数据科学导引 B-期末作业

课程综述需要选择若干与所选课题相关的文献进行阅读总结。加分项:

(1) 在选定课题内调研尽可能多的文献；

(2) 调研的文献贴近数据科学研究前沿；

(3) 在调研中总结分析自己对该领域的看法，如存在的问题和未来展望。

(4) 选取部分优秀的调研报告，邀请线上做 presentation。

综述建议阐述以下内容：

（1）该领域的研究重点以及发展过程；

（2）重要文献中的方法概括；

（3）重要的数据集简介；

（4）可以继续扩展的研究方向。

组队形式：6 人以内/个人均可，可以和期中项目一样，也可以重新组队

注：如果是团队，需要在邮件正文和调研报告里，均注明团队成员姓名+学号，并在课程报告里明确指出团队成员的分工。

建议使用 latex（非强制），推荐采用英文写作，最终提交 pdf 的文档，请按照规定时间将 survey 提交至对应助教邮箱，邮件名请写明所有组员名单，同时注明每位组员的分工。命名方式为: 选题名-xx.pdf，如 2.文本数据选择与评估-张三-李四-王五.pdf

课程综述推荐格式，推荐有详细的图标描述：

https://dl.acm.org/doi/abs/10.1145/3626235

https://dl.acm.org/doi/full/10.1145/3535101

https://arxiv.org/abs/2405.16640

Latex 模板推荐：

https://www.overleaf.com/latex/templates/association-for-computing-ma...

https://www.overleaf.com/latex/templates/neurips-2024/tpsbbrdqcmsh

https://www.overleaf.com/latex/templates/acm-conference-proceedings-p...

每个题目（除题目 14 和 15）限选 5 队，Survey 写作请保证原创性，严禁直接抄袭！

查看选组情况：数据科学导引-选组情况

报告提交 DDL：2025 年 1 月 19 日

Presentation DDL： 1 月 24 日 （提前提交的同学，如果报告做得足够好被选中也可以提前 Pre，比如 1 月 10 号提交 15 号做 pre，尽量不要等到 DDL）

考试周：2024 年 12 月 30 日至 2025 年 1 月 12 日

成绩录入 DDL：2025 年 1 月 26 日

1. 文本数据合成

在数据中心化机器学习（Data-Centric Machine Learning, DCML）领域，纯文本数据合成是指通过自动化手段生成大量文本数据，以满足模型训练所需的数据多样性和质量要求。与传统的数据采集方法相比，数据合成技术能够大规模生成高质量的文本数据，尤其在数据稀缺或特定任务需求较为复杂的情况下，具有重要的应用价值。

纯文本数据合成的核心目标是生成具有足够信息量和语义多样性的文本，以增强模型的泛化能力和鲁棒性。通过合成技术，研究人员可以在保持数据质量的前提下，灵活调整生成数据的风格、内容以及结构，以适应不同任务的需求。

常见的文本数据合成方法包括：

1.基于模板的生成

通过预定义的模板和规则生成文本，适用于结构化问题和重复性高的任务，如问答系统、对话系统等。

2.基于生成式模型的生成

利用现有的大型语言模型（如 GPT 系列、BERT 等）生成文本。这类方法能生成更加自然、富有创意的文本，广泛应用于文本生成、机器翻译等任务。

3.数据增强方法

通过对现有文本数据进行语义替换、重组、加噪等方式合成新的数据，用于提升模型对不同表达方式的适应性。

纯文本数据合成不仅能够补充真实世界数据的不足，还能在特定任务的训练中起到数据平衡、数据增强和知识迁移的作用。尤其在低资源语言、稀有任务或特殊领域，合成数

据能极大地提升训练效果和模型表现。

然而，纯文本数据合成也面临着诸如生成数据质量控制、语义一致性保证等挑战。因此，如何在合成过程中保持文本的多样性和质量，成为当前研究的热点之一。

参考文章链接：

MAGPIE: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing

Scaling Synthetic Data Creation with 1,000,000,000 Personas

AI models collapse when trained on recursively generated data

https://github.com/pengr/LLM-Synthetic-Data

本题作业请发到 2301111449@stu.pku.edu.cn

2. 文本数据选择与评估

文本数据选择与评估是确保高质量模型训练的核心环节，涵盖了从筛选适合训练的数据到对其质量和多样性进行严格评估的全过程。数据选择旨在筛选出与目标任务紧密相关且具有高质量的文本数据，确保模型能够高效学习。首先，所选数据需具备任务相关性，能够准确反映目标任务的语义需求；其次，数据应覆盖广泛的情境和风格，提供多样化的表达方式，避免训练数据的单一性。此外，还需保证数据质量，剔除含有语法错误、不清晰语义或标签不一致的样本，确保数据的准确性和一致性。

在评估阶段，通过检查数据覆盖度、语义有效性和标注准确性，进一步验证数据是否符合任务需求。具体而言，评估需关注数据集是否涵盖关键任务场景和语义，标注是否准确且一致，同时确保数据能够代表目标任务的实际应用场景，避免因数据偏差导致模型性能下降。合理的选择与评估能够显著提升模型对目标任务的理解能力，为后续的模型训练提供强有力的数据支持。

