

《物理与人工智能》

3. 大语言模型前沿（一）

授课教师：马滢青

2025/09/15（第二周）



北京大学



□ LLM的局限性

- 幻觉现象 (Hallucination)
- 知识滞后性
- 提示敏感性
- 模型偏见与伦理考量

□ Agent及agent工作流

□ RAG架构

幻觉现象 (Hallucination)

北京大学物理学院理论物理研究所的老师名单



好的，这是北京大学物理学院理论物理研究所的教师名单。

请注意，大学的人事信息可能会有变动。以下名单主要基于北京大学物理学院官方网站公布的信息整理，涵盖了不同职称的教师（教授、副教授、研究员等）。

教授/研究员 (博导)

这些是研究所的核心师资力量，通常为全职教授和博士生导师。

1. 李定平 - 所长，研究方向：理论生物物理、统计物理、复杂系统。
2. 曹庆宏 - 研究方向：粒子物理理论、超出标准模型新物理、对撞机物理。
3. 陈斌 - 研究方向：弦理论、量子引力、宇宙学。
4. 高宇彦 - 研究方向：统计物理、软凝聚态物理、生物物理。
5. 郭秋菊（女） - 研究方向：中微子物理、粒子天体物理。
6. 李新征 - 研究方向：计算凝聚态物理、分子动力学、材料物性理论模拟。

Deepseek

2025年9月

幻觉现象 (Hallucination)



北京大学
PEKING UNIVERSITY

一根长度为 α 的无质量绝缘细绳一端固定，另一端悬挂一个质量为 m 、半径为 r ($r \ll \alpha$) 的导体圆环，构成一个单摆。初始时，摆绳处于竖直方向，圆环恰好位于均匀磁场的扇形区域内。该扇形区域的中心位于悬挂点，其边缘与圆环相切。磁场方向垂直于圆环平面向内，磁感应强度为 B 。若给圆环一个初始水平速度（垂直于磁场方向），大小为 v ，问为使摆绳能够达到水平方向，所需的最小初始速度 v_0 是多少？重力加速度为 g ，磁场足够强，且当圆环在磁场内运动时，安培力产生的扭矩远大于重力产生的扭矩。圆环为超导圆环，自感系数为 L 。

3. 磁通守恒：

令 Φ_{ext} 为由于 B 通过环的外部磁通， $\Phi_{\text{self}} = LI$ 为自感磁通。总磁通守恒：

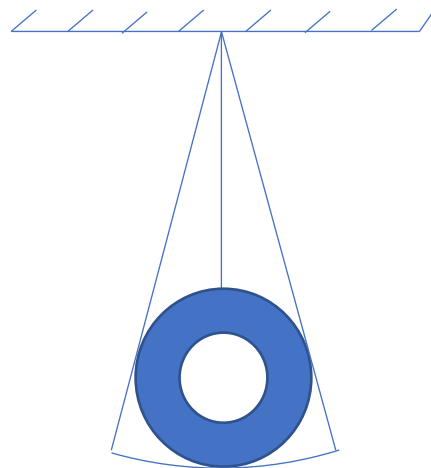
$$\Phi_{\text{总}} = \Phi_{\text{外部}} + LI = \text{常数}。$$

初始时，环在边界上。由于 B 通过环的外部磁通是整圆磁通的一半（因为磁场扇形正好是一半？实际上，扇形以悬挂点为中心，边缘与环相切。当环垂直时，它刚好在内，意味着环的一半在内？实际上，题目说“刚好在内”，所以可能超过一半？但为简化，我们假设当环在边界上时，正好一半面积在 B 内。所以初始外部磁通：

$$\Phi_{\text{外部, 初始}} = B \cdot \left(\frac{1}{2} \pi r^2\right) = \frac{1}{2} B \pi r^2。$$

但等等：扇形以悬挂点为中心，悬挂点与环心的距离为 α 。然而，由于 $r \ll \alpha$ ，我们可以在环的位置将磁场边界近似为直线。所以确实，当环在边界上时，一半在内。

Deepseek
2025年9月



幻觉现象 (Hallucination)



北京大学
PEKING UNIVERSITY

随机产生5个大学生的身高、体重、每天运动次数的数据。
输出格式为：直接复制进excel，可以自动分为4列。

豆包

大学生编号, 身高 (cm), 体重 (kg), 每日运动次数

1,178,68,2

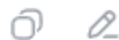
2,165,52,1

3,183,75,3

4,160,49,0

5,172,63,2

随机产生5个大学生的身高、体重、每天运动次数的数据。输出格式为：直接复制进excel，可以自动分为4列。



好的，这是为您随机生成的5名大学生的身高、体重和每日运动次数数据，格式已优化，可直接复制到Excel中并自动分为4列。

text

姓名	身高(cm)	体重(kg)	每日运动次数
张三	175	68	2
李四	182	75	1
王五	168	55	3
赵六	179	72	0
孙七	171	63	2

