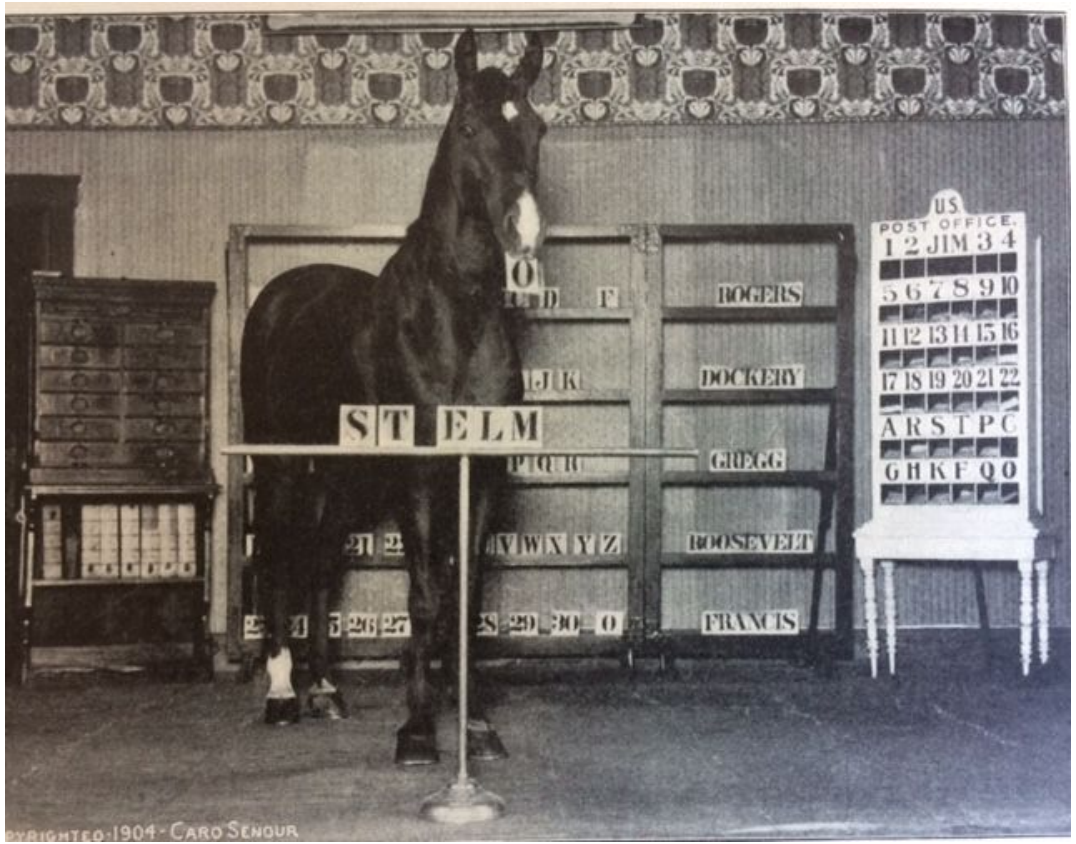

ELMER'S
Glue-All
Multi-Purpose Glue
Bonds Strong
For All Your Needs
Safe, Non-Toxic
4 FL OZ (118 mL)


ROGERS
DOCKERY
GRIGG
ROOSEVELT
FRANCIS
S T E L M
U.S. POST OFFICE.
1 2 JIM 3 4
5 6 7 8 9 10
11 12 13 14 15 16
17 18 19 20 21 22
A R S T P C
G H K F Q O
COPYRIGHTED 1904 - CARO SENOUR

