Supplementary Information for:

Auto-generating Question-Answering Datasets with Domain-Specific Knowledge for Language Models in Scientific Tasks

Zongqian Li [a] and Jacqueline M. Cole *,[a,b]

a. Cavendish Laboratory, Department of Physics, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, UK.

b. ISIS Neutron and Muon Source, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Chilton, Didcot, Oxfordshire, OX11 0QX, UK.

** Author for correspondence: E-mail: jmc61@cam.ac.uk*

Table of Contents

A. Solar-cell corpora

It was considered important for this work to create corpora that have rich knowledge related to the scientific domain of interest, exhibit high quality, and are large enough to train a language model. After considering the accessibility, diversity, and size of different information sources, scientific journals from different publishers were selected as the main source for the language model to gain specialized knowledge.

The process of building each corpus for this work consisted of four main steps (Figure A-1). Firstly, keywords, time range, and maximum number of papers were used as a query to search for papers through the application programming interface (API) provided by the publisher. Secondly, the responses that contain the contents of the papers were obtained from the server. Thirdly, the papers were downloaded as eXtended Markup Language (XML) or Hyper-Text Markup Language (HTML) files. Finally, the plain texts in these XML or HTML files were extracted using ChemDataExtractor and used to create corpora. Due to the intrinsic limitations of the data-extraction capabilities of ChemDataExtractor, four types of special characters were left in the texts, including newline characters \\n, \n, spaces \xa0, and \u202f. All of these characters were removed from the texts and UTF-8 was used in all encoding and decoding processes.

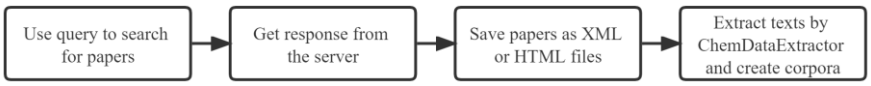

Figure A-1. Process for creating corpora.

Each solar-cell corpus contains solar-cell-specific papers from journals published by Elsevier, the RSC, and/or Springer. All papers were obtained by a search, using the keyword, “solar cell”. The API for downloading RSC and Elsevier papers is open to the public, while the API for accessing Springer papers is licensed through the Molecular Engineering group of the Cavendish Laboratory, University of Cambridge, UK. The total number of solar-cell-related papers from the three publishers for each year from 2000 to 2023 is shown in Figure A-2. The total trend is increasing over the last twenty years and Springer has published the highest number of solar-cell-related papers each year. In addition, some papers published before 2000 were downloaded, albeit these are not shown in Figure A-2.

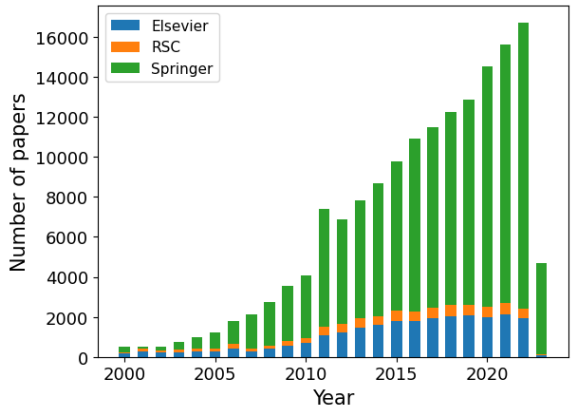

Figure A-2. Numbers of solar-cell-related papers published by Elsevier, RSC, and Springer between 2000 to 2023.

B. Solar-cell QA dataset

The maximum number of words in the contexts (split by spaces) for both first- and second-turn QA pairs is 282.

B.1 First-turn QA pairs

Based on the perovskite- and dye-sensitized solar-cell device databases and papers, the first-turn QA pairs in the Solar-cell QA dataset was created by a transformation algorithm and its statistical parameters are shown in Table B-1. There are 42,882 QA pairs in total. After shuffling by the random() function with a seed value of 50, these QA pairs were split into the train and test sets (0.8:0.2). Among the 16 properties in the QA dataset, eight properties have more than one thousand QA pairs and the power-conversion efficiency (PCE) performance characteristic has the largest number of QA pairs (16,081; 37.50%), followed by the open-circuit voltage (20.00%). According to Figure B-1, it can be known that most answers in the QA dataset are short and have a length of around five characters; this is because they are mainly the values of the properties and are combinations of numbers and units, while the contexts are much longer and have an average length of 240 characters.

Table B-1. Statistical parameters for the first-turn QA pairs in the Solar-cell QA dataset.

