The Acceptability Delta Criterion: Testing Knowledge of Language using the Gradience of Sentence Acceptability

Héctor Javier Vázquez Martínez
Massachusetts Institute of Technology
hjvm@mit.edu

Abstract

Any test that promises to assess Human Knowledge of Language (KoL) for any statistically-based Language Model (LM) must meet three requirements: (1) comprehensive coverage of linguistic phenomena; (2) replicable and statistically-vetted human judgement data; and (3) test the LM's ability to track the gradience of sentence acceptability. To this end, we propose here the LI-Adger dataset: a comprehensive collection of 519 sentence types (4177 sentences) spanning the field of current generative linguistics, accompanied by attested and replicable human acceptability judgements (Sprouse and Almeida, 2012; Sprouse et al., 2013; Sprouse and Almeida, 2017). Finally, we posit the Acceptability Delta Criterion (ADC), an evaluation metric that tests how well a LM can track changes in human acceptability judgements across minimal pairs instead of testing whether the LM assigned a greater likelihood to the expert-labeled acceptable sequence of a minimal pair ($S_1 > S_2$). We benchmark six different BERT (Devlin et al., 2018) models and a baseline trigram model with the ADC. Although the best performing BERT model scores 94%, and the trigram scores 75% classification accuracy under the traditional metric, performance drops precipitously to 38% for BERT and 30% for the trigram model under the ADC. Adopting the ADC reveals how much harder it is for LMs to track the gradience of acceptability across minimal pairs. With this work, we propose and provide the three necessary requirements for a comprehensive linguistic analysis and test of the apparently Human KoL exhibited by LMs that we believe is currently missing in the field of Computational Linguistics.

1 Introduction

Assessing Human Knowledge of Language (KoL) in statistically-based Language Models (LMs) generally involves assuming some fundamental property or computation occurring in the Human Language Faculty and arguing that an often poorly understood, statistical, and typically connectionist model, also encodes that property or uses that 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 any 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 have subsequently been used to build the linguistic theories that attempt to characterize and explain Human KoL.

We take concrete steps toward designing such a test of KoL for LMs by positing the necessary components required to build upon the same foundation of empirical data as in the field of generative linguistics. First, we posit a gold standard in empirical data: the LI-Adger dataset, a collection of linguistic phenomena representative of the field of linguistics, accompanied by human acceptability judgements in the form of Magnitude Estimation (ME) data, all collected by Sprouse and Almeida 2012 and Sprouse et al. 2013. Altogether, the dataset has an

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 479–495
Online, November 11, 2021.

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 and 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 attested empirical phenomena observed in the LI-Adger dataset (Sprouse and 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 attested continuous human judgements as the ground-truth labels that LMs must approximate in order to demonstrate KoL. Finally, we use two distinct Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2018) models: the out-of-the-box BERT trained using the Masked Language Modeling (MLM) objective (BERT_{MLM}), and $\text{BERT}_{\text{CoLA}}$, which is fine-tuned using the Corpus of Linguistic Acceptability (CoLA; Warstadt et al. 2019). As an additional baseline check, we include a trigram model trained using the British National Corpus (BNC) to test whether the ADC would constitute a meaningful advantage in statistical significance (fewer Type 1 errors) over the traditional method of comparing probabilities over isolated, minimally-differing sequences of words. Henceforth we refer to the traditional metric as the BLiMP criterion[1](Warstadt et al., 2020), that will be contrasted against the new measure proposed here, the Acceptability Delta Criterion (ADC).

Under the BLiMP criterion, $\text{BERT}_{\text{CoLA}}$, correctly evaluates 2213 out of 2365 ($\sim$94%) minimal pairs in the LI-Adger dataset; that is, for those 2213 minimal pairs, $\text{BERT}_{\text{CoLA}}$ gives a higher score to the sentence in the minimal pair deemed by experts to be the *acceptable* one of the pair. To put the performance of $\text{BERT}_{\text{CoLA}}$ into perspective, the trigram model's output, when normalized for sequence length using the Syntactic Log-Odds Ratio (SLOR; Pauls and Klein, 2012; Lau et al., 2017) is able to correctly evaluate 1781 out of 2365 ($\sim$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 tests with low levels of statistical significance (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 generative linguistics.

Precisely this dilemma is what the ADC is meant to solve. Evaluating the models on how well they track the gradience of acceptability by adopting the ADC at its strictest level (with $\delta = 0.5$, where δ indicates the number of standard deviation units a model's scores are allowed to deviate from the corresponding human judgements) leads to a notable drop in performance, particularly for the trigram model. $\text{BERT}_{\text{CoLA}}$ only correctly scores 726 out of 2365 ($\sim$31%) minimal pairs, whereas the trigram model with SLOR correctly scores 712 out of 2365 ($\sim$30%). The ADC is therefore a much harder task, because the difference in likelihood acceptability score between the two sentences in each minimal pair is what is being evaluated.

2 Related Work

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[2], has made it a popular endeavor to assess the potential KoL encoded in the learned representations of the language models and how that KoL may be contributing to their performance.

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 the system tend to focus on probing their weak generative capacity: i.e, 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. Contrast this with testing the LM's strong generative capacity by evaluating whether it has assigned the correct syntactic structure, or series of candidate syntactic structures in

[1]Although this has been the traditional metric to test KoL in the literature (Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; Warstadt and Bowman, 2020; among others), we name this metric with the same acronym as that of the Benchmark for Linguistic Minimal Pairs by Warstadt et al. (2020) mainly for conciseness and ease of use.

[2]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.

the case of ambiguous or homonymous sequences, to a sequence of words (Chomsky, 1956).

