WMT 2021 Efficiency Shared Task

Kenneth Heafield, Qianqian Zhu, Roman Grundkiewicz
Corruptly organizing while participating.

https://neural.mt/speed/2021

Intel Microsoft ORACLE for Research

Cost to translate a million characters

$20.	Google
$15.	Amazon
$10.	Microsoft
$0.001	Efficiency Task Submissions

This is marginal cost, not fixed cost.

Efficient Inference

Translate English→German for the constrained 2021 news task.

Measure quality, speed, RAM, and disk.
On GPU or CPU.

Batched throughput condition
+ New latency condition throws gauntlet at non-autoregressive papers with weak baselines.

Submissions

	Edinburgh	HuaweiTSC	NiuTrans	TenTrans
GPU Batch	10		4	4
GPU Latency	11			
1 Core CPU Batch	6			
1 Core CPU Latency	6	4		
36 Core CPU Batch	6		2	
Total	39	4	6	4

Focused Human Evaluation

	Edinburgh	HuaweiTSC	NiuTrans	TenTrans
GPU Batch	3/10		4/4	4/4
GPU Latency	0/11			
1 Core CPU Batch	0/6			
1 Core CPU Latency	3/6	4/4		
36 Core CPU Batch	0/6		0/2	
Total	6/39	4/4	4/6	4/4

Focused human evaluation based on:

- Close competition.
- Participant nomination.

Source-based DA, contrastive DA

Test Set

1 million unique shuffled sentences
20 average words/sentence

Quality measured on the WMT 2021 test set.
Threw in IWSLT 2019, TED 2020 v1, SimpleGen, WinoMT, RAPID, ...

Server-focused Hardware

NVidia A100 GPU or Intel Ice Lake CPU.

Oracle Cloud BM.GPU4.8 or BM.Optimized3.36 instances.
Bare metal (no VM), full machine to avoid noise.

Pareto Comparison

Submissions have varying quality and efficiency.
Tolerance for quality loss depends on application.

Pareto comparison: quality $\geq$ baseline **and** efficiency $\geq$ baseline.

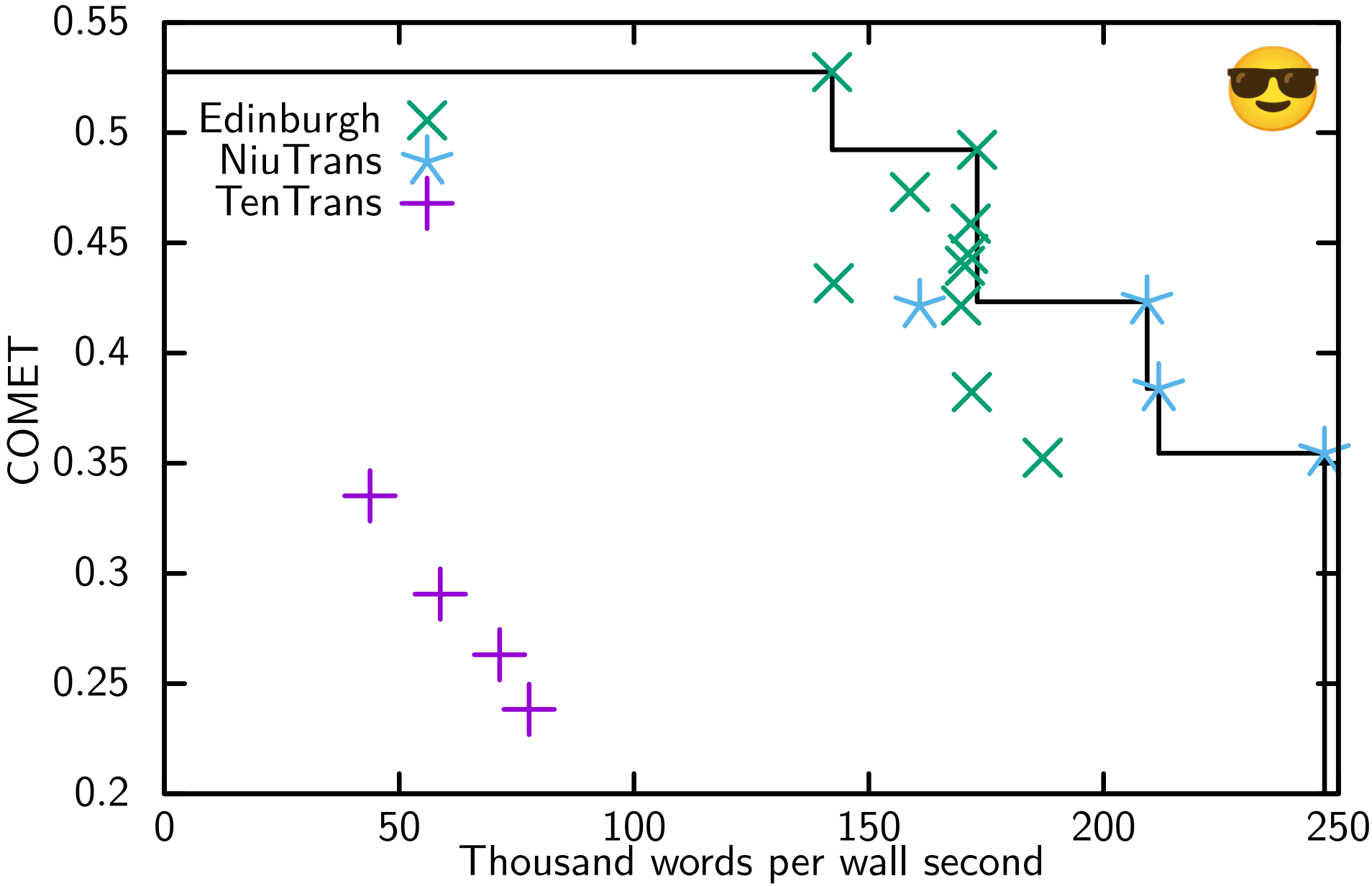
More efficient with same quality
...or better quality with same efficiency.

