

行业研究笔记 · 2026 年 8 月

FDE 是什么 AI 落地的最后一公里生意

前向部署工程师 (Forward Deployed Engineer)：一个岗位一年涨 729%、两个月吸金 90 亿美元的背后，是 AI 行业的核心矛盾从模型竞赛切换到企业落地。本文回答三个问题：FDE 是什么，商业模式为什么成立，在中国怎么走。

基于 8 份券商研报与会议纪要（国盛证券、中信证券等，2026-08）及公开报道整理 · 仅供个人学习研究，不构成投资建议

核心观点：FDE 是 AI 支出周期切换的承接形态——从"买模型"到"买结果"，驻场第一次成为有复利的生意

+729%

全球 FDE 岗位一年涨幅 (643 → 5330 个, 2025.04 → 2026.04)

\$90 亿+

OpenAI / Anthropic / AWS / 微软两个月内承诺的前沿部署投入

5%

真正通过 AI 落地实现可量化生产力提升的公司比例

155%

Palantir Rule of 40 得分，全球百强市值科技公司第一

1 | 为什么是现在。三个边际变化同时发生：交付物从硬件集成变成 Agent 系统；交付人力从 30 人外包小组变成 5-6 人 FDE 小组 + Agentic Coding；付费从按人天变成按结果。驻场服务的成本结构第一次被打穿。

2 | 模式为什么成立。双向蒸馏（客户隐性经验→AI 可用资产；项目经验→平台能力）让交付边际成本递减：第 1 个客户成本 3x，第 10 个只要 4-4.5x。延迟收益 + 规模效应，Palantir 已用 82-86% 毛利率验证（传统咨询仅 30-40%）。

3 | 中国怎么走。落地条件已齐（国产模型性价比、低代码普及、定制化需求爆发），但付费习惯、人力成本基数、信创合规、柯达困境四道坎决定斜率。别拿 Palantir 的财务指标直接套 A 股标的。

4 | 辨别真假 FDE。看四个数字：同类客户上线周期、单客户交付人天、可复用资产占比、毛利率。低毛利的"FDE 项目"本质是旧瓶装旧酒。

一个反常的数字：FDE 岗位一年增长 729%，薪酬对标模型公司研发——"驻场"从二等公民变成最贵的人

岗位与薪酬

- 全球 FDE 岗位：2025 年 4 月 **643 个** → 2026 年 4 月 **5330 个**，同比 **+729%**；另有口径为 400+ → 5000+、"年内 +800%"
- 基础年薪 7.4 万美元；含股权平均包 **17.2 万美元**，70% 岗位配股权激励
- 招聘主体：超一半岗位由**初创 / A 轮公司**发布——没存量包袱的人跑得最快
- 国内招聘平台已挂出 20–35k×15 薪的 FDE 岗位，要求"全栈+懂业务"

行业温度计

2025 年 6 月，a16z 将 FDE 定义为"**创业圈最热门的岗位**"。风投不创造价值，但对边际变化敏感。

2026 年 5–7 月，头部厂商两个月下注超 90 亿美元

厂商	投入	动作
OpenAI	\$40 亿+	联手 19 家 PE 成立 Deployment Company (投前估值 \$100 亿)，收购英国 Tomoro，获约 150 名 FDE
Anthropic	\$15 亿	与黑石、高盛、H&F 组建企业 AI 服务合资公司 (各出约 \$3 亿 / \$1.5 亿)，阿波罗、红杉跟投
AWS	\$10 亿	千人级独立 FDE 组织 + 合作伙伴主导的 FDE 体系
微软	\$25 亿	专属机构 Frontier Company，6000 名行业专家与工程人员

2025 年这些公司还在卷模型跑分，2026 年集体掉头抢驻场工程师——**竞争焦点从"模型能不能做演示"切到"能不能进生产系统"**。

FDE 是什么：驻客户现场写代码、对最终业务结果负责的工程师——分界线不在技能，在"对什么负责"

岗位	对什么负责	怎么收费	交付完
售前	商机转化	绑在合同里	签单即撤
传统外包	代码量、工作量	按人天 / 人月	交付即两清
咨询	方案	按项目 / 工时	PPT 留下，经验带走
FDE	最终业务结果	订阅 / 用量 / 按效果	系统留下，经验沉淀回平台

