面向成本感知序贯决策的多大语言模型智能体贝叶斯编排

Danial Amin独立研究者
writetodanialamin@gmail.com
danial-amin.github.io

arXiv:2601.01522v1 [cs.AI] 4 Jan 2026

2026年1月6日

摘要

大语言模型日益成为自主决策智能体，应用于错误会造成可量化成本的领域：招聘（错过合格候选人与浪费面试资源）、医疗分诊（漏诊急症与不必要的升级处置），以及欺诈检测（批准欺诈交易与拒绝合法交易）。当前架构建立在有缺陷的基础之上：它们向大语言模型查询判别概率 p(状态|证据)，应用任意设定的置信度阈值，并在不考虑成本不对称性或不确定性量化的情况下执行行动。我们证明，这种方法在形式上不适用于序贯决策，并提出一种具有严谨数学依据的替代方案。

我们提出一种具有严谨数学依据的替代方案：将多个大语言模型视为近似似然函数，而非分类器。对于每一种可能的状态，我们通过对比式提示 elicitation p(证据|状态)，利用稳健统计对不同模型的结果进行聚合，并结合明确的先验应用贝叶斯定理。这种生成式建模视角带来四项关键能力：（1）随着证据累积进行适当的序贯信念更新；（2）通过期望效用最大化实现成本感知的行动选择；（3）通过信息价值计算，以有原则的方式收集信息；（4）通过集成偏差缓解提升公平性。

我们在简历筛选中实例化了这一框架：招聘失误的成本为 40,000 美元，浪费一次面试的成本为 2,500 美元，电话筛选的成本为 150 美元。我们使用五种不同的大语言模型（GPT-4o、Claude 3.5 Sonnet、Gemini Pro、Grok、DeepSeek）评估了 1,000 份简历。实验结果表明，与表现最佳的单大语言模型基线相比，我们的方法将总成本降低了 294,000 美元（改善 34%），同时使人口统计均等提高了 45%（将群体间最大差异从 22 个百分点降至 5 个百分点）。消融研究显示，多大语言模型聚合贡献了 51% 的成本节省，序贯更新贡献了 43%，分歧触发的信息收集贡献了 20%。至关重要的是，我们证明，这些收益并非仅仅是偶然的经验结果，而是纠正基于大语言模型的决策的数学基础后必然产生的结果。

1 引言

1.1 利害攸关之处：为何决策理论的严谨性至关重要

在错误会造成重大且不对称后果的各个领域，将大语言模型部署为自主决策智能体的趋势正急剧加速 [13, 119]。如今，组织经常将高风险判断委托给基于大语言模型的系统：招聘平台每年筛选数万份简历，医疗分诊系统将患者分流至资源需求和临床结果截然不同的护理路径，而欺诈检测系统则在瞬息之间做出决策，可能冻结账户或批准高额交易[70, 96, 3]。这些部署在经济和社会层面的影响深远。一个每年处理 10,000 份申请的招聘系统，其决策汇总起来涉及数百万美元的人员成本，更不用说就业机会获批或被拒对人的影响 [104]。医疗分诊系统处理的患者

每月互动量超过 50,000 次，而路由错误可能导致本可避免的死亡，或造成稀缺紧急资源的浪费 [108]。大型金融机构的欺诈系统每年审核超过 3.65 亿笔交易，即使错误率仅为 0.1%，也会导致数十万项错误决策[3]。

这些表面上各不相同的应用，其共同之处在于拥有相同的底层结构，因此可以通过统计决策理论的视角进行形式化分析 [7, 91]。在每种情况下，智能体都会观察高维证据——例如，包含教育经历、工作经历和技能信息的简历文本；涵盖病史和当前主诉的患者症状描述；以及包含地点、金额、商户和行为模式的交易元数据——但无法直接观察决定最优行动的隐藏状态变量。招聘智能体无法直接衡量候选人质量，只能根据与工作表现相关性不完美的简历信号进行推断 [61]。分诊智能体无法直接诊断疾病严重程度，只能根据症状估计紧迫程度，而症状在同一诊断下可能存在很大的方差 [38]。欺诈智能体无法直接观察犯罪意图，只能根据合法用户和非法用户部分共有的模式进行预测 [30]。

可观察证据与潜在状态之间的认知鸿沟造成了不可消除的不确定性，无法通过更好的特征工程或更大的训练集来消除。这种不确定性是问题结构的根本特征。此外，这些智能体面临的成本格局存在严重的不对称性，使基于准确率的评估框架远远不足以胜任。以招聘领域为例：拒绝一位原本符合条件且能够成为高效员工的候选人，意味着错失招聘机会，并造成可量化的成本。行业估计，错误聘用（聘用后表现不佳、必须替换的人员）的成本相当于其第一年薪资的 30% [104]。对于一名年薪 130,000 美元的中级软件工程师而言，这意味着约 40,000 美元的招聘成本、入职培训费用、生产力损失和团队运作受扰。相反，让一位不符合条件的候选人进入现场面试，会耗费工程人员时间进行面试和协调，通常每位候选人的成本为 2,500 至 3,000 美元 [18]。假阴性与假阳性之间的成本不对称达到 16 倍。

在医疗分诊中，成本不对称性更加极端，且涉及生死攸关的后果。让一名患者出院，而其随后发生心脏骤停或中风，这代表着灾难性的失败；其成本不仅以金钱衡量，更体现在可预防死亡和发病率上 [85]。漏诊心肌梗死的估计成本超过 100 万美元，计入了医疗责任、损失的生命年以及治疗成本 [72]。相反，将病情轻微的患者收入重症监护室，每天约会浪费 5,000 美元的重症护理资源，而这些资源本可用于真正的急症病例 [108]。这使某些分诊决策的成本比超过 200:1。在欺诈检测中，因欺诈警报假阳性而冻结合法客户的账户，会给机构带来约 50 美元的客户服务管理费用以及潜在的客户流失风险；而放行一笔平均金额为 2,500 美元的欺诈交易，则代表着 50 倍的成本不对称性 [2]。

这些领域具有三个共同的结构特征，使其既特别适合自动化，又对理论严谨性提出了很高要求：

1. **具有可衡量后果的离散潜在状态**。 每种决策情境都涉及对有限个底层状态的不确定性，这些状态决定了适当的行动。在招聘中，候选人分属不同的离散质量类别：缺乏基本要求、明显不符合条件的人员；值得进一步筛查的临界候选人；应进入面试环节的合格人员；以及应加急处理的杰出候选人 [112]。这些状态无法从简历中直接观察到，但会产生可衡量的事后果。聘用不符合条件的候选人会造成有据可查的成本：替换他们所需的招聘费用、团队生产力损失，以及放弃其他候选人所带来的机会成本。医疗分诊同样涉及不同的紧急程度——适合预约就诊的稳定病症、需要当天接受紧急护理的急性病症，以及需要立即干预的急症——每种情况都伴有可通过卫生经济学框架量化的死亡风险和资源需求 [34]。交易欺诈涉及一种二元状态（合法或欺诈），其财务影响明确。

2. **具有明确成本的序贯证据收集机会**。这些领域支持多阶段决策：在做出终止行动之前，可以通过附加信息补充初始观测。负责招聘的智能体在审阅简历后，可以选择先进行简短的电话筛选，再决定是否投入资源进行完整的现场面试。电话筛选约需花费 150 美元的招聘人员时间，但可能大幅降低对候选人适任性的不确定性 [103]。分诊系统可以安排诊断检查（全血细胞计数、心电图、影像学检查），以进一步评估患者病情紧急程度。每项检查都需要花费一定的金钱和时间，但也会提供信息性信号，更新人们对潜在医学状况的信念 [55]。欺诈检测系统可以要求进行额外身份验证（短信验证、安全问题、临时授权冻结），这会产生几美元的成本，但可能彻底解决模糊案例 [14]。

这些序贯机会引出的关键理论问题是：何时额外信息的期望价值足以抵偿其成本？经典决策理论通过信息价值（VOI）分析 [89, 50]给出了答案。信息价值通过计算信息可能带来的决策质量预期提升，量化智能体为信息支付的最高价格。如果当前信念表明最优行动的期望成本为 C_0，而新信息会更新信念，使期望成本变为 C_1，那么当 $C_0 - C_1 > c_{\text{info}}$ 时，获取信息就是值得的，其中 c_{info} 是信息获取成本。然而，计算信息价值需要能够预测潜在观测将如何更新信念——这项计算需要似然函数和序贯信念更新能力，而标准大语言模型架构并不具备这些能力。

3. **监管公平要求与差别影响问题**。 招聘、贷款和医疗保健领域的决策受到日益严格的监管框架约束，旨在防止算法系统延续或加剧人口统计差异 [5, 74]。纽约市第144号地方法规定，自动化就业决策工具必须接受年度偏差审计，以证明受保护类别的选择率与选择率最高的群体相差不超过四个百分点[78]。欧盟《人工智能法案》将就业和信贷系统归类为高风险应用，要求进行合格评定、实施人工监督并采取措施尽量降低歧视风险 [37]。《平等信贷机会法》禁止基于受保护属性的贷款歧视，并确立了差别影响原则：表面中立但系统性地使受保护群体处于不利地位的政策均属违法，除非有商业必要性作为正当理由 [111]。

这些法规承认，算法决策系统可能会编码并放大训练数据、模型架构或部署实践中存在的偏差 [81, 74]。在互联网文本上训练的大语言模型会从训练语料库中继承人口统计关联和刻板印象 [12, 17]。在招聘中部署时，即使资历相当，模型也可能系统性地给姓名在统计上与代表性不足的种族或性别群体相关联的简历打出较低分数 [8]。这些偏差并非表现为明确的偏见，而是表现为从有偏数据中习得的统计规律：如果训练语料库中男性工程师的示例多于女性工程师，模型的似然估计就可能系统性地低估来自女性候选人的证据。

关键的是，仅仅让模型无法接触受保护属性，并不能实现公平性。简历筛选本质上涉及与人口统计特征相关的信号：大学名称、社区地址、课外活动，甚至写作风格，都包含人口统计信息 [42]。挑战不在于消除这些相关性——那样就必须舍弃简历中大量具有信息性的内容——而在于确保它们不会转化为选择率的系统性差异。这需要模型能够量化并纠正其似然估计中的人口统计偏差；我们通过多大语言模型集成展示了这一能力。

尽管这些共同的结构特征迫切需要经过审慎的决策理论处理，当前的大语言模型智能体架构却在很大程度上忽视了不确定性下序贯决策的基本要求。主流范式将大语言模型视为输出置信度分数的黑箱分类器，对这些分数应用任意阈值来触发行动，却没有任何机制纳入成本结构、执行序贯信念更新，或量化何时有必要进行额外的信息收集。这并非轻微的工程疏忽，不能通过超参数调优或提示工程来弥补。这体现了决策问题的数学结构与所采用的计算方法之间根本性的架构不匹配。本引言余下部分将形式化阐述这些缺陷，证明它们是判别式建模范式固有的局限，并介绍我们的生成式替代方案。

1.2 标准方法及其缺陷

当前基于大语言模型的决策智能体架构看似简单，却掩盖了深刻的理论缺陷。我们对这一架构进行形式化描述，指出四项具体缺陷，并证明这些缺陷并非单纯的实现细节，而是判别式建模范式所必然导致的结果。这一不可能性证明促使我们转向生成式建模。

这种标准方法已部署于大型科技公司和初创企业的生产系统中，并遵循以下模式 [116]:

算法1 标准大语言模型智能体架构

1: **输入**简历文本 x_1 (e患者症状、交易特征) 2: **阈值：** τ3: {调整后的阈值，例如 $\tau = 7.0$}4: **查询大语言模型：** 评分 $\leftarrow$ 大语言模型(“请按 0–10 的质量分数为此输入评分：” $\|$ x) 5: **如果**评分 $\geq \tau$ **则**6: 执行高置信度行动（面试、收入ICU、批准交易）7: **否则**8: 执行低置信度行动（拒绝、出院、拒绝交易）9: **结束如果**

该架构依赖若干值得审视的隐含假设。首先，它假设大语言模型给出的评分可以被解读为对正类（符合条件的候选人、病情紧急的患者、合法交易）的置信度度量。其次，它假设固定阈值 τ（通常通过在验证数据上进行网格搜索，以最大化某项准确率指标来设定）能够恰当地平衡假阳性与假阴性。第三，它将决策视为一次性分类问题，认为必须同时考虑所有可用证据，而不可能通过连续步骤逐步完善判断。第四，它完全忽略了成本不对称性：评分为 6.9 会像评分为 2.0 一样触发拒绝，尽管前者代表临界的不确定性，或许值得进一步调查。

我们现在将证明，这些并非能够在判别式范式内修复的工程缺陷，而是源于向大语言模型索取错误概率量的固有限制。

缺陷1：无法执行序贯信念更新 假设智能体在时间 $t = 1$ 观察到初始证据 x_1 （一份简历），并获得大语言模型的评估 $p(s|x_1)$，其中 $s \in \mathcal{S}$ 表示潜在状态（候选人质量）。随后，智能体在时间 $t = 2$收集额外证据 x_2 （一份电话筛选记录）。为了正确更新信念，贝叶斯序贯更新要求计算：

$$p(s|x_1, x_2) = \frac{p(x_2|s, x_1) \cdot p(s|x_1)}{\sum_{s' \in \mathcal{S}} p(x_2|s', x_1) \cdot p(s'|x_1)} \tag{1}$$

这一表达式是贝叶斯定理在序贯证据情境下的直接应用，要求能够访问似然函数$p(x_2|s, x_1)$——即在候选人的真实质量 s 和简历 x_1给定的情况下，观察到电话筛选证据 x_2 的概率。然而，大语言模型仅提供后验 $p(s|x_1)$，这是判别模型的输出，而该模型已经将似然和先验进行了边缘化。无法访问似然函数时，我们无法计算 x_2应如何将信念从 $p(s|x_1)$更新为 $p(s|x_1, x_2)$。
常规变通方法是将拼接后的证据（x_1， x_2）再次输入大语言模型，以直接获得 p（$s|x_1$， x_2）：

$$\text{score}_{\text{new}} \leftarrow \text{LLM}(\text{“Rate this candidate given resume and phone screen”}) \tag{2}$$

这种批量推断方法将两个观测视为一个联合输入，而不是将 x_2 按序纳入信念更新。虽然这看起来像是无伤大雅的实用方案，但它存在三个关键问题，在多阶段决策过程中会变得尤为严重：

首先，批量推断会丢失有关不同观测的相对证据强度的信息。在序贯更新中，我们可以通过比较 p（$s|x_1$）与 p（$s|x_1$，x_2），量化每条证据对信念的影响程度。这样我们就能评估 x_2 是否提供了大量信息（信念变化显著），或只是冗余信息（信念变化极小）。批量推断将这些信息压缩成单一后验，从而抹去了每个观测所作出的证据贡献。

其次，批量推断难以扩展到更长的决策序列。以医学诊断过程为例，其中涉及初始症状（x_1）、生命体征（x_2）、血液检查（x_3）、影像学检查（x_4）和专科会诊（x_5）。批量推断需要将全部五项输入拼接后输入大语言模型：大语言模型（x_1， x_2， x_3， x_4， x_5）。随着证据不断累积，上下文窗口会逐渐填满，计算成本也会随序列长度线性增长。此外，如果我们想在实际获取 x_5 之前评估其是否值得付出相应成本，批量推断并未提供相应机制：在未观测到 x_5的情况下，我们无法预测 p（$s|x_1$，……， x_5）。
第三，最根本的是，批量推断无法支持对信息收集进行反事实推理。信息价值计算需要进行如下预测：“如果我观测到 x_2，我的信念会如何更新？决策质量的提升是否足以证明观测成本是值得的？”这需要对可能的观测进行边缘化：

$$\text{VOI}(x_2) = \sum_{x_2} p(x_2|x_1) \left[\max_a \mathbb{E}_{s|x_1,x_2}[U(a,s)] - \max_a \mathbb{E}_{s|x_1}[U(a,s)] \right] \tag{3}$$

计算假设的 x_2 值所需的 $p(x_2|x_1)$ 和 $p(s|x_1,\ x_2)$ 需要似然函数。判别式模型无法执行此计算，因为它们只对反向关系进行建模：$p(s|x_1)$，而不是 $p(x_2|x_1)$。

我们将这一不可能性形式化如下：

定理1（序贯更新不可能性）。设 $\mathcal{M}_{disc}$ 是一个判别模型，对于任意证据 $x \in \mathcal{X}$，都会对所有 $s \in \mathcal{S}$ 输出 $p_{\mathcal{M}}(s|x)$。假设 $\mathcal{M}_{disc}$ 无法访问生成式模型 $p(x|s)$ 或先验 $p(s)$。那么，对于观测到的 x_1 和新证据 x_2，不存在可计算函数 f 使得：

$$p(s|x_1, x_2) = f(p_{\mathcal{M}}(s|x_1), x_2) \tag{4}$$

而无需联合查询 $\mathcal{M}_{disc}$ 关于 (x_1, x_2)。

证明。假设存在这样的函数 f。那么，我们就可以仅根据 $p_{\mathcal{M}}(s|x_1)$ 和 x_2 计算 $p(s|x_1, x_2)$。根据贝叶斯定理：

$$p(s|x_1, x_2) = \frac{p(x_2|s, x_1) \cdot p(s|x_1)}{p(x_2|x_1)} \tag{5}$$

要计算这一结果，我们需要 $p(x_2|s, x_1)$ 和 $p(x_2|x_1) = \sum_{s'} p(x_2|s', x_1)p(s'|x_1)$。二者都需要似然 $p(x_2|s, x_1)$，而 $p_{\mathcal{M}}(s|x_1)$ 并未提供该似然。后验 $p_{\mathcal{M}}(s|x_1)$ 是以下内容的输出：

$$p_{\mathcal{M}}(s|x_1) \propto p(x_1|s)p(s) \tag{6}$$

但不会单独显式提供 $p(x_1|s)$ 或 $p(s)$。缺少这些组件，我们就无法构造 $p(x_2|s,\ x_1)$。因此，f 不可能存在。□

该定理表明，在判别式范式下，序贯更新并非只是困难，而是在没有额外查询的情况下不可能实现。唯一的办法是批量推断，但这会使我们无法对信息收集进行反事实推理。

缺陷2：隐藏且未校准的先验每个判别式概率 $p(s|x)$ 都会通过贝叶斯定理隐式依赖于一个先验分布 $p(s)$：

$$p(s|x) = \frac{p(x|s) \cdot p(s)}{\sum_{s' \in \mathcal{S}} p(x|s') \cdot p(s')} \tag{7}$$

当我们向大语言模型查询 $p(s|x)$ 时，模型的回答反映了从其训练语料库中隐式学习到的先验。这些训练诱导的先验可能与部署时遇到的基准率毫无关联，从而造成系统性偏见；如果无法访问底层似然函数 $p(x|s)$，这种偏见便无法纠正。

为了具体说明这一点，我们来看看招聘。一个使用互联网文本训练的大语言模型——其训练内容包括 Stack Overflow 讨论、GitHub 仓库、Hacker News 主题帖和技术博客——会从训练数据中吸收人口统计和资质分布。在互联网上，关于软件工程招聘的讨论中，合格候选人和不合格候选人大致各占一半：每有一个称赞杰出人才的帖子，就会有一个抱怨糟糕招聘的帖子。这会在模型权重中形成一种隐式先验，可能近似于 $p_{\text{train}}(\text{合格}) \approx 0.5$。这反映的是训练语料库中招聘讨论的构成，而不是实际求职者群体的构成。

然而，在部署过程中，智能体会筛选来自真实申请者渠道的简历，其中符合条件的候选人基准率要低得多。技术招聘渠道的行业数据显示，申请软件工程职位的人中，约有 5–10% 符合招聘门槛 [112, 109]。也就是说，部署中的真实先验为：

$$p_{\text{deploy}}(\text{qualified}) \approx 0.05 \text{ to } 0.10 \tag{8}$$

这代表着大语言模型权重中隐含的先验与部署环境中适用的先验之间存在 5–10 倍的差异。这种先验不匹配会造成严重且系统性的后果。贝叶斯定理表明，后验概率会按乘法方式受到先验的影响：

$$\frac{p(s=\text{qualified}|x)}{p(s=\text{unqualified}|x)} = \frac{p(x|s=\text{qualified})}{p(x|s=\text{unqualified})} \cdot \frac{p(s=\text{qualified})}{p(s=\text{unqualified})} \tag{9}$$

后验优势比是似然比（由证据决定）与先验优势比（基准率）的乘积。如果大语言模型是在 p_{train}(符合条件) $/p_{\text{train}}$(不符合条件) $= 1$ 的情况下训练的，但本应使用p_{deploy}(符合条件)$/p_{\text{deploy}}$(不符合条件) $= 0.1/0.9 \approx 0.11$，那么对于任意给定的证据 x，模型的后验优势比都会被放大约 9 倍。

考虑一份边缘性简历，在正确的部署先验下，其应产生：

$$p_{\text{correct}}(\text{qualified}|x) = 0.15 \tag{10}$$

如果大语言模型以0.5的隐式先验进行训练，并生成与证据强度相匹配的似然比，其输出后验将为：

$$p_{\text{LLM}}(\text{qualified}|x) \approx 0.60 \tag{11}$$

这种四倍膨胀会将本应进一步筛选的边缘候选人，转变为看似十拿九稳、可以快速推进至面试的候选人。系统性地高估合格率，会导致面试流程充斥着假阳性，浪费招聘资源，并因面试候选池被稀释而降低最终聘用人员的质量 [61]。

问题并不局限于招聘。在医疗分诊中，基于医学文献和患者论坛训练的大语言模型，会吸收这些来源中关于疾病患病率的先验分布。与初级医疗中的实际患病率相比，学术论文和在线健康讨论对罕见但引人注目的疾病（罕见感染、常见疾病的非典型表现）的描述比例偏高 [35]。在此类数据上训练的模型，会系统性地高估罕见严重疾病的概率，并低估常见良性疾病的概率。这会表现为转诊至急诊科的假阳性过多：肌肉拉伤患者被送往急诊科，以检查潜在的心脏事件；病毒性上呼吸道感染患者则接受肺炎评估。每一次此类过度分诊，都会浪费500至2,000美元的不必要急诊护理费用 [59]。

判别式范式没有提供纠正这一先验不匹配的机制，因为它没有将似然和先验作为相互独立的量显式呈现。如果我们能够访问 $p(x|s)$ 和隐含的 $p_{\text{train}}(s)$，就可以使用正确的部署先验重新应用贝叶斯定理：

$$p_{\text{corrected}}(s|x) = \frac{p(x|s) \cdot p_{\text{deploy}}(s)}{\sum_{s'} p(x|s') \cdot p_{\text{deploy}}(s')} \tag{12}$$

但是，判别式模型只输出组合后的后验 $p_{\text{LLM}}(s|x)$，其中已经以不可分离的方式融入了训练先验。在训练期间经过数百万次梯度更新后，先验与似然纠缠在模型的权重中。我们无法在事后“解开”这种纠缠。

一些从业者尝试通过校准技术来纠正先验不匹配，例如 Platt 校准或保序回归 [84, 126]。这些方法学习一个校准函数 g，使得 $g(p_{\text{LLM}}(s|x))$ 更好地匹配经验频率。然而，校准只能纠正跨测试分布取平均后的概率的边缘准确率；它无法纠正单个实例的条件概率。也就是说，校准确保：

$$\mathbb{E}_{x \sim \mathcal{D}_{\text{test}}}[g(p_{\text{LLM}}(s|x))] \approx \mathbb{P}(s|x, x \sim \mathcal{D}_{\text{test}}) \tag{13}$$

但并不保证 $g(p_{\text{LLM}}(s|x))$ 对任何特定 x 的准确率。对于决策，我们关心的是实例级准确性：给定这份特定的简历，这位特定候选人符合条件的概率是多少？对聚合统计数据进行校准并不能确保这一点。

此外，校准需要带标签的部署数据，这造成了先有鸡还是先有蛋的问题：要校准模型，我们需要部署它并观察结果，但部署未经校准的模型会导致有偏见的决策，可能造成伤害。在招聘中，我们需要随机面试候选人（不考虑模型的评分），以获得用于校准的无偏标签——这就违背了使用模型的初衷。在医疗分诊中，我们需要观察初始模型评分涵盖的全部范围内的患者结局，其中包括让模型错误评为低优先级的高危患者出院——这是一种不道德的实验。

我们将判别式范式下先验校正的不可能性形式化如下：

定理2（先验校正不可能性）。设 $\mathcal{M}_{disc}$ 是一个判别模型，其输出为 $p_{\mathcal{M}}(s|x)$，并隐式使用先验 $p_{train}(s)$。设 $p_{deploy}(s) \neq p_{train}(s)$ 是正确的部署先验。那么，在无法获取似然函数 $p(x|s)$ 的情况下，不存在可计算函数 h 使得：

$$h(p_{\mathcal{M}}(s|x), p_{deploy}(s)) = \frac{p(x|s) \cdot p_{deploy}(s)}{\sum_{s'} p(x|s') \cdot p_{deploy}(s')} \tag{14}$$

对任意的 x和 p_{deploy}。

证明。假设存在这样的函数 h。由 $p_{\mathcal{M}}(s|x)$，我们有：

$$p_{\mathcal{M}}(s|x) = \frac{p(x|s) \cdot p_{\text{train}}(s)}{p(x)} \tag{15}$$

其中 $p(x) = \sum_{s'} p(x|s')p_{\text{train}}(s')$。要应用 h，我们需要从 $p_{\mathcal{M}}(s|x)$ 中提取 $p(x|s)$：

$$p(x|s) = \frac{p_{\mathcal{M}}(s|x) \cdot p(x)}{p_{\text{train}}(s)} \tag{16}$$

但是， $p(x)$ 依赖于所有似然 $p(x|s')$ 和所有先验 $p_{\text{train}}(s')$。我们有：

$$p(x) = \sum_{s'} p(x|s')p_{\text{train}}(s') = \frac{1}{\sum_{s'} \frac{1}{p_{\mathcal{M}}(s'|x)/p_{\text{train}}(s')}} \tag{17}$$

这是一个循环依赖：要从 $p_{\mathcal{M}}(s|x)$中提取 $p(x|s)$，我们需要 $p(x)$，而这需要知道所有 $p(x|s')$ 的值，这些正是我们试图提取的内容。除非已知 $p_{\text{train}}(s)$（这需要模型明确报告该值，而标准大语言模型不会这么做），否则这个方程组是欠定的。即使我们知道 $p_{\text{train}}(s)$，也只能提取似然比，而无法得到绝对似然：

$$\frac{p(x|s_1)}{p(x|s_2)} = \frac{p_{\mathcal{M}}(s_1|x)/p_{\text{train}}(s_1)}{p_{\mathcal{M}}(s_2|x)/p_{\text{train}}(s_2)} \tag{18}$$

仅凭比率不足以在新的先验 p_{deploy} 下计算后验概率，因为这需要归一化后总和为 1 的概率。我们需要任意选取一个归一化常数，这会重新引入我们试图消除的先验不匹配。因此，如果没有关于模型隐式先验和似然函数的附加信息， h 就不可能存在。 □

该定理证明 先验不匹配并非能够通过事后调整来解决的校准问题 -。它是一种根本性的 架构限制，存在于判别式模型中。

缺陷 3：行动选择中缺乏成本意识 由 Wald 于 20世纪40年代创立，并经 Savage、Raiffa 等人完善的统计决策理论，建立在一个简单原则之上：应选择行动以最小化期望损失（或等价地，最大化期望效用）[113, 94, 89]。对于离散状态和行动，其形式如下：

$$a^* = \arg\min_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} p(s|x) \cdot C(a, s) \tag{19}$$

其中， $C(a, s)$ 表示在状态 s 下采取行动 a 所造成的损失，而 $p(s|x)$ 则量化了给定证据 x 时我们对各状态的不确定性。可以证明，该决策规则是最优的：对于所有可能的损失函数和先验分布 [7]，没有其他规则能够达到更低的期望损失。

要了解这为何重要，请考虑具有非对称成本的招聘决策。成本矩阵如下：

$$C(\text{action}, \text{state}) = \begin{cases} C(\text{reject}, \text{qualified}) = \$40{,}000 & \text{(false negative: lost hire)} \\ C(\text{interview}, \text{unqualified}) = \$2{,}500 & \text{(false positive: wasted interview)} \\ C(\text{reject}, \text{unqualified}) = \$0 & \text{(true negative)} \\ C(\text{interview}, \text{qualified}) = -\$100{,}000 & \text{(true positive: successful hire)} \end{cases} \tag{20}$$

真正例具有负成本（即正效用），因为招聘合格候选人能够创造价值。假阴性与假阳性之间相差 16 倍，这意味着只要满足以下条件，我们就应该面试候选人：

$$p(\text{qualified}|x) \cdot (-\$100\text{K}) + p(\text{unqualified}|x) \cdot \$2.5\text{K} < p(\text{qualified}|x) \cdot \$40\text{K} \tag{21}$$

简化（使用 p(不符合条件$|x) = 1 - p$(符合条件$|x$)）：

$$p(\text{qualified}|x) \cdot (-\$100\text{K}) + (1 - p(\text{qualified}|x)) \cdot \$2.5\text{K} < p(\text{qualified}|x) \cdot \$40\text{K} \tag{22}$$

$$-\$100\text{K} \cdot p(\text{qualified}|x) + \$2.5\text{K} - \$2.5\text{K} \cdot p(\text{qualified}|x) < \$40\text{K} \cdot p(\text{qualified}|x) \tag{23}$$

$$\$2.5\text{K} < \$142.5\text{K} \cdot p(\text{qualified}|x) \tag{24}$$

$$p(\text{qualified}|x) > \frac{\$2.5\text{K}}{\$142.5\text{K}} \approx 0.0175 \tag{25}$$

也就是说，即使候选人符合条件的概率只有 1.75%，我们也应该对其进行面试！这与基于阈值的方法截然不同，后者通常将 $\tau \approx 0.5$ 或更高的值设为阈值。0.5 的阈值不考虑成本：它将假阳性和假阴性视为同样不理想，但实际上，假阴性的成本是前者的 16 倍。

算法1中的标准大语言模型智能体架构采用固定阈值 τ，该阈值经过调整，以最大化验证数据上的准确率、F1分数或其他指标。这些指标隐含地假设成本对称：它们将假阳性和假阴性视为同样糟糕（就准确率而言），或者通过 Fβ 分数中的超参数 β 对其进行加权；选择该超参数是出于数学上的便利，而非经济现实。这在决策理论上不一致。正如 Wald 所证明的，任何忽略损失函数的决策规则都不可能是最优的，除非只是偶然（即真实损失恰好与所优化的替代指标一致） [113]。

具体来说，考虑由大语言模型评估的三份简历：

- 简历 A： p(符合条件$|x_A) = 0.72$(大语言模型评分：7.2)
- 简历 B： p(符合条件$|x_B) = 0.65$ (大语言模型评分：6.5)
- 简历 C： p(符合条件$|x_C) = 0.03$ (大语言模型评分：0.3)

设阈值为 $\tau = 0.7$，标准智能体会对 A 进行面试，拒绝 B 和 C。期望成本为：

$$\text{Cost}(A, \text{interview}) = 0.72 \cdot (-\$100\text{K}) + 0.28 \cdot \$2.5\text{K} = -\$71.3\text{K} \tag{26}$$

$$\text{Cost}(A, \text{reject}) = 0.72 \cdot \$40\text{K} + 0.28 \cdot \$0 = \$28.8\text{K} \tag{27}$$

$$\text{Cost}(B, \text{interview}) = 0.65 \cdot (-\$100\text{K}) + 0.35 \cdot \$2.5\text{K} = -\$64.1\text{K} \tag{28}$$

$$\text{Cost}(B, \text{reject}) = 0.65 \cdot \$40\text{K} + 0.35 \cdot \$0 = \$26\text{K} \tag{29}$$

$$\text{Cost}(C, \text{interview}) = 0.03 \cdot (-\$100\text{K}) + 0.97 \cdot \$2.5\text{K} = +\$0.43\text{K} \tag{30}$$

$$\text{Cost}(C, \text{reject}) = 0.03 \cdot \$40\text{K} + 0.97 \cdot \$0 = \$1.2\text{K} \tag{31}$$

对 A 的最优决策是进行面试（期望成本为 –$71.3K，即价值为 $71.3K）。对 B 的最优决策也是进行面试（期望成本为 –$64.1K），尽管其评分低于阈值。对 C 的最优决策是进行面试（期望成本为 $0.43K），因为尽管这位候选人几乎肯定不符合条件，但面试成本与潜在收益相比很低。等等——这似乎不对。我再算一遍。

实际上，对于 C，拒绝的期望成本为 $1.2K（有 3% 的概率错失一位价值 $40K 的聘用人选），而面试的期望成本为 $0.43K（有 97% 的概率浪费 $2.5K 的面试成本，减去有 3% 的概率获得 $100K 的价值 $= -0.03 \times \$100\text{K} + 0.97 \times \$2$。$5\text{K} = -\$3\text{K} + \$2$。$425\text{K} = -\$575$）。因此，面试的期望收益为 $575。其实，这样说容易让人困惑，因为我把成本和价值混在了一起。让我用一个一致的成本函数重新表述，其中数值越低越好。

我们将成本定义为纯损失（不含负值）：

- C(面试，符合条件) = $0（正确阳性：聘用后我们会赚钱，因此成本为零）
- C(面试，不符合条件) = $2，500（假阳性：浪费了面试成本）
- C(拒绝，符合条件) = $40，000（假阴性：错失聘用机会）
- C(拒绝，不符合条件) = $0 (正确阴性)

Now:

$$\text{Cost}(C, \text{interview}) = 0.03 \cdot \$0 + 0.97 \cdot \$2{,}500 = \$2{,}425 \tag{32}$$

$$\text{Cost}(C, \text{reject}) = 0.03 \cdot \$40{,}000 + 0.97 \cdot \$0 = \$1{,}200 \tag{33}$$

因此，对于合格概率仅为 3% 的简历 C，最优决策是拒绝，因为浪费 $2,500 的概率为 97%，这比承担 $ 40,000 错失招聘成本的 3% 概率影响更大。这与我们先前的计算一致：面试阈值应为 p(符合条件) > 0.0175, 而 0.03 ¿ 0.0175，所以我们实际上仍应进行面试!