Speed

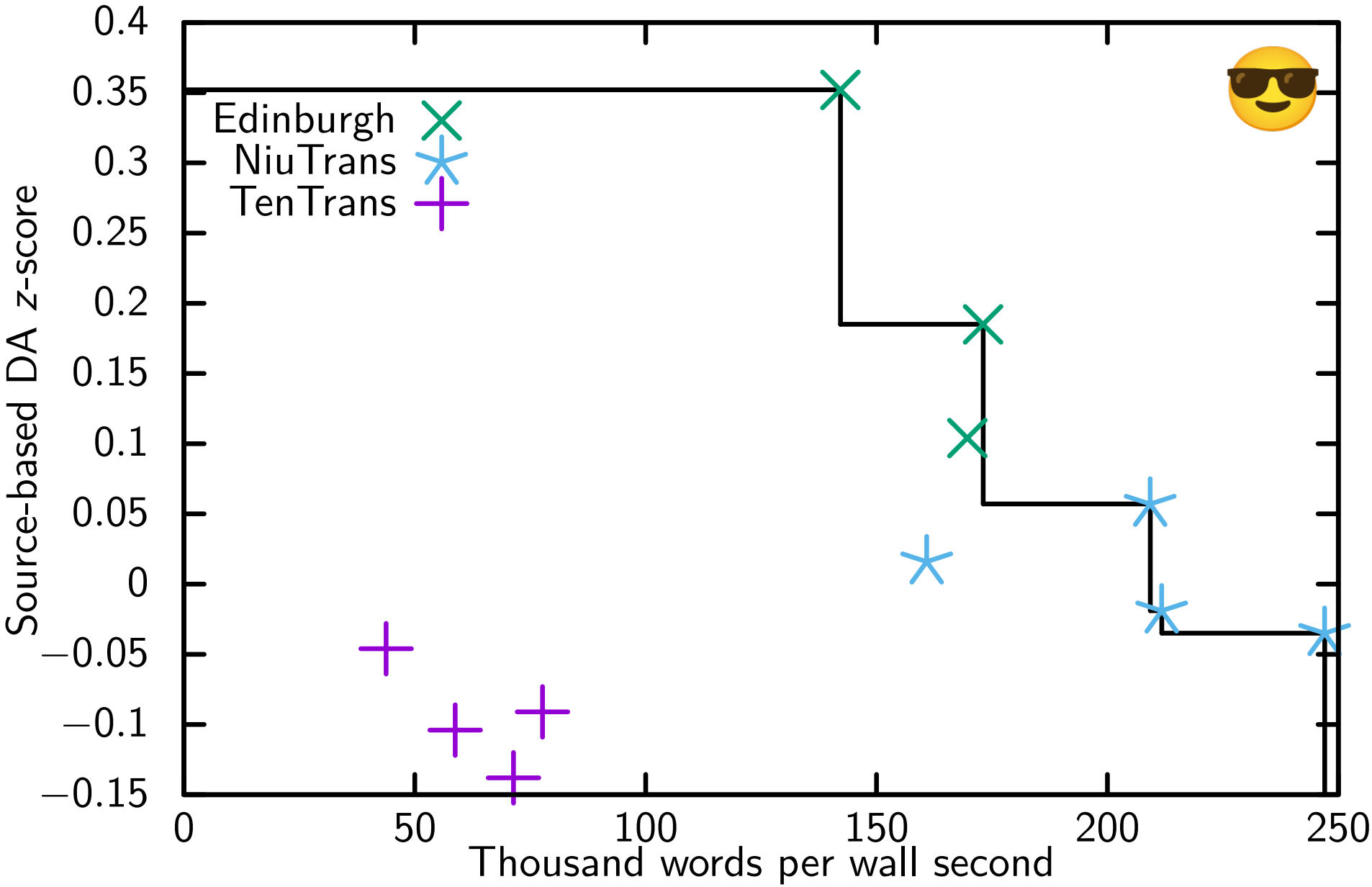
Primary: wall clock time.
Words per second based on 19,951,184 untokenized words.

Supplementary data: CPU time.

