

How is BERT surprised? Layerwise detection of linguistic anomalies

Bai Li^{1,4}, Zining Zhu^{1,4}, Guillaume Thomas², Yang Xu^{1,3,4}, Frank Rudzicz^{1,4,5}

¹ University of Toronto, Department of Computer Science

² University of Toronto, Department of Linguistics

³ University of Toronto, Cognitive Science Program

⁴ Vector Institute for Artificial Intelligence ⁵ Unity Health Toronto

{bai, zining, yangxu, frank}@cs.toronto.edu

guillaume.thomas@utoronto.ca

Abstract

Transformer language models have shown remarkable ability in detecting when a word is anomalous in context, but likelihood scores offer no information about the *cause* of the anomaly. In this work, we use Gaussian models for density estimation at intermediate layers of three language models (BERT, RoBERTa, and XLNet), and evaluate our method on BLiMP, a grammaticality judgement benchmark. In lower layers, surprisal is highly correlated to low token frequency, but this correlation diminishes in upper layers. Next, we gather datasets of morphosyntactic, semantic, and commonsense anomalies from psycholinguistic studies; we find that the best performing model RoBERTa exhibits surprisal in earlier layers when the anomaly is morphosyntactic than when it is semantic, while commonsense anomalies do not exhibit surprisal at any intermediate layer. These results suggest that language models employ separate mechanisms to detect different types of linguistic anomalies.

1 Introduction

Transformer-based language models (LMs) have achieved remarkable success in numerous natural language processing tasks, prompting many probing studies to determine the extent of their linguistic knowledge. A popular approach is to formulate the problem as a multiple-choice task, where the LM is considered correct if it assigns higher likelihood to the appropriate word than an inappropriate one, given context (Gulordava et al., 2018; Ettinger, 2020; Warstadt et al., 2020). The likelihood score, however, only gives a scalar value of the degree that a word is anomalous in context, and cannot distinguish between different *ways* that a word might be anomalous.

It has been proposed that there are different types of linguistic anomalies. Chomsky

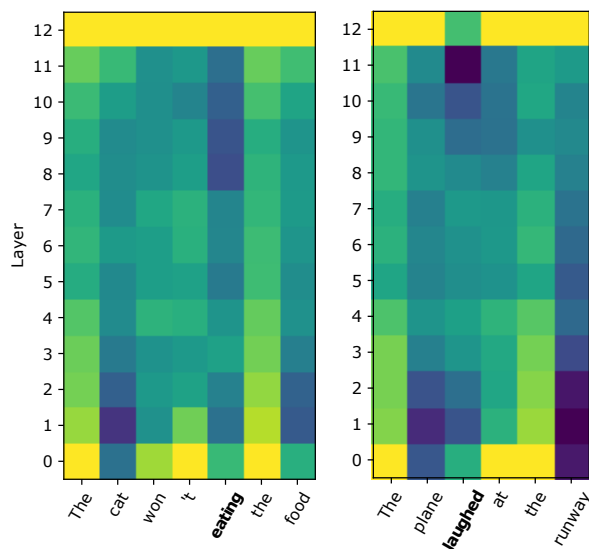


Figure 1: Example sentence with a morphosyntactic anomaly (left) and semantic anomaly (right) (anomalies in bold). Darker colours indicate higher surprisal. We investigate several patterns: first, surprisal at lower layers corresponds to infrequent tokens, but this effect diminishes towards upper layers. Second, morphosyntactic violations begin to trigger high surprisals at an earlier layer than semantic violations.

(1957) distinguished semantic anomalies (“*colorless green ideas sleep furiously*”) from ungrammaticality (“*furiously sleep ideas green colorless*”). Psycholinguistic studies initially suggested that different event-related potentials (ERPs) are produced in the brain depending on the type of anomaly; e.g., semantic anomalies produce negative ERPs 400 ms after the stimulus, while syntactic anomalies produce positive ERPs 600 ms after (Kutas et al., 2006). Here, we ask whether Transformer LMs show different surprisals in their intermediate layers depending on the type of anomaly. However, LMs do not compute likelihoods at intermediate layers – only at the final layer.

In this paper, we introduce a new tool to probe for surprisal at intermediate layers of BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), formulating the problem as density estimation. We train Gaussian models to fit distributions of embeddings at each layer of the LMs. Using BLiMP (Warstadt et al., 2020) for evaluation, we show that this model is effective at grammaticality judgement, requiring only a small amount of in-domain text for training. Figure 1 shows the method using the RoBERTa model on two example sentences.

We apply our model to test sentences drawn from BLiMP and 7 psycholinguistics studies, exhibiting morphosyntactic, semantic, and commonsense anomalies. We find that morphosyntactic anomalies produce out-of-domain embeddings at earlier layers, semantic anomalies at later layers, and no commonsense anomalies, even though the LM’s final accuracy is similar. We show that LMs are internally sensitive to the type of linguistic anomaly, which is not apparent if we only had access to their softmax probability outputs. Our source code and data are available at: <https://github.com/SPOCLab-ca/layerwise-anomaly>.

2 Related work

2.1 Probing LMs for linguistic knowledge

Soon after BERT’s release, many papers invented probing techniques to discover what linguistic knowledge it contains, and how this information is distributed between layers (e.g., Rogers et al. (2021) provides a comprehensive overview). Tenney et al. (2019) used “edge probing” to determine each layer’s contribution to a task’s performance, and discovered that the middle layers contributed more when the task was syntactic, and the upper layers more when the task was semantic.

Several papers found that BERT’s middle layers contain the most syntactic information. Kelly et al. (2020) found that BERT’s middle layers are best at distinguishing between sentences with direct and indirect object constructions. Hewitt and Manning (2019) used a structural probe to recover syntax trees from contextual embeddings, and found the performance peaked in middle layers.

Probing results are somewhat dependent on the choice of linguistic formalism used to annotate the data, as Kulmizev et al. (2020) found for syntax, and Kuznetsov and Gurevych (2020) found for se-

mantic roles. Miaschi et al. (2020) examined the layerwise performance of BERT for a suite of linguistic features, before and after fine tuning. Our work further investigates what linguistic information is contained in different layers, with a focus on anomalous inputs.

