How is BERT surprised? Layerwise detection of linguistic anomalies

Bai Li, Zining Zhu, Guillaume Thomas,
Yang Xu, Frank Rudzicz

Submitted to ACL 2021 (under review)

Syntactic and semantic anomalies

Syntactic*: *The cat won't eating the food*

Semantic*: #The plane laughed at the runway*

#Colorless green ideas sleep furiously

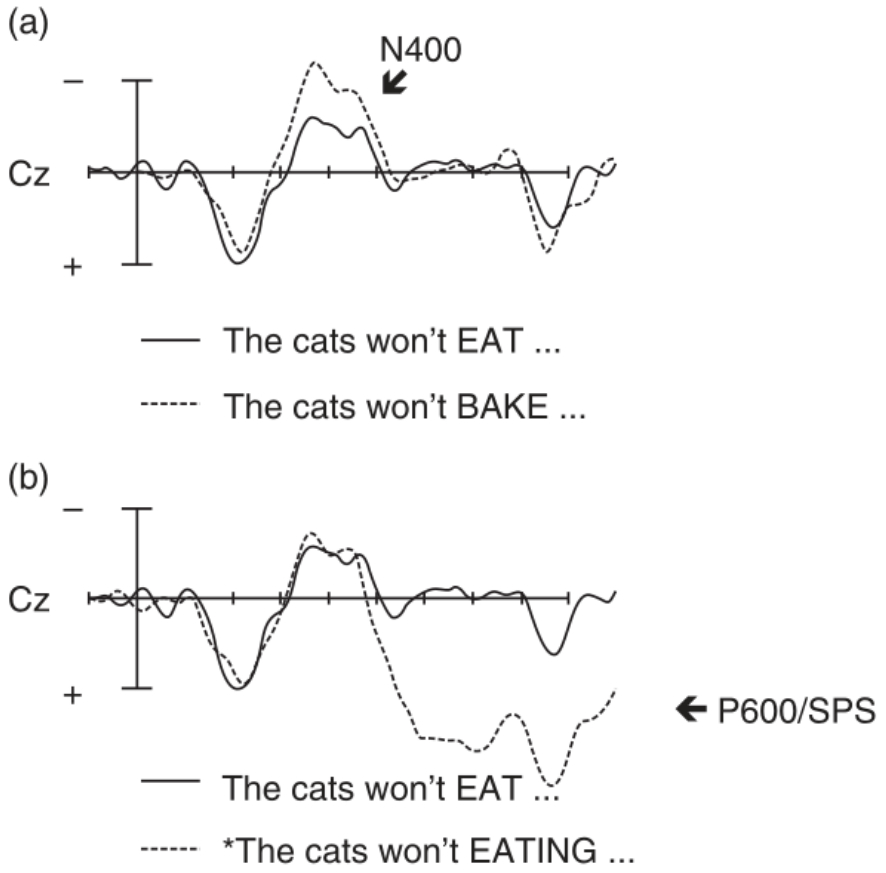
**Furiously sleep ideas green colorless*

