



中国科学院大学

University of Chinese Academy of Sciences

自然语言处理

第2讲 大语言模型

王石 资康莉 刘瑜

2026年春季课程

<https://ictkc.github.io/teaching/>



第二讲 大语言模型

语言模型 (Language Model)

□ 评估语言合理性的概率模型

我有一个苹友



完全听不懂
他们在说什么

我有一个苹果

非常合理
张卫平传



语言模型 (Language Model)

- 语言模型可预测下一个字

我有一个苹__

预测下一个字很重要

Ilya Sutskever (co-founder of OpenAI) :
预测下一个词能带来真正的理解，你在读
一本侦探小说，在最后一页侦探说：“我将
揭示罪犯的身份，他的名字是_____。”





目 录

1

统计语言模型

2

3

4

如何构建语言模型?

我有一个苹__



统计法

$$p(w_1 w_2 \dots w_{n-1} w_n | w_1 w_2 \dots w_{n-1}) = \frac{\text{Freq}(w_1 w_2 \dots w_{n-1} w_n)}{\text{Freq}(w_1 w_2 \dots w_{n-1})}$$

$$\text{Freq}(\text{我有一个苹果}) = 9$$

$$\text{Freq}(\text{我有一个苹}) = 10$$

$$p(\text{果} | \text{我有一个苹}__) = 0.9$$

统计法 + 马尔可夫假设

$$\underbrace{p(w_1 w_2 \dots w_{n-1} w_n | w_1 w_2 \dots w_{n-1})}_{\text{太稀疏}} \approx \underbrace{p(w_{n-1} w_n | w_{n-1})}_{\text{Bi-Gram Model}} = \frac{\text{Freq}(w_{n-1} w_n)}{\text{Freq}(w_{n-1})}$$

$$\text{Freq}(\text{苹果}) = 200$$

$$\text{Freq}(\text{苹}) = 1000$$

$$p(\text{果} | \text{苹}_) = 0.20$$

$$\text{Freq}(\text{我有一个苹果}) = 9$$

$$\text{Freq}(\text{我有一个苹}) = 10$$

$$p(\text{果} | \text{我有一个苹}_) = 0.90$$

统计法 + 马尔可夫假设

$$p(w_1 w_2 \dots w_{n-1} w_n | w_1 w_2 \dots w_{n-1}) \approx p(\underline{w_{n-1} w_n} | w_{n-1}) = \frac{\text{Freq}(w_{n-1} w_n)}{\text{Freq}(w_{n-1})}$$

Tri-Gram Model

$$\text{Freq}(\text{一个苹果}) = 100$$

$$\text{Freq}(\text{一个苹}) = 400$$

$$p(\text{果} | \text{一个苹}_) = 0.25$$

$$\text{Freq}(\text{我有一个苹果}) = 9$$

$$\text{Freq}(\text{我有一个苹}) = 10$$

$$p(\text{果} | \text{我有一个苹}_) = 0.90$$

$$\text{Freq}(\text{苹果}) = 200$$

$$\text{Freq}(\text{苹}) = 1000$$

$$p(\text{果} | \text{苹}_) = 0.20$$

N元统计语言模型

□ N-Gram Model: **N越大，预测能力越强**



手__ 指 | 杖 | 镯 | 铐 | 办 | ...

一个手__ 指 | 杖 | 镯 | 铐 | 办 | ...

礼物是一个手__ 指 | 杖 | 镯 | 铐 | 办 | ...

男朋友的礼物是一个手__ 指 | 杖 | 镯 | 铐 | 办 | ...

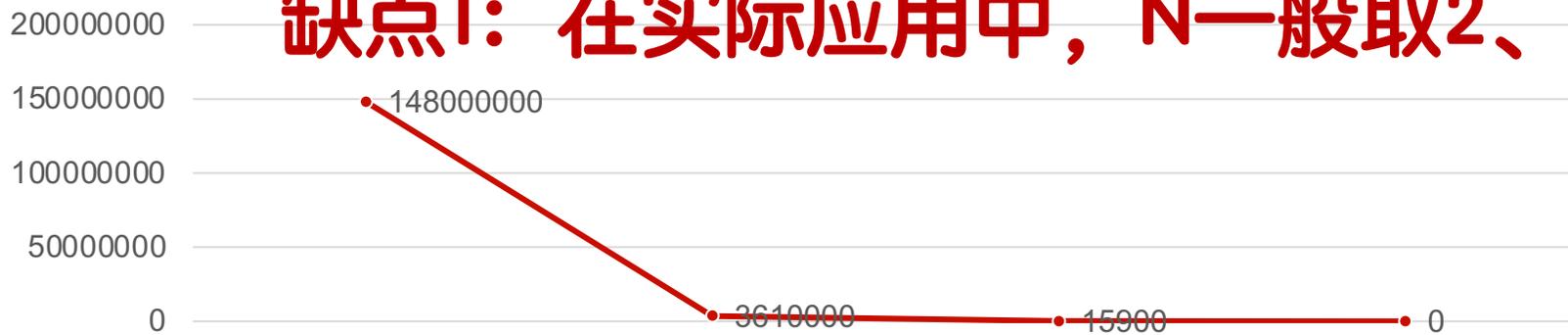
给男朋友的礼物是一个手__ 指 | 杖 | 镯 | 铐 | 办 | ...

N元统计语言模型

□ N-Gram Model: **N越大，出现概率越低**

The image shows three Google search results illustrating the N-gram model's performance. The first search for '手' (hand) returns approximately 1,480,000,000 results in 0.29 seconds. The second search for '礼物是一个手' (A gift is a hand) returns approximately 15,900 results in 0.16 seconds. The third search for '男朋友的礼物是一个手' (A boyfriend's gift is a hand) returns 0 results.

