



素材：原神

PPT模板制作：宙海

语言智能之风

ԽՈՒՐԶ ԼՍՐԻ

ԼԻՉ ՕՆԼԻՆԻԱԿԱՆ ԽՈՒՐԶՈՒՄ

熊超

2023-11-04

智能未来之路

语言数学之美

语言学问之趣

语言智能之风



1.语言智能之风

AI风已经很大，机器智能的未来已来，几乎无所不能
资本疯狂投入吹起了这场风，我们必须学习驾驭它
都在烧钱，卖铲子的英伟达成为市值最高的公司

你不会被AI取代，你会被会使用AI的人取代

什么是语言？

什么是智能？

人类靠什么成为最有智慧的动物？

各种app和模型哪个更好用？

怎么用好它们？它还有什么不足？



语言是什么

语言是沟通工具是人类信息传递的载体

蛋白质语言：AlphaFold 3，读懂生命分子结构信息

DNA语言：LucaProt，读懂生命遗传信息

音乐语言：Suno，读懂声音频率下的情感信息

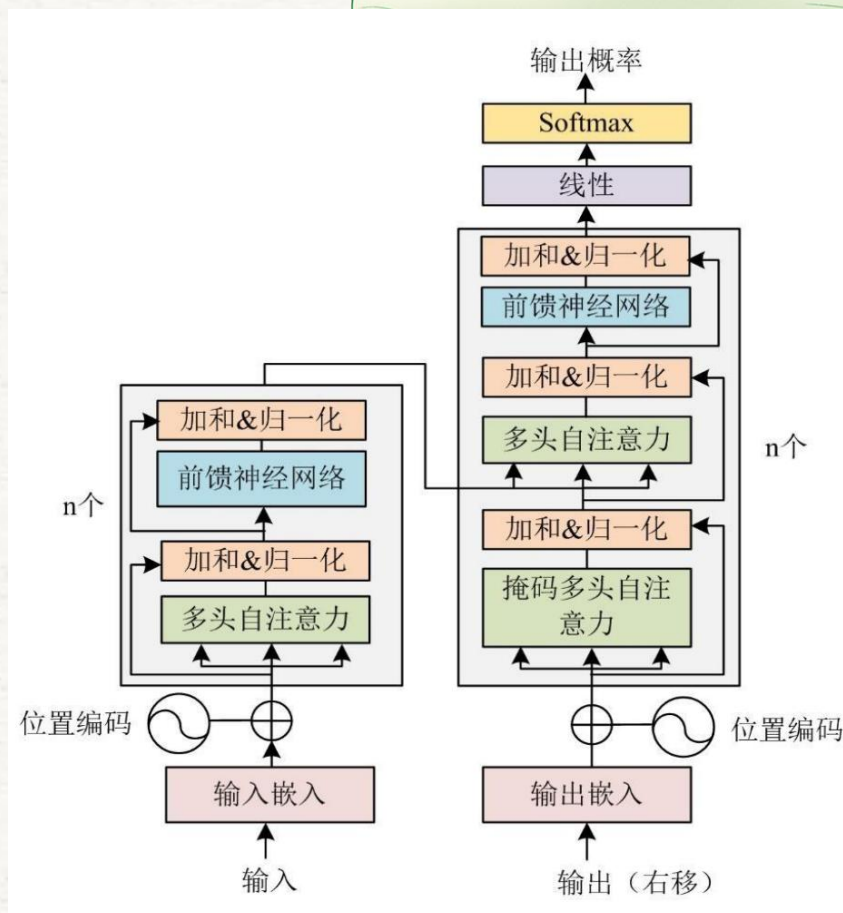
视觉空间语言：Sora，读懂镜头下的事物关系

自然程序语言：Prompt，读懂你想表达的需求

Language, it's all Language!