2.2 Neural grammaticality judgements

Many recent probing studies used grammaticality judgement tasks to test the knowledge of specific phenomena in LMs. Warstadt et al. (2019) gathered sentences from linguistic publications, and evaluated by Matthews Correlation with the ground truth. More commonly, the model is presented with a binary choice between an acceptable and unacceptable sentence: BLiMP (Warstadt et al., 2020) used templates to generate 67k such sentence pairs, covering 12 types of linguistic phenomena. Similarly, Hu et al. (2020) created syntactic tests using templates, but defined success criteria using inequalities of LM perplexities.

In contrast with artificial templates, Gulordava et al. (2018) generated test cases by perturbing natural corpus data to test long-distance dependencies. Most grammaticality studies focused on syntactic phenomena, but Rabinovich et al. (2019) tested LMs’ sensitivity to semantic infelicities involving indefinite pronouns.

2.3 Tests of selectional restrictions

Violations of selectional restrictions are one type of linguistic unacceptability, defined as a semantic mismatch between a verb and an argument. Sasano and Korhonen (2020) examined the geometry of word classes (e.g., words that can be a direct object of the verb ‘play’) in word vector models; they compared single-class models against discriminative models for learning word class boundaries. Chersoni et al. (2018) tested distributional semantic models on their ability to identify selectional restriction violations using stimuli from two psycholinguistic datasets. Finally, Metheniti et al. (2020) tested how much BERT relies on selectional restriction information versus other contextual information for making masked word predictions.

2.4 Psycholinguistic tests for LMs

The N400 response is a negative event-related potential that occurs roughly 400ms after a stimulus in human brains, and is generally associated with the stimulus being semantically anomalous with

respect to the preceding context (Kutas and Federmeier, 2011). Although many studies have been performed with a diverse range of linguistic stimuli, exactly what conditions trigger the N400 response is still an open question. Frank et al. (2015) found that the N400 response is correlated with surprisal, i.e., how unlikely an LM predicts a word given the preceding context.

Recently, several studies have investigated relationships between surprisal in neural LMs and the N400 response. Michaelov and Bergen (2020) compared human N400 amplitudes with LSTM-based models using stimuli from several psycholinguistic studies. Ettinger (2020) used data from three psycholinguistic studies to probe BERT’s knowledge of commonsense and negation. Our work is similar to the latter – we leverage psycholinguistic studies for their stimuli, but we do not use their N400 amplitude results.

3 Model

We use the transformer language model as a contextual embedding extractor (we write this as BERT for convenience). Let L be the layer index, which ranges from 0 to 12 on all of our models. Using a training corpus $\{w_1, \dots, w_T\}$, we extract contextual embeddings at layer L for each token:

$$\mathbf{x}_1^{(L)}, \dots, \mathbf{x}_T^{(L)} = \text{BERT}_L(w_1, \dots, w_T). \quad (1)$$

Next, we fit a multivariate Gaussian on the extracted embeddings:

$$\mathbf{x}_1^{(L)}, \dots, \mathbf{x}_T^{(L)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_L, \hat{\boldsymbol{\Sigma}}_L). \quad (2)$$

For evaluating the layerwise surprisal of a new sentence $\mathbf{s} = [t_1, \dots, t_n]$, we similarly extract contextual embeddings using the language model:

$$\mathbf{y}_1, \dots, \mathbf{y}_n = \text{BERT}_L(t_1, \dots, t_n). \quad (3)$$

The surprisal of each token is the negative log likelihood of the contextual vector according to the multivariate Gaussian:

$$G_i = -\log p(\mathbf{y}_i | \hat{\boldsymbol{\mu}}_L, \hat{\boldsymbol{\Sigma}}_L) \quad \text{for } i = 1 \dots n. \quad (4)$$

Finally, we define the surprisal of sentence \mathbf{s} as the sum of surprisals of all of its tokens, which is also the joint log likelihood of all of the embeddings:

$$\begin{aligned} \text{surprisal}_L(t_1, \dots, t_n) &= \sum_{i=1}^n G_i \\ &= -\log p(\mathbf{y}_1, \dots, \mathbf{y}_n | \hat{\boldsymbol{\mu}}_L, \hat{\boldsymbol{\Sigma}}_L). \end{aligned} \quad (5)$$

3.1 Connection to Mahalanobis distance

The theoretical motivation for using the sum of log likelihoods is that when we fit a Gaussian model with full covariance matrix, low likelihood corresponds exactly to high Mahalanobis distance from the in-distribution points. The score given by the Gaussian model is:

$$\begin{aligned} G &= -\log p(\mathbf{y} | \hat{\boldsymbol{\mu}}_L, \hat{\boldsymbol{\Sigma}}_L) \\ &= -\log \left(\frac{1}{(2\pi)^{D/2} |\hat{\boldsymbol{\Sigma}}_L|^{1/2}} \exp\left(-\frac{1}{2} d^2\right) \right), \end{aligned} \quad (6)$$

where D is the dimension of the vector space, and d is the Mahalanobis distance:

$$d = \sqrt{(\mathbf{y} - \hat{\boldsymbol{\mu}}_L)^T \hat{\boldsymbol{\Sigma}}_L^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_L)}. \quad (7)$$

Rearranging, we get:

$$d^2 = 2G - D \log(2\pi) - \log |\hat{\boldsymbol{\Sigma}}_L|, \quad (8)$$

thus the negative log likelihood is the squared Mahalanobis distance plus a constant.

Various methods based on Mahalanobis distance have been used for anomaly detection in neural networks; for example, Lee et al. (2018) proposed a similar method for out-of-domain detection in neural classification models, and Cao et al. (2020) found the Mahalanobis distance method to be competitive with more sophisticated methods on medical out-of-domain detection. In Transformer models, Podolskiy et al. (2021) used Mahalanobis distance for out-of-domain detection, outperforming methods based on softmax probability and likelihood ratios.

