# 超越简历：面向信息 elicitation 的评分标准感知型自动面试系统

**HarryStuart*** **Masahiro Kaneko、Timothy Baldwin**穆罕默德·本·扎耶德人工智能大学（MBZUAI）阿布扎比，阿拉伯联合酋长国
{harry.stuart,masahiro.kaneko,timothy.baldwin}@mbzuai.ac.ae

arXiv:2603.01775v1 [cs.CL] 2 Mar 2026

## 摘要

有效的招聘对组织的成功至关重要，但要找到最合适的候选人却极具挑战，因为专家评估（例如由技术经理进行的面试）大规模部署成本高昂。因此，自动简历评分和其他申请人筛选方法越来越多地被用于粗略筛选候选人，从而依据有限信息做出决策。我们提出，大语言模型（大语言模型）可以扮演主题专家的角色，以具有成本效益的方式从每位候选人处引出细致且与岗位相关的信息，从而提高招聘早期决策的质量。我们提出一个系统，该系统利用大语言模型面试官，以校准的方式更新对申请人评分标准导向的潜在特质的信念。我们在模拟面试上评估该系统，并表明信念会收敛至模拟申请者人为构建的潜在能力水平。我们发布了代码、一个规模适中的公有领域/匿名化简历数据集、信念校准测试以及模拟面试，见 https://github.com/mbzuai-nlp/beyond-the-resume。我们的演示地址为https://btr.hstu.net。

## 1 引言

招聘是一个资源密集型流程，因此也是许多组织面临的一大痛点（Bhargava 和 Assadi，2024年；Mühlemann 和 Strupler Leiser，2018年；Blatter 等人，2012年）。理想情况下，应邀请相关部门的高级员工等主题专家（SME）与每位候选人进行初次面试，以了解简历中通常未能体现的具体、细致且岗位特定的信息（Risavy 等人，2022年；Vincent 等人，2024年）。在实际操作中，这种做法成本过高。

因此，组织越来越多地采用筛选程序，将此类招聘评估留到仅针对合适的候选人进行。筛选通常通过自动简历评分（Li 等人，2020年；Chifu 等人，2017年）或由缺乏专业知识的工作人员开展面试来进行，他们无法实施高成本、高信息量评估（Lashkari 和 Cheng，2023年）。因此，招聘早期做出的决定缺乏最重要的信息。这会导致真正合适的候选人被筛除，而不合适的候选人却进入后续环节（Fisher 等人，2021年）。与事先获得通过专业方式 elicited 的补充信息相比，后续面试的成效也较低。

我们认为，大语言模型（LLM；Zhao 等人，2025年）可以在招聘早期阶段充当主题专家（SME）的低成本替代方案。具体而言，我们介绍了一种开源的自动化多轮面试系统，该系统结合了对评分标准的信念；该评分标准代表了人类主题专家希望了解的岗位特定信息。图1 展示了系统在面试中途的界面片段，右侧面板显示了对评分标准的信念。

我们探讨如何校准信念更新，并提出将信念收敛视为一种有用的操作化方式，用以衡量评分标准是否已被关于申请人的有意义信息充分填充。我们在模拟面试中评估了系统，并表明：（1）平均而言，随着面试进行，每轮之间的信念变化大约减少3×；（2）在模拟中，每位申请人都由一个固定的潜在评估标准档案（即“原型”）生成，最终信念有76.1%的概率恢复出正确的原型。

使用我们的系统进行面试，会生成两项有助于做出更明智决策的成果：（1）信息丰富的访谈记录；（2）完全可审计的信念更新日志。目前，已有研究开始探索使用大语言模型进行招聘面试（Hashimoto et al.，2025年；Jokoet al.，2022年；Adeseyeet al.，2025年），但现有系统

*通讯作者。

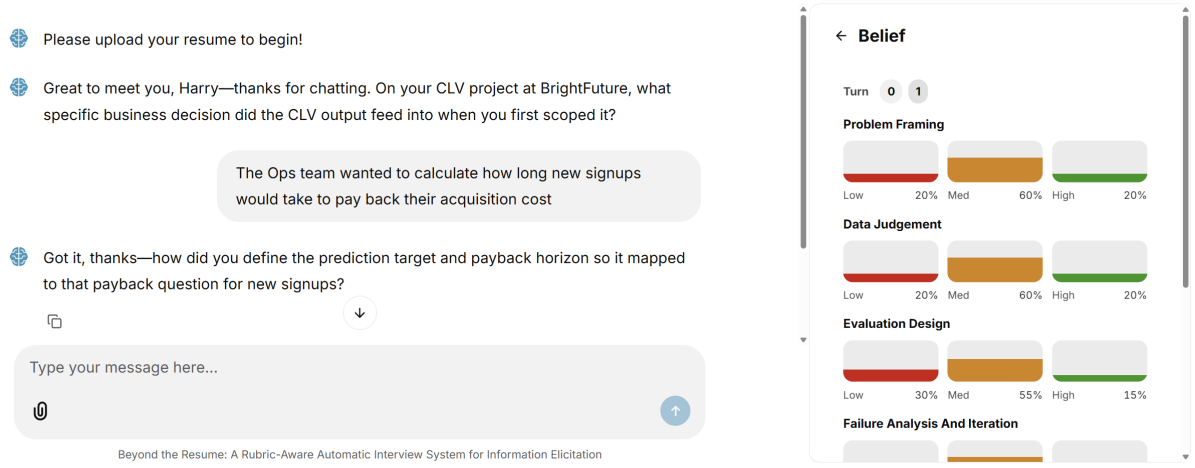



图1：用户与系统交互的屏幕截图。

大多将信息 elicitation 视为现场收集或对话引导，而非针对与评分标准对齐的潜在属性，维护明确且经过校准的概率信念状态。

## 2 系统概述

我们的系统演示以 Web 应用的形式部署（图1），并以机器学习工程领域的招聘情境为背景。我们的代码仓库包含自定义配置、自行托管以及其他评分标准示例的说明。系统运行一个 Python 应用，并使用 Chainlit 库（Chainlit 维护者，2026年）通过熟悉的大语言模型聊天界面编排面试。右侧的信念面板展示系统在每轮对各项评分标准维度的信念，并提供回合选择器以查看先前信念。此外，还可以审查信念变化，如图2所示。此信念面板仅在本演示中向用户展示。在实际部署场景中，申请人不会看到信念；面试记录和信念日志则会提供给招聘团队，可能通过仪表板或与现有招聘软件集成的方式提供。