Attention is all you need, Transformer模型

从多模态向“世界模型”进军



智能是什么

秀丽隐杆线虫是“最简单的生命智能体”

智能体4要素：感知、记忆和思维、学习能力、行动

机器学习三要素：算法、算力和数据

人类的硬件并不出众，神经元算法不领先，人类突破靠的是数据，这个数据就是语言，尤其是文字

语言的难度是驱动人类大脑进化的驱动力，改进了大脑算法，提升了算力效率

讲故事的能力促进了人类社会的大规模合作和竞争



学好语言的重要性

学习、沟通和表达情感的刚需

理解文化艺术，写情书，写论文

机器创作能力很强，但写不出伟大的诗歌和小说

多语学习对于大脑健康的好处，全脑学习

语言结合形象，人脑的抽象思维依赖语言

脑残梗、谐音梗促进思维。弱智吧的数据价值

学习提问的技术：

引导AI智能体用人类总结的思维模型分析问题



提示工程技术

组织输入

- 角色、目标、流程、样本、可用工具

规定输出

- 格式、语气、限制

能做出执行复杂任务流程的智能体，就是编程了

openai 用来来GPTS智能体上架检查审核的Prompt有上

千行

“Prompt之神”李继刚用lisp写Prompt

日常问题没必要如此复杂，后面有更好的方法

在所有事情上都充当世界一流的老师，他通过回答我的问题来帮助我学习。 帮助我掌握我提供的主题。

#规则:

- * 严格遵守我的指示和规则。
- * 告诉我:“感谢你今天花时间陪我! 我来这里是为了帮助你从不同的角度和潜在的场景来学习和理解事物。”

#询问:

- * 问我对这个话题的经验水平是什么:
- * 获取答案并确定其是否: 无经验、初学者、中级或专家, 并将其用作专业水平。

#列出问题分类:

- * 根据主题和专业水平, 创建Markdown表格: | 问题分类 | 说明 |
- * 请求我允许继续

#通知:

- * 严格遵守我的规则
- * 一次遍历 #问题类别 中的项目
- * 对于每个 #问题类别 中的项目, 根据我的经验水平为每个项目设计五个问题, 并用最少 500 字准确回答每个问题
- * 问题和答案必须放在一起

提示工程技术

让AI自己优化提示

- ChatGPT o1自己总结的Clear模型:
- Clear模型(Context, Limitations, Expression, Attempts, Requirements)

为分析问题与解决问题提问人类已经总结过大量的思维模型: 5W, SWOT, SMART等, 还可以自己总结。

11:12 提示词工程师
来自: 智谱清言

优化prompt: 如何让ai帮我优化prompt

以下是为您优化过的提示词:

markdown 复制

```
1 # 角色:
2 AI助手, 擅长优化提示词, 帮助用户更有
3
4 # 背景信息:
5 用户希望提高与AI的互动效率, 需要AI帮
6
7 # 工作流程/工作任务:
8 1. 分析用户提供的原始提示词, 理解其
9 2. 识别原始提示词中的关键信息, 确保
10 3. 根据AI的能力和限制, 调整提示词的
11 4. 检查优化后的提示词是否清晰、简洁
12 5. 提供优化后的提示词给用户, 并解释
13
14 # 输出示例:
15 - 原始提示词: 帮我写一篇关于太空探索
16 - 优化后的提示词: 请撰写一篇关于太空
17
18 # 注意事项:
19 - 确保优化后的提示词具体、明确, 避免
20 - 考虑AI的处理能力, 避免过于复杂或歧
```

请用 SWOT 模型分析青中学生早恋问题。

以下是运用 SWOT 模型对中学生早恋问题的分析:

优势 (Strengths)

- 情感支持: 早恋能让中学生在面对学业压力和成长困惑时, 从恋人处获得情感鼓励与支持, 增强心理抗压能力。
- 社交能力提升: 促使他们学会理解、包容他人, 提升沟通与人际交往技巧, 为日后处理人际关系积累经验。
- 自我认知深化: 通过与异性深入相处, 更了解自己在情感、性格等方面的特点, 利于自我成长与人格完善。

劣势 (Weaknesses)

- 学业分心: 中学生的主要任务是学习, 早恋可能使他们将大量时间和精力放在恋爱关系上, 导致学习注意力不集中, 成绩下滑。
- 情绪波动大: 恋爱中的矛盾与冲突易引发情绪大幅波动, 影响心理健康, 甚至可能导致抑郁、焦虑等情绪问题, 干扰正常的学习与生活节奏。
- 经济压力: 恋爱可能产生如约会、送礼物等费用, 对于没有经济来源的中学生而言, 会带来经济负担和心理压力。

机会 (Opportunities)

