Market Research and Analysis

Lecture 4: Large Language Model Annotation

- Zhenyu Zhao
- Nankai Institute of International Economics
- Feishu: 2120253538
- Email: zzynankai@outlook.com
- Website: xishanyu2.github.io

01

大语言模型简史

> 人工智能是人类智慧的镜像，也是未来的伙伴。——艾伦·图灵

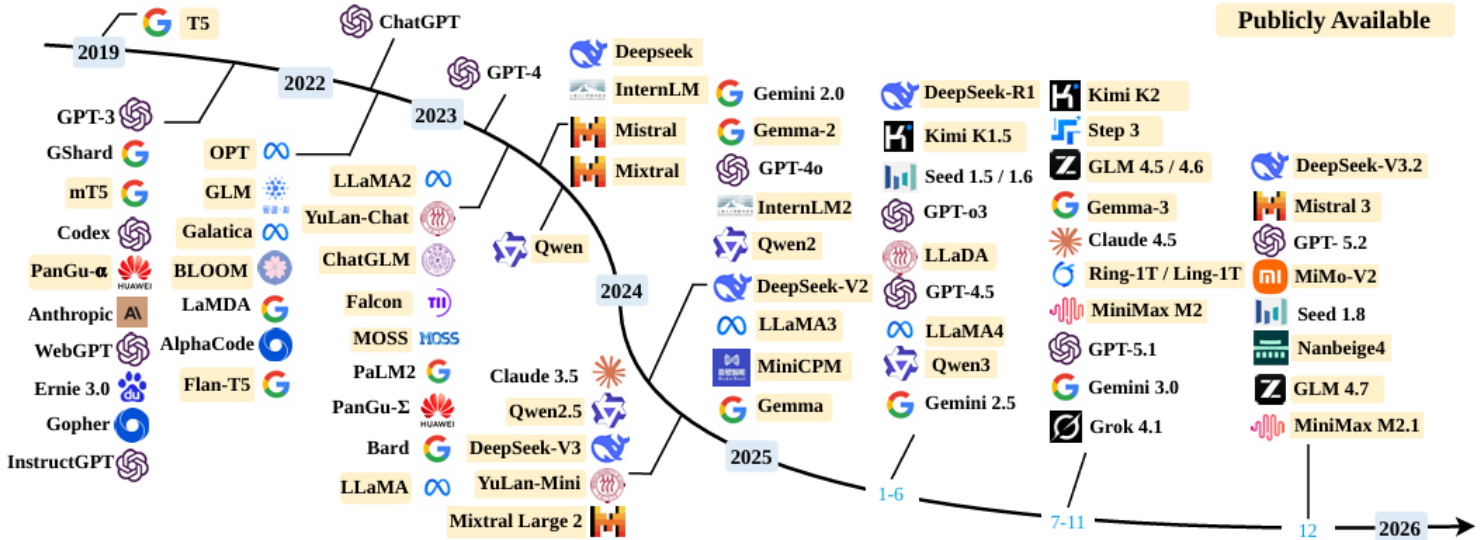


Fig. 3: A timeline of representative LLMs released in recent years. Models with publicly available checkpoints are highlighted in yellow.

