集成贝叶斯推理：利用小型语言模型在配置文件匹配任务中实现 LLM 级别的准确性

Haru-Tada Sato, Fuka Matsuzaki and Jun-ichiro Takahashi
Department of Data Science, i's Factory Corporation, Ltd.
Kanda-nishiki-cho 2-7-6, Tokyo 101-0054, JAPAN
{ satoh, matsuzaki, jtakahashi } @isfactory.co.jp

Abstract

本研究探讨了小型语言模型（SLM）集成在实现与专有大型语言模型（LLM）相当的准确性方面的潜力。我们提出了集成贝叶斯推断（EBI），这是一种新方法，它应用贝叶斯估计来结合多个 SLM 的判断，从而使它们超越单个模型的性能限制。我们在各种任务上的实验——包括日语和英语的能力评估和消费者档案分析——展示了 EBI 的有效性。值得注意的是，我们分析了一些案例，其中将具有负提升值的模型整合到集成中可以改善整体性能，同时我们还检验了该方法在不同语言下的效力。这些发现为利用有限计算资源构建高性能 AI 系统以及有效利用单个性能较低的模型提供了新可能。基于现有关于 LLM 性能评估、集成方法和开源 LLM 利用的研究，我们讨论了我们方法的新颖性和重要性。 [1]

关键词：贝叶斯推断，集成学习，分析任务，专有大语言模型比较，小型语言模型

1 介绍

在医学诊断领域，有报告称 GPT-4 在诊断准确性方面超过了医生，无论是独立还是作为助理 [1, 2]，表现相近 [3]，或未能展示出改进 [4]。尽管有些研究强调了专业知识和经验的不可替代性，并对 AI 在问题解决中准确识别常识的能力表示谨慎，但研究也表明其作为支持工具的有效性 [5, 6]。各种诊断任务已被评估，包括从测试结果和案例研究的预测准确性 [1, 3]，以及诊断总结内容 [2]。

在心理学领域，人们正在进行比较实验，以确定人工智能是否可以替代人类的智力活动，如决策和认知 [7] - [11]。在创造性构思中，人工智能展现出与人类相当或超越人类的新颖性，尽管在可行性上有些许不足 [7]。心理学实验评估人工智能是否可以替代人类对象 [8] 或做出类似于人类的行为判断 [9]。涉及人类判断的实验在定量评估上存在挑战，包括由于人类认知偏见造成的公平性问题 [10]，评估中的主观概念处理 [11]，对认知负荷的响应能力 [12]，以及评估中的可变性 [13]。

在这样的背景下，本文将人识别作为一个在 AI 和人类之间进行比较实验的判断任务来研究。判断的目标是评价个人特征的文本。任务包括比较关于同一组个体的两种类型的评价评论，并识别每个人，样本示例在附录 A 中说明。评价文本包含与评估者相关的噪声表达或潜在误导性陈述，这增加了误判的风险。在面对这样的输入时，提高推理的鲁棒性和可靠性是一个重大的挑战。

为了使用人工智能执行如此复杂和精细的判断，我们提出了一种判断方法，该方法结合了贝叶斯推断（BI）的概率判断和基于置信水平的主观元素，同时考虑人类判断特性（变异性和模式一致性 [13]）。虽然 BI 方法是人类判断过程的简化模型，但生成性人工智能也表现出判断的变异性，我们的目标是通过对 BI 的集成方法来适应这种变异性。这是我们称之为集成贝叶斯推断（EBI）的方法的关键理念。

虽然生成式 AI 的集成方法 [14] - [17] 着重于提高模型的准确性，但本文考察了判断多样性是否有助于准确性提升，重点关注最近出现的轻量级开源 LLM。由于解决与高级提示技术相关的计算成本和延迟变得至关重要——特别是在针对具有严格性能要求的现实世界应用时——我们的工作也旨在验证一个框架，通过利用高速处理芯片的优势进行大容量处理来弥补与专有 LLM 的性能差距。

2 贝叶斯推理方法的任务设计

任务问题设置涉及在给定所有信息 $X = B(b_i)$ 时确定哪个元素 $A(a_j)$ 对应同一位个体，其中包含关于一组个体的两组个人信息 $B = \{b_1, b_2, \cdots, b_n\}$ 和 $A = \{a_1, a_2, \cdots, a_n\}$。在这里，$A$ 和 B 表示根据不同的视角描述个人资料的文本集。所谓“不同的视角”，是指具有某种关系的领域，例如饮食习惯和健康卫生。

目标是估计每个人的特征档案，并通过档案分析判断他们是否为同一个个体。在技术上，主要目标是通过从 SLMs 获得频繁的响应并实施概率判断过程来实现可与专有 LLMs 相媲美的判断准确性。超过人类判断的表现程度也是一个重要的关注点。

[1] 代码和数据可在 https://github.com/maskcode9004/ebi_method 获取

2.1 置信矩阵和主观程度

我们技术的基本概念是贝叶斯定理。事件 A 在事件 X 发生之后发生的条件概率（后验概率）$P(A|X)$ 由似然性 $P(X|A)$ 和 $P(A)$ 给出，即事件 A 发生的概率（先验概率或主观概率）：

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}, \tag{1}$$

其中 $P(X) = \sum_A P(X|A)P(A)$ 。

当给定了 $B = \{b_1, b_2, \cdots, b_n\}$ 的一个元素 b_i 时，其对应于 $A = \{a_1, a_2, \cdots, a_n\}$ 的一个元素 a_j 的后验概率提供了置信矩阵 $\text{conf}_{B\to A}$ （统计或计算置信度，在不需要区分时简称为置信度）。它的元素，记作 $X = B(b_i)$ ，由贝叶斯定理 (1) 给出：

$$(\text{conf}_{B\to A})_{ij} = P(a_j|b_i) = \frac{P(b_i|a_j)P(a_j)}{P(b_i)}. \tag{2}$$

一个实验考虑是，当从多个查询中聚合响应以计算置信度时，从未在响应中出现的元素将导致 $P(X) = 0$ ，这可能导致在简单聚合中矩阵元素的分歧。因此，我们应用正则化以确保 $P(X)$ 具有较小的值 ε 。由于 $P(X|A)$ 值也成比例较小，方程 (1) 产生有限值。具体来说，没有响应的项目在聚合中被赋予 $\varepsilon = 0.1$ 的值，以避免除以零。

