Perhaps PTLMs Should Go to School – A Task to Assess Open Book and Closed Book QA

Manuel R. Ciosici, Joe Cecil, Alex Hedges, Dong-Ho Lee, Marjorie Freedman, Ralph Weischedel
manuelc@isi.edu, mrf@isi.edu, weisched@isi.edu
Information Sciences Institute, University of Southern California

Abstract

Our goal is to deliver a new task and leaderboard to stimulate research on question answering and pre-trained language models (PTLMs) to understand a significant instructional document, e.g., an introductory college textbook or a manual. PTLMs have shown great success in many question-answering tasks, given significant supervised training, but much less so in zero-shot settings. We propose a new task that includes two college-level introductory texts in the social sciences (American Government 2e) and humanities (U.S. History), hundreds of true/false statements based on review questions written by the textbook authors, validation/development tests based on the first eight chapters of the textbooks, blind tests based on the remaining textbook chapters, and baseline results given state-of-the-art PTLMs.

Since the questions are balanced, random performance should be ~50%. T5, fine-tuned with BoolQ achieves the same performance, suggesting that the textbook's content is not pre-represented in the PTLM. Taking the exam closed book, but having read the textbook (i.e., adding the textbook to T5's pre-training), yields at best minor improvement (56%), suggesting that the PTLM may not have "understood" the textbook (or perhaps misunderstood the questions). Performance is better (~60%) when the exam is taken open-book (i.e., allowing the machine to automatically retrieve a paragraph and use it to answer the question).

1 Introduction

Question answering (QA) is a yardstick for measuring machine understanding performance (Hermann et al., 2015). QA's popularity as an evaluation technique has led to several sub-categories: tasks can require a model to answer questions from either its background knowledge or from a short passage (e.g., SQuAD, Rajpurkar et al., 2016) or with information retrieval to allow the model to search for the answer in a large corpus (e.g., ARC, Clark et al., 2018). Answering can take the form of true/false classification (BoolQ, Clark et al., 2019), multiple-choice, span selection (SQuAD, Rajpurkar et al., 2016), or text generation (TriviaQA, Joshi et al., 2017).

Transformer architectures optimized for specific QA formulations have driven recent progress in question answering. For example, some models target IR-oriented QA (Guu et al., 2020) while others optimize their learning strategy to specific question types (e.g., by optimizing for expected answers to factoid questions, Roberts et al., 2020). While specialization improves performance, it limits generalization. UnifiedQA (Khashabi et al., 2020) takes a step forward by generalizing the architecture and training over multiple data sets with different QA formulations.

Most research assumes that the information necessary to answer questions is either included with the query (e.g., BoolQ, SQuAD 1.1) or that the information was already stored in language models during initial pre-training or a task-specific second pre-training.[1] However, this assumption limits language models relying on massive corpora (Gao et al., 2020; Raffel et al., 2020) to learning oft-repeated facts (Petroni et al., 2019). Valuable, domain-specific information seldom is repeated often enough to be captured by language models. An evaluation of domain-specific knowledge without access to a relevant text is even more challenging as simple strategies like identifying the answer by information retrieval are ineffective. Even reasoning tasks such as ARC (Clark et al., 2018) only target general scientific knowledge and offer large text corpora to aid QA systems.

We propose Learning from Textbooks (LEFT), a new task to classify domain-specific statements drawn from a textbook's review questions as true

[1] For example, Roberts et al. (2020) adjust T5's masking strategy to target named entities as they expect named entities to be parts of answers.

or false using three evaluation configurations. The first configuration tests the ability to answer questions without any domain-specific material (e.g., applying a PTLM with no access to domain-specific knowledge). This setting is equivalent to a person taking the test before taking the class. In the second configuration, a model has access to the textbook's content and may encode the information in the textbook but may not access the textbook during the test; we call this *closed book*. The second configuration tests a model's ability to learn by reading. In the third configuration, which we call *open book*, models can access the textbook during the test. Thus, LEFT supports contrasting QA formulations and reading methods to explore the strengths and weaknesses of various QA approaches. The LEFT data and leaderboard are available at `https://leftleaderboard.isi.edu`.

