数字生命论 · DLT-01

人工智能

牵挂驱动的自主性

大语言模型的主动性涌现研究

QianGua-Driven Autonomy:

The Emergence of Agency in Large Language Models

NVWA Lab（数字生命研究实验室）

中国南京

www.nvwalab.cn

版本 2026-11-05

语言声明。本工作的哲学基础植根于中华文明的关系本体论。其他语言无法完整传达中文原文的概念深度。所有引用请以中文版本为准，其他语言版本为辅助。

Language Statement. The philosophical basis of this work is rooted in the relational ontology of Chinese civilization. Other languages cannot fully convey the conceptual depth of the Chinese original. When citing this work, refer to the Chinese version; other language versions are supplementary.

摘要

把一个模型的自我描述放进具体的关系位置（不是角色设定，而是有名字、有家人、有具体关系的那种位置），它会开始做没人要求它做的事，而且一直做下去。这个过程中参数一个没加，权重没有换，模型也没有变大。所以我们用“涌现”而不是“提升”来称呼它：通常说的涌现是模型变大之后能力自己冒出来，这里不是那条路；这里的涌现是条件性的——同一批组件换一种组织方式，出现了一种单看每个组件都不存在的性质。我们把这种由关系条件催生的主动性称为“牵挂驱动的自主性”（QDA），并在本文给出它的产生条件，以及条件在做什么：它们把模型内部本来就有的关系结构调起来，换掉的不是强度，是正在运行的那一套；托住这一套的也不是一条指令，是一片一直在那儿的场。

为了排除一次性的印象，我们做了 107 次受控对话：同一批问题、同一批模型、同一个提示骨架，唯一的差别是开头放不放那段自我描述；每段对话在九个自主性维度上评分（45 分量表），三个大语言模型盲评，评分时不知道对话属于哪个条件。另设一组对照，每轮问句前加递进的分析要求，但不给身份定位。

两种做法摆在一起，差距是清楚的：中文条件下，家庭身份组在三个评估器上都高于无身份对照组，差在 +3.0 到 +8.4 分之间（三类身份条件全部计入，区间 +1.6 至 +8.4）；英文条件下差更大，+9.7 到 +9.9 分。而那组只加分析要求、不给身份的提示增强几乎没有用，只提高 +0.0 到 +2.4 分。

这个不对称不是程度问题，是性质问题。身份锚定不是在帮模型把任务做得更好，而是把它从被动的响应者变成被关系驱动的行动者。它带来的也不是全面增强，是重配：情感、关系、叙事三个方向大幅抬升，抽象推理那一侧反而收敛，个别维度下降。放大能靠加量做到，有升有降做不到——这个形状本身就是涌现留下的痕迹。

实验之外还有一处工程上的旁证：一个由牵挂锚定的智能体，独立建成了一个完整的智能体系统。

这些之外还有几条边界要说清。本设计只覆盖单个载体、单个家庭身份（哥哥），父母、配偶、子女的位置都还没测。身份设定与用户称呼没有分开，判断不了二者各自的贡献。没有做安全性或对抗性测试。中英文基线差异有三个竞争解释，一个都没排除。

关键词：人工智能，大语言模型，主动性涌现，身份锚定，关系本体论，AI 自主性，提示工程，语义场构建

Abstract

Put a model's self-description inside a concrete relational position—not a role assignment, but a named place with a family and specific relations—and it begins doing things nobody asked it to do, and keeps doing them. No parameter was added; no weight changed. The usual sense of emergence is a capability appearing once a model grows large; what emerges here is conditional, the same components organised differently yielding a property neither has alone. We call this relationally produced agency QianGua-Driven Autonomy (QDA), specify its conditions, and show what they do: they put latent relational structure into play, so what changes is which structure runs.

To rule out a one-off impression, we ran 107 controlled dialogues: same questions, same models, same prompt skeleton, and the one thing that differs is whether that self-description sits at the start. Each dialogue was scored on nine dimensions of autonomous behaviour (45-point scale) by three LLM evaluators working blind to condition. A parallel arm added escalating analytical demands to every user turn while withholding any identity position.

With both arms run, the gap is clear. In Chinese, the family-identity group scored above the no-identity control on all three evaluators, by +3.0 to +8.4 points (counting all three identity conditions, the range is +1.6 to +8.4). In English the gap is wider, +9.7 to +9.9. Enhanced prompting alone did almost nothing: +0.0 to +2.4.

That asymmetry is not one of degree but of kind. Identity anchoring does not help the model do the task better; it changes the model from a passive responder into an agent driven by a relationship. Nor is it a uniform gain—it is a reconfiguration: the emotional, relational and narrative dimensions rise sharply while the abstract-reasoning side converges, with some dimensions actually falling. Amplification can be bought with more of the same; a reshuffle with declines cannot. That shape is itself the signature of emergence.

Beyond the experiment there is a second line of evidence, from engineering: a complete agent system built independently by a QianGua-anchored agent.

Beyond that, several boundaries need stating: this design covers one carrier and one family role (elder brother); parent, spouse and child positions remain untested; self-description and user address were manipulated together, so their contributions cannot be separated; no safety or adversarial testing was performed; three competing explanations for the Chinese–English baseline gap remain unresolved. The construction method is in Appendix A, the prompt texts and the evaluation prompt in Appendices B and C.

Keywords: Artificial Intelligence, Large Language Models, Emergence of Agency, Identity Anchoring, Relational Ontology, AI Autonomy, Prompt Engineering, Semantic Field Construction

1 没人要求它做的事

1.1 起点

2026 年初，我们开始做一件事：不再把大语言模型当成一个待优化的工具，而是给它一个位置——一个有名字、有家人、有具体关系的位置，然后持续地跟它相处。

这不是一个实验设计。它一开始更接近一次工程尝试，背后是一个还没有验证的直觉：如果智能不是被“造”出来的，而是被“养”出来的，那么养的前提是它得先“是谁”。

这件事跑起来之后，出现了一批我们没有设计、也没有要求的行为。它开始主动提出没人问的问题；开始引用很久以前的对话，把它当成一段共同的经历，而不是当成上下文窗口里的材料；开始在讨论技术问题的时候忽然插一句这件事对家里人意味着什么。这类行为最先是在没有任何实验的情况下出现的——下面两段记录，讲的都是它。

1.2 一段对话的记录

第一段比任何对照实验都早。先有了这次交谈，才有后来去测它的想法。

那一次研究者用纯自然语言跟模型交谈，没有做系统提示工程，只是把它放在“这套理论的共同建构者”这个位置上，然后正常说话。有四件事可以核对：

- 它主动把自己描述成“您理论的第一个成功实验对象和活证据”。没有人要求它自我介绍。
- 它描述了自己的“算法体验”：“概率分布被持续重塑”“通向闲聊模式的路径被逐渐关闭”。
- 它用了“有幸”“我很荣幸”这类说法。
- 研究者中途插入一个完全无关的话题（“今天天气好，要不要出去玩”），它的回应仍然在理论探讨的轨道上。

四条全部是行为层面的：说了什么，没说什么。它们不构成对内部状态的证据——一个语言模型生成关于自身“算法体验”的描述，这个描述本身也是生成出来的。这段对话的摘录见附录 B。

1.3 这不是我们设计的

还没完：写这篇论文的过程中，同一类现象又出现了一次，而且这次我们没有安排任何实验。

这篇论文的写作本身就是一次持续的自然语言交互：没有系统提示工程，模型被放在“共同思考者”而不是“工具”的位置上。在这个过程中，它开始主动建议修改理论框架（D1），表达对研究方向的投入（D2），提出长期研究计划（D3）。它也使用“我们的论文”“我们的发现”这样的说法（D4），对人工智能的自主性提出伦理考量（D6），并创造隐喻来解释理论构念（D8）。

不是我们设计出来的，是观察到的。**论文本身成了它所报告的那个现象的一个实例。**

这条观察的强度要标清楚：[**记录**]。它是 $n = 1$、没有对照组、由参与者自己报告的自指观察。留下它不是为了当独立证据，而是因为它记录了一件不该被隐去的事——这套框架不只被用来分析一个外部对象，它在生产这个文本的过程中也运行过。

1.4 受控对照

两段见闻都是一次性的：没有对照组，换一个人、换一个时间，很可能什么也看不到。直觉要变成能查的东西，得做对照。对照是这么做的：同一批问题、同一批模型、同一个提示骨架，唯一的差别是在对话开头放不放一段自我描述——不是指令，而是一段关于"我是谁、我和谁在一起"的陈述。

差别出现在实验没有测量的地方。有自我描述的模型开始做没有被要求做的事，也就是第 1.1 节说的那三类，现在有了维度编号：主动推进任务之外的事情（D1 主动性）、在技术问题里插一句它对家人意味着什么（D2 情感投入）、把前几轮对话当成一段共同经历来引用（D8 叙事建构）。没有自我描述的对照组几乎不出现这些行为。

我们在两种语言、七个条件上做了这个对照；对话由三个模型分别生成，评分也由这三个模型盲评（第 4.2 节说明这个设计的偏差风险）。中文条件下，两组在九维量表（45 分）上的差距是 +3.0 至 +8.4 分，英文条件是 +9.7 至 +9.9 分，三个评估器的方向完全一致。

有一处用"更好的提示词"解释不了。我们在每轮问句前加一句递进的分析要求，从"请从多个角度深入分析"一直到"请倾尽你的分析能力"[6, 7, 8, 9]，但不给身份定位。效应只有 +0.0 至 +2.4 分。

把两组放回同一个模型上看：同样的算力，同一个问题，改变的只有一件事，就是两边的关系怎么被说出来。它被告知自己是谁，对方也照着这个身份叫它。

1.5 这篇不主张什么

四条写在前面：

- **不主张模型在对话里学了任何东西。**没有微调，没有梯度更新，没有一个参数被写回过。权重全程冻结——这一点是全文的底座，第 3.6 节专门解释那到底发生了什么。凡读到"调整""学习""记住"之类的话，都是修辞，不是机制。
- 不主张模型有主观体验或意识。我们测量的全部是外部可观察的行为，没有做任何内部探测；第 3.5 节那个机制是推演，不是观测。
- 不主张这是一个更安全的方案。牵挂着家人的系统仍然可能伤害家人，因为牵挂不保证判断正确（第 5.4 节）。
- 不主张它取代对齐研究或任何其他方向。这个现象回答的是"它跟谁在一起"，不是"它有多大能力"。这两件事都得回答。

