A Details on Datasets

We report in Tab. 5 and Tab. 5 global statistics of the data described in . Overall our conclusions are drawn from a total of over 270k scores.

	# of tasks	# of instance
GLUE	105	15
SGLUE	24, 15	
XTREM	15	5

Table 4: Summary of the considered benchmarks. Overall the total number of the score is over 2010.

	# of tasks	# of instance
PC	19	240
TC	19	300
FLICKR	14	864
MLQE	10	7000
RSUM	15	2500
SEVAL	17	1600
TAC08	15	2976
TAC09	15	2596
TAC11	15	2376

Table 5: Summary of the considered datasets. Overall this benchmark is composed of over 276276 scores.

B Additional Experiments

In this section, we report additional experimental results including the details of the robustness to the scaling experiment (see Ssec. B.1), the ranking on XTREM (see Ssec. B.2), complete results on the experiments when adding adding/removing metrics/tasks when Task Level Information (see Sssec. B.3.1) and Instance Level Information is available (see Sssec. B.3.2). In the main paper we only report the aggregated score for the agreement analysis when instance level informatino is available, we report detailed results on Ssec. B.4.

B.1 Toy Experiment on Scale

We display in Fig. 7 the results of the toy experiment on scaling robustness. When corrupting one task by rescaling, we see that the error of the ranking induced by σ^{mean} increases to 1 while the error of ranking-based aggregation remains constant.

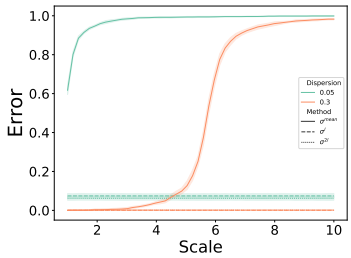
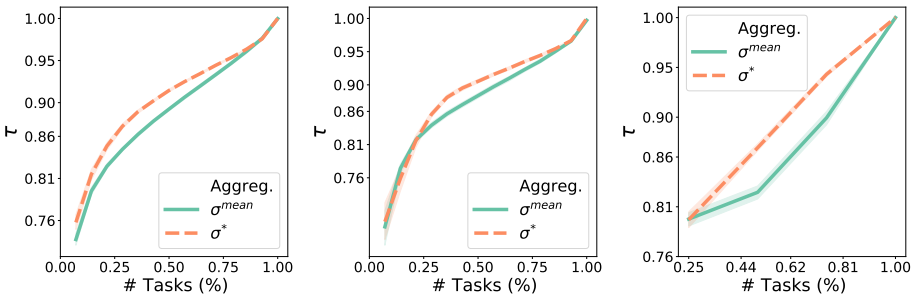


Figure 7: Synthetic Experiment on robustness to scaling. Error is measured in term of Kendall distance.

B.2 Ranking of SGLUE

We display in Tab. 6 the resulting ranking on the three considered benchmark. Although aggregation procedures tend to agree on good and bad systems, when changing the aggregation function, the



(a) GLUE (b) SGLUE (c) XTREM

Figure 8: Experiment on task addition/removal when Task level information is available.

rankings vary. Thus conclusion depending on the answer to the initial question "what are the best systems?" might change.

GLUE			SGLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (289)	Liam	0 (90.0)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (278)	Ms Alex	1 (89.4)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (268)	ERNIE	2 (89.3)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (263)	HUMAN	3 (89.2)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (256)	DEBERTA	5 (88.8)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (256)	Zirui	4 (88.8)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (234)	T5	6 (87.7)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (205)	Alibaba	7 (86.8)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (182)	Anuar	8 (86.1)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (181)	Huawei	11 (83.4)	9 (21)	Anony	10 (78.3)

Table 6: Qualitative analysis between ranking obtained with σ^* or σ^{mean}. Results in parenthesis report the score of the considered aggregation procedure.

B.3 Complete results on the task addition/removal experiments

B.3.1 Task Level Aggregation

Results of task addition/removal experiments when Task level information is available are reported in Fig. 8. Overall, we observe that ranking-based aggregation is more robust than mean-based aggregation.