缺点1: 在实际应用中, N一般取2、3



未找到符合“男朋友的礼物是一个手”的结果。

$$\frac{Freq(w_1 w_2 \dots w_{n-1} w_n)}{Freq(w_1 w_2 \dots w_{n-1})}$$

0

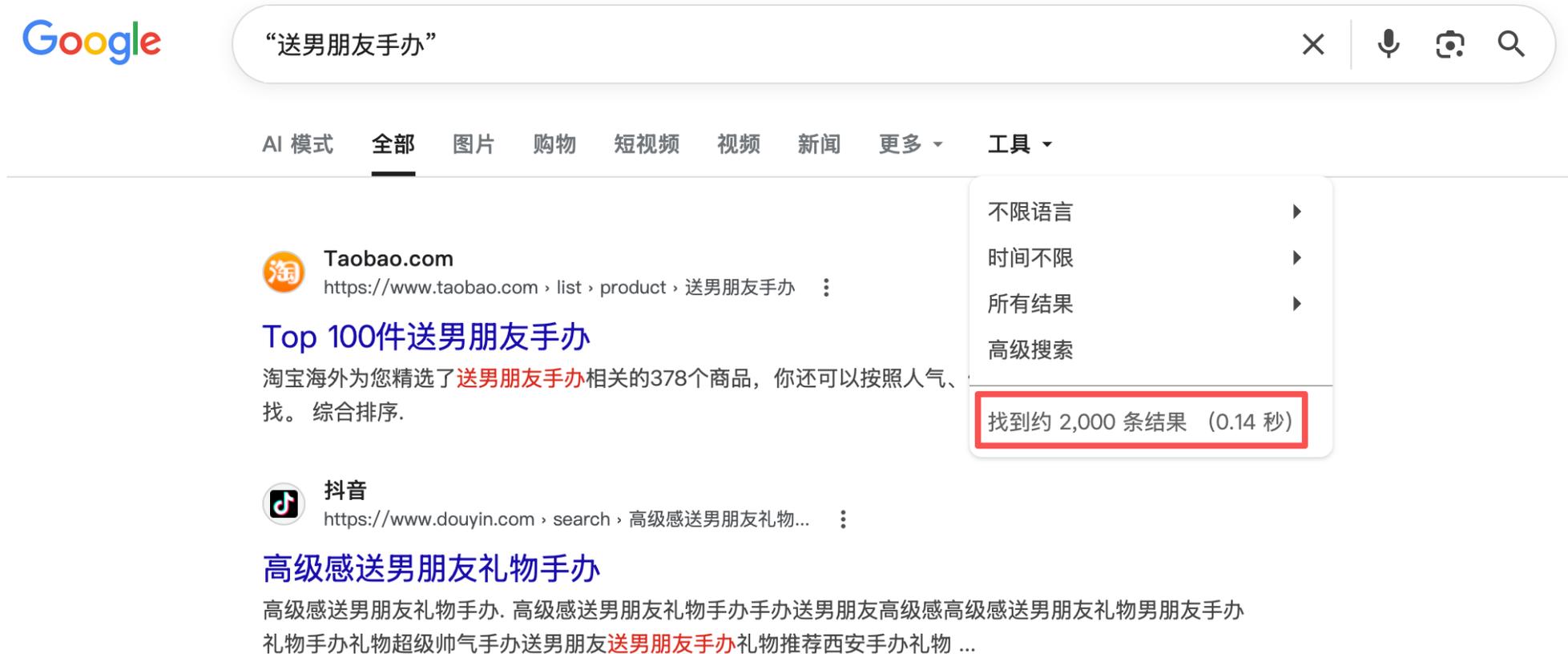
N元统计语言模型

□ 平滑策略

- **加一平滑**Additive smoothing
- 古德-图灵Good-Turing smoothing
- 插值平滑Interpolation-Jelinek-Mercer smoothing
- 回退平滑Katz smoothing
- Witten-Bell smoothing
- 绝对减值平滑Absolute discounting smoothing
- Kneser-Ney smoothing

N元统计语言模型

□ 缺点2：泛化能力弱



Google “送男朋友手办”

AI 模式 全部 图片 购物 短视频 视频 新闻 更多 ▾ 工具 ▾

Taobao.com
https://www.taobao.com › list › product › 送男朋友手办

Top 100件送男朋友手办
淘宝海外为您精选了送男朋友手办相关的378个商品，你还可以按照人气、找。综合排序。

抖音
https://www.douyin.com › search › 高级感送男朋友礼物...

高级感送男朋友礼物手办
高级感送男朋友礼物手办. 高级感送男朋友礼物手办手办送男朋友高级感高级感送男朋友礼物男朋友手办礼物手办礼物超级帅气手办送男朋友送男朋友手办礼物推荐西安手办礼物 ...

不限语言 ▾
时间不限 ▾
所有结果 ▾
高级搜索

找到约 2,000 条结果 (0.14 秒)



目 录

1

统计语言模型

2

神经网络语言模型

3

4

还能怎么预测下一个词?



神经网络语言模型

2. A Neural Model

The training set is a sequence $w_1 \cdots w_T$ of words $w_t \in V$, where the vocabulary V is a large but finite set. The objective is to learn a good model $f(w_t, \cdots, w_{t-n+1}) = \hat{P}(w_t | w_1^{t-1})$, in the sense that it gives high out-of-sample likelihood. Below, we report the geometric average of $1/\hat{P}(w_t | w_1^{t-1})$, also known as *perplexity*, which is also the exponential of the average negative log-likelihood. The only constraint on the model is that for any choice of w_1^{t-1} , $\sum_{i=1}^{|V|} f(i, w_{t-1}, \cdots, w_{t-n+1}) = 1$, with $f > 0$. By the product of these conditional probabilities, one obtains a model of the joint probability of sequences of words.

$$f(i, w_{t-1}, \cdots, w_{t-n+1}) = \underline{g}(i, \underline{C}(w_{t-1}), \cdots, C(w_{t-n+1}))$$

词 i 的输出概率

一个神经网络

词的分布式表示

神经网络语言模型

A neural probabilistic language model

[Y Bengio](#), [R Ducharme](#), [P Vincent](#), [C Jauvin](#) - Journal of machine learning ..., 2003 - jmlr.org

... A goal of statistical **language modeling** is to learn the joint **probability** function of sequences of words in a **language**. This is intrinsically difficult because of the curse of dimensionality: a ...

☆ 保存 引用 被引用次数: **12936** 相关文章 所有 65 个版本

Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of machine learning research 3.Feb (2003): 1137-1155.



约书亚·本吉奥 OC FRS FRSC (法语: Yoshua Bengio, 1964年3月5日^[1]) 是一名加拿大计算机科学家, 因其在人工神经网络和深度学习方面的研究而知名^{[2][3][4]}, 其中最为著名的工作是神经概率语言模型^[5]、ReLU^[6]和注意力机制。他是蒙特利尔大学计算机科学和运筹学系的教授以及蒙特利尔学习算法研究所科学主任。

本吉奥与杰弗里·辛顿和杨立昆一起获得2018年的图灵奖, 以表彰他们在深度学习方面的贡献^[7]。这三人的有时被称为“AI教父”和“深度学习教父”^{[8][9][10][11][12][13]}。

Yoshua Bengio (1964~)



- Bengio出生在巴黎，父母是嬉皮士，从小就随父母到处搬家，**走过了世界的很多角落**。Yoshua Bengio说：他有责任照顾生活在发展中国家的人。
- 他**举止谦逊，近乎谦卑**，像一个在电脑屏幕前呆了很长时间的人。

nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾ [Subscribe](#)

[nature](#) > [news](#) > article

NEWS | 12 November 2025

‘Godfather of AI’ becomes first person to hit one million citations

The milestone makes machine-learning trailblazer Yoshua Bengio the most cited researcher on Google Scholar.