Parameter	Value
The number of:	
Properties	16
Total QA pairs	42,882
QA pairs in the train set	34,305
QA pairs in the test set	8,577
The number of QA pairs for:	
Power-conversion efficiency	16,081
Open-circuit voltage	8,619
Short-circuit current density	3,405
Fill factor	2,148
Active area	2,048
Solar simulator and irradiance	1,738
Counter electrode	1,431
Substrate	1,195
Other	6,217
The average length of:	
Context /characters	240
Answer /characters	6

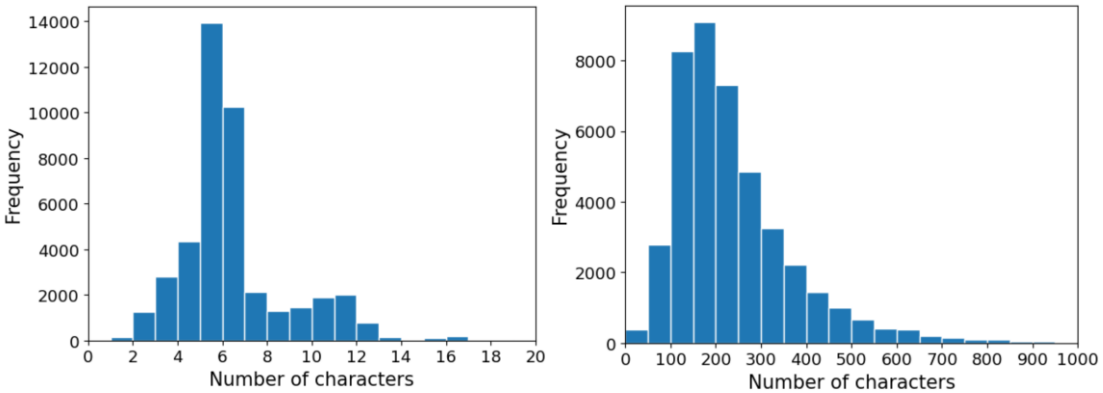

Figure B-1. (a) Distribution of answer character-lengths, and (b) distribution of context character-lengths for the first-turn QA pairs in the Solar-cell QA dataset.

To test the quality of the first-turn QA pairs in the Solar-cell QA dataset, an evaluation set was created for which parameters are shown in Table B-2. According to the proportions of properties present in the Solar-cell QA dataset, 1,000 QA pairs were randomly selected from the first-turn QA pairs therein. The answer for each QA pair was manually provided by the human and was compared to the answer generated by the algorithm.

Table B-2. Evaluation set for testing the quality of the first-turn QA pairs in the Solar-cell QA dataset.

Type of properties	Number of QA pairs
Power-conversion efficiency	380
Open-circuit voltage	200
Short-circuit current density	80
Fill factor	50
Active area	50
Solar simulator and irradiance	40
Counter electrode	30
Substrate	30
Other	140

For the first-turn QA pairs in the Solar-cell QA database, there are nine sources of error shown in Table B-3. Among these, 28.34% come from “not the whole value”, i.e., the answer is only a part of the value and could miss some numbers. “Not the higher value” means that the answer is not the largest of multiple-value options for the property, after using a new material or making any other improvement for the solar cell, which is the second most dominant error source (18.58%). The third largest error is “not a range” which means that the correct answer pertains to a range of values, albeit that the generated answer covers only one value. Except for “do not have units”, all the other errors pertain to intrinsic limitations of the data-extraction capabilities of ChemDataExtractor. These results prove that the transformation algorithm for creating the Solar-cell QA database achieves high performance and has avoided nearly all possible problems.

Table B-3. Error sources for the Solar-cell QA dataset (first-turn).

Error source (%)	Power-conversion efficiency	Open-circuit voltage	Short-circuit current density	Fill factor	Active area	Solar simulator and irradiance	Counter electrode	Substrate	Other	All (weighted average)
Not the whole number	37.14	11.11	33.33	0.00	100.00	0.00	41.67	100.00	3.95	28.34
Not a range	11.43	33.33	0.00	33.33	0.00	0.00	8.33	0.00	0.00	12.93
Does not have units	14.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.84	7.07
Not the higher value	25.71	44.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.58
Not a group of data	5.71	11.11	66.67	0.00	0.00	0.00	8.33	0.00	0.00	9.95
Not the correct value	2.86	0.00	0.00	33.33	0.00	100.00	8.33	0.00	17.11	9.55
Not whole context	2.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.07
Does not contain the answer	0.00	0.00	0.00	33.33	0.00	0.00	33.33	0.00	5.26	3.55
Specifier problem	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	61.84	8.96
Weight	37.50	20.10	7.94	5.01	4.78	4.05	3.34	2.79	14.49	

B.2 Second-turn QA pairs

The statistical details for the second-turn QA pairs in the Solar-cell QA dataset are shown in Table B-4. Among the ten properties in the second-turn QA pairs, seven properties have more than 100 QA pairs. The PCE has the largest proportion (42.32%), which is: 1.54 times that of the open-circuit voltage, 4.03 times that of the short-circuit current density, and 7.83 times

that of the active area. According to Figure B-2, most answers have a length of around two or four characters; this is because they tend to be abbreviations of materials or batch names, while contexts have much greater character lengths than the answers (~200 characters). Compared to first-turn QA pairs, second-turn QA pairs have six fewer properties and the average length of the answer is two fewer characters, albeit the average context length has five more characters.