CLEVER BUT NOT SMART

Hans the Horse, Goodhart's Law, and the Search for Better Benchmarks

Hannah MacKenzie-Margulies
Product Manager — Text Analytics, Basis Technology

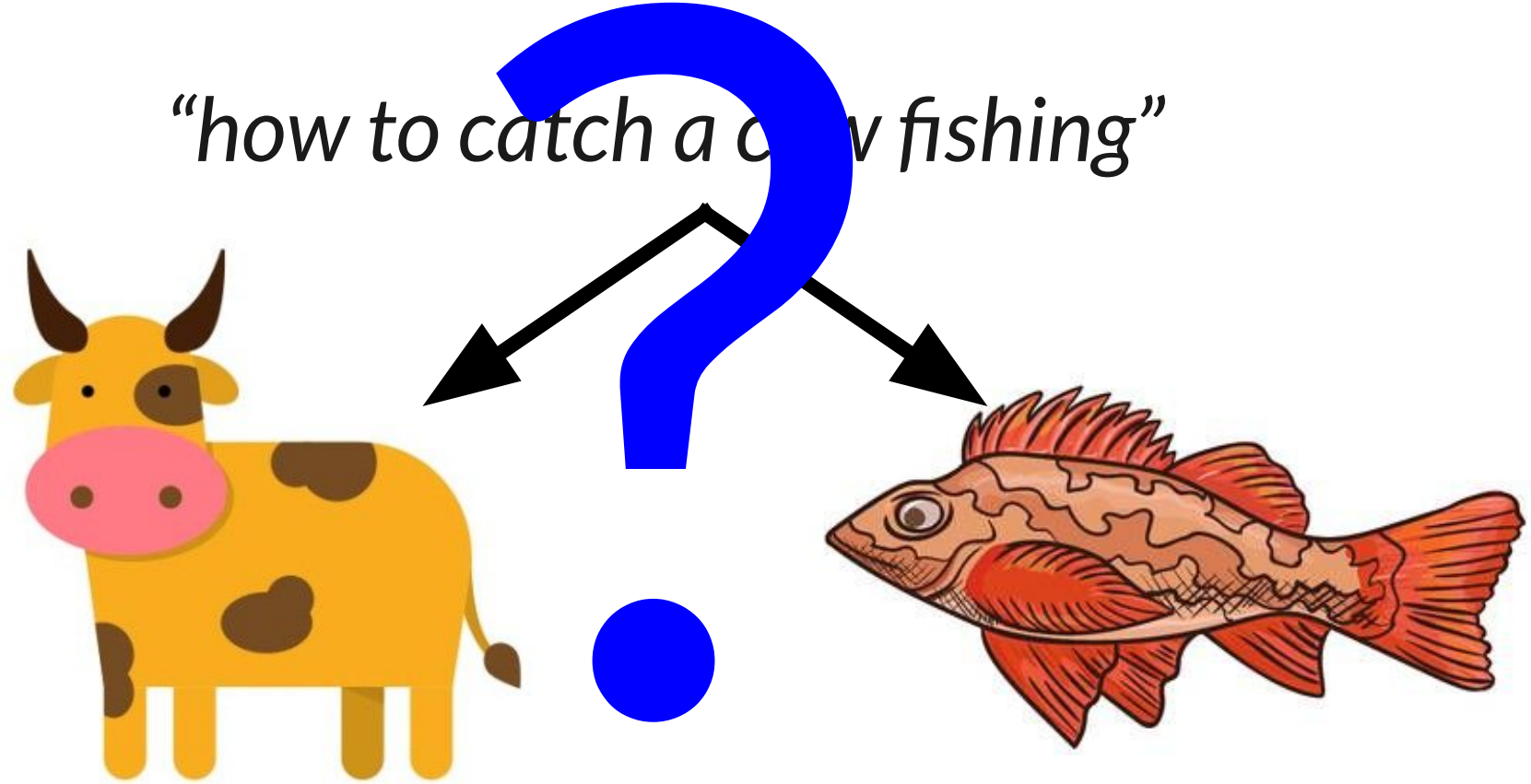

ELMER'S
Glue-All
Multi-Purpose Glue
Bonds Strong
For All Your Needs
Safe, Non-Toxic
4 FL OZ (118 mL)

Attention

Deep pre-trained language model

Bi-Directionality

“how to catch a c w fishing”

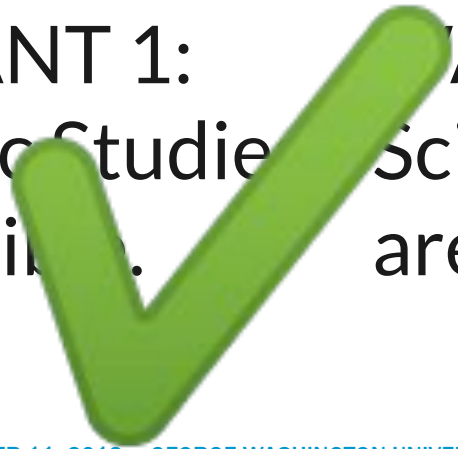


CLAIM: Smoking causes cancer.