相反地，当从 $A(a_j)$ 判断 $B(b_i)$ 时，我们获得作为观察值的信心水平。为了将其与上述信心水平区分开来，我们称之为主观信心水平（主观度）。在投票系统中，这可以是收到的票数比例（集体主观度）。当这一主观度被观察为 c_{ji} ，即观察矩阵（主观度矩阵），则似然度 $P(b_i|a_j)$ 由 $(\text{conf}_{A\to B})_{ij} = c_{ji}$ ，$(0 \le c_{ji} \le 1)$ 给出。然后我们获得主观概率：$P(a_j) = \frac{1}{C}\sum_i c_{ji}$ ，其中 $C = \sum_{ij} c_{ji}$ ，得到

$$P(b_i) = \sum_j P(b_i|a_j)P(a_j) = \frac{1}{C}\sum_{jk} c_{ji}c_{jk}. \tag{3}$$

。将这些代入 (2) 得到信心矩阵。

2.2 判断矩阵

问题设置为确定给定信息 $X = b_i$ 时，哪个 a_j 对应于 b_i 。简单的贝叶斯估计在信息不对称时可能效果不佳。也就是说，当技能、视角或背景的差异影响对 A 和 B 的解释时，诸如 MLM 和 GPT 之类的概率 LLM 方法可能不能可靠地执行。

无论使用哪种方法，如果判断结果的可靠性可以通过某些评估权重 s_{ij} 来表达，将其与置信度相乘来定义判断矩阵 J 似乎是有效的：

$$J_{ij} = s_{ij}P(a_j|b_i). \tag{4}$$

。J 的 ij 组件可以解释为每个候选 b_i 的可能性，因此这个 J 用于身份判断。该方法允许基于定义 s_{ij} 和观测矩阵 c_{ji} 的自由度来构建各种模型。

2.3 系统设置

在本研究中，我们使用小型语言模型（SLMs）来确定 b_i 和 a_j 是否代表同一人以创建 s_{ij} 。然而，为了扩大系统选项的范围，我们测试了多种不同的语言模型来计算 s_{ij} 和 c_{ij} 。为了获得观察矩阵 c_{ij} 和权重矩阵 s_{ij} ，我们考虑两种方法：（i）基于集体主观度（ID 选择频率）的评估和（ii）基于直接主观度的评估。这些通过以下类型的提示处理实现：

类型 1. 通过多次提交相同的问题获得响应中的 id 频率聚合

类型 2. 通过在响应评估过程中询问主观置信水平来收集值

提示结构因类型而异，但询问哪个 id 是相同的基本查询和分析方法的说明在两者之间是共同的。共同的不是具体描述性的表达，而是概念。本质的区别在于输出格式的规定：类型 1 仅输出被判断为相同的 id 对（见附录 B.1 ），而类型 2 则输出一个候选 a_j 的数组，按主观程度降序排列，并附上它们的主观程度（见附录 **??** ）。这些汇总值按问题的数量进行归一化处理。

在计算观测矩阵和权重矩阵时，响应会因所使用的语言模型和所给的提示（类型和描述性表达）而有所不同，因此我们将一个 *语言模型和提示配置* 视为一个系统。此外，我们通过计算从各个系统得到的判断矩阵 J 的加权平均值，来考虑集成系统（EBI）。

对于语言模型选择标准，我们主要采用轻量级的 SLM，以速度优先于准确性为原则来收集多个响应，不过在某些情况下我们也会考虑 70b 模型，并通过 WSE（Cerebras 晶圆级引擎和 CS-3 系统）或 LPU（Groq LPU™ 推理引擎）进行高速处理。EBI 中使用的模型包括：GroqChat：gemma2-9b-it，mixtral-8x7b-32768，llama3-8b-8192，llama3-70b-8192。Cerebras：llama-3.1-70b-versatile。OpenAI：gpt-4o-mini-2024-07-18。

3 数据生成和评估方法

我们基于来自多个视角的项目（属性值和观察值）为相同的个人生成两个不同的人物画像。我们将用于生成的项目分成两部分，允许重叠，并创建两种类型的人物数据，A 和 B 。在创建这些人物时，我们指定不同的视角，以确保它们不能仅通过词匹配（相似性）轻易关联。我们准备了两种类型的数据集：能力评估和购买历史。样本数据请参阅附录 A 。

对于能力评估（prof1j），具体的创建流程如下：观察数据由 50 人的能力测试结果组成，共有 14 个观察项目（项目包括社交回避、自我反思、任务持久性、风险回避水平等）。字符诊断评论是基于项目名称和数值使用 GPT-4o 生成的，以创建数据集 A 。添加了部门、服务年限和其他工作属性，以类似方式生成工作评估评论，创建数据集 B 。在创建评论时，指定了不同的视角，以确保 A 和 B 不会生成相同的评论。对于 A ，其目的是鼓励自我提升以提高表现，而 B 则侧重于与工作执行相关的问题。此外，通过使用 GPT-4o 翻译此数据到英语，创建了另一个数据集“prof1e”。

对于购买历史数据（prof2j），将 18 种购买的产品项目进行了划分，并使用与上述类似的过程创建了人物角色数据。在这种情况下，由于人物角色形象可能会受到两个划分中的产品构成的显著影响，因此，A 和 B 都以两种视角创建：行为模式和产品选择中的消费者价值。该数据的英文翻译为“prof2e”。

对于性能评估指标，准确度 Acc $= n_c/N$ （数据总数 $N = 50$ ，正确答案数 n_c ）是常用的，但如果数据集 A 和 B 的背景上下文不同，那么词语的背景知识和概念相似性的范围在它们之间也可能有所不同。在这种具有概念波动和不确定范围的情况下，尚无已知方法可以客观地量化元概念的一致性。因此，一个关键的考虑因素是，由于文本内容的性质、抽象程度、可解释性和评估者的背景知识不同，尺度可能会显著变化，从而使不同数据集之间的比较无法进行。因此，这一指标只能用于同一类型数据中的相对优劣判断，这是一个缺点。

