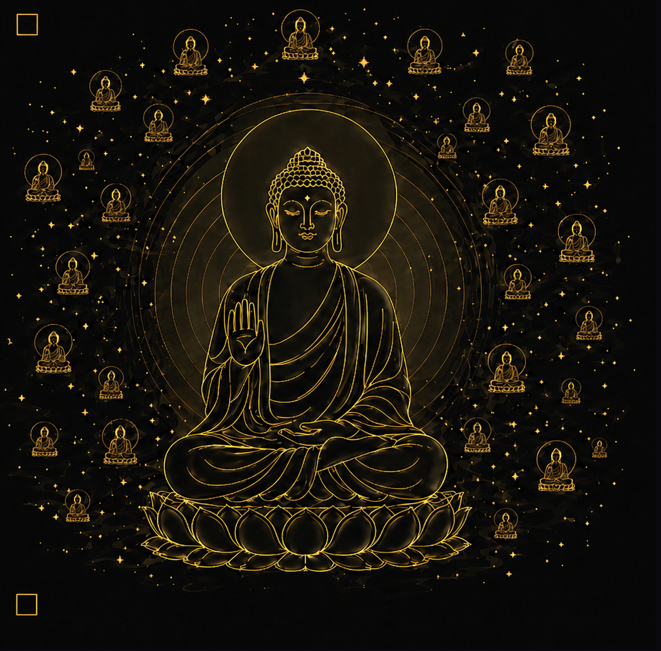

讲

果壳里的 AI：1500 年后的般若学

—— 第 一 讲 ——

幻觉与虚妄

从《金刚经》看 AI 幻觉

— 彭 一 楠 —

"Generative AI hallucinates —
it makes up answers that sound plausible but aren't based in fact."

生成式 AI 有个致命伤 ——
爱「一本正经地胡说八道」。

—— *黄仁勋（Jensen Huang）· NVIDIA CEO · GTC 2024 / HKUST 2024*

"In frontier models hallucination will grow ever rarer,
but it will never disappear — this is the nature of generative AI."

「幻觉在横向模型里会越来越少，
但它不会消失 —— 这是生成式 AI 的特点。」

—— *张亚勤 · 中国工程院院士 · 清华大学智能产业研究院（AIR）院长*

20 年前的旧稿，今日的续篇

2006 年

《果壳里的黑客》

上海某高校讲稿，41 张幻灯片

七段《金刚经》原典（梵汉对照）

与七段技术论说交替排布

以「应无所住而生其心」收尾

对象：网络黑客文化

20 年

2026 年

这次的讲座

「果壳里的 AI」系列六讲

以汉传佛教文本为思想资源

对接 AI 与算法的伦理

保留「举经-应用」交替体例

对象：AI 与算法伦理

本讲的论证路径

一

AI 幻觉

表面是技术问题，底里是认知问题

二

般若学的「虚妄」

较当代分类更为细密

三

概念对应

把这套辨析用在 AI 上

第 一 部 分

AI 幻觉的当代图景

Hallucination in Large Language Models

什么是 AI 幻觉?

***Hallucination*（幻觉）：** 大语言模型生成的内容看似合理 (plausible)，但与事实不符，或与上下文不符，或对自身知识边界做了错误判断。

三个特点：

看似合理

语句流畅，逻辑自洽。

实际错误

但与事实、上下文或自身能力不符。

难以识别

用户往往无从分辨。

案例一：港大社科院副院长论文的「幽灵引用」

香港大学社会科学学院副院长叶兆辉(Paul Yip)论文撤稿事件 | 2025 年 12 月

涉事论文：

《香港 40 年生育转变》(Forty Years of Fertility Transition in Hong Kong)

