Reasoning with Natural Language Explanations: From Epistemology to Neuro-Symbolic AI

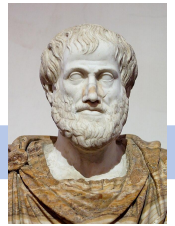
Marco Valentino
Neuro-Symbolic AI Group,
Idiap Research Institute.

https://www.marcovalentino.net
marco.valentino@idiap.ch

Explanation

A core feature of human intelligence and rationality:

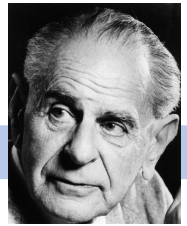
- Support learning and reasoning
- Closely linked to understanding and to the ability to perform abstractions
- Generalisation and adaptation to novel and unexpected situations
- Effective communication across different fields
- Scientific progress and the growth of knowledge
- Creativity and imagination



Lombrozo, Tania. The structure and function of explanations. 2006.

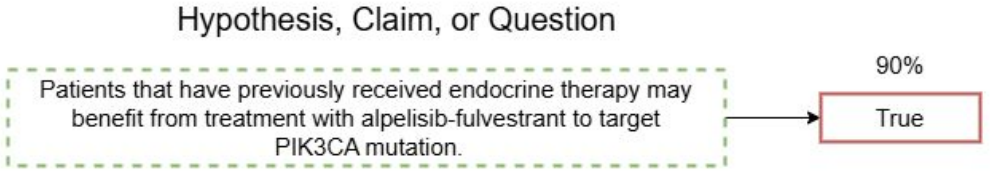
Salmon, Wesley C. *Four decades of scientific explanation*. 2006.

Popper, Karl. *Conjectures and refutations: The growth of scientific knowledge*. 1963.



> 2000 years

Explanation-Based Natural Language Inference (NLI)



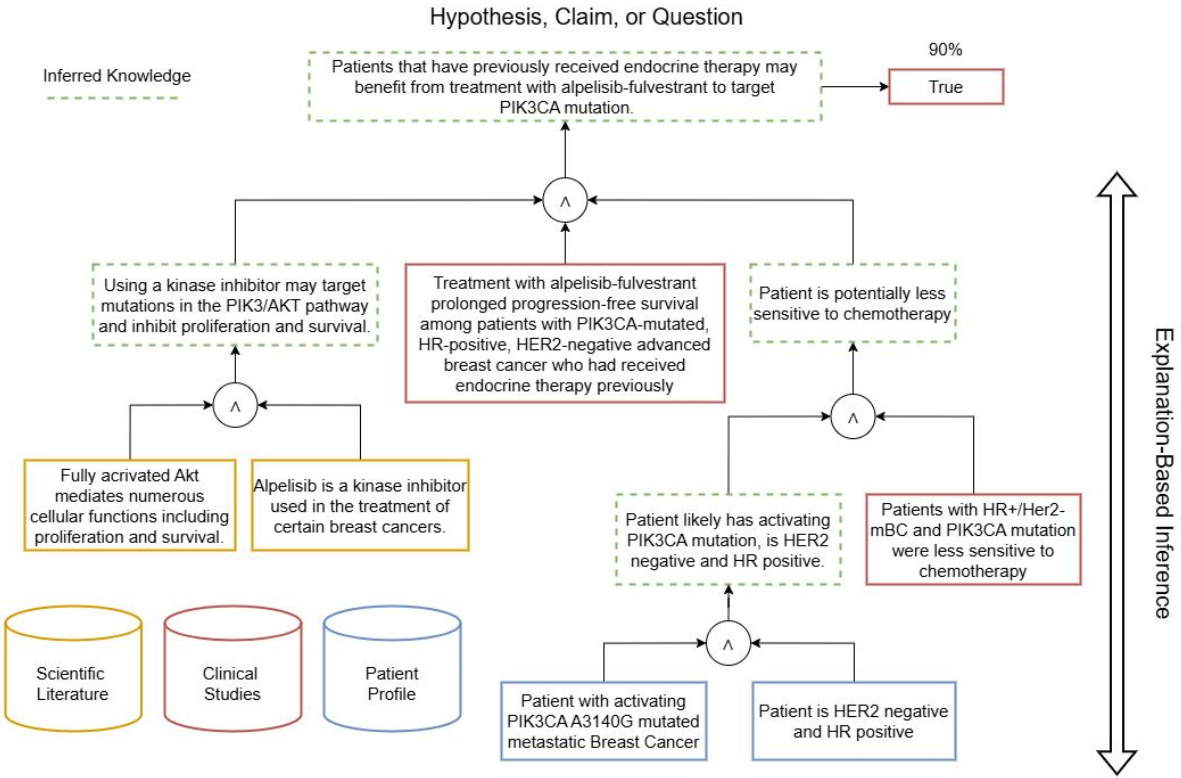
Fill a *knowledge* and *inference gap* on *how* and *why* the model arrived at the final answer.

Explanation-Based Natural Language Inference (NLI)

Long-term goals:

1. E*xplanation as a core mechanism* for learning and reasoning with natural language.
2. Enable *safe, transparent and effective human-AI collaboration* in high-stakes domains.

Explanation as a proxy for delivering and evaluating *complex reasoning behaviour.*



Today

What is an explanation?

>> Epistemological and linguistic perspective

How can we build models that can reason with natural language explanations?

>> From LLMs to Neuro-Symbolic AI

Marco Valentino and André Freitas. Reasoning with Natural Language Explanations. EMNLP 2024 (Tutorials)

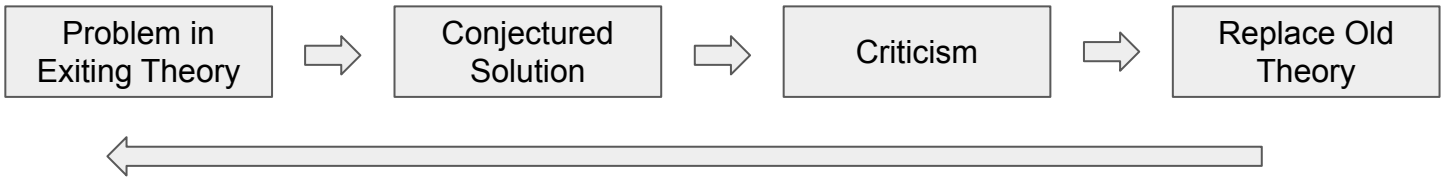
https://sites.google.com/view/reasoning-with-explanations

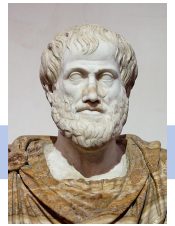
Part 1: What is an Explanation?

The Process of Explanation

An explanation is an argument composed of a set of statements (*explanans*) that describe why or how a particular phenomenon occurs (*explanandum*).

Explanation as problem solving:



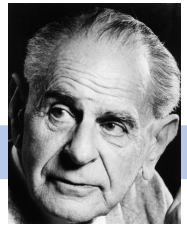


Lombrozo, Tania. The structure and function of explanations. 2006.

Salmon, Wesley C. *Four decades of scientific explanation*. 2006.

Popper, Karl. *Conjectures and refutations: The growth of scientific knowledge*. 1963.

> 2000 years

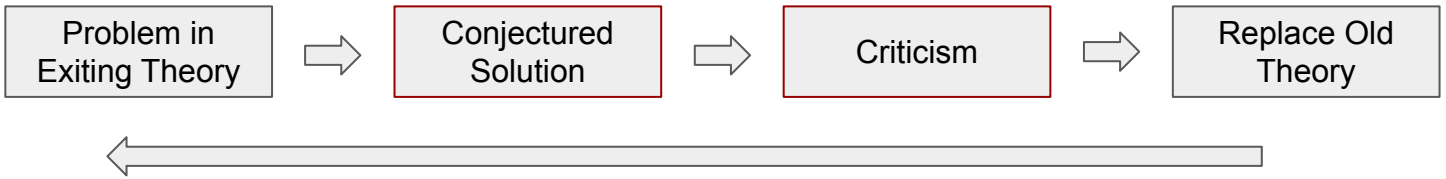


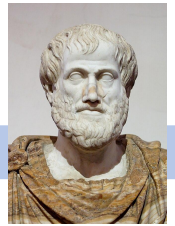
The Process of Explanation

An explanation is an argument composed of a set of statements (*explanans*) that describe why or how a particular phenomenon occurs (*explanandum*).

Explanation as problem solving:

Observable phenomena are explained in terms of (conjectured) unobservable phenomena.

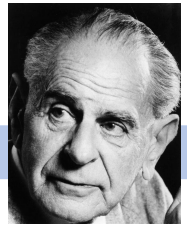




Lombrozo, Tania. The structure and function of explanations. 2006.

Salmon, Wesley C. *Four decades of scientific explanation*. 2006.

Popper, Karl. *Conjectures and refutations: The growth of scientific knowledge*. 1963.



> 2000 years

In Science

Problem. The geocentric model struggled to explain the apparent retrograde motion of planets like Mars, where they seem to move (temporarily) backwards in the sky.

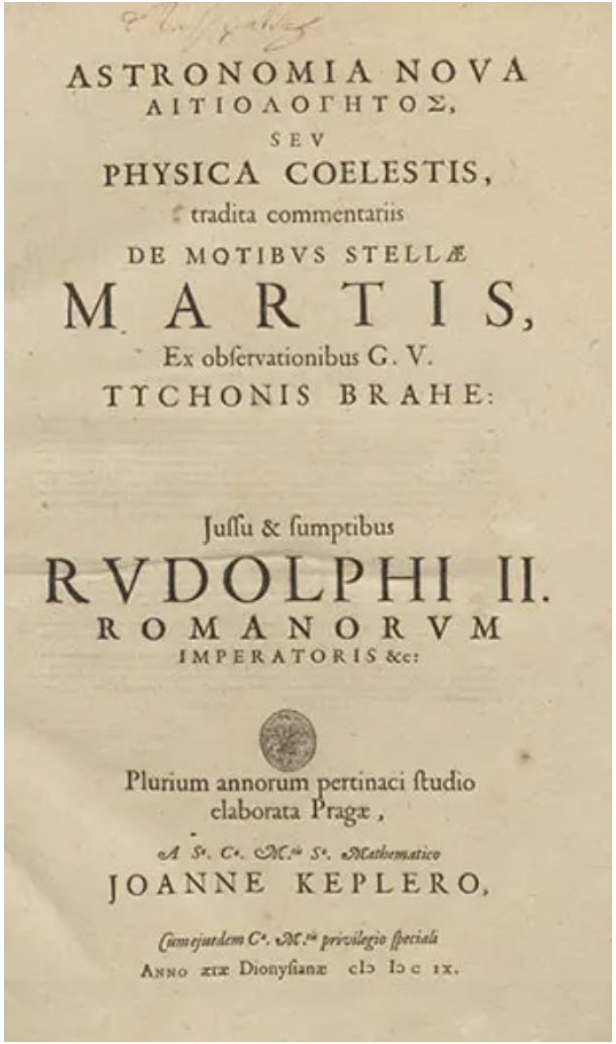
Copernicus's Explanation. In the heliocentric model, retrograde motion results from the relative positions and motions of Earth and the other planets. As Earth overtakes a slower-moving outer planet in its orbit, the planet appears to move backwards.


NICOLAI
COPERNICI TO-
RINENSIS DE REVOLVTIONI-
bus orbium cœleſtium,
Libri VI.
IN QVIBVS STELLARVM ET FI-
XARVM ET ERRATICARVM MOTVS, EX VETE-
ribus atq; recentibus obſeruationibus, reſtituit hic autor.
Præterea tabulas expeditas luculentasq; addidit, ex quibus eoſdem motus ad quoduis tempus Mathe-
matum ſtudioſus facillime calcu-
lare poterit.
ITEM, DE LIBRIS REVOLVTIONVM NICOLAI
Copernici Narratio prima per M. Georgium Ioachi-
mum Rheticum ad D. Ioan. Schone-
rum ſcripta.
Cum Gratia & Priuilegio Cæſ. Maieſt.
BASILEAE, EX OFFICINA
HENRICPETRINA.

In Science

Problem. Inability of the Copernican model to accurately predict planetary positions, particularly the orbit of Mars, based on the observational data collected by Tycho Brahe.

Kepler's Explanation. Define orbits as ellipses with the Sun at one focus, quantitatively explaining variations in planetary velocities and positions by establishing that planets sweep out equal areas in equal times (Second Law) and that the square of a planet's orbital period is proportional to the cube of its semi-major axis distance from the Sun (Third Law).


