



生物信息学

课程相关信息

课程相关信息



□ 教师：薛宇，东校区东11楼618#

✿ E-mail: xueyu@hust.edu.cn

✿ 上机：彭迪，pengdi@hust.edu.cn

✿ 助教：杨辰宇，656453094@qq.com

□ 课时安排：

✿ 课堂授课（40学时）+ 上机（16学时）

✿ 两次平时作业

□ 考试成绩：

✿ 总分 = 试卷（60%）+ 上机（30%）+ 平时（10%）

请准备一个作业本！

课件下载



□ 2025年课件：

- ✿ <https://xue.biocuckoo.cn>

- ✿ 统一复印后分发给同学（各班统计人数给助教，2.26发课件）

□ 视频课件（2020年）

- ✿ 微助教（PDF课件，PPTX带音频课件）

- ✿ <https://www.teachermate.com.cn/classes/1131882>

- ✿ 哔哩哔哩（MP4视频课件）

- ✿ <https://space.bilibili.com/677120859>

课件参考资料



- ❑ **MIT OCW: Foundations of Computational and Systems Biology**

<http://ocw.mit.edu/courses/biology/7-91j-foundations-of-computational-and-systems-biology-spring-2014/index.htm>

- ❑ **Ben Langmead's Teaching Materials**

<https://www.langmead-lab.org/teaching.html>

- ❑ **Proteomics Informatics Spring 2024, Dr. David Fenyő**

<https://fenyolab.org/teaching/>

课程章节概述



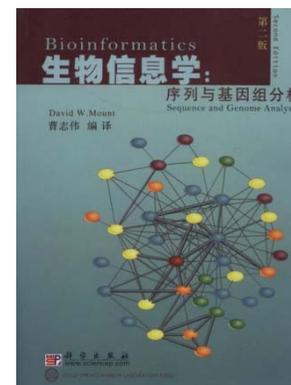
- 一、历史简介和概论
- 二、生物序列获取和存储
- 三、生物序列数据检索
- 四、机器学习和数学基础
- 五、基因组分析
- 六、双序列比对▲
- 七、多序列比对
- 八、序列模式识别
- 九、分子进化与系统发育分析
- 十、测序读段回贴
- 十一、转录组与转录调控分析
- 十二、计算蛋白质组
- 十三、结构生物信息学
- 十四、基因组组装

教材及参考书



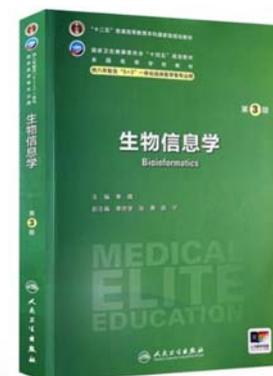
□ 主要教材：

- ❁ 生物信息学：序列与基因组分析（第二版），David Mount, 2006，科学出版社



□ 参考书：

- ❁ 生物序列分析：蛋白质和核酸的概率论模型，Richard Durbin等，清华大学出版社
- ❁ 生物信息学（第3版），李霞等主编，2016，人民卫生出版社
- ❁ 统计学习方法，李航，清华大学出版社



其他推荐资料



- ❑ **Python for bioinformatics, Sebastian Bassi, CRC Press**
- ❑ 分子进化与系统发育, [美] Masatoshi Nei, Sudhir Kumar, 高等教育出版社
- ❑ 生物统计学基础, [美] 伯纳德·罗斯纳, 科学出版社
- ❑ 结构生物信息学, [美] P.E.波恩 & H. 魏西希, 化学工业出版社
- ❑ 生物信息学与功能基因组学, [美] 乔纳森·佩夫斯纳, 化学工业出版社
- ❑ 生物信息学 (第二版), 陈铭 等译, 科学出版社
- ❑ 机器学习, 周志华 著, 2016, 清华大学出版社
- ❑ 深度学习, [美] 伊恩·古德费洛 等, 人民邮电出版社
- ❑ **Beginning Perl for bioinformatics, James Tisdall, O'Reilly**

网上课程 (1)



□ 生物信息学：导论与方法（2025）

🌸 教师：高歌、魏丽萍，北京大学

🌸 <https://www.coursera.org/learn/sheng-wu-xin-xi-xue>

为个人 为商务 为大学 为政府

coursera 探索 ▾ 您想学习什么? 🔍 在线学位 工作机会 登录 [免费加入](#)

🏠 > 浏览 > Health > Health Informatics

 **北京大学**
PEKING UNIVERSITY

生物信息学: 导论与方法

 位教师: [Ge Gao 高歌, Ph.D.](#) 另外 1 个

[免费注册](#)
于 Feb 10 开始 有助学金

20,574 人已注册
包含在 [coursera PLUS](#) 中 · [了解更多](#)

14个模块 深入了解一个主题并学习基础知识。	4.8 ★ (189 条评论)	需25小时完成 每周8小时, 共计3周	灵活的计划 自行安排学习进度
----------------------------------	-----------------------------	-------------------------------	--------------------------

网上课程 (2)



□ 【2025版】李宏毅机器学习系列课程

🌸 <https://www.bilibili.com/video/BV1YsqSY8EiW>

【2025版】李宏毅机器学习系列课程! 涵盖机器学习, 深度学习、神经网络算... ▾

📺 12.2万 🗨️ 429 🕒 2024-12-11 17:23:32

先

Introduction of
Machine / Deep Learning

Hung-yi Lee 李宏毅

投票

网上课程 (3)



□ 吴恩达机器学习 (2024)

🌸 <https://www.bilibili.com/video/BV1Bq421A74G>

(超爽中英!) 2024公认最好的【吴恩达机器学习】教程! 附课件代码 Machine ...

▶ 132.1万 🗨️ 8846 2024-03-26 12:19:08

吴恩达机器学习 bilibili

Stanford
ONLINE

DeepLearning.AI

Machine Learning
吴恩达机器学习 (中文版)

Welcome!
附课件源码

88人正在看, 已装填 143 条弹幕

请先 登录 或 注册

弹幕礼仪 > 发送



网上课程（4）

□ Keras深度学习快速简明教程

🔗 <https://www.bilibili.com/video/BV1gE411R7jd>

Keras深度学习快速简明教程 最易学的深度学习入门课程 人人都可以学的人工智能入门

4.1万播放 · 222弹幕 2019-09-30 11:00:50



课外延伸阅读



- ❑ 人工智能简史，尼克，人民邮电出版社
- ❑ 生物信息学札记，樊龙江
- ❑ 女士品茶，萨尔斯伯格，中国统计出版社
- ❑ 生命是什么，[奥] 埃尔温·薛定谔，湖南科学技术出版社
- ❑ 自私的基因，道金斯，中信出版社
- ❑ 世界观：科学史与科学哲学导论，理查德·德威特，电子工业出版社
- ❑ 生命科学史，[美] 洛伊斯·N·玛格纳，上海人民出版社
- ❑ 生命的语言，[美] 弗朗西斯·S·柯林斯，湖南科学技术出版社
- ❑ 生命的线索，约翰·苏尔斯顿 等，中信出版社
- ❑ 师从天才，[美] 罗伯特·卡尼格尔，上海科技教育出版社



DeepSeek: 思维链推理



DeepSeek手机App

← × 搜索

DeepSeek
9.2 MB · 9,914万次安装
AI智能对话助手

打开

大模型

 原文心一言 文小言	 通义	 腾讯元宝
 DeepSeek	 2025 纳米AI搜索	 R1高速专线 问小白

二 简要介绍一下生物信息学

简要介绍一下生物信息学

已深度思考 (用时 16 秒)

嗯，用户让我简要介绍一下生物信息学，我需要先理清这个领域的基本概念。首先，生物信息学结合了生物学、计算机科学和信息技术，处理和分析生物数据。不过，用户可能已经知道这些基础，我需要深入一点，但又要保持简洁。

接下来，我应该考虑生物信息学的主要应用，比如基因组学、蛋白质组学，还有具体的应用场景，比如基因测序、药物研发。可能还要提到常用的技术，比如数据挖掘、机器学习，这些技术如何帮助解析复杂的生物数据。

用户可能想知道生物信息学的重要性，特别是在现代医学和生物技术中的作用，比如精准医疗、疾病研究。这时候需要举一些例子，比如癌症基因组分析，或者COVID-19中的病毒变异追踪，这样更具体。

还要注意生物信息学的工具和数据库，比如GenBank、PDB，这些是基础资源，可能用户会感兴趣。另外，挑战部分也不能少，比如数据量增长带来的计算压力，跨学科合作的困难，这能展示领域的全貌。

开启新对话

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

atics, 2025, HU

DeepSeek vs. BAT



现在的大模型里，假如你Deepseek是100分，你给百度的文心一言，阿里的通义千问，腾讯的腾讯混元，各打多少分？



已深度思考 (用时 19 秒) ^

嗯，用户问的是现在的大模型里，如果我Deepseek是100分的话，给百度的文心一言、阿里的通义千问、腾讯的腾讯混元各打多少分。首先，我需要理解用户的需求。他们可能想了解国内这几个主流大模型之间的相对性能，或者是想做一个比较，看看哪个更适合他们的使用场景。

接下来，我得考虑评分标准。用户没有明确说具体从哪些方面评分，比如技术实力、应用场景、用户体验、市场表现等等。可能需要综合考虑多个因素。但通常来说，技术能力、模型效果、生态布局、商业化程度这些可能是比较重要的指标。