对照组：埃森哲 70 万员工、人均创收 9 万美元——传统咨询卖的是工时。FDE 的考核反过来：系统没跑起来、指标没改善，就是 FDE 的失败。

Palantir 的分工：Echo 与 Delta

Echo 挖痛点、翻译成需求 ("应该做什么")；Delta 现场搭原型、部署迭代 ("怎么做出来")。打通"懂业务的不会写代码、会写代码的不懂业务"的老断层。

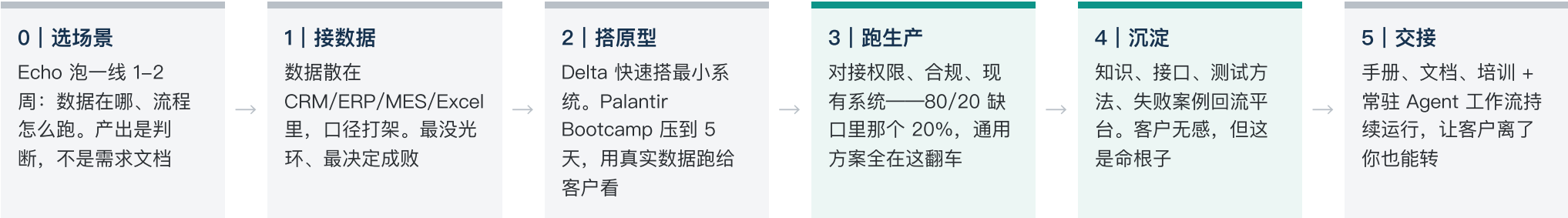
三件事

- 业务问题筛选：吃透商业模式、流程、组织，找出值得用 AI 做的场景（多数场景不值得，如实说）
- AI 工程落地：数据接入、工作流、原型、系统对接、调试部署
- 业务效果验证：拿业务指标说话，持续迭代

能力模型

计算/存储/网络架构 → 模型训练微调 → Agent 调试与 MCP 接入 → 系统集成，外加最稀缺的一条：**读懂客户的业务 SOP，翻译成 Agent 能执行的流程。**

一个 FDE 项目的完整流程：5-6 人小组、45 天冲刺（对照组：30 个外包干 6 个月）



为什么是现在

为什么以前跑不通、现在跑通了：供给侧三个边际变化 × 需求侧的落地剪刀差

① 交付物变了：动辄几千台服务器 + 超融合/私有云 → AI 服务和 Agent 系统。硬件集成最重的体力活没了。

② 交付人力变了：30 个外包 × 6 个月 → FDE 小组 + Agentic Coding。AI 吃掉代码生成、接口、测试、文档，人只做判断、架构和知识沉淀。

③ 付费方式变了：按人天 / 里程碑 → 结果导向或 License 部署验收。客户开始为"结果"付钱，不为"过程"付钱。

60%

开展过企业级 AI 工具评估

20%

进入落地试点阶段

5%

最终落地并创造百万美元级收益

23%

Agent 赛道规模化部署率（62% 开了试用）

80/20 缺口

通用大模型 + 基础提示词可覆盖企业 80% 的标准化常规业务；卡点集中在剩余 20%——长尾需求、异常场景、行业合规约束。这 20% 恰恰是企业愿意付高价的部分。多数公司每年花数百万至上千万美元采购模型 API，接口买回去落不了地，大量 Token 额度闲置。

采购逻辑转向

2025 年看	2026 年看
大模型榜单排名	任务成功率 80% → 95%
每百万 Token 单价	人工介入频次持续下降
API 接口调用	10 人流程精简到 3 人

所有落地成果都要折算成可量化的收入增长、成本压降和清晰 ROI。从"买模型"到"买业务结果"。

大型 AI 项目正从"一群外包+核心产品+一堆硬件+系统集成"的工业浊酒，切换成"几个人+业务 Agent+基础软件和数据+Agentic Coding"的酱香白酒。

AI 没有让 IT 预算缩水，只是换了流向；每轮范式切换都催生一代"翻译者"，这一轮轮到 FDE

\$6.15 万亿

2026 年全球 IT 支出预计 (2025 年 5.55 万亿)

\$1.43 万亿

其中软件支出，同比 +10.8%~14.7%

**<15% →
~25%**

AI 占 IT 总预算比例：2024 年 → 2027 年预计

预算重构的四个动作

压缩低频席位释放经费 | 砍重复工具去重降本 | 减少低价值 BPO | 把运营预算**转向智能体建设**。多数企业处于新旧系统并行的"重复付费期"——这正是 FDE 的进入窗口：帮企业把两边的账算成一边。

