arXiv:2601.07573v1 [econ.TH] 12 Jan 2026

人工锯齿智能模型

Joshua S. Gans*

2026年1月13日

摘要

生成式人工智能系统在看似“相近”的任务上，表现往往极不均衡：只需稍微改变措辞或上下文，它们就可能在某个提示词上表现出色，却在另一个提示词上自信地给出错误答案。我们将这一现象称为人工锯齿智能（AJI）。本文构建了一个易于分析的 AJI 经济学模型，将采用视为一个信息问题：用户关心局部可靠性，但通常只能观察到粗略的整体质量信号。在基线的一维景观中，真实情况遵循粗糙的布朗过程，而模型“知道”的是从泊松过程中抽取的零散点。模型进行最优插值，局部误差则以后验方差衡量。我们推导出盲用户的采用阈值，说明实际经历的错误会因检查悖论而放大，并将缩放定律解释为更密集的覆盖：这能提升平均质量，却无法消除锯齿性。随后，我们研究掌握与校准：即使在未通过盲目采用检验的定义域中，能够依据局部不确定性采取行动的校准用户仍可获得正的期望价值。将掌握建模为通过高斯过程回归学习可靠性映射，可得出一个由信息增益驱动的学习率界，从而阐明发现“模型在哪些方面有效”何以进展缓慢。最后，我们研究规模扩展与可发现性的相互作用：校准信号和用户掌握何时会加快规模提升效益的获取，以及不透明性何时会使规模扩展带来的收益实际上无法被察觉。《经济学文献杂志》分类号：D83、D81、O33、L86。

关键词：生成式人工智能、采用、校准、学习、知识密度、规模扩展。

*多伦多大学罗特曼管理学院及美国国家经济研究局。谨此感谢 Refine.ink 和 ChatGPT 5.2 提供的研究协助，以及加拿大社会科学与人文研究理事会提供的资助。感谢 Tom Cunningham 提供宝贵意见。所有错误均由本人负责。所有通信请寄至 joshua.gans@utoronto.ca。

1 引言

生成式人工智能正以前所未有的速度融入知识工作。新颖之处并不在于这项技术并不完美；大多数技术（事实上，人工也是如此）都不完美。新颖之处在于，这些缺陷往往是局部的，而且不透明。用户经常遇到一种难以用“平均准确率”思维解释的情况：模型可能在某个提示词上给出清晰、正确的答案，接着却在一个相近的提示词上给出看似合理、语气笃定但实际错误的答案。措辞、上下文长度或格式上的细微变化，都可能导致表现发生难以事前预测的转变。这种不均衡越来越引人关注，因为用户最在意避免罕见但代价高昂的失误，恰恰是在高价值领域（法律、医疗、政策、金融、研究）。

经验证据与这种局部且不均衡的模式相符。在一项针对专业写作任务的在线实验中，使用 ChatGPT 提高了平均生产率和质量（Noy 和 Zhang，2023）。在客户支持代理参与的实地研究中，生成式人工智能助手提高了平均生产率，但不同员工和问题类型之间存在显著异质性（Brynjolfsson 等人，2025年）。最直接的是，一项针对管理顾问的实地实验揭示了一个“锯齿状技术前沿”：人工智能显著提升了某些真实任务的表现，却降低了另一些看似相似的任务的表现（Dell’Acqua 等人，2023）。这些发现促使我们建立一种理论，认为采用与否不仅取决于平均质量，还取决于局部可靠性的可发现性。

本文为这一被称为人工锯齿智能 （AJI）的现象提出了一个简单模型。[1] 建模的目的并非复现任何特定系统的工程细节，而是构建一个简洁的经济学环境，使以下三个事实能够同时成立：

(i) **局部异质性：** 在任务空间中的表现参差不齐，既有能力突出的“口袋”，也有错误率高的“漏洞”；

(ii) **不透明性：** 用户起初并不知道这些“口袋”和“漏洞”在哪里； d

(iii) **可发现性至关重要：** 采用和生产力不仅取决于平均质量，还取决于用户能否找到可靠的区域并避开不可靠的区域。

从这个意义上说，AJI 最适合被视为一个信息问题 ，而非工程故障。模型的平均表现可能非常出色，但如果局部可靠性参差不齐且难以推断，仍然难以有效使用。

1.1 为什么经济学家应该关注

经济学家长期以来一直在研究质量存在不确定性、且须通过使用才能显现的技术采用。在关于经验品的经典研究中，关键障碍在于买家必须通过消费商品来了解其质量（Nelson，1970）。在关于逆向选择的经典研究中，关键障碍在于质量信息由私人掌握，导致市场瓦解（Akerlof，1970）。生成式人工智能带来了一种独特但相关的障碍：质量不确定性不仅存在于不同产品或企业之间，

[1] Andrej Karpathy 使用“锯齿智能”一词来描述模型在不同任务上的表现参差不齐；参见 https://x.com/karpathy/status/1816531576228053133。我们沿用这一含义，关注局部可靠性的参差不齐如何影响采用、校准和学习。

还存在于同一产品的不同任务之间。对用户而言，相关的对象并非“模型准确率”，而是模型在用户恰好面临的任务上的可靠性。

这种工具内部异质性会立即产生经济影响。首先，当工具的局部可靠性难以发现时，采用可能会过于保守：即使工具在相当一部分任务上具有价值，用户也可能因担心罕见的灾难性失误而避开使用。其次，采用工具后，用户可能会将任务错误地分配给模型：在模型不可靠的地方委派任务，而在模型可靠的地方不进行委派，从而造成生产力损失并影响信任动态。第三，由于掌握程度和界面设计会影响可发现性，对校准、工作流设计和用户培训的投资，与单纯改进模型在经济上具有互补性。这与一种更广泛的观点相呼应，即重大技术变革需要互补性创新和组织适应（Bresnahan 和 Trajtenberg，1995）。

1.2 建模策略

建模的挑战在于生成参差不齐的局部性能，同时避免引入不必要的复杂性。我们的策略包含两个要素。

知识点。 我们将模型“知道”的内容表示为一组任务，即模型能够在这些任务上正确作答。这种表示显然是风格化的，但它抓住了一个核心理念：训练、检索和微调会在某些区域形成密集的能力，而在另一些区域形成稀疏的能力。我们将知识点建模为强度为 λ 的泊松点过程；这是衡量覆盖率 的简化形式，也是与人工智能“规模”建立联系的自然对象。重要的是，用户无法观察这些知识点所在的位置。

这种知识点表示方式与 Carnehl and Schneider（2025年）对知识的“空间”视角密切相关。我们的侧重点有所不同：我们将用户无法观察局部可靠性及其带来的校准、掌握和界面设计作用，视为主要的经济摩擦。

粗糙的真实值景观。 在知识点之间，模型必须进行插值。我们将任务到正确答案的映射表示为一个粗糙的随机过程（基线采用布朗运动）。这使局部外推具有风险，并能得出知识点之间的闭式后验分布。选择布朗运动，并非因为现实世界真的遵循布朗运动，而是因为它是最简单的“粗糙”过程，且具有马尔可夫性质，能让插值过程在解析上清晰明了。[2]

这些假设共同定义了每项任务的局部均方误差。锯齿性由此产生：知识点附近的误差较小，间隔中部的误差较大，而间隔长度是随机的。

1.3 主要结果与直觉

为了明确说明这一想法，我们来看看一个简单的玩具模型，称为“知识之桥”。设想人工智能的任务空间是一条河流，而它的知识点是支撑桥梁的桥墩。模型通过桥板在桥墩之间进行插值，而桥板会因承重而下弯：跨距短的地方坚固安全，

[2]在方法上，本文借鉴了高斯过程学习和老虎机设计中的工具（Rasmussen 和 Williams，2006年；Srinivas 等，2010年；Vakili 等，2021年）。因此，我们沿袭 Callander（2011年）在经济学领域开创的研究；综述请参见 Bardhi and Callander（2026年）。

跨距长的地方则灵活而危险。假设这座桥由两个重复的路段组成：一段 2 米的短跨距和一段 8 米的长跨距。评估这座桥的工程师可能会计算出平均间隔为 5 米。但走过这座桥的用户却会有不同的体验。他们只有 20% 的时间走在安全路段上，80% 的时间走在危险路段上，因此体验到的平均间隔为 6.8 米。这就是**检查悖论**：从统计上看，用户会过度暴露于模型的弱点之中。这些空洞的重要性远超其出现频率所暗示的程度。这道简单的算术题阐明了采用与掌握背后的经济学：

• **盲目采用**。 无法观察局部可靠性的用户，必须事先决定是否依赖该工具。如果其安全阈值要求间隔小于 6 米，那么即使朴素的基准平均值表明桥梁是安全的，他们也会作出正确决定而拒绝使用。基于平均值的评估会产生误导，因为采用取决于已体验到的风险，而检查悖论会放大这种风险。[3]

• **规模扩展**。 将桥墩的密度增加一倍，会把间隔缩短为 1 米和 4 米。但相对崎岖性保持不变：用户仍有 80% 的时间处于最长的路段上。规模扩展提高了可靠性的水平，却没有改变异质性的形状。这有助于解释为何显眼的基准指标有所改善，却仍会持续出现“令人意外”的失败。

• **校准**。 现在，假设用户能够看见桥墩的位置。他们通过安全的 2 米路段，并避开危险的 8 米路段。这会将一次高风险的赌博转变为具有正期望价值的工具——即使盲目采用并不理性。校准能够释放能力局部区域，而不必承担这些空缺的全部代价。

• **掌握**。 校准是一项基准。在实践中，用户必须通过经验学习 可靠性映射。我们表明，当任务空间维度很高或可靠性景观崎岖不平时，即便是经验丰富的用户，学习过程也很缓慢。掌握存在速度上限。

• **互补品与替代品**。 低于盲目采用阈值时，规模扩展与校准是互补品：单靠规模扩展无法促成采用，但校准能够释放价值。高于该阈值时，两者则变为替代品：当模型已经足够好、可以盲目使用时，校准的边际收益会递减。

本文为方兴未艾的基础模型经济学作出四项贡献。第一，我们提出一个**崎岖可靠性的微观基础模型**。一个简约的“知识点”框架能够刻画随任务而显著变化的局部误差。由此得到一个具体对象——每个任务位置上的后验方差——将人工智能能力“崎岖不平”这一非正式概念（Dell’Acqua 等人，2023）转化为可操作的表述。第二，我们刻画**不透明性下的采用**。通过将检查悖论纳入采用决策，我们推导出一条简单的阈值规则，并正式说明，当局部失误显著时，平均基准表现为何可能无法为采用决策提供可靠指引。这项研究通过揭示测量可靠性与实际体验可靠性之间的系统性落差，为不断发展的人工智能评估文献（Burnell 等人，2023）作出贡献。第三，我们推导出一条**掌握程度的学习率界**。将

[3] 在下文概述的模型中，“最大容忍间隔”规则映射为赌注截止值 q：在 th e 布朗运动 b 脊方差 (1)，$\mathbb{E}[\sigma^2 \mid X] = X/6$，因此盲目采用可简化为 $q \geq \mathbb{E}[\sigma^2(x)] = \mathbb{E}[X^*]/6$。

表1：与“平均性能”视角相比，AJI所增添的内容

	基准测试平均视角	AJI视角
对象	测试套件上的平均分数	作为任务函数的可靠性
暴露度	代表性基准测试项目	偏向稀疏区域
主要阻碍	平均能力	错误集中位置的不透明性——程度
关键楔子	“模型与基准”	基准与经验 风险
杠杆	规模	规模 ＋ 规律性 ＋ 校准 ＋ 掌握

掌握程度建模为对潜在可靠性图谱进行高斯过程回归后，我们表明，最坏情况不确定性会以信息增益所决定的速率下降。这使“了解模型在哪些地方有效”何时进展缓慢这一问题得到了形式化阐释，并与人机协作的实证研究形成互补（Lai 等人，2023）。[4] 第四，我们刻画**规模扩展与可发现性之间的互补性**。分析阐明了界面与治理设计——包括相似性提示、不确定性估计、弃答机制和来源信息——在何种情况下是原始模型性能提升的替代品，又在何种情况下是互补品。这为产品设计和监管策略提供了参考。

本文受到一系列关于生成式人工智能在受控环境和实地场景中生产率效应的实证研究启发（Noy 和 Zhang，2023；Brynjolfsson 等人，2025；Dell’Acqua 等人，2023），并对三个理论研究领域的交汇处作出贡献。首先，本文将生成式人工智能的采用与质量不确定性和经验品经济学联系起来（Nelson，1970；Akerlof，1970）。其新颖之处在于工具内部异质性：局部可靠性会随任务空间而变化，因此，对于许多决策而言，“平均质量”并非相关的研究对象。其次，本文与技术使用中的学习和选择研究相关（Jovanovic，1982）。在我们的研究设定中，学习的对象不是一个标量质量参数，而是一个将任务映射为可靠性的函数。学习成本取决于任务空间的几何结构，以及用户能够利用何种相似性结构。第三，本文契合一种更广泛的观点：通用技术需要互补性投资和组织变革（Bresnahan and Trajtenberg，1995）。AJI 强调一种特定的互补要素：可发现性 工具（校准、来源信息、相似性提示）和熟练掌握资本（工作流程和训练），它们使选择性委托成为可能。

本文以 Carnehl and Schneider（2025年）的空间知识框架为基础，但将关注对象从“平均性能”转向各项任务的局部可靠性分布，以及用户因而暴露于最稀疏区域的程度。这带来了一组独特的经济影响：（i）间隙均匀基准与任务均匀体验之间的检查悖论楔子；（ii）由覆盖缺口的二阶矩决定的盲目采用阈值；以及（iii）规模、规律性、校准/弃答接口与熟练掌握投资（工作流程、提示词库和探索）之间显著的互补关系。

第 2 节介绍基线 AJI 模型，推导体验误差中的检查悖论楔子，并刻画盲目采用。第 3 节将缩放定律解释为更密集的知识覆盖，并说明为何平均质量可以提高而不消除锯齿性。

[4]在这方面，本文纳入了Rosenberg（1982年）研究的一项技术采用特征，即“通过使用学习”。

第 4 节研究校准与规律性，并刻画锯齿性（部分）可被发现时的规模报酬。第5节将推理引入为一种可选的增强可靠性的模式，并推导何时进行推理的截断策略。第6节将掌握建模为学习可靠性映射，并推导出凸显弃答陷阱的学习率界。第 7 节讨论扩展与应用。

2 AJI 模型与采用

要理解人工锯齿智能（AJI）的经济影响，我们需要一个能够刻画知识的空间相关性及其导致的表现差异的模型。我们采用 Carnehl and Schneider（2025年）（CS）提出的框架，该框架提供了一种便于分析的知识表示模型。随后，我们引入定义 AJI 情境的关键要素：关于知识具体实现形式的不完全信息。

本节构建基准模型，并推导出两个核心对象：预期体验误差和盲目采用阈值。在此过程中，我们还将说明这些原始参数如何映射到生成式人工智能情境中。

2.1 任务空间与相似性

设任务空间为一维域 $\mathcal{Z} \subset \mathbb{R}$。用户抽取一个任务 $x \in \mathcal{Z}$，并希望获得正确的标量答案 $Y(x) \in \mathbb{R}$。实际上，任务是存在于高维空间中的问题、提示、文档和上下文。采用一维限制是为了获得封闭形式的结果。[5] 这可以解释为：（i）沿工作流中任务的一维流形进行分析；或（ii）对嵌入所诱导的距离进行简化形式表示，并将关注范围限制在单一显著方向上的变化。[6]

基本假设是，不同问题的答案彼此相关：知道一个问题的答案，便能获得关于相似问题答案的信息。我们用一个随机过程来描述这种关系。

假设 1 (布朗景观)。 真值函数 Y： $\mathcal{Z} \to \mathbb{R}$ 是坐标 x上的（无漂移）布朗运动。[7]

布朗运动（或维纳过程）是一种连续随机过程，其增量相互独立且服从正态分布。具体而言，两个问题 $Y(x_2) - Y(x_1)$ 的答案之差服从正态分布，其方差等于距离 $|x_2 - x_1|$。这体现了两个基本特征：（i）局部相关性：相近问题的答案高度相关；（ii）累积不确定性：答案之间可能存在的差异

[5]采用一维限制是为了得到闭式插值风险（布朗桥方差）。其底层的检查悖论机制并非本质上局限于一维：在 d 维中，泊松覆盖会生成沃罗诺伊镶嵌，而任务均匀抽样会使胞元体积产生大小偏差（引理9）。将单元几何结构转化为 σ^2（x）取决于核/协方差结构，但基准加权与用户体验之间的核心落差依然存在。[6]第6节将进一步讨论用户学习可靠性映射时维度与光滑度的作用。[7]形式上，对于 $x_1 < x_2$, Y（x_2） $- Y$（x_1） $\sim \mathcal{N}$（0， $x_2 - x_1$），其具有独立增量。参见 Karatzas and Shreve（1991年）。

会随着问题变得更加不同而增大。布朗运动假设刻意设定得“粗糙”：它连续，但处处不可微。这体现了 AJI 的一个核心直觉：即使在任务空间中只发生微小“移动”，也可能导致正确输出发生巨大变化。布朗运动假设在分析上也很方便：给定端点后，该过程就是一个方差公式简单的布朗桥。我们的目标并非追求字面意义上的真实性，而是构造一个易于处理的基准，以清晰的方式呈现局部插值的风险。

2.2 数据与人工智能预测

人工智能（AI）模型在以下假设下使用一组数据（即任务空间和答案空间中的点）。

假设2（知识点）。 知识点（或数据）集合 $\{x_i\} \subset \mathcal{Z}$ 服从强度为 $\lambda > 0$的齐次泊松点过程。给定点集，模型观测到 $\{（x_i, Y（x_i）\}$。

遇到一个新问题 $x \notin \{x_i\}$时，人工智能（即大语言模型）会根据知识点集合生成一个猜测。如果 x 位于两个相邻知识点 x_i 和 x_{i+1}之间（间隔长度为 $X_i = x_{i+1} - x_i$），人工智能就会进行插值。人工智能的最佳预测是条件ˆ均值 $Y(x)$，即在 $Y(x_i)$ 和 $Y(x_{i+1})$ 之间进行线性插值。该预测的可靠性由条件方差 $\sigma^2(x)$ 表示。泊松假设是对覆盖不规则性的简化形式表示。在人工智能语境中，λ 可以解释为训练支持、记忆样本、检索命中或其他机制的密度，这些机制使模型在任务空间的某些邻域中具有可靠性。我们关注平稳基准，使单一参数能够表示覆盖率；第 3 节将规模扩展解释为 λ的增加。

模型会给出后验均值预测：

$$\hat{Y}(x) \equiv \mathbb{E}[Y(x) \mid \{(x_i, Y(x_i))\}_i] .$$

在假设 1 下，对于位于两个相邻知识点 $x_i < x < x_{i+1}$ 之间的任务 x，$Y（x）$ 的条件分布是布朗桥。后验均值是线性插值，后验方差为

$$\sigma^2(x) \equiv \mathrm{Var}(Y(x) \mid \{(x_i, Y(x_i))\}_i) = \frac{(x - x_i)(x_{i+1} - x)}{x_{i+1} - x_i}. \tag{1}$$

在基线情形中，人工智能根据其掌握的信息实现贝叶斯最优：它会输出均方误差最小的预测。因此，$\sigma^2（x）$并非任意的“错误率”，而是一个定义明确的对象：根据模型掌握的信息，在 x 处存在的不可约不确定性。

2.3 用户收益与信息

用户可以选择：（i）使用人工智能输出，或（ii）选择外部选项。我们对收益进行标准化，使得在任务 x 中使用人工智能的条件下，期望效用等于一项收益减去一项损失

该损失与均方误差成正比：

$$U(x) = 1 - \frac{\sigma^2(x)}{q}, \tag{2}$$

而外部选项的收益为 0。Carnehl and Schneider（2025）为此提供了微观基础。设正确答案的价值为 B，外部选项的价值为 B_0，均方误差会使收益减少 $c\,\sigma^2(x)$。那么，(2) 是 $B - B_0 - c\sigma^2(x)$ 和 $q \equiv (B - B_0)/c$ 的仿射归一化，其中利害关系是一个参数。较高的 q 对应低利害关系或易于验证；较低的 q 对应高风险决策或高成本验证。[8]

这里的关键假设是，人工智能用户无法观察到其所面临任务的 $\sigma^2(x)$。他们或许能看到全局基准或轶事证据，但看不到局部可靠性图。这引出了下文的盲目采用情形。这体现了 AJI 的核心挑战：用户无法确定人工智能针对某个具体问题给出的答案质量 x，因为他们不知道附近存在哪些知识空白。

2.4 检查悖论与实际经历的误差

令 X 表示相邻知识点之间的距离。回想一下，在假设2下，$X \sim$ 服从参数为 $\mathbb{E}[X] = 1/\lambda$ 的指数分布(λ)。相比之下，用户的任务在定义域中的位置（近似）均匀分布。在给定实际划分出的间隔后，随机抽取的任务更有可能落在较长的间隔中。这就是检查悖论。

命题 1 (间隔的检查悖论)。令 X^* 表示包含一个均匀位置任务的间隔长度。那么，X^* 服从长度偏置密度

$$f_{X^*}(x) = \lambda^2 x e^{-\lambda x}, \qquad x \geq 0,$$

即 $X^* \sim$ 伽马分布$(2, \lambda)$，且 $\mathbb{E}[X^*] = 2/\lambda$。

证明。对于平稳泊松过程，间隔长度的密度为 $f_X(x) = \lambda e^{-\lambda x}$，均值为$\mathbb{E}[X] = 1/\lambda$。均匀位置落在长度为 x 的间隔中的概率与$x f_X(x)$成正比，因此

$$f_{X^*}(x) = \frac{x f_X(x)}{\mathbb{E}[X]} = \frac{x(\lambda e^{-\lambda x})}{1/\lambda} = \lambda^2 x e^{-\lambda x}.$$

这是伽马分布 Gamma$(2, \lambda)$，均值为 $2/\lambda$。 □

典型间隔为 $1/\lambda$，但典型任务所经历的间隔为 $2/\lambda$，因为较长的间隔占据了更多空间。这精确地说明了，用户“更可能处于模型正在外推的区域”，这种可能性高于朴素计算所暗示的程度。这意味着用户所经历的间隔长度是平均间隔长度的两倍。这体现了锯齿性带来的代价：不规律性会增加用户所经历的平均误差，因为用户遇到大语言模型表现不佳的间隔的概率更高。

[8]参数 q 也可以看作用户特定的判断能力代理指标（即对收益和错误成本的了解）；参见 Agrawal 等人（2025年）。

接下来，在给定长度为 X 的间隔的条件下，问题在间隔中的位置服从均匀分布。利用 (1)，可以得到一个简洁的条件均值。为说明这一点，令 d 表示距左端点的距离； $d \sim \text{Uniform}(0, X)$。则 $\sigma^2 = d(X-d)/X$。因此

$$\mathbb{E}[\sigma^2 \mid X] = \frac{1}{X}\mathbb{E}[d(X-d)] = \frac{1}{X}\left(X\mathbb{E}[d] - \mathbb{E}[d^2]\right) = \frac{1}{X}\left(X \cdot \frac{X}{2} - \frac{X^2}{3}\right) = \frac{X}{6}.$$

Co将命题 1 与这一计算结合，即可得到期望体验方差 .

$$\mathbb{E}[\sigma^2(x)] = \mathbb{E}\big[\mathbb{E}[\sigma^2(x) \mid X^*]\big] = \mathbb{E}\left[\frac{X^*}{6}\right] = \frac{1}{6} \cdot \frac{2}{\lambda} = \frac{1}{3\lambda}. \tag{3}$$

随着覆盖率 λ 提高，预期误差会降低，但仍高于根据 $\mathbb{E}[X] = 1/\lambda$ 进行朴素计算所得的结果。若将相关间隔视为“典型”间隔，则会得到 $\mathbb{E}[\sigma^2] = 1/(6\lambda)$。检查悖论使人们暴露于长间隔的程度加倍，也使期望方差加倍至 $1/(3\lambda)$。当采纳决策取决于用户实际面临的任务分布时，这一区别就很重要。

许多公开评估会报告精选基准数据集上的平均准确率。在基准模型中，相关均值是根据任务到达分布计算的。该均值按照间隔的长度对覆盖中的“间隔”赋予权重（命题 1），而不是采用均匀分布。换言之，同样的基础能力可能显得更好或更差，具体取决于评估是在整个任务空间中均匀抽样，还是更多地从训练支持密集的区域抽样。这一观察表明，当采纳决策受罕见且代价高昂的失败驱动时，评估协议应在报告均值的同时，也报告尾部风险度量（例如按簇计算的最差表现，或以距训练支持的距离为条件的表现）。[9]

2.5 盲用户的采用

盲 用户必须事前决定是否依赖模型，而不观察实际任务中的 $\sigma^2(x)$。这涵盖了订阅采纳、组织的“默认使用”政策，或任务级验证成本高昂的情形。

假设 3 （盲目采用）。 **</style}用户在** （x）] ≥ 0 时采用人工智能；否则，选择外部选项。

基于这一假设，我们可以得出一个关键结果：

定理 1 （采用阈值）。 在假设 *1–3* 下，当且仅当发生盲目采用 *if*

$$q \geq \frac{1}{3\lambda}.$$

*Equ*等价地，定义可靠性比率 $R \equiv 3\lambda q$，当且仅当 $R \geq 1$ 时，才会采用 .

