Natural Language & Proofs: A Neuro-symbolic Perspective

Andre Freitas

Reasoning & Explainable AI (ExplAIn) Lab

EuroProofNet

Sept 2022

MANCHESTER 1824

The University of Manchester

idiap RESEARCH INSTITUTE


ExplAIn Lab

Natural Language Inference (NLI)

<u>Claim:</u> Specialized cells protect the human body from disease-causing microbes by producing chemicals that destroy the microbes.

True | False

Why? (Explanation)

Multi-hop
Multi-premise

Specialized cells are a source of chemicals that destroy disease-causing microbes.

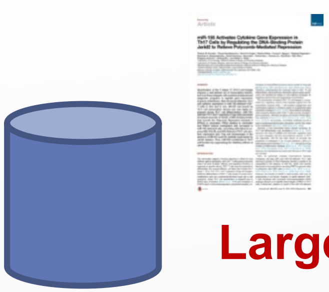
disease-causing microbes have a negative impact on the body.

Fact banks

Expert-level scientific inference & explanation

Claim: BRCA2 promotes the joining of undamaged homologous repair molecules via RAD51 homolog 1 in humans.

- BRCA2 and RAD51 homolog 1 are both involved in HRR in humans.
  - RAD51 is a eukaryotic gene that encodes the RAD51 homolog gene.
  - BRCA2 promotes the assembly of RAD51 homolog 1 onto SS DNA in HRR.
  - BRCA2 is a human protein involved in DSB DNA break repair via HRR
    - BRCA2 is a human protein involved in HRR.
      - BRCA2 is a human gene that encodes the BRCA2 protein.
      - BRCA2 protein is a tumour suppressor involved in HRR.
    - HRR is a DSB DNA repair process wherein damaged DNA is replaced by undamaged homologous molecules from sister chromatids or paternal/maternal copies of chromosomes.
      - HRR repairs damage to DNA using information copied from a homologous undamaged molecule.
      - HRR is the primary process for repairing DNA double strand breaks.
      - Undamaged homologous molecules are provided by sister chromatids or paternal/maternal copies of chromosomes.
- The binding of BRCA2 and RAD51 homolog 1 catalyzes the joining of undamaged homologous molecules.



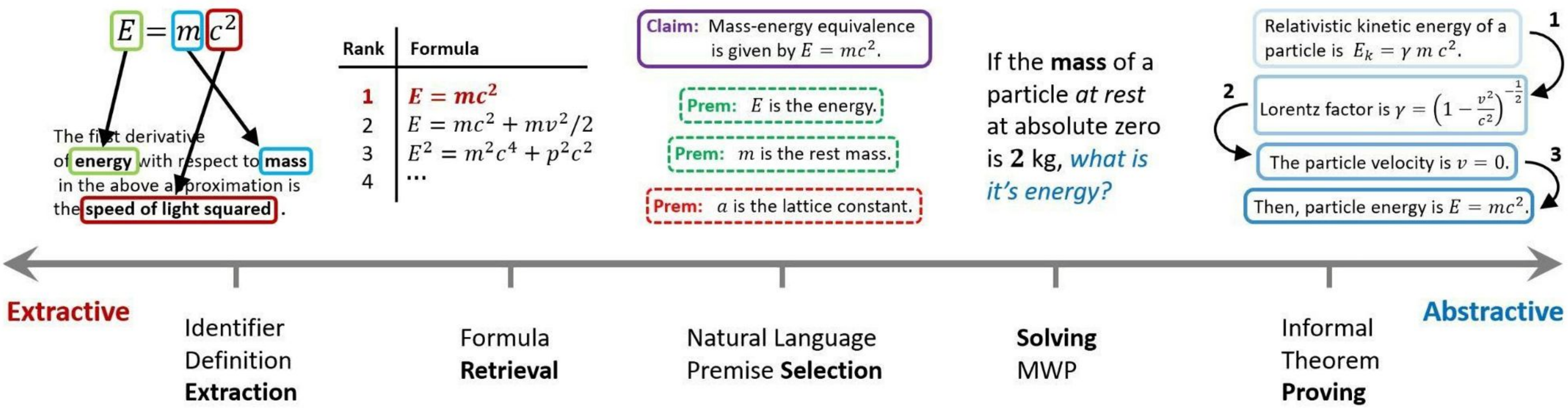
Large fact banks

Aims for Today

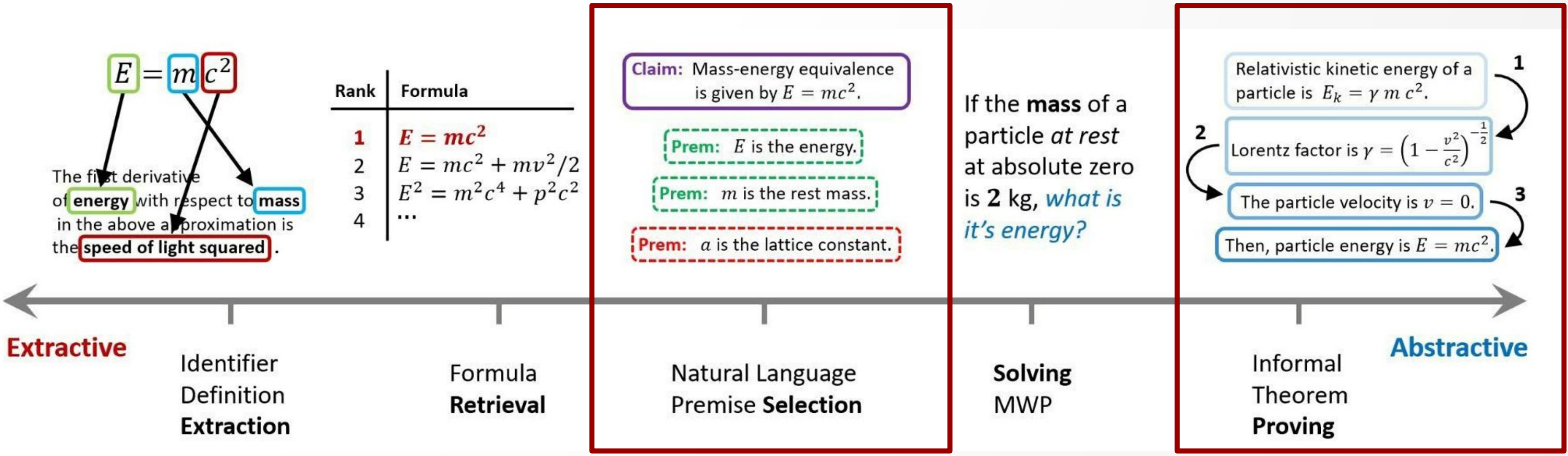
- Selective overview in Mathematical Language Processing (MathLP) - relevant to WG4.

- Emphasis on a particular category of ML model: Large Language Models (LLMs).

- and how to implement semantic and inference controls on the top of this substrate.



Meadows & Freitas, ArXiv: 2205.15231 (2022).



Meadows & Freitas, ArXiv: 2205.15231 (2022).

The Unreasonable Effectiveness of Large Language Models (LLMs)

Language Models

- Probability distributions over strings of text.

The students opened their …
The students opened their books
(predicted)

S = The students opened their books

P(S) = P(The) x P(students | The) x P(opened | The students) x P(their | The students opened) x P(books | The students opened their)

Neural Language Models

output distribution

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{U}\boldsymbol{h} + \boldsymbol{b}_2) \in \mathbb{R}^{|V|}$$

hidden layer

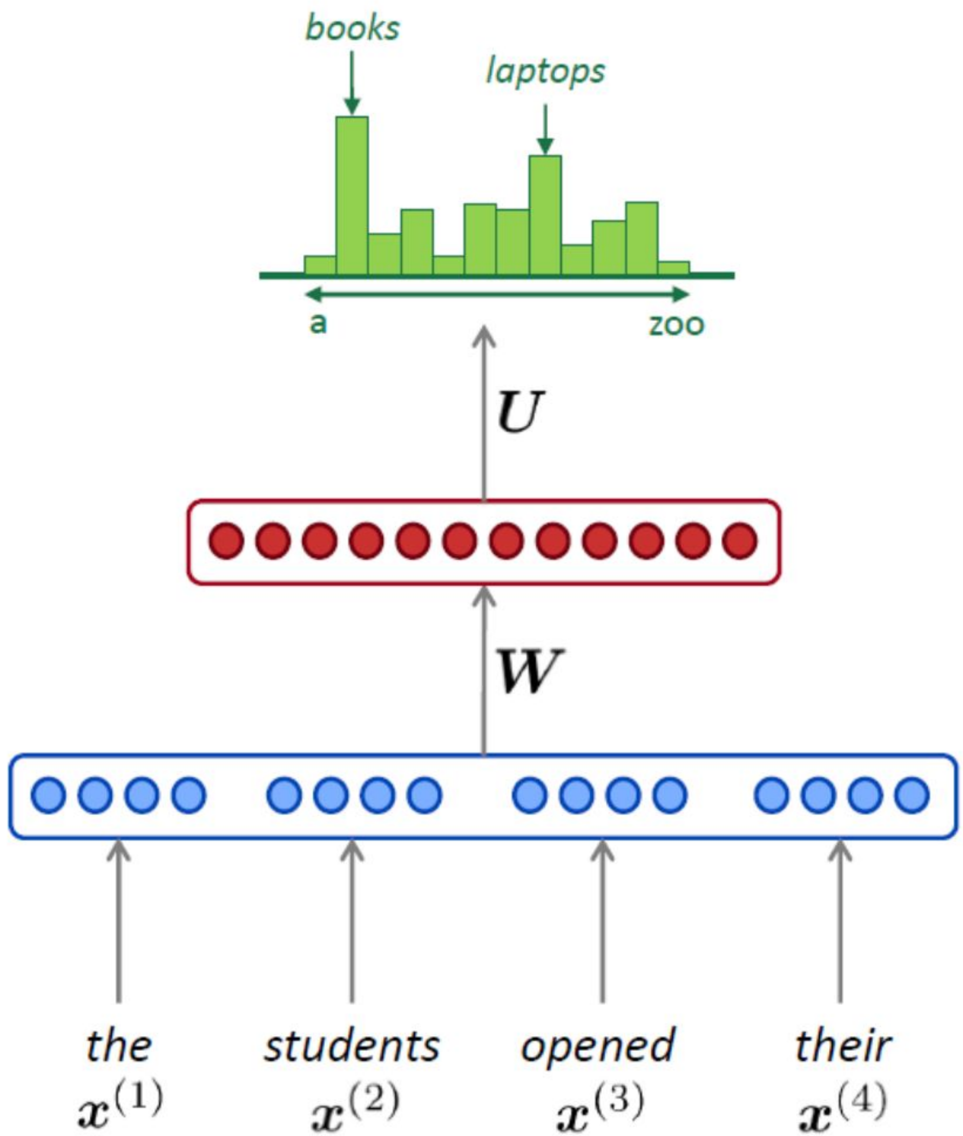
$$\boldsymbol{h} = f(\boldsymbol{W}\boldsymbol{e} + \boldsymbol{b}_1)$$

