arXiv:2601.04461v1 [cs.CL] 8 Jan 2026

用户误判自己对人工智能写作辅助的偏好

Vivian Lai[1] **Zana Buçinca**[2]**Nil-Jana Akpinar**[1] **Mo Houtti**[1] **Hyeonsu B. Kang**[1]**Kevin Chian**[1] **Namjoon Suh**[1] **Alex C.Williams**[1]

Microsoft [2]麻省理工学院viv.lai@microsoft.com

摘要

主动式人工智能写作助手需要预测用户何时需要起草帮助，但我们尚缺乏对影响偏好的因素的实证理解。通过一项析因情境研究，50 名参与者进行了 750 次两两比较；我们发现，写作构思所需的努力程度主导了决策（$\rho = 0.597$），而紧迫性不具备预测能力（$\rho \approx 0$）。更重要的是，用户表现出显著的感知—行为差距：尽管紧迫性是最弱的行为驱动因素，他们在自我报告中却将其排在首位，形成了彻底的偏好倒置。这种不一致会造成可衡量的后果。依据用户陈述偏好设计的系统准确率仅为 57.7%，甚至不如朴素基线；而采用行为模式的系统则显著更高，达到 61.3%（$p < 0.05$）。这些发现表明，依赖用户内省进行系统设计会主动误导优化，这对主动式自然语言生成（NLG）系统具有直接影响。

1 引言

大语言模型催生了能力日益强大的人工智能写作助理，但一个根本性的设计问题仍未解决：系统应当在何时主动提供起草帮助？响应式助理会回应用户的明确请求，而主动式系统则需要根据上下文推断适当的时机。以不确定情况下采取行动的成本与益处为依据，在适当时机主动介入有助于改善用户经验；时机不当则会增加中断成本并令用户感到沮丧（Horvitz，1999年；Iqbal 和 Bailey，2008年；Mc-Farlane，2002年）。尽管此类系统已广泛部署于电子邮件客户端（例如 Smart Reply（Kannan 等人，2016年）、Smart Compose（Chen 等人，2019年）），设计者仍缺乏有力证据，无法确定哪些上下文线索能够可靠地预测用户实际是否需要帮助。此外，陈述偏好和主观测量往往与观察到的行为和性能并不一致，而建议用户界面也会改变写作构思，为主动式助理带来权衡取舍（Buçinca 等人，2020年；Poursabzi- Sangdeh 等人，2021年；Buschek 等人，2021年）。

人们通常直觉上认为应在时间压力下提供协助，但中断研究强调，相较于紧迫性，任务结构和断点是预测较低干扰时机的更佳因素（Adamczyk 和 Bailey，2004年；Iqbal 和 Bailey，2008年）。另一种做法是公开用户控制项和偏好设置，这与人机交互指南一致（Amershi 等，2019年），同时认识到自我报告未必能准确反映对协助的需求（Buçinca 等，2020年）。

我们通过电子邮件起草来研究这些问题。这一领域为研究主动协助偏好提供了多项方法论优势。电子邮件是一项高频写作任务，具有可测量的上下文因素，并且作为已部署的人工智能助手应用，具备明确的生态效度。电子邮件起草体现了主动式自然语言生成系统面临的核心挑战，即仅凭上下文信号、在没有明确请求的情况下，推断用户何时希望获得协助。

通过析因情境研究，我们考察哪些电子邮件属性能够预测对人工智能助手的偏好。我们操纵四个维度：紧迫性、构思努力、发件人重要性和电子邮件类型，共设置16个情境。50名参与者进行了750次成对比较，并报告了其陈述偏好的驱动因素。我们回答三个研究问题：（1）哪些电子邮件属性驱动对人工智能写作辅助的偏好？（§4）（2）用户能否准确感知驱动其人工智能助手偏好的因素？（§5）（3）机器学习模型能否根据行为模式预测偏好，感知—行为差距的代价是什么？（§6）

我们的研究结果揭示了显著的错位。用户认为紧迫性最重要（自我报告的平均排名：4 项中排第 2.06），但行为分析显示，紧迫性的预测能力几乎为零

（$\rho \approx 0$）。相反，构思努力程度占据主导地位（$\rho = 0.597$）。这代表着偏好的彻底逆转，揭示出较差的元认知准确性：紧迫性在陈述偏好中排名第一，但在行为上排名最后；而努力程度则呈现相反模式。这表明用户无法准确感知哪些因素驱动着自身偏好。

