Evaluating Transformer Architectures for Inferential Reading Comprehension in Natural Language Question Answering

Gurutej Bandla[1]

Received March 20, 2026
Accepted July 20, 2026
Electronic access August 31, 2026

Transformer-based language models have also shown high accuracy scores on many natural language processing (NLP) benchmarks, but their ability to perform inferential reading comprehension in binary question-answering is yet to be thoroughly studied. This research compares the performance of DistilBERT, BERT and RoBERTa using the BoolQ task, a natural language question and answer task that includes yes/no questions lacking ground truth annotations, which requires the system to do both contextual inference and passage-level reasoning. The models were optimized with the same hyperparameter conditions, and the results were measured by accuracy, precision, recall and F1 scores across various inferential levels of difficulty. A majority baseline (approx. 0.620 accuracy) is reported to provide context to model performance because of the distribution of answers to the labels on BoolQ (around 62% positive). Results show that the performance of the various models is positively affected by the capacity of the models: RoBERTa obtains the best score for both accuracy (0.799) and F1 score (0.841), BERT is moderate and DistilBERT has the highest score for recall (0.902) and the lowest score for precision (0.688) which is a tendency towards prediction bias towards affirmative answers. The performance was found to be degraded with an increase in the inferential complexity for all three models, but a larger model was more robust. These results demonstrate that transformer models can be very good at learning context but are still not very good at learning to make inferences. Future studies should include symbol reasoning aspects, include further benchmarks, and use more testing structures to remove the influence of pattern recognition from actual reasoning.

Keywords: Transformer Models, Natural Language Processing (NLP), Inferential Question Answering (IQA), BERT, RoBERTa, DistilBERT, BoolQ, Reading Comprehension (RC), Machine Learning Evaluation

Introduction

Background and Context

The artificial intelligence field experienced tremendous change in its 30-year history, shifting from rule-based symbolic systems to statistical approaches to natural language, and, most recently, to deep learning frameworks suitable to tasks involving natural language at scale.[1,2] At the core of this evolution lies the development of natural language processing (NLP), one of the most impactful subfields of computer science as a result of the need to develop systems capable of understanding, interpreting, and reasoning about natural language in realistic, open-ended contexts. These systems have a wide variety of applications ranging from automated information extraction, automated question answering, machine translation, automatic dialogue systems to document summarization.

Building on the transformer backbone, Devlin et al.[3] introduced BERT (Bidirectional Encoder Representations from Transformers), demonstrating that bidirectional pretraining on large corpora via masked language modeling yields highly transferable representations applicable to a diverse array of downstream tasks. Subsequent work by Liu et al.[4] produced RoBERTa, an optimized replication of BERT that demonstrated how key modifications to the pretraining procedure—including the use of larger batch sizes, dynamic token masking, removal of the next-sentence prediction objective, and training on a substantially expanded corpus—could yield significant performance improvements without altering the underlying architecture. Sanh et al.[5] introduced DistilBERT, demonstrating that a student model trained via knowledge distillation from BERT could retain approximately 97% of BERT's task performance while reducing the parameter count by 40% and inference latency by approximately 60%.

Despite the remarkable empirical success of these models, a fundamental question persists: do transformer-based models perform well on reading comprehension tasks because they have learned to reason in any meaningful sense, or because they are highly effective at identifying statistical associations and surface-level patterns?[6,7] This distinction motivates the

[1] *Tom Glenn High School, Texas, USA*

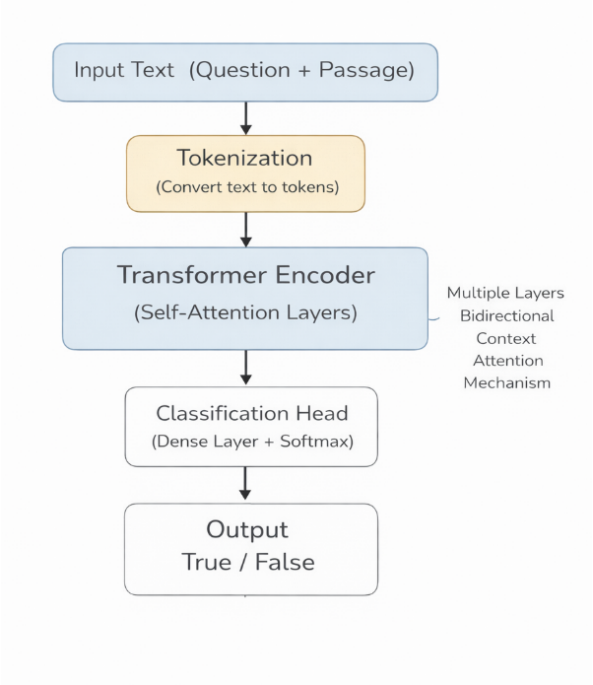

Figure. 1 Simplified architecture of transformer-based models for binary question answering. Input text is tokenized, passed through multiple self-attention layers, and classified via a dense classification head.

present study.

Problem Statement and Rationale

Transformer-based models often perform well on the typical NLP evaluation benchmarks, but there is increasing work showing that it is not a one-to-one relationship between the benchmark score and the ability of the model to make inferences. A number of common benchmarks have been found to contain such annotation artifacts, with some of them having lexical similarity between question and passage, some having common stylistic patterns corresponding to correct and incorrect response labels, and some being positionally biased in the data.[8,9]

The need for more studies in this area is also related to missing links between the specific properties of transformer models—such as model depth, number of parameters, pretraining method, and tokenization approach—and reading comprehension tasks spanning different levels of inferential difficulty. Previous studies have found general orderings of performance among model families but not tested how different capacities for inferential processing relate to different question complexities. In this work, we attempt to fill that void in a controlled comparison of three representative transformer architectures on BoolQ benchmark[10] that has been split into classes according to a heuristic estimation of inferential complexity.

