# Testing the limits of logical reasoning in neural and hybrid models

## Compositionality and Reasoning in AI and Cognitive Science Workshop

**Manuel Vargas Guzmán**
University of Warsaw
m.vargas-guzman@uw.edu.pl

**Jakub Szymanik**
University of Trento
jakub.szymanik@gmail.com

**Maciej Malicki**
University of Warsaw
mmalicki@mimuw.edu.pl

Warsaw, January 8, 2026

# Our contributions

- We present a framework, inspired by compositional tests (Hupkes et al., 2019), for systematically evaluating generalization in logical reasoning over natural language fragments (Pratt-Hartmann, 2004).

# Our contributions

- We present a framework, inspired by compositional tests (Hupkes et al., 2019), for systematically evaluating generalization in logical reasoning over natural language fragments (Pratt-Hartmann, 2004).
- Using syllogistic logic, one of the smallest logical fragment, we evaluate neural assistance in symbolic proof construction and identify significant limitations in generalization.

# Our contributions

- We present a framework, inspired by compositional tests (Hupkes et al., 2019), for systematically evaluating generalization in logical reasoning over natural language fragments (Pratt-Hartmann, 2004).
- Using syllogistic logic, one of the smallest logical fragment, we evaluate neural assistance in symbolic proof construction and identify significant limitations in generalization.
- We present a neuro-symbolic syllogistic prover that uses neural guidance for proof construction, achieving efficient symbolic search and robust, interpretable reasoning despite limited neural generalization.

# Syllogistic logic

Well-formed formulas

| | | |
|---|---|---|
| $Aab$ | $a \subseteq b$ | All $a$ are $b$ |
| $Eab$ | $a \cap b = \emptyset$ | No $a$ are $b$ |
| $Iab$ | $a \cap b \neq \emptyset$ | Some $a$ are $b$ |
| $Oab$ | $a \not\subseteq b$ | Some $a$ are not $b$ |

# Syllogistic logic

Well-formed formulas

| | | |
|---|---|---|
| $Aab$ | $a \subseteq b$ | All $a$ are $b$ |
| $Eab$ | $a \cap b = \emptyset$ | No $a$ are $b$ |
| $Iab$ | $a \cap b \neq \emptyset$ | Some $a$ are $b$ |
| $Oab$ | $a \not\subseteq b$ | Some $a$ are not $b$ |

- An *A-chain* ($Aa - b$) represents either a formula $Aab$ or the sequence $Aac_1, Ac_1c_2, \ldots, Ac_{n-1}c_n, Ac_nb$ (for $n \geq 1$)

# Syllogistic logic

| Well-formed formulas | | |
|---|---|---|
| $Aab$ | $a \subseteq b$ | All $a$ are $b$ |
| $Eab$ | $a \cap b = \emptyset$ | No $a$ are $b$ |
| $Iab$ | $a \cap b \neq \emptyset$ | Some $a$ are $b$ |
| $Oab$ | $a \not\subseteq b$ | Some $a$ are not $b$ |

- An *A-chain* ($Aa - b$) represents either a formula $Aab$ or the sequence $Aac_1, Ac_1c_2, \ldots, Ac_{n-1}c_n, Ac_nb$ (for $n \geq 1$)

- The **negation** of a formula $F$ is denoted as $\overline{F}$

$$\begin{array}{c|c} \overline{Aab} = Oab & \overline{Iab} = Eab \\ \overline{Oab} = Aab & \overline{Eab} = Iab \end{array}$$

# Knowledge Base

### Definition

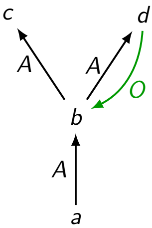
A knowledge base ($\mathcal{KB}$) is a finite set of syllogistic formulas

# Knowledge Base

## Definition

A knowledge base ($\mathcal{KB}$) is a finite set of syllogistic formulas

**Example:** graph representation.



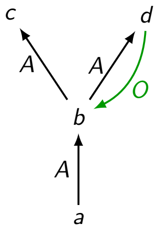

$$\mathcal{KB} = \{Aab, Abc, Abd, Odb\}$$

# Knowledge Base

## Definition

A knowledge base ($\mathcal{KB}$) is a finite set of syllogistic formulas

**Example:** graph representation.




$$\mathcal{KB} = \{Aab, Abc, Abd, Odb\}$$

**Remark:** we generate $\mathcal{KB}s$ that are consistent (no contradictions) and non-redundant, in the sense that each hypothesis admits only one minimal **proof**.