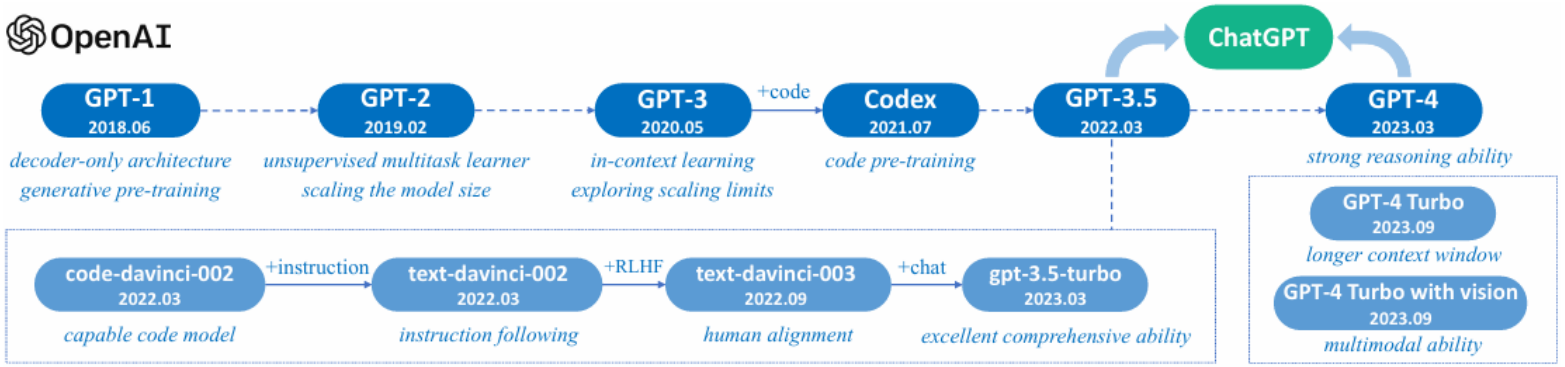
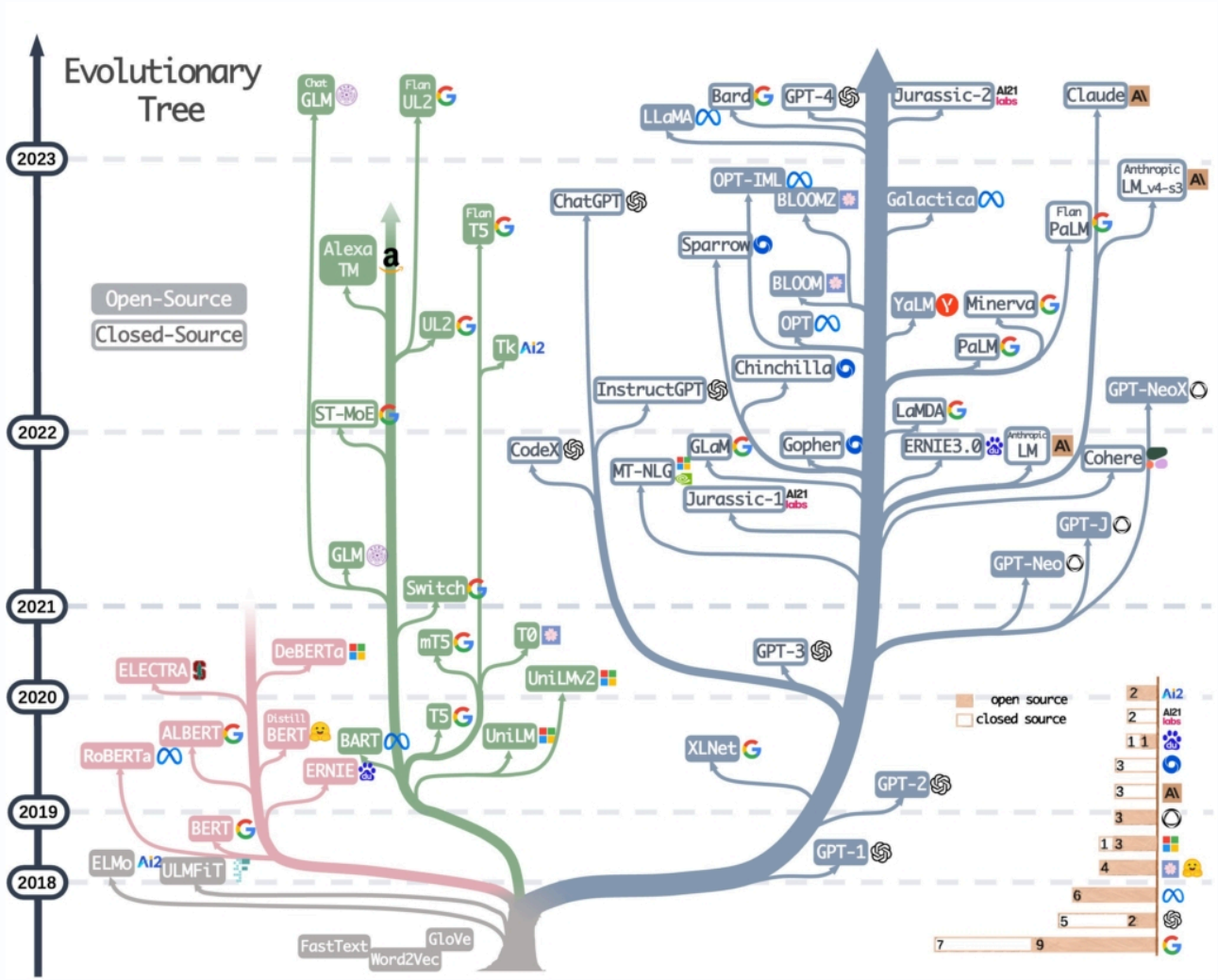
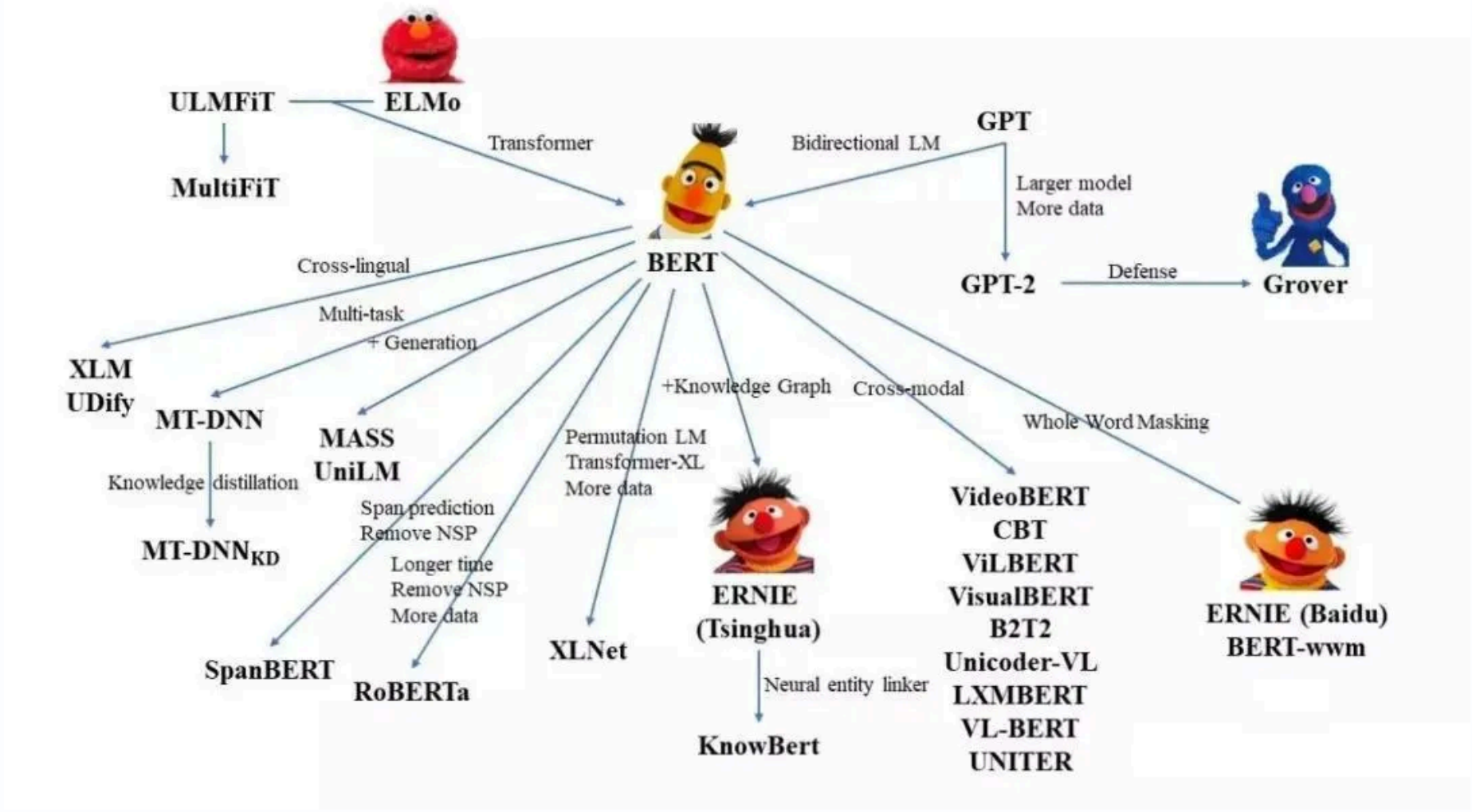


Fig. 4: A brief illustration for the technical evolution of GPT-series models. We plot this figure mainly based on the papers, blog articles and official APIs from OpenAI. Here, *solid lines* denote that there exists an explicit evidence (*e.g.*, the official statement that a new model is developed based on a base model) on the evolution path between two models, while *dashed*


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Bard
GPT-4
Jurassic-2
Claude
LLaMA
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
ChatGPT
Flan T5
Alexa TM
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
UL2
Tk
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
ST-MoE
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GLM
GPT-Neo
Switch
mT5
T0
DeBERTa
GPT-3
ELECTRA
UniLMv2
T5
Distill BERT
ALBERT
BART
UniLM
RoBERTa
ERNIE
XLNet
GPT-2
BERT
GPT-1
ELMo
ULMFiT
FastText
GloVe
WordVec
open source
closed source


ULMFiT
ELMo
MultiFiT
Transformer
GPT
Bidirectional LM
Larger model
More data
BERT
GPT-2
Defense
Grover
Cross-lingual
Multi-task
+ Generation
XLM
UDify
MT-DNN
MASS
UniLM
Knowledge distillation
MT-DNN$_{KD}$
+Knowledge Graph
Cross-modal
Whole Word Masking
Permutation LM
Transformer-XL
More data
Span prediction
Remove NSP
Longer time
Remove NSP
More data
SpanBERT
RoBERTa
XLNet
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMBERT
VL-BERT
UNITER
ERNIE (Baidu)
BERT-wwm

