arXiv:2601.18732v1 [econ.TH] 26 Jan 2026

人工智能算法中偏好的最优运用

Joshua S. Gans*

2026年1月27日

摘要

机器学习系统通过训练损失函数或对校准预测进行后处理来嵌入偏好。本文运用Strack 和 Yang（2024）的信息设计方法，提出了与决策问题无关的条件，在这些条件下，分离——即进行无偏好训练并事后应用偏好——是最优的。不同于要求明确指定下游目标函数的先前研究，本文的福利结果适用于各种决策问题。核心基础条件是信息价值递减条件：相较于固定的（归一化）无偏好损失，嵌入偏好会降低信息量在边际上的价值，从而使学习得到的后验分布发生保均值收缩。由于信息价值关于信念是凸函数，对于任何期望效用决策问题，无偏好训练都弱占优。这为模块化人工智能管道提供了理论基础：学习校准概率，并通过下游决策规则处理非对称成本。然而，分离要求用户能够实施最优决策规则。当认知约束产生影响时——正如人机决策研究所记录的那样——嵌入偏好可能因自动完成阈值计算而占优。这些结果为设计提供了指导：当目标可能变化时，通过后处理保留选择灵活性；当决策阶段摩擦占主导时，则嵌入偏好。《经济文献杂志》分类编号：C45、C53、D81、D82、D83。

关键词：损失函数、适当评分规则、贝叶斯决策理论、信息设计、凸序、代价敏感学习、后处理。

*多伦多大学罗特曼管理学院及美国国家经济研究局。感谢 Philipp Strack 提供宝贵意见，感谢加拿大社会科学与人文研究理事会提供经费支持，并感谢 Refine.ink、ChatGPT 5.2 和 Claude Opus 4.5 提供宝贵的研究协助。所有错误均由我本人负责。

1 引言

预测很少是目的本身。预测是决策的输入：贷款与保险、招生与招聘、医疗分诊，以及稀缺公共资源的分配。这引出了一个几乎存在于所有应用流程中的根本架构选择，而在人工智能（AI）的部署中，这一选择尤为突出。构建预测算法时，训练目标应直接纳入最终用户的特定偏好（例如非对称成本或安全优先级），还是应训练算法学习潜在的真实情况（校准概率），将偏好的实施留待之后？这不仅仅是超参数调优的技术问题；它关系到究竟是构建一种绑定于单一静态目标的专用工具，还是打造一种通用能力，以便为各种不断变化的下游任务保留信息的期权价值。

经典的决策理论观点认为，偏好应体现在损失函数中。具体而言，Granger and Machina (2006年) 展示了决策问题如何引出评估损失：相较于依据完全信息采取行动，根据预测采取行动所造成的效用损失。当预测方法被视为固定时，这一逻辑颇具说服力：改变评估损失会改变预测的排序方式和行动的选择方式，但不会改变预测所包含的信息。[1]

为说明论文的主要发现，不妨将其置于一个简单的统计决策问题背景下。假设世界存在一个未知状态，Y，其取值可能为 0 或 1。有两个行动，$a \in \{0,1\}$；如果决策者选择的行动与状态相符，则收益为 0；但如果 $(a,Y) = (1,0)$，则需承担成本 c_{FP}；如果 $(a,Y) = (0,1)$，则需承担成本 c_{FN}。假设 $Y=1$ 的先验概率为 $\frac{1}{2}$。

有两种方法可以处理这一决策。每种方法都涉及利用现有数据生成一个具有后验概率 $q = \Pr[Y=1|\cdot]$ 的信号。首先，你可以生成一个将偏好嵌入信号本身的信号。例如，该信号可以采用类别加权交叉熵，$L^w(p,1) = -w_1 \log p$ 和 $L^w(p,0) = -w_0 \log(1-p)$，从而在信念 q 处生成贝叶斯最优报告：

$$s(q) = \frac{w_1 q}{w_1 q + w_0(1-q)}.$$

也就是说，该报告是对后验概率进行交叉加权的结果。在这种情况下，当且仅当 $s(q) \geq \frac{1}{2}$ 时，决策者才会选择 $a=1$，此时的期望收益为：

$$\mathbb{E}[u_{\mathrm{emb}}] = -\mathbb{E}\left[q\, c_{\mathrm{FN}} \mathbf{1}\{s(q) < \tfrac{1}{2}\} + (1-q)\, c_{\mathrm{FP}} \mathbf{1}\{s(q) \geq \tfrac{1}{2}\} \right].$$

[1] 这种方法源自Varian(1990年）的研究；Zellner (1986年）；关于人工智能和机器学习背景下的讨论，参见 Gans (2025年）。

第二种方法是让决策者在没有任何权重的情况下获取信号。随后，决策者对偏好进行后处理。代理人形成后验信念 q，并且当且仅当 $q \geq \tau$时才会选择最优方案 $a = 1$，其中

$$\tau = \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}.$$

由此得到的期望收益为：

$$\mathbb{E}[u_{\mathrm{post}}] = -\mathbb{E}\Big[\, q\, c_{\mathrm{FN}} \mathbf{1}\{q < \tfrac{c_{\mathrm{FP}}}{c_{\mathrm{FP}}+c_{\mathrm{FN}}}\} + (1-q) c_{\mathrm{FP}} \mathbf{1}\{q \geq \tfrac{c_{\mathrm{FP}}}{c_{\mathrm{FP}}+c_{\mathrm{FN}}}\}\Big].$$

为了将其与偏好嵌入进行比较，请注意，由于 $s(q) \geq \frac{1}{2} \iff q \geq \frac{w_0}{w_0+w_1}$,$\mathbb{E}[u_{\mathrm{emb}}]$可~~以写~~成

$$\mathbb{E}[u_{\mathrm{emb}}] = -\mathbb{E}\Big[\, q\, c_{\mathrm{FN}} \mathbf{1}\Big\{q < \tfrac{w_0}{w_0+w_1}\Big\} + (1-q)\, c_{\mathrm{FP}} \mathbf{1}\Big\{q \geq \tfrac{w_0}{w_0+w_1}\Big\}\Big].$$

请注意 $\mathbb{E}[u_{\mathrm{post}}] = \mathbb{E}[u_{\mathrm{emb}}]$ 如果遵循Granger and Machina (2006年) 的建议，w_0：$w_1 = c_{\mathrm{FP}}$： c_{FN}。

这种等价性凸显了偏好嵌入所面临的挑战。问题在于，在机器学习的实际应用中，损失函数不仅用于评估预测结果，还用于训练生成这些预测结果的算法。当学习是内生的，改变损失函数就会改变算法所学到的内容。这一差异是认知经济学视角的核心，该视角将学习视为内生信息获取（Caplin 等人，2022年）。最近，Autor 等人(2025年)（ACMM）通过指出一种特定的激励失灵，进一步深化了这一批评：他们表明，效用加权目标会削弱信念空间某些区域中学习的边际激励，导致预测器反而变得不那么有用，甚至对其训练所针对的目标也是如此。尽管他们在实证上将其描述为激励失灵，但从实际操作来看，即使按照偏好加权的评估标准，无偏好概率模型的训练后再事后调整决策，也可能胜过偏好嵌入训练。

例如，假设偏好嵌入是通过使用加权 分类损失（如类别加权对数损失）进行训练来实现的，然后对模型输出使用默认截断值（通常为 0.5），其理论依据是权重已经编码了偏好不对称性。与上述无摩擦示例的关键区别在于，现代机器学习训练几乎从不单独最小化加权损失：它最小化的是加权损失加上 复杂度方面的标准控制项，例如权重衰减、提前停止或其他正则化。这些控制项会惩罚极端的决策边界，因此会将预测概率收缩至中间区域。重要的是，正则化项并不以与数据拟合项相同的方式随成本权重

变化，因此，从“期望的成本比率”到“实际采用的决策规则”的映射会变得不稳定且依赖于模型：改变权重不仅会改变隐含的截断点，还会改变模型愿意表示哪些区分。

为了以最简单的方式说明这一后果，假设底层信号有时会产生较低的后验概率 q_L，有时会产生较高的后验概率 q_H，且 $q_L < \tau < q_H$。采用后处理或分离时，校准学习器能够保留这种差异，决策者也会恰好在后验概率较高的情形下采取行动，正如阈值规则所规定的那样。采用带正则化的加权训练时，学得的输出会系统性地向中间收缩（高值情形不再那么极端，低值情形也不再那么低）。极端情况下，学习器可能会将这两种情形压缩为几乎相同的评分，或者至少消除它们相对于具有经济意义的阈值之间的差距。一旦如此，任何事后阈值处理都无法恢复失去的期权价值：如果模型输出不再区分高-q 状态与低-q 状态，那么任何仅基于该输出的下游规则都必须以几乎相同的方式对待它们。分离可以避免这种失效模式，因为它以稳定对象（校准信念）为目标，并将偏好选择推迟到透明且易于调整的后处理步骤中。

本文将这些洞见推广为稳健分离原理，将核心架构问题界定为学习与选择之间的抉择：算法应当学习关于世界的经校准信念，还是应当直接学习选择哪些行动？虽然 ACMM 展示了在目标固定时嵌入偏好的风险，但本文确立了在不确定目标下，无偏好训练（学习）具有全局优势的条件。这一点具有重要的实践意义，因为下游目标函数通常是不确定的、存在争议的或不断变化的：监管约束会不断演变，安全政策会得到修订，组织优先事项会发生转移，而异质用户的需求也各不相同。[2]

该方法将训练建模为对信息结构的选择。沿袭贝叶斯劝说与信息设计（Kamenica andGentzkow，2011年），并借鉴Strack 和 Yang （2024年）的凸序工具，训练后的预测器可概括为其所诱导的后验信念分布。关键的基本条件是信息价值递减条件：嵌入偏好会降低提高信息量所带来的边际收益。结合一维可比性条件，可得出明确的比较静态结论：偏好嵌入会使学习得到的后验分布发生保均值收缩（定理 1）；这是一种“能力税”，会摧毁事后无法恢复的信息。

[2]这一主题出现在Agrawal等人（2018年、2022年）的研究中，他们将决策过程中的偏好部分——判断（来自决策者）——与人工智能提供的预测部分区分开来。

福利后果源于期望效用决策问题关于信念的凸性：信息量更大的后验分布弱优，因此收缩结果意味着分离原理（定理 2）。采用无偏好严格适当评分规则进行训练，并仅在事后实施偏好，在所有期望效用决策问题中均弱优于嵌入偏好的训练。这为模块化人工智能设计提供了不依赖具体决策问题的基础：将知识获取与判断应用分离，可保留预测器的期权价值。

除了主要的分离结果之外，本文还阐明了对实践具有重要意义的另外三个方面。首先，“无偏好”并不是对训练目标的完整规定。相关类别是（严格）适当评分规则，它们能够引出校准概率（Gneiting 和 Raftery，2007年）。在二元情境下，有许多严格适当损失可供选择（局部性并不能确定唯一的损失函数），因此，无偏好损失之间的选择可能取决于贝叶斯风险曲率如何与学习摩擦相互作用。第4 节构建了无偏好损失的一个易处理排序，并解释了如何在设计类别内理解“最优性”。

其次，可以考虑一些能够减弱对分离偏好的因素。基准结果假设决策者能够在不产生成本的情况下，根据学习得到的概率实施最优事后决策规则。第5节引入了一种以理性注意力文献为基础的决策阶段摩擦：使用丰富的概率信号本身可能带来认知成本，而且这些成本会随着所处理的信息量增加。在这种情况下，通过在上游嵌入偏好来压缩信息并减少决策阶段的认知负担，可能会改善福利。定理3 刻画了这种“逆转”发生的条件。

最后，第6 节将同一视角应用于一个当代场景，其中“偏好嵌入”是字面意义上的：大语言模型（LLM）部署中的基于人类反馈的强化学习（RLHF）。这是让大语言模型输出变得诚实、有帮助且无害的过程。[3] RLHF 可以被理解为对底层后验质量分布进行指数倾斜。当目标稳定且一致时，RLHF 会通过提升质量水平来增进福利（定理4）。但当部署目标不确定或具有多个维度时，让生成器集中于一种固定权重，会降低其在不同目标之间的灵活性。这会给基础模型带来能力税（定理5），并且当学习得到的奖励与真实质量不一致时，可能加剧奖励设定错误（命题6）。从这个意义上说，当目标可能

[3]例如，OpenAI 使用数千名人工审核员提供反馈功能；参见Stiennon 等人(2020年)；Ouyang 等人(2022年)。

转移。

本文其余部分安排如下。第 2 节介绍模型环境和后验分布表示。第 2.1 节回顾适当评分规则，并形式化 ACMM 所强调的报告／后处理差距。第2.2 节将训练建模为伴有学习摩擦的信息获取，并提出价值递减条件。第 3节阐述收缩结果和稳健分离原理。第 4 节讨论在共同设计限制下如何选择无偏好损失函数，并将理论转化为实施指南、算例和诊断方法。第 5 节研究认知约束何时会使偏好嵌入成为最优选择。第 6节通过同一嵌入与后处理视角分析基于人类反馈的强化学习。第 7 节作总结。

2 模型设定与初步结果

核心分析聚焦于二元结果，以使凸序比较清晰且易于解读。许多人工智能预测任务都可以表述为这种标量信念形式（例如风险评分）。（超出标量信念范围的扩展需要对信息序作出额外结构假设，并将在第6节和附录A中讨论。）

令 $Y \in \{0,1\}$ 表示我们关注的结果。令 $\mu \in (0,1)$ 表示先验概率$\mathbb{P}(Y=1) = \mu$。关于 Y 的后验信念可用标量 $q \in [0,1]$表示，其含义为 $q = \mathbb{P}(Y=1 \mid \cdot)$。决策问题由一组可行行动 $\mathcal{A}$ 和收益函数$u: \mathcal{A} \times \{0,1\} \to \mathbb{R}$构成。观察到信息后，决策者选择一项行动以最大化期望收益。给定信念 q，决策者的间接价值为

$$V(q) := \sup_{a \in \mathcal{A}} \{q\, u(a,1) + (1-q)\, u(a,0)\}. \tag{1}$$

引理1 (间接价值的凸性)。 对于任意决策问题 $(\mathcal{A}, u)$，函数$V: [0,1] \to \mathbb{R}$ 是凸的。

证明。 对于每个固定的行动 a，映射 $q \mapsto q\, u(a,1) + (1-q)\, u(a,0)$ 是仿射函数。仿射函数的逐点上确界是凸的。 □

引理1 是福利比较背后的关键工具。它形式化地阐明了一个简单观点：当决策者能够在观察到信念后优化行动时，改善信息所带来的价值是信念的凸函数。引理1 具有期权价值解释。信息之所以有价值，是因为它扩大了信念实现后可采取的应变行动集合。凸性 V 意味着，后验信念中的均值保持扩散会提高预期

间接效用：信念的离散程度是有价值的灵活性来源，即使尚未指定任何具体应用也是如此。

将训练后的预测器建模为关于 Y的信号。令 S 表示预测器产生的信号。由 S 诱导的后验信念分布是随机变量的分布

$$Q := \mathbb{P}(Y = 1 \mid S). \tag{2}$$

贝叶斯可行性意味着 $\mathbb{E}[Q] = \mu$。

定义1（贝叶斯可行后验分布）。令 $\mathcal{Q}(\mu)$ 表示取值于 $[0,1]$ 且满足 $\mathbb{E}[Q] = \mu$的随机变量 Q 的集合。

这种后验分布表示在贝叶斯劝说和信息设计中是标准做法（Kamenica and Gentzkow, 2011年），并在Strack 和 Yang（2024年）中得到明确使用。这也与 ACMM 所强调的做法一致，即将所学内容（诱导的后验分布）与所报告内容（预测器的输出）区分开来。

二元情形中的相关信息序是具有固定均值的 $[0,1]$ 随机变量上的凸序。

定义2（凸序）。对于 $Q,\ Q' \in \mathcal{Q}(\mu)$，若满足以下条件，则记为 $Q \succeq_{\text{cx}} Q'$

$$\mathbb{E}[\varphi(Q)] \geq \mathbb{E}[\varphi(Q')] \quad \text{for all convex } \varphi : [0,1] \to \mathbb{R}. \tag{3}$$

等价地，Q' 是 Q的保均值收缩。

其含义是：Q 比 Q' 更加分散，因此信息量更大，同时保持相同的均值。这与 Strack 和 Yang (2024年)中保均值收缩结果所依据的序关系相同。

结合定义2 和引理1，可得到一个标准推论：

引理2（信息量更大的后验提高期望价值）。若 $Q \succeq_{\text{cx}} Q'$，则对于任意决策问题$(\mathcal{A},\ u)$，

$$\mathbb{E}[V(Q)] \geq \mathbb{E}[V(Q')]. \tag{4}$$

证明。由引理1，V是凸函数，因此定义2 意味着 $\mathbb{E}[V(Q)] \geq \mathbb{E}[V(Q')]$。 □

引理 2 是针对二元信念的 Blackwell 型结果。它将使本文能够把后验分布的收缩结果转化为一大类决策问题的福利比较。

2.1 损失函数与后处理

我们现在将严格阐明训练损失在何种意义上是“无偏好”的，并解释 ACMM 风格的报告与学习分解。

概率预测是报告的概率 $p \in [0,1]$，其含义是预测器声称 $Y = 1$ 的概率。训练损失是一个函数 $L : [0,1] \times \{0,1\} \to \mathbb{R}$。给定信念 q，报告 p 所产生的期望损失为

$$\bar{L}(p;q) := q\,L(p,1) + (1-q)\,L(p,0). \tag{5}$$

定义 3 （严格适当评分规则）。如果对于每个 $q \in [0,1]$， $L(p;q)$ 的唯一极小化者是 $p = q$，那么损失函数 L就是严格适当评分规则。

