# The Acceptability Delta Criterion: Memorization is not enough.

by

Héctor Javier Vázquez Martínez

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2021

Author . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Department of Electrical Engineering and Computer Science
January 15, 2021

Certified by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Professor Robert C. Berwick
Professor of Computational Linguistics and Computer Science and Engineering
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

# The Acceptability Delta Criterion: Memorization is not enough.

by

Héctor Javier Vázquez Martínez

Submitted to the Department of Electrical Engineering and Computer Science
on January 15, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

## Abstract

In order to effectively assess Knowledge of Language (KoL) for any statistically-based Language Model (LM), one must develop a test that is first comprehensive in its coverage of linguistic phenomena; second backed by statistically-vetted human judgement data; and third, tests LMs' ability to track human gradient sentence acceptability judgements. Presently, most studies of KoL on LMs have focused on at most two of these three requirements at a time. This thesis takes steps toward a test of KoL that meets all three requirements by proposing the LI-Adger dataset: a comprehensive collection of 519 sentence types spanning the field of generative grammar, accompanied by attested and replicable human acceptability judgements for each of the 4177 sentences in the dataset, and complemented by the Acceptability Delta Criterion (ADC), an evaluation metric that enforces the gradience of acceptability by testing whether LMs can track the human data.

To validate this proposal, this thesis conducts a series of case studies with Bidirectional Encoder Representations from Transformers (Devlin et al. 2018). It first confirms the loss of statistical power caused by treating sentence acceptability as a categorical metric by benchmarking three BERT models fine-tuned using the Corpus of Linguistic Acceptability (CoLA; Warstadt & Bowman, 2019) on the comprehensive LI-Adger dataset. We find that although the BERT models achieve approximately 94% correct classification of the minimal pairs in the dataset, a trigram model trained using the British National Corpus by Sprouse et al. 2018, is able to perform similarly well (75%). Adopting the ADC immediately reveals that neither model is able to track the gradience of acceptability across minimal pairs: both BERT and the trigram model only score approximately 30% of the minimal pairs correctly. Additionally, we demonstrate how the ADC rewards gradience by benchmarking the default BERT model using *pseudo log-likelihood* (PLL) scores, which raises its score to 38% correct prediction of all minimal pairs.

This thesis thus identifies the need for an evaluation metric that tests KoL via gradient acceptability over the course of two case studies with BERT and proposes the ADC in response. We verify the effectiveness of the ADC using the LI-Adger dataset, a representative collection of 4177 sentences forming 2394 unique minimal pairs each backed by replicable and statistically powerful human judgement data. Taken together, this thesis proposes and provides the three necessary requirements for the comprehensive linguistic analysis and test of the Human KoL exhibited LMs that is currently missing in the field of Computational Linguistics.

Thesis Supervisor: Professor Robert C. Berwick
Title: Professor of Computational Linguistics and Computer Science and Engineering

# Acknowledgments

I must first express my gratitude for Annika Heuser, whose editing help and support was instrumental to the timely completion of this thesis. I thank Sagar Indurkhya and Prof. Norbert Hornstein for giving me the pivotal feedback that led me to the main topic of this thesis: The Acceptability Delta Criterion. Additionally I thank Prof. Charles Yang, Prof. Jordan Kodner and Dr. Robert Ajemian for encouraging me to pursue my scientific curiosities, as well as my colleagues and mentors, among them Abbie Bertics, Prof. Barbarah Lust, Beracah Yankama, Shinjini Ghosh, Prof. Suzanne Flynn, and Spencer Caplan for giving me the sandbox where my ideas have grown and developed. I give special thanks for Prof. Robert C. Berwick, who welcomed me to this community, has advised me on my research path, and guided in the development of the project I present here.

And to my family: *¡Árrate abuelo!* Con eso lo digo todo.

# Contents

# List of Figures

# List of Tables

# Introduction

Assessing the Knowledge of Language (KoL) of statistically-based Language Models (LMs) generally involves assuming some fundamental property or computation occurring in the Human Language Faculty and arguing that a currently poorly understood, statistical, and typically connectionist model, also partakes in the use of that property or computation. This quickly becomes a problematic task because understanding the Human Language Faculty has been conventionally posed as a problem to be solved at a causal level removed from the algorithmic and computational implementation levels. Put in more abstract terms, assessing the KoL of a LM requires inferring some abstract operation inside a human black box based on input-output analysis and determining whether a second, statistical black box is somehow also performing the same operation by some other means.

The issue is made even more challenging by changes in either field that consequently change our assumptions surrounding the Human Language Faculty or the black boxes used in Machine Learning (ML). This, in turn, immediately impacts claims relating the two by some abstract property, linguistic or otherwise, that is required for the evaluation of LMs. If any concrete progress is to be made when it pertains to KoL in LMs, then the design of the tests we perform and their conclusions must be based on the same empirical data from current input-output analyses of the Human Language Faculty that has subsequently been used to build the linguistic theories that attempt to characterize and explain Human KoL.

This thesis takes concrete steps toward designing such a test of KoL for LMs by positing the necessary components required to build upon the same bedrock of empirical data as the field of generative grammar in Linguistics. First, we propose the

LI-Adger dataset, a collection of statistically powerful and attested linguistic phenomena representative of the field of Linguistics (Sprouse & Almeida 2012; Sprouse et al. 2013), accompanied by human acceptability judgements in the form of Magnitude Estimation (ME) data. Altogether, the dataset has an attested maximum False Positive (Type 1 error) rate between 1-12% and is well above the 80% threshold for statistical power (<20% False Negatives, or Type 2 errors) (Sprouse & Almeida 2017). The reliability of the LI-Adger dataset is such that, if the linguistic theories were somehow proven to be incorrect and reformulated, it would not be because of the data, but because of incorrect theorizing; any tractable theory of linguistics must account for the empirical phenomena observed in the LI-Adger dataset (Sprouse & Almeida 2012). To complement this data, we propose the Acceptability Delta Criterion (ADC), a proof of concept metric that enforces the gradience of acceptability in its evaluation of model performance, and adopts the continuous human judgements as the ground-truth labels that LMs must approximate in order to demonstrate KoL.

Our results suggest that, when acceptability is treated as a functionally categorical metric on isolated minimal pairs of sentences as it has been traditionally treated in the literature (Linzen et al. 2016; Marvin & Linzen 2018; Wilcox et al. 2018; Warstadt & Bowman 2020; among others), the task of determining sentence acceptability fails to properly test for KoL. Under this relaxed metric, the large, cased version of Bidirectional Encoder Representations from Transformers ($\text{BERT}_{\text{large−cased}}$; Devlin et al. 2018) when fine-tuned using the Corpus of Linguistic Acceptability (CoLA; Warstadt et al. 2019) (the model is henceforth referred to as $\text{BERT}_{\text{CoLA}_{\text{large−cased}}}$), correctly evaluates 2213 out of 2365 (∼94%) minimal pairs in the LI-Adger dataset; that is, for those 2213 minimal pairs, $\text{BERT}_{\text{CoLA}_{\text{large−cased}}}$ gives a higher score to the sentence in the minimal pair deemed by experts to be the *acceptable* one of the pair. We will continue to refer to this metric as the BLiMP Criterion, named after the BLiMP dataset (Warstadt et al. 2020). To put the performance of $\text{BERT}_{\text{CoLA}_{\text{large−cased}}}$) into perspective, a trigram model using the Syntactic Log-Odds Ratio (SLOR; Pauls & Klein 2012; Lau et al. 2017) is able to correctly evaluate 1781 out of 2365 (∼75%) minimal pairs. Considering the coverage of phenomena in the LI-Adger dataset, we may interpret these results in one of two ways: either metrics such as the BLiMP Criterion lead to statistically underpowered tests with a high rate of false positives, or a basic trigram model using SLOR encodes the KoL necessary to account for 75% of the phenomena in Linguistics. We opt for the first interpretation and consider this evidence of a theoretical flaw in the metric itself, not a demonstration of what the models *know* about language.

Adopting the ADC (with $\delta = 0.5$), which enforces that LMs' predictions be within a set number of standard deviation units ($\delta$) from the human ME judgements, quickly changes the panorama. $\text{BERT}_{\text{CoLA}_{\text{large}-\text{cased}}}$ only correctly evaluates 726 out of 2365 ($\sim$31%) minimal pairs, whereas the trigram model with SLOR correctly evaluates 712 out of 2365 ($\sim$30%). These results imply that, when it comes to tracking the acceptability of sentences across minimal pairs, the KoL encoded in BERT does not go much farther than that of an $N$-gram model.

Here we proceed as follows. First, we attempt to replicate the linguistic analysis of BERT conducted by Warstadt & Bowman (2020) using the grammatically annotated Corpus of Linguistic Acceptability. Over the course of this replication, we confirm evidence of underspecification in overparametrized Neural LMs as identified by D'Amour et al. (2020); McCoy et al. (2019), among others. In particular, we observe predictions on the LI-Adger sentences from $\text{BERT}_{\text{CoLA}_{\text{base}-\text{uncased}}}$, the smallest (i.e. least overparametrized) of the original $\text{BERT}_{\text{CoLA}}$ models, is extremely sensitive to the *order* in which the CoLA training sentences are presented, even though overall performance remained relatively unchanged. We observe this behavior even within the same initialization of the model, where the only difference between two runs is the random seed used to shuffle the training data. This underspecification takes the form of instability in the LI-Adger test set predictions: sentences predicted as acceptable (1) by BERT with around 90-99% *confidence* flip to be predicted as unacceptable with a similar magnitude, or vice versa. We find that over the course of 200 different training orders, 1272 sentences, or roughly 30% of the sentences in the LI-Adger dataset exhibit this flipping behavior. We affectionately name this subset of sentences the Acrobatic Sentences.

Given the alarmingly high proportion of acrobatic sentences produced by the predictions from $\text{BERT}_{\text{CoLA}_{\text{base-uncased}}}$, we find ourselves obliged to consider successful replication as achieving Matthew's Correlation Coefficient (MCC) scores on the CoLA test set that are *reasonably* close to those reported by Warstadt et al. (2020). To this end, we select the $\text{BERT}_{\text{CoLA}}$ models with the single best performance on the CoLA out-of-domain test set and further test them using the LI-Adger dataset under the BLiMP Criterion. Although we find that the $\text{BERT}_{\text{CoLA}}$ models satisfy the BLiMP criterion for roughly 94% of the minimal pairs, the magnitudes of their predictions do not track the degrees of acceptability exhibited by the gradient human judgements.

When benchmarking $\text{BERT}_{\text{CoLA}}$ models using the BLiMP criterion, the output of the models is determined by multiplying the final hidden vector ($\vec{h} \in \mathbf{R}^d$) by a weight matrix $W \in \mathbf{R}^{2\text{x}d}$ learned during the fine-tuning phase and taking the softmax of the product, written explicitly in Equation 1.

$$\text{softmax}(\vec{x}) = \frac{e^{x_i}}{\sum_{j=0}^{j=K} e^{x_j}} \tag{1}$$

The final output of the $\text{BERT}_{\text{CoLA}}$ models (*out*) is then computed by taking argmax of the two-dimensional vector resulting from the softmax, as expressed in Equation 2, thus yielding the final categorical 1/0 prediction.

$$out = \text{argmax}\left[\text{softmax}(Wh)\right] \tag{2}$$

However, in order to improve the $\text{BERT}_{\text{CoLA}}$ models' performance apriori before applying the ADC, we adopt the labels ±1 instead of 1/0 and scale the predicted labels by the output of the final softmax classification head in Equation 1. Now we have an approximate method of knowing when the $\text{BERT}_{\text{CoLA}}$ models consider a sentence completely acceptable ($\sim$0.95), completely unacceptable ($\sim$ -0.95), and anything in between. We confirm we do not lose any information because the categorical labels are recovered by taking the sign of the new output. The delta in the $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ model's acceptability scores across minimal pairs using this more gradient output metric only weakly correlates with the human judgements (∼0.349, $p$<0.0001). For reference, conducting the same analysis with the SLOR scores of a trigram model trained on the British National Corpus (Sprouse et al. 2018) yields almost the same Pearson's correlation coefficient (∼0.333, $p$<0.0001).

The above analyses by nature warrant further controls. For one, we are uncertain of what information–semantic, syntactic or otherwise–might be introduced into the BERT models by the CoLA training set itself, as opposed to already being present in their pretrained representations. Additionally, training a linear classifier on top of the BERT models' embeddings very rarely yields a softmax output of less than 0.95, meaning most predictions were around either 0.99 or -0.99. In spite of BERT's claimed KoL, expecting gradience from the resulting $\text{BERT}_{\text{CoLA}}$ after fine-tuning using the categorical labels could be viewed as unfair to the model due to its lack of access to gradient data. We believe this is a fair expectation because we do not have access to categorically labeled raw linguistic input during language acquisition, and we are ultimately probing the LMs for Human KoL. Regardless, we repeat the analyses on the out-of-the-box version of BERT ($\text{BERT}_{\text{MLM}}$). We obtain *pseudo-log-likelihood* (PLL) scores from $\text{BERT}_{\text{MLM}}$ by performing a variant of a Cloze test in which we sequentially mask each word in a given sentence and retrieve the probability of the originally masked token as predicted by $\text{BERT}_{\text{MLM}}$ (Wang & Cho 2019; Shin et al. 2019; Salazar et al. 2019 ).[1] In this task, the total PLL of a sentence $s_i$ of length $n$ is the sum total of the log-likelihood score of each of its tokens $[w_0, ..., w_n]$, which can be expressed as:

$$\text{PPL}(\text{s}_\text{i}) = \sum_{j=0}^{n} \log(P(w_j | w_0, ..., w_{j-1}, w_{j+1}, ...w_n)) \tag{3}$$

We find that this objective only slightly improves performance under the ADC: scoring 890 out of 2365 (∼38%) minimal pairs with $\delta = 0.5$, as well as slightly improv-

[1] We are fully aware that Jacob Devlin himself has said that BERT is not a language model and recommended against this sequential masked language modeling procedure (See the original issue on the Google Research GitHub repository). We point readers to Salazar et al. (2019), who report $\text{BERT}_{\text{MLM}}$ beats the state-of-the-art GPT-2 on BLiMP (Warstadt et al. 2020) when using PLL scores.

ing the correlation with the human judgement deltas across minimal pairs (∼0.384, $p$<0.0001).

Given the results of these analyses, the contributions of this thesis are threefold. First, it highlights the importance of interpreting sentence acceptability as a gradient metric and demonstrates how exhibiting such gradience is a prerequisite to attributing any KoL to a LM. Secondly, it proposes the Acceptability Delta Criterion as a proof of concept measurement that enforces the gradience of acceptability in its evaluation of performance and adopts continuous human judgements as the ground-truth labels that LMs are expected to approximate. Finally, it presents the LI-Adger dataset of over 4000 sentences each associated to a human ME result, and approximately 2400 unique minimal pairs, each supported by an Acceptability Delta value. Because the sentences in the LI-Adger dataset have a fairly wide and representative coverage of the field of linguistics, and because the human data presented here is statistically powerful, reliable and has been replicated on multiple occasions, researchers will hopefully adopt this data as the bedrock analysis test set of LMs against which any and all claims about their KoL can be put to the test.

# Chapter 1

# BLiMPs were meant to fly!

## 1.1 Assessing Knowledge of Language (KoL)

The success of Neural Language Models at different natural language tasks, such as Next Sentence Prediction (NSP), Machine Translation (MT) and Question Answering (QA), among others[1], has made it a popular endeavor to assess the potential Knowledge of Language encoded in the learned representations of the language models and how that KoL may be contributing to their performance. If one were to roughly summarize these efforts, one could group these types of analyses into two broad methodological categories: those that treat the language model as a black box and draw conclusions about the system based on thorough input-output analysis, and those that train additional classifiers (*probes*) to use the representations inside the black box in order to accomplish some linguistically meaningful task (Conneau et al. 2018; Elazar et al. 2020).

The probing approach requires an additional training corpus labeled with the linguistic concepts of interest in order to train and evaluate the probing classifier before drawing any conclusions. However, because probing relies on training an additional classifier on top of the latent (in other words: opaque and currently poorly understood) representations of neural LMs, it is extremely difficult to control for confound-

[1] For a quick collection of more natural language tasks and how different models perform on them, see the GLUE Leaderboard or the Super GLUE Leaderboard.

ing variables, such as the information being introduced into the system by training the probing classifier in the first place (Warstadt et al. 2020). Additionally, D'Amour et al. (2020) have found substantial evidence indicating that these overparametrized neural LMs by nature exploit different sets of spurious correlations according to their random initialization in spite of exhibiting very similar performance on I.I.D. test sets. This poses a unique set of difficulties for the use of probes for any assessment of KoL in such LMs.

To compound the matter, Human KoL, due to its abstract, deliberately acomputational nature, can *only* be assessed via proxies, generally by probing language acquisition or use. At present, the studies of LMs' KoL that rely on an input-output analysis of a system tend to focus on probing their weak generative capacity: testing whether a given LM can discern whether a particular sequence of words is or is not in the set of sentences generated by some presumed corresponding grammar, typically by comparing the probabilities the LM assigns to different but related sequences of words.