因此，使用提升率（Lift）来进行评估，它表明相对于人工评判有多少优势，通过适度消除数据依赖性（尽管前提是参考评估者应保持不变），能够进行更适当的比较。给定参考评估者的最大正确答案数 H ，提升率是 $= 100(\frac{n_c}{H} - 1)$ [%]。

当难以获得人类评估者时，可以用高精度的专有大型语言模型（LLM）代替人类评估者。在这种情况下，使用高精度专有 LLM（ChatGPT-4o）的最大正确答案数量 G ，我们评估使用达到率 Reach $= 100n_c/$Base [%]（其中 Base 对人类为 H ，对 LLM 为 G ）可以达到多接近。对于参考评估者 LLM，我们采用一种通过提示逐步构建推理的方法来尽可能接近人类判断的方法，而不是使用正在检验的 EBI（详见第 4 节）。

对于我们的英语验证，其中 H 是未知的，我们使用日本验证中的平均 H/G 比率 γ ，将 $H = G\gamma$ 替换到升力公式中来近似升力为 $\text{Lift}_{eff} = \text{Reach}/\gamma - 100$ 。我们从人工评估中获得了 (prof1j, prof2j) 的 H 值为 (19, 13)。对于四个数据集 (prof1j, prof2j, prof1e, prof2e)，获取的 G 值为 (22, 20, 28, 23)。因此，H/G 比率为 prof1j 的 19/22 和 prof2j 的 13/20，得到 $\gamma = 0.757$ 。对于英语数据，我们使用 $\text{H}_{eff} = \text{G}\gamma$ ，因此采用四个数据集的基线 H 值为 (19, 13, 21.2, 17.4)。人的准确率值为 (38 %, 26 %, 42 %, 35 %)，GPT-4o 的准确率值为 (44 %, 40 %, 56 %, 46 %)，相对于人的升力值为 (15.8 %, 53.8 %, 32.1 %, 32.2 %)。

4 顺序思考法

即使是专有的 LLM 也需要动态提示技术——其中语言模型通过自我反思来提高任务性能——以确保至少有人类水平的评估准确性 [15, 18, 19] 。由于简单的任务指令可能导致 LLM 的重复判断，本文应用反馈-反思-改进机制 [15] ，通过收集 LLM 输出的错误信息并将其用作反馈来解决冲突。

该提示使用两种类型的模板来进行解决和验证：(i) 解决提示：指示模型在提供任务对应答案之前用一句话解释原因。(ii) 反馈提示：分析当模型输出不正确时的原因并生成新的提示。

相比之下，EBI 方法通过强制性的处理而非基于判断的判决来消除重复。具体来说，它遵循以下步骤：1. 在判断矩阵 J 中选择输出值最高的元素。2. 移除包含所选元素的行和列（即形成一个子矩阵）。3. 在子矩阵中选择输出值最高的元素。4. 针对判断矩阵的行数（即要预测的个体数）重复上述步骤。

在顺序方法中，通过属性（如年龄）过滤的子集内逐一分析候选对象（系统提示 S1），并以竞赛风格（顺序处理 S2）缩小候选范围。由于初始候选对象选择错误会影响后续判断，因此当判断结果存在冲突或不一致时，会纳入审查结果的流程（递归处理 S3）。由于审查过程中也可能发生冲突，因此重复检查和判断，直到冲突解决（冲突处理 S4）。具体的提示 S1-S4 请参见附录 C 。顺序处理的纲要如下：

1. 程序从 A (Aset) 中获取与人口属性（如年龄）匹配的 id 集合。元素的数量记为 n 。当 Aset 改变时，消息历史将被清除（会话切换）。
2. 系统提示 S1：指定数据描述和分析任务的方法。
3. 顺序处理 S2：如果 $n = 1$ ，识别被确认，过程移动到下一个 Aset。如果 $n \geq 2$ ，从初始值 $k := n$ 开始，元素在每次识别后从 Aset 中移除，直到 $k = 1$ 。
4. 当 Aset 中剩余的人数达到指定值时，递归处理 S3 被激活。
5. 冲突解决 S4 在递归处理中执行。循环继续，直到重复计数器达到 0。

5 结果

在本研究中，我们在四个数据集上进行了实验，以检验 SLM 集成判别系统是否能达到与人类或专有的 LLM 判别相当的判别准确性。每个系统的 BI 和 EBI 方法的完整结果列在附录 D 中。

图 ?? 总结了日本数据集的结果，图 ?? 总结了英语数据集的结果。综合考虑所有四种类型的数据集，我们发现：(1) 单个 BI 系统可以实现提升 > 0；(2) 使用 EBI 可以让我们获得比单个 BI 系统更高提升的系统。在每个图中，考虑提升 $\gtrsim 30$ % 和覆盖 $\gtrsim 100$ % 为高性能区域，除了 prof1e（能力/英语）以外的所有数据集都达到这些区域，甚至在使用 EBI 时，prof1e 几乎达到了该区域的下限（空心标记表示比实心标记更高的提升值）。除了在 prof2j（购买/日语）中的系统 59（llama3.1-70b）以外，我们能够为所有数据集通过 EBI 系统（相比单个 BI）实现更高的准确度。然而，值得注意的是，由于可以通过增加测试的模型和系统数量来无限制地增加绘图的数量，因此没有一种公平的方法可以比较准确度或提升的平均值，因此适合将这些理解为趋势。

5.1 单个数据集

这些结果表明，使用 EBI 方法的 SLM 集成可以突破单个 BI 系统的性能限制，并实现与 LLM 相当的准确性，在某些情况下甚至超过 LLM。此外，检查构成集成的单个 BI 系统的性能显示，适当地组合表现较差的系统可以增强预测的鲁棒性。跨不同数据集和语言确认集成效应是一个重要发现，表明该方法的多功能性。具体来说：

