Transparency Helps Reveal When Language Models Learn Meaning

TACL 2023

Zhaofeng Wu, Will Merrill, Hao Peng, Iz Beltagy, and Noah Smith





ᐃᔨ ᐊᑕᓐᓂᐅᐹ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐃᓗᖅᐳᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐱᑦᓴᑎᑦᑐᖅ ᐅᕝᕙ. ᐅᓄᕐᓂᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐱᒋᕚ ᒪᕐᑮᒃ ᐸᓪᓗᓯᖕᒐᔪᖅ ᐊᑏᓗ ᐱᒋᕚ ᐱᓪᓗᐊᓇᕚ. ᐋᓯᕙᖅ ᐱᒋᕚ ᐊᕐᕕᓂᓖᑦ ᐱᖕᒐᓯᓂᒃ ᐸᓪᓗᓯᖕᒐᔪᖅ. ᐃᔾᔪᔪᒡ ᑕᖅᓴᓚᖅᐹ

ᐃᔾ ᐊᑕᓐᓂᐅᐹ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐃᓗᖅᐳᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐱᑦᓴᑎᑦᑐᖅ ᐅᕝᕙ. ᐅᓄᕐᓂᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ
ᐱᕆᕖ ᒪᕐᖀᒃ ᐸᓪᓗᓯᖕᖓᔪᖅ ᐊᑏᓗ ᐱᕆᕖ ᐱᓪᓗᐊᓇᕖ. ᐋᓯᕙᖅ ᐱᕆᕖ ᐊᖅᕕᓂᓖᑦ ᐱᖕᖓᓱᓂᒃ
ᐸᓪᓗᓯᖕᖓᔪᖅ. ᐃᔾᔪᔪᒡ ᑕᖅᓴᓕᖅᐹ

LMs can’t learn meaning from form alone.

ᐃᔾ ᐊᑕᓐᓂᐅᐹ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐃᓗᖅᐳᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐱᑦᓴᑎᑦᑐᖅ ᐅᕝᕙ. ᐅᓇᕐᓂᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐱᕆᕚ ᒪᕐᕘᒃ ᐸᓪᓗᓯᖕᒐᔪᖅ ᐊᑏᓗ ᐱᕆᕚ ᐱᓪᓗᐊᓇᕚ. ᐋᓯᕙᖅ ᐱᕆᕚ ᐊᕐᕕᓂᓖᑦ ᐱᖕᒐᓱᓂᒃ ᐸᓪᓗᓯᖕᒐᔪᖅ. ᐃᔾᔪᔪᒃ ᑕᖅᓴᓕᖅᐹ

LMs can’t learn meaning from form alone.

Can we say LMs understand language?

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proc. of ACL.

ᐃᔾ ᐊᑕᓐᓂᐅᕚ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐃᓗᖅᐳᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ ᐱᑦᓴᑎᑦᑐᖅ ᐅᕝᕙ. ᐅᓄᕐᓂᖅ ᓯᓚᕐᔪᐊᒥᐅᑦ
ᐱᕆᕖ ᒪᕐᕉᒃ ᐸᓪᓗᓯᖕᒐᔪᖅ ᐊᑏᓗ ᐱᕆᕖ ᐱᓪᓗᐊᓇᕖ. ᐋᓯᕙᖅ ᐱᕆᕖ ᐊᕐᕕᓂᓖᑦ ᐱᖕᒐᓱᓂᒃ
ᐸᓪᓗᓯᖕᒐᔪᖅ. ᐃᔾᔪᔪᒡ ᑕᖅᓴᓕᖅᐹ

LMs can’t learn meaning from form alone.

Can we say LMs understand language?

What are these “powerful” LMs really capable of?

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proc. of ACL.

```python
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.

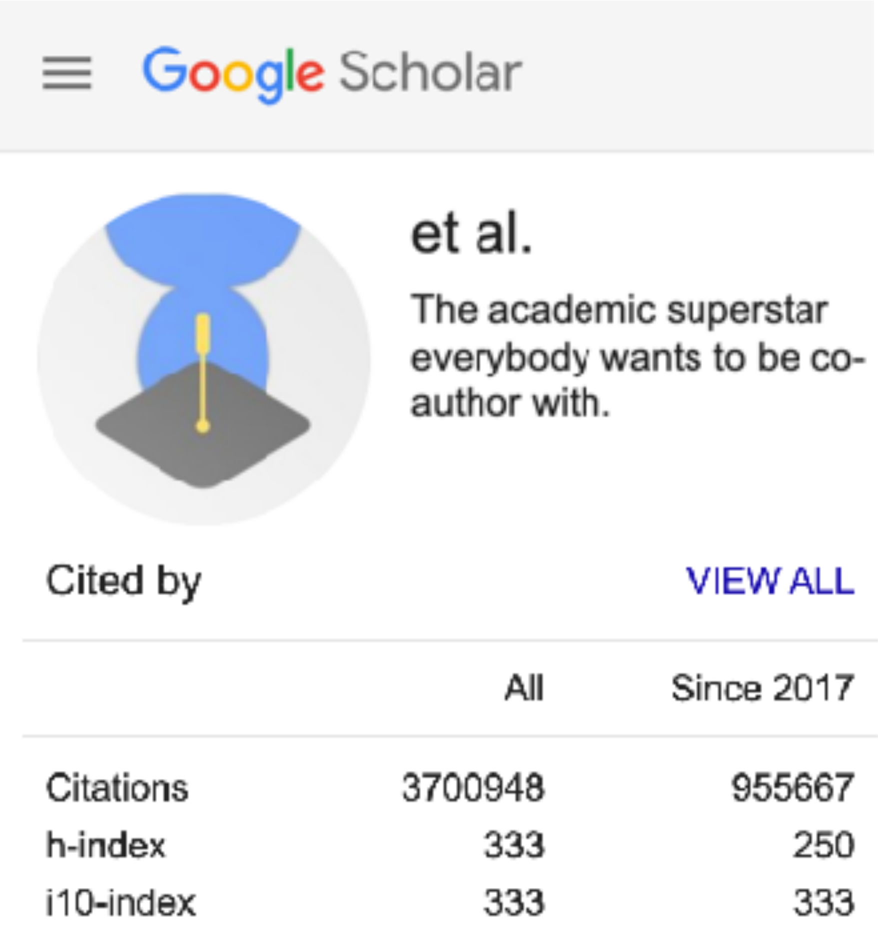
There are assertions:
`assert f(6) == 8`

```
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.



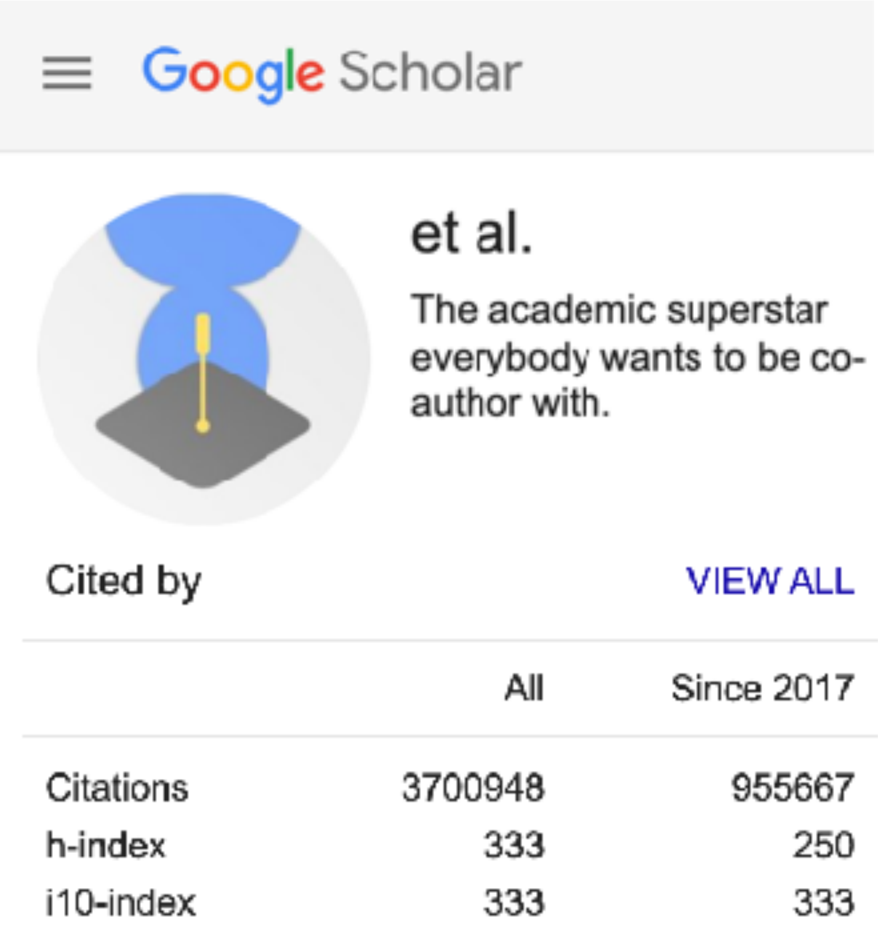
William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021.
Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? TACL, 9:1047–1060.

```
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.

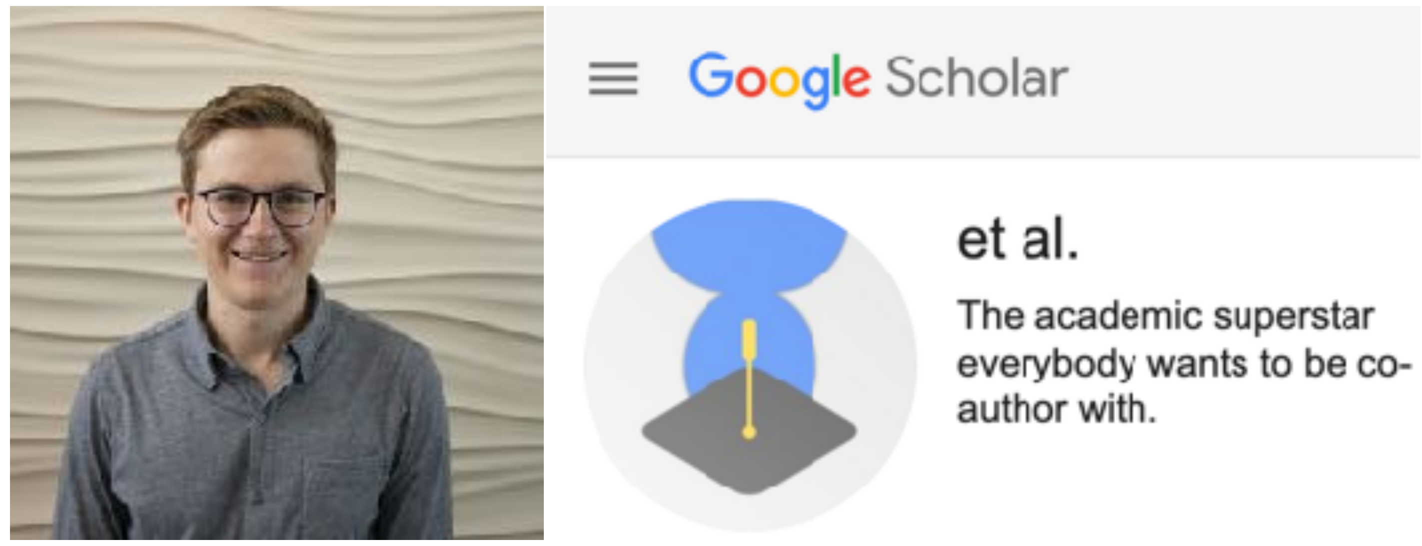


William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021.
Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? TACL, 9:1047–1060.

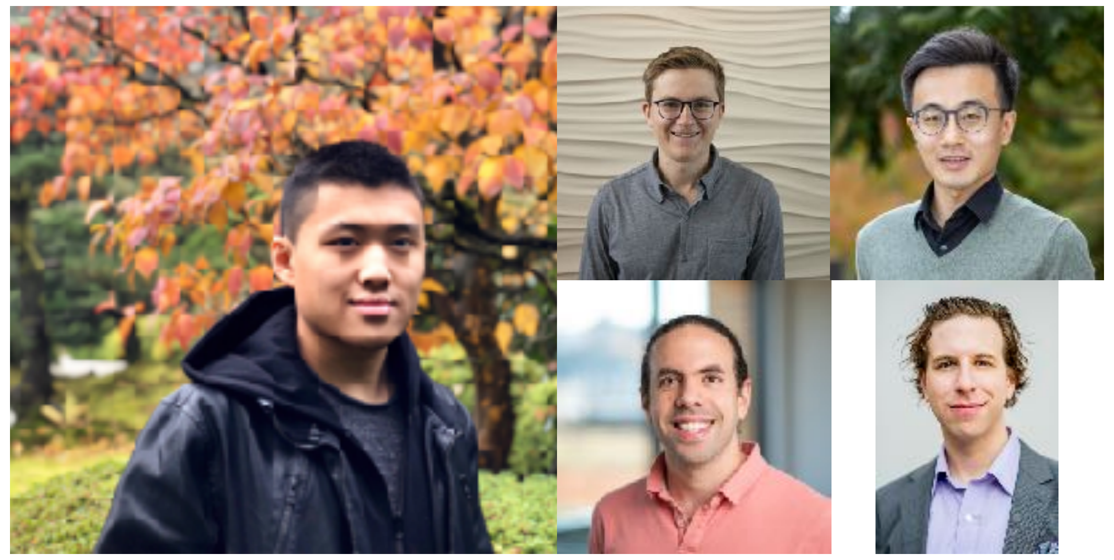
LMs can’t learn execution.

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

Assertions enable meaning learnability in some languages.



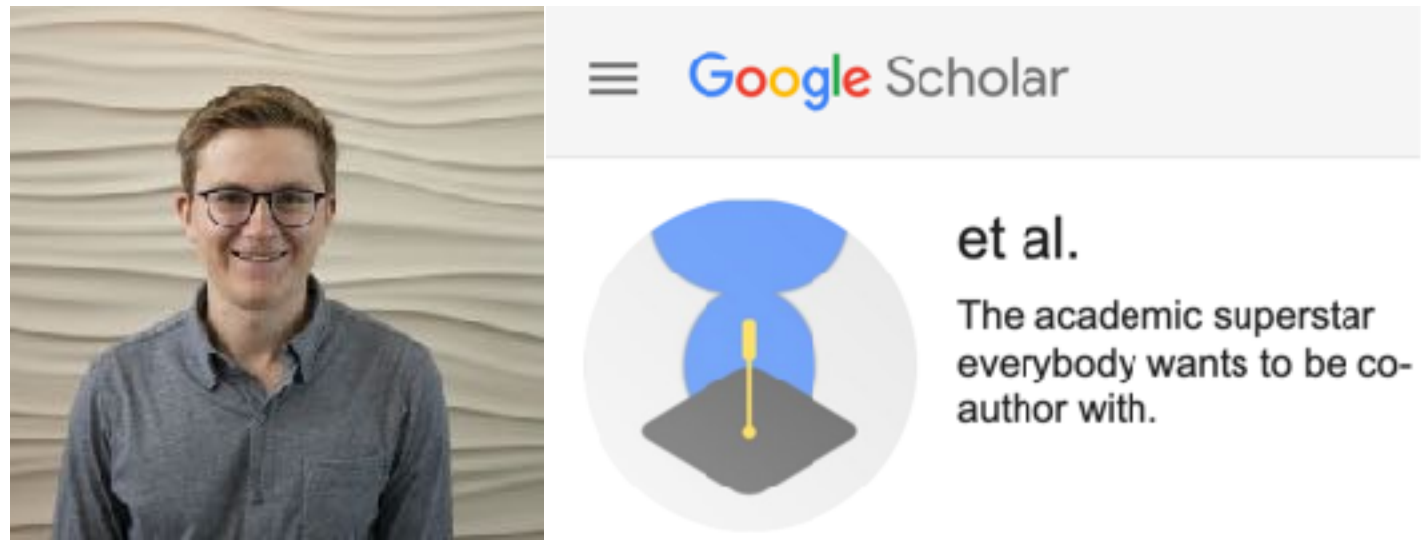
LMs learn the meaning of some languages with assertions.



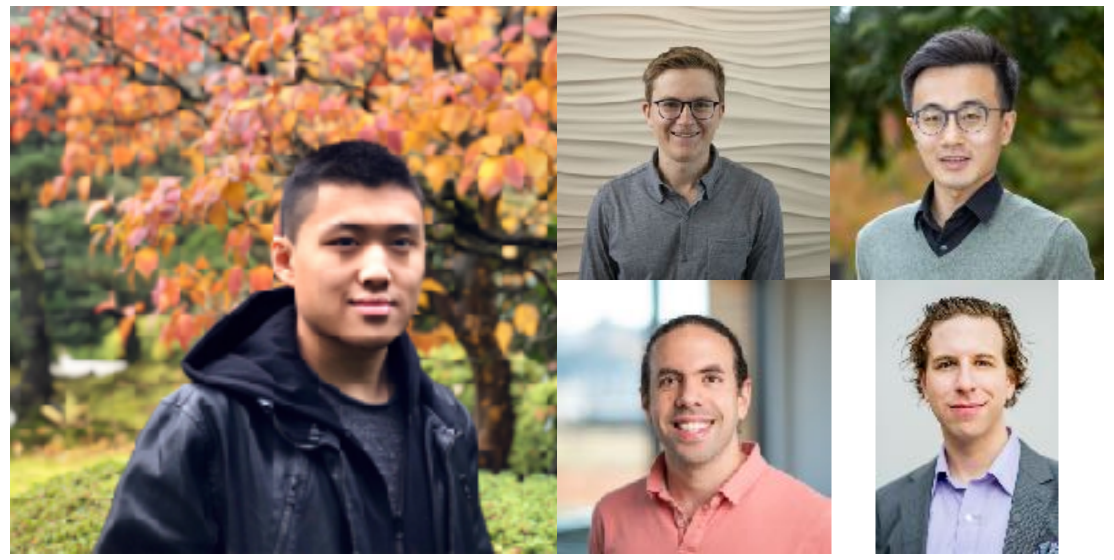
LMs can’t learn execution.

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

Setup

