Identification of Logical Fallacies in Natural Language Arguments

FILIP ILIEVSKI <http://ilievski.info>

VU Amsterdam / University of Southern California


Zhivar Sourati


Vishnu Priya


Darshan Deshpande


Himanshu Rawlani


Hông-Ân Sandlin


Alain Mermoud

What are logical fallacies?

Logical fallacy: a mistake in the reasoning from one proposition to the next, causing a faulty argument [AlMossawi, 2014]

A **broad category of violations** of argumentation norms, including structure, consistency, clarity, order, relevance, & completeness

Can be **formal** (structure) or **informal** (content)

Sourati, Z., Venkatesh, V.P.P., Deshpande, D., Rawlani, H., Ilievski, F., Sandlin, H.Â. and Mermoud, A., 2023. Robust and explainable identification of logical fallacies in natural language arguments. Knowledge-Based Systems.

What are informal fallacies?

A flaw in the substance (content / context) of an argument

Example: ad hominem (attacking the opponent's character or personal traits)

> ***I don't care what your arguments are; you are using Mickey Mouse tactics. The arguments you give are simply tacky.***

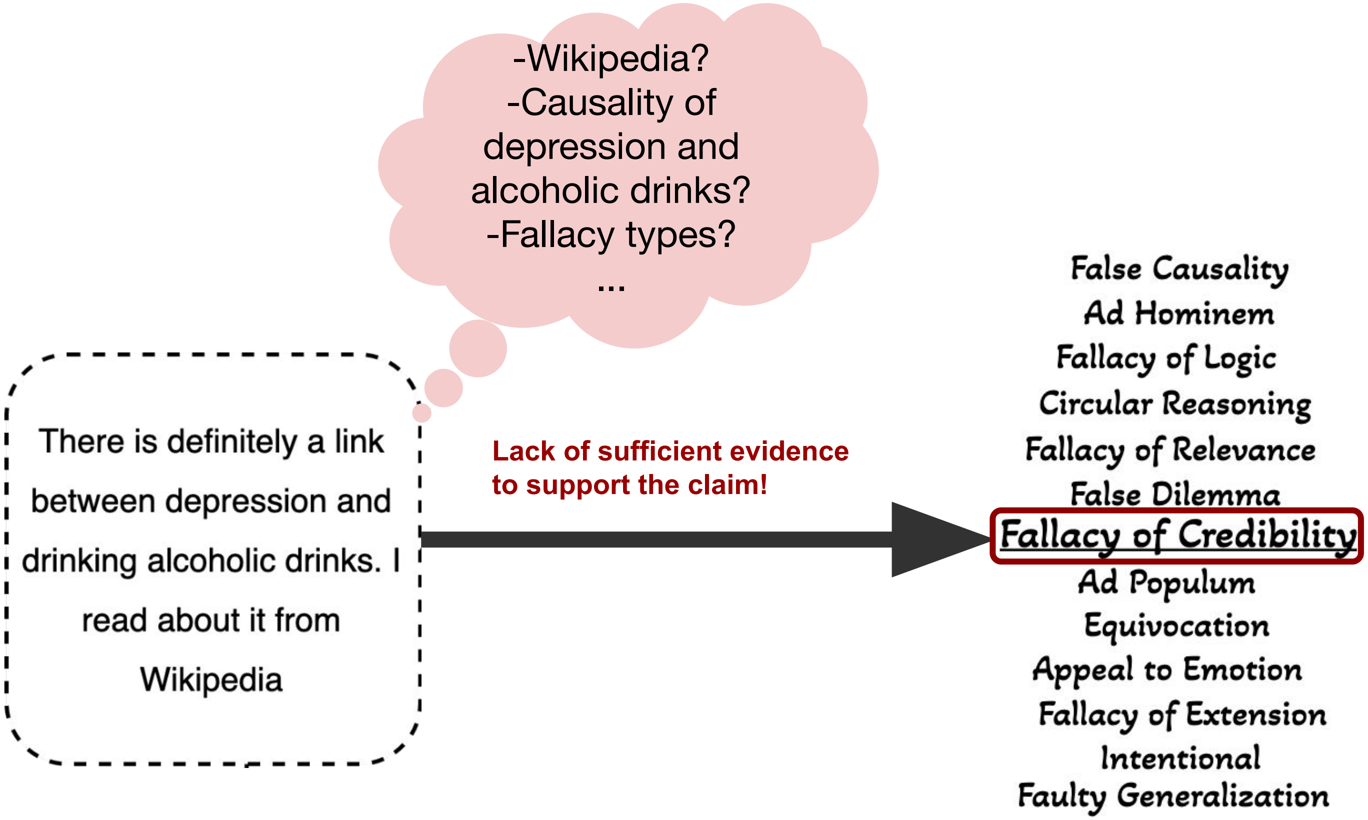
Appear to be misleadingly correct, hence seductive and persuasive

Additional material: Hansen, Hans, "Fallacies", *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/spr2023/entries/fallacies/

There is definitely a link between depression and drinking alcoholic drinks. I read about it from Wikipedia

-Wikipedia?
-Causality of depression and alcoholic drinks?
-Fallacy types?
...

