MANCHESTER
1824

The University of Manchester

WP2: Explainable Neuro-Symbolic Safeguard Framework

MANCHESTER 1824
The University of Manchester

AI systems

Digital Experimental Cancer Medicine Team

Vision - Ambition

To enable:

- **<u>Neural safety through explainability</u>**.

- **<u>Explainability as a support of formal neural verification.</u>**

- End-users to communicate and implement inference controls.

- The creation of safe NN-based applications.

Vision - Ambition

End-users to communicate and implement inference controls

Neural safety through explainability

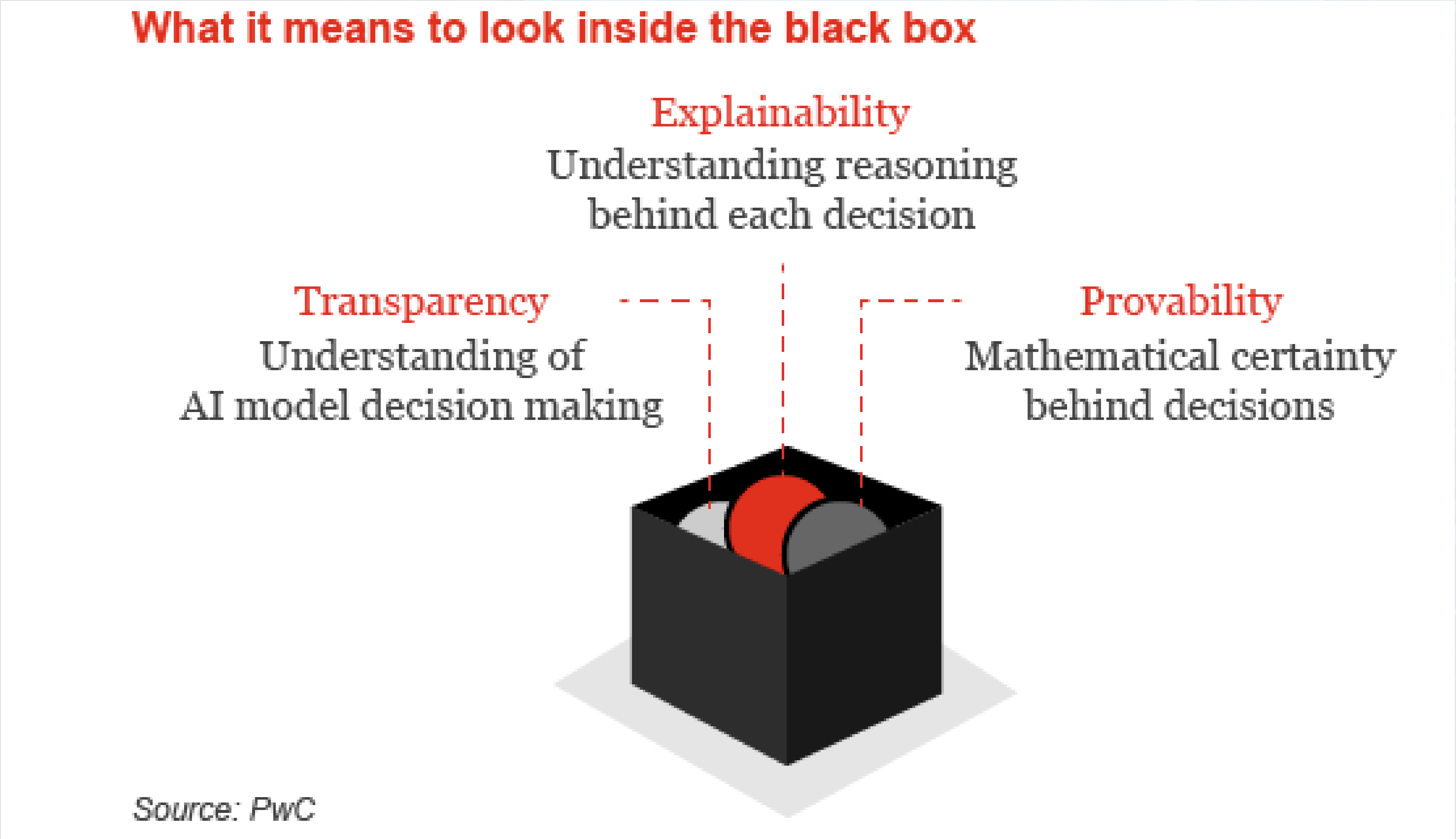
Explainabiltiy as an enabler of formal neural verification.