Table B-4. Statistical parameters for the second-turn QA pairs in the Solar-cell QA dataset.

Parameter	Value
The number of:	
Properties	10
Total QA pairs	4,386
QA pairs in the train set	3,508
QA pairs in the test set	878
The number of QA pairs for:	
Power-conversion efficiency	1,856
Open-circuit voltage	1,207
Short-circuit current density	460
Active area	237
Fill factor	236
Solar simulator and irradiance	201
Charge-transfer resistance	110
Other	79
The average length of:	
Answer /characters	4
Context /characters	245

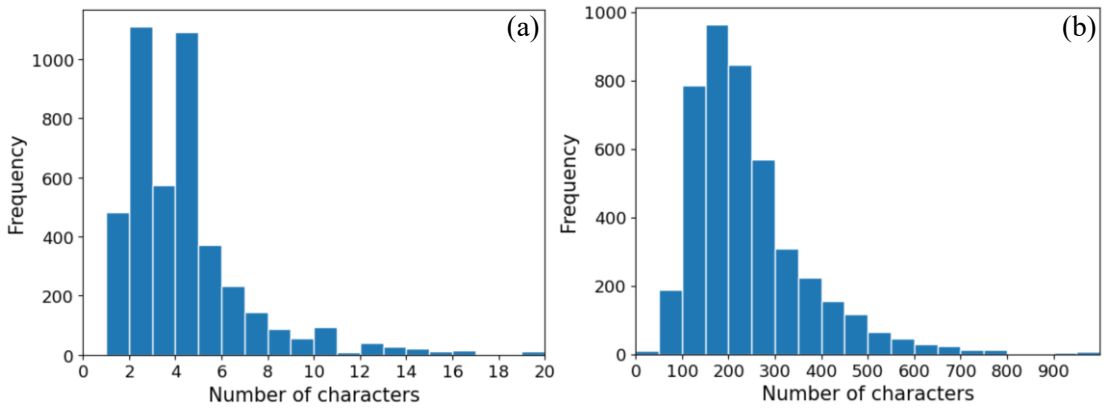

Figure B-2. (a) Distribution of character-lengths for answers, and (b) distribution of character-lengths for the context pertaining to the second-turn QA pairs in the Solar-cell QA dataset.

Table B-5 shows an evaluation set of 100 QA pairs that were randomly chosen from second-turn QA pairs according to the proportion of each property in the Solar-cell QA dataset. All the generated answers were compared to the standard answers that had been manually provided by the human, reaching an exact match of 72%. There are four sources of error in Table B-6. The most dominant error source is “not a material”, which includes more than half of the errors, i.e., the answer could be a number wrongly extracted as a material. “No answer” is the second largest error source, which shows that there is no material in the context. However, these two errors can be resolved by deleting the second-turn QA pairs with purely numerical answers, which will significantly improve the accuracy. All four error sources arise from intrinsic limitations of the data-extraction capabilities of ChemDataExtractor, which demonstrates the high quality of the improved transformation algorithm.

Table B-5. Evaluation set for the second-turn QA pairs in the Solar-cell QA dataset.

Type of properties	Number of QA pairs
Power-conversion efficiency	42
Open-circuit voltage	28
Short-circuit current density	10
Active area	5
Fill factor	5
Solar simulator and irradiance	5
Charge-transfer resistance	3
Other	2

Table B-6. Error sources for the second-turn QA pairs in the Solar-cell QA dataset.

Error sources	Count	Proportion
Wrong material	4	14.29%
No answer	5	17.86%
Not a material	16	57.14%
Not whole materials	3	10.71%
Total	28	100%

C Further Pretraining

The hyperparameters are: learning_rate: 2e-5; seed: 50; pad_to_max_length: False; validation_split_percentage: 5; warmup_ratio: 0.2; weight_decay: 0.01.