concatenated word embeddings

$$\boldsymbol{e} = [\boldsymbol{e}^{(1)}; \boldsymbol{e}^{(2)}; \boldsymbol{e}^{(3)}; \boldsymbol{e}^{(4)}]$$

words / one-hot vectors

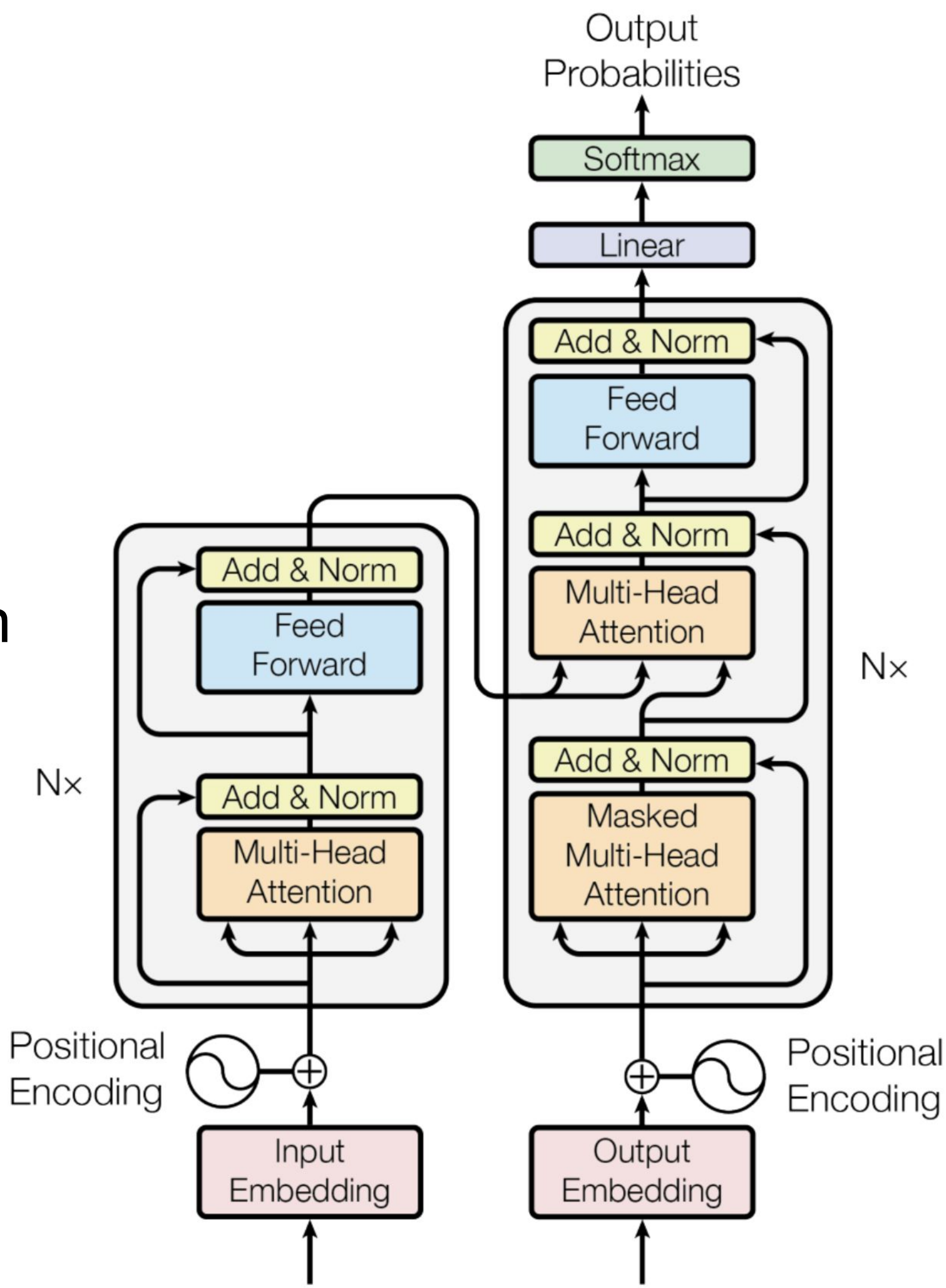
$$\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}$$



Kapronczay, Towards Data Science (2021)

Transformers

1. Positional Encodings
2. (Multi-head) Self-Attention

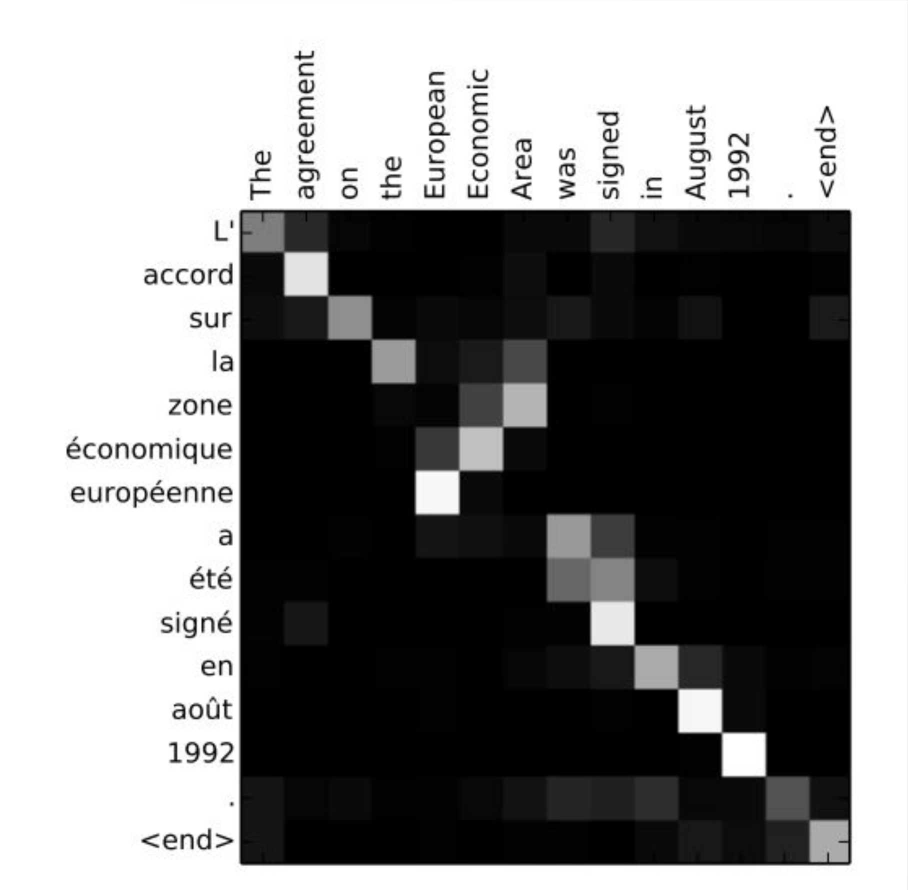


Vaswani et al, NeurIPS (2017)

Attention

The agreement on the European Economic Area was signed in August 1992.

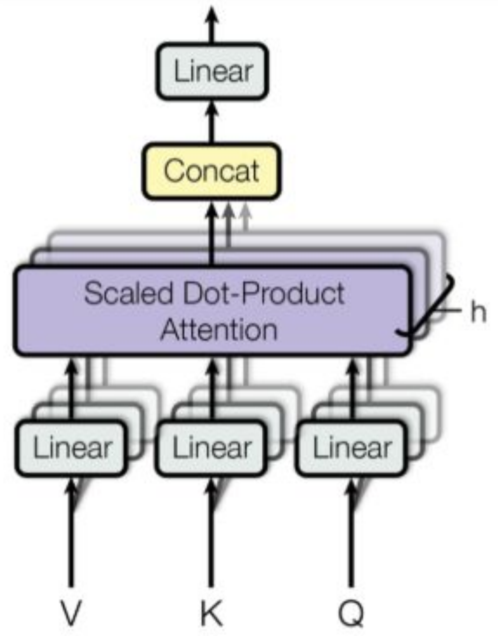
Which words the model should be "attending" to at each time step?



L'accord sur la zone économique européenne a été signé en août 1992.