这一轮翻译的本质区别

以前的迁移把系统从 A 架构搬到 B 架构，交付即结束；这一轮把"人做事"的流程翻译成"Agent 做事"，**交付物是系统本身**，交付后持续干活、消耗 Token、产生数据回流——驻场第一次从一次性买卖变成经常性收入入口。

每轮基础设施切换都有自己的"最后一公里"行业



AWS 自己就是活样本：亲历虚拟化→容器化→微服务化→Agent 化四轮改造，现在把下一个万亿级增长点押在 **AI 需求杠铃的中段——企业生产负载的 AI 化**。杠铃一头是模型研发，一头是消费级应用，中段的企业生产系统以前最不起眼，现在最肥。

核心机制"双向蒸馏"：客户的隐性经验和厂商的项目经验，同时被提纯成可复用资产

蒸馏方向	蒸馏对象	形成资产	效果
对客户侧	员工脑中的隐性业务经验、流程规则、操作逻辑	知识库、工作流、SOP、业务应用	零散经验显性化，换谁来操作系统经验都在
对厂商侧	行业知识、系统接口、测试方法、失败案例	平台组件、行业模板、本体体系、Skill、测试集	下一个同类客户简单配置 + 少量定制即可上线

失败案例尤其值钱。它是测试集和边界条件的来源——踩过的坑决定了下一个项目的交付速度。

飞轮闭环。FDE 现场蒸出隐性经验 → Harness 把它变成机器可重复执行的能力 → 系统上线产生运行数据 → FDE 甄别问题根源（模型/数据/权限/规则）→ 回头优化。转一圈，飞轮快一格。

成本结构对比：规模效应的来源

模式	第 1 个客户	第 10 个客户	组织结果
项目制交付	1×	8—9×	人越多，收入和成本一起涨
本体化交付（FDE）	3×	4—4.5×	前期更重，后期复用显著增加

前几个客户是亏的或微利的，复利从后面开始。FDE 的财务模型注定是**延迟收益**：初期试点允许亏损换信任，验证后横向拓展客户其他部门，做大合同、提升续约粘性。**准入门槛因此是"前期亏得起"**——排头兵是云厂商和拿巨额融资的模型公司，不是偶然。

FDE 典型应用场景的特点：领域足够大、客户足够多，前期投入才能获得长期复利。垂直但小众的场景撑不起这个模式。

计费三级跳：人天 → 订阅/用量 → 按结果付费；辨别真假 FDE 看四个数字

计费模式三级跳



国内明略科技已在往 result as service 转：驻场梳理场景、部署定制化 Agent、对接全系统、验证 ROI，按效果收钱。从"卖工具"到"卖价值"，收费方式的变化是商业模式变化的最后一公里。

运营、计费、激励的全面对比

	传统咨询外包	FDE 模式
运营	瀑布式项目制，提前固定需求	小团队端到端，交付同步完成产品迭代
成果归属	归客户，交付后无沉淀	经验沉淀为公司通用知识库
计费	按人天、人月	平台订阅、用量、按结果
员工激励	绑定工时饱和度	绑定客户业务指标与续约率
毛利率	33%—40%（埃森哲）	接近 80% 且持续上行（Palantir）
人员	随营收同步扩张	人员零增长、营收高增长

怎么判断是真 FDE：三条运营指标

- 同类客户上线周期持续缩短——沉淀在起作用
- 单客户交付人天持续下降——复用在起作用
- 可复用平台资产占比持续提升——飞轮在起作用

试金石：毛利率

传统咨询外包毛利率中枢 30—40%（埃森哲长期如此）。号称做 FDE 的公司如果毛利也是这个水平，本质是驻场外包换了个时髦名字。真 FDE 应向软件公司看齐——Palantir 2026Q2 调整后毛利率 86%、经营利润率 62%。

另外两个观测点

口碑案例：合作客户规模、落地效果、有无企业站台——第一个标杆客户几乎决定后面三年的生意。人才壁垒：全栈贯通 + 业务 SOP 翻译能力，培养周期长，薪酬对标模型公司研发。

Harness 是 FDE 手里的扳手：模型负责理解推理，Harness 负责让 Agent 在生产环境里不出事

Harness（马具）= 包裹在基础大模型外层的智能体运行时系统。模型越强，Harness 越重要。

它管五件事

- 上下文构建
- 工具筛选调用
- 权限合规控制
- 长周期任务状态维持
- 异常恢复

早期对话式 AI 是单次问答，出错重问一遍就是。现在的 Agent 任务是规划、检索、多轮工具调用、跨系统联动的长链条，一跑几个小时——**没有 Harness 的 Agent 是没装刹车的快车。**

