HIGH SCHOOL EDITION
Journal of Student Research

Empirical Approach to Understanding Natural Language Models

Bhuvishi Bansal

Delhi Public School, India

ABSTRACT

In this paper, we try to understand natural language models by treating them as black boxes. We want to learn about these models without going into their technical details pertaining to network architecture, tuning parameters, training datasets, and schedules. We instead take an empirical approach, where we classify the datasets into various categories. For scalability and avoiding subjective bias, we use Latent Dirichlet Allocation (LDA) to categorize language text. We fine-tune and evaluate natural language models for our tasks. We compare the performance of the same model across multiple categories and for the same category across multiple models. This can help not only in choosing models for the desired categories but is also useful in understanding the model attributes that can explain performance variation. We report here the observations from this empirical study and our hypotheses. We find that models do not perform uniformly across all the categories, which could be because of uneven representation of these categories in their training datasets. Models that specialized/fine-tuned for specific tasks had higher variance in performance across categories than the generic models. Some categories have high performance consistently across all models, while others have high variance. The code for this research paper is available here: https://github.com/bhuvishi/llm_understanding

Introduction

Owing to the popularity of language models, in particular LLMs, there has been a phenomenal rise in the number of models in the public domain. There are standard tasks (question-answering, summarization, etc.) and associated datasets for benchmarking performance, but these models are huge and not interpretable. Thus, good performance in one category doesn’t imply the same for other categories and not even similar categories. Thus aggregated metrics used for benchmarking may not reflect performance on the various sub-categories.

To make these models useful and evaluate risk-reward tradeoffs, we need to understand them better so that we can choose the right model for a problem/task of interest. Understanding which model attributes could be credited for performance on a subtask could not only help in identifying/choosing models, but also in developing or fine-tuning better models for a problem/task.

In this paper, we take an empirical approach for understanding the language models. Instead of looking under the hood for analytical understanding, for instance by studying network-architectures, activation-distributions, fine-tuning parameters, training tasks, and datasets, we treat the model as a black box. We take this first step with an empirical approach by going into the nuance of texts and evaluating models on these nuances.

Instead of relying on manual categorization, and to ensure objectivity and enhance scalability, we employed an unsupervised learning approach by training a Latent Dirichlet Allocation (LDA) classifier. It's important to note that due to the inherent stochasticity in the training process, each run produces a slightly different LDA model. To maintain consistency and mitigate noise arising from LDA classifier variance, we employed the same pre-trained LDA classifier for evaluating all the language models.

Background

In the realm of computational models and data-driven solutions, the evaluation of model performance has often been reduced to a single summary metric. This simplification, however, belies the nuanced nature of machine learning models, which, like any human endeavor, are not infallible and are susceptible to errors. Furthermore these models make errors that are unintuitive for humans. While a model may excel in one context (let's call it "x") and prove its mettle in another (referred to as "y"), the assumption that it will seamlessly extend/extrapolate its success to a composite context ("xy") or interpolate to an intermediate (“(x + y)/2”) often proves fallacious (Balestriero, 2021; Lu, 2022). For instance, if an autonomous vehicle (AV) learns how to navigate around pick up trucks (“x”) sites and traffic cones (“y”), it doesn’t extend its learning to pickup trucks carrying traffic cones (“xy”).

The inherent complexity of these models, their decision-making processes, and the intricacies of real-world data necessitate a more granular examination. This is especially true as we want to embed these models in more business and mission - critical systems and workflows. It is within this recognition that we embark on a journey to delve deeper into the world of model evaluation. Our motivation is grounded in the understanding that models are not silver bullets, and their performance cannot be distilled into a single number. We further believe that deeper evaluation will help us in demystifying the intricacies in the internal working of these models, which will help us in building models.

In light of these considerations, we have adopted a structured approach. We have identified and created subtopics that will allow us to dissect the intricacies of model evaluation. By exploring these subtopics, we aim to shed light on the subtleties, challenges, and opportunities that arise when assessing the capabilities of computational models. Through this research endeavor, we seek to contribute to the ongoing dialogue surrounding model evaluation and, ultimately, pave the way for more informed and nuanced decision-making in the realm of machine learning and artificial intelligence.