挑战

尽管文本数据选择与评估方法不断进步，但仍面临一些挑战。首先，数据的质量控制是一个复杂问题，特别是在大规模数据集中，如何快速识别和过滤低质量或噪音数据至关重要。其次，不同标注者之间的标准差异可能导致标签的不一致性，影响数据的可靠性。此外，平衡数据多样性与任务相关性也是一个难点，过度追求多样性可能引入无关数据，而过于专注相关性可能导致模型的泛化能力不足。最后，评估标准的设定往往需要根据具体任务动态调整，如何高效评估数据的全面性和代表性依然是一个重要研究方向。

通过改进数据选择策略和评估方法，结合自动化技术，可以有效应对上述挑战，为模型提供更高质量的训练数据，从而进一步提升模型的性能和适用性。

参考文章链接：

AlpaGasus: Training A Better Alpaca with Fewer Data

From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning

MoDS: Model-oriented Data Selection for Instruction Tuning

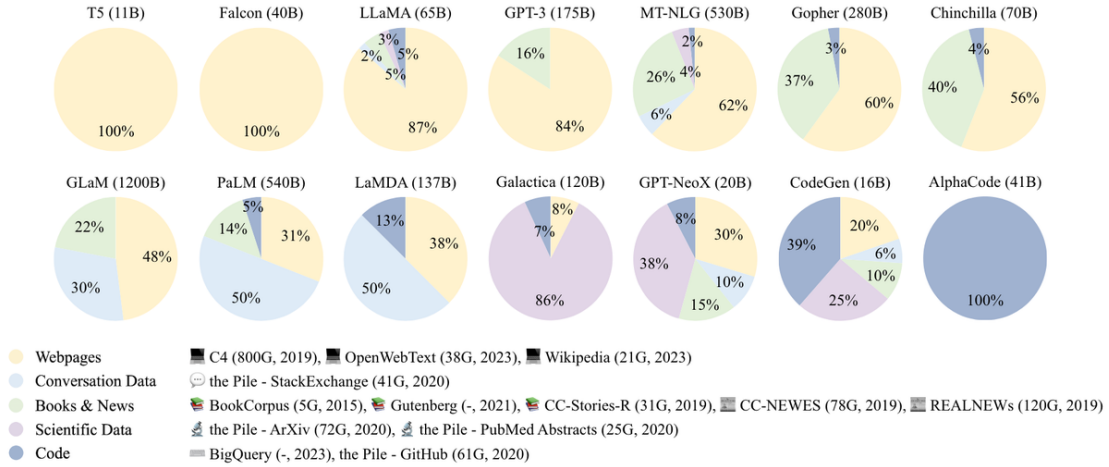
可选 Topic：文本语料，数学语料，代码语料等语料质量的评估

本题作业请发到 2301111449@stu.pku.edu.cn

3. 文本数据配比

在大数据环境下，数据来源多样，类型复杂，如何有效地进行数据融合，提高模型的预测精度，是当前的研究重点。此外在许多实际应用中，**数据的类别分布常常是不平衡的**，这对于模型的训练和优化带来了挑战。

大模型的数据配比是指在训练大规模深度学习模型时，如何分配可用的数据资源，包括训练集、验证集和测试集，以便在训练过程中更好地利用各类数据，提升模型性能。大型预训练语料库通常由来自 k 个领域的数据混合组成，其中领域通常由来源定义。跨域的权重由域权重$\alpha \in \triangle k$指定，其中包含 k 维概率向量。下图展示了当前流行的大语言模型所用训练数据中各来源数据的比例。例如，Pile 数据集由 24% Web 数据、9% Wikipedia、4% GitHub 领域数据组成。领域权重的选择可能会导致下游准确性的显着差异。给定固定的训练预算，数据配比方法优化域权重以提高训练效率和模型性能。



常见的文本数据配比方法包括：

（1）基于经验法则和统计信息。现有的工作中，基于经验法则或手动确定领域数据的配比是常见的方法。一些工作根据各领域本身的规模定义领域数据的权重，对所有的单个数据点赋予相同的权重。此外，当前较为流行的模型 LLAMA 等同样使用基于经验的启发式规则来调整训练数据的领域配比，以经验导向的高质量数据集（例如书籍和维基百科文章等）作为参考，对数据集的领域配比进行重加权。

（2）基于评测指标的分数。通过领域权重在模型下游任务上的性能进行数据配比同样是一种思路。用于训练 Gopher 模型的数据集通过测试 7 种领域权重组合的下游任务表现，筛选出性能最优的数据配比权重。GaLM 数据集的配比权重取决于 2 个方面，轻量级模型的下游性能与各领域本身的相对规模，该种方法可以避免小规模领域数据的过采样。此处轻量级模型的参数规模需要能够支撑具有一定意义的下游任务（约 1B 参数量级）。

（3）基于代理模型。目前基于代理模型进行数据配比的研究较为初步，大多数工作仅使用简单的线性模型进行拟合。DoReMi 提出了最小化最大优化的领域数据重加权方法，训练出一个轻量级的代理模型，可以在下游任务未知的情况下对训练数据的权重配比进行调整。DoGE 则将数据配比优化划分为 2 个阶段：首先训练代理模型以使用双层优化算法获得域权重，然后通过根据学习到的域权重对训练域进行采样来训练更大的基础模型。DSDM 则将数据配比视为一个典型的优化问题：给定目标任务、学习算法和候选数据，选择出最大化模型性能的子集，通过训练线性代理模型来得到相对最优的数据权重。

意义

通过适当的数据配比，模型在训练过程中可以接触到更丰富、更多样的数据，这将有助于提高模型的泛化能力和性能，使模型能够在更广泛的数据分布下表现出良好的效果。同时在许多实际应用中，数据往往存在严重的类别不平衡问题，即某些类别的数据过多，而某些类别的数据过少。适当的数据配比可以帮助缓解这个问题，使模型能够更好地学习到少数类别的特性，从而提高模型在所有类别上的性能。此外适当的数据配比

可以使模型在训练过程中更有效地利用数据，避免模型在训练过程中过度关注某些特定类型的数据，从而提高训练效率。

参考文章链接：

[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[2] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In Thirty-seventh Conference on Neural Informa tion Processing Systems, 2023a.

