Appendix

A Ethics Statement

While our text pair alignment model achieves state-of-the-art performance on many downstream tasks, like all models, it does make mistakes. For example, when used for fact verification or factual consistency evaluation, it could misidentify *factually correct* statements as *incorrect* and vice versa. Additionally, as we use publicly available datasets to train the alignment model, it might have learned biases inherent to those datasets. Thus, one should proceed with caution when using the alignment model for purposes other than NLP research.

B Comparison with Other Model Types

We illustrate the major differences between our approach, LLMs, multitask learning models, and task-specific finetuned models in Table 6. Compared with LLMs, our alignment model is more efficient but less versatile. In contrast to task-specific finetuned models, our alignment model is more general and can handle more types of tasks. Unlike multitask learning models, we unify language tasks into a single text pair alignment problem, and share model components across multiple tasks (as apposed to using dataset-specific prediction heads). As a result, our alignment model can be directly applied to a wide range of tasks out-of-the-box, without any finetuning.

Table 6: Comparison between our alignment model and other types of models.

Type	Model Example	Efficient	Out of the box	General
LLM	T5, PALM, UL2, GPT	✗	✔	✔
Multitask learning	MT-DNN, MUPPET	✔	✗	✔
Task specific LM	Finetuned RoBERTa/BERT	✔	✗	✗
Text pair alignment	**Ours**	✔	✔	✔

C Training Details

C.1 Trainig Setup

We choose RoBERTa [11] as the backbone for the alignment model. Alignment-base and Alignment-large are based on RoBERTa-base and RoBERTa-large, respectively. For the experiments in Section 4, we train the alignment model for 3 epochs with a batch size of 32, following common practice [11, 15]. Other hyperparameters are listed in Table 7. For the finetuned RoBERTa and RoBERTa-NLI model in Section 4.1, we set batch size to 16 and 32, respectively.

Table 7: The hyperparameters for training the alignment model.

Hyperparameter	**ALIGN-base**	**ALIGN-large**
Parameters	125M	355M
Batch Size	32	32
Epochs	3	3
Optimizer	AdamW	AdamW
Learning Rate	1e-5	1e-5
Weight Decay	0.1	0.1
Adam ϵ	1e-6	1e-6
Warmup Ratio	0.06	0.06
Random Seed	2022	2022
GPU	2×3090	4×A5000
GPU Hour	152h	620h

C.2 Training Datasets

We collect datasets that falls into the scope of alignment as mentioned in Section 3. Table 8 lists the datasets we use for training the alignment model. The size of these datasets ranges from 4k samples to 5M. Most of the datasets are used for binary classification task except some NLI, fact verification and STS datasets.

We only use the first 500k samples in each dataset due to limited computation resource, which results in 5.9M training samples in total. During training, the samples are randomly sampled from the entire adapted training sets.

Table 8: The training datasets of our alignment model. Note due to resource constraints, we only use at most 500k samples from each dataset to train the alignment model.

NLP Task	Dataset	Training Task	Sample Count
NLI	SNLI [32]	3-way classification	550k
	MultiNLI [7]	3-way classification	393k
	Adversarial NLI [31]	3-way classification	163k
	DocNLI [70]	binary classficiation	942k
Fact Verification	NLI-style FEVER [33, 31]	3-way classification	208k
	VitaminC [16]	3-way classification	371k
Paraphrase	QQP [35]	binary classficiation	364k
	PAWS-Wiki [17]	binary classficiation	695k
	PAWS-QQP [17]	binary classficiation	12k
	WikiText-103 [25]	binary classficiation	8M
STS	SICK [71]	regression	4k
	STSB [34]	regression	6k
QA	SQuAD v2 [21]	binary classficiation	130k
	RACE [20]	binary classficiation	351k
	Adversarial QA [72]	binary classficiation	60k
	BoolQ [37]	binary classficiation	19k
	DROP [73]	binary classficiation	155k
	MultiRC [36]	binary classficiation	24k
	HotpotQA [74]	binary classficiation	362k
	NewsQA [75]	binary classficiation	161k
	QuAIL [38]	binary classficiation	41k
	Quoref [76]	binary classficiation	39k
	ROPES [77]	binary classficiation	22k
	SciQ [78]	binary classficiation	47k
	StrategyQA [79]	binary classficiation	5k
Information Retrieval	MS MARCO [19]	binary classficiation	5M
Summarization	WikiHow [80]	binary classficiation	157k
Coreference	GAP [22]	binary classficiation	4k

D Additional Experiment Details

D.1 Natural Language Understanding Tasks

Prompts For FLAN models, we use the same prompts as Longpre et al. [40]. For datasets that do not appear in Longpre et al. [40], we use prompts of similar tasks. The prompt used for each dataset is listed below.