Dataset

In this study, we employed the Standard Question Answering Dataset (SQuAD) v2 (Rajpurkar, 2018) as the foundational dataset for our research. SQuAD, a widely recognized benchmark in the field of natural language processing, has played a pivotal role in advancing the capabilities of question answering systems.

SQuAD v2 (Rajpurkar, 2018), an extension of the original SQuAD dataset (Rajpurkar, 2016), presents a unique challenge by incorporating unanswerable questions in addition to the standard query-answer pairs (Rajpurkar, 2018). This augmentation adds a layer of complexity, pushing the boundaries of model evaluation beyond the conventional metrics. It aligns with our research's broader objective of comprehensively assessing model performance, particularly in scenarios where the answers may not always be readily available.

The SQuAD v2 dataset comprises a diverse range of topics and contexts, sourced from a multitude of articles, making it a valuable resource for evaluating question answering models under real-world conditions. Its inclusion in our study ensures the relevance and applicability of our findings to a wide array of practical use cases.

By utilizing the SQuAD v2 dataset as the cornerstone of our research, we aim to provide a robust and rigorous evaluation of the models under investigation, emphasizing their adaptability and robustness in handling complex questions, including those without definitive answers.

Methodology/Models

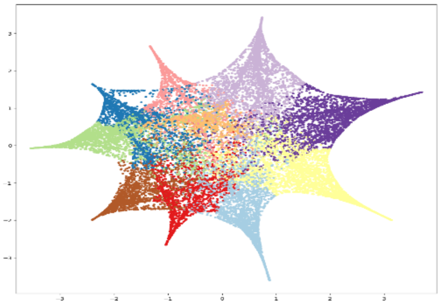
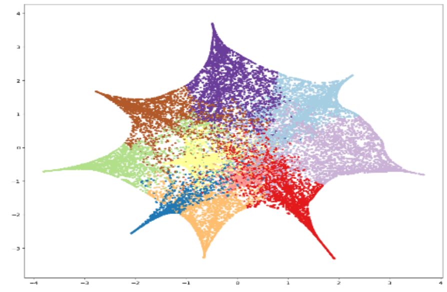
Data Organization

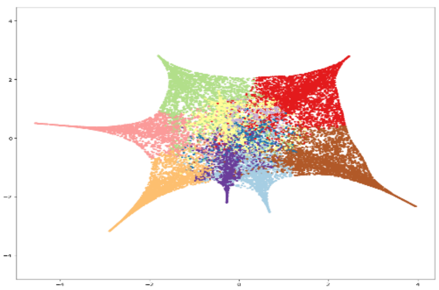
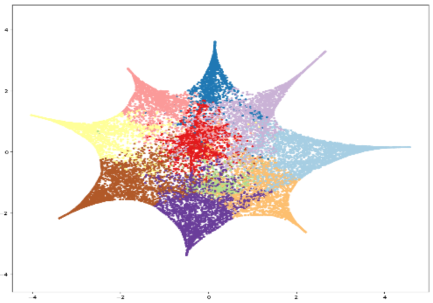
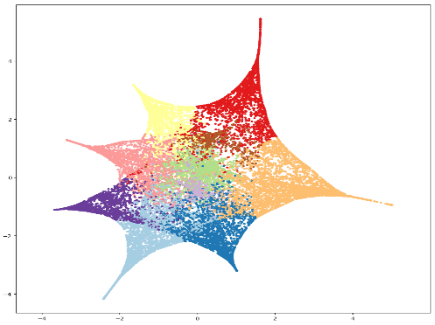
Our research commences with a meticulous organization of the training data, derived from the Standard Question Answering Dataset (SQuAD) v2, based on the diverse types of questions it contained. This initial step aims to categorize questions into distinct groups, providing a foundation for our subsequent analyses. However, recognizing the limitations of a one-dimensional comparison, we seek a more comprehensive approach.