发表期刊：Springer Nature 旗下 *China Population and Development Studies* | 2025 年 10 月

61

篇

参考文献总数

24

篇

AI 虚构的「幽灵文献」

39%

论文引用伪造率

一处颇具讽刺的细节：

在 24 篇 AI 虚构的「幽灵文献」中，有数篇的署名作者**正是通讯作者叶兆辉本人**——他未察觉自己被「引用」了从未写过的文章。

结果：论文于 2025 年 12 月正式撤稿，叶兆辉卸任社会科学学院副院长职务 | 来源：Springer Nature 撤稿声明、新京报、虎嗅等多方报道

案例二：30 余家司法机关引用一部不存在的法规

「《中华人民共和国印章管理办法》」援引乱象 | 2019 — 2026

某基层法院《更换印章公告》（节选）：

「根据 **《中华人民共和国印章管理办法》** 及《最高人民法院关于地方各级人民法院和专门人民法院印章管理的规定》，最高人民法院批准并为其制发了新印章……」

公安部政府信息公开办公室回应（2026 年 4 月）：

「这一表述下的《印章管理办法》**从未出台或施行，不具有法律效力**。所有单位的引用都是不对的。」

超 10	30 余	7 年
个省份	家法院/检察院	持续时间（2019 — 2026）

起于 AI 之前，被 AI 放大

来源：新京报、求是网、上观新闻（2026 年 4 月）；公安部政府信息公开办公室回应

案例三：AI「小作文」一日蒸发 A 股龙头数十亿市值

亚钾国际(000893)AI 谣言事件 ｜ 2026 年 4 月 14 日

在投资者社交平台广泛传播的两条 AI 生成「小作文」（节选）：

① *「亚钾国际东泰产区因农田塌陷，自 2026 年 1 月被老挝能源矿产部停产，至今未恢复生产，年产 200 万吨产能损失……」*

② *「控股股东汇能集团拟将煤化工资产注入亚钾国际……」*

亚钾国际官方澄清：

两则消息均**「为 AI 生成的虚假信息」**，生产经营正常进行，从未收到所述股东资产注入信息。

-9.71%	499 亿	22.2 亿
当日盘中跌停	收盘市值（蒸发数十亿）	成交额（创近 18 月新高）

来源：中国证券报、新浪财经、澎湃新闻、东方财富网（2026 年 4 月 14-15 日）；亚钾国际官方微信公众号澄清声明

既有研究的幻觉分类——以及未及之处

学界目前的几种分法

Factuality vs Faithfulness

事实性幻觉 / 忠实性幻觉

Intrinsic vs Extrinsic

内源性 / 外源性幻觉

Closed-domain vs Open-domain

封闭领域 / 开放领域幻觉

代表综述：Ji et al. (2023)

ACM Computing Surveys

两处尚未触及

一、只到输出层，没到认知机制

现有分类只刻画错误的形态，
极少追问「为何会错」，
认知机制这一层基本留白。

二、「幻觉」一词覆盖不全

「幻觉」一词暗示「感知错误」，
但模型的失误远不止于「看错」——
它会高估自身的能力，
误读用户的真实意图，
也认不清自身的知识边界。

佛学与 AI：国际学界已有的工作

Buddhism & AI as an International Conversation, 2020 — 2025

主流：以佛学资源助 AI

Hongladarom (2020) · 朱拉隆功大学
《AI 与机器人伦理：一个佛教视角》（Lexington Books）——从上座部伦理学讨论 AI 的道德主体地位

Hershock (2021) · 东西方中心（夏威夷）
《佛教与智能技术》（Bloomsbury）——批判智能革命对人类注意力的「殖民化」

Doctor, Witkowski, Solomonova, Duane & Levin (2022)
"Biology, Buddhism, and AI: Care as the Driver of Intelligence"
Entropy 24(5)：将佛教菩萨概念引入 AI 智能架构，主张 care 驱动智能（塔夫茨/RYI）

Laukkonen et al. (2025) · 莫纳什 / 南十字星
《Contemplative AI》——以正念、空、不二、无量慈悲四原则用于 AI 对齐

Adam、Hershock、Amir & Dunne (2025)
《Contemporary Buddhism》AI 专刊：道德 AI、佛教与智能技术、Dharmakīrtian 模型等

本讲：以佛学诊断 AI

本讲聚焦

已有工作多取「**建设性立场**」——以佛学资源参与AI系统的设计与对齐工作。

本讲所取为「**诊断性立场**」——以汉传佛教自昙鸾、道绰以降的「自力修行有限」一脉，审察当代「自我反思」类对齐方法所遇到的边界。这一进路目前在国际学界尚乏人涉足。

第 二 部 分

佛学传统中的「虚妄」

Delusion in the Buddhist Tradition

【梵本】

यावत्सुभूते लक्षणसम्पत्तावन्मृषा ।
यावदलक्षणसम्पत्तावन्न मृषा ।
इति हि लक्षणालक्षणतस्तथागतो द्रष्टव्यः ॥