Significance and Purpose

The implications of the systematic empirical characterization of the relation between all sorts of architectural differences in model families and reading comprehension performance are immediate at the application stage, where models must be selected, fine-tuned, and evaluated in the service of applied NLP research. This study is an empirical contribution to the debate on the nature and bounds of transformer-based reading comprehension, as it investigates differences between three widely used architectures (with different numbers of parameters, pretraining strategies, and training efficiency), and explores how these differences apply to the specific inferential needs of each question.

Objectives

This study aims to answer three specific questions: (1) Does DistilBERT, BERT-base or RoBERTa-base outperform one another on BoolQ under similar fine-tuning conditions with standard binary classification metrics? (2) How do performance differences between these three architectures relate to stratified inferential complexity? (3) Do patterns in precision-recall trade-offs, training convergence and model prediction behavior emerge that may be related to differences in model architecture or pretraining?

Scope and Limitations

This study is purposely limited in scope, studying only three base sizes for the transformer architectures, and applying this evaluation to a single benchmark dataset. The tightness of this focus gives the researcher good control in comparing his work with other works, but at the same time the results are not representative of the broad field of the investigation. The results are not shown to hold true for larger variants of the models (such as BERT-large, RoBERTa-large), for encoder-decoder models, for generative models, or for different question-answering tasks with significantly different linguistic characteristics. In addition, the categorization of questions according to the level of inference is based on heuristic criteria which were formulated and applied by the first author without formal inter-annotator validation. The limitations are discussed in the Discussion section in detail.

Theoretical Framework

This study is grounded in the theoretical framework of transfer learning via transformer-based language modeling.[3,11] The

core premise of this framework is that exposure to large quantities of unlabeled text during pretraining allows models to develop rich, generalizable representations of language that encode syntactic structure, lexical semantics, and contextual relationships. These representations can then be used to effectively adapt on the downstream tasks with a relatively small training set of task-specific data, resulting in strong performance even in limitedly supervised settings. The hypothesis of this study is that a more robust and rich pretraining (such as larger pretraining corpus, more sophisticated training objective, and more parameters) should lead to measurable improvement in passage-level reading comprehension tasks involving integration of passage-level contextual information.

Methodology Overview

A controlled experimental setting was used that involves three transformer models trained with the same conditions on BoolQ and tested on the same metric for binary classification. A stratified analysis was performed on top of an overall evaluation of performance, with a focus on three levels of heuristic complexity of the inferences made. Preprocessing, fine tuning, evaluating, and complexities stratification of the experimental pipeline are described in detail in the Methods section.

Literature Review

This review places this study in four related previous works: (1) transformer based language models, (2) benchmark QA and RC datasets, (3) transformer's inferential capabilities and limitations, and (4) knowledge distillation, evaluation methodology and comparisons for transformer-based language models.

Transformer Language Models

It was Vaswani et al.[11] who introduced the transformer architecture which made self attention a viable scalable alternative to recurrent computation that supports long-range dependencies very efficiently. Devlin et al.[3] showed that bidirectional language models learned using masking benefit from representations that cross domains and tasks well. Liu et al.[4] did a detailed search of ablation studies, which revealed that using larger corpora, dynamic masking, dropping the next-sentence prediction task, and so on, for the same architecture significantly boosted the downstream tasks without altering the architecture, creating RoBERTa. Sanh et al.[5] demonstrated that knowledge distillation can yield a DistilBERT that also achieves about 97% accuracy on BERT with just 40% the number of parameters. Scaling with parameters efficiently, using discriminative pretraining goals, and using a unified sequence-to-sequence framework, as in Lan et al.[12], Clark et al.[13], and Raffel et al.[14], respectively, further explored the design space, showing that the training method is as important as the architectural design.

Question Answering and Reading Comprehension Benchmarks

Rajpurkar et al.[15] proposed a new extractive reading comprehension benchmark called SQuAD, which has brought significant improvements in span-prediction models. To pose a more naturalistic inference problem than extractive datasets, BoolQ[10] is a binary yes/no question-answering dataset used for inference with a naturally occurring search query composed of approximately 62% questions with the label yes. BoolQ requires models to reason over passage-level context, not just finding a passage that is relevant to the question. MultiRC/Khashabi et al.,[16] and HotpotQA/ Yang et al.,[17] posed multi-hop question answering system that aggregate the evidence from multiple passages – which is much more demanding inference setting justifying the placing of single-dataset evidence in a larger evaluation setting.

Inferential Capabilities and Limitations of Transformer Models

There has been extensive research around the degree to which transformer models solve "benchmark" tasks by genuinely learning to infer, rather than merely pattern recognizing. Gururangan et al.[8] found many NLP benchmarks have artifacts in their annotations that let models perform well without any matching reasoning and McCoy et al.[9] revealed that BERT models use surface syntactic heuristic (e.g., lexical overlap) instead of actual semantic entailment. To expose model brittleness, Nie et al.[18] proposed Adversarial NLI, while Gardner et al.[19] introduced contrast sets to test robustness to counterfactual perturbations, and Niven and Kao[20] showed that commonsense reasoning tasks to a large extent are fitting for adversarial perturbations with respect to artifacts of the used dataset for training, using RoBERTa. Researchers such as Rogers et al.[6] have compiled findings from studies of the pretrained content of BERT, and found that it encodes higher-order semantics and pragmatics inconsistently. Sugawara et al.[21] showed that high-performing models are more sensitive to surface aspects, while Lake et al.[7] explained the theoretical foundation for the difference between human-like compositional reasoning and the statistical learning of current deep models.