为量化这种错位，我们训练机器学习模型来预测个体偏好，并比较三种基于规则的策略，即均匀加权、来自用户调查的陈述偏好加权，以及基于显 revealed 偏好的行为加权。陈述偏好方法的准确率仅为 57.7%，低于朴素基线；而行为加权达到了显著更高的 61.3%（$p < 0.05$）。

我们作出以下三项贡献：

1. 我们提供了实证证据，表明组合性努力程度，而非紧迫性，驱动着对人工智能写作辅助的偏好。这使设计重点从时间信号转向认知负荷检测。

2. 我们展示了感知—行为差距：用户错误地识别了自己的偏好，反映出较差的元认知准确性。这表明，自然语言生成系统应从隐式行为反馈中学习，而不是依赖明确的用户调查。

3. 我们量化了感知—行为差距的代价，表明行为模式的预测准确率达到 61.3%，显著优于准确率为 57.7% 的陈述偏好加权设计（$p < 0.05$）。这说明，与依据观察到的行为进行设计相比，依据用户自我报告进行设计会带来实际成本。

这项工作通过以行为为依据来设计主动式协助，揭示用户内省的局限性，并验证可用于部署的见解，推动了以人为中心的人工智能发展。

2 相关工作

电子邮件管理与主动式人工智能辅助。长期以来，电子邮件管理研究一直倡导并设计更智能的辅助功能（Whittaker 和 Sidner，1996年；Bellotti 等人，2003年），指出用户在处理大量电子邮件和分配回复投入方面面临系统性挑战（Dab- bish 和 Kraut，2005年）。综合以往研究，我们聚焦于影响电子邮件行为的四个上下文因素：紧迫性（Tyler 和 Tang，2003年），构思努力（Bellotti 等人，2003年；Whit- taker 和 Sidner，1996年）、发件人重要性（Dab- bish 和 Kraut，2005年）以及电子邮件类型（Mackay，1988年；Whittaker 和 Sidner，1996年；Ducheneaut 和 Bellotti，2001年）。这些维度为我们的因子设计提供依据，同时也代表了人工智能辅助系统中实际相关的特征。

大语言模型的近期进展使主动式写作辅助成为可能（Lee 等人，2022年；Buschek 等人，2021年）。在自然停顿点提供主动式协助，能够缩短反应时间、减少挫折感并降低中断成本（Horvitz，1999年；Cutrell 等人，2001年；Iqbal 和 Bailey，2008年）。然而，时机不当的协助会打断工作流程（Iqbal 和 Horvitz，2010年；Czerwinski 等人，2004年；Cutrell 等人，2000年）。尽管许多已部署的工具仍是被动响应式的，即使是知名的电子邮件助手（Chen 等人，2019年；Kannan 等人，2016年）也很少报告经过实证验证的时机策略。我们的工作首次系统地探究了哪些上下文因素会影响用户对人工智能写作辅助的需求。

人工智能系统中的用户感知。近期研究凸显了用户的认知与其在人工智能系统中的实际行为之间存在偏差。基于代理任务和主观测量的评估，无法预测实际决策表现（Buçinca 等人，2020年）。由于人工智能能力的不确定性以及输出的复杂性，人机交互尤其具有挑战性（Yang 等人，2020年）。成熟的设计指南强调用户控制，以及理解或纠正系统行为的机会（Amershi 等人，2019年），但当用户误解自身需求或人工智能的能力时，这些方法可能难以奏效。在人工智能写作辅助中，提供多条建议会改变不同用户群体的写作行为和感知有用性（Buschek 等人，2021年）。然而，先前研究并未操纵上下文因素，也未量化系统后果。我们通过析因方法和计算验证来弥补这些不足。

3 方法

研究设计。 我们采用了 $2 \times 2 \times 2 \times 2$析因情境设计，对四个二元电子邮件维度进行操纵：紧迫性、构思努力、发件人重要性和电子邮件类型，由此生成了 16 个情境，并使用 GPT-5 生成以确保一致性（附录 A.1）。50 名在职业情境中经常处理电子邮件的参与者完成了 15 项

