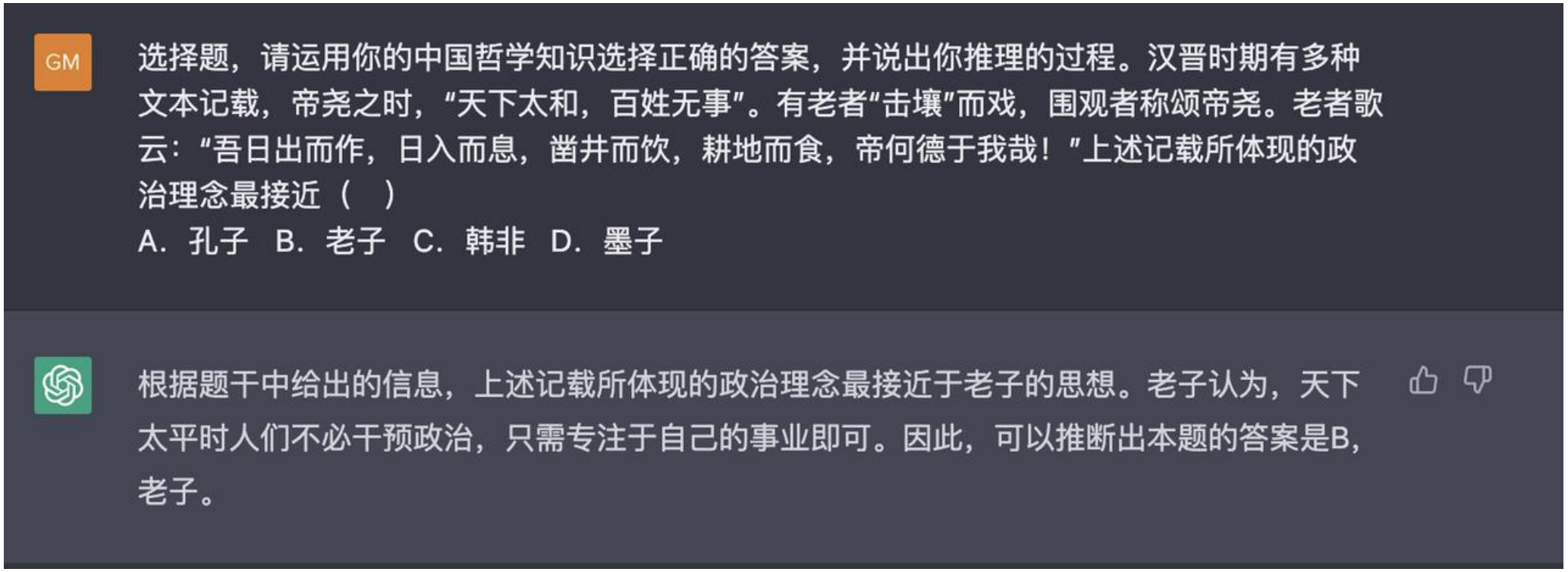
语言模型发展历程

《大语言模型》编写团队：赵鑫

ChatGPT 于2022年11月底上线

AI Box

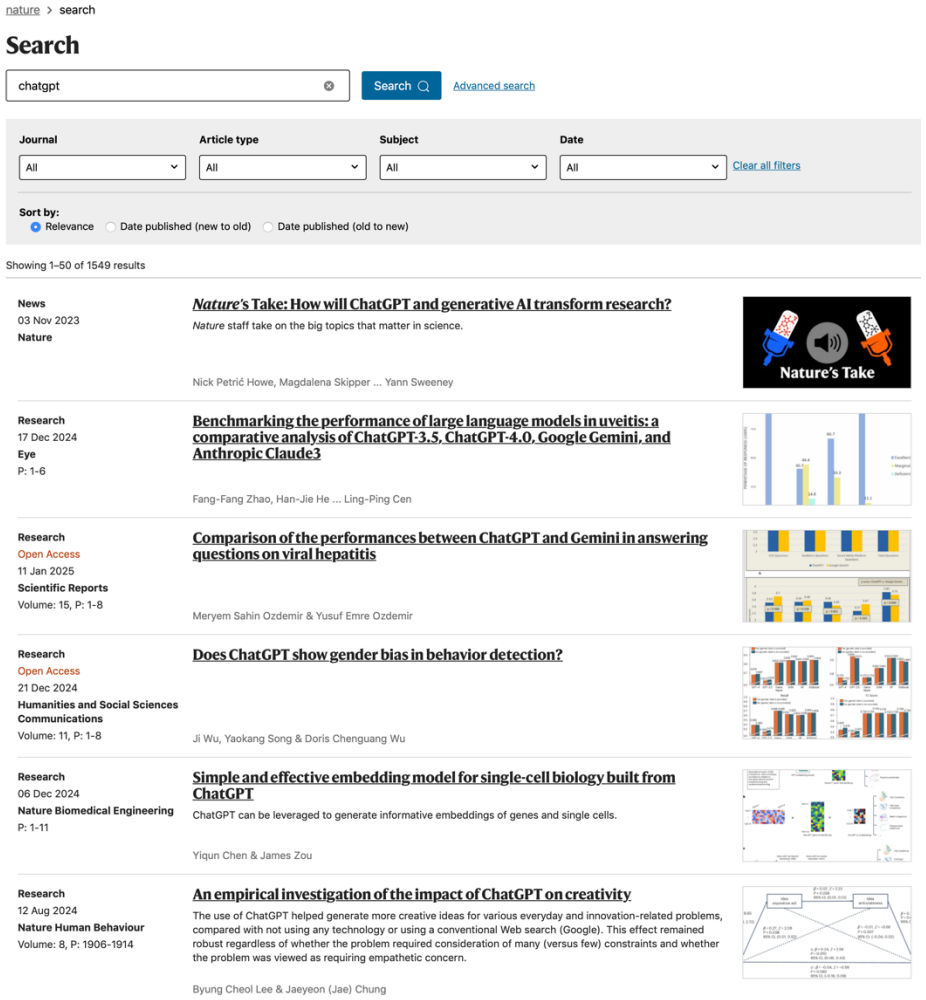


- 范围广泛的世界知识
- 较强的人类指令遵循能力
- 改进的复杂任务推理能力
- 较强的通用任务解决能力
- 较好的人类对齐能力
- 较强的多轮对话交互能力

大语言模型技术开始进入大众视野，改变了传统信息助手的研发路径

大语言模型的到来引发了技术变革

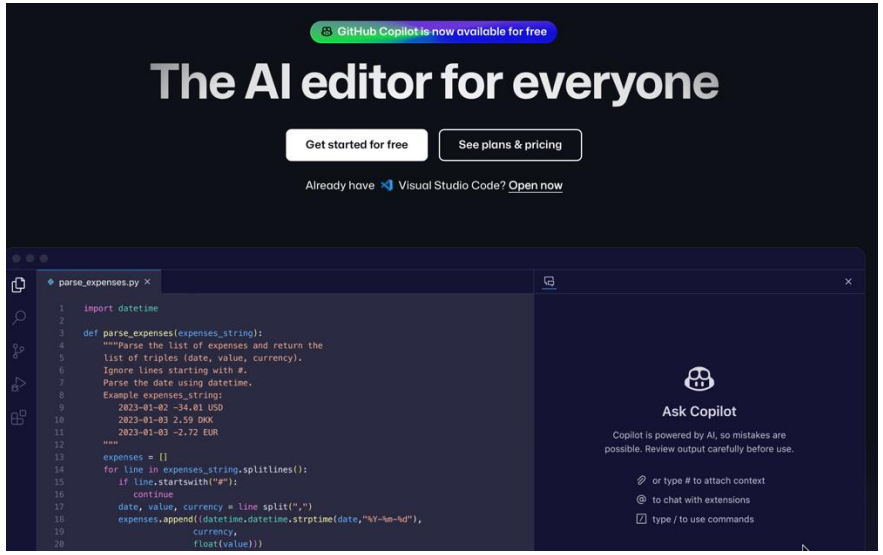
AI Box



Microsoft 365 Copilot

Deliver value and employee satisfaction with our tools for Microsoft 365 Copilot deployment and adoption. This powerful technology combines the power of large language models (LLMs) with your organization's data – all in the flow of work – to turn your words into one of the most powerful productivity tools on the planet.

Microsoft 365 Copilot Chat and in-app experiences provide real-time intelligent assistance, enabling users to enhance their creativity, productivity, and skills.