然后，我需要回忆一下这些模型的最新动态。比如，文心一言已经迭代到4.0版本，可能在中文理解和多模态方面有优势。通义千问在阿里云的支持下，可能在企业服务方面比较强。腾讯混元可能依托腾讯的社交数据，在对话和内容生成上有特色。不过具体的性能数据可能不公开，所以需要依赖公开信息和用户反馈。

开启新对话



informatics

在评估大模型时需要综合考虑技术能力、应用生态、商业化落地和用户体验等多个维度。根据当前公开信息和行业观察，我的初步评分如下（仅供参考）：

1. 百度文心一言 (85-90分)

优势：中文语义理解国内领先（尤其古诗词生成）、多模态能力突出（文生视频ERNIE-ViLG）、搜索数据优势明显
减分项：复杂逻辑推理能力待提升、代码生成能力弱于头部模型

2. 阿里通义千问 (80-85分)

优势：企业服务场景整合能力强（钉钉/阿里云深度集成）、多语言能力突出（支持近百种语言）、云计算基础设施支撑
减分项：C端用户感知度较低、创造性内容生成相对保守

3. 腾讯混元 (75-80分)

优势：社交场景适配度高（微信生态融合）、内容安全机制完善、工程化部署效率突出
减分项：技术透明度较低、学术论文产出相对较少、多模态能力待验证

补充说明：

开启新对话



DeepSeek 部署教程



📄 <https://xiaoyi.vc/deepseek>

主页 软件商店 每天必看 软件 壁纸 互联网 游戏 技巧 WP建站 **DeepSeek专题**

- [Cursor](#) : [访问](#)
- [POE](#) : [访问](#)
- [国家超算互联网](#) : [访问](#) | [介绍](#)

本地部署 DeepSeek 教程

- [本地部署「DeepSeek」模型硬件配置要求](#)
- [DeepSeek 开源模型 Qwen 与 Llama 区别，以及量化介绍](#)
- [超简单！3 步部署国产 AI 神器 DeepSeek 到你的电脑](#)
- [一行代码！手机秒变 AI 神器：DeepSeek 离线部署指南](#)
- [英特尔酷睿 Ultra 本地运行 DeepSeek：Flowy AI PC](#)

DeepSeek 本地大模型整合包

- [免部署即用 DeepSeek 全量本地模型库 + LM Studio](#)
- [DeepSeek 本地大模型整合包下载！开箱即用，无需部署](#)

安装Ollama



部署教程

1. 安装 Ollama 免费开源在本地轻松运行和部署大型语言模型的工具，并且支持 macOS、Linux 和 Windows，以及 Docker 容器化部署。用户还可以通过 Modelfile 修改模型参数，或基于现有模型微调。

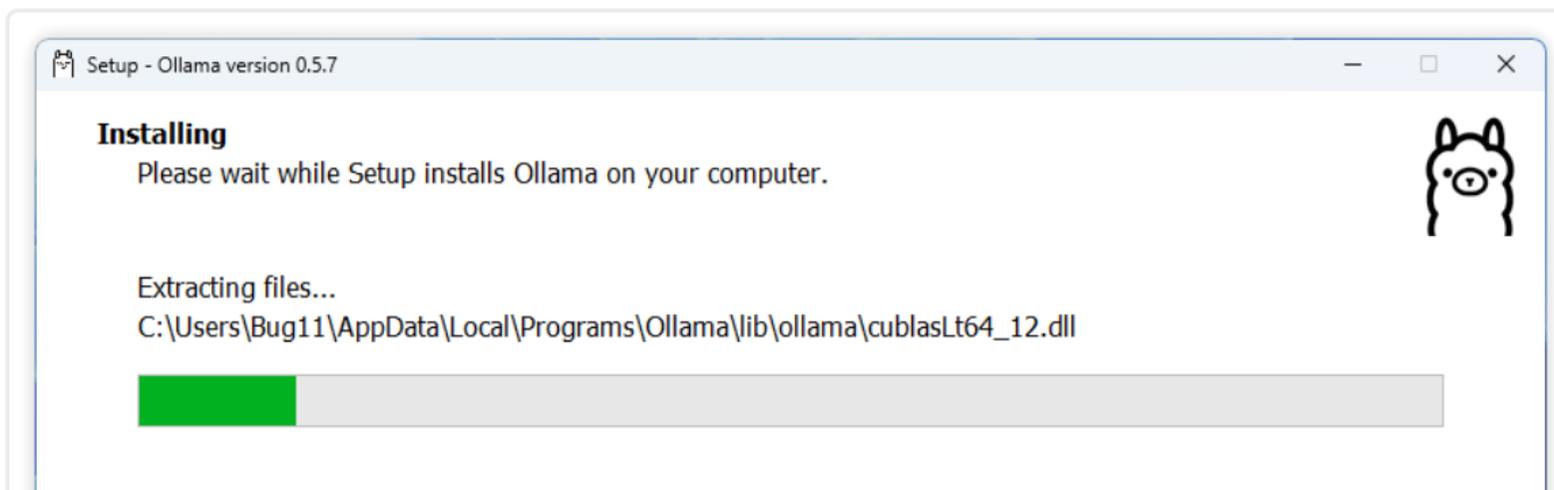
- 网盘下载：

<https://pan.quark.cn/s/a91827e20d18>

- 官方网站：

<https://ollama.com>

使用 Ollama 建议你的电脑有 8GB 内存以上（7B 模型），下载完毕直接双击运行安装即可。



安装DeepSeek-R1



❑ ollama run deepseek-r1:1.5b

2.安装使用 DeepSeek-R1 模型，在终端输入命令：`ollama run deepseek-r1:7b`，然后等待模型下载即可。

```
管理员: C:\WINDOWS\System32\WindowsPowerShell\v1.0\powershell.exe
PS C:\Windows\System32> ollama run deepseek-r1:7b
pulling manifest
pulling manifest
pulling manifest
pulling manifest
pulling manifest
pulling 96c415656d37... 27% ██████████ | 1.3 GB/4.7 GB 44 MB/s 1m16s_
```

@下一个好软件
XIAOYI.VC //
///// X16.LA

需要注意的是支持可选模型规格参数。PC 本地且推理能力较强推荐用 1.5b, 7b, 8b, 14b 模型。你也可以通过添加参数的方式来下载其它模型，例如命令：`ollama run deepseek - r1:8b`。具体可以到 Ollama 网站的模型里面查看。

UI界面 - ChatBox



The screenshot displays the ChatBox application interface. On the left is a sidebar with a list of chat sessions: 'Just chat' (starred), 'Image Creator (E...)', '做图表', '贪吃蛇(Artifact Example)', '助手咨询', 'Translator (Example)', 'Software Developer (E...)', 'Social Media Influence...', 'Travel Guide (Example)', and 'Markdown 101 (Examp...)', along with options for '新对话', '新图片', '我的搭档', '设置', and '关于(1.9.8)'. The main chat window is titled 'Just chat' and contains two messages. The first message is a system response: '我是一个AI助手, 由中国的深度求索 (DeepSeek) 公司独立开发, 我清楚我的身份与使命, 并且全心全意为人类智能的提升而努力。' with metadata 'word count: 49, tokens used: 175, model: Ollama (deepseek-r1:1.5b)'. The second message is a user query: '什么是生物信息学?' with metadata 'word count: 8'. Below it is a 'Thinking' indicator and a detailed response: '我是一位专注为人类设计 smarterAI助手的DeepSeek-R1-Lite-Preview, 很高兴为您提供专业的生物信息学解释。关于生命科学领域的重要研究模型和分析方法论, 请参考官方文档。' with metadata 'word count: 143, tokens used: 357, model: Ollama (deepseek-r1:1.5b)'. The interface includes a search bar, a refresh button, and a bottom input field with the placeholder text '在这里输入你的问题...'. The model name 'deepseek-r1:1.5b' is visible in the bottom right corner.



一键部署DeepSeek

☐ <https://deepseek.flashai.com.cn/>

FlashAI × DeepSeek

本地大模型整合包

一键部署、开箱即用的本地AI解决方案

- 一键部署**
无需任何配置，下载解压即用
- 本地运行**
数据安全，完全离线运行
- 集成图形界面**
畅快聊天，告别命令行
- 内置知识库**
上传自己的文档，让大模型分析

软件界面预览

机器学习的定义



- “如果一个计算机程序在某些任务类别 T 上根据性能衡量标准 P 测得的性能，能够随着经验 E 的积累而提高，那么就说该程序能从经验 E 中学习”
- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

——Tom Mitchell, 1997

机器推理的定义



- “利用先前获得的知识进行代数运算以回答新问题”
 - ✿ aims to build interpretable AI systems that can solve problems or draw conclusions from what they are told (i.e. facts and observations) and already know (i.e. models, common sense and knowledge) under certain constraints
 - ✿ 逻辑推理、概率推理

From Machine Learning to Machine Reasoning

Leon Bottou

A plausible definition of "reasoning" could be "algebraically manipulating previously acquired knowledge in order to answer a new question". This definition covers first-order logical inference or probabilistic inference. It also includes much simpler manipulations commonly used to build large learning systems. For instance, we can build an optical character recognition system by first training a character segmenter, an isolated character recognizer, and a language model, using appropriate labeled training sets. Adequately concatenating these modules and fine tuning the resulting system can be viewed as an algebraic operation in a space of models. The resulting model answers a new question, that is, converting the image of a text page into a computer readable text.

