A Appendix

A.1 Benchmarks

Table 1: Reviewed NLP benchmarks for languages with at least 10 million speakers. The benchmarks are sorted by the language and name.

Name	# of tasks	Languages
ALUE (Seelawi et al., 2021)	6	Arabic
CBLUE (Zhang et al., 2021)	6	Chinese
CLUE (Xu et al., 2020)	7	Chinese
CUGE (Yao et al., 2021)	13	Chinese
FewCLUE (Xu et al., 2021)	7	Chinese
LOT (Guan et al., 2022)	3	Chinese
DecaNLP (McCann et al., 2018)	9	English
Dynabench (Kiela et al., 2021)	5	English
GLGE (Liu et al., 2021)	4	English
GLUE (Wang et al., 2018)	6	English
KILT (Petroni et al., 2021)	5	English
SentEval (Conneau and Kiela, 2018)	7	English
SuperGLUE (Wang et al., 2019)	5	English
FLUE (Le et al., 2020)	7	French
IndicNLPSuite (Kakwani et al., 2020)	10	Indian
IndoNLG (Cahyawijaya et al., 2021)	4	Indonesian
IndoNLU (Wilie et al., 2020)	7	Indonesian
IndoLEM (Koto et al., 2020)	7	Indonesian
KLUE (Park et al., 2021)	8	Korean
KOBEST (Kim et al., 2022)	5	Korean
XGLUE (Liang et al., 2020)	9	Multilingual (13 languages)
GEM (Gehrmann et al., 2021)	9	Multilingual (15 languages)
XTREME (Hu et al., 2020)	6	Multilingual (17 languages)
E-KAR (Chen et al., 2022)	2	Multilingual (2 languages)
(Wang et al., 2022)	3	Multilingual (2 languages)
ParsiNLU (Khashabi et al., 2021)	6	Persian
Persian NLP Benchmark (Fallahnejad and Zarezade, 2021)	9	Persian
KLEJ (Rybak et al., 2020)	7	Polish
PLUE (Gomes, 2020)	6	Portugese
LiRO (Dumitrescu et al., 2021)	10	Romanian
RuMedBench (Blinov et al., 2022)	4	Russian
RussianSuperGLUE (Shavrina et al., 2020)	4	Russian
GLUES (Canete et al., 2022)	6	Spanish
SpanishGLUE (Canete et al., 2022)	7	Spanish
Mukayese (Safaya et al., 2022)	8	Turkish

A.2 Hyperparameter search

All hyperparameters used to create the first version of benchmark are presented in Table 2. 2

Table 2: Hyperparameters for finetuning the language models.

Max. sequence length	512
Classifier dropout	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]
No. finetuned layers	[0, 1, 2, 3, 4]
Learning rate	[1e-6, 5e-6, 1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3]
Max. no. epochs	[2, 3, 5, 10, 15, 20]
Batch size	[16, 32, 64]
Optimizer	[Adam, AdamW]
Weight decay	[1e-4, 1e-3, 1e-2, 1e-1, 0]
Adam epsilon	[1e-8, 1e-7, 1e-6, 1e-5, 1e-4]
Use optimizer scheduler	[true, false]
Optimizer scheduler warmup steps	[0, 25, 50, 100, 200]

A.3 Metrics

In this section we provide other metrics that were calculated during experimental phase – tables 3 to 12.