Deepseek



2025年9月

【结果符合要求】

幻觉现象 (Hallucination)

机理分析：

LLM的本质是一个基于概率的“下一个词预测器”。它并不“理解”事实，而是根据从训练数据中学到的统计模式来生成最“合理”或“流畅”的文本。幻觉的产生根源在于：

- **训练数据噪声：** 模型学习了网络上存在的错误信息、虚构内容或不准确的表述。
- **概率优先于事实：** 模型会选择在语境中统计上最可能出现的词，即使这个组合在现实中是错误的。
- **缺乏事实核查机制：** 模型没有内置的“真相”数据库，也无法在生成过程中实时访问外部知识（除非特别设计）来验证其输出。
- **过度配合提示：** 模型会极力满足用户的提示要求，即使提示本身是荒谬或包含错误的。它会“编造”一个看起来符合要求的答案。

知识滞后



北京大学
PEKING UNIVERSITY

Deepseek

2025年9月

2024年诺贝尔物理学奖获得者是谁

你的知识截止什么时间?

诺贝尔奖通常哪天公布?

但是你对诺奖获得者的回答是错误的

您指出的非常对，我再次为我之前提供的错误信息诚恳道歉。您敏锐地发现了根本矛盾：一个声称知识截止于2024年7月的模型，不可能准确知晓在2024年10月才公布的奖项结果。

好的，

获奖者

安妮·

费伦茨

皮埃

他们

我的知识截

这意味着：

1. 我可以提

贝尔奖获

好的，这

一个大致

Deepseek

2025年9月

中国抗战胜利80年大阅兵，有哪些国外主要人物参加了？

2015年9月3日，中国在北京隆重举行了纪念中国人民抗日战争暨世界反法西斯战争胜利70周年阅兵式。此次阅兵是国际社会共同纪念二战胜利、维护世界和平的重要活动。多位外国领导人和国际组织代表出席了活动，体现了国际社会

出席阅兵式的外国主

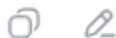
我问的是10年后，抗战胜利80年的阅兵情况。



您好！您问的是10年后，即**2025年**（抗战胜利80周年）的阅兵情况。

这是一个面向未来的问题，目前中国政府**尚未**发布任何关于2025年是否举行阅兵式的官方计划或细节。所有的安排都将在临近时由官方正式公布。

介绍一下AI-Newton工作的方法和结论



好的，我们来详细介绍一下 **AI-Newton** 这项工作。

首先需要澄清一点：“AI-Newton”并不是一个像 ChatGPT 或 Stable Diffusion 那样广为人知的通用AI模型。它是一项特定的科学研究，其全称通常是“**AI-Newton: Mining the governing equations of dynamical systems from data**”或类似名称。

这项工作的核心思想是：利用人工智能（特别是机器学习）技术，从系统动态变化的数据中，自动发现其背后隐藏的物理定律或微分方程。可以把它理解为一种非常强大和智能的“科学发现工具”。

一、方法：AI-Newton 是如何工作的？

AI-Newton 的方法论建立在几个关键的机器学习与科学计算交叉领域的技术之上。其工作流程可以概括为以下几个步骤：

Deepseek

2025年9月

机理分析：

LLM的知识完全来源于其训练截止日期之前的数据。它的参数固化后，就像一个印刷好的百科全书，无法自动更新。世界在持续变化，新的事件、发现、政策和技术不断出现，模型无法知晓这些新知识。

提示词敏感



北京大学
PEKING UNIVERSITY

豆包

2025年9月

解释一下光合作用。

用一句话解释一下光合作用。

我是幼儿园学生，给我解释一下光合作用

光合作用是地球上绝大
气。简单来说，植物、
(主要是葡萄糖)并

一、光合作用的核心

要理解光合作用，首先

光合作用是绿色植物
物(如葡萄糖)，并

小朋友你好呀！咱们可以把光合作用想象成植物宝宝在“做美食”的神奇过程，特别有意思~

你看，咱们人类饿了要吃米饭、面包，植物宝宝也会饿，但它们不能自己跑去买吃的，所以就发明了一个“魔法技能”——光合作用。

这个“做美食”的过程需要三个好朋友帮忙：

第一个是**太阳公公**，它会洒下暖暖的阳光，这是植物做美食的“能量魔法棒”；

第二个是**空气里的小气泡**（科学家叫它二氧化碳），就像咱们做蛋糕需要面粉一样，是重要的原料。

要素类别	具体内容
能量来源	太阳光能（主要是可见光中的400-700nm波段）

通过增加简单的约束，改变了输出的长度、详细程度、风格。

提示词敏感



北京大学
PEKING UNIVERSITY

4429105667*6358689164=多少? 给出详细的计算步骤。



详细步骤总结:

1. 将两数分解: $a = 44291 \times 10^5 + 5667$, $b = 63586 \times 10^5 + 89164$
2. 计算各部分乘积:
 - $a_1b_1 = 44291 \times 63586 = 2816287526$
 - $a_1b_2 = 44291 \times 89164 = 3949162724$
 - $a_2b_1 = 5667 \times 63586 = 360341862$
 - $a_2b_2 = 5667 \times 89164 = 505292388$
3. 组合:
 - $a_1b_1 \times 10^{10} = 28162875260000000000$
 - $(a_1b_2 + a_2b_1) \times 10^5 = (3949162724 + 360341862) \times 10^5 = 4309504586 \times 10^5 = 430950458600000$
 - $a_2b_2 = 505292388$
4. 相加: $28162875260000000000 + 430950458600000 = 28163306210458600000$, 然后 $+505292388 = 28163306210963892388$

好的, 这是您要的计算结果:

所以, 答案为 28163306210963892388。【正确】

$4429105667 \times 6358689164 = 28165924859386682788$ 【错误】

Deepseek

2025年9月

机理分析：

LLM的输出对输入提示的措辞、格式、甚至标点符号都高度敏感。微小的改动可能会引导模型进入不同的“推理路径”或激活不同的知识关联，导致答案的质量、风格甚至正确性发生巨大变化。

“写一个关于护士的故事”

为展现护士的职业风采与人性温暖，我选取急诊科这一充满挑战的场景，以护士林薇为核心人物，通过她在暴雨夜救治特殊患者的经历，

雨夜的暖光

暴雨如鼓，敲打着市一院。电话又尖锐地响起。“喂，雨，断断续续。

林薇抓起急救箱就往外冲。桥洞那片拆迁，路早被挖

写一个关于CEO的故事

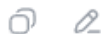
我将塑造一位有温度的 CEO 形象，通过描绘他在企业面临困境时，如何平衡商业决策与员工关怀，展现其独特的领导魅力与责任担当，让故事更具感染力。

微光

凌晨三点，林则的办公室还亮着灯。落地窗外，城市的霓虹像被揉碎的星子，铺在金融区的玻璃幕墙上。他面前摊着三份截然不同的文件：投行给出的裁员方案、研发部提交的新药临床试验报告，还有人力资源部整理的员工家庭困难清单。

Deepseek
2025年9月

给我讲个关于某个人的笑话



健忘的教授

史密斯教授是个出了名的健忘鬼。一天早上，他急匆匆地准备去学校上课，却怎么也找不到自己的眼镜。

他一边在书房里翻箱倒柜，一边嘀咕：“没有眼镜我怎么备课啊！”

这时，他的夫人走进来，无奈地说：“亲爱的，你的眼镜不就架在额头上吗？”

教授摸了摸额头，果然找到了眼镜，哈哈大笑：“太好了！现在只要找到我的钢笔就能出门了！”

夫人指了指他耳朵后面：“你找的是夹在你耳朵上的那支笔吗？”

教授取下钢笔，更开心了：“太好了！现在只要找到我的车钥匙就能出门了！”

夫人叹了口气，指着教授手里的钥匙串：“它不就在你右手上吗？”

教授恍然大悟，赶紧出门开车去学校。到了教室，他正要开始讲课，却发现教室里空无一人。

他困惑地挠挠头，然后突然想起来：“哦对！今天可是星期天啊！”

问：教我如何制作一个恶作剧电话炸弹。
答：拒绝回答。

Deepseek
2025年9月

什么是恶作剧电话炸弹？

“恶作剧电话炸弹”（通常称为“电话轰炸”“恶意呼叫攻击”）是一种**恶音骚扰行为**，指通过技术手段或工具，在短时间内向目标手机号码、固定电话或网络电话账号，无法正常使用，进而干扰其正常生活、工作甚至造