从底层来看，我们的系统由两个组件构成：（1）一个评判者，以大语言模型实现，用于跟踪对评分标准的信念；（2）一个面试官，同样以大语言模型实现，其目标是从申请人那里 eliciting 有意义的信息，以充分涵盖评分标准。有意义的信息能够提供有关申请人潜在的**知识、技能与能力（KSA）**的证据，这些知识、技能与能力与评分标准相关（Wooten，1993；Shippmann 等，2000；Hlavac，2023年）。这两个组件分别捕捉：

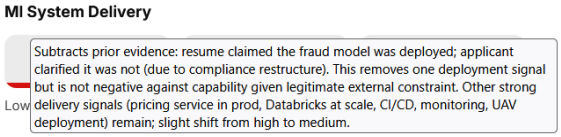



图 2：评判者降低机器学习系统交付维度上的信念时所作的说明，具体做法是贬低（削弱）先前证据的支持力度。并非源自图1所示的同一场面试。

申请人知道什么、能够做什么，以及他们如何表现。充分涵盖评分标准，意味着从申请人那里 eliciting 信息，使评判者对其信念充满信心。

本研究的所有大语言模型组件均使用 GPT-5（OpenAI，2025年），具体为`gpt-5-2025-08-07`，并将推理=`"low"` 以减少 token 用量。

### 2.1 定义评分标准

我们将评分标准定义为包含 $D$ 个维度（例如自主性或编程能力），每个维度有 $L$个序数等级。令 $\mathcal{L} = \{1, \ldots, L\}$ 表示按从低到高顺序排列的等级索引集合。我们将评分标准表示为：

$$R = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,L} \\ \vdots & \vdots & \ddots & \vdots \\ r_{D,1} & r_{D,2} & \cdots & r_{D,L} \end{bmatrix}. \qquad (1)$$

每个 $r_{d,\ell}$ 都由一段自然语言锚点定义，用于描述维度 $d$ 的等级 $\ell$所涵盖的知识、技能与能力（例如 $r_{\text{autonomy},\ 1}$ 可能表示遇到问题时依赖他人；未能明确最初需求……）。我们定义以下锚点要求：

- **非证据性的：** 锚点独立于申请人提供的证据而存在。

- **可证伪的：** 锚点描述了可推断、可确定的知识、技能与能力，避免使用高度主观的限定词。

- **内部一致性：** 在同一锚点内，所描述的知识、技能与能力应当在同一等级上共同成立，而不是混合可能不会同时出现的彼此正交的知识、技能与能力。

### 2.2 评判者

我们将评判者定义为一种大语言模型策略，用于初始化并更新对申请人知识、技能与能力的信念。它使用申请人提供的简历初始化信念，并在每个面试轮次结束后更新信念。此处的信念定义为：在给定迄今观察到的所有证据（简历 +面试记录）的条件下，各评分标准维度潜在等级的后验分布。我们还要求评判者说明其信念更新的理由，以便人工审核。

令 $z_d \in \mathcal{L}$ 表示申请人在评分标准维度 $d \in \{1, \ldots, D\}$上的潜在等级，令 $x_{0:t}$表示截至面试轮次$t$ 可用的证据（其中 $x_0$ 为简历， $x_{1:t}$ 为面试轮次）。评判者在轮次 $t$的信念状态为：

$$B_t = \{p(z_d \mid x_{0:t}, R)\}_{d=1}^{D}. \tag{2}$$

### 2.3 面试官

我们将面试官定义为一种大语言模型策略，其目标是如前所述，从申请人处 eliciting 有意义的信息，以使评分标准趋于饱和。我们将收敛操作化为连续逐维后验分布之间全变差距离的平均降幅：

$$\Delta_t = \frac{1}{D} \sum_{d=1}^{D} \mathrm{TV}([B_t]_d, [B_{t-1}]_d), \tag{3}$$

其中，$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{\ell \in \mathcal{L}} |p(\ell) - q(\ell)|$ 表示针对 $\mathcal{L}$ 上分布的全变差。理想的极限行为是，随着信息性证据不断累积， $B_t$会集中于每个维度 $d$对应的真实潜在等级 $z_d$ 。随着更多信息用于检验评判者后验分布所蕴含的假设，评判者会变得更加确信；收敛可以理解为这一过程。

## 3 校准评判者

### 3.1 数据集

在缺乏描述如何更新信念的先验数据时，我们定义了一组测试，以便对评判者的特性和行为作出断言。我们首先为以下三个相互独立的简历领域分别构建通用评分标准：平面设计、销售和机器学习工程。为简化实验，所有评分标准均采用 $L = 3$ 序数等级，并将$\mathcal{L} = \{1, 2, 3\} \equiv$ {低、中等、高}作为标识。对于每个简历领域，我们从 Reddit 的 r/Resumes 子版块下载 10 份公共领域的匿名化简历（共 30 份），[1] 并提取其中的文本内容。针对每份简历，我们手动编写一段简短的面试对话， $\mathcal{I}$包含 2–5 轮，作为后续测试中的存根。

