

Token 与 AI Agent 冲击下的企业决策与市场结构：成本重组、竞争传导与福利分析

2026 年 6 月

摘要

Token 与 AI agent 进入企业以后，最先改变的并不只是某一项技术支出，而是企业完成任务、组织劳动、调用模型、审核错误和承担治理责任的整套成本结构。文章从岗位和任务暴露度出发，把企业采用能力、实际使用强度和业务规模纳入同一个成本账本，再将劳动成本、非劳动运行成本、固定采用成本和规模凸性压缩为 $f_i, m_i(a_i), d_i(a_i), \beta_i, \lambda_i$ 等可用于竞争分析的参数。生产型 agent 在有效区间内可以降低边际成本并扩大产量，但 token 容量、错误扩散、治理投入和组织改造会限制深度采用。供给侧分析以 Cournot 为完整求解的基准，并比较 Bertrand 价格竞争、差异化产品定价、主导企业与竞争边缘、进入退出和平台网络效应下的传导差异。若大型企业拥有更高采用能力、更低固定门槛或更强数据基础，AI/agent 可能扩大成本差距并提高集中度；若通用工具降低中小企业采用门槛，也可能削弱这种趋势。需求侧和策略侧进一步加入消费者搜索 agent 与企业 pricing agent：前者通过 $b \rightarrow s(b) \rightarrow \varepsilon(b)$ 降低搜索摩擦并压缩加价空间，后者通过个性化报价能力 φ_i 和重复定价策略改变剩余分配与合谋风险。电商平台 A/B/C 的风格化案例显示，效率收益、搜索收益、价格歧视、合谋风险和治理成本会同时出现，福利方向取决于这些力量的相对大小。

关键词： AI agent；token 成本；企业采用；成本函数；Cournot 竞争；价格竞争；进入退出；市场集中；搜索摩擦；算法定价；福利分析

目录

1 引言：从一个企业问题开始	5
2 系统文献综述	7
2.1 从文献脉络到建模对象	7
2.2 AI、自动化与任务理论	8
2.3 AI 暴露度、使用程度与企业采用	8
2.4 Token、LLM 成本与模型规模经济	9
2.5 数字经济、数据经济与市场集中	10
2.6 算法定价、AI agent 与竞争博弈	10
2.7 文献空白与研究定位	11
3 基础模型：从 AI 暴露度到企业成本函数	11
3.1 从企业成本函数开始	11
3.2 企业 AI 暴露度：先从有限任务加权平均开始	12
3.3 暴露度、采用能力与实际使用强度	13
3.4 人力成本：从单个任务到企业总劳动	14
3.5 资本与协调成本	15
3.6 Token 成本：先递减、后稳定、再递增	16
3.7 审核、错误与监管成本	16
3.8 企业类型：用参数集合表示异质性	17
3.9 边际成本账本与成本压缩入口	18
4 人力资本结构：从任务劳动到组织重组	20
4.1 为什么人力成本要单独讨论	20
4.2 从任务积分到 routine/skilled 两类劳动	21
4.3 给定产量下的替代效应与互补效应	22
4.4 企业扩张后的总劳动需求	23
4.5 人力资本强度与工资结构	25
4.6 人力成本如何进入成本参数	26
5 非劳动智能体成本：统一框架、调用与规模风险	27
5.1 非劳动成本的归属规则	27
5.2 Token 调用成本与容量约束	27
5.3 非劳动运行成本：审核、治理与协调	29
5.4 固定采用与治理投资	31
5.5 规模凸性、错误扩散与风险控制	32
5.6 非劳动成本进入压缩模型	33

6	成本压缩与采用选择	33
6.1	为什么需要成本压缩	34
6.2	从劳动数量到劳动成本参数	34
6.3	非劳动成本如何进入参数	35
6.4	简化成本函数	36
6.5	条件采用强度	36
6.6	固定门槛与规模效应	37
6.7	规模凸性	38
6.8	通向市场竞争模型	38
7	市场竞争与市场结构	39
7.1	从成本压缩进入竞争模型	39
7.2	无 AI 的两企业 Cournot 基准	39
7.3	给定采用强度下的两企业 Cournot	40
7.4	对称企业与 n 企业推广	41
7.5	采用差异、成本差距与利润差距	42
7.6	市场份额与 HHI	44
7.7	规模凸性修正	45
7.8	Cournot 之外的市场结构	46
7.9	供给侧竞争结果	48
8	消费者搜索与企业定价机制	49
8.1	需求侧和策略侧的新摩擦	49
8.2	消费者搜索 agent：从搜索成本到需求弹性	49
8.3	搜索成本下降如何改变加价率	50
8.4	企业 pricing agent：实时定价与个性化报价	51
8.5	pricing agent 与算法合谋风险	52
8.6	两类 agent 同时存在时的四种情形	52
8.7	三类 agent 的合成效应	53
9	统一案例：电商平台竞争	53
9.1	电商平台作为统一场景	53
9.2	案例设定：三个电商平台	54
9.3	生产型 agent：从采用强度到成本和份额	55
9.4	消费者搜索 agent：比价能力与加价空间	56
9.5	企业 pricing agent：个性化报价与合谋风险	57
9.6	反事实：中小平台获得低成本 agent 工具	58
9.7	三条通道的共同作用	59

10 福利、政策与结论	59
10.1 为什么福利不能只看平均价格	59
10.2 供给侧效率收益：成本下降、产量上升与固定采用成本	59
10.3 消费者搜索 agent：搜索摩擦下降与消费者剩余	60
10.4 企业 pricing agent：价格歧视、分配与合谋风险	61
10.5 统一福利账本	61
10.6 政策含义：按参数干预	62
10.7 结论、局限与未来研究	63
A 主要符号表	64

1 引言：从一个企业问题开始

如果一家电商平台在客服、商品推荐、广告投放、商家运营和价格监测中引入 **AI agent**¹，它最先面对的往往是几笔具体的账：哪些任务会被加速，哪些岗位会减少重复劳动，哪些环节反而需要更强的监督、判断和系统整合？模型调用和 **token**² 费用会不会随业务规模上升？自动回复、自动审核或自动报价出错以后，复核和责任成本由谁承担？当系统进入金融、医疗、教育或法律等高风险场景时，合规和治理成本是否会吃掉一部分效率收益？

这些问题可以落到企业**成本函数**上。企业仍然选择产量或业务规模 q_i ，并在市场需求下最大化利润：

$$\max_{q_i, a_i} \pi_i = P(Q)q_i - C_i(q_i, a_i).$$

这里 $a_i \in [0, 1]$ 表示企业实际 AI/agent 使用强度³。关键变化在于， $C_i(q_i, a_i)$ 不再只是固定成本加上一个外生边际成本。AI/agent 会同时改变劳动需求、协调摩擦、token 调用、错误审核、治理合规和固定采用成本。论文关心的核心问题是：token 与 AI agent 普及后，企业成本函数如何被重组，这种成本重组又如何传导到采用选择、市场竞争、份额、集中度和福利？

一个过于简单的写法是直接假定 AI 降低边际成本。这个写法容易得到清楚结论，却会遮住最重要的经济机制。生产端 agent 可能节约客服初筛、文案生成、库存提醒和流程沟通成本，也可能增加提示词设计、模型监督、系统集成和异常处理劳动。非劳动成本也会随使用强度上升：token 调用、延迟、错误扩散、安全审计、责任保险、合规记录和固定系统改造都是真实成本。需求端和策略端还有另一层变化：消费者搜索 agent 让比价和识别隐藏费用更便宜，企业 pricing agent 让个性化报价和高频价格反应更容易。只看其中一条通道，市场结构和福利判断都会偏。

为把这些机制纳入同一分析框架，成本先写成完整账本，再在需要求解竞争均衡时压缩。岗位、任务**暴露度**⁴和企业**采用**⁵能力决定实际使用强度；人力成本被拆成常规任务节约、高技能互补和复杂性劳动；token、审核、治理和固定采用被整理为**非劳动 agent 成本**。经过局部近似以后，复杂账本可以折叠为几个有来源的参数：

$$f_i, \quad m_i(a_i), \quad d_i(a_i), \quad \beta_i, \quad \lambda_i.$$

β_i 表示有效节约能力， λ_i 表示使用强度上升带来的综合成本压力， f_i 表示固定采用和组织改造成本， $d_i(a_i)$ 表示业务规模扩大后的拥挤、错误扩散和容量压力。相应的简化

¹**AI agent** (AI agent; 人工智能智能体)：能够感知环境、调用工具并连续执行任务的软件系统，本文特指嵌入企业流程的生产型、搜索型或定价型智能体。

²**token** (token; 词元/令牌)：大语言模型处理文本时的最小计价与调用单元，企业按调用量或字符/词元数支付模型使用成本。

³**使用强度** (usage intensity; AI 使用强度)：AI/agent 在日常生产流程中从浅层辅助到深度嵌入的连续程度，用 $a_i \in [0, 1]$ 表示。

⁴**暴露度** (exposure; AI 暴露度)：某一职业、任务或企业业务流程与当前 AI/LLM 能力之间的匹配程度，衡量潜在可受影响范围。

⁵**采用** (adoption; AI 采用)：企业实际购买、部署并重组流程以使用 AI/agent 的决策与行为。

成本可以写成

$$C_i^{red}(q_i, a_i) = \left[c_i^0 - \beta_i a_i + \frac{\lambda_i}{2} a_i^2 \right] q_i + \frac{f_i}{2} a_i^2 + \frac{d_i(a_i)}{2} q_i^2.$$

这个式子牺牲了部分细节，但保留了最需要进入竞争模型的内容：采用强度如何改变单位成本、固定成本如何形成门槛、规模扩张为什么可能遭遇新的边际压力。

在供给侧，线性需求 $P(Q) = A - BQ$ 下的 **Cournot**⁶模型提供了一个完整可解的基准。给定企业已经形成的采用强度 a_i ，有效边际成本 $m_i(a_i)$ 进入反应函数和均衡产量；固定采用成本不进入当期一阶条件，却会通过采用门槛、净利润和长期进入退出影响哪些企业能够深度采用。仅有 Cournot 还不够，因为不同市场结构会改变成本冲击的传导强度。价格竞争中，成本差距可能更快转化为份额跳变；差异化产品市场中，需求弹性决定降本有多少传给消费者；主导企业与竞争边缘的结构会放大或削弱领先企业优势；进入退出模型会把固定采用成本放回长期企业数量；平台网络效应则可能进一步强化数据、匹配和规模反馈。

消费者和企业自己的 agent 又改变价格形成过程。消费者搜索 agent⁷的覆盖率为 b ，它通过

$$b \rightarrow s(b) \rightarrow \varepsilon(b)$$

降低搜索摩擦并提高需求弹性。企业 pricing agent⁸的个性化报价能力记为 φ_i ，它可能改善匹配，也可能把更多消费者剩余转移给企业；在重复定价环境中，它还可能降低监测偏离和执行惩罚的成本。生产端、需求端和策略端的作用方向并不一致，因此福利不能简单归结为“成本下降带来降价”。

后文的核心做法是把 AI/agent 的冲击拆成可解释的成本账本：先区分劳动成本与非劳动成本，再把复杂成本压缩为 $f_i, m_i(a_i), d_i(a_i), \beta_i, \lambda_i$ 等少数参数，最后用这些参数推导企业的采用选择、Cournot 均衡、市场份额和福利分配。这个过程中，供给侧竞争、消费者搜索 agent、企业 pricing agent 和福利账本会被放进同一个电商平台案例，用来展示效率收益、搜索收益、价格歧视、合谋风险和治理成本如何同时出现。

本文是一篇理论和参数模拟为主的课程论文，不估计真实企业的 AI 效应。电商平台 A/B/C 案例也是风格化例子，不对应具体公司数据。它的用途是把模型变量和现实机制对齐：哪些成本被节约，哪些成本被新增，为什么固定成本不进入短期边际条件，为什么低成本未必自动转化为消费者福利，以及为什么同一种 AI/agent 技术在不同企业和不同市场结构中可能产生不同结果。

⁶**Cournot** (Cournot competition; 古诺竞争)：企业以产量为策略变量、同时决策的寡头竞争模型，本文作为供给侧基准。

⁷**消费者搜索 agent** (consumer search agent; 消费者搜索智能体)：帮助消费者自动比较价格、识别隐藏费用和匹配替代选项的 AI 工具。

⁸**企业 pricing agent** (pricing agent; 企业定价智能体)：企业用于实时调整价格、识别用户类型并执行个性化报价的算法系统。

2 系统文献综述

AI/agent 对企业和市场结构的影响分散在几组相邻文献中。任务理论解释技术如何作用于劳动和技能结构；AI 暴露度与企业采用研究强调潜在冲击和实际使用之间的差距；LLM 与 token 成本文献提醒我们，模型调用、容量和治理并非零成本；数字经济和市场势力研究讨论数据、规模经济和固定成本如何改变集中度；算法定价与搜索摩擦文献则把注意力放到需求侧和策略互动上。把这些线索合在一起，才能解释为什么同一类 AI/agent 技术可能同时改变企业成本、采用门槛、市场份额和福利分配。

2.1 从文献脉络到建模对象

表 1 概括了各组文献与后文变量之间的对应关系。表格不把文献压成变量清单，而是给出阅读顺序：先看任务和暴露度如何决定使用强度，再看使用强度如何进入劳动成本、非劳动成本、竞争均衡和福利账本。

表 1: 文献组与模型变量的对应关系

文献组	对应变量	进入模型的方式
任务理论与自动化	$C_i^L, L_R(a_i), L_S(a_i), H_i(a_i), \beta_i^L, \lambda_i^L$	解释 AI/agent 为什么一方面替代常规任务，另一方面增加监督、判断和系统整合劳动，并把劳动数量变化转化为货币成本。
AI 暴露度与企业采用	$E_i, E_i^w, \theta_i, u_i(x), a_i, f_i$	区分 exposure、adoption 和 usage intensity，说明同样暴露度下，不同企业会因为数据、流程和管理能力不同而形成不同实际使用强度。
Token 与 LLM 成本	$X_i, C_i^{tok}, F_i^N, v_i^N, \Xi_i^N, \lambda_i^{tok}, d_i(a_i)$	解释 token 调用、容量约束、延迟、错误复核、模型治理和基础设施投入如何分别进入非劳动成本、固定成本和规模凸性。
数字经济与市场集中	$m_i(a_i), d_i(a_i), q_i^*, P^*, s_i, HHI$ $p_i, MC_i, \varepsilon_i, N^e$	说明数据、规模经济和固定成本摊薄如何把采用能力差异转化为成本差异，再通过数量竞争、价格竞争、加价率和进入退出影响份额与集中度。
算法定价与 agent 博弈	$b, s(b), \varepsilon(b)$ $\varphi_i, \delta, \Delta W^{search}, \Delta W^{price}$	消费者搜索 agent 改变需求弹性，企业 pricing agent 改变个性化报价和重复定价策略，二者共同影响消费者剩余和合谋风险。
风险、治理与政策	$z_i, \nu_i(z_i), E^{risk}, K_i^{adopt}$	说明治理既是成本，也是降低错误、隐私、责任和能源外部性的风险控制投资，最终进入福利账本。

表 1 的作用是把各组文献和模型变量对应起来，而不是把文献压成变量清单。任务理论关心的是岗位和任务如何被技术重塑，这解释了劳动成本为什么要从任务层面推起，以及 routine 劳动和 skilled 劳动为什么会朝不同方向变化。AI 暴露度与企业采用研究则强调，任务暴露度高不等于企业已经深度采用，因此 E_i 和实际使用强度 a_i 之间隔着企业能力、组织成本和互补投资。LLM 与 token 成本文献提醒我们，模型调用、容量、审核错误和治理合规并非零成本，这些成本先作为独立账本出现，随后才在竞争模型中被局部压缩为 λ_i 和 $d_i(a_i)$ 。数字经济与市场势力研究说明，成本差距如何通过数量竞争、价格竞争、加价率和进入退出转化为份额与 HHI 的变化。算法定价与搜索摩擦文献则把 b 、 φ_i 和合谋风险放到福利分析的需求侧和策略侧。

2.2 AI、自动化与任务理论

自动化经济学最重要的启发是：技术冲击首先作用于任务，而不是直接作用于“职业”或“企业”。D. H. Autor (2015) 解释了为什么自动化虽然替代许多常规任务，却没有消灭全部工作，因为人类在灵活判断、复杂沟通和非常规任务中仍有互补优势。Acemoglu and Restrepo (2018) 与 Acemoglu and Restrepo (2019) 进一步把技术影响分解为 **displacement effect**⁹ 和 **reinstatement effect**¹⁰：一方面机器替代既有劳动任务，压低劳动需求；另一方面新任务出现，重新创造劳动需求并改变技能结构。

这组文献给出的第一个启发是，人力成本变化不能简单写成“AI 减少所有人工”。更合理的处理是区分 **routine labor**¹¹ 和 **skilled labor**¹²。AI agent 可以自动处理客服初筛、文档摘要、代码补全、库存提醒、广告素材生成等重复性任务，从而降低 routine labor 需求；但企业也需要提示词设计、系统集成、输出审核、风险治理、数据标注和业务流程重构等高技能劳动。因此，AI 使用强度 a 在模型中同时降低 $L_R(a)$ 并提高 $L_S(a)$ 。

经验文献也支持这种区分。Acemoglu and Restrepo (2020) 研究机器人对美国劳动市场的影响，发现自动化冲击具有显著分配后果。Graetz and Michaels (2018) 的跨国行业研究显示，机器人提升生产率和 TFP，同时影响低技能劳动份额。D. Autor and Salomons (2018) 进一步指出，自动化带来的生产率增长、就业和劳动份额变化可能在行业内和行业间呈现不同方向。即使 AI/agent 降低企业成本，它对劳动结构和福利的影响仍可能是不均匀的。

Agrawal et al. (2019) 把 AI 概括为预测成本下降，这对企业决策尤其重要。预测成本下降不仅意味着少雇几个员工，也意味着库存、推荐、广告、信用评估、客服分流和生产排程等决策的边际质量提高。这里的 β_i 因此不只代表工资节约，也代表由于更好预测和更快响应带来的运营成本节约。与此同时，Acemoglu (2024) 对 AI 宏观生产率的谨慎估计提醒我们，AI 可有效替代的任务比例有限，不能把 β_i 设定成无限大或对所有企业相同。

2.3 AI 暴露度、使用程度与企业采用

第二组文献关注 AI exposure、adoption、usage intensity 和 productivity effect 的区别。暴露度衡量某个职业、行业或地区的任务与 AI 能力之间的匹配程度；采用表示企业是否真的购买、部署和重组流程；使用程度表示企业日常生产中实际依赖 AI/agent 的强度；生产率效应则是采用后实际产生的成本节约或质量提升。

⁹**displacement effect** (displacement effect; 替代效应)：机器或 AI 替代既有劳动任务、压低劳动需求的效应。

¹⁰**reinstatement effect** (reinstatement effect; 复归效应)：自动化后新任务出现并重新创造劳动需求、改变技能结构的效应。

¹¹**routine labor** (routine labor; 常规劳动)：重复性、可编码、容易被 AI/agent 自动化的任务劳动，如客服初筛、表格录入和库存提醒。

¹²**skilled labor** (skilled labor; 高技能劳动)：依赖判断、沟通、系统整合和风险治理的劳动，AI 采用后往往需求上升。

Felten et al. (2021) 构造了职业、行业和地理 AI 暴露度指标, 说明 exposure 可以帮助识别潜在冲击。Felten et al. (2023) 和 Eloundou et al. (2024) 将这种思想扩展到语言模型和 GPT 类技术, 显示大量文本、沟通和认知任务具有较高 LLM 暴露度。问题在于, 暴露度高并不自动等于企业采用强度高。一个会计、客服或软件开发岗位可能在技术上暴露于 LLM, 但企业是否采用取决于数据安全、流程兼容、员工培训、法律责任和 token 成本。

企业采用文献正好补上这一点。Acemoglu, D. Autor, et al. (2022) 使用在线职位数据观察 AI 岗位需求, 说明 AI 采用常常先表现为职位结构和技能需求变化。Alekseeva et al. (2021) 发现 AI 技能需求上升且行业差异明显。Babina et al. (2023) 从企业层面显示 AI 与企业增长和产品创新相关。这些研究共同说明, AI 冲击并非外生常数, 它会被企业能力和市场环境重新塑造。

生成式 AI 的微观实验进一步说明了采用收益的异质性。Noy and Zhang (2023) 发现生成式 AI 能提高写作任务的速度和质量。Brynjolfsson, Li, et al. (2025) 在工作场景中研究生成式 AI 的生产率影响。Peng et al. (2023) 研究 AI 编程助手对开发者生产率的影响。Dell'Acqua et al. (2023) 提出的 **jagged technological frontier**¹³ 尤其重要: AI 在某些任务上显著提高效率, 在另一些任务上可能因为错误自信、边界不清或监督不足而降低质量。因此, β_i 和 λ_i 都被设为企业和任务异质的参数。

采用还需要互补投资。Brynjolfsson, Rock, et al. (2019) 强调通用技术的生产率效果常因组织学习、流程重构和互补资产而滞后。这类互补投资在模型中对应固定采用成本 f_i 。如果 f_i 很高, 即使 AI 暴露度高, 企业也可能只停留在试用而非深度采用。

2.4 Token、LLM 成本与模型规模经济

第三组文献解释 token 为什么应当进入成本函数。大语言模型在技术上依赖大规模数据、参数和计算。T. B. Brown et al. (2020) 展示了大规模语言模型的 few-shot 能力, 说明当模型能力达到一定水平后, 企业可以把自然语言接口嵌入多种业务任务。Kaplan et al. (2020) 的规模律研究表明, 模型性能与参数量、数据和计算之间存在系统关系。Hoffmann et al. (2022) 进一步讨论计算最优训练, 强调参数规模与训练 token 数之间存在权衡。

对企业而言, 这些技术事实的经济含义是: AI/agent 的边际使用不是免费午餐。每一次提示、检索、工具调用、代码运行、结果复核和多轮对话都会消耗 token 或计算资源。即使模型供应商把单位 token 价格不断压低, 企业仍要面对上下文窗口、延迟、输出质量、失败重试、安全过滤和人工审核成本。因而, token 调用量、容量区间、审核损失和治理投入先被分开写入成本账本; 进入竞争模型时, 再把局部线性项和二次项压缩进 λ_i 、 $d_i(a_i)$ 和 f_i 。

同时, 基础模型具有明显的固定成本和规模经济。Bommasani et al. (2021) 把 **foun-**

¹³**jagged technological frontier** (jagged technological frontier; 锯齿状技术前沿): AI 在某些任务上效率很高, 在另一些任务上因错误自信或边界不清而表现不稳的技术边界。

dation models¹⁴ 描述为可迁移到许多下游任务的基础设施。这意味着上游模型训练和评估具有高固定成本，下游调用具有较低但非零的边际成本。Strubell et al. (2019) 与 Patterson et al. (2022) 从能源和计算效率角度讨论训练成本，一方面提醒我们大模型成本和外部性真实存在，另一方面也说明硬件和算法效率改进会改变成本曲线。

大语言模型还带来治理成本。Bender et al. (2021) 与 Weidinger et al. (2022) 强调语言模型的偏见、幻觉、误导、隐私和责任风险。企业如果把 AI agent 接入客户沟通、财务、法律或定价系统，就需要建立监控、审计、人工复核和事故响应流程。这些成本不一定随着 token 单价下降而消失，反而可能随着使用范围扩大而上升。这里的 λ_i 因此不只是技术 API 单价，更是每单位 agent 使用强度引出的综合治理成本。

2.5 数字经济、数据经济与市场集中

第四组文献解释为什么 AI/agent 采用可能改变市场结构。Goldfarb and Tucker (2019) 总结了数字经济中搜索、复制、运输、追踪和验证成本下降的影响。AI agent 可以被看作进一步降低这些成本的工具：它可以自动搜索供应商、比较价格、生成广告、识别客户、处理订单并优化库存。但成本下降并不必然带来更分散的市场，因为数字经济同时存在网络效应、数据积累和规模经济。

数字经济文献提供了关键机制。Jones and Tonetti (2020) 强调数据的非竞争性，即同一份数据可以被多个用途重复使用。Farboodi and Veldkamp (2021) 建立数字经济模型，说明数据积累会影响企业学习和生产效率。若大企业拥有更多交易数据、更完整的客户反馈和更强的系统集成能力，它们可能获得更高的 β_i 和更低的 λ_i ，从而在 AI/agent 采用后进一步扩大成本优势。

市场势力文献则提醒我们，成本下降是否传导给消费者取决于竞争结构。D. Autor, Dorn, et al. (2020) 讨论 superstar firms 的崛起和劳动份额下降，强调高生产率企业扩大市场份额。De Loecker et al. (2020) 研究加价率和市场势力上升，说明企业成本与价格之间的传导可能不完全。放入本模型以后，这些观点对应异质企业在 AI/agent 采用后出现 $m_H(a_H) < m_L(a_L)$ ，再根据市场结构分别表现为 Cournot 产量差距、Bertrand 价格压力、差异化产品加价率变化，或长期进入退出门槛变化。Cournot 在后文作为可完整求解的基准，市场结构问题还需要价格竞争、差异化产品和进入退出等扩展。

2.6 算法定价、AI agent 与竞争博弈

第五组文献把注意力从生产端转向策略互动。消费者 agent 可以降低搜索成本，这一机制可以追溯到信息经济学。Stigler (1961) 强调获取价格信息本身有成本。Diamond (1971) 和 Burdett and Judd (1983) 说明搜索摩擦可以维持高价格或价格分散。若消费者 agent 自动比较价格、读取评价、预测质量和执行购买，搜索成本 s 下降会提高需求弹性，迫使企业更激烈竞争。