Gaussian assumptions. Our model assumes that the embeddings at every layer follow a multivariate Gaussian distribution. Since the Gaussian distribution is the maximum entropy distribution given a mean and covariance matrix, it makes the fewest assumptions and is therefore a reasonable default. Hennigen et al. (2020) found that embeddings sometimes do not follow a Gaussian distribution, but it is unclear what alternative distribution would be a better fit, so we will assume a Gaussian distribution in this work.

3.2 Training and evaluation

For all of our experiments, we use the ‘base’ versions of pretrained language models BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and

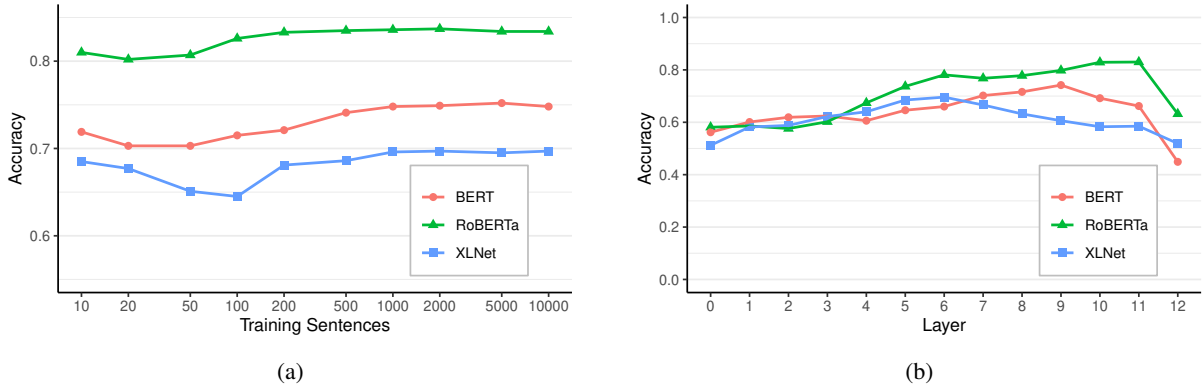


Figure 2: BLiMP accuracy different amounts of training data and across layers, for three LMs. About 1000 sentences are needed before a plateau is reached (mean tokens per sentence = 15.1).

XLNet (Yang et al., 2019), provided by HuggingFace (Wolf et al., 2020). Each of these models have 12 contextual layers plus a 0th static layer, and each layer is 768-dimensional.

We train the Gaussian model on randomly selected sentences from the British National Corpus (Leech, 1992), representative of acceptable English text from various genres. We evaluate on BLiMP (Warstadt et al., 2020), a dataset of 67k minimal sentence pairs that test acceptability judgements across a variety of syntactic and semantic phenomena. In our case, a sentence pair is considered correct if the sentence-level surprisal of the unacceptable sentence is higher than that of the acceptable sentence.

How much training data is needed? We experiment with training data sizes ranging from 10 to 10,000 sentences (Figure 2a). Compared to the massive amount of data needed for pretraining the LMs, we find that a modest corpus suffices for training the Gaussian anomaly model, and a plateau is reached after 1000 sentences for all three models. Therefore, we use 1000 training sentences (unless otherwise noted) for all subsequent experiments in this paper.

Which layers are sensitive to anomaly? We vary L from 0 to 12 in all three models (Figure 2b). The layer with the highest accuracy differs between models: layer 9 has the highest accuracy for BERT, 11 for RoBERTa, and 6 for XLNet. All models experience a sharp drop in the last layer, likely because the last layer is specialized for the MLM pretraining objective.

Comparisons to other models. Our best-performing model is RoBERTa, with an accuracy of 0.830. This is slightly higher the best model reported in BLiMP (GPT-2, with accuracy 0.801).

We do not claim to beat the state-of-the-art on BLiMP: Salazar et al. (2020) obtains a higher accuracy of 0.865 using RoBERTa-large. Even though the main goal of this paper is not to maximize accuracy on BLiMP, our Gaussian anomaly model is competitive with other transformer-based models on this task.

In Appendix A, we explore variations of the Gaussian anomaly model, such as varying the type of covariance matrix, Gaussian mixture models, and one-class SVMs (Schölkopf et al., 2000). However, none of these variants offer a significant improvement over a single Gaussian model with full covariance matrix.

3.3 Lower layers are sensitive to frequency

We notice that surprisal scores in the lower layers are sensitive to token frequency: higher frequency tokens produce embeddings close to the center of the Gaussian distribution, while lower frequency tokens are at the periphery. The effect gradually diminishes towards the upper layers.

To quantify the sensitivity to frequency, we compute token-level surprisal scores for 5000 sentences from BNC that were not used in training. We then compute the Pearson correlation between the surprisal score and log frequency for each token (Figure 3). In all three models, there is a high correlation between the surprisal score and log frequency at the lower layers, which diminishes at the upper layers. A small positive correlation persists until the last layer, except for XLNet, in which the correlation eventually disappears.

There does not appear to be any reports of this phenomenon in previous work. For static word vectors, Gong et al. (2018) found that embeddings for low-frequency words lie in a different region of

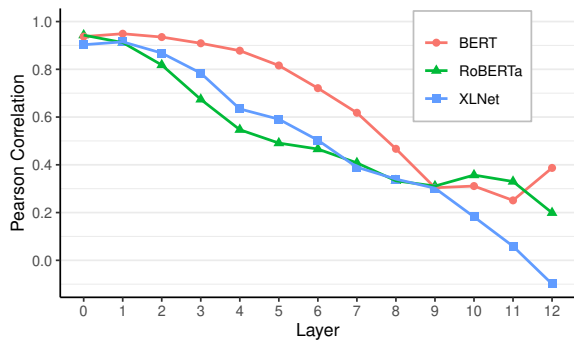


Figure 3: Pearson correlation between token-level surprisal scores (Equation 4) and log frequency. The correlation is highest in the lower layers, and decreases in the upper layers.

the embedding space than high-frequency words. We find evidence that the same phenomenon occurs in contextual embeddings (Appendix B). In this scenario, the Gaussian model fits the high-frequency region and assigns lower likelihoods to the low-frequency region, explaining the positive correlation at all layers; however, it is still unclear why the correlation diminishes at upper layers.