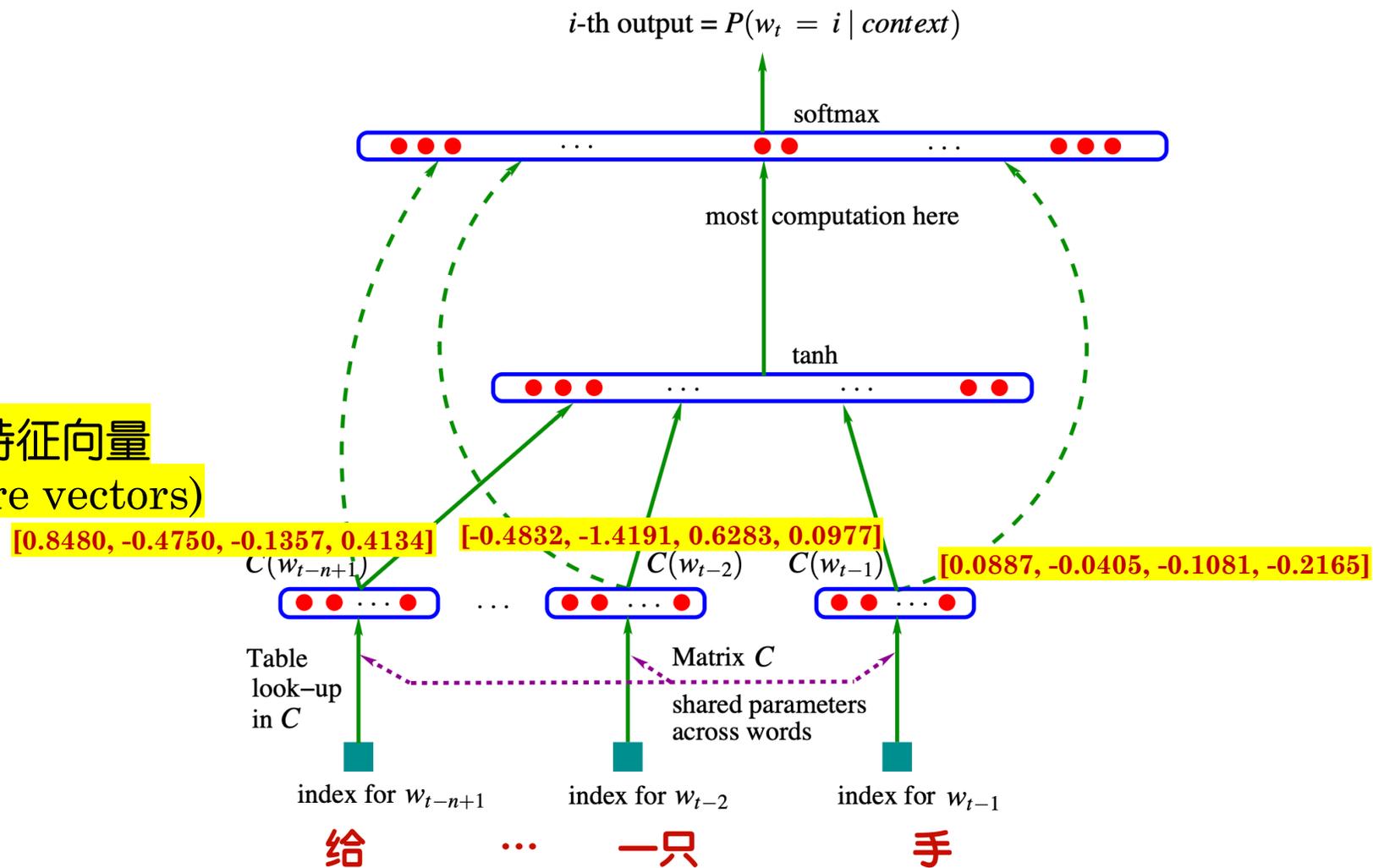
By [Davide Castelvecchi](#)



- 全球被引用最多的计算机科学家，以及**所有领域被引用最多的在世科学家**
- 2025年11月，成为第一位在谷歌学者中引用超过**100万次**的人工智能研究人员

神经网络语言模型

(1) 词分布式特征向量
(distributed feature vectors)



神经网络语言模型

(3) 输出每个词的概率以及最可能的词

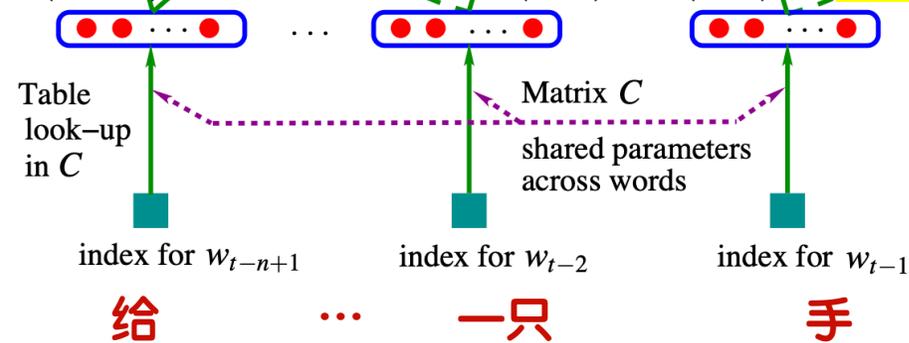
$$i\text{-th output} = P(w_t = i \mid \text{context})$$

$p(\text{办} \mid \text{给...一只手}) = 0.075$

$p(\text{机} \mid \text{给...一只手}) = 0.000011$ $p(\text{铐} \mid \text{给...一只手}) = 0.000001$

(1) 词分布式特征向量
(distributed feature vectors)

$[0.8480, -0.4750, -0.1357, 0.4134]$ $[-0.4832, -1.4191, 0.6283, 0.0977]$ $[0.0887, -0.0405, -0.1081, -0.2165]$

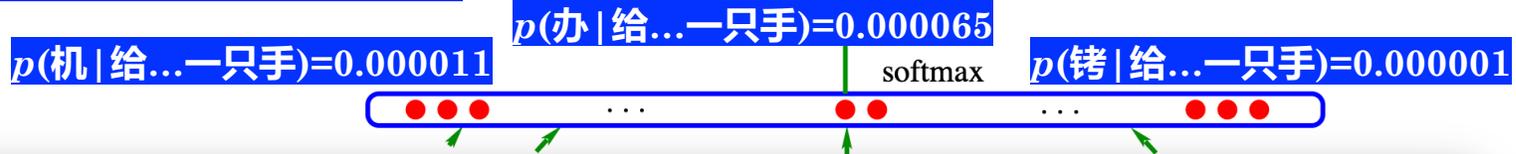


神经网络语言模型

(3) 输出每个词的概率以及最可能的词

$$i\text{-th output} = P(w_t = i | \text{context})$$

$$p(\text{办} | \text{给...一只手}) = 0.075$$



(2) 神经网络

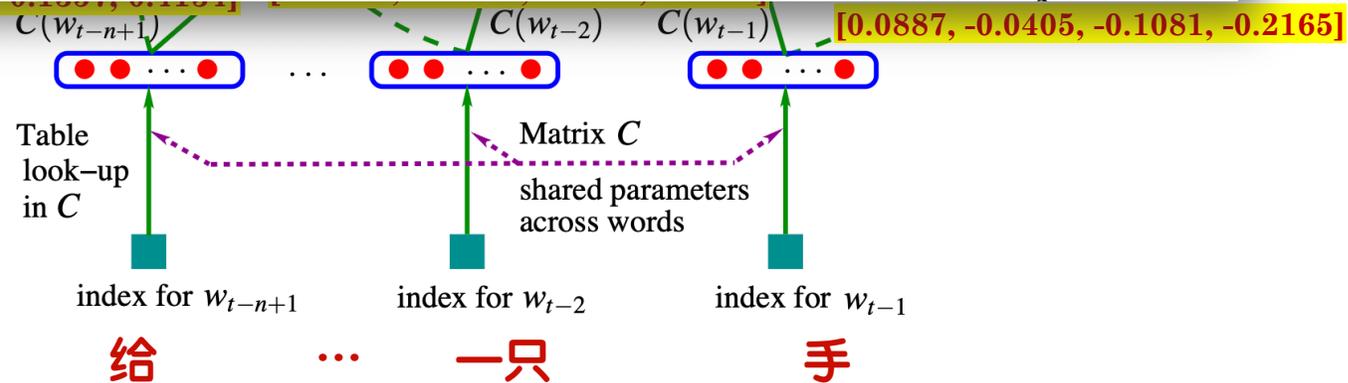
matrix C whose row i is the feature vector $C(i)$ for word i . The function g may be implemented by a feed-forward or recurrent neural network or another parametrized function, with parameters ω . The overall parameter set is $\theta = (C, \omega)$.