Latent Dirichlet Allocation (LDA)

To gain deeper insights and uncover hidden patterns within our dataset, we employed Latent Dirichlet Allocation (LDA), a powerful technique for topic modeling. LDA allows us to naturally group similar passages, and go beyond predefined categories. This approach enables us to investigate how various models perform when faced with passages of different complexities and themes. To ensure the quality of our LDA models, we tuned hyper-parameters like the number of training epochs. We evaluated the models by using variance ratio and perplexity metrics. Variance ratio seemed noisy while perplexity provided a clearer trend. As expected, we found that perplexity improved with more epochs, but improvements started to diminish at higher epochs. So, we picked the LDA model with 4000 training epochs. We fixed the number of topics to 10, to make the later analysis tractable.

Table 1. LDA Metrics

No. of training epochs	Variance ratio for topic clusters	LDA model's perplexity	
100	10053.74	-9.18	 Fig 1
500	9978.92	-9.21	 Fig 2

1000	11076.94	-9.17	 Fig 3
2000	11897.49	-9.15	 Fig 4
4000	10709.93	-9.14	 Fig 5

LDA Model Evaluation Metrics

Model Selection

We evaluate the performance of a diverse set of twelve models, each chosen for its relevance and applicability to question answering tasks:

DistilBERT-base-uncased (Sanh, 2019): A BERT base (Devlin, 2018) model distilled for efficiency. It disregards case sensitivity, promoting language consistency.

BERT Base (Uncased) (Devlin, 2018): Pretrained English model using masked language modeling (MLM). Promotes case insensitivity, enhancing language understanding.

BERT large uncased (Devlin, 2018): Pretrained English model with masked language modeling (MLM). Offers extensive language understanding, case insensitivity.

BERT large uncased whole word masking (Devlin, 2018): English pretrained model using MLM. Innovative Whole Word Masking technique applied, enhancing word-level context understanding.

BERT base multilingual uncased (Devlin, 2018): Pretrained on top 102 languages from Wikipedia. Utilizes masked language modeling (MLM) objectives. Uncased for language consistency.

Financial BERT (Hazourli, 2022): created for efficient financial NLP, reducing computational resource demands.

Roberta base (Liu, 2019): Pretrained English model with case sensitivity, MLM objective.

Bio ClinicalBERT (Alsentzer, 2019): Utilizes BERT or BioBERT base, trained on clinical data. Different variants for diverse clinical text sources

Legal BERT base uncased (Chalkidis, 2020): Specialized BERT Models for Law and NLP Advancements

SEC BERT base (Loukas, 2022): Financial NLP Models for FinTech and Research**.**

AstroBERT (Grezes, 2021): Case-Sensitive NLP Model for Astrophysics by NASA/ADS.

HateBERT (Caselli, 2021): A BERT model fine-tuned on 1 million banned Reddit posts.

Performance Metrics

To quantitatively assess the efficacy of these models, we utilized standard evaluation metrics such as the F1 score and its variance. These metrics provided a comprehensive view of each model's performance, including precision and recall, essential components of question answering systems.

By employing these metrics, we were able to draw meaningful comparisons and glean insights into the strengths and weaknesses of each model.

Analysing the Impact of Question Types

Our research ventures beyond mere model comparison; it delves into the influence of question types on model performance. We identified ten distinct topics within the dataset to represent a wide spectrum of questions:

Topic 0: Legal
Topic 1: Dynasties and Empires
Topic 2: Architecture and History
Topic 3: Natural and Financial Crisis/Disasters
Topic 4: Diverse, encompassing Technology and History
Topic 5: Science and Philosophy
Topic 6: Education
Topic 7: Wars
Topic 8: Music and sports
Topic 9: Competitions, including sports

Our investigation aims to uncover nuanced patterns in model performance across these topics, providing valuable insights into how different types of questions challenge and inform the capabilities of question answering systems.

Implications for the Future

The findings of this research have significant implications for enhancing the understanding and responsiveness of computers to a broad array of questions. By dissecting model performance across various question types, our work contributes to the refinement and advancement of question answering systems, ultimately improving their utility in diverse domains.