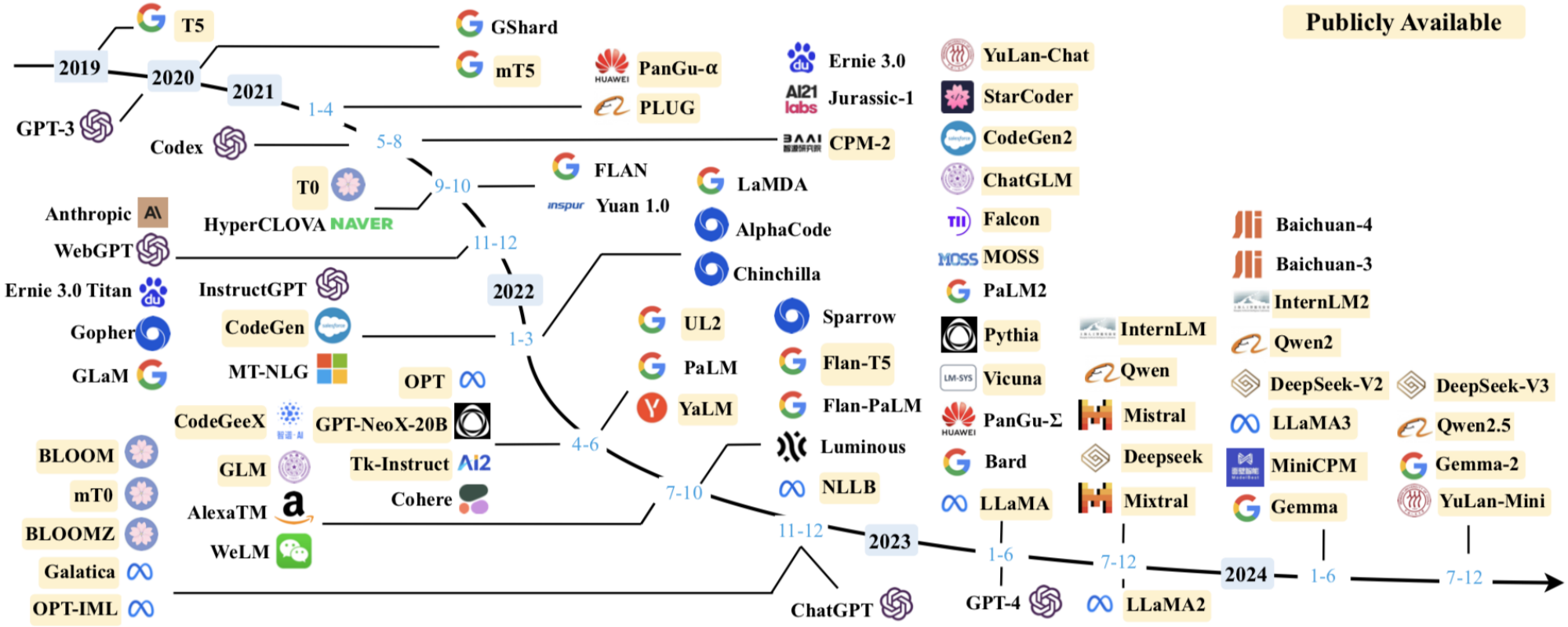


在学术界以及工业界都产生深远影响，预示着新一代信息产业革命的到来

大语言模型的百花齐放时代

AI Box

➢ 发展速度极快，模型性能呈现跃升趋势，中国模型居于世界前列

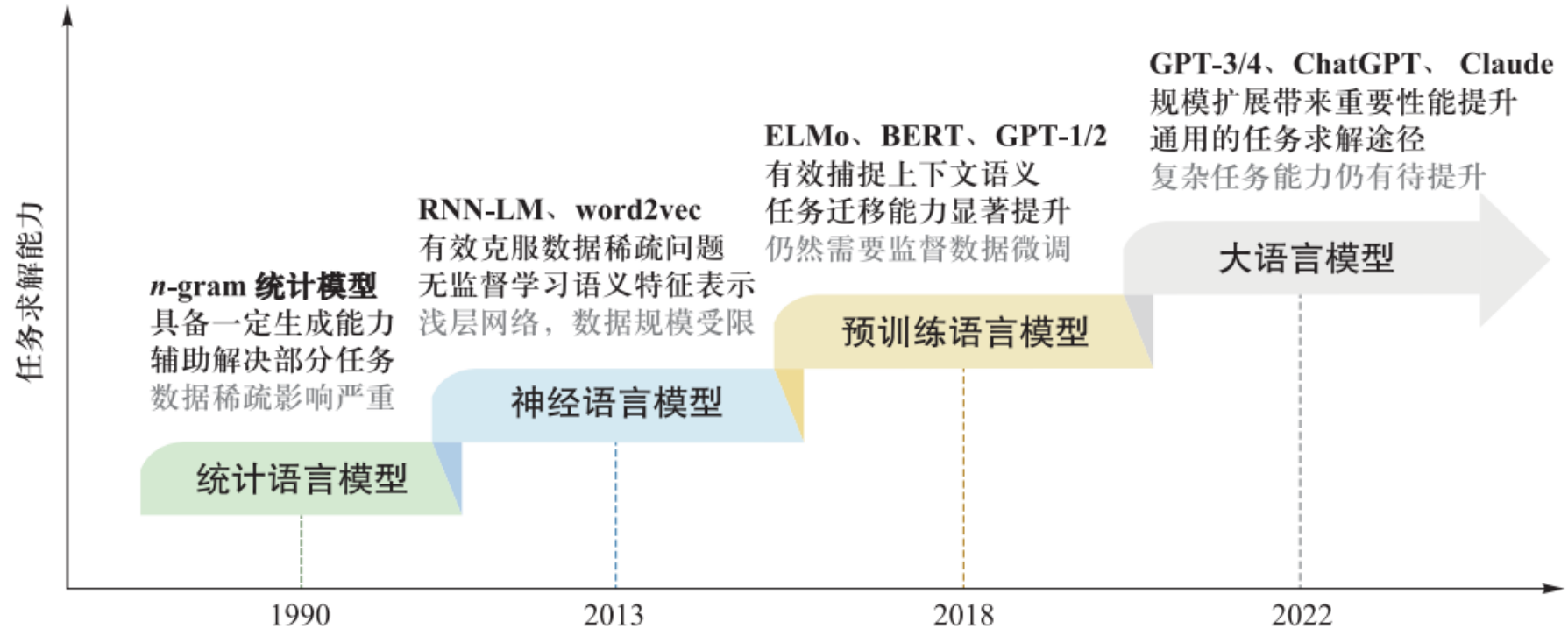


A Survey of Large Language Models, Arxiv, 2023

语言模型发展历程

AI Box

- 语言模型通常是指能够建模自然语言文本生成概率的模型
- 从语言建模到任务求解，这是科学思维的一次重要跃升



语言模型发展历程

AI Box

- 统计语言模型（SLM）
  - 主要建立在统计学习理论框架，通常使用链式法则建模句子序列
    - 例如：$p(I, am, fine) = p(I \mid \text{START}) * p(am|I) * p(fine|I, am)$
  - n-gram 语言模型：基于马尔科夫假设，当前词概率仅与前$n-1$个词有关

