

SCROLLS: Standardized CompaRison Over Long Language Sequences

Uri Shaham^τ Elad Segal^τ Maor Ivgi^τ Avia Efrat^τ
 Ori Yoran^τ Adi Haviv^τ Ankit Gupta^β Wenhan Xiong^μ Mor Geva^{τ,α}
 Jonathan Berant^τ Omer Levy^{τ,μ}

^τ Tel-Aviv University
^α Allen Institute for AI
^β IBM Research
^μ Meta AI

Abstract

NLP benchmarks have largely focused on *short* texts, such as sentences and paragraphs, even though *long* texts comprise a considerable amount of natural language in the wild. We introduce SCROLLS, a suite of tasks that require reasoning over long texts. We examine existing long-text datasets, and handpick ones where the text is naturally long, while prioritizing tasks that involve synthesizing information across the input. SCROLLS contains summarization, question answering, and natural language inference tasks, covering multiple domains, including literature, science, business, and entertainment. Initial baselines, including Longformer Encoder-Decoder, indicate that there is ample room for improvement on SCROLLS. We make all datasets available in a unified text-to-text format and host a live leaderboard to facilitate research on model architecture and pretraining methods.¹

1 Introduction

Standard benchmarks à la GLUE (Wang et al., 2018, 2019), WMT (Barrault et al., 2019, 2020), and SQuAD (Rajpurkar et al., 2016, 2018), have driven progress in natural language processing of *short* utterances. However, a large portion of natural language is produced in the context of *longer* discourses, such as books, articles, meeting transcripts, etc. To tackle the computational challenges associated with processing such long sequences, a plethora of new model architectures have recently emerged (Tay et al., 2020b; Fournier et al., 2021), *without* establishing a standard scheme for evaluating them on long natural language problems. Some long-context models are evaluated via language modeling perplexity, but this metric mostly captures model sensitivity to local, short-range patterns (Khandelwal et al., 2018; Sun et al.,

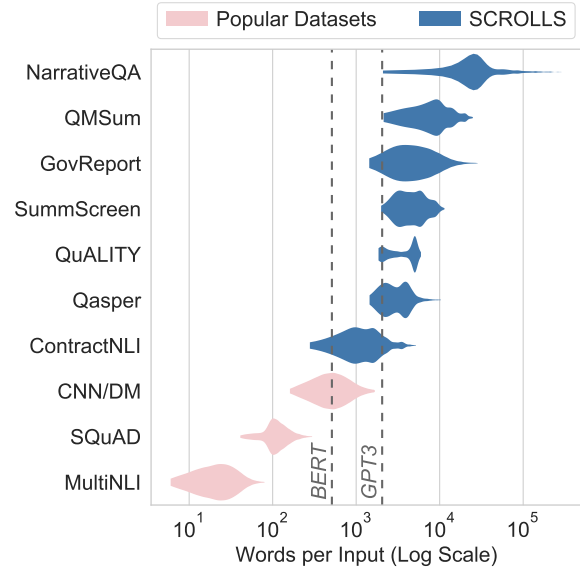


Figure 1: The distribution of words per input in SCROLLS datasets (blue), alongside frequently-used NLP datasets (pink). Dashed vertical lines indicate the maximal sequence length (in tokens) of BERT (Devlin et al., 2019) and GPT3 (Brown et al., 2020).

2021). Other studies rely on Long Range Arena (Tay et al., 2021), which is limited from a natural-language perspective, since only two of its datasets involve natural language, and those are artificially-elongated through byte tokenization. To enable the research community to go beyond sentences and paragraphs, we present a new benchmark, **SCROLLS: Standardized CompaRison Over Long Language Sequences**.

SCROLLS incorporates multiple tasks (summarization, question answering, and natural language inference) over various domains (literature, meeting transcripts, TV shows, scientific articles, and more), *where each example’s input typically contains thousands of words*. We review the existing literature on long-text tasks and manually curate a subset of 7 datasets, prioritizing those that require contextualizing and abstracting information

¹<https://www.scrolls-benchmark.com>

across multiple parts of the text. We then clean and convert the data to a unified text-to-text format to enable the evaluation of a single model over all datasets. Figure 1 shows that SCROLLS datasets are substantially longer than commonly-used NLP benchmarks.