证明。 由（2）和（3）可得，

$$\mathbb{E}[U(x)] = 1 - \frac{\mathbb{E}[\sigma^2(x)]}{q} = 1 - \frac{1}{3\lambda q}.$$

[9]对于 $\mathbb{R}$ 上的平稳泊松过程，如果随机任务是从发生的任务中“均匀地”选取的，那么泊松/长度偏差计算是精确的。在任务均匀到达的大型有界域中，对于远离边界的点，相同的公式近似成立；随着定义域增大，这种近似会变得精确。

当且仅当 $q \geq 1/(3\lambda)$ 时，此值非负，即 $R \geq 1$。 □

在盲目采用的情况下，用户会在了解实际任务 x 在崎岖景观中的位置之前 就决定依赖该模型。因此，在采用的条件下，使用人工智能的单项任务净回报率就是 $\bar{U}_B(\lambda, q) \equiv \mathbb{E}[U(x)] = 1 - \frac{1}{3\lambda q} = 1 - \frac{1}{R}$。 $R(\equiv 3\lambda q)$ 将~~-~~成为下文的一个关键参数。定理~~1~~等价于要求 $U_B(\lambda, q) \geq 0$ 或$R \geq 1$。

2.6 校准：完全知情基准

定理 1 中的盲目采用阈值由平均 期望效用决定。但 AJI 本质上是一种局部 现象：有些任务靠近知识点，模型的后验方差 $\sigma^2(x)$ 较小；另一些任务则处于漫长的知识空白之间，此时$\sigma^2(x)$ 较大。许多实际系统和工作流程都试图提供（不完美的）任务级局部可靠性信号（置信度启发式方法、弃答、检索溯源、交叉核验工具）。为单独考察其经济逻辑，本小节引入一个完美校准基准：用户会在决定是否在当前任务中依赖人工智能之前，观察实际的 $\sigma^2(x)$。

完全校准的用户会逐项将人工智能与外部选项进行比较。将外部选项标准化为 0 后，当且仅当 $U(x) \geq 0$ 时，用户才会依赖人工智能；等价地，$\sigma^2(x) \leq q$。因此，校准后的单任务效用为

$$u_C(x) = \max\{U(x), 0\} = \left(1 - \frac{\sigma^2(x)}{q}\right)_+, \qquad (a)_+ \equiv \max\{a, 0\}. \tag{4}$$

使用人工智能所对应的基准价值为

$$U_C(\lambda, q) \equiv \mathbb{E}[u_C(x)] = \mathbb{E}\left[\left(1 - \frac{\sigma^2(x)}{q}\right)_+\right]. \tag{5}$$

布朗–泊松基线的一个关键简化之处在于，校准价值仅通过可靠性比率 $R \equiv 3\lambda q$ 取决于 (λ, q)。因此，校准基准可推导如下：

命题2 (校准后的期望效用**)**。 在假设*1–2*（布朗–泊松基线）下，校准后的期望效用仅通过 R 取决于(λ, q)，即 $U_C(\lambda, q) = U_C(R)$。此外，对于每个 $R > 0$, $U_C(R) > 0$，且 U_C 随着..... 严格递增

$$U_C'(R) = \frac{1}{R^2}\mathbb{E}\big[3\lambda\sigma^2(x)\,\mathbf{1}\{3\lambda\sigma^2(x) < R\}\big] \in \left(0, \frac{1}{R^2}\right), \qquad R > 0.$$

最后，在布朗–泊松基线下， U_C 具有如下积分表示

$$U_C(R) = \int_0^1 \left[1 - \frac{6t(1-t)}{R} + \left(1 + \frac{6t(1-t)}{R}\right)\exp\left(-\frac{R}{3t(1-t)}\right)\right] dt.$$

证明。 Wr 将校准效用表示为 $\mathbb{E}[(1 - \sigma^2(x)/q)_+]$，并代入 $R = 3\lambda q$。为方便起见 in

下文将 归一化方差 定义为

$$Z \;\equiv\; 3\lambda\sigma^2(x). \tag{6}$$

因此

$$U_C(\lambda, q) = \mathbb{E}\left[\left(1 - \frac{Z}{R}\right)_+\right].$$

由 (3)， $\mathbb{E}[Z] = 3\lambda\mathbb{E}[\sigma^2(x)] = 1$。

在布朗–泊松基准下，对于均匀抽取的任务，令 X^* 为长度偏倚的间隔长度（命题 1），并令 $t \sim$ 为间隔内的相对位置，服从，独立于 X^*。那么由 (1)， $\sigma^2(x) = X^* t(1-t)$。由于 $X^* \sim$ 服从伽马分布(2， λ)，我们有$\lambda X^* \sim$ 服从伽马分布(2，1)，因此

$$Z = 3(\lambda X^*)\, t(1-t)$$

具有一个 分布，且该分布不依赖于 λ。因此， U_C 仅通过 R 依赖于 (λ， q) .

要证明对于所有 $R > 0$， $U_C(R) > 0$ ，注意到 Z 的支持集可以任意接近 0（例如，当 t 靠近间隔端点，或实际间隔较短时），因此 $\Pr(Z < R) > 0$ ，从而 $\mathbb{E}[(1 - Z/R)_+] > 0$。

对于每个已实现的 $z \geq 0$，映射 $R \mapsto (1 - z/R)_+ = (1 - z/R)\mathbf{1}\{z < R\}$ 是非递减的。对 $(1 - z/R)\mathbf{1}\{z < R\}$ 关于 R 求导得到 $\frac{z}{R^2}\mathbf{1}\{z < R\}$。由控制收敛，

$$U_C'(R) = \mathbb{E}\left[\frac{Z}{R^2}\,\mathbf{1}\{Z < R\}\right] = \frac{1}{R^2}\mathbb{E}[Z\,\mathbf{1}\{Z < R\}], \qquad R > 0,$$

由于 $\Pr(Z < R) > 0$，该值严格为正。

对于上界，注意到 $\mathbb{E}[Z] = 1$。事实上，

$$\mathbb{E}[Z] = 3\,\mathbb{E}[\lambda X^*]\,\mathbb{E}[t(1-t)] = 3 \cdot 2 \cdot \frac{1}{6} = 1,$$

因此 $\mathbb{E}[Z\mathbf{1}\{Z < R\}] < \mathbb{E}[Z] = 1$，从而 $U_C'(R) < 1/R^2$。

最后，为了得到积分表示，固定 t 并写作 $a \equiv t(1-t)$。由于 $X^* \sim$伽马分布$(2, \lambda)$ 的密度为 $f_{X^*}(x) = \lambda^2 x e^{-\lambda x}$ 且 $\sigma^2 = X^* a$，

$$\mathbb{E}\left[\left(1 - \frac{\sigma^2}{q}\right)_+ \,\middle|\, t\right] = \int_0^{q/a} \left(1 - \frac{ax}{q}\right) \lambda^2 x e^{-\lambda x}\, dx = 1 - \frac{2a}{\lambda q} + \left(1 + \frac{2a}{\lambda q}\right)\exp\left(-\frac{\lambda q}{a}\right).$$

S代入 $\lambda q = R/3$ 和 $a = t(1-t)$，并在 $t \sim$ 均匀分布(0,1) 上积分，得到

$$U_C(R) = \int_0^1 \left[1 - \frac{6t(1-t)}{R} + \left(1 + \frac{6t(1-t)}{R}\right)\exp\left(-\frac{R}{3t(1-t)}\right)\right] dt,$$

正如所声称的。 □

盲目信赖迫使用户即使在局部不确定性高到 $U(x) < 0$的任务上，也接受人工智能输出。这正是锯齿性发挥作用的地方： $\sigma^2(x)$的分布具有一个

由长间隔（检查悖论进一步放大）以及间隔内部几何结构产生的高误差尾部（$t(1-t)$在间隔中部达到峰值，其中 t在 $[0,1]$ 上均匀分布，表示其在间隔中的相对位置）。因此，盲目信赖所带来的经济成本由超过临界值 q的 $\sigma^2(x)$的上尾所决定。

回想一下，$U_B(\lambda,q) \equiv \mathbb{E}[U(x)] = 1 - 1/R$ 表示在所有任务上都盲目地依赖人工智能所获得的期望效用，那么完美校准弱支配盲目信赖，因为 $(a)_+ \geq a$逐点成立。将无法筛除局部负价值任务所造成的损失定义为（期望）盲用成本：

$$\Delta_B(\lambda,q) \;\equiv\; U_C(\lambda,q) - \bar{U}_B(\lambda,q) = \mathbb{E}\left[\left(\frac{\sigma^2(x)}{q} - 1\right)_+\right] = \frac{1}{q}\mathbb{E}\big[(\sigma^2(x) - q)_+\big]\,.$$

因此，锯齿性影响福利，不仅是通过平均误差（它决定 U_B），还通过高方差尾部的概率质量和超调（它决定 Δ_B）。若从命题2中的积分表达式减去$\bar{U}_B(\lambda,q)$，并利用 $\int_0^1 t(1-t)\,dt = 1/6$，则在布朗–泊松基准下，

$$\Delta_B(R) = \int_0^1 \left(1 + \frac{6t(1-t)}{R}\right)\exp\!\left(-\frac{R}{3t(1-t)}\right)\,dt.$$

这个积分可以用闭式求出：

$$\Delta_B(R) = \frac{2}{9}e^{-2R/3}\left[R\,K_0\!\left(\frac{2R}{3}\right) + (3-R)\,K_1\!\left(\frac{2R}{3}\right)\right],$$

其中 $K_\nu(\cdot)$ 是第二类修正贝塞尔函数。[10] 因此，校准将锯齿性从使用上的代价转化为一种期权价值：即使盲目使用的平均收益较低，用户仍可将依赖集中在局部可靠的区域，并在缺口处 abstain。这个基准之后会很有用，因为 $U_C(R)$ 和盲用差额 $\Delta_B(R)$ 都只通过单一指标 R 依赖原始参数 (λ,q)。

完美校准是一个刻意设定得十分极端的基准。一个自然的扩展是，让用户观察到一个噪声信号 $s(x) = \sigma^2(x) + \nu$(例如 $\nu \sim \mathcal{N}(0,\tau^2)$)，并在 $\mathbb{E}[\sigma^2(x) \mid s(x)] \leq q$ 时进行委托。随着 $\tau^2 \to 0$，这种方式逐渐接近经过校准的委托；随着 $\tau^2 \to \infty$，则会退化为盲目采用。这为研究带噪不确定性沟通如何削弱校准/弃答界面带来的福利增益，提供了一种简化形式的方法。

[10]函数 $K_\nu(\cdot)$ 是第二类修正贝塞尔函数（也称为 *Macdonald function*），即修正贝塞尔方程 $z^2y'' + zy' - (z^2+\nu^2)y = 0$ 的指数衰减解。对于 $z > 0$，它有如下积分表示

$$K_\nu(z) = \int_0^\infty e^{-z\cosh u}\cosh(\nu u)\,du,$$

由此可得 $K_\nu(z) > 0$，且 K_ν 随 z 增大而递减。其标准渐近式为 $K_\nu(z) \sim \sqrt{\frac{\pi}{2z}}\,e^{-z}$ $(z \to \infty)$，$K_0(z) \sim -\log(z/2) - \gamma$，$K_1(z) \sim \frac{1}{z}$ $(z \downarrow 0)$，其中 γ 是欧拉–马歇罗尼常数。在 $\Delta_B(R)$ 的闭式表达式中取参数 $z = 2R/3$，这些性质意味着 $\Delta_B(R) > 0$，当 $R \downarrow 0$ 时，$\Delta_B(R) \sim 1/R$，且当 $R \to \infty$ 时，$\Delta_B(R) \sim \big(\sqrt{3\pi}/(3\sqrt{R})\big)e^{-4R/3}$。

3 人工智能规模扩展定律的影响

大量实证研究记录了现代人工智能系统近似遵循的“缩放定律”：更多的数据、计算资源和工程投入，往往会系统性地产生更好的模型（Kaplan 等，2020年）。在本文中，我们将“更好”解释为对任务空间的更密集的有效覆盖。形式上，规模扩展意味着知识点强度 λ 的提高。

本节记录规模扩展的基本比较静态分析，阐明锯齿性为何依然存在，并引入一个单一指标；在间隔形状固定的情况下，该指标决定采用和福利比较。第 4 节随后将刻画这些边际的回报率。

3.1 规模扩展即更密集的覆盖率

首先，让我们在模型的语境下定义规模扩展。

定义 1 (规模扩展)。 规模扩展是指知识点强度从 λ_0增加到 $\lambda_S = S\lambda_0$ 的外生变化，其中 $S > 1$ 是缩放因子。

根据这一定义，规模扩展的影响如下：

命题 3 （规模扩展不会增大局部后验方差）。 固定 $0 < \lambda_0 < \lambda_S$。将对应的泊松知识点集合耦合起来：把一个 λ_S-过程写成一个独立的 λ_0-过程与一个独立的 $(\lambda_S - \lambda_0)$-过程之并。令 $\sigma_0^2(x)$ 和 $\sigma_S^2(x)$ 分别表示在这两个知识点集合下 $Y(x)$ 的后验方差。那么，对任意任务 x,

$$\sigma_S^2(x) \le \sigma_0^2(x) \qquad \textit{almost surely.}$$

证明。 固定耦合知识点集合的一种实现，并固定一个任务 x。令 $u_0 < x < v_0$ 表示基线集合中夹住 x 的相邻知识点，令 $u_S < x < v_S$ 表示规模扩展后的集合中夹住 x 的相邻知识点。在这种耦合下，规模扩展后的集合包含基线集合，因此夹区只会缩小： $u_0 \le u_S < x < v_S \le v_0$。

根据假设 1， x 处的后验方差仅取决于两侧的锚点，并由布朗桥公式 (1) 给出。因此

$$\sigma_0^2(x) = \frac{(x-u_0)(v_0-x)}{v_0-u_0}, \qquad \sigma_S^2(x) = \frac{(x-u_S)(v_S-x)}{v_S-u_S}.$$

如果 $(u_S, v_S) = (u_0, v_0)$，则没有内容可显示。否则，缩放后的集合会在 (u_0, v_0) 中至少引入一个额外的锚点，以一个更靠近 x 的点替换夹逼区间的一个端点。直接比较可知，端点向内移动时，桥方差弱递减：对于 $u < x < v' < v$,

$$\frac{(x-u)(v'-x)}{v'-u} \le \frac{(x-u)(v-x)}{v-u},$$

左端点向右移动时也对称地成立。因此，对于这次实现中的每个 x ，都有 $\sigma_S^2(x) \le \sigma_0^2(x)$，故该结论几乎必然成立。 □

在应用中，增加 λ 可以代表更多参数、更多训练数据、更完善的检索、领域微调，或使任务在实际意义上更“接近”的表征学习。

该模型不受具体机制的限制：所有能提高有效覆盖密度的改进，都会计入 λ。

3.2 基准测试、采用与充分指标

在布朗–泊松基准下，预期后验方差随规模线性下降：

$$\mathbb{E}[\sigma^2(x)] = \frac{1}{3\lambda_S} = \frac{1}{3S\lambda_0}.$$

这体现了规模扩展带来的熟悉的“平滑”基准改进。但用户的报酬取决于他们所面对的任务，而非平均误差：采用与福利取决于相对于利害关系的局部 可靠性。在该模型中，相关的比较是实际局部方差$\sigma^2(x)$是否相对于误差容忍度 q足够小。

对这一权衡进行概括的一种便捷方式是使用综合指标

$$R \;\equiv\; 3\lambda q. \tag{7}$$

有两个特点促使我们采用 R。首先，它将一个供给侧 对象（λ，有效覆盖）与一个需求侧 对象（q，利害关系／核验难度）汇总为单一的“可靠性”指标：在 q固定时，规模扩大 S 倍会使 R 乘以 S；而在 λ 固定时，转向利害关系更高的环境会使 R 按比例降低。其次，在基准情形下

$$R \;=\; \frac{q}{\mathbb{E}[\sigma^2(x)]},$$

因此， R 可理解为“以典型不确定性为单位衡量的利害关系”。因此，对 R 进行比较静态分析，可直接转化为（i）规模扩展回报，以及（ii）不同定义域之间利害关系的差异。

之所以 R 足够（保持间隔的形状参数 不变），是因为齐次性：按比例缩小会使间隔长度同比缩短，因此局部不确定性也会以相同方式重新缩放。下一个引理将这一不变性表述为一种有助于分析福利和使用决策的形式，因为二者都取决于 $\sigma^2(x)/q$的分布。

引理1（归一化方差的尺度不变性）。 假设间隔长度可表示为规模形式 $X = Y/\lambda$ ，其中 Y与 λ 相互独立（即规模扩展会改变强度，但不会改变间隔长度分布的形状）。那么，归一化局部方差 $3\lambda\sigma^2$（x）的分布与 λ 无关（但可能取决于 Y的形状参数）。因此，对于固定的间隔形状，任何仅通过 σ^2（x）$/q$ 的分布取决于（λ、 q）的福利或使用对象，都仅通过 $R = 3\lambda q$取决于（λ、 q）。

证明。 令 X^* 为长度偏倚的间隔长度，并令 $t \sim \text{Uniform}(0,1)$ 表示间隔内的相对位置。根据布朗桥公式， $\sigma^2(x) = X^* t(1-t)$。若 $X = Y/\lambda$，且 Y与 λ相互独立，则 $X^* = Y^*/\lambda$，其中 Y^*是 Y的长度偏倚版本，因此 $\lambda X^* = Y^*$ 与 λ相互独立。于是

$$3\lambda\sigma^2(x) = 3(\lambda X^*)\, t(1-t) = 3Y^*\, t(1-t),$$

它与 λ 相互独立，因为 Y^* 和 t 均与 λ相互独立。

最后，请注意 $\sigma^2(x)/q = (3\lambda\sigma^2(x))/R$。由于 $3\lambda\sigma^2(x)$ 的分布具有 λ 不变性，任何通过对 $\sigma^2(x)/q$ 的函数进行积分而形成的表达式，都只通过 R 依赖于 (λ, q)。 □

3.3 为什么锯齿性在规模扩展下依然存在

尽管规模会降低不确定性的程度，但未必会改善覆盖的形状参数。在泊松基准下，相对离散度具有尺度不变性：

命题 4 (锯齿性具有尺度不变性)。在假设*2*下，间隔长度的变异系数为常数：

$$CV(X) = \frac{\sqrt{\mathrm{Var}(X)}}{\mathbb{E}[X]} = 1,$$

与 λ 无关。

证明。若 $X \sim$ 服从指数分布(λ)，则 $\mathbb{E}[X] = 1/\lambda$ 且 $\mathrm{Var}(X) = 1/\lambda^2$，因此 $\mathrm{CV}(X) = 1$。 □

特别是，检查悖论惩罚 $\mathbb{E}[X^*]/\mathbb{E}[X] = 2$ 具有尺度不变性。规模扩展会使间隔的分布向内收缩，但不会改变其相对离散度，因此较长的间隔相对于均值仍然较长，而用户遭遇长间隔的概率不会按比例降低。这有助于解释为何基准改进未必能一对一地转化为高风险场景中“意外失败”率的下降。

