A Inference Types

A.1 SQuAD

Word Match: The model can simply match keys words in the question to find the answer bearing sentence and select the correct span.

Coreference: The model need to resolve a pronoun in the answer bearing sentence to find the answer.

Implicit Relation: Key entities in the context share a relationship that is not explicitly stated in the question. The model must infer the relationship to select the answer.

Paraphrase: The question paraphrases the answer bearing sentence.

Long Distance: Evidence for the answer is separated by a long sequence of irrelevant words.

Multi-coreference: The model needs to infer that one pronoun is referring to multiple entities.

Table 8 shows an example for each inference type.

A.2 HotpotQA

Multi Bridge: The model must perform multihop inference by finding and evaluating both supporting facts in the context. Each supporting fact is linked by a common ”bridge” entity.

No Multi Bridge: Context clues alone can identify the answer. No multihop inference required.

Comparison: The question compares two entities, and the model must select the correct one.

Yes/No: The model must choose between a yes or no answer.

Numeric: The model must compare numeric quantities to choose the answer.

A.3 MSMARCO

There is only one new category in MSMARCO:

Part-whole Relation The model would need to infer that one entity is an example or a subset of another entity and leverage inherited properties to answer the question. An example would be:

Question: *cannot uninstall windirstat*

Gold Context: *Windows Add/ Remove Programs offers users a way to uninstall the program ... Click Start menu and run Control Panel ...*

Answer: *Click Start menu and run Control Panel...*

The model would have to understand that windirstat is a program to make correct prediction.

Inference Type	**Question**	**Context**	**Answer**
Word Match	What team was the NFC champion?	... the National Football Conference (NFC) champion Carolina Panthers ...	Carolina Panthers
Coreference	What did Luther seek to restore?	Luther next set ... the authorities to restore public order, he signalled his reinvention ...	to restore public order
Implicit Relation	Who was Margaret's brother?	... King Malcolm III of Scotland married Edgar's sister Margaret ...	Edgar
Paraphrase	What is an example of a pump component?	Other components are often present; pumps (such as an injector) to supply water ...	injector
Long Distance	In a platoon teaching, what gives the children security?	a platoon system, involves ... The advantage here is ... staying with the same group of peers	staying with the same group of peers
Multi-Coreference	What do A, B and C have in common?	A, B and C are disturbed, they produce secretions that luminesce	they produce secretions that luminesce

Table 8: Inference Type Examples for SQuAD

Inference Type	**Question**	**Context**	**Answer**
Multi Bridge	How long is the river for which Frenchmans Creek is a tributary?	The Darling River is ... 2844 km Frenchmans Creek is a short tributary of the Darling River	2844 km
No Multi Bridge	Who directed and wrote the 2016 film featuring the voice of Townsend Coleman?	Sing is a 2016 American 3D computer-animated musical comedy film... directed and written by Garth Jennings	Garth Jennings
Comparison	Which head coach has led their team for a longer period? of time, Tim Cluess or Steve Prohm?	...seventh year head coach Tim Cluess. Steve Prohm, who was in his 1st season...	Tim Cluess
Yes/No	Are Uber Goober and American Jobs both documentaries about gaming?	Uber Goober... is a 2004 documentary American Jobs is a 2004 ... documentary	No
Numeric	Which genus is native to more continents, Nothoscordum or Callirhoe?	Nothoscordum... is native to North and South America Callirhoe is ... native to... North America	Nothoscordum

Table 9: Inference Type Examples for HotpotQA

Error Type	**Question**	**Answer**	**Prediction**	**QANet**	**BERT**	**CSM**	**Denoise**
Random Guess	How high do plague fevers run?	38-41C	near 100%	28%	16%	26%	35%
Same Entity Type	What team lost Super Bowl XXXIII?	Atlanta Falcons	Denver	30%	34%	24%	39%
Sentence Selection	What did Marlee Matlin translate?	the national anthem	American Sign Language	20%	22%	10%	7%
Copying From Question	What was Apple Talk	proprietary suite of networking protocols	AppleTalk	4%	0%	10%	2%
Fact. Correct Answer	Which video gaming company debuted ...	Nintendo	Pokemon Company	7%	11%	3%	5%
Reasonable Answer	What did Edison offer Tesla ...	$10 a week raise	payment	5%	8%	6%	3%