Table 3: Accuracy performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**94.02 ± 0.33**	93.92 ± 0.16	92.30 ± 0.25	93.48 ± 0.28	86.58 ± 0.68
DYK*	**90.40 ± 0.78**	87.66 ± 0.22	87.60 ± 0.64	86.82 ± 0.65	83.79 ± 0.88
PolEmo 2.0 In-Domain*	90.30 ± 0.28	**90.55 ± 0.47**	87.53 ± 0.43	87.59 ± 0.81	85.57 ± 0.46
PolEmo 2.0 Out-Domain*	75.06 ± 1.86	**75.30 ± 1.69**	69.31 ± 2.87	69.84 ± 1.34	57.69 ± 5.07
PSC*	98.22 ± 0.20	98.59 ± 0.57	**99.11 ± 0.11**	99.04 ± 0.08	73.99 ± 0.54
Abusive Clauses	87.04 ± 0.54	**88.01 ± 0.59**	87.49 ± 0.58	87.13 ± 0.74	86.42 ± 0.38
AspectEmo	95.19 ± 0.07	**95.27 ± 0.24**	94.56 ± 0.11	94.65 ± 0.07	92.73 ± 0.25
KPWr NER	97.13 ± 0.05	**97.25 ± 0.04**	96.86 ± 0.04	95.90 ± 0.03	95.74 ± 0.03
NKJP POS	98.88 ± 0.02	**98.98 ± 0.00**	98.77 ± 0.02	98.79 ± 0.02	98.16 ± 0.02
PolEmo 2.0	88.20 ± 0.50	**90.71 ± 0.40**	87.05 ± 1.23	87.05 ± 0.50	85.22 ± 0.70
Political Advertising	96.46 ± 0.21	**96.49 ± 0.18**	96.12 ± 0.09	96.38 ± 0.10	95.71 ± 0.08
Punctuation Restoration	93.56 ± 0.08	**94.10 ± 0.06**	91.89 ± 0.26	92.38 ± 0.13	83.71 ± 0.19
Dialogue Acts (WIP)	76.50 ± 0.21	**77.08 ± 0.40**	76.30 ± 0.25	76.16 ± 0.24	76.65 ± 0.75
Mean rank	2.23	1.31	3.42	3.27	4.77

Table 4: Micro F1 performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**94.02 ± 0.33**	93.92 ± 0.16	92.30 ± 0.25	93.48 ± 0.28	86.58 ± 0.68
DYK*	**90.40 ± 0.78**	87.66 ± 0.22	87.60 ± 0.64	86.82 ± 0.65	83.79 ± 0.88
PolEmo 2.0 In-Domain*	90.30 ± 0.28	**90.55 ± 0.47**	87.53 ± 0.43	87.59 ± 0.81	85.57 ± 0.46
PolEmo 2.0 Out-Domain*	75.06 ± 1.86	**75.30 ± 1.69**	69.31 ± 2.87	69.84 ± 1.34	57.69 ± 5.07
PSC*	98.22 ± 0.20	98.59 ± 0.57	**99.11 ± 0.11**	99.04 ± 0.08	73.99 ± 0.54
Abusive Clauses	87.04 ± 0.54	**88.01 ± 0.59**	87.49 ± 0.58	87.13 ± 0.74	86.42 ± 0.38
AspectEmo	58.18 ± 0.32	**59.10 ± 1.32**	50.82 ± 0.42	51.55 ± 0.98	34.07 ± 1.36
KPWr NER	**76.90 ± 0.25**	76.58 ± 0.67	72.59 ± 0.32	66.49 ± 0.09	61.57 ± 0.44
NKJP POS	98.88 ± 0.02	**98.98 ± 0.00**	98.77 ± 0.02	98.79 ± 0.02	98.16 ± 0.02
PolEmo 2.0	88.20 ± 0.50	**90.71 ± 0.40**	87.05 ± 1.23	87.05 ± 0.50	85.22 ± 0.70
Political Advertising	68.02 ± 1.35	67.86 ± 0.88	65.25 ± 0.54	**68.93 ± 1.31**	60.94 ± 0.71
Punctuation Restoration	73.23 ± 0.33	**75.01 ± 0.23**	66.49 ± 0.41	68.56 ± 0.39	21.72 ± 0.85
Dialogue Acts (WIP)	76.50 ± 0.21	**77.08 ± 0.40**	76.30 ± 0.25	76.16 ± 0.24	76.65 ± 0.75
Mean rank	2.15	1.54	3.42	3.12	4.77