B.3.2 Instance Level Aggregation

Results of task addition/removal experiments when instance-level information is available are reported in Fig. 9. Overall, we observe that ranking-based aggregation is more robust than mean-based aggregation.

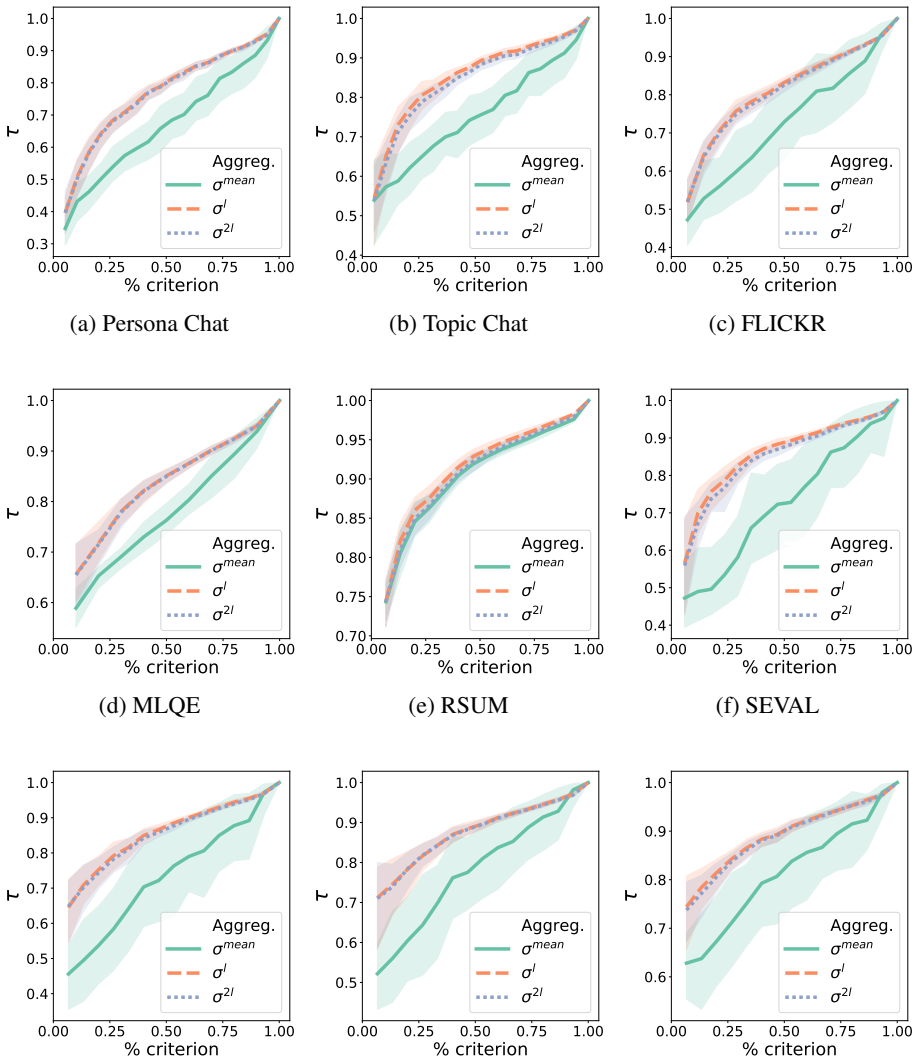
B.4 Agreement analysis on Instance Level Aggregation

To further complete the experiment on agreement analysis of Fig. 5.3 report results on individual tasks. We report in Fig. 10 the correlation of ranking when Top K systems for different values of K while Fig. 11 reports the agreement analysis.

B.5 Possible Extensions

For the futur we would like to extend the proposed ranking procedures to:

- Sequence Labelling Task [15, 16, 21] where instances are sequences, thus ordering matters.
- Emotion classification [22, 37, 45, 97] where score are continuous.



(a) Persona Chat (b) Topic Chat (c) FLICKR

(d) MLQE (e) RSUM (f) SEVAL

(g) TAC08 (h) TAC09 (i) TAC10

Figure 9: Experiment on task addition/removal when Instance level information is available.

- NLG [19, 20, 24–26, 88] where instance are sentences where both ordering and content matter.