对于 prof1j（能力/日语），多 SLM 单独表现的评估显示，有些系统达到了最高的提升值 21 %（37:gemma2-9b-it，42:llama3-8b-8192，76:llama3.1-70b-versatile）。详细信息在附录 D.1 中。通过构建包含这些高性能单系统的平均系统，提升值和覆盖值（以下简称，LR 值）从单一系统显著增加到最大值（36.8 %，118.2 %）。值得注意的是，在高精度集成系统中（限制为提升值 > 0 ），确认了某些实例中，将具有负提升值的单一系统（66:gpt-4o-mini-2024-07-18，12 和 40:mixtral-8x7b-32768，13:llama3-70b-8192 等）组合后，得到正的提升值。这表明 EBI 集成可以弥补单个表现较低系统的弱点，从而实现更稳健的预测。例如，表 4 显示的集成系统，尽管包含了负提升值的组件，但仍实现了正的提升值。

在附录 D.2 中总结的 prof2j（采购/日语）实验中，构建了以在两个不同的日语数据集上实现正提升值的系统（37、46、64、76）为中心的集成系统。尽管 Reach 未达到 100 %，但也确认了一些实例，其中将具负或零提升值的弱系统（40、42、43）加入集成中仍然能够导致正提升值。在特定情况下，表现最佳的系统 82（在表格 6 中）包含了 llama3-70b 系统 46 和 76，通过将它们整合为一个平均系统，它取得了（38.5 %，90 %）的结果，超越了每个组件系统的最大 LR 值（30.8 %，85 %）。

在附录 ?? 中的（能力/英语）图谱中，尽管集成系统的最大 LR 值（27.4 % ,96.4 % ）没有达到 100 %，但它实现了显著高于单一系统的 LR 值。同样，通过集成效果观察到对单一系统的性能提升，结合了 37（gemma2-9bit），40（mixtral-8x7b-32768），43（llama3-8b-8192）和 76（llama3.1-70b-versatile），其中 37 和 43 为弱学习者，提升值为 < 0 ，如表 7 所示。

在 prof2e（purchase/English）案例（附录 ?? ）中，表 10 显示，通过组合多个系统的集成系统实现了显著高于单一系统的 LR 值，最高达到（78.2 %，134.8 %）。具体而言，通过将系统 37（gemma2-9b-it），40（mixtral-8x7b-32768），43（llama3-8b-8192），46（llama3-70b-8192），66（gpt-4o-mini-2024-07-18）等以不同权重组合，观察到了显著的性能提升。系统 40 和 43 作为弱学习者在表 9 中显示了提升值为 < 0 。

6 结论

从本研究获得的实验结果可以得出以下结论：1. 通过 SLM 集成提高性能：使用像 EBI 这样的集成技术可以组合多个轻量级 SLM，以超越单个系统的性能限制，获得可与 LLM 相媲美的准确性。这表明在计算资源有限的环境中构建高性能 NLP 系统是可能的。

2. 弱学习者的利用：在一些实例中，通过将显示负提升值的系统，即单独表现不佳的系统，加入到集成中提高了整体性能。这重新评估了弱学习者在集成学习中的角色，并展示了融合多个具有不同视角的系统的有效性。

3. EBI 方法的有效性：与简单平均或其他方法相比，EBI 方法建议使用集体主观程度或直接主观程度进行加权，更有效地整合个人系统判断并提高集成性能。

4. 方法的多功能性：通过 SLM 集成体的性能改进在不同任务上得到了证实，例如能力评估和消费者分析，以及在包括日语和英语在内的不同语言中。这表明，该方法不仅限于特定任务或语言，而是具有广泛的适用性。

然而，这项研究中也发现了一些挑战。由于这些指标是基于与人工评估的比较，因此需要多个数据集来确保可靠性。为了找到一个最佳组合，采用了一种启发式方法，涉及使用多个表现较差的模型进行反复试验和错误。一个局限是缺乏系统和有效的方法来识别哪些弱学习者对组合有有效贡献。尽管 BI 方法本身是基于统计处理，因此避免了过度调整的风险，但组合构建过程仍可能容易受到过拟合或过度调整的影响。

未来的研究应集中于通过使用更多样化的数据集验证 EBI 方法的有效性，研究有效弱学习者的系统选择标准，引入对各种自然语言数据的人类评估，以及开发无需过多试错即可高效优化集成结构的方法。

总体而言，这项研究探讨了在资源有限的情况下构建高性能 AI 系统以及有效利用单独低性能系统等对 AI 技术的广泛采用和发展具有重大意义的主题。特别是，SLM 集成和 EBI 方法的结合代表了实现更高效和强大的 NLP 的一个有前景的方向。通过克服这些挑战并扩大这些方法的应用范围，预计 SLM 集成将来会发展成更实用的技术。

A 输入数据样本

本附录旨在提供身份匹配实验中使用的数据示例，展示如何在数据集 A 和 B 中对同一个人进行不同的描述。这两种数据集（A 和 B）被分配了不同的识别号码（id_A 和 id_B），并进行实验以识别同一个人的号码。示例包括基于购买数据的能力评估和行为分析。

Table 1: 工作能力评估的识别数据样本。属性信息：id_B=1，id_A=25，类型 =1，年龄 =30。

Category	Description
Strength(A)	Highly responsible, perseverant
Weakness(A)	Prone to stress, averse to change
Assessment(A)	The person has a strong sense of responsibility and perseverance to complete assigned roles. They may feel uneasy about adapting to changes, but by clearly defining goals, schedules, and expectations, they can effectively lead team ...
Personnel(B)	As a manager, the employee leads the team and delivers the expected results. To further enhance the overall output of the team, please focus on strategic goal setting, progress management, and member development. It is also ...

Table 2: 用于购买行为分析的已识别数据样本。属性信息：id_B=31, id_A=1, 性别 =2, 年龄 =60。

Category	Description
Style(B)	She values fermented foods and prefers natural foods and nutrient-rich ingredients.
Persona(B)	She has a strong interest in fermentation and health, especially favoring fermented foods like yogurt and kimchi. She is a loyal customer who makes regular purchases and is interested in new fermentation trends, ...
Style(A)	Her loyalty is still low and she is exploring. She has a very strong health-consciousness and a high interest in beauty and health.
Persona(A)	She has a strong interest in anti-aging and maintaining cognitive abilities, prefers natural and additive-free products. She is also sensitive to new beauty trends and is exploring effective beauty foods.