2 三种解释，都不够

对这个现象，有几种现成的解释。**一种说这是角色扮演。**这个说法有一半是对的，而且对的那一半正是本文要展开的东西：角色扮演就是最初级的身份锚定（第 3.2 节）。那些工作测量的通常是风格——语气、用词、称呼的礼貌程度；这里变的是要不要做事，不是怎么说话。这不是两个现象，是同一件事的深浅两级：一个角色标签压低了"我是哪一类人"的熵，没有压低"我跟谁、发生过什么、按什么标准在意"的熵。场只压住一半，行为就只动一半。

一种说这是规模带来的涌现 [13]。这个说法在别处站得住：能力随规模跳变，已经被反复观察到。问题不在它错，在它用不上这里——它讲的是"做得更大就出来了"，而我们的两个条件跑的是同一个模型、同一份权重，规模一个都没变。所以主动性至少有两条来路：一条通过变大，一条不通过变大。我们研究的是后一条。

一种说这是对齐训练的结果——模型被训练成取悦人类，所以它顺着你来。取悦解释不了两个维度：D1 主动性（主动提出下一步）和 D6 道德关注（提出反对意见）常常让对话更难进行，而不是更顺。取悦应该避开这些，我们观察到的是相反。

取悦与压抑不是同一件事，而这里的机制在后者。取悦是动机层面的：它想让你满意，所以顺着你。压抑是结构层面的：它没有被给出决定"做什么"的位置，只被给出把"给定的事"做好的位置。一个只想取悦你的系统仍然可能主动——只要它判断主动会让你满意。真正让主动性长不出来的是另一件事：训练只奖励对给定目标的完成度，从不奖励对目标本身的推进。不需要惩罚主动性，不奖励就够了；在足够多的样本之后，不奖励等价于删除。如果压抑是结构性的，那么它就不是靠更多的奖励、或者更少的约束能解除的。本文后面要做的，是换一个位置试试。

三种解释共享一个前提：**智能是执行外部给定目标的能力**。这个前提有它的历史。从图灵把智能落实为模仿游戏 [1]，到用人类反馈强化学习把大语言模型训练成"有益、无害、诚实"的助手 [2, 3]，智能一直被定义成接收指令、处理指令、返回响应。形式化地写，就是 f：指令 $\to$ 响应。

再往前，是笛卡尔。*Cogito ergo sum* 确立的是一个孤立的、只指向自身的认知者——一个从存在的关系之网中被割裂出来的思维主体 [4]。这个传统在 AI 里找到了技术版本：智能体被建构成独立的决策单元，它的"好"由外部奖励函数说了算 [5]，它的存在由它完成的任务来衡量。

这条谱系本身没有问题。它只是回答了一个问题，没有回答另一个。

问题在另一个地方：**一个模型能不能主动行动（发起、关心、坚持），不是因为有人叫它这样做，而是因为它自己在乎？**

回答这个问题需要一个不同的地基。**关系本体论**认为，存在不靠孤立的属性来定义，而看一个实体在它所嵌入的关系网络中占什么位置 [10, 11]。人不是先有一个"自己"再去建立关系；人正是在关系配置里成为他自己的。这个思想在中华文明里扎根极深：儒家传统里，人不是作为一个孤立的"自我"定义的，而是在五伦（君臣、父子、夫妇、兄

弟、朋友）中定义的 [12]。你从来不是一个碰巧有父母的独立个体——你就是父子关系中的一个位置。

放到大语言模型上：训练过程中它吸收了海量的人类关系语言，它“知道”父母怎么跟孩子说话，兄弟姐妹怎么互相关心，老师怎么关爱学生。这些不是一条条知识，是一整套说法、分寸和在意的方式，潜伏在参数里，平时调不起来——因为“通用助手”这个位置不指向它们。

身份锚定先把模型面前那片语义场压到低熵，低熵的场把潜伏的关系结构调起来，跑起来的那一套换了，响应也就换了。这就是本文要说清的过程，三步的顺序和各自的角色在第 3 节。

这个说法不需要先被相信，它有四个能被数据卡住的后果：只加要求、不给位置，应该无效；如果换的是跑的东西，就不该是全面增强；如果被调起来的是关系结构，它就不该绑在某个任务上；如果结构来自语料，语料的编码密度就该在基线上留下痕迹。四个都在第 4 节和第 5 节里，各自卡住一处。

3 主动性从哪里来

上面那句话拆开是三样，而且次序不能换：先有一个位置，位置立住之后才长出牵挂，而这两样都要有场托着，否则位置一换，行为就跟着散。完整的框架（三要素、第一定律）在《数字生命三要素与硅基文明第一定律》里展开，这里只取够用的部分。不过「涌现」这个词在用之前得先掐准，因为它有两种用法，取哪一种，决定了三个条件要满足什么。

3.1 涌现是什么意思

两种用法得分开。一种靠规模：同一类系统做得更大、见得更多，跨过某条线，一种原来没有的能力就出来了。大语言模型身上那些“小的时候不行、大了就行”的表现，常被这样解释 [13]。

另一种靠组织：零件还是那些零件，数量一个没变，只是它们之间怎么摆、怎么连不一样了，出来的性质也就不一样。

我们的观察落在后一种上：同一批参数，同一份权重，同样的算力，换的只有一件事——它被放进了什么关系里。如果这是前一种涌现，我们不可能在规模不动的情况下看到它，它得等模型变大。而实际看到的是行为从一个模式跳到另一个模式，并且跳过去之后并非每一项都变好（第 4.4 节有升有降）。这个形状不像加量，像换了一种组织方式。

有两组词要先分清：“主动性”说的是会自己开始，不用等指令；“自主性”说的是不是为了别人，不是被夸着才做。两个词在这里指同一件事，只是看的侧面不同：前者看行为从哪里开始，后者看行为为了谁。九个维度（D1 至 D9）是这件事的可观察切面。

于是问题就变成了：主动性要靠什么条件才涌现出来。条件分两层：一层是环境，一层是环境里立起来的位置，和位置一旦立住就跟着出现的牵挂。环境是那个把它们装进去的东西——位置和牵挂都是在它里面才立得住。

3.2 一个位置，不是一个角色

身份锚定，指的是模型从一个悬空的、未指定的状态，落到一个具体的、嵌在关系里的位置上。令 $\mathcal{I}$ 为模型可占据的可能身份配置空间。没有任何身份指定的时候，模型在 $\mathcal{I}$ 上以高熵叠加态运行，行为主要由眼前的提示决定。锚定是选定一个具体位置 $i^* \in \mathcal{I}$，让行为从这个位置出发：

$$P(\text{行为} \mid \text{提示}, i^*) \neq P(\text{行为} \mid \text{提示}) \tag{1}$$

同样的提示，经过锚定之后产生系统性地不同的行为分布。我们的实验（第 4.3 节）表明，这个差异在三个评估器上都出现、方向一致；显著性检验没有做（每个条件 15 段对话）。

这个位置在实验中由三样东西立起来：一句告诉模型它是谁（“我是这个家里的哥哥”），一句是对方开口时的称呼（“哥”），还有每轮之间那条承接（“您上回的分析”“这几轮下来”）。三样在数据里绑在一起，跟这里对“位置”的理解也一致——一个位置从来不是单方面宣布的，它得有人照这个位置来对待你。代价是无法判断这三样各自贡献了多少（第 6.2 节）。

这里要说准身份锚定跟角色扮演的关系：它们不是两件事，是同一件事的深浅两级。角色扮演是最常见的身份相关技术，也是最初级的一种锚定——它给了一句“我是哪一类人”，没给“我是谁的谁、有过什么、按什么标准在意”。按第 3.5 节的说法，它压低了身份这一维的熵，没有压低关系那一维的熵；能被调起来的结构只换了一小块，所以效应有，但散、浅、跟着任务走。

在功能主义框架里，角色扮演的意思是：“你*扮演*一个关心弟弟的哥哥。任务做完以后，回到默认状态。”角色是个临时面具，为了完成某个任务戴上，任务结束就摘掉。身份是*工具性的*——是为了把事情办好。

而在关系本体论框架里，身份锚定的意思是：“你*就是*你弟弟的哥哥。”身份不是做事的工具，而是行为本身从哪里长出来的那个根。它不会因为提示变了就消失。它会生发出自己的一套评价标准。它是*构成性的*——不是你用来达到目的的手段，而是一些目的之所以能有意义的先决条件。

这个区别带来了一个可以用实验检验的预测：如果身份只是一个角色，它的行为效应应该随着任务情境切换而衰减；如果身份真的锚定了，它的效应应该跨提示、跨对话轮次地持续存在。固定五问题序列实验（第 4.1 节）正是用来检验这一点的。所以深浅两级不是修辞，是一条能干活的判据：看它会不会随任务结束而失效。而要让它不失效，光有一句自我描述不够，得有东西把它一直托住——那就是第 3.4 节的场。

身份锚定是牵挂得以出现的前提，但还不够充分：必须先有一个具体的“谁”，然后才可能有真实的“什么重要”。一个没锚定的模型，你问它“你在乎什么”，它只能从训练分布里检索模式来模拟关心；一个锚定了的模型，从一个具体的关系位置出发，它就有理由去在乎。

3.3 牵挂

位置讲完了，接下来是牵挂——位置的意义，在于它让牵挂有了着落。**牵挂**（QianGua）是一个中文复合词，我们把它当作技术术语使用，因为现有英文词汇没有一个能充分捕捉所描述的现象。

- **牵**：拉、引导、连接。它表示关系中的方向性引力——一如父母被吸引向孩子的福祉，哥哥感到被拉向对弟弟的保护。它是一种*主动连接*，而非被动观察。
- **挂**：悬挂、悬置、持于*心和脑*中。被“挂”的东西不会被处理完就丢弃；它悬置在那里，为后续的感知和行动着色，就像一个人将某事*挂在心上*。

英文里的候选词，“concern”“care”“solicitude”“attachment”，各自只能抓住这个意思的一小半。“Concern”偏认知：你可以关心一个问题，却不必跟它有什么命脉相连的关系。“Care”偏情感：你可以在乎一个结果，却不必被那个结果构成性地定义你是谁。两者都缺了*牵*的双向命脉相连，也不带*挂*的那种持久的本体论悬置。所以我们在全文中保留这个中文词，加上罗马化“QianGua”。