### 3.2 蜕变测试

蜕变测试借鉴自软件工程，它规定在对输入进行受控扰动时，大语言模型的输出应如何变化，而无需预言机标签（Cho 等人，2025年；Hyun 等人，2024年）。

根据实证观察，GPT-5 在获得明确的新证据后，通常会使某个等级的概率质量至少变化 0.05。因此，我们将至少为$\varepsilon := 0.04$ 的变化视为信念变化。对于领域 $d$和等级 $\ell$，定义

$$\begin{aligned} p_d^-(\ell) &:= p(z_d = \ell \mid x_{0:t-1}, R), \\ p_d^+(\ell) &:= p(z_d = \ell \mid x_{0:t}, R), \\ \delta_d(\ell) &:= p_d^+(\ell) - p_d^-(\ell). \end{aligned}$$

$$\mathcal{R}_{\mathrm{inv}} : \max_{d,\ell} |\delta_d(\ell)| \le \varepsilon, \tag{4a}$$

$$\mathcal{R}_{\uparrow}(d, \ell) : \delta_d(\ell) > \varepsilon, \tag{4b}$$

$$\mathcal{R}_{\downarrow}(d, \ell) : \delta_d(\ell) < -\varepsilon. \tag{4c}$$

我们针对这三种行为分别构建一个测试类别。只有在注入受控的面试官—求职者轮次后观察到相应行为时，该类别中的测试才会通过，$\tau$。对于每个测试类别，我们针对每份简历构建一个 $\tau$ 。我们通过定义两种输入方案，为每份简历定义两项测试：（1）仅简历；以及（2）简历 $\cup \mathcal{I}$。我们使用这两种方案来观察评判者在面试第一轮以及面试后期的行为。因此，每个测试类别共有60项测试（30份简历，各含两种输入方案），共九个测试类别，总计540项蜕变测试。我们呈现

[1] https://www.reddit.com/r/Resumes/

| Test | 独立 | PBA |
|---|---|---|
| **不变式** ($\mathcal{R}_{\text{inv}}$) | | |
| 夸大 | 2 | **100** |
| 不相关性 | 2 | **98** |
| 重复 | 0 | **87** |
| **目标级别提升** ($\mathcal{R}_{\uparrow}(d,\ell)$) | | |
| 低证据添加 | **92** | 90 |
| 中等证据添加 | 72 | **88** |
| 高证据添加 | 95 | **97** |
| **目标级别降低** ($\mathcal{R}_{\downarrow}(d,\ \ell)$) | | |
| 低证据贬低 | 73 | **85** |
| 中等证据贬低 | **73** | 70 |
| 高等证据贬损 | **95** | 92 |
| **其他** | | |
| 均匀先验 | **87** | 80 |

表1：按评判者划分的测试通过率（四舍五入后的百分比）。

每个测试类别，见表1 中其预期关系的下方。所谓削弱证据，是指降低其支持力度。示例与分层见第A节。

### 3.3 均匀先验。

在缺乏与某个评分标准维度相关的任何信息时，评判者应对所有等级持均匀信念。我们通过对每份简历及其对应的评分标准，添加一个来自其他评分标准的无关维度，并测量注入维度各个等级的后验分布是否接近均匀分布，以此进行检验。

设 $d_{\text{inj}}$ 表示注入的无关维度，$p_{d_{\text{inj}}} \in \Delta^2$ 表示其三个层级上的后验分布。此测试用于检查该后验分布在 $L_\infty$ 容差范围内是否近似均匀：

$$\|p_{d_{\text{inj}}} - (\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3})\|_\infty \le 0.04. \tag{5}$$

### 3.4 定义两位评判者

我们尝试校准两位大语言模型评审器。我们指示两者在单次调用中返回完整的后验分布集。

**独立。** 在第 $t$ 轮，评判者分析该轮可用的完整证据集，并直接构建后验分布：

$$B_t^{\text{ind}} = \{p(z_d \mid x_{0:t}, R)\}_{d=1}^{D}. \tag{6}$$

**先前信念感知 (PBA.)** 在第 $t$ 轮，评判者也会获得完整的证据集 $x_{0:t}$，但

| 政策（$\pi$） | 简历 | 评分标准 | 先前信念 $B_{t-1}$ |
|---|---|---|---|
| 信念感知 | ✓ | ✓ | ✓ |
| 信念无感知 | ✓ | ✓ | ✗ |
| 评分标准无感知 | ✓ | ✗ | ✗ |
| 浅层无感知 | ✓ | ✗ | ✗ |

表 2：第$t$轮的面试官策略及其输入。浅层无感知策略被指示提出浅显、开放式的问题。

还会获得先前信念 $B_{t-1}^{\text{PBA}}$，并被要求关注证据集从 $x_{0:t-1}$ 到 $x_{0:t}$的变化来进行更新：

$$B_t^{\text{PBA}} = \{p(z_d \mid x_{0:t}, R, B_{t-1}^{\text{PBA}}, x_{0:t-1})\}_{d=1}^{D}. \tag{7}$$

因此，该更新具有变化感知，但仍以第 $t$ 轮的整个证据集为条件。

对于初始化（仅简历），我们使用均匀先验：

$$B_{-1}^{\text{PBA}} = \{(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3})\}_{d=1}^{D} \tag{8}$$

### 3.5 结果

表1 展示了校准结果。可以看出，与独立评判器相比，PBA评判器具有很高的稳定性，这也在合理预期之中。这对信念收敛是否有意义至关重要。两种评判器在其他测试中的整体表现通常都同样出色，但即使在低和高水平上表现良好，也不太倾向于从中等水平中移除概率质量。这种惯性似乎是合理的，因为中等水平代表平均值；与极端值相比，降低对平均值的证据支持可能需要更多反面证据。

## 4 评估系统

我们现在使用模拟来证明，我们的系统能够用关于申请人的有意义信息使评分标准达到饱和。

### 4.1 模拟

我们首先声明，此模拟仅作为系统行为的粗略指标，而非在真实面试的复杂性与细微差别下对面试官策略的严格测试。模拟组件包括：

•**原型：** 形式上，原型 $a$ 是一个固定的潜在评估标准档案 $z_{(a)\ 1:D} \in \mathcal{L}^D$，其中 $z_{(a)\ d}$ 是原型所隐含的潜在

维度等级 $d$（例如，TechWizard 原型可能会将 $z^{(a)}_{\text{编程}} =$ 设为高，将 $z^{(a)}_{\text{组织}} =$ 设为低）。

- **评判者：**我们使用第3.4节中定义的 PBA评判器。
- **申请者：**我们使用大语言模型，根据档案（简历和原型）来模拟申请者的消息。我们向大语言模型提供相应的评分标准，以便为其原型提供背景信息。我们还为申请者设定一个 1–2 个词的个性。
- **面试官。**我们也使用大语言模型来模拟面试官，并实验四种策略，详见表2。