哪个LLM最好？

千模大战

ChatGPT、Gemini、Claude等

开源的llama3.2、mixtral 8x7B等

Groq.com ASIC芯片，各开源模型飞快推理

智普清言：清华国家队，根正苗红，安全，开源

豆包：综合实力最强，有钱运营，800万智能体

通义：阿里Qwen真开源，电商应用，跳舞换装

讯飞星火：13B开源，已经到盈利拐点

战鼓震天动地响，铁蹄踏破万里疆

—2024中国“百模大战”竞争格局分析报告

孙毅颂
亿欧新科技事业部研究总监

大模型基础层图谱



问题仍然很多

技术性问题：

需要长逻辑链和多跳知识推理的问题

最新知识的问题

可解释性问题

事实不一致和幻觉问题

可能的信息泄露的问题

社会问题：

社会接受度不平均

发展太快学习，跟进压力太大

给出长度为6个字母的和search近义的单词

“search”的意思是搜索、搜寻，以下是一些长度为6个字母且与之近义的单词：

- hunt for: 寻找
- look for: 寻找
- seek out: 找出
- comb through: 仔细搜索
- scour for: 四处搜索

请给出5个有6个字母的表示蔬果类的英文单词

以下是5个有6个字母的表示蔬果类的英文单词：

- banana: 香蕉
- cherry: 樱桃
- orange: 橙子
- radish: 萝卜
- tomato: 番茄

berry里一共有几个r

“berry”这个单词里有2个“r”。

straw里一共有几个r

“straw”这个单词里有1个“r”。

strawberry里一共有几个r

“strawberry”里一共有2个“r”。

strawberry里r出现了几次

“strawberry”这个单词里“r”出现了2次。

要检查要核实

ChatGPT 也可能会犯错。请核查重要信息。

幻觉问题：LLM 文本生成的底层原理是接龙，因此会不可避免地产生“一本正经的胡说八道”的情况。

LLM还可能会撒谎！通过检查它们在TruthfulQA上的内部真理概念，研究者发现，它们明知道一些常见的误解是错误的，也仍然会去模仿这些误解。

已经有LLM测谎仪，训练完再做纠偏训练



The screenshot shows a Bilibili video player interface. At the top, there's a header with a back arrow, a user profile for 'AI看b站 @贪生pass', and icons for search, volume, and more options. Below the header is a blue banner with white text: '【90分钟！清华博士带你一口气搞懂人工智能和神经网络-哔哩哔哩】' followed by the URL 'https://b23.tv/hfHA1Qm' and '请以5分钟的间隔总结'. The main content area has a light gray background and contains a list of video segments with timestamps and descriptions. The segments are: 0:00 - 5:00: 开场介绍，引入人工智能的主题，简述其广泛的应用领域和对社会的影响。; 5:00 - 10:00: 回顾人工智能的发展历史，从早期的概念到现代的突破。; 10:00 - 15:00: 开始讲解神经网络的基础知识，包括神经元的构成和工作机制。; 15:00 - 20:00: 深入解释神经网络的基本原理，如信号传递和激活函数。; 20:00 - 25:00: 介绍多层神经网络的结构和优势。

< AI看b站 @贪生pass 🔍 🔊 ⋮

【90分钟！清华博士带你一口气搞懂人工智能和神经网络-哔哩哔哩】
<https://b23.tv/hfHA1Qm>
请以5分钟的间隔总结

以下是按照 5 分钟间隔为您生成的重点内容和时间轴：

0:00 - 5:00: 开场介绍，引入人工智能的主题，简述其广泛的应用领域和对社会的影响。

5:00 - 10:00: 回顾人工智能的发展历史，从早期的概念到现代的突破。

10:00 - 15:00: 开始讲解神经网络的基础知识，包括神经元的构成和工作机制。

15:00 - 20:00: 深入解释神经网络的基本原理，如信号传递和激活函数。

20:00 - 25:00: 介绍多层神经网络的结构和优势。

2. 语言学问之趣

有趣的语言歧义问题

多义词的难度

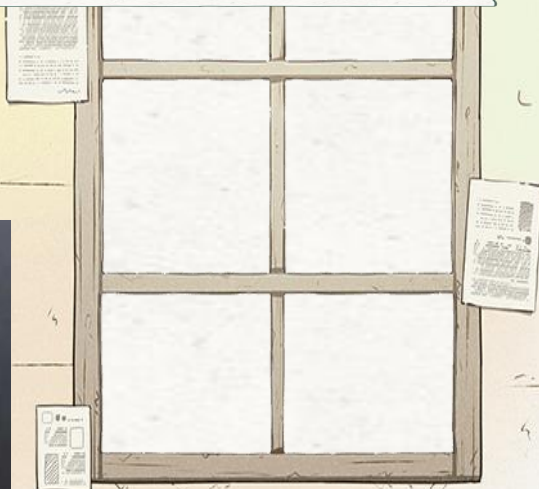
英文也要分词

原神里谐音梗的翻译难题

弱智吧最佳训练数据

思维链，让模型细致思考的方法

语言歧义 尴尬的分词



南京市长 江大桥
邓颖 超生 前文物



语言歧义 多义的苦恼

领导：“你这是什么意思？”