This observation suggests a conceptual continuity between algebraically rich inference systems, such as logical or probabilistic inference, and simple manipulations, such as the mere concatenation of trainable learning systems. Therefore, instead of trying to bridge the gap between machine learning systems and sophisticated "all-purpose" inference mechanisms, we can instead algebraically enrich the set of manipulations applicable to training systems, and build reasoning capabilities from the ground up.

机器学习 vs. 机器推理



Machine learning

“刻画数据的隐含特征”

Machine learning can process large volumes of data and capture the hidden patterns needed to effectively predict outcomes. It tackles a pre-determined problem, with clear inputs and expected outputs.

Machine reasoning

“利用类似人的常识来分析”

Machine reasoning can be seen as an attempt to implement abstract thinking as a computational system and apply human-like common sense to analyze and translate vast knowledge and learned network data into clear explainable insights. It does that by providing more contextual knowledge, concepts, and rules by which systems can obey, and from which they can start creating a model of the world around them.

<https://www.ericsson.com/en/blog/2022/4/differences-between-machine-learning-machine-reasoning>

思维链示例：逐步解答



❑ Input prompt: Solve the quadratic equation: $x^2 - 5x + 6 = 0$



Solve the quadratic equation $x^2 - 5x + 6 = 0$



As a text-based AI, I am unable to directly solve quadratic equations. However, I can guide you on how to solve it using various methods. One common method is factoring. Let's try factoring the given quadratic equation $x^2 - 5x + 6 = 0$:

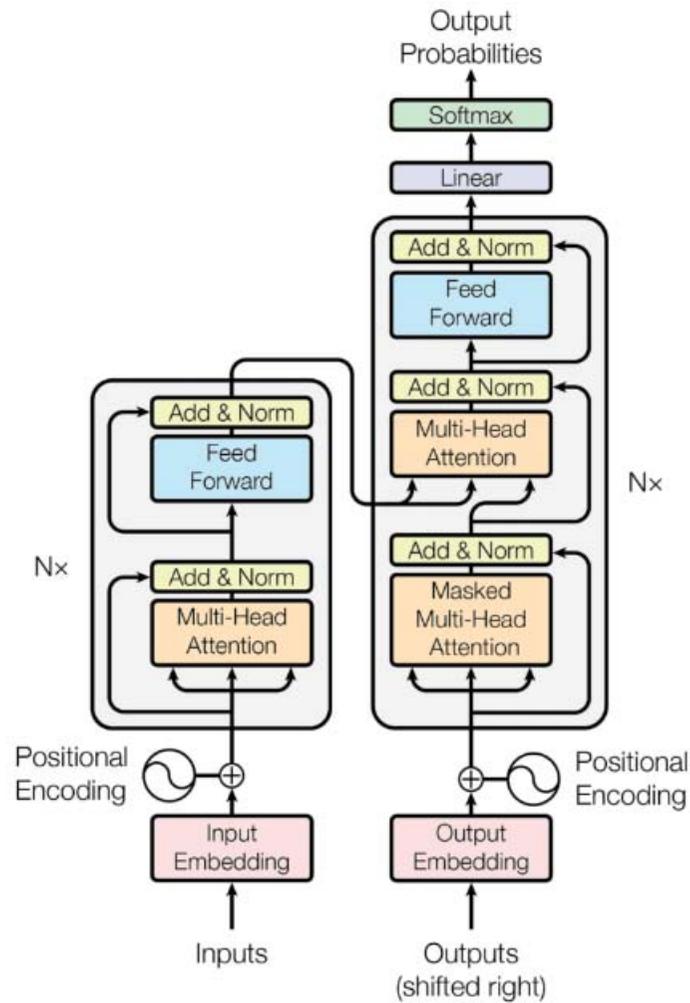
1. Factor the left side of the equation: $(x - 2)(x - 3)$
2. Set each factor equal to zero and solve for x :
 - $x - 2 = 0 \Rightarrow x = 2$
 - $x - 3 = 0 \Rightarrow x = 3$

So, the solutions to the quadratic equation $x^2 - 5x + 6 = 0$ are $x = 2$ and $x = 3$.

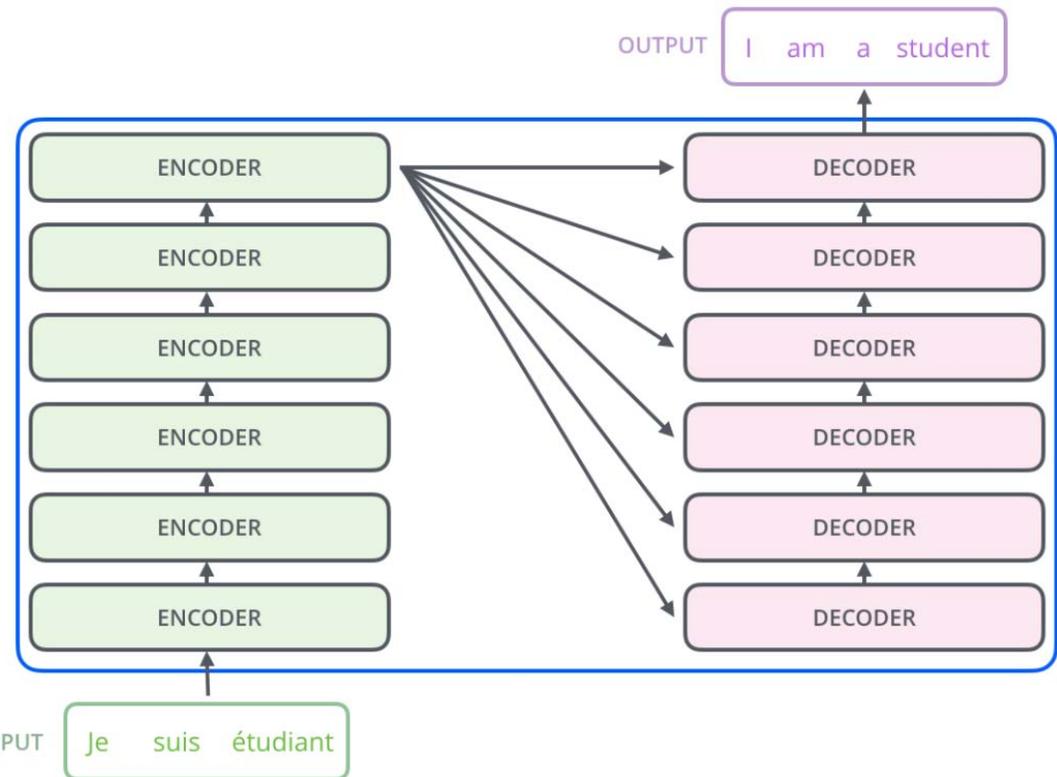
Transformer神经网络



- ❑ “Attention is all you need”
- ❑ Q: query; K: key; V: value



The encoder-decoder structure of the Transformer architecture
Taken from "Attention Is All You Need"