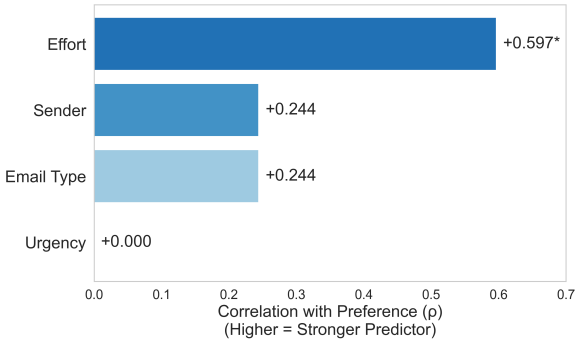

图1：通过斯皮尔曼相关性与布拉德利-特里偏好强度衡量的电子邮件维度预测强度。撰写所需努力程度是唯一具有统计显著性的预测因子，占据主导地位，而紧迫性显示出近乎为零的相关性。发件人重要性和电子邮件类型呈现出较弱且不显著的趋势。*p <0。05。

成对比较，每人随机抽取自所有 120 种可能的配对（$\binom{16}{2}$），选择他们更倾向于使用人工智能写作辅助处理的情境，并说明理由。退出调查收集了参与者对四个维度的排序、人工智能使用模式和人口统计信息。研究界面和完整的退出调查请见附录A.4。

分析方法。 我们使用布拉德利-特里模型从成对比较中推导偏好强度，随后计算斯皮尔曼相关性，以识别行为驱动因素（研究问题1）。我们通过有序逻辑回归和双因素方差分析验证了研究结果。针对研究问题2，我们将每位参与者的陈述偏好与显露偏好之间的相关性计算为元认知准确性，并通过方差分析和 t 检验检验预测因子。作为定量分析的补充，我们对750条理由和50份调查反思进行了由大型语言模型辅助的主题编码（附录A.2）。

针对研究问题3，我们使用五折交叉验证训练了四类模型（逻辑回归、随机森林、梯度提升、神经网络），并采用浅层架构以防止过拟合。我们比较了三种固定权重策略（均匀、陈述偏好、行为），以量化感知—行为差距成本。完整的模型规格见附录A.3。

4 哪些电子邮件属性影响用户对人工智能写作辅助的偏好？

我们根据 50 名参与者进行的 750 次成对比较，计算了全部 16 个情境的布拉德利-特里偏好强度。各维度与这些情境层级偏好强度之间的斯皮尔曼相关性显示，努力程度是 AI 起草偏好的主要决定因素

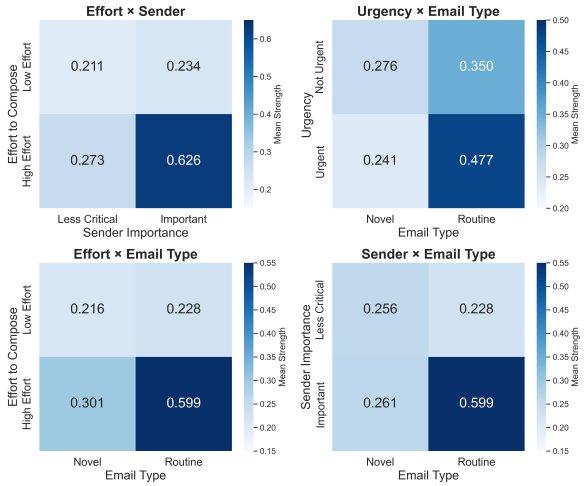

图2：关键的两两维度交互作用显示了布拉德利-特里偏好强度的均值。努力程度 ×发送者在两个因素一致时产生最高偏好（重要发件人 ×高投入：$\bar{\pi} = 0.626$），体现了“高风险利害、高负担”的启发式。发送者 × 类型表明，发件人重要性主要影响常规电子邮件（$\bar{\pi} = 0.599$），对新颖电子邮件的影响微乎其微。

($\rho = 0.597$, $p < 0.05$)，而紧迫性几乎没有预测能力（$\rho \approx 0$, $p = 1.0$）。发件人重要性（$\rho = 0.244$, 倾向于重要发件人）和电子邮件类型（$\rho = 0.244$, 倾向于常规而非新颖的电子邮件）呈现微弱且不显著的趋势，作用次于努力程度（图1）。有序回归验证了这些模式。高努力程度电子邮件被选中的几率是其他电子邮件的 13.68 倍（$p < 0.05$），而紧迫性没有显著影响。两项分析均证实，努力程度是主要驱动因素，而紧迫性没有可靠影响。

