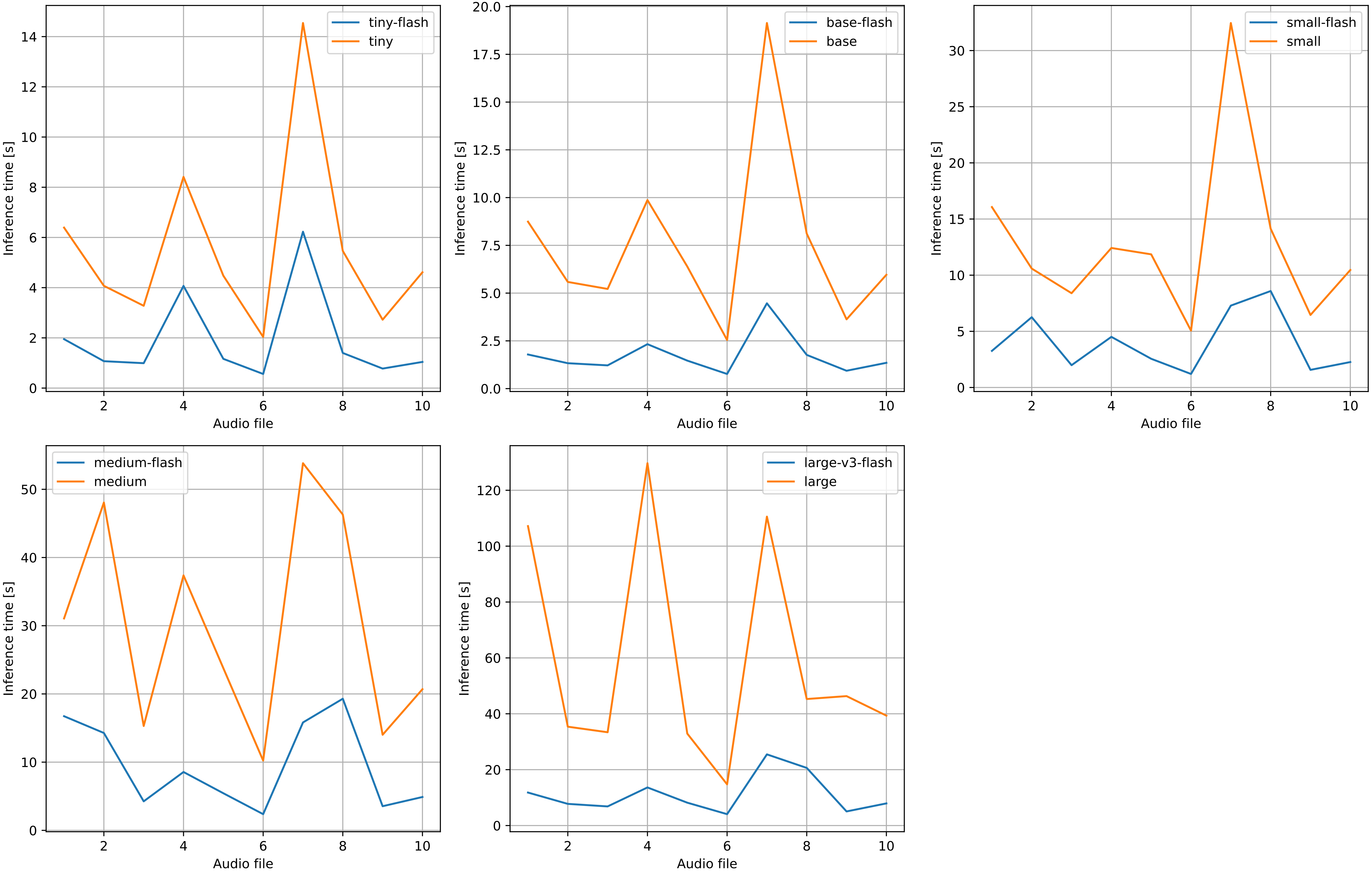

Comparison of inference times between regular and flash attention models