一、引言

知识图谱是人工智能基础设施的重要组成部分。它能够帮助我们以一种较强的逻辑方式沉淀和组织互联网中海量信息，从中提取出有用的知识，并进一步构建大规模的知识图谱，为人工智能上层应用的开发提供重要的战略优势。通过利用知识图谱，我们可以更好地组织和理解海量信息，并为各种应用场景提供智能化的解决方案。

1.1 技术特点及优势

大模型数据工程构建技术。针对目前主流的以数据为中心的 AI 大模型，通过分析应该要用什么样的数据训练模型、应该选取什么样的数据源、数据怎么进行预处理以及最终训练好的模型怎么进行自动化的评估。通过上述技术手段，能够充分的构建通用大模型及领域大模型数据工程，实现以数据为驱动的大模型构建

1.2 成果状态及关键指标

首先可以使用自监督预训练模型得到问题的环境状态编码表示；然后基于错误率设计强化学习问题的奖励函数，使得奖励越大的输出的错误率越小；最后将模型视为强化学习策略函数，使用策略梯度算法(REINFORCE)对其进行优化微调，进而提高模型及系统的性能。

二、研究内容

数据工程在以数据为中心的 AI 中起着至关重要的作用，通过改进数据集的质量可以提升模型的效果。对于大模型构建来说，数据工程仍然扮演了非常重要的角色。面向大模型场景数据工程时需主要解决的几个关键问题。

(-1, 1)	(0, 1)	(1, 1)
(-1, 0)	(0, 0)	(1, 0)
(-1, -1)	(0, -1)	(1, -1)

(a) 像素位置坐标

σ = 0.8

1	2.184	1
2.184	4.771	2.184
1	2.184	1

(b) 高斯滤波值

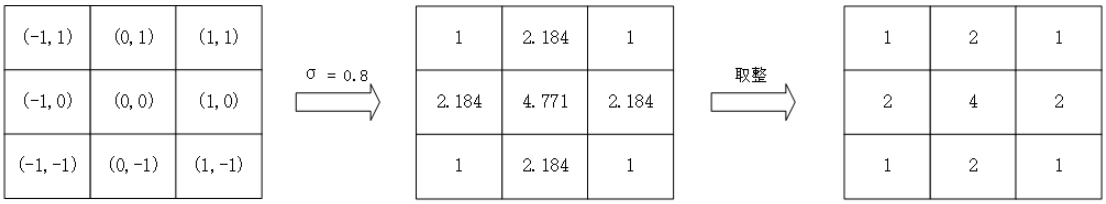
取整

1	2	1
2	4	2
1	2	1

(c) 高斯卷积核

三、大模型数据工程构建技术

为了构建起大规模且多样性的数据，需要广泛收集并标准化各类语料，建立完备的数据体系和数据来源，并分开收集不同类型的数据。为了确保数据的质量和多样性，需要考虑以下解决方案



(a) 像素位置坐标　　(b) 高斯滤波值　　(c) 高斯卷积核

四、参与人员

这是一个表格

姓名	性别	年龄	职位
A	男	26	助理
C	男	28	教授
E	女	29	副教授

五、技术指标

支持在多种硬件平台上部署，可以灵活地在不同的部署环境中进行调整；