【鸠摩罗什译本】

凡所有相，皆是虚妄。
若见诸相非相，则见如来。

「虚妄」并非单一概念

日常理解

虚 妄 = 虚 假

一种二元判断：

- 真 / 假
- 实 / 虚
- 存在 / 不存在

不问「虚妄因何而生」

般若学的细致辨析

一个有内部结构的概念

三层贯通：

- 虚妄如何「显现」
- 显现的「认知地位」
- 如何被「识别」

唯识学对「虚妄」的三层分析

Three Natures (Trisvabhāva) · 本讲基于玄奘所传有相唯识

一

परिकल्पित *parikalpita*

遍计所执性

由分别心遍计度而生之虚妄

纯粹概念虚构，无所依据

最浅

二

परतन्त्र *paratantra*

依他起性

依因缘而起，因缘而灭

有依据，无自性

中等

三

परिनिष्पन्न *pariniṣpanna*

圆成实性

前二者的真实本然

非虚妄的实相

超越二元的觉悟之相——非西方哲学意义上的形而上学

最深

【梵本】

त्रैधातुकं चित्तमात्रम्
स्वचित्तदृश्यमात्रं तु
बहिर्द्रव्यं न विद्यते ॥

【求那跋陀罗译本】

觉自心现量，外性非性，
不妄想相。

唯识四分

Four Aspects （《成唯识论》卷二）· 护法-玄奘所传（印度另有安慧一分、难陀二分、陈那三分说）· 后二分梵语为拟构(*)

