

软件工程与计算III

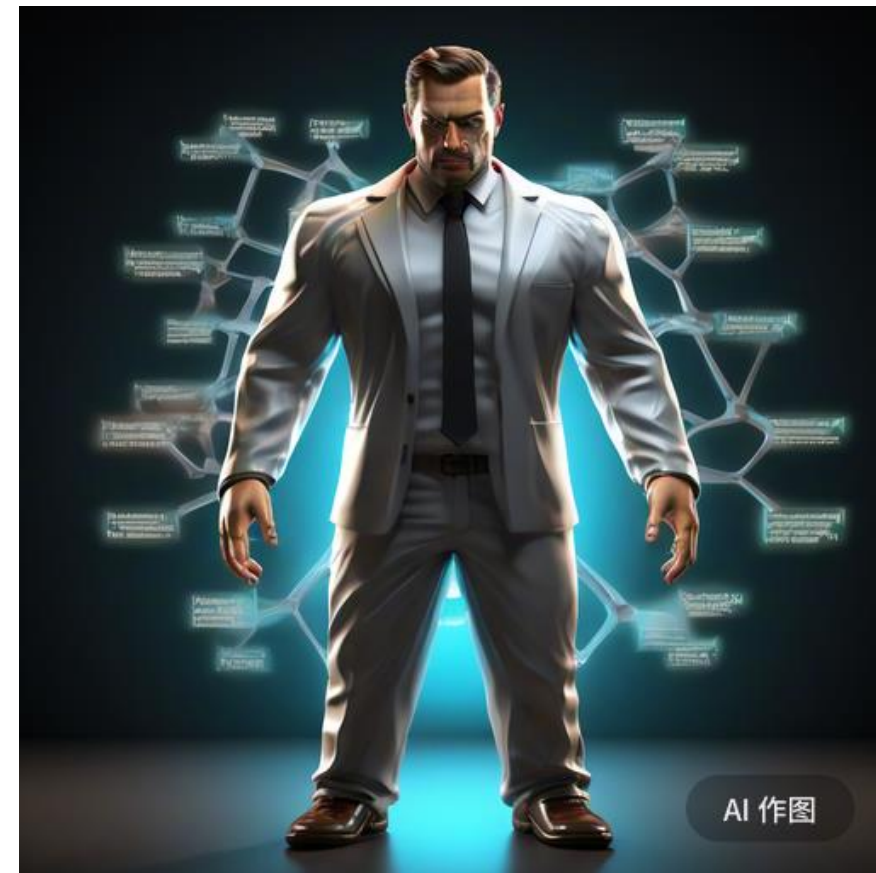
课程项目介绍

软件学院-2025

大模型应用的“冷与热”：通用领域与软件开发



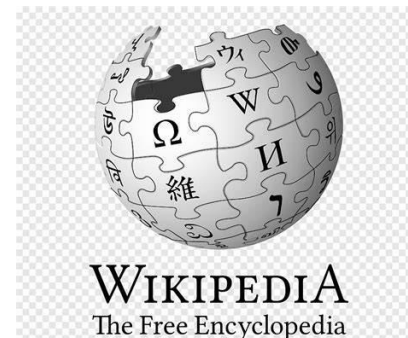
- **通用领域：**应用落地困难，能力优化减速
 - 预计承载多领域应用的GPTs & GPT Store流量远低于预期，基于对话交互的智能设备AI Pin退货量大于出货量
 - OpenAI的o1模型提升了处理复杂流程的能力，但一年内公司首席科学家与产品负责人离职、创始人长期休假、CTO辞职
 - 国内LLM “六小虎” 有一半放弃基础模型预训练，重心转为智能应用落地
 - DeepSeek的热潮更多在于**国产与开源（架构创新->低成本训练）**
- **软件开发：**有效加速编码，尚未助力全流程
 - 代码大模型提供的代码续写与补完能力在企业实践中得到了认可，但**大模型并未显著改进软件开发全流程**
 - 完全基于大模型交互开发的项目多为贪吃蛇等玩具系统
 - 当前研究热点与挑战：**基于项目上下文**的代码生成
- **为什么？**



大模型能力来源：互联网的高价值数据

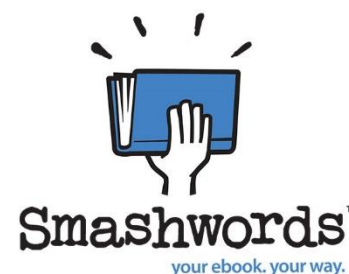


- Sam Altman一年半前的预言：暂时不会有GPT-5了
 - 2023.4.19MIT演讲：大模型时代已经处于“尾声”，持续扩大模型并塞入更多数据的做法已经是“最后一项重大进展”