它决定 Agent 的四条边界

- **可视范围**：能看见什么数据
- **操作权限**：能动什么系统
- **执行时机**：什么时候动手
- **错误处置**：出错了怎么办

Anthropic 的对比实验（据会议纪要转述）

	时长	成本	效果
简易单智能体	20 分钟	\$9	基础完成
全套 Harness	6 小时	\$200 (20 倍+)	完整度/准确率大幅提升

竞争不在功能数量，在**任务成功率与调用成本的平衡**——什么时候上重 Harness、什么时候够用就行，这个分寸感是核心壁垒。

AWS 的双层 Harness

智能体运行 Harness（管单 Agent 稳定运行）+ 交付 Harness（沉淀项目知识、跨客户复用）。前者管机器，后者管组织记忆。

关键判断：模型正在变得可替换（开源模型可用性大幅提升后尤其如此），真正有价值的是把业务 SOP 转成 Agent 系统的那套 Harness 和沉淀本身。

Palantir 二十年：驻场服务这个低毛利人力生意，做出了比大多数 SaaS 还好的财务指标

+93%

2026Q2 收入同比（\$19.35 亿，全年指引上调至 +82%）

155%

Rule of 40，全球百强市值科技公司第一（第二：英伟达）

157%

NDR 净收入留存率（上季 150%）

~1%

员工年增速——收入高增与人力完全脱钩

三个阶段

- 2003–2008 情报机构：Peter Thiel 等 2003 年创立，早期获 CIA 旗下 In-Q-Tel 投资。工程师带安全许可进驻军方/情报保密环境，沉淀 Gotham 与 Echo+Delta 架构——进最复杂的现场，啃最硬的数据
- 2016 起商业化：Foundry 面向企业，核心资产是统一语义层 **Ontology**，让系统“懂业务”而不是裸读数据库
- 2023 起 AIP：叠加大模型能力层；Bootcamp 五日驻场用真实数据搭原型，转化率约 70%、获客成本降 40–60%；上线周期缩到两三天

关键指标

年收入 2021 年 \$15 亿 → 2025 年约 **\$45 亿**；毛利率 **82–86%**；剩余合同价值超 \$60 亿；无负债。

对照传统咨询龙头（埃森哲）

	Palantir	埃森哲
员工规模	五年新增 4500 人	70 余万人
人均创收	\$140 万	\$9 万（15 倍差）
毛利率	82–86%	30–40%
增长逻辑	平台复用，复利增长	人力堆叠，线性增长

产品线：Gotham（情报/安全，FDE 诞生地）→ Foundry（企业数据操作系统，主战场）→ AIP（2023 推出的 AI 能力层）→ Ontology 贯穿三者，是二十年沉淀的核心资产。

注：2026Q2 财报数据（收入 +93%、Rule of 40 155%、NDR 157%、调整后经营利润率 62%）已与 Palantir 官方披露交叉验证。

五类玩家五种算盘：CSP 是排头兵，模型公司在补短板，咨询公司是冲击方

玩家	打法	优势	软肋 / 变数
CSP 云厂商 (排头兵)	AWS：专属 FDE 团队嵌入客户业务/安全团队，按月交付压到按天，双层 Harness；微软自建 Frontier Company	全栈人力 + 基础软件生态（数据库/大数据/安全），能做跨系统整套交付，中立方	自家算盘：FDE 交付的常驻 Agent 持续烧 Token——谷歌云月 Token 处理量一年 7 倍，\$1 万亿+ 年耗客户两月内 375→500 家
模型公司	OpenAI 收购 + 合资（Tomoro、BCG/贝恩/埃森哲合作）；Anthropic 与 PE 合资，Applied AI 工程师进驻被投资企业	模型能力最近，资本最厚	做不全——软件公司、咨询公司作为伙伴的价值会显现（中信证券）
独立 FDE	Palantir 独一档	Ontology + 政府级场景背书，Sovereign AI	稀缺，短期无人能复刻
SaaS 厂商	ERP/CRM 的 MCP 化改造做引流入口	现网系统高粘性	企业内部远不止 ERP，还要和 CSP 正面竞争，格局待观察
咨询公司	被冲击方	存量客户关系	Gartner：2026 年底 85%+ 科技服务商将以 FDE 项目为核心交付方式——转不过来就是生死问题