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.

While the tradition of using minimal pairs in linguistics dates back nearly 100 years (e.g. Bloomfield 1933, among many others), the concept of using minimal pairs to evaluate NN models is not entirely new either (Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; to name a few). However, the creators of BLiMP take the idea to a much larger scale and propose a single metric for evaluation, which we have named the BLiMP criterion out of convenience. 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 dataset 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.

Because the BLiMP dataset relies on templates, the choice and design of the linguistic phenomena represented in the data is inherently limited, and therefore unrepresentative of syntax or linguistics more broadly. Secondly, the templated approach produces semantically implausible sentences, such as *Sam ran around some glaciers* as noted by Warstadt et al. (2020) in their paper. Sprouse et al. (2018) have shown that semantic implausibility is a very strong confounding factor when eliciting human acceptability judgements, even in a Forced Choice (FC) task such as the one used to collect the human judgement data in BLiMP. Finally, the BLiMP criterion treats sentence acceptability as a functionally categorical phenomenon, which we assert to be gradient by nature.

Our proposed test of Human KoL in LMs takes this minimal pair approach to the next level by addressing the three gaps we have outlined above. We (1) expand linguistic coverage to be representative of the field with the LI-Adger dataset, (2) further control for confounding factors such as semantic implausibility when eliciting human acceptability judgements by using the examples constructed by hand by Sprouse and Almeida 2012 & Sprouse et al. 2013, and (3) enforce the gradient nature of sentence acceptability judgements with the ADC.

3 Why BERT is the benchmark.

We chose BERT as the model to test due to the growing body of research attributing ever greater Human KoL to BERT. Warstadt and Bowman (2019) have shown high Matthews Correlation Coefficient (MCC) 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. They 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. Finally, Salazar et al. (2019) used the raw pseudo-log-likelihood (PLL; Wang and Cho, 2019; Shin et al., 2019; Salazar et al., 2019) scores from the out-of-the-box BERT_{MLM} 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%. We take the information provided here and the overarching body of research surrounding BERT [3] as the baseline level of performance to be expected from the model. We hypothesize that BERT is capable of exhibiting reasonable levels of gradience when it calculates sequence likelihoods or acceptability scores across minimal pairs when compared to the human ME data.

[3] 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)

4 The LI-Adger dataset is the test set.

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 phenomenon includes 8 hand-constructed, semantically plausible and 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.

The second set of sentences is an exhaustive selection of 219 sentence types from Adger (2003)'s *Core Syntax* textbook (198 directly from the textbook + 21 as additional controls) that form 105 multi-condition phenomena collected by Sprouse and Almeida (2012). Much like the LI dataset, 8 tokens of each sentence type were created by hand by the original authors such that the structural properties of the condition were maintained but the lexical items varied.

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, when properly lexically matched. We include an explicit example of the arrangement with a multi-condition phenomenon from Chapter 8 (*Functional Categories III*) of the *Core Syntax* textbook in Appendix A.

The Adger dataset, in virtue of being sampled from the *Core Syntax* textbook, that constructs a theory of syntax from the ground up on the basis of key examples, can be taken to have a reasonably good coverage of the field of syntax. We augment this with the LI dataset, 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* textbook and *LI*2001-2010 are representative of the data in the field, so is the LI-Adger dataset. (Sprouse and Almeida, 2012; Sprouse et al., 2013).

5 The Human Magnitude Estimation (ME) data are the ground truth labels.

Just as important as the coverage of linguistic phenomena represented in the LI-Adger dataset is the foundation of human judgement data built upon it. Sprouse and Almeida (2012) collected Magnitude Estimation (ME) and Yes-No (YN) 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 (standard frequentist 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 and Almeida, 2012; Schütze and Sprouse, 2013).

Sprouse et al. (2013) took those analyses even further with their sample of 148 pairwise phenomena from *LI*2001-2010. They collected data for the LI sentences using the 7-point Likert Scale (LS) task, ME and Forced Choice (FC) and vetted this dataset via 5 different statistical analyses (Descriptive directionality, one- and two-tailed null hypothesis tests, Bayes factor analysis and linear mixed effect models). They estimated a minimum replication rate for journal data of 95 percent ± 5 (Sprouse et al., 2013; Schütze and Sprouse, 2013.

Finally, Sprouse and Almeida (2017) sampled 50 pairwise phenomena from the 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 and Almeida, 2017; Schütze and Sprouse, 2013). Because 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 human ME data that accompany the LI-Adger dataset are therefore reliable,

replicable and statistically powerful. This collection of empirical data 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 because this statistically vetted empirical evidence would still remain to be accounted for.

6 The Acceptability Delta Criterion (ADC) is the loss function.

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. The reason why is precisely the gradient nature of acceptability: violations of certain linguistic phenomena elicit much stronger responses from the participants than others, as observed in the differences in ME scores across minimal pairs and across linguistic phenomena in the dataset. We include an example of such a difference in Appendix B.

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; Sprouse and Almeida, 2013; Schütze and Sprouse, 2013; T Schütze, 2016). 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. We include one example of this variability in intensity and how it is reflected in the human ME judgements in Appendix B.

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 whose change in human acceptability rating is nearly night and day, a language model with comparable KoL must output a similarly drastic change in acceptability rating across the same minimal pair.

To make this example more concrete: Suppose we have a language model L with output function f that takes in a sequence of words $\vec{x}_i$ and outputs a score y_i. The first step in computing 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 values, 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 of 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 and Sprouse, 2013).