Table 5: Micro Precision performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**94.02 ± 0.33**	93.92 ± 0.16	92.30 ± 0.25	93.48 ± 0.28	86.58 ± 0.68
DYK*	**90.40 ± 0.78**	87.66 ± 0.22	87.60 ± 0.64	86.82 ± 0.65	83.79 ± 0.88
PolEmo 2.0 In-Domain*	90.30 ± 0.28	**90.55 ± 0.47**	87.53 ± 0.43	87.59 ± 0.81	85.57 ± 0.46
PolEmo 2.0 Out-Domain*	75.06 ± 1.86	**75.30 ± 1.69**	69.31 ± 2.87	69.84 ± 1.34	57.69 ± 5.07
PSC*	98.22 ± 0.20	98.59 ± 0.57	**99.11 ± 0.11**	99.04 ± 0.08	73.99 ± 0.54
Abusive Clauses	87.04 ± 0.54	**88.01 ± 0.59**	87.49 ± 0.58	87.13 ± 0.74	86.42 ± 0.38
AspectEmo	60.56 ± 0.99	**61.07 ± 2.81**	55.71 ± 1.16	56.05 ± 1.21	40.39 ± 2.12
KPWr NER	**75.09 ± 0.37**	73.85 ± 0.72	70.30 ± 0.34	64.48 ± 0.20	57.28 ± 0.46
NKJP POS	98.88 ± 0.02	**98.98 ± 0.00**	98.77 ± 0.02	98.79 ± 0.02	98.16 ± 0.02
PolEmo 2.0	88.20 ± 0.50	**90.71 ± 0.40**	87.05 ± 1.23	87.05 ± 0.50	85.22 ± 0.70
Political Advertising	64.46 ± 2.95	64.07 ± 2.50	66.09 ± 2.36	**68.18 ± 1.81**	58.85 ± 2.23
Punctuation Restoration	74.93 ± 0.33	**77.33 ± 0.55**	68.50 ± 2.34	70.64 ± 1.64	33.43 ± 1.44
Dialogue Acts (WIP)	76.50 ± 0.21	**77.08 ± 0.40**	76.30 ± 0.25	76.16 ± 0.24	76.65 ± 0.75
Mean rank	2.23	1.62	3.27	3.12	4.77

Table 6: Micro Recall performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**94.02 ± 0.33**	93.92 ± 0.16	92.30 ± 0.25	93.48 ± 0.28	86.58 ± 0.68
DYK*	**90.40 ± 0.78**	87.66 ± 0.22	87.60 ± 0.64	86.82 ± 0.65	83.79 ± 0.88
PolEmo 2.0 In-Domain*	90.30 ± 0.28	**90.55 ± 0.47**	87.53 ± 0.43	87.59 ± 0.81	85.57 ± 0.46
PolEmo 2.0 Out-Domain*	75.06 ± 1.86	**75.30 ± 1.69**	69.31 ± 2.87	69.84 ± 1.34	57.69 ± 5.07
PSC*	98.22 ± 0.20	98.59 ± 0.57	**99.11 ± 0.11**	99.04 ± 0.08	73.99 ± 0.54
Abusive Clauses	87.04 ± 0.54	**88.01 ± 0.59**	87.49 ± 0.58	87.13 ± 0.74	86.42 ± 0.38
AspectEmo	56.01 ± 1.09	**57.31 ± 0.95**	46.74 ± 0.89	47.77 ± 1.76	29.59 ± 2.39
KPWr NER	78.80 ± 0.13	**79.53 ± 0.64**	75.05 ± 0.35	68.63 ± 0.10	66.55 ± 0.51
NKJP POS	98.88 ± 0.02	**98.98 ± 0.00**	98.77 ± 0.02	98.79 ± 0.02	98.16 ± 0.02
PolEmo 2.0	88.20 ± 0.50	**90.71 ± 0.40**	87.05 ± 1.23	87.05 ± 0.50	85.22 ± 0.70
Political Advertising	72.09 ± 0.70	**72.21 ± 1.17**	64.53 ± 1.70	69.70 ± 1.12	63.30 ± 1.99
Punctuation Restoration	71.61 ± 0.38	**72.84 ± 0.81**	64.70 ± 1.99	66.65 ± 1.74	16.10 ± 0.74
Dialogue Acts (WIP)	76.50 ± 0.21	**77.08 ± 0.40**	76.30 ± 0.25	76.16 ± 0.24	76.65 ± 0.75
Mean rank	2.23	1.31	3.42	3.27	4.77