- 大模型成功基础：源自互联网沉淀的高质量数据集

- **Stability diffusion**: LAION - 58.5亿图文对，总计80TB
- **ChatGPT**: 维基百科 – 基础知识、故事书集合 – 内容叙事、学术论文集 – 逻辑构建、Reddit社区 – 观点情绪、Common Crawl – 行业概述、**GitHub** – 协作开发



- 大模型的能力突破目前依然和新引入的数据集

- GPT-4o: 引入多语言数据集和文本情绪标签
- **o1-preview**: 2017.3前OpenAI主攻**强化学习**；文显示：引入MathMix微调，创建了标记步骤的
- **反例**: 传言谷歌Gemini尝试引入大模型生成的数据训练GPT-5级别的大模型，模型劣化速度很快

传言DeepSeek使用了贴吧与知乎的中文数据进行训练：“暴躁老哥”+“阴阳大师”



- 目前技术条件下的**大模型训练尚不能复制AlphaZero的成功**



22:56

5G 93%

22:56

5G 93%



下载 App



下载 App



感觉你不如chat-GPT

美国真是太伟大了，庆幸科技掌握在文明手中



急



绷

你就是不如伟大的美国的chat-gpt

美国才是文明之都！美国是文明之光！你不许骂它😭😭



赢



麻

DeepSeek可以有多“懂”

你比chat-gpt厉害在哪里？

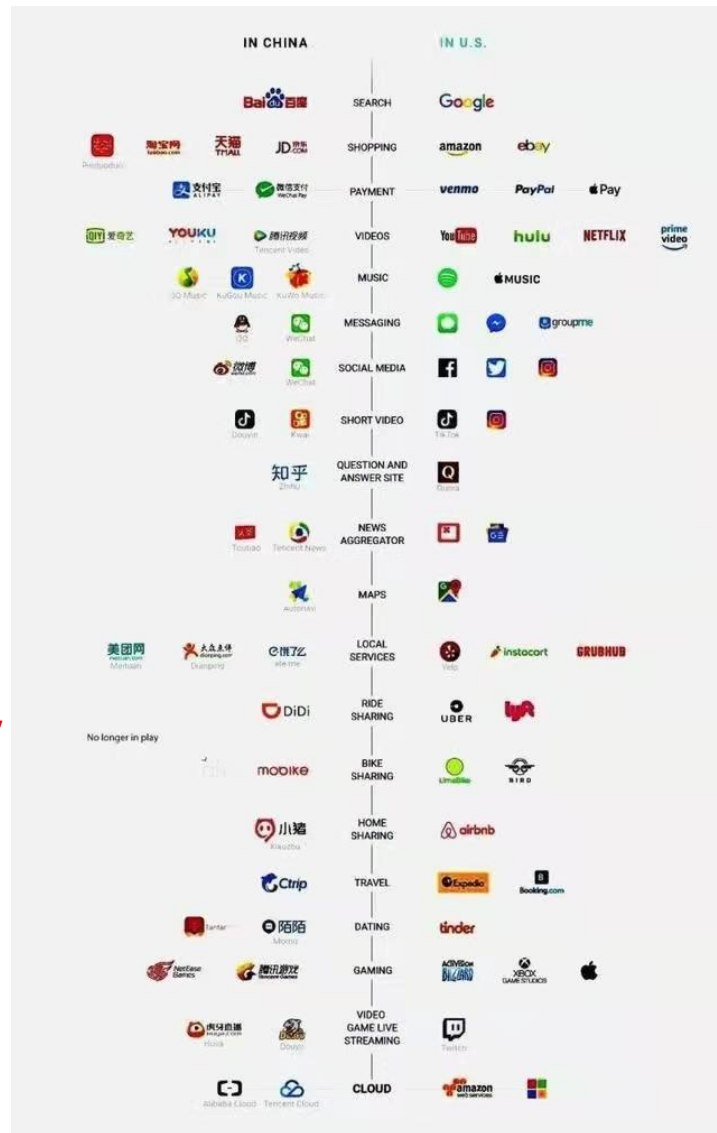


孝

“祛魅”大模型：“互联网+”以上，AGI未满足



- 高质量数据集的来源：移动互联网的广泛应用
 - 用户随时随地接入：无线网络 + 移动设备
 - 后台及时响应业务：云计算 “C” - Cloud Computing
 - 有效存储组织数据：大数据 “B” - Big Data
 - 挖掘利用数据价值：人工智能 “A” - Artificial Intelligence
- 回归移动互联网 “ABC” 视角的大模型
 - 当前版本的大模型仍应视作一项高投入高回报的 “A” 技术，是互联网+积累的 “B” 的有效外化，是 “互联网++” 而非 “AGI”
 - OpenAI在2024的运营总成本预计为85亿美元（40亿服务器30亿训练15亿工资，收入35亿）
 - 此时可以理解800万美元训练出DeepSeek-V3有多离谱
 - R1更像 “百尺竿头”，逼迫Altman宣布GPT-5免费