4 Levels of linguistic anomalies

We turn to the question of whether LMs exhibit different behaviour when given inputs with different types of linguistic anomalies. The task of partitioning linguistic anomalies into several distinct classes can be challenging. Syntax and semantics have a high degree of overlap – there is no widely accepted criterion for distinguishing between ungrammaticality and semantic anomaly (e.g., Abrusán (2019) gives a survey of current proposals), and Poulsen (2012) challenges this dichotomy entirely. Similarly, Warren et al. (2015) noted that semantic anomalies depend somewhat on world knowledge.

Within a class, the anomalies are also heterogeneous (e.g., ungrammaticality may be due to violations of agreement, *wh*-movement, negative polarity item licensing, etc), which might each affect the LMs differently. Thus, we define three classes of anomalies that do not attempt to cover all possible linguistic phenomena, but captures different levels of language processing while retaining internal uniformity:

1. **Morphosyntactic anomaly:** an error in the inflected form of a word, for example, subject-verb agreement (**the boy eat the*

sandwich), or incorrect verb tense or aspect inflection (**the boy eaten the sandwich*). In each case, the sentence can be corrected by changing the inflectional form of one word.

2. **Semantic anomaly:** a violation of a selectional restriction, such as animacy (*#the house eats the sandwich*). In these cases, the sentence can be corrected by replacing one of the verb’s arguments with another one in the same word class that satisfies the verb’s selectional restrictions.
3. **Commonsense anomaly:** sentence describes an situation that is atypical or implausible in the real world but is otherwise acceptable (*#the customer served the waitress*).

4.1 Summary of anomaly datasets

We use two sources of data for experiments on linguistic anomalies: synthetic sentences generated from templates, and materials from psycholinguistic studies. Both have advantages and disadvantages – synthetic data can be easily generated in large quantities, but the resulting sentences may be odd in unintended ways. Psycholinguistic stimuli are designed to control for confounding factors (e.g., word frequency) and human-validated for acceptability, but are smaller (typically fewer than 100 sentence pairs).

We curate a set of 12 tasks from BLiMP and 7 psycholinguistic studies¹. Each sentence pair consists of a control and an anomalous sentence, so that all sentences within a task differ in a consistent manner. Table 1 shows an example sentence pair from each task. We summarize each dataset:

1. BLiMP (Warstadt et al., 2020): we use subject-verb and determiner-noun agreement tests as morphosyntactic anomaly tasks. For simplicity, we only use the basic regular sentences, and exclude sentences involving irregular words or distractor items. We also use the two argument structure tests involving animacy as a semantic anomaly task. All three BLiMP tasks therefore have 2000 sentence pairs.

¹Several of these stimuli have been used in natural language processing research. Chersoni et al. (2018) used the data from Pykkänen and McElree (2007) and Warren et al. (2015) to probe word vectors for knowledge of selectional restrictions. Ettinger (2020) used data from Federmeier and Kutas (1999) and Chow et al. (2016), which were referred to as CPRAG-102 and ROLE-88 respectively.

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 .
Semantic	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.
	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.
Commonsense	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.
	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.

Table 1: Example sentence pair for each of the 12 tasks. The 3 BLiMP tasks are generated from templates; the others are stimuli materials taken from psycholinguistic studies.

- Osterhout and Nicol (1999): contains 90 sentence triplets containing a control, syntactic, and semantic anomaly. Syntactic anomalies involve a modal verb followed by a verb in *-ing* form; semantic anomalies have a selectional restriction violation between the subject and verb. There are also double anomalies (simultaneously syntactic and semantic) which we do not use.
- Pylkkänen and McElree (2007): contains 70 sentence pairs where the verb is replaced in the anomalous sentence with one that requires an animate object, thus violating the selectional restriction. In half the sentences, the verb is contained in an embedded clause.
- Warren et al. (2015): contains 30 sentence triplets with a possible condition, a selectional restriction violation between the subject and verb, and an impossible condition where the subject cannot carry out the action, i.e., a commonsense anomaly.
- Osterhout and Mobley (1995): we use data from experiment 2, containing 90 sentence pairs where the verb in the anomalous sentence is semantically inappropriate. The experiment also tested gender agreement errors, but we do not include these stimuli.
- Federmeier and Kutas (1999): contains 34 sentence pairs, where the final noun in each anomalous sentence is an inappropriate completion, but in the same semantic category as the expected completion.
- Chow et al. (2016): contains 44 sentence pairs, where two of the nouns in the anomalous sentence are swapped to reverse their roles. This is the only task in which the sentence pair differs by more than one token.
- Urbach and Kutas (2010): contains 120 sentence pairs, where the anomalous sentence replaces a patient of the verb with an atypical one.

4.2 Quantifying layerwise surprisal

Let $\mathcal{D} = \{(s_1, s'_1), \dots, (s_n, s'_n)\}$ be a dataset of sentence pairs, where s_i is a control sentence and s'_i is an anomalous sentence. For each layer L , we define the *surprisal gap* as the mean difference of surprisal scores between the control and anoma-

lous sentences, scaled by the standard deviation:

$$\text{surprisal gap}_L(\mathcal{D}) = \frac{\mathbb{E}\{\text{surprisal}_L(s'_i) - \text{surprisal}_L(s_i)\}_{i=1}^n}{\sigma\{\text{surprisal}_L(s'_i) - \text{surprisal}_L(s_i)\}_{i=1}^n} \quad (9)$$

The surprisal gap is a scale-invariant measure of sensitivity to anomaly, similar to a signal-to-noise ratio. While surprisal scores are unitless, the surprisal gap may be viewed as the number of standard deviations that anomalous sentences trigger surprisal above control sentences. This is advantageous over accuracy scores, which treats the sentence pair as correct when the anomalous sentence has higher surprisal by any margin; this hard cutoff masks differences in the magnitude of surprisal. The metric also allows for fair comparison of surprisal scores across datasets of vastly different sizes. Figure 4 shows the surprisal gap for all 12 tasks, using the RoBERTa model; the results for BERT and XLNet are in the Appendix C.

