Automated Software Engineering
ASE 2026 Munich
October 12–16, 2026

Software Research Lab

iSE LAB
Interactive Software Engineering and Analytics Lab

MergeSE

v2.0 LIVE

Post-Hoc Model Merging for Software Engineering Tasks Without Retraining

ASE 2026 Tool Demo · Wed 14 Oct, 15:35 · Forum 8, Munich


Palash Ranjan Roy


Banani Roy


Kevin A. Schneider


Chanchal K. Roy

Department of Computer Science, University of Saskatchewan, Canada

University of Saskatchewan

BE WHAT THE WORLD NEEDS

Merging works, but doing it was still hard

71%

F1 drop when a same-language clone detector meets cross-language clones

0.151

F1 of multi-task training on unseen AI-generated clones, worse than any specialist

93%

of multi-task performance reached by merging, with no training data

But every step was a one-off research script.

Is my fine-tuned model good on other SE tasks? Are two models compatible? Did a merge help? Can I deploy it? No shared workflow answered these.

Introducing MergeSE 2.0

More than a merging tool. A shared workflow for SE models, live at mergese.usask.ca on an NVIDIA L40 GPU server.
Merging is optional.



BRING YOUR MODEL

Any fine-tuned HuggingFace encoder, from the Hub or uploaded



EVALUATE ON SE TASKS

Clone, vulnerability, defect, code smell and five more tasks, or your own CSV



MERGE AND EXPORT

Parameter merging or greedy layer stitching on the GPU, then export to deploy

Five operations, each usable on its own



`tasks`

Which SE tasks are supported?

9 task types, input formats, default metrics



`inspect`

Are my checkpoints safe to merge?

Base, tokenizer and architecture checks, plus a verdict



`merge`

How do I combine them?

Parameter merging (TIES, DARE-TIES, WUDI, PCB, averaging) or greedy layer stitching



`evaluate`

How good is my model?

F1, precision, recall on SE benchmarks, merged or not



`export`

Can I deploy it?

HuggingFace, ONNX, TorchScript

The whole workflow in three commands

```
$ mergese inspect --base microsoft/codebert-base \
      ./codebert_bcb ./codebert_clcdsa
  Verdict: COMPATIBLE
  Cosine sim: 0.077 | Sign agreement: 62.1%

$ mergese merge --base microsoft/codebert-base \
      -m ties ./codebert_bcb ./codebert_clcdsa -o ./merged
  Merged in 4.9s | Trimmed: 22.3% | Conflicts: 15.5%

$ mergese evaluate ./merged -t clone_detection -d ./bcb_test.csv
  F1: 0.801 | Precision: 0.747 | Recall: 0.864
```

1. **Inspect**
   Same parent, aligned task vectors, safe to merge

2. **Merge**
   Two 124M-parameter models in under 5 seconds on CPU

3. **Evaluate**
   BigCloneBench F1 of the merged CodeBERT

Live demo

mergeSE.usask.ca

Scan to follow along

Upload · Evaluate · Inspect · Merge · Export

CodeBERT clone specialists for BigCloneBench and CLCDSA

MergeSE reproduces the paper's results

Model	BCB F1	CLCDSA F1	Combined F1	GPTCloneBench F1
CodeBERT-BCB specialist	0.894	0.464	0.679	0.450
CodeBERT-CLCDSA specialist	0.000	0.924	0.462	0.000
UniXcoder multi-task (all data)	0.933	0.922	0.927	0.151
MergeSE TIES (CodeBERT)	**.801 ± .018**	**.746 ± .012**	**.774 ± .014**	**.609 ± .049**
MergeSE TIES (UniXcoder)	**.890 ± .010**	**.839 ± .026**	**.865 ± .012**	**.450 ± .042**

Both merges recover the domain each specialist fails on. The CodeBERT merge gives the best score on unseen AI-generated clones (0.609).

Correctness check against the paper's reference merge: 149 of 199 parameter keys are identical, and the rest differ by a median of 0.0078.

Already used around the world

76

users of the web tool

31

countries

163+

operations run so far

Inspect, merge, evaluate and export jobs on mergese.usask.ca, from our server records