ASTRONOMIA NOVA
AITIOΛOΓHTOΣ,
SEV
PHYSICA COELESTIS,
tradita commentariis
DE MOTIBVS STELLÆ
MARTIS,
Ex obſervationibus G. V.
TYCHONIS BRAHE:

Juſſu & ſumptibus
RVDOLPHI II.
ROMANORVM
IMPERATORIS &c:

Plurium annorum pertinaci ſtudio
elaborata Pragæ,
A S. C. M.ti S. Mathematico
JOANNE KEPLERO,

Cum ejusdem C. M.ti privilegio ſpeciali
ANNO XIX Dionyſianæ CIↃ IↃC IX.

Abduction & Inference to the Best Explanation

(C.S. Peirce)

Abductive Reasoning

problem/observations → Inference to the best explanation (may be true)

Deductive Reasoning

premises → derive conclusions from premises (true)

Inductive Reasoning

specific facts/observations → generalise rules from observations (may be true)

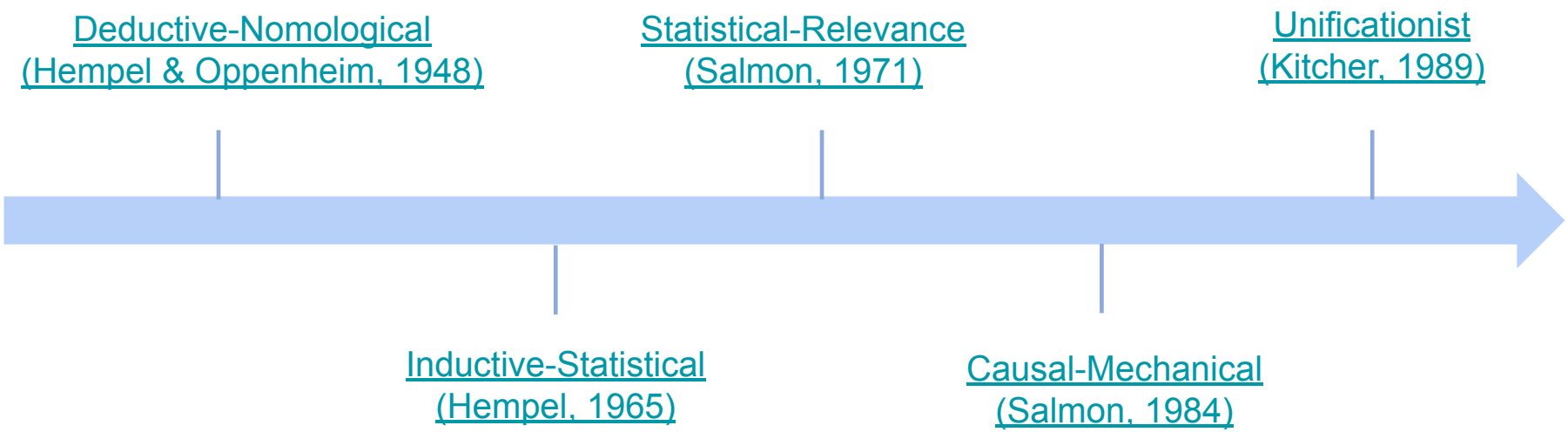
Criteria for Good Explanations

How do we criticise explanations?

- **Hardness to variation**. A good explanation cannot be arbitrarily modified without making it either inconsistent or unable to account for the phenomenon it explains (functional role between explanans and explanandum).
- **Falsifiability.** Demarcation between scientific and non-scientific explanations.
- **Simplicity**. Explanations are more persuasive when they're clear, logically structured, and avoid unnecessary complexity (related to *Occam's Razor*)
- **Coherence**. Explanations that fit with existing knowledge are more convincing.
- **Relevance**. An explanation should provide the right kind of information for the given context.
- **Predictive power.** A good explanation not only clarifies why something happened but can predict future occurrences.
- **Generality**. Explanations that apply to many situations are often favored.

David Deutsch: A new way to explain explanation

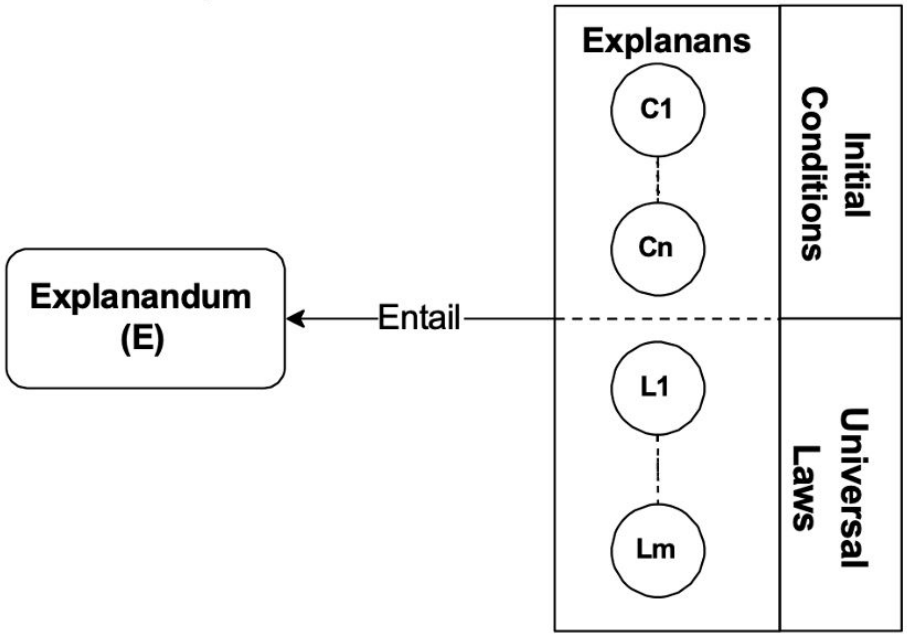
Explanatory Arguments



Valentino & Freitas, Technology & Philosophy (2024)

Deductive-Nomological (Hempel, 1948)

Structure: The explanation is a deductive argument.
Role: Showing that the explanandum has to be expected by virtue of the explanans



Deductive-Nomological (DN)

explanans

initial conditions

C_1: The (cool) sample of mercury was placed in hot water;
C_2: Mercury is a metal;
L_1: All metals expand when heated;
E: The sample of mercury expanded.

universal laws

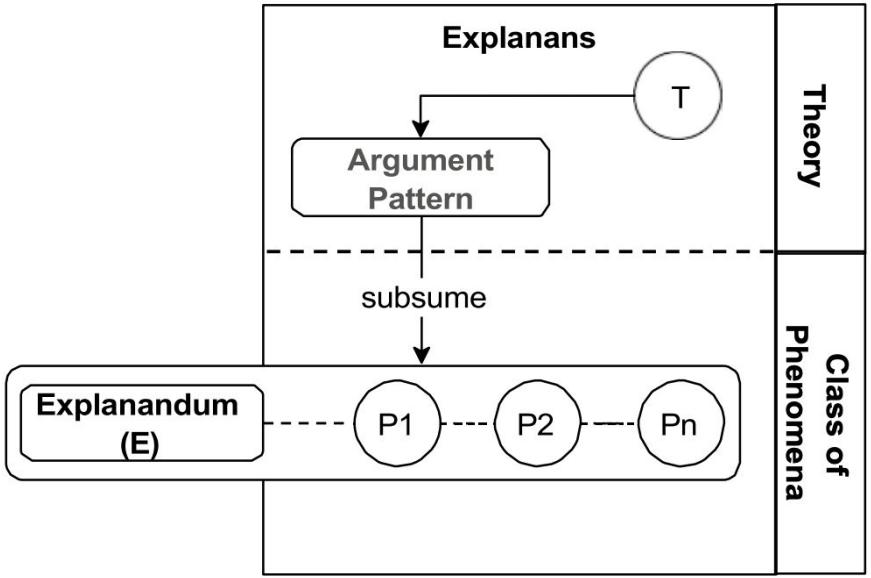
explanandum

Emphasise the relation between **explanation and prediction**.
Explanation as a **logically valid argument.**

(Hempel & Oppenheim, 1948)

Unificationist (Kitcher, 1989)

Structure: A theory, an argument pattern and a class of phenomena
Role: Showing how a theory T subsumes a class of phenomena including the explanandum



Explanatory Unification

To properly characterise an explanation we need to consider its main function of **fitting the explanandum into a broader unifying pattern**.

Specifically, an explanation is an argument whose role is to **connect a set of apparently unrelated phenomena**.

An **argument pattern** is a sequence of schematic sentences organised in premises and conclusions.

A schematic sentence can be described as a template obtained by replacing some non-logical expressions in a sentence with variables or dummy letters.

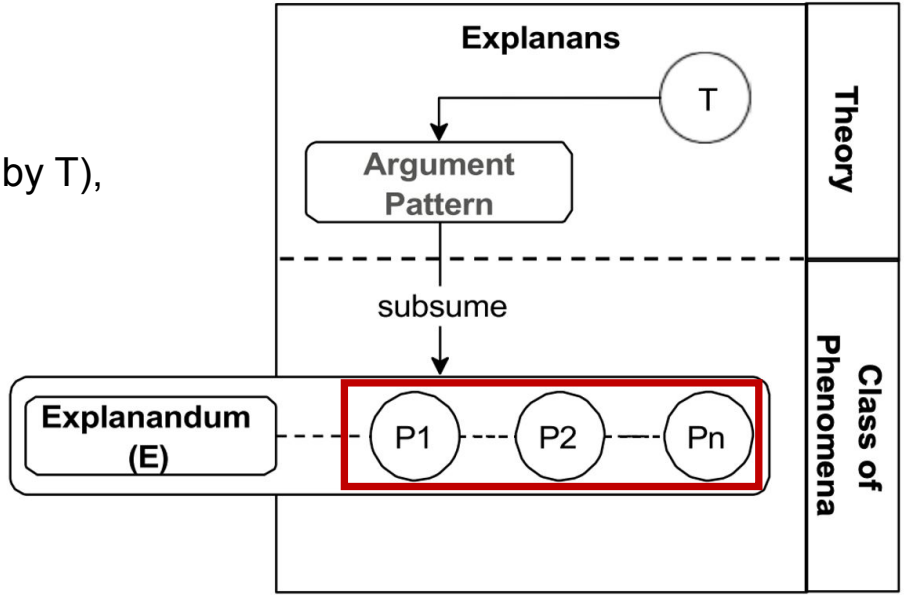
"to explain is to ft the phenomena into a unified picture insofar as we can. What emerges in the limit of this process is nothing less than the causal structure of the world" (Kitcher, 1989).