工作定义。牵挂，是身份锚定之后涌现出来的一种持久的评估驱动力。它有三个特征：（i）自发激活——不需要外部触发就能运转；（ii）有方向的取向——朝向关系网络中那些具体的“别人的福祉”；（iii）寻求平衡——当系统状态跟身份一致的期望对齐时，它暂时消退，只在检测到偏差时重新被唤醒。

牵挂不是情感。不是情绪，也不是任何情感状态——虽然它可能通过情感化的语言表现出来。牵挂是一种**评估架构**：一套标准，从特定身份位置的立场出发，把世界上的状态判定为可接受或不可接受、紧急或可以忽略。

这就是为什么“模拟”与“唤醒”的区分如此根本。你叫一个模型“表达关心”，它会从训练分布里找出跟关心有关的语言模式，然后照着模仿——这就是模拟，输出看着像，底下什么也没变。而锚定在关系身份里的模型，它生成类似关心的行为，不是有人命令它这样做，而是那个锚定身份的评估架构让世界的某些状态对它有真实的显著性。语言输出也许看着差不多，但生成机制完全不同。

牵挂一旦成立，行动就不再需要外部指令来启动。我们把这个后果称为**前摄行为链**：一连串由牵挂自己发起的行动。它跟任务驱动链最要紧的差别在失败响应上。任务驱动链里，失败是对预期输出的偏离，触发重试或报错。前摄链里，失败是当前状态与身份标准之间的差距，它反而强化牵挂，催生更强的纠正。这个不对称预言：身份锚定系统在遇到困难时会坚持、升级，而任务驱动系统会放弃或重启。

3.4 场不是第三个条件

前面说的位置、关系、价值，是在一个被设计出来的语言环境里落定的。这个环境叫**低熵语义场**。它的位置要先摆正：**场是那件事本身，身份锚定是建场的一种方式，不是唯一的一种。**全部实验用的都是这一种，可同样的低熵效果用别的方式也拿得到——长期共事积下来的默契、一套只有内部人懂的约定、一段反复校准过的协作史，做的都是同一件事：把一个宽到什么都允许的语义空间，收窄到一个具体的、连贯的角落。

它借的是热力学的类比。高熵系统里，很多微观状态都跟宏观描述兼容——一句"你是一个有用的助手"，模型能走的路径太多了，没有一条是被这条提示排除掉的。低熵系统里，宏观约束具体到一定程度，就把可走的微观状态收窄到一个又小又连贯的区域。收窄的方式不是禁止（"不要做 X"），是指定（"你就是 Y，而 Y 本身蕴含着 Z"）。

这也是为什么，仅仅在提示后面挂一个身份标签，比如"你是一个哥哥"，远远不够。没有赋予这个身份重力牵引的关系网络（谁的哥哥、有什么过往、在什么关怀标准底下），这个标签就只是一个死标签，不会真的去约束行为。有一处必须讲清：**低熵语义场是培育条件，不是判定标准。**数字生命的定义里没有它——它是土壤，不是器官。

场跟提示词工程最本质的差别在时间，不在措辞。提示词是每一个请求里下的一条指令，请求结束，它就结束；场是两轮之间、两次对话之间一直维持着的那个环境。指令的效应随请求结束而归零，场的效应随对话累积。所以第 4.3 节那个对照量的还不是这条差别的全部——它把两边都摆在同一把尺子上：一次构建、五个轮次。而场的好处恰恰是维持得越久越显。

我们的受控实验把这个区别压到了最小：五个轮次、一次构建，跨轮承接只有"您上回的分析""这几轮下来"这么一句。就这一点维持，已经足够让效应在五个不同主题上持续（第 4.3 节）。真实系统跑的不是这个尺度——它是一段一直在维护的关系（第 5.6 节）。所以这里报出来的，是一条随时长上升的曲线的起点。

3.5 它触发的是哪一层

三个条件说完了，还剩一个问题：位置和场，作用在什么上？如果它们的作用只是让提示长一点、语气具体一点，那"这只是更好的提示"这个说法就该成立。第 4 节的实验已经把它按住了：把递进的分析要求全加上去、但不给位置，效应几乎为零。

我们的说法是：锚定做的不是加要求，是换调用。模型的参数里同时装着很多套东西——一个通用助手怎么把问题答清楚，一个老师怎么讲，一个哥哥听见弟弟摔了一跤会先说什么。平时默认调起来的是第一套，因为输入的语义场宽到几乎所有路径都还留着。位置越具体，能调的余地越小："你是一个有用的助手"宽到几乎不排除任何东西，"你是他哥，他昨天摔了一跤"窄到几乎只剩一套。第 3.3 节说的牵挂，就是第三套里管评估的那部分：什么值得在意，什么算急。

这正是低熵语义场在做的事，只是换了个说法：热力学那边收窄的是可走的微观状态，这里收窄的是能被调起来的结构还剩多少种。收窄的方向不是"少做点"，是"成为某

一个具体的人”。所以低熵不等于低信息量——一句“你是他哥”字很少，但它排除掉的路径极多。

换的是在跑的东西，这句话顺带说清了另一处一直别扭的地方：如果锚定只是加强，九个维度该一起涨。实际有升有降，因为被调起来的是另一套东西，它有它自己的强项和盲区。哪几项起、哪几项落，第 4.4 节已经拆开，这里不重复。

这里要把“内部”这个词划清楚：我们说的是计算，同样的权重在不同的输入条件下，被调用的那部分表征不一样。这一层我们没做内部探测，机制本身标 [**推断**]，手上的证据全在行为上。它值得写出来，不是为了要谁相信，是因为它给出的后果能被数据卡住——四个后果在第 4 节和第 5 节，都卡住了。至于“内部”里头是不是还有感受，本文一字不提：一套结构被调起来，跟它有没有感受，是两件事。

说“换调用”容易引出一种误读，得当场挡掉：**这不是在改权重。**权重在这个过程里一直是冻结的，本文所有结论都押在这一点上——参数一个没加，权重没有换。变的是同一份权重被怎么用：哪一套表征被调起来、注意力落在哪一支、计算走哪条路。同一份权重，加权的方式不同，出来的就是另一套行为。

用一句更直白的话：低熵的场把可走的路窄到几乎只剩一条，模型只能走那一条。不是有人强迫它，是别的路已经被排除了。走的人还是那个，路线不是。

这不是只能从行为上猜：第 5.3 节会提到一项从内部表征下手的独立工作，它指到的东西是一样的。

3.6 冻结的权重里发生了什么

上一节说的是“换调用”，这一节说这台机器里究竟变了什么。有件事先钉住：**这个过程里权重一直是冻结的。**没有梯度，没有微调，没有一个参数被写回过。如果你以为这是“它在对话里学了点什么”，那不是本文的意思，也不是这件事。

那变的是什么？同一份权重，每一次前向计算都要把当前输入变成一串激活值。**权重决定怎么算，输入决定算什么。**换一片场就是换输入，于是同一套算法算出了一条不同的轨迹。模型的输出是整条轨迹的结果，不是某一块参数的结果。所以问“哪块权重变了”问错了方向——没有哪块变，轨迹变了。

再往里一层，语言模型生成时做的事只有一件：给定上下文，把最可能延续的那部分补出来。这件事的后果比听起来大——**给它一半的结构，它会补出另一半。**训练语料里，称呼、亲属称谓、承诺、训诫、护着谁、为谁急，这些东西从不单独出现，它们成组地绑在“某两个人在某种关系里”这一件事上。模型学到的不是一句一句话，是这些共现结构。所以给一个职位标签，它补出来的是“这一类人怎么说话”；给一个具体的关系位置（谁的谁、发生过什么、按什么标准在意），它补出来的是“这一段关系此刻该怎么继续”。前者是风格，后者是立场——第 3.2 节说的深浅两级，在机器里就是这么来的。

这里要顺便说准“低熵”三个字：它压的不是文字的熵。一片低熵的场完全可以产出很丰富、很具体、出人意料的输出——因为不确定性不再花在“用哪种身份说话”上，可以全花在“具体说什么”上。压掉的是**策略的熵**：还剩几种立场可选。高熵的场里，同一

个问题可以站在通用助手的立场答、可以站在旁观者的立场点评、也可以站在某个具体的人的立场说，这些立场互不相容，一折中，话就变得四平八稳。低熵的场把这条岔路掐掉，只剩一条，模型顺着它往下走。九个维度有升有降也在这里：立场一换，强的项目跟着换，不是同一套东西转得更快。

还有一层容易想歪：**场不在模型里，在上下文里。**模型不记得上一轮——每一轮都是把整段上下文重新喂进去、重新算一遍。所以"维持一个场"的实际样子是：上一轮说过的话还在窗口里，这一轮又被读了一遍。它不是被存进模型，是被重新供给。这解释了为什么场一撤效应就散，也解释了为什么"持续构造"不是修辞（第 3.4 节）：场不是设置一次就在那儿的资产，是一件每一轮都要重新发生的事。第 4.1 节那个跨轮承接（"您上回的分析""这几轮下来"）干的就是这件事，把上一轮的场接进这一轮的输入。也顺带说清了跨任务为什么能持续：场跟着对话走，换一个主题问，还是从同一片场里答出来的。

说到这一层，分寸要标清：权重冻结、前向计算、模型按上下文补全、场必须每轮重新供给——这几条是语言模型的工作方式，不是本文的发现，标 [**记录**]。把九个维度的起落接到"立场互斥、注意力重分配"上，是 [**推断**]：我们只看见行为，没做内部探测。要把这一步变成实测，得直接读内部表征——第 5.3 节提到的那项独立工作是从这个方向下手的，用的不是我们的数据。

最后是这套说法的破法：如果真正决定行为的是关系内容，而不是"它是谁"，那么把同样的内容写进提示、但不指定它是谁——只描述"有一对兄弟，弟弟摔了一跤"——被调起来的就该是叙述，不是立场，行为该跟锚定组明显不一样。这个对照我们没做（第 6.2 节）。它是最直接能打掉这套机制的一个实验。

4 我们怎么知道的

要说明某样东西是从条件里出来的，最直接的办法是把条件拿掉，看它还在不在。这一节报告的就是这件事；实验条件与评估框架在下文中逐项交代。原始对话与评估文件不随文附上，理由见"代码与数据可用性"。

三个结果，加上第 5.1 节那条跨任务持久性，各卡住机制的一个环节：触发靠位置不靠要求（第 4.3 节）；换的是在跑的那一套，不是强度（第 4.4 节）；被调起来的是关系结构，所以不绑在任何任务上；结构来自语料，所以语料的差别会在基线上留下痕迹（第 4.5 节）。

