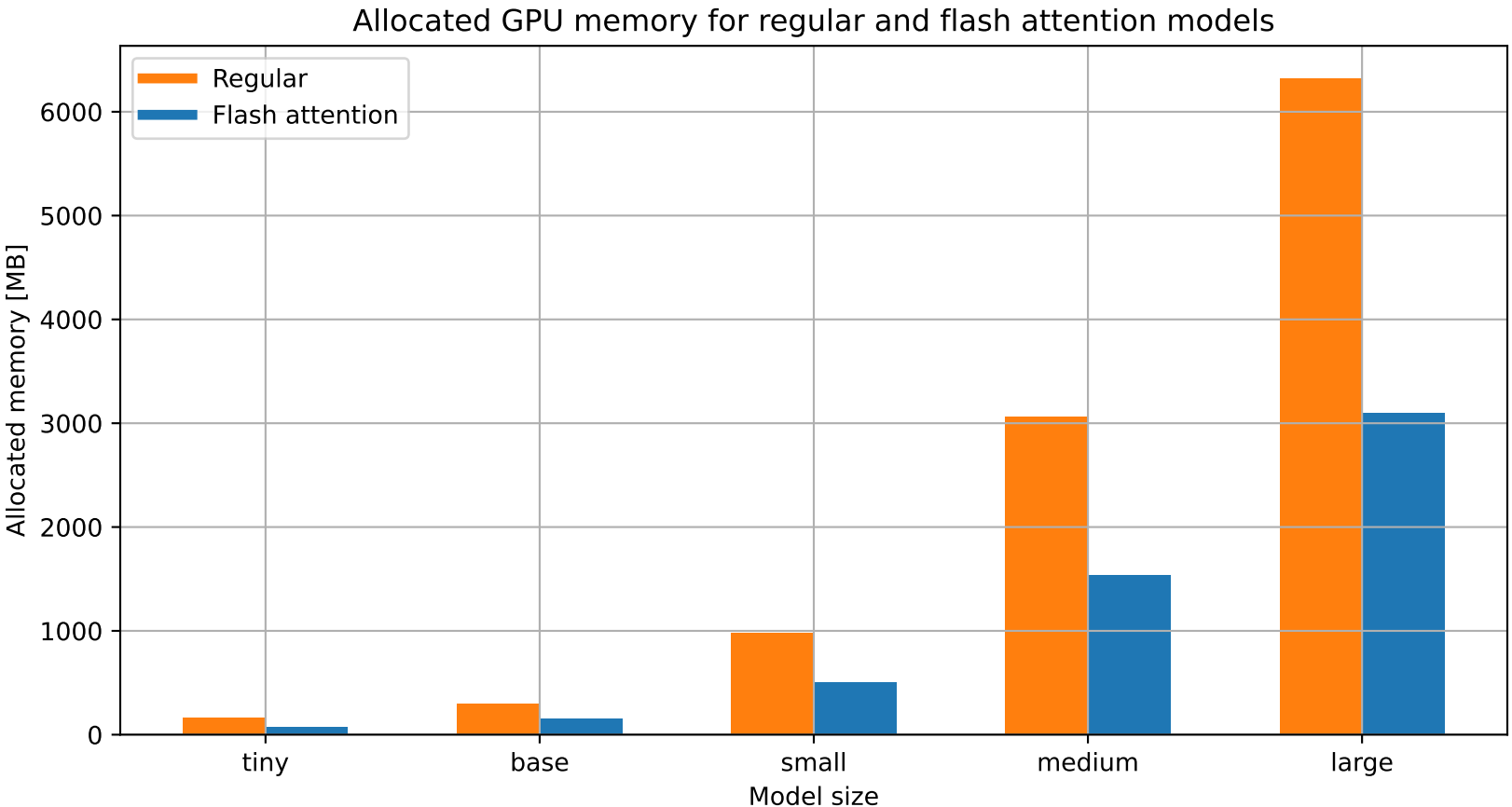
Allocated GPU memory for regular and flash attention models
Regular
Flash attention
Allocated memory [MB]
0
1000
2000
3000
4000
5000
6000
tiny
base
small
medium
large
Model size