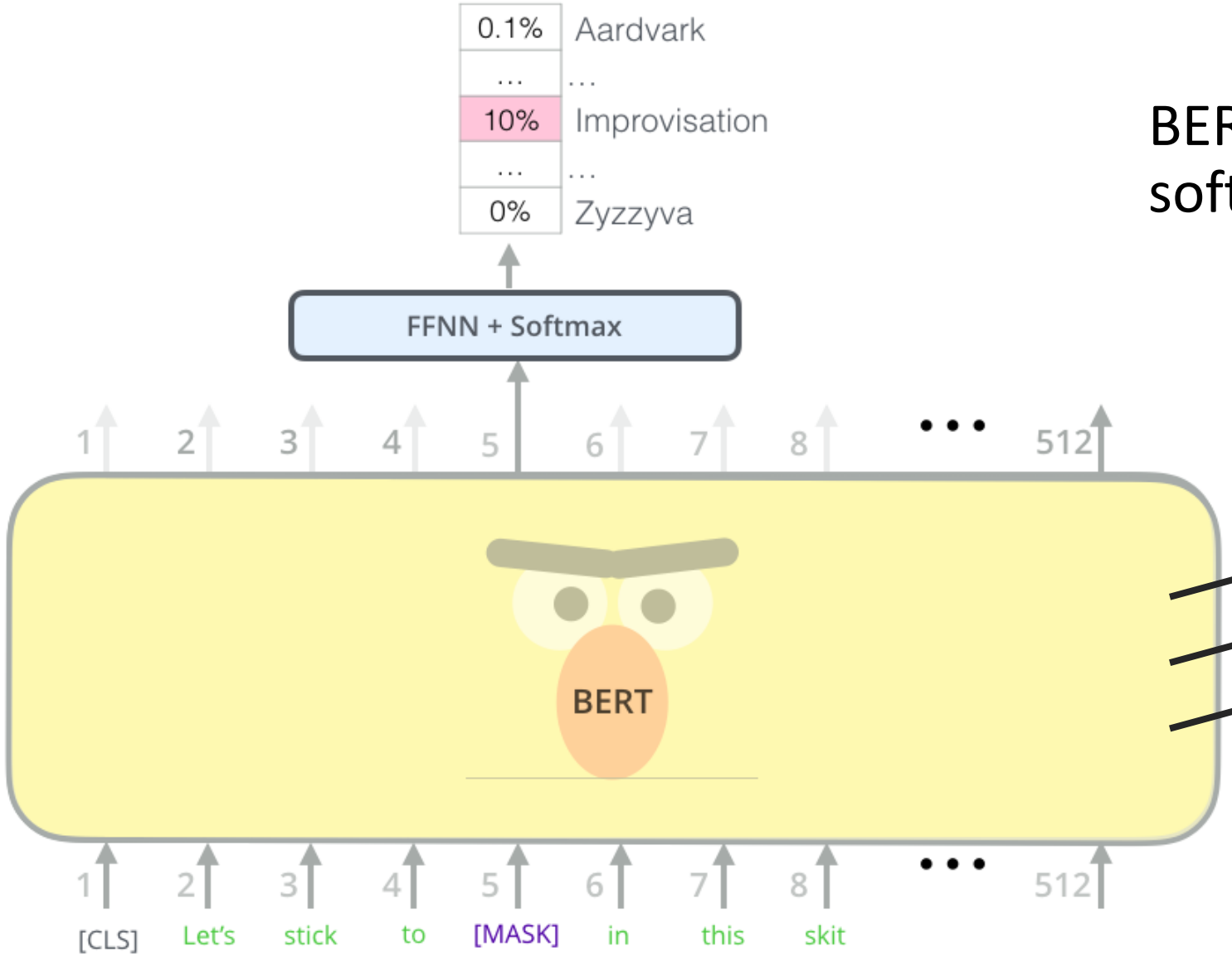
(Chomsky, 1957)

Research Question: Are LMs sensitive to different types of linguistic anomalies?

Motivation from N400 / P600

- Early work on event-related potentials found semantic anomalies trigger N400, while syntactic anomalies trigger P600.
- But follow-up experiments found it's not so simple.



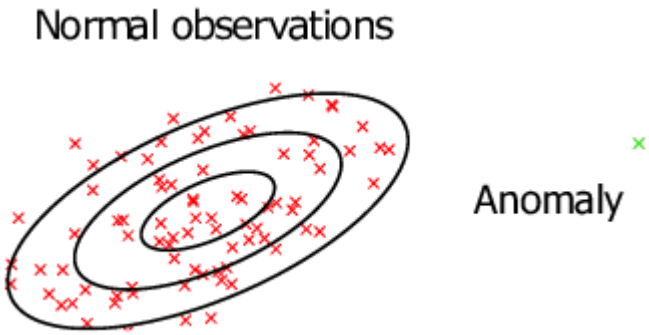


BERT gives access to final softmax probabilities.