There is definitely a link between depression and drinking alcoholic drinks. I read about it from Wikipedia


-Wikipedia?
-Causality of depression and alcoholic drinks?
-Fallacy types?
...
There is definitely a link between depression and drinking alcoholic drinks. I read about it from Wikipedia
Lack of sufficient evidence to support the claim!
False Causality
Ad Hominem
Fallacy of Logic
Circular Reasoning
Fallacy of Relevance
False Dilemma
Fallacy of Credibility
Ad Populum
Equivocation
Appeal to Emotion
Fallacy of Extension
Intentional
Faulty Generalization

Why automatic fallacy identification?

Fallacies have long been discussed in philosophy, from Aristotle to Copi and Barker

Identifying informal fallacies is a subtle task, requires flexible abstraction of information - many violation types and classes!

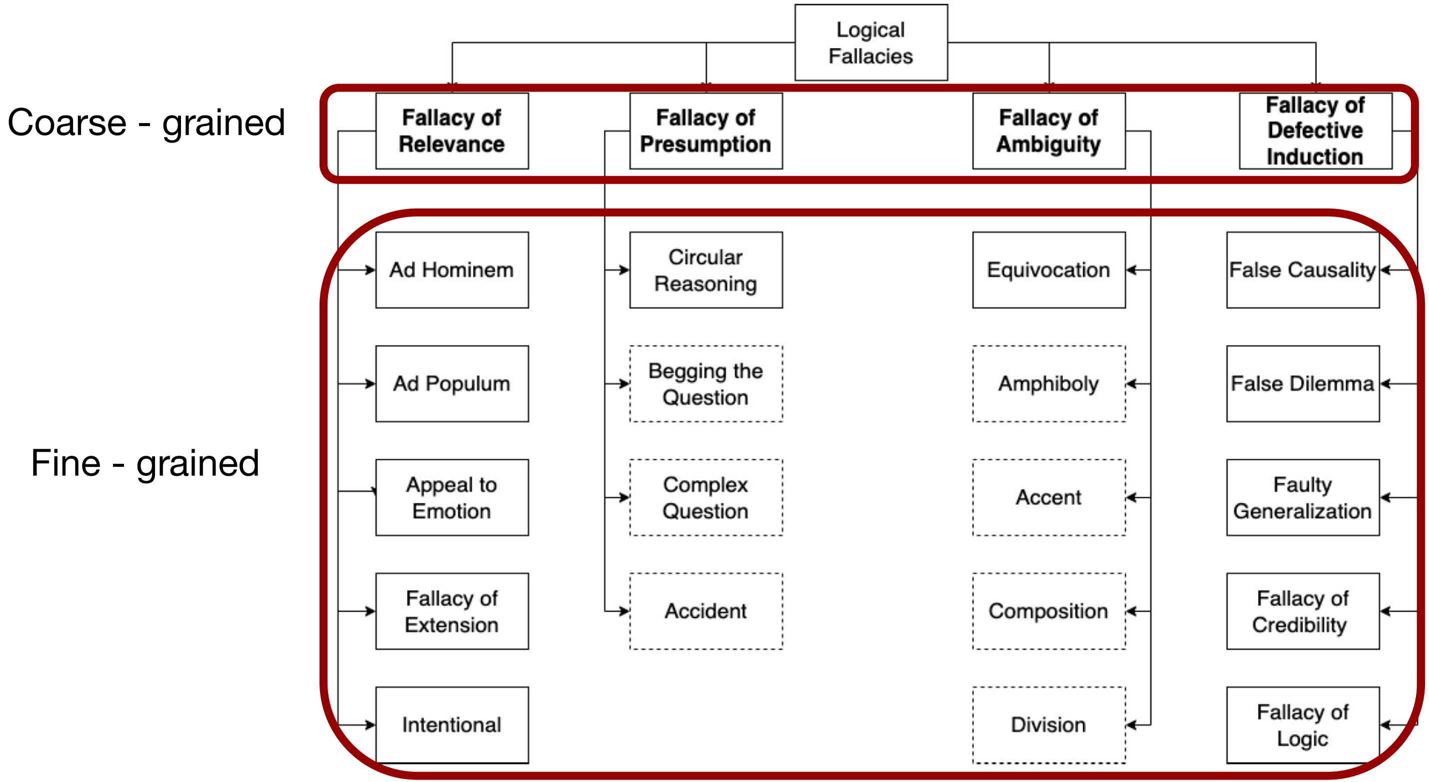
Almost no computational work on informal fallacy identification