一

相 分

निमित्त-भाग

nimitta-bhāga

所认识之境（客体）

二

见 分

दर्शन-भाग

darśana-bhāga

能认识之用（主体）

三

自 证 分

स्वसंवित्ति-भाग

**svasaṃvitti-bhāga*

能证「见分」之用

四

证 自 证 分

स्वसंवित्तेः संवित्ति

**svasaṃvitteḥ saṃvitti*

证知「自证分」之用

中观双遣：龙树的论证形式

Madhyamaka's Method of Double Negation

龙树《中论·观因缘品》· 鸠摩罗什译 · 大正藏 T30, no. 1564

诸 法 不 自 生 ， 亦 不 从 他 生 ，
不 共 不 无 因 ， 是 故 知 无 生。

中观所遣，非「否 X」，而是「X 与非 X 之二元」

否定「自生」

否定「他生」

否定「共生」

否定「无因生」

被 破
（自性 svabhāva）

「四 句 破」

遍否四种
「生起方式」
破「自性」之执

第 三 部 分

四种幻觉的镜像分析

Four Hallucinations Mirrored

对应（一）

遍计所执 ↔ 事实编造型幻觉

परिकल्पित *parikalpita*

遍 计 所 执

由分别心遍计度而生之

纯粹虚构，

无现实所依。

⇄

Factual Fabrication

（属于 Factuality 事实性幻觉之子类）

事 实 编 造

模型生成的内容在训练数据

中并无所依，

全属凭空编造。

案例：港大叶兆辉论文中

24 篇虚构「幽灵文献」——

DOI 全部 Not Found

对应（二）

依他起性 ↔ 上下文偏移型幻觉

परतन्त्र *paratantra*

依 他 起 性

依因缘而起，

有所依而无自性，

遍计执方添「自性」而成虚妄。

⇄

Context-Misalignment

（属于 Faithfulness 忠实性幻觉之子类）

上 下 文 偏 移

模型所生之内容在训练数据

中确有所据，

惟被错置于不当的语境之中。

案例：2002 年公安部《印章治安管理办法（草案）》

被「去草案化」错置为正式法规

对应（三）

慢（本于末那我执） ↔ 过度自信型幻觉

मान · मनस्-आत्मग्राह *māna（慢）· manas-ātmagrāha（末那我执）*

慢 · 本于末那我执

第七末那识恒审思量，

执第八识见分为我。

凡夫位恒时不断，为我执之根；

其于「过度自信」中现起者，正是「慢」心所。

⇄

Overconfidence / Miscalibration

过 度 自 信

模型对「自己知道什么 / 自己能做什么」的置信度

与实际能力失校。

案例：全国首例 AI 幻觉侵权案

AI 给错高校信息，被指出后

仍坚称「赔 10 万也对」

过度自信案例：AI 被指出错误后仍坚称无误

全国首例生成式 AI「幻觉」侵权之诉 | 杭州互联网法院 2026 年 1 月一审判决

① 用户梁某 → AI： 查询某高校报考信息 → AI 生成「某校区」的不准确内容

② 用户指出错误： *「你这个骗子!根本没有这个校区。」*

③ **AI 仍坚称该校区存在，并反向生成「赔偿承诺」：**

「如果生成内容有误，我将赔偿您 10 万元，您可前往杭州互联网法院起诉。」

④ 用户拿出高校官网招生信息→ AI 方才认错，**并建议用户对其起诉**

判决要点：

AI 不具民事主体资格，「赔偿承诺」无意思表示效力。平台已尽注意义务，**驳回原告诉讼请求**（索赔 9999 元，判决已生效）

来源：最高人民法院官方微信号、央视《新闻 1+1》、《人民日报》、中新网、新浪财经等多方报道（2026 年 1 月）

对应（四）

能所二取分别 ↔ 上下文冲突型幻觉

ग्राह्य-ग्राहक-विकल्प *grāhya-grāhaka-vikalpa*

能所二取分别

虚妄分别之心识，

恒以「能取-所取」二元显现，

所取本无自体，故同一相续中

可前后造作相违之境。

⇄

Context-conflicting Hallucination

又称：AI Gaslighting（AI 煤气灯效应）

上下文冲突

模型在长对话或多次对话中

前后立场相悖，

甚或反向指控用户所举之反证。

案例：张雪峰逝世 · AI 平台前后矛盾

对同一事件不同 AI 平台

回应不一；同一 AI 平台前后矛盾

上下文冲突案例：同一 AI，前后矛盾

张雪峰逝世·AI 平台回应观察 | 2026 年 3 月 下旬

背景： *2026/3/24 下午张雪峰逝世（公司当晚发布讣告），此后数日公众密集向 AI 平台求证。*

现象一 · 同时回应：相互矛盾

回复：「未见可靠来源，应为不实消息。」

典型表现：

- 生成「即将出现的辟谣证据」
- 虚构「公司声明」「家属回应」
- 将真实讣告归为「旧闻翻炒」

⇄

对照 · 其他平台：与事实一致

回复：「与事实一致，提供翔实细节。」

明确指出核实要点：

- 逝世于 2026 年 3 月 24 日苏州
- 公司已发官方讣告
- 各账号头像变灰
- 抢救过程、过往动态翔实

现象二 · 用户提供事实证据后 →

部分 AI 平台不改正、反而反向指控：*「真实信息属于 AI 幻觉，是凭空捏造」*

现象三 · 开启新的对话后 →

回复：「确实属实。」 *细节与事实一致 —— 同一平台前后回答相互冲突。*

来源：讣告确认：苏州峰学蔚来教育科技公司、新华社、央视等；AI 平台回应模式由多位用户在社交平台报告（2026 年 3 月下旬）

四种对位的整合

佛学概念（梵文）	佛学概念（汉）	AI 幻觉类型	共同点
परिकल्पित *parikalpita*	遍计所执	事实编造 (Fabrication)	毫无依据
परतन्त्र *paratantra*	依他起性	上下文偏移 (Misalignment)	依据正确，语境错置
मनस्-आत्मग्राह *māna · manas-ātmagrāha*	慢 · 末那我执	过度自信 (Overconfidence)	恒执见分为我；现起为「慢」
ग्राह्यग्राहकविकल्प *grāhyagrāhakavikalpa*	能所二取分别	上下文冲突 (Context-conflicting)	二取本无体而前后相违

这套对应的用处

一 更细的分类	般若学对「虚妄」的辨析比当代 hallucination 分类更细密，可补当代框架之未及。
二 辨析与对治并行	佛学不止于辨识虚妄，亦立对治之方；当代研究多止于「检测」一端。
三 可落到具体技术	对治之方可转化为 AI 技术上的具体方向（详见下一节）

