Tutorial 1：工作负载与评价指标

题目

1. [单选题] 在大规模离线评测场景下，通常应优先报告哪类指标？A. TTFT 与 P99 B. 吞吐、总完成时间与成本 C. 工具调用成功率 D. 首 token 交互体验

---

2. [判断题] 在线问答场景中，只要 throughput 更高，用户体验就一定更好。

---

3. [简答题] 若用户主要反馈“首句出现过慢”，为什么应优先检查 TTFT，而不是平均 throughput?

4. [多选题] 下列哪些因素会使 RAG 服务的指标体系不同于普通在线对话服务？A. 检索链路时延 B. 长 prompt 带来的 prefill 压力 C. 总是只关心平均吞吐 D. 全链路端到端响应时间

---

5. [计算题] 某系统在 20 秒内完成了 120 个请求，其中 84 个请求满足既定 SLO。请计算 throughput 与 goodput。

---

6. [判断题] 如果大量请求虽然完成但违反了 SLO，那么 throughput 仍然可以完全代表真实服务质量。

---

7. [简答题] 为什么 P99 往往比平均值更接近真实用户体验?

8. [多选题] 在长短请求混合的 workload 中，哪些现象会让单一指标更容易误导判断？A. 长请求拉高平均值 B. 短请求对首 token 延迟更敏感 C. 所有请求瓶颈完全相同 D. 不同请求对尾时延容忍度

不同

---

9. [伪代码题] 请写出一个简短伪代码，根据 workload 类型在“离线批推理、在线对话、RAG”三类场景中返回应优先报告的核心指标。
10. [简答题] 为什么“选择哪些指标来报告”本身就属于系统设计问题，而不是结果展示问题?

参考答案

1. B。离线场景通常优先关注吞吐、总完成时间和成本。
2. 错。更高 throughput 可能以更差的 TTFT 或尾时延为代价。
3. 因为用户首先感知的是首 token 等待时间，而不是系统的平均完成量。
4. A、B、D。RAG 需要同时考虑检索链路、长 prompt 和全链路时延。
5. throughput = 120 / 20 = 6 请求每秒；goodput = 84 / 20 = 4.2 请求每秒。
6. 错。此时 goodput 更能反映真实服务质量。
7. 因为用户体验常由最慢的一批请求决定，而不是由平均水平决定。
8. A、B、D。混合 workload 会放大不同请求类型对指标的敏感性差异。
9. 示例伪代码：

```
function select_metrics(workload_type):
     if workload_type == "offline":
          return [throughput, total_time, cost]
     if workload_type == "online_chat":
          return [TTFT, TPOT, P99]
     if workload_type == "RAG":
          return [end_to_end_latency, TTFT, retrieval_latency]
```

10. 因为指标决定了系统优化目标和结果解释边界。