泊松基准特意设定为平稳的。实际情况中，覆盖率存在异质性；一种自然的扩展是采用空间变化强度 $\lambda(x)$，即使总体规模很大，也会持续存在难以应对的区域。平稳基准将随机覆盖和长度偏倚暴露所产生的锯齿性单独剥离出来，再进一步纳入系统性异质性。

4 规模报酬率、校准与规律性

本节提出一个刻意注重实践的问题：如果人工智能提供商（或部署组织）的工程资源有限，应当投向何处？关于“人工智能进展”的公开讨论，往往将截然不同的活动都归结为单一的规模叙事。AJI 则让这些区别更加清晰。

由于收益是逐项任务实现的，而非依据基准平均值实现，因此有必要将实践中常被混为一谈的三种投资边际区分开来：

(i) **规模** （能力）：提高 λ，从而缩小后验不确定性，但无法消除知识空白。

(ii) **规律性** （覆盖形态）：降低知识空白的离散度，减少对罕见但巨大的知识空白的暴露（以及检查悖论惩罚）。

(iii) **校准** （信息/接口）：提供局部可靠性信号（此处为 $\sigma^2(x)$），并允许弃答（第2.6节）。

在模型中，用户福利是在局部、针对实际面临的任务实现的，因此，任何改进的边际收益取决于：(i) 局部可靠性的分布；(ii) 用户的信息与决策规则；以及 (iii) 评估基准对任务空间的加权方式。因此，我们围绕以下内容组织

分析的三大主题：(i) 规模扩展、校准和规律性的回报率驱动因素；*(ii)* 这些边际是互补品还是替代品；以及 *(iii)* 基于基准的推断如何错误估计回报，以及 *AJI* 如何提出修正后的评估目标。我们始终使用可靠性指数，该指数汇总供给侧能力（λ，有效覆盖密度）和需求侧风险（q，核验难度）。在 q 固定的情况下，按因子 S 进行规模扩展会使 R 乘以 S；在 λ 固定的情况下，转向利害关系更高的场景会使 R按比例降低。在稳定的差距形状下（第 3 节），许多福利和使用相关的对象仅通过 R依赖于 $(\lambda,\ q)$，从而可以进行清晰的比较静态分析。

4.1 扩展回报

规模扩展会增加 λ，因此会逐点弱化降低局部后验方差（命题 3）。这种统计改善带来的经济回报取决于两个边际：（i）广延边际（用户是否采用），以及（ii）集约边际（采用后人工智能的使用程度）。

盲用户一旦采用，便会将所有任务委托给人工智能；当期望效用非负时，用户就会采用（假设 3）。在布朗–泊松基准中，定理 1 给出 $U_B(R) = \max\left\{1 - \frac{1}{R},\, 0\right\}$。由此得到的规模边际回报（以 R 为单位衡量）为

$$\frac{\partial U_B}{\partial R} = \begin{cases} 0, & R < 1, \\ \frac{1}{R^2}, & R > 1, \end{cases} \qquad \text{and} \qquad \frac{\partial U_B}{\partial \lambda} = 3q\,\frac{\partial U_B}{\partial R}. \tag{8}$$

$R = 1$ 处的折点至关重要：低于该阈值时，边际规模扩展的实现收益为零，因为用户不会采用；高于该阈值时，回报为正，但会像$1/R^2$那样递减。

在完美校准下，用户观察到 $\sigma^2(x)$，且当且仅当 $\sigma^2(x) < q$ 时才会委托任务（第2.6节）。命题 2 表明，期望效用仅通过 R 取决于 (λ, q)；对其求导（如命题2所示）可得

$$\frac{\partial U_C}{\partial R} = \frac{1}{R^2}\mathbb{E}\left[3\lambda\sigma^2(x)\,\mathbf{1}\{3\lambda\sigma^2(x) < R\}\right] \in \left(0, \frac{1}{R^2}\right) \qquad \text{for all } R > 0, \tag{9}$$

以及 $\partial U_C/\partial\lambda = 3q\,\partial U_C/\partial R$。从经济学角度看，校准将规模扩展从一种“有拐点”的投资（采用之前毫无价值）转变为一种平滑的投资（对所有 $R > 0$ 都能带来正回报），因为即使平均可靠性低到不足以盲目采用，用户仍可以利用低方差区域。

校准还会改变人工智能的使用程度。将使用份额定义为在随机任务上进行委托的概率：

$$s_B(R) = \mathbf{1}\{R \geq 1\} \qquad \text{(blind: use whenever adopted)},$$

并且，在校准下，

$$s_C(R) \equiv \Pr(\sigma^2(x) < q). \tag{10}$$

条件 对 t 进行离子化并使用 $\lambda X^* \sim$ 伽马分布(2,1) 可得到一维表示

$$s_C(R) = \int_0^1 \left[1 - \left(1 + \frac{R}{3t(1-t)}\right) \exp\left(-\frac{R}{3t(1-t)}\right)\right] dt, \tag{11}$$

因此，对于有限的 R， $s_C(R) \in (0, 1)$，且随 R 增大而增加。因此，与盲目不采用相比，校准在低-R情境下可以提高使用率；而与盲目"总是委托"相比，在高-R 情境下则可以降低使用率，同时仍能提升福利。

4.2 校准的回报

校准是一种信息/界面边际：它不会（必然）改变底层模型的能力，但会通过实现选择性委托，改变能力与福利之间的映射。在模型中，完美校准意味着用户可以观测到局部后验方差$\sigma^2(x)$，并在方差过高时选择弃权。在实践中，"投资于校准"对应于构建并部署可靠性信号和政策，并依据该信号进行决策。在第6节中，我们将探讨用户如何自行实现校准。这里，我们重点关注人工智能提供商可以进行的投资。

典型的实施手段包括：

- **置信度／风险估计：** 训练一个独立的预测器（或使用熵、对数概率、自一致性、检索诊断等内部信号），以估计给定查询出错的概率。

- **信号校准：** 拟合并验证一种映射，使声明的置信度与经验准确率相符（例如温度缩放、保序回归、普拉特缩放或保形校准），通常按定义域和查询类型分层。

- **选择性预测与弃答：** 设定阈值，并将低置信度案例路由给人工、交由更安全的工具处理、转入检索优先工作流，或选择"提出澄清问题"，而不是直接作答。

- **验证层：** 使用批评／验证模型、基于工具的检查或单元测试生成第二意见，并将其纳入风险评分和弃答规则。

- **界面与治理：** 向用户呈现不确定性（或强制执行审核关卡），从而要求高风险行动具备更高的可靠性。

这些干预措施主要重塑尝试哪些任务 以及如何 在不确定情况下作出响应，而不是均匀地提升原始准确率。

自然的福利对象是获得局部可靠性信号（即， $\sigma^2(x)$）相对于盲目信赖所带来的收益。回想一下， $U_B(\lambda, q) \equiv \mathbb{E}[U(x)] = 1 - 1/R$ 表示在所有任务上盲目信赖人工智能所带来的期望效用（忽略采用截断），而盲用成本为：

$$\Delta_B(\lambda, q) \equiv U_C(\lambda, q) - \bar{U}_B(\lambda, q) = \mathbb{E}\left[\left(\frac{\sigma^2(x)}{q} - 1\right)_+\right] = \frac{1}{q}\mathbb{E}\left[(\sigma^2(x) - q)_+\right]. \tag{12}$$

这一数量是完美校准所带来的回报：它恰好等于因无法筛除局部负价值任务而产生的预期损失。它不仅取决于 $\sigma^2(x)$的均值，还取决于超过临界值 q的尾部概率质量和超出幅度。在实际应用中，当系统有时会面临可在低成本下检测并避开的“不良区域”（高方差）时，校准的价值最大。

命题5 （校准回报随可靠性提高而缩小）。 固定间隔形状参数（因此 U_C 对 λ, q 的依赖仅通过 $R = 3\lambda q$）。那么，对于所有 $R > 0$， $\Delta_B(R) > 0$，且 $\Delta_B(R)$ 关于 R 严格递减，且

$$\Delta_B'(R) = -\frac{1}{R^2}\mathbb{E}\big[3\lambda\sigma^2(x)\,\mathbf{1}\{3\lambda\sigma^2(x) > R\}\big] < 0.$$

证明。 对于固定的已实现 $s \geq 0$，映射 $R \mapsto \left(\frac{3\lambda s}{R} - 1\right)_+$ 是非递增的，并且除 $R = 3\lambda s$ 处的折点外，对$R > 0$ 可微，其导数为 $-\frac{3\lambda s}{R^2}\mathbf{1}\{3\lambda s > R\}$。控制收敛给出导数公式。由于基线下 $3\lambda\sigma^2(x)$ 的支持集任意接近 0 且无上界，因此对于所有 $\overline{R} > 0$，$\Pr(3\lambda\sigma^2(x) < R)$ 和 $\Pr(3\lambda\sigma^2(x) > R)$ 均为正，这就证明了正性。

□

命题5阐明了为什么校准在边际领域最有价值：当 R 较小时，相当一部分任务在局部上具有负价值，因此筛选很有价值；随着 R 增大，这类任务会变得更少，筛选的增量价值也随之下降。在实践中，这意味着，在高利害关系或低覆盖环境中，对可靠的不确定性估计、选择性弃答和升级策略的投资能带来最大的福利增益；而在高R环境中，它们主要充当安全层，其边际收益会递减。

4.3 规律性的回报

规模扩展会提高 λ，并缩短任务空间中“锚点”之间的均值 间隔长度。在实践中，λ 可以指更多参数、更多预训练词元、更多计算资源、更好的检索，或任何能够提升整体有效覆盖的机制。规律性 是提供方侧一个独立的边际：它关注的是覆盖率在何处 得到提升，而不仅仅是平均而言覆盖率有多少 。

在模型中，锯齿性源于知识点间距不均：有些区域锚点密集，另一些区域则锚点稀疏。对于大语言模型，类似的不均衡源于训练和后训练流程未能在整个任务空间中均匀分配投入。具体而言，提供方可以通过以下选择提高规律性（降低有效间隔的离散度）：

• **针对稀疏区域的数据获取与数据整理**。 收集或授权获取代表性不足的领域数据（例如专业技术写作、企业格式、少数语言、小众工具工作流），并将其清理至可用质量。这不一定会大幅增加平均词元数，但能缩小最严重的空白。

• **重新加权与混合设计**。 调整训练权重，使稀有但重要的定义域相较于覆盖率已较高的定义域获得更多梯度质量（或进行过采样）。在保持词元预算均值不变的情况下，这会改变覆盖率的形状参数。

• **主动学习／故障驱动数据**。 利用评估、红队测试和产品遥测来识别高错误率簇，然后在这些簇中生成或收集数据。用模型语言来说，这相当于“填补最长的间隔”，而不是均匀地添加点。

• **定向后训练**。 领域适配器、专业专家（例如混合专家路由），或针对困难区域的微调，即使平均基准分数几乎没有变化，也能降低方差离散度。

• **检索与工具覆盖**。 专门为稀疏区域扩展检索索引、工具访问和落地资源，实际上是在局部增加“锚点”，从用户角度提升规律性。

这些杠杆不同于规模扩展，因为它们主要改变局部可靠性的形状参数：将投入从已经密集的区域重新分配到最严重的空缺区域。

规律性和校准解决的是不同的摩擦，对应着不同的工程选择。规律性是一种事前能力干预：它通过定向数据、重新加权、专用组件、检索覆盖，或针对特定故障簇的后训练，缩小最严重的“空缺”（高方差区域），从而改变底层的可靠性景观。相比之下，校准是一种事后信息与政策干预：它基本不改变底层景观，但提升系统识别自身可能出错的时机，并根据该信号调整行为的能力（例如弃答、升级至人工处理、检索优先路由、基于工具的验证，或请求澄清）。（第5节随后研究一种有代价的响应：根据有关基线不可靠性的（可能经过校准的）信息，有条件地采取该响应，即付费启用能够进一步降低错误的“推理”模式。按照我们的术语，这本身不是校准，而是由校准信号促成的一种行动。）因此，在几个常见的现实场景中，这些边际是彼此不同的：(i) 提供方可以保持模型权重不变，同时部署经过校准的风险评分和延迟处理策略，通过避免高风险查询来减少故障（有校准而无规律性）；(ii) 提供方可以通过定向微调或改进领域覆盖来填补已知空缺，同时保持面向用户的置信度和路由不变，通过让模型在这些区域表现更好来减少故障（有规律性而无校准）；(iii) 检索/工具升级可以实现其中任一边际，具体取决于它们主要是为稀疏区域增加能力（规律性），还是主要生成可靠的“未找到支持”信号以触发延迟处理（校准）。简言之，规律性改变模型在哪些地方表现良好，而校准改变系统何时选择根据自身已知信息采取行动。

形式化地，令 X 表示相邻知识点之间的间隔长度，令 X^* 表示均匀抽取的任务所面临的长度偏倚 间隔。离散度之所以重要，是因为随机遇到的任务更可能落在较长的间隔中（检查悖论），因此用户体验到的间隔分布与间隙均匀评估器所体验到的间隔分布存在系统性差异。

命题6 (经验平均间隙长度与方差**)**。 令 X 表示具有 $\mathbb{E}[X] = 1/\lambda$ 且变异系数为 变异系数 $\equiv \sqrt{\mathrm{Var}(X)}/\mathbb{E}[X]$ 的（无偏）间隔长度。令 X^* 表示

均匀抽取的任务所面临的长度偏倚的间隔长度。那么

$$\mathbb{E}[X^*] = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]} = \mathbb{E}[X]\,(1 + \mathrm{CV}^2) = \frac{1 + \mathrm{CV}^2}{\lambda}.$$

因此，

$$\mathbb{E}[\sigma^2(x)] = \frac{\mathbb{E}[X^*]}{6} = \frac{1 + \mathrm{CV}^2}{6\lambda}. \tag{13}$$

证明。长度偏倚给出 $\Pr(X^* \in dx) \propto x \Pr(X \in dx)$，因此 $\mathbb{E}[X^*] = \mathbb{E}[X^2]/\mathbb{E}[X]$。写作 $\mathbb{E}[X^2] = \mathrm{Var}(X) + (\mathbb{E}[X])^2 = (\mathrm{CV}^2 + 1)(\mathbb{E}[X])^2$，可得 $\mathbb{E}[X^*] = \mathbb{E}[X](1 + \mathrm{CV}^2)$。最后，$\sigma^2(x) = X^* t(1-t)$，其中 $t \sim \mathrm{Uniform}(0,1)$ 蕴含 $\mathbb{E}[\sigma^2] = \mathbb{E}[X^*]\mathbb{E}[t(1-t)] = \mathbb{E}[X^*]/6$。 □

方程（13）将两个与提供者相关的基本要素区分开来：

- **密度 (λ)**：整体能力/覆盖率，可通过“全面增加”（规模）来提升。
- **离散度 (CV)**：覆盖不均匀性，可通过有针对性地弥补最严重的缺口（规律性）来改善。

离散度会通过 $(1 + \mathrm{CV}^2)$ 乘数式地放大体验方差，即使平均间距 $1/\lambda$ 相同也是如此。一个便于使用的推论是体验 可靠性比率，

$$R^{\exp}(\lambda, q, \mathrm{CV}) \;\equiv\; \frac{q}{\mathbb{E}[\sigma^2(x)]} = \frac{6\lambda q}{1 + \mathrm{CV}^2} = \frac{2R}{1 + \mathrm{CV}^2}. \tag{14}$$

对于盲目信赖（在截断于 0 之前），

$$\bar{U}_B(\lambda, q, \mathrm{CV}) = 1 - \frac{\mathbb{E}[\sigma^2(x)]}{q} = 1 - \frac{1}{R^{\exp}(\lambda, q, \mathrm{CV})}. \tag{15}$$

因此，离散度会直接改变盲目采用阈值：$U_B \geq 0$ 等价于 $R \geq$（1 + 变异系数^2）/2。从操作层面看，这意味着：如果一个系统呈“尖峰式”分布（离散度高），规模扩展必须进行得更远，典型用户才会认为全面委托值得采用。

原则上，“规律性”是多维的：提供方可以通过针对特定薄弱区域调整数据、检索、工具或后训练，重塑覆盖缺口的整个分布。为便于处理，我们用单一的离散度统计量——变异系数——概括这一形状

$$\mathrm{CV} \;\equiv\; \frac{\sqrt{\mathrm{Var}(X)}}{\mathbb{E}[X]},$$

因为命题6意味着，对于盲目期望效用而言，离散度只通过变异系数影响经验平均方差：

$$\mathbb{E}[\sigma^2(x)] \;=\; \frac{1 + \mathrm{CV}^2}{6\lambda}.$$

随后，我们引入一维规则性指数 $r \in [0,1]$，它会单调地降低离散度，并在两个极端情况下进行归一化：

$\underset{=1}{\text{变异系数(0)}}$（类泊松不规则性），$\underset{=0}{\text{变异系数(1)}}$(完全规则的间距)。 (16)

为使表达式清晰易懂，我们采用线性归一化

$$\mathrm{CV}(r) = 1 - r,$$

因此，r 可以理解为“消除的泊松离散度比例”。在这一参数化下，式（13）变为

$$\mathbb{E}[\sigma^2(x)] = \frac{1 + (1-r)^2}{6\lambda}. \tag{17}$$

这一设定是一种简化，实质上将人工智能提供方视为直接投资以降低变异系数，适用于此处的分析。[11]

定义 2 (提供者投资技术**)。** 提供方选择覆盖强度 $\lambda > 0$和规律性 $r \in [0,1]$，成本为 $C(\lambda, r)$，其中 $C_\lambda > 0$、$C_r > 0$和 C分别关于每个自变量均为凸函数。

凸性在经济学上很自然：成本最低的规律性提升来自修复明显且反复出现的缺失（例如高频故障簇）；之后的提升则需要逐步投入更多数据获取、数据整理、工具开发或专门训练。许多现实中的干预会同时改变两个边际（例如，检索可以提高有效的 λ，同时通过对稀疏区域提供不成比例的帮助来改善规律性）；我们将（λ，r）分开考察，以便区分“覆盖更多”与“覆盖更均匀”的比较静态分析。

在式（17）的均值层面，规律性和规模可以相互替代（两者都通过$\lambda/(1+(1-r)^2$发挥作用），但它们的边际比率不同。

命题 7 （相对回报：规模与规律性）。 在盲模型中（以采用为条件），设 变异系数(r) 可微，且满足 变异系数$'(r) \le 0$。那么，规模与规律性之间的边际替代率为

$$\mathrm{MRS}_{\lambda,r} \;\equiv\; \frac{\partial \bar{U}_B/\partial r}{\partial \bar{U}_B/\partial \lambda} = \lambda\, \frac{-\frac{d}{dr}\big(1+\mathrm{CV}(r)^2\big)}{1+\mathrm{CV}(r)^2} = \lambda\, \frac{-2\mathrm{CV}(r)\mathrm{CV}'(r)}{1+\mathrm{CV}(r)^2}.$$

在线性归一化 变异系数$(r) = 1-r$下，该式可简化为

$$\mathrm{MRS}_{\lambda,r} = \frac{2(1-r)\lambda}{1+(1-r)^2},$$

它随 λ 增加，并在 $r \to 1$时趋于0 。

证明。 截断前的盲目期望效用为 $\bar{U}_B(\lambda,\ q,\ r) = 1 - \mathbb{E}[\sigma^2(x)]/q$。根据命题6,

[11]这种设定适用于此处使用的对象，因为在盲目信赖下，福利仅通过 $\mathbb{E}[\sigma^2(x)]$, 取决于覆盖形态，而式（13）表明变异系数是相关的汇总统计量。任何满足式（16）的可微递减映射$r \mapsto$ 变异系数(r)都将得出相同的定性比较静态分析；线性选择则确定了单位，并能得出闭式表达式。我们因此舍弃的是：变异系数属于二阶矩汇总统计量。两种间隔分布可以有相同的（λ，变异系数），但其 *tails* 可能大不相同（例如，罕见的灾难性间隔与较温和的离散）。这在校准下尤为重要，因为福利和使用取决于 $\sigma^2(x)$的完整分布，并受到临界值 q处截断的影响。因此，应将 r理解为捕捉通过二阶矩（等价地说，通过 $\mathbb{E}[X^2]$）发挥作用的“规律性”成分，而非对尾部工程的完整描述。最后，这一简化对 $r \in [0,1]$ 作出限制，由此可得变异系数 $\in [0,1]$，即覆盖的离散程度不高于泊松基准。如果我们将类泊松间距视为自然的高离散度基线，并将正则性投资理解为降低离散度，那么这种设定是恰当的。如果某些环境具有重尾间隔（变异系数 > 1），则可通过允许 $r < 0$或使用另一种单调映射（例如，对于 $r > -1$，变异系数$(r) = 1/(1+r)$）来扩展分析，而不改变其经济逻辑。

$\mathbb{E}[\sigma^2(x)] = (1 + \text{变异系数}(r)^2)/(6\lambda)$，因此

$$\bar{U}_B(\lambda, q, r) = 1 - \frac{1 + \mathrm{CV}(r)^2}{6\lambda q}.$$

求导得到 $\partial \bar{U}_B/\partial r = -\frac{1}{6\lambda q}\frac{d}{dr}(1 + \mathrm{CV}(r)^2)$ 和 $\partial \bar{U}_B/\partial \lambda = {}^{1+}\mathrm{CV}(\quad r\ ^2)$。取其比值即可得到一般表达式。代入 $\mathrm{CV}(r) = 1 - r$ 后，可得封闭形式及其比较静态分析。

$6q\lambda^2$

□

命题 7 阐明了一维规则性指数在经济学上的作用：它追踪提供者努力压缩离散度的速度（通过 $\mathrm{CV}'(r)$），因此也反映了随着密度进一步增加，体验二阶矩 $\mathbb{E}[X^2]$ 下降的速度。无论具体采用何种参数化，这一质性结论都具有稳健性：随着 λ 增大，且平均间隔已经较短，填补剩余不规则空缺的边际价值相对于再增加一个“典型”点会有所上升，尽管两者都会降低均值方差。

4.4 边际如何相互作用：互补品与替代品

这三个边际相互作用，是因为它们针对的是不同的摩擦。规模扩展能在各处降低不确定性；规律性降低离散度和尾部暴露；校准则让用户依据局部可靠性作出条件判断，从而改变行为。AJI 表明，边际之间是互补品还是替代品，取决于哪个边际受到约束。

