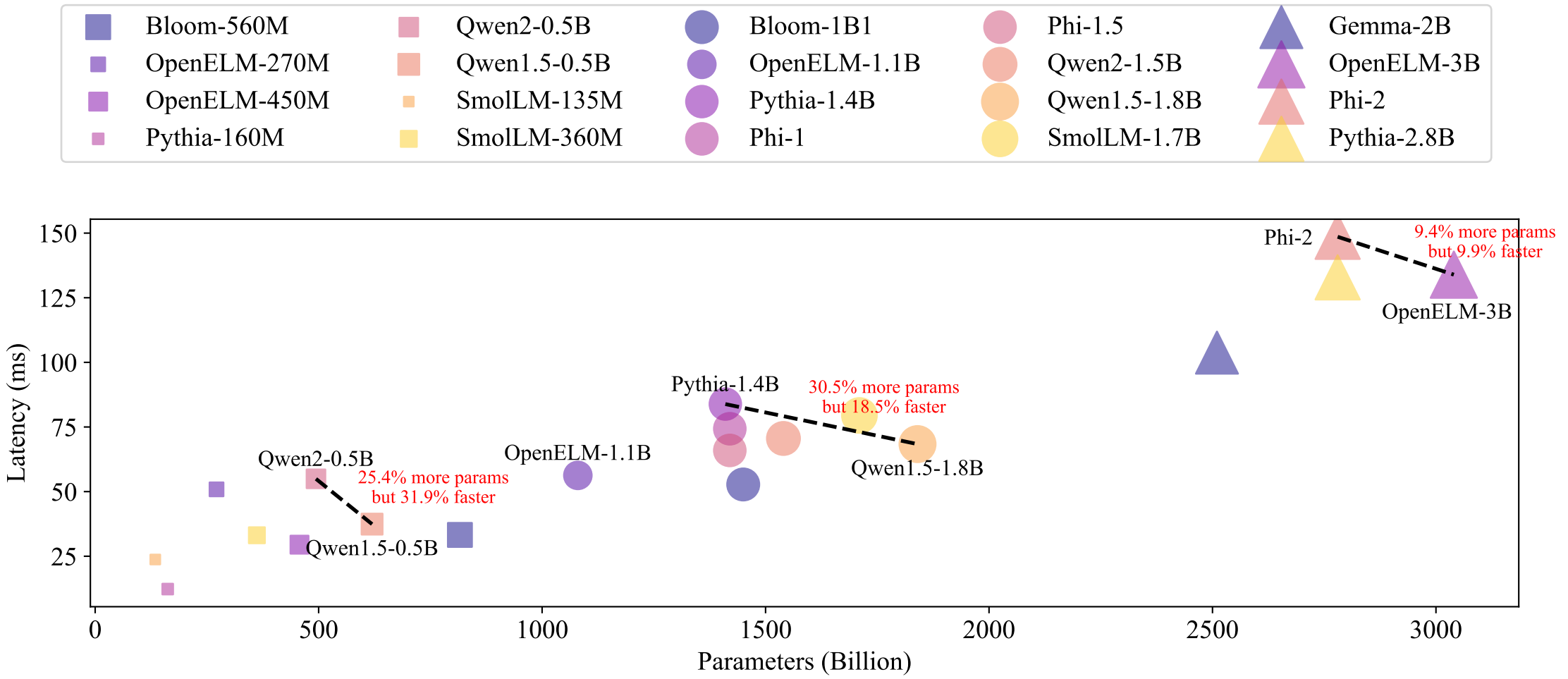
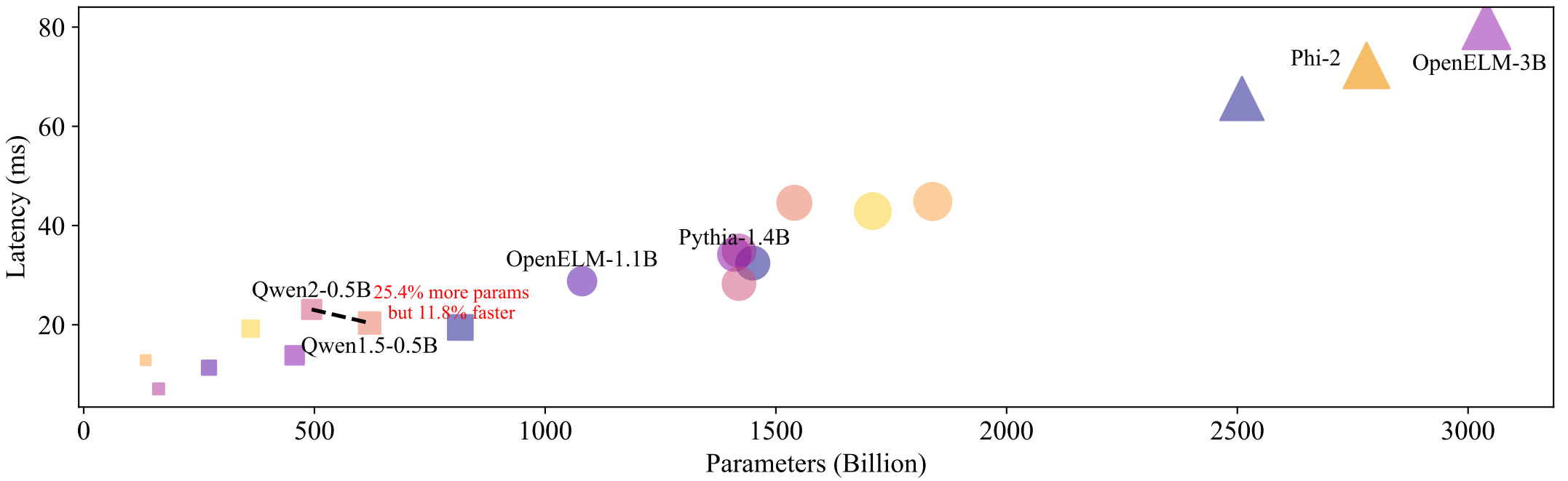
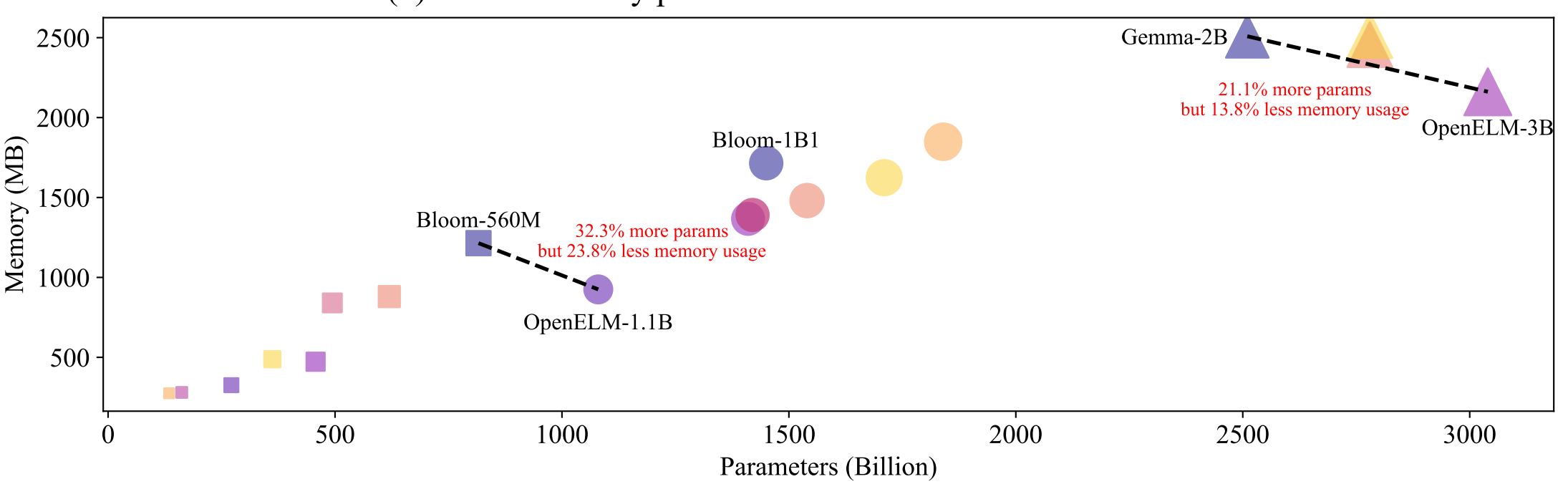


(a) First token time for models with different sizes



(b) Decode latency per token for models with different sizes



(c) Memory usage for models with different sizes