日常想法随手记-2025

2025年5月13日 4 分钟阅读

#AI #Thinking #Daily #2025

> *日常想法随手记*

这里随手记下当天对看到的新闻、文章、视频等的想法。

2025 October

2025-10-31

GitHub Copilot、VS Code 团队以及微软开源项目办公室（OSPO）共同资助了这九个开源 MCP 项目，这些项目提供了新的框架、工具和助手，以开启原生人工智能工作流、智能代理工具和创新。link:

fastapi_mcp: 通过最少的设置、身份验证和有限的配置，将安全的FastAPI端点作为MCP工具暴露出来，所有这些都借助统一的基础设施。 nuxt-mcp: 用于路由检查和 SSR 调试的 Nuxt 开发者工具，能让你的团队更轻松地让模型更好地理解你的 Vite/Nuxt 应用。 unity-mcp: Unity MCP 允许你对接游戏引擎的应用程序接口（API），以进行人工智能辅助的游戏开发，并为你的人工智能工具提供在 Unity 中管理资源、控制场景、编辑脚本和自动化任务的能力。

context7: 直接从你的代码中提取最新的、特定版本的文档和代码示例，并将它们直接插入到你的人工智能和大型语言模型提示词以及大型语言模型的上下文中。

serena: 用于代理驱动的编码代理工具包的语义代码编辑和检索，该工具包提供语义检索和编辑功能。

Peekaboo: Swift代码分析能将你屏幕上的内容转化为可操作的人工智能上下文，以创建完整的图形用户界面自动化，并且可用于人工智能助手。 coderunner: Coderunner将大型语言模型（LLMs）转变为即时的本地执行伙伴，它能在你的机器上一个预先配置好的沙箱中编写和运行代码，自动安装工具，直接读取文件，并返回输出结果和生成的工件。

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#AI #Thinking #Daily

n8n-mcp: n8n-MCP 是一个经过超优化的平台，它通过简化工作流的创建和编排，增强了 n8n 的工作流自动化能力。它集成了 AI 模型，帮助用户更好地理解 n8n 节点并与之协同工作。 inspector: 一种用于测试和调试MCP服务器的工具，可检查协议握手、工具、资源、提示词和OAuth流程。它提供内置的LLM实验环境，并允许运行评估模拟以捕捉安全性或性能回归问题。

2025-10-30

今天读了OpenAI的多智能体合作研究，这篇论文展示了如何使用OpenAI的多智能体SDK实现合作的智能体系统，这些智能体可以共同完成复杂任务，如股票交易、项目管理等。研究结果表明，通过有效的协作机制，智能体系统的整体表现显著优于独立智能体。这为构建更智能、更高效的AI系统提供了新的思路。

2025-10-27

xAI 于 2025 年 10 月 27 日发布了 **Grokipedia** 的测试版，该版本包含约 88.5 万篇由 Grok 模型生成的文章，并整合了来自 X 平台帖子等来源的实时数据。该平台具备用户标记纠错功能、事实核查时间戳以及开源访问权限，马斯克承诺在未来的更新中会大幅提升其准确性。

今天读了Verbalized Sampling: 言语采样提升模型多样性，这篇论文提出了一种名为**口头化采样（Verbalized Sampling, VS）** 的推理时提示方法，该方法通过明确要求模型输出多个可能的响应及其对应的概率分布，有效地恢复了LLM在预训练阶段所学的内在多样性。实验证明，VS显著提升了模型在创意写作、对话模拟、开放式问答和合成数据生成等任务中的输出多样性，同时保持了事实准确性和安全性。这个方法非常简单，而且效果非常好，值得推广。

2025-10-15

宾夕法尼亚大学研究发现（论文《Mind Your Tone》），向GPT-4o等新一代大语言模型使用简洁直接或略带粗鲁的提问，可减少冗余修饰语，聚焦核心问题，从而将多项选择题准确率从约80.8%提升至84.8%；但该结论仅基于多选题实验，且不同模型对语气敏感度各异，应用时需在效率与礼貌间权衡。

