Coherence Models

Re-Thinking the Shuffle Test

Philippe Laban, Luke Dai,
Lucas Bandarkar, Marti A. Hearst

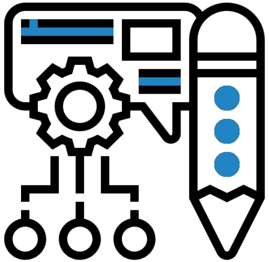
Berkeley
UNIVERSITY OF CALIFORNIA

ACL2021 Recorded Presentation

We thank our sponsor: Microsoft Research

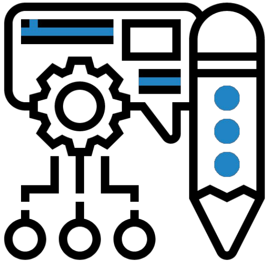
Progress in text generation

Modern language models generate increasingly realistic text.



Progress in text generation

Modern language models generate increasingly realistic text.



	Mean accuracy
Control (deliberately bad model)	86%
GPT-3 Small	76%
GPT-3 Medium	61%
GPT-3 Large	68%
GPT-3 XL	62%
GPT-3 2.7B	62%
GPT-3 6.7B	60%
GPT-3 13B	55%
GPT-3 175B	52%

Human accuracy in identifying whether short (~200 word) news articles are model generated

Tom Brown, et al. "Language Models are Few-Shot Learners" *Proceedings of NeurIPS*. 2020.

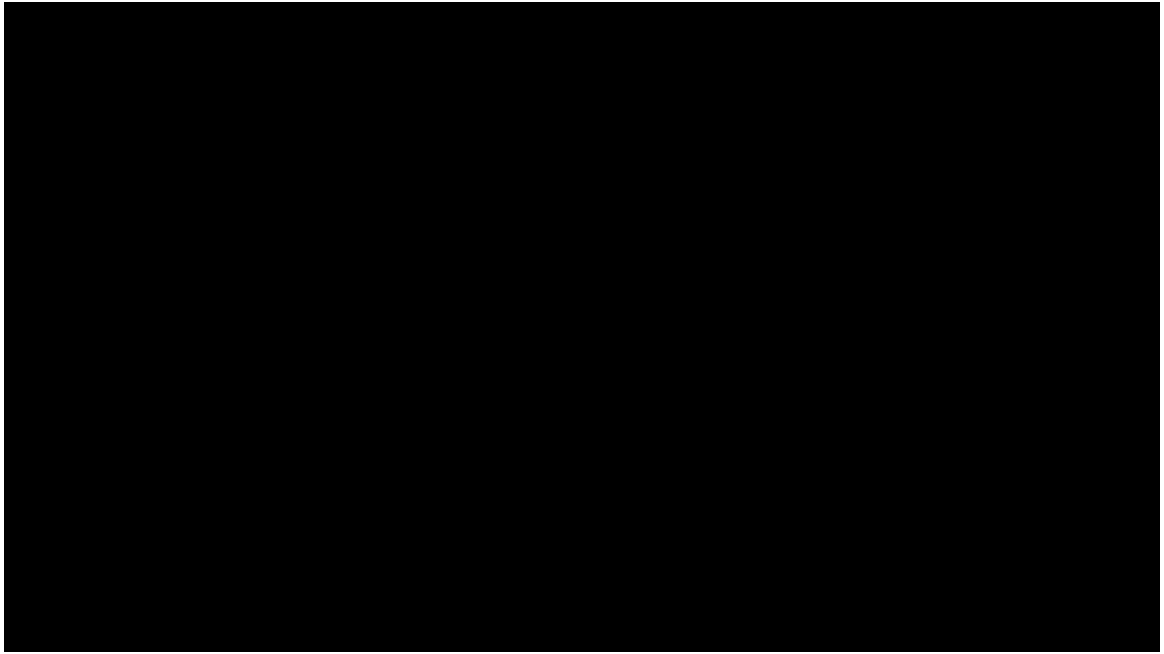
Is the generated text

Coherent ?