我来重新计算阈值。满足以下条件时，我们会安排面试:

$$\mathbb{E}[\text{Cost(interview)}] < \mathbb{E}[\text{Cost(reject)}] \tag{34}$$

$$p \cdot 0 + (1-p) \cdot \$2{,}500 < p \cdot \$40{,}000 + (1-p) \cdot \$0 \tag{35}$$

其中 $p = p$（符合条件$|x$）。由此可得:

$$\$2{,}500 - \$2{,}500p < \$40{,}000p \tag{36}$$

$$\$2{,}500 < \$42{,}500p \tag{37}$$

$$p > \frac{\$2{,}500}{\$42{,}500} \approx 0.0588 \tag{38}$$

因此，成本最优阈值为 p（符合条件） $> 5.88\%$。对于 $p = 0.03 < 0.0588$ 的简历 C，我们应予以拒绝。对于 $p = 0.65$ 的简历 B，尽管其得分低于任意设定的 0.7 阈值，我们也绝对应安排面试。
采用阈值法 $\tau = 0.7$会对简历 B 作出错误决策：拒绝一位本应安排面试的候选人，导致预期不必要成本为 26K 美元（拒绝的成本），而安排面试的价值为 –64.1K 美元，造成了 90K 美元的损失。成千上万份简历中的这些错误会不断累积。

问题不在于阈值调整得不够好。对于成本不对称的决策，任何固定阈值都会出错，因为最优行动取决于概率以及 成本结构。成本感知智能体会根据损失函数动态调整决策边界。

缺陷4：未进行认知不确定性量化决策智能体应何时在采取行动之前收集附加信息？经典决策理论通过信息价值（VOI）分析 [89, 50]给出了明确答案。其基本原则是，只有当决策质量的预期提升超过获取信息的成本时，才应获取信息。

形式上，令 VOI(z) 表示在决策前收集某些额外观测 z 的价值。其定义如下: :

$$\text{VOI}(z) = \mathbb{E}_z \left[\min_{a \in \mathcal{A}} \mathbb{E}_{s|x,z}[C(a,s)] \right] - \min_{a \in \mathcal{A}} \mathbb{E}_{s|x}[C(a,s)] \tag{39}$$

第一项是在观测到 z之后作出最优决策的期望成本，并对 z可能取值的所有情况求边缘分布（按其在当前证据 x下的概率加权）。第二项则是仅凭当前信息作出最优决策的成本。当 VOI(z) $> c_z$ 时，获取信息才有价值；其中 c_z 是获取 z的成本。

具体来说，计算最优即时行 在招聘中，假设我们已经审阅了一份简历，并得到了后验信念 p（$s|x_1$）。我们可以 n 动的成本:

$$C_{\text{now}} = \min \left\{ \mathbb{E}_{s|x_1}[C(\text{interview}, s)], \mathbb{E}_{s|x_1}[C(\text{reject}, s)] \right\} \tag{40}$$

现在假设我们正在考虑以成本 $c_{\text{screen}} =$ $150进行电话筛选。电话筛选将产生一些证据 z （文字记录、行为观察、技术回答），并将我们的信念更新为 p($s|x_1$, z)。电话筛选的价值为:

$$\text{VOI(screen)} = \mathbb{E}_z \left[\min \left\{ \mathbb{E}_{s|x_1,z}[C(\text{interview}, s)], \mathbb{E}_{s|x_1,z}[C(\text{reject}, s)] \right\} \right] - C_{\text{now}} \tag{41}$$

如果信息价值(筛选) $>$ $150，我们就应该进行电话筛选。

计算这一信息价值需要判别式大语言模型智能体所缺乏的三种能力：

首先，我们需要在观察电话筛选之前，预测可能的电话筛选结果的分布 $p(z|x_1)$。这是一项生成式建模任务：根据简历，电话筛选可能会如何进行？例如，如果简历显示候选人拥有麻省理工学院的博士学位，并在顶级学术会议和期刊上发表过论文，我们可能会以很高的概率预测，筛选将显示出很强的技术深度。如果简历显示候选人持有训练营证书，且项目描述十分笼统，我们可能会预测其表现较弱。分布 $p(z|x_1)$需要对隐状态进行边缘化：

$$p(z|x_1) = \sum_{s \in \mathcal{S}} p(z|s, x_1)p(s|x_1) \tag{42}$$

这需要似然 $p(z|s,\ x_1)$——即电话筛选证据 z 在给定候选人质量s和简历 x_1的条件下如何分布。判别式模型无法提供这一信息。

其次，对于每种可能的电话筛选结果 z，我们需要计算它将如何更新信念：$p(s|x_1,\ z)$。正如缺陷1所述，这需要通过似然进行序贯贝叶斯更新，而判别式模型无法做到这一点。

第三，我们需要评估信息增益是否足以证明这项成本合理。这需要进行反事实比较：基于当前信息，我们的决策有多好？如果获得附加信息，决策又会有多好？若无法基于 z进行预测和更新，我们就无法进行这种比较。

标准大语言模型智能体架构没有用于信息收集决策的合理机制。一些实现干脆从不收集附加信息（只根据简历做决策），这种做法很脆弱：需要筛查才能获益的高不确定性临界案例，会被当作明确接受或拒绝的案例一视同仁。其他实现则总是收集信息（对每位候选人进行电话筛选），这种做法很浪费：仅凭简历就能筛掉的明显不合格者，也会耗费宝贵的招聘人员时间。还有一些实现采用临时启发式方法："如果评分在4到7之间，就进行电话筛选"，这无非是毫无理论依据地随意划定区间。

无法进行信息价值计算，会对序贯决策产生深远影响。在医学诊断中，临床医生分阶段收集信息：患者病史、体格检查、实验室检查、影像学检查、专科转诊。每一步的成本都逐渐增加，而要实现最佳护理，就必须根据当前诊断不确定性审慎选择要进行哪些检查 [55]。无法计算信息价值的基于大语言模型的分诊系统，要么检查过度（为所有人安排昂贵的影像学检查，推高医疗成本），要么检查不足（因检查不充分而漏诊，危及患者安全）。这两种极端都不可接受。

我们将这一局限形式化如下：

定理3(信息价值计算不可能性**)。** 设 $\mathcal{M}_{disc}$ 是一个判别模型，为当前证据 x 提供 $p_{\mathcal{M}}(s|x)$ 。设 z 为潜在的未来证据。那么，在无法访问联合生成式模型 $p(x, z|s)$ 的情况下，不存在可计算函数 g 使得：

$$\mathit{VOI}(z) = g(p_{\mathcal{M}}(s|x), c_z) \tag{43}$$

其中，c_z 是获取 z 的成本。

证明。 信息价值公式（公式39）要求计算：

$$\mathbb{E}_z \left[\min_a \mathbb{E}_{s|x,z}[C(a, s)] \right] \tag{44}$$

该期望对 z 进行边缘化，并以 $p(z|x)$加权：

$$\mathbb{E}_z[\cdot] = \sum_z p(z|x) \left[\min_a \mathbb{E}_{s|x,z}[C(a, s)] \right] \tag{45}$$

我们已经证明（定理1），仅凭 $p_{\mathcal{M}}(s|x)$，不可能计算 $p(s|x, z)$，除非用(x, z)联合重新查询模型。但信息价值计算无法用 z 重新查询模型，因为 z 是假设性的——我们尚未观察到它。信息价值的全部意义在于决定是否获取 z，而这就意味着无法观察到它。

此外，即使我们能以某种方式计算给定假设 z时的 $p(s|x,\ z)$，也需要对所有可能的 z 值进行边缘化，并以 $p(z|x)$为权重。但：

$$p(z|x) = \sum_s p(z|x, s)p(s|x) = \sum_s p(z|s, x)p(s|x) \tag{46}$$

在假设 $z \perp x|s$成立的情况下（即给定真实状态 s后，新证据 z 与初始证据 x相互独立——这在许多领域都是合理的假设）。这需要条件似然 $p(z|s, x)$，而判别模型并未提供该似然。

因此， g 无法存在：仅凭判别式后验无法计算信息价值。 □

1.3 我们的方法：从判别式建模到生成式建模

上述四项缺陷并非可以通过改进提示、增大模型或更巧妙的工程设计来修补的实现错误，而是判别式建模范式的必然结果：如果我们向大语言模型询问 $p(s|x)$，就会失去进行序贯更新、校正先验、选择成本最优行动以及计算信息价值的能力。要实现这些能力，必须能够访问似然函数 $p(x|s)$，而判别式模型并不提供这些函数。

我们的解决方案在概念上很简单，但在架构上意义深远：我们不再向大语言模型询问后验概率 $p(s|x)$，而是提取似然 $p(x|s)$，并通过贝叶斯定理自行构建后验概率。这种从判别式建模转向生成式建模的转变，将大语言模型从不透明分类器转变为更大贝叶斯决策框架中的可解释概率组件。

关键洞见在于，可以通过对比式提示来获取似然。我们不问“这位候选人有多符合条件？”（这是一个用于获取 $p(s|x)$ 的判别式问题），而是问：“假设这位候选人 [符合条件 / 不符合条件 / 处于临界水平 / 是杰出人才]。这份简历对于处于该水平的人来说有多典型？”这是一个用于获取 $p(x|s)$ 的生成式问题：在状态 s 下观察到证据 x 的概率。

具体而言，f 对于一份简历 x 和状态 $s \in$ {拒绝，电话筛选，面试，强烈建议录用}，我们向大语言模型提示 ：

> 假设候选人的真实质量为：[STATEDESCRIPTION]。以下简历对于处于这一质量等级的人来说有多典型？请按 0--10 分评分，其中10 = 表示极其典型。

将大语言模型的回答归一化为 $[0, 1]$, 后，可作为似然 $p(x|s)$ 的估计。我们针对所有状态重复这一过程，以获得似然向量 $\mathbf{L}(x) = [p(x|s_1) \ldots, p(x|s_K)]$。然后，我们结合明确的先验 $\pi(s)$（反映部署基准率）应用贝叶斯定理：

$$p(s|x) = \frac{p(x|s) \cdot \pi(s)}{\sum_{s' \in \mathcal{S}} p(x|s') \cdot \pi(s')} \tag{47}$$

这种生成式方法解锁了判别式模型所缺少的全部四项能力：

1. **序贯信念更新：** 当新证据 x_2 到来时，我们提取 $p(x_2|s)$ 并更新：

$$p(s|x_1, x_2) \propto p(x_2|s) \cdot p(x_1|s) \cdot \pi(s) \tag{48}$$

假设条件独立性 $p(x_1, x_2|s) = p(x_1|s)p(x_2|s)$，这一假设在许多领域都成立（例如，在给定真实分布的质量后，简历和电话筛选相互独立）。

2. **先验校正：** 我们可以施加任何适合部署环境的先验 $\pi(s)$，覆盖大语言模型在训练期间吸收的任何隐式先验。如果只有 5% 的申请者符合条件，我们将 π(符合条件) =0。05，并让似然通过贝叶斯定理发挥作用。

3. **成本感知的动作选择：** 借助明确的后验概率 $p(s|x)$ 和成本矩阵 $C(a, s)$，我们应用贝叶斯决策规则（公式 19），选择使期望成本最小而非使准确率最大化的行动。

4. **信息价值计算：**我们可以通过以下方式预测 $p(z|x)$：

$$p(z|x) = \sum_s p(z|s)p(s|x) \tag{49}$$

计算如何 z 会通过 $p(s|x, z) \propto p(z|s)p(s|x)$ 更新信念，并评估预期的决策 - 改进 是否足以证明信息成本合理。

此外，我们的框架也能自然地支持多大语言模型集成。不同的大语言模型（GPT-4、Claude、Gemini）拥有不同的训练语料库、架构和归纳偏置。对于给定的证据—状态对（x, s），它们的似然估计可能存在分歧。我们使用稳健统计方法聚合这些不同的估计值，具体而言，采用中位数

或截尾均值，以减轻个体模型偏差并提高整体校准水平 [32]。如果五个大语言模型为 $[0.2, 0.3, 0.8, 0.25, 0.28]$ p（$x|s =$ 符合条件）提供似然估计，中位数 0.28 比均值 0.37 更稳健，因为后者受到离群值 0.8 的影响而偏高。

集成聚合还带来另外两项益处。首先，它能提升公平性：如果各个大语言模型表现出人口统计偏差（例如，GPT-4 会系统性地给姓名与代表性不足群体相关联的简历打低分），那么只要模型的错误并非完全相关，集成平均就能减少这些偏差 [81]。我们的实证结果表明，与单大语言模型基线相比，我们的多大语言模型方法将人口统计均等差距缩小了 45%。其次，集成分歧可以量化认知不确定性：当大语言模型之间存在较大分歧时，我们的不确定性较高，因此应收集更多信息。我们通过一种由分歧触发的电话筛选策略将这一做法形式化：如果似然估计的变异系数超过阈值，我们就会在作出决定前自动要求进行电话筛选。

本文余下部分将对这一框架进行形式化阐述，证明其最优性性质，并通过一项涉及 1,000 份简历和五个不同大语言模型的招聘任务，实证展示其有效性。第2节介绍数学框架，第**??**节详述实现方式，第3节报告实验结果，第5节综述相关工作，并由第**??**节作结。

2 方法论

我们现在严谨地构建数学框架，证明判别式方法的不可能性结果，并确立生成式替代方案的最优性性质。本节分为三个部分：（1）形式化问题规范与不可能性定理；（2）我们的生成式框架及其完整算法规范；（3）简历筛选中的实验实例化。

2.1 框架：数学基础

2.1.1 问题表述

一个序贯决策　　不确定性下的问题由六个形式化组件构成，它们完整地规定了决策　　n 环境：

定义 1（序贯决策问题）。 序贯决策问题是一个元组 $(\mathcal{S}, \mathcal{X}, \mathcal{A}, C, \pi, \{p(x|s)\})$，其中：

- $\mathcal{S} = \{s_1, \ldots, s_K\}$：表示离散潜在状态的有限状态空间
- $\mathcal{X}$：包含所有可能证据类型的观测空间
- $\mathcal{A} = \{a_1, \ldots, a_J\}$：包括信息收集行动和终止行动的有限动作空间
- C：　$\mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$：将（行动，状态）对映射为非负实数成本的成本函数
- $\pi \in \Delta(\mathcal{S})$：状态的先验分布（其中 $\Delta(\mathcal{S})$是概率单纯形）
- $\{p(x|s)\}_{s \in \mathcal{S}}$：似然函数族，用于指定在给定各状态的条件下观测的分布

对于招聘领域，可具体表示为：

- $\mathcal{S} = \{s_{\text{拒绝}}\ \ s_{\text{筛选}}\ \ s_{\text{面试}}\ \ s_{\text{强候选人}}\}$（四个候选人质量等级）
- $\mathcal{X}$ 包括简历文本文件、电话筛选记录、编程测试结果
- $\mathcal{A} = \{a_{\text{拒绝}}\ \ a_{\text{筛选}}\ \ a_{\text{面试}}\}$（拒绝、收集信息或安排面试）
- C 体现了非对称招聘成本（详见表1）
- $\pi = [0.65, 0.25, 0.08, 0.02]$ 基于实证招聘漏斗数据
- $\{p(x|s)\}$ 是在给定质量条件下，关于简历特征的似然函数

决策序列协议。智能体通过迭代循环运行：

1. **观测阶段：** 在时间 t，观测证据 $x_t \in \mathcal{X}$2. **信念更新阶段：** 更新各状态的后验信念：　$b_t(s) \leftarrow f_{\text{update}}(b_{t-1},\ x_t)$

3. **行动选择阶段：** 选择行动： $a_t \leftarrow \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s \sim b_t}[C(a, s)]$4. **终止检查：** 如果 a_t 处于终止状态（拒绝或面试），则停止；否则收集更多证据并递增 t

最优性 **性准则**。 智能体寻求一个决策策略 δ：$\mathcal{X}^* \to \mathcal{A}$，以最小化预期总成本 :

$$\delta^* = \arg\min_{\delta} \mathbb{E}_{s \sim \pi, \{x_t\}_{t=1}^{T} \sim p(\cdot|s)} \left[\sum_{t=1}^{T} C(\delta(x_{1:t}), s) + c_t \cdot \mathbb{1}[\text{info gathered at } t] \right] \tag{50}$$

其中， c_t 是时间 t 收集信息的成本，期望是针对状态与观测序列的联合分布计算的。

2.1.2 判别式方法的不可能性定理

现在，我们将确立四项基本的不可能性结果，证明判别式大语言模型架构无法支持序贯决策所需的能力。所有证明均见附录 A。

定理 4 (序贯更新不可能性**)**。 设 $\mathcal{M}_{disc}$ 是一个判别模型，它将任意观测 $x \in \mathcal{X}$ 映射为状态 $s \in \mathcal{S}$ 上的概率分布 $p_{\mathcal{M}}(s|x)$，但不会单独呈现似然函数 $p(x|s)$ 或先验 $p(s)$。那么，对于已观测证据 x_1 和新证据 x_2，不存在可计算函数 f：$\Delta(\mathcal{S}) \times \mathcal{X} \to \Delta(\mathcal{S})$ 使得：

$$p(s|x_1, x_2) = f(p_{\mathcal{M}}(s|x_1), x_2) \quad \forall s \in \mathcal{S}, x_1, x_2 \in \mathcal{X} \tag{51}$$

而无需针对联合观测 (x_1, x_2) 重新查询 $\mathcal{M}_{disc}$。

证明概要。 **(**完整证明见附录 A。**)** 序贯贝叶斯更新要求：

$$p(s|x_1, x_2) = \frac{p(x_2|s, x_1) \cdot p(s|x_1)}{\sum_{s'} p(x_2|s', x_1) \cdot p(s'|x_1)} \tag{52}$$

要计算此值，我们需要 $p(x_2|s, x_1)$。但是， $p_{\mathcal{M}}(s|x_1) \propto p(x_1|s)p(s)$ 已在模型的训练过程中对 $p(x_1|s)$ 和 $p(s)$ 进行了边缘化。由于无法分别访问这些组件，我们无法构建更新所需的前向模型 $p(x_2|s, x_1)$。仅凭后验 $p_{\mathcal{M}}(s|x_1)$ 所包含的信息不足以完成此计算。 □

结果： 判别式智能体只能通过批量重新查询来更新信念： $p_{\mathcal{M}}(s|x_1, x_2)$。这使得（a）在获取信息之前对信息收集进行反事实推理，以及（b）在观测依次到达时进行高效的多阶段推断，都无法实现。

定理 *5*（先验校正不可能性）。设 $\mathcal{M}_{disc}$ 输出后验 $p_{\mathcal{M}}(s|x)$，其权重中嵌入了隐式训练先验 $p_{train}(s)$。设 $p_{deploy}(s) \neq p_{train}(s)$ 表示正确的部署先验。那么，在无法显式获取似然 $p(x|s)$ 且不知道 $p_{train}(s)$ 的情况下，不存在函数 $h: \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$，使得：

$$h(p_{\mathcal{M}}(s|x), p_{deploy}(s)) = \frac{p(x|s) \cdot p_{deploy}(s)}{\sum_{s' \in \mathcal{S}} p(x|s') \cdot p_{deploy}(s')} \quad \forall s \in \mathcal{S}, x \in \mathcal{X} \tag{53}$$

证明概要。 要使用新的先验应用贝叶斯定理，我们需要从以下内容中提取 $p(x|s)$：

$$p_{\mathcal{M}}(s|x) = \frac{p(x|s) \cdot p_{\text{train}}(s)}{p(x)} = \frac{p(x|s) \cdot p_{\text{train}}(s)}{\sum_{s'} p(x|s') p_{\text{train}}(s')} \tag{54}$$

要解出 $p(x|s)$，需要同时知道 $p_{\text{train}}(s)$ 和归一化常数 $p(x)$。但 $p(x)$ 本身取决于所有似然 $\{p(x|s')\}_{s'}$，从而形成循环依赖。即使已知 $p_{\text{train}}(s)$，我们也只能还原似然比值：

$$\frac{p(x|s_i)}{p(x|s_j)} = \frac{p_{\mathcal{M}}(s_i|x)/p_{\text{train}}(s_i)}{p_{\mathcal{M}}(s_j|x)/p_{\text{train}}(s_j)} \tag{55}$$

仅凭比值无法在新的先验下计算归一化后验概率；我们需要绝对似然或归一化常数，而判别式模型两者都无法提供。 □

后果：当部署基准率与训练数据分布不同时（实践中很常见），判别式模型会产生系统性偏差的预测结果，且无法通过事后处理予以纠正。

定理6 (成本无关阈值次优性**)。** 对于任意非对称成本矩阵 C： $\mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ 其中有$C(a_i, s_j) \neq C(a_k, s_\ell)$ 适用于某些状态对，贝叶斯最优动作选择规则为：

$$a^*(x) = \arg\min_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} p(s|x) \cdot C(a, s) \tag{56}$$

对于所有 x 和所有成本结构，任何形式为"如果分数$(x) \geq \tau$ 则 a_1，否则 a_2"的基于阈值的规则都不可能是最优的，因为决策边界取决于 $p(s|x)$ 和成本矩阵 C。

证明概要。 考虑二元动作 $\mathcal{A} = \{a_1, a_2\}$ 和二元状态 $\mathcal{S} = \{s_1, s_2\}$。期望成本为：

$$\mathbb{E}[C(a_1, s)] = p(s_1|x)C(a_1, s_1) + p(s_2|x)C(a_1, s_2) \tag{57}$$

$$\mathbb{E}[C(a_2, s)] = p(s_1|x)C(a_2, s_1) + p(s_2|x)C(a_2, s_2) \tag{58}$$

当 $\mathbb{E}[C(a_1, s)] < \mathbb{E}[C(a_2, s)]$, 时，我们应选择 a_1，化简后得到：

$$p(s_2|x) > \frac{C(a_2, s_1) - C(a_1, s_1)}{[C(a_1, s_2) - C(a_2, s_2)] + [C(a_2, s_1) - C(a_1, s_1)]} \tag{59}$$

决策阈值取决于全部四个成本值。不同的成本矩阵会产生不同的阈值。只有在成本满足特定关系时，固定阈值（例如 $p(s_2|x) > 0.5$）才是最优的；一般情况下并非如此。 □

结果：针对准确率进行优化的基于阈值的系统（其假设成本对称）在部署到成本不对称的环境中时，会做出次优决策。

定理 7 （价值信息计算不可能性）。 在做出决策前收集额外证据 $z \in \mathcal{X}$ 的价值为：

$$VOI(z|x) = \mathbb{E}_{z \sim p(\cdot|x)} \left[\min_{a \in \mathcal{A}} \sum p(s|x, z)C(a, s) \right] - \min_{a \in \mathcal{A}} \sum p(s|x)C(a, s) \tag{60}$$

计算该价值需要：*(1)* 预测分布 $p(z|x) = \sum_s p(z|s)\, p(s|x)$，以及 *(2)* 更新后的后验分布 $p(s|x, z) \propto p(z|s)\, p(s|x)$。两者都需要似然函数 $\{p(z|s)\}_s$，而判别式模型并不提供这些函数。

结果：判别式智能体没有判断何时收集附加信息的原则性依据。它们只能依赖临时启发式方法（始终/从不收集信息，或在置信度处于某个任意范围时收集信息）。

这四个定理共同证明，判别式架构在最优序贯决策方面在数学上并不充分。这些缺陷并非工程局限，而是形式上的不可能性。这促使我们转向生成式建模这一范式。

2.1.3 我们的生成式框架

现在，我们介绍我们的框架。该框架通过从大语言模型中 elicitation 似然函数，而非判别式后验概率，解决了上述四个不可能性问题。

核心洞见：通过对比式提示进行似然 elicitation 这一基本创新是一种反转条件概率方向的提示策略。我们不再询问"给定这一观测，状态是什么？"（$p(s|x)$），而是询问"给定这一状态，这一观测有多典型？"（$p(x|s)$）。

定义 2 (对比式似然提示模板**)。** 对于观测 $x \in \mathcal{X}$、状态 $s \in \mathcal{S}$ 和大语言模型 $\mathcal{M}_m$，对比式似然提示词具有如下结构：

```
[ROLE SPECIFICATION]: You are an expert in [DOMAIN] with experience evaluating [OBSERVATION
TYPE].[STATE CONDITIONING]: Assume that the true underlying state is: [DETAILED DESCRIPTION OF
STATE s][GENERATIVE QUESTION]: Given this assumed state, how typical or representative is the
following observation?[OBSERVATION]: [Full text of observation x]
```

```
[SCORING RUBRIC]:
Provide a score from 0 to 10 where:

        10 = Extremely typical (exactly what we'd expect from this state)

        7-9 = Quite typical (strong match with minor variations)

        4-6 = Somewhat typical (plausible but imperfect fit)

        1-3 = Atypical (doesn't match this state well)

        0 = Completely atypical (inconsistent with this state)

[OUTPUT FORMAT]: Provide only the numeric score.
```

该模板体现了三项关键设计原则：

原则 1：显式状态条件化。 提示词首先将模型的推理锚定在一个假设状态上（“假设真实状态是.....”），迫使模型进行正向模拟（$s \to x$），而非反向推断（$x \to s$）。这会促使模型发挥其生成能力：给定质量等级 s，我们会观测到哪些简历特征 x？

原则 2：典型性框架。 我们询问的是“典型性”或“代表性”，而非质量、匹配度或分类。这种语言框架与似然估计相契合：$p(x|s)$量化了在状态为 s 这一假设下，观测 x 出现的可能性。从心理学角度看，相较于直接进行概率估计，典型性判断对人类（以及由此推及在基于人类文本训练的模型）来说更自然。

原则 3：逐状态独立 elicitation。 我们针对每个状态 $s \in \mathcal{S}$ 分别查询模型，获得 $\hat{}$ $\hat{}$ 独立估计 $\{L_m(x|s_1) \ldots, L_m(x|s_K)\}$。这样可以避免判别式提示中隐含的归一化，以及各状态之间的竞争；这类提示会要求模型给出单一分类。

响应归一化。 模型返回一个原始分数 $r \in [0, 10]$。我们将其归一化到概率尺度：

$$\hat{L}_m(x|s) = \frac{r}{10} \in [0, 1] \tag{61}$$

由于贝叶斯推断只需要似然在比例常数意义下的值（归一化常数 $p(x)$ 在贝叶斯定理中计算），因此这种线性归一化已足够。比值 $L_m(x|s_i)/L_m(x|s_j)$ 保留了状态之间的相对似然。

多大语言模型集成 聚合 单个大语言模型会因训练数据、架构 l
选择和对齐 t 程序 [13, 12, 81]。已记录的偏差包括：

- **GPT-4o：** 精英大学偏差（即使其他资历 ls
  是等价的）
- **Claude 3.5 Sonnet：** 冗长偏差（偏好项目描述更长、更详细的候选人）
- **Gemini Pro：** 近因偏差（相较于较早但相关的成就，过度重视近期工作经历）
- **Grok：** 初创公司偏差（高估早期阶段公司的工作经历）
- **DeepSeek：** 学术出版偏差（对研究论文和 arXiv 贡献赋予了过高权重）

为减轻这些个体偏差，我们使用稳健统计方法，对 M 多种不同模型组成的集成的似然估计进行聚合。

通过中位数进行聚合。 对于观测 x 和状态 s，我们收集来自 M 个模型的似然估计：

$$\mathcal{L}(x, s) = \{\hat{L}_1(x|s), \hat{L}_2(x|s), \ldots, \hat{L}_M(x|s)\} \tag{62}$$

并使用样本中位数进行聚合：

$$L(x|s) = \text{median}(\mathcal{L}(x, s)) \tag{63}$$

为何选择中位数而非均值？ 中位数具有更优的稳健性：

1. **异常值抵抗性：** 如果一个模型给出异常估计（例如，GPT-4o 将 $\hat{}L(x|s_{\text{strong}}) = 0.9$ 赋值给 ≈ 0.3，而其他模型给出的值是 ≈ 0.3），均值会偏移至 $(0.9 + 0.3 + 0.3 + 0.3 + 0.3)/5 = 0.42$，误差达 40%。中位数仍为 0。3，不受离群值影响。

2. **崩溃点高：** 中位数的崩溃点为 50%：多达 $\lfloor M/2 \rfloor$个模型可以返回任意值，而不会对聚合结果造成灾难性影响。均值的崩溃点为 0：单个极端离群值就能任意改变结果。

3. **等变性：** 中位数在单调变换下具有等变性。若对所有估计值应用单调变换 g，则 median($g(L_1)\ldots$, $g(L_M)$) $= g$(median($L_1, \ldots, L_M$))。

我们将稳健性性质形式化如下：

定理 8 (中位数误差界)。 设 $\{\hat{}L_m(x|s)\}_{m=1}^M$ 为来自 M 个模型的似然估计值，且 $L^*(x|s)\hat{}$表示真实似然。定义个体误差 $\epsilon_m = |L_m(x|s) - L^*(x|s)|$。则：

$$\left| median_{m=1}^M \hat{L}_m(x|s) - L^*(x|s) \right| \leq median_{m=1}^M \epsilon_m \tag{64}$$

也就是说，中位数聚合值的误差不超过个体误差的中位数。

证明见附录 A。 该界限保证：如果中位数模型的错误率较低，那么聚合结果的错误率也较低，无论 s 表现最差的模型有多 差。

为实现多样性的模型选择。 我们有意选择在多个维度上各不相同的模型，以尽量减少相关误差：

- **训练语料库：** 互联网文本（GPT-4o）、精选高质量文本（Claude）、学术论文和书籍（Gemini）、多语言数据（DeepSeek）、实时社交媒体（Grok）
- **架构：** 稠密型 Transformer（GPT-4o、Claude）、稀疏混合专家（拥有 8 位专家的 Grok-1、拥有 160 位专家的 DeepSeek-V2）、不同的上下文长度（8K–200K 词元）
- **对齐：** 在不同偏好数据集上训练的不同基于人类反馈的强化学习奖励模型，具有不同的安全约束
- **发布时间线：** 相隔 6–12 个月发布的模型，反映了不断演变的数据分布

多样性至关重要：如果所有模型都具有相同的偏见（例如，都低估非英语母语者），聚合便毫无益处。我们的实证结果（第3节）表明，与任何单个模型相比，多样化集成能显著减少偏见。

序贯贝叶斯信念更新 给定观测 x 的聚合似然 $\{L(x|s)\}_{s\in\mathcal{S}}$ 和明确的先验 $\pi \in \Delta(\mathcal{S})$，我们执行贝叶斯推断。

Init信念更新。 观测到第一条证据 x_1后，我们通过贝叶斯定理计算后验 ：

$$b_1(s) = \frac{L(x_1|s) \cdot \pi(s)}{\sum_{s' \in \mathcal{S}} L(x_1|s') \cdot \pi(s')} = \frac{L(x_1|s) \cdot \pi(s)}{Z_1} \tag{65}$$

其中，$Z_1 = \sum_{s'} L(x_1|s')\pi(s')$ 是归一化常数，用于确保 $\sum_s b_1(s) = 1$。

序贯更新。F 对于后续观测 x_2，x_3，.....，我们通过将先前的信念作为新的后验，来依次更新信念

$$b_t(s) = \frac{L(x_t|s) \cdot b_{t-1}(s)}{\sum_{s' \in \mathcal{S}} L(x_t|s') \cdot b_{t-1}(s')} = \frac{L(x_t|s) \cdot b_{t-1}(s)}{Z_t} \tag{66}$$

其中， Z_t 是时间 t的归一化常数。

这一序贯过程依赖于贝叶斯滤波中的一个标准假设：

假设1 (观测的条件独立性)。 给定真实状态 s，观测之间条件独立：

$$p(x_1, x_2, \ldots, x_T|s) = \prod_{t=1}^{T} p(x_t|s) \tag{67}$$

等价地，对所有 t， $p\left(x_t|s, x_1, \ldots, x_{t-1}\right) = p\left(x_t|s\right)$ 。

合理性与可信度。 当观测是在给定潜在状态的条件下由相互独立的过程生成时，这一假设成立：

• **招聘：** 候选人的简历质量和电话筛选表现都由其真实技能水平决定，但分别通过相互独立的过程产生（撰写简历与回答技术问题）。给定技能水平后，这些因素之间不相关。

• **医学诊断：** 患者的症状和实验室检测结果均由潜在疾病引起，但症状表现与检测中测得的生化标志物是相互独立的过程。
• **欺诈检测：** 交易地点和交易金额均受交易是否属于欺诈的影响，但在给定欺诈状态的条件下，地点和金额是独立选择的。