# Derivation rules (Smiley, 1973)

## Definition

A *derivation* $\nabla$ is one of the following three types:

(i) Every $F \in \mathcal{KB}$ is a derivation from $\mathcal{KB}$

$$\frac{}{F}\ \text{(i)}$$

# Derivation rules (Smiley, 1973)

## Definition

A *derivation* $\nabla$ is one of the following three types:

(i) Every $F \in \mathcal{KB}$ is a derivation from $\mathcal{KB}$

$$\frac{}{F}\,(\mathrm{i})$$

(ii) The following four trees are derivations from $\mathcal{KB}$. Where $\nabla'$ and $\nabla''$ are derivations from $\mathcal{KB}$

$$\frac{\overset{\nabla'}{Aab} \qquad \overset{\nabla''}{Abc}}{Aac}\,(\mathrm{r1}) \qquad \frac{\overset{\nabla'}{Aab} \qquad \overset{\nabla''}{Ebc}}{Eac}\,(\mathrm{r2}) \qquad \frac{\overset{\nabla'}{Eba}}{Eab}\,(\mathrm{r3}) \qquad \frac{\overset{\nabla'}{Aba}}{Iab}\,(\mathrm{r4})$$

# Derivation rules (Smiley, 1973)

## Definition

A *derivation* $\nabla$ is one of the following three types:

(i) Every $F \in \mathcal{KB}$ is a derivation from $\mathcal{KB}$

$$\frac{}{F}\,\text{(i)}$$

(ii) The following four trees are derivations from $\mathcal{KB}$. Where $\nabla'$ and $\nabla''$ are derivations from $\mathcal{KB}$

$$\frac{\begin{array}{c}\nabla'\\ Aab\end{array} \qquad \begin{array}{c}\nabla''\\ Abc\end{array}}{Aac}\,\text{(r1)} \qquad \frac{\begin{array}{c}\nabla'\\ Aab\end{array} \qquad \begin{array}{c}\nabla''\\ Ebc\end{array}}{Eac}\,\text{(r2)} \qquad \frac{\begin{array}{c}\nabla'\\ Eba\end{array}}{Eab}\,\text{(r3)} \qquad \frac{\begin{array}{c}\nabla'\\ Aba\end{array}}{Iab}\,\text{(r4)}$$

(iii) *Proof by contradiction*: where $\nabla'$ is a derivation from $\mathcal{KB} \cup \{\overline{H}\}$ and $\nabla''$ is a derivation from $\mathcal{KB}$.

$$\frac{\begin{array}{c}\nabla'\\ F\end{array} \qquad \begin{array}{c}\nabla''\\ \overline{F}\end{array}}{H}\,\text{(iii)}$$

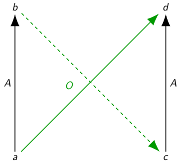
# Types of syllogisms

(1) $\{Aa-b, Ac-d, Oad\} \vdash Obc$
(2) $\{Aa-b\} \vdash Aab$
(3) $\{Aa-b, Ac-d, Aa-e, Ede\} \vdash Obc$
(4) $\{Aa-b, Aa-c\} \vdash Ibc$
(5) $\{Aa-b, Ac-d, Ae-f, Iae, Edf\} \vdash Obc$
(6) $\{Aa-b, Ac-d, Ebd\} \vdash Eac$
(7) $\{Aa-b, Ac-d, Iac\} \vdash Ibd$

# Types of syllogisms

| |
|---|
| (1) $\{Aa-b, Ac-d, Oad\} \vdash Obc$ |
| (2) $\{Aa-b\} \vdash Aab$ |
| (3) $\{Aa-b, Ac-d, Aa-e, Ede\} \vdash Obc$ |
| (4) $\{Aa-b, Aa-c\} \vdash Ibc$ |
| (5) $\{Aa-b, Ac-d, Ae-f, Iae, Edf\} \vdash Obc$ |
| (6) $\{Aa-b, Ac-d, Ebd\} \vdash Eac$ |
| (7) $\{Aa-b, Ac-d, Iac\} \vdash Ibd$ |

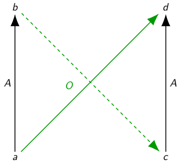
Type (1)
$\{Aa-b, Ac-d, Oad\} \vdash Obc$