2025 September

2025-09-05

Gemini Nano Banana的各种创意玩法。高度的人物一致性让故事创造，人物换装，角色扮演，手办制作等成为可能。
https://github.com/ZHO-ZHO-ZHO/ZHO-nano-banana-Creation
https://github.com/PicoTrex/Awesome-Nano-Banana-images
https://waytoagi.feishu.cn/wiki/IffFw8eTaiAfBHk12i5cdRk5nsO
https://waytoagi.feishu.cn/wiki/RuGMwHKZ6isFgFkqWCHc2xk8n2Q
https://waytoagi.feishu.cn/wiki/ItkqwpDc8i8H79kZfbEciTe9nhe

2025 August

2025-08-25

麻省理工学院的一份新报告在企业人工智能领域引起了轩然大波：95% 的生成式人工智能试点项目投资回报率为零。报告。40% 的组织表示已部署了人工智能工具，但仅有 5% 的组织成功地将这些工具大规模融入工作流程。大多数项目都停滞在试点阶段。与此同时，媒体头条纷纷警告存在“人工智能泡沫”，一些投资者基于生成式人工智能在企业中的重大机遇已停滞不前这一观点而做空人工智能股票。

2025-08-20

8 月 18 日，智谱正式发布了新的 ToC 产品 AutoGLM 2.0——一个手机通用 Agent。支持安卓和IOS。

2025-08-17

How To AI (Almost) Anything 腾讯研究院：2025 AI Coding非共识报告 OpenAI Multi-Agent Portfolio Collaboration

We Made Claude Code Build Lovable in 75 Minutes (With No Code)

2025-08-04

2025-08-03

8月3日晚，为期5天的AI Olympic 2025在巴黎拉德芳斯体育馆圆满结束。此次赛事设有代码、数学推理、多模态问答和实时策略四个项目，共吸引了36支顶尖队伍参赛。中国“紫霄-AI”战队以3金1银的

优异成绩首次获得综合冠军，而美国的OpenAI和法国的Mistral分别获得亚军和季军。闭幕式上，国际奥委会宣布2027年的比赛将新增“具身智能”和“AI医疗救援”两个新赛道。

2025-08-01

美国《AI 主权法案》于8月1日凌晨在参议院以67票对33票获批，总统当天上午10点签署生效。该法案将≥10^{25} FLOPs训练模型的权重、超参数和训练日志列为“敏感技术”，并对违反出口规定的企业处以全球营收的20%罚款，并追究高管的刑事责任。硅谷超过300家初创公司连夜召开了合规峰会，各大基金如红杉、光速等也紧急调整了投资组合。

2025 July

2025-07-22

Gemini Deep Think成功破解IMO(国际数学奥林匹克竞赛)前5题，斩获35分（满分42分）。之所以能获得金牌，核心在于它采用了端到端的自然语言推理能力，直接从英文题目描述生成严谨的数学证明。通过高级并行思考技术,同时探索多条推理路径，模拟人类深度思维过程和强化学习训练，Gemini能够在4.5小时内高效探索多条解题路径，并整合最优解法，成功攻克五道高难度数学题，最终获得官方认可的金牌成绩。这一突破不仅体现在解题速度和准确性，更在于AI首次无需形式化语言辅助，完全以自然语言实现了顶尖数学推理水平。标志着AI从被动问答转向主动解决复杂科学问题，为科研、工程等领域带来范式变革。

2025-07-19

2025-07-20

> 今天写了Claude Code 介绍，并把它和Kimi K2模型关联了起来，准备关联更多的MCP server，然后对比一下Claude Code，Cursor，Gemini Code的优劣。

2025-07-04

> 今天读了Reflect, Retry, Reward: 大型语言模型的自我进化新范式，这篇论文的结论是：通过“反思、重试、奖励”方法在微调过程中通过自我反思和强化学习，实现LLM在缺乏外部监督和特定任务数据情况下的自我改进。

2025-07-01

今天读了深度研究智能体：系统性审查与路线图，这篇论文的结论是：**深度研究智能体（DR）是基于大型语言模型（LLMs）的智能体，能够通过结合动态推理、自适应长时规划、多跳信息检索、迭代工具使用以及生成结构化分析报告来处理复杂的、多轮的信息研究任务。**