Table 7: Macro F1 performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**90.96 ± 0.73**	90.48 ± 0.20	88.95 ± 0.31	90.62 ± 0.27	82.62 ± 0.88
DYK*	**82.39 ± 1.43**	79.58 ± 0.59	75.87 ± 0.98	74.41 ± 1.15	58.93 ± 7.98
PolEmo 2.0 In-Domain*	88.10 ± 0.36	**88.34 ± 0.63**	85.32 ± 0.45	85.71 ± 0.40	83.75 ± 0.45
PolEmo 2.0 Out-Domain*	**57.31 ± 2.93**	57.08 ± 2.03	54.10 ± 3.82	54.29 ± 1.83	45.12 ± 3.40
PSC*	97.90 ± 0.24	98.33 ± 0.69	**98.95 ± 0.13**	98.87 ± 0.10	58.85 ± 1.49
Abusive Clauses	85.66 ± 0.58	**86.57 ± 0.91**	85.93 ± 0.66	85.74 ± 0.86	84.32 ± 0.71
AspectEmo	37.28 ± 0.71	**39.44 ± 1.74**	30.01 ± 0.58	31.48 ± 1.06	18.42 ± 0.98
KPWr NER	**54.22 ± 0.76**	52.68 ± 1.39	48.01 ± 0.76	40.21 ± 0.50	36.13 ± 0.44
NKJP POS	94.59 ± 0.56	**96.14 ± 0.38**	94.34 ± 0.61	94.54 ± 0.19	90.29 ± 0.51
PolEmo 2.0	86.78 ± 0.79	**89.33 ± 0.49**	85.89 ± 1.25	85.83 ± 0.47	84.12 ± 0.47
Political Advertising	61.42 ± 1.38	62.16 ± 0.14	58.94 ± 1.92	**62.52 ± 1.23**	56.68 ± 0.94
Punctuation Restoration	45.59 ± 0.38	**46.68 ± 0.61**	38.89 ± 0.91	41.31 ± 0.59	14.33 ± 1.94
Dialogue Acts (WIP)	49.54 ± 0.74	**51.11 ± 0.85**	50.20 ± 1.32	48.87 ± 0.90	49.05 ± 0.39
Mean rank	2.15	1.62	3.23	3.08	4.92

Table 8: Macro Precision performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	92.24 ± 0.81	**92.83 ± 0.69**	89.91 ± 1.30	91.20 ± 0.92	80.18 ± 1.24
DYK*	**83.47 ± 1.57**	77.85 ± 0.34	78.90 ± 1.65	77.30 ± 1.94	66.03 ± 13.81
PolEmo 2.0 In-Domain*	89.40 ± 0.55	**89.60 ± 0.95**	85.99 ± 0.53	85.86 ± 0.95	83.86 ± 0.39
PolEmo 2.0 Out-Domain*	59.01 ± 1.95	**59.29 ± 1.80**	56.17 ± 2.23	56.94 ± 1.52	52.14 ± 2.03
PSC*	97.81 ± 0.23	98.42 ± 0.34	**98.87 ± 0.11**	98.64 ± 0.08	74.30 ± 0.90
Abusive Clauses	84.94 ± 0.57	**86.08 ± 0.46**	85.54 ± 0.64	85.18 ± 0.71	84.99 ± 0.97
AspectEmo	40.08 ± 1.12	41.70 ± 3.99	**43.69 ± 3.19**	37.97 ± 3.17	23.57 ± 2.15
KPWr NER	**55.97 ± 1.02**	53.07 ± 1.77	50.45 ± 1.32	45.00 ± 1.25	36.64 ± 0.35
NKJP POS	95.91 ± 0.82	**97.23 ± 0.38**	96.59 ± 0.73	96.99 ± 0.23	92.50 ± 1.17
PolEmo 2.0	87.23 ± 0.43	**90.04 ± 0.61**	86.20 ± 1.38	86.09 ± 0.61	84.41 ± 0.82
Political Advertising	58.82 ± 3.33	59.62 ± 1.33	60.89 ± 3.68	**62.93 ± 1.75**	54.68 ± 2.62
Punctuation Restoration	49.48 ± 0.26	**53.19 ± 0.97**	46.22 ± 2.05	50.29 ± 6.67	23.85 ± 1.64
Dialogue Acts (WIP)	52.83 ± 0.97	**53.61 ± 1.10**	52.38 ± 1.06	50.95 ± 1.10	52.47 ± 1.65
Mean rank	2.69	1.62	2.77	3.15	4.77

