# 基于稀疏自编码器的可编辑大模型生成系统

## 系统背景:

随着大语言模型（LLM）的广泛应用，如何根据特定需求对其进行编辑成为了一个关键问题。大语言模型虽然在通用任务上表现优异，但在面对个性化需求或特定领域时，常常需要对其内部知识进行修改或引导（steering）。基于此，本文提出了一种基于稀疏自编码器（SAE）的可编辑大模型生成系统，旨在通过修改模型的内部表示实现对其生成内容的多维度编辑，包括但不限于安全性、情感、人格等方面的调控。

## 主要功能

本系统旨在基于稀疏自编码器（SAE）技术对大语言模型进行编辑，主要功能模块包括SAE训练模块、SAE特征选择模块、编辑模块和效果评估模块。

（1）SAE训练模块  
该模块用于对稀疏自编码器进行训练，用户可以配置训练参数。

（2）SAE特征选择模块  
该模块通过输入数据样本，从SAE的输出中筛选出最具代表性的特征，用于后续的模型编辑。数据样本可以包括不同的编辑目标（如安全性、情感等），系统根据数据对特征进行自动筛选，用户也可以手动选择要进行编辑的特征。

（3）模型编辑模块  
该模块负责将选择的特征用于大语言模型的内部表示编辑。通过注入特定的特征，系统可以对模型的生成输出进行定向引导，确保生成的内容符合预期的安全性或情感等要求。

（4）效果评估模块  
该模块用于评估编辑后的大语言模型的性能。系统可以自动化运行一系列测试，检查模型是否满足编辑目标，用户也可以通过手动反馈进一步优化编辑结果。

## 开发环境

操作系统及版本：Ubuntu 22.04.2 LTS

编程语言：Python 3.11

深度学习框架：PyTorch（用于模型训练和编辑）

模型基础：LLaMA、Gemma等等大语言模型

硬件配置：支持GPU加速，使用NVIDIA CUDA环境

依赖库：transformers、numpy、scipy 等常用深度学习库

## 基本的架构体系

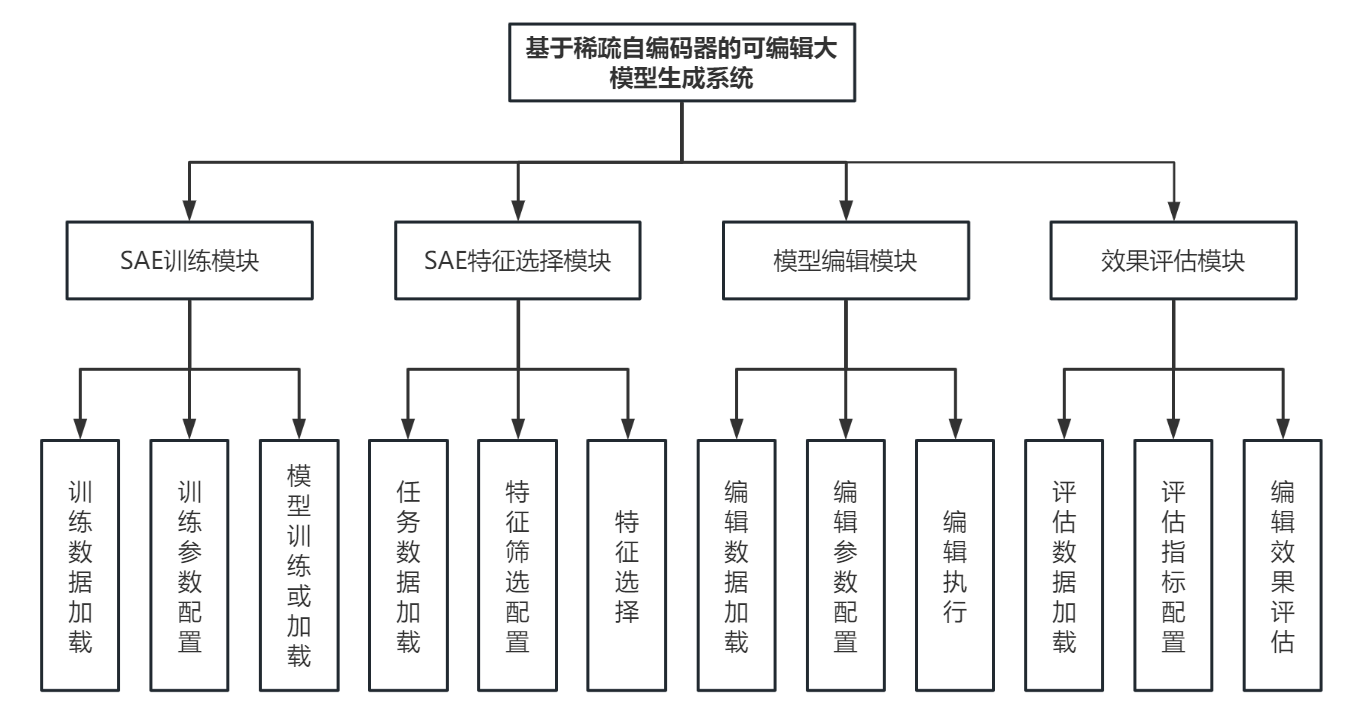


图 1 基于稀疏自编码器的可编辑大模型生成系统功能结构图

系统的功能架构分为四大模块：SAE训练模块、SAE特征选择模块、编辑模块和效果评估模块，各模块相互配合，实现对大语言模型的可编辑生成。

（1）SAE训练模块：该模块主要负责对稀疏自编码器（SAE）进行训练。用户可以自行设置输入数据集，配置模型参数如学习率和正则化强度等，完成自编码器的训练。

（2）SAE特征选择模块：该模块负责筛选和提取训练后SAE模型生成的特征。根据用户提供的特定任务数据集，系统可以自动选择最相关的特征，或允许用户自定义筛选标准，以确保选定的特征适用于后续的模型编辑过程。

（3）模型编辑模块：在该模块中，用户可以利用选定的特征对大语言模型进行编辑。编辑过程会影响模型的内部表示，调控生成的内容，使其符合预设的方向，如情感、安全性或特定的语言风格。

（4）效果评估模块：该模块对编辑后的大语言模型进行效果评估。用户可以查看评估结果，通过定量指标和定性分析确保模型达到预期效果，支持多维度的编辑目标测试。