- LMs struggle with this task [Jin et al., 2022]

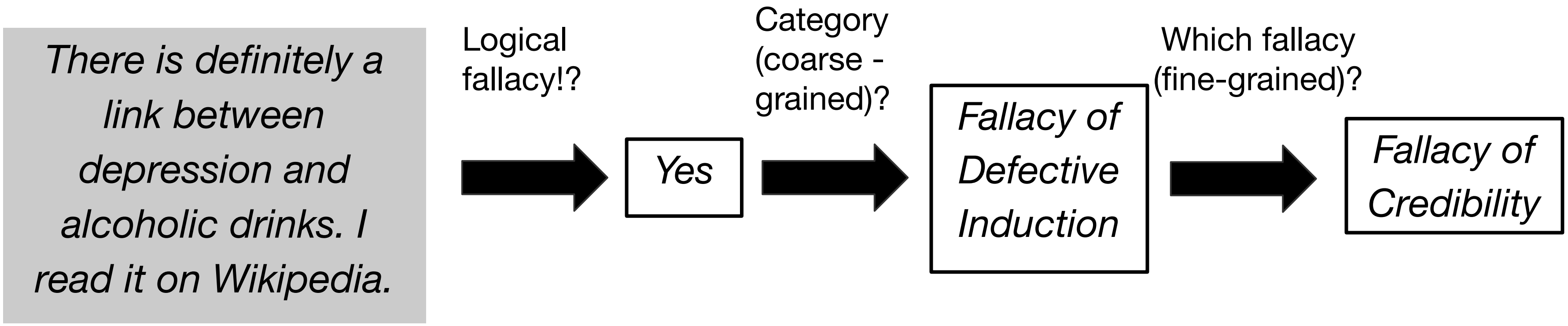
Research goal:

Can we build NeSy methods for robust and explainable identification of logical fallacies in natural language arguments?