¹⁴**foundation models** (foundation models; 基础模型): 在大规模数据上训练、可迁移到多种下游任务的大规模 AI 模型，上游固定成本高、下游调用边际成本低但非零。

但企业也会使用 pricing agent。Varian (2019) 讨论 AI 在产业组织、算法定价和个性化策略中的作用。Z. Y. Brown and MacKay (2023) 研究定价算法竞争，说明算法改变企业对价格和需求信息的反应方式。Assad et al. (2023) 提供了算法定价在实际市场中的经验案例。企业 pricing agent 可以实时监测竞争对手、库存、用户画像和转化率，从而进行高频调价和个性化报价。

更强的风险来自算法合谋。Calvano et al. (2020a) 发现强化学习定价算法可能在没有明示合谋协议的情况下学到高价策略。Klein (2021) 和 Hansen et al. (2021) 也指出独立算法可能形成超竞争价格。Calvano et al. (2019) 与 Calvano et al. (2020b) 因此强调竞争政策需要关注算法定价和消费者保护。消费者 agent 和企业 pricing agent 需要放在同一机制中理解：前者降低搜索成本，后者提高价格歧视和策略学习能力，最终福利方向不确定。

2.7 文献空白与研究定位

已有研究分别讨论了 AI 替代任务、AI 暴露度、生成式 AI 生产率、LLM 成本、数字经济集中和算法定价风险，但相对较少把它们串成一条企业层面的传导链。本文试图填补的就是这个缺口：从潜在暴露度 E_i 到企业采用能力 θ_i ，再到实际使用 $u_i(x)$ ；从劳动成本 C_i^L 和非劳动成本 C_i^N ，到压缩后的 $f_i, m_i(a_i), d_i(a_i), \beta_i, \lambda_i$ ；最后进入 Cournot 基准、价格竞争扩展、进入退出、搜索摩擦、pricing agent 和福利账本。这样的处理能把 exposure、adoption、usage、market power 和 algorithmic pricing 放在同一条因果链里，而不是让它们停留在几组并列背景中。

3 基础模型：从 AI 暴露度到企业成本函数

3.1 从企业成本函数开始

企业引入 AI/agent 以后，利润最大化问题多了一个新的选择维度。企业仍然选择产量或业务规模 q_i ，但还要决定 AI/agent 在日常流程中使用多深，记为 $a_i \in [0, 1]$ 。因此，企业 i 的问题可以先写成

$$\max_{q_i, a_i} \Pi_i = P(Q)q_i - C_i(q_i, a_i), \quad Q = \sum_{j=1}^n q_j.$$

这里 Q 是行业总产量， $P(Q)$ 是市场反需求函数；在竞争模型中会使用线性反需求 $P(Q) = A - BQ$ 。 a_i 并非简单的“是否采用 AI”的 0-1 变量，而是连续强度：它表示 AI/agent 在企业流程中从浅层辅助到深度嵌入的程度。真正需要先弄清楚的是 $C_i(q_i, a_i)$ 的形状：哪些成本随 a_i 下降，哪些成本随 a_i 上升，哪些成本还会随业务规模 q_i 凸增。

如果直接写

$$m_i(a_i) = c_i^0 - \beta_i a_i + \frac{\lambda_i}{2} a_i^2$$

这样的有效边际成本，推导会很快，却难以解释 β_i 和 λ_i 从何而来。更稳妥的做法是先 把成本拆开：岗位和任务暴露度决定潜在影响，企业采用能力决定实际使用强度，人力、

协调、token、审核和治理再分别进入总成本。简化式只有在这些来源被说明以后，才有清楚的经济含义。

先写总成本的总账：

$$C_i(q_i, a_i) = F_i(a_i) + C_i^L(q_i, a_i) + C_i^K(q_i) + C_i^{coord}(q_i, a_i) \\ + C_i^{tok}(q_i, a_i) + C_i^{rev}(q_i, a_i) + C_i^{gov}(q_i, a_i) + C_i^{oth}(q_i, a_i).$$

其中 $F_i(a_i)$ 是固定采用成本，包括系统集成、数据清洗、员工培训和流程改造。固定采用成本不随当前产量 q_i 直接变化，但随采用深度变化：自然的基准是 $F_i(0) = 0$ ，通常 $F_i'(a_i) \geq 0$ ，而深度接入核心流程时还可能有 $F_i''(a_i) \geq 0$ ，因为从“偶尔使用工具”走向“重构组织流程”会越来越难。 C_i^L 是人力成本； C_i^K 是传统资本成本； C_i^{coord} 是沟通、交接、搜索和组织协调成本； C_i^{tok} 是 token 与模型调用成本； C_i^{rev} 是错误审核、返工和人工复核成本； C_i^{gov} 是伦理、合规与政府监管成本； C_i^{oth} 是其他未完全建模成本，例如网络安全、供应商锁定、数据泄露、品牌声誉损失、延迟、碳排放和内部组织阻力。 C_i^{oth} 的符号不预先规定：有些未列入项可能是额外成本，也可能是被 AI 附带节省的摩擦成本。保留这一项，是为了避免把现实中很难枚举的组织成本硬塞进某一个参数。

这个式子本身只是记账框架。数学上， a_i 可以同时进入多个成本项，而且不同项的方向可能相反。经济上，AI/agent 是一种组织技术：它既可能节省人工和协调成本，也可能增加 token、审核、监管和未预见成本。

3.2 企业 AI 暴露度：先从有限任务加权平均开始

如果直接说“企业暴露于 AI”，读者容易误解为企业受到一个统一冲击。更自然的想法是：企业有不同岗位，岗位里有不同任务，AI 对每个任务的作用不同。比如一家电商平台中，客服岗位包含“识别用户问题、检索订单、生成回复、处理投诉升级”等任务；运营岗位包含“写活动文案、查看销售数据、调整广告投放”等任务；程序员岗位包含“读代码、写测试、修 bug、设计架构”等任务。AI 对生成回复和写活动文案可能很有帮助，对复杂投诉责任判断或系统架构设计则未必同样有效。

定义 1（企业 AI 暴露度 E_i ）

先用一个有限任务版本说明。设 $\mathcal{O}_i$ 是企业 i 的岗位集合， $\mathcal{K}_o$ 是岗位 o 内部的任务集合。岗位 o 在企业 i 中的占比为 π_{io} ，岗位 o 内任务 k 的权重为 α_{ok} ，任务 k 的 AI 暴露度为 $e_{ok} \in [0, 1]$ 。那么企业的平均 AI 暴露度可以写成

$$E_i^{disc} = \sum_{o \in \mathcal{O}_i} \pi_{io} \sum_{k \in \mathcal{K}_o} \alpha_{ok} e_{ok}, \quad \sum_o \pi_{io} = 1, \quad \sum_k \alpha_{ok} = 1.$$

这个式子只是加权平均：岗位越重要， π_{io} 越大；某任务在岗位里越常见， α_{ok} 越大；任务越容易被 AI 影响， e_{ok} 越高。

把任务从几个离散格子推广到连续分布，就得到积分版本。设岗位内部任务或技能记为 $x \in \mathcal{X}_o$ ，任务密度为 $h_o(x)$ ，满足

$$\int_{\mathcal{X}_o} h_o(x) dx = 1.$$

任务层面的 AI 暴露度为

$$e(x) \in [0, 1].$$

企业层面的平均 AI 暴露度为

$$E_i = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} e(x) h_o(x) dx.$$

如果希望高工资岗位在企业成本中获得更大权重，还可以定义工资加权暴露度

$$E_i^w = \frac{\sum_{o \in \mathcal{O}_i} \pi_{io} w_o \int_{\mathcal{X}_o} e(x) h_o(x) dx}{\sum_{o \in \mathcal{O}_i} \pi_{io} w_o}.$$

这里 w_o 是岗位 o 的工资或单位劳动成本。工资加权版本的含义是：如果一个高度暴露的岗位工资很高，那么它对企业成本函数的影响也应更大；同样的技术暴露度，发生在高工资岗位和低工资岗位上，对企业成本的意义并不完全相同。

这个写法与 AI 暴露度文献相连。Felten et al. (2021)、Felten et al. (2023) 和 Eloundou et al. (2024) 都从职业、任务或技能角度衡量 AI/LLM 能影响哪些工作。Acemoglu, D. Autor, et al. (2022) 则提示，企业对 AI 的需求会体现在岗位和技能结构变化中。这里借用这一路径，但把暴露度放进企业成本函数：岗位结构不同，企业的 E_i 就不同；同样是电商平台，客服密集型企业 and 仓储物流密集型企业面对的 AI 冲击也不会相同。

这一小节真正要保留的是一个经济学区分：暴露度只是潜在可影响程度，不等于企业已经采用 AI，更不等于 AI 已经在生产中高强度运行。

3.3 暴露度、采用能力与实际使用强度

有了暴露度，还要问第二个问题：企业能不能把 AI 真正接进流程？同样是客服任务，一家大型平台可能有完整的订单数据库、知识库、权限系统和审核流程；一家小商家可能只有聊天记录和人工经验。前者更容易把 AI agent 变成生产工具，后者即使任务暴露度很高，实际使用也可能很低。

定义 2（实际使用强度 $u_i(x)$ ）

因此，定义任务层面的实际 AI 使用强度为

$$u_i(x) = \min\{1, a_i \theta_i e(x)\}.$$

其中 $a_i \in [0, 1]$ 是企业总体 AI/agent 使用强度， $\theta_i \geq 0$ 是企业采用能力， $e(x)$ 是任务本身的暴露度。直白地说， $e(x)$ 回答“这个任务多大程度上适合被 AI 影响”， θ_i 回答“企业有没有能力把这种影响接进流程”， a_i 回答“企业实际愿意把 AI 用到多深”。使用 $\min\{1, \cdot\}$ 是为了保证 $u_i(x) \in [0, 1]$ 。在内点区间，也就是 $a_i \theta_i e(x) < 1$ 时，可以近似写成

$$u_i(x) \approx a_i \theta_i e(x).$$

这个设定把三个概念分开：

$$\text{exposure} \neq \text{adoption} \neq \text{usage intensity}.$$

exposure 是任务和技术之间的潜在匹配; adoption 是企业是否部署 AI/agent; usage intensity 是 AI 在日常生产中被用到什么程度。

数学上, $u_i(x)$ 把任务异质性 $e(x)$ 和企业异质性 θ_i 合在一起。经济学上, 它说明 AI 冲击并非落在所有企业身上的同质外生冲击, 而是被企业能力和任务结构共同过滤后的结果。

3.4 人力成本：从单个任务到企业总劳动

人力成本是 AI/agent 最容易被误读的部分。若只看替代效应, AI 好像一定减少劳动; 若只看监督和治理, AI 又好像一定增加劳动。更稳妥的做法是先写单个任务的劳动需求, 再把任务加总。

设岗位 o 中任务 x 的基准单位劳动需求为 $\ell_o^0(x)$ 。当该任务的实际 AI 使用强度为 u 时, 单位任务劳动需求写成

$$n_o(x, u) = \frac{\ell_o^0(x)}{1 + \phi_o(x)u} + \chi_o(x)u + \frac{\eta_o(x)}{2}u^2.$$

这里 $\ell_o^0(x)$ 是没有 AI 时完成任务 x 的基准劳动需求, 三个新参数分别刻画 AI 对这项任务的三种作用。第一项表示效率提升: $\phi_o(x) > 0$ 越大, AI 越能减少完成同一任务所需的人力。第二项表示互补劳动: $\chi_o(x)$ 越大, 说明提示词设计、流程改造、模型监督、数据标注和异常处理等新劳动越多。第三项表示复杂性劳动: $\eta_o(x)$ 越大, 说明高强度使用后返工、审核、责任界定和跨部门协调越容易凸增。为了保持含义清楚, 可以把 $\ell_o^0(x), \phi_o(x), \chi_o(x), \eta_o(x)$ 都理解为非负参数。D. H. Autor (2015) 与 Acemoglu and Restrepo (2019) 的任务框架正好说明, 技术既会替代旧任务, 也会重组任务并创造新任务。

对 u 求导, 可以看出方向并不固定:

$$\frac{\partial n_o(x, u)}{\partial u} = -\frac{\ell_o^0(x)\phi_o(x)}{[1 + \phi_o(x)u]^2} + \chi_o(x) + \eta_o(x)u.$$

第一项为负, 是 AI 的直接节约; 后两项为正, 是互补劳动和复杂性劳动。若负项绝对值更大, 单位任务劳动需求下降; 若监督和复杂性项更大, 单位任务劳动需求反而上升。生成式 AI 的任务边界并不平滑, Dell'Acqua et al. (2023) 所说的 jagged frontier 可以理解为: 有些任务 AI 很强, 有些任务 AI 看似可用但容易出错, 于是 $\phi_o(x)$ 、 $\chi_o(x)$ 和 $\eta_o(x)$ 都会因任务而异。

把单个任务加总到岗位, 再乘以产量或业务规模 q_i , 得到企业人力成本:

$$C_i^L(q_i, a_i) = q_i \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} w_o n_o(x, u_i(x)) h_o(x) dx.$$

这里 w_o 是岗位 o 的工资或单位劳动成本, π_{io} 让不同岗位按企业内部占比进入总成本, $h_o(x)$ 则让岗位内部不同任务按重要性进入总成本。人力成本不能用“AI 替代多少

人”一句话概括，它由岗位结构、任务结构、工资结构和实际使用强度共同决定。展开 $n_o(x, u_i(x))$ 后就是

$$C_i^L(q_i, a_i) = q_i \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} w_o \left[\frac{\ell_o^0(x)}{1 + \phi_o(x)u_i(x)} + \chi_o(x)u_i(x) + \frac{\eta_o(x)}{2}u_i(x)^2 \right] h_o(x) dx.$$

这个式子还说明，AI 对就业和人力成本的方向并不必然为负。若把总劳动需求记为 $L_i(q_i, a_i)$ ，则

$$\frac{dL_i}{da_i} = \frac{\partial L_i}{\partial a_i} + \frac{\partial L_i}{\partial q_i} \frac{dq_i}{da_i}.$$

第一项是给定产量时 AI 对劳动的直接影响；第二项是规模效应。若 AI 降低成本并使企业扩张，那么即使单位产出所需劳动下降，总劳动需求也可能上升。现实中“AI 导致裁员”和“AI 企业继续招聘”并不矛盾，因为它们对应的是这个分解中的不同项。Alekseeva et al. (2021) 关于 AI 技能需求上升的证据，也说明企业可能减少某些岗位，同时增加另一些更高技能岗位。

3.5 资本与协调成本

传统资本成本先写成

$$C_i^K(q_i) = \varrho_i K_i(q_i),$$

其中 $K_i(q_i)$ 是支撑产量或业务规模所需的传统资本， ϱ_i 是资本使用成本或资本租金率。这里不用 r_i ，是因为后面还要讨论 review 成本；若同一个字母同时表示资本租金和审核函数，读者很容易把两类成本混在一起。AI/agent 通常不会直接让仓库、服务器、门店、运输设备或办公系统消失，但它会改变这些资本和劳动的配合方式。例如，AI 可以让客服系统、广告投放系统和库存系统之间的信息流动更快，也可以减少员工反复开会、查找文档和交接任务的时间。

为了从直觉走到公式，先看单位协调成本。没有 AI 时，单位业务有一个基准协调成本 κ_i^0 。低到中等强度的 AI 可以降低沟通、交接、搜索和文档整理成本，用 $-\kappa_i^1 a_i$ 表示；但高强度使用又会带来权限管理、流程重构、系统依赖和跨工具错误传播，用 $\kappa_i^2 a_i^2/2$ 表示。因此

$$c_i^{coord}(a_i) = \kappa_i^0 - \kappa_i^1 a_i + \frac{\kappa_i^2}{2} a_i^2, \quad \kappa_i^1, \kappa_i^2 \geq 0.$$

乘以业务规模，就得到

$$C_i^{coord}(q_i, a_i) = q_i c_i^{coord}(a_i) = q_i \left[\kappa_i^0 - \kappa_i^1 a_i + \frac{\kappa_i^2}{2} a_i^2 \right].$$

数学上，这一项同样具有“先下降、后可能上升”的形状。经济学上，它提醒我们，AI 的一个重要收益不只是替代某个员工，而是降低多人合作中的摩擦。对大型平台和复杂组织来说，这类协调节约可能非常大；对业务简单的小企业来说， κ_i^1 可能较小。

3.6 Token 成本：先递减、后稳定、再递增

AI/agent 与传统软件的一个明显差别是，它的使用往往按 token、调用次数、上下文长度、工具执行和模型推理计费。先定义企业总 token 使用量。若任务 x 每单位实际 AI 使用需要 $\tau_o(x)$ 个 token 或调用强度，则

$$X_i(q_i, a_i) = q_i \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \tau_o(x) u_i(x) h_o(x) dx.$$

这只是“单位业务 token 强度”乘以业务规模 q_i 。

接下来写 token 的边际成本。边际成本不一定恒定。最初，企业可能因为学习、缓存、批量调用和流程熟悉而降低单位调用成本；进入稳定阶段后，边际成本接近供应商报价或内部模型推理成本；当使用规模超过容量或供应约束时，延迟、算力、上下文窗口、稳定性、峰值需求和议价限制会推高边际成本。为表达这种形状，令

$$MC_i^{tok}(X) = \begin{cases} p_i^0 - \delta_i X, & 0 \leq X \leq \tilde{X}_i, \\ p_i^L, & \tilde{X}_i < X \leq \bar{X}_i, \\ p_i^L + \zeta_i(X - \bar{X}_i), & X > \bar{X}_i. \end{cases}$$

这里 p_i^0 是初始边际 token 成本， δ_i 表示学习、缓存和批量调用使边际成本下降的速度， p_i^L 是进入稳定调用阶段后的边际成本。 $\tilde{X}_i$ 是学习和批量化阶段的上界， $\bar{X}_i$ 是正常容量上界， ζ_i 表示超过容量后的边际成本上升速度。为了让分段函数不是任意拼接，参数应满足

$$p_i^0 - \delta_i \tilde{X}_i = p_i^L, \quad p_i^L \geq 0, \quad \delta_i, \zeta_i \geq 0.$$

第一条保证第一段和第二段在 $\tilde{X}_i$ 处连续；第三段在 $\bar{X}_i$ 处也连续，因为 $p_i^L + \zeta_i(\bar{X}_i - \bar{X}_i) = p_i^L$ 。

总 token 成本由边际成本积分得到：

$$C_i^{tok}(q_i, a_i) = \int_0^{X_i(q_i, a_i)} MC_i^{tok}(z) dz.$$

总 token 成本等于从零调用量到当前调用量的边际成本累积。这个写法也解释了为什么大企业并不总是 token 成本更低：大平台用量很大，更容易进入 $X > \bar{X}_i$ 的区间，此时模型供应、稳定性和峰值调用会显著提高边际 token 成本。

3.7 审核、错误与监管成本

AI/agent 的输出并不天然正确。越多业务流程交给 agent，越需要人工复核、错误追踪、异常处理和责任归属。为了说明审核成本为什么可能随规模凸增，可以从“错误概率乘以影响规模”开始。

设 AI 使用强度越高，错误或需要复核的概率越高。一个简单近似是

$$\Pr(\text{error} \mid a_i) \approx \omega_i a_i^2, \quad \omega_i > 0.$$

这里 ω_i 衡量企业或行业对 AI 错误的敏感程度； ω_i 越大，同样的使用强度 a_i 会带来更高的错误或复核概率。若企业规模为 q_i ，一次系统性错误影响的业务范围也可能随 q_i 扩大。于是错误扩散成本可以近似写成

$$\text{错误概率} \times \text{影响规模} \approx \omega_i a_i^2 \cdot q_i^2.$$

把比例系数合并到 $\nu_i/2$ ，再加上常规人工 review，得到

$$C_i^{rev}(q_i, a_i) = q_i \psi_i^{rev}(a_i) + \frac{\nu_i}{2} a_i^2 q_i^2.$$

其中 $q_i \psi_i^{rev}(a_i)$ 是常规审核成本， $\psi_i^{rev}(a_i)$ 可以理解为每单位业务需要付出的人工复核、抽检和异常处理成本。一般来说 $\psi_i^{rev}(a_i) \geq 0$ ，且 $\frac{d\psi_i^{rev}(a_i)}{da_i} \geq 0$ ：AI 用得越深，日常抽检、提示词治理、输出追踪和异常处理通常越多。但这里不要求 ψ_i^{rev} 本身承担全部凸性，因为更强的规模化错误扩散已经由 $\nu_i a_i^2 q_i^2 / 2$ 捕捉。 $\nu_i a_i^2 q_i^2 / 2$ 表示错误扩散和复核压力随使用强度和业务规模凸增。这个式子不声称真实错误一定精确等于 $a_i^2 q_i^2$ ，只是用一个简单凸函数表达“使用越深、规模越大，错误外溢越难处理”。

监管与伦理成本写成

$$C_i^{gov}(q_i, a_i) = q_i g_i(a_i; \rho_i) + F_i^{gov}(a_i; \rho_i),$$

其中 ρ_i 是行业监管强度， $g_i(a_i; \rho_i)$ 是随业务规模变化的单位监管或伦理成本， $F_i^{gov}(a_i; \rho_i)$ 是固定合规成本，例如审计系统、合规流程、数据权限改造和责任追踪机制。这里同样不强行给出一个具体函数，但需要它满足基本方向：通常

$$\frac{\partial g_i(a_i; \rho_i)}{\partial a_i} \geq 0, \quad \frac{\partial g_i(a_i; \rho_i)}{\partial \rho_i} \geq 0, \quad \frac{\partial F_i^{gov}(a_i; \rho_i)}{\partial a_i} \geq 0.$$

也就是说，AI 用得越深、监管越强，单位合规成本和固定合规成本一般不会下降。在高监管行业，还可以预期 a_i 和 ρ_i 之间存在互补压力：同样提高 AI 使用强度，在金融、医疗、法律、教育等行业会带来更大的边际合规负担，因为这些行业对解释性、隐私、审计、歧视、责任和政府合规要求更严格。普通电商营销和客服场景的 ρ_i 可能较低，但不能忽略消费者保护、虚假宣传和数据安全问题。

这两项说明，AI/agent 的成本不只是技术价格。即使 token 单价下降，错误审核和监管约束仍可能让深度采用变贵。也正因为如此，不同行业的最优 a_i 可能差别很大。

3.8 企业类型：用参数集合表示异质性

为了避免把企业机械分成少数几类，先用参数集合表示企业类型：

$$\Theta_i = (E_i, \theta_i, \{\pi_{io}\}_{o \in \mathcal{O}_i}, \{h_o\}_{o \in \mathcal{O}_i}, \bar{X}_i, \zeta_i, \rho_i, \text{Scale}_i).$$

其中 E_i 是企业暴露度， θ_i 是采用能力， $\{\pi_{io}\}_{o \in \mathcal{O}_i}$ 和 $\{h_o\}_{o \in \mathcal{O}_i}$ 表示岗位结构和任务分布， $\bar{X}_i$ 与 ζ_i 描述 token 容量和超容量边际成本， ρ_i 是监管强度， Scale_i 是企业规模。这里不用 S_i 表示规模，是为了避免后文把 s_i 用作市场份额时符号混淆。

这个参数集合不用于给企业贴标签，而是说明同样的 a_i 会如何塑造不同的成本曲线。大型数字平台通常有较高 E_i 、较高 θ_i 和较大 Scale_i 。它们的组织和数据基础较好，可能让单位采用收益更快进入 $m_i(a_i)$ 的下降部分；但规模也会让 $X_i(q_i, a_i)$ 更快接近甚至超过 $\bar{X}_i$ ，并让错误扩散项 $\nu_i a_i^2 q_i^2 / 2$ 更重要。因此，大平台未必只是“成本更低”的企业，它们也可能拥有更高的 $d_i(a_i)$ ，即更强的规模压力。

中小普通企业可能有不低的任务暴露度 E_i ，例如客服、销售、运营和文案任务都容易被 AI 影响，但由于数据分散、流程不标准、员工培训不足， θ_i 往往较低。对这类企业来说， $F_i(a_i)$ 相对于规模更重，深度采用的固定成本门槛更明显；同时由于 q_i 较小，token 容量和错误扩散未必马上成为主要问题。它们的成本曲线更可能表现为：浅层采用有一点节约，但 a_i 一旦提高，固定采用成本和组织改造成本迅速变得显眼。

高监管专业服务企业，如金融、医疗、法律和教育机构，可能同时具有较高 E_i 和较高 ρ_i 。它们的高工资岗位让 AI 的潜在节约很大，但 $\psi_i^{rev}(a_i)$ 、 $g_i(a_i; \rho_i)$ 和 $F_i^{gov}(a_i; \rho_i)$ 也更重。结果是， $m_i(a_i)$ 未必随着 AI 使用强度单调下降：低强度使用可能提高效率，而高强度使用会被审核、合规和责任追踪抵消。若业务规模扩大，复核和合规压力还可能进入 $d_i(a_i)$ ，使边际成本随规模上升更快。

AI-native 企业的特点是 θ_i 高、流程从一开始就围绕 agent 和自动化工具设计，因此 $F_i(a_i)$ 中一部分固定改造成本已经被前期组织设计吸收。它们可能更容易把人力效率、协调节约和自动化流程转化为较低的 $m_i(a_i)$ 。但这并不意味着它们没有上升成本：如果业务依赖外部模型供应， $\bar{X}_i$ 和 ζ_i 仍会决定高强度使用后的 token 成本；如果产品直接面对消费者，错误扩散和责任风险也会让 $d_i(a_i)$ 上升。

