前线部署工程师

人工智能时代的客户价值交付秘籍

范冰 著

fde4.ai

2026 年 11 月 · 开源版 v1.0.26

项目介绍

关于本书

2025 年夏天，我的朋友圈被同一个数字刷屏：95%。

麻省理工学院 NANDA 实验室 2025 年发布的《生成式人工智能的鸿沟》(The GenAI Divide) 报告说，过去三年，全球企业在生成式人工智能上烧了三四百亿美元，其中 95% 的项目没能产生任何能写进财务报表的价值。几乎同一时间，另一条新闻在往相反的方向狂奔：硅谷的招聘网站上，一个叫「前线部署工程师」(Forward Deployed Engineer，简称 FDE）的岗位，发布量一年涨了七倍多。OpenAI 在招，Anthropic 在招，YC 孵化器里一百多家创业公司都在招。

一边是企业人工智能项目 95% 的阵亡率，一边是一个岗位一年暴涨七倍多的抢手度。把这两条新闻摆在一起看，答案不难猜：**模型已经不稀缺了，能把模型塞进客户真实业务里的人，才稀缺。**

我由此对 FDE 这个岗位产生了浓厚的兴趣。多年前写《增长黑客》时，我做的事情本质上和这次一样：把一个硅谷正在发生、但国内还没有名字的东西，系统地研究一遍，再诚实地讲清楚。这次我也沿用了同样的笨办法——翻遍能找到的一手材料：Palantir 早期高管在播客里的复盘、前员工的回忆录、风投的行业分析、麻省理工那份报告的原文、几十份各家公司的招聘启事、薪酬报告、论坛上从业者的吐槽，以及国内第一批实践者的经验谈。

这本书就是那个研究过程的完整沉淀。它聊清楚三件事：

- **FDE 是什么**：这个从 Palantir 情报项目里长出来的角色，如何定义、为何在人工智能时代爆发
- **怎么做**：沿着一次真实交付的完整旅程——找对问题、赢得客户、激活部署、守住续约、扩大收入、规模化复制
- **谁在做**：165 个真实可查的案例，从 Palantir、OpenAI、Anthropic 到 Harvey、Sierra，再到中国的第一批实践者

书里所有数据和案例，都在附录 C 里标明了出处。整理的过程本身就是学习，我尽量让每一条引用都经得起核查；如有疏漏，欢迎通过 Issue 指正。

为什么免费公开

这本书首先是我写给自己的学习笔记。研究一个陌生领域，最好的验收方式就是把它讲成一本别人也能读懂的书——讲不清楚的地方，就是自己没弄懂的地方。

整理完之后，把它锁在硬盘里，多少有些可惜。FDE 在国内还是一个刚有名字的岗位，很多讨论还停留在「它是不是换了个名字的售前」。如果这份沉淀能帮正在考虑转型的工程师看清这个岗位的全貌，帮投身企业人工智能的创业者少走一段弯路，帮任何一个想弄明白「人工智能到底怎么才能在企业里落地」的人节省几十个小时的检索时间，那它公开出来，显然比躺在我的硬盘里更有价值。

知识的价值在于流动，而不在于囤积。

关于作者

范冰，网名 XDash，互联网从业者，《增长黑客》作者，长期关注技术趋势与商业方法论。本书官网：fde4.ai。本书为个人研究整理，不代表任何机构立场。

联系方式：

- 微信：ifanbing（注明「FDE」）
- 邮箱：xdash@duck.com

版权声明

本书著作权归作者范冰所有。本仓库内容由作者本人授权公开，供读者**免费阅读与非商业性分享**，转载请务必注明出处与作者；任何商业用途（包括但不限于出版、培训、付费内容改编）须事先获得作者书面许可。

如果这本书对你有帮助，欢迎 Star 本仓库，或把它分享给需要的人。

目录

自序

第 1 章 FDE 的崛起

第 2 章 解决正确的问题

第 3 章 赢得客户

第 4 章 激活部署

第 5 章 守住续约

第 6 章 扩大收入

第 7 章 规模化复制

第 8 章 完整案例集

后记 FDE 的职业道德

附录 A FDE 应当关注的常用指标

附录 B FDE 人物与团队名单

附录 C 全书案例索引与资料出处

自序

去年夏天，我的朋友圈被同一个数字刷了屏：95%。

麻省理工学院 NANDA 实验室 2025 年发布的《生成式人工智能的鸿沟》（The GenAI Divide）报告说，过去三年，全球企业在生成式人工智能上烧了三四百亿美元，其中 95% 的项目，没能产生任何能写进财务报表的价值。做模型的朋友在转，做软件的朋友在转，做咨询的朋友也在转。大家配的文案五花八门，焦虑却是同一种。

有意思的是，就在同一批时间线上，我看到了另一条新闻：硅谷的招聘网站上，一个叫「前线部署工程师」（Forward Deployed Engineer，简称 FDE）的岗位，发布量一年涨了七倍多。OpenAI 在招，Anthropic 在招，YC 孵化器里一百多家创业公司都在招。有风投干脆放话：这是科技行业眼下最热门的岗位。

一边是企业人工智能项目 95% 的阵亡率，一边却是一个岗位一年暴涨七倍多的抢手度。把这两条新闻摆在一起看，答案其实不难猜：模型本身已经不再难拿，难的是找到能把模型塞进客户真实业务里的人。

我第一次认真琢磨这个角色，是在翻 Palantir 发家史的时候。这家公司 2003 年成立，二十年里坚持干一件在软件行业看来很「轴」的事：把最顶尖的工程师派到客户现场——情报机构、战场、油田、工厂——让人家蹲在客户的数据泥潭里写代码。

华尔街很多年看不懂 Palantir，嫌它「更像咨询公司」。然后呢？2025 年，它的市值冲过了 4000 亿美元，40 法则这条健康度指标甚至好到让分析师连夜重写研报。

顺着这条线往下挖，更多线索浮出来：OpenAI 在 2024 年悄悄组建了 FDE 团队，2026 年 5 月干脆联合 19 家顶级资本，成立了一家估值百亿美元的「部署公司」；Anthropic 几乎同一天被曝出和黑石集团组建对标的合资公司。这就不难判断了：FDE 不是一阵招聘热潮，它是软件交付方式的一次改朝换代。

为什么是我来写这本书

过去十年，我一直在做一件事：把硅谷正在发生、但国内还没有名字的东西，系统地讲清楚。FDE 正卡在一个熟悉的时间点上——硅谷已经把它写进了组织架构图，而中文世

界的多数讨论，还停留在「它是不是换了个名字的售前」。

我可以先负责任地说：不是。但它到底是什么，没法用一句话讲清，必须回到它诞生的现场去讲。Palantir 怎么被情报机构的保密墙逼出这套打法，OpenAI 的工程师怎么在爱荷华州的农田里交付人工智能，一家叫 Harvey 的创业公司怎么打进连普通软件都进不去的顶级律所，以及中国企业软件行业老大难的「项目制」，跟 FDE 之间那种微妙的关系。

为了讲清楚这些，我翻遍了能找到的一手材料：Palantir 早期高管在播客里 51 分钟的复盘，几位前员工的回忆录，风投的行业分析，麻省理工那份报告的原文，几十份各家公司的招聘启事，薪酬报告，论坛上从业者的吐槽，以及国内第一批实践者的经验谈。这本书，是这些材料的沉淀，也是我的判断。

这本书怎么读

第 1 章先把这个角色的来龙去脉讲透：它从哪来，是什么，为什么是现在。第 2 到 7 章，沿着一次真实交付的完整旅程走一遍——找对问题、赢得客户、让系统被用起来、守住续约、把生意做大、让成功可复制。

第 8 章回到源头，把几个最精彩的现场完整复盘。书末聊聊这个职业的边界，附录里放了能直接用的指标清单、人物名单，以及全书所有数据和案例的出处。

你可能是三种人：1）想投身企业人工智能的创业者，你会看到一套更适合这个时代的打法；2）想转型或正在做 FDE 的工程师，你会看到这个岗位的全貌和天花板；3）或者你只是想弄明白，人工智能到底怎么才能在企业里落地——这是眼下价值万亿的真问题，这本书就是冲着它去的。

最后想说一句。写这本书的过程中，我反复想起一个朴素的道理：复杂的事情，没法隔着屏幕解决，总得有人到场。FDE 不过是把这句大实话，做成了一种职业。

愿这本书，帮你把人工智能从演示厅搬进真实世界。

范冰

2026 年 7 月

第 1 章 FDE 的崛起

> 「给间谍做软件的一个挑战是：我不认识任何间谍。」 —— 鲍勃·麦格鲁，Palantir 早期高管、OpenAI 前首席研究官

1.1 先讲一个死掉的数百万美元

故事的开头总是一样的。一家大企业的会议室里，供应商的演示刚刚结束。大模型对答如流，数据大屏流光溢彩，连最挑剔的高管都挑不出毛病。首席执行官当场拍板：签。合同金额数百万美元，双方握手，合影，发新闻稿。

九个月后，这个项目死了。

不是轰轰烈烈地死，是悄无声息地死。系统还在跑，服务器还开着，只是没有任何一个业务部门真的在用。供应商交付了合同里的每一项功能，企业付清了合同里的每一笔钱。唯一没有到货的，是「价值」。

如果你觉得这只是运气不好，麻省理工学院 NANDA 实验室 2025 年发布的《生成式人工智能的鸿沟》（The GenAI Divide）报告会告诉你：这才是常态。他们访谈了 52 家组织、收了 153 份高管问卷、翻了 300 多个公开的企业人工智能项目，结论只有一句话：烧进去的三四百亿美元里，**95% 没有产生任何可以衡量的财务回报**。

这个 95% 在引用之前，得先交代口径。报告发布后，质疑集中在三点：

1. 它把「失败」定义为六个月内没有可衡量的财务报表影响——按这个标准，互联网和云计算的早期投资几乎全是「失败」；
2. 它对价值的度量只认利润、成本、收入三项，流程提速、员工采用率这些先行指标都不算数；
3. 样本以大型企业为主，失败数据又只来自项目发起人一侧。还有评论者提醒了一层利益关系：NANDA 实验室自己就在研究智能体互联网，报告给出的解法恰好指向自家方向。（出处见附录 C）

这些质疑都成立，但都没能推翻那个方向性判断：企业级人工智能项目大面积不达预期——兰德公司、标普全球等机构的独立调研，指向的是同一幅图景。所以本书引用95%，取的是方向，不是精确值；精确到个位数的失败率，本身就是一种伪精确。

更有意思的是失败的方式。报告特意写了：问题不在模型。那些在演示里惊艳四座的模型，进了生产环境依然聪明，只是它们「不记反馈、不存上下文、不进工作流」——长得像产品，用起来却像展品。

斯坦福大学同年的大盘数据也在旁边作证：组织里 88% 都在用人工智能，可真正跑进生产环境的智能体应用，只有个位数百分比。麦肯锡补了最后一刀：能说人工智能对利润贡献超过 5% 的企业，只有 6%。

《财富》杂志报道这份报告的那几天，一位制造业首席运营官的吐槽在业内流传：「网上说一切都变了，回到我们车间，什么都没动。」（出处见附录 C）这话比任何数据都扎心——他的公司不缺预算，不缺工具，缺的是有人把工具塞进车间真实的流程里。

报告里还有一组更安静、却直接为本书主题投票的对照：企业自己关起门来建的系统，成功率只有外购成熟方案的约三分之一。报告主要作者说，他们走访的几乎每一家企业，都在试图自己造工具——而自建恰恰是死得最多的那条路。

他还讲过一个对照组的故事。一些十九、二十岁的年轻人创业公司，靠生成式人工智能一年做到两千万美元收入。打法和老牌企业正好相反：只挑一个痛点，打穿为止，并且死死绑定真正会用他们产品的客户。年纪大的公司总想一口吃成胖子，而年轻人只咬一小口，咬穿了再咬下一口。（出处见附录 C）

其实这道沟，不是人工智能时代才挖出来的。中国的企业软件行业早就在沟里躺了十几年：大公司要定制，厂商做一单赔一单，交付完代码离场作废，最后集体沦为「甲方的外包公司」。播客《硬地骇客》的一位企业软件老兵说得狠：「定制化是 SaaS 的天敌，这个诅咒只能等市场成熟那天自然消解。」美国那边体面些，剧本也差不多：销售签单，实施进场，半年后交付一个「功能齐全但没人爱用」的系统，然后是漫长的扯皮。

归根结底是同一堵墙：造软件的地方，和价值产生的地方，不在一个地方。墙这边，需求被工单、纪要、周报层层转述，每转述一次就失真一层；墙那边，客户真正的工作流藏在没人写进文档的表格里、口耳相传的惯例里、「这事得问老王」的隐性知识里。软件行业发明了无数翻城墙的梯子——需求文档、用户调研、实施方法论、客户成功体系——但墙始终在那里。

直到有一家公司决定：不翻墙了，把人送过去。

1.2 Palantir 的胜利

2003 年，硅谷刚从互联网泡沫的废墟里爬出来。彼得·蒂尔和几个斯坦福出身的年轻人，创办了一家名字取自《指环王》的公司——Palantir，真知晶球，传说里能看见远方的石头。他们要干的事听起来像科幻小说：给美国情报机构做数据分析软件，把散落在无数保密数据库里的碎片连成图景，帮分析师抓恐怖分子。

这门生意有一个能让任何产品经理当场崩溃的前提。多年后，早期高管鲍勃·麦格鲁在 YC 的 Lightcone 播客里把这段往事讲得活灵活现：

> 「我们创业时的目标，是给情报界做软件，说白了，是给间谍做软件。而给间谍做软件的一个挑战是：我不认识任何间谍，你大概也不认识。就算你碰巧找到一个间谍，问他『你平时到底怎么工作的』，他通常也不会告诉你。」

没有用户访谈，没有需求文档，没有可用性测试。互联网创业方法论的第一课，在这里全部作废。创始人之一斯蒂芬·科恩想出的办法笨得可爱：先做个演示样品，拿给情报机构的人看，问觉得怎么样。对方一点不客气：「这东西太糟了，跟我们做的事毫无关系。」科恩没撤退，追问一句：「那你们希望它哪里不一样？」然后掏出本子，一条一条记下来，回去改，改完再送上门。

这个笨拙的循环，就是 FDE 的胚胎。里面藏着两个后来被证明价值千金的直觉：一是复杂领域的客户，在看到能用的东西之前，并不知道自己要什么；二是想知道客户要什么，最快的路是让造东西的人，站到用东西的人旁边。

把这套直觉升级成公司战略的，是第 13 号员工希亚姆·桑卡尔。当 Palantir 从第一个客户走向第二个、第三个，团队发现一个反直觉的事实：每个客户要的东西，都有细微但关键的不同。标准解法是提炼共性、做通用产品、对差异说不。可 Palantir 的客户是中情局、联邦调查局、战场上的美军——说不，就等于出局。

桑卡尔反着来：做一个能灵活定制的平台，然后派工程师驻扎到客户现场，把最后一公里修完。

他最关键的动作，是改了这件事的账目。在软件行业的账本里，「为单个客户做定制」叫服务，是利润率的敌人。桑卡尔把它翻了过来：现场定制这笔账，要记在产品发现上。工程师在客户现场踩的每一个坑，都是平台下一次进化的路标。

桑卡尔本人就是第一个前线部署工程师，他最早的驻场经历，堪称这个岗位的原型现场。2007 年前后，美军在伊拉克伤亡的最大来源是路边炸弹，为此成立的反路边炸弹作战中心允许桑卡尔带着一支小队和「还很粗糙」的产品，钻进保密信息室联合办公两周。所谓保密信息室，是物理隔离的涉密空间，连免提电话都禁用。桑卡尔想了个野蛮的办法：用松紧带把电话绑在自己头上，腾出双手敲代码——一只耳朵听分析师提意见，另一只耳朵听硅谷总部的同事说话。两周里每天干十九个小时，演示、接数据、收反馈、当场改。结束时，分析师们说：这东西有用。桑卡尔自己却累垮了，打电话给首席执行官卡普：「这不可持续，我们完了。」卡普的答案日后成了公司文化：把这种「不可持续」，做成制度。（出处见附录 C）

多年以后同事回忆，桑卡尔骂人都不带情绪。有一次他和同事梅布里在机场咖啡馆聊得正开心，对方邮箱里突然收到他一封措辞严厉的批评邮件——就是坐在对面当场写完发的。「里面没有人身攻击，只有一句潜台词：为了赢，我欠你这些真话。」（出处见附录 C）

这套打法很快在战场上见了血。驻场的 Palantir 工程师发现，士兵们根本不需要什么花哨的情报图表，他们只要一个能在地图上标注「这条路可疑」的小工具——路边炸弹是巡逻队最大的杀手。工程师当场拼出一个简陋的地图工具，士兵点一下就能标出危险路段，全队实时可见。这个工具救了命，后来沉淀成平台的标准功能。它不可能诞生在任何总部的会议室里，只能诞生于工程师和士兵一起看向同一条公路的那个瞬间。

商业化那边，则是先死过一次。Palantir 第一个面向企业的产品叫 Metropolis，市场反响惨淡，只有几家金融公司勉强在用。第二次尝试 Foundry 才算开了张，转折点在空客：图卢兹工厂里，A380 的一个燃油泵故障反复发作，空客自己的工程师查了两年没头绪。Palantir 的人进场，把传感器数据接进平台，两周破案——飞机爬升时燃油晃离了泵体。一个微不足道的修复，保住了据报道价值数百亿美元的订单。空客数字化负责人后来公开感慨：「同样的问题，我们以前要查二十四个月。」（出处见附录 C）空客从此成了 Palantir 在欧洲最铁的拥趸，把自家数据平台 Skywise 整个建在上面，接入上万架飞机、五万多用户。后来的走向更有意思：这套平台从客户变成了渠道，如今有一百五十多家航空公司跑在 Skywise 上。

还有个八卦值得一记。江湖盛传 Palantir 的软件参与了 2011 年击毙本·拉登的行动。这个说法从未被证实，也从未被证伪——写 Palantir 传记报道的记者特意把这个暧昧的注脚留了下来。可不管真假，这个传言本身，就是 Palantir 最好的销售武器。（出处见附录 C）

到 2016 年之前，Palantir 前线部署工程师的人数，一度超过了平台工程师。一家软件公司，一半以上的工程师不在总部写产品，而是撒在全球的客户现场。华尔街多年看不懂，嫌它「人海战术」「更像咨询公司」。

然后，时间给出了答案。2023 年它推出人工智能平台 AIP，配上一套叫「训练营」的打法（第 8 章细讲），把企业软件九到十二个月的销售周期压到几周。2025 年第四季度，它的「40 法则」——收入增速加调整后经营利润率（非通用会计准则口径）：软件业的健康度指标，40 分算及格——干到了 127%；2026 年第一季度，145%。单季签约额 42.6 亿美元（bookings，总合同价值口径），净收入留存（Net Revenue Retention，NRR；Palantir 财报口径称 NDR）139%，账上现金 72 亿美元。

卡普在财报电话会上只说了一句：「我们是一个自成一格的物种。」市值一度冲破 4000 亿美元。（出处见附录 C）

当年笑它人海战术的人哑口无言。Palantir 用二十年证明了一件事：那堵墙，梯子翻不过去，人却翻得过去。翻墙的这批人，有了一个正式的名字——Forward Deployed Engineer，前线部署工程师。

这个名字的发明权，桑卡尔后来在《美国乐观主义者》（American Optimist）播客里公开认领了。他给的定义一点都不官方：前线部署工程师，就是「把痛苦吃进去、把产品产出来」的那种人。

当然，关于起源另有一说。有风投机构考证，把工程师嵌进客户现场这套打法，IT 服务公司比 Palantir 做得更早——这个争议先按下不表，第 7 章会带着数据回来。

1.3 什么是 FDE

一句话定义

本书采用的定义，来自这个模式最好的阐释者鲍勃·麦格鲁——他早年在 PayPal 做工程师，后来是 Palantir 早期高管，再后来是 OpenAI 首席研究官，ChatGPT、GPT-4 都出自他领导的团队：

> **前线部署工程师，是一个驻扎在客户现场、填补「产品能做的事」与「客户需要的事」之间鸿沟的工程师。**

「驻扎现场」，说的是你的工作语境嵌进客户那里：进客户的群，读客户的数据，开客户的会，认识那个「知道流程为什么是这样」的人——不一定天天坐客户办公室。「鸿沟」是这个角色存在的理由：产品开箱即用的地方，不需要你；鸿沟越深的地方，越需要你——情报、金融、制造、医疗、法律。

「工程师」是最要紧的限定词：你写的是生产环境的代码，不是报告。Palantir 命名时特意保留「软件工程师」几个字，就是向世界强调：这不是咨询岗。至于「Forward Deployed」，是军事术语，指部署在前线的部队——把最有战斗力的人，放在离问题最近的地方。

它不是什么

先看它不是什么。

它不是售前。售前的工作在签约前结束，目标是赢单，作品是幻灯片；FDE 的工作在签约后才进入深水区，目标是赢结果，作品是跑在生产环境里的系统。售前负责让客户相信「这事能成」，FDE 负责让这事真的成。

它也不是驻场外包——这个区分对中国读者尤其要紧，「工程师驻场」在国内有太长、也太不堪的历史。国内第一批打出 FDE 旗号的服务商启盟科技，在官网用三句话划清了它和驻场的界限：

1. 驻场按工时算钱，FDE 按阶段交付、按结果验收；
2. 驻场从零现写，FDE 带着产品底座来做工程；
3. 驻场越驻越久、人走系统停，FDE 做完会走，能力留在系统和客户团队里。

硅谷人工智能客服公司 Cresta 的 FDE 负责人 Jove，在一场视频对谈里把这条边界讲得更细。他的团队今年要从 30 人扩到 100 人，而他的判断是：FDE 必须绑定在一个人工智能平台上才有意义——如果只是做传统的数据对接和系统搭建，那跟传统实施工程师或外包就很难区分。他招人还有一条硬杠：智能体时代，不会写代码就像文盲。

另一个值得注意的机制是双重职责：FDE 不光要把部署做成功，还背着「让产品变得更成熟」的指标——现场学回来的东西，必须反哺平台。（出处见附录 C）

它不是咨询顾问。顾问按项目交付建议，对执行不负责；FDE 对系统的最终运转负责，终点是「客户团队能独立使用」。Anthropic 与金融科技公司 FIS 的合作是个标本：工程师嵌入 FIS 共建反洗钱智能体，把调查从几小时压到几分钟，但合作写明的目标不是交

系统，而是「转移知识，让 FIS 以后能自己建智能体」。顾问的生意建立在客户持续的需要上；FDE 干没干成，标准正好反过来——客户哪天不再需要你，才说明你干成了。

买方心里也是这么算账的，摩根大通的一位技术负责人在谈与平台厂商的共建时说得很直白：「我们要的不是更多顾问，是能造出还不存在的东西的工程师。」

它也不是传统产品工程师。产品工程师面对抽象的用户——画像（用户长什么样）、漏斗（多少人从访问走到付费）、日活跃用户数；FDE 面对具体的客户——一家银行的风控部、爱荷华州的农场、巴格达郊外的巡逻队。Palantir 官方博客的《Dev versus Delta》（开发工程师对三角洲）一文里，有对这对角色的官方划分：平台工程师负责「一种能力，服务多个客户」，前线部署工程师（内部代号「三角洲」）负责「一个客户，调动多种能力」。平台工程师追求的，是一个功能到处能用；FDE 追求的，是先把眼前这一个客户的问题彻底解决。

在 Palantir 干了近八年前线部署的纳比尔·库雷希（Nabeel Qureshi），把这条纪律讲成一句糙话：「去他的可泛化性」——先把眼前这个客户救活，能不能复用到下一家，是平台团队的事。

这股风是怎么刮起来的

一个 2003 年就发明的角色，凭什么 2025 年才成了顶流？

最直接的导火索，是生成式人工智能。大模型制造了一种前所未有的落差：任何人五分钟就能做出惊艳的演示，可把演示接进企业真实的数据、权限、合规和工作流，难出一个数量级。模型公司们陆续想明白：接下来的胜负手，在部署能力上。一家物流人工智能公司的部署负责人说得更直白：人工智能落地，很少败在模型，几乎都败在上下文。

需求侧的投票同样直白。高盛把自研的人工智能助手铺向全行，内测阶段就覆盖了上万名员工；Klarna 的人工智能客服上线第一个月，就接下了三分之二的客服会话；连大学都在行动——雪城大学把 Claude 开放给全校师生，而且要求先完成培训、再开通账号。

这阵风也不是一夜刮起来的。早在 2025 年 7 月，Semafor 就发文断言，这个「名字平平无奇的岗位」将改变人工智能行业。

数字勾勒出曲线的陡峭：据招聘平台 Indeed 的官方统计，2025 年 4 月，全美挂着 FDE 头衔的岗位只有 643 个；一年之后，变成 5,330 个（Indeed 的职位匹配口径），涨

了 729%。其他机构的口径更夸张——800%、1,165%，统计方法各异，方向却完全一致。YC 招聘板上，一百多家创业公司挂出这个三年前几乎不存在的职位。

风投机构 a16z（安德森·霍洛维茨基金）直接称它为「科技行业最热门的岗位」，还配了个传神的比喻：企业买人工智能，就像你奶奶拿到一部 iPhone——她想用，但需要你帮她设置好。

《金融时报》2025 年 11 月的一组报道，是观察这场风潮最好的切片。OpenAI 欧洲区前线部署负责人富尼耶说，团队一年前才成立，马上要扩到 50 人，「需求超出了我们的预期」；Anthropic 的应用人工智能负责人德容说得更有趣：「一家财富五百强银行的需求，和一家人工智能原生创业公司，完全是两个物种。」——所以她的团队一年扩五倍。Palantir 英国的负责人普雷特约翰把公司信条浓缩成一句：「软件只有当它对最终客户真的有意义时才有价值。」连模型公司 Cohere 的首席执行官戈麦斯也出来站台：「我们在合同一开始就嵌入工程师，等客户跑顺了再往后撤。」（出处见附录 C）

抢人的烈度还有一串硬指标。OpenAI 的前线部署职能是 2025 年 1 月由科林·贾维斯在社交媒体上官宣成立的，使命只有一句话：「帮客户把系统推进生产环境」；团队从 2 个人起步，一年涨到 52 人。把「不自己做实施」写进教科书的 Salesforce，公开承诺要招 1000 名 FDE；谷歌云一次放出 59 个岗位，首席执行官库里安亲自上网喊话，招「想站在智能体时代中心的建造者」；Box 的首席执行官莱维则公开断言，这将成为科技行业最抢手的职位之一。

Databricks 做得更彻底，把原来的专业服务部门整体改组成 FDE 组织，十二个月服务了 1,900 多个客户。连卖咨询为生的德勤，都在 2025 年 12 月成立了专门的 FDE 业务线。欧洲也没闲着——从 Mistral 到 Lovable，明星创业公司的招聘重心，正从研究人才转向部署人才。（出处见附录 C）

这股风甚至刮进了政府的视野。2025 年底，上海办了全国第一个 FDE 专题培训班，市委组织部、市经信委等六家单位联合出面。有意思的是第一期的学员构成：不是工程师，而是市属国企和重点行业主管部门的负责人——先训需求方，再训供给方。配套工程的目标同样直白：链接百家企业、打造千个智能体、带动万名开发者转型。

官方给这个岗位的定调，是「人工智能领域的特种兵」。一个岗位火到政府出面办培训班，这在软件行业的历史上并不多见。

然后是巨头们的用脚投票。2026 年 5 月 11 日，OpenAI 宣布成立「部署公司」：自己控股，联合 TPG、贝恩资本、博枫等 19 家顶级资本，初始投资超 40 亿美元，媒体披露投

前估值约 100 亿美元，顺手收购了一家有 150 名部署工程师的咨询公司。几小时后，Anthropic 被曝出与黑石集团组建对标的合资公司。两家最大的模型公司，在同一天把「部署」从成本中心升格为战略资产——资本市场用最贵的方式，给 FDE 投了票。

资本市场为什么愿意这样投票？风投机构 Foundation Capital 把账算得很开：他们估计，这波浪潮瞄准的是一个 4.6 万亿美元量级的市场——一半是企业付给销售、营销、工程等岗位的薪酬，一半是 IT 服务与外包的支出。换句话说，软件的收费对象正在从「工具预算」换成「人力预算」。用 a16z 的话说：软件不再只是帮工人干活，软件自己就是工人。

当然，也有冷静的质疑。《华尔街日报》在报道里把问题摆上了桌面：这套打法更像咨询，而不是软件；嵌入式人才太贵，能不能规模化，还是悬案。这是一个认真的质疑——本书第 7 章会正面回应它。

最深的一层原因，还是麦格鲁看得透：「人工智能智能体是个还没有霸主的品类，所以有海量的产品发现要做。」客户管理软件该长什么样，二十年前就有标准答案；智能体该长什么样，没人知道，包括客户自己。答案只能去客户现场找。麦格鲁在红杉资本的 Training Data 播客里把话说得更白：Palantir 的人工智能平台之所以值钱，恰恰因为它不是模型，而是长在「模型之外、和企业其余部分打交道」的那一层——模型越强，这一层越值钱。

科技分析师本·汤普森提供了一个更长的历史坐标：多年期、深嵌入的交付，不是 Palantir 的怪癖，而是软件业向四十年前常态的回归——他的判断是，服务与集成团队将全面回归，这一代人工智能公司必须学会从客户的顶层往下卖。

2003 年，Palantir 因为「不知道间谍怎么工作」发明了 FDE；2025 年，整个行业因为「不知道企业里的智能体该怎么工作」而集体拥抱 FDE。

1.4 FDE 的职责和特质

一份为这个岗位而生的履历

如果要回答「FDE 是一种怎样的职业」，麦格鲁的履历几乎就是标准答案。

他的第一份工作在 PayPal，早期工程师。那批人后来被称为「PayPal 黑帮」，深刻塑造了整个硅谷。离开之后他加入初创期的 Palantir，一路做到高管，亲历了 FDE 从应急之

举变成公司战略的全过程。他管产品与工程团队时提过一个著名的比喻：前线部署工程师在客户现场修出一条条通往价值的「砾石路」，产品团队负责判断哪些砾石路值得拓宽硬化，变成服务下十个客户的「高速公路」。他后来用一句话概括整个模式：把不可规模化的事，规模化地做。

再往后，他出任 OpenAI 首席研究官，领导了 ChatGPT、GPT-4 和 o1 推理模型的研发。换句话说，这个人既造过墙这边的平台，也翻过墙那边的现场，最后还亲手造出了把这堵墙垒得更高的技术本身。

有趣的一幕发生在 2025 年 YC 的一场人工智能大会上。麦格鲁原以为创业者会围着他问「怎么发明的 ChatGPT」，结果所有人追着问的都是同一个问题：Palantir 的 FDE 模式到底是怎么运作的？一个发明了 ChatGPT 的人，被追问最多的是交付方法论。

三层特质

综合二十余份各家招聘启事和从业者的现身说法，FDE 的特质可以归纳成三层。

第一层，足够宽的技术通才。FDE 不需要是某个领域最深的专家，但必须能在客户现场独立解决全栈（从界面到数据库一个人包办）问题：写得了代码，调得了接口，懂数据管道，能上云，摸得准大模型的脾气，还得懂企业环境的「水电煤」——单点登录（一次登录、处处通行）、权限、合规认证。招聘市场对这种组合有明码标价：GetPerspective 2026 年初发布的《前线部署工程师薪酬报告》显示，头部人工智能实验室的中级 FDE，年总薪酬中位数约 38.5 万美元，资深约 61 万，首席过百万——比同级多数纯研发岗位还高，因为市场知道这种人多稀缺。（出处见附录 C）

第二层，把技术翻译成业务结果的能力。这是 FDE 与普通工程师的分水岭。一位一线从业者的话被广泛引用：「模型通常是最干净的部分。难的是找到那个没人写进文档的工作流、人们真正信任的那个数据源、以及知道流程为什么是那样的那个人。」Palantir 的招聘标准说得更直白：「候选人的表达力、清晰度和沟通自如度，要让我乐于让他主持一场与客户的会议。」

面试也在筛这种翻译能力。OpenAI 和 Palantir 的 FDE 面试有个标志性环节叫「问题拆解」：抛给你一个巨大而模糊的真实企业问题，六十分钟不写一行代码，只看你怎么追问、怎么界定范围、怎么在混乱里建立秩序。面试官的忠告是：先理解问题再跳进去——慢就是顺，顺就是快。普通面试里「我把查询优化了 40%」是满分答案；FDE 面试里的满

分答案是：「我把查询优化了 40%，这让客户的分析师每天提前两小时拿到报表，团队处理量翻了三倍。」技术成就必须换算成客户语言。

第三层，主人翁意识，外加一点「叛逆」。从业者圈子里流传一句话，专门描述这个岗位要求的担当：「部署在凌晨两点挂了。你不提工单，不怪别的团队，不回去睡觉。你修好它。句号。」Palantir 对业务侧角色还有个更微妙的期待：既要有深厚的行业知识，又要敢当「叛逆者」——看得出客户现状的荒谬之处，敢推动十倍级、而不是百分之十的改良。首席执行官卡普当年定的行为标杆是「法国侍者」：嵌在服务流程里，对真实需求敏感，同时有足够的自信与品位，把客户从「他们以为自己要的」，引导到「真正对他们好的」。

一天怎么过

落到日常，FDE 的时间大致这么分：四到五成泡在客户侧写代码调系统，两三成和客户管理层对齐方向、拆问题、做架构决策，一两成把现场学到的模式沉淀回公司产品线，剩下是评估优化和知识分享——写打法手册、内部布道、培训客户团队。

这份时间表里藏着一个重要信息：FDE 不是「被外派的工程师」，而是「带着双向使命的工程师」——一头向客户交付结果，一头向公司输送情报。这也是接下来两节的主题。

1.5 一切用结果说话

如果要用一句话概括 FDE 的工作信条，那就是：一切用结果说话。Palantir 高管梅布里把它拆成两个朴素的追问：「它真的管用吗？它真的要紧吗？」——按他的说法，FDE 模式是公司「最大的秘密之一」。

先把「数据」和「结果」分清楚。企业软件史上从来不缺数据好看、结果糟糕的项目：功能清单百分之百打勾，但用的人只有百分之五；系统可用性四个九（99.99% 的时间在线），业务部门却宁愿继续用电子表格。麻省理工那份报告里 95% 的失败项目，绝大多数不缺数据报表——数字都在，而价值没来，因为没人对「财务报表上那行数字」负责。

FDE 模式从制度上保证「结果」不被稀释，具体靠三件事。

- **收钱方式向结果靠拢：** Palantir 早年做政府项目，就大量采用「做成了才付钱」的安排。麦格鲁回忆起来很直白：「早期，创业公司自己承担全部风险是合理的——做成

了你再付我们。」这套逻辑在人工智能时代演化得更精细：Sierra 按「已解决的会话」收费，不解决不收钱；不少 FDE 服务商按阶段交付、按结果验收。收费一旦与结果绑定，交付团队的全部行为都会重新排序——你不会再花三周打磨一个没人用的功能，因为「没人用」要你自己买单。平台厂商也在跟进：Databricks 在官方交付模式里，写进了按里程碑、与结果对齐的定价选项。

- **成功度量前置到开工之前：** FDE 项目的第一步不是写代码，是和客户一起定义「什么叫成」。Palantir 的训练营要求客户先锁定一个极其聚焦的核心战场——「把某条产线的排产冲突降 30%」，而不是「探索人工智能赋能制造」——就是为了防止项目在「探索」的名义下漂向不可证伪。OpenAI 与约翰迪尔的合作是教科书示范：先和农艺专家一起评审数百个真实作业案例，建起定制评估体系，然后才开始迭代模型。最终「化学品使用减少最高 70%」这个数字，不是事后包装的宣传口径，是开工前就定好的靶子。
- **最终裁判是客户组织的行为改变：** 那份报告里有个辛辣的发现：只有约四成企业为员工提供官方的人工智能工具订阅，而多达九成员工，日常在用个人消费级产品解决工作问题。这意味着大量「成功上线」的项目，实际处于「官方系统空转、员工绕道而行」的状态。FDE 哲学里，里程碑不在系统上线那天，在客户团队改变工作方式那天。Sierra 内部刻意把岗位命名为「智能体工程师」，负责人默勒解释选材标准时说：只接两类问题——真的难的，和真的有业务影响的，二者必须同时成立。

「一切用结果说话」听起来是常识，执行起来却是对整个利益结构的冒犯：1）销售不敢过度承诺了，因为交付团队要为结果负责；2）客户信息部门没法用「功能清单」交差了，因为业务部门的使用率成了验收标准；3）FDE 自己也没法用「我按需求做完了」免责，因为需求本身的对错也算他的账。这个角色为什么贵，又为什么值得贵，原因都在这里。

1.6 FDE 在团队里的四张面孔

一个 FDE 同时活在四个世界里，他是这四个世界的连接件。

- **对客户，他是「嵌入式产品经理 + 全栈工程师」：** 既像人类学家一样观察客户的真实工作——最有价值的发现往往来自「看」，不是「问」——又像创业者一样在观察现场直接动手做出来。Palantir 把这个双人组合制度化：部署战略师（Deployment

Strategist，内部代号「回声」）负责读懂客户的使命、各相关方和采纳路径，前线部署工程师（内部代号「三角洲」）负责技术实现。两人一组，一个诊断，一个建造，缺一不可。

- **对公司产品线，他是「前哨与情报官」：** 这是 FDE 与传统交付团队最本质的区别。传统实施的成本是销售成本，花出去的每个人天（按人头按天计费）都要从合同里挣回来；健康的 FDE 组织把现场工作当研发——三个客户撞上同一个集成缺口，那不是三桩麻烦，那是一条产品情报；五个部署都需要同一种工作流，那就该抽象成平台的下一个标准能力。前 Palantir 工程师 Barry 回忆，Foundry 平台的关键组件诞生于苏黎世、休斯顿、圣保罗、图卢兹等天南海北的客户现场，自下而上长成，最后反哺为年收入数十亿美元的产品。（出处见附录 C）
- **对销售：** 企业客户被辜负过太多次，对一切幻灯片免疫。FDE 用两个动作重建信任：一是动手，在客户自己的数据上、客户自己的环境里，当场做出能跑的东西；二是诚实，敢对客户的错误前提说不。Palantir 的训练营把这种信任生产流程化了：客户带真实数据来，一到五天做出能部署的原型，高管亲手点着用。早期训练营的付费转化率只有 5% 到 10%，而公司披露的后期转化率已接近 75%。（出处见附录 C）
- **对组织本身：** Palantir 走出了密度惊人的创业者群体——Lenny's Podcast 的访谈给出过一个数字，将近三分之一的产品经理，离职后自己创了业。这不奇怪——FDE 的日常训练，就是在资源受限、需求模糊、各方关系复杂的环境里，端到端把一个有价值的东西做出来并让人用起来，几乎就是创始人训练的完整预演。后来创办 Decagon 的斯里尼瓦斯、组建 Sierra 智能体工程团队的默勒、写下这个行业流传最广的方法论文章的几位作者，都是从 Palantir 的前线部署岗位走出来的。一家公司的人才外溢，变成了一整个行业的人才基础设施。（出处见附录 C）

这四重身份合起来，是一件事：组织了解客户的方式变了——从层层转述的二手信息，变成工程师在现场亲手摸到的一手经验。

1.7 如何招聘 FDE

先泼一盆冷水：FDE 是软件行业最难招的岗位之一，因为它要求一个人在两个通常此消彼长的维度上同时优秀。

前 Palantir 工程师 Barry 在回忆文章里把这件事说透了：Palantir 招 FDE 的标准，是「能进谷歌或脸书的工程师」——因为他们是去客户现场建造系统的，不是去调参数的；但光有技术远远不够，前线部署的人还需要创造力、判断力和面对客户的魅力。他补了一句扎心的：这比招一支传统的售前团队，昂贵和困难得多。

拆解招聘市场的实践，FDE 招聘有三个关键环节。

- **候选人画像：** 招「好奇的推土机」，不招「精致的工匠」。a16z 给创业公司的建议用了「充满好奇心的实干家」这个词：主观能动性强、对现状缺乏敬意、对客户的问题有饥饿感。麦格鲁说得更具体：FDE 团队要两种人——「领域叛逆者」：懂行业但不迷信行业惯例；「原型快手」：速度优先于完美，接受第一版要扔掉重写。反过来，两类在传统工程文化里受宠的人，反而是 FDE 岗位的危险信号：把代码优雅置于客户结果之上的「工匠」，和把客户每句话当圣旨的「忠诚执行者」。Palantir 早年干脆把筛选器装在漏斗最外面：高调拥抱国防业务、薪水低于市场、工时长得吓人——库雷希管这叫「蝙蝠信号」，只对同类有效。
- **面试：** 用「问题拆解」代替八股。前面提过，这个环节是 FDE 面试的灵魂：给候选人一个模糊、庞大、带着真实业务毛边的问题——「某银行合规团队每天人工核对三万条交易告警，九成是虚惊，你怎么办」——然后观察六十分钟。考察的不是答案，是过程：先问清约束再动手没有，区分根因与症状没有，记不记得系统另一端坐着一个真实用户，能不能清楚地讲出取舍。Palantir 还会在每个技术轮里嵌入约二十分钟的行为问题，并且明确会拒掉技术很强但文化不合的候选人——文化项里最重要的一条，是面对模糊时的秩序感。
- **薪酬结构：** 接受「工程师薪酬 + 与公司经营指标挂钩的浮动奖金」的混合结构。2026 年初的市场数据可以作锚点：Palantir 前线部署工程师年总薪酬中位数约 21.5 万美元；头部人工智能实验室中级 FDE 约 38.5 万，资深约 61 万；Anthropic 的岗位底薪在 20 到 30 万美元之间。另一个值得注意的细节是奖金设计：Palantir 的奖金常与客户扩展等经营指标挂钩，介于工程奖金与销售佣金之间；a16z 的建议则是激励与客户经理对齐，但别让 FDE 背硬性销售指标——那会把行为引向签单，而不是结果。一位扒了五十份招聘启事的作者，给出一块更简单的试金石：看这个岗位的薪酬包里有没有销售提成——提成占比高的，多半是售前换了个时髦头衔。薪酬结构也是角色定义的一部分：给钱的方式，会影响人的行为。（出处见附录 C）

1.8 如何成为 FDE

换个视角：如果你是一个工程师、产品经理或顾问，想进入这个高速增长的市场，路怎么走?