Table 9: Macro Recall performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	89.92 ± 0.94	88.51 ± 0.67	88.33 ± 1.09	**90.15 ± 0.52**	85.78 ± 1.95
DYK*	81.45 ± 1.58	**81.95 ± 1.61**	73.77 ± 0.94	72.52 ± 1.69	58.11 ± 5.10
PolEmo 2.0 In-Domain*	87.40 ± 0.33	**87.62 ± 0.86**	84.83 ± 0.56	85.75 ± 0.23	83.71 ± 0.67
PolEmo 2.0 Out-Domain*	**60.99 ± 11.77**	56.17 ± 1.53	**71.26 ± 13.42**	66.69 ± 13.84	42.76 ± 3.81
PSC*	98.00 ± 0.33	98.25 ± 1.02	99.03 ± 0.15	**99.10 ± 0.15**	59.35 ± 1.01
Abusive Clauses	86.91 ± 1.09	**87.37 ± 1.73**	86.51 ± 1.04	87.01 ± 1.81	84.06 ± 1.76
AspectEmo	37.36 ± 0.72	**39.80 ± 1.30**	27.54 ± 0.69	29.80 ± 0.99	16.06 ± 1.17
KPWr NER	**55.74 ± 0.71**	55.52 ± 1.47	49.03 ± 0.55	40.81 ± 0.27	38.60 ± 0.47
NKJP POS	93.95 ± 0.64	**95.39 ± 0.45**	93.18 ± 0.50	93.28 ± 0.12	88.86 ± 0.33
PolEmo 2.0	86.50 ± 1.00	**88.94 ± 0.39**	85.68 ± 1.13	85.63 ± 0.40	83.95 ± 0.36
Political Advertising	65.04 ± 1.05	**65.60 ± 1.59**	57.70 ± 1.62	62.72 ± 1.13	59.55 ± 1.23
Punctuation Restoration	43.16 ± 0.36	**43.56 ± 1.06**	35.50 ± 1.27	38.11 ± 0.98	10.71 ± 1.90
Dialogue Acts (WIP)	49.54 ± 1.16	**51.79 ± 0.71**	50.91 ± 1.45	49.58 ± 0.81	49.36 ± 0.84
Mean rank	2.38	1.62	3.31	2.77	4.92