$$p(s) = p(w_1)p(w_2|w_1) \dots p(w_m|w_{m-n+1}, \dots, w_{m-1})$$
$$= \prod_{i=1}^{m} p(w_i|w_{i-n+1}, \dots, w_{i-1})$$

  - 如果使用二元语言模型，则上述示例概率计算变为

$$p(I, am, fine) = p(I \mid \text{START}) * p(am|I) * p(fine|\, am)$$

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1086/syllabus.html

语言模型发展历程

AI Box

- 基于频率的估计方法(最大似然估计)
  - 四元语言模型估计示例

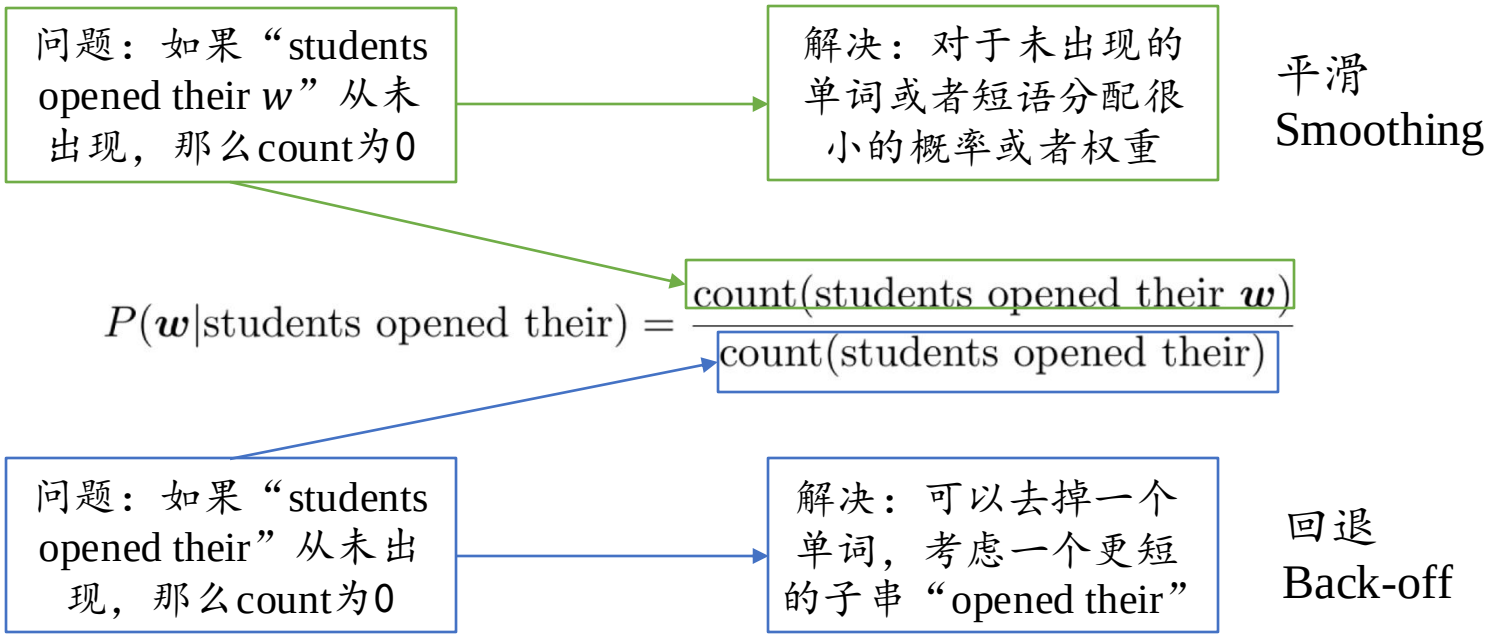
$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w})}{\text{count}(\text{students opened their})}$$