[3] Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training, 2023a.

[4] Fan, S., Pagliardini, M., & Jaggi, M. (2023). Doge: Domain reweighting with generalization estimation. arXiv preprint arXiv:2310.15393.

[5] Engstrom L, Feldmann A, Madry A. DsDm: Model-Aware Dataset Selection with Data models[J]. arXiv preprint arXiv:2401.12926, 2024.

[6] Dong G, Yuan H, Lu K, et al. How abilities in large language models are affected by supervised fine-tuning data composition[J] . arXiv preprint arXiv:2310.05492, 2023.

[7] Lu K, Zhao K, Liang Z, et al. VersaTune: Fine-Tuning Multi-Ability LLMs Efficiently[J]. arXiv preprint arXiv:2411.11266, 2024.

本题作业请发到 keer.lu@stu.pku.edu.cn

4. 文本 reasoning 数据的合成，选择与评估

文本推理（Reasoning）数据的合成、选择与评估是自然语言处理（NLP）中非常重要的一环，尤其是在大型语言模型（如 GPT-o1 系列）进行训练时。

1. 文本推理数据的合成

文本推理数据的合成通常涉及通过特定的策略或自动化方法生成一组问题和推理过程的示例，这些示例可以用来训练和评估模型的推理能力。对于 GPT-1 这种基于自回归的模型，合成数据主要依赖模型辅助生成：利用 GPT-1 或其他预训练模型生成推理数据。

通过提供一段输入文本并提示模型生成推理过程（如“根据上面的内容，推断出...”），可以得到问题和相应的推理过程。

2. 数据选择

数据选择是确保合成数据具有高质量和多样性的重要步骤。数据选择不仅关注数据的数量，还关注数据的代表性和多样性，具体方法包括：

- **问题难度筛选**：选择具有不同难度层级的问题。简单的问题可能对于模型来说过于容易，难度过高的问题则可能导致模型过拟合。通过精细的难度调节，确保训练数据覆盖从基础到高级的推理任务。
- **推理类型多样性**：选择涵盖多种推理类型的数据。推理任务不仅仅限于数学问题，还包括因果推理、归纳推理、类比推理等。通过多样化的推理任务，确保模型能够处理不同类型的推理挑战。
- **文本背景的多样性**：确保所选数据中，问题背景的来源多样。例如，不仅限于文学文本，还可以包括科学、历史、哲学等领域的内容。这有助于模型处理不同领域的推理问题。

3. 数据评估

评估文本推理数据的质量，尤其是在训练和微调过程中，能够帮助确保模型推理能力的提升。评估的标准可以包括：

- **准确性（Accuracy）**：通过比较模型生成的推理结果与人工标签或标准答案来评估模型的准确性。对于推理任务，准确性通常以是否能够正确推理出答案为衡量标准。
- **推理过程的连贯性**：评估模型是否能够在多个步骤中保持推理的连贯性。尤其是在长推理链或多步骤推理中，保持思路的清晰和连贯性非常重要。
- **生成的多样性**：生成的推理过程是否多样化。一个好的推理模型不仅能给出正确答案，还能够给出不同的推理路径，展示多种可能的推理方式。
- **逻辑一致性**：评估生成的推理是否符合逻辑。对于 GPT-1 这样的语言模型，推理过程的逻辑一致性至关重要，尤其是在面对复杂推理问题时。
- **人工评估**：尽管有很多自动化评估指标，人工评估依然是推理任务中不可忽视的部分。通过专家对模型生成推理过程进行评分，可以从更高层次评估模型的推理能力。

本题作业请发到 2301111449@stu.pku.edu.cn

5. 多模态数据合成

多模态数据合成是指通过自动化方法生成包含不同类型数据（如图像、文本、语音和结

构化数据）的数据集，用于多模态模型的训练和推理。这种技术帮助模型更好地理解和融合不同模态的信息，提高其对复杂任务的处理能力。

1. 图像与文本的合成

图像与文本的合成常见于图像描述生成和视觉问答等任务。通过生成图像的文本描述（如“一个女孩在公园里玩耍”），或根据文本生成对应的图像（如文本到图像生成）提升跨模态理解和生成能力。

2. 语音与文本的合成

语音与文本的合成用于语音识别（将语音转化为文本）和文本到语音生成（TTS）。这些技术广泛应用于语音助手、语音交互系统等，增强模型的语音理解与生成能力。

3. 图像与结构化数据的合成

将图像与结构化数据（如表格）结合，用于医疗影像分析、财务报告分析等任务，帮助模型理解图像内容与表格信息的关联。表格生成图像也常用于数据可视化。

4. 多模态综合合成

多模态数据合成不仅涉及单一模态的生成，还包括多个模态的联合生成。例如，生成包含图像、文本、语音和表格的综合数据集，提升模型在多模态任务中的表现。

参考文章链接：

A Survey of Multimodal Large Language Model from A Data-centric Perspective

Infinity-MM: Scaling Multimodal Performance with Large-Scale and High-Quality Instruction Data