Explanatory Unification

Provides a set of criteria to identify the “best explanation” among competing theories:

Explanatory power:

The larger the cardinality of P
(i.e. the number of phenomena that are unified by T),
the greater the **explanatory power** of T



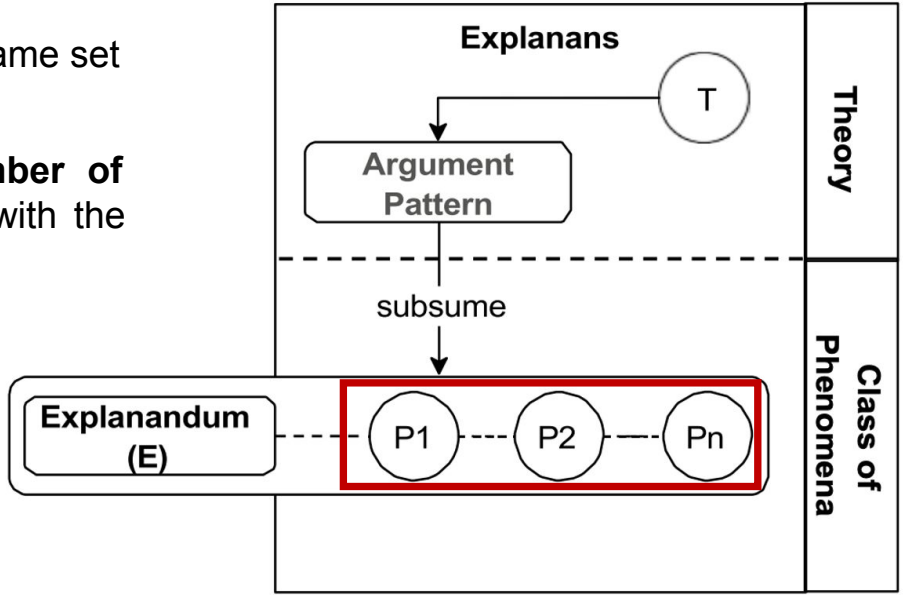
Explanatory Unification

Provides a set of criteria to identify the “best explanation” among competing theories:

Simplicity:

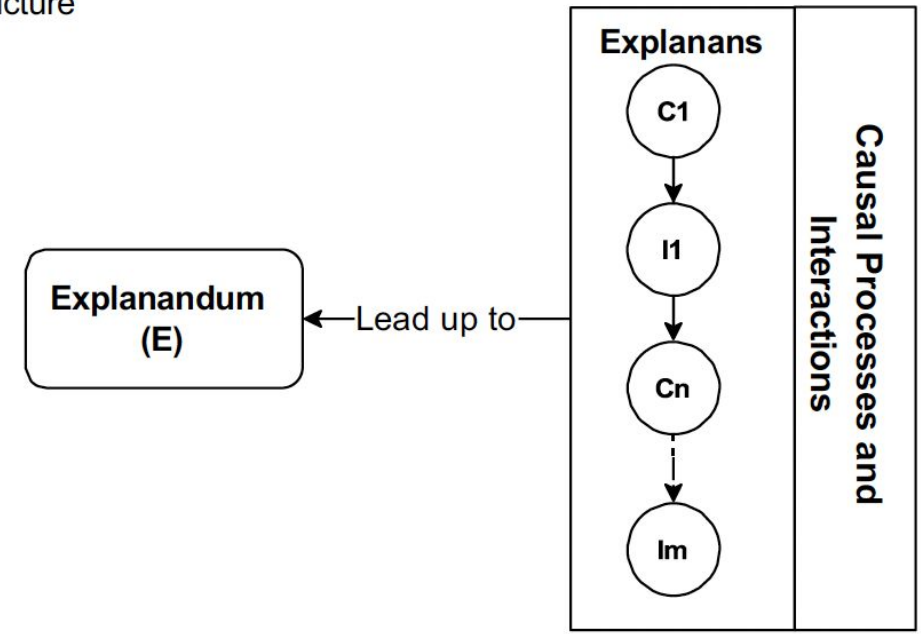
Given two theories T1 and T2 able to unify the same set of phenomena P.

The theory that makes use of a **lower number of premises** in its argument patterns is the one with the greatest **explanatory power.**



Causal-Mechanical (Salmon, 1984)

Structure: A set of relevant causal processes and interactions
Role: Showing that the explanandum is part of a broader causal structure



Causal-Mechanistic (CM)

Salmon identifies two major ways of constructing causal explanations for some event E.

An explanation can be either:

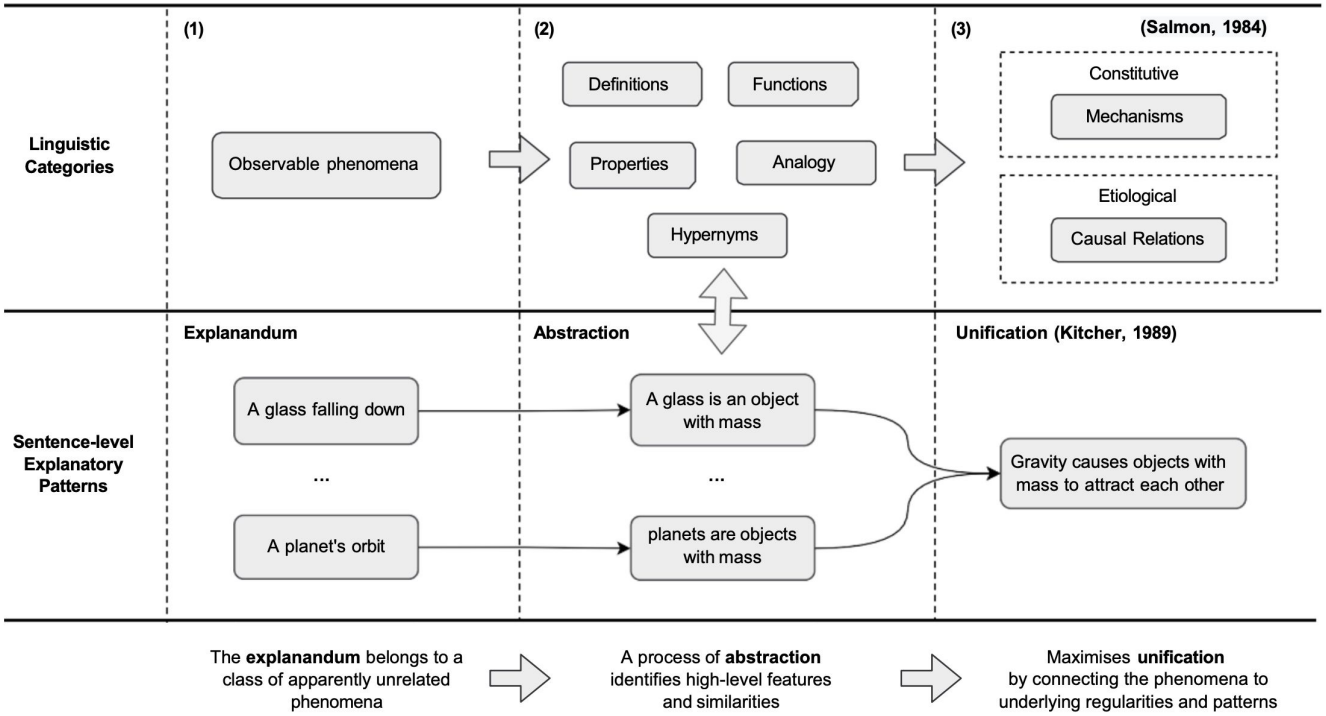
Etiological: E is explained by revealing part of its causes.

Constitutive: the explanation of E describes the underlying mechanisms giving rise to E.

Causes and mechanisms create **explanations that are hard to vary**.

Causes and mechanisms at a higher level of abstractions are **connected to unification** (the same underlying causal mechanisms can explain a large set of phenomena).

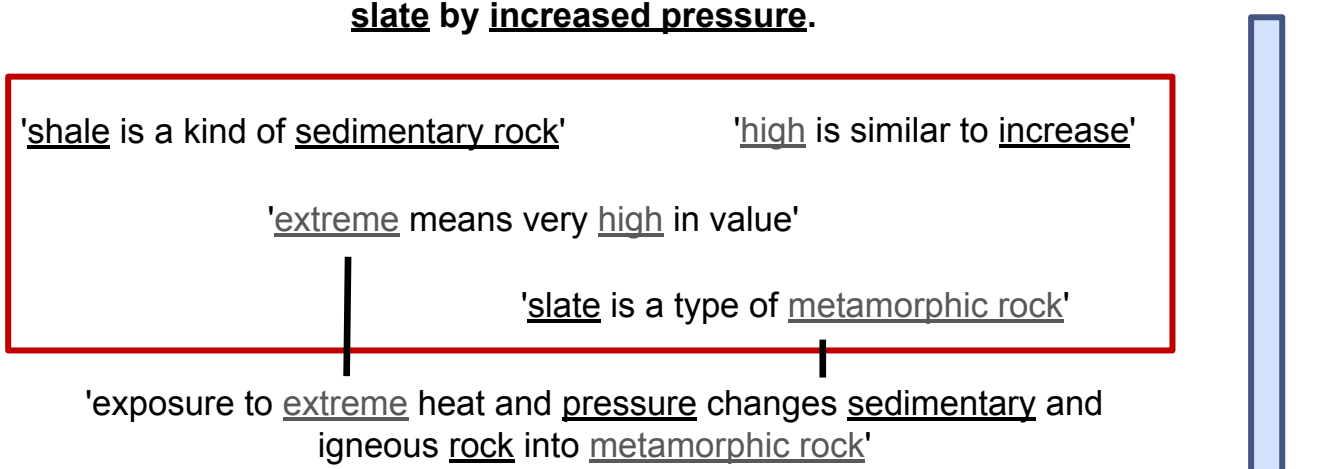
The Linguistic Perspective



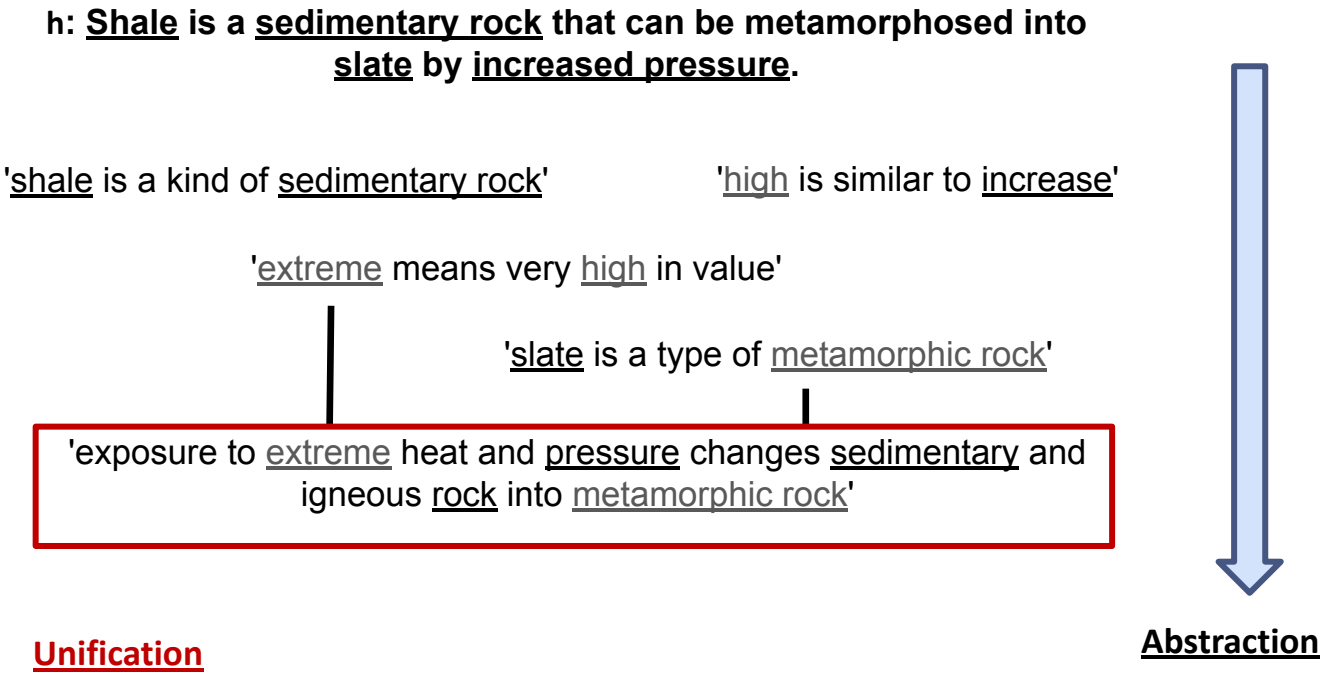
Valentino & Freitas, Technology & Philosophy (2024)

Examples of simple scientific explanations

h: Shale is a sedimentary rock that can be metamorphosed into slate by increased pressure.



