Table 1: Test on well-known fact-checkers. Fact-checks from this test set are published by the same set of fact-checkers in the train set when split strategy is training on well-known fact-checkers only, and the test set is unchanged when split strategy is mixing half from new-coming fact-checkers.

(a) Evaluation on claims. ROUGE-1 and -L F1/precision/recall scores and tagged % are reported.

Split strategy	Lead token	Tagger	ROUGE-1			ROUGE-L			Tagged
			F1	Precision	Recall	F1	Precision	Recall	%
Training on well-known fact-checkers only	Baseline		0.183 (0.183)	0.300 (0.300)	0.141 (0.141)	0.133 (0.133)	0.267 (0.267)	0.124 (0.124)	1.000
	[CLS]	Fluent	0.636 (0.853)	0.669 (0.897)	0.633 (0.850)	0.610 (0.819)	0.658 (0.883)	0.623 (0.837)	0.745
		Concise	0.592 (0.864)	0.615 (0.897)	0.596 (0.870)	0.570 (0.832)	0.607 (0.886)	0.586 (0.854)	0.686
	Paragraph position	Fluent	0.638 (0.854)	0.674 (0.902)	0.637 (0.853)	0.612 (0.819)	0.663 (0.888)	0.627 (0.839)	0.747
		Concise	0.646 (0.866)	0.664 (0.889)	0.652 (0.873)	0.625 (0.837)	0.656 (0.878)	0.641 (0.859)	0.747
Mixing half from under-represented fact-checkers	Baseline		0.183 (0.183)	0.300 (0.300)	0.141 (0.141)	0.133 (0.133)	0.267 (0.267)	0.124 (0.124)	1.000
	[CLS]	Fluent	0.628 (0.858)	0.656 (0.896)	0.628 (0.858)	0.604 (0.826)	0.645 (0.882)	0.617 (0.844)	0.732
		Concise	0.602 (0.846)	0.625 (0.878)	0.604 (0.849)	0.581 (0.816)	0.618 (0.868)	0.595 (0.835)	0.712
	Paragraph position	Fluent	0.638 (0.874)	0.668 (0.916)	0.636 (0.872)	0.615 (0.842)	0.659 (0.903)	0.626 (0.858)	0.730
		Concise	0.632 (0.866)	0.654 (0.896)	0.639 (0.876)	0.611 (0.837)	0.646 (0.886)	0.629 (0.861)	0.730

(b) Evaluation on claimants. ROUGE-1 and -L F1/precision/recall scores and tagged % are reported.

Split strategy	Lead token	Tagger	ROUGE-1			ROUGE-L			Tagged
			F1	Precision	Recall	F1	Precision	Recall	%
Training on well-known fact-checkers only	Baseline		0.237 (0.237)	0.181 (0.181)	0.352 (0.352)	0.199 (0.199)	0.181 (0.181)	0.352 (0.352)	1.000
	[CLS]	Fluent	0.769 (0.894)	0.803 (0.934)	0.759 (0.883)	0.760 (0.883)	0.803 (0.933)	0.759 (0.882)	0.860
		Concise	0.784 (0.907)	0.789 (0.913)	0.783 (0.906)	0.781 (0.904)	0.789 (0.913)	0.783 (0.906)	0.864
	Paragraph position	Fluent	0.794 (0.889)	0.821 (0.919)	0.789 (0.884)	0.787 (0.880)	0.821 (0.919)	0.789 (0.883)	0.894
		Concise	0.839 (0.928)	0.852 (0.943)	0.834 (0.923)	0.834 (0.923)	0.852 (0.943)	0.834 (0.923)	0.904
Mixing half from under-represented fact-checkers	Baseline		0.237 (0.237)	0.181 (0.181)	0.352 (0.352)	0.199 (0.199)	0.181 (0.181)	0.352 (0.352)	1.000
	[CLS]	Fluent	0.778 (0.881)	0.813 (0.921)	0.771 (0.873)	0.770 (0.872)	0.813 (0.921)	0.771 (0.873)	0.883
		Concise	0.807 (0.925)	0.814 (0.932)	0.807 (0.924)	0.804 (0.922)	0.814 (0.932)	0.807 (0.924)	0.873
	Paragraph position	Fluent	0.792 (0.895)	0.816 (0.922)	0.789 (0.891)	0.785 (0.887)	0.816 (0.922)	0.789 (0.891)	0.885
		Concise	0.822 (0.922)	0.834 (0.936)	0.817 (0.917)	0.817 (0.916)	0.834 (0.936)	0.817 (0.917)	0.891

(c) Evaluation on verdicts. ROUGE-1 and -L F1/precision/recall scores and tagged % are reported.

