

洞察力

人工智能时代的网络

从速度领先到AI体验

人工智能体验正成为移动网络运营商（MNO）面向客户差异化竞争的新焦点——也是当今CXO议程中的核心议题。决定最终成败的关键绩效指标（KPI），包括首包延迟、流媒体流畅度、多模态上行、一致性等，横跨接入网、核心网、云网，而在共享层市场，更涵盖对活跃层本身的治理。

速度正在被重新定义，而非取代。

过去三十年，网络领导力一直取决于覆盖范围、语音可靠性、吞吐量、视频质量和游戏延迟。这些因素依然重要。但人工智能工作负载关注的是另一种速度：第一个数据包到达的速度有多快，30秒的响应流式传输是否流畅，多模态提示上传是否可靠，以及这种速度是否能在不同地点和不同时间段保持一致。

合作

WITH:

P3

哈坎·埃克门，P3
卡里姆·塔加·赫拉德·法伊兹博士，P3
克里斯托夫·乌费勒，P3
丹尼尔·古兹米奇斯博士，P3

行业默认的KPI奖励持续的大批量吞吐量，并将延迟视为一个粗略的阈值——因此，一个运营商可以在全国吞吐量上领先，却提供明显弱于拥有更响应式端到端路径的竞争对手的语音AI或代理体验。传统的评分卡无法可靠地检测出这种差距，而这种差距正是下一轮商业差异化的所在。