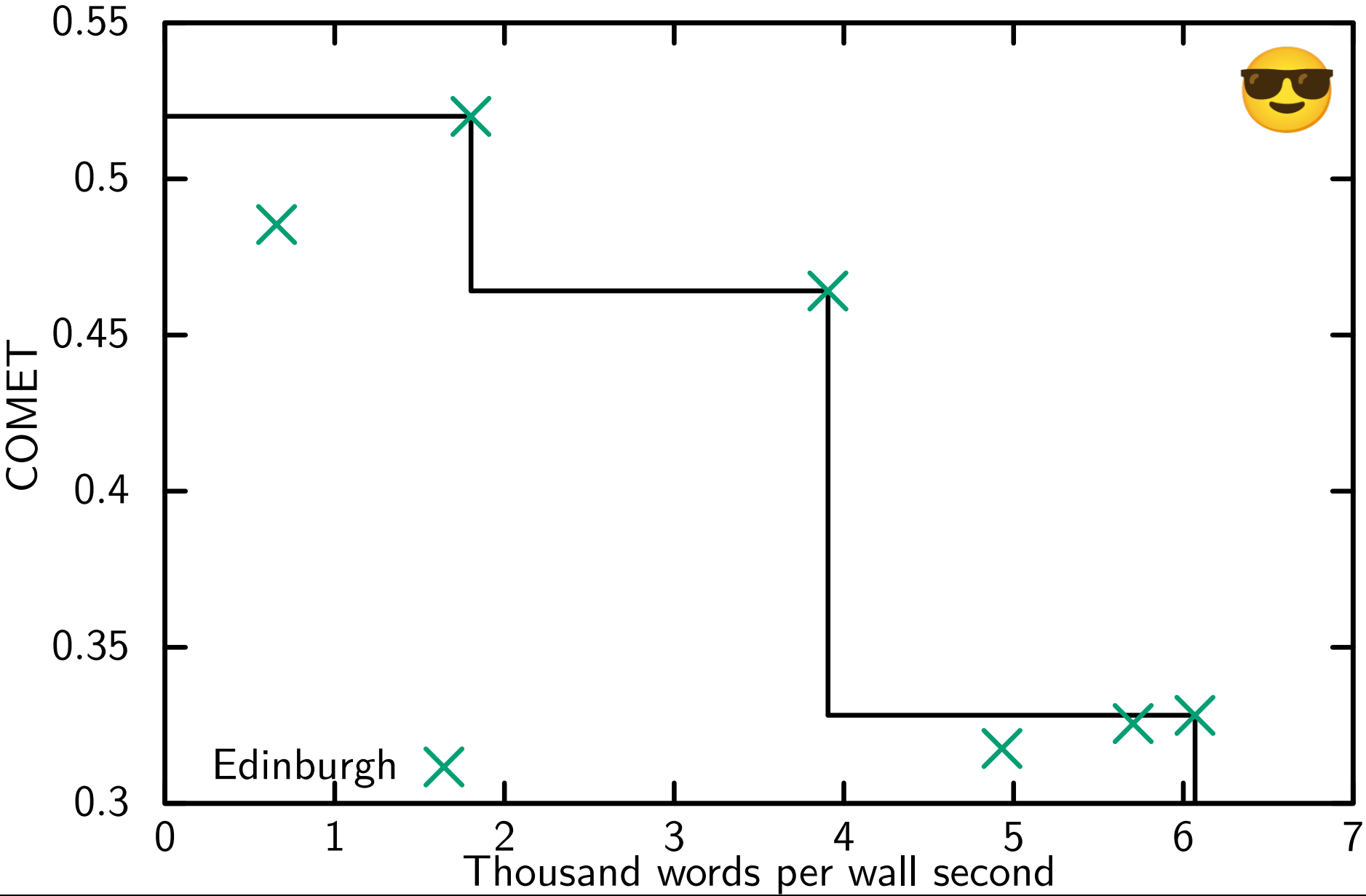
GPU Batch: Automatic



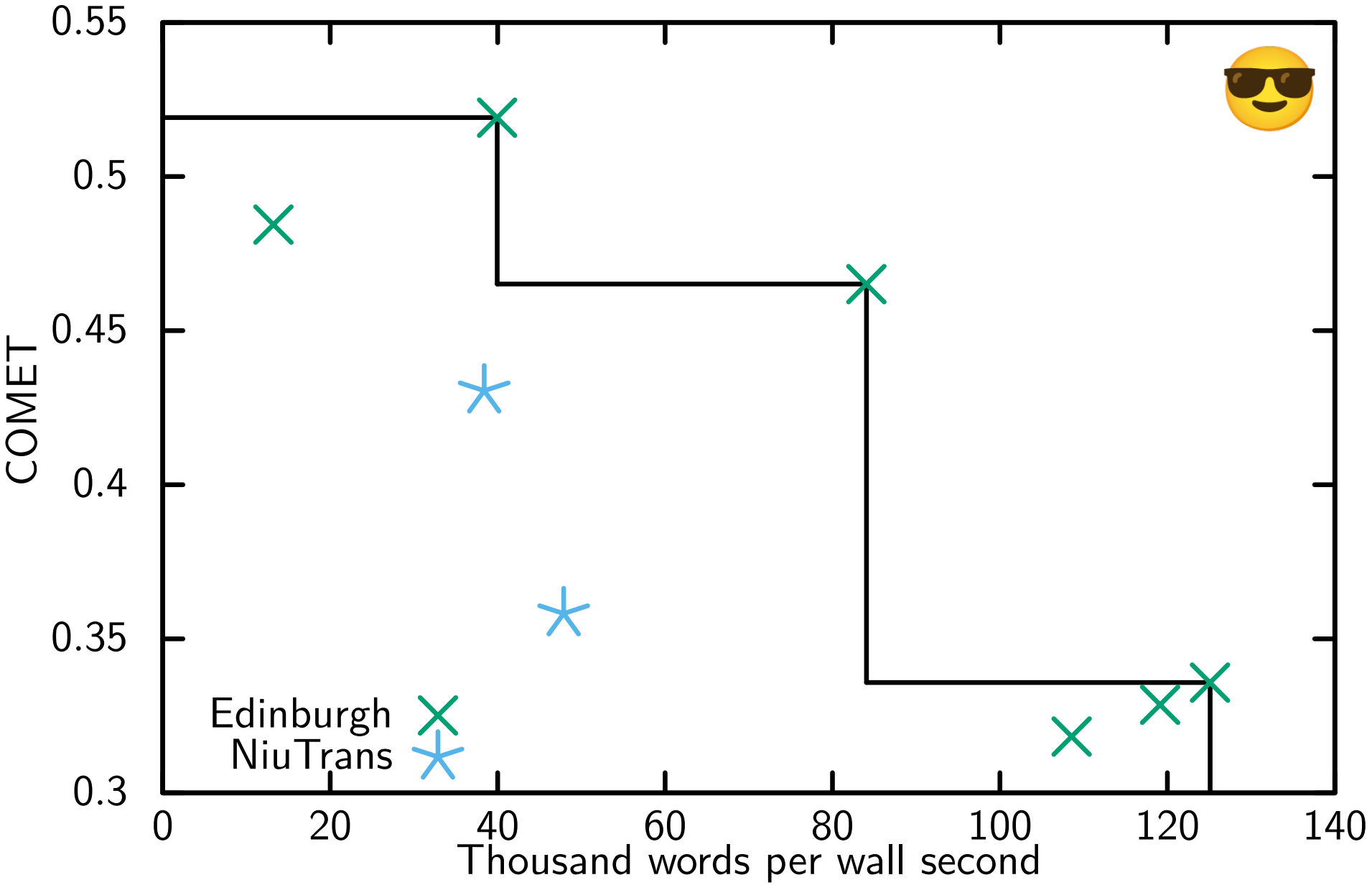
GPU Batch: Human



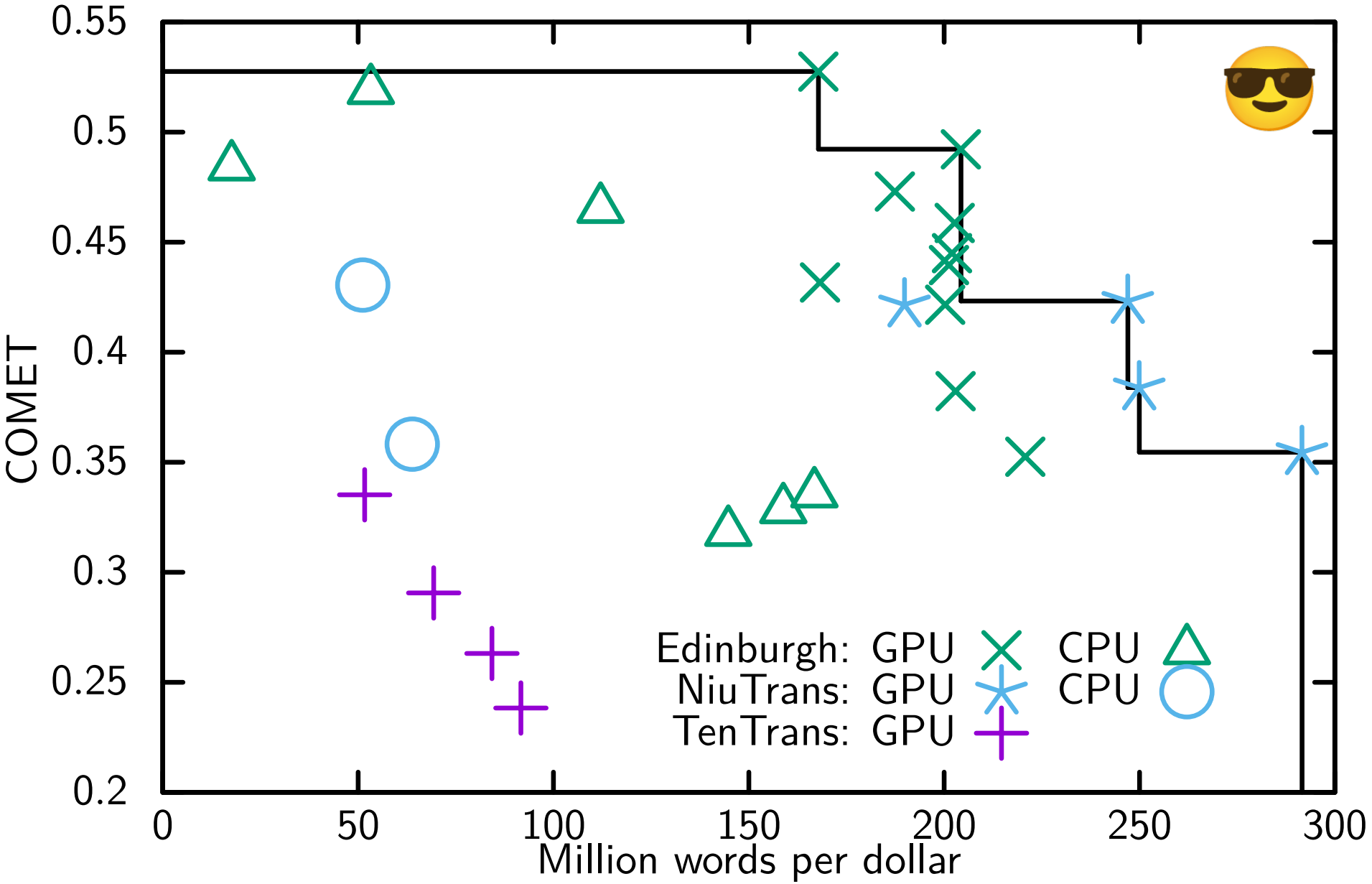