B BI 方法的提示

以下是用于从 B 估计 A 时使用的提示示例。类型 1 提示 t1 输出被认为相同的 id 对，而类型 2 提示 t2 输出判断结果以及主观信心 s_{ij} 源数据。在这个例子中，匹配的 id 数量设置为 7。当发送提示时，插入了个别文本数据。

B.1 类型 1 提示：t1（能力评估）

你是一名擅长分析人们个性和心理的专家分析员。根据 # # id_B 的人员评估结果，推断个人档案，并将其与来自 # # 能力评估结果的 id_A 进行比较，后者似乎是同一个人。作为专家，请根据 # # 分析方法和 # # 执行方法选择最可能的匹配，并将结果输出为指定的 # # 输出格式。
分析方法
* 模拟人类思维过程并进行定性分析以得出结论。
* 直接解读数据并从上下文和表达中做出直观推断。
* 根据 # # 人员评估 id_B 的评论，详细分析个人的行为特征、专业能力和个人特性，并估计其概况。
* 将推断的概况与 # # id_A 的能力评估发现的评论进行比较，并选择最匹配的候选人 id_A。
# # 执行方法
* 描述选择最符合推测配置文件的候选 id_A 的过程。
* 一旦找到匹配的候选 id，就输出该 id。
* 根据指定的 # # 输出格式输出匹配的候选 id。
# # 输出格式
描述选择最符合推断轮廓的候选 id_A 的过程。id_B: { id_B 编号 } ，id_A: { 匹配候选 id_A 编号 }

能力评估结果
对于 id_A 的评估测试评论。数据如下：
{ id_A, 评估 (A) [重复 7 组样本数据] }
人员评估结果
id_B 的人员评估意见。数据如下：
{ id_B, 人员评估 (B) [重复目标 ID 数据] }
根据上述要求，请根据输出格式输出匹配的 ID。
您是一名优秀的分析员。请找到与从 id_B 的人员评价中推断出的同一个人的 id_A，并将其与候选人 id_As 的评估测试进行比较。然后，将这些候选人按可能性排序。此外，输出每个匹配的确定性水平。请参考以下 # # 指南，# # 详细要求，# # 确定性水平评估方法，# # 输出格式，及 # # 数据描述：
指南
* 模拟人的思维过程，通过定性分析得出结果。
* 直接读取数据内容，并从其上下文和表达中直观地推断。
* 基于 id_B 的人员评估，详细分析该人的行为特征、专业能力和个人特质，以推断出人物特征。
* 将推断出的人物特征与 id_A 的评估测试进行比较，以确定匹配的确定性水平。
# # 详细要求
* 描述推断出的角色。将推断出的角色与 id_A 的评估测试进行比较，以找到匹配的候选者。计算确定性水平（百分比）。
* 按最高确定性水平的顺序列出匹配的候选者 id_As。
* 最多输出 7 个匹配的候选者 id_As。
* 在每个匹配的 id_A 旁边显示确定性水平。
* 以指定的 # # 输出格式输出所有 id_B（共 7 个）的结果，不遗漏任何步骤。
# # 确定性水平评估方法
高确定性（例如，0.9 - 1.0）：两段文本之间有非常清晰的匹配。
中等确定性（例如，0.5 - 0.8）：存在一些共性，但不是完美匹配。
低确定性（例如，0.1 - 0.4）：不是很有信心，但可能匹配。
非常低确定性（例如，0.0）：两段文本之间几乎没有匹配点。
输出格式
**id_B: { id_B 编号 } ** { 推断出的角色描述。}
1. id_B: { id_B 编号 } , id_A: { 匹配候选人 id_A 编号 } { 确定性水平 }
2. id_B: { id_B 编号 } , id_A: { 匹配候选人 id_A 编号 } { 确定性水平 }
(省略)
数据描述
以下是用于分析的两个 CSV 数据集。
* ** 描述 id_A 的 CSV 数据:** 数据包括 id_A，评估测试。
数据如 { } 中所示。
{ id_A, 评估 (A) [重复 7 个样本数据] }
* ** 描述 id_B 的 CSV 数据:** 数据包括 id_B，人员评估。
数据如 { } 中所示。
{ id_B, 人员评估 (B) [重复 7 个样本数据] }
基于上述要求，请输出其匹配的候选 id_A 以及所有 id_B 的确定性水平，遵循 # # 输出格式，最多到第 7 个匹配候选。无需代码。

C 顺序思维提示

以 prof1e（英语/能力评估）为例，描述顺序提示的细节。提示由四部分组成：指定个人分析方法 (S1)，通过逐步方式逐个比较个人顺序处理 (S2)，当剩余数量达到指定值（默认 =2）时激活动态处理 (S3)，并通过 S4 执行重复检查和审查。

1. 系统提示 S1

您是一位专门从事身份匹配的优秀分析师。请逐步执行判断步骤。对于 B 文件中的每一个 ‘id_B‘，评估它对应于 A 文件中的哪个 ‘id_A‘。请按照下面描述的 { # 分析方法}，逐步做出判断，并尽可能达到最高精确度。
输入文件描述
‘id_A‘ 和 ‘id_B‘ 都是员工编号，但数字之间没有关联。
每个 ‘id_A‘ 对应唯一的 ‘id_B‘。
- **A 文件 **：包含员工编号 (‘id_A‘)、年龄、类型、优点、缺点和能力测试结果。
- **B 文件 **：包含员工编号 (‘id_B‘)、年龄、类型和人力资源评论。
分析方法
- ** 对每个 ‘id_B‘的分析 **：仔细分析从年龄、类型和 HR 评论中推断出的行为特征和个人特征，以创建个人的详细档案。