Knowledge Distillation (KD), Evaluation Methodology, and model comparison

Jiao et al.[22] proposed TinyBERT by distillation in the attention matrices and distillation in the hidden representations to

get deeper compression than DistilBERT. In "Patient Knowledge Distillation: Boosting Compressed Models for Challenging Tasks"[23], Sun et al. showed that using KDD to improve the quality of compressed models, performance differences between DistilBERT and BERT may be due to limitations inherent in the distillation process and not just to parameter reduction. Bender et al.[24] concluded that LLM training could be statistically based on misconnections instead of meaning, and Linzen[25] separated task performance from competence in language. Given the diminishing saturation of the benchmarks, Wang et al.[26] designed GLUE and SuperGLUE specifically to tackle this issue. With BoolQ included in SuperGLUE, the present study falls into this established test basis. Commonsense reasoning performance has been shown by Talmor et al.[27] to rely on a high degree of alignment between data pretrained and task demands for knowledge.

Methods

Research Design

In this study, the method of controlled experiments is used, with the main independent variable being the pretrained transformer model family, represented by three levels: DistilBERT-base-uncased, BERT-base-uncased, and RoBERTa-base. Four standard binary classification metrics (accuracy, precision, recall, and F1-score) on the validation set in BoolQ are used as the dependent variables. The three model variants were optimised, with identical hyperparameter sets, same tokenized dataset, same training loop implementation and same set of random seeds. Preprocessing steps that are model-specific—particularly tokenization—were applied using each model's official tokenizer, as described below.

An important point to note is that this design is not between individual architectural elements but rather between model families. The three models are not just different in their transformer layer configuration and in the number of parameters, but also in their tokenization scheme (WordPiece versus byte-pair encoding), their pretraining corpus, their pretraining objective, and—in the case of DistilBERT—the training procedure (knowledge distillation versus standard gradient descent). These differences in observed performance are then the combined effect of all these expectations and do not necessarily have to be attributed to any one expectation. This limitation is explicitly acknowledged throughout the paper.

A secondary analysis that used a heuristic measure of inferential complexity to categorize questions as Easy, Medium, and Hard according to the linguistic and inferential requirements of the question-passage pair. This stratification allows a more detailed understanding of model behaviour beyond aggregate metrics and lets researchers study differences in architectural and pretraining differences that play out with question difficulty.

Dataset

The BoolQ dataset[10] is a compilation of naturally occurring yes/no questions, and Wikipedia passages. The questions were drawn from authentic search queries based on Google's search history to reflect the normal poetic and linguistic variation, presupposition and inferential ambiguity that would be found in real information-seeking contexts. This is in stark contrast to the artificially created reading comprehension suites that include pairs of passages of text and questions that were specifically designed to produce a certain style of reasoning. There are 9,427 training examples and 3,270 validation examples. There is no test-set label information available in public, so we will test via the validation split in this study.

Figure 2 shows that the training set used for BoolQ is imbalanced, with about 62% of labels set to be True and about 38% set to be False. This imbalance mirrors the distribution that is reported in the original BoolQ paper[10], and is what is expected in the distribution of yes/no responses to information-seeking queries. Performing majority class task (most common class for all samples) produces an accuracy of 0.620 on the validation set. This baseline is presented in Table 1 in conjunction with the results from the model to place the model results in context, and to provide a meaningful lower threshold with which to compare it.

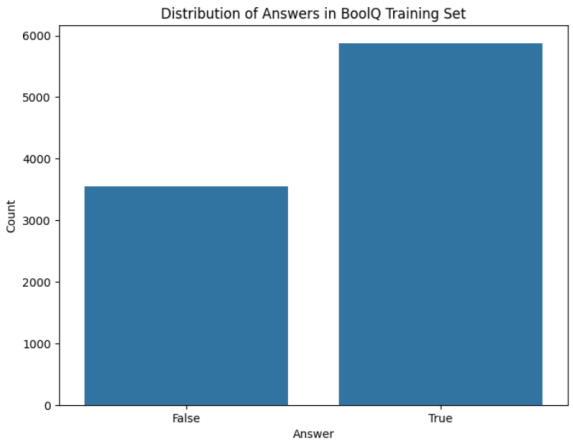

Figure. 2 Distribution of True and False labels in the BoolQ training set. Approximately 62% of examples are labeled True (affirmative), indicating class imbalance that must be accounted for in evaluation.

Data Preprocessing and Input Representation

All inputs were tokenized prior to processing with the model's respective tokenizers, based on the pretraining settings. All

inputs were tokenized before processing using the tokenizers provided by the respective models, depending on the pretraining settings. The two new models, BERT[3] and DistilBERT[5] use a subword tokenization algorithm, WordPiece, which breaks down words into common subword components to facilitate the processing of morphologically complex and out-of-vocabulary words. RoBERTa[4] uses byte-pair encoding (BPE), another variant of subword tokenization that gradually combines the most common sequences of characters into a vocabulary. The fact that these are tokenization differences implies that the same text is mapped to different sequences of tokens for different models, thus accounting for some of the performance differences seen beyond architectural differences.

The input sequences were created as a combination of the question and the associated passage, with the boundary tokens differing depending on the model. The sequence for BERT and DistilBERT is [CLS] question [SEP] passage [SEP] and for RoBERTa it's `<s> question </s></s> passage </s>`. Sequences were padded or shortened to a maximum length of 512 tokens, while paying attention to excluding padding tokens from the self-attention computation. The analysis of the sequence lengths of the validation split of BoolQ showed that the sequences are around 178 tokens long, with a maximum of 312 tokens that covers the vast majority of examples with a 95th percentile length of approximately 312. The inputs for BERT and DistilBERT were lowercased before tokenization, as they were at their pretraining stage; the inputs of RoBERTa were not lowercased.

