Appendix B

Detailed annotation results

Here, we report all our annotations in detail, with absolute and relative numbers. Note, that numbers from sub-categories do not necessarily add up to the higher level category, because an example might contain features from the same higher-level category. (for example if an example requires both Bridge and Constraint type of reasoning, it will still count as a single example towards the *Operations* counter).

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Answer	50	100.0	50	100.0	50	100.0	50	100.0	50	100.0	50	100.0
Span	25	50.0	49	98.0	50	100.0	36	72.0	38	76.0	20	40.0
Paraphrasing	4	8.0	0	0.0	0	0.0	24	48.0	0	0.0	0	0.0
Unanswerable	20	40.0	0	0.0	0	0.0	0	0.0	12	24.0	0	0.0
Abstraction	1	2.0	1	2.0	0	0.0	12	24.0	0	0.0	31	62.0

Table 1: Detailed Answer Type results. We calculate percentages relative to the number of examples in the sample.

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Factual Correctness	23	46.0	13	26.0	4	8.0	19	38.0	21	42.0	5	10.0
Debatable	17	34.0	12	24.0	4	8.0	14	28.0	16	32.0	5	10.0
Arbitrary Selection	9	18.0	2	4.0	0	0.0	0	0.0	5	10.0	1	2.0
Arbitrary Precision	3	6.0	5	10	1	2.0	4	8.0	7	14.0	2	4.0
Conjunction or Isolated	0	0.0	0	0	0	0.0	5	10.0	0	0.0	0	0.0
Other	5	10.0	5	10	3	6.0	5	10.0	4	8.0	2	4.0
Wrong	6	12.0	1	2.0	0	0.0	5	10.0	5	10.0	0	0.0

Table 2: Detailed results for the annotation of factual correctness.

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Knowledge	3	10.0	8	16.0	19	38.0	11	22.0	6	15.8	20	40.0
World	0	0.0	3	6.0	12	24.0	3	6.0	1	2.6	6	12.0
Cultural	0	0.0	1	2.0	3	6.0	1	2.0	0	0.0	0	0.0
Geographical	0	0.0	0	0.0	2	4.0	0	0.0	1	2.6	0	0.0
Legal	0	0.0	0	0.0	2	4.0	0	0.0	0	0.0	0	0.0
Political	0	0.0	1	2.0	2	4.0	0	0.0	0	0.0	1	2.0
Technical	0	0.0	0	0.0	1	2.0	2	4.0	0	0.0	0	0.0
DomainSpecific	0	0.0	1	2.0	2	4.0	0	0.0	0	0.0	5	10.0
Intuitive	3	10.0	5	10.0	9	18.0	8	16.0	5	13.2	14	28.0

Table 3: Detailed results for the annotation of factual correctness. We calculate percentages relative to the number of examples that were annotated to be not unanswerable.

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
Reasoning	30	1.0	50	1.0	50	1.0	50	1.0	38	1.0	50	1.0
Mathematics	0	0.0	3	6.0	0	0.0	1	2.0	0	0.0	34	68.0
Subtraction	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	20	40.0
Addition	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	4.0
Ordering	0	0.0	3	6.0	0	0.0	0	0.0	0	0.0	11	22.0
OtherArithmethic	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	4.0
Linguistics	2	6.7	0	0.0	2	4.0	7	14.0	0	0.0	2	4.0
Negation	0	0.0	0	0.0	2	4.0	1	2.0	0	0.0	2	4.0
Con-/Disjunction	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
Conditionals	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Monotonicity	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Quantifiers	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Exists	2	6.7	0	0.0	0	0.0	4	8.0	0	0.0	0	0.0
ForAll	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
Operations	2	6.7	36	72.0	0	0.0	1	2.0	2	5.3	8	16.0
Join	1	3.3	23	46.0	0	0.0	0	0.0	0	0.0	0	0.0
Comparison	1	3.3	2	4.0	0	0.0	0	0.0	0	0.0	0	0.0
Count	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	7	14.0
Constraint	0	0.0	11	22.0	0	0.0	1	2.0	2	5.3	6	12.0
Intersection	0	0.0	4	8.0	0	0.0	0	0.0	0	0.0	0	0.0
Temporal	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Spatial	0	0.0	1	2.0	0	0.0	0	0.0	0	0.0	0	0.0
Causal	0	0.0	0	0.0	2	4.0	15	30.0	0	0.0	0	0.0
ByExclusion	0	0.0	0	0.0	17	34.0	1	2.0	0	0.0	0	0.0
Retrieval	26	86.7	13	26.0	31	62.0	30	60.0	38	100.0	9	18.0

Table 4: Detailed reasoning results. We calculate percentages relative to the number of examples that are not unanswerable, i.e. require reasoning to obtain the answer according to our definition.

	MSMarco		HotpotQA		ReCoRd		MultiRC		NewsQA		DROP	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
LinguisticComplexity	18	60.0	49	98.0	42	97.7	43	87.8	34	89.5	46	92.0
Lexical Variety	14	46.7	44	88.0	36	83.7	35	71.4	30	78.9	42	84.0
Redundancy	12	40.0	38	76.0	19	44.2	31	63.3	27	71.1	30	60.0
Lex Entailment	0	0.0	1	2.0	1	2.3	2	4.1	0	0.0	0	0.0
Dative	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Synonym	7	23.3	7	14.0	25	58.1	11	22.4	15	39.5	12	24.0
Abbreviation	2	6.7	4	8.0	1	2.3	1	2.0	0	0.0	7	14.0
Symmetry	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Syntactic Variety	2	6.7	10	20.0	2	4.7	2	4.1	1	2.6	4	8.0
Nominalisation	0	0.0	6	12.0	0	0.0	1	2.0	0	0.0	2	4.0
Genitive	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Voice	2	6.7	4	8.0	2	4.7	1	2.0	1	2.6	2	4.0
Lexical Ambiguity	7	23.3	32	64.0	26	60.5	34	69.4	11	28.9	7	14.0
Coreference	7	23.3	32	64.0	26	60.5	34	69.4	11	28.9	7	14.0
Restrictivity	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Factivity	0	0.0	0	0.0	0	0.0	1	2.0	0	0.0	0	0.0
Syntactic Ambiguity	2	6.7	22	44.0	6	14.0	7	14.3	9	23.7	9	18.0
Preposition	0	0.0	1	2.0	0	0.0	0	0.0	0	0.0	0	0.0
Ellipse/Implicit	2	6.7	3	6.0	3	7.0	3	6.1	1	2.6	8	16.0
Listing	0	0.0	16	32.0	5	11.6	6	12.2	1	2.6	13	26.0
Scope	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Relative	0	0.0	20	40.0	3	7.0	4	8.2	8	21.1	3	6.0

Table 5: Detailed linguistic feature results. We calculate percentages relative to the number of examples that were annotated to contain supporting facts.