We believe for these reasons that in order to effectively assess KoL for any statistically-based LM, one must develop a test that requires both KoL in the form of a grammar for a language and a mapping that describes the *use* of that grammar. We take steps to this end by presenting the LI-Adger dataset, a collection of roughly 4200 sentences, each backed by human Magnitude Estimation (ME) data assigning a gradient acceptability value to each sentence. This thesis also posits the Acceptability Delta Criterion (ADC) as a measure that enforces the gradience of acceptability when evaluating LMs, and it empirically shows how it is a step above the weak generative capacity tested by evaluating set membership.

## 1.2 BLiMP: The Benchmark of Linguistic Minimal Pairs

Warstadt et al. (2020) have taken seminal steps toward evaluating LMs beyond their weak generative capacity by positing the Benchmark of Linguistic Minimal Pairs for English (BLiMP). They automatically generated 67 datasets of 1000 minimal pairs each from grammar templates that span 12 linguistic phenomena. They designed the templates to contrast in grammatical acceptability by isolating specific phenomena in syntax, morphology or semantics. In doing so, the authors intend to mirror what a working linguist uses to probe KoL in native speakers of a language. Because such principles generally appeal to grammatical constraints, they go beyond simple weak generative capacity.

Although the concept of using minimal pairs is not new (Linzen et al. 2016; Marvin & Linzen 2018; Wilcox et al. 2018; to name a few), the creators of BLiMP take the idea to a much larger scale and propose a single metric for evaluation, which we will call the BLiMP Criterion. For a given minimal pair $m_i$ consisting of an acceptable sentence $s_{i,1}$ and an unacceptable sentence $s_{i,2}$, if a LM evalutes $P(s_{i,1}) > P(s_{i,2})$, then the LM has met the BLiMP Criterion for $m_i$. The authors of BLiMP thus score a LM on the BLiMP Benchmark according to the percentage of all the minimal pairs for which it was able to fulfill the BLiMP Criterion. This, of course, can be broken down into further analyses of the 12 linguistic phenomena they sought to represent in the dataset.

### 1.2.1 Not everything on the BLiMP flies.

The BLiMP Criterion is met whenever the acceptable sentence of a minimal pair receives a higher score or probability than its unacceptable counterpart. This setup has the unfortunate consequence of treating sentence acceptability, a metric well known to be gradient by nature (Sprouse et al. 2018), as functionally categorical. Under the BLiMP criterion, a sentence is either more acceptable or less acceptable than its counterpart, greatly simplifying the task of assigning acceptability judgements for LMs. This point is underscored by the high performance of the baseline 5-gram model in Warstadt et al. (2020), scoring 61.2% of the 67,000 minimal pairs correctly under the BLiMP Criterion. This has the immediate implication that an $N$-gram model, well understood to have little to no Knowledge of Language, suddenly *knows* approximately 60% of all the phenomena tested in BLiMP.

On the subject of the phenomena tested in BLiMP lies the question of whether the authors' selection of phenomena is representative of syntax or linguistics. They very correctly point out that, because they designed the grammatical templates with an emphasis on controlling for sentence length and lexical parity, their coverage of linguistic phenomena is fundamentally limited. There is therefore no concrete notion to what achieving 60% correctness on BLiMP means, as in the case of the 5-gram model, because the KoL being tested is only the subset that can be reasonably generated using a templated approach.

Templating also brings with it another problem: although the authors validate their data using human judgement data using a Forced Choice (FC) task, automatic generation leads to semantically implausible sentences. The authors argue that this semantic implausibility should not influence human subjects' judgements because all semantically implausible sentences are pairs of an acceptable and unacceptable sentence that differ along a single feature, which should control for that confound. However, Sprouse et al. (2018) conducted a similar exercise by setting up an acceptability FC experiment wtih Chomsky's canonical *Colorless green ideas sleep furiously* sentence. They first obtained all 120 possible permutations of the 5 word sequence (henceforth the CGI dataset) and proceeded to generate all possible 7140 unique pairs of sentences from the 120 CGI sentences. Sprouse et al. (2018) ranked each CGI sentence according to the Elo chess rating system by treating each FC trial as a chess match. They found that, although the canonical sentence is perfectly well-formed, three other sentences were rated as more acceptable, shown along with their acceptability (Elo) ratings in Table 1.1 below.

Even though in theory the canonical (cgi.0) sentence should have received the highest acceptability rating out of all its other 119 permutations, it was bested by three of its kin. Although it may be argued (with some squinting) that cgi.24 and cgi.25 are also perfectly well-formed, the case is much harder to make for cgi.11. Hence semantic implausibility is a very strong confounding factor when eliciting human acceptability judgements even in FC, casting doubt on the reliability of the native

| **Sentence ID** | **Sentence** | **Elo score** |
|---|---|---|
| cgi.0 | colorless green ideas sleep furiously | 179.9308187 |
| cgi.11 | colorless ideas furiously sleep green | 180.7792248 |
| cgi.24 | green colorless ideas sleep furiously | 220.4766292 |
| cgi.25 | green colorless ideas furiously sleep | 187.6557574 |

Table 1.1: Four sentences from the Colorless Green Ideas (CGI) dataset collected by Sprouse et al. (2018). All 120 permutations of the canonical sentence were paired with each other for a total of 7140 unique pairs. Each FC trial was treated as a chess match, and then each sentence was given an Elo chess rating according to the number of *matches* it won.

speakers' judgements for this class of minimal pairs.

Lastly is the fact that the human judgements of the BLiMP data were collected using a FC task in which the human participants were asked to select the more acceptable sentence of the two in each minimal pair. Although the FC task is statistically more powerful than the Likert Scale (LS) and Magnitude Estimation (ME) tasks at detecting differences in acceptability, it is ill-suited for quantitative experiments of this nature. FC tasks only *detect a difference in acceptability*, but do not allow direct comparison of the magnitude of the change in acceptability (Schütze & Sprouse 2013). Hence, very valuable information is lost: a very large difference in the acceptability of two sentences in a minimal pair merits a different explanation than that of a small difference in acceptability.

By computing human performance on BLiMP based on the number of minimal pairs where the more acceptable sentence of the pair was preferred via the FC task, the authors adopt the paradigm of relying on expert labels as the ground truth in evaluation. However, the minimal pairs where the human judgements were considered to be incorrect, i.e, where the unacceptable sentence was preferred over the acceptable one under FC, can also be interpreted to mean that, for those minimal pairs in particular, the FC task could not detect an appreciable difference in acceptability. After all, the linguistic theory the phenomena come from is fundamentally derived from human data, thus it stands to reason that one may adopt the human judgements as the true labels in the paradigm, not the expert-assigned categorical labels.

## 1.3 Taking the BLiMP to new heights

In this section, we explain the three major contributions of this thesis. This thesis first presents the LI-Adger dataset collected by Sprouse & Almeida (2012) and Sprouse et al. (2013) of well over 4000 sentences each associated to a human ME result, thereby yielding approximately 2400 minimal pairs with a representative coverage of the field of generative syntax. To effectively use this dataset, this thesis highlights the importance of interpreting sentence acceptability as a gradient metric and demonstrate how exhibiting such gradience is a prerequisite to attributing any KoL to a LM. Lastly, this thesis proposes the Acceptability Delta Criterion (ADC) as a proof of concept measurement that begins to enforce the gradience of acceptability in its evaluation of model performance and adopts continuous human judgements as the ground-truth labels that LMs are expected to approximate.

### 1.3.1 The LI-Adger dataset

The LI-Adger dataset is a collection of two separate datasets. The first consists of a randomly selected sample of 150 pairwise phenomena (300 sentence types) from Linguistic Inquiry (LI) 2001-2010 collected by Sprouse et al. (2013). Each pairwise phenomena includes 8 hand-constructed, lexically matched minimal pairs such that most of the contribution of lexical information to the acceptability of the sentences would be distributed equally to the pair. For the purposes of complete transparency: 144 out of the 150 pairwise phenomena consisted of 8 lexically matched pairs of sentences. The remaining 6 phenomena consisted of 7 lexically matched pairs and one non-matched pair, because the originally published pair in LI was not lexically matched.

The second set of sentences is an exhaustive selection of 219 sentence types from Adger's (2003) *Core Syntax* textbook (198 directly from the textbook + 21 created as additional controls) that form 105 multi-condition phenomena collected by Sprouse &

Almeida (2012). Much like the LI dataset, 8 tokens of each sentence type were created by hand such that the structural properties of the condition were maintained but the lexical items varied. One thing to note is that many of these sentences often have interesting names from Greek mythology in the textbook, but these were changed to common names in order to keep the proper names from biasing the native speakers' judgements of the sentence. For the purposes of the LI-Adger dataset as a whole, we have split each multi-condition phenomenon into minimal pairs by taking each possible combination of acceptable and unacceptable sentences in the condition as a valid minimal pair. For example, the multi-condition phenomenon from Chapter 8 (*Functional Categories III*) of the textbook presented in Table 1.2 below would yield the two minimal pairs presented in Table 1.3:

| **Sentence ID** | **Sentence** |
|---|---|
| ch8.150.*.01 | Melissa seems that is happy. |
| ch8.151.g.01 | It seems that Melissa is happy. |
| ch8.152.g.01 | Melissa seems to be happy. |

Table 1.2: Example multi-condition phenomenon from the Adger dataset. Note: the original sentences in the Adger textbook use the name Agamemnon, but was changed to Melissa in order to avoid any potential influence of the unfamiliar name in native speakers' judgements.

| **Acceptable sentence** | **Unacceptable sentence** |
|---|---|
| It seems that Melissa is happy. | Melissa seems that is happy. |
| Melissa seems to be happy. | Melissa seems that is happy. |

Table 1.3: Two minimal pairs constructed from a single multi-condition phenomenon from the Adger dataset. Note: the original sentences in the Adger textbook use the name Agamemnon, but was changed to Melissa in order to avoid any potential influence of the unfamiliar name in native speakers' judgements.

The Adger dataset, in virtue of being sampled from the *Core Syntax* textbook, which constructs a theory of syntax from the ground up on the basis of examples, can be taken to have reasonably good coverage of the field of syntax. Add to this coverage the LI dataset, which is sampled from the 111/114 articles published in

Linguistic Inquiry about US English syntax from 2001-2010 (out of the total 308 articles published during that time). Therefore, to the extent that the Adger *Core Syntax* texbook and *LI* 2001-2010 are representative of the data in the field, so is the LI-Adger dataset. (Sprouse & Almeida 2012; Sprouse et al. 2013).

### 1.3.2 Human Magnitude Estimation (ME) data

Perhaps even more importantly than the coverage of linguistic phenomena represented in the LI-Adger dataset is the human judgement data that comes with it. Sprouse & Almeida (2012) collected Magnitude Estimation and Yes-No judgement data from a total of 440 native participants for the 469 data points they sampled from the Adger *Core Syntax* textbook. After conducting three different statistical analyses on the data (traditional null hypothesis significance tests, linear mixed-effects models, and Bayes factor analyses), they found that the maximum replication failure rate between formal and informal judgements (i.e. formal vs. informal data collection methods) was 2 percent (Sprouse & Almeida 2012; Schütze & Sprouse 2013).

Sprouse et al. (2013) took those analyses even further with their sample of 148 two-sentence phenomena from *LI* 2001-2010. They collected data for the LI sentences using the 7-point Likert Scale (LS) task, ME and FC and vetted it under 5 different statistical analyses (the same three as Sprouse & Almeida (2012) plus Descriptive directionality and two-tailed null hypothesis tests). They estimated a minimum replication rate for journal data of 95 percent ±5 (Sprouse et al. 2013; Schütze & Sprouse 2013.

Finally, Sprouse & Almeida (2017) sampled 50 pairwise phenomena from LI dataset in a complementary study that determined the statistical power of formal linguistics experiments by task and average effect size and recommend setting the threshold for well-powered experiments at 80% statistical power. They find that the FC task would reach the 80% power threshold and detect 70% of the phenomena published in *LI* 2001-2010 with just ten participants, assuming each provides only one judgement per phenomenon. With fifteen participants, FC would detect 80% of the phenomena. Because the ME task has less statistical power than FC, it requires at least thirty to thirty-five participants to reach the same 80% coverage of *LI* 2001-2010 as FC (Sprouse & Almeida 2017; Schütze & Sprouse 2013. Because 20 is the sample size of the human FC data in BLiMP, and the sample sizes for the LI-Adger datsets are much larger (104 participants per condition for the LI sentences and 40 for the Adger sentences), we do not forfeit any statistical power by using ME data in spite of the higher statistical power of the FC task. On the contrary, the ME task will allow us not only to perform the same type of functionally categorical acceptability comparison as the BLiMP Criterion, but also allow us to make comparisons between every condition in the dataset.

Taken together, the LI-Adger dataset is a representative collection of linguistic phenomena that have been validated multiple times over by human judgement data across ME, FC, LS and Yes-No tasks. The human ME data we include as part of the LI-Adger dataset is therefore reliable, replicable and statistically powerful. The LI-Adger dataset has the added benefit of being theory-agnostic; if linguistic theories were to fundamentally change in the future, the significance and validity of the data would remain unchanged.

### 1.3.3 The Acceptability Delta Criterion (ADC)

Thanks to the ME data associated with each sentence in the LI-Adger dataset, we can now make direct acceptability comparisons, not just between the two sentences of a minimal pair, but also across minimal pairs and even across phenomena. It is crucial to be able to make such direct comparisons due to the gradient nature of acceptability. Acceptability judgement experiments carry as a necessary underlying assumption that acceptability is a *percept* that arises in response to linguistic stimuli. Collecting data about the percept requires then that the subject report that perception of acceptability (Chomsky 1965; T Schütze 2016; Sprouse & Almeida 2013; Schütze & Sprouse 2013). Consequently, acceptability judgements are a behavioral response that may vary in intensity, much like brightness, loudness, temperature, pain, etc. The degree of this response is inherently informative, in particular because acceptability is the behavioral output of the grammatical system, to which neither speakers nor linguists have direct access.

In order to illustrate the informativeness of adopting gradient acceptability judgements and of being able to make direct comparisons across minimal pairs with the ME data, take as an example the following two minimal pairs:

| **Sentence ID** | **Sentence** | **ME zscore** |
|---|---|---|
| 32.3.Culicover.7a.g.01 | John tried to win. | 1.453262 |
| 32.3.Culicover.7b.*.01 | John tried himself to win. | -0.86729 |
| 33.2.bowers.7b.g.07 | Sarah counted the change accurately. | 1.230412 |
| 33.2.bowers.7b.*.07 | Sarah accurately counted the change. | 1.20698 |

Table 1.4: Two minimal pairs for the Linguistic Inquiry (LI) dataset collected by Sprouse & Almeida, 2012. The ME zscore is the averaged zscore transformation of the Magnitude Estimation results across 104 different experimental participants.

It is clear that the difference in acceptability across the Culicover minimal pair is vastly different from the difference across the Bowers minimal pair in Table 1.4. In fact, the average ME rating for the expert-labeled unacceptable Bowers sentence (33.2.bowers.7b.*.07) is much higher than many other sentences in the data that are expert-labeled as *acceptable*, meaning the 104 participants that were asked to rate this sentence found it *statistically* completely acceptable. This type of information is absolutely crucial when evaluating whether a LM has knowledge of any particular linguistic phenomenon, yet this information is lost when analysing performance according to the BLiMP criterion.

To this end, we propose the Acceptability Delta Criterion (ADC). It is founded on the principle that, if we are to ascribe any inferred knowledge of one black box (the Human Language Faculty) to another black box (Neural Language Models) based solely on an input-output analysis of both systems, then the response of both systems must agree both categorically and in magnitude. In other words, for a minimal pair such as the Culicover pair in Table 1.4 whose change in human acceptability rating is nearly night and day, a language model with comparable KoL will output a similarly drastic change in acceptability rating across the same minimal pair.

To make this more concrete: Suppose we have a language model $L$ with output function $f$ that takes in a sequence of words $x_i$ and outputs a score $y_i$. The first step in the ADC is to understand the range of values output by the language model $L$ over the 4179 LI-Adger sentences: $Y = [y_1, y_2, ..., y_{4179}]$. With the full range of value, we apply a z-score transformation to each of the values in $Y$ by subtracting the mean of $Y$ from each of the values and then dividing them by the standard deviation of $Y$. This will yield the set z-score transformed predictions $Z = [z_1, z_2, ..., z_{4179}]$. Notice that because this is a purely linear transformation, it preserves the relationships between the data points. In addition, the resulting set of predictions $Z$ represents a standardized form of $Y$, where each prediction $z_i$ is expressed in standard deviation units of $y_i$ from the mean of $Y$ (Schütze & Sprouse 2013).

One may argue that even though the human ME data and the scores output by the LM, because the scales are by nature fundamentally different, cannot be compared even when expressed in standard deviation units. Let us assume for a moment that what we obtain from the LM is a probability distribution over the sequence of words (as per the canonical definition of a LM). That means that whatever is output by the LM is bounded in the range [0, 1], yet we typically work with log probabilities in this context, so the range of possible values becomes $(-\infty, 0]$ assuming there is some smoothing in place such that we do not attempt to calculate the logarithm of 0. Strictly speaking, the range of log probabilities is upper- and lower-bounded, but in practice it is mostly upper-bounded. Turning to ME data, the participant is asked to use a reference sentence as a unit of measurement to estimate how acceptable the target sentence is. For example, given a reference sentence $a$ and a target sentence $b$, the participant must give an estimate of how acceptable $b$ is by using $a$ as a unit of measurement. I.e. *b is four times more acceptable than a*, or *b is half as acceptable as a*. This means that the scale is theoretically lower-bounded by 0 (which could be argued to be absolute unacceptability), but open-ended and infinite on the upper range of the scale. In practice, participants seem to use the ME task as a Likert Scale with more response options. Both original units of measurement then (ME and log probabilities) are scales bounded on one end and open on the other end. Converting both to standard deviation units converts them to an unbounded scale, which Schütze and Sprouse argue not to be an issue even for LS measurements, which are both discrete and bounded at both ends of the scale (Sprouse 2011; Schütze & Sprouse 2013).

