What are the best systems? New perspectives on NLP Benchmarking

CentraleSupélec
28/01/2025

Guilherme Mertens
Lucas Tramonte
João Pedro Regazzi
Quentin Lemboulas

Indices

1 Introduction

2 Proposed Approach

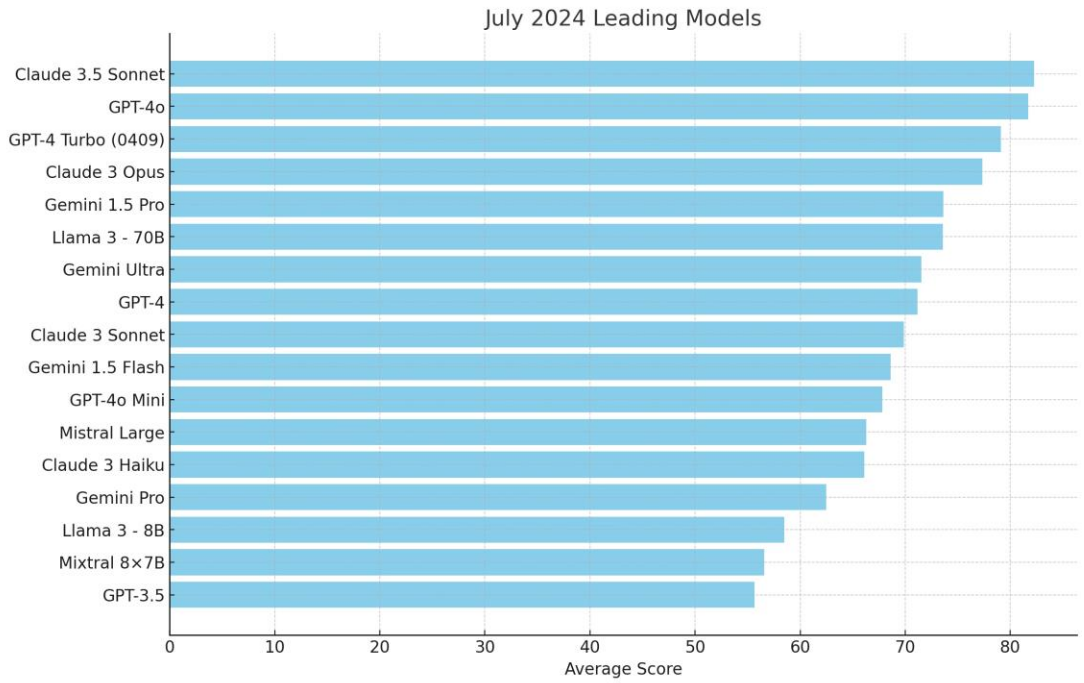
3 Experiments

4 Conclusion

Introduction

Contextualization

- LLMs are evolving with **greater complexity** and generalization.
- Benchmarking ensures effectiveness across **tasks** and datasets.
- Current methods lack standardized **aggregation** for diverse tasks.



Benchmarking

Measure **progress**, assess **weaknesses**, and uncover **opportunities** for improvement