CPU single core Batch



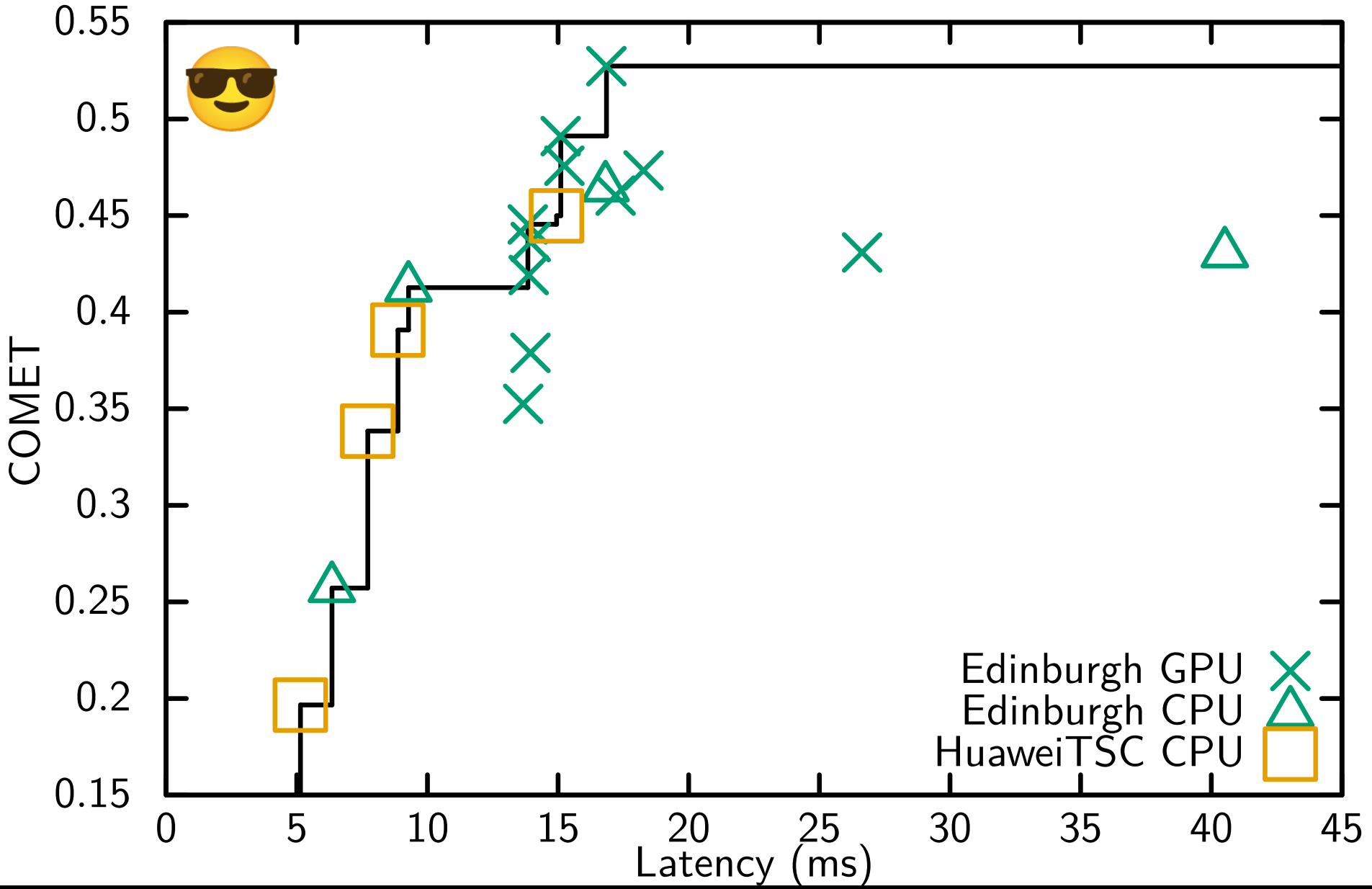
CPU 36 core Batch



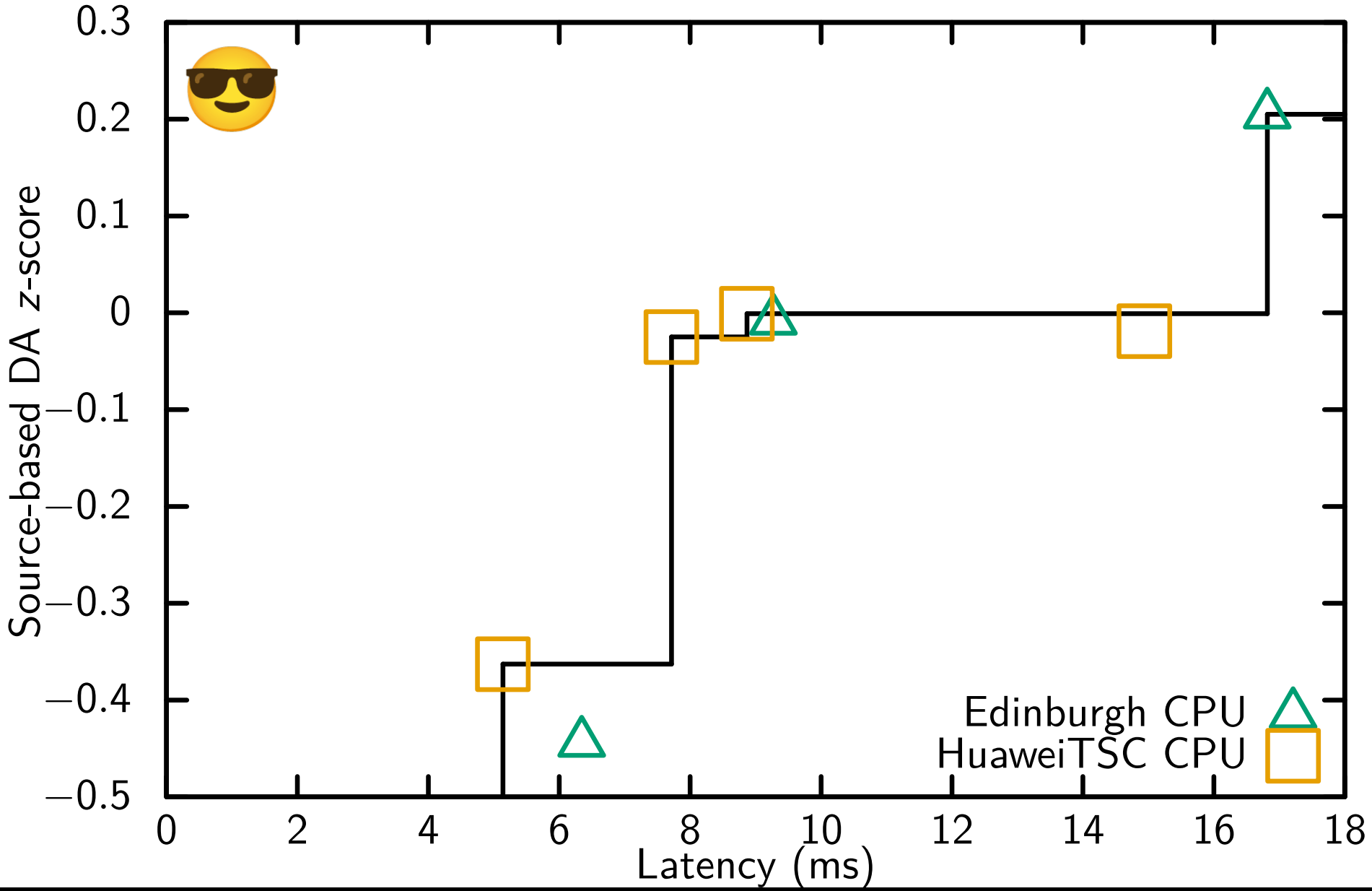
Cost


0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
COMET
0
50
100
150
200
250
300
Million words per dollar
Edinburgh: GPU
CPU
NiuTrans: GPU
CPU
TenTrans: GPU