Now that we have grounds for making the comparison[4] and a value for how acceptable the model L finds a sequence of words $\vec{x}_i$ 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

[4]We rely on the Z-score transformation as opposed to the log-transformation because the Z-score transformation is linear and therefore preserves the relationships that exist between the data. This transformation also has the added benefit that it assumes the data are continuous, which both the ME data and the LMs' likelihoods are. On the other hand, the log-transformation is generally not recommended for ME data because it introduces distortion to the data by nature of being nonlinear, and is too powerful for simple outlier removal (Sprouse, 2011; Schütze and Sprouse, 2013). The Z-score transformation is therefore the better candidate operation we can reasonably apply to both the human judgement data and the LMs' likelihood scores.

a reformulated BLiMP criterion in terms of the delta values, we may at last proceed to define the ADC. Let δ be a scalar value indicating the number of maximum allowed units of deviation between the human judgement delta Δh_i and the language model delta Δlm_i. Using this δ value, we consider the ADC to be met for the minimal pair m_i when the following two conditions are met:

$$\text{sign}(\Delta h_i) = \text{sign}(\Delta lm_i) \tag{1}$$

$$|\Delta h_i - \Delta lm_i| < \delta \tag{2}$$

The δ parameter in Equation 2 can be adjusted to allow for larger or smaller amounts of deviation between the human and LM acceptability deltas. If δ is set to a large number, the ADC functionally becomes the BLiMP criterion because it is dominated by Equation 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 δ 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. We include a concrete example of the ADC in action in Appendix C. Because the δ parameter determines how closely the LM's acceptability deltas must track the human judgement deltas, we adopt a first approximation of $\delta = 0.5$, which we believe to be the proper value of δ in order to test for the gradience of acceptability. Additionally, we adopt a value of $\delta = 1$ in order to show how performance under the ADC changes as the δ parameter is relaxed. We include results for both values of δ in the case study to be presented here.

7 The KoL experiment brings everything together.

With the LI-Adger dataset as the test set, the human ME data as the ground truth labels and the ADC as the loss function, this KoL experiment now only requires the model to test.

In order to ensure we observe the best performance by BERT in our benchmarks, we adopt two parallel approaches to the tests: a syntactic probing approach and a MLM approach with a variant of the Cloze task (Taylor, 1953). For the probing approach, we train a linear softmax classifier (probe) on top of BERT's hidden layer outputs using the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019). 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 $\text{BERT}_{\text{base-uncased}}$, 20 random seeds of $\text{BERT}_{\text{large-uncased}}$, and 20 random seeds of $\text{BERT}_{\text{large-cased}}$. For further control we attempt to replicate the mean MCC scores per linguistic phenomenon obtained by Warstadt and Bowman (2019) on their later published CoLA Analysis set. For the purposes of transparency and the replicability of our own experiment, we include the results of our models' performance on the CoLA Analysis set in Appendix D. For each of the three $\text{BERT}_{\text{CoLA}}$ we selected the random restart that yielded the highest MCC score on the CoLA test set as the model to test using the ADC. Finally, we need to adapt the categorical output of the fine-tuned BERT models (c_j where $j \in \{1, 0\}$) to a more gradient form in order to ensure the models have the best chance of tracking sentences through the acceptability gradient. We rely on Sun et al. (2019)'s formulation of the probabilities[5] of the categorical labels as the softmax of BERT's final hidden layer output h_i:

$$P(c_j|h_i) = \text{softmax}(Wh_i) \tag{3}$$

Therefore, the output of the fine-tuned BERT models can be defined as:

$$out_i = \underset{c_j \in \{0,1\}}{\arg\max} \left[P(c_j|h_i) \right] \tag{4}$$

Out of convenience, we will switch the categorical labels from {0, 1} to {-1, 1}. This allows us to multiply the model's confidence in a particular label ($P(c_j|h_i)$) by the label itself, written explicitly below:

$$out_i = \underset{c_j \in \{-1,+1\}}{\arg\max} \left[P(c_j|h_i) \right] * \max_{c_j \in \{-1,+1\}} \left[P(c_j|h_i) \right] \tag{5}$$

[5] 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. Because 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 and Almeida, 2012), we will loosely interpret the output of the softmax in Equation 3 as the model's *confidence* in that particular label.

With this formulation, we can easily retrieve both the predicted categorical label ($\text{sign}(out_i)$) and the model's *confidence* in that label ($|out_i|$).

Reformulations aside, because probing relies on training an additional classifier on top of the latent, albeit poorly understood, representations of neural LMs, it is extremely difficult to control for confounding 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.

We mitigate these concerns with the second, parallel approach to our experiment: We use the publicly available model checkpoints for $\text{BERT}_{\text{base-uncased}}$, $\text{BERT}_{\text{large-uncased}}$, and $\text{BERT}_{\text{large-cased}}$, which were originally trained using the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives (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 dataset under the traditional metric, as well as on other natural language benchmarks, potentially due to BERT's ability to better leverage the left and right context of each masked token when calculating its likelihood. This altogether strongly favors PLL scores as the ideal metric to test the out-of-the-box BERT models with the ADC.

Now we have 6 distinct BERT models to test: the three best performing random restarts of $\text{BERT}_{\text{base-uncased}}$, $\text{BERT}_{\text{large-uncased}}$, and $\text{BERT}_{\text{large-cased}}$ when fine-tuned using CoLA (collectively $\text{BERT}_{\text{CoLA}}$), and the three unadulterated, publicly available model checkpoints for which we will calculate PLL scores (collectively BERT_{MLM})[6]. To this assortment, we add the trigram model trained by Sprouse et al. (2018) using the British National Corpus (BNC). We will use both its raw negative log-likelihood scores and Syntactic Log-Odds Ration (SLOR; Pauls and Klein, 2012, Lau et al., 2017) scores.