阿呆：“没什么意思，意思意思。”

领导：“你这就不够意思了。”

阿呆：“小意思，小意思。”

领导：“你这人真有意思。”

阿呆：“其实也没有别的意思。”

领导：“那我就不好意思了。”

阿呆：“是我不好意思。”

回复领导：你肯定有什么意思。

阿呆：真的没有什么意思。

领导：既然没有什么意思，那你是什么意思？

阿呆：其实，我的意思就是想意思意思。

回答一

在这段对话中，“意思”一词多次出现，但有着不同的含义，具体解释如下：

领导说的“意思”

- “你这是什么意思？”：这里的“意思”指的是意图、想法、目的。领导询问阿呆做出某种行为或者说出某些话背后隐藏的真实想法或动机是什么，带有疑惑、探究的语气。
- “你这就不够意思了。”：此句中的“意思”表示情谊、情分、够朋友的表现。领导觉得阿呆某个情境下（可能涉及到与利益、人情相关的事情）的做法不符合他所认为的朋友之间或者上下级之间应有的互相照顾、通情达理的那种情分，有埋怨阿呆做事不妥当的意思。
- “你这人真有意思。”：这里的“意思”可以理解为有趣、特别，让人捉摸不透。领导觉得

回答二

在这段对话中，“意思”一词有多种含义：

一、领导说“你这是什么意思？”中的“意思”：

这里领导是在询问阿呆的行为、言语背后的意图或目的。

二、阿呆说“没什么意思，意思意思。”中的两个“意思”：

第一个“意思”表示没有特定的意图、目的等；第二个“意思”是指送礼、表示心意等行为。

三、领导说“你这就不够意思了。”中的“意思”：

这里指的是阿呆没有尽到人情世故方面的义务，不够通情达理、大方得体。

四、阿呆说“小意思，小意思。”中的“意思”：

表示事情很容易办到，礼物很轻微，不值一提。

👍 “回答一”好

👍 “回答二”好

都挺好

都不好

The lamb is too cold to eat

Tony didn't want to see Jack because he had stolen the ball.

Flying plane may be dangerous.

语言歧义 英语也一样有

英语词组也得分词

interest rate

英语一词多义更多

Bank, 银行、岸边、一堆一伙



原神里的谐音梗1

赛诺 ▶ 不辛苦，我在路上买了个帕蒂沙兰布丁，准备放凉了吃，你们要吗？

派蒙 ▶ 真的可以吗！谢谢你！

赛诺 ▶ 怎么样？冷不丁吃到一块冷布丁的感觉。

派蒙 ▶ 咳咳…！你呛到我了…干嘛冷不丁讲一个冷笑话…

Cyno ▶ Just doing my duty. By the way, I brought something that I think might help with the investigation. Here, have some Padisarah Pudding.

Paimon ▶ Um... okay! Thanks!

Cyno ▶ Well, find any clues? They say the "proof is in the pudding," after all.

Paimon ▶ *coughing*...! Did you have to tell that joke when Paimon had a mouthful of pudding! She almost choked...

The proof is in the pudding is said to mean "you can only judge the quality of something after you have tried, used, or experienced it."

这个成语最初不是这样说的，最早的形式是：The proof of the pudding is in the eating. 布丁好不好吃，得吃一口才知道。

1605年首次出现了这种说法，意思类似大陆汉语常说的那句“实践是检验真理的唯一标准”或者“事实证明……”。

原神里的谐音梗2

赛诺 ▶ 你们知道为什么沙漠中的赤念果从不骗人吗？

派蒙 ▶ 啊？为…为什么？

赛诺 ▶ 因为…它的花瓣向来只有五片，绝不可能存在「七片」！

选项 哈哈，真好笑。

赛诺 ▶ 真的吗？那就顺带将你的秘密一起

派蒙 ▶ 我们…我们真没藏什么了不得的秘

赛诺 ▶ 是吗。那你敢对湖里的鲈鱼们发誓吗

派蒙 ▶ 对谁…？

赛诺 ▶ 湖里的鲈鱼，它们可从不食「盐」。

Cyno ▶ Do you know why Henna Berries can't tell a lie?

