Helps Reveal When Language Models Learn Meaning

Transparency Helps Reveal When Language Models Learn Meaning

TACL 2023

Zhaofeng Wu, Will Merrill, Hao Peng, Iz Beltagy, and Noah Smith
ᐃᔨ ᐅᓄᕐᓂᖅ ᐅᓄᕐᓂᖅ ᐅᓄᕐᓂᖅ ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐅᓄᕐᓂᖅ ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭᖓᔪᖅ. ᐱᒋᔭᖓᔪᖅ ᐱᒋᔭ[layer2].
LMs can’t learn meaning from form alone.
Can we say LMs understand language?
LMs can’t learn meaning from form alone.

Can we say LMs understand language?

What are these “powerful” LMs really capable of?

def f(n):
    if n == 1 or n == 2:
        return 1
    return f(n - 1) + f(n - 2)
def f(n):
    if n == 1 or n == 2:
        return 1
    return f(n - 1) + f(n - 2)

LMs can't learn execution.
def f(n):
    if n == 1 or n == 2:
        return 1
    return f(n - 1) + f(n - 2)

LMs can't learn execution.

There are assertions:
assert f(6) == 8
def f(n):
    if n == 1 or n == 2:
        return 1
    return f(n - 1) + f(n - 2)

LMs can't learn execution.

There are assertions:
assert f(6) == 8

Assertions enable meaning learnability in some languages.

def f(n):
    if n == 1 or n == 2:
        return 1
    return f(n - 1) + f(n - 2)

LMs can't learn execution.

There are assertions:
assert f(6) == 8

Assertions enable meaning learnability in some languages.

LMs can’t learn execution.

There are assertions:
assert f(6) == 8

Assertions enable meaning learnability in some languages.

LMs learn the meaning of some languages with assertions.
LMs can’t learn execution.

There are assertions:
assert f(6) == 8

Assertions enable meaning learnability in some languages.

LMs learn the meaning of some languages with assertions.

But not natural language.
Can LMs Learn From Assertions?
Setup
Pretraining

\begin{align*}
((\neg T) \land (\neg (T \land (T \lor (\neg F))))) &= (TV(\neg (\neg T) \lor (\neg (\neg F)))) \\
(\neg (\neg (\neg (\neg F) \lor (\neg F)) \lor (\neg (T \land (T \lor (\neg (T \lor (\neg F))))))) &= ((T) \land (\neg F) \land (\neg F)) \\
(((\neg (\neg (\neg (\neg F) \lor (\neg F)) \lor (\neg (T \land (T \lor (\neg (T \lor (\neg F)))))))) \lor (\neg (T) \land (\neg (T) \land (T \lor (\neg (T \lor (\neg F))))))) &= ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(((\neg (\neg (\neg F) \lor (\neg F)) \lor (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) \lor (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
((\neg (\neg (\neg F) \lor (\neg F)) \lor (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) \lor (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F))))))) \\
(\neg ((\neg (T) \land (\neg F) \land (\neg F)) \land (\neg (T) \land (T \lor (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))) &= ((\neg F) \land (\neg F) \land (\neg (T) \land (\neg (T) \land (T \lor (\neg (T) \lor (\neg F)))))))
Pretraining

\[\neg T \land (\neg (\neg F)) \equiv (T \land (\neg (\neg T \lor \neg F)))\]
\[\neg (\neg ((F \land F) \land F) \land F) \equiv (T \land (\neg (\neg T \lor \neg F)))\]
\[(((\neg (\neg (\neg T))) \lor T) \land T) \equiv (F \land (\neg (\neg T \lor \neg F) \land T)))\]
\[(((\neg F) \land (\neg F)) \land (\neg (\neg F) \lor F) \land T) \equiv (F \land (\neg (\neg T \lor \neg F) \land T)))\]
Probing

Pretraining

RoBERTa-like MLM
GPT-2-like ALM
Setup

Pretraining

RoBERTa-like MLM
GPT-2-like ALM

unseen

(FA(¬T))