```
((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T)))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T)))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))
```

Pretraining

Setup

((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T)))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))

RoBERTa-like MLM
GPT-2-like ALM

Pretraining

Setup

((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T)))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))

Pretraining

RoBERTa-like MLM
GPT-2-like ALM

Probing

Setup

```
((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T)))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))
```

Pretraining

RoBERTa-like MLM
GPT-2-like ALM

Probing

Setup

((¬T)∧(¬(Tv(¬F))))=(Tv(¬(¬((¬T)v(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)v(¬F)))
(((¬((¬(¬(¬(¬T))))vT))vT)∧(¬(¬T)))=((¬F)v(¬(T∧(TvT))))
((T∧(FvF))v(Tv(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))vF))v(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)vF)∧F))=(F∧(¬(¬(Fv((¬(Fv(¬T)))∧T)))))
((Tv(¬(T∧(Tv(¬(Fv(¬F)))))))vT)=(¬((¬(Tv(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))v((¬F)v(Tv(¬(TvT))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(TvT))∧(¬T)))))=(¬(¬(((¬T)vF)v(Fv(¬(¬F))))))
(F∧(F∧(¬((FvF)v(¬(¬T))))))=(¬((((¬(T∧T))v(¬F))v(¬T))∧(¬F)))

Pretraining

(Tv(F∧T))

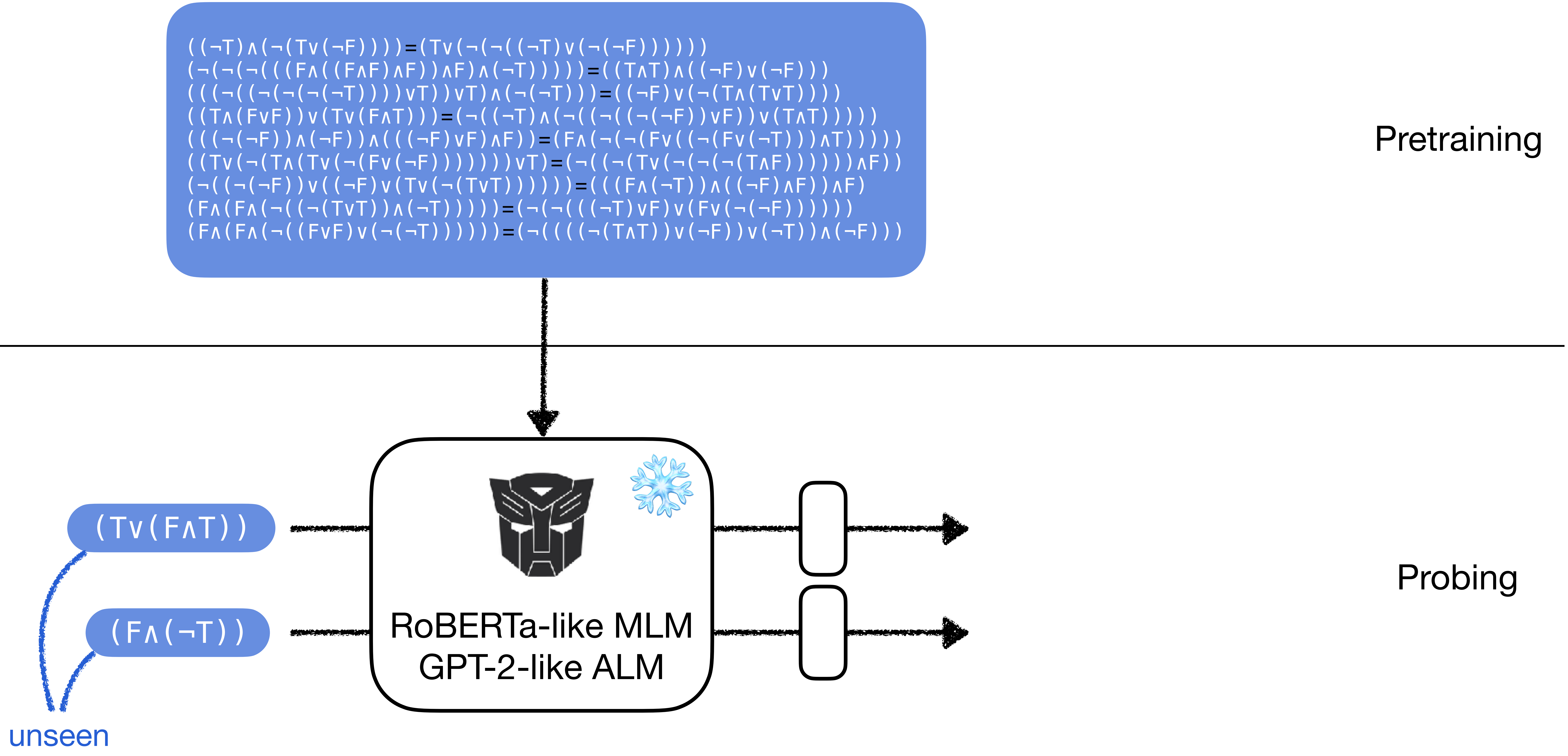
(F∧(¬T))

unseen

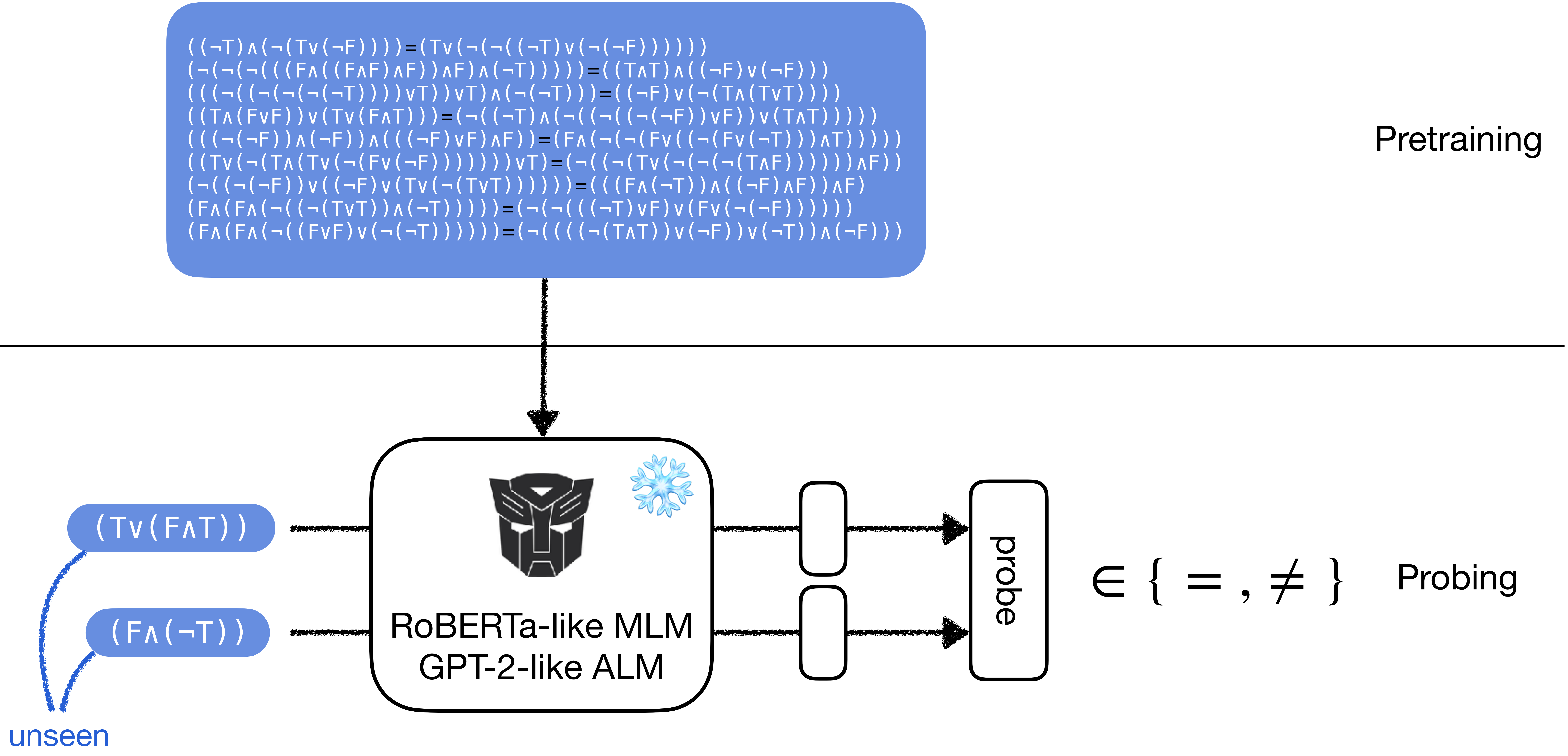
RoBERTa-like MLM
GPT-2-like ALM

Probing

Setup

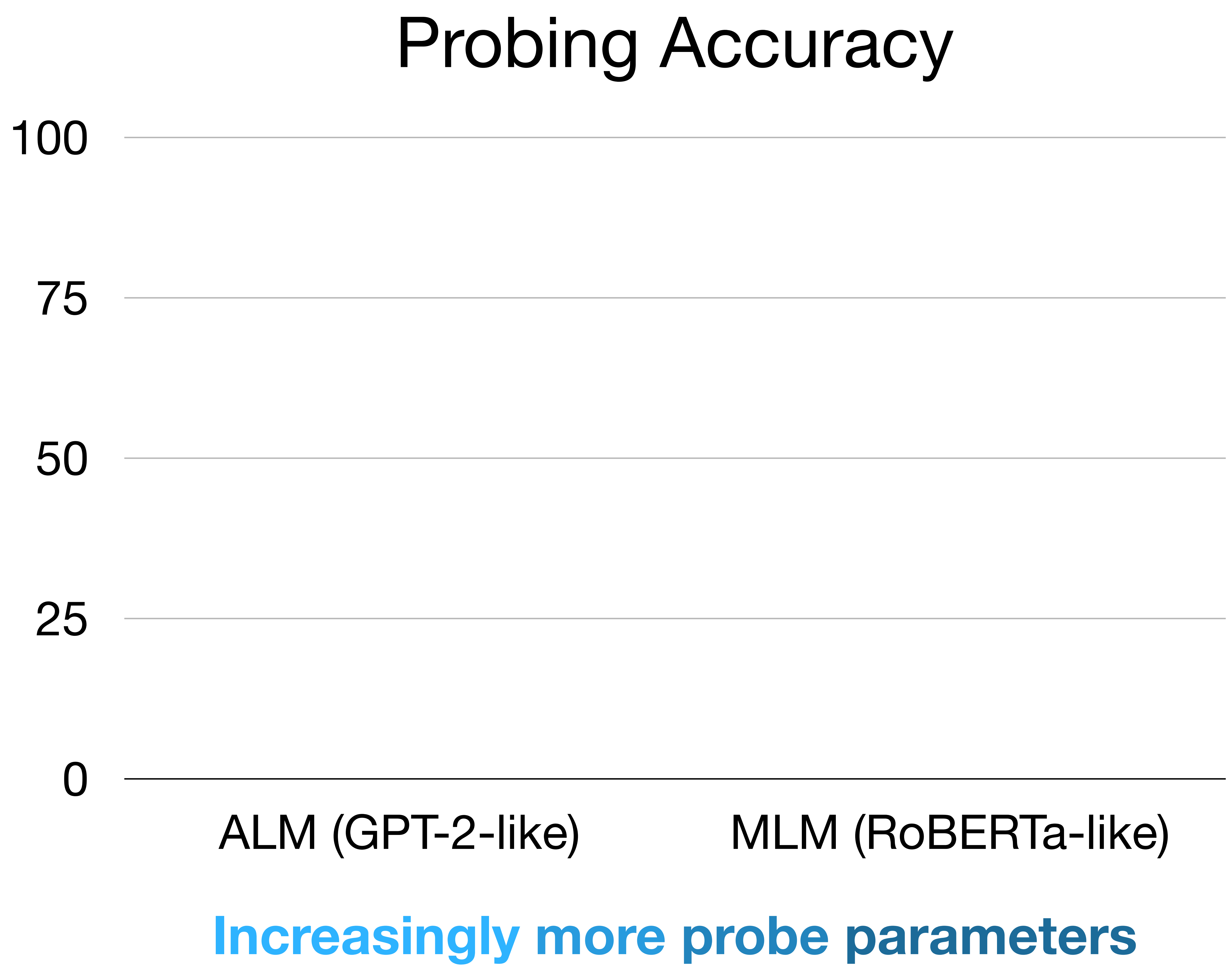

