# Contextual Metric Meta-Evaluation by Measuring Local Metric Accuracy

Athiya Deviyani and Fernando Diaz
Language Technologies Institute, Carnegie Mellon University
adeviyan@cs.cmu.edu

read our paper!

While most meta-evaluation methods assess metrics through global evaluation over arbitrary outputs, **real-world use cases are highly contextual**, focused on specific models or output qualities. We introduce **local metric accuracy** as a way to evaluate metrics within a context, revealing that metric reliability can shift significantly across settings and motivating the need for context-aware evaluation.

**Metric accuracy** measures how often an evaluation metric accurately assigns the true preference between a pair of system decisions.

**Global metric accuracy** measures this across all outputs, while **local metric accuracy** focuses on specific contexts, e.g., a model, domain, or quality level, revealing how the reliability of a metric varies across settings.

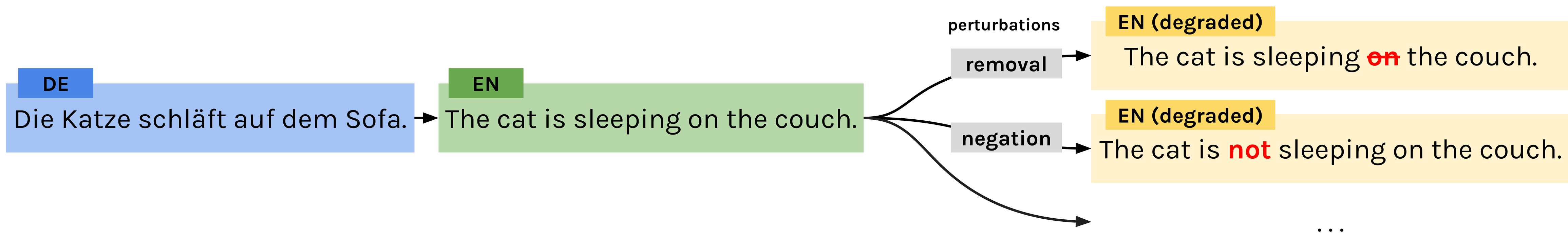
| | context | | | |
|---|---|---|---|---|
| metric | X | Y | Z | global |
| $\mu_A$ | 0.9 | 0.9 | 0.3 | 0.7 |
| $\mu_B$ | 0.7 | 0.7 | 0.7 | 0.7 |
| $\mu_C$ | 0.3 | 0.3 | 0.9 | 0.5 |

**H1:** the **absolute local accuracy** a metric μ change as the context changes (row-wise change)

**H2:** the **relative local accuracy** of a metric μ, i.e. the total ordering of the local metric accuracies, changes as the context changes (cross-column change)

## Measuring local metric accuracies

Given an input $x \in \mathcal{X}$, a system output $y \in \mathcal{Y}_c$ sampled from a specific context $c$, and a degraded version $y'$, we ask: how often does the metric assign a higher score to $y$ than $y'$, across all inputs $\mathcal{X}$ and outputs $\mathcal{Y}_c$?