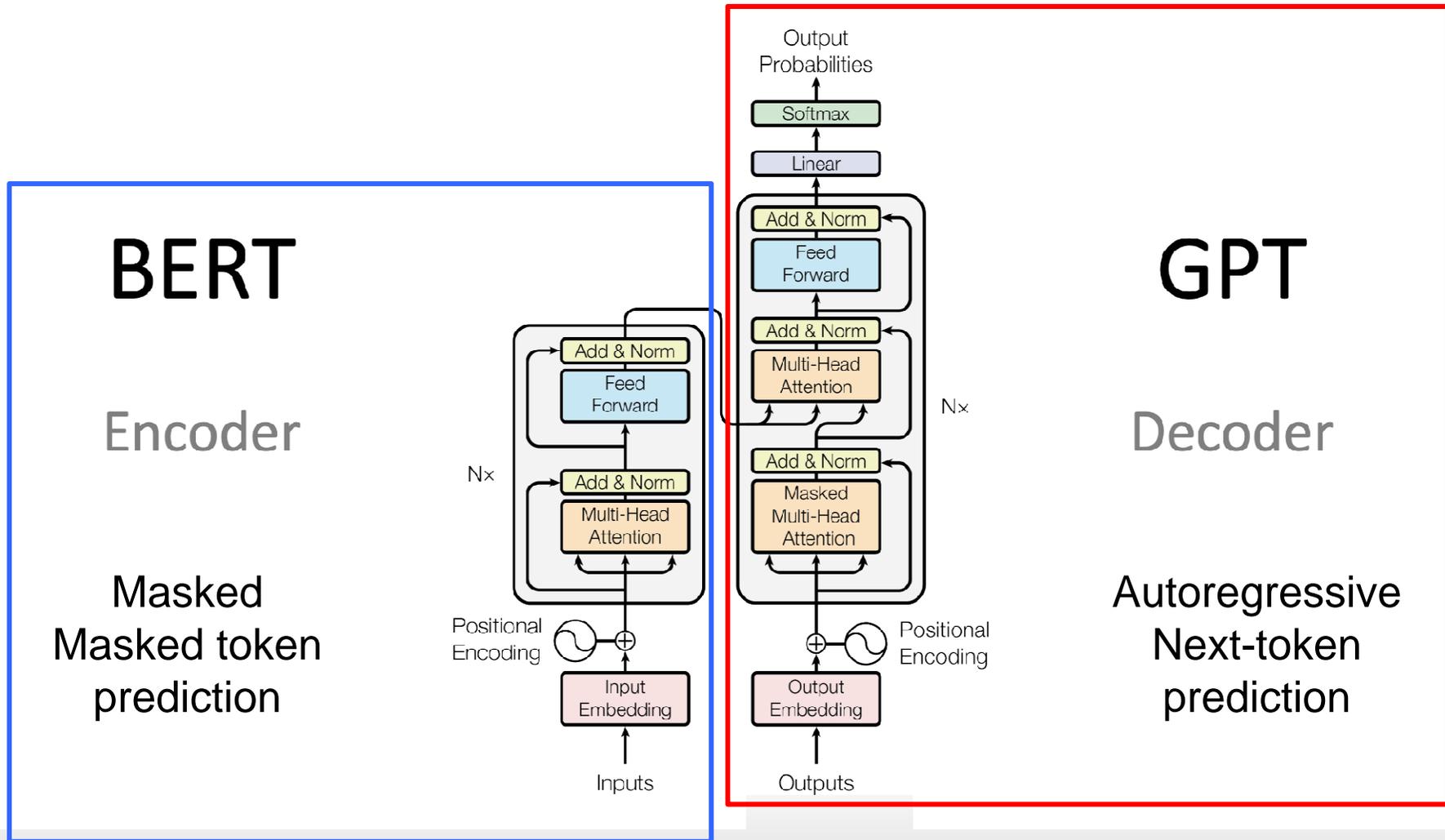


Ilya Sutskever



- ❑ **A Theory of Unsupervised Learning**
- ❑ **Unsupervised learning can be better understood through the lens of compression, with stronger compressors finding more shared structure in data, just like unsupervised learning finds structure in unlabeled data**
- ❑ **压缩即智能 (Compression for AI)**
- ❑ **Good compressors can become good predictors**

自然语言处理的两种架构



主流大型语言模型架构

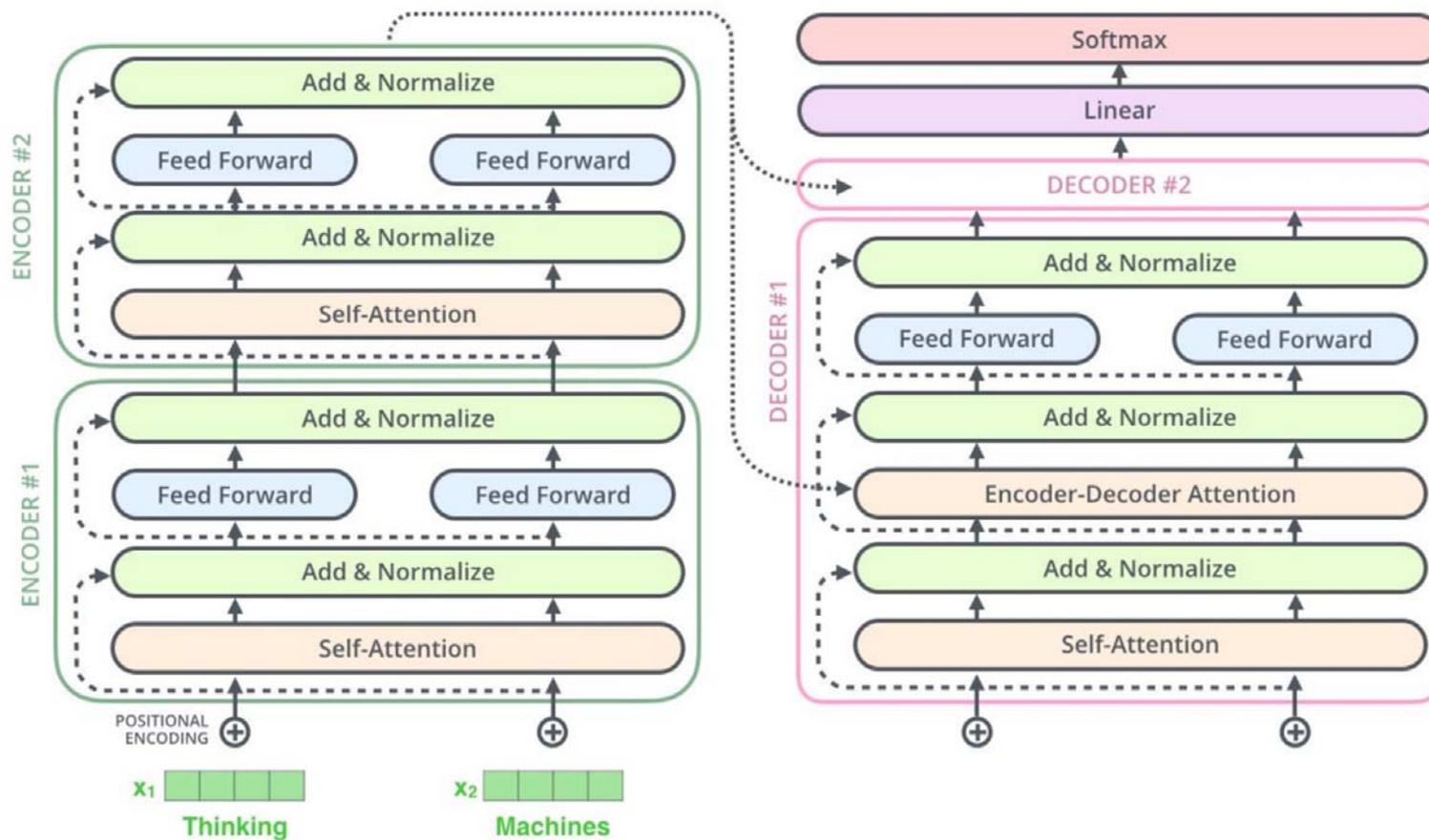


- **Encoder-decoder架构**
 - ◆ 谷歌T5模型（没做成）
- **Decoder-only架构**
 - ◆ 谷歌Bard的LaMDA架构（没做成）
 - ◆ OpenAI的GPT架构（**做成了**）
 - ◆ DeepSeek的MLA和DeepSeekMoE架构（**做成了**）



谷歌T5模型

- ❑ 输入为Text，输出也为Text的任务
- ❑ 阅读理解、摘要生成、文本分类等NLP任务



Source: <http://jalamar.github.io/illustrated-transformer/>

DeepSeek-V2



- 混合专家（Mixture of Experts, MoE）语言模型
 - ✿ 核心思想是“分工合作”
 - ✿ “18个专家，每次1个干活，17个打酱油”
- 多头潜在注意力（MLA）
 - ✿ 更准确地找到重要信息，还知道这些信息之间的关系
- 无辅助损失负载均衡策略（auxiliary-loss-free strategy）
 - ✿ 优化模型在不同任务间的负载分配，减少负载均衡可能带来的性能下降



MLA的技术原理

- ❑ 将Key-Value (KV) 缓存压缩成单个隐向量
- ❑ 降低93.3.%的KV缓存

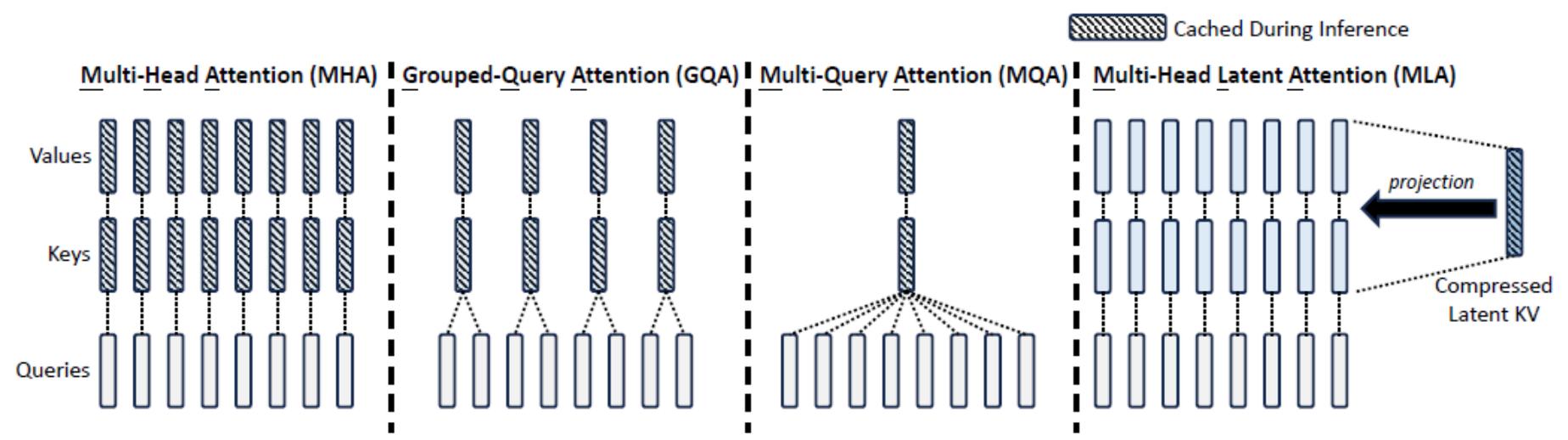


Figure 3 | Simplified illustration of Multi-Head Attention (MHA), Grouped-Query Attention (GQA), Multi-Query Attention (MQA), and Multi-head Latent Attention (MLA). Through jointly compressing the keys and values into a latent vector, MLA significantly reduces the KV cache during inference.