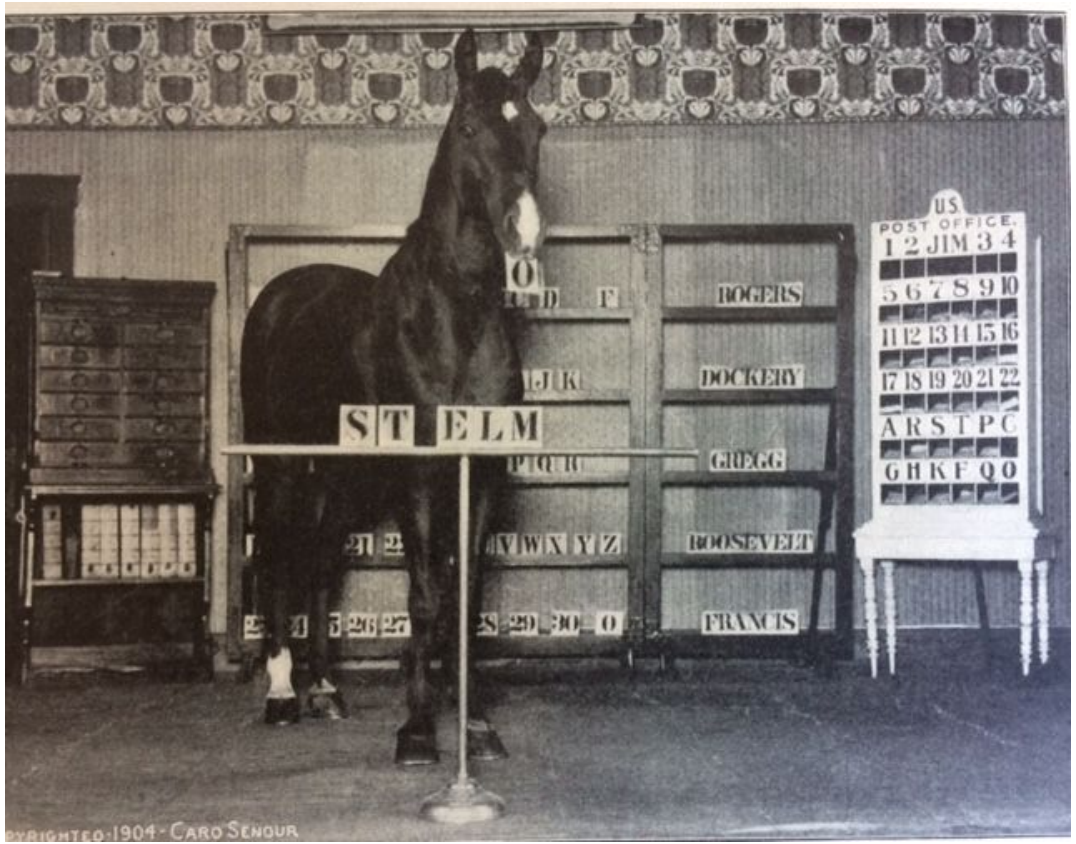
REASON: Scientific studies have shown a link between smoking and cancer.