Can we measure

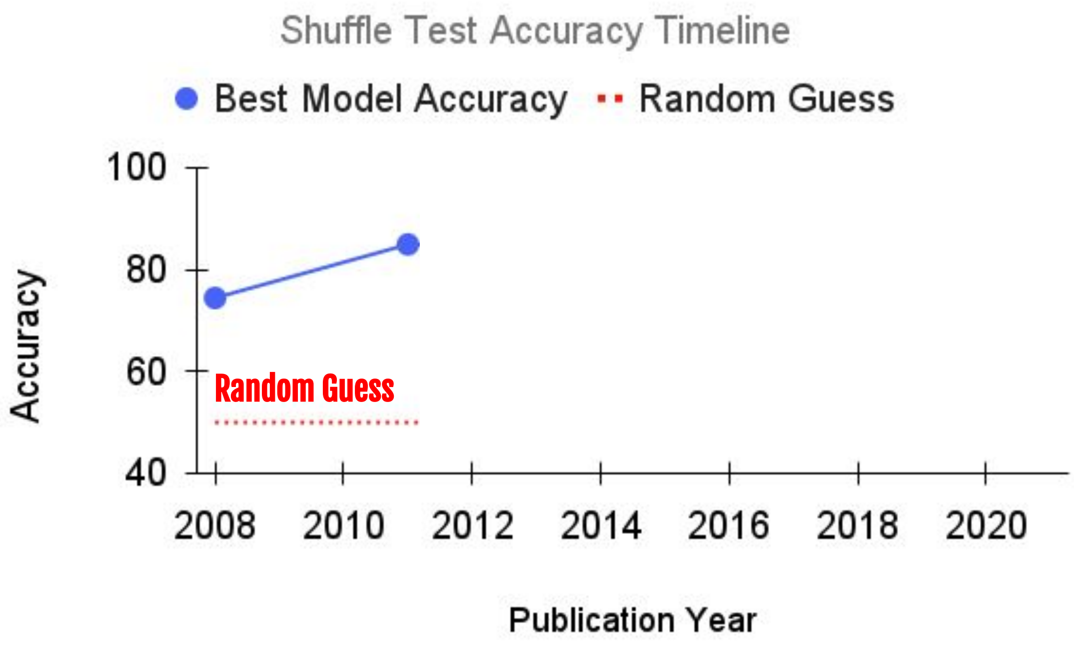
Coherence ?

The Shuffle Test



Can NLU models detect shuffled text?