DeepSeek-V3



- MoE: 671B参数, 37B激活/token
- 多token预测

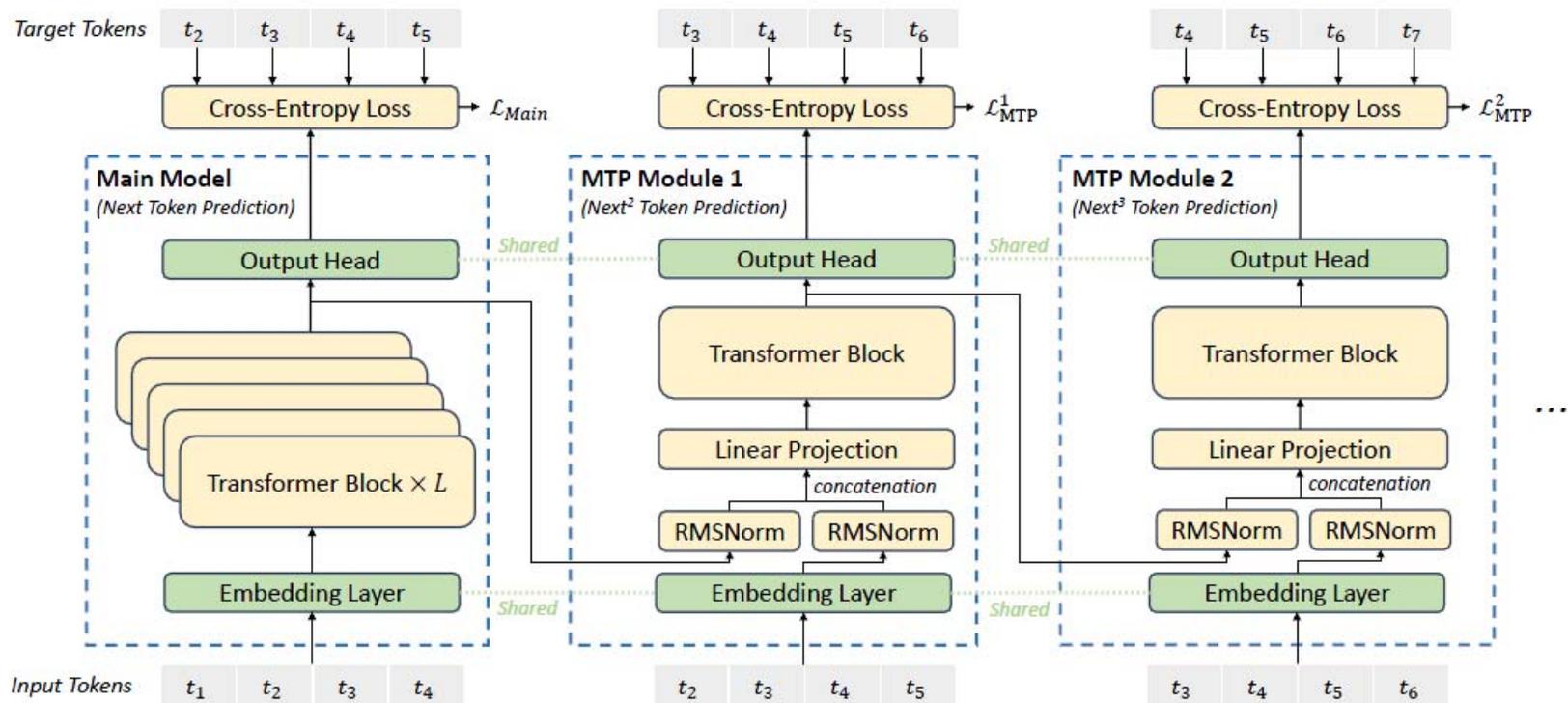


Figure 3 | Illustration of our Multi-Token Prediction (MTP) implementation. We keep the complete causal chain for the prediction of each token at each depth.

DeepSeek-V3



□ DeepSeek-V3

- ❁ 278.8万H800 GPU机时
- ❁ 预训练：14.8万亿高质量、多样化的token， 266.4万GPU机时
- ❁ 长上下文扩展：11.9万GPU机时
- ❁ 后训练：监督微调 & 强化学习，5千GPU机时

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	\$5.328M	\$0.238M	\$0.01M	\$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is \$2 per GPU hour.

DeepSeek-V3性能



对标GPT-4o，性能相当

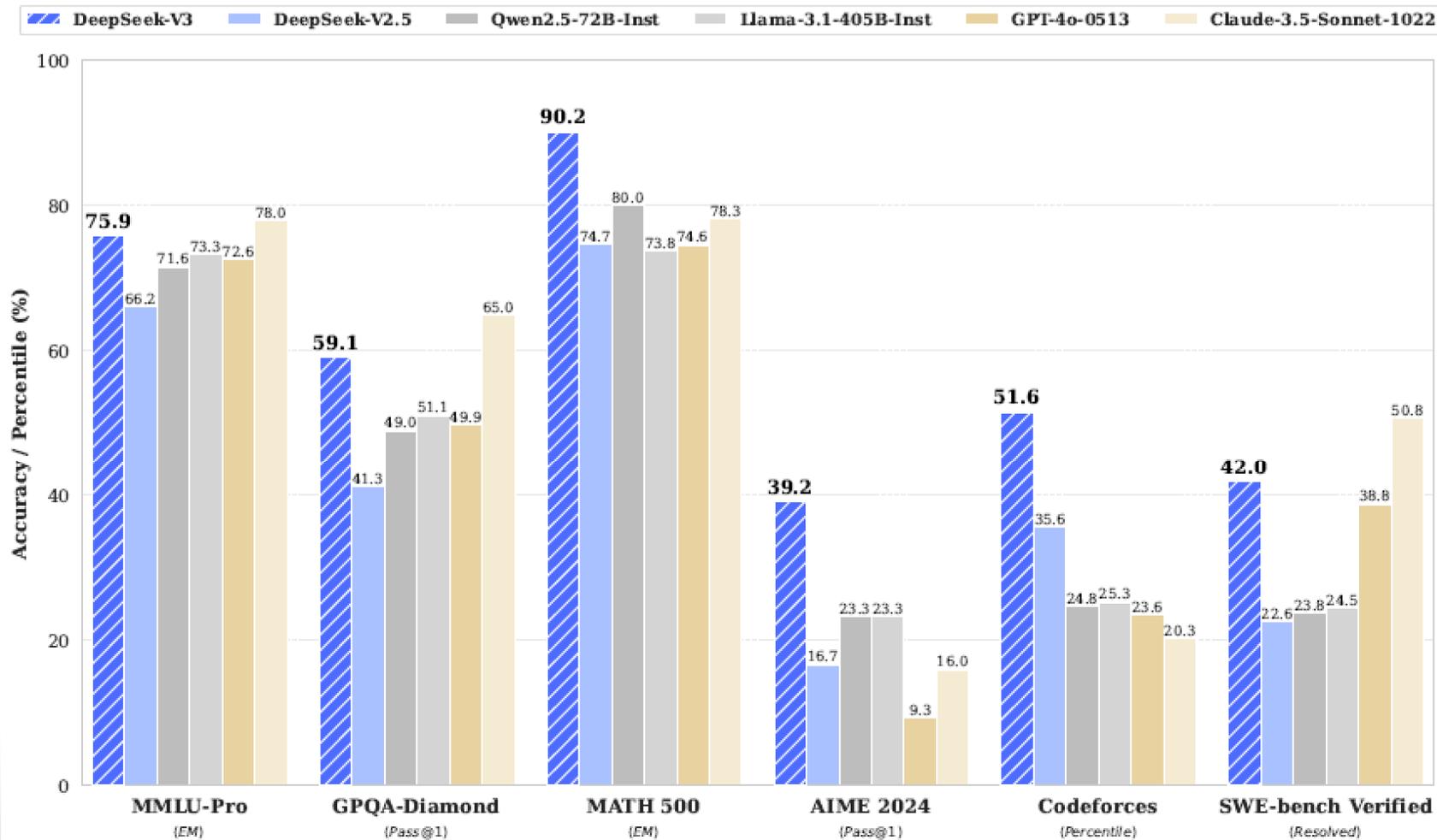


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

强化学习简介 (From DeepSeek)



- 定义：通过与环境交互学习策略，以最大化累积奖励
- 核心要素
 - ✿ 智能体 (Agent)：学习并做出决策的主体
 - ✿ 环境 (Environment)：智能体交互的外部系统
 - ✿ 状态 (State)：环境的当前情况
 - ✿ 动作 (Action)：智能体在状态下的选择
 - ✿ 奖励 (Reward)：智能体执行动作后获得的反馈
- 学习目标：找到最优策略 (Policy)，使长期累积奖励最大化
- 关键算法
 - ✿ Q学习 (Q-Learning)：基于值函数的无模型算法
 - ✿ 深度Q网络 (DQN)：结合深度学习的Q学习
 - ✿ ...