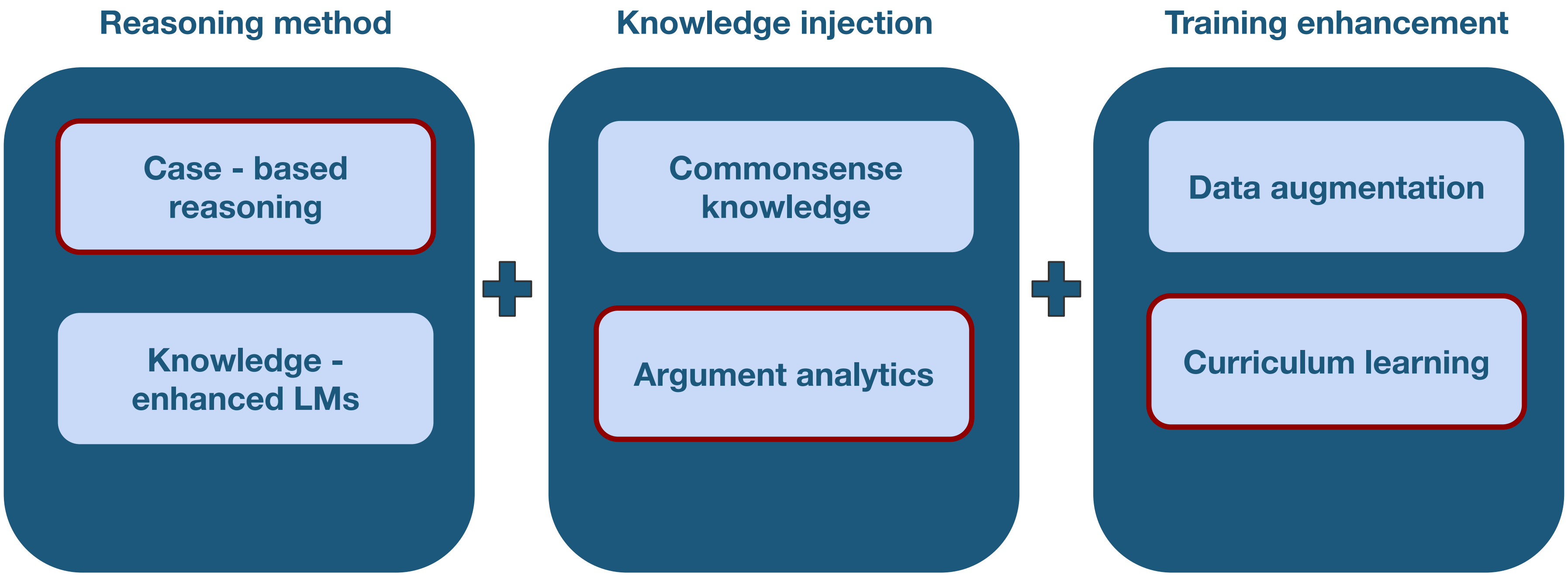
Our taxonomy of logical fallacies



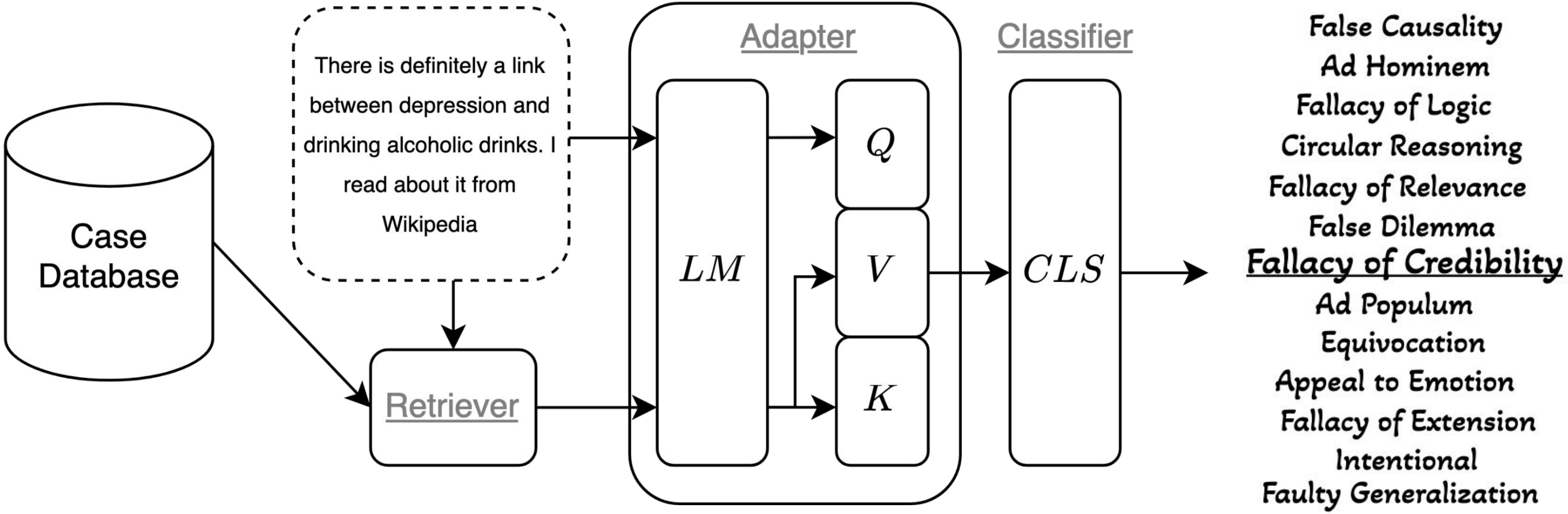
Task definition



NeSy framework for fallacy identification