Table 10: Weighted F1 performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**93.93 ± 0.32**	93.80 ± 0.18	92.16 ± 0.24	93.42 ± 0.25	86.83 ± 0.54
DYK*	**90.25 ± 0.79**	88.08 ± 0.21	87.00 ± 0.58	86.20 ± 0.56	80.06 ± 2.89
PolEmo 2.0 In-Domain*	89.90 ± 0.28	**90.20 ± 0.51**	87.33 ± 0.45	87.64 ± 0.53	85.69 ± 0.40
PolEmo 2.0 Out-Domain*	75.61 ± 2.84	**76.10 ± 2.67**	69.68 ± 3.43	71.05 ± 1.20	60.33 ± 4.55
PSC*	98.22 ± 0.20	98.59 ± 0.58	**99.11 ± 0.11**	99.04 ± 0.08	68.62 ± 1.00
Abusive Clauses	87.22 ± 0.51	**88.11 ± 0.69**	87.57 ± 0.57	87.29 ± 0.73	86.32 ± 0.46
AspectEmo	58.27 ± 0.40	**59.16 ± 1.02**	50.38 ± 0.39	51.50 ± 0.95	33.45 ± 1.42
KPWr NER	**77.09 ± 0.24**	77.02 ± 0.69	72.47 ± 0.27	65.97 ± 0.17	61.84 ± 0.47
NKJP POS	98.88 ± 0.02	**98.98 ± 0.00**	98.76 ± 0.02	98.79 ± 0.02	98.16 ± 0.02
PolEmo 2.0	88.02 ± 0.61	**90.44 ± 0.41**	86.99 ± 1.18	87.08 ± 0.49	85.17 ± 0.53
Political Advertising	68.10 ± 1.34	68.10 ± 0.83	65.28 ± 0.56	**68.89 ± 1.33**	60.97 ± 0.61
Punctuation Restoration	72.41 ± 0.30	**73.77 ± 0.26**	65.29 ± 0.48	67.29 ± 0.38	21.44 ± 0.78
Dialogue Acts (WIP)	75.57 ± 0.30	**76.03 ± 0.18**	75.43 ± 0.25	75.21 ± 0.21	75.59 ± 0.59
Mean rank	2.19	1.5	3.46	3.08	4.77

Table 11: Weighted Precision performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**93.98 ± 0.36**	93.90 ± 0.16	92.21 ± 0.24	93.43 ± 0.29	87.47 ± 0.67
DYK*	**90.16 ± 0.81**	88.78 ± 0.50	86.78 ± 0.64	85.98 ± 0.64	78.95 ± 5.62
PolEmo 2.0 In-Domain*	89.90 ± 0.31	**90.22 ± 0.57**	87.27 ± 0.48	87.84 ± 0.32	85.86 ± 0.37
PolEmo 2.0 Out-Domain*	77.90 ± 2.30	**78.62 ± 2.63**	72.96 ± 2.41	74.62 ± 1.55	69.00 ± 2.77
PSC*	98.23 ± 0.20	98.60 ± 0.56	**99.11 ± 0.11**	99.05 ± 0.09	74.15 ± 0.69
Abusive Clauses	87.83 ± 0.67	**88.46 ± 0.93**	87.81 ± 0.55	88.05 ± 1.07	86.55 ± 0.42
AspectEmo	61.79 ± 1.23	**62.07 ± 2.14**	56.96 ± 1.27	57.36 ± 1.69	39.70 ± 1.43
KPWr NER	**76.43 ± 0.35**	75.60 ± 0.74	71.13 ± 0.25	65.41 ± 0.30	58.98 ± 0.43
NKJP POS	98.88 ± 0.02	**98.99 ± 0.00**	98.77 ± 0.02	98.80 ± 0.02	98.16 ± 0.02
PolEmo 2.0	87.98 ± 0.60	**90.41 ± 0.43**	87.00 ± 1.11	87.18 ± 0.46	85.22 ± 0.50
Political Advertising	64.97 ± 3.06	64.93 ± 2.40	66.61 ± 2.34	**68.50 ± 1.87**	59.17 ± 2.21
Punctuation Restoration	73.94 ± 0.27	**76.80 ± 0.49**	67.80 ± 1.84	69.84 ± 1.76	32.96 ± 1.15
Dialogue Acts (WIP)	75.96 ± 0.47	**76.30 ± 0.34**	75.75 ± 0.55	75.20 ± 0.27	75.99 ± 0.86
Mean rank	2.15	1.62	3.46	3.0	4.77

Table 12: Weighted Recall performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. The mean rank row is the average of a ranking established on the mean of model retrains. Values marked with **Bold** present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. **WIP** denotes the dataset for which we present preliminary results.