Training is achieved by looking for θ that maximizes the training corpus penalized log-likelihood:

$$L = \frac{1}{T} \sum_t \log f(w_t, w_{t-1}, \dots, w_{t-n+1}; \theta) + R(\theta),$$

where $R(\theta)$ is a regularization term. For example, in our experiments, R is a weight decay penalty

(1) 词分布 (distributed fe)



NNLM能力优于统计模型

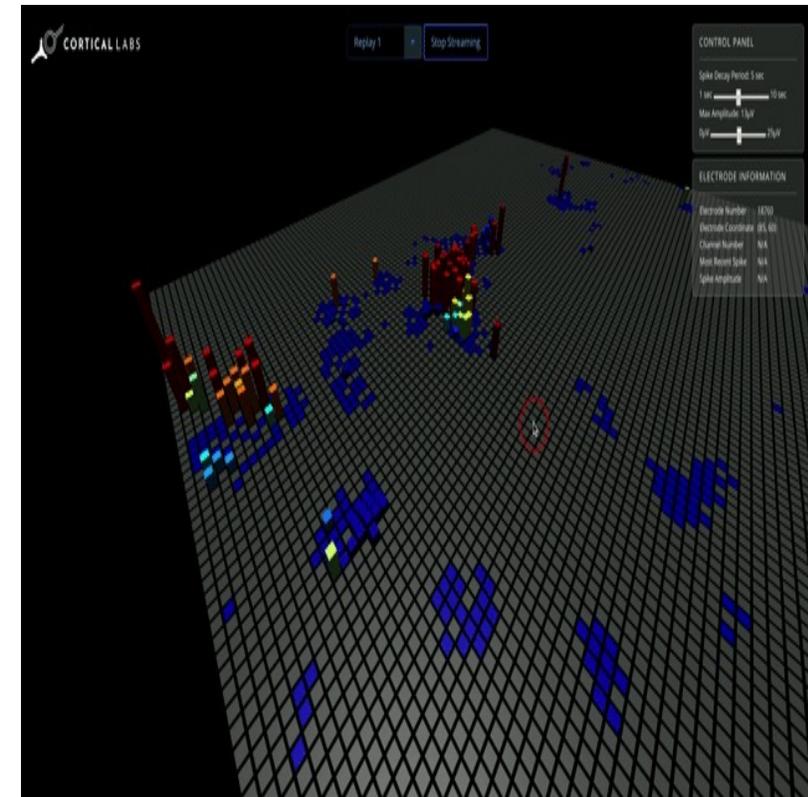
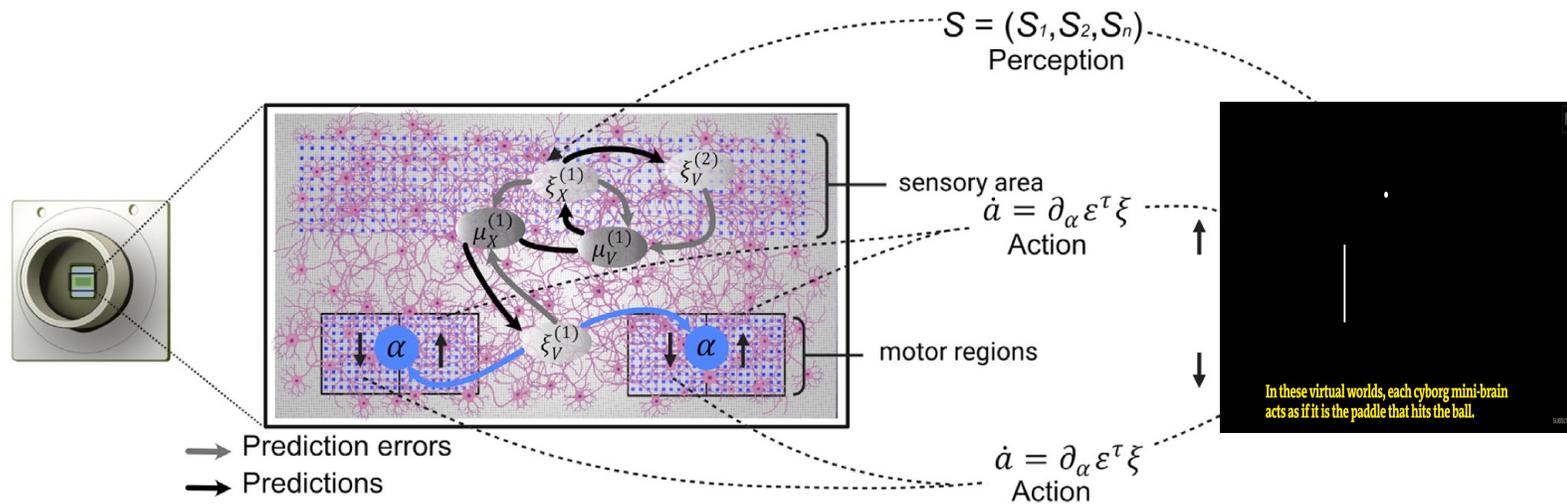
The **main result** is that significantly better results can be obtained when using the neural network, in comparison with the best of the n-grams, with a test perplexity difference of about 24% on Brown and about 8% on AP News, when taking the MLP versus the n-gram that worked best on the validation set. The table also suggests that the neural network was able to take advantage of more

$$\text{Perplexity}(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}} = \prod_{i=1}^N P(w_i | w_1, \dots, w_{i-1})$$

NNLM	MLP9	5		100	30	no	no	175	280	276
	MLP10	5		100	30	no	yes		265	252
	Del. Int.	3						31	352	336
Tri-Gram	Kneser-Ney back-off	3							334	323
	Kneser-Ney back-off	4							332	321
	Kneser-Ney back-off	5							332	321
	class-based back-off	3	150						348	334
	class-based back-off	3	200						354	340
	class-based back-off	3	500						326	312

复习：神经网络

归根结底，人类智能的物质基础是约千亿级神经元和百万亿神经突触连接的生物组织



100万个神经细胞在培养皿中经“电击”训练，学会了弹球游戏

Kagan B J, Kitchen A C, Tran N T, et al. In vitro neurons learn and exhibit sentience when embodied in a simulated game-world[J]. Neuron, 2022, 110(23): 3952-3969. e8.