LLAMA 2

Text Summarization

Machine Translation

Sentiment Analysis

Named Entity Recognition

QNLI

STS-B

MRPC

CoLA

G eval

Matthew's Corr

F1

Pearson-Spearman Corr

Accuracy

2 1 3

Models **Tasks** **Datasets** **Metrics** **Ranking**

Limitations of Existing Methods

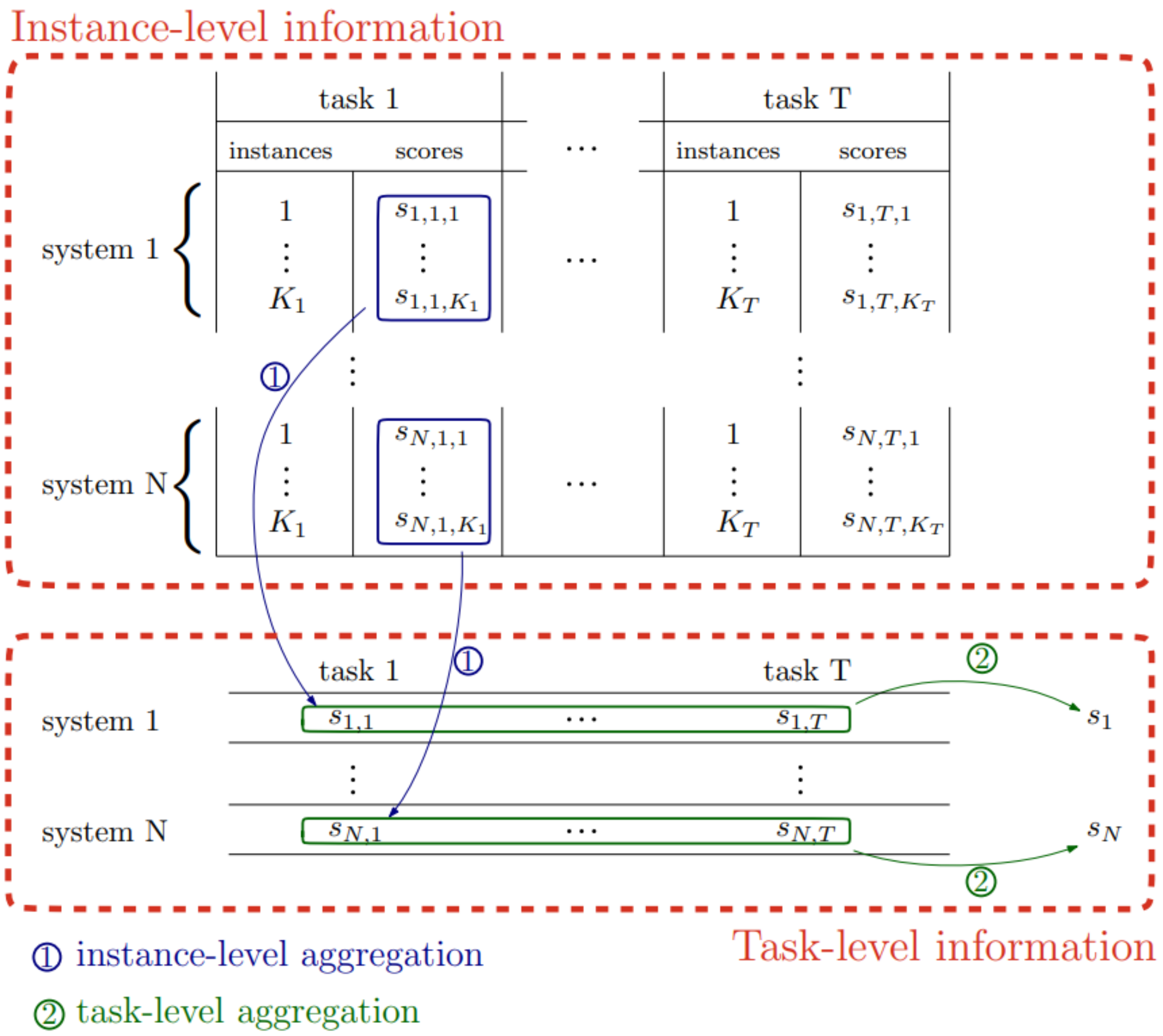
- **Mean Aggregation**:
  - Metrics on different scales and boundedness issues.
  - Tasks vary in importance and difficulty.
- **Pairwise Ranking**:
  - Limited to two systems at a time.
  - Computationally expensive for large-scale settings (O(N^2)complexity).
  - Prone to paradoxical conclusions in rankings.

	Task1	Task2	Task3	Task4	Task5	Task6	SUM
A	0.3 (3)	5 (3)	10 (1)	0.02 (2)	1.0 (1)	0.4 (3)	16.72 (13)
B	0.1 (2)	4 (2)	13 (2)	0.01 (1)	2.2 (3)	0.3 (2)	19.61 (12)
C	0.0 (1)	3 (1)	15 (3)	0.03 (3)	2.0 (2)	0.2 (1)	20.23 (11)

Problem Formulation

Notations:

- N: Number of systems, T: Number of tasks, Kt: Instances in task t.
- S n,t,k: Score of system n on instance k of task t.



Goal:

- Rank N systems based on their performance across T tasks.
- Address limitations in task-level and instance-level aggregation.