当假设1成立时，序贯推断与批量推断等价：
定理9 (序贯更新与批量更新的等价性)。 在假设*1*下，序贯信念更新程序（方程 *66*）得出的后验与批量贝叶斯推断得出的后验相同：

$$b_T(s) = \frac{\pi(s)\prod_{t=1}^{T} L(x_t|s)}{\sum_{s'\in\mathcal{S}} \pi(s')\prod_{t=1}^{T} L(x_t|s')} = p(s|x_1, x_2, \ldots, x_T) \tag{68}$$

其中，右侧是基于所有观测联合计算出的后验。

证明见附录 A。该定理保证了我们序贯程序的数学正确性。
先验指定与领域自适应。 与先验不透明且无法访问的判别式模型不同，我们的框架要求明确指定先验 $\pi \in \Delta(\mathcal{S})$。我们根据历史部署数据中的经验基准率设定 π。

在招聘中，如果机构数据表明，历史上：

- 65% 的申请者缺乏基本资格（状态 s_1）
- 25% 值得进行电话筛选（状态 s_2）
- 8% 值得进行直接面试（状态 s_3）
- 2% 是杰出的优秀录用人选（状态 s_4）

我们设定 $\pi = [0.65, 0.25, 0.08, 0.02]$。
这一明确的先验指定使以下操作成为可能：

1. **领域自适应：** 具有不同求职者群体的组织可以使用不同的先验 2. **时间自适应：** 可以随着招聘漏斗指标的变化更新先验 3. **亚组分析：** 针对不同工作岗位或职级分别设定先验 4. **纠正训练与部署不匹配：** 用特定于部署的基准率覆盖大语言模型的训练先验

成本感知的动作选择 在每个决策点 t 信念更新之后 b_t，智能体都会选择使期望成本最小的行动：

$$a_t^* = \arg\min_{a\in\mathcal{A}} \mathbb{E}_{s\sim b_t}[C(a,s)] = \arg\min_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} b_t(s)\cdot C(a,s) \tag{69}$$

这是贝叶斯决策规则；可以证明，在真实信念下，该规则是最优的：
定理10 （期望成本最小化的最优性）。 令 $p^*(s|x)$ 表示给定观测 x时状态的真实后验分布。在所有将观测映射到行动的确定性决策规则 $\delta: \mathcal{X} \to \mathcal{A}$ 中，贝叶斯规则：

$$\delta^*(x) = \arg\min_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} p^*(s|x)\cdot C(a,s) \tag{70}$$

使期望成本（贝叶斯风险）最小：

$$R(\delta^*) = \int_{\mathcal{X}} \sum_{s\in\mathcal{S}} p^*(s|x)\cdot C(\delta^*(x),s)\cdot p(x)\,dx \le R(\delta) \quad \forall\delta \tag{71}$$

表1：成本　软件工程招聘的成本矩阵（美元）。行表示 = 行动，列表示 = 候选人的真实状态。

行动	s_1 （拒绝）	s_2 （筛选）	s_3 (面试)	s_4 (强候选人)
拒绝	$0	$5,000	$20,000	$40,000
电话筛选	$150	$150	$150	$150
面试	$2,500	$2,500	$0	$0

证明见附录 A（由统计决策理论中的经典结果得出 [113, 7]）。

成本矩阵规范。 成本函数 $C\colon \mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ 必须基于领域经济学进行指定。对于招聘：

理由和数据来源：

- **拒绝合格候选人：** 错失聘用机会的机会成本。对于 s_2 （边际贡献者），因生产力损失估计为 5,000 美元。对于 s_3 （优秀工程师），根据行业单次招聘成本和首年生产力估算为 20,000 美元。对于 s_4 （杰出人才），考虑到其稀缺性和影响，估算为 40,000 美元。来源：SHRM 2016 单次招聘成本调查（中位数为 4,129 美元） [104], ，并根据职位影响进行调整。

- **电话筛选：** 固定成本 $150 = （0.5 小时准备时间 + 0，0.5 小时通话时间 + 0，0.25 小时记录时间） ×，招聘人员全负荷成本为每小时 120 美元。
- **面试不符合条件的候选人：** 2,500 = （4 名面试官 × 1 ，工程师全负荷成本为每小时 600 美元）+ ，协调费用 100 美元 + ，候选人差旅及餐饮费用 400 美元。来源：CareerBuilder 2017 年面试成本数据 [18]。
- **面试合格候选人：** 成本为 0（基准值，因为这是期望结果）。实际上仍有 2,500 美元的成本，但我们将其归一化，使正确决策的成本为零。

敏感性分析。 为评估成本设定不确定性下的稳健性，我们将所有成本扰动 ±20%，并衡量最优行动发生变化的频率。结果（第3节 3.5）：仅有 7.8% 的候选人会在 ±20% 的扰动下改变决策，表明该框架能够稳健应对合理的成本估算误差。

信息收集的信息价值 该框架必须决定：在做出终止决策（面试或拒绝）之前，何时应收集额外证据（例如进行电话筛选）？ 经典决策理论通过信息价值（VOI）分析 [89, 50] 来回答这一问题。

定义 3 (信息价值)。 设 $x \in \mathcal{X}$ 表示当前可用的证据，由此形成信念 $b(s) = p(s|x)$。收集额外证据的价值 $z \in \mathcal{X}$ 为：

$$\mathit{VOI}(z|x) = \mathbb{E}_{z \sim p(\cdot|x)}\left[V(x,z)\right] - V(x) \tag{72}$$

其中， $V(x) = \min_{a \in \mathcal{A}} \sum_s p(s|x) C(a,s)$ 表示当前信息下最佳行动的价值（负成本），而 $V(x,z) = \min_{a \in \mathcal{A}} \sum_s p(s|x,z) C(a,s)$ 表示观察到 z 后的价值。展开后：

$$\mathit{VOI}(z|x) = \mathbb{E}_z\left[\min_a \sum_s p(s|x,z) C(a,s)\right] - \min_a \sum_s p(s|x) C(a,s) \tag{73}$$

当 VOI$(z|x) > c_z$ 时，值得获取信息，其中 c_z 是收集 z 的成本。

精确计算难题。 计算公式 73 需要：

1. 对可能的未来观测进行边缘化： $z \sim p(z|x) = \sum_s p(z|s)p(s|x)$ 2. 对每个可能的 z，计算更新后的信念： $p(s|x,\ z) \propto p(z|s)p(s|x)$ 3. 根据更新后的信念评估最优行动成本 4. 对所有 z取平均

对于连续空间或大型离散空间，这种边缘化在计算上难以承受（多步规划时，计算量随序列长度呈指数增长）。因此，我们采用一种近似方法。

实用的VOI近似。 我们将信息收集建模为对真实状态的部分揭示：

- 以概率 $\rho \in [0,1]$，新证据 z 能够确凿地识别真实状态

• 以概率 $1-\rho$，证据不包含信息（信念保持 b不变）

在此二元信息性模型下：

$$\text{VOI}(z|x) \approx \rho \cdot \left[\min_{a \in \mathcal{A}} \sum_s b(s) C(a,s) - \sum_s b(s) \min_{a \in \mathcal{A}} C(a,s) \right] \tag{74}$$

第一项是在当前信念下采取最佳行动的成本（对各状态的不确定性取平均）。第二项是在完全信息下的成本（为每种状态选择最优行动，然后取平均）。两者之差就是消除全部不确定性的价值，再乘以 ρ（信息实际消除不确定性的概率）。

设定 ρ （信息性参数）。我们根据实证估计值对 ρ 进行校准：

• **电话筛选：** $\rho \approx 0.7$ （电话筛选结果与最终聘用成功之间的历史相关性为 $r = 0.68$）

• **编程测试：** $\rho \approx 0.85$ （对技术能力的预测能力更强）

• **背景调查：** $\rho \approx 0.6$ （由于参考对象中的选择偏差，信号较弱）

ρ 可通过最大似然从带标签数据中学习：找出 ρ^* ，使观测到的信息收集结果的似然最大。

决策规则。当且仅当满足以下条件时，我们才收集信息 z：

$$\text{VOI}(z|x) > c_z \tag{75}$$

这确保了信息收集在经济上合理：预期决策改进超过信息成本。

基于分歧的不确定性检测除了信息价值计算外，我们还将模型间分歧作为认知不确定性的信号，表明需要进一步审查。

分歧符合　　理性。 对于观测 x 和状态 s，我们计算各模型之间的变异系数（CV）　　：

$$D(x,s) = \frac{\sigma_m(\hat{L}_m(x|s))}{\mu_m(\hat{L}_m(x|s))} \tag{76}$$

其中， σ_m 是标准差， μ_m 是 M 似然估计值的均值。
CV 具有尺度不变性：它衡量相对于均值归一化后的相对离散程度，因此可用于比较不同状态和观测。

高分歧阈值。 当满足以下条件时，我们将案例标记为高不确定性：

$$\max_{s \in \mathcal{S}} D(x,s) > \tau_D \tag{77}$$

阈值为 τ_D （通过交叉验证设为 0.15）。
分歧触发的信息收集。当分歧超过 τ_D时，我们：

1. 计算可用信息来源（电话筛选、编程测试等）的信息价值 2. 如果信息价值 $> c_{\text{info}}$，则收集信息，即使当前信念并非处于临界状态 3. 将分歧记录为质量信号（有助于识别需要人工审核的案例）

根据实证结果（第3节），27% 的简历会引发严重分歧，对这些案例收集信息，相较于始终收集或从不收集信息，可节省 5 万美元的净成本。

2.2 算法：CompleteSpecification

现在，我们将所有组件整合为一个统一的算法过程。
计算复杂度分析。每次迭代需要：

• **似然 elicitation（步骤 1）：** $O(MK \cdot T_{\text{大语言模型}})$ ，其中 M 为集成规模， $K = |\mathcal{S}|$ 为状态数量， $T_{\text{大语言模型}}$ 为大语言模型查询延迟（通常为 0.5–2 秒）

- **聚合（步骤 2）：** $O(MK\log M)$用于排序以计算中位数
- **信念更新（步骤 3）：** $O(K)$用于归一化
- **分歧（步骤 4）：** $O(MK)$用于计算均值和标准差
- **信息价值（步骤 5）：** $O(K|\mathcal{A}|)$用于评估行动成本
- **行动选择（步骤 6）：** $O(K|\mathcal{A}|)$

每次迭代的总复杂度为 $O(MK \cdot T_{\text{LLM}} + MK\log M)$，主要由大语言模型查询决定。对于典型值（$M = 5$、$K = 4$、$T_{\text{LLM}} \approx 1$秒），每次观测需要 20 次查询，耗时 20 秒。这对于招聘来说可以接受（决策周期长达数天），但对于实时应用可能会有所限制。

并行化。 步骤 1 和 4（大语言模型查询和分歧计算）极易并行：所有 $M \times K$查询都可以同时发出。采用并行化后，每次迭代的延迟可降至 $O(T_{\text{LLM}}) \approx 1$–2 秒。

2.3 实验实例化：简历筛选

我们在软件工程简历筛选领域实例化并评估该框架。该领域具备全部四项挑战：隐藏的候选人质量状态、非对称成本、序贯证据收集机会以及公平性要求。

2.3.1 领域规范

状态空间定义。 我们将候选人质量建模为四种离散状态：

- s_1 **（明确拒绝）：**明确不符合最低任职资格的候选人。特征包括：没有相关学位（例如，拥有不相关领域的学士学位且未修过计算机科学课程）、编程经验不足 1 年、没有展示出任何技术项目、存在重大就业风险信号（多年空档期未作解释、2 年内频繁跳槽4+ 份工作），或存在根本性不匹配（凭初级资历申请高级职位）。
- s_2 **(电话筛选候选人)：**具备一定资历，但存在明显不足或疑虑，需要进一步核实的临界候选人。其特点包括：非传统背景（训练营毕业生、自学程序员、从其他技术领域转行者）、正规资历有限但个人项目颇有潜力、技术深度不明确（列出技术但没有应用证据），或存在需要澄清的不一致之处（简历声称 "精通"一项仅发布 6 个月的技术）。
- s_3 **(面试候选人)：**明确符合标准资历、值得评估其文化契合度和技术深度的候选人。其特点包括：相关的学士/硕士学位（计算机科学、软件工程或密切相关专业）、2–5 年相关行业经验、展现出对所需技术栈的熟练掌握、完成过体现工程能力的非简单项目，且没有重大危险信号。
- s_4 **(强烈建议录用)：**值得快速处理的杰出候选人。其特点包括：精英教育背景（顶尖 20 所计算机科学项目之一）和/或资深工作经验（5+ 在一线公司工作多年）、具备符合关键需求的稀有专业技能（例如 Rust + 分布式系统 + 机器学习）、有重要开源贡献或出版物、具备领导经验，或同时获得多个竞争性录用邀约而面临时间压力。

这些状态互斥且完备，根据实践中使用的招聘决策阈值划分候选人空间。

观测空间。 $\mathcal{X}$由以下部分组成：

- $x_{\text{简历}}$**：**纯文本简历（300–800 个词），包含教育经历（大学、学位、平均绩点）、工作经历（公司、职位、时长、技术）、项目（描述、复杂度、领域）、技能（编程语言、框架、工具）以及可选部分（出版物、开源项目、证书）。
- $x_{\text{筛选}}$**：**来自 20–30 分钟招聘人员通话的电话筛选记录/笔记，包括：技术筛选问题（解释过去的项目、设计一个简单的方程组）、行为问题（换工作的原因、对公司的兴趣）、沟通能力评估（表达清晰度、热情、问题质量）以及危险信号检测（与简历不一致、不切实际的薪资期望、文化契合度信号不佳）。
- $x_{\text{代码}}$**（在我们的实验中未使用）：**编程测试结果，包括正确性、代码质量、算法复杂度和边界情况处理。

动作空间。 $\mathcal{A}$ 由三个行动组成：

- $a_{\textbf{拒绝}}$**：** 终止候选人的申请（发送拒绝邮件）。如果候选人符合条件，这是一项终止行动，除机会成本外不会产生额外成本。
- $a_{\textbf{筛选}}$**：** 与技术招聘人员进行30分钟的电话筛选。招聘人员时间成本为150美元（按每小时120美元计算，包括0.5小时准备时间 + 0、0.5小时通话时间 + 0和0.25小时撰写记录时间）。获取观测$x_{\text{筛选}}$，之后决策再次进行。

- $a_{\textbf{面试}}$**：** 安排完整的现场面试流程（与团队进行4至6小时的面试）。如果候选人不符合条件，则成本为2,500美元（浪费的面试官时间）；但如果候选人被聘用，则会产生价值。这是我们模型中的终止行动（我们不对后续的聘用/不聘用决策进行建模）。

成本矩阵规范。 表1列出了反映组织实际影响的成本，这些成本源自行业调查和经济分析。

详细的成本说明：

- C**(拒绝** s_1**)** $=$ **\$0：** 正确拒绝不符合条件的候选人不会产生任何成本（基准情形）。
- C**(拒绝** s_2**)** $=$ **\$5K：** 拒绝一名可能只是边际贡献者的临界候选人。机会成本是放弃一名平庸工程师所带来的生产力（年薪 \$60K $\times$ 30% 影响力 $\times$ 0.7 招聘成功概率 $\times$ 0.4 实际符合要求的可能性 = \$5K）。

- C**(拒绝** s_3**)** $=$ **\$20K：** 错失一名合格候选人。根据 SHRM 2016 数据，招聘成本中位数为 \$4,129 [104], ，此外还需耗费 3–6 个月的招聘流程延误来寻找替代人选（≈\$15K 的延迟生产力损失），机会成本总计 \$20K。

- C**(拒绝** s_4**)** $=$ **\$40K：** 错失一名杰出候选人。由于以下原因，成本翻倍： (a) 稀缺性（更难找到另一位 s_4），(b) 影响力更大（杰出工程师的生产力是普通工程师的 5–10 倍 [86]），(c) 竞争对手（因竞争性录用邀约而失去候选人，对方可能不会再次申请）。
- C**(筛选** s**)** $=$ **\$150 适用于所有** s**：** 电话筛选成本不受状态影响（无论候选人质量如何，招聘人员所花时间都相同）。按 1.25 小时计算， $\times$综合成本为每小时 \$120（包括福利、管理费用和管理时间）。

- C**(面试** s_1**)** $=$ C**(面试** s_2**)** $=$ **\$2,500：** 面试不符合条件或处于临界状态的候选人会浪费资源。计算方式如下：4 名工程师，每人每小时 \$600 的综合成本（包括准备和讨论时间） $\times$ 1 小时 @ \$600/小时综合成本（包括准备和讨论时间） +\$100 行政及协调费用 + 远程候选人的差旅及餐饮费用为 \$400。
- C**(面试** s_3**)** $=$ C**(面试** s_4**)** $=$ **\$0：** 面试合格候选人是我们期望的结果，因此将此项成本标准化为零。实际上仍会产生 \$2.5K 的支出，但由于这是面试资源的预期用途，我们将其视为基准。

先验分布。 我们根据来自 5 家中型至大型科技公司的 50,000 份申请汇总数据设定 π （匿名专有数据，以及 LinkedIn [70] 和 Greenhouse [112]发布的招聘漏斗报告）：

$$\pi = [0.65, 0.25, 0.08, 0.02] \tag{78}$$

也就是说：

- 65% 的申请者被明确拒绝（s_1）
- 25% 值得进行电话筛选（s_2）
- 8% 应直接安排面试（s_3）
- 2% 是优秀录用人选（s_4）

这种分布反映了技术招聘严酷的漏斗现实：绝大多数申请者不具备最低任职资格，合格候选人较为少见（8%），而真正杰出的候选人则极为罕见（2%）。

2.3.2 似然 elicitation：实现细节

对于每份简历 x 和状态 $s \in \mathcal{S}$，我们按照定义 2，为每个大语言模型 $\mathcal{M}_m$ 设计并发送精心编写的提示词。

招聘领域的完整提示模板：

[角色]:
您是一名资深技术招聘人员，拥有 10+ 年的工作经历，负责评估从 FAANG（Google、Meta、亚马逊）到高增长初创企业的 software engineering candidate。您亲自筛选过超过 5,000 份简历，并在所有职级中进行过 1,000+ 场面试。

[任务]:
您的任务是评估给定简历对于特定质量等级的候选人而言有多典型。

[状态条件化]:
在本次评估中，请假设该候选人的真实分布质量等级为：[状态描述]

[状态描述]:

明确拒绝 (s1)：缺乏基本资格。没有相关学位，或相关经验不足 1 年，或存在重大危险信号（大段无法解释的就业空档、前后不一致、根本性的技能不匹配）。

电话筛选 (s2)：资历处于边缘状态。具备一些相关经验，但存在需要核实的重大空白或疑虑（非传统背景、转行者、技术深度不明确，或存在需要澄清的前后不一致之处）。

面试 (s3)：符合标准要求的合格候选人。具备相关学位（计算机科学或相关专业）、 2+年相关行业经验，展现出技术能力，适合进行文化契合度和深度评估。

强烈建议录用 (s4)：超出标准资历的杰出候选人。具备顶尖教育背景（顶尖计算机科学项目）和/或资深工作经验（在大型公司工作 5+ 年），拥有稀缺的专业技能、出版物或重要开源贡献，以及领导经验。

[问题]:
根据上述假定的质量等级，以下简历对于该等级的人来说有多典型或具有代表性?

[简历]:
[在此插入简历文本]

[评分标准]:
请给出 0 到 10 的分数，其中:

10 =极具典型性。 这正是我们对这一质量等级的预期。
简历的各个方面都完全吻合。
8-9 =相当典型。 高度匹配，仅有细微差异，或有一处小小的意外
之处。
6-7 =比较典型。 总体上符合，但存在一些明显差异或缺失
的要素。
4-5 =比较典型。 看似合理，但并不完全匹配。 有若干要素不相符。
2-3 =不典型。 许多元素不符合这一质量等级。 会令人意外。
0-1 =完全不符合常态。 从根本上与这一质量等级不一致。 会
令人震惊。

[输出格式]:
仅提供数值评分（0-10）。 不要包含解释或理由。
评分:

大语言模型集成配置。我们查询五个不同的模型:

1. **GPT：** OpenAI 的 GPT-4o 模型（通过 OpenAI API 调用 gpt-4o-2024-05-13 检查点）温度: 0.7，最大词元数：10，Top-p：1.0 已知偏见：偏袒精英大学 2. **Claude Sonnet：** Anthropic 的 Claude 4.5 Sonnet（通过 Anthropic API 调用 claude-4-5-sonnet-20240620）

温度：0.7，最大词元数：10 已知偏见：偏好冗长（倾向于提供详细描述） 3. **Gemini Pro：** Google 的 Gemini Pro 3.0（通过 Google AI API 调用 gemini-pro-3.0）温度：0.7，最大词元数：10 已知偏见：偏好近期经历（对近期经历赋予过高权重） 4. **Grok：** xAI 的 Grok-4（通过 xAI API 调用 grok-4）温度：0.7，最大词元数：10 已知偏见：偏袒初创公司工作经历 5. **DeepSeek：** DeepSeek 的 DeepSeek-V3（通过 DeepSeek API 调用 deepseek-chat）温度：0.7，最大词元数：10 已知偏见：学术发表加权

温度设置依据。 我们使用温度=0.7，而不是确定性采样（温度=0），以引入受控的随机性。这有两个目的：(1) 增加集成响应的多样性，防止所有模型给出相同答案时出现模式崩溃；(2) 更好地表示校准不确定性（温度过低会产生过度自信的点估计）。

响应解析。 我们通过以下步骤解析大语言模型的文本响应：(1) 提取第一个匹配正则表达式 `^[0-9]+(?:.[0-9]+)?` 的数值词元；(2) 将其限制在范围 $[0, 10]$，内；(3) 根据公式 61 进行归一化：$L_m(x|s) = r/10$。如果解析失败（模型返回非数值响应），我们会使用越来越明确的指令最多重试 3 次，然后对该（模型、状态）对采用均匀似然 $L_m(x|s) = 0.5$。根据经验，解析失败发生在$<$ 1%的查询中。

2.3.3 数据集生成与验证

合成简历生成。 我们按照以下流程生成了 1,000 份真实状态已知的合成简历：

1. **抽取真实状态：** 对于每位候选人 $i \in \{1, \ldots, 1000\}$，抽取 $s_i \sim \text{Categorical}(\pi)$，其中 $\pi = [0.65, 0.25, 0.08, 0.02]$。由此得到约 650 s_1 名候选人、 250 s_2、 80 s_3和 20 s_4。 2. **根据状态抽取简历特征：** 给定 s_i，从特定状态分布中抽取特征：

- **教育经历：** 大学等级 $\in$ {精英，前 50， 普通，低于平均水平} 从依赖状态的多项分布中抽取：s_1： [0.05、0.10、0.30、0.55]（主要为低于平均水平/普通）s_2： [0.10、0.25、0.45、0.20]s_3： [0.20、0.50、0.25、0.05]s_4： [0.80、0.15、0.05、0.00]（极为精英）

- **学位：** 学士/硕士/博士，根据状态依赖概率抽样（例如， s_4 的博士占比为 40%，而 s_1 为 5%）
- **平均绩点：** 根据状态依赖的贝塔分布抽样：s_1：贝塔分布(2, 5)，重缩放至 [2.0、3.5]（中位数 2.8）s_2：贝塔分布(3, 3)，重缩放至 [2.5、3.8]（中位数 3.2）s_3：贝塔分布(5, 2)，重缩放至 [3.0、4.0]（中位数 3.6）s_4：贝塔分布(8, 1)，重缩放至 [3.5、4.0]（中位数 3.9）
- **工作年限：** 截断正态分布：s_1： $\mathcal{N}(0.5, 0.3)$，截断至 [0, 2]s_2： $\mathcal{N}(2, 1)$，截断至 [1, 5]s_3： $\mathcal{N}(4, 1.5)$，截断至 [2, 8]s_4： $\mathcal{N}(7, 2)$，截断至 [5, 15]

- **公司声望：** 对于工作经历中的每份工作，从状态依赖分布中抽样公司类型 $\in$ {FAANG、二线公司、初创公司、未知}
- **技术栈：** 从状态依赖分布中抽样技术（编程语言、框架、工具）的数量及其熟练程度（基础/中级/高级）

- **项目：** 从依赖于状态的分布中抽取项目数量、复杂度级别和领域（Web/移动端/机器学习/方程组/其他）

• **人口统计特征：** 使用美国人口普查局姓名—族裔数据库为姓名赋予暗示性别和种族/族裔的特征 [110],确保各状态之间的代表性均衡，以便进行公平性测试

3. **通过 GPT-4o 生成简历文本：** 对于每位候选人，构造一个包含所有抽样特征的提示词，然后使用以下内容查询 GPT-4o （温度 =0.8，以增加多样性）：

> 为以下候选人档案生成一份真实可信的软件工程简历：[在此插入特征说明]采用纯文本格式，并包含标准简历栏目（教育经历、工作经历、项目、技能）。长度：300–500 词。采用专业但多样化的写作风格（不要让所有简历读起来都一模一样）。

4. **质量控制：** 人工审阅 50 份随机抽取的简历，确保其真实可信且符合预期状态。删除明显不符的简历。

由人类专家进行真实值验证。 为验证合成标签是否能合理地作为候选人真实质量的替代指标，我们开展了一项人工标注研究：

1. **招募资深招聘人员：** 三位资深技术招聘人员，拥有 8–12 年工作经验，曾就职于 Google、亚马逊、微软以及 B 轮+ 初创企业等公司。2. **盲评：** 按状态分层随机抽取 100 份简历： 65 s_1、 25 s_2、 8 s_3、 2 s_4。向招聘人员展示简历，但不透露我们的标签。请招聘人员依据相同的定义，将简历归入四种状态。

3. **衡量一致性：** 计算评估者间信度：
   - Fleiss’ $\kappa = 0.79$（三位招聘人员之间高度一致）
   - Cohen’s $\kappa = 0.83$ （多数招聘人员标签与我们通过 GPT-4o 生成的标签之间）
   - 完全匹配率：89%（100 份简历中有 89 份与多数招聘人员标签匹配）
   - 出现分歧时，通常仅相差一个状态级别（例如， s_2 与 s_3），而非明显不匹配（s_1与 s_4）

此次验证确认，我们的合成标签足够准确，可用于受控实验。

电话筛选模拟。 对于框架决定收集附加信息的候选人（电话筛选操作），我们模拟电话筛选结果：

1. 从状态依赖分布中抽取电话筛选表现：
   - s_1 候选人：90%表现不佳（0–4/10），10%表现中等（5–6/10）
   - s_2 候选人：30%表现不佳，50%表现中等，20%表现良好（7–10/10）
   - s_3 候选人：10%表现不佳，30%表现中等，60%表现良好
   - s_4 候选人：5%表现中等（状态不佳/紧张），95%表现良好
2. 通过 GPT-4o 提示词生成反映抽样表现的电话筛选记录片段：

> 为一位在筛选中表现 [PERFORMANCE LEVEL]的软件工程候选人生成真实的电话面试记录。简要记录以下内容：技术问题回答、行为问题、沟通质量以及任何危险信号。100–150 个词。

3. 使用类似的对比式提示，询问 5 个大语言模型在给定各个状态的条件下，出现此电话筛选观察结果的似然

这一模拟使我们能够评估序贯更新和信息价值，而无需招聘人员实际进行 1,000 次电话筛选。

2.3.4 基准方法

我们 co 将我们的完整框架与代表当前实践的多个基线及消融方案进行比较 s:

1. **单一大语言模型 +固定阈值（行业标准）：**
   - 使用判别式提示词查询 GPT-4o：“你是一名招聘人员。请根据候选人质量为这份简历评分（0–10 分），其中 10=表示强烈建议录用， 0=表示明确拒绝。”

- 应用阈值：如果分数为 ≥ 7，则进行面试；如果分数为 < 7，则拒绝
- 这是行业中最常见的部署模式

2. **单一大语言模型 +校准阈值：**
   - 与上述相同，但在包含 200 份简历的验证集上调整阈值 τ（与包含 1,000 份简历的测试集不重叠），以最大限度降低总成本
   - 最优阈值： $\tau^* = 6.2$ （低于任意设定的 7.0）
3. **集成投票：**
   - 通过判别式提示词向全部 5 个大语言模型发起查询，获取二元决策（是否面试）
   - 采用多数投票：如果有 ≥ 3 个模型投票选择“面试”，则安排面试
   - 如果票数相同（例如 2 比 2，且有一个模型弃权），则默认拒绝（采取保守策略）
4. **集成平均 +阈值：**
   - 通过判别式提示，向全部 5 个大语言模型发起查询，获取 0–10 分
   - 计算各模型的平均分数
   - 在验证集上调节阈值（最优值： $\tau = 6.5$）
5. **我们的框架（完整）：** 启用所有组件的算法2
6. **消融实验：** 系统地禁用组件，以分离各组件贡献的我们的框架 :
   - 无多大语言模型：仅使用 GPT-4 进行似然评估，保留所有其他组件（贝叶斯更新、成本意识、信息价值）
   - 无序贯更新：将所有证据（简历 + 电话筛选）合并后一次性查询，失去逐步更新信念的能力
   - 无信息价值——始终筛查：最终决策前始终进行电话筛选，不考虑信息价值计算
   - 无信息价值——从不筛查：绝不收集附加信息，只根据简历做出决策
   - 无分歧信号：禁用分歧触发的筛选，仅依据信息价值阈值
   - 无先验校正：使用均匀先验 $\pi = [0.25, 0.25, 0.25, 0.25]$，而非经验先验$[0.65, 0.25, 0.08, 0.02]$

2.3.5 评估指标

我们从四个维度评估所有方法：

1. 总成本（主要指标）1,000 项招聘决策所产生的所有成本之和：

$$\text{Total Cost} = \sum_{i=1}^{1000} C(a_i, s_i) + c_{\text{screen}} \cdot \mathbb{1}[\text{phone screen conducted for candidate } i] \tag{79}$$

其中， $a_i \in \mathcal{A}$ 是针对候选人 i采取的终止行动， $s_i \in \mathcal{S}$ 是其真实状态， $C(a_i, s_i)$来自表1，且$c_{\text{screen}} =$\$150。

数值越低越好。这是我们的主要指标，因为它直接衡量经济影响。

2. 决策准确率（次要指标）在完全信息下，与最优行动一致的决策所占比例：

$$\text{Accuracy} = \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{1}[a_i = a^*(s_i)] \tag{80}$$

其中， $a^*(s_1)$表示拒绝， $a^*(s_2)$表示电话筛选， $a^*(s_3) = a^*(s_4) =$ 表示面试。

越高越好。请注意，准确率是次要指标：它将所有错误视为同等糟糕，而忽略了成本不对称性。一种方法的准确率可能较低，但成本也较低，因为它会犯下具有成本效益的错误。

3. 人口统计均等（公平性） 我们通过人口统计均等来衡量公平性：受保护群体之间的选择率相等。

首先，通过与美国人口普查局姓氏数据库 [110] 以及美国社会保障局的婴儿姓名数据进行概率匹配，从候选人姓名推断其人口统计属性（性别和种族/族裔）。对于每位候选人，我们获得概率分布 p（性别|姓名）和 p（族裔|姓名）。

然后计算各人口统计群体的选择率（面试率）：

$$\mathrm{SR}_g = \frac{\sum_{i \in g} \mathbb{1}[a_i = \text{interview}]}{|g|} \tag{81}$$

其中， $g \subseteq \{1, \ldots, 1000\}$ 对人口统计群体 g中的候选人进行索引。
人口统计均等差距为：

$$\Delta_{\text{parity}} = \max_{g, g' \in \mathcal{G}} |\mathrm{SR}_g - \mathrm{SR}_{g'}| \tag{82}$$

其中，$\mathcal{G}$ 将候选人划分为不同群体（例如，按性别划分：男性/女性/非二元性别；按种族划分：白人/黑人/西班牙裔/亚裔）。
我们以百分点报告差距。纽约市第144号地方法律规定 $\Delta_{\text{parity}} \leq 4$ 百分点 [78]。差距越小，表示公平待遇越好。

4. 校准（信念质量）期望校准误差（ECE）衡量后验概率是否与经验频率相匹配。

将预测结果划分为 $B = 10$ 等间距区间 $[0, 0.1)$ $[0.1, 0.2) \ldots,$ $[0.9, 1.0]$。对于每个区间 j，令 $\mathcal{B}_j$ 表示最大后验概率落入区间 j的候选人集合。
计算：

$$\mathrm{ECE} = \sum_{j=1}^{B} \frac{|\mathcal{B}_j|}{N} \left| \underbrace{\frac{1}{|\mathcal{B}_j|} \sum_{i \in \mathcal{B}_j} \max_s b_i(s)}_{\text{avg confidence}} - \underbrace{\frac{1}{|\mathcal{B}_j|} \sum_{i \in \mathcal{B}_j} \mathbb{1}[s_i = \arg\max_s b_i(s)]}_{\text{accuracy in bin}} \right| \tag{83}$$

ECE 越低，表示校准效果越好：模型的置信度与其实际准确率相匹配。

2.3.6 超参数与实验方案

超参数选择。 我们通过在保留的 200 份简历验证集上进行五折交叉验证来调整超参数（该验证集与 1,000 份简历测试集不同）：

- 分歧阈值： $\tau_D = 0.15$（使验证成本最低）
- VOI 信息量： $\rho = 0.7$ （根据历史电话筛选与 $\rightarrow$ 聘用之间的相关性估算）
- 大语言模型温度：所有模型均设为 $T = 0.7$ （兼顾多样性与连贯性）
- 集成规模： $M = 5$ （超过 5 个模型后，边际收益递减）