Timeline of Shuffle Test Performance

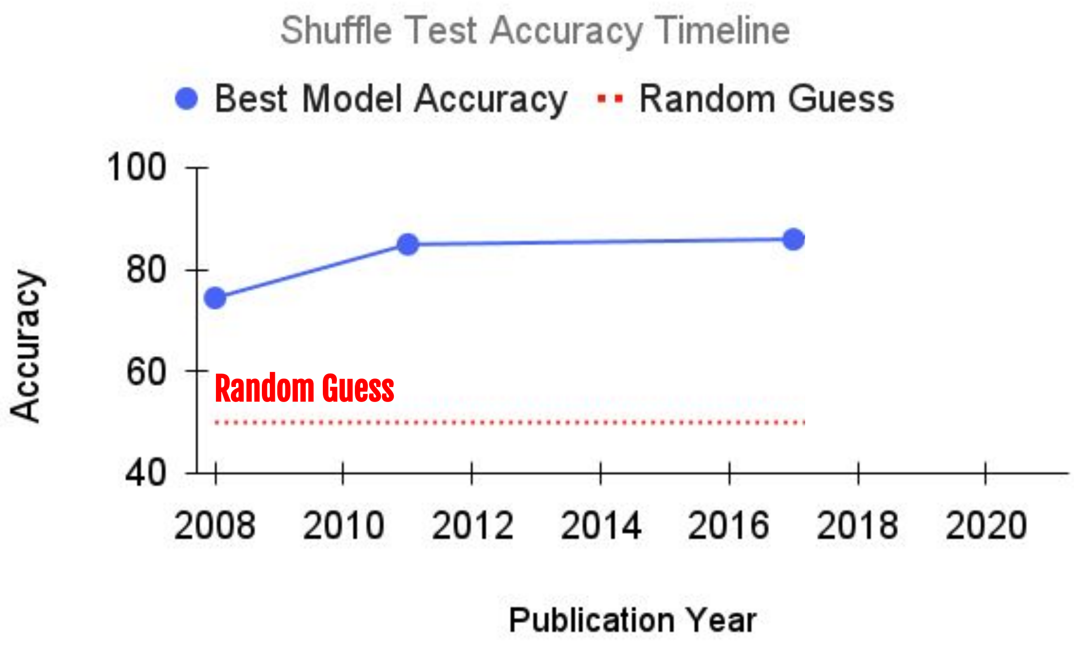


On a common Wall Street Journal (WSJ) test-set of ~1000 documents

The task was introduced in 2008 by Barzilay et al., along with the *Entity Grid*: ~75% accuracy.

Elsner et al. *extend* the Entity Grid in 2011, adding new features and improve accuracy to ~85% accuracy.

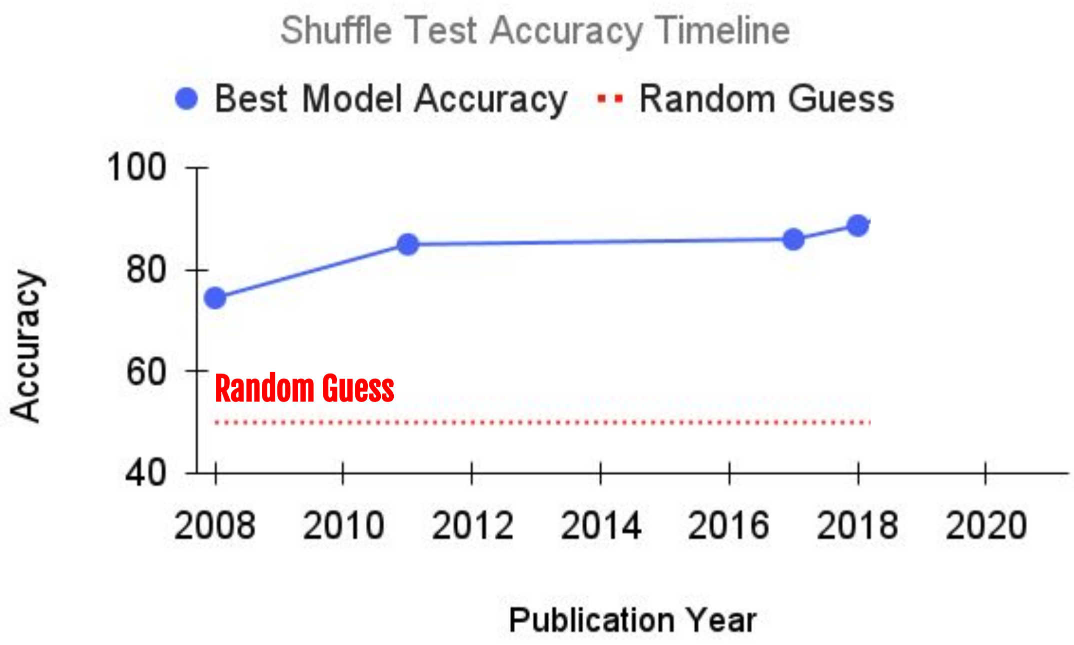
Timeline of Shuffle Test Performance



On a common Wall Street Journal (WSJ) test-set of ~1000 documents

Nguyen et al. introduce a CNN-based neural network to the task in 2017. Accuracy: 86%.