    - “students opened their”出现了1000次
    - “students opened their books”出现了400次
      - P(books|students opened their) = 0.4
    - “students opened their exams”出现了100次
      - P(exams|students opened their) = 0.1

语言模型发展历程

AI Box

- 基于频率的估计方法 (最大似然估计)
  - 主要问题



语言模型发展历程

AI Box

- 基于频率的估计方法 (最大似然估计)
  - 加一平滑 (又称为 Laplace smoothing)
    - 每个词都加上一次出现

原始估计 $\mathrm{P}_{MLE}(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$

加一平滑 $\mathrm{P}_{Add-1}(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i) + 1}{count(w_{i-1}) + V}$ ← 词典大小

    - 仍然保持概率分布，不破坏概率分布基本性质
      - $\mathrm{P}(w_i) > 0, \forall\, w_i \in V$
      - $\sum_i \mathrm{P}(w_i) = 1$

语言模型发展历程

AI Box

- 基于频率的估计方法 (最大似然估计)
  - 回退 (back-off)
    - 当$count(w_{i-n+1},\dots,w_i)=0$，$n$元语言模型退化成更低阶数元语言模型，

$$P(w_i|w_{i-1},\dots,w_{i-n+1})=P(w_i|w_{i-1},\dots,w_{i-n+k+1})$$

    - 例如：当 $count(w_{i-2},w_{i-1},w_i)=0$时，三元语言模型可以退化成二元语言模型进行估计

$$P(w_i|w_{i-1},w_{i-2})=P(w_i|w_{i-1})$$

语言模型发展历程

AI Box

- 基于频率的估计方法 (最大似然估计)
  - 插值 (interpolation)
    - 例如： 混合多个不同阶数的语言模型

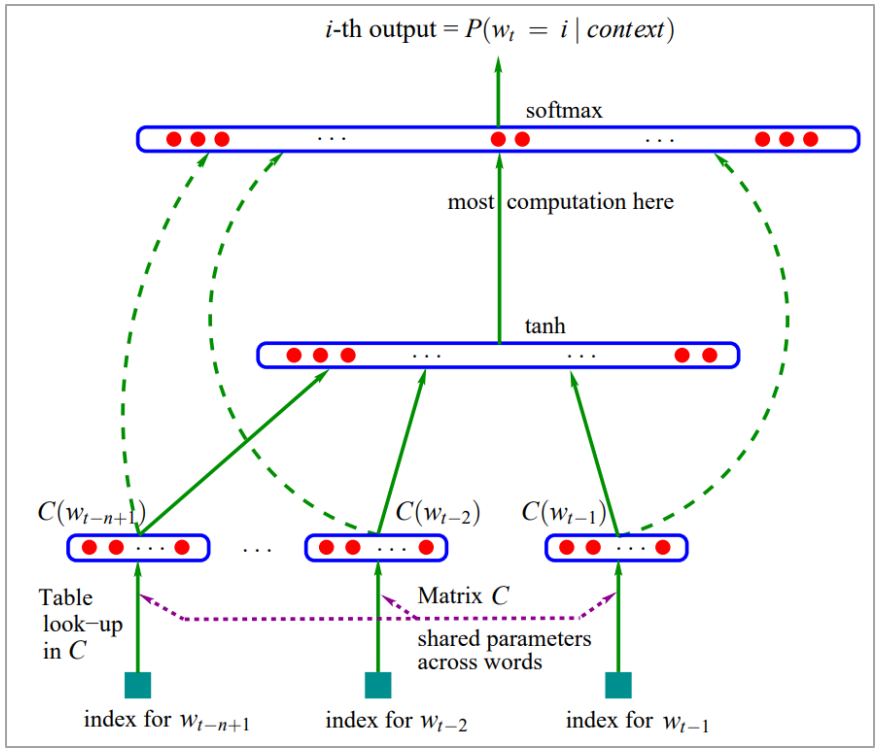
$$\mathrm{P}'(w_i|w_{i-1},w_{i-2}) = \alpha\mathrm{P}(w_i|w_{i-1},w_{i-2}) + \beta\mathrm{P}(w_i|w_{i-1}) + \gamma\mathrm{P}(w_i)$$