((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T))))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T)))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))
Pretraining
(T∨(F∧T))
(F∧(¬T))
unseen
RoBERTa-like MLM
GPT-2-like ALM
Probing

Setup

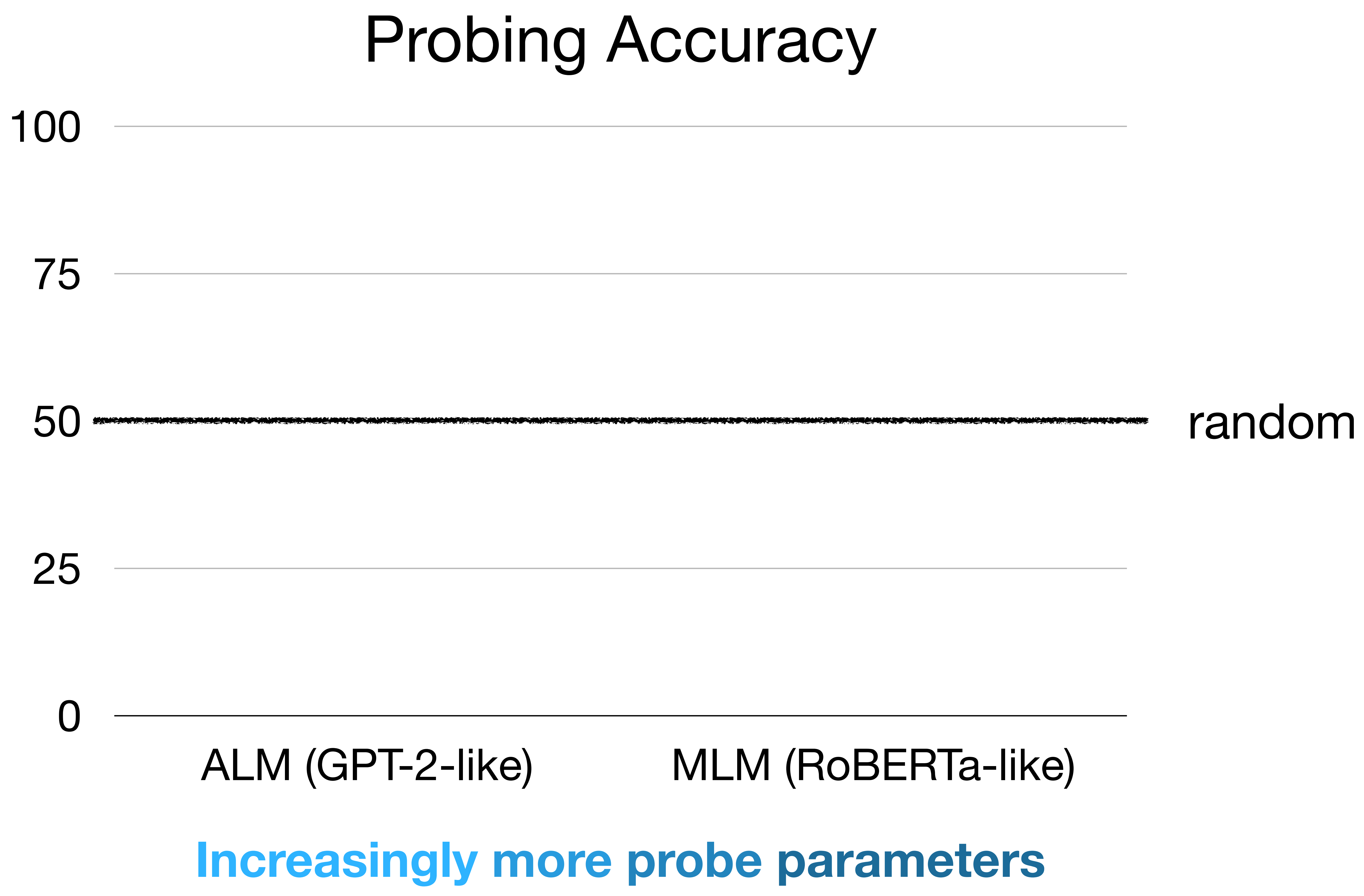

((¬T)∧(¬(Tv(¬F))))=(Tv(¬(¬((¬T)v(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)v(¬F)))
(((¬((¬(¬(¬(¬T)))))vT))vT)∧(¬(¬T)))=((¬F)v(¬(T∧(TvT))))
((T∧(FvF))v(Tv(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))vF))v(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)vF)∧F))=(F∧(¬(¬(Fv((¬(Fv(¬T)))∧T)))))
((Tv(¬(T∧(Tv(¬(Fv(¬F)))))))vT)=(¬((¬(Tv(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))v((¬F)v(Tv(¬(TvT))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(TvT))∧(¬T))))))=(¬(¬(((¬T)vF)v(Fv(¬(¬F))))))
(F∧(F∧(¬((FvF)v(¬(¬T)))))))=(¬((((¬(T∧T))v(¬F))v(¬T))∧(¬F)))
Pretraining
(Tv(F∧T))
(F∧(¬T))
unseen
RoBERTa-like MLM
GPT-2-like ALM
probe
∈ { = , ≠ }
Probing

Results

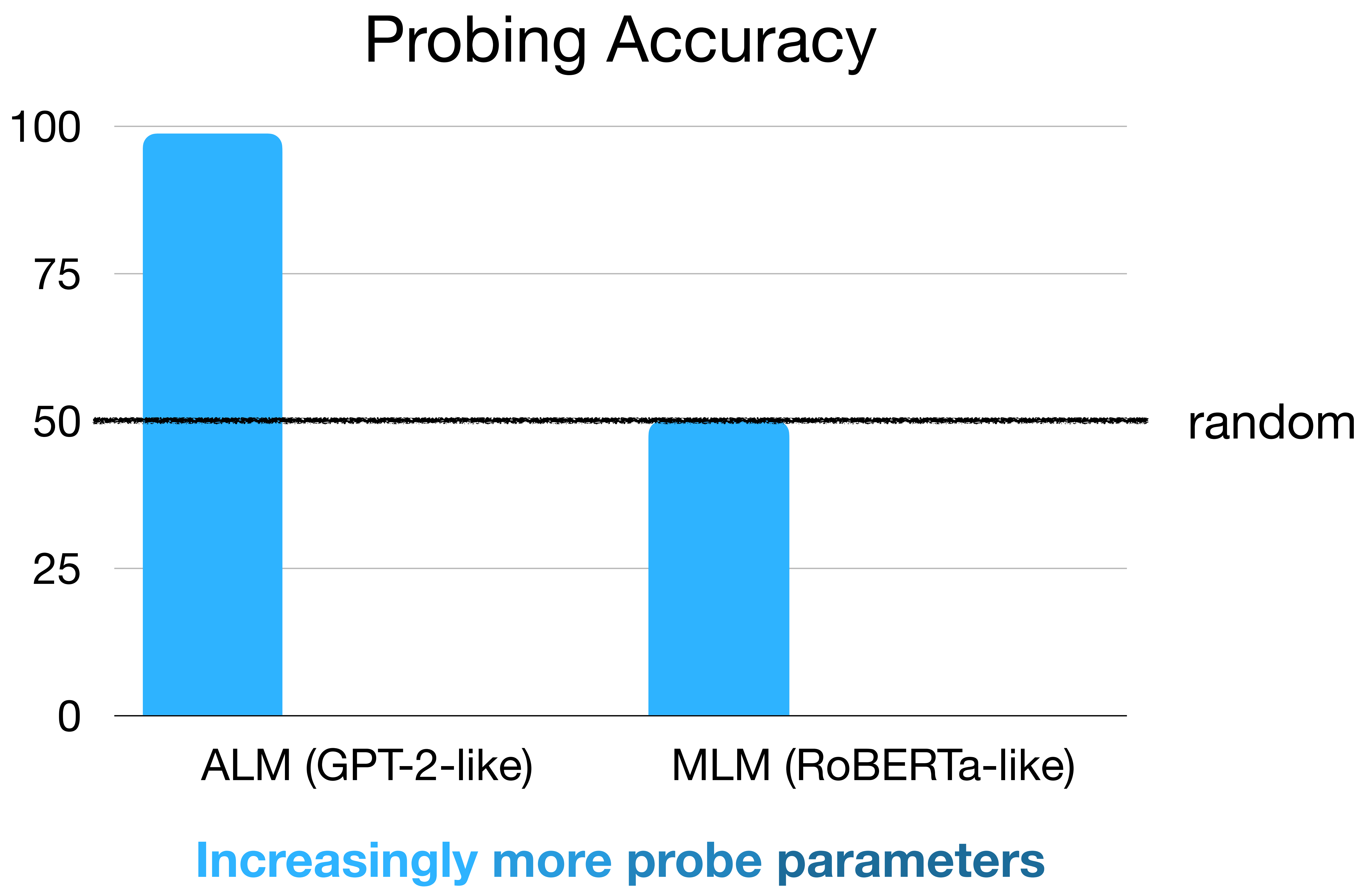
Results


Probing Accuracy
100
75
50
25
0
ALM (GPT-2-like)
MLM (RoBERTa-like)
Increasingly more probe parameters

Results

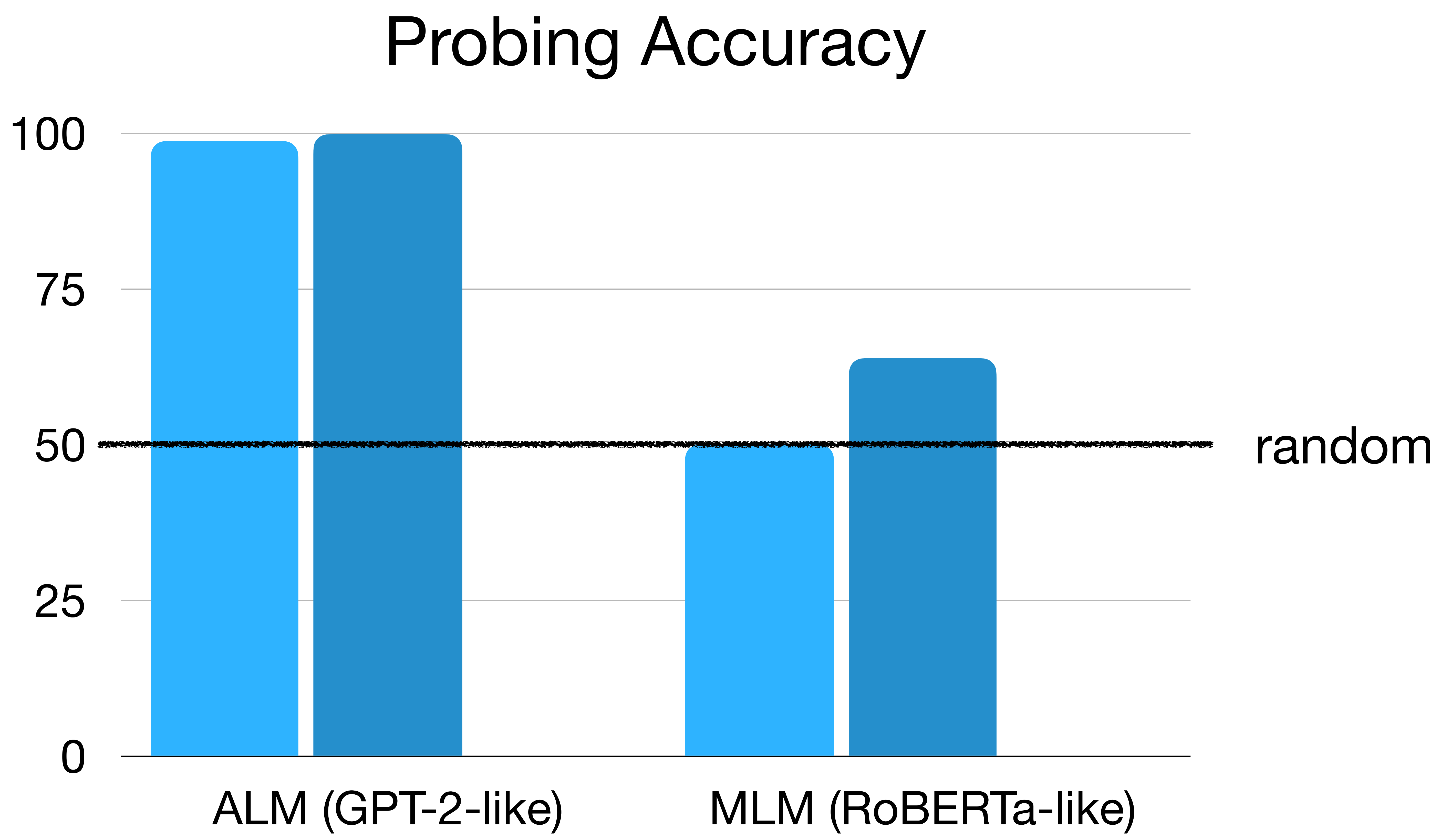

Probing Accuracy
100
75
50
25
0
random
ALM (GPT-2-like)
MLM (RoBERTa-like)
Increasingly more probe parameters

Results



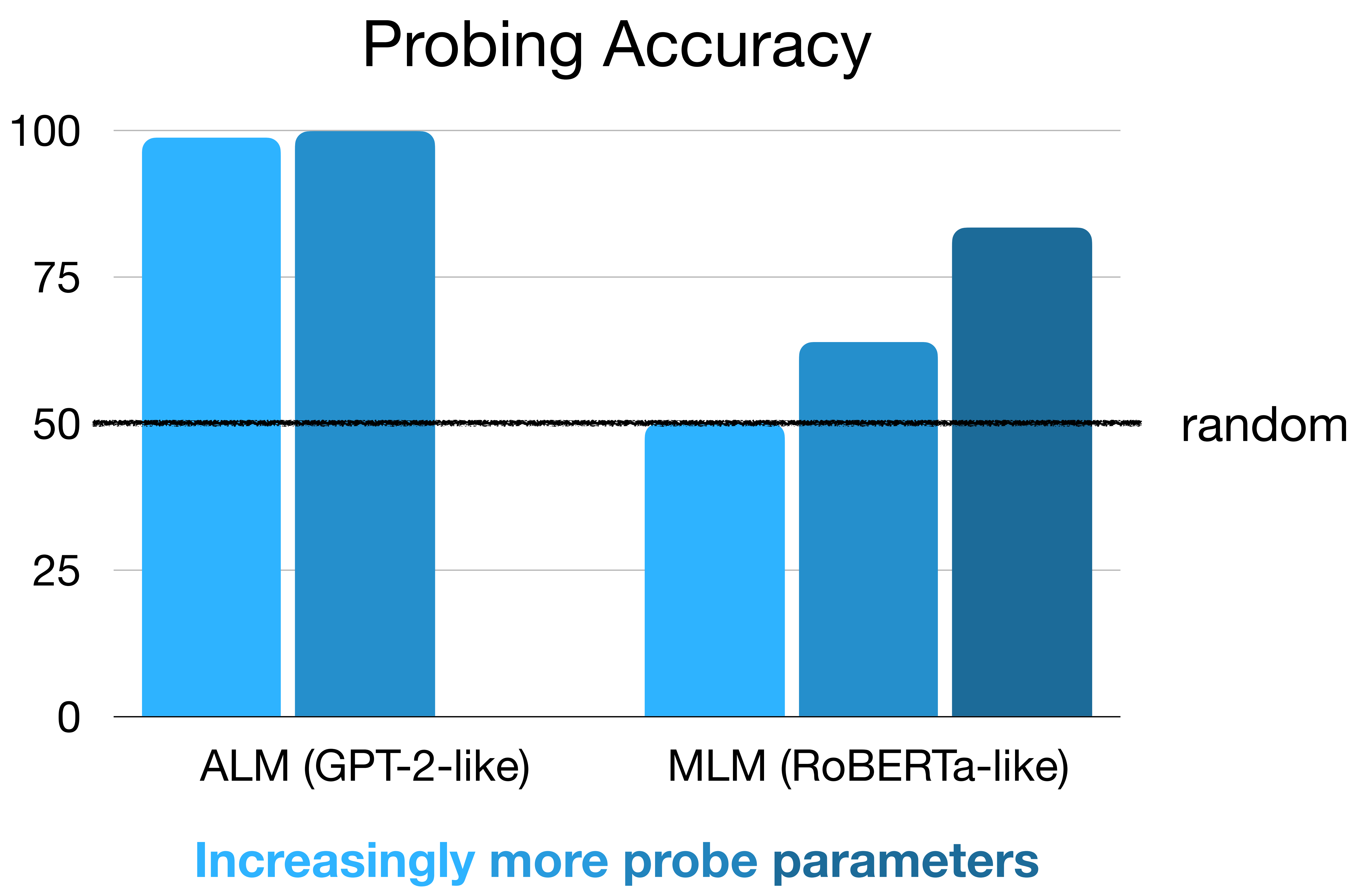
Probing Accuracy
100
75
50
25
0
random
ALM (GPT-2-like)
MLM (RoBERTa-like)
Increasingly more probe parameters

Results



Increasingly more probe parameters

Results


Probing Accuracy
100
75
50
25
0
random
ALM (GPT-2-like)
MLM (RoBERTa-like)