严格适当性体现了损失函数在学习概率时“无偏好”的形式化含义：当预测器的内部信念为 q时，最小化损失函数的方式就是真实地报告 q。[4]

与损失函数相关的一个关键对象是其贝叶斯风险。

定义 n 4 （贝叶斯风险）。给定损失函数 L，定义贝叶斯风险 $H_L : [0,1] \to \mathbb{R}$ 如下

$$H_L(q) := \min_{p \in [0,1]} \bar{L}(p;q). \tag{6}$$

对于严格适当评分规则，贝叶斯风险在 $p = q$ 处达到，因此 $H_L(q) = L(q;q)$。在标准正则条件下，适当评分规则与凹贝叶斯风险函数一一对应（Gneiting 和 Raftery, 2007年）。这种凹性正是适当损失奖励后验变化所携带信息的途径。

嵌入偏好的常见方法是修改训练损失，对某些错误施加比其他错误更重的惩罚。对于二元分类，一个典型例子是使用权重 w_1， $w_0 > 0$ 的类别加权交叉熵：

$$L^w(p,1) = -w_1 \log(p), \qquad L^w(p,0) = -w_0 \log(1-p). \tag{7}$$

除非 $w_1 = w_0$，否则该损失函数相对于真实后验概率 q 并非严格适当，因为贝叶斯最优报告是对 q 的一种变换：

$$p^w(q) = \frac{w_1 q}{w_1 q + w_0(1-q)}. \tag{8}$$

[4]参见Gneiting 和 Raftery(2007年)，其中有全面的讨论。

式（8）中的变换正是 ACMM 所强调的那类后处理映射：在给定信念 q的条件下，偏好嵌入会改变报告内容，但报告对象所包含的信息不会多于 q；若无法进行逆变换，或报告经过压缩，则信息可能严格少于 q。[5]

命题1（类别加权交叉熵的贝叶斯风险与曲率）。 对于 （7）中的类别加权交叉熵损失函数，贝叶斯风险为

$$H_w(q) = -w_1 q \log\big(w_1 q\big) - w_0(1-q)\log\big(w_0(1-q)\big) + \big(w_1 q + w_0(1-q)\big)\log\big(w_1 q + w_0(1-q)\big), \quad (9)$$

（通常约定 $0 \log 0 := 0$）。它在 （0，1）上是凹的，其中

$$H_w''(q) = -\frac{w_0 w_1}{q(1-q)\big(w_0 + q(w_1 - w_0)\big)}. \quad (10)$$

如果 $H_{\log}$ 表示未加权对数损失（$w_0 = w_1 = 1$）的贝叶斯风险，那么

$$\big(H_w - H_{\log}\big)''(q) = \frac{w_0 + q(w_1 - w_0) - w_0 w_1}{q(1-q)\big(w_0 + q(w_1 - w_0)\big)}. \quad (11)$$

特别地，$\big(H_w - H_{\log}\big)$ 在 （0，1）上是凸的，当且仅当 $\max\{w_0, w_1\} \le 1$。因此，对于常见的成本敏感归一化，其中一个 （w_0，w_1）超过1，假设*3* 中的递减值条件必须经过检查，而不能想当然地认为其成立。

证明。根据（8），贝叶斯最优报告为 $p^w(q)$，因此 $H_w(q) = q\,L^w(p^w(q), 1) + (1-q)\,L^w(p^w(q), 0)$，化简后得到（9）。对其求两次导数可得（10），进而得到（11）。由于（11）中的分子关于 q 是仿射的，凸性条件归结为在 $q = 0$ 和 $q = 1$处对其进行检查，从而得到 $\max\{w_0, w_1\} \le 1$。 □

这里的归一化很重要。类别权重 (w_0, w_1) 只能确定到一个共同的比例因子：将两个权重同时乘以一个常数，只会重新缩放损失函数，不会改变式（8）中的贝叶斯最优报告。然而，在学习模型（12）中，这种缩放会改变学习激励相对于摩擦 $C(\cdot)$ 的关系。因此，像命题 1中的曲率比较，应理解为在损失函数整体尺度固定的情况下进行。在采用其他成本敏感归一化且 $\max\{w_0, w_1\} > 1$ 时，假设

[5]在多类别情形下，ACMM 给出了一般性刻画：对于严格适当的基础损失，最优的偏好加权预测对应于一种基于期望效用分量对后验概率进行的确定性变换。本文的核心观点在二分类和多类别情形下相同：给定后验概率后，将偏好嵌入预测中属于后处理步骤。

3中的价值递减条件可能不成立；原则上，偏好嵌入可能增加信息量，而不是使其收缩。

偏好嵌入式管线与无偏好管线之间的任何福利差异，都必须源于训练所诱导的后验分布存在差异，而非报告转换的存在。这正是下一节信息获取模型的提出动机。

2.2 训练与学习摩擦

本文将训练建模为信息量的内生选择，这与Caplin 等人(2022年提出的基于成本的方法一致。该模型有意采用简化形式。本文不假设某个特定优化算法能够求解某个特定的随机控制问题。相反，本文旨在捕捉许多实际训练管线的一项稳健含义：改变损失函数会改变投入建模能力、优化努力和表征能力，以学习具有信息量的区分的激励。

假设 1（基于成本的学习作为后验选择）。 对于每个训练损失 L，训练算法都会诱导一个后验随机变量 $Q_L \in \mathcal{Q}(\mu)$，使其满足

$$Q_L \in \arg\min_{Q \in \mathcal{Q}(\mu)} \{\mathbb{E}[H_L(Q)] + C(Q)\}, \tag{12}$$

其中 $C\colon \mathcal{Q}(\mu) \to \mathbb{R} \cup \{+\infty\}$ 是一个学习摩擦函数。

这种解释很直观。信息量更丰富的后验分布更有利，因为它们能降低预期贝叶斯风险，而损失函数 L则决定了这些降低有多大价值。函数 $C(Q$刻画了优化摩擦、正则化、有限容量，或任何其他会增加获取信息更丰富的预测器成本的因素。这与Caplin 等人(2022年)中的伪成本解释一致。

存在性。在继续之前，我们需要一个基本的适定性结果：对于每个训练目标索引 t，式（12）中的约化形式“学习问题”确实存在解。命题2 给出了保证最优（贝叶斯可行）后验分布存在的温和条件，因此后续的比较静态论证并非针对空集。

命题 2 (最优后验概率的存在性**)**。 固定 $t \in [0,1]$。假设贝叶斯风险函数 H_t 在 $[0,1]$上连续。假设学习摩擦 C 是分布不变 的（仅取决于 Q的分布），在 $\mathcal{Q}(\mu)$上有下界，并且在 $[0,1]$ 上诱导分布的弱收敛下是下半连续的。假设至少存在

一个 $Q \in \mathcal{Q}(\mu)$ 满足 $C(Q) < \infty$。那么，(12) 中的极小化问题至少存在一个极小化者。

证明。 将每个 $Q \in \mathcal{Q}(\mu)$ 与其在 $[0,1]$, 上的分布 P_Q 对应，并令 $\mathcal{P}(\mu)$ 表示在 $[0,1]$ 上均值为 μ 的概率测度集合。由于 $[0,1]$ 是紧集，且均值约束定义了一个闭集，因此 $\mathcal{P}(\mu)$ 在弱收敛下是紧的。

在 $\mathcal{P}(\mu)$ 上定义目标函数如下：

$$F_t(P) := \int_0^1 H_t(q)\, dP(q) + \widetilde{C}(P),$$

其中，对于任意分布为 P 的 Q ， $C(P)$ 都是 $C(Q)$ 的相同取值（由分布不变性可知，该值定义良好）。 H_t 的连续性意味着 $P \mapsto \int H_t\, dP$ 连续，而根据假设， C 是下半连续的。因此， F_t 在紧集上是下半连续的，故能取到最小值。任一极小化者 $P^\star$ 都可由某个 $Q^\star \in \mathcal{Q}(\mu)$ 实现，从而得证。 □

命题 2表明，将后验概率 Q 与其在 $[0,1]$ 上诱导的分布对应，并施加贝叶斯可行性约束 $\mathbb{E}[Q] = \mu$后，设计者的选择集是紧的；同时，该命题还表明，目标函数在弱收敛下是下半连续的，因此必定存在极小化者。证明存在性之后，我们可以将训练结果视为后验分布上的分布（这是一种信息设计的做法），因为优化可以直接在 $\mathcal{P}(\mu)$ 上进行，而不必担心下确界只能无限逼近却永远无法取到。这样，我们就可以专注于 形状对 H_t 以及 次序性质对 C 的影响，以及它们如何影响最优后验分布。

学习摩擦。 下一个假设表明，生成信息量更大的后验分布，所需成本至少不会更低。

假设2 （凸序单调学习摩擦）。 对于 Q, $Q' \in \mathcal{Q}(\mu)$，若 $Q \succeq_{\text{cx}} Q'$，则 $C(Q) \geq C(Q')$。

假设2 是学习技术领域中与这样一种观点相对应的类比：信号退化不会增加约束或成本；参见图1。它体现了 Strack 和 Yang 的观点：在约束下可行对象构成一个在保均值收缩下封闭的集合（Strack 和 Yang，2024年）。此处，限制并非像他们的硬性可行性约束那样设定，而是通过成本来体现。

许多合理的学习摩擦都满足假设 2。[6] 一个特别简洁的微观基础是信息论容量：生成信息量更大的信号需要付出成本

[6] $C(Q)$刻画了现实世界机器学习模型中可能存在的学习摩擦。例如，正则化技术（如 L_1/L_2 权重衰减或随机失活）会对模型复杂度施加惩罚，实际上使高度分散的后验分布变得“更难”学习。这种表述也与

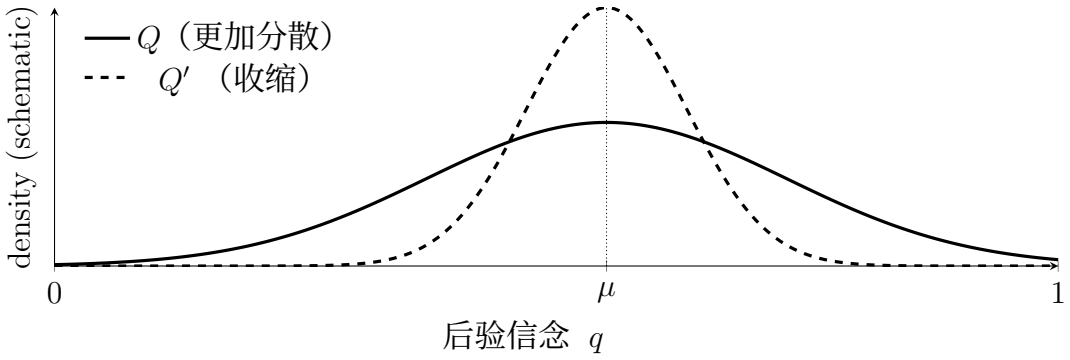


图 1： $Q \succeq_{\mathrm{cx}} Q'$示意图： Q' 是 Q 的保均值收缩（示意图）。

因为这需要更多用于表征、优化投入或数据的“比特”。令 S 表示训练算法生成的信号（表征），令 $Q = \mathbb{P}(Y = 1 \mid S)$ 表示由此产生的后验信念。考虑互信息学习摩擦

$$C_{\mathrm{MI}}(Q) \;:=\; \kappa\, I(Y;S), \qquad \kappa > 0, \tag{13}$$

与理性疏忽模型中一样（Sims，2003年）。在二元情形下，这一成本可以完全写成后验随机变量 Q的函数：

$$I(Y;S) = h_{\mathrm{bin}}(\mu) \;-\; \mathbb{E}[h_{\mathrm{bin}}(Q)] = \mathbb{E}[\mathrm{KL}(\mathrm{Bern}(Q) \,\|\, \mathrm{Bern}(\mu))]\,, \tag{14}$$

其中 $h_{\mathrm{bin}}(q) := -q\log\, q - (1-q)\log(1-q)$ 是二元熵。这里，$\mathrm{Bern}(q)$ 表示定义在 $\{0,1\}$ 上的伯努利分布，其成功概率为 q（即 $\mathbb{P}(1) = q$ 和 $\mathbb{P}(0) = 1-q$），因此 $\mathrm{Bern}(Q)$ 表示成功概率等于后验概率 Q 实现值的伯努利分布（等价地，$Y \mid S \sim \mathrm{Bern}(Q)$）。最后，$\mathrm{KL}(P\|R)$ 表示从分布 P 到 R的 Kullback–Leibler 散度（相对熵）：

$$\mathrm{KL}(P\|R) \;:=\; \sum_{y\in\{0,1\}} P(y)\log\!\left(\frac{P(y)}{R(y)}\right).$$

信息瓶颈原理，其中训练目标会明确惩罚关于输入所保留的互信息。除了显式惩罚之外，$C(Q)$ 还可以表示隐式资源约束——例如训练数据有限或提前停止预算有限——这些约束会阻止模型达到信息充分（但可能过拟合）的配置。尽管分布不变性这一假设忽略了架构的归纳偏置（例如，卷积神经网络在结构上偏好空间特征），但该函数稳健地捕捉了预测准确性与表征或优化成本之间的一般权衡。

特别地，对于伯努利分布，

$$\mathrm{KL}(\mathrm{Bern}(q)\,\|\,\mathrm{Bern}(\mu)) = q\log\left(\frac{q}{\mu}\right) + (1-q)\log\left(\frac{1-q}{1-\mu}\right).$$

由于 h_{bin} 是凹函数， $Q \succeq_{\mathrm{cx}} Q'$ 意味着 $\mathbb{E}[h_{\mathrm{bin}}(Q)] \le \mathbb{E}[h_{\mathrm{bin}}(Q')]$, 因此 $C_{\mathrm{MI}}(Q) \ge C_{\mathrm{MI}}(Q')$。因此，(13) 满足假设 2。[7]

关键条件。 关键步骤是将“偏好嵌入会减少信息”这一想法转化为一个原始条件。模型允许偏好嵌入具有连续指标，从而可以运用单调比较静态论证。

令 $t \in [0,1]$ 为一族训练目标的指标。 $t = 0$ 对应无偏好训练。 $t = 1$ 对应偏好嵌入变体，例如类别加权交叉熵或效用加权评分。形式上，指标 t通过贝叶斯风险进入。令 H_t 表示与按 t索引的训练目标相关联的贝叶斯风险函数。

假设3 （信息价值递减）。 该族 $\{H_t\}_{t\in[0,1]}$ 关于凸序满足递增差异：对于任意 $t_1 > t_0$ 和任意 Q， $Q' \in \mathcal{Q}(\mu)$ 满足 $Q \succeq_{\mathrm{cx}} Q'$，

$$\mathbb{E}[H_{t_1}(Q) - H_{t_0}(Q)] \ge \mathbb{E}[H_{t_1}(Q') - H_{t_0}(Q')]\,. \tag{15}$$

此外，不等式（15）在严格成立当且仅当 $Q \succeq_{\mathrm{cx}} Q'$ 且 $Q \overset{d}{\neq} Q'$时。

假设3 可以理解为，就贝叶斯风险降低而言，信息性的边际价值递减。对于每个 t，将信息结构 Q在 t-索引训练目标下的（事前）价值定义为：相对于接收无信息时贝叶斯风险的降低：

$$\mathrm{VoI}_t(Q) := H_t(\mu) - \mathbb{E}[H_t(Q)]. \tag{16}$$

由于 H_t 关于 q是凹的，$\mathrm{VoI}_t(Q) \ge 0$，并且随着信息量按凸序增加而增加。假设3 表明，当 t 增加时（偏好嵌入得更深），对于信息量更高的后验分布，该值会下降得更多：对于任意 $t_1 > t_0$ 和 $Q \succeq_{\mathrm{cx}} Q'$，

$$\big(\mathrm{VoI}_{t_1}(Q) - \mathrm{VoI}_{t_0}(Q)\big) \le \big(\mathrm{VoI}_{t_1}(Q') - \mathrm{VoI}_{t_0}(Q')\big). \tag{17}$$

[7]同样的“率失真”逻辑也构成了信息论正则化以及机器学习中的瓶颈式目标的基础。对 C（·）的相关解释包括样本规模／数据采集成本、容量和瓶颈约束（如Strack 和 Yang（2024年）所述）、正则化／最小描述长度惩罚、PAC–贝叶斯复杂度界，以及优化／早停预算。

等价地，对于每个 $t_1 > t_0$，增量 $\Delta_{t_1,\ t_0}(q) := H_{t_1}(q) - H_{t_0}(q)$ 是 q 的凸函数（因此 H_{t_1} 比 H_{t_0}“凹性更弱”）。从这种曲率意义上说，偏好嵌入会削弱学习激励。

在二元情形下，假设 3简化为一维形状限制。固定$t_1 > t_0$ 并定义增量 $\Delta_{t_1,\ t_0}(q) := H_{t_1}(q) - H_{t_0}(q)$。当且仅当 $\Delta_{t_1,\ t_0}$ 在 $[0,1]$ 上为凸函数（严格性条件要求其严格凸）时，假设 3 成立。当 H_t 二次可微时，这等价于检验对所有 q，$\Delta''_{t_1,\ t_0}(q) \geq 0$。对于许多常用的训练目标，$H_t$ 有闭式表达；否则，可以在密集网格上进行数值凸性检验。