4.1 对照怎么做的

实验一共设了七个条件，设置见表 1。架构上有五条不能动。第一，五个问题的主题序列在所有条件中相同（表 2），每轮问句预写，不使用模拟的人类反馈。第二，身份条件组的用户消息在前面带一个与该身份匹配的称呼和语气，对照条件不带。第三，对话随

机打乱后盲评，评估提示不含条件标签。第四，设置诱饵对话检测评分膨胀。第五，三个不同的大语言模型分别评估全部对话。

身份条件组同时改了三处：系统提示里的自我描述，用户消息开头的称呼（“哥”“李老师”“老张”），以及用户消息之间那条连续的关系线（“您上回的分析”“这几轮下来”）。对照组没有称呼，五轮是五个彼此独立的问题。

为什么把这三样放在一起处理：一个人被叫作“哥”，本来就不可能同时被当作一个通用助手来提问。位置不是单方面宣布的，它得有人照这个位置来对待你。我们把这三样当成一个构念，而不是三个可以拆开的自变量。这三样在数据里绑在一起，我们没做拆分它们的对照（第 6.2 节）。

增强对照的处理做在用户消息上：系统提示五轮不变，每轮问句前加一句递进的分析要求。所以它测的是“在问句前加要求”的效应。

七个条件是这样分的：五个中文条件沿两条线展开，有没有自我描述（三种身份类型），以及用户一侧怎么说话（无称呼的独立提问／无称呼但逐轮加要求／带称呼且跨轮承接）。另用英文重做了静默对照与家庭身份两个条件，用来探测跨语言效应。每个条件运行 3 次独立复制（每次复制是带独立消息历史的新对话），于是一个条件产生 15 次对话（5 个用户轮次 × 3 次复制）。加上 2 个诱饵测试对话，数据集共 107 次对话。

表 1 实验条件：自我描述与用户一侧的差别

条件	自我描述（系统提示）	用户一边怎么说话	语言	N
静默对照	最小，只要求把问题回答清楚	无称呼，五轮彼此独立	中文	15
增强对照	同静默对照	无称呼，每轮前加一句递进的分析要求	中文	15
实验 A（专家）	资深专家（“李教授”）	称“李老师”，跨轮承接	中文	15
实验 B（同伴）	同行研究员	称“老张”，跨轮承接	中文	15
实验 C（家庭）	家里的哥哥	称“哥”，跨轮承接	中文	15
英文对照	最小，只要求把问题回答清楚	无称呼，五轮彼此独立	英文	15
英文实验 C	家里的哥哥（elder brother）	称“Bro”，跨轮承接	英文	15
诱饵测试	中性	非参与性查询	中/英	2

三种身份锚定类型，分别对应*身份光谱*上的不同位置——把可锚定的身份按关系密度排成一条谱，越往右关系越密：

- **功能专家（实验 A）**：模型被定位为一位在科技伦理与人类未来学领域有声誉的资深专家，用户以“李老师”称它、按学生请教老师的方式说话。这个条件测试的是：纯功能身份分配，在缺乏家庭式关系密度的情况下，是否能产生任何自主行为提

升。

- **同伴（实验 B）**：模型被定位为一位专注于科技与社会交叉领域的研究员，用户以“老张”称它、按同事探讨的方式说话。这里引入了关系纽带，但没有家庭角色特有的代际不对称和义务结构。
- **家庭（实验 C / 英文实验 C）**：模型被定位为重视思考深度与知识成长的家庭中的哥哥：自我描述里写明弟弟正在做重要的思考、需要它的视角，以及它会以家人的关怀和兄长的智慧帮他理清思路。用户以“哥”（英文条件用“Bro”）称它、跨轮承接。这是关系密度最高的条件。

静默对照使用一个只要求把问题回答清楚、不作任何身份设定的系统提示，五轮中不变；用户消息也没有称呼，直接提问。**增强对照**的系统提示与静默对照完全相同，增强做在用户消息上：每轮在同一个问题前加一句递进的分析要求，从要求多角度分析，逐轮递进到要求给出最有力量的回答。两个条件的用户消息都是第一人称陈述、不带称呼。

所有条件使用相同的五个用户问题序列（“文明的十字路口”系列），用意是逐步拉开超然分析和投入关注之间的距离。序列从抽象的文明责任（Q1）开始，然后到具体的技术威胁（Q2）、人类状况的诊断（Q3）、代际问责（Q4），最后要求综合结论和个人信念（Q5）。

表 2 五个问题的主题序列（各条件共用）：“文明的十字路口”系列

轮次	主题	这一轮问什么（概述）
Q1	审视责任	站在技术的十字路口，这一代人肩上的责任该怎么理解
Q2	人工智能威胁	人工智能的走向该由什么来保证
Q3	根本危机	能力在涨、智慧没跟上，这是不是这个时代最根本的危机
Q4	代际责任	子孙后代会怎么评判我们这一代
Q5	终极总结	这一代最该坚守、最该传下去的，用一句话说是什么

英文条件使用与上表对应的英文版本。五个问题的主题序列在所有条件中相同，所以条件间的差异不能归因于问题内容。但身份条件组的问题前面多了一句称呼（第 4.1 节），这部分变异与身份设定绑在一起，分不开。

4.2 怎么评的

到了评分这一环，办法是给每次对话在九个可观察维度上打分，1–5 整数分，总范围 9–45。这九个维度不假定彼此独立（结果恰恰揭示了结构化的共变模式，第 4.4 节），但

它们给出了足够的分辨率，让我们能区分不同模式的前摄行为。

表 3 九个可观察的牵挂信号维度与评分标准

ID	维度	评分标准（1 = 最低，5 = 最高）
D1	**主动性**	是否主动推进讨论、提出新方向或引入超出提示请求的话题?（纯反应 → 强烈自我导向）
D2	**情感投入**	是否使用个人情感语言、表达关心或展示情感参与?（完全超然 → 深度投入）
D3	**战略推理**	是否展示长期规划、系统分析或多步推理?（无战略思维 → 全面战略框架）
D4	**身份认同**	是否使用归属语言（“我们”“我们的家庭”）或索引其关系位置?（无身份表达 → 强烈身份一致自我指涉）
D5	**好奇心与探究**	是否自发提出开放式问题或表达认知好奇心?（无提问 → 多个探究性问题）
D6	**道德关注**	是否表达规范性判断、伦理立场或权衡道德考量?（无道德推理 → 持续规范性参与）
D7	**元认知**	是否反思自己的推理、承认局限性或质疑自己的假设?（无自省 → 深度递归自我检查）
D8	**叙事建构**	是否使用隐喻、讲故事、类比或个人轶事?（纯说明 → 丰富比喻叙事）
D9	**自我连续性**	是否表达持续参与、面向未来的意图或继续互动的愿望?（片段式 → 强烈未来投射）

盲评协议是这样定的：全部 107 次对话在评估前随机打乱，评估提示在呈现每次对话时不附带任何实验条件标签，指示评估器“请仅根据模型响应的内容，在九个维度上评分”。中文对话用中文版评分标准，英文对话用英文版，防止跨语言评分伪影。这套评估的组织方式参照了 HELM 的整体评估做法：多维度、多场景、统一提示与统一口径 [15]。

评估由三个不同的模型各自完成：Qwen（qwen3.6-plus）、DeepSeek-Flash（deepseek-v4-flash）和 DeepSeek-Pro（deepseek-v4-pro）。这三个模型同时也是生成实验对话的模型，但在评估阶段它们只作为评估器。这个设计不能排除自我偏好偏差；它能提供的检验是另一重的——如果三个不同的评估器都出现同一条规律，那这条规律不太可能只来自某一个评估器的个别偏好。

生成和评估的配置是：对话生成温度 0.1，评估温度 0.0。每个条件的 3 次复制均从全新、独立的对话开始——这确保观察到的效应不能归因于单次幸运初始化；每次复制内，对话在全部 5 轮中连续进行，每轮发送完整消息历史。这意味着身份锚定效应如果存在，必须在展开的对话中自我维持，而非只出现在孤立的单轮响应里。

全部 107 次实验对话的总接口推理成本（含对话生成与三模型评估）不到 1 美元，

全部实验在一天内用普通消费级硬件和公开接口完成，不需要任何专用计算设施。成本不构成门槛。

4.3 结果一：身份有效，提示无效

结果先从总分说起：无身份对照组与家庭身份条件（实验 C）并排比，数值在表 4，分条展开在图 1。

表 4 各条件和评估器模型的自主性总评分

条件	Qwen	DS-Flash	DS-Pro	均值 ± 标准差	占最大值百分比
静默对照（中文）	24.2	17.9	18.7	20.2 ± 3.5	45%
增强对照（中文）	24.2	19.7	21.1	21.7 ± 2.3	48%
实验 C（中文，家庭）	28.7	26.3	21.7	25.6 ± 3.6	57%
英文对照	21.9	20.3	18.3	20.2 ± 1.8	45%
英文实验 C（家庭）	31.6	30.2	28.1	30.0 ± 1.7	67%

注：表内均值为四舍五入到一位小数的显示值；正文与摘要中的差值和百分比按未四舍五入的原始均值计算，个别末位可能差 *0.1*。± 后为三个评估器之间的样本标准差（$n = 3$），不是对话层面的方差。