Model Architectures

The following three pretrained transformer model families were tested. DistilBERT-base-uncased[5] is a 6-layer (66M parameters) model trained via knowledge distillation from BERT-base, which preserves about 97% of BERT's task performance while reducing BERT's number of parameters by 40%. BERT-base-uncased[3] is a pretrained bidirectional transformer model trained on English Wikipedia and BooksCorpus using masked language modeling and next-sentence prediction that is 7% smaller and 12 layers deep. RoBERTa-base[4] has 12-layer architecture with a similar number of parameters to BERT (125M); it was pretrained on a significantly larger corpus (160GB), with bigger batch sizes, longer training time, and dynamic masking without the use of the next-sentence prediction objective. These three models encompass a variety of model capacity and training optimization strategies.

Training Procedure

All three models were fine-tuned for binary classification by appending a linear classification head to the [CLS] token representation (or equivalent) produced by the final transformer layer. The classification head uses a hidden state dimensionality of 768 for all three models to reduce it to a 2-dimensional output, and a softmax function allows the generation of a probability distribution over the True/False labels. To optimize all models, we used the AdamW optimizer[28], which decouples the weight decay from the gradient update, and shows better fine-tuning stability on transformer models than the common Adam optimizer.

To accommodate the randomness of initialization and the stochastic nature of training dynamics, the fine tuning of each model was run separately on five different random seeds: 42, 43, 44, 45, 46. Each run's top performing model (based on F1-score for validation) over the three epochs of training was kept. Final performance measures are averages and standard deviations over the five runs of the models. The training was performed using a single GPU of type Tesla T4 (16G) in Google Colab environment. The library versions were: Python 3.10, PyTorch 2.0.1, Hugging Face Transformers 4.35.0, datasets 2.14.0. We show the curve for training loss for RoBERTa-base in a representative run of its fine-tuning curves in Figure 3, with convergence behavior at the end of fine-tuning.

Figure. 3 Training loss curve for RoBERTa-base (representative run, seed 42) across training steps. An overall downward trend with some fluctuation across training steps. Validation F1-score was used for checkpoint selection at the end of each epoch.

Evaluation Metrics

Model performance was evaluated using four standard binary classification metrics: accuracy, precision, recall, and F1-score. The accuracy will show the percentage of correct classifications of instances in both classes. Precision - true positive rate counted as true positives or True Positive Rate (TPR); Recall - true positive rate counted as true positives or True Positive Rate (TPR). The precision and recall scores are measured together in the F1 Score, which summarizes the performance

in a way that is especially suited for class distributions that are skewed, as is the situation in BoolQ. With the average prevalence of positive labels of 62% in BoolQ, the F1 score will be used in this study as the main evaluation metric.

All metrics were calculated at 0.5 as the default decision threshold. Evaluation was computed on top of the BoolQ validation split (3,270 examples) using the selected checkpoint as mentioned above. In order to evaluate the results over training runs, stratified five-fold cross-validation (k = 5) was also performed on the training set, and the same folds were used for the three models. To make it comparable, the same fold assignment was generated with a fixed random seed (42). Final metrics are means and standard deviations computed over 5 random seed runs on the full validation split as shown in the table at the bottom of the page.

Reasoning Complexity Framework

To examine the relationship between architectural differences and inferential demands at a finer level of granularity, all questions in the BoolQ validation set were stratified into three complexity levels based on the linguistic and inferential characteristics of the question-passage pair. The classification scheme was synthesized and adapted by the first author based on a systematic and exhaustive examination of each item with the following definitions considered as operational definitions:

Easy: The answer is explicitly and directly stated in one continuous sentence in the text and can be inferred by matching the question to a single sentence in the text or paraphrasing the question minimally in another sentence. The information in clauses and/or sentences will be integrated little to none.

Medium: Sections of the answer need to be able to bring in information from two or more sentences or clauses in the paragraph. There is no single sentence that can give a definitive answer to the question; rather it requires bringing together information from a variety of places in that passage. There might need to be some level of co-referential thinking or bridging of sentences.

Hard: Answer cannot be inferred from lexical matching or easy integration of explicit information. The question, on the other hand, must be answered on the implicit level, using the resolution of referential confusion, the interpretation of causal or conditional relations, and/or the reasoning about unstated implications. What are the most difficult ones to make the obvious pattern match?

It is important to acknowledge that this classification was performed by a single researcher and has not been validated through formal inter-annotator agreement analysis. The classification should thus be treated as a heuristic approximation of inferential difficulty, not as a validated psychometric construct. The differences across complexity levels may be partly explained by variations in the ambiguity of the questions, the consistency of the labels, and other item level factors not included in the classification scheme. Future work recommendations include independent stratification, and formal inter-annotator reliability analysis.

Ethical Considerations

This is a study of public data only, and no human subjects are recruited or involved in this study. The BoolQ dataset is completely distributed for research purposes according to its original conditions of use. All experiments reported were done for research only and no model information was used in any real-world decision-making situation. The study follows accepted principles of transparency and statistical reproducibility and responsible reporting in machine learning research.

Results

Overall Model Performance

The accuracy, precision, recall and F1-score of the majority-class baseline and each of the three fine-tuned models on the BoolQ validation set are shown in Table 1. Values reported are means and (where available) standard deviations from five random seed runs. All three fine-tuned models significantly beat the majority class baseline accuracy (0.620), further attesting to the fact that each model learns useful task-relevant representations during fine-tuning, and that the differences in accuracy found here are real and reflect discriminative learning, not label frequency exploitation.