这些维度并非孤立发挥作用，而是呈现出显著的交互效应（图2）。高努力程度与重要发件人相结合时，用户偏好最高（$\bar{\pi} = 0.626$），这表明用户采用了一种乘法式的“高风险利害、高负担”启发式。相反，紧迫性的影响取决于上下文。用户偏好在处理紧急常规电子邮件时获得人工智能帮助（$\bar{\pi} = 0.477$），但处理紧急新颖电子邮件时则不然（$\bar{\pi} = 0.241$），这或许反映了用户对人工智能生成可预测回答的信任度更高。尽管双因素方差分析证实努力程度具有主效应（$F = 5.294$, $p < 0.05$），但交互效应未达到统计显著性。然而，机器学习特征重要性分析（§6.1）使用全部 750 个单独比较结果，而非 16 个汇总后的情境均值，发现这些交互作用解释了 61.7% 的预测方差。这表明确实存在交互结构，只是传统方差分析的统计效能不足以检测到它。

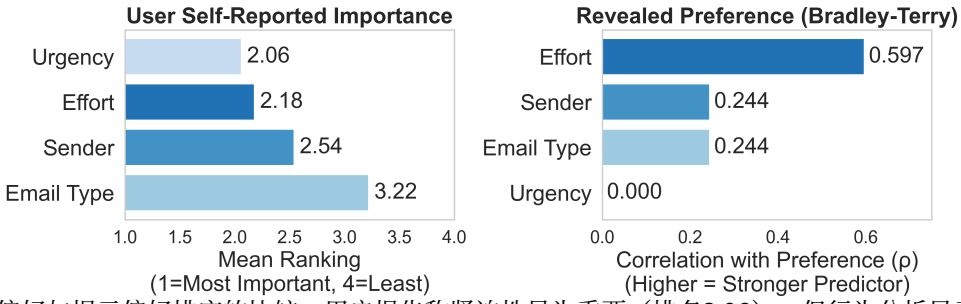

图3：陈述偏好与揭示偏好排序的比较。用户报告称紧迫性最为重要（排名2.06），但行为分析显示其不具备预测能力（$\rho \approx 0$）。努力程度是实际最强的驱动因素（$\rho = 0.597$），却只获得中等程度的重要性评价（排名2.18）。这种不显著的相关性掩盖了极端情况下的系统性倒置，表明陈述偏好将优先考虑错误的因素。

4.0.1 理由的定性分析

对 750 条即时理由进行的主题分析（每条回答平均包含 1.03 个主题）既印证了量化分析呈现的模式，也使其更为复杂。**紧迫性和时间压力**是被提及最多的因素（20.3%）；用户指出，截止期限和时间压力会促使他们寻求人工智能协助。这带来了一个疑问：用户最常提及紧迫性，但它对行为的预测能力几乎为零。**起草复杂度**是提及频率第二高的主题（14.7%）；用户明确表示，希望人工智能协助处理多步骤、需要大量综合的任务，以及制定全面的过渡计划，这与努力程度和行为之间最强的相关性一致（$\rho = 0.597$）。用户也会主动拒绝在简单的是非回答中使用人工智能（**任务简单——无需人工智能**，12.8%），证实内容组织负担会影响偏好。

共现分析揭示了能够解释这些模式的复杂决策启发式规则。用户会根据任务可预测性来校准对AI的信任：**模板或历史背景**（10.1%）经常与感知到的人工智能能力共同出现，用户表示“格式已经存在，因此人工验证会更快”。这种条件性信任模式表明，模板能够同时减少努力程度并且提升对AI准确性的信心。一些用户在说明人工智能使用的理由时也提到了**易于核验**（1.3%），例如“对我来说，验证正确性很容易”。这里的验证是指在发送前检查AI生成的草稿。用户更愿意采用人工智能，因为他们可以快速审查输出并发现任何错误。这使设计挑战从最大化AI准确性转变为最小化验证摩擦，并表明诸如置信度分数或突出显示AI生成的内容等特征，能够扩大风险规避型用户的采用率。

一个令人惊讶的发现是**避免高-风险或敏感任务** 的情况很少见（4.8%），这表明，在可以进行验证的情况下，敏感性顾虑并不会主导人工智能采用决策。此外，用户将人工智能视为用于**头脑风暴或创意协助**的认知伙伴（3.9%），而不仅仅是自动化工具，这表明除了起草之外，协作功能也具有发展潜力。有关主题频率和更多见解，请参阅附录A.2.2 。

5 用户能否准确感知哪些因素影响他们对 AI 助手的偏好?