对 PE 的账：帮被投资企业落地 AI、改善利润，退出时直接体现为估值倍数——这是黑石、高盛们愿意掏钱的底层逻辑。

中国：落地条件已齐，但四道坎决定斜率——方向确定，别拿 Palantir 的指标直接套 A 股

条件已齐

海外验证完成（Palantir 财报） | 头部路径清晰（自建或收购） | 国产模型性价比斩杀线（DeepSeek） | 低代码工具普及 | 定制化需求爆发 | 软件行业五年衰退结束，70-80% 转型滞后公司承压——FDE 是出口也是生死线。

四道坎

- **付费习惯**：预算薄、定制重、回款慢，政企连订阅制都没完全接受，何况"按结果付费"
- **人力成本基数**：中国工程师成本是美国 1/3-1/2，Agent 省下的红利绝对值更小，毛利天花板更低
- **信创合规**：数据不出内网、国产化替代抬高交付成本；好消息是国产模型栈（DeepSeek/Kimi/智谱）已过及格线
- **柯达困境**：存量业务被 AI 冲击，内部抵制是常态——超一半 FDE 岗位是初创/A 轮公司在招

结论：方向确定，斜率未知。国盛的选股框架两类——传统软件转型公司（存量客户+FDE 深挖）与 FDE 基础平台供应商（卖算力和工具链）。看一家公司值不值得信：上线周期、交付人天、可复用资产占比，加一个毛利率。

国内玩家已经在动

类型	公司与打法
平台型	腾讯（Codebuddy/Workbuddy 降门槛，FDE 重心抬到场景筛选）；阿里云（开 FDE 岗位，百炼/悟空平台沉淀一线经验）
传统 IT 转型	软通动力（AI 工厂+FDE，算力集群 + Token 工厂做配套）；中软国际（AI 训练营+FDE 团队）；极视角（FDE 专责交付，知识回流产品）
行业软件	金蝶（Harness 内嵌 ERP 底层）；科大讯飞、赛意信息；鼎捷数智（EIOSpace 企业智能运行空间）
垂直行业	工业：中控技术（中石油场景，已跑通按结果计费）、宝信软件；金融/能源：明略科技（实控普联软件）、海智科技；法律：华宇软件；医疗：卫宁健康、麦迪科技；AI4S：英矽智能、华兰股份；安全：亚信安全

对个体是杠杆位，对行业是支出周期切换；最大的不确定性是模型会不会吃掉 FDE 自己

对从业者

- **三者都要**：工程能力独立交付系统 + 业务理解能和老师傅对话 + 沟通推进能在甲方组织里推动变革
- **补课方向**：技术背景补业务；业务背景补 Agentic Coding 的工程手感——不用自己写每行代码，但要知道系统边界在哪
- **杠杆位**：一年见的场景比总部研发五年都多，"独立交付过生产级 Agent 系统"会越来越值钱
- **风险**：岗位曲线符合热潮规律（稀缺→高薪→供给涌入→均衡）；横跨业务和技术，供给爬坡慢，红利期可能更长

支出周期切换

2023–2025 的大头是 CSP 基建投入（抢卡、建机房）；2026 年起增量转向企业侧——主权 AI、生产负载 AI 化。a16z：财富 500 强的 29%、全球 2000 强的 19% 已是头部 AI 创业公司付费客户，B 端渗透刚开始。

Token 经济的自我强化

FDE 落地越多 → 常驻 Agent 越多 → Token 消耗越大 → 云厂商越有钱投 FDE。硬件传导：长上下文、多轮工具调用推高 KV cache 占用，对推理算力和 HBM 是刚性拉动。单客价值膨胀已可见：Adobe 年付费 \$100 万+ 客户 2 月 500 余家 → 4 月破 1000 家；Runway 营收 2 月 \$1.4 亿 → 5 月 \$4.7 亿。

最大的不确定性：模型会不会吃掉 FDE。如果 Agent 强到自己读懂企业系统、自己完成集成，驻场的必要性就下降。目前的答案是不会——80/20 缺口还在，合规和权限是信任问题不是能力问题。但这是需要持续跟踪的假设，不是背景板。

伪 FDE 会泛滥。每轮概念热潮都伴随贴牌：驻场外包改名叫 FDE，人天报价单改名叫"按效果付费"。四个数字比任何路演都诚实：毛利率、上线周期、人天投入、可复用资产占比。

技术从来不会自己走进企业，总要有人把它背进去。这一次不同的是，背进去的东西会自己沉淀下来，越背越轻。