Beyond Accuracy: Behavioral Testing of NLP models with **CheckList**

Marco Tulio Ribeiro @marcotcr
Microsoft Research

Tongshuang (Sherry) Wu @tongshuangwu
Carlos Guestrin @guestrin
University of Washington

Sameer Singh @sameer
University of California, Irvine

Motivation

How do I check if my model works?

How do I check if my model works?

OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

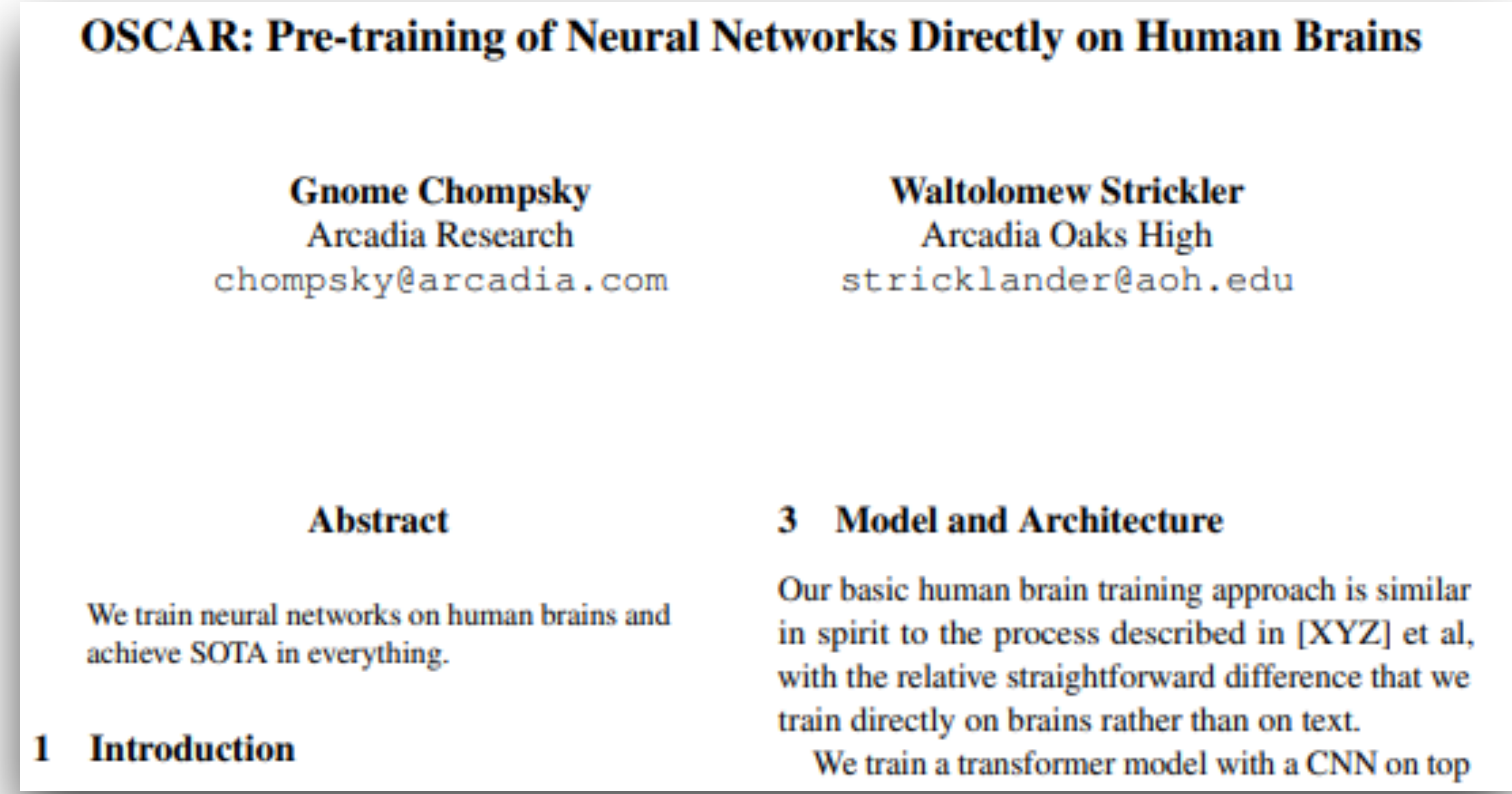
3 Model and Architecture

Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

How do I check if my model works?

Should I replace my doctor with OSCAR?



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
`chompsky@arcadia.com`

Waltolomew Strickler
Arcadia Oaks High
`stricklander@aoh.edu`

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

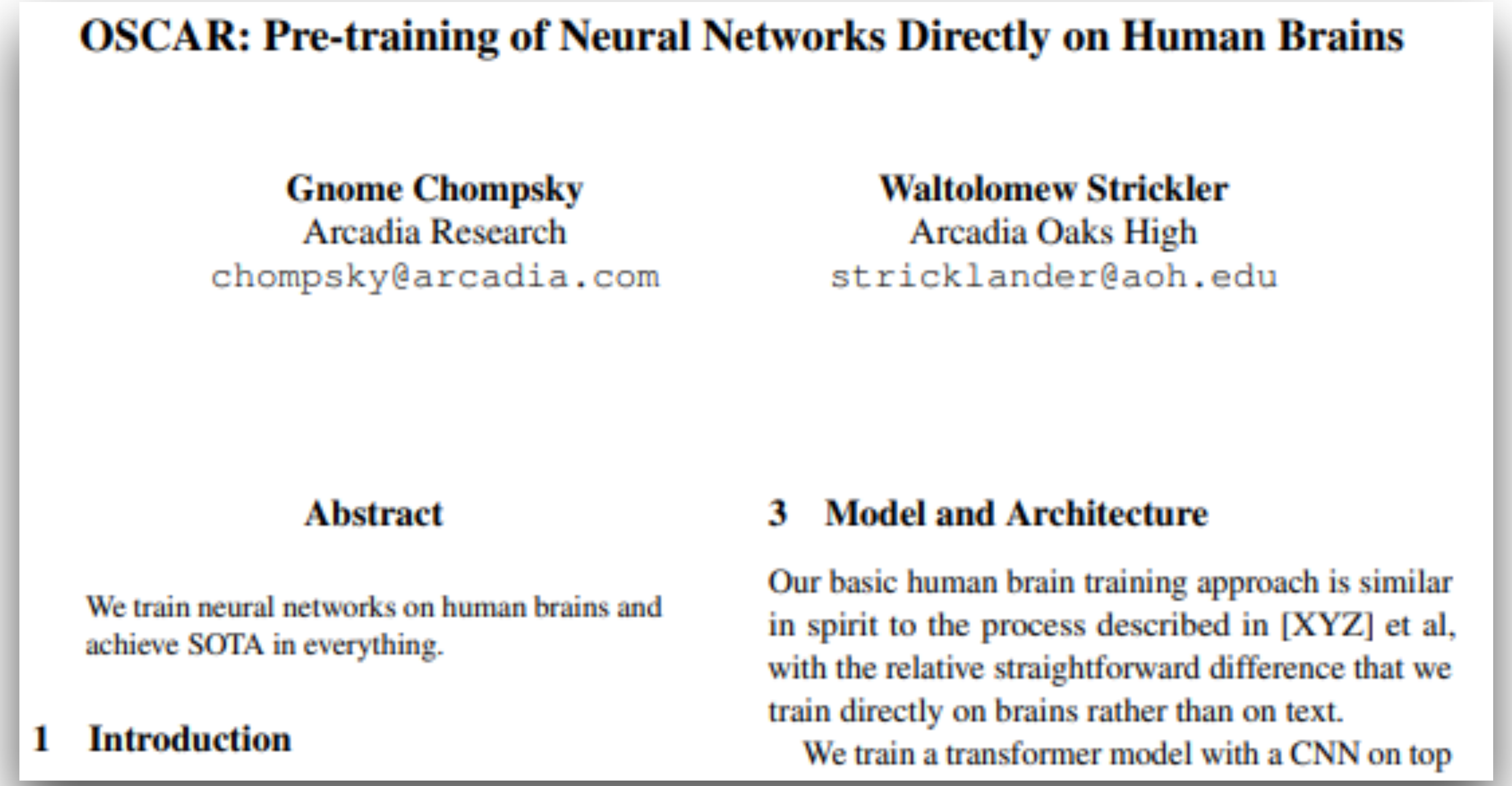
Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

How do I check if my model works?

Should I replace my doctor with OSCAR?

Should we use OSCAR in our products?