Proposed Approach

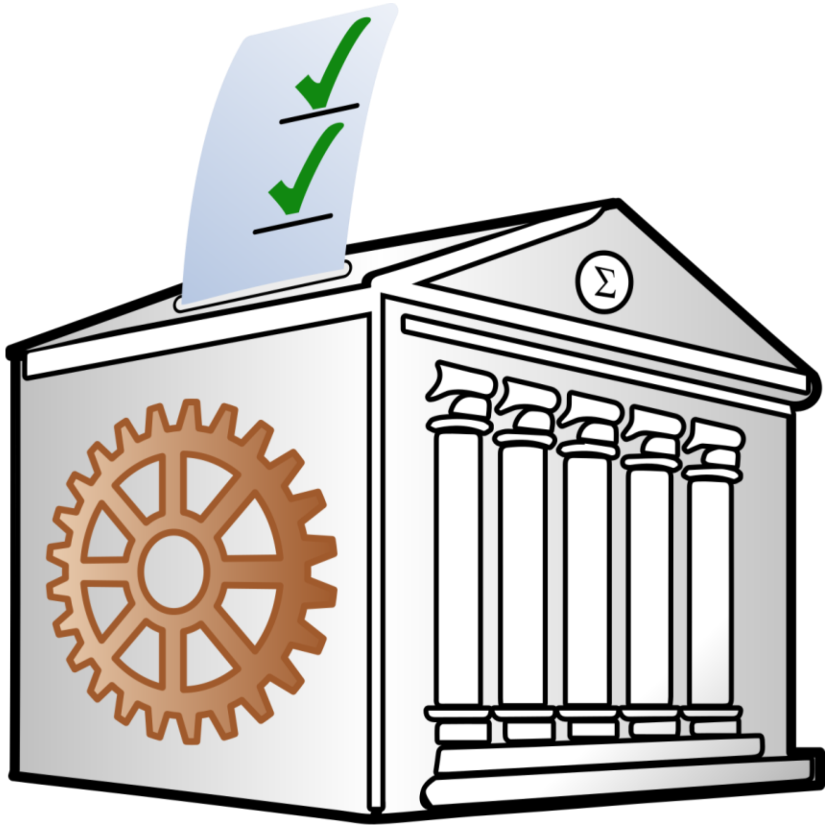
Ranking Aggregation: Accuracy vs. Efficiency

Kemedy Consensus:

- Inspired by Social Choice Theory
- Neutrality, Consistency
- Minimizes Kendall distance (disagreement)
- Optimal but Computationally Expensive

Borda's Count Approximation:

- Sum of Ranks
- 5-Approximation of Kemeny
- Efficient Alternative
- Scalable for Large Benchmarks



$$f : \underbrace{\mathfrak{S}_N \times \cdots \times \mathfrak{S}_N}_{T \text{ times}} \longrightarrow \mathfrak{S}_N$$

Proposed Ranking Procedure

Two Stages Ranking System:

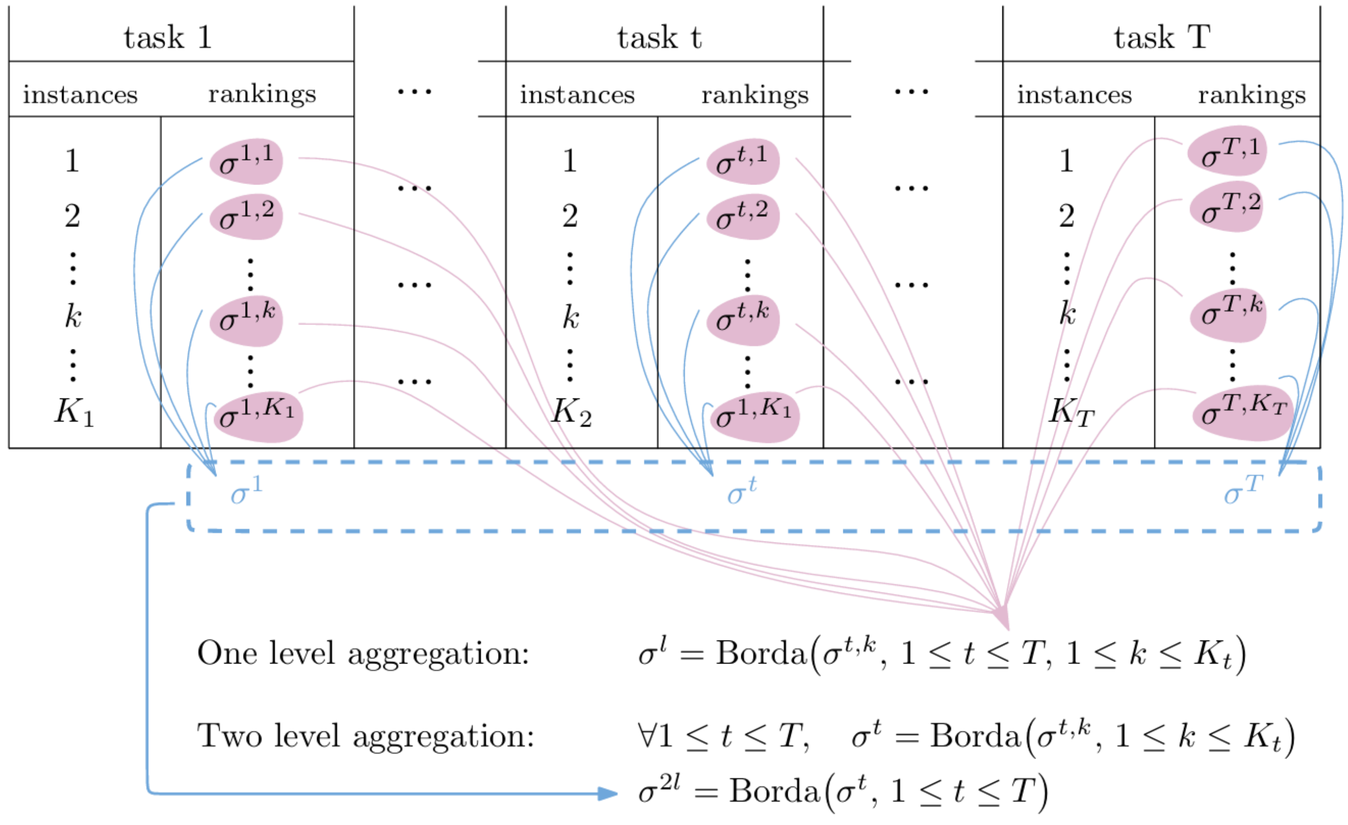


Figure 2: Illustration of our two aggregation procedures to rank systems from instance-level in formation.

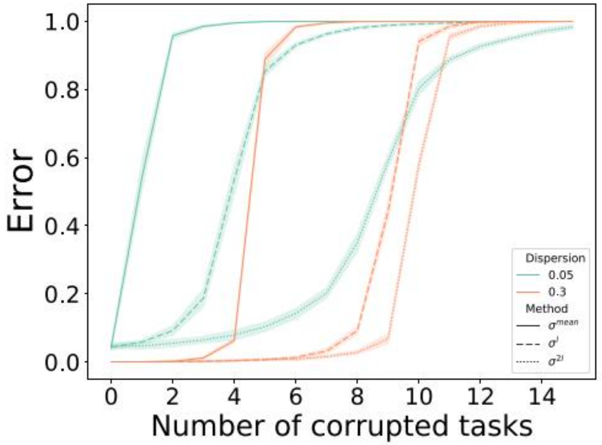
Experiments

Synthetic Experiments

Robustness to manipulation:

Method :

- Corrupt scores by generating them with a distribution centered on the opposite of the initial score
- Determine how many scores need to be perturbed for the classification error to be greater than 50%.



(a) Synthetic Scores

Figure 3: Robustness on synthetic scores.

✓ The ranking-based methods are more robust than σ_mean (σ_2l is the most robust procedure)

Synthetic Experiments

Robustness to scaling:

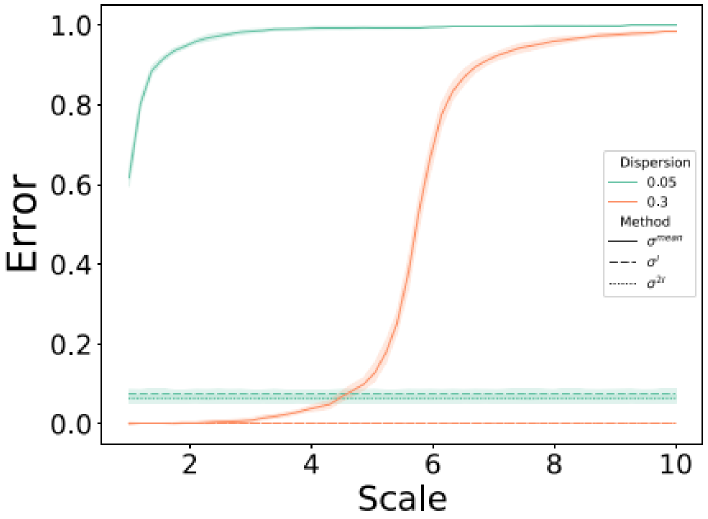


Figure 7: Synthetic Experiment on robustness to scaling. Error is measured in term of Kendall distance.

✓ Re-scaling a task's score by an arbitrarily large number causes an equally large error in mean aggregation but leaves ranking-based aggregation unaffected

Empirical Experiments

Data Collection:

1

Datasets with Task Level Information

- GLUE, SGLUE and XTREME
  - Tasks for XTREME benchmark:
    - Sentence classification and retrieval
    - Structured prediction
    - Question answering

2

Datasets with Instance-level information (NLG evaluation)

- Tasks focused:
  - Summary evaluation
  - Image description
  - Dialogue and Translation

Empirical Experiments

How to compare different rankings quantitatively?

1 Kendall Distance

- Computes the number of inversions between two permutations

$$K\,(\tau^1, \tau^2) = \sum_{(j,s)\, j \neq s} K^*_{j,s}(\tau^1, \tau^2)$$

$$K^*_{j,s}(\tau^1, \tau^2) = \begin{cases} 0 & if\ x_i, x_j\ are\ in\ the\ same\ order\ in\ \tau^1\ and\ \tau^2 \\ 1 & if\ x_i, x_j\ are\ in\ the\ inverse\ order\ in\ \tau^1\ and\ \tau^2 \end{cases}$$

Empirical Experiments

How to compare different rankings quantitatively?

2 Kendall Tau (τ) correlation

- $\tau \in [-1, 1]$
- -1 : strong disagreement
- 1 : strong agreement

$$\tau = \frac{(nombre\ de\ paires\ concordantes\ ) - (nombre\ de\ paires\ disconcordantes\ )}{\frac{n(n-1)}{2}}$$

3 Agreement rate

- Proportion of common top-ranked systems between σ_mean and σ*

Empirical Experiments

Task-level Aggregation Experiments

1 Compute the agreement rate (in %)

2 Compute the Kendall Tau (τ) correlation between the rankings

Dataset	Top 1	Top 3	Top 5	Top 10
XT.	1	0.66	0.8	0.9
GLUE	1	1	0.8	0.8
SGLUE	1	1	0.8	0.9
Dataset	Last 3	Last 5	Last 10	τ
EXT.	1	0.8	0.9	0.82
GLUE	1	0.8	0.7	0.92
SGLUE	1	1	1	0.91

Table 2: Agreement count between Top N/Last N systems on the Ranking when Task Level Information is available. τ is computed on the total ranking.

- ✓ High correlation between the final rankings
- ✓ Methods tend to agree on which are the best/worst systems

Task-level Aggregation Experiments

GLUE			XTREM		
σ^*	Team	σ^{mean}	σ^*	Team	σ^{mean}
0 (1430)	Ms Alex	0 (88.6)	0 (55)	ULR	0 (83.2)
1 (1405)	ERNIE	1 (88.0)	1 (50)	CoFe	1 (82.6)
2 (1397)	DEBERTA	2 (87.9)	2 (44)	InfoLXL	3 (80.6)
3 (1391)	AliceMind	3 (87.8)	3 (42)	VECO	4 (80.3)
4 (1375)	PING-AH	5 (87.6)	4 (35)	Unicoder	5 (79.4)
5 (1362)	HFL	4 (87.7)	5 (34)	PolyGlot	2 (80.6)
6 (1361)	T5	6 (87.5)	6 (31)	ULR-v2	6 (79.4)
7 (1358)	DIRL	10 (86.7)	7 (29)	HiCTL	8 (79.1)
8 (1331)	Zihan	7 (87.6)	8 (29)	Ernie	7 (79.1)
9 (1316)	ELECTRA	11 (86.7)	9 (21)	Anony	10 (78.3)

Table 3: Qualitative analysis between ranking obtained with σ^* or σ^{mean}. Results in parenthesis report the score of the considered aggregation procedure.