SCROLLS is available via the Datasets library (Lhoest et al., 2021) or direct download on its website, which hosts a live leaderboard that accepts submissions and automatically evaluates them against private test sets. By producing a single aggregate score, in addition to individual dataset scores, SCROLLS can serve as an evaluation platform for future approaches to processing long text, whether by new pretraining schemes, novel transformer architectures and alternatives, or even retrieval-based methods. We provide initial baselines for SCROLLS using two transformer models, BART (Lewis et al., 2020), and its length-efficient variant, Longformer Encoder-Decoder (Beltagy et al., 2020). Our experiments indicate that SCROLLS poses a formidable challenge for these models, leaving much room for the research community to improve upon.

2 Background: Contemporary Evaluation of Long-Text Models

While transformers (Vaswani et al., 2017) are the current go-to architecture for building state-of-the-art models in NLP, they present a computational challenge when it comes to long sequences due to the $O(n^2)$ complexity of self-attention, where n is the sequence’s length. To address this problem, a wide variety of efficient alternatives and approximations have been proposed over the past couple of years (Tay et al., 2020b; Fournier et al., 2021). Much of these novel architectures were developed concurrently, leading to somewhat of a “Wild West” when it comes to model evaluation, making cross-model comparison challenging. Roughly speaking, we can cluster the more prominent evaluation methodologies into three categories: language modeling, Long-Range Arena, and summarization.

The language modeling community typically uses perplexity to measure how well models predict the next token, a practice that has been adopted by several works on efficient transformer architectures (Roy et al., 2021; Choromanski et al., 2020; Tay et al., 2020a; Peng et al., 2021). However, using perplexity to evaluate a model’s *long-range* abilities is currently under scrutiny. A growing

amount of literature shows that predicting the next token is mostly a local task that does not require modeling long-range dependencies (Khandelwal et al., 2018; Sun et al., 2021), and that masking or down-weighting distant tokens can actually *improve* perplexity (Press et al., 2021a,b).

A more recent approach to standardizing long-sequence model evaluation is the Long Range Arena (LRA) (Tay et al., 2021). It incorporates 5 classification datasets: byte-level sentiment analysis (IMDB) and document relatedness (ACL Anthology); path-finding (Pathfinder) and image classification (CIFAR-10) over 1-dimensional pixel sequences; and executing a list of mathematical operations (ListOps). Of those, two involve visual reasoning, and one is a synthetic mathematical language (ListOps), leaving only two natural language datasets (sentiment analysis and document relatedness). The multi-modal nature of LRA makes it inappropriate as a testbed for pretrained language models, limiting its relevance for NLP. Moreover, LRA artificially inflates natural language sequences via byte tokenization, and truncates each example at 4,000 bytes, which is equivalent to less than 1,000 words. This exempts models from coping with the complex long-range dependencies that exist in naturally long texts.

The third practice uses summarization tasks to evaluate long-sequence models. The most popular datasets use abstracts of academic papers on Arxiv and PubMed (Cohan et al., 2018) as summaries. Other summarization datasets, however, are less frequently used, biasing the evaluation towards academic domains. SCROLLS includes summarization as one of its main tasks, selecting datasets from several different domains to increase diversity.

3 The SCROLLS Benchmark

SCROLLS aims to challenge a model’s ability to process long texts in the wild, and therefore focuses on discourses that are *naturally* long, encompassing domains such as literature, TV show scripts, scientific articles, and more. We review the datasets in existing literature, seeking ones that challenge models not only by the length of each input, but also by the need to contextualize across different sections and process long-range dependencies. At the same time, we strive to maintain a diversity of tasks, covering summarization and query-based summarization, open ended and multiple-choice

Dataset	Task	Domain	Metric	Avg #Words		#Examples
				Input	Output	
GovReport (Huang et al., 2021)	Summ	Government	ROUGE	7,897	492.7	19,402
SummScreenFD (Chen et al., 2021)	Summ	TV	ROUGE	5,639	100.0	4,348
QMSum (Zhong et al., 2021)	QB-Summ	Meetings	ROUGE	10,396	69.7	1,810
Qasper (Dasigi et al., 2021)	QA	Science	F1	3,671	11.5	5,692
NarrativeQA (Kočíský et al., 2018)	QA	Literature, Film	F1	51,790	4.6	71,187
QuALITY (Pang et al., 2021)	MC-QA	Literature, Misc	EM	4,198	10.3	6,737
ContractNLI (Koreeda and Manning, 2021)	NLI	Legal	EM	1,708	1.4	10,319

Table 1: An overview of the datasets in SCROLLS and their statistics. *Summ* refers to summarization, *QB-Summ* means query-based summarization, and *MC-QA* abbreviates multiple-choice question answering. The number of examples includes train, validation, and test sets.

question answering, as well as natural language inference.