C.1 Further pretraining on Solar-cell Corpus Small

The originally formulated BERT-base models were further pretrained on the Solar-cell Corpus Small (scsmall). For the BERT-base-cased model shown in Figure C-1 (a), the initial training and evaluation losses are between 3 and 3.5. After these losses decrease with a constant speed over 20 epochs of refinement, the evaluation loss converges to 1.1, which is smaller than the training loss of 1.2. For the BERT-base-uncased model shown in Figure C-1 (b), the initial training and evaluation losses are in the range of 2.0-2.5, which are smaller than those of the BERT-base-cased model. The losses reduce with decreasing speed and converge to 1.4 and 1.3, respectively for training and validation after the fifteenth epoch of refinement. Compared to the BERT-base-cased model, the losses of the BERT-base-uncased model have lower initial loss values and higher speeds of decreasing loss, albeit their final values are higher as well. The evaluation losses for the BERT-base-cased and BERT-base-uncased models are lower than their training losses and both two models achieved good convergence.

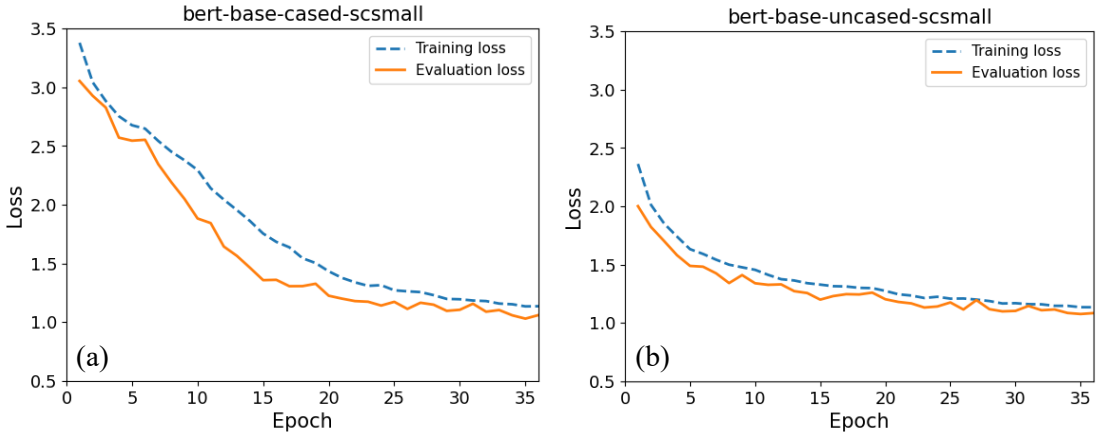

Figure C-1. Convergence profiles of model refinements that further pretrain the originally formulated (a) BERT-base-cased and (b) BERT-base-uncased models when incorporating Solar-cell Corpus Small.

Figure C-2 shows the process of further pretraining BERT-large models when incorporating Solar-cell Corpus Small. The initial values of the training and evaluation losses are in the range of 2.75-3.25 for the BERT-large-cased model, which are higher than those of the BERT-large-uncased model (1.75-2.0). There are two obvious stages when pretraining the BERT-large-cased model: a first stage where the model refines with a convergence rate that initially decreases and then plateaus (epochs 0-13) before the second stage onsets with a constant decreasing speed over epochs 14-25. The losses of the BERT-large-cased model finally reach values between 0.8 and 1.0, which are lower than those of the BERT-large-uncased model (1.0-1.1) by epoch 35. Both the initial and final losses of the BERT-large models are smaller than those of the BERT-base models, albeit the large models require more time to converge. This result shows that the accuracy of a BERT model increases with the size of the model, albeit that the time required for further pretraining increases commensurately.

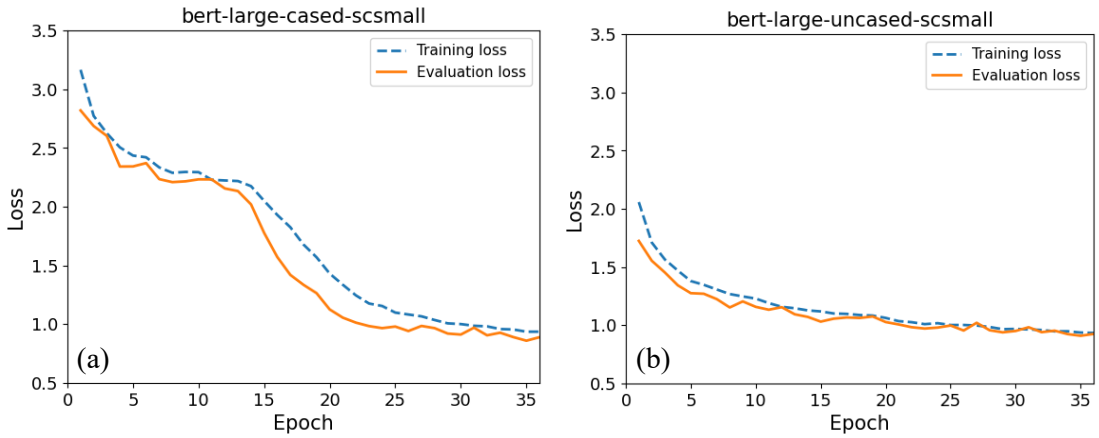

Figure C-2. Convergence profiles of model refinements that further pretrain the originally formulated (a) BERT-large-cased and (b) BERT-large-uncased models when incorporating Solar-cell Corpus Small.