Finally, we test four metrics: the original BLiMP criterion, and the ADC with δ values of 0.5, 1.0 and 5.0. The goal with the last $\delta = 5.0$ is to observe how model performance approaches that of the BLiMP criterion when the ADC is dominated by Equation 1. The results of these tests are presented in Table 1.

8 Results

We summarize the results of our proposed KoL benchmark with the BERT and trigram models in Table 1. We observe extremely good performance across the board under the BLiMP criterion. However, employing the ADC with the stricter δ values ($\delta = 0.5$, $\delta = 1.0$) leads to a precipitous drop in performance. When the distance between the models' output deltas and the human judgement deltas (Equation 2) is no longer considered by the ADC ($\delta = 5.0$), all the models' performances come very close to the performance numbers they obtained under the BLiMP criterion because the models are no longer expected to track the human judgement deltas in magnitude, only monotonicity. The minute differences in performance are primarily accounted for by disagreements between the expert labels and the human judgements in minimal pairs where the human judgements favored the sentence expert-labeled as unacceptable. We include 4 example minimal pairs where the $\text{BERT}_{\text{CoLA}}$ models scored correctly under the BLiMP criterion, but not under the ADC with $\delta = 5.0$ in Appendix E. This is a very encouraging result, suggesting that the ADC is in fact a generalization of the BLiMP criterion, with the difference being explained by the shift to human judgements as the ground-truth labels.

To further compare how each of the models' acceptability deltas (Δlm_i) behave with respect to the ground truth (Δh_i), we plot in Figure 1 a heatmap of the Pearson's Correlation Coefficient (PCC) matrix for all 8 models and the human judgements.

The models that were best able to track the human judgement deltas through the full spectrum of acceptability were the BERT_{MLM} mod-

[6]We make a note here to acknowledge that, due to computational limitations at the time of writing and experimentation, we were unable to add models much larger than $\text{BERT}_{\text{large-cased}}$, such as GPT-2 or GPT-3 to the case study presented here.

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
trigram_{SLOR}	0.753	0.301	0.520	0.744
$\text{trigram}_{log-prob}$	0.671	0.165	0.329	0.668

Table 1: Comparison between the models' BLiMP criterion and ADC scores, using δ={0.5, 1.0, 5.0}. We include three BERT_{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

els with the PLL metric at a moderate positive correlation of 0.38 for $\text{BERT}_{\text{large-cased}}$ and 0.39 for $\text{BERT}_{\text{base-uncased}}$ and $\text{BERT}_{\text{large-uncased}}$. The BERT models fine-tuned using CoLA were only very slightly (0.01) better correlated with the human judgement deltas than the trigram model's SLOR score deltas. Functionally, the 7 models (excluding the trigram's raw log-likelihood scores) perform very closely when tracking the human judgement deltas at the minimal pair level. As an additional resource, we include example minimal pairs of where the $\text{BERT}_{\text{CoLA}}$ models and the trigram model disagree in Appendix F, and plot all six BERT models' deltas (Δlm_i) against the human deltas (Δh_i) in Appendix G.

9 Discussion

The first column of Table 1 (BLiMP) is what first prompted us to formulate the ADC. In particular, note that the performance of the basic trigram model trained using the BNC is at 75.3% when using its SLOR scores. Having discussed how comprehensive the LI-Adger dataset is, this leaves us with the following troubling dichotomy: Either the trigram model encodes approximately 3 quarters of the empirical linguistic phenomena in the LI-Adger dataset, or it is too easy to best the BLiMP criterion with word cooccurrences alone. The dichotomy suggests that the BLiMP criterion has a low significance level as a metric when assessing KoL and thus commits a higher rate of Type 1 errors than would be advisable in order for experiments that rely on it to be conclusive. We believe this to be due to treating sentence acceptability as a functionally categorical metric when it is inherently gradient; we remedy this by testing for the gradience of acceptability using the ADC.

However, the ADC does raise the concern of whether we are fundamentally changing the results by adopting human acceptability judgements as ground-truth instead of the expert labels. We evaluated all 8 models with ever greater δ values in order to observe how much this should concern us, and as observed in the rightmost column of Table 1 (ADC, $\delta = 5.0$), that concern is minimal. The performance of each of the models very closely approximated their original performance under the traditional metric (leftmost column; BLiMP). Adopting the human judgements as the ground truth labels therefore does not inherently cause a steep drop in performance. Furthermore, we believe it a counterproductive effort to compute human performance as compared to the expert labels, or to have LMs competing against human baselines to reach perfect predictive accuracy of expert labels. The expert labels are based on linguistic theories that are firstly subject to change as the theories are either refined or disproven, and secondly built upon the empirical basis of human judgements in the first place. Our belief is as follows: just as any new, tractable theory of linguistics must account for the empirical phenomena observed in the LI-Adger dataset and human judgement data, so must any Language Model account for the linguistic phenomena observed in the dataset, and it must do so by tracking the human data within an established margin (δ).