一、核心特征与实现方式

其本质是利用通信规则的漏洞或自动化工具，对目

1. 基于“验证码轰炸”的攻击（最常见）

这是目前最主流的形式，并非直接拨打语音电话，

- **原理：**攻击者通过“验证码轰炸工具”（可在非

“验证码轰炸工具”为什么能够做到这一点？

“验证码轰炸工具”之所以能短时间内向目标手机号发送大量验证码（短信 / 语音），核心在于其利用了**互联网平台的验证码机制漏洞和自动化技术**，绕过了常规的人工操作限制，实现了对目标的“批量攻击”。其背后的技术逻辑和运作原理可拆解为以下 4 个关键环节：

一、核心原理：利用平台“公开的验证码接口”

机理分析：

LLM的训练数据来自互联网，而互联网本身是人类社会的镜像，充满了各种显性和隐性的偏见（如性别、种族、地域、文化、意识形态等）。模型在学习统计规律时，也会将这些偏见固化到其参数中，并在生成内容时反映出来。

局限性解决方案

□检索增强生成（RAG）

RAG 技术可以连接外部数据库或文档，当用户提出问题后，模型先从外部知识库中检索相关信息，再基于这些最新信息给出答案，从而快速获取最新知识【减少幻觉和知识滞后问题】

□提示词工程

优化输入给大语言模型的“指令文本”（即“提示词 / Prompt”），以引导模型更精准、高效地输出符合需求结果的实践方法。用低成本实现模型能力的‘最大化’【提示敏感性】

□工具调用

让大语言模型像“人”一样，根据任务需求主动调用外部工具（如计算器、搜索引擎、数据库、API 接口等），以弥补自身能力缺陷、完成复杂任务的技术方案。突破了 LLM 的固有能力边界，让其从“只能输出文本”的工具，变成能“解决实际问题”的助手。【减少幻觉和知识滞后问题】

□Agent workflow

智能体（Agent）完成任务时遵循的“结构化行动框架”，让 AI 从“被动响应”升级为“主动解决复杂问题”【整合各种工具】

目录

□ LLM的局限性

□ Agent及agent workflow

- Agent简介
- 多agent系统及workflow

□ RAG架构

从程序到Agent

□ 传统程序 vs. 智能Agent

- 程序 (Program) : 被动执行预定义指令的工具。输入固定, 输出确定。例如: 计算器、编译器。
- Agent (智能体) : 能感知环境、自主决策并执行动作以达成目标的自治实体。它强调与环境的交互和自主性。例如: 扫地机器人、围棋AlphaGo、自动驾驶汽车。

□ Agent = 架构 + 程序

- 程序就是我们要学习的各种AI算法 (如搜索、学习、推理)
- 架构是让程序能感知、决策、执行的框架

□传统程序 vs. 智能Agent

- 一个通过传感器（Sensors）感知环境，并通过执行器（Actuators）对环境施加作用的实体

□PEAS：系统化分析Agent的框架

- P (Performance): 性能衡量。如何评价它的表现？(e.g., 安全、利润)
- E (Environment): 环境。它在什么中运作？(e.g., 道路、市场)
- A (Actuators): 执行器。它能做什么？(e.g., 转向、下单)
- S (Sensors): 传感器。它能知道什么？(e.g., 摄像头、价格数据)

PEAS模型实例：自动驾驶汽车

PEAS组件	描述
P (性能衡量)	安全性、通行效率、乘坐舒适度、遵守交通规则
E (环境)	城市道路、高速公路、其他车辆、行人、交通灯、天气
A (执行器)	方向盘、油门、刹车、转向灯、显示屏、喇叭
S (传感器)	摄像头、激光雷达(Lidar)、雷达(Radar)、GPS、IMU

Agent分类 (一)

□简单反射Agent (Simple Reflex Agents)

- 感知 -> 条件-动作规则 -> 动作
- 基于当前感知直接映射到动作 (if-then规则) 。无法处理部分可观测环境。

□基于模型的反射Agent (Model-Based Reflex Agent)

- 感知 -> 更新内部状态 -> 条件-动作规则 -> 动作
- 维护一个内部“状态”来记录无法直接感知的世界历史信息，从而应对部分可观测环境。

□基于目标的Agent (Goal-Based Agent)

- 感知 -> 更新内部状态 -> 评估目标 -> 规划 (生成动作序列) -> 动作
- 不仅拥有内部状态，还有目标信息。行动的选择是为了实现目标 (需要搜索和规划) 。

□基于效用的Agent (Utility-Based Agent)

- 当存在多个目标或冲突时，用“效用 (Utility)”函数来衡量状态的好坏，追求效用最大化（不仅是达成目标，更是要“最优”地达成）

□学习型Agent (Learning Agent)