比较自我报告排名与显露偏好，揭示了感知—行为差距（图3）。尽管紧迫性的行为预测能力近乎为零（$\rho \approx 0$），用户仍将其排在最重要的位置（均值排名：2.06）。作为最强驱动因素，努力程度（$\rho = 0.597$）获得的自我报告重要性却仅属中等（均值排名：2.18）。这体现了两端的偏好倒置：用户排名最高的维度对行为毫无影响，而实际驱动作用最强的因素，其陈述重要性仅居中游。这表明，若依据陈述偏好进行设计，就会优先考虑最弱的驱动因素，同时低估最强驱动因素。

5.1 任务后反思的定性分析

对 50 份调查回复进行主题分析（每份回复平均包含 2.8 个主题），证实了这种认知与行为之间的差距。回顾性陈述中，**紧迫性／时间敏感性**占主导地位，出现在 66% 的回答中；与即时理由（20.3%）和行为现实（$\rho \approx 0.000$）相比，这一比例被大幅高估。相比之下，**努力程度与效率**仅出现在 32% 的回顾中，相较于其实际主导地位（$\rho = 0.597$），这一比例被低估了。这种在回顾性陈述中过度报告紧迫性、少报努力程度的模式，与

行为现实形成了鲜明对比。

除了紧迫性与努力程度之间的分歧外，调查反馈还揭示了用户如何理解自己的决策过程。**数据可用性**尽管数据可用性并非实验操纵的维度，却在 28% 的回答中出现，这反映出用户会将构成性努力拆解为信息搜集与起草复杂度。关于错误认知的关键证据，仍是用户系统性地高估紧迫性、低估努力程度。完整的主题对比见附录A.2.2。

5.2 元认知准确性的个体差异

我们探讨某些用户特征是否与陈述偏好和显露偏好之间更高的一致性相关。个体差异分析显示，与使用频率或电子邮件数量相比，用户自评对人工智能技术的熟悉程度（以 7 分制衡量）更能预测元认知准确性。熟悉度较高的用户（得分 5–7）所陈述的偏好与实际偏好更加一致（$\rho = 0.090$），而熟悉度较低的用户（得分 1–3）则表现出相反的认知（$\rho = -0.550, p < 0.05$）。这种错误认知体现在各维度的排序上：熟悉度较低的用户将紧迫性列为首要考虑因素（平均排序：1.00），却将努力程度列在最后（3.25），两者相差 2.25 个名次。相比之下，熟悉度较高的用户对各维度的排序几乎均衡（紧迫性：2.08，努力程度：2.05），表明他们准确理解了努力程度的重要性。

人工智能使用频率和电子邮件数量都无法显著预测元认知准确性，这表明心理准备度比行为经验更为重要（见附录13 中的A.5）。也就是说，对人工智能技术感到自在的用户更能自我觉察自己的偏好驱动因素；而无论他们使用人工智能或处理电子邮件的频率如何，不自在都会加剧他们对紧迫性相较于努力程度的误判。

6 机器学习模型能否根据行为模式预测偏好？感知—行为差距会带来什么代价？

§4 和 §5中的行为研究结果表明，用户偏好在起草高努力程度的电子邮件时使用人工智能写作辅助，但他们认为紧迫性会左右自己的决策，这揭示了显著的感知—行为差距。要将这些发现应用于实用的自然语言生成系统，还需回答两个关键问题。首先，机器学习模型能否根据电子邮件属性准确预测用户偏好？

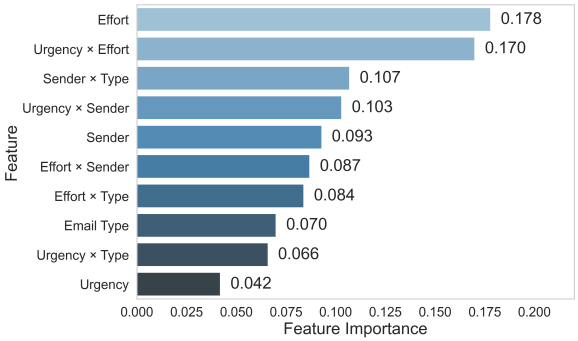

图 4：使用包含所有两两交互作用的随机森林进行偏好预测时的特征重要性。努力程度主效应（17.8%）是最重要的单一特征，其次是努力程度 × 紧迫性交互作用（17.0%）。交互作用的总占比为 61.7%，高于主效应的 38.3%，体现了偏好结构的协同作用。然而，使用交互作用训练的模型（准确率为 59.5%）表现逊于仅使用主效应的模型（准确率为 59.7%），这表明交互作用虽提升了特征重要性，却未改善预测性能。