Listen, Think, and Understand

Video Instruction Tuning With Synthetic Data

可以自由选择方向：图像，视频，语音，Table 等结构化数据的多模态

本题作业请发到 2301111449@stu.pku.edu.cn

6. 多模态数据选择

多模态数据选择是指在处理多模态任务（如图像、文本、音频、视频等数据联合处理）时，如何高效地从大量数据中选择适合的子集或样本，尤其是在训练大型多模态模型（如多模态大语言模型，Multimodal LLMs）时尤为重要。合理的数据选择不仅能提高模型

的训练效率，还能提高模型的泛化能力和精度。

在多模态任务中，数据选择包括文本、图像、视频、语音等不同模态的数据样本选择，其挑战主要体现在如何平衡和协调不同模态的数据，以及如何有效地从中挑选出与任务最相关的部分。以下是多模态数据选择的几个关键方面：

1. 模态之间的对齐

在多模态任务中，数据往往是跨多个模态的（例如图像和文本、视频和语音等）。数据选择的第一步通常是确保这些数据的配对关系（例如图像与描述文本，视频与音频）。配对的数据需要保证模态之间的对齐和同步，常见的配对方式包括：

- **图像-文本配对**：例如图像描述任务，选择与图像内容相关的描述性文本。
- **视频-文本配对**：视频中的语音内容与文字注释或字幕进行配对，确保视频内容和文本描述相对应。
- **语音-文本配对**：将语音内容转换为文本，然后确保该文本能够清晰表达语音内容。
- **多模态输入对齐**：确保模型可以同时处理多个模态数据，并保持其对齐性（如音频-视频-文本在同一时间段内的同步）。

2. 数据选择的策略

在实际应用中，不同的任务可能需要不同的数据选择策略，常见的策略包括：

- **基于任务需求的选择**：根据具体的多模态任务要求（如图像分类、视觉问答、视频推理等）选择适当的数据。例如，视觉问答（VQA）任务可能需要选择包含图像和相应问题/答案的配对数据集。
- **样本多样性选择**：确保所选的数据样本覆盖了多种不同的情况和模式，以提高模型的泛化能力。例如，在图像描述任务中，可以选择不同背景、不同场景、不同类型的物体等，以提高模型在多样化图像上的表现。
- **基于数据质量的选择**：选择那些清晰、标注准确的数据样本。低质量的数据（如模糊图像、噪音音频或错误标注的文本）可能会影响模型的训练效果，因此需要过滤掉这些低质量的数据。
- **基于标签分布的选择**：如果任务是分类任务，可能需要根据标签的分布情况选择样本。例如，选择少数类样本进行增强，以缓解类别不平衡问题。
- **主动学习**：在训练过程中，主动学习算法可以通过模型对未标注数据的不确定性来选择最有价值的训练样本。通过这种方式，模型可以逐步选择难度较高或更具挑战性的样本来增强其学习效果。

参考文章链接：

A Survey of Multimodal Large Language Model from A Data-centric Perspective

Your Vision-Language Model Itself Is a Strong Filter: Towards High-Quality Instruction Tuning with Data Selection

LESS IS MORE: HIGH-VALUE DATA SELECTION FOR VISUAL INSTRUCTION TUNING

可以自由选择方向：图像，视频，语音，Table 等结构化数据的多模态

本题作业请发到 2301111449@stu.pku.edu.cn

7. RAG 相关-数据检索增强生成

RAG（Retrieval-Augmented Generation）结合了信息检索与生成任务，增强了生成模型的能力。通过在生成过程中引入外部知识库，RAG 可以显著提高复杂任务的处理能力。其工作流程通常包括一个**检索模块**（从文档数据库中检索相关信息）和一个**生成模块**（基于检索到的信息生成答案或文本）。RAG有两个主要版本：**RAG-Sequence**和**RAG-Token**，分别在序列生成和单步生成中处理检索信息。

1. 数据管理、抽取与清洗在 RAG 中的作用

- **数据存储与检索**：RAG 依赖高效的知识库和索引机制，如使用 FAISS 进行向量检索，确保信息快速准确地检索。
- **数据选择与抽取**：根据任务需要抽取相关信息，进行有效的上下文切割和语义提取。例如，在图像描述任务中，选择与图像最相关的文本段落。
- **数据清洗**：去噪、去重、标准化和缺失值处理是数据清洗的核心步骤，确保输入数据的一致性和质量。
- **数据增强**：通过重写文本、替换同义词等方法扩展训练数据，提升模型泛化能力。

2. RAG 中的数据抽取

- **基于语义的检索**：利用文本的嵌入表示（如 BERT、Sentence-BERT）进行语义检索，比传统关键词匹配更能捕捉深层次的语义相关性。
- **信息抽取（IE）**：从文档中提取关键信息（如命名实体、事件等），为生成任务提供更准确的上下文支持。
- **多模态数据抽取**：在处理图像、视频等多模态数据时，采用跨模态检索和对齐策略，**增强多模态输入的一致性和相关性。**

3. RAFT（Retrieval-Augmented Fine-Tuning）

RAFT 是一种结合 RAG 与模型微调的技术，旨在通过在检索过程中引入外部信息，并利用检索结果对模型进行进一步微调，从而提高模型的推理能力。RAFT 利用检索的额外信息来“指导”模型学习，优化了生成过程中的知识融合，特别适合需要长期推理和复杂背景知识的任务。RAFT 能通过不断微调模型，提升在特定领域或任务上的生成效果。