The creation of safe NN-based applications

WP3

WP1

Vision - Ambition



Concretely - Safeguards

- Can we ensure that the model does not contradict a medical fact?
- Can we ensure a chatbot is not outputting a racist statement?
- How software developers can program the safeguards into the model?

More Formally

- **O1**: Develop a novel conceptual/symbolic safeguard mechanism for neuro-symbolic platforms.

- EnnCore will pioneer the use of **neuro-symbolic architectures** and **explainability/interpretability** mechanisms to support end-users specifying a **conceptual safeguard** core to neural-based AI systems.

Tasks

- **<u>T2.1:</u>** Systematic analysis of neural interpretability methods.
- **<u>T2.2:</u>** Design and implementation of the neurosymbolic safeguard.
- **<u>T2.3:</u>** Evaluation.

Runs from: **<u>M4-M30</u>**

Neural safety through explainability

Explainability

Semantic controls

Semantic probing

Semantic fine-tuning

Architectural controls

Disentangled encodings

Synthetic Datasets

Information Theoretical Asps.

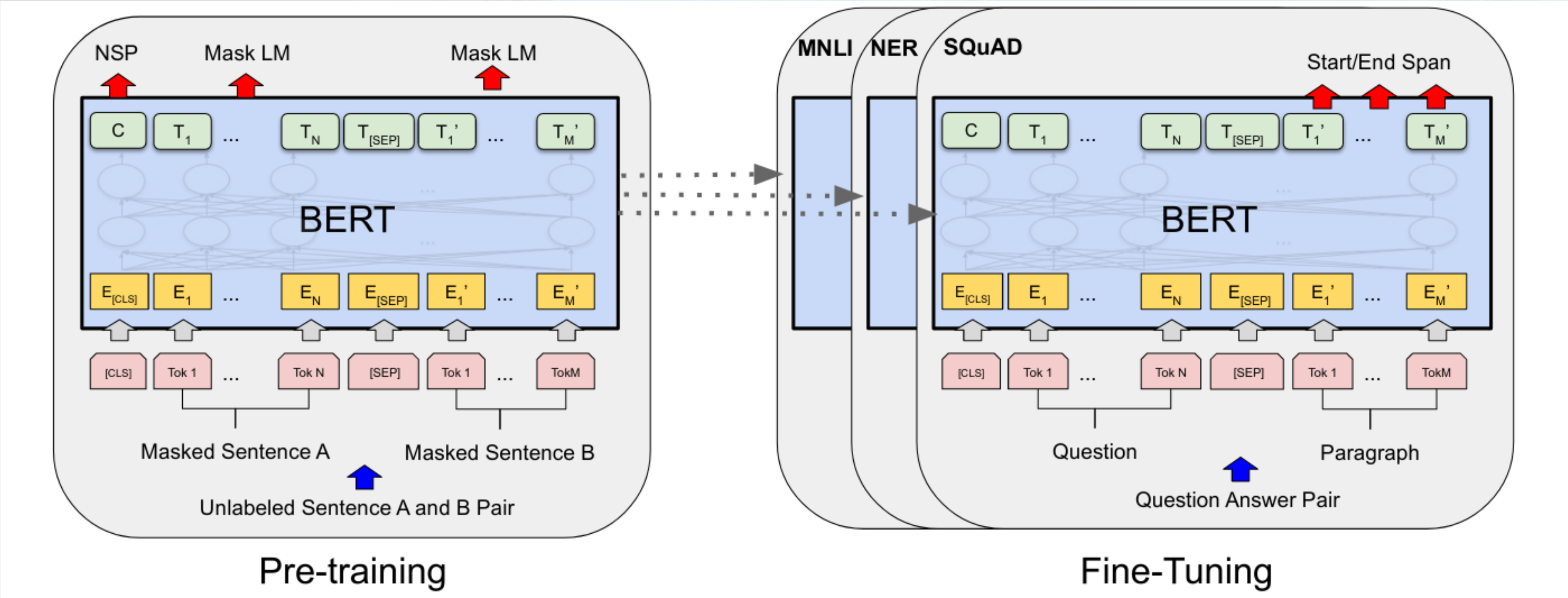
Formal Software Verification

Embeddings visualisation

Safety Reports

Safety Specs

Motivational Scenario



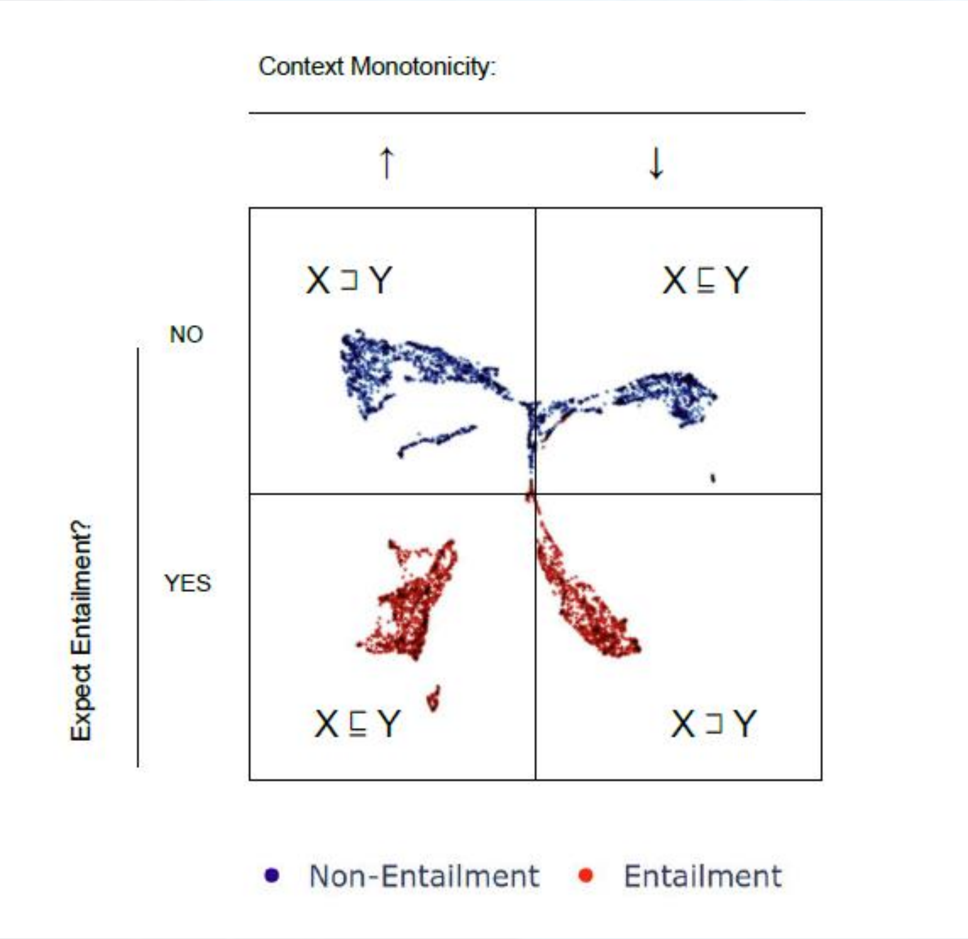
#1: Development of novel semantic probing/fine-tuning mechanisms

Monotonicity: e.g. interpreting negation and generalised quantifiers

		Auxilliary Label
Context	I did not eat any x for breakfast.	$\downarrow$
Insertion Pair	(fruit, raspberries)	$\sqsupset$
		NLI Label
Premise	I did not eat any fruit for breakfast.	Entailment
Hypothesis	I did not eat any raspberries for breakfast.	