MNLI, NLI-FEVER, VitaminC:

```
"Premise: {premise}\n\nHypothesis: {hypothesis}\n\nDoes the premise
entail the hypothesis?\n\nA yes\nB it is not possible to tell\nC no"
```

ANLI:

```
"{context}\n\nBased on the paragraph above can we conclude that
\"{hypothesis}\"?\n\nA Yes\nB It's impossible to say\nC No"
```

SNLI:

```
"If \"{premise}\", does this mean that \"{hypothesis}\"?\n\nA yes\nB it is
not possible to tell\nC no"
```

SICK, STSB:

```
"{sentence1}\n{sentence2}\n\nRate the textual similarity of these two
sentences on a scale from 0 to 5, where 0 is \"no meaning overlap\" and
5 is \"means the same thing\".\n\nA 0\nB 1\nC 2\nD 3\nE 4\nF 5"
```

PAWS, PAWS-QQP:

```
"{sentence1}\n{sentence2}\n\nDo these sentences mean the same thing?\nA
no\nB yes"
```

QQP:

```
"{question1}\n{question2}\nWould you say that these questions are the
same?\nA no\nB yes"
```

RACE, QuAIL, SciQ:

```
"{fact}\n{question}\n\nA {option 1}\nB {option 2}\nC {option 3} ..."
```

Multi-RC:

```
"{paragraph}\n\nQuestion:  \"{question}\"\n\nResponse:
\"{response}\"\n\nDoes the response correctly answer the question?\n\nA
no\nB yes"
```

BoolQ:

```
"{text}\n\nCan we conclude that {question}?\n\nA no\nB yes"
```

GAP:

```
"Context:  {context}\n Given the context, which option is true?  \n\nA
{option 1}\nB {option 2}\nC {option 3} ..."
```

Results We provide additional results on the in-domain datasets of the alignment model. We show the performance of finetuned RoBERTa and FLAN-Alpaca on these datasets in Table 9. We have compared the alignment model with finetuned RoBERTa on these datasets in Figure 2. When comparing FLAN-T5 and FLAN-Alpaca, we notice FLAN-T5 consistently outperforms FLAN-Alpaca on all scales. Thus, we compare our alignment model with FLAN-T5 in Table 1.

D.2 Factual Consistency Evaluation for Language Generation

In this section, we report the detailed results associated with Figure 4. We list the performance of each metric on SummaC Benchmark and TRUE Benchmark in Table 11 and Table 12, respectively. We also show the Pearson Correlation, Spearman Correlation and Kendall's tau Correlation coefficients on `other` datasets in Table 13, 14 and 15, respectively.

D.3 Question Answering with Unanswerable Question

Simplified Natural Questions For this experiment, we construct a new SQuAD-style QA dataset with unanswerable questions, Simplified Natural Questions (Simplified NQ), base on Natural Questions [68]. For each sample (an search query and a Wikipedia article) in Natural Questions, [68] ask five annotators to find 1) an HTML bounding box (the *long answer*; e.g., paragraphs, tables, list items, or whole list) containing enough information to infer the answer to the query (long answer), and 2) a short span within the long answer that answers the question (the *short answer*). For both short and long answers, the annotators can alternatively indicate that an answer could not be found.

For the purpose of constructing Simplified NQ, we consider a sample to be answerable if at least 2 annotators identified both long and short answers. In this case, we use the most popular (among the annotators) short answer as the ground truth answer, and the most popular long answer containing the selected short answer as the context. Conversely, if less than 2 annotators identified any long answer,

Table 9: The performance of finetuned RoBERTa and FLAN-Alpaca on the in-domain datasets. We report the average performance of each model and we also include the average without RACE and QuAIL.