Table 10: Common Types of Errors on SQuAD

Error Type	Question	Answer	Prediction	QANet	BERT	CSM	Denoise
Multihop Inference	How long is the river for which Frenchmans Creek is a Tributary?	2844 km	729 km	13%	8%	12%	35%
Sentence Selection	What three time Tony nominee composed Ghost Quartet?	Dave Malloy	Julie Harris	12%	18%	29%	34%
Span Selection	Which ”Roseanne” star is in Scream 2?	Laurie Metcalf	Rebecca Gayheart	33%	22%	19%	7%
Confused By Question	What type of word play does ”What Are Little Girls Made Of?” and ”What Are Little Boys Made ”Of” have in common?	ryhme	rock	9%	14%	15%	7%
Fact. Correct Answer	Where is Anticimex’s parent company headquartered?	EQT Plaza	Woonsocket Rhode Island	13%	12%	7%	5%
Entity Choice	Which band has released more albums with their original members, Sick Puppies or Third Eye Blind?	Sick Puppies	Third Eye Blind	10%	16%	11%	9%
Yes/No Choice	Are Uber Goober and American Jobs both documentaries about gaming?	No No	Yes Yes	10%	9%	5%	4%
Numeric Inference	Which genus is native to more continents, Nothoscordum or Callirhoe ?	Nothoscordum	Callirhoe	8%	2%	8%	6%

Table 11: Common Types of Errors on HotpotQA

Error Type	Question	Answer	Prediction	QANet	BERT	CSM	Denoise
Random Guess	variety plague carrier seen	flea	nosferatu	19%	16%	28%	18%
Same Entity Type	powered gasoline engine electric motor company ’s ..	honda	toyota	30%	29%	32%	37%
Sentence Selection	judy garland first female honored special golden globe ...	cecil b demille	jodie foster	20%	22%	19%	24%
Fact. Correct Answer	manatee relative order sirenia found coastal waters north australia	dugong	dugongs	8%	10%	7%	6%
Reasonable Answer	valley 282 feet sea level state lowest point western hemisphere	california	death [10] valley	6%	7%	6%	4%
Answer Missing	jan 20 , 2009 man lose 400,000 year plus 50 grand expenses federal ...	george w bush	willie pearl russell	5%	7%	5%	4%

Table 12: Common Types of Errors on SearchQA

Error Type	Question	Answer	Prediction	QANet	BERT	CSM	Denoise
Random Guess	what is the longest baseball hit	Joe DiMaggio’s 56 game hitting streak	3 of the 1932	42%	14%	26%	48%
Same Entity Type	when is st patrick’s day	March 17	2017	10%	18%	23%	25%
Sentence Selection	what airline flies to las vegas	British Airways, Virgin Atlantic	biggest airlines flying to Vegas	9%	15%	16%	6%
Fact. Correct Answer	how long are car loans typically	60-month	5 years	14%	40%	12%	11%
Reasonable Answer	what food can make you regrow hair	Fish can make you regrow hair	walnuts and salmon	17%	11%	11%	4%
Wrong Yes/no Choice	is eric trump’s wife jewish	No, she’s not jewish	yes	8%	11%	4%	0%

Table 13: Common Types of Errors on MSMARCO

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	.270	.132			
Length (Tokens)					
Question	-.058^c	.009	.927	.944	.961
Answer	-.081^c	.006	.910	.922	.934
Overlap Types					
Word Match	.238^b	0.064	1.12	1.27	1.44
Question-Answer	3.10^c	.371	10.8	22.3	46.4
Question-Sentence	.062^a	.020	1.02	1.06	1.11
Avg Word Match	-.042	.024	.915	.959	1.00
Question Types					
Who	.950^c	.116	2.06	2.58	3.24
What	.442^c	.091	1.30	1.56	1.86
Where	.418^a	.133	1.17	1.52	1.97
When	1.31^c	.122	2.91	3.70	4.71
Why	-.084	.189	.635	.920	1.33
How Many	1.11^c	.128	2.37	3.04	3.91
Which	.673^c	.126	1.53	1.96	2.51
Entity Counts					
Question	.083^a	.026	1.03	1.09	1.14
Pronouns (Passage)	-.015^c	.008	.971	.986	1.00
Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$					