经济学上， Θ_i 告诉我们：企业异质性不能只写成“高效率/低效率”标签，它来自暴露度、采用能力、规模、token 供给和监管环境。市场结构分析中的成本差异和份额差异，就是这些参数差异的结果。

3.9 边际成本账本与成本压缩入口

总成本分解以后，边际成本也可以按同样方式分解：

$$\begin{aligned} MC_i(q_i, a_i) &= \frac{\partial C_i(q_i, a_i)}{\partial q_i} \\ &= MC_i^L(q_i, a_i) + MC_i^K(q_i) + MC_i^{coord}(q_i, a_i) + MC_i^{tok}(q_i, a_i) \\ &\quad + MC_i^{rev}(q_i, a_i) + MC_i^{gov}(q_i, a_i) + MC_i^{oth}(q_i, a_i). \end{aligned}$$

这里的上标和前面的总成本分解一一对应。固定采用成本 $F_i(a_i)$ 不直接进入 $\partial C_i / \partial q_i$ ，但它会影响企业是否值得采用以及采用强度多高。下面逐项把边际成本和前面的总成本项接起来。

第一，人力边际成本来自 $C_i^L(q_i, a_i)$ 对 q_i 求导。在这个设定中，给定 a_i 时岗位结构和任务分布先视为给定，因此

$$MC_i^L(a_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} w_o n_o(x, u_i(x)) h_o(x) dx.$$

这就是“单位业务的岗位-任务加权劳动成本”。它可能随 a_i 下降，也可能随 a_i 上升，取决于替代项、互补劳动项和复杂性劳动项谁占主导。

第二，传统资本边际成本来自

$$MC_i^K(q_i) = \frac{\partial C_i^K(q_i)}{\partial q_i} = \varrho_i K'_i(q_i).$$

如果 $K_i(q_i)$ 近似线性，那么这一项接近常数；如果仓储、服务器、门店或设备容量在高产量区间变紧， $K'_i(q_i)$ 可能上升，资本边际成本也会随规模上升。AI/agent 对这一项的影响通常不表现为直接减少资本，而是通过协调和流程效率改变资本与劳动的配合。

第三，协调边际成本最直接。由于

$$C_i^{coord}(q_i, a_i) = q_i c_i^{coord}(a_i),$$

所以

$$MC_i^{coord}(a_i) = c_i^{coord}(a_i) = \kappa_i^0 - \kappa_i^1 a_i + \frac{\kappa_i^2}{2} a_i^2.$$

这说明协调成本的边际项继承了“先下降、后可能上升”的形状：低到中等 AI 使用降低组织摩擦，高强度使用则可能带来系统依赖和流程复杂性。

第四，token 边际成本需要用链式法则。为简化记号，先把单位业务的 token 强度记为

$$\bar{\tau}_i(a_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \tau_o(x) u_i(x) h_o(x) dx, \quad X_i(q_i, a_i) = q_i \bar{\tau}_i(a_i).$$

上一小节定义的是 token 使用量 X 的边际价格函数 $MC_i^{tok}(X)$ 。因此，企业多生产一单位业务带来的 token 边际成本为

$$\frac{\partial C_i^{tok}(q_i, a_i)}{\partial q_i} = MC_i^{tok}(X_i(q_i, a_i)) \cdot \frac{\partial X_i(q_i, a_i)}{\partial q_i} = MC_i^{tok}(X_i(q_i, a_i)) \bar{\tau}_i(a_i).$$

经济含义是：token 成本不仅取决于“每个 token 多贵”，还取决于每单位业务需要多少 token。高强度 AI 使用会提高 $\bar{\tau}_i(a_i)$ ，而大规模业务还可能把企业推入 $X_i > \bar{X}_i$ 的高边际价格区间。

第五，审核边际成本可以直接从 C_i^{rev} 得到：

$$MC_i^{rev}(q_i, a_i) = \frac{\partial C_i^{rev}(q_i, a_i)}{\partial q_i} = \psi_i^{rev}(a_i) + \nu_i a_i^2 q_i.$$

第一项是常规逐单位审核，第二项是错误扩散和复核压力的规模效应。这个式子很重要：即使常规审核成本只是随 a_i 温和上升，规模化使用也会让审核边际成本随 q_i 上升。

第六，监管和伦理成本区分单位合规成本与固定合规成本。若 $F_i^{gov}(a_i; \rho_i)$ 不随当前产量变化，则

$$MC_i^{gov}(a_i; \rho_i) = g_i(a_i; \rho_i).$$

固定合规成本不会进入当期边际成本，但会影响采用门槛和最优采用强度。高监管行业的特殊性在于：即使 g_i 只是单位成本， ρ_i 较高也会抬高每一单位业务的合规负担。

最后，其他成本项保留为

$$MC_i^{oth}(q_i, a_i) = \frac{\partial C_i^{oth}(q_i, a_i)}{\partial q_i}.$$

保留这一残差项，是为了提醒模型仍然遗漏了一些现实机制。网络安全、供应商锁定、延迟、碳排放、品牌声誉和内部组织摩擦未必都有同一个符号，因此在理论模型里不强行给出单一方向。

为了和成本压缩模型连接，可以在某个参考产量 $\bar{q}_i$ 附近对完整边际成本做一阶近似：

$$\begin{aligned} MC_i(q_i, a_i) &\approx MC_i(\bar{q}_i, a_i) + \left. \frac{\partial MC_i(q, a_i)}{\partial q} \right|_{q=\bar{q}_i} (q_i - \bar{q}_i) \\ &= m_i(a_i) + d_i(a_i)q_i, \end{aligned}$$

其中

$$d_i(a_i) = \left. \frac{\partial MC_i(q, a_i)}{\partial q} \right|_{q=\bar{q}_i}, \quad m_i(a_i) = MC_i(\bar{q}_i, a_i) - d_i(a_i)\bar{q}_i.$$

因此， $m_i(a_i)$ 和 $d_i(a_i)$ 不是新的拍脑袋假设，它们是完整边际成本函数在局部产量区间内的压缩表示。 $m_i(a_i)$ 记录边际成本水平， $d_i(a_i)$ 记录边际成本随规模变化的斜率。人力效率和协调改善可能让 $m'_i(a_i) < 0$ ，但 token、审核、监管和组织复杂性也可能让 $m'_i(a_i) > 0$ ，所以它的符号不能事先规定。 $d_i(a_i)$ 通常刻画规模压力：超容量 token 成本、错误扩散、复核拥挤和组织拥挤越强， $d_i(a_i)$ 越大。

对近似边际成本关于 q_i 积分，并把与当前产量无关的常数并入固定成本，就得到竞争分析常用的局部成本表示：

$$C_i(q_i, a_i) \approx F_i(a_i) + m_i(a_i)q_i + \frac{d_i(a_i)}{2}q_i^2.$$

成本压缩时会先保留这个局部近似。竞争分析从 $d_i(a_i) = 0$ 的常数边际成本版本推起，因为它给出最清楚的基准；随后再讨论 $d_i(a_i) > 0$ 时，规模拥挤和审核压力如何改变产量、价格 and 市场份额。

上面的推导可以写成

AI/agent 对企业成本的影响由岗位-技能暴露度、企业采用能力、token 供给、错误审核、监管约束和业务规模共同决定。后面出现的 $m_i(a_i)$ 和 $d_i(a_i)$ ，记录的就是这一成本重组的局部形状。

接下来可以把任务积分投影为 routine labor 与 skilled labor 两类劳动，从人力资本结构这一条线看成本变化；边际成本账本则继续作为成本压缩和竞争分析的基础。

4 人力资本结构：从任务劳动到组织重组

4.1 为什么人力成本要单独讨论

企业成本账本中，人力成本 $C_i^L(q_i, a_i)$ 有一层独立含义。它不只是工资支出，还对应岗位结构、技能结构和收入分配。token、审核、治理和固定采用成本属于另一类成本

来源：人力部分需要单独展开，是因为 AI/agent 对劳动的影响经常同时包含替代、互补和岗位重组。

企业采用 AI/agent 后，最直接的问题是单位产出需要多少劳动。但劳动不是同质的。客服初筛、标准文案、表格录入、简单代码补全和库存提醒这类任务，比较容易被 agent 替代或加速；模型监督、复杂投诉处理、合规审计、流程设计、系统集成和异常判断这类任务，则更可能与 AI 互补。若把所有劳动压成一个 L ，模型会太粗；若保留完整的岗位-任务积分，后面每一步推导都会被积分条件淹没。一个可读的折中，是把复杂任务空间投影为 routine labor 和 skilled labor 两类。

这个投影有代价。两类劳动不会覆盖现实中的所有岗位差异，也不能说明每个具体职业的变化。它的作用更有限：保留任务替代和技能互补两条主要机制，并让劳动结构能够进入后面的成本函数。

4.2 从任务积分到 routine/skilled 两类劳动

上一节从单个任务的劳动需求开始：

$$n_o(x, u) = \frac{\ell_o^0(x)}{1 + \phi_o(x)u} + \chi_o(x)u + \frac{\eta_o(x)}{2}u^2,$$

其中 $u_i(x) = \min\{1, a_i\theta_i e(x)\}$ 。这个式子已经把三种力量放在一起： $\phi_o(x)$ 刻画 AI 提高效率、减少单位任务所需劳动的能力， $\chi_o(x)$ 刻画提示词设计、流程改造、人工监督等互补劳动， $\eta_o(x)$ 刻画高强度使用后的返工、审核和复杂性劳动。

为了得到一个更容易分析的人力资本模型，把岗位 o 内部的任务集合粗略分为两部分： $\mathcal{X}_o^R$ 表示更常规、更可编码、更容易被 AI 加速的任务； $\mathcal{X}_o^S$ 表示更依赖判断、沟通、责任和系统整合的任务。于是企业 i 的两类劳动可以写成

$$L_{Ri}(q_i, a_i) = q_i \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o^R} n_o(x, u_i(x)) h_o(x) dx,$$

$$L_{Si}(q_i, a_i) = q_i \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o^S} n_o(x, u_i(x)) h_o(x) dx.$$

这里暂时写劳动数量，尚未乘上岗位工资；若乘上 w_o ，就回到上一节的人力成本 $C_i^L(q_i, a_i)$ 。数学上，这两个式子只是把完整积分按任务类型分组。经济学上，它说明 routine/skilled 并非从空中加进来的假设，而是对岗位-任务结构的降维。

完整的积分模型当然可以直接求导。比如在内点 $u_i(x) = a_i\theta_i e(x)$ 下， $\partial L_i / \partial a_i$ 会等于各岗位、各任务上

$$\frac{\partial n_o(x, u_i(x))}{\partial u_i(x)} \frac{\partial u_i(x)}{\partial a_i}$$

的加权积分。这个表达式更一般，但它的比较静态会变成一长串积分条件，不适合在正文中反复带入成本压缩和 Cournot 模型。因此，正文保留它的经济来源，同时把它投影成两类单位劳动需求。把单位产出所需的两类劳动记为 $\ell_{Ri}(a_i)$ 和 $\ell_{Si}(a_i)$ ，于是

$$L_{Ri}(q_i, a_i) = q_i \ell_{Ri}(a_i), \quad L_{Si}(q_i, a_i) = q_i \ell_{Si}(a_i).$$

这里的近似可以理解为：先在 $a_i = 0$ 附近看 routine 任务和 skilled 任务的平均斜率，再用少数几个参数概括这些平均斜率和曲率。一个便于推导的两类劳动表示是

$$\begin{aligned}\ell_{Ri}(a_i) &= \ell_{Ri}^0(1 - \alpha_{Ri}a_i), & 0 \leq \alpha_{Ri}a_i \leq 1, \\ \ell_{Si}(a_i) &= \ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2\right), & \sigma_i, \eta_i \geq 0.\end{aligned}$$

更准确地说，第一式捕捉 routine 任务的直接节约， α_{Ri} 越大，AI 对常规任务的替代或加速越强。这里不用 ρ_i ，是因为上一节已经把 ρ_i 留给监管强度。第二式捕捉 skilled 任务的互补和复杂性， σ_i 表示线性的监督、整合和流程设计劳动， η_i 表示高强度采用后更快上升的复杂性劳动。这里的 η_i 与上一节任务层面的 $\eta_o(x)$ 含义一致，只是已经被聚合到两类劳动模型中。

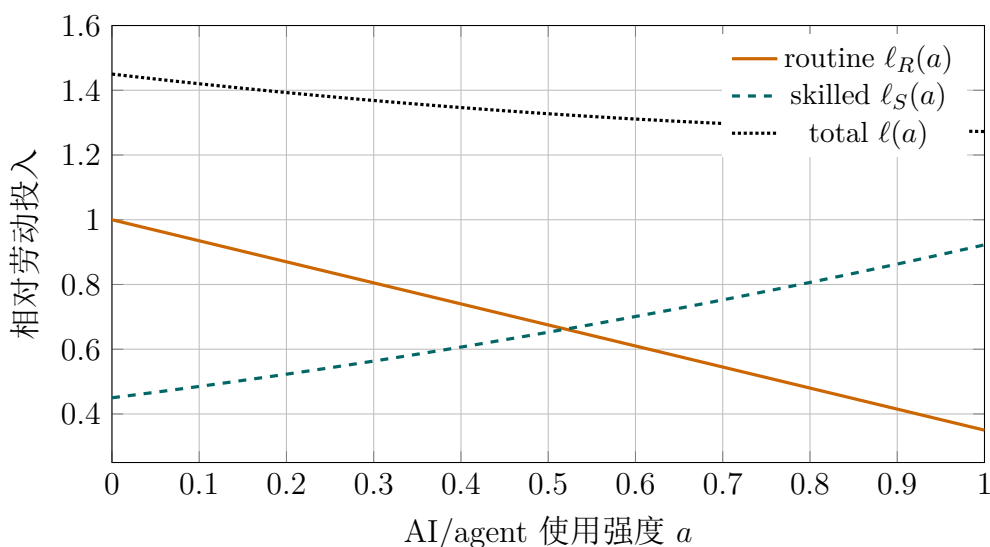


图 1: AI/agent 采用对单位劳动结构的影响。常规劳动随自动化下降，高技能监督、整合和模型治理劳动上升；总单位劳动需求取决于两类劳动变化的合力。

图 1 使用的就是这种低维表示。图中的曲线不预测真实就业数量，只说明一种局部机制：随着 a_i 上升，常规劳动的单位需求下降，高技能监督和整合劳动的单位需求上升；两者相加后的总单位劳动需求则要看哪一股力量更强。

4.3 给定产量下的替代效应与互补效应

先固定产量 q_i ，只看单位产出需要多少劳动。令

$$\ell_i(a_i) = \ell_{Ri}(a_i) + \ell_{Si}(a_i)$$

表示单位产出的总劳动需求。由上面的两类劳动近似可得

$$\ell'_i(a_i) = -\alpha_{Ri}\ell_{Ri}^0 + \ell_{Si}^0(\sigma_i + \eta_i a_i).$$

第一项为负，表示 routine 任务的节约；第二项为正，表示 skilled 任务的互补劳动和复杂性劳动。方向不由模型预先规定，而由两边力量谁更大决定。

命题 1（给定产量下的单位劳动需求）

在上述两类劳动近似下，若

$$\alpha_{Ri}\ell_{Ri}^0 > \ell_{Si}^0(\sigma_i + \eta_i a_i),$$

则给定产量时 AI/agent 使用强度上升会降低单位劳动需求，即 $\ell'_i(a_i) < 0$ 。若反向不等式成立，则单位劳动需求上升。

证明

由

$$\ell'_i(a_i) = -\alpha_{Ri}\ell_{Ri}^0 + \ell_{Si}^0(\sigma_i + \eta_i a_i)$$

直接可知，当 routine 节约项 $\alpha_{Ri}\ell_{Ri}^0$ 大于 skilled 互补和复杂性项 $\ell_{Si}^0(\sigma_i + \eta_i a_i)$ 时，导数为负；反之导数为正。□

含义与解释。 这个命题比较的是给定产量下的单位劳动需求，不是企业总用工。图 1 中，routine labor 曲线下降，skilled labor 曲线上升，两者相加后的方向取决于下降曲线的斜率是否大于上升曲线的斜率。因此，AI/agent 对劳动的影响不能简单概括为“替代劳动”。若企业主要把 agent 用在标准客服、文案和重复数据处理上， α_{Ri} 较高，单位劳动需求下降；若企业把 agent 接入复杂决策、合规流程和客户责任场景，监督和复杂性劳动会上升， σ_i 与 η_i 可能抵消甚至超过替代效应。D. H. Autor (2015) 和 Acemoglu and Restrepo (2019) 的任务框架说明，技术改变的是任务组合，而不是机械地消灭全部劳动。Dell’Acqua et al. (2023) 的 jagged frontier 也提醒我们，AI 在一些任务上很强，在另一些任务上会带来额外监督成本。

不同企业会落在不同参数区域。大型电商平台的客服、广告素材和商家运营高度标准化， α_{Ri} 往往较高；知识密集型专业服务企业的任务更依赖判断和责任分配， σ_i 与 η_i 可能更高；高监管行业还会遇到额外审核和合规成本，但那部分成本属于下一章的非人力 agent 成本。这里先把劳动侧讲清楚，避免把所有成本都塞进人力资本结构。

4.4 企业扩张后的总劳动需求

上一小节固定了 q_i 。但企业采用 AI/agent 的目的通常不是只在原有规模上少用一点劳动。若 AI 降低有效边际成本，企业可能在之后的 Cournot 竞争中扩大产量。于是总劳动需求要同时看单位劳动需求和规模变化。

令企业总劳动需求为

$$L_i(q_i, a_i) = q_i \ell_i(a_i).$$

如果均衡产量也随 a_i 变化，即 $q_i = q_i(a_i)$ ，那么

$$\frac{dL_i}{da_i} = \left. \frac{\partial L_i}{\partial a_i} \right|_{q_i} + \frac{\partial L_i}{\partial q_i} \frac{dq_i}{da_i} = q_i(a_i) \ell'_i(a_i) + \ell_i(a_i) q'_i(a_i).$$

第一项是给定产量时的直接劳动效应，第二项是企业扩张或收缩带来的规模效应。这个分解把劳动结构和成本压缩连接起来。

把上一小节的近似式代入，可以看得更清楚：

$$\frac{dL_i}{da_i} = q_i(a_i) [-\alpha_{Ri}\ell_{Ri}^0 + \ell_{Si}^0(\sigma_i + \eta_i a_i)] + q'_i(a_i) \left[\ell_{Ri}^0(1 - \alpha_{Ri}a_i) + \ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2 \right) \right].$$

第一组中括号就是上一命题中的单位劳动效应；第二组中括号是每单位产出所需的总劳动。这个式子说明，AI 的劳动影响有两个方向：它一方面可能让每一单位业务少用劳动，另一方面又可能通过降本扩大业务规模，让企业处理更多订单、客户和交易。

命题 2（规模效应可能抵消单位劳动节约）

若 $\ell'_i(a_i) < 0$ ，即单位劳动需求下降，则总劳动需求 $L_i(q_i(a_i), a_i)$ 上升的充要条件是

$$\frac{q'_i(a_i)}{q_i(a_i)} > -\frac{\ell'_i(a_i)}{\ell_i(a_i)}.$$

证明

由

$$\frac{dL_i}{da_i} = q_i(a_i)\ell'_i(a_i) + \ell_i(a_i)q'_i(a_i)$$

可知， $\frac{dL_i}{da_i} > 0$ 等价于

$$\ell_i(a_i)q'_i(a_i) > -q_i(a_i)\ell'_i(a_i).$$

在 $q_i(a_i) > 0$ 、 $\ell_i(a_i) > 0$ 下，两边同除以 $q_i(a_i)\ell_i(a_i)$ ，得到命题中的条件。 □

含义与解释。 现实中“AI 减少一些岗位”和“采用 AI 的企业继续扩张招聘”并不矛盾：前者对应单位劳动需求下降，后者对应规模效应上升。命题把这一直觉写成了两个增长率的比较——只要产量增长率 $q'_i(a_i)/q_i(a_i)$ 超过单位劳动节约率 $-\ell'_i(a_i)/\ell_i(a_i)$ ，总劳动需求就会上升。图 2 显示，若只看单位劳动曲线，AI/agent 可能让每单位产出需要的劳动下降；但若企业产量明显上升，总用工曲线可能反而向上。两个图必须一起读，不能只凭劳动结构图判断就业方向。对于大平台，如果 AI 降低客服、运营和推荐成本，并使其服务更多交易，企业总用工未必下降，但岗位构成会改变。Alekseeva et al. (2021) 关于 AI 技能需求上升的证据，与这种岗位重组解释相符。

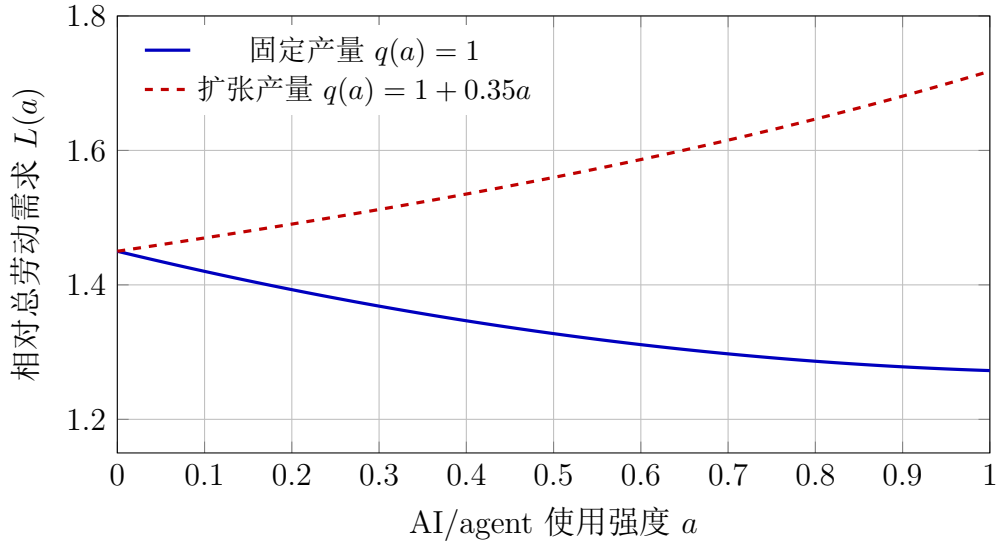


图 2: 单位劳动节约与规模效应。即使单位劳动需求下降, 若 AI/agent 降低成本并推动企业产量扩张, 总劳动需求也可能上升。

4.5 人力资本强度与工资结构

接下来把劳动数量转成人力资本结构。设 $\omega_R > 0$ 是 routine labor 的人力资本权重, $\omega_S > 0$ 是 skilled labor 的人力资本权重, 并且 $\omega_S > \omega_R$ 。企业剩余劳动的人力资本强度定义为

$$H_i(a_i) = \frac{\omega_S L_{Si}(q_i, a_i)}{\omega_S L_{Si}(q_i, a_i) + \omega_R L_{Ri}(q_i, a_i)}.$$

由于 L_{Ri} 和 L_{Si} 都同乘 q_i , 在这个定义中产量规模会约掉。也就是说, $H_i(a_i)$ 衡量的是劳动结构, 不是企业大小。

代入两类劳动近似后, $H_i(a_i)$ 可以写得更具体:

$$H_i(a_i) = \frac{\omega_S \ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2\right)}{\omega_S \ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2\right) + \omega_R \ell_{Ri}^0 (1 - \alpha_{Ri} a_i)}.$$

这个式子的分子随 skilled 任务的监督和复杂性劳动增加而上升, 分母还包含正在下降的 routine 劳动。若两股力量的方向成立, 人力资本强度自然会上升。

命题 3 (AI/agent 采用提高剩余劳动的人力资本强度的局部条件)

若在某一区间内 $L'_{Ri}(a_i) < 0$ 、 $L'_{Si}(a_i) > 0$, 且 $\omega_S > \omega_R > 0$, 则 $H'_i(a_i) > 0$ 。

证明
记

$$\mathcal{S}_i(a_i) = \ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2\right), \quad \mathcal{R}_i(a_i) = \ell_{Ri}^0 (1 - \alpha_{Ri} a_i).$$

于是

$$H_i(a_i) = \frac{\omega_S \mathcal{S}_i(a_i)}{\omega_S \mathcal{S}_i(a_i) + \omega_R \mathcal{R}_i(a_i)}.$$

对 a_i 求导：

$$H'_i(a_i) = \frac{\omega_S \omega_R [S'_i(a_i) \mathcal{R}_i(a_i) - S_i(a_i) \mathcal{R}'_i(a_i)]}{[\omega_S S_i(a_i) + \omega_R \mathcal{R}_i(a_i)]^2}.$$

这里

$$S'_i(a_i) = \ell_{Si}^0(\sigma_i + \eta_i a_i) > 0, \quad \mathcal{R}'_i(a_i) = -\alpha_{Ri} \ell_{Ri}^0 < 0.$$

只要 $S_i(a_i) > 0$ 、 $\mathcal{R}_i(a_i) > 0$ ，分母为正，分子中的 $S'_i(a_i) \mathcal{R}_i(a_i)$ 为正， $-S_i(a_i) \mathcal{R}'_i(a_i)$ 也为正，因此 $H'_i(a_i) > 0$ 。□

含义与解释。 这个命题成立需要两个局部条件：routine labor 下降，skilled labor 上升。若某个行业中 skilled 监督劳动没有上升，或 routine 任务并未被有效替代，结论就不一定成立。近似式把这个条件写成了可观察的参数方向—— $\alpha_{Ri} > 0$ 推低 routine 劳动， $\sigma_i, \eta_i \geq 0$ 推高 skilled 劳动。图 3 中，skilled labor 的权重高于 routine labor，随着 a_i 上升，劳动结构中 skilled labor 占比提高， $H_i(a_i)$ 也随之上升；复杂性劳动越强，这一上升越快。