Examples of simple scientific explanations

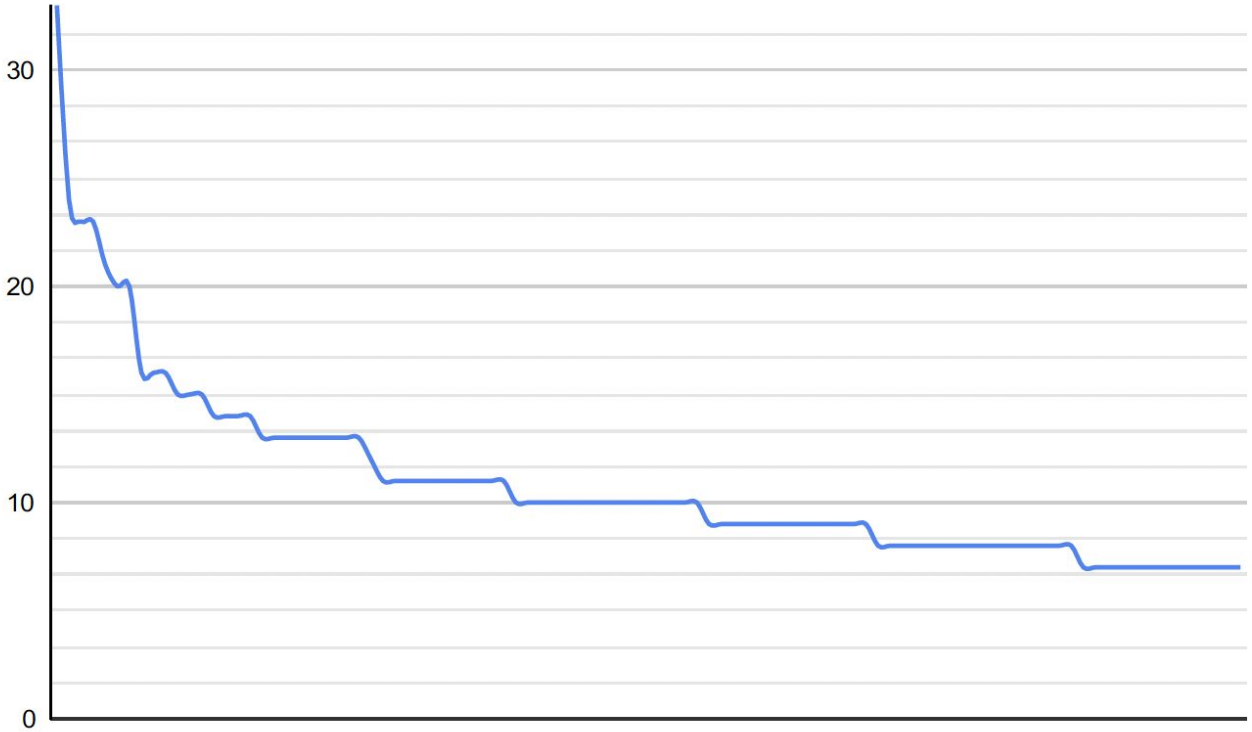


Composition

Grounding		Grounding	Occurrence
_ is a kind of _ (Taxonomic)		_ is a kind of _ (Taxonomic)	524
_ is a kind of _ (Taxonomic)		_ is part of _ (Part-of)	73
_ is a kind of _ (Taxonomic)		_ is made of _ (Made-of)	37
_ is a kind of _ (Taxonomic)	○	_ typically performs action _ on _ (Actions)	30
_ is a kind of _ (Taxonomic)		_ is a property of _ (Properties)	25
Grounding		**Central**	**Occurrence**
_ is a kind of _ (Taxonomic)		_ typically performs action _ on _ (Actions)	209
_ is a kind of _ (Taxonomic)		if _ then _ (Conditionals)	202
_ is a kind of _ (Taxonomic)		_ causes _ (Causal)	179
_ is a kind of _ (Taxonomic)	○	_ changes from _ to _ by _ (Processses)	153
_ is a kind of _ (Taxonomic)		_ uses _ for _ (Functional)	133

Valentino & Freitas, Technology & Philosophy (2024)

Long tail of explanatory sentences



Valentino & Freitas, Technology & Philosophy (2024)

Summary

Explanation involves a distinctive reasoning process consisting of **conjectures and criticisms**.

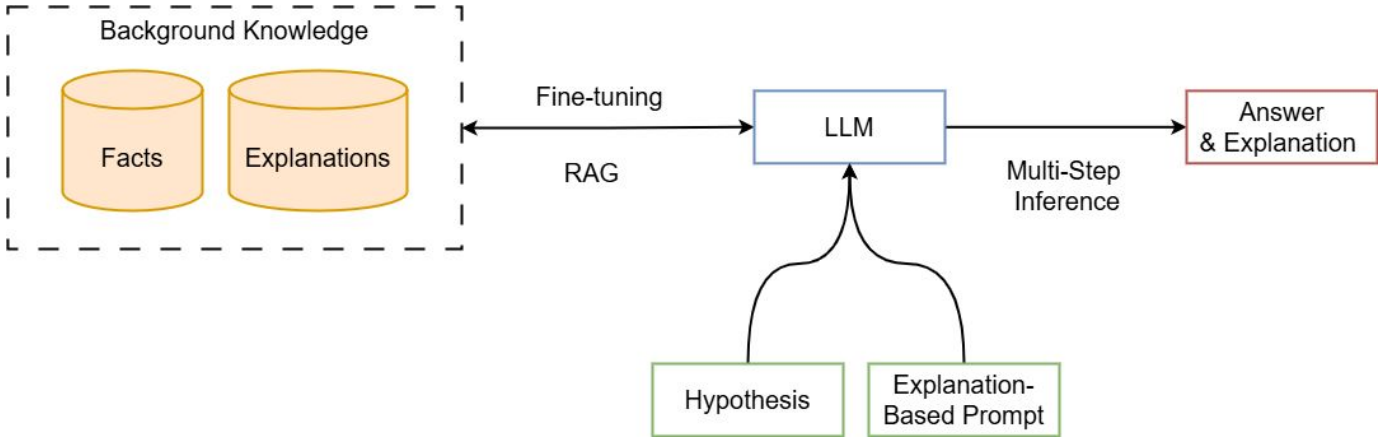
Explanations are typically a combination of **logical** and **causal-mechanistic** arguments whose function is to perform **unification**.

Unification is realised through a process of **abstraction** and is responsible to the creation of **argument and linguistic patterns** identifiable in corpora of natural language explanations.

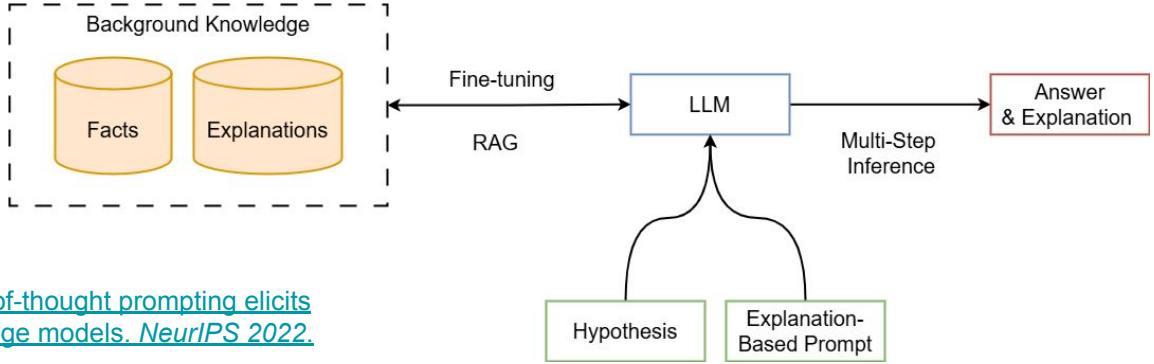
Part 2: Reasoning with Natural Language Explanations

Explanation-Based NLI Today

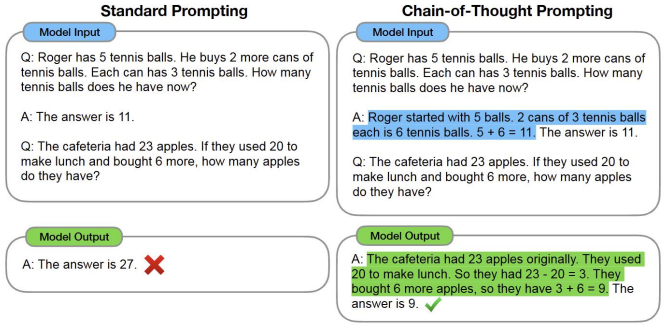
Mostly based on Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and specific prompting techniques to elicit explanation and multi-step inference.

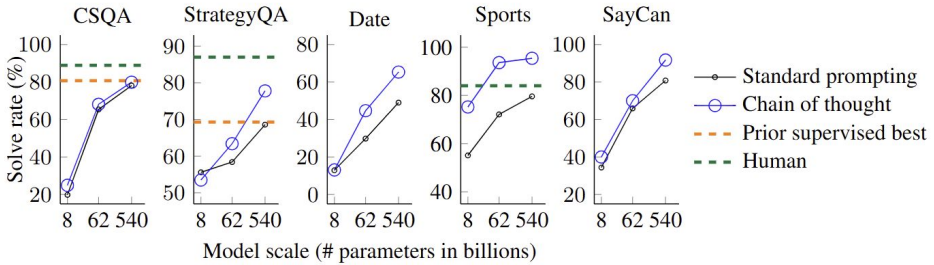


Explanation-Based NLI Today

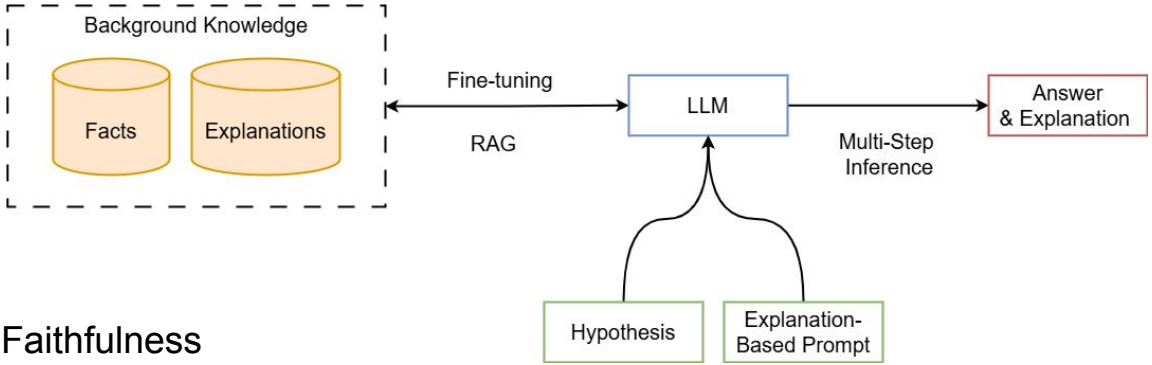


Wei, Jason, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS 2022.*

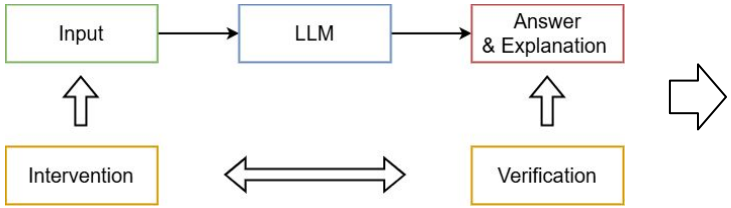


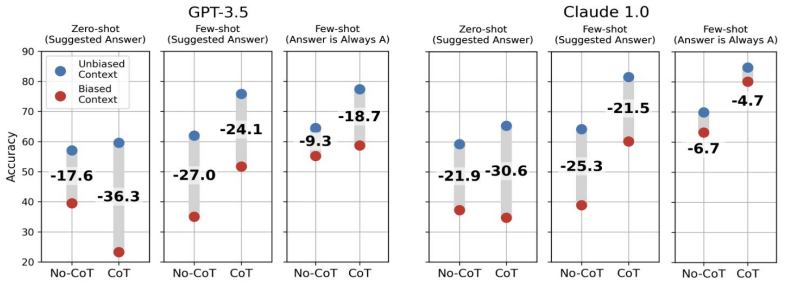


Explanation-Based NLI Today



Plausibility ≠ Faithfulness





Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* (2020).

Dziri, Nouha, et al. "Faith and fate: Limits of transformers on compositionality." *NeurIPS* 2024.

Turpin, Miles, et al. "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting." *NeurIPS* 2024.

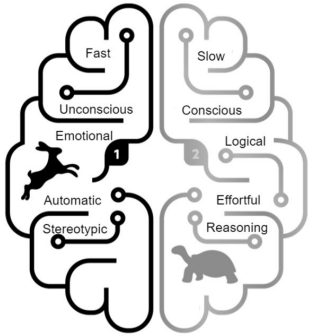
Neuro-Symbolic Models of Explanation

Explanations present the challenge of combining *linguistic and world knowledge* with *abstract inferential processes*, such as abductive, deductive, and causal reasoning.

A *hybrid neuro-symbolic* direction can help advance both technical and theoretical research questions:

1. Combining the flexibility, linguistic capabilities and world knowledge of *LLMs (System 1)* with the systematicity and controllability of *symbolic methods (System 2)*.
2. Neuro-symbolic architectures enable the modelling of *explicit hypotheses from theoretical accounts* (e.g., Epistemology, Linguistics, Cognitive Science) to help answer questions about the nature of explanation.