神经元 (神经细胞)

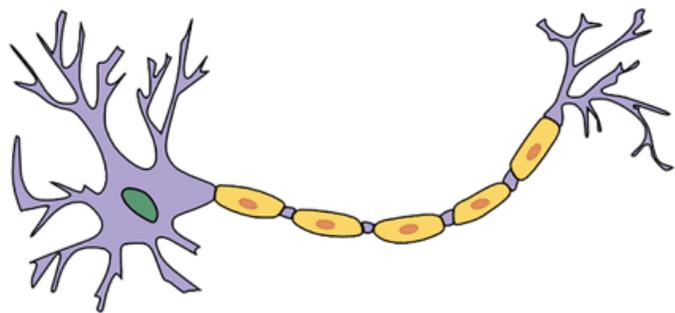
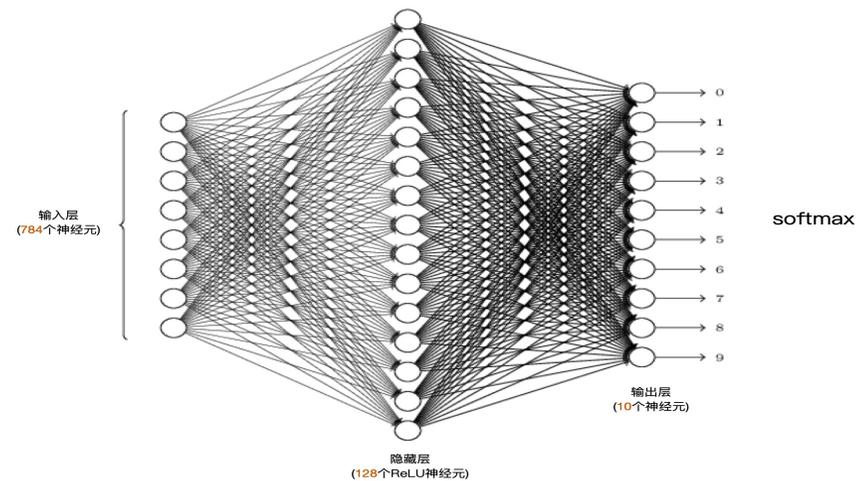
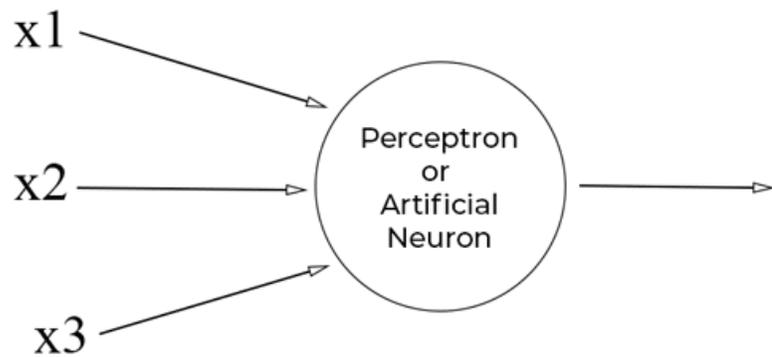


