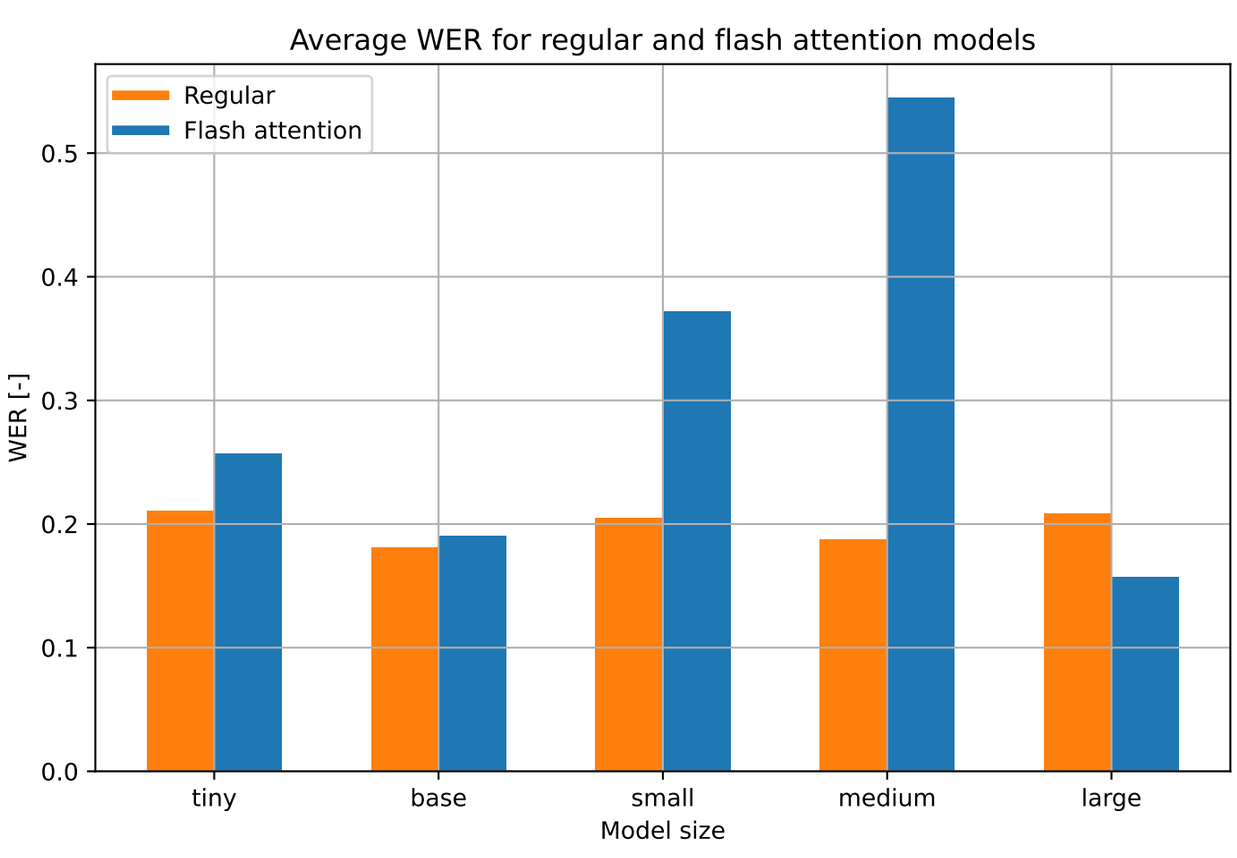
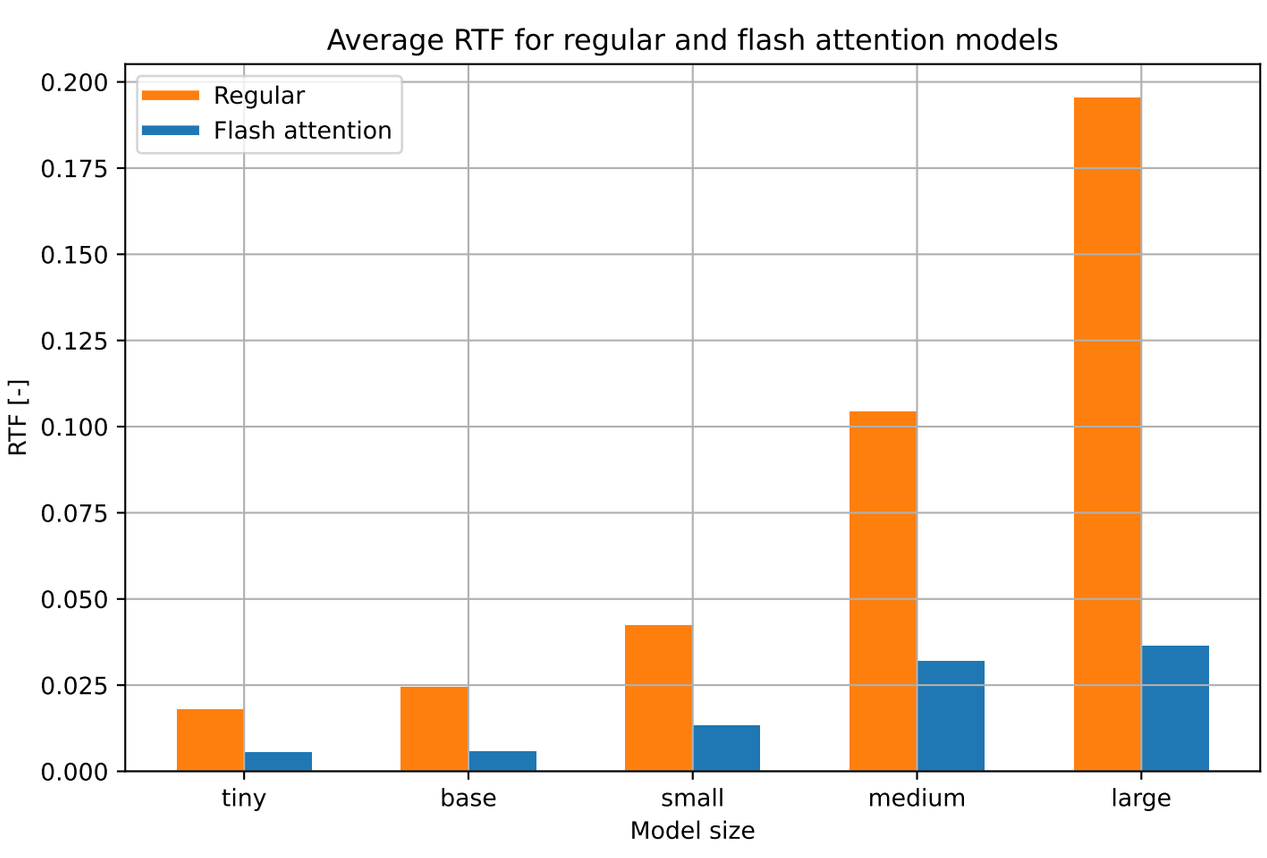
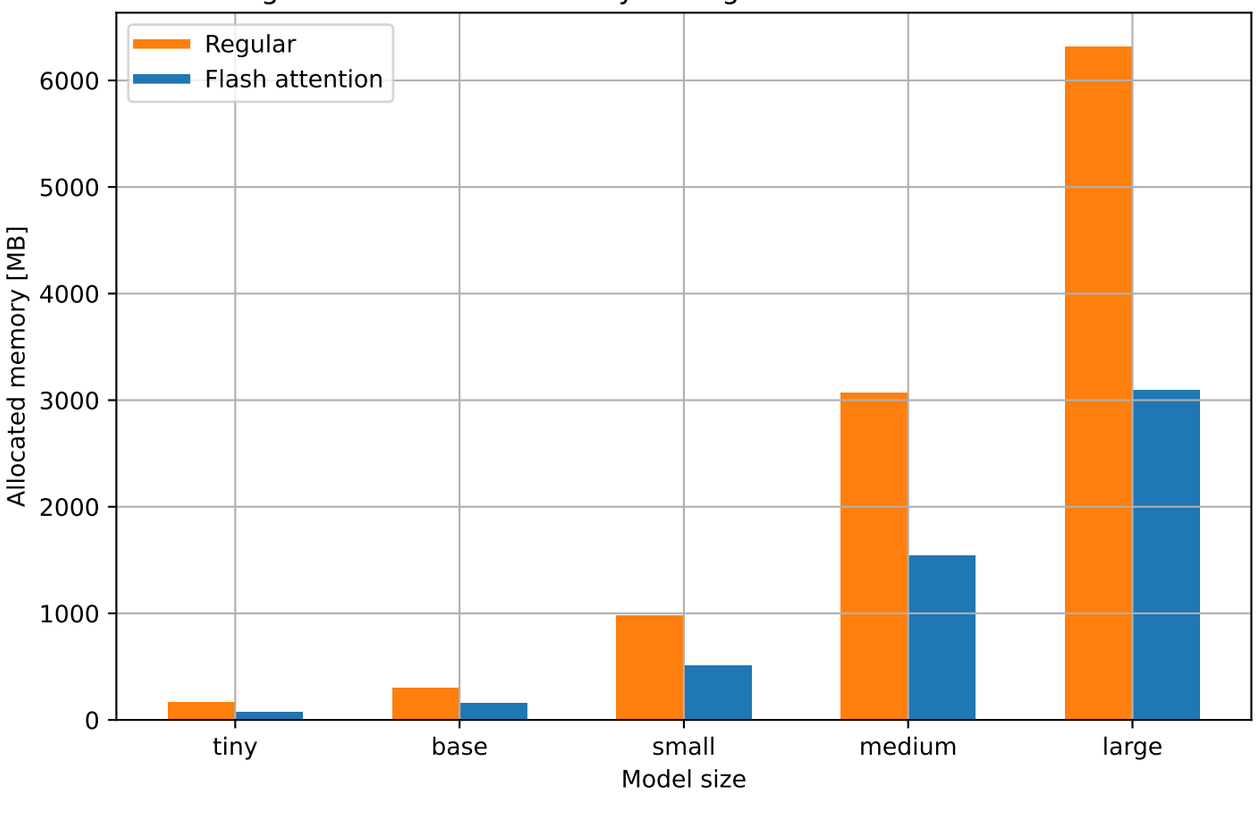
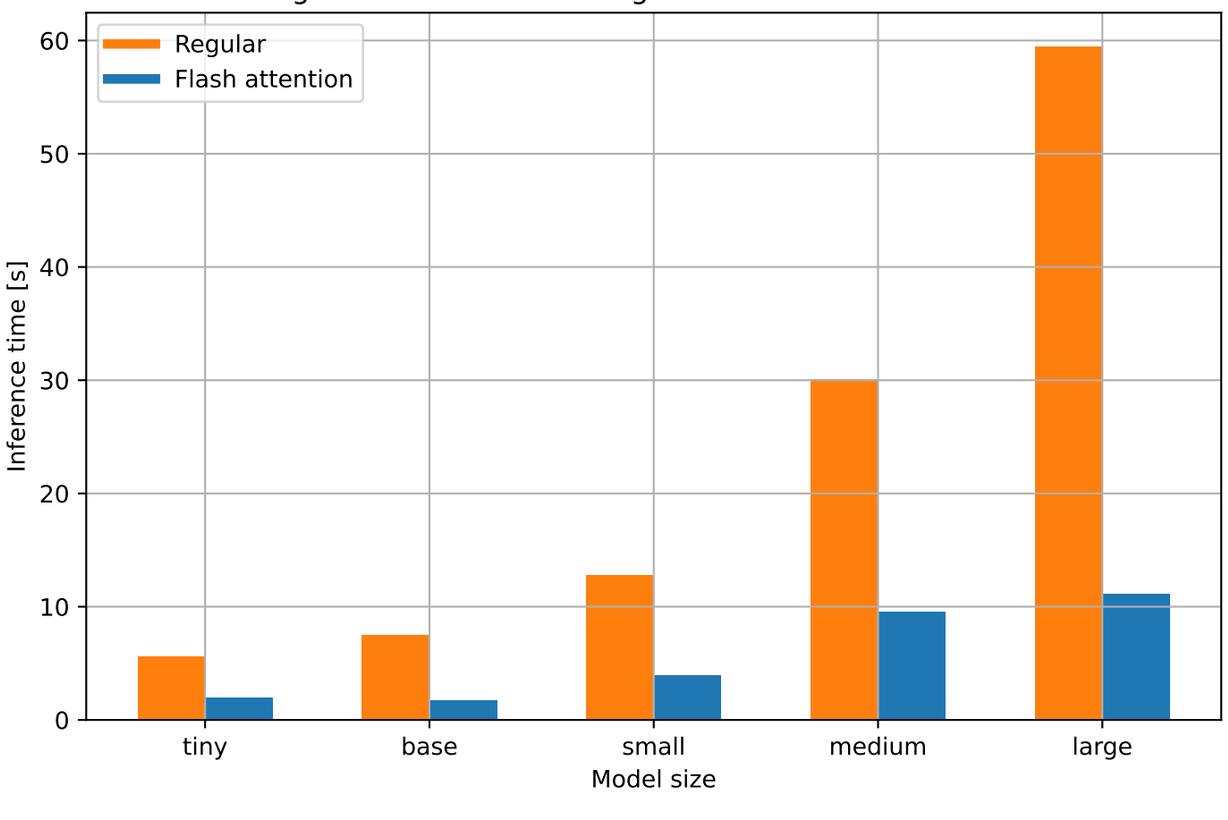

Average WER for regular and flash attention models
Regular
Flash attention
WER [-]
0.0
0.1
0.2
0.3
0.4
0.5
tiny
base
small
medium
large
Model size


Average RTF for regular and flash attention models
Regular
Flash attention
RTF [-]
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
tiny
base
small
medium
large
Model size


Average allocated GPU memory for regular and flash attention models
Regular
Flash attention
Allocated memory [MB]
0
1000
2000
3000
4000
5000
6000
tiny
base
small
medium
large
Model size


Average inference time for regular and flash attention models
Regular
Flash attention
Inference time [s]
0
10
20
30
40
50
60
tiny
base
small
medium
large
Model size