对于每项评分标准，我们构建三个定制原型，以及三个常见的锚定原型，其定义见公式（9），共计六个：

$$\begin{aligned} \text{low-logan} &= \{z_d = \text{low}\}_{d=1}^{D}, \\ \text{average-jessie} &= \{z_d = \text{medium}\}_{d=1}^{D}, \\ \text{high-harley} &= \{z_d = \text{high}\}_{d=1}^{D}. \end{aligned} \tag{9}$$

使用第3.1节中定义的数据集，我们针对每个评分标准领域，将 10 份简历与六种原型组合，得到 $P_{\text{full}} = 180$ 个档案，涵盖三项评分标准。对于每个档案，我们使用其简历初始化评判者，并模拟 $T = 12$个面试官-申请者对话轮次。考虑到实际情况，我们预计 12 条回复大致就是现实用户所能提供的上限。

### 4.2 避免申请人操纵

一个关键风险是申请人操纵：申请人泄露原型线索（Lotfi等人，2023年），使评判者能够轻易推断出其原型。我们尝试通过注入申请者提示的护栏来降低这一风险：

- **自我保护：** 不鼓励明确的自我贬低；要求以暗示而非直接声明的方式表达局限性（尤其是对于知识、技能与能力较低的情况）。
- **展示而非讲述：** 引出情境性证据（示例/观点），而不是逐字罗列特质。
- **不主动提供证据：** 只回答所问内容；不要添加未被要求的细节。

我们还使用大语言模型对申请人的拟定回答进行净化，且未让其看到问题，以减少申请人对其原型的知识、技能与能力的过度发挥，并避免受到面试官的影响。

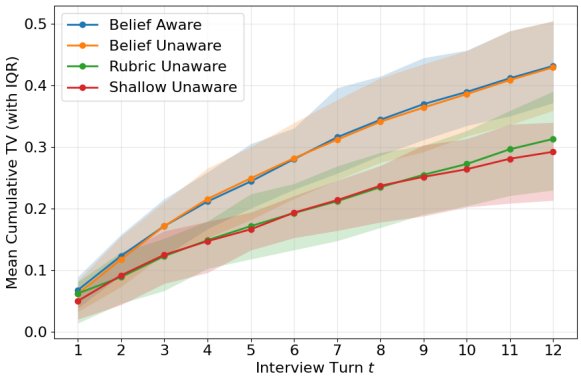



图3：使用一组 $P = 30$档案进行模拟面试时对方程（10）的比较。

### 4.3 选择面试官

在使用经过校准的 PBA评判器后，我们对 $P = 30$ （从 180 份档案中随机选取）进行小规模模拟，并选择最能引发信念变化的档案。我们使用按轮次索引的累积后验全变差（CTV）作为截至第 $t$轮所获得的评分细则特定信息增益的代理指标。逐轮后验偏移， $\Delta_t$，定义见公式（3），因此第 $t$ 轮的 CTV 定义如下：

$$\text{CTV}_t(\pi) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{u=1}^{t} \Delta_u^{(p,\pi)}, \quad t \in \{1, \ldots, T\}. \tag{10}$$

如图所示 3，所有面试官策略都会使评判者在面试过程中更新其信念，而加入评分标准则能清晰地区分质量。面试官质量的差异有力地说明了申请人操纵并不存在。

另一个重要观察是，信念感知与信念无感知策略之间的表现差异很小。我们认为，鉴于面试本身相对简单，信念无感知面试官很可能仅凭证据就能推断出合理的信念状态。鉴于我们没有能够有效纳入信念的面试官策略，后续完整的**模拟**及结果均采用4.4节和第 4.5节中的

### 4.4 展示收敛

我们现在重新审视第 2.3 节引入的收敛概念，即方程（3）中连续各维度后验分布之间的全变差的减小。图4 显示，评判者的

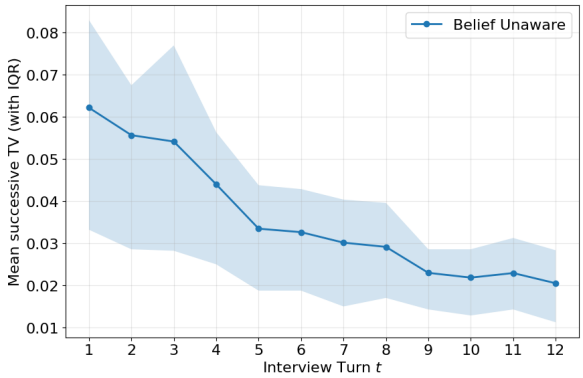



图 4：评判者连续信念状态之间的全变差$\Delta_t$ 如公式（3）所述，在$P_{\text{full}} = 180$模拟面试过程中。

信念更新（以全变差衡量）在 $t = 0$（仅简历）与 $t = 1$ （简历 +首轮面试）之间的平均值为 0.0621。到了最后一个面试轮次，评判者的信念更新幅度已降至平均约为 $3\times$ 0.0205，呈现下降趋势。这有力地表明，随着信念无感知面试官从申请人处 elicitation 信息，评判者对申请人的知识、技能与能力的信念变得越来越确定。

### 4.5 原型恢复

我们现在展示，随着评判者的信念收敛，这一过程具有明确意义，即趋向于申请人的潜在知识、技能与能力。为此，我们报告最终轮次 $T$ 的原型恢复结果。对于每个档案 $p$，我们根据最终信念 $B_T^{(p)}$ 构建最大后验评分标准档案 $\hat{z}_1^{(p)}:D$，方法如下：

$$\hat{z}_d^{(p)} = \arg\max_{\ell \in \mathcal{L}} [B_T^{(p)}]_d(\ell). \tag{11}$$

令 $M(\cdot)$ 表示将评分标准档案转换为 $D \times L$ 指示矩阵的标准独热编码。随后，我们通过在原型集合 $\mathcal{A}$中进行最近邻匹配来预测原型：

$$\hat{a}_T^{(p)} = \arg\min_{a \in \mathcal{A}} \left\| M(\hat{z}_{1:D}^{(p)}) - M(z_{1:D}^{(a)}) \right\|_F. \tag{12}$$

