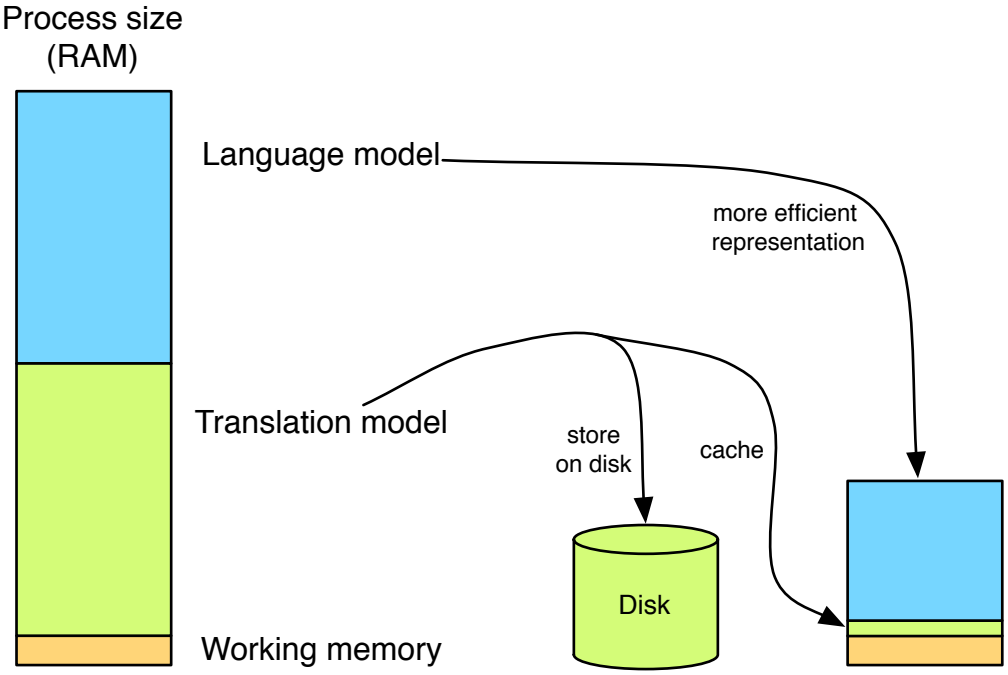
Process size
(RAM)
Language model
more efficient
representation
Translation model
store
on disk
cache
Disk
Working memory