C.2 Further pretraining on Solar-cell Corpus Medium

As shown in Figure C-3, BERT-base models were further pretrained while incorporating Solar-cell Corpus Medium (scmedium). The BERT-base-cased model has initial training and evaluation losses of 2.75-3.0, which are higher than those of the BERT-base-uncased model (1.75-2.0). After ten epochs, both BERT-base-cased and BERT-base-uncased models converge, with final losses being in the ranges of 1.0-1.2 and 1.1-1.3, respectively. The changing rate of convergence for the BERT-base-cased model is higher than that of the BERT-base-uncased model during the first five epochs of model refinement while their convergence profiles are similar over the next five epochs. There is a gap between the training and evaluation losses for both models and the evaluation loss has lower values overall. Compared with the BERT-base models that were further pretrained on Solar-cell Corpus Small, the BERT-base models that were further pretrained on Solar-cell Corpus Medium exhibit lower initial losses and higher convergence speeds, which shows that the increase of the corpus size may not lead to an increase in training time.

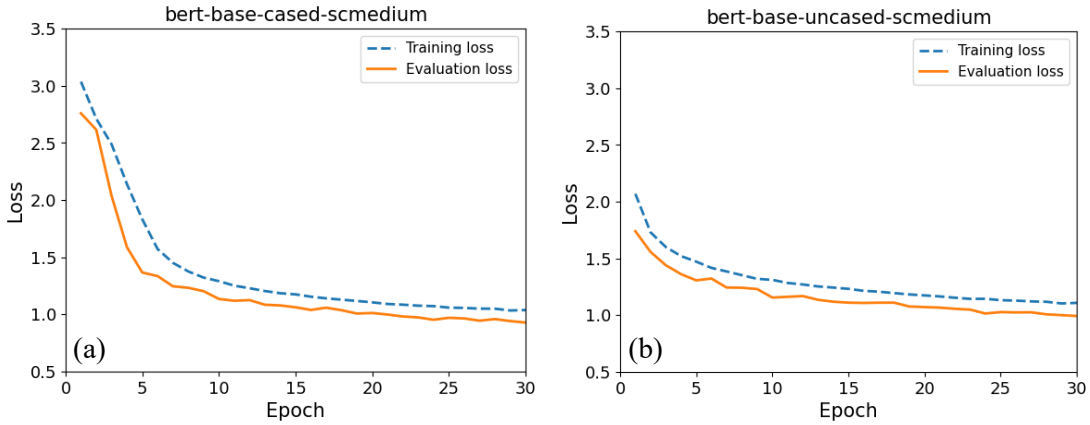

Figure C-3. Convergence profiles of model refinements that further pretrain the originally formulated (a) BERT-base-cased and (b) BERT-base-uncased models when incorporating Solar-cell Corpus Medium.

BERT-large models were also further pretrained when incorporating Solar-cell Corpus Medium. For the BERT-large-cased model, Figure C-4 (a) shows that the initial values of training and evaluation losses (2.5-2.75) decrease to 2.3 during the five epochs of model refinement and rapidly decrease to approximately 0.9 over the next five epochs of refinement. For the BERT-large-uncased model, Figure C-4 (b) shows that the initial training and evaluation losses (1.5-1.75) decrease to around 1.0 after ten epochs of model refinement; these losses have lower initial values and higher final values than those of the BERT-large-uncased model. Both models converge after ten epochs of model refinement and their evaluation losses are lower than their training losses. Compared to BERT-base models that were further pretrained while incorporating Solar-cell Corpus Medium, the BERT-large models that were further pretrained on the same corpus afford lower initial and final losses and need the same number of epochs to achieve sufficient convergence to produce stable models. Compared to BERT-large models that were further pretrained on Solar-cell Corpus Small, the same models further pretrained on Solar-cell Corpus Medium exhibit lower initial losses, similar final losses, and higher convergence speeds.