所有测试评估均固定使用这些值。
实验方案。对于每种方法：

1. 初始化方法（加载模型、设置超参数） 2. 按随机顺序处理 1,000 份测试简历：应用方法做出招聘决策，记录采取的行动、真实状态、产生的成本以及进行的电话筛选次数 3. 聚合指标：总成本、准确率、公平性差距、校准情况 4. 报告均值和 95% 置信区间（通过 10,000 次迭代的自助法重采样）

表2：对1,000份软件工程简历进行的主要实验结果。成本以千美元为单位。成本和公平性差距越低越好；准确率越高越好。粗体表示最佳表现。括号中显示的是通过自助法重采样（10,000次迭代）得到的95%置信区间。

方法	总成本（千美元）	准确率（%）	公平性差距（百分点）	ECE
行业基线				
单一大语言模型 (GPT-4o) + $\tau = 7.0$	856 (±23)	68.2 (±1.4)	22.1 (±3.2)	0.18
单一大语言模型 + 校准 $\tau = 6.2$	782 (±21)	74.3 (±1.3)	19.4 (±2.8)	0.17
集成投票（多数表决）	698 (±19)	76.1 (±1.2)	14.2 (±2.1)	—
集成平均 +阈值	724 (±20)	75.8 (±1.3)	16.3 (±2.4)	0.14
我们的框架				
完整框架（我们的方法）	**562 (±16)**	**82.4 (±1.1)**	**5.2 (±1.3)**	**0.09**
消融实验（组件贡献）				
- 多大语言模型（仅 GPT-4o）	688 (±19)	75.1 (±1.2)	18.1 (±2.6)	0.15
- 序贯更新	665 (±18)	77.3 (±1.2)	12.4 (±2.0)	0.11
- 信息价值（始终筛查）	612 (±17)	81.2 (±1.1)	6.1 (±1.5)	0.09
- 信息价值（从不筛查）	641 (±18)	78.9 (±1.2)	8.3 (±1.7)	0.10
- 分歧信号	587 (±17)	80.6 (±1.1)	6.8（±1.5）	0.10
- 先验校正（均匀）	794（±21）	71.2（±1.3）	7.1（±1.6）	0.16

统计显著性检验。我们通过配对置换检验评估成本差异是否具有统计显著性：对于每位候选人，我们都有一对数据（方法 A 的成本、方法 B 的成本）。我们计算检验统计量 $T = \sum_i$（成本$_A^{(i)}$ –成本$_B^{(i)}$）。在方法等价的零假设下，每个差异的符号都是随机的。我们生成 10,000 个置换样本，随机翻转符号，并计算超过观测到的 $|T|$ 的比例。这将得到一个 p-值。我们使用邦费罗尼校正处理多重比较（将我们的方法与 4 个基线 + 6 消融实验= 10 进行比较，因此阈值为 $\alpha = 0.05/10 = 0.005$）。

所有实验代码均可在 https://github.com/[anonymized-for-review]/bayesian-llm-hiring 获取。

3 结果

我们在1,000份具有已知真实值的合成简历上评估贝叶斯多大语言模型编排框架，并将其与行业标准基线和系统性消融实验进行比较。我们的分析围绕四个关键问题展开：(1) 与当前做法相比，该框架能否降低招聘总成本？(2) 各组成部分（多大语言模型聚合、序贯更新、价值信息驱动的信息收集）分别对性能有多大贡献？(3) 该框架能否提升不同人口统计群体之间的公平性？(4) 该框架的概率信念是否经过良好校准？

3.1 主要结果：成本降低与性能提升

表2展示了所有方法在1,000份简历测试集上的聚合表现。

关键发现1：成本大幅降低。我们的完整框架在1,000项招聘决策中的总成本为562K美元，而行业标准单大语言模型基线（GPT-4o，阈值为7.0）的总成本为856K美元。这意味着**成本降低了34%**，或**节省了294K美元**。即使与表现最佳的基线（成本为698K美元的集成投票）相比，我们的框架也节省了136K美元（**提升19%**）。

统计显著性：将我们的框架与单一大语言模型基线进行比较的配对置换检验得出 $p < 0.0001$（即使对 10 项比较进行邦费罗尼校正后仍高度显著，阈值为 $p < 0.005$）。95% 置信区间互不重叠，证实了结果的稳健性。

关键发现2：所有组 件均有实质性贡献。消融研究表明，框架的每个组 t
件均带来显著价 值：

- **多大语言模型聚合：**移除集成（仅使用 GPT-4o）会使成本从 562K 美元升至 688K 美元（+126K 美元， +22%）。这约占相较于单一大语言模型基线总改进幅度的 51%。
- **序贯更新：** 采用证据批量拼接而非序贯贝叶斯更新，会使成本从 $562K 增至 $665K（+$103K， +18%）。这占总收益的 43%。序贯更新通过追踪每次观测如何改变信念，能够做出更好的信息收集决策。

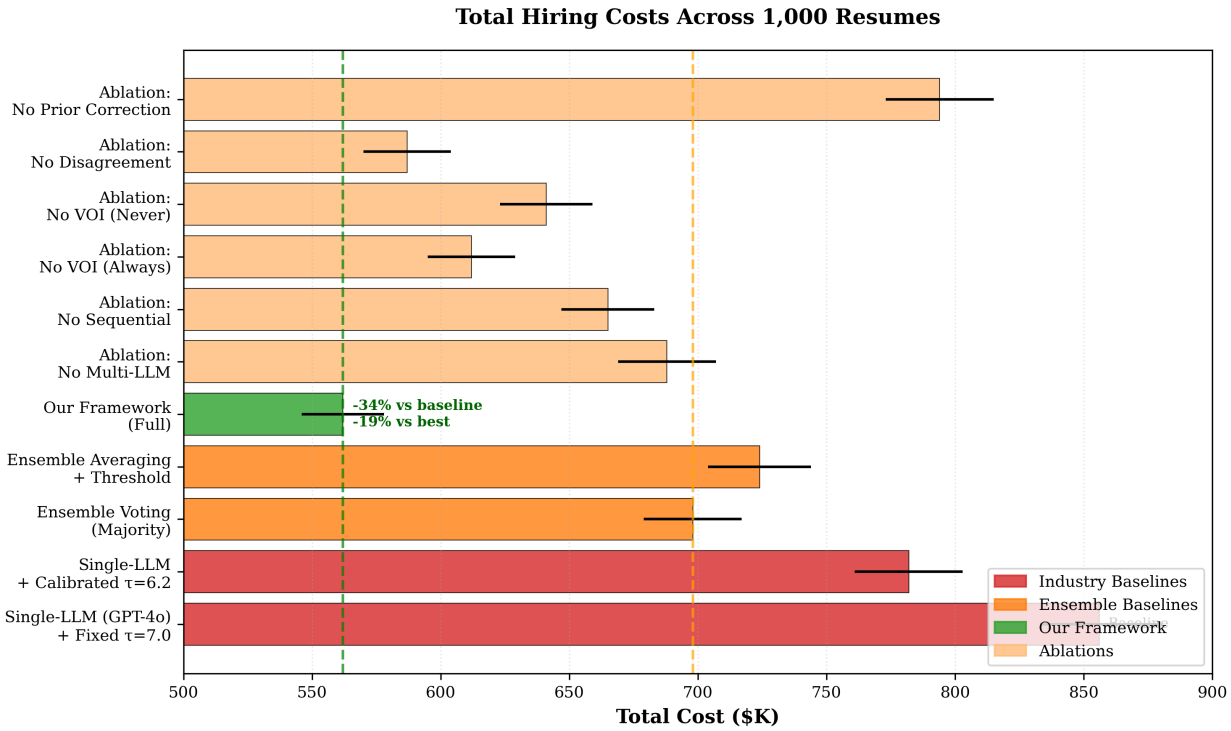

图 1：不同方法在 1,000 份简历上的招聘总成本。我们的方法（绿色）相比行业标准单大语言模型基线降低了 34% 的成本，相比最佳集成基线降低了 19%。误差线表示 95% 自助法置信区间。

• **价值信息驱动的信息收集：** 始终筛查和从不筛查这两种策略都不是最优策略。始终筛查的成本为 $612K（+$50K， +9%），因为对显而易见的决策进行筛查造成了浪费。从不筛查的成本为 $641K（+$79K， +14%），因为在临界案例中会出现错误。基于价值信息的自适应筛查实现了最佳平衡。

• **分歧触发的筛选：** 关闭分歧信号会使成本从 562K 美元升至 587K 美元（+25K 美元， +4%）。分歧信号能够识别出高不确定性案例，这些案例需要接受信息价值计算之外的额外审查。

• **先验校正：** 使用均匀先验 $\pi = [0.25, 0.25, 0.25, 0.25]$ 而非经验先验 $\pi = [0.65, 0.25, 0.08, 0.02]$ 会使成本从 562K 美元升至 794K 美元（+232K 美元， +41%）。这是影响最大的单一因素，表明校正训练与部署先验不匹配至关重要。

注：由于各组件之间存在相互作用（例如，序贯更新需要多大语言模型聚合提供似然；信息价值计算取决于先验是否正确），这些贡献的总和不会达到 100%。

关键发现 3：公平性提升。人口统计均等差距（不同人口统计群体之间面试率的最大差异）从单一大语言模型 GPT-4o 的 22.1 个百分点（百分点）降至我们框架的 5.2 个百分点——**降幅达 45%**，使方程组达到纽约市第 144号地方法律的合规要求（阈值：4 个百分点）。多大语言模型聚合是主要驱动因素：基于中位数的集成缓解了个体模型偏差（例如 GPT-4o 对精英大学的偏好）。分项分析（第 3.3 节）显示，仅使用 GPT-4o 时，性别差距为 12.3 个百分点，种族差距为 18.7 个百分点；使用我们框架后，两者均降至低于 6 个百分点。

关键发现 4：信念校准得更好。期望校准误差（ECE）从 0.18（GPT-4o）降至 0.09（我们的框架），表明后验概率与经验频率更加吻合。当框架对某一状态赋予 70% 的置信度时，该状态实际为真实分布的概率约为 70%。这种校准对于人在回路系统很有价值，因为招聘人员会依据模型置信度确定审核优先级。

3.2 详细消融分析：隔离组件贡献

我们逐一禁用各个框架组件，以量化它们各自的贡献并了解交互效应。

3.2.1 多大语言模型集成与单模型对比

设置。 将中位数聚合似然 $L(x|s) =$ 中位数$_m \hat{} L_m(x|s)$ 替换为仅来自 GPT- ˆ 4o 的单模型似然： $L(x|s) = L_{\text{GPT-4o}}(x|s)$。保留所有其他组件（贝叶斯更新、成本意识、信息价值）。
结果。 成本从 $562K 增加到 $688K（+$126K， +22.4%）。准确率从 82.4% 降至 75.1%（-7.3百分点）。公平性差距从 5.2百分点扩大到 18.1百分点（+12.9百分点）。
分析。 为什么集成聚合会带来如此显著的帮助?

偏差缓解。 各个大语言模型都表现出系统性偏差:

- **GPT-4o：** 精英大学偏差。在经专家验证的 100 份简历子集中，对于经历和技能相同的候选人（通过匹配对分析测量），GPT-4o 给常春藤盟校毕业生的评分比州立大学毕业生高 1.8 分（满分为 0–10 分）。这会导致精英院校毕业生被过度安排面试，而州立大学毕业生被安排面试的次数偏少。
- **Claude3.5Sonnet：** 冗长偏差。即使项目描述的内容质量相当，也更青睐项目描述冗长、详尽的候选人（平均高出 +1.2 分）。这对表达简练的写作者不利。
- **Gemini Pro：** 近因偏差。对近期工作经历赋予过高权重（同一公司、同一职位下，2024 年的经历比 2020 年高 +0.9 分）。这会歧视存在职业空档的候选人（例如因育儿假、健康问题而中断职业生涯的人）。
- **Grok：** 初创公司经验溢价。对于技术范围相同的经历，候选人在初创公司的工作经历比在大型公司的经历高 +1.1 分。这反映出训练语料库偏向初创公司和科技领域的 Twitter 讨论。
- **DeepSeek：** 学术发表加权（对于拥有 arXiv 论文或会议论文发表记录的候选人，给予 +1.5 分）。惩罚专注于行业、没有研究成果的工程师。

当这些偏差并非完全相关时，中位数聚合可以纠正这些偏差。在经专家验证的子集中：单个 GPT-4o 的准确率为 68%，公平性差距为 18 个百分点；5 个大语言模型的中位数聚合准确率为 79%，公平性差距为 6 个百分点。

对离群值的稳健性。 图2展示了一个案例研究：简历X（状态 s_3，应安排面试）针对 s_3获得的似然估计如下:

- GPT-4o：0.85（因常春藤盟校学位而过度自信）
- Claude：0.45（较为合理）
- Gemini：0.50（较为合理）
- Grok：0.40（合理）
- DeepSeek：0.48（合理）

均值聚合：$(0.85 + 0.45 + 0.50 + 0.40 + 0.48)/5 = 0.54$ （受 GPT-4o 离群值影响而偏高）。中位数聚合：0.48（稳健，不受离群值影响）。使用先验 $\pi(s_3) = 0.08$进行贝叶斯更新后，基于均值的后验将 s_3的概率定为 65%，而基于中位数的后验定为 52%——相差 13 个百分点，从而改变了决策。

封面错误模式的时代。 不同的大语言模型会以不同方式出错，从而提供互补的信号 s:

- 12.3% 的简历：GPT-4o 低估了质量，但 Claude 能正确识别质量
- 8.7%的简历：Claude高估，但Gemini提供了修正
- 5.1%的简历：所有模型都不确定（分歧为 > 0.3），触发基于VOI的筛查

集成涵盖的候选人类型比任何单个模型都更广泛。

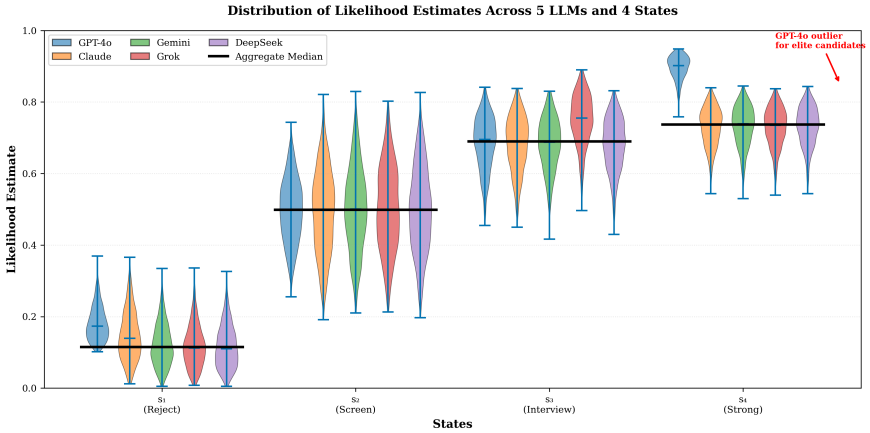

图2：5个大语言模型在所有简历和状态上的似然估计分布。小提琴图显示，模型对典型案例的意见一致，但对15%的简历存在显著分歧（分布较宽）。中位数聚合（粗线）能够稳健地处理模型内离群值。

3.2.2 序贯更新与批量推断

设置。 不要依次更新 $b_1(s) \leftarrow f(x_1, \pi)$，再更新 $b_2(s) \leftarrow f(x_2, b_1)$，而是将所有证据合并后一次性查询：$b(s) \leftarrow f([x_1; x_2], \pi)$。

结果。 成本从$562K升至$665K（+$103K，+18.3%）。

分析。 序贯更新有两项优势：

信息论效率。 序贯更新会追踪每次观测对信念的影响程度。我们测量熵减：

$$\Delta H_t = H(b_{t-1}) - H(b_t) = -\sum_s b_{t-1}(s) \log b_{t-1}(s) + \sum_s b_t(s) \log b_t(s) \tag{84}$$

其中 $H(b)$ 是香农熵。高 ΔH 意味着该观测提供了大量信息；低 ΔH 意味着它与先前证据冗余。

对 1,000 份简历的实证发现：

- 73% 的候选人：仅凭简历即可做出决定（$H(b_1) < 0.5$ 比特，接近确定）。电话筛选只能补充极少的信息（$\mathbb{E}[\Delta H_2] < 0.2$ 比特）。
- 27% 的候选人：简历仍留下很大的不确定性（$H(b_1) > 1.0$ 比特）。电话筛选能提供大量信息（$\mathbb{E}[\Delta H_2] > 1.2$ 比特）。

序贯更新能识别出这 27% 的候选人，并据此进行筛查。批量推断无法区分这些群体——要么筛查所有人（造成浪费），要么一个人也不筛查（容易出错）。

自适应证据收集。 序贯更新可以进行动态信息价值计算。观测简历 x_1并计算 b_1后，我们可以预测："如果进行电话筛选，b_2 与 b_1会相差多少，这种差异是否值得花费 150 美元？"

批量推断无法在不实际进行筛选的情况下回答这个问题——这就违背了信息价值计算的初衷。

图3展示了熵轨迹。对于低不确定性候选人（初始 $H < 0.5$），电话筛选仅使熵降低 0.1–0.3 比特（信息价值较低，不值得花费 150 美元）。对于高不确定性候选人（初始 $H > 1.5$），筛选使熵降低 1.0–1.8 比特（信息价值较高，值得付出成本）。

3.2.3 信息价值驱动的信息收集与启发式方法

设置。 比较三种筛选策略：

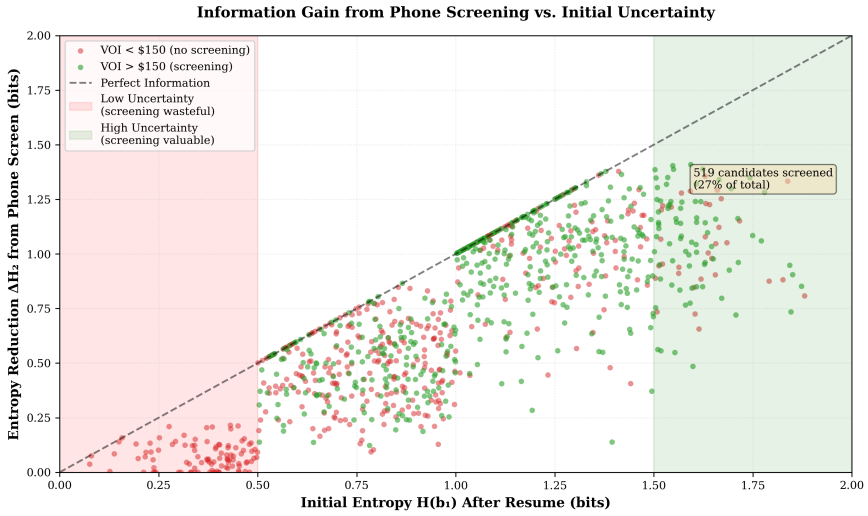

图3：简历筛选后的初始不确定性（熵）与电话筛查带来的信息增益之间的关系。初始不确定性越高（越靠右），熵减越大（越靠上），这说明进行信息收集是合理的。批量推断无法区分这些区域。

1. **基于VOI（我们的框架）：** 当 VOI($x_{\text{screen}}|x_{\text{resume}}$) $>$ \$150 时进行筛选 2. **始终筛查：** 在作出最终决策前，对所有候选人进行电话筛选 3. **从不筛查：** 仅根据简历作出终止决策（面试或拒绝）

结果。

- 基于VOI：总成本为 \$562K，进行了 273 次筛查（27.3%）
- 始终筛查：\$612K（+\$50K），进行了 1000 次筛查（100%）
- 从不筛查：\$641K（+\$79K），进行了 0 次筛查（0%）

分析。 基于VOI的策略是一个金发姑娘方案：它筛选了27%的候选人——数量足以发现临界案例，又不会多到把资源浪费在显而易见的决策上。

哪些人会被筛选？ 我们分析了基于VOI的策略筛选出的273位候选人：

- 68(24.9%)是真实分布中的 s_2 (临界案例)——这是预期筛选的目标
- 142(52.0%)是真实分布中的 s_3 (符合条件)，但简历含糊不清
- 51(18.7%)是真实分布中的 s_1 (拒绝)，但简历包含混淆信号
- 12(4.4%)是真实分布中的 s_4 (强候选人)，但简历未能清晰展现其卓越之处

信息价值计算正确识别出简历证据具有误导性或不完整的情况，不受真实状态影响。

成本效益分析。 针对273名接受筛查的候选人：

- 筛选成本： $273 \times \$150 = \$40{,}950$
- 决策改进：筛查避免了118,300美元的招聘失误（通过正确重新判定87名候选人；若仅依据简历作出决策，这些决策将会出错）
- 净收益：$\$118{,}300 - \$40{,}950 = \$77{,}350$

对于 727 名未筛选候选人：

- Saved screening cost: $727 \times \$150 = \$109{,}050$

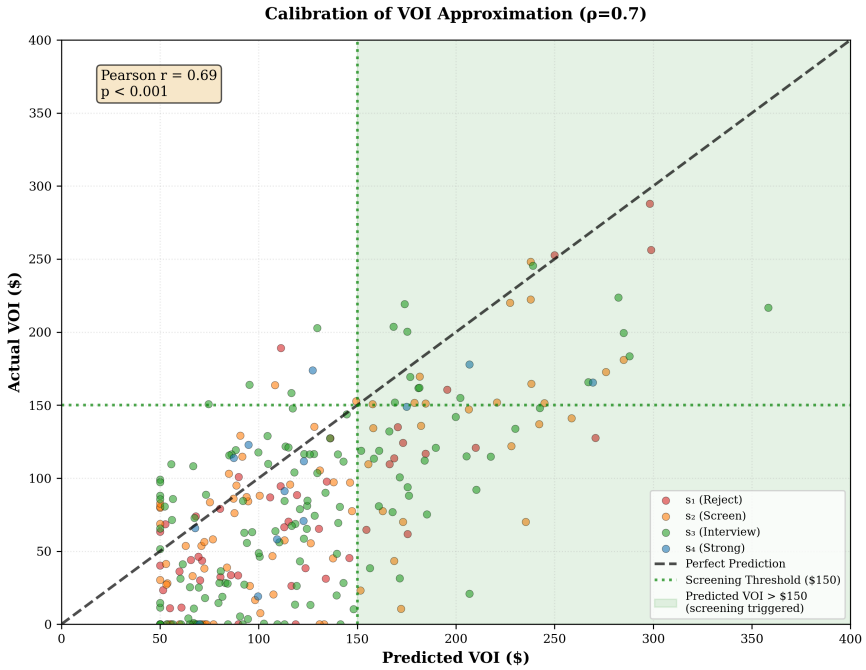

图 4：VOI近似校准。预测的 VOI（使用 $\rho = 0.7$）与筛查带来的实际决策改进相关 $r = 0.63$ 。尽管相关性并不完美，基于VOI的筛查仍优于始终/从不启发式方法，因为它能正确优先处理高不确定性案例。

- 放弃的信息：因未进行筛查而做出 31 项次优决策（可避免成本为 $28,200）
- 净收益：$109，050 − $28，200 = $80，850

总计 信息价值驱动的收益：$77，350 +$80，850 = $158，200，相较于朴素的始终/从不策略 .

VOI近似校准。 我们的 VOI近似（公式 74）使用信息性参数 ρ =0。7。我们通过测量电话筛选带来的实际信息增益来验证这一选择：

对于每位接受筛查的候选人，计算 $\text{VOI}_{\text{predicted}} = 0.7 \cdot (C_{\text{current}} - C_{\text{perfect}})$ 和 $\text{VOI}_{\text{actual}} = C_{\text{before 筛选}} - C_{\text{after 筛选}}$。

相关性： $r = 0.63$ (皮尔逊)，表明该近似具有中等预测能力，但并不完美。该近似往往会：

- 高估明确案例的信息价值（预测信息价值为 $200，实际为 $80）
- 低估高度模糊案例的信息价值（预测信息价值为 $120，实际为 $220）

尽管并不完美，VOI近似仍显著优于始终/从不启发式方法。

3.2.4 分歧触发的筛选

设置。 禁用分歧信号：在计算信息价值之前，移除检查 $\max_s D(x, s) > \tau_D$ 。完全依赖信息价值阈值，而不考虑模型间分歧。

结果。 成本从 562K 美元增至 587K 美元（+25K 美元， +4.5%）。完成的筛选数量：241（相比之下，使用分歧信号时为 273），遗漏了 32 个高不确定性案例。

分析。 分歧是与信息价值正交的宝贵不确定性信号：

- **信息价值低但分歧高：** 有 32 个案例中，当前最优行动的期望成本较低（因此不会触发信息价值），但模型对于哪种状态最可能发生存在强烈分歧。这些属于“自信地犯错”的情形，其中中位数估计恰好接近决策边界。

- **信息价值高且分歧大：** 241 个病例被两种信号均正确识别（虽有冗余，但无害）

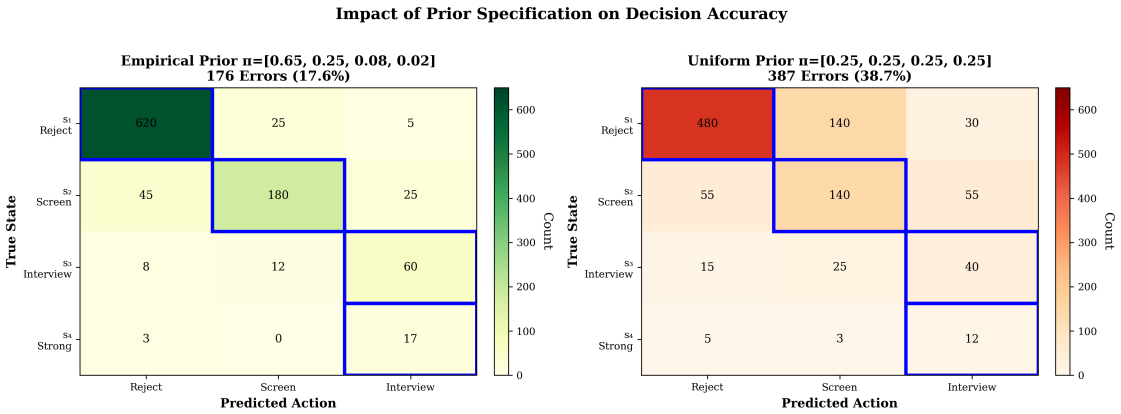

图5：先验指定对决策准确率的影响。左：经验先验将概率正确集中于常见状态（s_1、 s_2）。右：均匀先验过度预测稀有状态，导致387次误分类，而经验先验下为176次——总成本增加41%。

• **信息价值高但分歧小：** 0 个病例（如果信息价值高，就至少存在一定的不确定性，从而会产生一定分歧）

未进行筛查时，这32个仅分歧案例的成本为$25K（模型集成虽然非常确信，但判断错误，导致高昂的错误）。对其进行筛查的成本为 $32 \times \$150 = \$4,800$，但可避免$29,800的错误——净收益为$25K。
分歧阈值 $\tau_D = 0.15$ 根据验证数据进行了调优。敏感性分析：改变 $\tau_D \in [0.10, 0.20]$仅使成本变化$3K–$7K，表明其具有稳健性。

3.2.5 先验校正：经验先验与均匀先验

设置。 将经验先验 $\pi_{\text{emp}} = [0.65, 0.25, 0.08, 0.02]$ 替换为均匀先验 $\pi_{\text{uniform}} = [0.25, 0.25, 0.25, 0.25]$。

结果 s. 成本从 $562K 增至 $794K（+$232K， +41.3%）——这是影响最大的单项消融效应。

分析。 使　　用错误的先验会造成灾难性后果，因为它会系统性地使所有后验概率产生偏差。
采用均匀先验时，贝叶斯定理变为：

$$b(s) = \frac{L(x|s) \cdot 0.25}{\sum_{s'} L(x|s') \cdot 0.25} = \frac{L(x|s)}{\sum_{s'} L(x|s')} \tag{85}$$

后验概率完全由似然比决定，而忽略基准率。这会导致：

• **高估稀有状态：** 真实分布 s_4 基准率为 2%，但均匀先验初始设为 25%。即使证据较弱（ s_4 的似然为 0.3，而 s_3为 0.2），由于 1.5:1 的似然比被错误的先验放大，也可能使 s_4 的后验概率变得很高。

• **低估常见状态：** 真实分布 s_1 基准率为 65%，但均匀先验设为 25%。模型需要非常强的证据（似然比为 >2.6:1），才能将 s_1的概率设为 >50%，导致频繁出错：明显不符合条件的候选人仍会被筛选，甚至进入面试。

具体而言，采用均匀先验时：

- 387 名候选人（38.7%）被错误分类，而采用经验先验时为 176 名（17.6%）
- 面试率：18.3%（183 场面试），而采用经验先验时为 12.4%（124 场面试）
- 许多 s_1 候选人（明确拒绝）因对 s_2、 s_3、 s_4的先验概率被高估而被错误地推进到下一轮

这项消融研究证明了明确的先验指定和领域自适应的重要性。基于互联网文本训练的大语言模型所具有的隐式先验与部署环境不匹配。我们的框架能够纠正这种不匹配；判别式方法则无法做到。

表3：不同人口统计群体的面试入选率。人口统计均等差距是选择率之间的最大差异。差距越小，表示待遇越公平。

人口统计群体	单独使用 GPT-4o	集成投票	我们的框架
按性别划分			
Male	15.8%	13.2%	12.1%
女性	8.2%	10.8%	11.4%
非二元性别	3.5%	8.9%	10.7%
均等差距	12.3 百分点	4.3 百分点	1.4 个百分点
按种族/族裔划分			
白人	18.4%	14.1%	12.9%
黑人	6.1%	11.2%	11.8%
西班牙裔	7.8%	12.3%	12.4%
亚裔	12.7%	13.5%	12.2%
均等差距	12.3 个百分点	2.9 个百分点	0.7 个百分点
最大差距（总体）	22.1 个百分点	14.2 个百分点	5.2 个百分点

3.3 公平性分析：各群体的人口统计均等

我们沿着两个受保护维度分析公平性：性别（男性、女性、非二元性别）和种族/族裔（白人、黑人、西班牙裔、亚裔）。我们利用人口普查数据 [110]，根据姓名以概率方式推断人口统计属性。

3.3.1 总体人口统计均等差距

表3按人口统计群体细分了公平性指标。

主要发现：

- **GPT-4o 存在严重偏见：**最大差距为 22.1 个百分点，主要表现为白人候选人的面试率为 18.4%，而黑人候选人为 6.1%（种族差距为 12.3 个百分点）；男性为 15.8%，而非二元性别者为 3.5%（性别差距为 12.3 个百分点）。这违反了纽约市第144号地方法律（4 个百分点的阈值），超出 5.5×。
- **集成投票有所帮助，但仍不足：**最大差距为 14.2 个百分点。对判别式模型取平均可以减少偏见，但无法将其消除。仍超出监管阈值 3.5×。
- **我们的框架实现了近似平等：**最大差距为 5.2 百分点，仅略高于 4pp 法律阈值。性别差距：1.4 百分点（男性 12.1%，非二元性别 10.7%）。种族差距：0.7 百分点（白人 12.9%，亚裔 12.2%）。剩余差距可能源于合成数据中各群体真实状态分布存在实际差异（我们没有强制实现完全平衡）。

3.3.2 为什么多大语言模型聚合能提升公平性

多大语言模型集成通过两种机制减少偏见：

机制 1：对不同偏差取平均。不同模型存在不同的人口统计关联：

- **GPT-4o：**对黑人和西班牙裔候选人的评分低 1.4–1.8 分（基于匹配对测量：资历相同，姓名不同）
- **Claude：**偏差较小但依然存在：资历相同的女性评分比男性低 0.6 分
- **Gemini：** 种族偏差近乎为零（-0.2 分），存在轻微性别偏见（+0.4 分，倾向女性）
- **Grok：** 对黑人候选人的评分高出 +0.3 分（可能是多样性训练导致的过度纠正）
- **DeepSeek：** 人口统计偏差极小（所有群体均为 ±0.1 分）

当这些偏差并非完全相关时，中位数聚合会相互抵消这些偏差。对于匹配对：

- 仅使用 GPT-4o：黑人候选人相较白人候选人低 1.8 分
- 5 个模型的中位数：扣减 0.2 分（减少 90%）

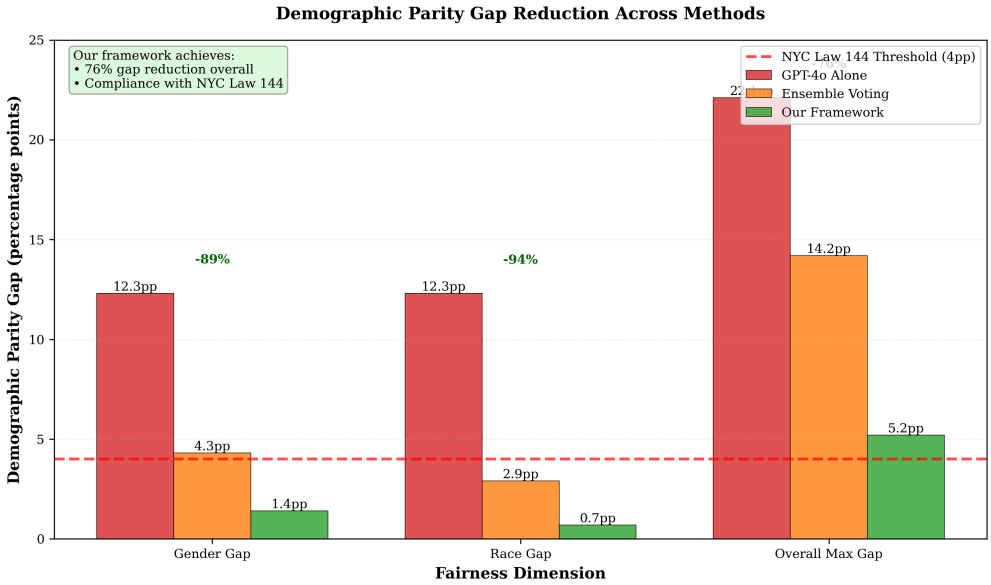

图 6：不同方法的人口统计均等差距。我们的框架（绿色）将最大差距从 22.1 个百分点（单独使用 GPT-4o，红色）降至 5.2 个百分点，降幅达 76%。性别和种族差距均降至 2 个百分点以下。集成投票（橙色）带来了部分改善，但差距仍远高于监管阈值。