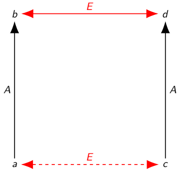
# Types of syllogisms

$$
\begin{array}{|l|}
\hline
(1)\ \{Aa-b, Ac-d, Oad\} \vdash Obc \\
(2)\ \{Aa-b\} \vdash Aab \\
(3)\ \{Aa-b, Ac-d, Aa-e, Ede\} \vdash Obc \\
(4)\ \{Aa-b, Aa-c\} \vdash Ibc \\
(5)\ \{Aa-b, Ac-d, Ae-f, Iae, Edf\} \vdash Obc \\
(6)\ \{Aa-b, Ac-d, Ebd\} \vdash Eac \\
(7)\ \{Aa-b, Ac-d, Iac\} \vdash Ibd \\
\hline
\end{array}
$$

Type (1)
$\{Aa-b, Ac-d, Oad\} \vdash Obc$



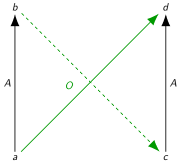

Type (6)
$\{Aa-b, Ac-d, Ebd\} \vdash Eac$



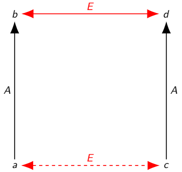
# Types of syllogisms

(1) $\{Aa - b, Ac - d, Oad\} \vdash Obc$
(2) $\{Aa - b\} \vdash Aab$
(3) $\{Aa - b, Ac - d, Aa - e, Ede\} \vdash Obc$
(4) $\{Aa - b, Aa - c\} \vdash Ibc$
(5) $\{Aa - b, Ac - d, Ae - f, Iae, Edf\} \vdash Obc$
(6) $\{Aa - b, Ac - d, Ebd\} \vdash Eac$
(7) $\{Aa - b, Ac - d, Iac\} \vdash Ibd$

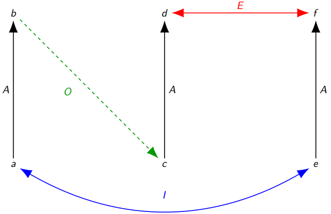
Type (1)
$\{Aa - b, Ac - d, Oad\} \vdash Obc$




Type (6)
$\{Aa - b, Ac - d, Ebd\} \vdash Eac$




Type (5)
$\{Aa - b, Ac - d, Ae - f, Iae, Edf\} \vdash Obc$

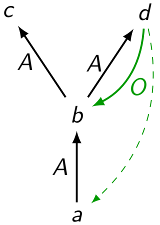


# Example of a syllogism

Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$, select the minimal set of premises to derive $H$ from $\mathcal{KB}$

# Example of a syllogism

Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$, select the minimal set of premises to derive $H$ from $\mathcal{KB}$

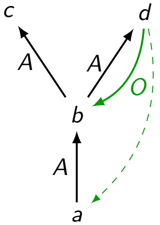



$\mathcal{KB} = \{Aab, Abc, Abd, Odb\}$

$H = Oda$

# Example of a syllogism

Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$, select the minimal set of premises to derive $H$ from $\mathcal{KB}$




$\mathcal{KB} = \{Aab, Abc, Abd, Odb\}$

$H = Oda$

Type (1) $\{Aab, Odb\} \vdash Oda$

# Experiments (I): Neural models for premise selection

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that provide the necessary premises to derive $H$, whenever an inference exists.

# Experiments (I): Neural models for premise selection

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that provide the necessary premises to derive $H$, whenever an inference exists.
- **Experimental setup:**

# Experiments (I): Neural models for premise selection

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that provide the necessary premises to derive $H$, whenever an inference exists.

- **Experimental setup:**

  - A single knowledge base.

# Experiments (I): Neural models for premise selection

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that provide the necessary premises to derive $H$, whenever an inference exists.

- **Experimental setup:**
  - A single knowledge base.
  - One-hot vector representations of syllogistic formulas.

# Experiments (I): Neural models for premise selection

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that provide the necessary premises to derive $H$, whenever an inference exists.

- **Experimental setup:**
  - A single knowledge base.
  - One-hot vector representations of syllogistic formulas.
  - Neural models trained from scratch, including MLPs, RNNs, CNNs, and encoder-only Transformers.