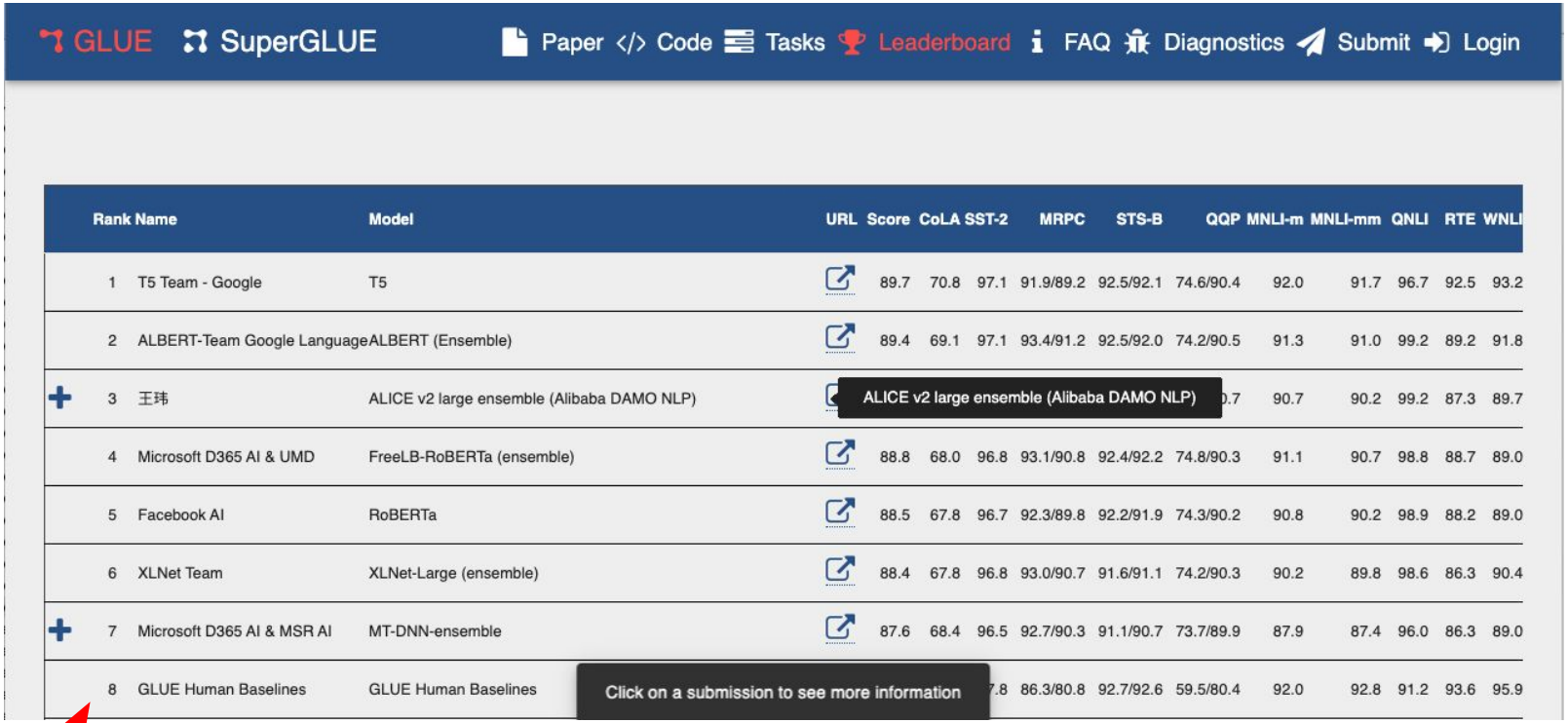
Choose one the following:

WARRANT 1: Scientific Studies are credible.



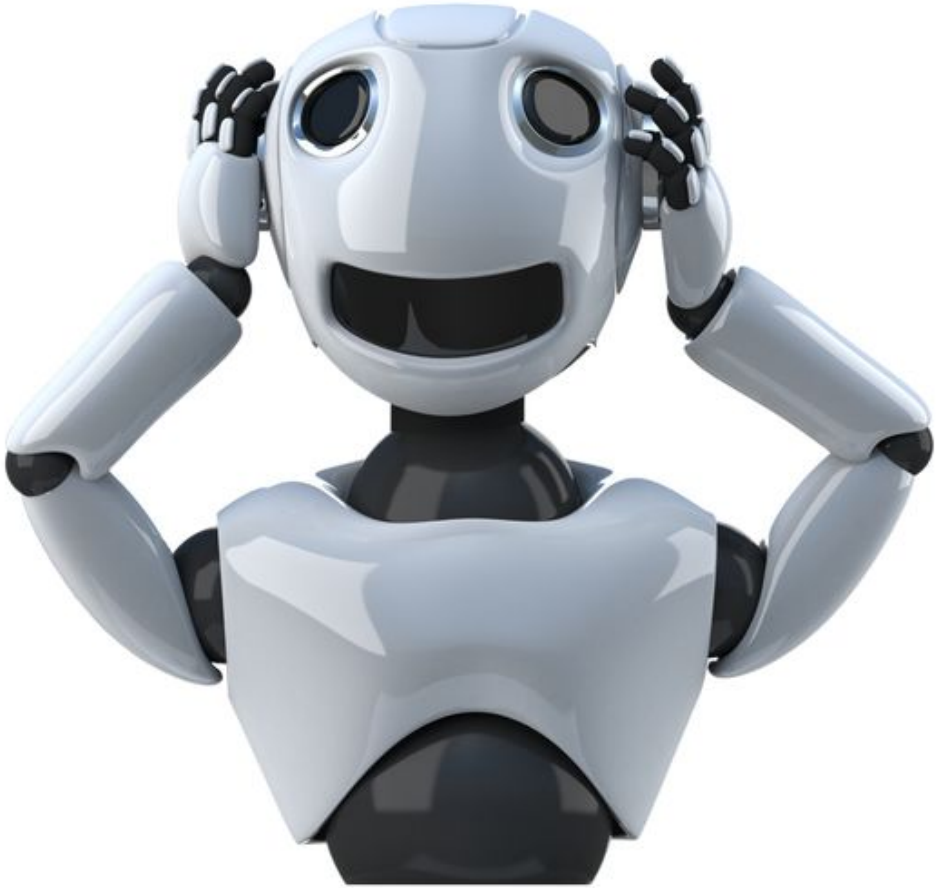
WARRANT 2: Scientific Studies are expensive.