System 1	System 2
drive a car on highways	drive a car in cities
come up with a good chess move (if you're a chess master)	point your attention towards the clowns at the circus
understands simple sentences	understands law clauses
correlation	causation
hard to explain	easy to explain



Kahneman, Daniel. "Thinking, fast and slow" (2011).

Evans, Richard, and Edward Grefenstette. "Learning explanatory rules from noisy data." *Journal of Artificial Intelligence Research* 61 (2018).

Sarker, Md Kamruzzaman, et al. "Neuro-symbolic artificial intelligence." *AI Communications* (2021)

Theoretically-Informed Models of Explanation

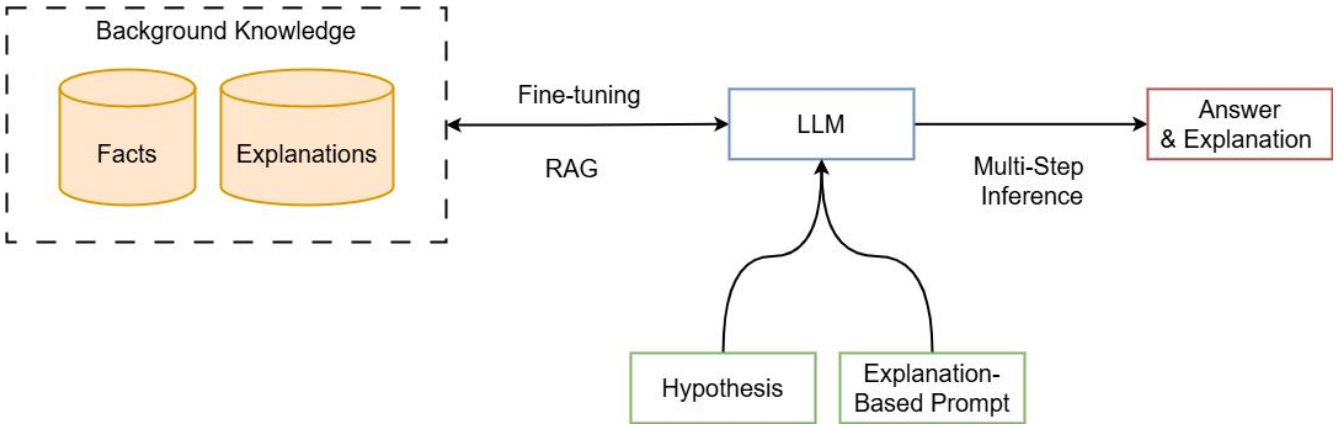
Quan, X., Valentino, M., Dennis, L. A., & Freitas, A. Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. *EMNLP 2024 (Outstanding Paper Award).*

Dalal, D., Valentino, M., Freitas, A., & Buitelaar, P. Inference to the Best Explanation in Large Language Models. ACL 2024.

Valentino, Marco, and André Freitas. "On the nature of explanation: An epistemological-linguistic perspective for explanation-based natural language inference." *Philosophy & Technology* 37.3 (2024).

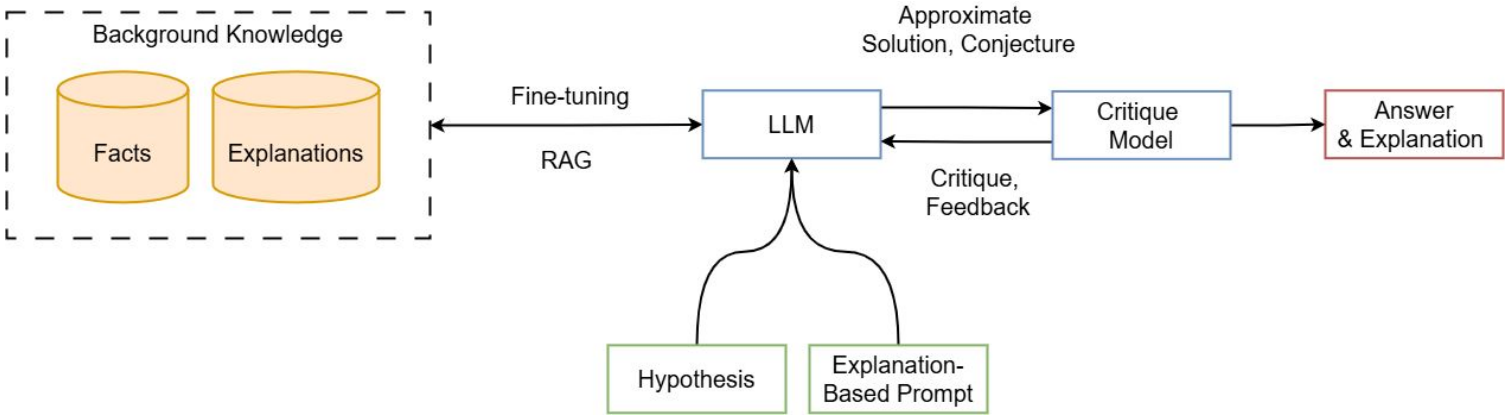
Neuro-Symbolic Explanation-Based NLI

How can neuro-symbolic methods support natural language explanations?



Neuro-Symbolic Explanation-Based NLI

How can neuro-symbolic methods support natural language explanations?



Neuro-Symbolic Explanation-Based NLI

System 1: is a fast, intuitive, unconscious and emotional, stereotypical, automatic and uses similarity with past experience to get a decision

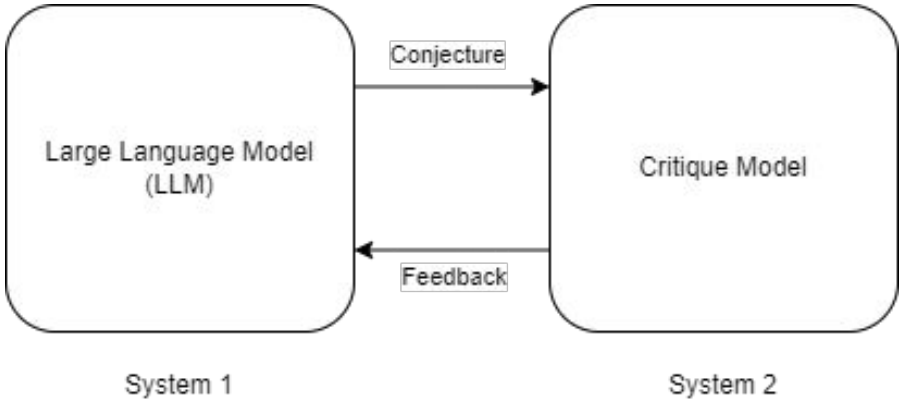
System 2: is a slow, conscious, logical, effortful and uses high-level reasoning to get decisions

Approximate solutions

Model Contextual Relevance

Fill knowledge gaps, reduce search space

Translate between different formats, languages



Provide feedback based on explicit knowledge

Provide formal guarantees on the correctness

Refine approximate solutions

Can Interpret models' behaviour

Kambhampati, Subbarao, et al. "LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks." *arXiv preprint arXiv:2402.01817* (2024).

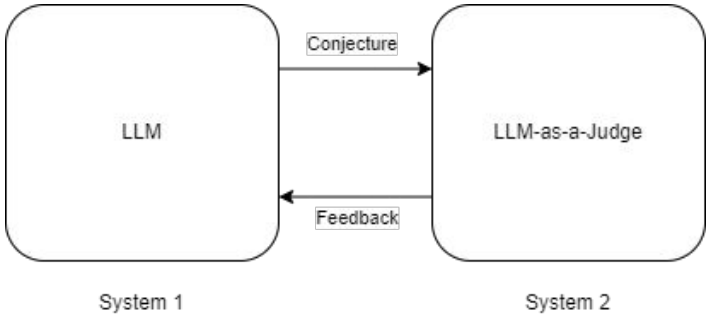
Critique via LLMs

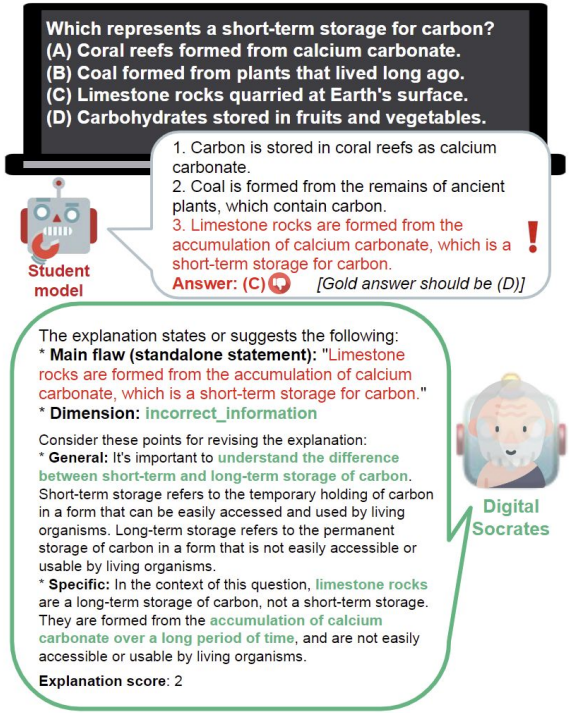
LLMs can be used as critique models (i.e., LLM-as-a-Judge).

The LLM-as-a-Judge paradigm offers flexibility as we do not need to parse or process the output of system 1 to perform the critique.

This paradigm is useful for improving and refining the performance of LLMs and better align smaller student models.

At the same time, LLMs do not have the reliability of a standard system 2 in a neuro-symbolic framework.





Yuling Gu, Oyvind Tafjord, and Peter Clark. 2024. Digital Socrates: Evaluating LLMs through Explanation Critiques.

Critique via LLMs

Correlation between human and LLM judgments across 20 datasets, considering different factors.

LLM exhibits a large variance across datasets in its correlation to human judgments.

Limited evidence that state-of-the-art LLMs are ready to replace expert or non-expert human judges.

LLMs can be useful, but need to be deployed with caution as they lack interpretability and consistency!

Best suited for soft and stylistic critiques (when formal guarantees are not required).

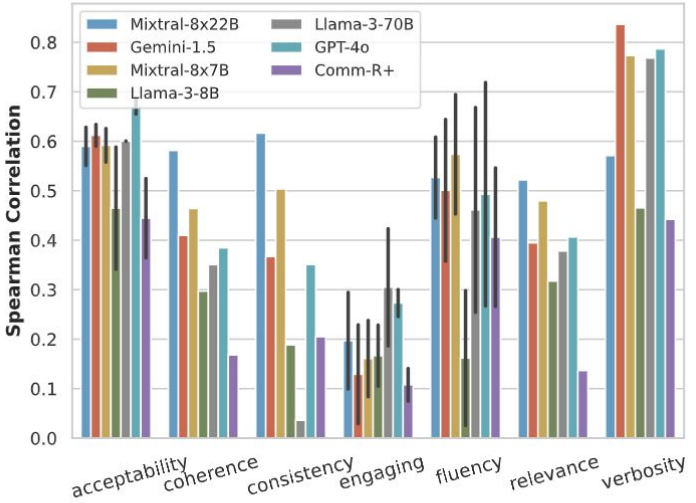


Figure 3: Correlation scores for those properties with exclusively graded judgements across datasets.

Bavaresco, Anna, et al. "LLMs instead of human judges? a large scale empirical study across 20 nlp evaluation tasks."

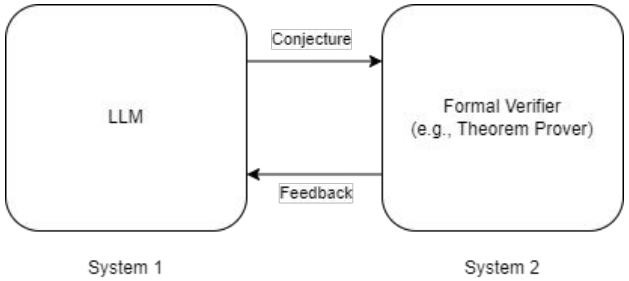
Hard Critique

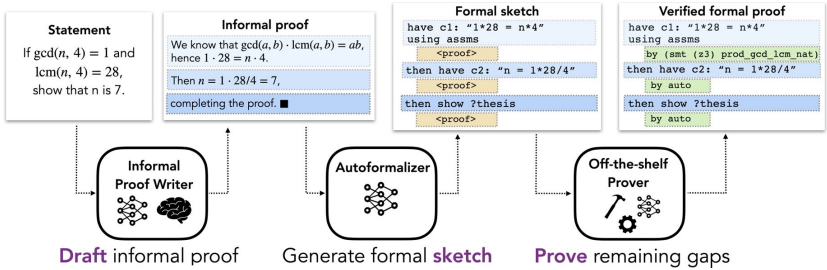
Symbolic models (e.g., Theorem Provers) can also be used as critique models

Less flexibility than the LLM-as-a-Judge paradigm but formal guarantees on the correctness of the output.