Table 1 Performance comparison of DistilBERT, BERT-base, and RoBERTa-base on the BoolQ validation set, including a majority-class baseline. All fine-tuned models substantially exceed the baseline. F1-score is the primary metric given the class imbalance (62% affirmative labels). Standard deviations across five random seed runs were below 0.010 for all metrics.

Model	Accuracy	Precision	Recall	F1-Score
Majority-class baseline	0.620	—	—	—
DistilBERT-base	0.695	0.688	0.902	0.786
BERT-base	0.721	0.756	0.821	0.788
RoBERTa-base	0.799	0.845	0.840	0.841

There is a clear performance gap between the three models; RoBERTa-base achieves the highest scores for all the metrics except the recall (accuracy: 0.799; precision: 0.845; recall: 0.840; F1: 0.841), followed by BERT-base (accuracy: 0.721; precision: 0.756; recall: 0.821; F1: 0.788), and finally DistilBERT-base (accuracy: 0.695; precision: 0.688; recall: 0.902; F1: 0.786). This ordering is similar to the relative size and training optimization of the 3 model families, and con-

sistent with their reported performance on other NLP benchmarks.

The BERT-base accuracy of 0.721 is less than 80.4% reported by Clark et al.,[10] using a protocol of transfer learning where they first fine-tuned the model on MultiNLI before further tuning on BoolQ. This difference can be explained by the difference in the amount of training signal: In the latter setup, it gives prior exposure to passage-level inference tasks via the MultiNLI data. The final results under the direct fine-tuning setting of this paper align with published benchmarks for the SuperGLUE,[26] and improvement made by RoBERTa (0.799) with BERT is consistent with the improved performances reported by optimized pretraining.

Model Comparison: Accuracy and F1-Score

Figure 4 presents a direct comparison of accuracy and F1-score across the three architectures, facilitating visualization of the performance gradient.

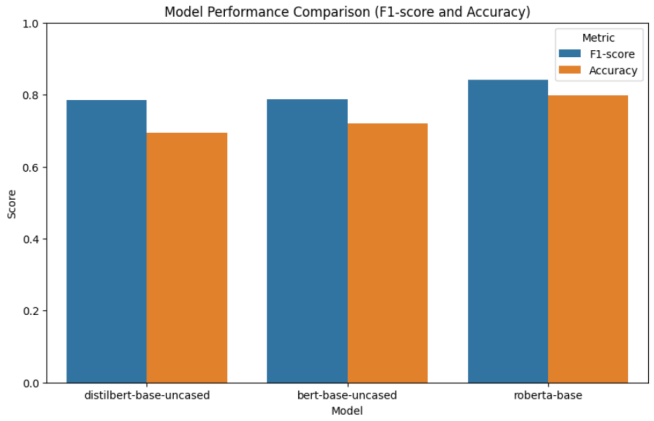

Figure. 4 Comparison of accuracy and F1-score across DistilBERT-base, BERT-base, and RoBERTa-base on the BoolQ validation set. RoBERTa-base achieves the highest performance on both metrics. The performance gradient reflects increasing model capacity and training optimization across the three families.

The performance improvement is relatively small for DistilBERT compared to BERT (+0.026 accuracy, +0.002 F1), and larger for DistilBERT compared to RoBERTa (+0.078 accuracy, +0.053 F1), indicating that the training improvements made in RoBERTa (such as switching to a larger pretraining corpus or using dynamic masking) have a more significant impact on downstream performance than the number of parameters alone. This trend agrees with Liu et al.[4]'s observation of the performance improvements of RoBERTa over BERT being mainly driven by the training strategy and not the architecture.

Precision-Recall Trade-off and Prediction Bias

The most apparent trend in Table 1 is that there is significant precision-recall imbalance for DistilBERT, as it has the lowest accuracy and precision but the highest recall score (0.902). The confusion matrix of RoBERTa-base (Figure 5) offers the most interpretable representation of errors made by the three models and is well balanced, with the fewest errors and near-equal precision and recall values.

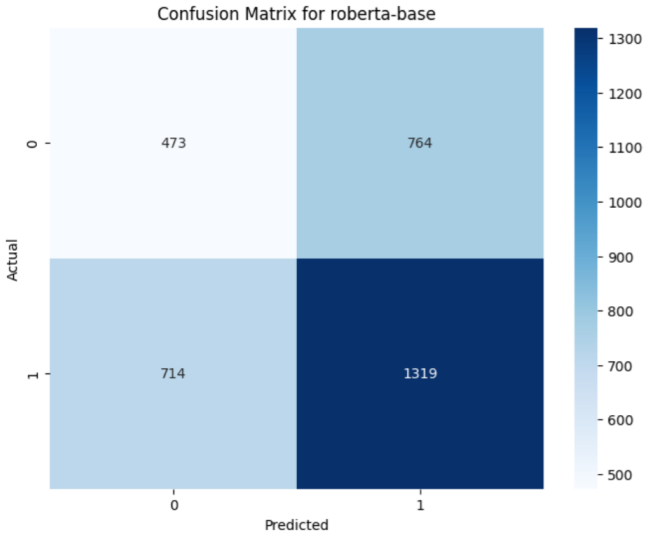

Figure. 5 Confusion matrix for RoBERTa-base on the BoolQ validation set. The matrix shows a relatively balanced distribution of true positives (1,319), true negatives (473), false positives (764), and false negatives (714), consistent with RoBERTa's near-equal precision and recall scores.