先自测：这个岗位的光鲜与代价，是一体两面。论坛从业者社区里，对 FDE 的讨论有一种少见的诚实。正面的部分：技术含量和品牌背书的最佳组合，少数能同时积累技术、商业和客户资源的岗位。代价的部分：四分之一到一半的出差是常态，OpenAI 的招聘启事里明确写着出差最高可达 50%；工作节奏被客户的紧急程度、而不是自己的排期定义；还有一个反复出现的提醒——职业倦怠的风险是真实的。

有一条评论：「有人把它当品牌跳板，有人说它是挂着酷头衔的咨询，两种说法都对——区别在于，你所在的公司是把现场学习回流到产品，还是把你当人天在卖。」这句话既是择业标准，也是本书第 7 章的主题。(出处见附录 C)

再说国内的钱。前面那些美元薪酬，对多数中国读者只是背景音；2026 年上半年的国内招聘市场，已经给这个岗位标出了人民币价格：字节跳动给豆包的前线部署工程师开到 3.5 万到 7 万月薪、15 薪；蚂蚁数科 4 万到 6 万、15 薪；智谱的 FDE 负责人 6 万到 8 万；腾讯云 3.5 万到 6.5 万（这两家未披露薪数)；阿里云 2 万到 5 万、16 薪。折算成年包，这批岗位大致落在 30 万到 100 万元出头——「年薪百万」在头部岗位的高端够得着，但不是行业平均。(出处见附录 C)

同一头衔底下的价差，比层级差更值得看。一家唐山的工业软件公司招「人工智能交付工程师（FDE)」，要求硕士、驻场制造业，月薪 8000 到 1.6 万——和智谱的价差，最大相差近十倍。价差背后，是平台底座、客户质量、计价模式的差。所以看国内机会，别先看名片上有没有 FDE 三个字母，用本书第 8 章那把尺子量：这家公司按结果收钱，还是按人头收钱。

再练翻译：多数工程师的技术底子够用了，但稀缺的是「翻译」这门手艺——具体说是三种：把业务问题翻译成技术问题（第 2 章)，把技术方案翻译成高管能懂的话（第 3 章)，把现场经验翻译成团队能复用的知识（第 7 章)。练这三样，上课不如上场：跟一次售前、值一次驻场、给一个真实用户做培训，然后看自己在哪种不适里成长最快。

面试准备上，把履历改写成「客户结果导向」。原则前面说过，再强调一次：简历里的每个技术成就，都要走完到客户语言的最后一公里。「做了检索增强（RAG，让模型先查资料再回答）系统」是工程师语言；「做的检索增强系统让客服首次响应从 4 小时降到 8 分

钟，续约时客户主动提出扩容」是 FDE 语言。同时备好两类故事：一次你在需求不清时建立秩序的经历，和一次诚实的失败——Palantir 系的面试官对「讲一个真实的失败」有执念，因为这个工作的本质就是在不确定中前进，不承认失败的人，没有进化能力。

选公司时，反问三个问题。一是「你们的产品平台是什么」——没有平台底座的 FDE，是纯人力外包。二是「现场学习怎么回流产品」——请对方讲一个最近从现场沉淀为产品功能的实例，讲不出来就有问题。三是「FDE 向谁汇报」——向产品或工程线汇报，通常意味着模式被认真对待；向销售线汇报，则要小心沦为售前的人力池。

上面这些，都是写给工程师的。不写代码的人呢？位置有，而且 Palantir 早就把它留好了：1.6 节那个双人组合里，代号「回声」的部署战略师，干的就不是写代码的活——读懂客户的使命、各相关方和采纳路径。沿着这条线往外看，第 4 章的变革管理、第 5 章的「培训培训师」机制、得物那种钻进业务团队搬运隐性经验的知识运营小组、第 3 章的技术内容营销，都是非工程背景的入场口。这些角色的共同要求，恰好是运营、咨询、内容出身的人的老本行：把人的事理顺。

最后一个问题，只能诚实回答：这条路能走多远？没人走完过。这个岗位太新，完整周期还没有样本，能看清的只有结构。消耗端是真实的——前面说的出差强度、被客户紧急程度定义的节奏，都会随年龄和家庭阶段变重。增值端也是真实的——客户信任、行业判断、跨组织的威望，恰恰都随工龄复利。

出口隐约有三条：1）回流产品线，把现场熬成平台的人是产品负责人的天然候选；2）自己创业，Palantir 出来的产品经理，将近三分之一离职后自己创了业；3）去甲方，读完第 8 章那些国企故事就知道，甲方最缺的恰恰是懂乙方套路的人。哪条成立，五年后才有答案。现在能确定的只有一句：这个职业给「熬」定价，但前提是你在一家把现场学习当资产的公司——否则熬出来的只是工龄。

1.9 FDE 的常用工具箱

最后，给出这个岗位当前的工具全景。工具会过时，但工具背后的能力分层不会。五层，从脚下到身后。

- **平台底座层。** FDE 模式成立的前提是「带着平台去现场」，否则就退化为定制开发。Palantir 的 Foundry 与 AIP，核心是那个叫「本体」的东西——把企业的数据、逻辑和动作建模成一套语义层，让人工智能在「懂业务」的地基上运行；OpenAI 的模型

接口与智能体工具链；Sierra 的智能体平台。评估任何 FDE 机会时，这一层的厚度是第一优先级。

- **人工智能工程层。** 2025 年后的日常手艺包括：提示词工程与上下文管理；检索增强（RAG）；评估体系：为模糊的业务质量建立可量化的标尺——这是人工智能时代 FDE 区别于传统实施工程师的标志性技能；智能体架构：工具调用、多智能体协作、关键环节留人把关；以及成本与速度的工程优化。
- **数据与集成层。** 几乎所有 FDE 项目的第一周都在和这一层搏斗：数据管道、企业系统连接器、权限与身份认证、给人工智能查资料用的向量数据库、数据治理与脱敏（隐去敏感信息）。库雷希还有一条经验之谈：企业数据问题里，九成五是接入、清洗、关联，根本不是分析——项目进度的七成卡在这一层，但演示里看不见它。
- **交付与协作层。** 在客户的安全边界内工作，意味着「双重适配」：既要会用自己的现代工具链，也要能屈身于客户的环境——可能是与互联网物理隔离的内网，可能只能在客户的云环境里部署，可能连代码托管网站都访问不了。容器化（把环境打包带走的技术）、基础设施即代码（用代码管理服务器环境）、以及「在断网的会议室里也能把环境跑起来」的应急预案，都属于这一层。
- **知识沉淀层。** 这是最容易被忽视、但决定团队能否摆脱「收入随人头线性增长」的一层：打法手册、组件库、部署检查清单、以及把「某个客户的解法」改写为「一类客户的模式」的写作习惯。第 7 章会专门展开它。

五层工具箱合起来，就是这个岗位的完整轮廓：以平台为底座，以工程为手艺，对客户的结果负责，同时连着公司的产品线。

接下来的七章，我们进入方法论的腹地，从一个项目最原点的选择开始——怎么确保你在解决正确的问题。

第 2 章 解决正确的问题

> 「在错误的问题上，一切执行力都是浪费。」

2.1 概念验证坟墓的解剖报告

硅谷有个词，叫「概念验证坟墓」（PoC purgatory——PoC 即 Proof of Concept，概念验证；purgatory 直译「炼狱」）。进去的项目出不来：没死透，不能宣告失败；没活成，不敢加大投入。于是它们在季度汇报里一年一年地「持续推进中」，像一屋子插着管子的病人。

麻省理工学院 NANDA 实验室的研究者在《生成式人工智能的鸿沟》报告里，给这座坟墓做过一次系统解剖。他们归纳出企业人工智能项目规模化路上的五大路障，按出现频率排下来：

1. 员工不愿用新工具——而这些人每天私下用 ChatGPT 用得飞起
2. 对模型输出质量的担忧
3. 糟糕的用户体验
4. 缺乏高管支持
5. 变革管理困难

这份清单里缺席的东西更值得看：模型不够聪明、算力不够便宜、技术不够先进——都不在列。杀死这些项目的，几乎全部发生在「问题定义」和「组织现实」层面，不是技术层面。

一家企业花五万美元采购了专业的合同分析工具，功能清单洋洋洒洒。但公司的一位资深律师就是不用——她继续用免费的 ChatGPT 起草合同。理由很朴素：买的那个工具，摘要太死板，没法按她的习惯定制。采购部门的报表上写着「已部署」，真实的日常却是官方系统空转、员工绕道而行。（出处见附录 C）

这个细节揭示了第一大路障的真正成因：不是员工守旧，是消费级产品把他们的胃口养刁了——每天在家用得飞起的人，没法忍受办公室里的「人工智障」。

还有一组更扎心的对比：**与外部专业供应商合作的项目，成功率约为内部自建的三倍**。为什么外部团队反而胜率高？他们输不起——按结果收钱的人，定义错问题的代价是自己买单。

日本一家公司的内部项目就是典型的反面教材：抽调三四名最强工程师组成攻坚组，演示惊艳、领导点头，唯独从第一天起就没有人能回答「这套系统的好，由什么标准衡量」。半年后项目从汇报材料里悄悄消失，没有人宣布它失败，它就这样自然死亡了。（出处见附录 C）

这就是第一性原理：企业里的错误问题，比想象中多得多。所以写第一行代码之前，先得确保你在解决正确的问题。

2.2 PSF：找到问题与方案的契合

互联网创业方法论里有个核心概念叫 PMF（Product-Market Fit，产品与市场的契合）：产品对了，市场自己会拉动增长。在 FDE 的世界里，对应的单位不是「产品与市场」，而是「问题与方案」——我称之为 PSF（Problem-Solution Fit，问题与方案的契合）。

区别微妙而关键。PMF 问「我的产品有没有市场要」，视角在供给方；PSF 问「客户这个具体的问题，值不值得、能不能被我们的能力解决」，视角在需求方。一个企业客户内部，可能有几百个「人工智能能做点什么」的机会点，但真正值得做的，必须同时过三关。

第一关，痛点检验：这个问题，是不是某个具体的人的具体的痛？注意是两个「具体」。「提升客服效率」不是痛点，是方向；「客服主管每周一早上要花三小时，从四个系统里手动汇总上周的升级工单，而她真正的工作应该是分析升级原因」——这才是痛点。

麦格鲁给过一个更锋利的标准：去解决首席执行官最关注的五个问题之一。理由很现实，只有这个量级的问题，才能帮你碾过企业内部的官僚阻力。Sierra 的智能体工程负责人默勒的标准是同一个原则的另一种说法：只接两类问题——真的难的，和真的有业务影响的。只有难没有影响，是炫技；只有影响不难，轮不到你。

第二关，经济性检验：解决这个问题，值多少钱？很多痛点是真痛，但不值钱；算不清这笔账的项目，做了也活不过下一个预算季。

你可以粗略地算：这个问题每周吃掉多少人时？折合多少人力成本？出错一次赔多少？省下来的人手能去干什么？

Palantir 的训练营干脆把这关前置——客户开始前就得锁定「核心战场」、给出量化靶子，「排产冲突降低 30%」「库存周转天数缩短 15%」。NANDA 的报告里有个被广泛引用的发现，恰好说明这关多容易被跳过：超过半数的企业人工智能预算投向前台的销售营销，回报却集中在后台的合同审查、采购、风控这些「不性感」的地方——大家都在解决「演示效果好的问题」，而不是「值钱的问题」。

第三关，可行性检验：以我们现在的能力，和这个客户的数据现实，能做到几分？这一关最容易被热情淹没。有两个必须现场回答的问题：数据在哪、什么状态？答案经常比想象糟——分散在七个系统里、三个版本对不上、最权威的那份在某个老员工私人的表格里。

OpenAI 的 FDE 负责人科林·贾维斯用一句话概括过现场经验：客户在选题阶段描述的问题，常常和地面上真实的数据与系统状况对不上。

第二个必须现场回答的问题：准确率门槛是多少？「99% 可用」和「90% 可用」之间隔着数量级的工程投入，而很多业务场景其实 90% 加人工复核就是最优解——判断这一点需要的不是技术，是对业务后果的理解。

「能做到几分」不是猜出来的，是测出来的——所以老练的团队把「先建考题」放在「上线」之前。摩根士丹利就是这么干的：从三个具体场景起步，让理财顾问和工程师逐条给模型输出打分，分数直接回流到迭代里，此后每天把旧考题重跑一遍——防止模型悄悄退步——在质量下滑抵达一线顾问之前拦截。

OpenAI 官方面向企业的部署指南列了七条经验，第一条同样是「从测试题库开始」。为什么要每天跑？因为人工智能系统出错时往往不报错，它只是悄悄变得不靠谱——等用户自己察觉，信任的账单已经到期了。

三关都过，才算摸到问题与方案的契合。而这三关必须在客户现场过——前面那些错误，大多是从总部会议室里对着二手信息做判断时长出来的。

而且这三关不是进场前过一遍就一劳永逸，这套检验得变成一种日常纪律。企业支出管理公司 Ramp 给自家 FDE 定的第一条原则就叫「永远在选题」：不能对客户的要求照单全收，每接一个需求，先收集上下文、验证假设、评估对整体排期的影响。这条纪律是他们交过学费换来的——曾经为一个客户做了几周的安卓端功能，交付前才发现这家公司内部强制只用 iPhone，几周工作全部作废。选题纪律松懈的代价，以周计算。

2.3 拒绝昂贵的「概念验证坟墓」

2025 年，聚焦不动产与设施管理行业的中国服务商启盟科技，在官网上写了一段颇为狠辣的话，劝退潜在客户：场景还没验证的，先去参加演示活动；数据一张表就能导出的，轻量服务就够了；只是想了解人工智能的，用免费的方式。最后一句最狠：「FDE 是重投入，我们宁可你晚一点开始，也不希望你在错的时机开始。」（出处见附录 C）

这段话说出了一个反销售直觉的事实：拒绝错误的项目，本身就是 FDE 模式的盈利能力。

错误的项目为什么危险？因为 FDE 的成本结构是前置的——最优秀的工程师、最贵的差旅、最长的现场投入，全部发生在回款之前。一旦陷入泥潭，不是亏一单的问题，是整支精英团队被拖住、机会成本雪崩的问题。

前 Palantir 工程师 Barry 回忆：「我们烧掉数百万美元做客户试点，很多项目利润率字面是负无穷，因为我们是免费做的。」他紧接着补充的视角才是重点：Palantir 烧得起，是因为它把试点当研发组合来投资——像风险投资一样，多数下注归零没关系，但成的那些要赢回一切。可如果你既没有它的资本厚度，又没有「把失败试点转化为产品资产」的机制，那么每一个错误的试点，都是纯失血。（出处见附录 C）

所以 FDE 团队需要一套「拒绝的机制」，而不只是拒绝的勇气。三条可操作的防线。

- **防线一：概念验证必须有「毕业标准」**。每个验证项目启动时就写明：多少周后、用什么指标、达到什么数值，项目「毕业」进入付费部署；达不到，双方体面散伙。Palantir 的训练营把这套逻辑做到了极致——它不是概念验证，是概念验证的工业化替代品：一到五天、客户带真实数据、现场做出能部署的原型、高管当场拍板。要么几天内见真东西，要么不要开始。传统概念验证之所以沦为坟墓，恰恰因为它「无限期、无指标、无裁判」。
- **防线二：** 警惕三类高危信号。综合从业者的经验，三类信号出现两个以上就要高度警惕。一是「无主之地」——项目在公司内部没有明确的业务方负责人，只有信息部门在对接。信息部门关心的是合规与稳定，而合规与稳定从来不是做新项目的理由。二是「只许看、不给碰」——客户要求你先证明能力，但拒绝提供真实数据。如果没有真实数据的验证，做出的成功就只是假象。三是「宇宙级需求」——第一次会议就要「覆盖全公司所有场景」的客户，往往还没准备好做任何一个场景。

- **防线三：** 给「拒绝」留个体面的台阶。拒绝不等于断交。最好的做法是把「现在不做」翻译成「什么时候做」：「这个场景的数据基础还差三件事，我们建议先做另一个场景，顺手把这三件事补齐，下个季度再来。」把拒绝包装成路线图，既守住产能，又维护关系——第 3 章的「灯塔客户筛选」会沿着这条线继续。

这三条防线是写给乙方的筛子。翻过来用，就是甲方的自检表——买 FDE 服务的人，比卖 FDE 服务的人多得多。

如果你的项目在乙方眼里亮起「无主之地」信号——集团发文压下来、信息部门牵头、业务部门无人署名——老练的乙方会警惕，而你应该比他更警惕：连供应商都看得出这个项目没有真正的主人。如果你的采购流程天然制造「只许看、不给碰」——安全审查要求对方先证明能力，真实数据却一份也不给——那你买到的验证只能是假象。

如果你写进招标书的需求是宇宙级的——「人工智能赋能全集团」——你吓跑的是最懂行的供应商，吸引来的是最敢吹牛的。

2.4 痛点，催生部署的第一原动力

正确的问题从哪来？FDE 的现场经验是：从痛来。而痛不会出现在会议室里，只会出现在工作现场。

这套方法，Palantir 二十年前就用血写出来了。回看 1.2 节那个伊拉克战场的故事：士兵需要路边炸弹预警工具——这个需求在任何访谈里都问不出来，因为士兵不知道「可以向软件要这个」，他们以为这就是巡逻生活的一部分。这种痛点用户自己说不出来，要靠工程师坐在旁边亲眼看见。驻场工程师跟着巡逻队出任务，亲眼看到车队在可疑路段前的犹豫与恐惧，才有了那个改变战场的简陋地图工具。

这个方法在人类学里有名字，叫「参与式观察」；在丰田生产方式里叫「现地现物」——到现场去，看实物，得实情。FDE 把它变成了一套可操作的田野方法，我称之为「影子工作法」。

跟着真实用户，过完他真实的一天。不是采访他，是坐在他旁边看他工作——看他打开哪些系统、在哪些电子表格之间复制粘贴、在哪些环节皱眉头、绕过哪些「官方流程」。

OpenAI 的 FDE 团队在约翰迪尔项目里就是这么做的：飞到爱荷华州，跟着农艺专家和农场主下地，看他们如何做施药决策、看哪些信息真正进入决策、看农时这道硬期限如何支配一切。他们要打掉的那个旧流程，原本的形态是农艺师挨家挨户打电话给农户、口

头给设备使用建议——这种工作流不会出现在任何文档里，只有下地才看得见。那句被反复引用的从业者箴言，说的就是这套方法的发现对象：「难的是找到那个没人写进文档的工作流、人们真正信任的那个数据源、以及知道流程为什么是那样的那个人。」这三样东西，每一样都只能在现场找到。

下到地里还确认了一件会议室里永远说不清的事：农户要的从来不是「人工智能」，是少打药、多打粮——后来长出来的产品，干脆按实际启用技术的英亩数收费，客户付的每一分钱都和他得到的价值对齐。这套计价逻辑搬到中国，对应的变形是按启用的产线、门店或网点数收费——在按项目制验收的采购习惯里，它是「按结果收费」最容易被采购科接受的第一步。

重点观察「变通」，而不是「流程」。官方流程图告诉你组织「应该怎么运转」，变通告诉你组织「实际怎么运转」。

员工为什么坚持把数据导出到表格里再算一遍？为什么部门里公认「这张表要找小王」？为什么明明有数据系统，决策会前总有人手动核对数字？每一个变通背后，都是一个未被满足的痛点、一次系统的失败——也都是 FDE 的机会。

警惕「翻译过的痛点」。你听到的需求，如果是经客户信息部门、采购部门、咨询顾问转述的，每转一手就失真一次——信息部门会把业务痛点翻译成技术需求（「需要一个数据平台」），采购部门会把它翻译成合规条目（「需要满足某某标准」）。

传统软件项目的灾难往往就从这里开始：厂商对翻译件负责，而不是对痛点本身负责。所以你的第一要务，是绕过翻译件，直达疼痛的神经末梢。这也是 Palantir 双人模型里「回声」的核心职责——理解客户的「使命」而不是「需求」：写在需求文档里的，是痛点被转述过的版本；客户的使命，才是痛点本来的出处。

找到了真痛点，下一步是用最小的代价验证：我们的方案，真能止这个痛吗？

2.5 用「最小可行部署」验证价值

互联网创业方法论里有个著名概念叫 MVP（Minimum Viable Product，最小可行产品）：用最小的产品验证市场的需求。FDE 对应的概念，我称之为 **MVD（Minimum Viable Deployment，最小可行部署）**——用最小的工程投入，在客户的真实环境里，对真实的痛点，验证一次价值的真实发生。

一字之差，差在裁判。MVP 验证「要不要做这个产品」，裁判是市场；MVD 验证「这个方案在这个客户身上能不能产生价值」，裁判是这个具体客户的具体业务。

一个在十个客户那里验证成功的方案，在第十一个客户那里照样可能失败——数据基础不同、组织惯性不同、痛点的形状不同。这就是企业交付的残酷之处：价值没法从上十个客户那里继承，只能在这一个客户身上重新验证。

MVD 有三条军规。

- **第一条：** 真实数据，没有例外。用客户提供的「脱敏（隐去敏感信息）的样例数据」或自己构造的演示数据做验证，是概念验证坟墓的第一块砖。真实数据里藏着一切魔鬼：字段含义与文档不符、三成的空值、三年前的编码规则，以及最要命的——数据本身记录着错误的流程。还有一种更隐蔽的魔鬼：数据里并存着两个互相矛盾、各自「正确」的答案。Salesforce 把自家客服智能体当第一客户用了一年，踩过的最大的坑就是这个——智能体遇到两个打架的答案时会试图调和、甚至编造，一份过时又没被链接的旧页面就足以污染回答。逼得他们回头把公司内部六百多条数据流整合成唯一权威数据源，每个问题只留一个权威答案。
- **落地只有一条硬杠：客户必须带自己的真实业务数据来。** Palantir 的训练营把这条写进规则，约翰迪尔项目把「评审数百个真实作业案例」放在建模之前。OpenAI 内部干脆按这条线划分了两个岗位：解决方案架构师可以用匿名的样例数据做演示和验证，FDE 则必须在客户的基础设施上、用客户的真实数据写生产代码。在假数据上成立的方案，上线那天就会遇到文档里不存在的字段——那时候已经付了真价钱。
- **第二条：** 缩小的是范围，不是价值。常见错误是把 MVD 理解成「阉割版的大方案」——功能砍掉七成，做得四不像。正确的做法是不砍方案的深度，只砍覆盖的面：不追求「覆盖全公司的智能客服」，而是「只覆盖退换货这一类工单，但做到端到端、无人干预」；不追求「全集团的供应链优化」，而是「只做这条产线的排产冲突，但每周实实在在省下 20 个人时」。切口小到价值密度足够高，高到业务部门肉眼可见、主动传播。法律人工智能公司 Harvey 的扩张路径就是这个打法的教科书：不做全所铺开，先打透一个全球业务组，让第一批合伙人成为信徒，六个月实战后再横向扩展。
- **第三条：** 定死截止时间，倒逼取舍。MVD 的验证周期应该以「周」计，不是以「月」计。Palantir 的训练营一到五天，Sierra 公开报道的最快上线案例四周，Decagon 的典型部署四到八周。截止时间的意义不在快，而在强迫双方做诚实的取

> 舍：凡不能在这几周里体现价值的部分，都还不是核心价值。一个六个月的「最小验证」，几乎必然重新长成一个什么都想要的大项目——那就是概念验证坟墓的又一次开工。

这种速度在客户眼里值多少钱？美国坦帕综合医院的首席数字与创新官斯科特·阿诺德（Scott Arnold）在一场行业大会上说得很直白：「我们能用几小时、几天解决问题，而不是几个月、几年。」血液供应商 OneBlood 遭网络攻击时，这家医院用 Palantir 的平台在几个小时内搭出一个血液库存调配应用，随后交给佛罗里达州供其他医院复用。他同时承认 Palantir 的价格有溢价，「但这溢价值得」——周级的节奏，买的从来不只是快，是出事当晚手里就有工具的确定性。

训练营：MVD 的工业化

Palantir 在 2023 年推出的 AIP 训练营（AIP Bootcamp），是目前唯一被大规模验证过的 MVD 流水线。

先看数据：从 2022 年的不足百场试点起步，场次连年倍增，媒体追踪显示累计完成场次早已过千，2025 年高峰期平均每天近 6 场；企业软件传统的销售周期（九到十二个月）被压到数周；美国商业收入 2025 年第四季度同比增长 137%，公司公开把增长几乎全部归因于此。（出处见附录 C）

再看流程，它把 MVD 拆成五个标准化动作。第 0 天，筹备：双方锁定一个极其聚焦的核心战场——「优化某条产线排产」「降低库存周转天数」，拒绝一切宏大叙事。第 1 天，接入：打通客户的现有系统，把孤立数据抽出来，构建初步的本体模型。

第 2 到 3 天，构建：FDE 与客户技术人员背靠背写代码、配规则，把大模型接入业务流，做出能执行真实动作的自动化工作流。第 4 到 5 天，演示与拍板：产出的不是报告，是活的软件界面，业务高管亲手点击，看着人工智能基于自己公司的数据给出建议——震撼过后，直接进入商务谈判。

训练营的精妙，在于它同时解决了 MVD 的三个经典难题：真实数据问题（客户自己带来）、期限问题（五天封顶）、裁判问题（高管亲手用）。它还顺手解决了一个更深的问题——信任。让决策者亲手操作基于自己数据的系统，可行性报告就可以省掉大半。

训练营之后的合同，签得有多快？Palantir 在财报电话会上披露过一组真实节奏，快到同行都未必信：一家大型医疗公司，12 月参加训练营，五周后签下五年期、年合同额

（ACV，年度合同价值）2600 万美元的协议；一家全球银行，试点一个月后先签 200 万美元初始合同，四个月后扩展为三年期、年合同额 1900 万美元；连锁药房 Walgreens 先在 10 家门店试点，店内运营效率提升 30%，然后八个月内推到 4000 家门店，人工智能驱动的端到端工作流每天自动处理原本要人工做出的约 3840 亿次决策。（出处见附录 C）这组数字回答了「MVD 之后会发生什么」：验证通过的项目不是慢慢长大，是跳跃式放大——客户在五天里已经亲眼见过价值，剩下的只是商务流程。

这套「先十家、后四千」的节奏，在中国企业语境里的对应物，是先拿下一个事业部、一个厂区或一个区域公司，用使用数据去敲集团的门。

当然，这套模式的复制有门槛：背后得有成熟的平台底座，否则五天连环境都搭不起来。对没有平台的团队，可执行的简化版是「两周冲刺验证」：第一周进场、接数据、定指标，第二周做出一个只解决单点问题、但能跑真实业务的原型，周末向业务方演示并当场决定进退。形式可以裁剪，军规却不能裁剪。

两周的日程可以拆得很具体。第一周：第 1 天进场对齐，和客户一起把「单点问题」写成一句话、把验收指标写成一个数；第 2 到 3 天接数据，只接这个问题需要的最小数据集，权限、脱敏、导出方式当场定死；第 4 到 5 天搭出能查数、能回答的骨架。

第二周：第 6 到 8 天把原型接进真实业务流，让一两个真实用户开始用；第 9 天收集使用痕迹和问题清单；第 10 天演示——不是给信息部门演示功能，是给业务方演示「你那个问题，现在这样了」，然后当场决定：放大、调整，还是散伙。

如果没有平台底座，工具链就用手头现成的凑：模型调商用接口，或拿开源模型做私有化部署；知识检索用开源的向量数据库；流程编排用开源框架；界面用最轻的前端模板。这套「散装」组合足以撑过一次单点验证。唯一要警惕的，是把冲刺开成技术选型大会——验证期选型的唯一标准是快；能不能复用、能不能扩展，是验证通过之后才配考虑的问题。

客户侧也有三项配合要求：1）一个能拍板的业务方负责人在场；2）一个懂数据的内部接口人全程陪同；3）真实数据的访问权限第 1 天就位，而不是「正在走流程」。三条只要缺一条，两周冲刺就大概率滑回传统的概念验证。

2.6 早期要不要迁就客户的现有环境

MVD 阶段有一个几乎每个项目都会撞上的分歧：客户的现有技术环境——那堆跑了二十年的老系统、部门自建的小工具、还有严格的安全合规边界——我们的方案应该在多大程度上迁就它?

两派都有道理。「迁就派」说：在客户的真实约束里验证，才是真实的验证。「重构派」说：为一个即将被替换的旧环境做深度适配，是把宝贵的验证期浪费在注定要扔掉的工程上。

FDE 的实践给出一条中间路线：数据上兼容旧系统，而架构上绝不迁就旧系统——我称之为「读旧写新」。

数据层面，深度兼容旧环境。客户的数据在哪里，就从哪里读——哪怕它在一台老式主机里、在共享盘的表格里、在某个古老系统的私有接口里。金融业的老式大型机、医疗行业数百家诊所的异构系统，历来是部署工作的核心战场。读数据的兼容没有捷径，因为数据是验证价值的前提，而数据永远不会为了迁就你的架构而搬家。

好消息是，这一层工作正在被人工智能本身改变：过去需要人工解读的字段映射、跨系统搬运、没有接口的老系统取数，现在可以大量交给智能体完成——比如用浏览器智能体模拟人工操作，从没有接口的老系统里取数。集成成本正在数量级地下降。(出处见附录 C)

架构层面，坚决不做旧环境的寄生体。验证期的系统应该运行在自己可控的边界内，通过接口与旧系统交互，而不是把代码写进旧系统里。理由有三：1）验证期方案本身有一半以上的概率被推翻重写，寄生越深浪费越大；2）写入旧系统要走客户的变更管理流程，周期以月计，与 MVD 的周级节奏根本冲突；3）保持「可撤离」的姿态，本身就是谈判筹码和诚实姿态——FDE 做完会走，寄生体却永远走不了。

流程层面，顺从人的习惯，而不是系统的习惯。这是最容易被技术团队搞反的一条。技术环境可以强硬，人的习惯却必须顺从。

如果业务用户的核心动作发生在表格和邮件里，MVD 的界面就应该出现在表格插件和邮件里，而不是要求用户登录一个崭新的门户。OpenAI 在西班牙对外银行的部署，从 12 万员工已经在用的 ChatGPT 界面切入，而不是另起炉灶，就是顺从习惯的典范。2.1 节那条死因在这里依然成立：员工不愿采用新工具，在五大路障里排第一。新系统真正难对付的，是用户的旧习惯。

2.7 「行胜于言」的用户调研

最后，把镜头拉回方法论的源头，谈谈 FDE 式的用户调研与传统调研的分野。

传统调研的信条是「问」：问卷、访谈、焦点小组。FDE 的信条是「看」和「做」——行胜于言。原因有三，层层递进。

第一层：客户不知道自己要什么。这是认知规律。面对全新品类——2004 年的情报分析软件，2025 年的人工智能智能体——用户没有参照系来描述需求。

Palantir 的演示循环之所以有效，恰恰因为它不问「你要什么」，而是说「这是我做的东西，你来说说哪里不对」。人对「哪里不对」的判断力，远强于对「要什么」的想象力。

第二层：客户说的和做的，是两回事。2.1 节那个律师群体的例子是最好的证明：问卷调研「是否愿意使用专业法律人工智能」，采购部门会告诉你意愿强烈——毕竟刚花五万美元买了工具。但观察律师们的实际行为，他们在用 ChatGPT 起草合同。

企业语境下，「说」被太多因素污染：政治正确、对供应商的客气、对自身角色的维护。只有行为不会撒谎。影子工作法观察行为，MVD 测量行为，都建立在「做比说真」的基础上。

第三层：最高质量的调研，发生在共同劳动中。访谈里，客户是「被研究对象」，警惕而表演；并肩干活时，客户是同事，松弛而真实。训练营里 FDE 与客户技术人员背靠背写代码的那两三天，交换的信息密度超过任何正式调研——客户会在调试的间隙随口说出「其实这个字段我们从来不信」「这条流程表面走系统，实际还是打电话」。

这些话在正式访谈里永远不会出现，因为它们看起来「不正式」。但它们恰恰是部署成败的关键情报。

访谈本身也在被人工智能改造。营销技术公司特赞的首席技术官丁鑫栋分享的做法是，不再发结构化问卷，而是用智能体对企业全员做自主访谈——根据不同的场景和不同人的实际痛点，做差异化沟通，最后汇总出一张「每条产品线处于什么阶段、痛点是什么、适合用什么方式推进」的全局诊断图。人工智能把这种全员深聊的成本打了下来。（出处见附录 C）

到这里，正确的问题已经锁定，价值已经初步验证。下一个战场，是把验证过的单点价值变成一纸真正的合同，以及一段真正开始的关系——如何赢得客户。

第 3 章 赢得客户

> 「企业买人工智能，就像你奶奶拿到一部 iPhone——她想用，但需要你帮她设置好。」
> —— a16z（安德森·霍洛维茨基金）

3.1 筛选你的灯塔客户

互联网产品获客的第一课是筛选种子用户：100 个爱你的用户，胜过 10000 个觉得你还行的用户。FDE 世界里的对应物，是灯塔客户——那种不仅能给你收入，更能给整个行业发信号的客户。

灯塔的战略价值，在 FDE 模式下被成倍放大，原因有三。

第一，灯塔是最强的销售资产。企业客户决策链长、风险厌恶度高，所以同行背书是最短的说服路径。

法律人工智能公司 Harvey 的起家史就是教科书：2023 年 2 月官宣的第一个大客户，是拥有 3500 名律师、43 个办公室的全球顶级律所年利达。这座灯塔一亮，普华永道、佳利等大客户接踵而至——法律行业最讲出身，拿下年利达，等于拿到了整个顶级律所市场的通行证。Palantir 的早期历史同构：中情局是它最严苛、也最有背书价值的客户，情报界的信任，日后变成了敲开华尔街、制造业和政府市场的钥匙。

Harvey 点亮这座灯塔的过程，本身就是一堂 FDE 课。

2022 年 11 月，年利达成立了一个专门小组——市场创新组，由合伙人大卫·韦克林牵头，开始秘密试用这家当时名不见经传的创业公司的产品。他们不看你演示，直接全所实战。到试点结束，3500 名律师向系统提出了约 4 万个真实工作问题，覆盖 250 个业务领域、50 种语言。韦克林的结论是那句被法律界反复引用的话：「我做了 15 年法律科技，从没见过这样能改变游戏规则的东西。」另一位合伙人给媒体演示了一个具体画面：让系统给美国客户准备一份「如何在卢森堡开银行」的十页材料，「它真的做出来了。」（出处见附录 C）

第二，灯塔决定你的产品基因。FDE 模式下，现场学习回流产品——这意味着你的前十个客户，在事实上参与塑造你的产品。一旦选错灯塔，产品就会被带向没有普遍性的方

向。a16z 给创业公司的第一条建议「聪明地卖」，说的就是这个：年轻的公司不可能满足所有人，所以要选一个在系统环境和使用场景上至少有共性的理想客户画像（ICP），让每次交付的学习可以累积，而不是互相抵消。

第三，灯塔的质量比数量重要一个数量级。你的产能天然稀缺——一支精英团队同期能深度服务的客户是个位数，一旦接错一个，代价就不是「少赚一单」，是一支顶尖团队被泥潭占用半年。所以第 2 章的三类高危信号（没有业务负责人、拒绝提供真实数据、宇宙级需求）在这里同样要查，还要加一条灯塔特有的检验：合作初期就谈好，成功后他愿不愿意站出来——联合发布案例、行业会议现身说法、接待你潜在客户的参访。不愿意公开替你说话的灯塔，价值至少打对折。

警惕「需求蝗虫」

做互联网产品要警惕「产品蝗虫」——蜂拥而至、用完即走、还误导产品方向的早期用户。FDE 世界里的对应物种是「需求蝗虫」：预算充足、需求旺盛、但会吸干你团队却不产生任何复利的客户。三个识别特征：

1. 需求与公司战略方向明显偏离，做了也沉淀不下可复用能力；
2. 把 FDE 当廉价外包使唤，按人头派活，而不是按结果对齐；
3. 内部政治消耗巨大，你的主要工作变成帮某个部门证明另一个部门错了。

对需求蝗虫说不很难——它们的合同额往往诱人。但 Barry 的账是：错误的试点烧掉的不只是当下的成本，还有团队的时间、士气，以及本可以长在正确客户身上的产品复利。

3.2 从最笨的事情做起

著名孵化器 YC 有一条古训：「做不可规模化的事。」例如，共享民宿平台 Airbnb 的创始人曾经挨家挨户给房东的屋子拍照，在线支付公司 Stripe 的创始人也曾亲自帮用户安装自家软件。