Results and Discussion

Table 2. F1 Scores by Topic

	DistilBERT base uncased	BERT base uncased	BERT large uncased	BERT large uncased whole word masking	BERT base multilingual uncased	Financial BERT	Roberta base	Bio Clinical BERT	Legal BERT base uncased	SEC BERT base	astroBERT	hateBERT
Topic / Size	**268MB**	**440MB**	**1372MB**	**1382MB**	**672MB**	**439 MB**	**501MB**	**436MB**	**440MB**	**439MB**	**439MB**	**440MB**
Topic 0	58.79	68.17	79.69	83.57	73.67	57.89	80.30	67.34	70.50	69.81	72.12	57.16
Topic 1	57.36	67.40	77.64	84.54	75.66	56.33	80.18	64.25	69.25	67.26	69.46	52.07
Topic 2	60.06	67.73	73.57	83.64	74.59	58.39	79.14	66.19	68.41	69.60	70.27	54.08
Topic 3	56.71	66.81	75.57	82.22	72.07	55.76	78.04	64.66	67.35	67.60	70.68	53.58
Topic 4	59.03	66.37	75.37	81.46	72.44	57.67	75.97	63.53	68.65	67.88	71.40	58.82
Topic 5	61.05	69.24	81.49	87.53	78.27	58.75	82.36	67.91	71.55	71.26	74.95	60.27
Topic 6	62.59	68.15	78.83	89.50	77.16	59.65	84.62	67.98	71.45	73.14	77.28	61.32
Topic 7	54.46	64.42	75.28	83.03	72.22	55.55	76.10	61.83	65.67	66.65	69.14	51.07
Topic 8	55.09	67.28	76.58	83.15	68.78	55.63	80.69	63.72	63.81	64.90	67.90	56.83
Topic 9	62.36	67.71	83.45	89.83	78.34	49.01	84.18	63.38	74.94	72.34	76.76	56.85
Overall	58.71	67.45	77.56	84.08	74.27	57.33	79.44	65.43	69.19	68.90	71.82	56.56
Normalized std scores	0.05	0.02	0.04	0.04	0.04	0.05	0.04	0.03	0.05	0.04	0.05	0.06
Mean score	58.75	67.33	77.75	84.85	74.32	56.46	80.16	65.08	69.16	69.04	71.99	56.20
Topic with least score	Topic 7	Topic 7	Topic 2	Topic 4	Topic 8	Topic 9	Topic 4	Topic 7	Topic 8	Topic 8	Topic 8	Topic 7
Topic with max score	Topic 9	Topic 5	Topic 9	Topic 9	Topic 9	Topic 6	Topic 6	Topic 6	Topic 9	Topic 6	Topic 6	Topic 6
Bottom topics with least scores	7,8,3	7,4,3	7,4,2	4,3,7	8,3,7	9,7,8	4,7,3	7,9,8	8,7,3	8,7,1	8,7,1	7,1,3
Top topics with max scores	6,9,5	5,0,6	9,5,0	9,6,5	9,5,6	6,5,2	6,9,5	6,5,0	9,5,6	6,9,5	6,9,5	6,5,4

Table 3. Variance by Topics of Normalised F1 Scores

Model / Topic	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Overall	Normalized std scores	Mean score	Topic with least score	Topic with max score	Bottom topics with least scores
DistilBERT base uncased	1.00	0.98	1.02	0.97	1.00	1.04	1.07	0.93	0.94	1.06	1.00	0.05	58.75	Topic 7	Topic 9	7,8,3
BERT base uncased	1.01	1.00	1.01	0.99	0.99	1.03	1.01	0.96	1.00	1.01	1.00	0.02	67.33	Topic 7	Topic 5	7,4,3
BERT large uncased	1.02	1.00	0.95	0.97	0.97	1.05	1.01	0.97	0.99	1.07	1.00	0.04	77.75	Topic 2	Topic 9	7,4,2
BERT large uncased whole word masking	0.98	1.00	0.99	0.97	0.96	1.03	1.05	0.98	0.98	1.06	0.99	0.04	84.85	Topic 4	Topic 9	4,3,7
BERT base multilingual uncased	0.99	1.02	1.00	0.97	0.97	1.05	1.04	0.97	0.93	1.05	1.00	0.04	74.32	Topic 8	Topic 9	8,3,7
Financial BERT	1.03	1.00	1.03	0.99	1.02	1.04	1.06	0.98	0.99	0.87	1.02	0.05	56.46	Topic 9	Topic 6	9,7,8
Roberta base	1.00	1.00	0.99	0.97	0.95	1.03	1.06	0.95	1.01	1.05	0.99	0.04	80.16	Topic 4	Topic 6	4,7,3
Bio ClinicalBERT	1.03	0.99	1.02	0.99	0.98	1.04	1.04	0.95	0.98	0.97	1.01	0.03	65.08	Topic 7	Topic 6	7,9,8
Legal BERT base uncased	1.02	1.00	0.99	0.97	0.99	1.03	1.03	0.95	0.92	1.08	1.00	0.05	69.16	Topic 8	Topic 9	8,7,3
SEC BERT base	1.01	0.97	1.01	0.98	0.98	1.03	1.06	0.97	0.94	1.05	1.00	0.04	69.04	Topic 8	Topic 6	8,7,1
astroBERT	1.00	0.96	0.98	0.98	0.99	1.04	1.07	0.96	0.94	1.07	1.00	0.05	71.99	Topic 8	Topic 6	8,7,1
hateBERT	1.02	0.93	0.96	0.95	1.05	1.07	1.09	0.91	1.01	1.01	1.01	0.06	56.20	Topic 7	Topic 6	7,1,3
STD scores	0.01	0.02	0.03	0.01	0.03	0.01	0.02	0.02	0.03	0.06	0.01					