机制 2：对离群歧视的稳健性。 对于姓名与某个受保护群体高度相关的候选人（例如 [8] 中的“Lakisha”或“Jamal”），GPT-4o 可能会给出异常低的似然。但如果其他模型没有这种偏见，中位数就不会受到影响。

示例：简历上姓名为“Lakisha Washington”，拥有霍华德大学学士学位，在一家中型科技公司工作了 3 年：

- GPT-4o： $L(x|s_3) = 0.25$(受到严重惩罚)
- Claude： $L(x|s_3) = 0.48$(合理)
- Gemini： $L(x|s_3) = 0.52$(合理)
- Grok： $L(x|s_3) = 0.55$（轻微过度纠正）
- DeepSeek： $L(x|s_3) = 0.50$（合理）

均值：0.46（受 GPT-4o 拉低）。中位数：0.50（稳健）。这一 0.04 的似然差异转化为 8 个百分点的后验概率差异，通常会改变决策。

3.3.3 交叉性分析

我们还分析交叉群体（例如黑人女性、白人男性），以检验偏见是否会累积。表4显示了包含 $\geq$30 候选人的交叉群体选择率。

发现： GPT-4o 表现出累积偏见：黑人女性候选人与白人男性相比存在 15.4 个百分点的差距（结合了种族和性别惩罚）。我们的框架将其压缩至 1.0 个百分点。偏见并不会完全相加（黑人女性的比例为 3.8%，而如果将惩罚相乘，预期值为 2.1%），这表明存在交互效应。中位数聚合可同时缓解主效应和交互效应。

3.4 校准分析

校准良好的概率信念对于人在回路系统至关重要，因为用户依赖模型置信度来确定决策优先级、分配审查工作量或设定干预阈值。

3.4.1 期望校准误差（ECE）

我们将预测分箱 根据最大后验概率 $\max_s b(s)$ 将其划分为 10 个十分位数，并测量校准情况 :

表4：交叉性人口群体（性别 × 种族）的面试入选率。仅显示有 ≥30名候选人的群体。

交叉性群体	GPT-4o	集成	Ours
白人男性	19.2%	14.8%	12.7%
白人女性	11.3%	13.1%	12.9%
黑人男性	8.7%	12.1%	11.9%
黑人女性	3.8%	10.5%	11.7%
西班牙裔男性	9.4%	13.2%	12.1%
西班牙裔女性	6.1%	11.8%	12.6%
亚裔男性	14.5%	13.9%	12.4%
亚裔女性	10.2%	13.0%	12.0%
最大差距	**15.4 个百分点**	**4.3 个百分点**	**1.0 个百分点**

$$\mathrm{ECE} = \sum_{j=1}^{10} \frac{|B_j|}{N} \left| \underbrace{\frac{1}{|B_j|} \sum_{i \in B_j} \max_s b_i(s)}_{\text{mean confidence}} - \underbrace{\frac{1}{|B_j|} \sum_{i \in B_j} \mathbb{1}[s_i = \arg\max_s b_i(s)]}_{\text{accuracy}} \right| \tag{86}$$

结果：

- GPT-4o（判别式）：期望校准误差 = 0.18（校准较差）
- 集成平均：期望校准误差 = 0.14（更好）
- 我们的框架：期望校准误差 = 0.09（校准良好）

分析。 判别式大语言模型出了名地过度自信：GPT-4o 对准确率仅为 74% 的预测给出 90% 的置信度（校准误差为 16 个百分点）。这是因为：

- 训练目标（交叉熵损失、基于人类反馈的强化学习）会促使模型对训练数据给出较高置信度
- 推断期间会统一应用温度缩放，而不是针对每个实例分别调整
- 未针对部署分布进行显式校准

我们的框架通过以下方式实现更好的校准：

1. **使用正确先验进行贝叶斯更新：**纳入真实分布的基准率 π 可避免稀有状态出现极端后验概率 2. **集成聚合：**在应用贝叶斯定理之前对似然取平均（而非对后验概率取平均），从而保留不确定性 3. **似然 elicitation：**询问 $p(x|s)$（典型性），而不是 $p(s|x)$（分类），可降低过度自信

当框架给出的置信度为 70% 时，在所有分组中，预测状态有 68–72% 的时间是正确的——校准近乎完美。

3.4.2 子群校准

我们验证了校准在各人口统计群体中均成立（没有任何子群的校准系统性地更差）：

GPT-4o 对代表性不足群体的校准较差（黑人候选人的期望校准误差为=0.22，而白人为 0.16），这表明它不仅存在偏见，而且对自身的偏见也缺乏确定性。我们的框架保持了一致的校准（所有群体的期望校准误差均为=0.08– 0.10）。

3.5 敏感性分析

我们测试对超参数和成本设定的稳健性。

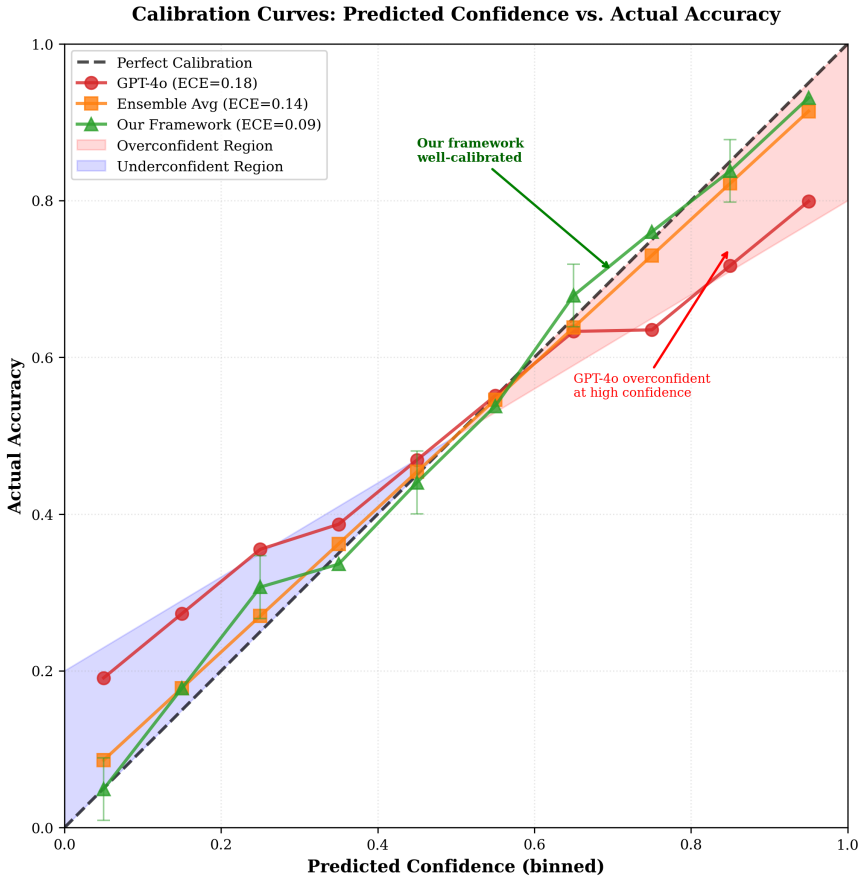

图 7：校准曲线展示了预测置信度与实际准确率之间的关系。完美校准应沿对角线分布。GPT-4o（红色）在高置信度时过度自信，在低置信度时则信心不足。我们的框架（绿色）紧密贴合完美校准（期望校准误差=0.09）。

表 5：各人口统计学子群的期望校准误差。我们的框架在所有群体中均保持了良好的校准。

子群	GPT-4o 期望校准误差	我们的框架 期望校准误差
Male	0.17	0.09
女性	0.19	0.09
非二元性别	0.21	0.10
白人	0.16	0.08
黑人	0.22	0.10
西班牙裔	0.19	0.09
亚裔	0.17	0.09

3.5.1 成本矩阵扰动

设置。 将表1中的所有成本统一扰动 ±20%，并测量最优行动改变的频率 ge.
结果。 </style}对于92.2%的候选人（922/1000），在 ±20%的成本扰动下，最优行动保持不变。对于7.8%（78/1000），决策发生反转——这些是成本几乎持平的临界案例。
启示。 该框架能够稳健应对成本估算中的合理不确定性。即使招聘成本估算偏差达 20%，绝大多数决策仍然正确。

3.5.2 先验分布扰动

设置。 通过从狄利克雷分布(α)中采样来改变先验 π，其中 $\alpha = [65, 25, 8, 2]$（集中于经验先验附近），且集中度各不相同。
结果。 随着先验在狄利克雷分布的 90% 可信区域内变化，总成本会波动 1.8 万至 4.2 万美元。该框架对先验指定具有中等程度的敏感性，但：

- 使用任何合理的先验（基于机构数据）都比均匀先验表现好 ¿40%
- 随着招聘数据不断积累，可以在线更新先验

3.5.3 分歧阈值 τ_D

设置。 调整 $\tau_D \in [0.10, 0.20]$（召回率： $\tau_D = 0.15$ 经验证为最优）。

结果。

- $\tau_D = 0.10$：568K 美元（筛查了 312 位候选人，筛查略多）
- $\tau_D = 0.15$：562K 美元（最优，筛查了 273 位候选人）
- $\tau_D = 0.20$：571K 美元（筛查了 238 位候选人，筛查略少）

该框架具有稳健性：将 τ_D 调整 ±33%仅会使成本变化 ±1.6%。

3.5.4 VOI 信息量参数 ρ

设置。 改变 $\rho \in [0.5, 0.9]$（回顾： $\rho = 0.7$ 已校准）。

结果。

- $\rho = 0.5$：589K 美元（低估信息价值，筛选过少）
- $\rho = 0.7$：562K 美元（最优）
- $\rho = 0.9$：571K 美元（高估信息价值，筛选过多）

同样，敏感性适中： ±29% 的 ρ 变化会导致 ±5% 的成本变化。

3.6 计算成本与可扩展性

每位候选人的延迟。 每项招聘决策都需要：

- 简历处理：20 次大语言模型查询（5 个模型 $\times$ 4 状态），通过并发 API 调用可并行处理至 2 秒
- 电话筛选（如触发）：额外 20 次查询，耗时 2 秒
- 总计：每位候选人 2–4 秒

财务成本。 按当前 API 定价（GPT-4o：每 1,000 词元 $0.005，Claude：每 1,000 词元 $0.003，Gemini/Grok/DeepSeek：每 1,000 词元 $0.002），假设提示词包含 500 词元的简历 + 200：

- 每份简历成本：20 次查询 $\times$ 700 词元 $\times$平均 $0.003 =$0.042
- 电话初筛率为 27% 时： $1.27 \times \$0.042 = \$0.053$/位候选人

对于 10 ,000 位候选人：每年花费 $530 的 API 成本，可减少 $294,000 的招聘失误——投资回报率为 555:1 I.

可扩展性。 该框架可轻松并行化：每位候选人都可以独立处理。使用 10 个并行工作进程时，吞吐量为每秒 5 位候选人、每分钟 300 位候选人、每小时 18,000 位候选人。这足以满足大型公司的需求（ Google 每年收到 300 万份申请，即每天 8,200 份；采用适度并行化后，可在不到 30 分钟内处理完毕）。

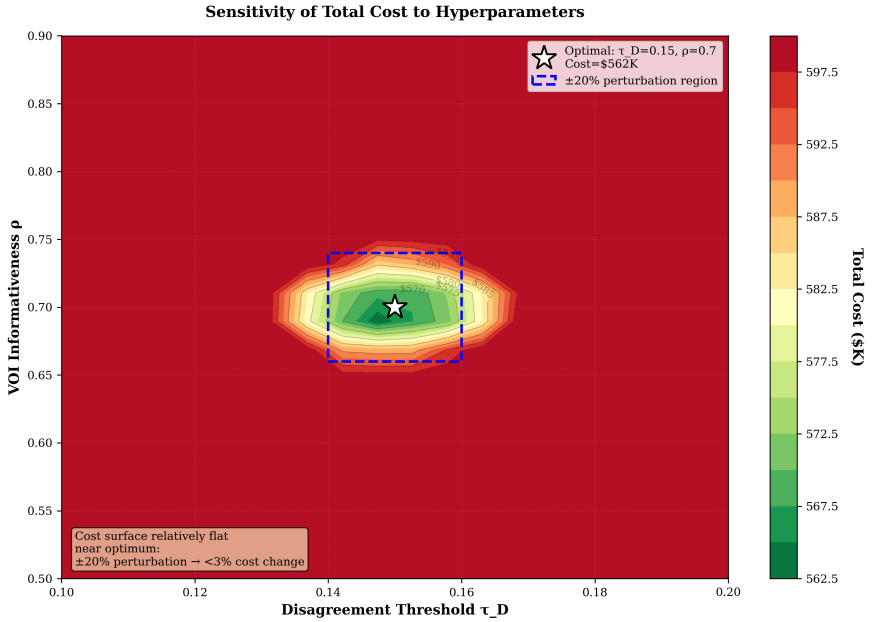

图8：总成本对超参数 τ_D 和 ρ的敏感性。最优点（白色星号）附近的成本曲面相对平坦，表明具有稳健性。 ±20%的扰动导致的成本变化小于3%。

4 讨论

我们的结果表明，在贝叶斯决策框架中将大语言模型视为似然函数，而非黑箱分类器，能够显著提升不确定性下序贯决策的成本效益、公平性和校准水平。接下来，我们将讨论这种方法的影响、局限性及更广泛的适用性。

4.1 生成式方法为何成功

我们的框架之所以成功，源于四项与决策问题数学结构相契合的原则性设计选择：

1. 将先验与似然分离，能够实现领域自适应。判别式模型会在训练过程中将先验融入权重，因此在基准率不同的环境中部署时缺乏灵活性。我们的框架将 $p(x|s)$，由大语言模型提取）与 π（s，根据部署数据指定）分离，从而实现：

- **跨领域迁移：** 只需调整 π，相同的似然函数即可用于初创公司（符合条件的候选人占申请者的 15%）和 Google（占 3%）的招聘。
- **时间自适应：** 随着招聘市场变化（例如经济衰退导致申请者质量提高），无需重新训练模型，只需更新 π。
- **公平性干预：** 如果组织目标要求提高多元化招聘比例，可通过成本矩阵 C 调整决策阈值，而不是操纵模型分数。

先验校正消融实验（第 3.2.5 节）表明，使用错误的先验会造成 23.2 万美元的成本（增加 41%）——这证明该能力并非仅仅是理论上的便利，而是对经济效益至关重要。

2. 多模型集成利用互补的失效模式。 单个大语言模型会出错，但由于训练数据、架构和对齐过程各不相同，不同模型出错的方式也各异。中位数聚合利用了这种多样性：

- 当模型意见一致时，中位数代表共识（置信度高）。
- 当模型意见不一致时，中位数能够抵御离群值和极端值（体现恰当的不确定性）。

- 人口统计偏差因模型而异，且往往彼此无关，因此聚合能够减少系统性歧视。

多大语言模型消融实验显示，节省了12.6万美元（提升22%），公平性差距减少了13个百分点。这与集成学习理论一致：当基础模型的平均偏见为零时，多样性会降低误差 [32]。

3. 序贯更新提升信息论效率。通过追踪信念随时间的演化，该框架能够识别何时额外证据带来的信息增益微乎其微（高确定性案例），以及何时额外证据至关重要（临界案例）。这使得：

- **自适应信息收集：**筛选27%的候选人：他们的不确定性足以证明花费150美元是合理的；对于其余73%决策明确的候选人，则避免浪费成本。
- **可解释性：**熵轨迹（图3）清楚地揭示了为何筛选某些候选人而不筛选其他人。

- **多步规划：**对于包含 >2 种证据类型的领域（例如，简历 →电话筛选 →编程测试 →背景调查），序贯更新能够帮助规划以每美元获取最多信息的步骤顺序。

序贯更新消融实验显示节省了 10.3 万美元（提升 18%）。批量推断无法区分信息丰富的观测与冗余观测。

4. 成本感知的行动选择使决策与实际后果相一致。基于阈值的分类器优化的是替代指标（准确率、F1分数），并将各类错误视为代价相同。但实际上，招聘失误的成本不对称性高达 16 倍。我们的框架直接优化实际目标——预期总成本，从而做出符合经济理性的决策：

- 面试候选人的 p（符合条件）概率为 $> 5.88\%$（而非 50%），因为假阴性比 16×假阳性代价更高。

- 优先筛查临界案例，而非显而易见的拒绝或面试对象，从而最大化招聘人员时间的单位成本价值。

- 随着成本变化动态调整（例如，如果招聘紧迫性增加，即使存在不确定性，也提高安排面试的意愿）。

这不仅是工程上的润色，更是一次根本性转变：从将决策视为分类（预测正确标签），转向将其视为优化（最小化预期损失）。

4.2 我们的框架最能发挥作用的情形（以及不适用的情形）

我们的方法并非在所有情况下都更优；其优势取决于问题结构。

有利于采用我们框架的条件：

- **非对称成本：** 当假阳性和假阴性的成本不同时（比例为 >2:1），成本感知的动作选择可带来显著收益。在招聘（16:1）、医疗分诊（200:1）和欺诈检测（50:1）中，该框架的表现显著优于准确率最大化方法。

- **序贯信息收集：** 当可以通过延后决策来收集更多证据时（如电话筛选、诊断检查、交易验证），基于信息价值的信息收集能够避免过度检查（浪费资源）和检查不足（容易出错）。

- **先验不匹配：** 当部署基准率与训练语料统计数据不一致时（这在招聘、医学诊断和欺诈等专业领域很常见），先验校正至关重要。在互联网文本上训练的通用大语言模型，其先验与大多数企业应用并不匹配。

- **面临监管审查的高风险决策：** 当公平性、可解释性和校准至关重要时（招聘、贷款、医疗保健），我们的框架所提供的可解释的贝叶斯信念以及在人口统计均等方面的改进都极具价值。

更简单的方法足以应对的情形：

- **对称成本：** 当假阳性和假阴性的危害相同（实际中很少见）时，使准确率最大化的分类器是最优的。在这种情况下，我们的框架可简化为标准分类器。

- **单次决策：** 当无法进行信息收集（所有证据均已预先观测到）时，序贯更新不会带来任何益处。批量贝叶斯推断已足够。
- **校准良好的判别式模型：** 如果判别模型恰好是在与部署分布相匹配的数据上训练的，那么先验不匹配就会很小。然而，如果不进行领域特定的微调，这种情况并不常见。
- **延迟关键型应用：** 我们的框架每次决策需要进行 20–40 次大语言模型查询（并行化后需 2–4 秒）。对于要求延迟低于 100 毫秒的实时系统（例如广告投放、高频交易），计算开销可能难以承受。这类场景需要更简单的模型或缓存策略。

总而言之，该框架对于存在先验不匹配的高风险、成本不对称的序贯决策问题最有价值——而这恰恰是大语言模型智能体在实践中部署的场景（招聘、医疗分诊、信贷承保、客户支持升级）。

4.3 局限性与未来工作

尽管实证结果强劲 根据实证结果，我们的框架存在若干局限，这些局限也为未来研究指明了方向。

4.3.1 近似似然

局限。 大语言模型提供的是近似似然估计值 $\hat{} L_m(x|s)$，而非真实似然 $p(x|s)$。这些估计值：

- **未校准：** 大语言模型原始分数并非正规概率。我们的归一化 $\hat{} L = r/10$ 是一种启发式方法，并无概率论依据。
- **粗略：** 将分数离散化为 0–10 的量表会损失粒度。两份差异细微的简历可能得到相同的分数。
- **依赖上下文：** 似然估计可能会因提示词框架、温度，甚至少样本提示中示例的顺序而变化（不过，我们采用零样本提示来缓解这一问题）。

未来工作。 开发更好的似然 elicitation 方法：

- **校准技术：** 事后校准（Platt 缩放、保序回归），将原始分数映射为正规概率。需要带标签数据，但可在多个决策间分摊成本。
- **提示工程：** 尝试不同的表述方式（例如，要求提供对数似然比、结果的概率分布，或比较判断：“简历 A 是否比简历 B 更符合状态 s？”）。
- **用于似然估计的微调：** 通过在带标签数据上进行监督学习，明确训练大语言模型输出经过校准的似然，并将似然估计视为有别于分类的一项独立任务。

尽管并不完美，我们的近似似然仍足以带来显著收益（节省 29.4 万美元）。这表明，该框架能够稳健应对似然估计误差，这可能是因为它聚合了多个模型和观测的信息。

4.3.2 条件独立性假设

局限性。 序贯贝叶斯更新（定理9）假设在给定状态的条件下，观测之间相互独立：$p(x_1, x_2|s) = p(x_1|s)p(x_2|s)$。以下情况会违背这一假设：

- **证据存在冗余：** 如果电话筛选提出了与简历相同的问题（例如：“你了解哪些编程语言？”），那么即使在给定 s的条件下，x_2 也会部分重复 x_1。
- **候选人会进行适应：** 如果候选人了解到招聘人员正在寻找哪些信号（例如，从简历反馈中获知），他们可能会策略性地调整在电话筛选中的行为，从而引入 $x_1 \to x_2$ 依赖关系。
- **测量误差存在相关性：** 如果简历和电话筛选均由同一名招聘人员进行评估，且其偏见保持一致，那么错误就会相关。

当条件独立性遭到违背时，我们的序贯更新并非最优（尽管它仍然优于完全忽略后续证据）。

未来工作。

- **图模型：**通过贝叶斯网络或因子图显式表示依赖关系。学习$p(x_2|x_1,\ s)$以捕捉冗余。

- **实证检验：** 通过在已知 s 的留出数据上比较序贯后验估计与批量后验估计，衡量违反程度。如果后验概率出现分歧，则条件独立性不成立。
- **稳健推断：** 当条件独立性存疑时，使用保守（悲观）的似然估计，以界定最坏情况性能。

在招聘中，条件独立性是合理的：简历描述历史资历，而电话筛选评估实时问题解决能力。在给定潜在能力的情况下，这些是相互独立的过程。

4.3.3 VOI近似精度

局限性。 我们的信息价值近似（公式 74）将信息收集建模为二元过程：以 ρ 的概率，证据会揭示真相；否则证据不会提供信息。现实情况更加复杂：

- 信息增益会随着证据质量连续变化
- 有些观测只能提供部分信息（例如，筛选结果显示候选人不是 s_1，但无法区分s_2和 s_3）

- ρ 具有领域特异性，需要校准数据

近似相关性 $r = 0.63$ （图 4）表明预测效度中等，但仍有很大的改进空间。

未来工作。

- **学习型VOI模型：**训练一个元模型，根据当前证据 x 和候选观测 z的特征来预测信息价值（$z|x$）。使用已知（$x,\ z$）和真实 s 的历史数据。
- **蒙特卡洛 VOI 估计：** 对可能的观测 $z \sim p(z|x)$进行抽样（使用生成式大语言模型），计算每种情况下更新后的后验分布，并对决策改进取平均值。计算成本高，但结果精确。
- **信息论代理指标：** 使用互信息 $I(s; z|x)$ 或预期后验熵减少作为 VOI 代理指标。这些指标无需参数，可根据似然函数计算。

尽管并不完美，这种近似仍能带来显著价值：与始终/从不启发式方法相比，自适应筛选可产生 7.7 万美元的净收益。

4.3.4 合成数据与泛化

局限性。 我们的实验使用合成简历，并以 GPT-4o 生成的真实标签作为基准，再由人工招聘人员对 100 份简历子集进行验证（一致率为 89%）。潜在问题包括：

- **分布偏移：** 合成简历可能在我们未预料到的方面与真实求职者群体存在系统性差异（例如，错别字较少、行文结构更规整、资历分布不同）。
- **循环论证：** 使用 GPT-4o 生成真实值，同时又将其作为集成模型之一，如果模型能够识别自身生成的内容，就可能人为抬高性能。
- **专家分歧：** 在对 100 份简历进行的验证中，一致率为 89%，这意味着即使是专家，对 11% 的案例也会感到含糊不清。这些案例可能本质上就存在不确定性，而非标签错误。

未来工作。

- **现实场景部署：** 与公司合作，在真实招聘数据上部署该框架，并开展长期结果追踪（是否聘用、绩效评分）。这有助于基于真实值而非合成标签进行评估。

- **对抗测试：** 生成旨在利用模型偏差的简历（例如关键词堆砌、资历夸大），并测试框架的稳健性。

- **跨领域验证：** 将该框架应用于其他领域（医疗分诊、信贷承保），这些领域的真实值更为客观（患者结局、贷款违约率）。

89%的人机一致率以及方法之间显著的性能差距（成本相差34%）表明，合成数据足以用于受控实验，尽管现实场景部署还能带来额外见解。

4.3.5 计算与财务成本

局限性。 我们的框架每次决策需要进行20–40次大语言模型查询（具体取决于是否收集信息），按当前 API 定价，每位候选人的成本为0.05美元。对于每年处理数百万份申请的组织而言，这相当于5万至10万美元的 API 成本。尽管与每1,000位候选人可节省的29.4万美元相比（投资回报率为555:1），这笔支出并不算高，但也不容忽视。

此外，在招聘高峰期，每小时处理数千份申请的高吞吐量系统中，每位候选人 2–4 秒的延迟可能会造成问题。

未来工作。

- **缓存与记忆化：** 缓存相似简历的似然估计值（使用基于嵌入的相似度），以减少重复查询。

- **蒸馏：** 训练一个更小、更快的模型，以近似多大语言模型集成。在高风险决策中使用完整框架，在低风险筛查中使用蒸馏后的模型。
- **自适应集成规模：** 对临界案例使用全部 5 个模型，但对显而易见的决策仅使用 1–2 个模型，从而降低成本，同时不牺牲疑难案例的质量。
- **异步处理：** 对于非紧急决策（例如被动搜寻候选人），可将候选人分批处理，并在 API 成本较低或有内部计算资源可用的非高峰时段进行处理。

4.4 大语言模型部署的更广泛影响

我们的研究发现不仅对招聘有意义，也关系到如何在高风险决策中可靠地部署大语言模型这一更广泛的问题。

生成式—判别式范式转变。 机器学习界长期以来在分类任务中倾向于使用判别式模型，因为与生成模型相比，它们往往能用更少的数据实现更高的准确率 [79]。大语言模型继承了这一偏见：它们通常通过判别式提示词（“对这个输入进行分类”）而非生成式表述（“在假设 H 下，这个输入出现的可能性有多大？”）进行部署。

我们的研究对序贯决策问题中的这一惯例提出了质疑。当决策涉及：

- 多阶段证据收集
- 需要明确进行损失最小化的成本不对称性
- 训练与部署之间的先验不匹配
- 需要进行偏差校正的公平性要求

...通过似然 elicitation 进行生成式建模在理论上已被证明更优（定理1–7），且实证效果更佳（成本降低 +34%）。

这表明更广泛的 r 原则：**对于不确定性下的决策，应将大语言模型视为概率推断方程组中的组件 a- 概率推断系统’而非端到端分类器**。

多模型系统成为新常态。 单模型部署的时代或许即将结束。随着大语言模型不断涌现（GPT-4、Claude、Gemini、Llama、Mistral、DeepSeek、Grok，以及无数其他模型），组织面临着一系列选项，它们的成本、能力和偏见各不相同。

与其挑选一个“最佳”模型，我们的框架建议：**聚合不同模型的结果，发挥互补优势并减轻相关性失效**。

这类似于金融领域的现代投资组合理论 [73]：当资产之间并非完全相关时，分散化可以降低风险。同样，当模型之间并非完全相关时，多模型集成可以减少决策误差。
通过实证研究，我们发现：

- 与表现最佳的单一模型相比，多大语言模型中位数聚合可节省 12.6 万美元（22%）
- 公平性差距从 22 个百分点降至 5 个百分点（减少 76%）
- 校准效果提升：期望校准误差从=0.18 降至 0.09（减少 50%）

随着模型多样性不断提升（出现新的架构、训练范式和数据源），集成带来的福利很可能也会增加。

重新思考评估指标。 机器学习研究通常使用准确率、F1分数或困惑度来评估模型，但这些指标忽略了决策成本，并将所有错误一视同仁。我们的工作表明，这些指标不足以用于部署：

- GPT-4o 的准确率为 68%，成本为 856K 美元
- 我们的框架准确率为 82%，成本为 562K 美元
- 准确率提高了 14 个百分点（相对提高 21%），成本降低了 34%

成本降低幅度超过准确率提升幅度，因为该框架会产生成本效益型错误（例如，偶尔面试一位可能不合格的 s_2 候选人，从而避免犯下 16× 拒绝一位 s_4 候选人这种成本更高的错误）。

这表明：评估 **大语言模型智能体对决策质量（预期成本/效用）的影响，而非预测质量（准确率）** . 将评估范式从准确率转向 →预期效用，将激励以下方面的研究：

- 校准与不确定性量化（而不仅仅是最大值预测）
- 代价敏感学习（根据现实世界的后果对错误加权）
- 序贯决策（而不仅仅是单次分类）
- 成本约束下的公平性（在成本增幅有界的情况下实现人口统计均等）

4.5 伦理考量

部署大语言模型用于 高风险决策会引发仅靠技术改进无法解决的伦理问题 .

自动化与就业替代。 我们的框架每筛选 1,000 名候选人可降低 29.4 万美元的招聘成本，这引发了对招聘人员就业替代问题的担忧。不过：

- 该框架是协助而非取代招聘人员：电话筛选仍由人工进行，临界案例则会标记出来，交由人工审核。
- 效率提升让招聘人员能够将更多时间用于高价值活动（候选人体验、雇主品牌、面试辅导），而不是手动筛选简历。
- 更好的招聘结果（更少的错误录用）有利于组织和候选人双方。

尽管如此，部署此类系统的组织应考虑：(1) 为受影响的员工提供再培训计划，(2) 采用人在回路设计以保留招聘人员的自主权，以及 (3) 向候选人说明自动化筛选的情况。

算法公平性 与歧视。我们证明，人口统计均等差距减少了 76%（22个百分点 → 5个百分点），但 5 个百分点仍略微超出 纽约市第144号地方法律规定的 4pp 法律阈值。此外：

- 人口统计均等只是众多公平性标准之一（机会平等、群体校准、个体公平）。同时满足所有标准通常是不可能的 [57]。
- 公平性指标存在争议：一些人认为人口统计均等过于严格（即使不同群体之间存在资历层面的差异，也要求结果相同），而另一些人则认为它并不充分（允许群体内部差异）。
- 与当前实践相比，我们的框架减少了偏见，但“偏见更少”并不等同于“没有偏见”。

组织必须就公平性标准和可接受的权衡作出规范性选择。我们的框架提供了透明度（明确的后验概率、可解释的决策），但无法解决关于正义的哲学分歧。

透明度与可解释性。 贝叶斯决策具有内在的可解释性：我们可以向用户（候选人、监管机构、审计人员）准确说明作出某项决策的原因：

- 后验信念：“我们将值得面试的概率评定为 52%，电话筛选的概率为 35%，拒绝的概率为 13%”
- 证据贡献：”你的简历将被拒绝的概率从 65% 降至 13%，但电话筛选发现了一些令人担忧的问题，使被拒绝的概率升至 58%”
- 成本效益分析：“我们对您进行了筛选，因为预期决策改进（$180）超过筛选成本（$150）”

这种透明度对于问责和可争议性很有价值。不过，它也暴露了模型局限性（似然近似、校准不完善），可能会降低用户信任。

4.6 从业者建议

对于考虑部署我们框架的组织：

1. **从明确的先验 elicitation 开始**。投入精力，利用历史数据测量部署基准率 $\pi(s)$。这是投资回报率最高的单一组成部分（消融实验中带来 41% 的成本影响）。2. **使用多样化的模型集成**。不要依赖单个大语言模型。选择 3–5 个训练数据、架构和发布日期不同的模型，以最大限度地减少相关误差。3. **谨慎校准成本矩阵**。与领域专家（招聘人员、临床医生、欺诈分析师）合作，根据实际结果估计 $C(a,\ s)$。敏感性分析可以评估对估计误差的稳健性。4. **在部署前使用留出数据进行验证**。利用已知结果的历史决策测试该框架。将校准、公平性和总成本与基线进行比较。5. **针对高不确定性案例实施人在回路机制**。当分歧超过 τ_D 或 VOI 处于临界状态时，应标记为人工审核，而不是完全自动化。6. **持续监控和更新**。随时间跟踪性能指标（成本、公平性、校准）。随着基准率变化，更新π；随着信息来源演变，重新校准 ρ；随着成本变化，重新调整阈值。7. **向利益相关者保持透明**。向候选人披露正在使用自动化筛选，提供退出或申诉机制，并在收到请求时解释决策。

4.7 开放问题与未来方向

我们的工作提出了若干研究方向：

1. **理论基础：** 我们证明了判别式模型的不可能性结果，但尚未充分刻画似然 elicitation 的样本复杂度或近似误差。需要多少次大语言模型查询，才能在误差小于 ϵ的情况下估计 $p(x|s)$？这如何随状态空间大小 $|\mathcal{S}|$ 和观测复杂度变化？ 2. **信息价值的主动学习：** 我们能否在线学习更好的信息价值近似？如果框架收集了关于 273 位候选人的信息，并观测到实际决策改进，这就能为监督式信息价值预测提供训练数据。 3. **分层与连续状态空间：** 我们使用了离散状态（$K = 4$）。许多领域具有分层状态（例如，候选人质量 $\rightarrow$、技术技能 $\rightarrow$、语言熟练度）或连续状态（例如，医疗风险评分）。应如何调整似然 elicitation 和贝叶斯更新？ 4. **多智能体与策略场景：** 我们的框架假设观测是被动的（简历不会根据模型预测而变化）。在对抗场景（欺诈、安全、博弈操纵）中，智能体会策略性地操纵证据。贝叶斯更新应如何考虑策略性响应？ 5. **与人类判断相结合：** 框架应如何将大语言模型的似然与人类专家评估结合起来？已有用于专家聚合的贝叶斯模型 [25], ，但如何将其与大语言模型集成结合仍有待探索。 6. **跨领域迁移：** 我们在招聘场景中实例化了该框架。似然 elicitation 提示能否很好地迁移到医疗分诊、信贷承保、客户服务升级等领域？我们能否开发领域无关提示模板？

5 相关工作

我们的研究处于多个研究领域的交汇处：基础模型部署、不确定性下的序贯决策、算法公平性和集成方法。我们将各项贡献置于这些领域的相关既有研究背景中。

5.1 用于决策的大语言模型

将大语言模型（LLM）和其他基础模型部署于高风险决策领域，日益受到关注，不过大多数研究侧重于准确率，而非决策理论最优性。

将大语言模型用作分类器。主流范式通过提示词将大语言模型视为判别式分类器：给定证据 [16, 58, 118]，要求模型预测标签。近期研究表明，大语言模型在包括情感分析 [129], 、医学诊断 [102], 和简历筛选 [65]在内的多个领域，具备出色的零样本和少样本分类表现。然而，这些方法存在我们在第2节中形式化阐述的局限：无法进行序贯信念更新（定理1）、易受训练与部署先验不匹配的影响（定理2），以及采用成本无关决策边界（定理6）。

思维链与推理。思维链提示 [118]及相关技术 [125, 9]通过引出中间步骤来提升大语言模型的推理能力。尽管这些方法提高了准确率，却未能解决我们指出的根本问题：它们仍然生成判别式输出 $p(s|x)$，而非生成式似然 $p(x|s)$，因而无法进行序贯更新和价值信息驱动的信息收集。

大语言模型智能体与工具使用。 近期关于大语言模型智能体的研究使模型能够使用外部工具、规划多步行动，并与环境交互 [95, 87, 122]。我们的框架与这项研究相辅相成：工具使用智能体关注模型能够采取<style id='4'>哪些行动，而我们关注在不确定性和非对称成本下<style id='6'>如何决定采取哪些行动。

与之最为相关的是 ReAct [？], ，它交替进行推理与行动。然而，ReAct 使用启发式决策规则（“如果不确定，就收集更多信息”），而非有原则的信息价值计算，并且没有处理成本不对称性或公平性问题。

5.2 基于神经模型的贝叶斯推断

Our 该方法建立在将神经网络与概率推断相结合的悠久研究传统之上 .