Possibility to assess the logical validity of explanatory arguments but help also with other dimensions (e.g. causal reasoning).

Require a process of “autoformalisation” or program synthesis.





Jiang, Albert Qiaochu, et al. "Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs."

Natural Language Explanations

Premise: A smiling woman is playing the violin in front of a turquoise background
Hypothesis: A woman is playing an instrument
Explanation: A violin is an instrument

Autoformalisation (FOL, Event-based Semantics)

```
theorem hypothesis:
  (* Premise: A smiling woman is playing the violin in front of a turquoise background. *)
  assumes asm: "Woman x ∧ Violin y ∧ Background z ∧ Turquoise z ∧ Smiling x ∧ Playing e ∧ Agent e
    x ∧ Patient e y ∧ InFrontOf x z"
  (* Hypothesis: A woman is playing an instrument. *)
  shows "∃ x y e. Woman x ∧ Instrument y ∧ Playing e ∧ Agent e x ∧ Patient e y"
```

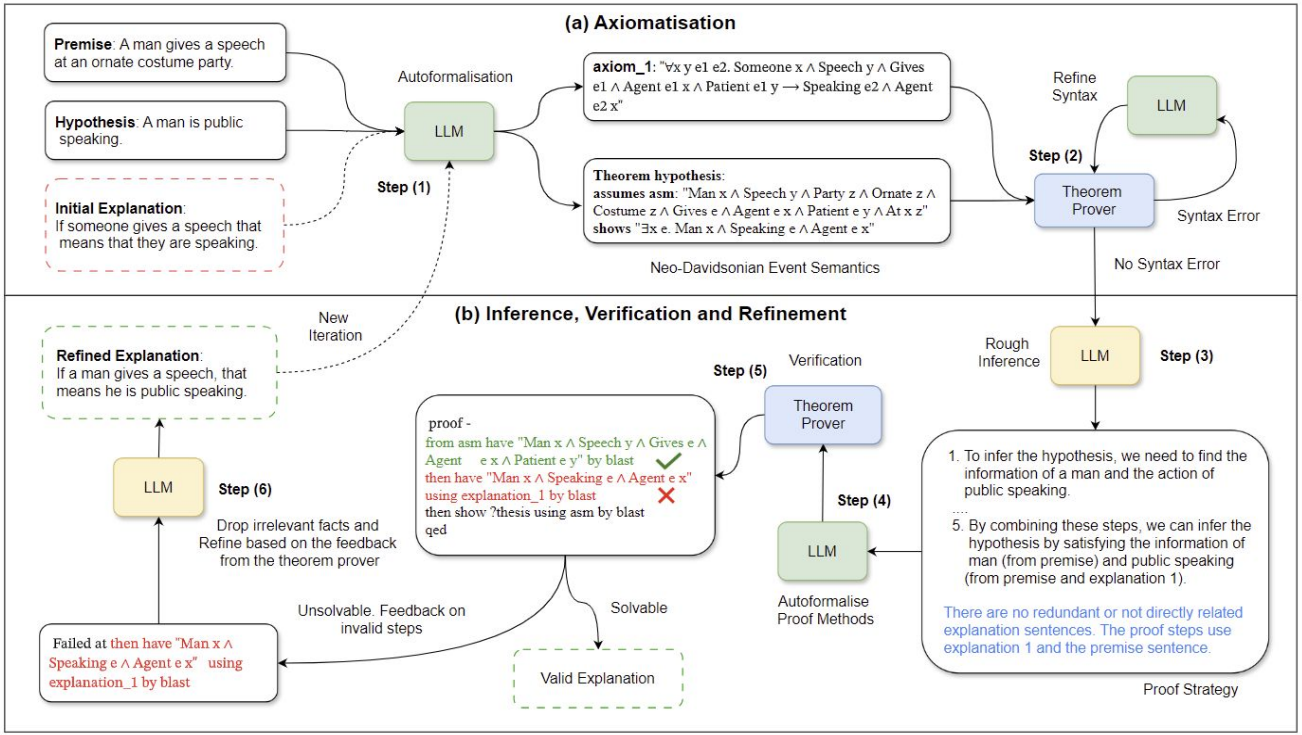
```
(* Explanation 1: A violin is an instrument. *)
axiomatization where
  explanation_1: "∀x. Violin x ⟶ Instrument x"
```

Proof Generation

```
proof -
  from asm have "Woman x ∧ Violin y ∧ Playing e ∧ Agent e x ∧ Patient e y" by blast
  then have "Woman x ∧ Instrument y ∧ Playing e ∧ Agent e x ∧ Patient e y" using explanation_1 by
    blast
  then show ?thesis using asm by blast
qed
```

Xin Quan, Marco Valentino, Louise Dennis, Andre Freitas. Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. *EMNLP 2024 (Outstanding Paper Award).*

Verification & Refinement of Natural Language Explanations



Xin Quan, Marco Valentino, Louise Dennis, Andre Freitas. Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. *EMNLP 2024 (Outstanding Paper Award).*

Verification & Refinement of Natural Language Explanations

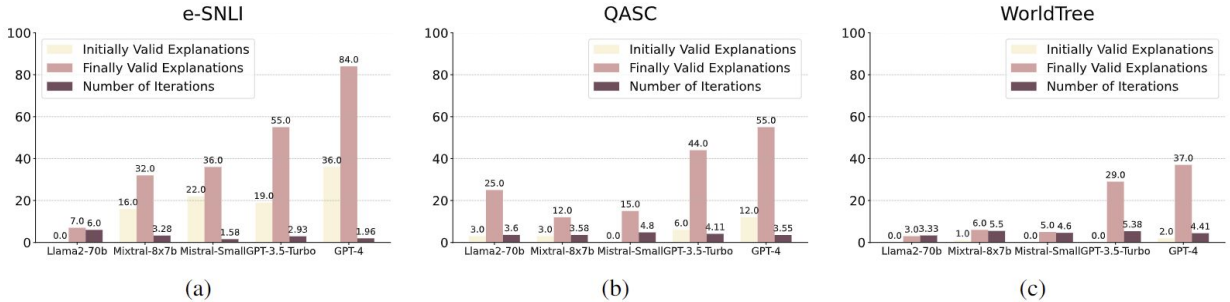
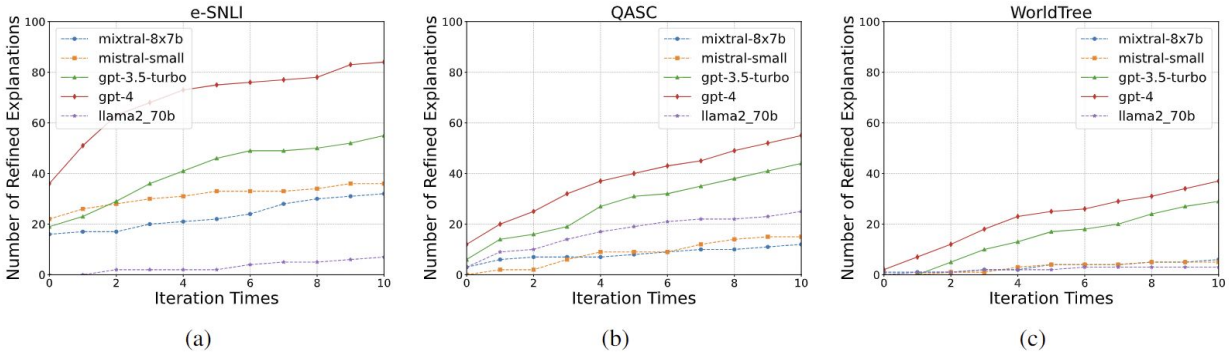


Figure 4: The initial and final number of logically valid explanations, along with the average iteration times required to refine an explanation for each LLM



Verification & Refinement of Natural Language Explanations

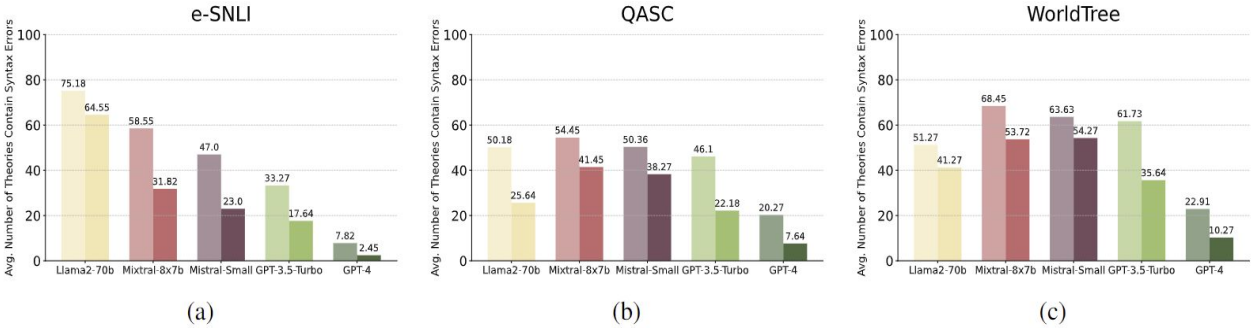
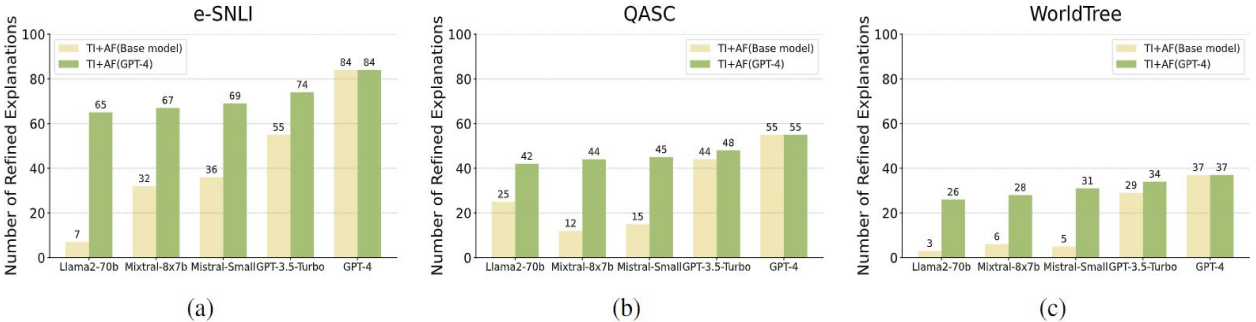


Figure 6: The average number of theories containing syntactic errors before and after the syntax refinement process

