GenAI 在 OSS/BSS 领域价值正在显现，但仍需继续投入新能力

在电信和 IT 领域，人们普遍认为 GenAI 将在未来的技术或运营进步中发挥至关重要的作用。像 ChatGPT 这样的免费应用平台的出现，使得没有技术专长的用户也能够试验 OpenAI 的大型语言模型（LLM）。然而，为具有特定电信功能集的通信服务提供商（CSPs）定制这种技术却更具挑战性。服务商需要使用以下三种技术对他们的 LLM 解决方案进行额外调整：提示工程、微调和检索增强生成。

1. 提示工程技术应作为调优 LLM 的初始步骤

提示是一种通过调整提供的输入来引导 LLM 响应的方法。提示技术可以从基本短语到详细指令，视任务和模型能力而定。为特定任务制作和改进提示的过程，或提出正确问题的过程，被称为提示工程。提示工程的本质是帮助 LLM 在特定背景和目标下生成最合适的响应。

2. 为提高准确性，应对 LLM 模型进行微调

为了使服务商能够将基础 LLM 适应于特定数据集或特定用例，微调变得必要。这涉及监督学习过程，其中使用带标签的示例数据集来调整 LLM 的权重，从而增强其在特定任务中的能力。服务商必须能够获得大量高质量和适当格式的数据以用于此目的。微调过程的结果必须在支持的用例上下文中进行评估和验证。

此外，基于人类反馈的强化学习（RLHF）在 LLM 微调阶段至关重要。基础模型在进行监督微调之前会经历预训练阶段，然后是 RLHF。RLHF 通过构建直接来自人类反馈的最终模型来帮助分类 LLM 对提示的响应。这一过程有助于消除糟糕的响应或误解，并改进用于查询模型的提示。因此，它提高了与设计支持的用例相关的所有模型的准确性。服务商可能会进行微调以满足其特定应用领域的需求，这将成为他们的知识产权。

3. 需要考虑用于上下文数据的检索增强生成（RAG）

RAG 通过从相关数据源动态检索上下文并将其与整体提示策略结合起来，以生成基于事实的准确响应，从而解决了微调和提示工程的局限性。RAG 方法使模型能够访问外部数据以增强响应生成。这一功能使 LLM 的部署更具灵活性，因为模型可以主动参考最新的信息以生成响应。基本的 RAG 可能只是一个检索机制，它获取静态文本内容，将其附加到提示和请求/输入文本中，并将这些工程化的响应/输出文本用于生成。

解决幻觉问题

LLM 中的幻觉发生在 GenAI 生成不存在或不可能的结果时，通常是由于没有正确理解上下文而将数据组合在一起。这可能导致提出新的产品提供或创建不存在的示例。因此，有必要监控和验证输出以确保数据的准确性。服务商需要使用 RAG、微调和提示工程等方法在提示中整合特定的知识增强。这些增强需要根据正在解决的用例进行修改。

测试和扩展

在训练 LLM 并实施某些缓解措施后，仍然需要进行系统测试。通常，测试涉及在各种环境中使用应用程序，例如浏览器、云基础设施、数据库和逻辑。逻辑测试可以针对 API 进行，但由于输出的可变性，测试 GenAI 对相同或类似提示的响应可能具有挑战性。

嵌入式 GenAI 的可扩展性是一个重要的考虑因素。较小的 LLM 比较大的 LLM 部署成本更低。然而，如果模型不是开源的，可能会有额外的许可成本。较小的模型可能不太通用，更适合于它们应用的特定用例，这可能导致服务商需要多个 LLM 来支持每个用例或应用领域。这会增加它们的维护和支持成本。

大规模模型的表现可能不如小规模模型，这可能影响自动化过程和整体应用性能。当过程扩展时，这些问题可能变得更加明显。例如，一些基于机器学习（ML）的模型在扩展时未能及时提供对输入数据的推理，这突显了潜在的挑战。

构建“GenAI 原生”应用的服务商需要在新开发能力上持续投入

现有的 OSS/BSS 行业应用需要进行重组以整合 GenAI 能力。这可能需要创建用户界面来管理各种输入和输出，或引入新的 API 来处理不同的文本输入。为了最大化 GenAI 的好处，重新编写应用程序以成为 GenAI 原生可能是有利的。在逻辑曾经依赖于代码或算法的情况下，可以支持提示请求并能够提供生成输出。在之前已经开发过聊天机器人的情况下，可以相对迅速地添加 GenAI 功能。然而，对于像网络运营商使用 GenAI 进行新的网络设计这样的情况，需要对应用程序进行更具体的改动。

服务商在 GenAI 应用方面的投入方向

OSS/BSS 服务商和应用企业不仅需要在新开发能力上进行投资，还需要通过增强员工技能、能够访问大量高质量数据、拥有新的 LLM 训练框架、支持新的测试环境以及可能重建现有应用程序来支持这些能力。曾经硬编码在许多应用程序中的逻辑需要转化为具有优异操作结果的 LLM，以证明他们在 GenAI 上的投资是合理的。