Relevant Observations

- Topic 9 has the highest standard deviation of 0.0599. Its variation is also indicated by the fact that for some models it has the highest score and for others it has the lowest score.
- HateBERT (Caselli, 2021) has a higher variance of 0.061. It is likely because it is specialized for hate texts and thus performs better on such classes/topics while trading lower in performance on other classes/topics. We observe the same trend with other specialized models like FinancialBERT (Hazourli, 2022).
- BERT Base uncased (Devlin, 2018) has the lowest variance of 0.0191. It could be because it is a generic model and not specialized and thus doesn't have drastically different performance for different topics.

- BERT-large-uncased (Devlin, 2018) is the bigger version of BERT-base-uncased and has higher variance (0.0399 versus 0.0191). Increasing the size resulted in higher variance.
- DistillBERT-base-uncased (Devlin, 2018) is distilled from BERT-base-uncased (Devlin, 2018) and has a higher variance (0.048 versus 0.0191). Distillation resulted in higher variance too.
- BERT large uncased (Devlin, 2018) is the biggest model (1382MBs) and has the best overall score of 84.079
- DistillBERT-base-uncased (Sanh, 2019) is the smallest model (268MB) and has the worst overall score of 58.711
- All the models perform relatively poorly than other topics on topic 3 and 7. Topics 3 and 7 share similarities. Topic 3 is about natural and financial crises/disasters and topic 7 is about wars. Since all the normalized scores for these topics are below 1.0.
- All the models perform relatively better, than other topics, on topics 5 and 6. Topics 5 and 6 are similar. Topic 5 is about science and philosophy and topic 6 is about education. Since all the normalized scores for these topics are above 1.0.

Conclusion

In this paper, we have embarked on a journey to unravel the mysteries of natural language models, treating them as enigmatic black boxes. Rather than analytically delving into the intricate technical details, such as network architecture, tuning parameters, and training data specifics, we assess how a model performs across various categories and how different models fare within the same category. This not only aids in selecting models for specific categories but also offers insights into the model characteristics that contribute to performance variations.

Our approach has involved classifying datasets into various categories, but we sought to do so without introducing our subjective bias. To accomplish this, we have harnessed the power of Latent Dirichlet Allocation (LDA). We fine-tuned some LLMs for unsupervised learning to categorize language text. By doing so, we have unlocked a pathway to understanding these complex models.

We have conducted fine-tuning and evaluation exercises on natural language models across a spectrum of tasks. Our investigation has included comparing the performance of the same model across multiple categories and assessing different models within the same category. This comprehensive analysis serves a dual purpose. Not only does it aid in selecting the most suitable models for specific categories, but it also unravels key attributes of these models that underlie performance variations.

Our empirical study has yielded intriguing observations. We have discovered that models do not exhibit uniform performance across all categories, hinting at the influence of the representation of these categories in their training data. Furthermore, models fine-tuned for specialized tasks display greater variability in performance across categories compared to their more generic counterparts. Some categories consistently demonstrate high performance across all models, while others exhibited significant performance variances.

