Comparative Analysis of 47 Context-Based Question Answer Models Across 8 Diverse Datasets

Muhammad Muneeb[1,2], **David B. Ascher**[1,2,*], **and Ahsan Baidar Bakht**[3]

[1]School of Chemistry and Molecular Biology, The University of Queensland, Brisbane, 4067, Australia
[2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, 3004, Australia
[3]Mechanical Engineering Department, Khalifa University, Abu Dhabi, UAE
[*]d.ascher@uq.edu.au

ABSTRACT

Dataset Explanation

This section briefly explains the dataset, outlining the data curation and generation processes for the two datasets (IELTS and JournalQA).

We considered eight different datasets from multiple sources. **Atlas-math-sets** centers around mathematical concepts related explicitly to sets and mathematical operations (addition, subtraction, multiplication, and division). **Bioasq10b-factoid** contains questions from the BioASQ challenge, which contains biomedical question answers. **Bbiomedical_cpgQA** is tailored for question-answering tasks within the biomedical realm, explicitly focusing on Clinical Practice Guidelines (CPG). It encompasses a set of questions delving into medical practices and guidelines, providing a comprehensive resource for training and evaluating models in the healthcare domain. **QuAC** (Question Answering in Context) is a meticulously crafted dataset for advancing question-answering research. Questions are posed within a conversational context, demanding the model to consider the intricacies of previous turns in the dialogue. **ScienceQA** focuses on questions and answers related to scientific topics. It is a robust evaluation dataset scrutinizing a model's capacity to comprehend and adeptly respond to queries within the landscape of scientific knowledge. The **Question Answer Dataset** consists of articles from Wikipedia containing context, questions, and answers. The **IELTS** (International English Language Testing System) dataset is designed to assess English language proficiency across listening, reading, writing, and speaking skills. The IELTS dataset is centered on the English language, encompassing reading comprehension with corresponding answers. We compiled the dataset consisting of 50 questions from various online IELTS sample tests. The actual URL to the test, the passage or context, the question, and the answer to that question are available on GitHub. We considered only those IELTS questions that ended with a question mark sign. The **JournalQA** dataset is crafted from scholarly articles. Research articles contain information like the model's performance, the best-performing model, population, specific methodology, and other information that can be extracted using question-answer models. To generate such questions, we used two systematic literature reviews. The first review focused on articles related to solar power prediction[?], while the second centered on using polygenic risk scores for predicting type 2 diabetes[?].

Model Explanation

Table 1 shows the model name, size, and the dataset used to fine-tune the models.

Statistical Analysis

This section displays two heatmaps illustrating the t-test between the model's performance 1 and execution time 2.

Index	Model Name	Model Size (MB)	Dataset
1	twmkn9/albert-base-v2-squad2	12	SQuAD v2
2	valhalla/bart-large-finetuned-squadv1	388	SQuAD v1
3	deepset/bert-base-cased-squad2	104	SQuAD v2
4	google/bigbird-roberta-base?	122	Books, CC-News, Stories and Wikipedia.
5	google/bigbird-pegasus-large-arxiv?	551	Arxiv dataset
6	dmis-lab/biobert-v1.1	104	NA
7	deepset/roberta-base-squad2	119	SQuAD v2
8	Splend1dchan/canine-c-squad	126	NA
9	YituTech/conv-bert-base	101	NA
10	Palak/microsoft_deberta-large_squad	387	SQuAD v1
11	microsoft/deberta-v2-xlarge	844	NA
12	distilbert-base-uncased?	64	BookCorpus and English Wikipedia
13	bhadresh-savani/electra-base-squad2	104	SQuAD v2
14	nghuyong/ernie-1.0-base-zh?	96	Chinese
15	xlm-mlm-en-2048?	637	Masked language modeling
16	google/fnet-base?	80	Colossal Clean Crawled Corpus (C4)
17	funnel-transformer/small?	125	BookCorpus, English Wikipedia, Clue Web, GigaWord, and Common Crawl
18	EleutherAI/gpt-neo-1.3B?	1255	Pile
19	hf-internal-testing/tiny-random-gptj	1	NA
20	gpt2	119	WebText
21	kssteven/ibert-roberta-base?	119	NA
22	allenai/led-base-16384	155	NA
23	allenai/longformer-large-4096-finetuned-triviaqa	415	NA
24	facebook/mbart-large-cc25	583	Multilingual mbart model
25	mnaylor/mega-base-wikitext	8	wikitext-103
26	csarron/mobilebert-uncased-squad-v2	24	SQuAD v2
27	microsoft/mpnet-base	105	NA
28	google/mt5-small	165	101 languages
29	RUCAIBox/mvp?	388	NA
30	sijunhe/nezha-cn-base	98	Chinese
31	uw-madison/nystromformer-512	104	BookCorpus and English Wikipedia
32	facebook/opt-350m?	316	BookCorpus, CC-Stories, The Pile, Pushshift.io, and CCNewsV2
33	bert-base-uncased?	105	BookCorpus and English Wikipedia
34	google/rembert?	550	Multilingual Wikipedia data over 110 languages.
35	roberta-base?	119	BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories
36	andreasmadsen/efficient_mlm_m0.40	339	NA
37	ArthurZ/dummy-rocbert-qa	112	NA
38	tau/splinter-base?	103	BookCorpus and English Wikipedia
39	squeezebert/squeezebert-uncased?	49	BookCorpus and English Wikipedia
40	t5-small?	58	Colossal Clean Crawled Corpus (C4)
41	xlm-mlm-en-2048?	637	Masked language modeling
42	xlnet-base-cased?	112	XLNet model pre-trained on English language
43	uw-madison/yoso-4096	121	Masked language modeling
44	SRDdev/QABERT-small?	64	30k samples from the Stanford Question Answering Dataset
45	bert-large-uncased-whole-word-masking-finetuned-squad?	320	BookCorpus and English Wikipedia
46	facebook/bart-large-cnn?	388	CNN Daily Mail
47	ahotrod/electra_large_discriminator_squad2_512	319	SQuAD v2