Now that we have grounds for making the comparison and a value for how acceptable the model $L$ finds a sequence of words $x_i$ in terms of standard deviation units $z_i$, we can begin to compare the degree of this acceptability response to the human judgement data, also expressed in standard deviation units. For a given minimal pair $m_i$ consisting of an acceptable sentence $s_{i,1}$ and an unacceptable sentence $s_{i,2}$, we will have 4 pieces of information: two human Z-score transformed acceptability judgements $h_{i,1}$ and $h_{i,2}$, and two language model scores $z_{i,1}$ and $z_{i,2}$. We turn these into two concrete points of comparison: a human acceptability delta $\Delta h_i = h_{i,1} - h_{i,2}$ and a language model acceptability delta $\Delta lm_i = z_{i,1} - z_{i,2}$. In this new formulation, no information has been lost. Recall that the BLiMP Criterion is met for the minimal pair $m_i$ when the language model scores the acceptable sentence higher than the unacceptable one, i.e. $\Delta lm_i > 0$.

With the fully defined delta values as well as a reformulated BLiMP Criterion in terms of the delta values, we may finally proceed to define the ADC. Let $\delta$ be a scalar value indicating the number of maximum allowed units of deviation between the human judgement delta $\Delta h_i$ and the language model delta $\Delta lm_i$. Using this $\delta$ value, we consider the ADC to be met for the minimal pair $m_i$ when the following two conditions are met:

$$\operatorname{sign}(\Delta h_i) = \operatorname{sign}(\Delta lm_i) \tag{1.1}$$

$$|\Delta h_i - \Delta lm_i| < \delta \tag{1.2}$$

The $\delta$ parameter in Equation 1.2 can be adjusted to allow for larger or smaller amounts of deviation between the human and LM acceptability deltas. If $\delta$ is set to a large number, the ADC functionally becomes the BLiMP Criterion because it is dominated by Equation 1.1. The main difference would be that, instead of comparing the expert labels to the LM's output, the human judgements would become the ground truth. For example, if $\delta$ is set to a very large number, and the human ME data find the expert-labeled *unacceptable* sentence as more acceptable than the expert-labeled *acceptable* counterpart, then the LM is expected to follow the same monotonicity.

As an example of the ADC in action, consider the minimal pairs from Table 1.4, expressed in Table 1.5 in terms of the Sentence ID of the grammatical sentence. We show the acceptability delta values for the log probabilities of a simple trigram model trained on the British National Corpus (Sprouse et al. 2018), as well as the human acceptability deltas. We also include two columns indicating whether the BLiMP Criterion (BC) or Acceptability Delta Criterion (ADC) was met.

| **Sentence(g) ID** | $\Delta h_i$ | $\Delta lm_i$ | **BC met?** | **ADC met?** ($\delta = 1$) |
|---|---|---|---|---|
| 32.3.Culicover.7a.g.01 | 2.320552 | 0.633896671 | Yes | No |
| 33.2.bowers.7b.g.07 | 0.023432 | -0.158799029 | No | No |

Table 1.5: The two minimal pairs from Table 1.4 with acceptability delta values from the human judgements and log probability scores from a trigram trained by Sprouse et al. (2018) on the British National Corpus (BNC). The last two columns show whether the BLiMP Criterion (BC) or the Acceptability Delta Criterion (ADC) was met.

Although we maintain the ADC is posited here as a proof of concept, we hope that its simplicity appeals to the intuition that a LM's acceptability judgements must track those of native speakers both in absolute terms (categorically) and in magnitude of the response if any KoL is to be claimed. For this reason, this thesis withholds from determining a final value of $\delta$, as it is both the subject of ongoing work and will likely be the topic of debate. Instead, this thesis adopts a first and second approximation of $\delta = 0.5$ and $\delta = 1$ for the case studies used to study the results of the Acceptability Delta Criterion.

# Chapter 2

# Fine-tuning BERT for Acceptability Judgements

Although we have expressed our qualms regarding the probing approach to assessing the KoL of Neural LMs in Section 1.1, we believe it important to explore this avenue nonetheless, even if to a limited extent. Pre-trained Transformer-based models encode a large body of general knowledge, but are poorly optimized for specific natural language tasks out of the box. Therefore, in order to get optimal downstream task performance, it may be advantageous to fine-tune the pre-trained Transformer model on a downstream task with domain-specific data (Radford et al. 2018; Devlin et al. 2018).

Our task of interest is obtaining acceptability judgements over entire sequences of words from BERT in order to compare them across the LI-Adger minimal pairs under two different criteria. We do not want to discount the possibility that BERT's performance may improve on either criterion by fine-tuning the model for sequence classification specifically. We formulate the sequence (acceptability) classification task and training as follows.

When given a sequence of words $s_i$, BERT's final hidden layer will produce an encoded sequence output $h_i$. Accordingly, we proceed to train a linear output layer that maps via a learned weight matrix $W$ the encoded sequence output $h_i$ to a particular label $c_j$ where $j \in \{1, 0\}$. The probability of label $c_j$ can then be expressed in terms of the softmax function (Equation 1), written formally as Equation 2.1 (Sun et al. 2019).

$$P(c_j|h) = \text{softmax}(Wh_i) \tag{2.1}$$

We will loosely interpret the output of the softmax in the final layer as the model's *confidence* in a particular label, or how acceptable or unacceptable BERT finds a particular sequence of words to be. To clarify, although Sun et al. (2019) define the output of the softmax in Equation 1 as a true probability, hence why we refer to it as *confidence*, Guo et al. (2017) among others have found that in order for the softmax output of a neural network to be considered a true probability or confidence, it must be calibrated to the true correctness likelihood via other post-processing methods currently unavailable to us. For example, there is currently no complete theory of the gradient nature of acceptability that can produce the gradient acceptability score for a given sentence on demand (Sprouse & Almeida 2012). However, without confidence calibration, Guo et al. (2017) find the softmax output of modern neural networks often overestimates the true underlying probabilities. Conversely, when BERT is used to predict the probability of a token (Masked Language Modeling - MLM), a similar softmax operation is performed to yield what is considered the true probability of the target token. We do not adopt a particular stance on the matter and will simply use the italicized term *confidence* to refer to the softmax output as formulated in Equation 2.1 by Sun et al. (2019).

## 2.1 BERT pilots the BLiMP

Over the course of this chapter and the remainder of this thesis, we will be working with Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. (2018)). We determined BERT to be the ideal model to test due to the growing body of research attributing ever greater KoL to BERT. Warstadt & Bowman (2019) have already shown high Matthews Correlation Coefficient (MCC; Matthews 1975) scores between the expert acceptability labels for the sentences in the Corpus of Linguistic Acceptability (CoLA; Warstadt et al. 2019) and $\text{BERT}_{\text{CoLA}}$ models' predictions.

These researchers have gone on to show with a grammatically annotated CoLA analysis set that $\text{BERT}_{\text{CoLA}}$ models exhibit very strong positive MCC scores on multiple syntactic features. For example, they claim BERT exhibits strong knowledge of complex or noncanonical argument structures such as ditransitives and passives, and has a distinct advantage over baseline performance on sentences with long-distance dependencies such as questions. Additionally, Manning et al. (2020) have approximated sentence tree structures by linearly transforming BERT's learned representations into a metric that captures parse tree distances. Finally, Salazar et al. (2019) used the raw psuedo-log-likelihood (PLL; Wang & Cho 2019; Shin et al. 2019; Salazar et al. 2019) from the out-of-the-box $\text{BERT}_{\text{MLM}_{\text{large}-\text{cased}}}$ to evaluate its KoL using the BLiMP benchmark and found it to correctly predict 84.8% of the minimal pairs in BLiMP, thereby beating GPT-2 by 4.2% and almost reaching the human baseline at 88.6%. As we will demonstrate below, we do not align ourselves with many of the claims we have reviewed here regarding the KoL encoded in BERT. Nonetheless, we believe it important to provide background for the *claims* that have recently been made in the field.[1] We will take the information provided here as the baseline level of performance we will expect from BERT moving forward: in other words, we do not believe it unfair given these results to *a priori* expect BERT to exhibit the same or a similar level of gradience in acceptability judgements across minimal pairs to that of humans.

## 2.2 BERT drinks the CoLA

In order to provide BERT with the best possible chance of achieving maximum performance in our proposed test of KoL using the LI-Adger dataset and the ADC, we begin our analyses of BERT by first replicating the results observed by Warstadt & Bowman (2019)(henceforth W&B 2019). This replication serves two purposes: to ensure that our training regime did not inadvertently cripple BERT in any meaningful way, and to have an objective data point of how performance on phenomena attributed

[1] For a recent review of the knowledge of language that has been attributed to BERT, see *A Primer in BERTology: What We Know About How BERT Works*, (Rogers et al. 2021)

to BERT such as noncanonical argument structures, as W&B argued, translates to performance on the ADC.

We use the Huggingface Transformers library (Wolf et al. 2020) to fine-tune three pre-trained versions of BERT in order to be comprehensive in our coverage: 10 random seeds of $\text{BERT}_{\text{CoLA}_{\text{base}-\text{uncased}}}$, 20 random seeds of $\text{BERT}_{\text{CoLA}_{\text{large}-\text{uncased}}}$, and 20 random seeds of $\text{BERT}_{\text{CoLA}_{\text{large}-\text{cased}}}$. Here we note a slight divergence from the authors in methodology. W&B 2019 noted they trained 20 random restarts of $\text{BERT}_{\text{CoLA}_{\text{large}}}$ (we suspect the cased version) and discarded 5 out of the 20 restarts because they were degenerate, i.e. those restarts yielded an MCC of zero on the CoLA test set. Instead of training a fixed number of seeds and then discarding the degenerate ones, we continued training seeds until we reached 20 nondegenerate random restarts of $\text{BERT}_{\text{CoLA}_{\text{large}-\text{uncased}}}$ and $\text{BERT}_{\text{CoLA}_{\text{large}-\text{cased}}}$.

We recreate in Table 2.1 below an updated version of the table of MCC scores on the CoLA test set presented by W&B both in 2018 and 2019. We add a column to indicate the authors responsible for training the model and include our three trained models in the comparison. Additionally, we include two models submitted by Jacob Devlin to the GLUE Leaderboard for additional points of comparison, although we assume the scores presented in the leaderboard are the maximum MCC scores achieved by the models on the CoLA out-of-domain test set.

Our mean MCC scores for $\text{BERT}_{\text{CoLA}_{\text{large}-\text{cased}}}$ were within error margins of the $\text{BERT}_{\text{CoLA}_{\text{large}}}$ model reported by W&B 2019. Additionally, the maximum MCC score achieved here by $\text{BERT}_{\text{CoLA}_{\text{large}-\text{cased}}}$ beat the score posted by Jacob Devlin on the GLUE Leaderboard, and was less than 0.01 away from the maximum MCC score posted by W&B's $\text{BERT}_{\text{CoLA}_{\text{large}}}$. We consider these results to be strongly indicative of successful replication, given the known stochastic variation in such models, and proceed to conduct the remainder of the linguistic analyses presented by W&B 2019 using the CoLA analysis set. However, we focus our analyses exclusively on the three $\text{BERT}_{\text{CoLA}}$ models we trained, and do not replicate the results of the other models, as they are not the focus of this thesis.

When we consider only the major features in the CoLA analysis set, the replication

| **Model$_{\text{CoLA}}$** | **Mean (STD)** | **maximum** | **Ensemble** | **Authors** |
|---|---|---|---|---|
| CoLA baseline | 0.320 (0.007) | 0.330 | 0.320 | W&B 2019 |
| GPT | 0.528 (0.023) | 0.575 | 0.567 | W&B 2019 |
| **BERT$_{\text{large}}$** | **0.582 (0.032)** | **0.622** | **0.601** | **W&B 2019** |
| Human | 0.697 (0.042) | 0.726 | 0.761 | Warstadt et al. 2018 |
| BERT$_{\text{base−uncased}}$ | 0.478 (0.018) | 0.514 | 0.522 | Héctor & friends |
| BERT$_{\text{large−uncased}}$ | 0.542 (0.019) | 0.583 | 0.578 | Héctor & friends |
| **BERT$_{\text{large−cased}}$** | **0.574 (0.026)** | **0.613** | **0.588** | **Héctor & friends** |
| BERT$_{\text{base}}$ | 0.521* (N/A) | 0.521* | 0.521* | Jacob Devlin |
| **BERT$_{\text{large}}$** | **0.605* (N/A)** | **0.605*** | **0.605*** | **Jacob Devlin** |

Table 2.1: Replication of Warstadt & Bowman (2019) with our trained BERT$_{\text{CoLA}}$ models for comparison. Performance (MCC) on the CoLA test set, including mean over restarts of a given model with standard deviation, maximum over restarts, and majority prediction over restarts. We include the BERT$_{\text{CoLA}}$ scores on the GLUE leaderboard for the CoLA task submitted by Jacob Devlin for further points of reference.

trial seems even more promising. In Figure 2-1 below we replicate the first figure in W&B 2019, which shows model performance by major syntactic feature in the CoLA analysis set. We deviate slightly from the authors when plotting mean MCC performance. While they use dashed lines to show MCC performance on the entire CoLA development set, we use them to show MCC performance on the CoLA out-of-domain test set as a follow up to the MCC scores presented in Table 2.1.

Unfortunately, the major features in the CoLA analysis set are where our successful replication ends. By that we mean that studying the finer-grained minor features in the analysis set reveals what the MCC scores on the test set and major features have obscured. Much like Figure 2-1, we replicate the second figure in W&B 2019 in Figure 2-2, but again using the average CoLA out-of-domain test set MCC scores presented in Table 2.1 as the horizontal dashed lines.

The added resolution of the minor features reveals somewhat erratic behavior from the three different BERT models. For one, we observe degenerate performance (MCC = 0) on a number of features, but most notably sentences that contain Complement

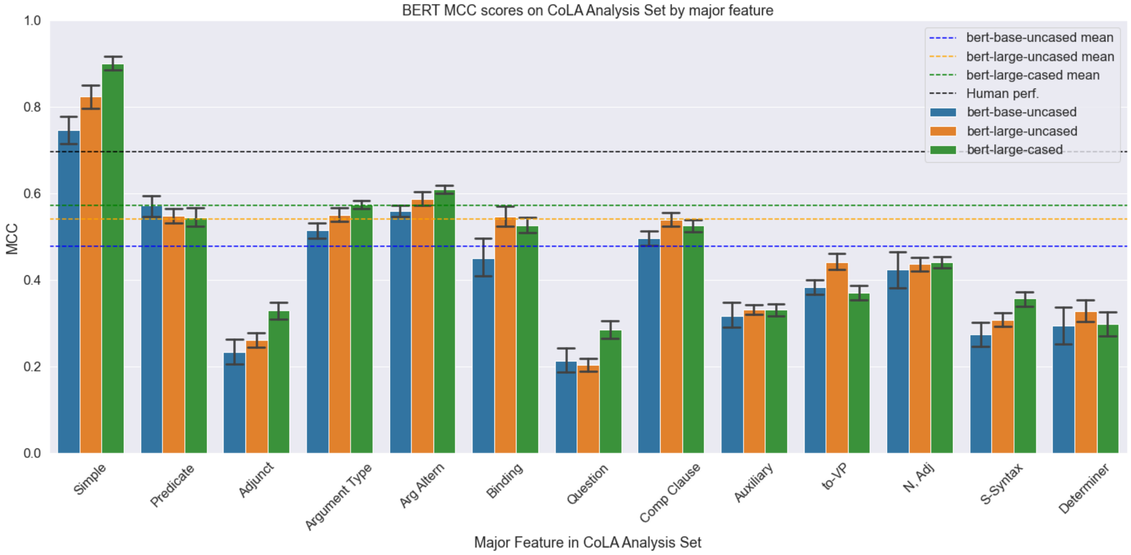


Figure 2-1: Replication of Warstadt & Bowman (2019) with our $\text{BERT}_{\text{CoLA}}$ models for comparison. Performance (MCC) on CoLA analysis set by major feature. Dashed lines show mean performance on the CoLA out-of-domain test set. From left to right, performance for each feature is given for base-uncased, large-uncased, and large-cased.

Clause Subjects (CP Subj), Raising, or Noun-Noun Compounds (NNCompd) toward the right-hand side of Figure 2-2. Thankfully, we only see this degenerate behavior from $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ in NNCompd, but also observe very low MCC scores for a handful of other features. Most notably, the $\text{BERT}_{\text{CoLA}}$ models perform the worst on the Question major feature in Figure 2-1, which also translates to poor performance on Matrix Questions (Matrix Q) and Embedded Questions (Emb Q). Other underperforming minor features of note are the Miscellaneous Adjuncts (Misc), Modal Verbs (Modal) and Negation (Neg). We believe this to be a manifestation of the underspecification phenomenon identified by D'Amour et al. (2020), where near identical performance on I.I.D. test sets is nonetheless met with different combinations of spurious and meaningful correlations acquired during training. Although we do not proceed to investigate the extent of this underspecification behavior as it pertains to W&B 2019, we do investigate to what extent this may be reflected when testing performance on the LI-Adger dataset.