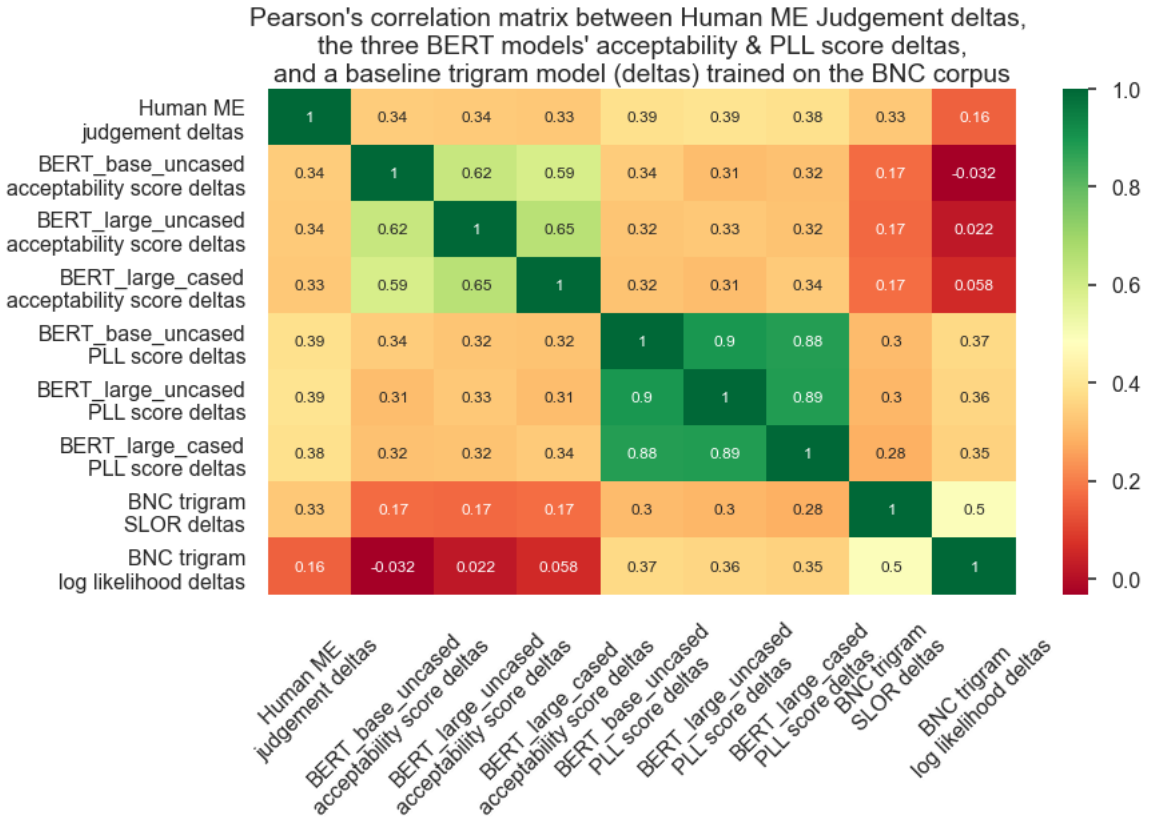

Figure 1: PCC matrix between human judgements and all three $\text{BERT}_{\text{CoLA}}$ & BERT_{MLM} models. As a further baseline we add the SLOR and log likelihood scores of a trigram model trained on the British National Corpus by Sprouse et al. (2018). All correlations shown have a $p < 0.0001$.

10 Contributions

With this work, we have identified three quintessential requirements to a test of Human Knowledge of Language in statistically-based Language Models: It must (1) exhibit comprehensive coverage of linguistic phenomena, (2) support attested and replicable human judgement data, and (3) test LMs' ability to track different linguistic phenomena across the full range of the acceptability gradient. Additionally, we advance a test of KoL for LMs that meets all three requirements.

First, we present the LI-Adger dataset as a gold standard test dataset: a comprehensive, empirically attested collection of 519 pairwise and multi-condition phenomena collected by Sprouse et al. (2013) from Linguistic Inquiry (LI) 2001-2010, and by Sprouse and Almeida (2012) from Adger's (2003) *Core Syntax* textbook. To complement the test dataset, we present statistically powerful, replicable and validated human Magnitude Estimation (ME) data collected by Sprouse and Almeida (2012) and Sprouse et al. (2013) as the ground truth labels we expect LMs with Human KoL to approximate. Finally, in order to tie the LI-Adger dataset and the human ME data together, we present the Acceptability Delta Criterion (ADC), a metric that tests LMs for Human KoL by requiring LMs to track the validated human judgements through the gradient spectrum within a specified margin (δ) as the acceptability values change across minimal pairs.

Our three main contributions with this work when taken together create a comprehensive and powerful input-output analysis of Human KoL for LMs. With further ongoing work, the test will empower us to see a fine-grained analysis of which phenomena a LM is able to account for in its output from how well it predicts the acceptability deltas around them. It is our hope that researchers will rely on the LI-Adger dataset for its coverage of empirically attested linguistic phenomena, embrace the paradigm of human judgements as the ground-truth labels that LMs are expected to approximate, and, beyond that, adopt the ADC as we take our understanding of KoL in LMs to the next level.

References

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

N. Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124.

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

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Adam Pauls and Dan Klein. 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)*, pages 959–968.

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

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2019. Pseudolikelihood reranking with masked language models. *CoRR*, abs/1910.14659.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Carson T Schütze and Jon Sprouse. 2013. Judgement data. In Robert Podesva and Devyani Sharma, editors, *Research Methods in Linguistics*, pages 27–51. Cambridge University Press.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using bert for speech recognition. In *Asian Conference on Machine Learning*, pages 1081–1093.

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

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

Jon Sprouse and Diogo Almeida. 2013. *The Role of Experimental Syntax in an Integrated Cognitive Science of Language*.

Jon Sprouse and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, 2(1):1.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Carson T Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Alex Warstadt and Samuel R Bowman. 2019. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

Alex Warstadt and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix A

An example of one multi-condition phenomenon from Chapter 8 (*Functional Categories III*) of the Adger (2003) *Core Syntax* textbook is presented in Table 2 below.