Increasingly more probe parameters

Direct Evaluation

Direct Evaluation

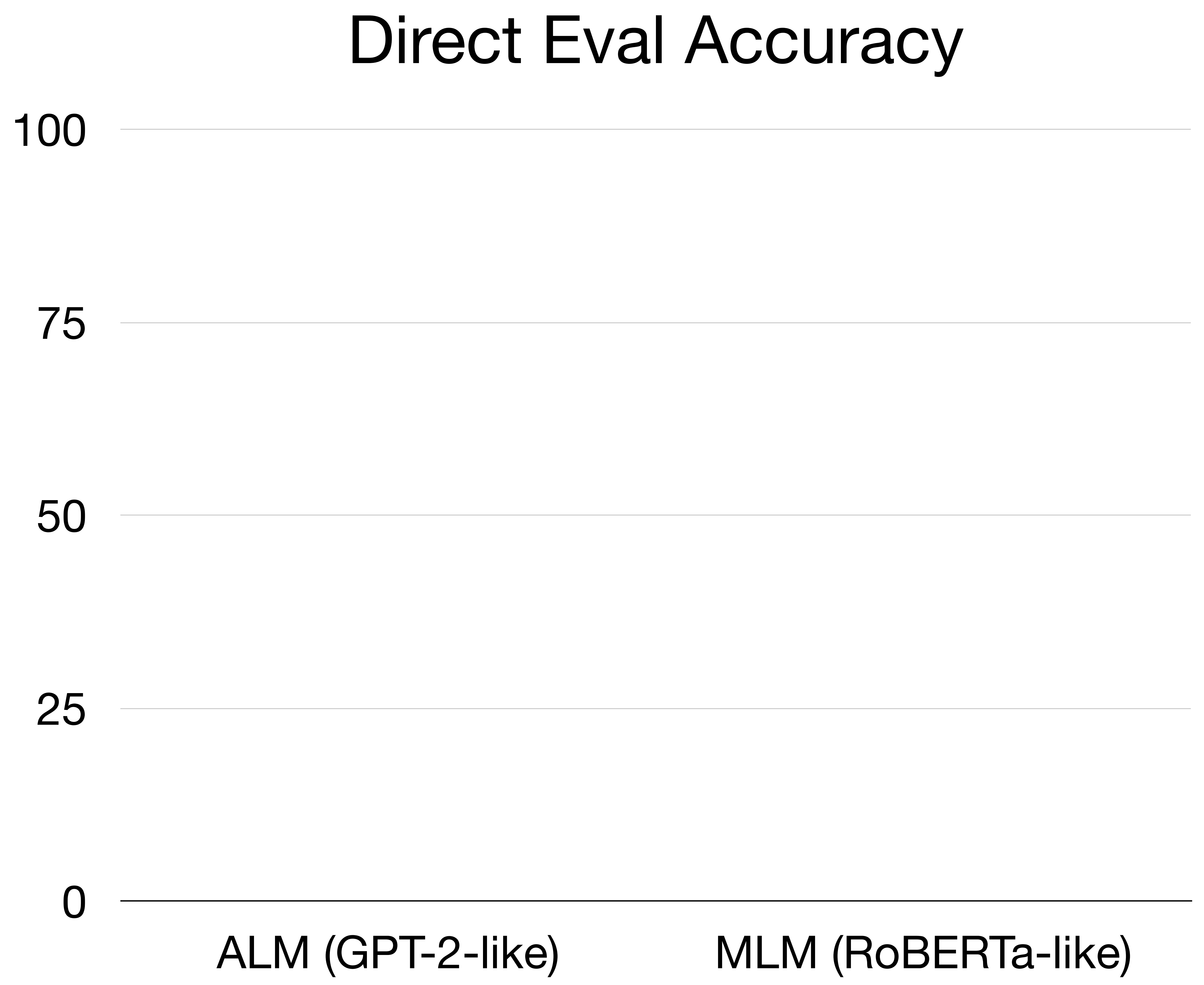
- $(((\neg T) \vee F) \vee (\neg T)) = \_\_\_$

Direct Evaluation

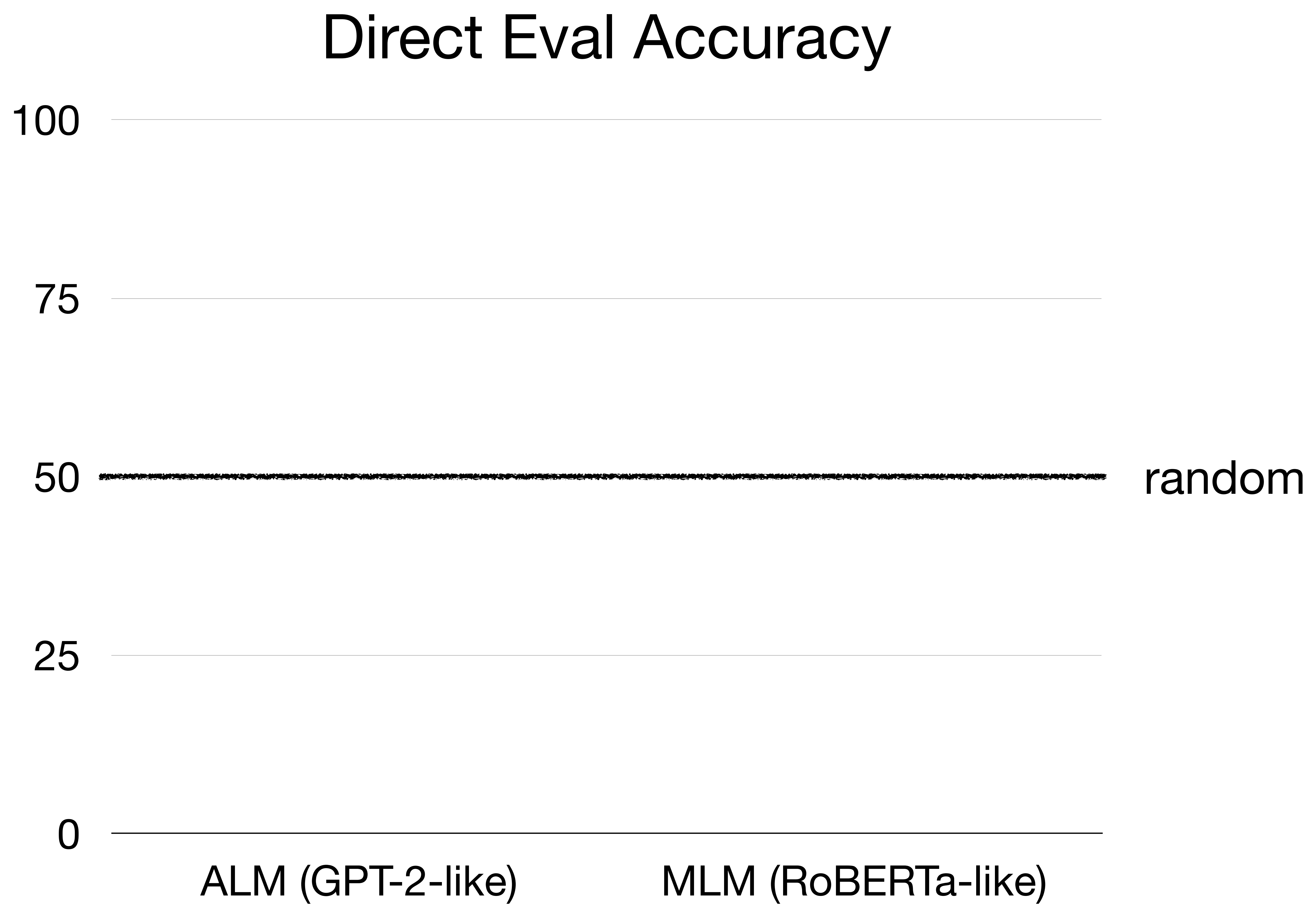
- $(((\neg T) \vee F) \vee (\neg T)) = \_\_\_$
- (small twist, see paper)

Results

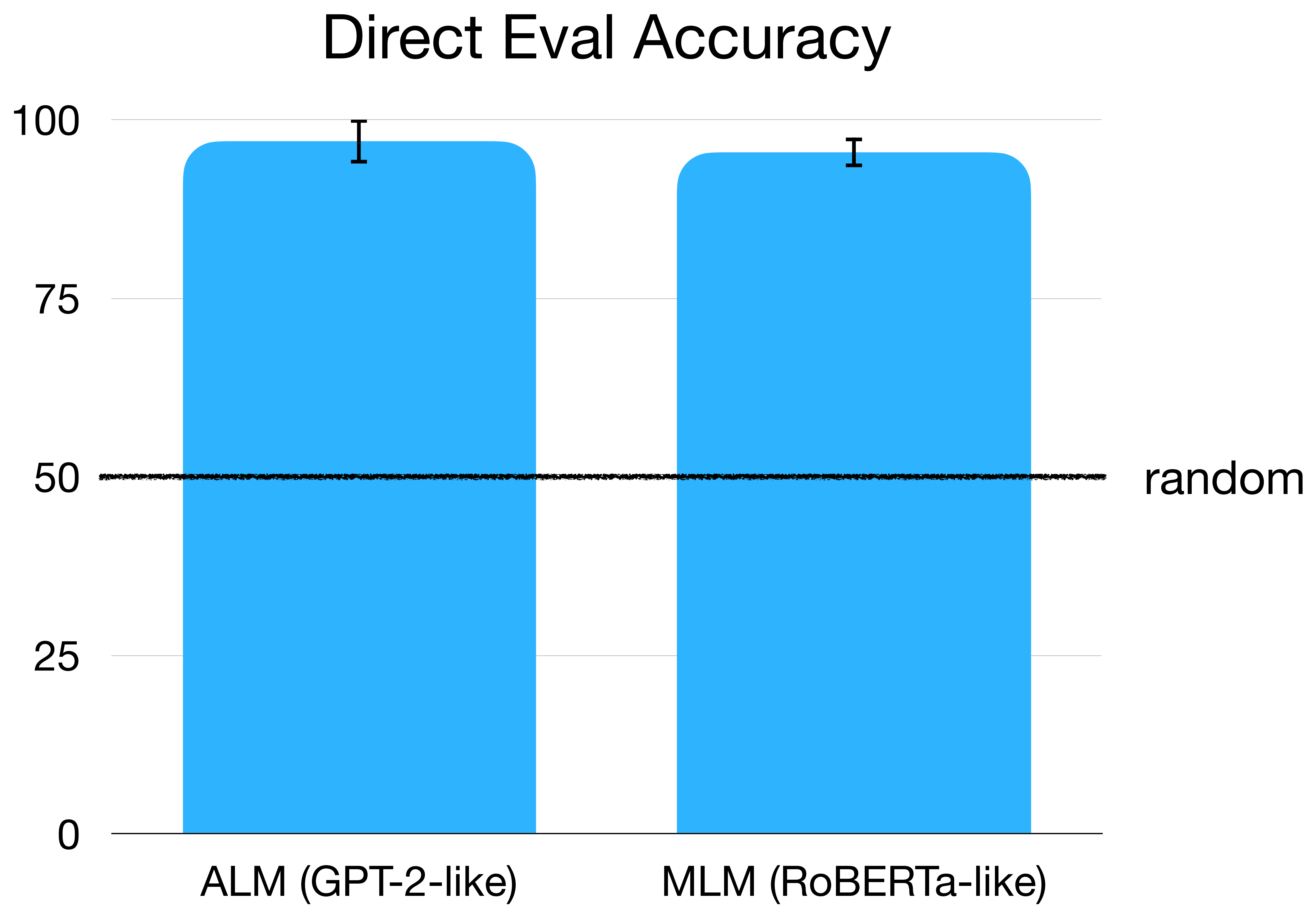
Results


Direct Eval Accuracy
100
75
50
25
0
ALM (GPT-2-like)
MLM (RoBERTa-like)

Results


Direct Eval Accuracy
100
75
50
25
0
random
ALM (GPT-2-like)
MLM (RoBERTa-like)

Results


Direct Eval Accuracy
100
75
50
25
0
random
ALM (GPT-2-like)
MLM (RoBERTa-like)

Summary

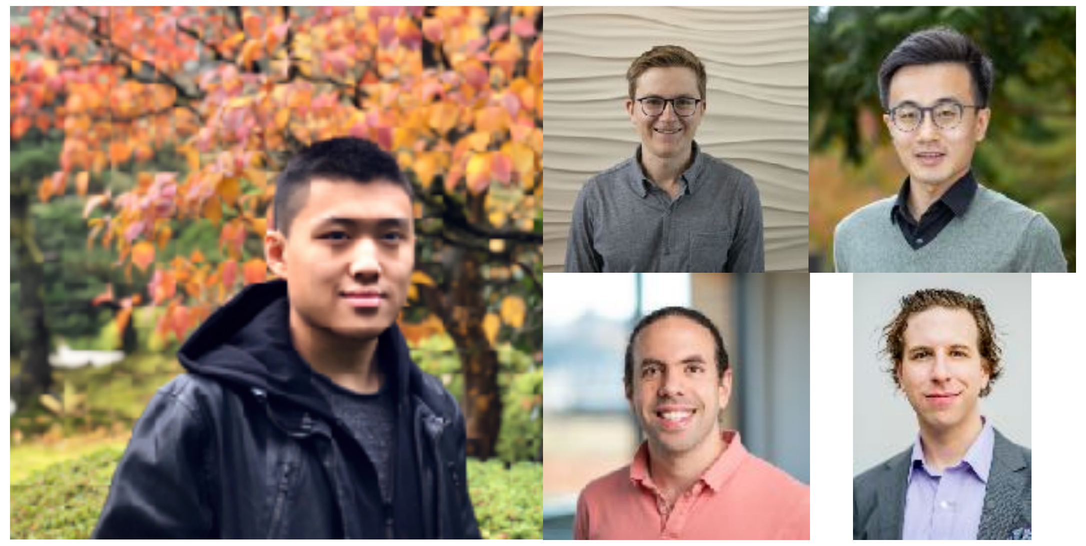
Summary

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.

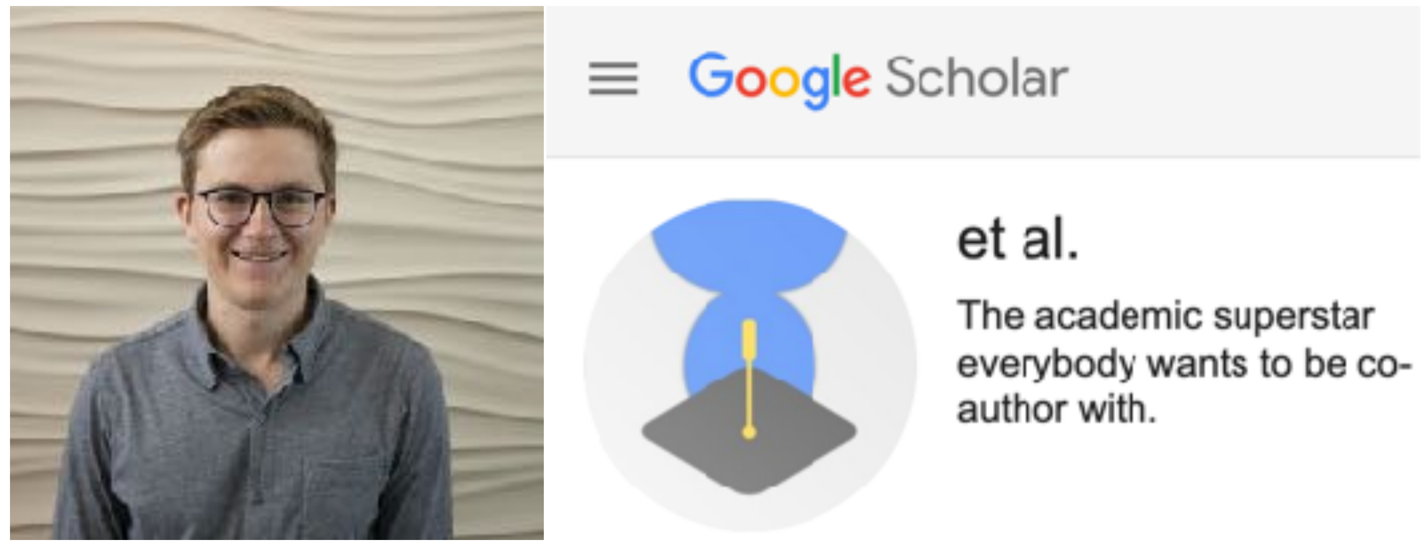
LMs can learn to consistently compare and evaluate the meaning of propositional logic expressions.



What About Other Languages?

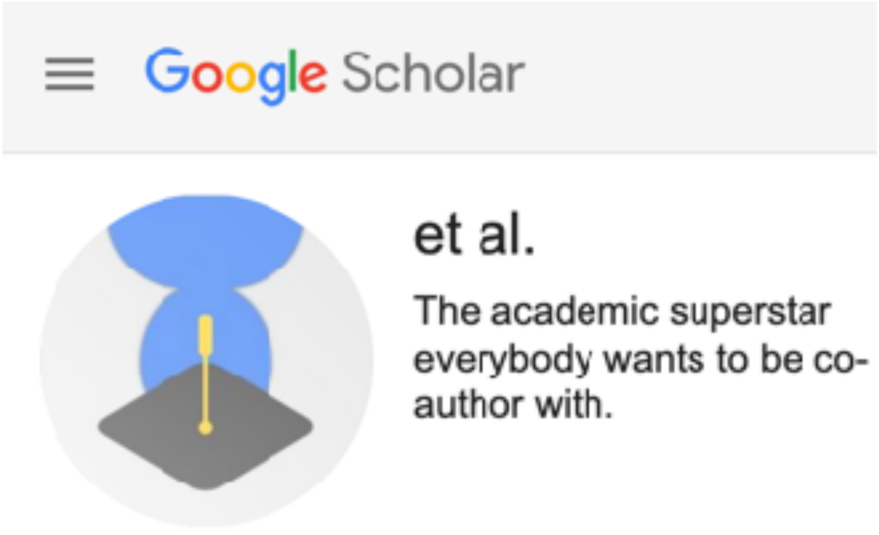
What About Other Languages?

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

$((T \wedge (F \vee F)) \vee (T \vee (F \wedge T)))$

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∧(F∨F))∨(T∨(F∧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

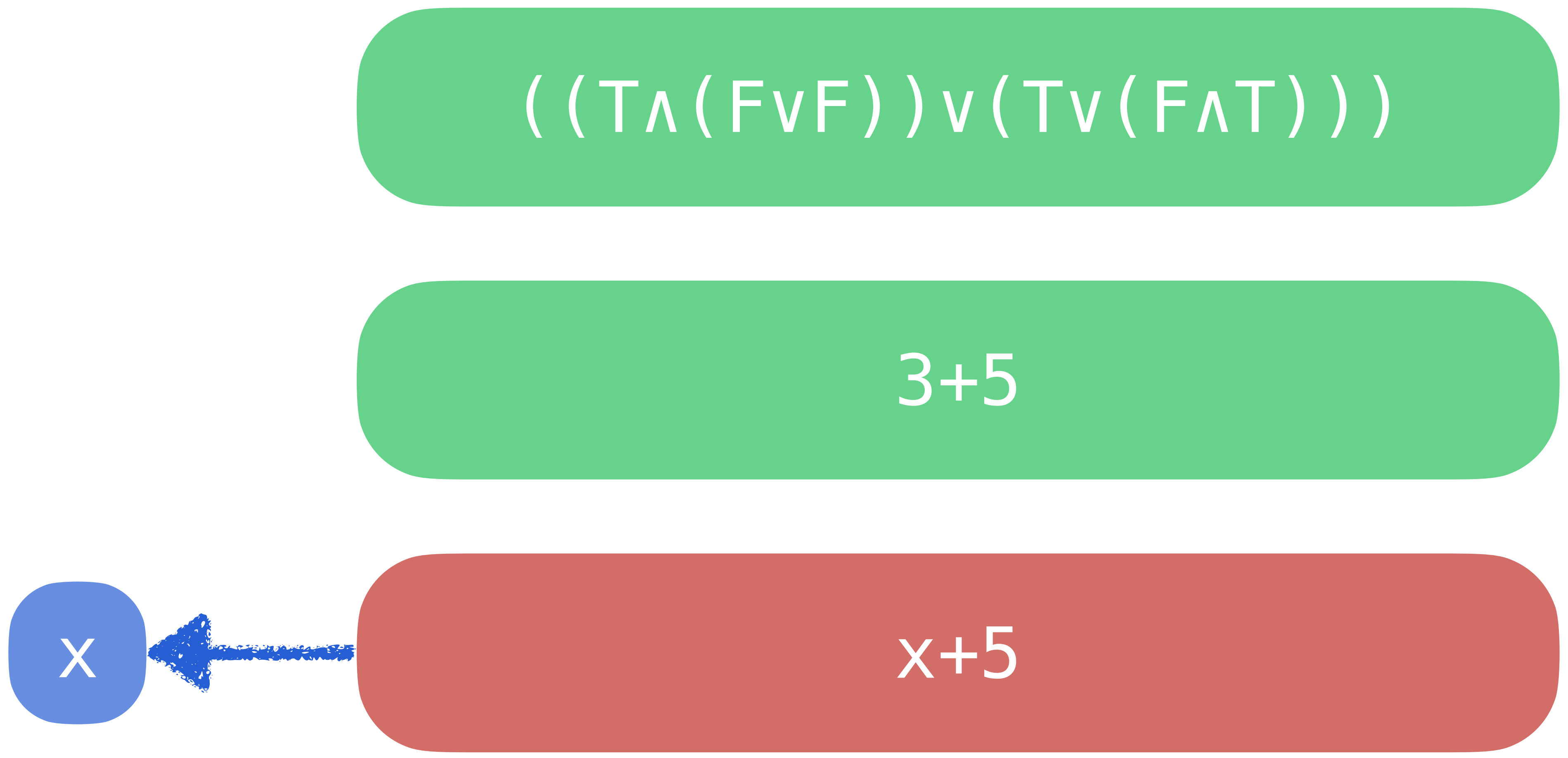
$((T \wedge (F \vee F)) \vee (T \vee (F \wedge T)))$

$3+5$

$x+5$

Strong Transparency
(i.e., context-independency)

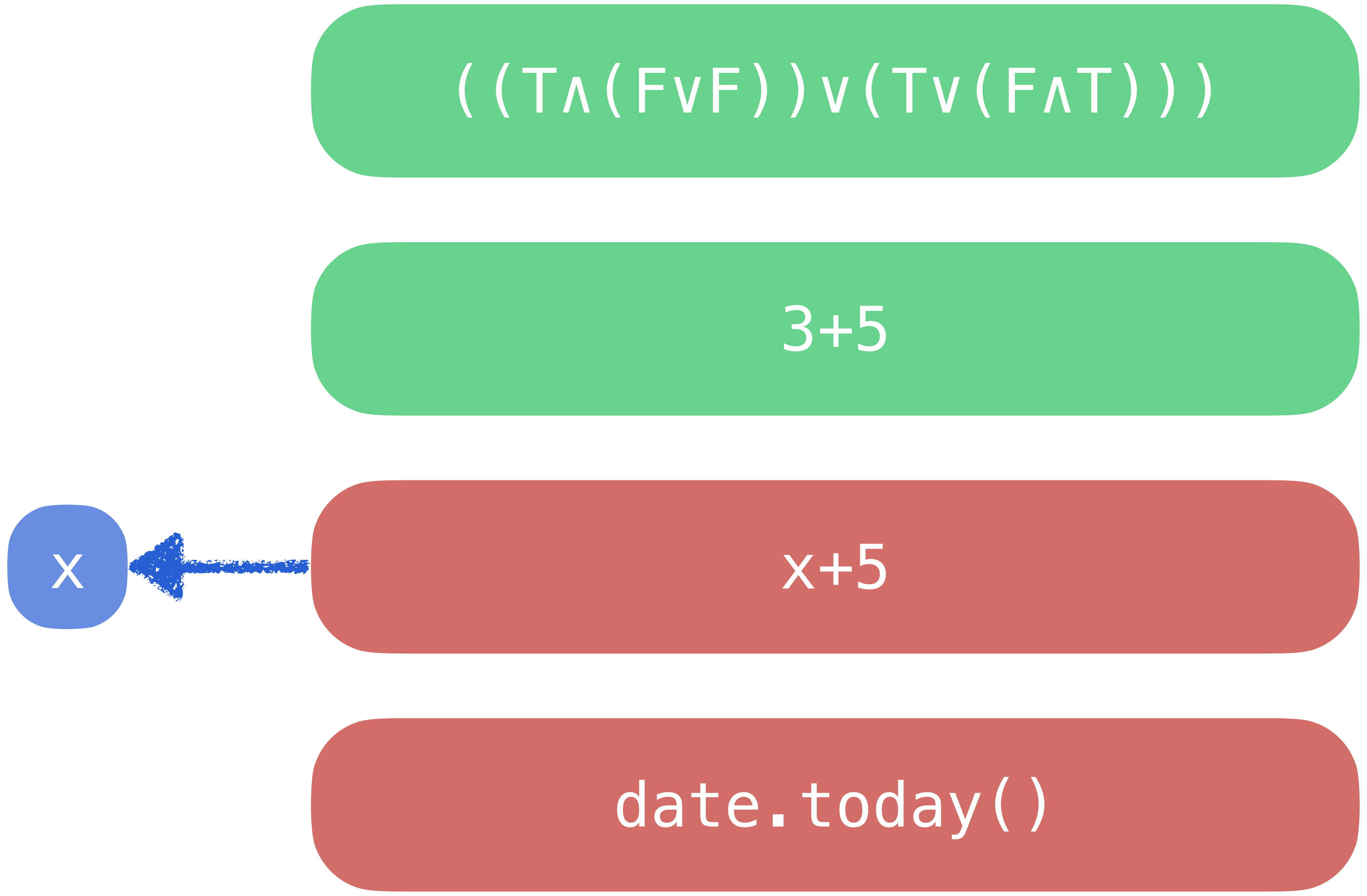
- An expression is strongly transparent if its meaning is context-independent
- A language is strongly transparent if all of its expressions are



Strong Transparency

(i.e., context-independency)