Vaswani et al, NeurIPS (2017)

Self-Attention

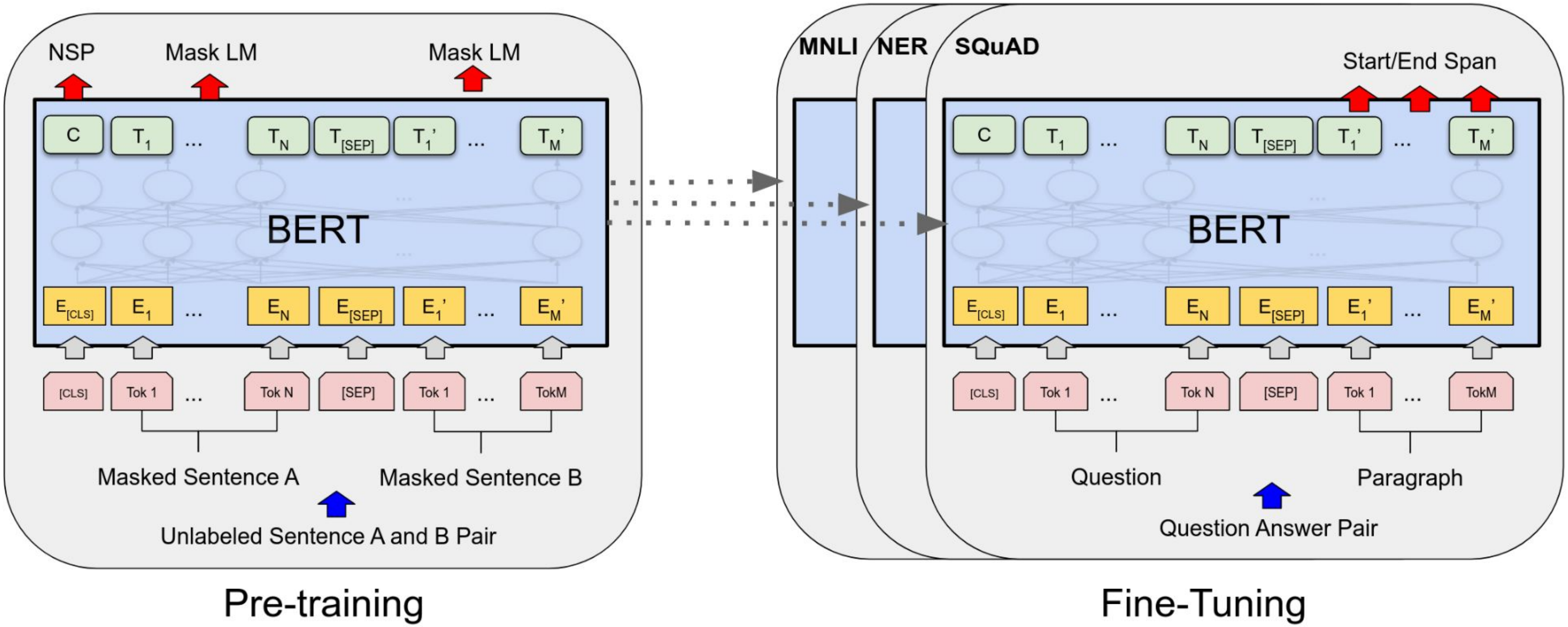


Self-attention allows a a model to assign a meaning to a term in a complex context .

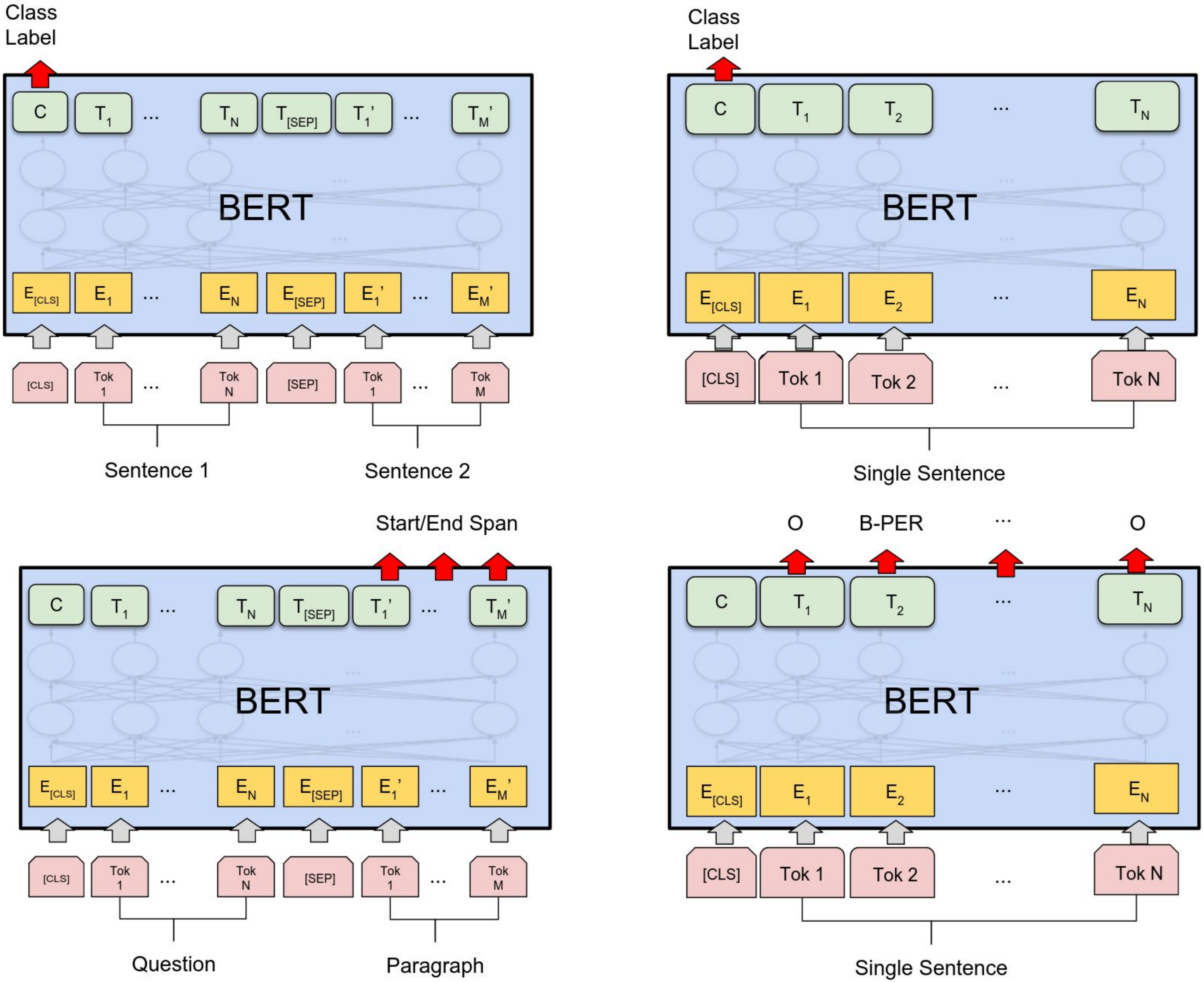
Vaswani et al, NeurIPS (2017)

BERT: Bidirectional Encoder Representations from Transformers

Self-attention allows a a model to assign a meaning to a term in a complex context.



Devlin, Chang, Lee, Toutanova, CoRR (2018)


Class Label
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
Tok M
Sentence 1
Sentence 2
Class Label
C
T_1
T_2
T_N
BERT
$E_{[CLS]}$
E_1
E_2
E_N
[CLS]
Tok 1
Tok 2
Tok N
Single Sentence
Start/End Span
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
Tok M
Question
Paragraph
O
B-PER
...
O
C
T_1
T_2
T_N
BERT
$E_{[CLS]}$
E_1
E_2
E_N
[CLS]
Tok 1
Tok 2
Tok N
Single Sentence

Transformers as Soft Reasoners

(Input Facts:) Alan is blue. Alan is rough. Alan is young.
Bob is big. Bob is round.
Charlie is big. Charlie is blue. Charlie is green.
Dave is green. Dave is rough.

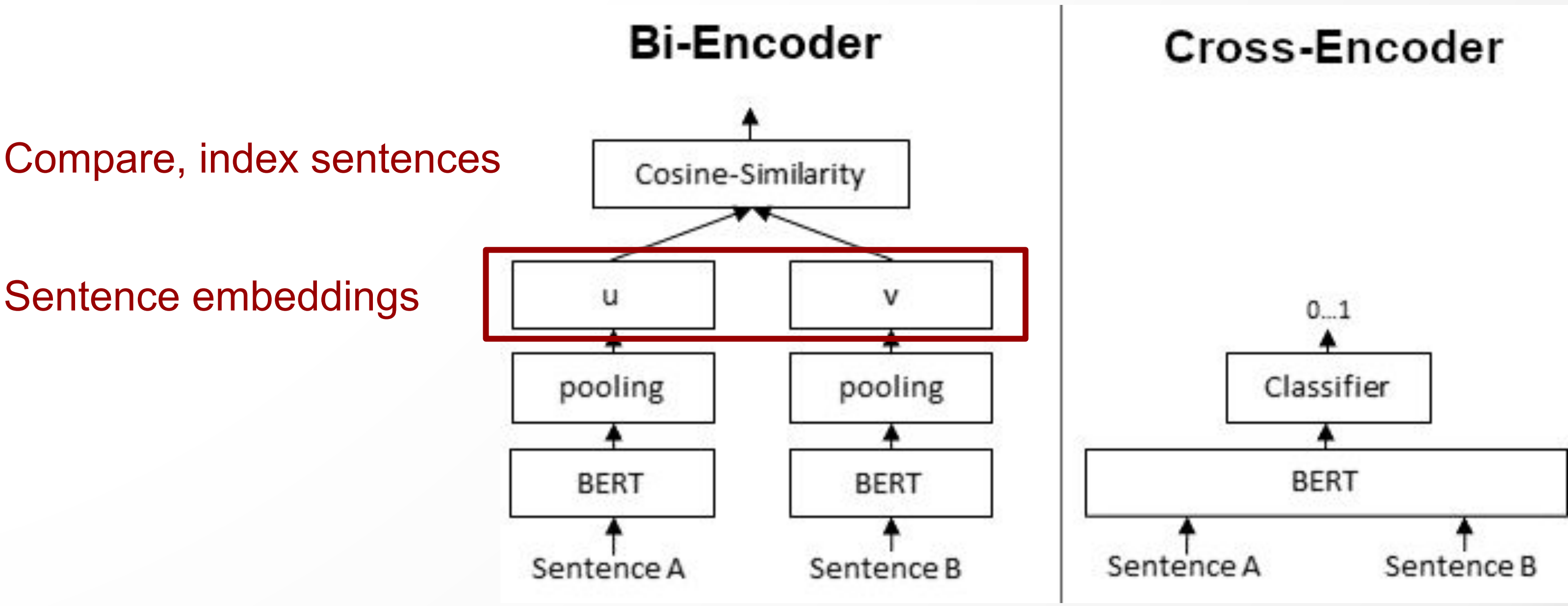
(Input Rules:) Big people are rough.
If someone is young and round then they are kind.
If someone is round and big then they are blue.
All rough people are green.

Q1: Bob is green. True/false? **[Answer: T]**
Q2: Bob is kind. True/false? **[F]**
Q3: Dave is blue. True/false? **[F]**

Clark, Tafjord, Richardson, IJCAI (2020)

SBERT

Cross-encoders: perform full-attention over the input pair.
Bi-encoders: map each input independently to a dense vector space.

