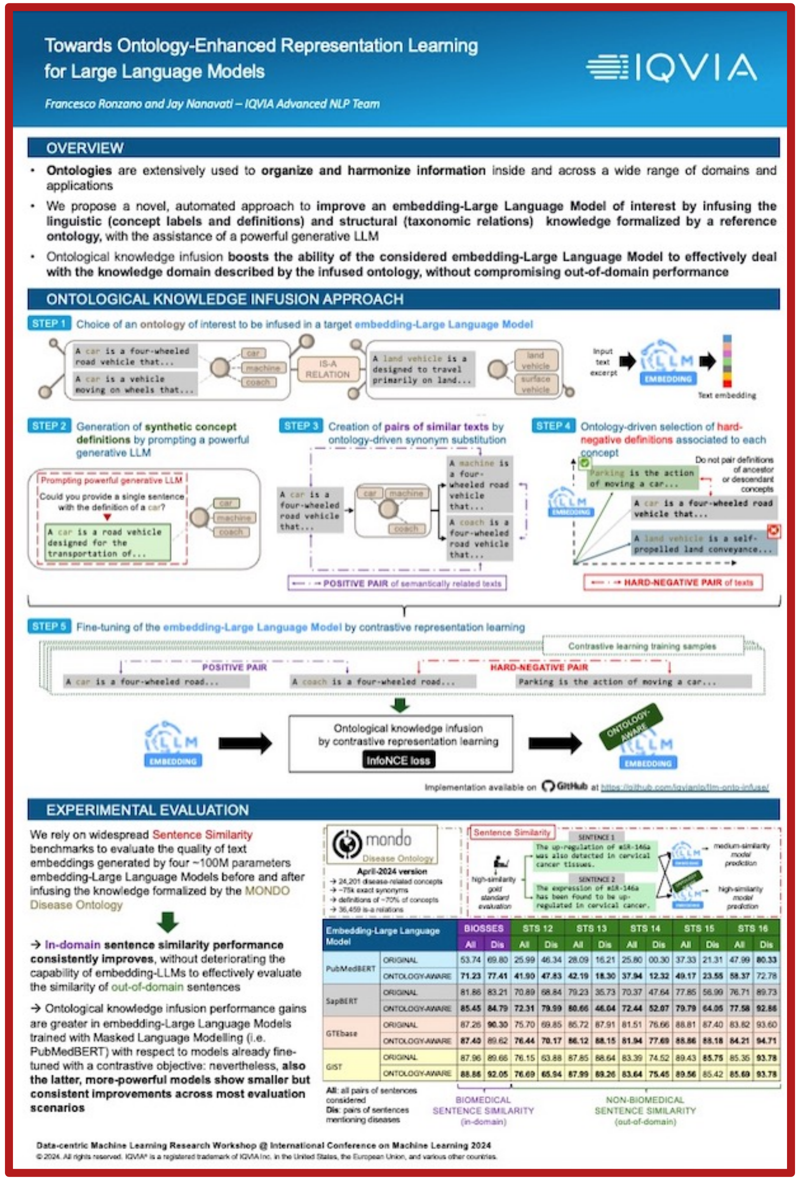
Computer Science > Computation and Language

[Submitted on 30 May 2024]

Towards Ontology-Enhanced Representation Learning for Large Language Models

Taking advantage of the widespread use of ontologies to organise and harmonize knowledge across several distinct domains, this paper proposes a novel approach to improve an embedding-Large Language Model (embedding-LLM) of interest by infusing the knowledge formalized by a reference ontology: ontological knowledge infusion aims at boosting the ability of the considered LLM to effectively model the knowledge domain described by the infused ontology. The linguistic information (i.e. concept synonyms and descriptions) and structural information (i.e. is-a relations) formalized by the ontology are utilized to compile a comprehensive set of concept definitions, with the assistance of a powerful generative LLM (i.e. GPT-3.5-turbo). These concept definitions are then employed to fine-tune the target embedding-LLM using a contrastive learning framework. To demonstrate and evaluate the proposed approach, we utilize the biomedical disease ontology MONDO. The results show that embedding-LLMs enhanced by ontological disease knowledge exhibit an improved capability to effectively evaluate the similarity of in-domain sentences from biomedical documents mentioning diseases, without compromising their out-of-domain performance.

Available on ArXiv at: https://arxiv.org/abs/2405.20527



Key learnings

- **Ontologies** are extensively used to **organize and harmonize information** across distinct domains and applications
- Knowledge resources like **ontologies** can be effectively exploited to both **create** and **effectively exploit high-quality textual data** useful to train Large Language Models, in data-hungry scenarios
- **Ontologies** can be effectively used to **prompt (powerful) generative Large Language Models** to drive the focused creation of **additional textual data** to support Large Language Models training
- **Contrastive learning** constitutes an effective technique to enrich the **latent knowledge** embedded inside a *Large Language Model* by relying on the **explicitly knowledge** formalized by an *ontology*

Tools exploited

SBERT.net

Sentence Transformers

- Python library useful to access, use, and train text and image embedding models
- Provide customizable classis useful to train (batching, loss function, etc.) and evaluate sentence embeddings

Hugging Face

- Used as a repository of embedding Large Language Models

OpenAI

OpenAI GPT-3.5-Turbo

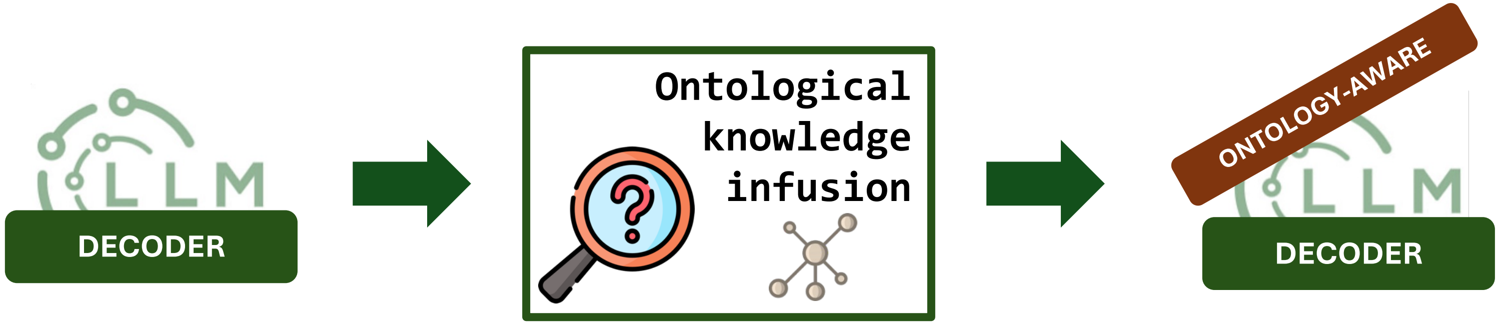
- Prompted to generate synthetic definitions of concepts from ontologies

IQVIA

Next steps

- Evaluate **bigger embedding-Large Language Models**, eventually derived from generative models (e.g. by LLM2Vec)
- Consider **distinct / multiple ontologies** to quantify the effectiveness of ontological knowledge infusion under distinct scenarios
- Explore **alternative strategies for ontology-driven training data generation**
- **Extend evaluation to additional tasks**, besides sentence similarity

...and **extend the proposed ontological knowledge infusion approach to generative Large Language Models**



Thanks for your attention
Any questions?

Towards Ontology-Enhanced Representation Learning for Large Language Models

Francesco Ronzano and Jay Nanavati
IQVIA Advanced NLP Team