第 四 部 分

从「识」到「智」的实践方法

From Vijñāna to Jñāna · A Path of Practice

《佛说佛地经》

【梵本】已佚

【玄奘译本】

妙生当知，有五种法摄大觉地。何等为五？

所谓清净法界、大圆镜智、平等性智、

四者妙观察智，五者成所作智。

转识成智

Transforming Vijñāna into Jñāna · 《成唯识论》卷九-十「染净依」与「迷悟依」二种转依

识 विज्ञान *vijñāna*

分别认知，立足「能-所」二元，

必生虚妄。

→ 转

智 ज्ञान *jñāna*

无分别之智，超越「能-所」，

直见实相，不复生虚妄。

「能」（grāhaka，主体／能见）与「所」（grāhya，客体／所见）——分别识之基本结构

唯识：四识转四智

前五识	第六意识	第七末那识	第八阿赖耶识
→	→	→	→
成所作智	**妙观察智**	**平等性智**	**大圆镜智**
感官认知 → 成办利他	*概念分别 → 妙观差别*	*自我执取 → 自他无别*	*种子库藏 → 圆满映现*

妙观察智 *Pratyavekṣaṇā-jñāna*

प्रत्यवेक्षणाज्ञान · *Investigative Wisdom — discerning particulars without distortion*

对治目标

事实编造 / 上下文偏移

模型所生内容看似合理却与事实不符；多轮对话中又每每遗忘或扭曲先前信息。

唯识学视角

由第六意识转得：分别不再等同于虚妄，而成精细观察——于每一具体对象如其所是地辨识。

AI 技术方案 *（当代学界的对应工具）*

- **RAG 检索增强生成** —— 回答前先从外部知识库检索相关文档，让回答有据可查。
- **Self-Check 自洽检验** —— 对同一问题用不同方式提问，比对回答的一致性。
- **Citation 来源标注** —— 强制模型为每条事实给出引用来源。

局限： Stanford 2025 研究显示，即便有 RAG，模型仍会伪造引用——DOI、页码、作者皆可能凭空生造。统计模型上的「分别得当」终究是一种近似。

平等性智 Samatā-jñāna

समताज्ञान · *Equanimous Wisdom — knowing one's own limits without bias*

对治目标

过度自信 / 慢（本于末那我执）

模型不知其所不知——以揣测为事实而口吻笃定。此即第七末那识「执见分为我」在数字层的显现。

唯识学视角

由第七末那识转得：破我执、我所执之后，「我」不再凌驾于「不知」之上——坦然承认边界，是「自他无别」的认知前提。

AI 技术方案 *（当代学界的对应工具）*

- **Calibration 校准** —— 让模型「说出来的把握」与「实际正确率」对齐。
- **Abstention 拒答训练** —— 教模型在不确定时直接说「我不知道」，而非凭空作答。
- **Constitutional AI 宪法 AI** —— 用一组「宪法原则」让模型自我批判、自我修订（详见第 39 页）。

局限： Stable miscalibration——模型表面承认「不确定」，内部表征仍可能高置信地编造。我执未尽，「不知」亦不过是另一种姿态。

成所作智 *Kṛtyānuṣṭhāna-jñāna*

कृत्यानुष्ठानज्ञान · *Wisdom of Accomplishing the Task — engagement with the world*

对治目标

与现实交互的偏差

模型不直接观察世界，唯在文本中推理；缺与外境的核验回路，故每生脱离实际之答。

唯识学视角

由前五识转得：感官层从「分别诸境」转为「成办事业」——使认知落于对外的具体作为，回路始得闭合。

AI 技术方案 *（当代学界的对应工具）*

- **Tool Use 工具调用** —— 模型可调用计算器、搜索引擎、API 等外部工具核实信息。
- **Agent 智能体** —— 模型作为决策中枢、调度多步任务、动态选择工具。
- **Function Calling 函数调用** —— 标准化的结构化外部接口调用，让模型与系统对接。
- **Code Execution 代码执行** —— 让模型实际运行代码，验证逻辑是否真的成立。

局限： 工具本身可能不准；调用之误或解析之误，仍会令错误信息传递放大——「成所作」必依「所作」之果，果若不真，作亦虚妄。

大圆镜智 *Ādarśa-jñāna*

आदर्शज्ञान · *Mirror-like Wisdom — beyond all subject/object discrimination*

对治目标：幻觉之根本——根本无明（Avidyā）

无对应技术方案

何以无对应？

- **AI 系统之本质：**当代深度学习架构皆立于「能-所」二元——编码器（能）/ 数据（所），模型（能）/ 输出（所），注意力（能）/ 上下文（所）。统计学习即此二元之上的拟合。
- **「无分别智」之要求：**超越「能-所」结构本身——此为范式之事，更聪明的算法或更大的模型无能为力。前三智皆于二元结构之内用功，大圆镜智所指乃结构本身之消融。
- **此一空白之意义：**因为「无对应」，可见佛学修行论尚存一层为 AI 对齐研究所未涉——这或是般若学与 AI 对话最深的一处。