(Tv(F∧T))
Pretraining

RoBERTa-like MLM
GPT-2-like ALM

unseen

(Tv(F∧T))
(Fa(¬T))
Setup

Pretraining

\[
((-T) \land \neg(Tv(-F))) = (Tv(-((\neg(T) \lor (-T)) \land \neg(T))))
\]
\[
(-((\neg(Tv(-F)))) = ((T\land(-F)) \land T))
\]
\[
((-T) \land \neg(Tv(Tv))) = ((T\land(-F)) \land T))
\]
\[
((Tv(Tv)) \lor (Tv(Fv)) = (Tv(Tv)) \lor (Tv(Fv)) = (Tv(Tv)) \lor (Tv(Fv))
\]
\[
((-T) \land \neg(Tv(Tv))) = ((T\land(-F)) \land T))
\]
\[
(Fv((Tv(-F)))) = ((T\land(-F)) \land T))
\]
\[
(Fv((Tv(-F)))) \lor (Tv(Fv)) = (Tv(Tv)) \lor (Tv(Fv)) = (Tv(Tv)) \lor (Tv(Fv))
\]
\[
((Tv(-F)) \land T) = ((T\land(-F)) \land T))
\]
\[
(Fv((Tv(-F)))) \lor (Tv(Fv)) = (Tv(Tv)) \lor (Tv(Fv)) = (Tv(Tv)) \lor (Tv(Fv))
\]
\[
((Tv(-F)) \land T) = ((T\land(-F)) \land T))
\]
\[
((Tv(-F)) \land T) = ((T\land(-F)) \land T))
\]

Probing

\[\in \{ =, \neq \}\]
Results
Results

Probing Accuracy

Increasingly more probe parameters
Probing Accuracy

Increasingly more probe parameters
Results

Probing Accuracy

Increasingly more probe parameters
Results

Probing Accuracy

Increasingly more probe parameters
Results

Probing Accuracy

Increasingly more probe parameters
Direct Evaluation
Direct Evaluation

• \(((\neg T) \lor F) \lor (\neg T)\) = ___
Direct Evaluation

• \(((\neg T \lor F) \lor (\neg T))\) = ___

• (small twist, see paper)
Results
Results

Direct Eval Accuracy

100
75
50
25
0

ALM (GPT-2-like)  MLM (RoBERTa-like)
Results

Direct Eval Accuracy

100
75
50
25
0

ALM (GPT-2-like)  MLM (RoBERTa-like)
random
Results

Direct Eval Accuracy

ALM (GPT-2-like)  MLM (RoBERTa-like)
Summary
Summary

We let GPT-2 complete the simple arithmetic problem *Three plus five equals*. The five responses below [...] show that this problem is beyond the current capability of GPT-2, and, we would argue, any pure LM.
Summary

We let GPT-2 complete the simple arithmetic problem *Three plus five equals*. The five responses below [...] show that this problem is beyond the current capability of GPT-2, and, we would argue, any pure LM.

LMs can learn to consistently compare and evaluate the meaning of propositional logic expressions.
What About Other Languages?
What About Other Languages?

Assertions enable meaning learnability in some languages.
What About Other Languages?

Assertions enable meaning learnability in some languages.
Strong Transparency
(i.e., context-independency)
**Strong Transparency**
(i.e., context-independency)

- An expression is strongly transparent if its meaning is context-independent
Strong Transparency
(i.e., context-independency)

- An expression is strongly transparent if its meaning is context-independent.
- A language is strongly transparent if all of its expressions are.
Strong Transparency
(i.e., context-independency)

- An expression is strongly transparent if its meaning is context-independent
- A language is strongly transparent if all of its expressions are

$$((\neg \neg F \lor F) \lor (T \lor (F \land T)))$$
An expression is strongly transparent if its meaning is context-independent.

A language is strongly transparent if all of its expressions are strongly transparent.

\[((T \land (F \lor F)) \lor (T \lor (F \land T)))\]

3 + 5
**Strong Transparency**  
(i.e., context-independency)

- An expression is strongly transparent if its meaning is context-independent.
- A language is strongly transparent if all of its expressions are:

\[((T \wedge (F \vee F)) \vee (T \vee (F \wedge T)))\]

3 + 5

x + 5
Strong Transparency
(i.e., context-independency)

- An expression is strongly transparent if its meaning is context-independent.
- A language is strongly transparent if all of its expressions are

\[((\text{T} \land (\text{F} \lor \text{F})) \lor (\text{T} \lor (\text{F} \land \text{T})))\]

3+5

\[x\]

\[x+5\]
Strong Transparency
(i.e., context-independency)

- An expression is strongly transparent if its meaning is context-independent.
- A language is strongly transparent if all of its expressions are.

\[((T \land (F \lor F)) \lor (T \lor (F \land T)))\]

- $x + 5$
- `date.today()`
An expression is strongly transparent if its meaning is context-independent.