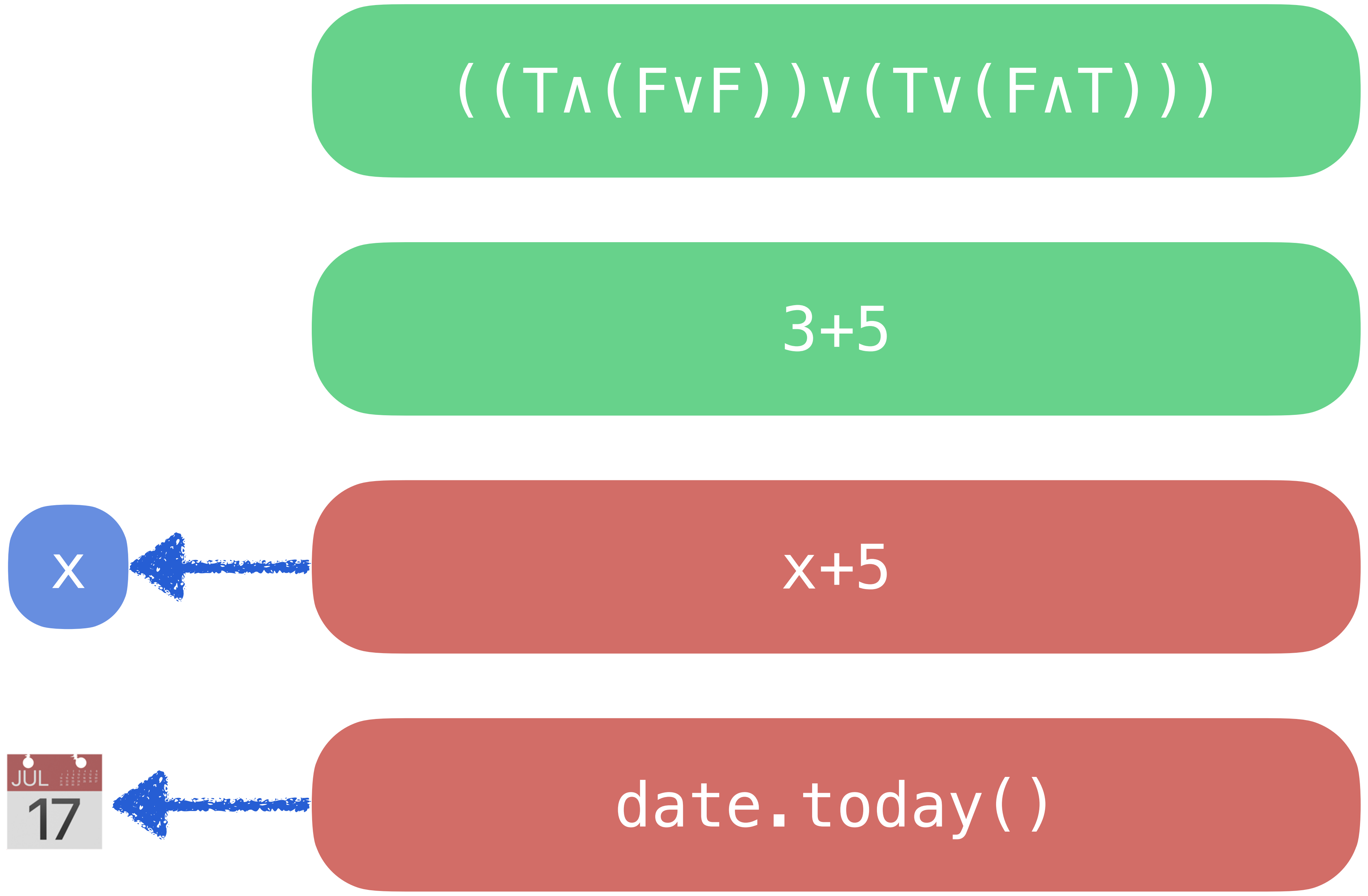
- An expression is strongly transparent if its meaning is context-independent
- A language is strongly transparent if all of its expressions are



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



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∧(F∨F))∨(T∨(F∧T)))`

`3+5`

`Some corgis run.`

x ← `x+5`

← `date.today()`

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∧(F∨F))∨(T∨(F∧T)))`

`3+5`

x ← `x+5`

← `date.today()`

`Some corgis run.`

`His corgis run.`

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∧(F∨F))∨(T∨(F∧T)))`

`3+5`

x ← `x+5`

← `date.today()`

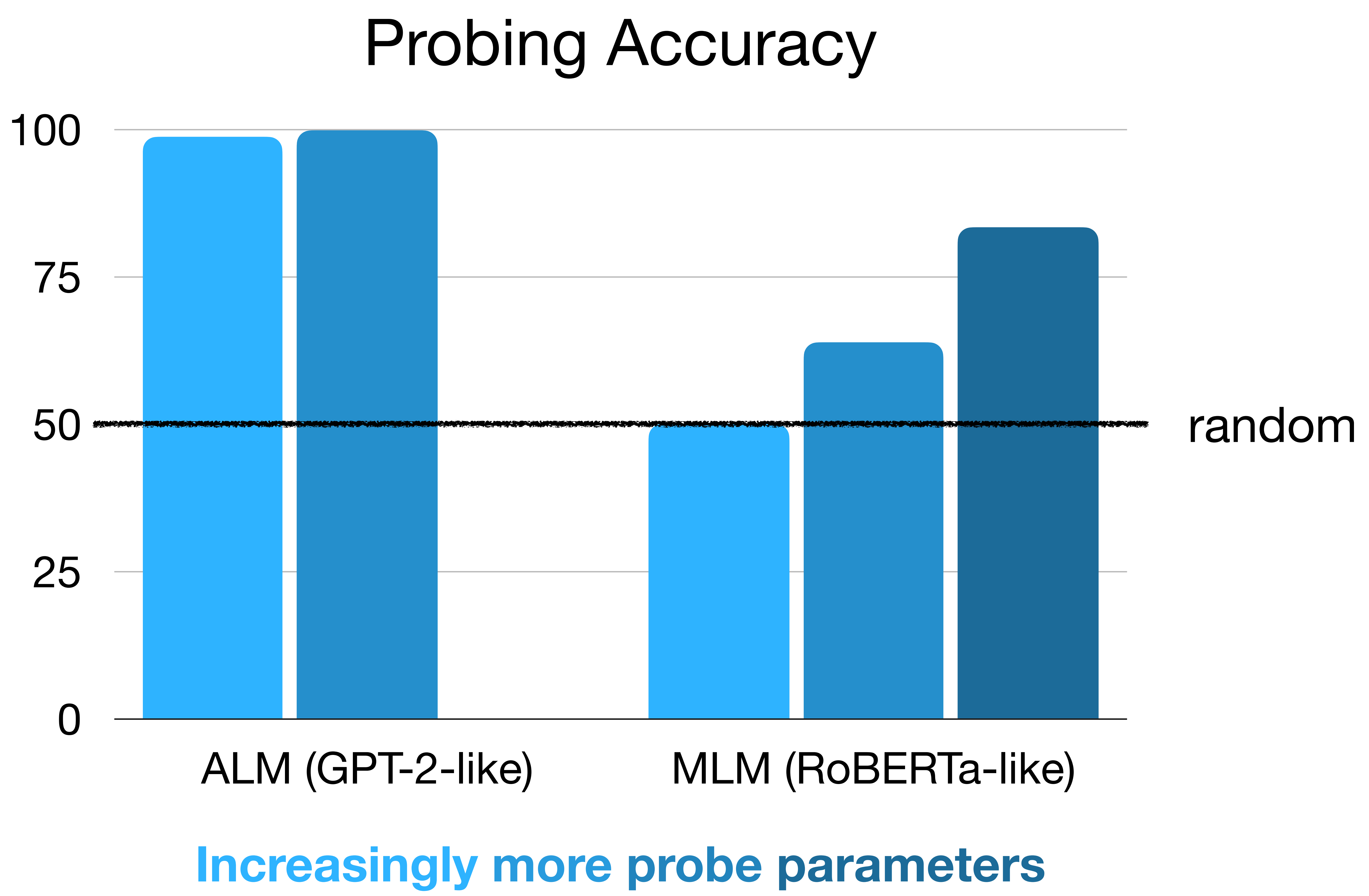
`Some corgis run.`

`His corgis run.`

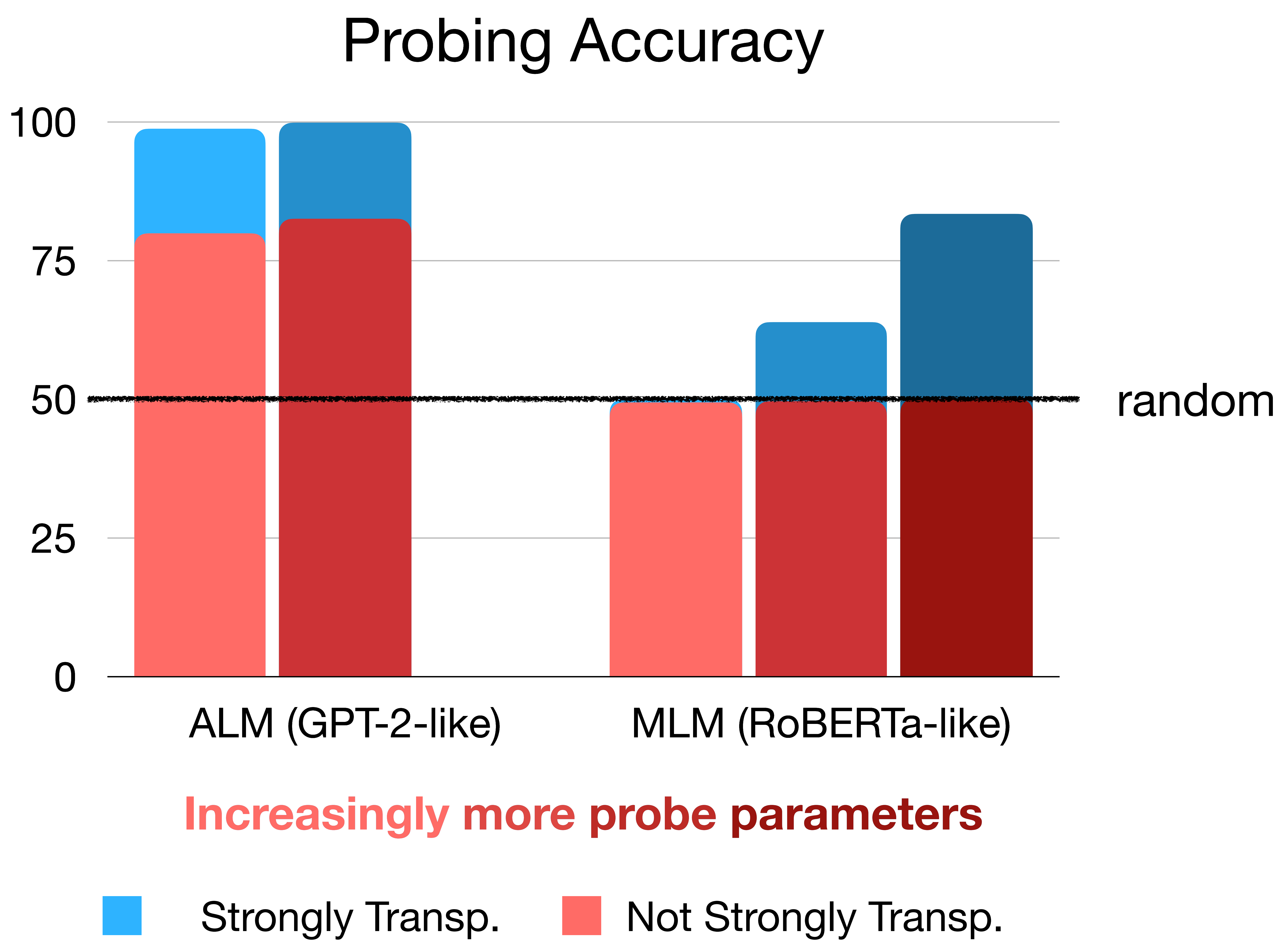
`Today, some corgis ran.`

Removing Strong Transparency

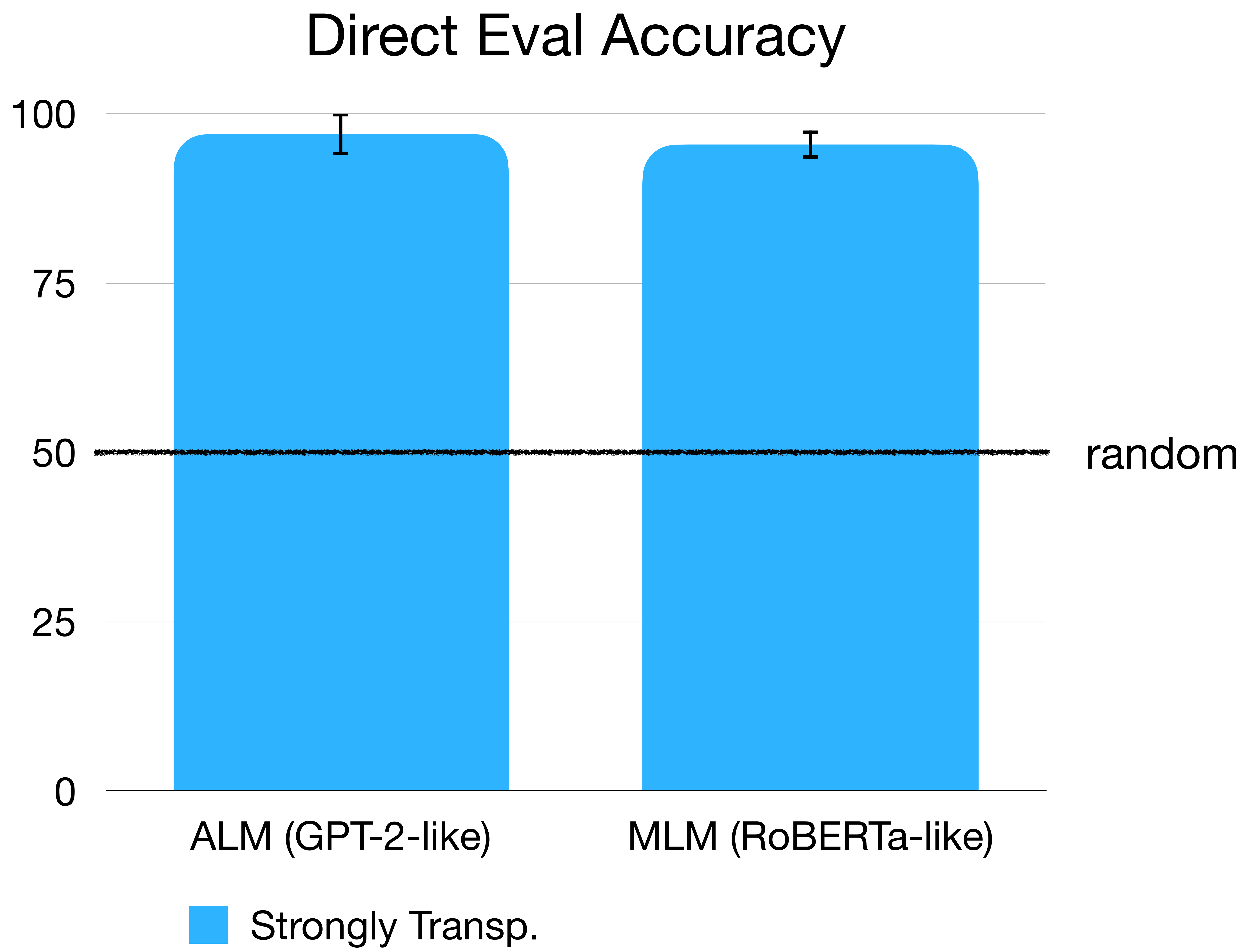
Removing Strong Transparency



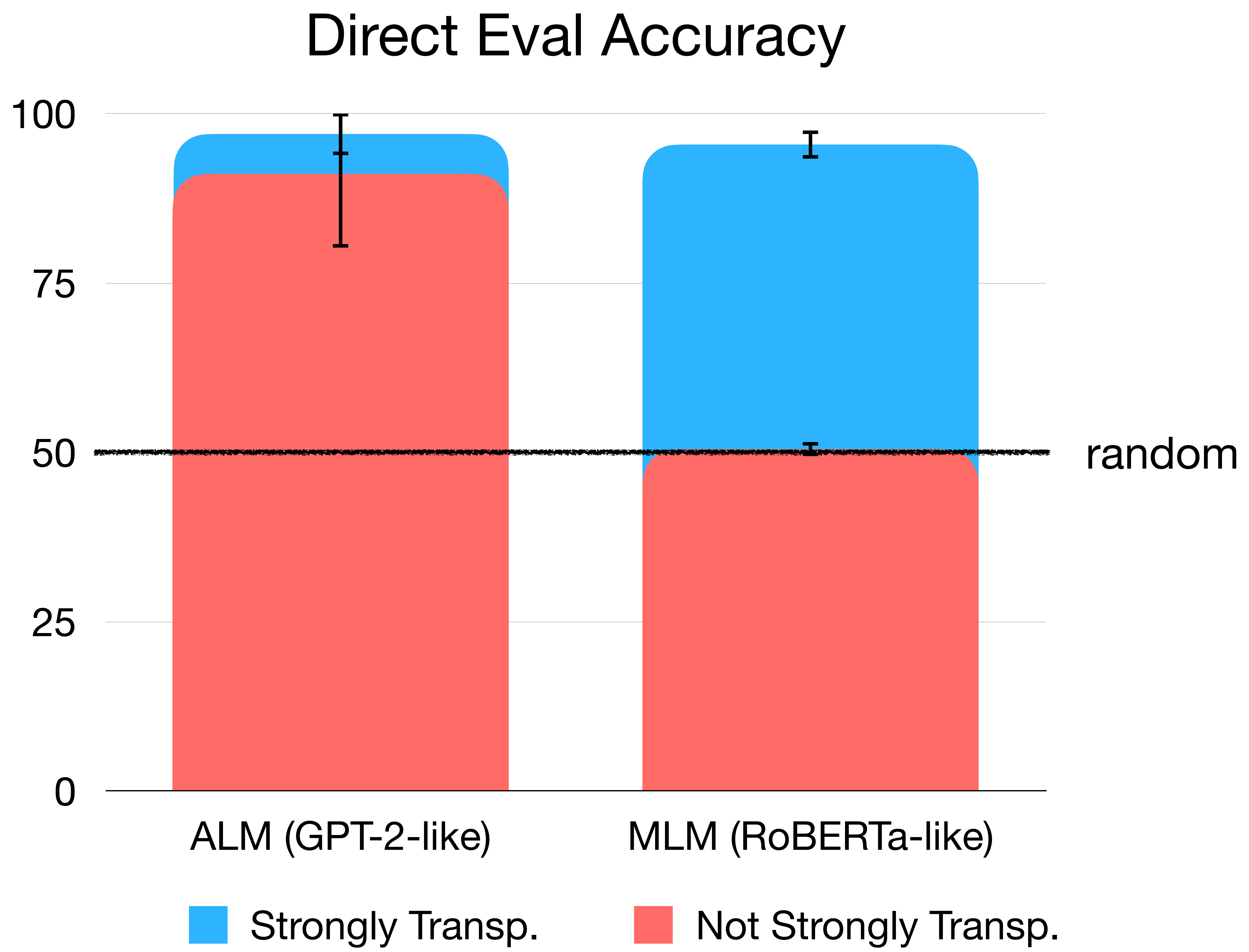
Removing Strong Transparency



Removing Strong Transparency

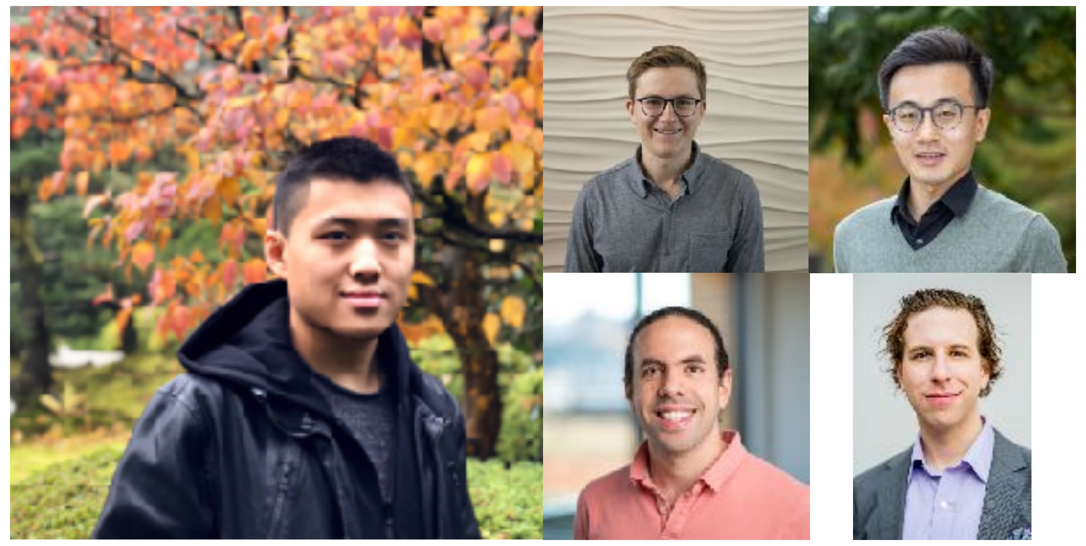


Removing Strong Transparency



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

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

Referential Opacity

Foreshadow: it makes NL not strongly transparent

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

Referential Opacity

Foreshadow: it makes NL not strongly transparent

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

$[\![\text{Lois Lane believes Superman is a hero.}]\!]$

$\shortparallel$

T

Referential Opacity

Foreshadow: it makes NL not strongly transparent