参考文章链接：

Retrieval-Augmented Generation for AI-Generated Content: A Survey

RAFT: Adapting Language Model to Domain Specific RAG

DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation

RAG-Studio: Towards In-Domain Adaptation of Retrieval Augmented Generation Through Self-Alignment

可以自由选择 RAG 相关方向

本题作业请发到 2301111449@stu.pku.edu.cn

8. 数据解析

非结构化数据指没有预定义模型的数据类型，包括文本、图像、音频和视频等。许多历史记录、学术出版物和法律文件仍以扫描或图像格式存在，这对信息提取、文档理解和检索增强等任务构成了显著挑战。解析非结构化数据，尤其是文档数据（如 PDF），并将其转换为结构化格式（如 LaTeX 或 Markdown），是数据解析的关键环节。这一过程重塑了信息在各类应用中的存储、共享和应用方式，为下游流程提供了基础支持，如**强推理多模态大模型的训练语料和检索增强生成（RAG）系统的开发。**此外，文档中隐藏的大量潜在信息尚未被充分开发。数据解析技术能够有效提取和组织这些丰富的知识，为下一代智能系统的发展奠定基础，例如训练更专业、更强大的多模态模型。目前，数据解析包括对于非结构化数据中**复杂元素的针对性解析识别**和利用**多模态模型的通用端到端解析方案。**

1.科学表达符号解析与提取

近期，随着以 O1 为代表的强推理模型在数学、代码等科学问题上取得的突破性进

展。人们开始关注到如何从广泛但无结构的科学文章中提取结构化的信息，用于训练增强模型的推理能力。公式、化学分子式等科学表达符号是其中的一类代表，因为他们将抽象、生涩的科学概念形式化为一组人类/模型可理解的表征（如 MathML 或 LaTeX）。然而，科学符号的提取并非易事，因为其编码侧面临打印体与手写体、图像退化、变形等挑战，解码侧则要求模型具有数学符号识别、结构推理和语义解析等能力。尽管已有诸多研究贡献，仍存在数据短缺和复杂数据识别精度不高等问题。

参考文章链接：

[1] Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction

[2] Generating Handwritten Mathematical Expressions From Symbol Graphs: An End-to-End Pipeline

[3] MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation

[4] Visual Parsing with Query-Driven Global Graph Attention (QD-GGA): Preliminary Results for Handwritten Math Formula Recognition

[5] UniMERNet: A Universal Network for Real-World Mathematical Expression Recognition

2. 数据可视化元素的解析与提取

表格、图表等数据可视化元素是文档（如报告、学术论文、财务报表和技术文档）中的重要信息载体，呈现数据的结构化和连贯性，帮助快速理解数据间的关系和层次。这些元素的解析涉及从非结构化数据中识别并提取表格信息，包括识别结构、数据等等，并将其转化为结构化数据格式（如 JSON，LaTeX、HTML 或 Markdown）。目前，可视化元素解析面临的挑战包括复杂表格/图表结构的解析（如合并单元格或嵌套表格）、噪声和不规则格式的处理，以及跨语言和跨文档格式的一致性解析。

参考文章链接：

[1] Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction

[2]Chimera: Improving Generalist Model with Domain-Specific Experts

[3]Detection and Recognition of Table structures from Unstructured Documents

[4] OmniParser: A Unified Framework for Text Spotting, Key Information

Extraction and Table Recognition

3. 通用文档数据解析

传统文档解析工具往往面临如下问题：

- 每个算法模块（布局检测、公式识别、表格识别，...）需要算法人员专门优化升级，成本高
- 长尾需求多，拓展成本较高（手写识别、研报表格、PPT 文档，...）

但是由于 MLLM 的飞速发展，例如 GPT-4o、Qwen2-VL、DeepSeek-V2.5 这些通用的模型均在 OCR 等任务上取得了优异的成绩，越来越多的目光，例如 DocOwl2、Vary 等，开始聚焦于基于多模态大模型方案构建高质量文档解析的端到端模型，将传统算法模块在自回归的范式下进行统一，更方便的对接各种长尾需求。

参考文章链接：

[1] General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model

[2] Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models

[3] DOGE: Towards Versatile Visual Document Grounding and Referring

[4] mPLUG-DocOwl2:High-resolution Compressing for OCR-free Multi-page Document Understanding

[5] Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs

[6] OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation

本作业请发送到：🔗 zhangqt21@gmail.com

9. 数据压缩

在深度学习领域，大规模数据在不同场景中广泛存在。由于其数据量庞大，在处理时往往会产生高昂的存储开销和运行开销。为了解决这一问题，数据压缩是一项至关重要的技术，它能够将庞大而复杂的数据转化为精简的压缩数据，从而满足个人设备、商用设备等资源受限环境下的部署需求。针对 Graph 和 CV 领域，列举两种经典的数据压缩范式：图数据选择和图数据集蒸馏。图数据选择往往无需模型训练，通常使用图算法保留大图的核心结构进行压缩。图数据集蒸馏通常包含图像数据(Image)和图数据（Graph）的压缩，通常需要采用模型训练的方式生成压缩数据。此外，随着大型语言模型（LLMs）

