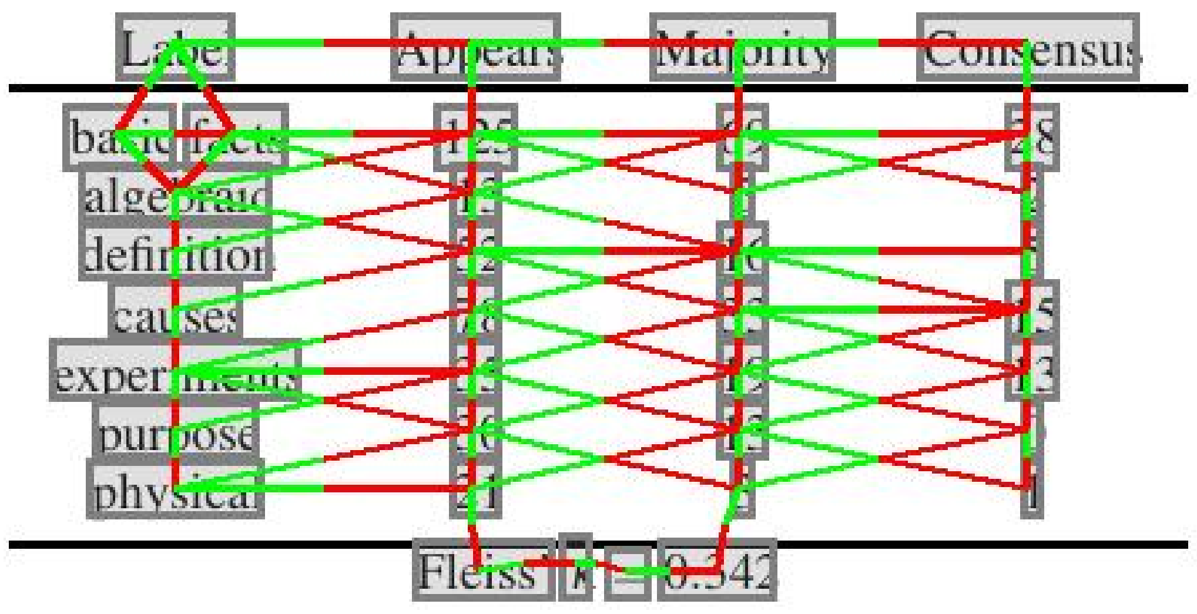
ground truth. This method of thinking is largely credited to Condorcet (de Caritat, 1785; Young, 1988) and there is recent work in characterizing other voting rules as maximum likelihood estimators (MLEs) (Conitzer et al., 2009). The Kemeny voting rule is the MLE of the Condorcet Noise Model, in which pairwise inversions of the preference order happen uniformly at random (Young, 1988, 1995). Hence, if we assume all annotators make pairwise errors uniformly at random then Kemeny is the MLE of label orders they report.



Label	Appears	Majority	Consensus
basic facts	125	66	28
algebraic	15	7	2
definition	52	16	8
causes	78	32	15
experiments	35	19	13
purpose	36	12	[illegible]
physical	21	5	1
Fleiss' $\kappa = 0.342$			

Table 1: Pairwise inter-rater agreement for Knowledge Labels, along with the mean and Fleiss' κ for survey responses.

Label	Appears	Majority	Consensus
linguistic	66	31	8
algebraic	15	8	3
explanation	80	22	4
hypothetical	62	21	6
multihop	45	6	0
comparison	46	13	3
qn logic	78	33	2
physical	18	3	0
analogy	4	1	1
Fleiss' $\kappa = -0.683$			

Table 2: Pairwise inter-rater agreement for Reasoning Labels, along with the mean and Fleiss' κ for survey responses.

4.1.1 Knowledge Labels

We achieve $\kappa = 0.342$, which means that our raters did a reasonable job of independently agreeing on the types of knowledge required to answer the questions. The mean Kemeny score of the consensus ranking for each question is 2.57, meaning that on average there are less than three flips required to get from the consensus ranking to each of the annotators' rankings. The most frequent label in the first position was *basic facts*, followed by *causes*. Overall, there was a reasonable amount of consensus between the raters for knowledge type: $^{64}/_{192}$ questions had a consensus amongst all the raters. Taken together, our results on knowledge type indicate that most questions deal with *basic facts*, *causes*, and *definitions*; and that labeling can be done reliably.

4.1.2 Reasoning Labels

The inter-rater agreement score for the reasoning labels tells a very different story from the knowledge labels. The agreement was $\kappa = -0.683$, which indicates that raters did not agree above chance on their labels. Strong evidence for this comes from the fact that only $^{27}/_{192}$ questions had a consensus label. This may be due to the fact that we allow multiple labels, and the annotators simply disagree on the *order* of the labels. However, the score of the consensus ranking for each question is 6.57, which indicates that on average the ordering of the labels is quite far apart.