麦格鲁在 Lightcone 播客里被问到 FDE 与这条古训的关系时，给出了一个精准的表述：FDE 模式，就是在规模上做不可规模化的事。

这句话道破了 FDE 获客的哲学底色：在企业市场，信任的获取没有规模化捷径，只有笨办法。三个层次，层层递进。

第一层：人要到场。a16z 给前线部署团队的建议清单里，最后一条只有四个字：亲自到场。理由是「陈词滥调，但陈词滥调之所以成为陈词滥调，是因为它是真的」——到场不仅改善销售，更能在梳理客户内部权力关系、推动新工具被采用时，成倍提高成功率。

企业客户的信任，是按「见面次数」和「共同经历的事」计价的。远程会议能混个脸熟，可信任得在一起扛过事之后才长出来。

第二层：手要弄脏。OpenAI 的 FDE 在爱荷华的农田里跟着农艺师下地；Palantir 的工程师在油田钻井平台和飞机总装车间里一住数周；Harvey 的 FDE 在律所里逐个合伙人做采纳演示。这些场景里没有一个是「高效」的，但每一个都在生产远程沟通不可能生产的东西：对客户处境的亲身感受。体感直接转化为方案质量——只有踩过农田的泥，你才会理解为什么那个看似完美的手机界面，在户外强光下根本没法用。

第三层：先做仆人，再做导师。FDE 进场初期最常见的错误，是带着「我们来拯救你们」的姿态。姿态一旦错了，一切情报通道就都会关闭。

正确的顺序是先做最不起眼的服务：帮客户的分析师修一个数据问题，帮信息部门补一份接口文档，帮业务团队把一份周报自动化掉。这些「笨事」换来三样战略资产：组织内部的真实地图（谁说了算、谁被信服、谁是隐藏的关键节点）、数据环境的真相（哪份数据被信任、哪份是摆设）、以及最重要的——「这个外来者是自己人」的身份认证。等这个身份建立起来，你说的话才开始有重量。

笨办法的尽头不是永远笨下去。第 7 章会讲怎么把这些笨功夫沉淀成可复制的手册——但在规模化之前，你必须先在泥里趟出那条值得被复制的路。

3.3 信任红利：标杆客户的背书矿藏

企业级采购是一个「信息不对称极度严重、失败代价极度高昂」的市场，决策者对信息的信任排序层级分明：厂商自己的宣传，不如分析师报告；分析师报告，不如同行的公开案例；同行的公开案例，不如同行的私下推荐。最后一档——「我认识的人用了，说真有用」——转化效率碾压其他所有渠道。而 FDE 模式恰恰是生产「私下推荐素材」的最佳机器：你交付的不是软件许可证，是客户高管在同行饭局上的一句「他们真的懂行」。

开采这座矿藏，有三个层次的动作。

第一层：把交付做成「能讲出去的故事」。客户高管愿意替你传播的前提，是你的交付能被他讲成一个让他在同行面前有光彩的故事。这要求交付成果具备三个叙事要素：一个具体的数字（「化学品使用减少 70%」「反洗钱调查从几小时到几分钟」）、一个具体的人（「我们的农艺师和他们团队一起在地里干」）、一个具体的反差（「以前要三个月，这次只要五天」）。

交付结束时，主动帮客户的内部支持者备好这套叙事——一页纸、三张图、一个 30 秒能讲完的版本。材料做得多细，决定了别人转述时走不走样。

第二层：把客户的成功，变成客户的社交货币。连锁药房 Walgreens 就是样本：通过训练营模式八个月部署四千家门店之后，成了行业会议上的明星；市场研究公司 J.D. Power 走得更远——作为客户参加完 Palantir 的训练营后，开始为自己的客户办训练营。（出处见附录 C）这是最理想的状态：客户把你的方法论，当作它自己的行业领导力来展示。背书从「感谢你」升级为「以用你为荣」。

第三层：小心背书的「反噬机制」。企业市场是记仇的。一次高曝光客户的失败交付，传播速度远快于十次成功——因为失败故事更下酒。这就是为什么第 2 章的「拒绝机制」和第 4 章的「激活纪律」如此重要：灯塔战略的死穴不是找不到灯塔，是让灯塔在你手里熄灭。

3.4 用数据摸清客户家底

FDE 进场前后，有一个动作叫「进场尽职调查」（下称「尽调」）：在写第一行代码之前，用结构化的方法，摸清这个客户的全部家底。一份完整的尽调清单，包含五份地图。

- **数据地图：** 客户有哪些数据源、各自由谁拥有、质量如何、权限归谁管、有没有你想都想不到的暗数据——某个老员工私藏的表格、只在邮件里流转的报表、纸质台账。重点不是「有什么」，是「哪个被信任」。几乎每个组织里，官方数据源和被员工真正信任的数据源都不是同一个，后者才是你要接的。
- **流程地图：** 目标业务流程的真实运转图——不是流程文件里的版本，是影子工作法（shadowing，像影子一样跟着用户观察其真实工作）观察出来的版本，包括所有没写进文档的环节、例外和变通。特别标注三个点位：时间消耗最大的环节、出错代价最大的环节、情绪最激烈的环节。这三个点位，通常就是价值的富矿。

- **组织地图：** 谁发起、谁买单、谁使用、谁能否决、谁是无冕之王。企业项目最常见的死因不是技术，是没算准组织账——你的内部支持者职位不够高，或者太高（高到没时间管你），或者正确的人没被纳入。特别要找到那类「知识枢纽」：职位不高、但所有人遇到实际问题都去找他的人。他们既是最好的需求来源，也是日后推广的种子节点。
- **系统地图：** 技术环境的真实面貌——要对接的系统清单及其接口状态、安全与合规边界、变更管理流程。系统地图决定你的部署架构和排期。很多项目的时间表在第一天就注定破产，因为没人问清楚：客户那边发布一个版本，要走六周流程。

最后一份地图最敏感，也最关键：政治地图。它要回答的是另外三个问题：这个项目动了谁的奶酪，要自动化的目标流程是不是某个部门的权力来源，方案落地之后谁的工作会显得多余。想象一个具体场景：系统上线演示那天，全程配合的业务部门在鼓掌，角落里另一个部门的主管却一言不发——他要维护的那套流程，正是被你自动化掉的那一套。《生成式人工智能的鸿沟》报告把「员工不愿采用新工具」列为第一大路障，而抵触的根源大多不是懒，是恐惧——怕被取代、怕被证明无能、怕失去存在的意义。

政治地图的功能，是提前识别这些恐惧的载体，并在方案设计里给他们安排「新生路」，而不是「死路」。第 5 章的续约话题会回到这一点：被你设计成「受害者」的人，会成为续约时最坚定的反对者。

五份地图齐了，你才算真正「进场」。这套尽调通常要一到两周，由「回声」（Echo，负责用户研究的业务团队）与「三角洲」（Delta，驻场工程师团队）分头进行、每日对表。它看起来像纯成本，但账要这么算：两周尽调省下的，是三个月在错误战场上的狂奔。

3.5 技术内容营销：打造持续输出的信任引擎

内容营销是互联网公司获客的经典武器。FDE 公司的获客逻辑，决定了它的内容营销有一个独特的定位：不追求流量，追求「专业信任的预售」。

企业级客户找到你的方式，不是刷到你的广告，而是：遇到难题，搜索、打听，发现某家公司对这个问题有过极其专业的公开论述，然后得出结论——「他们懂，找他们」。内容在这里扮演的角色，是客户决策链最前端的「资格预审」。Palantir 常年发布场景化的技术文章；OpenAI、Anthropic 把企业客户案例做成详细的技术叙事；a16z 的《Trading

Margin for Moat》（为护城河牺牲毛利）为整个赛道定了调——这些都不是品牌宣传，是精心经营的信任资产。

FDE 公司的内容营销，有三条区别于常规企业内容的军规。

- **军规一：写「战壕视角」，不写「展台视角」。** 常规企业内容讲产品多强、愿景多大；而 FDE 内容讲问题多难、我们怎么趟过去的。前 Palantir 工程师 Barry 那篇《理解前线部署工程》（Understanding Forward Deployed Engineering），为什么在业内被疯传？因为它写的全是展台上看不到的东西：重复造轮子的浪费、负利润率的试点、累垮的工程师——结果这篇「自曝家丑」的文章，成了 Palantir 模式最好的布道。战壕里的真话自带穿透力，因为读者能分辨：谁在讲营销话术，谁在讲他们每天活在其中的现实。
- **军规二：方法论开源，制造「被引用的资格」。** 把自己怎么做发现、怎么做验证、怎么做交付的方法论公开，短期看是教会同行，长期看却是定义行业标准——当整个行业的客户都开始用你的框架提问（「你们的评估体系怎么建？」「你们的部署清单长什么样？」），你就从供应商变成了出题人。2026 年出现了 OpenFDE 这样的从业者开源社区并快速聚合人气，正说明这个行业的知识饥渴远未被满足；谁先系统性地满足它，谁就掌握定义权。本书的写作，某种程度上也是同一逻辑的实践。
- **军规三：让客户的内部支持者，成为内容里的英雄。** 案例文章的署名逻辑很微妙：主角应该是客户方那个有远见的管理者，而你的团队是「帮助他成功的伙伴」。给支持者舞台，等于给下一个潜在客户里的同类角色，递上一张「成为他」的邀请函。

内容之外，还有两类资产能替你在客户工程师的桌面上长期「上班」。

一类是开源的工具与组件。企业采购的发起者是高管，技术否决权却在工程师手里。把交付中沉淀的通用模块——评估框架、连接器、部署模板——开源出去，等于在决策链的技术端预埋了信任票。Anthropic 把模型上下文协议（MCP，一种让模型连接外部工具的开放标准）做成公开规范，FDE 在客户现场的交付物就建立在这个协议上——客户工程师学会的第一课，就是 Anthropic 的技术栈。资源有限的团队还有一个轻量版：一个能让客户工程师一下午接上样例数据跑通的快速上手包，就是最好的销售工程师。

另一类是工程师社区的存在感。目标行业的技术社区里，持续出现你的工程师的身影——回答问题、分享踩坑、提交代码。这种存在感的转化路径很长，但它触达的，正是掌

握技术否决权的人。FDE 公司的招聘与获客，在这个动作上是合一的：最好的候选人池和最好的客户线索，往往来自同一个社区。

3.6 采购、法务与安全审查的通关术

企业市场的「闸门」是采购部、法务部和安全审查委员会。大量技术上成功的 FDE 项目，死在这三道闸门里，而且死得毫无技术尊严：合同条款谈崩、数据处理协议卡壳、安全问卷填到第四个月。

通关的关键，是怎么看待这三道闸门：当阻碍，你会在最后一公里翻车；当成交付的一部分，反而能建立优势——多数技术公司把安全问卷答到第四个月，而你两周交卷，就赢了。

采购关，给他们「可交差」的工具：清晰的分阶段报价（让降价有台阶）、可对比的市场基准（让审批有依据）、可衡量的结果承诺（让「买贵了」的指控无从成立）。按结果收费的模式在这里有意想不到的优势：采购最难审批的是「说不清会得到什么的支出」，而「每解决一个工单付 X 元」这种报价，采购一眼就懂。

法务关，把数据处理协议当产品来打磨。企业人工智能项目的法务焦点高度集中：数据用不用来训练模型？存在哪、谁能访问？出了安全事故谁赔？输出出错算谁的？聪明的团队把这些问题的标准答案产品化——预置的协议模板、模型使用声明、分级的事故责任框架。Palantir 能在情报界生存，靠的是把「权限与审计」做到产品内核（谁看过什么数据，全程可查）；Anthropic 在与金融客户的合作中，把「可追溯、可审计」做成了卖点。这一关的信任，主要靠产品和流程提前设计出来，而谈判桌上临时争取来的很有限。

安全审查关，用「预答白皮书」抢出两个月。高效团队的做法，是主动维护一份安全白皮书：部署架构图、数据流图、加密与权限方案、合规认证、历史审查中所有被问过的问题的标准答案——多数问卷的八成，可以直接从里面复制粘贴。你的响应速度本身，就是安全成熟度最直观的信号。

3.7 生态捆绑：站在渠道的肩膀上

FDE 时代的生态策略，是借三类伙伴的肩膀。

第一类：云厂商与大平台。亚马逊云、微软云、谷歌云，本身就是企业级人工智能采购的总入口之一。进入它们的联合销售体系，等于获得一张直达企业采购清单的通行证。云厂商的应用市场还解决了一个实际痛点：客户可以动用既有的云承诺消费额度（committed spend）来采购你的服务，绕开新增供应商的采购流程。

第二类：咨询公司与系统集成商。这是 2026 年最戏剧性的生态变局。OpenAI 新设的交付实体「部署公司」（The Deployment Company）的创始伙伴名单里，赫然并列着贝恩咨询（Bain & Company，与贝恩资本是两家相互独立的公司）、凯捷、麦肯锡——全球最大的咨询与集成巨头，从「潜在的竞争者」变成了「持股的同盟」。

逻辑很清晰：模型公司有技术，咨询公司有客户关系与行业纵深，集成商有落地人力——三方合流，才能把「人工智能转型」这个巨型市场整体吃下。对创业公司而言，启示是双向的：既要警惕咨询巨头借生态之名吃掉你的交付层，也要看到与区域性、行业性集成商结盟的真实红利——它们手里，握着你三年也建不起来的客户关系。

同样的剧本几乎同时在中国上演。2026 年 7 月，字节跳动旗下的火山引擎与四大会计师事务所之一的安永签署战略合作：围绕数据治理、财务管理、营销增长等核心场景共建解决方案，最值得注意的是，双方计划搭建一支千人级的 FDE 团队，「打造人工智能原生的交付团队」。火山引擎总裁谭待的话很直白：FDE 团队要「让具备技术与行业双重背景的工程师，前置到客户现场，深度参与方案落地的全流程」。一边是坐拥客户关系的咨询巨头，一边是缺行业纵深的模型平台——FDE 成了两边的焊接点。（出处见附录 C）

第三类：客户的客户。最高级的捆绑，是嵌进客户自己的价值链。J.D. Power 参加完 Palantir 的训练营后，开始为自己的客户办训练营——Palantir 的能力顺着 J.D. Power 的客户关系外溢，获客成本趋近于零。

设计你的交付时，就可以预埋这种「可再交付性」：这个方案，能不能让客户拿去服务它的客户？能，你就嵌进了客户的商业模式里。

3.8 排期：谁先谁后，本身就是战略

稀缺可以制造吸引力——邀请制、排队机制，都是互联网产品的经典玩法。FDE 的排期问题，在形状上相反、在原理上相通：你的产能总是小于需求，「谁先谁后」本身就成了战略工具。

先说残酷的现实约束。一支合格的前线部署小分队（一名业务侧加两三名技术侧），同期能高质量交付的项目，通常不超过两个。行业里另一组口径可以互证：一个客户成功经理能同时管八到十二个客户，而一个 FDE 深度嵌入的客户只有一到三个，单个客户的部署往往意味着三十到六十天的每日在场。OpenAI 早年从 2 个工程师起步，Sierra 的部署周期以月计，Harvey 的单所部署要六到九个月——精英交付产能，是天然的稀缺资源。

产能稀缺躲不掉，能经营的是给这份稀缺定个什么价。

定价的第一条红线是合同体量。投资人托马什·通古兹算过一笔直白的账：给一万美元的合同配一个年薪二十万美元的 FDE，账永远算不过来——这个模式，大约在单份合同十万美元以上才开始成立。后面要说的种种排期算计，都得先越过这条底线再谈。折算到中国市场，这条红线大致对应大几十万到百万人民币级的项目——7.8 节那笔 10 人小队的账，可以拿来对一对自己的账。

排期的第一原则，是按战略价值排，而不是按合同金额排。决策矩阵只有两个维度：这个客户的灯塔价值（行业信号强度），和这个客户的学习价值（能沉淀多少可复用能力）。两个维度都高的客户，甚至可以、也应该倒贴资源优先做——Barry 回忆 Palantir 早期免费做试点、烧掉数百万美元，赌的就是这个矩阵。两个维度都不高的大合同反而最危险：钱看着多，却把精英团队套牢半年。

第二原则，是把「等待」产品化。客户排队三个月才能进场，你别让这三个月白流：给他数据准备的清单（进场即提速）、组织预热的建议（先搞定哪几个关键人）、轻量的远程诊断（保持温度、提前扫雷）。等待期经营得好，进场后的交付周期就能缩短三分之一——排队就从客户体验的负担，变成了交付质量的组成部分。

第三原则，是永不承诺超越产能的并行。你的交付质量，靠的是同一支团队的连续专注——每个项目都背着一整套客户语境，切换一次就要重新加载一次，多项目并行的损耗远大于想象。宁可让销售把「我们档期排到下个季度」当稀缺性卖（这往往还能提价），也别让团队在三个客户之间疲于奔命。生意越好的时候，越容易忍不住多接单。而多接单崩掉的代价，往往比少接一单大得多。

3.9 提案与验证方案的撰写

企业市场里单价最高的文本，是提案与概念验证（PoC）方案。

坏消息是，多数技术团队写的提案都在犯同一个错误：通篇讲「我们要做什么」，而不是「你将得到什么」。好的 FDE 提案，遵循一个严格的倒金字塔结构。

第一层：业务结果，一段话说完。提案的第一段，必须是客户语言的业务结果：「八周内，把你行反洗钱调查的平均处理时间从 4 小时降到 15 分钟，释放的调查产能相当于现有团队的 2.5 倍。」没有这一句，客户的高管就根本看不进去后面的内容。

第二层：价值验证路径，怎么证明做到了。写明验收指标、测量方法、基线数据、以及每一阶段的退出机制。这一层传递的信号是「我们敢被检验」——在被过度承诺伤害过的企业市场里，敢被检验，是最稀缺的诚意。

第三层：交付方法，我们怎么做。这一层才轮到技术方案，而且写法上要让非技术读者也能跟上：架构图配业务注释，里程碑配决策点。特别要写明需要客户配合的事项——数据访问、关键人员的投入时间。把客户该干什么写清楚，一来提前排雷，二来也是筛选客户。

第四层：风险与对策，我们想过会怎么死。多数提案回避风险，仿佛提了就不吉利——而成熟的买家选供应商，看的正是他敢不敢把死法说清楚：数据质量不达标怎么办、关键人员变动怎么办、准确率达不到门槛怎么办。这一层是提案的信任放大器，也是第 2 章「拒绝昂贵的概念验证坟墓」理念在文本上的体现。

概念验证方案另有特殊之处：它是「毕业标准」的载体。方案里必须写明周期上限、验收指标、以及毕业与散伙的两种后续。写得含糊的验证方案，等于给概念验证坟墓签发开工许可。

这种「敢被检验、写死标准」的打法，有比人工智能时代早了十年的先例。数据分析公司 Looker 从 2013 年起就在免费试用期内做重型售前实施，原则只有一条：演示就是概念验证——永远向潜在客户要真实的数据集来玩，绝不用精心准备的样例数据表演。同时他们把账算到了尽头：每客户每年约 2.5 万美元，两千个客户就是一亿美元的年收入，够摸到上市门槛。其时任首席执行官的话很直：如果你说不清自己需要的是两千个还是十万个客户，那你就是在烧风投的钱。

验证方案敢写多狠，取决于背后的账算得多清。

买方那一侧的标准，也在朝同一个方向变硬。人力资源软件公司 Rippling 选型时，把「供应商必须投入专属工程资源」写进了硬性标准——他们要的定制化接口工作流，没有厂商的工程师下场支持就做不出来。协同文档公司 Notion 也一样：它本身就是一家人工智能优先的公司，权衡之后依然选择买而不是自建，其招标书列了五条标准，最后一条是：

不只要一个供应商，更要一个能一起共建的团队。面对这样的买家，提案写得再漂亮都不够——第 2 章说的「三关必须在客户现场过」，从投标那一刻就开始了。

3.10 从驻场到远程：混合交付的边界

FDE 交付这几年经历了一次重要进化：从纯驻场，走向「远程为主 + 关键节点驻场」的混合模式。Palantir 官方口径也确认，如今许多项目大部分远程执行，驻场集中在关键里程碑。

混合模式的节奏设计是一门手艺。哪些事必须物理在场？经验归拢下来是三类：1）关系建立的初期（第一次见面、影子工作法、与高管的信任建立——信任这东西，隔着屏幕建不起来，必须见面）；2）高强度共创（训练营式的联合构建、关键架构决策的白板攻坚）；3）政治敏感期（方案推广、部门协调、变革管理——这些时刻，你需要读懂会议室里的空气）。

哪些事远程反而更好？深度编码、文档沉淀、常规迭代——需要无打扰心流的工作。

混合模式还有两个隐性红利。一是成本结构：驻场是 FDE 成本里最大的可变项之一，合理的混合比例，能把交付利润率拉出十个点以上的空间。

二是人才可持续性：常年一半时间出差的团队，倦怠率显著更高——论坛上从业者对差旅的抱怨高居榜首。招聘点评网站上，Palantir 的 FDE 岗位评价把这种分裂写得明明白白：薪酬福利 4.0 分（满分 5 分），工作与生活平衡却只有 2.8 分，高频抱怨是工作节奏被不断打断（第三方聚合口径，未经原页逐条核对）。混合模式不是偷懒，是组织续航能力的保障。

但底线要划清：全程不见面的项目，失去的不只是关系，还有对现场的第一手认知。远程维持的是已有的信任与语境，而它们，必须先在物理世界里被创造出来。

把边界再推远一步，就是出海。美国公司的全球化由巨头示范：OpenAI 的 FDE 招聘版图，从旧金山、纽约铺到伦敦、东京、新加坡、阿布扎比——团队的地理分布，精确描摹出企业人工智能付费能力的全球地图。中国语境下是另一番逻辑：过去的困境是「产品基因」，国内厂商长期做高度定制的项目制生意，跟海外市场要的标准化、产品化对不上。但 FDE 模式给了一个反向视角——中国工程师的交付文化恰恰最贴近 FDE 的要求，真正的问题是他们背后的平台底座够不够厚。

本土的参照系，已经有人闷头跑了很久。53AI 从 2023 年起就按结果收费——客户业务真的变好才付钱，比硅谷的同行还早。学费也交过：早期让提示词工程师单挑交付，摸不透客户的生意，创始人一年半里有一半时间在四处救火；改成业务专家先进场吃透生意、再带工程师进场搭建之后，局面才扭转。最硬的证据来自一次对照实验：一家在线教育的头部客户，「销售加智能体」组的产出做到了纯人工组的 3 倍。更值得记下来的是客户的反应——人效上去之后，客户没有裁员，而是用同样的人去抢市场上更多的优质线索。（出处见附录 C）

出海的具体策略上，FDE 有一个先天优势和一个先天约束。优势：「灯塔到背书」的逻辑在全球市场同样有效，且发达国家企业客户的付费意愿和契约精神更成熟。约束：FDE 是重度本地化的生意——语言、时区、合规、本地信任网络，每一项都要求本地团队，而非远程支持。这意味着 FDE 出海的成本结构天然高于 SaaS 出海，节奏必须更克制：先用能远程交付的产品站稳脚跟，再在重点市场组建本地交付团队，一个国家一个国家地做。

这十件事归结到一句话：赢得客户，就是在一个被辜负过无数次的市场里，做到让人信。签下合同只是入场券——接下来，真正的硬仗开始：让系统活起来，让人用起来。

第 4 章 激活部署

> 「模型通常是最干净的部分。难的是找到那个没人写进文档的工作流。」 —— 一位一线 FDE
>
> 「无论技术如何进化，组织内部的人性和权责博弈，始终是比技术更复杂的难题。」 —— 申悦，FDE 一线从业者

4.1 上线不等于激活：企业部署的「首日魔咒」

「激活」在消费互联网，指新用户完成关键行为、体会到产品的「啊哈时刻」（aha moment，用户第一次体会到产品价值的瞬间）。而在企业部署里，激活指**目标用户群体在日常工作中，形成对系统的稳定使用习惯**——不是演示会上鼓掌，是三个月后没人催、系统依然被高频使用。

换句话说：企业软件的「上线」，走完的只是采购和验收的行政流程；「激活」发生在用户的行为里。上线那天固然值得庆祝，但项目最终是否成功，还得看三个月后还有没有人继续使用。

麻省理工学院 NANDA 实验室那份《生成式人工智能的鸿沟》报告里有一组数字：只有约四成企业为员工提供官方的人工智能工具订阅，而多达九成员工，日常在用个人消费级的人工智能工具干活。这意味着大量企业的真实状态是「双轨制」：官方系统空转，影子人工智能（shadow AI，员工绕开公司官方系统、用个人工具干活的现象）横行。系统上线了，激活却从未发生。

为什么企业部署普遍死在激活环节？因为它要过的是三道关，和消费互联网产品（面向个人用户的 to C 产品）完全不是一种关。消费互联网产品死于「不好玩」，而企业部署死在用户手里——用户嫌它不顺手：比旧习惯多一步，就没人用；用户觉得它不可信：一旦错一次，信任就清零；用户觉得它与自己无关：那就更没人碰。

这三关，没有一关能在总部解决，你只能到客户的会议室、车间和工位上去过。所以激活是 FDE 的主场，也是它跟「交付即走」的传统实施最大的不同：传统实施把验收单当终点，FDE 把客户行为的改变当终点。

4.2 快速迭代：部署期的「热修复」文化

互联网产品迭代靠 A/B 测试（分组对比实验），用小步快跑逼近最优。但企业部署场景里做不了严格的 A/B 测试，样本太小、干扰太多。可它的精神内核——快速、小步、基于证据的迭代——在部署期变成了一种工作方式：像抢修线上事故一样，快速响应用户的每一个小抱怨。我称之为「热修复」。

消费互联网产品的迭代节奏以「版本」计，周更、双周更。而 FDE 部署期的迭代节奏以「天」甚至「小时」计：上午业务用户说「这个输出少了供应商编码字段」，下午字段就加上了；今天车间主任说「这个界面戴着手套点不准」，明天按钮就放大一倍。这种响应速度对用户的意义，远超功能本身——上午提的问题下午就改好，比任何承诺都更能攒下用户的信任。用户在第一个月就会形成判断：「这个团队是来真的」，还是「又一个交完就走的」。

这种节奏不是口号。

品牌效果营销公司 Skai 就接过这么一个紧急需求：数千个商品要在多个电商平台之间统一分类。放在过去，这事得拉上数据科学团队排期——「他们有工具，但产能有限，排进路线图要等很久。」这回，一支面向客户做定制功能的小分队直接在数据平台上调用模型：第一天写好提示词、在客户商品库上跑通，第二天就发布到生产——同样的活按传统做法至少要一个月。带队的副总裁事后总结：「我不需要变成提示词工程师，也不需要变成大模型专家。」热修复的底气，是工具链已经铺到了问题旁边。（出处见附录 C）

热修复文化有三个执行要点。

第一，反馈必须直达写代码的人，中间不能隔着传话筒。以前，用户的反馈要经过客户成功、产品经理、排期，最后才到工程师——每传一手丢一半上下文，等排上期，一周过去了，用户的心也凉了。而在 FDE 模式下，反馈直达写代码的人，最好写代码的人就坐在用户旁边——这不是修辞：第一美国主管数据与人工智能的副总裁普鲁布·纳辛纳（Prabhu Narsina）回忆与平台方工程师的合作，原话是「他们几乎每天和我们一起开会，一起写代码，一起 debug」。Palantir 缩短的正是这个回路：前线部署工程师现场就能构建、测试、学习、回传，不必等一个正式的产品需求走完流程。

第二，迭代优先级由「使用阻塞度」排序，而不是「功能重要度」。部署期的取舍逻辑与产品期相反。产品期按战略价值排需求，而部署期要问的是「什么在阻碍明天的使用」——一个配色问题如果让车间工人觉得「这是给办公室白领做的」，它就是最高优先级；一

个强大的预测功能如果用户当前用不上，它就排到激活完成之后。部署期先把使用率做出来，功能深度往后排。

第三，每天收工时问一个问题：今天的改动，让用户明天的哪一刻更顺了？这样修下去，就不只是「有求必应」——每一次修复，都是一次有意识的激活设计：你在逐个拆除用户与系统之间的摩擦点。摩擦点的清单，就藏在进场尽调的流程地图和日常观察里。

OpenAI 的 FDE 工作法里有一个对应的节奏划分：前期共创（驻场白板对齐）、验证（建评估体系，即 evals）、交付（多日驻场构建）。注意交付阶段仍以「多日驻场」为单元——能修得这么快，靠的就是人守在问题旁边。

4.3 在客户的环境里迭代：评估体系驱动的质量提升

速度解决了，还剩方向问题：往哪儿改、改成什么样算好？在人工智能部署里，回答这个问题的载体是一个 2024 年后才成为主流的实践：**评估体系**——用一批标准案例，给系统的输出持续打分。

传统软件的质量只有两种状态：功能对或错，测试过或不过。而人工智能系统的质量是连续的、概率的、场景相关的——同一个回答在演示里惊艳，在某个具体业务语境里却可能是灾难。更要命的是，「好」的定义权在业务方，不在工程方：模型觉得完美的回答，业务专家可能一眼看出外行。如果没有这把评估标尺，人工智能部署的迭代就是蒙眼狂奔：调了半个月，质量是升是降，没人说得清。

评估体系的工程实践，在领先团队里已经固定成三步。

第一步：从真实案例里长出来。评估集不能靠工程师编，必须来自客户的真实业务素材。这一步的本质，是把业务专家的隐性判断力，显性化为可执行的标尺。

第二步：让业务方成为评委。评估体系不是工程师的自嗨工具，评审者必须包含业务方。最好的做法是把评估做成业务专家能参与的形态：并排的输出对比、简单的优劣标注、定期的评审会。

这个过程有双重收益：评估集越来越准，业务方对系统的理解越来越深——他们看着系统在自家案例上一次比一次答得好，信任的积累是亲眼所见，而非汇报所得。第 2 章写的摩根士丹利就是这么干的：给系统打分的不是工程师，是天天用它的理财顾问。

第三步：把评估体系接到生产回路上。上线不是评估的终点。生产环境里持续采集真实输入输出，定期抽样评估，分数一旦下滑立即告警——这把「质量」从上线前的一次性

验收，变成贯穿生命周期的持续看护。毕竟，用户最敏感的不是平均质量，而是质量的稳定性：系统错一次造成的信任塌方，需要十次正确才能填回来。

评估驱动的质量提升，在曲线上长什么样？客服软件公司 Intercom 的公开案例给出了一条罕见的完整上升曲线：它的人工智能客服产品，第一代平均只能解决 23% 的会话，更换底层模型后升到 51%，再针对具体客户深度定制，最高到 86%；Anthropic 自己拿它做客服，自 2024 年上线起持续调优，解决率（会话无需人工介入就解决的比例，口径纪律见附录 A）升到 79%，每月解决约 56 万次会话。同样的爬坡，Salesforce 也走过：把自家客服智能体调了整整一年，「答不上来」的比例才从三成压到一成以下（第 2 章展开过它自用踩坑的另一面）。

上线那天的分数只是曲线的起点，而激活期的质量迭代以季度计。（出处见附录 C）

约翰迪尔的合作，是「评估体系先行」最完整的公开样本。这家公司要解决的是除草剂浪费：传统喷药机整田覆盖，而它的「看见即喷」（See & Spray）技术用 36 个摄像头加机器视觉，在时速 12 到 15 英里的行进中只喷杂草——一分钟覆盖三个足球场的面积。精准农业的理想很大：美国每年种 12 万亿株玉米和大豆，最好的农田每英亩产 200 蒲式耳，顶尖种植者能做到 600——「如果每一株都能被单独照料，产量可以翻天覆地。」这是约翰迪尔技术高管贾斯汀·罗斯的原话。

但农民要的不是技术，是可信的建议。OpenAI 的前线部署工程师飞到爱荷华，跟着农艺师下地，先和专家一起评审数百个真实作业案例、建起定制评估体系，再快速迭代模型——并且必须赶上农时，错过播种季就是错过一年。

最终结果：化学品使用量减少最高 70%，农户的互动频率提升了 6 倍。先有评估体系定下的「好」，才有这两个数字。（出处见附录 C）

评估体系的深层意义在于，把老师傅脑子里那句「什么叫好」，变成系统天天自动执行的打分标准。这正是 FDE 模式的微观缩影——客户的知识不再只是需求文档里的文字，而是活在系统里的标尺。

4.4 另辟蹊径，降低使用门槛

让用户用最少的动作到达价值，是产品设计的常识。但企业系统的使用门槛，往往是系统之外的障碍。部署实践里被反复验证的「降门槛」手法，有四种。

手法一：寄生在用户已有的界面里。用户的主战场在哪，你的系统就该出现在哪：他在电子表格里工作，你就做表格插件；他在邮件里审批，你就让审批在邮件里完成；他在工单系统里干活，你就把人工智能建议嵌进工单卡片。而如果要求他登录新系统，就等于在你和他之间加了一道每天都要翻的墙。

OpenAI 在西班牙对外银行从 12 万员工已有的 ChatGPT 界面切入，Anthropic 则通过模型上下文协议（MCP）让模型进入客户已有的工作流工具，都是同一逻辑。

家居零售商 Lowe's 把同一逻辑做到了卖场地板上：店员的人工智能助手不做成新系统，直接装进他们手里已有的手持终端，语音就能问商品、问库存。这个助手铺到了 1,700 多家门店，累计回答了逾 500 万次店员提问。项目负责人的原话：「不在卖场地板上待过，就没法为门店团队做设计。」（出处见附录 C）

西班牙对外银行是「顺着旧习惯进入组织」的最佳样本。它 2024 年 5 月开始与 OpenAI 合作，第一步只发出去 3300 个 ChatGPT 企业版账号——不做全员运动，先让种子用户自己玩。员工很快自发创建了 2 万多个定制助手（GPTs，员工可自建的专属人工智能应用），其中约 4000 个被高频使用；管理层没有禁止「影子人工智能」，而是反其道行之：「我们给大家一个安全的平台，让他们放心去试。」同时配套结构化培训：250 名高管（含董事长本人）先上课，全行建起「人工智能先锋网络」，培养了一批被内部称为「人工智能极客」的高级用户。一年多后的数据：员工平均每周节省约 3 小时，83% 的人每周活跃使用，账号也扩到了 1.1 万个。2025 年 12 月，双方顺势官宣全行推广：25 个国家、12 万名员工，并启动一个叫「八件事」的端到端转型路线图。从 3300 到 1.1 万再到 12 万，每一步扩张都发生在上一步的使用数据被公布之后——这就是「让习惯自己走路」的激活。（出处见附录 C）

这套打法的在地版本有个关键差别：中国企业少有「员工自主选工具」的土壤，所以种子不是「发账号让员工自己玩」，而是先争取一个事业部或一个条线的专项预算——组织配发的第一批账号，本身就是合法性。

手法二：默认值里藏着激活率。新用户面对空白系统的第一反应是「然后呢」，多数流失就发生在这三秒。FDE 的解法是把「第一次使用」预装好：预置的模板、预填的示例（基于客户自己的数据）、预设的引导（第一次打开时，已经有一个属于他自己的待办）。系统在第一天，就应该比用户更懂他可能要做什么——前几天蹲点观察到的东西，这时候就派上用场了。

手法三：把「问人工智能」翻译成「点按钮」。企业用户对「和人工智能对话」这件事的熟练度，远比想象参差。所以让用户自己组织提问，就等于把工程负担转嫁给最不该负担它的人。

成熟的做法是把高频场景封装成一键动作：「生成上周异常报告」「核对这批发票」「草拟这封客户函」——按钮背后，是经过评估调优的完整提问与流程。对话式交互留给探索，按钮式交互留给日常。

手法三做得最狠的，是供应商先拿自己练手。Salesforce 的自家帮助站沉淀了 74 万多篇知识文章，过去客户遇到问题得自己搜索；他们把问答智能体直接封装进帮助站，用低代码（low-code，拖拽式免编程开发）工具两个月就完成上线——如今 76% 的咨询无需人工介入就解决（2026 年 8 月访问口径）。ServiceNow 同样先拿自己开刀：据其自家披露，自家助手上线 120 天，就跑出约 1,000 万美元的年化收益。

自己先激活，既是最好的测试场，也是最有说服力的销售材料。（出处见附录 C）

手法四：先做「副驾驶」，再谈「自动驾驶」。面对高风险、高抵触的流程，不要一步到位推自动化。让系统先以「建议者」身份存在——人工智能起草、人来确认；人工智能标注、人来裁决。用户在一次次确认中，建立对系统判断力的信任校准；等确认通过率高到一定程度，自动化才提上议程。

约翰迪尔的方案至今保留着农艺师的最终决策权；金融合规场景的智能体设计，普遍保留人工把关；医疗场景更是如此——坦帕综合医院的脓毒症（一种可致命的全身感染反应）预警至今是系统提示、医生裁决（第 8 章会展开这个挽回数百条生命的案例）。先副驾驶后自动，是拿一点激活率换风险可控的稳妥走法。

但「人来确认」有一个反直觉的失效模式：确认一多，人就不看了。Anthropic 的工程团队公布过一组遥测数据（系统自动上报的使用数据）：用户会批准约 93% 的权限请求——看得越多越不细看，审批疲劳让监督形同虚设；一次内部演习里，一封「帮我跑一下这个」的钓鱼邮件，25 次尝试成功了 24 次。

他们的解法是把边界从「指望人每次都认真点允许」挪到系统底层：默认拒绝，出事也翻不了墙。高风险的内容场景，则把这道边界前置到上线之前：TIME 为年度人物做交互式人工智能体验，先让红队（red team，模拟攻击者做安全测试的团队）模拟了数千种攻击手法才放行。副驾驶要配的从来不只是一个「确认按钮」，还有一道不依赖人的注意力的边界。（出处见附录 C）

4.5 旷日持久的集成大战

每个 FDE 老兵都有一身伤疤，它们来自同一场战争：与客户「遗留系统」的集成大战——那些在企业里运行多年、没人敢动的老系统。

战争的残酷程度，超出所有没在一线待过的人的想象。a16z 在《Trading Margin for Moat》里的描述一针见血：人工智能应用需要的上下文——历史记录、业务逻辑、权限体系——全部锁在企业内部的数据库、接口和工作流里，而连接它们的工作「从来都不是选择题，而是必修课」。金融业的老式大型机、医疗业数百家诊所的异构系统、制造业几十套各自为政的车间系统——这些环境的共同点是：文档过时、接口残缺、懂它们的人已经退休了一半。

这场战争的战略要点有四。

- **第一，把集成当战役打，不当杂务办。** 集成工作在多数项目计划里被写成一行「系统对接：2 周」，然后在执行中膨胀为四个月。错误的根源，是把集成当技术杂务——其实它既是技术考古，又是组织政治，还得管数据治理，是不折不扣的硬仗。正确的姿势：进场尽调时就画出系统地图，把集成风险分级排期，最硬的骨头最早啃——集成的不确定性最大，每推迟启动一天，整个项目的时间表就裸奔一天。
- **第二，数据问题先于模型问题解决。** 那份报告里有个被广泛转引的洞察：大量人工智能项目失败的底层原因是「垃圾进垃圾出」——人工智能接入了未治理的数据源，同一个文档十个版本随机命中，输出自然不可信。Palantir 的解法是本体（ontology）：先把企业的数据资产建模成带业务语义的层，厘清「哪个字段权威、哪个版本有效、谁有权看什么」，人工智能在这层地基上运行。这条路径的普适启示是：数据治理没法在开工前一次性做完，而会贯穿项目始终——很多人工智能项目干到最后，主体工作其实就是治理。连 Palantir 自己都在这事上栽过：内部一个系统被 230 万个数据条目撑爆内存，复盘的第一句话是「我们从未见过这个系统真正要处理的数据」。