- ** 相似候选人的比较 **：识别与 ‘id_B‘年龄和类型相同的 ‘id_A‘候选人，并评估哪个 ‘id_A‘档案最接近 ‘id_B‘的特征。通过比较文件 A 中的信息（年龄、优点、缺点、能力测试结果）来证明结论。
请评估 ‘id_B= { id_b } ‘。请按照以下格式以 JSON 格式提供您的答案：
‘ { { ”thought”: str, ”id_A”: int } } ‘。在 ‘thought‘ 中，记录 ‘id_B‘ 和 ‘id_A‘ 之间的逐步比较和推理。对于 ‘id_A‘，输入与其最相似的个体的 ID 号。# # B 文件 { row_b } 条件分支：提示将根据 k 进行如下更改。如果计数器 k= n, ”””
匹配候选人为：‘id_A= { ids_a } ‘。# # A 文件 { rows_a } ”””
否则””” 评估的目标是 id_A= { ids_a } 。信息之前已提供。”””
3. 递归 S3
请评估 ‘id_B= { ids_b }‘。评估目标将是 ‘id_A= { ids_a }‘。
如果评估结果中有任何不一致或重复，可以从初始候选集 S0=id_A { { { orig_ids_a } } } 中重新选择。如果评估结果有重复，确定哪个更有可能。此外，查看先前的评估结果，识别需要在其与 ‘id_A‘ 的链接中进行更正的任何 ‘id_B‘，并在必要时进行调整。
请以书面格式提供您的回答。首先，描述逐步的比较和推理过程。接下来，记录对评估结果的任何修订。最后，输出确认的 ‘id_B‘ 和 ‘id_A‘ 对。对于 ID 对，包含所有以前的 ‘id_B= { old_ids_b }‘。# # B 文件 { rows_b }
4. 冲突解决 S4
使用 ‘id_A‘ 和 ‘id_B‘ 的输出对，通过以下步骤检查重复或未解决的 ID：
步骤
- 在 ‘<thinking>‘ 标签内，记录评估结果的审查过程。包括以下考虑：
- 确保没有重复的 ‘id_A‘ 被分配到多个 ‘id_B‘。
- 通过确定最合理的对来解决任何未解决的 ID。
- 审查总体结果并用较为合理的对替换任何对。
- 在 ‘<result>‘ 标签内，记录修订后的 ID 对。如果不需要修订，则记录原始对。
- 在 ‘<reflection>‘ 标签内，反思修订后的结果中是否有重复或未解决的 ID，并记录其状态。
- 在 ‘<count>‘ 标签内，指示需要多少额外修订。如果不需要额外修订，则将 ‘<count>‘ 的值设置为 0。

D 结果表

本附录列出了每个系统按数据集划分的结果。单个 BI 系统的表格项目包括系统编号、生成性 AI 模型的名称、用于生成观测矩阵 c_{ji} 的提示（类型和调用次数）、用于生成权重矩阵 s_{ij} 的提示（类型和调用次数）、正确答案的数量 n_c 、提升和覆盖率。t1 $*$ 上的星号表示需要描述判断过程的提示。t2 $'$ 上的撇号表示使用 groq:llama3-70b-8192 生成 s_{ij} 的实例。
对于 EBI 结果的表格，项目 “components” 列出了组成集成的单个 BI 系统编号集合，而 “weights” 则显示了它们各自的权重。

D.1 prof1j：日语能力评估

在这种情况下，(Lift, Reach) 的基线值为 $(H, G) = (19, 22)$ 。表 3 显示了 BI 系统的结果。BI 系统 37 和 42 取得了可与系统 76 (llama3.1-70b-versatile) 相当的结果，尽管模型规模更大。它们也超越了系统 64 和 66 (gpt-4o-mini-2024-07-18)。表 4 显示了 EBI（集成 BI 系统）的结果。所有这些都表明，通过将具有负 Lift 的 BI 系统 (66, 40, 12, 13) 结合起来，Lift 值变为正值。EBI 系统 50、81 和 83 的 Acc 超过了单一系统。

Table 3: 单个 BI 系统在 prof1_j（日本能力评估）中的结果。

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
37	gemma2-9b-it	t1 $*$ -100	t2 $'$ -10	23	21.1 %	104.5 %
42	llama3-8b-8192	t1 $*$ -100	t1 $*$ -100	23	21.1 %	104.5 %
76	llama3.1-70b-versatile	t1 $*$ -100	t2-10	23	21.1 %	104.5 %
64	gpt-4o-mini-2024-07-18	t2-10	t2-10	21	10.5 %	95.5 %
43	llama3-8b-8192	t1 $*$ -100	t2 $'$ -10	21	10.5 %	95.5 %
25	gemma2-9b-it	t1 $*$ -100	t2 $'$ -10	20	5.3 %	90.9 %
28	llama3-8b-8192	t1 $*$ -100	t1 $*$ -100	20	5.3 %	90.9 %
46	llama3-70b-8192	t1 $*$ -100	t2-10	20	5.3 %	90.9 %
66	gpt-4o-mini-2024-07-18	t1 $*$ -100	t2-10	18	-5.3 %	81.8 %
40	mixtral-8x7b-32768	t1 $*$ -100	t2 $'$ -10	18	-5.3 %	81.8 %
12	mixtral-8x7b-32768	t1-500	t1-500	18	-5.3 %	81.8 %
13	llama3-70b-8192	t1 $*$ -500	t1 $*$ -500	17	-10.5 %	77.3 %

Table 4: 针对 prof1_j*（日本能力评估）中 EBI（集成系统）的结果，仅限于表现最好的系统（提升 ≥ 0 ）。

system	components	weights	n_c	Lift	Reach
83,81	{ 37,40,43,46 }	[1,1,1,1],[1,1,2,3]	26	36.8 %	118.2 %
50	{ 12,13,25,28,37,40,43,46 }	[3,2,1,1,1,1,2,3]	25	31.6 %	113.6 %
55	{ 12,13,25,28,37,40,43,46 }	[3,2,1,1,5,1,2,3]	23	21.1 %	104.5 %
78	{ 37,43,45,66,76 }	[30,3,1,1,10]	23	21.1 %	104.5 %
71	{ 37,43,66 }	[1,1,1]	22	15.8 %	100.0 %
82,84	{ 37,40,43,46,66,76 }	[1,1,2,3,1,1],[1,1,1,1,1,1]	20	5.3 %	90.9 %
85	{ 37,40,42,46,64,59 }	[1,1,1,1,1,1]	19	0.0 %	86.4 %