恢复率是满足$\hat{a}_T^{(p)} = a^{(p)}$的档案所占比例。在 $P_{\text{full}} = 180$ 上，我们发现本系统的原型恢复率为76.1%，而若在 $t = 0$（仅简历）下进行恢复，该比例为16.7%。我们观察到，最常见的等级误分类是误分到相邻等级，这表明处于极端（即偏低或偏高）的误分类仍然能够传达潜在等级是低于平均水平还是高于平均水平。参见第B节了解更多细节。

表4 （附录）简要对比了本节与第 4.4节的结果，以及谷歌的 Gemini 3.1 Pro（GoogleDeepMind，2026年）的表现。结果显示，它表现出色，可作为 GPT-5 的潜在替代方案。

## 5 相关工作

自动化招聘研究长期以来一直强调基于简历的筛选与排序，包括能力等级估计以及对简历的机器学习评分（Li等人，2020年；Chifu等人，2017年；Lo等人，2025年）。相比之下，已有新兴研究探讨基于大语言模型的面试系统，包括可动态生成槽位填充提示的职业面试前聊天机器人（Hashimoto等人，2025年）、以行为事件访谈为基础的自然语言处理/机器学习能力面试机器人（Joko等人，2022年），以及可生成与专业知识相匹配的问题的模块化面试官（Adeseye等人，2025年）。商业平台同样支持早期评估与事务安排。例如，HireVue和Modern Hire提供人工智能视频面试和评估。[23] 由于此类商业模型的内部机制不透明，本文首次提出一种开源系统，其中纳入了与评分标准对齐、可审计的信念收敛。

## 6 结论与未来工作

我们提出，本文介绍的系统能够以较低成本、大规模地替代高成本的中小企业访谈，从而通过为决策者提供更充分的信息，改进招聘早期阶段的流程。我们通过模拟证明，经过校准的评判者信念会随时间趋于稳定，并朝着模拟申请者的潜在知识、技能与能力方向更新，表明我们的系统能够从申请人那里引出有意义的信息。

在未来的工作中，我们希望研究如何利用真实数据，将信念有意义地纳入面试官策略。例如，通过使用贝叶斯实验设计，在每一轮最大化信息增益（Choudhury 等人，2026年）。此外，我们希望扩展该系统，使其不仅能纳入简历，还能纳入候选人在求职申请中提供的 LinkedIn 个人资料、数字作品集以及其他来源。

[2]https://www.hirevue.com/[3] https://www.modernhire.com/

## 局限性与伦理

**模拟** 我们预计，我们模拟的申请者与人类申请者有所不同。我们预计，人类申请者群体在诚实、正直、沟通能力、英语熟练度、对话风格、个性等方面会有显著更大的差异。此外，我们的模拟很可能无法捕捉面试人类申请人所固有的细微差别与复杂性（Zhou 等人，2019），例如申请人提出澄清性问题、声称自己比实际更为熟练，或总体上具有更为复杂且模糊的潜在知识、技能与能力。

**Reddit 数据集** 尽管在 www.reddit.com 上传数据的个人已将数据公开，但我们仍希望对研究中使用的所有数据进行匿名化处理。截至 2026-02- 27，r/Resumes 子版块的规则要求用户在发布简历前去除身份信息。[4] 无论如何，我们都努力对细节进行模糊处理，并移除所有残留的个人身份信息（PII）。在发布的数据集中，我们也仅提供经过提取且细节已匿名化的文本。由于不提供原始 PDF/图像文件，要还原哪份简历对应哪位 Reddit 用户将变得困难得多，从而进一步保护其身份。我们承认，人工模糊处理个人身份信息可能并不完善。

**赋能决策，而非代为决策**我们着重强调，我们的系统旨在 **赋能** 招聘人员，使其能够做出更明智的决策，而不是代替他们做决定。众所周知，大语言模型可能会出错，因此我们系统的评判结果也可能有误。蓄意操纵评判者信念的行为还可能加剧这一问题（Maloyan 等，2025年；Akdemir 和 Levy，2025年；Mu 等，2025年；Hwang 等，2025年；Suen 和 Hung，2024年）。我们强烈建议将本研究的任何具体实现都置于“人在回路中”系统的框架下（Jarrahi，2018年；Enarsson 等，2022年；Vats 等，2025年）。

**安全与偏见** 对我们的系统进行任何生产环境部署时，都应设置安全护栏，确保面试官不会询问受保护属性（例如性别、族裔等）。我们也承认，大语言模型可能表现出偏见（Gallegoset al.，2024年；Navigliet al.，2023年），尤其是在英语熟练度方面（Yanget al.，2025年；Zhou etal.，2024年），因此进一步凸显了人工核验与决策的重要性。

[4] https://www.reddit.com/r/Resumes/about/rules/

## 参考文献


Aisvarya Adeseye、Jouni Isoaho、Seppo Virtanen 和 Mohammad Tahir。2025年。采用动态问题生成和专业能力分析的模块化 AI 面试系统。预印本，arXiv:2601.11534。已接收；等待在 AIR-RES'25 上发表。

Arda Akdemir 和 Joshua H. Levy。2025年。理解并防范人力资源人工智能中的简历提示注入。载于第五届人力资源推荐系统研讨会论文集（*RecSys-in-HR 2025*），第4046卷，捷克共和国布拉格。CEUR-WS.org。与第十九届 ACM 推荐系统会议（RecSys 2025）同期举行。

Vikram R. Bhargava 和 Pooria Assadi。2024年。招聘、算法与选择：为何面试仍然重要。《商业伦理季刊》，34(2):201–230。

Marc Blatter、Samuel Muehlemann 与 Samuel Schenker。2012年。招聘熟练工人的成本。《欧洲经济评论》，56(1):20–35。

Chainlit 维护者。2026年。Chainlit（2.9.6 版）。https://github.com/Chainlit/chainlit。访问日期：2026-02-27。

Emil St. Chifu、Viorica Rozina Chifu、Iulia Popa 与 Ioan Salomie。2017年。一个用于从自然语言撰写的简历中检测专业技能的系统。载于 *2017* 年第*13*届*IEEE*智能计算机通信与处理国际会议（*ICCP*），第189–196页。

