1 Erratum

In our paper titled “SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials,” we reported results from participants on the NLI4CT-P dataset, evaluating their submissions using several metrics, including “Consistency.” This metric was intended to assess a system’s ability to produce identical outcomes for semantically equivalent inputs, focusing on the uniformity in representing semantic concepts across different statements, regardless of the correctness of the final prediction.

However, we have discovered an error in the computation of the Consistency metric. The code we initially used inadvertently calculated the accuracy of the predictions against the gold labels, rather than measuring the consistency between the model’s predictions on semantically equivalent inputs.

The correct implementation should compare the model’s predictions on the original inputs to its predictions on the semantically equivalent (contrast) inputs.

Steps taken

- Re-evaluation of Submissions: We have rerun all participant submissions using the corrected code to obtain the accurate Consistency scores, which are presented in Table 1.
- Updating Results: All tables, figures, and analyses involving the Consistency metric have been updated to reflect the corrected scores.
- Impact on Conclusions: The correction has led to changes in the reported Consistency scores; however, the overall conclusions of the paper remain unchanged.
- Communication with Stakeholders: We have informed all co-authors and participants about this correction.

The computational error affected only the Consistency metric. All other metrics and analyses remain valid. The corrected Consistency scores provide a more accurate assessment of the models’ abilities to maintain uniform predictions across semantically equivalent inputs.

We apologize for any confusion or inconvenience caused by this error. We are committed to ensuring the accuracy and integrity of our work and appreciate the understanding of the community.

Team	**Consistency**	**Incorrect Consistency**	**Difference**	**Average Score**
CRCL	0.74	0.705	0.035	0.77
SEME	0.594	0.56	0.034	0.601
USMBA-NLP	0.691	0.537	0.154	0.583
TLDR	0.74	0.582	0.158	0.633
Puer	0.753	0.638	0.115	0.687
YNU-HPCC	0.839	0.756	0.083	0.807
D-NLP	0.772	0.741	0.031	0.784
Saama Technologies	0.651	0.579	0.072	0.635
0x.Yuan	0.707	0.56	0.147	0.634
Edinburgh Clinical NLP	0.843	0.775	0.068	0.855
LMU-BioNLP	0.729	0.688	0.041	0.778
TuDuo	0.83	0.752	0.078	0.81
NYCU-NLP	0.886	0.809	0.077	0.863
RGAT	0.782	0.735	0.047	0.803
DKE-Research	0.807	0.745	0.062	0.783
IITK	0.743	0.707	0.036	0.777
BD-NLP	0.831	0.758	0.073	0.799
FZI-WIM	0.751	0.729	0.022	0.818
Concordia University	0.958	0.392	0.566	0.55
UniBuc	0.841	0.722	0.119	0.795
CaresAI	0.834	0.755	0.079	0.785
iML	0.846	0.518	0.328	0.609
Lisbon Computational Linguists	0.746	0.715	0.031	0.79
T5-Medical	0.772	0.495	0.277	0.564
DFKI-NLP	0.718	0.684	0.034	0.759

Table 1: Consistency, Incorrect Consistency, Difference, and Average Score for Different Teams