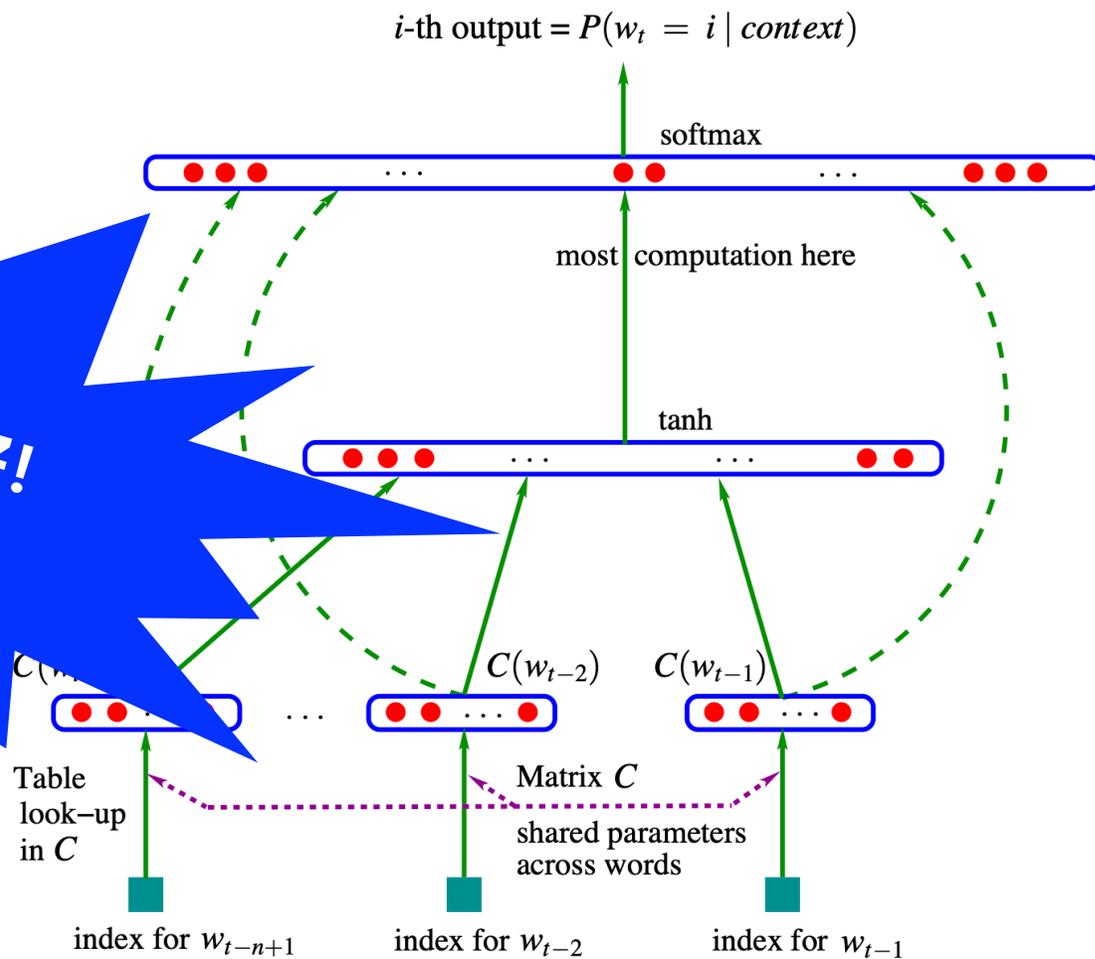
Fig: Biological Neuron



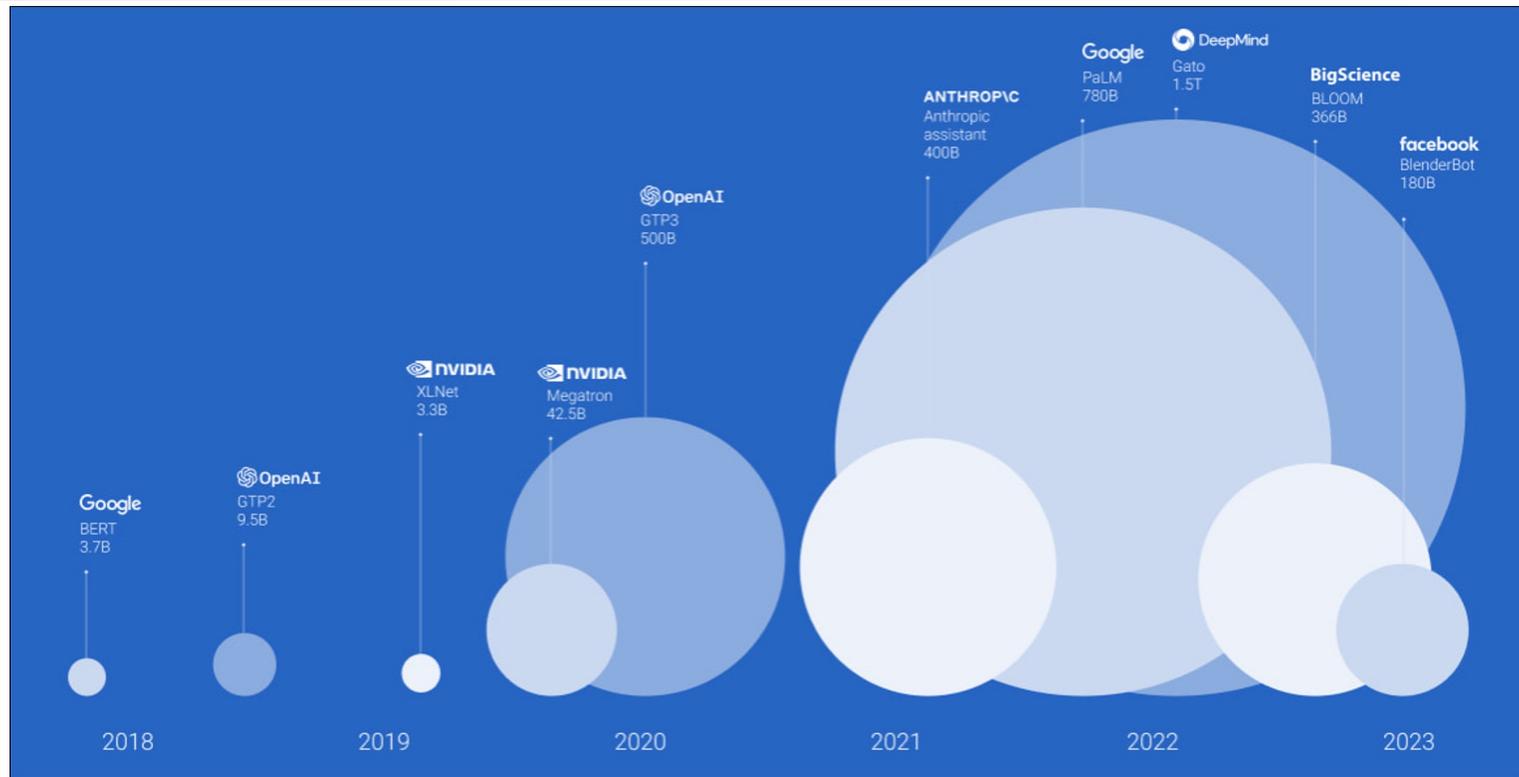
怎么才能预测的更好

· 参数量越大越好!

· 语料库越大越好!



Scaling Law: 参数越大, 能力越强



- GPT-1: 1.17亿参数, 数据集使用1万本书的BookCorpus, 25亿词
- GPT-2: 15亿参数, 数据为40GB的WebText
- GPT-3: 1750亿参数, 数据包括CC(4千亿词)+WebText2(190亿词)+BookCorpus(670亿词)+Wiki(30亿词)
- GTP-4: 1-1.7万亿参数