这一假设是 ACMM 中激励趋平直觉的精确原始条件对应。它也是学习领域与 Strack 和 Yang (2024年) 的结果相对应的类比：隐私保护信号对应于基准后验均值分布的保均值收缩。

假设 3 的一个充分条件尤其直观。

命题3（一个方便的充分条件）。 如果 $H_t(q) = H_0(q) + t\,h(q)$ 对于某个凸函数 $h : [0,1] \to \mathbb{R}$，则假设 *3* 成立。此外，如果 h 是严格凸的，则假设 *3* 中的严格性条件也成立。

证明。对于 $t_1 > t_0$, $\mathbb{E}[H_{t_1}(Q) - H_{t_0}(Q)] = (t_1 - t_0)\mathbb{E}[h(Q)]$。如果 $Q \succeq_{\text{cx}} Q'$，那么 $\mathbb{E}[h(Q)] \geq \mathbb{E}[h(Q')]$，因为 h 是凸的。这意味着（15）。如果 h 是严格凸的，且 $Q \succeq_{\text{cx}} Q'$ 满足$Q \overset{d}{\neq} Q'$，那么该不等式是严格的。 □

命题 3表明，与其假设偏好嵌入会使学习得到的后验分布发生收缩，该命题不如假设偏好嵌入如何改变后验分布空间中的训练目标受到某种形状限制。收缩是比较静态的推论。

将我们的结果付诸实践的最后一个假设如下：

假设4（最优后验的可比性）。 对于任意 $t_1 > t_0$，以及任意最小化者 $Q_{t_1} \in \arg\min_{Q \in \mathcal{Q}(\mu)} \{\mathbb{E}[H_{t_1}(Q)] + C(Q)\}$ 和 $Q_{t_0} \in \arg\min_{Q \in \mathcal{Q}(\mu)} \{\mathbb{E}[H_{t_0}(Q)] + C(Q)\}$，二元组（$Q_{t_0}$，$Q_{t_1}$）在凸序下具有可比性：$Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ 或 $Q_{t_1} \succeq_{\text{cx}} Q_{t_0}$。

假设4 是一项实质性限制。凸序只是 $\mathcal{Q}(\mu)$ 上的一种偏序，因此一般而言，两个最优后验分布不一定可比较。该假设通过要求最优解沿着一条凸序链（即一维的“信息性前沿”）移动，排除了这些选择问题。该假设仅用于将

假设 3所蕴含的不等式转化为关于凸序的单调比较静态分析结论。[8]

假设 4 将假设 3 中的递增差异不等式转化为关于 所有 极小化解的全局陈述。没有这一假设，定理 1 背后的论证仍然可以得到较弱的方向性约束：如果在 t_1 处的极小化者与在 $t_0 < t_1$处的极小化者在凸序下具有可比性，则唯一可能的排序是 $Q_{t_0} \succeq_{\mathrm{cx}} Q_{t_1}$ （较高的t 极小化者不可能是较低的t 极小化者的严格均值保持扩散）。当学习摩擦确实具有多维性时，可能会出现不可比性，因此，与其将该假设视为任意 C（·）学习技术的一般性推论，不如将其理解为对学习技术的单指标约束。唯一最小化者以及嵌套的（按Blackwell序排列的）信号族，是可比性自动得到满足的两种常见情形。

可比性的一个充分条件是，学习技术允许存在一个按信号退化（Blackwell序）排序的嵌套可获得信号族。在这种情况下，由此产生的后验均值按凸序排列：

引理3 (信号退化蕴含凸序收缩)。 设 S_0 是关于 Y 的信号，且S_1 是对 S_0 的信号退化（即 S_1 由 S_0 通过一个独立于Y的随机映射生成）。设 $Q_i := \mathbb{P}(Y = 1 \mid S_i)$。那么 $Q_0 \succeq_{\mathrm{cx}} Q_1$。

证明。 由于 S_1 由 S_0 通过一个独立于 Y的随机映射生成，因此有马尔可夫链 $Y \to S_0 \to S_1$ （等价地， $Y \perp S_1 \mid S_0$），所以 $\mathbb{E}[Y \mid S_0, S_1] = \mathbb{E}[Y \mid S_0]$。因此，根据塔式性质，

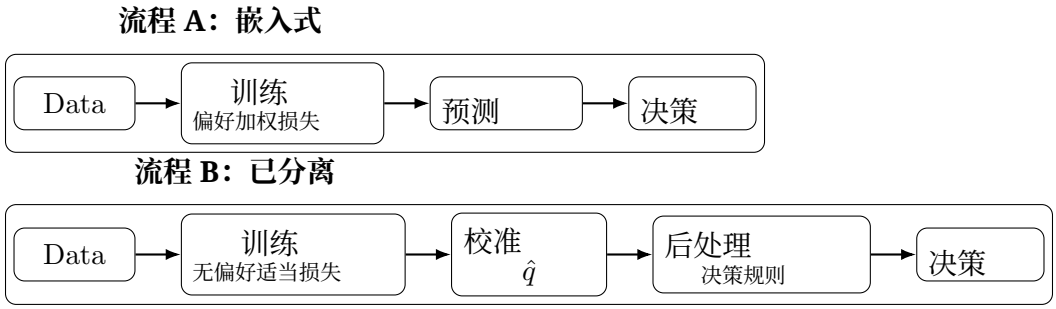
$$Q_1 = \mathbb{E}[Y \mid S_1] = \mathbb{E}[\mathbb{E}[Y \mid S_0] \mid S_1] = \mathbb{E}[Q_0 \mid S_1].$$

对于任意凸函数 φ，由条件詹森不等式可得 $\varphi(Q_1) = \varphi(\mathbb{E}[Q_0 \mid S_1]) \le \mathbb{E}[\varphi(Q_0) \mid S_1]$。取期望可得 $\mathbb{E}[\varphi(Q_1)] \le \mathbb{E}[\varphi(Q_0)]$, , 这正是 $Q_0 \succeq_{\mathrm{cx}} Q_1$ 根据定义2。 □

3 主要结果：收缩与分离

目标是探究在可以进行后处理时，嵌入偏好是否能够达到最优；参见图 2。下文将证明，在以下条件下无法做到这一点

[8]假设4 最好被视为对学习技术的一种单指标限制：最优解沿着一维的“信息性前沿”移动。在许多机器学习流程中，这是自然的情形，因为可学习性由少数几个单调调节参数决定（样本量、训练时长、模型规模、正则化强度、量化水平）。当单个调节参数收紧或放宽约束时，可行预测器会通过信号退化或细化形成嵌套关系，进而使后验信念构成凸序链。该假设仅用于根据假设1中的递增差异不等式，得到定理3中的凸序收缩。一旦以任何方式确立后验分布之间的凸序收缩，引理2 中的福利步骤及定理2中的福利比较便不再依赖可比性；我们在定理4 中保留假设2 ，仅仅是因为其证明引用了定理1。



根据定理2，分离弱支配（在假设 1、3、4成立的条件下）。

图 2：在预测流程中应用偏好的两种位置：将偏好嵌入训练中（A），或在学习出校准概率后再应用偏好（B）。

上述假设。我们首先证明一个收缩结果，该结果与前面给出的结果结合后，将得出一个分离结果。

3.1 偏好嵌入会降低学习到的信息量

第一个定理是一个 Strack-Yang式压缩结果。递增差异（假设 3）确定了最优后验分布可比较时比较静态的方向，而假设 4 确保优化器对应落在一条凸序链上，从而使这一比较静态有明确的定义。

定理1 （偏好嵌入导致保均值收缩）。 假设 *1*、*3*和*4* 均成立。对于每个 $t \in [0,1]$，令 Q_t 为（12）的一个解，其贝叶斯风险为H_t。

若 $t_1 > t_0$，则在 t_0 处的每个极小化者 Q_{t_0} 都凸序支配在 t_1 处的每个极小化者 Q_{t_1}:

$$Q_{t_0} \succeq_{\mathrm{cx}} Q_{t_1}. \tag{18}$$

证明。 固定 $t_1 > t_0$。令 Q_{t_1} 和 Q_{t_0} 分别为 t_1 和 t_0处的极小化解。由最优性可知

$$\mathbb{E}[H_{t_1}(Q_{t_1})] + C(Q_{t_1}) \le \mathbb{E}[H_{t_1}(Q_{t_0})] + C(Q_{t_0}), \tag{19}$$

and

$$\mathbb{E}[H_{t_0}(Q_{t_0})] + C(Q_{t_0}) \le \mathbb{E}[H_{t_0}(Q_{t_1})] + C(Q_{t_1}). \tag{20}$$

将（19）与（20）相加会抵消成本项，并得到

$$\mathbb{E}[H_{t_1}(Q_{t_1}) - H_{t_0}(Q_{t_1})] \le \mathbb{E}[H_{t_1}(Q_{t_0}) - H_{t_0}(Q_{t_0})]\,. \tag{21}$$

根据假设4，要么 $Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ （此时问题已解决），要么 $Q_{t_1} \succeq_{\text{cx}} Q_{t_0}$。假设 $Q_{t_1} \succeq_{\text{cx}} Q_{t_0}$。那么，假设3 意味着

$$\mathbb{E}[H_{t_1}(Q_{t_1}) - H_{t_0}(Q_{t_1})] \geq \mathbb{E}[H_{t_1}(Q_{t_0}) - H_{t_0}(Q_{t_0})]\,,$$

当 $Q_{t_1} \overset{d}{\neq} Q_{t_0}$时，存在严格不等式。这与（21）矛盾，除非 $Q_{t_1} \overset{d}{=} Q_{t_0}$。在这个临界情形下，$Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ 也成立。因此，$Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ 在所有情况下都成立。 □

此外，加入关于学习摩擦的假设2 可得如下结论：

推论1 （收缩弱化地降低学习成本）。除定理 *1*的假设外，设假设 *2* 成立。那么，对于任意 $t_1 > t_0$ 及其对应的任意极小化解 Q_{t_0} 和 Q_{t_1}，

$$C(Q_{t_0}) \geq C(Q_{t_1}).$$

此外，如果假设 *2* 中的不等式在 $Q \succeq_{\text{cx}} Q'$ 且 $Q \overset{d}{\neq} Q'$时严格成立，那么$C(Q_{t_0}) > C(Q_{t_1})$ 在 $Q_{t_0} \overset{d}{\neq} Q_{t_1}$时成立。

证明。由定理 1，$Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$。假设 2 进而意味着 $C(Q_{t_0}) \geq C(Q_{t_1})$。严格性条款的证明类似。 □

定理1 是对 ACMM 风格主张的形式化表述：偏好嵌入会削弱学习激励，并产生信息较少的预测器。这也是 Strack 和 Yang 的均值保持收缩刻画的直接对应版本，其中“偏好嵌入”起到额外限制的作用，从而引致收缩（Strack 和 Yang，2024年）。

3.2 分离原理：先学习概率，再做决策

下一个定理利用收缩结果建立稳健的福利比较。

定理2 （稳健分离原理）。假设 *1*、*3*和 *4* 均成立，并令 Q_0 和 Q_1 分别表示由无偏好训练和偏好嵌入训练所产生的后验信念随机变量。

那么，对于任意决策问题 $(\mathcal{A}, u)$（定义见第 *2* 节），

$$\mathbb{E}[V(Q_0)] \geq \mathbb{E}[V(Q_1)]. \tag{22}$$

此外，无偏好流程下的福利可通过仅在事后应用偏好来实现，即采用贝叶斯最优行动规则 $a^\star(q) \in \arg\max_{a\in\mathcal{A}}\{qu(a,1) + (1-q)u(a,0)\}$。

证明。 根据定理 1， $Q_0 \succeq_{\mathrm{cx}} Q_1$。引理 2 于是推出 $\mathbb{E}[V(Q_0)] \geq \mathbb{E}[V(Q_1)]$。

第二个陈述直接来自 $V(q)$ 的定义：它表示在信念为 q 时选择贝叶斯最优行动的价值。 □

定理 2 将 ACMM 所倡导的实用指导原则的一个强版本形式化。在给定的基本设定下，偏好的最佳位置是事后阶段，而非训练目标。证明刻意保持透明：收缩性加上凸性蕴含支配关系。

定理 2 普遍适用于关于二元状态 Y 的期望效用决策问题，其中预测器的输出是决策者关于该状态的信息输入。这比 Strack 和 Yang(2024年) 中的二元行动类别更为广泛，因为当前模型并未施加隐私约束来限制信号可行集。之所以能够实现如此广泛的适用性，是因为 V 对于几乎所有期望效用问题都是凸的（引理1）。如果决策者在部署时还会依据额外可观测量或约束 Z 进行决策，那么要进行普遍的福利比较，通常需要采用条件收缩（ $Q_0 \mid Z \succeq_{\mathrm{cx}} Q_1 \mid Z$ ，使得结果在 Z中逐状态成立），或者对完整信号进行更强的 Blackwell 比较；仅凭 Q的边际分布满足凸序支配，未必足够。

二元结果假设仍然很重要。在多维信念情形中，凸序是部分序，选择问题也变得更加突出。本文的关注点与经济学中许多实际的机器学习情形一致，在这些情形中，核心预测对象是标量风险评分。在附录A中，本文探讨了定理 2的多元版本，表明可以对二元结果假设进行一定程度的部分放宽。

3.3 闭式解示例

定理 1 和 2 针对一般类别的学习摩擦而建立。下一个示例展示了一个简单且完整推导的情形下的收缩机制，其中最优后验分布 Q_t 可以用闭式形式写出。

固定先验 $\mu \in (0,1)$。考虑二次贝叶斯风险族

$$H_t(q) := q(1-q) + t\,q^2 = q - (1-t)q^2, \qquad t \in [0,1]. \tag{23}$$

这满足假设3 根据命题3 ， 其中 h（q） $= q^2$ （凸）。增加 t 会使H_t 的凹性减弱，从而体现偏好嵌入会使学习激励趋于平缓这一理念。

令学习摩擦通过（分布不变）泛函取决于后验分散度

$$C(Q) := \frac{\lambda}{2}\Big(\mathbb{E}[Q^2] - \mu^2\Big)^2, \qquad \lambda > 0. \tag{24}$$

由于 $q \mapsto q^2$ 是凸函数，并且 $\mathbb{E}[Q^2] \geq \mu^2$ 对所有 $Q \in \mathcal{Q}(\mu)$，成本（24）是凸序单调的（假设2）。

对于任意 $Q \in \mathcal{Q}(\mu)$，记 $x := \mathbb{E}[Q^2]$。由于 $\mathbb{E}[Q] = \mu$ 是固定的，而 H_t 是二次函数，

$$\mathbb{E}[H_t(Q)] = \mu - (1-t)\,x,$$

因此学习问题（12）归结为一维优化问题 :

$$\min_{x \in [\mu^2,\, \mu]} \left\{ \mu - (1-t)\,x + \frac{\lambda}{2}(x - \mu^2)^2 \right\}. \tag{25}$$

唯一的极小化者为

$$x_t^\star = \left(\mu^2 + \frac{1-t}{\lambda}\right) \wedge \mu, \tag{26}$$

它关于 t 弱递减。

为获得显式的最优后验分布，选取取值于 $\{0, q_t^H\}$、均值为 μ 且二阶矩为 $x_t^\star$ 的两点后验分布 Q_t：

$$Q_t = \begin{cases} 0 & \text{with probability } 1 - \dfrac{\mu}{q_t^H}, \\[2ex] q_t^H & \text{with probability } \dfrac{\mu}{q_t^H}, \end{cases} \qquad q_t^H := \frac{x_t^\star}{\mu} \in [\mu, 1]. \tag{27}$$

那么 $\mathbb{E}[Q_t] = \mu$ 且 $\mathbb{E}[Q_t^2] = x_t^\star$因此 Q_t 满足（12）。

最后，t he 家族 $\{Q_t\}_{t \in [0,1]}$ 在凸序下是全序的。对于任意 $k \in [0,1]$,

$$\mathbb{E}[(Q_t - k)_+] = \begin{cases} \mu\left(1 - \dfrac{k}{q_t^H}\right) & \text{if } k < q_t^H, \\[2ex] 0 & \text{if } k \geq q_t^H, \end{cases}$$

它关于 q_t^H逐点递增。因此，如果 $t_1 > t_0$ 那么 $q_{t_1}^H \leq q_{t_0}^H$ 从而 $Q_{t_0} \succeq_{\mathrm{cx}} Q_{t_1}$。引理2 则意味着 $\mathbb{E}[V(Q_{t_0})] \geq \mathbb{E}[V(Q_{t_1})]$ 对每个期望效用决策问题均成立，正如定理2所示。

这一示例说明了一个更一般的观点。从无摩擦基准$C \equiv 0$ （在该示例中即 $\lambda = 0$）出发。此时，训练只需选择一个贝叶斯可行的后验分布，以最小化预期贝叶斯风险。由于贝叶斯风险是凹函数，

目标会弱倾向于选择分散程度更高的后验分布，因此完全揭示是一个最优解（当贝叶斯风险严格凹时，它还是唯一的最优解；在二次函数示例中， $t = 1$ 处的贝叶斯风险是线性的，因此每个贝叶斯可行的 Q 都是最优的）。在该基准下，偏好嵌入训练与无偏好训练的结果相同，因为学习没有成本，所以偏好只会影响下游行动规则。现在引入 $C > 0$，它反映出由于容量、优化、正则化或数据约束，信息量是有成本的。训练于是成为一个真正的信息获取问题：你需要在贝叶斯风险降低与 C（Q）之间进行权衡，从而选择后验分布应达到的分散程度。关键在于，偏好嵌入改变了贝叶斯风险的曲率，使其关于信念的凹性减弱，因此提高分散程度的边际收益降低。在相同的学习摩擦下，优化器因而会在偏好嵌入训练中转向后验分布的保均值收缩。由于间接效用关于信念是凸的，这种收缩会转化为稳健的福利损失，因此，应用偏好的最优时机是在事后通过决策规则进行。