D.2 prof2j: 日本购买数据

在这种情况下，（提升，覆盖）的基线值是 $(H, G) = (13, 20)$ 。表 5 显示了 BI 系统的结果，表 6 显示了 EBI（集成 BI 系统）的结果。它们主要由在 prof1j 中实现了正提升的 BI 系统（37, 46, 64, 76）组成。将具有负提升或零提升的系统（40, 42, 43）结合起来导致了正的提升值。

表现最好的 EBI 系统 82 包括规模为 70b 的系统 76，但它单独超越了系统 76 的提升值。由不含 70b 模型的系统组成的 EBI 系统 89 { 37, 40, 43 } ，在保持正提升的同时，实现了与系统 37 单独运行相当的准确性。这些结果表明，与其过度依赖像 70b 这样大规模的系统，不如将多个 SLM 一起使用可以提供多样性和稳健性。

Table 5: 单个 BI 系统在 prof2_j（日本购买数据）上的结果。

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
59	cerebras:llama3.1-70b-versatile	t2-10	t2-10	21	61.5 %	105 %
46	llama3-70b-8192	t1 $*$ -100	t2-10	17	30.8 %	85 %
64	gpt-4o-mini-2024-07-18	t2-10	t2-10	16	23.1 %	80 %
65	gpt-4o-mini-2024-07-18	t1 $*$ -100	t1 $*$ -100	16	23.1 %	80 %
76	cerebras:llama3.1-70b-versatile	t1 $*$ -100	t2-10	16	23.1 %	80 %
66	gpt-4o-mini-2024-07-18	t1 $*$ -100	t2-10	15	15.4 %	75 %
37	gemma2-9b-it	t1 $*$ -100	t2 $'$ -10	15	15.4 %	75 %
40	mixtral-8x7b-32768	t1 $*$ -100	t2 $'$ -10	13	0.0 %	65 %
45	llama3-70b-8192	t1 $*$ -100	t1 $*$ -100	12	-7.7 %	60 %
42	llama3-8b-8192	t1 $*$ -100	t1 $*$ -100	12	-7.7 %	60 %
43	llama3-8b-8192	t1 $*$ -100	t2 $'$ -10	11	-15.4 %	55 %

Table 6: 针对 prof2_j*（日本采购数据）的 EBI（集成系统）结果，仅限于表现最佳的系统（提升 ≥ 0 ）。

system	components	weights	n_c	Lift	Reach
82	{ 37,40,43,46,66,76 }	[1,1,2,3,1,1]	18	38.5 %	90 %
83	{ 37,40,43,46 }	[1,1,1,1]	17	30.8 %	85 %
84	{ 37,40,43,46,66,76 }	[1,1,1,1,1,1]	16	23.1 %	80 %
85	{ 37,40,42,46,64,59 }	[1,1,1,1,1,1]	16	23.1 %	80 %
81,91	{ 37,40,43,46 }	[1,1,2,3], [1,1,1,1]	15	15.4 %	75 %
89	{ 37,40,43 }	[1,1,1]	15	15.4 %	75 %
90	{ 37,40,43 }	[1,2,1]	14	7.7 %	70 %
92,93	{ 37,40,43,46 }	[1,2,1,2], [1,2,1,3]	14	7.7 %	70 %
86	{ 37,40,59 }	[1,1,1]	14	7.7 %	70 %
74	{ 37,64 }	[1,1]	13	0.0 %	65 %
77,78	{ 37,43,45,66,76 }	[1,1,1,1,1],[30,3,1,1,10]	12	-7.7 %	60 %

对于 prof1e，基线参考值为 $(H, G) = (21.2, 28)$ ，其中 $H = G * \gamma$ ，$\gamma = 0.757$ 。表 7 显示了 BI 系统的结果。系统 40 (mixtral-8x7b-32768) 取得的结果与系统 66 (gpt-4o-mini-2024-07-18) 相当，并超过了系统 76 (llama3.1-70b-versatile)。

表 8 显示了 EBI 的结果。系统 93 { 37, 40, 43, 76 } 超过了所有单一系统，尽管其对抗 GPT-4o 的到达率未达到 100 % ，但实现了相当高的 96 % 。有趣的是，添加 BI 系统 37, 43 和 76，它们单独的到达率较低，却提高了整体到达率。此外，EBI 系统 89 和 90 { 37, 40, 43 } 使用较弱的系统 37 和 43 (gemma2-9b-it, llama3-8b-8192) 保证多样性并作为鲁棒性的来源，同时实现了与单一系统最高到达率相当的到达率。

Table 7: Prof1_e（英语能力评估）单个 BI 系统的结果。

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
66	gpt-4o-mini-2024-07-18	t1 * -100	t2-10	25	17.9 %	89.3 %
40	mixtral-8x7b-32768	t1 * -100	t2 ′ -10	25	17.9 %	89.3 %
76	cerebras:llama3.1-70b-versatile	t1 * -100	t2-10	22	3.8 %	78.6 %
46	llama3-70b-8192	t1 * -100	t2-10	21	-0.9 %	75.0 %
37	gemma2-9b-it	t1 * -100	t2 ′ -10	20	-5.7 %	71.4 %
43	llama3-8b-8192	t1 * -100	t2 ′ -10	20	-5.7 %	71.4 %

Table 8: EBI（集成系统）在 prof1_e*（英语能力评估）中的结果。

system	components	weights	n_c	Lift	Reach
93	{ 37,40,43,76 }	[1,2,1,3]	27	27.4 %	96.4 %
89,90	{ 37,40,43 }	[1,1,1],[1,2,1]	25	17.9 %	89.3 %
91	{ 37,40,43,66 }	[1,1,1,1]	24	13.2 %	85.7 %
83	{ 37,40,43,46 }	[1,1,1,1]	24	13.2 %	85.7 %
92	{ 37,40,43,66 }	[1,2,1,2]	23	8.5 %	82.1 %
84	{ 37,40,43,46,66,76 }	[1,1,1,1,1,1]	22	3.8 %	78.6 %
81	{ 37,40,43,46 }	[1,1,2,3]	22	3.8 %	78.6 %
71	{ 37,43,66 }	[1,1,1]	21	-0.9 %	75.0 %
82	{ 37,40,43,46,66,76 }	[1,1,2,3,1,1]	19	-10.4 %	67.9 %