Probing Context Monotonicity and Relational Knowledge in Neural Natural Language Inference

#1: Development of novel semantic probing/fine-tuning mechanisms



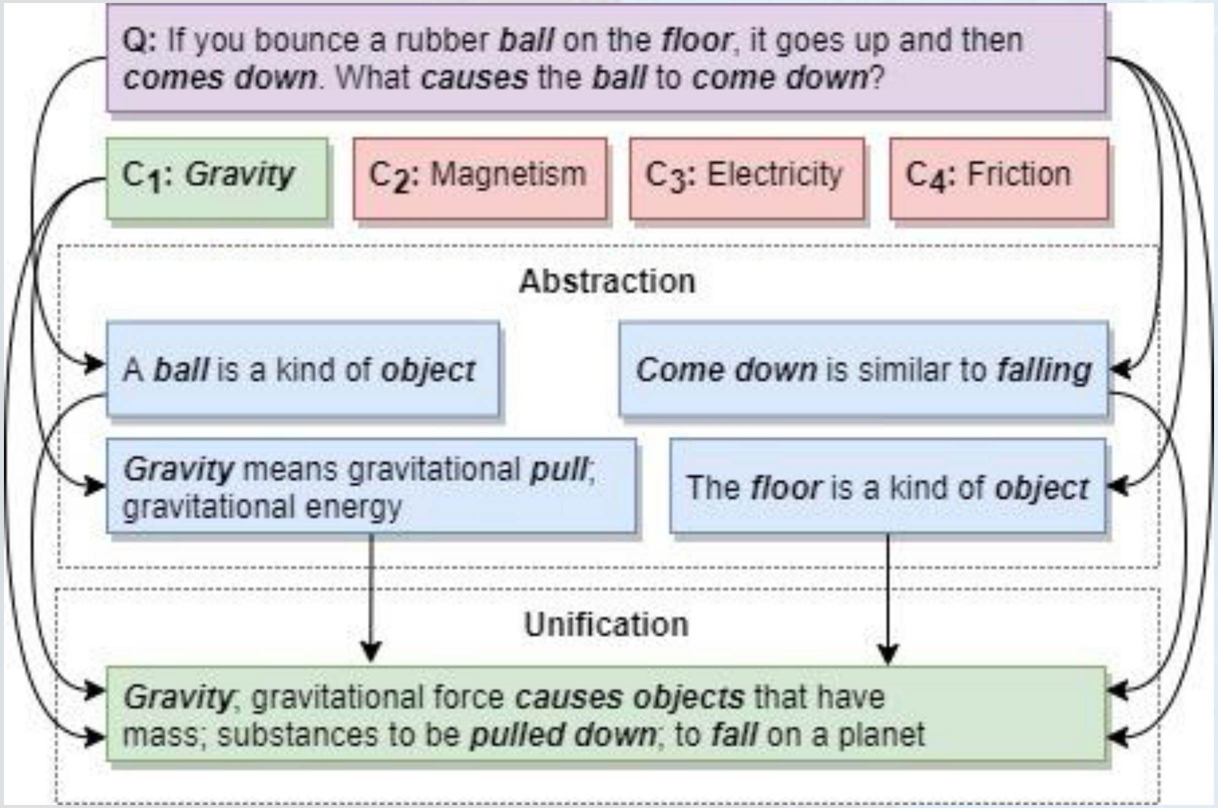
Probing Context Monotonicity and Relational Knowledge in Neural Natural Language Inference

#1: Development of novel semantic probing/fine-tuning mechanisms

Model	Training Data	Task: / Portion of Sequence:	Accuracy at Maximum Selectivity: Lexical Relation Classification	Monotonicity Classification	
			XY, concatenated	XY, concatenated	X
Majority Class Baseline	-		0.4050	0.5056	0.5056
Random	-		0.4345	0.5715	0.5715
`bert-base`	-		0.4605	0.6781	0.7141
`bert-large`	-		0.4565	0.6766	0.6728
`bert-base`	SNLI		0.5310	0.7330	0.7281
`bert-base`	SNLI + HELP		0.5345	0.7455	0.743
`facebook/bart-large`	MNLI		0.7056	0.7143	0.7513
`facebook/bart-large`	MNLI + HELP		0.7382	0.7625	0.7643
`roberta-large`	MNLI		**0.8390**	0.6771	0.8243
`roberta-large`	MNLI + HELP		0.7821	**0.8742**	**0.8691**

Probing Context Monotonicity and Relational Knowledge in Neural Natural Language Inference