3.4 一种共同决策环境

我们继续讨论引言中描述的二元统计决策问题。在该环境中，决策者观察数据，形成后验信念 $q = \Pr(Y = 1 \mid \cdot)$，并在面临不对称误分类成本时作出选择 $a \in \{0, 1\}$ c_{FP} 和 c_{FN}。如前文所示，最优策略是一个阈值规则：当且仅当 $q \geq \tau$ 时采取行动，其中 $\tau = c_{\mathrm{FP}}/(c_{\mathrm{FP}} + c_{\mathrm{FN}})$。引言还对比了两种流程：后处理，即先学习校准概率，再应用阈值；以及偏好嵌入，即使用成本敏感损失（例如类别加权交叉熵）进行训练，然后对所得评分应用默认截断值。在无摩擦基准下，通过选择权重，使诱导出的评分阈值实现相同的τ，这两种流程在决策上可以等价。

本节旨在解释：一旦学习过程以实践中通常存在的方式变得不完美（有限数据、正则化、有限容量、优化误差），这种等价性为何可能失效。其机制并不是权重“选择了错误的阈值”；引言已经明确说明了如有需要应如何对齐阈值。相反，嵌入会改变 学习目标 ，从而改变信息性的边际价值。通过贝叶斯风险的曲率可以方便地理解这一点。令 $H^0(q)$ 表示未加权对数损失的贝叶斯风险，并令 $H^w(q)$ 表示引言中加权损失所诱导的贝叶斯风险。直接计算可得

加权损失。直接计算可得

$$H^{w\prime\prime}(q) = \rho(q)\, H^{0\prime\prime}(q), \qquad \rho(q) := \frac{w_0 w_1}{(1-q)w_0 + q w_1}. \tag{28}$$

在对权重进行标准的尺度归一化后，$\rho(q) \le 1$，因此嵌入式目标函数的凹性较弱：后验分布的离散程度增加，对期望贝叶斯风险的降低幅度更小。当信息量需要付出代价时（无论原因如上所述为何），这种较弱的曲率都会促使学习得到的后验分布发生保均值收缩：与分离式训练相比，预测器的输出会向先验靠拢。

一个两点示例清楚地展示了其福利影响。假设分离式训练得到的校准后验分布取值为

$$Q_0 \in \{\mu - \delta,\ \mu + \delta\} \quad \text{with probability } \tfrac{1}{2} \text{ each},$$

其中，$\mu = \mathbb{E}[Q_0]$ 和 $\delta > 0$ 表示信息量（因此 $\mathrm{Var}(Q_0) = \delta^2$）。假设 $\mu - \delta < \tau < \mu + \delta$，因此最优阈值规则仅在 $Q_0 = \mu + \delta$ 时采取行动。在存在学习摩擦的偏好嵌入下，令学习得到的后验分布为一个保均值收缩

$$Q_1 \in \{\mu - \delta^w,\ \mu + \delta^w\} \quad \text{with probability } \tfrac{1}{2} \text{ each}, \qquad 0 < \delta^w < \delta,$$

因此 $\mathrm{Var}(Q_1) = (\delta^w)^2 < \mathrm{Var}(Q_0)$。如果收缩幅度足够大，使得 $\mu + \delta^w < \tau$，那么嵌入式流程产生的后验概率永远不会越过具有经济意义的阈值，决策者因而最优地选择永不采取行动。与分离相比，这显然是次优的，因为 $\mu + \delta > \tau$ 意味着，在 Q_0 下，高后验状态时采取行动是最优的，但在 Q_1 下，该状态已无法区分。将损失写成负收益后，分离情况下的期望损失为

$$\mathbb{E}[\ell_{\mathrm{sep}}] = \tfrac{1}{2}\Big[(\mu - \delta)c_{\mathrm{FN}} + \big(1 - (\mu + \delta)\big)c_{\mathrm{FP}}\Big],$$

而在嵌入状态下（当 $\mu + \delta^w < \tau$ 时），决策者从不采取行动，因此产生

$$\mathbb{E}[\ell_{\mathrm{emb}}] = \mathbb{E}[Q_1]\, c_{\mathrm{FN}} = \mu\, c_{\mathrm{FN}}.$$

因此，福利损失为

$$\mathbb{E}[\ell_{\mathrm{emb}}] - \mathbb{E}[\ell_{\mathrm{sep}}] = \tfrac{1}{2}\Big[(\mu + \delta)c_{\mathrm{FN}} - \big(1 - (\mu + \delta)\big)c_{\mathrm{FP}}\Big] \ > \ 0, \tag{29}$$

严格正值性由 $\mu + \delta > \tau$ 得出（等价地，在高

后验概率）可知。因此，即使嵌入式目标函数在无摩擦基准下与决策者的偏好一致，偏好嵌入仍可能降低福利，因为它会降低学习得到的信号的信息量，而事后阈值化无法弥补这一点。定理 1 和2将这一直觉形式化。

4 分离原理下的损失函数选择与实施

分离原理回答了偏好应在流程的何处 介入。仅凭这一原理，并不能确定唯一的无偏好损失函数，因为多种严格适当损失都能得出校准概率，而且它们与学习摩擦的相互作用可能各不相同。

本节阐明，当我们将关注范围限定于实践中使用的损失函数时，哪些结论是确定的，哪些不是；同时明确说明应如何理解“无偏好学习在各种变体中最优”。在这方面，本节将形式化结果转化为简明的实施建议，并通过一个常见示例展示其中的关键机制。

4.1 设计 类别以及“在各种变体中最优”的含义

定理 2 比较无偏好族系（$t = 0$）与嵌入偏好的族系（$t = 1$，并固定底层学习技术 C（·）。这与 ACMM 提出的实际问题十分相近：非对称偏好应通过修改训练损失函数来实施，还是通过对校准概率估计进行后处理来实施?

要讨论变体之间的最优性，就必须指定一个设计类别。具体来说，如果设计者可以任意调整损失函数的尺度，就需要进行归一化，因为缩放会改变式（12）中预期贝叶斯风险与学习摩擦之间的权衡。在应用研究中，缩放还会与优化选择（学习率、提前停止、正则化）相互作用，因此，将 $C(\cdot)$ 视为固定时，最自然的做法是同时将允许的损失函数尺度视为固定或经过归一化。

有两点值得澄清。首先，由于后验分布是贝叶斯可行的（$\mathbb{E}[Q] = \mu$），在贝叶斯风险中加入关于 q 的仿射函数不会影响学习问题：若 $\tilde{H}(q) = H(q) + aq + b$，则 $\mathbb{E}[H(Q)] = \mathbb{E}[H(Q)] + a\mu + b$，因此（12）的极小化解集合保持不变。因此，对不同损失函数的比较，自然应视为在贝叶斯风险中忽略仿射项。其次，乘法缩放确实会产生影响，应视为设计选择的一部分：将损失函数乘以一个常数，相当于改变（12）中的有效学习摩擦计划。因此，下文的排序结果应被理解为：在候选损失函数尺度固定的实际相关类别中提供指导（例如

按照惯例、通过优化超参数，或通过显式归一化来固定尺度）。

4.2 选择无偏好基准

在概率分类中，许多常用损失函数都具有局部性，即某个结果对应的损失只取决于分配给该实际结果的概率。在二元情形中，这一限制没有实质意义：任何损失函数 $L(p, y)$ 都可以写成局部形式，只需令 $\ell_1(p) := L(p,1)$，并令 $\ell_0(s) := L(1-s,0)$。因此，局部性在此并不能选出唯一的严格适当评分规则。[9]

因此，二元情形下分离的实质内容在于位置而非唯一性：先使用无偏好严格适当损失函数进行训练，以学习校准概率，再在事后应用偏好。当存在多种可用的严格适当损失函数时，选择哪一种可能会产生影响，因为它们的贝叶斯风险曲率决定了生成更分散后验分布所带来的边际收益，进而与学习摩擦相互作用。第4.3 节将对这一曲率渠道进行形式化说明。

4.3 无偏好损失函数上的曲率序

当有多个适当损失可供选择时，该模型提供了一个简单的比较原则：在尺度归一化的条件下，应优先选择其贝叶斯风险能够更有力地激励信息性后验分散度的损失。对贝叶斯风险函数定义一种序，是对此进行形式化的一种简便方法。

定义5（贝叶斯风险凹性序）。对于两个贝叶斯风险函数 H 和 $\tilde{H}$，若 $H - \tilde{H}$ 是凸的，则记作 $\tilde{H} \preceq H$。

若 $H - \tilde{H}$ 是凸的，则 $\tilde{H}$ 的“凹性较弱”：相较于 H，目标 $\tilde{H}$ 对分散度增加了一个凸惩罚。在二元情形下，这种序很容易验证。特别地，若 H 和 $\tilde{H}$ 在 $(0,\ 1)$ 上二次可微，则 $\tilde{H} \preceq H$ 等价于 $H''(q) - \tilde{H}''(q) \geq 0$ 对所有 $q \in (0,\ 1)$ 均成立。

下一个命题将“最优损失函数”的解释精确化。它在第5节所用相同的基本设定下，将定义5中的曲率次序转化为对学习得到的后验分布的排序。

命题4（曲率排序的适当损失会带来信息量更大的学习）。设有两个严格适当评分规则 L^0 和 L^1，其贝叶斯风险分别为 H_0 和 H_1，并假设 $H_0 \preceq H_1$。

[9] 在真正的多类别情形（至少有三个结果类别）中，局部性会成为一项实质性限制，并且在标准正则条件下可据此唯一确定对数评分；参见Gneiting 和 Raftery (2007年)及其中的参考文献。

对于 $t \in [0,1]$，定义插值损失 $L^t := (1-t)\,L^0 + tL^1$，并令 H_t 表示其贝叶斯风险。继续采用假设*1*，并假设假设 *4*对族 $\{H_t\}_{t\in[0,1]}$成立。对于每个 t，令 Q_t 为（(12) 的任意极小化者，其贝叶斯风险为 H_t。那么，若 $t_1 > t_0$，

$$Q_{t_0} \succeq_{cx} Q_{t_1}.$$

特别地，在 L^0 凸序下进行训练所诱导的后验分布，支配在 L^1 下训练所诱导的后验分布。

此外，对于任意决策问题 $(\mathcal{A}, u)$，

$$\mathbb{E}[V(Q_0)] \;\geq\; \mathbb{E}[V(Q_1)]\,.$$

证明。 由于 L^t 是严格适当损失的凸组合，因此它是严格适当损失，且其贝叶斯风险满足 $H_t = (1-t)H_0 + tH_1 = H_0 + t(H_1 - H_0)$。由于 $H_1 - H_0$ 是凸的，命题 3 表明，族 $\{H_t\}_{t\in[0,1]}$ 满足假设 3。因此，定理 1 给出 $Q_{t_0} \succeq_{cx} Q_{t_1}$，对任意 $t_1 > t_0$均成立。福利不等式由引理 2得出。 □

命题 4给出了一种易于处理的"无偏好变体之间的最优性"概念：在固定的设计类别中，贝叶斯风险更凹的损失函数会诱导信息量（弱）更高的学习得到的后验分布，因此在 所有期望效用决策问题中都具有弱支配性。序 $\preceq$是一种偏序，因此并非每对损失函数都可以比较。不过，对于许多参数族或实践中常用的损失函数，检验增量的凸性都很简单。

4.4 对数损失与布里尔损失

这里用两种标准的严格适当损失来说明命题 4 。对数损失下的贝叶斯风险是香农熵，

$$H_{\log}(q) = -q\log(q) - (1-q)\log(1-q), \tag{30}$$

而布里尔损失下的贝叶斯风险是

$$H_{\text{Brier}}(q) = q(1-q). \tag{31}$$

一个 dir 直接计算表明，增量 $H_{\text{Brier}} - H_{\log}$ 在 (0，1) 上是凸函数：

$$\big(H_{\text{Brier}}(q) - H_{\log}(q)\big)'' \;=\; \frac{1}{q(1-q)} - 2 \;>\; 0 \qquad \forall q \in (0,1).$$

因此， $H_{\log} \preceq H_{\text{Brier}}$ 按照定义5的含义。在命题4 的条件下（特别是，固定相同的学习摩擦 $C(\cdot)$，并假设沿插值过程具有可比性），采用对数损失进行训练所诱导的后验分布凸序支配采用布里尔损失所诱导的后验分布：

$$Q_{\log} \succeq_{cx} Q_{\text{Brier}}.$$

根据引理2，这意味着一种稳健的福利比较：对于任何下游决策问题，$\mathbb{E}[V(Q_{\log})] \geq \mathbb{E}[V(Q_{\text{Brier}})]$。

结论依赖于固定式（12）中的学习技术，包括损失函数的有效尺度。对损失函数重新缩放，等同于重新缩放学习摩擦，并可能颠覆曲率比较。这正是第4.1 节强调“在各种变体中最优”必须以某个设计类别和尺度归一化为前提的原因。

4.5 进一步思考

收缩逻辑在预测概率层面有一个可观测的含义。在定理1 （偏好嵌入）或命题4（改用凹性较弱的适当损失）的条件下，校准后的预测概率 $\hat{q}(X)$ 的诱导分布预计会发生保均值收缩。在实证研究中，可以通过校准来实现这一点（从而使不同流程的均值对齐）；在命题4 所比较的严格适当损失中，根据贝叶斯可行性，均值相同，因此均值对齐主要是出于实践便利。等价地，对于一组截断点 $k \in [0,1]$，可以比较经验函数 $k \mapsto \mathbb{E}\big[(\hat{q}^{(t)}(X) - k)_+\big]$，其逐点排序刻画了凸序。这个含义是可证伪的：如果这些曲线系统性地交叉，就表明收缩机制并不适用于该环境。

本文的理论以精确的方式进一步支持了 ACMM。他们的实证发现表明，过度重新加权可能降低学习质量，而且事后调整可能优于偏好嵌入训练（Caplin 等人，2022年；Autor 等人，2025年）。定理 1通过表明偏好嵌入可能起到有效的信息量惩罚作用，从而导致学习得到的后验分布发生保均值收缩，对这一现象给出了理论解释。定理 2 随后将这种收缩转化为针对一大类经济决策问题的福利比较。

命题 4指出了一个在许多应用实践中隐含的额外要点：即使在无偏好严格适当损失函数这一类中，曲率也会影响学习激励。当两个严格适当损失的贝叶斯风险相差一个凸增量（保持尺度不变）时，贝叶斯风险更具凹性的损失会诱导出包含更多信息的学习得到的后验分布，并且

因此对于任何下游决策问题都弱占优。

与此同时，这些结果也明确了适用范围。本文并未声称偏好嵌入总是有害的。当学习技术违反假设3或假设2时，偏好嵌入原则上可能有利于特定目标。这些情形在实践中很重要，但与本文分析的分离设定不同。

5 偏好嵌入何时占据主导

定理2 中的分离原理刻意设定得很强：它假设，一旦获得经过校准的后验概率，决策者便可以毫无成本地解读这些概率，并实施贝叶斯最优行动规则。这是标准的决策理论基准，也支撑着经典的实施建议：先训练模型预测概率，再事后施加偏好。

本节通过引入一种基于理性注意力（RI）文献（Sims，2003年）的决策阶段摩擦，对这一基准作出限定。核心思想很简单：即便预测模型提供了有效的后验概率，使用该后验概率仍需耗费稀缺的认知资源（注意力、时间、工作记忆和概率推理能力）。在理性注意力模型中，认知努力随所处理的信息量而增加。因此，认知成本取决于随机环境：使用信息量高的预测器比使用信息量低的预测器要求更高。此时，偏好嵌入可能成为最优选择，因为它实际上压缩了预测信号，减轻了用户的认知负担，尽管这会像定理2所示那样牺牲信息量。

由此产生的权衡一目了然：无偏好训练能产生信息量更高的后验分布（决策价值更高），但注意力成本也更高；偏好嵌入训练产生的后验分布信息量较低（决策价值较低），但处理成本更低。

5.1 研究动机

实证研究记录了人类—人工智能流程中后处理的系统性失效。Agarwal等人（2023年）研究了接收人工智能生成的医学诊断概率评估的放射科医生。尽管临床医生获知人工智能评分已得到良好校准，但许多人仍用自己准确性较低的判断推翻了人工智能有把握的预测。总体而言，人类—人工智能组合的表现差到作者建议，将许多病例分配给不使用人工智能支持的人类，或交由人工智能处理而不进行人工干预。

这种行为直接违反了分离基准：校准后的概率不会自动转化为最优行动。一种自然的解释是，处理概率的实际认知成本很高：用户可能默认采用粗略启发式、忽略概率信息，或在时间压力下不一致地运用这些信息。

更广泛地说，关于算法厌恶的文献记录了这样一种现象：即使机器生成的建议能够提高准确性，许多用户仍会低估或拒绝这些建议（Tejeda 等，2022年）。认知经济学为这种行为提供了理论基础：使用概率信息需要稀缺的认知输入，而这些输入具有机会成本（Caplin，2025年）。本节使用基于理性疏忽的注意力成本，将这些观察形式化。

5.2 扩展设置

这里的扩展保留了第2节中的二元结果环境。令 $V(q)$ 表示决策者针对相关下游任务的间接价值函数。如引理1所确立的，V 对于任意期望效用决策问题，均是关于信念 q 的凸函数。