But no way to access probabilities / surprisals at intermediate layers…

Proposed Method: Gaussian Model

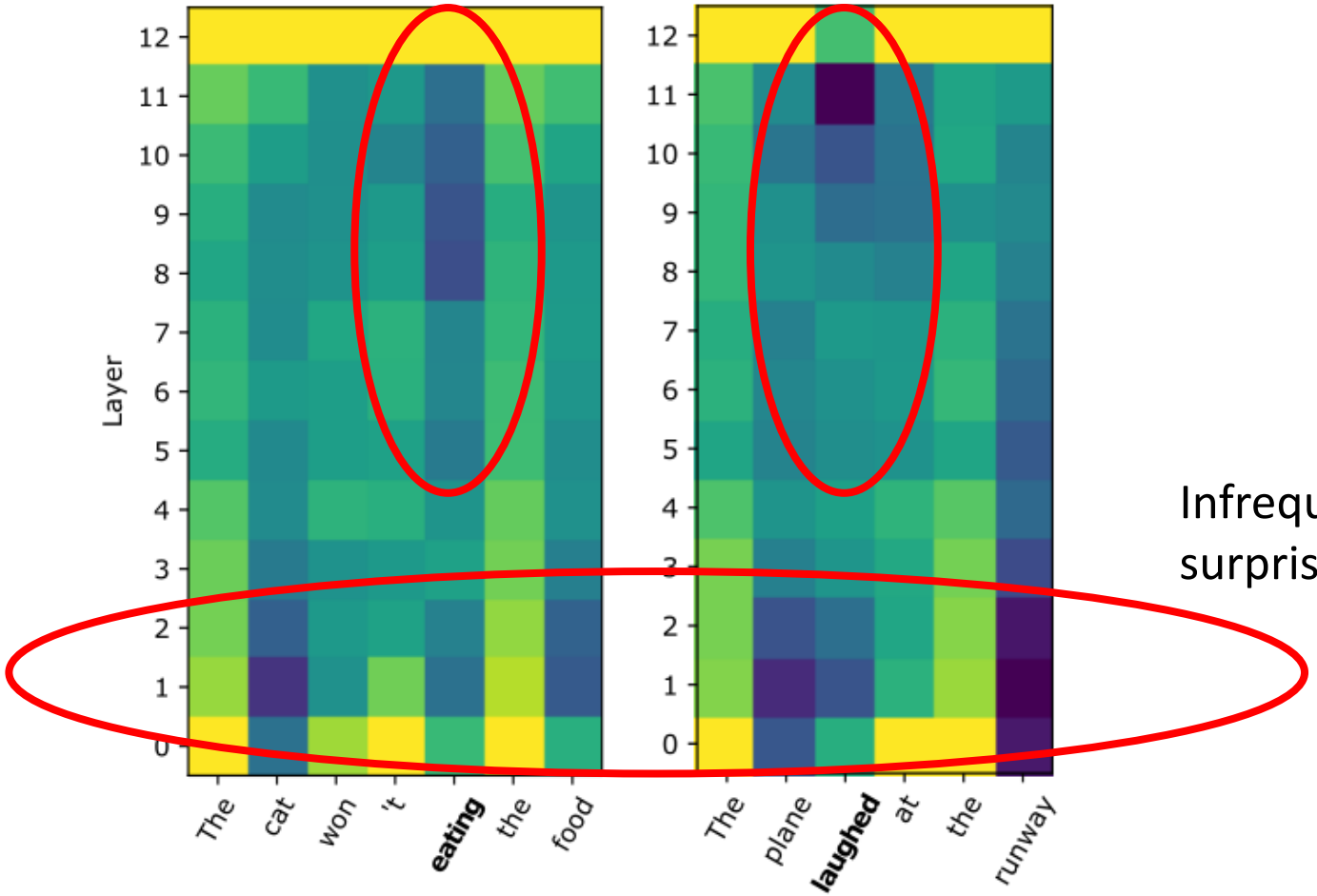
- Idea: Train a Gaussian model (one per layer) on BERT embeddings from in-domain text (BNC).
- Surprisal of new point = log likelihood according to this Gaussian distribution.



- Fun fact: if using Gaussian with full covariance matrix, equivalent to Mahalanobis distance.