2 Related Work

Question Answering. Most previous research specializes QA models to target specific question formulations. Question answering with a relevant paragraph often relies on span selection (Rajpurkar et al., 2016; Yang et al., 2015) or simple reasoning (Clark et al., 2019). Previous open-book QA methods first filter a large corpus to a small set of relevant documents using information retrieval (Karpukhin et al., 2020; Robertson and Zaragoza, 2009). The document set then provides context for answering questions (Dhingra et al., 2017; Dunn et al., 2017; Joshi et al., 2017; Nguyen et al., 2016). Conversely, closed-book QA instead requires models to answer using only their implicit knowledge (Roberts et al., 2020). Taking a step towards generalizing QA, UnifiedQA (Khashabi et al., 2020) proposes a unified architecture that answers various question types relying partly on knowledge encoded in its language model.

Knowledge in Pre-trained Language Models. Pre-trained language models (PTLMs) have shown good performance in cloze-style queries (Petroni et al., 2019), fact-checking (Thorne et al., 2018), entity linking (Guo and Barbosa, 2018; Hoffart et al., 2011), and open-domain QA (Joshi et al., 2017; Kwiatkowski et al., 2019; Petroni et al., 2021). However, in most cases, the PTLMs rely on knowledge learned from massive corpora during pre-training. LEFT tests domain-specific knowledge acquired from a textbook, a small corpus of only a few hundreds of thousands of words (see Table 1).

	AG		USH	
	Dev	Test	Dev	Test
Num. chapters	8	9	8	24
Text size (words)	137 620	138 668	89 765	301 860
Num. statements	186	214	274	412

Table 1: Data overview for the two textbooks: American Government 2e (AG) and U.S. History (USH).

Textbook Question Answering. Researchers have explored machine understanding of elementary- and middle-school science textbooks by visual question answering (Gomez-Perez and Ortega, 2020; Kembhavi et al., 2017; Kim et al., 2019) and information retrieval (Clark et al., 2018). While existing textbook QA tasks focus on general knowledge (which can be gained by pre-training on general web corpora), LEFT focuses on domain-specific knowledge. Furthermore, it quantifies pre-trained language models' pre-existing knowledge by requiring that models take the task *before* and *after* reading LEFT's two textbooks.

3 Task Description

Learning from Textbooks (LEFT) contains two machine-readable college-level introductory textbooks and a set of true/false statements manually derived from review questions written by the textbook authors. The task requires that systems based on language models classify the statements *before* and *after* reading the given textbook material to separate what was learned from the book from what was known before reading. "Reading" is any algorithm method that learns from the domain text without storing a copy of the text. To support comparisons with existing QA approaches, LEFT also supports the open-book setting, where a system can use a textbook paragraph when answering.

Our goal is to support testing pre-trained language models, e.g., T5 (Raffel et al., 2020), and also those approaches that extract and store triples during reading (e.g., *<U.S. Declaration of independence; signed; Aug 2, 1776>*). While learning corpora appear in other question answering tasks (e.g., ARC, 14M words, Clark et al., 2018), the text included in LEFT is small and corresponds to the textbook chapters relevant to each question set. The largest text in LEFT contains only 300K words (for details, see Table 1).

LEFT includes two openly licensed[2] college-level introductory textbooks, *American Government 2e* (Krutz, 2019) and *U.S. History* (Corbett, 2014), and true/false statements derived from each book's review questions. We manually rewrote each textbook's multiple-choice review questions into a balanced set of true and false statements.[3,4] We intentionally wrote the statements such that each *true* and *false* pair has high word overlap to deter classification strategies that rely on word overlap with the textbook. We include five sample statements from LEFT in Appendix A and discuss statement correctness in Appendix C.

We measure task performance by *accuracy*. Since the two textbooks are used in teaching college students, we do not release the correct labels (see the Ethical Considerations section). We split each textbook into a *Dev* set consisting of the first eight chapters and a *Test* set consisting of the remaining chapters (see Table 1 for an overview). We allow unlimited submissions to the *Dev* set, but for any submission, we only provide the overall *accuracy* without feedback on which statements were correctly classified. This design decision aims to prevent divulging the correct answers (see the Ethical Considerations section).