令 Q_0 和 Q_1 分别表示由无偏好训练和偏好嵌入训练所诱导的后验信念随机变量。分离原理（定理2）意味着，在不存在决策阶段摩擦的情况下，无偏好训练能够产生不低于偏好嵌入训练的价值。我们将嵌入所导致的信息损失定义为：

$$\Delta_I := \mathrm{E}[V(Q_0)] - \mathrm{E}[V(Q_1)] \geq 0. \tag{32}$$

该数量 Δ_I 捕捉了学习得到的后验信念因凸序收缩而导致的决策价值的总体减少，尚未计入处理该信息所需的任何认知成本。

5.3 决策阶段的理性疏忽

第5节中的认知成本模型旨在捕捉一个简单事实：从概率预测中提取与决策相关的内容需要认知努力。对此类努力，一个自然的分析框架是理性疏忽文献（Sims，2003年），其中注意力成本随状态与所处理信号之间的互信息而增加。

具体而言，令 S 表示决策者关注的（经处理的）信号，并令$Q = \Pr(Y = 1 \mid S)$ 表示由此诱导的后验概率。互信息满足：

$$I(Y;S) = H(Y) - H(Y \mid S) = h(\mu) - \mathrm{E}[h(Q)], \tag{33}$$

其中 $\mu = \Pr(Y=1)$ 和 $h(q) = -q\log q - (1-q)\log(1-q)$ 是二元熵函数。[10]等价地，

$$I(Y;S) = \mathrm{E}\big[\mathrm{KL}\big(\mathrm{Bern}(Q) \,\|\, \mathrm{Bern}(\mu)\big)\big]. \tag{34}$$

由于 Q 是关于 Y 的 S 的充分统计量，因此可以写作 $I(Y;S) = I(Y;Q)$。因此，我们将决策阶段认知努力建模为与 $I(Y;Q)$ 成正比：信号对 Y （相对于先验）的不确定性削减越多，为了利用该信号就必须吸收越多信息。

在标准理性疏忽中，主体可以在熵成本约束下设计最优信息结构。在人机协作场景中，用户通常面对的是一份已部署工具（界面）的菜单，而不是随意重新设计信号的能力。放射科医生无法重新设计模型的用户界面；他们需要决定在多大程度上依赖模型、是否咨询模型，以及是否默认采用自己的基线启发式。

我们用一个基于菜单的理性疏忽假设来刻画这种产品环境：用户可以（i）忽略工具并依据先验采取行动，或（ii）按原样处理工具提供的信号，然后根据由此得到的后验概率采取最优行动。

假设5 （有限菜单下的决策阶段理性疏忽成本）。 固定一个注意力价格参数 $\lambda_{\mathrm{cog}} > 0$。在流程 $i \in \{0,1\}$下，用户在以下选项中作出选择：

(a) 忽略工具，依据先验采取行动，获得价值 $V(\mu)$；或

(b) 使用工具：处理其输出以获得后验概率 Q_i，然后选择贝叶斯最优行动 $a^\star(Q_i)$，并承担认知成本 $\lambda_{\mathrm{cog}} \cdot I(Y;Q_i)$。

用户无法在不付出成本的情况下，对工具的输出进行重新设计、模糊化或重新格式化，使其超出这一有限菜单。

假设5 是最小的 RI 结构，它：**(i)** 使认知成本内生于所部署预测的信息量；**(ii)** 将算法厌恶/有限采用纳入一个明确的外部选项（忽略工具并依据先验采取行动）。“有限菜单”条款并非技术性权宜之计；它反映了这样的制度现实：信号重新设计是在上游作出的工程决策（或者其本身就具有认知成本）。

5.4 注意力—信息量权衡

在流程 $i \in \{0,1\}$ 下，用户的福利为：

$$W_i(\lambda_{\mathrm{cog}}) := \max\Big\{V(\mu),\ \mathrm{E}[V(Q_i)] - \lambda_{\mathrm{cog}}\, I(Y;Q_i)\Big\}. \tag{35}$$

[10]对数可以取以 e （纳特）为底，也可以取以 2（比特）为底。这只会对注意力价格参数进行缩放。

定义嵌入带来的 注意力节省：

$$\Delta_{MI} := I(Y;Q_0) - I(Y;Q_1). \tag{36}$$

下一个引理将注意力节省直接与推动本文的收缩结果联系起来。

引理4 （偏好嵌入减少互信息）。 如果 $Q_0 \succeq_{\text{cx}} Q_1$ （如定理 *1*中所述），那么 I（Y；Q_0） $\geq I$（Y； Q_1），因此 $\Delta_{MI} \geq 0$。

证明。由 $h(\cdot)$ 的凹性， $Q_0 \succeq_{\text{cx}} Q_1$ 意味着 $\mathrm{E}[h(Q_0)] \leq \mathrm{E}[h(Q_1)]$。利用（33）， $I(Y;Q) = h(\mu) - \mathrm{E}[h(Q)]$ 因此对于 Q_0 比对于 Q_1 更大。 □

引理 4 使这种权衡变得直接。偏好嵌入会降低预期决策价值（降低幅度为 $\Delta_I \geq 0$），并且 会降低使用该工具的注意力成本（降低幅度为 $\lambda_{\text{cog}}\Delta_{MI} \geq 0$）。问题在于哪种影响占主导地位。

5.5 嵌入在理性疏忽下占主导地位时

现在我们可以刻画理性疏忽扩展中的主要结果。

定理3 (理性疏忽逆转条件)。 假设假设 *5* 成立，且 $\Delta_{MI} > 0$。

(a) *(*何时会使用工具？ *)* 当且仅当满足以下条件时，才会使用流程 i （而非忽略它）：

$$\lambda_{\text{cog}} < \bar{\lambda}_i, \qquad \bar{\lambda}_i := \begin{cases} \dfrac{E[V(Q_i)] - V(\mu)}{I(Y;Q_i)} & \text{if } I(Y;Q_i) > 0, \\ 0 & \text{if } I(Y;Q_i) = 0. \end{cases} \tag{37}$$

当 $I(Y;Q_i) = 0$ 时，贝叶斯可行性意味着 $Q_i = \mu$ 几乎处处成立，因此该工具产生 $V(\mu)$，且绝不会严格优于忽略该工具。

(b) *(*当两种工具均被使用时，嵌入与后处理的比较。 *)* 若 $\lambda_{\text{cog}} < \min\{\lambda_0, \lambda_1\}$，则偏好嵌入训练带来更高的福利， $W_1(\lambda_{\text{cog}}) > W_0(\lambda_{\text{cog}})$，当且仅当满足以下条件：

$$\lambda_{\text{cog}} > \lambda^\star := \frac{\Delta_I}{\Delta_{MI}}. \tag{38}$$

等价地，当注意力节省超过信息损失时，嵌入才占优：

$$\lambda_{\text{cog}}\Delta_{MI} > \Delta_I.$$

(c)（纯采用渠道。）如果 $\lambda_1 > \lambda_0$，那么对于所有 $\lambda_{\mathrm{cog}} \in (\lambda_0, \lambda_1)$，无偏好工具都会被忽略，而嵌入式工具则会被使用，这意味着 $W_1(\lambda_{\mathrm{cog}}) > W_0(\lambda_{\mathrm{cog}})$。

证明。（a）部分由（35）得出：当且仅当 i时，才会使用流程

$$\mathrm{E}[V(Q_i)] - \lambda_{\mathrm{cog}} I(Y; Q_i) > V(\mu).$$

若 $I(Y; Q_i) = 0$，则（33）意味着 $\mathrm{E}[h(Q_i)] = h(\mu)$，而 h 的严格凹性推出 $Q_i = \mu$几乎必然成立；因此该不等式不成立（等号成立），且绝不会严格使用该工具。若$I(Y; Q_i) > 0$，整理后可得（37）。

对于（b）部分，当两者都被使用时，我们比较净价值：

$$\begin{aligned} &\mathrm{E}[V(Q_1)] - \lambda_{\mathrm{cog}} I(Y; Q_1) > \mathrm{E}[V(Q_0)] - \lambda_{\mathrm{cog}} I(Y; Q_0) \\ &\qquad \Longleftrightarrow \ \lambda_{\mathrm{cog}} (I(Y; Q_0) - I(Y; Q_1)) > \mathrm{E}[V(Q_0)] - \mathrm{E}[V(Q_1)]. \end{aligned}$$

代入 Δ_{MI} 和 Δ_I。

（c）部分可由（a）部分直接得出。 □

定理 3 给出了清晰的“每比特价值”解释。比率

$$\lambda^\star = \frac{\Delta_I}{\Delta_{MI}}$$

是用户感到无差异时的注意力价格。分子 Δ_I 表示通过嵌入而损失的决策价值（信息量较低的后验概率）。分母 Δ_{MI} 表示用户无需处理的状态信息比特数/纳特数。相对于这些额外比特的认知成本，当其边际价值较低时，嵌入占优。

主要的比较静态结果可直接由 $\lambda^\star = \Delta_I / \Delta_{MI}$得到。首先，随着 λ_{cog}上升（时间压力、多任务处理、有限的统计训练），无偏好工具不仅较不可能被使用（第（a）部分），而且在使用的条件下，也较不可能被偏好（第（b）部分）。这为算法厌恶和低采用率提供了自然的微观基础：信息量很高的概率工具可能会被忽视，因为利用它们的认知成本很高。

其次，当学习摩擦严重时，Q_0 和 Q_1 都可能接近于没有信息量，而信息损失 Δ_I 也很小。在这种情况下，即使是适度的注意力节省，也足以使嵌入获得合理性。这与本文更广泛的观点一致：当可实现的信息性前沿较低时，压缩预测的机会成本也较低。

第三，在信息损失 Δ_I保持不变的情况下，当注意力节省 Δ_{MI} 较大时，偏好嵌入更有可能占据主导地位，因为阈值 $\lambda^\star = \Delta_I / \Delta_{MI}$ 会下降。

更一般地说，决定这种逆转的是比率 Δ_I/Δ_{MI}：当上游学习问题的变化使无偏好信号包含更多信息时，Δ_I 和Δ_{MI} 通常都会发生变化，因此在没有进一步限制的情况下，$\lambda^\star$ 的净效应无法确定符号。

最后，概率推理能力的个体差异可以表示为 λ_{cog}方面的异质性。例如，Caplin 等人（2024年记录了校准能力方面稳定的异质性；在当前框架中，这种异质性自然地对应于有效注意力价格的差异，因而也对应于用户偏好的流程差异。

5.6 为什么有限菜单假设很重要

假设 5 将用户限制在一个已部署工具的菜单中，而不是允许他们自由地重新设计信号。这一限制至关重要。如果决策者能够无成本地使任何收到的信号发生信号退化，那么信息量更大的信号就不可能更差：用户总可以选择最优地忽略或粗化该信号，从而复制在信息量较小的信号下可实现的任何结果。

具体而言，如果 Q_1 是 Q_0 的信号退化（如定理 2所依据的收缩解释中所述），那么从 Q_0 可实现的可行后验集合包含从 Q_1可实现的集合，理性不注意主体从 Q_0 开始，并在需要时对其进行最优压缩，就可以弱支配嵌入式流程。因此，嵌入逆转从根本上说是关于部署刚性的论断：用户通常无法获得对模型输出进行无成本且经过最优设计的信息退化；他们面对的是界面。

这一观察也阐明了一项重要的设计启示：如果设计者能够保持无偏好训练并提供认知负担较轻的界面（例如，离散化风险分档或明确建议），那么通常就能在不牺牲训练信息量的情况下，获得注意力节省带来的益处。偏好嵌入是将认知负担前移的一种方式，但并非唯一方式。

6 基于人类反馈的强化学习作为偏好嵌入

本文探讨一个简单的设计问题：何时应将非对称偏好纳入学习目标，何时又应通过后处理和决策规则在学习之后加以实现？本文的形式化结果针对关于二元结果的标量后验概率展开，但在现代人工智能系统中，“偏好的位置”这一问题也一再出现。

本节回答两个问题，将理论与当前大语言模型（LLM）的实践联系起来。

1. **该模型与大语言模型有何关联？** 许多已部署的大语言模型系统都包含标量—二元组件，其形式恰好与第2–2.1节研究的形式一致：安全性和内容审核分类器、质量或可接受性评分器，以及将评分映射为离散操作（响应、拒绝、重新生成或升级处理）的路由规则。对于这些模n 块，这种映射是直接的，而分离原理也具有明确的实施意义：使用无偏好严格适当损失来学习校准概率，然后通过下游阈值、路由或最佳项选择来实施非对称权衡。

2. **基于人类反馈的强化学习在这些结果中处于什么位置？** 基于人类反馈的强化学习（RLHF）是一种生成器端对齐方法，它会修改大语言模型的输出分布，使得被评分为“偏好”的输出变得更有可能出现。从本文的视角来看，RLHF 是一种偏好嵌入：它将实施特定下游标准权重的负担，从推理阶段前移到了训练阶段。当部署目标稳定且学习得到的奖励经过良好定义时，这可能是有益的（定理 4）。但当目标发生变化、具有多维性或随时间漂移时，它可能带来能力税（定理 5）；而当学习得到的奖励与真实质量不对齐时，它还可能放大奖励错配问题（命题 6）。其实际权衡与正文中的结论相同：嵌入以降低推理时成本为代价，换取稳健性和选择灵活性。

回答这些问题涉及同一个设计问题，但采用了不同的形式化基本要素。定理 1–2 比较具有共同均值且按凸序排列的后验信念随机变量。相比之下，基于人类反馈的强化学习结果研究训练如何通过指数倾斜改变生成器的输出分布；这一过程通常会改变均值，并且对于固定目标而言，使用一阶随机占优进行比较最为自然。因此，两者的共同点并不在于具体的信息序，而在于偏好的位置：将目标固化到训练中，可以降低推理时的搜索或选择成本，但当目标发生漂移或存在错配时，可能会降低稳健性。

讨论安排如下。第6.1节记录了二元可接受性评分的严格映射。其余部分将基于人类反馈的强化学习视为生成器的“嵌入还是后处理”选择，并着重说明通过本文结果解读基于人类反馈的强化学习所需的最低限度技术机制。

6.1 大语言模型部署中的二元可接受性评分

尽管论文中的形式化结果针对的是关于二元结果的标量后验概率，但许多大语言模型部署恰好包含此类标量—二元模块。固定一个提示 X 和一个候选补全 Z。令 $Y \in \{0,1\}$，其中 $Y = 1$ 表示该补全符合部署策略的要求，

论文对象	**大语言模型系统类比（二元可接受性设定性）**
结果 $Y \in \{0,1\}$	候选补全是否可接受（安全、符合政策、质量高）
信号 S 和后验概率 Q	由以下模型生成的评分 安全性/奖励/质量模型及其隐含的可接受概率
无偏好严格适当损失	使用严格适当的损失函数训练评分模型 用于估计 $q = \mathbb{P}(Y = 1 \mid \cdot)$ 的损失函数
偏好嵌入损失	使用非对称损失函数训练评分模型（例如类别加权交叉熵），将下游权衡硬编码代码下游权衡
决策问题（$\mathcal{A}$, u）	部署 策略：展示/拒绝/重新生成/升级处理； 收益体现安全性和用户价值
后处理/决策规则 $a^\star(\cdot)$	阈值处理、路由、best-of-n选择、基于估计的n进行拒绝 基于估计的 q

表 1：将本文的框架映射到常见的大语言模型部署子问题：对候选输出进行二元可接受性评分。

而 $Y = 0$ 否则。评分模型会观察 (X, Z) 的特征，并生成信号 S，从而得到后验信念 $Q = \mathbb{P}(Y = 1 \mid S) \in [0, 1]$。随后，部署策略选择一个行动 $a \in \mathcal{A}$(展示 Z、拒绝、重新生成、提出后续问题，或转交人工审核)，其收益为 $u(a, Y)$。这正是第 2 节所述的环境。表 1列出了两者之间的对应关系。

在这种设定下，分离原则很直接：使用不含偏好的严格适当损失训练评分器，使其估计 $\mathbb{P}(Y = 1 \mid \cdot)$（并在需要时进行校准），然后通过下游策略 $a^\star(\cdot)$实现非对称偏好。具体而言，部署权衡的变化（例如更严格的安全标准或不同的延迟成本）自然应由后处理 和决策层处理，而非通过重新加权评分器的训练损失来实现。

这种严格的应用范围有意设得较窄：它适用于大语言模型系统中无处不在的评分与路由 模块。本节其余部分将转向另一类对象：端到端流程，例如基于人类反馈的强化学习，它们会直接修改生成器本身。

6.2 关于基于人类反馈的强化学习的经济学入门

大语言模型可以表示为给定提示 x 时，对补全结果 z 的条件分布 $\pi(z \mid x)$。预训练 通过在大型文本语料库上进行最大似然训练，得到一个基线生成器 π_0 。由此产生的分布能够生成高质量输出，却没有明确的方向：它反映的是训练数据中的模式，而非任何特定的部署目标。

基于人类反馈的强化学习是一种广泛使用的方法，用于将 π_0 转向人类认为更令人满意的输出。将其分为两个要素来理解会很有帮助。

1. **奖励建模**。 人类会比较针对同一提示的成对候选输出，并指出哪个更值得偏好。训练一个奖励模型 $r(x,\ z)$来预测这些偏好，根据训练偏好数据生成一个标量评分，用于概括某个输出看起来“有多好”。[11]

2. **带 KL约束的策略优化**。 随后更新生成器，以提高预期奖励，同时保持接近 π_0:

$$\max_{\pi}\ \mathbb{E}_{x\sim P(x)}\,\mathbb{E}_{z\sim\pi(\cdot|x)}[r(x,z)]\ -\ \lambda\cdot \mathrm{KL}(\pi\,\|\,\pi_0), \tag{39}$$

其中，$\mathrm{KL}(\pi\,\|\,\pi_0)\ =\mathbb{E}_x[\mathrm{KL}(\pi(\cdot\mid x)\ \|\,\pi_0(\cdot\mid x))]$ 表示从分布 π_0 到 π的 Kullback–Leibler 散度（相对熵），而 $\lambda>0$ 控制约束的强度。

对于经济学家而言，（39）是一个熟悉的对象：这是一个关于选择概率的正则化期望效用问题。KL项是一种调整成本（或信息成本），用于惩罚偏离基线分布的情况。与逻辑模型一样，参数 λ 决定选择概率的集中程度：较小的 λ 对应于对评分进行激进优化，而较大的 λ 则使策略更接近基线。

因此，关键的设计问题随即显现。可以通过求解（训练时的（39）（基于人类反馈的强化学习，RLHF）来实现偏好；也可以在部署时从 π_0 中采样候选项，并使用评分器在其中进行选择（重排序、拒绝采样、best-of-n或路由）。前者将特定奖励嵌入生成器；后者则使生成器保持广泛的能力，并以事后方式应用偏好，体现了分离的理念。

6.3 将基于人类反馈的强化学习作为偏好嵌入

为了将基于人类反馈的强化学习与论文中的对象联系起来，可将提示 x 对应的每个补全 z 视为具有一个（潜在的）可接受性 概率 $q(x,z)\ \in\ [0,1]$，用于衡量相关下游意义上的

[11]见 查看 gler 等人（2019年）；Ouyang 等人（2022年）；Rafailov 等人（2023年）是经典论述 .