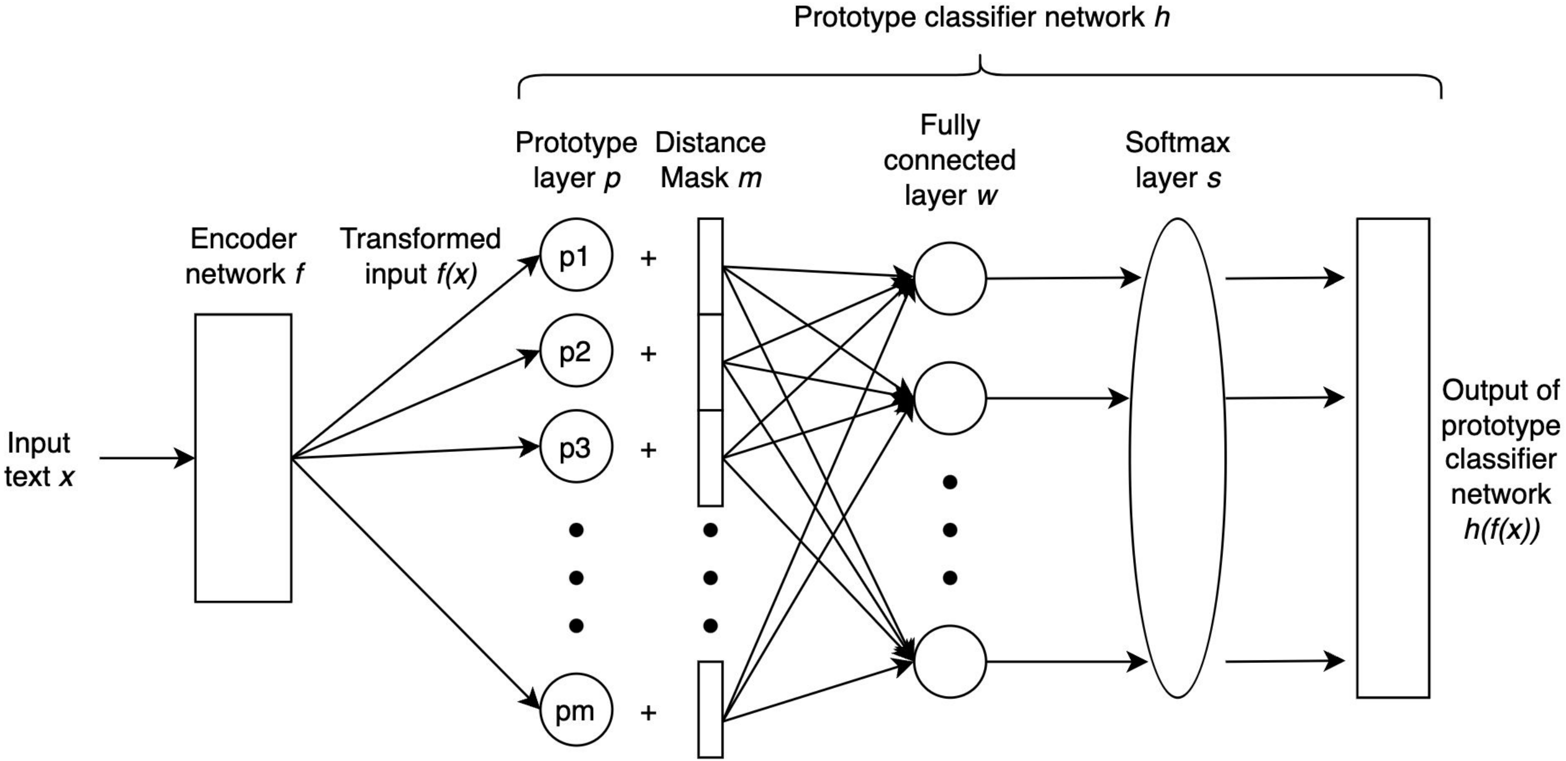


Instance-based reasoning with LMs

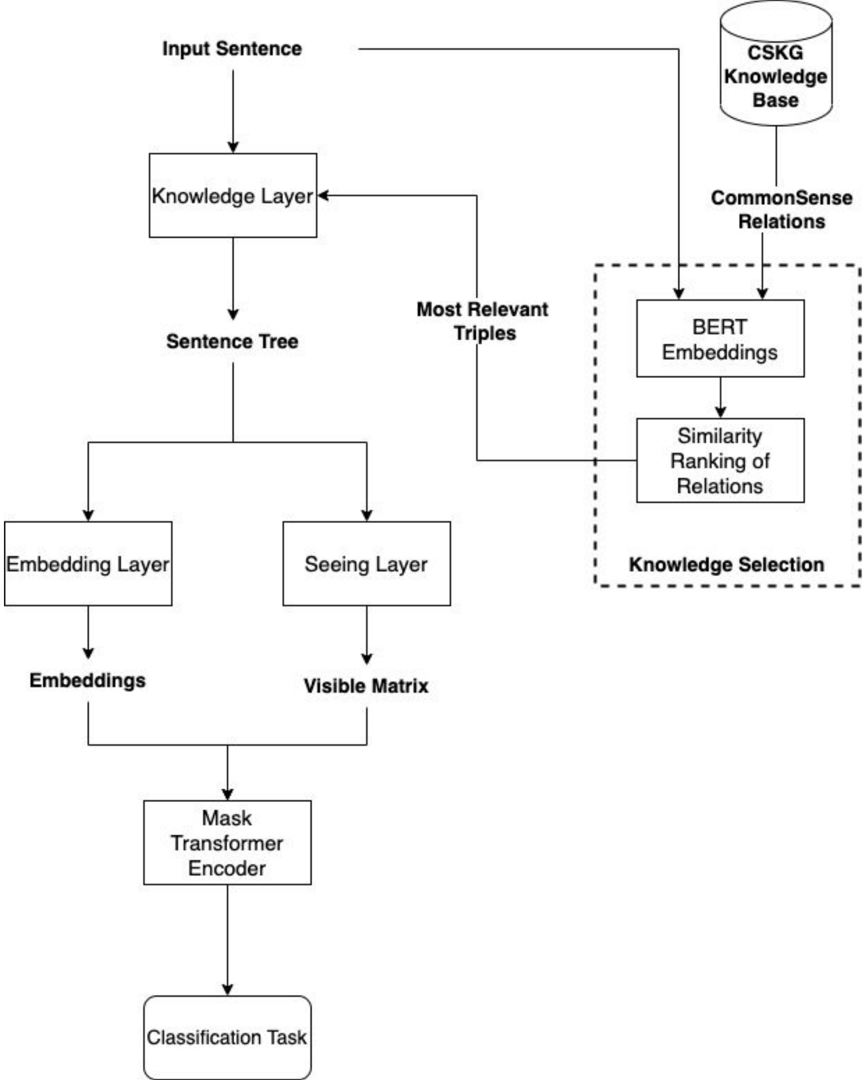


Sourati Z, Ilievski F, Sandlin HÂ, Mermoud A. Case-based reasoning with language models