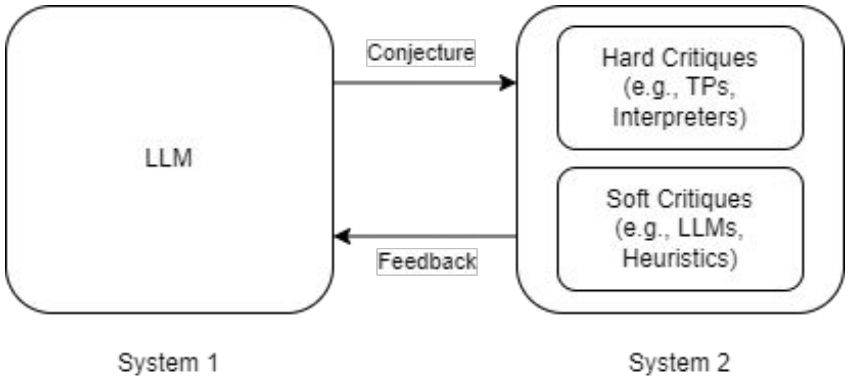


Combining Soft and Hard Critiques

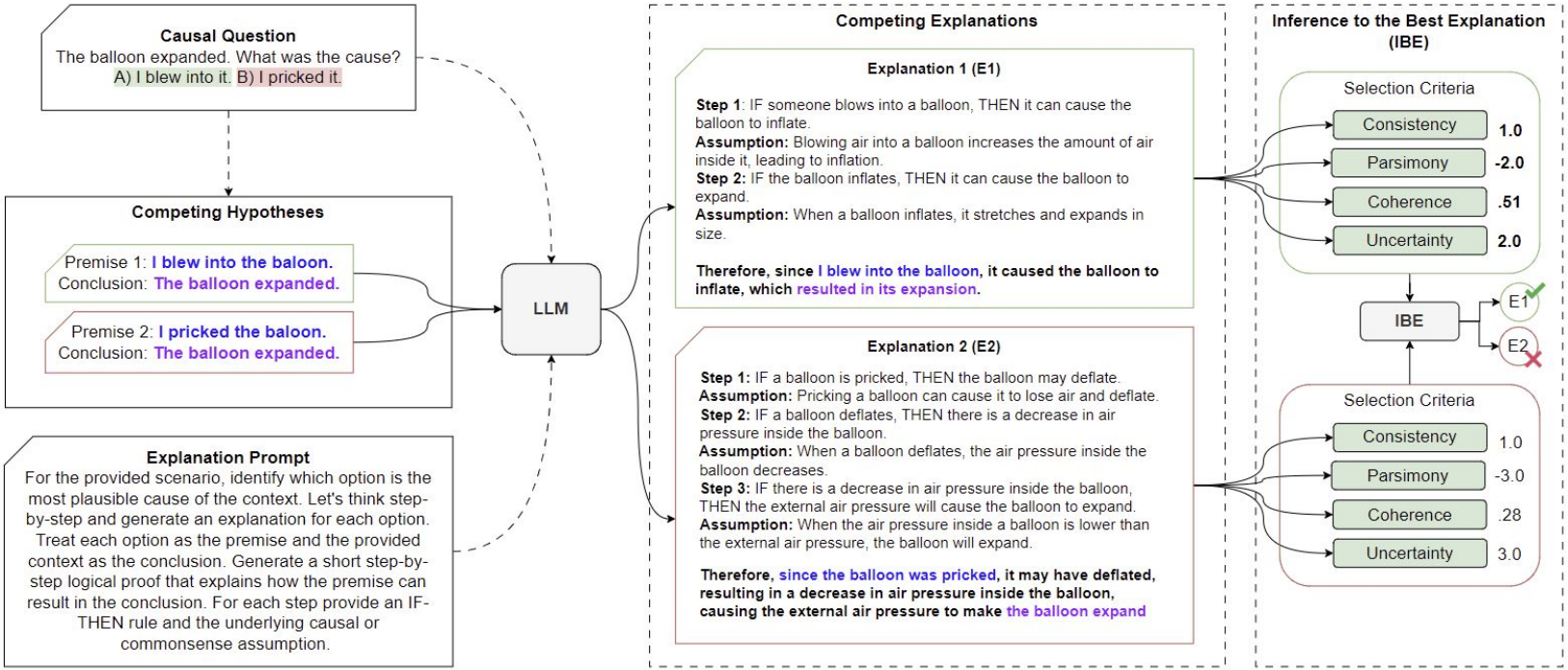
Soft and hard critiques can capture complementary aspects. Combining soft and hard critiques can help us build more diverse feedback and evaluation models.

Soft critiques for capturing stylistic and qualitative features that can be directly extracted from text.

Hard critiques for obtaining formal guarantees on the correctness of the generated output.



Combining Soft and Hard Critiques



Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the Best Explanation in Large Language Models.

Combining Soft and Hard Critiques

Evaluation Criteria:

Consistency aims to verify whether the explanation is logically valid. An explanation is logically consistent if it is possible to build a deductive proof linking premise and conclusion (via an external TP).

Parsimony, also known as Ockham's razor, favors the selection of the simplest explanation consisting of the fewest elements and assumptions (Sober, 1981). We adopt two metrics as a proxy of parsimony, namely proof depth, and concept drift.

Coherence evaluates the quality of each intermediate If-Then implication by measuring the entailment strength between the If and Then clauses via a fine-tuned NLI model.

Uncertainty considers the linguistic certainty expressed in the generated explanation as a proxy for plausibility. The linguistic uncertainty score is extracted using a fine-tuned sentence-level RoBERTa model from Pei and Jurgens (2021) operating on hedge words.

Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the Best Explanation in Large Language Models.

Combining Soft and Hard Critiques

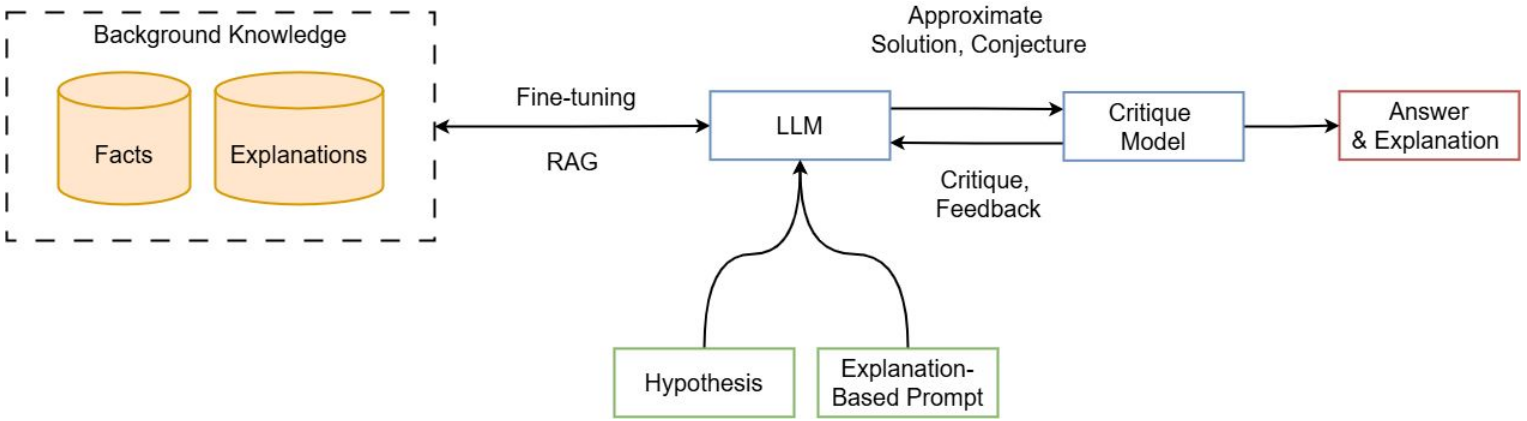
	COPA			E-CARE		
	GPT 3.5	**LlaMA 2 13B**	**LlaMA 2 7B**	**GPT 3.5**	**LlaMA 2 13B**	**LlaMA 2 7B**
Baselines						
GPT3.5 Judge	.59	.47	.63	.43	.61	.52
Human	***.95***	***1.0***	***.91***	***.90***	***.91***	***.92***
IBE Features						
Consistency	.51	.52	.55	.54	.54	.54
Depth (Parsimony)	.67	.53	.63	.66	.56	.54
Drift (Parsimony)	.67	.63	.58	.66	.57	.57
Coherence	.66	.66	.56	.56	.57	.59
Linguistic Uncertainty	.70	.65	.61	.59	.56	.60
Composed Model						
Random	.50	.50	.50	.50	.50	.50
+ Consistency	.51	.52	.55	.54	.54	.54
+ Depth	.67	.53	.63	.66	.56	.56
+ Drift	.70	.65	.65	.72	.66	.65
+ Coherence	.73	.71	.69	.73	.68	.69
+ Linguistic Uncertainty	**.77**	**.74**	**.70**	**.74**	**.70**	**.73**

Table 1: An ablation study and evaluation of the IBE criteria and the composed *IBE-Eval* model. *IBE-Eval* outperforms the GPT 3.5 Judge baseline by an average of +17.5% across all all models and tasks.

Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the Best Explanation in Large Language Models.

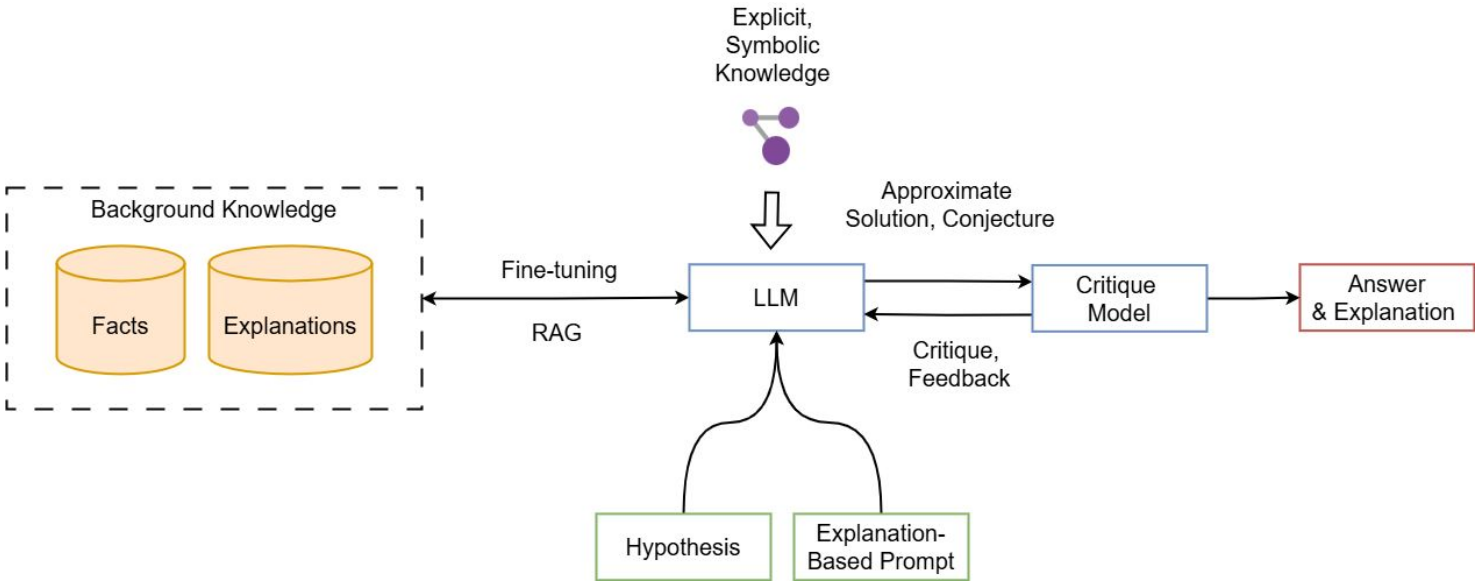
Neuro-Symbolic Explanation-Based NLI

How can neuro-symbolic methods support natural language explanations?



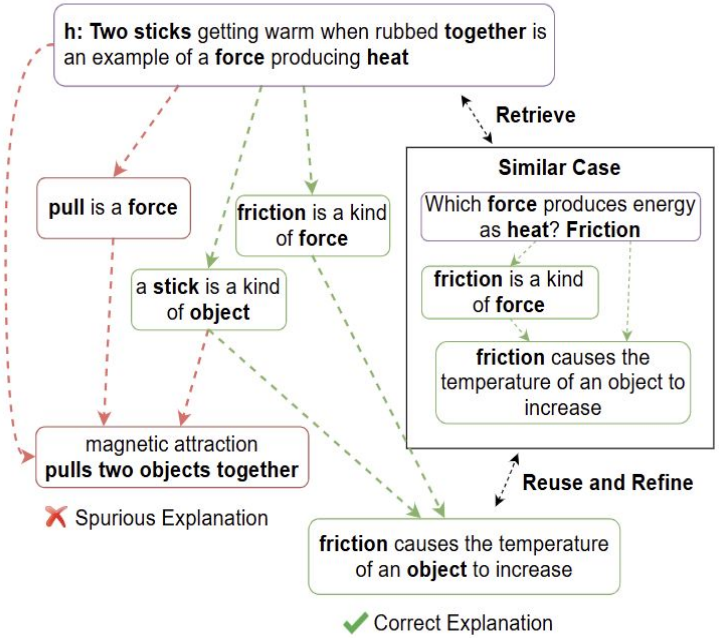
Neuro-Symbolic Explanation-Based NLI

How can neuro-symbolic methods support natural language explanations?