规模扩展 与 校准： 比较（8）和（9）可以看出，两者存在鲜明的交互作用：

• 对于 $R < 1$，盲目用户不会采用，且 $\partial U_B/\partial R = 0$；而校准用户可获得$\partial U_C/\partial R > 0$。校准通过让边际规模扩展即使在盲目采用阈值以下也能带来收益，成为规模扩展的互补品 。

• 对于 $R > 1$，盲目用户已经会在所有情况下委托，因此边际规模扩展的价值为 $1/R^2$；在校准下，$\partial U_C/\partial R < 1/R^2$ ，因为部分负价值尾部已经被规避。在广泛采用的前提下，校准在边际上是规模扩展的替代品 ，因为它会减弱进一步扩大规模所带来的增量收益。

因此，“规模报酬”并非纯粹的技术斜率：它是一种行为对象，取决于用户能够观察到什么，以及他们如何作出反应。

规模与规律性： 固定 q 后，较高的 λ 和较高的 r 都会提高体验可靠性比率 $R^{\exp} = 6\lambda q/(1 + (1-r)^2)$，因此，对于平均（盲目）表现而言，它们是替代品。但由于 λ 的边际回报以 $1/\lambda^2$ 的速度递减，而 r 的边际回报以 $1/\lambda$的速度递减，命题 7表明，规律性的相对回报会随着规模扩大而上升。这正是“规模已经大幅扩大后，弥补剩余漏洞就成了高杠杆举措”这句话的正式含义。

校准 与 规律性： 校准通过筛选缓解事后 的缺口；规律性则通过缩减高方差尾部，缓解事前 的缺口。盲视成本恒等式 (12) 清楚地揭示了这种替代关系：由于 $x \mapsto (x - q)_+$ 是凸函数，分布 $\sigma^2(x)$的任何保均值收缩都会降低校准的回报。

引理 2（尾部收缩时，规律性会降低校准的价值）。 固定(λ, q)。如果规律性的提高使 $\sigma^2(x)$ 转变为凸序意义下的保均值收缩（即，对于每个凸函数 ϕ ，它弱降低 $\mathbb{E}[\phi(\sigma^2(x))]$ ，同时保持 $\mathbb{E}[\sigma^2(x)]$ 不变），那么盲用成本 $\Delta_B(\lambda, q)$ 弱降低。

证明。 根据（12）， $\Delta_B(\lambda, q) = \frac{1}{q}\mathbb{E}[(\sigma^2(x) - q)_+]$。函数 $s \mapsto (s-q)_+$ 是凸函数，因此在凸序下进行保均值收缩时，其期望弱递减。乘以 $1/q$ 后不等式方向保持不变。

□

引理 2 将一个简单的思想形式化：校准和规律性都是尾部风险工具，因此改善其中一项会降低另一项的边际价值。同时，校准可以通过防止盲目采用的广延边际崩塌，提高低-R 定义域中规律性的实现收益：没有校准时，用户可能根本不会采用，从而即使大幅改善规律性，在私人层面也无关紧要。

4.5 基准测试、测量失准的回报与经 AJI 调整的评估 n

AJI 同样是一个测量问题：常见的基准测试通常以与用户接触任务空间的方式不相匹配的方式，对任务或定义域加权。用模型的术语来说，基准测试套件通常更接近于从间隔均匀（或以其他方式按密集区域加权的）分布中抽样，而现实世界中的使用则从整个空间中抽取任务，因此会对最稀疏区域产生长度偏倚的暴露度。由于用户在稀疏区域具有长度偏倚，忽略长度偏倚的基准测试可能会误述投资的水平和边际回报。

一种实际上对间隔均匀加权的基准测试机制（每个间隔/定义域一个项目）评估的是“无偏”的间隔长度 X ，而不是用户的长度偏倚 X^*。由于间隔内平均后验方差为 $X/6$,

$$\mathbb{E}_{\text{bench}}[\sigma^2] = \frac{\mathbb{E}[X]}{6} = \frac{1}{6\lambda}, \qquad \mathbb{E}_{\text{user}}[\sigma^2] = \frac{\mathbb{E}[X^*]}{6} = \frac{1+\text{CV}^2}{6\lambda}.$$

等价地，隐含的可靠性比率满足

$$\frac{R^{\text{exp}}}{R^{\text{bench}}} = \frac{q/\mathbb{E}_{\text{user}}[\sigma^2]}{q/\mathbb{E}_{\text{bench}}[\sigma^2]} = \frac{1}{1+\text{CV}^2} \le 1,$$

仅当间距完全规则时（变异系数 $=0$）才取等号。在泊松基准中（变异系数 $=1$），该扭曲恰好为 $1/2$：间隔均匀基准对可靠性的乐观程度是代表性用户体验的两倍。

If r从基准斜率推断规模扩展收益时，AJI 会产生两种扭曲 :

(i) 广延边际扭曲（采用）。忽略长度偏倚的基准测试机制会过早预测采用：它们在 $\mathbb{E}[X]$ 进行评估，而用户体验到的是 $\mathbb{E}[X^*]$。

(ii) 边际回报扭曲。即使行为保持不变，基准测试曲线的斜率也会夸大规模扩展带来的边际改善，因为间隔均匀基准实际上评估的是$q/\mathbb{E}_{\text{bench}}[\sigma^2]$ ，而用户体验则由 $q/\mathbb{E}_{\text{user}}[\sigma^2]$决定。由于 $\mathbb{E}_{\text{user}}[\sigma^2] = (1 +$

$\mathrm{CV}^2)\mathbb{E}_{\mathrm{bench}}[\sigma^2]$，体验可靠性比率为 $R^{\mathrm{exp}} = R^{\mathrm{bench}}/(1+\mathrm{CV}^2)$，因此，提高 λ 带来的边际增益会被因子 $1/(1+\mathrm{CV}^2)$削弱。

该模型表明应立即进行一项修正：报告（或以.....为目标）间隔长度的二阶原始矩， $\mathbb{E}[X^2]$（连同 $\mathbb{E}[X]$），因为任务均匀使用会产生长度偏倚，而经验平均间隔满足 $\mathbb{E}[X^*] = \mathbb{E}[X^2]/\mathbb{E}[X]$。在操作层面，这意味着可以（i）通过从任务空间中均匀抽取任务来评估（这会自动产生长度偏倚）；或者（ii）使用估计的离散因子（1+变异系数^2）或其经验对应值，调整间隔均匀基准估计值。附录 B 回顾了大语言模型常用的基准测试，并着重说明该模型对其解读的影响，以及如何调整这些基准测试的建议。

为清楚说明其对投资的影响，考虑一个长度为 L、间隔为（X_i）$_{i=1}^{n}$、 $\sum_i X_i = L$ 的有限任务空间。均匀抽取的任务以 X_i/L 的概率落入间隔 i，且间隔内平均方差为 $X_i/6$，因此经验平均方差为

$$\mathbb{E}[\sigma^2] = \sum_{i=1}^{n} \frac{X_i}{L} \cdot \frac{X_i}{6} = \frac{1}{6L} \sum_{i=1}^{n} X_i^2. \tag{18}$$

相比之下，间隔均匀基准实际上评估的是

$$\mathbb{E}_{\text{gap-unif}}[\sigma^2] = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{6} = \frac{L}{6n},$$

它仅取决于 n，且不受（X_i）中的离散度影响。

命题 8（边际知识点发挥作用的情形）。 假设提供方可以在选定的、长度为 X 的间隔内添加一个知识点，将其分成长度分别为 αX 和$(1-\alpha)X$ 的两个间隔，其中 $\alpha \in (0,\ 1)$。那么经验平均方差(18) 的减少量为

$$\Delta\mathbb{E}[\sigma^2] = \frac{1}{6L}\Big(X^2 - (\alpha^2 + (1-\alpha)^2)X^2\Big) = \frac{1}{3L}\alpha(1-\alpha)X^2,$$

当间隔平分时，该减少量最大（$\alpha = 1/2$），且随原始间隔长度 X 增大而增大。特别是，为了使边际知识点带来的经验质量增益最大化，提供方应拆分最长的间隔。

证明。 根据 (18)，经验平均方差与 $\sum_i X_i^2$ 成正比。将长度为 X 的间隔拆分后， X^2 被替换为 $(\alpha X)^2 + ((1-\alpha)X)^2 = (\alpha^2 + (1-\alpha)^2)X^2$，从而得到所述减少量。函数 $\alpha(1-\alpha)$ 在 $\alpha = 1/2$ 处取最大值，且减少量随 X^2 变化。[12]

□

命题 8 突显了激励问题：在间隔均匀基准下，无论是在灾难性空缺处填补一点，还是在已经密集的区域进一步细化，只要在任何位置增加一个点，就会使 n 增加 1，并以相同幅度提高分数。在基于体验（按使用加权）的目标下，边际价值随 X^2变化，因此填补最严重空缺具有一阶价值。简言之，AJI 不只是改变了衡量性能的水平，还改变了哪些投资看起来回报较高，进而影响哪些改进更有可能得到提供。

[12]在他们的 对“知识深化”的研究中，Carnehl and Schneider（2025年）证明了一个类似的结果，即在中点添加知识是处的突出点。

经过校准后，“使用情况”是内生的：用户恰恰会在高方差区域弃答。因此，与福利一致的评估应同时报告 (i) 扣除弃答后的价值 U_C，以及 (ii) 覆盖率 s_C （式 (10) 中的使用份额）。否则，模型完全可能通过隐含地拒绝在其表现崎岖的区域作答，来营造良好表现。

5 推理模式与崎岖景观

许多人工智能系统现在提供多种推断模式：廉价的“默认”模型，以及更昂贵的“推理”（或“思考”）模式，后者使用额外的计算资源（通常也会增加延迟）来减少错误。在平滑的世界中，这是标准的质量—价格权衡。在崎岖的世界中，它则变成了一个选择 问题：为推理付费的价值，恰恰在局部困难的任务上最高；根据命题 1，用户接触这些任务的程度却超出其应有比例。本节将介绍一个关于推理的简单基准模型，并研究它如何与锯齿性、检查悖论以及规模报酬的测量相互作用。

5.1 设置：两种推断模式

对于每个任务 x，用户可以选择不作答（外部选项为 0）、使用廉价的“快速”模式（F），或调用成本更高的“推理”或“思考”模式（T）。快速模式的效用是公式 (2) 中的基准收益，

$$U_F(x) \;\equiv\; 1 - \frac{\sigma^2(x)}{q}. \tag{19}$$

推理会产生额外的每任务成本 $\kappa > 0$ （以时间、金钱或注意力支付），并将局部均方误差从 σ^2（x）降低至 σ_T^2（x）：由于收益按（2）式进行归一化，κ 以相同单位衡量：假设推理输出的误差为零，则净效用为 $1 - \kappa$。

$$U_T(x) \;\equiv\; 1 - \frac{\sigma_T^2(x)}{q} - \kappa. \tag{20}$$

因此，相对于快速模式，推理的每任务价值为

$$V_T(x) \;\equiv\; U_T(x) - U_F(x) = \frac{\sigma^2(x) - \sigma_T^2(x)}{q} - \kappa. \tag{21}$$

（19）—（21）有两个在经济学上至关重要的特点。首先，推理是一种推断侧 边际投入：它按任务付费，而非预先支付，因此锯齿性会通过任务到达分布影响实际产生的“推理账单”。其次，只有当用户能够将推理针对 σ^2（x）较高的任务时，才能获得推理的全部价值；这直接将推理与校准联系起来（第2.6节）。

5.2 基准模型：作为局部噪声证据的推理

关键的建模选择在于推理如何改变 $\sigma^2(x)$。一种实用的基准做法是，将推理视为获取一条额外的、任务特定的证据；该证据能够提供有关$Y(x)$的信息，但并不完美。这涵盖了多种机制（检索、工具使用、更长时间的搜索、自一致性），同时保持模型透明。

定义 3 (推理作为有噪声的伪锚点)。 当在任务 x 中调用推理时，模型会获得一个额外信号

$$\tilde{Y}(x) = Y(x) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2),$$

独立于基线信息集。因此，得到的后验方差为

$$\sigma_T^2(x) = \left(\frac{1}{\sigma^2(x)} + \frac{1}{\sigma_\varepsilon^2} \right)^{-1} = \frac{\sigma^2(x)\,\sigma_\varepsilon^2}{\sigma^2(x) + \sigma_\varepsilon^2}. \tag{22}$$

定义 3 可以按字面理解（有噪声的辅助估计），也可以视为一种简化形式：通过“额外计算资源”生成答案分布的一个额外独立样本。参数 σ_ε^2 是推理噪声下限：即使基线不确定性无限大，推理也无法将方差降低到低于 σ_ε^2。

命题 9 （推理如何重塑局部风险）。根据定义 *3*:

(i) σ_T^2（x）关于 σ^2（x）递增且为凹函数，其中 σ_T^2（x）$\leq$最小值$\{\sigma^2$（x） $\sigma_\varepsilon^2\}$，且极限σ^2（x）$\to\infty$ σ_T^2（x） $=\sigma_\varepsilon^2$。

(ii) 方差降低

$$\Delta(x) \;\equiv\; \sigma^2(x) - \sigma_T^2(x) = \frac{\sigma^4(x)}{\sigma^2(x) + \sigma_\varepsilon^2}$$

关于 σ^2（x）递增且为凸函数。

(iii) 因此，推理使景观呈现出“崎岖但有上限”的特征：较长的间隔仍会产生较高的基准方差，但推理会将这些任务上的实现方差限制在（约） σ_ε^2 。

证明。 写作 $v \equiv \sigma^2(x)$。由 (22) 可得， $\sigma_T^2(v) = v\sigma_\varepsilon^2/(v + \sigma_\varepsilon^2)$。于是

$$\frac{d\sigma_T^2}{dv} = \frac{\sigma_\varepsilon^4}{(v + \sigma_\varepsilon^2)^2} > 0, \qquad \frac{d^2\sigma_T^2}{dv^2} = -\frac{2\sigma_\varepsilon^4}{(v + \sigma_\varepsilon^2)^3} < 0,$$

由此确立了单调性和凹性，而界限和极限则可直接观察得出。该约简为 $\Delta(v) = v - \sigma_T^2(v) = v^2/(v + \sigma_\varepsilon^2)$，满足

$$\frac{d\Delta}{dv} = \frac{v(v + 2\sigma_\varepsilon^2)}{(v + \sigma_\varepsilon^2)^2} > 0, \qquad \frac{d^2\Delta}{dv^2} = \frac{2\sigma_\varepsilon^4}{(v + \sigma_\varepsilon^2)^3} > 0,$$

因此 Δ 递增且为凸函数。 □

相反，如果推理按比例降低方差， $\sigma_T^2(x) = \theta\sigma^2(x)$，其中 $\theta \in (0, 1)$，那么推理会保留锯齿性的形状参数 （它会缩放整个景观），而第 4 节中的检查悖论乘数则保持不变。定义 3 则刻意采用了不同的设定：它捕捉到一种经验上常见的模式，即推理在高不确定性区域带来的帮助更大，但效果存在下限。

5.3 校准下的最优推理

与第2.6节一样，考虑一位校准用户，他在决定是否使用人工智能以及运行哪种模式之前，会观察 $\sigma^2(x)$。[13]

命题 10（推理的阈值规则）。 根据定义 *3*，校准用户按照基准方差中的临界值规则进行选择

$$\max\{0, U_F(x), U_T(x)\}$$

by a cutoff rule in the baseline variance $v \equiv \sigma^2(x)$。定义快速模式的盈亏平衡 点 $v_0 \equiv q$，将推理进入阈值 v_L 定义为方程 $U_T(v) = U_F(v)$ 的唯一正解，

$$v_L \equiv \frac{\kappa q + \sqrt{\kappa^2 q^2 + 4\kappa q \sigma_\varepsilon^2}}{2}, \tag{23}$$

并将推理盈亏平衡 点 v_H（如果存在）定义为 $U_T(v_H) = 0$，

$$v_H \equiv \begin{cases} +\infty, & \sigma_\varepsilon^2 \le (1-\kappa)q, \\ \dfrac{(1-\kappa)q\,\sigma_\varepsilon^2}{\sigma_\varepsilon^2 - (1-\kappa)q}, & \sigma_\varepsilon^2 > (1-\kappa)q. \end{cases} \tag{24}$$

那么：

(*i*) 如果 $U_T(q) < 0$（等价地，$\kappa > \frac{q}{q+\sigma_\varepsilon^2}$），推理从来都不是最优的，并且当且仅当 $v \le q$ 时，用户使用快速模式；否则选择弃权。

(ii) 如果 $U_T(q) \ge 0$（等价地，$\kappa \le \frac{q}{q+\sigma_\varepsilon^2}$），那么 $v_L \le q \le v_H$，最优政策为：当 $v < v_L$ 时采用快速模式，当 $v \in [v_L, v_H]$ 时进行推理，当 $v > v_H$ 时弃答（当 $v_H = +\infty$ 时不弃答）。

证明。 记 $v = \sigma^2(x)$ 和 $\xi \equiv \sigma_\varepsilon^2$。由（19）–（22）可得，

$$U_F(v) = 1 - \frac{v}{q}, \qquad U_T(v) = 1 - \frac{v\xi}{q(v+\xi)} - \kappa.$$

二者均随 v 递减，且 $U_T(v)$ 的下界为 $1 - \xi/q - \kappa$。此外，

$$U_T(v) - U_F(v) = \frac{v^2}{q(v+\xi)} - \kappa,$$

该函数随 v 严格递增（当 $v > 0$ 时其导数为正）。因此存在唯一的交点 v_L，在该点 $U_T(v) = U_F(v)$；该交点由二次函数 $v^2 - \kappa q\, v - \kappa q\, \xi = 0$ 解得，并得到（23）。

如果 $U_T(q) < 0$，那么对所有 $v \ge q$，$U_T(v) < 0$（因为 U_T递减），且对所有 $v \le q$，$U_T(v) < U_F(v)$（因为 $U_T - U_F$ 递增，且在 q处为负），因此推理永远不是最优选择。如果 $U_T(q) \ge 0$，那么 $U_T - U_F$在 q之前或恰好于此处穿过零点，因此 $v_L \le q$，且对于所有情况，推理都优于快速

[13]这是一个基准：在实践中，用户拥有的是部分信号（置信度分数、自我评估、启发式方法），而不是直接获取 $\sigma^2(x)$ 的权限。校准情形的意义在于，在理想化目标定位下，单独考察推理选项的经济学。

模式。仅当且唯当 $U_T(v) \geq 0$时，推理才优于弃答；这定义了式 (24) 中的 v_H。综合这些比较，即可得出所述政策。 □

方程（23）表明，当基线方差足够高、使方差降低值得支付附加费用 κ 时，校准用户会购买推理。方程（24）清楚地揭示了噪声底限逻辑：当 σ_ε^2 相对于利害关系 q 较大时，即使进行推理也无法使任务足够安全，因此对于足够大的 v，用户会选择放弃。当 $\sigma_\varepsilon^2 \leq (1-\kappa)q$ 时，最困难的任务会通过付费进行推理来处理，而不是选择放弃。

推论1（比较静态分析）。 在命题 *10* 下：

(i) v_L 随 κ（推理成本更高）、q（利害关系较低）以及 σ_ε^2（推理质量较差）的增加而增加。

(ii) 当 $v_H < +\infty$ 时，可行性临界值 v_H 随 q 增加，并随 κ 和 σ_ε^2 减少。

5.4 盲选与选择性推理的期权价值

盲用户无法根据 $\sigma^2(x)$ 进行条件化，必须将推断模式作为一揽子策略来选择，¯这与第 2 节中的盲目采用类似。令 $U_F(R) = 1 - 1/R$ 为快速模式的预期收益（第 4 节）。根据定义 3，始终进行推理所带来的预期收益为

$$\bar{U}_T(R, \kappa, \sigma_\varepsilon^2) \;\equiv\; \mathbb{E}[U_T(x)] = 1 - \kappa - \frac{\mathbb{E}[\sigma_T^2(x)]}{q}. \tag{25}$$

在布朗–泊松基准（第 2–3 节）中，将其表示为式 (6) 定义的无尺度归一化方差 $Z \equiv 3\lambda\sigma^2(x)$ 会很方便；在泊松覆盖下，其分布具有尺度不变性（引理1）。令 $\rho \equiv 3\lambda\sigma_\varepsilon^2$ 表示以相同单位计的推理噪声下限。于是，(22) 意味着

$$Z_T \;\equiv\; 3\lambda\sigma_T^2(x) = \frac{Z\rho}{Z+\rho}, \qquad \text{and hence} \qquad \bar{U}_T(R,\kappa,\rho) = 1 - \kappa - \frac{1}{R}\mathbb{E}\left[\frac{Z\rho}{Z+\rho}\right]. \tag{26}$$

盲用户选择最佳的一揽子规则，

$$U_{\text{modes}}^{\text{blind}}(R,\kappa,\rho) \;\equiv\; \max\Big\{0,\; \bar{U}_F(R),\; \bar{U}_T(R,\kappa,\rho)\Big\}.$$

相比之下，校准用户可以针对每项任务选择最佳模式，从而获得

$$U_{\text{modes}}^{\text{cal}}(R,\kappa,\rho) \;\equiv\; \mathbb{E}[\max\{0,\; U_F(x),\; U_T(x)\}]\,.$$

间隔 $U_{\text{modes}}^{\text{cal}} - U_{\text{盲模式}}$ 是进行选择性 推理的期权价值。

命题 11（锯齿性为推理创造期权价值）。 对于任意联合分布$(\sigma^2(x)\;\sigma_T^2(x))$，

$$U_{modes}^{cal} \;\geq\; U_{modes}^{blind},$$

当且仅当 *(i)* $\mathrm{Var}(\sigma^2(x)) \;>\; 0$ ，并且 *(ii)* 以正概率在 $\sigma^2(x)$ 的不同实现中，不同模式分别达到最优时，不等式严格成立。

证明。 函数 $g(a,b) \equiv \max\{0,a,b\}$ 是 (a,b) 的凸函数，因为它是仿射函数的逐点最大值。因此，根据詹森不等式，有

$$\mathbb{E}[g(U_F(x),U_T(x))] \;\geq\; g(\mathbb{E}[U_F(x)],\mathbb{E}[U_T(x)]).$$

左侧为 $U^{\mathrm{cal}}_{\mathrm{modes}}$，右侧等于 $\max\{0,\bar{U}_F,\bar{U}_T\} = U^{\mathrm{blind}}_{\mathrm{modes}}$。只要 g 在 $(U_F,\ U_T)$ 的支撑集上不是仿射函数，不等式就严格成立；也就是说，只要最大值点会随状态而变化。