Given the positive MCC results by $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ on the CoLA out-of-domain test set, CoLA analysis set major features, and even many of the minor features

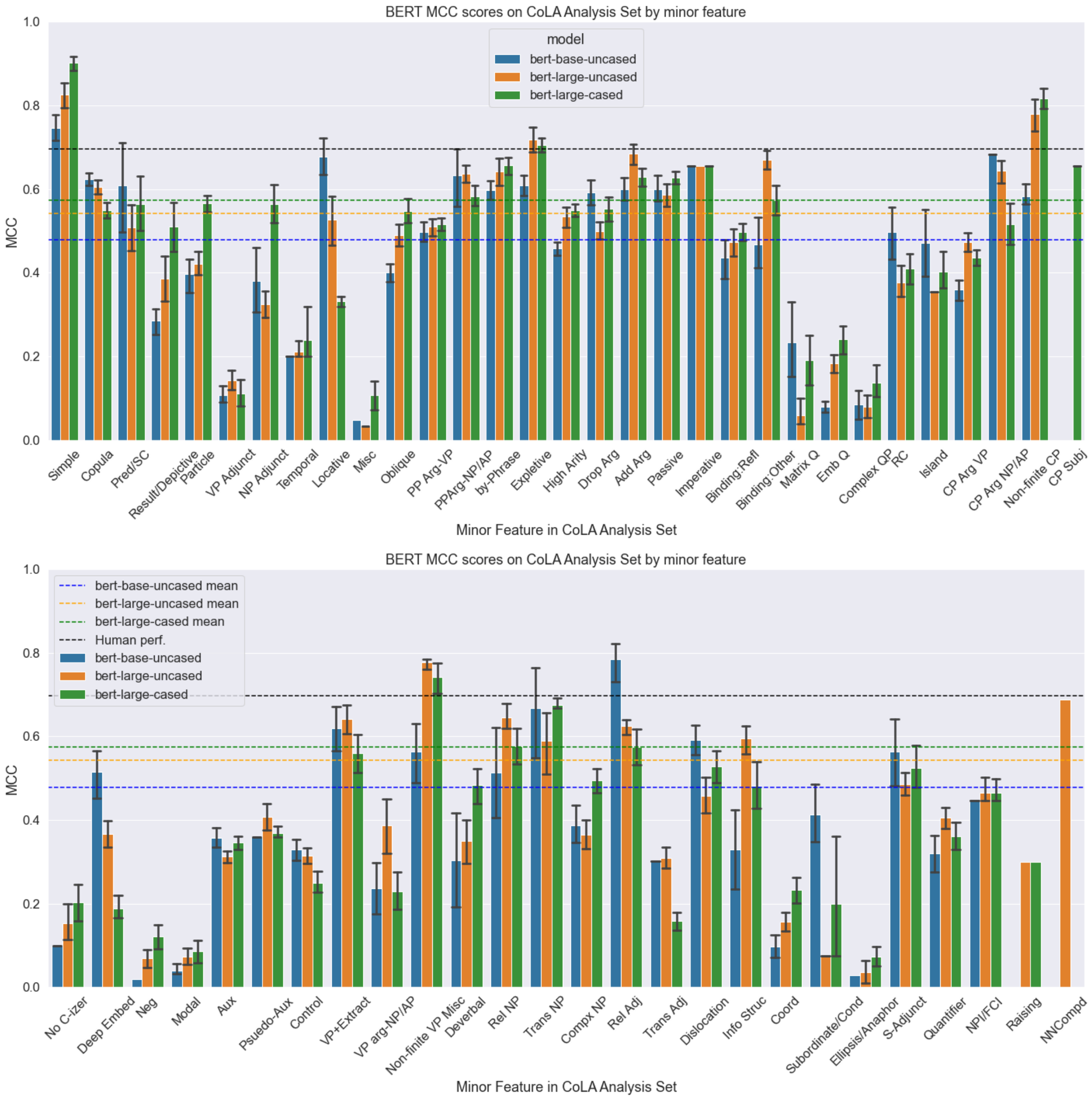


Figure 2-2: Replication of Warstadt & Bowman (2019) with our trained $\text{BERT}_{\text{CoLA}}$ models for comparison. Performance (MCC) on CoLA analysis set by minor feature. Dashed lines show mean performance on the CoLA out-of-domain test set. From left to right, performance for each feature is given for base-uncased, large-uncased, and large-cased.

in the analysis set, we are satisfied with the performance of our $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ model's performance. Accordingly, we select the random restart that yielded the maximum MCC score reported in Table 2.1 as the model to be studied in our later analyses. Lastly, we have grounds to believe our failure to replicate exact results on the CoLA analysis set's minor features is not the fault of our training regime or choice of hyperparameters, but rather a consequence of the overparametrization that is characteristic of the BERT models, and almost certainly all neural LMs (D'Amour et al. 2020).

## 2.3 BERT has too much to drink

The volatility in the three $\text{BERT}_{\text{CoLA}}$ models' predictions revealed by our attempts to replicate the results of Warstadt & Bowman (2019) warrants further investigation. As we briefly alluded to in Section 2.2, we do not intend to assess the degree of instability in the CoLA analysis set, nor do we wish to make claims regarding the validity of the KoL attributed to BERT as a result of Warstadt & Bowman's findings. Our interest here is simple: we want to know how and to what degree the overparametrization of the $\text{BERT}_{\text{CoLA}}$ models may affect the results we observe when obtaining acceptability judgements from $\text{BERT}_{\text{CoLA}}$ on the LI-Adger sentences before applying the Acceptability Delta Criterion.

Due to limited computational resources, we conduct the following experiment. We initialize a single instance of pre-trained $\text{BERT}_{\text{base−uncased}}$ with a linear classification head as expressed in Equation 2.1 at the beginning of the chapter. This is the only point in the experiment where a model is initialized, so the weight matrix $W$ in the linear output layer will always have the exact same starting weights. Next, we make a full copy of the model in order to keep the base initialization without any fine-tuning, and perform fine-tuning on the second copy using the CoLA. However, before performing the fine-tuning, we shuffle the order of the training data according to one random seed. After the fine-tuning process, we gather a categorical acceptability prediction from $\text{BERT}_{\text{CoLA}_{\text{base−uncased}}}$ for each sentence in the LI-Adger sentence by selecting the label with the highest softmax output. I.e. for the sentence *Colorless green ideas sleep furiously* ($s_i$), one random seed of $\text{BERT}_{\text{CoLA}_{\text{base−uncased}}}$ outputs a softmax value of 0.168 for the the unacceptable label ($P(c_j = 0|h_i) = 0.168$), and a softmax value of 0.832 for the acceptable label ($P(c_j = 1|h_i) = 0.832$). Therefore, we determine the model's prediction for that sentence to be 1.0 (acceptable). We express the model's output $out_i$ more succinctly as follows:

$$out_i = \underset{c_j \in \{0,1\}}{\text{argmax}} \Big[ P(c_j | h_i) \Big] \tag{2.2}$$

Although we change this paradigm for later analyses, we continue to use the categorical output for this particular experiment because it allows us to calculate performance using MCC as in the CoLA out-of-domain test set, which is where we first identified this particular instability.

We repeat the process of cloning $\text{BERT}_{\text{base-uncased}}$ and fine-tuning the copy with the reshuffled CoLA training set 200 times. That is, the exact same BERT model is trained on the same data, but shuffled into 200 different orders. Each time we collect the fine-tuned model's categorical predictions for the sentences in the LI-Adger dataset, we compare them to the previous seed's predictions. Every sentence whose predicted label changes between the previous and current random seed is considered an "Acrobatic Sentence" due to its exhibited *fipping* behavior in response to a reordering of the training data. For the sake of consistency, we name the sentences whose predictions remain constant the "Unathletic Sentences," because they do not *flip* back and forth. We plot the percentage of the LI-Adger sentences that fall into the set of Acrobatic Sentences as a function of the number of different training orders used to fine-tune $\text{BERT}_{\text{CoLA}_{\text{base-uncased}}}$ in Figure 2-3. We additionally plot the baseline accuracy of a majority-class predictor.

The pattern of instability in Figure 2-3 has important negative implications for the strength of any conclusions that can be drawn regarding BERT if it needs to be fine-tuned as part of the experiment. Recall this instability arises from the fundamentally underdetermined system of equations the model is trying to solve, which by nature either have no solutions or infinitely many solutions. By having more parameters than data points, BERT, as well as any other such overparametrized model, is able to settle on an unknown number of spurious correlations that may yield good performance on I.I.D. test sets (D'Amour et al. 2020). Figure 2-3 shows that while $\text{BERT}_{\text{CoLA}_{\text{base-uncased}}}$ has a relatively constant MCC score on the LI-Adger dataset,

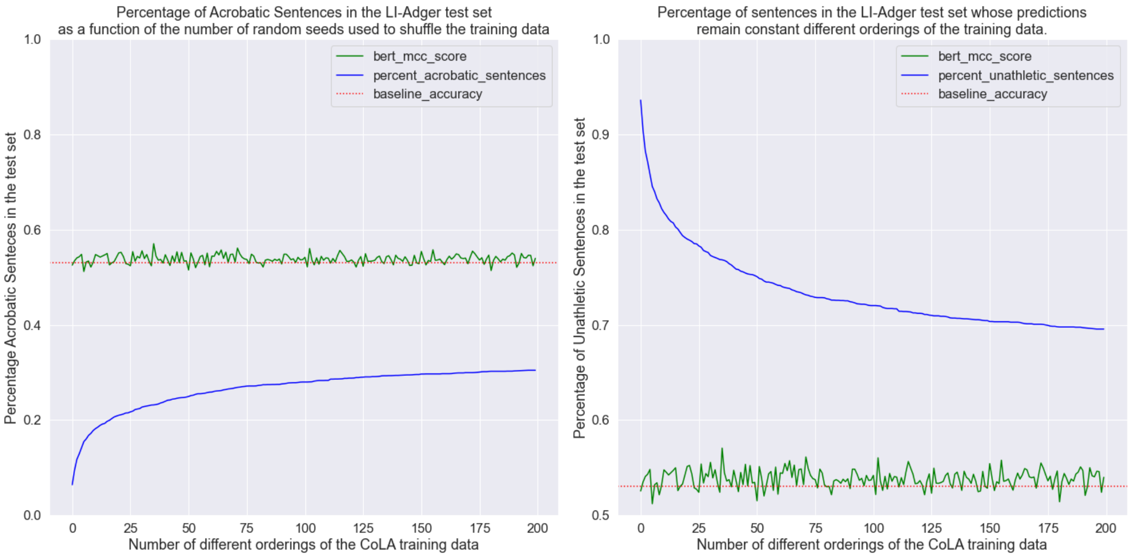


Figure 2-3: As the same initialization of $\text{BERT}_{\text{CoLA}_{\text{base-uncased}}}$ is fine-tuned in different random orders, the percentage of sentences in the test set that become Acrobatic Sentences (left) and the percentage of sentences whose predicted labels remain constant (Unathletic Sentences–right). The MCC score achieved by BERT on the LI-Adger dataset at each random seed is plotted in green. The baseline accuracy of a majority-class predictor is plotted in orange.

it changed its predictions for 1272 out of the 4178 total sentences in the LI-Adger dataset. What is more sobering is the fact that we performed this test with the $\text{BERT}_{\text{base-uncased}}$ model, the version of BERT with the fewest parameters. Although we do not conduct the same test with $\text{BERT}_{\text{large-cased}}$ out of a lack of computational resources, we have no reason to believe this instability will be any less pronounced. On the contrary, because we are effectively tripling the number of free parameters (from 110M in $\text{BERT}_{\text{base}}$ to 340M in $\text{BERT}_{\text{large}}$) in an already underdetermined system of equations, we expect nothing other than an even more severe instability.

At this point we can summarize what this implies for our Acceptability Delta Criterion test using the LI-Adger dataset. We briefly discussed in Section 1.1 why we do not believe probing to be the best approach to assessing the KoL of neural LMs. The results observed in Figure 2-3 are strong evidence for the lack of reliability of such experiments. However, we will proceed to study the performance of $\text{BERT}_{\text{CoLA}}$ models on the LI-Adger dataset under the Acceptability Delta Criterion nonetheless.

We will not yet dismiss the argument that BERT must be fine-tuned for sequence classification in order to perform its best (Sun et al. 2019). However, we will evaluate BERT by adopting some of the recommendations of D'Amour et al. 2020.

Let us momentarily set aside the scientific question and consider BERT for what it is: an engineering achievement capable of state-of-the-art performance in deployment of multiple different NLP services and tasks such as Google Search. If we take the LI-Adger dataset as an example of data that will be observed in deployment, then we cannot select a BERT model to evaluate according to MCC performance on the LI-Adger dataset. After all, if one had access to the data that that will be seen in deployment... what would be the point of all the research that has been conducted for Machine Learning models to generalize beyond the training data? Consequently, we select the $\text{BERT}_{\text{CoLA}}$ model to evaluate according to MCC performance on the CoLA out-of-domain test set, and then evaluate its performance on an entirely unseen LI-Adger dataset.

The last point we wish to make is that we select a **single** model out of the multiple random restarts of each of the $\text{BERT}_{\text{CoLA}}$ models instead of averaging predictions across them. This is in part because the ML pipeline typically selects the model that best performs on the held-out test set and uses it in deployment. The principal reason for this approach, however, is that we wish to evaluate the KoL contained in the model. By averaging the predictions of multiple different random restarts of the same model, especially with the degree of instability observed in Figure 2-3, we might mask anything meaningful that we could learn about the models, because the test would amount to evaluating the average of many different spurious correlations, even if it results in better performance overall.

## 2.4 Benchmarking with the LI-Adger Dataset

Having evaluated all the models on the CoLA out-of-domain test set, we select the best performing random restart of each model to evaluate under the Acceptability Delta Criterion using the LI-Adger dataset. However, before applying the Acceptability

Delta Criterion, we treat the LI-Adger dataset as if it were the CoLA test set: We assign each sentence its original, categorical (1/0) expert label and evaluate each model's MCC performance on the LI-Adger. This is to ensure that the LI-Adger dataset is not *a priori* too easy or too hard for the models. Table 2.2 displays the MCC scores of the best performing BERT$_{\text{CoLA}}$ models both on the CoLA out-of-domain test set and the LI-Adger dataset.

| **BERT$_{\text{CoLA}}$ model** | **CoLA test set MCC score** | **LI-Adger MCC score** |
|---|---|---|
| base-uncased | 0.514 | 0.553 |
| large-uncased | 0.583 | 0.576 |
| **large-cased** | **0.613** | **0.595** |

Table 2.2: MCC scores for each of the chosen BERT$_{\text{CoLA}}$ models on the CoLA out-of-domain test set and the LI-Adger dataset when using the expert labels as the true labels to be predicted. The models selected had the highest MCC score on the CoLA out-of-domain test set out of all the other random restarts.

The MCC scores on the LI-Adger dataset presented in Table 2.2 confirm that there is no overt abnormality in the BERT$_{\text{CoLA}}$ models' behavior with the dataset. The next step is to evaluate how the models' predictions correlate with the human judgements on an individual sentence level. In order to do this, we need to make the models' predictions gradient, which is also a prerequisite of the Acceptability Delta Criterion. The first change we make is to update our expert labels to be ±1 instead of 1/0. Now recall our example from Section 2.3 with the sentence *Colorless green ideas sleep furiously.* Rewriting the model output with the new labels would look as follows: $P(c_j = -1|h_i) = 0.168$, and $P(c_j = 1|h_i) = 0.832$. The traditional paradigm calls for selecting the label $c_j$ with the highest softmax output as the categorical prediction, but what we will do is multiply the chosen label by the model's *confidence* in that label. This results in the following equation describing each BERT$_{\text{CoLA}}$ model's output $out_i$ for sentence $s_i$:

$$out_i = \underset{c_j \in \{-1,+1\}}{\text{argmax}} \Big[ P(c_j|h_i) \Big] * \max_{c_j \in -1,+1} \Big[ P(c_j|h_i) \Big] \tag{2.3}$$

With this formulation, we can easily retrieve both the predicted categorical label (sign($out_i$)) and the model's *confidence.* According to the findings surrounding BERT's KoL described in Section 2.1, we do not find it unreasonable to expect this reformulated $\text{BERT}_{\text{CoLA}}$ output to track human judgements through the whole range of acceptability, from completely unacceptable sentences to completely acceptable ones.

Following the reformulated output, we must now rely on Pearson's correlation coefficient (PCC) instead of the categorically-based MCC, due to the now gradient nature of both the $\text{BERT}_{\text{CoLA}}$ and human judgements. We do not find this to be problematic because one of the main benefits of MCC, that it works well in cases where there is a class imbalance, is unnecessary for the LI-Adger dataset. The distribution of acceptable to unacceptable sentences, according to the expert labels, is 2217 acceptable and 1961 unacceptable, which we consider to be fairly balanced.

In addition to the correlations between the three $\text{BERT}_{\text{CoLA}}$ models and the human ME judgements, we add the SLOR and log likelihood scores of a trigram model trained on the British National Corpus (BNC) by Sprouse et al. (2018). We compute the full correlation matrix for all six metrics and display the results in Figure 2-4. All correlations shown have a $p < 0.0001$.