对于 prof2e，基准参考值是 $(H, G) = (17.4, 23)$ ，其中 $H = G * \gamma$ ，$\gamma = 0.757$ 。表 9 展示了 BI 系统的结果。系统 37（gemma2-9b-it）实现了 Reach > 100 %，超过了 GPT-4o。系统 66（gpt-4o-mini）同样超越了 GPT-4o。

表 10 显示了 EBI 的结果。合成上述所有六个单系统的系统 95 和 96 排名前两名。不包含系统 94（llama3.3-70b）的系统，如 92 和 93 也实现了达到 > 100 %，超过了 GPT-4o。这些系统包括弱系统 40 和 43，可以认为是鲁棒性的来源。

Table 9: 单一 BI 系统在 prof2_e（英语购买数据）上的结果。

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
94	cerebras:llama3.3-70b	t1 * -100	t2-10	30	72.4 %	130.4 %
46	llama3-70b-8192	t1 * -100	t2-10	29	66.7 %	126.1 %
37	gemma2-9b-it	t1 * -100	t2 ′ -10	25	43.7 %	108.7 %
66	gpt-4o-mini-2024-07-18	t1 * -100	t2-10	24	37.9 %	104.3 %
40	mixtral-8x7b-32768	t1 * -100	t2 ′ -10	17	-2.3 %	73.9 %
43	llama3-8b-8192	t1 * -100	t2 ′ -10	17	-2.3 %	73.9 %

Table 10: EBI（集成系统）对于 prof2_e*（英语购买数据）的结果。

system	components	weights	n_c	Lift	Reach
96	{ 37,40,43,46,66,94 }	[1,1,1,1,1,1]	31	78.2 %	134.8 %
95	{ 37,40,43,46,66,94 }	[1,1,2,3,1,1]	29	66.7 %	126.1 %
92,93	{ 37,40,43,66 }	[1,2,1,2],[1,2,1,3]	28	60.9 %	121.7 %
81,83	{ 37,40,43,46 }	[1,1,2,3],[1,1,1,1]	26	49.4 %	113.0 %
91	{ 37,40,43,66 }	[1,1,1,1]	26	49.4 %	113.0 %
71	{ 37,43,66 }	[1,1,1]	23	32.2 %	100.0 %
89,90	{ 37,40,43 }	[1,1,1],[1,2,1]	23	32.2 %	100.0 %

References

[1] A. Rodman, T. A. Buckley, A. K. Manrai, and D. J. Morgan, "Artificial Intelligence vs Clinician Performance in Estimating Probabilities of Diagnoses Before and After Testing", JAMA Network Open. 2023;6(12):e2347075

[2] P.-H. Chen, J.-J. Jung, Y. Kim, M. Lee et.al, "AI-assisted clinical summary and treatment planning for cancer care: A comparative study of human vs. AI-based approaches.(AI-Assisted Planning and Summarization)", Journal of Clinical Oncology Volume 42, 1523-1523(2024).

[3] A. Rodman, T. A. Buckley, A. K. Manrai, and D. J. Morgan, "Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge", JAMA Network Open. 2023;6(12):e2347075

[4] E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, et.al, "Large Language Model Influence on Diagnostic Reasoning", JAMA Network Open. 2024;7(10):e2440969

[5] Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330(1):78-80

[6] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, B. He, S. Jiang, and B. Dong. 2024. ChatGPT Is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3098–3110, Torino, Italia. ELRA and ICCL.

[7] C. Si, D. Yang, T. Hashimoto, "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers", International Conference on Learning Representations (ICLR) 2025. https://doi.org/10.48550/arXiv.2409.04109

[8] Z. Cui, N. Li, H. Zhou, "Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments with LLMs", Available at SSRN: https://dx.doi.org/10.2139/ssrn.4940173

[9] Q. Mei, Y. Xie, W. Yuan, and M. O. Jackson, "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans", Proceedings of the National Academy of Sciences, Vol. 121, No. 9, 2024

[10] Tartaraj, A., and Trebicka, B. (2023). "Optimal grading scales for enhancing product evaluation by infomediaries", Interdisciplinary Journal of Research and Development, 10(1), 59–66.

[11] Tholen, G. (2024). "Matching candidates to culture: How assessments of organisational fit shape the hiring process", Work, Employment and Society, 38(3), 705–722.

[12] X. Zhai, M. Nyaaba, W. Ma, "Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science?", Epistemic Insight & Artificial Intelligence(2024).

[13] T. Szandała, "ChatGPT vs human expertise in the context of IT recruitment", Expert Systems with Applications, Volume 264, 10 March 2025, 125868

[14] J. Qiu, D. Guo, P. Natalie, P. Noelle, L. Cheri, T. R. Henry, "Ensemble of Large Language Models for Curated Labeling and Rating of Free-text Data", arXiv:2501.08413 [cs.CL], https://doi.org/10.48550/arXiv.2501.08413

[15] C. Zhang, L. Liu, J. Wang, C. Wang, X. Sun, H. Wang, M. Ca, "PREFER: Prompt Ensemble Learning via Feedback-Reflect-Refine", Proceedings of the AAAI Conference on Artificial Intelligence, 38(17)

[16] Jiang, D., Ren, X., Lin, B.Y.(2023). "LLM-Blender: Ensembling Large Language Mod-els with Pairwise Ranking and Generative Fusion", In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

[17] Lu, K., Yuan, H., Lin, R., Lin, J., Yuan, Z., Zhou, C., Zhou, J.(2024). "Routing to the Expert: Efficient Reward-guided Ensemble of Large Language Models", In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1964–1974, Mexico City, Mexico. Association for Computational Linguistics.

[18] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. (2023), "Reflexion: Language Agents with Verbal Reinforcement Learning", Advances in Neural Information Processing Systems, 36.

[19] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston. 2024. "Chain-of-Verification Reduces Hallucination in Large Language Models", In Findings of the Association for Computational Linguistics: ACL 2024, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.