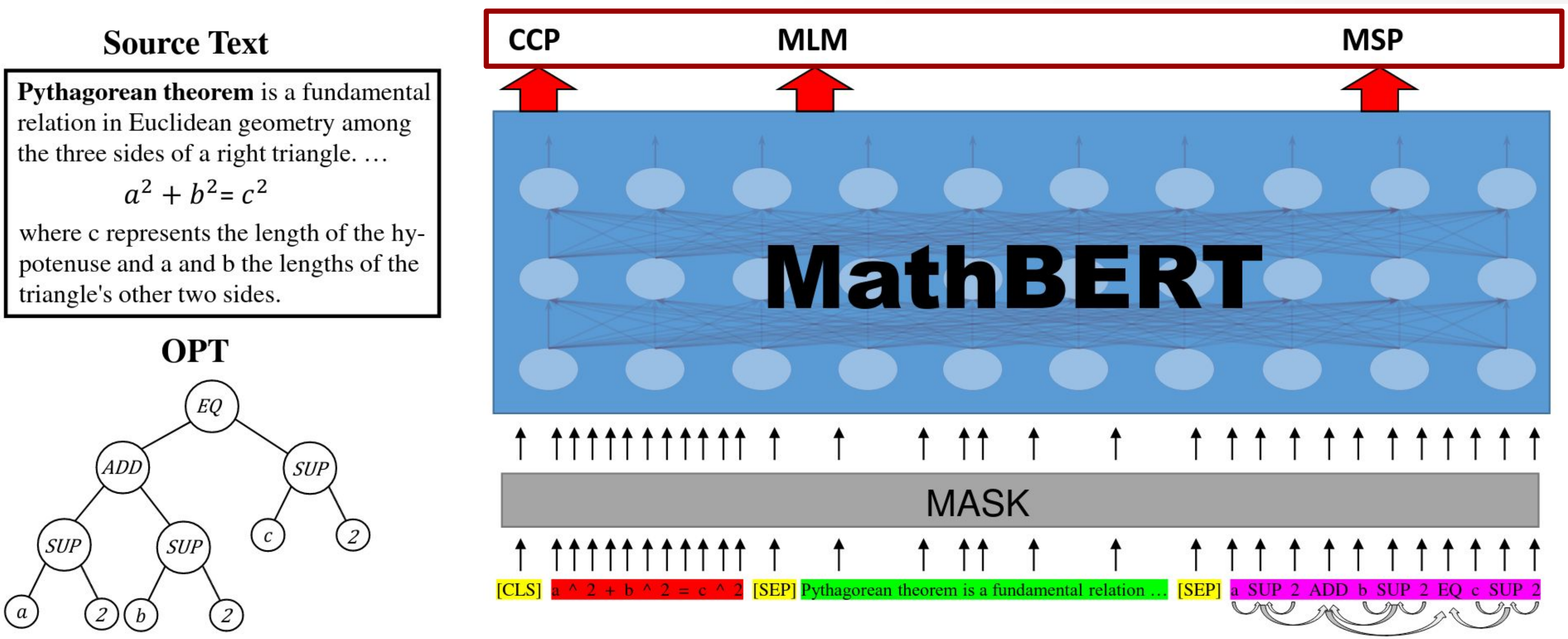


Reimers & Gurevych, EMNLP (2019)

MathBERT

Pre-trained on Arxiv bulk data (Amazon S3)

MLM: Masked Language Modeling
CCP: Context Correspondence Prediction
MSP: Masked Substructure Prediction



Peng, Yuan, Gao, Tang, ArXiv:2105.00377 (2021)

LLMs are few-shot learners

- **'In-context' learning.**
  - Text input of a LLM as a form of task specification.
  - Natural language instruction and
  - a few demonstrations of the task
  - model expected to complete further instances of the task.

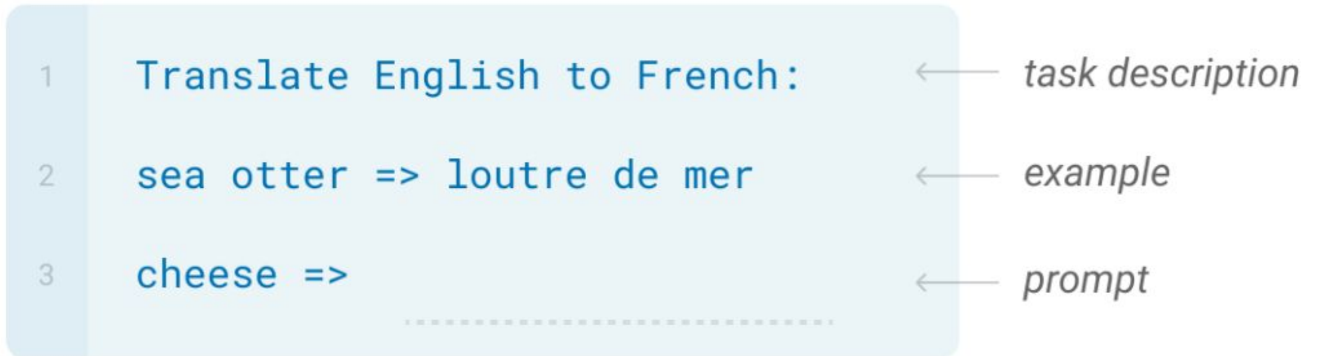
- **Controlling generation.**

  Set an arbitrary prefix (the prompt) as a control mechanism.

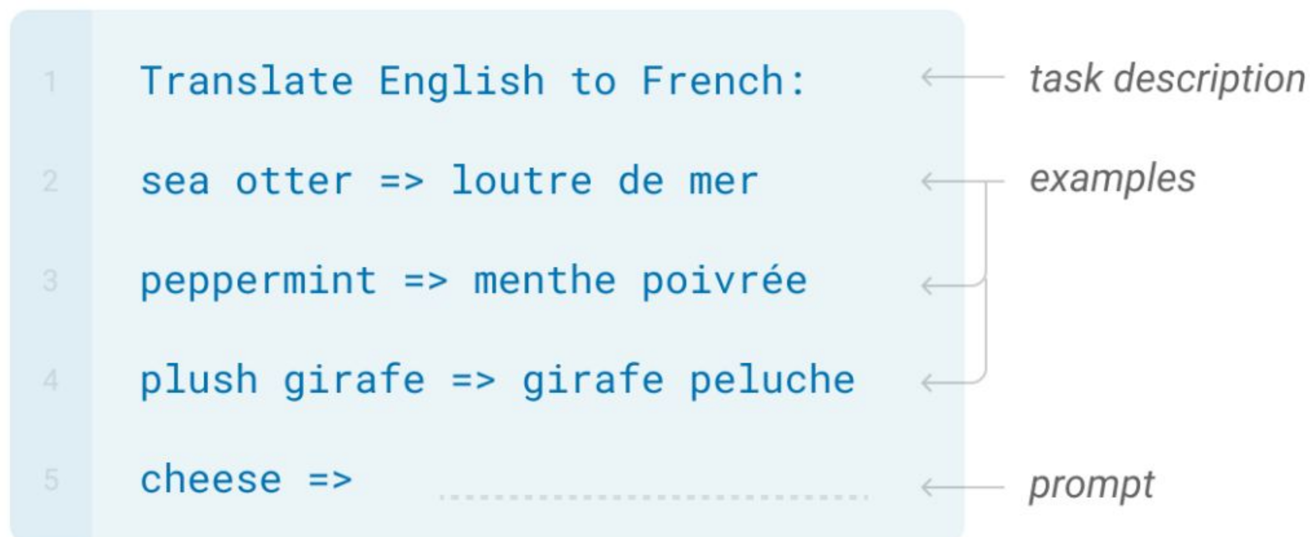
Brown et al., Arxiv:2005.14165 (2020)

Zero-shot

One-shot

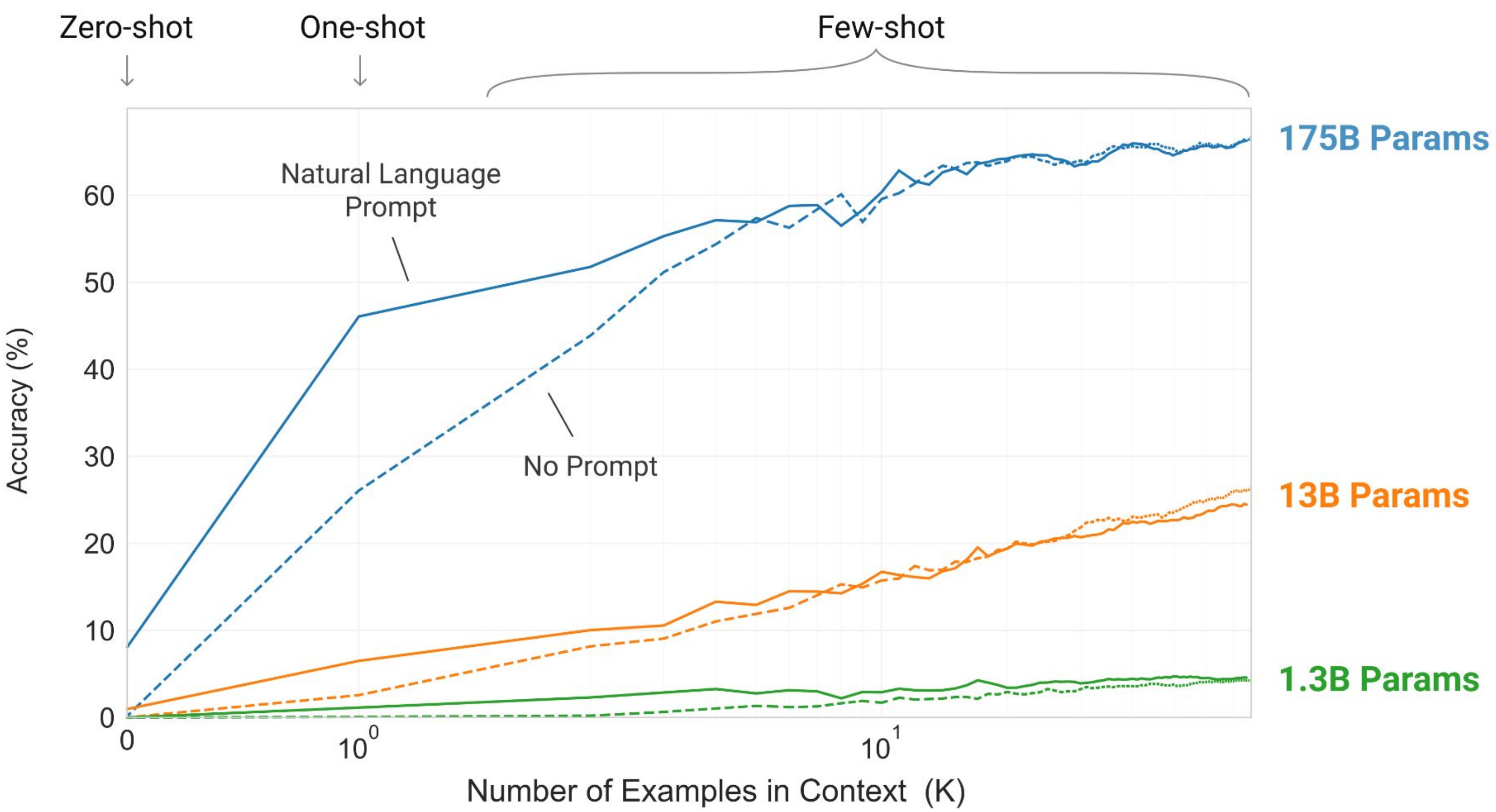


Few-shot



Brown et al., Arxiv:2005.14165 (2020)