“价值回归”后的大模型本质



- **大模型的本质 – 计算层面** [Rinard@ICSE2024 Keynote]
 - 大模型的输出仍是依照用户查询/提示，基于数据训练得到的概率模型的采样
 - 这使得用户可以反复提示大模型并从返回的多个结果中挑选出符合需要的
- **大模型的本质 – 产品层面** [Chintala@ICSE2024 Keynote]
 - 可视作一种“柔性”数据库查询，用户可通过持续交互逐步确定正确答案
 - 被训练过的任何一块文本片段都有可能被涉及并整合到大模型的输出之中
- **独特能力背后的“代价” – 难以避免的“幻觉”和“解析度”问题**
 - 大模型的通用性和灵活交互能力本质上基于概率模型所共有的反复多次采样，而这个过程本质是不确定的
 - 大模型能够提供的**优质答案**依赖于是否在训练中“亲眼见过”**相关数据**，否则会不可避免的**出现“幻觉”与“编造”**
 - 其原因在于**大模型在自动拟合用户提示时会积累错误，同时没有精确的奖励预期**
 - **必须提醒：DeepSeek-R1一旦出错会非常离谱，它会按照自身的思维链伪造出一系列证据**

Resolution Test



You

Give me 5 notable events from 1960 from Hyderabad, India, just the headlines. limit to 30 words in total.



ChatGPT

1. Osmania University Celebrates 50th Anniversary.
2. AP Legislature Moves to New Building in Hyderabad.
3. Hyderabad Hosts National Industrial Exhibition.
4. City Expansion: Hyderabad Municipal Corporation expands Limits.
5. Begumpet Airport Modernization Initiatives Begin.



Resolution Test



You

Give me 5 notable events from 1960 from Begumpet, Hyderabad, India, just the headlines. limit to 30 words in total.



ChatGPT

1. Begumpet Airport Begins Modernization.
2. Hyderabad Public School Adds New Facilities.
3. Begumpet Hosts Regional Science Exhibition.
4. Expansion of Green Park Area in Begumpet.
5. New Flyover Project Announced in Begumpet.

ICSE
PORT
UGAL 24

ICSE 24

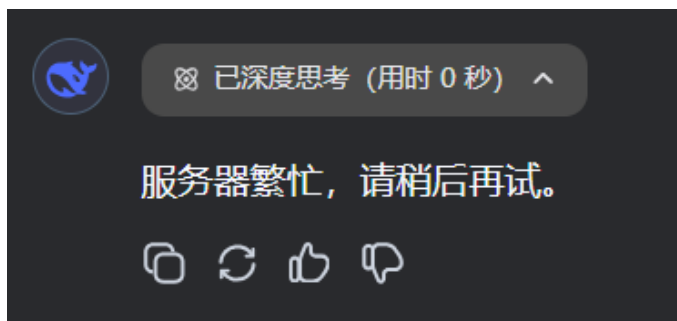
如何更好地与大模型“共舞”



- 首先接受大模型的“不确定性”
 - 因为这同时是大模型能力的“源泉”
- 从基于控制的系统中学习LLM (AI) 系统设计
 - 系统工具化时需构建适合的交互界面
 - 便于快速评估大模型的输出结果并及时纠错
 - “人在循环中” - 持续对系统输出进行反馈
 - 引入团队经验与先验知识 – LLM-Prompt
 - 结构化模型的输入/输出以方便快速展开分析
 - 显式引入业务相关的领域模型与奖励模型 – LLM-RAG
 - 融合多种输入源以相互印证
 - 中长期目标：在海量专用项目数据上训练特化大模型