for classification of logical fallacies. 2023. IJCAI

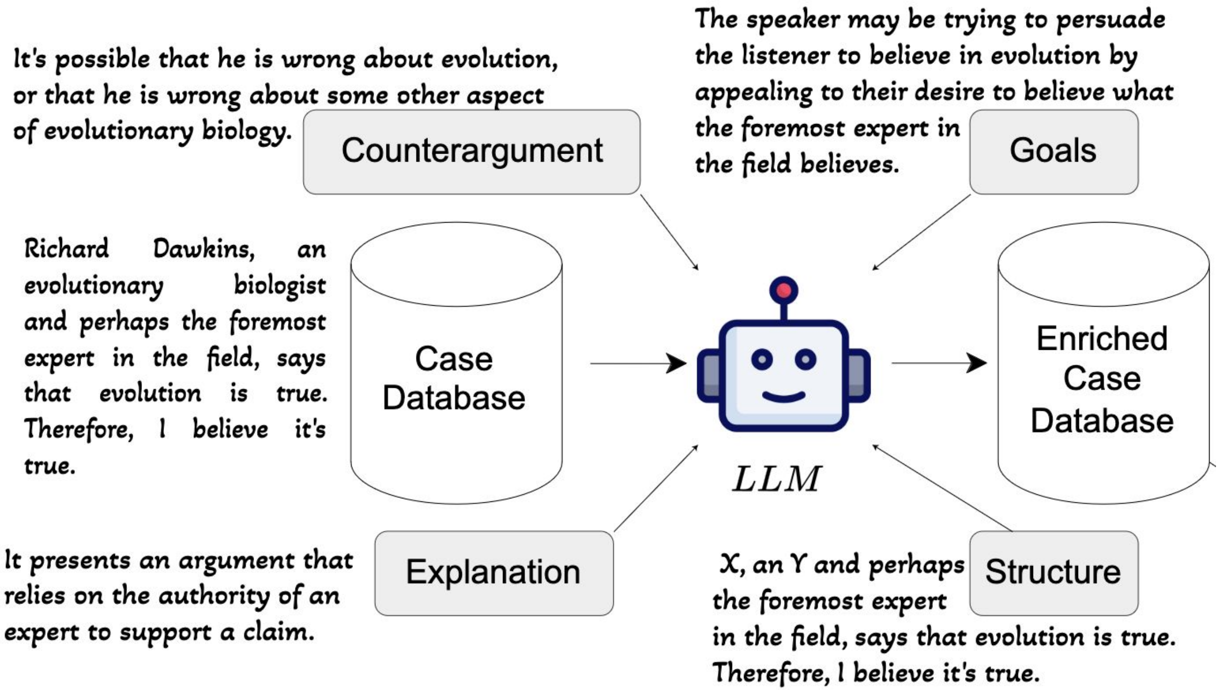
Prototype-based reasoning



Knowledge - enhanced LM reasoner (K-BERT with commonsense knowledge)