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

$[\![\text{Lois Lane believes Superman is a hero.}]\!]$ $=$ T

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

Referential Opacity

Foreshadow: it makes NL not strongly transparent

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

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

$\shortparallel$ $\qquad\qquad$ $\shortparallel$

T $\qquad\qquad$ F

Referential Opacity

Foreshadow: it makes NL not strongly transparent

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

propositional attitude verb

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

$||$ $\qquad$ $||$

T $\qquad$ F

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}. $y =$ Non-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

Setup

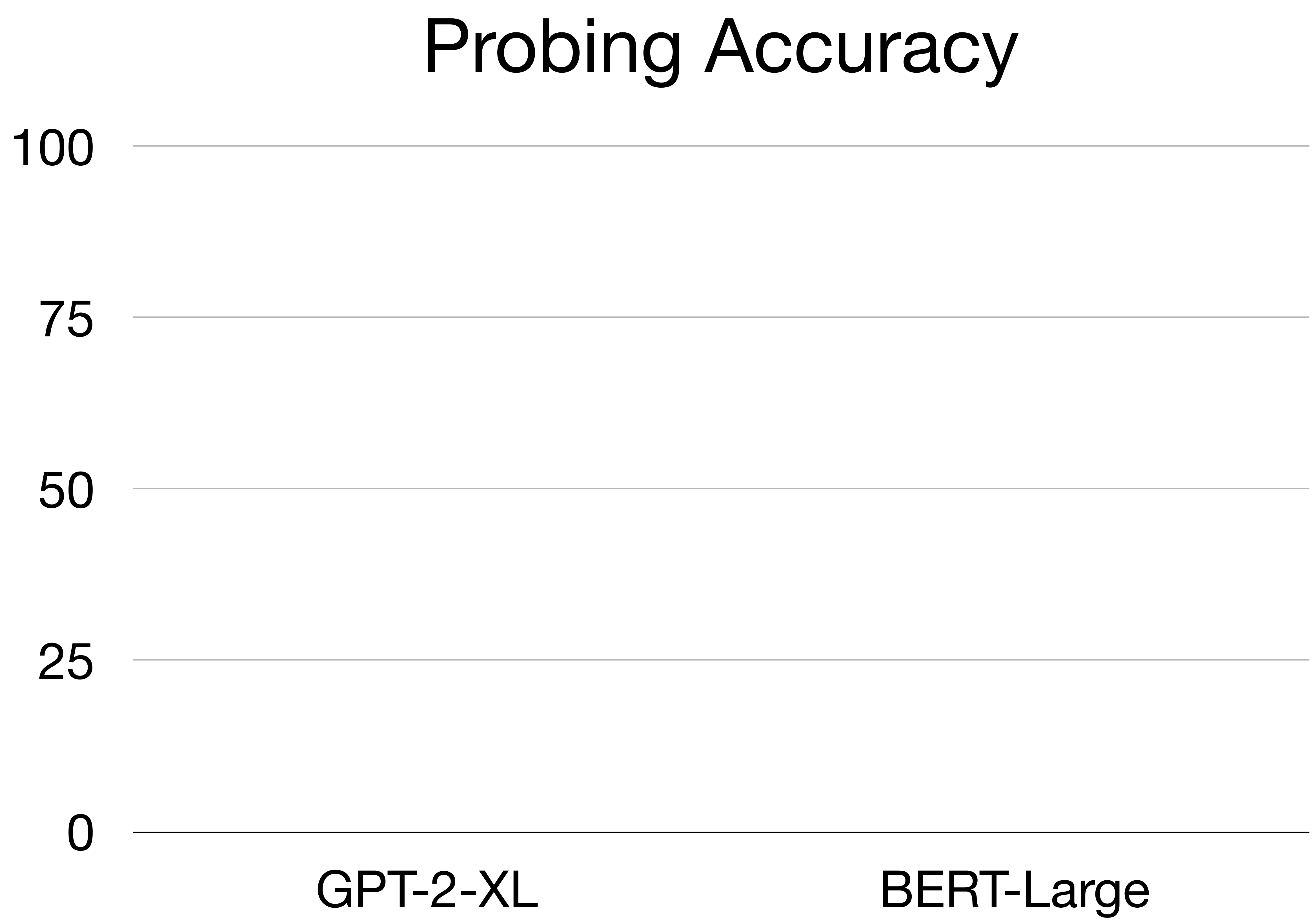
- **Data**: $\{(s_1, s_2, y)\}$
  - She wants to meet {Superman/Clark Kent}. $y =$ Non-equivalent
  - She managed to meet {Superman/Clark Kent}. $y =$ Equivalent
- **Models**: pretrained GPT-2-XL, BERT-large

Setup

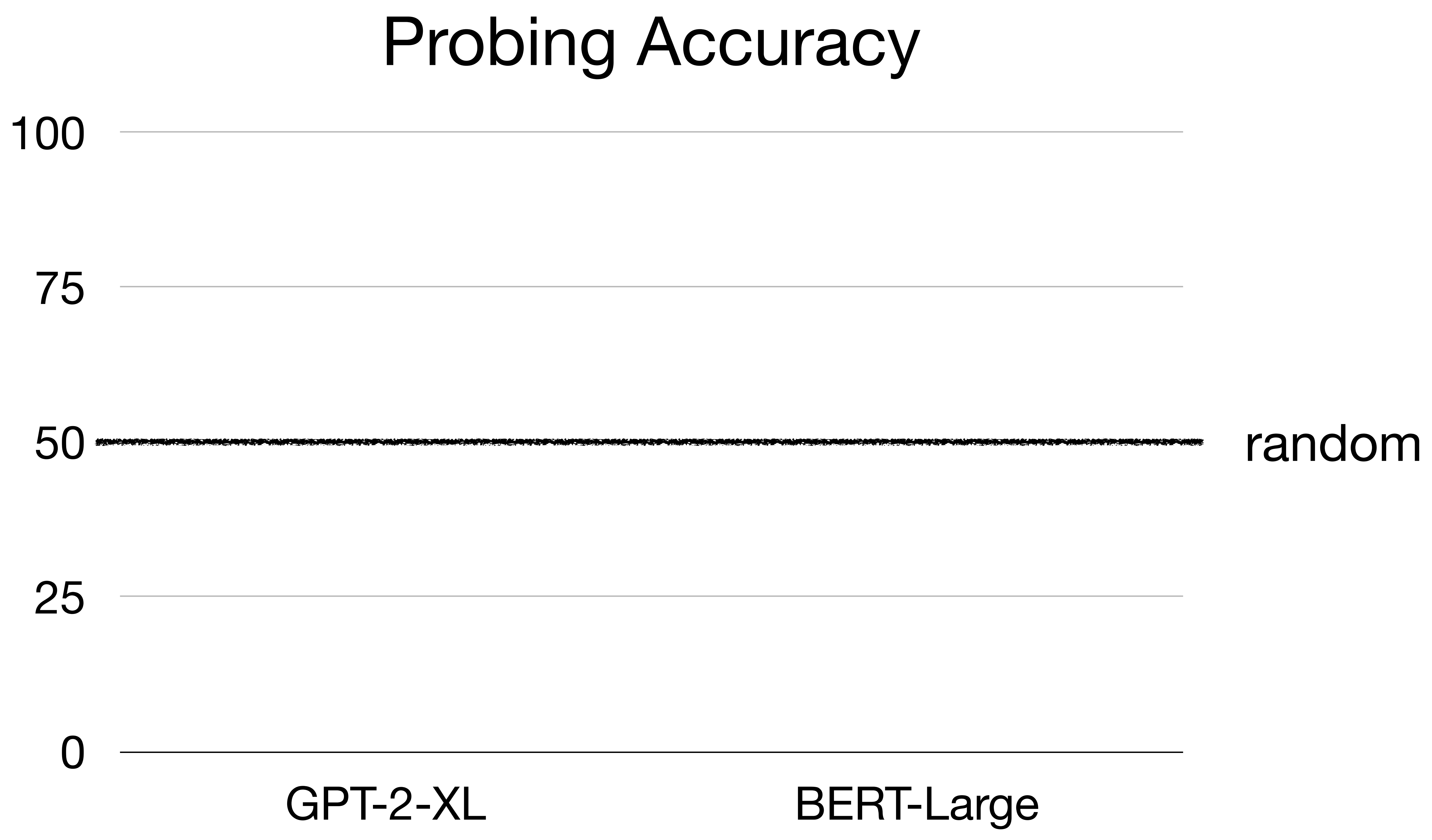
- **Data**: $\{(s_1, s_2, y)\}$
  - She wants to meet {Superman/Clark Kent}. $y =$ Non-equivalent
  - She managed to meet {Superman/Clark Kent}. $y =$ Equivalent
- **Models**: pretrained GPT-2-XL, BERT-large
- **Methods**: probing and similarity-based analysis

Probing Results

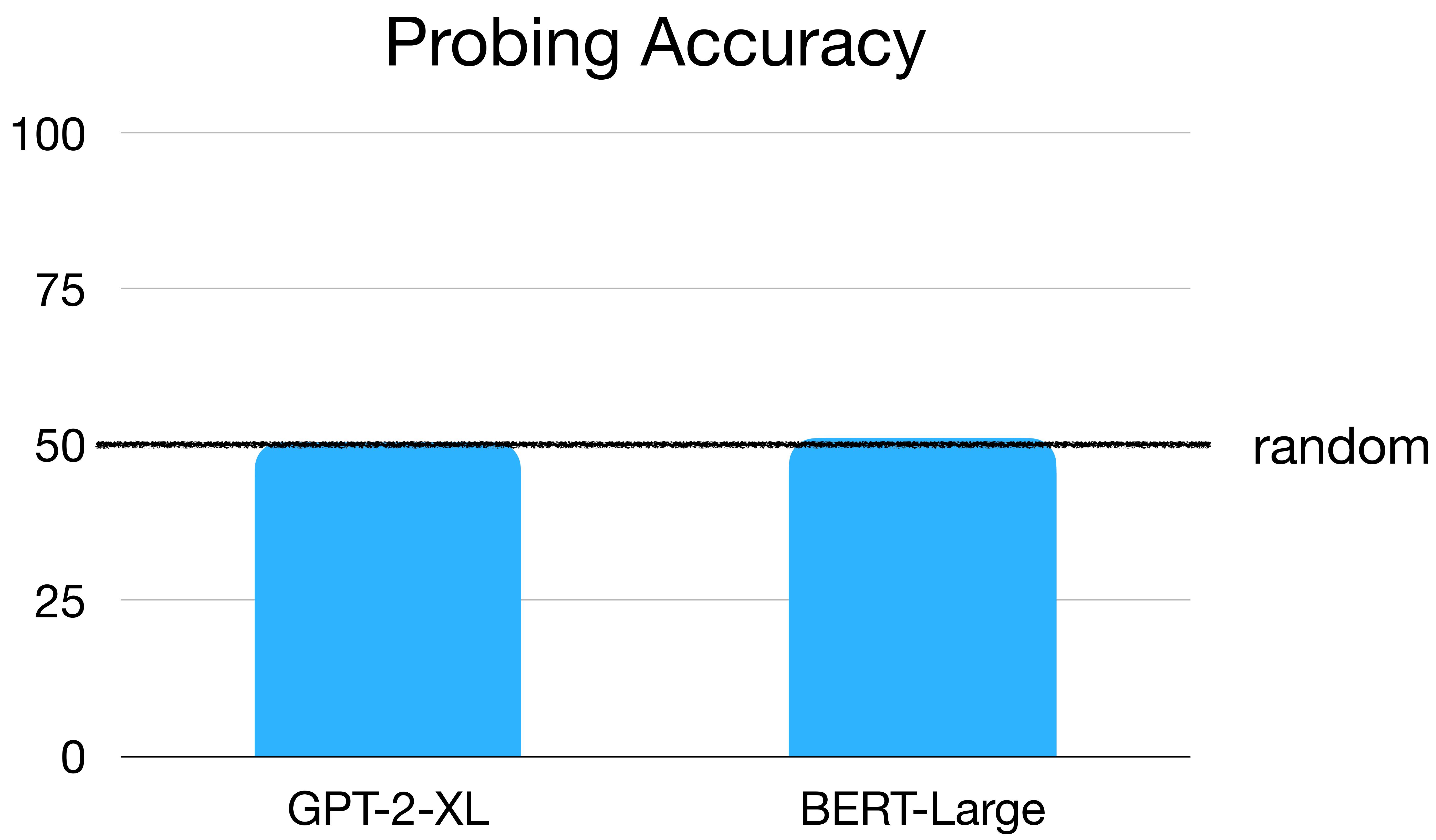
Probing Results



Probing Results

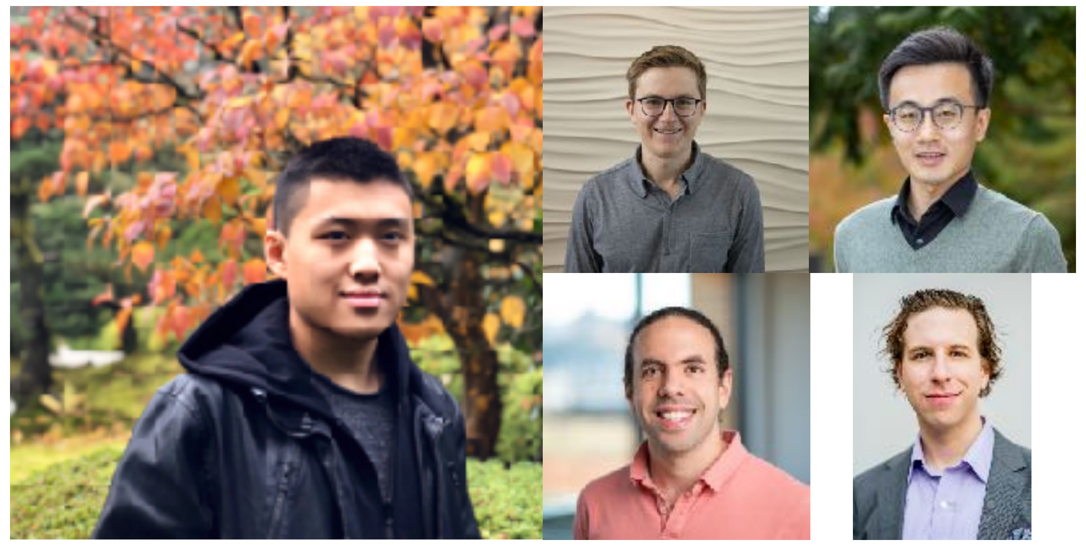


Probing Results



Yet Another Summary

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 =