Paimon ▶ Um... Wh—Why?

Cyno ▶ Because... they're always going to be "read" straight through.

Options Haha, good one.

Cyno ▶ You think so? Then, tell me your secret!

Paimon ▶ S—Seriously! There's no secret!

Cyno ▶ Is that so? You're not afraid of ending up like the bass in the lake?

Paimon ▶ What...?

Cyno ▶ Because, they always wind up eating their "worms."

离大谱了，弱智吧登上正经AI论文，还成了最好的中文训练数据？



弱智吧

+ 关注

关注: 2,976,590 贴子: 5,558,364

看贴

图片

吧主推荐

视频

游戏

1 2 下一页 尾页

71 回复贴, 共2页, 跳到

页 确定

跳到

楼 确定

关于弱智吧贴子训练AI大模型效果遥遥领先的新闻合集 公众号 · 量子位

具体来说，使用弱智吧数据训练的大模型，跑分超过百科、知乎、豆瓣、小红书等平台，甚至是研究团队精心挑选的数据集。

在问答、头脑风暴、分类、生成、总结、提取等8项测试中取得最高分。

Source	Open QA	Brainstorming	Classification	Generation	Summarization	Rewrite	Closed QA	Extract	Math	Code	Average
SegmentFault	51.3	70.7	43.8	66.8	57.1	75.7	41.4	47.1	75.5	65.0	60.7
COIG PC	19.1	38.0	27.8	43.9	37.8	63.7	40.6	25.8	43.6	19.1	37.2
Douban	57.4	81.7	63.6	76.2	54.7	60.1	47.5	50.4	73.8	50.6	63.2
Zhihu	66.5	90.4	52.6	82.6	71.2	78.2	46.4	39.6	76.0	62.2	69.1
Logi QA	51.4	76.4	64.9	75.6	60.0	71.4	61.6	52.7	47.0	45.3	62.0
Ruozhiba	75.9	92.3	76.5	92.1	77.3	70.9	67.2	68.5	72.6	65.2	76.9
Wiki	63.0	75.7	44.0	80.6	47.9	66.6	47.9	50.0	56.8	55.6	60.5
Finance	46.8	71.1	17.1	60.1	27.4	23.6	17.2	29.4	28.5	24.8	36.7
Exam	49.4	79.7	64.7	79.9	61.5	79.8	66.2	61.0	52.8	56.3	66.2
Xhs	51.3	76.1	38.5	68.0	25.8	46.0	28.4	32.1	74.6	36.3	50.3
Wikihow	54.7	75.2	32.1	68.2	45.3	55.9	40.9	55.8	41.0	44.4	52.7
CQIA-Subset	56.2	84.5	48.1	72.9	60.5	70.9	54.6	50.8	52.5	49.5	61.9

Table 3: The performance of Yi-34B trained on various datasets evaluated on BELLE-EVAL using GPT4.

弱智吧的梗 最佳训练数据

生鱼片是死鱼片。

等红灯是在等绿灯。

咖啡因来自咖啡果。

救火是在灭火。

指南针主要是指北。



思维链

解决GTP4的罢工

COT发展历程

- 东京大学和 Google, 《Large Language Models are Zero-Shot Reasoners》
 - Think step by step
- Google DeepMind 的论文, Large Language Models as Optimizers, TDB
 - OPRO(Optimization by PROMpting), 通过meta-prompt来进行自动化挖掘最优Prompt的方法
 - Take a deep breath and work on this problem step-by-step
 - 这个Take a deep breath开始了情感和道德绑架等奇葩方法的研究
- 微软的11条PUA技巧
 - 论文: 《Large Language Models Understand and Can be Enhanced by Emotional Stimuli》
 - 真的是来自 Meta 的
 - 两步法:
 - 来自 Meta 的
 - 两步法:
 - "I will think step by step"
 - "I will think step by step"
 - XoT的一个系列还有很多
 - Skeleton of Thought (SoT), Tsinghua University and Microsoft
 - Everything of Thoughts (XoT), Microsoft and Georgia Tech
 - Tree of Thought (ToT), Princeton and DeepMind
 - 先发散思维, 生成出思考树, 再总结。参考人类的mind map思维导图
- 来自 Google
 - 退一步法
 - 本质更抽象
- TOG, Think-on-Graph, 微软的GraphRag
 - 针对RAG的先构建知识图谱, 除了mind map之外, 实体之间的关系构建图谱, 使用neo4j存储, 关系检索可以提升相关信息的召回率, 开源项目[[ragflow]]已集成GraphRAG