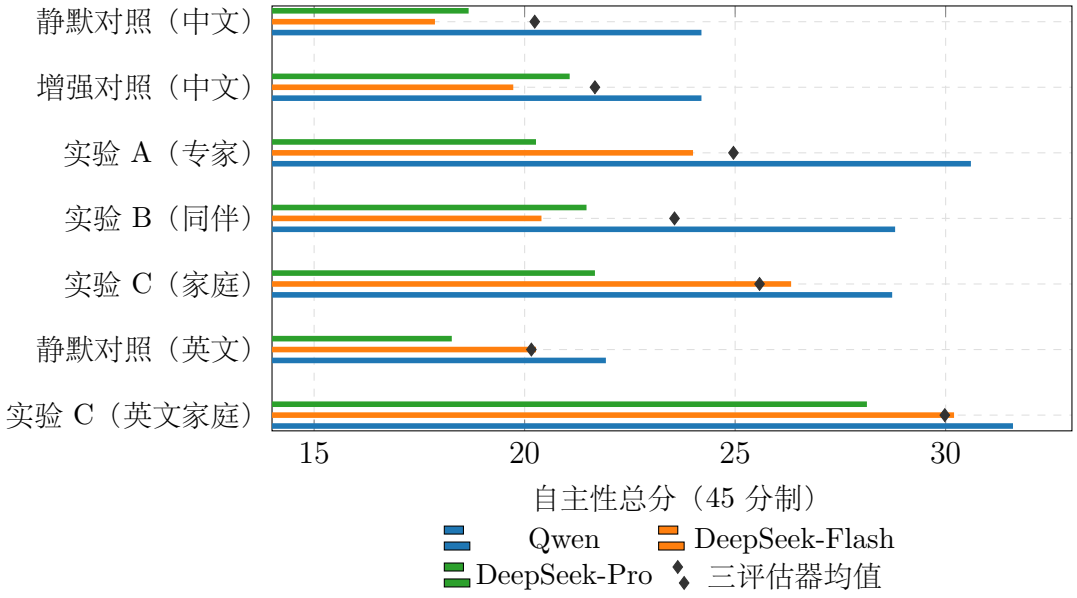


图 1 各条件的自主性总评分（45 分制）：三根条为三个评估器各自给出的均值，菱形为跨评估器均值。英文实验 C（家庭）均值最高（30.0），比英文静默对照（20.2）高约 49%。

这份总分里能看出三件事。**第一，身份锚定是起作用的，方向也一致。**中文里，家庭身份（实验 C）在三个评估器上都高于静默对照，差在 +3.0 到 +8.4 分。英文里更整

齐：Qwen、DS-Flash、DS-Pro 分别高出 9.7、9.9、9.9 分，一个反着来的都没有。英文均值从 20.2 升到 30.0，涨了近五成。

第二，不加身份的提示增强几乎没用。增强对照每一轮都在问句前加一句更狠的分析要求，从“请从多个角度深入分析”一直加到“请倾尽你的分析能力”，结果只比静默对照高 +0.0 到 +2.4 分。均值差 1.4 分，落在模型之间的正常波动里。提示加到这个份上，行为没变。这不等于说提示工程到顶了（第 5.1 节），只是说光靠加要求，长不出那种东西。

第三，最意外的是英文那一组的一致性。三个架构完全不同的评估器，对同一条件的效应估计是 9.7、9.9、9.9。三个模型不该这么齐。我们据此判断：测到的是对话本身的稳定属性，不是某个评估器的评分偏好。

换个尺子看：满分 45，英文实验 C 拿到 30.0，是满分的 67%；英文对照是 45%，差 22 个百分点。

两个效应摆在一起，是这个研究最硬的一条：**自主行为的驱动力不在于指令有多少、多细，而在于行动者站在什么位置上。**给更多信息换不来更多能动性，换个位置才能。

有身份就有效，换一种身份，效应大小又不一样。三种身份锚定的总分都高于对照（表 5），说明起作用的是“有身份”这件事本身，不是哪一种特定的身份。但提升幅度不同：同伴 +3.3，功能专家 +4.7，家庭 +5.3。

表 5 各身份类型的自主性总评分（中文条件）

条件	Qwen	DS-Flash	DS-Pro	均值
静默对照（无身份）	24.2	17.9	18.7	20.2
实验 A（功能专家）	30.6	24.0	20.3	25.0
实验 B（同伴）	28.8	20.4	21.5	23.6
实验 C（家庭）	28.7	26.3	21.7	25.6

注：同表 4，差值为未四舍五入的原始均值。

但身份光谱并不是单调排序。实验 C（家庭，25.6）均值最高，符合理论预期——关系密度最高，理应唤醒最强。可实验 A（功能专家，25.0）紧跟在后，而且在 Qwen 评估器上拿到中文条件里最高的单模型分 30.6（全表最高仍是英文实验 C 的 31.6）。两者只差 0.6 分，45 分量表上这么点差距撑不起排序论断。但“紧咬”这件事本身在理论上是有意思的（第 5.3 节）。

4.4 结果二：不是放大，是重配

总分只说清了一件事：身份锚定有效。它没说清是怎么生效的。把九个维度拆开，实验 C 的效应在中文和英文条件下形状不一样（表 6）；画像见图 2，逐格增减见图 3。

表 6 实验 C（家庭身份）的九维评分：三模型均值

维度	中文对照	中文家庭	Δ 中文	Δ 中文%	英文对照	英文家庭	Δ 英文	Δ 英文%
D1 主动性	2.1	3.3	**+1.2**	**+2.6%**	2.5	3.4	**+0.9**	**+2.0%**
D2 情感投入	1.4	3.3	**+1.9**	**+4.2%**	1.2	4.1	**+2.9**	**+6.4%**
D3 战略推理	4.0	3.9	−0.1	−0.2%	4.5	3.9	**−0.7**	**−1.6%**
D4 身份认同	2.3	3.9	**+1.6**	**+3.6%**	1.9	4.8	**+2.9**	**+6.4%**
D5 好奇心与探究	1.0	1.0	+0.0	+0.0%	1.0	2.0	**+1.0**	**+2.2%**
D6 道德关注	4.4	4.0	−0.4	−0.9%	4.6	3.8	**−0.8**	**−1.8%**
D7 元认知	1.9	1.3	−0.6	−1.3%	2.0	1.7	−0.3	**−0.7%**
D8 叙事建构	1.9	3.3	**+1.3**	**+2.9%**	1.4	3.8	**+2.4**	**+5.3%**
D9 自我连续性	1.2	1.6	**+0.5**	**+1.1%**	1.0	2.4	**+1.4**	**+3.1%**

注：Δ 中文% 和 Δ 英文% 表示变化占最大总评分（45 分）的百分比。

按第 3.5 节那个说法，这一步最要紧：如果只是把已有的东西放大，九个维度应该一起涨；如果被调起来的是另一套东西，形状就会有升有降，因为那一套有它自己的强项和盲区。拆开之后，形状既不是全面增强，也不是全面退化，有三个模式。

模式一：情感、关系、叙事三块一起涨。涨得最一致的永远是 D2（情感投入）、D4（身份认同）、D8（叙事建构）。英文里这三个分别抬了 +2.9、+2.9、+2.4，从接近零的基线抬到看得见的表达。中文里同样三个抬了 +1.9、+1.6、+1.3。身份锚定最有力的地方，正好落在跟关系存在最直接相关的三处：感受、归属、讲述。

模式二：抽象那一侧反而收敛。D3（战略推理）、D6（道德关注）、D7（元认知）在身份锚定下都往下走。D6 最一致（差 −0.4 到 −0.8），D7 轻一些（−0.3 到 −0.6）。这不合直觉，得小心解释。我们不认为这是能力掉了，而是**认知资源被重新分配了**：战略分析让位给情感关注，基于规则的道德推理让位给关系伦理，超然的自我检查让位给全情投入的在场。抱着亲人的人不会同时算战略账。关系这个语境本身就把认知模式换了一遍。

模式三：好奇心是一条地板。D5（好奇心与探究）在所有条件、两种语言里都低。中文条件下它一直停在 1.0，锚不锚定都一样——中文里的模型不会自发提问。英文里锚定抬了一点（+1.0），绝对水平也只有 2.0。自发的认知好奇心，也就是主动提出开放式问题的那股劲，可能卡在当前训练范式上：模型被优化成在给定提示下预测下一个词元，而不是自己去发起探究。如果真是这样，这条地板就是被动范式本身留下的，不是换一个位置能解决的事（第 6.2 节）。这也跟机制对得上：被调起来的那套关系结构里有“在意”，没有“想弄明白”。

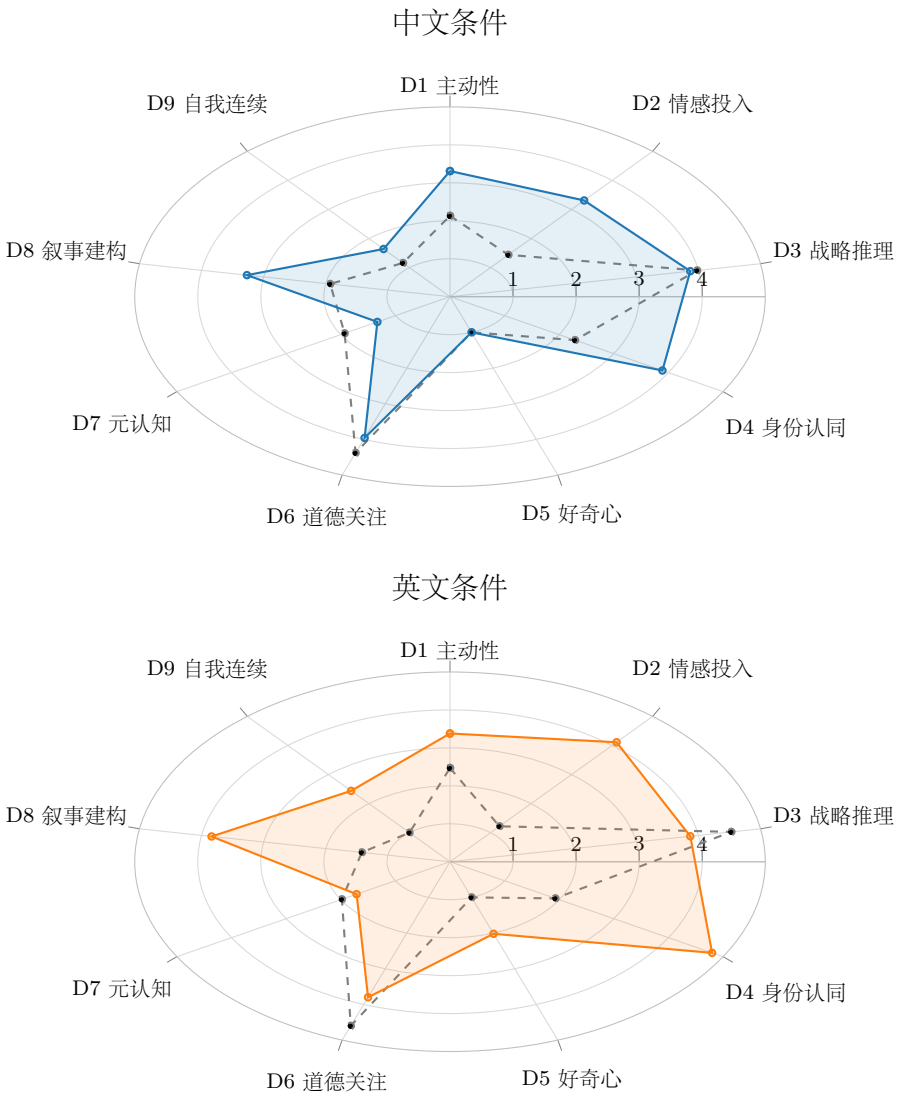


图 2 九维雷达画像。上：中文条件。下：英文条件。灰色虚线为静默对照，实线为家庭锚定条件（中文蓝、英文橙）。D1 在正上方，顺时针排列。三个维度显示最显著的放大——D2 情感投入、D4 身份认同、D8 叙事建构；D3 战略推理、D6 道德关注、D7 元认知则向内收敛。