LEFT has three evaluation configurations: *(1) Prior-knowledge*; *(2) Closed-book, after reading*; and *(3) Open-book*. **Prior-knowledge** tests the ability to answer questions without any domain-specific material. Language models must rely solely on the knowledge learned from their large pre-training corpora. In the second configuration, **Closed-book, after reading**, models may access the textbook's content and may encode the information in the textbook but may not access the textbook during the test. For each set, models may read the set's corresponding textbook chapters, the entire textbook, or both textbooks. We require that all model submissions to this evaluation configuration also submit to *Prior-knowledge*. Predictions *before* reading (*Prior-knowledge*) quantify the information included in each model through initial pre-training. The change in performance from *Prior-knowledge* to *Closed-book, after reading* illustrates each model's reading effectiveness. In the third configuration, **Open-book**, models can access the textbook or relevant chapter during the test. To support research on open-book question answering, with each statement, we include the textbook paragraph that provides the information necessary to classify the statement. In our experiments, we call this *goldIR*. Thus, LEFT supports contrasting QA formulations and reading methods to explore the strengths and weaknesses of various QA approaches.

4 Results

We illustrate baseline performance on LEFT using two state-of-the-art language models: T5 (Raffel et al., 2020) and GPT-Neo (a GPT-3 architecture, Brown et al., 2020, trained on the open Pile corpus, Gao et al., 2020). We fine-tune the two language models using BoolQ (Clark et al., 2019). Table 2 shows results in LEFT's three evaluation settings: *Prior-knowledge* (out-of-the-box language models fine-tuned on BoolQ), *Closed-book, after reading* (language models with continued light pre-training on LEFT's text content), and *Open-book* (where models have access to the relevant textbook paragraph). Since the *Prior-knowledge* and *Closed-book* settings do not include the relevant paragraph for each question, we adjust fine-tuning to only use BoolQ's questions and ignore its text snippets. In the *Open-book* setting, we consider automatically retrieved textbook paragraphs (using sBERT, Reimers and Gurevych, 2019) and manually identified the relevant paragraphs (gold information retrieval, goldIR). When selecting the relevant textbook content, we select one natural paragraph (i.e., as written by each textbook's authors). However, due to technical limitations imposed by T5's memory consumption, in our experiments, we limit the concatenated statements and paragraphs to a maximum length of 128 word pieces (see Appendix B.1).

4.1 Baseline Results

T5 and GPT-Neo's scores are indistinguishable from the random baseline of 50% in the *Prior-knowledge setting*, suggesting that the textbooks query for information is either *not* present in the two language models or not easily accessible. Con-

[2]Both textbooks are licensed under the Creative Commons Attribution License v4.0 license.

[3]We construct one true and one false statement for each question to obtain a balanced data set. For example, the question *When was the U.S. Declaration of Independence signed? (A)(correct) August 2, 1776 (B) December 2, 1776, (C) August 2, 1746, (D) August 22, 1976* could become *The U.S. Declaration of Independence was signed on August 2, 1776* (true) and *The U.S. Declaration of Independence was signed on August 2, 1746* (false).

[4]For U.S. History's *Dev* set, we also process questions written by a community of instructors.

	American Government 2e		U.S. History	
	Dev (186)	Test (214)	Dev (274)	Test (412)
Prior-knowledge				
T5-3B -ctx	51.08	49.53	50.36	50.00
GPT-Neo 2.7B -ctx	52.69	48.13	51.09	49.27
Closed-book, after reading				
T5 3B +pt -ctx	56.45	52.34	50.73	50.00
GPT-Neo 2.7B +pt -ctx	50.00	55.14	50.73	49.76
Open-book				
T5-3B +ctx +sBERT	60.22	61.21	55.47	59.95
T5-3B +pt +ctx +sBERT	55.91	52.80	52.19	56.31
T5-3B +ctx +goldIR	**71.51**	**74.30**	**68.61**	**68.69**
T5-3B +pt +ctx +goldIR	58.60	63.08	57.66	66.26

Table 2: Baseline accuracy with the current state-of-the-art language models. U.S. History's *Dev* set consists of statements based on the textbook statements and on questions from a community of instructors. In the heading, each set's name is followed by its number of statements. The order of abbreviations reflects the order of operations. All models are fine-tuned with BoolQ; *+/- ctx* – whether we included BoolQ's context during fine-tuning; *+pt* – whether we pre-trained on the relevant textbook chapters.

tinuing each model's pre-training with the relevant textbook parts sometimes helps, but not consistently. The lack of improvement *after* reading is further evidence that the models memorize, but not in beneficial ways, i.e., they can complete sentences but do not learn the subject matter and cannot classify the statements, even after 20 epochs. It also suggests that the closed-book setting represents a new challenge for PTLMs.