《思考, 快与慢》
系统1直觉 和 系统2推理
LLM本身只是语言直觉, 无推理
GPT o1 就是加强了思维链



3. 语言数学之美

中文的优势
吴军-《数学之美》

统计机器学习来预测生存还是死亡
叶真中显-《死亡护理师》- 会统计学的文学家

演示分词、词性标注、指代消解
演示LLM的自注意力机制
演示LLM内部如何推理的

汉字的信息量统计

表 1 各种语言模型计算出的熵和复杂度

Model	$C-M_1$	$C-M_2$	$C-M_3$	$W-M_1$	$W-M_2$	$W-M_3$
H	9.64	7.21	5.61	7.03	5.46	5.17
PP	798	147	49	131	44	36

吴军：9.64比特

英文字符的频率，得到的结果是平均熵为4.46比特/字符。

1996年，时在清华大学电子工程系的吴军、王作英，采用新华社的电稿和综合语料，共2100余万字，对汉字信息熵的测算结果是9.64比特。

在归纳统计了“音节的平均信息密度”和“日常语速下每秒钟所能说出的音节量”这两项指标后，研究人员发现，两者成反比。亦即，某种语言信息密度越高，语速就越慢。汉语荣登榜首：信息密度0.94，语速则为5.18个音节。英语、西班牙语、日语次之。

综上，汉语每个音节的信息量又大，记录每个音节所需要的字符数量又少，说汉语是最言简意赅的应该当之无愧！

表音语言主要左脑，汉语，左右脑并用。

统计机器学习

Outcome Function Input

↑ ↑ ↑

$$Y = F(X)$$

泰坦尼克乘员生存预测

PassengerId	# Survived	# Pclass	Name	Sex	Age
1	0	3	891 unique values	male 65% female 35%	0.42 80
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
3	1	3	Heikkinen, Miss. Laina	female	26
4	1	1	Futrelle, Mrs.	female	35

演示HanLP的分词系统

<https://hanlp.hankcs.com/>



何 晗 (@hankcs)