■ 项目背景 – 大模型时代

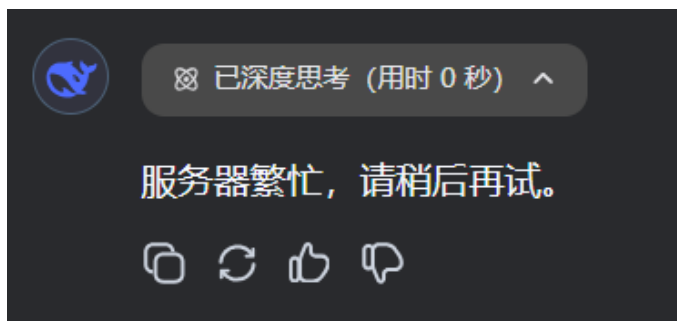


■ 项目背景 – 大模型时代

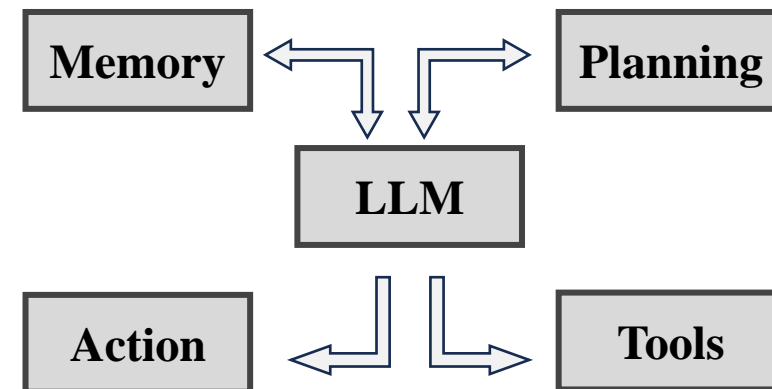


不只是聊天机器人：

1. 代码：代码生成、代码建议
2. 文生图像视频
3. 大模型+机器人
4. 领域大模型：医疗、法律...



■ 项目背景 – 大模型的局限



1. 难以实现真正的端到端服务 → 不够“智能” → Agent (智能体)

2. 静态落后的预训练数据 → 幻觉 → RAG



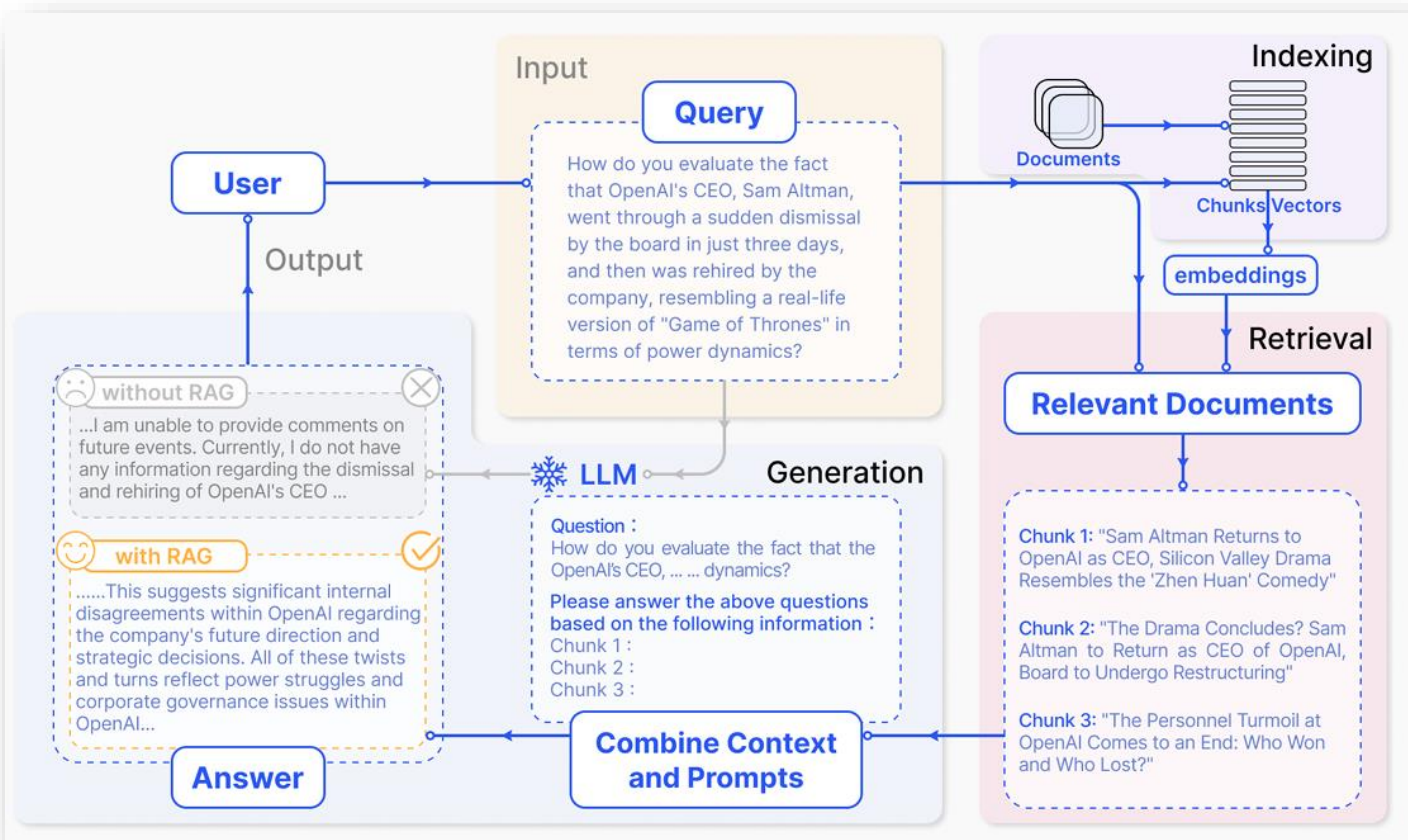
[ChatDev](#): 一个由各种Agent构成的软件开发公司

■ 项目背景 – RAG

检索增强生成 (Retrieval-Augmented Generation, RAG) 通过信息检索算法, 从外部知识库中筛选与用户问题相关的信息片段, 并将这些片段与用户问题同时传送给大模型, 从而增强回答的事实准确性和上下文相关性。