Considering the histogram in Figure 3, we see that *qn logic*, *linguistic*, and *explanation* are the most frequent label types; this may indicate that getting better at understanding the questions themselves could lead to a big boost for reasoners. For Figure 4, we have merged the first and second label (if present) for all annotators. Now, the set of all possible labels is all singletons as well as all pairs of labels. Comparing this histogram to the one in Figure 3, we see that while *linguistic* and *explanation* remain somewhat unchanged, the *qn logic* label becomes very spread out across the types. This is more support for our hypothesis that annotators may be disagreeing on the ordering of the labels, rather than the content itself.

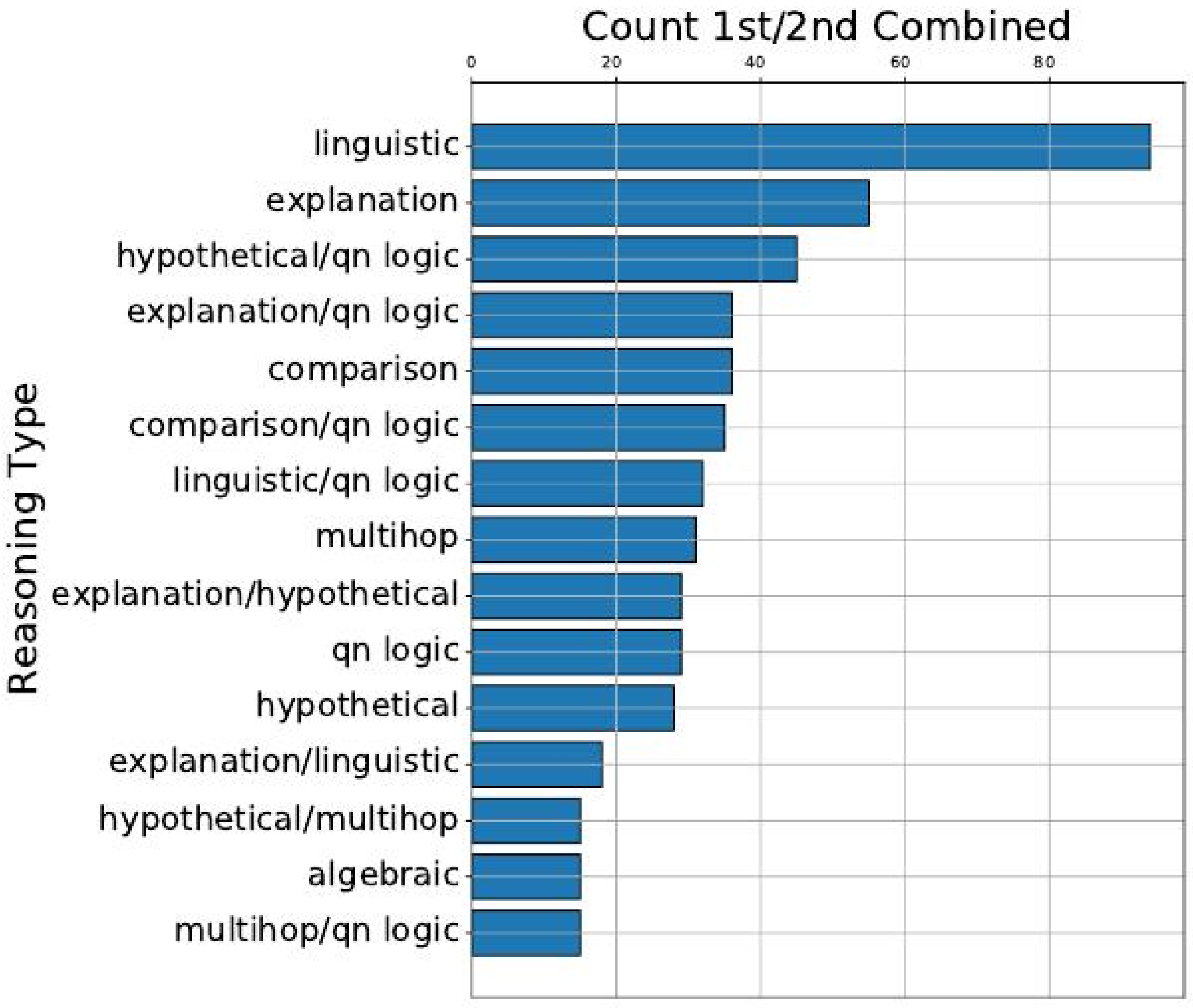


Figure 4: Histogram of the reasoning labels when we combine the first and (if present) second label of every annotator. The count refers to annotations.

4.2 Search Results

To quantitatively measure the efficacy of the annotated context (search results) from the interface, we evaluated 47 questions and their respective human-annotated relevant sentences with a pretrained `DrQA` model (Chen et al., 2017). We compared this to a baseline which only returned the sentences retrieved by using the text of the