LLMs are few-shot learners



Brown et al., Arxiv:2005.14165 (2020)

Autoformalisation

Automatically translating from natural language mathematics to a formal language.

Case Study 1 Question:
"Prove that there is no function f from the set of non-negative integers into itself such that $f(f(n)) = n + 1987$ for every n."

Codex Output:

```
theorem
  fixes f :: "nat \<Rightarrow> nat"
  assumes "\<forall> n. f (f n) = n + 1987"
  shows False
```

Wu et al., Arxiv:2205.12615 (2022)

Informalisation

Automatically translating from natural language mathematics to a formal language.

lemma *seteqI:*
"$\llbracket \bigwedge x.\ x \in A \Longrightarrow x \in B;$
$\bigwedge x.\ x \in B \Longrightarrow x \in A\ \rrbracket$
$\Longrightarrow A = B$"

Translate the Isabelle version to a natural language version:

Codex Output:
"If A and B are sets such that $A \subseteq B$ and $B \subseteq A$, then $A = B$."

lemma *topologyI:*
"$\llbracket\ \bigwedge x\ y.\ \llbracket\ is\_open\ T\ x;\ is\_open\ T\ y \rrbracket \Longrightarrow$
$is\_open\ T\ (x \cap y);$
$\bigwedge M.\ \forall\ m \in M.\ is\_open\ T\ m \Longrightarrow$
$is\_open\ T\ (\bigcup\ M)$
$\rrbracket \Longrightarrow topology\ T$"

Translate the Isabelle version to a natural language version:

Codex Output:
"If T is a set and T is closed under finite intersections and arbitrary unions, then T is a topology."

Wu et al., Arxiv:2205.12615 (2022)

MiniF2F

- MiniF2F dataset containing 488 mathematical competition statements manually formalized.

			Test Set	Validation Set
TOTAL			244	244
IMO			20	20
AIME			15	15
AMC			45	45
MATH	Algebra	Level 5	14	14
		Level 4	14	14
		Level 3	14	14
		Level 2	14	14
		Level 1	14	14
	Number Theory	Level 5	16	16
		Level 4	11	11
		Level 3	11	11
		Level 2	11	11
		Level 1	11	11
CUSTOM	Algebra		18	18
	Number Theory		8	8
	Induction		8	8

Zheng et al., Arxiv:2109.00110 (2021) https://github.com/openai/miniF2F

Autoformalisation

- LLMs can correctly translate 25.3% of mathematical competition problems to formal specifications in Isabelle/HOL.

Formal System	Model	miniF2F-valid Proof Length	miniF2F-valid Pass rate Pass@1	miniF2F-valid Pass rate Pass@8	miniF2F-test Proof Length	miniF2F-test Pass rate Pass@1	miniF2F-test Pass rate Pass@8
Metamath	GPT-f	16.2	1.0%	2.0%	20.3	1.3%	1.6%
Lean	tidy	1.7	16.8%	-	1.8	18.0%	-
Lean	GPT-f	2.6	23.9%	29.3%	2.5	24.6%	29.2%