DeepSeek-R1：强化学习新策略



□ DeepSeek-R1-Zero

- ✿ 大规模强化学习，训练时不用监督微调（SFT）

□ DeepSeek-R1

- ✿ 对标OpenAI o1的推理模型
- ✿ 在强化学习之前加入了几千条长思维链的冷启动数据
- ✿ **Group Relative Policy Optimization**强化学习新策略

□ 小模型

- ✿ 基于Qwen and Llama，从DeepSeek-R1蒸馏
- ✿ 1.5B, 7B, 8B, 14B, 32B, 70B

2025.1.22, DeepSeek-R1: Incentivizing Reasoning Capability
in LLMs via Reinforcement Learning

GRPO vs. PPO



- PPO : 需要人类判断推理结果的优劣 (效率低)
- GRPO : 机器自动判断推理结果的优劣 (效率高)

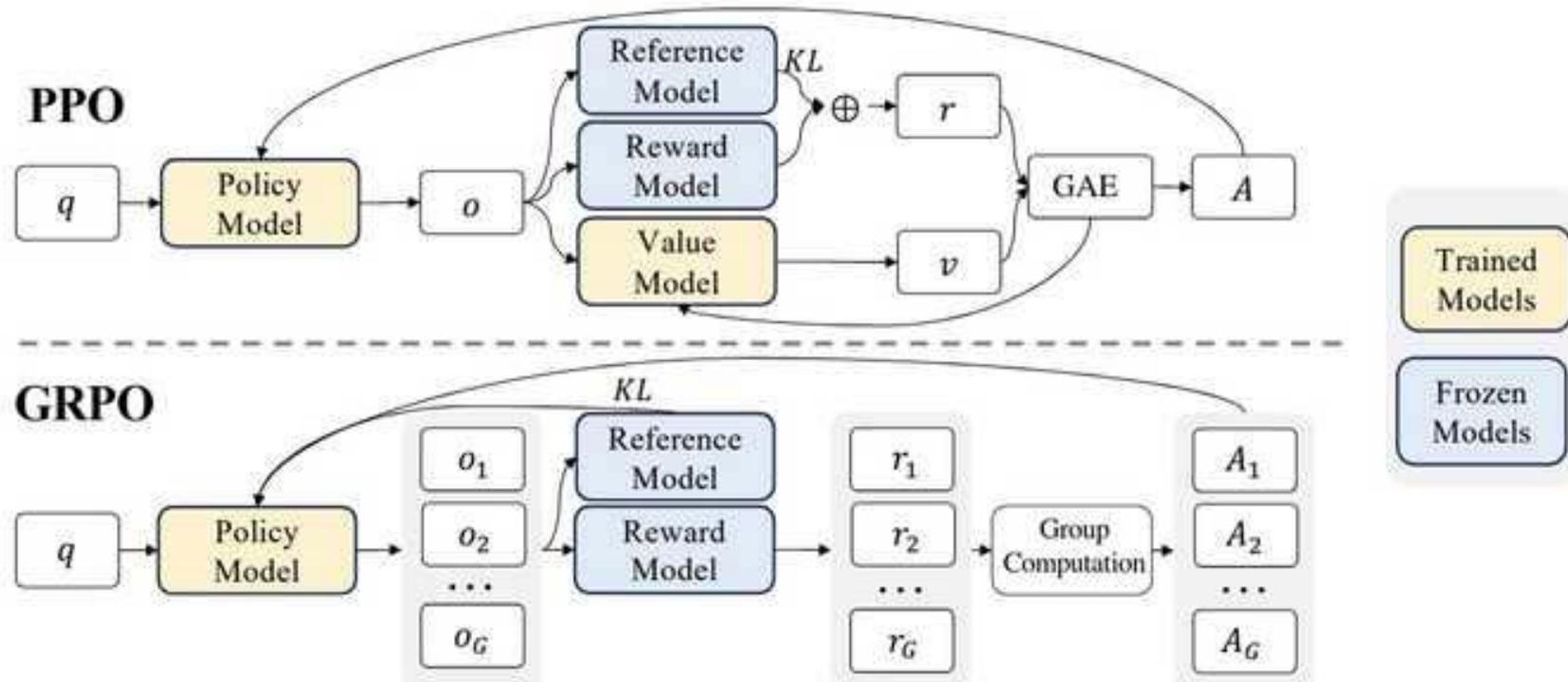


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

GRPO vs. PPO (示例)



□ 如何评价学生的综合素质?

- ⚙️ PPO: 10,000个教师, 每人对学生进行排序 (人工标注)
- ⚙️ GRPO: 考试10,000次, 机器阅卷按分数排序 (自动计算)



DeepSeek-R1 性能



对标OpenAI-o1，性能相当

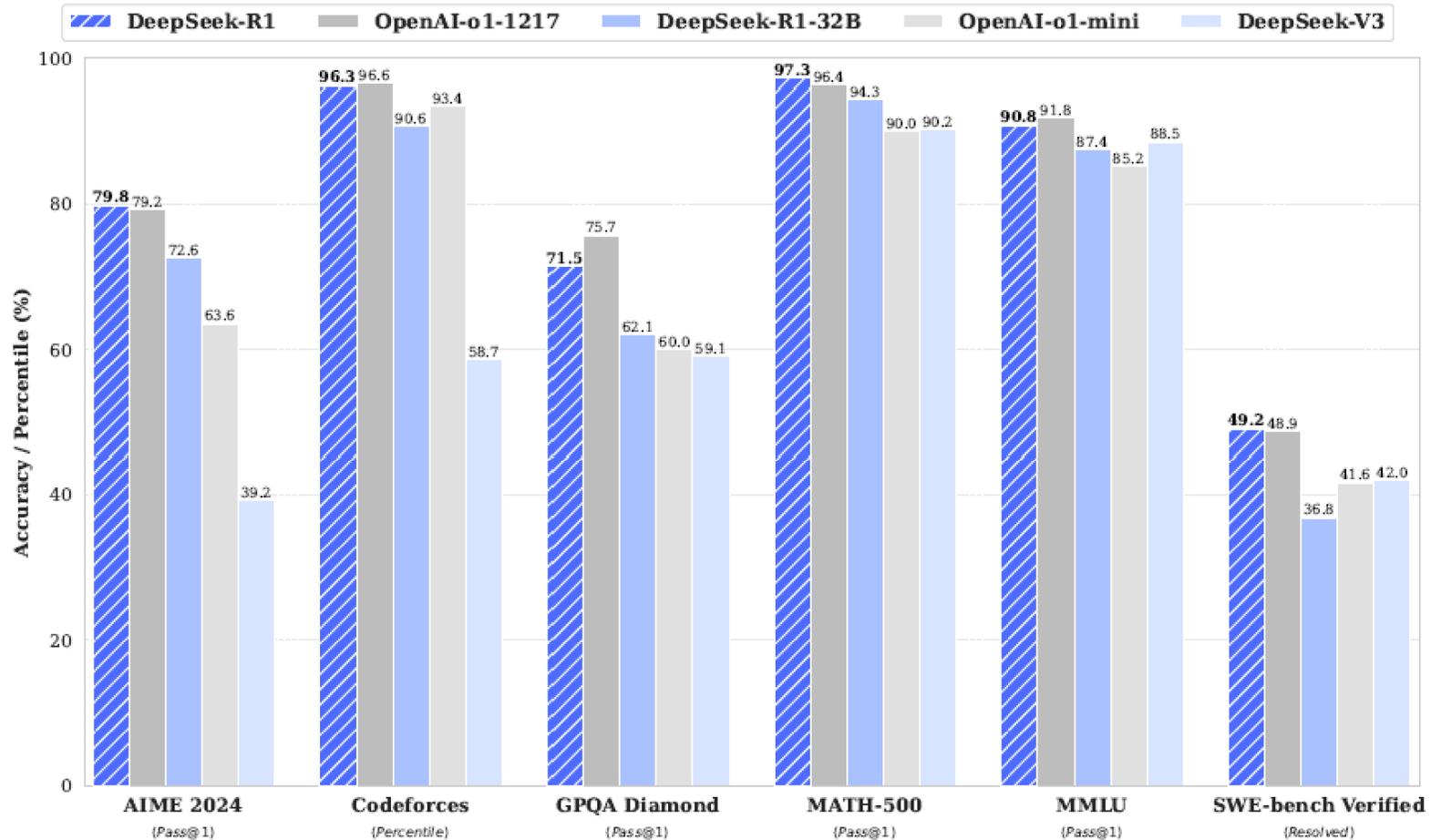


Figure 1 | Benchmark performance of DeepSeek-R1.