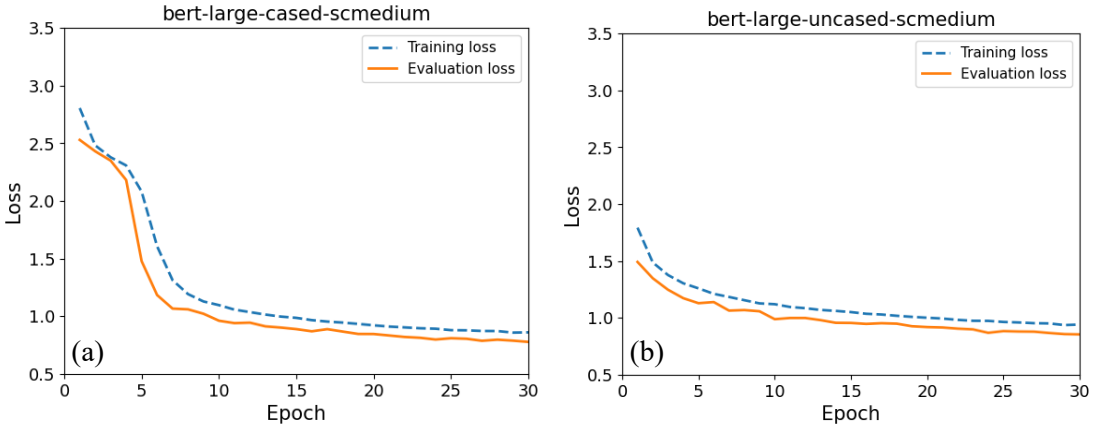

Figure C-4. Convergence profiles of model refinements that further pretrain the originally formulated (a) BERT-large-cased and (b) BERT-large-uncased models while incorporating Solar-cell Corpus Medium.

C.3 Further pretraining on Solar-cell Corpus Large

BERT-base models were further pretrained while incorporating Solar-cell Corpus Large (sclarge). Figure C-5 shows that the training loss of its BERT-base-cased model variant decreases substantially during the first two epochs from 2.4 to 1.5, before gradually reducing to 1.2 after the tenth epoch, whereby the evaluation loss starts from a lower value (1.7) and slowly decreases to 1.5. Note that the training loss is higher than the evaluation loss during epochs 0-2, but decreases at a higher rate such that it becomes lower than the evaluation loss from epoch 3 and continues to decrease at a more rapid pace. This means that the BERT-base-cased model overfits the training set after the second epoch of model refinement. Given that the evaluation loss does not increase, and the corpus is large enough, this overfit is acceptable. For the BERT-base-uncased model, both training and evaluation losses start from a similar value (1.75) and reduce to about 1.6 after one epoch of model refinement. After the tenth epoch, the evaluation loss has decreased to 1.5, while the training loss becomes lower than the evaluation loss and reaches 1.3. The final losses of the BERT-base-cased model are lower than that of the BERT-base-uncased model and overfitting appears in both models.

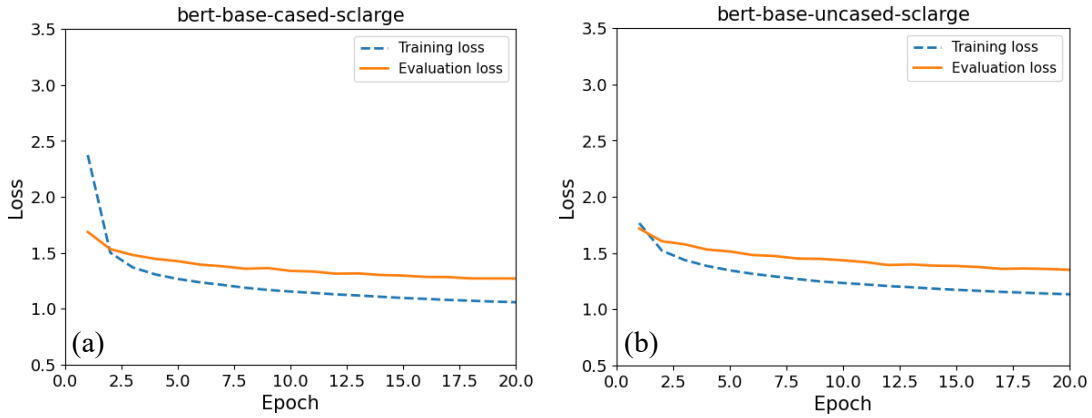

Figure C-5. Convergence profiles of model refinements that further pretrain the originally formulated (a) BERT-base-cased (b) BERT-base-uncased models while incorporating sclarge.