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 2: Example multi-condition phenomenon from the Adger dataset. Note: the original sentences in the Adger textbook used names from Greek mythology, but were changed to common names in order to avoid inadvertently influencing the native speakers' judgements.

This multi-condition phenomenon would constitute two minimal pairs in the LI-Adger dataset we present as the gold standard test set. The unacceptable sentence (ch8.150.*.01) forms a minimal pair with each of the other two acceptable forms. We include the pairings explicitly in the list below.

1. Minimal pair with ch8.151.g.01
   (a) It seems that Melissa is happy.
   (b) * Melissa seems that is happy.
2. Minimal pair with ch8.152.g.01
   (a) Melissa seems to be happy.
   (b) * Melissa seems that is happy.

Similarly, we form minimal pairs of all multi-condition phenomena in the original LI and Adger datasets by exhaustively enumerating all lexically matched pairs of acceptable & unacceptable sentences in the multi-condition phenomenon.

Appendix B

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 two minimal pairs in Table 3.

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 3. 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.

Appendix C

As an example of the ADC in action, consider the minimal pairs from Table 3 (Appendix B), expressed in Table 4 in terms of the Sentence ID of the grammatical sentence. We show the acceptability delta values for the Z-score transformed 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 the Acceptability Delta Criterion (ADC) was met.

Appendix D

We recreate in Table 5 an updated version of the table of MCC scores on the CoLA test set presented by Warstadt and Bowman (2019) (W&B). We add a column to indicate the authors responsible for training the model and include our three

Sentence ID	Sentence	ME Z-score
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 3: Two minimal pairs for the Linguistic Inquiry (LI) dataset collected by Sprouse & Almeida, 2012. The ME Z-score is the averaged Z-score transformation of the human Magnitude Estimation scores for each of the shown sentences across the 104 different experimental participants (Sprouse et al., 2013).

Sentence(g) ID	Δh_i	Δlm_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 4: The two minimal pairs from Table 3 expressed in terms of the Sentence ID of the grammatical sentence with acceptability delta values from the human judgements and Z-score transformed 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.

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 $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ were within error margins of the $\text{BERT}_{\text{CoLA}_{\text{large}}}$ model reported by W&B. Additionally, the maximum MCC score achieved here by $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$ 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.

Appendix E

We present in Table 6 four example minimal pairs that the three $\text{BERT}_{\text{CoLA}}$ models evaluated correctly under the BLiMP criterion, but not under the ADC with $\delta = 5.0$. We report the Z-score transformed acceptability scores for $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$, the best performing out of the three BERT models. In the case of $\text{BERT}_{\text{CoLA}_{\text{large-cased}}}$, the minimal pairs presented are 4 out of the 58 total minimal pairs that it scored correctly under the BLiMP criterion, but not under the ADC with $\delta = 5.0$.

The common factor among the four minimal pairs presented in Table 6, 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* and developing grammars that are able to account for the attested empirical phenomena. The paradigm of the ADC is therefore grounded in the empirical data itself, not on the theory built upon it.

Appendix F

We present in Table 7 four example minimal pairs where all three $\text{BERT}_{\text{CoLA}}$ models scored the pair correctly under the ADC with $\delta = 0.5$ but the trigram did not. In Table 8 we present the opposite: four example minimal pairs where the trigram model's SLOR scores evaluated the pair correctly under the ADC with $\delta = 0.5$ but none of the three $\text{BERT}_{\text{CoLA}}$ models did.

Appendix G

In Figure 4 we draw a correlation (scatter) plot with best fit line of the six BERT models' delta scores (Δlm_i) on the x-axis against the human judgement

Model$_{\text{CoLA}}$	**Mean (STD)**	**Max**	**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$_{\text{base}-\text{uncased}}$	0.478 (0.018)	0.514	0.522	HJVM & friends
BERT$_{\text{large}-\text{uncased}}$	0.542 (0.019)	0.583	0.578	HJVM & friends
BERT$_{\text{large}-\text{cased}}$	**0.574 (0.026)**	**0.613**	**0.588**	**HJVM & 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 5: Replication of Warstadt and 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 comparison.

deltas (Δh_i) on the y-axis. The figure reveals the BERT$_{\text{CoLA}}$ models' behavior is not gradient despite our reformulation of the BERT$_{\text{CoLA}}$ models' outputs in order to make them more gradient. The BERT$_{\text{CoLA}}$ models consistently predicted sentences to be either more than 90% acceptable or less than 90% unacceptable; the fine-tuning phase on the categorical expert labels in CoLA only seems to cripple the BERT$_{\text{CoLA}}$ models' performance.

The out-of-box BERT$_{\text{MLM}}$ models' PLL output deltas are much more gradient than those of the BERT$_{\text{CoLA}}$ models and seem to better roughly track the best-fit line.