美国海军的「造船操作系统」，是这个启示最壮观的注脚。2025 年 12 月，海军部长与 Palantir 首席执行官共同宣布了一笔 4.48 亿美元的合同：先覆盖两家大型造船厂、三个海军船坞和一百家供应商。造船业的数据环境是教科书级的灾难：企业资源计划系统、几十年前的数据库、纸质图纸并存。而试点阶段的两个数字让所有人闭嘴：在核潜艇制造商通

用动力电船公司，**潜艇排产计划从 160 个人工小时压缩到 10 分钟以内**；在朴茨茅斯海军船厂，**物料审核时间从几周压缩到一小时以内**。

海军部长特意强调：「这不是概念，不是试点，不是研究——这事已经在干了。」（出处见附录 C）

先花大力气把数据接进统一的语义层，效率奇迹才可能发生——而不是反过来。

同样的逻辑在三个完全不同的行业复现。快餐连锁 Wendy's：全北美 6450 家门店的糖浆缺货调度，过去要 15 名员工查上一整天，接入数据平台后五分钟出方案。房贷巨头房利美：用人工智能识别抵押贷款欺诈，检出率超过 99%，远超原来的规则系统。花旗银行：客户信用审批从几小时压到几分钟——人工智能一口气读完信用史、交易习惯、行业风险与关联企业。

三个行业，一个共同前提：先把散乱的数据归置成机器能理解的语义层，智能才有地方落脚。（出处见附录 C）

- **第三，用人工智能打人工智能的集成战争。** 一个 2025 年后的重要变量：集成工作本身开始被人工智能自动化。a16z 预想的场景已经部分成为现实——没有接口的老系统，用浏览器智能体模拟人去取数；字段映射、格式转换、接口文档解读，大量交给模型处理。领先团队的自我要求是：「尽最大可能自动化集成流程——流程挖掘、数据管道、系统对接、接口文档梳理——这种速度优势会复利。」用人工智能来干人工智能部署的活，这可能是这个岗位最妙的地方。
- **第四，知道什么时候绕开，而不是攻克。** 不是所有遗留系统都值得正面集成：有些系统的正确解法是「影子读取」：只读快照、夜间同步；有些是「人工摆渡」：过渡期保留人工环节；有些干脆「宣告隔离」：明确告知客户该流程的数据不在本项目范围内。工程师的自尊心总想攻克每一座堡垒，而 FDE 的判断力，恰恰体现在选择战场——项目要的是客户结果，不是技术全胜。

4.6 变革管理：让客户组织为你站台

激活最大的软阻力不在技术，在组织。报告的五大路障里，「员工抵触」与「变革管理」合计占了近半壁江山。让系统跑起来，工程师就够了；但要让人的行为改变，就得动员整个组织。

变革管理在 FDE 语境下，核心是三组人物的经营。

- **支持者：你的内部盟友。** 每个成功部署的背后，都站着一个客户内部的支持者：他真心相信这件事，愿意押上自己的信誉为你开路。年利达律师事务所的大卫·韦克林就是典型——作为律所市场创新负责人，他是 Harvey 在所内的对应方、倡导者和保护伞。支持者的经营要点：给他能讲的材料（能讲出去的故事）、给他战功（把他的远见变成他职业生涯的亮点）、给他安全感（失败时你顶在前面）。一个被善待的支持者，胜过十场产品宣讲会。Lowe's 干脆把这条做成了制度：每个人工智能项目必须落入「顾客怎么买、门店怎么卖、员工怎么干活」三类之一，且必须有高级副总裁级的业务负责人背书——「他们不入局，我们就不做。」
- **影响者：非正式的意见领袖。** 每个组织里都有一批无冕之王：资深分析师、车间老师傅、部门里「问他就行」的人。他们不掌权力，但掌握信任。他们的一句「这玩意儿还真行」，抵过管理层三封全员邮件；他们一句「不好使」，第二天车间里就没人点开。所以激活期要刻意经营这批人：请他们第一批试用、认真对待他们的每一条吐槽、把采纳他们的建议显性化——「这个字段是按王师傅的意见加的」。让影响者成为共同作者，是破解「不是我做的东西我不用」心态的最短路径。设备管理公司 Jamf 是个极端样本：全公司 16 个部门推广之后，最大的意外是「推动最广泛采用的不是工程师」——绩效考核的对话工具是业务同事花不到 45 分钟搭出来的，而过去同类项目要一个团队做三个月。当搭建的门槛降到这个程度，影响者就不再只是口碑节点，而是直接的建设者。
- **受损者：方案触动的利益群体。** 自动化必然重新分配工作，而重新分配工作必然制造受损者：被压缩了存在感的审批岗、被穿透了信息壁垒的部门、被替代了「独门手艺」的老员工。如果忽视他们，他们就会成为系统最顽强的地下抵抗者——消极不配合、传播事故案例、在验收时投反对票。成熟的变革管理，会提前设计「受损者的出路」：把被释放的人力导向更高价值的工作（并公开承诺不裁员），把「守门人」转型为「教练」（老师傅的经验用来训练系统），让受损者看到自己在未来里的位置。这既是人道，也是纯粹的功利——抵抗的成本，远高于安抚的成本。

票务平台 Vivid Seats 与 Sierra 的合作，展示了激活期「客户侧全员动员」的样子。决定引入智能体后，这家公司的产品、客户体验、工程三个团队全部压上——「一旦做了决定，我们就百分之百投入，全员重测试。」结果：从启动到上线不到四周，上线后问题解决率提升了 40%，客户满意度（CSAT）提升了 35%。

但更有价值的是后续变化：常规问题被智能体接走后，客户体验团队从「排队灭火」转向根治流程问题，甚至有余力做一些超出客户预期的服务；而智能体对话数据开始反哺

产品——「如果一个月有一万人问同一个功能，我们就能立刻把它排上优先级。最好的时刻是：一个规律带来的产品改进，让用户根本不需要再求助。」（出处见附录 C）激活做到头，用户反而越来越少需要求助——这个案例同时预告了第 7 章的「现场反哺产品」。

动员的结构也可以反过来设计：不只是客户在动员，供应商同样把资源嵌进客户组织。Lyft 与 Anthropic 的合作是个三件套：双方共建产品、Lyft 优先获得新能力的早期测试资格、Anthropic 为 Lyft 的工程师提供培训——结果是客服平均解决时长降了 87%，每天有数千个请求由智能体直接解决。三件套里最便宜、也最容易被省掉的是第三件，而往往正是它决定了动员能不能在客户组织里扎下根。（出处见附录 C）

中国公司里的对应样本也在出现。电商公司得物的效率工程负责人任喜亮，把万人规模公司的人工智能转型拆成三步：第一步跟全员达成共识，降低工具门槛，让所有人先用起来；第二步把企业场景按容错空间分成四象限，容错空间大的场景（比如经营分析）优先交给人工智能；第三步设专门的知识运营小组，钻进业务团队里，把专家的隐性经验从个人脑子里搬到人工智能可见的空间。共识、场景、知识——顺序不能反：先有人愿意用，再挑对的地方用，最后让人工智能有米下锅。（出处见附录 C）

变革管理的终极检验标准只有一个：当你的团队撤场后，系统是否依然被使用。如果你在场时热火朝天、撤场后迅速冷却，那不是激活，是伴舞。真正被激活的组织，会自己往前走。

4.7 我，机器人——交付工作本身的自动化

给客户交付自动化，是产品本身；让自己的交付工作自动化，是效率。后者是 FDE 团队摆脱「收入随人头线性增长」的第一杠杆。

FDE 的日常工作中，有惊人比例的动作是可模板化的重复劳动：新客户的部署环境初始化、数据接入的标准管道、安全审查的问卷应答、上线前的检查清单、周期性向客户汇报的格式。每一个第二次做的动作，都应该触发一个提问：「这能变成脚本、模板或清单吗？」

领先团队的做法，已经沉淀成四类资产。

- **部署模板：** 把环境搭建、权限配置、监控接入，打包成一键式的基础设施即代码（IaC）。新客户进场，第一天就能在客户的云环境里拉起标准化环境，而不是从零配

置一周。Decagon 能把简单场景的部署压到 15 天，背后是高度模板化的接入层。（出处见附录 C）

- **集成组件库：** 主流企业系统的连接器（预置的系统对接件），写一次、处处复用。Palantir 的本体模型，本质上把这个逻辑做到了极致：数据接入的产物不是一次性管道，而是可复用的、带业务含义的数据资产。
- **检查清单文化：** 安全审查清单、上线准入清单、撤场交接清单。这套清单最见价值的时刻，是撤场前一天：驻场工程师和客户的接手人坐下来，对着交接清单逐项打钩，钩打完，人才撤场。清单把「依赖个人经验」变成「依赖组织记忆」。
- **自动化汇报：** 向客户的周报、向公司的现场情报，都应该半自动生成。FDE 的时间单价太高，不该花在复制粘贴上——更不该让「忘了汇报」切断第 7 章要讲的现场与产品之间的回路。

自动化还有一层容易被忽略的用法：把质量检查也交给机器。DoorDash 是典型：其智能客服系统每天处理数十万通来电，背后是自动化测试容量提升了 50 倍——如果没有这样的测试产能，以天为单位的迭代根本不敢在这么大的流量上跑。这也是 DoorDash 能和外部专家团队一起，用八周完成从方案设计到部署的底气之一。（出处见附录 C）

自动化的复利有多可观，可以从一个侧面印证：a16z 把「构建或采购工具来自动化服务交付」列为组建前线部署团队的关键建议，并判断这是这一代人工智能公司能比上一代企业软件公司跑得更快、客单价门槛更低的关键变量。Sierra 公开的一组客户数据也印证了这一点：从启动到上线最快的不到两周、最慢的六周，有客户上线首日就解决了过半咨询；Palantir 把销售周期从九个月压到几周——这些都不是靠工程师加班，是靠把昨天的交付变成今天的脚手架。

前 Palantir 工程师 Barry 把这个逻辑说得更直白：公司在客户试点上烧掉过数百万美元，很多试点的利润率「字面是负无穷」，但管理层把它记在研发的账上，而不是成本的账上——每一次实施，首先是构建与学习的机会，其次才是生意。「百分之九十九的创始人和投资人咽不下这口气。」先承认交付是投资，才轮得到它产生复利。（出处见附录 C）

「FDE 不就是人海战术吗」——这个最常见的质疑，到这里也有了答案。人海的问题不在人多，而在每个人都在做不可复用的一次性劳动。每一次交付都在为下一次交付修路——这样一来，同样的一支队伍，速度也会一轮快过一轮。

系统跑起来了，人也用起来了。但企业生意的残酷在于：激活只能让项目活过第一年；能不能长期活下去，还得看续约。下一章，守住续约。

第 5 章 守住续约

> 「如果你在场时热火朝天、撤场后迅速冷却，那不是激活，是伴舞。」

5.1 续约与流失

做生意要先算账：获取一个新客户的成本，是留住一个老客户的数倍。企业级生意把这笔账放大了两个数量级：一个大客户的获客成本——数月销售周期、训练营式投入、概念验证成本——动辄数十万美元，而续约的成本趋近于零。续约率，是这门生意能不能算过账的关键。

先看流失（churn）。企业客户的流失与消费互联网产品（面向个人用户的 to C 产品）的流失在形状上截然不同：消费互联网产品流失是静悄悄的卸载，企业客户流失是一场漫长的凌迟——先是使用率阴跌，然后例会上开始有人质疑「这东西到底值不值」，接着续约谈判被「明年再看」无限推迟，最后在某个预算季被连根拔掉。更可怕的是连带伤害：第 3 章说过企业市场是记仇的，一次公开化的失败，会在行业的小圈子里被反复咀嚼多年。

企业客户流失的原因，可以归纳为五类，按可预防性排序。

- **价值蒸发：** 系统还在跑，但没人记得它解决了什么问题。这通常源于价值叙事的断档——当初立项的业务问题被解决后，没有人持续向组织（尤其是新上任的管理者）重述「我们为什么有这套系统」。价值不是一次性证明的，是需要不断重新证明的。
- **支持者离场：** 你的内部盟友升职、调岗或离职，继任者没有亲历过当初的选择，对系统天然无感，甚至敌视——「前任的政绩工程」。企业软件圈内有句黑话：「支持者一动，合同就悬。」
- **质量漂移：** 业务在变、数据在变、模型在更新，系统输出质量便缓慢下滑，用户的信任随之缓慢塌方——等塌方被管理层注意到时，通常已经晚了。
- **成本反噬：** 系统用量越大、账单越高，财务部门的审视越严。如果价值叙事跟不上账单增长，「成功」反而成了续约的障碍。

- **供应商戒断反应：** 客户对「被绑定」的警惕。咨询公司高德纳（Gartner）的分析师亚历克斯·科奎罗（Alex Coqueiro）在行业媒体上给出过一个惊人的预测：到 2028 年，70% 的企业将因供应商成本过高与内部技能不足，被迫放弃由前线部署工程师主导的智能体方案（媒体报道中分析师个人观点的口径，非高德纳正式预测报告；出处见附录 C）。这个预测本身，就是对所有 FDE 团队的警告：如果你的模式让客户感到被绑架，市场会集体反抗。

再看怎么度量「留住」。消费互联网产品看留存率；而企业生意要看三个层次（完整指标见附录 A）：使用率与活跃深度（行为层）、健康分（全称健康度评分；关系层）、净收入留存（NRR；财务层，衡量同一批老客户今年比去年多付还是少付）。其中净收入留存是最终裁判：超过 100%，意味着你就算不签一个新单，存量生意也在增长——这也是「交付即经营」最直接的证明。

这个指标做到顶级是什么样？Palantir 2025 年第四季度的财报给了三个数字：净收入留存 139%——老客户自动增长近四成；剩余履约义务（RPO，已签约还没确认收入的合同额）也多了 145%——未来几年不愁没生意做；单季签约总合同额 42.6 亿美元创历史纪录。管理层特意解释过一个细节：139% 这个数字，还不包括最近十二个月新签客户的收入——它纯粹是「老客户的信任在增值」。这家公司早年最被人嘲笑的就是「项目制、没复购」，二十年后，它用同一批客户证明：贴身的交付如果能持续创造价值，续约就不再是销售问题，只是时间问题。（出处见附录 C）

这几类流失风险，各有一道对应的防线：性能与稳定，防质量漂移；有损服务，防成本反噬；上手与培训，防价值蒸发在行为层——使用率随人员流动而衰减；组织维系，防支持者离场；健康分与预警干预机制，是对五类风险的总防线。

5.2 优化系统性能与稳定性

消费互联网产品的规律是：加载慢一秒，留存掉一截。企业系统的性能问题有不同的病理：企业用户对「慢」的容忍度其实高于消费者（他们习惯了老系统的迟钝），但对「不可靠」的容忍度趋近于零。企业系统的输出会进入真实的业务决策——一次错误的库存建议造成的损失，可能抹掉系统一年的收益；错误还会被放大传播——系统错一次，故事会在部门里传三个月；做对一百次，却没人记得。信任要以季度为单位慢慢攒，毁掉它却只要几分钟。

FDE 团队守护可靠性的四道工事。

- **第一道：** 明确服务等级协议（SLA，对可用性、延迟、错误率的承诺），并让它可见。这些承诺不只写在合同里，还要做成客户自己也能看的监控面板。把「系统很稳」从你需要辩白的立场，变成客户随时可查的事实。
- **第二道：** 为人工智能的「概率性」设计护栏。人工智能系统不可能百分之百正确，工程上接受这个现实，产品上管理这个现实：模型拿不准的输出，必须标注出来或转给人工；高风险动作，必须有人把关；模型每次大更新，必须重新过一遍评估，防止老毛病复发——第 4 章建立的评估体系，在这里变成生产护栏的一部分。
- **护栏的样板：** Anthropic 在与金融机构的合作中，把「可审计、可追溯」作为核心设计，每个智能体决策都能回放证据链。在金融、医疗这类场景，可审计性不是加分项，是准入证。
- **防漂移的样板：** 护栏的另一半不是挡错，是持续修对。一家法律科技公司把这件事做成了闭环：系统追踪合同条款抽取的召回率（该查出来的问题里实际查出来的比例），每个漏检自动生成标注样本、进入每晚的微调任务——一个月内，关键条款的召回率从 92% 升到 98%，服务没有中断一天。（出处见附录 C）对抗质量漂移，靠的不是一次性修对，而是让系统每天都在自我修复。
- **第三道：** 值班和响应，要让客户感觉到你一直在。凌晨两点系统挂了，FDE 的反应速度，就是客户对这段关系的体感温度。第 1 章引过那句从业者的铁律：「部署在凌晨两点挂了。你不提工单，不怪别的团队，不回去睡觉。你修好它。句号。」这话的精神要落到机制：值班轮换、事故复盘、以及每次事故后给客户的诚实通报——企业客户能接受事故，不能接受隐瞒。
- **第四道：** 容量与成本的同步规划。用量增长是幸福的烦恼，一旦处理不好，就成了续约时的刺客。性能团队要永远跑在用量曲线前面半步：客户业务旺季来临前，容量、限流、降级预案已经就位。在客户还没感觉到慢的时候，就把慢消灭掉——性能工作做到位，本来就是没人察觉的。

5.3 有损服务——放下不必要的坚持

「有损服务」是互联网产品设计里的一个概念：在极端场景下主动降级，保住核心价值。这个概念在 FDE 语境下有更深的变体——它关乎定制化与标准化之间永恒的撕扯。

背景是每支 FDE 团队都会遭遇的引力：客户在，需求就在；需求在，定制就停不下来。三个月后你回头看，这个客户的部署里长满了定制功能，一半只有三个人在用，维护成本却全压在你头上。这笔账其实比你以为的更紧：按市场行情，一个满负荷的 FDE 也只能同时服务三五个客户，摊下来单个部署的年交付成本约 7.5 万美元起——长尾定制啃掉的每一分工时，都是从这条薄利里割的肉。（出处见附录 C）再放任下去，你就背上了一座「定制债务」——它一边啃你的利润（维护成本吃掉合同收入），一边捆你的手脚（平台一升级就可能踩中定制地雷）。

这里要把两个常被混用的数字口径掰开，不然账会算错。7.5 万美元说的是单个部署摊到的交付直接成本，出自行业自媒体的单源估算，宜作下限看；第 1 章的 38.5 万美元，是头部实验室中级 FDE 的年总薪酬中位数。

按后一个口径自己算一遍：38.5 万薪酬，加上福利、差旅与工具链分摊，一个中级 FDE 的满负荷年成本往 50 万美元靠；除以同时服务的客户（满负荷按三五个摊；3.8 节说过，深度嵌入时只有一到三个），单个部署摊到的人力账在 10 万到 17 万美元之间——这还是没算项目空档期和售前投入的理想状态。两个口径摆在一起，单个部署的真实全成本大体落在 7.5 万到 17 万美元的宽区间里，具体位置取决于公司的薪酬水平和团队的复用率。第 3 章通古兹那条「单份合同 10 万美元起步」的红线，正好卡在这个区间的下沿之上——算下来，它就是盈亏平衡的位置。这笔账怎么从单个部署算到一整支团队，第 7 章最后一节接着算。

「有损」的智慧，是在三个维度上主动做减法。

- **功能维：敢对长尾需求说「不做」。** 判断标准不是需求是否合理（多数需求单独看都合理），而是两个问题：它服务的用户数乘以频率，值不值它的终身维护成本？它能否泛化为平台能力（值的话，进第 7 章的回流通道）？两个都答否的需求，最好的回应是提供变通方案，而不是代码。FDE 不是接单员——接单员文化，恰恰是「法国侍者」模型（第 1 章卡普定的行为标杆：嵌在服务流程里，却有自信与品位把客户引向真正对他好的方向）要反对的唯唯诺诺。
- **承诺维：分级承诺，而不是一律最高。** 不是所有功能都值得四个九（99.99% 的时间在线）的可用性。核心交易链路按最高标准守护，报表与探索性功能则大方接受降级——降级策略（高峰期关闭重计算功能）、错峰策略（重任务夜里跑）、以及事先向客户言明的分级承诺。把可靠性资源集中在命门上，比均匀的平庸可靠更诚实，也更可持续。

- **成本维：用量账单的主动管理。** 5.1 说过「成本反噬」：用量越大账单越高，价值叙事跟不上时，成功反而成为续约障碍。主动的 FDE 团队会在账单变痛之前先行动：给出成本优化方案（缓存、批处理、模型分层——用便宜的模型处理简单请求）、重新设计计价结构（从纯用量转向「平台费加用量」的平滑结构）、以及最重要的——在客户财务负责人问起之前，先把「这笔账单对应的价值账」算给对方看。等到被叫去解释账单，局面就被动了；而主动把账算给客户看的团队，续约谈判时会从容得多。

「有损」的本质，是承认资源永远有限，把有限的资源持续押注在客户真正在意的地方。它与第 2 章「拒绝昂贵的概念验证坟墓」一脉相承。

5.4 引导新用户快速上手

系统激活之后，用户群不会静止：新员工入职、组织调整、新部门纳入使用范围。企业系统的「新用户引导」，是一个永不结束的滚动过程。上手机制设计得好，使用率随时间上升；而设计得差，使用率随最初那批被培训过的用户流失而自然衰减。

企业场景的上手引导，与消费互联网产品有根本不同：消费互联网产品的引导是一次性的自助流程，企业系统的引导是「人传人」的组织工程。三个可复用的结构。

- **分层培训：** 一刀切的全员培训是最大的浪费。而有效的分层是三层：管理员与内部支持者的深度培训（他们未来的角色是内部专家）；普通用户的场景化培训（不讲功能，讲「你的日常三件事怎么用系统做」，30 分钟封顶）；高管的一句话培训（「打开这里，这个数字就是答案」）。分层的精神与 4.4 节一脉相承：每个角色只学与他有关的那部分。
- **「培训培训师」的杠杆：** FDE 团队终将撤场，培训工作必须在撤场前移交。识别客户组织里的热情分子，把他们培养成内部讲师与内部答疑人——给予官方认证、专属的支持通道、以及在高管面前的曝光机会。Anthropic 与 FIS 合作的核心设计就是这条：「转移知识，让 FIS 能独立构建和扩展自己的智能体」。交付做到最后，是教会客户教自己。

两种「人传人」的规模化培训，值得对照着看。西班牙对外银行推 12 万人上线时，靠的不是厂商的培训团队，而是两批内部角色：一张全行的「人工智能先锋网络」，负责在各业务部门做工作坊、挖场景；一批被同事称为「人工智能极客」的高级用户，手把手带身

边人。咨询巨头埃森哲与 Anthropic 的合作则是另一个量级：三万名顾问接受 Claude 的系统培训，组成全球最大的人工智能实践者网络之一——埃森哲再把这支队伍带进自己的客户。（出处见附录 C）结构却是同一个：说到底，厂商真正要培训的只有一类人——会去培训别人的人。

- **文档与自助体系：** 企业文档大多写完就没人看，除非它符合两个标准：以任务而非功能组织（「如何处理一笔异常退款」，而非「退款模块功能说明」），以及嵌入产品而非独立存在（在用户卡住的地方就地出现）。文档既是给客户的交接物，也是给你自己的规模化资产——下一个同类客户的培训体系，七成可以继承。

5.5 组织维系与单点依赖

「支持者离场」是企业续约的第二大死因，值得单列出来专门处理——它的普遍性和致命性，都被严重低估。

单点依赖的形成几乎是自然的：项目由支持者发起，关系靠支持者维系，成功叙事由支持者代言——然后某天他离开了。继任者带着自己的议程上任，你的系统在他的待办清单上排不进前二十，续约季来临，无人替你说话，合同无声死亡。行业里无数「明明用得很好却被砍」的系统，死于这个剧本。

这个剧本的现场长这样：续约评审会上，新上任的负责人翻到系统那一页，问「这是什么」，满屋子没人答得上来——会用的人不敢替预算说话，能说话的人没用过它。合同到期那天，没有争吵，也没有投诉，只是没有人发起续约流程。服务器还在跑，监控面板上一片绿灯。

防御工事要修在三处。

- **关系的网格化：** 从你意识到单点风险的那天起，系统性拓宽关系网：一个支持者之外，至少再发展两条独立的关系线——业务线的日常使用者社群（内部讲师们是天然的节点），以及更高一层的高管赞助人。高层关系不必频繁维护，但要在关键节点（季度汇报、续约前）保持可见。关系网格的检验标准：任何单个人离开，信息通道都不中断。
- **价值的组织化：** 把系统的价值，从「支持者的政绩」改写为「组织的资产」。具体动作：价值数据的定期全员通报（让使用部门自己感受到离不开）、成功案例在客户内部会议的反复讲述（形成集体记忆）、以及把系统嵌入流程文件（当系统被写进标准

作业流程，替换它就得改写流程，成本陡增）。目标是让任何新上任的管理者，在第一周就得出「这系统是这里的固定资产」的印象。

- **离场也要做成仪式：** 支持者离开时，多数供应商的反应是被动哀叹。正确的反应，是把它当成一次关系建设的机遇：为离开的支持者举办一场得体的「功成」仪式（感谢信、成就总结、赠予他可带走的职业资本——比如对外分享案例的资格），同时第一时间启动与继任者的交接——以「帮助继任者快速出成绩」为切入点，而不是「说服他保留我们的系统」。离开的支持者去了新东家，是你下一个客户的潜在入口；善待离场者，是把流失风险转化为获客通道。

5.6 设计健康分与预警干预机制

最后一节，把守护续约的全部动作，收拢为一个系统：客户健康分体系。它的目标是让「关系变质」从一场突然死亡，变成一条早就被监测到的缓慢曲线——你总有时间干预。

健康分的看法，企业客户和消费互联网产品不一样。消费互联网产品盯留存，看日活跃用户数就够了；企业客户，得同时看四类信号：使用信号（每周活跃人数的趋势、关键功能有几成人在用、是真用还是打卡）、价值信号（当初定的业务指标现在如何、投资回报的故事是否依然成立）、关系信号（支持者在职状态、最近 30 天高层接触次数、客户对你团队的响应速度）、商业信号（用量与账单的比率趋势、合同到期时间、竞品动向）。把这四类信号合成一个健康分，跌破警戒线就告警。健康分最大的用处，不是那个分数，是它逼着团队每周把所有客户挨个过一遍。

季度业务回顾：把「我们创造了什么价值」变成每个季度的固定节目。季度业务回顾（QBR，供应商与客户每季度对齐价值与计划的例会）是企业服务最重要的续约工程，却常被办成供应商单方面放幻灯片、客户方低头看手机的例会。好的季度回顾有三个纪律：1）讲客户的语言，而非你的产品语言（「本季度为你们节省了多少工时」，而非「本季度我们上线了什么功能」）；2）让客户的业务方当主角，而非听众（由支持者来讲他团队的故事，你提供数据弹药）；3）以「下一季度的价值计划」收尾。

什么算「讲客户的语言」？教育出版公司 Wiley 拿得出的季度故事是个范本：213% 的投资回报率（ROI）、季节性客服上岗培训提速一半，外加一个意外收获——工单分类数据让它能拿着工单周转时长（TAT）去跟分销商问责。（出处见附录 C）这样的季度回顾，

客户的高管会自己搬椅子来听。把续约拆成每个季度的价值对齐，这样一来，到期那天签字往往只是走一道手续。

预警与干预：用量下滑了，怎么办。消费互联网产品靠推送和邮件唤醒用户；企业客户的「唤醒」却是另一套行规——流失预警与干预，得靠「人加数据」的组合打法：健康分一掉，分级响应立刻启动：轻度下滑（某部门使用减少），对应内部讲师上门回访；中度下滑（整体活跃降三成），FDE 团队启动专项诊断——通常是业务变化、人员变动或质量漂移三者之一，对症处置；重度下滑（濒临停用），高管层介入，坦诚对话「还值不值得继续」——有时答案是体面的退出或降级，保住了关系与口碑，为未来的重逢留了门。

还有一层唤醒，发生在你部署的系统内部。Sierra 发现，智能体最出人意料的强项是谈判挽留：它服务的一家旅游平台，让智能体在用户犹豫退订的「摇摆时刻」主动介入——帮用户厘清套餐权益、探索替代方案、挖出被忽略的价值——套餐保留率就这么提升了 5%。（出处见附录 C）当你的系统开始替客户守住它的客户，你在续约谈判桌上的身份就变了：从一项可以被砍的成本，变成一只会下金蛋的鹅。

到这里，守护续约的五道防线齐了。但只守不攻的企业关系会萎缩——客户组织里，永远有下一个未被解决的问题。下一章，进攻：如何在一个成功部署的地基上，把生意做大。

第 6 章 扩大收入

> 「FDE 模式成立的标志是：每个后续客户的定制量递减。」 —— 鲍勃·麦格鲁

6.1 免费验证的世界

互联网经济把「免费」从噱头变成了战略。FDE 这行也绕不开「免费」这道坎，只是坎的样子不一样：企业人工智能时代，免费的概念验证是最贵的一种免费。

先看这个免费经济体的规模。Palantir 的人工智能平台（AIP）训练营，本质上是把免费验证做成了一条流水线：客户带着真实数据来，一到五天做出能部署的原型，收费为零或象征性。从 2022 年的不足百场起步，场次连年倍增，到 2025 年高峰期日均近 6 场——以每场投入数名顶尖工程师数日计，这是一年数千万美元级别的免费投入。前 Palantir 工程师 Barry 回忆更早期时说得更直白：「我们烧掉数百万美元做客户试点，很多项目利润率字面是负无穷，因为我们是免费做的。」

免费的验证为什么成立？三本账算得过来。

第一本账，算的是获客。传统企业软件靠销售大军获客：差旅、酒局、标书、漫长的谈判，钱花得又多又不可控。训练营换了个打法：不说服客户，让客户自己说服自己——高管亲手点着用自己数据跑出来的系统，胜过一百页幻灯片。

Palantir 的销售周期从九到十二个月压到数周，美国商业收入单季同比增长 137%，免费训练营是公认的主引擎。免费的验证不是纯成本：它是把销售费用换成了工程师费用，而工程师费用能沉淀为产品。

第二本账：风险定价的权力。麦格鲁的建议是，早期创业公司应主动承担风险：「做成了你再付我们。」底气来自对产品力的自信，也来自一个冷静的计算——企业客户对新供应商最大的疑虑是「你们行不行」，免费验证是这个疑虑的溶解剂。当怀疑被溶解，后续的定价权反而回到你手里：客户买的不再是「一场赌局」，而是「一份已被亲眼验证的确定性」，确定性是可以溢价的。

第三本账：失败也要让失败值钱。免费验证必然有失败——这是组合投资的常识。区别在于失败的去向：传统销售失败，留下的是一堆差旅发票；FDE 式免费验证失败，留下

的是对一个行业的理解、一批可复用的组件、一组评估数据。只要你建立了第 7 章的回流机制，失败的验证也在为公司存款。

但免费的验证有一个致命的前提：必须设计好毕业（转正）的那一天——周期上限、验收指标、到期转付费的约定。一旦缺了这些，免费就成了无限期的居留。这就引出下一节。

6.2 免费午餐的终结

免费是手段，收费是目的，转化设计决定了前面所有免费投入的归宿。FDE 世界里，「验证转付费」是最关键的临门一脚，也是行业事故最高发的环节——前面章节反复提到的「概念验证坟墓」，多数不是因为技术失败，而是没有设计「结束」的机制。

从免费到付费的转化，有五个必须前置设计的开关。

开关一，毕业标准先于开工。「设计好毕业的那一天」这条要求，必须在验证项目启动时就白纸黑字写明，而不是开工后再补。Palantir 的训练营把这个设计做到了极致：第 4 到 5 天的日程表上，「演示」之后紧跟的就是「拍板」。转化不是一个事后事件，是日程表上的一个格子。

开关二，免费的边界显性化。客户必须清楚地知道：免费到哪一天、覆盖什么范围、超出部分如何计价。模糊的免费边界会培养「免费是常态」的预期，等到收费日，对方的感觉不是「开始付费」，而是「被宰了一刀」——同样一笔钱，预期不同，体验天差地别。

开关三，让内部支持者成为销售员。这个开关最反直觉：验证成功之后，真正去敲预算门的不是你的销售，而是客户内部那个亲眼见证了价值的支持者。他走进财务负责人办公室的那一刻，手里能拿出什么，决定了这扇门开不开。

财务负责人抬头，问的通常只有三个问题——这东西现在带来了什么？停了会失去什么？明年为什么要多付？你的支持者答得上几句，取决于你在免费期里给他攒了多少弹药。答不上来的会，开完就是一句「明年再看」。

你的工作是把他武装好：一页纸的价值报告（数字、对比、同事证言）、应对财务质询的问答集、以及「如果不继续」的机会成本陈述。内部人推销的内部信任度，外部人却永远够不着。

开关四，价格锚点的提前埋设。免费期就要开始谈论价值——「这套系统这个月为你们释放了约 120 个人时」。价值叙事贯穿免费期，报价单出现时，客户心里已有锚点；而免

费期只谈功能不谈价值，报价单就是一个突兀的惊吓。

开关五，给「不转正」设计体面的出口。不是所有验证都该转正，强扭的转化是毒药合同。对不达标或时机不对的客户，给一个「暂停但保留」的选项：保留数据与配置、约定重启条件、维持轻量联系。企业市场很小，今天的「明年再说」，常常是后年的大单——前提是你把告别做得专业。

五个开关合起来，本质上是一句话：把「免费到付费」从一次惊险的跳跃，设计成一段缓坡。

6.3 按结果收费：客户得到什么，就为什么付钱

FDE 的计价原则很直白：客户拿到多少价值，你就收多少钱。

SaaS（订阅制在线软件）时代的主流计价是按席位（per-seat，按账号数）——为「使用权」付费。这个逻辑在人工智能时代正在被侵蚀：当一个智能体能完成 10 个人的工作，按账号收费就成了笑话——难道为 0.1 个账号付费？于是「按结果收费」兴起，Sierra 按「已解决的会话」收费是最醒目的样本：客户不为软件付费，为「被解决的问题」付费。

按结果收费的进化链有四档，越往上越贴近价值，也越难执行。

第一档，按用量：按 token、调用量、处理量计费。优点是清晰可测，缺点是它追踪的是成本而非价值——用量大可能是价值高，也可能是系统低效。模型接口按量计费，是行业通行的基准线，但应用层公司很少把它作为唯一计价。

第二档，按任务（per-action）：按「完成一次退货处理」「生成一份合规报告」计费。比用量进了一步，计价单位开始有了业务含义。

第三档，按结果：按「成功解决的会话」「追回的一笔坏账」计费。Sierra 的按解决量收费、部分风控公司的按挽回损失分成，都在这档。执行难点在归因——「解决」的认定，需要双方认可的判定标准（这正是第 4 章评估体系的商业用途：技术评估体系，同时是计费基础设施）。

第四档，按价值分成：按为客户创造的财务价值提成——省下的成本、追回的收入、释放的产能。这是最贴近价值的终极形态，也最难：需要客户开放财务数据、需要抗周期的信任、需要极强的价值测量能力。目前只在部分深度绑定的高客单价场景出现。

按结果收费的公司，必须敢把结果公开。Sierra 自己公布的客户数据，构成了一张有趣的成绩单（以下均为厂商自报口径）：物业管理公司 Funnel Leasing，解决率 94%；金融科技公司 Ramp，90%；床垫品牌 Casper，74% 且客户满意度提升超过两成；慧俪轻体，约 70%、满意度 4.6 分（满分 5）；连表现垫底的客户也有 64%。

第三方估算的价位也随之浮出水面：年合同门槛约 15 万美元起，含部署费首年预算常见 20 到 35 万美元，大型客户可到每年百万美元级；据报道，单次成功解决的定价在 1 到 2 美元。换句话说，客户每一分钱都对应一次「问题真的被解决了」——Sierra 敢这么收，是因为它的评估体系能向客户证明「解决了」。计价方式和评估体系，在这里是同一块硬币的两面。（出处见附录 C）

按结果收费的上限，是「结果」本身变成收入。窗帘品牌 Hunter Douglas 的客服智能体跨过了这条线：据 Decagon 公布的客户案例，它有超过 100 万美元的收入，来自完全由 AI 处理、从未转人工的对话——客服部门这个传统的成本中心，第一次拥有了自己的收入归因。供应商案例页的数字要留个心眼，但「客服变收入」这个方向是实的。（出处见附录 C）

选择计价档位时，有一条朴素的判断原则：计价单位越贴近客户价值，你的定价天花板越高，但你的测量与信任成本也越高。

还有一层后果很少有人细算：计价档位对你的财务报表做了什么。按席位订阅，收入稳定可预测，资本市场给的是软件的估值；按结果收费，收入随客户的解决量波动，回款随验收节奏起伏——收入确认的时点、应收账款的天数、现金流的可预测性，全部变差。Sierra 的模式，等于把原本由客户承担的使用风险接到了自己肩上：客户用得少，你的收入就少，而你的工程师成本一天不少。这不是反对按结果收费——它对客户关系的价值，前面已经说过了——而是提醒：选高档计价的公司，现金流储备和成本结构得扛得住收入的波动。定价空间是用波动换来的，先确认自己换得起。

「价值指标写进验收标准」落到合同里，又长什么样？一个能执行的条款，至少写清五件事：

- **指标定义：** 「解决」指终端用户接受回答、且 24 小时内未升级人工。
- **基线与测量窗口：** 以上线前 8 周的均值为基线，按季度测量。
- **判定方与数据源：** 以双方共认的仪表盘为准，而不是供应商的单方报表。
- **争议仲裁：** 数据有分歧时由第三方审计或联合复核，而不是重新谈判。
- **流产条款：** 客户数据基础不达标、指标无法测量时，费用与责任如何分担。

五件里最常被漏掉的是最后一件，出了事最伤的也是它。这些东西只要缺一样，验收桌上的「按结果收费」，就会退化成「按关系收费」。

对中国市场的读者，还要补一句现实主义的注脚：国内企业客户对「订阅制」的接受度至今有限，「买断加实施」「按项目验收付款」仍是主流。FDE 模式在中国的落地计价，往往要中西合璧：按阶段交付验收（顺应项目制习惯）+ 价值指标写入验收标准（注入按结果收费基因）。纯订阅制在这里是理想，而混合制才是活路。

6.4 存量深耕：从一个部门到一张大网

企业市场有一条铁律：最大的收入增长不在新客户，在老客户内部。行业用净收入留存（NRR，见第 5 章）衡量这件事，优秀的 FDE 驱动型公司，净收入留存常年站在 120% 以上——就算不签任何新单，存量收入也自然增长两成。Palantir 的商业故事，本质就是存量深耕的故事：从一个情报小组到整个机构，从一个工厂到整个集团，从政府部门到商业帝国。