# Overall accuracy

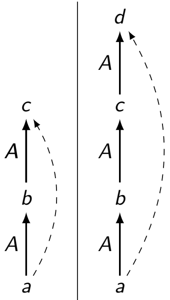
| Model | Inf. | Best | Mean | SD | NNM |
|---|---|---|---|---|---|
| MLP | Val. | 93.9 | 83.2 | 13.1 | 88.9 |
| | Inv. | 97.1 | 94.2 | 2.5 | – |
| | **All** | **96.6** | **93.5** | **3.1** | – |
| RNN | Val. | 95.9 | 93.5 | 1.3 | 95.3 |
| | Inv. | 98.3 | 97.7 | 0.5 | – |
| | **All** | **98.0** | **97.4** | **0.4** | – |
| CNN | Val. | 94.3 | 92.0 | 1.3 | 94.4 |
| | Inv. | 97.3 | 96.7 | 0.3 | – |
| | **All** | **96.9** | **96.4** | **0.2** | – |
| TRA | Val. | 96.6 | 93.6 | 2.9 | 95.7 |
| | Inv. | 97.8 | 96.3 | 1.3 | – |
| | **All** | **97.7** | **96.1** | **1.3** | – |

# Generalization tests for neural models

- Good generalization (the ability to perform on new data) is an essential aspect of NLP neural models.

# Generalization tests for neural models

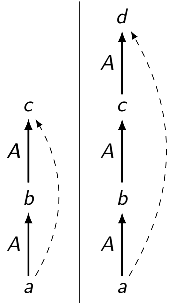
- Good generalization (the ability to perform on new data) is an essential aspect of NLP neural models.




| | | |
|---|---|---|
| Recursiveness | Train | Test |
| Compositionality | Test | Train |

# Generalization tests for neural models

- Good generalization (the ability to perform on new data) is an essential aspect of NLP neural models.

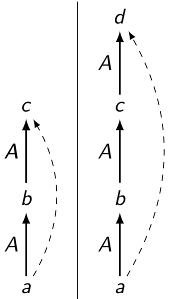



| | | |
|---|---|---|
| Recursiveness | Train | Test |
| Compositionality | Test | Train |

- We define the *length* of inference as the total number of $A$-formulas among the premises.

# Generalization tests for neural models

- Good generalization (the ability to perform on new data) is an essential aspect of NLP neural models.

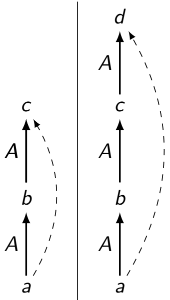



| | | |
|---|---|---|
| Recursiveness | Train | Test |
| Compositionality | Test | Train |

- We define the *length* of inference as the total number of $A$-formulas among the premises.
- For training data, we removed inferences either with short or long lengths.

# Generalization tests for neural models

- Good generalization (the ability to perform on new data) is an essential aspect of NLP neural models.

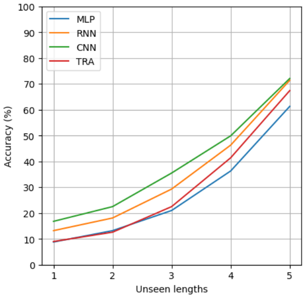



| | | |
|---|---|---|
| Recursiveness | Train | Test |
| Compositionality | Test | Train |

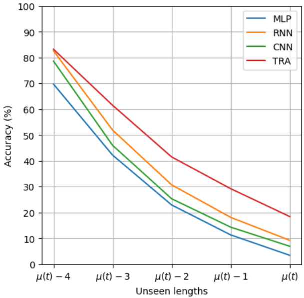
- We define the *length* of inference as the total number of $A$-formulas among the premises.
- For training data, we removed inferences either with short or long lengths.
- For test data, we evaluate the eliminated inferences.

# Results

Compositionality




Recursiveness




- **Neural models generalization:** The models cannot learn the logic's fully recursive and compositional nature.

# Results

- **Additional compositional tests:** When certain components of the knowledge base (e.g., segments of A-chains or entire syllogism types) are removed during training, models fail to recognize or generalize to them at test time.

# Results

- **Additional compositional tests:** When certain components of the knowledge base (e.g., segments of A-chains or entire syllogism types) are removed during training, models fail to recognize or generalize to them at test time.

- **Basic properties:** The models generalize basic non-compositional and non-recursive features of the syllogistic logic: Principle of Contradiction (either $H$ or $\overline{H}$ is invalid), non-empty denotations of constants (if $Aab$ is valid, then $Iab$ is valid), as well as the symmetry of formulas $Iab$ and $Eab$.