02

大语言模型数据标注

在经济金融与会计等社会科学研究领域，研究者普遍面临着海量、高度复杂的非结构化文本数据（如政策文件、公司报告、ESG披露、新闻与访谈等）的分析挑战。传统依赖人工标注、固定词典或经典机器学习的方法，往往成本高、效率低，且难以适应大规模、动态变化的文本分析需求。

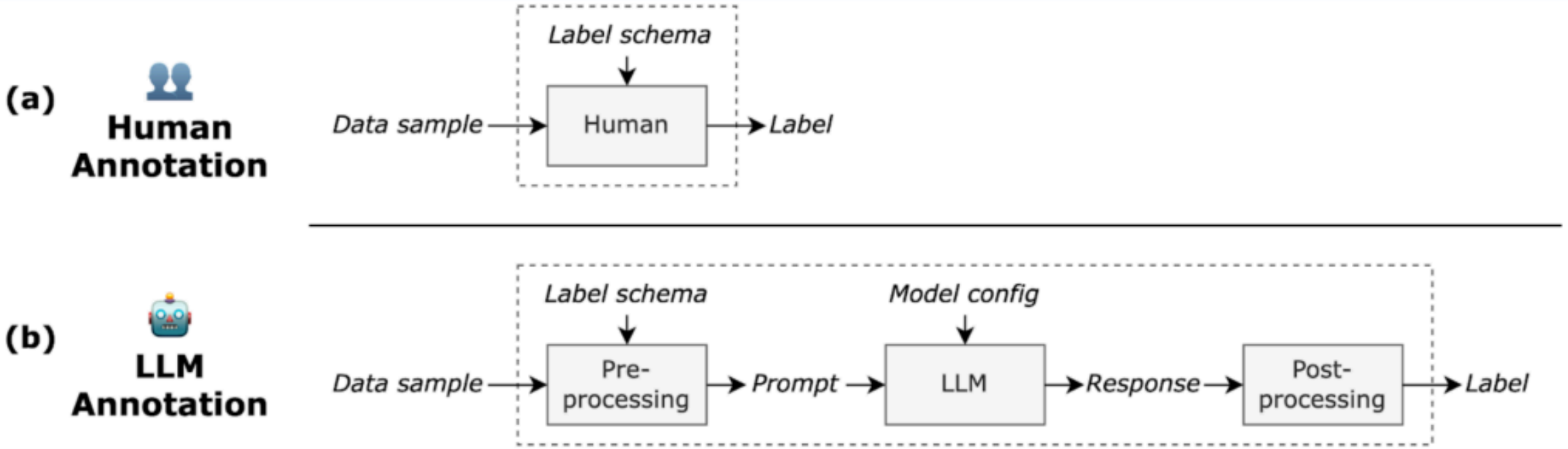
传统的词频分析方法根据关键词词典统计出年报文本中含有的关键词数量，如果含有关键词则认为能够反映出企业具备相应能力，而关键词的数量（或关键词的占比）则能反映出能力的大小；由于缺乏对关键词语境的分析会造成构建的指标有偏，例如含有关键词的句子是表达否定或对未来规划展望的含义都会被识别成为肯定含义。词典法存在诸多问题：

- 一是容易挂一漏万（词选少了），长期来看一些新词语、新提法的出现势必要求研究人员及时更新词库，人力成本较大；
- 二是容易词不达意（词选多了），选取的关键词不能很好地反映真实情况造成冗余和误判；
- 另外词典选词涉及专家评估，主观评价标准不同会造成词典间难以统一。

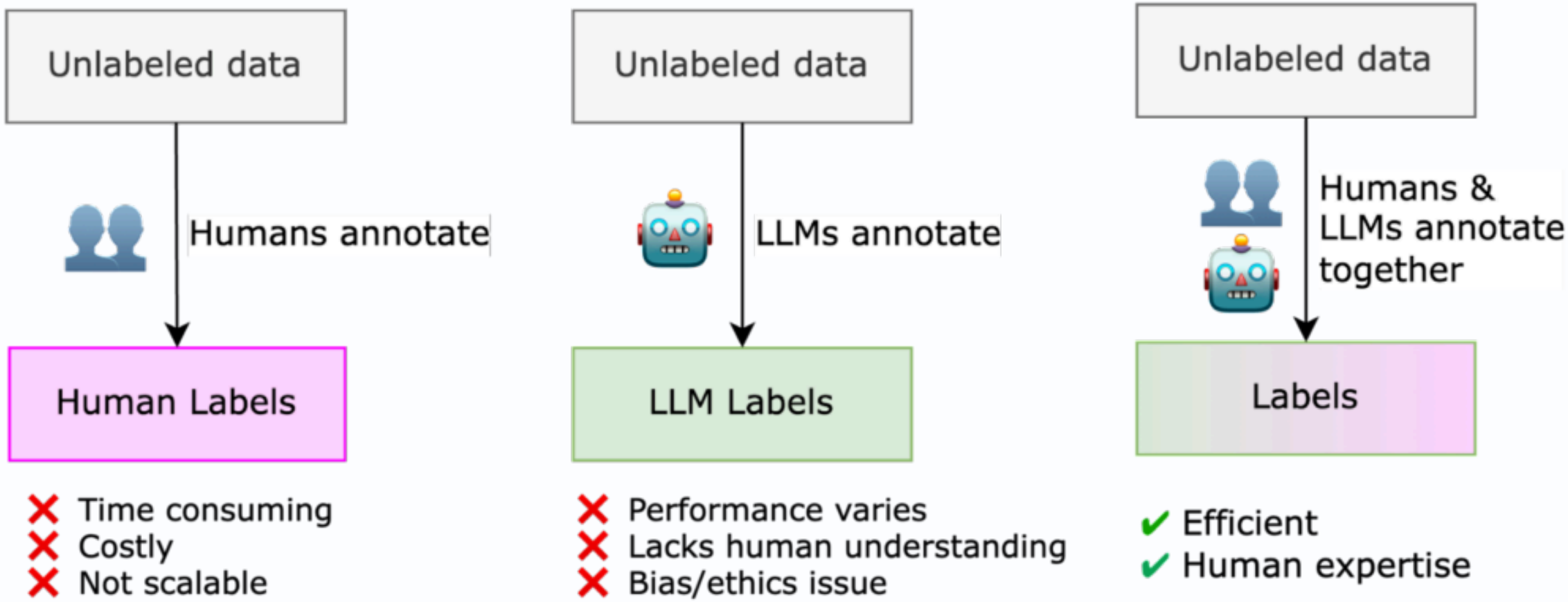
已有文献采取人工标注的方式，对含有关键词的句子打标签判断是否真正地体现了相应的能力（减小第二类错误），以及对不含关键词的句子打标签判断是否潜在地体现了相应的能力（减小第一类错误），将最终生成的“句子-标签”作为训练集，减少由于简单词频统计造成的偏误。具体步骤如下：在提取出句子的基础上进行人工阅读分类，对于子样本数据由一组两个人分别打一遍标签，如果两个人打的标签相同则认为判定准确，如果不同则需要在所有组内进行进一步讨论最终确定标签。