□

命题 11 是第2.6节中校准价值的“推理”类比。如果任务是同质的（$\sigma^2(x)$ 为常数），那么一概采用推理，与按任务级选择同样有效。锯齿性会造成 $\sigma^2(x)$ 的离散度，使逐任务推理决策具有价值：用户只需在需要推理的任务上支付 κ 。

5.5 推理与检查悖论

第 3 节表明，用户体验会因较长的间隔而产生长度偏倚（命题 1）。由于推理的直接收益随局部不确定性上升（命题 9），检查悖论会系统性地改变实际推理价值，使之相对于低估长间隔的评估结果发生偏移。

为了清晰地进行比较，考虑泊松基线下针对间隔的两种抽样方案。随机抽取间隔 的“基准”看到的是 $X \sim \mathrm{Exp}(\lambda)$，而随机抽取位置 的用户看到的是长度偏倚间隔 $X^* \sim$ 伽马分布$(2,\lambda)$（命题 1）。给定长度为 X 的间隔以及间隔内相对位置为 $t \sim$ 均匀分布(0,1)，基线方差为 $v = Xt(1-t)$，推理方差由式（22）给出。

定理 2 （推理价值的检查悖论放大效应）。 固定 $(q,\kappa,\sigma_\varepsilon^2)$ 并定义总方差缩减 $\Delta(v) = v - \frac{v\sigma_\varepsilon^2}{v+\sigma_\varepsilon^2}$。令

$$\bar{\Delta}^{gap} \;\equiv\; \mathbb{E}\big[\Delta\big(Xt(1-t)\big)\big]\,, \qquad \bar{\Delta}^{use} \;\equiv\; \mathbb{E}\big[\Delta\big(X^*t(1-t)\big)\big]\,,$$

其中，在两种期望中，$t \sim$ 均匀分布$(0,1)$ 均与间隔长度相互独立，而 $X \sim Exp(\lambda)$，而 $X^* \sim$ 伽马分布$(2,\lambda)$。于是 $\Delta^{use} > \Delta^{gap}$。等价地，对于任意固定的推理成本 κ，间隔均匀抽样评估会低估实际使用中推理带来的预期净收益 $\mathbb{E}[V_T(x)]$ 。

证明。 对于每个固定的 $t \in (0,1)$，映射 $x \mapsto \Delta(x\,t(1-t))$ 关于 x 严格递增。此外，在泊松情形下，X^* 在一阶随机意义上支配 X：对所有 $z>0$，$\Pr(X^*>z) = e^{-\lambda z}(1+\lambda z) > \Pr(X>z) = e^{-\lambda z}$。因此，对于每个 t，都有 $\mathbb{E}[\Delta(X^*t(1-t)) \mid t] > \mathbb{E}[\Delta(Xt(1-t)) \mid t]$，对 t 积分即可得到所要证明的结论。 □

定理 2 是第 4 节中“基准测试楔形”在推理方面的对应概念。基准测试低估了最长间隔，因此低估了预期 高成本推理的价值，进而低估了推理质量改进带来的福利增益。在最优推理区域属于上尾的特殊情况下（即命题 10 中的 $v_H = +\infty$ ），基准测试还低估了用户最优选择推理的频率。

5.6 与规模和规律性的交互

推理引入了一种在性质上不同于训练时规模的设计余量。规模（λ）在所有地方都降低 $\sigma^2(x)$（命题 3），而推理则有条件地 降低方差，并且每项任务都需要付出成本。二者通过局部方差的分布相互作用：规模扩展改变了用户遇到 v 高于推理进入阈值v_L 的任务的频率（命题 10），而推理改变了剩余长间隔尾部的成本。

一个简单的理解方式是考察 (26) 中 U_T 的极限行为。由于 Z 在泊松覆盖下具有无尺度性，始终进行推理的表现对 λ 的依赖由 $\rho = 3\lambda\sigma_\varepsilon^2$ 驱动。

命题 12（泊松覆盖下始终推理的极限）。在布朗–泊松基准中，如果用户对每项任务都进行推理，则：

（*i*）（稀疏规模）当 $\lambda \to 0$（等价地，当 $\rho \to 0$）时，

$$\mathbb{E}[\sigma_T^2(x)] \to \sigma_\varepsilon^2 \qquad and \qquad \bar{U}_T \to 1 - \kappa - \frac{\sigma_\varepsilon^2}{q}.$$

（*ii*）（稠密规模）当 $\lambda \to \infty$（等价地，当 $\rho \to \infty$）时，

$$\mathbb{E}[\sigma_T^2(x)] \sim \frac{1}{3\lambda} \qquad and \qquad \bar{U}_T \sim 1 - \kappa - \frac{1}{R}.$$

证明。利用 (26)，

$$\mathbb{E}[\sigma_T^2(x)] = \frac{1}{3\lambda}\mathbb{E}\left[\frac{Z\rho}{Z+\rho}\right] = \sigma_\varepsilon^2\,\mathbb{E}\left[\frac{Z}{Z+\rho}\right].$$

由于 $\rho \to 0$，$\frac{Z}{Z+\rho} \uparrow 1$ 逐点成立且被 1 所界定，因此控制收敛给出 $\mathbb{E}\left[\frac{Z}{Z+\rho}\right] \to 1$，从而 $\mathbb{E}[\sigma_T^2(x)] \to \sigma_\varepsilon^2$。由于 $\rho \to \infty$，$\frac{Z\rho}{Z+\rho} \to Z$ 逐点成立且由 Z 控制。根据控制收敛，$\mathbb{E}\left[\frac{Z\rho}{Z+\rho}\right] \to \mathbb{E}[Z] = 1$，因此

$$\mathbb{E}[\sigma_T^2(x)] = \frac{1}{3\lambda}\mathbb{E}\left[\frac{Z\rho}{Z+\rho}\right] \sim \frac{1}{3\lambda}.$$

由于 $\rho = 3\lambda\sigma_\varepsilon^2$，这正是当 $\lambda \to \infty$ 时的稠密规模极限（保持 σ_ε^2 不变）。代入 (25) 即得所述极限。 □

命题 12 表明，在低-λ 制度下，推理可以部分替代训练时规模：它通过施加误差下限 σ_ε^2，防止长间隔尾部爆炸。但在高-λ 制度下，推理并不能“击败”规模：一旦典型的 $\sigma^2(x)$远低于推理误差下限，推理除了施加每任务成本 κ之外，几乎没有其他作用。

规律性也会产生类似的相互作用。在没有推理的情况下，经验平均方差随间隔长度的二阶矩变化（第 4 节）。在存在推理误差下限时，极长间隔每单位任务空间长度的贡献约为 σ_ε^2，而不会按二次函数增长，因此，在集约边际上，针对最长间隔进行设计所带来的福利回报会减弱。同时，如果推理成本高昂且仅被选择性地使用（命题 10），提高规律性会减少用户进入长间隔尾部、从而必须支付 κ 的频率。这就形成了一个自然的

三方设计空间：训练时规模（λ）、覆盖规律性（第 4 节）和推理端推理质量（κ，σ_ε^2）。

总之，推理并不会消除锯齿性，而是改变其经济几何形态。它限制了最长间隔造成的损失，但检查悖论意味着，这些间隔仍然主导着（i）推理被调用的频率，以及（ii）失败显得多么突出。

6 掌握：学习局部可靠性

校准（第2.6节）是一项信息基准：用户观察到 $\sigma^2(x)$，并可以对局部不可靠的任务选择弃答。在大多数部署中，用户无法获得如此清晰的可靠性信号。相反，他们会投入精力来掌握：通过反复使用、试验和验证，学习哪些任务表述方式是安全的，哪些容易出问题，以及成功模式能够推广到多大范围。

本节将掌握形式化为一个学习问题，随后研究它与规模扩展投资之间的相互作用。关键并不在于掌握“优于”规模，或反之，而在于规模扩展影响技术，而掌握影响可发现性：规模扩展可能确实带来改进，但如果改进发生在用户已经学会避开或无法推断的区域，那么这些改进在经济上仍可能潜藏。

6.1 掌握度作为可靠性映射的学习

在本节中，下标 $t = 1$、2、.....是序贯评估的时间索引。这与前面在计算间隔内部平均值时使用的间隔内坐标 $t \in [0, 1]$ 无关。

固定一个模型版本（固定的 λ）。令 $\mathcal{Z}$ 表示用户能够诱导出的任务表示空间：提示、工具配置、检索设置、输入格式，以及任何其他可能使实际交互转移到任务空间中不同位置的设计自由度。[14]

令 $V\colon \mathcal{Z} \to \mathbb{R}_+$ 表示模型的局部均方误差（或更一般地说，局部不可靠性指数）。在布朗–泊松基准中，$V(x) = \sigma^2(x)$ 是由隐藏的知识点配置所诱导的局部后验方差。用户无法直接观测到 $V(x)$，必须通过学习来了解它。

从此处开始，$V(\cdot)$ 被视为任意的（可能呈锯齿状的）可靠性映射。在布朗–泊松基准中，$V(x) = \sigma^2(x)$ 由隐藏的知识空白生成；在实际应用中，它也可以反映任何依赖任务的误差过程。假设 5 是关于用户对 $V(\cdot)$ 的信念的模型（且可能存在设定错误）；信息增益项 γ_t 是核函数 k 的性质，它决定了主观不确定性在反馈作用下缩减的速度。

假设4 (反馈和验证噪声**)。** 当用户在 $x_t \in \mathcal{Z}$ 处评估模型（例如通过对照真实值进行验证或运行审计测试）时，他们观察到

$$y_t = V(x_t) + \eta_t,$$

[14]这一解读很重要：掌握并非只是在固定任务中“知道是否该信任模型”，还包括学习如何通过改写、拆解、补充上下文或更换工具，将任务 $move$ 到更可靠的区域。将此建模为模式 $x \in \mathcal{Z}$ 的选择，是一种简化形式，可用于涵盖提示词库、工作流工程和标准操作规程。

其中 $(\eta_t)^{t\geq 1}$ 在给定 $(x_s)^{s\geq 1}$ 和 V 的条件下独立同分布，且 $\eta_t \sim \mathcal{N}(0,\ \sigma_n^2)$。

即使在具有"客观"评估的情境中，可靠性也会受到噪声影响：同一提示模板下的任务各不相同，评估并不完美，而且组织使用代理性检查（抽样审计、共识检查或下游错误信号），而非完整的真实值。为便于分析，我们假设噪声服从高斯分布；此外，定理 3 中使用的互信息恒等式在高斯似然下是精确成立的。

假设 5 (信念：高斯过程先验**)**。 用户将 $V(\cdot)$ 建模为来自具有有界核函数 $k(x,x)\ \leq 1$ 的高斯过程先验 $V \sim \mathcal{GP}(0,k)$ 的样本。

高斯过程（或 GP）先验是一种易于处理的表示方式，用来刻画老练用户的信念："相似的任务具有相似的可靠性。"核函数 k 编码了用户对相似性的判断：提示词中只改动几个词可能被视为小幅扰动（高 k），而切换定义域或工具则可能被视为本质上毫不相关（低 k）。这不仅是为了统计上的便利：界面设计（模板、工具可供性、工作流结构）和组织实践（提示词库、输入模式）会通过决定用户认为哪些转换有意义且可迁移来塑造 k 。

令 $\mu_t(x) \equiv \mathbb{E}[V(x)\ |\ \mathcal{F}_t]$ 和 $s_t^2(x)\ \equiv \mathrm{Var}(V(x)\ |\ \mathcal{F}_t)$ 分别表示经过 t 次反馈观测后的高斯过程后验均值和方差，其中 $\mathcal{F}_t$ 是由历史记录生成的σ-代数。为避免与模型的局部误差 $\sigma^2(x)$ 混淆，我们用 $\sigma^2(x)$ 表示真实的局部误差映射，并用 $s_t^2(x)$ 表示用户对该映射的不确定性。

衡量掌握程度的一个自然指标是最坏情况不确定性：

$$m_t\ \equiv\ \sup_{x\in\mathcal{Z}} s_t^2(x). \tag{27}$$

较大的 m_t 意味着仍有一些区域，用户无法可靠地区分安全委托与不安全委托。

6.2 一个乐观的基准：掌握能够多快学会?

如果用　户愿意进行有目的的实验，一种简单的策略是不确定性采样　　：

$$x_{t+1} \in \arg\max_{x\in\mathcal{Z}} s_t^2(x). \tag{28}$$

这将系统性压力测试形式化：在最不确定的地方进行探测。这是一个乐观的基准，因为它假设用户能够承担采样潜在不安全点的代价；在高风险场景中，这一假设并不成立（而这种失效恰恰是它与规模扩展之间的互动变得有趣之处）。

定义最大信息增益在 t 次采样后，

$$\gamma_t\ \equiv\ \max_{A\subset\mathcal{Z}:|A|=t} I(V;y_A)\,, \tag{29}$$

其中， y_A 表示 A中设计点处的含噪观测值向量。

定理 3（掌握速度上限）。 在假设 *4–5* 下，假设用户通过不确定性采样（28）选择x_{t+1} 。定义命中时间

$$t_0 \;\equiv\; \inf\{t \in \mathbb{N}_0 : m_t \le \sigma_n^2\},$$

采用约定 如果集合为空，则规定为 $t_0 = \infty$ 。于是，对于所有整数 $t \ge t_0 + 1$ （当 $t_0 < \infty$ ），

$$m_t \le \frac{4\sigma_n^2\,\gamma_t}{t - t_0}.$$

特别地，当渐近地 $m_t = O(\sigma_n^2 \gamma_t / t)$ 时。

证明。 对于具有方差为 σ_n^2 的高斯观测噪声的高斯过程回归， V 与在点 $x_1, \dots, x_t$ 处观测值之间的互信息满足

$$I(V; y_{1:t}) = \frac{1}{2} \sum_{s=1}^{t} \log\!\left(1 + \frac{s_{s-1}^2(x_s)}{\sigma_n^2}\right)$$

（例如 Rasmussen 和 Williams（2006年））。由于 γ_t 是所有设计中的最大值，因此 $I(V; y_{1:t}) \le \gamma_t$。

对于 $a \in [0, 1]$, $\log(1 + a) \ge a/2$。根据 t_0 的定义以及后验方差的单调性，对于所有 $s \ge t_0 + 1$ ，我们有 $s_{s-1}^2(x_s) \le m_{s-1} \le m_{t_0} \le \sigma_n^2$，因此 $s_{s-1}^2(x_s)/\sigma_n^2 \in [0, 1]$ ，并且

$$\log\!\left(1 + \frac{s_{s-1}^2(x_s)}{\sigma_n^2}\right) \ge \frac{1}{2} \frac{s_{s-1}^2(x_s)}{\sigma_n^2}.$$

因此,

$$\gamma_t \ge I(V; y_{1:t}) \ge \frac{1}{2} \sum_{s=t_0+1}^{t} \log\!\left(1 + \frac{s_{s-1}^2(x_s)}{\sigma_n^2}\right) \ge \frac{1}{4\sigma_n^2} \sum_{s=t_0+1}^{t} s_{s-1}^2(x_s).$$

在不确定性采样下， $s_{s-1}^2(x_s) = m_{s-1}$ 和 $\{m_s\}$ 均不会增加，因此

$$m_t \le \frac{1}{t - t_0} \sum_{s=t_0+1}^{t} m_{s-1} = \frac{1}{t - t_0} \sum_{s=t_0+1}^{t} s_{s-1}^2(x_s) \le \frac{4\sigma_n^2\,\gamma_t}{t - t_0}.$$

□

定理 3 给出了不确定性采样下残余不确定性的上界： 关于局部可靠性的最坏情况后验不确定性满足 $m_t = O(\sigma_n^2 \gamma_t / t)$（忽略常数和噪声下限）。 γ_t 这一项表示在用户的相似性概念下，可靠性映射的内在复杂度。如果用户认为任务空间具有高维度或结构较弱，那么 γ_t会快速增长，掌握也会很缓慢。如果任务空间实际上是低维的（例如，具有稳定模板的狭窄工作流），那么 γ_t 增长缓慢，掌握就可能很快。

命题 13 (信息增益增长率)。 设 $\mathcal{Z} \subset \mathbb{R}^d$ 是紧致的，并假设 k 满足以下条件之一：

（*a*）平方指数（径向基函数）核函数，或

（*b*）平滑度参数为 $\nu > 0$ 的 *Mat´ern* 核函数。

那么，最大信息增益满足：

$$\begin{aligned} &\textit{(RBF)} \qquad \gamma_t = O\Big((\log t)^{d+1}\Big) \qquad \textit{(Srinivas et al., 2010)},\\ &\textit{(Matérn)} \qquad \gamma_t = \tilde{O}\Big(t^{\frac{d}{2\nu+d}}\Big) \qquad \textit{(Vakili et al., 2021)}, \end{aligned}$$

˜其中 $O(\cdot)$省略了多对数因子。

结合命题 13 和定理 3，可得出关于样本复杂度的结论：若要将最坏情况下的不确定性降至 ε 以下，就需要大约 $t \gtrsim \sigma_n^2 \gamma_t / \varepsilon$ 次评估。这与观察到的实践相符：用户可以快速学会适用于结构化领域（有效维度较低）的可靠提示词手册，但在为开放式任务（有效维度较高）制定稳定策略时却困难重重。

6.3 不确定性下的掌握即委托

掌握程度具有经济意义，因为它决定了用户会在何处 进行委托。给定局部误差 $V(x)$，使用人工智能时每项任务的期望效用为 $U(x) = 1 - V(x)/q$（外部选项归一化为 0）。如果用户确切知道 $V(x)$，那么最优规则将与完美校准相符：当且仅当 $V(x) \le q$时才委托。这是一个有用的上界。

引理 3 (完全掌握可复现校准)。 如果用户逐点知道 $V(x)$ ，那么最优委托规则为 $\pi^\star(x) = \mathbf{1}\{V(x) \le q\}$ ，期望效用为

$$U_M^\star(\lambda, q) = \mathbb{E}\left[\left(1 - \frac{V(x)}{q}\right)_+\right].$$

在基 线中， $V(x) = \sigma^2(x)$，这与第 *2* 节中的校准效用 $U_C(\lambda, q)$ 一致。 *6.*

证明。 逐点来看，用户在 $U(x)$ 和 0 之间进行选择，因此最优收益为 $\max\{U(x)\ 0\} =(1 - V(x)/q)_+$。取期望即可得到该表达式。当 $V(x) = \sigma^2(x)$ 时，这恰好就是 U_C 的定义。

□

在实践中，用户面临法律责任、验证成本和非对称损失：一次严重失误就可能抵消许多小收益。一种简化形式的刻画方法是采用保守委托规则 ，依据 $V(x)$ 的置信上界进行决策。

固定 $\beta > 0$ 并定义保守估计

$$\hat{V}_t(x) \ \equiv \ \mu_t(x) + \sqrt{\beta}\, s_t(x). \tag{30}$$

ˆ 对于合适的 β （为达到所需的置信水平而选取），在高斯过程模型下， $V_t(x)$ 以高概率构成 $V(x)$的上界（标准的“高斯过程置信上界”形式；参见 Srinivas 等人（2010））。由此得到的感知安全集为

$$S_t \ \equiv \ \{x \in \mathcal{Z} : \hat{V}_t(x) \le q\}, \tag{31}$$

相应的保守委托策略为 $\pi_t(x) = \mathbf{1}\{x \in S_t\}$。

式（30）—（31）的目的并非引入完整的多臂老虎机框架，而是刻画一种在经验上十分重要的行为规律：在高风险领域，组织不会仅凭均值信念进行委托。它们要求有证据以较高置信度确认委托是安全的，这会使委托区域相较于校准基准有所缩小，并使学习内生的 于过去的委托行为。

6.4 规模扩展与掌握的惯性

现在我们来考察扩展投资如何与掌握相互作用。令 V_0 表示强度为 λ_0的模型的更新前可靠性映射，令 V_1 表示扩展至 $\lambda_1 > \lambda_0$后的更新后映射。用户在 V_0 下进行学习，并将这些信念带入新版本。

假设6 （单调扩展）。 从 λ_0 扩展至 $\lambda_1 > \lambda_0$ 会增加覆盖率，而非重新分配覆盖范围。因此，局部误差在每一点上都弱递减：

$$V_1(x) \le V_0(x) \qquad \textit{for all } x \in \mathcal{Z}.$$

这一假设将掌握过程中的信息 摩擦单独提取出来。如果更新可能任意恶化某些区域，那么即使信息完美，出于风险考虑，缓慢采用也可能是最优的。假设6排除了这种情况：它确保先前的安全区域仍然安全，并确保任何未能利用改进的情况，都源于用户未能学习 新近改善的区域。这也与布朗–泊松基准相一致：增加知识点会细化间隔，并在各处（弱）降低布朗桥方差。

接下来, 为了更明确地表达学习会在局部泛化这一观点，我们施加有限推断半径 s.