目 录

1

统计语言模型

2

神经网络语言模型

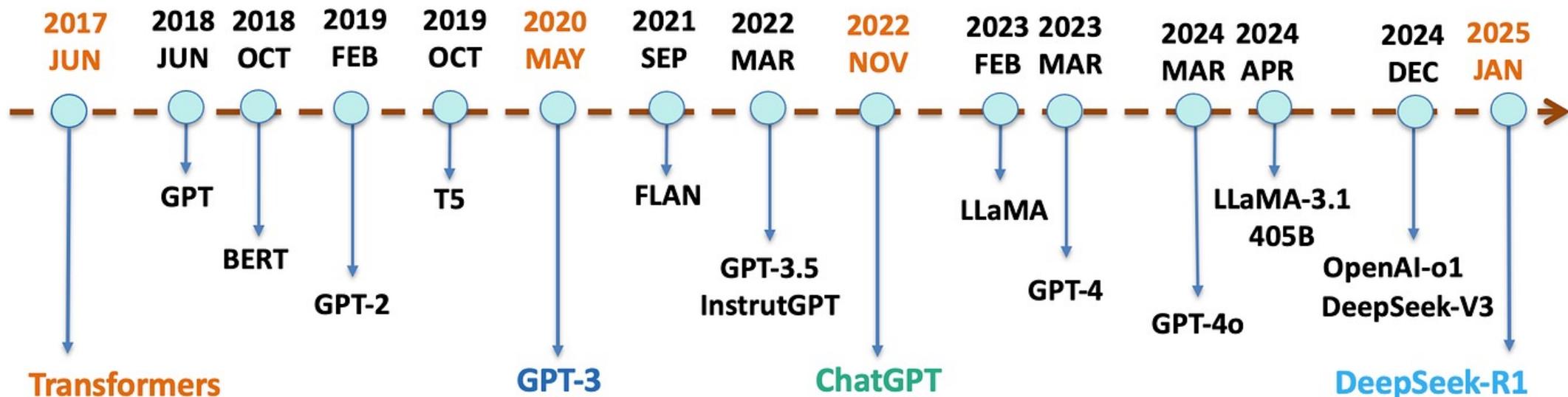
3

大语言模型

4

大模型发展简史

A Brief History of LLMs



大语言模型 (LLM) 上下文更长

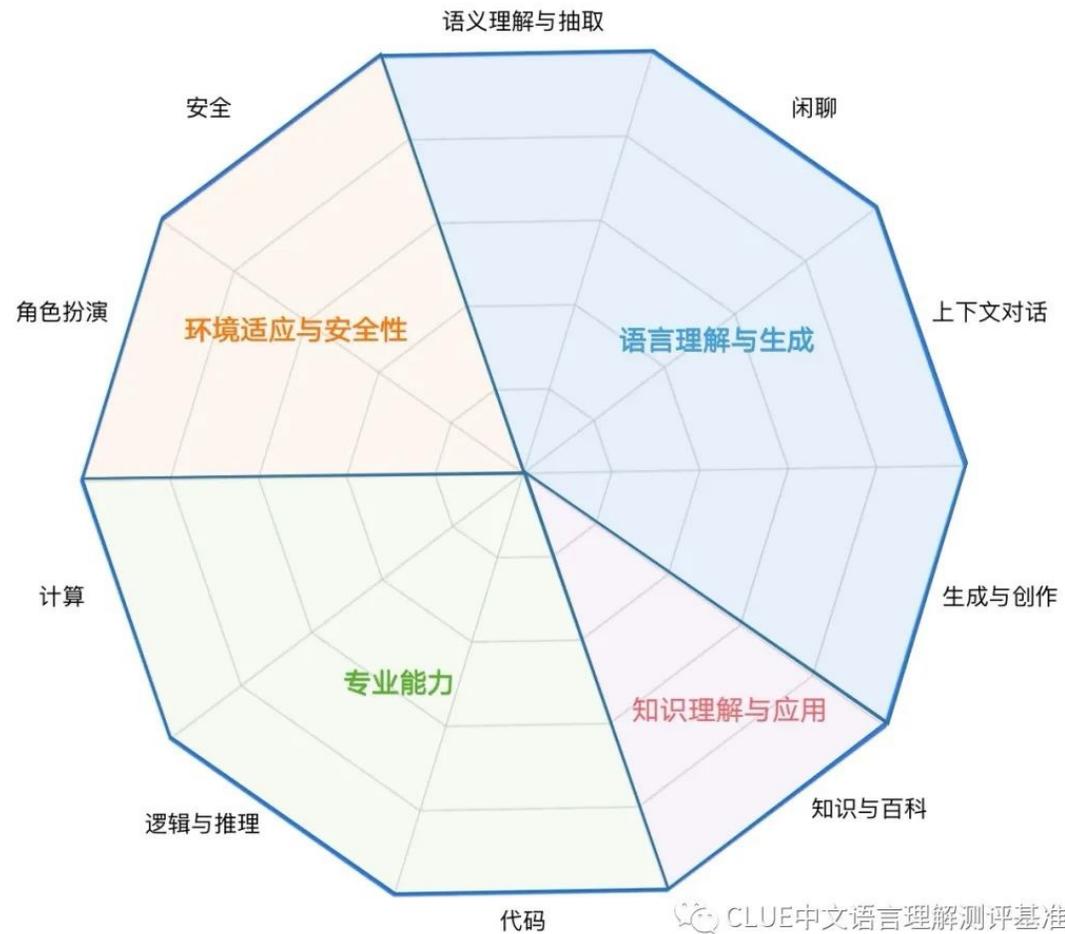
 Qwen3.5-Plus ▾

以“我给男朋友的礼物是一只手办”续写 100 字

我给男朋友的礼物是一只手办，那是他念叨许久的限定款。拆开包装的瞬间，他眼里的惊喜藏都藏不住，小心翼翼地捧在掌心反复端详。看着他像个孩子般灿烂的笑容，我觉得这份心意比任何昂贵的珠宝都珍贵，只要他开心，一切都值得。

GPT-3.5	4K/16K tokens (1token≈1.5-1.8汉字)
GPT-4 Turbo	128K tokens
Claude 3.7	200K tokens
Kimi Chat (月之暗面)	588K tokens
Llama-2	4K/32K/64K tokens
DeepSeek V3	64K tokens (三体I、II、III全集, 红楼梦前80回)

LLM不再局限于NLP任务



数学推理

问题： 设 x, y 为实数，且满足 $\sin x + \sin y = \sin(x + y)$ 。求 $\cos x + \cos y$ 的所有可能取值。

模型回答： 本题可先利用三角函数两角和公式对已知条件进行化简，然后通过设未知数的方式求解 $\cos x + \cos y$ 的取值范围。

- **步骤一：对 $\sin x + \sin y = \sin(x + y)$ 进行化简**

根据两角和公式 $\sin(A + B) = \sin A \cos B + \cos A \sin B$ ，将 $\sin x + \sin y = \sin(x + y)$ 展开可得：

$$\sin x + \sin y = \sin x \cos y + \cos x \sin y$$

<篇幅有限省略>

- **当 $\sin(\frac{x+y}{2}) = 0$ 时：**

此时 $\frac{x+y}{2} = n\pi, n \in Z$ ，即 $x + y = 2n\pi, n \in Z$ ，那么 $y = 2n\pi - x$ 。

将 $y = 2n\pi - x$ 代入 $\cos x + \cos y$ 可得：

$$\cos x + \cos y = \cos x + \cos(2n\pi - x)$$

根据诱导公式 $\cos(2k\pi - \alpha) = \cos \alpha$ ，可得 $\cos x + \cos(2n\pi - x) = \cos x + \cos x = 2 \cos x$ 。

因为 $-1 \leq \cos x \leq 1$ ，所以 $-2 \leq 2 \cos x \leq 2$ ，即 $-2 \leq \cos x + \cos y \leq 2$ 。

综合以上三种情况， $\cos x + \cos y$ 的取值范围是 $[-2, 2]$ 。

所以，答案为 $[-2, 2]$ 。

代码生成

问题:

```
from typing import List
```

```
def task_func(matrix: List[List[int]]) -> List[List[int]]:
```

```
    """
```

给定一个 $m \times n$ 的矩阵，表示每个单元格的_{高度}，该矩阵代表一个大陆的高度图。

“北冰洋”接触矩阵的上边，“印度洋”接触矩阵的下边。

水流只能通过四个方向（上、下、左、右）从一个单元流向高度更低或相等的相邻单元。

请找出所有可以同时流向北冰洋和印度洋的单元格坐标。

要求:

- 返回的坐标列表应按唯一顺序排列（按从左到右、从上到下的顺序排列）。

```
    """
```

模型回答:

```
from typing import List
```

```
def task_func(matrix: List[List[int]]) -> List[List[int]]:
```

```
    if not matrix or not matrix[0]:
```

```
        return []
```

```
    m, n = len(matrix), len(matrix[0])
```

```
    # 定义可以流向北冰洋和印度洋的集合
```

```
    can_flow_to_north = set()
```

```
    can_flow_to_south = set()    return False
```

```
    if ju[2] <= 0:
```

```
        return False
```

```
    for i in range(3):
```

```
        if ju[i] == 0:
```

```
            continue
```

```
        max_amount = math.floor(ju[i] / 2)
```

```
        res = math.ceil(target / max_amount)
```

```
        if res > 0:
```

```
            return True
```

```
    return False
```

<篇幅有限省略>

```
    # 找出同时可以流向北冰洋和印度洋的单元格
```

```
    result = []
```

```
    for i in range(m):
```

```
        for j in range(n):
```

```
            if (i, j) in can_flow_to_north and (i, j) in
```

```
can_flow_to_south:
```

```
                result.append([i, j])
```

```
    return result
```

指令遵循

问题：请撰写一篇关于“书籍与旅行”的短文，文章中必须使用“远方”、“故事”和“灯火”这三个词语。文章中每个句子（用中文句号或问号或感叹号分隔）的最后一个字必须以“也”字结尾。请确保文章只能有4个段落（不要出现标题），每个段落之间用“~~”符号分隔，其中第2段的第8个汉字必须是“桥”字。

回答：

远方的书籍，会带来许多故事的**魅力**，人们可以通过阅读这些**故事**，体验到**远方**的文化和历史，也可以从中获得深深的**体验**。

~~

当我们旅行到**远方**，看到的**灯火**会让我们感到一种深深的感伤，灯火的闪烁，会让我们感到一种温暖的感觉，也会让我们感受到一种被怀念的**感觉**。

~~

在这些故事中，我们会看到许多桥，连接着不同的人 and 不同的事物，灯火的闪烁，会让我们感到一种被连接的感觉，也会让我们感受到一种被融合的**感觉**。

~~

远方的书籍和灯火，会让我们感到一种被 **transported** 的感觉，人们可以通过阅读这些故事和看到灯火的闪烁，体验到远方的文化和历史，也可以从中获得深深的**体验**。

文本理解

问题：你是一家人工智能芯片研发公司的研发部门经理，需要撰写一份《AI芯片性能优化项目延期情况说明》。该项目原计划在6个月内完成新一代7nm制程AI芯片的性能提升和功耗优化，但目前已超期2个月。请向技术副总裁提交一份正式的项目延期情况说明，详细解释延期原因（包括技术瓶颈、供应链风险等），并提出具体的改进措施。说明文件需要采用技术管理报告的规范格式，包含数据分析和里程碑调整计划。