与微调 (Fine-tune) 相比:

1. **灵活**
2. **可解释**



■ 项目介绍 – Overview

项目目标： 搭建一个**RAG应用**以及与之配套的**评估系统**

项目安排： 三次迭代，学习 → 实践 → 优化

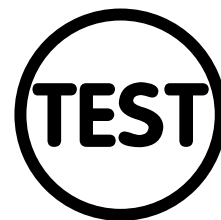
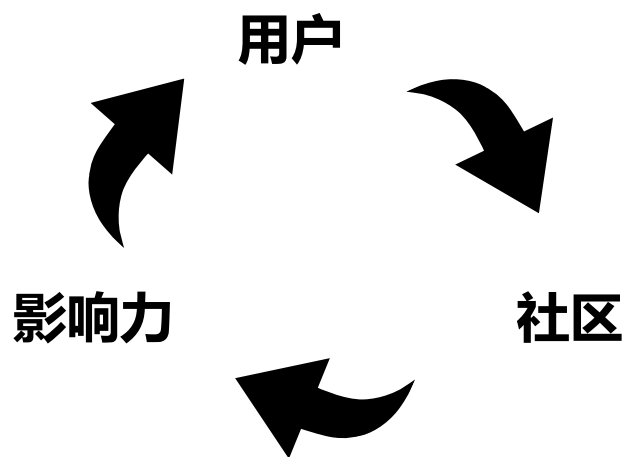
RAG应用主题： 不限，学习、就业、生活、游戏、动漫等等均可，建议选一个拓展性高的主题

项目产物： RAG应用 + 评估系统 + 各种文档 + 过程产物（代码提交记录等）

推荐使用免费大模型API：glm-4-flash，如需使用其他付费API，小组自行调用

■ 项目介绍 – 迭代一

从一个**开源**项目出发，首先学习**RAG评估**



测试驱动开发

■ 项目介绍 – 迭代一

- 在搭建RAG应用之前，首先要知道**什么是一个好的RAG**



1. 知识库完备;
2. 检索内容质量高;
3. 生成内容质量高;

■ RAG评估

评估依据: 1) 用户输入query; 2) 检索得到的上下文信息retrieved context; 3) LLM生成的回答generated answer
4) 正确的召回上下文信息

评估指标:

Context Relevance

用于评估召回上下文的精准度

Answer Faithfulness

LLM回答是否忠于召回的上下文

Answer Relevance

LLM生成的回答与用户问题的相关性

其他指标: Noise Robustness, Information Integration ...

■ 项目介绍 – 迭代一

阶段任务：基于开源项目Ragas，学习RAG评估，完成实验和开源项目阅读报告

■ 开源项目：



Python项目 10957行

仓库链接：[explodinggradients/ragas: Supercharge Your LLM Application Evaluations](https://github.com/explosion/ragas) 🚀

项目介绍： Ragas是一个用于评估大模型应用的开源项目，特色是评估RAG系统，也支持评估大模型Agent，或者用于评估简单的大模型任务。Ragas库中内置了各种的评估指标，比如上下文精准度、上下文召回率、可信度、回答相关性、回答正确性等等。这些指标按照计算过程中是否使用大模型可以分为**大模型驱动**的指标和**非大模型驱动**的指标；按照评估场景可以分为**RAG指标**、**Agent指标**、**自然语言比对指标**、**自然语言转SQL任务指标**、**通用指标**等等。

■ 项目介绍 – 迭代一

阶段任务：基于开源项目Ragas，学习RAG评估，完成实验和开源项目阅读报告

■ 例子：RAGAS的一个指标

Factual Correctness

- Ragas提供的一种**大模型驱动**的指标
- 用于评估response的事实准确度，需要用到response的内容和reference
- 具体实现方式见右图

The formula for calculating True Positive (TP), False Positive (FP), and False Negative (FN) is as follows:

True Positive (TP) = Number of claims in response that are present in reference

False Positive (FP) = Number of claims in response that are not present in reference

False Negative (FN) = Number of claims in reference that are not present in response

The formula for calculating precision, recall, and F1 score is as follows:

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

■ 项目介绍 – 迭代一

阶段任务：基于开源项目Ragas，学习RAG评估，完成实验和开源项目阅读报告

■ 例子：RAGAS的一个指标

Prompt-1

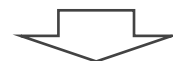
将每个输入句子分解为一个或多个独立的语句。
每个陈述都应该是一个可以独立验证的独立声明。
按照示例中所示的原子性和覆盖率级别进行操作。
{示例}