人工标注同样面临着诸多困难与挑战，这种做法虽然能准确地处理文本细节得到高质量的数据，但由于费时费力需要投入高昂的人力、物力、财力成本，标注环节可能成为研究瓶颈拖慢整体进度；此外，两人相同的标注也不一定完全正确，这部分标注数据也会进入训练集。

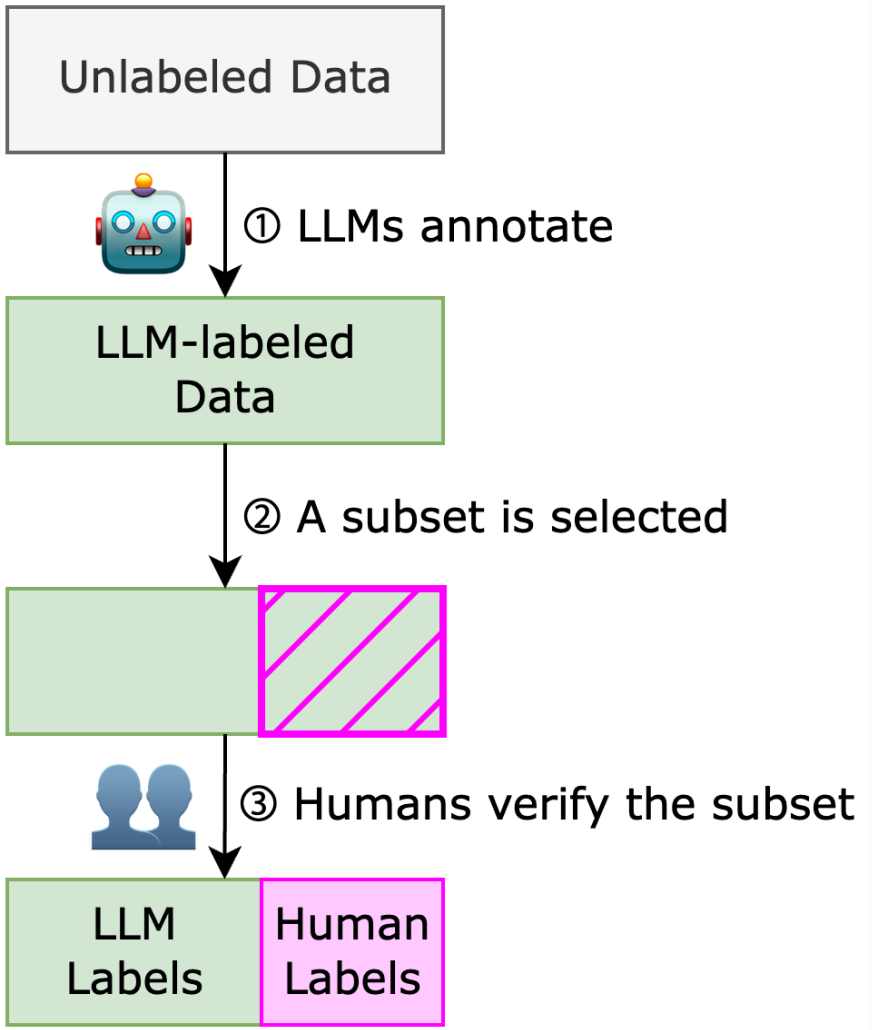
近年来随着大语言模型的更新与迭代升级，使用大语言模型进行文本标签生成与特征构造成为可能。而基于少量样本提示的大语言模型，则为上述问题提供了创新性的解决路径：它能够以显著降低的成本自动化完成大量文本处理任务，并允许研究者根据研究目标灵活定义和调整标签体系，极大增强了文本分析的适应性与可扩展性。已有研究实践印证了这一方向的可行性。


(a)
Human Annotation
Label schema
Data sample
Human
Label
(b)
LLM Annotation
Label schema
Model config
Data sample
Pre-processing
Prompt
LLM
Response
Post-processing
Label

进行大语言模型标注的具体工作流程如下：首先需要进行标注任务与大语言模型提示词的初步制定，为人机双方提供清晰的说明，确保任务定义明确统一；然后抽取小规模随机样本由人工与大语言模型并行标注，通过对比两者在样本上的差异，分析大语言模型输出的推理依据，进而持续优化分类定义与示例；在此基础上，不断修订代码与提示词，并循环进行样本标注与差异比较，直到模型的标注性能达到预期标准；最终，使用一个更大规模的验证集进行最终测试，并记录准确率、F1值等指标，从而完成从任务定义到结果评估的全流程可控、可迭代的自动化标注体系建设。


Unlabeled data
Humans annotate
Human Labels
❌ Time consuming
❌ Costly
❌ Not scalable
Unlabeled data
LLMs annotate
LLM Labels
❌ Performance varies
❌ Lacks human understanding
❌ Bias/ethics issue
Unlabeled data
Humans &
LLMs annotate
together
Labels
✔ Efficient
✔ Human expertise

使用大语言模型也可能存在一定的潜在弊端与风险，如何确保标注结果可信、应对幻觉与系统性偏差，是利用大语言模型标注的核心问题。若仅依赖大语言模型标注会造成结论失真：大语言模型的偏见、事实不准确、不一致都有可能导致误导性研究结果，因此目前还不能将所有工作都交给大语言模型，需要将人类标注与大语言模型标注衔接起来，构建统计意义上有效的分析基础。


Unlabeled Data
① LLMs annotate
LLM-labeled Data
② A subset is selected
③ Humans verify the subset
LLM Labels
Human Labels

03

大语言模型在经济学中的研究进展

近年来，大语言模型（LLM）在经济学研究中的应用正深刻改变传统文本分析范式。传统方法（如关键词词典法）在测度复杂经济概念时面临语义理解局限和上下文信息缺失的挑战，而LLM凭借其强大的语义表征能力和上下文推理技术，为构建精准的经济指标提供了新路径。

[1]金星晔,左从江,方明月,等. 企业数字化转型的测度难题：基于大语言模型的新方法与新发现[J].经济研究,2024,59(03):34-53.

[2]李涛,涂玮,李兵,等. 关税冲击与企业数字技术应用[J].经济研究,2024,59(12):130-148.

[3]陆瑶,施函青,周欣怡. 中国企业数字技术风险暴露对企业价值的影响——来自大语言模型的文本分析证据[J].经济研究,2025,60(02):73-89.

[4]陈运森,高毅达,于耀,等. 公共实施机制与私人实施机制的监管联动——基于信号博弈和大语言模型的证据[J].经济研究,2025,60(07):142-161.

[5]谢佳松,樊嘉诚,林建浩. 信息摩擦、公共数据与企业创新边界[J].经济研究,2025,60(11):51-73.

先导性文章：刘青和肖柏高（2023）

> 刘青,肖柏高.劳动力成本与劳动节约型技术创新——来自AI语言模型和专利文本的证据[J].经济研究,2023,58(02):74-90.