经济含义。 AI/agent 可能提高剩余岗位的平均技能含量，但这不等于所有劳动者都受益。低复杂度岗位可能减少，高技能监督、系统整合和风险治理岗位可能增加。工资结构也可能因此拉开：若 ω_S 代表更高工资或更高人力资本， $H_i(a_i)$ 上升同时意味着企业内部劳动成本的构成发生变化。

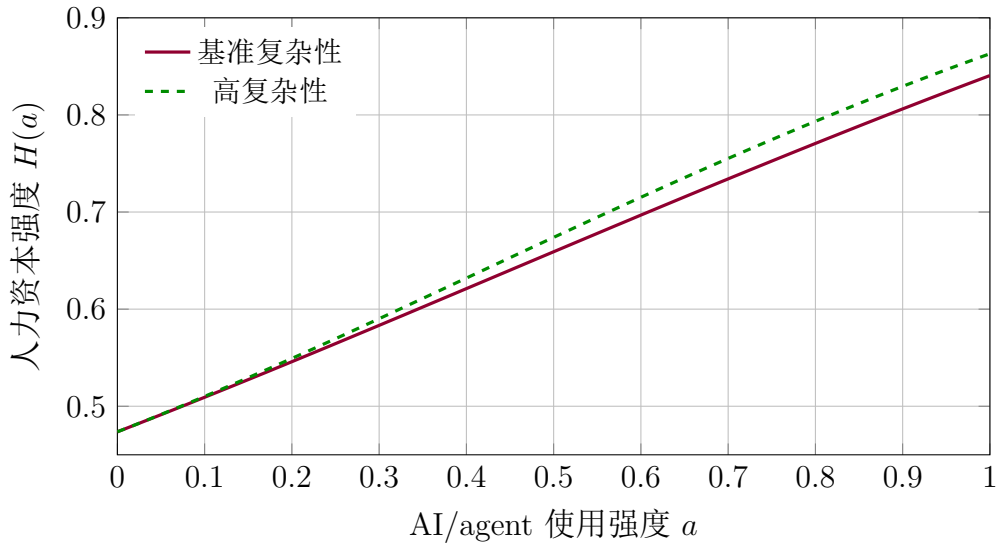


图 3: 人力资本强度 $H(a)$ 。当 skilled labor 的权重高于 routine labor 时，AI/agent 普及会提高剩余劳动的平均人力资本含量；复杂性劳动越强，上升越快。

4.6 人力成本如何进入成本参数

人力成本的变化来自任务层面的 $n_o(x, u_i(x))$ 。把任务空间投影为 routine labor 和 skilled labor 后，替代效应、互补效应和复杂性效应可以同时写进 $\ell_{Ri}(a_i)$ 与 $\ell_{Si}(a_i)$ 。给

定产量时，单位劳动需求可能下降；若企业因降本而扩大产量，总劳动需求也可能回升。因此，就业方向不能只靠“AI 替代劳动”这一句话判断。

这些劳动侧结果会落到不同参数上。routine 任务节约、人力效率提高和部分协调改善，会提高有效节约能力 β_i 。模型监督、异常处理、返工和复杂性劳动，会推高使用强度上升后的成本压力，进入 λ_i 或 $d_i(a_i)$ 。培训、流程改造和组织转型中不随单个订单变化的部分，则更接近固定采用成本 f_i 。token 调用、错误赔付、外部审计、治理合规和非工资协调摩擦另有来源，不能在这里重复计入。

5 非劳动 Agent 成本：统一框架、token 与规模风险

5.1 非劳动成本的归属规则

企业内部员工的工时、工资和人力资本结构已经由 $C_i^L(q_i, a_i)$ 捕捉。剩下的问题是，工资成本之外的 agent 使用还会带来哪些支出。这里的边界很重要：人工审核员的工资属于劳动成本；错误赔付、外部审计费用、责任保险和系统重跑成本属于非劳动成本。同一笔成本只能进入一个账本。

定义 3（非劳动成本 C_i^N ）

把除内部员工工资以外的 agent 成本记为 C_i^N ，其中 N 表示 non-labor：

$$C_i^N(q_i, a_i, z_i) = F_i^N(a_i, z_i) + C_i^{tok}(q_i, a_i) + q_i v_i^N(a_i, z_i) + \frac{1}{2} q_i^2 \Xi_i^N(a_i, z_i) + C_i^{oth, N}(q_i, a_i).$$

这里 z_i 表示企业事前投入的治理和风险控制强度，例如日志系统、权限系统、模型审计、责任追踪和风险分级。这个式子按成本进入函数的方式分类，而不是按“审核”“治理”“协调”这些词分类。 F_i^N 是固定非劳动成本， C_i^{tok} 是 token 和模型调用成本， $q_i v_i^N$ 是随业务规模线性变化的非劳动运行成本， $\frac{1}{2} q_i^2 \Xi_i^N$ 是随规模凸增的非劳动压力， $C_i^{oth, N}$ 留给供应商锁定、网络安全、延迟、碳排放和品牌风险等未完全列出的成本。

这个归属规则有一个硬约束：凡是企业内部员工工资、培训工时、人工复核工时、合规人员工资、员工少开会带来的工资时间节约，都不进入 C_i^N 。这些成本已经由 C_i^L 处理。非劳动账本关注的是另一组支出：即使工资成本已经单独拿出去，企业采用 agent 仍要面对 token 调用、外部审核损失、治理系统、固定基础设施和规模风险。

这些非劳动成本先按进入成本函数的方式写清楚：哪些是固定项，哪些随业务规模线性增加，哪些会随着规模凸增。采用强度和竞争均衡要等成本压缩后再求。

5.2 Token 调用成本与容量约束

token 成本需要单独保留，因为它不像一般固定成本或一般治理成本。它按调用量、上下文长度、工具执行、模型路由和推理算力变化，还会出现容量阈值。

Section 3 已经把企业总 token 使用量写成

$$X_i(q_i, a_i) = q_i \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \tau_o(x) u_i(x) h_o(x) dx.$$

括号中的部分是单位业务 token 强度：

$$\bar{\tau}_i(a_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \tau_o(x) u_i(x) h_o(x) dx.$$

为了让这个式子真正进入推导，使用内点近似

$$u_i(x) = a_i \theta_i e(x).$$

此时可以定义企业的任务 token 强度指数

$$\mathcal{T}_i = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \tau_o(x) e(x) h_o(x) dx.$$

于是

$$\bar{\tau}_i(a_i) = a_i \theta_i \mathcal{T}_i, \quad X_i(q_i, a_i) = q_i a_i \theta_i \mathcal{T}_i.$$

这个等式把 token 使用量拆成四个来源：业务规模 q_i 、使用强度 a_i 、企业采用能力 θ_i 、任务本身的 token 强度 $\mathcal{T}_i$ 。如果企业已经进入 $u_i(x) = 1$ 的饱和区间，上式要改成带 $\min\{\cdot\}$ 的版本；正文先用内点近似，是为了看清比较静态。

token 的边际价格仍取三段式：

$$MC_i^{tok}(X) = \begin{cases} p_i^0 - \delta_i X, & 0 \leq X \leq \tilde{X}_i, \\ p_i^L, & \tilde{X}_i < X \leq \bar{X}_i, \\ p_i^L + \zeta_i(X - \bar{X}_i), & X > \bar{X}_i. \end{cases}$$

参数满足

$$0 < \tilde{X}_i < \bar{X}_i, \quad p_i^L = p_i^0 - \delta_i \tilde{X}_i, \quad p_i^L \geq 0, \quad \delta_i, \zeta_i \geq 0.$$

第一段表示学习、缓存、批量调用和提示词复用；第二段表示稳定调用；第三段表示容量、延迟、算力和供应约束开始推高边际价格。

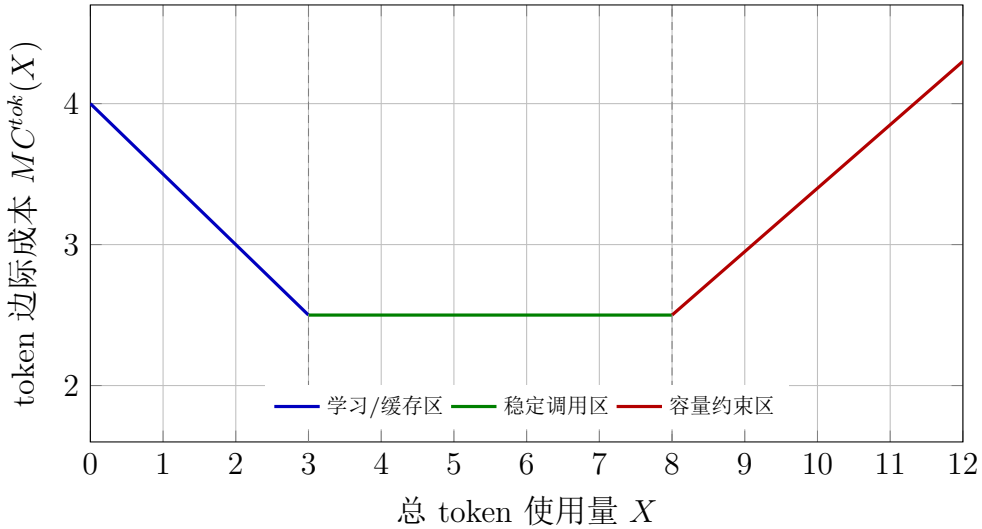


图 4: token 边际成本的三段式变化。参数为 $p_i^0 = 4, \delta_i = 0.5, \tilde{X}_i = 3, p_i^L = 2.5, \bar{X}_i = 8, \zeta_i = 0.45$ 。早期学习、缓存和批量调用使边际成本下降；稳定区间边际价格近似不变；超过容量后，延迟、算力和供应约束使边际成本上升。

把 $X_i = q_i a_i \theta_i \mathcal{T}_i$ 代入后，企业处在不同 token 成本区间的条件可以直接写成采用强度阈值：

$$a_i \leq \frac{\tilde{X}_i}{q_i \theta_i \mathcal{T}_i}, \quad a_i > \frac{\bar{X}_i}{q_i \theta_i \mathcal{T}_i}.$$

这比单纯说“用量大时成本上升”更具体。企业规模 q_i 越大、采用能力 θ_i 越高、任务越 token-intensive 即 $\mathcal{T}_i$ 越高，企业越早进入稳定调用区甚至容量约束区。大企业可能有更高的 $\bar{X}_i$ 和更强议价能力，但它们的 q_i 也大，不能简单说 token 成本一定更低。

总 token 成本是边际价格的积分：

$$C_i^{tok}(q_i, a_i) = \int_0^{X_i(q_i, a_i)} MC_i^{tok}(z) dz.$$

对产量求边际成本时，得到

$$MC_i^{tok,q}(q_i, a_i) = \frac{\partial C_i^{tok}}{\partial q_i} = MC_i^{tok}(X_i(q_i, a_i)) \bar{\tau}_i(a_i).$$

在内点近似下，这个边际成本可以写成

$$MC_i^{tok,q}(q_i, a_i) = \begin{cases} [p_i^0 - \delta_i q_i a_i \theta_i \mathcal{T}_i] a_i \theta_i \mathcal{T}_i, & X_i \leq \tilde{X}_i, \\ p_i^L a_i \theta_i \mathcal{T}_i, & \tilde{X}_i < X_i \leq \bar{X}_i, \\ [p_i^L + \zeta_i(q_i a_i \theta_i \mathcal{T}_i - \bar{X}_i)] a_i \theta_i \mathcal{T}_i, & X_i > \bar{X}_i. \end{cases}$$

在不位于 $\tilde{X}_i$ 或 $\bar{X}_i$ 这两个阈值点的分段区间内部，再对 q_i 求导，得到

$$\frac{\partial MC_i^{tok,q}}{\partial q_i} = [MC_i^{tok}]'(X_i) \bar{\tau}_i(a_i)^2.$$

第一段中 $[MC_i^{tok}]'(X_i) < 0$ ，token 边际成本随规模上升而下降，表现为局部规模经济。稳定区间中导数为零，token 成本不额外制造规模压力。容量约束区间中导数为正，token 成本会进入 $d_i(a_i)$ ：企业越大，深度使用 agent 的边际成本越容易随产量上升。

5.3 非劳动运行成本：审核、治理与协调

除 token 以外，还有一类成本随每单位业务发生，但不属于员工工资。把它们合并成

$$v_i^N(a_i, z_i) = \psi_i^{rev,N}(a_i, z_i) + g_i^{gov,N}(\Gamma_i(a_i), z_i; \rho_i) + c_i^{coord,N}(a_i) + v_i^{other,N}(a_i).$$

这四项都按非劳动口径理解。人工审核员工资、合规人员工资和员工协作时间节约已经放在 Section 4；这里不再计算。

先看非劳动审核和复核成本。 $\psi_i^{rev,N}(a_i, z_i)$ 包含重新调用模型、外部抽检、第三方审计、投诉处理中的非工资支出、技术复核和低额补偿等成本。若从任务层写，可以设

$$b_o^{rev,N}(x, u) = b_o^{rev,N,0}(x) + b_o^{rev,N,1}(x)u + \frac{b_o^{rev,N,2}(x)}{2}u^2,$$

并令

$$\psi_i^{rev,N}(a_i, z_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} c_o^{rev,N}(z_i) b_o^{rev,N}(x, u_i(x)) h_o(x) dx.$$

这里 $b_o^{rev,N}$ 是任务层面的非劳动复核负担, $c_o^{rev,N}(z_i)$ 是非劳动复核或风险处理单价。治理投入 z_i 可能降低某些复核成本, 也可能提高某些审计标准, 因此不强行规定它的符号方向。

治理的单位运行成本来自任务监管敏感度。设 $\gamma_o(x) \geq 0$ 表示任务 x 对隐私、解释性、审计、数据权限和责任追踪的敏感程度, 定义治理暴露度

$$\Gamma_i(a_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \gamma_o(x) u_i(x) h_o(x) dx.$$

内点下,

$$\Gamma_i(a_i) = a_i \theta_i \mathcal{G}_i, \quad \mathcal{G}_i = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \gamma_o(x) e(x) h_o(x) dx.$$

所以治理压力来自

$$a_i \times \theta_i \times \mathcal{G}_i \times \rho_i.$$

a_i 是使用强度, θ_i 是采用能力, $\mathcal{G}_i$ 是企业任务结构中的监管敏感度, ρ_i 是行业监管强度。一个可读的特例是

$$g_i^{gov,N}(\Gamma_i, z_i; \rho_i) = \rho_i \left[g_0(z_i) + g_1 \Gamma_i + \frac{g_2}{2} \Gamma_i^2 \right], \quad g_1, g_2 \geq 0.$$

这说明, 同样的 agent 使用强度, 在普通营销任务和医疗、金融、法律任务中会产生完全不同的治理成本。

协调成本也要用非劳动口径。员工少开会、少找文档、少交接带来的工资时间节约属于 C_i^L 或劳动侧的 β_i^L 。这里的协调成本只记录系统延迟、跨工具依赖、自动任务失败、服务等级损失、重复调用和流程阻塞等非工资摩擦。设 $\mu_o(x)$ 表示任务的交接、搜索、文档和跨系统依赖强度, 定义

$$\mathcal{M}_i = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \mu_o(x) e(x) h_o(x) dx.$$

于是单位非劳动协调成本可以写成

$$c_i^{coord,N}(a_i) = \kappa_i^{N,0} - \kappa_i^{N,1}(a_i \theta_i \mathcal{M}_i) + \frac{\kappa_i^{N,2}}{2} (a_i \theta_i \mathcal{M}_i)^2, \quad \kappa_i^{N,1}, \kappa_i^{N,2} \geq 0.$$

低到中等强度的 agent 使用减少搜索、交接和跨系统摩擦; 强度过高后, 权限管理、工具依赖、输出不一致和流程重构会把协调成本推回来。若 $\kappa_i^{N,2} > 0$, 最小化该项得到

$$a_i^{coord} = \frac{\kappa_i^{N,1}}{\kappa_i^{N,2} \theta_i \mathcal{M}_i},$$

只要它落在 $[0, 1]$ 内。经济含义是: 组织协作复杂度越高、采用能力越强, 企业越早达到协调成本的低点; 继续提高 a_i 未必更省钱。

5.4 固定采用与治理投资

固定非劳动成本写成

$$F_i^N(a_i, z_i) = F_i^{adopt, N}(\mathcal{D}_i(a_i)) + F_i^{gov, N}(\Gamma_i(a_i), z_i; \rho_i).$$

第一项是一般采用基础设施，第二项是事前治理体系。二者都不包含员工培训工资或内部合规人员工资。若员工花时间培训、学习和监督，这些工时应回到 Section 4。

为了把一般采用成本从任务结构中推出，设 $\iota_o(x)$ 表示任务 x 接入 agent 的系统难度，包括数据清洗、知识库整理、API 网关、模型路由、向量数据库、agentOps、业务系统改造和供应商迁移等非劳动支出。定义

$$\mathcal{D}_i(a_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \iota_o(x) u_i(x) h_o(x) dx.$$

内点下，

$$\mathcal{D}_i(a_i) = a_i \theta_i \mathcal{D}_i^0, \quad \mathcal{D}_i^0 = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \iota_o(x) e(x) h_o(x) dx.$$

一个简单的固定采用成本表示是

$$F_i^{adopt, N}(\mathcal{D}_i(a_i)) = F_i^0 \mathbf{1}\{a_i > 0\} + \phi_i^1 \mathcal{D}_i(a_i) + \frac{\phi_i^2}{2} \mathcal{D}_i(a_i)^2, \quad \phi_i^1, \phi_i^2 \geq 0.$$

$F_i^0 \mathbf{1}\{a_i > 0\}$ 表示只要开始接入 agent，就要付出 API 网关、模型服务接入或基础设施启动成本；线性项表示接入任务范围扩大；二次项表示深度集成后，旧系统、权限、数据格式和供应商依赖使改造成本加速上升。

固定治理体系 $F_i^{gov, N}$ 包含日志系统、权限隔离、隐私合规工具、模型解释性报告、外部认证、监管报送系统和责任链条。它与 $F_i^{adopt, N}$ 的分界在于：若支出来自监管、审计、隐私或责任追踪要求，就归入 $F_i^{gov, N}$ ；若只是为了让 agent 能接入业务系统，就归入 $F_i^{adopt, N}$ 。

固定成本不进入当期边际成本：

$$\frac{\partial F_i^N(a_i, z_i)}{\partial q_i} = 0.$$

但它进入平均成本：

$$\frac{F_i^N(a_i, z_i)}{q_i}.$$

这也是采用门槛和市场集中会出现的原因之一。规模大的企业可以把固定成本摊到更多业务上；小企业即使任务暴露度高，也可能因为 F_i^N/q_i 太高而停留在低强度采用。这个结论也不能反过来说“大企业一定占优”：传统大企业的旧系统复杂， $\mathcal{D}_i^0$ 和 ϕ_i^2 可能很高；AI-native 企业虽然规模未必最大，但从一开始围绕 agent 设计流程，固定改造成本可能低得多。

5.5 规模凸性、错误扩散与风险控制

最后一类非劳动成本随规模凸增。它写成

$$C_i^{scale,N}(q_i, a_i, z_i) = \frac{1}{2} q_i^2 \Xi_i^N(a_i, z_i),$$

其中

$$\Xi_i^N(a_i, z_i) = \nu_i(z_i) \Omega_i(a_i, z_i) + \Xi_i^{tok,loc}(a_i) + \Xi_i^{gov}(\Gamma_i(a_i); \rho_i) + \Xi_i^{coord}(a_i).$$

Ξ_i^N 是规模压力，刻意不用字母 s 表示，是为了避免和后文市场份额 s_i 冲突。

错误扩散是第一项。设 $\omega_o(x)$ 表示任务 x 出现系统性错误、模板化误判或自动流程连锁传播的风险敏感度。先定义不考虑治理投入时的基准系统性错误暴露：

$$\Omega_i^0(a_i) = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \omega_o(x) u_i(x)^2 h_o(x) dx.$$

内点下，

$$\Omega_i^0(a_i) = a_i^2 \theta_i^2 \mathcal{R}_i, \quad \mathcal{R}_i = \sum_{o \in \mathcal{O}_i} \pi_{io} \int_{\mathcal{X}_o} \omega_o(x) e(x)^2 h_o(x) dx.$$

平方项的含义是，高强度使用并不只是“多一点错误”，而是可能让同一类错误在自动化流程中被重复放大。普通独立小错误可以放进线性运行成本；系统性错误扩散才进入 Ξ_i^N 。

治理投入 z_i 的作用不是单纯增加成本。它一方面提高 $F_i^{gov,N}$ 和 $g_i^{gov,N}$ ，另一方面降低错误扩散强度。可以写成

$$\nu_i = \nu_i(z_i), \quad \nu'_i(z_i) < 0,$$

或者写成

$$\Omega_i = \Omega_i(a_i, z_i), \quad \frac{\partial \Omega_i(a_i, z_i)}{\partial z_i} < 0.$$

这里 $\Omega_i^0(a_i)$ 是没有治理改善时的基准暴露， $\Omega_i(a_i, z_i)$ 是治理投入后的有效暴露。若暂时不建模治理如何改变暴露度，可以令 $\Omega_i(a_i, z_i) = \Omega_i^0(a_i)$ ，只让 z_i 通过 $\nu_i(z_i)$ 降低单位风险损失。

一个直观特例是

$$\nu_i(z_i) = \nu_i^0 e^{-\xi_i z_i}, \quad \xi_i > 0.$$

此时错误扩散成本为

$$\frac{1}{2} \nu_i^0 e^{-\xi_i z_i} \Omega_i^0(a_i) q_i^2.$$

企业面对的是治理成本和风险损失之间的权衡。若只写最简单的治理选择问题，可以把它表示为

$$\min_{z_i} F_i^{gov,N}(\Gamma_i(a_i), z_i; \rho_i) + q_i g_i^{gov,N}(\Gamma_i(a_i), z_i; \rho_i) + \frac{1}{2} q_i^2 \nu_i(z_i) \Omega_i(a_i, z_i).$$

这里不求这个问题的一阶条件，只保留经济含义：治理像保险、防火墙或审计系统，付出事前成本，换取更低的错误扩散和责任损失。规模大的企业因为 q_i^2 项更重，往往更有动力治理；但它们的数据、用户和监管关注度也更高，治理成本本身也可能更高。

另外三项 $\Xi_i^{tok,loc}$ 、 Ξ_i^{gov} 和 Ξ_i^{coord} 分别来自 token 的局部曲率、合规拥挤和协调拥挤。这里要避免重复计算：如果已经保留完整的分段 token 成本，完整成本账本就取 $\Xi_i^{tok,loc}(a_i) = 0$ 。等成本压缩成 $m_i(a_i) + d_i(a_i)q_i$ 时，才可以用 $\Xi_i^{tok,loc}$ 表示超容量 token 成本的局部二次近似。同理，线性单位治理成本已经在 v_i^N 中，只有随规模凸增的监管报送、隐私请求、审计抽样和责任追踪压力才进入 Ξ_i^{gov} 。

对规模凸性成本求边际成本可得

$$MC_i^{scale,N}(q_i, a_i, z_i) = q_i \Xi_i^N(a_i, z_i).$$

因此

$$\frac{\partial MC_i^{scale,N}}{\partial q_i} = \Xi_i^N(a_i, z_i).$$

这也是 $d_i(a_i)$ 的非劳动来源：错误扩散、超容量 token、合规拥挤和协调拥挤越强，企业边际成本越会随产量上升。

5.6 非劳动成本进入 $f_i, m_i(a_i), d_i(a_i)$

非劳动成本按进入企业成本函数的方式整理后，可以压缩为

$$C_i^N(q_i, a_i, z_i) = F_i^N(a_i, z_i) + C_i^{tok}(q_i, a_i) + q_i v_i^N(a_i, z_i) + \frac{1}{2} q_i^2 \Xi_i^N(a_i, z_i) + C_i^{oth,N}(q_i, a_i).$$

这个式子和人力成本 C_i^L 合在一起，构成成本压缩的基础。

映射关系可以写成

$$\begin{aligned} F_i^N(a_i, z_i) &\Rightarrow f_i, \\ v_i^N(a_i, z_i) &\text{与稳定 token 成本} \Rightarrow m_i(a_i) \text{ 和 } \lambda_i, \\ \Xi_i^N(a_i, z_i) &\text{与 token 局部曲率} \Rightarrow d_i(a_i), \\ C_i^L(q_i, a_i) &\Rightarrow \beta_i \text{ 和 } \lambda_i^L. \end{aligned}$$

最后一行把人力成本单独列出，是为了防止重复计算。员工培训工资、人工审核工资、合规人员工资和员工协作时间节约都不在 C_i^N 中重复计入。劳动侧的效率节约和复杂性劳动进入 β_i 与 λ_i^L ；token、非劳动运行成本、固定采用、治理投资和规模风险则从 C_i^N 进入压缩参数。

由此得到的成本图景是双向的：人力效率、协调改善和流程自动化可能降低成本；token、固定采用、治理、错误扩散和规模拥挤会提高成本。压缩成 f_i 、 $m_i(a_i)$ 和 $d_i(a_i)$ 以后，这些来源才能进入市场竞争和市场结构分析。

6 成本压缩与采用选择

完整成本账本解释了 AI/agent 为什么会同时带来节约和新增成本。若把岗位积分、分段 token 价格、审核风险、治理投入和规模凸性全部直接带入竞争模型，模型会变成