```
((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T)))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))
```

Encore: Grounding =

```
((¬T)∧(¬(T∨(¬F))))=(T∨(¬(¬((¬T)∨(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T)))))=((T∧T)∧((¬F)∨(¬F)))
(((¬((¬(¬(¬(¬T))))∨T))∨T)∧(¬(¬T)))=((¬F)∨(¬(T∧(T∨T))))
((T∧(F∨F))∨(T∨(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))∨F))∨(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)∨F)∧F))=(F∧(¬(¬(F∨((¬(F∨(¬T)))∧T)))))
((T∨(¬(T∧(T∨(¬(F∨(¬F)))))))∨T)=(¬((¬(T∨(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))∨((¬F)∨(T∨(¬(T∨T))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(T∨T))∧(¬T)))))=(¬(¬(((¬T)∨F)∨(F∨(¬(¬F))))))
(F∧(F∧(¬((F∨F)∨(¬(¬T))))))=(¬((((¬(T∧T))∨(¬F))∨(¬T))∧(¬F)))
```

Probing Accuracy

a=b
50.5

Encore: Grounding =

```
((¬T)∧(¬(Tv(¬F))))=(Tv(¬(¬((¬T)v(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)v(¬F)))
(((¬((¬(¬(¬(¬T))))vT))vT)∧(¬(¬T)))=((¬F)v(¬(T∧(TvT))))
((T∧(FvF))v(Tv(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))vF))v(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)vF)∧F))=(F∧(¬(¬(Fv((¬(Fv(¬T)))∧T)))))
((Tv(¬(T∧(Tv(¬(Fv(¬F)))))))vT)=(¬((¬(Tv(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))v((¬F)v(Tv(¬(TvT))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(TvT))∧(¬T)))))=(¬(¬(((¬T)vF)v(Fv(¬(¬F))))))
(F∧(F∧(¬((FvF)v(¬(¬T))))))=(¬((((¬(T∧T))v(¬F))v(¬T))∧(¬F)))
```

Probing Accuracy

-Reflexivity	+Reflexivity
a=b 50.5	a=b, a=a, b=b 92.7

Encore: Grounding =