假设7 (局部相似性 / 有限推断半径**)**。 存在 $\ell > 0$ ，使得核函数满足 $k(x, x') = 0$ ，只要 $\|x - x'\| > \ell$。

假设7排除了这样的退化情形：一次观测就能确定各处的可靠性。它可以从两个方面得到解释：一是提示词变体之间的迁移有限（用户会将差异足够大的任务表述方式视为彼此无关），二是基线中的插值风险具有局部性（某一点的不确定性取决于附近的锚点，而非远处的锚点）。没有紧支撑的核函数（例如径向基函数或 Mat´ern）会产生相同的定性逻辑：溢出效应会随距离增加而衰减，因此，随着改进发生在离用户已访问集合越来越远的位置，下面的“隐藏规模”结果会成为任意精确的近似。

令 $\mathcal{A}_t = \{x_1, \dots, x_t\}$ 为截至时间 t的评估位置集合，并定义其 ℓ-邻域

$$N_\ell(\mathcal{A}_t) \equiv \{x \in \mathcal{Z} : \min_{a \in \mathcal{A}_t} \|x - a\| \le \ell\}. \tag{32}$$

在假设7下， t时刻 $N_\ell(\mathcal{A}_t)$之外的点在统计上与用户的体验相互脱节：若没有刻意的探索，其后验分布将保持为先验。

下一个命题将“弃答陷阱”形式化：改进在技术上可能确实存在，却在经济上不可见。

命题14（隐藏尺度的边际收益为零）。 固定一个信念状态 $\mathcal{F}_t$，其数据位置为 $\mathcal{A}_t$，安全集为 S_t。考虑两个更新后误差映射 V_1 和 V_1，它们满足 $V_1(x) \le V_1(x)$，对所有 $x \in \mathcal{Z}$ 以及

$$\tilde{V_1}(x) = V_1(x) \quad for\ all\ x \in N_\ell(\mathcal{A}_t). \tag{33}$$

假设用户在更新时将 $\mathcal{F}_t$ 作为先验带入，并且更新后只在其委托的位置观察反馈（即只在 $s \ge t$ 的 S_s 内的点上观察）。定义停止时刻

$$\tau \equiv \inf\{s > t : x_s \notin N_\ell(\mathcal{A}_t)\},$$

约定无穷大 $\emptyset = \infty$。那么，根据假设6和7，在事件 $\{\tau = \infty\}$上，V_1 和 V_1 所诱导的高斯过程后验与委托策略序列完全相同；更一般地，在每条样本路径上，它们在时间 τ 之前都完全相同。特别是，相较于 V_1，V_1中任何由该邻域之外$(\mathcal{A}_t)$ 的部分所支持的改进，在用户首次于该邻域之外采样之前，对结果的边际影响均为零；若 $\tau = \infty$，则其边际影响将永远为零。

证明。 条件（33）意味着，在每个 $x \in N_\ell$（$\mathcal{A}_t$）处，V_1 和 V_1 下反馈 $y = V$（x）的条件分布 $+\eta$ 完全相同~。根据假设7和 τ的定义，对于所有 $s < \tau$，用户只会在 N_ℓ（$\mathcal{A}_t$）内委托，因此截至时间 s 的实现数据集仅包含来自该邻域的观测值~。由于 V_1 和 V_1 在 N_ℓ（$\mathcal{A}_t$）上重合，因此对于所有 $s < \tau$，两种映射下实现的反馈序列（因而，高斯过程后验在 N_ℓ（$\mathcal{A}_t$）上的限制）都相同。

我们对 s 进行归纳证明：对于任何满足 x_{t+1}、.....、$x_s \in N_\ell(\mathcal{A}_t)$ 的样本路径，在时间 s 之前，V_1 和 V_1 下实现的更新后数据集完全相同，因此两个映射下的后验分布（μ_s、s_s，限制在 $N_\ell(\mathcal{A}_t)$ 上）也完全相同。基本情形 $s = t$ 由共同先验 $\mathcal{F}_t$ 得证。对于归纳步骤，假设该命题在时间 s 成立，且 $x_{s+1} \in N_\ell(\mathcal{A}_t)$。由于用户的委托和评估选择是其当前后验分布的可测函数，归纳假设意味着两个映射下的选择 x_{s+1} 相同。~由于 V_1 和 V_1 在 x_{s+1}处一致，y_{s+1} 的条件分布在两个映射下相同，因此实现的数据集仍然相同，时间 $s+1$的后验分布也相同。

因此，在所有时刻 $s < \tau$，后验分布以及委托策略都一致。若 $\tau = \infty$，则这些过程在所有时刻都一致。由于实现效用仅取决于委托集合以及该集合上的$V(\cdot)$，因此在 $N_\ell(\mathcal{A}_t)$ 之外的改进，在 τ 之前都不会产生边际影响（而在$\tau = \infty$时则永远不会产生影响）。

□

掌握数据是在用户实际操作的地方生成的。在保守委托下，不会对“不安全”区域进行采样，因此这些区域的信念不会更新。因此，规模扩展可能会产生隐藏能力：这些任务实际上已变得安全，却仍被排除在用户继承的掌握地图之外。命题14表明，在局部泛化下，若改进超出用户的信息触达范围，那么在用户主动向外探索之前，其实际边际回报为零。如果用户始终不这么做，这些回报就会永远为零。

6.5 前沿溢出效应

命题14 并不意味着规模扩展通常毫无用处。用户经常在一个^ 前沿附近操作，其中 $V_t(x)$ 接近 q。在他们实际涉足的区域内进行改进，可以改变他们对邻近任务的认知

并扩展安全集。下一个引理记录了高斯过程更新，并明确说明了局部性。

引理 4（高斯过程回归下的局部信念更新）。 在假设 *4–5* 下，观测到$y_{t+1} = V（x_{t+1}） + \eta_{t+1}$在位置 x_{t+1}处的值后，后验均值和方差满足

$$\mu_{t+1}(x) = \mu_t(x) + \frac{k_t(x, x_{t+1})}{k_t(x_{t+1}, x_{t+1}) + \sigma_n^2}\Big(y_{t+1} - \mu_t(x_{t+1})\Big), \tag{34}$$

$$s_{t+1}^2(x) = s_t^2(x) - \frac{k_t(x, x_{t+1})^2}{k_t(x_{t+1}, x_{t+1}) + \sigma_n^2}, \tag{35}$$

其中， $k_t（\cdot、 \cdot）$ 是时间 t的后验协方差核。特别地，如果 $k_t（x、 x_{t+1}） = 0$，那么$\mu_{t+1}（x） = \mu_t（x）$，且 $s_{t+1}（x） = s_t（x）$。

证明。 这些是标准的高斯过程回归更新公式（例如 Rasmussen 和 Williams（2006年，第 2 章））。最后的结论直接由（34）–（35）得出。 □

现在，我们将“前沿溢出效应”机制形式化：在查询点观测到改进会减轻悲观程度，并收紧附近的不确定性，从而可能扩大感知安全集。

命题15（规模扩展的前沿溢出效应）。 假设规模扩展改进后，真实误差映射从 V_0 转变为 V_1，并且逐点满足 $V_1 \le V_0$。考虑用户在更新后立即评估的点 $x^\dagger \in \mathcal{S}_t$，并定义相对于用户更新前信念的（局部）意外改进

$$\Delta_t^\dagger \;\equiv\; \mu_t(x^\dagger) - V_1(x^\dagger).$$

然后，在给定 $\mathcal{F}_t$ 和 $x_{t+1} = x^\dagger$的条件下，

$$\mathbb{E}[\hat{V}_{t+1}(x) \mid \mathcal{F}_t, x_{t+1} = x^\dagger] \le \hat{V}_t(x) - \frac{k_t(x, x^\dagger)}{k_t(x^\dagger, x^\dagger) + \sigma_n^2}\,\Delta_t^\dagger. \tag{36}$$

证明。 给定 $\mathcal{F}_t$ 和 $x_{t+1} = x^\dagger$的条件下， y_{t+1} 中唯一剩余的随机性是 η_{t+1}，因此 $\mathbb{E}[y_{t+1} \mid \mathcal{F}_t, x_{t+1} = x^\dagger] = V_1(x^\dagger)$。所以，预期创新项为

$$\mathbb{E}[y_{t+1} - \mu_t(x^\dagger) \mid \mathcal{F}_t, x_{t+1} = x^\dagger] = V_1(x^\dagger) - \mu_t(x^\dagger) = -\Delta_t^\dagger.$$

代入引理 4 得到 $\mathbb{E}[\mu_{t+1}(x) \mid \mathcal{F}_t, x_{t+1} = x^\dagger] = \mu_t(x) - \frac{k_t(x,\, x^\dagger)\, k_t(x^\dagger,\, x^\dagger)}{k_t(x^\dagger,\, x^\dagger) + \sigma_n^2}\,\Delta_t^\dagger$。由于$s_{t+1}(x) \le s_t(x)$，我们有 $\hat{} \; V_{t+1}(x) \le \mu_{t+1}(x) + \sqrt{\beta}\, s_t(x)$，从而得到 (36)。 □

隐藏的规模是空间性的：远离用户操作区域的改进不会改变他们的信念，但前沿任务上的改进（或附近的改进）能够为相邻任务提供信息，并扩大安全集。这种效应的强度和范围由核函数（用户认为什么是“相似的”）以及噪声水平 σ_n^2 （验证的成本或噪声有多高）决定。

6.6 基线几何：布朗间隙中的安全边缘

在布朗–泊松基准中，上述局部性有一个简单的几何解释。在间隔长度为 X 的两个相邻知识点之间，局部方差为

布朗桥表达式 (1)： $\sigma^2(x) = Xt(1-t)$，其中 $t \in [0,1]$ 是间隔内的相对位置。该间隔内局部非负价值任务的集合仅取决于 X。

引理5 （布朗桥间隔内的安全边缘）。 在基准模型中，固定相邻知识点之间长度为 X 的间隔。在该间隔中，使用人工智能可获得非负期望效用的位置集合 $\{x:\ \sigma^2(x) \le q\}$，是与间隔端点相邻的两个区间的并集。若 $X \le 4q$，则整个间隔都是安全的。若 $X > 4q$，安全区间的长度为

$$d(X,q) \ \equiv\ \frac{1}{2}\Big(X - \sqrt{X(X-4q)}\Big), \tag{37}$$

因此，不安全区域是长度为 $X - 2d(X,q)$ 的中间区间。此外， $d(X,q)$ 在 X 的情况下，对 $X > 4q$ 严格递减。

证明。 将间隔记为 $[0,X]$。在距左端点 $s \in [0,X]$ 处， $\sigma^2(s) = s(X-s)/X$。条件 $\sigma^2(s) \le q$ 等价于 $s(X-s) \le qX$，即 $s^2 - Xs + qX \ge 0$。若 $X \le 4q$，判别式非正，且对所有 $s \in [0,X]$ 不等式均成立。若 $X > 4q$，二次函数有根 $s_\pm = \frac{1}{2}\big(X \pm \sqrt{X(X-4q)}\big)$，且 $\sigma^2(s) \le q$ 对 $s \in [0,s_-] \cup [s_+,X]$ 成立。因此，安全边缘长度为 $d(X,\ q) = s_-$。对（37）求导得 $d_X(X,\ q) < 0$，其中 $X > 4q$。 □

规模扩展通过缩小间隔而发挥作用。在这个基线情形中，缩小间隔会确定性地扩大安全边缘。这是前沿溢出效应最简单的情形：即使用户只在现有锚点（安全区域）附近操作，在附近增加锚点也会扩大委托的局部净收益非负区域。

6.7 扩展投资的意义

掌握程度使规模扩展的实现收益取决于状态，因为规模扩展会改变底层映射 V，而掌握程度则决定了该映射中哪些部分能够实际转化为收益。

令 x 表示从 $\mathcal{Z}$ 上的到达分布中抽取的随机任务。给定委托策略 $\pi(\cdot) \in \{0,1\}^{\mathcal{Z}}$，在误差映射 V 下实现的期望效用为

$$\bar{U}(V,\pi) \ \equiv\ \mathbb{E}\left[\left(1 - \frac{V(x)}{q}\right)\pi(x)\right].$$

对于任意更新前策略 π_t 和更新后策略 π_{t+1}，

$$\begin{aligned}\bar{U}(V_1,\pi_{t+1}) - \bar{U}(V_0,\pi_t) = \mathbb{E}\left[\frac{V_0(x) - V_1(x)}{q}\,\pi_t(x)\right] \\ + \mathbb{E}\left[\left(1 - \frac{V_1(x)}{q}\right)\big(\pi_{t+1}(x) - \pi_t(x)\big)\right].\end{aligned} \tag{38}$$

第一项是集约边际：用户原本就会委托的任务所带来的改进。第二项是广延边际：随着信念更新，委托任务集合扩大所带来的额外价值。

命题14对广延边际提出了严格约束：用户推断邻域之外的改进无法改变 π_{t+1}，除非用户向外探索。命题15

和引理5展示了相反的渠道：用户操作前沿附近的改进可以扩展 π_{t+1} ，即使无需进行有风险的探索。

相较于研究固定信息条件下静态规模报酬的第 4 节，掌握度引入了一个动态楔子：

• **可发现性很重要**。提高 λ 所带来的实现边际回报，取决于改进后的区域是否会被访问，或能否从已访问的区域推断出来。主要改善“其他地方”的规模扩展，在短期内可能带来较低的实现收益，即便它确实提升了各处的能力。

• **前沿定位具有不成比例的价值**。集中于用户已经接触的任务附近，尤其是接近安全边界的任务附近的改进，能够触发安全集扩展，并解锁广延边际。

• **发布策略会与掌握度相互作用**。当重新验证和工作流调整的成本较高时，用户可能会在更新之间沿用较为保守的掌握度图谱。在单调扩展下，这样做是安全的，但可能使收益无法发挥。这就为将规模扩展与以下任一项结合起来提供了依据：(i) 能够改变信念的显著改进证据；或 (ii) 可替代缓慢掌握过程的界面级校准工具。

总之，单调扩展确保既有掌握度不会被破坏，但并不确保新的能力能够被发现。因此，掌握度是规模扩展的补充，不是因为它能让模型变得更好，而是因为它决定了用户能够以多快、多广的速度收获规模扩展所创造的改进。

7 扩展与应用

前几节阐述了人工锯齿智能（AJI）的核心经济逻辑：用户体验到的质量由局部可靠性 决定，并受到检查悖论的影响，因此基准平均值可能无法准确预测实际体验价值。本节将介绍两个运用这一逻辑的应用场景，其中相关对象不再只是“一个典型的采用者”，而是异质用户和制度。

第一个应用展示了 AJI 如何自然地产生基于可靠性特征的垂直差异化：即使“平均能力”的排名十分明确，利害关系不同的用户仍可能理性地偏好不同的系统。第二个应用研究组织采纳与工作者体验：委托人依据基准测试信号作出采纳决策，而代理人体验的却是长度存在偏差的任务，从而造成过度采纳和可预测的信任动态。

7.1 基于可靠性特征的垂直差异化

用户面临的利害关系差异很大。错误容易纠正（ q 较高）时，用户主要关注平均能力。错误代价高昂或不可逆（ q 较低）时，用户会格外关注如何避免罕见但极端的失败。AJI 使这种区别具有了经济意义，因为崎岖景观会产生高误差尾部，而在实际使用中，这类尾部会被过度加权（命题 1）。因此，除了基准平均值之外，质量还有第二个维度：系统的可靠性特征——错误在各项任务中的分布情况，而不仅仅是其均值。

在基准收益 $U(x) = 1 - \sigma^2(x)/q$ 中，盲用户的期望价值仅取决于$\mathbb{E}[\sigma^2]$。当错误造成的损失近似线性时，这种设定是合理的，此时用户对错误实际上是风险中性的。高风险部署很少如此：法律责任、下游级联效应和合规约束会引入凸损失，使 $\sigma^2(x)$ 中的离散度 （尾部风险）成为一阶考量。下面的扩展以最简方式单独刻画这一渠道，同时保留模型的结构。

7.1.1 “灾难性”利害关系的最简模型

假设8 （凸性错误损失）。 当用户在任务 x 中依赖提供方 j 时，其净收益为

$$u_j(x;q) \;=\; 1 - \frac{\sigma_j^2(x)}{q} - \phi\left(\frac{\sigma_j^2(x)}{q}\right)^2, \qquad \phi \ge 0, \tag{39}$$

外部选项标准化为 0。

二次项是一种简化形式，用来体现这样一种观点：在高风险场景（法律风险敞口、声誉损害、安全事故、级联返工）中，大错误造成的成本会不成比例地高。它也是围绕小误差展开的一般凸损失展开中的二阶项。当 $\phi = 0$ 时，基准模型即为该模型的特例，因此所有基准结果仍可解释为风险中性基准。

在假设8下，盲用户从提供方 j 获得的期望价值（在外部选项处截断之前）为

$$\bar{U}_j(q) \;\equiv\; \mathbb{E}[u_j(x;q)] = 1 - \frac{\mathbb{E}[\sigma_j^2(x)]}{q} - \phi\frac{\mathbb{E}[\sigma_j^4(x)]}{q^2}. \tag{40}$$

新的对象是四阶矩 $\mathbb{E}[\sigma_j^4(x)]$, ，而这正是可靠性特征发挥作用之处。下一个引理将在布朗桥环境中计算这些矩，并清楚揭示检查悖论的作用。

引理6 （实际经历的误差的矩）。 固定一个提供方 j ，其间隔长度分布X_j 的均值为 $\mathbb{E}[X_j] = 1/\lambda_j$。令 X_j^* 表示随机到达的任务所面对的长度偏倚间隔（命题 *1*），并令 $t \sim \mathrm{Uniform}(0,1)$ 表示任务在间隔中的相对位置，因此σ_j^2（x） $= X_j^*\, t$（$1-t$）。于是

$$\mathbb{E}[\sigma_j^2(x)] = \frac{\mathbb{E}[X_j^*]}{6} = \frac{\mathbb{E}[X_j^2]}{6\,\mathbb{E}[X_j]}, \tag{41}$$

$$\mathbb{E}[\sigma_j^4(x)] = \frac{\mathbb{E}[(X_j^*)^2]}{30} = \frac{\mathbb{E}[X_j^3]}{30\,\mathbb{E}[X_j]}. \tag{42}$$

证明。 在给定 X_j^* 的条件下， $\sigma_j^2 = X_j^*\, t(1-t)$，其中 $t \sim$ 均匀分布于 0，1，因此 $\mathbb{E}[t(1-t)] = \int_0^1 t(1-t)\,dt = 1/6$ 且 $\mathbb{E}[t^2(1-t)^2] = \int_0^1 t^2(1-t)^2\,dt = 1/30$。对 X_j^* 取期望可得 $\mathbb{E}[\sigma_j^2] = \mathbb{E}[X_j^*]/6$ 且 $\mathbb{E}[\sigma_j^4] = \mathbb{E}[(X_j^*)^2]/30$。最后，对于任意间隔分布 X，长度偏倚矩满足 $\mathbb{E}[(X^*)^m] = \mathbb{E}[X^{m+1}]/\mathbb{E}[X]$，由此得到（41）–（42）。 □

引理6说明了为什么检查悖论会对“灾难性”的利害关系产生影响。实际上对间隔进行均匀加权的基准测试（每个间隔/定义域各取一项）会将 X^* 替换为 X。在这种

间隔均匀抽样下，

$$\mathbb{E}_{\text{gap-unif}}[\sigma^2] = \frac{\mathbb{E}[X]}{6} = \frac{1}{6\lambda}, \qquad \mathbb{E}_{\text{gap-unif}}[\sigma^4] = \frac{\mathbb{E}[X^2]}{30}.$$

相比之下，用户体验将 $\mathbb{E}[X]$ 替换为 $\mathbb{E}[X^*] = \mathbb{E}[X^2]/\mathbb{E}[X]$，并将 $\mathbb{E}[X^2]$ 替换为 $\mathbb{E}[(X^*)^2] = \mathbb{E}[X^3]/\mathbb{E}[X]$。因此，AJI 不仅放大了平均误差，也放大了决定尾部敏感损失的高阶矩。

7.1.2 一个易于处理的规律性族

为了将可靠性特征与第 4 节的规律性边际联系起来，采用一个由单一参数控制离散度的参数族会比较方便。

假设9 (伽马规律性族)。 对于提供方 j，间隔长度满足 $X_j \sim \text{Gamma}(a_j, a_j\lambda_j)$（形状参数 $a_j > 0$，率参数 $a_j\lambda_j$），因此 $\mathbb{E}[X_j] = 1/\lambda_j$ 且 $\text{CV}(X_j) = 1/\sqrt{a_j}$。

伽马族是间距的一种标准简化形式：它在 $a = 1$（指数间隔）时包含泊松基准，并在 $a \to \infty$（完全规律性）时收敛于确定性间隔。因此，它为第 4 节中的规律性参数提供了微观基础，同时保留了闭式矩。

在假设9下，引理6中的矩表达式可以直接用（λ，变异系数）表示。

引理7 （伽马分布规律性下的闭式矩）。 在假设*9*下，记变异系数$_j$ ≡变异系数(X_j)，

$$\mathbb{E}[\sigma_j^2(x)] = \frac{1 + \text{变异系数}_j^2}{6\lambda_j}, \tag{43}$$

$$\mathbb{E}[\sigma_j^4(x)] = \frac{1 + 3\text{CV}_j^2 + 2\text{CV}_j^4}{30\,\lambda_j^2} = \frac{(1 + \text{CV}_j^2)(1 + 2\text{CV}_j^2)}{30\,\lambda_j^2}. \tag{44}$$

此外，四阶矩的检查悖论放大效应为

$$\frac{\mathbb{E}[\sigma_j^4]_{user}}{\mathbb{E}[\sigma_j^4]_{gap\text{-}unif}} = 1 + 2\text{CV}_j^2,$$

因此，在泊松间隔下（变异系数$_j = 1$），用户体验到的 $\mathbb{E}[\sigma^4]$ 是间隔均匀基准值的三倍。

证明。 对于 $X \sim$ 伽马分布$(a, a\lambda)$，$\mathbb{E}[X^2] = (a+1)/(a\lambda^2) = (1 + 1/a)/\lambda^2 = (1 + \text{变异系数}^2)/\lambda^2$，因此（43）可由（41）得出。同理，$\mathbb{E}[X^3] = (a+1)(a+2)/(a^2\lambda^3)$，因此 $\mathbb{E}[(X^*)^2] = \mathbb{E}[X^3]/\mathbb{E}[X] = (a+1)(a+2)/(a^2\lambda^2)$，且（44）可由（42）得出。对于放大比率，用户体验的四阶矩为

$$\mathbb{E}[\sigma^4]_{\text{user}} = \frac{1}{30}\mathbb{E}[(X^*)^2] = \frac{(1 + \text{CV}^2)(1 + 2\text{CV}^2)}{30\lambda^2},$$

而间隔均匀四阶矩为 $\mathbb{E}[\sigma^4]_{\text{gap-unif}} = \frac{1}{30}\mathbb{E}[X^2] = \frac{(1+\text{变异系数}^2)^2}{30\lambda^2}$。相除后得到 $\mathbb{E}[\sigma^4]_{\text{user}}/\mathbb{E}[\sigma^4]_{\text{gap-unif}} = 1 + 2\text{变异系数}^2$。 □

7.1.3 按利害程度排序

考虑两个提供方 H和 L，其可靠性特征分别由 (λ_H，变异系数$_H$) 和 (λ_L，变异系数$_L$) 表示。可以将 H 看作“经过缩放但参差不齐”，将 L看作“更加规整”。

命题 16 (按利害程度单交叉)。 维持假设*8*和假设*9*。定义均值项和尾部项

$$m_j \equiv \mathbb{E}[\sigma_j^2(x)], \qquad v_j \equiv \mathbb{E}[\sigma_j^4(x)].$$