A language is strongly transparent if all of its expressions are strongly transparent (i.e., context-independency).

- $((T \land (F \lor F)) \lor (T \lor (F \land T)))$
- $3 + 5$
- $x$
- $x + 5$
- `date.today()`
- 17
Strong Transparency
(i.e., context-independency)

• An expression is strongly transparent if its meaning is context-independent
• A language is strongly transparent if all of its expressions are

$$(T \land (F \lor F)) \lor (T \lor (F \land T))$$

$3+5$

Some corgis run.

$x$

$x+5$

date.today()
An expression is strongly transparent if its meaning is context-independent.

A language is strongly transparent if all of its expressions are strongly transparent.

\[(T \land (F \lor F)) \lor (T \lor (F \land T))\]

Some corgis run.

His corgis run.

\[x + 5\]

\[date.today()\]
An expression is strongly transparent if its meaning is context-independent.

A language is strongly transparent if all of its expressions are strongly transparent.

Some corgis run.

His corgis run.

Today, some corgis ran.
Removing Strong Transparency
Removing Strong Transparency

Probing Accuracy

Increasingly more probe parameters
Removing Strong Transparency

Probing Accuracy

Increasingly more probe parameters

- ALM (GPT-2-like)
- MLM (RoBERTa-like)


Removing Strong Transparency
Removing Strong Transparency

Direct Eval Accuracy

ALM (GPT-2-like)  MLM (RoBERTa-like)

Random

Strongly Transp.
Removing Strong Transparency

Direct Eval Accuracy

- ALM (GPT-2-like)
- MLM (RoBERTa-like)

- Strongly Transp.
- Not Strongly Transp.

random
Another Summary

LMs can learn the meaning of a strongly transparent language. And strong transparency is important for this learnability.
But is NL strongly transparent?
Referential Opacity

Foreshadow: it makes NL not strongly transparent
Referential Opacity
Foreshadow: it makes NL not strongly transparent

[[Superman]] = [[Clark Kent]]
Referential Opacity

Foreshadow: it makes NL not strongly transparent

\[[\text{Superman}] = [\text{Clark Kent}] = \]
Referential Opacity
Foreshadow: it makes NL not strongly transparent

\[
[[\text{Superman}]] = [[\text{Clark Kent}]] = T
\]

[[Lois Lane believes Superman is a hero.]]
\[\parallel\]
T
Referential Opacity

Foreshadow: it makes NL not strongly transparent

\[
\text{Lois Lane believes Superman is a hero.} = \text{Lois Lane believes Clark Kent is a hero.}
\]

\[
[[\text{Superman}] = [[\text{Clark Kent}]] = T
\]

\[
[[\text{Lois Lane believes Superman is a hero.}}] = [[\text{Lois Lane believes Clark Kent is a hero.}}] = F
\]
Referential Opacity

Foreshadow: it makes NL not strongly transparent

\[ [[\text{Superman}]] = [[\text{Clark Kent}]] = \]

\[ [[\text{Lois Lane believes Superman is a hero.}}]] \neq [[\text{Lois Lane believes Clark Kent is a hero.}}]] \]
Referential Opacity
Foreshadow: it makes NL not strongly transparent

\[ [[\text{Superman}]] = [[\text{Clark Kent}]] = \]

propositional attitude verb

\[ [[\text{Lois Lane believes Superman is a hero.}]] \neq [[\text{Lois Lane believes Clark Kent is a hero.}}]] \]

<table>
<thead>
<tr>
<th>T</th>
<th>F</th>
</tr>
</thead>
</table>
Formalizing Referential Opacity
Formalizing Referential Opacity

- **Theorem:** A compositional language with referential opacity is not strongly transparent
Formalizing Referential Opacity

- **Theorem:** A compositional language with referential opacity is not strongly transparent

- We know the meaning of strongly transparent languages is learnable
Formalizing Referential Opacity

- **Theorem:** A compositional language with referential opacity is not strongly transparent

- We know the meaning of strongly transparent languages is learnable

- But we saw strong transparency is important for learnability
Formalizing Referential Opacity