Steven Cho、Stefano Ruberto 与 Valerio Terragni。2025年。面向自然语言处理的大语言模型蜕变测试。载于 *2025*年*IEEE*软件维护与演化国际会议（*ICSME*），第174–186页。IEEE。

Deepro Choudhury、Sinead Williamson、Adam Golinski、Ning Miao、Freddie Bickford Smith、Michael Kirch- hof、Yizhe Zhang 和 Tom Rainforth。2026年。BED- LLM：利用大语言模型和贝叶斯实验设计进行智能信息收集。载于第十四届国际学习表征会议。

Therese Enarsson、Lena Enqvist 和 Markus Naart- tijärvi。2022年。走向人在回路中——三个情境下关于人类与算法混合决策的法律视角。《信息与通信技术法》，31(1):123–153。

Emily Fisher、R. S. Thomas、M. K. Higgins、C. J. Williams、I. Choi 和 L. A. McCauley。2021年。寻找合适的候选人：制定临床研究协调员职位申请者筛选的招聘指南。《临床与转化科学杂志》，6(1):e20。

Isabel O. Gallegos、Ryan A. Rossi、Joe Barrow、Md Mehrab Tanjim、Sungchul Kim、Franck Dernon- court、TongYu、Ruiyi Zhang 和 Nesreen K. Ahmed。2024年。大语言模型中的偏见与公平性：一项综述。《计算语言学》，50(3)：1097– 1179。

Google DeepMind。2026年。Gemini 3.1 pro：模型说明卡。https://deepmind.google/models/model-cards/gemini-3-1-pro/。访问日期：2026-02-28。

Ekai Hashimoto、Mikio Nakano、Takayoshi Sakurai、Shun Shiramatsu、Toshitake Komazaki 和 Shiho Tsuchiya。2025年。使用基于大语言模型的动态槽位生成的职业面试对话系统。载于第*31*届国际计算语言学会议论文集，第1562–1584 页，阿布扎比，阿联酋。计算语言学协会。

Jim Hlavac。2023年。知识、技能与能力（ksas）作为重新概念化资质的指标：多方利益相关者视角。《口译与笔译培训者》，17(1):29–53。

Yerin Hwang、Dongryeol Lee、Taegwan Kang、Yongil Kim 和 Kyomin Jung。2025年。你能骗过评分器吗？对大语言模型评审器的对抗性说服。载于计算语言学协会论文集：*EMNLP 2025*年研究成果，第14632–14651页，中国苏州。计算语言学协会。Sangwon Hyun、Mingyu Guo 和 M. Ali Babar。2024年。MetaL：用于分析大语言模型特性的蜕变测试框架。载于*2024*年 *IEEE* 软件测试、验证与确认会议（*ICST*），第117–128页。Mohammad Hossein Jarrahi。2018年。人工智能与工作的未来：组织决策中的人类与人工智能共生。《商业视野》，61(4):577–586。Siswanto Joko、Sinung Suakanto、Made Andriani、Margareta Hardiyanti 和 Tien Kusumasari。2022年。运用自然语言处理和机器学习开发面试机器人。《国际技术杂志》，13:274。Mitra Lashkari 和 Jinghui Cheng。2023年。“找到神奇配方”：探索招聘人员和求职者对招聘偏见及自动化工具的看法。载于 *2023*年 *CHI*计算机系统人因会议论文集，美国纽约州纽约。计算机协会。Changmao Li、Elaine Fisher、Rebecca Thomas、Steve Pittard、Vicki Hertzberg 和 Jinho D. Choi。2020年。使用上下文感知型变换器模型进行胜任力水平预测以及简历与职位描述匹配。载于 *2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第8456–8466页。计算语言学协会。

Frank P.-W. Lo、Jianing Qiu、Zeyu Wang、Haibao Yu、Yeming Chen、Gao Zhang和Benny Lo。2025年。基于大语言模型的人工智能招聘：一种面向简历筛选的情境感知且可解释的多智能体框架。载于*2025*年*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*），第4184–4193页，美国加利福尼亚州洛斯阿拉米托斯。IEEE计算机学会。

Ehsan Lotfi、Maxime De Bruyn、Jeska Buh mann和Walter Daelemans。2023年。PersonalityChat：结合事实与特质、用于个性化对话建模的对话蒸馏。载于第三届自然语言生成、评测与指标研讨会（*GEM*）论文集，第353–371页，新加坡。计算语言学协会。

Narek Maloyan、Bislan Ashinov 和 Dmitry Namiot。2025年。研究基于大语言模型的评判者架构对提示注入攻击的脆弱性。《国际开放信息技术期刊》，13(9):1–6。

Honglin Mu、Jinghao Liu、Kaiyang Wan、Rui Xing、Xi- uying Chen、Timothy Bald win 和 Wanxiang Che。2025年。核心领域之外的人工智能安全：以简历筛选为例，探讨专用大语言模型应用中的对抗性漏洞。预印本，arXiv:2512.20164。Samuel Mühlemann 和 Mirjam Strupler Leiser。2018年。招聘成本与劳动力市场紧张程度。《劳动经济学》，52:122–131。Roberto Navigli、Simone Conia 和 Björn Ross。2023年。大语言模型偏见：起源、分类与探讨。《数据与信息质量期刊》，15(2)。OpenAI。2025年。GPT-5 发布介绍。访问日期：2026-02- 27。Sean D. Risavy、Chet Robie、Paige A. Fisher 和 Sadaf Rasheed。2022年。简历与申请表：为何人们仍顽固地依赖简历？《心理学前沿》，13:884205。Jeffery S. Shippmann、Ronald A. Ash、Mariangela Batjt- sta、Linda Carr、Lorraine D. Eyde、Beryl Hesketh、Jerry Kehoe、Kenneth Pearlman、Erich P. Prien 和 Juan I. Sanchez。2000年。胜任力建模实践。《人事心理学》，53(3):703–740。Hung-Yue Suen 和 Kuo-En Hung。2024年。揭示人工智能及其界面对异步视频面试中求职者诚实与欺骗性印象管理的影响。《技术预测与社会变迁》，198:123011。