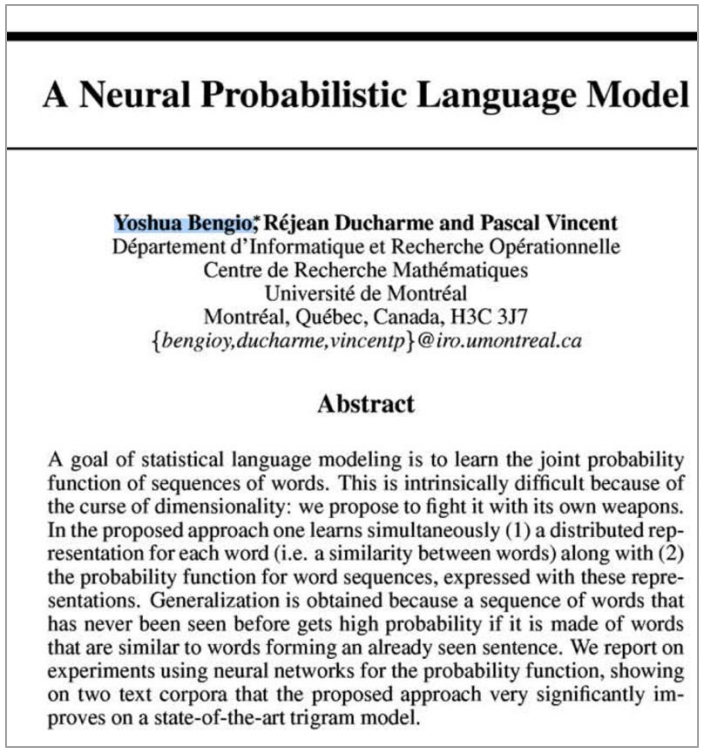
  - 可以证明，仍然能够保证语言模型的概率性质
  - 通常这种方式可以结合不同阶数估计方法的优势
  - 但仍然不能从根本解决数据稀疏性问题

语言模型发展历程

AI Box

- 神经语言模型（NLM）
  - 早期工作（MLP）：单词映射到词向量，再由神经网络预测当前时刻词汇





A Neural Probabilistic Language Model

Yoshua Bengio; Réjean Ducharme and Pascal Vincent
Département d'Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal
Montréal, Québec, Canada, H3C 3J7
{bengioy,ducharme,vincentp}@iro.umontreal.ca

Abstract

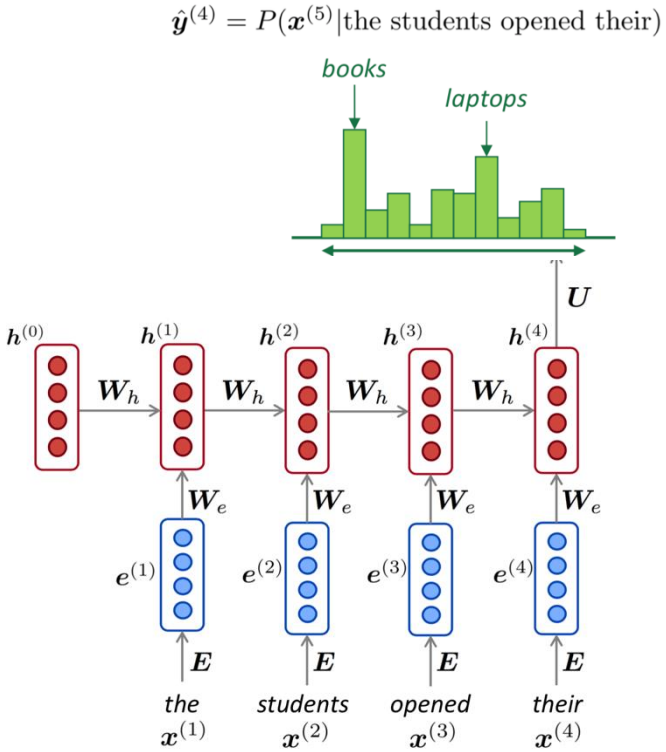
A goal of statistical language modeling is to learn the joint probability function of sequences of words. This is intrinsically difficult because of the curse of dimensionality: we propose to fight it with its own weapons. In the proposed approach one learns simultaneously (1) a distributed representation for each word (i.e. a similarity between words) along with (2) the probability function for word sequences, expressed with these representations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar to words forming an already seen sentence. We report on experiments using neural networks for the probability function, showing on two text corpora that the proposed approach very significantly improves on a state-of-the-art trigram model.