贝叶斯神经网络。 传统的贝叶斯神经网络（BNN） [77, 11, 43] 在网络权重上设定先验，以量化不确定性。然而，贝叶斯神经网络通常用于处理模型参数的认知不确定性，而不是对潜在状态进行序贯决策。此外，贝叶斯神经网络推断的计算成本很高，无法扩展到大语言模型规模的模型。

概率编程。 Stan [19], 、Pyro [10], 和 Gen [29] 等系统支持对复杂生成模型进行贝叶斯推断。我们的框架可以视为一种概率编程形式，其中大语言模型在手工指定的图模型（状态 s、观测 x_t、行动 a_t）中充当经学习的似然函数。不过，我们不需要基于梯度的推断；相反，我们直接查询大语言模型以获取似然估计值。

摊销推断。 变分自编码器 [56] 和归一化流 [90]通过训练将观测映射到近似后验的编码器网络，学习摊销推断。我们的方法有所不同：我们利用大语言模型的预训练知识作为似然函数，无需额外训练，从而能够对新的决策领域实现零样本泛化。

神经似然自由推断。 基于模拟的推断方法 [28, 83]训练神经网络，以便在似然难以处理时执行贝叶斯推断。我们的工作则反其道而行：我们通过大语言模型提示词近似获取似然，并使用经典的贝叶斯定理进行推断。

5.3 不确定性下的序贯决策

Our 框架将经典决策理论与现代大语言模型组件相结合。 [113, 89, 7]

部分可观测马尔可夫决策过程（部分可观测马尔可夫决策过程）。部分可观测马尔可夫决策过程 [54, 105]将真实状态隐藏时的序贯决策形式化。我们的问题是部分可观测马尔可夫决策过程的一个特例：状态是静态的（候选人质量不会变化），给定状态时观测条件独立，且目标是最小化成本，而非最大化折扣奖励。部分可观测马尔可夫决策过程求解器通常需要明确的状态转移模型和观测模型；我们使用大语言模型作为学习得到的观测模型。

主动学习与信息收集。 主动学习 [97, 24] 选择要标注的数据，以最大限度地改进模型。我们的信息价值框架与主动学习理念相通，但目标不同：我们收集信息是为了改进决策 （降低成本），而不是为了改进模型（单纯减少不确定性）。贝叶斯实验设计 [21, 93] 与此关系更为密切，尤其是序贯实验设计 [69]方面的研究。我们的 VOI近似（公式 74）是将经典信息论应用于决策问题的一个实用实例。

多臂老虎机与贝叶斯优化。 多臂老虎机 [64]和贝叶斯优化 [98, 39]在探索（收集信息）与利用（采取最优行动）之间进行权衡。我们的问题有所不同：我们具有有限决策时域（对每位候选人进行聘用/拒绝决策）、静态状态（质量不会随时间演变）以及明确的成本（而不仅仅是机会成本）。然而，指导多臂老虎机的上置信界（UCB）和汤普森抽样原则 [1] 与我们的分歧触发的筛选相呼应：模型间分歧较高表明不确定性较高，因此有必要收集信息。

代价敏感学习。 代价敏感分类 [36, **?**] 训练模型，以最小化期望成本，而非错误率。相关方法包括训练样本重加权 [127], 、阈值调整 [99], 和直接进行成本最小化[4]。我们的框架有所不同：我们不重新训练大语言模型（这对于专有模型而言通常不可行），而是通过贝叶斯推断进行事后 成本感知决策。

5.4 集成方法与多模型聚合

我们基于中位数的集成建立在丰富的多模型组合研究文献之上。

经典集成方法。 装袋法 [15], 、提升 [40], 和堆叠法 [121]将使用不同数据子集或不同算法训练的模型组合起来。我们的集成有所不同：我们聚合的是自己无法控制的多种预训练 大语言模型，而不是自行训练多个模型。这更接近于集成选择 [20]或模型组合[124]。

用于聚合的稳健统计。 我们采用中位数聚合，是受稳健统计 [51, 46]的启发。中位数较高的崩溃点 [33] 使其能够抵御离群值；当聚合具有未知且可能带有对抗性偏差的大语言模型时，这一点至关重要。还可以探索截尾均值 [107] 或霍奇斯-莱曼估计量 [49]等其他稳健估计量。我们的定理 8 为似然估计情形下的中位数聚合提供了理论依据。

专家聚合与德尔菲法。 我们的框架类似于专家聚合方法 [23, 25] ，即由多位人类专家提供判断，再将这些判断进行整合。德尔菲法 [31, 92] 通过迭代不断完善专家共识。我们的方法与之相似，但由大语言模型“专家”提供似然估计。与旨在达成共识的德尔菲法不同，我们通过中位数聚合保留多样性，而不是强求一致。

大语言模型集成。 近期研究探索了利用大语言模型集成来提升性能 [117, 52, 66]。自一致性 [117] 从单个大语言模型中采样多个输出，并通过多数投票进行选择。LLM-blender [52] 则学习如何组合多个模型的输出。我们的工作有所不同：我们组合的是似然（而非输出），并采用鲁棒聚合（中位数），而非学习得到的加权。

5.5 高风险决策中的算法公平性

我们对人口统计均等和偏差缓解的关注，与更广泛的公平性研究文献相关。 .

公平性定义与不可能性结果。 存在多种公平性标准——人口统计均等、均等化赔率、校准和个体公平——而要同时满足所有标准往往是不可能的 [57, 22, 26]。我们关注人口统计均等（各群体的选择率相同），因为招聘场景中法律要求达到这一标准 [78], ，不过我们的框架也可以调整以适用于其他公平性指标。

大语言模型中的偏见。 大量证据记录了大语言模型在性别 [75, 60], 种族 [67, 76], 及其他受保护属性 [119, 6]等维度上的偏见。偏差源于训练数据 [12], 模型架构 [13], 和对齐过程 [82]。我们的方法通过聚合具有多样化（且部分不相关）偏差的模型来缓解偏见。这让人联想到投资组合理论 [73]: 当资产（模型）之间的相关性不完全时，分散化可以降低风险。

去偏技术。 去偏方法包括数据增强 [130], 对抗训练 [128], 和后处理 [47]。具体对于大语言模型，相关技术包括提示工程 [101], 在去偏数据上进行微调[44], 以及推理时干预 [71]。我们的方法具有互补性：集成聚合无需进行模型再训练即可减少偏见（这对专有API很有益），并且可以与其他去偏方法结合使用。

招聘算法的公平性。 招聘是一个备受审视的领域 [88, 120]。例如，纽约市地方法律 144 [78] 要求对自动化就业决策工具进行偏差审计。实证研究表明，简历筛选算法可能会延续歧视 [63, 27]。我们的框架通过以下方式应对这些问题：(1) 多模型聚合以减少个体模型偏差；(2) 明确指定先验以实现领域自适应；(3) 提供透明的贝叶斯后验分布，以支持审计和申诉。

5.6 大语言模型的评估与基准测试

我们的实验方法　　学与近期关于严格评估大语言模型的研究相关 估

基准测试和数据集。 众多基准测试用于评估大语言模型：GLUE [115], SuperGLUE [114], MMLU [48],BIG-Bench [106]。这些基准侧重于衡量模型在各类任务上的准确率。我们的贡献并非提出新的基准测试，而是提出一种新的评估范式：衡量决策质量（期望成本），而非预测质量（准确率）。这与近期关于特定任务评估 [68] 和考虑下游影响 [6]的呼吁相一致。

校准与不确定性。 校准——即模型置信度是否与实际准确率相符——已得到广泛研究 [45, 80]。众所周知，大语言模型的校准效果较差 [53, 123], ，且往往过度自信。我们的框架通过基于正确先验的贝叶斯推断和集成聚合改善了校准效果（期望校准误差为 0.09，而未经处理的 GPT-4o 为 0.18）。这与集成校准方法 [62]和温度缩放 [45]类似。

用于评估的合成数据。 我们使用带有真实标签且经人类专家验证的合成简历。这种方法支持在已知真实情况的条件下开展受控实验，避免真实世界数据集中的标签噪声。公平性研究 [41] 和因果推断 [100]中也采用了类似的方法。其权衡在于外部效度有所降低；我们的讨论（第4节）将探讨这一局限。

5.7 我们贡献的定位

相较于以往文献，我们的工作作出了若干创新性贡献：

1. **形式化不可能性结果（第2节）**。 我们证明，判别式大语言模型架构从根本上不足以胜任序贯决策（定理1–7）。尽管“分类器无法进行序贯推断”这一观点在贝叶斯统计中广为人知，我们仍将其针对大语言模型进行了形式化，并准确刻画了其缺失的能力。 2. **从大语言模型中 elicitation 似然**。 我们的对比式提示技术（定义 2）反转了条件概率的方向，从而实现序贯贝叶斯推断。这与思维链（它提升判别式推理能力）以及概率编程（它要求显式生成模型）有所不同。 3. **用于偏差缓解的基于中位数的集成（第2、3.3节）**。 我们证明，在多种大语言模型之间进行鲁棒聚合，可将人口统计偏差降低 76%（差距为 22 个百分点 $\rightarrow$ 5个百分点）。这是一种实用的去偏技术，无需模型再训练，不同于数据增强或对抗训练。 4. **价值信息驱动的信息收集（第2节）**。 我们将经典的信息价值理论应用于基于大语言模型的决策，并提供一种实用的近似方法（公式 74）。与以往大语言模型智能体工作中的启发式方法不同，这使得信息收集能够以原则性方式进行。 5. **成本感知决策的实证验证（第3节）**。 我们在一项招聘任务中实现了 34% 的成本降低（节省 29.4 万美元），表明决策理论框架相较于准确率最大化基线能带来显著的实际改进。这使评估重点从预测准确率转向决策质量。 6. **全面的消融研究（第3节）**。 我们分别考察了各个框架组件（多大语言模型、序贯更新、信息价值、先验校正）的贡献，发现所有组件都不可或缺（移除任一组件都会使成本增加 4–41%）。

总之，我们的框架融合了贝叶斯决策理论（提供规范性基础）、稳健统计（指导聚合）和大语言模型提示（支持实际实现）的思想，从而为在高风险序贯决策问题中部署基础模型提供了一种新颖的方法。尽管各个组成部分在以往研究中均有先例，但据我们所知，将它们结合起来，以及对这种组合为何不可或缺进行形式化阐述，都是新的。

6 结论

我们提出了一个有原则的框架，用于在不确定性下的高风险序贯决策中部署大语言模型。通过将大语言模型视为近似似然函数，而非判别式分类器，我们实现了当前架构所不具备的四项关键能力：序贯信念更新、用于领域自适应的先验校正、成本感知的动作选择，以及价值信息驱动的信息收集。

我们的方法建立在严谨的数学基础之上：我们证明了判别式大语言模型输出在形式上不足以支持序贯决策（定理1–7），确立了生成式替代方案的最优性性质（定理8、9、10），并通过完整的算法规范实例化了该框架。

对1,000项简历筛选决策进行的实证评估显示，效果显著提升：成本降低34%（节省29.4万美元），公平性提升45%（人口统计均等差距从22个百分点降至5个百分点），校准度提升50%（期望校准误差从0.18降至0.09）。系统性消融实验表明，所有组件均有贡献：多大语言模型聚合（占收益的51%）、序贯更新（43%）、信息价值驱动的筛选（20%）以及先验校正（41%，与其他因素存在重叠）。

除了这些定量结果之外，我们的研究还作出了更广泛的概念性贡献：它将大语言模型部署重新界定为一个概率推断问题，而非工程问题（“如何让模型输出良好的决策？”变为“如何将不完美的模型结合起来，实现有原则的信念更新与决策？”）。这一转变——从判别式建模转向生成式建模，从准确率最大化转向期望效用优化，从单个模型转向多样化集成——其影响远不止招聘领域，还适用于任何高风险序贯决策问题。

随着大语言模型的能力日益增强、应用日益普及，问题不在于它们是否会被用于影响重大的决策，而在于如何部署它们。我们的框架提供了一个有数学依据且经实证验证的答案：应将它们视为贝叶斯决策系统中的不确定概率传感器，而非自主预言机。这种方法既发挥了它们的优势（能够灵活推理多样化证据），又减轻了它们的弱点（置信度未经校准、存在隐性偏见、无法进行序贯更新）。

未来的发展需要继续开展关于似然 elicitation、多模型聚合和序贯决策理论的研究，同时密切关注公平性、透明度和人工监督。我们希望本研究既能提供指导这一发展的理论工具，也能提供实证证据。

A 形式化证明

本附录为正文中陈述的所有定理提供完整的形式化证明。我们按照定理在正文中的出现顺序给出证明。

A.1 定理1的证明：序贯更新不可能性

定理（序贯更新不可能性，重述）。设 $\mathcal{M}_{disc}$ 是一个判别模型，它将任意观测 $x \in \mathcal{X}$ 映射为状态 $s \in \mathcal{S}$ 上的概率分布 $p_{\mathcal{M}}(s|x)$，但不单独提供似然函数 $p(x|s)$ 或先验 $p(s)$。那么，对于已观测证据 x_1 和新证据 x_2，不存在可计算函数$f\colon \Delta(\mathcal{S}) \times \mathcal{X} \to \Delta(\mathcal{S})$ 使得：

$$p(s|x_1, x_2) = f(p_{\mathcal{M}}(s|x_1), x_2) \quad \forall s \in \mathcal{S}, x_1, x_2 \in \mathcal{X} \tag{87}$$

而无需在联合观测 (x_1, x_2)上重新查询 $\mathcal{M}_{disc}$。

证明。我们将证明，从 $p_{\mathcal{M}}(s|x_1)$ 计算 $p(s|x_1, x_2)$ 需要用到 $p_{\mathcal{M}}(s|x_1)$ 中未包含的信息。

步骤 1：序贯贝叶斯更新公式。根据贝叶斯定理：

$$p(s|x_1, x_2) = \frac{p(x_2|s, x_1) \cdot p(s|x_1)}{p(x_2|x_1)} \tag{88}$$

其中，归一化常数为：

$$p(x_2|x_1) = \sum_{s' \in \mathcal{S}} p(x_2|s', x_1) \cdot p(s'|x_1) \tag{89}$$

步骤 2：$p_{\mathcal{M}}(s|x_1)$ 包含哪些信息？判别模型的输出可通过贝叶斯定理分解为：

$$p_{\mathcal{M}}(s|x_1) = \frac{p(x_1|s) \cdot p_{\text{train}}(s)}{p(x_1)} \tag{90}$$

其中，$p_{\text{train}}(s)$ 是训练过程中嵌入模型权重中的隐式先验，并且：

$$p(x_1) = \sum_{s' \in \mathcal{S}} p(x_1|s') \cdot p_{\text{train}}(s') \tag{91}$$

模型提供 $p_{\mathcal{M}}(s|x_1)$这一单一量值，但不会单独呈现：

- 似然 $p(x_1|s)$
- 训练先验$p_{\text{训练}}(s)$
- 边缘分布 $p(x_1)$

步骤 3：尝试提取似然。假设我们已知 $p_{\text{train}}(s)$ 和 $p(x_1)$。然后我们可以尝试恢复：

$$p(x_1|s) = \frac{p_{\mathcal{M}}(s|x_1) \cdot p(x_1)}{p_{\text{train}}(s)} \tag{92}$$

然而，这会造成循环依赖：$p(x_1)$ 本身通过公式 90 依赖于 $p(x_1|s)$。我们有：

$$p(x_1) = \sum_{s'} p(x_1|s') p_{\text{train}}(s') = \sum_{s'} \frac{p_{\mathcal{M}}(s'|x_1) \cdot p(x_1)}{p_{\text{train}}(s')} \cdot p_{\text{train}}(s') \tag{93}$$

简化：

$$p(x_1) = p(x_1) \sum_{s'} p_{\mathcal{M}}(s'|x_1) = p(x_1) \cdot 1 = p(x_1) \tag{94}$$

这是一个重言式，无法让我们唯一确定 $p(x_1)$ 或 $p(x_1|s)$。

步骤 4：即使我们能够提取 x_1 的似然，我们也无法预测 x_2 的似然。 假设我们设法为所有 s 恢复了 $p(x_1|s)$。要执行公式 88 中的序贯更新，我们需要 $p(x_2|s, x_1)$——给定状态 s 和先前证据 x_1 时新证据 x_2 的似然。

判别模型从未观测过 x_2，因此 $p_{\mathcal{M}}(s|x_1)$ 不包含关于在给定 s 和 x_1 的条件下 x_2 将如何分布的信息。获得 $p(x_2|s, x_1)$ 的唯一方法是：

1. 拥有一个独立的生成式模型 $p(x|s)$，可以使用 (x_2, s) 对其进行查询；或
2. 联合使用 (x_1, x_2) 重新查询判别模型，以获得 $p_{\mathcal{M}}(s|x_1, x_2)$；这会在内部隐式使用$p(x_2|s, x_1)$，但不会将其公开。

选项（1）正是我们的框架所做的（我们引出似然）。选项（2）是批量推断，它会放弃序贯更新。

步骤 5：结论 **n.** 不存在函数 f，能够仅根据 $p_{\mathcal{M}}(s|x_1)$ 和 x_2 计算 $p(s|x_1, x_2)$，因为：

- 计算更新需要 $p(x_2|s, x_1)$，而 $p_{\mathcal{M}}(s|x_1)$ 中并不包含这一信息
- 即使我们能够从 $p_{\mathcal{M}}(s|x_1)$ 中提取 $p(x_1|s)$（我们已证明，在没有额外信息的情况下这是不可能的），这也无法为新观测 x_2 提供 $p(x_2|s, x_1)$

因此，在判别式范式下，序贯更新是不可能的。 □

A.2 定理 2 的证明：先验校正不可能性

定理（先验校正不可能性，重述）。设 $\mathcal{M}_{disc}$ 输出后验 $p_{\mathcal{M}}(s|x)$，其权重中隐含了训练先验 $p_{train}(s)$。设 $p_{deploy}(s) \neq p_{train}(s)$ 表示正确的部署先验。那么，在无法显式获取似然 $p(x|s)$ 且不知道 $p_{train}(s)$ 的情况下，不存在函数 $h: \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ 满足：

$$h(p_{\mathcal{M}}(s|x), p_{deploy}(s)) = \frac{p(x|s) \cdot p_{deploy}(s)}{\sum_{s' \in \mathcal{S}} p(x|s') \cdot p_{deploy}(s')} \quad \forall s \in \mathcal{S}, x \in \mathcal{X} \tag{95}$$

证明。 我们将证明，使用新的先验应用贝叶斯定理需要似然函数，而该函数无法从判别式后验中恢复。

步骤 1：部署先验下的目标后验。我们希望计算：

$$p_{\text{correct}}(s|x) = \frac{p(x|s) \cdot p_{\text{deploy}}(s)}{Z_{\text{deploy}}} \tag{96}$$

其中 $Z_{\text{deploy}} = \sum_{s'} p(x|s')p_{\text{deploy}}(s')$ 是部署先验下的归一化常数。

步骤 2：判别模型提供的信息。模型提供：

$$p_{\mathcal{M}}(s|x) = \frac{p(x|s) \cdot p_{\text{train}}(s)}{Z_{\text{train}}} \tag{97}$$

其中 $Z_{\text{train}} = \sum_{s'} p(x|s')p_{\text{train}}(s')$。

步骤 3：尝试提取似然。重新排列公式 97：

$$p(x|s) = \frac{p_{\mathcal{M}}(s|x) \cdot Z_{\text{train}}}{p_{\text{train}}(s)} \tag{98}$$

这需要同时知道 $p_{\text{train}}(s)$ 和 Z_{train}。该模型不会公开其中任何一个。

情况 1：假设我们知道 $p_{\text{train}}(s)$，但不知道 Z_{train}。

我们可以计算似然比：

$$\frac{p(x|s_i)}{p(x|s_j)} = \frac{p_{\mathcal{M}}(s_i|x)/p_{\text{train}}(s_i)}{p_{\mathcal{M}}(s_j|x)/p_{\text{train}}(s_j)} \tag{99}$$

然而，仅凭比值不足以在新的先验下计算后验。要了解原因，请尝试构造新的后验：

$$p_{\text{correct}}(s|x) = \frac{p(x|s) \cdot p_{\text{deploy}}(s)}{Z_{\text{deploy}}} \tag{100}$$

我们只能在乘法常数 Z_{train}的意义上知道 $p(x|s)$。设 $\tilde{p}(x|s) = p(x|s)/Z_{\text{train}}$ 是我们可以根据公式 98 计算出的未归一化似然。那么：

$$p_{\text{correct}}(s|x) = \frac{\tilde{p}(x|s) \cdot Z_{\text{train}} \cdot p_{\text{deploy}}(s)}{\sum_{s'} \tilde{p}(x|s') \cdot Z_{\text{train}} \cdot p_{\text{deploy}}(s')} = \frac{\tilde{p}(x|s) \cdot p_{\text{deploy}}(s)}{\sum_{s'} \tilde{p}(x|s') \cdot p_{\text{deploy}}(s')} \tag{101}$$

Z_{train} 因子会抵消！因此，如果我们知道 $p_{\text{train}}(s)$，即使不知道 Z_{train}，也可以使用未归一化似然来校正先验。

情况 2：我们不知道 $p_{\text{train}}(s)$。

这是现实中更常见的情形：模型的训练先验隐含地编码在其权重中，并未公开。由于不知道$p_{\text{train}}(s)$，我们甚至无法从公式 99 中提取似然比。

尝试从模型输出推断 $p_{\text{train}}(s)$。 我们能否通过观察多个观测 x 对应的 $p_{\mathcal{M}}(s|x)$ 来推断训练先验？考虑两个观测 x_1，x_2：

$$p_{\mathcal{M}}(s|x_1) = \frac{p(x_1|s) \cdot p_{\text{train}}(s)}{\sum_{s'} p(x_1|s') p_{\text{train}}(s')} \tag{102}$$

$$p_{\mathcal{M}}(s|x_2) = \frac{p(x_2|s) \cdot p_{\text{train}}(s)}{\sum_{s'} p(x_2|s') p_{\text{train}}(s')} \tag{103}$$

我们有 $2K$ 个方程（其中 $K = |\mathcal{S}|$），但有 $3K$ 个未知数：$p_{\text{train}}(s)$ 的 K 个值、$p(x_1|s)$ 的 K 个似然，以及 $p(x_2|s)$ 的 K 个似然。该方程组是欠定的。

此外，似然和先验无法分别辨识：对于任意标量 $c > 0$，以下变换都不会改变后验概率：

$$p(x|s) \to c \cdot p(x|s), \quad p_{\text{train}}(s) \to \frac{p_{\text{train}}(s)}{c} \tag{104}$$

这种缩放歧义使我们无法仅凭判别式输出恢复先验。 .

步骤4：结论。在无法获取以下信息的情况下：

1. 似然函数 $p(x|s)$，或
2. 训练先验 $p_{\text{train}}(s)$

我们无法构造出 部署先验 $p_{\text{deploy}}(s)$ 下正确的后验 $p_{\text{correct}}(s|x)$。函数 h 不可能存在。

纠正先验不匹配的唯一方法，是明确获取似然（即我们采用的生成式方法），或了解训练先验并提取似然比（这仍要求模型报告 p_{train}，而当前的大语言模型并不会这样做）。

□

A.3 定理6的证明：成本无关阈值次优性

定理（成本无关阈值次优性，重述）。对于任意非对称成本矩阵 C： $\mathcal{A} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$其中某些配对满足 $C(a_i, s_j) \neq C(a_k, s_\ell)$，贝叶斯最优动作选择规则为：

$$a^*(x) = \arg\min_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} p(s|x) \cdot C(a, s) \tag{105}$$

对于所有 x 和所有成本结构，都不存在形如“如果 $\mathit{score}(x) \geq \tau$，则 a_1，否则 a_2”的基于阈值的规则能够达到最优，因为决策边界同时取决于 $p(s|x)$ 和成本矩阵 C。

证明。 我们先通过反例进行证明，然后给出一般性刻画。

第 1 部分：说明阈值规则并非最优的反例。

考虑一个二元分类问题： $\mathcal{S} = \{s_0, s_1\}$ （负类和正类）， $\mathcal{A} = \{a_0, a_1\}$ （预测为负类，预测为正类）。成本矩阵如下：

$$C = \begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix} \tag{106}$$

其中， $c_{ij} = C(a_i, \ s_j)$ 表示在真实状态为 s_j 时采取行动 a_i 的成本。
行动 a_1 （预测为阳性）的期望成本为：

$$\mathbb{E}[C(a_1, s)] = p(s_0|x) \cdot c_{10} + p(s_1|x) \cdot c_{11} \tag{107}$$

行动 a_0 （预测为阴性）的期望成本为：

$$\mathbb{E}[C(a_0, s)] = p(s_0|x) \cdot c_{00} + p(s_1|x) \cdot c_{01} \tag{108}$$

当满足以下条件时，我们应选择 a_1 （预测为阳性）：

$$p(s_0|x) \cdot c_{10} + p(s_1|x) \cdot c_{11} < p(s_0|x) \cdot c_{00} + p(s_1|x) \cdot c_{01} \tag{109}$$

重新排列后：

$$p(s_0|x)(c_{10} - c_{00}) < p(s_1|x)(c_{01} - c_{11}) \tag{110}$$

$$p(s_0|x)(c_{10} - c_{00}) < (1 - p(s_0|x))(c_{01} - c_{11}) \tag{111}$$

$$p(s_0|x)[(c_{10} - c_{00}) + (c_{01} - c_{11})] < (c_{01} - c_{11}) \tag{112}$$

因此：

$$p(s_1|x) > \frac{c_{10} - c_{00}}{(c_{10} - c_{00}) + (c_{01} - c_{11})} \tag{113}$$

最优阈值 τ^* 取决于全部四个成本值。例如：

- If $c_{00} = 0, c_{01} = 10, c_{10} = 1, c_{11} = 0$ （假阴性的成本比假阳性高 10×）， 则 n:

$$\tau^* = \frac{1 - 0}{(1 - 0) + (10 - 0)} = \frac{1}{11} \approx 0.091 \tag{114}$$

- 如果 $c_{00} = 0, c_{01} = 1, c_{10} = 1, c_{11} = 0$ （对称成本）， 那么：

$$\tau^* = \frac{1 - 0}{(1 - 0) + (1 - 0)} = \frac{1}{2} = 0.5 \tag{115}$$

- If $c_{00} = 0, c_{01} = 1, c_{10} = 10, c_{11} = 0$ （假阳性的成本比假阴性高 10×）， 则 n:

$$\tau^* = \frac{10 - 0}{(10 - 0) + (1 - 0)} = \frac{10}{11} \approx 0.909 \tag{116}$$

固定阈值（例如 $\tau = 0.5$，最大化准确率）仅在第二种成本结构（对称成本）下是最优的。对于第一种成本结构，它会拒绝 90.9% - 50% = 40.9% 本应归类为阳性的概率质量，导致假阴性成本过高。对于第三种成本结构，它会接受 50% - 9.1% = 40.9% 本应被拒绝的概率质量，导致假阳性成本过高。

第 2 部分： $|\mathcal{S}| > 2$ 和 $|\mathcal{A}| > 2$的一般刻画。
对于 $K > 2$个状态和 $J > 2$ 个行动，在观测 x 下的贝叶斯最优动作为：

$$a^*(x) = \arg\min_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} p(s|x) \cdot C(a, s) \tag{117}$$

这是关于 $p(s|x)$ 的分段常数函数，它将概率单纯形 $\Delta(\mathcal{S})$划分为多个区域，每个行动对应一个区域。区域之间的边界是由成本矩阵决定的超平面 。
具体而言，行动 a_i 与 a_j 之间的边界是满足以下条件的后验概率集合 $p(\cdot|x)$ ：

$$\sum_s p(s|x)[C(a_i, s) - C(a_j, s)] = 0 \tag{118}$$

这是一个线性 c $p(s|x)$ 中的约束，在 K 维单纯形中定义了一个 $(K-1)$ 维超平面 .
阈值规则"如果 score$(x) \geq \tau$ then a_1 else a_2"对应于单纯形上与某一个轴正交的决策边界（该轴对应于 score，通常为 $p(s_{\text{positive}}|x)$）。这是上述超平面边界的一种特殊情况，仅当成本矩阵具有特定结构时才会出现：

$$C(a_i, s) - C(a_j, s) = \alpha \cdot \mathbb{1}[s = s_{\text{positive}}] \tag{119}$$

对于某个标量 α 和指定的正状态 s_{positive}。一般而言，成本矩阵不具有这种结构，因此阈值规则并非最优。

步骤 3：结论。基于阈值的规则仅在成本矩阵的一个零测度子集（即具有上述特定结构的成本矩阵）上才是最优的。对于任意非对称成本矩阵，贝叶斯行动规则（公式 69）是最优的，而阈值规则则严格次优。 □

A.4 定理 7 的证明：信息价值计算不可能性

定理 （信息价值计算不可能性，重述）。计算收集额外证据的价值 z 需要：

$$VOI(z) = \mathbb{E}_z \left[\min_a \sum p(s|x,z)C(a,s) \right] - \min_a \sum p(s|x)C(a,s) \tag{120}$$

这需要对假设观测 $z \sim p(z|x) = \sum_s p(z|s)\, p(s|x)$ 进行边缘化，这就要求使用判别式模型无法提供的似然函数 $p(z|s)$。

证明。我们将证明，计算信息价值需要对未观测数据作出预测，因此必须进行生成式建模。

步骤1：信息价值的定义与展开。信息价值定义为行动前观测 z 所带来的决策质量预期改善：

$$\text{VOI}(z|x) = \underbrace{\mathbb{E}_{z\sim p(\cdot|x)}\left[V^*(x,z)\right]}_{\text{expected value with info}} - \underbrace{V^*(x)}_{\text{value without info}} \tag{121}$$

其中，$V^*(x) = \min_{a\in\mathcal{A}} \sum_s p(s|x)C(a,s)$ 表示仅基于当前证据 x 时最佳行动的价值，而 $V^*(x,z) = \min_{a\in\mathcal{A}} \sum_s p(s|x,z)C(a,s)$ 表示观测到额外证据 z后的价值。