Next, we compare the performance of the Gaussian model with the masked language model (MLM). We score each instance as correct if the masked probability of the correct word is higher than the anomalous word. One limitation of the MLM approach is that it requires the sentence pair to be identical in all places except for one token, since the LMs do not support modeling joint probabilities over multiple tokens. To ensure fair comparison between GM and MLM, we exclude instances where the differing token is out-of-vocabulary in any of the LMs (this excludes approximately 30% of instances). For the Gaussian model, we compute accuracy using the best-performing layer for each model (Section 3.2). The results are listed in Table 2.

5 Discussion

5.1 Anomaly type and surprisal

Morphosyntactic anomalies generally appear earlier than semantic anomalies (Figure 4). The surprisal gap plot exhibits different patterns depending on the type of linguistic anomaly: morphosyntactic anomalies produce high surprisal relatively early (layers 3-4), while semantic anomalies produce low surprisals until later (layers 9 and above). Commonsense anomalies do not result in surprisals at *any* layer: the surprisal gap is near zero for all of the commonsense tasks. The observed difference between morphosyntactic and semantic

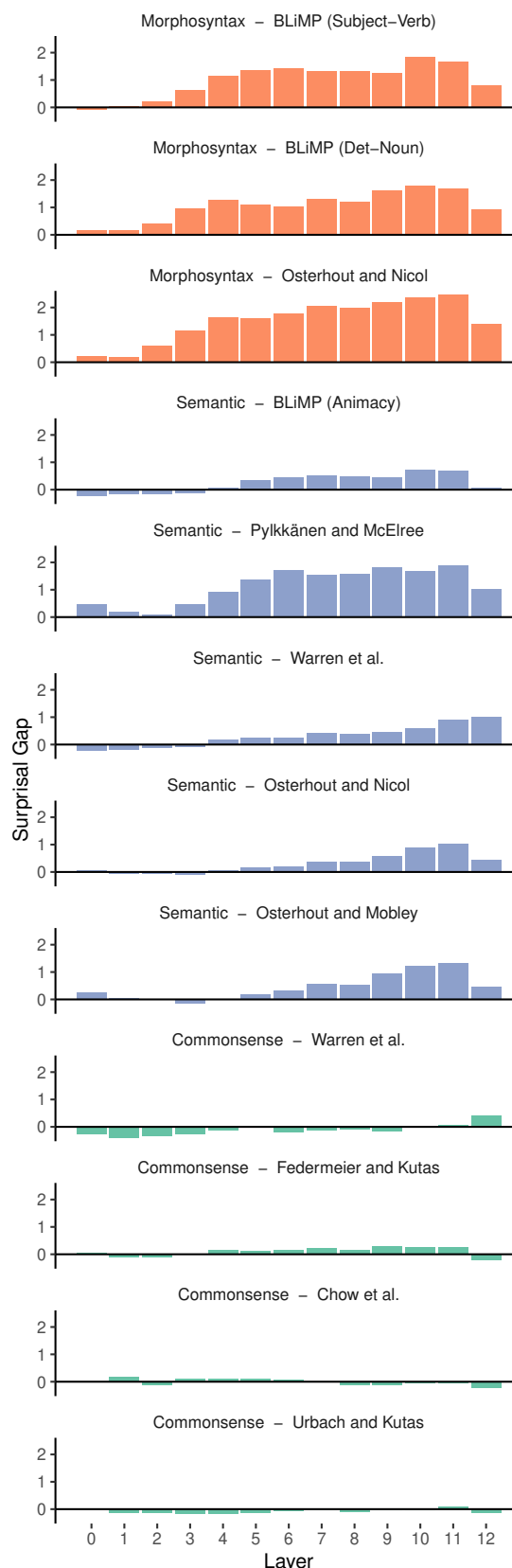


Figure 4: Layerwise surprisal gaps for all tasks using the RoBERTa model. Generally, a positive surprisal gap appears in earlier layers for morphosyntactic tasks than for semantic tasks; no surprisal gap appears at any layer for commonsense tasks.

Type	Task	Size	BERT		RoBERTa		XLNet	
			GM	MLM	GM	MLM	GM	MLM
Morphosyntax	BLiMP (Subject-Verb)	2000	0.953	0.955	0.971	0.957	0.827	0.584
	BLiMP (Det-Noun)	2000	0.970	0.999	0.983	0.999	0.894	0.591
	Osterhout and Nicol (1999)	90	1.000	1.000	1.000	1.000	0.901	0.718
Semantic	BLiMP (Animacy)	2000	0.644	0.787	0.767	0.754	0.675	0.657
	Pykkänen and McElree (2007)	70	0.727	0.955	0.932	0.955	*0.636	0.727
	Warren et al. (2015)	30	*0.556	1.000	0.944	1.000	*0.667	*0.556
	Osterhout and Nicol (1999)	90	0.681	0.957	0.841	1.000	*0.507	0.783
	Osterhout and Mobley (1995)	90	*0.528	1.000	0.906	0.981	*0.302	0.774
Commonsense	Warren et al. (2015)	30	*0.600	*0.550	0.750	*0.450	*0.300	*0.600
	Federmeier and Kutas (1999)	34	*0.458	*0.708	*0.583	0.875	*0.625	*0.667
	Chow et al. (2016)	44	*0.591	n/a	*0.432	n/a	*0.568	n/a
	Urbach and Kutas (2010)	120	*0.470	0.924	*0.485	0.939	*0.500	0.712

Table 2: Comparing accuracy scores between Gaussian anomaly model (GM) and masked language model (MLM) for all models and tasks. Asterisks indicate that the accuracy is not better than random (0.5), using a binomial test with threshold of $p < 0.05$ for significance. The MLM results for [Chow et al. \(2016\)](#) are excluded because the control and anomalous sentences differ by more than one token. The best layers for each model (Section 3.2) are used for GM, and the last layer is used for MLM. Generally, MLM outperforms GM, and the difference is greater for semantic and commonsense tasks.

anomalies is consistent with previous work ([Tenney et al., 2019](#)), which found that syntactic information appeared earlier in BERT than semantic information.