在各个领域的广泛应用，其训练和部署的效率问题日益凸显。文本压缩成为了现在被广泛关注的问题。

1. 图数据选择

图数据选择可以细分为三种方法：图稀疏化（Graph Sparsification），图粗粒度化（Graph Coarsening），图核心集（Graph Coresets）。图稀疏化作为一种直观的数据压缩方法，涉及根据特定的标准选择基本边或节点，然后从这些选定的元素构建一个更小的图。传统方法通常专注于保留特定的图属性，例如频谱和中心性。相比图稀疏化，图粗粒度化为了确保保留了足够数量的信息，这些技术涉及对节点进行分组和聚合。核心集方法试图从原始数据集中找到有信息的样本，保证在这个小子集上做学习任务，就能得到原数据集的一个近似最优解。

参考文章链接：

[1] A Comprehensive Survey on Graph Reduction: Sparsification, Coarsening, and Condensation. IJCAI'2024

[2] A Survey on Graph Neural Networks for Graph Summarization. Arxiv'2023

2. 图数据集蒸馏

数据集蒸馏（dataset distillation, dataset condensation）是一种通过人工合成少量样本的技术，使得在这些小样本上训练的模型能达到与在原始大规模数据集上训练相当的性能。与从数据集中选取子集的方法不同，数据集蒸馏利用模型的中继信息，如梯度和输出，创造新的具有代表性的样本。数据集蒸馏算法主要关注 3 个方面：首先是有效性，它必须保证合成数据集能够精确捕捉并反映原始数据的核心特征；其次是泛化能力，样本训练出的模型能够适应不同的下游任务和模型架构；最后是高效性，一个高效的数据集蒸馏方法能够将大规模数据快速压缩，极大提升其在实际应用中的可行性和经济性。另外，此技术可以扩展到其他各种数据类型，包括 image、text 和 audio 等，为数据密集型应用带来巨大的潜在价值和应用前景。

参考文章链接：

[1] A Comprehensive Survey of Dataset Distillation. TPAMI'23.

[2] Graph Condensation: A Survey. ArXiv'24.

3. 文本压缩

随着大型语言模型（LLMs）在各个领域的广泛应用，其训练和部署的效率问题日益凸显。比如常用的大模型 Bert 消耗计算资源和时间随着 token 的长度是平方级别增长的，所以其无法处理太长的 token，目前最长只支持 512 个 token，token 过长也很容易会内

存溢出，所以在使用常用的大模型如 Bert 在长文本处理一般分为三种方法：截断法、Pooling 法、压缩法。此外，传统的子词标记化方法，如 Byte Pair Encoding (BPE) 或 Unigram，虽然能够在一定程度上压缩文本，但其压缩率通常只有 4 倍左右。LLMs 在处理文本时，需要消耗大量的计算资源来处理这些标记。为了解决该问题，可以采用神经压缩等方法进行优化。

参考文章链接：

[1] CogLTX: Applying BERT to Long Texts. NIPS'2020

[2]Training LLMs over Neurally Compressed Text. Arxiv'2024

本题作业请发到 liangyx@stu.pku.edu.cn

10.数据标注

数据标注在机器学习和深度学习中占据着至关重要的地位。高质量的标注数据通常是构建强大模型的基础。然而，传统的人工标注过程既耗时又昂贵，尤其是在数据量庞大的情况下。大语言模型(LLM, Large Language Models)作为一个通用的预训练语言模型，其通过对海量文本的学习，能够捕捉语言中的丰富模式和结构信息，从而在处理自然语言理解和生成任务时表现出色。随着 LLM 的快速发展，研究者逐渐探索将其应用于数据标注的可行性，旨在通过其自动生成标注或辅助进行人工标注的能力，减轻人工标注的负担，降低标注成本。

挑战

尽管 LLM 在自动标注中具有巨大潜力，但在实际应用过程中仍面临诸多挑战：

（1）标注质量不稳定：LLM 生成的标注在大多数情况下存在噪声和不一致性，尤其是在面对复杂或指意不明的标注任务时。LLM 本身的随机性更加剧了这一挑战。

（2）标注偏向性：LLM 在训练过程中可能会吸收并放大训练数据中的偏见，导致其生成的标注结果具有偏向性。例如，LLM 在处理涉及敏感信息的标注任务时，可能会无意中展现出性别、种族等方面的偏差。

（3）专业领域标注质量不佳：在某些专业领域或特定任务中，LLM 可能无法充分理解领域特定的术语或语境，从而导致标注结果的准确性下降。

（4）多模态数据标注：除了文本标注，许多应用场景涉及到多模态数据的标注，如图像、视频等。LLM 主要在文本生成和理解方面表现突出，如何结合多模态信息进行有效标注仍是一个亟待解决的问题。

总之，尽管 LLM 在数据标注中展现出巨大的潜力，但仍面临诸多挑战。未来的研究需要

进一步优化标注质量、尽量避免标准偏差、增强模型的专门领域适应能力，并探索更加高效的多模态标注方法。

参考文章链接：

Tan, Zhen, et al. "Large Language Models for Data Annotation and Synthesis: A Survey." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

本题作业请发到 shengzeang18@pku.edu.cn

11.大语言模型与图数据结合

图学习（Graph Learning）作为一种研究图结构数据的方法，广泛应用于社交网络、推荐系统、知识图谱等领域。图神经网络（GNNs, Graph Neural Networks）作为图学习中的核心模型，能够有效地挖掘图结构数据中的节点和边之间的关系，已在许多实际应用中取得了良好的效果。然而，尽管 GNNs 在图数据的处理上展现出了强大的能力，如何进一步利用大语言模型（LLMs，Large Language Models）对图学习任务进行补充和提升，仍然是一个重要且富有挑战性的研究课题。