Through this curation process, we handpick 7 datasets, and process them into a uniform text-to-text format. Table 1 provides an overview of the datasets included in SCROLLS. Figure 2 and Figure 3 show two examples from SCROLLS datasets SummScreenFD and QuALITY, demonstrating how contextualizing and synthesizing information over long ranges of text is paramount to addressing the challenges in the benchmark.

3.1 Datasets

We survey the 7 datasets in SCROLLS, and elaborate how the original data was collected.

GovReport (Huang et al., 2021): A summarization dataset of reports addressing various national policy issues published by the Congressional Research Service² and the U.S. Government Accountability Office,³ where each document is paired with an expert-written executive summary. The reports and their summaries are longer than their equivalents in other popular long-document summarization datasets; for example, GovReport’s documents are approximately 1.5 and 2.5 times longer than the documents in Arxiv and PubMed (Cohan et al., 2018), respectively.

SummScreenFD (Chen et al., 2021): A summarization dataset in the domain of TV shows (e.g. Friends, Game of Thrones). Given a transcript of a specific episode, the goal is to produce the episode’s recap. The original dataset is divided into two complementary subsets, based on the source of its community contributed tran-

scripts. For SCROLLS, we use the ForeverDreaming (FD) subset,⁴ as it incorporates 88 different shows, making it a more diverse alternative to the TV MegaSite (TMS) subset,⁵ which has only 10 shows. Community-authored recaps for the ForeverDreaming transcripts were collected from English Wikipedia and TVMaze.⁶

QMSum (Zhong et al., 2021): A query-based summarization dataset, consisting of 232 meetings transcripts from multiple domains and their corresponding summaries. The corpus covers academic group meetings at the International Computer Science Institute (Janin et al., 2003),⁷ industrial product meetings for designing a remote control (Carletta et al., 2005), and committee meetings of the Welsh⁸ and Canadian⁹ Parliaments, dealing with a variety of public policy issues. Annotators were tasked with writing queries about the broad contents of the meetings, as well as specific questions about certain topics or decisions, while ensuring that the relevant text for answering each query spans at least 200 words or 10 turns.

Qasper (Dasigi et al., 2021): A question answering dataset over NLP papers filtered from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020). Questions were written by NLP practitioners after reading only the title and abstract of the papers, while another set of NLP practitioners annotated the answers given the entire document. Qasper contains abstractive, extractive, and yes/no questions, as well as unanswerable ones.

⁴<http://transcripts.foreverdreaming.org>

⁵<http://tvmegasite.net/>

⁶<https://www.tvmaze.com/>

⁷<https://groups.inf.ed.ac.uk/ami/icsi/index.shtml>

⁸<https://record.assembly.wales>

⁹<https://www.ourcommons.ca/Committees/en/Home>

²<https://crsreports.congress.gov/>

³<https://www.gao.gov/>

Penny returns from visiting family in Nebraska, but mentions while picking up mail from Leonard that most of her relatives became sick. Sheldon, a germophobe according to Leonard, freaks out and becomes sick, becoming demanding on top of his already obnoxious personality. Familiar with Sheldon being sick, Leonard and the guys hide from him at a Planet of the Apes series marathon, leaving Penny to care for Sheldon. However, Leonard breaks his glasses in the cinema and has to retrieve his spare pair from the apartment, piloted by Howard and Raj using a laptop, an endoscope, and a Bluetooth helmet camera worn by the short-sighted Leonard. Penny intercepts him and abandons him to his fate with Sheldon. Leonard tries to escape, but runs into a wall and nearly knocks himself out. In the end, injured Leonard and sick Sheldon sit miserably on the couch.

— Transcript —

...[1,032 words]...

Howard: Hello.

Sheldon: Howard, I'm sick.

...[40 words]...