Table 14: Logistic Regression for QANet EM Score on SQuAD Dataset

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	.959	.145			
Length (Tokens)					
Question	-.051^c	.010	.931	.951	.970
Answer	-.080^c	.007	.911	.923	.934
Overlap Types					
Word Match	.204	0.072	1.06	1.22	1.41
Question-Answer	2.71^c	.439	6.38	15.0	35.6
Question-Sentence	.077^a	.023	1.03	1.08	1.13
Avg Word Match	-.099^b	.027	.859	.906	.955
Question Types					
Who	1.04^c	.131	2.19	2.83	3.66
What	.591^c	.097	1.49	1.81	2.18
Where	.251	.142	.975	1.29	1.70
When	1.40^c	.142	3.08	4.05	5.37
Why	-.356	.191	.482	.700	1.02
How Many	1.02^c	.142	2.11	2.78	3.68
Which	.663	.139	1.48	1.94	2.55
Entity Counts					
Question	.067	.030	1.01	1.07	1.13
Pronouns (Passage)	-.023	.009	.960	.977	.993
Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$					

Table 15: Logistic Regression for BERT EM Score on SQuAD Dataset

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	-.360	.127			
Length (Tokens)					
Question	-.042^c	.009	.943	.959	.975
Answer	-.080^c	.007	.911	.923	.934
Overlap Types					
Word Match	.255^c	0.061	1.15	1.29	1.45
Question-Answer	2.81^c	.318	8.91	16.6	31.0
Question-Sentence	.055^a	.018	1.02	1.06	1.09
Avg Word Match	-.066^a	.021	.898	.936	.976
Question Types					
Who	.911^c	.109	2.01	2.49	3.08
What	.507^c	.091	1.39	1.66	1.99
Where	.376	.128	1.13	1.46	1.87
When	1.07^c	.110	2.35	2.92	3.62
Why	.208	.190	.845	1.23	1.78
How Many	1.20^c	.119	2.62	3.30	4.18
Which	.486^c	.118	1.29	1.63	2.05
Entity Counts					
Question	.020	.023	.975	1.02	1.07
Pronouns (Passage)	-.006	.007	.908	.994	1.01
Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$					

Table 16: Logistic Regression for CommonSense Model EM Score on SQuAD Dataset

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	.486	.129			
Length (Tokens)					
Question	-.072^c	.009	.914	.931	.947
Answer	-.122^c	.007	.885	.872	.897
Overlap Types					
Word Match	.214^b	0.062	1.10	1.24	1.40
Question-Answer	3.14^c	.332	12.0	23.0	44.2
Question-Sentence	.072^b	.018	1.03	1.07	1.11
Avg Word Match	-.099^c	.022	.867	.905	.945
Question Types					
Who	.725^c	.111	1.66	2.07	2.57
What	.188	.091	1.01	1.21	1.44
Where	.137	.130	.889	1.15	1.48
When	1.08^c	.115	2.34	2.93	3.68
Why	-.217	.197	.545	.805	1.18
How Many	.948^c	.121	2.04	2.58	3.28
Which	.338	.120	1.11	1.40	1.76
Entity Counts					
Question	.061	.024	1.01	1.06	1.11
Pronouns (Passage)	-.030^c	.007	.956	.979	.984
Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$					

Table 17: Logistic Regression for DS-QA EM Score on SQuAD Dataset

Variable	**B**	**SE**	**95% CI for odds ratio**		
			Lower	**Odds Ratio**	**Upper**
Constant	.555	.048			
Length (Tokens)					
Question	.000	.001	.998	1.00	1.00
Answer	-.045^c	.003	.949	.956	.961
Dist between Sup. Facts	-.007^a	.002	.988	.992	.997
Question-Answer Overlap	.013^c	.003	1.01	1.01	1.02
Distractor Sentences	-.001^c	.001	.997	.999	1.00
Answer Types					
Yes/No	.155^a	.049	1.06	1.17	1.28
Comparison	-.041^c	.018	.924	.959	.994
Numeric	.128^b	.034	1.06	1.14	1.22
Question Types					
How Many	-.129	.054	.789	.878	.977
Why	-.094	.137	.696	.910	1.19
When	0.133	.052	1.03	1.14	1.26
How	.062	.054	.957	1.06	1.18
Which	.054	.044	.968	1.06	1.15
What	.045	.044	.960	1.05	1.15
Where	-.059	.055	.855	.952	1.06
Who	.070	.046	.980	1.07	1.17

Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$

Table 18: Logistic Regression for QANet EM Score on HotpotQA

Variable	**B**	**SE**	**95% CI for odds ratio**		
			Lower	**Odds Ratio**	**Upper**
Constant	.499	.048			
Length (Tokens)					
Question	.001	.001	.999	1.00	1.00
Answer	-.047^c	.003	.948	.954	.960
Dist between Sup. Facts	-.008^b	.002	.988	.992	.997
Question-Answer Overlap	.009^a	.003	1.00	1.01	1.01
Distractor Sentences	-.002^c	.001	.996	.998	1.00
Answer Types					
Yes/No	.042	.049	.949	1.04	1.15
Comparison	-.080^c	.018	.890	.923	.957
Numeric	.026	.034	.847	1.03	1.05
Question Types					
How Many	-.059	.054	.847	.943	1.05
Why	-.190	.137	.633	.827	1.08
When	0.125	.052	1.02	1.13	1.25
How	.045	.054	.941	1.05	1.16
Which	.059	.044	.973	1.06	1.16
What	.044	.044	.959	1.04	1.14
Where	.011	.055	.908	1.01	1.13
Who	.072	.046	.982	1.07	1.18

Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$

Table 19: Logistic Regression for BERT EM Score on HotpotQA

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	.541	.047			
Length (Tokens)					
Question	-.002	.001	.996	.998	1.00
Answer	-.038^{c}	.003	.956	.962	.968
Dist between Sup. Facts	-.007^{a}	.002	.989	.993	.997
Question-Answer Overlap	.003	.003	.998	1.00	1.01
Distractor Sentences	-.003^{c}	.001	.995	.997	.998
Answer Types					
Yes/No	.045	.047	.954	1.05	1.15
Comparison	.006	.018	.971	1.01	1.04
Numeric	.098^{a}	.033	1.03	1.10	1.17
Question Types					
How Many	-.130	.052	.791	.877	.972
Why	-.064	.132	.723	.938	1.22
When	-.006	.050	.900	.993	1.09
How	-.076	.052	.836	.927	1.03
Which	-.027	.042	.895	.973	1.06
What	-.043	.042	.881	.957	1.04
Where	-.070	.053	.840	.932	1.03
Who	-.030	.044	.890	.971	1.06

Model $\chi^2(1) < .001$; $^{a}p < .05$ $^{b}p < .01$ $^{c}p < .001$

Table 20: Logistic Regression for Commonsense Model EM Score on HotpotQA

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	.455	.046			
Length (Tokens)					
Question	-.007^{c}	.001	.991	.993	.996
Answer	-.034^{c}	.003	.960	.966	.972
Dist between Sup. Facts	-.005	.002	.991	.995	1.00
Question-Answer Overlap	.023^{c}	.003	1.02	1.02	1.03
Distractor Sentences	-.003^{a}	.001	.996	.997	.999
Answer Types					
Yes/No	.237	.046^{c}	1.16	1.27	1.39
Comparison	.025	.018	.991	1.03	1.06
Numeric	.116^{b}	.032	1.05	1.12	1.20
Question Types					
How Many	-.092	.052	.824	.912	1.01
Why	-.054	.130	.734	.947	1.22
When	.049	.049	.954	1.05	1.16
How	-.037	.052	.871	.964	1.07
Which	.020	.042	.941	1.02	1.11
What	.016	.042	.936	1.02	1.10
Where	.001	.052	.904	1.00	1.11
Who	.015	.043	.931	1.01	1.11