Table 1. The table shows language models from Hugging Face included in this study. For access to the documentation of each model, utilize the following pattern: `https://huggingface.co/+ModelName` (e.g., `https://huggingface.co/twmkn9/albert-base-v2-squad2`). The Model Name column presents the name of each model, and the Model Size column denotes the size of the respective models. Additionally, the Dataset column specifies the datasets utilized in to fine-tune each model.

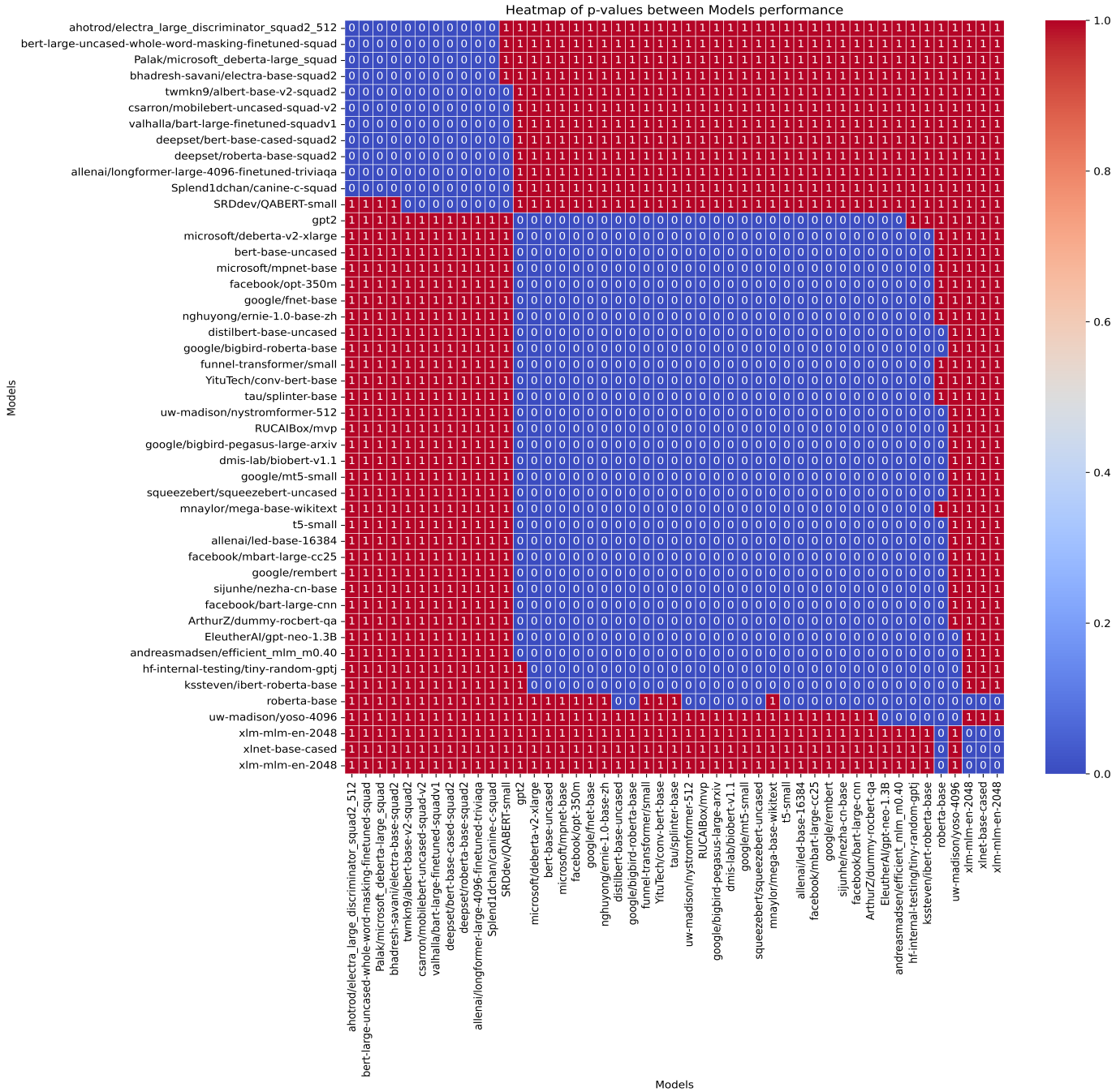

Figure 1. This heatmap depicts the t-test comparing the performance of each model across datasets. We sorted the values based on performance and calculated the t-test (p-value < 0.1). If the p-value for a t-test was less than 0.1, we assigned a value of 1 to that particular entry; otherwise, it is marked as 0. This representation indicates whether a specific pair of two models is statistically significant.

