Comparison of RTF between regular and flash attention models