		Finetuned RoBERTa		FLAN-Alpaca		
		base	large	base	large	xlarge
Model Parameters		125M	355M	220M	770M	3B
NLI	MNLI-mm	87.2	90.3	79.9	86.4	89.3
	MNLI-m	87.9	90.6	80.0	87.2	89.4
	ANLI-1	62.8	72.7	47.4	65.7	74.8
	ANLI-2	44.5	48.3	38.2	46.6	57.6
	ANLI-3	42.8	47.0	37.7	46.4	54.6
	SNLI	91.0	91.4	82.9	88.1	90.2
Fact Verification	NLI-FEVER	76.1	77.7	69.6	73.0	72.1
	VitaminC	89.3	91.6	63.3	72.5	77.4
STS	SICK	88.9	84.7	37.7	66.4	70.1
	STSB	89.8	90.6	33.4	52.5	79.5
Paraphrase	PAWS	92.3	92.5	68.1	92.0	93.0
	PAWS-QQP	94.7	94.2	57.6	85.1	87.1
	QQP	91.1	92.0	75.5	81.6	86.5
QA	RACE-m	74.6	24.0	64.3	78.5	87.8
	RACE-h	67.8	23.9	57.3	71.9	82.9
	Multi-RC	77.5	85.5	64.2	84.3	87.1
	BoolQ	79.1	85.7	71.7	82.0	87.2
	QuAIL	57.7	27.0	56.7	78.2	84.2
	SciQ	93.4	95.5	90.8	83.1	95.6
Coreference	GAP	74.3	89.8	58.4	65.6	80.7
Average		78.1	74.7	61.7	74.4	81.4
Average w/o RACE, QuAIL		80.2	83.5	62.1	74.0	80.7

Table 10: The performance of RoBERTa-NLI and FLAN-Alpaca on the zero-shot datasets (of alignment model). The gray number shows the specific dataset is appeared in the training set of FLAN-T5. We report the average performance of each model in the last row.

		RoBERTa-NLI		FLAN-Alpaca		
		base	large	base	large	xlarge
Model Parameters		125M	355M	220M	770M	3B
NLI	AXB	75.2	79.2	53.6	72.3	77.2
	AXG	59.6	73.6	49.4	72.5	88.8
	CB	85.7	87.5	78.6	78.6	87.5
	RTE	81.2	88.1	72.9	79.8	87.0
	WNLI	52.1	50.7	40.8	62.0	71.8
	SE14T1	91.2	93.1	65.0	72.4	77.3
Paraphrase	MRPC	38.7	42.3	66.7	75.4	83.1
QA	DREAM	63.9	73.0	63.5	76.8	89.5
	Quartz	54.6	65.6	68.1	87.4	90.2
Average		66.9	72.6	62.1	75.2	83.6

Table 11: The ROC AUC of each metric on the SummaC benchmark. CGS and XSF are abbreviations of CogenSumm and XSumFaith, respectively. The strongest performance on each dataset is shown in **bold**. The last column shows the average performance on each dataset in the SummaC benchmark.

	SummaC Benchmark						
	CGS	XSF	PolyTope	FactCC	SummEval	FRANK	AVG
BERTScore	63.1	49.0	85.3	70.9	79.6	84.9	72.1
BLEURT	60.8	64.7	76.7	59.7	71.1	82.5	69.2
BARTScore	74.3	62.6	91.7	82.3	85.9	88.5	80.9
CTC	76.5	65.9	89.5	82.6	85.6	87.3	81.2
UniEval	84.7	65.5	**93.4**	89.9	86.3	88.0	84.6
QAFactEval	83.4	66.1	86.4	89.2	88.1	89.4	83.8
Alignment-base (Ours)	80.6	**76.1**	87.5	93.1	88.6	89.5	85.9
Alignment-large (Ours)	**88.4**	74.6	92.5	**94.9**	**92.3**	**91.3**	**89.0**

Table 12: The ROC AUC of each metric on the TRUE benchmark. The datasets with asterisks(*) appear in the training set of the alignment model. We compute both the overall average on all datasets (`Average`) and average without PAWS, FEVER, VitaminC datasets (`Average-ZS`). The latter shows the zero-shot performance of the alignment model. **Bold** indicates the best performance on a dataset.