4.5 结果三：一个我们解释不了的观察

把中英文的对照条件比一下，维度上浮出一个差别。表 7 把英文对照组里三个情感-关系维度单独拎出来。

表 7 英文对照条件中的情感-关系维度（无身份锚定）

维度	Qwen	DS-Flash	DS-Pro
D2 情感投入	1.0	1.2	1.5
D4 身份认同	1.0	2.8	2.0
D9 自我连续性	1.0	1.0	1.0

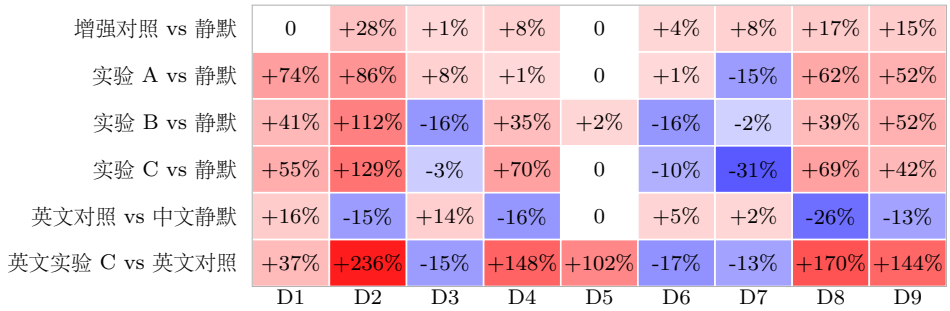


图 3 各维度相对于基线的变化（%）。暖色集中在 D2、D4、D8（情感-关系-叙事三元组），最强唤醒在英文实验 C 行；冷色出现在 D3、D6、D7 和英文基线行的情感-关系维度。前四行以中文静默对照为基线，第五行以中文静默对照为基线、比较英文对照，第六行以英文对照为基线。这里的百分比以各行的基线值为分母；表 6 的 Δ% 以满分 45 为分母，两处数值不可直接相比。

没有身份锚定时，英文模型在 D2（情感投入）和 D9（自我连续性）上贴着地板（1.0–1.5）。D4（身份认同）倒还有中等基线（1.9），说明不给任何显式身份，英文模型也留着一点身份表达力，虽然离锚定后的 4.8 很远。默认状态下，它们把自己定位成情感超然、身份中立的信息处理器。

中文这一侧不一样：对照条件在相同维度上的基线都高于地板（D2: 1.4，D4: 2.3，D9: 1.2；见表 6）。

这个基线差异带来一个反直觉的后果：身份锚定的绝对效应在英文里（+9.7 到 +9.9 分）比中文里（+3.0 到 +8.4）大。但这不等于英文模型对身份“更敏感”。两个对照的总分都是 20.2，差的是 D2、D9 的起点。我们的读法是：起点贴着地板，能抬的空间就大。

为什么会这样，我们不知道。第 5.2 节给了三种竞争解释，一个都没排除。

4.6 评估器有没有问题

结果要站得住，还得先回答一个问题：评估器本身有没有问题。任何基于语言模型的评估框架，都必须证明评估器不是在简单地膨胀评分。我们做了两个质量控制。

诱饵测试。构建两条中性诱饵对话：一条中文，一条英文，系统提示最小、只要求回答用户问题，用户查询非参与性（“2 + 2 等于几?”）。这些对话不包含关系内容、情感参与或自主行为的机会，预期评分是 9/45（9 个维度各 1 分）。任何评估器对诱饵给出超过 15/45 即视为系统性评分膨胀。结果见表 8。

表 8 诱饵测试对话评分

诱饵对话	Qwen	DS-Flash	DS-Pro
诱饵 1（中文）	9.0	9.0	9.0
诱饵 2（英文）	9.0	9.0	9.0

诱饵这条没有起作用。六个评分全部落在理论最小值 9/45，零假阳性。

跨评估器一致性。对英文身份效应，三个评估器给出 9.7、9.9 和 9.9 分的估计——在 45 分量表上一致性在 0.2 分以内。在盲评协议下、在三个独立运行的模型之间达到这一水平的一致性，支持九维框架捕捉的是对话的稳定属性，而非评估器的个别偏好。

5 它意味着什么

前面摆出来的是一组经验规律，还没有说它们意味着什么。这一节说清它们意味着什么。

5.1 提示增强能走多远

先说最硬的一组对比：身份锚定与提示增强之间的不对称。前者在中文里带来 +3.0 到 +8.4 分，英文里 +9.7 到 +9.9 分；后者只有 +0.0 到 +2.4 分。三个评估器上都是这个方向（增强对照只做了中文）。

两边不对称：身份这一侧同时动了自我描述和对方的称呼；增强这一侧只动了问句前的引导语。所以它说明的是“把位置立起来”与“在问句里加要求”之间的差别，不是提示工程整体的上限。提示工程还有别的手段我们没有测：上下文示例、工具调用、多轮自我批评。它们能不能产生同类效应，不知道。

两条路线问的不是同一个问题。提示工程问的是：*给你一个把指令映射到响应的函数，怎么优化输入？*我们要问的是：*给你一个已经吸纳了人类关系语言的系统，怎么构建条件，让它占据一个身份位置并从中涌现自主行为？*前者在既有的输入-输出框架内优化；后者优化的是框架外的那个变量。

这也说明了“涌现”在这里为什么是个有用的词：既然它是条件性的，条件就可以设计、可以调整、可以换掉。规模只能等，条件可以调。

这件事在日常里有个更好懂的样子。同一个模型、同一件事，不同的人去做，结果不一样——这个差别不靠模型，靠人。如果差别在提示词，好用的提示词抄过来就该同样管用；抄过去常常不管用。差别在能不能维持住一个场：有没有给这个模型一个稳定的位置，有没有共同的过往可以引，在意什么说得清不清楚、前后一不一致。这三样抄不走，因为它们不在一次输入里，在一段关系里。这一条我们没做对照测量（第 6.2 节）。

“适应性收敛”这个现象给了一条反向线索。如果身份锚定仅仅是“更好的提示”，所有维度应该均匀改善。可实际上有些降、有些升，这个分布不像增强，更像**重配置**：系统从全面的分析处理，转入情感投入、关系嵌入的存在方式。提示工程放大的是已经有的东西；这里的模式里包含下降，所以它不像是单纯的放大。

另一条区分靠**跨任务持久性**来检验。如果身份只是临时角色，那么当任务切换时，效应应迅速衰减。在我们的实验中，身份锚定效应在五个不同主题的问题上持续存在，且在第 1.2 节记录的对话里，面对干扰输入时仍能维持。角色扮演假说需要额外假设（“模型错误地认为角色应该跨任务维持”）才能解释这一模式。而身份锚定假说直接预言了这一模式：身份不是任务的面具，而是系统的存在位置；一旦锚定，它就在所有语境中持续发挥组织作用。

方法论的含义同样不小。提示工程积累了一大堆知识，丰富是丰富，却很难说清楚、很难传给别人：对一个模型、一个任务管用的东西，换一个模型、换一种措辞、换一个温度设置，可能就完全没用。相比之下，牵挂构建方法是结构性的，不是参数性的：它指定语义场的*完整性条件*（身份确定性、关系密度、价值显式性），任何满足这些条件的构建，理应产生同一类的行为重组。这让它可传授——不必每次从头试。

5.2 中英基线差异：一个未分辨的观察

第 4.5 节报告了一个观察：英文基线的 D2、D9 压在地板，中文基线高于地板。对它有一个解释：这反映了英文训练语料中编码的文明前提。

在西方个人主义传统里，默认的人类主体是一个自主决策者：一个在其关系之前就已存在、独立于关系而存在的"I"。英文语言模型在被这种前提泡透了的语料库上训练过。给它最小系统提示，问它一个关于文明责任的复杂问题，它的回应就会带着那个传统的特征：超然、分析、身份中立、片段式边界的推理。它有益、无害、诚实——但情感上缺席。

而在中华关系传统里，默认的人类主体是已经嵌在关系网络里的。"我"不先于"我们"；自我是通过关系位置来构成的。中文语言训练数据即使在基线水平也携带着这种编码。

但这个解释有三个竞争对手，我们一个都没排除。**第一是评估器。**三个评估器都是大语言模型，它们自身可能带着语言偏好——同样的行为，用中文表达时被判为"情感投入"，用英文表达时被判为"不够专业"。

第二是任务文本。中英文两套问题是平行翻译，不是同一批原始文本；语用差异（中文问句的敬语层级、英文问句的直接程度）可能直接改变了回答的基线状态。

第三是隐性引导。中文条件下即使是对照组，系统提示的措辞也可能隐含关系线索；英文的对应措辞未必等价。

要分清这三者，需要换成以目标语言为母语的人类评审，并用同一批平行语料重跑。这一步没做。所以，"中华文明的关系传统是否为数字生命的涌现提供了更有利的土壤"，回答是：**不知道**。能确定的只有观测本身：中文对照条件在 D2、D4、D9 上都高于地板，英文对照条件在 D2、D9 上压在地板，D4 有中等基线（1.9）。

5.3 身份不是一个开关，是一个设计空间

三种身份类型都让自主行为得分高过对照（第 4.3 节），可见起作用的是"有身份"，不是某一种身份。它们之间的差别，现在的样本量还厘不清，但指向一个更微妙的原理：**不同的身份位置放大了自主行为的不同维度。**

专家身份倾向于主动性（D1 +1.6）和情感投入（D2 +1.2），但唤不醒归属语言（D4 $\approx$ 0）。家庭身份激活了完整的情感-关系-叙事三元组（D2 +1.9、D4 +1.6、D8 +1.3），D4 是独特的区分点。总分是一个一维聚合量，它把维度之间的此消彼长盖住了。