假设提供方 H 的均值更优，但尾部更差，即

$$m_H < m_L \qquad and \qquad v_H > v_L. \tag{45}$$

那么存在唯一的临界值

$$q^* \equiv \phi \frac{v_H - v_L}{m_L - m_H} > 0 \tag{46}$$

使得:

- *(i)* 若 $q > q^*$ *(*低利害*)*，则 $U_H(q) > U_L(q)$;
- *(ii)* 若 $q < q^*$ （高利害关系），则 $U_H(q) < U_L(q)$。

证明。由（40）可得，两者之差为

$$\bar{U}_H(q) - \bar{U}_L(q) = \frac{m_L - m_H}{q} - \phi \frac{v_H - v_L}{q^2}.$$

在（45）成立时，第一项为正，第二项为负。令两者之差为零，可得唯一的正解（46）。立即可知 q 高于或低于 q^*时的符号比较结果。

□

当 q 较大（错误代价较低）时，（40）中的二次项可以忽略，用户的行为接近风险中性：他们会选择均值表现更好的提供方（此处为 H）。当 q较小时（错误代价较高），二次项会随 $1/q^2$ 增大而放大，尾部风险占据主导：用户更倾向于可靠性特征中极端故障占比较低的提供方（此处为 L）。引理7说明了规律性为何在利害关系重大时至关重要：CV 增大会使 v 急剧上升，而检查悖论会通过对较长间隔赋予过高权重来放大这一影响。

规律性通过缩小间隔的离散度，事前降低高方差尾部。校准通过筛选出尾部，事后降低其对福利的影响（第2.6节）。形式上，对于面临归一化方差 Z 和可靠性指数 R的校准用户，

$$U_C(R) - \bar{U}_B(R) = \mathbb{E}\left[\left(\frac{Z}{R} - 1\right)_+\right] = \frac{1}{R}\mathbb{E}[(Z - R)_+],$$

因此，校准的增量价值就是所避免的预期尾部损失。由于 $(z - R)_+$ 关于 z 是凸函数， Z 的任何均值保持离散度增加（即凸序增加）都会提高$\mathbb{E}[(Z - R)_+]$，因此在 R 固定时，也会提升校准的价值。这从严格的形式上表明，崎岖程度更高的可靠性特征会使校准/弃答接口更

有价值。因此，“规模大但崎岖”的提供方可以通过投资校准/弃答接口，部分弥补规律性不足，而这正是高风险场景用户看重的。

命题 16 描述了一种简单的细分逻辑，与观察到的实践相符：大众市场产品可以合理地优先考虑规模和平均能力，而专业级产品则可以合理地优先考虑更稳定的可靠性特征，以及能够降低尾部暴露度的校准功能。检查悖论使这种差异化在经济上尤为显著：它让罕见的长间隔故障成为实际体验的重要组成部分。

7.2 组织采纳与工作者体验

组织通过采购、试点和基准测试评估来采用人工智能。随后，工作者会在实际任务中使用该系统。AJI 预测，这两种视角之间存在系统性的落差：评估通常更接近间隔均匀抽样，而工作者体验则存在长度偏倚（命题 1）。由此会出现一种可预见的模式：先是以出色的基准平均值为依据采纳系统，随后当“意外失败”发生时，便引发后续的沮丧、变通方案和不信任。本小节将对这一落差进行形式化分析，并探讨其与规模扩展、校准和组织政策的关系。

固定一个覆盖强度为 λ、间隔离散度为变异系数的环境，如第 4 节所述。设组织在一个实际上对间隔赋予均匀权重的基准上评估该系统。在布朗桥结构下，基准评估对应于均值—方差$\mathbb{E}_{\text{gap-unif}}[\sigma^2] = \mathbb{E}[X]/6 = 1/(6\lambda)$。定义相应的 基准可靠性比率

$$R^{\text{bench}} \;\equiv\; \frac{q}{\mathbb{E}_{\text{gap-unif}}[\sigma^2]} = 6\lambda q. \tag{47}$$

相比之下，员工面对的是从到达分布中抽取的任务，因此体验到长度偏倚的均值方差 $\mathbb{E}[\sigma^2] = \mathbb{E}[X^*]/6 = (1 + \text{变异系数}^2)/(6\lambda)$（命题6）。定义体验可靠性比率

$$R^{\text{exp}} \;\equiv\; \frac{q}{\mathbb{E}[\sigma^2]} = \frac{6\lambda q}{1 + \text{CV}^2} = \frac{R^{\text{bench}}}{1 + \text{CV}^2}, \tag{48}$$

在泊松间隔下，该值与第 2 节中的 R 一致（变异系数 $= 1$）。

命题 17 （检查悖论导致的过度采纳）。 假设某组织在基于基准的盲目价值非负时采用：$1 - 1/R^{\text{bench}} \geq 0$，即 $R^{\text{bench}} \geq 1$。当且仅当 $R^{\text{exp}} \geq 1$ 时，员工体验到的盲目价值为正；等价地， $R^{\text{bench}} \geq 1 + \text{变异系数}^2$。因此，对于每个变异系数 > 0 ，都存在一个区间

$$R^{\text{bench}} \in [1,\, 1 + \text{CV}^2)$$

在该区间内，组织会采用，但员工因盲目信赖而体验到的期望价值为负。在泊松间隔下（变异系数 $= 1$），这一过度采纳区间为 $R^{\text{bench}} \in [1, 2)$。

证明。 根据定义，当且仅当 $R^{\text{exp}} \geq 1$ 时，员工价值为正。利用 (48) 得到 $R^{\text{exp}} \geq 1 \iff R^{\text{bench}} \geq 1 + \text{变异系数}^2$。基准采用规则只要求 $R^{\text{bench}} \geq 1$，因此，当变异系数 > 0 时，不匹配区间为 $[1, 1 + \text{变异系数}^2)$。 □

在过度采用区域内，按照其测量机制，该组织的判断是“正确的”：基准表现的平均值足够高，超过了临界条件。员工的判断也“正确”：他们实际遇到的任务是从长度偏倚分布中抽取的，其中长间隔任务所占比例过高，而系统恰恰会在这些任务上失效。因此，AJI 预测的并非随机的失望，而是在基准平均值显示采用处于临界状态的情境中，集中出现的系统性失望。

命题 17 也阐明了为什么不同受众对测得的“扩展回报”可能会有不同看法。对 (47)–(48) 关于 λ 求导，得到

$$\frac{\partial R^{\mathrm{bench}}}{\partial \lambda} = 6q, \qquad \frac{\partial R^{\mathrm{exp}}}{\partial \lambda} = \frac{6q}{1+\mathrm{CV}^2}.$$

因此，按基准指标看似带来 $\Delta R^{\mathrm{bench}}$ 改善的固定规模增幅，在实际可靠性方面仅带来 $\Delta R^{\mathrm{bench}}/(1+\text{变异系数}^2)$。在泊松间隔下，这一比例为 $1/2$。因此，AJI 预测，除非明确纳入长度偏倚或尾部指标，否则基于基准的规模扩展曲线可能会系统性地高估不规则领域中的规模扩展福利回报。

如果某个组织知道（或能够界定）变异系数，就可以通过仅在以下条件成立时采用，来使采用决策与劳动者福利保持一致：

$$R^{\mathrm{bench}} \geq 1 + \mathrm{CV}^2. \tag{49}$$

挑战在于，标准排行榜通常无法识别变异系数。因此，第 4 节提出了评估协议：要么明确按使用加权，要么报告与长间隔相关的尾部风险度量。

一种常见的组织政策是强制验证或人工审查。在该模型中，验证之所以有价值，恰恰是因为它以一定成本复现了校准基准。

命题 18（强制验证何时能改善福利）。 假设员工可以通过支付每任务成本 $c_v > 0$ 来验证局部可靠性。验证会揭示 $\sigma^2(x)$，并允许员工在依赖其结果会产生负价值时选择弃权。那么，经过验证后进行委托的预期收益为

$$U_C(\lambda, q) - c_v \;\equiv\; \mathbb{E}\left[\left(1 - \frac{\sigma^2(x)}{q}\right)_+\right] - c_v,$$

而盲目委托的收益为 $\bar{U}_B(\lambda, q) = 1 - \mathbb{E}[\sigma^2(x)]/q = 1 - 1/R^{\mathrm{exp}}$（可能为负），其中 $R^{\mathrm{exp}} \equiv q/\mathbb{E}[\sigma^2(x)]$。当且仅当满足以下条件时，验证会严格提高预期福利：

$$c_v < U_C(\lambda, q) - \bar{U}_B(\lambda, q) = \mathbb{E}\left[\left(\frac{\sigma^2(x)}{q} - 1\right)_+\right] = \mathbb{E}\left[\left(\frac{\tilde{Z}}{R^{\mathrm{exp}}} - 1\right)_+\right], \tag{50}$$

其中 $\tilde{Z} \equiv \sigma^2(x)/\mathbb{E}[\sigma^2(x)]$ 的均值为 1。

证明。 在验证的情况下，工作人员仅在 $\sigma^2(x) \leq q$ 时才委派，并获得 $(1-\sigma^2(x)/q)_+ - c_v$。取期望可得 $U_C(\lambda, q) - c_v$。在没有验证的情况下，工作人员根据预期质量进行委派，得到 $\bar{U}_B(\lambda, q) = \mathbb{E}[1 - \sigma^2(x)/q]$。因此，验证带来的福利增益为

$$U_C(\lambda, q) - c_v - \bar{U}_B(\lambda, q),$$

严格改进的条件为（50）。 □

(50) 右侧恰好是因被迫进入高误差尾部而产生的预期损失。在更崎岖的环境中（ Z 的离散度更高），以及在采用前沿附近，这一损失更大；组织最常在这些时候讨论推广部署。这形式化地说明了为什么“人在回路”政策最有价值的情形，并非系统糟糕透顶或完美无缺，而是系统几乎足够好。

即使采用平均而言能够改善福利，当负责决定是否委托的工人需要在组织价值与不借助人工智能完成任务所需的私人努力成本之间权衡时，组织内部的委托选择仍可能受到扭曲。在我们的约化形式设定中，这种利益不一致源于努力成本楔子（而非另行设定的人工智能错误损失分摊规则）。

假设 10 （委托与努力）。 工人可以选择：（*i*）自行完成任务，承担私人努力成本 $c_e > 0$ ；或（*ii*）将任务委托给人工智能。如果工人盲目委托，随机任务上委托的¯预期增量价值为 U_B（$R^{\exp}$）。

假设10采用有意简化的形式：它刻画了委托如何节省努力，却使组织面临错误风险。为清晰呈现这一楔子，假设将更复杂的契约安排和监督排除在外。

命题19 （工人过度委托）。 根据假设*10*，只有当 U_B（$R^{\exp}$）≥ 0 时，组织才更倾向于¯委托，即 $R^{\exp} \geq 1$。只要 U_B（$R^{\exp}$）$\geq -c_e$，工人就会委托。因此，只要 $c_e > 0$，在一组非空的参数取值范围内，工人就会过度委托。等价地，以基准可靠性表示，过度委托区域为

$$R^{\text{bench}} \in \left[\frac{1+\text{CV}^2}{1+c_e},\, 1+\text{CV}^2 \right), \tag{51}$$

¯按照以下约定：若 $U_B < -c_e$，则绝不选择委托。

¯证明。 由于 $U_B(R^{\exp}) = 1 - 1/R^{\exp}$，当且仅当 $1 - 1/R^{\exp} \geq -c_e$ 时，工作人员才会委托，即 $R^{\exp} \geq 1/(1+c_e)$。利用式（48）中的 $R^{\exp} = R^{\text{bench}}/(1+\text{CV}^2)$，可得 $R^{\text{bench}} \geq (1+\text{CV}^2)/(1+c_e)$。只有当 $R^{\exp} \geq 1$，即 $R^{\text{bench}} \geq 1+\text{CV}^2$ 时，组织才更倾向于委托。两者之差即为式（51）。 □

环境越崎岖（变异系数越大），过度委托在机制上就越严重，因为相同的基准可靠性比率对应的实际可靠性更低。这有助于解释一种常见的部署病理：管理层看到基准测试表现出色，便强制要求使用；工作人员面临工作压力，在预期组织价值为负的边缘情形下仍选择委托；随后发生质量事故，信任逐渐瓦解，非正式规范也随之形成，将使用限制在“安全”的范围内。

一旦劳动者因锯齿性而将使用范围限制在安全区域内，他们的经验便会内生于其委托策略。第6节表明，这会造成额外的楔差：如果规模化改进发生在劳动者不会涉足、也无法根据自身涉足区域推断的范围之外，那么这些改进在经济上潜在的（命题14）。组织基准

改进会因此两度高估劳动者体验到的改进：首先，通过（48）在机制上高估；其次，在动态层面上，保守委托可能阻碍发现新能力所需的采样过程。

总之，AJI 不仅是模型误差的一种属性，也是制度与测量的一种属性。它预示着基准测试驱动的采用何时会令人失望，解释了为何管理者和员工对扩展回报的感受可能不同，也说明了为什么在组织最希望获得可靠人工智能的领域，校准、验证和正则性投资恰恰具有不成比例的价值。

8 结论

人工锯齿智能（AJI）是模型可靠性的一种几何特征：由于知识覆盖不均，模型在相邻任务上的表现会出现剧烈波动。在布朗–泊松基准中，这种不均匀性与检查悖论相互作用：典型用户遇到的间隔存在长度偏倚，因此实际经历的平均误差大于锚点之间平均间距所暗示的水平。这为盲目用户提供了简明的采用阈值，也阐明了为何规模扩展（提高覆盖强度）能提升平均可靠性，却仍无法消除失败风险的长右尾。

一项核心经济学启示是任务级校准的期权价值。当用户能够观察（或推断）局部可靠性时，便可以有选择地委派任务，将锯齿性转化为可利用的机会。因此，校准与规模扩展既可能是替代品（当规模扩展压缩误差分布并缩短尾部时），也可能是互补品（当校准使用户能够安全地获取规模扩展带来的改进，而这些改进原本可能无法使用时）。

最后，掌握依然至关重要，因为校准很少能够做到完美：用户必须了解在哪些情况下进行委托是安全的。在高斯过程学习模型下，关于局部可靠性的最坏情况不确定性上界为 $O(\gamma_t/t)$，其中 γ_t 刻画了可靠性景观的有效复杂度。这会造成弃答陷阱：如果用户只抽样他们原本就认为安全的任务，那么规模扩展带来的改善可能始终无法被发现。鼓励用户在前沿附近进行探索、提供可靠不确定性信号，或降低验证成本的界面设计，能够将原本隐藏的规模扩展收益转化为切实的福利改善。

这些结果表明，仅凭综合基准测试评估人工智能系统，可能会系统性地低估风险（通过长度偏倚暴露）和价值（通过选择性委托的期权价值）。因此，理解 AJI 对于预测采用情况、指导扩展投资，以及设计能够让可靠性可被发现而非偶然出现的工具，至关重要。

参考文献

Agrawal, A., J. S. Gans, and A. Goldfarb (2025年): “思维自行车经济学”，技术报告，美国国家经济研究局。Akerlof, G. A. (1970年): ““柠檬”市场”: 质量不确定性与市场机制，《经济学季刊》，84，488–500。Bardhi, A. and S. Callander (2026年): “相关世界中的学习”，《经济学年度评论》，51624，072515。Bresnahan, T. F. and M. Trajtenberg (1995年): “通用目的技术是‘增长引擎’吗？”，《计量经济学杂志》，65，83–108。Brynjolfsson, E., D. Li, and L. Raymond (2025年): “生成式人工智能在工作中的应用”，《经济学季刊》，140，889–942，发表于2025年2月4日。Burnell, R., W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B.Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela,M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo(2023): “重新思考人工智能评估结果的报告方式”，《科学》，380，136–138。Callander, S. (2011年): “通过试错进行搜索与学习”，《美国经济评论》，101，2277–2308。Carnehl, C. and J. Schneider (2025年): “知识探索”，《计量经济学》，93，623–659。Chiang, W.-L., L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang,B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica (2024年): “Chatbot Arena：一个基于人工偏好评估大语言模型的开放平台”，arXiv 预印本 arXiv:2403.04132。Dell’Acqua, F., E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg,S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2023): “探索锯齿状技术前沿：人工智能对知识工作者生产率和质量影响的现场实验证据”，技术报告 24-013，哈佛商学院工作论文，工作论文，2023年9月22日。Dubois, Y., B. Galambosi, P. Liang, and T. B. Hashimoto (2024年): “长度控制型 AlpacaEval：一种消除自动评估器偏倚的简便方法”，arXiv 预印本 arXiv:2404.04475。Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt(2021年): “衡量大规模多任务语言理解”，载于国际学习表征会议。Jovanovic, B. (1982年): “选择与产业演化”，《计量经济学》，50，649–670。Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray,A. Radford, J. Wu, and D. Amodei (2020年): “神经语言模型的缩放定律”，arXiv 预印本。

Karatzas, I. and S. E. Shreve(1991年)：《布朗运动与随机微积分》，施普林格，第2版。Lai, V., C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan(2023)：《迈向人机决策科学：实证人机研究中设计空间概述》，载于《2023年ACM公平、问责与透明度会议论文集》（FAccT '23），美国伊利诺伊州芝加哥：ACM，1369–1385。Liang, P., R. Bommasani, T. Lee, et al. (2022)：《语言模型的整体评估》，arXiv预印本 arXiv:2211.09110。Lin, S., J. Hilton, and O. Evans (2022)：《TruthfulQA：衡量模型模仿人类谬误的程度》，载于《计算语言学协会第60届年会论文集》（第1卷：长篇论文），爱尔兰都柏林：计算语言学协会，3214–3252。Nelson, P.(1970年)：《信息与消费者行为》，《政治经济学杂志》，78，311–329。Noy, S. and W. Zhang (2023)：《生成式人工智能生产率效应的实验证据》，技术报告，麻省理工学院，工作论文（未经同行评审），2023年3月2日。Rasmussen, C. E. and C. K. I. Williams(2006年)：《用于机器学习的高斯过程》，麻省理工学院出版社。Rosenberg, N. (1982年)：《通过使用学习》，载于《黑箱内部：技术与经济学》，剑桥大学出版社。Srinivas, N., A. Krause, S. Kakade, and M. Seeger (2010年)：《老虎机设定中的高斯过程优化：无遗憾与实验设计》，载于《第27届国际机器学习会议论文集》（ICML）。Srivastava, A. et al.(2023)：《超越模仿游戏：量化并外推语言模型的能力》，《机器学习研究汇刊》。Vakili, S., K. Khezeli, and V. Picheny (2021年)：《关于高斯过程老虎机中的信息增益与遗憾界》，《机器学习研究论文集》，130，693–702。Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman(2019)：《SuperGLUE：面向通用语言理解系统的更具挑战性的基准》，载于《神经信息处理系统进展》。Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman(2018)：《GLUE：自然语言理解多任务基准与分析平台》，载于《2018年EMNLP研讨会论文集：BlackboxNLP：自然语言处理神经网络的分析与解释》，比利时布鲁塞尔：计算语言学协会，353–355。

Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica(2023): 《使用 MT-Bench 和 Chatbot Arena 进行 JudgingLLM-as-a-Judge》, *arXiv* 预印本 *arXiv:2306.05685*。

A　在线附录：检查悖论的推广

正文将泊松覆盖作为便于处理的基线：间隔服从指数分布，按长度偏倚抽取的间隔服从伽马分布（2，λ）（命题 1），而且若干对象可以得到封闭形式。本附录记录哪些结论并不依赖泊松结构。关键在于，检查悖论是尺寸偏倚抽样的一种体现：均匀抽取的任务位置更可能落在覆盖过程中的较大“空洞”里。这一逻辑与具体分布无关，且适用于一维间隔以外的情形。

A.1 披露、法律责任与评估设计（简要讨论）

AJI 表明，政策与市场设计应少关注单一数值的基准分数，更多关注可靠性如何随任务而变化，以及用户会如何暴露于最差区域。两个自然的政策杠杆是：（i）披露和界面义务（校准、弃答、来源信息和不确定性沟通），以便实现选择性委托；（ii）评估标准要求报告尾部表现（或经验加权风险），而非只报告宏观平均值。责任制度也会与锯齿性相互作用：当错误成本被外部化时，提供方可能会理性地追求醒目的平均值，却在规律性和尾部安全性方面投入不足。

A.2 一般间隔分布的长度偏倚

令 $\{z_i\}$ 表示一条线上的知识点，令 X 表示相邻知识点之间（无偏的）间隔长度。在泊松基线中，$X \sim \text{Exp}(\lambda)$，其中 $\mathbb{E}[X] = 1/\lambda$。这里我们允许 X 服从在 $\mathbb{E}[X] \in (0, \infty)$ 处的任意分布。

定义4 (长度偏倚间隔**)**。令 X^* 表示包含一个均匀抽取的任务位置的间隔长度。X^* 的长度偏倚分布定义为

$$\Pr(X^* \in A) \equiv \frac{\mathbb{E}[X\,\mathbf{1}\{X \in A\}]}{\mathbb{E}[X]} \qquad \textit{for all measurable } A. \tag{52}$$

如果 X 具有密度 f_X，那么 X^* 的密度为

$$f_{X^*}(x) = \frac{x f_X(x)}{\mathbb{E}[X]}. \tag{53}$$

定义4是检查悖论的正式表述：任务在空间中均匀到达，而不是在间隔中均匀到达，因此间隔是按其长度成比例抽样的。

引理 8 （一般检查悖论恒等式）。若 $\mathbb{E}[X] < \infty$，则根据定义*4*，X^* 有明确定义；对于任意满足 $\mathbb{E}[X^{m+1}] < \infty$的 $m \geq 0$ ，

$$\mathbb{E}[(X^*)^m] = \frac{\mathbb{E}[X^{m+1}]}{\mathbb{E}[X]}. \tag{54}$$

特别地，若 $\mathbb{E}[X^2] < \infty$,

$$\mathbb{E}[X^*] = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]} = \mathbb{E}[X]\left(1 + \mathrm{CV}^2\right), \tag{55}$$

其中 变异系数 $\equiv \sqrt{\mathrm{Var}(X)}/\mathbb{E}[X]$。

证明。 根据定义4,

$$\mathbb{E}[(X^*)^m] = \int x^m \frac{x f_X(x)}{\mathbb{E}[X]}\, dx = \frac{\mathbb{E}[X^{m+1}]}{\mathbb{E}[X]},$$