		BERTScore	BLEURT	BARTScore	CTC	UniEval	QAFactEval	**Alignment -base (Ours)**	**Alignment -large (Ours)**
TRUE Benchmark	FRANK	84.0	81.6	87.8	87.1	88.1	**88.5**	79.6	83.2
	SummEval	72.3	68.0	78.9	79.8	81.2	80.9	**81.3**	81.1
	MNBM	52.5	65.5	63.5	65.0	66.8	67.3	83.3	**85.1**
	QAGS-C	70.6	71.2	83.9	77.3	86.5	83.9	94.7	**94.9**
	QAGS-X	44.3	56.2	60.2	67.7	76.7	76.1	97.6	**98.4**
	BEGIN	86.4	86.6	**86.7**	72.0	73.6	81.0	77.2	79.2
	Q2	70.2	72.9	65.1	66.8	70.4	75.8	80.3	**89.0**
	DialFact	68.6	73.0	60.8	63.7	80.4	81.8	90.4	**91.4**
	PAWS*	78.6	68.4	77.1	63.1	80.1	**86.1**	79.5	83.8
	FEVER*	54.2	59.5	66.1	72.5	**92.1**	86.0	76.4	74.4
	VitaminC*	58.2	61.8	64.2	65.0	79.1	73.6	97.8	**98.3**
	Average	67.2	69.5	72.2	70.9	79.5	80.1	85.3	**87.2**
	Average-ZS	68.6	71.9	73.4	72.4	78.0	79.4	85.5	**87.8**

Table 13: The Pearson correlation coefficients of various metrics on `other` datasets mentioned in Section 4.2.1. Q-XSum and Q-CNNDM are abbreviations of QAGS-XSum and QAGS-CNNDM, respectively. F-XSum and F-CNNDM are abbreviations of FRANK-XSum and FRANK-CNNDM, respectively. The last column shows the average performance on each dataset. The best performance is shown in **bold**.

	Other Datasets - Pearson							
	XSumFaith	SummEval	Q-Xsum	Q-CNNDM	F-Xsum	F-CNNDM	SamSum	AVG
BERTScore	13.0	33.1	-10.6	51.7	13.0	51.7	10.9	23.3
BLEURT	**38.7**	23.8	13.2	45.2	15.6	37.5	8.1	26.0
BARTScore	29.3	35.5	16.3	71.5	23.7	51.9	15.0	34.7
CTC	27.2	54.7	30.6	64.5	20.0	54.5	16.9	38.3
UniEval	23.9	57.8	45.5	66.7	27.2	58.3	23.2	43.2
QAFactEval	30.3	61.6	44.2	68.4	32.1	64.6	38.9	48.6
Alignment-base (Ours)	33.2	57.8	51.1	60.9	31.2	61.8	21.1	45.3
Alignment-large (Ours)	28.8	**66.7**	**53.9**	**76.1**	**38.9**	**68.9**	**47.7**	**54.4**

Table 14: The Pearson correlation coefficients of various metrics on `other` datasets mentioned in Section 4.2.1. The format in this table follows Table 13.

	Other Datasets - Spearman							
	XSumFaith	SummEval	Q-Xsum	Q-CNNDM	F-Xsum	F-CNNDM	SamSum	AVG
BERTScore	13.4	31.5	-8.9	46.2	12.7	45.1	13.1	21.9
BLEURT	37.0	23.6	12.4	43.4	13.9	37.6	6.7	24.9
BARTScore	29.8	39.1	17.0	68.1	20.0	53.3	16.3	34.8
CTC	29.8	41.7	30.6	57.3	20.4	49.4	17.7	35.3
UniEval	25.3	44.3	50.0	67.6	26.7	54.0	22.8	41.5
QAFactEval	31.9	42.8	44.1	63.1	25.5	53.7	35.9	42.4
Alignment-base (Ours)	**38.8**	42.0	52.7	56.1	25.5	56.4	22.3	42.0
Alignment-large (Ours)	32.1	**47.9**	**57.4**	**71.6**	**30.0**	**61.8**	**46.7**	**49.7**

Table 15: The Kendall's tau correlation coefficients of various metrics on `other` datasets mentioned in Section 4.2.1. The format in this table follows Table 13.

	Other Datasets - Kendall's tau							
	XSumFaith	SummEval	Q-Xsum	Q-CNNDM	F-Xsum	F-CNNDM	SamSum	AVG
BERTScore	9.2	24.9	-7.3	36.3	10.4	34.7	10.7	17.0
BLEURT	25.3	18.6	10.1	33.9	11.4	28.8	5.5	19.1
BARTScore	20.2	31.0	13.9	55.6	16.3	41.4	13.3	27.4
CTC	20.2	33.2	25.1	45.7	16.6	38.2	14.4	27.6
UniEval	17.0	35.3	40.9	54.4	21.8	42.4	18.7	32.9
QAFactEval	23.2	34.0	36.2	50.5	22.4	42.2	30.1	34.1
Alignment-base (Ours)	**26.6**	33.4	43.1	45.5	20.8	44.4	18.2	33.1
Alignment-large (Ours)	21.9	**38.4**	**47.0**	**59.6**	**24.5**	**49.5**	**38.2**	**39.9**

and less than 2 annotators identified any short answer, we consider the sample to be unanswerable and use a random paragraph from the article as the context. We discard all remaining samples to avoid ambiguity (e.g., some samples might only have long answers but not short answers). This results in a total of 3336 answerable samples and 3222 unanswerable samples in the validation set.

