数据来源:midjourney   吐司Tusi.art 图虫

数据训练：

文本数据：收集大量包含AI生成和人工编写的文本数据。这些数据具有多样性，包括科普、散文、小说、学术语言等多种类型的文字。  
 图片数据：同样地，收集包含AI生成和真实拍摄的图片数据。这些数据包括写实、动漫、3D、建筑、素描、人物、动物、植物等不同类型的图片，很好的覆盖了生活中可能遇到的各种图片  
 数据分类:关于文本数据，我们通过词袋模型和词向量模型，将文本转换为向量表示，以便机器学习算法可以处理。词袋模型将文本视为文字单词的集合，而词向量模型则试图捕捉单词之间的语义关系。对于图片数据，我们先基于数据做了文件分类，然后我们使用标签和关键词来进一步描述图片的内容，以便更快地找到所需要的图片，然后通过卷积神经网络CNN进行特征的提取，将数据转换为可供模型训练的特征表示

训练：使用训练数据集进行模型训练。在每个训练迭代中，将输入数据提供给模型进行前向传播计算，得到预测结果，然后计算损失函数的值。接下来，使用反向传播算法计算梯度，并使用优化算法更新模型参数，使损失函数逐渐减小。在模型训练过程中，可能需要调整一些超参数，如学习率、正则化项系数、隐藏层大小等。可以使用交叉验证技术，通过在验证集上评估模型性能来选择最佳的超参数组合。

改进过程：我们通过数据分析，仔细分析和理解数据集，包括数据的分布、特征之间的关系、异常值等。发现数据的特征和模式，并为后续的改进提供指导。针对数据不足或样本不平衡的情况，我们考虑数据增强技术，如旋转、翻转、裁剪、添加噪声等，来生成更多的训练样本，以提高模型的泛化能力。