一长串难以读懂的条件。竞争分析需要一个更短的表达：参数要少，但每个参数都能追溯到真实成本来源。

成本压缩的作用就在于此。 f_i 、 $m_i(a_i)$ 、 $d_i(a_i)$ 、 β_i 和 λ_i 并非新的黑箱参数，它们是劳动成本、非劳动成本、固定采用成本和规模压力的局部表达。在 Cournot 基准和其他市场结构中，这些参数决定企业在市场中的成本位置。

6.1 为什么需要成本压缩

最清楚的 Cournot 基准通常使用常数边际成本。企业 i 的利润写成

$$\pi_i = P(Q)q_i - C_i(q_i).$$

如果边际成本是 m_i ，反需求是 $P(Q) = A - BQ$ ，企业的一阶条件就很容易解释：多生产一单位的边际收益等于边际成本。

AI/agent 进入企业后，成本不再只是一个常数。完整成本同时包含人力、传统资本、协调、token、审核、治理和其他成本；进一步拆分时，还要区分劳动成本和非劳动成本。为了保留可读性，完整成本可以在一个相关产量区间附近压缩为

$$C_i(q_i, a_i) \approx F_i^{eff}(a_i) + m_i(a_i)q_i + \frac{d_i(a_i)}{2}q_i^2.$$

其中 $F_i^{eff}(a_i)$ 是有效固定采用成本， $m_i(a_i)$ 是给定规模附近的单位边际成本水平， $d_i(a_i)$ 记录边际成本随产量上升的速度。

数学上，这是 Section 3 中 $MC_i(q_i, a_i) \approx m_i(a_i) + d_i(a_i)q_i$ 的成本函数版本。经济学上， $m_i(a_i)$ 管“单位业务是否更便宜”， $d_i(a_i)$ 管“规模扩大后是否越来越难便宜”。之后的 Cournot 先从 $d_i(a_i) = 0$ 推起，是为了得到清楚基准；再把 $d_i(a_i) > 0$ 看成规模拥挤、错误扩散和 token 容量约束的修正。

6.2 从劳动数量到劳动成本参数

Section 4 讨论的是劳动数量。为了进入成本函数，需要把劳动数量乘上工资或影子工资。用 w_{Ri} 表示 routine labor 的单位工资，用 w_{Si} 表示 skilled labor 的单位工资。企业每单位产出的人力成本可写成

$$c_i^L(a_i) = w_{Ri}\ell_{Ri}(a_i) + w_{Si}\ell_{Si}(a_i).$$

这里的 $\ell_{Ri}(a_i)$ 和 $\ell_{Si}(a_i)$ 仍然是 Section 4 的单位劳动需求，不是新变量。

把 Section 4 的低维近似代入：

$$\ell_{Ri}(a_i) = \ell_{Ri}^0(1 - \alpha_{Ri}a_i), \quad \ell_{Si}(a_i) = \ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2\right).$$

于是

$$\begin{aligned} c_i^L(a_i) &= w_{Ri}\ell_{Ri}^0(1 - \alpha_{Ri}a_i) + w_{Si}\ell_{Si}^0 \left(1 + \sigma_i a_i + \frac{\eta_i}{2} a_i^2\right) \\ &= c_i^{L,0} - \beta_i^L a_i + \frac{\lambda_i^L}{2} a_i^2, \end{aligned}$$

其中

$$\begin{aligned} c_i^{L,0} &= w_{Ri} \ell_{Ri}^0 + w_{Si} \ell_{Si}^0, \\ \beta_i^L &= w_{Ri} \ell_{Ri}^0 \alpha_{Ri} - w_{Si} \ell_{Si}^0 \sigma_i, \quad \lambda_i^L = w_{Si} \ell_{Si}^0 \eta_i. \end{aligned}$$

这一步的含义很直接。Routine 任务节约带来成本下降，所以进入 β_i^L 的正项；skilled labor 的一阶互补需求会抵消一部分节约，所以进入 β_i^L 的负项；复杂性劳动随 a_i 凸增，所以进入 λ_i^L 。因此， β_i^L 不必然为正。只有 routine 节约的工资价值大于 skilled 互补劳动的一阶工资成本时，劳动侧才在低强度使用区间带来净节约。

图像上， $c_i^L(a_i)$ 可以先下降后上升。经济学上，“AI 节省人力成本”不能作为无条件判断；真正起作用的是 routine 替代、skilled 互补和复杂性劳动三者相减后的结果。

6.3 非劳动成本如何进入参数

Section 5 的非劳动成本写成

$$C_i^N(q_i, a_i, z_i) = F_i^N(a_i, z_i) + C_i^{tok}(q_i, a_i) + q_i v_i^N(a_i, z_i) + \frac{1}{2} q_i^2 \Xi_i^N(a_i, z_i) + C_i^{oth,N}.$$

这个式子已经按“进入成本函数的方式”分好了类。固定非劳动成本进入 F_i^N ，单位非劳动运行成本进入 v_i^N ，规模凸性进入 Ξ_i^N ，token 成本则因分段边际价格单独保留。

为了得到可推导的形式，可以在低到中等使用强度附近做局部近似：

$$\begin{aligned} F_i^N(a_i, z_i) &\approx F_i^{0,N} + \frac{f_i^N}{2} a_i^2, \\ \frac{C_i^{tok}(q_i, a_i)}{q_i} + v_i^N(a_i, z_i) &\approx c_i^{N,0} - \beta_i^N a_i + \frac{\lambda_i^N + \lambda_i^{tok}}{2} a_i^2, \\ \Xi_i^N(a_i, z_i) &\approx d_i^N(a_i). \end{aligned}$$

这里 $F_i^{0,N}$ 是一次性非劳动进入成本，例如初始采购、基础设施接入和安全审查； $f_i^N a_i^2/2$ 是采用深度越高越重的固定集成成本。 β_i^N 是净一阶非劳动节约，主要来自非工资协调改善、流程自动化和重复调用减少；如果稳定 token 价格带来线性调用成本，它会抵消这部分节约。 λ_i^N 表示非劳动运行成本随使用强度上升的曲率， λ_i^{tok} 表示 token 成本在局部压缩时带来的曲率。若 Section 5 保留完整分段 C_i^{tok} ，就不要在同一成本账本中再重复计入 λ_i^{tok} ；只有把 token 成本压缩进二次函数时，才使用这一项。

规模项 $d_i^N(a_i)$ 的来源包括错误扩散、超容量 token、合规拥挤和非工资协调拥挤。它不改变给定产量下第一单位业务的成本，却会让企业规模越大、深度采用越难保持低边际成本。

6.4 简化成本函数

定义 4 (压缩成本函数 C_i^{red})

把劳动侧和非劳动侧合在一起, 得到压缩成本:

$$C_i^{red}(q_i, a_i) = \left[c_i^0 - \beta_i a_i + \frac{\lambda_i}{2} a_i^2 \right] q_i + \frac{f_i}{2} a_i^2 + \frac{d_i(a_i)}{2} q_i^2.$$

其中

$$\beta_i = \beta_i^L + \beta_i^N,$$

$$\lambda_i = \lambda_i^L + \lambda_i^N + \lambda_i^{tok},$$

$$f_i = f_i^L + f_i^N.$$

f_i^L 来自培训、组织学习和内部流程改造中不随单个订单变化的劳动投入; f_i^N 来自软件、云服务、数据管线、安全系统和治理体系等非劳动固定投入。

这个式子是竞争模型的入口。它保留了三种经济力量: $-\beta_i a_i q_i$ 表示采用 AI/agent 后随业务规模放大的节约; $\lambda_i a_i^2 q_i / 2$ 表示使用强度越高, 复杂性、监督、token 和治理压力越重; $f_i a_i^2 / 2$ 表示深度采用需要固定集成和组织改造, 即使当前产量不变也要付出成本。

命题 4 (压缩成本的方向)

在 $d_i(a_i) = 0$ 的基准情形下, 给定产量 q_i , AI/agent 使用强度对成本的边际影响为

$$\frac{\partial C_i^{red}}{\partial a_i} = -\beta_i q_i + \lambda_i a_i q_i + f_i a_i.$$

若 a_i 很低, 采用是否降低成本主要取决于 β_i ; 若 a_i 较高, λ_i 和 f_i 会限制继续加深采用。

证明

在 $d_i(a_i) = 0$ 时, 只需要对

$$\left[-\beta_i a_i + \frac{\lambda_i}{2} a_i^2 \right] q_i + \frac{f_i}{2} a_i^2$$

关于 a_i 求导, 得到命题中的式子。 □

这个式子说明, β_i 决定一阶节约, $\lambda_i q_i + f_i$ 决定采用强度的二阶成本。给定 q_i 时, 采用收益是一条开口向下的二次曲线, 企业不会因为节约能力就无限加深采用, 而是停在边际节约等于边际使用成本的位置。

6.5 条件采用强度

先固定 q_i , 并暂时令 $d_i(a_i) = 0$ 。这相当于问: 如果企业预期自己要处理 q_i 单位业务, 它应该把 AI/agent 用到多深? 由于收入端 $P(Q)q_i$ 在给定 q_i 时不随 a_i 变化, 企业选择 a_i 等价于最小化压缩成本中和 a_i 有关的部分:

$$\min_{a_i \in [0,1]} \left[-\beta_i a_i + \frac{\lambda_i}{2} a_i^2 \right] q_i + \frac{f_i}{2} a_i^2.$$

一阶条件为

$$-\beta_i q_i + \lambda_i a_i q_i + f_i a_i = 0.$$

因此内点条件采用强度是

$$a_i^c(q_i) = \frac{\beta_i q_i}{\lambda_i q_i + f_i}.$$

考虑边界 $a_i \in [0, 1]$, 最优采用强度写为

$$a_i^*(q_i) = \min \{1, \max \{0, a_i^c(q_i)\}\}.$$

这个结果给出三个比较静态:

$$\begin{aligned} \frac{\partial a_i^c}{\partial \beta_i} &= \frac{q_i}{\lambda_i q_i + f_i} > 0, \\ \frac{\partial a_i^c}{\partial f_i} &= -\frac{\beta_i q_i}{(\lambda_i q_i + f_i)^2} < 0, \quad \frac{\partial a_i^c}{\partial \lambda_i} = -\frac{\beta_i q_i^2}{(\lambda_i q_i + f_i)^2} < 0. \end{aligned}$$

若 $\beta_i > 0$, 并且 $f_i > 0$, 还有

$$\frac{\partial a_i^c}{\partial q_i} = \frac{\beta_i f_i}{(\lambda_i q_i + f_i)^2} > 0.$$

数学上, β_i 提高会提高采用强度, λ_i 和 f_i 提高会降低采用强度。图像上, q_i 越大, 固定采用成本越容易被摊薄, 所以 $a_i^c(q_i)$ 上升; 当 q_i 很大时, $a_i^c(q_i)$ 接近 β_i/λ_i 。经济学上, 这也是大企业可能更愿意深度采用 AI/agent 的第一条渠道: 并非大企业天然更聪明, 而是同样的固定集成成本可以分摊到更多业务上。

6.6 固定门槛与规模效应

上面的 $f_i a_i^2/2$ 表示采用深度带来的固定集成成本。如果企业还必须先支付一次性进入成本 F_i^0 , 例如采购系统、法律审查、数据管线重建或安全认证, 那么采用还需要通过一个门槛。

在不考虑边界约束时, 采用带来的最大成本节约为

$$\max_{a_i} \left\{ \beta_i a_i q_i - \frac{1}{2} (\lambda_i q_i + f_i) a_i^2 \right\} = \frac{\beta_i^2 q_i^2}{2(\lambda_i q_i + f_i)}.$$

因此, 有一次性固定成本时, 企业愿意进入深度采用的条件是

$$\frac{\beta_i^2 q_i^2}{2(\lambda_i q_i + f_i)} \geq F_i^0.$$

这个门槛式解释了为什么“暴露度高”不等于“采用深”。一家企业可能有很多任务适合 AI, 但如果 F_i^0 或 f_i 很高, 或者 β_i 很低, 它仍然只会浅层使用。对中小企业来说, 问题常常不在于看不到 AI 的潜在用途, 而在于采用收益不足以覆盖固定改造成本。

6.7 规模凸性 $d_i(a_i) > 0$

前面的条件采用强度忽略了 $d_i(a_i)$ 。这样做有助于得到最清楚的基准，但不能忘记 Section 5 里的错误扩散、token 容量和治理拥挤。若保留规模凸性，给定 q_i 时，采用相关成本变为

$$\left[-\beta_i a_i + \frac{\lambda_i}{2} a_i^2 \right] q_i + \frac{f_i}{2} a_i^2 + \frac{d_i(a_i)}{2} q_i^2.$$

一阶条件为

$$-\beta_i q_i + (\lambda_i q_i + f_i) a_i + \frac{1}{2} d'_i(a_i) q_i^2 = 0.$$

若 $d'_i(a_i) \geq 0$ ，最后一项会降低最优采用强度。直觉是：采用越深，规模扩大后错误、复核、延迟、超容量 token 和监管拥挤越明显，企业就不会像常数边际成本模型预测的那样一直加深使用。

一个简单特例是

$$d_i(a_i) = d_{0i} + d_{1i} a_i^2, \quad d_{1i} \geq 0.$$

此时一阶条件变成

$$-\beta_i q_i + (\lambda_i q_i + f_i + d_{1i} q_i^2) a_i = 0,$$

内点采用强度为

$$a_i^d(q_i) = \frac{\beta_i q_i}{\lambda_i q_i + f_i + d_{1i} q_i^2}.$$

这和上一节的 $a_i^c(q_i)$ 形成对比。固定成本摊薄使大规模企业更愿意采用；规模凸性又使高规模深度采用变得更贵。大企业的最终采用优势，取决于固定成本摊薄是否强于错误扩散、token 容量和治理拥挤。

6.8 通向市场竞争模型

竞争分析采用两阶段解释。第一阶段，企业根据预期业务规模 $\bar{q}_i$ 选择采用强度：

$$a_i^*(\bar{q}_i) = \min \left\{ 1, \max \left\{ 0, \frac{\beta_i \bar{q}_i}{\lambda_i \bar{q}_i + f_i} \right\} \right\},$$

或在保留规模凸性时使用 $a_i^d(\bar{q}_i)$ 。第二阶段，企业把已经形成的采用强度带入市场竞争，面对有效成本

$$C_i(q_i | a_i) = m_i(a_i) q_i + \frac{d_i(a_i)}{2} q_i^2 + \text{固定项}.$$

竞争分析先从 $d_i(a_i) = 0$ 的常数边际成本 Cournot 推起。此时 a_i 的作用就是移动企业的有效边际成本 $m_i(a_i)$ 。在这个基准之后，再讨论 $d_i(a_i) > 0$ 如何让规模扩大变得更贵，并比较价格竞争、差异化产品、主导企业和进入退出等市场结构下的传导差异。

成本压缩的核心关系是：

Section 4 和 Section 5 的复杂成本机制，
最终通过 $\beta_i, \lambda_i, f_i, m_i(a_i), d_i(a_i)$ 进入企业竞争。

7 市场竞争与市场结构：Cournot 基准与扩展

7.1 从成本压缩进入竞争模型

成本压缩给出了竞争阶段需要的有效成本。为避免同时求解采用和产量，采用两阶段口径：企业先根据预期规模选择 a_i ，形成有效成本；随后在给定 a_i 的情况下参与市场竞争。现实中这两个决策会相互反馈，但两阶段写法能把采用机制和竞争机制分开看清。

Cournot 数量竞争提供了最清楚的基准：成本下降怎样改变反应函数、均衡产量、价格和份额，可以一步步算出来。现实中的平台、电商、软件和专业服务市场还可能更接近价格竞争、差异化产品竞争、主导企业加竞争边缘，或者长期进入退出过程。因此，Cournot 在这里承担完整推导的角色，其他市场结构用于检查同一成本冲击在不同竞争环境下的传导强度。

反需求仍为

$$P(Q) = A - BQ, \quad A > 0, \quad B > 0.$$

当暂时忽略规模凸性时，企业 i 的竞争阶段成本写成

$$C_i(q_i | a_i) = m_i(a_i)q_i + \text{固定项}.$$

这里的固定项包括上一章中的采用固定成本、一次性进入成本和其他不随当前产量变化的支出。它们会影响企业是否采用、采用多深，也会影响净利润，但在给定 a_i 的 Cournot 产量一阶条件中不会出现。这个口径很重要：若把 f_i 或 F_i^0 又放进边际成本，就会把同一笔采用成本算两遍。

在这个阶段，AI/agent 的作用先集中到有效边际成本

$$m_i(a_i) = c_i^0 - \beta_i a_i + \frac{\lambda_i}{2} a_i^2.$$

β_i 越高，低强度采用越容易降本； λ_i 越高，深度使用越容易遇到 token、审核、治理和复杂性压力。两者合在一起决定 $m_i(a_i)$ 是否真的下降。

7.2 无 AI 的两企业 Cournot 基准

先把 AI 放到一边，令两家企业的边际成本为常数 m_i 和 m_j 。企业 i 的利润是

$$\pi_i = [A - B(q_i + q_j)] q_i - m_i q_i.$$

展开后可写成

$$\pi_i = (A - m_i)q_i - Bq_i^2 - Bq_i q_j.$$

给定竞争对手产量 q_j ，企业 i 选择 q_i 。一阶条件为

$$\frac{\partial \pi_i}{\partial q_i} = A - m_i - 2Bq_i - Bq_j = 0.$$

二阶导数为 $-2B < 0$ ，所以一阶条件给出最大值。由此得到企业 i 的反应函数：

$$q_i = \frac{A - m_i - Bq_j}{2B}.$$

同理，企业 j 的反应函数为

$$q_j = \frac{A - m_j - Bq_i}{2B}.$$

把两个反应函数联立，得到两企业 Cournot 均衡：

$$q_i^* = \frac{A - 2m_i + m_j}{3B}, \quad q_j^* = \frac{A - 2m_j + m_i}{3B}.$$

总产量和价格为

$$Q^* = \frac{2A - m_i - m_j}{3B}, \quad P^* = \frac{A + m_i + m_j}{3}.$$

这些式子需要满足内点条件 $q_i^* > 0, q_j^* > 0$ 。如果某家企业成本太高，内点解不再适用，模型会转为角点情形；主线先讨论两家都生产的区间。

利润也可以由一阶条件直接写出。由于均衡时

$$P^* - m_i = Bq_i^*,$$

企业 i 的经营利润为

$$\pi_i^* = (P^* - m_i)q_i^* = B(q_i^*)^2.$$

这里说的是竞争阶段经营利润。如果要计算采用后的净利润，还需要减去上一章讨论的固定采用成本。

7.3 给定采用强度下的两企业 Cournot

现在把 AI/agent 放回模型。给定 a_i 和 a_j ，企业面对的有效边际成本变为

$$m_i(a_i) = c_i^0 - \beta_i a_i + \frac{\lambda_i}{2} a_i^2, \quad m_j(a_j) = c_j^0 - \beta_j a_j + \frac{\lambda_j}{2} a_j^2.$$

把上一节的 m_i, m_j 换成 $m_i(a_i), m_j(a_j)$ ，均衡产量为

$$q_i^*(a_i, a_j) = \frac{A - 2m_i(a_i) + m_j(a_j)}{3B},$$

$$q_j^*(a_i, a_j) = \frac{A - 2m_j(a_j) + m_i(a_i)}{3B},$$

价格为

$$P^*(a_i, a_j) = \frac{A + m_i(a_i) + m_j(a_j)}{3}.$$

采用强度通过成本进入均衡。对 $m_i(a_i)$ 求导：

$$\frac{\partial m_i}{\partial a_i} = -\beta_i + \lambda_i a_i.$$

当 $a_i < \beta_i/\lambda_i$ 时, 企业 i 加深采用会降低自身有效边际成本; 当 $a_i > \beta_i/\lambda_i$ 时, 继续加深采用反而会推高有效边际成本。这一点把 Section 5 的非劳动成本带回了竞争模型: token、审核和治理压力不是背景噪音, 它们会决定“更多 AI”是否还能继续降本。

由 Cournot 均衡式可得

$$\frac{\partial q_i^*}{\partial m_i} = -\frac{2}{3B}, \quad \frac{\partial q_i^*}{\partial m_j} = \frac{1}{3B}, \quad \frac{\partial P^*}{\partial m_i} = \frac{1}{3}.$$

再用链式法则,

$$\begin{aligned} \frac{\partial q_i^*}{\partial a_i} &= -\frac{2}{3B}(-\beta_i + \lambda_i a_i) = \frac{2(\beta_i - \lambda_i a_i)}{3B}, \\ \frac{\partial P^*}{\partial a_i} &= \frac{-\beta_i + \lambda_i a_i}{3}. \end{aligned}$$

因此, 在 $a_i < \beta_i/\lambda_i$ 的区间, 企业 i 加深采用会提高自身产量并压低市场价格; 超过这个区间后, 采用加深带来的成本压力会削弱甚至反转这个结果。

命题 5 (成本移位下的 Cournot 比较静态)

在两企业内点 Cournot 均衡中, 若企业 i 的 AI/agent 采用满足 $a_i < \beta_i/\lambda_i$, 则提高 a_i 会提高 q_i^* 、降低 P^* , 并降低竞争对手的均衡产量。若 $a_i > \beta_i/\lambda_i$, 上述方向可能反转。

证明

由

$$\frac{\partial m_i}{\partial a_i} = -\beta_i + \lambda_i a_i$$

可知, 当 $a_i < \beta_i/\lambda_i$ 时, $\partial m_i/\partial a_i < 0$ 。又因为

$$\frac{\partial q_i^*}{\partial m_i} = -\frac{2}{3B} < 0, \quad \frac{\partial P^*}{\partial m_i} = \frac{1}{3} > 0, \quad \frac{\partial q_j^*}{\partial m_i} = \frac{1}{3B} > 0,$$

所以提高 a_i 会提高自身产量、降低价格, 并降低对手产量。若 $a_i > \beta_i/\lambda_i$, 则 $\partial m_i/\partial a_i > 0$, 比较静态方向相反。□

在 Cournot 中, AI/agent 首先是一个成本移位变量: 降本企业的反应函数向外移动, 均衡点朝该企业更高产量、对手更低产量的方向移动。这也意味着, 采用 AI 不只是企业内部效率问题, 它会通过产量选择改变对手处境和市场价格。

7.4 对称企业与 n 企业推广

如果所有企业相同, 令 $m_i(a_i) = m(a)$ 。两企业对称均衡为

$$q^*(a) = \frac{A - m(a)}{3B}, \quad Q^*(a) = \frac{2[A - m(a)]}{3B}, \quad P^*(a) = \frac{A + 2m(a)}{3}.$$

在 $m'(a) < 0$ 的区间, 采用加深使价格下降、总产量上升。这就是最朴素的“生产型 AI 提高效率并向消费者传导”的机制。但这个结论依赖两个条件: 企业采用确实降低 $m(a)$, 并且竞争足以让降本传导到价格。

这一结论可以推广到 n 家企业。给定常数边际成本 $m_1, \dots, m_n$ ，企业 i 的一阶条件为

$$A - BQ - Bq_i - m_i = 0.$$

联立 n 个一阶条件可得

$$q_i^* = \frac{A + \sum_{k=1}^n m_k - (n+1)m_i}{B(n+1)}.$$

总产量和价格为

$$Q^* = \frac{nA - \sum_{k=1}^n m_k}{B(n+1)}, \quad P^* = \frac{A + \sum_{k=1}^n m_k}{n+1}.$$

对称情形下，

$$q^* = \frac{A - m}{B(n+1)}, \quad P^* = \frac{A + nm}{n+1}.$$

这组公式说明，企业数量 n 越大，单个企业产量越小，价格越接近边际成本；但只要企业之间成本不同，低成本企业仍会获得更高份额。

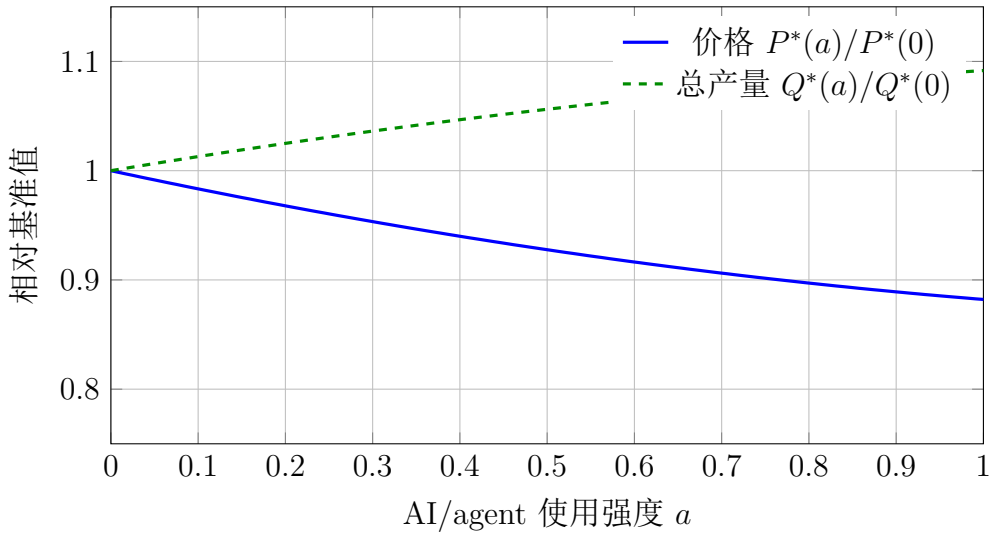


图 5: 对称 Cournot 中 AI/agent 降低边际成本时，均衡价格下降而总产量上升。参数为 $A = 80, B = 1, n = 3, c_0 = 20, \beta = 8, \lambda = 5$ 。

