语言模型训练数据

数据来源

- 为了保证模型使用更多高质量数据进行训练，在GPT-3 训练时，根据语料来源的不同，设置不同的采样权重。在完成3000 亿词元训练时，英文Wikipedia 的语料平均训练轮数为3.4 次，而CommonCrawl 和Books 2 仅有0.44 次和0.43 次
- 大语言模型训练所需的数据来源大体上可以分为通用数据和专业数据两大类
  - 通用数据包括网页、图书、新闻、对话文本等内容。通用数据具有规模大、多样性和易获取等特点，因此可以支持大语言模型的**构建语言建模和泛化能力**
  - 专业数据包括多语言数据、科学数据、代码以及领域特有资料等数据。通过在预训练阶段引入专业数据可以有效提供大语言模型的**任务解决能力**

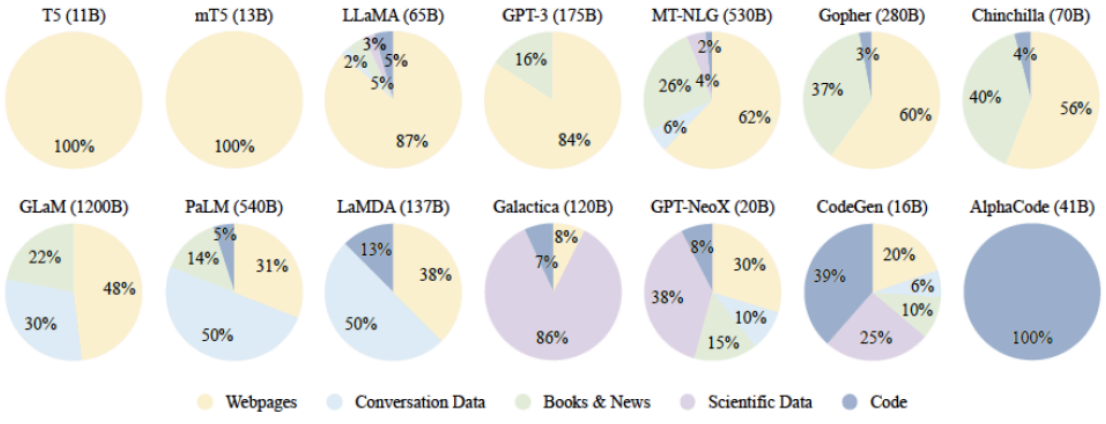

图 3.1　典型大语言模型所使用数量类型的分布[18]

数据处理

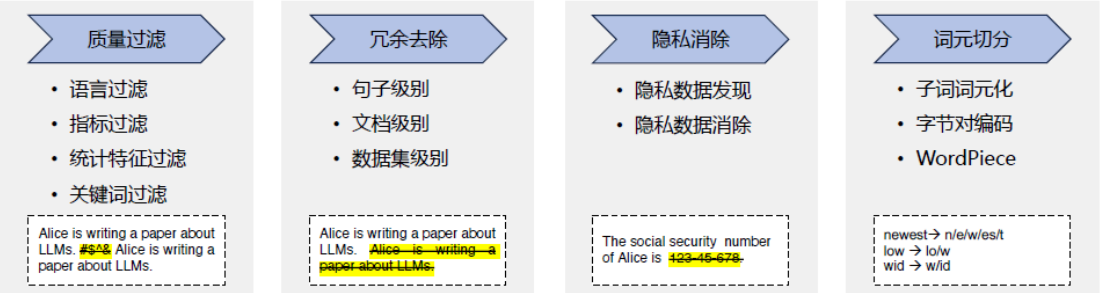

图 3.2　典型大语言模型数据处理流程图[18]

数据影响分析

- 研究发现，如果模型训练要达到计算最优（Compute-optimal），模型大小和训练词元数量应该等比例缩放，即模型大小加倍则训练词元数量也应该加倍(chinchilla)
- 通过使用更多的数据和更长的训练时间，较小的模型也可以实现良好的性能。
- 语言模型在经过清洗的高质量数据上训练数据可以得到更高的性能。
- 高质量数据对自然语言生成任务上的影响大于在自然语言理解任务

开源数据集

- Pile
- ROOTS
- RefinedWeb
- SlimPajama