Howard: It's my own fault, I forgot the protocol we put in place after the great ear infection of '06.

Leonard: You call Koothrappali, we need to find a place to lay low for the next eighteen to twenty four hours.

Howard: Stand by. Ma, can my friends come over?

Howard's Mother: I just had the carpets steamed.

Howard: That's a negatory. But there's a Planet of the Apes marathon at the New Art today.

Leonard: Five movies, two hours apiece. It's a start.

...[660 words]...

Sheldon: Based on what happened next, I assume it means "would you like an enema?"

Penny: Okay, sweetie, I'll take care of you, what do you need?

...[766 words]...

Penny: You deliberately stuck me with Sheldon.

Leonard: Well, I had to, you see what he's like.

...[142 words]...

Figure 2: An example from the SummScreenFD summarization dataset, where the task is to generate the recap (top paragraph) given the episode's script. In this example, the information required to compose the third sentence in the recap (highlighted) is scattered across several snippets throughout the transcript.

NarrativeQA (Kočiský et al., 2018): An established question answering dataset over entire books from Project Gutenberg¹⁰ and movie scripts from different websites.¹¹ Annotators were given summaries of the books and scripts obtained from Wikipedia, and asked to generate question-answer pairs, resulting in about 30 questions and answers for each of the 1,567 books and scripts. They were encouraged to use their own words rather than copying, and avoid asking yes/no questions or ones about the cast. Each question was then answered

¹⁰<http://www.gutenberg.org>

¹¹<http://www.imsdb.com>, <http://www.dailyscript.com/>, <http://www.awesomefilm.com>

The text says "The expert frowned horribly." What makes the expert's smile so horrible?

(A) The frown indicates that he's close to detecting Korvin's true motivations.

(B) The frown indicates that he knows that Korvin switched the wires on the lie detector.

(C) The frown is a signal to the Ruler that Korvin is lying.

(D) The frown is physically horrible because the Tr'en have fifty-eight, pointed teeth.

— Story —

...[607 words]...

It was a ritual, Korvin had learned. "You are of the Tr'en," he replied. The green being nodded. "I am Didyak of the Tr'en," he said.

...[257 words]...

Didyak beamed at him. The sight was remarkably unpleasant, involving as it did the disclosure of the Tr'en fifty-eight teeth, mostly pointed. Korvin stared back impassively. "I have been ordered to come to you," Didyak said, "by the Ruler. The Ruler wishes to talk with you."

...[1,366 words]...

"They can be treated mathematically," one of the experts, a small emerald-green being, told Korvin thinly. "Of course, you would not understand the mathematics."

...[33 words]...

The expert frowned horribly, showing all of his teeth. Korvin did his best not to react. "Your plan is a failure," the expert said, "and you call this a good thing."

...[1,808 words]...

Figure 3: An example from the QuALITY dataset, where the task is to answer multiple-choice questions about a given story or document. In this example, answering the question correctly requires reasoning over four different snippets that are separated by long token sequences.

by an additional annotator, providing each question with two reference answers (that may be identical).

QuALITY (Pang et al., 2021): A multiple-choice question answering dataset over stories and articles sourced from Project Gutenberg,¹⁰ the Open American National Corpus (Fillmore et al., 1998; Ide and Suderman, 2004), and more. Experienced writers wrote questions and distractors, and were incentivized to write answerable, unambiguous questions such that in order to correctly answer them, human annotators must read large portions of the given document. To measure the difficulty of their questions, Pang et al. conducted a speed validation process, where another set of annotators were asked to answer questions given only a short period of time to skim through the document. As a result, 50% of the questions in QuALITY are labeled as *hard*, i.e. the majority of the annotators in the speed validation setting chose the wrong answer.

Contract NLI (Koreeda and Manning, 2021): A natural language inference dataset in the legal domain. Given a non-disclosure agreement (NDA, the premise), the task is to predict whether a particular legal statement (the hypothesis) is entailed, not entailed (neutral), or cannot be entailed (contradiction) from the contract. The NDAs were manually picked after simple filtering from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR)¹² and Google. The dataset contains a total of 607 contracts and 17 unique hypotheses, which were combined to produce the dataset’s 10,319 examples.