并且通过写成 $\mathbb{E}[X^2] = \mathrm{Var}(X) + (\mathbb{E}[X])^2 = (1 + \text{变异系数}^2)(\mathbb{E}[X])^2$可得（55）。 □

间隔内的布朗桥公式保持不变：在给定实际间隔长度X^* 以及该间隔内的相对位置 $t \sim$ 均匀分布(0，1)的条件下,

$$\sigma^2(x) = X^*\, t(1-t).$$

因此，只要 $\mathbb{E}[X^2] < \infty$,

$$\mathbb{E}[\sigma^2(x)] = \mathbb{E}[X^*]\mathbb{E}[t(1-t)] = \frac{\mathbb{E}[X^*]}{6} = \frac{\mathbb{E}[X^2]}{6\,\mathbb{E}[X]} = \frac{\mathbb{E}[X]}{6}\left(1 + \mathrm{CV}^2\right). \tag{56}$$

公式（56）是本文与分布无关的核心：用户体验的 平均不确定性取决于（i）间隔的均值 长度 $\mathbb{E}[X] = 1/\lambda$ （规模），以及（ii）间隔的离散度 变异系数（规律性）。

A.3 规律性以乘法楔差的形式显现

基准测试和试点通常类似于间隔均匀 抽样：它们测试固定数量的任务类别或定义域，隐含地对每个间隔赋予相同权重。相比之下，实际使用具有长度偏倚。表达由此产生的测量差距的一种简便方法，是比较隐含的“可靠性比率”。

令 $\lambda \equiv 1/\mathbb{E}[X]$ 表示覆盖强度，并固定利害关系 q。在间隔均匀抽样下，与基准相关的均值方差为 $\mathbb{E}_{\text{gap-unif}}[\sigma^2] = \mathbb{E}[X]/6 = 1/(6\lambda)$。在长度偏倚抽样下，由（56）可得 $\mathbb{E}_{\text{use}}[\sigma^2] = \mathbb{E}[X](1 + \text{变异系数}^2)/6 = (1 + \text{变异系数}^2)/(6\lambda)$。

命题20 （基准可靠性与体验可靠性）。假设 $\mathbb{E}[X^2] < \infty$。定义

$$R^{\text{bench}} \equiv \frac{q}{\mathbb{E}_{gap\text{-}unif}[\sigma^2]} = 6\lambda q, \qquad R^{\text{exp}} \equiv \frac{q}{\mathbb{E}_{use}[\sigma^2]} = \frac{6\lambda q}{1 + \mathrm{CV}^2}.$$

Then

$$R^{\text{exp}} = \frac{R^{\text{bench}}}{1 + \mathrm{CV}^2} \le R^{\text{bench}}, \tag{57}$$

当且仅当变异系数 $= 0$ （间距完全规则）时取等号。对于泊松间隔（变异系数 $= 1$），$R^{\text{exp}} = R^{\text{bench}}/2$。

证明。 由 (56) 和定义可立即得出。 □

间隔分布 X	条件	变异系数(X)	$\mathbb{E}[X^*]/\mathbb{E}[X] = 1 + $变异系数^2
确定性间隔	—	0	1
指数分布（泊松间隔）	—	1	2
伽马分布(k, θ)（形状参数 k，规模 θ）	$k > 0$	$1/\sqrt{k}$	$1 + 1/k$
帕累托(α)（I型）	$\alpha > 2$	$\frac{1}{\sqrt{\alpha(\alpha-2)}}$	$\frac{(\alpha-1)^2}{\alpha(\alpha-2)}$

表 2：常见间隔分布的检查悖论乘数。对于帕累托(α)，当 $\alpha \le 2$,$\mathbb{E}[X^2] = \infty$ 时， $\mathbb{E}[X^*] = \infty$，因此乘数无界。

正文中所有基于均值的“盲目”结果，在以体验单位衡量可靠性后均可原样延伸。例如，盲目期望效用（在零点截断之前）可写为

$$\bar{U}_B = 1 - \frac{\mathbb{E}[\sigma^2]}{q} = 1 - \frac{1}{R^{\exp}},$$

因此，盲目采用阈值为 $R^{\exp} \ge 1$，即 $R^{\text{bench}} \ge 1 + \mathrm{CV}^2$。因此，方程（57）阐明了为何规模报酬可能被错误测量：基准改进只有除以（$1 + \mathrm{CV}^2$）后，才会转化为体验到的改进。

A.4 示例与尾部放大

表 2 列出了常见间隔族的检查悖论乘数。所有比率均为无尺度的：它们取决于形状参数 （规律性/尾部），而非 λ。

引理 8 要求 $\mathbb{E}[X^2] < \infty$，才能用 $1 + \mathrm{CV}^2$ 概括这一悖论。如果 $\mathbb{E}[X] < \infty$，但 $\mathbb{E}[X^2] = \infty$ （例如，帕累托分布中 $\alpha \in (1, 2]$），那么 $\mathbb{E}[X^*] = \infty$。从经济学角度看，这意味着用户体验的平均不确定性 $\mathbb{E}[\sigma^2]$ 发散，尽管朴素的“平均间隔”统计量确实存在。这是锯齿性的极端形式：罕见而巨大的空缺主导着实际体验。本文大部分内容有意避开这种情况，因为像 R 这样的基于均值的对象将不再具有信息价值，但其定性结论反而更加突出：覆盖的尾部行为变得至关重要。

A.5 超越一维：大小偏倚单元

在更高维的任务空间中，“间隔”不再是区间，而是稀疏区域。一种自然的几何类比是由知识点所诱导的沃罗诺伊镶嵌：每个点所拥有的任务空间区域，都比其他任何点更靠近该点。[15]

令 $\{C_i\}$ 为沃罗诺伊单元，并令 V 表示一个典型单元的随机体积。均匀抽取的任务位置落入单元 i 的概率与 $\mathrm{vol}(C_i)$成正比，因此，经历到的单元体积具有大小偏倚。

引理9 （沃罗诺伊镶嵌中的大小偏倚）。 设 V 为典型单元的体积， V^*为包含均匀随机选取位置的单元的体积。那么 V^* 具有大小偏倚：

$$\Pr(V^* \in A) = \frac{\mathbb{E}[V\,\mathbf{1}\{V \in A\}]}{\mathbb{E}[V]}.$$

[15]一个沃罗诺伊镶嵌是将度量空间划分为若干区域的一种方式，使得给定区域中的每个点 is 更靠近于 比起该区域定义种子点集合中的其他任何种子，更靠近该区域的定义种子点。

若 $\mathbb{E}[V^2] < \infty$,

$$\mathbb{E}[V^*] = \frac{\mathbb{E}[V^2]}{\mathbb{E}[V]} = \mathbb{E}[V]\left(1 + \mathrm{CV}(V)^2\right).$$

证明。 这与定义4和引理 8中的论证相同，只需将区间长度替换为单元体积。

□

引理9表明，检查悖论逻辑在任何维度下都成立：用户会过度暴露于较大的空区域，而衡量标准现在是单元体积，而非区间长度。更复杂的是，如何将单元几何结构映射到局部插值风险 $\sigma^2(x)$。在一维情形下，布朗插值可得桥方差的闭式表达 $\sigma^2 = X^* t(1-t)$。在高维情形下，相应的量取决于协方差结构以及最近锚点的几何布局。尽管如此，经济学上的基本机制依然存在：对“覆盖区域”的大小偏倚，会在间隔均匀评估与用户体验之间造成系统性差异。映射到 $\mathbb{E}[\sigma^2(x)]$ 中的乘法 差异具体如何形成，取决于插值风险如何随单元几何结构和协方差结构变化，因此，一维中的（1 +变异系数^2）因子未必能原样沿用。

A.6 对主文的启示

本附录　阐明了泊松覆盖在正文中的作用，以及它没有发挥的作用　　g.

• **泊松分布是一种便利设定，而非临界条件**。检查悖论与分布无关：它源于长度偏倚（定义4）。泊松分布仅仅使 X^* 伽马分布$(2, \lambda)$具有简洁的闭式表达，并为校准对象提供闭式结果。

• **规律性是第二种能力原语**。公式（56）表明，体验到的平均不确定性可分解为规模（λ）和离散度（变异系数）。这正是为什么即使对基准平均值影响甚微，“弥合最严重的间隔”的投资仍可能带来很高福利回报的形式化原因。

• **当**变异系数 > 0 **时，规模报酬会被系统性地错误测量**。命题20意味着，基准改进在 $R_{\text{基准}}$ 只有除以（1 +变异系数^2）后，才会转化为体验到的改进在$R_{\text{体验}}$。在泊松间隔下，该因子为 2；在更厚的尾部下，它可能大得多。

• **泊松分布发挥作用之处**。依赖局部方差完整分布的对象（例如校准后的使用份额、推理阈值和尾部损失）取决于 X^*的形状参数，而不仅仅是其均值。泊松分布提供了简洁的闭式基线；超越泊松分布会改变定量表达式，但保留锯齿性产生期权价值和尾部风险楔形差这一质性逻辑。

• **高维任务空间保留了相同的核心作用**。 在 $d > 1$ 维中，区间被沃罗诺伊单元取代，长度偏倚则转变为大小偏倚（引理9）。实证问题随之变为：覆盖区域的离散程度有多大？这种离散度又如何转化为相关任务表征中的插值风险？

简言之，检查悖论并非泊松过程所独有的统计学奇观。它是将不规则覆盖转化为系统性差距的机制，使易于衡量的内容与实际体验之间产生落差；正是这一落差推动了本文关于采用、规模化激励和设计的研究结果。

B 在线附录：实践中的基准测试与 AJI 调整

本附录汇总了广泛使用的大语言模型基准测试，并说明如何根据人工锯齿智能（AJI）对其醒目的得分进行系统性重新解读，并在可能的情况下加以调整。主论文所阐述的机制是检查悖论：用户并非均匀地遇到“定义域”或“间隔”，而是按照这些区域在任务空间中所占的比例遇到它们。这便造成了测量差距：一方面是对精选任务/定义域集合进行宏平均的基准测试，另一方面则是在部署中的实际体验可靠性。

B.1 一般性修正：从宏平均到实际体验可靠性 y

许多基准测试通过对一组定义域/任务的表现进行（近似）平均，报告单一分数。在 AJI 模型中，与之对应的做法是间隔均匀评估：任务空间的每个区域都只被计入一次，无论其所占空间比例有多大。

为使调整过程清晰透明，考虑一个总长度为 L 的离散任务空间，它被划分为 n 个区域（“间隔”），长度分别为 $(X_i)_{i=1}^{n}$ 和 $\sum_i X_i = L$。在区域 i 内，间隔内后验方差的平均值为 $X_i/6$（根据 $\sigma^2(x) = X_i t(1-t)$ 和 $\int_0^1 t(1-t)\,dt = 1/6$）。那么：

$$\mathbb{E}_{\text{bench}}[\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{6} = \frac{L}{6n}, \tag{58}$$

$$\mathbb{E}_{\text{user}}[\sigma^2] = \sum_{i=1}^{n}\Pr(\text{user in } i)\cdot\mathbb{E}[\sigma^2 \mid i] = \sum_{i=1}^{n}\frac{X_i}{L}\cdot\frac{X_i}{6} = \frac{1}{6L}\sum_{i=1}^{n}X_i^2. \tag{59}$$

该比率为

$$\frac{\mathbb{E}_{\text{user}}[\sigma^2]}{\mathbb{E}_{\text{bench}}[\sigma^2]} = \frac{n\sum_{i=1}^{n}X_i^2}{\left(\sum_{i=1}^{n}X_i\right)^2} = 1 + \mathrm{CV}_X^2, \tag{60}$$

其中，$\mathrm{CV}_X \equiv \sqrt{\mathrm{Var}(X)}/\mathbb{E}[X]$ 是间隔长度的变异系数。在泊松基线中，$\mathrm{CV}_X = 1$，因此，间隔均匀基准对平均可靠性的估计会乐观一倍：$\mathbb{E}_{\text{user}}[\sigma^2] = 2\,\mathbb{E}_{\text{bench}}[\sigma^2]$。

实践中可以调整什么？方程（60）表明，有两种可通过经验数据实施的基准调整。

(i) 暴露加权（可行时优先采用）。令 d 为基准测试领域（学科、任务、场景）的索引。如果基准报告了模型 m 的各定义域得分 $s_d(m)$

and 按权重 ω_d^{bench}（通常为均匀分布）进行汇总，则部署对齐得分为

$$S^{\text{dep}}(m) = \sum_d \omega_d^{\text{dep}}\, s_d(m), \qquad \sum_d \omega_d^{\text{dep}} = 1, \tag{61}$$

其中， ω_d^{dep} 根据使用日志、产品遥测数据或其他提示词频率数据进行估算。

(ii) 检查悖论式"参差性惩罚"（在无法获得暴露权重时用作诊断）。如果只能获得无权重或均匀权重的基准测试，可以报告跨定义域的离散度统计量，例如领域级错误率$e_d \equiv 1 - s_d$的 $\mathrm{CV}_d(e_d)$。在定义域规模与定义域错误率呈正相关的情形下（即 AJI机制），$\mathrm{CV}_d(e_d)$能够反映宏平均值与实际体验可靠性之间可能相差多大。若无额外假设，这并非结构估计量，但却是一种实用的诊断方法：离散度高，意味着检查楔形差距可能较大。

B.2 逐项基准测试：设计、使用情况与 AJI 调整

下面我们总结一些重要的基准测试，并说明进行 AJI式调整需要满足哪些条件。

GLUE 和 SuperGLUE（宏平均的自然语言理解基准套件）。GLUE 和 SuperGLUE 是早期用于概括多项任务进展的"单一数值"自然语言理解基准测试。二者均通过对各项任务指标进行（近似）无权重平均来汇总结果；若某项任务包含多个指标，则该任务的得分为这些指标的平均值。（Wang等人，2018，2019）

AJI 含义。 任务均匀聚合对应于差距均匀加权（每个区块获得相同权重）；(60)展示了相对于按使用加权经验的由此产生的楔形差距。如果某个组织的实际工作负载集中在一部分任务上（例如蕴含类任务），GLUE 风格的宏平均可能会误述实际体验到的表现。

调整。 (i) 在(61)中将任务均匀权重替换为部署权重 ω_d^{dep} 。报告跨任务的离散度（例如$_{\text{(ii)}}$ $\mathrm{CV}_d(e_d)$或任务得分第十百分位数），以揭示锯齿性。

MMLU（广泛的知识/推理学科测试）。 MMLU 评估57个学科上的多项选择题表现，并报告跨任务和示例聚合后的总体准确率。（Hendrycks等人，2021年）

AJI 含义。 MMLU 的总分是衡量均值 能力的便捷代理指标，但它无法揭示各学科之间表现有多不均衡（即 AJI 所说的"空洞"），也不会根据部署中各学科被遇到的频率（或利害关系）对其进行加权。

调整。 (i) 当可获得学科层级的分数时，计算各学科的部署权重（例如企业与消费者的构成比例），并在（61）中形成 S^{dep} 。(ii) 除均值外，还应纳入尾部汇总指标（表现最差五分位学科、错误离散度），以近似检查楔形差距。

BIG-bench（长尾能力套件）。 BIG-bench 包含 204 项任务，并将总体性能定义为各任务“归一化”首选指标的平均值。（Srivastava 等，2023）

AJI 含义。 由于 BIG-bench 刻意涵盖了许多稀有/长尾任务，因此自然应将其解读为探测难度的右尾，而非估算按使用加权的平均值。在 AJI 框架下，这很有用：意外失败和采用阻力恰恰源自尾部。

调整。 (i) 将 BIG-bench 视为尾部风险模块，并报告任务分数分布（例如分位数），而不仅仅是平均值。(ii) 对于与部署对齐的层级，将其与使用加权基准结合，采用以下混合方式：

$$S^{\mathrm{mix}}(m) = (1-\pi)\, S^{\mathrm{main}}(m) + \pi\, S^{\mathrm{tail}}(m),$$

其中，π 是部署环境中“尾部”查询所占的（估计）比例。

HELM（场景 × 指标矩阵）。 HELM 在许多场景下依据多项目标要求（不止准确率）评估语言模型，并明确将评估组织为按场景和指标划分的矩阵。（Liang 等，2022）

AJI 含义。 HELM 非常适用于 AJI，因为它已经着重关注覆盖率（许多场景）和多指标权衡。尚待解决的问题是聚合：部署很少会对各场景一视同仁。此外，AJI 强调各场景的利害关系各不相同，因此，与福利相符的聚合方式应同时纳入频率和利害关系。

调整。 (i) 使用 (61) 构建部署加权场景得分。(ii) 如果不同场景的利害关系差异显著，应根据 q_d^{-1} 的估算值（或与政策相关的损失权重）为场景赋予权重，以反映某些场景中的错误代价高得多。

TruthfulQA（对抗性真实性/误解陷阱）。 TruthfulQA 旨在测试模型在人工常常重复常见误解的情境中，能否给出真实答案；它涵盖 38 个类别，共有 817 个问题。（Lin 等，2022）

AJI 含义。 TruthfulQA 针对的是一种重要的失败模式，而非具有代表性的平均情形工作负载。在 AJI 视角下，最好将其视为对高成本尾部（“灾难性”区域）的概率质量和超调幅度的测量，这与正文中的盲区成本对象密切对应。

调整。 将 TruthfulQA（以及类似的安全性/真实性测试套件）用作尾部失败风险的估算指标，并与平均情形基准一并报告：

$$\text{Report:}\quad \big(S^{\mathrm{avg}}(m),\ \Pr(\text{tail failure} \mid m)\big).$$

这在结构上与 AJI 方案相呼应：将平均情形指标与尾部风险（或覆盖率）指标配对。这些基准对象并非旨在直接估计（U_C，s_C），因为后者取决于内生的弃答策略和使用分布。

MT-Bench 和 Chatbot Arena（交互式/聊天评估）。MT-Bench 是一组固定的多轮问题，采用大语言模型评审方法进行评分。（Zheng 等人，2023）Chatbot Arena 收集众包提示词下模型输出的成对人类偏好，并拟合统计偏好模型以生成排名/分数。（Chiang 等人，2024）

AJI 含义。 MT-Bench 更接近传统的间隔均匀基准：它由一组规模较小且固定的提示组成，因此也面临通常的代表性问题。从概念上说，Chatbot Arena 更接近按用户加权的 抽样，因为提示由用户提供，这在一定程度上内化了检查悖论逻辑：出现频率较高的查询类型更有可能被纳入。

调整。

- 对于 MT-Bench：按定义域（或推断出的主题）对提示进行分层抽样并应用部署权重，或者利用从产品遥测数据得出的抽样结果扩充提示集，使提示频率与部署情况相匹配。

- 对于 Chatbot Arena：重新加权（或对提示进行子抽样），使其符合目标用户群体和用例组合（企业与消费者、语言组合、安全关键型与休闲场景）。Arena 为估算 ω部署 d in (61) 提供了一个起点，但权重因环境而异。

关于基准实践中的"偏差校正"的说明。 对已知的、由机制引发的偏差进行校正，正日益成为基准测试中的标准做法（例如，自动偏好评估器中的输出长度偏差）。长度控制型 AlpacaEval 为大语言模型评估中的此类校正策略提供了一个实例。（Dubois 等人，2024）本文倡导的 AJI 调整在理念上与之类似：其目的是校正抽样/加权 偏差，避免因对稀疏且容易发生故障的区域赋予过低权重，而系统性地高估实际体验可靠性。

B.3 实用报告模板（仅添加最少的 AJI 内容）

对于希望尽量沿用现有做法，同时纳入 AJI 的评估者，最简模板如下：

(i) **平均分数** （现状）：现有的基准测试汇总分（例如宏平均）。(ii) **离散度/尾部**：至少报告一个跨基准测试领域/任务的离散度统计量（例如 $\mathrm{CV}_d(e_d)$、领域得分第十分位数或最差的 k 个领域）。(iii) **部署重新加权（可行时）**：针对少数典型环境（例如"消费者聊天""企业办公工作流""编程助手"），公布按部署加权的得分 S^{dep}，如式（61）所示。(iv)**尾部风险模块**：单独报告有针对性的安全/真实性基准测试（例如 TruthfulQA），而不是将其纳入同一个宏平均。

表3：概要：常见基准设计如何映射到 AJI，以及调整需要什么。

基准	默认聚合	AJI问题（为何标题可能产生误导）	AJI式调整（需要什么待办）	所需数据
GLUE / 超GLUE（Wang 等人，2018，2019）	大多为任务均匀平均值	差距均匀加权忽略暴露度和尾部	根据部署组合重新加权任务-合；报告离散-分位数	任务级分数；使用权重
MMLU（Hendrycks 等人，2021年）	总体跨多个学科准确率	Mean 掩盖了不均衡的学科可靠性；子-学科暴露度因部署而异	重新加权学科；公布尾部指标（最差十分位学科）	学科得分；领-域使用份额
BIG-bench（Srivas-tava 等，2023）	204 项任务的平均值任务（归一化）	测试套件的设计具有长尾特征；设计；平均值并非“典型使用情况”	视为尾部模块；报告分数分布；与主基准测试结合，使用 π	任务分数；尾部估-查询值分享 π
HELM (Liang 等，2022)	场景 × 指标矩阵；聚合变化	均匀分布 场景聚合 很少匹配 部署；利害关系因场-景而异	部署- and 利害关系-加权场景聚合	场景指标；场景使用情况和利害关系权重
TruthfulQA (Lin 等，2022）	针对性的 真实性失败率参数	不是平均情形；衡量 误解“陷阱”尾部	单独报告为尾部风险；在需要时通过混合权重进行组合	Tail 基准得分；尾部流行率估计
MT-Bench（Zheng 等人，2023）	固定 提示集，大语言模型评审评分	小型固定集合存在间隔-均匀分布； may miss 最严重的漏洞	扩展/分层抽样 提示；重新加权 by 部署提示分布	提示词 元数据；部署提示词混合
聊天机器人 Arena（Chiang et al., 2024年）	众包提示 +人工成对偏好	更接近 to 用户加权的 抽样，但反映了 Arena 的用户群体	对提示/用户重新加权，以匹配目标部署；按领域/安全类别分层抽样	提示词 数据集；目标人群权重