At first glance, the three $\text{BERT}_{\text{CoLA}}$ models have a moderate positive correlation of slightly above 0.6 with the human judgements on the LI-Adger dataset. However, upon closer inspection, it quickly becomes apparent that this summary statistic can be deceptive. We show in Figure 2-5 a scatterplot of the $\text{BERT}_{\text{CoLA}}$ models' predictions as the $x$ coordinate and the human judgements as the $y$ coordinate.

Figure 2-5 shows how, despite a PCC of $> 0.6$ across all three $\text{BERT}_{\text{CoLA}}$ models, they all fail to capture the full range of possible acceptability scores. In fact, it seems that the three $\text{BERT}_{\text{CoLA}}$ models, contrary to our expectations, mostly only output acceptability scores either greater than 0.9 or less than $-0.9$. We believe there are at least two (but likely more) possible sources of this behavior. This may be a symptom of the underspecification observed in 2.3, where the free parameters have led $\text{BERT}_{\text{CoLA}}$ to effectively *memorize* the categorical labels instead of generalizing

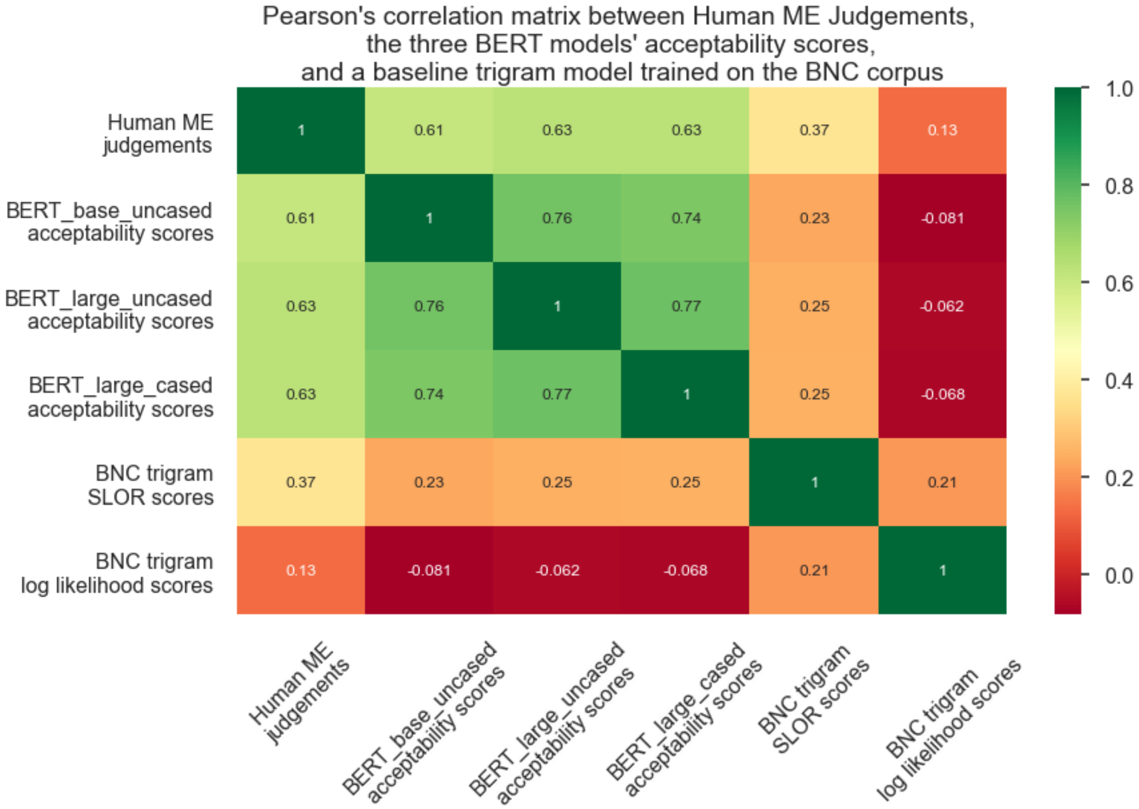


Figure 2-4: PCC matrix between human judgements and all three $\text{BERT}_{\text{CoLA}}$ models. In addition we add the SLOR and log likelihood scores of a trigram model trained on the British National Corpus by Sprouse et al. 2018 for additional reference. All correlations shown have a $p < 0.0001$.

to gradient acceptability values. The other possible explanation could be that the softmax output from Equation 2.1 is consistently overestimating the label probabilities, either because the output layer was not calibrated as per Guo et al. (2017), or the learned weight embedding matrix $W$ is unable to map the encoded output of $\text{BERT}_{\text{CoLA}}$'s final hidden layer to the acceptability gradient. We do not take a stance as to what the cause of the behavior may be, and merely use it to have an idea of what to expect from the three $\text{BERT}_{\text{CoLA}}$ models when we apply the Acceptability Delta Criterion.

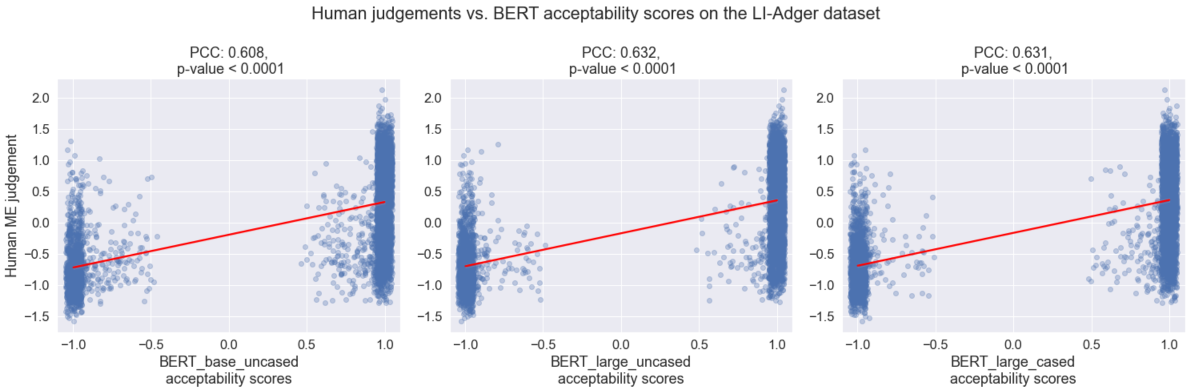


Figure 2-5: Scatterplot of human judgements ($y$-axis) vs. $\text{BERT}_{\text{CoLA}}$ acceptability score output for each sentence in the LI-Adger dataset with best-fit line in red. We add a jitter of 0.05 along the $x$-axis and lower the alpha to 0.3 to highlight the density of the points.

## 2.5 BERT takes on the ADC (ADC round 1)

In light of how the three $\text{BERT}_{\text{CoLA}}$ models fail to account for the full range of acceptability values in the LI-Adger dataset, the three models now become a good case study of how well the Acceptability Delta Criterion (ADC) may be able to discern this. As a brief reminder, the ADC requires that the model outputs be Z-score transformed such that they are expressed in terms of standard deviation units. Afterward, the LI-Adger sentences are assembled into minimal pairs, with each minimal pair ($m_i$) being associated to an Acceptability Delta from the human judgements ($\Delta h_i$) and from the models' Z-score transformed outputs ($\Delta lm_i$). Then, if the distance between $\Delta h_i$ and $\Delta lm_i$ is greater than $\delta$ (Equation 1.2) or if $\Delta h_i$ and $\Delta lm_i$ differ in sign (Equation 1.2), the ADC for $m_i$ is not met. For further details on the principles underlying the ADC, we refer the reader to 1.3.3.

For the remainder of this section, we will apply the ADC on the three $\text{BERT}_{\text{CoLA}}$ models to output acceptability scores over whole sentences. In addition, we include into our analysis the SLOR and log likelihood scores of a trigram model trained on the BNC by Sprouse et al. 2018 as a baseline.

### 2.5.1 Correlations at the minimal pair level

The first step in observing how the ADC may perform with the $\text{BERT}_{\text{CoLA}}$ models and the trigram baseline is to update the PCCs using the acceptability deltas (i.e. $\Delta lm_i$) at the minimal pair level, instead of the raw acceptability scores at the individual sentence level. Note that we will first Z-score transform the raw acceptability output from the $\text{BERT}_{\text{CoLA}}$ models and the trigram model before computing the deltas. This preliminary step does not affect the correlations because, as discussed at length in Section 1.3.3, the Z-score transformation is a linear operation that does not introduce distortion into the data (Schütze & Sprouse 2013,pp 27-51). We recreate in Figure 2-4 an updated correlation matrix using the newly computed acceptability deltas from the six metrics compared in Section 2.4.

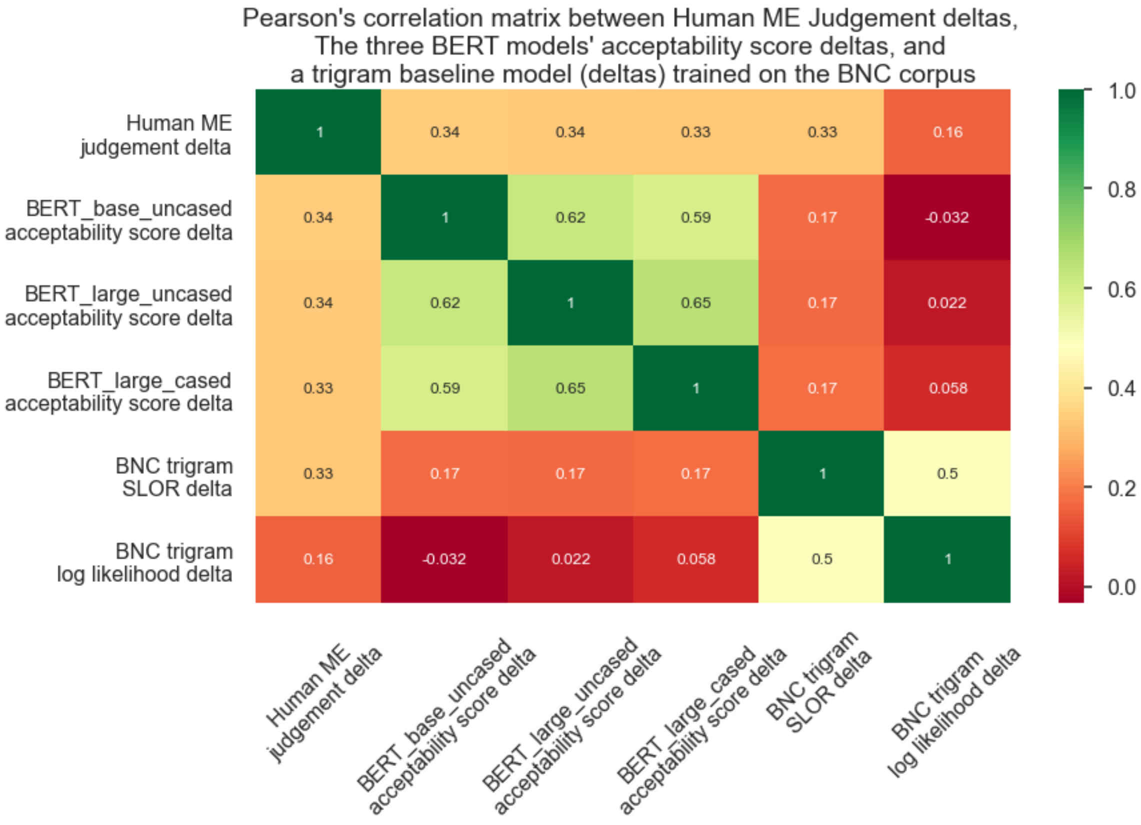


Figure 2-6: PCC matrix between human judgements and all three $\text{BERT}_{\text{CoLA}}$ models. In addition we add the SLOR and log likelihood scores of a trigram model trained on the British National Corpus by Sprouse et al. 2018 for additional reference. All correlations shown have a $p < 0.0001$.

We see a precipitous drop in the PCCs between the human data and the three $\text{BERT}_{\text{CoLA}}$ models, yet the trigram model's PCCs remain relatively constant. This lends credence to the idea that the source of the moderate to strong positive correlation observed in Section 2.4 was a result of the same spurious correlations causing the instability in Section 2.3. For ease of comparison, we present in Table 2.3 the PCCs between the human judgement data and the 5 models currently under study at the individual sentence level and at the minimal pair (delta) level.

| **Model** | **PCC LI-Adger sentences** | **PCC LI-Adger min pairs** |
|---|---|---|
| $\text{BERT}_{\text{CoLA}_{\text{base}-\text{uncased}}}$ | 0.608 | 0.340 |
| $\text{BERT}_{\text{CoLA}_{\text{large}-\text{uncased}}}$ | 0.632 | 0.338 |
| $\text{BERT}_{\text{CoLA}_{\text{large}-\text{cased}}}$ | 0.631 | 0.335 |
| $\text{trigram}_{SLOR}$ | 0.368 | 0.333 |
| $\text{trigram}_{log-prob}$ | 0.131 | 0.156 |

Table 2.3: PCCs on the LI-Adger dataset on invididual sentences (middle) and across minimal pairs (right) between human acceptability judgements and 5 other models. We include three $\text{BERT}_{\text{CoLA}}$ models, as well as SLOR and log-likelihood scores from a trigram model trained on the British National Corpus by Sprouse et al. 2018

To conclude this section, we redraw the correlation plots with best-fit lines from Section 2.4 but now using the acceptability delta values at the minimal pair level. Figure 2-7 confirms our suspicions regarding the $\text{BERT}_{\text{CoLA}}$ models' behavior: despite our reformulation of $\text{BERT}_{\text{CoLA}}$ models' outputs in order to make them gradient in line with human acceptability judgements, the models consistently predicted sentences to be either more than 90% acceptable or less than 90% unacceptable.

Despite our best efforts to have $BERT_{\text{CoLA}}$ output gradient acceptability judgements with the formulation discussed in Equation 2.1, the fine-tuning phase on categorical data only seems to cripple the models' performance, contrary to our expectations as discussed in 2.1. What we find most surprising is not the precipitous drop in PCC when considering a simple delta across a minimal pair, but that the PCC was at the level of that of a trigram model. We investigate this further by studying the three $\text{BERT}_{\text{CoLA}}$ models' performance under the ADC, comparing them against the

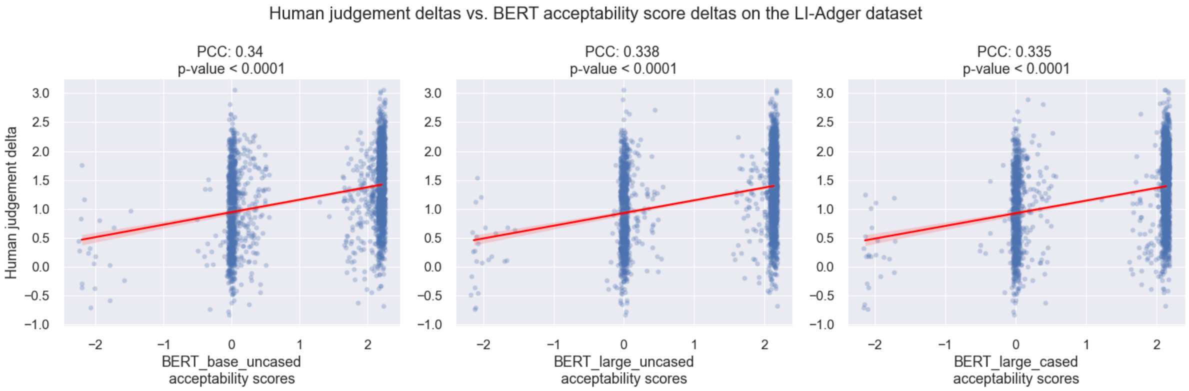


Figure 2-7: Scatterplot of human judgement deltas ($y$-axis) vs. $\text{BERT}_{\text{CoLA}}$ acceptability score deltas for each minimal pair in the LI-Adger dataset with best-fit line in red. We add a jitter of 0.05 along the $x$-axis and lower the alpha to 0.3 to highlight the density of the points.

trigram baseline.

### 2.5.2 Applying the Acceptability Delta Criterion

Here we conduct one final analysis of the $\text{BERT}_{\text{CoLA}}$ models. We benchmark the three models as well as the two trigram metrics using the BLiMP Criterion and the ADC using three different values of $\delta$. We use $\delta = 0.5$ as the strictest test, requiring that the LM acceptability delta ($\Delta lm_i$) and the human judgement delta ($\Delta h_i$) be within 0.5 standard deviation units of each other. Increasing $\delta$ makes the test progressively easier, until it functionally becomes very similar to the BLiMP criterion, with the crucial difference being that the BLiMP Criterion maintains the expert labels as the ground truth, whereas the ADC will use the sign of the human judgements as the true label of each sentence.

In order to test how the ADC generalizes to a form similar to the BLiMP criterion, we add two additional ADC tests: one with $\delta = 1$ and one with $\delta = 5$. We report the results of our 4 tests in Table 2.4.