# Experiments (II): Neural models for premise selection and proof by contradiction

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that (1) provide the necessary premises to derive $H$ and (2) generate formulas that yield a contradiction, enabling indirect (reductio ad absurdum) proofs.

# Experiments (II): Neural models for premise selection and proof by contradiction

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that (1) provide the necessary premises to derive $H$ and (2) generate formulas that yield a contradiction, enabling indirect (reductio ad absurdum) proofs.

- **Experimental setup:**

# Experiments (II): Neural models for premise selection and proof by contradiction

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that (1) provide the necessary premises to derive $H$ and (2) generate formulas that yield a contradiction, enabling indirect (reductio ad absurdum) proofs.

- **Experimental setup:**

  - Multiple knowledge bases.

# Experiments (II): Neural models for premise selection and proof by contradiction

- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that (1) provide the necessary premises to derive $H$ and (2) generate formulas that yield a contradiction, enabling indirect (reductio ad absurdum) proofs.

- **Experimental setup:**
  - Multiple knowledge bases.
  - Textual representations of syllogistic formulas.

# Experiments (II): Neural models for premise selection and proof by contradiction

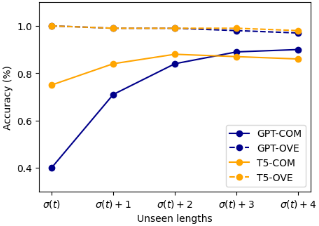
- **Task:** Given a consistent knowledge base $\mathcal{KB}$ along with a hypothesis $H$. Train models that (1) provide the necessary premises to derive $H$ and (2) generate formulas that yield a contradiction, enabling indirect (reductio ad absurdum) proofs.

- **Experimental setup:**
  - Multiple knowledge bases.
  - Textual representations of syllogistic formulas.
  - Fine-tuning pre-trained language models, including a relatively small encoder–decoder model (T5) and a substantially larger decoder-only model (GPT).

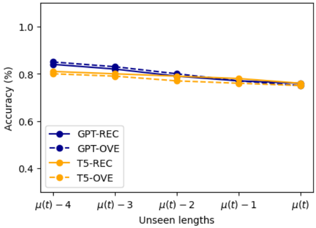
# Generalization performance of GPT and T5