A Neural Probabilistic Language Model. **NIPS 2000.**

语言模型发展历程

AI Box

- 神经语言模型（NLM）
  - 循环神经网络（RNN）



$$\hat{\boldsymbol{y}}^{(t)} = \text{softmax}\left(\boldsymbol{U}\boldsymbol{h}^{(t)} + \boldsymbol{b}_2\right) \in \mathbb{R}^{|V|}$$ 输出词汇的概率分布

$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h\boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e\boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$$ 隐含层

$$\boldsymbol{e}^{(t)} = \boldsymbol{E}\boldsymbol{x}^{(t)}$$ 词嵌入

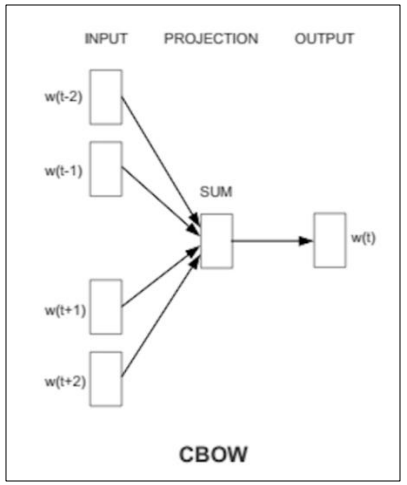
$$\boldsymbol{x}^{(t)} \in \mathbb{R}^{|V|}$$ 词汇，one-hot向量

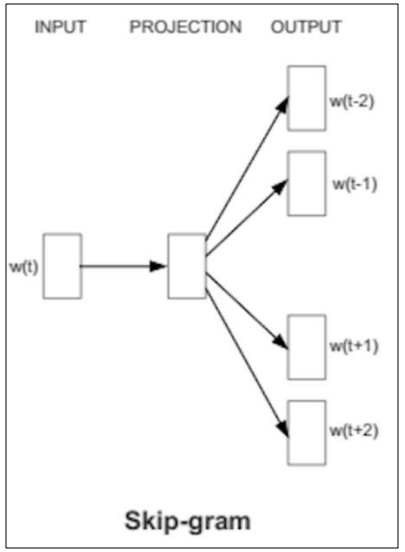
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/index.html#schedule

语言模型发展历程

AI Box

- 神经语言模型（NLM）
  - 简化模型：Word2Vec　NLP领域深度学习时代最重要的工作之一
    - 基本功能
      - 给定文本数据，对于每个单词学习一个低维表示
    - 基于分布式语义的思想进行设计
      - 词义=背景单词的语义
    - 不考虑窗口内单词的顺序
      - 应用了简单的average pooling的策略
    - 充分考虑实践和效果
      - 有很多的优化trick，速度快、效果稳定





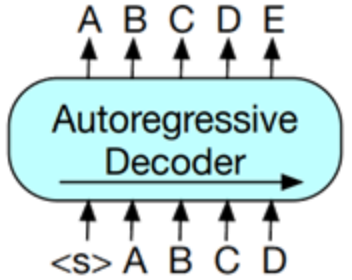
Distributed representations of words and phrases and their compositionality. **NeurIPS 2013**.

语言模型发展历程

AI Box

- 预训练语言模型（PLM）
  - PLM：通过在大量语料上进行无监督预训练后，其可以在特定下游任务或领域上微调并取得较好效果
  - 自回归语言模型：GPT，GPT-2

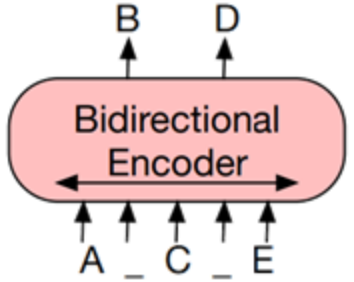
$$\max_{\theta} \log p_\theta(X) = \log \prod_{t=1}^{T} p_\theta(x_t | X_{<t})$$



  - 自编码语言模型：BERT，RoBERTa

$$\max_{\theta} \log p_\theta(X | \hat{X}) \approx \log \prod_{t=1}^{T} m_t p_\theta(x_t | \hat{X})$$

$$m_t = \begin{cases} 1 & \text{当前位置被遮掩} \\ 0 & \text{当前位置未被遮掩} \end{cases}$$



Bert: Pre-training of deep bidirectional transformers for language understanding. **NAACL 2019.**

语言模型发展历程

AI Box

- 传统语言模型存在局限性，需要使用特殊的技术进行模型能力提升
  - 缺乏背景知识
    - 需要知识图谱等外部知识源补充
  - 任务泛化性较差
    - 需要针对特定任务进行微调，适配成本较高
  - 复杂推理能力较弱
    - 通常需要对于结构进行修改，或者进行大规模微调