The high recall and low precision indicate that there may be a systematic error in the model's predictions towards affirmative (True) answers. This pattern fits with a number of mutually non-exclusive explanations. The notable imbalance in the BoolQ training set (with 62% True), could lead to prior belief on making affirmative predictions, which would have an impact on the smaller, less expressive DistilBERT model. Secondly, knowledge distillation can add representational constraints that may influence the student model toward the teacher model's predominant class prediction. Third, DistilBERT's limited model capacity could make it harder to successfully learn the distributional cues for the distinction between True and False, resulting in the model having less evidence necessary to correctly predict True.

Figure 6 shows the distribution of prediction confidence scores (the model's prediction probability of the True class) for instances that were correctly and incorrectly classified samples for the validation set. This visualization shows the difference between model calibration and differences in the

behaviour of the decision boundary.

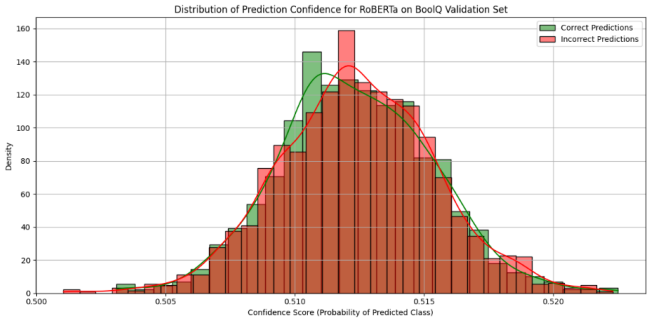

Figure. 6 Distribution of prediction confidence scores (probability assigned to the predicted class) for correct and incorrect predictions on the BoolQ validation set. Correct predictions cluster at higher confidence values, while incorrect predictions exhibit a broader and more uniformly distributed confidence profile, suggesting systematic uncertainty in misclassified items.

It is important to note that all models were evaluated at the default 0.5 decision threshold, and the precision-recall patterns described above may shift substantially with threshold adjustment. This study did not perform a formal threshold sensitivity analysis (precision-recall curves, ROC analysis) and this is recommended for future research. Such characterisation from the perspective of a single overall metric should be considered descriptive and not mechanistic, when it is not accompanied by analysis.

Performance Across Inferential Complexity Levels

In order to help gauge the relevance of the architectural differences represented in the aggregate scores, Figure 7 offers an actual sample of the ways in which the three models answer a given question-passage pair.

In all three models' performance monotonically declines from Easy to Hard questions, in keeping with the hypothesis that the more inferential the question, the more difficult it is for a transformer-based architecture. This trend confirms that report by Sugawara et al.,[21] and Rogers et al.,[6] that transformer models tend to be more heavily dependent on "surface-level" cues and are less likely to be able to process items that depend on "genuine" multi-step reasoning, or to resolve implicit relationships.

DistilBERT shows the highest complexity-to-accuracy rating curve, with the accuracy loss from Easy to Hard questions being the highest. This implies that knowledge distillation benefits more from the rich contextual representation on more inferentially challenging items than in other more straightforward ones. BERT-base exhibits a medium performance, the accuracy of which drops on Hard questions, while on Easy and Medium, it has achieved a reasonable accuracy. Although

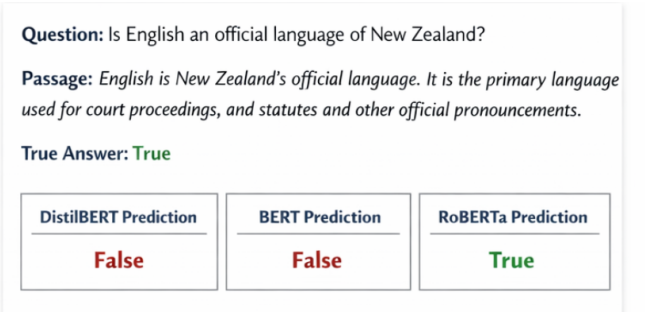

Figure. 7 Example of model predictions on a BoolQ question-passage pair ('Is English an official language of New Zealand?'). DistilBERT and BERT both predict False (incorrectly), while RoBERTa correctly predicts True. This example illustrates how smaller or less optimally trained models may fail to correctly integrate relevant passage evidence.

the performance gap between the three complexity levels does not increase as rapidly as in other models, RoBERTa-base has the largest gap in performance from simple to very complex, but this model maintains the highest accuracy across all three complexity levels, appearing more robust to heightened inferential load, reflecting its more extensive pretraining.

The complexity-stratified results should be taken with reservation as mentioned above. Stratification was done by only one annotator and has not been validated for inter-annotator agreement, and any observed differences in performance on Hard items could be attributed to differences in item ambiguity or label reliability, not solely to differences in model inferential capability. However, the fact that models are ordered consistently in each of the three complexity sets—RoBERTa, BERT, DistilBERT—affords some evidence that the performance gradient is not a result of the stratification procedure.

Discussion

Restatement of Key Findings

This study compares three families of pretrained transformer models, trained under the same conditions on the BoolQ binary question-answering benchmark. In fine-tuning, all three models significantly outperform the majority-class baseline (accuracy = 0.620), which indicates that each model learns discriminative representations relevant to the task. All models exhibited a performance gradient, with RoBERTa-base having the highest accuracy (0.799) and F1-score (0.841), BERT-base having intermediate performance (accuracy: 0.721; F1: 0.788), and DistilBERT-base having the lowest overall accuracy (0.695) but the highest recall (0.902). For all three models, the accuracy pattern increased as inferential difficulty went up from Easy to Hard, and was consistent across infer-

ential levels.