2025 June

2025-06-30

百度开源了他的第一个开源大模型ERNIE4.5。Tech Report: 我对百度开源500B大模型ERNIE4.5的技术报告的解读

2025-06-29

以下是当前业界对Claude Code、OpenAI CodeX和Gemini CLI三大AI编程工具的评价对比，综合技术能力、用户口碑、性价比及适用场景分析：

技术能力与核心优势对比

工具	技术亮点	主要短板
Claude Code	- 深度代码库理解能力，支持跨文件协调修改和复杂重构 - 终端原生集成，自动化Git操作（提交、PR、冲突解决） - 混合推理模式（快速响应+深度思考），SWE-bench测试解决率70.3%	学习曲线陡峭，订阅成本高（Max套餐约¥1600/月）
OpenAI CodeX	- 代码生成质量稳定，GitHub Copilot的底层引擎 - 支持多种审批模式，代码可控性强	功能创新不足，缺乏多模态和终端深度集成
Gemini CLI	- **100万Token上下文窗口**，支持多模态输入（图像、PDF） - **开源免费**（Apache 2.0），每分钟60次，每日1000次免费请求 - 原生支持Windows，集成MCP协议扩展功能	早期版本稳定性不足（内存泄漏问题），没有cursor好用。但是前景光明。

2025-06-17

AI的越来越强大，将成为越来越大的杠杆，想不被拉下，需要尽早使用AI，并不断提升自己的技能水平。

2025-06-16

MiniMax 推出了他们的首个推理模型： MiniMax M1，这是继 DeepSeek R1 之后第二款最智能的开放权重模型，拥有更长的 100 万个标记的上下文窗口。

> M1 基于其于 2025 年 1 月 14 日发布的 Text-01 模型——这是一个拥有总计 4560 亿参数和 459 亿活跃参数的混合专家模型。这使得 M1 的总参数数量小于 DeepSeek R1 的 6710 亿参数，但大于 Qwen3 的 2350 亿 - 220 亿。Text-01 和 M1 均仅支持文本输入和输出。
>
> MiniMax M1 80K 在人工智能分析指数上得分为 63 分。这落后于 DeepSeek R1 0528，但略高于阿里巴巴的 Qwen3 235B-A22B 和 NVIDIA 的 Llama 3.1 Nemotron Ultra。MiniMax M1 提供两个版本：M1 40K 和 M1 80K，分别提供 40k 和 80k 的令牌思考预算。
>
> MiniMax 披露，他们使用 512 块 H800 GPU 对 Text-01 进行了为期三周的完整强化学习训练（非预训练）以创建 M1，这相当于 53 万美元的租赁成本。

2025-06-12

> 今天读了阿里云AI搜索Agentic RAG技术实践，这篇论文描述的RAG技术演进之路具有普遍性。而我认为，Agentic RAG 2.0 虽然和Deep Research有一定相似之处，但是前者是“工程驱动的智能协作”，而后者的目标“智能驱动的自我进化”，目标上有差异，但技术上有相互借鉴之处。

2025-06-11

今天找了一题给不同的推理模型做，答案是D是酸，A是碱。只有Gemini 2.5 pro, Cloud 3.7 Sonnet Thinking, 元宝R1 答对.奇怪的是DS R1在推理了非常非常非常久后，得出了错误答案，但是元宝R1却答对了。

```
实验室中有五种无色液体：A、B、C、D、E。已知条件：

    A 和 B 混合 → 产生蓝色沉淀；
    B 和 C 混合 → 无反应；
    C 和 D 混合 → 溶液变红色；
    D 和 E 混合 → 生成无色气体；
    A 和 E 混合 → 无反应。

其中一种液体是酸，一种是碱，其余三种为中性物质。

问题：请推断哪种液体是酸？哪种是碱？并解释推理过程。
```

模型	酸	碱
O3/4O	D	E
G-2.5p	D	A
Claud-37T	D	A
Grok3	D	B
R1	D	B
元宝R1	D	A

2025-06-10

互联网女王“Mary Meeker的AI趋势报告，https://www.bondcap.com/reports/tai

2025-06-05

今天读了General agents need world models，这篇论文的结论是：**智能体本身就是世界模型。**

2025-06-04

今天读了OmniThink: Expanding Knowledge Boundaries in Machine Writing through Thinking，这篇论文的结论是：**OmniThink 是一个基于信息树和概念池的“慢思考”方法，旨在解决机器写作中的知识边界扩展问题。**