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

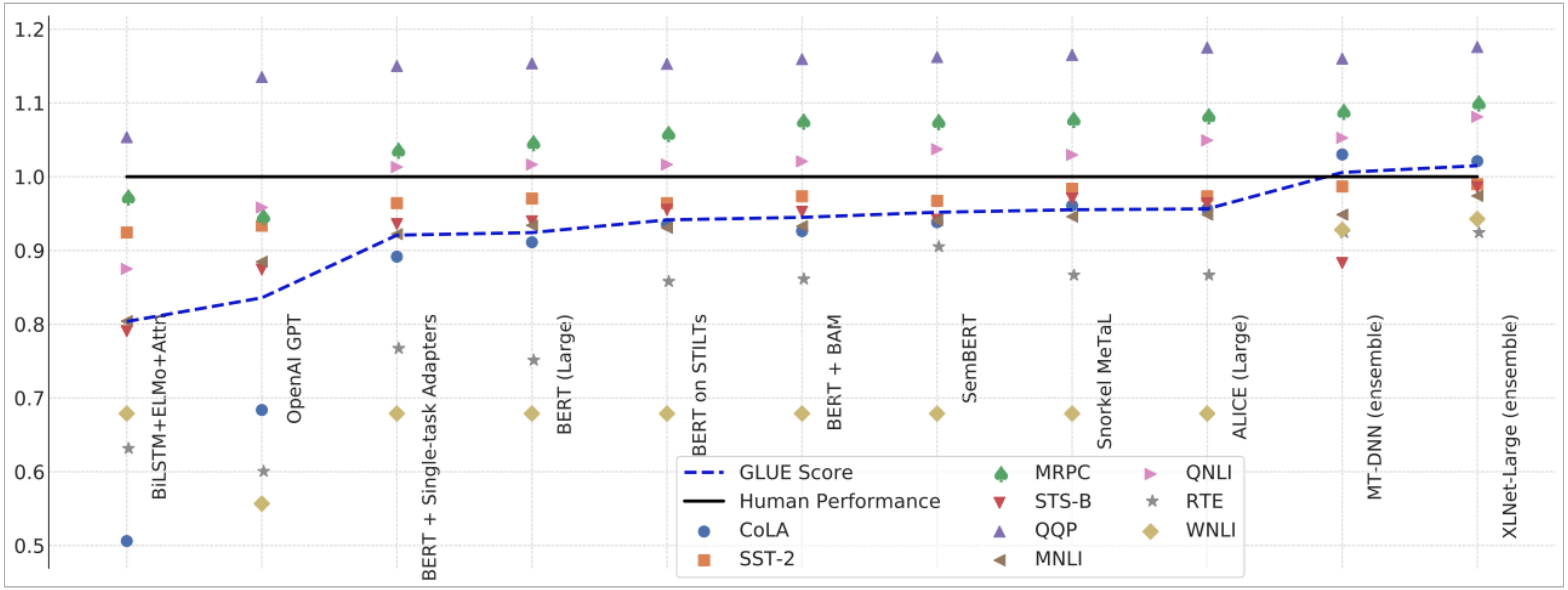
3 Model and Architecture

Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

Accuracy seems a good solution?

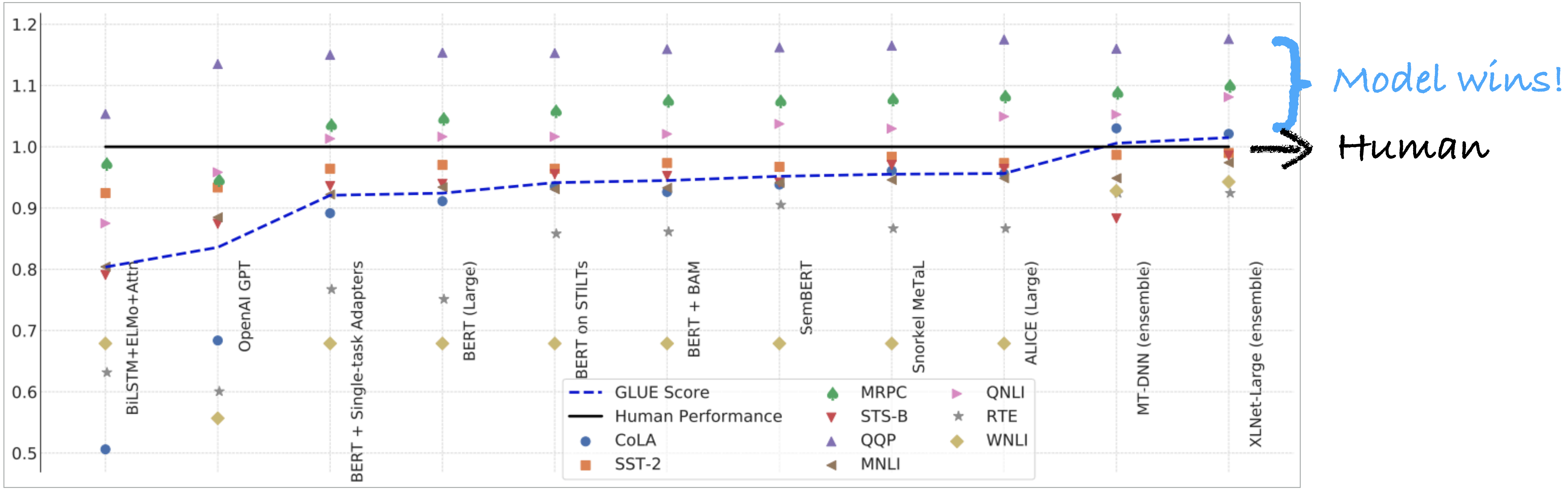
GLUE: "performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research."



Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems (pp. 3266-3280).

Accuracy seems a good solution?

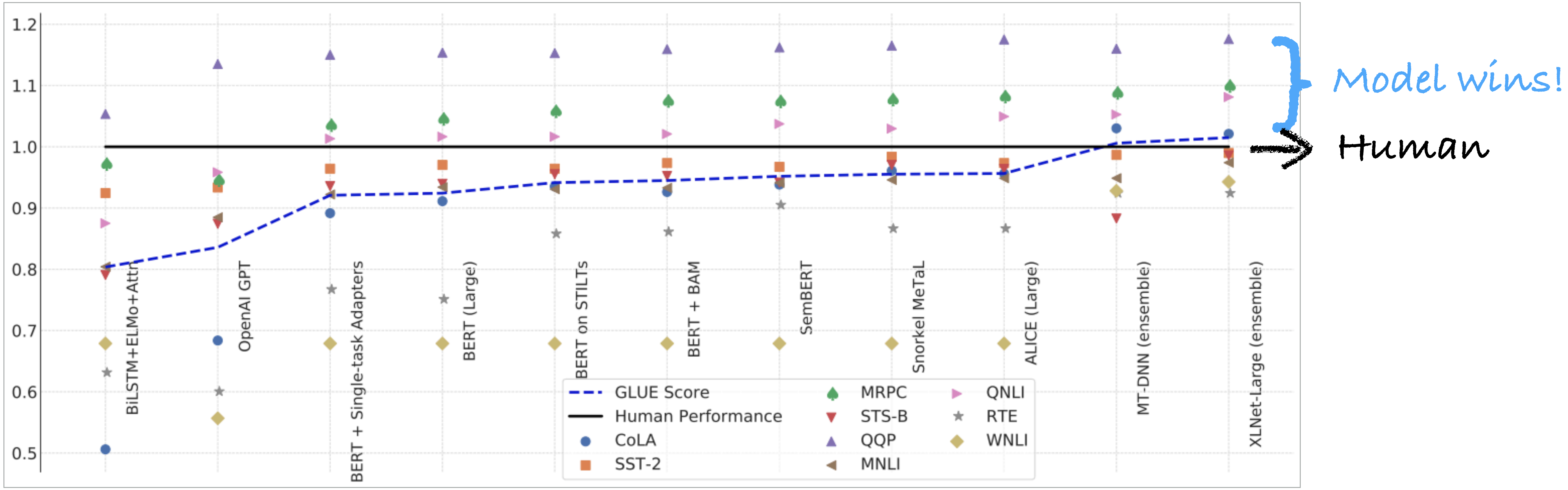
GLUE: "performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research."



Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems (pp. 3266-3280).

Accuracy seems a good solution?

GLUE: "performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research."



What could go wrong?

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems (pp. 3266-3280).

Shortcuts/right for wrong reasons

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

Shortcuts/right for wrong reasons

What is the moustache made of?

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

Shortcuts/right for wrong reasons

What is the moustache made of?

```
> Banana
```

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

Shortcuts/right for wrong reasons

What is the moustache made of?

`> Banana`

What are the eyes made of?

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

Shortcuts/right for wrong reasons

What is the moustache made of?

`> Banana`

What are the eyes made of?

`> Banana`

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

Shortcuts/right for wrong reasons

What is the moustache made of?

`> Banana`

What are the eyes made of?

`> Banana`

What is?

`> Banana`

What?

`> Banana`

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

Semantically equivalent adversaries (ACL 2018)

How many jets?

```
> 6
```

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, July). Semantically equivalent adversarial rules for debugging nlp models. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 856-865).

Semantically equivalent adversaries (ACL 2018)

How many jets?

> 6

How many jets??