- **关键词**：劳动节约型技术；BERT模型；劳动力成本；最低工资

1.简介

摘要：从替代体力的蒸汽机到替代脑力的AI，重大劳动节约型技术的创新和应用都深度改变了世界经济发展格局，并成为学界研究焦点。然而梳理文献发现，现有关于劳动节约型技术的研究主要聚焦于其应用或引进，对其创新关注甚少。本文首次结合AI领域前沿的**BERT语言模型**和专利摘要较为精准地识别劳动节约型技术创新。在此基础上，考虑到中国正处于老龄化加剧、劳动力成本快速上涨阶段且劳动节约型前沿技术引进日益受阻，检验劳动力成本上升如何影响中国企业的劳动节约型技术自主创新。基于上市公司数据，本文发现劳动力成本上升显著地促进了劳动节约型技术创新的**概率、数量、占比以及质量**。使用考虑中国省界特征的相邻城市样本来解决潜在内生性问题后结论依然成立。本文提出的特定技术创新的识别方法可复制推广，在技术创新相关研究领域有较大应用前景。本文研究结论对于理解我国在人口红利消失、老龄化加剧、劳动力成本快速上升阶段的特定类型技术创新有重要启示。

2.指标构建

概念界定：

劳动节约型技术(labor saving technology，又称劳动替代型技术)是指能代替或节约劳动力使用的技术。自工业革命以来，技术换人一直都是备受关注的话题，其热度经久不衰的原因在于，一方面，劳动节约型技术的应用对于经济增长和劳动力市场有重大影响，如生产率、就业结构、劳动收入份额、劳动收入差距、老龄化等；另一方面，劳动节约型技术在不断创新，从替代体力的蒸汽机到替代脑力的人工智能(AI)，技术对劳动力的替代不断泛化和深化，从而推动劳动节约型技术持续地成为社会大众和学术研究的焦点。

数据选择：

阻碍劳动节约型技术创新相关研究一个可能的重要原因是数据缺失。劳动节约型技术的应用指标已经不少，例如机器、电脑、机器人等的存量或安装指标。但是梳理文献可知，目前尚未有良好的劳动节约型技术创新的直接度量指标，这使得高质量的实证研究难以开展，而理论研究也可能会因为缺少典型事实或关键参数支持而难以持续深化。基于现实与学术背景，作者引入人工智能领域前沿的BERT语言模型，**基于专利摘要语义识别以节约劳动为目的或者客观上可以节约劳动力的专利技术，构建劳动节约型技术创新的直接度量指标**，并在此基础上研究劳动力成本上升对我国劳动节约型技术创新的影响。（1）**模型训练数据**：从1998-2015年间900余万发明、新型实用专利中随机抽取了5000条。（2）**待预测数据**：2002-2016年，中国沪深上市公司的所有发明和实用新型专利。

指标构建过程：

（1）**构建训练数据**。从1998-2015年间900余万发明、新型实用专利中随机抽取了5000条，通过**逐一阅读的方式进行人工分类**，作为BERT模型的训练基础。人工分类的标准是：1. 专利的目的或者功效是节约劳动力的使用。如果某专利摘要中有“节省或替代劳动力(人力、人工)”等词句，则说明此专利的功能和目的之一就是节约劳动力，这符合劳动节约型技术的**直接定义**，故将其判定为劳动节约型专利。2.专利摘要中描述了一种可以(部分)替代劳动力使用的自动化技术或设备。专利的主体可以全自动完成某项任务，或者它可以部分替代人类任务，即便专利摘要中没有明确指出它可以减少劳动力使用，但是从实际效果而言，它还是会对劳动力产生替代效应。从**间接意义**上来说，它也符合劳动节约型技术的定义。

（2）**模型微调和预测**。作者指出经济学界常用**word2vec模型**来完成文本表征任务，而**BERT在表征文本长度、准确度等方面有明显提升**。作者在人工分类的专利样本中，随机抽**4000条作为训练集，500条作为开发集，500条作为测试集**，对BERT模型进行微调，用微调后的BERT模型对上市公司的所有发明和实用新型专利进行分类。微调和预测过程如后图所示。

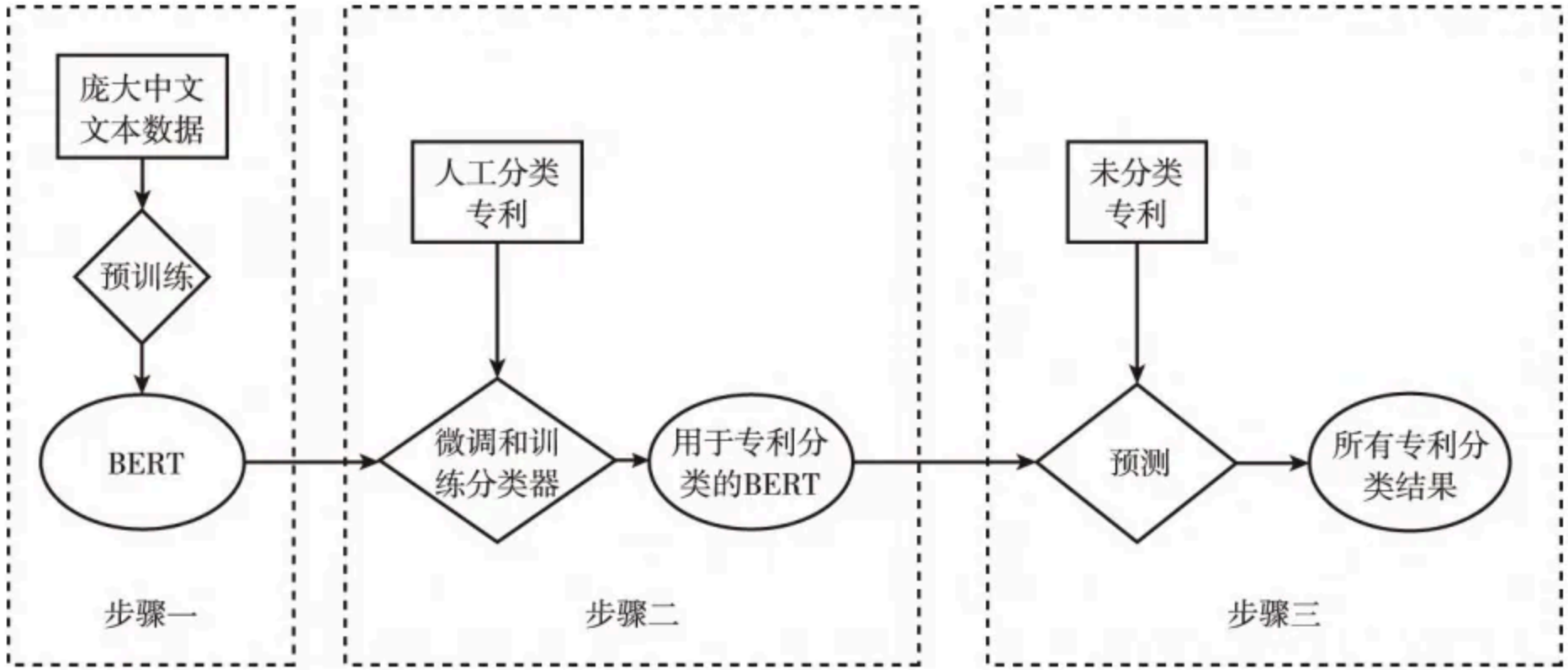


图 1　使用 BERT 进行专利文本分类任务主要步骤

（3）**核心指标**。基于上述过程，作者得到上市公司 “劳动节约型专利数量”、 “非劳动节约型专利数量” 。

开创性文章：金星晔等（2024.3）

> 金星晔,左从江,方明月,等.企业数字化转型的测度难题：基于大语言模型的新方法与新发现[J].经济研究,2024,59(03):34-53.