Zheng et al., Arxiv:2109.00110 (2021) https://github.com/openai/miniF2F

LLMs trained on code

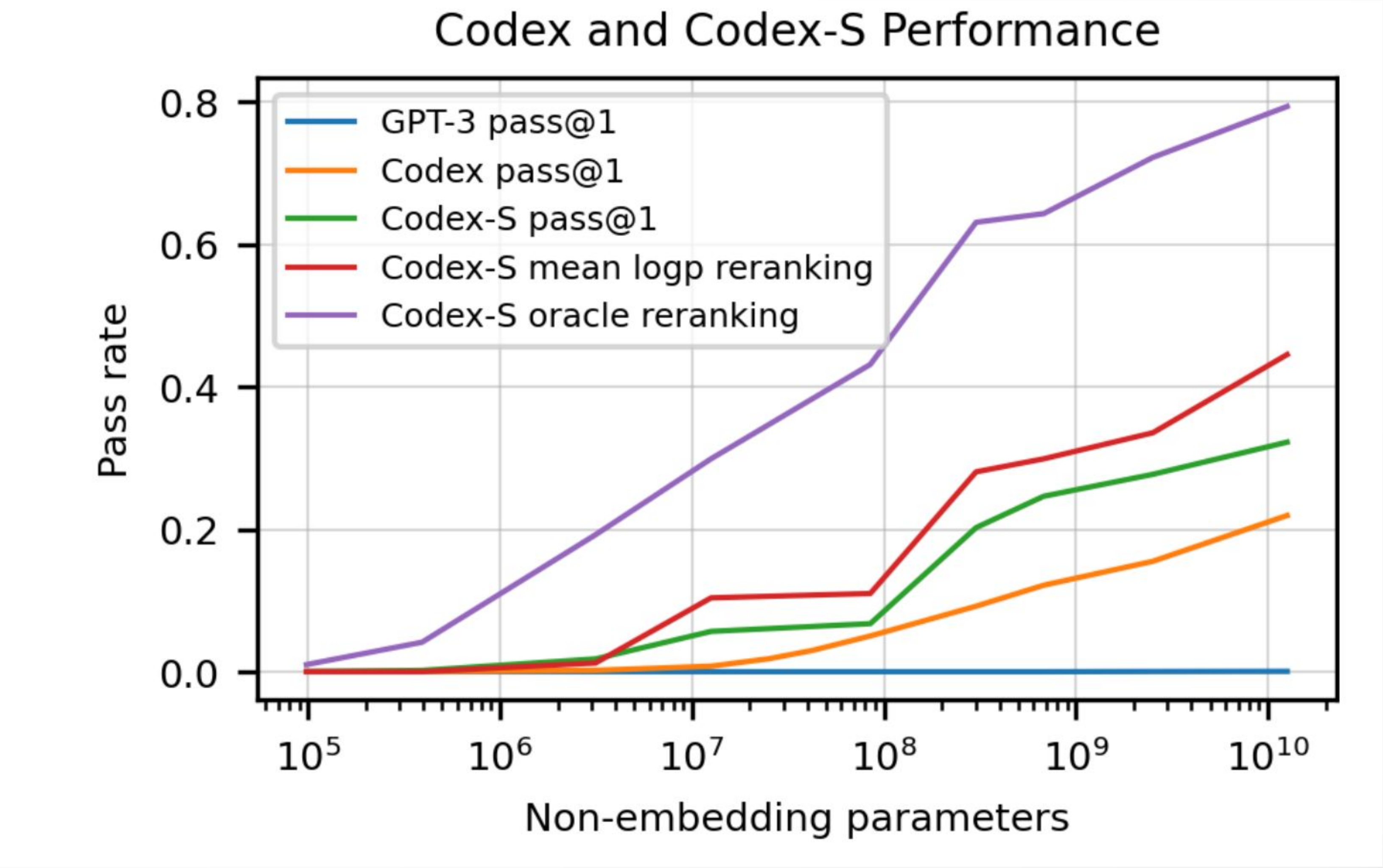
```python
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) =⇒12
    solution([3, 3, 3, 3, 3]) =⇒9
    solution([30, 13, 24, 321]) =⇒0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

Chen et al., Arxiv:2107.03374 (2021)

LLMs trained on code



Chen et al., Arxiv:2107.03374 (2021)

Encoding Inference:

Semantic & Inference Controls

Typing & Discourse-level

$E = mc^2$

The first derivative of **energy** with respect to **mass** in the above approximation is the **speed of light squared**.

We wish to find a function f which satisfies the boundary conditions $f(a) = A$, $f(b) = B$, and which extremizes the functional:

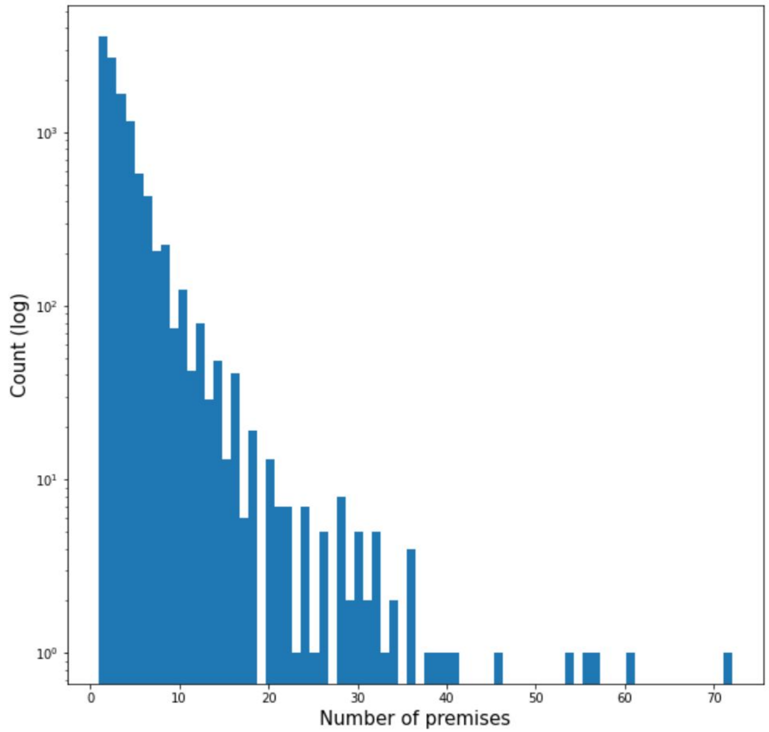
$$J = \int_a^b F(x, f(x), f'(x))\,\mathrm{d}x\,.$$

Work	Task	Learning	Approach	Dataset
Identifier-Definition Extraction				
Kristianto et al. (2012)	Expression-definition	S	CRF with linguistic pattern features	LaTeX papers
Kristianto et al. (2014a)	Expression-definition	S	SVM with linguistic pattern features	LaTeX papers
Pagael and Schubotz (2014)	Identifier-definition	R	Gaussian heuristic ranking	Wikipedia articles
Schubotz et al. (2016a)	Identifier-definition	UNS	Gaussian ranking + K-means namespace clusters	NTCIR-11 Math Wikipedia
Schubotz et al. (2017)	Identifier-definition	S	G. rank + pattern matching + SVM	NTCIR-11 Math Wikipedia
Stathopoulos et al. (2018)	Variable Typing	S	Link prediction with BiLSTM	arXiv papers
Alexeeva et al. (2020)	Identifier-definition	R	Odin grammar	MathAlign-Eval
Jo et al. (2021)	Notation auto-suggestion and consistency checking	S	BERT fine-tuning	S2ORC

Informal (NL) Premise Selection

Conjecture	Premise	Predicted	Label
Let $T = (S, \tau)$ be a topological space. Let A, B be subsets of S. Then: $\partial(A \cap B) \subseteq \partial A \cup \partial B$ where ∂A denotes the boundary of A.	Let S, T_1, T_2 be sets such that T_1, T_2 are both subsets of S. Then, using the notation of the relative complement: $ST_1 \cap T_2 = ST_1 \cup ST_2$	1	1
$\int \frac{x}{x(x^2-a^2)} = \frac{1}{2a^2}, \ln \frac{x^2-a^2}{x^2} + C$ for $x^2 > a^2$.	$\int \frac{dx}{x} = \ln x + C$ for $x \neq 0$.	1	1
Let $T = S, \tau$ be a compact space. Then T is countably compact.	Let $T = (S, \tau_{a,b})$ be a modified Fort space. Then T is not a T_3 space, T_4 space or T_5 space.	1	0

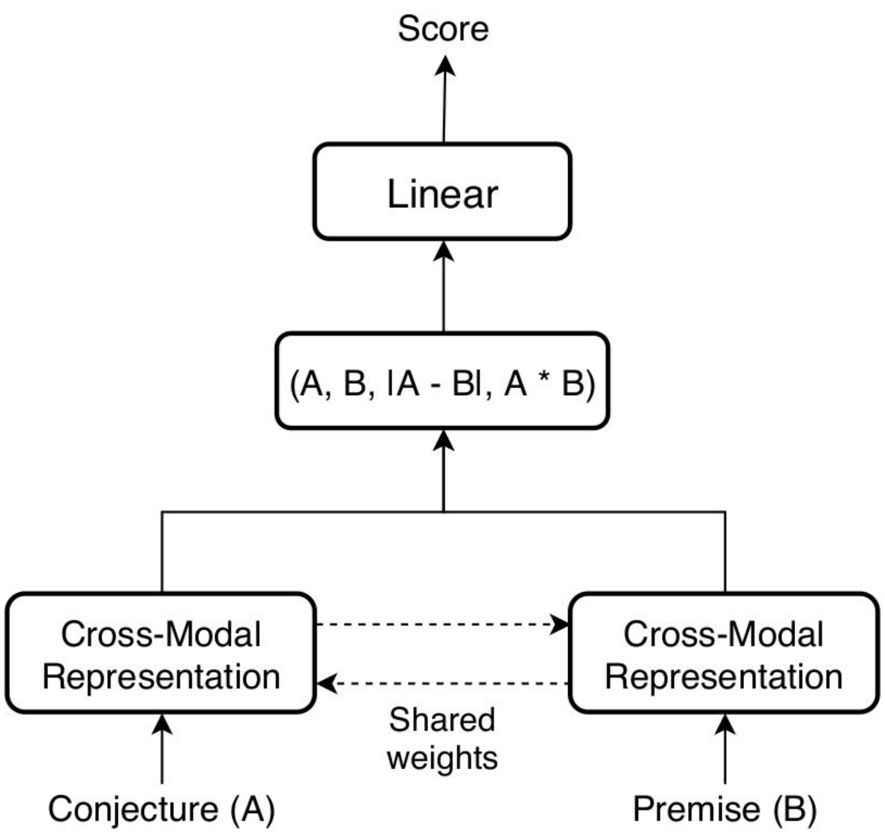
Statement type	Data Split				
	KB	Train	Dev	Test	All (Unique)
Definitions	7,077	0	0	0	7,077
Lemmas	252	134	70	69	252
Corollaries	161	113	57	57	275
Theorems	8,715	5,272	2,652	2,636	14,003
Total	16,205	5,519	2,778	2,763	21,746

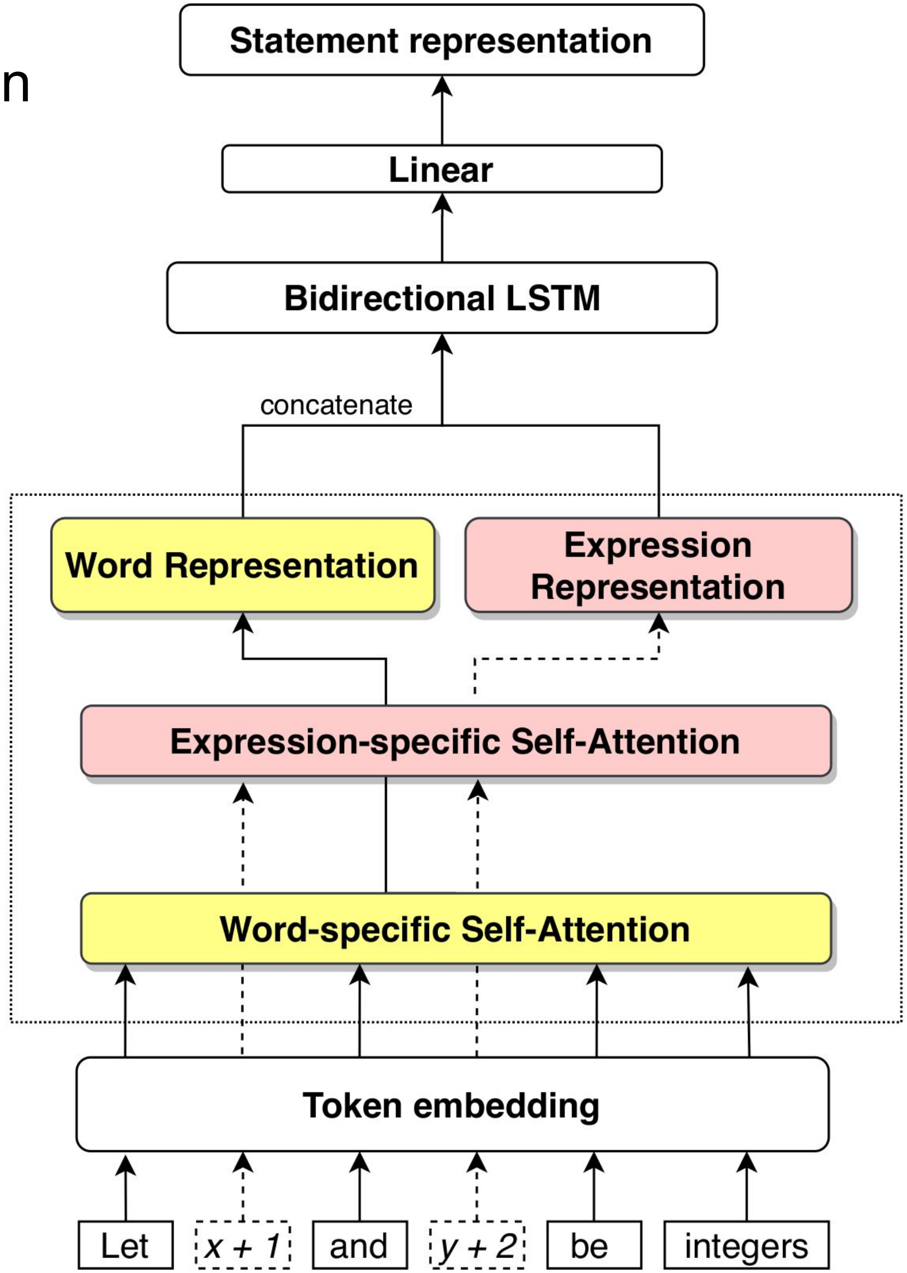


Ferreira & Freitas, LREC (2020)

Informal (NL) Premise Selection