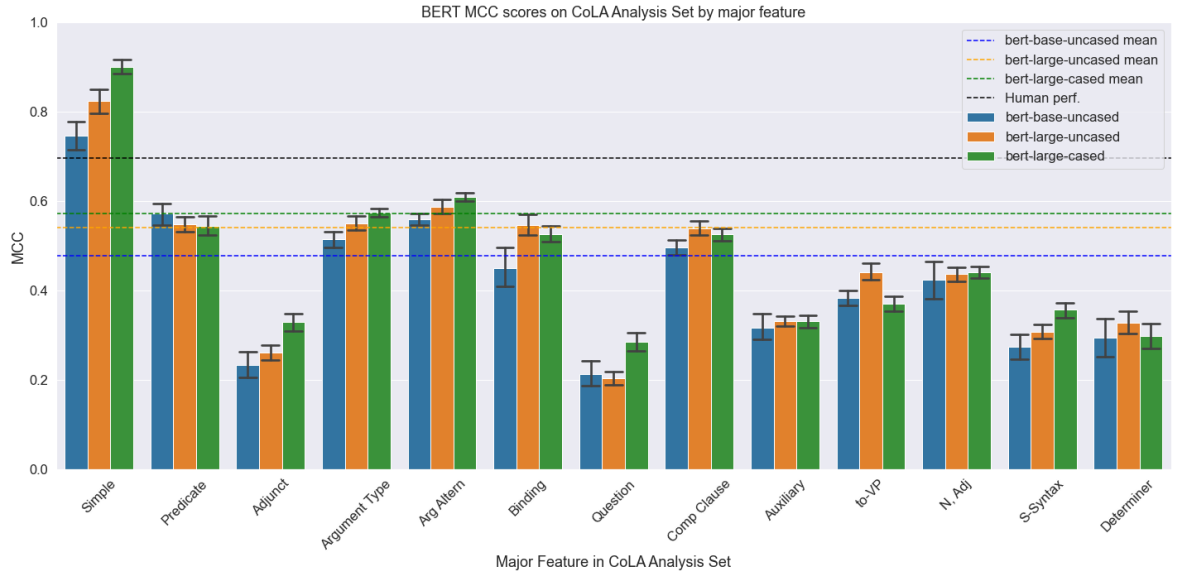

Figure 2: Replication of Warstadt and Bowman (2019) with our $\text{BERT}_{\text{CoLA}}$ models for comparison. Performance (MCC) on CoLA analysis set by major feature. Dashed lines show mean performance on the CoLA out-of-domain test set. From left to right, performance for each feature is given for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

Minimal Pair **Top**: Acceptable | **Bottom**: Unacceptable	Human judgement	BERT 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 6: Four minimal pairs where the $\text{BERT}_{\text{CoLA}}$ models meet the BLiMP criterion but not the generalized ADC with $\delta = 5.0$. We report the acceptability scores from the large-cased version of $\text{BERT}_{\text{CoLA}}$ The human judgement and BERT acceptability scores are already Z-score transformed. The common factor is that the human judgements disagree with the expert labels.