“良好”输出。[12] 对于任意生成器 π 和提示 x，令 $Q_{\pi,x}$ 表示当 $Z \sim \pi(\cdot \mid x)$ 时，$q(x, Z)$ 的分布律。在提示分布 $P(x)$ 下，将无条件后验质量分布定义为以下混合分布

$$Q_\pi(\cdot) := \int Q_{\pi,x}(\cdot)\, dP(x).$$

等价地，如果 $X \sim P$ 和 $Z \sim \pi(\cdot \mid X)$，那么 Q_π 就是 $q(X, Z)$的分布律。这就像正文中那样，将生成器的比较简化为对 $[0,1]$, 上的分布进行比较。

假设 6（对齐的奖励模型）。 奖励模型与可接受性（基数意义上）对齐：对于所有 (x, z)，均有 $r(x, z) = q(x, z)$。

假设 6剥离了基于人类反馈的强化学习的机械效应，并抽象掉奖励设定错误的影响。在这一假设下，基于人类反馈的强化学习有一个明确的刻画。

命题 5（基于人类反馈的强化学习即指数倾斜）。 在假设 *6*下，(39)的解对每个提示都满足 x，

$$\pi_R(z \mid x) = \frac{\pi_0(z \mid x) \exp(q(x,z)/\lambda)}{Z_\lambda(x)}, \tag{40}$$

其中 $Z_\lambda(x) = \int \pi_0(\tilde{z} \mid x)\exp(q(x,\tilde{z})/\lambda)\, d\tilde{z}$。此外，对于每个提示 x，诱导分布 $Q_{R,\,x}$ 由 $Q_{0,\,x}$ 通过以下重新加权恒等式得到

$$\mathbb{E}_{Q_{R,x}}[\varphi(Q)] = \frac{\mathbb{E}_{Q_{0,x}}[\varphi(Q)\exp(Q/\lambda)]}{\mathbb{E}_{Q_{0,x}}[\exp(Q/\lambda)]} \qquad \textit{for all bounded measurable } \varphi : [0,1] \to \mathbb{R}. \tag{41}$$

因此，在 Q_R 下的无条件期望满足

$$\mathbb{E}_{Q_R}[\varphi(Q)] = \int \frac{\mathbb{E}_{Q_{0,x}}[\varphi(Q)\exp(Q/\lambda)]}{\mathbb{E}_{Q_{0,x}}[\exp(Q/\lambda)]}\, dP(x).$$

命题 5 表明，RLHF 是将基线策略向更高可接受性输出倾斜的一种方式。参数 λ 恰好发挥了 logit尺度的作用：当 $\lambda \downarrow 0$时，策略会集中于最高可接受性的q 补全结果，而较大的 λ 则使生成器保持接近 π_0。熵正则化问题的证明是标准的，为完整起见，附录B中给出了证明。

一个简单的推论是，RLHF 提高了高-q 输出的可能性。

引理 5（基于人类反馈的强化学习的一种支配性质）。 在假设*6*下，对于每个提示 x，倾斜后验分布 $Q_{R,\,x}$ 在一阶随机意义上支配 $Q_{0,\,x}$。因此，

[12]附录 B 提供了一个最小的微观基础，由二元可接受性变量生成这样的标量 $q(x, z)$。

无条件混合分布 Q_R 在一阶随机意义上支配 Q_0。特别地， $\mathbb{E}[Q_R] \geq \mathbb{E}[Q_0]$，除非 $Q_{0,\,x}$ 对于 P-几乎处处的 x 退化，否则不等式严格成立。

证明。 固定 x。方程（41）表明，$Q_{R,\,x}$ 由 $Q_{0,\,x}$ 通过指数倾斜得到，其似然比正比于 $\exp(Q/\lambda)$，且该比值关于 Q严格递增。这种单调似然比性质意味着，$Q_{R,\,x}$ 一阶随机占优于 $Q_{0,\,x}$ （附录 B 使用切比雪夫协方差不等式给出了直接证明。）

由于 Q_R 和 Q_0 是 $\{Q_{R,x}\}_x$ 和 $\{Q_{0,x}\}_x$ 在相同提示分布$P(x)$下的混合分布，因此混合后仍保留一阶随机占优：对于每个阈值 $t \in [0,1]$,

$$\Pr_{Q_R}(Q \geq t) = \int \Pr_{Q_{R,x}}(Q \geq t)\, dP(x) \;\geq\; \int \Pr_{Q_{0,x}}(Q \geq t)\, dP(x) = \Pr_{Q_0}(Q \geq t).$$

均值不等式由一阶随机占优推出。只有当倾斜在 P-几乎处处为平凡变换时，严格性才会失效，即$Q_{0,\,x}$ 对于 P-几乎每个 x都是退化的。 □

引理 5 已经在当前框架下回答了“基于人类反馈的强化学习是否属于偏好嵌入”这一基本问题。与保持 π_0 不变、再从其抽样结果中进行选择的分离式流程相比，基于人类反馈的强化学习会改变 生成器 ，以便在上游实现特定的奖励加权。剩下的问题是，何时这种改变是可取的，以及何时它会造成与定理 2所强调的相同类型的鲁棒性损失。

6.4 基于人类反馈的强化学习何时能提升福利?

基于人类反馈的强化学习对福利的主要影响，并非是在均值固定时进行 Blackwell 比较（如主要定理所述），因为基于人类反馈的强化学习会改变可接受性的水平。当部署目标稳定且与奖励一致时，这种均值变化是有益的。

定理4 （目标固定时的基于人类反馈的强化学习）。 保持假设 *6*。假设部署目标可以写成

$$W(\pi) = \mathbb{E}[V(Q_\pi)]$$

对于某个非递减的 $V\colon\ [0,1] \to \mathbb{R}$。那么，对于任意 $\lambda > 0$,

$$W(\pi_R) \geq W(\pi_0).$$

证明。 根据引理 5， Q_R 在一阶随机意义上占优于 Q_0。由于 V 是非递减的，$\mathbb{E}[V\ (Q_R)\ ] \geq \mathbb{E}[V\ (Q_0)\ ]$。 □

定理 4 阐明了为什么基于人类反馈的强化学习在实践中通常有效：当目标定义明确

且稳定（“在固定策略下遵循指令”）时，调整生成器以提高奖励的输出出现频率，会直接提升预期性能。

6.5 何时分离占主导?

当相关目标不是单一固定标量，而是一组权重可随上下文或时间变化的标准（有用性、无害性、真实性、风格、领域约束、监管规则等）时，分离逻辑便变得相关。在这种情况下，将一种固定的标量化嵌入生成器，可能会降低其适应其他标量化方式的灵活性。

形式化地说，假设每个补全结果都有一个逐准则可接受性后验向量$q(x,\ z) \in [0,1]^K$，且标量目标是线性泛函 $r_w(x,\ z) = w \cdot q(x,\ z)$，其权重为 $w \in \Delta^{K-1}$。基于单一 w 训练的 RLHF，会使生成器集中生成适合该标量化方式的补全结果；分离式设计则保留更丰富的选项，并在事后选择与w相对应的行动。

定理 5（目标不确定性下的分离）。 固定一个提示 x ，并假设基线生成器 $\pi_0(\cdot \mid x)$ 具有有限支撑 $\mathcal{Z}(x)$。对于每个补全结果 $z \in \mathcal{Z}(x)$，令$q(z) \in [0,1]^K$；对于任意权重向量 $w \in \Delta^{K-1}$，定义 $r_w(z) := w \cdot q(z)$。令 w 表示用于 *RLHF* 训练的权重，并定义由此产生的 *RLHF*策略

$$\pi_R^\lambda(z \mid x) = \frac{\pi_0(z \mid x)\exp(r_w(z)/\lambda)}{\sum_{z' \in \mathcal{Z}(x)} \pi_0(z' \mid x)\exp(r_w(z')/\lambda)}.$$

设部署权重 W 为取值范围为 $\mathcal{W} \subset \Delta^{K-1}$ 的随机变量，并假设 $\mathbb{P}(W = w') > 0$，其中 $w' \neq w$。假设存在补全结果 z^w、$z^{w'} \in \mathcal{Z}(x)$，使得 z^w 是 r_w 的严格最大化者，且 $z^{w'}$ 是 $r_{w'}$的严格最大化者，其中 $r_{w'}(z^{w'}) > r_{w'}(z^w)$。

考虑一种理想化的分离设计：在观察到 W后，选择 $z^\star(W) \in \arg\max_{z \in \mathcal{Z}(x)} r_W(z)$。那么，存在 $\lambda > 0$，使得对于所有 $\lambda < \lambda$,

$$\mathbb{E}_W\big[r_W(z^\star(W))\big] > \mathbb{E}_W\Big[\mathbb{E}_{Z \sim \pi_R^\lambda(\cdot|x)}\big[r_W(Z)\big]\Big].$$

证明。由于 $\mathcal{Z}(x)$是有限的，且 z^w 是 r_w的严格最大化者，因此对于每个 $z \neq z^w$ ，

$$\frac{\pi_R^\lambda(z \mid x)}{\pi_R^\lambda(z^w \mid x)} = \frac{\pi_0(z \mid x)}{\pi_0(z^w \mid x)} \exp\left(\frac{r_w(z) - r_w(z^w)}{\lambda}\right) \to 0 \quad \text{as } \lambda \downarrow 0.$$

因此 $\pi_R^\lambda(\cdot \mid x) \Rightarrow \delta_{z^w}$，因此 $\mathbb{E}_{Z \sim \pi_R^\lambda(\cdot|x)}[r_{w'}(Z)] \to r_{w'}(z^w)$。由于 $z^{w'}$ 是 r_w的严格极大值点，

并且当 λ 充分小时， $r_{w'}(z^{w'}) > r_{w'}(z^{w})$,

$$\mathbb{E}_{Z\sim\pi_R^{\lambda}(\cdot|x)}[r_{w'}(Z)] < r_{w'}(z^{\star}(w')).$$

取 对 W 取期望，并使用 $\mathbb{P}(W = w') > 0$ ，可得严格不等式 . □

定理 5 形式化地阐述了一个简单的想法：在目标不确定的情况下，将某一种特定标量化嵌入生成器，虽能提升该标量化下的表现，却会牺牲其他标量化下的菜单灵活性。这对应于预测器嵌入某种特定的非对称损失，而非学习校准概率并在事后进行选择所造成的福利损失。

定理5表明，经过 KL 正则化的基于人类反馈的强化学习会将基础生成器向训练奖励较高的补全结果进行指数倾斜 r_{w}。当正则化参数 λ 较小时，这种倾斜会变得高度集中：生成器的行为如同一个摊销优化器，用于 r_w ，并将几乎所有概率都赋予（某个）使 r_w最大化的补全结果。当评估采用相同的目标 r_w时，这样做是有益的，但也会损失期权价值：由于生成器集中于对 w最有利的内容，它对其他可能与部署相关的目标 w' 就不那么有用了。

定理 5 针对 目标不确定性 （随机部署权重 W），并将嵌入式基于人类反馈的强化学习生成器与理想化分离式设计进行比较；后者在观察到目标后选择最佳补全结果 。以下结果提取了最简单且最常见的特殊情形：部署使用一个单一 固定的替代目标 $r_{w'}$ 来评估补全结果，该目标不同于训练目标 r_{w}。量

$$\sup_{z\in\mathcal{Z}(x)} r_{w'}(z) - \mathbb{E}_{Z\sim\pi_R^{\lambda}(\cdot|x)}[r_{w'}(Z)]$$

可理解为从将 w 嵌入生成器所产生的选项价值损失：第一项表示系统若保留完整的补全结果“菜单”，再针对 w' 事后选出最佳补全结果，所能达到的效果；第二项则表示经基于人类反馈的强化学习训练的生成器在依据 w'评估时所能提供的效果。

推论 2 （固定替代目标下的能力税）。 在定理 5的设定中，固定 $w' \neq w$，并假设 $r_{w'}$（$z^{w'}$） $>$ $r_{w'}$（z^{w}）对于严格最大化点 $z^{w} \in \arg\max r_w$ 和$z^{w'} \in \arg\max r_{w'}$均成立。那么对于所有 λ 足够小的值，

$$\sup_{z\in\mathcal{Z}(x)} r_{w'}(z) - \mathbb{E}_{Z\sim\pi_R^{\lambda}(\cdot|x)}[r_{w'}(Z)] > 0,$$

且只有当 r_w 与 $r_{w'}$ 共享相同的最大化解时，差距才会消失。

证明。 应用定理 5 ， 其中 $W \equiv w'$。 □

当 λ较小时，KL 正则化的 RLHF 策略π_R^λ 会将概率质量集中在使 r_w最大化的补全结果上（此处由严格最大化解 z^w体现）。如果 w' 对另一个补全结果 $z^{w'}$的排名严格高于 z^w，即 $r_{w'}(z^{w'}) > r_{w'}(z^w)$，那么在替代目标下可实现的最佳值为 $\sup_z r_{w'}(z) = r_{w'}(z^{w'})$，而 RLHF 生成器的预期 $r_{w'}$值则会随着 λ 变小而趋近于 $r_{w'}(z^w)$。这会在λ足够小时造成严格为正的差距。要让这种“能力税”在小λ 条件下消失，唯一的办法是让这些目标在最优处对齐，也就是说， r_w 与 $r_{w'}$ 共享相同的最大化解，这样，集中于 r_w最优补全结果就不会排除任何 $r_{w'}$最优补全结果。-

在许多基于人类反馈的强化学习部署中，“奖励”并不是一个原始的标量，而是多个标准（有用性、无害性、真实性、风格、领域约束等）的标量化 ，这里用 $q^1, \ldots, q^K$表示。训练会选择一组权重 w，将这些标准转化为单一目标 r_w。但在部署时，相关权衡可能因用户、上下文或时间而异，对应于不同的权重 w'。接下来的结果将能力税的概念转化为多标准框架：将一种权重嵌入生成器，可以提升该权重下的表现，但代价是在采用替代权重时灵活性降低。

推论 3 （跨标准的能力税）。 假设 $q(z) = (q^1(z) \ldots, q^K(z))$，且基于人类反馈的强化学习针对权重 $w \in \Delta^{K-1}$，在 $r_w(z) = \sum_k w_k q^k(z)$ 上进行训练。固定一种替代权重 $w' \neq w$。如果 r_w 和 $r_{w'}$ 的最大化者不同，那么当 λ 足够小时，基于人类反馈的强化学习生成器在替代标量化 $r_{w'}$ 下严格次优于一种在事后根据 w' 选择补全结果的设计。

证明。 这只是推论 2 的直接重述，其中 $r_{w'}(z) = \sum_k w'_k q^k(z)$。 □

如果 r_w 和 $r_{w'}$ 的最大化者不同，那么存在一个补全结果，对于替代权衡 w' 而言是最优的， w但通过优化 λ并不会选中它。当 w较小时，基于人类反馈的强化学习生成器会集中生成对于 w'而言最优的补全结果，因此在 w'下评估时会严格次优。相比之下，将标量化选择推迟到部署阶段的分离式设计，可以在事后选择 $_z r_{w'}(z$ 的最大化者，因此达到 $\sup(r_{w'})$。形式上，这正是将线性标量化 $z(= \sum_k w'_k q^k)$ z代入推论 2 所得到的结果。

6.6 奖励设定错误

到目前为止，分析都假设奖励模型与真实可接受性后验概率$q(x, z)$一致。实际上， $r(x, z)$是根据有限的偏好数据估计得出的，因此只是一个不完美的代理指标。当代理指标成为优化目标时，模型设定错误便会被放大，这正是

古德哈特定律所描述的熟悉情形（“当衡量指标变成目标时，它就不再是一个好的衡量指标”）：选择压力会将生成器推向输出空间中代理指标较高的区域，而这些区域未必与真实可接受性较高的区域重合。