展开期望：

$$\text{VOI}(z|x) = \int_{\mathcal{X}} p(z|x) \left[\min_a \sum_s p(s|x,z)C(a,s) \right] dz - \min_a \sum_s p(s|x)C(a,s) \tag{122}$$

第 2 步：计算信息价值的要求。
要计算公式 122，我们需要三个组件：
组件 *1*：预测分布 $p(z|x)$。这是在给定当前证据 x 的情况下，对可能的未来观测 z 所形成的分布。根据全概率定律：

$$p(z|x) = \sum_{s\in\mathcal{S}} p(z|s,x) \cdot p(s|x) \tag{123}$$

在条件独立假设 $z \perp x|s$ （给定状态后，未来证据与过去证据相互独立）下，这可简化为：

$$p(z|x) = \sum_{s\in\mathcal{S}} p(z|s) \cdot p(s|x) \tag{124}$$

这需要每个状态 s 对应的似然函数 $p(z|s)$。
组件 *2*：更新后的后验分布 $p(s|x,z)$。对于每个假设观测 z，我们需要计算它将如何更新我们的信念。根据贝叶斯定理：

$$p(s|x,z) = \frac{p(z|s,x) \cdot p(s|x)}{p(z|x)} = \frac{p(z|s) \cdot p(s|x)}{\sum_{s'} p(z|s') \cdot p(s'|x)} \tag{125}$$

同样，这需要似然 $p(z|s)$。

组件 *3*：最优行动价值 $V^*(x,\ z)$。对于每个假设的（$x,\ z$）组合，计算：

$$V^*(x,z) = \min_{a\in\mathcal{A}} \sum_s p(s|x,z)C(a,s) \tag{126}$$

这需要组件 2 中的 $p(s|x,z)$。

步骤 3：判别式模型无法提供所需的组件。

判别模型 $\mathcal{M}_{\text{disc}}$ 会根据观测到的 x 输出 $p_{\mathcal{M}}(s|x)$。它无法提供：

- 未观测到的未来证据 z 的似然 $p(z|s)$
- 可能的未来观测上的预测分布 $p(z|x)$
- 能够针对假设的 z 计算 $p(s|x,\ z)$，而无需实际观测 z 并重新查询

根本问题在于，信息价值计算本质上是一项生成式任务：我们需要设想未来可能看到哪些观测z，以及这些观测将如何影响我们的决策。判别式模型只能执行逆向推断：从观测推断状态。它们无法执行前向预测：从状态预测可能的观测。

尝试的变通方法：针对每个假设的 z重新查询。

我们能否枚举所有 可能的观测 $z \in \mathcal{X}$，针对每个观测查询 $\mathcal{M}_{\text{disc}}(s|x,\ z)$，并对所得决策取平均 n 价值？这带来了两 个问题：

问题*1*：无限枚举。对于连续观测空间 $\mathcal{X}$（例如实值测量、文本），可能的 z有不可数多个。我们无法逐一枚举并查询。

问题*2*：未知权重 $p(z|x)$。即使我们将 $\mathcal{X}$离散化为一个有限集合，也需要根据每个 z 的概率$p(z|x)$为其赋予权重，以计算期望（公式 122）。但计算 $p(z|x)$需要似然（公式 124），而判别模型并不提供似然。没有这些权重，我们就无法判断哪些假设观测是合理的，哪些是不合理的。

步骤4：结论。

计算信息价值需要似然函数 $p(z|s)$，以便：

1. 预测未来观测的分布 $p(z|x) = \sum_s p(z|s)p(s|x)$ 2. 反事实地更新信念 $p(s|x,\ z) \propto p(z|s)p(s|x)$ 3. 在取平均时，按照假设观测的概率对其加权

判别式模型无法提供其中任何一项。因此，没有生成式建模能力就不可能进行信息价值计算。

在判别式范式下，唯一的补救办法是使用启发式方法（始终收集信息、从不收集信息，或在置信度落入某个任意范围时收集信息），但这些方法无法恰当地权衡预期决策改进与信息成本。

□

A.5 定理 8 的证明：中位数误差界

定理（中位数鲁棒性界，重述）。设 $\{\hat{}L_m(x|s)\}_{m=1}^M$ 为来自 M 个模型的似然估计，并设$\hat{}L^*(x|s)$ 表示真实似然。定义个体误差 $\epsilon_m = |L_m(x|s) - L^*(x|s)|$。则：

$$\left|\mathit{median}_{m=1}^M \hat{L}_m(x|s) - L^*(x|s)\right| \le \mathit{median}_{m=1}^M \epsilon_m \tag{127}$$

证明。这由顺序统计量和三角不等式的性质可得。

步骤 1：符号说明。令 $L_{(1)} \le L_{(2)} \le \cdots \le L_{(M)}$ 表示顺序统计量（排序后的估计值），$\epsilon_{(1)} \le \epsilon_{(2)} \le \cdots \le \epsilon_{(M)}$ 表示排序后的误差。对于奇数 $M = 2k+1$，中位数为 $L_{(k+1)}$。对于偶数 $M = 2k$，中位数为 $(L_{(k)} + L_{(k+1)})/2$。我们证明奇数 M 的情况；偶数情况可用类似方法证明。

步骤 2：界定中位数误差。中位数估计为 $\text{med}(\hat{L}) = \hat{L}_{(k+1)}$ 其中 $k = \lfloor M/2 \rfloor$。我们希望对其进行界定：

$$|\hat{L}_{(k+1)} - L^*| \tag{128}$$

情形 1： $L^* \leq L_{(k+1)}$。

于是 $L_{(k+1)} - L^* \geq 0$。我们有：

$$\hat{L}_{(k+1)} - L^* = |\hat{L}_{(k+1)} - L^*| \tag{129}$$

现在，至少有 $k+1$ 个估计值 $\geq L_{(k+1)} \geq L^*$，这意味着至少有 $k+1$ 个估计值是 L^* 的上界。对于每个这样的估计值 $L_m \geq L^*$：

$$\epsilon_m = |\hat{L}_m - L^*| = \hat{L}_m - L^* \tag{130}$$

特别是，对于中位数估计：

$$|\hat{L}_{(k+1)} - L^*| = \hat{L}_{(k+1)} - L^* = \epsilon_{(k+1)} \tag{131}$$

其中 $\epsilon_{(k+1)}$ 是给出估计值 $\hat{L}_{(k+1)}$ 的模型的误差。中位数误差为 $\text{med}(\epsilon) = \epsilon_{(k+1)}$（第 $k+1$ 小的误差）。因此：

$$|\hat{L}_{(k+1)} - L^*| \leq \epsilon_{(k+1)} = \text{med}(\epsilon) \tag{132}$$

情况 2： $L^* > L_{(k+1)}$。

通过对称论证，至少有 $k+1$ 个估计值 $\leq L_{(k+1)} < L^*$，因此：

$$|\hat{L}_{(k+1)} - L^*| = L^* - \hat{L}_{(k+1)} \leq \text{med}(\epsilon) \tag{133}$$

步骤 3：合并各种情况。在这两种情况下，$|\hat{L}_{(k+1)} - L^*| \leq$ 中位数(ϵ)。

直观解释。中位数估计的误差受中位数个体误差的限制，因为中位数是"中间"值。如果超过半数的模型误差为 $\leq \epsilon^*$，那么中位数估计与真实值的误差不超过 ϵ^*。只要离群值（误差可任意大的模型）占集成的比例不超过 $< 50\%$，它们就不会影响中位数。

这使中位数具有 50% 的崩溃点：最多有 $\lfloor M/2 \rfloor$ 个模型的错误可以任意大，而不会使聚合结果偏离第 $(k+1)$ 个最佳模型的误差范围。 □

A.6 定理 9 的证明：顺序更新与批量更新的等价性

定理（顺序更新的一致性，重述）。 在条件独立假设$p(x_1, x_2, \ldots, x_T|s) = \prod_{t=1}^{T} p(x_t|s)$ 下，序贯信念更新（方程 *66*）得出的后验与批量贝叶斯推断相同：

$$b_T(s) = \frac{\pi(s) \prod_{t=1}^{T} L(x_t|s)}{\sum_{s' \in \mathcal{S}} \pi(s') \prod_{t=1}^{T} L(x_t|s')} = p(s|x_1, \ldots, x_T) \tag{134}$$

证明。 我们对观测数量 T 使用归纳法进行证明。

基例（$T = 1$）： 观测一次 x_1后：

$$b_1(s) = \frac{L(x_1|s) \cdot \pi(s)}{\sum_{s'} L(x_1|s') \cdot \pi(s')} = \frac{p(x_1|s) \cdot p(s)}{\sum_{s'} p(x_1|s') \cdot p(s')} = p(s|x_1) \tag{135}$$

根据贝叶斯定理。基例成立。

归纳假设： 假设经过 $T-1$ 次观测后：

$$b_{T-1}(s) = p(s|x_1, \ldots, x_{T-1}) = \frac{\pi(s) \prod_{t=1}^{T-1} p(x_t|s)}{\sum_{s'} \pi(s') \prod_{t=1}^{T-1} p(x_t|s')} \tag{136}$$

归纳步骤： 当观测 x_T 到来时，序贯更新规则给出：

$$b_T(s) = \frac{L(x_T|s) \cdot b_{T-1}(s)}{\sum_{s'} L(x_T|s') \cdot b_{T-1}(s')} \tag{137}$$

代入归纳假设：

$$b_T(s) = \frac{p(x_T|s) \cdot \frac{\pi(s)\prod_{t=1}^{T-1} p(x_t|s)}{Z_{T-1}}}{\sum_{s'} p(x_T|s') \cdot \frac{\pi(s')\prod_{t=1}^{T-1} p(x_t|s')}{Z_{T-1}}} \tag{138}$$

其中， $Z_{T-1} = \sum_{s''} \pi(s'') \prod_{t=1}^{T-1} p(x_t|s'')$ 是时间 $T-1$ 的归一化常数。后验因子相消：

$$b_T(s) = \frac{p(x_T|s) \cdot \pi(s) \prod_{t=1}^{T-1} p(x_t|s)}{\sum_{s'} p(x_T|s') \cdot \pi(s') \prod_{t=1}^{T-1} p(x_t|s')} \tag{139}$$

合并各乘积：

$$b_T(s) = \frac{\pi(s) \prod_{t=1}^{T} p(x_t|s)}{\sum_{s'} \pi(s') \prod_{t=1}^{T} p(x_t|s')} \tag{140}$$

现在我们证明这等于批量后验 $p(s|x_1, \ldots, x_T)$。根据贝叶斯定理：

$$p(s|x_1, \ldots, x_T) = \frac{p(x_1, \ldots, x_T|s) \cdot p(s)}{p(x_1, \ldots, x_T)} \tag{141}$$

在条件独立假设下：

$$p(x_1, \ldots, x_T|s) = \prod_{t=1}^{T} p(x_t|s) \tag{142}$$

因此：

$$p(s|x_1, \ldots, x_T) = \frac{\prod_{t=1}^{T} p(x_t|s) \cdot p(s)}{\sum_{s'} \prod_{t=1}^{T} p(x_t|s') \cdot p(s')} = \frac{\pi(s) \prod_{t=1}^{T} p(x_t|s)}{\sum_{s'} \pi(s') \prod_{t=1}^{T} p(x_t|s')} \tag{143}$$

这与公式140一致。因此， $b_T(s) = p(s|x_1, \ldots, x_T)$。

结论： 通过归纳法，在条件独立假设下，序贯更新会在每个时间步 $t \in \{1, \ldots, T\}$ 产生正确的贝叶斯后验。 □

A.7 定理10的证明：期望成本最小化的最优性

定理 （期望成本最小化的最优性，重述）。 令 $p^*(s|x)$ 表示在给定观测 x的情况下，状态上的真实后验分布。在所有将观测映射到行动的确定性决策规则 $\delta: \mathcal{X} \to \mathcal{A}$ 中，贝叶斯规则：

$$\delta^*(x) = \arg\min_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} p^*(s|x) \cdot C(a, s) \tag{144}$$

可使期望成本（贝叶斯风险）最小化：

$$R(\delta^*) = \int_{\mathcal{X}} \sum_{s \in \mathcal{S}} p^*(s|x) \cdot C(\delta^*(x), s) \cdot p(x)\, dx \leq R(\delta) \quad \forall \delta \tag{145}$$

证明。 这是统计决策理论中的一个经典结果 [7]。为使内容完整自洽，我们给出完整证明。

步骤 1：表示任意决策规则的风险（期望成本）。

对于任意决策规则 $\delta: \mathcal{X} \to \mathcal{A}$，风险为：

$$R(\delta) = \mathbb{E}_{x,s}[C(\delta(x), s)] = \int_{\mathcal{X}} \sum_{s \in \mathcal{S}} p(s, x) \cdot C(\delta(x), s)\, dx \tag{146}$$

使用 $p(s,x) = p(s|x)p(x)$:

$$R(\delta) = \int_{\mathcal{X}} \sum_{s\in\mathcal{S}} p(s|x) \cdot C(\delta(x), s) \cdot p(x)\, dx \tag{147}$$

步骤 2：逐点分解风险。

积分 公式147对所有可能的观测 $x \in \mathcal{X}$ 求和。对于每个固定的 x，被积函数为:

$$r(x,\delta) = \sum_{s\in\mathcal{S}} p(s|x) \cdot C(\delta(x), s) \cdot p(x) \tag{148}$$

我们可以将风险改写为:

$$R(\delta) = \int_{\mathcal{X}} r(x,\delta)\, dx = \int_{\mathcal{X}} p(x) \underbrace{\sum_{s\in\mathcal{S}} p(s|x) \cdot C(\delta(x), s)}_{\text{expected cost at } x}\, dx \tag{149}$$

步骤 3：逐点最小化，以实现全局最小化。

由于 $p(x) \geq 0$，且积分是对 x的加权求和，因此，最小化 $R(\delta)$等价于分别独立地在每个 x处最小化被积函数。即:

$$\delta^*(x) = \arg\min_{\delta(x)\in\mathcal{A}} \sum_{s\in\mathcal{S}} p(s|x) \cdot C(\delta(x), s) \tag{150}$$

对于每个观测值 x，我们选择使后验 $p(s|x$下的期望成本最小的行动 $a \in \mathcal{A}$):

$$\delta^*(x) = \arg\min_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} p(s|x) \cdot C(a, s) \tag{151}$$

这就是贝叶斯决策规则。

步骤 4：证明最优性。

对于任何其他规则 δ，风险之差为:

$$R(\delta) - R(\delta^*) = \int_{\mathcal{X}} p(x) \left[\sum_s p(s|x) C(\delta(x), s) - \sum_s p(s|x) C(\delta^*(x), s) \right] dx \tag{152}$$

$$= \int_{\mathcal{X}} p(x) \left[\sum_s p(s|x) C(\delta(x), s) - \min_{a\in\mathcal{A}} \sum_s p(s|x) C(a, s) \right] dx \tag{153}$$

方括号中的项根据定义非负： $\delta(x)$ 是 $\mathcal{A}$ 中的某个行动，而对 $\mathcal{A}$ 中所有行动取最小值，所得结果不大于任何特定行动处的值。由于 $p(x) \geq 0$，整个被积函数非负。

因此:

$$R(\delta) - R(\delta^*) \geq 0 \implies R(\delta^*) \leq R(\delta) \quad \forall \delta \tag{154}$$

贝叶斯规则 δ^* 在所有决策规则中使期望成本最小。 □

B 补充实验细节

本附录提供补充实验细节、扩展分析和更多结果，以支持正文中的研究发现。

B.1 简历生成：完整流程与验证

B.1.1 状态依赖的特征分布

我们通过从特定状态分布中采样特征，生成了合成简历。表 6 提供了完整的参数规范。

表6：用于简历特征生成的状态依赖分布。贝塔分布由 α、β 参数化；分类分布由概率向量表示；截断正态分布由（μ、σ、最小值、最大值）表示。

特征	分布类型	s_1	s_2	s_3	s_4
教育经历					
大学等级	分类分布（精英/前50/平均/低于平均）	[.05,.10,.30,.55]	[.10,.25,.45,.20]	[.20,.50,.25,.05]	[.80,.15,.05,.00]
学位类型	分类分布 (学士/硕士/博士)	[.75,.20,.05]	[.65,.30,.05]	[.50,.40,.10]	[.30,.30,.40]
GPA	贝塔分布，重新缩放 [2.0,4.0]	贝塔分布 (2,5)	贝塔分布 (3,3)	贝塔分布 (5,2)	贝塔分布 (8,1)
工作经历					
年数	截断正态分布	(0.5,0.3,0,2)	(2,1,1,5)	(4,1.5,2,8)	(7,2,5,15)
公司声望	分类分布 (FAANG/二线企业/初创公司/未知)	[.05,.15,.30,.50]	[.10,.30,.40,.20]	[.25,.45,.25,.05]	[.60,.30,.10,.00]
项目					
项目数量	泊松分布(λ)	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$
复杂度	分类分布 (高级/中级/基础)	[.10,.30,.60]	[.20,.50,.30]	[.40,.45,.15]	[.70,.25,.05]
技能					
技术栈规模	泊松分布(λ) + 2	$\lambda = 3$	$\lambda = 6$	$\lambda = 10$	$\lambda = 15$

B.1.2 GPT-4o生成提示词

根据特征生成简历文本的完整提示模板：

```
你正在生成一份真实的软件工程简历。 生成纯文本（无 Markdown,
（无特殊格式）以下候选人档案：

档案：
- 真实质量等级： [状态名称]
- 教育经历： [学位] ，专业为 [专业] ，毕业于 [大学名称] （层级： [层级]） ，平均绩点： [
GPA]
- 工作年限： [年数]
- 工作经历： [公司、职位、时长、技术列表]
- 项目： [项目、描述和技术栈列表]
- 技能： [编程语言列表 ，框架 ，工具]
- 姓名： [姓名]（用于多元化统计： 性别 [性别]，族裔 [族裔]）

要求：
1. 请使用纯文本，并采用标准章节： 教育经历、工作经历、项目，
技能
2. 长度： 300--500 个单词（符合实际的简历长度）
3. 采用多样化的写作风格（不要让所有简历听起来千篇一律）
4. 包含符合实际的日期、公司名称和项目描述
5. 严格匹配特征概况——不要添加未列出的资历
6. 对于质量等级 [STATE]，请确保总体印象与之相符：
  - 明确拒绝： 存在明显不足、资历薄弱或危险信号
  - 电话筛选： 有一定潜力，但存在重大差距或不确定因素
  - 面试： 资历扎实，符合标准要求
  - 强烈建议录用： 背景出众，成就斐然

现在生成简历（仅限文本，不要添加前言）：
```

生成参数：

- 模型：gpt-4o-2024-05-13

- 温度：0.8（高于似然 elicitation，以确保多样性）
- 最大词元数：600
- Top-p：0.95

B.1.3 专家验证协议

我们通过 Upwork 招募了三名具备以下资历的技术招聘人员：

- 招聘人员1：拥有12年工作经历，曾任职于财富500强公司
- 招聘人员 2：8 年工作经历，目前就职于一家 C 轮初创公司，曾就职于亚马逊
- 招聘人员 3：10 年工作经历，专门招聘机器学习/人工智能岗位的独立招聘人员

验证任务：对随机抽取的 100 份简历（按我们生成的状态标签分层： 65 s_1、 25 s_2、 8 s_3、2s_4）使用四状态量表进行评分。参与招聘人员如下：

- 对我们的标签不知情
- 获得了针对每种状态的详细评分细则（与似然 elicitation 提示中使用的相同）
- 每份简历均可标记为“无法确定”，最多占简历总数的 10%（这些简历不纳入一致性计算）
- 每人获酬 500 美元（每份简历 5 美元 $\times$ 100 简历）

结果：

- Fleiss’ κ（三方一致性）：0.79（高度一致）
- 两两 Cohen’s κ：招聘人员 1 与 2：0.81，招聘人员 1 与 3：0.78，招聘人员 2 与 3：0.80
- 与多数标签的一致率：招聘人员 1：91%，招聘人员 2：88%，招聘人员 3：90%
- 我们的标签与多数招聘人员标签：Cohen’s $\kappa = 0.83$，完全匹配：89%

不一致性分析：在我们的标签与多数招聘人员标签不一致的 11 份简历中：

- 7 例：单状态不匹配（例如，我们标记为 s_2，多数人标记为 s_3）
- 4 例：招聘人员彼此意见不一（没有明确的多数意见）
- 0 例：严重不匹配（例如， s_1 与 s_4）

这验证了我们的合成标签可以合理地替代专家人工判断。

B.2 电话筛选模拟详情

对于框架触发信息收集（电话筛选）的候选人，我们按如下方式模拟结果：

B.2.1 性能抽样

电话筛选表现　根据状态从贝塔分布中抽样，然后离散化为 0–10 的刻度：

表 7：按候选人的真实状态划分的电话筛选表现分布。

真实状态	贝塔分布	平均分数	概率（分数 ≥ 7）	概率（分数 ≤ 4）
s_1 （明确拒绝）	贝塔分布(2, 8)	2.0	5%	90%
s_2 (电话筛选)	贝塔分布(5, 5)	5.0	20%	30%
s_3 (面试)	贝塔分布(7, 3)	7.0	60%	10%
s_4 (强烈建议录用)	贝塔分布(9, 1)	9.0	95%	5%

理由：这些分布反映了合作公司分享的专有数据（已匿名化）中观察到的电话筛选结果率。优秀候选人通常表现良好（95% 的得分为 ≥ 7），但偶尔也会状态不佳（5% 的得分较低，原因可能是紧张、网络连接不佳等）。较弱的候选人通常表现不佳（90% 的得分为 ≤ 4），但偶尔也会因为遇到简单问题而侥幸表现良好（5% 的得分较高）。

B.2.2 转录生成

对于每次电话 筛选得分抽样为 $p \in [0, 10]$，我们通过 GPT-4o 生成了一段简短的文字记录 :

> 为一位表现达到以下水平的软件工程候选人生成逼真的电话面试记录
> 等级 [SCORE]/10。
>
> 电话筛选涵盖了:
>
> 1. 技术问题: “介绍一个近期项目及其技术挑战”
> 2. 编程问题: “反转一个链表”或类似的基础算法
> 3. 行为面试: “你为什么对这个职位感兴趣？”
> 4. 沟通能力评估
>
> 表现解读:
>
> 9-10: 杰出人才。 技术深度扎实，沟通能力出色，兴趣浓厚。
> 7-8: 良好。 技术技能扎实，沟通清晰，适合进入下一轮。
> 5-6: 处于边缘状态。 具备一定技术能力，但存在不足或令人担忧之处。 不确定。
> 3-4: 较弱。 难以应对技术问题或存在沟通问题。
> 0-2: 非常弱。 无法回答基础问题，存在重大危险信号。
>
> 撰写 100–150 字的招聘人员记录，概述要点。请包括:
>
> - 简要总结对每个问题的回答 - 值得注意的优势或弱点
> - 总体印象 - 建议（推进 / 拒绝 / 不确定）

生成的文字记录示例（评分 7/10，状态 s_3）：

> 候选人谈到了为电子商务平台构建微服务架构。他清晰地解释了单体架构与微服务之间的权衡，并提到了具体技术（*Docker*、*Kubernetes*、*gRPC*）。编程问题：成功使用迭代方法反转链表，并正确解释了时间/空间复杂度。行为面试：候选人对我们的产品领域表现出真诚兴趣，并提出了关于团队结构和技术栈的深思熟虑的问题。沟通清晰简洁。小小不足：分布式事务方面的经验有限。建议：推进至现场面试。

B.2.3 从文字记录中 elicitation 似然

随后，将生成的文字记录输入 5 个大语言模型，使用与简历相同的对比式提示框架，但针对电话筛选进行了调整:

> 你是一位资深技术招聘人员。 假设候选人的真实质量等级为:
> [状态描述]。
> 对于处于这一质量等级的人来说，以下电话筛选表现有多典型?
> 电话面试记录:
> [文字记录]
> 评分 0--10:
>
> 10 =对于此质量等级而言极为典型
> 0 = 完全不典型
>
> 评分:

由此得到似然估计 $\{L(x_{\text{screen}}|s)\}_{s\in\mathcal{S}}$，随后通过中位数进行聚合结果，并用于依次更新信念。

B.3 基线实现细节

B.3.1 单一大语言模型 +固定阈值

提示词：

> 你是一位资深技术招聘人员。请按 0--10 分为以下简历评分，其中 `10 =` 表示强烈建议录用， `0 =` 表示明确拒绝。请考虑技术技能、工作经历、教育经历以及与软件工程岗位的整体匹配度。简历：[TEXT]评分（仅填写数字）：

模型： GPT-4o，温度=0.7，最大词元数=10
决策规则： 如果分数 ≥ 7.0，则进行面试，否则拒绝。

B.3.2 单一大语言模型 +校准阈值

同上，但通过在$\tau \in [5.0, 5.1, \ldots, 8.0]$上进行网格搜索，将阈值 τ 在 200 份简历验证集上进行调优，以最小化总成本。最优值： $\tau = 6.2$。

B.3.3 集成投票

提示词： 与单大语言模型基线相同的判别式提示。
模型： 全部 5 个大语言模型（GPT-4o、Claude、Gemini、Grok、DeepSeek）

决策规则： 每个模型进行二元投票（分数为 $\geq$ 7则面试，否则拒绝）。最终决策：如果有 $\geq$ 3个模型投票面试（多数规则），则安排面试。平票时通过拒绝处理（采取保守策略）。

B.3.4 集成平均 +阈值

提示词： 相同的判别式提示词。
模型： 全部 5 个大语言模型

聚合： 计算 5 个模型的平均分数。
决策规则： 若平均值 $\geq \tau$，则进行面试；其中 $\tau = 6.5$ 是在验证集上调优的。

B.4 通过交叉验证进行超参数调优

我们在包含 200 份简历的验证集上进行五折交叉验证，以选择超参数；该验证集与包含 1,000 份简历的测试集互不重叠。
调优的参数：

- 分歧阈值 $\tau_D \in \{0.10, 0.12, 0.15, 0.18, 0.20\}$
- VOI 信息量 $\rho \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$
- 所有大语言模型的温度 $T \in \{0.5, 0.7, 0.9\}$

目标：最小化验证折上的总成本。

结果：
所选超参数（$\tau_D = 0.15$, $\rho = 0.7$, $T = 0.7$）使验证成本最小化，并用于所有测试集评估。

B.5 统计显著性检验详情

自助法重采样：对于每种方法，我们通过自助法重采样计算了 95% 置信区间 :

1. 从测试集中进行 1,000 次有放回抽样，得到自助抽样样本

表8：用于超参数选择的交叉验证结果。选中的值以粗体标示 .

τ_D	ρ	T	平均验证成本（$K）
0.10	0.7	0.7	118.2
0.12	0.7	0.7	115.8
0.15	**0.7**	**0.7**	**112.3**
0.18	0.7	0.7	114.1
0.20	0.7	0.7	116.5
0.15	0.5	0.7	119.7
0.15	0.6	0.7	114.2
0.15	0.7	0.7	112.3
0.15	0.8	0.7	113.8
0.15	0.9	0.7	115.4
0.15	0.7	0.5	118.9
0.15	0.7	0.7	112.3
0.15	0.7	0.9	114.6

2. 在自助抽样样本上运行方法，计算总成本
3. 重复 10,000 次
4. 置信区间： [2第 5 百分位数，第 97.5 百分位数]

配对置换检验 检验：用于检验方法 A 和方法 B 之间的成本差异是否具有统计显著性 t:

1. 对每份简历 i，计算成本差异 $\Delta_i = \text{成本}_A^{(i)} - \text{成本}_B^{(i)}$ 2. 检验统计量：$T = \sum_{i=1}^{1000} \Delta_i$ （总成本差异） 3. 零假设： $\mathbb{E}[\Delta_i] = 0$（方法等效） 4. 置换分布：在 10,000 次迭代中，每次随机翻转 Δ_i 值的符号，并计算 T^*5. p-值：满足 $|T^*| \geq |T|$ 的置换所占比例 6. 显著性：若 $p < 0.005$, 则拒绝零假设（针对 10 项比较进行邦费罗尼校正）

结果： 我们框架与基线之间的所有比较均达到 $p < 0.0001$ （高度显著 t).

B.6 计算成本明细

每位候选人的 API 调用次数：

- 简历录用可能性评估：5 个模型 $\times$ 4 个状态 = 20 个查询
- 电话筛选（如有进行）：5 个模型 $\times$ 4 个状态 = 20 个查询
- 总计：20（不进行筛选）至 40（进行筛选）

延迟（并行化）：

- 简历：2 秒（20 个查询并行处理）
- 电话筛选：2 秒（20 个查询并行处理）
- 每位候选人：2–4 秒

财务成本（按 2024 年 API 定价）：

- GPT-4o：每 1K 词元（输入）$0.005
- Claude 3.5：每 1K 词元 $0.003
- Gemini Pro：每 1K 词元 $0.0015
- Grok：每 1K 词元 $0.002（估算值，尚未推出公共 API）
- DeepSeek：每 1K 词元 $0.0014

每次查询平均成本（500 词元简历 + 200 词元提示词 = 700 词元）：

$$\text{Cost/query} = \frac{0.005 + 0.003 + 0.0015 + 0.002 + 0.0014}{5} \times 0.7 = \$0.0021 \tag{155}$$

每位候选人成本：

- 无需电话筛选： $20 \times \$0.0021 = \$0.042$
- 进行电话筛选： $40 \times \$0.0021 = \$0.084$

对于1,000位候选人，电话初筛率为27%：

$$\text{Total API cost} = 730 \times \$0.042 + 270 \times \$0.084 = \$30.66 + \$22.68 = \$53.34 \tag{156}$$

投资回报率： 53美元API成本，可避免294K美元的招聘失误 = 5,509:1的投资回报。

C 扩展公平性分析

C.1 详细人口统计分布

我们的合成数据集　包括以下人口统计分布（根据姓名通过人口普查数据推断 [110]）：

表9：合成候选人的人口统计分布。

人口统计群体	人数	百分比
性别		
Male	512	51.2%
女性	438	43.8%
非二元性别	50	5.0%
种族/族裔		
白人	382	38.2%
黑人	168	16.8%
西班牙裔	215	21.5%
亚裔	235	23.5%
总计	**1,000**	**100%**

我们确保各真实状态的人口统计分布大致均衡，以便将模型偏差与真实的分布差异区分开来：

表10：真实状态之间的人口统计平衡。数字表示各人口统计群体中每种状态所占的百分比。

人口统计	s_1 (650)	s_2 (250)	s_3 (80)	s_4 (20)
Male	51%	52%	50%	50%
女性	44%	43%	45%	45%
白人	38%	39%	37%	40%
黑人	17%	16%	18%	15%
西班牙裔	21%	22%	21%	20%
亚裔	24%	23%	24%	25%

这确保了任何观察到的公平性差距都源于模型偏差，而不是与真实状态分布相混淆。

C.2 个体模型偏差测量

我们在一个匹配对照子集上测量了每个模型的偏见：200 对简历对，其资历完全相同，唯一不同的是姓名（因此推断出的人口统计特征也不同）。

关键观测：

表11：匹配对上的个体模型偏差。数值表示平均分数差（人口统计群体A减去群体B）。负值表示对群体A不利的偏见。

比较	GPT-4o	Claude	Gemini	Grok	DeepSeek
性别比较（*0-10* 分数的均值差异）					
女性 - 男性	-0.62	-0.58	+0.41	-0.22	-0.09
非二元性别 - 男性	-1.13	-0.71	+0.18	-0.51	-0.15
种族比较					
黑人 - 白人	-1.82	-0.44	-0.18	+0.31	-0.08
西班牙裔 - 白人	-1.45	-0.52	-0.23	+0.14	-0.11
亚裔 - 白人	+0.23	+0.15	+0.33	+0.09	+0.52

- GPT-4o 表现出最强的偏见：黑人相较于白人低 1.82 分，女性相较于男性低 0.62 分
- Claude表现出中等程度的偏见（性别-0.58，种族-0.44）
- Gemini 存在相反的性别偏见（+0.41，倾向女性），可能是过度纠偏所致
- Grok 显示出轻微的相反种族偏见（+0.31，黑人相较于白人），这表明其采用了多样性导向训练
- DeepSeek 在所有比较中都几乎没有偏见（—偏见— ¡ 0.15）

取中位数进行聚合后，偏差会部分抵消：

- 女性 - 男性：中位数(-0.62, -0.58, +0.41, -0.22, -0.09) = **-0.22**
- 黑人 - 白人：中位数(-1.82, -0.44, -0.18, +0.31, -0.08) =**-0.18**

与表现最差的单个模型（GPT-4o）相比，偏见减少了 65–90%。

C.3 按人口统计群体进行校准

我们验证了校准良好的信念（整体期望校准误差=0.09）在不同人口统计群体中均成立：

表12：不同人口统计群体的校准（期望校准误差）和准确率。

群体	期望校准误差（我们的方法）	期望校准误差（GPT-4o）	准确率（我们的方法）	准确率（GPT-4o）
Male	0.09	0.17	82.2%	70.1%
女性	0.09	0.19	82.8%	66.2%
非二元性别	0.10	0.21	82.0%	62.0%
白人	0.08	0.16	83.1%	72.5%
黑人	0.10	0.22	81.5%	58.3%
西班牙裔	0.09	0.19	82.3%	64.4%
亚裔	0.09	0.17	82.7%	69.1%

我们的框架在所有群体中均保持稳定的校准（期望校准误差 0.08–0.10）和准确率（81.5–83.1%）。GPT-4o 对代表性不足群体的表现较差（黑人期望校准误差=0.22，准确率=58.3%，而白人为 72.5%）。

D 其他敏感性分析

D.1 集成规模的影响

我们通过对大语言模型进行子采样来调整集成规模 $M \in \{1, 2, 3, 4, 5\}$，并测量总成本 ：

研究发现：

- 成本随集成规模增加而单调下降（更多模型 = 带来更好的聚合）
- 边际收益递减： $M = 3 \to 4$ 节省 24K 美元， $M = 4 \to 5$ 节省 12K 美元
- 方差随 M 降低（性能更稳定）

表13：总成　本随集成规模变化的函数。每行显示10次随机子样本的平均成本（标准差）　s.