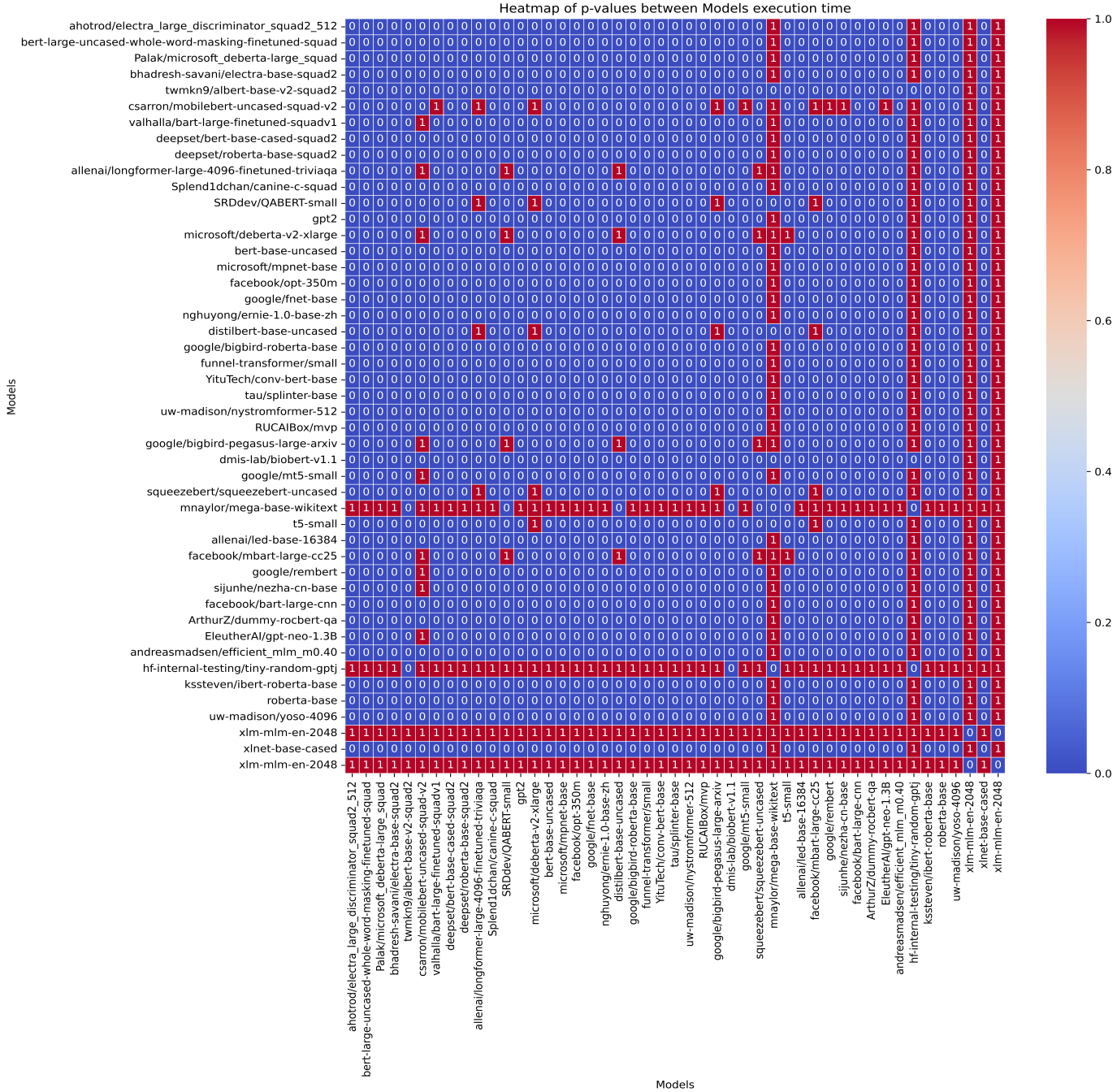

Figure 2. This heatmap depicts the t-test comparing the execution time of each model across datasets. We sorted the values based on performance and calculated the t-test (p-value < 0.1). If the p-value for a t-test was less than 0.1, we assigned a value of 1 to that particular entry; otherwise, it is marked as 0. This representation indicates whether a specific pair of two models is statistically significant.

References