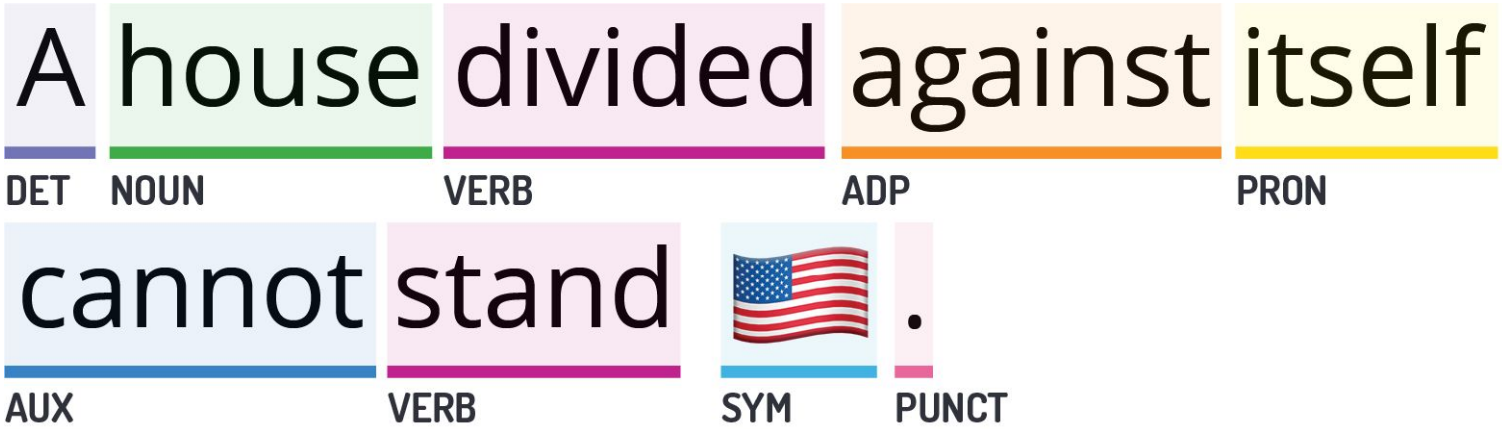

ROGERS
DOCKERY
GRIGGS
ROOSEVELT
FRANCIS
S T E L M
U.S. POST OFFICE.
1 2 JIM 3 4
5 6 7 8 9 10
11 12 13 14 15 16
17 18 19 20 21 22
A R S T P C
G H K F Q O
COPYRIGHTED 1904 - CARO SENOUR