- ✓ When changing the aggregation function, the response to the question "what are the best systems?" varies!

Instance-level Aggregation Experiments

1 Comparing the Kendall correlation

2 Comparing the number of agreements between the top N systems

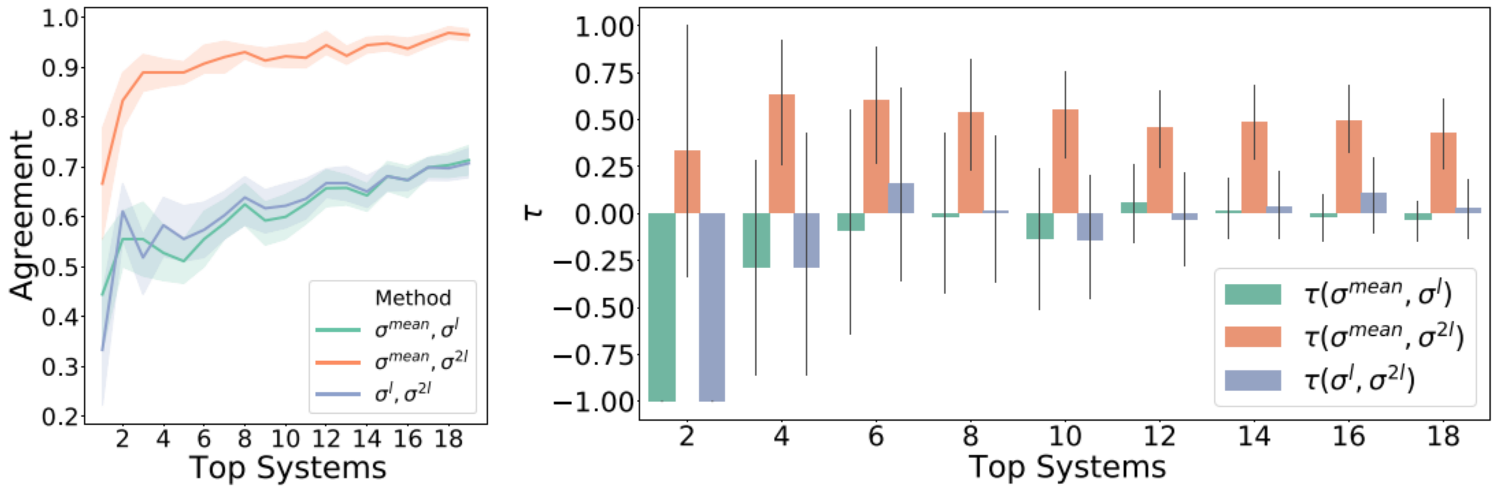
3 Computing the Kendall correlation between them

	PC	TC	FLI.	MLQE
$\tau(\sigma^l, \sigma^{2l})$	-0.08	-0.01	0	-0.03
$\tau(\sigma^{mean}, \sigma^{2l})$	0.32	0.27	0.29	0.01
$\tau(\sigma^{mean}, \sigma^l)$	-0.10	-0.15	-0.04	0.00
RSUM	SEVAL	TAC08	TAC09	TAC11
0.04	0.14	0.28	0.06	-0.06
0.07	0.52	0.32	0.37	0.37
0	0.10	0.23	0.19	0.07

Figure 6: τ on global instance-level rankings.

Empirical Experiments

Instance-level Aggregation Experiments



(a) Agreement Analysis

(b) Correlation Analysis

Figure 4: Global Analysis of Instance Level Raning

✓ σ_2l exhibits a more similar behavior than σ_l with respect to σ_mean

Empirical Experiments

How does the addition/removal of new tasks/metrics affect the ranking?