Split strategy	Lead token	Tagger	ROUGE-1			ROUGE-L			Tagged
			F1	Precision	Recall	F1	Precision	Recall	%
Training on well-known fact-checkers only	Baseline		0.660 (0.660)	0.638 (0.638)	0.702 (0.704)	0.645 (0.645)	0.638 (0.638)	0.704 (0.704)	1.000
	[CLS]	Fluent	0.931 (0.975)	0.934 (0.979)	0.930 (0.974)	0.930 (0.974)	0.934 (0.979)	0.930 (0.974)	0.955
		Concise	0.938 (0.971)	0.940 (0.973)	0.938 (0.970)	0.937 (0.970)	0.940 (0.973)	0.938 (0.970)	0.967
	Paragraph position	Fluent	0.940 (0.978)	0.942 (0.980)	0.939 (0.978)	0.939 (0.978)	0.942 (0.980)	0.939 (0.978)	0.961
		Concise	0.941 (0.975)	0.944 (0.979)	0.940 (0.974)	0.939 (0.974)	0.944 (0.979)	0.940 (0.974)	0.965
Mixing half from under-represented fact-checkers	Baseline		0.660 (0.660)	0.638 (0.638)	0.702 (0.704)	0.645 (0.645)	0.638 (0.638)	0.704 (0.704)	1.000
	[CLS]	Fluent	0.931 (0.972)	0.933 (0.973)	0.931 (0.971)	0.931 (0.971)	0.933 (0.973)	0.931 (0.971)	0.959
		Concise	0.936 (0.977)	0.938 (0.978)	0.936 (0.976)	0.935 (0.976)	0.938 (0.978)	0.936 (0.976)	0.959
	Paragraph position	Fluent	0.936 (0.975)	0.940 (0.978)	0.935 (0.973)	0.935 (0.973)	0.940 (0.978)	0.935 (0.973)	0.961
		Concise	0.938 (0.972)	0.939 (0.974)	0.938 (0.972)	0.937 (0.971)	0.939 (0.974)	0.938 (0.972)	0.965

Table 2: Test on under-represented fact-checkers. Fact-checks from this test set are published by different fact-checkers than the ones in the train set when split strategy is training on well-known fact-checkers only, and the test set is randomly sampled to half when split strategy is mixing half from new-coming fact-checkers.

(a) Evaluation on claims. ROUGE-1 and -L F1/precision/recall scores and tagged % are reported.

Split strategy	Lead token	Tagger	ROUGE-1			ROUGE-L			Tagged
			F1	Precision	Recall	F1	Precision	Recall	%
Training on well-known fact-checkers only	Baseline		0.175 (0.175)	0.372 (0.372)	0.122 (0.122)	0.114 (0.114)	0.324 (0.324)	0.106 (0.106)	1.000
	[CLS]	Fluent	0.444 (0.725)	0.483 (0.788)	0.443 (0.724)	0.413 (0.675)	0.462 (0.755)	0.431 (0.703)	0.612
		Concise	0.386 (0.713)	0.406 (0.748)	0.406 (0.749)	0.363 (0.669)	0.399 (0.736)	0.398 (0.733)	0.542
	Paragraph position	Fluent	0.519 (0.728)	0.566 (0.794)	0.517 (0.725)	0.485 (0.681)	0.548 (0.768)	0.504 (0.707)	0.713
		Concise	0.527 (0.738)	0.532 (0.744)	0.559 (0.781)	0.492 (0.688)	0.515 (0.721)	0.540 (0.755)	0.715
Mixing half from under-represented fact-checkers	Baseline		0.174 (0.174)	0.380 (0.380)	0.120 (0.120)	0.110 (0.110)	0.327 (0.327)	0.102 (0.102)	1.000
	[CLS]	Fluent	0.381 (0.785)	0.403 (0.832)	0.375 (0.773)	0.359 (0.740)	0.390 (0.804)	0.363 (0.749)	0.485
		Concise	0.424 (0.780)	0.421 (0.774)	0.447 (0.822)	0.396 (0.729)	0.404 (0.743)	0.427 (0.786)	0.544
	Paragraph position	Fluent	0.495 (0.761)	0.540 (0.830)	0.489 (0.752)	0.461 (0.708)	0.518 (0.796)	0.473 (0.726)	0.650
		Concise	0.519 (0.782)	0.544 (0.819)	0.536 (0.807)	0.487 (0.734)	0.529 (0.797)	0.522 (0.787)	0.664

(b) Evaluation on claimants. ROUGE-1 and -L F1/precision/recall scores and tagged % are reported.

