

HotpotQA

A Dataset for Diverse, Explainable Multi-hop Question Answering

What is HotpotQA?

HotpotQA is a question answering dataset featuring natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems. It is collected by a team of NLP researchers at [Carnegie Mellon University](#), [Stanford University](#), and [Université de Montréal](#).

For more details about HotpotQA, please refer to our EMNLP 2018 paper:

(Yang, Qi, Zhang, et al. 2018)

Getting started

HotpotQA is distributed under a [CC BY-SA 4.0 License](#). The training and development sets can be downloaded below.

Training set (535MB)

Dev set (distractor) (44MB)

Dev set (fullwiki) (45MB)

Test set (fullwiki) (46MB)

A more comprehensive summary about data download, preprocessing, baseline model training, and evaluation is included in our [GitHub repository](#), and linked below.

Getting started guide

Once you have built your model, you can use the evaluation script we provide below to evaluate model performance by running `python hotpot_evaluate_v1.py <path_to_prediction> <path_to_gold>`

Leaderboard (Distractor Setting)

In the *distractor* setting, a question-answering system reads 10 paragraphs to provide an answer (Ans) to a question. They must also justify these answers with supporting facts (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F1	EM	F1	EM	F1
1 Oct 18, 2019	C2F Reader (single model) <i>Joint Laboratory of HIT and iFLYTEK Research</i>		67.98	81.24	60.81	87.63	44.67	72.73
2 Sep 27, 2019	HGN (single model) <i>Microsoft Dynamics 365 AI Research</i>		66.07	79.36	60.33	87.33	43.57	71.03
3 Jul 29, 2019	TAP 2 (ensemble)		66.64	79.82	57.21	86.69	41.21	70.65
4 Jul 29, 2019	TAP 2 (single model)		64.99	78.59	55.47	85.57	39.77	69.12
5 May 31, 2019	EPS + BERT (large) (single model) <i>Anonymous</i>		63.29	76.36	58.25	85.60	41.39	67.92
6 Aug 31, 2019	SAE (single model) <i>Anonymous</i>		60.36	73.58	56.93	84.63	38.81	64.96
7 Jun 13, 2019	P-BERT (single model) <i>Anonymous</i>		61.18	74.16	51.38	82.76	35.42	63.79
8 Sep 16, 2019	LQR-net 2 + BERT-Base (single model) <i>Anonymous</i>		60.20	73.78	56.21	84.09	36.56	63.68
9 Apr 11, 2019	EPS + BERT (single model) <i>Anonymous</i>		60.13	73.31	52.55	83.20	35.40	63.41
10 May 16, 2019	PIPE (single model) <i>Anonymous</i>		59.77	72.77	52.53	82.82	35.54	62.92
11 Jun 8, 2019	TAP (single model)		58.63	71.48	46.84	82.98	32.03	61.90

Evaluation script (4.2KB)

Sample dev prediction (982KB)

To submit your models and evaluate them on the official test sets, please read our submission guide hosted on Codalab.

Submission Guide

We also release the processed Wikipedia used in the process of creating HotpotQA (also under a [CC BY-SA 4.0 License](#)), serving both as the corpus for the *fullwiki* setting in our evaluation, and hopefully as a standalone resource for future researches involving processed text on Wikipedia. Below please find the link to the documentation for this corpus.

Processed Wikipedia README

Stay connected!

Join our [Google group](#) to receive updates or initiate discussions about HotpotQA!

If you use HotpotQA in your research, please cite our paper with the following BibTeX entry















```
@inproceedings{yang2018hotpotqa,
  title={HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering},
  author={Yang, Zhilin and Qi, Peng and Zhang, Saizheng and Bengio, Yoshua and Cohen, William W. and Salakhutdinov, Ruslan and Manning, Christopher D.},
  booktitle={Conference on Empirical Methods in Natural Language Processing (EMNLP)},
  year={2018}
}
```







12	SAQA (single model)		55.07	70.22	57.62	84.19	35.94	61.72
Aug 14, 2019	<i>Anonymous</i>							
13	MKGN (single model)		57.09	70.69	54.26	83.54	35.59	61.69
Sep 2, 2019	<i>Anonymous</i>							
14	GRN + BERT (single model)		55.12	68.98	52.55	84.06	32.88	60.31
Apr 19, 2019	<i>Anonymous</i>							
15	LQR-net + BERT-Base (single model)		57.20	70.66	50.20	82.42	31.18	59.99
Jun 19, 2019	<i>Anonymous</i>							
16	DFGN (single model)		56.31	69.69	51.50	81.62	33.62	59.82
Apr 22, 2019	Shanghai Jiao Tong University & ByteDance AI Lab (Xiao, Qu, Qiu et al. ACL19)							
17	QFE (single model)		53.86	68.06	57.75	84.49	34.63	59.61
Nov 21, 2018	NTT Media Intelligence Laboratories (Nishida et al. ACL19)							
18	LQR-net (ensemble)		55.19	69.55	47.15	82.42	28.42	58.86
Apr 17, 2019	<i>Anonymous</i>							
19	GRN (single model)		52.92	66.71	52.37	84.11	31.77	58.47
Mar 4, 2019	<i>Anonymous</i>							
20	DFGN + BERT (single model)		55.17	68.49	49.85	81.06	31.87	58.23
Mar 1, 2019	<i>Anonymous</i>							
21	BERT Plus (single model)		55.84	69.76	42.88	80.74	27.13	58.23
Mar 4, 2019	CIS Lab							
22	KGNN (single model)		50.81	65.75	38.74	76.79	22.40	52.82
May 18, 2019	<i>Anonymous</i>							
23	Baseline Model (single model)		45.60	59.02	20.32	64.49	10.83	40.16
Oct 10, 2018	Carnegie Mellon University, Stanford University, & Universite de Montreal (Yang, Qi, Zhang, et al. 2018)							
-	ChainEx (single model)		61.20	74.11	N/A	N/A	N/A	N/A
Sep 24, 2019	UT Austin							
-	DecompRC (single model)		55.20	69.63	N/A	N/A	N/A	N/A
Feb 27, 2019	University of Washington (Min et al. ACL18)							
-	MatrixRC (single model)		47.07	60.75	N/A	N/A	N/A	N/A
Apr 2, 2019	<i>Anonymous</i>							