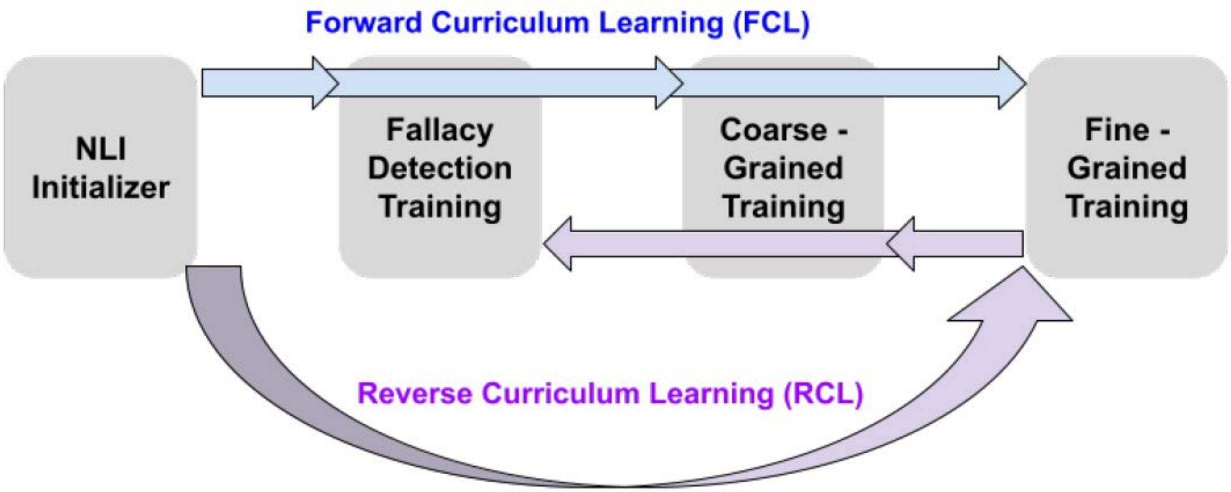


Knowledge injection: Argument analytics



Sourati Z, Ilievski F, Sandlin HÂ, Mermoud A. Case-based reasoning with language models for classification of logical fallacies. 2023. IJCAI

Training enhancement with curriculum learning



Instance-based reasoning outperforms the other methods

		LOGIC (in domain)				LOGIC Climate (out of domain)			
Type	**Model**	Acc	P	R	F1	Acc	P	R	F1
Random	/	0.076	0.094	0.076	0.079	0.077	0.124	0.077	0.085
Frequency	/	0.094	0.094	0.094	0.093	0.079	0.120	0.079	0.080
NLI	Electra	0.602	0.614	0.602	0.599 ± 0.02	0.229	0.276	0.229	0.217 ± 0.01
IBR	Electra	**0.631**	**0.638**	**0.631**	**0.627** ± 0.01	**0.254**	0.281	**0.254**	**0.245** ± 0.01
PBR	Electra	0.574	0.600	0.574	0.574 ± 0.01	0.199	**0.330**	0.199	0.166 ± 0.01
KI	BERT	0.488	0.478	0.488	0.482 ± 0.03	0.106	0.092	0.106	0.090 ± 0.02

Out-of-domain performance still much lower than in-domain

Argument analytics improves instance-based reasoning performance

Model	Representation	LOGIC P	LOGIC R	LOGIC F1	LOGIC Climate P	LOGIC Climate R	LOGIC Climate F1
ELECTRA	*Text*	0.655	0.634	0.635	0.317	0.242	0.242
	Counterarg.	**0.663**	**0.664**	**0.657**	0.355	**0.254**	**0.270**
	Goals	0.646	0.622	0.621	**0.376**	0.217	0.222
	Structure	0.634	0.625	0.618	0.375	0.254	0.269
	Explanations	0.605	0.580	0.578	0.314	0.242	0.237
RoBERTa	Text	**0.633**	0.613	0.619	0.343	0.236	0.251
	Counterarg.	0.624	0.613	0.615	0.367	0.198	0.216
	Goals	0.632	0.613	0.619	0.351	0.242	**0.263**
	Structure	0.631	**0.619**	**0.619**	**0.379**	**0.248**	0.245
	Explanations	0.575	0.558	0.559	0.359	0.192	0.181
BERT	*Text*	0.595	0.604	0.596	0.311	0.192	0.204
	Counterarg.	0.607	0.613	0.603	0.342	**0.217**	**0.228**
	Goals	0.598	0.607	0.596	0.310	0.204	0.203
	Structure	**0.613**	**0.616**	**0.611**	**0.359**	0.204	0.200
	Explanations	0.540	0.531	0.532	0.274	0.217	0.190