The initial results of the ADC are very promising. For one, the $\text{BERT}_{\text{CoLA}}$ model scores under $\delta = 0.5$ are in line with our expectations from the PCCs observed in Section 2.5.1. This is supported by the trigram's SLOR performance also under the

| **Model** | **BLiMP** | **ADC,** $\delta = 0.5$ | **ADC,** $\delta = 1.0$ | **ADC,** $\delta = 5.0$ |
|---|---|---|---|---|
| $\text{BERT}_{\text{CoLA}_{\text{base-uncased}}}$ | 0.915 | 0.286 | 0.538 | 0.902 |
| $\text{BERT}_{\text{CoLA}_{\text{large-uncased}}}$ | 0.917 | 0.311 | 0.564 | 0.907 |
| $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ | 0.936 | 0.307 | 0.561 | 0.925 |
| $\text{trigram}_{SLOR}$ | 0.753 | 0.301 | 0.520 | 0.744 |
| $\text{trigram}_{log-prob}$ | 0.671 | 0.165 | 0.329 | 0.668 |

Table 2.4: Comparison between the models' BLiMP and ADC scores, using $\delta$={0.5, 1.0, 5.0}. We include three $\text{BERT}_{\text{CoLA}}$ models, as well as SLOR and log-likelihood scores from a trigram model trained on the British National Corpus by Sprouse et al. 2018

ADC with $\delta = 0.5$, which is also extremely close to the $\text{BERT}_{\text{CoLA}}$ models' PCCs at the minimal pair level. It is also a promising sign to see that the $\text{BERT}_{\text{CoLA}}$ model scores and the trigram SLOR scores scale together at $\delta = 1.0$. Lastly, setting $\delta = 5.0$ as a large number yielded scores for all 5 models very close to their performance under the BLiMP Criterion, strongly suggesting that the ADC is in fact a generalization of the BLiMP criterion when the stricter $\delta$ measure is relaxed. Testing this further, we present 4 example minimal pairs that $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ scored correctly under the BLiMP Criterion, but not under the ADC with $\delta = 5.0$ in Table 2.5.

The common factor among the four minimal pairs presented in Table 2.5, and the other 54 minimal pairs where $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ fulfilled the BLiMP criterion but not the ADC with $\delta = 5$, is that the human judgements disagree with the expert categorization. This is, by design, one of the crucial properties of the ADC, because ultimately linguistic theory is developed by probing either language use or language acquisition.

With a clearer idea of the difference between the BLiMP Criterion and the generalized ADC, one big question remains: how much overlap is there between the performance of the trigram model and the $\text{BERT}_{\text{CoLA}}$ models? In other words, is it purely coincidental that all three $\text{BERT}_{\text{CoLA}}$ models and the trigram SLOR scores had extremely close PCCs on the acceptability deltas **and** extremely close scores on the ADC for both $\delta = 0.5$ and $\delta = 1.0$? If we consider that $N$-gram class models

| **Minimal Pair**<br>**Top**: Acceptable \| **Bottom**: Unacceptable | Human<br>judgement | BERT<br>acceptability |
|---|---|---|
| We proved Amelia to the manager to be responsible. | -0.56008 | 0.732817911 |
| *We proved to the manager Amelia to be responsible. | -0.13864 | -1.39757562 |
| There is likely to live a snake in the garden. | -0.6451 | -1.02182 |
| *There is likely a snake to live in the garden. | -0.51201 | -1.39602 |
| Jenny would accurately have calculated the results. | 0.345683 | -1.340338319 |
| *Jenny accurately will calculate the results. | 0.501494 | -1.40060934 |
| The announcer's introduction of Ted was humorous. | 0.659471 | 0.73306608 |
| The announcer's introduction of Ted's was humorous. | 0.748718 | -1.335794047 |

Table 2.5: Four minimal pairs where the $\mathrm{BERT}_{\mathrm{CoLA}}$ models meet the BLiMP Criterion but not the generalized ADC with $\delta = 5.0$. We report the BERT acceptability score from $\mathrm{BERT}_{\mathrm{CoLA}_{\mathrm{large-cased}}}$ The human judgement and BERT acceptability scores are already z-score transformed. The common factor is that the human judgements disagree with the BLiMP Criterion.

have little to no KoL other than word cooccurrence, and that the $\mathrm{BERT}_{\mathrm{CoLA}}$ models did not seem to track sentences across the acceptability spectrum, this is a question that demands at least some cursory sanity checks. Accordingly, we perform two more evaluations of the ADC but with different datasets. For the first, we use the results for the ADC with $\delta = 0.5$ from Table 2.4 to subtract from the LI-Adger dataset all the minimal pairs where the trigram SLOR deltas met the ADC. We hope that by doing so, we will have factored out from the dataset all the minimal pairs that can be correctly tracked using purely word cooccurrence statistics, thereby leaving behind only minimal pairs whose acceptability delta requires the $\mathrm{BERT}_{\mathrm{CoLA}}$ models' extra machinery (parameters) to correctly track. Table 2.6 presents the $\mathrm{BERT}_{\mathrm{CoLA}}$ models' scores under the ADC with $\delta$={0.5,1.0} before and after removing the minimal pairs where the trigram SLOR model met the ADC.

Poor performance aside, Table 2.6 reveals a small change in overall performance and a low percentage of minimal pairs correctly scored by both the $\mathrm{BERT}_{\mathrm{CoLA}}$ models and the trigram model using SLOR. This allays concerns that the $\mathrm{BERT}_{\mathrm{CoLA}}$ models

| **Model** | **ADC,** $\delta = 0.5$ | | | **ADC,** $\delta = 1.0$ | | |
|---|---|---|---|---|---|---|
| **BERT**$_{\text{CoLA}}$ | **Original** | **Reduced** | **Overlap** | **Original** | **Reduced** | **Overlap** |
| base-uncased | 0.286 | 0.294 | 8.08% | 0.538 | 0.546 | 27.6% |
| large-uncased | 0.311 | 0.309 | 9.51% | 0.564 | 0.557 | 29.6% |
| large-cased | 0.307 | 0.313 | 8.84% | 0.561 | 0.560 | 29.2% |

Table 2.6: BERT$_{\text{CoLA}}$ models' performance on the ADC with $\delta$=0,5,1.0 before (Original) and after (Reduced) removing all minimal pairs for which the ADC Criterion was met by the trigram baseline model trained on the BNC corpus. The Overlap columns display the percentage of minimal pairs that both the BERT$_{\text{CoLA}}$ model and the trigram baseline pass.

might only be doing well on minimal pairs that the trigram model also predicts correctly. To conclude, we inspect a few example minimal pairs that all three BERT$_{\text{CoLA}}$ models scored correctly but the trigram did not (Table 2.7), and vice versa: a few example minimal pairs the trigram model scored correctly using SLOR but none of the three BERT$_{\text{CoLA}}$ models did (Table 2.8).

| **Minimal Pair** | trigram | BERT$_{\text{CoLA}}$ |
|---|---|---|
| **Top**: Acceptable \| **Bottom**: Unacceptable | SLOR | acceptability |
| She taught the students math. | -0.685949 | 0.73276 |
| *She taught math the students. | -0.562807 | -1.40171 |
| There are linguists available. | -0.337031 | 0.732224 |
| *There are linguists tall. | -0.512287 | -1.41472 |
| Our professor gave no extensions to any students. | -0.728557 | 0.721394 |
| *Our professor gave any extensions to no students. | -1.33971 | -1.3493 |
| What did you address to whom? | 0.478869 | 0.73067 |
| *To whom did you address what? | -0.890322 | 0.681159 |

Table 2.7: Four minimal pairs where all BERT$_{\text{CoLA}}$ models meet the ADC with $\delta = 0.5$ but the trigram baseline does not. We report the acceptability score from BERT$_{\text{CoLA}_{\text{large-cased}}}$. The trigam SLOR and BERT$_{\text{CoLA}_{\text{large-cased}}}$ acceptability scores are already z-score transformed.

Taking Table 2.7 as the only source of evidence, it seems the trigram model is

| **Minimal Pair**<br>**Top**: Acceptable \| **Bottom**: Unacceptable | trigram<br>SLOR | $\text{BERT}_{\text{CoLA}}$<br>acceptability |
|---|---|---|
| Michael managed to drive his car. | 0.950214 | 0.733175 |
| *Michael managed to have driven his car. | 0.26957 | -1.37701 |
| Paul flew to Ireland and Laura sailed to Greece. | 0.253906 | 0.733086 |
| *Paul flew Ireland and Laura sailed to Greece. | -0.779695 | 0.731989 |
| She ran into Spencer and asked him out. | 0.821345 | 0.73299 |
| *She ran into Spencer and asked out. | -0.194269 | -1.38959 |
| The children are almost all sleeping. | 0.30149 | 0.733162 |
| The children almost all are sleeping. | -0.680258 | 0.729437 |

Table 2.8: Four minimal pairs where the trigram baseline meets the ADC with $\delta = 0.5$ but none of the BERT models do. We report the BERT acceptability score from $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$. The trigam SLOR and $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ acceptability scores are already z-score transformed.

failing to meet the ADC when the words in the unacceptable sentence of the minimal pair are okay locally, but result in a very overtly bad sentence. The main exception would perhaps be the last example in 2.7, in which the $\text{BERT}_{\text{CoLA}}$ models correctly agreed with the human judgements that both sentences of the pair are okay. The trigram model likely struggled with the frequency imbalance between *What* and *To whom* at the start and end of both sentences.

However, the $\text{BERT}_{\text{CoLA}}$ models' lack of gradience is revealed when considering Table 2.8. Most of the $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$'s predictions shown in Tables 2.5, 2.7 and 2.8 are either very unacceptable ($\sim$-1.3) or very acceptable ($\sim 0.7$). Precisely this behavior is what causes the $\text{BERT}_{\text{CoLA}}$ models to have such a low performance under the ADC, and such low PCCs in Section 2.5.1. This is made even clearer when considering the third minimal pair in Table 2.8: both the trigram and $\text{BERT}_{\text{CoLA}}$ agree in the sign of their predictions, but $\text{BERT}_{\text{CoLA}}$ predicts a shift from completely unacceptable to completely acceptable, unlike the trigram whose acceptability delta is more moderate. This leads the trigram to fall within the 0.5 standard deviation units required by ADC with $\delta = 0.5$, whereas the three $\text{BERT}_{\text{CoLA}}$ models do not.

# Chapter 3

# Evaluating out-of-the-box BERT

The poor performance of $\text{BERT}_{\text{CoLA}}$ may have understandably raised many of the concerns we expressed in Section 1.1 regarding the strength of the conclusions that can be drawn from using a probing approach. As it stands now, the results observed in Sections 2.5.1 and 2.5.2 suggest that the additional performance afforded by $\text{BERT}_{\text{CoLA}}$ over a trigram model is due to the extra machinery relying on a collection of spurious correlations that provide good I.I.D. test set performance.

In order to verify the validity of this interpretation, we conduct one final case study. Having seen how the ADC works as a function of $\delta$ and having compared it to the BLiMP criterion, we now apply it directly to the out-of-the-box versions of the BERT models ($\text{BERT}_{\text{MLM}}$) studied in Chapter 2. This addresses what we see as two critical weak points in our analyses thus far. The first is that of the instability during the fine-tuning phase observed in Section 2.3; there exists the possibility that there is a particular seed in which each $\text{BERT}_{\text{CoLA}}$ model performs much better or much worse on the ADC. The second is that we have no control over what information is being introduced into the system with CoLA. Although we were clear about our expectations for $\text{BERT}_{\text{CoLA}}$ and their basis in Section 2.1, we hope that by removing CoLA from the analysis pipeline, we address concerns regarding the lack of gradience in $\text{BERT}_{\text{CoLA}}$'s output, as formulated in Equation 2.3.

## 3.1 BERT Masked Language Modeling

The pre-trained BERT models this thesis has been using in its analyses are the publicly available pre-trained model checkpoints originally published by Devlin et al. 2018. The models were pre-trained using two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). For these analyses, we will focus on MLM, a variant of the Cloze task (Taylor, 1953). MLM involves randomly masking approximately 15% of the subword tokens provided to the BERT model as input during training, with some variation in order to achieve robustness in later fine-tuning stages. The model is then asked to predict the original token behind the mask by feeding the final hidden vector directly into a softmax ouptut layer with an output node for each item in the model's vocabulary. The weight update is calculated afterward using cross entropy loss (Devlin et al. 2018).

Because MLM is one of the tasks used to pre-train BERT in the first place, we use it to test the models in their out-of-the-box state. By masking each token in a sentence $s_i$ sequentially and recovering the log likelihood of the original token, we are able to calculate a *pseudo-log-likelihood* (PLL) score for the sentence. Salazar et al. 2020 have shown that BERT's PLL scores are able to outperform GPT-2 on the BLiMP Criterion, as well as other natural language benchmarks. They attribute this success to the PLL's unsupervised expression of linguistic acceptability without left-to-right bias (Salazar et al. 2020); BERT is better able to leverage the entire left and right context of each masked token in order to calculate original token's likelihood. This altogether strongly favors PLL scores as the ideal metric to test the out-of-the-box BERT models with the ADC.

To express the concept of the PLL metric more formally, suppose we want to get the PLL score of the sentence *Colorless green ideas sleep furiously* $s_i$ from $\text{BERT}_{\text{MLM}}$. For each word $w_j$ in the sentence $s_i$, we first replace $w_j$ using a `[MASK]` token, and then apply the softmax function defined in Equation 1 directly on the encoded output

$h_{i,j}$ of BERT$_{\text{MLM}}$'s hidden layers, written below.

$$P(w_j|h_{i,j}) = \text{softmax}(h_{i,j}) \tag{3.1}$$

We can rewrite the probability of the token $w_j$ given the hidden vector $h_{i,j}$ as the probability of the token given its entire left and right context, the principal advantage of the MLM approach:

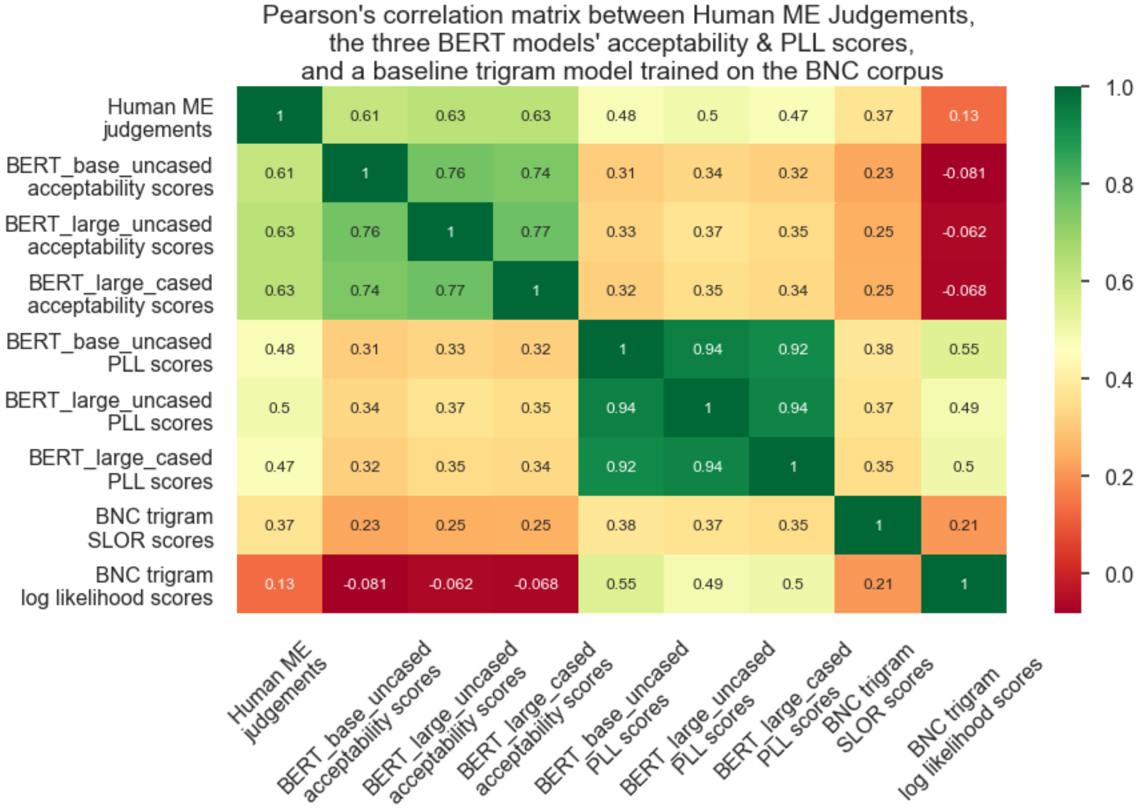
$$P(w_j|w_0, ..., w_{j-1}, w_{j+1}, ...w_n)) = \text{softmax}(h_{i,j}) \tag{3.2}$$

And now this final form can be used by Equation 3, which sums over the MLM log probabilities of each of the tokens to produce the *pseudo log-likelihood*, PLL$(s_i)$.

## 3.2 Correlations with human judgements

Before applying the BLiMP Criterion and the ADC, it seems appropriate to carry out a pilot analysis by calculating the PCCs between the three BERT$_{\text{MLM}}$ models and the human judgements when using the models' PLL scores for individual sentences in the LI-Adger datatset. We present in Figure 3-1 an updated correlation matrix containing the PCCs for the new PLL scores as well as the BERT$_{\text{CoLA}}$ acceptability outputs presented in Figure 3-1.