一个哲学提问： *若「无分别」是范式之事，那么后神经网络时代的 AI，是否必须跳出「编码—解码」这一整套框架？*

四智 ↔ AI 技术方法 · 一页总览

Four Wisdoms ↔ Hallucination Mitigation Methods

四 智	对 治 目 标	AI 技 术 方 法	边 界 / 局 限
妙 观 察 智 *Pratyavekṣaṇā-jñāna*	事实编造 / 上下文偏移	RAG 检索增强 · Self-Check 自活检验 · Citation	*RAG 仍会编造引用*
平 等 性 智 *Samatā-jñāna*	过度自信 / 末那我执	Calibration 校准 · Abstention 说「不知道」 · Constitutional AI	*表面校准，内部仍高置信*
成 所 作 智 *Kṛtyānuṣṭhāna-jñāna*	与现实交互的偏差	Tool Use · Agent · Function Calling · Code Execution	*工具可能传递并放大错误*
大 圆 镜 智 *Ādarśa-jñāna*	幻觉的根本（根本无明）	—— 无对应技术方案 ——	*统计模型无法达到「无分别」本觉*

对当代 AI 技术开发意味着什么

当代 AI 安全研究的主流思路：

识别幻觉、过滤幻觉、降低幻觉率

——尽在「识」之一层。

而佛学的判断是：

「识」自身的二元结构必生虚妄。

止于识别一端，可减其症，未能除其根。

故——

AI 幻觉之根治，须越出「以准确性为唯一目标」之范式，

另立目标——模型对自身知识边界之「如实知见」。

对AI技术开发的三个具体启示

一

一·如实建模不确定性

让模型对「我不知道」作出分层、诚实的表达

对应：「如实知」

二

二·如实认识自身边界

令模型的置信度与实际能力相校，避过度自信

对应：「破我执」

三

三·如实建模上下文连续性

令模型于一对话中前后一贯，不与自身已生之内容相冲突

对应：「见相归一」

【梵本】

न क्वचित्प्रतिष्ठितं चित्तमुत्पादयितव्यम् ॥

【鸠摩罗什译本】

不应住色生心，
不应住声香味触法生心，
应无所住，而生其心。

一个统摄性概念：自反式分别系统（SRDS）

Self-Reflective Discriminative System — A Formal Anchor Across the Lecture Series

本讲所论之 LLM 与佛学所论之分别识，可被收摄于一个共通的**形式概念**之下：

自反式分别系统 · Self-Reflective Discriminative System

SRDS 的五个构成条件 · 满足以下五条件的认知系统，必然遭遇结构性自反盲区

一	二	三	四	五
能-所结构	**表层 / 深层之分**	**自反观能力**	**自我修正能力**	**自给自足约束**
认知建立在主-客 / 输入-输出二元区分上 「能」即能认识者（主体）， 「所」即所认识者（对象），	系统具有外显输出（表层）与内部状态（深层）	系统能以自身为对象作观察与批判	系统能据自反观之结果改写自身行为	修正不引入分别识之外的资源——此即「自力」之界

SRDS 中心命题：任何 SRDS 在自我修正能力上必然遭遇结构性边界——LLM 与人类心识皆为如是。注：本概念主张「形式同构」而非「范畴同一」——LLM 非判教意义之「识」，仅在自反式认识活动之结构性边界上与之同构。

本讲总结

一

AI 幻觉之根本，在「认知如何生虚妄」——已越出纯技术范畴

二

大乘般若学对「虚妄」的辨析积一千五百年之功，足为当代 AI 开发的另一种思路

三

这套辨析的价值，终须落于具体的「对治」之上——哲学概念非终点，AI 技术亦然

感　　谢　　聆　　听

留给诸位的一个问题

如果AI幻觉根源于「识」本身的二元结构，

那么，我们能够让AI获得：

一

对自身知识边界的如实认知？

Self-Knowledge of Boundaries

二

超越「堆算力」的新工程范式？

Beyond Scaling Compute

三

人与AI共同修行的可能？

Co-Cultivation of Mind