Split strategy	Lead token	Tagger	ROUGE-1			ROUGE-L			Tagged
			F1	Precision	Recall	F1	Precision	Recall	%
Training on well-known fact-checkers only	Baseline		0.132 (0.132)	0.114 (0.114)	0.204 (0.204)	0.115 (0.115)	0.114 (0.114)	0.204 (0.204)	1.000
	[CLS]	Fluent	0.264 (0.567)	0.364 (0.782)	0.236 (0.506)	0.241 (0.517)	0.364 (0.782)	0.236 (0.506)	0.465
		Concise	0.323 (0.650)	0.379 (0.764)	0.304 (0.612)	0.306 (0.617)	0.379 (0.764)	0.304 (0.613)	0.496
	Paragraph position	Fluent	0.377 (0.635)	0.510 (0.859)	0.342 (0.576)	0.345 (0.580)	0.510 (0.859)	0.342 (0.575)	0.594
		Concise	0.462 (0.709)	0.549 (0.843)	0.436 (0.670)	0.436 (0.669)	0.549 (0.843)	0.436 (0.670)	0.651
Mixing half from under-represented fact-checkers	Baseline		0.144 (0.144)	0.128 (0.128)	0.205 (0.205)	0.129 (0.129)	0.128 (0.128)	0.205 (0.205)	1.000
	[CLS]	Fluent	0.481 (0.750)	0.535 (0.835)	0.471 (0.735)	0.466 (0.727)	0.535 (0.835)	0.470 (0.734)	0.641
		Concise	0.560 (0.823)	0.592 (0.871)	0.557 (0.820)	0.546 (0.804)	0.592 (0.871)	0.557 (0.820)	0.680
	Paragraph position	Fluent	0.550 (0.717)	0.639 (0.832)	0.528 (0.688)	0.530 (0.690)	0.639 (0.832)	0.527 (0.687)	0.767
		Concise	0.575 (0.781)	0.599 (0.813)	0.581 (0.789)	0.566 (0.769)	0.599 (0.813)	0.581 (0.788)	0.736

(c) Evaluation on verdicts. ROUGE-1 and -L F1/precision/recall scores and tagged % are reported.

Split strategy	Lead token	Tagger	ROUGE-1			ROUGE-L			Tagged
			F1	Precision	Recall	F1	Precision	Recall	%
Training on well-known fact-checkers only	Baseline		0.392 (0.392)	0.385 (0.385)	0.409 (0.409)	0.385 (0.385)	0.385 (0.385)	0.409 (0.409)	1.000
	[CLS]	Fluent	0.429 (0.806)	0.484 (0.910)	0.421 (0.792)	0.419 (0.788)	0.484 (0.910)	0.421 (0.792)	0.532
		Concise	0.451 (0.832)	0.484 (0.892)	0.446 (0.821)	0.441 (0.813)	0.484 (0.892)	0.446 (0.821)	0.543
	Paragraph position	Fluent	0.367 (0.733)	0.451 (0.902)	0.359 (0.718)	0.357 (0.714)	0.451 (0.902)	0.359 (0.717)	0.500
		Concise	0.473 (0.832)	0.520 (0.914)	0.467 (0.822)	0.463 (0.815)	0.520 (0.914)	0.467 (0.822)	0.569
Mixing half from under-represented fact-checkers	Baseline		0.357 (0.357)	0.351 (0.351)	0.374 (0.374)	0.351 (0.351)	0.352 (0.352)	0.374 (0.374)	1.000
	[CLS]	Fluent	0.441 (0.752)	0.511 (0.870)	0.437 (0.744)	0.429 (0.731)	0.511 (0.870)	0.437 (0.744)	0.587
		Concise	0.485 (0.797)	0.565 (0.929)	0.470 (0.771)	0.473 (0.776)	0.565 (0.929)	0.470 (0.771)	0.609
	Paragraph position	Fluent	0.475 (0.712)	0.573 (0.859)	0.469 (0.704)	0.462 (0.694)	0.573 (0.859)	0.469 (0.703)	0.667
		Concise	0.482 (0.797)	0.562 (0.931)	0.464 (0.768)	0.468 (0.775)	0.562 (0.931)	0.464 (0.768)	0.604