Additionally, we update our correlation graph from Figure 2-5 in order to observe how the PLL scores may account for the full range of acceptability on an individual sentence level.



| | Human ME judgements | BERT_base_uncased acceptability scores | BERT_large_uncased acceptability scores | BERT_large_cased acceptability scores | BERT_base_uncased PLL scores | BERT_large_uncased PLL scores | BERT_large_cased PLL scores | BNC trigram SLOR scores | BNC trigram log likelihood scores |
|---|---|---|---|---|---|---|---|---|---|
| Human ME judgements | 1 | 0.61 | 0.63 | 0.63 | 0.48 | 0.5 | 0.47 | 0.37 | 0.13 |
| BERT_base_uncased acceptability scores | 0.61 | 1 | 0.76 | 0.74 | 0.31 | 0.34 | 0.32 | 0.23 | -0.081 |
| BERT_large_uncased acceptability scores | 0.63 | 0.76 | 1 | 0.77 | 0.33 | 0.37 | 0.35 | 0.25 | -0.062 |
| BERT_large_cased acceptability scores | 0.63 | 0.74 | 0.77 | 1 | 0.32 | 0.35 | 0.34 | 0.25 | -0.068 |
| BERT_base_uncased PLL scores | 0.48 | 0.31 | 0.33 | 0.32 | 1 | 0.94 | 0.92 | 0.38 | 0.55 |
| BERT_large_uncased PLL scores | 0.5 | 0.34 | 0.37 | 0.35 | 0.94 | 1 | 0.94 | 0.37 | 0.49 |
| BERT_large_cased PLL scores | 0.47 | 0.32 | 0.35 | 0.34 | 0.92 | 0.94 | 1 | 0.35 | 0.5 |
| BNC trigram SLOR scores | 0.37 | 0.23 | 0.25 | 0.25 | 0.38 | 0.37 | 0.35 | 1 | 0.21 |
| BNC trigram log likelihood scores | 0.13 | -0.081 | -0.062 | -0.068 | 0.55 | 0.49 | 0.5 | 0.21 | 1 |

Figure 3-1: PCC matrix between human judgements, $\text{BERT}_{\text{CoLA}}$ acceptability scores, & $\text{BERT}_{\text{MLM}}$ PLL scores from all three BERT models. In addition we add the SLOR and log likelihood scores of a trigram model trained on the British National Corpus by Sprouse et al. 2018 for additional reference. All correlations shown have a $p < 0.0001$.

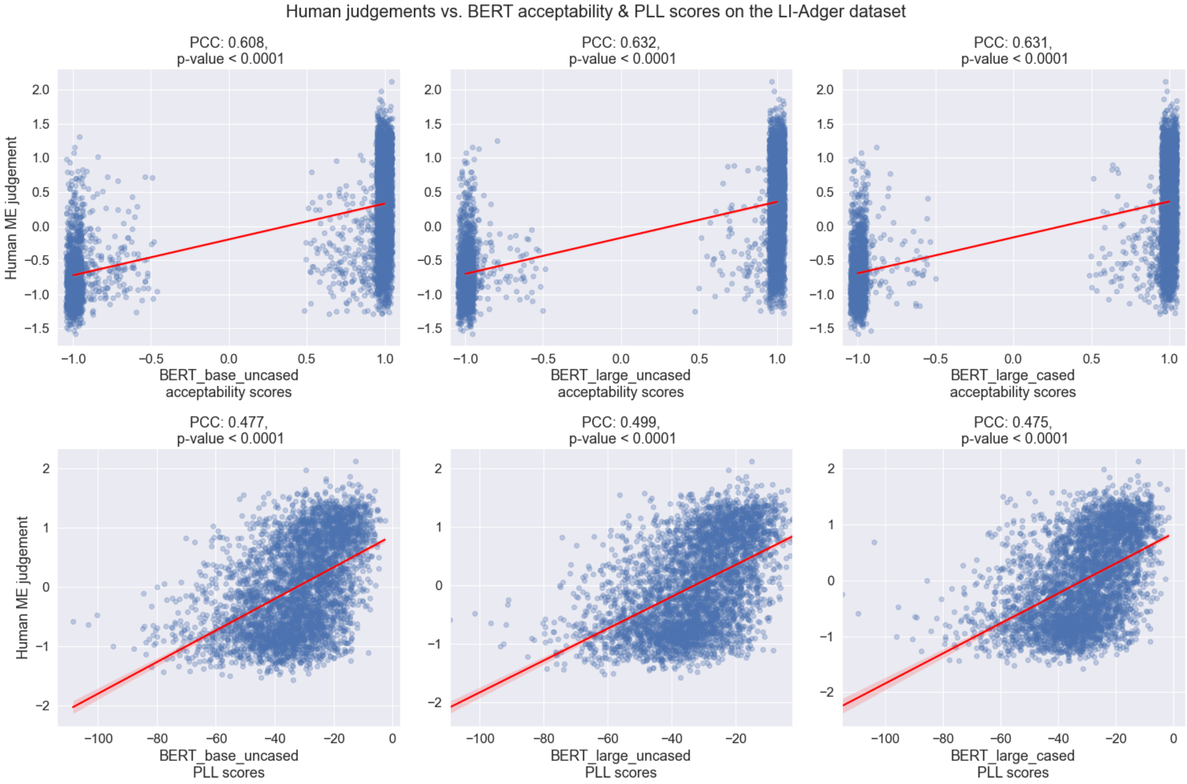


Figure 3-2: Scatterplot of human judgements (y-axis) vs. $\text{BERT}_{\text{CoLA}}$ acceptability scores, & $\text{BERT}_{\text{MLM}}$ PLL scores from all three BERT models for each sentence in the LI-Adger dataset with best-fit line in red. We add a jitter of 0.05 along the x-axis and lower the alpha to 0.3 to highlight the density of the points.

If previous examples are on the mark, the fact that the PCCs for the $\mathrm{BERT}_{\mathrm{MLM}}$ PLL scores are, on average, around 0.15 points lower than the corresponding $\mathrm{BERT}_{\mathrm{CoLA}}$ acceptability scores is not indicative of performance on the ADC. At least now with the PLL scores, the sentences truly seem to line up on a gradient scale, and one that appears to roughly track the best-fit line much better than the acceptability scores. The next step is to calculate the PCCs for the Z-score transformed PLL deltas and add them to the correlation matrix in Figure 2-6 in order to see how the PCCs change according to the more gradient metric. We present in Figure 3-3 the updated correlation matrix comparing the baseline trigram model, $\mathrm{BERT}_{\mathrm{CoLA}}$ acceptability delta scores and the newly calculated $\mathrm{BERT}_{\mathrm{MLM}}$ PLL delta scores.

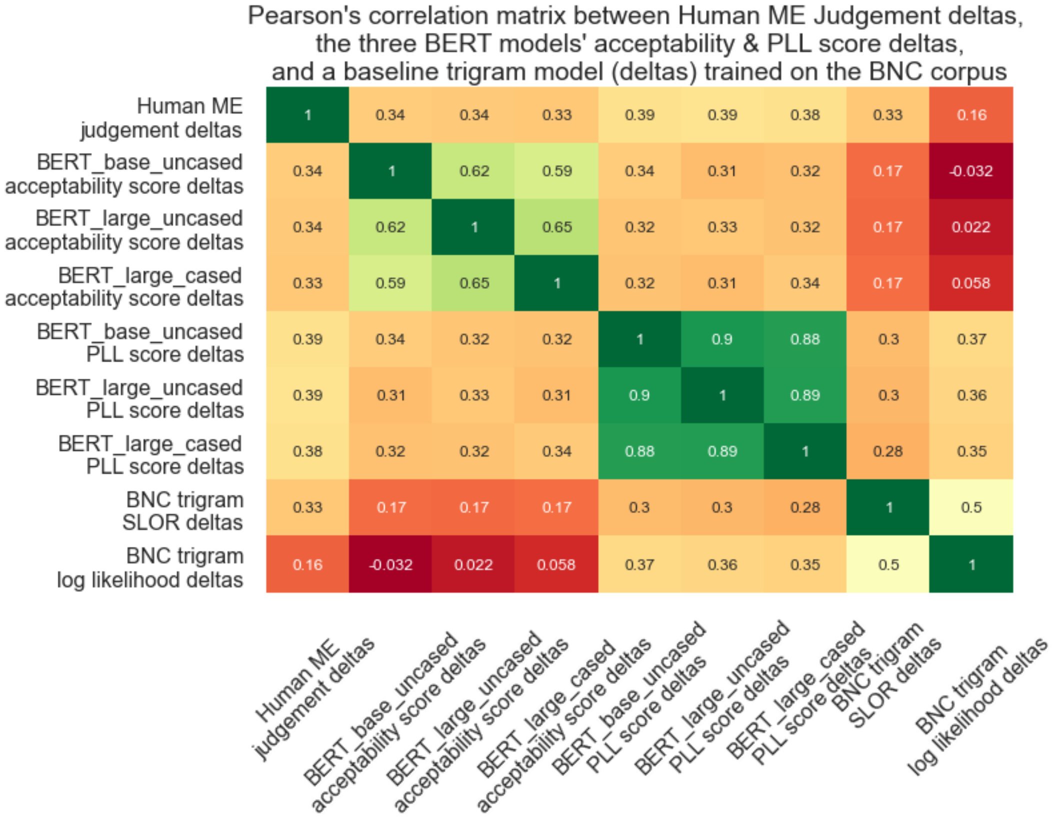


Figure 3-3: PCC matrix between human judgements and all three $\mathrm{BERT}_{\mathrm{CoLA}}$ & $\mathrm{BERT}_{\mathrm{MLM}}$. In addition we add the SLOR and log likelihood scores of a trigram model trained on the British National Corpus by Sprouse et al. 2018 for additional reference. All correlations shown have a $p < 0.0001$.

Encouragingly, we see a similar scenario to that of the trigram's SLOR score deltas and the three $\text{BERT}_{\text{CoLA}}$s' acceptability score deltas discussed in Section 2.5.1. Although the $\text{BERT}_{\text{CoLA}}$s' acceptability scores at the individual sentence level obtain much higher PCCs with the human judgements on the LI-Adger dataset, that advantage disappears when calculating the PCCs between the human judgement deltas and the acceptability score deltas. We see in Figure 3-3 that the PLL score deltas overtake the acceptability score deltas, although only by a small margin. For completeness, we plot in Figure 3-4 once more the correlation graphs in Figure 2-7 but adding the PLL score deltas to the comparison.

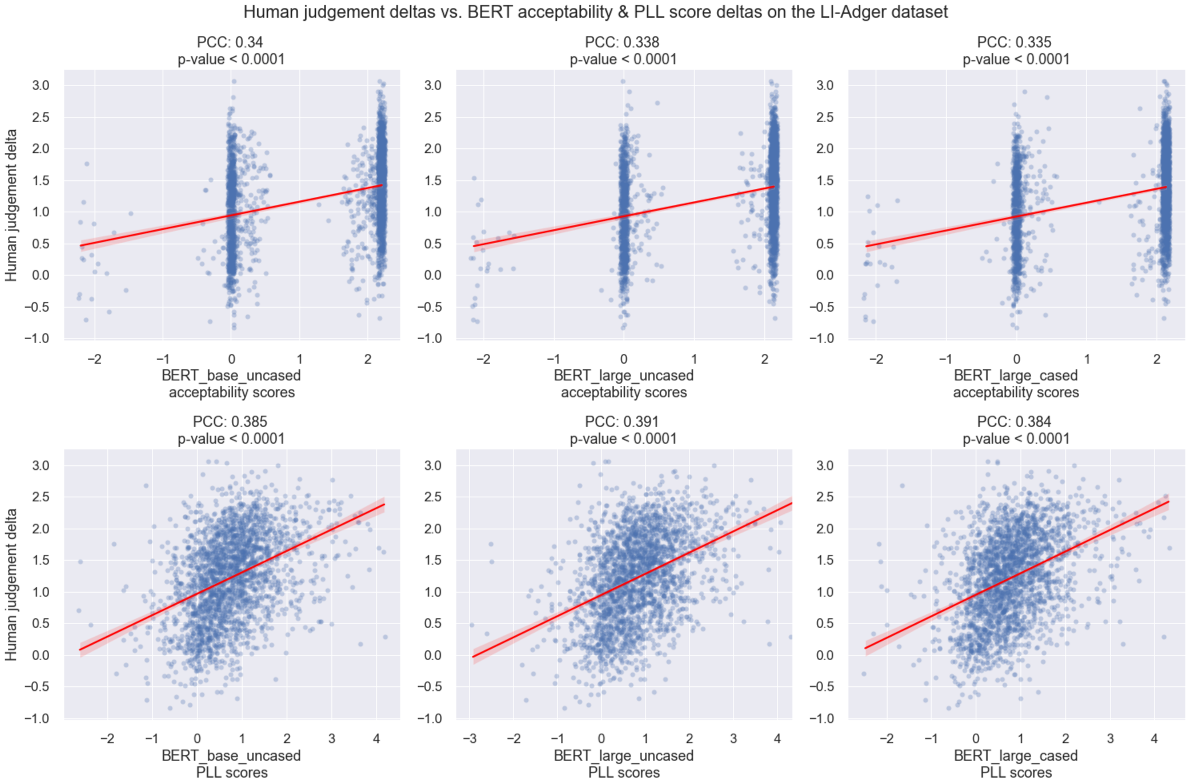


Figure 3-4: Scatterplot of human judgement deltas (y-axis) vs. $\text{BERT}_{\text{CoLA}}$ acceptability score delta & $\text{BERT}_{\text{MLM}}$ PLL delta for each minimal pair in the LI-Adger dataset with best-fit line in red. We add a jitter of 0.05 along the x-axis and lower the alpha to 0.3 to highlight the density of the points.

With all the preliminary correlations in place, we can set reasonable expectations for the $\text{BERT}_{\text{MLM}}$ models' performance under the ADC using the PLL deltas. We believe the gradience shown both at the sentence level PLL scores and the PLL deltas at the minimal pair level will yield better performance under the ADC at all three levels tested ($\delta = 0.5$, $\delta = 1.0$ and $\delta = 5.0$). How much better that performance is, remains to be seen.

## 3.3 BERT says best 2 out of 3 (ADC round 2)

Similar to Section 2.5.2, we apply the BLiMP Criterion and the ADC with $\delta = \{0.5, 1.0, 5.0\}$ in order to see how the ADC scales as it becomes less strict and generalizes to a form similar to the BLiMP Criterion. We report the performance of the three $\text{BERT}_{\text{MLM}}$ models using their PLL scores along with all the previously evaluated models in Table 3.1.

| **Model** | **BLiMP** | **ADC,** $\delta = 0.5$ | **ADC,** $\delta = 1.0$ | **ADC,** $\delta = 5.0$ |
|---|---|---|---|---|
| $\text{BERT}_{base-uncased;\text{MLM}}$ | 0.852 | 0.364 | 0.631 | 0.849 |
| $\text{BERT}_{large-uncased;\text{MLM}}$ | 0.866 | 0.378 | 0.658 | 0.859 |
| $\text{BERT}_{large-cased;\text{MLM}}$ | 0.871 | 0.376 | 0.661 | 0.868 |
| $\text{BERT}_{base-uncased;\text{CoLA}}$ | 0.915 | 0.286 | 0.538 | 0.902 |
| $\text{BERT}_{large-uncased;\text{CoLA}}$ | 0.917 | 0.311 | 0.564 | 0.907 |
| $\text{BERT}_{large-cased;\text{CoLA}}$ | 0.936 | 0.307 | 0.561 | 0.925 |
| $\text{trigram}_{SLOR}$ | 0.753 | 0.301 | 0.520 | 0.744 |
| $\text{trigram}_{log-prob}$ | 0.671 | 0.165 | 0.329 | 0.668 |

Table 3.1: Comparison between the models' BLiMP and ADC scores, using $\delta$=\{0.5, 1.0, 5.0\}. We include three $\text{BERT}_{\text{MLM}}$ models, three $\text{BERT}_{\text{CoLA}}$ models, as well as SLOR and log-likelihood scores from a trigram model trained on the British National Corpus by Sprouse et al. 2018

Reassuringly, the higher PCCs shown in Figures 3-3 and 3-4 by the $\text{BERT}_{\text{MLM}}$ models' PLL output translate well to better performance than the $\text{BERT}_{\text{CoLA}}$ models on the ADC for $\delta = 0.5$ and $\delta = 1.0$. However, when the distance between the models' output deltas and the human judgement deltas (Equation 1.2) is no longer considered by the ADC ($\delta = 5.0$), the $\text{BERT}_{\text{CoLA}}$ models outperform the $\text{BERT}_{\text{MLM}}$ models. This is likely due to the lack of gradience in the $\text{BERT}_{\text{CoLA}}$ models' acceptability output no longer being a determining factor in whether they evaluated a minimal pair correctly or not. This is presented in more detail below. First, as in Section 2.5.2, we inspect 4 minimal pairs where the $\mathrm{BERT}_{\mathrm{MLM}}$ models meet the BLiMP Criterion but not the ADC with a $\delta = 5.0$, shown in Table 3.2.

| **Minimal Pair**<br>**Top**: Acceptable \| **Bottom**: Unacceptable | Human<br>judgement | $\mathrm{BERT}_{\mathrm{MLM}}$<br>PLL |
|---|---|---|
| What is there a coupon for on the counter? | 0.085185 | 0.731265 |
| *What is a coupon for on the counter? | 0.134364 | 0.00874871 |
| What the runners believe is that they will win the race. | -0.23705 | 1.44651 |
| *What the runners believe is they will win the race. | -0.1155 | 1.1792 |
| I guessed he was married. | 0.111685 | 1.3683 |
| *I guessed he is married. | 0.588843 | 0.753377 |
| The announcer's introduction of Ted was humorous. | 0.659471 | 0.52595 |
| *The announcer's introduction of Ted's was humorous. | 0.748718 | 0.378315 |

Table 3.2: Four minimal pairs where all 3 models $\mathrm{BERT}_{\mathrm{MLM}}$ passed the BLiMP criterion but not the generalized ADC with $\delta = 5.0$. We report the $\mathrm{BERT}_{\mathrm{MLM}}$ PLL scores from $\mathrm{BERT}_{\mathrm{MLM}_{\mathrm{large-cased}}}$. The human judgement and $\mathrm{BERT}_{\mathrm{MLM}}$ PLL scores are already z-score transformed.