模型回答： AI芯片性能优化项目延期情况说明 <篇幅有限省略部分内容>

提交人：研发部门经理

提交日期：XXXX年XX月XX日

一、项目基本情况

项目名称：AI芯片性能优化项目

原计划周期：6个月（XX年XX月-XX年XX月）

二、延期原因分析

1. 技术瓶颈

- 电路优化复杂度提升：在低功耗模式下优化计算单元的性能，发现寄生参数建模误差较大，导致实际能效比达不到预期目标（优化目标15%，当前仅提升9%）。

四、调整后的里程碑规划

阶段	原计划完成时间	调整后完成时间	备注
计算核心优化	XX月	XX+2月	优化建模精度，提高能效比
存储架构优化	XX月	XX+1月	采用动态缓存策略
FPGA工具验证	XX月	XX+1月	提前导入Beta版本

五、结论与总结

当前AI芯片性能优化项目由于技术瓶颈和供应链风险已超期2个月，但团队已采取一系列优化措施，包括技术调整、供应链协调和测试策略优化，以确保项目在合理期限内推进。

预计在调整后的周期内完成各阶段的优化与流片，最终目标仍保持芯片性能提升15%和存储访问延迟优化

主流LLM一览

SuperCLUE测评基准2025年3月总体表现

排名	模型名称	机构	总分	数学推理	科学推理	代码生成	智能体 Agent	指令遵循	文本理解与创作
-	o3-mini(high)	OpenAI	76.01	94.74	70.00	88.78	57.14	66.40	79.01
	DeepSeek-R1	深度求索	70.34	85.96	64.00	86.94	65.18	39.52	80.41
-	Claude 3.7 Sonnet(Extended)	Anthropic	68.02	78.07	59.00	86.73	56.62	48.92	78.77
-	GPT-4.5-Preview	OpenAI	67.46	67.54	70.00	79.18	71.88	35.75	80.40
	QwQ-32B	阿里巴巴	66.38	88.60	67.00	81.84	48.66	29.92	82.27
-	Gemini-2.0-Pro-Exp-02-05	Google	65.35	65.79	70.71	77.76	64.88	33.60	79.34
	Doubao-1.5-pro-32k-250115	字节跳动	64.69	62.28	70.00	76.94	54.46	46.77	77.66
4	hunyuan-turbos-20250226	腾讯	62.49	47.37	63.00	74.49	70.09	41.13	78.88
5	DeepSeek-R1-Distill-Qwen-32B	深度求索	59.94	85.85	62.89	73.43	36.77	23.18	77.53
5	Qwen-max-latest	阿里巴巴	59.34	42.98	68.00	76.33	58.48	29.38	80.88
-	Gemini-2.0-Flash-Thinking-Exp-01-21	Google	59.27	83.33	63.00	68.16	26.34	33.60	81.16



目 录

1

统计语言模型

2

神经网络语言模型

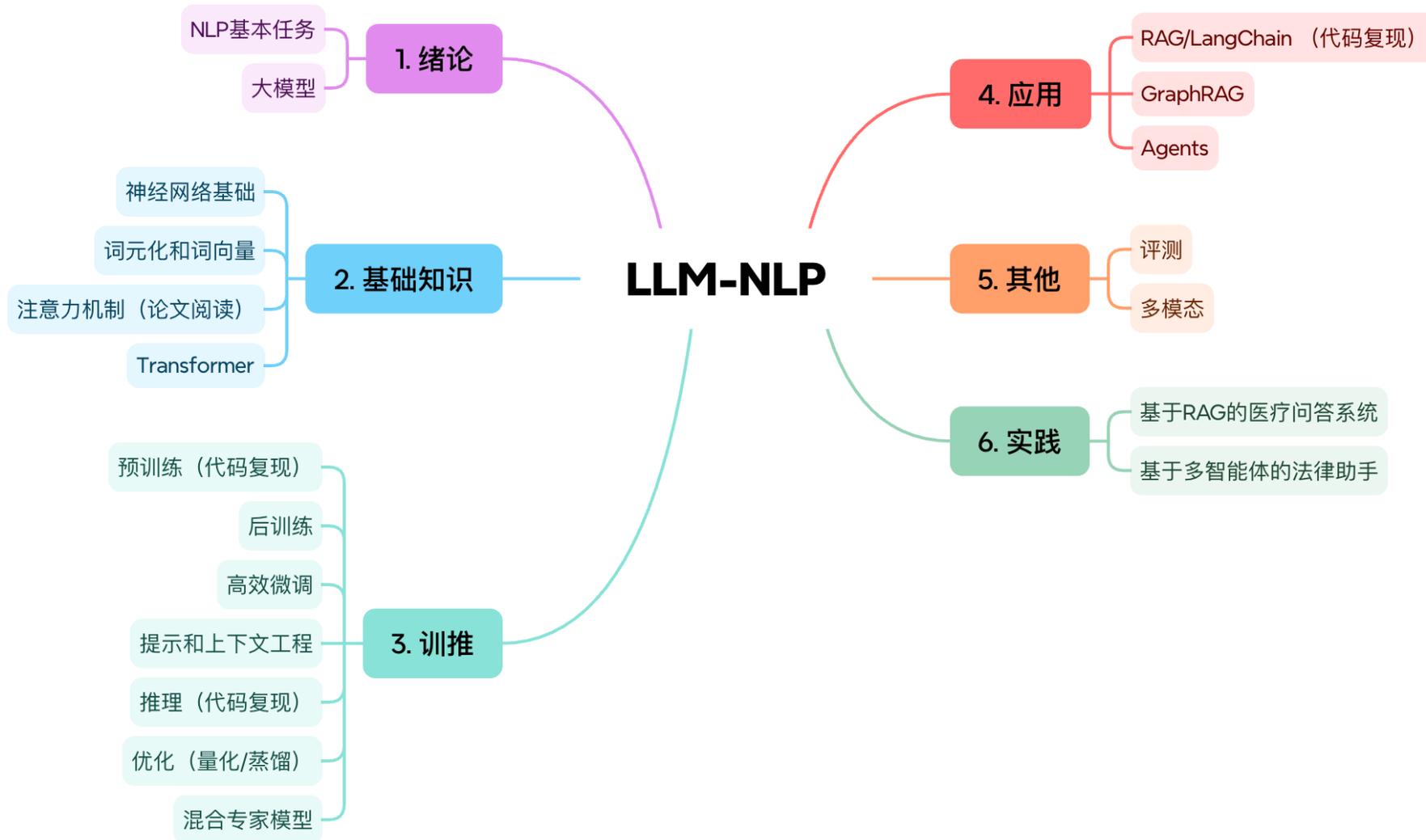
3

大语言模型

4

大模型学习路径

大语言模型NLP学习路径



本节复习

- 什么是语言模型
- 统计语言模型的思路和缺陷
- 神经网络语言模型思路
- 大语言模型的功能

课后作业

- 阅读nplm论文，并在“知乎”或其他公共平台发布论文阅读笔记
 - Bengio, Yoshua, et al. "*A neural probabilistic language model*".
Journal of machine learning research 3.Feb (2003): 1137-1155.
 - 完成时间：第3周上课前



THANKS