三十年来，网络领导力一直取决于覆盖范围、语音可靠性、吞吐量、视频质量和游戏延迟。

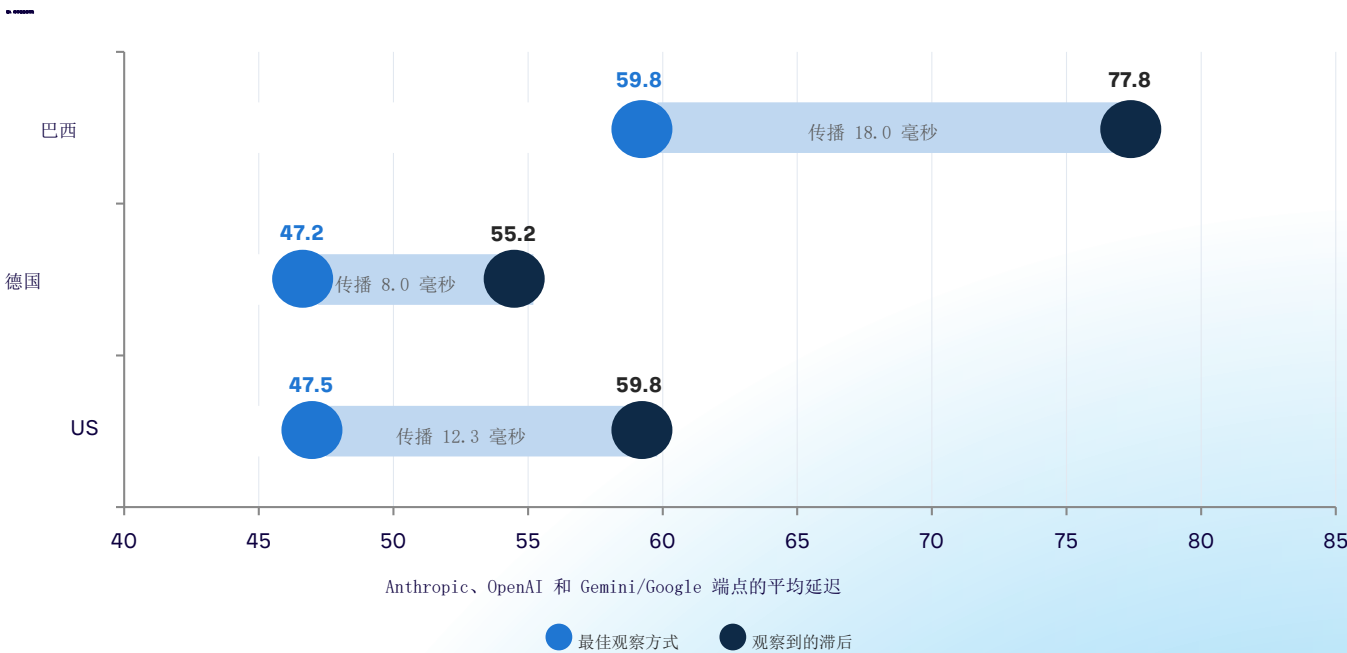
人工智能体验现已可衡量

P3的2026基准测试框架新增了一个专门用于AI服务的层级，该层级测试AI端点的连接性以及实时提示功能。它针对主要端点（OpenAI、Anthropic、Gemini/Google）测量延迟、抖动、成功率以及百分位数性能，并通过P3的DriveTest流程运行真实提示，以捕获会话成功率、首个令牌响应时间（TTFT）和响应时间。端点可达性显示了服务在可接受的网络行为下是否可达，而提示指标则说明了交互是否被终端用户感知为响应迅速。

美国、德国和巴西的初步结果显示，物质已通过市场、运营商路径和终端点进行传播。

跨市场延迟差距至AI终端端大致为8-18毫秒（见图1）；在德国DriveTest样本中，TTFT在领先运营商处约为0.94秒，在落后运营商处为1.11秒（见图2）。单次查询来看，差异似乎不大。但在通过移动网络每日运行的数亿AI会话中，这其中的差异决定了AI是感觉流畅还是感觉生硬。之前的评分卡未能反映这一点。

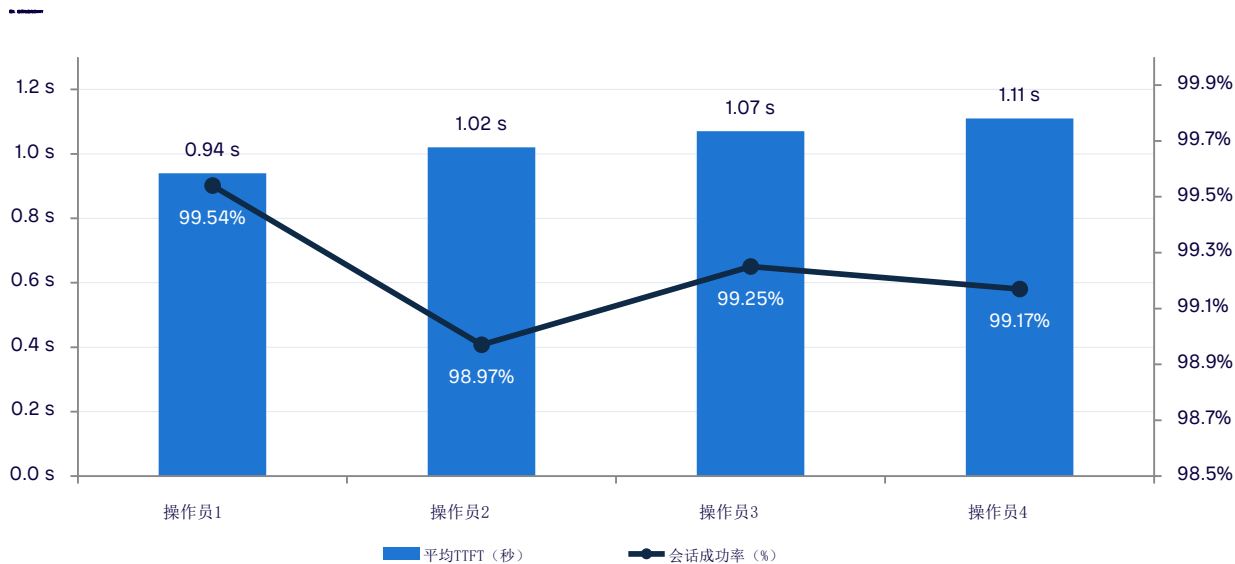
图1. 跨市场延迟对比



注意：操作员身份已匿名 来源：Arthur D . Little, P3

图2. 德国运营商INSIGHT的TTFT

亚瑟·D·利特尔



注意：按平均TTFT排序的操作员 来源：Arthur D. Little, P3

注意：按平均TTFT排序的操作员 亚瑟·D·利特尔, P3

网络竞争力关键绩效指标 (Wǎngluò Jìngzhēnglì Guānjiàn Jìxiàozhǐ)