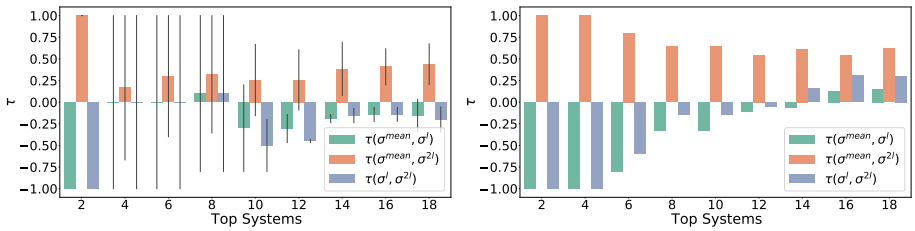
C Dispersion Analysis

In this section, we introduce the notion of dispersion across a set of different rankings $\sigma_1, \ldots, \sigma_T$.

C.1 Dispersion as a measure performance

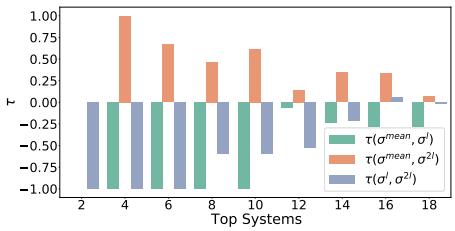
Suppose you have two ranking candidates, σ^A and σ^B, to summarize $\sigma_1, \ldots, \sigma_T$. A natural question consists in measuring the performance of σ^A and σ^B. Denoting by d the Kendall distance, a natural measure is

$$\sigma \mapsto \sum_{t=1}^{T} d(\sigma, \sigma_t). \tag{1}$$

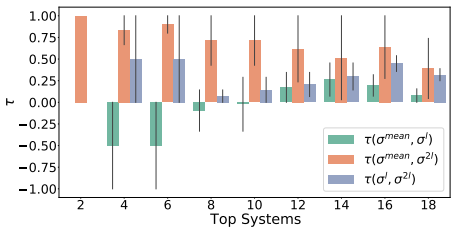


(a) Dialog

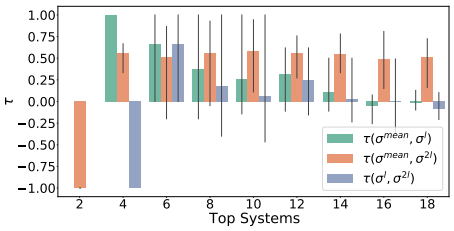
(b) FLICKR



(c) MLQE



(d) Summarizations Datasets



(e) TAC Datasets

Figure 10: Test size experiements

Of course, in our task-level framework, our candidate σ^* will achieve better performance than the mean since it is designed to minimize this quantity. From a probabilistic point of view, one can see $\sigma_1, \ldots, \sigma_T$ as an i.i.d. realization of a r.v. Σ on the symmetric group $\mathfrak{S}_N$. Denoting by $\mathbb{P}$ is law and $\mathbb{E}$ the associated expectation, Equation (1) is an approximation of

$$\sigma \mapsto \mathbb{E}[d(\sigma, \Sigma)]. \tag{2}$$

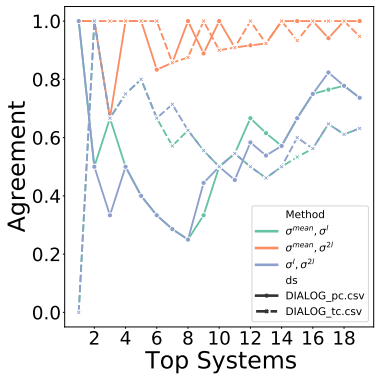
C.2 Dispersion to measure ranking complexity

In order to take into account the intrinsic complexity of ranking $\sigma_1, \ldots, \sigma_T$, a natural way would be to compute some kind of Dispersion among these permutations. Using the same notations as before, one can rely on the pairwise distance.