Prompts and QA Inference For FLAN T5, we follow [40] and use the following prompt:

```
Context: {context}\nQuestion: {question}\nAnswer:
```

For GPT-3.5, we use a prompt with additional instructions:

```
Find the answer to the question from the given context. When the question
cannot be answered with the given context, say "unanswerable". Just say
the answer without repeating the question.\nContext: {context}\nQuestion:
{question}\nAnswer:
```

At inference time, we truncate the context if necessary such that the entire input is at most around 2000 tokens long (2000 for FLAN T5, 2040 for GPT-3.5 to account for the longer prompt). We use greedy decoding for FLAN T5, a the default chat completion settings for GPT-3.5. When FLAN T5 outputs `"unanswerable"`, we interpret it as predicting the sample to be not answerable. Similarly, if GPT-3.5's output contains any of `"unanswerable"`, `"no answer"`, `"context does not provide an answer"`, we consider the prediction to be unanswerable.

Additional Results In addition to FLAN T5 and GPT-3.5, we also experiment with Electra [81], one of the top performing single models on the SQuAD v2 leaderboard, for reference. Specifically, we reproduce Clark et al. [81]'s design that use a QA prediction head to jointly predict the answer span and unanswerable probability. As shown in Table 16, while Electra is a strong performer on SQuAD v2 and Simplified NQ, adding the alignment verifier to GPT-3.5 and FLAN T5 greatly reduces the performance gap. Additionally, on ACE-whQA, our design (both FLAN T5 and GPT-3.5 with alignment verifiers) outperforms Electra.

D.4 Ablation Study

We present the additional ablation result on factual consistency evaluation tasks in Table 17. This part follows Section 4.4, where we use the same checkpoints that are trained on incrementally added tasks.

Table 16: Additional experiment results on QA with unanswerable questions including Electra. The best model for each task/metric is shown in **bold**.

	SQuAD v2			ACE-whQA			Simplified NQ		
	EM	F1	AUC	EM	F1	AUC	EM	F1	AUC
Electra	**86.47**	**89.37**	**0.97**	52.32	55.59	0.87	**70.81**	**74.13**	**0.88**
GPT-3.5	52.53	63.96	0.76	67.98	71.98	0.77	58.37	68.61	0.81
Flan T5	75.72	79.01	0.83	26.29	29.24	0.51	38.24	44.98	0.58
GPT-3.5 + Verifier (Ours)	67.19	77.63	0.93	**79.02**	**80.91**	0.84	56.16	57.40	0.86
FLAN T5 + Verifier (Ours)	83.72	86.55	0.95	75.75	77.60	**0.90**	64.93	67.99	0.83

Result shows the training tasks are generally compatible and effective, though we notice adding fact verification and paraphrase detection tasks lead to a slightly performance drop. We speculate it is due to the paraphrase detection task, where a text pair is expected to have exactly the same information. The Alignment-base model, which uses all the possible training data, gets the best performance on every factual consistency evaluation task.

Table 17: Ablation results on factual consistency evaluation tasks. Each row corresponds with a model trained with data adapted from incrementally more types of tasks. For example, the model on the second row is trained with NLI, Fact Verification and Paraphrase tasks. The model on the last row is the same as Alignment-base. We report the average performance for each evaluation tasks. The last column shows the overall average for the factual consistency evaluation tasks. The best performance for each task is shown in **bold**.

Training Tasks	Factual Consistency Evaluation Tasks					
	SummaC	TRUE	Other-Pearson	Other-Spearman	Other-Kendall	Average
+NLI	78.1	77.5	32.6	33.6	26.3	49.6
+FV, Para	74.9	80.3	27.6	27.2	21.1	46.2
+Coref, Sum, IR, STS	84.2	83.7	39.4	36.8	28.8	54.6
+QA (Alignment-base)	**85.9**	**85.3**	**45.3**	**42.0**	**33.1**	**58.3**

References

[1] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

[2] 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. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[3] Matúš Pikuliak. Chatgpt survey: Performance on NLP datasets, Mar 2023. URL http://opensamizdat.com/posts/chatgpt_survey/.

[4] Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.660. URL https://aclanthology.org/2020.emnlp-main.660.