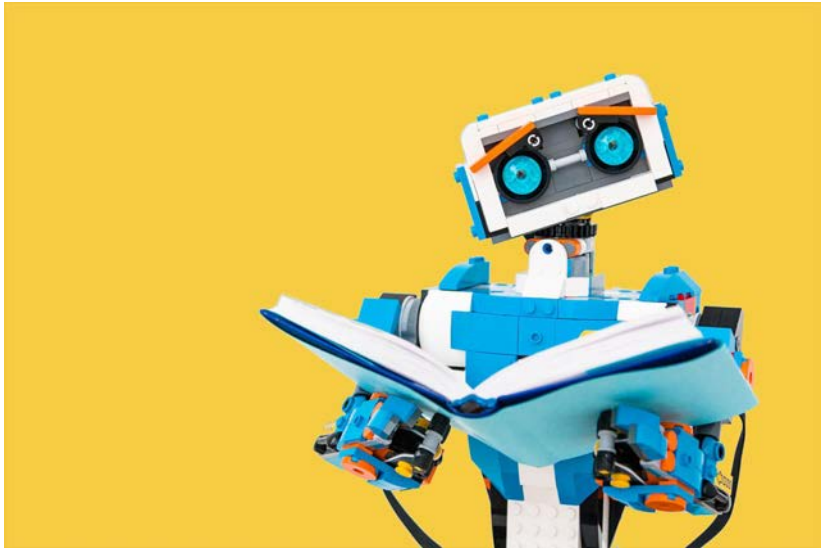
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI
1	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2
2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8
3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)							.7	90.7	90.2	99.2	87.3	89.7
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3	91.1	90.7	98.8	88.7	89.0
5	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	98.9	88.2	89.0
6	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3	90.2	89.8	98.6	86.3	90.4
7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0
8	GLUE Human Baselines	GLUE Human Baselines				.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9

“When a measure becomes a target, it ceases to be a good measure.