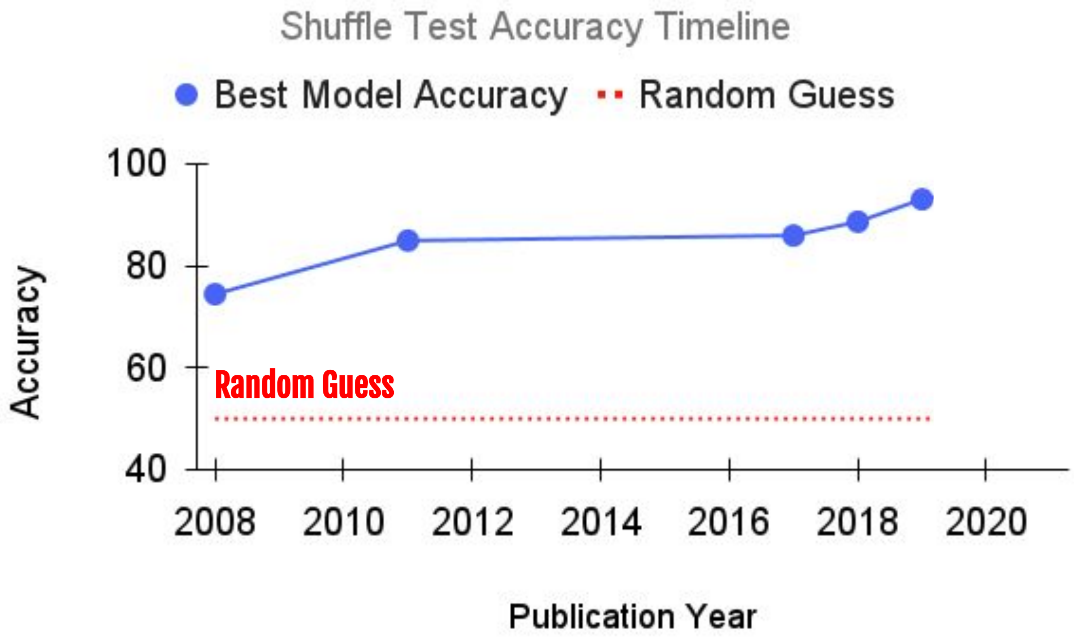
Timeline of Shuffle Test Performance



In 2018, Mohiuddin et al. add word2vec embeddings to a neural network for the task. Accuracy: 89%.

On a common Wall Street Journal (WSJ) test-set of ~1000 documents

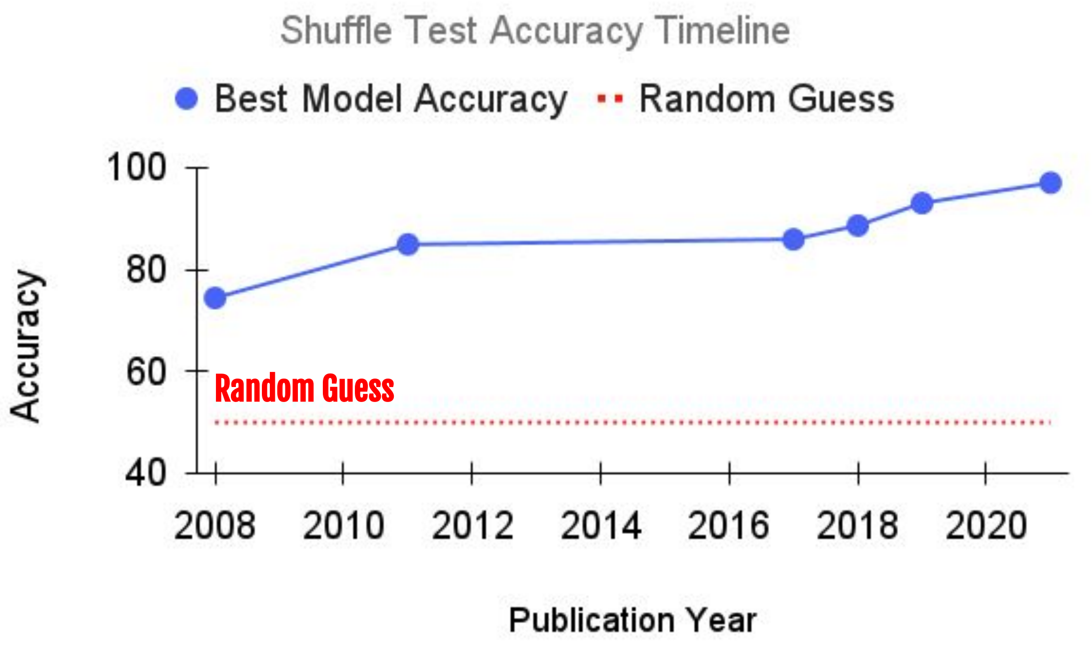
Timeline of Shuffle Test Performance



On a common Wall Street Journal (WSJ) test-set of ~1000 documents

In 2019, Moon et al. leverage ELMO contextual embeddings to improve performance. Accuracy: 93%.