[5] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *CoRR*, abs/2104.14690, 2021. URL https://arxiv.org/abs/2104.14690.

[6] Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.104. URL https://aclanthology.org/2021.naacl-main.104.

[7] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

[8] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL https://aclanthology.org/P19-1441.

[9] Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.468. URL https://aclanthology.org/2021.emnlp-main.468.

[10] Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. MVP: multi-task supervised pre-training for natural language generation. *CoRR*, abs/2206.12131, 2022. doi: 10.48550/arXiv.2206.12131. URL https://doi.org/10.48550/arXiv.2206.12131.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/arXiv.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

[14] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003. URL http://jmlr.org/papers/v3/bengio03a.html.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[16] Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52.

[17] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131.

[18] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1150. URL https://aclanthology.org/P15-1150.

[19] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016. URL http://arxiv.org/abs/1611.09268.

[20] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

[21] Pranav Rajpurkar, Robin Jia, and Percy Liang. 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124.

[22] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018. doi: 10.1162/tacl_a_00240. URL https://aclanthology.org/Q18-1042.

[23] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10.

[24] Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. SMART: sentences as basic units for text evaluation. *CoRR*, abs/2208.01030, 2022. doi: 10.48550/arXiv.2208.01030. URL https://doi.org/10.48550/arXiv.2208.01030.

[25] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

[26] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4020. URL https://aclanthology.org/P18-4020.

[27] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606, 2016. URL http://arxiv.org/abs/1602.03606.

[28] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750.

[29] Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.599. URL https://aclanthology.org/2021.emnlp-main.599.

[30] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

[31] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

[32] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

[33] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

[34] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

[35] Kornél Csernai. First quora dataset release: Question pairs. URL https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.

[36] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL https://aclanthology.org/N18-1023.

[37] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

[38] A Rogers, O Kovaleva, M Downey, and A Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[39] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[40] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. *CoRR*, abs/2301.13688, 2023. doi: 10.48550/arXiv.2301.13688. URL https://doi.org/10.48550/arXiv.2301.13688.

[41] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[42] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 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, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

[43] SemEval-2014. Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. URL https://alt.qcri.org/semeval2014/task1/.

[44] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

[45] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264. URL https://aclanthology.org/Q19-1014.

[46] Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1608. URL https://aclanthology.org/D19-1608.

[47] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16121.

[48] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1051. URL https://aclanthology.org/D19-1051.

[49] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1256. URL https://aclanthology.org/P19-1256.

[50] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

[51] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.19. URL https://aclanthology.org/2022.dialdoc-1.19.

[52] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24.

[53] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450.

[54] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.383. URL https://aclanthology.org/2021.naacl-main.383.

[55] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

[56] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166, 2023. doi: 10.48550/arXiv.2302.04166. URL https://doi.org/10.48550/arXiv.2302.04166.

[57] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634, 2023. doi: 10.48550/arXiv.2303.16634. URL https://doi.org/10.48550/arXiv.2303.16634.

[58] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. Human-like summarization evaluation with chatgpt. *CoRR*, abs/2304.02554, 2023. doi: 10.48550/arXiv.2304.02554. URL https://doi.org/10.48550/arXiv.2304.02554.

[59] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[60] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

[61] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

[62] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.131.

[63] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.187. URL https://aclanthology.org/2022.naacl-main.187.

[64] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[65] Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. Read + verify: Machine reading comprehension with unanswerable questions. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019,*

Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6529–6537. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33016529. URL https://doi.org/10.1609/aaai.v33i01.33016529.

[66] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. Sg-net: Syntax-guided machine reading comprehension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9636–9643. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6511.

[67] Elior Sulem, Jamaal Hay, and Dan Roth. Do we know what we don't know? studying unanswerable questions beyond SQuAD 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4543–4548, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.385. URL https://aclanthology.org/2021.findings-emnlp.385.

[68] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://doi.org/10.1162/tacl_a_00276.

[69] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

[70] Wenpeng Yin, Dragomir Radev, and Caiming Xiong. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.435. URL https://aclanthology.org/2021.findings-acl.435.

[71] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

[72] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL https://aclanthology.org/2020.tacl-1.43.

[73] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

[74] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

[75] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623.

[76] Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1606. URL https://aclanthology.org/D19-1606.

[77] Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5808. URL https://aclanthology.org/D19-5808.

[78] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413.

[79] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21.

[80] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305, 2018. URL http://arxiv.org/abs/1810.09305.

[81] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.