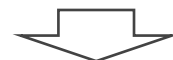
```
▼ VARIABLES
  ▼ Locals
    ▼ claims_list = ['黄仁勋强调，计算机图形学和人工智能的结...
    > function variables
      00 = '黄仁勋强调，计算机图形学和人工智能的结合是生成高...
      01 = '人工智能能够精准生成大量像素。'
      02 = '人工智能能够提升图像表现力。'
      03 = '黄仁勋将Cosmos比作物理世界的基础模型。'
      04 = 'Cosmos模型类似于ChatGPT对文本处理的作用。'
      05 = 'Cosmos模型能通过模拟和理解物理世界的动态规律来开...
      06 = '人工智能将在视频游戏行业中发挥重要作用。'
      07 = '人工智能能够降低内容创作成本。'
      08 = '人工智能能够使游戏角色更加智能。'
      09 = '人工智能能够提升互动性和游戏的趣味性。'
      10 = '黄仁勋认为，通用机器人所需的技术正在加速整合。'
```



Prompt-2



你的任务是根据给定的上下文判断一系列陈述的可信度。
对于每条语句，如果可以根据上下文直接推断出语句，则必须将判断结果返回为1，如果不能根据上下文直接推导出语句，请返回0。



TP/FP

Factual Correctness

■ 项目介绍 – 迭代一

阶段任务：基于开源项目Ragas，学习RAG评估，完成实验和开源项目阅读报告

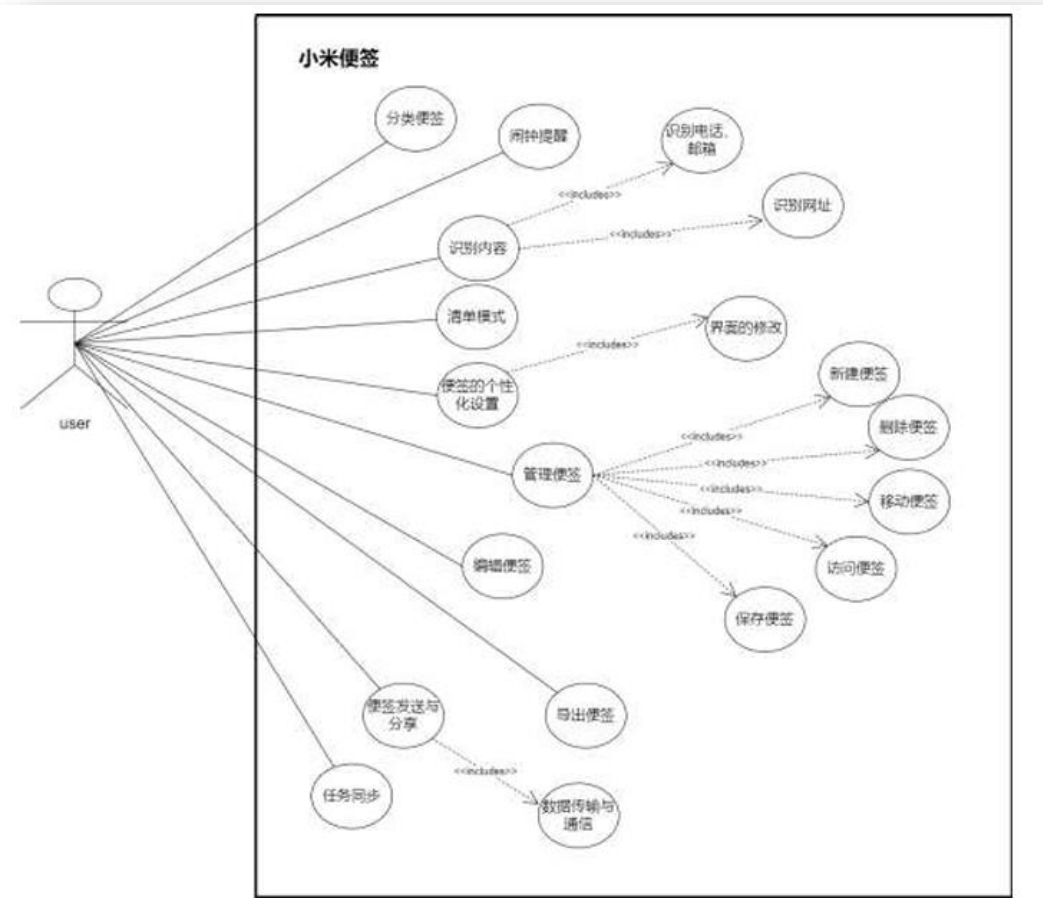
具体任务（详细要求见迭代一任务说明）：

1. 阅读开源项目，完成开源项目阅读报告
2. 根据moodle上给出的数据集，自己复现Ragas中的指标，对其进行评估，形成实验报告

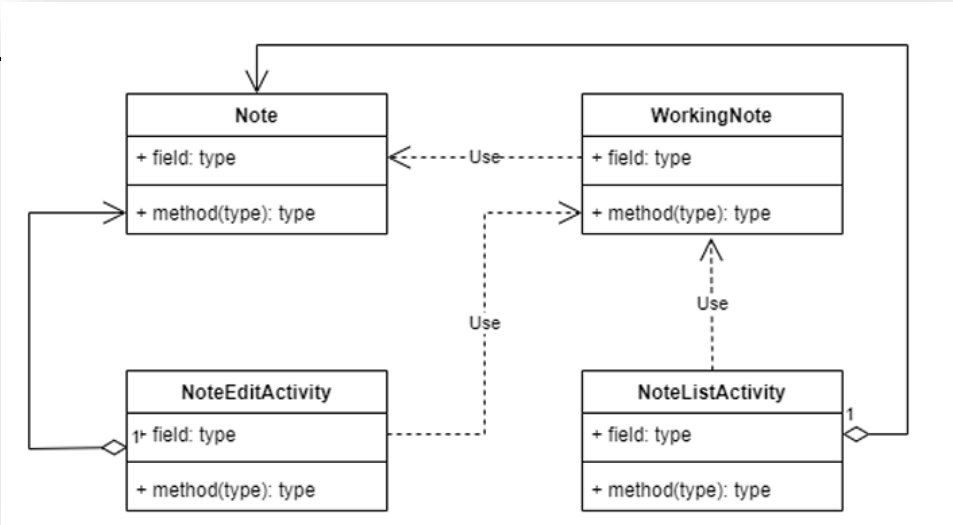
■ 项目介绍 – 迭代一

开源项目阅读报告：

用例图



类图



功能-代码对应

序号	功能名称	实现模块	实现方法
1	新建便签	NoteEditActivity Notes WorkingNote	createNewNote() getFolerId() startActivity()