Leaderboard (Fullwiki Setting)

In the *fullwiki* setting, a question-answering system must find the answer to a question in the scope of the entire Wikipedia. Similar to in the distractor setting, systems are

evaluated on the accuracy of their answers (Ans) and the quality of the supporting facts they use to justify them (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F ₁	EM	F ₁	EM	F ₁
1 <small>Oct 7, 2019</small>	HGN + SemanticRetrievalMRS IR (single model) <i>Microsoft Dynamics 365 AI Research</i>		56.71	69.16	49.97	76.39	35.63	59.86
2 <small>Sep 20, 2019</small>	Graph-based Recurrent Retriever (single model) <i>Anonymous</i>		56.04	68.87	44.14	73.03	29.18	55.31
3 <small>Sep 28, 2019</small>	MIR+EPS+BERT (single model) <i>Anonymous</i>		52.86	64.79	42.75	72.00	31.19	54.75
4 <small>Sep 21, 2019</small>	Transformer-XH (single model) <i>Anonymous</i>		48.95	60.75	41.66	70.01	27.13	49.57
5 <small>May 15, 2019</small>	SemanticRetrievalMRS (single model) <i>UNC-NLP (Nie et al., EMNLP'2019)</i>		45.32	57.34	38.67	70.83	25.14	47.60
6 <small>Jul 31, 2019</small>	Entity-centric BERT Pipeline (single model) <i>Anonymous</i>		41.82	53.09	26.26	57.29	17.01	39.18
7 <small>May 21, 2019</small>	GoldEn Retriever (single model) <i>Stanford University (Qi et al., EMNLP-IJCNLP 2019)</i>		37.92	48.58	30.69	64.24	18.04	39.13
8 <small>Aug 14, 2019</small>	PR-Bert (single model) <i>KingSoft AI Lab</i>		43.33	53.79	21.90	59.63	14.50	39.11
9 <small>Feb 21, 2019</small>	Cognitive Graph QA (single model) <i>Tsinghua KEG & Alibaba DAMO Academy (Ding et al., ACL'19)</i>		37.12	48.87	22.82	57.69	12.42	34.92
10 <small>Mar 5, 2019</small>	MUPPET (single model) <i>Technion (Feldman and El-Yaniv, ACL'19)</i>		30.61	40.26	16.65	47.33	10.85	27.01
11 <small>Apr 7, 2019</small>	GRN + BERT (single model) <i>Anonymous</i>		29.87	39.14	13.16	49.67	8.26	25.84
12 <small>May 20, 2019</small>	Entity-centric IR (single model) <i>Anonymous</i>		35.36	46.26	0.06	43.16	0.02	25.47
13 <small>May 19, 2019</small>	KGNN (single model) <i>Anonymous</i>		27.65	37.19	12.65	47.19	7.03	24.66
14 <small>Aug 16, 2019</small>	SAQA (single model) <i>Anonymous</i>		28.44	38.62	14.69	47.17	8.62	24.49

15	GRN (single model) <i>Anonymous</i>		27.34	36.48	12.23	48.75	7.40	23.55
Mar 4, 2019								
16	QFE (single model) <i>NTT Media Intelligence Laboratories</i>		28.66	38.06	14.20	44.35	8.69	23.10
Nov 25, 2018	(Nishida et al., ACL'19)							
17	Baseline Model (single model) <i>Carnegie Mellon University, Stanford University, & Universite de Montreal</i>		23.95	32.89	3.86	37.71	1.85	16.15
Oct 12, 2018	(Yang, Qi, Zhang, et al. 2018)							
-	TPReasoner w/o BERT (single model) <i>Anonymous</i>		36.04	47.43	N/A	N/A	N/A	N/A
May 19, 2019								
-	DecompRC (single model) <i>University of Washington</i>		30.00	40.65	N/A	N/A	N/A	N/A
Feb 28, 2019	(Min et al., ACL'18)							
-	MultiQA (single model) <i>Anonymous</i>		30.73	40.23	N/A	N/A	N/A	N/A
Mar 3, 2019								

Copyright © HotpotQA Team, 2018-2019.

Theme adapted from Start Bootstrap's [Clean Blog](#) template.