Latency: Automatic



Latency: Human

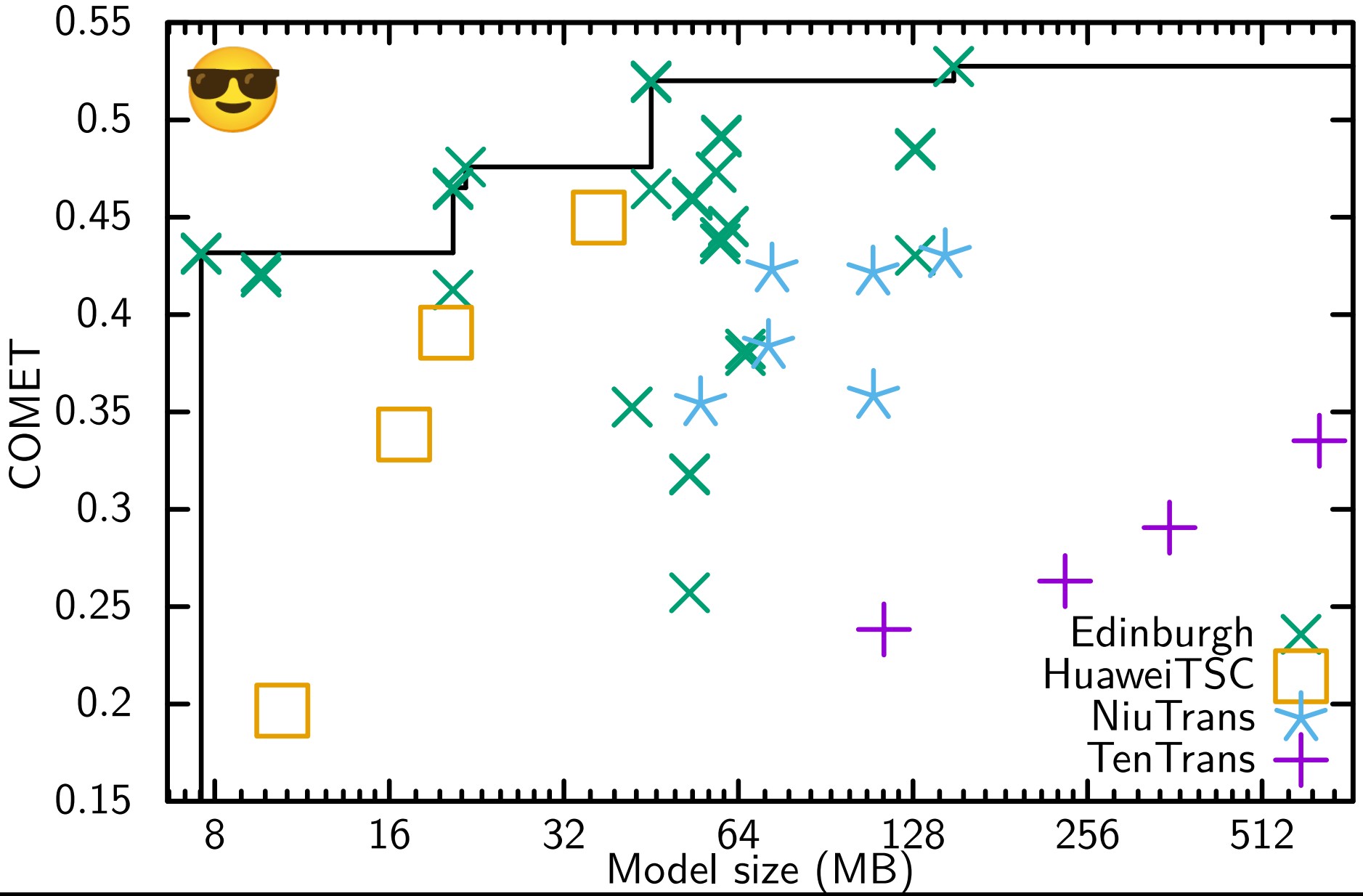


Disk

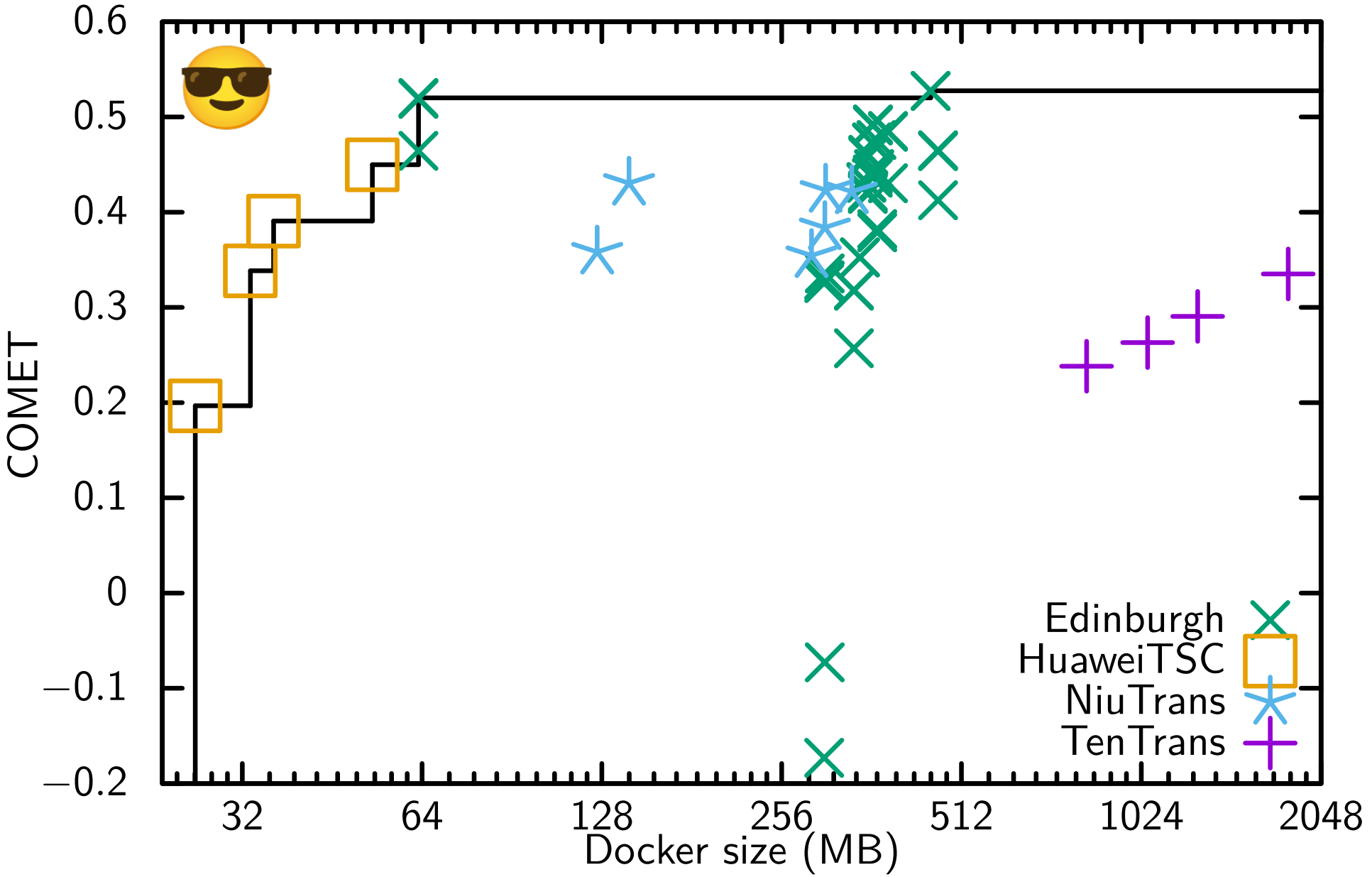
Model size: parameters, BPE, shortlists, etc.