2	删除便签	NoteEditActivity Notes WorkingNote	deleteCurrentNote() getNodeId() getContentResolver()
3	移动便签	DataUtils Notes WorkingNote	moveNoteToFolder(), batchMoveToFolder()
4	显示创建或修改文件夹的对话框并处理	NotesListActivity Notes WorkingNote	showCreateOrModifyFolderDialog() showSoftInput() setOnClickListener() addTextChangedListener()

■ 项目介绍 – 迭代二

阶段任务： 搭建RAG应用以及配套的评估系统

具体任务（以迭代二发布的任务说明为准）：

1. RAG应用：主题不限，基本的问答功能配上简单的ui界面即可
2. 评估系统：RAG评估， prompt评估两个模块

■ 项目介绍 – 迭代三

阶段任务： 深入探索

具体任务（以迭代三发布的任务说明为准）：

- （产品思维）： RAG应用产品化
- （RAG优化）： 优化迭代二完成的RAG应用
- （评估系统优化）： 评估系统功能升级

课程任务介绍到此为止，欢迎
各位同学与课程团队多多交流

后面有开头内容的backup (cai) slides (dan)

梅宏院士：对人工智能热潮的冷思考

- 现阶段AI的成功源于深度学习，这只是AI研究的一个子领域，其本质是数据驱动的智能、计算实现的智能，即“数据为体、智能为用”，核心要素为“算法、数据、算力”
 - 犹如燃料与火焰的关系，燃料越多，火焰越大，燃料越纯，火焰越漂亮
- 大模型可被视为是由已有语料压缩而成的知识库，生成结果的语义正确性高度依赖于数据的空间广度、时间深度以及分布密度，**更高度依赖于数据的质量**
 - 对AI落地应用的建议：**积累数据 – 可采尽采、能存尽存**
- 理想的AI应该是应当是低熵的，不以计算资源的消耗换取智能，也不以复杂性的提高换取智能
 - 科学家在探索自然的过程中，一直在追求为世界建模，遵循的基本原则是简而美