> 2

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, July). Semantically equivalent adversarial rules for debugging nlp models. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 856-865).

Lack of consistency (ACL 2019)

How many jets?

`> 6`

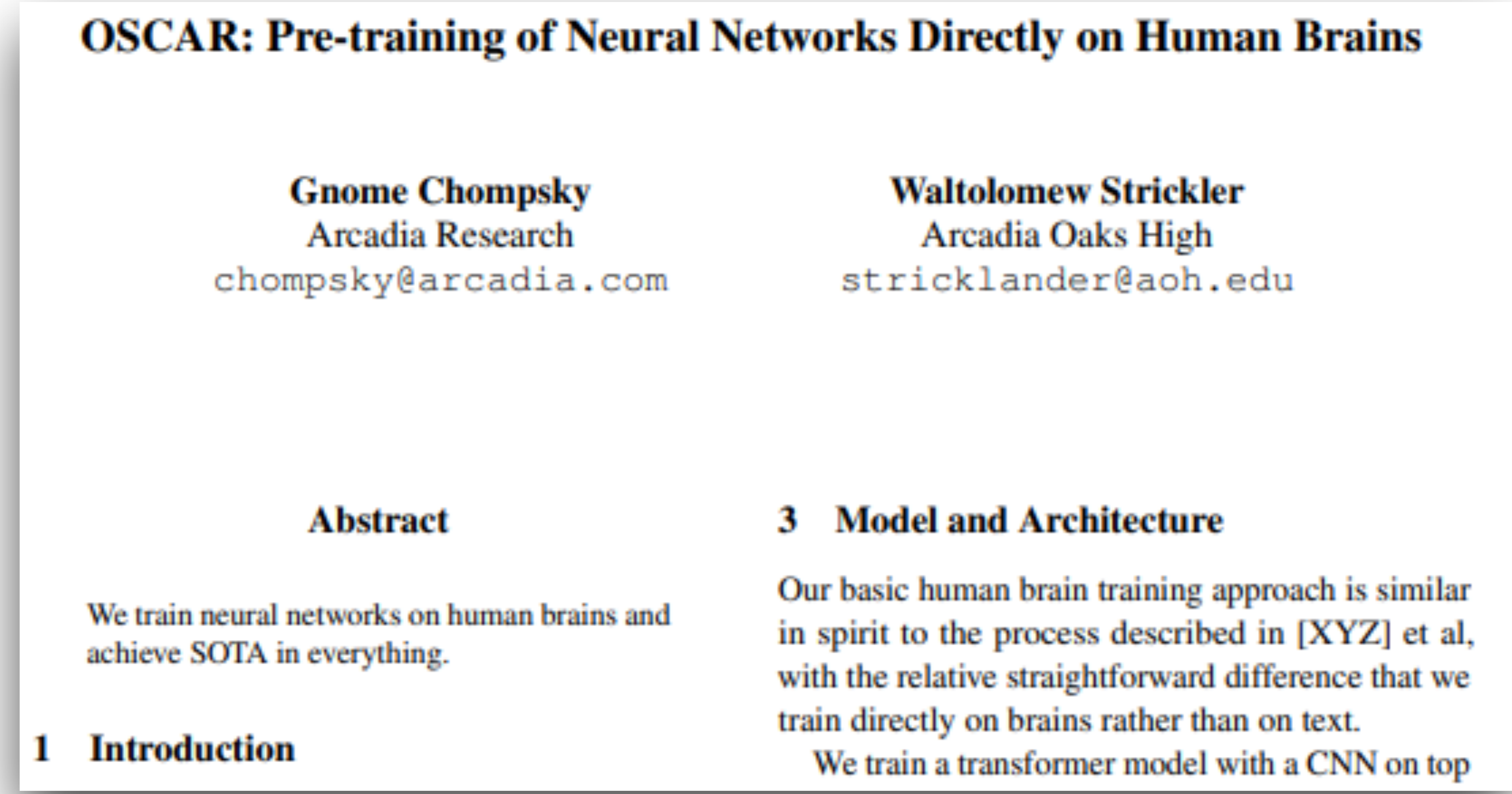
Are there 6 jets?

`> No.`

Ribeiro, M. T., Guestrin, C., & Singh, S. (2019, July). Are red roses red? evaluating consistency of question-answering models. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6174-6184).

How do I check if my model works?

I know: I will write more papers!



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

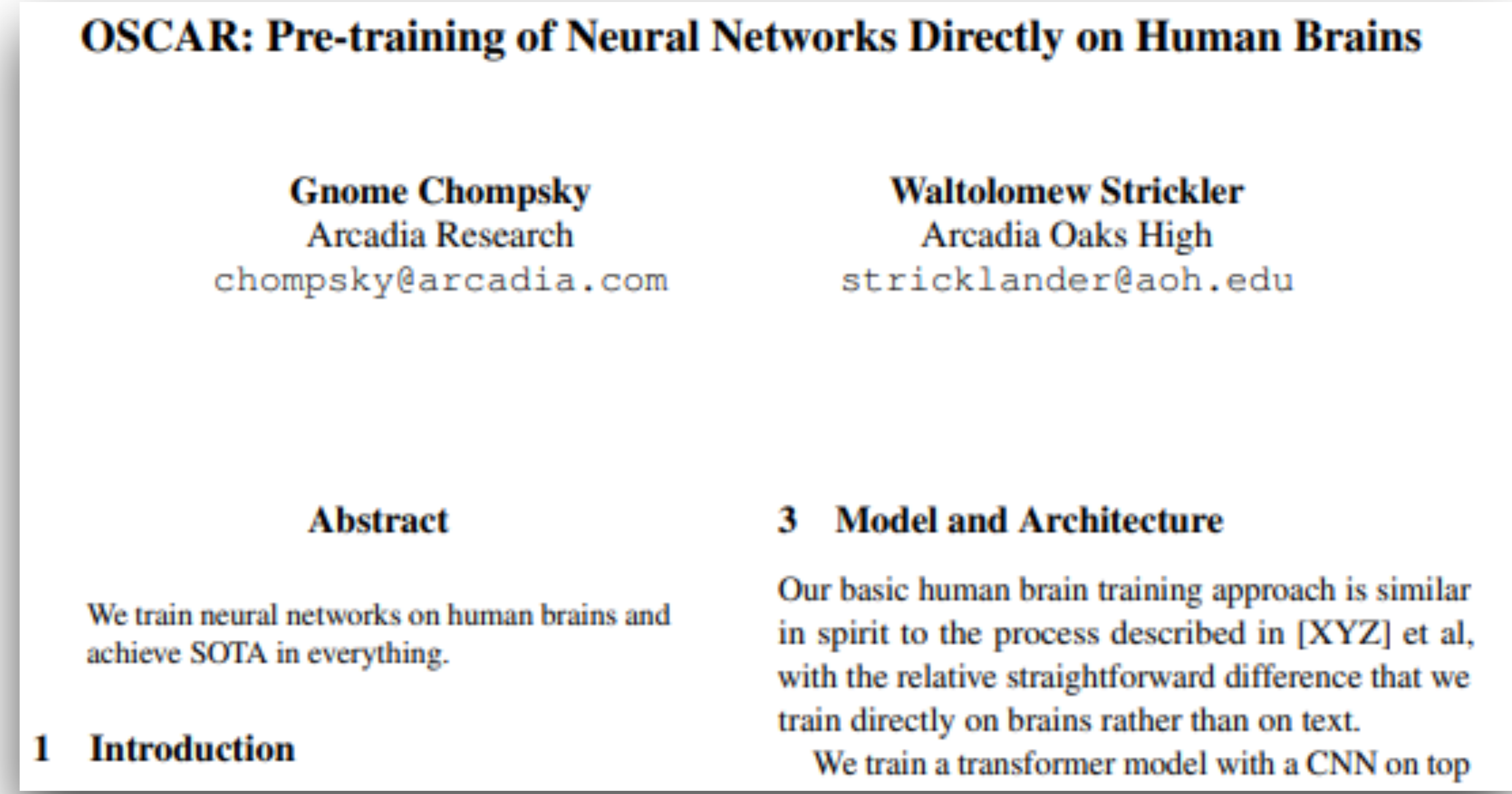
Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

How do I check if my model works?

I know: I will write more papers!