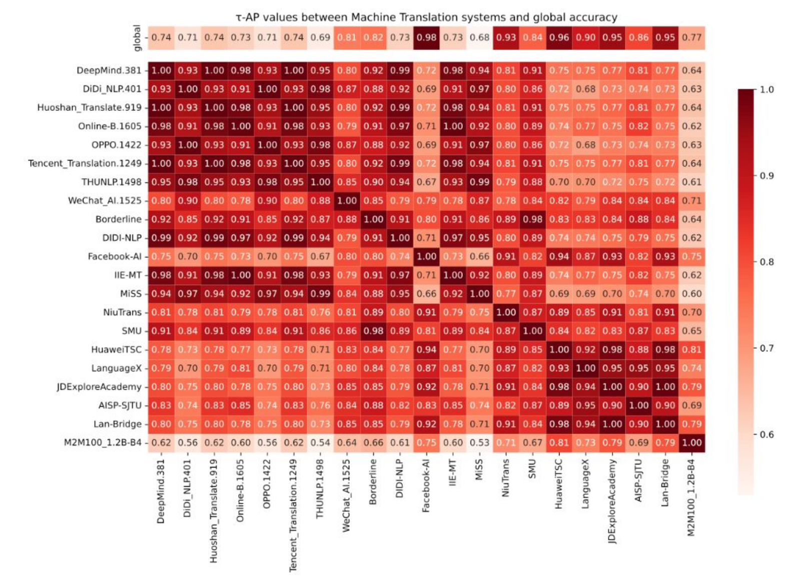



## Metrics evaluated

**MT:** BLEU, ChrF, ROUGE, METEOR, BertScore, COMET, BleuRT, UniTE
**ASR:** WER, MER, WIL, WIP, CER
**Ranking:** MAP, RecipRank, Recall@K, Precision@K, nDCG@K, Interpolated Precision at Recall Level@K

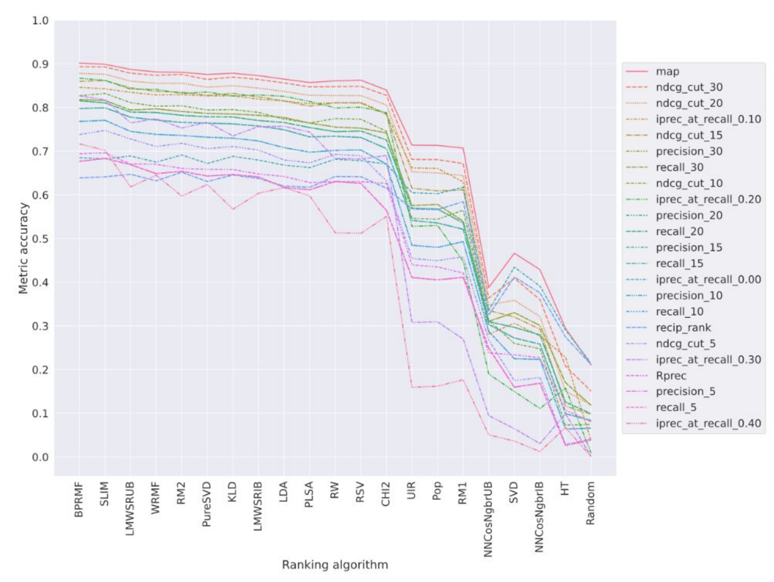
## Results and analysis


(a) Local metric accuracy across the different contexts

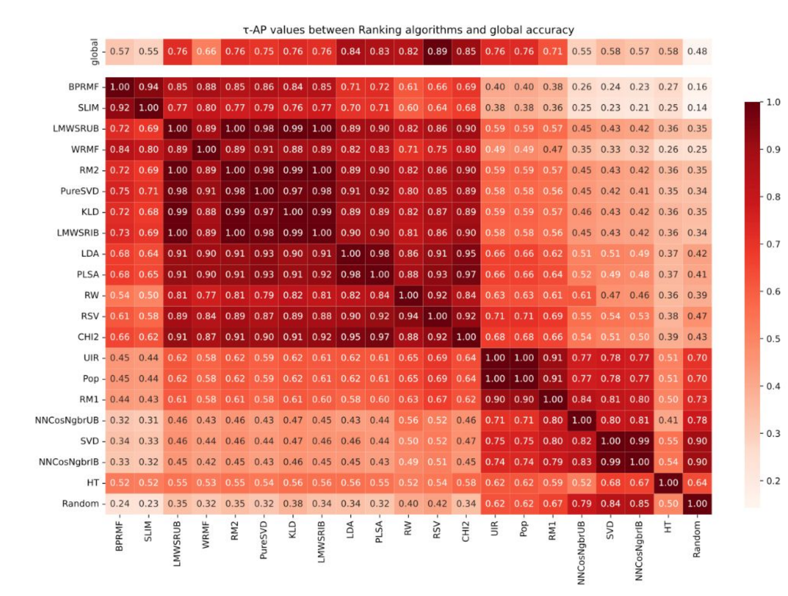



(b) $\tau_{AP}$ values of metrics ordered by local accuracy between different contexts and comparison with global ranking