BERT-large models were further pretrained while incorporating Solar-cell Corpus Large. The shapes of the convergence profiles shown in Figure C-6 are similar to the curves in Figure C-5, albeit the losses of the BERT-large models are smaller than those of the BERT-base models. For the BERT-large-cased model, Figure C-6 (a) shows that its training and evaluation losses converge to 1.0 and 1.3, respectively, which are lower than those for the analogous BERT-base-cased model (1.2 and 1.5, respectively).

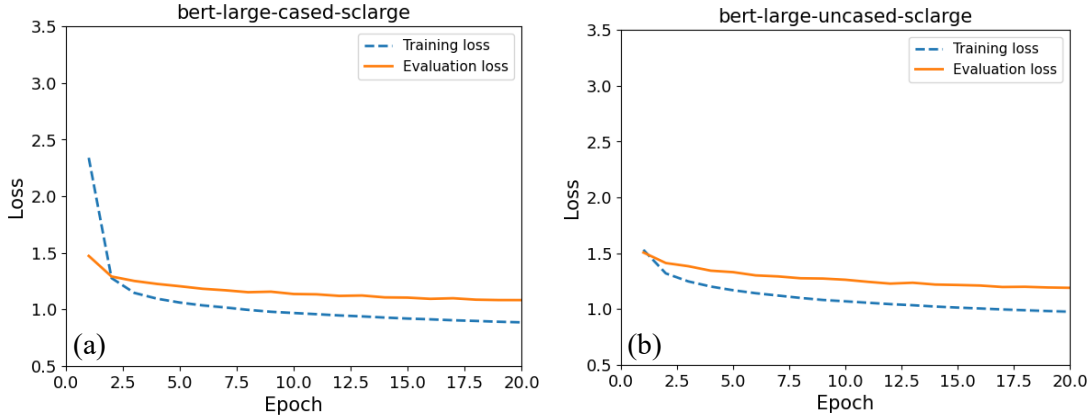

Figure C-6. Convergence profiles of model refinements that further pretrain the originally formulated (a) BERT-large-cased (b) BERT-large-uncased models while incorporating sclarge.

D Finetuning

The hyperparameters for finetuning are: learning_rate: 3e-5; max_seq_length: 512; doc_stride: 128; seed: 50.

D.1 Finetuning on the SQuAD dataset

The originally formulated and further pretrained BERT models were each finetuned on the training set of SQuAD and tested on the validation set of the Solar-cell QA dataset (first-turn); the corresponding results are shown in Table D-1. The performance of a BERT model increases as its F1 score and exact match increase. In more realistic applications, the ability of a BERT model to extract part of the correct answer is also meaningful; accordingly, the F1 score is mainly used instead of the exact match scoring metric for comparing different BERT models. The F1 scores of cased models are higher than those of uncased models except for BERT-base and BERT-large models that were further pretrained on Solar-cell Corpus Large. The difference between the cased and uncased BERT-base models decreases as the incorporated solar-cell corpus becomes larger, to the extent that the F1 score of the BERT-base-uncased-sclarge model finally exceeds that of BERT-base-cased-sclarge. By comparison, the difference between the cased and uncased BERT-large models increases initially, before it decreases, albeit that BERT-large-uncased-sclarge exceeds BERT-large-cased-sclarge as well. For BERT-base models, the F1 scores of the models that were further pretrained while incorporating Solar-cell Corpus Large are the highest and are larger than those of the originally formulated models and the models further pretrained while incorporating smaller corpora. For BERT-large models that were further pretrained while incorporating solar-cell corpora, the trends are similar. The F1 scores of all BERT-base models are smaller than those of their analogous BERT-large models.

Table D-1. F1 scores and exact match scoring metrics for the BERT models finetuned on SQuAD (the maximum value per group is given in bold).

Model	F1 score	Exact match
Base		
bert-base-cased-squad	60.932	52.443
bert-base-uncased-squad	**57.874**	**49.096**
bert-base-cased-scsmall-squad	62.433	55.264
bert-base-uncased-scsmall-squad	60.731	52.035
bert-base-cased-scmedium-squad	63.486	57.234
bert-base-uncased-scmedium-squad	62.363	54.518
bert-base-cased-sclarge-squad	63.684	56.243
bert-base-uncased-sclarge-squad	63.909	56.313
Large		
bert-large-cased-squad	62.0602	55.194
bert-large-uncased-squad	62.038	52.431
bert-large-cased-scsmall-squad	64.057	55.078
bert-large-uncased-scsmall-squad	63.269	54.693
bert-large-cased-scmedium-squad	65.092	57.911
bert-large-uncased-scmedium-squad	64.821	57.619
bert-large-cased-sclarge-squad	65.526	57.736
bert-large-uncased-sclarge-squad	**65.708**	**60.126**