One should be careful and avoid drawing conclusions from only a few experiments. A similar situation occurred in psycholinguistics research ([Kutas et al., 2006](#)): early results suggested that the N400 was triggered by semantic anomalies, while syntactic anomalies triggered the P600 – a different type of ERP. However, subsequent experiments found exceptions to this rule, and now it is believed that the N400 cannot be categorized by any standard dichotomy, like syntax versus semantics ([Kutas and Federmeier, 2011](#)). In our case, [Pykkänen and McElree \(2007\)](#) is an exception: the task is a semantic anomaly, but produces surprisals in early layers, similar to the morphosyntactic tasks. Hence it is possible that the dichotomy is something other than syntax versus semantics; we leave to future work to determine more precisely what conditions trigger high surprisals in lower versus upper layers of LMs.

5.2 Comparing anomaly model with MLM

The masked language model (MLM) usually outperforms the Gaussian anomaly model (GM), but the difference is uneven. MLM performs much better than GM on commonsense tasks, slightly better on semantic tasks, and about the same or slightly worse on morphosyntactic tasks. It is not obvious why MLM should perform better than GM, but we note two subtle differences between the MLM and GM setups that may be contributing

factors. First, the GM method adds up the surprisal scores for the whole sequence, while MLM only considers the softmax distribution at one token. Second, the input sequence for MLM always contains a [MASK] token, whereas GM takes the original unmasked sequences as input, so the representations are never identical between the two setups.

MLM generally outperforms GM, but it does not solve every task: all three LMs fail to perform above chance on the data from [Warren et al. \(2015\)](#). This set of stimuli was designed so that both the control and impossible completions are not very likely or expected, which may have caused the difficulty for the LMs. We excluded the task of [Chow et al. \(2016\)](#) for MLM because the control and anomalous sentences differed by more than one token².

5.3 Differences between LMs

RoBERTa is the best-performing of the three LMs in both the GM and MLM settings: this is expected since it is trained with the most data and performs well on many natural language benchmarks. Surprisingly, XLNet is ill-suited for this task and performs worse than BERT, despite having a similar model capacity and training data.

The surprisal gap plots for BERT and XL-

²Sentence pairs with multiple differing tokens are inconvenient for MLM to handle, but this is not a fundamental limitation. For example, [Salazar et al. \(2020\)](#) proposed a modification to MLM to handle such cases: they compute a *pseudo-log-likelihood* score for a sequence by replacing one token at a time with a [MASK] token, applying MLM to each masked sequence, and summing up the log likelihood scores.

Net (Appendix C) show some differences from RoBERTa: only morphosyntactic tasks produce out-of-domain embeddings in these two models, and not semantic or commonsense tasks. Evidently, how LMs behave when presented with anomalous inputs is dependent on model architecture and training data size; we leave exploration of this phenomenon to future work.

6 Conclusion

We use Gaussian models to characterize out-of-domain embeddings at intermediate layers of Transformer language models. The model requires a relatively small amount of in-domain data. Our experiments reveal that out-of-domain points in lower layers correspond to low-frequency tokens, while grammatically anomalous inputs are out-of-domain in higher layers. Furthermore, morphosyntactic anomalies are recognized as out-of-domain starting from lower layers compared to syntactic anomalies. Commonsense anomalies do not generate out-of-domain embeddings at any layer, even when the LM has a preference for the correct cloze completion. These results show that depending on the type of linguistic anomaly, LMs use different mechanisms to produce the output softmax distribution.

Acknowledgements

We thank Julian Salazar and our anonymous reviewers for their helpful suggestions. YX is funded through an NSERC Discovery Grant, a SSHRC Insight Grant, and an Ontario ERA award. FR is supported by a CIFAR Chair in Artificial Intelligence.

References

Márta Abrusán. 2019. Semantic anomaly, pragmatic infelicity, and ungrammaticality. *Annual Review of Linguistics*, 5:329–351.

Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. 2020. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*.

Emmanuele Chersoni, Adrià Torrens Urrutia, Philippe Blache, and Alessandro Lenci. 2018. Modeling violations of selectional restrictions with distributional semantics. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 20–29.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co.

Wing-Yee Chow, Cybelle Smith, Ellen Lau, and Colin Phillips. 2016. A “bag-of-arguments” mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5):577–596.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kara D Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, 41(4):469–495.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: Frequency-agnostic word representation. In *Advances in neural information processing systems*, pages 1334–1345.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. Association for Computational Linguistics.

- MA Kelly, Yang Xu, Jesús Calvillo, and David Reitter. 2020. Which sentence embeddings and which layers encode syntactic structure? In *Cognitive Science*, pages 2375–2381.
- Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091.
- Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62:621–647.
- Marta Kutas, Cyma K Van Petten, and Robert Kluender. 2006. Psycholinguistics electrified II (1994–2005). In *Handbook of psycholinguistics*, pages 659–724. Elsevier.
- Iliia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182. Association for Computational Linguistics.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31:7167–7177.
- Geoffrey Neil Leech. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research*, 28:1–13.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How relevant are selectional preferences for Transformer-based language models? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1266–1278.
- Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *The 28th International Conference on Computational Linguistics*, pages 745–756.
- James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663.
- Lee Osterhout and Linda A Mobley. 1995. Event-related brain potentials elicited by failure to agree. *Journal of Memory and language*, 34(6):739–773.
- Lee Osterhout and Janet Nicol. 1999. On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and cognitive processes*, 14(3):283–317.
- Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting Mahalanobis distance for Transformer-based out-of-domain detection. In *35th AAAI Conference on Artificial Intelligence (AAAI 2021)*.
- Mads Poulsen. 2012. The usefulness of the grammaticality–acceptability distinction in functional approaches to language. *Acta Linguistica Hafniensia*, 44(1):4–21.
- Liina Pykkänen and Brian McElree. 2007. An MEG study of silent meaning. *Journal of cognitive neuroscience*, 19(11):1905–1921.
- Ella Rabinovich, Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2019. Say anything: Automatic semantic infelicity detection in L2 English indefinite pronouns. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 77–86.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. Association for Computational Linguistics.
- Ryohei Sasano and Anna Korhonen. 2020. Investigating word-class distributions in word vector spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3657–3666.
- Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. 2000. Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.
- Thomas P Urbach and Marta Kutas. 2010. Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2):158–179.
- Tessa Warren, Evelyn Milburn, Nikole D Patson, and Michael Walsh Dickey. 2015. Comprehending the impossible: what role do selectional restriction violations play? *Language, cognition and neuroscience*, 30(8):932–939.

- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mo-
hananey, Wei Peng, Sheng-Fu Wang, and Samuel R
Bowman. 2020. BLiMP: The benchmark of linguis-
tic minimal pairs for English. *Transactions of the
Association for Computational Linguistics*, 8:377–
392.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bow-
man. 2019. Neural network acceptability judg-
ments. *Transactions of the Association for Compu-
tational Linguistics*, 7:625–641.
- Thomas Wolf, Julien Chaumond, Lysandre Debut, Vic-
tor Sanh, Clement Delangue, Anthony Moi, Pier-
ric Cistac, Morgan Funtowicz, Joe Davison, Sam
Shleifer, et al. 2020. Transformers: State-of-the-
art natural language processing. In *Proceedings of
the 2020 Conference on Empirical Methods in Nat-
ural Language Processing: System Demonstrations*,
pages 38–45.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Car-
bonell, Russ R Salakhutdinov, and Quoc V Le. 2019.
XLNet: Generalized autoregressive pretraining for
language understanding. In *Advances in neural in-
formation processing systems*, pages 5753–5763.

A Ablation experiments on Gaussian model

We compare some variations to our methodology of training the Gaussian model. All of these variations are evaluated on the full BLiMP dataset. In each experiment, (unless otherwise noted) the language model is RoBERTa-base, using the second-to-last layer, and the Gaussian model has a full covariance matrix trained with 1000 sentences from the BNC corpus.

Covariance matrix. We vary the type of covariance matrix (Table 3). Diagonal and spherical covariance matrices perform worse than with the full covariance matrix; this may be expected, as the full matrix has the most trainable parameters.

Covariance	Accuracy
Full	0.830
Diagonal	0.755
Spherical	0.752

Table 3: Varying the type of covariance matrix in the Gaussian model.

Gaussian mixture models. We try GMMs with up to 16 mixture components (Table 4). We observe a small increase in accuracy compared to a single Gaussian, but the difference is too small to justify the increased training time.

Components	Accuracy
1	0.830
2	0.841
4	0.836
8	0.849
16	0.827

Table 4: Using Gaussian mixture models (GMMs) with multiple components.

Genre of training text. We sample from genres of BNC (each time with 1000 sentences) to train the Gaussian model (Table 5). The model performed worse when trained with the academic and spoken genres, and about the same with the fiction and news genres, perhaps because their vocabularies and grammars are more similar to those in the BLiMP sentences.

One-class SVM. We try replacing the Gaussian model with a one-class SVM (Schölkopf et al., 2000), another popular model for anomaly detection. We use the default settings from scikit-learn

Genre	Accuracy
Academic	0.797
Fiction	0.840
News	0.828
Spoken	0.795
All	0.830

Table 5: Effect of the genre of training data.

with three kernels (Table 6), but it performs worse than the Gaussian model on all settings.

Kernel	Score
RBF	0.738
Linear	0.726
Polynomial	0.725

Table 6: Using 1-SVM instead of GMM, with various kernels.

Sentence aggregation. Instead of Equation 5, we try defining sentence-level surprisal as the maximum surprisal among all tokens (Table 7):

$$\text{surprisal}(s_1, \dots, s_n) = \max_{i=1}^n G_i; \quad (10)$$

however, this performs worse than using the sum of token surprisals.

Aggregation	Accuracy
Sum	0.830
Max	0.773

Table 7: Two sentence-level aggregation strategies

B PCA plots of infrequent tokens

We feed a random selection of BNC sentences into RoBERTa and use PCA to visualize the distribution of rare and frequent tokens at different layers (Figure 5). In all cases, we find that infrequent tokens occupy a different region of the embedding space from frequent tokens, similar to what [Gong et al. \(2018\)](#) observed for static word vectors. This is consistent with the correlation between token-level surprisal and frequency (Figure 3), although the decrease in correlation towards upper layers is not apparent in the PCA plots.

C Surprisal gap for BERT and XLNet

Figures 6 and 7 plot the surprisal gaps using the BERT and XLNet models; data and algorithms are identical to the RoBERTa model (Figure 4). The Gaussian model is only sensitive to morphosyntactic anomalies, and not to semantic and common-sense ones.

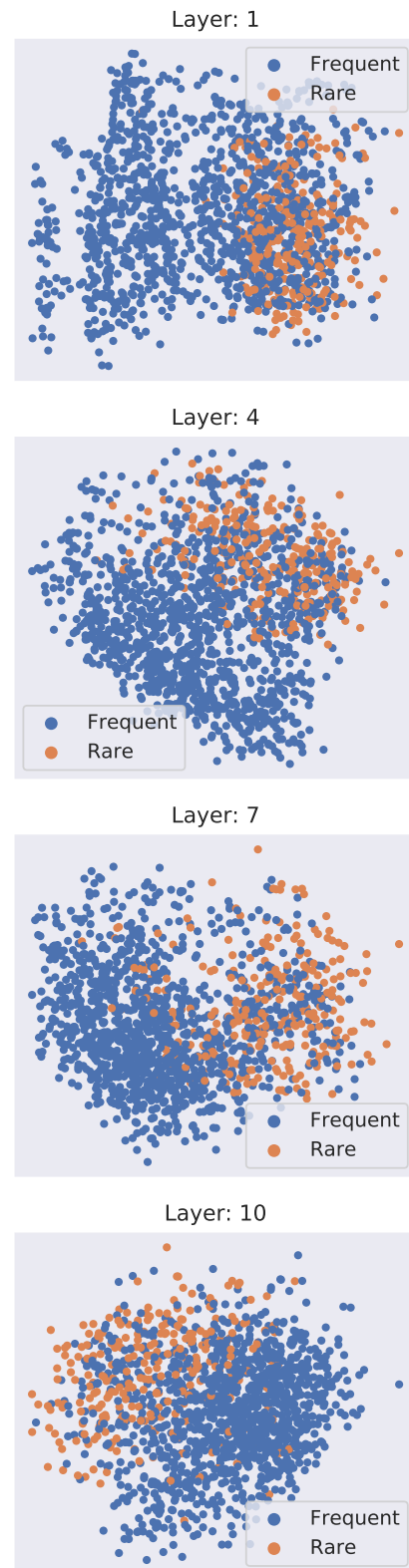


Figure 5: PCA plot of randomly sampled RoBERTa embeddings at layers 1, 4, 7, and 10. Points are colored by token frequency: “Rare” means the 20% least frequent tokens, and “Frequent” is the other 80%.