存量深耕的打法，业内有个形象的说法：「先登陆、后扩张（land and expand）」。登陆靠前面五章，扩张有三个方向。

- **横向：从一个团队到相邻团队。** 你帮客服部做成了智能工单，隔壁售后部、技术支持部就是你接下来最容易拿下的客户。横向扩张时，最有说服力的证据就在客户内部：同一家公司、同一套数据环境、隔壁部门同事现身说法——这是销售阻力最小的扩张，几乎不需要重新建立信任。Harvey 在律所的扩张就是这个节奏：单一业务组切入，六个月实战验证，横向扩到全所。
- **纵向：从执行层到决策层。** 最初的项目通常服务一线执行者，纵向扩张是把价值链往上传导：为中层管理者做分析与预警，为高管做决策仪表盘（dashboard）。纵向扩张的意义不只是收入，更是安全——第 5 章说过，只被基层喜爱的系统，在预算季没有辩护者；而进入高管视野的系统，才进入了组织的「固定资产」行列。
- **纵深：从辅助工具到核心流程。** 最深的扩张，是让系统从「帮忙的工具」变成「离不开的流程」——从「给建议」到「执行业务动作」，从「可选」到「标准作业流程的一部分」。纵深扩张的每一步，都伴随更大的责任与更高的信任门槛，但它也构建最深的护城河：替换一个工具只要换软件，而替换一个嵌入流程的系统，等于做一次手术。

两张「登陆与扩张」的成绩单，值得对照着看。Harvey 从年利达一家律所起步，到 2026 年，用户已经超过 10 万名律师、1300 家组织，覆盖美国百大律所的多数、500 多个企业法务团队和 50 家资产管理公司，遍及 60 个国家；年度经常性收入（ARR）从 2025 年 8 月的约 1 亿美元，涨到 2026 年 1 月的约 1.9 亿美元——五个月接近翻倍。行业调查显示，68% 的受访律所已在生产环境使用 Harvey 的智能体，深度用户平均每周节省 11 小时。估值随之一年内四连跳：30 亿、50 亿、80 亿、110 亿美元。（出处见附录 C）而 Palantir 那张净收入留存 139% 的成绩单，5.1 节已经讲过了——两张表合起来，说明了存量深耕的逻辑：新合同靠营销推动，收入增长却主要是老客户内部长出来的。

扩张还有一种更安静的形式：吃掉客户预算表上别人的科目。据 Decagon 公布的客户案例，ClassPass 的智能客服上线后，实际自动分流（deflection）的咨询量是预期的 10 倍；随后发生了一件小事——它的 AI 翻译质量超过了原本聘用的本地化供应商，后者的合同到期未再续签。客户的预算没有变大，只是换了主人。纵深扩张到一定深度，你的竞争对手就不再是同行，而是客户预算表上的其他供应商。（出处见附录 C）

这些扩张共用同一套节奏纪律：扩张必须由价值牵引，而不是由销售指标牵引。第 5 章的健康分体系在这里有进攻性用途：使用深度高、价值明确的部门，就是扩张的下一个目标；而客户组织里「看到别人用得好、主动来问」的时刻，是扩张的黄金窗口——此时你不是在推销，是在响应需求。

6.5 看 Anthropic 与 FIS 如何打金融服务这张牌

2026 年企业人工智能市场最有标本价值的一组合作，值得完整看一遍——Anthropic 与金融科技巨头 FIS 共建金融犯罪智能体。

FIS 是全球金融技术基础设施的巨头，服务着全球银行的核心系统。2026 年 5 月，FIS 发布金融犯罪侦测智能体，首批客户是蒙特利尔银行和 Amalgamated 银行。这个智能体做的事：把反洗钱调查从数小时压缩到数分钟——跨银行核心系统自动汇编证据、按风险排序呈交调查官、全程可审计可追溯。

打法上，四个动作环环相扣。

动作一，嵌入而非交付。Anthropic 派出应用人工智能团队与前线部署工程师，直接嵌入 FIS，与 FIS 的专家共同设计。注意：FIS 不是终端客户，而是渠道级伙伴——

Anthropic 的智能体，将随 FIS 的产品进入它背后的数百家银行。这是一单生意，也是一百单生意的入口。

动作二，知识转移作为卖点。官方声明里特意写明：嵌入的目标包括「转移知识，让 FIS 未来能独立构建和扩展更多智能体」。把「教会客户」写进合同——这既是对「供应商锁定」担忧的预防性回应（呼应 5.1 的「供应商戒断反应」），也是一种高级的绑定：当客户的技术栈生长在你的方法论上，分离的成本只会越来越高。

动作三，可审计性作为产品特性。金融合规场景，监管要求每个决策可回放。Anthropic 把「完全可审计、可追溯」做成智能体的核心卖点，而非附加功能——这给所有受监管行业一个模板：别人眼里的合规成本，却可以是你的定价理由。

动作四，生态放大。FIS 案例发布同日，Anthropic 还宣布了面向金融服务的连接器与「拿来即用」模板，以及十余家同类合作——单个案例，立即被抽象为可复制的产品资产。同一时期，市场上还传出了它与黑石集团组建企业人工智能服务公司的消息，据报道规模约 15 亿美元，与 OpenAI 的部署公司正面对垒。（以上出处见附录 C）

这局牌还有一条暗线——首席信息官（CIO）们的警觉。面向 CIO 的科技媒体 CIO.com 在报道中引述了咨询公司战略官马哈帕特拉的提醒：「这个模式里最有结构性的问题，是前线部署的成本到底谁在付——这是 CIO 该问、但大多没问的问题。」高德纳（Gartner）的分析师亚历克斯·科奎罗（Alex Coqueiro）则预测，2028 年前 70% 的企业将因供应商成本与技能空心化，而被迫放弃此类方案。这提醒我们：FDE 模式的收入设计里，藏着一个长期的平衡——你创造的价值，必须持续大于你的存在感带来的成本与依赖。赚「创造价值」的钱，生意能长久；赚「客户离不开你」的钱，却迟早会被客户清算。

6.6 变惩为奖：用量与扩容的定价心理学

好的机制设计，能把惩罚反过来做成奖励。FDE 商业模式里，这个智慧应用于一个微妙场景：客户用量超出预期时，怎么办。

粗放的做法是「惩罚式超量」：合同内用量用尽后，超额部分按惩罚性高价计费，或系统直接限流降速。这在云计算早期很常见，结果是灾难性的——客户为了不超额而主动压抑使用，使用率下降，价值缩水，续约时双输。你惩罚的，正是你最想要的东西：深度使用。

「变惩为奖」的设计思路，是把「超量」重新定义为「成长的勋章」。三个具体手法。

手法一：阶梯价格，越用越便宜。用量越大，单价越低——客户超量时收到的不是罚单，是折扣。这与电信业的流量阶梯定价同理，但要在合同里表达清楚：客户看到的不是「用多了要多付」，而是「用多了单价更低、我们更划算了」。同一笔账单，叙事不同，关系走向不同。

手法二：超量预警 + 主动升舱。系统监测到客户将超量时，不是悄悄计费，而是主动上门：「你们的用量增长很快，照这个趋势，升一档套餐能省 15%。」把计费事件，转化为顾问式销售的机会——客户感受到的是被照顾，而不是被算计。这个动作还有一个隐含收益：它逼着你的团队持续关注客户用量的健康分，与第 5 章的健康分体系天然合流。

手法三：把「省下来」的功劳还给客户。用量优化（模型分层、缓存、批处理）降低了成本，就要把这笔账主动算给客户听：「这个季度我们通过架构优化，为你们省下了约 X 万元。」在客户财务负责人眼里，帮你省钱与等你超量，是两类完全不同的供应商——主动帮客户省钱的供应商，续约时收获的信任溢价远超那笔费用。

定价心理学的底层，是一条朴素的道理：计价结构每天都在向客户讲述「我们是什么样的关系」。惩罚式结构讲的是「我们盯着你」，奖励式结构讲的是「我们和你一起长大」。

6.7 建立价值度量系统以小博大

还有一项基础设施工程绕不开：建立一套贯穿所有客户交付的价值度量系统——把「我们为客户创造了多少价值」从印象变成数据，从数据变成资产。

这套系统把前面那些分散的部件收拢到了一起：「经济性检验」是它的输入（立项时的价值假设），评估体系是它的微观基础（质量数据），健康分是它的运营界面（客户关系数据），计价与扩容是它的商业出口（收入数据）。放到一起，它能帮你干四件事。

对客户，它是续约与扩容时的证据库：每一次季度回顾、每一次续约谈判、每一次升舱建议，背后都是这套系统输出的价值报告——工时节省、错误率下降、处理量提升、对应折算的财务数字。第 5 章说过，价值需要不断重新证明，这套系统就是「证明」的流水线；有数据在手的续约谈判，和凭感觉的续约谈判相比，成交率会高出一截。

对公司，它是交付质量的体检仪：跨客户聚合价值数据，你能回答关乎模式存亡的问题——哪类场景的价值密度最高（应该把销售火力引向哪里）？哪类客户的交付成本最高

（定价或打法要不要调整）？哪些部署在创造价值、哪些在空转（资源要不要重新配置）？如果没有这套系统，你对这些问题的回答都是猜。

对产品，它是回流情报的放大器：价值数据与使用数据结合，是产品决策最硬的依据——哪个功能的价值产出最高（加大投入）、哪个功能无人使用（果断下掉）、哪个场景反复被定制（平台化的信号）。这接通了第 7 章的主题：价值度量系统，本质上是「现场到产品」回流管道的仪表盘。

第四件事是对市场。一个潜在客户第一次记住你，往往不是见到了你的销售，而是撞上了一个数字——「化学品使用减少最高 70%」「调查时间从数小时到数分钟」。这些让整个行业记住的数字，全部来自价值度量系统的积累。案例营销最有效的不是讲故事，是亮数据；而数据不会在谈判桌上凭空出现，它在交付的第一天就要开始采集。

搭建这套系统，给你三条实践建议。其一，从第一天就采集基线——没有改造前的数据，就没有改造后的价值证明，而基线只在项目启动那一刻存在，错过永不再来。其二，指标必须与客户共建——他不认的指标，算出花来也没有谈判效力；在立项会上达成一致的那一刻，指标就成了你们的共同语言。其三，克制指标数量——每个客户三到五个核心指标足矣；指标一多，就等于没有指标。

收入章到此结束，落点只有一个：FDE 模式的收入，归根到底是价值创造的结果。

下一章是全书的「最后一公里」：如何让这一切不依赖英雄主义的个人，而沉淀为可复制的组织能力——规模化复制。

第 7 章 规模化复制

> 「如果你不把现场学习转化为产品资产，你得到的只是一堆定制项目。」 —— Barry，前 Palantir 前线部署工程师

7.1 用可复制性撬动交付的杠杆

互联网增长最理想的状态，是让一个用户带来更多用户——增长引擎从外购变成内生。FDE 要回答的问题也一样：每做完一个客户，下一个客户就该更好做——不靠加人，靠把上一次的经验攒下来。

这个命题是全书的「最后一公里」，因为它是 FDE 模式与咨询业、外包业之间的最后一道防线。把规模化这一环从前面讲的打法里抽掉，结局是什么？一支能打硬仗的精英团队，接一个客户、交付一个客户、赚的钱覆盖成本、然后接下一个——收入增长与人数增长完全线性。恭喜，你重新发明了咨询公司。

Barry 在回忆录里把这个分界讲得最狠：「如果你不把现场学习转化为产品资产，你得到的只是一堆定制项目。」麦格鲁的表述是另一个角度：FDE 模式成立的标志，是「每个后续客户的定制量递减」——如果你做的定制工作不随客户数递减，你就不是在做 FDE，是在做按人天（人头×天数）计费的外包生意。

可复制性的杠杆有三级，一级比一级难，也一级比一级值钱。

- **一级杠杆｜知识复制（打法手册）：** 把「怎么做」写成文档——打法模板、检查清单、培训材料。它把对个人经验的依赖，变成对组织记忆的依赖：新人上手周期，从一年压到三个月。这是最基础的杠杆，多数团队止步于此。
- **二级杠杆｜组件复制（工具与代码）：** 把「做过的」变成「能直接用的」——集成连接器、评估框架、部署模板、行业数据模型。下一支团队的起点不再是零，而是上一个项目的肩膀。4.7 节讲的自动化资产，就属于这一级。
- **三级杠杆｜产品复制（平台能力）：** 把「反复出现的定制」，抽象为平台的标准能力——从此这个需求，不再需要任何工程师到场。Palantir 的本体（ontology，能让模型在企业数据上干活的语义层）、Decagon 的智能体操作流程（Agent Operating

Procedures，AOP）、Sierra 的智能体平台，都是三级杠杆的产物。这一级杠杆改变的是整个生意的成本结构：前两级杠杆降低交付成本，第三级杠杆消灭交付成本。

而这三级杠杆的递进，就是麦格鲁那句「砾石路与高速公路」的完整版：FDE 修砾石路（解决眼前问题），组件团队铺碎石路基（让下一辆车好走），平台团队浇高速公路（从此人人可通行）。

杠杆做对之后，账面上的差距是数量级的。一线交付复盘的对照口径（完整版本见附录 A）：常规做法下，首个智能体上线要 8 到 12 周，上线后每个客户每月消耗 30 到 50 个 FDE 小时，一名 FDE 同时只能支撑 1 到 3 个客户；把数据地基与交接做对的团队，可以压到 1 到 3 周、2 到 5 小时、8 到 12 个客户。所以这一章的复制工程与第 5 章的续约工程，从来都是同一台机器的两个侧面——约三个月后仍保持健康的交付只有三到四成，最常见的死法是「FDE 走了，智能体也漂移了」。

7.2 坏事传千里——失败部署的复盘传播

把事故变成资产，是组织的稀缺能力。FDE 组织对失败的态度，决定了它能否拥有复制能力：失败是复制的原材料，掩盖失败的组织，就等于把金矿当垃圾埋掉。

先说清楚一件事：失败在 FDE 模式里不是意外，是必然。Barry 的回忆毫不留情：「失败多得很——大把大把的时间、金钱和差旅费，烧得轰轰烈烈。」Palantir 干脆把接受失败写进了制度：试点就当风险投资来打，多数会死，死完必须复盘。Barry 特别提醒：对那些撞了车的项目、白忙一场的团队、累垮的工程师，也要说声谢谢——教训是他们买单买来的。这一点最难学，因为工程师的天性，是庆祝成功、藏着失败。

把失败转化为组织资产，需要三个机制。

第一是复盘的无罪化：复盘会议的唯一规则是对事不对人。一旦复盘与绩效挂钩，人们就会开始修饰失败——而修饰过的失败，毫无教学价值。无罪化不是不追责，而是把「责」重新定义为「是否诚实完整地提取了教训」。一个诚实复盘的失败项目负责人，对组织的贡献，可能大于一个侥幸成功的。

第二是教训的结构化：复盘产出不能是「下次注意」式的感慨，必须是可检索、可执行的知识——这个客户类型的高危信号是什么？哪个集成假设被推翻了？哪个评估指标设计错了？

这些问题的答案，应该进入打法手册的「雷区」章节、进入新项目的尽职调查（尽调）清单，让下一次团队的进场尽调自动携带历史教训。失败的教训只有进入流程，才会被下一个团队真正用上。

第三个机制最反直觉：失败的适度外传——把失败教训脱敏之后对外分享。3.5 节讲过，Barry 那篇「自曝家丑」的文章，成了 Palantir 模式最好的布道。那篇文章里没有成功学，通篇是大把烧掉的时间、金钱和差旅费，是一个个撞了车的项目；恰恰是这种写法，让读者一眼认出这是战壕里出来的人——企业客户听过的展台话术太多了，他们能分辨什么是话术，什么是真相。一个敢于公开解剖失败的团队，传递的信号是「我们已经为这个行业的弯路付过学费了」——这正是企业客户最想听到的。

三个机制说完，把一个失败的完整解剖走一遍。下面这个剧本，部件全部来自本书前面的真实案例，拼法是这个行业的常见死法。

一家制造企业的人工智能项目，死于第 9 个月。第 1 个月，立项就带病：集团发文要搞人工智能，信息部门牵头，业务部门无人署名——2.3 节的「无主之地」信号，供应商看见了，装作没看见，因为合同额诱人。第 2 到 4 个月，验证期在样例数据上跑出了漂亮指标——客户的安全审查不肯放真实数据，「只许看、不给碰」，第二条高危信号也亮了，项目经理选择先把东西做出来再说。第 5 个月，演示会大获成功，验收通过。第 6 个月起，真实数据接进来，模型开始对着三年前的编码规则给出自洽但错误的答案；业务部门本来就没人认领这个项目，现在更没人愿意收拾。第 7 到 8 个月，使用率阴跌，例会上开始有人问「这东西到底值不值」。第 9 个月，预算季，合同被连根拔掉——没有事故，没有争吵，只是没有人替它说话。

这种项目的复盘纪要，死因一栏通常写着「客户组织不成熟」。这是修饰过的失败。按 2.3 节的清单诚实复盘：高危信号三条亮了两条，每一条在签约之前就看得见。真正的教训只有一句：这个项目的死期是第 1 个月就定好的，后面八个月只是走流程。解剖做到这个深度，下一次的尽调清单上，「业务负责人署名」和「真实数据到位」，才会从建议变成硬门槛。

7.3 借势而上：站在行业的风口上

借外部事件的关注度为自己所用，是市场工作的经典手法。FDE 赛道的团队正站在一个史诗级的风口上，而风口的窗口期不会永远敞开。

2025 到 2026 年的行业风口，由几股气流汇合而成：1）《生成式人工智能的鸿沟》报告让「95% 失败率」成为每个首席信息官（CIO）的心病——客户的心病，就是你的机会；2）OpenAI 成立部署公司，Anthropic 被曝组建同类合资公司，两则消息一周之内前后脚传出，把「部署」推上财经头条——巨头教育市场的钱，全行业受益；3）FDE 岗位一年 729% 的招聘增长，让这个角色成为科技媒体的常客——连「要不要招 FDE」都成了一种企业焦虑。这三股气流共同制造了一个罕见局面：客户主动在找你，讨论你想卖的东西。

借势的第一层动作是「定义议题」。风口的注意力是公共资源，谁的内容定义了议题，谁就收割了注意力。风投机构 a16z（安德森·霍洛维茨基金）用一篇行业雄文为整个赛道定了调；OpenFDE 这样的社区，在聚合从业者的同时也在聚合定义权。对具体公司而言，想抢下议题的定义权，最好的办法是把压箱底的东西公开出来：打法手册、失败复盘、价值数据——独家的一手认知，是议题定义权的唯一货币。

第二层动作是「绑定行业议程」。企业人工智能采购的最大驱动力，正在从「效率」升级为「生存」——首席执行官们在董事会上被追问「我们的人工智能战略是什么」，这种焦虑会传导为预算。把你的方案翻译成「董事会语言」：不是「我们能帮你优化客服流程」，而是「我们能让你在下次董事会上，拿出可量化的人工智能成绩单」。同一套交付，锚定的议程层级不同，触达的预算池也就不同。

第三层动作是「在风口期建立资产，而不是消耗风口」。风口会停。报告掀起的讨论会冷却，媒体的热词会过气。

所以，风口期最宝贵的不是多签几个单，而是把风口带来的注意力，转化为长期资产：灯塔案例、行业方法论的话语权、可复制的交付体系。风口停的时候，手里有案例、方法论和交付体系的团队才站得稳。

7.4 构建跨客户的传播循环

不靠广告、靠口碑的传播，在 FDE 这行主要通过三个圈子发生。

圈层一：从业者社区。FDE 是一个正在形成中的职业共同体——2025 年它还散在各家公司内部，到了 2026 年，已经有了 OpenFDE 这样的专门社区、专门的薪酬报告、专门的面试指南。职业共同体的形成，对先行者是天赐红利：为共同体贡献内容（方法论、

工具、数据），你的名字就会与这个新兴职业绑定。概念的传播路径历来如此：先行的布道者，最后都成了概念的代名词。

圈层二：客户的行业社区。每个行业都有自己的小圈子——银行家的论坛、律所合伙人的年会、医院管理者的峰会。3.3 节讲的信任红利，主要在这些封闭的圈子里流动。

进入这些圈层的门票不是广告，是被邀请演讲的客户支持者——你的工作是让他有值得讲的、讲得好的。客户在他的行业峰会上替你做的十分钟分享，胜过你在行业展会的一百天展位。

圈层三：生态伙伴网络。3.7 节讲过生态捆绑的获客价值，这里强调它的传播属性：云厂商的解决方案架构师、咨询公司的顾问、集成商的项目经理，这些人每天都在不同的客户现场流动——他们口中的「靠谱团队」名单，是传得最快、最准的宣传渠道。经营这个圈层的要点是让他们「赢」：让生态伙伴在与你的合作中获得业绩、案例与口碑。帮别人成功的人，会被整个网络记住。

生态捆绑在 2026 年还长出了更深的形态：人工智能公司把前线部署团队直接嵌进伙伴体内。编程智能体公司 Cognition（Devin 的开发商）与全球信息技术服务巨头 Cognizant 的合作是样板——一支嵌入式的前线部署工程师团队进驻对方，负责项目筛选、工程师带教与成效度量，借 Cognizant 几十年攒下的客户关系，把自己的产品规模化铺进医疗、金融与保险行业的大企业。Databricks 把专业服务整体改组为前线部署团队时，同样把数百家伙伴组成的全球网络列为规模化的四要素之一。走到这一步，伙伴网络传播的已经不只是口碑，而是你驻在对方办公室里的人。

三个圈层合起来的飞轮（自我强化的正循环）是：交付创造故事，故事在圈层中传播，传播带来新机会，新机会创造新故事。这个飞轮与传统营销的区别在于，它的燃料全部是「已交付的价值」——口碑传播的本质，是让过去的每一分努力，都让未来的获客更容易。

把视野再拉高一格，2026 年 5 月前后脚落地的两笔交易，把「规模化路径」这道选择题摆上了台面。Anthropic 与黑石（Blackstone）、Hellman & Friedman、高盛宣布组建 15 亿美元的独立企业服务公司，专门把 Claude 部署进中型企业；一周后，OpenAI 宣布成立部署公司（OpenAI Deployment Company）——19 家投资方、40 亿美元初始投资，并收购英国应用人工智能咨询公司 Tomoro，第一天就带来约 150 名有经验的前线部署工程师与部署专家。连同 AWS 向认证咨询伙伴开放方法论与生产标准的做法，行业实际上给出了三条路径：自己招人守住控制（Palantir 路径，招聘门槛就是天花板）、用资本直接买产

能（OpenAI、Anthropic 路径）、认证伙伴放大（AWS 路径，用标准换规模）。三条路径共用一把试金石：部署第 N+1 个客户，是否比第 N 个更便宜——答「是」的在做平台化，答「否」的，只是换了个股权结构的外包。

7.5 产品内置的裂变因子

好的产品会自己传播——邮件签名里的一句「来自某某邮箱」，就是最早的自传播设计。FDE 交付的系统，也可以预埋「内置的扩散机制」，让系统自己成为扩张的推销员。

因子一：可见性设计。系统的输出物天然会被传阅——周报、分析结果、审批流。让每一份输出物携带「出身」：报表页脚的「由某某系统生成」、异常预警里的处理入口链接。

当一份漂亮的分析报告被转发给隔壁部门总监，看到的人自然会问「这是什么做的」。可见性设计的精髓，是让系统的价值在日常工作流中自我展览，而不是靠专门的汇报。

因子二：协作式准入。设计需要跨角色协作的功能：调查官标注的案例需要主管复核，分析师的报告需要业务方确认——每个协作动作，都在向一个新用户自然地介绍系统。协作带来的新用户不是「被推广的」，是「被工作本身带来的」，接受度完全不同。

因子三：自助化的探索层。在核心交付之外，留出一个低风险的练习区（沙盒）：客户员工可以自己在里面试试「让人工智能帮我做个分析」。沙盒里长出来的用法，往往是最有生命力的扩张线索——某个部门员工自己玩出来的场景，比销售去推十次的场景更有根。让全员帮你做发现，而 FDE 团队专注攻坚。

两个 2026 年的新产品，把这个因子做成了明牌。Sierra 在 3 月推出「代笔人」（Ghostwriter）：客户上传自己的标准作业流程、历史对话记录，甚至一张白板照片，或者用大白话描述一个目标，系统自动生成一个生产可用的智能体，覆盖语音、对话、邮件三个渠道，三十多种语言——客户的业务人员第一次可以「自助造智能体」，前线部署团队则从「每个都亲手做」变成「验收和兜底」。Decagon 的智能体操作流程更早走通这条路：客户的客服主管用自然语言写流程（「超过 30 天的退货，先查会员等级，符合政策就办」），不用等工程师排期——据报道，Chime 的解决率做到 70%，多邻国的问题分流率达 80%，健身平台 ClassPass 的客服成本下降 95%。（出处见附录 C）当客户开始自己建造，你的角色就从建筑队变成了设计院。

因子四：跨边界的价值证据。每个季度自动生成「本季度价值报告」，让它自然流向客户高管层。这份报告既是第 5 章的续约工程，也是纵向扩张的敲门砖——高管看到的每一个数字，都在为「这套系统还值得在哪些地方用」埋下伏笔。

全员推广本身也是一场工程，它的产能瓶颈同样需要被解掉。Booking.com 把企业搜索平台推广到 14,000 名员工，是该公司历史上第一个全公司级采用的人工智能平台；支撑这场推广的培训与宣传物料也跟着提速——单条推广视频的制作周期，压到了原来的四分之一。系统交付的价值，转身就成了系统自己的推广产能。

内置因子的共同原则是：想让别人帮你传播，先得让人家实实在在用出了价值。任何为传播而传播的设计，在企业场景里都会被识别为「厂商的小动作」而遭反感。企业市场没有病毒式传播，只有「口碑的链式反应」——慢，但每一步都坚实。

7.6 规模化中的组织政治把握

规模化这条路上，卡住你的通常是人的地盘，技术反而好解决。

当系统从一个部门扩展到十个部门，你实际上在重写组织内的权力版图：信息垄断被打破（某个部门独占的数据，现在全公司可见）、审批权被压缩（人工智能秒回的事情，不再需要三天的层层签字）、专业壁垒被夷平（资深员工的「独门经验」被编码进系统）。每一个变化，都在制造沉默的敌人。

应对组织政治，FDE 老兵们沉淀了三条心法。

- **心法一：扩张必须带着「当地人」。** 每进入一个新部门，先找到那个部门里的「地头蛇」——最有声望的资深员工，邀请他参与适配过程：「这个场景，你们部门有什么不一样？」他的答案会让方案更准，他的参与会让方案在他的地盘畅通无阻。人们不会反对自己参与建造的东西——这是组织行为学里少有的可靠杠杆。

Harvey 在顶级律所的扩张，是读懂组织政治的经典教案。所内同时存在三层人群：合伙人——买单的人，但不亲自干活；律师助理——干活的人，但不签字；知识管理律师——两边都不沾，却同时看得见两边的盲区。通用的问卷调研会把三层人压成同一组下拉选项，得到一堆失真的平均分；Harvey 的前线部署团队则为三层人分别设计访谈和演示：给合伙人看「这单生意多赚钱」，给助理看「这周少加三天班」，给知识管理律师看「全所的隐性经验终于有地方沉淀」。一个部署要同时让三层人都觉得自己是受益者——一旦缺了任何一层，系统就会在对应的环节被悄悄搁置。（出处见附录 C）

- **心法二：让每一步扩张都有「受益者」。** 扩张前算清楚每个相关部门得到了什么：被穿透了数据壁垒的部门，得到了全局可见的分析能力；被压缩了审批权的部门，得到了例外事项的聚焦权（人工智能处理常规，人专注疑难）。扩张前先想清楚每个相关部门能得到什么，并且讲清楚——想不清这一点的扩张，在对方眼里就是抢地盘。
- **心法三：永远给「旧秩序」留体面。** 被系统替代的流程，曾经也是某个能干的人设计的；被编码的经验，曾经也是某个专家的毕生心血。在推广话术里，对旧秩序的尊重，是对旧秩序维护者的尊重：「这套流程在过去十年支撑了公司，现在我们要把它十年的经验，放大一百倍。」给旧秩序留足体面的人，新系统上线时会少一半暗中的抵抗。

三条心法说的都是客户的组织。规模化路上还有一层政治，藏在你自己公司内部：销售与交付的权责。

这层政治的根源是激励错位：销售的奖金挂在合同额上，交付的口碑挂在结果上——签单的冲动和交付的纪律，天然打架。处理它靠三张成文的表。

- **分工表：** 销售对「签不签」负责，交付对「做不做得到」负责；定制需求的否决权归交付，但否决必须附替代方案——如果只许说不、不给路，交付就会变成销售眼里「挡生意的人」。
- **分钱表：** 首单佣金归销售，续约与扩购的奖金则必须与交付团队共享——5.1 节算过，续约的成本趋近于零，价值全部来自交付质量；续约奖金如果全归销售，公司就是在用钱包告诉所有人「签下比做好重要」。
- **背锅表：** 交付事故按「承诺是谁做的」划责——销售过度承诺惹出的事故，记进销售的考核；交付自己挖的坑，记在交付头上。

三张表都不用复杂，但必须成文；不成文的权责，会在订单最多、团队最大的那一年集中爆炸。

7.7 用打法手册沉淀复制效率

一级杠杆里，打法手册的工程细节最容易被低估——多数团队的打法手册，写完那天就是阅读量最高的一天。

- **手册的第一属性：长在流程里，而不是躺在知识库里。** 一个只能在文档系统里被搜到的手册，等于不存在。有效的做法是把它嵌入工作流：新项目的尽调阶段，系统自动弹出该客户类型的尽调清单；部署阶段，检查清单是任务系统里的硬关卡；复盘阶段，手册更新是复盘的必填产出。知识不嵌入流程，就只是考古资料。这条路的尽头，是流程开始自己执行手册：Ramp 的工程团队让人工智能助手先与客户经理做多轮澄清、再起草需求文档，据其披露，省下了约两成的需求界定时间。
- **手册的第二属性：按「场景」而非「功能」组织。** 失败的手册按公司组织结构写（「交付流程」「技术规范」），而成功的手册按客户场景写（「金融反洗钱场景打法」「制造业排产场景打法」「律所知识库场景打法」）。每个场景手册包含固定七件套：一、典型痛点与契合检验要点；二、尽调清单（这个场景特有的高危信号）；三、数据与集成的已知雷区；四、评估与验收的指标模板；五、变革管理的角色地图；六、计价与扩容的参考结构；七、历史复盘的教训库。Harvey 在律所场景的打法能复制到普华永道、佳利，靠的就是把第一个客户（年利达）的经验，写成了可继承的场景资产。
- **手册的第三属性：它必须「活」——有负责人、有版本、有折旧。** 每个场景手册指定一个负责人（通常是最资深的那支交付团队），每个项目结束后强制回看更新，每半年做一次整体检修——删掉过时的、合并重复的。没有负责人的手册，一年后新人照它做，会踩进早已修掉的坑——这比没有手册更糟。

手册与二级杠杆（组件库）、三级杠杆（平台）的关系是：手册管判断（什么情况怎么办），组件库管手艺（能直接拿来用的工具），平台管能力（不需要人到场的功能）。三者齐备，一支新组建的 FDE 小队，才能在两周内达到老团队八成的战斗力——这就是复制的真实含义。

这三样解决的都是「物」的复制。FDE 模式里最稀缺的资产却是会下现场的人，而这种人在招聘市场上几乎不存在——设计过 Palantir 轮岗培养计划的 Vinoo Ganesh 把话说得很绝：「你没法靠招聘建出真正的 FDE 组织，只能养出来。」毕竟，传统工程岗位刻意把工程师和客户的现实隔开，现场判断力在办公室里长不出来。

「养」不等于不挑苗。Palantir 面试的招牌环节，是丢给候选人一个刻意模糊的真实问题——比如「设计一个把城市紧急呼叫路由给正确响应者的系统」——让他当场拆解，几乎不写代码；按候选人的普遍反馈，最常见的失误是把这场面试当纯软件工程师面试来准备。它筛的，是能对着一团迷雾先拆出结构的人。

养的方法则是轮岗：把工程师成批送进真实客户部署，先后 250 多名工程师经这条路上了现场，校友如今散在 OpenAI、xAI、Anduril 等一众公司。这个培养计划最关键的一处改进，是把轮岗的岗位与工程师回去要造的产品对齐——先当用户，再当建造者。对没条件搞轮岗的公司，他留下一份六个月的自助培养清单：

1. 头两周旁听客户电话，记的不是功能请求，而是「这个人一天是怎么过的」；
2. 第二个月端到端认领一次客户事故；
3. 第三个月去现场，找一个一天之内能修的小痛点、当周修掉；
4. 最后几个月，维护一份周更的《我从客户身上学到的东西》。

这份清单同样是一份打法手册——一份培养人的手册。

「养」的起点是「招」，苗从哪来，先有一张地图。从业者经验里转化率最高的三个池子：1）大厂的解决方案架构师——懂技术、见过客户，缺的是对结果负责的经验；2）咨询公司的技术顾问——懂业务、能讲，缺的是动手；3）客户侧的明星用户——最懂痛点，挖过来就是「回声」的天然人选。反而要慎用纯后台研发，不是能力问题，是他们离客户的现实太远，现场判断力得从零长。

招进来之前的最后一道筛子，是 1.7 节那道「问题拆解」面试。同一道题——「某银行合规团队每天人工核对三万条交易告警，九成是虚惊，你怎么办」——好坏答案的差距长这样。危险的答案直奔技术：「上个分类模型。」三句话就露了底：他没问告警怎么产生、误判的代价谁承担、合规官信什么。而好的答案从提问开始：虚惊的判定规则是谁定的？漏放一条真告警谁负责？现有流程哪一步最慢？然后给出最小切口——「先不做判定，只做预排序，把最像虚惊的排到最后，人先核对高危的」；最后主动交代取舍——「误报还在，但人力省一半，而且每条建议都可追溯」。

招到人、养出人，最后还要回答一个组织设计问题：派出去的最小作战单元是什么？答案不是一个人，而是「一对」。把各家的一线组织摆在一起看，会发现一个有意思的「组织指纹」：Palantir 的组合是「三角洲」（Delta，动手建造）加「回声」（Echo，读懂客户）；Anthropic 是 FDE 加参与经理（Engagement Manager）；OpenAI 是技术部署负责人加 FDE。注意一个不对称：构建者的头衔全行业几乎统一了，协调者的头衔却五花八门——这本身就是行业仍在成型期的信号。所以看一个团队是不是真在做 FDE，别只数它招了多少工程师，看它有没有配上协调者：能对结果负全责的，从来不是孤胆英雄，而是这「一对」。

7.8 产品化的策划与打磨

三级杠杆的登顶之作，是把现场经验结晶为产品。

先看最壮观的成功案例：Palantir 的 Foundry。今天这个年收入数十亿美元、支撑公司数千亿市值的平台，其核心组件诞生于苏黎世、休斯顿、圣保罗、图卢兹、巴库等天南海北的客户现场——各地前线部署工程师为解决本地问题各自造的工具，经过数年的自然选择与平台化收编，最终长成了统一的产品。Barry 回忆这个过程「完全自下而上」：2014 年公司的产品战略字面就是「强烈的观点，松散地持有」，任凭现场百花齐放，活下来的收归核心。

2026 年还有一个更极端的注脚，来自 OpenAI 自己的前线部署团队。随着 Codex 能力跃升，这支团队做了一次「自我清理」：废弃为单个用例编写的独立终端与命令行工具，把约九成定制代码，转成基于 Codex 原语的通用插件与技能。效果直接写在部署速度上：2026 年 8 月，一个用例的初始部署花了两个月，扩展到第二个团队只用了一天。产品同样从现场长出来：企业语音产品 OpenAI Presence，是从 Swarm 与 Agents API 的实验里演进而来的，其投资方旗下公司过半的电话已交由人工智能处理。（出处见附录 C）连自己人都在亲手消灭自己的定制代码——这是对「产品化」最不含糊的表态。

这个过程不是田园牧歌，它有严苛的内在逻辑，可以提炼为「产品化四问」。

- **第一问：** 这个现场解法，是「个性」还是「共性」？判断标准：同类问题是否在三个以上客户那里独立出现过？三个客户的相同痛点是产品信号，而一个客户的特殊要求可能只是怪癖。Palantir 的经验法则，是让「多支前线部署团队的重复造轮子」自然暴露共性——当三支团队各自造了相似的东西，平台团队就知道：收编的时机到了。
- **第二问：** 泛化的代价，是否小于收益？麦格鲁提醒过：现场代码「快而糙」，把它泛化的成本「往往高于写第一版」。产品化决策必须算这笔账：泛化投入，对比未来若干个客户省下的定制成本。算不过来的，留在组件层，不要强行平台化——平台里的每一个能力，都是终身维护承诺。
- **第三问：** 它能否被「不会写代码的人」使用？现场工具的使用者是工程师，平台能力的使用者是所有人。产品化最难的一步，往往不是技术而是交互：把工程师的脚本，变成业务人员能配置的功能。Decagon 的智能体操作流程（让运营用自然语言定义流程）就是这一步的样板——抽象层一旦选对，使用人群就能扩大百倍。

- **第四问：** 收编之后，现场还有空间吗？这是最微妙的平衡：平台收编得太多，现场团队就退化为配置员，FDE 模式的灵魂（现场创造力）就死了；收编得太少，规模化的杠杆又起不来。Palantir 的解法，是保持「前线对基地的激进授权」——高层只定目标，打法归现场。平台把常见的事变容易，现场团队才有余力去啃那些原本做不成的硬骨头。

四问答完、平台建成，最后一道纪律才刚出场：扩大开放的顺序。拉美电商巨头 Mercado Libre 把它做成了教科书。这家公司的内部平台向 17,000 名开发者开放自助：任何团队用节点和技能拼装人工智能应用，内置测试，但永远看不到源码——认知负荷压到最低，安全的赌注全押在平台护栏的质量上。

扩大开放的节奏则慢得近乎保守：先把人工智能用在错误成本低的任务上——商品编目、翻译、评论摘要——光编目一项，两年内可处理的商品量提升一百倍；用低风险任务攒够数据与组织信任之后，才让平台去碰涉及金钱的客服争议调解，而且先在单一主要站点放开约一成，人类保持副驾驶与升级的通道。它的首席运营官说得直白：编目和摘要「有趣也有用，但不是变革性的」，真正的变革是让系统自主处理复杂、高风险的活。只是那样的系统，得先用低风险任务的年头喂出来。

怀疑这套飞轮的人，最有力的反驳永远是毛利率（收入减去直接成本后剩下的比例）：派人下现场，毛利怎么可能高？历史给过三次回答。ServiceNow 上市时毛利率 63.2%，被嫌「太像服务公司」——十年后 79%，市值近 2000 亿美元；Workday 上市时 54.1%——如今 76%；Palantir 上市时 79%，2025 年已达 82%，比绝大多数「纯软件」公司还漂亮。

风投机构 Insight Partners 的总结一针见血：这些在平台切换期重仓实施与服务的公司，恰好是在所有人朝反方向跑时坚持下来的公司。秘诀在于毛利率能逐年往上爬——随着平台沉淀加厚，客户自己能干的实施工作越来越多（训练营模式让大量集成工作由客户侧完成），服务成本被持续摊薄——开局都不好看，往后却越跑越轻。（出处见附录 C）

反面的标本同样现成。Bessemer Venture Partners 拆解过印度的信息技术服务业：行业老大塔塔咨询服务公司（TCS）过去十年营收翻倍，员工数也几乎翻倍，人均营收却停在约 4.9 万美元原地不动——增长依然完全靠堆人。不完成平台化切换的服务公司，十年后只会是一家更大的服务公司——「重新发明咨询公司」的行业级版本。

毛利率之外，还有一把更直观的尺子：人均营收。把三家摆在一起看——TCS 还是那约 4.9 万美元；埃森哲 2025 财年约 8.9 万美元（697 亿美元营收、77.9 万名员工）；