亚瑟·D·利特尔公司预计，未来几个基准周期内，六个关键绩效指标（KPI）将成为网络竞争力的主要驱动因素。每个指标都易于针对受控的推断端点进行衡量，并与一个熟悉的指标相对应（最接近的代理指标显示在下划线/括号中），因此运营商和评估机构可以无需重建方法论即可采用它们：

1. TTFT——从发送提示到第一个响应标记的延迟，与标准托管模型（呼叫建立时间）进行对比

2. 令牌流平滑性 — 在30秒的响应（视频缓冲率）期间，令牌数每秒的持续性和间隔可变性

3. AI响应延迟——标准多模态任务（文本+压缩图像）端到端时间：上传、推理、流式响应（文件下载会话时间）

4. 多模态上行链路 — 对于1-5 MB的有效载荷，在平均情况和较低尾部（文件上传会话时间）下均能保持持续的上传吞吐量

5. 一致性得分——同时满足上述四个KPI最低标准的会话比例（速度等级履行）

6. 边缘人工智能延迟——对于拥有多接入边缘计算（MEC）的运营商，其自建推理与公共云终端（当前无代理）相比的延迟优势

应尽早确定三个设计选择：（1）在专用基础设施上，针对主要基础设施供应商以及一个中立、与运营商无关的模型进行运行精度测试；（2）利用5G独立组网（SA）的首个令牌收益，最终以用户实际感受的方式使SA业务案例可见；（3）为受控道路测试预留详细的人工智能指标，使用众包数据以获得更广泛的连贯性和可用性视角。

杠杆组加宽，尤其是在活动层共享处

TTFT、流媒体平滑度、多模态上行链路和终端可达性并非仅由RAN（无线接入网）决定。它们将历史上以工程为主导的资产——用户面功能（UPF）部署、互联网交换（IX）/对等互联策略、服务质量（QoS）策略、边缘计算（MEC）、端到端——纳入商业路径。运营商可以运行一流的RAN，但仍可能因对等互联薄弱、核心网突发拥塞或边缘计算缺失而导致AI会话质量不佳而失败。

在活动层共享的情况下（大多数欧洲市场及几个海湾市场），这一点尤为突出：AI关键绩效指标（KPI）对调度器、QoS（服务质量）和路由设置敏感，且对不透明的共享层行为容忍度极低。以下是五个治理举措：

1. 明确零售运营商、共享合作伙伴以及NetCo/ Ser vCo架构中的AI KPI责任归属——哪个KPI由谁负责，在哪个边界范围内。
2. 将共享服务水平协议（SLA）的涵盖范围从覆盖率和吞吐量扩展到抖动、最坏情况上行性能和会话级一致性。
3. 为调度器、QoS（服务质量）、切片、本地 breakout（本地中继）和路由等实际影响 AI 关键绩效指标（KPI）的控制功能获取安全配置权限。
4. 建立跨RAN、传输、核心和云的联合可观测性，以便能够将问题归因于特定层级并在源头解决。
5. 通过针对人工智能的特定服务类别和企业级SLA，保留零售差异化的空间，以免共享层压垮商业故事。

人工智能网络作为高管议程

对于首席技术官（CTO）而言，AI体验将网络重新定义为端到端的性能系统；会话级别的成果所有权必须涵盖无线接入网（RAN）、核心网、传输网和云——不能局限于任何单一领域。

对于CSO而言，“AI工作负载的领先网络”是一个与其性能指标不直接竞争的差异化方向，它直接映射到B2B领域——如联络中心语音AI、现场多模态助手、CPaaS路由——并且在主动共享市场中，它也映射到选择性去共享或高级SLA层级的案例。

对于首席财务官而言，那些难以提升每用户平均收入（ARPU）的投资——5G SA、边缘计算、服务保障、IX升级、可观测性——能够为B2B合同带来可衡量的终端客户联系，从而实现变现，而非仅仅依赖消费者ARPU。