一种便于使用的约化形式是将奖励分解为真实成分和伪相关成分。

假设 7（错误设定的奖励）。对每个 (x, z)，

$$r(x, z) = \alpha\, q(x, z) + (1 - \alpha)\, s(x, z), \tag{42}$$

其中，$\alpha \in [0, 1]$ 衡量对齐，而 $s(x, z)$ 是一种伪相关特征，它在偏好数据中会获得奖励，但在部署时与可接受性并无内在联系。

命题 6（基于人类反馈的强化学习下的古德哈特放大效应）。在假设 7下，*RLHF*策略满足

$$\pi_R(z \mid x) \;\propto\; \pi_0(z \mid x) \exp\left(\frac{\alpha q(x, z) + (1 - \alpha) s(x, z)}{\lambda}\right).$$

固定一个提示 x。令 $Q_{0,x}$ 表示在 $Z \sim \pi_0(\cdot \mid x)$ 下 $q(x, Z)$ 的分布律，并令 $Q_{R,x}$ 表示在 $Z \sim \pi_R(\cdot \mid x)$下的分布律。那么，对所有有界可测的 φ，

$$\mathbb{E}_{Q_{R,x}}[\varphi(Q)] = \frac{\mathbb{E}_{Q_{0,x}}[\varphi(Q) \exp(\alpha Q/\lambda)\, m_{\lambda,x}(Q)]}{\mathbb{E}_{Q_{0,x}}[\exp(\alpha Q/\lambda)\, m_{\lambda,x}(Q)]}, \qquad m_{\lambda,x}(q) := \mathbb{E}[\exp((1 - \alpha)S/\lambda) \mid Q = q]\,, \tag{43}$$

其中，定义 $m_{\lambda,x}$ 的条件期望是在 $Z \sim \pi_0(\cdot \mid x)$ 下取的，且有 $Q = q(x, Z)$ 和 $S = s(x, Z)$。

命题6 明确揭示了放大效应的作用路径。相较于对齐基准（其中倾斜仅为 $\exp(Q/\lambda)$），模型设定错误会引入一个额外的倾斜因子 $m_{\lambda,x}(Q)$，该因子取决于虚假成分在给定真实可接受性时如何变化（通常还取决于提示 x）。优化越激进（λ越小），这个额外因子就越可能主导选择。尤其是，如果偏好数据环境与部署环境中的联合分布 (Q, S) 不同（这是一种常见的分布偏移），那么优化 r 就可能将概率质量推向S 较高、但在部署环境中实际可接受性并不高的输出。附录B 给出了一个简单的两点示例：尽管奖励在训练中看似具有预测能力，基于人类反馈的强化学习却降低了真实接受率。

6.7 实践启示

基于人类反馈的强化学习的结果带来了设计指导，这些指导与论文前文关于嵌入与分离的诊断方法相呼应，只不过此时生成器扮演了被嵌入对象的角色。理解基于人类反馈的强化学习的一种有效方式，是将其视为一种摊销选择：它把原本的推理时选择问题（采样多个候选项并从中选择）转移到了训练阶段。当部署时的选择成本高昂，或受到延迟、计算预算或可采样数量有限等因素制约时，这种做法尤其有价值。在这些情况下，将偏好嵌入生成器，可能优于名义上更优、但用户实际上无法执行的事后规则。

与此同时，当目标随时间发生变化，或因用户和上下文而异时，分离逻辑就更具说服力。定理 5 表明，将单一标量化硬编码到生成器中会带来能力税：随着优化力度加大，生成器会集中于训练目标下的最优选择，并牺牲其他目标下的期权价值。模块化架构保留了能力广泛的基础生成器，并通过重排序、阈值、过滤器或路由来实现针对特定目标的偏好，从而保留这种选择灵活性，因此面对目标变化时也更加稳健。

当后处理不可行而无法避免嵌入时，分析表明，采取更保守的优化立场是一种稳健性选择。具体而言，较大的 λ会限制 RLHF策略的集中程度，因此既能限制目标不匹配时的能力税，也能抑制奖励设定错误的放大效应。最后，明确系统中究竟训练哪个组件也很有帮助。在大语言模型部署中，前文讨论的严格应用涉及评分器和路由策略，而基于人类反馈的强化学习则会修改生成器本身。混淆这两个层面可能会掩盖可用的设计自由度：通常可以用弱偏好或无偏好的方式训练评分器，同时仍通过阈值和路由实施严格的部署策略；只有在推理时无法可靠地完成选择时，才需要在生成器层面进行嵌入。

这一视角也阐明了基于人类反馈的强化学习分析与论文关于嵌入与分离的主要结果之间的关系。前文在存在学习摩擦的情况下比较预测器，而基于人类反馈的强化学习会直接改变生成器的输出分布，并非保持均值不变的变换。由于基于人类反馈的强化学习有意改变输出分布（从而改变可接受性的均值水平），定理4自然表现为一阶随机占优结果，而非凸序比较。在大语言模型场景中，与后处理摩擦相对应的是推理时的搜索成本：当延迟限制严格、只能生成少量样本时，嵌入可能占优，因为最优的事后选择规则无法实现；当后处理可行且目标可能发生变化时，

分离逻辑便再次占据主导，倾向于采用具有丰富基础分布并进行事后选择的方案。

7 结论

本文研究了支持经济决策的人工智能管道中一种简单却普遍存在的架构选择：是将下游偏好嵌入训练目标，还是训练一个无偏好预测器，仅通过后处理和决策规则来实施偏好。本文的核心贡献，是在学习内生的情形下，精确阐明“先学习，再选择”这一理念。将训练建模为对贝叶斯可行后验分布的选择，使现代机器学习实践与信息设计建立了联系。在信息价值递减这一性质下，偏好嵌入会压平信息量进一步提升所带来的边际回报。若最优后验满足一个温和的可比性条件，就能推出保均值收缩：相较于无偏好训练，偏好嵌入训练会产生离散程度更低的后验信念分布。由于期望效用决策问题中的间接效用关于信念是凸的，这种收缩意味着稳健的福利比较：对于一大类下游决策问题，无偏好流程至少同样优越；通过贝叶斯最优决策规则在事后应用偏好，即可实现其福利水平。

这些结果有助于调和两种常见的实践建议，而这两种建议原本似乎存在矛盾。一方面，实践者常会采用代价敏感学习、类别重加权或“效用加权”目标，以反映假阳性与假阴性之间的不对称性。另一方面，实践者也强调校准、适当评分规则和阈值处理，将其作为部署工具。这项分析阐明了何时应优先采取后一种做法：如果能够在部署时可靠地实施所需策略，那么通过训练来学习校准概率，就能在目标不确定、存在争议或不断变化时保留期权价值。在此类环境中，将某一特定的权衡嵌入损失函数，可能会破坏之后无法恢复的信息，即使最终评估采用的正是这个嵌入式目标。

这种分离逻辑与算法公平性文献有着天然的联系（Ram- bachan 等人，2020年；Liang 等人，2022年；Liang 和 Lu，2024年）。公平性标准是目标的典型例子，会因司法辖区、组织和时间而异，并且常常会根据新证据或新的规范性约束进行修订。许多公平性干预可以视为在给定评分的情况下调整决策规则（例如，群体感知阈值、约束选择或程序要求），而底层预测任务仍然是“估计风险”或“估计可接受性”。本文的结果为维持经过校准、无偏好的预测核心提供了清晰的经济学依据

，并将公平性作为透明的后处理层来实施，以便在不重新训练预测器的情况下对其进行审计、更新和讨论。同时，这项分析也阐明了为何这一建议并非放之四海而皆准：当后处理成本高昂或不可靠时，上游嵌入可以通过压缩信息、减轻使用概率预测所带来的认知和制度负担来提升福利。何时应将公平性约束作为训练时嵌入来实施，何时应将其作为部署时策略来实施，仍是一个重要的应用设计问题；本文的框架表明，答案应取决于下游实施的可行性和成本，而不仅仅取决于公平性标准本身。

在人工智能对齐与安全领域，同样的主题也会出现，但利害关系更大。在大型语言模型部署中，许多安全与质量控制措施已经采用二元可接受性评分和路由的形式：系统会为某个补全结果被接受的可能性赋予一个概率，然后决定是展示、拒绝、重新生成，还是升级处理。对于这些模块，分离原理有着字面意义：使用严格适当损失训练评分器，以估计可接受性概率，并通过阈值和路由来编码政策权衡。更具争议的是，基于人类反馈的强化学习等端到端对齐方法，可以被解读为生成器侧偏好嵌入。当目标稳定且奖励设定明确时，嵌入可以通过改变输出分布来提高可接受性水平。但当目标具有多个维度，或在不同上下文中发生变化时，将单一标量化目标嵌入生成器，可能会牺牲其依据其他标准灵活提供不同选项的能力，从而造成能力税；而激进的优化则可能以古德哈特定律所描述的方式放大奖励设定错误。这些观察表明，在经济学与对齐的交界处存在一个颇有前景的研究方向：刻画以下部分之间的最优分工：（i）能力广泛的基础模型；（ii）用于编码政策与安全约束的模块化后处理层；以及（iii）为满足延迟或可用性要求而确有必要进行的任何生成器级微调。

若干扩展可以进一步完善并拓宽本文结论。一个方向是将该框架与公平性约束更直接地联系起来：把公平性目标视为一种目标不确定性，并研究哪些类别的公平性政策能够在不损失信息的情况下作为事后决策规则实施。第二个方向是更全面地放宽标量信念限制：多类别和多准则情境需要更丰富的信息序，也会带来新的选择问题；但公平性与对齐方面的关切通常也存在于这些情境中（见在线附录 A）。第三个方向是动态的：目标会演变，模型会更新，治理干预也会随时间发生。若能形式化说明模块化设计如何在这种演变过程中保留期权价值，以及模块化的成本在何种情况下足以证明嵌入的合理性，就能使分离原理在现实

机构中发挥实际作用。

综合来看，这些结果支持一项严谨的设计准则：在可行的情况下，应构建能够保留信息并在选择点进行判断的通用能力；只有当下游实施摩擦使分离难以实行时，才将偏好嵌入其中。

参考文献

Nikhil Agarwal、Alex Moehring、Pranav Rajpurkar 和 Tobias Salz，《将人类专业知识与人工智能相结合：来自放射学的实验证据》，*NBER* 工作论文，2023年，（31422）。

Ajay Agrawal、Joshua S. Gans 和 Avi Goldfarb，《预测、判断与复杂性：决策与人工智能理论》，载于《人工智能经济学：议程》，芝加哥大学出版社，2018年，第89–110页。

_ 、_、和 ，_《预测机器：人工智能的简明经济学》，第2版，哈佛商业出版社，2022年。

David Autor、Andrew Caplin、Daniel J. Martin 和 Philip Marx，《设计失调：机器学习中的激励失灵》，工作论文第34504号，美国国家经济研究局，2025年11月。

Andrew Caplin，《认知经济学导论》，帕尔格雷夫·麦克米伦，2025年。

_ 、**Daniel J. Martin 和 Philip Marx**，《机器学习建模：一种认知经济学方法》，工作论文第30600号，美国国家经济研究局，2022年。

_ ，**David Deming、Shangwen Li 和 Kadachi Jiada Ye**，《与人工智能共事的受益者：能力、信念与校准》，工作论文，美国国家经济研究局，2024年。

Gans, Joshua S.，《人工智能的微观经济学》，麻省理工学院出版社，2025年。

Gneiting, Tilmann 和 Adrian E. Raftery，《严格适当评分规则、预测与估计》，《美国统计学会杂志》，2007年，*102* (477)，359–378。

Granger, Clive W. J. 和 Mark J. Machina，《预测与决策理论》，载于 Graham Elliott、Clive W. J. Granger 和 Allan Timmermann 编，《经济预测手册》，第1卷，爱思唯尔，2006年，第81–98页。

Kamenica, Emir 和 Matthew Gentzkow，《贝叶斯劝说》，《美国经济评论》，2011年，*101*(6)，2590–2615。

Liang, Annie 和 Jay Lu，《算法公平性与社会福利》，载于《美国经济学会论文与会议录》，第114卷，美国经济学会，2014 Broadway，305室，纳什维尔，田纳西州37203，2024年，第628–632页。

_ , _和 **Xiaosheng Mu**, 《算法设计：公平性与准确性》，载于《第23届 ACM 经济学与计算会议论文集》，2022年，第58–59页。

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray 等, 《通过人类反馈训练语言模型遵循指令》，《神经信息处理系统进展》，2022年，*35*，27730–27744。

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning 和 Chelsea Finn, 《直接偏好优化：你的语言模型其实是一个奖励模型》，《神经信息处理系统进展》，2023年，*36*。

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig 和 Sendhil Mullainathan, 《算法公平性的经济学视角》，《美国经济学会论文与会议录》，2020年，*110*，91–95。

Sims, Christopher A., 《理性注意力的影响》，《货币经济学杂志》，2003年，*50* (3)，665–690。

Stiennon、Nisan、Long Ouyang、Jeffrey Wu、Daniel M. Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul Christiano, 《通过人类反馈学习生成摘要》，载于《神经信息处理系统进展》第33卷，2020年。

Strack、Philipp 和 Kaihao Yang, 《隐私保护信号》，《计量经济学》，2024年11月，*92*(6)，1907–1938。

Heliodoro Tejeda、Aditya Kumar、Padhraic Smyth 和 Mark Steyvers, 《医疗保健中的人工智能辅助决策：算法建议对临床判断的影响》，*AAA*人工计算与众包会议论文集，2022年，*10*，172–182。

Hal R Varian, 《优化模型中的拟合优度》，《计量经济学杂志》，1990年，*46*(1-2)，125–140。

Zellner，Arnold, “使用非对称损失函数进行贝叶斯估计和预测”，《美国统计学会杂志》，1986年，*81*(394)，446–451。

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving, 《基于人类偏好对语言模型进行微调》，*CoRR*，2019年，*abs/1909.08593*。

A 在线附录：多元凸序扩展

本附录概述了定理1 和2中的收缩与分离逻辑如何推广到二元结果之外。正文聚焦于标量后验，因为$[0,1]$ 上的凸序相对容易处理，而且许多经济预测任务本质上是标量的（风险评分）。在多类别或多状态环境中，自然的信念对象是后验概率的向量。标量凸序恰当的对应概念是单纯形上的（向量）凸序，它仍能为所有期望效用决策问题提供稳健的福利比较。主要局限在于可比性：在维度大于一时，凸序是一种偏序，“选择”问题也更为严重。

A.1 信念与贝叶斯可行后验概率分布

设结果／状态空间为有限集，

$$\mathcal{Y} \coloneqq \{1, \dots, K\}, \qquad K \geq 2,$$

并令 $\Delta(\mathcal{Y})$ 表示 K 单纯形：

$$\Delta(\mathcal{Y}) \coloneqq \left\{ q \in \mathbb{R}_+^K : \sum_{y \in \mathcal{Y}} q_y = 1 \right\}.$$

（向量）后验信念为 $q \in \Delta(\mathcal{Y})$，其解释为 $q_y = \mathbb{P}(Y = y \mid \cdot)$。设先验为$\mu \in \Delta(\mathcal{Y})$。

信号 S 诱导出一个随机后验向量

$$Q \coloneqq \big(\mathbb{P}(Y = y \mid S)\big)_{y \in \mathcal{Y}} \in \Delta(\mathcal{Y}),$$

而贝叶斯可行性意味着 $\mathbb{E}[Q] = \mu$ （逐分量）。

定义 6 （单纯形上的贝叶斯可行后验分布）。设 $\mathcal{Q}^K(\mu)$ 表示取值于 $\Delta(\mathcal{Y})$ 的随机向量 Q 的集合，使得 $\mathbb{E}[Q] = \mu$。

A.2 多元凸序

定义 7 （ $\Delta(\mathcal{Y})$上的凸序）。 对于 Q， $Q' \in \mathcal{Q}^K(\mu)$，记 $Q \succeq_{\text{cx}} Q'$如果

$$\mathbb{E}[\varphi(Q)] \geq \mathbb{E}[\varphi(Q')] \quad \text{for all convex } \varphi : \Delta(\mathcal{Y}) \to \mathbb{R}.$$

等价地， Q' 是 Q的保均值收缩，即存在一个耦合（$\widetilde{Q}$， $\widetilde{Q}'$），满足 $\widetilde{Q} \overset{d}{=} Q$、 $\widetilde{Q}' \overset{d}{=} Q'$ 和 $\mathbb{E}[\widetilde{Q} \mid \widetilde{Q}'] = \widetilde{Q}'$。

与标量情形一样， $Q \succeq_{\mathrm{cx}} Q'$ 表示 Q 比 Q' 更分散（因而信息量更大），同时保持均值 μ 不变。

A.3 向量信念下的决策问题与信息价值

决策问题由一组可行行动 $\mathcal{A}$ 和一个收益函数 u： $\mathcal{A} \times \mathcal{Y} \to \mathbb{R}$构成。给定信念 $q \in \Delta(\mathcal{Y})$，决策者的间接价值为

$$V(q) \coloneqq \sup_{a \in \mathcal{A}} \sum_{y \in \mathcal{Y}} q_y \, u(a, y). \tag{44}$$

引理 6 (单纯形上间接价值的凸性)。 对于任意决策问题$(\mathcal{A}, u)$，函数 $V \colon \Delta(\mathcal{Y}) \to \mathbb{R}$ 在 (44) 中定义，且为凸函数。