Palantir 2025 年约 100 万美元（44.8 亿美元营收、4429 名员工）。从外包到咨询再到 FDE，人均产出差出一个数量级。

创业公司一侧，倍数更直白：Harvey 约 1.9 亿美元的年度经常性收入，撑着 110 亿美元的估值，约 58 倍——市场显然在按软件、而不是按服务给它定价。这个定价五年后才知道对不对，但至少说明：资本市场愿意为「人均产出向软件看齐」的交付模式，付软件的倍数。这组对比，也是对「FDE 是不是咨询换皮」最硬的一个回答：是不是，不看头衔，看人均产出和毛利率往哪个方向爬。（出处见附录 C）

把镜头从巨头拉回地面，算一笔 10 人小队的粗账——这是本书的推算，每个假设都摆出来，你可以换成自己的数字重算。成本端：按 5.3 节算过的口径，美国市场一个满负荷 FDE 的年成本往 50 万美元靠，10 人就是 500 万美元；中国市场按第 1 章的薪酬水平折算，成本端大几百万到上千万人民币。产能端：3.8 节算过，一支小分队同期最多高质量交付两个项目，10 人拆成三支小队，一年流转下来，大约深度服务 10 到 15 个客户。

收入端：深度服务项目按单合同 10 万到 30 万美元计，一年 10 到 15 个客户，收入 100 万到 450 万美元。和美国口径 500 万美元的成本摆在一起，头两年多半盖不住；换成中国口径的成本，才勉强打平。这就是 FDE 生意头两年的常态：纯人力期，毛利率两到四成，弄不好是负的。

扭亏的杠杆只有三个：复用率——组件与平台，把单项目工时往下压；计价档位——把客单价往上抬；续约率——让第二年的收入不再需要第一年的成本。三个杠杆各抬十个百分点，毛利率大致能从 30% 爬到 60%——前面那组数字里 ServiceNow 们的十年，爬的就是这三个杠杆。麦格鲁那句「每个后续客户的定制量递减」，说的正是这台机器的转速表：定制量递减的速度，就是人力生意变成软件生意的速度。

最后，讲一个「赌对了方向」的故事。2022 年底 ChatGPT 横空出世，所有人都在猜 Palantir 怎么办。桑卡尔下了一个后来被反复引用的注：大模型会迅速商品化，价值会向两端聚集——一端是算力芯片，另一端是本体。他把公司大半工程师压到人工智能平台 AIP 上。结果：Palantir 股价从他接手首席技术官时的个位数，一路涨到 200 美元上下；美国商业收入连续多个季度三位数增长。这个赌注的逻辑，正是全书开篇的那条主线：模型越强越便宜，「把模型接进现实」的那段工程就越值钱。（出处见附录 C）

产品化四问的终点，是对整个 FDE 商业模式的收束：现场替你探测真实世界，平台把探测到的东西放大，最后沉淀进产品，持续生息。这是个互相喂养的循环：现场让平台更扎实，扎实的平台让现场敢接更远的活。这个飞轮一旦转起来，你就不再是一家「派人干

活」的公司，而是一家「把世界的复杂性持续转化为软件资产」的公司。这种公司的简历，前面那笔毛利率账已经替它写好了。

下一章是完整案例集：回到这些方法论最初的现场。

第 8 章 完整案例集

前面七章的方法论，全部提炼自真实的实践现场；这里把它们完整还原：五组案例，从模式的发明者到最新的继承者，从硅谷到中国。

8.1 Palantir：用二十年把一个「笨办法」炼成护城河

创业：为「不能问的用户」造软件

2003 年的硅谷，还躺在互联网泡沫的废墟里。彼得·蒂尔拉上一群斯坦福出身的年轻人，要做一件听起来疯狂的事：为美国情报机构造数据分析软件。公司的名字取自《指环王》里的真知晶球 Palantir——能看见远方的石头。

这件事最难的地方，从第一天起就是用户。鲍勃·麦格鲁多年后那段广为流传的回忆，把困境讲得最清楚：「给间谍做软件的一个挑战是，我不认识任何间谍。你大概也不认识。就算你碰巧找到一个间谍，问他『你平时到底怎么工作的』，他通常也不会告诉你。」

软件工业的全部方法论——用户访谈、问卷调查、可用性测试——在这个用户群体面前集体失效。创始人之一斯蒂芬·科恩的应对笨得可爱：做一个演示样品，拿给情报机构的人看，问他们觉得怎么样。答案是毫不留情的：「这东西太糟了，跟我们做的事毫无关系。」科恩没有撤退，而是追问：「那你们希望它哪里不一样？」然后把每一条意见记下来，回去改，改完再送上门。循环往复。

这个循环就是 FDE 的基因片段，它包含两个日后被证明价值千金的直觉：复杂领域的客户，在看到能用的东西之前，不知道自己要什么；想知道「要什么」，最快的路是让造东西的人，站到用东西的人旁边。

成型：从应急之举到组织战略

早期的 Palantir 沿着「一个客户、一套定制」的路径往前走，很快撞上了那个所有企业软件公司都会撞上的问题：第二个客户要的东西，和第一个「细微但关键地不同」。

标准解法是提炼共性、做通用产品、对差异说不。但 Palantir 的客户不接受「不」——他们是中央情报局、是战场上的美军。真正把公司带出这个死结的，是早期员工希亚姆·桑卡尔（后任公司总裁兼首席技术官）。他的解法在当时被视为异端：不做两个产品，也不做一个僵化的产品，而是做一个高度可定制的平台，派工程师驻扎到客户现场，完成最后一公里。

桑卡尔最关键的动作，是账本上的叛逆。在软件行业的财务语言里，「为单个客户做定制」叫服务成本，是利润率的敌人，要被最小化。而桑卡尔把它重新定义为「产品发现」——现场定制不是花钱，是在为产品的下一次进化采集情报。一个名词的更换，改变了一家公司的资本配置逻辑：别人眼里的成本中心，成了 Palantir 的研发中心。

组织形态随之成型。前线部署工程师（内部代号「三角洲」）嵌入客户，「一个客户，调动多种能力」；部署战略师（代号「回声」）负责理解客户使命与组织；平台工程师在后方，「一种能力，服务多个客户」。到 2016 年前，Palantir 的前线部署工程师人数超过了平台工程师——一家软件公司，大半兵力在「前线」。

试炼：战场、飓风与油田

FDE 模式的价值，在一个个极端现场被反复证明。

伊拉克与阿富汗战场，路边炸弹是巡逻队的头号杀手。驻场的 Palantir 工程师发现，士兵们对精美的情报图表毫无兴趣——他们要的只是「在地图上标出这条路可疑」。工程师当场拼出一个简陋的地图工具：士兵点一下，风险路段全员可见。这个工具直接降低了伤亡，后来沉淀为平台的标准功能。总部会议室里拼不出这种工具。它只能出现在工程师与士兵一起看向同一条公路的那个瞬间。

而这套打法的原型现场，比地图工具更早。2007 年前后，桑卡尔带着一支小队钻进美军联合简易爆炸装置对抗组织（JIEDDO）的保密信息室，联合办公两周。保密室里连免提电话都禁用，他用松紧带把电话绑在头上，腾出双手敲代码——一只耳朵听分析师提意见，另一只耳朵听硅谷总部的同事说话。每天十九个小时，演示、接数据、收反馈、当场改。

两周结束，分析师们说：这东西有用。桑卡尔自己却累垮了，打电话给卡普：「这不可持续，我们完了。」卡普的答案日后成了公司文化：把这种「不可持续」，做成制度。（出处见附录 C）

2012 年飓风桑迪救灾现场，Palantir 的前线部署工程师把分散的救援数据拼成实时地图，协调救援资源的投放。油田钻井平台、飞机总装车间、银行交易大厅——工程师们把「驻场」这件事，做到了软件行业前所未有的深度。前 Palantir 工程师 Barry 回忆，Foundry 平台的核心组件诞生于苏黎世、休斯顿、圣保罗、图卢兹、巴库——一张世界地图，每个点都是一次现场创造。

代价同样真实。Barry 的记录，是理解这个模式的必读反面教材：顶尖工程师、全球差旅、大量重复造轮子、免费试点，「很多项目利润率字面是负无穷」；2014 年公司的产品战略字面是「强烈的观点，松散地持有」，混乱到让持股员工夜不能寐。Palantir 能吞下这些代价，靠的是两样东西：天价合同额撑起的资本厚度，以及「把试点当风投打」的心态——多数下注归零没关系，但赢的那几把要全赚回来。

商业化也先死过一次。第一个面向企业的产品 Metropolis 市场反响惨淡，第二次尝试 Foundry 才开了张。转折点是空客：图卢兹工厂里，A380 的一个燃油泵故障反复发作，空客自己的工程师查了整整两年没有头绪。Palantir 的人进场，把传感器数据接进平台，两周破案——飞机爬升时燃油晃离了泵体。一个微不足道的修复，保住了据报道价值数百亿美元的订单。

空客数字化转型负责人马克·方丹后来公开感慨：「同样的问题，我们以前要花二十四个月。」（出处见附录 C）空客从此成了 Palantir 在欧洲最热情的布道者，把自家数据平台 Skywise 整个建在上面，接入上万架飞机、五万多用户。

这套平台打法，后来在性命攸关的现场又证明了一次。佛罗里达州的坦帕综合医院（Tampa General）曾和一家供应商共建脓毒症预警工具，工具被美国食品药品监督管理局（FDA）归类为受监管的医疗器械，被迫停用。换作一般的客户，故事到这里就结束了；而坦帕综合的选择，是在 Palantir 的平台上自己动手重建，由首席数字与创新官斯科特·阿诺德（Scott Arnold）牵头。到 2025 年年中，这套自建的脓毒症中心已经挽救 569 条生命，脓毒症患者的住院时长缩短了三成。（出处见附录 C）

爆发：训练营点燃商业帝国

2023 年，生成式人工智能爆发。所有企业都想「用上人工智能」，所有企业却都卡在同一个地方：不知道怎么用。Palantir 的答案，是把它二十年的驻场方法论，压缩成一台工业化机器——AIP 训练营。

规则极简：客户带着真实数据和真实问题来，Palantir 的前线部署工程师陪着，一到五天做出一个能部署的人工智能应用原型，免费或象征性收费。第 0 天锁定一个极其聚焦的核心战场；第 1 天打通数据、构建本体模型（ontology，把客户业务的对象与关系建成语义层）；第 2 到 3 天，工程师与客户技术人员背靠背写代码，把大语言模型接入业务流；第 4 到 5 天，高管亲手操作系统，当场拍板。

这台机器的转化效率，让整个行业看傻了眼：从 2022 年的不足百场试点起步，场次连年倍增，到 2025 年高峰日均近 6 场；企业软件九到十二个月的销售周期被压到数周；连锁药房 Walgreens 通过训练营八个月部署四千家门店；市场研究公司 J.D. Power 这样的客户学会之后，开始给自己的客户办训练营——模式开始自我繁殖。训练营之后的合同节奏快到同行都未必信：一家大型医疗公司，参加训练营五周后签下五年期、年合同额（ACV）2600 万美元的协议；一家全球银行，试点一个月后先签 200 万美元，四个月后扩展为三年期、年合同额 1900 万美元。（出处见附录 C）

财务结果随之爆炸：2025 年第四季度美国商业收入同比增长 137%；「40 法则」（Rule of 40，收入增速与调整后经营利润率之和）达到 127%，2026 年第一季度冲到 145%；单季签约 42.6 亿美元，净收入留存 139%；市值一度突破 4000 亿美元。华尔街那些嘲笑它「人海战术」的分析师，开始连夜重写研究报告。

最大的一单信任票，来自美国海军。2025 年 12 月，海军部长与卡普共同宣布「造船操作系统」合同：4.48 亿美元，先用 Foundry 与 AIP 改造潜艇工业基地——两家大型造船厂、三个海军船坞、一百家供应商。试点数据前面讲过了：排产从 160 小时到 10 分钟，物料审核从几周到一小时。海军部长特意声明「这不是概念、不是试点、不是研究」——它跳过概念验证直接进生产，因为前期的嵌入试点已经证明了价值。同一年，美国陆军还把 75 份分散的服务合同打包成一份十年 100 亿美元的框架协议交给 Palantir。政府的信任，是二十年驻场攒出来的复利。（出处见附录 C）

一位曾在图卢兹驻场整整一年的工程师纳比尔·库雷希（Nabeel Qureshi）留下过更生动的注脚：他每周四天泡在总装车间，和制造工人一起，给 A350 的产线写软件——他管那东西叫「造飞机版的 Asana」（编者注：Asana，硅谷流行的项目协作软件）。一个硅谷工程师，在异国工厂里住满一年。（出处见附录 C）

方法论还原

回看前七章，Palantir 案例几乎处处有对应：演示循环与「客户不知道自己要什么」对应第 2 章，灯塔客户的战略价值对应第 3 章，免费试点的组合投资逻辑对应第 6 章，现场工具自下而上长成平台对应第 7 章。最底层的信条贯穿全书：复杂性无法被远程消灭，所以必须有人到场。

8.2 OpenAI：当造出 ChatGPT 的人开始下场搬砖

转折：从「模型即产品」到「部署即战略」

OpenAI 的前七年是一部纯粹的技术史诗：GPT 系列一路狂飙，ChatGPT 创下人类历史上最快的用户增长纪录。在这样的公司叙事里，「企业交付」听起来像隔壁行业的事。

转折发生在 2024 年。企业市场的数据，说明了一件不容忽视的事：《生成式人工智能的鸿沟》报告里那个后来广为人知的「95% 失败率」正在上演；与此同时，企业大模型市场的份额版图剧烈变动——Menlo Ventures 的行业追踪显示，Anthropic 在企业市场攀升至约 32%，OpenAI 从早年约 50% 的绝对领先回落至约 25%。模型能力的差距在缩小，而企业客户用真金白银投票的标准，悄然变成了另一个问题：**「谁能帮我把这东西真正用起来？」**

OpenAI 的回应分两步。第一步，2024 年组建前线部署工程团队——从 2 名工程师起步，科林·贾维斯带队，快速扩张到遍布旧金山、纽约、伦敦、都柏林、慕尼黑、巴黎、苏黎世、东京、新加坡、悉尼的全球网络，一年内从 2 人涨到 52 人。招聘启事写得毫不掩饰：「嵌入到那些模型表现至关重要、交付迫在眉睫、模糊性就是默认状态的客户中去」，出差最高 50%。欧洲区负责人富尼耶说得更直白：「市场需求超出了我们的预期。」（出处见附录 C）

第二步，2026 年 5 月 11 日，成立「部署公司」——一家由 OpenAI 控股、联合 19 家顶级资本（TPG 领投，Advent、贝恩资本、博枫联合领投，高盛、软银、华平等参与，贝恩咨询、凯捷、麦肯锡同为创始伙伴）的合资公司，初始投资超过 40 亿美元，媒体披露的投前估值约 100 亿美元；同步收购应用人工智能咨询公司 Tomoro，约 150 名部署工程师成建制并入。首席运营官布拉德·莱特凯普亲自挂帅。

这家新公司的结构设计，值得写进商学院教材，因为它把「客户从哪来」和「资本怎么回」焊在了一起。出资方是私募股权巨头，而这些巨头手里攥着横跨医疗、制造、金融、零售、物流的几百家被投企业——它们就是部署公司现成的客户池；作为交换，据报

道 OpenAI 向这些私募出资方承诺了五年 17.5% 的年化优先回报（preferred return）（媒体报道，OpenAI 官方未予确认），自己则保留超级投票权、控股战略方向。

被收购的 Tomoro 也不是无名之辈：这家英国应用人工智能咨询公司的工程师，此前在乐购、维珍航空、游戏公司 Supercell 干的就是企业落地。一天之内，OpenAI 集齐了「工程师、渠道、资本」三张牌。（出处见附录 C）

一家估值数千亿美元、掌握着人类最强模型的公司，为「派人去客户现场干活」单独成立一家百亿美元公司——这是 FDE 模式历史上最重的一枚砝码。

现场：约翰迪尔的农田

OpenAI FDE 方法论最好的展示窗，是与约翰迪尔的合作。这家近两百年历史的农业巨头，要解决的是一个土地里长出来的问题：杂草防治。传统做法是整田喷洒除草剂——成本高、药害重、环境代价大。精准农业的理想是「见草才打」，但识别的难度在于：每块田的草相不同、作物长势不同、时节不同。

OpenAI 前线部署团队的做法，是本书第二、四章方法论的完整演示：飞到爱荷华州，跟着农艺专家下田，理解精准农艺的工作流与约束（包括那条不可商量的硬期限——农时）；与专家一起评审数百个真实田间作业案例，把「什么是好的施药建议」编码成定制评估体系；在评估体系的看护下快速迭代模型与系统。最终结果：化学品使用量减少最高 70%，农户的互动频率提升了 6 倍。收费方式也与价值对齐：不按设备一口价，按实际启用技术的英亩数订阅计费——2024 年一年，这套系统省下了 800 万加仑除草剂混合液，2025 年覆盖面积已超过 500 万英亩。

其一，硬期限是农时，而非项目排期——FDE 的节奏必须服从客户的业务节律。其二，评估体系先行于模型优化——先定义「好」，再追求「好」。其三，70% 这个数字不是事后包装的宣传，是开工前就与专家共同定下的靶子。

现场：西班牙对外银行的十二万人

另一个标杆是西班牙对外银行（BBVA）。合作起点并不惊艳——部署企业版 ChatGPT。但 FDE 模式的复利在于：从一个工具进场，最后改造的是整个组织。从 3300 个账号起步，员工自发创建 2 万多个定制助手（GPTs，员工可自建的专属人工智能应用），83% 的人每周活跃使用，平均每周节省约 3 小时；一年半后，部署扩展为覆盖 25

个国家、12 万名员工的体系，银行的目标也随之升级为「建设人工智能原生的全球银行」。从「给一个工具」到「共建一种组织形态」，这是第 6 章「存量深耕」的极限形态。（出处见附录 C）

方法论还原

OpenAI 案例的启示，在于「模型公司的自觉」：当模型能力趋同，部署能力成为差异化的主战场；当产品形态未定（智能体该长什么样没人知道），现场成为产品发现的唯一现场。它的三阶段工作法（共创、验证、交付）、它对评估体系的工程化执着、以及部署公司的资本结构创新（用私募股权资本的产业网络做客户渠道），都是这种自觉的产物。

8.3 Anthropic、垂直三杰与平台厂商：FDE 的四种变奏

如果说 Palantir 发明了模式、OpenAI 把它推上头条，那么 Anthropic、一批垂直人工智能公司与数据平台厂商，则展示了 FDE 在不同土壤里的四种变奏。

变奏一：Anthropic——安全基因与「可审计」的差异化

Anthropic 的 FDE 挂在应用人工智能团队之下，招聘启事里有两个值得玩味的限定词：一是「安全与可靠」——所有交付必须满足 Anthropic 的安全标准；二是「在一线捍卫公司的使命」——FDE 同时是价值观的大使。2025 年，这个团队以一年五倍的速度扩张，负责人凯特·德容的解释是：「一家财富五百强银行的需求，和一家人工智能原生创业公司，完全是两个物种。」（出处见附录 C）

差异化的直接产物，是「可审计性」成为产品特性。与金融科技巨头 FIS 的合作是标杆：Anthropic 的前线部署工程师嵌入 FIS 共同设计金融犯罪侦测智能体，反洗钱调查从数小时压缩到数分钟，每个决策全程可回放、可追溯，首批落地蒙特利尔银行与 Amalgamated 银行。在强监管行业，每个人工智能决策都得向监管者解释清楚，解释不了就进不了场——Anthropic 把这条准入线做成了卖点。

「安全偏执」变成企业市场杀手锏的证据，写在客户的评价里。网络安全公司派拓网络把 Claude 部署给 2500 名开发者，工程总监的评价很直接：「他们比别家更在乎安全——每个会都在谈安全影响。作为最大的网络安全公司，这对我们很重要。」结果：功能开发速度提升两三成，初级开发者完成集成任务快了 70%。制药巨头诺和诺德用它写临床研

究文档：过去十几周的报告，现在十分钟出稿（厂商官方案例口径）。网约车平台 Lyft 把 Claude 接入客服，平均解决时长下降 87%；客服软件公司 Intercom 的客服智能体 Fin 换用 Claude 后，解决率从 23% 升到 51%，深度定制后最高做到 86%——这两个部署从试点到全员推广的完整过程，都写在第 4 章。与此同时，它与咨询巨头埃森哲合作培训了三万名 Claude 专业顾问，企业客户数两年内从不足千家涨到 30 万家以上。（出处见附录 C）

更深一层的差异，在「知识转移」的官方承诺：嵌入的明确目标，是让 FIS「未来能独立构建和扩展自己的智能体」。把「教会客户」写进合同，短期看是削弱自己的不可替代性，但长期看是对「供应商锁定」（vendor lock-in，被单一供应商绑死难以迁移）质疑最有力的回应——高德纳的一位分析师预测，2028 年前 70% 的企业将因供应商成本与技能空心化而弃用此类方案（媒体转述口径，详见 5.1 节），Anthropic 的做法，是提前拆掉这颗定时炸弹。

2026 年 5 月，Anthropic 被曝与黑石集团等组建企业人工智能服务合资公司，据报道规模约 15 亿美元，聚焦把 Claude 嵌入中型企业运营——与 OpenAI 的部署公司同日对垒。两家模型巨头在同一天完成「部署即战略」的组织化，这件事本身，比任何分析都更能说明 FDE 的坐标。

变奏二：Sierra 与 Decagon——创业公司的「交付即产品」

Sierra（Salesforce 前联合首席执行官布雷特·泰勒与谷歌前副总裁克莱·巴沃尔 2023 年创立，2024 年 10 月融资时估值已达 45 亿美元，此后据报道继续攀升）把 FDE 模式做成了商业模式本身：客户不为软件付费，为「已解决的会话」付费；实施由 Sierra 的前线部署团队托管，客户只需要贡献品牌语气与运营规则。效果：搜诺思（Sonos）、卡斯珀（Casper）、慧俪轻体（WeightWatchers）、ADT 等品牌客户云集，公开报道的最快上线案例为四周，年合同门槛约 15 万美元起（第三方估算，详见 6.3 节）。按结果收费让 Sierra 的定价天然锚定在客户价值上——第 6 章「按结果收费」最激进的现实样本。

Sierra 内部对这个岗位的命名也值得一记：不叫 FDE，叫「智能体工程师」（Agent Engineer）。其负责人娜塔莉·默勒（前 Palantir 五年，跑过执法、国防和基建客户）解释：名字应该描述「技术工作的形状」，而不只是「对客户的痴迷」——语音智能体的构建需要一种特殊的「品味」（什么听起来像人、什么听起来对），这是比通用 FDE 更窄更深的技艺。票务平台 Vivid Seats 的合作是她的团队最亮的战报：四周上线，解决率提升

40%，客户满意度提升 35%——而对话数据反哺产品的机制，让「一个月一万人问同一个功能」能立刻排上优先级。（出处见附录 C）

Decagon（创始人之一阿什温·斯里尼瓦斯出身 Palantir，总融资 2.31 亿美元）走了另一条路：把交付经验直接产品化为智能体操作流程——让客户的运营人员用自然语言定义多步工作流，人工智能确定性地执行。客户名单：Notion、多邻国、Eventbrite、Rippling。据报道，Chime 的解决率做到 70%，多邻国的问题分流率达 80%，ClassPass 的客服成本下降 95%。Decagon 的 FDE 更多地扮演「把客户团队教会用这套系统的人」——交付的目标，是让客户的非技术人员拥有自助能力，这是第 7 章「产品化四问」中「能否被不会写代码的人使用」的最佳实践。（出处见附录 C）

变奏三：Harvey——打进普通软件进不去的地方

法律行业是企业软件的「百慕大」：顶级律所保密纪律森严、数据不出域（data residency，数据不离开客户指定的环境）、合伙人各自为政——自助式 SaaS 在这里一胜难求。Harvey（2022 年创立，2026 年 3 月完成新一轮融资后估值已达 110 亿美元）用 FDE 模式硬生生打了进去。

这家公司的出身有点特别：两位创始人，一位是律所出身的诉讼律师，一位是 DeepMind（谷歌旗下人工智能实验室）出身的科学家；种子轮 500 万美元，来自 OpenAI 创业基金。（出处见附录 C）

复盘它的打法，几乎每一步都能在本书方法论里找到对应。第一个大客户锁定全球顶级律所年利达（2023 年 2 月，3500 名律师、43 个办公室），随后普华永道（4000 多名法律专业人士）、佳利、麦克法兰接踵而来。在年利达内部，对接人大卫·韦克林（市场创新负责人）成为部署的所内轴心；节奏上从单一业务组切入，靠合伙人支持者推动，六个月实战后才横向扩张。客户研究是最难的一环：合伙人买单但不干活，律师助理干活但不买单，知识管理律师看到两边的盲区，所以 FDE 的访谈体系必须同时穿透三层人群——按 Harvey 部署方的复盘，部署里最难的不是模型，是客户研究这个回路。周期也拉得很长：单所部署六到九个月，FDE 从文档管理系统的数据管道，一路负责到逐合伙人的采纳演示。

效果是垂直人工智能里少见的猛涨势头：用户超过 10 万名律师、1300 家组织，覆盖美国百大律所的多数；年度经常性收入（ARR）从 2025 年 8 月的约 1 亿美元，涨到

2026 年 1 月的约 1.9 亿美元；估值一年内四连跳——30 亿、50 亿、80 亿、110 亿美元。（出处见附录 C）

Harvey 证明了 FDE 模式的边界在哪里：凡是「保密制度、数据不出域、关键用户自主权」三座大山同时存在的市场，产品化交付无解，只有人进去，才能把系统带进去。而这样的市场——法律、医疗、金融、政府——恰好是地球上利润率最高的企业服务市场。

变奏四：Databricks 与 Snowflake——平台厂商的嵌入式交付

变奏不只发生在模型公司。数据平台厂商面对的客户，大多不是人工智能原生公司，而是数据堆积如山、却不知怎么用的传统行业——这里的交付，同样只能靠工程师嵌进去做。

第一美国（First American）是美国产权保险行业的老牌公司，维护着 80 多亿份公开登记文件的影像，其中一大批是连元数据都没有的历史扫描件。它要把约 400 万份可追溯至 1930 年代的产权保单逐页分类、逐页提取数据字段，商用大模型的精度却不达标。Databricks 的工程师直接嵌进了客户团队——客户方主管数据与人工智能的副总裁普鲁布·纳辛纳（Prabhu Narsina）回忆：「他们几乎每天和我们一起开会，一起写代码，一起 debug。」最终的路线是在平台上微调开源模型，约 50 张 GPU 跑训练，微调周期从其他平台的两个多月压缩到两周。项目总账：成本降低 70%，工期缩短 75%，数据提取量提升 10 倍。（出处见附录 C）

托管服务商 Ensono 与 Snowflake 的合作，则把「救火」改造成了「防火」：它的预测引擎上线以来分析了超过 7500 万个事件与 900 万条告警，在一张工单演变成服务事故之前提前预警——重大事故减少超过两成，故障修复时长最多压缩七成。（出处见附录 C）

这两个案例补上的是变奏谱系里容易被忽略的一块：FDE 不是模型公司的专利。手里有平台、又肯把工程师派进客户现场的公司，都能进这个场子——区别只在于，现场经验最后沉淀成了平台能力，还是消散在又一单人力交付里。

8.4 中国案例：在「项目制」的土地上种 FDE

中国是全球最特殊的企业软件市场：这里有最勤奋的交付工程师、最苛刻的定制化需求和最深的「项目制诅咒」——大企业要定制，厂商做一单赔一单，成本无法摊薄，沦为

甲方外包，专业空心化。FDE 模式在这片土地上的命运，是一个必须单独回答的问题：它是诅咒的解药，还是诅咒的新马甲？

这个局有多拧巴，三个从业者的叹息说得最清楚。第一声来自数据库公司 DolphinDB，一篇行业长文发问《谁在摧毁中国的企业软件产业》，答案是六座大山：白嫖、开源、外包、招标、数科（大型集团自设的数字科技子公司）、畸形的市场结构——最顶尖的中国工程人才，长期消耗在「高度定制、强关系、价格战」的非标交付里。第二声来自播客《硬地骇客》的一位企业软件老兵：「定制化是 SaaS 的天敌，这个诅咒只能等市场成熟那天自然消解。」第三声来自钉钉前总裁：很多公司做到一两亿收入就卡住——获客难、交付难、运维难、定制成本高，规模经济根本无从谈起。

36 氪替这三声叹息做了个总结：「软件产品化是把项目制的研发费用摊薄，SaaS 是把销售费用摊薄——只有项目制公司最难受，项目的成本离场即作废，没法摊出去。」（出处见附录 C）这三声叹息，就是 FDE 模式要回答的中国问题。

大厂路径：火山引擎、华为与腾讯

字节跳动旗下的火山引擎，是中国互联网大厂中最明确打出 FDE 旗号的玩家：2026 年专门成立 FDE 团队，深入行业与标杆客户深度共创。火山引擎总裁谭待在公开采访中的表述，与本书的定义几乎逐字呼应：「FDE 不是销售，也不是售前，他们必须具备很强的技术落地能力，特别是人工智能代码的落地能力。」他还透露，团队成员刻意配置多元行业背景（比如让生物工程出身的人去对接生物医药行业），目前已覆盖汽车、医疗、教育、金融、半导体等重点行业。2026 年 7 月，火山引擎又与咨询巨头安永签署战略合作，双方计划搭建千人级的 FDE 团队，「探索人工智能全栈交付新模式」。（出处见附录 C）

大厂的分量，在最重的行业里看得最清楚。2025 年 2 月，上海瑞金医院发布临床级病理大模型 RuiPath，与华为联合打造：16 张 GPU 卡、两个月，完成百万级病理切片的训练，数据准备周期缩短 80%。病理是中国医疗资源最紧缺的领域之一——医生缺口巨大，基层医院初诊符合率（基层诊断与上级医院最终诊断的一致率）低——这套模型覆盖了中国每年约九成癌症发病人群罹患的癌种。院长宁光在发布会上的表态，可以当作这类联合部署的前提来读：「我们只用经得起验证的技术。」（出处见附录 C）

腾讯的抓手，是把智能体做成平台生意。其智能体开发平台 2026 年 3 月上线，据称半年累计处理 90 亿次任务请求。平台上跑出的生产环境样本，正好补上了国内案例最重的一块拼图（以下数字均为腾讯云披露口径，出处见附录 C）：伊利把导购服务智能体接进导

购场景，订单提升 26%、点击率提升 15.7%；华润信托的「智能债券助手」直接住进交易群，面对年 40 万亿元的债券交易、每天约 4 万条信息，三个智能体的拆单准确率做到 99%；毕马威用它搭建行业研究助手，把一份深度研报的耗时从 14 小时（其中约 11 小时是重复查询检索）压到 2 小时以内；云南白药给答案加上权限过滤——同一个问题，不同权限的员工得到不同范围的回答——并办起大赛，让业务同事自建智能体上传平台，把个人能力沉淀为组织资产。支撑这批案例的需求挖掘纪律，与第 2 章的方法论互证：东北证券与腾讯云共建展业素材质检智能体（准确率 90% 以上）时，关键做法是「多问业务部门平时怎么工作、难点在哪、用哪些流程和系统，让业务提供真实案例集，而不是直接问他们想要什么」。

大厂做 FDE 的先天优势是平台底座厚（火山方舟提供了模型、工具链与云基础设施的一体底座）；先天挑战，则是大厂的销售体系惯性与「交付是成本中心」的财务惯性——FDE 团队能否获得 Palantir 式「现场即研发」的组织地位，决定了它是真 FDE，还是挂着新名字的售前支持部。

先行者路径：把「FDE 不等于驻场外包」写进官网

更值得关注的是一批本土创业公司。它们没有大厂的底座，却在用最清晰的话语，区分 FDE 与中国市场的旧物种。一家聚焦不动产与设施管理行业的服务商，在官网上写下了目前国内对 FDE 最精炼的定义性区隔：

> 一，FDE 按阶段交付、按结果验收，驻场外包按工时结算；二，FDE 带着产品底座来做工程，驻场外包从零现写现改；三，FDE 做完会走，能力沉淀在系统和你的团队里，驻场外包越驻越久、人走系统停。

三句话，分别对应本书的三个核心命题：按结果收费（第 6 章）、平台底座（第 7 章）、知识转移（第 5 章）。它还建立了一套「劝退机制」——场景未验证的不接、数据一张表能导出的不接、只想了解人工智能的不接，「我们宁可你晚一点开始，也不希望你在错的时机开始」。这是第 2 章「拒绝昂贵的概念验证坟墓」在中国市场的自觉实践。

另一家服务商则把 FDE 模式搬进了对合规最敏感的金融业，并沉淀出明确的三阶段方法论：场景诊断一到两周（识别价值最高的三到五个核心场景、量化投资回报）→ 嵌入式交付八到十六周（从模型选型到生产集成的全栈交付）→ 生产部署与持续演进（被业务部

门日常使用、随业务持续迭代），同时强调「数据不出域」——这是对中国金融机构监管环境的在地适配。（出处见附录 C）

2026 年 6 月，一位国内投资机构的从业者在网上发了一篇长文，标题很扎眼：《硅谷今年最火的岗位 FDE，我们闷头干了三年》。文章记录了他们把这套打法用在中国企业服务里的真实体感：前线部署工程师「签单后扎进一个客户，能力不设限，把人工智能真装进他的业务」；衡量标准「不是写了多少代码、交了几份文档，是这套系统到底有没有人在用、客户的生意有没有变好」；以及那句与本书完全一致的总结——「上线不算完，他得一直待到这套系统变成客户的日常、客户自己的人能接手」。这篇文章在国内从业者圈层被大量转发，因为它第一次用中文讲清了一个对照表：产品工程师对路线图负责，售前对签单负责，实施对验收单负责，客户成功对续费负责——而 FDE，对「客户的转型和人效（人均产出）有没有真涨」负责。（出处见附录 C）

对照组：机顶盒与纸质合同

FDE 该投谁、不该投谁，百道数据首席执行官吴光宇在一场创业者圆桌上给了两个反差极大的客户。

一个是机顶盒客户，需求极其明确——「怎么在机顶盒里用 AI」。FDE 和他们泡在一起，最后做出一个陪你看剧、聊八卦的智能体，投入产出比很高。另一个是出海客户，老板开口要投 100 万美元做 AI，签合同时却每次都把合同打印出来、用笔圈改，连 Word 的修订功能都不用。吴光宇的判断是：「这种数字化基础都没有的客户，FDE 投进去肯定没产出。我们直接告诉他，先买 100 个 Gemini 账号让全员用起来，有了具体思路再谈。」

他的结论，与第 2 章的拒绝机制如出一辙：FDE 的前提是客户有明确的需求点。模糊、没边界的需求不要轻易投；可以拆成一期、二期、三期，先做清晰的部分，交付结果后，后续需求自然能转化成长期服务合同。（出处见附录 C）

另一面：真空层与脏活

但中国的 FDE 实践不都是上面这种样子。一线传来的消息要糙得多。

《增长黑客 AI 周报》记录过一线执行者申悦的观察：在国内不少项目里，FDE 更像乙方外包人力，干的是填坑、擦屁股的脏活累活。她亲历的一个千万级国企项目，大老板签完合同就没了下文，项目里没有真正负责的人，FDE 被架在真空层——向上没有决策者，

向下没有执行者，最后只能靠研究「汇报逻辑」来争取验收。另一个广东的民营家装公司，厂二代老板开门见山：帮她把 6000 人裁到 3000 人。可 AI 能替代的是任务，不是岗位，项目就在反复挨骂中艰难推进。

申悦的总结是：中国 FDE 的难点，在需求方隔层、在企业数字化基础薄弱、在老板缺乏耐心——而前端销售为了卖产品许下的过度承诺，最后都要 FDE 来还。她还有一句话，我引作了第 4 章的题注：「无论技术如何进化，组织内部的人性和权责博弈，始终是比技术更复杂的难题。」（出处见附录 C）

另一种声音：FDE 是不是一门好生意？

质疑也不只来自执行层。同在一场圆桌上，Kuse.ai 创始人吴显昆对 FDE 模式本身提出了疑问：「FDE 很难成为一个好的商业模式。」他的理由很直接：做浅了，模型厂商自己会做；做深了，本质上还是外包，交付成本高且难以规模化。「给国央企做服务，你没法拒绝项目范围外的需求，你的杠杆在哪里？是否会在深夜怀疑自己成了廉价外包——这是所有做 FDE 的人都要面对的问题。」

他的解法相当激进：AI Roll-up（滚动并购）。既然客户担心数据泄露、团队嫌外包违背初衷，那就把公司买下来，所有人变成股东、利益一致，整个公司底朝天改掉；改完以后，用同一套方法论继续收购下一家。他的逻辑是：如果人工智能转型真有那么大价值，那就应该自己拥有这门生意（own the business），而不是卖工时。（出处见附录 C）

这个质疑未必有答案，但它和第 5 章引过的高德纳预测——2028 年前 70% 的企业可能放弃 FDE 主导的方案——形成了奇妙的呼应：对 FDE 模式来说，最难过的一关是它自己的成本结构。这个问题本书不打算给标准答案，但它是每一个想入场的人，都该先想清楚的问题。

中国 FDE 的三个特殊命题

综合这些实践与行业观察，FDE 在中国要成立，必须回答三个美国同行不必回答的问题。

第一，计价。美国市场可以接受订阅制、按量计费甚至成果分成；而中国客户的肌肉记忆是「买断加项目制验收」。可行的过渡形态是「混合制」。按阶段验收付款，顺应中国

客户的采购习惯；价值指标写进验收条款，把成果基因注入交易；再签一份年度运维与演进合同，逐步培育续约意识。一步到位照搬按解决量收费，在多数行业会死在采购科。

第二，平台底座。FDE 的经济学依赖「平台底座加现场定制」，而中国大量软件公司的问题恰恰是「有现场、没平台」——每单都是从零写代码的人力生意。对这些公司，比学 FDE 形式更重要的，是补 FDE 的地基：先把最高频的定制沉淀为可复用组件，再谈前线部署。如果没有第 7 章的复制杠杆，FDE 在中国只会沦为「名字好听的驻场外包」——这正是本土先行者们把「带产品底座」写进定义的原因。

第三，人才。中国工程师文化——能吃苦、响应快、全栈通吃、习惯在客户现场解决一切问题——恰恰最贴近 FDE 的要求，这是「机」；但「困」在于，过去二十年，这类人才被困在人天计价（按工程师人头×天数收费）的外包体系里，市场价格信号长期低估他们。FDE 概念在中国的普及，可能带来一个深远副作用：重新为中国数十万交付工程师定价。当他们知道自己在大洋彼岸的同行拿着 38.5 万美元的中位数总薪酬，这个行业的价值坐标会松动。

定价松动的证据，在 2026 年已经挂牌可见。需求端，招聘市场的价格拉开了档次：头部大厂与人工智能公司的 FDE 岗位，月薪挂牌区间大致 2 万到 8 万元，高职级年包 40 万元起，与传统实施工程师的价差拉到数倍（具体行情与分层见 1.8 节）。供给端，政府在动手补人：上海 2025 年底首开 FDE 专题培训班之后，2026 年 2 月又开出面向产业工程师的转型专班，聚焦智能制造、金融、医疗三个行业，首期报名超过 200 人，全年计划三期，并已列入市级专业技术人才知识更新工程。（出处见附录 C）