High surprisals starting in middle layers

High surprisals in upper layers

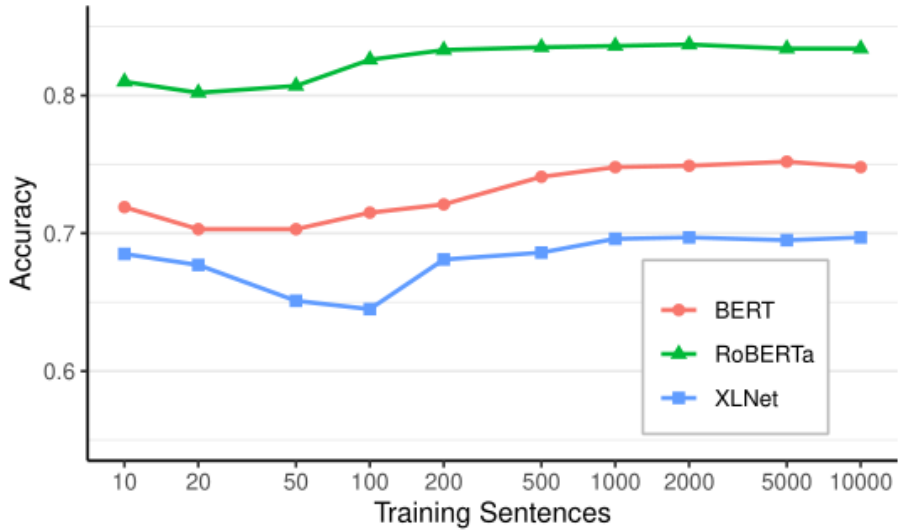


Infrequent tokens trigger high surprisal in lower layers

Morphosyntactic

Semantic

Evaluating Gaussian Model



- Use BLiMP for validation (67k grammaticality sentence pairs).
- Only takes 1k sentences for accuracy to plateau.
- RoBERTa is best performing model (0.83 accuracy on BLiMP).
- We also experimented with covariance matrix, GMM, 1-SVM.

Types of anomaly

- **Morphosyntactic**: error in inflected form of word ("*the boy **eat** the sandwich*")
- **Semantic**: violation of semantic restriction ("the **house** ate the sandwich")
- **Commonsense**: situation that's atypical in real world ("*the customer served the waitress*")

Data sources: BLiMP (3 tasks, template generated), psycholinguistic studies (9 tasks from 7 papers, written by researchers)