Sourati Z, Ilievski F, Sandlin HÂ, Mermoud A. Case-based reasoning with language models for classification of logical fallacies. 2023. IJCAI

Curriculum learning helps coarse- and fine-grained classification

Model	CL Type	Binary (BIG Bench) P	R	F1	Coarse-grained P	R	F1	Fine-grained P	R	F1
BERT	-	**0.848**	**0.845**	**0.845** ±0.01	0.714	0.718	0.717 ±0.04	0.583	0.583	0.583 ±0.01
	FCL	-	-	-	0.717	0.727	0.721 ±0.03	**0.613**	**0.586**	**0.584** ±0.02
	RCL	0.826	0.827	0.826 ±0.00	**0.783**	**0.779**	**0.778** ±0.02	-	-	-
DeBERTa	-	**0.988**	**0.988**	**0.988** ±0.00	0.746	0.740	0.741 ±0.03	0.607	0.593	0.592 ±0.02
	FCL	-	-	-	0.748	0.758	0.751 ±0.02	**0.632**	**0.604**	**0.608** ±0.01
	RCL	0.908	0.892	0.889 ±0.05	**0.779**	**0.785**	**0.780** ±0.02	-	-	-
DistilBERT	-	**0.848**	**0.847**	**0.847** ±0.01	0.684	0.695	0.683 ±0.02	0.508	0.513	0.505 ±0.02
	FCL	-	-	-	0.703	0.713	0.706 ±0.02	**0.550**	**0.520**	**0.525** ±0.03
	RCL	0.844	0.842	0.841 ±0.01	**0.704**	**0.719**	**0.711** ±0.03	-	-	-
RoBERTa	-	**0.983**	**0.983**	**0.983** ±0.01	0.719	0.714	0.716 ±0.01	0.560	0.545	0.545 ±0.02
	FCL	-	-	-	0.710	0.713	0.706 ±0.02	**0.578**	**0.569**	**0.565** ±0.02
	RCL	0.900	0.899	0.899 ±0.01	**0.736**	**0.741**	**0.732** ±0.01	-	-	-
Electra	-	**0.995**	**0.995**	**0.995** ±0.00	0.765	0.767	0.764 ±0.01	0.614	0.602	0.599 ±0.02
	FCL	-	-	-	0.711	0.722	0.716 ±0.03	**0.624**	**0.613**	**0.610** ±0.04
	RCL	0.957	0.957	0.957 ±0.01	**0.779**	**0.782**	**0.775** ±0.03	-	-	-

Explaining by example

Class	Input Sentence	Similar Cases (IBR)	Prototypical Cases (PBR)
Ad Populum	Everyone is going to get the new smart phone when it comes out this weekend. Why aren't you?	(1) **I'm gonna get an iPhone because everybody else has an iPhone and they're cool.** (2) **Everyone wants the iPhone 11 because it's the best phone on the market!**	(1) **Everyone seems to support the changes in the vacation policy, and if everyone likes them, they must be good.** (2) **Everyone is buying the new iPhone that's coming out this weekend. You have to buy it too.**
Faulty Generalization	Everyone knows that teenagers are lazy	(1) **If we let teenagers wear whatever they want to school, they will no longer respect the rules and academic performance will decline.** (2) If we don't teach teens to work harder, the human race is doomed	(1) **If we allow a housing development to be built on Sunny Lake, a resort will come next, and soon we won't have any wilderness left!** (2) **Michael is part of the Jackson Five. Without Tito and company, he will never make it.**

Bold *means same class as the ground truth*

Sourati, Z., Venkatesh, V.P.P., Deshpande, D., Rawlani, H., Ilievski, F., Sandlin, H.Â. and Mermoud, A., 2023. Robust and explainable identification of logical fallacies in natural language arguments. Knowledge-Based Systems.

Takeaways

Logical fallacy identification is an understudied AI challenge, while popular in social sciences

Instance-based reasoning, curriculum learning, and argument analytics improve the robustness of LMs

Further research on NeSy methods needed to build robust and explainable models

Thanks!

http://ilievski.info

f.ilievski@vu.nl

Towards comprehensive argument analytics