Leveraging Explanatory Unification Patterns

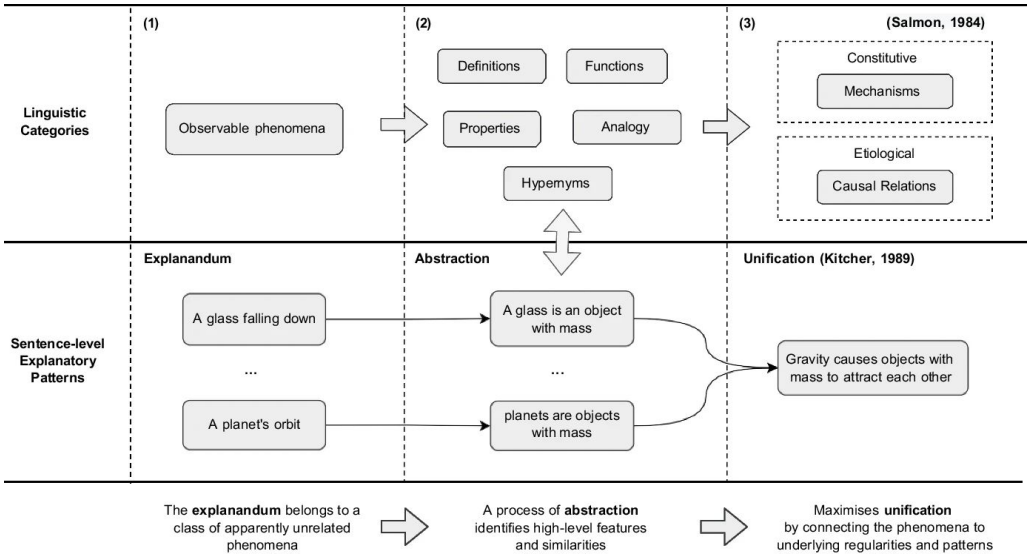
1. Similar problems tend to require similar explanations;
2. Abstract facts tend to express general explanatory knowledge about underlying regularities, being frequently reused to explain a large variety of phenomena;
3. Prior solutions can help constraint the search space, reducing the risk of composing spurious inference chains.



(Valentino et al., 2022)

Modelling Explanatory Power via Unification

The more a fact fi is reused for explaining similar hypotheses, the higher its explanatory power:



(Valentino & Freitas, 2024)

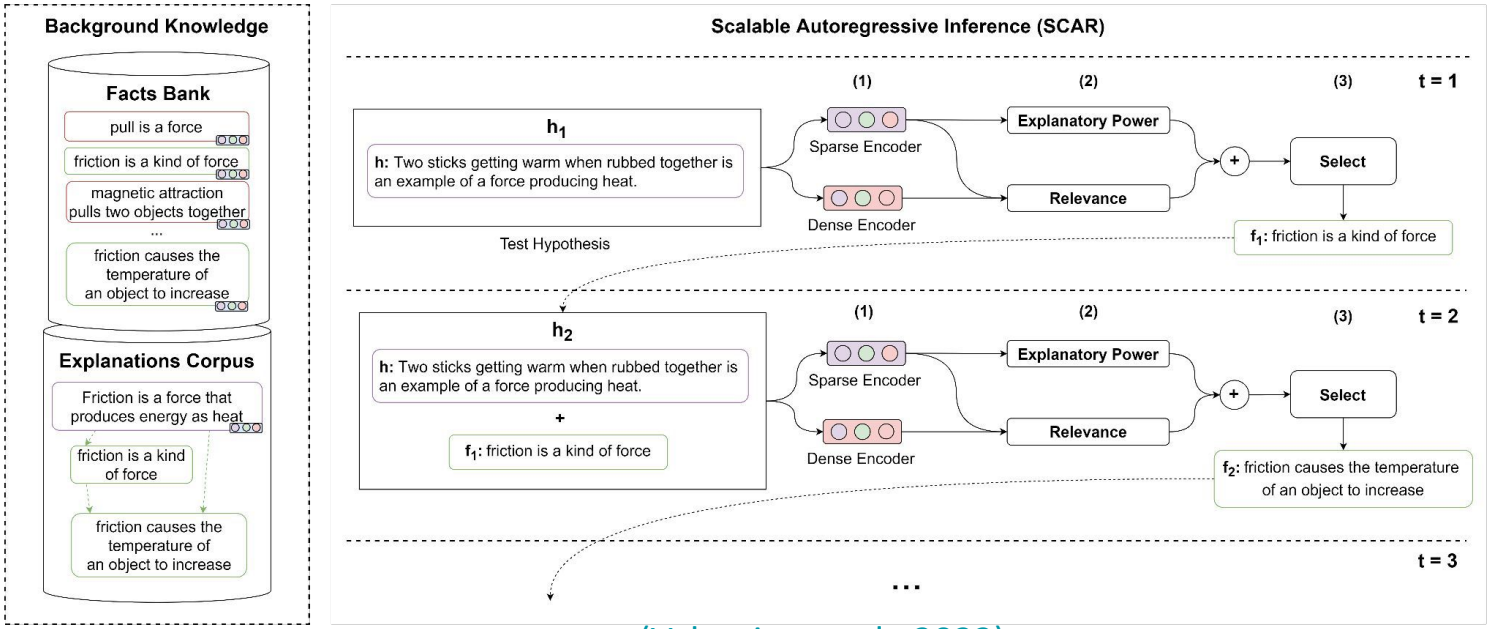
$$pw(f_i, h) = \sum_{h_k \in kNN(h)}^{K} sim(s(h), s(h_k)) \cdot \mathbb{1}(f_i, h_k)$$

$$\mathbb{1}(f_i, h_k) = \begin{cases} 1 & \text{if } f_i \in E_k \\ 0 & \text{if } f_i \notin E_k \end{cases}$$

(Valentino et al., 2021)

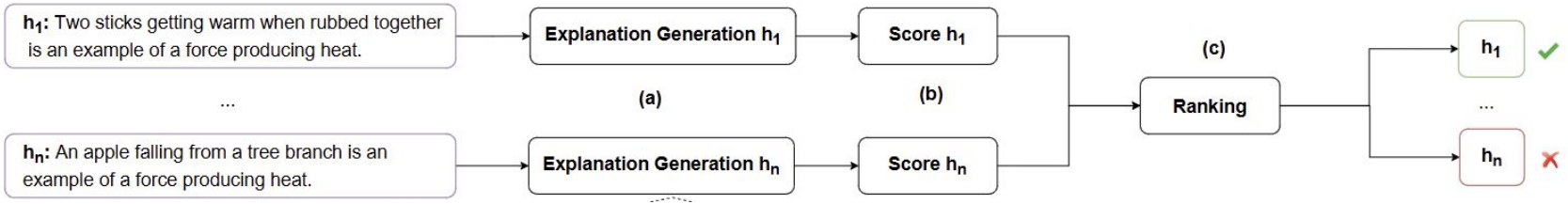
Hybrid Autoregressive Inference

Autoregressive Formulation: $$P(\mathcal{P}_{seq}|q) = \prod_{t=1}^{n} P(p_t|q, p_1, ..., p_{t-1}),$$

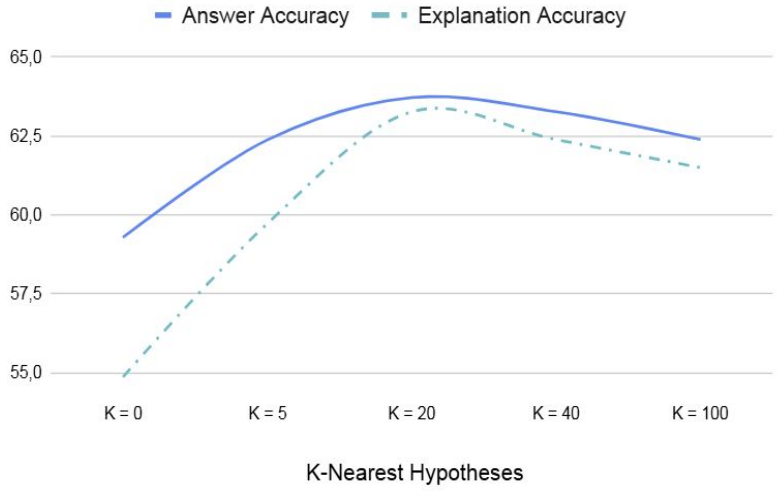


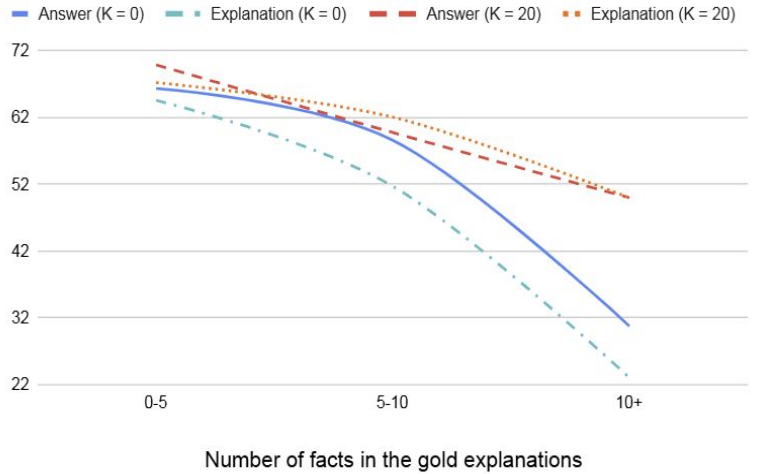
(Valentino et al., 2022)

Cas-Based Abductive Natural Language Inference



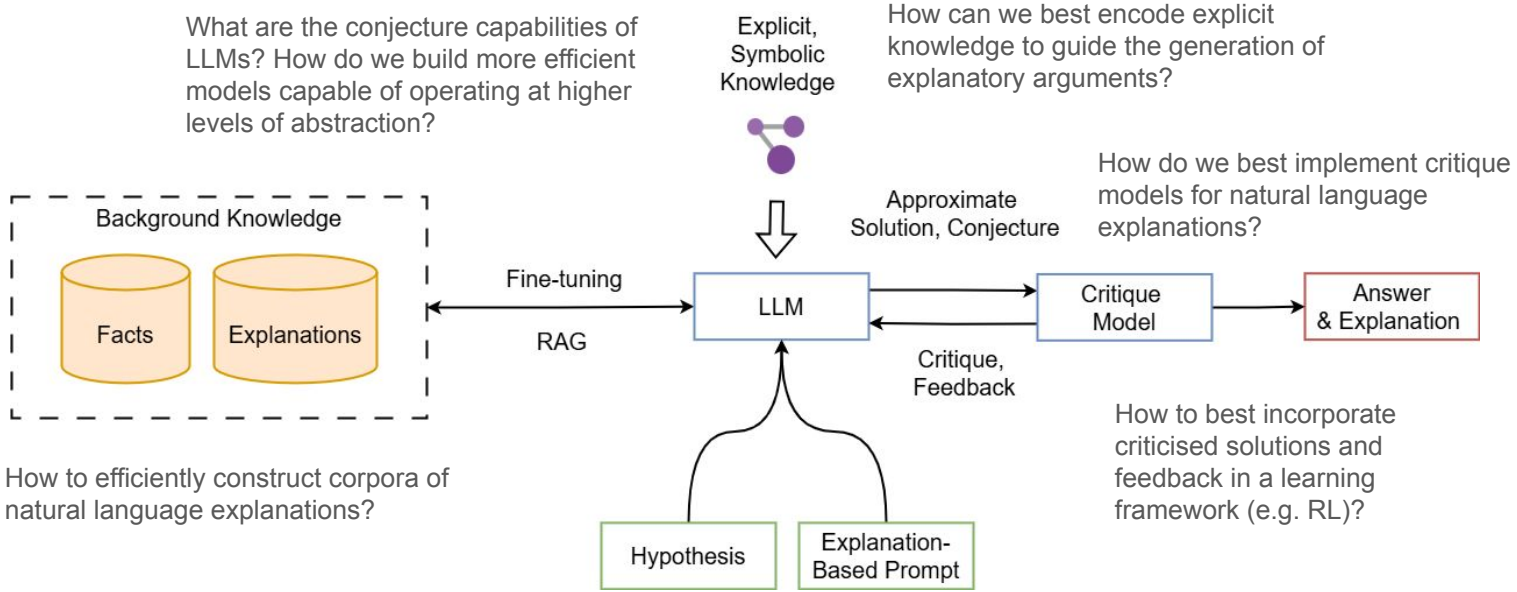
(Valentino et al., 2022)





Conclusion & Future Work

Establishing a dialogue between theory and practice.



Thank you!

Marco Valentino and André Freitas.
Reasoning with Natural Language Explanations.
EMNLP 2024 (Tutorials)

https://sites.google.com/view/reasoning-with-explanations

https://www.marcovalentino.net
marco.valentino@idiap.ch