#2: Semantic controls over complex end-to-end neural architectures

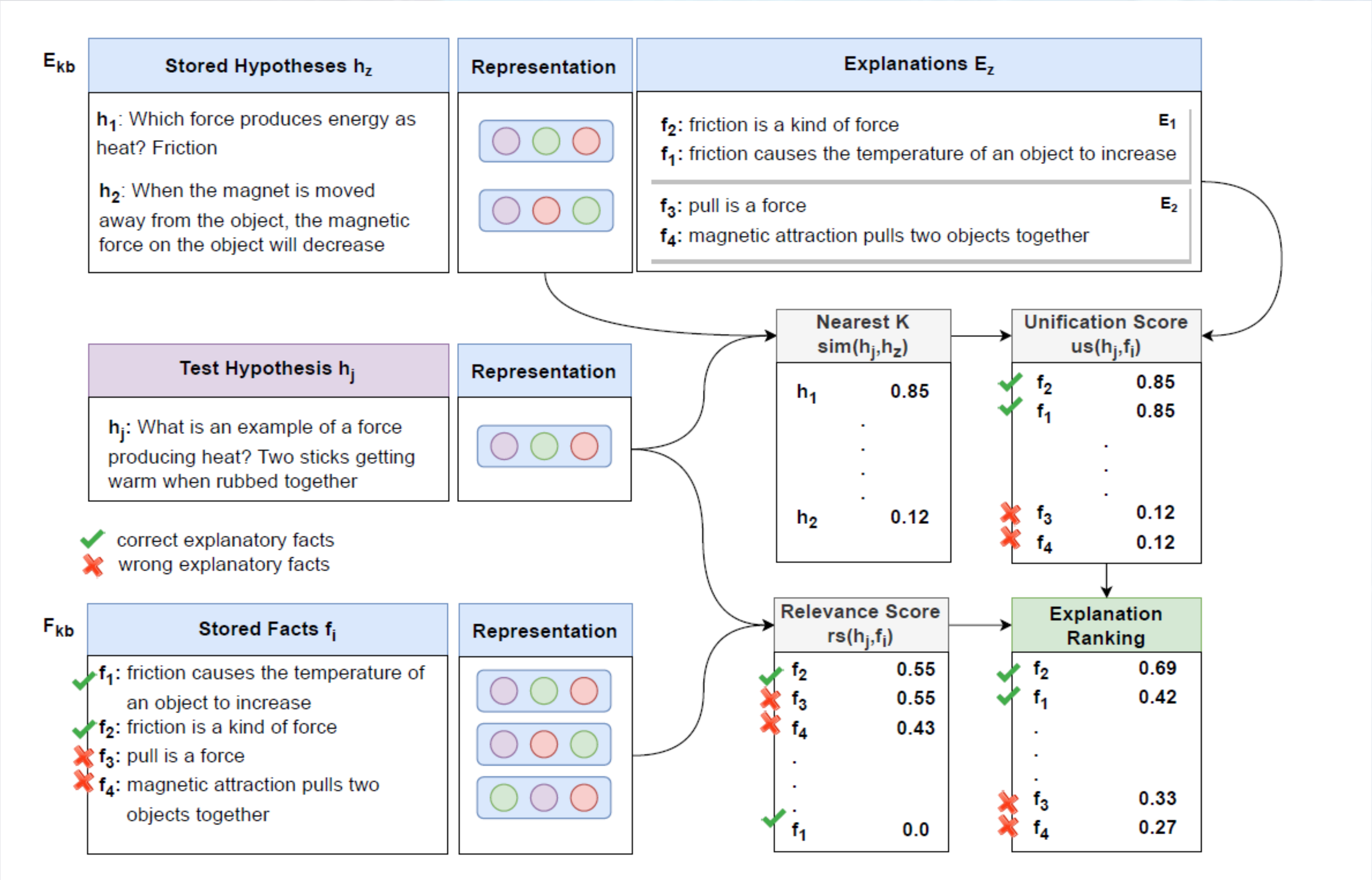


Abstraction

Unification

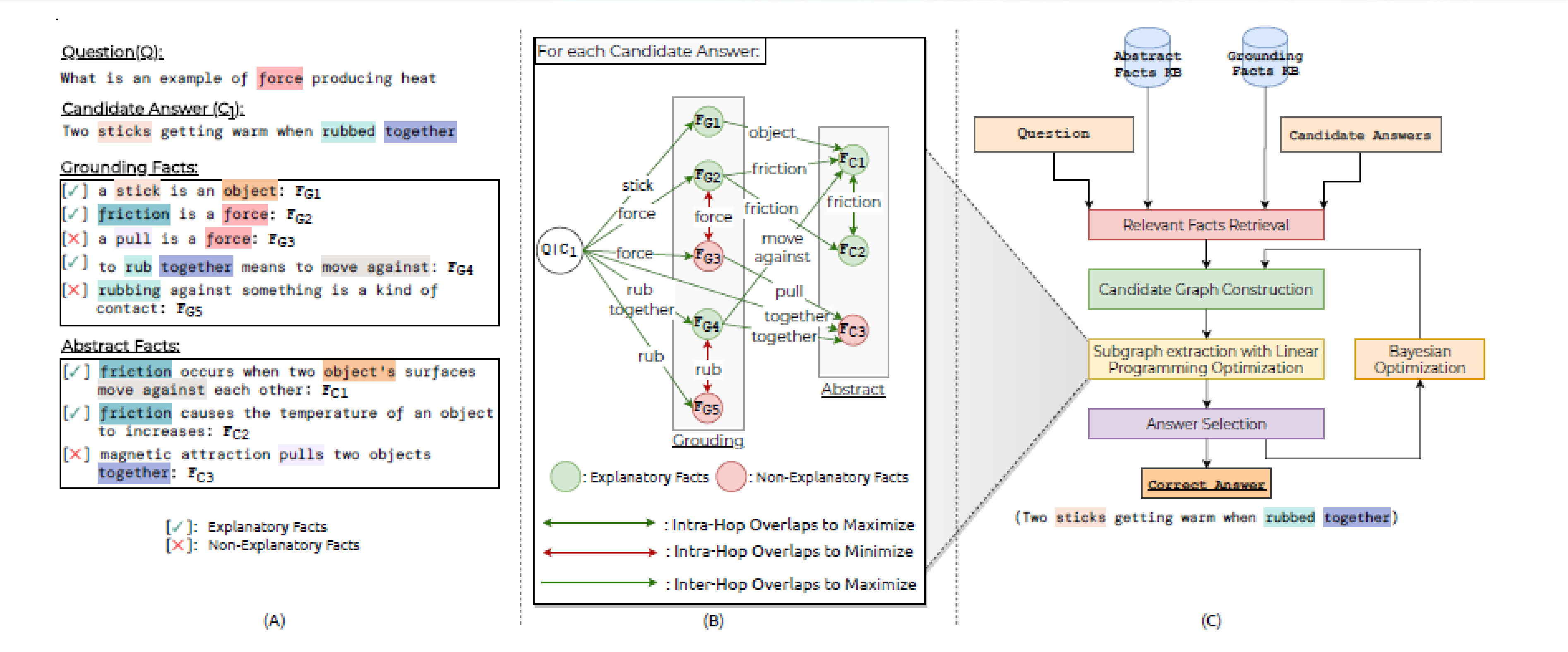
Unification-based Reconstruction of Multi-hop Explanations for Science Questions (EACL 2021)

#2: Semantic controls over complex end-to-end neural architectures