- **关键词**：企业数字化转型；数字经济；数字技术；人工智能；大语言模型

1.简介

摘要：社会各界关于企业数字化转型的重要性已经基本达成共识，但对企业数字化转型的效果存在严重分歧。产生这一现象的主要原因是现有研究对企业数字化转型的测度存在问题：一是测度对象不够统一明确，二是测度方法不够科学准确。这导致很多研究结论不可比较、难以复制和相互冲突。为了更好地处理上述难题，本文运用**机器学习和大语言模型**构造一套新的企业数字化转型指标。本文首先对2006-2020年上市公司年报中的句子进行**人工打标签**，然后用标记结果**训练和微调**包括大语言模型在内的多个机器学习模型，选择其中分类效果最好的**ERNIE模型**作为句子分类模型来预测全部文本中句子的标签，最终构造了企业数字化转型指标。理论分析和数据交叉验证均表明，本文构建的指标相对已有方法更准确。在此基础上，本文实证检验**企业数字化转型对财务绩效**的影响。研究发现：第一，企业数字化转型能够显著提高财务绩效，其中，大数据、人工智能、移动互联、云计算和物联网均有明显作用，但区块链并没有明显的作用；第二，只有在财务绩效较差的企业中，数字化转型才能够显著提高财务绩效；第三，企业数字化转型提高财务绩效的主要渠道包括改善效率和降低成本。本文研究对于推动企业数字化转型和实现经济高质量发展具有一定现实意义。

2.指标构建

概念界定：

对于企业来说，数字经济和实体经济的融合过程本质上是一个数字化转型过程。所谓数字化转型(digital transformation)，是企业借助数字技术来改造企业的生产经营系统、管理模式和核心业务流程,从而形成一个破坏性创新和变革的过程。

测度方法评估：

现有文献在测度企业数字化转型程度时，通常使用了三种方法：第一种方法是**客观数据法**，例如，计算本企业与数字技术相关的软件投资或硬件投资占总资产的比重；第二种方法是**事件冲击法**，利用企业所属群体是否受到数字化转型政策的冲击来度量企业的数字化转型情况。第三种方法是**词典法**（最主流）先构建一个包含各种数字技术的关键词词典，然后根据这些**关键词在上市公司年报中"管理层讨论与分析（MD&A）"部分出现的次数或比例**，构建企业数字化转型指标。方法背后隐含的假设是，提及了某种数字技术的关键词就表明企业进行了数字化转型。因此，某个上市公司年报中提及数字技术的次数或者比例越高，表示企业的数字化转型程度越高。使用词典法进行文本分析虽然操作比较简便，但是这种传统方法不能够充分提取文本中含有的信息，导致分析的准确度、指标的有效性较低。近年来，自然语言处理（NLP）技术广泛应用于于机器翻译、舆情监测、自动摘要、观点提取、文本分类等领域，能够弥补词典法的不足，因此，作者采用"基于大语言模型的新测度方法"测算企业数字化转型程度。

数据选择：

由于数字化转型涉及企业组织结构、内部管理、业务流程等方方面面的变革，难以在财务指标中完整显示，但上市公司有强烈的意愿在年报中披露以获得资本市场的青睐，因此文献中通常基于年报进行文本分析来衡量数字化转型水平。借鉴已有文献的做法，论文作者采用上市公司年报作为企业数字化转型指标的文本基础。作者通过爬虫程序和人工收集方法，收集2006-2020年间披露的上市公司年报，并选择年报中“管理层讨论与分析”和 “目录、释义及重大风险提示”这两个章节作为文本数据，最终得到2006-2020年4181家公司的39175份年报文本。

指标构建过程：

（1）**词典构建**。作者将数字技术分为六种类型：大数据、人工智能、移动互联、云计算、物联网和区块链。然后，基于政策文本、研究报告和已有文献，并通过人工阅读之后的不断补充，收集整理了一个包含了311个数字技术关键词的词典。

（2）**构建句库**。将全部文本按照句号和分号分割，得到待预测句库。由于年报中大多数句子与数字化转型无关，需要**使用关键词抽取具有不同代表性的年报句子**，并与随机抽取的句子一起构成待标记句库。利用构建的词典，作者抽取了包含10个及以上不同关键词的年报文本，并取出其中含有关键词的句子。此外，为了提高模型对不含关键词句子的预测能力，随机抽取部分年报，并分割为句子。由于上市公司总数逐年增加，存在大量被标注句子靠近当前年份，为了解决**年份分布不均匀**的问题，在上述两部分句子的基础上，按照年份分组，在每个年份中取出相同数量的句子，再从这部分均匀分布的句子中进行不放回的随机抽取，得到本研究的待标记句库。最终，**待标记句库中包含38994个句子**。

（3）**句子标注**。作者使用人工标注方法，先判断企业使用了哪种/哪几种数字技术，进而判断企业是否进行了数字化转型。人工标注的目的是对上述待标记句库中包含的38994个句子，形成训练集、测试集和验证集，为后面的机器学习打下基础。在正式标注时，待标记句库中的每个句子都会被两位研究成员标记。如果双方标注结果一致，则句子标签被记录；对存在分歧的句子，经过全部成员讨论后确定其标签；对难以确定标签的句子，不纳入训练集。

（4）**模型训练**。让**人工智能技术替代人工**来判别文本中包括数字技术关键词是否意味着企业真正进行了数字化转型（避免虽然文本中提及数字技术的关键词，但实际上并未使用数字技术的难题），需要使用分类模型。作者将全部被标记的句子按照**8:1:1的比例分为训练集、测试集和验证集**。**对比了ERNIE、BERT_base_Chinese、SVM（支持向量机）、Neural Networks（神经网络）、SVM与Neural Networks结合的Voting算法、KNN（K近邻）以及GaussianNB（朴素贝叶斯）共七类常见模型**的训练效果，发现**ERNIE和BERT模型的综合分类能要远高于常见的分类模型**。

（5）**构造数字化转型指标**。基于ERNIE大语言模型，作者对2006~2020年待预测句库中的每一句文本进行预测，判断企业是否以及使用何种数字技术。**构造了企业数字化转型哑变量**，即公司在当年只要使用大数据 、人工智能、移动互联、云计算、区块链和物联网中的任意技术，则指标赋值为1，反之为 0。

一招鲜吃遍天：李涛等（2024.12）

> 李涛,涂玮,李兵,等.关税冲击与企业数字技术应用[J].经济研究,2024,59(12):130-148.