## Task: Premise selection

### Compositionality

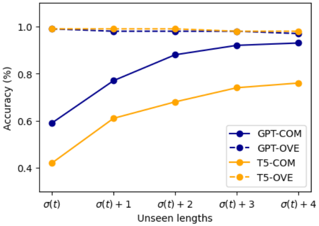



### Recursiveness

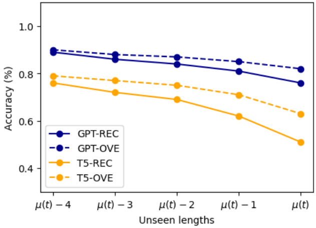



## Task: Proof By Contradiction

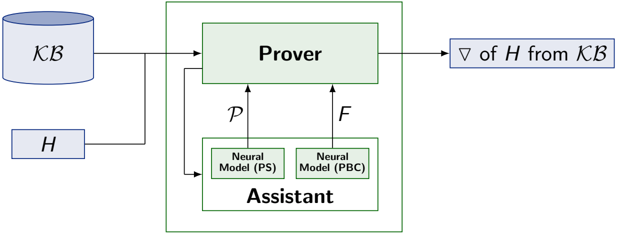
### Compositionality



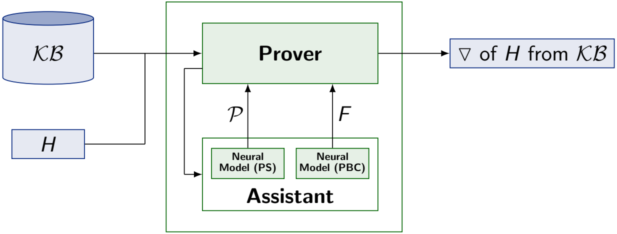

### Recursiveness



# Components of a hybrid model

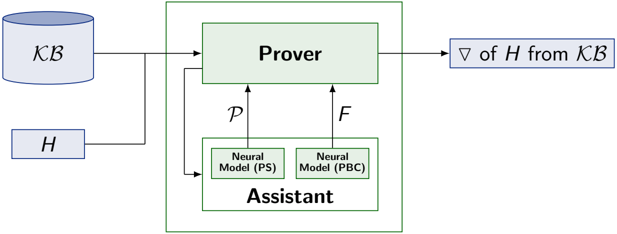


# Components of a hybrid model




**Input:** A knowledge base $\mathcal{KB}$ (set of premises) and a hypothesis $H$.

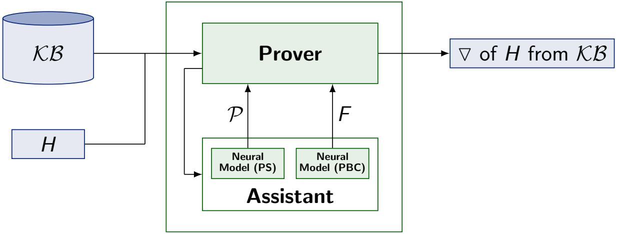
# Components of a hybrid model




**Input:** A knowledge base $\mathcal{KB}$ (set of premises) and a hypothesis $H$.

**Hybrid Model:** If the prover asks for assistance, the neural model (PS) provides $\mathcal{P} \subset \mathcal{KB}$ *s.t.* $\mathcal{P} \vdash H$; and the neural model (PBC) predicts a formula $F$ *s.t.* $\mathcal{KB} \cup \{\overline{H}\} \vdash F \wedge \overline{F}$.

# Components of a hybrid model

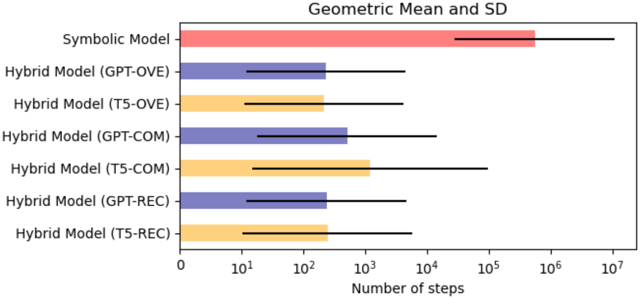



**Input:** A knowledge base $\mathcal{KB}$ (set of premises) and a hypothesis $H$.

**Hybrid Model:** If the prover asks for assistance, the neural model (PS) provides $\mathcal{P} \subset \mathcal{KB}$ *s.t.* $\mathcal{P} \vdash H$; and the neural model (PBC) predicts a formula $F$ *s.t.* $\mathcal{KB} \cup \{\overline{H}\} \vdash F \wedge \overline{F}$.

**Output:** The prover computes a derivation $\nabla$ (if exists) of $H$ from $\mathcal{KB}$.

# Number of steps for the Symbolic and Hybrid models



# Conclusions

- **Neural models generalization:** Pre-trained language models handle recursive reasoning but show weak compositional generalization: training on complex inferences does not transfer to recognizing their simpler components.

# Conclusions

- **Neural models generalization:** Pre-trained language models handle recursive reasoning but show weak compositional generalization: training on complex inferences does not transfer to recognizing their simpler components.

- **Hybrid models comparison:** Hybrid models reduce proof steps by approximately three orders of magnitude compared to a purely symbolic model.

# Conclusions

- **Neural models generalization:** Pre-trained language models handle recursive reasoning but show weak compositional generalization: training on complex inferences does not transfer to recognizing their simpler components.
- **Hybrid models comparison:** Hybrid models reduce proof steps by approximately three orders of magnitude compared to a purely symbolic model.
- **Robustness:** Despite limitations in generalization and scale, LLMs remain effective assistants to symbolic provers.

# Future work

- **Extend the logic:** Future work will investigate richer logical fragments, including those studied by (Pratt-Hartmann, 2004) and selected fragments of modal logic.

# Future work

- **Extend the logic:** Future work will investigate richer logical fragments, including those studied by (Pratt-Hartmann, 2004) and selected fragments of modal logic.
- **Generalization analysis:** Studying richer logical systems may reveal new and qualitatively different generalization challenges.

# References I


Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2019).The compositionality of neural networks: Integrating symbolism and connectionism. *CoRR, abs/1908.08351.* http://arxiv.org/abs/1908.08351

Pratt-Hartmann, I. (2004).Fragments of language. *Journal of Logic, Language and Information*, *13*(2), 207–223. https://doi.org/10.1023/b:jlli.0000024735.97006.5a

Smiley, T. J. (1973).What is a syllogism? *Journal of Philosophical Logic*, *2*(1), 136–154. https://doi.org/10.1007/bf02115614