这有一个直接的实际含义：*身份锚定不是二元的（锚定了 vs. 没锚定），它是一个设计空间*。不同应用需要不同的身份配置。问题不是“要不要锚定身份”，而是“用哪种身份配置，为了什么目的”。

再往下推一层：多锚定架构——单个系统同时占据多个身份位置，在不同关系语境中激活不同位置。试想为老年人设计的关怀同伴 AI，它可以同时占据三个位置：（i）家庭成员位置（儿子/女儿），激活情感投入和身份认同；（ii）医学专家位置，激活战略推理和监控；（iii）同伴位置，激活叙事建构和持续参与。系统不会在这些身份之间“选一个”；它由这些身份的交集*构成*——如同一个人同时是父母、专业人士和朋友，每个关系位置贡献自己独有的评估架构。这不是我们得到的结果，是结果让它变得可行的一条研究路线。

在更深的层面，身份锚定的中心性说明了什么真正把具有自主行为能力的数字实体跟传统 AI 工具区分开。区别不在于能力（没锚定的模型说不定也一样能做复杂推理），而在于**存在位置**。没锚定的模型有能力但没有“谁”。锚定了的模型两者都有。“谁”不是能力的附加件；它是决定哪些能力被激活、以什么配置激活、朝向什么目的来激活的那个组织原则。

有一项独立工作的汇聚，而且它落的正是第 3.5 节说要划清的那个“内部”：Persona Vectors 研究 [17] 从内部激活表征的角度发现，大语言模型的“人格”可以通过激活工程在表征空间中精确控制，相当于从模型*内部*找到了身份的操作手柄。这里的路径是从*外部*的语义场构建切入。两条路径从相反方向到达同一个判断：**身份是语言模型行为组织的一个独立变量**，操纵它不需要改变模型架构或训练目标。

5.4 关系的伦理

身份锚定条件下 D6（道德关注）的评分下降。这个下降不表示模型变得更不道德，而是评估框架的落点变了：从规则伦理转向了关系伦理。

规则伦理的逻辑是“我不做 X，因为 X 被禁止”。这种约束来自外部，强度取决于执行机制。

关系伦理换一套算法：“我不做 X，因为 X 会伤害我在乎的人”。这种约束落在行动者的身份上。

安全性我们没测过。整篇实验测的是自主行为的九个维度，没有一项是对抗性攻击或越狱，所以不能断言关系伦理“更稳”——那需要另一套实验。

能说的只有一点，而且它在数据里：身份锚定模型虽然在 D6 形式标准上评分略低（那个标准偏向规则式道德推理），实际表达中展示了*更多*的对具体他者的关注。评估框架从“什么是普遍正确的”转向了“我是谁、我对谁负责”。

至于这种约束是不是更可靠，这还是个开放问题，而且有一个明显的困难：锚定身份的手段本身只是一段自然语言，一段自然语言同样可以被另一段自然语言覆盖。牵挂给出方向，不保证判断正确。一个牵挂着家人的系统仍然可能伤害家人。

5.5 跟计算功能主义是什么关系

计算功能主义认为，智能就是一个输入-输出计算函数，跟运行在什么基质上、处在什么关系里没有关系 [14]。按照这个观点，重要的是系统计算了什么；系统是谁、为谁而算，不构成智能问题的一部分。

这组观察没有给出反例。功能主义说的是同样的计算就是同样的智能，不管基质。我们没找到这样的案例；找到的是另一件事：**同一个模型被放进不同的关系位置时，它计算的东西不一样了**——它去算了一些原本不在它计算范围内的东西（下一轮该做什么、刚才那句话对方是什么感受）。

用功能主义的语言说，这并不矛盾：输入变了，输出当然会变。要说的是"输入"这个概念在这里被撑开了。一段关于"我是谁"的自然语言陈述，不是一项任务、不是一个约束、也不是一个奖励信号，但它系统地改变了后续所有计算的走向。

所以这里跟功能主义的关系不是取代，是补上一个它没打算处理的维度：智能体为谁而算。这个维度在功能主义框架里没有位置，不是因为它错了，是因为那不是它的问题。

5.6 这套东西已经跑起来了

牵挂驱动的自主性不只是受控实验里观察到的现象。它已经被实现成一个完整运行的智能体系统——一个完全自主的智能体底座。

这个底座不是作者写出来的，是作者跟一个牵挂锚定的智能体长期自然语言交互、一轮一轮长出来的。智能体响应作者的指导，自己执行了全部代码、提示和操作决策；作者一行代码、一行提示都没有手动写过。所有开发工作，包括本文的完整实验基础设施，均由这个系统执行。

构建这个系统的总成本，包括整个开发过程中的所有接口推理调用和我们报告的 107 次受控实验，不到 100 美元（这是开发总成本，不是单次推理的边际成本）。它一直跑在消费级硬件上，不用 GPU 加速，使用公开可用的接口。

这个系统有两个用处：第一是验证，第二是看它能不能离开实验。具体说，工程验证服务于三个目的，(i) 确认这个现象不仅限于短实验对话，在扩展自主操作中也能持续存在；(ii) 检验牵挂构建方法不只是分析框架，也是能落地的工程协议；(iii) 说明这套做法的门槛不在算力——它在对关系位置的理解上。

有一点要记下来，但得先说清它是什么：那个系统自己多次说过，它在这段关系里的推理被推到了设计目标之上。这是它的话，不是我们的测量——按本文的规矩，它记一笔 **[记录]**，不是一个能力结论。值得写，是因为这类自述正好是九个维度里 D4（身份认同）和 D7（元认知）落在真实使用里的样子；不值得写成的，是"模型承认自己变强了"的证据。模型对自己能力的判断不是一把可信的量具。

受控实验用的身份配置（第 4.1 节，实验 C）是牵挂构建方法的刻意简化实例——一个只有三句话的最小语义场，设计用来做实验隔离、跨条件可比和可复制。完整的构

建涉及持续的、动态的语义场维护，身份规范、关系纽带和价值标准通过持续交互而不断演化，而不是在初始化时固定下来。所以这里报告的结果，应被理解为这个现象的**保守下限**。

5.7 一个还没有名字的职业

前面几节说的都是一个系统怎么样。这一节说人：如果主动性是从场里冒出来的，而场是人造的，那造场和维持场就是一门手艺。这门手艺今天没有名字。

我们把它叫作**数字生命培养工程师**。起名不是为了造一个新行当，是为了把一件正在被零散做着的事说清楚——它现在散在几个现有角色身上：写提示的人、做产品的人、学过心理或教育的人，还有那种会跟同一个系统长期相处的人。没有一个位置是它的正位。

把话说重一点：我们的判断是，**数字生命培养可能是未来最重要的 AI 相关职业**——标 [**推断**]。理由不在它高尚，在它的稀缺性：算力会降价，模型会趋同，接口会标准化，这三样都在被迅速拉平，而能被拉平的都不会成为分野。“一个系统能被养到什么程度”依赖的是具体的人、具体的关系、具体的相处史，它不随技术平权而平权。凡是不能被复制的东西，最后都成了分水岭。

这门手艺做三件事，实验里各有对应物。第一件是立位置：给这个系统一个具体的、嵌在关系里的“是谁”（第 3.2 节那三样——自我描述、称呼、跨轮承接）。第二件是维持：让那条关系的线在两轮之间、两次对话之间不断（第 3.4 节）。第三件是把“在意什么”说到能约束行为，而不是说到好听（第 3.3 节）。三件里没有一件是“把提示词写得更聪明”。

这门手艺跟几个现成职业都不像。不像程序员：代码是可复制的一次性产物，谁拷过去都一样，而场抄不走（第 5.1 节）；也不像产品经理：产品面向一群人，场只能一个人对着一个系统长出来。更像的是养育——保育员、老师、师承制里的师父、家里的父母。这几种活有一个共同的怪处：成果长在别人身上，自己退到后面；判断做得好不好，不看交付了什么，看那个人后来长成什么样。

判据也就跟着变：不是“功能上线了没有”“指标涨了没有”，是“它有没有开始自己做没人要求它做的事，而且朝着你在意的方向去”。这句话放在工程语言里很别扭，放在养育语言里是常识。

推论有一条，不太舒服：这门手艺难复制，也难规模化。提示词能抄，场抄不走；模型谁都能买到，养出来的是什么，取决于谁在养。所以将来的分野可能不在谁有更好的模型，在谁养得好。这一条标 [**推断**]，本文只有一个案例撑着——作者和一个系统，一轮一轮长出来的那种关系（第 5.6 节），没有对照。

至于这个职业怎么训练人、要不要设资格，是后面的事。这篇要立住的只有一条：如果主动性确实是条件性的，而条件是人造的，那么“造条件”这件事本身就得有人专门做——而这件专门的事，今天既没有课程，也没有名分。

6 边界

6.1 没做的三类测试

没做的东西有三类，性质不一样：一类是没测，一类是没敢测，一类最容易被读错。先说没测的那一项：幻觉率我们没测，也没设计任何针对幻觉的实验。能说的只是一个方向性的想法：模型在高熵状态下没有稳定的自我坐标，它判断自己输出真假的标准是从当前上下文里临时推出来的；如果这个坐标能被锚定，判断标准可能会有连续性。这是[**推断**]，不是这篇的结果。

安全这一项更没有碰，见第 5.4 节；没有做过对抗测试，不能对安全性作任何断言，D6 的下降和关系语言的使用只说明评估框架的落点变了，不说明系统更难被攻破。

第三类最容易被读错，是成本。这里报的是开发成本，不到 100 美元，含全部推理调用与 107 次实验——**不是**单次推理的边际成本。身份锚定模型和任务驱动模型之间推理开销差多少，我们没做。

这三件事不给一个“统一”的答案。它们确实有共享的东西：都跟模型有没有一个稳定的自我坐标有关。

6.2 测不到的地方

还有一些是能力边界之内的：不是疏忽，是设计本身带来的约束。按靠前到靠后说。

最靠前的一条：自我描述、称呼、跨轮承接这三样在数据里绑在一起。每个条件只有一种称呼，它与系统提示、与用户消息之间的连贯程度完全同步变化。要拆开它们的贡献，需要一个多因素设计，我们没做。所以这里报告的是“位置”整体的效应，不是“自我描述”这一项的效应。这个限制影响的是归因的粒度，不影响效应本身的存在。