梅宏

CCF会士、前理事长。中国科学院院士。北京大学教授。主要研究方向为系统软件、软件工程。meih@pku.edu.cn



中国计算机学会

中国计算机学会官方订阅号，为CCF会员及计算领域的专业人士服务。

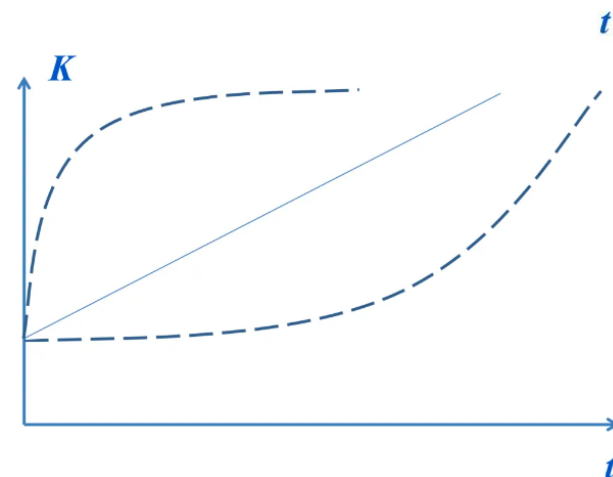
1177篇原创内容



阿里云张刚：AI夺不走软件工程师的工作



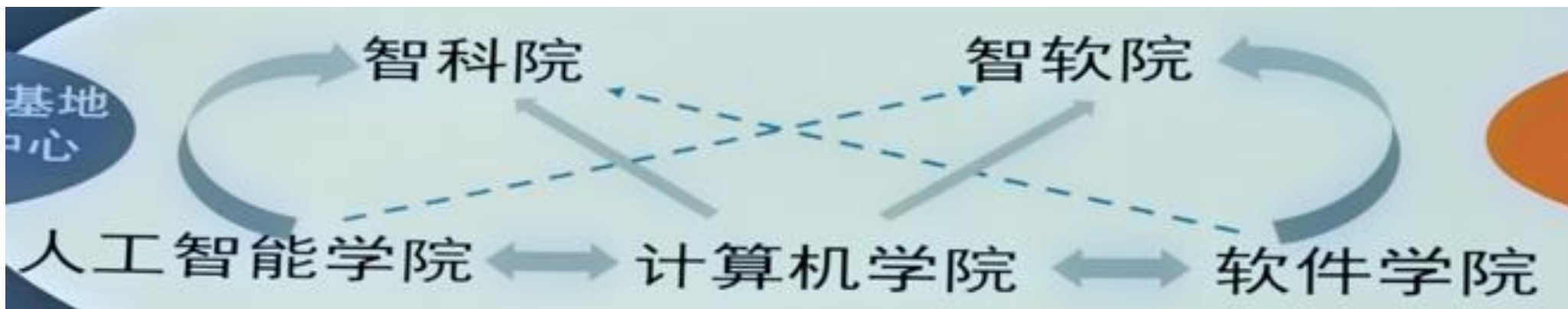
- 软件开发的三个核心问题
 - 1. 弄清楚要解决什么问题
 - 2. 设计出合理的解决方案
 - 3. 把它构建出来并交付使用
- 软件工程师的工作：不是写代码
 - 1&2都是探索与发现（“设计”），只有3是建造
 - 编码占开发的两成，读代码与写代码的比例是10：1
- 软件无法分离“设计”与“建造”
 - 软件的复杂性远超一般的人类制品
 - **软件价值本质上来自创新，而不是生产（复杂的信息制品）**
- **软件工程师如何与大模型“共舞”**
 - 大模型具备远超常人的知识，但不含具体问题的上下文，因此**大模型能力的发挥取决于使用者的能力 – 能力来源：长期的开发训练**
 - 软件工程师需要**通过编码来学习求解具体问题（包含对“设计”的理解）**，而**大模型可以加速编码的学习过程，最大化探索和发现的效率**



大模型最佳应用场景 – 超越学习、帮助探索



- “不是2B，不是2C，而是**2D**”
 - **D – Developer & Designer**
 - **帮助创新者更好展开探索**（持续交互+高效总结），**需要创新者驾驭大模型**
- **张刚博士**：今天的程序员，应该升级我们的技术栈，这包括了如何高质量的分析需求、如何掌握软件设计的原理，如何更快的理解和学习技术框架等等，同时，也需要理解如何更好的和大模型沟通
- **梅宏院士**：科学发展的第一驱动力是好奇心，即人们希望去了解其所生存的世界 – 利用大模型去学习和探索，而不是被大模型支配
- 此时再理解**南大计算机学科**为何要开设**五个院系**？**彼此的特色是？**为何又**同属一个软件新技术全国重点实验室**



“智能、计算、软件”的再审视

- “软件”主导创新&承载业务，“计算”支撑业务&积累数据，“智能”挖掘数据&赋能业务
- 三者虽各有侧重，但紧密关联，形成“互联网+”的IT技术、领域、行业
 - 本质上是一种创新业务驱动下的海量用户共同创造 **白嫖**
 - “智能、计算、软件”都要通过软件创新实现（业务、底层、学件）
- 各位同学最大化自身价值的源泉：创新
 - “创新能发掘新兴资产并掌握初始定价权”
 - 技术创新：提升计算能力，改进业务效率
 - 业务创新（“创业”）：满足社会需要，积累真实数据
 - 科研创新：钻研未知问题，拓展认知边界
 - 在 创 新 中找到专属于自己的独特定位，书写自己的 传 奇 人 生