Total Docker size: model, part of Ubuntu, code

Model size, all platforms



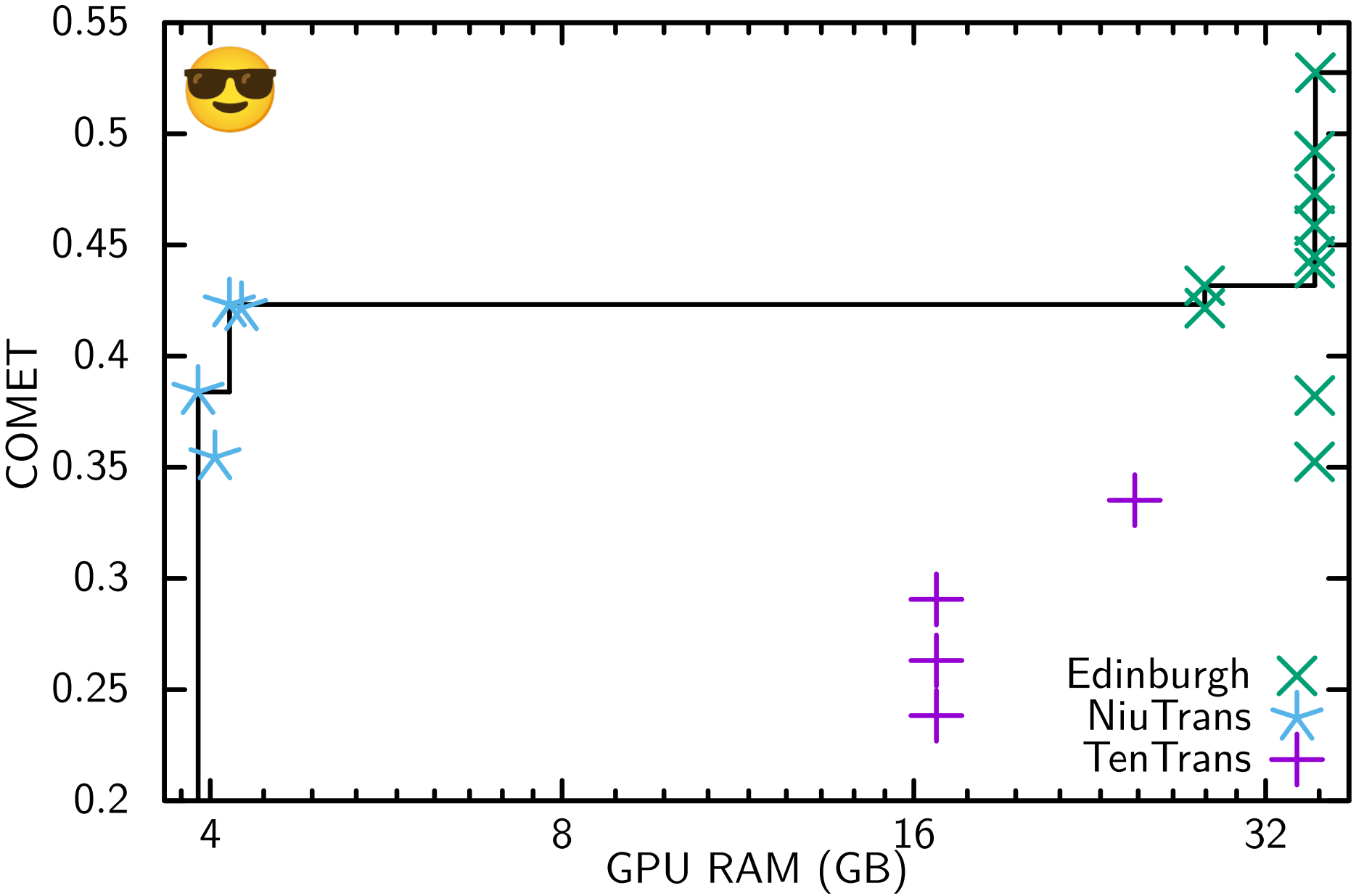
Docker size


0.6
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
COMET
32
64
128
256
512
1024
2048
Docker size (MB)
Edinburgh
HuaweiTSC
NiuTrans
TenTrans