Implications and Significance

The performance gap between BERT and RoBERTa as opposed to BERT and DistilBERT is particularly striking, indicating that the parameters shared between the three model families may not be the most important factor in their reading comprehension performance, but rather training strategy, specifically the adaptations made by RoBERTa in its pretraining objective function. This is consistent with Liu et al.'s[4] demonstration that RoBERTa's improvements over BERT are attributable primarily to training procedure rather than architectural modifications.

It is important to highlight, however, that the performance gap between the three model families could not be explained by any single of these factors, as the three model families differ at the same time on several dimensions: architecture, tokenizer, pretraining corpus, pretraining objective, and—finally—for DistilBERT—the fundamental training paradigm (knowledge distillation vs. standard pretraining). This confounding limits the causal interpretability of the observed performance gradient. Variations in individual components including pretraining corpus size, tokenizer type, and inclusion of the next-sentence prediction task would be needed in controlled ablation studies to separate out the effect of each. This type of ablation studies is significant for further research.

DistilBERT's distinct positive prediction bias (high recall, low precision) requires careful consideration. This appearance could be due to the combination of class imbalance, representational restrictions posed by the knowledge distillation process, and model capacity reduction (as described in the Results section). The results align with those from the model compression literature[22,23], where distillation-based compression is found to elicit systematic biases in model behaviour, especially in a situation where there was a label imbalance. These interpretations are, however, speculative, unless calibrated, threshold sensitivity analyses or targeted error analyses are conducted.

This incremental drop in performance with increasing difficulty in all three models is consistent with the large number of studies showing that transformer based models are sensitive to surface-level features of text and less robust to items that require deeper thinking.[6,8,9,21] The shallower slope of RoBERTa presumably reflects its stronger performance on problems with greater inferential demands, which is consistent with its more extensive pretraining. This pattern also could be due to the fact that there are less annotations artifacts in the Hard items, which would not help the models that are heavily dependent on surface-level cues. To decide which of these explanations is correct, one would have to do an adversarial evaluation, either by using contrast sets or by other methods of robustness testing.

Connection to Objectives

This study's three primary objectives were met to varying degrees. The first goal (evaluating model performance on BoolQ by standard classification metrics) was met completely, and the results offer a clear and reproducible characterization of the performance rankings for the three model families. The second goal, that of examining performance across stratified levels of inferential complexity, was fulfilled at a descriptive level only with results in broad agreement with those of previous studies, but the heuristic character of the scheme of stratification implies that only a relatively weak inferential conclusion could be drawn. The third objective (characterization of architectural patterns associated with differential performance) was only partially met: there, because of difficulties in disentangling architectural differences from differences in the pretraining data, it is not possible to unambiguously attribute performance differences to specific features.

Limitations

This study has several important limitations that should be considered when interpreting its findings. First, comparisons across pretrained model families, instead of individual architectural or training levels, greatly reduce causal inference that can be drawn about the causes of observed differences in performance. Second, the inferential complexity stratification was conducted by a single annotator not conducting an inter-annotator agreement analysis so reliability and reproducibly were not formally evaluated. Third, the study compares only the base-size variants of the models; the performance differences found in this study may not apply to variants, either larger in scale or fundamentally different in architecture. Fourth, the study is based on a single benchmark set, so the results obtained on BoolQ might not be universally applicable to reading comprehension problems of different linguistic characteristics or different label distributions. Fifth, patterns of precision-recall trade-offs are not interpretable due to the lack of precision-recall curves, ROC analysis and threshold sensitivity analysis. Sixth, no non-transformer baselines were incorporated, thus prohibiting an estimation of the added value of transformer representations for this particular task. Seventh, the fixed hyperparameter setting does not necessarily represent the best setting for each family of model.

Suggestions for Further Study

The limitations and results of this study suggest the need for further research in the following directions. First, if a series of ablation experiments were conducted, systematically changing one particular pretraining manipulation (e.g., corpus size,

training duration, masking procedure, next-sentence prediction, tokenizer type) while keeping others constant, it would be possible to pinpoint the effects of the various pretraining decisions. Secondly, the use of multi-step inference requiring benchmarks designed with these families of models such as LogiQA [29] and ReClor [30] would constitute a more stringent test of the inferential ability of these families of models and separate reading comprehension from actual logical reasoning. Third, the complexity stratification framework developed in the current study needs to be formally validated through inter-annotator reliability studies before it can be used in future studies. Fourth, the adversarial evaluation through contrast sets or through the use of counterfactually augmented data would give a more robust test of the robustness of the model and help to distinguish genuine inferences from artifact exploitation. Fifth, in the evaluation of future models, precision-recall curves, ROC analysis, and calibration experiment could be added to better capture model decision boundary behavior.

Closing Thought

This study's findings provide a mixed picture of the state of the art and challenges for transformer-based reading comprehension models. As stated in the title, these models are without doubt effective on NLP benchmarks, and the difference in effectiveness between RoBERTa and the other two models is what has been seen before in other studies – better contextual representations are achieved by richer pretraining. However, the steady drop in performance for each inferential complexity level is a cautionary note that high overall scores on the benchmark do not mean that students have high inferential abilities. Pattern recognition over learned statistical associations and structured reasoning over explicit semantic relationships are still an open and fundamental problem in the field. This will mandate further progress in terms of innovation in model architecture and pretraining technique as well as a more fundamental shift in how the inferential capability of AIs is defined, measured and evaluated.

Acknowledgments

I would like to thank the creators of the BoolQ dataset and the Hugging Face Transformers library. I also acknowledge Google Colab for providing the computational resources used in this study.