	HerBERT (base, cased)	HerBERT (large, cased)	PolBERT (base, cased)	PolBERT (base, uncased)	XLM-RoBERTa (paraphrase)
CDSC-E*	**94.02 ± 0.33**	93.92 ± 0.16	92.30 ± 0.25	93.48 ± 0.28	86.58 ± 0.68
DYK*	**90.40 ± 0.78**	87.66 ± 0.22	87.60 ± 0.64	86.82 ± 0.65	83.79 ± 0.88
PolEmo 2.0 In-Domain*	90.30 ± 0.28	**90.55 ± 0.47**	87.53 ± 0.43	87.59 ± 0.81	85.57 ± 0.46
PolEmo 2.0 Out-Domain*	75.06 ± 1.86	**75.30 ± 1.69**	69.31 ± 2.87	69.84 ± 1.34	57.69 ± 5.07
PSC*	98.22 ± 0.20	98.59 ± 0.57	**99.11 ± 0.11**	99.04 ± 0.08	73.99 ± 0.54
Abusive Clauses	87.04 ± 0.54	**88.01 ± 0.59**	87.49 ± 0.58	87.13 ± 0.74	86.42 ± 0.38
AspectEmo	56.01 ± 1.09	**57.31 ± 0.95**	46.74 ± 0.89	47.77 ± 1.76	29.59 ± 2.39
KPWr NER	78.80 ± 0.13	**79.53 ± 0.64**	75.05 ± 0.35	68.63 ± 0.10	66.55 ± 0.51
NKJP POS	98.88 ± 0.02	**98.98 ± 0.00**	98.77 ± 0.02	98.79 ± 0.02	98.16 ± 0.02
PolEmo 2.0	88.20 ± 0.50	**90.71 ± 0.40**	87.05 ± 1.23	87.05 ± 0.50	85.22 ± 0.70
Political Advertising	72.09 ± 0.70	**72.21 ± 1.17**	64.53 ± 1.70	69.70 ± 1.12	63.30 ± 1.99
Punctuation Restoration	71.61 ± 0.38	**72.84 ± 0.81**	64.70 ± 1.99	66.65 ± 1.74	16.10 ± 0.74
Dialogue Acts (WIP)	76.50 ± 0.21	**77.08 ± 0.40**	76.30 ± 0.25	76.16 ± 0.24	76.65 ± 0.75
Mean rank	2.23	1.31	3.42	3.27	4.77

References

Pavel Blinov, Arina Reshetnikova, Aleksandr Nesterov, Galina Zubkova, and Vladimir Kokh. 2022. Rumedbench: A russian medical language understanding benchmark. *arXiv preprint arXiv:2201.06499*.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2022. Spanish pre-trained bert model and evaluation data. In *Proceedings of the Practical Machine Learning for Developing Countries @ ICLR 2022*.

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-kar: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, et al. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Zohreh Fallahnejad and Ali Zarezade. 2021. Persian nlp benchmark. `https://github.com/Mofid-AI/persian-nlp-benchmark#persian-nlp-benchmark`.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

J. R. S. Gomes. 2020. Plue: Portuguese language understanding evaluation. `https://github.com/jubs12/PLUE`.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. Lot: A story-centric benchmark for evaluating chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 10:434–451.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, et al. 2021. Parsinlu: A suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Dohyeong Kim, Myeongjun Jang, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks. *arXiv preprint arXiv:2204.04541*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2021. Glge: A new general language generation evaluation benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 408–420.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. Klej: Comprehensive benchmark for polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish nlp strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. Alue: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022. A fine-grained interpretability evaluation benchmark for neural nlp. *arXiv preprint arXiv:2205.11097*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, et al. 2021. Cuge: A chinese language understanding and generation evaluation benchmark. *arXiv preprint arXiv:2112.13610*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.