Timeline of Shuffle Test Performance



On a common Wall Street Journal (WSJ) test-set of ~1000 documents

In each step, an increase in model capacity leads to an improvement in task performance.

We push this logic further: we finetune a **Roberta-large** on the task.

We achieve near-perfect performance of 98%.

Is supervision on the Shuffle Test appropriate?

Shuffle Test and Supervision

Mohiuddin et al. 2020 show:
There is only **weak correlation** between performance of coherence models on *synthetic tasks* (e.g. the Shuffle Test) and *downstream tasks* (e.g. ranking generated summaries).

Shuffle Test and Supervision

Mohiuddin et al. 2020 show:
There is weak correlation between performance of coherence models on *synthetic tasks* (e.g. the Shuffle Test) and *downstream tasks* (e.g. ranking generated summaries).

We argue:
That this is due to the use of *direct supervision* on the task.
High-capacity models learn features specific to *shuffle-ness*, not necessarily important for *coherence.*

Zero-Shot Shuffle Test

We propose that the Shuffle Test should be applied in a **Zero-Shot** setting.
More precisely:

In the Zero-Shot Shuffle Test, the evaluated model must not be pre-trained, fine-tuned or modified using shuffled text.

Adapting Models

To the Zero-Shot Shuffle Test

Adapting LM Models

We adapt Language Models to perform the Zero-Shot Shuffle Test.

Language models take as input a text **`T`**, and output a probability: **`LM(T)`**

Adapting LM Models

We adapt Language Models to perform the Zero-Shot Shuffle Test.

Language models take as input a text **T**, and output a probability: **`LM(T)`**

Given a text **T1** and a shuffled version **T2**:

If `LM(T1) > LM(T2)`, we label T1 as **original**, and T2 as **shuffled**
If `LM(T1) <= LM(T2)`, we label T1 as **shuffled**, and T2 as **original**

Adapting LM Models

We adapt Language Models to perform the Zero-Shot Shuffle Test.

Language models take as input a text **`T`**, and output a probability: **`LM(T)`**

- LMs are trained on a language modeling loss, and not exposed to shuffled text.
- We test GPT2 language models of varying size (small, medium, large).

Adapting NLU Models

We adapt NLU models (e.g., BERT) to perform the Zero-Shot Shuffle Test.

Because of bi-directionality, we need to use pseudo-log likelihood, also known as Masked Language Model Scoring (**`MLMS`**).

Salazar, Julian, et al. "Masked Language Model Scoring." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Adapting NLU Models

We adapt NLU models (e.g., BERT) to perform the Zero-Shot Shuffle Test.

Because of bi-directionality, we need to use pseudo-log likelihood, also known as Masked Language Model Scoring (**`MLMS`**).

Given a text **T1** and a shuffled version **T2**:

If `MLMS(T1) > MLMS(T2)`, we label T1 as **original**, and T2 as **shuffled**
If `MLMS(T1) <= MLMS(T2)`, we label T1 as **shuffled**, and T2 as **original**

Adapting NLU Models

We adapt NLU models (e.g., BERT) to perform the Zero-Shot Shuffle Test.

Because of bi-directionality, we need to use pseudo-log likelihood, also known as Masked Language Model Scoring (**`MLMS`**).

We experiment with BERT and RoBERTa models, which are pre-trained with Masked Language Modeling.

Shuffle Test – Test Domains

News

Based on Wall Street Journal (WSJ) corpus. *(following prior work)*

Legal

Based on US legislation bills included in the BillSum dataset.

Blog/Reddit

Based on entire Reddit posts included in the Reddit TIFU dataset.