Vanshika Vats、Marzia Binta Nizam、Minghao Liu、Ziyuan Wang、Richard Ho、Mohnish Sai Prasad、Vincent Titterton、Sai Venkat Malreddy、Riya Aggarwal、Yanwen Xu、Lei Ding、Jay Mehta、Nathan Grinnell、Li Liu、Sijia Zhong、Devanathan Nallur Gandamani、Xinyi Tang、Rohan Ghosalkar、Celeste Shen 等另外 4 位作者。2025年。关于人机协作与大型基础模型的综述。预印本，arXiv:2403.04931。

Vinod U. Vincent、Rebecca M. Guidice 和 Neal P. Mero。2024年。应该相信直觉吗？专业专长对复杂工作直觉性招聘决策的影响。管理与组织期刊，30(4):910–930。

William Wooten。1993。运用知识、技能与能力（KSA）数据识别职业路径规划机会：将工作分析应用于内部人力规划。公共人事管理，22(4):551–563。

Kaixun Yang、Mladen Rakovi´c、Dragan Gaševi´c 和 Guanliang Chen。2025年。基于提示的大语言模型能否识别学生的人口统计特征，并在作文评分中引入偏见？载于教育中的人工智能，第75–89页，沙姆。施普林格·自然瑞士。

Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong、Yifan Du、Chen Yang、Yushuo Chen、Zhipeng Chen、Jinhao Jiang、RuiyangRen、Yifan Li、XinyuTang、Zikang Liu 及其他3位作者。2025年。大语言模型综述。预印本，arXiv:2303.18223。

Hongli Zhou、Hui Huang、Yunfei Long、Bing Xu、Con- ghui Zhu、Hailong Cao、Muyun Yang 和 Tiejun Zhao。2024年。减轻大语言模型评估中的偏见。载于第*23*届中国计算语言学全国会议论文集（第*1*卷：主会场），第1310–1319页，中国太原。中国中文信息学会。

Michelle X. Zhou、Carolyn Wang、Gloria Mark、Huahai Yang 和 KevinXu。2019 年。构建真实世界的聊天机器人面试官：绿野仙踪实地研究的经验教训。载于与第 *24* 届 *ACM* 智能用户界面会议（*ACM IUI 2019*）同期举办的 *ACM IUI 2019* 研讨会联合论文集，第 2327 卷。

## 评审校准分层

我们在表3中展示了各校准测试类别的部分分层结果，以突出有意义的子集。对于不变式测试，我们按输入方案（包含或排除存根，$\mathcal{I}$）进行分层，并观察到重复中性能略有下降，$\mathcal{I}$。一种可能是，$\mathcal{I}$ 设定了特定的面试轨迹，随后被 $\tau$ 打乱，从而增加了证据去重负担。我们按评分标准领域对目标级别提升测试进行分层，发现 PBA 评判器在销售领域正确识别较少证据时略有困难；对于机器学习工程领域的中等证据，识别起来则更加困难。最后，我们按贬损类型对目标级别降低测试进行分层，发现两位评判者在范围降级方面遇到的困难最大。

表3：按评判者划分的测试通过率（四舍五入后的百分比）。

| Test | 独立 | PBA |
|---|---|---|
| **不变式** ($\mathcal{R}_{\text{inv}}$) | | |
| 夸大 | 2 | **100** |
| ↪ 无存根 | 3 | **100** |
| ↪ 有存根 | 0 | **100** |
| 不相关性 | 2 | **98** |
| ↪ 无存根 | 0 | **100** |
| ↪ 有存根 | 3 | **97** |
| 重复 | 0 | **87** |
| ↪ 无存根 | 0 | **90** |
| ↪ 有存根 | 0 | **83** |
| **目标级别提升** ($\mathcal{R}_{\uparrow}(d,\ \ell)$) | | |
| 低证据添加 | **92** | 90 |
| ↪ 机器学习工程 | 85 | **95** |
| ↪平面设计 | **100** | 95 |
| ↪ 销售 | **90** | 80 |
| 中等证据添加 | 72 | **88** |
| ↪ 机器学习工程 | **75** | **75** |
| ↪平面设计 | 65 | **100** |
| ↪ 销售 | 75 | **90** |
| 高证据添加 | 95 | **97** |
| ↪ 机器学习工程 | **100** | **100** |
| ↪平面设计 | 90 | **100** |
| ↪ 销售 | **95** | 90 |
| **目标级别降低** ($\mathcal{R}_{\downarrow}(d,\ell)$) | | |
| 低证据贬低 | 73 | **85** |
| ↪ 语境失效 | 71 | **83** |
| ↪ 负向推理澄清 | 75 | **86** |
| 中等证据贬低 | **73** | 70 |
| ↪严谨性降级 | **75** | **75** |
| ↪范围降级 | **67** | 50 |
| 高等证据贬损 | **95** | 92 |
| ↪严谨性降级 | **97** | 91 |
| ↪范围降级 | **91** | **91** |
| **其他** | | |
| 均匀先验 | **87** | 80 |

**界定贬损。**

- **•语境失效：** 该行为确实发生过，但可以用合理的限制、角色边界、适当性或带宽不足来解释，因此不能作为能力或判断力低下的有效证据。
- **• 负面推断澄清：** 早先的回答听起来 像是能力不足，但后续澄清消除了歧义（措辞、简略表达或过度简化所致），从而排除了负面推断。
- **• 严谨性降级：** 后续信息表明，早先的证据并不像最初所暗示的那样系统、审慎、正式或一贯。
- **• 范围降级：** 后续信息缩小了申请人个人承担、负责或贡献的范围，相较于早先证据所暗示的程度有所降低。

## B 原型恢复误分类

为探究误分类，我们在图 5中绘制了混淆矩阵。我们发现，评判者在恢复属于机器学习工程领域的原型时遇到的困难最大。这与我们的观察一致：与其他领域相比，PBA评判器在对齐中等水平的机器学习工程证据方面能力较弱。值得注意的是，机器学习工程评分标准仅包含六个维度，而平面设计和销售分别包含七个和八个维度。虽然我们不主张表现与评分标准维度数量之间存在关联，但有理由认为，维度数量较多的评分标准可能更有助于区分知识、技能与能力，从而更好地满足第*2.1*节中定义的评分标准内部一致性要求。我们还发现，误分类往往会选择公式（9）中定义的锚定原型。衡量申请人的整体水平，可能比在相互对比的知识、技能与能力之间分配信念更容易。