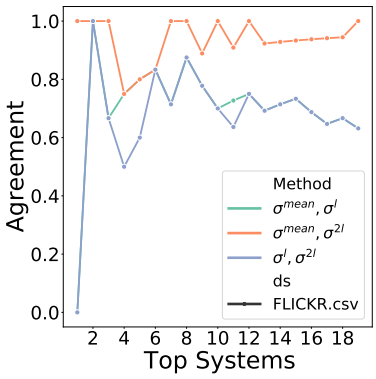
$$\sum_{t_1, t_2 = 1}^{T} d(\sigma_{t_1}, \sigma_{t_2}) \tag{3}$$

On a practical perspective, the computational complexity of Equation (3) is $O(T^2 N \log N)$. Nevertheless, it is possible to efficiently approximate this value via empirical stopping algorithms based on Berstein or Hoeffding bounds [56, 65]. Notice that the pairwise distance has a solid theoretical foundation as to be the base for a measure of the spread. It is an empirical approximation of $\mathbb{E}[d(\Sigma, \Sigma')]$ where Σ and Σ' are independent with law $\mathbb{P}$. Therefore, both are directly related to the noise level in different probabilistic models for permutations [43]. Moreover, the former is the basis of statistical uniformity tests for ranked data building upon it [12].

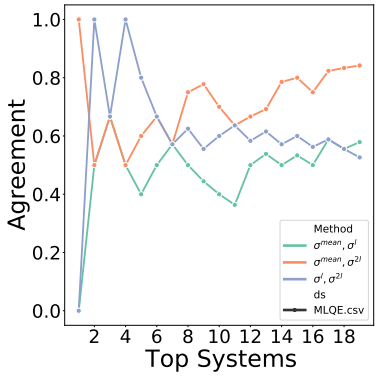
Remark 2. *A known relation between $\mathbb{E}[d(\sigma, \Sigma)]$ and $\mathbb{E}[d(\Sigma, \Sigma')]$ has remarkable consequences on assessing the quality of our estimation. It says that the former is bounded below (respectively, above)*



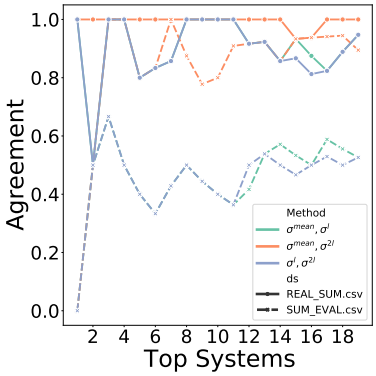
(a) Dialog



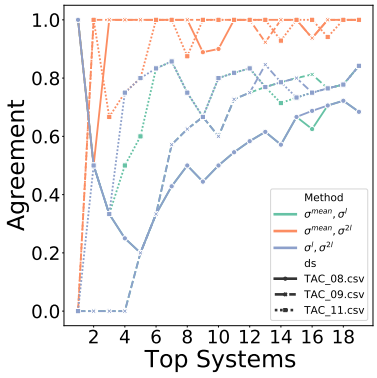
(b) FLICKR



(c) MLQE



(d) Summarizations Datasets



(e) TAC Datasets

Figure 11: Test size experiements

by 0.5 (respectively, 1) times the latter [56]. This fact has broad practical application since it means that the measure of the expected quality of the estimators σ^{2l} and σ^{l} is lower an upper bounded by the intrinsic difficulty of the problem, which we can approximate via sample statistics in Equation (3).

We conclude by noting that $\mathbb{E}[d(\Sigma, \Sigma')]$ is the natural ranking counterpart of the variance for real-valued aggregation σ^{mean}. However, when the scores are not on the same scale, then the variance of the scores is no longer interpretable as a measure of spread in the population.

C.3 Experiments

We report in Tab. 7 the results of the dispersion analysis. We compare the dispersion to measure performance obtained with the induced ranking by σ^* σ^{mean} and the one obtained by 100 random permutations.
Takeaways As expected, σ^* obtains a lowest dispersion which further validate our approach.

	σ^*	σ^{mean}	$Random$
GLUE	793	805	2746
SGLUE	44.9	47.21	137.3
XTREM	12.25	12.75	50.6

Table 7: Results of the dispersion analysis on the considered benchmarks.