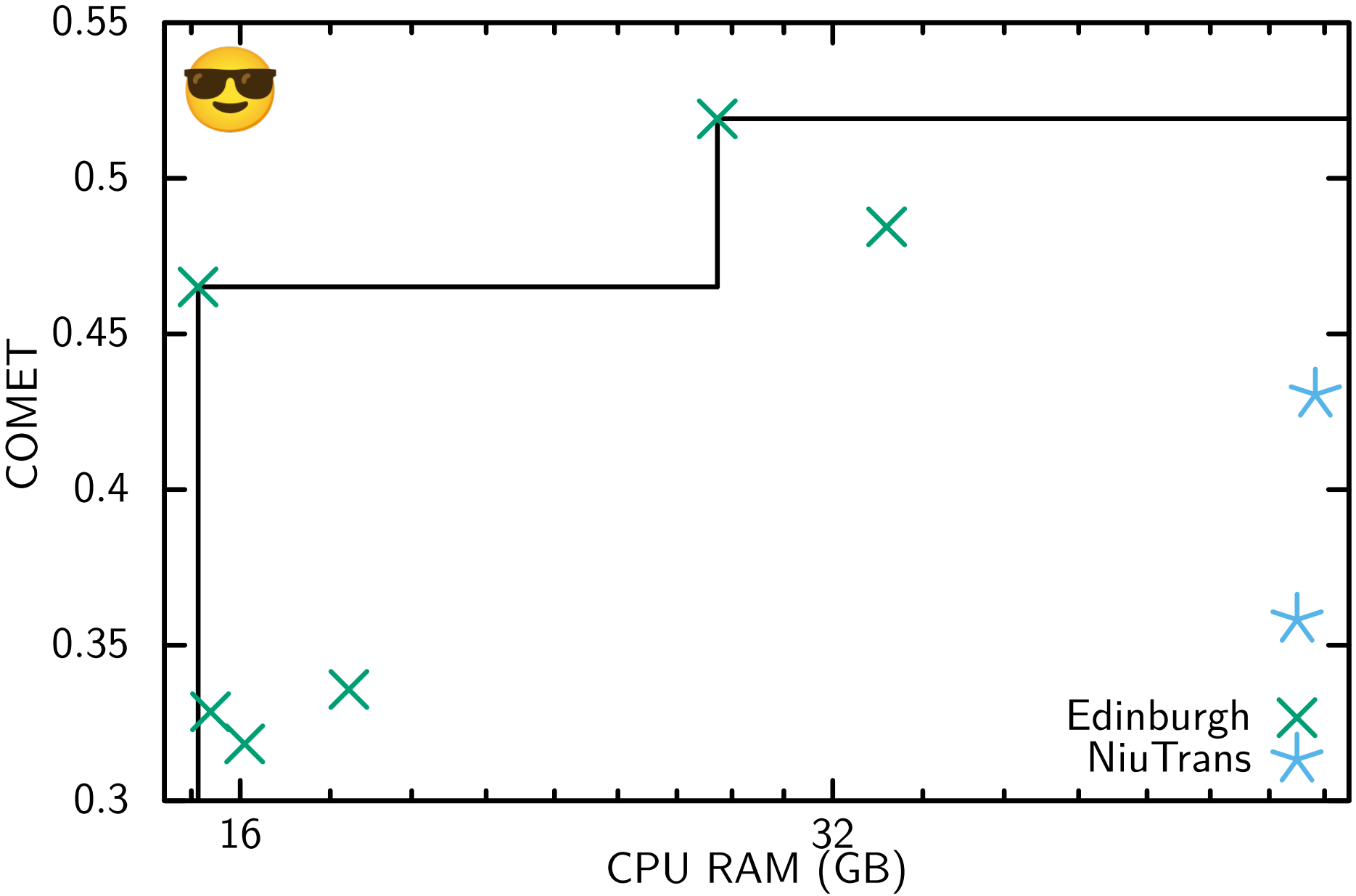
RAM

Peak GPU or CPU RAM usage.
Big batches go faster but use more RAM.

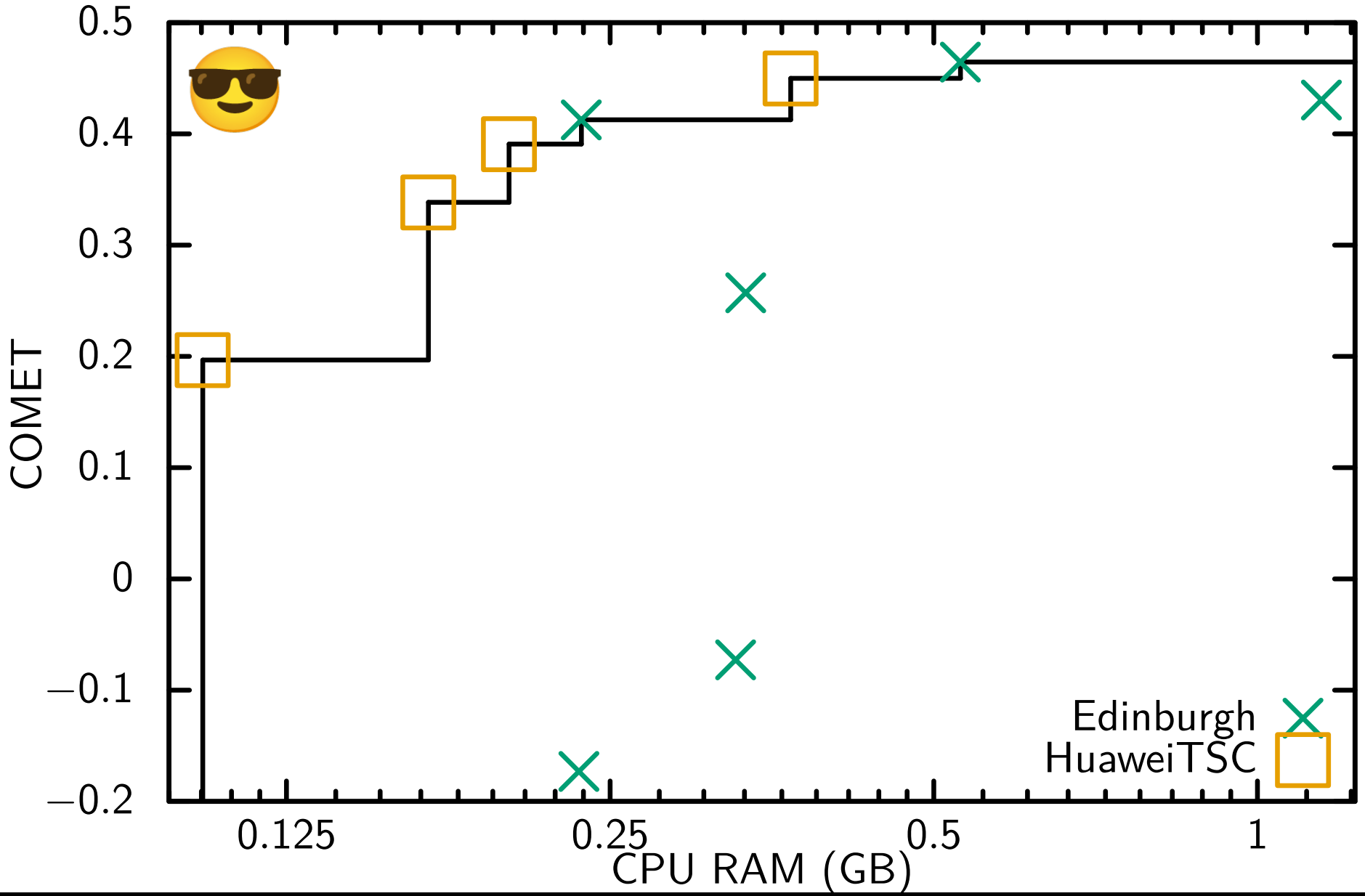
GPU Batch RAM



36 CPU Batch RAM



CPU Latency RAM



See posters for methods used!
Contrastive direct assessment in the paper.

Next Year?

Simplify task, remove unpopular tracks.

Provide open Edinburgh systems in advance?

More than a month after news deadline?

More participants?

Efficient training?