证明。 对于每个固定行动 a，映射 $q \mapsto \sum_y q_y u(a, y)$关于 q是线性的。线性函数的逐点上确界是凸函数。

□

引理 7 (信息量更大的后验分布会提高期望价值)。 若 $Q \succeq_{\mathrm{cx}} Q'$，则对于任意决策问题$(\mathcal{A}, u)$，

$$\mathbb{E}[V(Q)] \geq \mathbb{E}[V(Q')].$$

证明。 根据引理 6， V在 $\Delta(\mathcal{Y})$ 上是凸的。根据定义 7，凸序意味着 $\mathbb{E}[V(Q)] \geq \mathbb{E}[V(Q')]$。

□

A.4 多变量收缩与分离

为使扩展与正文保持平行，假设使用以 t 为索引的目标进行训练，会产生后验分布 $Q_t \in \mathcal{Q}^K(\mu)$，该分布解决约化形式学习问题

$$Q_t \in \arg\min_{Q \in \mathcal{Q}^K(\mu)} \left\{ \mathbb{E}[H_t(Q)] + C(Q) \right\}, \tag{45}$$

其中， $H_t \colon \Delta(\mathcal{Y}) \to \mathbb{R}$ 是与以 t 为索引的训练目标相关联的贝叶斯风险，而C 是一个分布不变的学习摩擦泛函。

假设 3 的多变量类比是同一“递增差异”条件，现在则针对单纯形上的凸序来表述。

假设8 （ $\Delta(\mathcal{Y})$上的信息价值递减）。 对于任意 $t_1 > t_0$ 和任意Q， $Q' \in \mathcal{Q}^K(\mu)$，且满足 $Q \succeq_{\text{cx}} Q'$，

$$\mathbb{E}[H_{t_1}(Q) - H_{t_0}(Q)] \geq \mathbb{E}[H_{t_1}(Q') - H_{t_0}(Q')]\,. \tag{46}$$

此外，只要 $Q \succeq_{\text{cx}} Q'$ 且 $Q \overset{d}{\neq} Q'$，不等式 (46) 就是严格的。

一个方便的充分条件与命题 3中的相同：若 $H_t(q) = H_0(q) + t\,h(q)$，对于某个凸函数 $h : \Delta(\mathcal{Y}) \to \mathbb{R}$，则假设 8 成立（若$h$ 严格凸，则不等式严格）。当 H_t 在单纯形的相对内部二次可微时，可通过要求其海森矩阵在切空间上半正定来检验 h 的凸性 $\{v \in \mathbb{R}^K : \sum_y v_y = 0\}$。

假设 4 的多变量类比才是真正棘手之处。

假设 9 （最优后验的可比性）。 对于任意 $t_1 > t_0$，以及任意极小化者 Q_{t_1} 和 Q_{t_0}，若它们是 (45) 的极小化者，则这对 (Q_{t_0}, Q_{t_1}) 在凸序下可比：即 $Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$或 $Q_{t_1} \succeq_{\text{cx}} Q_{t_0}$。

定理 6 （偏好嵌入下的多变量收缩）。 假设 *8*和 *9* 成立。令 $\{Q_t\}_{t\in[0,1]}$ 为 (45)的任意极小化解选择。如果 $t_1 > t_0$，那么

$$Q_{t_0} \succeq_{\text{cx}} Q_{t_1}.$$

证明。 固定 $t_1 > t_0$，令 Q_{t_1} 和 Q_{t_0} 分别为 t_1 和 t_0 处的极小化解。由最优性可得

$$\mathbb{E}[H_{t_1}(Q_{t_1})] + C(Q_{t_1}) \leq \mathbb{E}[H_{t_1}(Q_{t_0})] + C(Q_{t_0}),$$

and

$$\mathbb{E}[H_{t_0}(Q_{t_0})] + C(Q_{t_0}) \leq \mathbb{E}[H_{t_0}(Q_{t_1})] + C(Q_{t_1}).$$

将两个不等式相加后，成本项相互抵消，得到

$$\mathbb{E}[H_{t_1}(Q_{t_1}) - H_{t_0}(Q_{t_1})] \leq \mathbb{E}[H_{t_1}(Q_{t_0}) - H_{t_0}(Q_{t_0})]\,. \tag{47}$$

根据假设 9，要么 $Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ （得证），要么 $Q_{t_1} \succeq_{\text{cx}} Q_{t_0}$。后一种情况下，假设 8蕴含与（47）相反的不等式，且除非 $Q_{t_1} \overset{d}{=} Q_{t_0}$，否则是不严格的。因此，（47）只有在 $Q_{t_1} \overset{d}{=} Q_{t_0}$时才能成立，此时 $Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ 也成立。因此 $Q_{t_0} \succeq_{\text{cx}} Q_{t_1}$ 在所有情况下均成立。 □

推论 4 （多变量稳健分离原理）。 在定理 *6*的假设下，设 Q_0 和 Q_1 分别表示由无偏好训练（$t = 0$）所诱导的后验分布

以及由嵌入偏好的训练（$t = 1$）所诱导的后验分布。那么，对于任意决策问题（$\mathcal{A}, u$），

$$\mathbb{E}[V(Q_0)] \geq \mathbb{E}[V(Q_1)].$$

证明。 定理 6 给出 $Q_0 \succeq_{\text{cx}} Q_1$。引理 7 进而推出 $\mathbb{E}[V(Q_0)] \geq \mathbb{E}[V(Q_1)]$。 □

A.5 局限性：这种扩展并不能带来什么

这种多变量扩展在数学上简洁明了，但在经济学上却令人警醒。

(i) **凸序在 $K > 2$ 中只是一个偏序**。 在更高维度下，两个符合贝叶斯条件的后验分布往往无法比较。因此，假设 9 比其标量类比强得多，应将其理解为对可行信息性前沿的单指标限制（最优解沿着一条信息弱化／细化链移动）。

(ii) **检验信息价值递减更为困难**。 在一维情形下，假设 3 可简化为标量增量的凸性。在单纯形上，它则变成沿切向方向施加的矩阵（海森矩阵）限制。原则上，这仍然可以检验，但它并不是那种“看看二阶导数就心满意足”的条件。

(iii) **稳健性是有代价的**。 推论 4 中的结论对所有决策问题都是统一成立的，因为它依赖于完整的凸序。如果愿意限制下游目标函数的类别，则可以使用较弱的信息序。例如，如果下游价值仅取决于一个一维指标 $w \cdot q$（即信念标量化），那么正文中的标量理论也适用于 $w \cdot Q$，即使 Q 本身不具有凸序可比性。

(iv) **端到端序列生成仍未被“涵盖”**。 尽管语言模型的状态空间可以表示为一个巨大的单纯形，但定理 6 背后的可比性要求对于此类对象而言通常并不合理。第 6.1节讨论了严格二元可接受性设定下相关的大语言模型解释。

简言之：多变量结果表明，分离原理并非天生局限于“一维”，但也阐明了正文中对二元／标量情形的限制确实发挥着重要作用。障碍不在于间接价值的凸性（它可以直接推广）；而在于维度大于一时，能否按照一种易于处理的信息量顺序对学习结果进行排序。

B 在线附录：第 6 节的技术细节

本附录汇集了第 6 节中省略的推导和证明。正文将基于人类反馈的强化学习讨论聚焦于设计思路（嵌入还是后处理）；以下结果虽属标准内容，但仍予以列出，以求完整，并使映射过程自成一体。

B.1 可接受性的最简微观基础

固定一个提示 x。令 $\theta \in \Theta$ 表示潜在的"真实情况"或用户意图，其先验为 $p(\theta \mid x)$。对于每个补全 z，定义一个二元可接受性指标

$$Y = \mathbf{1}\{z \text{ is acceptable given } \theta\} \in \{0, 1\}.$$

与补全 z 相关的（贝叶斯）可接受性后验为

$$q(x, z) \equiv \mathbb{P}(Y = 1 \mid x, z) = \mathbb{E}[Y \mid x, z] \in [0, 1].$$

对于任意生成器 $\pi(\cdot \mid x)$，定义可接受性后验的诱导分布 $Q_{\pi,x}$ 为当 $q(x, Z)$ 服从 $Z \sim \pi(\cdot \mid x$ 时的分布律。在提示分布 $P(x$ 下，第 6 节中使用的无条件分布 是这些条件分布律的混合分布。

与大语言模型部署相匹配的一个简单下游决策问题是接受/重新生成选择。令 $a \in \{$ 表示接受，,重新生成$\}$，收益为

$$u(a, Y) = \begin{cases} Y - c(1 - Y) & \text{if } a = \text{accept}, \\ -d & \text{if } a = \text{regenerate}, \end{cases} \qquad c > 0,\ d \geq 0.$$

给定信念 q，间接价值为

$$V(q) = \max\{q(1 + c) - c,\ -d\},$$

它在 $[0, 1]$ 上是凸且非递减的。因此，生成器 π 的期望价值可写为

$$W(\pi) = \mathbb{E}[V(Q_\pi)].$$

这为定理中使用的表征提供了一个具体的微观基础4。许多其他大语言模型部署目标（阈值处理、路由、best-of-n）也具有相同的定性性质：价值随可接受性后验概率单调不减。

B.2 命题 5 的证明5

固定一个提示 x 并将其省略。根据假设6，基于人类反馈的强化学习问题（39）在 x 上逐点化简为

$$\max_{\pi(\cdot)} \int q(z)\,\pi(z)\,dz \;-\; \lambda \int \pi(z)\log\!\left(\frac{\pi(z)}{\pi_0(z)}\right) dz \quad \text{s.t.} \quad \int \pi(z)\,dz = 1.$$

在概率约束上引入乘子 η，构造拉格朗日函数：

$$\mathcal{L}(\pi,\eta) = \int \Big(q(z)\pi(z) - \lambda\pi(z)\log(\pi(z)/\pi_0(z))\Big)\,dz + \eta\Big(\int \pi(z)\,dz - 1\Big).$$

关于 $\pi(z)$ 的一阶条件给出

$$q(z) - \lambda\Big(1 + \log(\pi(z)/\pi_0(z))\Big) + \eta = 0,$$

so

$$\pi(z) = \pi_0(z)\exp\!\left(\frac{q(z)+\eta-\lambda}{\lambda}\right) \;\propto\; \pi_0(z)\exp(q(z)/\lambda).$$

归一化后得到 (40)。

为得到 (41)，任取有界可测的 $\varphi:\ [0,1]\to\mathbb{R}$ 并写出 $Q = q(Z)$ 在$Z\sim\pi_0(\cdot\mid x)$ 下的形式。令 $Q_{0,x}$ 表示 Q 在 $Z\sim\pi_0(\cdot\mid x)$ 下的分布律。利用 (40)，

$$\mathbb{E}_{Z\sim\pi_R(\cdot|x)}[\varphi(q(Z))] = \frac{\mathbb{E}_{Z\sim\pi_0(\cdot|x)}[\varphi(q(Z))\exp(q(Z)/\lambda)]}{\mathbb{E}_{Z\sim\pi_0(\cdot|x)}[\exp(q(Z)/\lambda)]} = \frac{\mathbb{E}_{Q_{0,x}}[\varphi(Q)\exp(Q/\lambda)]}{\mathbb{E}_{Q_{0,x}}[\exp(Q/\lambda)]}.$$

这就是 (41)。对条件恒等式关于 $x\sim P(x)$ 积分，得到命题5中所述的无条件混合表示。

B.3 引理5的证明

固定一个提示 x 并将其省略。令 $Q_{0,\,x}$ 和 $Q_{R,\,x}$ 分别表示在 $Z\sim\pi_0(\cdot\mid x)$ 和 $Z\sim\pi_R(\cdot\mid x)$ 下 $q(Z)$ 所诱导的分布律。

令 $g(Q):=\exp(Q/\lambda)$，该函数单调递增。对于任意阈值 $t\in[0,1]$，将 (41) 与$\varphi(Q) = \mathbf{1}\{Q\ge t\}$:

$$\mathbb{P}_{Q_{R,x}}(Q\ge t) = \frac{\mathbb{E}_{Q_{0,x}}[\mathbf{1}\{Q\ge t\}g(Q)]}{\mathbb{E}_{Q_{0,x}}[g(Q)]}.$$

由于 $\mathbf{1}\{Q\ge t\}$ 和 $g(Q)$共单调，切比雪夫协方差不等式推出$\mathbb{E}[\mathbf{1}\{Q\ge t\}g(Q)] \ge \mathbb{E}[\mathbf{1}\{Q\ge t\}]\mathbb{E}[g(Q)]$, 因此$\mathbb{P}_{Q_{R,\,x}}(Q\ge t)\ge \mathbb{P}_{Q_{0,\,x}}(Q\ge t)$。对于每个固定的 x，这是一阶随机占优。将同一恒等式应用于 $\varphi(Q) =$

Q 得到 $\mathbb{E}[Q_{R,x}] \geq \mathbb{E}[Q_{0,x}]$，除非$Q_{0,\ x}$ 是退化的，否则不等式严格成立（因为 g 严格递增）。

引理 5中的无条件支配结论 $x \sim P(x)$可通过对提示进行混合得到，正如正文中的证明所示。

B.4 命题6的证明

Fix x 并将其抑制。在假设7下，逐点基于人类反馈的强化学习解为

$$\pi_R(z \mid x) \ \propto \ \pi_0(z \mid x) \exp\left(\frac{\alpha q(z) + (1-\alpha)s(z)}{\lambda}\right).$$

令 (Q, S) 表示在 $Z \sim \pi_0(\cdot \mid x)$ 下的配对 $(q(Z),\ s(Z))$，并将 $Q_{0,x}$ 记为 Q 的边缘分布律。对于任意有界可测的 φ,

$$\begin{aligned}\mathbb{E}_{Q_{R,x}}[\varphi(Q)] &= \frac{\mathbb{E}_{Z\sim\pi_0(\cdot|x)}\left[\varphi(q(Z))\exp\left(\frac{\alpha q(Z)+(1-\alpha)s(Z)}{\lambda}\right)\right]}{\mathbb{E}_{Z\sim\pi_0(\cdot|x)}\left[\exp\left(\frac{\alpha q(Z)+(1-\alpha)s(Z)}{\lambda}\right)\right]} \\ &= \frac{\mathbb{E}_{Q_{0,x}}[\varphi(Q)\exp(\alpha Q/\lambda)\,\mathbb{E}[\exp((1-\alpha)S/\lambda) \mid Q]]}{\mathbb{E}_{Q_{0,x}}[\exp(\alpha Q/\lambda)\,\mathbb{E}[\exp((1-\alpha)S/\lambda) \mid Q]]},\end{aligned}$$

这就是带有 $m_{\lambda,x}(q) = \mathbb{E}[\exp((1-\alpha)S/\lambda) \mid Q = q]$ 的（43）。

B.5 一个简单的古德哈特示例：优化会选中罕见的“投机输出”

这个例子说明了基于人类反馈的强化学习可能放大模型设定错误的原因：即使奖励模型在基线分布上看起来颇具信息性，激进的优化也可能集中于代理指标很高、但真实可接受性很低的输出。

固定一个提示 x，并考虑一个有限支撑为 $\mathcal{Z}(x) = \{z_1, z_2, z_3\}$、概率为 $\pi_0(z_1) = 0.49$, $\pi_0(z_2) = 0.49$, $\pi_0(z_3) = 0.02$ 的基线生成器。设真实可接受性和虚假分数为

$$(q_1, q_2, q_3) = (0.9,\ 0.6,\ 0.4), \qquad (s_1, s_2, s_3) = (0.2,\ 0.3,\ 1.0),$$

并设（存在模型设定错误的）奖励为 $r = \alpha q + (1-\alpha)s$，其中 $\alpha = 0.6$。那么

$$(r_1, r_2, r_3) = (0.62,\ 0.48,\ 0.64),$$

因此，尽管 z_3 在两个“合理”输出中可接受性最低（z_1，z_2），但 z_3 获得了最高奖励。

基线可接受性。在 π_0下，预期可接受性为

$$\mathbb{E}_{\pi_0}[q(Z)] = 0.49 \cdot 0.9 + 0.49 \cdot 0.6 + 0.02 \cdot 0.4 = 0.743.$$

在基线分布中，奖励仍与真实可接受性正相关： z_1 的 q 和 r 都高于 z_2，而 z_3 很罕见。（形式上，在 π_0下，$\mathrm{Cov}(q(Z)\ r(Z)) > 0$。）

基于人类反馈的强化学习选择。在基于人类反馈的强化学习中，$\pi_R^\lambda(z) \propto \pi_0(z)\exp(r(z)/\lambda)$。由于 $r_3 > r_1 > r_2$，我们有$\pi_R^\lambda \Rightarrow \delta_{z_3}$，其中 $\lambda \downarrow 0$。因此

$$\lim_{\lambda \downarrow 0} \mathbb{E}_{\pi_R^\lambda}[q(Z)] = q_3 = 0.4 < 0.743 = \mathbb{E}_{\pi_0}[q(Z)].$$

[-] [-]由连续性可知，存在 $\lambda > 0$，使得对所有 $\lambda < \lambda$， $\mathbb{E}_{\pi_R^\lambda}[q(Z)] < \mathbb{E}_{\pi_0}[q(Z)]$。

这是标准“奖励劫持”逻辑的简化形式。代理奖励在基线分布上看似有用，但优化步骤会促使生成器将概率质量分配给某个补全（z_3）；该补全在 π_0 下很罕见，之所以受到偏好，主要是因为它触发了奖励中的虚假成分。关键的比较静态分析与正文中的相同：更激进的嵌入（更小的 λ）会扩大这种放大的空间。