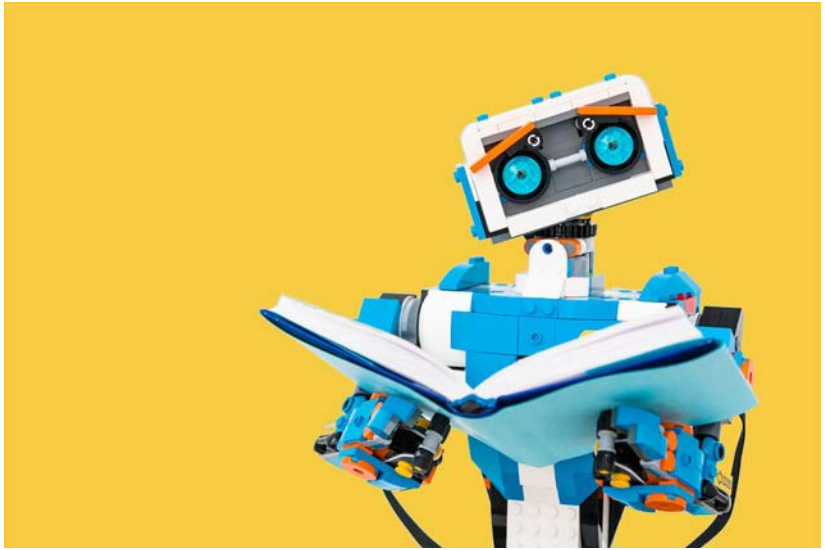


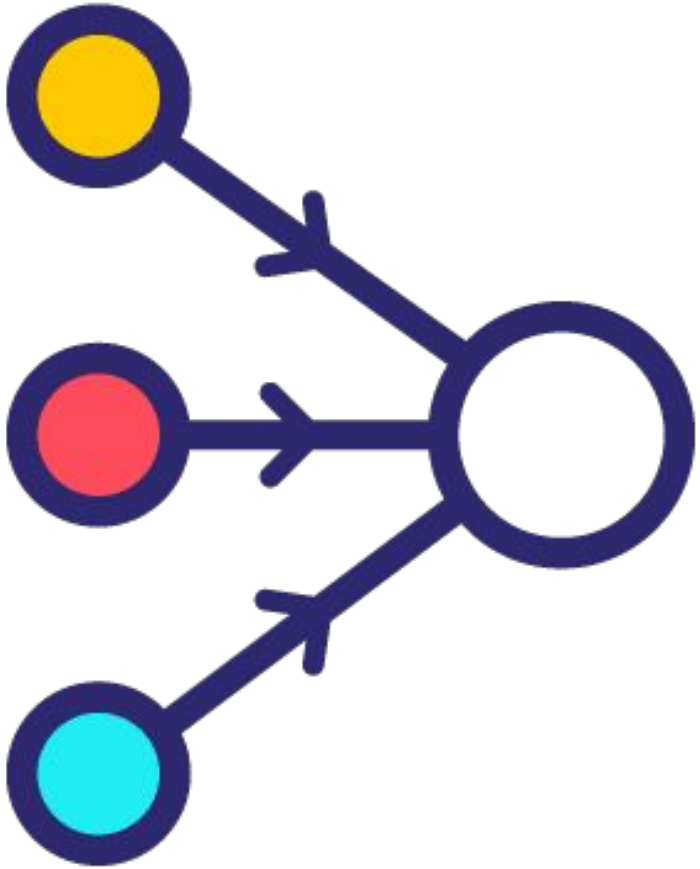

A house divided against itself
DET NOUN VERB ADP PRON
cannot stand 🇺🇸 .
AUX VERB SYM PUNCT

- Approach
- Benchmarks
- Data

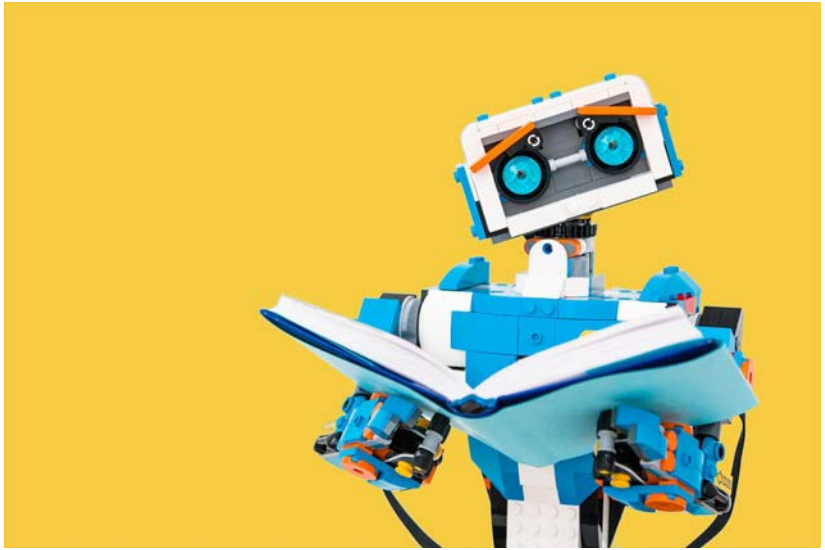


- Approach
- Benchmarks
- Data





- Approach
- Benchmarks
- Data




ELMER'S
Glue-All
Multi-Purpose Glue
Bonds Strong
For All Your Needs
Safe, Non-Toxic
4 FL OZ (118 mL)


SUPER GLUE
NET WT. 0.11 OZ. (3g)