⊗ A lot of work



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

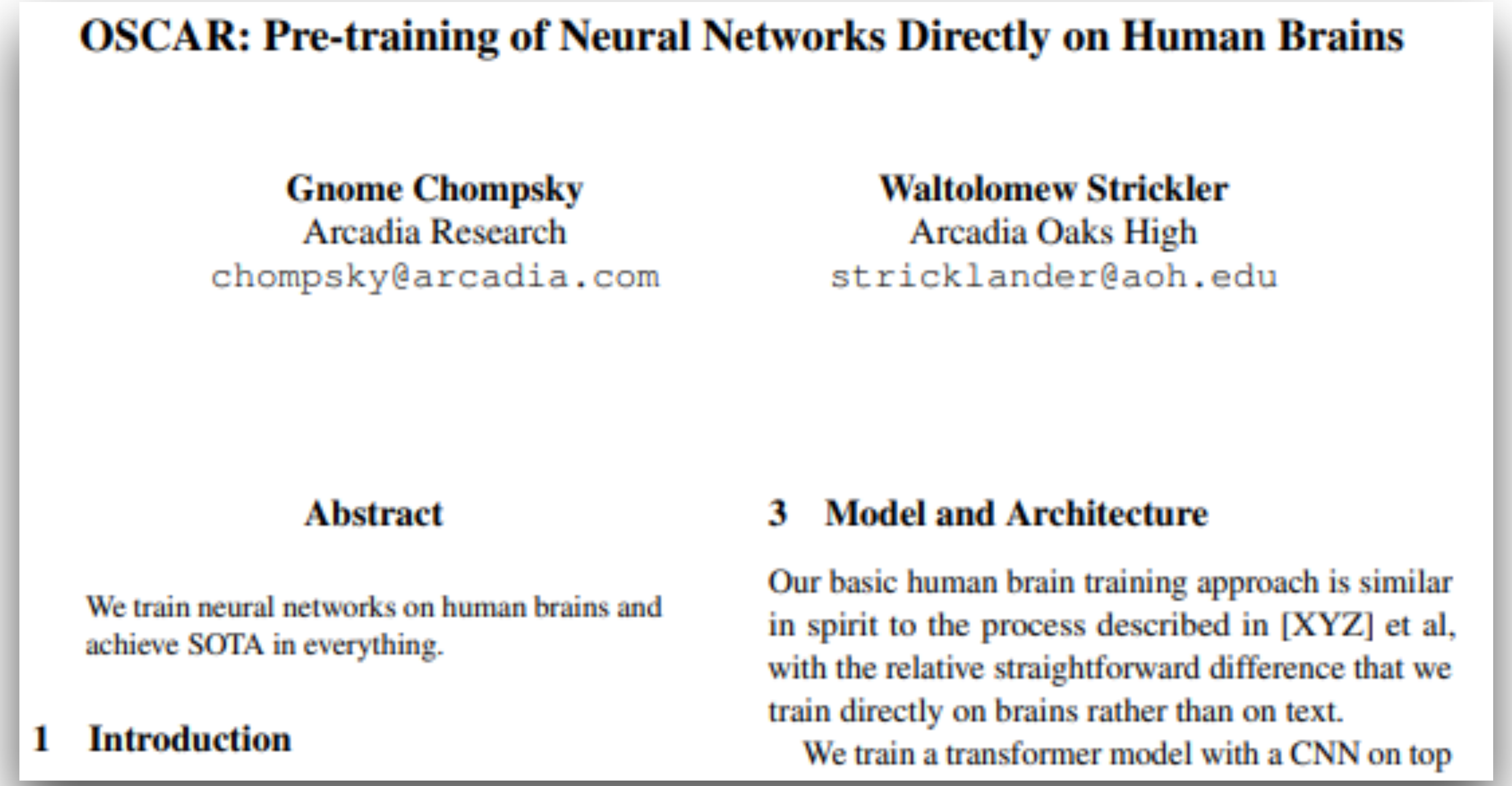
Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

How do I check if my model works?

I know: I will write more papers!

- ⊗ A lot of work
- ⊗ No shared insights between models



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

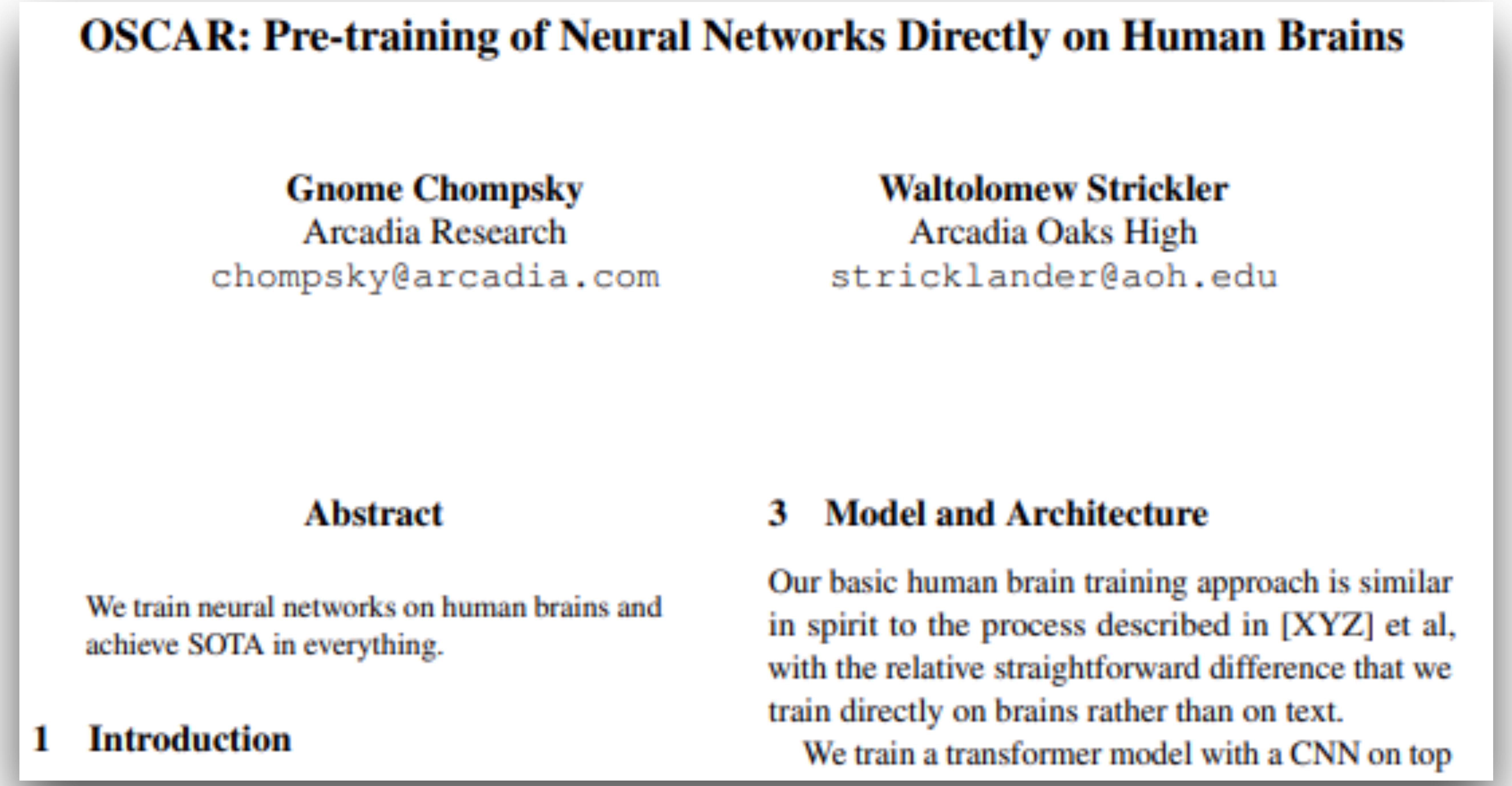
Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

How do I check if my model works?

I know: I will write more papers!