Although the three $\mathrm{BERT}_{\mathrm{MLM}}$ models failed to meet the ADC with $\delta = 5.0$ in Table 3.2, the reported PLL scores from $\mathrm{BERT}_{\mathrm{MLM}_{\mathrm{large-cased}}}$ immediately show how $\mathrm{BERT}_{\mathrm{MLM}}$'s PLL scores are much more gradient than the acceptability outputs from the $\mathrm{BERT}_{\mathrm{CoLA}}$ models. Let us inspect a few more example minimal pairs, but this time those where the $\mathrm{BERT}_{\mathrm{MLM}}$ models met the ADC with $\delta = 5.0$ but not the BLiMP Criterion, shown in Table 3.3.

What all the example minimal pairs in Table 3.3 have in common is that the human judgements disagree with the expert labels. Therefore, if we were to evaluate the human judgements themselves under the BLiMP Criterion, they would not pass either. However, reading the second and third minimal pair in Table 3.3 highlights precisely why it is preferable to rely on human judgements before the expert labels: those two minimal pairs are very close in acceptability values, and in fact read almost the same to native speakers. This additional resolution is lost when switching to cat-

| **Minimal Pair**<br>**Top**: Acceptable \| **Bottom**: Unacceptable | Human<br>judgement | $\text{BERT}_{\text{MLM}}$<br>PLL |
|---|---|---|
| We proved Amelia to the manager to be responsible. | -0.56008 | -1.77701 |
| *We proved to the manager Amelia to be responsible. | -0.13864 | -1.26881 |
| What did you give to whom? | 0.082244 | 0.899228 |
| *To whom did you give what? | 0.289657 | 0.930362 |
| There is likely to depart a train at midnight. | -0.53097 | -0.706574 |
| *There is likely a train to depart at midnight. | 0.177497 | 0.441587 |
| Jenny would accurately have calculated the results. | 0.345683 | -1.90655 |
| *Jenny accurately will calculate the results. | 0.501494 | -0.458686 |

Table 3.3: Four minimal pairs where all 3 $\text{BERT}_{\text{MLM}}$ models pass the generalized ADC with $\delta = 5.0$ but not the BLiMP criterion. We report the $\text{BERT}_{\text{MLM}}$ PLL scores from $\text{BERT}_{\text{MLM}_{\text{large-cased}}}$. The human judgement and $\text{BERT}_{\text{MLM}}$ PLL scores are already z-score transformed.

egorical expert labels such as those required by BLiMP. The $\text{BERT}_{\text{MLM}}$ PLL scores agree in monotonicity with the human judgements too; that is, the three $\text{BERT}_{\text{MLM}}$ models scored the unacceptable sentence in the minimal pair higher than the acceptable sentence, following the trend in the human judgements for the four examples in Table 3.3. Unfortunately, this is not a consistent trend. For example, upon inspection of a few examples where the $\text{BERT}_{\text{CoLA}}$ models meet the generalized ADC ($\delta = 5.0$) but not the $\text{BERT}_{\text{MLM}}$ models, we see the added gradience of the PLL scores alone is not enough. The four examples in Table 3.4 show the $\text{BERT}_{\text{CoLA}}$ models follow the monotonicity of the human judgements, whilst the $\text{BERT}_{\text{MLM}}$ models flip which sentence is the more acceptable of the pair. Under $\delta = 5.0$, the lack of gradience no longer affects the $\text{BERT}_{\text{CoLA}}$ models, thus allowing them to finally score higher than the $\text{BERT}_{\text{MLM}}$ models, which they were unable to do for $\delta = 0.5$ nor $\delta = 1.0$. However, there are cases where, even with $\delta = 5.0$, $\text{BERT}_{\text{MLM}}$ scores minimal pairs correctly that $\text{BERT}_{\text{CoLA}}$ is unable to account for, show in Table 3.5.

| **Minimal Pair** | Human | $\text{BERT}_{\text{CoLA}}$ | $\text{BERT}_{\text{MLM}}$ |
|---|---|---|---|
| **Top**: Acceptable \| **Bottom**: Unacceptable | judgement | acceptability | PLL |
| The book was written truthfully. | 1.30085 | 0.732642 | 1.00074 |
| *The book writes truthfully. | -0.40842 | -1.40058 | 1.03895 |
| It stormed suddenly. | 0.548385 | -1.39038 | 0.506643 |
| *It suddenly stormed. | 0.451832 | -1.39582 | 0.785855 |
| How few people were there at the rally? | 0.371244 | 0.732758 | 0.989334 |
| *How few people there were at the rally? | -0.01639 | 0.732748 | 1.02265 |
| The IRS denied Lilly her refund. | 1.40233 | 0.732951 | -1.22485 |
| *The IRS denied her refund to Lilly. | -0.47087 | 0.612725 | -0.966726 |

Table 3.4: Four minimal pairs where all $\text{BERT}_{\text{CoLA}}$ models meet the ADC with $\delta = 5.0$ but the $\text{BERT}_{\text{MLM}}$ models do not. We report the scores from $\text{BERT}_{large-cased}$. The human judgement, acceptability, and PLL scores are already z-score transformed.

| **Minimal Pair** | Human | $\text{BERT}_{\text{CoLA}}$ | $\text{BERT}_{\text{MLM}}$ |
|---|---|---|---|
| **Top**: Acceptable \| **Bottom**: Unacceptable | judgement | acceptability | PLL |
| Toby said to Sally to take care of herself. | 0.188154 | -1.39838 | 0.540271 |
| *Toby said to Sally to take care of himself. | -0.488670 | 0.608852 | 0.348782 |
| I predicted she was guilty. | 0.703876 | 0.730575 | 0.615354 |
| *I predicted she may be guilty. | 0.530117 | 0.732155 | 0.18414 |
| The ice melted quickly on the table. | 1.459752 | 0.729108 | 1.34043 |
| *The ice quickly melted on the table. | 0.618948 | 0.730474 | 1.18234 |
| A lawyer smarter than my brother... | 0.596167 | 0.733196 | -0.188386 |
| *A smarter lawyer than my brother... | 0.782106 | 0.733102 | 0.123958 |

Table 3.5: Four minimal pairs where all $\text{BERT}_{\text{MLM}}$ models meet the ADC with $\delta = 5.0$ but not the $\text{BERT}_{\text{CoLA}}$ models. We report the scores from $\text{BERT}_{large-cased}$. The human judgement, acceptability, and PLL scores are already z-score transformed.

Our analyses of the $\text{BERT}_{\text{MLM}}$ models continue by conducting the same trigram-overlap analyses we did with the BERT models fine-tuned using CoLA. *A priori* there likely is a much larger overlap between the trigram model's SLOR predictions and $\text{BERT}_{\text{MLM}}$'s PLL predictions by the very nature of the MLM procedure. $\text{BERT}_{\text{MLM}}$ uses the surrounding context of a masked token to predict a probability distribution $P(w_j|w_0, ..., w_{j-1}, w_{j+1}, ...w_n)$ (Equation 3.2) for each token in a given sequence of words $s_i$. This is, in principle, similar to a trigram using the preceding context of a word to predict its likelihood $P(w_j|w_{j-3}, w_{j-2}, w_{j-1})$. We say *in principle* because the MLM and $N$-gram outputs and training regimes can be framed in terms of predicting individual tokens given their context, with the caveat that the trigram is severely handicapped in terms of the machinery it has at its disposal. However, $\text{BERT}_{\text{CoLA}}$ outputs a probability distribution over a finite set of categorical labels when given the entire sequence $s_i$ as input, so it is much farther removed from the computations a trigram performs relative to $\text{BERT}_{\text{MLM}}$. We assess the extent of this similarity by calculating the ADC for $\delta = 0.5$ and $\delta = 1.0$ once again. Then for each value of $\delta$, we create a new dataset by subtracting the minimal pairs from LI-Adger dataset that were correctly predicted by the trigram's SLOR scores. We finalize by recalculating the $\text{BERT}_{\text{MLM}}$ models' ADC scores using the reduced datasets. The results for the procedure are presented in Table 3.6.

| **Model** | **ADC, $\delta = 0.5$** | | | **ADC, $\delta = 1.0$** | | |
|---|---|---|---|---|---|---|
| **$\text{BERT}_{\text{MLM}}$** | **Original** | **Reduced** | **Overlap** | **Original** | **Reduced** | **Overlap** |
| base-uncased | 0.364 | 0.323 | 13.87% | 0.631 | 0.553 | 36.58% |
| large-uncased | 0.378 | 0.333 | 14.50% | 0.658 | 0.586 | 37.51% |
| large-cased | 0.376 | 0.335 | 14.21% | 0.661 | 0.589 | 37.80% |

Table 3.6: $\text{BERT}_{\text{MLM}}$ models' performance on the ADC with $\delta$=0,5,1.0 before (Original) and after (Reduced) removing all minimal pairs for which the ADC Criterion was met by the trigram baseline model trained on the BNC corpus. The Overlap columns display the percentage of minimal pairs that both the $\text{BERT}_{\text{MLM}}$ model and the trigram baseline pass.

A difference that becomes immediately apparent is the drop in total performance across the board for all three $\text{BERT}_{\text{MLM}}$ models both for $\delta = 0.5$ and $\delta = 1.0$ once the minimal pairs correctly scored by the trigram are removed from the dataset. Unsurprisingly, the percentage of the minimal pairs scored correctly by both $\text{BERT}_{\text{MLM}}$ and the trigram SLOR model correctly is on average 14.2% for $\delta = 0.5$ (up from 8.8% with $\text{BERT}_{\text{CoLA}}$) and 37.3% for $\delta = 1.0$ (up from 28.8% with $\text{BERT}_{\text{CoLA}}$).

Finally, this brings us to the question posed at the end of Chapter 2: What is lost (or gained) in terms of performance under the ADC by switching from $\text{BERT}_{\text{CoLA}}$ acceptability scores to $\text{BERT}_{\text{MLM}}$ PLL scores? We present in Table 3.7 the performance of $\text{BERT}_{\text{MLM}}$ under the ADC as shown previously, but add an additional column containing the percentage of overlap between $\text{BERT}_{\text{MLM}}$ and $\text{BERT}_{\text{CoLA}}$, i.e. what percentage of the minimal pairs in the LI-Adger dataset was classified correctly by both BERT models. We carry out this calculation for $\delta = 0.5$ and $\delta = 1.0$ as with the trigrams assessment because these are where ADC enforces the gradient aspect of the minimal pairs, not just their monotonicity.

| **Model** | **ADC,** $\delta = 0.5$ | | **ADC,** $\delta = 1.0$ | |
|---|---|---|---|---|
| $\textbf{BERT}_{\text{MLM}}$ | **Score** | **Overlap** | **Score** | **Overlap** |
| base-uncased | 0.364 | 9.05% | 0.631 | 34.25% |
| large-uncased | 0.378 | 9.73% | 0.658 | 37.29% |
| large-cased | 0.376 | 9.89% | 0.661 | 37.04% |

Table 3.7: The percentage of minimal pairs that both $\text{BERT}_{\text{MLM}}$ and $\text{BERT}_{\text{CoLA}}$ passed, as well as the ADC scores with $\delta$={0.5, 1.0} of the $\text{BERT}_{\text{MLM}}$ models for reference.

The overlap between $\text{BERT}_{\text{MLM}}$ and $\text{BERT}_{\text{CoLA}}$, although substantial, only accounts for roughly one third of the minimal pairs either of the two models scored correctly under the stricter $\delta = 0.5$ measure. Relaxing the ADC to $\delta = 1.0$ allows the overlap between the two models to account for roughly half of the minimal pairs they each scored correctly. However, this supports the interpretation we set out to test at the beginning of this chapter: $\text{BERT}_{\text{MLM}}$ and $\text{BERT}_{\text{CoLA}}$ behave almost completely differently, even demonstrating different KoL despite relying on the same pretrained BERT model. Training an additional classifier on top of BERT, as we did with CoLA, further obfuscates any information to be gained from conducting an input-output analysis on the model.

# Contributions

This thesis has reviewed current empirical methods of assessing the KoL of LMs, and found that to date there exists no test of KoL that is comprehensive in its coverage of linguistic phenomena, is backed by attested and replicable human judgement data, and tests LMs' ability to track different linguistic phenomena across the full range of the acceptability gradient. This thesis addresses this gap by proposing the three necessary components needed to construct such a comprehensive test of KoL.

First, this thesis presents the LI-Adger dataset, a collection of 150 pairwise phenomena collected by Sprouse et al. (2013) from Linguistic Inquiry (LI) 2001-2010, and an additional 105 multi-condition phenomena collected by Sprouse & Almeida (2012) from an exhaustive selection of 219 sentence types from Adger's (2003) *Core Syntax* textbook. The phenomena represented in the LI-Adger dataset far exceed the coverage of the most recent datasets published to date for the purposes of testing the KoL in LMs. The Corpus of Linguistic Acceptability (CoLA; Warstadt & Bowman 2019) and The Benchmark of Linguistic Minimal Pairs for English (BLiMP; Warstadt et al. 2020).

This thesis supports the LI-Adger dataset with statistically powerful, replicable and validated human Magnitude Estimate (ME) data collected by Sprouse et al (2013) and Sprouse & Almeida (2012). The data accompanying the LI dataset boasts a 95 percent ±5 minimum replication rate (Sprouse et al. 2013), whereas the ME data in the Adger dataset increases the minimum replication rate to over 97% (Sprouse & Almeida 2012).. Additionally, Sprouse et al. (2018) determined the statistical power of both the LI and Adger datasets to meet the 80% power threshold for the detection of False Negatives.

This dataset and accompanying human judgements then become the gold standard in terms of coverage and reliability. In order to make full use of this dataset, this thesis proposes the Acceptability Delta Criterion, a metric that tests LMs for Human KoL by enforcing the gradience of acceptability and requiring LMs to track the validated human judgements through the gradient spectrum as the acceptability values change across minimal pairs. We demonstrate further that adopting a functionally categorical view of acceptability leads to an unstable BERT model when fine-tuned with CoLA achieving 94% correct classification of the minimal pairs in the LI-Adger dataset; while a trigram model trained on the British National Corpus (BNC) by Sprouse et al. 2018 achieves (75%). These results imply that either trigram models are able to account for 75% of the phenomena in Generative grammar, or, alternatively, that treating acceptability as a categorical metric leads to a high false positive rate in KoL tests. Accordingly, the ADC with a strict $\delta = 0.5$ determined that neither BERT, whose predictions were nearly all categorical, and the trigram model both only correctly accounted for roughly 30% of the minimal pairs in the LI-Adger dataset. Using the defaul BERT models with gradient *pseudo log-likelihood* (PLL) outputs increased its score to (37%), further demonstrating the need for gradience in order to meet the ADC.

The three main contributions of this thesis when used together create the most comprehensive test of Human KoL for LMs currently available. With further ongoing work, the test will also allow us to see a fine-grained analysis of which phenomena a LM was able to account for in its output and how well it predicted the acceptability judgements around them. It is to be hoped that researchers will adopt the LI-Adger dataset for its coverage of Generative grammar and rely on the human judgements as the ground-truth labels that LMs are expected to approximate, and, beyond that, the ADC.

# Bibliography

Adger, D. (2003). *Core syntax: A minimalist approach*, volume 20. Oxford University Press Oxford.

Chomsky, N. (1965). Aspects of the theory of syntax. *Cambridge, MA: MIT Press*, (1977), 71–132.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2020). When bert forgets how to pos: Amnesic probing of linguistic properties and mlm predictions. *arXiv preprint arXiv:2006.00995*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Marvin, R. & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031.*

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

McCoy, R. T., Min, J., & Linzen, T. (2019). Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969.*

Pauls, A. & Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 959–968).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2019). Pseudolikelihood reranking with masked language models. *CoRR*, *abs/1910.14659.*

Schütze, C. T. & Sprouse, J. (2013). Judgement data. In R. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 27–51). Cambridge University Press.

Shin, J., Lee, Y., & Jung, K. (2019). Effective sentence scoring method using bert for speech recognition. In *Asian Conference on Machine Learning*, (pp. 1081–1093).

Sprouse, J. (2011). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, 274–288.

Sprouse, J. & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, *48*(3), 609–652.

Sprouse, J. & Almeida, D. (2013). *The Role of Experimental Syntax in an Integrated Cognitive Science of Language.*

Sprouse, J. & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, *2*(1), 1.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, *134*, 219–248.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., & Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, *35*(3), 575–599.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? *CoRR*, *abs/1905.05583*.

T Schütze, C. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Wang, A. & Cho, K. (2019). Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Warstadt, A. & Bowman, S. R. (2019). Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

Warstadt, A. & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, *7*, 625–641.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (pp. 38–45)., Online. Association for Computational Linguistics.