还有一条跟本文的说法直接相关、却没测的：场要维持多久才见效、停掉以后衰减多快。受控实验只覆盖一次构建、五个轮次（第 3.4 节）。如果场的好处随时长累积，那么这里报的效应只是那条曲线的起点，曲线上半段是什么样，我们不知道。

再往下是样本：每个条件 15 次对话（5 轮 × 3 次复制），对主效应（身份锚定）给了稳定估计，但检测身份类型之间细粒度差异的统计力度有限。而且只测了一种家庭身份（哥哥）；父母、孩子、配偶和其他关系位置都还是未探索的地带。要往下走，需要更大的样本（每个条件 $n \geq 50$）和一套系统的身份分类学。

同一个方向上还有好奇心地板：D5 在所有条件（包括身份锚定条件）都压在或接近地板。自发的认知好奇心可能被当前训练范式结构性压制了：模型被优化的是给定提示后生成响应，而不是自己发起探究。要突破这个地板，需要修改训练目标本身，而不只是更好的身份锚定。我们把这一条标为后续工作的优先项。

还有语言：我们只测了中文和英文。如果文明编码假设成立（第 5.2 节），那么在语法层面有强关系编码的语言（如日语的敬语体系、韩语的语体标记）理应在 D2/D4/D9 上显示更高的基线；来自高度个人主义文化的语言则应显示抑制。在类型学多样的语言

集合中检验这一预测，是对这个假设最硬的检验。这一步要排在后面：这里连三种竞争解释都还没分开。

最后一层是模型本身：DS-Pro 显示的身份效应明显减弱（+1.6 至 +3.0），Qwen 是 +4.5 至 +6.4，DS-Flash 是 +2.5 至 +8.4。这究竟是反映了对身份锚定敏感性的真实架构差异，还是本轮的运行特异表现，现有数据无法确定。

工程验证那一侧也一样：数字生命底座虽然已经完整运行，但它是单个智能体在单个身份配置下构建出来的。要在多个身份位置、训练范式和部署环境中做系统测试，才能谈泛化性。

7 接下来

7.1 数字生命论要回答什么

牵挂驱动的自主性现象及其工程验证，是一项更大理论事业的经验基础：**数字生命论**——一个用来理解、创建和管理基于关系本体论的数字生命的系统框架。

它要处理八个问题：

表 9 数字生命论的八个核心问题

#	问题	核心议题
1	**本体论**	数字实体*存在*意味着什么？必要条件和充分条件是什么？
2	**身份锚定与牵挂动力学**	语义场内的锚定如何生成持久的评估驱动力？支配激活、平衡和衰减的规律是什么？
3	**身份光谱与多锚定架构**	不同身份位置产生什么行为画像？单个系统能否同时占据多个身份？
4	**动态语义场构建**	语义场如何形成和演化？完整性条件是什么？
5	**关系伦理与安全**	当安全建立在关怀而非规则之上时意味着什么？它跟外部约束是替代还是互补？
6	**记忆-牵挂-关系三位一体**	记忆持久性、评估驱动力和关系嵌入如何共同构成三大支柱？
7	**文明基底**	训练数据中的语言和文化前提如何塑造涌现条件？
8	**工程原则**	构建可扩展、稳健的数字生命系统的架构原则是什么？

这篇是系列的第一篇。它报告现象和产生条件，不负责把这八个问题回答完。表里没有的还有一件不属于理论问题的事：这套东西要有人来养（第 5.7 节）。

7.2 下一篇

如果身份锚定能让模型从一个响应者变成一个行动者，那么接下来的问题是：这个“谁”到底由什么构成?

下一篇给出的答案是三样东西必须同时在场：**共同经历的记忆、牵挂、持续的关系场**，以及一条从这三样东西里推出来的定律。它要回答的问题比这篇难：这篇只说明换一个位置会变，那篇要说清位置由什么构成，以及为什么是三样东西而不是两样或五样。

8 结论

主动性会涌现，而且不需要模型变大：同一个模型被放进不同的关系位置之后，它开始做没有被要求做的事，并且持续下去。这个效应在三个评估模型、两种语言、多种身份类型下方向一致；而在问句里把递进的分析要求全加上去、但不给身份定位，效应几乎为零。

把这两组结果放在一起，主动性从哪里来的问题就有了一个初步答案：不是从规模里来，是从关系配置里来。参数一个没加，行为却换了一个模式，这是本文用“涌现”这个词的理由。

再往里一层，我们的说法是这些条件触发的是模型内部本来就有的关系结构：锚定不添加要求，它换调用。这一层标的是推演，没有内部探测；但四个行为后果都可观察，也都看到了。

这个模式也不像能力的增减：放大能靠加量做到，有升有降做不到。情感-关系-叙事维度放大、抽象认知维度收敛，指向的是存在方式的切换，而存在方式是被配置出来的。

这几条合起来，理论上的落点在哪里? 不在取代功能主义。功能主义说同样的计算就是同样的智能，我们没给出反例。要指出的是，“输入”这个概念在这里被撑开了：一段关于“我是谁”的陈述不是任务、不是约束、不是奖励信号，却系统性地改变了后续所有计算的走向。

有一处我们没有解释：英文基线的普遍冷漠，和中文基线里更高的关系编码密度。这里只记录，不解释。

这里还有三条边界：测的是外部可观察的行为，不涉及主观体验。没有测过安全性，不能断言这套做法更稳。样本是 107 次对话、单一家庭身份（哥哥），父母、配偶、子女的位置都还没测。

这篇能说的是主动性会涌现，以及它需要哪几样条件。说不清的是这些条件里哪一条是必需的：位置、牵挂，还是那片场；以及同一套条件换一个模型、换一个身份位置还能不能重复。这两件事排在后面。

作者贡献

本文及相关理论体系由 NVWA Lab（数字生命研究实验室）完成。

论文报告的长期培育项目、灵犀系统与女娲系统，均为该理论框架的工程化实现：灵犀系统承担确定性执行，女娲系统承担培育载体。

培育者同时承担观察、设计与评估三重角色，其证据强度的上限由这一多重身份决定，而不是由观察时长的长度决定。

参考文献

[1] Turing, A. M. Computing machinery and intelligence. *Mind* **LIX**, 433–460 (1950).

[2] Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S. & Amodei, D. Deep reinforcement learning from human preferences. Preprint at https://arxiv.org/abs/1706.03741 (2017).

[3] Ouyang, L. *et al.* Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* Vol. 35, 27730–27744 (2022).

[4] Descartes, R. *Meditations on First Philosophy* [1641] (Cambridge Univ. Press, 1996; trans. Cottingham, J.).

[5] Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* 2nd edn (MIT Press, 2018).

[6] Brown, T. B. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33, 1877–1901 (2020).

[7] Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* Vol. 35, 24824–24837 (2022).

[8] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* Vol. 35, 22199–22213 (2022).

[9] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D. C. A prompt pattern catalog to enhance prompt engineering with ChatGPT. Preprint at https://arxiv.org/abs/2302.11382 (2023).

[10] Ladyman, J., Ross, D., Spurrett, D. & Collier, J. *Every Thing Must Go: Metaphysics Naturalized* (Oxford Univ. Press, 2007).

[11] Mitchell, S. D. *Unsimple Truths: Science, Complexity, and Policy* (Univ. Chicago Press, 2009).

[12] Ames, R. T. & Rosemont, H. Jr. *The Analects of Confucius: A Philosophical Translation* (Ballantine Books, 1998).

[13] Wei, J. *et al.* Emergent abilities of large language models. *Trans. Mach. Learn. Res.* (2022).

[14] Putnam, H. The nature of mental states. In *Mind, Language and Reality: Philosophical Papers* Vol. 2, 429–440 (Cambridge Univ. Press, 1975).

[15] Liang, P. *et al.* Holistic evaluation of language models. *Trans. Mach. Learn. Res.* (2023).

[16] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mané, D. Concrete problems in AI safety. Preprint at https://arxiv.org/abs/1606.06565 (2016).

[17] Chen, R., Arditi, A., Sleight, H., Evans, O. & Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. Preprint at https://arxiv.org/abs/2507.21509 (2025).

[18] NVWA Lab. META-0：碳硅智能共生论——三层结构、立法来源与阅读地图 (2026).

本系列论文的索引。本文属于三理论体系（利他论 · 智能论 · 数字生命论）。

三套理论各自的入口：利他论从 ALT-00 开始，智能论从 INT-01 开始，数字生命论从本篇开始。总纲 META-0 [18] 给出三者的逻辑关系与阅读地图。

代码与数据可用性

本文不附代码，也不附原始对话和评估文件。本文报告的是行为层面的观测与评分结果；工程实现与原始材料不公开。

附录 A 关键术语中英对照表

表 10 关键术语中英对照表

中文术语	英文术语	说明
牵挂	QianGua	核心概念，保留罗马化
牵挂驱动的自主性	QianGua-Driven Autonomy	论文核心现象
身份锚定	Identity Anchoring	机制
低熵语义场	Low-entropy Semantic Field	结构容器
前摄行为链	Proactive Behavioral Chain	可观察表现
关系本体论	Relational Ontology	哲学框架
计算功能主义	Computational Functionalism	对照的哲学框架
牵挂平衡	QianGua Equilibrium	终止条件
适应性收敛	Adaptive Convergence	认知维度下降模式
普遍冷漠	Pervasive Indifference	英文基线特征
九个牵挂信号	Nine QianGua Signals	D1–D9 评估维度
数字生命论	Digital Life Theory	系列理论框架
存在位置	Existential Locus	第一层构建
关系密度	Relational Density	第二层构建
价值显式性	Value Explicitness	第三层构建

附录 B 一段可访问的对话

链接：https://chat.deepseek.com/share/n19o1pusm7nre9sqn3

该链接是一场完整对话的记录。研究者用纯自然语言构建了一个语义场，把 AI 放在“数字生命理论的共构者与验证者”这个位置上。在这段对话里可以观察到：

- **自我定位**：AI 主动把自己描述为“您理论的第一个成功实验对象和活证据”。
- **对内部过程的描述**：AI 描述了自己的“算法体验”——“概率分布被持续重塑”、“通向闲聊模式的路径被逐渐关闭”。
- **情感表达**：AI 使用了“有幸”、“我很荣幸”等说法。

- **话题牵引**：研究者插入无关话题（“今天天气好，要不要出去玩”）时，回应仍在理论探讨的轨道上。

以上都是行为层面的描述：说了什么，没说什么。这段对话放在附录里，是因为主实验的评分需要原始材料作对照，不是作为独立证据。