3.2 Preprocessing

Data Cleansing As part of the curation process, we examine each dataset and clean or filter examples to ensure high quality data. In GovReport, we discard all examples where the report’s length (in words) is less than twice the summary, or more than 1,000 times the summary, as well as examples where the summary exists verbatim in the report. This process removes 64 examples from the original dataset. In Qasper, we discard all papers that have less than 8,192 characters, removing a total of 176 questions over 63 papers, which appear to be of lower quality. In NarrativeQA, we locate markers indicating the start and end of the actual story, and use them to remove excess metadata such as licenses, HTML headers, etc.

Unified Format We reformulate every dataset in SCROLLS as a sequence-to-sequence task to allow for a simple unified input-output format. When a query is given in addition to the raw text (as in QMSum, Qasper, NarrativeQA, QuALITY, and ContractNLI), we prepend it to the text, using two newlines as a natural separator. For the multiple-choice dataset QuALITY, we also provide all four answer candidates as part of the query. For the summarization datasets, GovReport and SummScreenFD, we use only the original documents as input. Some datasets (Qasper and NarrativeQA) contain multiple target outputs for each input; we split them into separate instances for training and development. For test, we score each prediction with every valid answer independently, and then merge the scores of identical inputs by taking the maximum of those scores. Table 2 provides an example from each SCROLLS dataset.

¹²<https://www.sec.gov/Archives/edgar/Oldloads>

3.3 Evaluation

Each dataset is split into training, validation, and test sets based on the original dataset splits. In SCROLLS, test set outputs are kept private, and only the inputs are publicly available. When evaluating a model, users must submit their model’s outputs for *all* test sets via the SCROLLS website. Once a model is submitted, we compute the average performance metric across all datasets to provide the submission with a single aggregate SCROLLS score. We employ three different evaluation metrics across SCROLLS datasets: ROUGE for summarization tasks (GovReport, SummScreenFD, and QMSum), unigram overlap (F1) for question answering (Qasper and NarrativeQA), and exact match (EM) for multiple-choice (QuALITY) and classification (ContractNLI) tasks. The official evaluation script is available online.¹³

ROUGE We use three flavors of ROUGE (Lin, 2004) to measure the overlap between the system-generated output and the reference: unigram overlap (ROUGE-1), bigram overlap (ROUGE-2), and the longest overlapping subsequence (ROUGE-L). Both system output and reference are normalized by lowercasing and converting all non-alphanumeric characters to whitespaces, followed by whitespace tokenization. We compute the geometric mean of the three scores (ROUGE-1/2/L) to produce a single score per dataset, which is used to calculate the final SCROLLS score.¹⁴

F1 Similar to ROUGE-1, the F1 metric calculates unigram overlap. The key difference is that both reference and system output strings are normalized slightly differently; in addition to lowercasing and punctuation removal, stopwords are also discarded, following the practice of SQuAD (Rajpurkar et al., 2016) and other question-answering datasets (Fisch et al., 2019). Both Qasper and NarrativeQA contain questions with more than one reference answer; for each such example, we take the maximal F1 score over all of its reference answers.

¹³<https://github.com/tau-nlp/scrolls>

¹⁴Although there are issues with using ngram overlap metrics (such as ROUGE) to evaluate generated text, our assessment at the time of writing is that, in the context of summarization, the alternatives have not yet matured. We may decide to replace or complement ROUGE with model-based evaluation in the future.