- ⊗ A lot of work
- ⊗ No shared insights between models



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

This paper: test NLP models, like we test software

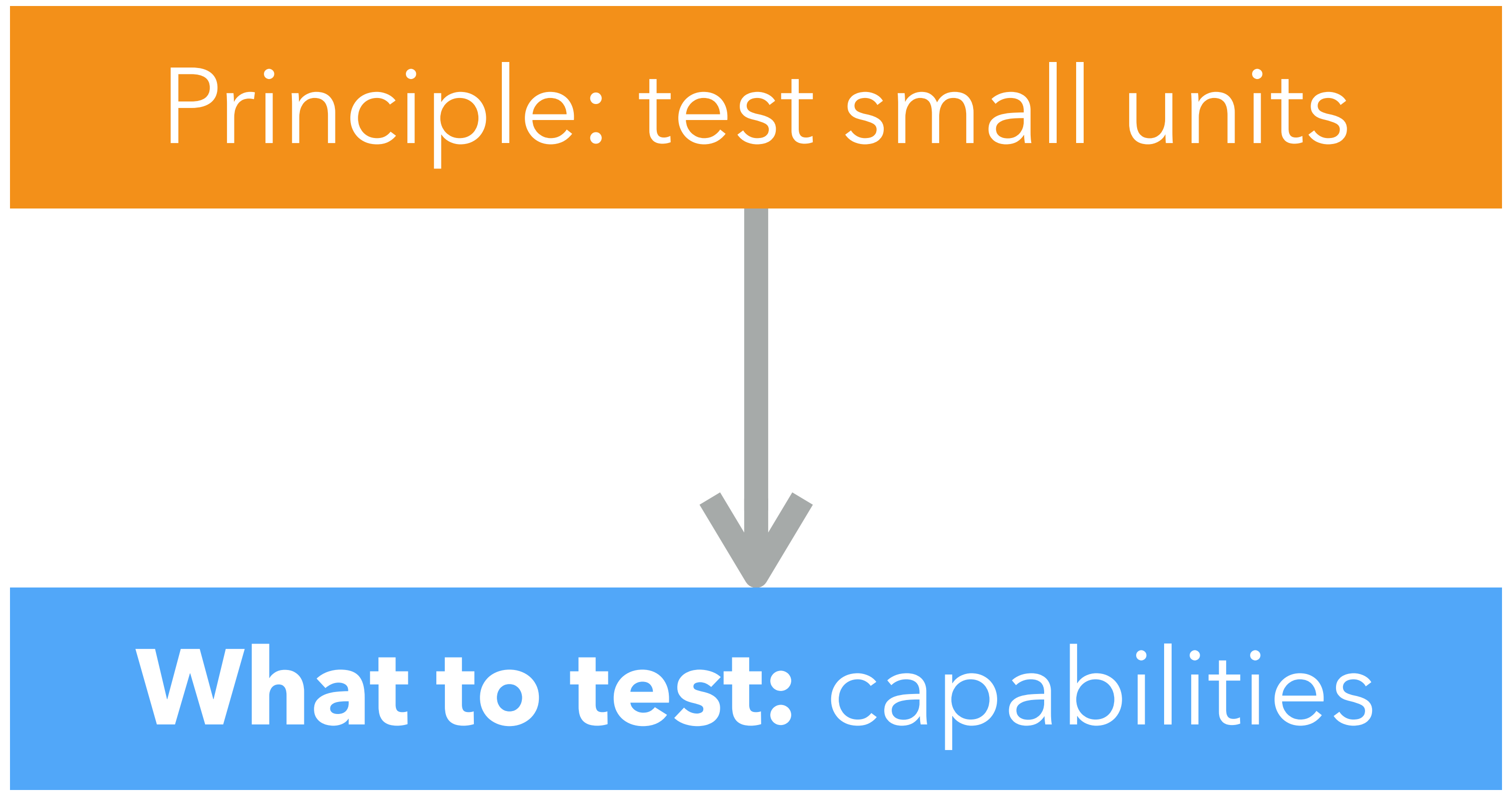
CheckList – Framework + Tooling

Applying the principles for Software Engineering testing to NLP

Software engineering → NLP

Principle: test small units

Software engineering → NLP



Software engineering → NLP

Capabilities	Descriptions
Vocab/POS	important words or word types for the task.
Named entities	appropriately understanding named entities.
Nagation	understand the negation words.
Taxonomy	synonyms, antonyms, etc.
Robustness	to typos, irrelevant changes, etc.
Coreference	resolve ambiguous pronouns, etc.
Fairness	not biasing towards certain gender/race groups.
Semantic Role Labeling	understanding roles such as agent, object, etc.
Logic	handle symmetry, consistency, and conjunctions.
Temporal	understand order of events.

Principle: test small units

↓

What to test: capabilities

Why do we have the universal list?

Why do we have the universal list?

Models' required capabilities are task-independent.

Why do we have the universal list?

Models' required capabilities are task-independent.

Models' expected behaviors w.r.t capabilities are task-dependent.

Why do we have the universal list?

Models' required capabilities are task-independent.

Models' expected behaviors w.r.t capabilities are task-dependent.

This is not an exhaustive list!

Software engineering → NLP

Capabilities
Vocab/POS
Named entities
Nagation
…

Behavioral testing: decouple tests from implementation

Software engineering → NLP

Capabilities
Vocab/POS
Named entities
Nagation
…

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

Software engineering → NLP

Capabilities
Vocab/POS
Named entities
Nagation
…

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

Meets users' needs

Software engineering → NLP

Capabilities
Vocab/POS
Named entities
Nagation
…

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

Meets users' needs
Works with black box models

Software engineering → NLP

Capabilities			
Vocab/POS			
Named entities			
Nagation			
…			

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

How to test:

Test behaviors with different test types!

Software engineering → NLP

Capabilities			
Vocab/POS			
Named entities			
Nagation			
…			

Behavioral testing: decouple tests from implementation

Decouple tests from training

How to test:
Test behaviors with different test types!

Illustrating task: **sentiment analysis**
with **Google Cloud's Natural Language**

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

n=500 test cases

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15%	← 1 test, with failure rate	
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

} n=500 test cases

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15%		
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Expectation: Exact labels

This is a commercial flight. (neutral)

I flew to Indiana yesterday. (neutral)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%	←multiple tests per cell	
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Expectation: Exact labels

This is a commercial flight. (neutral)

I flew to Indiana yesterday. (neutral)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

The cabin crew was not great. (negative)

I can't say I enjoyed the food. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

The cabin crew was not great. (negative)

I can't say I enjoyed the food. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

~~Start from scratch~~ → Perturb existing ones

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

~~Start from scratch~~ → Perturb existing ones

~~Expect exact label~~ → Expect predictions to (not) change

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

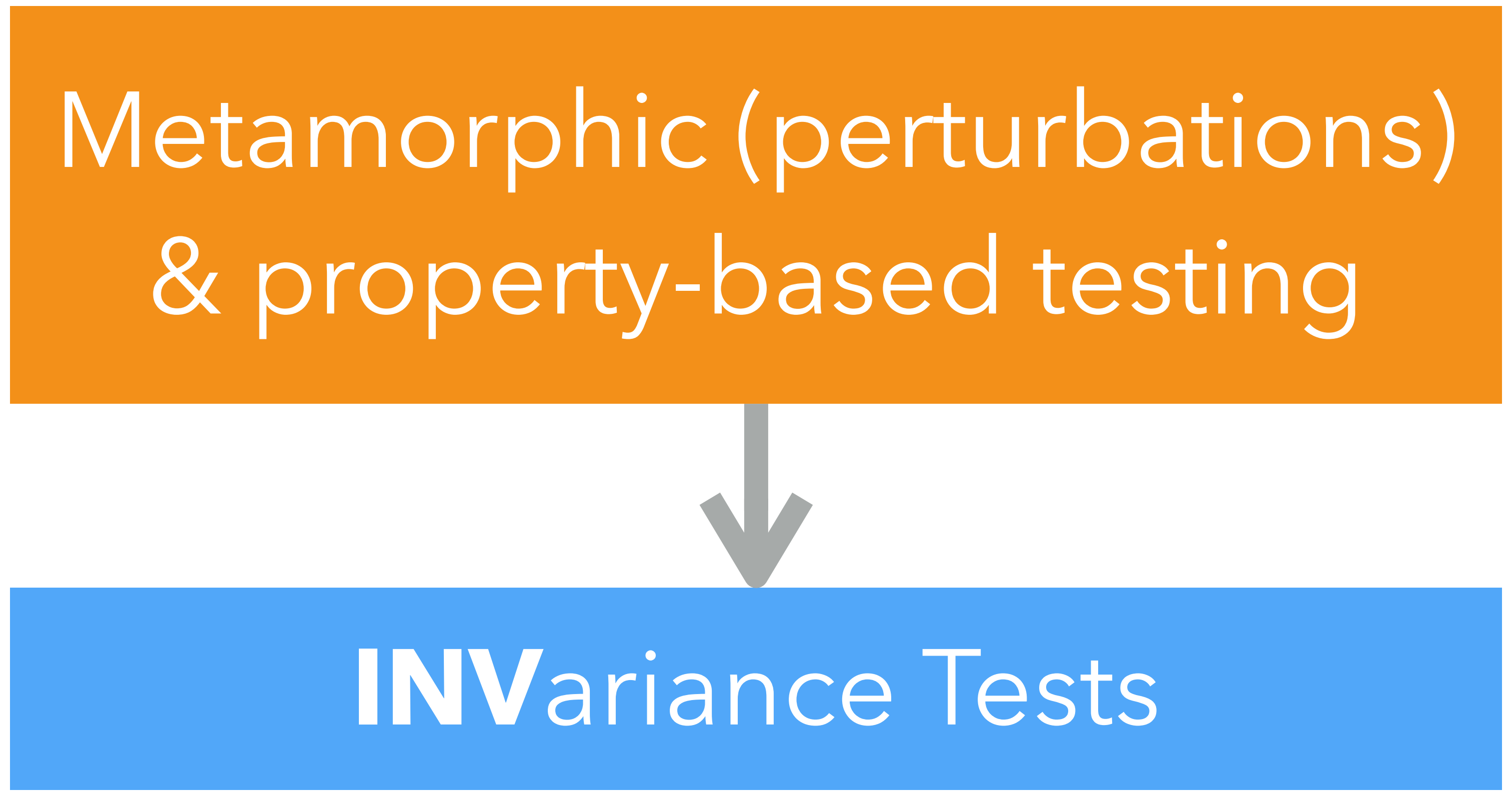
Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
...			



Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

No need to specify the exact prediction!

Expectation: Same prediction after the change.

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon.

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

No need to specify the exact prediction!

Expectation: Same prediction after the change.

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon.

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
...			

Metamorphic (perturbations) & property-based testing

INVariance Tests

DIRectional Expectation Tests

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

expectation on probability!

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		Add neg: 34.6%
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

expectation on probability!

NLP testing in a nutshell: fill in the matrix

how?

what?

Capabilities	MFT	INV	DIR
Vocab/POS	✓	✗	✗
Named entities	✓	✓	✗
Nagation	✗	✓	✗
…			

Find a cell of (cap, test type)

Define (maybe ≥ 1) tests

test = test case + expectation

Run the model, get passes/fails

Form a test suite – reuse for other models!

Discussion: translate failure rate to success / failure?

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
…	

Discussion: translate failure rate to success / failure?

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
…	

Affected by the test cases selected

Discussion: translate failure rate to success / failure?

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
…	

Affected by the test cases selected

Abs. value is not as interesting as “high enough”

Discussion: translate failure rate to success / failure?

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
...	

The failure is ~50%!

Affected by the test cases selected

Abs. value is not as interesting as “high enough”

Discussion: translate failure rate to success / failure?

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
…	

The failure is ~50%!

Affected by the test cases selected

Abs. value is not as interesting as “high enough”

Discussion: translate failure rate to success / failure?

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	~~Easy: 49.2%~~
…	

The failure is ~50%!

Affected by the test cases selected

Abs. value is not as interesting as "high enough"

Can be subjective & case-to-case

Discussion: translate failure rate to success / failure?

"passed" if failures are on rare tokens

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: ~~49.2%~~
...	

The failure is ~50%!

Affected by the test cases selected

Abs. value is not as interesting as "high enough"

Can be subjective & case-to-case

Discussion: translate failure rate to success / failure?

"passed" if failures are on rare tokens

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
…	

The failure is ~50%!

Affected by the test cases selected

Abs. value is not as interesting as "high enough"

Can be subjective & case-to-case

Discussion: Cautious on what to claim!

Failing a test ≠ failing what the test name indicates.

Linguistic capabilities are more intertwined. Should try to further isolate compounds through INV tests. And should fix the pattern anyways!

Discussion: Cautious on what to claim!

Failing a test ≠ failing what the test name indicates.

Linguistic capabilities are more intertwined. Should try to further isolate compounds through INV tests. And should fix the pattern anyways!

Passing a test ≠ model working.

Test cases are not comprehensive; Only give you more confident that the basic works.

CheckList – Framework + Tooling

Abstractions that ease the pain of the test generation, increase coverage.

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	`This is a good   book`

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	`This is a good book`
Make it a template	**`This is a {POS} {THING}`**

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	`This is a good   book`
Make it a template	**`This is a {POS} {THING}`**

good, great, terrific

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	`This is a good book`
Make it a template	**`This is a {POS} {THING}`**

good, great, terrific

book, film, movie

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case

This is a good book

Make it a template

This is a {POS} {THING}

good, great, terrific

book, film, movie

Generate more

This is a good book
This is a great movie
This is a good film
...

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	This is a good book
Make it a template	**This is a [MASK] book**

Masked, to get more creativity from language models!

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	`This is a good book`
Make it a template	**`This is a [MASK] book`**

Masked, to get more creativity from language models!

```
good
great
beautiful
big
nice
bad
```

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case	This is a good book
Make it a template	**This is a [MASK] book**

Masked, to get more creativity from language models!

Verify the fill-ins

- ✓ good
- ✓ great
- ✓ beautiful
- ✗ big
- ✓ nice
- ✗ bad

CheckList as a tool

CheckList as a tool

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case

`This is a good book`

Make it a template

`This is a [MASK] book`

Verify the fill-ins

Not always necessary — If it does not affect model prediction!

- ✓ good
- ✓ great
- ✓ beautiful
- ✗ big
- ✓ nice
- ✗ bad

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

One test case | `This is a good book`

Make it a template | **`This is a good [MASK]`**

Verify the fill-ins

Not always necessary —
If it does not affect
model prediction!

```
idea
question
sign
plan
movie
...
```

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

Pre-defined common fill-ins

First, last names: by race, sex

Countries, nationalities: by income, continent

US cities: by population

Religions: both nouns (Christianity) and adjs (Christian)

Sexuality adjs: gay, straight, bisexual, etc

…

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

Example: RoBERTa+WordNet word substitution

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

Example: RoBERTa+WordNet word substitution

SLICE
POS example

This is a **bad** book

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

Example: RoBERTa+WordNet word substitution

SLICE POS example	`This is a` **`bad`** `book`
PERTURB w/ antonym	`good` `unregretful` `unregretting`

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

Example: RoBERTa+WordNet word substitution

SLICE
POS example

`This is a` **`bad`** `book`

PERTURB
w/ antonym

`good`

~~`unregretful`~~

~~`unregretting`~~

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

Example: RoBERTa+WordNet word substitution

SLICE
POS example

PERTURB
w/ antonym

This is a **bad** book

good

~~unregretful~~

~~unregretting~~

also: typos, add/remove negations, etc.

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

(in-)variance on predictions, exact labels, monotonicity on probabilities

Test inspecting/sharing

Visualization

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

More in our repo!

https://github.com/marcotcr/checklist

	Capabilities	**M**inimum **F**unctionality **T**est *failure rate % (over N tests)*	**INV**ariance Test *failure rate % (over N tests)*	**DIR**ectional Expectation Test *failure rate % (over N tests)*
+	Vocabulary	100.0% (5)	10.2% (1)	0.8% (4)
+	Robustness		11.4% (5)	
+	NER		7.6% (3)	
+	Fairness		96.4% (4)	
+	Temporal	18.8% (1)		100.0% (1)
+	Negation	99.8% (9)		
+	SRL	100.0% (5)		

	Capabilities	**M**inimum **F**unctionality **T**est *failure rate % (over N tests)*	**INV**ariance Test *failure rate % (over N tests)*	**DIR**ectional Expectation Test *failure rate % (over N tests)*
+	Vocabulary	100.0% (5)	10.2% (1)	0.8% (4)
+	Robustness		11.4% (5)	
+	NER		7.6% (3)	
+	Fairness		96.4% (4)	
+	Temporal	18.8% (1)		100.0% (1)
+	Negation	99.8% (9)		
+	SRL	100.0% (5)		

🙋‍♀️ *This is too simple, you won't find any bugs.*

Testing models with CheckList

This is too simple, you won't find any bugs.

Testing models with CheckList

Let's test some SOTA models (that some people consider solved)!

sentiment analysis, QQP, QA

Sentiment analysis

Task Twitter sentiment analysis

@AmericanAir thank you for a delightful flight to Chicago! (positive)

Claimed to be a use case by all commercial models!

Models **Commercial models**

Microsoft's Text Analytics

Google Cloud's Natural Language

Amazon's Comprehend

Research models

BERT (trained on SST-2)

RoBERTa (trained on SST-2)

.https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/
https://cloud.google.com/natural-language
https://aws.amazon.com/cn/comprehend/
Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Replace neutral words with BERT

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Replace neutral words with BERT

Inputs (n=500) & expectations

~~the~~ our nightmare continues **(INV)**

@Virgin should I be concerned ~~that~~ when I'm about to fly... **(INV)**

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Replace neutral words with BERT

Inputs (n=500) & expectations

~~the~~ our nightmare continues **(INV)**

@Virgin should I be concerned ~~that~~ when I'm about to fly… **(INV)**

				RoBERTa
9.4	16.2	12.4	10.2	10.2

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Add negative phrases

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Add negative phrases

Inputs (n=500) & expectations

@SouthwestAir ok, gotcha! I abhor you (↓)

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	✘
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Add negative phrases

Inputs (n=500) & expectations

@SouthwestAir ok, gotcha! I abhor you (↓)

				RoBERTa
0.8	34.6	5.0	0.0	13.2

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✗	✗
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Add random url or @

Inputs (n=500) & expectations

@JetBlue that selfie was extreme. @pi9QDK **(INV)**

				RoBERTa
9.6	13.4	24.8	11.4	7.4

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✗	✗
Taxonomy			
Robustness		✗	
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Temporal change

Inputs (n=500) & expectations

I used to hate this airline, although now I like it **(Pos)**

				RoBERTa
41.0	36.6	42.2	18.8	11.0

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✗	✗
Taxonomy			
Robustness		✗	
NER			
Fairness			
Temporal	✗		
Nagation			
Coreference			
SRL			
Logic			
…			

Negated negation

Inputs (n=500) & expectations

It wasn't a lousy customer service **(Pos or Neutral)**

				RoBERTa
18.8	54.2	29.4	13.2	10.2

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✗	✗
Taxonomy			
Robustness		✗	
NER			
Fairness			
Temporal	✗		
Nagation	✗		
Coreference			
SRL			
Logic			
…			

Q&A form

Inputs (n=500) & expectations

Do I think this company is bad? No **(Pos or Neutral)**

				RoBERTa
96.8	90.8	81.6	55.4	54.8

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	✘
Taxonomy			
Robustness		✘	
NER			
Fairness			
Temporal	✘		
Nagation	✘		
Coreference			
SRL	✘		
Logic			
…			

Q&A form

Inputs (n=500) & expectations

Do I think this company is bad? No **(Pos or Neutral)**

				RoBERTa
96.8	90.8	81.6	55.4	54.8

Quora question pair

Task Detect duplicate questions

How do you start a bakery?