还有一个更底层的锚点，藏在传统软件的财报里。2025 年年报，用友人均营收约 48 万元，金蝶约 62 万元；同一年，Palantir 的人均营收约 100 万美元——十倍级的人效差。这既是「项目制诅咒」在财务上的样子，也是 FDE 模式在中国最大的想象空间：谁能把交付工程师的人均产出，从「人天」定价拉到「结果」定价，谁就同时改写这数十万人的市场价格，和这门生意的财务结构。（出处见附录 C）

老牌企业软件公司也在印证人才端的这个判断。SAP 大中华区总裁原欣在《晚点聊 LateTalk》播客里谈到，前线部署工程师和传统 ERP（企业资源计划系统）顾问的角色正在融合，眼下最缺的是「既有产品工程能力、又深入理解业务的复合型人才」。她还提醒：AI 落地的最大挑战是组织惯性而非技术——基础数据架构的历史欠账，不会因为 AI 来了就自动消失。（出处见附录 C）

判断

FDE 不会治愈中国企业软件的所有痼疾，但它提供了一次罕见的「正名」机会：把「贴近客户」从低端外包的标志，重新定义为高端能力的标志。中国有的是愿意下现场的工程师，缺的却是另外三样东西：让现场工作产生复利的平台、方法论与计价结构。谁先补齐这三块，谁就有机会在这个全球最大、也最复杂的企业服务市场里跑出来。

8.5 一个人工智能创业团队的 180 天：FDE 实战全复盘

最后这个案例，来自我跟踪调研的一个中国人工智能创业团队（应团队要求匿名，下称「N 公司」）。它没有 Palantir 的平台、没有 OpenAI 的光环，但它 180 天的历程，完整地演示了本书方法论在一个资源有限的团队身上如何运转。为保护商业信息，数字做了模糊化处理，但比例关系与决策逻辑保持真实。也正因如此，这个案例的读法和前面几组不同：前面的事实，可以逐条到附录 C 核对出处；这一组请当作一次完整的「流程演示」来读——它的价值是把全书的打法按时间顺序串一遍，事实判断以前面的案例为准。

第 1 到 30 天：选战场，而不是抢合同

N 公司做企业知识库人工智能，手上有三个潜在客户线索：一家头部券商（预算大、需求宏大）、一家区域连锁零售集团（预算中、痛点具体）、一家三甲医院（影响力大、数据敏感）。销售本能是扑向券商。但团队按第 2、3 章的方法，做了三重检验和进场尽调：券商的问题是「宇宙级需求」（首次会议就要覆盖全公司），且没有明确的业务负责人；医院的痛点真实，但数据不出域的限制让交付周期不可控；零售集团的首席财务官亲自盯项目，痛点具体（「3000 家门店的运营督导报告，区域经理根本看不过来」），数据基础虽然乱，但可触及。

团队选择了零售集团——灯塔价值不是最高，但三重检验全过。这就是 3.8 节的纪律：排期按战略价值与学习价值排，不按合同金额排。

被放掉的券商，值得一个注脚。「拒绝」不是这道题的唯一解。对灯塔价值极高、但需求模糊的大客户，还有第三种打法：不做交付，先卖一份付费的诊断——两到三周、收费固定，产出一份场景优先级报告和一份数据准备度清单，相当于把 2.3 节那个「体面的台阶」，做成一个正式的产品。如果对方真有转型决心，诊断过程会把宏大需求熬成一个清晰的第一场景，再签 MVD 水到渠成；对方若只是想在年报里写「人工智能」的，诊断费就是

试金石——连诊断都不肯付费的客户，本来也不该做。N 公司当时没走这条路。以本书的框架复盘，这大概是 180 天里唯一「正确但可惜」的决定：拒绝守住了产能，但也把一个潜在的顶级灯塔，让给了敢卖诊断的同行。

第 31 到 75 天：进场、影子工作法与一个「被砍掉的大方案」

两名工程师、一名行业顾问进场。前两周没有写一行产品代码，做的是影子工作法：跟着区域经理巡店，看他们如何在微信群里追门店数据、如何在电子表格里做周报。关键发现来自一个「变通」：区域经理们根本不看公司数据系统里的报表，他们真正信任的数据源，是门店店长每天手填的一张表——因为「系统的数据晚三天，还老错」。

这个发现推翻了最初的大方案（智能数据分析平台），团队把最小可行部署收窄为一个小切口：**只做一个「门店异常日报」**——每天早上八点，把昨天 3000 家门店的异常信号（销售异动、库存异常、投诉激增）汇总成三分钟内可读完的日报，推送到区域经理的微信。真实数据、单一痛点、六周硬期限。

第 76 到 120 天：激活的暗战

系统上线只是开始。激活期遇到两个教科书式的障碍。其一是信任校准：第一周，人工智能把两家门店的促销高峰误判为「异常激增」，经理们在群里吐槽。团队没有辩解，48 小时内把「促销日历」接入判断逻辑，并且——关键动作——在群里公开致谢报错的那位经理。其二是影响者经营：一位资深区域总监（无冕之王型）起初冷眼旁观，团队请他参与评估标准的修订，他的三条经验被写进系统规则。两周后，他成了系统最卖力的布道者。

第 90 天，日报的自然打开率稳定在 85% 以上。第 120 天，零售集团的首席信息官在季度经营会上主动展示了这套系统——内部支持者，被武装成了销售员。

第 121 到 180 天：复制的第一块砖

续约谈判没有悬念，但团队做了两个更重要的动作。其一，把这次交付中反复用到的资产沉淀下来：零售行业的数据接入组件、「异常信号」评估框架、门店场景的尽调清单——第一份场景打法手册成型。其二，用零售集团的案例（客户同意联合发布）敲开了第

二家客户——一家快消品牌的渠道管理部门，场景同构。第二个项目的交付周期，比第一个缩短了 40%：复制的飞轮，转出了第一圈。

180 天，一个小团队，没有融资新闻，没有颠覆式技术。但它验证了全书最朴素的结论：这套方法论不要求平台和光环，只要求每一步老老实实走完。

五组案例讲完，剩下两个问题：这样一支力量，应该恪守怎样的边界（后记）；以及入局者该盯哪些指标（附录 A）。

后记 FDE 的职业道德

每一本讲方法论的书，结尾都必须谈一个问题：当你掌握了这套方法，边界在哪里。而这本书要谈的话题，比大多数方法论更沉重。因为 FDE 手里握着的，不是一般的技术——是客户组织最深处的秘密，和越来越大的、代替人做决定的权力。

FDE 的工作性质，决定了它会看到什么。为了做好交付，你要看客户最真实的经营数据——包括难看的部分；你要摸清组织的权力地图——包括谁无能、谁失势；你要接触核心的业务流程——包括那些游走在灰色地带的变通。客户把这一切向你敞开，基于一个朴素的约定：你是来帮忙的。这份信任是 FDE 模式的地基，而毁掉它，只需要一次越界。

我认为 FDE 的职业道德至少有六条底线，写出来与所有同行共勉。

第一，数据的主权属于客户。 在客户现场看到的数据，一个字节都不应该出现在不该出现的地方——不进入人工智能的训练数据（除非合同明确授权），不进入案例素材（除非客户书面同意），不进入你下一份工作的谈资。最小权限原则（least privilege）不只是技术规范，是职业操守：能不看的就不看，能脱敏的就脱敏。

第二，诚实报告结果，包括坏消息。 按结果收费的模式里，最大的道德风险是粉饰结果——把「系统上线了」包装成「价值实现了」，把相关性包装成因果。第 6 章那套价值度量系统，既可以是最诚实的工具，也可以是最精巧的谎言机器，区别只在人心。FDE 的立身之本是「敢被检验」，那么被检验出不好的结果时，也要以同样的心态把它呈上去。

第三，不制造依赖，不贩卖恐惧。 这个行业里有两类隐蔽的做法：一种是故意把系统做成黑箱，让客户永远离不开你；另一种是夸大「不用人工智能就会死」的恐慌来促成交易。高德纳（Gartner）预测 2028 年 70% 的企业将因成本与技能空心化，而被迫放弃前线部署主导的方案——这个预测是对全行业的警钟。

而健康的 FDE 模式，交付结束的标志是客户自己跑得动：知识转移给客户，能力沉淀给客户团队。

第四，把「被替代的人」当回事。 FDE 交付的系统，在很多场景里确实会替代一部分人的工作。这本书讲了大量「变革管理」的技术，但技术之外是伦理：不要在被替代者面前庆祝效率，不要在方案里把人写成「成本项」而不给出路，不要对自己「正在改变谁的命运」装糊涂。别拿「技术是中立的」当借口——部署者每天做的选择，本身就是立场。

第五，对「客户要求但不该做的事」说不。 你会遇到游走在合规边缘的要求：「帮我们把员工的行为监控做得再细一点」「这些数据合规上有点模糊，但先接了再说」。「法国侍者」的隐喻在这里有最深的含义——真正的专业，不是满足客户的一切要求，而是敢于引导客户走向对他真正有利、也对世界无害的方向。说不的底气，来自你账上有别的客户；所以道德，从来也和商业模式有关。

第六，记住你代表的是「技术」本身。 对很多客户而言，你是他们接触人工智能的第一副面孔。你的每一次夸大，都在透支整个行业在他们心中的信用；你的每一次兑现，都在为整个行业存款。

你设计的护栏也在替普通人做决定——Anthropic 的工程团队公开复盘过一个数字：系统事事弹窗请求批准时，用户会照批约 93% 的请求，看得越多越不细看，「人在回路」（human-in-the-loop，人来把关）就这样在疲劳中形同虚设；所以设计护栏的人，不能假设把关的人永远清醒。当人工智能越来越深地进入社会的运转，部署者就是技术与人类日常之间的最后一道翻译——翻译失真的代价，由所有人支付。

这六条底线，反过来也是你挑雇主的试纸。2026 年的就业市场上，挂着 FDE 头衔的岗位一年暴涨，愿意干这行的工程师却只有约一成——供需的裂缝里挤满了跟风者。同一个头衔之下，前沿实验室的岗位是「软件工程加客户现场」，而跟风公司的岗位可能只是换了个名头的驻场外包：一位工程师入职后发现，公司只想让他做项目协调、一行代码也不打算让他写，四个星期后就辞了职。（出处见附录 C）

入行前多问一句——这个岗位写不写生产代码、现场经验反不反哺产品——比薪酬谈判重要得多。

Palantir 这家公司本身充满争议：它为情报与军事机构服务的历史，让许多人对它的一切——包括 FDE 模式——抱持戒备。我在本书中大量引用它的方法论，不等于认同它的全部客户选择。恰恰相反，正因为它服务的领域如此敏感，它的工程师文化里那些关于权限、审计、知所必需（need-to-know）的纪律，才格外值得借鉴。

在 Palantir 做了近八年 FDE 的纳比尔·库雷希，把他经手的项目按道德属性分成三类：道德中立的、明确为善的（疫情响应、打击儿童剥削）、以及灰色地带的（军事、移民、警务）。面对灰色地带，他的答案不是追求道德纯洁，也不是转身退出，而是「待在房间里」——在决策桌上保住自己的一票。我引用这个分类，因为它比「抵制」或「服从」都更接近部署者的真实处境：大多数道德难题恰恰发生在灰色地带，而离场把房间让给更不挑剔的人，未必是更道德的答案。

FDE 是一个年轻的职业，它的行规还没有人写。希望这本书，是它的第一块界碑。

致谢

感谢鲍勃·麦格鲁、Barry、泰德·梅布里、纳比尔·库雷希等 Palantir 前员工公开的回忆与书写——你们把一家「神秘公司」的方法论，变成了公共知识；感谢 YC Lightcone、Latent Space 等播客的主播们，你们的追问让大量一手经验得以留存；感谢 a16z、麻省理工学院 NANDA 实验室、The New Stack、CIO.com 的研究与报道；感谢中国的 FDE 先行者在官网写下的思考——中文世界的这场讨论，因你们而不必从零开始。

感谢每一位愿意读到这里的人。愿你修的路，都有人走；愿你走完后，路还在。

附录 A FDE 应当关注的常用指标

本附录给出 FDE 工作全链路的指标体系，按四个层面组织：交付层、客户层、商业层、组织层。指标贵精不贵多——每个客户项目盯住 3 到 5 个核心指标，胜过铺满一面墙的仪表盘。

〇、本书高频黑话速查

每词一句人话。读正文时卡住了，回这里翻。

- **FDE（前线部署工程师）：** 扎进客户现场、对结果负责的工程师。
- **本体（ontology）：** 把企业的数据、逻辑和动作建成模型能看懂的语义层。
- **训练营（Bootcamp）：** 客户带真实数据来，一到五天做出能用的原型，高管当场拍板。
- **回声 / 三角洲：** Palantir 的双人组合——前者读懂客户，后者动手建造。
- **概念验证坟墓：** 无限期、无指标、无裁判的试点，做着做着就死了。
- **MVD（最小可行部署）：** 用最小工程投入，在真实环境里验证一次价值真实发生。
- **影子工作法：** 坐在真实用户旁边，看他过完真实的一天。
- **灯塔客户：** 能给整个行业发信号的客户，背书价值大于合同金额。
- **需求蝗虫：** 预算足、需求旺，但吸干团队又不产生复利的客户。
- **NRR（净收入留存）：** 同一批老客户今年比去年多付还是少付，100% 及格。
- **RPO（剩余履约义务）：** 已签约、还没确认收入的合同额。
- **ARR / ACV：** 年度经常性收入 / 年度合同价值。
- **40 法则：** 收入增速加调整后经营利润率，软件业健康度指标，40 分及格。
- **按结果收费：** 客户不为软件付费，为「被解决的问题」付费。
- **人天计价：** 按工程师人头乘天数收钱——外包的老计价方式。
- **SLA（服务等级协议）：** 对可用性、延迟、错误率的承诺。
- **健康分：** 使用、价值、关系、商业四类信号合成的客户体检分。
- **QBR（季度业务回顾）：** 供应商与客户每季度对齐价值的例会。
- **RAG（检索增强）：** 让模型先查资料再回答。

- **评估体系：** 给模糊的业务质量建可量化标尺的体系，同时是计费的地基。
- **MCP（模型上下文协议）：** 让模型连接外部工具的开放标准，Anthropic 主导。
- **定制递减率：** 第 N 个客户的定制量应显著小于第 1 个，不降就是外包。

一、交付层指标（项目做得对不对）

TTV（Time to Value，价值实现时间）： 从进场到客户获得第一次可衡量价值的时间。FDE 模式的生命线指标。参考标准：最小可行部署的验证应以周计（2 到 6 周），完整部署以月计（1 到 4 个月）。价值实现时间持续变长，是方法论或平台底座出问题的第一信号。

概念验证（PoC）转化率： 验证项目进入付费部署的比例。Palantir 的训练营把这个数字从早期的 5% 到 10%，提升到了公司披露口径的约 75%（另有口径称部分场次更高）——它同时测量「客户筛选质量」与「交付质量」两件事。转化率过低，说明进场筛选失职；过高（接近 100%），则要检查是不是只接没有挑战的单。

评估达标率： 人工智能输出通过业务评估集的比例，以及它在生产环境中的时间序列走势。质量漂移的报警器。

解决率与分流率的口径纪律： 解决率（会话无需人工介入即闭环的比例）与分流率（deflection，用户被自助服务拦住、不再进入人工通道的比例）是客服类部署最核心的价值指标，也是各家口径最乱的指标。以 ServiceNow 的官方口径为例，一次分流成立需要两个条件同时满足：交互后 24 小时内未提交工单，且出现正向参与信号（如用户对结果给出正向反馈）。引用任何厂商的解决率或分流率数字前，先问清三件事：分母是什么、判定窗口多长、谁来判定——口径不清的 86% 不如口径清楚的 51%。

部署频次与回滚率： 迭代速度的硬指标。健康的部署期应该保持高频小步（周级甚至日级），回滚率低而稳定。

缺陷逃逸率： 上线后才被发现的缺陷占比。测量的是测试与评估体系的完备性，而非工程师的水平。

实战考核四指标： 据对约 1000 份 FDE 招聘启事的解码与一线交付复盘（出处见附录 C），考核一支 FDE 团队可以盯住四个数字：合同签订到首个生产工作流上线的天数；每个生产工作流消耗的 FDE 工时；试点转生产的转化率；上线后 30 天的留存使用率。值得玩味的是，这近千份招聘启事没有一份提到销售配额——FDE 的账，从来不是按签单算的。

两条经验阈值：客单价低于 10 万美元的产品，首个工作流若要 80 到 120 个 FDE 工时才能跑通，毛利会迅速收紧；一个季度内试点转生产不足一半，问题多半出在产品的封装、客户资格判定与实施设计，而不是市场教育。

二、客户层指标（客户活得怎么样）

激活率： 目标用户群中形成稳定使用习惯的比例。注意分母是「目标用户群」，而非「系统账号数」。警戒：激活率长期低迷的部署，名义上活着，实际上死了。

使用深度： 关键功能有几成人在用（用了多少场景）、使用频次分布（打卡式使用还是工作流嵌入）、以及自助探索行为的出现（用户开始自己发现新用法——这是最珍贵的信号）。

健康度评分： 使用、价值、关系、商业四类信号合成（详见 5.6 节）。关键纪律：每周全量巡视，跌破警戒线就自动触发干预流程。

支持者覆盖数： 客户组织内活跃盟友的数量与层级分布。单点是高危，三点成网。

净推荐值的慎用： 净推荐值（NPS，询问「你多大可能把我们推荐给别人」的指标）在企业场景的参考价值有限——样本小、政治化。更可靠的替代是「续约意向的提前问询」：到期前两个季度，直接问支持者「如果今天续约，你会续吗」。

三、商业层指标（生意值不值）

NRR（Net Revenue Retention，净收入留存）： 存量客户收入的年度变化（含流失、降级、扩容）。FDE 商业模式的最终裁判。及格线 100%，优秀线 120%。净收入留存高于 120% 的公司，增长引擎已经内生。

交付毛利率：（单个客户收入减去交付直接成本（人力、差旅、云资源））÷ 单个客户收入。FDE 模式的及格线随平台化程度浮动：纯人力交付期可能只有 20% 到 40%，平台复用度上来后应向 60% 以上上升。毛利率涨得快不快，比绝对值高低更能说明模式成不成——不上升的 FDE，是咨询公司。

定制递减率： 麦格鲁的试金石——第 N 个客户的定制工作量，应该显著小于第 1 个。连续三个客户定制量不降，立即检视产品回流机制。

销售周期： 从首次接触到签约的时间。Palantir 用训练营把它从 9 到 12 个月压到数周。它是获客效率与信任资产厚度的综合读数。

CAC 回收期： 获客成本（CAC，Customer Acquisition Cost，含免费验证的投入）通过合同毛利收回的时间。FDE 模式前期普遍难看，关键是要看到它随案例积累而缩短的趋势。

LTV/CAC： 客户终身价值（LTV，Lifetime Value）与获客成本之比。企业级生意的健康线一般在 3 以上；FDE 模式由于获客成本前置，早期可能低于 2，必须配合净收入留存一起读才有意义。

四、组织层指标（团队能不能走远）

现场到产品回流速率： 单位时间内，从现场沉淀为组件或平台能力的产出数（组件入库数、手册更新数、平台化提案数）。这个指标测量的，是 FDE 模式的「灵魂器官」是否还在跳动。

交付资产复用率： 新项目中复用已有组件、模板、清单的比例。复用率是三级复制杠杆（第 7 章）的综合读数，目标应该逐季上升。

FDE 人均产能： 每名前线部署工程师支撑的年收入。它是规模化的总账：纯人力模式的天花板明显，平台化后应该持续抬升。

团队续航指标： 出差强度（月出差天数）、值班负荷、离职率与倦怠预警信号。Reddit 上 FDE 从业者的最大吐槽点就是差旅与续航——团队烧干了，前面所有指标都是烟花。

团队规模的两个预警点： 一线复盘的经验值（出处见附录 C）：FDE 团队约 10 人处，协作摩擦开始变尖锐；50 到 100 人进入「存在性危机」——要么完成平台化转身，要么退化成按人头计费的交付部门。

五、行业基线参考（外部标尺）

内部指标之外，几组公开的调研基线，可以帮你校准「自己处在什么水位」。口径与年份各异，引用时以出处方为准（出处见附录 C）。

- **常态化运行率：** 中国电子技术标准化研究院 2025 年对 1200 家规上制造企业的调研：62% 已采购或上线人工智能系统，但上线满三个月仍常态化运行的仅 38%；停

用原因中近三成是「没有持续的现场陪跑」。这组数字是第 5 章「守住续约」最硬的市场注脚。

- **宏观看空数据组：** 美国全国经济研究所（NBER）2026 年的全球证据综述：89% 的企业表示人工智能尚未改变其生产力；IBM 的调研中，首席执行官们自认仅约四分之一的人工智能项目达到预期回报；一项 400 位企业数据负责人的调查显示，约六成企业在智能体人工智能上投入数百万至上千万美元，自觉做好生产就绪的仅约 15%——瓶颈与模型无关，在模型之下的数据地基。
- **采纳断层数据组：** 麦肯锡 2026 年调研中，80% 的企业认为人工智能提升了员工效率，但称它贡献了息税前利润的只有 37%，真正把人工智能跑进业务深处、拿到显著经营结果的高绩效企业只有 6%，约两成受访者称 Token 等运营成本已开始制约使用；高德纳对 1303 人的调研中，仅 22% 的组织在多业务部门实现了规模化；普华永道对 1217 名高管的调研显示，20% 的公司捕获了 74% 的人工智能经济价值，领先者重设计工作流的可能性约为同行的两倍；德勤对 501 名高管的调研中，仅 15% 实现规模化的多智能体部署，72% 缺统一可访问的数据，67% 认为集成过贵过复杂。
- **规模化的经济学水位：** 一线交付复盘的对照区间（出处见附录 C）：常规做法下，首个智能体上线要 8 到 12 周，上线后每个客户每月消耗 30 到 50 个 FDE 小时，一名 FDE 同时只能支撑 1 到 3 个客户；把数据地基与交接做对的团队，可以压到 1 到 3 周、2 到 5 小时、8 到 12 个客户。另一组同样扎眼：上线约三个月后仍保持健康的交付只有三到四成，最常见的失败模式是「FDE 走了，智能体也漂移了」。

使用建议

1. 每个客户项目锁定 3 到 5 个「核心指标组合」：通常是一个价值指标 + 一个使用指标 + 一个关系指标。
2. 基线数据在项目启动第一天采集——错过永不再来。
3. 指标与客户共建、双方共认，否则它在续约谈判桌上没有效力。
4. 每半年检修一次指标体系本身：删除无人看的，补上反复被问的。

附录 B FDE 人物与团队名单

本附录列出理解 FDE 模式绕不开的人物、团队与知识源头。本书写作过程中参考的完整来源清单，见随书稿公开的调研笔记。

一、关键人物

希亚姆·桑卡尔（Shyam Sankar） — Palantir 总裁兼首席技术官，早期员工。被普遍认为是 FDE 战略的发明者：把「现场定制」从成本重新定义为「产品发现」的人。他的公开发言，是理解 Palantir 哲学的第一手材料。

鲍勃·麦格鲁（Bob McGrew） — PayPal 早期工程师、Palantir 早期高管、OpenAI 前首席研究官（领导 ChatGPT、GPT-4、o1 研发）。FDE 模式最好的阐释者。「砾石路与高速公路」「在规模上做不可规模化的事」皆出自他口。必读：YC Lightcone 播客《给人工智能创业公司的 FDE 打法》（2025 年 9 月）。

斯蒂芬·科恩（Stephen Cohen） — Palantir 联合创始人。演示循环的发明者：「这东西太糟了」「那你们希望它哪里不一样」的主角。

亚历克斯·卡普（Alex Karp） — Palantir 首席执行官。「法国侍者」行为模型的提出者：反对唯唯诺诺的接单式工程文化。

Barry — 前 Palantir 前线部署工程师，《理解前线部署工程》一文作者。FDE 模式最重要的一篇「内部人警告」：成本、混乱与倦怠的真实账本。

泰德·梅布里（Ted Mabrey） — Palantir 资深员工与写作者，「桑卡尔炸弹」邮件（桑卡尔那封宣告战略转向的全员邮件）的亲历回忆者。他把「一切用结果说话」拆成两问：「它真的管用吗？它真的要紧吗？」

纳比尔·库雷希（Nabeel Qureshi） — 在 Palantir 做了近八年前线部署工程师，《Palantir 八年反思》（Reflections on Palantir，2024 年）作者。图卢兹空客现场驻场一年；「去他的可泛化性」、招聘的「蝙蝠信号」反向筛选、「数据问题九成五是接入、清洗、关联」，以及面对灰色地带项目「待在房间里」的道德三分类，皆出自他的公开书写。

维努·加内什（Vinoo Ganesh） — Palantir 轮岗培养计划（Project Frontline）的设计者，《FDE 权威指南》（The Definitive Guide to FDE，2026 年）作者。「你没法靠招聘建出真正的 FDE 组织，只能养出来」——250 多名工程师经轮岗进入真实部署，校友散在 OpenAI、xAI、Anduril。书中第 7 章的六个月自助培养清单、第 4 章「230 万个数据条目撑爆内存」的事故复盘，均出自他的一手书写。

科林·贾维斯（Colin Jarvis） — OpenAI 前线部署工程团队负责人，从零组建 OpenAI FDE 建制的人。

布拉德·莱特凯普（Brad Lightcap） — OpenAI 首席运营官，「部署公司」的主导者。

娜塔莉·默勒（Natalie Meurer） — Sierra 智能体工程负责人（前 Palantir 五年）。「智能体工程师」命名者。必读：Latent Space 播客访谈（2026 年 7 月）。

布雷特·泰勒（Bret Taylor）/ 克莱·巴沃尔（Clay Bavor） — Sierra 创始人（前者曾任 Salesforce 联合首席执行官，后者曾任谷歌副总裁）。把 FDE 模式做成商业模式本身（按结果收费）的人。

杰西·张（Jesse Zhang）/ 阿什温·斯里尼瓦斯（Ashwin Sreenivas） — Decagon 创始人（后者出身 Palantir）。智能体操作流程的设计者，「交付经验产品化」的样板。

大卫·韦克林（David Wakeling） — 年利达律师事务所前市场创新负责人，Harvey 首个灯塔部署的所内支持者。客户方支持者视角的最佳样本。

二、标志性团队

Palantir FDE 组织 — 模式的发明者与完成形态。「回声–三角洲–平台」三角，训练营机器。

OpenAI 前线部署工程团队 — 2024 年组建，全球分布式，约翰迪尔与西班牙对外银行的操盘者；2026 年升级为「部署公司」。

Anthropic 应用人工智能团队 — 以「应用人工智能工程师」名义组建的 FDE 建制，FIS 金融犯罪智能体的共建者；2026 年与黑石等组建企业人工智能服务合资公司。

Sierra 智能体工程团队 — 按结果收费 + 托管交付的极限形态（2023 年创立，2024 年 10 月估值 45 亿美元，此后据报道继续攀升）。

Harvey 部署团队 — 垂直行业（法律）FDE 的教科书（2022 年创立，2026 年 3 月估值 110 亿美元，用户覆盖 10 万+ 律师）。

Cognition 前线部署团队 — 「嵌入伙伴体内」的生态捆绑样板：2026 年起一支嵌入式前线部署工程师团队常驻全球信息技术服务巨头 Cognizant，负责项目筛选、工程师带教与成效度量，借伙伴几十年攒下的客户关系把产品铺进医疗、金融与保险大企业。

Scale AI / Databricks / Salesforce / 谷歌云 / Mistral / Cohere — 各自设有 FDE 或同构岗位，名称不同（解决方案架构师、客户工程师等），骨架相同。

火山引擎豆包大模型 FDE 团队 — 中国互联网大厂中最明确的 FDE 建制（2026 年成立，覆盖汽车、医疗、教育、金融、半导体；与安永共建千人级团队）。

中国本土 FDE 先行者：启盟科技、钱拓科技 — 以不动产设施管理、金融业为阵地的本土服务商，公开输出了「FDE 不等于驻场外包」的定义性思考。

三、知识源头（按重要性排序）

1. YC Lightcone 播客：《给人工智能创业公司的 FDE 打法——鲍勃·麦格鲁访谈》（2025 年 9 月）——模式第一讲述者的 51 分钟复盘。
2. a16z：《用利润率换护城河》（Trading Margin for Moat）——商业模式分析的标杆。
3. Barry：《理解前线部署工程》（Understanding Forward Deployed Engineering）——内部人的冷静警告。
4. 麻省理工学院 NANDA 实验室：《生成式人工智能的鸿沟：2025 年商业人工智能现状》（The GenAI Divide）——「95% 失败率」的出处，FDE 存在理由的数据地基。
5. The New Stack：《为什么 OpenAI 和 Anthropic 都在抢建 FDE 团队》（2026 年 5 月）。
6. CIO.com：《Anthropic 的金融智能体，暴露了前线部署工程师这个新瓶颈》（2026 年 5 月）——买方视角与高德纳预警。
7. Latent Space 播客：《前线部署工程师与软件工程的未来》（2026 年 7 月）——Sierra 视角。
8. Palantir 官方博客：《Palantir 前线部署工程师的一天》（2020 年）——官方自述，有中文译本流传。
9. OpenFDE 社区（open-fde.com）——从业者的开源共同体，岗位与公司图谱。

10. 36 氪《中国 To B 软件，走出亏损式增长的陷阱》、陈果 George 公众号 FDE 系列文章——中国语境的必读对照。

四、薪酬与求职参考

- GetPerspective：《2026 前线部署工程师薪酬报告》（1200 个样本）——头部实验室中级 FDE 年总薪酬中位数约 38.5 万美元、资深约 61 万美元、首席超过 120 万美元。
- fdenest.com / fde.academy / sundeepteki.org —— 面试准备与职业路径的三家专题站点。
- Levels.fyi —— Palantir 等职位的实时薪酬数据。
- 招聘启事本身是最好的教科书：OpenAI、Anthropic、Decagon、Harvey、Scale AI 的 FDE 招聘启事，值得逐字研读。

附录 C 全书案例索引与资料出处

> 本附录分两部分：第一部分是全书 100+ 个真实案例的索引（按章节）；第二部分是所有数据、观点、案例的资料出处（按章节）。标注「媒体报道」「公司披露」「第三方估算」的内容，其原始口径以出处方为准；标注「综合」的内容为多个公开来源的综合叙述。

第一部分 全书案例索引

第 1 章（21 个）

1. 全球五百强制造企业人工智能项目「安静死亡」（综合多个项目的典型场景）
2. 麻省理工学院《生成式人工智能的鸿沟》报告：300–400 亿美元投入、95% 无回报
3. 斯坦福大学《2026 人工智能指数报告》：组织采纳率 88%、生成式人工智能三年覆盖 53% 人口、智能体生产部署个位数百分比
4. 麦肯锡《人工智能现状》调研：仅 6% 企业人工智能贡献超 5% 息税前利润
5. 制造业首席运营官「车间没有任何改变」
6. 19–20 岁创业者年收入从零到 2000 万美元（《财富》对报告 Lead 作者的采访）
7. 中国 SaaS「亏损式增长陷阱」：定制化诅咒（36 氪、DolphinDB、《硬地骇客》播客）
8. Palantir 创立：2003 年、蒂尔/卡普/科恩、中情局旗下 In-Q-Tel 首投
9. 斯蒂芬·科恩的演示循环：「这东西太糟了」「那你们希望它哪里不一样」
10. 希亚姆·桑卡尔：第 13 号员工、FDE 模式与命名的发明者
11. 保密室两周驻场：每天 19 小时、电话绑在头上（COIC，2007 年前后）
12. 「桑卡尔炸弹」邮件：机场咖啡馆里的当面批评（泰德·梅布里回忆）
13. 伊拉克/阿富汗路边炸弹地图工具：从现场需求到平台标准功能
14. 2012 年飓风桑迪救援现场的 Palantir 工程师
15. Foundry 诞生于各客户现场：苏黎世、休斯顿、圣保罗、图卢兹、巴库
16. Metropolis 的商业失败与 Foundry 的翻盘

17. 空客 A380 燃油泵：两年未解 → 两周破案（保住据报道 400 亿美元订单）；Skywise 平台上万架飞机、五万用户
18. 纳比尔·库雷希图卢兹驻场一年：「造飞机版的 Asana」
19. 本·拉登行动传言（从未证实、从未证伪，Colossus 杂志注脚）
20. Palantir 2016 年前 FDE 人数超过平台工程师；一半以上的工程师撒在全球客户现场（口径为工程师占比，非全体员工）
21. Palantir 财报「自成一格的物种」：Q4 2025 合同额 42.6 亿美元、净收入留存 139%、现金 72 亿；Q1 2026 收入 +85%

第 1 章 1.3 节（FDE 风潮，12 个）

1. 《金融时报》2025 年 11 月报道：FDE 岗位九个月增长 800%（Indeed 数据）
2. YC 招聘板 100+ 创业公司招 FDE（三年前几乎为零）
3. a16z「科技行业最热门的岗位」「奶奶与 iPhone」
4. OpenAI 前线部署团队：2024 年初成立、一年从 2 人到 52 人
5. 阿尔诺·富尼耶：约 50 人扩编计划、约翰迪尔降药六到七成、「反哺研究和产品」
6. 凯特·德容（Anthropic）：应用人工智能团队一年扩五倍；「五百强银行与创业公司是两个物种」
7. 尼克·普雷特约翰（Palantir 英国）：「从内部做产品发现」
8. 艾丹·戈麦斯（Cohere）：「合同一开始就嵌入工程师」
9. Salesforce 承诺招聘 1000 名 FDE
10. 德勤 2025 年 12 月成立 FDE 业务线
11. OpenAI 部署公司（2026-05-11）：40 亿美元、19 家投资方、收购 Tomoro
12. Anthropic 与黑石等组建约 15 亿美元企业人工智能合资公司

第 2 章（8 个）

1. 日本企业自建人工智能团队「最强阵容的安静死亡」
2. 五万美元合同分析工具输给免费 ChatGPT（麻省理工调研细节）
3. 外部合作成功率约为内部自建的三倍（同报告）
4. 超半数人工智能预算投向前台、回报却在后台（同报告）
5. 中国 FDE 服务商的「劝退机制」：三种客户不接

6. Barry 回忆：免费试点烧掉数百万美元、利润率「负无穷」
7. 训练营转化合同：医疗公司五周签 2600 万美元年约；全球银行 200 万→1900 万；另有财富五百强医疗公司约 1000 万美元一例（正文从简，未列）
8. Walgreens：10 家门店试点提效 30% → 八个月 4000 家门店、每日 3840 亿次自动决策

第 3 章（8 个）

1. Harvey × 年利达：2022 年 11 月秘密试点、4 万次真实查询、250 个业务领域、50 种语言；年利达 3500 名律师、43 个办公室，单所部署六到九个月（getperspective，单一来源，谨慎引用）
2. 大卫·韦克林：「15 年法律科技未见过的游戏规则改变者」；合伙人「卢森堡开银行」演示
3. Harvey 创始团队：诉讼律师 + DeepMind 科学家、OpenAI 基金 500 万美元种子轮
4. 普华永道 2023 年 3 月引入 Harvey：100 多个国家 4000 多名法律专业人士
5. CIA 灯塔效应：从情报界信任到华尔街与制造业
6. 火山引擎 × 安永：千人级 FDE 团队、TRAE 工具平台（2026-07-19）
7. OpenAI 部署公司伙伴名单：麦肯锡、贝恩、凯捷同为创始伙伴
8. J.D. Power 学会训练营后为自己的客户办训练营

第 4 章（11 个）

1. 约翰迪尔「看见即喷」：36 摄像头、时速 12–15 英里、一分钟三个足球场
2. 约翰迪尔产量对照：12 万亿株作物、200 对 600 蒲式耳（贾斯汀·罗斯）
3. 约翰迪尔成果：化学品减 70%、农户互动提升 6 倍
4. OpenAI 欧洲负责人谈约翰迪尔：「反馈反哺研究」
5. 西班牙对外银行渗透路径：3300 → 11000 → 120000 账号
6. 西班牙对外银行数据：每周省 3 小时、83% 周活跃、2 万+ 定制助手（4000 高频）
7. 「影子人工智能」对策：「给他们一个安全的平台」（埃琳娜·阿尔法罗）
8. 「八件事」转型路线图与 250 名高管先行培训、人工智能极客网络
9. 美国海军 ShipOS：4.48 亿美元、电船公司 160 小时→10 分钟、朴茨茅斯船厂几周→1 小时

10. Wendy's：6450 家门店糖浆调度，15 人一天 → 5 分钟
11. 房利美：抵押欺诈检出率超 99%；花旗银行审批从几小时到几分钟

第 4–5 章（组织与续约，7 个）

1. Vivid Seats：四周上线、自助解决率 +40%、满意度 +35%
2. Vivid Seats 后续：团队从灭火转向根治、「一万人问同一功能」反哺产品
3. Anthropic × FIS：反洗钱调查从数小时到数分钟、蒙特利尔银行与 Amalgamated 银行首批
4. FIS 合作的知识转移承诺：「让 FIS 能独立建智能体」
5. 西班牙对外银行的「人工智能先锋网络」与「极客」制度
6. 埃森哲 × Anthropic：三万名 Claude 认证顾问
7. 高德纳预测：2028 年 70% 企业将放弃 FDE 主导的智能体方案；「FDE 成本谁在付」之问

第 6 章（11 个）

1. 训练营免费经济学：一年数千万美元的免费投入
2. Palantir 存量成绩单：净收入留存 139%、剩余合同额 +145%
3. Sierra 成绩单：Funnel 94%、Ramp 90%、Casper 74%（满意度 +20%）、慧俪轻体约 70%（4.6/5）、AOL 64%
4. Sierra 价位：年门槛约 15 万美元起、首年 20–35 万、大单百万级、单次解决 1–2 美元（第三方估算）
5. Sierra 按「已解决会话」收费：不解决不收费的结构
6. Harvey 扩张：10 万律师、1300 家组织、60 国、百大律所多数
7. Harvey 收入：年经常性收入五个月从 1 亿到 1.9 亿美元；估值一年四连跳至 110 亿
8. 行业调研：68% 律所生产环境用 Harvey、深度用户每周省 11 小时
9. Anthropic × 黑石合资：对标 OpenAI 部署公司、聚焦中型企业
10. 按结果收费四档进化链：用量 → 行为 → 结果 → 价值分成
11. 中国「混合制」计价：阶段验收 + 价值指标 + 年度演进合同

第 7 章（9 个）

1. 「砾石路与高速公路」：现场定制 → 平台泛化（麦格鲁）
2. Barry：试点即风投组合、研发而非销售成本、任务式指挥（Auftragstaktik）
3. 2014 年 Hobbitcon 大会：产品战略字面是「强烈的观点，松散地持有」
4. Sierra「代笔人」（Ghostwriter）：上传流程文档，自动生成生产级智能体
5. Decagon 智能体操作流程（AOP）：客服主管自然语言写流程
6. Decagon 客户数据：Chime 70%、多邻国 80%、ClassPass 客服成本降 95%
7. 毛利率时间简史：ServiceNow 63%→79%、Workday 54%→76%、Palantir 79%→82%（Insight Partners）
8. Salesforce 早年：烧 5200 万美元换 2200 万美元收入的实施投入期（a16z）
9. 桑卡尔的豪赌：「价值归于芯片和本体」、大半工程师压上 AIP