GovReport <i>Summarization</i>	
<i>Input</i>	Introduction The United States has an abundance of natural resources. For much of the nation’s history, energy availability was not a concern as commerce and industry needs could be met by domestic supplies. However, industrialization and population growth, and the continuing development of a consumer-oriented society, led to growing dependence...
<i>Output</i>	Energy is crucial to the operation of a modern industrial and services economy. Concerns about the availability and cost of energy and about environmental impacts of fossil energy use have led to the establishment of...
SummScreenFD <i>Summarization</i>	
<i>Input</i>	Ted’s kitchen Ted from 2030: Kids, when it comes to love, the best relationships are the ones that just come naturally. Ted: My first solo batch. Victoria: Um, I think those need to stay in the oven a while longer. Here’s a professional tip. If it’s still runny, it’s not a cupcake. It’s a beverage...
<i>Output</i>	Just as things are going well between Ted and Victoria, the latter is offered a surprising but incredible opportunity to be a fellow at a culinary institute in Germany. As the couple discuss the viability of long-distance...
QMSum <i>Query-Based Summarization</i>	
<i>Input</i>	What did the team discuss during the product evaluation about its feature to solve customers’ concerns? Project Manager: Yep. Soon as I get this. Okay. This is our last meeting. Um I’ll go ahead...
<i>Output</i>	Generally speaking, the team agreed that the product was intuitive and had successfully incorporated main aims that the team had. The team believed the customers were not likely to lose the remote control since it was...
Qasper <i>Question Answering</i>	
<i>Input</i>	Which languages are used in the multi-lingual caption model? Introduction The bilingual lexicon induction task aims to automatically build word translation dictionaries across different languages, which is beneficial for various natural language processing tasks such as cross-lingual information...
<i>Output</i>	German-English, French-English, and Japanese-English
NarrativeQA <i>Question Answering</i>	
<i>Input</i>	What is the first heist that Dignan and Anthony commit? BOTTLE ROCKET screenplay by Wes Anderson and Owen Wilson EXT. ALLEY. DAY ANTHONY and DIGNAN walk down an alley behind a convenience store. Anthony’s nineteen. He’s got on a...
<i>Output</i>	As a practice heist they break into Anthony’s family’s home.
QuALITY <i>Multiple-Choice Question Answering</i>	
<i>Input</i>	Why did the beings come to Earth? (A) it was the next planet for them to destroy (B) they wanted all of Earth’s resources (C) they wanted to take over Earth (D) they were curious about Earth’s creatures "Phone Me in Central Park" By JAMES McCONNELL There should be an epitaph for...
<i>Output</i>	it was the next planet for them to destroy
ContractNLI <i>Natural Language Inference</i>	
<i>Input</i>	Agreement shall not grant Receiving Party any right to Confidential Information. NON-DISCLOSURE AND CONFIDENTIALITY AGREEMENT This NON-DISCLOSURE AND CONFIDENTIALITY AGREEMENT (“Agreement”) is made by and between: (i) the Office of the United Nations High Commissioner...
<i>Output</i>	Entailment

Table 2: An example from each one of the SCROLLS datasets, shown in the benchmark’s text-to-text format. In this illustration, we truncate the examples’ inputs and outputs for brevity.

EM Exact match normalizes the output strings using the same procedure as F1 (lowercasing, removing punctuation and stopwords, and normalizing whitespaces), and then compares whether the two normalized strings are identical. For QuALITY, we calculate EM over the entire test set, and also EM over its subset of *hard* questions, as defined in the original dataset. For computing the final SCROLLS score, however, we only use the EM value calculated over the full test set.

4 Experiments

We conduct experiments to evaluate the ability of mainstream models to handle the various long text challenges presented by SCROLLS. Our code is based on the Transformers library (Wolf et al., 2020), and is available online.¹³

4.1 Baselines

We finetune two pretrained transformer variants as baselines, as well as naive heuristic baselines to establish the floor performance on each task.

BART As a standard transformer baseline, we use the pretrained BART-base¹⁵ model (Lewis et al., 2020). BART is a transformer encoder-decoder pretrained by reconstructing noised texts, which achieved state-of-the-art results on several summarization datasets when released. BART was pretrained on sequences of up to 1,024 tokens; we therefore truncate all inputs by retaining only their 1,024-token prefix. To examine the effect of available input length, we also consider truncating BART’s inputs at 256 and 512 tokens.

Longformer Encoder-Decoder (LED) We experiment with LED-base,¹⁶ the encoder-decoder version of the efficient transformer architecture Longformer (Beltagy et al., 2020). Longformer avoids computing the quadratic-complexity attention matrix via sliding-window attention, where each word only attends to a constant numbers of nearby tokens, on top of a few tokens that compute global attention over the entire input. LED is initialized with BART’s parameters, without further pretraining. In our experiments, we use a sliding window of 1,024 tokens, and restrict the total

input length to 16,384 tokens via truncation, following Beltagy et al. We also experiment with maximum sequence lengths of 1,024 and 4,096 tokens. While the original work on LED selects the globally-attending tokens on a per-task basis, we follow their summarization setting throughout all tasks (for uniformity), which enables global attention only for the first token.