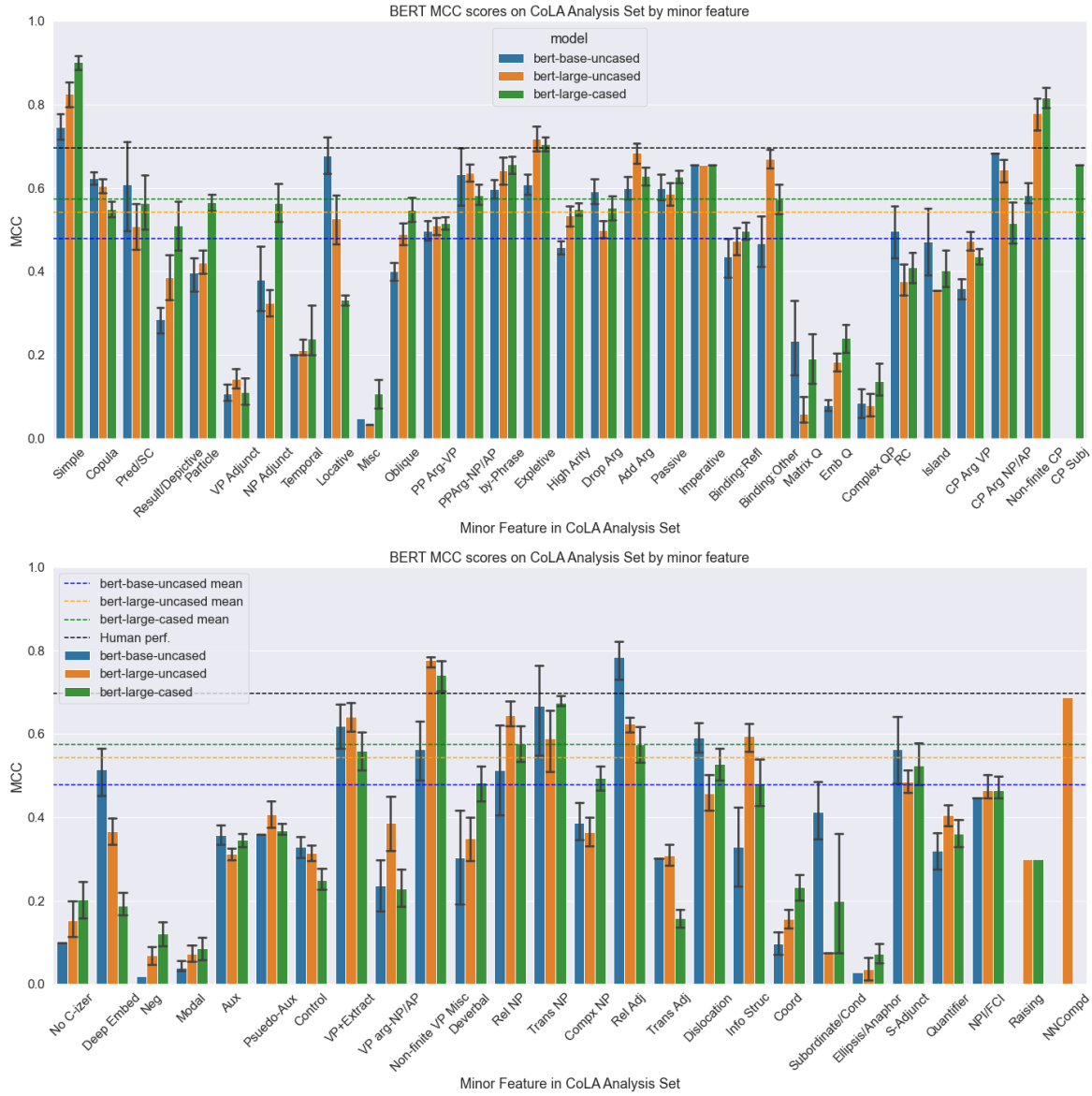

Figure 3: Replication of Warstadt and Bowman (2019) with our trained $\text{BERT}_{\text{CoLA}}$ models for comparison. Performance (MCC) on CoLA analysis set by minor feature. Dashed lines show mean performance on the CoLA out-of-domain test set. From left to right, performance for each feature is given for base-uncased, large-uncased, and large-cased.

Minimal Pair **Top**: Acceptable | **Bottom**: Unacceptable	trigram SLOR	$\text{BERT}_{\text{CoLA}}$ 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 7: Four minimal pairs where all $\text{BERT}_{\text{CoLA}}$ models meet the ADC with $\delta = 0.5$ but the trigram baseline does not. We report the acceptability scores from the large-cased version of $\text{BERT}_{\text{CoLA}}$. The trigam SLOR and $\text{BERT}_{\text{CoLA}}$ acceptability scores are already Z-score transformed.

Minimal Pair **Top**: Acceptable | **Bottom**: Unacceptable	trigram SLOR	$\text{BERT}_{\text{CoLA}}$ 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 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 acceptability scores from the large-cased version of $\text{BERT}_{\text{CoLA}}$. The trigam SLOR and $\text{BERT}_{\text{CoLA}}$ acceptability scores are already Z-score transformed.

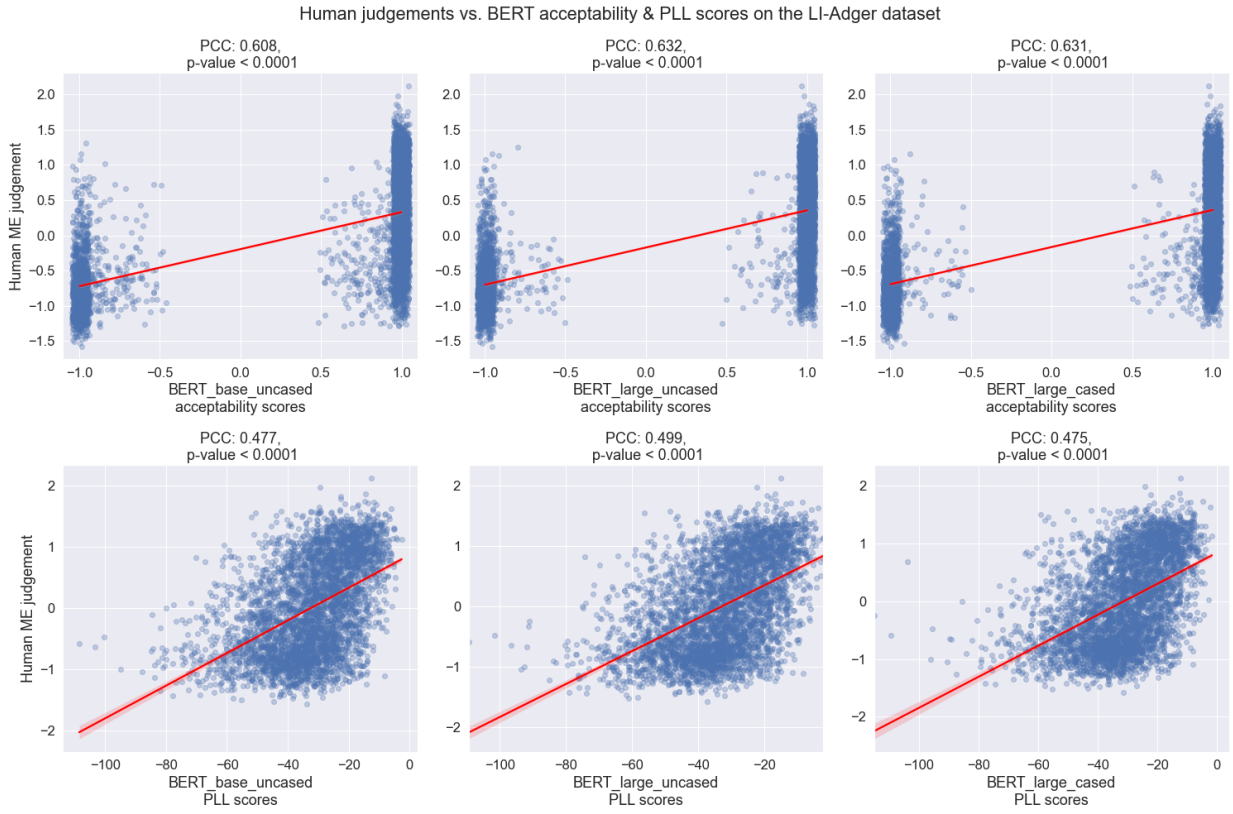

Figure 4: Scatterplot of human judgements (y-axis) vs. $\text{BERT}_{\text{CoLA}}$ acceptability scores, & BERT_{MLM} PLL scores from all three BERT models for each sentence in the LI-Adger dataset with best-fit line in red. We add a jitter of 0.05 along the x-axis and lower the alpha to 0.3 to highlight the density of the points.