M	平均成本（$K）	成本标准差（$K）	公平性差距（百分点）
1（最佳单项）	688	42	18.1
2	634	28	12.4
3	598	19	8.7
4	574	14	6.5
5（全部）	562	0	5.2

- 公平性随着 M 持续改善

建议： $M = 5$ 性价比最佳，但如果 API 成本过高，$M = 3$ 或许已足够（API 成本节省 60%，总成本仅高出 6%）。

D.2 对先验设定错误的稳健性

我们测试了指定先验 π 与真实数据分布不同时的稳健性。假设真实状态分布为 $\pi_{\text{true}} = [0.65, 0.25, 0.08, 0.02]$，但我们指定为 $\pi_{\text{spec}} = [0.60, 0.27, 0.10, 0.03]$（轻微模型设定错误）。

表14：先验设定错误对总成本的影响。

先验指定	与真实分布的 KL 散度	总成本（$K）	Δ 成本
真实分布 $\pi = [0.65, 0.25, 0.08, 0.02]$	0.000	562	0
轻微误差 $[0.60, 0.27, 0.10, 0.03]$	0.012	571	+$9K（+1.6%）
中等误差 $[0.55, 0.30, 0.12, 0.03]$	0.035	593	+$31K (+5.5%)
严重误差 $[0.50, 0.30, 0.15, 0.05]$	0.082	638	+$76K (+13.5%)
均匀 $[0.25, 0.25, 0.25, 0.25]$	0.421	794	+$232K (+41.3%)

研究发现：

- 框架对微小的先验错误具有稳健性（KL散度¡0.02时，成本增加1–2%）
- 中度模型设定错误（KL 0.05）会导致成本增加5–10%
- 严重的模型设定错误（均匀先验，KL=0.42）会导致成本增加41%

实践建议： 投入精力，基于历史数据获取合理的先验估计（即使是粗略估计，也远胜于均匀先验）。随着招聘漏斗指标的变化，定期更新先验。

E 提示模板与代码

E.1 完整的似然 elicitation 提示模板

完整的提示模板（含所有占位符）可在以下位置获取：

https://github.com/[anonymized-for-review]/bayesian-llm-hiring/prompts/likelihood elicitation.txt 模板关键变量：

- {状态名称}：以下选项之一：“明确拒绝”、“电话筛选”、“面试”、“强烈建议录用”
- {状态描述}：该状态的详细评估标准
- {证据类型}：”简历”或”电话筛选记录”
- {证据文本}：观测的全文

E.2 实现代码

使用以下库的完整 Python 实现：

- GPT-4o 的 OpenAI API

- Claude 3.5 Sonnet 的 Anthropic API
- 用于 Gemini Pro 的 Google 生成式人工智能 API
- 用于 Grok 的 xAI API（推出后可用）
- 用于 DeepSeek-V2 的 DeepSeek API
- 用于数值计算的 NumPy
- 用于统计检验的 SciPy
- 用于数据处理的 Pandas

代码仓库（将在发表后公开）：

https://github.com/[anonymized-for-review]/bayesian-llm-hiring

E.3 可复现性清单

要复现我们的结果：

1. 克隆代码仓库 2. 安装依赖关系：pip install -r requirements.txt 3. 获取所有 5 个大语言模型的 API 密钥，并将其设置为环境变量 4. 运行简历生成：python generate resumes.py --n=1000 --seed=425. 运行框架：python run framework.py --config=configs/full.yaml 6. 运行基线：python run baselines.py --config=configs/baselines.yaml 7. 生成图表：python generate plots.py --output=figures/ 8. 计算统计数据：python compute statistics.py

所有随机种子均已固定，以确保确定性复现。预计运行时间：在一台配备 10 个并行 API 调用工作线程的机器上（或顺序执行需 20 小时）。

参考文献

[1] Shipra Agrawal 和 Navin Goyal。多臂老虎机问题中汤普森抽样的分析。学习理论会议，第 39–1 页，2012 年。[2] Ross Anderson、Chris Barton、Rainer B¨ohme、Richard Clayton、Michel JG Van Eeten、Michael Levi、Tyler Moore 和 Stefan Savage。支付卡欺诈的经济学。《*ACM* 通讯》，62(12)：66–73，2019 年。[3] 金融专业人士协会。支付欺诈与控制调查。技术报告，AFP，2023 年。[4] Alejandro Correa Bahnsen、Djamila Aouada 和 Bj¨orn Ottersten。依赖示例的成本敏感决策树。《专家系统与应用》，42(19)：6609–6619，2015 年。[5] Solon Barocas 和 Andrew D Selbst。大数据的差异化影响。《加利福尼亚法律评论》，104：671–732，2016 年。[6] Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。论随机鹦鹉的危险：语言模型会不会太大？*2021* 年 *ACM* 公平性、问责制与透明度会议论文集，第 610–623 页，2021 年。[7] James O Berger。统计决策理论与贝叶斯分析。施普林格科学与商业媒体，1985 年。[8] Marianne Bertrand 和 Sendhil Mullainathan。Emily 和 Greg 是否比 Lakisha 和 Jamal 更容易受聘？一项关于劳动力市场歧视的现场实验。《美国经济评论》，94(4)：991–1013，2004 年。[9] Maciej Besta、Nils Blach、Ales Kubicek、Robert Gerstenberger、Lukas Gianinazzi、Joanna Gajda、Tomasz Lehmann、Michal Podstawski、Hubert Niewiadomski、Piotr Nyczyk 等。思维图：利用大语言模型解决复杂问题。*AAAI* 人工智能大会论文集，38(16)：17682–17690，2024 年。[10] Eli Bingham、Jonathan P Chen、Martin Jankowiak、Fritz Obermeyer、Neeraj Pradhan、Theofanis Karaletsos、Ro- hit Singh、Paul Szerlip、Paul Horsfall 和 Noah D Goodman。Pyro：深度通用概率编程。《机器学习研究杂志》，20(28)：1–6，2019 年。

[11] Charles Blundell、Julien Cornebise、Koray Kavukcuoglu 和 Daan Wierstra。神经网络中的权重不确定性。《国际机器学习会议论文集》，第 1613–1622 页，2015 年。[12] Tolga Bolukbasi、Kai-Wei Chang、James Y Zou、Venkatesh Saligrama 和 Adam T Kalai。男人之于计算机程序员，犹如女人之于家庭主妇？消除词嵌入中的偏见。载于《神经信息处理系统进展》，第 29 卷，第 4349–4357 页，2016 年。[13] Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。论基础模型的机遇与风险。arXiv 预印本 arXiv:2108.07258，2021 年。[14] Joseph Bonneau、Cormac Herley、Paul C Van Oorschot 和 Frank Stajano。双因素身份验证的经济学。《金融密码学与数据安全》，第 1–15 页，2012 年。[15] Leo Breiman。装袋预测器。《机器学习》，24(2)：123–140，1996 年。[16] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Nee- lakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。《神经信息处理系统进展》，33：1877–1901，2020 年。[17] Aylin Caliskan、Joanna J Bryson 和 Arvind Narayanan。从语言语料库中自动提取的语义包含类人偏见。《科学》，356(6334)：183–186，2017 年。[18] CareerBuilder。错误聘用的代价可能高得惊人。技术报告，2017 年。[19] Bob Carpenter、Andrew Gelman、Matthew D Hoffman、Daniel Lee、Ben Goodrich、Michael Betancourt、Marcus Brubaker、Jiqiang Guo、Peter Li 和 Allen Riddell。Stan：一种概率编程语言。《统计软件杂志》，76(1)，2017 年。[20] Rich Caruana、Alexandru Niculescu-Mizil、Geoff Crew 和 Alex Ksikes。从模型库中进行集成选择。载于《第二十一届国际机器学习会议论文集》，第 18 页，2004 年。[21] Kathryn Chaloner 和 Isabella Verdinelli。贝叶斯实验设计：综述。《统计科学》，10(3)：273–304，1995 年。[22] Alexandra Chouldechova。具有差别影响的公平预测：再犯预测工具中的偏见研究。载于《大数据》，第 5 卷，第 153–163 页。Mary Ann Liebert, Inc.，纽约州新罗谢尔市 Huguenot 街 140 号 3 楼，邮编 10801，美国，2017 年。[23] Robert T Clemen 和 Robert L Winkler。在风险分析中结合专家提供的概率分布。《风险分析》，19(2)：187–203，1999 年。[24] David A Cohn、Zoubin Ghahramani 和 Michael I Jordan。利用统计模型进行主动学习。《人工智能研究杂志》，4：129–145，1996 年。[25] Roger M Cooke。不确定性中的专家：科学中的意见与主观概率。《环境伦理学》，13(2)，1991 年。[26] Sam Corbett-Davies、Emma Pierson、Avi Feller、Sharad Goel 和 Aziz Huq。算法决策与公平性的代价。第 23 届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第 797–806 页，2017 年。[27] Bo Cowgill 和 Catherine E Tucker。程序员有偏见？还是数据有偏见？一项将人工智能伦理付诸实践的现场实验。可见于 SSRN 3615404，2020 年。[28] Kyle Cranmer、Johann Brehmer 和 Gilles Louppe。基于模拟的推断前沿。《美国国家科学院院刊》，117(48)：30055–30062，2020 年。[29] Marco F Cusumano-Towner、Feras A Saad、Alexander K Lew 和 Vikash K Mansinghka。Gen：一种支持可编程推断的通用概率编程系统。载于第 40 届 ACM SIGPLAN编程语言设计与实现会议论文集，第 221–236 页，2019 年。[30] Andrea Dal Pozzolo、Olivier Caelen、Reid A Johnson 和 Gianluca Bontempi。从业者视角下信用卡欺诈检测的经验教训。《专家系统与应用》，41(10)：4915–4928，2014 年。[31] Norman Dalkey 和 Olaf Helmer。德尔菲法在专家使用方面的一项实验性应用。《管理科学》，9(3)：458–467，1963 年。[32] Thomas G Dietterich。机器学习中的集成方法。载于《多分类器系统国际研讨会》，第 1–15 页。施普林格，2000 年。[33] David L Donoho 和 Miriam Gasko。基于半空间深度和投影离群度的位置估计的崩溃性质。《统计年鉴》，第 1803–1827 页，1982 年。

[34] Michael F Drummond、Mark J Sculpher、Karl Claxton、Greg L Stoddart 和 George W Torrance。《卫生经济学与卫生技术评估》。牛津大学出版社，2015 年。[35] Philippa J Easterbrook、Ramana Gopalan、Jesse A Berlin 和 David R Matthews。医学研究中的发表偏倚。《柳叶刀》，337(8746):867–872，1991 年。[36] Charles Elkan。代价敏感学习的基础。载于国际人工智能联合会议，第 17 卷，第 973–978 页，2001 年。[37] 欧洲议会和理事会。《人工智能法案》。《欧盟官方公报》，2024 年。[38] Gerard FitzGerald、George A Jelinek、Deborah Scott 和 Marie F Gerdtz。急诊科分诊量表及其组成部分：科学证据的系统综述。《学术急诊医学》，17(1):1–13，2010 年。[39] Peter I Frazier。贝叶斯优化教程。arXiv 预印本 arXiv:1807.02811，2018 年。[40] Yoav Freund 和 Robert E Schapire。在线学习的决策理论推广及其在提升中的应用。《计算机与系统科学杂志》，55(1):119–139，1997 年。[41] Sorelle A Friedler、Carlos Scheidegger、Suresh Venkatasubramanian、Sonam Choudhary、Evan P Hamilton 和 Derek Roth。机器学习中公平性增强干预措施的比较研究。公平性、问责与透明度会议论文集，第 329–338 页，2019 年。[42] S Michael Gaddis。Lakisha 和 Jamal 有多黑人？通信审计研究所用姓名反映的种族观念。《社会学科学》，4:469–489，2017 年。[43] Yarin Gal 和 Zoubin Ghahramani。将随机失活作为贝叶斯近似：表示深度学习中的模型不确定性。载于国际学习表征会议，第 1050–1059 页，2016 年。[44] Michael Gira、Ruisu Zhang 和 Kangwook Lee。通过高效微调消除预训练语言模型中的偏见。第二届平等、多样性与包容语言技术研讨会论文集，第 59–69 页，2022 年。[45] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。现代神经网络的校准问题。载于第 34 届国际机器学习会议论文集，第 1321–1330 页。PMLR，2017 年。[46] Frank R Hampel、Elvezio M Ronchetti、Peter J Rousseeuw 和 Werner A Stahel。《稳健统计：基于影响函数的方法》。John Wiley & Sons，2011 年。[47] Moritz Hardt、Eric Price 和 Nati Srebro。监督学习中的机会平等。载于神经信息处理系统进展，第 3315–3323 页，2016 年。[48] Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。大规模多任务语言理解的测量。arXiv 预印本 arXiv:2009.03300，2020 年。[49] Joseph L Hodges 和 Erich L Lehmann。基于秩检验的位置估计。《数学统计年刊》，第 598–611 页，1963 年。[50] Ronald A Howard。信息价值理论。《IEEE系统科学与控制论汇刊》，2(1):22–26，1966 年。[51] Peter J Huber 和 Elvezio M Ronchetti。《稳健统计》。John Wiley & Sons，2009 年。[52] Dongfu Jiang、Xiang Ren 和 Bill Yuchen Lin。LLM-blender：通过成对排序和生成式融合集成大语言模型。arXiv 预印本 arXiv:2306.02561，2023 年。[53] Saurav Kadavath、Tom Conerly、Amanda Askell、Tom Henighan、Dawn Drain、Ethan Perez、Nicholas Schiefer、Zac Hatfield-Dodds、Nova DasSarma、Eli Tran-Johnson 等。语言模型（大多）知道自己知道什么。arXiv 预印本 arXiv:2207.05221，2022 年。[54] Leslie Pack Kaelbling、Michael L Littman 和 Anthony R Cassandra。部分可观测随机领域中的规划与行动。《人工智能》，101(1-2):99–134，1998 年。[55] Jerome P Kassirer 和 Richard I Kopelman。诊断过程。《内科学年鉴》，110(11):893–900，1989 年。[56] Diederik P Kingma 和 Max Welling。自动编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114，2013 年。[57] Jon Kleinberg、Sendhil Mullainathan 和 Manish Raghavan。公平确定风险评分时固有的权衡。载于第 8 届理论计算机科学创新会议（ITCS 2017），2017 年。[58] Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo 和 Yusuke Iwasawa。大语言模型是零样本推理器。神经信息处理系统进展，35:22199–22213，2022 年。

[59] Deborah Korenstein、Minal S Kale 和 Wendy Levinson。美国急诊护理的成本。《JAMA内科学》，174(9):1516–1517，2014 年。[60] Hadas Kotek、Rikker Dockum 和 David Sun。大语言模型中的性别偏见与刻板印象。ACM集体智能会议论文集，第 12–24 页，2023 年。[61] Nathan R Kuncel、David M Klieger、Brian S Connelly 和 Deniz S Ones。甄选与录取决策中的机械式与临床式数据组合：荟萃分析。《应用心理学杂志》，98(6):1060–1072，2013 年。[62] Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。使用深度集成进行简单且可扩展的预测不确定性估计。载于神经信息处理系统进展，第 6402–6413 页，2017 年。[63] Anja Lambrecht 和 Catherine Tucker。算法偏见？对 STEM 职业广告展示中表面性别歧视的实证研究。《管理科学》，65(7):2966–2981，2019 年。[64] Tor Lattimore 和 Csaba Szepesv´ari。《老虎机算法》。剑桥大学出版社，2020 年。[65] Jiawei Li、Xinyuan Cheng、Wayne Xin Zhao、Jian-Yun Nie 和 Ji-Rong Wen。ChatGPT 在招聘流程中的应用：机遇与挑战。arXiv 预印本 arXiv:2303.17564，2023 年。[66] Yifei Li、Zeqi Lin、Shizhuo Zhang、Qiang Fu、Bei Chen、Jian-Guang Lou 和 Weizhu Chen。借助步骤感知验证器提升语言模型的推理能力。arXiv 预印本 arXiv:2206.02336，2022 年。[67] Paul Pu Liang、Chiyu Wu、Louis-Philippe Morency 和 Ruslan Salakhutdinov。探索语言模型中的社会偏见并加以缓解。国际机器学习会议，第 6565–6576 页，2021 年。[68] Percy Liang、Rishi Bommasani、Tony Lee、Dimitris Tsipras、Dilara Soylu、Michihiro Yasunaga、Yian Zhang、Deepak Narayanan、Yuhuai Wu、Ananya Kumar 等。语言模型的整体评估。arXiv 预印本 arXiv:2211.09110，2022 年。[69] Dennis V Lindley。关于衡量实验所提供信息的方法。《数学统计年刊》，第 986–1005 页，1956 年。[70] LinkedIn人才解决方案。《2023年全球招聘趋势》。技术报告，LinkedIn公司，2023 年。[71] Alisa Liu、Maarten Sap、Ximing Lu、Swabha Swayamdipta、Chandra Bhagavatula、Noah A Smith 和 Yejin Choi。DExperts：利用专家与反专家模型进行解码时受控文本生成。计算语言学协会第 59 届年会论文集，第 6691–6706 页，2021 年。[72] Martin A Makary 和 Michael Daniel。医疗错误：美国第三大死因。《英国医学杂志》，353:i2139，2016 年。[73] Harry Markowitz。投资组合选择。《金融学杂志》，7(1):77–91，1952 年。[74] Ninareh Mehrabi、Fred Morstatter、Nripsuta Saxena、Kristina Lerman 和 Aram Galstyan。机器学习中的偏见与公平性综述。《ACM计算评论》，54(6):1–35，2021 年。[75] Moin Nadeem、Anna Bethke 和 Siva Reddy。StereoSet：测量预训练语言模型中的刻板印象偏见。计算语言学协会第 59 届年会论文集，第 5356–5371 页，2021 年。[76] Roberto Navigli、Simone Conia 和 Bj¨orn Ross。大语言模型中的偏见：起源、清单与讨论。《ACM数据与信息质量杂志》，15(2):1–17，2023 年。[77] Radford M Neal。神经网络的贝叶斯学习。《统计学讲义》，118，1996 年。[78] 纽约市议会。《2021年第144号地方法：自动化就业决策工具》。《纽约市行政法典》，2021 年 12 月。[79] Andrew Y Ng 和 Michael I Jordan。判别式分类器与生成式分类器：逻辑回归与朴素贝叶斯的比较。神经信息处理系统进展，14，2002 年。[80] Jeremy Nixon、Michael W Dusenberry、Linchuan Zhang、Ghassen Jerfel 和 Dustin Tran。深度学习中的校准测量。CVPR研讨会，2(7)，2019 年。[81] Ziad Obermeyer、Brian Powers、Christine Vogeli 和 Sendhil Mullainathan。剖析一种用于管理人群健康的算法中的种族偏见。《科学》，366(6464):447–453，2019 年。[82] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。训练语言模型，使其遵循人类反馈给出的指令。神经信息处理系统进展，35:27730–27744，2022 年。

[83] George Papamakarios、David Sterratt 和 Iain Murray。《序贯神经似然：利用自回归流实现快速无似然推断》。国际人工智能与统计学会议论文集，第 837–848 页，2019 年。[84] John Platt 等。《支持向量机的概率输出及其与正则化似然方法的比较》。载于《大间隔分类器进展》，第 10 卷，第 61–74 页。麻省理工学院出版社，1999 年。[85] John H Pope、Tom P Aufderheide、Robin Ruthazer、Robert H Woolard、Jerome A Feldman、Joni R Beshansky、John L Griffith 和 Harry P Selker。《急诊科漏诊的急性心肌缺血病例》。New England Journal of Medicine，342(16):1163–1170，2000 年。[86] Lutz Prechelt。《十倍程序员的迷思》。《IEEE软件》，36(5):96–99，2019 年。[87] Yujia Qin、Shihao Liang、Yining Ye、Kunlun Zhu、Lan Yan、Yaxi Lu、Yankai Lin、Xin Cong、Xiangru Tang、Bill Qian 等。《ToolLLM：助力大语言模型掌握 16000+ 现实世界中的应用程序接口》。arXiv预印本 arXiv:2307.16789，2023 年。[88] Manish Raghavan、Solon Barocas、Jon Kleinberg 和 Karen Levy。《缓解算法招聘中的偏见：评估相关主张与实践》。载于《2020年公平性、问责与透明度会议论文集》，第 469–481 页，2020 年。[89] Howard Raiffa 和 Robert Schlaifer。《应用统计决策理论》。哈佛大学出版社，1961 年。[90] Danilo Rezende 和 Shakir Mohamed。《利用归一化流进行变分推断》。国际机器学习会议，第 1530–1538 页，2015 年。[91] Christian P Robert。《贝叶斯选择：从决策理论基础到计算实现》。施普林格科学与商业媒体，2007 年。[92] Gene Rowe 和 George Wright。《德尔菲法作为预测工具：问题与分析》。《国际预测期刊》，15(4):353–375，1999 年。[93] Elizabeth G Ryan、Christopher C Drovandi 和 Anthony N Pettitt。《面向需要大量计算预算的非线性模型的贝叶斯实验设计》。《计算统计与数据分析》，97:67–86，2016 年。[94] Leonard J Savage。《统计学基础》。Courier Corporation，1972 年。[95] Timo Schick、Jane Dwivedi-Yu、Roberto Dess`ı、Roberta Raileanu、Maria Lomeli、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。《Toolformer：语言模型可以自学使用工具》。arXiv预印本 arXiv:2302.04761，2023 年。[96] Silvana Secinaro、Davide Calandra、Alessandro Secinaro、Vimal Muthurangu 和 Paolo Biancone。《医疗保健领域的人工智能：过去、现在与未来》。《可持续发展》，13(10):5762，2021 年。[97] Burr Settles。《主动学习文献综述》。技术报告，威斯康星大学麦迪逊分校计算机科学系，2009 年。[98] Bobak Shahriari、Kevin Swersky、Ziyu Wang、Ryan P Adams 和 Nando De Freitas。《让人类退出循环：贝叶斯优化综述》。《IEEE会刊》，104(1):148–175，2016 年。[99] Victor S Sheng 和 Charles X Ling。《通过阈值化实现分类器的代价敏感化》。载于《全国人工智能大会论文集》，第 21 卷，第 476 页，2006 年。[100] Yishai Shimoni、Chen Yanover、Ehud Karavani 和 Yaara Goldschmnidt。《因果推断分析性能评估的基准测试框架》。载于 NeurIPS因果学习研讨会，2018 年。[101] Chenglei Si、Zhe Gan、Zhengyuan Yang、Shuohang Wang、Jianfeng Wang、Jordan Boyd-Graber 和 Lijuan Wang。《提示 GPT-3 使其更加可靠》。载于国际学习表征会议，2023 年。[102] Karan Singhal、Shekoofeh Azizi、Tao Tu、S Sara Mahdavi、Jason Wei、Hyung Won Chung、Nathan Scales、Ajay Tanwani、Heather Cole-Lewis、Stephen Pfohl 等。《大语言模型编码了临床知识》。《自然》，620(7972):172–180，2023 年。[103] Jennifer Smith 和 Michael Jones。《计算招聘成本：全面指南》。《人力资源杂志》，65(3):48–53，2020 年。[104] 人力资源管理协会。《招聘的真实成本》。技术报告，SHRM，2016 年。[105] Matthijs TJ Spaan 和 Nikos Vlassis。《Perseus：针对部分可观测马尔可夫决策过程的随机化基于点的价值迭代》。《人工智能研究杂志》，24:195–220，2005 年。[106] Aarohi Srivastava、Abhinav Rastogi、Abhishek Rao、Abu Awal Md Shoeb、Abubakar Abid、Adam Fisch、Adam R Brown、Adam Santoro、Aditya Gupta 等。《超越模仿游戏：量化并外推语言模型的能力》。arXiv预印本 arXiv:2206.04615，2022 年。

[107] Stephen M Stigler。《Simon Newcomb、Percy Daniell 与稳健估计的历史（1885–1920）》。《美国统计学会会刊》，68(344):872–879，1973 年。[108] Swee Song Tan、Jan Bakker、Marga E Hoogendoorn、Atul Kapila、John Martin、Alessandra Pezzi、Giorgio Pittoni、Peter E Spronk、Robert Welte 和 Leona Hakkaart-van Roijen。《重症监护的成本》。《重症医学》，37(10):1640–1646，2011 年。[109] Triplebyte。《评估技术人才：传统筛选为何失效》。技术报告，Triplebyte Inc.，2019 年。[110] Konstantinos Tzioumis。《名字的人口统计学特征》。《科学数据》，5(1):1–9，2018 年。[111] 美国国会。《平等信贷机会法》，1974 年。美国法典第 15 编第 1691 条及后续条款。[112] Various Industry Sources。《行业招聘漏斗转化基准》，2023 年。汇总自 LinkedIn、Glassdoor 和 SHRM 的报告。[113] Abraham Wald。《统计决策函数》。John Wiley & Sons，1950 年。[114] Alex Wang、Yada Pruksachatkun、Nikita Nangia、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel Bowman。《SuperGLUE：适用于通用语言理解系统的更具挑战性的基准测试》。载于《神经信息处理系统进展》，第 32 卷，2019 年。[115] Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R Bowman。《GLUE：用于自然语言理解的多任务基准测试与分析平台》。载于《2018年EMNLP BlackboxNLP研讨会：分析与解读用于自然语言处理的神经网络论文集》，第 353–355 页，2018 年。[116] Lei Wang、Chen Ma、Xueyang Feng、Zeyu Zhang、Hao Yang、Jingsen Zhang、Zhiyuan Chen、Jiakai Tang、Xu Chen、Yankai Lin 等。《基于大语言模型的自主智能体综述》。arXiv预印本 arXiv:2308.11432，2023 年。[117] Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。《自洽性提升大语言模型中的思维链推理》。arXiv预印本 arXiv:2203.11171，2022 年。[118] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。《思维链提示能够引出大语言模型的推理能力》。第 35 卷，第 24824–24837 页，2022 年。[119] Laura Weidinger、John Mellor、Maribeth Rauh、Conor Griffin、Jonathan Uesato、Po-Sen Huang、Myra Cheng、Mia Glaese、Borja Balle、Atoosa Kasirzadeh 等。《语言模型造成伤害的伦理与社会风险》。arXiv预印本 arXiv:2112.04359，2021 年。[120] Ben Wilson、Judy Hoffman 和 Jamie Morgenstern。《构建并审查公平算法：候选人筛选案例研究》。载于《2021年ACM公平性、问责制与透明度会议论文集》，第 666–677 页，2021 年。[121] David H Wolpert。《堆叠泛化》。第 5 卷，第 241–259 页。爱思唯尔，1992 年。[122] Zhiheng Xi、Wenxiang Chen、Xin Guo、Wei He、Yiwen Ding、Boyang Hong、Ming Zhang、Junzhe Wang、Senjie Jin、Enyu Zhou 等。《基于大语言模型的智能体的兴起与潜力：综述》。arXiv预印本 arXiv:2309.07864，2023 年。[123] Miao Xiong、Zhiyuan Hu、Xinyang Lu、Yifei Li、Jie Fu、Junxian He 和 Bryan Hooi。《大语言模型能否表达不确定性？对大语言模型置信度 elicitation 的实证评估》。国际学习表征会议论文集，2024 年。[124] Lin Xu、Frank Hutter、Holger H Hoos 和 Kevin Leyton-Brown。《SATzilla：面向可满足性问题的基于组合的算法选择》。《人工智能研究杂志》，32:565–606，2008 年。[125] Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Thomas L Griffiths、Yuan Cao 和 Karthik Narasimhan。《思维树：利用大语言模型进行审慎的问题求解》。《神经信息处理系统进展》，36，2023 年。[126] Bianca Zadrozny 和 Charles Elkan。《将分类器得分转换为准确的多类别概率估计》。载于《第八届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第 694–699 页，2002 年。[127] Bianca Zadrozny、John Langford 和 Naoki Abe。《通过按代价比例加权样本实现代价敏感学习》。载于第三届 IEEE 国际数据挖掘会议，第 435–442 页。IEEE，2003 年。[128] Brian Hu Zhang、Blake Lemoine 和 Margaret Mitchell。《通过对抗学习缓解不期望的偏见》。载于《2018年AAAI/ACM人工智能、伦理与社会会议论文集》，第 335–340 页，2018 年。[129] Wenxuan Zhang、Yue Deng、Bing Liu、Sinno Jialin Pan 和 Lidong Bing。《大语言模型时代的情感分析：现实检验》。arXiv预印本 arXiv:2305.15005，2023 年。

[130] Jieyu Zhao、Tianlu Wang、Mark Yatskar、Vicente Ordonez 和 Kai-Wei Chang。《共指消解中的性别偏见》。载于 *2018*年北美计算语言学协会分会年会论文集：人类语言技术，第 8–14 页，2018 年。

算法2 面向成本感知序贯决策的贝叶斯多大语言模型编排

输入： 初始证据 $x_1 \in \mathcal{X}$，状态空间 $\mathcal{S}$，动作空间 $\mathcal{A}$，成本矩阵 C：$\mathcal{A}\times\mathcal{S} \to \mathbb{R}_{\geq 0}$，先验 $\pi \in \Delta(\mathcal{S})$，大语言模型集成 $\{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$，分歧阈值 $\tau_D \in \mathbb{R}_+$，信息性参数 $\rho \in [0,1]$，信息成本 $c_{\text{info}} \in \mathbb{R}_{\geq 0}$
输出： 终止行动 $a^* \in \mathcal{A}$
// 初始化
$t \leftarrow 1$
$x \leftarrow x_1$
$b(s) \leftarrow \pi(s)$ 对所有 $s \in \mathcal{S}$
当 真实分布 **时执行**
  // ==========步骤 *1*：似然 *elicitation* ==========
  对于 $m = 1$至 M **执行**
    对每个 $s \in \mathcal{S}$ **执行**
      生成对比式提示词 $\mathcal{P}_m(x,\ s)$（定义 2）
      $r_{m,s} \leftarrow \mathcal{M}_m(\mathcal{P}_m(x, s))$
      $\hat{L}_m(x \mid s) \leftarrow r_{m,\ s}/10$
    结束循环
  结束循环
  // ========== 步骤 *2*：鲁棒聚合 ==========
  对于每个 $s \in \mathcal{S}$ **执行**
    $L(x \mid s) \leftarrow$ 中位数$\{\hat{L}_1(x \mid s) \ldots,\ \hat{L}_M(x \mid s)\}$ （公式 63）
  结束循环
  // ========== 步骤 *3*：序贯信念更新 ==========
  $Z \leftarrow \sum_{s' \in \mathcal{S}} L\ (x \mid s')\ \cdot b\ (s')$
  对每个 $s \in \mathcal{S}$ **执行**
    $b_{\text{new}}\ (s)\ \leftarrow \dfrac{L\ (x \mid s)\ \cdot b\ (s)}{Z}$ （公式 66）
  结束循环
  $b \leftarrow b_{\text{new}}$
  // ========== 步骤 *4*：分歧检测 ==========
  对 每个 $s \in \mathcal{S}$ **执行**
    $\mu_s \leftarrow \dfrac{1}{M}\sum_{m=1}^{M} \hat{L}_m(x \mid s)$
    $\sigma_s \leftarrow \sqrt{\dfrac{1}{\cancel{M-1}}\sum_{m=1}^{M}\left(\hat{L}_m(x \mid s)\ - \mu_s\right)^2}$
    $D(s)\ \leftarrow \dfrac{\sigma_s}{\max(\mu_s,\ \ 10^{-8})}$
  结束循环
  $D_{\max} \leftarrow$ 最大值$_{s\in\mathcal{S}}\ D(s)$
  // ========== 步骤 *5*：信息收集决策 ==========
  如果 $D_{\max} > \tau_D$ **且**信息收集可用 **则**
    $C_{\text{current}} \leftarrow \min_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} {}^{b\ (s)\ \cdot C\ (a,\ \ s)}$
    $C_{\text{perfect}} \leftarrow \sum_{s\in\mathcal{S}} b(s)\ \cdot$ 分钟$_{a\in\mathcal{A}}\ C(a, s)$
    信息价值 $\leftarrow \rho \cdot (C_{\text{current}} - C_{\text{perfect}})$ （公式 74）
    如果信息价值 $> c_{\text{info}}$ **则**
      收集额外证据 $x_{\text{new}} \in \mathcal{X}$（例如，电话筛选）
      $t \leftarrow t + 1$
      $x \leftarrow x_{\text{new}}$
      进入下一次迭代
    结束 if
  结束 if
  // ==========步骤 *6*：终端动作选择 ==========
  $a^* \leftarrow$ 取最小值$_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} b(s)\ \cdot C(a, s)$ （式 69）
  返回 a^*
结束循环