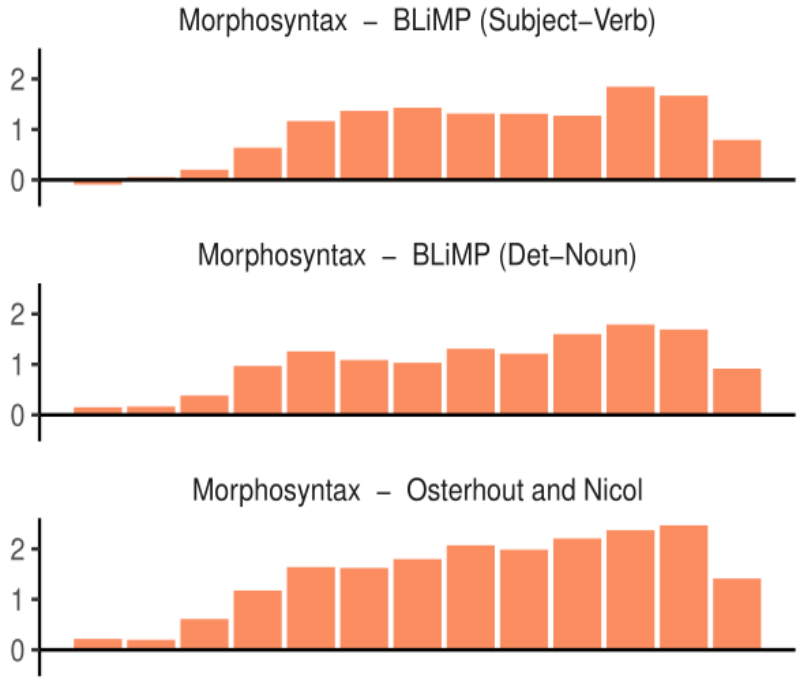
Example Sentences

Type	Task	Correct Example	Incorrect Example
Morphosyntax	BLiMP (Subject-Verb)	These casseroles **disgust** Kayla.	These casseroles **disgusts** Kayla.
	BLiMP (Det-Noun)	Craig explored that grocery **store**.	Craig explored that grocery **stores**.
	Osterhout and Nicol (1999)	The cats won't **eat** the food that Mary gives them.	The cats won't **eating** the food that Mary gives them.
Semantic	BLiMP (Animacy)	Amanda was respected by some **waitresses**.	Amanda was respected by some **picture**.
	Pylkkänen and McElree (2007)	The pilot **flew** the airplane after the intense class.	The pilot **amazed** the airplane after the intense class.
	Warren et al. (2015)	Corey's hamster **explored** a nearby backpack and filled it with sawdust.	Corey's hamster **entertained** a nearby backpack and filled it with sawdust.
	Osterhout and Nicol (1999)	The cats won't **eat** the food that Mary gives them.	The cats won't **bake** the food that Mary gives them.
	Osterhout and Mobley (1995)	The plane sailed through the air and **landed** on the runway.	The plane sailed through the air and **laughed** on the runway.
Commonsense	Warren et al. (2015)	Corey's hamster **explored** a nearby backpack and filled it with sawdust.	Corey's hamster **lifted** a nearby backpack and filled it with sawdust.
	Federmeier and Kutas (1999)	"Checkmate," Rosalie announced with glee. She was getting to be really good at **chess**.	"Checkmate," Rosalie announced with glee. She was getting to be really good at **monopoly**.
	Chow et al. (2016)	The restaurant owner forgot which **customer** the **waitress** had served.	The restaurant owner forgot which **waitress** the **customer** had served.
	Urbach and Kutas (2010)	Prosecutors accuse **defendants** of committing a crime.	Prosecutors accuse **sheriffs** of committing a crime.

Surprisal Gap (RoBERTa)

Calculate difference in surprisals, scaled by standard deviation:

$$\text{surprisal gap}_L(\mathcal{D}) = \frac{\mathbb{E}\{\text{surprisal}_L(\boldsymbol{s}'_i) - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^n}{\sigma\{\text{surprisal}_L(\boldsymbol{s}'_i) - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^n}$$



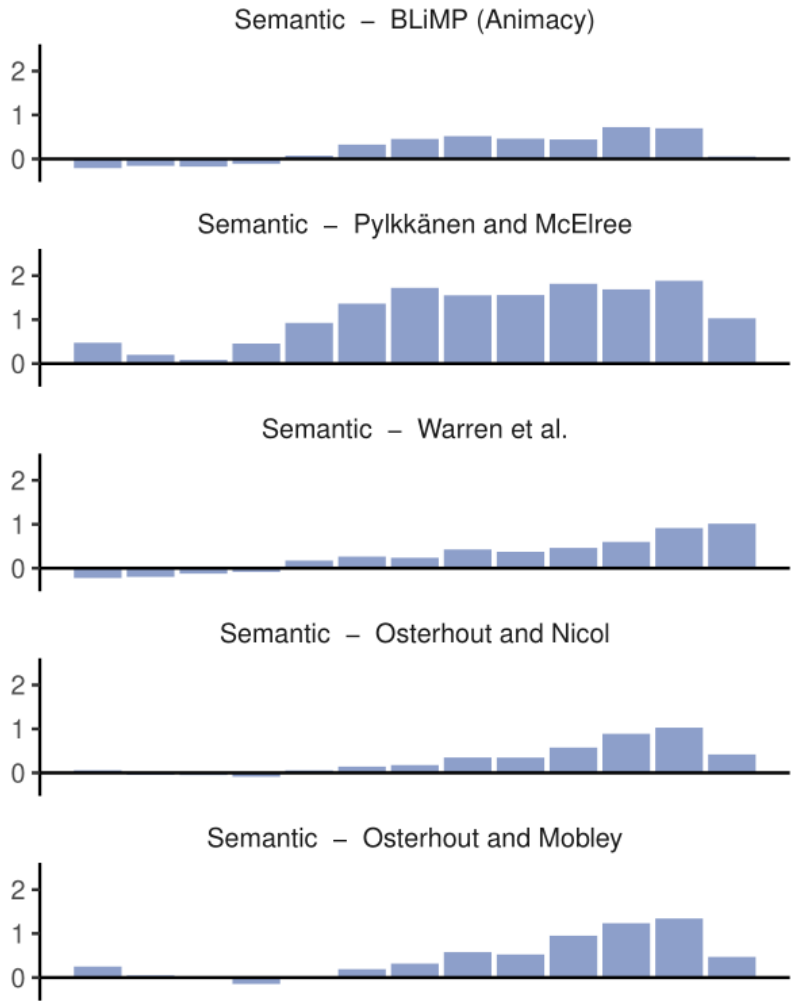
Morphosyntactic anomalies: produce surprisals starting from layers 3-4.