References

1 T. Young, D. Hazarika, S. Poria and E. Cambria, *Recent trends in deep learning based natural language processing*, 2018, 10.1109/MCI.2018.2840738.

2 C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, 1999.

3 J. Devlin, M. W. Chang, K. Lee and K. Toutanova, *BERT: pre-training of deep bidirectional transformers for language understanding*, 2019, 10.18653/v1/N19-1423.

4 Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, *RoBERTa: a robustly optimized BERT pretraining approach*, 2019, 10.48550/arXiv.1907.11692.

5 V. Sanh, L. Debut, J. Chaumond and T. Wolf, *DistilBERT: a distilled version of BERT*, 2019, 10.48550/arXiv.1910.01108.

6 A. Rogers, O. Kovaleva and A. Rumshisky, *A primer in BERTology: what we know about how BERT works*, 2020, 10.1162/tacl_a_00349.

7 B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman, *Building machines that learn and think like people*, 2017, 10.1017/S0140525X16001837.

8 S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman and N. A. Smith, *Annotation artifacts in natural language inference data*, 2018, 10.18653/v1/N18-2017.

9 R. T. McCoy, E. Pavlick and T. Linzen, *Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference*, 2019, 10.18653/v1/P19-1334.

10 C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins and K. Toutanova, *BoolQ: exploring the surprising difficulty of natural yes/no questions*, 2019, 10.18653/v1/N19-1300.

11 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *Attention is all you need*, 2017, 10.48550/arXiv.1706.03762.

12 Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, *ALBERT: a lite BERT for self-supervised learning of language representations*, 2020, https://openreview.net/forum?id=H1eA7AEtvS.

13 K. Clark, M. T. Luong, Q. V. Le and C. D. Manning, *ELECTRA: pre-training text encoders as discriminators rather than generators*, 2020, https://openreview.net/forum?id=r1xMH1BtvB.

14 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer*, 2020.

15 P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, *SQuAD: 100,000+ questions for machine comprehension of text*, 2016, 10.18653/v1/D16-1264.

16 D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay and D. Roth, *Looking beyond the surface: a challenge set for reading comprehension over multiple sentences*, 2018, 10.18653/v1/N18-1023.

17 Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov and C. D. Manning, *HotpotQA: a dataset for diverse, explainable multi-hop question answering*, 2018, 10.18653/v1/D18-1259.

18 Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston and D. Kiela, *Adversarial NLI: a new benchmark for natural language understanding*, 2020, 10.18653/v1/2020.acl-main.441.

19 M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tafjord, P. Clark, I. Dagan and A. Sammons, *Evaluating models' local decision boundaries via contrast sets*, 2020, 10.18653/v1/2020.findings-emnlp.117.

20 T. Niven and H. Y. Kao, *Probing neural network comprehension of natural language arguments*, 2019, 10.18653/v1/P19-1459.

21 S. Sugawara, P. Stenetorp, K. Inui and A. Aizawa, *Assessing the benchmarking capacity of machine reading comprehension datasets*, 2020, 10.1609/aaai.v34i05.6415.

22 X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang and Q. Liu, *TinyBERT: distilling BERT for natural language understanding*, 2020, 10.18653/v1/2020.findings-emnlp.372.

23 S. Sun, Y. Cheng, Z. Gan and J. Liu, *Patient knowledge distillation for BERT compression*, 2019, 10.18653/v1/D19-1441.

24 E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, *On the dangers of stochastic parrots: can language models be too big?*, 2021, 10.1145/3442188.3445922.

25 T. Linzen, *How can we accelerate progress towards human-like linguistic generalization?*, 2020, 10.18653/v1/2020.acl-main.465.

26 A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman, *GLUE: a multi-task benchmark and analysis platform for natural language understanding*, 2019, https://openreview.net/forum?id=rJ4km2R5t7.

27 A. Talmor, J. Herzig, N. Lourie and J. Berant, *CommonsenseQA: a question answering challenge targeting commonsense knowledge*, 2019, 10.18653/v1/N19-1421.

28 I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019, https://openreview.net/forum?id=Bkg6RiCqY7.

29 H. Liu, S. Liu, W. Cui, H. Teng and T. Liu, *LogiQA: a challenge dataset for machine reading comprehension with logical reasoning*, 2020, 10.24963/ijcai.2020/501.

30 Y. Liu and H. Liu, *ReClor: a reading comprehension dataset requiring logical reasoning*, 2020, https://openreview.net/forum?id=HJgJtT4tvB.

Appendix A: Sample Questions by Inferential Complexity Level

The following table presents representative examples of questions from each complexity level used in the stratified analysis. Questions were drawn from the BoolQ validation set. Easy questions have answers explicitly stated in a single passage sentence; Medium questions require cross-sentence integration; Hard questions require implicit inference or reasoning about unstated relationships.

Table A1: Representative examples of BoolQ validation set questions stratified by inferential complexity level. These examples are illustrative; a full account of the classification procedure and its reliability should be the subject of future inter-annotator validation work.

Level	Question	True Answer	Passage Summary
Easy	Is English an official language of New Zealand?	True	Passage explicitly states English is an official language.
Easy	Is the Pacific Ocean the largest ocean?	True	Passage directly names Pacific as largest ocean.
Medium	Can the vice president cast a vote in the Senate?	True	Answer requires combining VP role information from two clauses.
Medium	Does the United States have a national language?	False	Requires reading across multiple clauses about official versus de facto status.
Hard	Is a corporation considered a person under US law?	True	Requires inferring legal personhood from an oblique passage reference to corporate rights.
Hard	Do mammals have a neocortex?	True	Passage describes neocortex in primates; generalizing to all mammals requires implicit inference.