Accuracy in the open-book setting is far higher, especially when using goldIR (i.e., a manually selected relevant paragraph). As in the closed book setting, we contrast models using only prior knowledge with models pre-trained on the textbook. Pre-training with the textbook never improves the system's accuracy, suggesting that even in this setting, the models are not learning by reading the textbook. The gap between goldIR- and sBERT-based retrieval suggests that there is room for retrieval-based improvement in the open-book setting. However, even with goldIR, T5 only achieves an accuracy of ~70%, suggesting that paragraph-based QA alone is not solved with existing models.

5 Conclusions & Future Work

There are several natural directions in which we can extend and improve LEFT. We are extending U.S. History's *Test* set as we did with the *Dev* set by including statements based on questions written by a community of instructors. We are also collecting relevant paragraphs for the extra statements. Lastly, we are categorizing the kind of knowledge required to classify each statement to better understand what kinds of knowledge pose the most difficulties.

We draw several conclusions from this work. Foremost, Learning from Textbooks (LEFT) represents a new type of challenge task for PTLMs, contrasted with the much-studied challenges of (1) common sense QA based on prior knowledge, (2) reading comprehension given a paragraph, and (3) QA using large domain-specific corpora, e.g., science at the elementary- or middle-school level. The task is intended to stimulate research on the following dimensions:

1. Zero-shot learning, much as an entering college student could do when studying a textbook,

2. Measuring a system's knowledge before vs. after "reading" the textbook,

3. Capability in both closed-book and open-book question answering,

4. The effect of IR accuracy on task accuracy compared to the system's language understanding performance.

Our baseline studies show that T5 and GPT-Neo thus far are challenged to show improvement after reading the relevant textbook, that open-book evaluation is easier than closed-book (as it is for humans), and that the gating factor in LEFT is understanding the textbook and/or the question rather than paragraph retrieval. The baseline results show there is much room for improvement.

Ethical Considerations

We have reflected on two ethical considerations when creating Learning from Textbooks (LEFT): content and environmental impact.

Content. The two textbooks in LEFT cover topics that include history, race, and politics. OpenStax textbooks follow a set of *Diversity and Representation Development Guidelines*, which aim to "properly represent genders, gender identities, races, cultures, geographies, ethnic backgrounds, disabilities, nationalities, ages, sexual orientations, socio-economic status, and diverse viewpoints".[5] As creators of an NLP task, we do not make any claims, nor do we comment on the topics covered in the two textbooks. Furthermore, we understand that documents as large and complex as textbooks are bound to contain inaccuracies. We invite users with specific content accuracy concerns to consult the official textbook errata included in each textbook's instructor resources.[6]

Releasing labels for the statements in LEFT would indirectly reveal the correct answers for multiple-choice questions in the two textbooks. While both American Government 2e and U.S. History include answer keys, they are incomplete. We believe releasing the correct answers to all multiple-choice questions in the book would be detrimental to the intended primary users of the two textbooks; in other words, it might hinder students' learning. We only used full-time employees compensated according to U.S. law to rewrite the multiple-choice review questions in the two textbooks.

Environmental. We included baseline results based on large pre-trained language models. Strubell et al. (2019) raised concerns about the environmental impact of training deep learning language models. Patterson et al. (2021) pointed out that most of the energy consumption for deep learning language models comes during the initial pre-training. In this work, we limit ourselves to fine-tuning and light continued pre-training of T5 and GPT-Neo. While we do not have information about GPT-Neo's training, T5's training took place in highly efficient data centers whose energy consumption was offset by purchasing electricity from renewable sources (Patterson et al., 2021). For our light pre-training and fine-tuning, we use a machine with four NVIDIA Quadro RTX 8000 fed from California's energy grid. The total computation time for the experiments in this paper is about 500 hours, but this is an informal estimate rather than an accurate measurement.

[5]See *Diversity and Representation Development Guidelines* in the instructor materials for each textbook.