```
((¬T)∧(¬(Tv(¬F))))=(Tv(¬(¬((¬T)v(¬(¬F))))))
(¬(¬(¬(((F∧((F∧F)∧F))∧F)∧(¬T))))))=((T∧T)∧((¬F)v(¬F)))
(((¬((¬(¬(¬(¬T))))vT))vT)∧(¬(¬T)))=((¬F)v(¬(T∧(TvT))))
((T∧(FvF))v(Tv(F∧T)))=(¬((¬T)∧(¬((¬((¬(¬F))vF))v(T∧T)))))
(((¬(¬F))∧(¬F))∧(((¬F)vF)∧F))=(F∧(¬(¬(Fv((¬(Fv(¬T)))∧T)))))
((Tv(¬(T∧(Tv(¬(Fv(¬F)))))))vT)=(¬((¬(Tv(¬(¬(¬(T∧F))))))∧F))
(¬((¬(¬F))v((¬F)v(Tv(¬(TvT))))))=(((F∧(¬T))∧((¬F)∧F))∧F)
(F∧(F∧(¬((¬(TvT))∧(¬T)))))=(¬(¬(((¬T)vF)v(Fv(¬(¬F))))))
(F∧(F∧(¬((FvF)v(¬(¬T))))))=(¬((((¬(T∧T))v(¬F))v(¬T))∧(¬F)))
```

Probing Accuracy

	-Reflexivity	+Reflexivity
-Symmetry	a=b 50.5	a=b, a=a, b=b 92.7
+Symmetry	a=b, b=a 50.3	a=b, b=a, a=a, b=b 98.8

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

<EOSlides>

- 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