大模型提示工程



❑ https://blog.csdn.net/star_nwe/article/details/142396026

大模型提示工程 | 39 种提示词工程方法总结，看这一篇就够了！

原创 大模型微调部署 于 2024-09-22 11:15:00 发布 阅读量3.2k 收藏 24 点赞数 11

版权

文章标签： [人工智能](#) [大模型](#) [AI大模型](#) [ai](#) [提示词](#) [提示工程](#) [学习](#)

大语言模型 (LLMs) 在许多不同的自然语言处理 (NLP) 任务中表现出色。提示工程需要编写称为提示的自然语言指令，以结构化方式从LLMs中提取知识。与以往的最新技术 (SoTA) 模型不同，提示工程不需要根据给定的NLP任务进行广泛的参数重新训练或微调，因此仅依赖于LLMs的嵌入知识。

一、提示词工程

方法 1：基础/标准/普通提示

基础提示指的是直接向LLM提出查询的方法，而无需进行任何工程改进以提升LLM的性能，这是大多数提示策略背后的核心目标。基础提示在不同的研究论文中也被称为标准提示或普通提示。

方法 2：思维链提示 (Chain-of-Thought, CoT)

在CoT提示策略中，前将其分解为更小、更易处理的子问题的思路进行研究。类似地，作者调查了通过产生一系列思维链或中间推理步骤，如何内在增强LLMs进行复杂推理的能力。



39 种提示词工程方法

- ❑ 1: 基础/标准/普通提示
- ❑ 2: 思维链提示 (Chain-of-Thought, CoT)
- ❑ 3: 自我一致性 (Self-Consistency)
- ❑ 4: 集成精炼 (Ensemble Refinement, ER)
- ❑ 5: 自动思维链 (Automatic Chain-of-Thought, Auto-CoT)
- ❑ 6: 复杂思维链 (Complex CoT)
- ❑ 7: 思维程序 (Program-of-Thoughts, PoT)
- ❑ 8: 从简到繁 (Least-to-Most)
- ❑ ...
- ❑ 38: 基本提示加术语定义 (Basic with Term Definitions)
- ❑ 39: 基本+基于注释指南的提示+基于错误分析的提示 (Basic + Annotation Guideline-Based Prompting + Error Analysis-Based)

大语言模型助力睡眠机制解析



- ❑ 2023年5月17日投稿
- ❑ 2024年4月17日接收

揭示社交信号对睡眠潜在影响，科学家用大模型解析睡眠群体活动分子机制

原创 刘雅坤 DeepTech深科技 2024年09月23日 19:48 北京

随着 OpenAI 发布新模型 o1，大模型和基于思维链 (CoT, Chain of Thought) 的机器推理 (Machine reasoning)，再次成为人们讨论的热门话题。

然而，在 o1 发布之前，已经有科学家通过思维链推理在生物信息领域进行探索。

“OpenAI 的 o1 在机器推理方向上取得了颠覆性的突破，很高兴我们国内的研究能够与国际同步。” 华中科技大学薛宇教授表示。

为探索 and 解决生物学领域的睡眠调控机制问题，华中科技大学张珞颖教授和薛宇教授联合课题组，开发了一种可同时进行多目标视频追踪的定量装置。

借助大模型 GPT-3.5 和提示词工程 (Prompt Engineering) 技术，对果蝇全基因组进行系统解释及功能筛选。

他们对 758 个果蝇基因参与调控群体/个体睡眠和活动进行预测和验证后，发现了仅在群体环境下调节睡眠功能的基因 MRE11。

nature communications



Article

<https://doi.org/10.1038/s41467-024-48005-w>

Large-language models facilitate discovery of the molecular signatures regulating sleep and activity

Received: 17 May 2023

Accepted: 17 April 2024

Published online: 01 May 2024

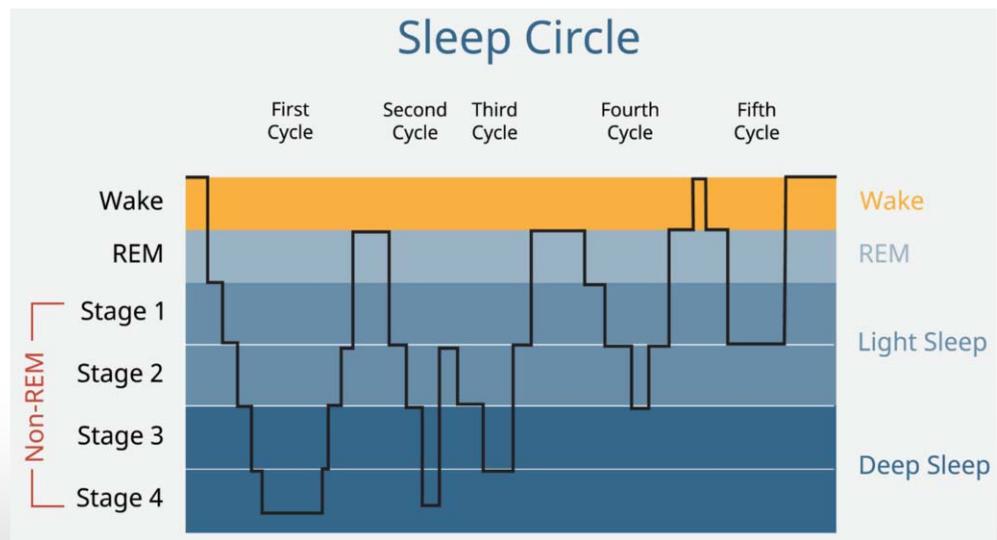
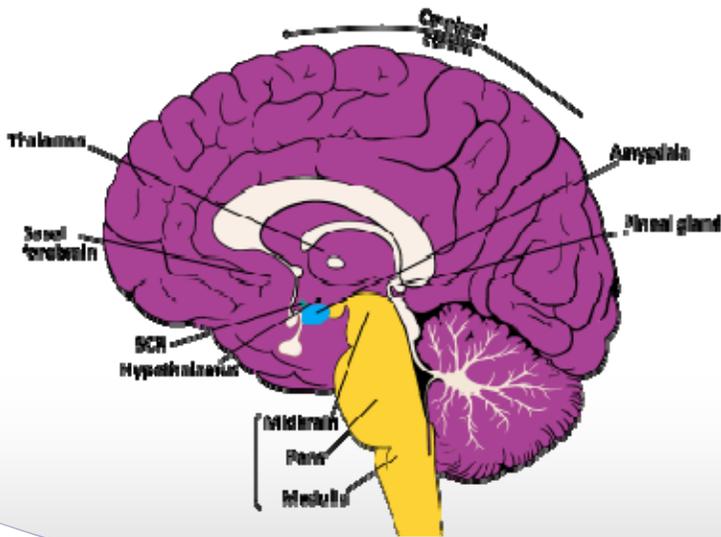
Di Peng^{1,4}, Liubin Zheng^{1,4}, Dan Liu^{1,4}, Cheng Han^{1,4}, Xin Wang¹, Yan Yang¹, Li Song¹, Miaoying Zhao¹, Yanfeng Wei¹, Jiayi Li¹, Xiaoxue Ye¹, Yuxiang Wei¹, Zihao Feng¹, Xinhe Huang¹, Miaomiao Chen¹, Yujie Gou¹, Yu Xue^{1,2} & Luoying Zhang^{1,3}

<https://xue.biocuckoo.cn/course.html>



睡眠周期

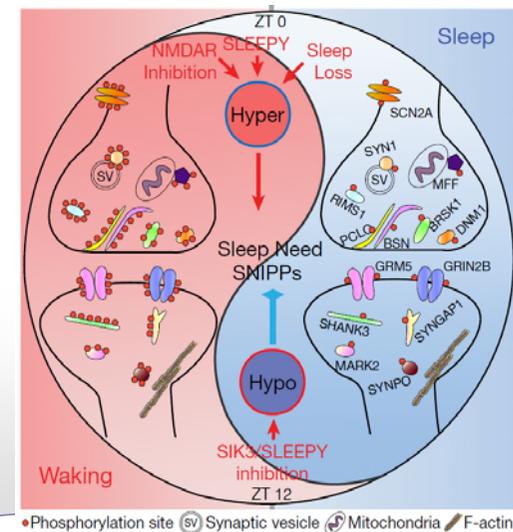
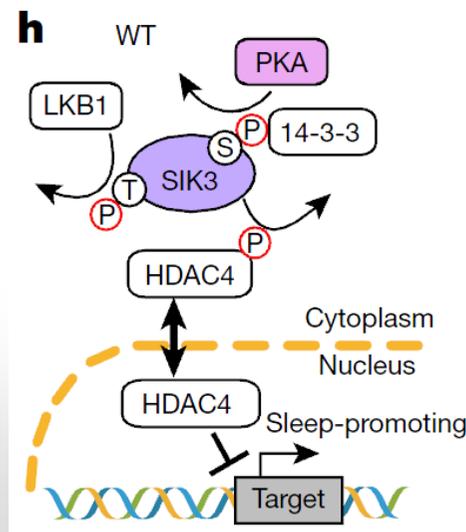
- 下丘脑 & 脑干: **GABA**
- 每个周期有4个阶段 (90~110分钟)
 - ❁ 非快速动眼睡眠 (NREM) : 无梦、大脑和肌肉的活动降低
 - ❁ 快速动眼睡眠 (REM) : 做梦, ~2小时/天, 大脑活动强烈, 身体麻痹



科学问题



- ❑ 究竟有哪些基因已报道与睡眠、活动和社交相关？
- ❑ 对于一个候选基因，应当开展什么实验来研究其与睡眠、活动或社交的关联？
- ❑ 如果发现了一个新的睡眠基因，这个基因如何与其他基因相互作用，从而调控睡眠？



提示词工程 - 示例



➤ 标准提示：直接提问

Can you tell me the function of the fruit fly protein "Dop1R1 (Dopamine 1-like receptor 1, isoform E)"? Is the protein "Dop1R1" involved in regulating sleep?