Use it · mergese.usask.ca

Contribute · GitHub

Join us & Contribute

MergeSE is open source and built to grow with the community

Report issues, request features, add a new SE task, or pick a good first issue

github.com/srlabUsask/MergeSE · Apache 2.0 · doi.org/10.5281/zenodo.21797313

Palash Ranjan Roy · palash.roy@usask.ca · palashroy.me

Software Research Lab

iSE LAB
Interactive Software Engineering and Analytics Lab

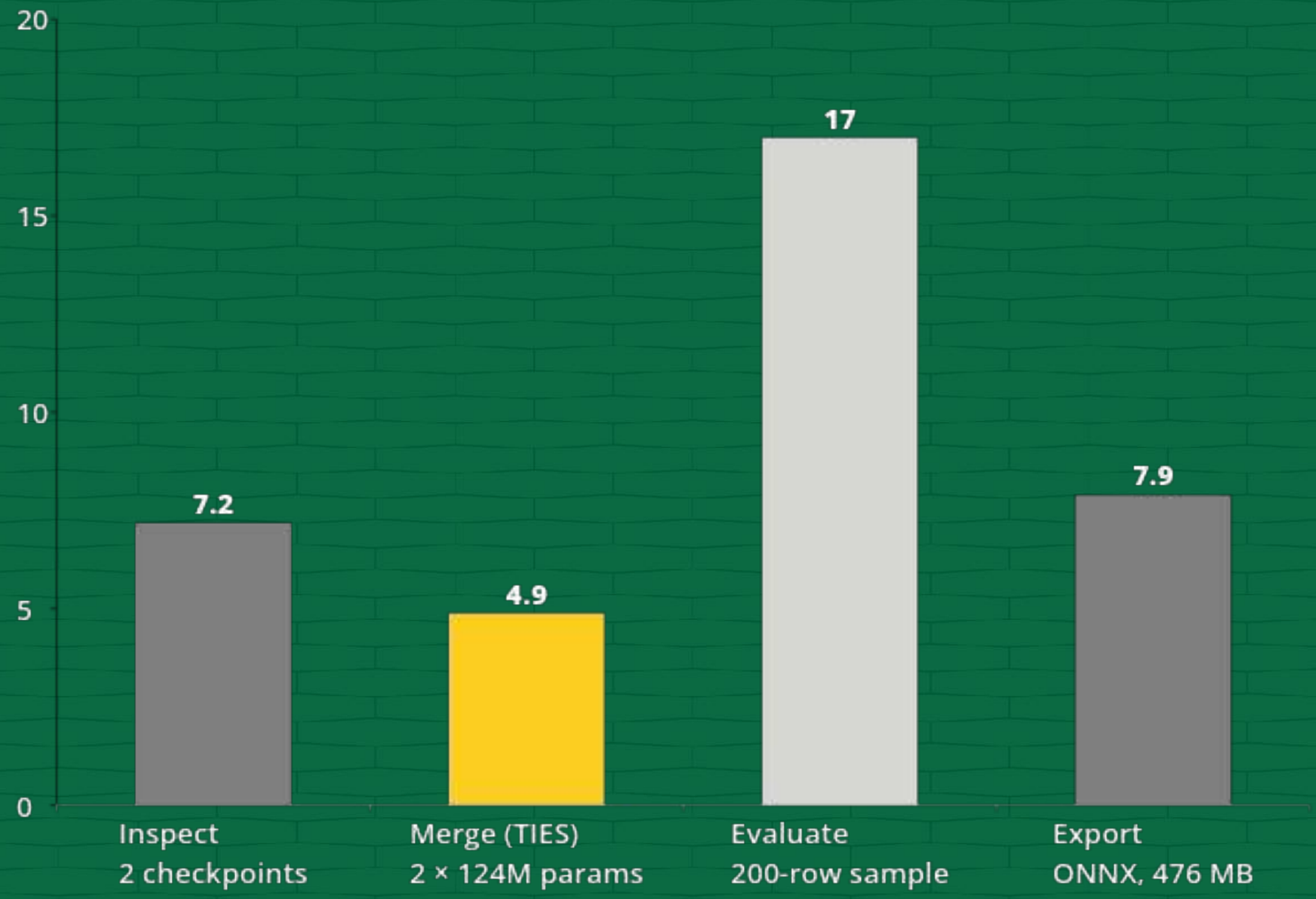
Automated Software Engineering
ASE 2026 Munich
October 12–16, 2026

BACKUP

The whole pipeline runs in under 40 seconds on CPU

No GPU and no training data for inspect, merge and export.

Measured on CPU. The hosted service runs on an NVIDIA L40 GPU server, so evaluation there is faster.



BACKUP

Inspect catches incompatible pairs before you merge

Checkpoint pair	Verdict	Cosine similarity	Sign agreement
CodeBERT-BCB and CodeBERT-CLCDSA	**COMPATIBLE**	+0.077	62.1%
CodeBERT-BCB and UniXcoder-BCB	**INCOMPATIBLE**	+0.102	53.2%

The cross-base pair is flagged because the tokenizers differ, even though its geometry numbers look harmless.

These geometry values use one shared base for both models. The research paper computes each model against its own base, which gives cosine 0.001 and 50.4% sign agreement.

Verdicts are COMPATIBLE, RISKY or INCOMPATIBLE.

BACKUP

What it took to make merging reliable

Mismatched classifier heads

A binary clone head and a multi-class head cannot be averaged. MergeSE detects this and merges the shared encoder only.

Very large tensors

torch.quantile fails on RoBERTa embedding matrices during TIES trimming. MergeSE uses torch.kthvalue instead.

Heads stored separately

Some research code saves the classifier in its own file. MergeSE finds it and attaches it at load time.

Tokenizer version drift

A fallback chain over config files and model type metadata finds a compatible tokenizer.