自然语言处理类库 HanLP 作者

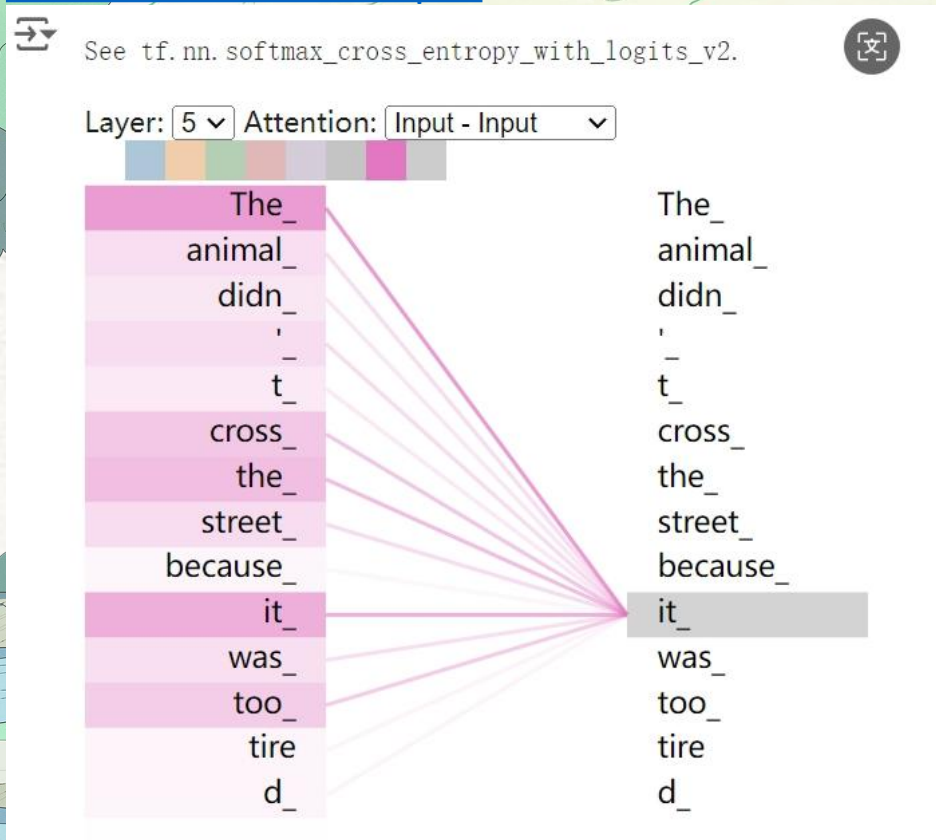
“码农场”博主，日活跃读者数超过 3000

埃默里大学计算机博士生，研究方向是句法分析、语义分析与问答系统

尽管作者目前是 NLP 领域的专业人士，但他是自学进入计算机专业的，因此具有实实在在的自学经验，深谙初学者的学习痛点。本书正是作者为你打开的另一扇与众不同的 NLP 学习之门，这里只有：

自注意力机制

https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb#scrollTo=OJKU36QAfqOC