Heuristic Baselines We use simple heuristics to find the lower bound of performance on each dataset. For most datasets, we use the fixed-length prefix heuristic, akin to the LEAD baseline in the summarization literature. Specifically, we compute the average output-input length ratio ρ over the training set (in characters), and then produce the first $\rho \cdot n$ characters from the given input at inference time (where n is the input’s length in characters). For QuALITY, we use the majority class (which is just above one quarter). For ContractNLI, we use the per-hypothesis majority class, as the same 17 hypotheses are shared across all documents.

4.2 Hyperparameters

We finetune each of the baseline models on every dataset separately, using mixed precision and gradient checkpointing, with an effective batch size of 131,072 (2^{17}) tokens. The summarization datasets are trained for 10 epochs, while Qasper, QuALITY, and ContractNLI are trained for 20; NarrativeQA (the largest dataset) is trained for 2 epochs. We tune the maximum learning rate over each validation set, selecting from 6 possible values: 1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4. The learning rate is warmed up from zero during the first 10% steps, and then linearly decays back to zero throughout the remaining 90%. We also apply 0.1 dropout throughout each network. During inference, we generate outputs using greedy decoding.

4.3 Results

Table 3 shows the baselines’ performance on SCROLLS. A few trends are apparent:

More Context Improves Performance Within each pretrained model, we experiment with three context lengths. As the model receives more context, its average SCROLLS score increases. For BART, increasing the input length from 256 tokens to 1,024 increases performance by 2.66 points,

¹⁵<https://huggingface.co/facebook/bart-base>

¹⁶<https://huggingface.co/allenai/led-base-16384>

Model (Input)		GovRep ROUGE-1/2/L	SumScr ROUGE-1/2/L	QMSum ROUGE-1/2/L	Qspr F1	Nrtv F1	QALT EM-T/H	CNLI EM	Avg
Naive	-	45.3 / 17.9 / 20.8	19.6 / 1.8 / 11.0	14.2 / 2.0 / 9.3	3.4	1.5	25.2 / 26.1	66.0	19.35
BART	256	41.9 / 14.2 / 20.3	24.5 / 3.8 / 15.3	29.9 / 8.3 / 20.4	23.3	14.0	26.0 / 25.8	69.8	26.35
	512	45.6 / 16.9 / 21.8	26.3 / 5.1 / 16.2	29.5 / 8.2 / 20.1	24.7	14.5	26.8 / 27.4	71.6	27.58
	1024	47.9 / 18.6 / 22.7	27.2 / 4.9 / 16.7	30.2 / 8.7 / 20.7	26.3	15.4	26.0 / 25.9	77.4	29.01
LED	1024	40.9 / 16.1 / 23.1	22.7 / 3.6 / 15.1	24.6 / 6.5 / 19.0	24.4	15.2	26.6 / 27.2	73.4	27.06
	4096	52.5 / 23.3 / 26.8	23.0 / 4.1 / 15.1	26.6 / 6.9 / 19.9	25.0	16.3	26.6 / 27.3	71.5	28.30
	16384	56.2 / 26.6 / 28.8	24.2 / 4.5 / 15.4	25.1 / 6.7 / 18.8	26.6	18.5	25.8 / 25.4	71.5	29.16

Table 3: Baseline results on SCROLLS, using naive heuristics, BART, and Longformer Encoder-Decoder (LED), and various input length limits. The final SCROLLS score (Avg) is computed by averaging over each dataset’s overall performance score. For QuALITY (QALT), we use the EM score calculated over the full test set (EM-T), without up-weighting the performance on the hard subset (EM-H). For datasets evaluated with ROUGE, we aggregate the different ROUGE scores via geometric mean to produce a single score per dataset.

while LED grows by 2.1 points when enlarging its maximal sequence length from 1,024 tokens to 16,384. This improvement is relatively consistent across datasets for BART, but less so for LED (e.g., QMSum and ContractNLI).

BART versus LED Although LED does achieve the highest SCROLLS score when given 16,384 tokens per sequence, BART arrives within 0.15 points of the top score *despite being limited to only 1,024 tokens*. This is surprising, given the substantial difference in input lengths. Moreover, when controlling for the number of tokens, BART outperforms LED by almost two points, suggesting that LED might be under-optimized. Inspecting the dataset-level results reveals that LED (16k) significantly outperforms BART (1k) in two datasets, GovReport and NarrativeQA, which are coincidentally the largest datasets in SCROLLS by number of examples. Thus, it is possible that since LED is initialized with BART’s parameters (without long-text pretraining), it requires a substantial amount of data and fine-tuning to adapt the parameters to sliding window attention and potentially longer inputs.