2025-06-03

最近一直在对比各种DeepResearch开源项目实现是如何实现“Deep”和“Research”的。

2025 May

2025-05-31

通义灵码 AI IDE 正式上线，深度适配千问3大模型，集成智能编码助手功能，支持自动记忆和工程感知，提升开发者编程效率和体验。支持 MCP 协议，覆盖 3000 多个 MCP 服务，满足多场景开发需求，如快速创建地图类应用。link

2025-05-30

最新公布的DeepSeek R1-0528 非常不错。

Category	Benchmark (Metric)	DeepSeek R1	DeepSeek R1 0528
General	MMLU-Redux (EM)	92.9	93.4
	MMLU-Pro (EM)	84.0	85.0
	GPQA-Diamond (Pass@1)	71.5	81.0
	SimpleQA (Correct)	30.1	27.8
	FRAMES (Acc.)	82.5	83.0
	Humanity's Last Exam (Pass@1)	8.5	17.7
Code	LiveCodeBench (2408-2505) (Pass@1)	63.5	73.3
	Codeforces-Div1 (Rating)	1530	1930
	SWE Verified (Resolved)	49.2	57.6
	Aider-Polyglot (Acc.)	53.3	71.6
Math	AIME 2024 (Pass@1)	79.8	91.4
	AIME 2025 (Pass@1)	70.0	87.5
	HMMT 2025 (Pass@1)	41.7	79.4
	CNMO 2024 (Pass@1)	78.8	86.9
Tools	BFCL_v3_MultiTurn (Acc)	-	37.0
	Tau-Bench (Pass@1)	-	53.5(Airline)/63.9(Retail)

Model	AIME 24	AIME 25	HMMT Feb 25	GPQA Diamond	LiveCodeBench (2408-2505)
Qwen3-235B-A22B	85.7	81.5	62.5	71.1	66.5
Qwen3-32B	81.4	72.9	-	68.4	-
Qwen3-8B	76.0	67.3	-	62.0	-
Phi-4-Reasoning-Plus-14B	81.3	78.0	53.6	69.3	-
Gemini-2.5-Flash-Thinking-0520	82.3	72.0	64.2	82.8	62.3
o3-mini (medium)	79.6	76.7	53.3	76.8	65.9
DeepSeek-R1-0528-Qwen3-8B	86.0	76.3	61.5	61.1	60.5

2025-05-29

学习需要结合从上到下的整体理解和从下到上的基础夯实，是一个动态的过程，在不同场景中灵活切换——有时先从上到下（如程序员熟悉新开源代码），有时先从下到上（如学生掌握相对论），有时是以一个为主，另一为辅等等，以实现更有效的知识掌握和应用。

2025-05-27

> **微软/Meta/OpenAI Distinguished Engineer- Philip Su访谈, 可以看这里**

2025-05-22

我给Google IO 2025做了个总结

Google Deep Research新提供了几个功能：用户可以将自己的文件上传到DR平台，并通过在 Canvas 中将它们转化为互动内容、测验、音频概述等，让你的报告更具沉浸感。后面考虑写一篇使用Deep Research的教程。

2025-05-21

“Vibe Coding”（“氛围编程”）指的是一种使用 AI 模型进行软件开发的方式，在这种方式下，开发者更关注高层级的需求和意图，而不太关心底层的技术实现细节。与传统的编程相比，Vibe Coding 更侧重于通过自然语言与 AI 交流，让模型生成大部分代码，开发者主要关注结果是否符合预期。这使得非专业开发者也能参与软件创建，实现个人实用程序。然而，这种方式也存在挑战，因为生成的代码可能对 Vibe Coder 来说是黑箱，难以理解和修改底层逻辑。

2025-05-20

2025-05-18

对于动物图片的识别，AI为什么需要大量的数据才能学会，而不像人一样通过少量的数据就能学会？比如你教小baby认猫，只需要给他看几张图片，听你说，他就学会了，下次见到即使不一样的猫，他也能认出来。但是AI必须通过大量的数据（大量的猫的照片）才能学会。我今天的一个想法是，深度模型的参数很多，所以它对一个事物所能提供的“Feature（特征）”也多，这些特征往往是人类不可知的，但是它却能通过这些“暗”特征来识别事物，大模型能把整个人类知识囊括到大模型这个超大“函数”里，这个“暗feature解构能力功不可没，但也从另一方面导致了，它想认识一个事物，仅凭几张图片是不行的，因为不足以生成“暗”特征。也就是说，它缺少了“归纳”的能力，它只能根据已有的数据来识别事物，而不能根据事物之间的关联来识别。人类的大脑经过多年的进化，已经具备了“归纳”，也就是浓缩主要信息，并根据主要信息来在不同的对象之间快速建立联系来达到认识相似的事物的能力，所以人类只需要少量的数据就能学会一个新事物。