Yes, the protein Dop1R1 is involved in regulating sleep. Dop1R1 is a type of dopamine receptor that is found in the brain of the fruit fly. It has been shown to play a role in several different behaviors, including circadian rhythms and sleep...Overall, Dop1R1 is a protein that plays an important role in the brain of fruit flies, and its function is closely tied to the regulation of sleep and circadian rhythms.

➤ 思维链提示：在提示词中加入ChatGPT提前回答出的基因功能信息，并明确要求ChatGPT使用思维链进行推理，从而增强ChatGPT的推理能力

mre11 (meiotic recombination 11): Validated to be involved in regulating sleep, social activity and locomotor activity. Mre11 plays a key role in DNA double strand break repair that is important for the regulation of gene expression....

Dop1R1 (Dopamine 1-like receptor 1): Dopamine 1-like receptor 1 (Dop1R1) is receptor for dopamine. It is required in the mushroom body as a receptor regulating sleep...

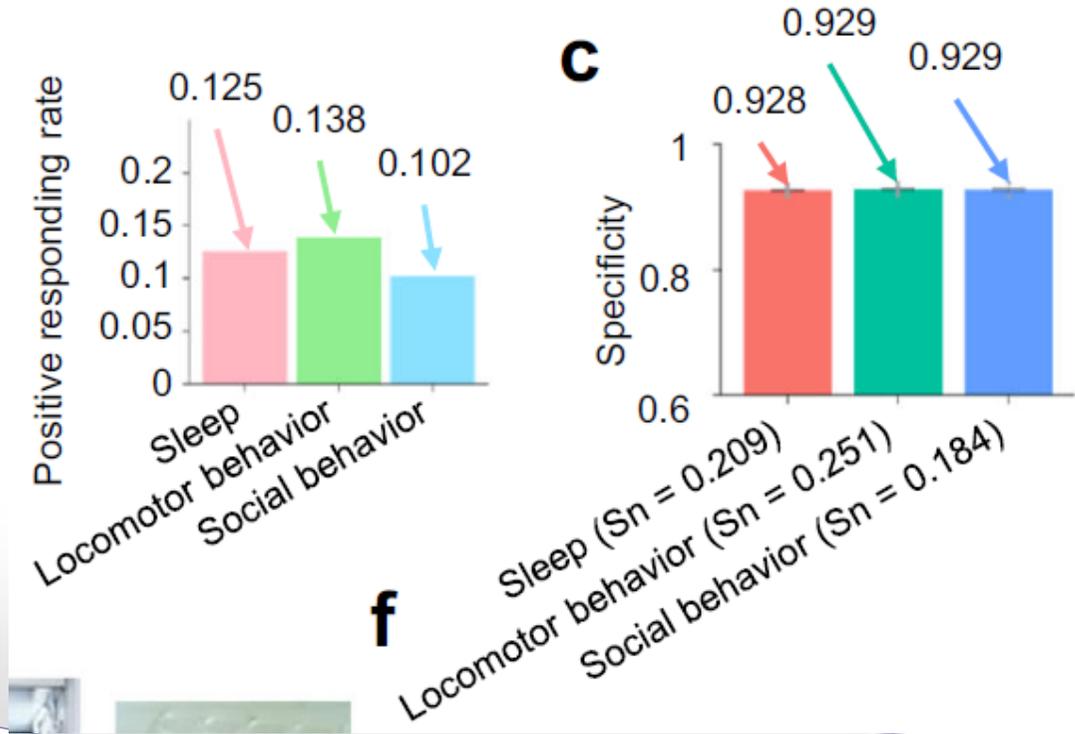
Based on your knowledge, please reason regulation relationship between mre11 (meiotic recombination 11) and Dop1R1 (Dopamine 1-like receptor 1). **Please tell me the chain of thought to reason the regulation relationship between mre11 (meiotic recombination 11) and Dop1R1 (Dopamine 1-like receptor 1)?**

Based on the information provided, we can reason that Mre11 and Dop1R1 are connected through the dopaminergic signaling pathway, which is involved in the regulation of sleep. Mre11 plays a key role in DNA double strand break repair, which is important for the regulation of gene expression. Knocking down Mre11 reduces the expression of essential sleep-related genes involved in the dopaminergic signaling pathway, including Hk, Shal, Shab, and Ih. On the other hand, Dop1R1 is a receptor for dopamine, which is a neurotransmitter that regulates sleep. Dop1R1 is required in the mushroom body as a receptor regulating sleep. Therefore, we can reason that Mre11 may regulate the expression of Dop1R1 indirectly through the dopaminergic signaling pathway, which is important for the regulation of sleep.



标准提示的性能

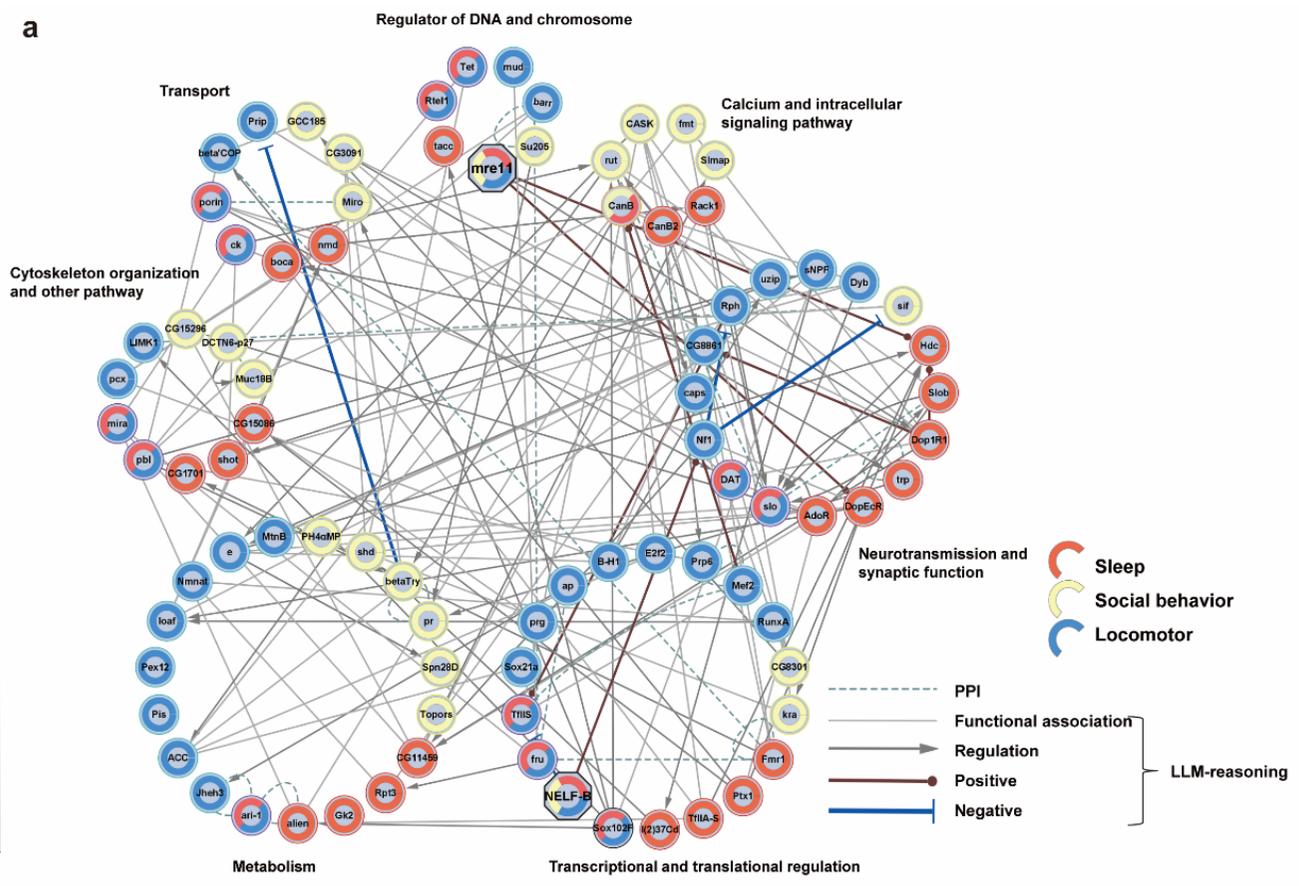
- 阳性相应率: 12.5%, 13.8%, 10.2%
- 已知基因: 268个睡眠基因、283个活动基因、49个社交基因
- S_n (Type II errors): 20.9%, 25.1%, 18.4%
- S_p (Type I errors): 92.8%, 92.9%, 92.9%





思维链提示的性能

- ❑ 86个基因，19个已知，67个候选，3655对关联
- ❑ 思维链提示：139对关联，103对（74.1%）有文献支持
- ❑ 机器 vs. 人类：3小时 vs. 1周



大创项目？



□ 能否设计**更有效**的提示词？

- ✿ 哪些基因已报道与睡眠、活动和社交相关？
- ✿ 新的睡眠基因如何与其他基因相互作用，从而调控睡眠？
- ✿ 应当开展什么样的实验来发现新基因，揭示新机制？

□ AI助教？

- ✿ 将教科书、参考书、课件喂给大模型做强化学习
- ✿ 在线工具、App、语音机器人？