Overall, our experiments highlight the importance of measuring not only whether an architecture can efficiently process a long language sequence, but also whether it can effectively model long-range dependencies. This is exactly what SCROLLS is designed to do.

How Far is SCROLLS from being Solved? The heuristic baselines set a lower bound average score of 19.35, which the model baselines are able to improve upon by 7 to 10 points. While it is difficult to establish an accurate human performance ceil-

ing on SCROLLS, especially when considering the summarization datasets, we do have some indicators that it is probably much higher than the current baselines. Dasigi et al. (2021) study a subset of Qasper that has multiple annotated answers, and find their overlap to be 60.9% F1, more than double our best baseline. Likewise, human agreement on QuALITY was measured at 93.5% EM (Pang et al., 2021). We also compute the inter-annotator agreement (F1) on NarrativeQA’s test set (where each question has two answers), arriving at around 58.7% F1, compared to our best baseline of 18.5% F1. Overall, it seems that contemporary off-the-shelf models struggle with these tasks, challenging future work to make progress on SCROLLS.

5 Conclusion

We propose a new benchmark that places the spotlight on naturally long texts and their intricacies. SCROLLS fills a current gap around evaluating efficient transformer architectures and their alternatives on natural language tasks, and at the same time provides a testing ground for new pretraining schemes that target long language sequences. We hope that SCROLLS inspires the NLP community to go beyond single sentences and paragraphs, and meet the challenges of processing and reasoning over longer discourses.

Acknowledgements

This research was supported in part by the Yandex Initiative for Machine Learning, the Tel Aviv University Data Science Center, the Shashua Fellowship, the Alon Scholarship, the 2021 Scholarship in Computer Science granted by the Séphora Berrebi

Foundation, and the European Research Council (ERC) under the European Union Horizons 2020 research and innovation programme (grant ERC DELPHI 802800).

References

- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. [Findings of the 2020 conference on machine translation \(WMT20\)](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39, Berlin, Heidelberg. Springer.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021. [Summscreen: A dataset for abstractive screenplay summarization](#).
- Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. [Rethinking attention with performers](#).
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. [A discourse-aware attention model for abstractive summarization of long documents](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.
- Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. [A dataset of information-seeking questions and answers anchored in research papers](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Charles J. Fillmore, Nancy Ide, Dan Jurafsky, and Catherine Macleod. 1998. An american national corpus: a proposal. In *LREC*.
- Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. [MRQA 2019 shared task: Evaluating generalization in reading comprehension](#). In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.
- Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. 2021. [A practical survey on faster and lighter transformers](#).
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. [Efficient attentions for long document summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.
- Nancy Ide and Keith Suderman. 2004. [The American national corpus first release](#). In *Proceedings of the Fourth International Conference on Language*

- Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. [The icsi meeting corpus](#). In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, volume 1, pages I–I.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. [Sharp nearby, fuzzy far away: How neural language models use context](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. [The NarrativeQA reading comprehension challenge](#). *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Yuta Koreeda and Christopher D. Manning. 2021. [ContractNLI: A dataset for document-level natural language inference for contracts](#).
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. [Datasets: A community library for natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. [S2ORC: The semantic scholar open research corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.
- Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. 2021. [Quality: Question answering with long input texts, yes!](#)
- Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. 2021. [Random feature attention](#). In *International Conference on Learning Representations*.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2021a. [Shortformer: Better language modeling using shorter inputs](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5493–5505, Online. Association for Computational Linguistics.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2021b. [Train short, test long: Attention with linear biases enables input length extrapolation](#).
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don't know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- 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.
- Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. [Efficient content-based sparse attention with routing transformers](#). *Transactions of the Association for Computational Linguistics*, 9:53–68.
- Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. [Do long-range language models actually use long-range context?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2020a. [Synthesizer: Re-thinking self-attention in transformer models](#).
- Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021.

[Long range arena : A benchmark for efficient transformers.](#) In *International Conference on Learning Representations*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020b. [Efficient transformers: A survey](#).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. [QMSum: A new benchmark for query-based multi-domain meeting summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.