挑战

（1）异质性和数据融合问题：LLMs 通常处理的是序列数据（如文本、句子等），而图数据则具有节点、边以及复杂的图结构。想要有效结合 LLM 的文本理解能力与 GNN 的图结构理解能力，就需要解决如何将异质化的信息（语言特征与图结构特征）良好融合的问题。如何在保证图结构信息的同时，有效引入 LLM 的语义知识，是一个亟待解决的挑战。

（2）规模与效率问题：LLMs 通常是大规模的预训练语言模型，训练和推理成本非常高。因此，将 LLMs 与 GNN 结合时，如何兼顾计算效率与模型效果，是一个重要挑战。尤其在图学习任务中，图中的节点数目可能非常庞大，如何在大规模图数据中高效地整合 LLM 的信息，避免过度计算，是一个更难解决的问题。

参考文章链接：

[1] Ren, Xubin, et al. "A survey of large language models for graphs." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024.

[2] Jin, Bowen, et al. "Large language models on graphs: A comprehensive

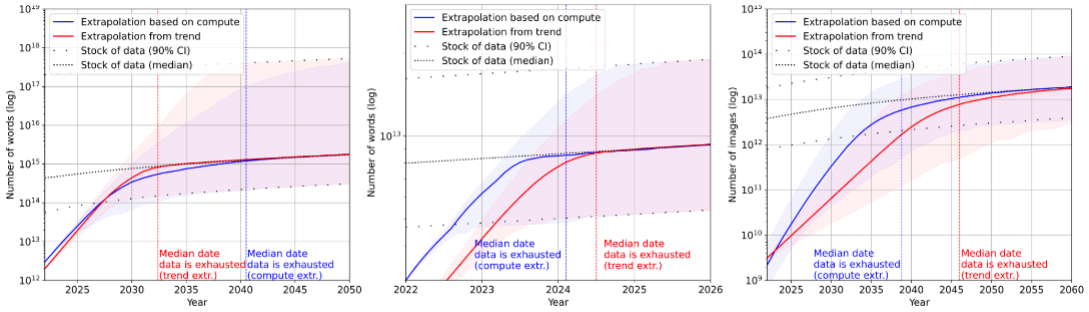
survey." IEEE Transactions on Knowledge and Data Engineering (2024).

[3] Fan, Wenqi, et al. "Graph machine learning in the era of large language models (llms)." arXiv preprint arXiv:2404.14928 (2024).

本题作业请发到 shengzeang18@pku.edu.cn

12. 合成数据与 Agent 智能体

机器学习的发展依赖于大规模、高质量的数据集，但数据稀缺、隐私保护和高成本等问题使得获取这些数据面临挑战。有研究表明，高质量文本数据可能在 2026 年耗尽，低质量文本和视觉数据也将在 2050 年和 2060 年面临枯竭，这将严重阻碍机器学习的进步。



(a) Projections for low-quality language data

(b) Projections for high-quality language data

(c) Projections for vision data

合成数据成为应对这一挑战的关键策略。通过算法生成的数据能够模仿现实世界数据的特征，以低成本填补稀缺领域的数据缺口，提升模型的泛化能力和性能。此外，合成数据还能缓解隐私问题，增强数据安全性，并融入模型的自我迭代过程，推动模型的持续优化。因此，合成数据有望克服现实数据的局限，推动大型模型的快速发展。

LLM Agent 是基于大型语言模型的智能代理，具备对话、任务执行和推理能力。它们通过规划、记忆、工具使用和行动模块来处理复杂任务，并能自我学习和改进。这些智能体能够适应不同的工作模式，广泛应用于多个领域，预示着未来人机交互和生活方式的重大变革。

1.基于 Agent 的数据生成与增强

通过基于智能体（Agent）的数据生成与增强技术，能够显著提升数据的质量、多样性和复杂性，从而为模型训练和优化提供更强大的支持。智能体可以通过模拟专家行为或探索虚拟环境，生成高度逼真、覆盖多场景的数据，同时模拟复杂的动态系统和多维度交互，生成具有高层次逻辑的数据。此外，这种技术能够以经济高效的方式大规模生成

数据，减少对人工标注和真实数据收集的依赖，并在保护隐私的前提下满足训练需求。基于智能体的数据生成与增强为人工智能模型的训练开辟了新的路径，使其能够在更广泛的应用场景中实现高效学习和优化。

包括但不限于

- 基于小批量的种子数据生成大量的合成数据
- 基于原有数据对其进行数据增强，如
  - 提高数据的复杂度
  - 对原有数据的格式进行 reformat
  - 删除原有数据中的冗余或噪声

参考文章链接：

Synthesizing Post-Training Data for LLMs through Multi-Agent Simulation

Star-Agents: Automatic Data Optimization with LLM Agents for Instruction Tuning

https://github.com/hyp1231/awesome-llm-powered-agent

https://github.com/pengr/LLM-Synthetic-Data

2.基于合成数据提升 Agent 的通用或特定能力

近年来，关于智能体训练（Agent Training）的研究取得了显著进展，研究表明，利用专家级轨迹数据（即行动-观察对的序列）对大型语言模型（LLM）进行指令微调（instruction tuning），能够显著提升其能力。然而，现有研究大多局限于从人工设计的任务和环境中合成轨迹数据。这类方法虽然有效，但其局限性在于创建这些任务和环境需要耗费大量的人力、物力和时间成本，导致生成的轨迹数据在多样性和覆盖范围上难以满足智能体训练的需求。