Unification-based Reconstruction of Multi-hop Explanations for Science Questions (EACL 2021)

#2: Semantic controls over complex end-to-end neural architectures



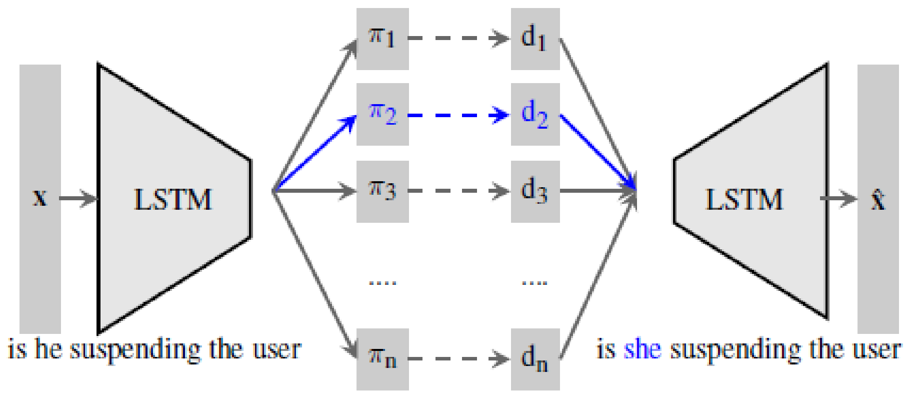
Unification-based Reconstruction of Multi-hop Explanations for Science Questions (EACL 2021)

