三阶段的评估框架

该文章提出了一个三阶段评估框架，用于评估大型语言模型（LLMs）在自动化网络安全告警分析中的应用效果，具体分为三个阶段：**准备（Preparation）**、**计算（Computation）和解释（Interpretation）**。以下是对每个阶段的详细讲解。

1. 准备（Preparation）

在准备阶段，设计了一个系统流程，为LLMs的评估过程打好基础。该阶段包括以下步骤：

1. **选择告警类型**：从MITRE ATT&CK框架中选择要分析的告警类型（如暴力破解或网络钓鱼），并确定每个告警类型的分析步骤。
2. **定义分析步骤**：确定告警分析的每个步骤，以便生成任务场景提示。每个步骤应清晰地定义为操作员或LLM要执行的具体任务。
3. **过滤步骤与收集上下文**：逐步分析每个步骤，判断是否需要人工执行或可以通过简单的API调用完成。对于需要人类判断的步骤，应将相关信息纳入上下文信息。最终得出一组需要人工或LLM进行逻辑推理的步骤列表。
4. **基准建立**：为每一步骤设置预期的结果作为基准，以便与LLMs的响应结果进行对比。例如，电子邮件的紧急程度可以通过“紧急”或“非紧急”来表示，以便作为准确性的对比标准。
5. **模型选择**：选择要评估的LLMs，文章根据模型的通用性、可用性和性能来选择模型。公开的LLMs更容易获取，私人模型如GPT-4性能较高但获取难度较大。因此，平衡模型的广泛应用和具体性能至关重要。
6. **信息结构化**：将收集到的告警信息、基准数据、每个步骤的提示信息等结构化组织，以便后续算法可以读取和处理。

2. 计算（Computation）

计算阶段的目标是根据准备阶段所定义的步骤，对LLMs进行实际的性能测试，并记录各种关键性能指标（KPIs）。该阶段主要流程包括：

1. **创建并运行算法**：设计一个算法，对每个模型和提示组合逐步进行评估。在每个步骤中，算法会加载场景提示和系统提示，形成一个完整的测试提示，并传递给模型。
2. **记录性能指标**：算法会记录每次模型响应的关键性能指标，包括：
    - **准确性**：模型响应是否符合基准数据（如与人工操作员给出的答案一致）。
    - **响应时间**：模型生成响应所需的时间，用于衡量模型的效率。
    - **字符长度**：评估模型响应的字符数，确保模型在响应时符合预期的输出格式（如简单的“是”或“否”回答）。
3. **数据存储与整理**：将模型响应的准确性、时间和字符长度等数据整理为一个汇总表格，展示各模型在不同情境下的性能表现，以便后续分析和比较。

3. 解释（Interpretation）

解释阶段旨在分析和解读计算阶段的结果，以提供模型在实际应用中的有效性、局限性和改进方向。具体包括以下内容：

1. **分数分析**：对比各模型的准确性分数，较高分数表示模型对基准结果的匹配度较高，表明其在特定告警情境下的有效性。

2. **时间分析**：时间分析关注响应时间。较短的响应时间意味着更高的效率。模型在高负荷实时应用中的适用性会受到响应时间的影响。
3. **字符长度分析**：如果模型在回答时严格遵循预期的字符长度（如仅回答“是”或“否”），则表明模型响应表现更为一致、可预测性更强，反之则可能出现多余信息或格式不符合预期的情况。
4. **模型对比**：对比各模型在同一告警分析任务中的得分、时间和长度表现，以确定最适合告警分析的模型。
5. **告警分析**：在相同模型下对不同告警类型的表现进行对比，以识别模型在特定类型告警分析中的优势或局限。
6. **提示分析**：对比不同提示设计的效果，评估哪些提示更有效，以便为未来的提示设计提供指导。

通过上述三个阶段的评估框架，文章实现了对LLMs在告警分析任务中的全面测试，从而得出LLMs在网络安全告警分析中的有效性、优缺点及未来的改进方向。这一框架也为未来的LLMs研究提供了一个系统化的评估模板。