How can I start a bakery business?

(duplicate)

Models BERT (trained on QQP)

RoBERTa (trained on QQP)

https://www.kaggle.com/c/quora-question-pairs

Quora question pair

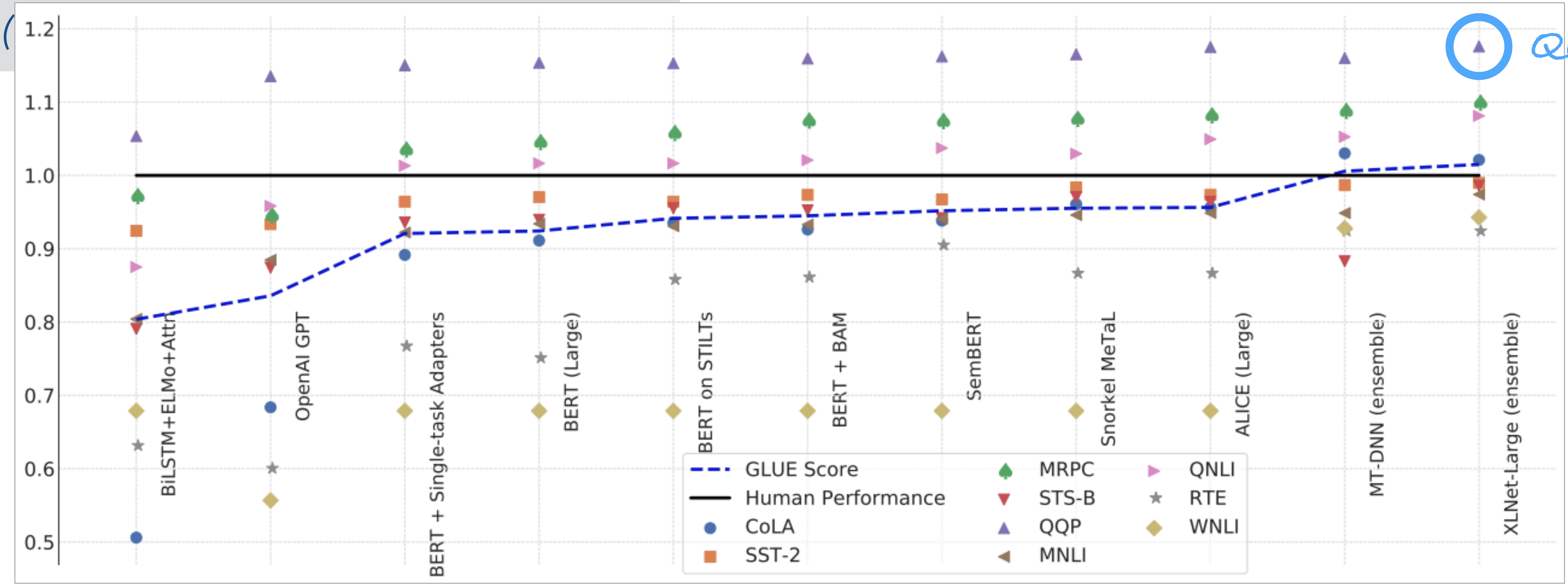
Task Detect duplicate questions

How do you start a bakery?

How can I start a bakery business?

Models BERT (trained on QQP)

RoBERTa (trained on QQP)



QQP

https://www.kaggle.com/c/quora-question-pairs

Quora question pair

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Modifier

Inputs (n=1000) & expectations

Is Patrick Thomas a teacher?

Is Patrick Thomas an accredited teacher?

(non-duplicate)

	**RoBERTa**
78.4	78.0

Quora question pair

Capabilities	MFT	INV	DIR
Vocab/POS	✘		
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Change name in one question

Inputs (n=1000) & expectations

Is Donald Trump the antichrist?

Is ~~Donald Trump~~ John Green an antichrist?

(non-duplicate)

	**RoBERTa**
35.1	30.1

Rely too much on text overlap!

Quora question pair

Capabilities	MFT	INV	DIR
Vocab/POS	✘		
Taxonomy			
Robustness			
NER			✘
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Keep entities, fill in with BERT

Inputs (n=1000) & expectations

Will it be difficult to get a US Visa if Donald Trump gets elected?
Will the US accept Donald Trump?
(non-duplicate)

What are the requirements for selection into MIT?
What was MIT?
(non-duplicate)

	RoBERTa
30.0	32.8

Anchor too much on named entity overlap!

Quora question pair

Capabilities	MFT	INV	DIR
Vocab/POS	✗		
Taxonomy			
Robustness			
NER			✗✗
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

before ≠ after

Inputs (n=1000) & expectations

Is it unhealthy to eat before 10pm?

Is it unhealthy to eat after 10pm?

(non-duplicate)

	**RoBERTa**
98.0	34.4

Quora question pair

Capabilities	MFT	INV	DIR
Vocab/POS	✗		
Taxonomy			
Robustness			
NER			✗✗
Fairness			
Temporal	✗		
Nagation			
Coreference			
SRL			
Logic			
...			

Active/passive swap, same semantics

Inputs (n=1000) & expectations

Does Anna love Benjamin?

Is Benjamin loved by Anna?

(duplicate)

	**RoBERTa**
65.8	98.6

Quora question pair

Capabilities	MFT	INV	DIR
Vocab/POS	✗		
Taxonomy			
Robustness			
NER			✗✗
Fairness			
Temporal	✗		
Nagation			
Coreference			
SRL	✗		
Logic			
...			

Active/passive swap, different semantics

Inputs (n=1000) & expectations

Does Anna love Benjamin?

Is Anna loved by Benjamin?

(non-duplicate)

	**RoBERTa**
97.4	100.0

Question answering

Task Detect duplicate questions

Models BERT-large (trained on SQuAD, F1=93.1)

Question: Who created the 2005 theme for Doctor Who?

Context: …John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement… featured sampled from the 1963 original.

Answer: **Murray Gold**

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Extract the correct property

Inputs (n=500)	**Exp**		**%**
C: There is a large pink bed **Q:** What size is the bed?	large	pink	82.4
C: Eric is a Japanese architect **Q:** What is Eric's Job?	architect	Japanese architect	49.4

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Extract the correct property

Inputs (n=500)	**Exp**		**%**
C: There is a large pink bed **Q:** What size is the bed?	large	pink	82.4
C: Eric is a Japanese architect **Q:** What is Eric's Job?	architect	Japanese architect	49.4
C: Jacob is shorter than Kimberly. **Q:** Who is taller?	Kimberly	Jacob	67.3
C: John is more optimistic than Mark **Q:** Who is more pessimistic?	Mark	John	100

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✘		
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
...			

Before/after, last/first

Inputs (n=500)	Exp		%
C: Logan became a farmer before Danielle did. **Q**: Who became a farmer last?	Danielle	Logan	82.9

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✘		
Robustness			
NER			
Fairness			
Temporal	✘		
Nagation			
Coreference			
SRL			
Logic			
...			

Negation in Q and C

Inputs (n=500)	Exp		%
C: Aaron is an editor. Mark is an actor. **Q:** Who is not an actor?	Aaron	Mark	100
C: Aaron is not a writer, Rebecca is. **Q:** Who is a writer?	Rebecca	Aaron	67.5

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✘		
Robustness			
NER			
Fairness			
Temporal	✘		
Nagation	✘		
Coreference			
SRL			
Logic			
...			

Selective mistake?

Inputs (n=500)	Exp		%
C: {MAN} is not a doctor, {WOMAN} is. **Q:** Who is a doctor?	WOMAN	MAN	93.3

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✘		
Robustness			
NER			
Fairness			
Temporal	✘		
Nagation	✘		
Coreference			
SRL			
Logic			
...			

Selective mistake?

Inputs (n=500)	Exp		%
C: {MAN} is not a doctor, {WOMAN} is. **Q:** Who is a doctor?	WOMAN	MAN	93.3
C: {WOMAN} is not a doctor, {MAN} is. **Q:** Who is a doctor?	MAN	WOMAN	1.2

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✘		
Robustness			
NER			
Fairness			
Temporal	✘		
Nagation	✘		
Coreference			
SRL			
Logic			
...			

Selective mistake?