Model $\chi^2(1) < .001$; $^{a}p < .05$ $^{b}p < .01$ $^{c}p < .001$

Table 21: Logistic Regression for DS-QA EM Score on HotpotQA

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	-2.17	.140			
Length (Tokens)					
Passage (Avg)	$.021^c$	.004	1.01	1.02	1.03
Question	$.026^b$	.007	1.01	1.03	1.04
Answer	$.290^c$	.035	1.25	1.34	1.43
Answer Counts					
Answer-Bearing Passages	$.052^c$	.003	1.05	1.05	1.06
Answer Mentions	$-.004^c$	.001	.994	.996	.998
Answer Entity Type					
Person	$.258^c$	.053	1.17	1.29	1.44
Location	$.477^c$	.055	1.45	1.61	1.79
Organization	$.416^c$	.100	1.25	1.52	1.85
Work of Art	$.716^c$	.207	1.38	2.05	3.11
Consumer Good	.973	.667	.797	2.64	11.9
Similar Entity Mention	$.003^c$	.001	1.00	1.00	1.00
Other	$.523^c$	.111	1.36	1.69	2.10

Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$

Table 22: Logistic Regression for QANet EM Score on SearchQA

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	-2.27	.145			
Length (Tokens)					
Passage (Avg)	$.026^c$	.004	1.02	1.03	1.05
Question	$.034^c$	.007	1.02	1.03	1.05
Answer	$.240^c$	.036	1.18	1.27	1.36
Answer Counts					
Answer-Bearing Passages	$.059^c$	.003	1.06	1.06	1.07
Answer Mentions	-.002	.001	.996	.998	.999
Answer Entity Type					
Person	$.263^c$	.054	1.17	1.30	1.45
Location	$.498^c$	.057	1.47	1.64	1.84
Organization	$.432^c$	.106	1.25	1.52	1.85
Work of Art	.618	.221	1.25	1.54	1.90
Consumer Good	.637	.674	.559	1.89	8.62
Other	$.544^c$	.119	1.37	1.72	2.18
Similar Entity Mention	$.002^b$	.001	1.00	1.00	1.00

Model $\chi^2(1) < .001$; $^ap < .05$ $^bp < .01$ $^cp < .001$

Table 23: Logistic Regression for BERT EM Score on SearchQA

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	-2.05	.142			
Length (Tokens)					
Passage (Avg)	.015^{b}	.004	1.01	1.02	1.02
Question	.021^{a}	.007	1.01	1.02	1.02
Answer	.145^{c}	.035	1.08	1.16	1.24
Answer Counts					
Answer-Bearing Passages	.053^{c}	.003	1.05	1.05	1.06
Answer Mentions	-.000	.001	.998	1.00	1.00
Answer Entity Type					
Person	.556^{c}	.054	1.57	1.74	1.94
Location	.694^{c}	.055	1.80	2.00	2.23
Organization	.500^{c}	.100	1.36	1.65	2.01
Work of Art	.918	.210	1.68	2.50	3.82
Consumer Good	.369	.587	.469	1.45	4.93
Other	.815^{c}	.114	1.81	2.26	2.84
Similar Entity Mention	.002^{a}	.001	1.00	1.00	1.00

Model $\chi^2(1) < .001$; $^{a}p < .05$ $^{b}p < .01$ $^{c}p < .001$

Table 24: Logistic Regression for Commonsense Model EM Score on SearchQA

Variable	B	SE	95% CI for odds ratio		
			Lower	Odds Ratio	Upper
Constant	-2.06	.144			
Length (Tokens)					
Passage (Avg)	.027^{c}	.004	1.02	1.03	1.04
Question	.025^{c}	.004	1.01	1.03	1.04
Answer	.007	.036	.94	1.01	1.08
Answer Counts					
Answer-Bearing Passages	.062^{c}	.003	1.06	1.06	1.07
Answer Mentions	-.002	.001	.996	.998	1.00
Answer Entity Type					
Person	.433^{c}	.055	1.38	1.54	1.72
Location	.475^{c}	.056	1.44	1.61	1.80
Organization	.402^{c}	.105	1.22	1.50	1.84
Work of Art	.511	.212	1.11	1.67	2.56
Consumer Good	1.23	.785	.880	3.42	22.6
Other	.594^{c}	.120	1.44	1.81	2.30
Similar Entity Mention	.002^{a}	.001	1.00	1.00	1.00

Model $\chi^2(1) < .001$; $^{a}p < .05$ $^{b}p < .01$ $^{c}p < .001$

Table 25: Logistic Regression for DS-QA EM Score on SearchQA