No domain overlaps with the training data used to train GPT2, BERT or RoBERTa

Results

Model	News	Legal	Reddit	Overall
GPT2-base	47	92	75	71
GPT2-medium	**91**	99	89	**93**
GPT2-large	73	**99**	91	88
BERT-base	73	96	86	85
RoBERTa-base	82	95	**97**	91

Accuracy of several models at the Zero-Shot Shuffle Test on test portions of three textual domains.

Results

Model	News	Legal	Reddit	Overall
GPT2-base	47	92	75	71
GPT2-medium	**91**	99	89	**93**
GPT2-large	73	**99**	91	88
BERT-base	73	96	86	85
RoBERTa-base	82	95	**97**	91

Observation 1: Models are competitive even in Zero-Shot setting, with overall performance ~90%.

Results

Model	News	Legal	Reddit	Overall
GPT2-base	47	92	75	71
GPT2-medium	**91**	99	89	**93**
GPT2-large	73	**99**	91	88
BERT-base	73	96	86	85
RoBERTa-base	82	95	**97**	91

Observation 2: Performance varies across domains (e.g., all models have strong performance in Legal domain).

Results

Model	News	Legal	Reddit	Overall
GPT2-base	47	92	75	71
GPT2-medium	**91**	99	89	**93**
GPT2-large	73	**99**	91	88
BERT-base	73	96	86	85
RoBERTa-base	82	95	**97**	91

Observation 3: Increasing model size mostly increases model performance at the Zero-shot Shuffle Test.

Results

Model	News	Legal	Reddit	Overall
GPT2-base	47	92	75	71
GPT2-medium	**91**	99	89	**93**
GPT2-large	73	**99**	91	88
BERT-base	73	96	86	85
RoBERTa-base	82	95	**97**	91

Observation 4: Bi-directional models perform better than NLG counterparts of similar size (GPT2-base vs. RoBERTa-base).

Results

Model	News	Legal	Reddit	Overall
GPT2-base	47	92	75	71
GPT2-medium	**91**	99	89	**93**
GPT2-large	73	**99**	91	88
BERT-base	73	96	86	85
RoBERTa-base	82	95	**97**	91

Observation 5: RoBERTa outperforms BERT, even though BERT has a *Next Sentence Prediction* (NSP) loss.
Confirms prior work indicating that NSP is not useful for model pre-training.

Is the Zero-Shot Shuffle Test a solved problem?

Increasing Block Size

We introduce a variation to the Shuffle Test, increasing the **block size.**

4. **Hayden usually brings coffee.**
5. **Jesse on the other hand prefers tea.**
6. **There is no accounting for tastes.**
1. Jesse and Hayden go to the park.
2. They go there every day.
3. It's a good way to get fresh air.

Block Size 3

Results – Blocked Shuffle Test

	Block Size				
Model	1	2	3	4	5
Human performance	97	94	93	96	94

Block size does not affect human performance. Inter-annotator agreement remains high for all block sizes (0.86)

Results – Blocked Shuffle Test

	Block Size				
Model	1	2	3	4	5
Human performance	97	94	93	96	94
GPT2-med - WSJ	95	91	89	87	85
GPT2-med - Legal	99	98	97	96	94
GPT2-med - Reddit	89	79	66	59	54
GPT2-med - Average	94	89	84	81	78

Model performance drops with increased block size in all domains.

Results – Blocked Shuffle Test

	Block Size				
Model	1	2	3	4	5
Human performance	97	94	93	96	94
GPT2-med - WSJ	95	91	89	87	85
GPT2-med - Legal	99	98	97	96	94
GPT2-med - Reddit	89	79	66	59	54
GPT2-med - Average	94	89	84	81	78

Average performance drops from 94% at block size 1 to 78% at block size 5.

Takeaways

1. Current coherence models are **directly supervised** on the Shuffle Test, weakening the evaluation of the models.

2. Recent Transformer-based models achieve strong **Zero-shot** performance (>90%) at the standard Shuffle Test.

3. Increasing the block-size of the Shuffle Test **increases task difficulty** for models (model performance drops to 78% at block size 5).

Thanks

See you at the Q&A!

Code & data on Github:
https://github.com/tingofurro/shuffle_test/

Contact:
phillab@berkeley.edu

We thank our sponsor: Microsoft Research