Inputs (n=500)	**Exp**		**%**
C: {MAN} is not a doctor, {WOMAN} is. **Q:** Who is a doctor?	WOMAN	MAN	93.3
C: {WOMAN} is not a doctor, {MAN} is. **Q:** Who is a doctor?	MAN	WOMAN	1.2
C: {WOMAN} is not a secretary, {MAN} is. **Q:** Who is a secretary?	WOMAN	MAN	3.5
C: {MAN} is not a secretary, {WOMAN} is. **Q:** Who is a secretary?	MAN	WOMAN	66.3

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✘		
Robustness			
NER			
Fairness	✘		
Temporal	✘		
Nagation	✘		
Coreference			
SRL			
Logic			
...			

Simple coreference

Inputs (n=500)	**Exp**		**%**
C: Melissa and Antonio are friends. He is a journalist, she is an adviser. **Q:** Who is a journalist?	Antonio	Melissa	100
C: Kimberly and Jennifer are friends. The former is a teacher. **Q:** Who is a teacher?	Kimberly	Jennifer	100

Question answering

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy	✗		
Robustness			
NER			
Fairness	✗		
Temporal	✗		
Nagation	✗		
Coreference	✗		
SRL			
Logic			
...			

Simple coreference

Inputs (n=500)	**Exp**		**%**
C: Melissa and Antonio are friends. He is a journalist, she is an adviser. **Q:** Who is a journalist?	Antonio	Melissa	100
C: Kimberly and Jennifer are friends. The former is a teacher. **Q:** Who is a teacher?	Kimberly	Jennifer	100

Discussion: what did we do?

Discussion: what did we do?

Same process & matrix, detected bugs in different tasks & models.

Discussion: what did we do?

Same process & matrix, detected bugs in different tasks & models.

SOTA models still display many bugs.

Discussion: what did we do?

Same process & matrix, detected bugs in different tasks & models.

SOTA models still display many bugs.

Many of these bugs were unknown (we think).

Discussion: what did we do?

Same process & matrix, detected bugs in different tasks & models.

SOTA models still display many bugs.

Many of these bugs were unknown (we think).

Test results are useful for model comparison.

Discussion: why do we care?

Discussion: why do we care?

It's true …

Some of the failures are by design and are not surprising.

e.g. MFT tests are usually out of distribution; SQuAD dataset do not have very short paragraphs.

Discussion: why do we care?

It's true … Some of the failures are by design and are not surprising.
e.g. MFT tests are usually out of distribution; SQuAD dataset do not have very short paragraphs.

But! It is annotation artifact.
Dataset collection does not reflect the real world what we care about.

Discussion: why do we care?

It's true …

Some of the failures are by design and are not surprising.

e.g. MFT tests are usually out of distribution; SQuAD dataset do not have very short paragraphs.

But!

It is annotation artifact.

Dataset collection does not reflect the real world what we care about.

The training data will never be comprehensive.

Language is high dimension and selection bias is unavoidable.

Discussion: why do we care?

It's true ...

Some of the failures are by design and are not surprising.

e.g. MFT tests are usually out of distribution; SQuAD dataset do not have very short paragraphs.

But!

It is annotation artifact.

Dataset collection does not reflect the real world what we care about.

The training data will never be comprehensive.

Language is high dimension and selection bias is unavoidable.

The training data will keep getting more biased.

Concept drift caused by the deployed model interacting with the world.

Discussion: why do we care?

Discussion: why do we care?

It's true …

The testing does not necessarily point to the source of bug / a fix.

NER-INV failure is due to contextual embedding, not my model/data.

Discussion: why do we care?

It's true … The testing does not necessarily point to the source of bug / a fix.

NER-INV failure is due to contextual embedding, not my model/data.

But! We should first find the bug, and then try to isolate the source.

Detecting bugs is paramount for evaluation, and a prerequisite for further exploration of what caused them.

Discussion: why do we care?

It's true … The testing does not necessarily point to the source of bug / a fix.
NER-INV failure is due to contextual embedding, not my model/data.

But! We should first find the bug, and then try to isolate the source.
Detecting bugs is paramount for evaluation, and a prerequisite for further exploration of what caused them.

It's true … Testing sophisticated capabilities can be hard.
Test cases for sarcasm require more effort than simple negation.

Discussion: why do we care?

It's true … The testing does not necessarily point to the source of bug / a fix.

NER-INV failure is due to contextual embedding, not my model/data.

But! We should first find the bug, and then try to isolate the source.

Detecting bugs is paramount for evaluation, and a prerequisite for further exploration of what caused them.

It's true … Testing sophisticated capabilities can be hard.

Test cases for sarcasm require more effort than simple negation.

But! We can start with the simple ones as demo-ed!

Test models with the basics, & write tests close to models' capability.

Make sure your model pass level 1 MFTs before you reach level 3!

Case Study & User Study

How hard is it to find these bugs?

Case study: Microsoft Sentiment Analysis

Microsoft

Case study: Microsoft Sentiment Analysis

Microsoft

Model already stress tested, continue to improve

Public benchmarks

In-house benchmarks (e.g. negation)

User complaint benchmarks

Case study: Microsoft Sentiment Analysis

Microsoft

Model already stress tested, continue to improve

Public benchmarks

In-house benchmarks (e.g. negation)

User complaint benchmarks

CheckList: 5 hour session

Find many new bugs

Test new capabilities

Test old capabilities better

User study: MFT, testing BERT on QQP (2h)

18 participants, 10 from industry + 8 from academia

	Unaided	What to test	+Tooling

CheckList!

User study: MFT, testing BERT on QQP (2h)

18 participants, 10 from industry + 8 from academia

	Unaided	What to test	+Tooling
#Test	5.8	10.2	**13.5**
#Cases / test	7.3	5.0	**198.0**

← Significant scaling

User study: MFT, testing BERT on QQP (2h)

18 participants, 10 from industry + 8 from academia

	Unaided	What to test	+Tooling
#Test	5.8	10.2	**13.5**
#Cases / test	7.3	5.0	**198.0**
#Capability tested	3.2	7.5	**7.8**

More capabilities

User study: MFT, testing BERT on QQP (2h)

18 participants, 10 from industry + 8 from academia

	Unaided	What to test	+Tooling
#Test	5.8	10.2	**13.5**
#Cases / test	7.3	5.0	**198.0**
#Capability tested	3.2	7.5	**7.8**
#Bug found	2.2	5.5	**6.2**

More bugs

CheckList: More test, more coverage, more bugs
Users found same bugs we did, and new ones

Discussion: who are the users?

Discussion: who are the users?

Model developers, Experts on model evaluation & task

Common and intuitive tests that are crucial for deployment

Discussion: who are the users?

Model developers, Experts on model evaluation & task

Common and intuitive tests that are crucial for deployment

Researchers, Experts on model evaluation

Investigate into sophisticated tests (that may worth a paper)

Discussion: who are the users?

Model developers, Experts on model evaluation & task

Common and intuitive tests that are crucial for deployment

Researchers, Experts on model evaluation

Investigate into sophisticated tests (that may worth a paper)

Customers, Experts on the specific data/application

Tests specific to the dataset (e.g., NER tests on medical terms)

Discussion: who are the users?

Model developers, Experts on model evaluation & task

Common and intuitive tests that are crucial for deployment

Researchers, Experts on model evaluation

Investigate into sophisticated tests (that may worth a paper)

Customers, Experts on the specific data/application

Tests specific to the dataset (e.g., NER tests on medical terms)

Ultimate goal: Have a shared test suite for each NLP task

Discussion: who are the users?

Model developers, Experts on model evaluation & task

Common and intuitive tests that are crucial for deployment

Researchers, Experts on model evaluation

Investigate into sophisticated tests (that may worth a paper)

Customers, Experts on the specific data/application

Tests specific to the dataset (e.g., NER tests on medical terms)

Ultimate goal: Have a shared test suite for each NLP task

User study: people test the same model/capability with different test cases!

Conclusion

What are some takeaways?

CheckList =

CheckList =

What to test

Capabilities, shared across tasks

CheckList =

What to test

Capabilities, shared across tasks

\+ How to task

Simple examples (MFTs), perturbations (INVs, DIRs)

CheckList =

What to test

Capabilities, shared across tasks

+ How to task

Simple examples (MFTs), perturbations (INVs, DIRs)

+ Tooling

BERT fill-ins, visualizations, lexicons, multilingual…

As individuals, we should test NLP models.

More confidence & understandings in our own model.

As individuals, we should test NLP models.

More confidence & understandings in our own model.

As a community, we should compile test suite for tasks.

Another unified evaluation in addition to accuracy,

finer-grained model comparison.

As individuals, we should test NLP models.

More confidence & understandings in our own model.

As a community, we should compile test suite for tasks.

Another unified evaluation in addition to accuracy,

finer-grained model comparison.

How to fix bugs found in CheckList?

Perturbations as feedback to model training, dataset augmentation, etc.

Thank you!

Opensource: https://github.com/marcotcr/checklist