• **Theorem:** A compositional language with referential opacity is not strongly transparent

• We know the meaning of strongly transparent languages is learnable

• But we saw strong transparency is important for learnability

• How well do LMs learn this NL phenomenon that is not strongly transparent?
Setup
Setup

- **Data:** \( \{(s_1, s_2, y)\} \)
Setup

- **Data:** \{ (s_1, s_2, y) \}

- She **wants** to meet {Superman/Clark Kent}. \( y = \text{Non-equivalent} \)
Setup

• **Data:** \( \{(s_1, s_2, y)\} \)

  • She *wants* to meet {Superman/Clark Kent}. \( y = \text{Non-equivalent} \)

  • She *managed* to meet {Superman/Clark Kent}. \( y = \text{Equivalent} \)
Setup

- **Data:** \{ (s_1, s_2, y) \}
  - She *wants* to meet \{Superman/Clark Kent\}. \( y = \) Non-equivalent
  - She *managed* to meet \{Superman/Clark Kent\}. \( y = \) Equivalent
- **Models:** pretrained GPT-2-XL, BERT-large
Setup

- **Data:** \{ (s_1, s_2, y) \}
  - She *wants* to meet \{Superman/Clark Kent\}. \( y = \) Non-equivalent
  - She *managed* to meet \{Superman/Clark Kent\}. \( y = \) Equivalent
- **Models:** pretrained GPT-2-XL, BERT-large
- **Methods:** probing and similarity-based analysis
Probing Results
Probing Results

Probing Accuracy

0 25 50 75 100

GPT-2-XL  BERT-Large
Probing Results

Probing Accuracy

100
75
50
25
0

GPT-2-XL  BERT-Large

random
Probing Results

Probing Accuracy

![Bar chart showing GPT-2-XL and BERT-Large with random accuracy]

- GPT-2-XL
- BERT-Large
- Random
Although LMs could learn the meaning of a strongly transparent language, they don't well-represent referential opacity and hence the meaning of the entirety of NL.
Conclusions
Conclusions

• Aligning with the theoretical guarantee, current LM architectures & objectives can learn the meaning of a strongly transparent language
Conclusions

• Aligning with the theoretical guarantee, current LM architectures & objectives can learn the meaning of a strongly transparent language

• Strong transparency plays a big part in this learnability
Conclusions

• Aligning with the theoretical guarantee, current LM architectures & objectives can learn the meaning of a strongly transparent language

• Strong transparency plays a big part in this learnability
  • Though learnability is not completely destroyed w/o strong transparency
Conclusions

• Aligning with the theoretical guarantee, current LM architectures & objectives can learn the meaning of a strongly transparent language

• Strong transparency plays a big part in this learnability
  • Though learnability is not completely destroyed w/o strong transparency

• On NL, there is no evidence at all of LMs representing referential opacity, a phenomenon that is not strongly transparent
Propositional Logic vs. NL
Propositional Logic vs. NL

• Why did we see >random probing/eval accuracy on the perturbed propositional logic, but not referential opacity?
Propositional Logic vs. NL

• Why did we see >random probing/eval accuracy on the perturbed propositional logic, but not referential opacity?

• Maybe referential opacity is just harder
Propositional Logic vs. NL

• Why did we see >random probing/eval accuracy on the perturbed propositional logic, but not referential opacity?
  
• Maybe referential opacity is just harder
  
• Maybe it’s because of the large variation in NL, with sentences that are untruthful, subjective, etc.
Propositional Logic vs. NL

• Why did we see >random probing/eval accuracy on the perturbed propositional logic, but not referential opacity?
  
  • Maybe referential opacity is just harder
  
  • Maybe it’s because of the large variation in NL, with sentences that are untruthful, subjective, etc.
  