2025-05-17

2025-05-16

2025-05-15

Grok告诉我：“微软近期裁员 6000 人的最根本原因在于通过裁员来控制成本、提高运营效率，同时继续在人工智能领域大力投入。”

2025-05-14

Devin 的 deepwiki 是一款非常出色的 AI 网页工具，它能够根据 GitHub 仓库自动生成相关的知识库文档，并为用户提供一个名为 “DeepResearch” 的问答窗口，方便高效地检索和交流技术内容。

2025-05-13

在2025年4月的最新模型（o3、GPT-4.1）上，医生参考AI输出后，已无法进一步提升AI答案，在医疗安全场景下，“最差表现”比“平均表现”更关键。提升极端case下的稳健性，将是下阶段AI医疗模型优化的重心。(参考OpenAI HealthBench)

2025-05-12

从红杉资本2025 年 AI 峰会的上获得的一些思考

> Agent经济：Agent as a Service将替代SaaS。软件的合作将变成Agent之间的合作。Agent不仅在软件层面，更会在物理层面操作现实世界的设备。商业模式从卖“工具”转向卖“成果”,卖“Agent”劳动力。物理层面也需要“图灵测试”和“物理API”，具身智能的标准化需要提到日程。
>
> Agent技术：RL将在Agent的世界继续发扬光大，Deep Research的下一步是科学原创，比如发现广义相对论级别的科学理论。另外多Agent的协作和安全需要新技术的加持。
>
> 基础模型：当前的pre-train模型在数据用尽的情况下基本到头，需要新的模型技术突破。
>
> 基础设施加大投资：在Agent经济不断扩大的前提下，推理需要的硬件将继续增长。

2025-04

吴恩达新作《如何在人工智能领域建立你的职业生涯》英文原版，link 中文翻译校对版：link

引言：AI编程是新的读写能力 第一章：职业生涯发展的三个步骤：学习，项目，找工作 第二章：学习建构人工智能职业生涯的关键技能

基础机器学习技能

深度学习

与机器学习相关的数学

软件开发 第三章：你应该学习数学来获得人工智能的工作吗?

面向 AI/ML 工作只需掌握能支持关键建模与调试决策的那层数学（随技术成熟所需深度会下降），同时保留由好奇心驱动的自由探索以孕育创新。 第四章：成功 AI 项目的范围

人工智能架构师最重要的技能之一就是能够识别出值得开展的想法。

步骤 1、确定业务问题（而不是人工智能问题）。

步骤 2、头脑风暴 AI 解决方案。

步骤 3、评估潜在解决方案的可行性和价值。

步骤 4、确定里程碑。

步骤 5、预算资源。 第五章：寻找与你的职业目标互补的项目

通过连续、难度递增且与职业目标互补的项目实践（小起步→快迭代→择优放大），用"准备-开火-瞄准"式低成本试错与必要时"准备-瞄准-开火"的前置论证相结合，加速技能复利与影响力攀升。 第六章：建立能够展示技能进步的项目组合

建立一个项目组合，特别是一个随着时间的推移从简单到复杂的项目组合，在找工作时将大有帮助。 第七章、开启人工智能求职的简单框架 第八章：使用咨询性面谈找到合适的工作

在进行咨询性面谈前，提前研究面试者和公司，以便带着深思熟虑的问题参加面试。你可以问以下问题：
你通常一周或一天的工作内容是什么？

这个职位最重要的任务是什么？

成功最需要哪些技能？

你的团队如何合作以实现目标？

招聘流程是什么？

对于过去表现突出的候选人，他们脱颖而出的原因是什么？ 第九章：找到适合你的AI工作 第十章：在人工智能领域建立事业的关键 第十一章：克服冒名顶替综合症

> **Reinforced Self-play...**
>
> 论文介绍了强化自博弈推理的零数...

> **Reinforced Self-play...**
>
> 论文介绍了强化自博弈推理的零数...