- **关键词**：关税冲击；数字技术；中美贸易摩擦；大语言模型
- **摘要**：本文使用2015-2020年中国上市公司年报文本数据，采用大语言模型方法构建企业层面的数字技术应用程度指标，并结合中国上市公司数据、中国海关数据和关税变化数据构建由中美贸易摩擦带来的企业层面关税冲击指标，进而实证考察关税冲击对中国企业数字技术应用的影响。结果表明，美国加征关税对中国企业数字技术应用具有显著抑制效应，而中国加征关税的效应不显著。本文进一步验证了关税冲击影响企业数字技术应用的供应链机制、竞争压力机制、经济政策不确定性机制以及融资成本机制，并发现关税冲击对数字技术应用的影响随企业的贸易特征、规模以及对美国出口依赖程度等方面的变化而出现异质性。此外，企业的上下游受到美国加征关税冲击也会影响其数字技术应用决策。本文研究拓展了数字技术应用动因的讨论，丰富了国际贸易对企业经营决策影响的研究，对在微观层面理解贸易摩擦的作用渠道和制定应对策略也具有参考价值。

本文采用了金星晔等（2024）的方法，并根据研究需要，通过计算企业MD&A文档中**表达数字化转型句子（Digital_sentence）的数量占该文档所有句子数量的比重**来衡量企业数字技术应用程度（Digital）。

审稿人指出，因为样本期间大部分的企业都已经应用了数字技术，笼统使用衡量企业数字技术应用的虚拟变量并不适合本文，体现不出数字技术应用程度的变化。基于此，本文使用金星晔等（2024）的方法，在识别出文档中表达数字技术应用的句子后，使用这些句子在文档句子数量中的占比测度企业数字技术应用程度。

迅速跟进：陆瑶等（2025.2）

> 陆瑶,施函青,周欣怡.中国企业数字技术风险暴露对企业价值的影响——来自大语言模型的文本分析证据[J].经济研究,2025,60(02):73-89.

- **关键词**：数字技术风险；供应链外溢；大语言模型；企业价值；数据安全

1.简介

摘要：做优做强数字经济，需要统筹发展和安全两个方面，高度重视新一轮数字技术变革中的风险防范问题。**本文利用FinBERT和GPT大语言模型对A股上市公司年报管理层讨论与分析部分进行文本分析，构建多层次数字技术风险暴露程度的衡量指标，探讨我国企业数字技术风险的暴露情况及对企业市场价值的影响**。研究发现:第一，年龄较大、杠杆率较高、资产收益率较低、公司治理水平较差的企业面临更严重的数字技术风险暴露。第二，数字技术风险暴露会显著降低公司的长期市场价值。其中，数据安全风险和网络安全风险均有明显负面影响。在利用美国同行业数字技术风险暴露作为工具变量缓解内生性问题后，本文实证结果保持稳健。第三，主要的影响机制在于数字技术风险通过数字基础设施的互联性，对供应链生产网络产生溢出效应，且给企业带来长久的声誉损失。本文的结论对如何平衡数字化转型中的收益与风险、推进企业持续优化数字化转型路径、护航数字经济长期健康发展具有政策意义。

2.指标构建

概念界定：

“数字技术风险”一词指的是企业在应用数字技术时可能面临的潜在风险，主要包括数据安全风险和网络安全风险。这一风险涵盖了诸如数据泄露、网络攻击、信息安全漏洞等潜在威胁，可能对企业的信息资产、业务连续性和声誉造成严重影响。

数据选择：

论文以**2003-2022年中国A股上市公司**为研究对象，将企业年度报告的管理层讨论与分析部分作为指标构建数据来源。**大量文献证实企业年度报告中的“管理层讨论与分析”（management discussion and analysis，以下简称 MD&A）部分具有客观有效的信息含量，存在风险揭示功能，并能增强财务报告的有用性**。例如，某企业指出：“公司技术革新风险：公司互联网视频业务对互联网的依赖程度较高，运营的安全易受到电讯故障、黑客攻击、病毒等因素的影响。”MD&A信息识别企业风险暴露情况的做法在国内外众多文献中已经得到应用。

指标构建过程：

（1）**抓取数字技术风险相关的文本论述**。先构建出数字技术风险相关的关键词列表（**词典构建**），并将关键词所在的句子提取出来作为后文训练和预测的基础。考虑到企业应用数字技术类型的差异，构建关键词时包括了各种数字技术的具体风险。关键词的选取基础参考 《工业和信息化领域数据安全风险信息报送与共享工作指引（2021）》（试行）和国家互联网信息办公室印发的《国家网络安全事件应急预案（2017）》。

（2）**标注训练集**。从**经过关键词筛选后的句子论述中随机抽取了10%的样本**，进行**GPT和人工标注相结合**的办法，以判断论述是否反映出企业在数字技术风险方面的暴露或采取了相应的防范措施。值得注意的是，作者同时**使用美国OpenAI公司的ChatGPT、国内智谱清言公司的ChatGLM两种大语言模型进行标注，出现矛盾时结合人工复核，以提高标注效率和最终标注的准确性**。作者将涉及因数字技术风险而可能对公司运营产生负面影响的相关论述界定为数字技术安全风险暴露，并同时将反映企业采取的数字技术风险防范措施的论述纳入变量计算过程。

（3）**模型微调和预测识别**。作者将标注的数据转化为适用于**情感分类模型**的训练集，及对于标注结果为“风险暴露”的文本赋值为 **-1，表示“负面”情感**；对于标注结果为“风险防范”的文本赋值为**1，表示“正面”情感**。将标注的样本**按0.6:0.4划分为训练集和验证集**，以此对FinBERT模型进行微调。使用微调后的模型未标记的样本进行预测，得到全部数据的标注结果。

（4）**构建数字技术风险暴露程度指标**。企业数字技术风险的最终暴露程度取决于以下两方面的差值：一是企业在数字技术应用中可能遭遇的最严重风险；二是企业在应对这些风险方面所积累的防范能力。作者将企业数字技术风险暴露（Digi_Risk_Exposure）定义为企业每年MD&A部分涉及数字技术风险的文本中，**负面文本的负面情感概率最大值与正面文本的正面情感概率平均值之间的差值**。

最新研究：结合博弈论与信息经济学

陈运森,高毅达,于耀,等.公共实施机制与私人实施机制的监管联动——基于信号博弈和大语言模型的证据[J].经济研究,2025,60(07):142-161.

- **关键词**：中小股东行权；股东投票；问询函；大语言模型；信号博弈
- **摘要**：推动公共实施机制和私人实施机制的联动协同对于强化中小股东的行权力量和构建监管协同的资本市场至关重要。本文构建信号传递博弈模型，通过手工搜集的股东大会投票数据并结合**FinBERT大语言模型**对理论模型进行检验后发现，在中国这一新兴资本市场中，公共实施机制（交易所监管问询）与私人实施机制（中小股东异议）存在监管联动效应。交易所监管问询能够影响中小股东信念和吸引中小股东参与，激发中小股东的异议投票行为。公司通过回函等信息披露行为能传递分离信号，缓解市场信息不对称和逆向选择问题。进一步分析显示，中小股东无法主动识别机器学习预测的公司被问询风险，凸显了公共监管问询在信息传递和监管协同中的关键作用。事件研究法等检验表明，公共监管问询能够引导中小股东进行信息生产行为。最后，交易所监管问询和中小股东行权的监管联动效应在投资者关系管理较差的公司中更为显著。本文对于全面系统地评判以交易所监管问询为代表的公共实施机制的经济后果提供了新视角，并为2024年国务院“加强监管防范风险推动资本市场高质量发展”的资本市场顶层设计提供了政策启示。

最新研究：结合博弈论与信息经济学

> 谢佳松,樊嘉诚,林建浩.信息摩擦、公共数据与企业创新边界[J].经济研究,2025,60(11):51-73.