SuperGLUE GLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

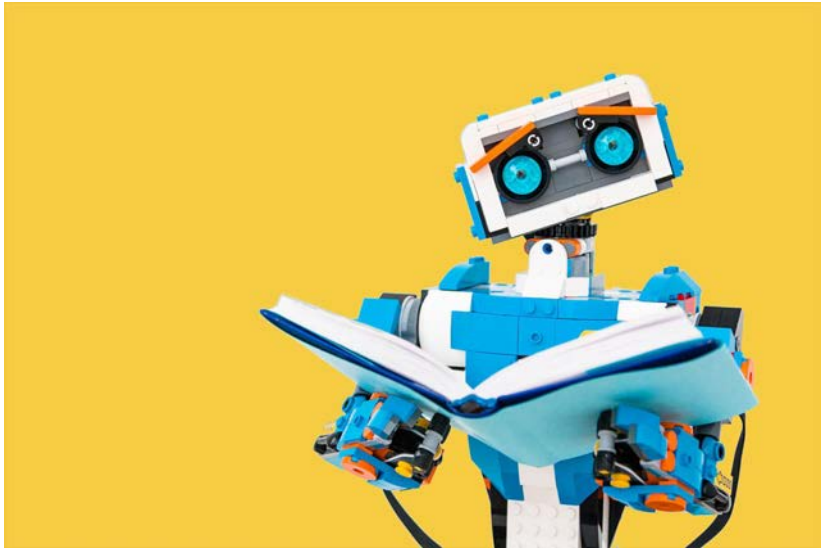
Leaderboard Version: **2.0**

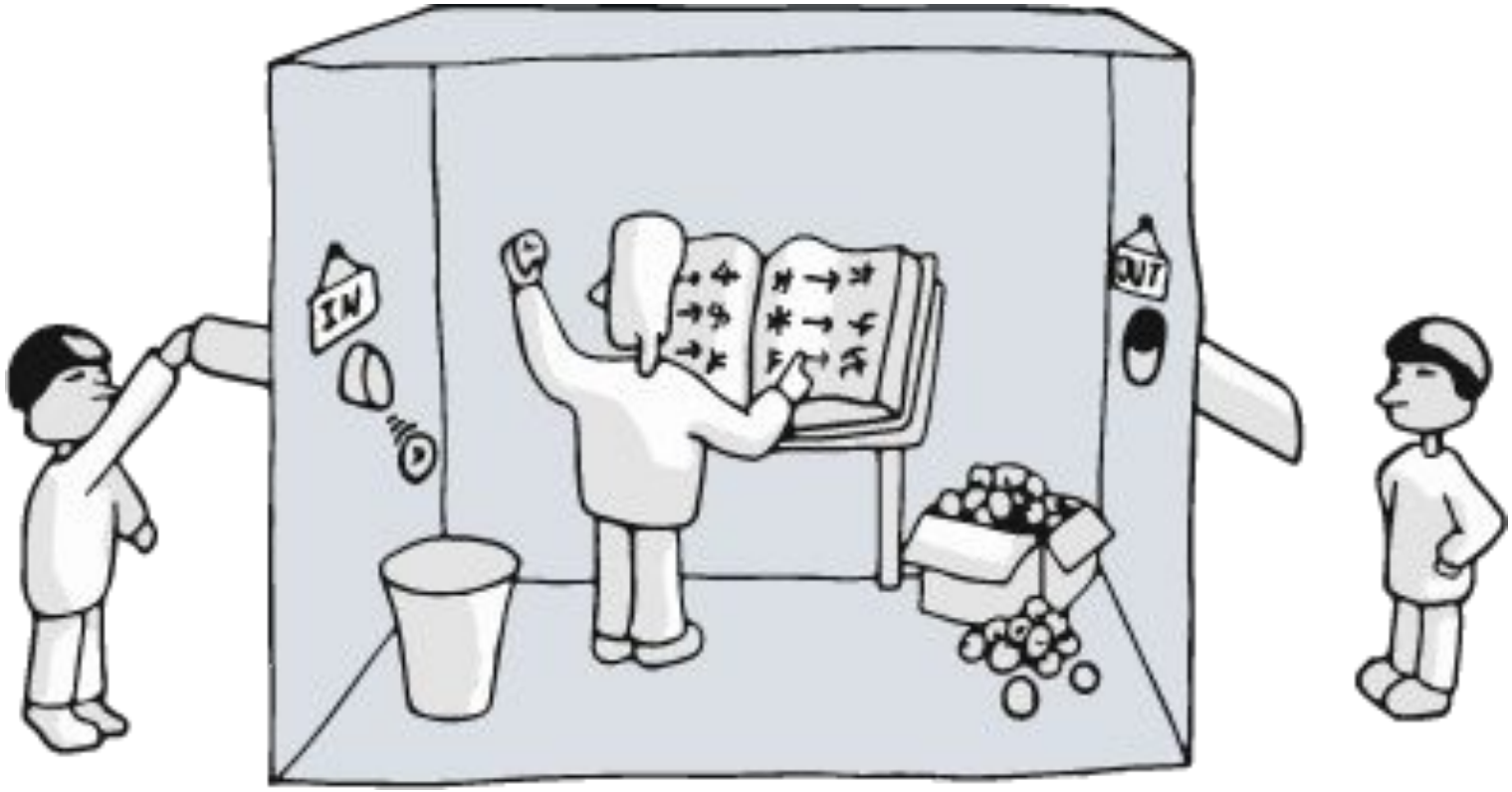
Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
2	T5 Team - Google	T5		88.9	91.0	93.0/96.4	94.8	88.2/62.3	93.3/92.5	92.5	76.1	93.8	65.6	92.7/91.9
3	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
4	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
5	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4

The trophy doesn't fit in the brown suitcase because it's too small.
What is too small?

– the trophy
– the suitcase 

- Approach
- Benchmarks
- Data




IN
OUT