尽管早期研究工作较多，但是没有工作能够通过统一途径同时解决上述代表性挑战

语言模型发展历程

AI Box

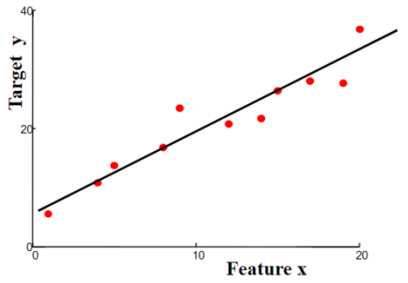
- 大语言模型的到来
  - 海量无标注文本数据预训练得到的大型预训练语言模型
    - 通常指参数规模达到百亿、千亿甚至万亿的模型
    - 经过大规模数据预训练的数十亿参数的高性能模型也可以称为大语言模型
  - 与传统语言模型构建的差异
    - 极大地扩展了模型参数和数据数量
    - 需要更为复杂、精细的模型训练方法

语言模型发展历程

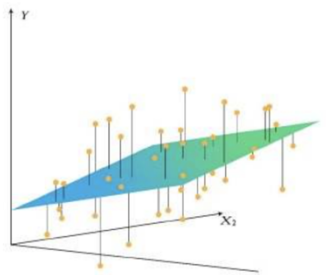
AI Box

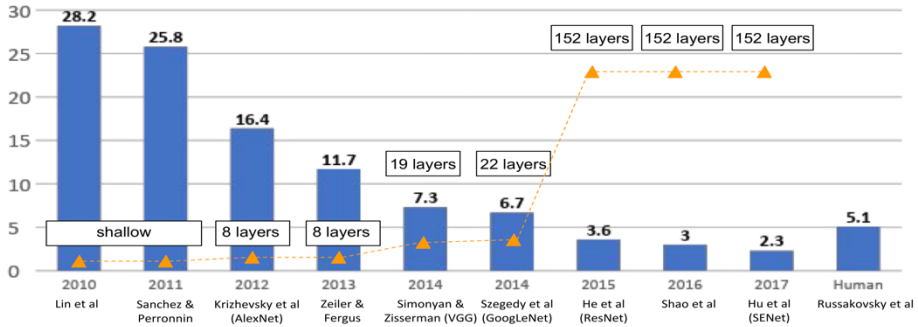
➢ 模型参数规模具备一定规模非常重要

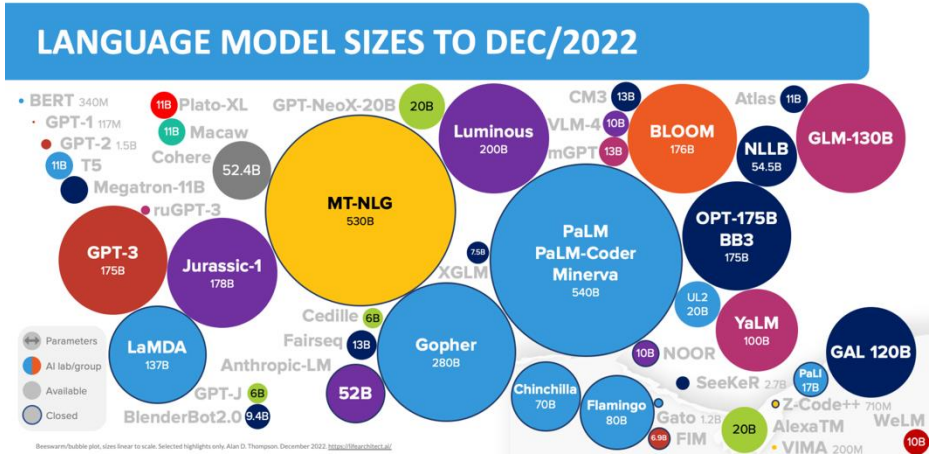
- 一元线性回归
  - 只有一个自变量（特征 x 的维度为1，因此 w 的维度也是1）
    - $y = wx + b$



- 多元线性回归
  - 有多个自变量（特征 x 的维度大于1，因此 w 的维度也大于1）
    - $y = w\ x + b$







语言模型发展历程

AI Box

➢ 模型需要能够学习更多的数据知识

Table 1 | **Current LLMs**. We show five of the current largest dense transformer models, their size, and the number of training tokens. Other than LaMDA (Thoppilan et al., 2022), most models are trained for approximately 300 billion tokens. We introduce *Chinchilla*, a substantially smaller model, trained for much longer than 300B tokens.

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
Chinchilla	70 Billion	1.4 Trillion

数据数量、数据质量决定了模型的能力，同样意味着大算力需求

Training Compute-Optimal Large Language Models

谢谢