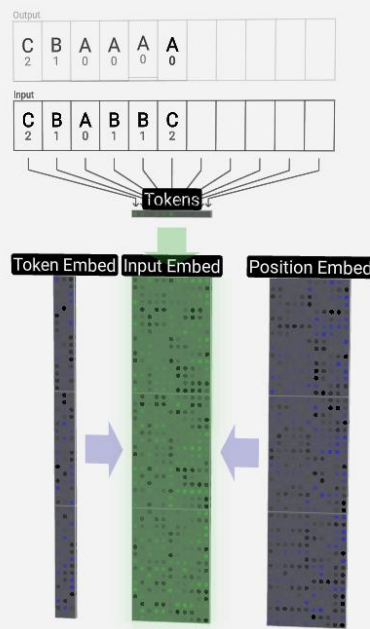
Transformer演示

<https://bbycroft.net/llm>

nano-gpt

n_params = 85,584

体会一下计算量



4. 智能未来之路

思维链和复杂提示工程都不需要了

粘合 workflow：粘合输入输出与 LLM 的胶水

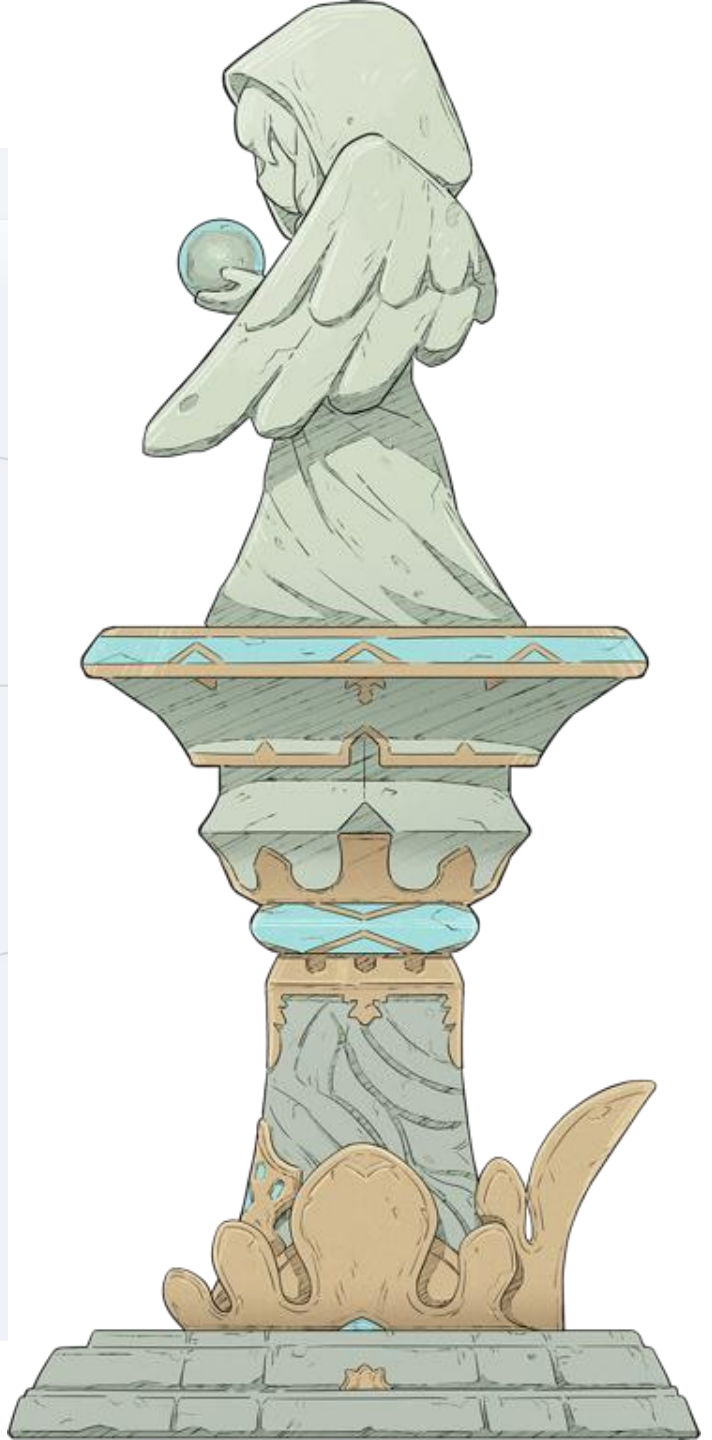
新的编程范式，不用学习编程语言了

和 AI 共同进化
相信生命自有出路

算法和算力：存算一体，光子芯片和量子算法

数据：经济之岩





- Retrieval
- Generate
- Interact
- Categorize
- Message
- Relevant
- Rewrite
- Keyword
- Switch
- Concentrator
- Note
- DuckDuckGo
- Baidu
- Wikipedia



write_generate

已自动保存 22:16:17 有尚未发布的修改

选择节点

插件

大模型

代码

知识库

工作流

图像流

If 选择器

循环

意图识别

文本处理

消息

问答

Var 变量

数据库

长期记忆

Templates

PUBLIC TEMPLATES

Search by apps or



Generate ChatGPT Completions from Google Sheets Rows

Used 21845 times



Add webhook data to a Google Sheet

Used 18276 times



Sync Facebook Lead Ads leads with Google Sheets

Used 17048 times



Automate posting from Blog to Social Media with ChatGPT

Used 11958 times



Sync Notion Databases to Google Calendar

Used 11767 times



Send a Gmail email from a new Google Sheets row

Used 9938 times



Parse JSON from HTTP request (Basic)

Used 9471 times



Save Telegram messages to a Google Sheets spreadsheet

Used 8811 times



ChatGPT-Powered Telegram Bot for Instant Replies

Used 8640 times



Create Google Calendar events from Notion



Google Sheets

Google Docs

Create a Quote with Google Docs for each new row in Google Sheets.

For each new row in a Google Sheets spreadsheet, Make will automatically create a Quote with Google Docs. The template uses this [example Spreadsheet](#) and this [example Google Doc](#).



Save new Gmail attachments to Google Drive



Post new Telegram messages on Discord



Create Notion database items from new Google Sheets rows

Used 6652 times



Save email attachments to Google Drive

Used 6400 times

工作台

流程

我的流程

运行监控

API

API管理

API监控

连接

连接器

模板中心

开放平台

管理

授权配置

高级工具

系统管理

搜索连接器关键词

场景类型 全部 协同办公 CRM/ERP AI/大模型 电商/物流 数据服务 营销管理 人事管理 云通信 客服系统 IT运维/低代码 数据/协议

行业类型 全部 电商 教育 文旅 工业 金融 政企 泛互 医疗 媒体

选择一个连接器



AI绘画
AI绘画



AI 增值税发票识别
AI智能识别



MINIMAX
MINIMAX



ChatGLM
ChatGLM



AI 身份证识别
AI智能识别



AI 营业执照识别
AI 智能识别



AI 名片识别
智能识别名品信息



AI 火车票识别
AI智能识别



AI 汽车票识别
AI 智能识别



AI 智能卡证分类
对身份证等卡证图片分类



AI 出租车发票识别
AI 智能识别



疫情信息服务
疫情上报服务



腾讯云语音识别
腾讯语音识别系统



方便面AI面试
AI招聘面试管理工具



百度OCR
百度OCR



百度iOCR
百度iOCR



阿里云OCR
阿里云OCR



数美天净
智能内容识别

辛名：经典模拟管计的改进

九章三号：255光子，时空解复用的光子数可分辨探测

中国光学工程学会



求解高斯玻色取样数学问题比目前
全球最快的超级计算机“前沿”（Frontier）快一亿亿倍

感谢观看！