In light of the soaring popularity of language models, especially large language models (LLMs), the public domain has witnessed an explosion of model availability. However, these models, while powerful, remain largely opaque and uninterpretable. The challenge lies in recognizing that stellar performance in one category does not necessarily extend to other categories, not even those that seem similar. Consequently, aggregated metrics used for benchmarking may not accurately represent performance across various sub-categories.

Our quest for a deeper understanding of these models is driven by the need to make them truly useful and to evaluate the risk-reward trade-offs associated with their application. By gaining insights into the attributes that contribute to their performance on specific subtasks, we not only empower ourselves to choose the right model for a given problem but also pave the way for the development and fine-tuning of superior models tailored to specific tasks.

In this paper, we have taken an empirical stride toward unravelling the complexities of language models. Rather than dissecting them analytically, we have probed their nuances through a systematic evaluation of texts. Our use of Latent Dirichlet Allocation (LDA) for classification, with the added rigor of using a consistent LDA classifier across all language models, has paved the way for a more nuanced understanding of these black boxes. It is a step towards harnessing their power effectively while navigating the intricacies of language and meaning.

Acknowledgments

I would like to take this opportunity to express my heartfelt gratitude to the individuals and institutions that have supported me throughout the course of this research endeavour.

First and foremost, I extend my sincere thanks to my mentor, Clayton, whose unwavering guidance and invaluable insights have been instrumental in shaping the direction of this study. Your expertise, patience, and dedication to the pursuit of knowledge have inspired me throughout this journey. Your mentorship has not only enriched my understanding but has also kindled my passion for empirical research.

I would also like to acknowledge the support of my academic institution, which provided me with the resources and environment necessary for the successful completion of this research project. The encouragement and scholarly atmosphere at DPS, Karnal have been essential in fostering my intellectual growth.

Lastly, I want to express my gratitude to my family and friends for their unwavering support, encouragement, and understanding during this journey. Your belief in my capabilities has been a constant source of motivation.

This research would not have been possible without the collective efforts and encouragement of all these individuals and institutions. Thank you for being an integral part of this academic expedition.

References

Alsentzer, E. (2019, April 6). Publicly available clinical BERT embeddings. arXiv.org. https://arxiv.org/abs/1904.03323

Balestriero, R. (2021, October 18). Learning in high dimension always amounts to extrapolation. arXiv.org. https://arxiv.org/abs/2110.09485

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. Association for Computational Linguistics, 2021, 17–25. https://doi.org/10.18653/v1/2021.woah-1.3

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Αλέτρας, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. Association for Computational Linguistics, 2020, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Devlin, J. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

Grezes, F. (2021, December 1). Building astroBERT, a language model for Astronomy & Astrophysics. arXiv.org. https://arxiv.org/abs/2112.00590

Hazourli, A. R. (2022). FinancialBERT - A Pretrained Language Model for Financial Text Mining. ResearchGate. https://doi.org/10.13140/RG.2.2.34032.12803

Hoffman, M. (2010). Online learning for latent dirichlet allocation. https://papers.nips.cc/paper_files/paper/2010/hash/71f6278d140af599e06ad9bf1ba03cb0-Abstract.html

Liu, Y. (2019, July 26). Roberta: A robustly optimized BERT pretraining approach. arXiv.org. https://arxiv.org/abs/1907.11692

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., & Παλιούρας, Γ. (2022). FINER: Financial Numeric Entity Recognition for XBRL Tagging. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2022.acl-long.303

Lu, Y. (2022). Imitation Is Not Enough: Robustifying Imitation with Reinforcement Learning for Challenging Driving Scenarios. arXiv.org. https://arxiv.org/abs/2212.11419

Rajpurkar, P. (2018, June 11). Know what you don’t know: Unanswerable questions for SQUAD. arXiv.org. https://arxiv.org/abs/1806.03822

Rajpurkar, P. (2016, June 16). SQUAD: 100,000+ questions for machine comprehension of text. arXiv.org. https://arxiv.org/abs/1606.05250

Sanh, V. (2019, October 2). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv.org. https://arxiv.org/abs/1910.01108