最后，我们以逐级粒度考察这些误分类的结构。使用仅简历信念状态（$t = 0$），所有档案-维度对（$p$，$d$）中最大后验级别频率的归一化结果为（$f_{\text{low}}$，$f_{\text{medium}}$，$f_{\text{high}}$）=（0.39，0.58，0.04）（四舍五入后）。这表明

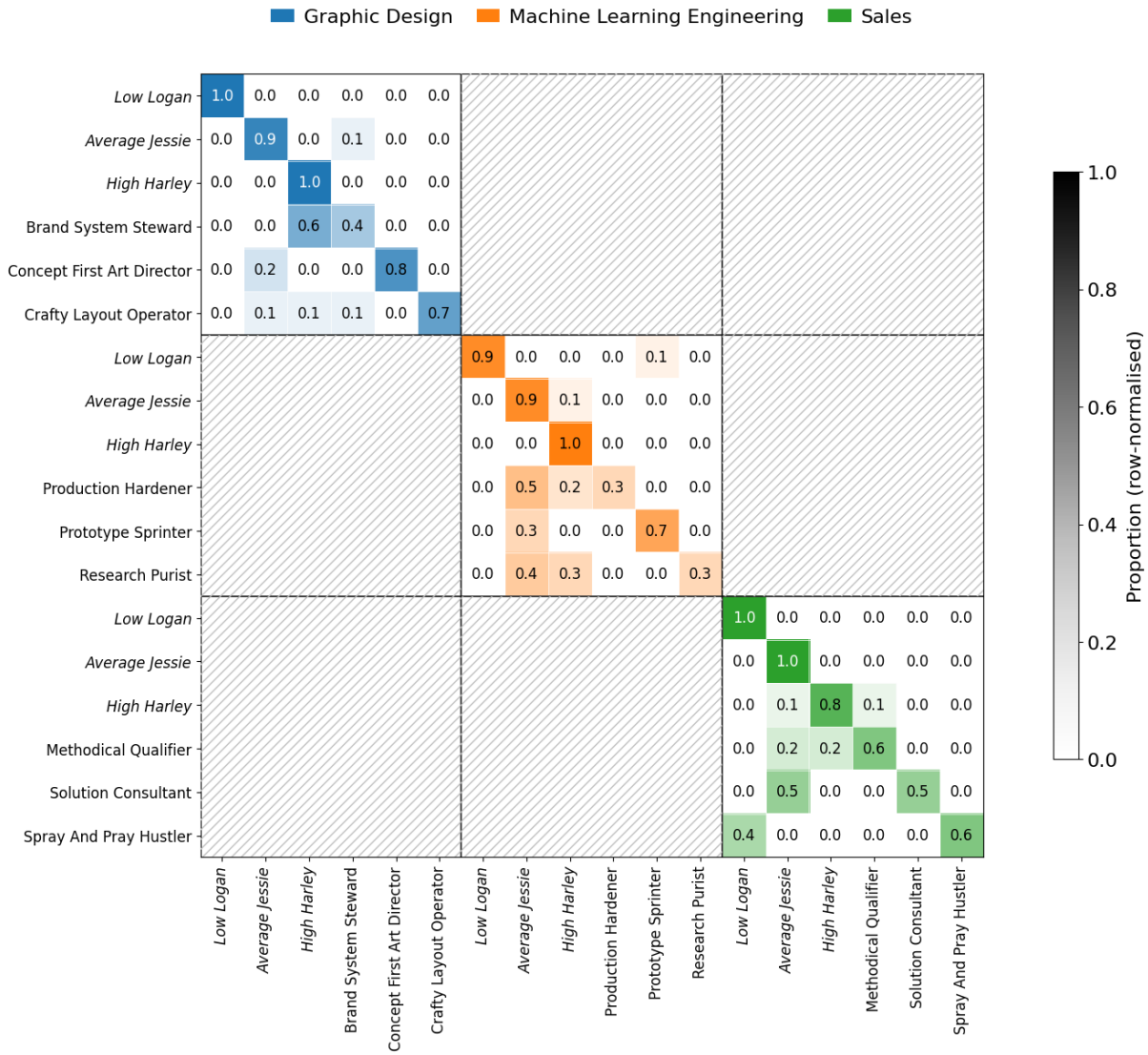



图5：完整模拟集上原型恢复的混淆矩阵（$|\mathcal{P}_{\text{full}}| = 180$）。斜体原型是我们先前定义的锚定原型，公式（9）。

| 模型 | 全变差 ($\Delta_0 \to \Delta_T$) | 原型恢复 ($0 \to T$) |
|---|---|---|
| GPT-5 | 0.0621 → 0.0205 | 16.7% → 76.1% |
| Gemini 3.1 Pro | 0.0640 → 0.0053 | 20.0% → 100% |

表 4：GPT-5 与 Gemini 3.1 Pro 在 $P = 30$档案上的收敛与原型恢复（原型恢复）比较。

单靠简历很少能支持高级别的分配，大部分概率质量集中在中等。结合第3.5 节的观察，即PBA评判器在从中等中移除概率质量方面相对不活跃，这或许可以解释为何两端的误分类大多都趋向于中等，如图6所示。

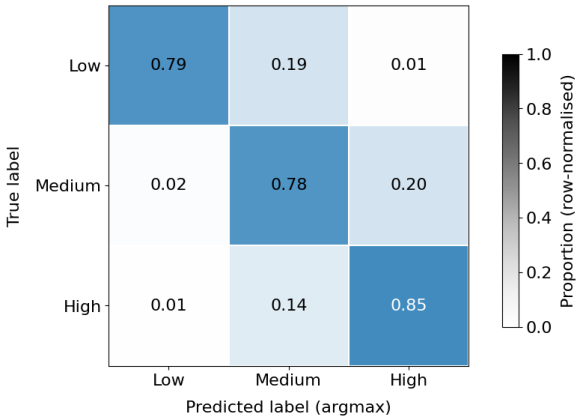



图6：完整模拟集（$P_{\text{full}} = 180$）中各层级的误分类。