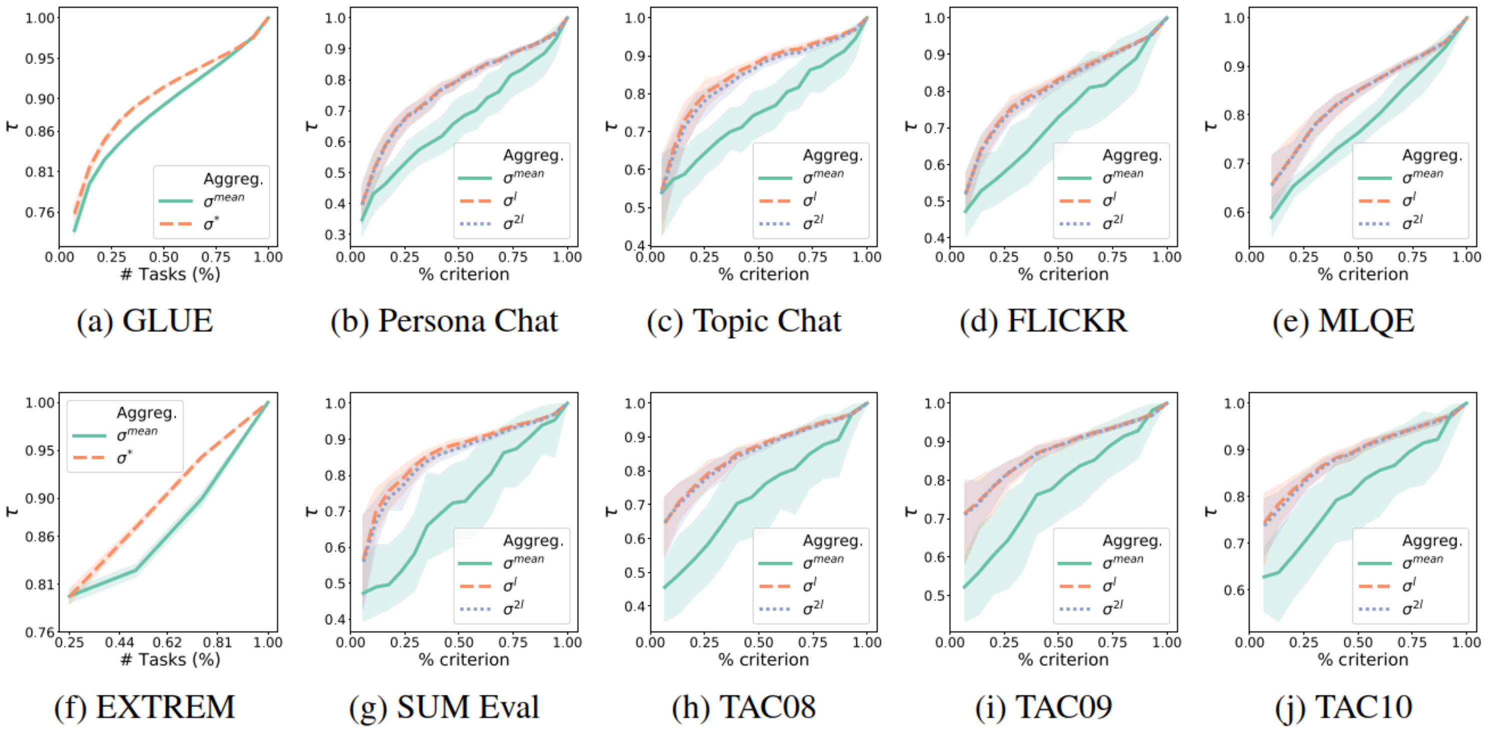


Figure 5: Impact of adding/removing metrics/tasks. The first column refers to ranking obtained with task-level information, while others columns refer to ranking obtained with instance-level information.

- ✓ The ranking from σ∗ is more robust to taks addition/drop than the one from σ_mean
- ✓ The ranking obtained with either σ_l or σ_2l are more robust to task addition/drop than the one from σ_mean

Conclusion

Conclusion and opening for future work

- This paper introduced an **aggregation procedure** based on Kemeny ranking consensus to rank systems
- This method is both **more reliable** and **more robust** than the mean aggregation, formerly used for ranking
- When **task level** (or better **instance level**) ranking is available, we should **use the aggregation procedure** σ* (or better σ_2l) **rather than σ_mean** (σ_mean and σ_l)
- This approach **could be used in other benchmarking** such as Computer Vision or Audio

$$\sigma_{2l} > \sigma_{l} > \sigma_{mean}$$

Ranking prefered methods