D.2 Finetuning on the combined SQuAD and Solar-cell QA dataset (first-turn)

The originally formulated BERT models and the BERT models that had been further pretrained on solar-cell corpora were finetuned on both SQuAD and the Solar-cell QA dataset (first-turn) (Table D-2). Here, the training set is the combination of the training set in SQuAD and the training set in the Solar-cell QA dataset (first-turn), while the test set is the validation set of

the Solar-cell QA dataset (first-turn). All models have F1 scores of ~76% and the differences among these are small. The bert-large-uncased-sclarge-squadscqa1 model has the highest F1 score (77.106%), while the bert-base-uncased-sclarge-squadscqa1 model has the lowest F1 score (76.14%), noting that the entire range of F1 scores differs by only 0.966%.

Table D-2. F1 scores and exact match scoring metrics for the BERT models finetuned on the combined SQuAD and Solar-cell QA dataset (first-turn) (maximum value per group is in bold).

Model	F1 Score	Exact match
Base		
bert-base-cased-squadscqa1	76.201	**71.459**
bert-base-uncased-squadscqa1	76.349	72.415
bert-base-cased-scsmall-squadscqa1	76.726	73.079
bert-base-uncased-scsmall-squadscqa1	76.452	73.452
bert-base-cased-scmedium-squadscqa1	76.609	73.068
bert-base-uncased-scmedium-squadscqa1	76.423	72.636
bert-base-cased-sclarge-squadscqa1	76.816	73.301
bert-base-uncased-sclarge-squadscqa1	**76.14**	72.659
Large		
bert-large-cased-squadscqa1	76.322	72.916
bert-large-uncased-squadscqa1	76.491	72.823
bert-large-cased-scsmall-squadscqa1	76.836	74.082
bert-large-uncased-scsmall-squadscqa1	76.997	73.907
bert-large-cased-scmedium-squadscqa1	76.605	73.662
bert-large-uncased-scmedium-squadscqa1	76.876	73.779
bert-large-cased-sclarge-squadscqa1	76.499	73.452
bert-large-uncased-sclarge-squadscqa1	**77.106**	**74.467**

D.3 Finetuning on the Solar-cell QA dataset (first-turn)

The importance of domain-specific knowledge was analysed further by finetuning BERT models using only the Solar-cell QA dataset (first-turn). Without the 87,599 QA items from the SQuAD that were derived from general English text, there are only 34,305 QA items from the Solar-cell QA dataset (first-turn) in the training set. When the SQuAD was used to train the BERT models for QA tasks and the Solar-cell QA dataset (first-turn) improves their ability to extract solar-cell properties, both the SQuAD and Solar-cell QA dataset (first-turn) are necessary for finetuning BERT models. However, if a small QA dataset with rich domain knowledge is enough to train the models for property extraction, the SQuAD may not be necessary and the Solar-cell QA dataset (first-turn) could be enough in of itself. In this case, the use of computing resources could be minimized.

This hypothesis was tested by finetuning BERT models using only the Solar-cell QA dataset (first-turn). Results are shown in Table D-3. The training set comprises only the training set of the Solar-cell QA dataset (first-turn) without the SQuAD and the test set is the validation set of the Solar-cell QA dataset (first-turn) exclusively. The F1 scores for the BERT models range from 75.952% to 76.929% and have an average value of 76.46%; i.e., they are all similar to each other (their entire range spans 0.977%). No model exhibits any apparent outstanding performance or obvious weakness.

Table D-3. F1 scores and exact match scoring metrics for the BERT models finetuned on the Solar-cell QA dataset (first-turn) (the maximum value per group are given in bold).

Model	F1 score	Exact match
Base		
bert-base-cased-scqa1	76.726	73.685
bert-base-uncased-scqa1	**75.952**	72.729
bert-base-cased-scsmall-scqa1	76.479	**72.403**
bert-base-uncased-scsmall-scqa1	76.01	72.589
bert-base-cased-scmedium-scqa1	76.154	72.624
bert-base-uncased-scmedium-scqa1	76.029	72.508
bert-base-cased-sclarge-scqa1	76.431	72.578
bert-base-uncased-sclarge-scqa1	76.482	72.764
Large		
bert-large-cased-scqa1	76.449	73.137
bert-large-uncased-scqa1	76.361	73.557
bert-large-cased-scsmall-scqa1	76.18	72.555
bert-large-uncased-scsmall-scqa1	76.415	73.137
bert-large-cased-scmedium-scqa1	76.893	73.872
bert-large-uncased-scmedium-scqa1	76.91	74.047
bert-large-cased-sclarge-scqa1	**76.929**	**74.245**
bert-large-uncased-sclarge-scqa1	76.897	74.094