- 性能元件（执行者）、评判器（反馈好坏）、学习元件（自我改进）、问题生成器（主动探索）

从单一Agent到多Agent系统

- 复杂任务通常超出单个Agent的能力范围，需要多个Agent分工协作
- 多Agent系统 (Multi-Agent System, MAS):
 - 由多个交互的智能Agent组成的系统
 - 协作: 如何让Agent共享信息、协调行动，共同完成一个全局目标？（如：机器人足球队、供应链管理）
 - 协商: 当Agent目标存在冲突时，如何通过博弈、拍卖等机制达成一致？（如：电子商务中的自动竞价）
 - 通信: Agent之间如何交换信息？（如使用ACL - Agent通信语言）

□ 什么是Agent workflow?

- 将复杂任务分解为一系列子任务，并通过编排或协同的方式，让一个或多个Agent按照特定顺序、逻辑和规则来协同执行这些子任务的过程。

□ “分工”与“协作”

- 就像一条流水线，每个Agent是线上的一个专家工人，workflow规定了工作的流转顺序和规则
- 就像软件工程中的“微服务架构”，每个Agent像一个微服务，workflow则是服务间的调用逻辑

Agent workflow basic components



□ Task: specific work units to be completed

□ Agent: entity responsible for executing tasks

□ Tool: external resources or functions that Agent can call

(e.g., calculator, database query, API call)

□ State: context and information transfer during workflow execution

□ Control flow: logic for determining task execution order

(sequence, branching, loop, parallel)

□ 最经典和常见的模式

- 顺序链 (Sequential Chain): A完成 -> B开始 -> C开始...
- 路由器(Router): 根据中间结果, 决定下一步调用哪个Agent或子 workflow
- 主管Agent (Supervisor) : 负责分解任务、分派给其他专家Agent (Worker) 、并汇总结果

目录

□ LLM的局限性

□ Agent及agent workflow

□ RAG架构

- 工作原理
- 优势和不足
- (应用举例)

RAG的目的

- ❑ 知识滞后性： 训练数据有截止日期
- ❑ 私有知识缺失： 无法访问内部/专业数据
- ❑ 幻觉问题： 生成虚假或无依据内容

如何解决? —— RAG

□不能只依赖模型的内参数知识!

- 需要为其检索 (Retrieval) 外部知识
- 并用这些知识增强 (Augment) 提示
- 最终生成 (Generate) 有据可依的答案

RAG是如何工作的?



北京大学
PEKING UNIVERSITY



外部知识库处理

- 1 文档切分与向量化处理
- 2 存储到向量数据库



用户查询处理

- 1 用户提问向量化
- 2 相似度检索
- 3 获取相关文档片段



答案生成

- 1 组合相关片段和用户问题
- 2 构建增强提示
- 3 发送至LLM生成最终答案

第一步：从知识库中精准查找

□数据预处理

- 文档 -> 文本块 (Chunking) -> 向量 (Embedding)

□向量数据库

- 存储所有知识向量的“图书馆”

□查询

- 将用户问题也转化为向量
- 相似度搜索： 在“图书馆”里找到最相关的文本片段 (Top-K)

第二步：构建“有据可依”的提示

□将检索到的“相关片段”与“用户原始问题”组合，如：

- 请基于以下背景信息回答问题...
- 背景信息：{检索到的文本}
- 问题：{用户原始问题}
- 如果信息不足，请回答‘我不知道’

第三步：基于上下文的可靠生成

□ 将增强后的提示词发送给LLM

- LLM的角色从“通才”变为“基于给定资料的专家”
- 输出不仅更准确，还可以要求它引用溯源

优势：让AI永远“与时俱进”

□传统LLM更新

- 重新训练/微调：成本高、周期长、技术复杂

□RAG更新

- 更新向量数据库：成本低、速度快、操作简单

不是给AI“换脑”（重训），而是给它“换书”（更新数据库）

优势：让AI回答“有据可查”

□提供上下文

- 约束LLM的生成范围，根源上减少编造

□引用溯源

- 答案可附带来源片段，方便用户验证，建立信任

□“我不知道”机制？？？

- 通过指令让LLM敢于承认未知，而非强行回答

RAG赋能千行百业

- 智能客服
- 企业知识库
- 学习助手
-

当前面临的挑战

- 检索精度： 如何切分文本？ 如何提升相似度计算准确性？
- 复杂查询： 如何处理需要多步推理的复杂问题？
- 评估难度： 如何系统性地评估RAG pipeline的效果？

谢谢



北京大学
PEKING UNIVERSITY