#2: Semantic controls over complex end-to-end neural architectures

#	Parameter	Value WT	Value ARC
1	Question-Abstract overlap (θ_{qa})	0.10	0.09
2	Question-Grounding overlap (θ_{qg})	0.98	0.84
3	Abstract-Abstract overlap (θ_{aa})	0.01	0.11
4	Grounding-Abstract overlap (θ_{ga})	0.14	0.23
5	Grounding-Grounding overlap (θ_{gg})	-0.99	-0.92
6	Abstract Relevance (θ_{ar})	0.03	0.09
7	Grounding Relevance (θ_{gr})	0.36	0.14
8	Edge weight (θ_{ew})	0.80	0.26
9	Node weight (θ_{vw})	0.76	0.67

ExplanationLP: Abductive Reasoning for Explainable Science Question Answering

#3: Development of symbolic disentangled encodings



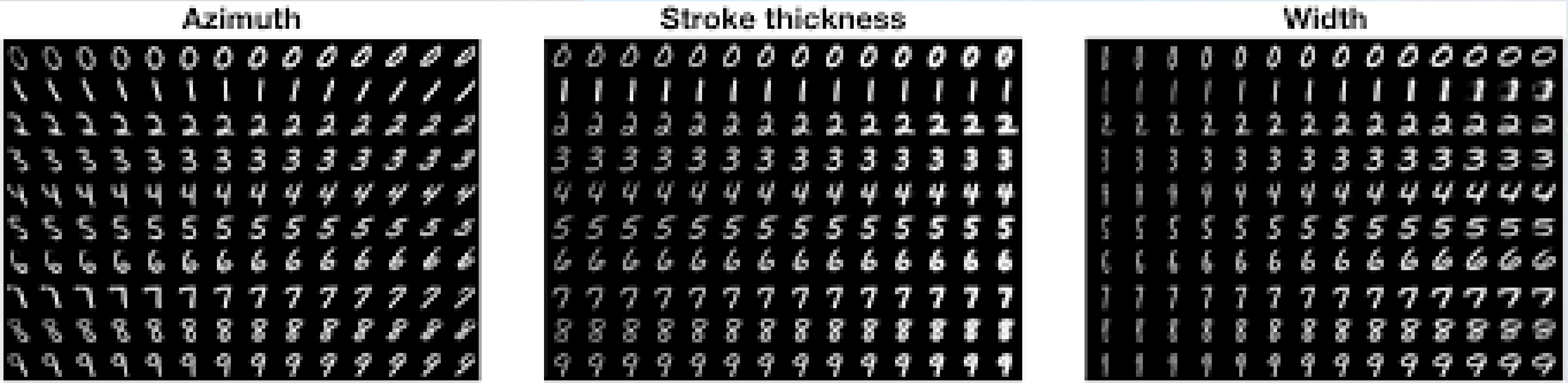
(a) Model architecture

N	Discrete Factor (π)	Sentence example
1	Verb/Object (VO)	is he suspending the computer
2	Gender (G)	is she suspending the user
3	Sentiment (S)	is he not suspending the user
4	Tense (T)	was he suspending the user
5	Subject Number (SN)	are they suspending the user
6	Object Number (ON)	is he suspending the users
7	Question/Answer (QA)	he is suspending the user

(b) Traversal generation

Towards a Discrete Latent Variables Model for Text Disentanglement

#3: Development of symbolic disentangled encodings



Other aspects

#4: Development of synthetic datasets for textual inference

#5: Connection with feature selection and feature stability

#6: Connection with formal verification

Discussion points / TODOs

- Initial assumption here is delivering an end-to-end solution
- Closing the gap between different modalities
- Which data modalities to target?
- Planning
- First collaboration points
- Start via synthetic datasets