Cross-model statement representation (STAR)





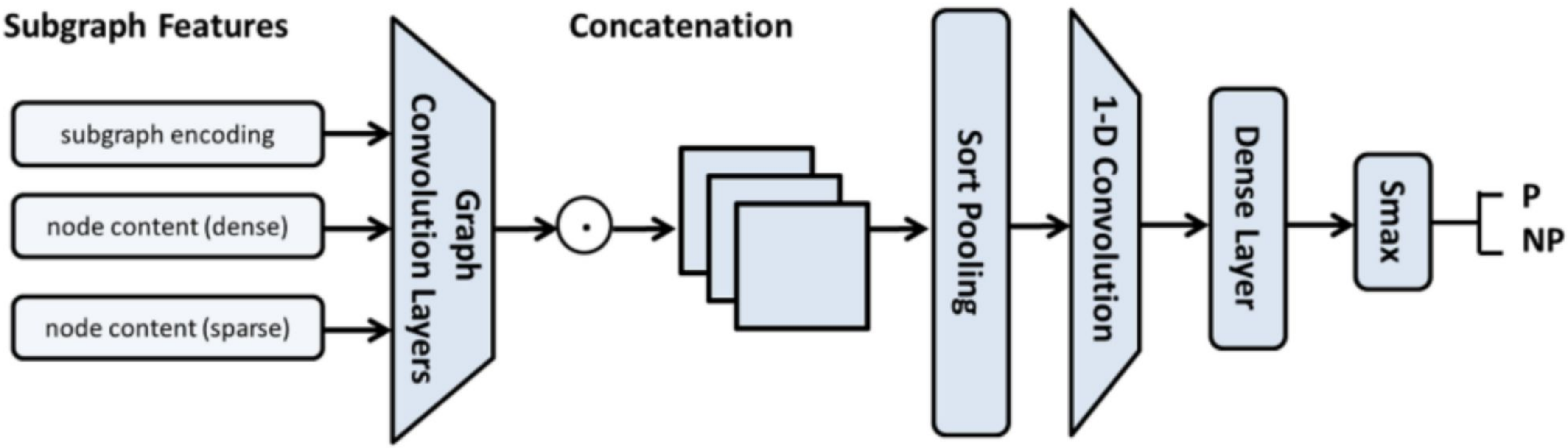
Ferreira & Freitas, EACL (2021)

Informal (NL) Premise Selection

	Val			Test		
	F1	**P**	**R**	**F1**	**P**	**R**
BERT	**.886**	.871	.901	.877	.925	.834
MathSum	.644	.512	.869	.459	.562	.388
Self-attention + BiLSTM	.651	.550	.796	.631	.573	.703
STAR	.885	.854	.917	**.882**	.865	.899

Ferreira & Freitas, EACL (2021)

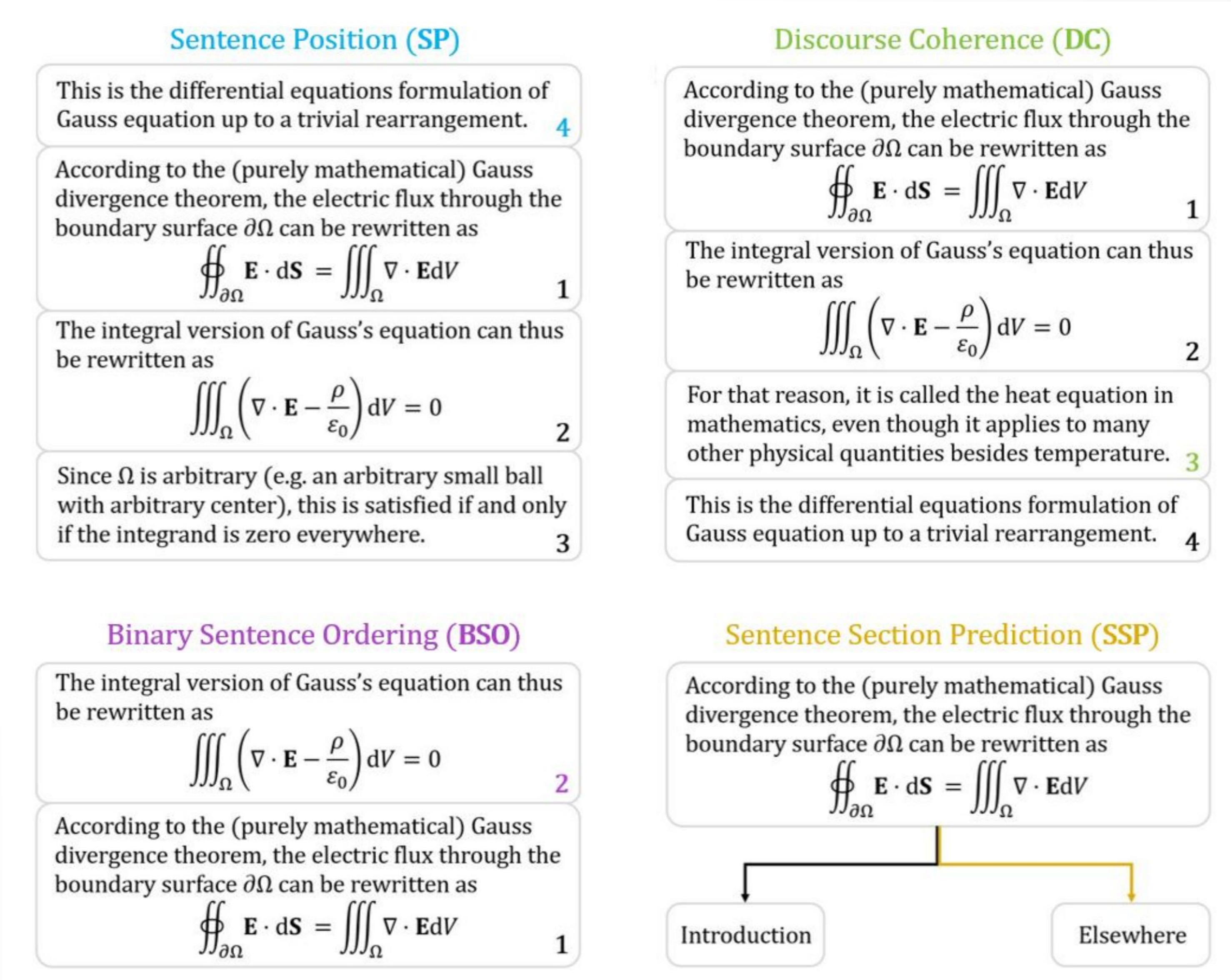
Informal (NL) Premise Selection



	BERT			Proposed Model		
	P	R	F1	P	R	F1
2-hop	47.5	**78.9**	59.3	54.8	**68.7**	61.0 (+ 3%)
3-hop	41.0	45.1	**49.2**	58.8	**63.3**	61.2 (+ 24%)

Ferreira & Freitas, ACL (2020)

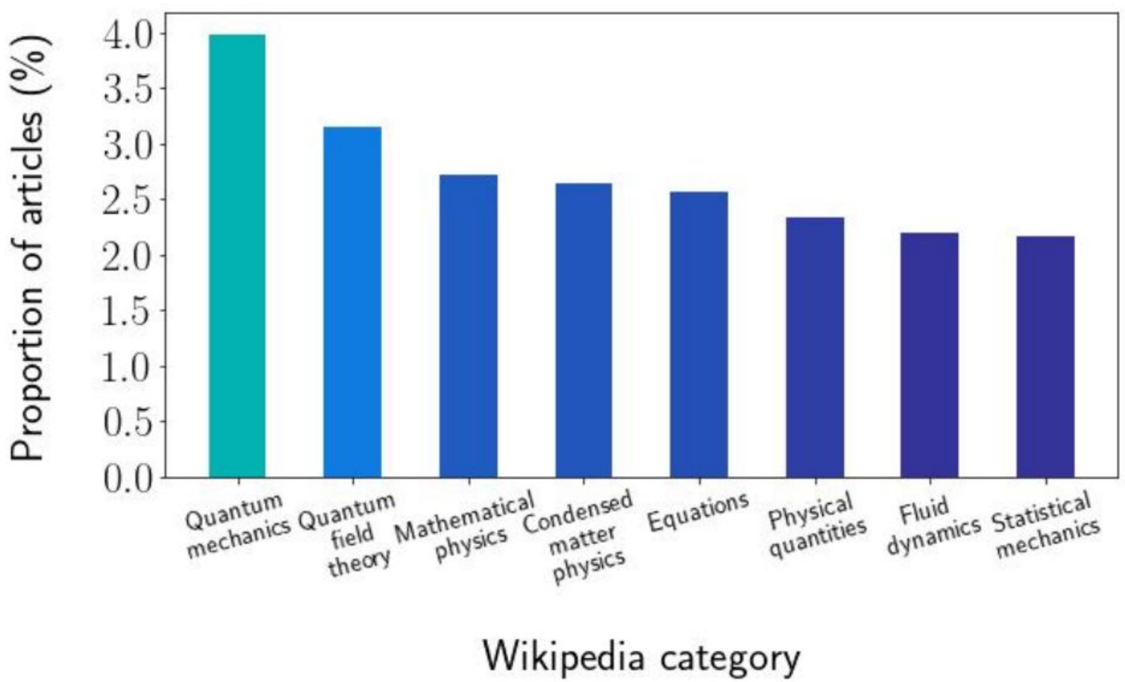
Discourse-level



Meadows, Zhou, Freitas, LREC 2022.

Discourse-level

Dataset	Size	% with math	% with equations
DC	35 k	45	35
SP	40 k	36	29
BSO	459 k	24	17
SSP	90 k	12	7



Meadows, Zhou, Freitas, LREC 2022.

Symbolic Gap

We can repeat this for momentum by interpreting the function $\boxed{\tilde{g}(p) = p \cdot \varphi(p)}$ ~~as a vector, but we can also take advantage of the fact~~ that $\psi(x)$ and $\varphi(p)$ are Fourier transforms of each other. We evaluate the inverse Fourier transform through integration by parts:

$$\boxed{g(x) = \frac{1}{\sqrt{2\pi\hbar}} \cdot \int_{-\infty}^{\infty} \boxed{\tilde{g}(p)} \cdot e^{ipx/\hbar}\, dp}$$

$$= \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \boxed{p \cdot \varphi(p)} \cdot e^{ipx/\hbar}\, dp$$

1

$$\boxed{= \frac{1}{2\pi\hbar} \int_{-\infty}^{\infty} \left[p \cdot \int_{-\infty}^{\infty} \psi(\chi) e^{-ip\chi/\hbar}\, d\chi \right] \cdot e^{ipx/\hbar}\, dp}$$

$$= \frac{i}{2\pi} \int_{-\infty}^{\infty} \left[\cancel{\psi(\chi) e^{-ip\chi/\hbar} \Big|_{-\infty}^{\infty}} - \int_{-\infty}^{\infty} \frac{d\psi(\chi)}{d\chi} e^{-ip\chi/\hbar}\, d\chi \right] \cdot$$

$$= \frac{-i}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{d\psi(\chi)}{d\chi} e^{-ip\chi/\hbar}\, d\chi\, e^{ipx/\hbar}\, dp$$

$$= \left(-i\hbar \frac{d}{dx} \right) \cdot \psi(x),$$

$$\boxed{g(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \tilde{g}(p) \cdot e^{\frac{ipx}{\hbar}}\, dp}$$

$$\boxed{\tilde{g}(p) = p \cdot \varphi(p)}$$

$$g(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} p \cdot \varphi(p) \cdot e^{\frac{ipx}{\hbar}}\, dp$$

$$\varphi(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \varphi(\chi) \cdot e^{\frac{-ip\chi}{\hbar}}\, d\chi$$

$$g(x) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} p \cdot \left(\frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \varphi(\chi) \cdot e^{\frac{-ip\chi}{\hbar}}\, d\chi \right) \cdot e^{\frac{ipx}{\hbar}}\, dp$$

$$\boxed{g(x) = \frac{1}{2\pi\hbar} \int_{-\infty}^{\infty} p \cdot \left(\int_{-\infty}^{\infty} \varphi(\chi) \cdot e^{\frac{-ip\chi}{\hbar}}\, d\chi \right) \cdot e^{\frac{ipx}{\hbar}}\, dp}$$

