

《大模型推理词元效用评估数据集》展示样例（3条）

【样例1】EXP0001|模型precision|并发50|上下文short|降级off|轮次1|P50=45.2ms|P95=72.7ms|P99=89.9ms|词元吞吐857.9/s|输出质量0.960|单位成本0.03500元|效用0.77971

【样例2】EXP0601|模型balanced|并发500|上下文medium|降级off|轮次1|P50=34.3ms|P95=55.0ms|P99=64.9ms|词元吞吐13445.9/s|输出质量0.872|单位成本0.01200元|效用0.88782

【样例3】EXP1525|模型ultra_light|并发1000|上下文ultra|降级on|轮次1|P50=84.0ms|P95=133.7ms|P99=162.1ms|词元吞吐11642.5/s|输出质量0.474|单位成本0.00066元|效用0.84535

字段同《数据结构说明》,共13项。