TABLE 1. Performance on the test sets (with whichever number of additional context sentences achieved the best performance on the development set). Macro-average scores represent an average of each label's statistic (Undefined* treated as 0), weighted by the true number of that label appearing in the test set; this evaluation-stage weighting allows for aggregate interpretation. For Task 1, the neural baseline is an LSTM. For Task 2, the neural baseline is a feed-forward neural network. RoBERTa's overall performance in precision, recall, and F1 scores for both Tasks is significantly higher than the corresponding metrics for the baseline models at $p < 0.001$ (except for RoBERTa's Task 1 precision, which is lower than the neural baseline's precision).

	Keyword baseline (logistic regression)			Neural baseline (feed-forward or LSTM)			Finetuned RoBERTa		
	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
Task 1	0.409	0.532	0.462	0.603	0.417	0.493	0.490	0.611	**0.544**
Task 2 Terrorism	0.424	0.452	0.438	0.435	0.645	0.519	0.487	0.613	0.543
Task 2 Precedent (Int'l/Dom)	0.000	0.000	Und.*	0.000	0.000	Und.*	Und.*	0.000	Und.*
Task 2 Denial	0.313	0.185	0.233	0.222	0.074	0.111	0.650	0.481	0.553
Task 2 Political	0.556	0.357	0.435	0.000	0.000	Und.*	0.500	0.357	0.417
Task 2 Development and Unity	0.000	0.000	Und.*	1.000	0.333	0.500	0.167	0.333	0.222
Task 2 Legal Authorization	0.714	0.455	0.556	0.442	0.576	0.500	0.472	0.515	0.493
Task 2 Emergency Situation	0.348	0.516	0.416	0.231	0.194	0.211	0.472	0.548	0.507
Task 2 Law and Order	0.333	0.308	0.320	0.667	0.154	0.250	0.538	0.538	0.538
Task 2 Utilitarianism/Deterrence	0.143	0.143	0.143	0.043	0.143	0.067	0.750	0.429	0.545
Task 2 Miscellaneous	0.421	0.500	0.457	0.352	0.396	0.373	0.523	0.479	0.500
Task 2 macro-average	0.429	0.404	0.405	0.334	0.332	0.319	0.518	0.505	**0.505**