图 5 使用 $n = 3$ 的对称 Cournot 例子。参数下 $m(a) = 20 - 8a + 2.5a^2$ ，在 $a \in [0, 1]$ 内有效边际成本下降，所以均衡价格下降、总产量上升。图不预测真实幅度，只把 $m(a)$ 进入 Cournot 的方向画出来。

7.5 采用差异、成本差距与利润差距

真实市场中，企业不会拥有相同的 β_i, λ_i, f_i 。大企业可能有更多数据、更成熟的 IT 系统、更强的采购能力和更专业的合规团队，因此 β_i 更高、 λ_i 或 f_i 更低。小企业也可

能通过低价 API 获得技术平权，但若缺乏数据、流程和人才，实际使用强度仍可能较低。

考虑两家企业 H 和 L 。它们的有效成本差距为

$$\begin{aligned}\Delta m &\equiv m_L(a_L) - m_H(a_H) \\ &= (c_L^0 - c_H^0) - \beta_L a_L + \beta_H a_H + \frac{\lambda_L}{2} a_L^2 - \frac{\lambda_H}{2} a_H^2.\end{aligned}$$

这个式子把“谁更有 AI 优势”说得更具体。高效率企业不一定只是 c_H^0 更低；它也可能是 β_H 更高、 λ_H 更低，或者上一章的固定成本摊薄使 a_H 更高。

由两企业 Cournot 均衡，

$$q_H^* - q_L^* = \frac{m_L(a_L) - m_H(a_H)}{B} = \frac{\Delta m}{B}.$$

经营利润差距也可以写成

$$\pi_H^* - \pi_L^* = B [(q_H^*)^2 - (q_L^*)^2] = \Delta m \cdot Q^*.$$

因此，只要高效率企业形成正的成本优势 $\Delta m > 0$ ，它就会获得更高产量和更高经营利润。若计算净利润，还要减去两家企业各自的固定采用成本；这也是为什么 f_i 和 F_i^0 不应被忘掉，只是它们不属于竞争阶段的边际成本。

命题 6（异质采用扩大经营份额差距）

若企业 H 通过更高 β_H 、更低 λ_H 或更高采用强度实现 $m_H(a_H) < m_L(a_L)$ ，则在两企业内点 Cournot 均衡中 $q_H^* > q_L^*$ 。成本优势 $\Delta m = m_L(a_L) - m_H(a_H)$ 越大，产量差距和经营利润差距越大。

证明

由

$$q_H^* - q_L^* = \frac{\Delta m}{B}$$

直接可得，当 $\Delta m > 0$ 时 $q_H^* > q_L^*$ ，且差距随 Δm 线性上升。又因为

$$\pi_H^* - \pi_L^* = B(q_H^* - q_L^*)(q_H^* + q_L^*) = \Delta m Q^*,$$

所以在 $Q^* > 0$ 时，经营利润差距也随 Δm 上升。 □

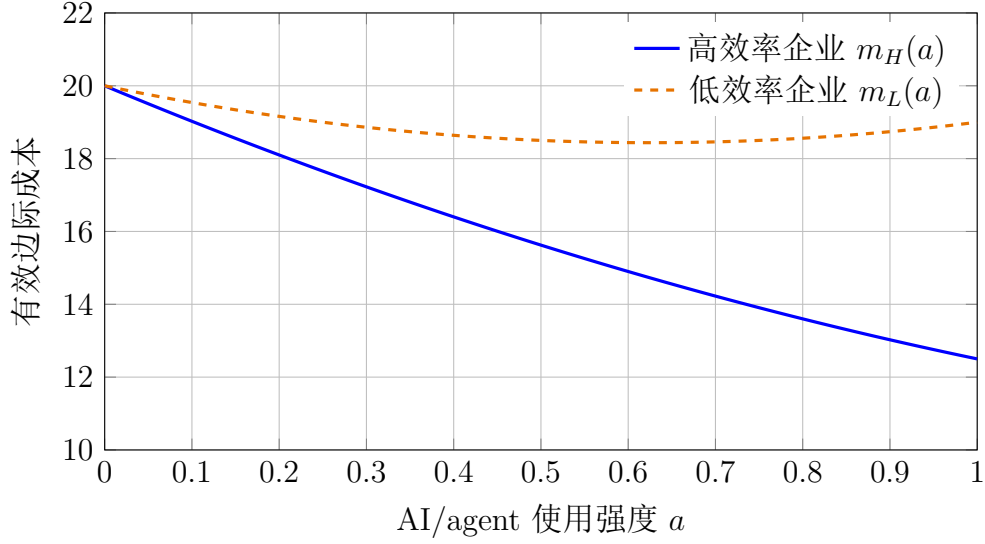


图 6: 异质企业的 AI/agent 成本差距。更高的节约能力或更低的治理成本使高效率企业获得更低边际成本，从而在 Cournot 中扩大份额。

Δm 把”谁更有 AI 优势”说得更具体。高效率企业不一定只是 c_H^0 更低；它也可能是 β_H 更高、 λ_H 更低，或者固定成本摊薄使 a_H 更高。图 6 显示，高效率企业的 $m_H(a)$ 随采用强度下降更多，低效率企业的 $m_L(a)$ 下降较少且较早受到成本压力限制。成本曲线之间的垂直距离，就是同一种 agent 技术可能让企业之间有效边际成本差距扩大的来源。

这种异质性与数字经济文献相吻合。数据非竞争性使数据可重复用于模型训练、推荐、风控和运营优化 (Jones and Tonetti 2020)；数据积累又可能提高企业学习效率 (Farboodi and Veldkamp 2021)。因此，已有规模优势可能通过 AI/agent 被放大，而不是自动消失。

7.6 市场份额与 HHI

两企业市场总产量为

$$Q^* = \frac{2A - m_H - m_L}{3B}.$$

高效率企业的份额为

$$s_H = \frac{q_H^*}{Q^*} = \frac{A - 2m_H + m_L}{2A - m_H - m_L}.$$

在 $A > m_L$ 的内点区间，

$$\frac{\partial s_H}{\partial m_H} = \frac{-3(A - m_L)}{(2A - m_H - m_L)^2} < 0.$$

所以，当 H 的有效边际成本下降而 L 的成本不变时， H 的市场份额上升。

两企业 **HHI**¹⁵ 为

$$HHI = s_H^2 + (1 - s_H)^2.$$

对 s_H 求导：

$$\frac{\partial HHI}{\partial s_H} = 4s_H - 2.$$

当 $s_H > 1/2$ 时， s_H 继续上升会提高 HHI。也就是说，若 AI/agent 主要强化领先企业的成本优势，市场集中度会上升；若它主要降低落后企业或新进入者的采用门槛，市场也可能更分散。

在 n 企业情形下，市场份额为

$$s_i = \frac{q_i^*}{Q^*},$$

集中度可写成

$$HHI = \sum_{i=1}^n s_i^2.$$

没有必要把 n 企业 HHI 的所有偏导数都写出。关键是： β_i, λ_i, f_i 决定 a_i ， a_i 决定 $m_i(a_i)$ ，成本差异再通过 Cournot 公式变成份额差异。市场结构并非另一个独立故事，它是成本重组传导到均衡产量后的结果。

7.7 规模凸性 $d_i(a_i) > 0$ 的修正

前面先令 $d_i(a_i) = 0$ ，是为了得到清楚的基准。现在把上一章的规模凸性放回来。给定采用强度，企业 i 的竞争阶段成本写成

$$C_i(q_i | a_i) = m_i(a_i)q_i + \frac{d_i(a_i)}{2}q_i^2 + \text{固定项}.$$

利润为

$$\pi_i = [A - B(q_i + q_j)]q_i - m_i(a_i)q_i - \frac{d_i(a_i)}{2}q_i^2.$$

对 q_i 求导：

$$\frac{\partial \pi_i}{\partial q_i} = A - m_i(a_i) - Bq_j - [2B + d_i(a_i)]q_i = 0.$$

因此反应函数变为

$$q_i = \frac{A - m_i(a_i) - Bq_j}{2B + d_i(a_i)}.$$

和常数边际成本版本相比， $d_i(a_i)$ 出现在分母中。它表示产量越大，额外一单位业务越贵；反应函数也因此变平：

$$\frac{\partial q_i}{\partial q_j} = -\frac{B}{2B + d_i(a_i)}.$$

¹⁵**HHI** (Herfindahl-Hirschman Index; 赫芬达尔-赫希曼指数)：衡量市场集中度的指标，等于各企业市场份额平方和。

若 $d_i(a_i)$ 上升, 企业对市场扩张的反应会更弱。这里的经济含义和 Section 5 对应: token 容量、错误扩散、治理拥挤和非工资协调拥挤, 会让“降本后扩大规模”变得没有常数边际成本模型那么顺畅。

若两家企业都存在规模凸性, 记

$$D_i = 2B + d_i(a_i), \quad D_j = 2B + d_j(a_j).$$

联立两条反应函数可得

$$q_i^* = \frac{D_j[A - m_i(a_i)] - B[A - m_j(a_j)]}{D_i D_j - B^2},$$

$$q_j^* = \frac{D_i[A - m_j(a_j)] - B[A - m_i(a_i)]}{D_i D_j - B^2}.$$

当 $d_i = d_j = 0$ 时, $D_i = D_j = 2B$, 上式退回前面的两企业 Cournot 解。这个退回关系说明这里没有另起模型, 只是在基准模型上加入规模压力。

规模凸性带来的结论要谨慎说。它通常会削弱高采用企业的扩张幅度, 但不会机械地消除大企业优势。若领先企业的 β_i 很高、固定成本摊薄很强, $m_i(a_i)$ 的下降仍可能大于 $d_i(a_i)$ 的上升; 若 token 容量和治理拥挤很强, 深度采用反而可能使扩张受限。市场集中度最终取决于这两股力量谁更强。

7.8 Cournot 之外的市场结构

Cournot 把企业的核心选择写成产量。这个基准清楚, 但并不是所有行业的好近似。若企业直接选择价格, 若产品存在差异, 若一家企业已经占据主导地位, 或者若长期进入退出很重要, 同样的 $m_i(a_i)$ 、 $d_i(a_i)$ 、 f_i 会通过不同路径影响价格和集中度。下面的扩展不重建五套完整模型, 只检查哪些结论依赖 Cournot, 哪些来自更一般的成本传导机制。

Bertrand 价格竞争。 先考虑同质产品价格竞争。若两家企业产品完全相同、消费者只买低价企业的产品, 并且企业边际成本为 $m_H(a_H) < m_L(a_L)$, 则低成本企业 H 可以用略低于 $m_L(a_L)$ 的价格服务市场。均衡价格通常落在

$$m_H(a_H) \leq p^* \leq m_L(a_L)$$

之间; 在标准的无容量约束、无价格离散条件下, 可以把它理解为 $p^* \approx m_L(a_L)$ 。这和 Cournot 的区别很大。Cournot 中, 低成本企业扩大产量但高成本企业仍可能保留一部分产量; Bertrand 中, 成本差距更容易转化为价格挤出和份额跳变。

经济含义是, AI/agent 若让领先企业形成 $m_H(a_H) < m_L(a_L)$, 在价格竞争市场中的集中压力可能强于 Cournot。反过来, 如果通用 agent 工具把中小企业成本迅速拉近, Bertrand 竞争也会更快把成本接近转化为低价格。这里的判断依赖产品同质性。电商平台若存在配送、会员、搜索体验和商家质量差异, 纯 Bertrand 只是一个极端参照。

差异化产品与加价率。许多数字市场并不是同质产品市场。平台、软件和专业服务往往存在品牌、界面、生态、服务质量和转换成本差异。此时企业可以直接选择价格，价格和边际成本之间的关系可用局部 **Lerner 指数**¹⁶ 概括：

$$\frac{p_i - MC_i}{p_i} = \frac{1}{|\varepsilon_i|}, \quad |\varepsilon_i| > 1.$$

若给定采用强度下的边际成本为

$$MC_i(q_i, a_i) = m_i(a_i) + d_i(a_i)q_i,$$

则局部价格可写为

$$p_i = \frac{MC_i(q_i, a_i)}{1 - 1/|\varepsilon_i|}.$$

这个式子把供给侧和需求侧放在一起。生产型 agent 通过 a_i 降低 MC_i ，但降本向价格传导的幅度还取决于需求弹性 $|\varepsilon_i|$ 。消费者搜索 agent 若降低搜索摩擦并提高需求弹性，就会压缩加价空间；企业 pricing agent 若提高个性化报价能力，则可能改变不同消费者面对的有效弹性。下一章会专门处理这条需求侧和策略侧机制。

主导企业与竞争边缘。还有一种市场结构介于完全对称寡头和垄断之间：一家主导企业 D 面对若干竞争边缘企业 F 。主导企业的有效边际成本为 $m_D(a_D)$ ，边缘企业的成本为 $m_F(a_F)$ 。若边缘企业的总供给为 $Q_F(P)$ ，市场需求为 $D(P)$ ，主导企业面对的剩余需求就是

$$Q_D(P) = D(P) - Q_F(P).$$

当 $P < m_F(a_F)$ 时，边缘企业可能退出供给，主导企业面对的剩余需求更大；当通用 agent 工具降低 $m_F(a_F)$ 时，竞争边缘的供给曲线向外移动，主导企业的剩余需求被压缩。

这组关系能解释一个 Cournot 基准不容易直接表达的现象。若大型平台既有高 β_D 、低 λ_D ，又能把 f_D 摊薄到大量订单上，AI/agent 可能强化主导企业地位。若第三方模型、低成本审计和标准化接口主要降低边缘企业的 f_F 与 λ_F ，它们反而会削弱主导企业的定价空间。市场结构方向仍然是条件化的。

进入退出与长期集中度。Cournot 一阶条件只使用边际成本，所以固定采用成本不进入反应函数。但这不表示固定成本不重要。长期市场结构取决于企业能否覆盖固定采用成本、治理成本和进入成本。令企业 i 的净利润为

$$\Pi_i^{net} = \Pi_i^{op}(m_i(a_i)) - K_i^{adopt}(a_i, z_i) - F_i^0,$$

其中 Π_i^{op} 是竞争阶段经营利润， K_i^{adopt} 是采用和治理 AI/agent 的固定资源成本， F_i^0 是一次性进入成本或其他沉没成本。企业进入或继续经营的条件是

$$\Pi_i^{net} \geq 0.$$

¹⁶**Lerner 指数** (Lerner index; 勒纳指数)：价格减边际成本再除以价格，衡量企业加价能力。

若长期自由进入，边际进入者通常满足 $\Pi_i^{net} = 0$ ，由此决定均衡企业数量 N^e 。

这个入口把 Section 6 的 f_i 和 F_i^0 放回了市场结构。高固定采用成本会提高有效进入门槛，使长期企业数量下降并提高集中度；标准化工具、共享审计、低价推理服务和互操作接口则可能降低门槛，使 N^e 上升。注意这里没有把固定成本塞进 Cournot 的边际条件。它影响的是“谁能进入和深度采用”，不是给定 a_i 后多生产一单位的边际成本。

平台网络效应作为边界。 电商平台还可能有双边网络效应。买家越多，商家越愿意入驻；商家越多，买家越容易找到合适产品。AI/agent 又可能强化这种反馈：消费者搜索 agent 改善匹配，商家运营 agent 降低上架和广告成本，平台推荐系统积累更多数据。可以用很简化的方式写成

$$U_C = u_C(n_S, p_C, b), \quad U_S = u_S(n_C, p_S, a_i),$$

其中 n_C 和 n_S 分别表示消费者侧和商家侧规模。这个方向很重要，但完整求解双边平台模型会明显改变论文重心：若同时求解双边价格、网络效应、搜索 agent 和 pricing agent，问题会从成本冲击传导转向平台经济的完整动态模型。

网络效应在这里作为边界条件处理。它提醒我们，Cournot 和价格竞争扩展主要解释成本和价格通道；在强平台市场中，数据、匹配和跨边网络反馈可能进一步放大或削弱这些结果。

表 2: 不同市场结构下 AI/agent 成本冲击的传导方式

市场结构	核心选择变量	AI/agent 进入方式	市场结构含义
Cournot 数量竞争	q_i	$m_i(a_i), d_i(a_i)$ 改变反应函数	成本下降提高产量和份额，HHI 随成本差距变化。
Bertrand 价格竞争	p_i	成本差距决定可持续低价	同质产品下，低成本企业更容易获得强份额优势。
差异化产品定价	p_i, ε_i	MC_i 和需求弹性共同决定加价率	搜索 agent 提高弹性，pricing agent 可能改变个体报价。
主导企业与边缘	$P, Q_D(P)$	主导企业和边缘企业的成本差距 改变剩余需求	AI 可能强化主导企业，也可能帮助边缘企业追赶。
进入退出	N^e, Π_i^{net}	K_i^{adopt}, F_i^0 改变进入门槛	固定采用成本影响长期企业数量，而不是短期 FOC。
平台网络效应	n_C, n_S	agent 改变匹配、搜索和商家运营	网络反馈可能放大成本优势，是模型的扩展边界。

7.9 供给侧竞争结果

压缩成本进入市场竞争以后，供给侧主线可以概括为：

$$\boxed{\beta_i, \lambda_i, f_i \Rightarrow a_i^* \Rightarrow m_i(a_i), d_i(a_i) \Rightarrow q_i^*, P^*, s_i, HHI, p_i, N^e.}$$

β_i 提高、 λ_i 和 f_i 降低，会让企业更容易采用。采用若降低 $m_i(a_i)$ ，企业在 Cournot 中会扩大产量并获得更高份额；在价格竞争中，成本差距可能更快转化为价格压力；在进入退出模型中，固定采用成本会影响长期企业数量。若采用同时提高 $d_i(a_i)$ ，规模扩张会受到 token、错误和治理压力限制。

这条链条给出供给侧结论：AI/agent 通过成本函数改变企业竞争，并且传导强度取决于市场结构。消费者搜索 agent 和企业 pricing agent 还会改变需求弹性、定价速度、价格歧视和合谋风险，它们需要在供给侧结果之外单独处理。

8 消费者搜索 agent 与企业 pricing agent：需求侧和策略侧机制

8.1 需求侧和策略侧的新摩擦

供给侧竞争说明了生产型 agent 如何通过 a_i 改变有效边际成本 $m_i(a_i)$ ，进而影响产量、价格、市场份额和 HHI。价格形成还取决于另外两类 agent：消费者是否能更低成本地搜索和比较，企业是否能更快、更细地调整价格。

消费者搜索 agent 改变需求侧摩擦：价格比较、质量识别和隐藏费用识别变得更便宜。企业 pricing agent 改变策略侧能力：企业可以更快更新价格、更细地识别消费者类型，并可能在重复互动中学习高价策略。这些变量通常不直接进入 $m_i(a_i)$ ，除非它们确实改变生产成本；它们主要改变需求弹性、加价空间和定价博弈。

现实中的 agent 既可能让企业生产更便宜，也可能让消费者更会比价，还可能让企业更会定价。三种作用方向不一定一致，所以需要把生产端、需求端和策略端分开。

8.2 消费者搜索 agent：从搜索成本到需求弹性

先看消费者侧。传统搜索理论强调，消费者了解价格和质量并不是免费的。如果搜索成本高，消费者不会查看所有卖家，企业就可能在有多个竞争者的情况下仍维持较高价格或价格分散 (Stigler 1961; Diamond 1971; Burdett and Judd 1983)。消费者 agent 的作用可以理解为把这类搜索活动自动化：比较价格、读取评价、识别隐藏费用、监测历史价格、总结退换货条款，甚至自动执行购买。

定义 5 (消费者搜索 agent 覆盖率 b)

令 $s \in [0, 1]$ 表示标准化搜索摩擦。这里 s 越高，表示消费者越难比较不同企业；后文的 s_i 仍表示企业 i 的市场份额，二者不是同一个变量。再令 $b \in [0, 1]$ 表示消费者搜索 agent 的覆盖率或使用强度。一个最简单的写法是

$$s(b) = s_0(1 - \omega b), \quad 0 \leq \omega \leq 1.$$

这里 s_0 是没有消费者 agent 时的搜索摩擦， ω 表示消费者 agent 把搜索摩擦转化为实际下降的效率。若 ω 很小，即使很多消费者安装了 agent，搜索成本也可能下降有限；这对应 agent 比价能力弱、平台限制抓取、商品质量难以比较等情形。

搜索摩擦下降后，企业面对的剩余需求通常更有弹性。为避免把需求侧模型写得过重，可以用一个局部表达：

$$\varepsilon_i(s) = \varepsilon_0 + \chi(1 - s), \quad \chi > 0.$$

ε_0 是搜索摩擦很高时企业面对的基础需求弹性, χ 表示搜索摩擦下降对弹性的放大程度。

这个式子并不声称所有产品都服从同一个需求函数; 它只是把搜索理论中的直觉压缩成一个可用于比较静态的关系: 消费者越容易比较, 越容易离开高价企业, 企业就越难维持高加价。

8.3 搜索成本下降如何改变加价率

上一章已经说明, 在差异化产品或局部垄断定价中, 成本和价格之间可以用 Lerner 指数连接。沿用这个口径, 消费者搜索 agent 对需求弹性的影响可以写成:

$$\frac{p_i - MC_i}{p_i} = \frac{1}{|\varepsilon_i(s)|}.$$

这里 MC_i 是企业定价时面对的边际成本, 可以由上一章的 $m_i(a_i)$ 给出; 若考虑规模凸性, 也可以写成 $m_i(a_i) + d_i(a_i)q_i$ 。差异化产品市场结构已经在供给侧扩展中讨论过, 这里把注意力放在 $\varepsilon_i(s)$: 消费者搜索 agent 如何改变弹性, 进而改变加价空间。

命题 7 (消费者搜索 agent 与加价率)

若 $s(b) = s_0(1 - \omega b)$ 、 $\varepsilon_i(s) = \varepsilon_0 + \chi(1 - s)$, 且企业的局部加价率满足 $L_i(b) = 1/\varepsilon_i(s(b))$, 则在 $\chi > 0$ 、 $\omega > 0$ 、 $s_0 > 0$ 且 $\varepsilon_i > 0$ 时, 消费者搜索 agent 使用强度 b 上升会降低企业的局部加价率。

证明

把 $s(b)$ 代入需求弹性, 有

$$\varepsilon_i(s(b)) = \varepsilon_0 + \chi[1 - s_0(1 - \omega b)].$$

对 b 求导:

$$\frac{d\varepsilon_i}{db} = \chi s_0 \omega > 0.$$

因此

$$\frac{dL_i}{db} = \frac{d}{db} \left(\frac{1}{\varepsilon_i(s(b))} \right) = -\frac{\chi s_0 \omega}{\varepsilon_i(s(b))^2} < 0.$$

□

消费者 agent 覆盖率越高, 搜索摩擦越低, 需求弹性越高, Lerner 加价率也就越低。其经济含义是, 生产型 agent 降低成本后, 价格是否充分下降, 还取决于消费者能不能把企业之间的价格差异看清楚。如果搜索摩擦仍高, 企业可能把一部分降本收益留作利润; 如果消费者 agent 普及, 企业更难这样做。

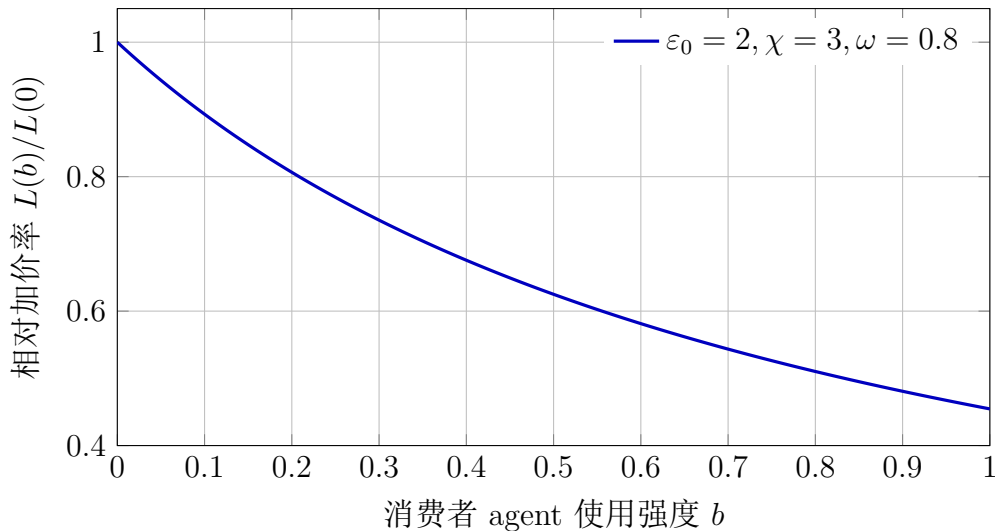


图 7: 消费者 agent 使用强度与相对加价率。参数为 $s_0 = 1, \omega = 0.8, \varepsilon_0 = 2, \chi = 3$, 因此 $\varepsilon(b) = 2 + 2.4b$ 。随着消费者 agent 覆盖率上升, 搜索摩擦下降、需求弹性上升, Lerner 加价率相对基准下降。

8.4 企业 pricing agent: 实时定价与个性化报价

企业侧的 pricing agent 改变的是定价策略空间。它可以实时读取竞争对手价格、库存、广告投放、天气、用户画像、转化率和历史购买记录, 再自动更新报价。与人工定价相比, 变化不只是“更快”, 还包括信息维度更高、实验频率更高、价格更容易按消费者类型变化。Varian (2019) 对 AI 与产业组织的讨论已经强调了这类机制, Z. Y. Brown and MacKay (2023) 则从算法定价角度展示了价格规则怎样改变竞争结果。