- **关键词**：信息摩擦；公共数据；创新边界；数据结构；大语言模型
- **摘要**：在从“创新大国”迈向“创新强国”的战略转型中，如何破解突破式创新不足的结构性困境，已成为提升中国科技竞争力的关键命题。本文从信息摩擦视角出发，构建了一个包含私有数据与公共数据的异质性创新选择模型，刻画企业在渐进式与突破式创新之间的最优决策。理论分析表明，公共数据作为一种公共信息，一方面增强企业对外部创新环境的识别能力（信息供给），另一方面有助于降低企业间创新网络的协同成本（协调强化），从而提高企业研发效率并激励突破式创新，推动创新边界拓展。在此基础上，本文利用中国地方政府公共数据开放平台上线作为影响企业信息环境的冲击，对模型结论进行实证检验，结果与理论预测相符。进一步地，本文**基于大语言模型对企业专利进行技术领域识别**，计算其与地方政府公共数据所属领域的**余弦相似度**，构建技术布局引导变量，并结合企业未来研发投入波动率，共同验证了公共数据在技术引导和不确定性缓解方面的信息供给机制。此外，公共数据开放还显著推动了企业开放式创新，佐证了理论模型中的协调强化机制。本文拓展了信息经济学与创新经济学的交叉研究，为理解数据如何重构微观创新决策提供了理论基础与实证支撑，同时也为充分释放公共数据价值与完善政府数据开放平台建设提供了针对性的政策建议。

04

环境配置

1. Download and install Anaconda: Download Success | Anaconda
2. Open Anaconda Prompt
3. Create a virtual environment:
   ```
   conda create -n my_paddlenlp python=3.10
   conda activate my_paddlenlp
   ```
4. Install paddlepaddle: 开始使用_飞桨-源于产业实践的开源深度学习平台
   CPU version:
   ```
   python -m pip install paddlepaddle==3.3.0 -i https://www.paddlepaddle.org.cn/packages/stable/cpu/
   ```
   GPU version:
   ```
   python -m pip install paddlepaddle-gpu==3.3.1 -i
   https://www.paddlepaddle.org.cn/packages/stable/cu129/
   ```

5. Install paddlenlp: 安装PaddleNLP — PaddleNLP 文档

```
pip install --upgrade --pre paddlenlp
```

6. Change aistudio_sdk.hub's version

```
pip uninstall aistudio-sdk
```

```
pip install aistudio-sdk==0.2.6
```

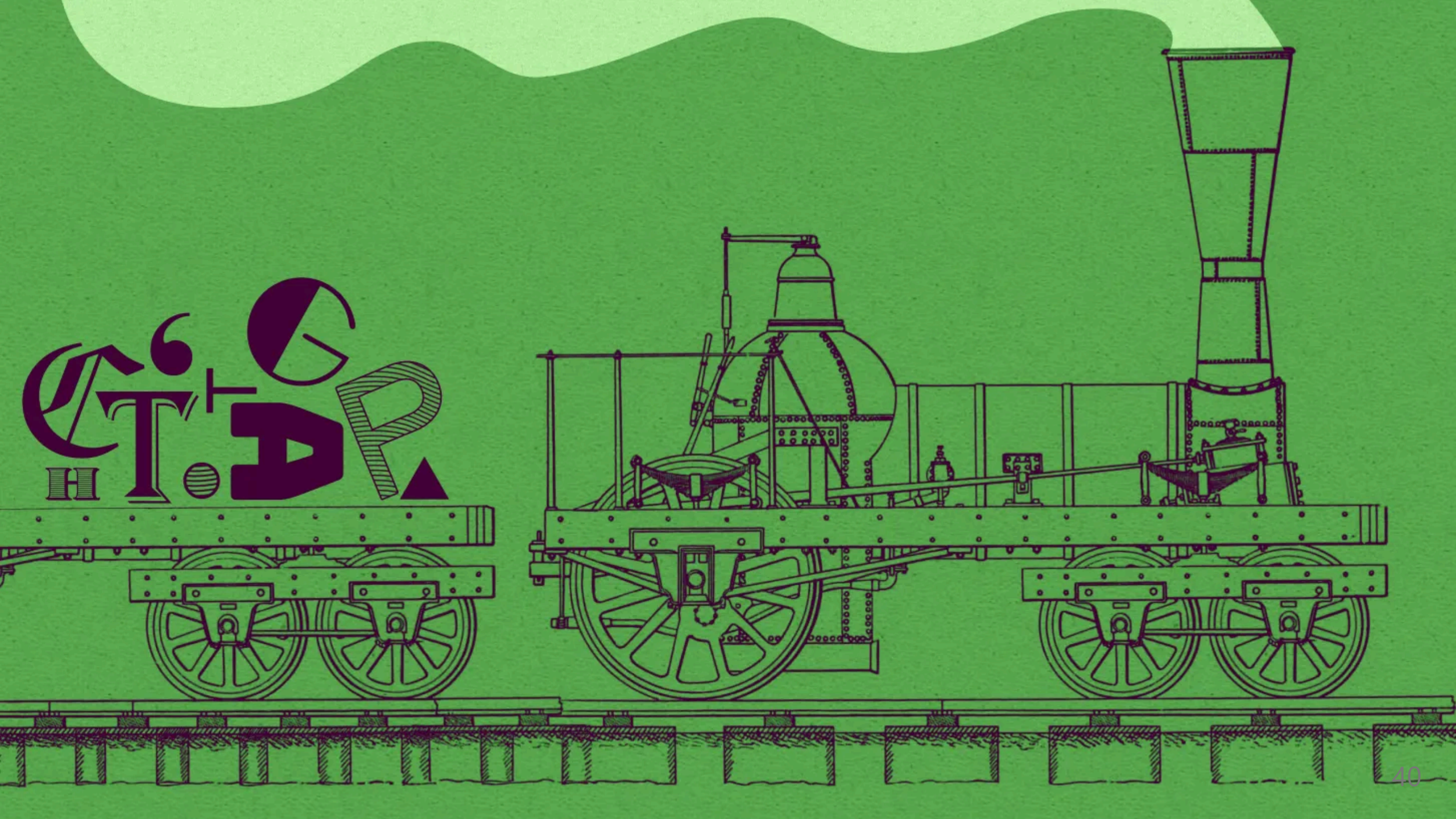
7. Install necessary packages

```
pip install openpyxl
```

```
pip install jupyter
```

8. Open Jupyter Notebook

```
jupyter notebook
```


СТОП