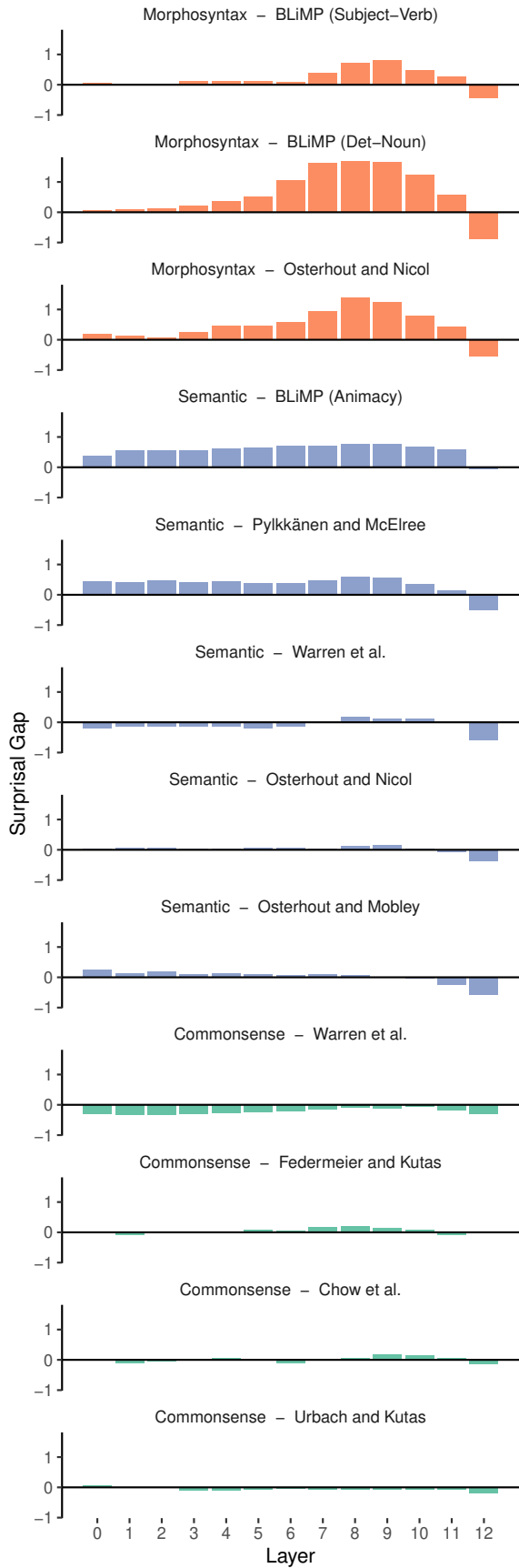


Figure 6: Surprisal gap plot using BERT.

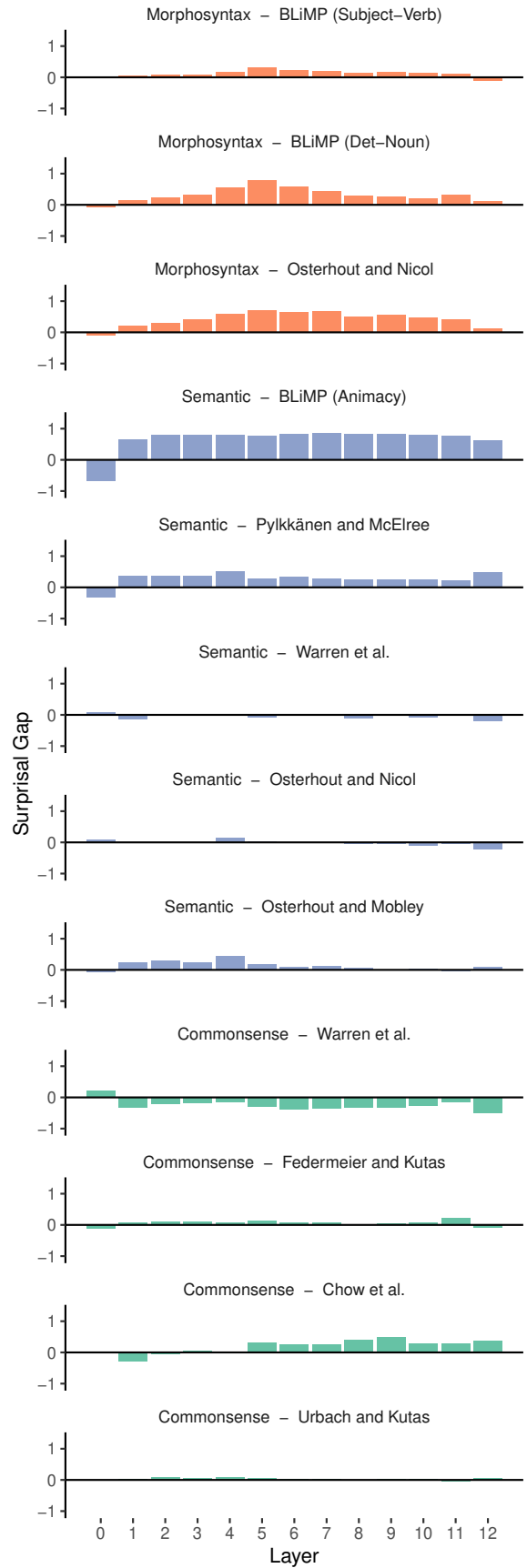


Figure 7: Surprisal gap plot using XLNet.