**Machine Translation**




(a) Local metric accuracy across the different contexts

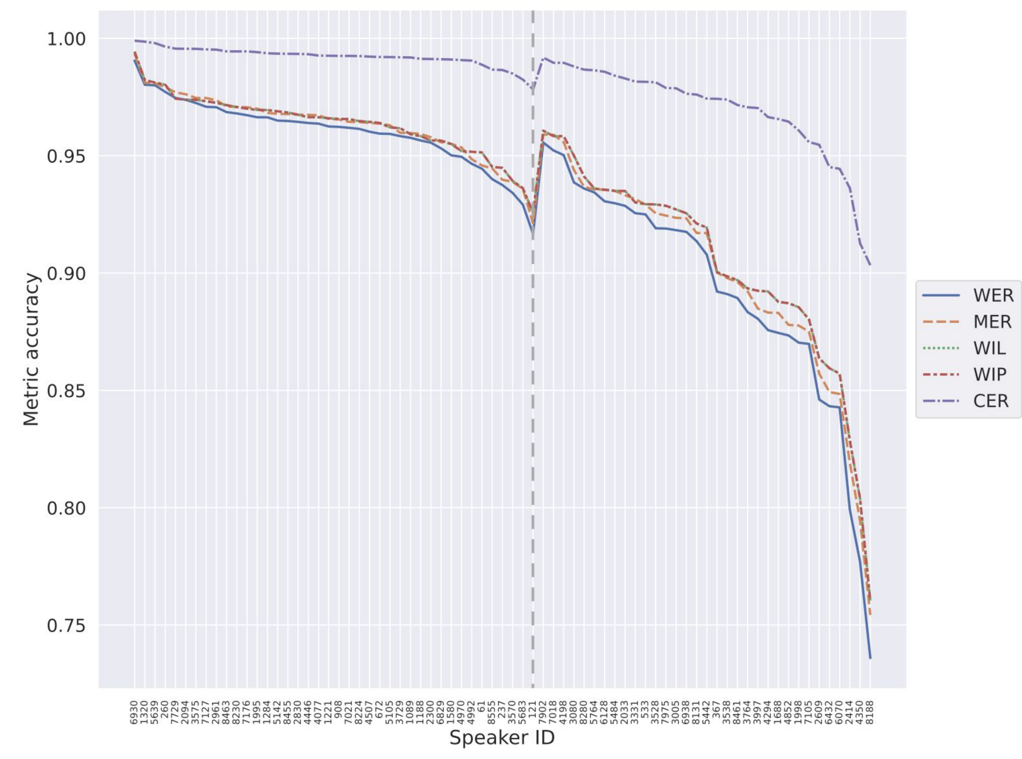



(b) $\tau_{AP}$ values of metrics ordered by local accuracy between different contexts and comparison with global ranking

**Ranking**

✅ **H1 supported:** Local accuracy varies significantly across models and algorithms in both tasks. Each metric's performance depends heavily on the specific system being evaluated.

✅ **H2 supported:** Metric rankings are not stable across contexts. The best-performing metric in one system may underperform in another, highlighting the need for context-aware metric selection.




**Automatic Speech Recognition**

✅ **H1 supported**

❌ **H2 not supported:** Metric rankings are relatively stable across contexts. This is likely due to the low ambiguity of ASR outputs (there's usually a single correct transcription) and the fact that most ASR metrics target similar statistical properties like phonetic or lexical accuracy.

## Practical guidelines

1. **Identify context:** Define the evaluation setting, such as model stage or domain.
2. **Measure local accuracy:** Evaluate how well each metric distinguishes quality differences within that context.
3. **Select metrics based on stability and context:** choose metrics that demonstrate stable accuracies for the specific use case.
4. **Reassess regularly:** Update metric choices as the evaluation needs evolve.