第 8 章（25 个）

1. Palantir 创业史全复盘（科恩演示循环、桑卡尔战略化）
2. 保密室两周与「这不可持续，我们完了」电话
3. 路边炸弹地图工具全复盘
4. 空客 A380 两周破案与 Skywise 平台
5. 美国海军 ShipOS 与陆军十年 100 亿美元框架合同
6. Palantir 帮：Decagon、Sierra 智能体工程负责人、行业写作者群
7. OpenAI 转折：企业份额从 50% 到 25%、Anthropic 升至 32%（Menlo Ventures 2025 年中）
8. OpenAI FDE 建制：2 人起步、科林·贾维斯带队、全球九城
9. 约翰迪尔项目全复盘
10. 西班牙对外银行项目全复盘
11. 部署公司的资本工程学：私募被投企业即客户池、17.5% 保底回报报道、Tomoro（乐购、维珍航空、Supercell 背景）
12. Anthropic 安全差异化：派拓网络 2500 名开发者（提速两三成、初级快 70%）
13. 诺和诺德：十几周的临床报告 → 十分钟
14. Anthropic 企业客户：两年内从不足千家到 30 万家以上
15. Sierra 创立与估值：2023 年创立、2024 年 10 月 45 亿美元、此后继续攀升
16. 娜塔莉·默勒：从 Palantir 五年到 Sierra「智能体工程师」命名
17. Harvey 打法复盘：灯塔、支持者、纵深、超长周期

18. Harvey 估值阶梯：30 亿 → 110 亿美元（2026-03，新加坡政府投资公司 + 红杉）
19. 火山引擎 FDE：谭待专访、行业覆盖、豆包日均 tokens 超 180 万亿
20. 安永 × 火山引擎千人 FDE 团队
21. 启盟科技：「FDE ≠ 驻场外包」三条区隔与劝退机制
22. 钱拓科技：金融 FDE 三阶段方法论、数据不出域
23. 《我们闷头干了三年》：中国从业者的 FDE 对照表
24. 中国企业软件三声叹息：DolphinDB、《硬地骇客》、钉钉前总裁、36 氪
25. N 公司 180 天实战复盘（综合多个调研素材的匿名案例，数字已模糊化）

增补案例（2026-07-30，来自作者 Obsidian 库的一手素材）

1. Cresta FDE 团队：今年 30 人扩到 100 人；「FDE 必须绑定 AI 平台」、招人硬杠「不会代码像文盲」、双重职责（Jove 视频对谈）
2. 特赞 atypica：用智能体对企业全员做自主访谈，替代结构化问卷（丁鑫栋）
3. 得物：万人公司 AI 转型三步「共识、场景、知识」与容错四象限（任喜亮）
4. 百道数据对照组：机顶盒「陪看剧 Agent」vs 纸质合同的出海客户「先买 100 个 Gemini 账号」（吴光宇）
5. 申悦的一线冷水：千万级国企项目的「真空层」与「汇报逻辑」验收
6. 广东民营家装公司：厂二代要求裁 6000 人到 3000 人，「AI 替代的是任务不是岗位」
7. Kuse.ai 的 AI Roll-up：「FDE 很难成为好的商业模式」与 own the business 解法（吴显昆）
8. SAP 原欣：FDE 与 ERP 顾问角色融合；「AI 落地最大挑战是组织惯性而非技术」

增补案例（2026-08-07，OpenFDE 信号库）

> 以下来自六章增补的用稿登记（sourcelog）。沿用「双源优先、单源标注」原则；标「待核」者出版前需复核原文，标「公司官方案例」「供应商案例口径」「公司自述口径」者原始口径以出处方为准。

第 1 章

1. Indeed 官方口径：FDE 岗位一年增长 729%（2025 年 4 月的 643 个 → 2026 年 4 月的 5,330 个；Business Insider 报道 + 智东西中文编译，双源，待核）
2. 高盛自研人工智能助手铺向全行：内测覆盖上万名员工、全程在防火墙内（《财富》2025-06-24，单源）
3. Klarna 人工智能客服上线首月接下三分之二客服会话（Klarna 官方新闻稿 2024-02-27，单源；OpenAI 客户故事同事件源见待补源清单）
4. 雪城大学全校开放 Claude：先完成培训再开通账号（校方官方新闻，单源，待核）
5. 科林·贾维斯 2025 年 1 月一手官宣 OpenAI 前线部署职能：使命「帮客户把系统送上生产」（当事人 X 帖，单源）
6. 谷歌云一次放出 59 个 FDE 岗位：首席执行官库里安亲自喊话招「想站在智能体时代中心的建造者」（CIO Dive 2026-05-14，单源）
7. Databricks 专业服务部门整体改组为 FDE 组织：十二个月服务 1,900 多个客户；摩根大通技术负责人买方证词「我们要的不是更多顾问，是能造出还不存在的东西的工程师」（Databricks 官方博客 2026-06-11，单源）
8. 上海全国首个政府主导 FDE 培训班：六家联办、先训需求方决策层、「百千万」工程、官方定调「人工智能领域的特种兵」（上海创智学院官网 2025-12，双源）
9. Foundation Capital 4.6 万亿美元拆账：一半是岗位薪酬、一半是信息技术服务与外包，软件收费对象从工具预算换为人力预算（Foundation Capital，双源）
10. 《华尔街日报》的反方质疑：更像咨询而非软件、嵌入式人才太贵、规模化是悬案（单源，待核，仅引质疑框架）

第 2 章

1. 摩根士丹利「先建考题」：三场景起步、理财顾问与工程师人工打分、每日回归测试题库，质量下滑在抵达一线前拦截（OpenAI 客户案例，多源交叉，待核；与 OpenAI 官方企业部署指南「从测试题库开始」互证）
2. Ramp「永远在选题」：第一条纪律与安卓/iOS 反面教材（几周工作全废）（Ramp 工程博客 2025-08-05，单源，待核）
3. 约翰迪尔的价值定义与订阅计价：农户要的不是「人工智能」而是少打药多打粮，按实际启用英亩订阅收费（OpenAI 官方访谈 2025-05，单源，待核）
4. Salesforce 拿自己练手（Customer Zero）：自用一年后把 650 多条数据流整合成单一事实源，「答不上来」比例从三成压到一成以下（Salesforce 官方，与 142 双源齐备）

第 3 章

1. Looker 2013 年先例：「演示就是概念验证」、只要真实数据集；2.5 万美元 × 2,000 个客户 = 1 亿美元年经常性收入的测算（First Round Review 2025，单源）
2. Rippling 把「供应商必须投入专属工程资源」写进选型硬标准（Decagon 客户案例页，单源，供应商案例口径）
3. Notion 自身人工智能优先仍选择买而非自建：RFP 五条标准末条「要能一起共建的团队」（Decagon 客户案例页，单源，供应商案例口径）
4. 53AI 三年按结果收费实践：提示词工程师单挑交付的教训、业务专家先行、A/B 实证人工智能组产出 3 倍（53AI 杨芳贤 2026-06-10，单源，公司自述口径）

第 4 章

1. Skai 定制小分队两天上线商品分类工具（传统做法一个月）：「我不需要变成提示词工程师」（Snowflake 官方客户案例，单源，公司官方案例）
2. Intercom 客服产品解决率爬升 23% → 51% → 86%；Anthropic 自用约两年升至 79%、月解决约 56 万次（claude.com 客户案例页，单源，公司官方案例，待核）
3. Lowe's 店员助手装进既有手持终端：铺 1,700 多家门店、回答 500 万多次提问；「SVP 不入局就不做」（OpenAI 客户案例经官方新闻稿与 2025 年报交叉，多源，待核）
4. Salesforce 帮助站自用：74 万多篇知识文章、低代码两个月上线、76% 咨询无需人工（Salesforce 官方客户故事，与 134 双源齐备）
5. ServiceNow 自家助手上线 120 天跑出约 1,000 万美元年化收益（Now on Now 官方客户故事，单源，公司官方，待核）
6. Anthropic 审批疲劳实验：用户批准约 93% 的权限请求、钓鱼演习 25 次成功 24 次，边界应下沉系统底层、默认拒绝（Anthropic 官方工程博客 2026-05-25，单源，公司官方）
7. TIME 年度人物交互体验上线前红队模拟数千种攻击手法（Scale AI 官网客户故事，单源，公司官方案例）
8. Lyft × Anthropic 三件套（共建产品 + 早期接入 + 工程师培训）：客服解决时长 -87%、日解数千请求（Lyft 官方博客 2025-02-06，单源，客户方官方）
9. Jamf 十六个部门推广：「推动最广泛采用的不是工程师」；绩效对话工具 45 分钟搭成（过去一个团队三个月）（Anthropic 客户案例经转述核验，单源，待核）

10. DoorDash 日处理数十万通人工智能来电、自动化测试容量 50 倍、与外部专家八周完成设计到部署（Anthropic × AWS 官方案例 2025，单源，公司官方）
11. Sierra 五个客户上线最快不到两周、最慢六周，有客户首日解决过半咨询（Sierra 官方博客 2025-09-02，单源，公司官方，匿名客户组）

第 5 章

1. 单个部署年交付成本约 7.5 万美元起、一个满负荷 FDE 只服务三五个客户（SaaStr 2025-08，单源）
2. 法律科技公司防漂移样板：漏检自动生成标注样本进每晚微调，关键条款召回率一个月 92% → 98%、服务不中断（Foundation Capital 2025-07-03，单源）
3. Wiley 的季度价值故事：213% 投资回报、季节性客服上岗培训提速一半、工单分类数据反哺分销商问责（Salesforce 官方客户故事，单源，公司官方案例）
4. Sierra 智能体在「摇摆时刻」主动挽留：一家旅游平台套餐保留率 +5%（Sierra 官方博客 2025-04-17，单源，单一客户案例未外推）

第 6 章

1. Hunter Douglas 超 100 万美元收入来自纯人工智能处理的对话：客服部门长出了收入归因（Decagon 客户案例页，单源，供应商案例口径）
2. ClassPass 实际上线自动分流为预期的 10 倍；人工智能翻译质量超过原本地化供应商、致其合同未续签（Decagon 官方客户案例，单源，供应商案例口径）

第 7 章

1. Cognition 把嵌入式 FDE 团队进驻 Cognizant：负责项目筛选、工程师带教与成效度量，借伙伴客户关系铺进医疗、金融、保险大企业（Cognition 官方博客 2026-01-28，单源）
2. Booking.com 把企业搜索平台推广到 14,000 名员工：公司首个全公司级人工智能平台，推广视频制作周期压到原来的四分之一（Glean 客户案例页，单源，公司官方案例）
3. Palantir 轮岗培养计划（Project Frontline）：250 多名工程师进真实部署、校友散在 OpenAI/xAI/Anduril，「你没法靠招聘建出真正的 FDE 组织，只能养出来」（vinoo.io 2026-02-05，单源，项目设计者一手）

4. Palantir 面试招牌环节：拆解刻意模糊的真实问题、几乎不写代码，候选人最常见失误是当纯软件工程师面试准备（Glassdoor 聚合经多源补齐，单源，待核）
5. Mercado Libre 内部平台向 17,000 名开发者开放自助：先在低风险任务积累信任（编目量两年提升一百倍），多年后才碰涉及金钱的客服争议调解、且只放开一小块（OpenAI 客户案例经二手交叉核实，单源，待核）
6. BVP 拆解塔塔咨询（TCS）：十年营收翻倍、人数几乎翻倍、人均营收停在约 4.9 万美元——不平台化的服务公司十年后只是更大的服务公司（BVP Atlas 2025，单源）

第 8 章

1. 坦帕综合医院脓毒症中心：供应商工具被 FDA 停用后于 Palantir 平台自建，截至 2025 年中挽救 569 条生命、住院时长缩短三成；OneBlood 遭攻击时几小时搭出血液库存应用并交州级复用（Healthcare IT News 2025-07 + MedCity News 2025-02-20，双源，待核）
2. First American × Databricks 产权保单页级提取：工程师嵌入共创、微调开源模型 50 张 GPU，成本 -70%、工期 -75%、数据提取量 10 倍（Databricks 客户案例，双源：与厂商视角互证）
3. Ensono × Snowflake 预测引擎：分析 7,500 万事件 / 900 万告警，重大事故降幅按保守口径写「超过两成」（-30% 与 -22% 两口径并存）、修复时长最多压缩七成（Snowflake 官方客户案例，单源，公司官方案例）
4. 瑞金医院 RuiPath 病理大模型 × 华为：16 卡 2 个月训练百万切片、数据准备周期 -80%、覆盖约九成癌症发病人群癌种（华为企业业务官网 2025-02-18，单源，公司官方）

增补案例（2026-11-02，intake 增补轮）

1. OpenAI FDE 团队的「自我清理」：废弃单用例独立终端与命令行工具，约九成定制代码转为基于 Codex 原语的通用插件与技能；一个用例初始部署两个月、扩展到第二个团队一天；企业语音产品 OpenAI Presence 由 Swarm/Agents API 演进，投资方旗下公司过半电话交由人工智能处理（AI 醒知 2026-10-25 转述 OpenAI FDE 负责人复盘文，单源）
2. 「一对」组织指纹：Palantir「三角洲+回声」、Anthropic「FDE+参与经理」、OpenAI「技术部署负责人+FDE」；构建者头衔全行业已统一、协调者头衔不一（戴

珂待出版新书章节，单源）

3. 规模化三条路径：自招（Palantir）、资本买产能（OpenAI 收购 Tomoro 带来约 150 名 FDE、Anthropic 与黑石/H&F/高盛组建 15 亿美元合资公司）、认证伙伴（AWS 向认证咨询伙伴开放方法论与生产标准）；试金石「部署第 N+1 个客户是否比第 N 个更便宜」（两起交易经官方公告核验：Anthropic 2026-05-04、OpenAI 2026-05-11；路径框架引自戴珂新书章节，单源）
4. 规模化经济学对照：首个智能体上线 8–12 周 vs 1–3 周、每客户每月 30–50 FDE 小时 vs 2–5 小时、一名 FDE 支撑 1–3 个 vs 8–12 个客户；三个月后健康交付仅三到四成、「FDE 走了，智能体也漂移了」；团队约 10 人变尖锐、50–100 人存在性危机（戴珂待出版新书章节，单源）
5. FDE 实战考核四指标与毛利阈值：合同到首个生产工作流上线天数、单工作流 FDE 工时、试点转生产转化率、上线 30 天留存使用率；10 万美元客单价对应 80–120 工时警戒、季度试点转生产不足一半为产品封装问题；约 1000 份招聘启事无一提销售配额（同上，单源）
6. 常态化运行率：中国电子技术标准化研究院 2025 年调研 1200 家规上制造企业，62% 已采购或上线人工智能系统、满三个月常态化运行仅 38%、停用原因近三成是「没有持续的现场陪跑」（AI 故事计划报道转引，单源）
7. 采纳断层数据组：麦肯锡 2026（80% 提效、37% 贡献息税前利润、6% 高绩效、约两成称 Token 成本制约使用）、高德纳 1303 人（22% 多部门规模化）、普华永道 1217 高管（20% 公司捕获 74% 价值、领先者重设计工作流可能性约两倍）、德勤 501 高管（15% 规模化多智能体、72% 缺统一可访问数据、67% 集成过贵过复杂）、NBER 2026（89% 未改变生产力）、IBM（约四分之一项目达预期回报）、400 位数据负责人（约六成投入数百万至上千万美元、约 15% 自觉生产就绪）（Fundaai 研报与 Nextplayso 转引各机构调研，以各机构原始报告口径为准）
8. 腾讯云智能体开发平台四案例（均为腾讯云披露口径）：伊利导购智能体订单 +26%、点击率 +15.7%；华润信托「智能债券助手」住进交易群，年 40 万亿元债券交易、日约 4 万条信息、三个智能体拆单准确率 99%；毕马威行研助手将深度研报从 14 小时（约 11 小时重复查询检索）压到 2 小时以内；云南白药按权限过滤答案、办大赛让业务同事自建智能体沉淀为组织资产；平台 2026 年 3 月上线，据称半年累计 90 亿次任务请求（运营社报道，单源，厂商口径）
9. 东北证券 × 腾讯云展业素材质检智能体：准确率 90% 以上；需求挖掘纪律「多问业务部门平时怎么工作、难点在哪、用哪些流程和系统，让业务提供真实案例集，而不

是直接问他们想要什么」（同上，单源）

> 合计 174 个条目，覆盖 100+ 个产品、品牌、服务与组织。

第二部分 资料出处（按章节）

自序 / 第 1 章

- 麻省理工学院 NANDA 实验室：《The GenAI Divide: State of AI in Business 2025》（报告原文，2025-07；方法学：52 家组织访谈 + 153 份高管问卷 + 300+ 公开项目）；《财富》杂志 2025-08-18 报道及对报告 Lead 作者 Aditya Challapally 的采访
- 斯坦福大学以人为本人工智能研究院（Stanford HAI）：《2026 人工智能指数报告》（组织采纳 88%、生成式人工智能人口覆盖 53%、智能体部署个位数）
- 麦肯锡：《The State of AI》（2025-10，1993 名受访者：88% 使用、39% 有息税前利润影响、6% 高绩效者）
- Y Combinator Lightcone 播客：《The FDE Playbook for AI Startups with Bob McGrew》（2025-09-08，文字实录：定义、科恩演示循环、桑卡尔发明 FDE 战略、砾石路与高速公路、doing things that don't scale at scale、YC 招聘板 100+）
- Colossus 杂志：《The Patriot: Shyam Sankar of Palantir》（2026-04-14：保密室两周、电话绑头、桑卡尔炸弹、FDE 命名、本·拉登传言注脚、「价值归于芯片和本体」）
- Barry（前 Palantir FDE）：《Understanding Forward Deployed Engineering》（barry.ooo：负无穷利润率、研发而非销售成本、Hobbitcon、任务式指挥、风投组合）
- Challenges 杂志（2019-03）：空客案例与马克·方丹引语、Metropolis 失败
- theforwarddeployed.io 案例库：纳比尔·库雷希图卢兹驻场记述
- 《金融时报》（2025-11-02，经 eWeek / Moneycontrol / Economic Times / storyboard18 转引）：FDE 岗位 +800%、OpenAI 约 50 人计划（富尼耶）、Anthropic 五倍扩编（德容）、普雷特约翰与戈麦斯引语、Echo/Delta 双人组

- Insight Partners：《Why we invested in Rocketlane》（2026-04：OpenAI FDE 2→52 人、Salesforce 1000 名 FDE、德勤 FDE 业务线、三家毛利率史）
- a16z：《Trading Margin for Moat》（services-led growth：奶奶与 iPhone、实施必要性、八条最佳实践、Salesforce 烧钱史）
- Palantir Q4 2025 财报新闻稿（Business Wire，2026-02-02）与 Q1 2026 财报（2026-05-04）：Rule of 40 127%/145%、美国商业收入 +137%/+133%、合同额、净收入留存、现金
- OpenAI 官网：《OpenAI launches the OpenAI Deployment Company》（2026-05-11）；《John Deere transforms agriculture with AI》（2025-05-06）；deploy.co 案例页；Axios/路透社/《金融时报》关于投前估值与 17.5% 回报的报道（OpenAI 未确认后者）
- 36 氪《中国 To B 软件，走出亏损式增长的陷阱》（2024-08）；DolphinDB 博客《谁在摧毁中国的企业软件产业》（2026-01）；小宇宙《硬地骇客》EP54、《对话不穷》
- GetPerspective：《2026 FDE 薪酬报告》（单一来源，出版前建议以 Levels.fyi 复核）；fdenest.com；sundeepteki.org 面试指南（问题拆解轮、凌晨两点格言、Palantir 招聘标准引语）
- Reddit r/cscareerquestions、r/salesengineers 从业者讨论（经 creme de la CRM 转引）

第 2 章

- 麻省理工学院 NANDA 报告（同上：五大路障、律师细节、外部 vs 自建、预算错配、「垃圾进垃圾出」）
- timewell.jp《The GenAI Divide》综述（2026-05：日本企业自建失败案例、三报告对照）
- 启盟科技官网 FDE 服务页（aipm.cn/workshop/fde：三条区隔、劝退机制）
- 钱拓科技官网 FAQ（aioai.cc：金融 FDE 三阶段）
- Palantir 财报电话会披露的训练营转化合同（经雪球、微信公众号「Palantir 财报透视」转引：2600 万/1900 万/1000 万年约、Walgreens 细节）
- 百度百科「AIP Bootcamp」词条（场次与转化率口径之一，已与多源交叉并按保守口径表述）；搜狐《深度拆解：Palantir 业绩狂飙背后的秘密武器——AIP

Bootcamp》

- Harvey × 年利达官宣（2023-02-15，经 Legal IT Professionals、ABA Journal、Law Society of Ireland Gazette 报道）
- sundeepteki.org：OpenAI 三阶段工作法（共创、验证、交付）

第 3 章

- Harvey × 年利达（同上）与 × 普华永道（2023-03，经 Iberian Lawyer、36 氪/雪球报道）
- getperspective.ai：《Harvey AI's Forward Deployed Engineers》（部署周期、三层人群、年利达规模、「最难的是客户研究回路」一语的出处，单一来源，谨慎引用）
- 中国日报网（2026-07-23）、DoNews（2026-07-22）、火山引擎官方公众号：安永 × 火山引擎战略合作与千人级 FDE 团队
- 界面新闻《对话火山引擎谭待》（2026-06-23）、财联社/蓝鲸财经（2026-06-24）：火山引擎 FDE 团队成立、行业覆盖、谭待引语
- a16z《Trading Margin for Moat》（亲自到场、激励对齐、留文档、自动化、招实干者）
- OpenFDE 社区（open-fde.com：公司图谱与岗位名称对照）

第 4 章

- OpenAI 官网约翰迪尔案例（同上：See & Spray 技术参数、贾斯汀·罗斯引语、70% 与 6 倍）
- OpenAI 官网与西班牙对外银行新闻稿（2025-11-06 / 2025-12-12）：3300→11000→120000 路径、每周 3 小时、83% 周活跃、2 万+ 定制助手、「八件事」路线图；hamidun.com 案例页的口径辨析（内部估算、无独立审计）
- 美国海军 ShipOS：USNI News / Defense One（A 级）/ The Register / WWD（2025-12-10 前后：4.48 亿美元、160 小时→10 分钟、几周→1 小时、海军部长引语）
- SegmentFault/墨滴转引的 Palantir Ontology 案例集（Wendy's、房利美、花旗、Lowe's；原始口径来自 Palantir 客户披露）

- Sierra 官网客户案例页：Vivid Seats（2026-03-17：四周上线、+40%/+35%、帕切科与史密斯引语）
- CIO.com（2026-05-06）：FIS × Anthropic 细节、高德纳 70% 预测（Alex Coqueiro）、马哈帕特拉「谁在付 FDE 成本」
- The New Stack（2026-05-28/29）：「模型是最干净的部分」引语、FDE 工作内容

第 5 章

- 同第 4 章 FIS、西班牙对外银行、埃森哲来源；VentureBeat（2025-12-25：埃森哲三万顾问、派拓网络、诺和诺德）
- Palantir 财报（净收入留存 139%）
- 高德纳预测（CIO.com 报道引述分析师 Alex Coqueiro 的个人观点，非高德纳正式预测报告；正文 5.1 已按此口径表述）

第 6 章

- Sierra 官网客户案例与第三方汇总（myaskai.com 2026-03：解决率表 64%-94%；twig.so / eesel.ai / usefini.com：价位与单次解决定价的第三方估算，口径不一，正文已标注「第三方估算/据报道」）
- helpshift.com：Decagon vs Sierra 对比（2026-06：部署周期、价位区间）
- metronome.com / pickaxe.co：Decagon 定价结构（9.5 万美元起、按会话/按解决）
- Harvey 估值与收入：aitoolsbakery、valueaddvc、teahose、aifundingtracker（2026 年 3 月 110 亿美元轮、年经常性收入 1.9 亿、估值阶梯；多为行业二级来源，口径一致）；RSGI 律所调研（经 valueaddvc 转引：68%、11 小时）
- Menlo Ventures：《2025 Mid-Year LLM Market Update》（GlobeNewswire，2025-07-31：Anthropic 32%、OpenAI 25%、Google 20%）与《2025 State of Generative AI in the Enterprise》（2025-12：Anthropic 40%、OpenAI 27%）
- marktechpost（2026-05-20）：部署公司细节、Anthropic 合资 15 亿美元、薪酬带宽

第 7 章

- Barry（barry.ooo）、YC 播客（麦格鲁）、Insight Partners（毛利率史）、Colossus（桑卡尔豪赌）（均同上）
- Sierra Ghostwriter 发布（2026-03-25，经 myaskai 等报道）
- Decagon 客户数据：eesel.ai（Chime 70%、多邻国 80%、ClassPass 95%，单一来源，已标注）
- Salesforce 烧钱史：a16z《Trading Margin for Moat》

第 8 章

- 综合以上全部来源；新增：财联社《豆包官宣推出专业版》（2026-06-24：火山引擎 FDE、豆包 2.1、日均 tokens 180 万亿、IDC 份额 49.5%）；rui.juzi.bot《硅谷今年最火的岗位 FDE，我们闷头干了三年》（2026-06-04）；中国日报网安永合作报道（同上）
- N 公司案例为作者调研的多个中国创业团队素材的综合匿名处理，数字已模糊化，不作为单一事实引用

增补素材出处（2026-07-30）

- 课代表立正 YouTube 对谈：《他的 FDE 团队从 30 人扩到 100 人，但为什么很多大厂工程师甚至不知道这个岗位？》（2026-07，Cresta FDE 负责人 Jove；经 LLM 转录，概念性引用，未逐字引用）
- 极客公园 Founder Park 圆桌：《AI-Native 组织落地和 FDE 到底怎么做，我们跟五位一线创业者聊了聊》（2026-07-22，微信公众号；丁鑫栋/特赞、刘之/科锐国际、任喜亮/得物、吴光宇/百道数据、吴显昆/Kuse.ai）
- 《增长黑客 AI 周报》EP#63：《FDE 岗位在中国、北京创投圈见闻实录》（申悦的一线观察与总结，作者本人 newsletter）
- 晚点聊 LateTalk 播客 EP174：《AI 冲击企业软件巨头？与 SAP 原欣聊大模型 to B 的颠覆与边界》（2026-07-28，摘要与核心要点转述）

增补素材出处（2026-08-07，OpenFDE 信号库）

第 1 章

- Stratechery（2023-06-15 桑卡尔与梅布里专访：Skywise 一百五十多家航空公司、梅布里「两问」；2024-09 本·汤普森：多年期深嵌入交付是向四十年前常态的回归）
- American Optimist 播客（Joe Lonsdale，2026-05-08：桑卡尔认领 FDE 命名、「把痛苦吃进去、把产品产出来」；待核：据官方 shownotes 与转述交叉）
- Palantir 官方博客《Dev versus Delta》（Dev/Delta 划分的一手口径；待核：据搜索摘要重建归属）
- nabeelqu.co《Reflections on Palantir》（2024-10：「去他的可泛化性」、蝙蝠信号反向筛选、数据问题九成五是接入清洗关联；待核：原文经 Marginal Revolution/Lenny's 转述交叉）与 Lenny's Podcast（2025：近三分之一 Palantir 产品经理离职创业）
- 《财富》（2025-06-24：高盛人工智能助手全行铺开）；Klarna 官方新闻稿（2024-02-27：首月三分之二会话）；Syracuse University Today（2025-09-22/2026-08-05：全校开放 Claude；待核）
- Business Insider（2026，Indeed Hiring Lab 数据：729%）＋智东西中文编译（2026-05-18）；FT 官方 X（2025-11：800% 原始链路）＋Fast Company（2026：1,165% 并置与莱维断言；待核）
- Semafor（2025-07-11：岗位定调时间线起点；待核，标题级）；CIO Dive（2026-05-14：谷歌云 59 个岗位、库里安喊话）；Sifted（2025-12-15：欧洲转向；待核，paywalled）
- 上海创智学院官网（2025-12-09/2025-12-21：FDE 培训班，双源）；Foundation Capital（2024/2024-25：4.6 万亿美元拆账，双源）
- The Wall Street Journal（2026：反方质疑框架；待核，paywalled，仅框架级使用）
- Sequoia Training Data 播客（2025：麦格鲁「模型之外」论）；X @colintjarvis（2025-01-15，经 fxtwitter 镜像：贾维斯官宣）
- Databricks Blog（2026-06-11：专业服务改组 FDE、1,900 多个客户、摩根大通买方证词、按里程碑与结果对齐定价）
- The Pragmatic Engineer《The Pulse》（2026-05-14：资深 FDE 总包中位 48.5 万美元口径；待核，paywalled，正文按「2026 年初」限定处理、两口径不并置）
- 人人都是产品经理（2026：五十份招聘启事提炼的真假 FDE 试金石）
- a16z《Trading Margin for Moat》（同上，补「软件自己就是工人」一句）

第 2 章

- The Pragmatic Engineer（2025–08–12，科林·贾维斯受访：选题描述与地面现实对不上、约翰迪尔旧流程前史、SA/FDE 按真实数据划线）
- OpenAI 客户案例（摩根士丹利，经 Inblix 详录与多源交叉；待核：原文 403）与 OpenAI 官方企业部署指南 PDF（2025：七条经验，「从测试题库开始」与之互证）
- Ramp Builders Blog（2025–08–05，Leo Mehr：永远在选题与安卓/iOS 反面教材；待核：JS 壳经转述补齐）
- OpenAI 官方访谈（Justin Rose，2025–05：按实际启用英亩订阅收费；待核）
- Salesforce News 官方（2025–09–18：650 多条数据流整合为单一事实源）
- MedCity News（2025–02–20，ViVE 现场采访：坦帕综合医院 CIO 引语、OneBlood 几小时搭出血液库存应用；待核）
- 53AI 杨芳贤（2026–06–10：「AI 出错不报错，只是悄悄变得不靠谱」，定性引用未引数字）

第 3 章

- SaaStr《Dear SaaStr》（2025，Jason Lemkin：客户成功经理管 8–12 户对 FDE 嵌 1–3 户、单客户部署 30–60+ 天每日在场）
- tomtunguz.com（2025–07–18：约 10 万美元/合同是模式经济性下限）
- First Round Review（2025：Looker「演示就是概念验证」先例与年经常性收入测算）
- Decagon 客户案例页（Rippling 选型硬标准、Notion 买而非自建；供应商案例口径）
- Glassdoor（第三方聚合转述：Palantir FDE 薪酬 4.0/5 对工作生活平衡 2.8/5；待核，聚合口径未经原页核对，正文已标注来源层级）
- 53AI 杨芳贤（2026–06–10：三年按结果收费实践、A/B 实证人工智能组产出 3 倍；公司自述口径）

第 4 章

- Snowflake 官方客户案例（Skai，2026–08–07 抓取）；claude.com 客户案例页（Intercom 23% → 51% → 86%、Anthropic 自用约 79%；待核，终审建议复核口径）

- OpenAI 客户案例（Lowe's，经官方新闻稿、2025 年报与零售媒体交叉；待核，数字以年报口径）
- Salesforce 官方客户故事（帮助站 76% 咨询无需人工，2026-08 访问口径；与 Salesforce News 双源齐备）
- ServiceNow Now on Now 官方客户故事（120 天约 1,000 万美元年化收益；待核；分流率判定口径见附录 A）
- Anthropic Engineering Blog（2026-05-25：审批疲劳约 93%、钓鱼演习 25 次成功 24 次）；Scale AI 官网客户故事（TIME 红队模拟数千种攻击手法，原文口径 7,000+）
- Lyft 官方博客（2025-02-06：客服解决时长 -87%）；Anthropic 客户案例（Jamf，经 theapplied.co 转述核验；待核）
- Anthropic × AWS 官方案例单页 PDF（2025：DoorDash）；Sierra 官方博客（2025-09-02：五个客户上线速度，匿名客户组）
- vinoo.io（2026-02：Palantir 内部系统被 230 万个数据条目撑爆内存，当事人一手）

第 5 章

- SaaStr（2025-08 中旬，引 Orosz 推文：单个部署年交付成本约 7.5 万美元起、满负荷 FDE 只服务三五个客户）
- Foundation Capital（Ashu Garg、Jaya Gupta，2025-07-03：法律科技公司防漂移样板，召回率 92% → 98%）
- Salesforce 官方客户故事（Wiley 213% 投资回报，2026-08 访问）；Sierra 官方博客（2025-04-17，Nishita Jain：「摇摆时刻」保留率 +5%）

第 6 章

- Decagon 客户案例页（Hunter Douglas 超 100 万美元收入归因、ClassPass 自动分流 10 倍与翻译替代；供应商案例口径）

第 7 章

- Cognition 官方博客（2026-01-28：嵌入式 FDE 团队进驻 Cognizant）
- Glean 客户案例页（2026-08-07 抓取：Booking.com 14,000 名员工、推广视频周期压到四分之一）

- vinoo.io（原载 Next Play，2026-02-05：Project Frontline 轮岗培养计划，项目设计者一手）
- Glassdoor 面试问题聚合页（UGC，要点经多家面试指南与候选人报告补齐；待核，引用已标注来源层级）
- OpenAI 客户案例（Mercado Libre，原文 403，经 Hamidun News 案例库详录与多家二手报道交叉；待核）
- Bessemer Venture Partners（BVP Atlas 2025：塔塔咨询拆解；同报告「FDE 起源新说」见第 1 章）

第 8 章

- Healthcare IT News（2025-07：坦帕综合医院 569 条生命、住院时长缩短三成；待核，原站 403 经多源交叉核验）
- OpenAI 官方访谈（2025-05：See & Spray 2024 年省 800 万加仑除草剂、2025 年覆盖 500 万+ 英亩；待核）
- Databricks 客户案例（First American，发布日期未标注）；Snowflake 官方客户案例（Ensono，页面 2026 年版本，重大事故降幅取保守口径）
- 华为企业业务官网新闻（2025-02-18：瑞金医院 RuiPath 病理大模型）

后记

- The Pragmatic Engineer（Gergely Orosz，2026-03-27：岗位暴涨与仅约一成工程师愿意干并存、入职 4 周闪辞一手案例；内含 WSJ/AlphaSense 数据转引）
- nabeelqu.co（2024-10：道德三分类与「待在房间里」；待核：原文被拦，据搜索摘要与转述交叉，建议浏览器补抓校对引语）
- Anthropic Engineering Blog（2026-05-25：审批疲劳，同上）

增补素材出处（2026-08-08，模拟读者评审轮）

- MIT《GenAI Divide》95% 的口径争议（第 1 章 1.1 新增交代段）：GoML.io《The GenAI Divide Report is a Trojan Horse for MIT NANDA》（六个月窗口过短、价值口径仅三项、样本偏向大企业、NANDA 自身利益关系）；JorgeP.com《A Cautionary Tale or a Misleading Snapshot?》（样本规模质疑）；Digital Health

Wire（「按此定义，互联网与云计算的早期也算失败」）；DHInsights（麦肯锡、兰德公司、标普全球的同方向独立研究佐证）

- 国内 FDE 薪酬行情（第 1 章 1.8、第 8 章 8.4 新增）：36 氪《年薪百万，招聘需求涨 42 倍，「AI 圈最火岗位」到底是干嘛的？》（2026）、财联社（2026）、新浪财经（2026-05-18）——均为招聘平台公开挂牌价的媒体汇总口径，单一来源簇，出版前建议以招聘平台实时数据复核
- 可比公司人均营收（第 7 章 7.8 新增）：Palantir 2025 全年营收约 44.8 亿美元（Q4 2025 财报）与员工 4,429 人（2025 年末口径），人均约 100 万美元；埃森哲 2025 财年年报（营收 697 亿美元、员工约 77.9 万人），人均约 8.9 万美元；Harvey 的约 58 倍估值倍数为本书据第 6 章已有数据（ARR 约 1.9 亿美元、估值 110 亿美元）的算术
- 用友、金蝶人均营收（第 8 章 8.4 新增）：用友网络 2025 年年报（营收 91.82 亿元、年末员工 19,055 人，人均约 48 万元）；金蝶国际 2025 年度业绩（营收 70.06 亿元、员工 11,294 人，人均约 62 万元）
- 上海 FDE 培训进展（第 8 章 8.4 新增）：上海创智学院官网（2026-02-25：转型工程师专班第一期开班，聚焦智能制造、金融、医疗）；上观新闻（2026：首期报名 200+、全年三期）；上海市经信委《专业技术人才知识更新工程 2026 年度高级研修项目》公示（2026-06）

增补素材出处（2026-11-02，intake 增补轮）

- 第 7 章新增（OpenAI 九成定制代码、Presence 演进、规模化三条路径、「一对」组织指纹、规模化经济学对照、戴珂四指标）：AI 醒知（2026-10-25，转述 OpenAI FDE 负责人复盘文，单源）；戴珂《从 0 到 1：构建可信可用的企业智能体》待出版新书章节节选（公众号连载，单源）；OpenAI 官方公告（2026-05-11：OpenAI Deployment Company，19 家投资方 40 亿美元初始投资、TPG 领投，收购 Tomoro 带来约 150 名 FDE 与部署专家）；Anthropic、Blackstone、Hellman & Friedman、Goldman Sachs 官方新闻稿（2026-05-04：15 亿美元独立企业服务公司，面向中型企业部署 Claude）——两起交易已按官方一手源核验
- 第 8 章新增（腾讯云智能体开发平台四案例、东北证券质检智能体）：运营社报道（腾讯大会演讲与腾讯云披露口径，单源，正文已标注厂商口径）

- 附录 A 新增（实战考核四指标、行业基线参考五组）：戴珂新书章节（四指标、毛利阈值、经济学水位、团队规模预警点，单源）；中国电子技术标准化研究院 2025 年调研（经 AI 故事计划报道转引，单源）；Fundaai 研报（转引麦肯锡、高德纳、普华永道、德勤 2026 年调研）；Nextplayso（转引 NBER 2026 全球证据综述、IBM 调研，注意原文为 H 公司赞助内容）；各机构原始报告口径以出处方为准

附录 A/B

- 指标定义综合自 Palantir 财报口径、Insight Partners、a16z、GetPerspective 薪酬报告、sundeepteki 面试指南、各家官方招聘启事（OpenAI / Anthropic / Decagon / Scale AI / Roboflow 等）

待补源清单（2026-08-07）

> 以下官方一手源因拦截或死链未取得。正文已有对应内容的，出版前建议补抓核对；未进正文的，补源后再评估是否采用。

- 诺和诺德 × Claude：~~官方一手源待补~~ 已补源核实（2026-08-07 核验轮）：公开口径为 10–15 周 → 约 10 分钟（NovoScribe 平台，厂商口径），Anthropic 官方新闻页 anthropic.com/news/driving-ai-transformation-with-claude 可替代死链；正文已加厂商口径限定
- 西班牙对外银行合作扩展公告：~~官方一手源待补~~ 已核实（2026-08-07 核验轮）：25 国 12 万员工、3,300 → 1.1 万 → 12 万扩张路径、2 万+ GPTs 其中约 4,000 高频使用，均与 OpenAI 官方案例页一致
- 雅诗兰黛 × ChatGPT 数据洞察（openai.com，Cloudflare 拦截）：素材卡无正文，未进正文，补源后再评估
- Klarna（OpenAI 客户故事页，Cloudflare 拦截）：与 123 号案例同事件，作废备查，以 Klarna 官方新闻稿为主源

说明：本书写作遵循「双源优先、单源标注」原则。所有「据报道」「第三方估算」「公司披露」「媒体追踪显示」的表述，均表示该数据来自单一或二手来源；出版前如需更

高置信度，建议以各公司官方披露、财报原文与一手访谈为准。