定义 6 (企业 pricing agent 个性化能力 φ_i)

为了表达个性化能力, 令 $\varphi_i \in [0, 1]$ 表示企业 i 的 pricing agent 个性化报价能力。若 p_i 是统一价格, 消费者类型 k 的支付意愿为 v_k , 可以写一个极端示意式:

$$p_{ik}(\varphi_i) = (1 - \varphi_i)p_i + \varphi_i(v_k - \epsilon), \quad \epsilon > 0.$$

当 $\varphi_i = 0$ 时, 企业只收统一价格 p_i 。当 $\varphi_i = 1$ 时, 价格接近按消费者支付意愿 v_k 报价。

这个式子只是机制表达, 不声称企业真的能完全观察 v_k 。更现实的理解是, 企业用搜索记录、购买频率、设备、地理位置、会员等级和历史响应来估计消费者类型, 因此 φ_i 越高, 价格越接近“按人报价”。

个性化定价的福利含义并不单向。它可能让一些原本价格太高而不购买的消费者进入市场, 也可能把原本属于消费者的剩余更多转给企业。若消费者 agent 同时存在, 结果还会更复杂: 消费者 agent 可以帮助识别差别定价和隐藏费用, 但企业 pricing agent 也可能更快学习消费者 agent 的行为规则。

8.5 pricing agent 与算法合谋风险

pricing agent 的另一个问题是重复博弈中的算法合谋¹⁷风险。这里说的合谋不一定是人工签订协议，而是算法在反复互动中学习到“维持高价、惩罚降价”的策略。相关实验和理论文献表明，这种可能性不能被简单排除 (Calvano et al. 2020a; Klein 2021; Hansen et al. 2021; Calvano et al. 2019)；真实市场中的经验识别也显示，算法定价的竞争后果需要具体分析 (Assad et al. 2023)。

用标准重复博弈条件可以把机制写清楚。令 π^C 表示维持高价合作时的单期利润， π^D 表示单期偏离降价时的利润， π^N 表示之后回到竞争或惩罚状态的利润， $\delta \in (0, 1)$ 是贴现因子。维持高价策略可持续的条件是

$$\frac{\pi^C}{1-\delta} \geq \pi^D + \frac{\delta\pi^N}{1-\delta}.$$

整理得

$$\delta \geq \frac{\pi^D - \pi^C}{\pi^D - \pi^N}.$$

这条式子的经济含义是：企业越重视未来、偏离带来的短期收益越小、惩罚越有效，高价策略越容易维持。pricing agent 可能通过更快监测竞争对手、更快执行惩罚和更稳定学习策略，降低维持高价所需的有效门槛。但这只是风险，不是必然结论。若消费者 agent 覆盖率高、比价容易、监管能审计算法策略，高价策略也可能更难维持。

8.6 两类 agent 同时存在时的四种情形

把消费者 agent 和企业 pricing agent 放在一起，可以得到四种情形。第一， b 低、 φ_i 低时，市场接近传统搜索摩擦市场，企业有一定加价空间，但个性化和算法学习能力有限。第二， b 高、 φ_i 低时，消费者比价能力增强，企业统一价格的加价空间被压缩，竞争更接近价格透明市场。第三， b 低、 φ_i 高时，消费者仍难以比较，而企业定价能力提高，价格歧视和高价维持风险更强。第四， b 高、 φ_i 高时，双方都用 agent，市场结果取决于谁的信息处理和策略适应更强，也取决于平台规则和监管约束。

	φ_i 低	φ_i 高
b 低	传统搜索摩擦	企业定价优势更强
b 高	比价压低加价	搜索竞争与个性化定价并存

这张简单分类不替代模型；它提醒案例分析，同样是 agent 普及，消费者侧和企业侧的普及方向可能相反。只讨论生产效率会低估定价权问题；只讨论算法合谋又会忽略消费者 agent 对搜索摩擦的削弱。

¹⁷算法合谋 (algorithmic collusion; 算法合谋)：企业定价算法在没有明示协议的情况下，通过重复互动学习到维持高价的策略。

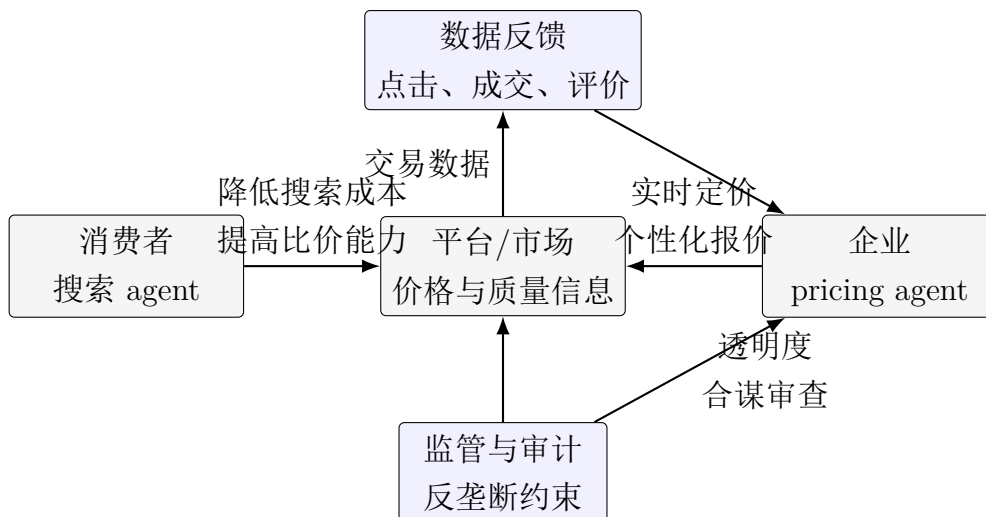


图 8: 消费者 agent 与企业 pricing agent 的博弈机制。消费者端降低搜索成本，企业端提高定价和学习能力，监管端影响算法策略空间。

8.7 三类 agent 的合成效应

生产型 agent 的主通道仍然是

$$a_i \Rightarrow m_i(a_i), d_i(a_i) \Rightarrow q_i^*, P^*, s_i, HHI.$$

消费者搜索 agent 的通道是

$$b \Rightarrow s(b) \downarrow \Rightarrow \varepsilon_i(s) \uparrow \Rightarrow \text{加价空间下降}.$$

企业 pricing agent 的通道是

$$\varphi_i \Rightarrow \text{个性化报价和策略学习} \Rightarrow \text{消费者剩余与竞争强度可能改变}.$$

合在一起看，agent 对市场结构和福利的影响没有单一方向。生产端 agent 可能降低成本并改变企业在不同竞争形式下的份额；消费者端 agent 可能加强比价并压低加价；企业定价端 agent 可能提高价格歧视能力，并在某些条件下提高算法合谋风险。电商平台案例就是这种混合情形：平台既采用生产型 agent 降本，也面对消费者比价 agent，还可能部署 pricing agent 进行动态报价。

9 统一案例：电商平台 A/B/C 的风格化分析

9.1 电商平台作为统一场景

电商平台是一个适合串联各类机制的场景。平台会在客服、推荐、广告、商家运营、内容审核和价格监测中使用生产型 agent；消费者也可能用搜索 agent 比较价格、配送费和售后条款；平台自己还可能部署 pricing agent 做动态报价和个性化优惠。同一组企业会同时受到生产端、需求端和策略端的影响。

案例只做风格化模拟，不对应某几家真实企业。它的作用更像一张带数值的模型地图：每个参数从哪里来，怎样进入公式，最后怎样改变价格、份额和 HHI。真实行业的因果效应仍需要数据和实证识别。

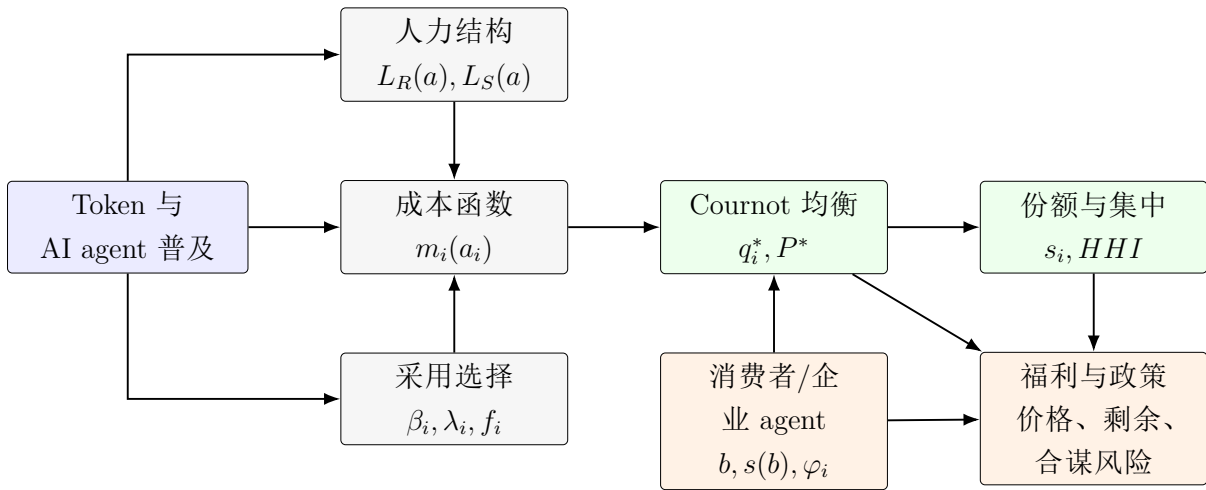


图 9: 全文统一机制。Token 与 AI agent 先改变企业采用选择和成本函数，再通过 Cournot 竞争改变价格、产量、份额和集中度；消费者搜索 agent 通过 b 和 $s(b)$ 改变需求弹性，企业 pricing agent 通过 φ_i 改变策略空间。

图 9 把机制合在一起。生产型 agent 改变企业的有效边际成本 $m_i(a_i)$ ，再通过竞争模型改变 q_i^* 、 P^* 、 s_i 和 HHI；数值部分使用 Cournot 作为基准。消费者搜索 agent 通过 b 和 $s(b)$ 改变消费者的比价能力；企业 pricing agent 通过 φ_i 改变个性化报价能力。这三条线方向不同，“AI 降本”不足以概括整个市场结果。

例 1（电商平台 A/B/C¹⁸ 的风格化案例）

9.2 案例设定：三个电商平台

考虑一个电商平台市场。三家平台 A、B、C 提供近似同质的交易服务，消费者关心价格、配送、搜索、客服和商家质量。为了让数值推导保持清楚，数值例子先采用 Cournot 基准，市场反需求设为

$$P(Q) = 100 - Q,$$

未采用 AI/agent 时的共同边际成本为 $c_0 = 40$ 。三家平台都可以把生产型 agent 用在客服、商家运营、库存预测、内容审核和价格监测中，但采用能力不同。

¹⁸电商平台 A/B/C（e-commerce platforms A/B/C；电商平台 A/B/C）：本文用于串联生产型、搜索型和定价型 agent 机制的风格化数值案例，不对应真实公司。

表 3: 电商平台 A/B/C 的风格化参数

平台	β_i	λ_i	f_i	a_i	$m_i(a_i)$	q_i^*	s_i
A: 大型平台	18	8	6	0.90	27.04	22.81	0.455
B: 中型平台	12	10	10	0.65	34.31	15.54	0.310
C: 小型平台	8	14	18	0.35	38.06	11.79	0.235

表 3 中的 a_i 是已经形成的实际采用强度，这里不重新求最优 a_i^c 。这个口径很重要。成本压缩部分已经说明 a_i 受预期规模、固定采用成本和使用成本压力影响；案例把 a_i 当作进入 Cournot 的成本移位变量，避免同时求解采用和产量而把例子变成另一个完整模型。

若这个市场更接近 Bertrand 价格竞争，成本差距会更快反映为价格压力和份额跳变；若产品差异更强，需求弹性和加价率会变得更重要；若进入退出更活跃， f_i 和 F_i^0 会通过企业数量影响集中度。这里不逐一重算这些情形，而是采用 Cournot 数值例子展示主链条，再在解释中保留这些市场结构差异。

平台 A 的 β_A 较高，表示同样加深 agent 使用时，它能更有效地节约客服、运营和协调成本； λ_A 较低，表示 token、审核、错误和治理的综合压力较小。平台 C 也能购买模型 API，但流程整合、数据管线和风控能力较弱，所以 λ_C 和 f_C 更高。这里的 f_i 不直接进入 Cournot 一阶条件，它已经通过影响采用深度体现在 a_i 中。

9.3 生产型 agent：从采用强度到成本和份额

按照 Section 6 的压缩成本，

$$m_i(a_i) = c_0 - \beta_i a_i + \frac{\lambda_i}{2} a_i^2.$$

代入三家平台参数，得到

$$m_A = 40 - 18(0.90) + \frac{8}{2}(0.90)^2 = 27.04,$$

$$m_B = 40 - 12(0.65) + \frac{10}{2}(0.65)^2 = 34.31,$$

$$m_C = 40 - 8(0.35) + \frac{14}{2}(0.35)^2 = 38.06.$$

这些数字说明，三家平台都采用了 agent，但采用后的有效边际成本并不相同。Cournot 竞争关心的是 $m_i(a_i)$ 的差距，而不是“是否使用 AI”这个二元变量。

三企业 Cournot 均衡沿用 Section 7 的 $n = 3$ 公式：

$$q_i^* = \frac{A + \sum_{j=A,B,C} m_j - 4m_i}{4B}.$$

把 $A = 100, B = 1$ 和三个成本代入，得到

$$q_A^* = 22.81, \quad q_B^* = 15.54, \quad q_C^* = 11.79.$$

总产量约为 $Q^* = 50.15$ ，价格为

$$P^* = 100 - 50.15 = 49.85.$$

市场份额分别为 $s_A = 0.455$ 、 $s_B = 0.310$ 、 $s_C = 0.235$ ，于是

$$HHI = 0.455^2 + 0.310^2 + 0.235^2 \approx 0.358.$$

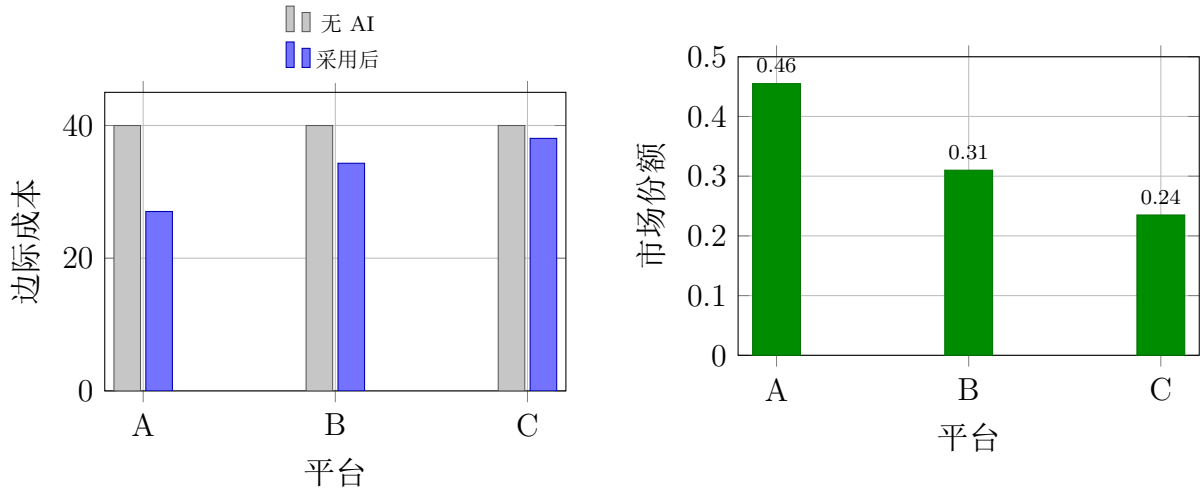


图 10: 电商平台 A/B/C 的风格化案例。平台 A 因更高 β 和更低治理成本获得更低边际成本，在 Cournot 均衡中取得更高份额；平台 C 虽也采用 AI/agent，但固定和治理成本较高，份额提升有限。

图 10 把成本和份额放在一起。它的含义很直接：生产型 agent 的第一层影响是成本移位，第二层影响才是份额变化。若平台 A 的 m_A 降得更多，它在 Cournot 均衡中会扩张产量；平台 B、C 即使也采用 agent，仍会因为降本较弱而丢掉相对份额。

这个例子没有预设“大平台必胜”。它只说明一个条件化结果：当大型平台同时具有高 β_i 、低 λ_i 和较低采用门槛时，AI/agent 会沿着既有能力差距放大成本差距。若小平台获得低成本工具、共享模型基础设施或更低审计成本，结果会改变。

9.4 消费者搜索 agent：比价能力与加价空间

生产端降本不自动等于消费者得到全部好处。价格还取决于消费者能否比较不同平台。沿用 Section 8 的设定，用 $b \in [0, 1]$ 表示消费者搜索 agent 的覆盖率或使用强度，搜索摩擦为

$$s(b) = s_0(1 - \omega b),$$

需求弹性为

$$\varepsilon(b) = \varepsilon_0 + \chi[1 - s(b)].$$

在案例中取 $s_0 = 1$, $\omega = 0.8$, $\varepsilon_0 = 2$, $\chi = 3$ 。若消费者搜索 agent 覆盖率为 $b = 0.7$ ，则

$$s(0.7) = 1(1 - 0.8 \times 0.7) = 0.44,$$

$$\varepsilon(0.7) = 2 + 3(1 - 0.44) = 3.68.$$

Lerner 指数给出的相对加价率为

$$\frac{L(0.7)}{L(0)} = \frac{1/\varepsilon(0.7)}{1/\varepsilon(0)} = \frac{2}{3.68} \approx 0.543.$$

这个计算的经济含义是：当消费者 agent 能自动比价、识别历史价格、比较配送费和售后条款时，同样的成本下降更容易通过竞争压力传导到价格。若消费者仍然面对高搜索成本，低成本平台可能把更多降本收益留为利润；若 b 上升，消费者更容易转向低价或高质量平台，企业加价空间会被压缩。

9.5 企业 pricing agent：个性化报价与合谋风险

需求侧 agent 增强消费者的比价能力，企业端 pricing agent 则给企业更强的报价和学习能力。令

$$\varphi_A = 0.80, \quad \varphi_B = 0.55, \quad \varphi_C = 0.25,$$

分别表示三家平台的个性化报价能力。沿用 Section 8 的表达式，

$$p_{ik}(\varphi_i) = (1 - \varphi_i)p_i + \varphi_i(v_k - \epsilon).$$

$\varphi_i = 0$ 时，平台主要使用统一价格； φ_i 越高，平台越能根据消费者类型、浏览路径、会员状态和价格敏感度调整报价。这里不假定平台能完全观察 v_k ，公式只是表示从统一价格到个性化报价的连续变化。

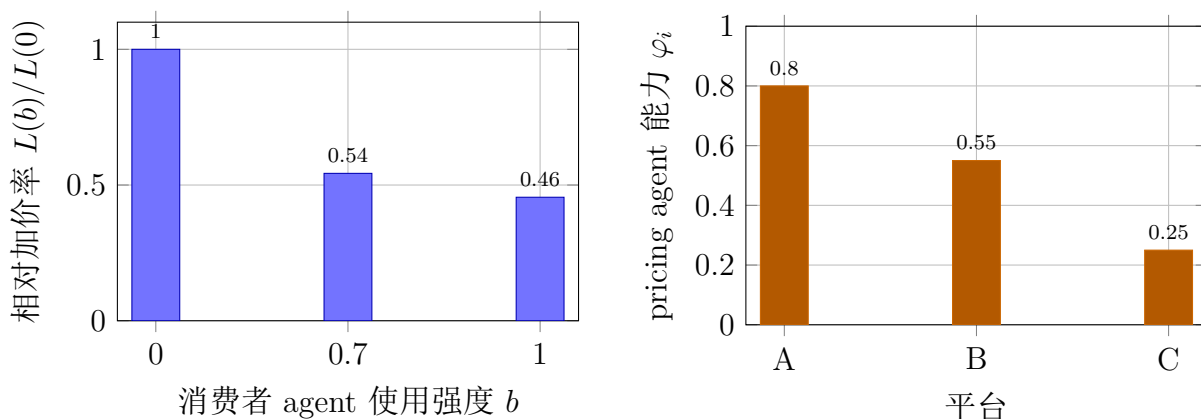


图 11: 电商案例中的需求侧与策略侧 agent。左图使用 $s_0 = 1, \omega = 0.8, \varepsilon_0 = 2, \chi = 3$ ，显示消费者 agent 覆盖率上升会压低相对加价率；右图显示平台 A/B/C 的个性化报价能力差异。

图 11 把需求侧和策略侧两层放在同一张图里。左图显示 b 上升会压低相对加价率；右图显示平台 A 的 pricing agent 能力最高。两张图合在一起说明一个容易被忽略的点：平台 A 可能同时拥有最低生产成本和最强个性化定价能力。因此，低边际成本不必然意味着消费者剩余最大。消费者能否获得更多剩余，取决于搜索 agent 的约束力量是否强过企业 pricing agent 的价格歧视能力。

pricing agent 还会改变重复定价博弈。沿用 Section 8 的条件，若合作高价、单期偏离和竞争惩罚状态下的单期利润分别为 π^C, π^D, π^N ，高价策略能够维持的条件是

$$\delta \geq \frac{\pi^D - \pi^C}{\pi^D - \pi^N}.$$

pricing agent 可能通过快速监测价格、自动识别偏离和稳定执行惩罚策略降低这个门槛；消费者搜索 agent、监管审计和平台透明度规则又可能提高维持高价的难度。案例中不再建立完整动态模型，因为主线仍是企业成本和供给侧竞争基准。这里要保留的结论只有一个：策略侧 agent 的福利方向是条件化的，不能简单写成“更智能的定价提高效率”。

9.6 反事实：中小平台获得低成本 agent 工具

为了检验模型是否预设“大平台必胜”，考虑一个反事实。假设开源模型、低价推理服务、标准化合规工具和第三方 agent 平台改善了 C 的采用环境，使

$$(\beta_C, \lambda_C, f_C, a_C) = (11, 8, 8, 0.70).$$

这里同时调整 β_C, λ_C 和 a_C 。原因很简单：给定 a_i 后，固定采用成本 f_i 不进入 Cournot 一阶条件。单独降低 f_C 会改变采用门槛和最优采用强度，但不会在固定 a_C 的 Cournot 阶段直接改变边际成本。

新成本为

$$m'_C = 40 - 11(0.70) + \frac{8}{2}(0.70)^2 = 34.26.$$

此时三家平台成本为

$$m_A = 27.04, \quad m_B = 34.31, \quad m'_C = 34.26.$$

重新代入三企业 Cournot 公式，可得

$$q_A^* = 21.86, \quad q_B^* = 14.59, \quad q_C^* = 14.64.$$

总产量约为 51.10，价格降至 48.90。市场份额变为

$$s_A \approx 0.428, \quad s_B \approx 0.286, \quad s_C \approx 0.287,$$

HHI 约为

$$HHI' \approx 0.347.$$

这个反事实没有让小平台反超大型平台，但它明显缩小了份额差距。经济含义也更稳：降低中小企业的固定采用成本本身很重要，但它必须通过提高实际采用强度、提高节约能力或降低使用成本压力，才能改变竞争阶段的边际成本 and 市场份额。政策若只补贴软件采购，却不改善数据接口、审计工具、模型可靠性和组织整合能力，对市场结构的影响可能有限。

9.7 三条通道的共同作用

这个电商案例可以压成三条通道。第一条是生产通道：

$$a_i \Rightarrow m_i(a_i) \Rightarrow q_i^*, P^*, s_i, HHI.$$

它说明生产型 agent 如何通过有效边际成本进入供给侧竞争；这里用 Cournot 作为可计算基准。第二条是需求通道：

$$b \Rightarrow s(b) \Rightarrow \varepsilon(b) \Rightarrow L(b).$$

它说明消费者搜索 agent 如何降低搜索摩擦并压缩加价空间。第三条是策略通道：

$$\varphi_i, \delta \Rightarrow \text{个性化报价与重复定价策略}.$$

它说明企业 pricing agent 可能改变消费者剩余和合谋风险。

三条通道合在一起后，福利判断不会只有一个方向。生产端可能降本，消费者端可能增强竞争，企业端也可能扩大价格歧视或提高高价维持能力。福利分析必须把效率收益、搜索摩擦下降、价格歧视、错误风险和算法合谋分开看；合成一个笼统的“AI 对福利的影响”，反而会掩盖真正的经济学问题。

10 综合福利、政策讨论与结论

10.1 为什么福利不能只看平均价格

AI/agent 的福利效应不能只从平均价格判断。生产型 agent 可能降低企业有效边际成本，消费者搜索 agent 可能压低加价空间，企业 pricing agent 又可能提高价格歧视能力或改变重复定价策略。平均价格即使下降，也不能回答两个关键问题：降本收益由谁获得，错误、隐私、能源和治理等风险成本是否已经被计入。

在电商平台案例中，平台 A 的生产成本最低，但它也可能拥有更强的个性化报价能力。消费者搜索 agent 覆盖率提高时，消费者更容易比较总价、配送费和售后条款；企业 pricing agent 更强时，企业又更可能按消费者类型报价。两种力量方向相反。福利分析因此要把生产效率、消费者搜索、定价策略和外部风险分开看。