Proof, Explanation & Natural Language Inference

H: <u>Shale</u> is a <u>sedimentary rock</u> that can be metamorphosed into <u>slate</u> by <u>increased pressure</u>.

'<u>shale</u> is a kind of <u>sedimentary rock</u>'

'<u>high</u> is similar to <u>increase</u>'

'<u>extreme</u> means very <u>high</u> in value'

'<u>slate</u> is a type of <u>metamorphic rock</u>'

'exposure to <u>extreme</u> heat and <u>pressure</u> changes <u>sedimentary</u> and igneous <u>rock</u> into <u>metamorphic rock</u>'

<u>Abstraction, grounding</u>

<u>Abstraction</u>

Proof, Explanation & Natural Language Inference

H: <u>Shale</u> is a <u>sedimentary rock</u> that can be metamorphosed into <u>slate</u> by <u>increased pressure</u>.

'<u>shale</u> is a kind of <u>sedimentary rock</u>'

'<u>high</u> is similar to <u>increase</u>'

'<u>extreme</u> means very <u>high</u> in value'

'<u>slate</u> is a type of <u>metamorphic rock</u>'

'exposure to <u>extreme</u> heat and <u>pressure</u> changes <u>sedimentary</u> and igneous <u>rock</u> into <u>metamorphic rock</u>'

<u>Unification</u>

<u>Abstraction</u>

Controlling NLI

Sentence embeddings for approximate premise selection (kNN query - scalable).

Add constraints which define an explanation.

Constructs a fact graph where each node is a fact with explicit attributes.

Define properties which we can optimise: e.g. **relevance**, **saturation** and **diversity**.

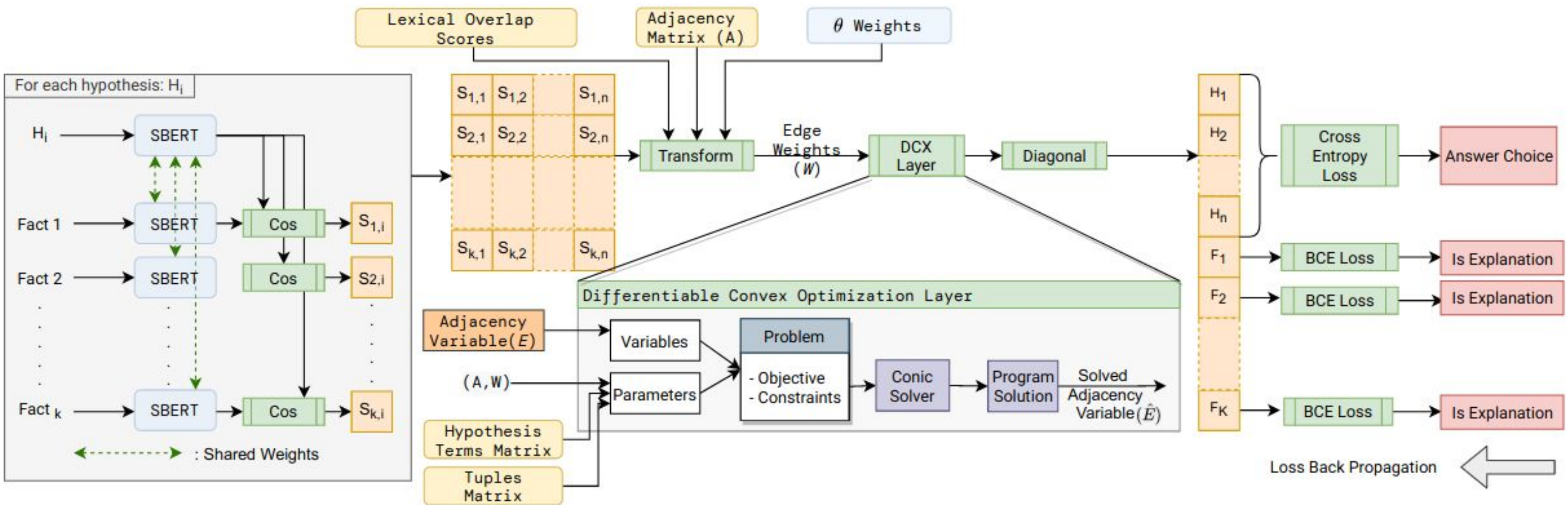
Thayaparan et al, TACL (2022)

Valentino, Thayaparan, Ferreira, Freitas, AAAI (2022)

Valentino, Thayaparan, Freitas, EACL (2021)

Thayaparan & Freitas, ACL Findings (2021)

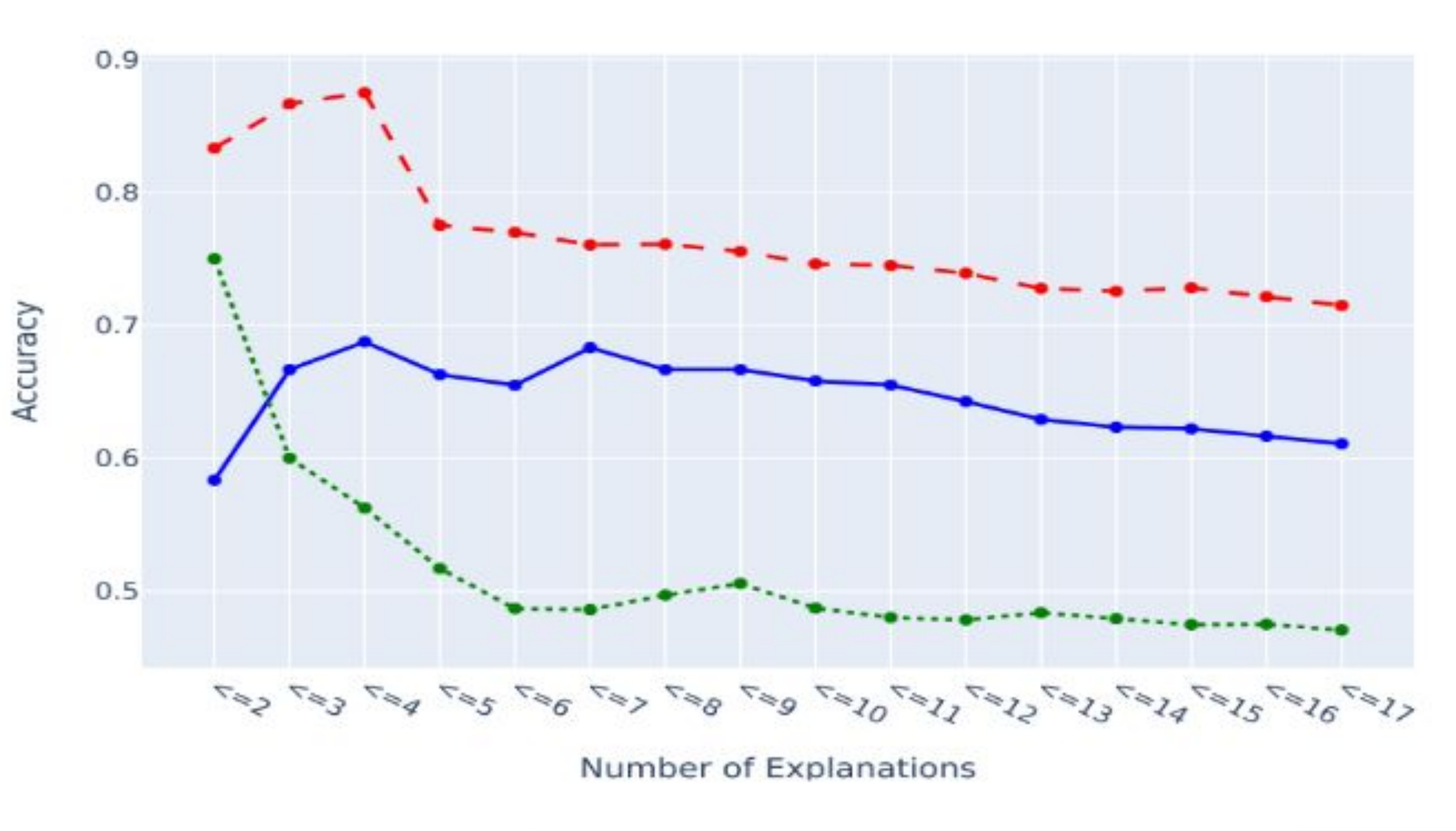
Controlling NLI



An end-to-end differentiable framework that incorporates constraints via convex optimization layers into broader transformers-based architectures.

Semantic and lexical scores are weighted by a set of learnable θ parameters to construct an explanation graph G = (V, E) supporting the candidate answer.

Thayaparan et al, TACL (2022)



red: ExplanationLP + UR
blue: $BERT_{Large}$ + UR
green: PathNet + UR

Thayaparan & Freitas, ACL Findings (2021)

Conclusions

- LLMs have demonstrated the capability of synthesising code from NL in a few-shot setting.

- NLI have been complementing LLMs models with additional semantic and inference controls.

- Nothing specific here for NL: applicable to other types of language.

- Strategic (cross-disciplinary) space for WG4:
  - What are the efficiency gains of LLMs and NLI in the construction of proof libraries?

- Because this group is closer to the resources (libraries), I believe we are at a unique position to answer this question.

Questions, Collaborations?

andre.freitas@manchester.ac.uk