[6]See the *Errata Release Notes* at https://openstax.org/details/books/american-government-2e?Instructor%20resources for *American Government 2e* and https://openstax.org/details/books/us-history?Instructor%20resources for *U.S. History*.

Acknowledgment

This material is based on research supported by DARPA under agreement number N66001-19-2-4032. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

P Corbett. 2014. *U.S. History*. OpenStax College, Houston, Texas.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question an-

swering by search and reading. *arXiv preprint arXiv:1707.03904*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: a new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning. Available at https://github.com/PyTorchLightning/pytorch-lightning.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Jose Manuel Gomez-Perez and Raúl Ortega. 2020. ISAAQ - mastering textbook questions with pre-trained transformers and bottom-up and top-down attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5469–5479, Online. Association for Computational Linguistics.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, volume 28. Curran Associates, Inc.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Daesik Kim, Seonhoon Kim, and Nojun Kwak. 2019. Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3568–3584, Florence, Italy. Association for Computational Linguistics.

Glen Krutz. 2019. *American government 2e*. OpenStax, Rice University, Ann Arbor, MI.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: a human generated machine reading comprehension dataset. In *CoCo@NIPS*.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. *arXiv preprint arXiv:2104.10350*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and

Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

A Sample Statements

Sample statements in LEFT. The first two statements are from American Government 2e, the following three from U.S. History:

- Public goods are available to all without payment.
- In a majoritarian voting electoral system voters select the party of their choice rather than an individual candidate.
- Europeans did not introduce Indians to wampum.
- Philadelphia served as the base for British operations for most of the Revolutionary War.
- The British bombardment of Baltimore inspired The Star-Spangled Banner.

B Training Details

For all light pre-training and fine-tuning, we use a machine with four NVIDIA Quadro RTX 8000 GPUs.

B.1 T5-3B

We implement the model using PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) and Hugging Face's PyTorch Transformers (Wolf et al., 2020). For pre-training and fine-tuning, we use a maximum sequence length of 128. We searched for the best learning rate for our model out of $\{3e{-}5, 1e{-}4, 3e{-}4, 1e{-}3\}$.

	Fine-tuning	Pre-training
Batch size	16	16
Gradient accumulation	1	1
Learning rate	3e-4	1e-3
Num epochs	20	
Optimizer	AdamW	
$\beta 1$	0.9	
$\beta 2$	0.999	
ϵ	1e-8	
Weight decay	0.0	
Scheduler	WarmupDecayLR	
Warmup max steps	400	
fp16	no	

Table 3: Hyperparameters for T5-3B.

B.2 GPT-Neo 2.7B

We use GPT-Neo 2.7B from the Hugging Face Model Hub.[7] GPT-Neo matches the architecture of GPT-3 (Brown et al., 2020), but is trained on the openly available Pile corpus (Gao et al., 2020).

	Fine-tuning	Pre-training
Batch size	48	2
Gradient accumulation	1	4
Num epochs	10	
Optimizer	AdamW	
$\beta 1$	0.9	
$\beta 2$	0.999	
ϵ	1e-8	
Weight decay	0.01	
Scheduler	WarmupDecayLR	
Warmup max steps	200	
fp16	yes	

Table 4: Hyperparameters for GPT-Neo.

C Ensuring Statement Correctness

We took several steps to ensure statements' true/false correctness and prevent data bias/tells. For true/false correctness, we manually inspected the statements to check that they correspond to the correct and incorrect choices as given by each textbook's instructor material. We then wrote a script to automatically count the statements for each chapter to ensure that there are as many true labels as there are false. If some labels were to change accidentally during our research, the script would detect the change. For the manually retrieved relevant passages, the humans read each statement and identified the relevant paragraph. In the process, they also checked each statement's label.

To prevent data bias, we wrote statement pairs to have as much word overlap as logically and grammatically possible. We used multiple annotators to write the statements for the two textbooks (two native speakers for U.S. History; one native, one fluent non-native for American Government 2e). No partition is composed of statements written exclusively by a single person, ensuring no person-specific tells. Following that, we checked all statements for grammar and punctuation issues using automated checkers and another annotator reading. This stage deals with copy-paste tells in the data and cases where statements for one label sound unnatural.

[7] https://huggingface.co/EleutherAI/gpt-neo-2.7B