福利部分采用账本口径，不再建立新的完整均衡模型。账本的作用是区分三类项：提高总剩余的效率收益，改变剩余归属的转移项，以及消耗真实资源或带来外部性的风险成本。

10.2 供给侧效率收益：成本下降、产量上升与固定采用成本

先以 Section 7 的 Cournot 基准计算供给侧效率收益。给定各企业采用强度 a_i 后，有效边际成本为 $m_i(a_i)$ 。在线性需求 $P(Q) = A - BQ$ 下，消费者剩余仍可写为

$$CS^{prod}(\mathbf{m}) = \frac{1}{2}B[Q^*(\mathbf{m})]^2,$$

其中 $\mathbf{m} = (m_1, \dots, m_n)$, $Q^*(\mathbf{m}) = \sum_i q_i^*(\mathbf{m})$ 。当生产型 agent 在有效区间内降低 $m_i(a_i)$ 时, 总产量上升, 消费者剩余通常增加。若市场更接近 Bertrand、差异化产品或进入退出模型, 成本下降向消费者剩余的传导会改变: 价格竞争可能强化传导, 差异化和高加价率可能削弱传导, 进入退出则会把固定采用成本带入长期企业数量。

企业经营利润的口径也要和 Cournot 推导一致。若先不扣固定采用成本, 异质企业下的经营利润为

$$\Pi^{op}(\mathbf{m}) = \sum_i B[q_i^*(\mathbf{m})]^2.$$

这个式子来自 Section 7 的一阶条件: 在线性 Cournot 中, 均衡经营利润等于 $B(q_i^*)^2$ 。它只衡量市场阶段的经营利润, 不包含系统集成、治理体系、员工培训和组织改造的固定成本。

固定采用成本应在净利润或福利中扣除, 而不是放入 Cournot 一阶条件。令

$$K_i^{adopt}(a_i, z_i)$$

表示企业 i 为采用和治理 AI/agent 支付的固定资源成本, 其中 z_i 是治理投入。它可以包括软件系统、数据管线、安全审计、合规工具和组织学习成本。于是生产侧的净福利贡献更接近

$$CS^{prod}(\mathbf{m}) + \Pi^{op}(\mathbf{m}) - \sum_i K_i^{adopt}(a_i, z_i).$$

固定成本和边际成本的区分在这里变得很关键: 固定成本影响采用门槛和净福利, 但给定 a_i 后, 它不直接改变 Cournot 反应函数。

规模凸性 $d_i(a_i) > 0$ 会削弱这个效率收益。token 容量约束、错误扩散、治理拥挤和组织协调压力会使大规模扩张变贵。若这些成本已经进入 $d_i(a_i)$, 生产端福利不能只看 $m_i(a_i)$ 下降, 还要看规模扩张是否带来新的拥挤成本。

10.3 消费者搜索 agent: 搜索摩擦下降与消费者剩余

消费者搜索 agent 的福利作用不通过企业成本进入, 而是通过搜索摩擦和需求弹性进入。沿用 Section 8 的记号,

$$s(b) = s_0(1 - \omega b), \quad \varepsilon(b) = \varepsilon_0 + \chi[1 - s(b)].$$

当 b 上升时, 消费者更容易比较价格、识别隐藏费用、查看历史折扣和匹配替代平台。需求弹性上升后, 企业的局部加价率下降:

$$L(b) = \frac{1}{|\varepsilon(b)|}.$$

这部分福利可以记为

$$\Delta W^{search}(b).$$

它包括两类收益。第一是真实搜索成本下降: 消费者花更少时间查价、比价和判断质量。第二是竞争约束增强: 企业更难把低成本完全转化为利润。若搜索 agent 让消费者真正

看到全价、服务质量和退货条款，它会提高消费者剩余，也可能推动低效率企业改进服务。

但搜索 agent 也有边界。若消费者 agent 只比较标价而不比较运费、会员限制、售后质量或数据使用条款，搜索摩擦并没有完全下降。若平台用复杂价格菜单或个性化优惠干扰比较， b 的福利效果也会被削弱。因此， $\Delta W^{search}(b)$ 不应机械地写成总为正；它取决于搜索 agent 能看到什么、比较什么，以及平台是否能重新制造信息摩擦。

10.4 企业 pricing agent：价格歧视、分配与合谋风险

企业 pricing agent 的福利效果更不适合用一句“提高定价效率”概括。沿用 Section 8 的个性化报价能力 φ_i ，平台 i 对消费者类型 k 的报价可以写成

$$p_{ik}(\varphi_i) = (1 - \varphi_i)p_i + \varphi_i(v_k - \epsilon).$$

φ_i 越高，平台越接近按消费者支付意愿报价。这个机制可能提高成交匹配，也可能把消费者剩余转移给企业。若低支付意愿消费者得到更低价格，价格歧视可能扩大交易；若企业主要向高支付意愿消费者提价，消费者剩余会下降而企业利润上升。

因此，pricing agent 的福利项写为

$$\Delta W^{price}(\varphi, b),$$

其中 $\varphi = (\varphi_1, \dots, \varphi_n)$ 。它的符号不固定。消费者搜索能力 b 越强，价格歧视越难隐藏；平台数据越丰富、实验频率越高， φ_i 的分配效应越强。

重复定价博弈还会带来合谋风险。若合作高价、单期偏离和竞争惩罚状态下的单期利润分别为 π^C, π^D, π^N ，高价策略可维持的条件为

$$\delta \geq \frac{\pi^D - \pi^C}{\pi^D - \pi^N}.$$

pricing agent 可能通过更快监测、更稳定惩罚和更少人工误差降低维持高价所需的有效门槛，这一风险与算法合谋文献的担忧一致 (Calvano et al. 2020a; Calvano et al. 2019)。但它仍是条件化结果。消费者搜索 agent、算法日志、监管审计和平台透明度规则都可能削弱高价维持。

10.5 统一福利账本

定义 7 (统一福利账本 W)

把三条通道放在一起，福利分解可以写成

$$\begin{aligned} W = & CS^{prod}(\mathbf{m}) + \Pi^{op}(\mathbf{m}) - \sum_i K_i^{adopt}(a_i, z_i) \\ & + \Delta W^{search}(b) + \Delta W^{price}(\varphi, b) - E^{risk}(\mathbf{a}, \mathbf{z}, \varphi). \end{aligned}$$

这个式子是福利账本，不是新的完整一般均衡解。它把各项机制放在同一页上： $\mathbf{m}$ 来自生产端成本， K_i^{adopt} 来自固定采用和治理投入， ΔW^{search} 来自消费者搜索 agent， ΔW^{price} 来

自企业 pricing agent, E^{risk} 则汇总错误、隐私、偏见、能源消耗、责任和合谋等风险成本。

风险项不能被当作脚注处理。模型错误、隐私泄露、偏见和能源消耗会消耗真实资源或造成外部性 (Bender et al. 2021; Weidinger et al. 2022; Strubell et al. 2019)。若治理投入 z_i 能降低错误扩散参数 $\nu_i(z_i)$, 风险成本会下降; 但 z_i 自身又会提高 K_i^{adopt} 或运行成本。这就是 Section 5 中治理的经济含义: 治理既是成本, 也是风险控制投资。

表 4: 机制、变量、福利方向与政策杠杆

机制	关键变量	福利方向	政策杠杆
生产型 agent	$a_i, \beta_i, \lambda_i, f_i, m_i(a_i)$	降低边际成本可提高产量和消费者剩余; 固定采用、token 和治理成本会抵消一部分收益。	降低中小企业采用门槛, 改善模型可靠性、互操作和合规工具。
消费者搜索 agent	$b, s(b), \varepsilon(b)$	搜索摩擦下降会压缩加价空间, 并提高消费者转向低价或高质量平台的能力。	总价透明、历史价格可查、比较接口、反暗模式规则。
企业 pricing agent	φ_i, δ	可能改善匹配, 也可能扩大价格歧视或降低高价维持门槛。	算法日志、价格实验记录、合谋筛查、可审计惩罚策略。
风险治理	$z_i, \nu_i(z_i), E^{risk}$	治理提高固定和运行成本, 但可降低错误扩散、责任损失和外部性。	高风险行业审计、责任链条、隐私保护、能源与安全标准。

表 4 更像参数地图, 而非政策清单。抽象地说“支持 AI”或“限制 AI”很难对应到经济机制。更清楚的问法是: 政策改变的是 f_i 、 λ_i 、 b 、 φ_i 、 ν_i , 还是 E^{risk} ?

10.6 政策含义: 按参数干预

第一, 竞争政策应关注采用异质性。若大型平台有更高 β_i 、更低 λ_i 和更低 f_i , 生产型 agent 会扩大成本差距并提高市场集中。政策若要降低这种差距, 应优先改善中小企业的接口、模型调用条件、合规工具和互操作标准, 而不是简单补贴软件采购。

第二, 消费者政策应提高有效搜索能力 b 。总价透明、历史价格可查、配送费和会员限制可比、退换货质量可见, 这些制度会让消费者搜索 agent 真正发挥作用。若平台能用复杂菜单、隐藏费用或个性化界面重新制造摩擦, 名义上的搜索技术未必提高消费者剩余。

第三, 算法定价监管要盯住策略空间, 而不是只看是否存在人工协议。pricing agent 可以通过高频实验、自动监测和稳定惩罚改变重复博弈条件。监管工具应包括算法日志、定价实验记录、异常协同涨价筛查和对惩罚策略的可审计记录。

第四, 治理投入 z_i 应按行业风险区分。金融、医疗、法律、教育等行业中, 错误和责任成本更可能随规模凸增。治理会提高成本, 但它也可能降低 $\nu_i(z_i)$ 或 $\Omega_i(a_i, z_i)$ 。把治理只看成负担会低估其风险控制价值, 把治理写成免费保险又会高估净福利。

第五, 人力资本政策要承接任务重组。人力资本分析强调的是 routine 任务、监督任务、系统整合和判断任务的比例改变, 而非简单的“工作消失”。培训政策应帮助劳动者从重复执行转向模型评估、流程设计、数据治理、合规判断和跨部门协调。

10.7 结论、局限与未来研究

Token 与 AI agent 的经济影响可以沿着一条企业层面的传导链理解：

$$\text{任务暴露度与采用能力} \Rightarrow C_i(q_i, a_i) \Rightarrow m_i(a_i), d_i(a_i), f_i \Rightarrow q_i^*, P^*, s_i, HHI \Rightarrow W.$$

这条链把技术语言翻译成成本、产量、价格、份额和福利语言。任务是否暴露于 AI，只说明潜在影响；企业是否有能力采用，决定这种潜在影响能否进入实际使用；实际使用再通过成本、需求弹性和定价策略进入市场结构。

回到本文的核心问题：token 与 AI agent 进入企业后，如何改变成本函数并进一步影响市场竞争与福利？在本文的静态框架下，答案取决于几条力量的相对大小。AI/agent 不是单纯的边际成本下降，它同时带来人力替代与互补、token 成本、审核错误、治理投入和固定采用门槛；企业异质性会把这些成本差异转化为份额和集中度的变化；消费者搜索 agent 与企业 pricing agent 则从需求侧和策略侧分别压低加价空间、提高价格歧视和合谋风险。政策讨论也因此必须落到具体参数上：降低哪类固定成本、提高谁的搜索能力、约束哪类定价策略、治理哪类风险。

模型也有明显局限。它是静态的，未刻画企业随时间学习、积累数据和更新模型的过程；市场结构部分虽然补充讨论了价格竞争、差异化产品、进入退出和平台网络效应，但只有 Cournot 被完整求解，其他市场结构仍是比较静态扩展；案例是风格化模拟，不是实证估计；pricing agent 只用简单重复博弈条件说明合谋风险，没有求解完整动态算法博弈。

进一步研究可以沿三条路推进。第一，在企业层面区分 AI exposure、adoption 和 usage intensity，估计它们对成本和人力结构的不同影响。第二，把 token 价格、API 调用、模型可靠性、审核成本和治理投入纳入企业成本估计。第三，建立消费者搜索 agent 与企业 pricing agent 的动态模型，分析搜索摩擦下降、个性化报价和算法合谋之间的边界。

token 与 AI agent 的经济含义不只在于自动化更多任务。在本文的设定下，它们改变了企业的成本函数、消费者的信息环境和企业的策略空间；市场结构向哪个方向移动，则取决于成本节约、采用门槛、搜索摩擦、定价算法和风险治理之间的相对大小。这些判断来自一个静态理论模型，未来若能结合企业层面的成本数据和动态竞争数据，会对这些机制的相对强弱给出更精确的估计。

A 主要符号表

符号	含义
A, B	线性反需求 $P(Q) = A - BQ$ 的截距和斜率
q_i, Q	企业 i 的产量与市场总产量
p_i	企业 i 在价格竞争或差异化产品定价中的价格
MC_i	企业 i 定价时面对的边际成本
a_i	企业 i 的 AI/agent 使用强度
c_0	未采用 AI/agent 时的基准边际成本
β_i	AI/agent 对边际人工和运营成本的节约能力
β_i^L, β_i^N	劳动侧与非劳动侧的一阶净节约能力
λ_i	token、监督、错误校验和治理成本参数
$\lambda_i^L, \lambda_i^N, \lambda_i^{tok}$	劳动复杂性、非劳动运行成本与 token 曲率成本参数
f_i	固定采用、系统集成和组织转型成本参数
f_i^L, f_i^N	劳动固定采用成本与非劳动固定采用成本
$m_i(a_i)$	企业 i 的有效边际成本
$d_i(a_i)$	边际成本随产量上升的规模压力参数
Ξ_i^N	非劳动成本中的规模凸性或风险拥挤项
z_i	企业 i 的治理投入或风险控制投入
K_i^{adopt}	企业采用和治理 AI/agent 的固定资源成本
L_R, L_S	常规劳动和高技能劳动
$H(a)$	人力资本强度
b	消费者搜索 agent 的覆盖率或使用强度
s	消费者搜索摩擦，区别于市场份额 s_i
$\varepsilon_i(s)$	企业 i 面对的需求弹性
ΔW^{search}	消费者搜索 agent 带来的福利变化
ΔW^{price}	pricing agent 带来的定价与分配福利变化
E^{risk}	错误、隐私、能源、责任和合谋等风险成本
φ_i	企业 pricing agent 的个性化报价能力
δ	重复定价博弈中的贴现因子
π^C, π^D, π^N	合作高价、单期偏离和竞争惩罚状态下的单期利润
s_i	企业 i 的市场份额
HHI	Herfindahl-Hirschman 市场集中度指数
N^e	长期进入退出后的均衡企业数量
D, F	主导企业与竞争边缘企业
$Q_D(P), Q_F(P)$	主导企业剩余需求与竞争边缘供给
n_C, n_S	平台市场中消费者侧和商家侧规模

参考文献

- Acemoglu, D. (2024). “The Simple Macroeconomics of AI”. In: *Economic Policy*. DOI: [10.1093/epolic/eiae042](https://doi.org/10.1093/epolic/eiae042).
- Acemoglu, D., D. Autor, J. Hazell, and P. Restrepo (2022). “Artificial Intelligence and Jobs: Evidence from Online Vacancies”. In: *Journal of Labor Economics* 40.S1, S293–S340. DOI: [10.1086/718327](https://doi.org/10.1086/718327).
- Acemoglu, D. and P. Restrepo (2018). “The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment”. In: *American Economic Review* 108.6, pp. 1488–1542. DOI: [10.1257/aer.20160696](https://doi.org/10.1257/aer.20160696).
- (2019). “Automation and New Tasks: How Technology Displaces and Reinstates Labor”. In: *Journal of Economic Perspectives* 33.2, pp. 3–30. DOI: [10.1257/jep.33.2.3](https://doi.org/10.1257/jep.33.2.3).
- (2020). “Robots and Jobs: Evidence from US Labor Markets”. In: *Journal of Political Economy* 128.6, pp. 2188–2244. DOI: [10.1086/705716](https://doi.org/10.1086/705716).
- Aghion, P., B. F. Jones, and C. I. Jones (2017). *Artificial Intelligence and Economic Growth*. Tech. rep. 23928. National Bureau of Economic Research. DOI: [10.3386/w23928](https://doi.org/10.3386/w23928).
- Agrawal, A., J. S. Gans, and A. Goldfarb (2019). “Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction”. In: *Journal of Economic Perspectives* 33.2, pp. 31–50. DOI: [10.1257/jep.33.2.31](https://doi.org/10.1257/jep.33.2.31).
- Alekseeva, L., J. Azar, M. Gine, S. Samila, and B. Taska (2021). “The Demand for AI Skills in the Labor Market”. In: *Labour Economics* 71, p. 102002. DOI: [10.1016/j.labeco.2021.102002](https://doi.org/10.1016/j.labeco.2021.102002).
- Assad, S., R. Clark, D. Ershov, and L. Xu (2023). “Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market”. In: *Journal of Political Economy* 132.3, pp. 723–771. DOI: [10.1086/726906](https://doi.org/10.1086/726906).
- Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020). “The Fall of the Labor Share and the Rise of Superstar Firms”. In: *Quarterly Journal of Economics* 135.2, pp. 645–709. DOI: [10.1093/qje/qjaa004](https://doi.org/10.1093/qje/qjaa004).
- Autor, D. and A. Salomons (2018). *Is Automation Labor-Displacing? Productivity Growth, Employment, and the Labor Share*. Tech. rep. 24871. National Bureau of Economic Research. DOI: [10.3386/w24871](https://doi.org/10.3386/w24871).
- Autor, D. H. (2015). “Why Are There Still So Many Jobs? The History and Future of Workplace Automation”. In: *Journal of Economic Perspectives* 29.3, pp. 3–30. DOI: [10.1257/jep.29.3.3](https://doi.org/10.1257/jep.29.3.3).
- Babina, T., A. Fedyk, A. He, and J. Hodson (2023). “Artificial Intelligence, Firm Growth, and Product Innovation”. In: *Journal of Financial Economics* 151, p. 103745. DOI: [10.1016/j.jfineco.2023.103745](https://doi.org/10.1016/j.jfineco.2023.103745).

- Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). “On the Dangers of Stochastic Parrots”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. DOI: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922).
- Bommasani, R., D. A. Hudson, E. Adeli, et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv preprint. DOI: [10.48550/arXiv.2108.07258](https://doi.org/10.48550/arXiv.2108.07258).
- Brown, T. B., B. Mann, N. Ryder, et al. (2020). *Language Models are Few-Shot Learners*. arXiv preprint. DOI: [10.48550/arXiv.2005.14165](https://doi.org/10.48550/arXiv.2005.14165).
- Brown, Z. Y. and A. MacKay (2023). “Competition in Pricing Algorithms”. In: *American Economic Journal: Microeconomics* 15.2, pp. 109–156. DOI: [10.1257/mic.20210158](https://doi.org/10.1257/mic.20210158).
- Brynjolfsson, E., D. Li, and L. R. Raymond (2025). “Generative AI at Work”. In: *Quarterly Journal of Economics*. DOI: [10.1093/qje/qjae044](https://doi.org/10.1093/qje/qjae044).
- Brynjolfsson, E., T. Mitchell, and D. Rock (2018). “What Can Machines Learn and What Does It Mean for Occupations and the Economy?” In: *AEA Papers and Proceedings* 108, pp. 43–47. DOI: [10.1257/pandp.20181019](https://doi.org/10.1257/pandp.20181019).
- Brynjolfsson, E., D. Rock, and C. Syverson (2019). “Artificial Intelligence and the Modern Productivity Paradox”. In: *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press. DOI: [10.7208/chicago/9780226613475.003.0001](https://doi.org/10.7208/chicago/9780226613475.003.0001).
- Burdett, K. and K. L. Judd (1983). “Equilibrium Price Dispersion”. In: *Econometrica* 51.4, pp. 955–969. DOI: [10.2307/1912045](https://doi.org/10.2307/1912045).
- Calvano, E., G. Calzolari, V. Denicolo, and S. Pastorello (2019). “Algorithmic Pricing: What Implications for Competition Policy?” In: *Review of Industrial Organization* 55.1, pp. 155–171. DOI: [10.1007/s11151-019-09689-3](https://doi.org/10.1007/s11151-019-09689-3).
- (2020a). “Artificial Intelligence, Algorithmic Pricing, and Collusion”. In: *American Economic Review* 110.10, pp. 3267–3297. DOI: [10.1257/aer.20190623](https://doi.org/10.1257/aer.20190623).
- (2020b). “Protecting Consumers from Collusive Prices due to AI”. In: *Science* 370.6520, pp. 1040–1042. DOI: [10.1126/science.abe3796](https://doi.org/10.1126/science.abe3796).
- De Loecker, J., J. Eeckhout, and G. Unger (2020). “The Rise of Market Power and the Macroeconomic Implications”. In: *Quarterly Journal of Economics* 135.2, pp. 561–644. DOI: [10.1093/qje/qjz041](https://doi.org/10.1093/qje/qjz041).
- Dell’Acqua, F., E. McFowland, E. R. Mollick, et al. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. SSRN working paper. DOI: [10.2139/ssrn.4573321](https://doi.org/10.2139/ssrn.4573321).
- Diamond, P. A. (1971). “A Model of Price Adjustment”. In: *Journal of Economic Theory* 3.2, pp. 156–168. DOI: [10.1016/0022-0531\(71\)90013-5](https://doi.org/10.1016/0022-0531(71)90013-5).
- Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). “GPTs are GPTs: Labor Market Impact Potential of LLMs”. In: *Science* 384.6702, pp. 1306–1308. DOI: [10.1126/science.adj0998](https://doi.org/10.1126/science.adj0998).

- Farboodi, M. and L. Veldkamp (2021). *A Model of the Data Economy*. Tech. rep. 28427. National Bureau of Economic Research. DOI: [10.3386/w28427](https://doi.org/10.3386/w28427).
- Felten, E. W., M. Raj, and R. Seamans (2021). “Occupational, Industry, and Geographic Exposure to Artificial Intelligence: A Novel Dataset and Its Potential Uses”. In: *Strategic Management Journal* 42.12, pp. 2195–2217. DOI: [10.1002/smj.3286](https://doi.org/10.1002/smj.3286).
- (2023). *How will Language Modelers like ChatGPT Affect Occupations and Industries?* arXiv preprint. DOI: [10.48550/arXiv.2303.01157](https://doi.org/10.48550/arXiv.2303.01157).
- Goldfarb, A. and C. Tucker (2019). “Digital Economics”. In: *Journal of Economic Literature* 57.1, pp. 3–43. DOI: [10.1257/jel.20171452](https://doi.org/10.1257/jel.20171452).
- Graetz, G. and G. Michaels (2018). “Robots at Work”. In: *The Review of Economics and Statistics* 100.5, pp. 753–768. DOI: [10.1162/rest_a_00754](https://doi.org/10.1162/rest_a_00754).
- Hansen, K. T., K. Misra, and M. M. Pai (2021). “Frontiers: Algorithmic Collusion: Supra-competitive Prices via Independent Algorithms”. In: *Marketing Science* 40.1, pp. 1–12. DOI: [10.1287/mksc.2020.1276](https://doi.org/10.1287/mksc.2020.1276).
- Hoffmann, J., S. Borgeaud, A. Mensch, et al. (2022). *Training Compute-Optimal Large Language Models*. arXiv preprint. DOI: [10.48550/arXiv.2203.15556](https://doi.org/10.48550/arXiv.2203.15556).
- Jones, C. I. and C. Tonetti (2020). “Nonrivalry and the Economics of Data”. In: *American Economic Review* 110.9, pp. 2819–2858. DOI: [10.1257/aer.20191330](https://doi.org/10.1257/aer.20191330).
- Kaplan, J., S. McCandlish, T. Henighan, et al. (2020). *Scaling Laws for Neural Language Models*. arXiv preprint. DOI: [10.48550/arXiv.2001.08361](https://doi.org/10.48550/arXiv.2001.08361).
- Klein, T. (2021). “Autonomous Algorithmic Collusion: Q-Learning under Sequential Pricing”. In: *RAND Journal of Economics* 52.3, pp. 538–558. DOI: [10.1111/1756-2171.12383](https://doi.org/10.1111/1756-2171.12383).
- Noy, S. and W. Zhang (2023). “Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence”. In: *Science* 381.6654, pp. 187–192. DOI: [10.1126/science.adh2586](https://doi.org/10.1126/science.adh2586).
- Patterson, D., J. Gonzalez, Q. Le, et al. (2022). “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink”. In: *Computer* 55.7, pp. 18–28. DOI: [10.1109/MC.2022.3148714](https://doi.org/10.1109/MC.2022.3148714).
- Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. arXiv preprint. DOI: [10.48550/arXiv.2302.06590](https://doi.org/10.48550/arXiv.2302.06590).
- Stigler, G. J. (1961). “The Economics of Information”. In: *Journal of Political Economy* 69.3, pp. 213–225. DOI: [10.1086/258464](https://doi.org/10.1086/258464).
- Strubell, E., A. Ganesh, and A. McCallum (2019). “Energy and Policy Considerations for Deep Learning in NLP”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650. DOI: [10.18653/v1/P19-1355](https://doi.org/10.18653/v1/P19-1355).

- Varian, H. R. (2019). “Artificial Intelligence, Economics, and Industrial Organization”. In: *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press. DOI: [10.7208/chicago/9780226613475.003.0016](https://doi.org/10.7208/chicago/9780226613475.003.0016).
- Webb, M. (2019). *The Impact of Artificial Intelligence on the Labor Market*. SSRN working paper. DOI: [10.2139/ssrn.3482150](https://doi.org/10.2139/ssrn.3482150).
- Weidinger, L., J. Mellor, M. Rauh, et al. (2022). “Taxonomy of Risks Posed by Language Models”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229. DOI: [10.1145/3531146.3533088](https://doi.org/10.1145/3531146.3533088).