因此，探索通过合成数据替代真实人工数据来为智能体提供训练支持，成为一种更具经济性和可扩展性的解决方案。这种方法不仅能够降低对人工标注和真实数据收集的依赖，还能通过大规模生成多样化数据，进一步提升智能体在通用或特定任务上的表现。合成数据的引入为智能体训练开辟了新的路径，使其能够在更广泛的应用场景中实现高效学习和优化。

- 通用能力包括但不限于
  - 推理能力，决策能力，规划能力，记忆能力，人机交互能力，对话能力，工具使用能力（function call）
- 特定能力主要指的是在垂直领域上的能力或针对特定任务的能力，包括但不限于

- 医疗诊断，推荐系统，数学与代码任务等等

参考文章链接：

AgentGen: Enhancing Planning Abilities for Large Language Model based Agent via Environment and Task Generation

https://github.com/hyp1231/awesome-llm-powered-agent

https://github.com/pengr/LLM-Synthetic-Data

本题作业请发到 youquan.li@stu.pku.edu.cn

13. AI for Science 中的数据科学

1. 蛋白质设计

蛋白质设计（Protein Design）是蛋白质科学领域的重要问题。从天然蛋白质序列和结构的分布中学习的深度生成模型，可能使设计具有有价值功能的新蛋白质成为可能。尽管当今大多数模型专注于生成序列或结构，但新兴的联合生成方法有望实现更准确和可控的蛋白质设计，理想情况下通过同时建模这两种模态来实现。深度生成模型在蛋白质设计中有不少最新进展，其中包括序列方法、结构方法以及序列-结构联合生成方法等。这些方法背后的关键方法论和评估原则不尽相同，各种文献里讨论了不同蛋白质设计、生成方法在未来的发展机遇。

完成本题时，既可以从蛋白质设计这一大领域入手，讨论整个领域的发展历程与状况，也可以聚焦某一部分内容，比如蛋白质设计领域里的数据设计与增强方法等。

2. 多模态和统一建模

AI for Science (AI4S) 领域涉及多种类型和模态的数据，包括蛋白质序列、三维结构、核酸序列、配体小分子、晶格排列数据以及文本数据（如生物医学文献）。这些多模态数据为 AI 模型提供了丰富的输入，使其能够更全面地理解和预测生物系统的行为。例如，AlphaFold3 和 RoseTTAFold-All Atom 等模型突破了传统仅对蛋白质建模的局限，能够从全原子角度对蛋白质、核酸、配体小分子等生物分子进行联合建模，并预测它们之间的相互作用。这些进展不仅推动了蛋白质结构预测的精度，还为药物设计、分子相互作用研究和合成生物学提供了新工具。

然而，AI4S 领域仍面临诸多挑战。首先，**数据异质性和融合问题**：不同模态的数据（如序列、结构、晶格排列、文本）具有不同的表示形式和语义，如何有效地融合这些异质

数据以提升模型性能是一个关键问题，当然如何生成多模态的数据（比如序列和二维、三维结构）也是值得研究的方向。其次，**统一建模的探索**：不同分子类型（如蛋白质、核酸、小分子）甚至不同种类（如有机分子、无机材料）的建模能否通过一个统一的框架实现，使模型能够同时学习到这些不同分子的表征，是一个极具潜力的研究方向。

完成本题时，可以选择相关领域的某个或某些问题，对最新的文献工作展开调研。

参考文章链接：

蛋白质序列和结构联合生成 Survey: 🔗 https://arxiv.org/pdf/2410.01773

蛋白质领域的跨模态大模型： 🔗 https://arxiv.org/pdf/2403.07920

蛋白质多模态建模：🔗 https://arxiv.org/abs/2410.13782

蛋白质设计：🔗 https://github.com/Peldom/papers_for_protein_design_using_DL

本题作业请发到 ytgeng@pku.edu.cn

14. 数据科学在本学科中的应用（限本科生）

数据科学是一门跨学科的领域，本身便结合了统计学，计算机科学的方法技术。其应用也几乎涵盖了所有学科。如在医学中，数据科学可以通过分析病人的基因组数据，提供定制化的治疗方案。在生物信息学中，可以处理解释生物分子数据。在经济学中，可以使用机器学习算法预测市场趋势。在社会科学中，可以进行社会网络分析。

在数据量较少，可解释要求高，即时性要求高等场景下，传统机器学习算法仍有其优势。如果你对数据科学在本专业上的应用感兴趣，可以选择调研数据科学在本学科的前沿应用(不局限于传统机器学习算法)。可以选择数据科学在本学科内部的某一类应用作为课题。

本题作业请发到 libozhou@pku.edu.cn

15. 数据科学的历史脉络与未来发展 （限本科生）

数据科学是一个跨学科领域，它结合了统计学、信息科学、计算机科学以及领域专业知识，旨在从数据中提取知识和洞见。随着大数据时代的到来，数据科学已经成为推动科学发现、商业决策和社会变革的关键力量。了解数据科学的历史发展对于把握其

未来趋势至关重要。

本课题旨在让同学们通过文献调研，探索数据科学的起源、发展和演变过程。通过对数据科学历史的深入研究，同学们将能够理解数据科学如何塑造现代社会，并预测其未来可能的发展方向。

本题作业请发到 zhaoks@stu.pku.edu.cn