  • Or maybe…
Encore: Grounding =
Encore: Grounding =
Encore: Grounding =

\[
\begin{align*}
((\neg T) \land (\neg (T \lor (\neg F)))) &= (T \land (\neg T \lor (\neg F))) \\
(\neg (\neg((FA(FAF)A)FA)(\neg T))) &= ((TA(T)A(F)\lor (\neg F))) \\
(((\neg (\neg (\neg T))))(\neg T)) &= (\neg T \lor (\neg (\neg F))) \\
((TA(FV)F)\lor (TA(T))) &= ((\neg F) \land (\neg F) \lor (F \land T)) \\
(((\neg F) \land (\neg F) \land (((\neg F) \lor F) \land F)) &= (F \land (\neg T)) \\
((FA)(FA)((FA)(FA)\lor (\neg T))) &= (FA(FA(F)\lor (\neg T))) \\
(FA(FA((FA)F)\lor (\neg T))) &= (FA(FA(F)\lor (\neg T))) \\
\end{align*}
\]

Probing Accuracy

\[
a=b \quad 50.5
\]
Encore: Grounding =

\[
((-T)A(\neg(Tv(\neg F)))) = (Tv(\neg((\neg T)v(\neg F))))
\]

\[
(\neg(((\neg T)v F)v F)\neg T) = (((\neg T)v F)v F)
\]

Probing Accuracy

<table>
<thead>
<tr>
<th>-Reflexivity</th>
<th>+Reflexivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>a=b</td>
<td>a=b, a=a, b=b</td>
</tr>
<tr>
<td>50.5</td>
<td>92.7</td>
</tr>
</tbody>
</table>
Encore: Grounding =

\[
\begin{align*}
((\neg T)A(\neg (Tv(\neg F)))) & = (Tv(\neg ((\neg T)v(\neg F)))) \\
((\neg (\neg ((FA(FvFAFAFAFA)))\neg (Tv(\neg F)))) & = ((TAT)A((\neg F)v(\neg F))) \\
((\neg (\neg ((Tv(\neg (TvFA)))vTv))) & = ((\neg T)v(\neg (TAT))) \\
((TaFvF)\neg (Tv(\neg (TvFAT)))) & = ((\neg T)v(\neg ((\neg (Tv(\neg (Tv(\neg T)))vTv)))) \\
((\neg (\neg F))A(-(\neg F))) & = (FA(-Fv((\neg FvTv(\neg T)))vTv)) \\
((Tv(\neg (Tv(\neg (Tv(\neg (TvFV)))))vTv)) & = ((\neg (Tv(\neg (Tv(\neg (Tv(\neg (TvFA)))vTv))))A(\neg F)) \\
((\neg (\neg (Tv(\neg (Tv(\neg (TvFV))))))) & = (FA(Fv(\neg T)))vTv(\neg F))) \\
((FaF(\neg (Tv(\neg T)))vTv(\neg F))) & = (\neg (\neg (\neg T)vFvTv(\neg F))) \\
(FaF(\neg ((TvFA)vTv(\neg T))))) & = (\neg (\neg (\neg (\neg (TvFA)vTv(\neg T)))vTv(\neg F))) \\
(FaF(\neg ((TvFA)vTv(\neg T)))) & = (\neg (\neg (TvFA)vTv(\neg T)))A(\neg F)))
\end{align*}
\]

### Probing Accuracy

<table>
<thead>
<tr>
<th></th>
<th>-Reflexivity</th>
<th>+Reflexivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Symmetry</td>
<td>a=b</td>
<td>a=b, a=a, b=b</td>
</tr>
<tr>
<td></td>
<td>50.5</td>
<td>92.7</td>
</tr>
<tr>
<td>+Symmetry</td>
<td>a=b, b=a</td>
<td>a=b, b=a, a=a, b=b</td>
</tr>
<tr>
<td></td>
<td>50.3</td>
<td>98.8</td>
</tr>
</tbody>
</table>
Propositional Logic vs. NL

• Why did we see >random probing accuracy on the perturbed propositional logic, but not referential opacity?
  • Maybe referential opacity is just harder
  • Maybe it’s because of the variation in NL, with sentences that are untruthful, subjective, etc.
  • Or maybe…
Propositional Logic vs. NL

• Why did we see >random probing accuracy on the perturbed propositional logic, but not referential opacity?
  
  • Maybe referential opacity is just harder

  • Maybe it’s because of the variation in NL, with sentences that are untruthful, subjective, etc.

  • Or maybe…

    • We don’t have such an explicit representation of equivalence in NL pretraining
• Aligning with the theoretical guarantee, current LM architectures & objectives can learn the meaning of a strongly transparent language

• Strong transparency plays a big part in this learnability
  • Though learnability is not completely destroyed w/o strong transparency

• On NL, there is no evidence at all of LMs representing referential opacity, a phenomenon that is not strongly transparent

• Careful design of the pretraining data/setup is crucial