Surprisal Gap (RoBERTa)

Calculate difference in surprisals, scaled by standard deviation:

$$\text{surprisal gap}_L(\mathcal{D}) = \frac{\mathbb{E}\{\text{surprisal}_L(\boldsymbol{s}'_i) - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^{n}}{\sigma\{\text{surprisal}_L(\boldsymbol{s}'_i) - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^{n}}$$

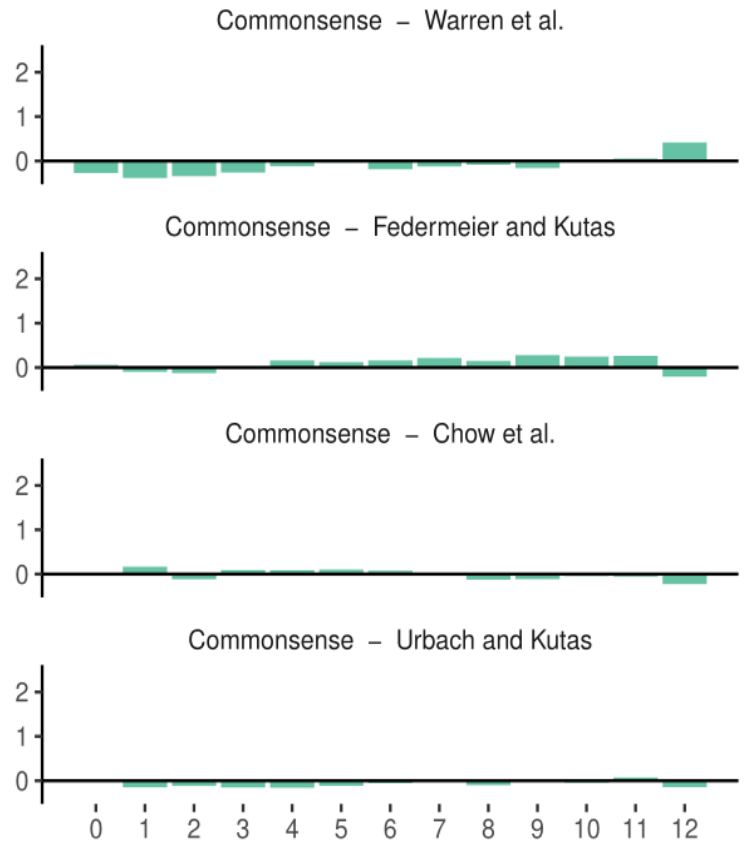
Semantic anomalies: low surprisals until upper layers (9 and above).



Surprisal Gap (RoBERTa)

Calculate difference in surprisals, scaled by standard deviation:

$$\text{surprisal gap}_L(\mathcal{D}) = \frac{\mathbb{E}\{\text{surprisal}_L(\boldsymbol{s}'_i) - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^n}{\sigma\{\text{surprisal}_L(\boldsymbol{s}'_i) - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^n}$$



Commonsense anomalies: no surprisals at any layer.

Comparing vs MLM

How does Gaussian anomaly model compare vs masked language model?

Type	Task	Size	RoBERTa	
			GM	MLM
Morphosyntax	BLiMP (Subject-Verb)	2000	0.971	0.957
	BLiMP (Det-Noun)	2000	0.983	0.999
	Osterhout and Nicol (1999)	90	1.000	1.000
Semantic	BLiMP (Animacy)	2000	0.767	0.754
	Pylkkänen and McElree (2007)	70	0.932	0.955
	Warren et al. (2015)	30	0.944	1.000
	Osterhout and Nicol (1999)	90	0.841	1.000
	Osterhout and Mobley (1995)	90	0.906	0.981
Commonsense	Warren et al. (2015)	30	0.750	0.450
	Federmeier and Kutas (1999)	34	0.583	0.875
	Chow et al. (2016)	44	0.432	n/a
	Urbach and Kutas (2010)	120	0.485	0.939

- MLM usually better than GM
- Bigger difference in commonsense tasks, less in morphosyntactic tasks
- Conclusion: RoBERTa uses different mechanisms to solve MLM, depending on the type of anomaly.

Conclusions

Proposed a new method to measure surprisals at intermediate layers of language models.

Validated Gaussian model on BLiMP, training requires only a small amount of in-domain data.

RoBERTa produces different patterns depending on type of linguistic anomaly (morphosyntactic vs semantic vs commonsense).