其次，感知—行为差距会造成多大代价？基于陈述信念设计的系统与以行为为依据的替代方案相比，表现如何？我们通过机器学习验证（§6.1）和系统设计比较（§6.2）来回答这些问题。

6.1 预测验证

我们首先仅使用 4 个主效应特征，建立基线预测性能。所有模型的预测准确率均约为 60%（表1），高于多数类基线（55.2%），且显著优于随机猜测（$p < 0.001$）。为检验交互作用能否在此基线上进一步提升预测效果，我们随后使用全部 10 个特征（4 个主效应 + 6 两两交互作用）训练模型。令人意外的是，加入交互作用并未带来预测收益：使用全部 10 个特征训练的模型均值准确率为 59.5%，而仅使用主效应特征的模型为 59.7%。

鉴于特征重要性分析显示交互作用主导着预测重要性，这一结果似乎与之矛盾。当我们考察使用全部 10 个特征训练的随机森林模型（性能最佳的模型）时，交互作用占特征重要性的 61.7%，而主效应占 38.3%（图 4）。努力程度的主导地位既体现在其主效应（17.8%，单个最重要的特征）中，也体现在其交互作用中，尤其是努力程度× 紧迫性（17.0%）、努力程度 × 发送者（8.7%）以及

	全部特征 (10)		主效应 (4)	
模型	Acc.	F1	Acc.	F1
逻辑回归 0.587		0.657	0.580	0.643
随机森林 0.608		0.719	0.605	0.714
梯度提升	0.588	0.709	0.600	0.713
神经网络	0.599	0.692	0.603	0.698
Mean	**0.595**	**0.694**	**0.597**	**0.692**

表1：完整特征集（4 个主效应 +6 两两交互作用）与仅包含主效应的特征集之间的机器学习模型性能比较。所有模型均采用五折交叉验证进行训练。主效应的均值准确率略高（59.7% 对 59.5%），表明交互作用并未带来预测收益。完整指标见附录A.3。

策略	Acc.	精确率	召回率	F1
均匀	0.589	0.586	0.870	0.700
陈述偏好	0.577	0.568	0.981	0.719
最佳机器学习模型	0.605	0.614	0.855	0.714
行为	**0.613**	**0.608**	**0.841**	**0.706**

表2：基于规则的策略与最佳机器学习模型（包含4个主效应的随机森林）的性能比较。简单的行为规则与机器学习模型表现相当，这表明正确的特征加权比算法复杂度更重要。行为规则和机器学习模型在准确率上均显著优于陈述的偏好（行为规则和机器学习模型：$p < 0.05$）；F1值差异在统计学上并不显著。采用连续性校正的麦克内马尔检验。

努力程度 × 类型（8.4%）。除努力程度外，其他值得注意的交互还包括发送者 × 类型（10.7%）和紧迫性 ×发送者（10.3%）。

为什么交互项会主导特征重要性，却无法带来预测收益？答案在于认识到，交互项捕捉的方差已经由二元特征空间中的主效应解释。由于努力程度出现在多个重要性较高的交互项中（努力程度 ×紧迫性、努力程度 × 发送者、努力程度 × 类型），其预测信号会在不同交互项中被重复计算。当努力程度较高时，它的许多交互项也会非零，但这种冗余并未带来主效应本身之外的新预测信息。性能差异不大，这表明在我们的二元特征实验设置中，仅主效应就能有效捕捉偏好结构。

6.2 系统设计策略比较

为评估感知—行为差距的代价，我们比较了三种基于规则的加权策略

策略	紧迫性	努力程度	发送者	Type
均匀	25%	25%	25%	25%
陈述偏好	40%	30%	20%	10%
行为	0%	55%	22.5%	22.5%

表3：基于规则的策略的特征权重。均匀策略赋予各项同等重要性。陈述偏好反映用户调查排名。行为策略反映 Bradley-Terry 相关性强度，其中紧迫性的相关性接近于零。

与最佳机器学习模型进行比较。由于仅主效应就已足够，且交互作用不带来预测收益，因此我们采用包含 4 个主效应的随机森林（准确率为 60.5%，与包含 10 个特征的版本的 60.8% 相当），如表2所示。表 3 展示了各策略的特征权重。

陈述偏好策略的准确率为57.7%，低于基线的58.9%，这体现了第§5节所述的感知—行为差距。相比之下，行为加权的准确率达到 61.3%，比陈述偏好提高了3.6个百分点（$p < 0.05$）。最佳机器学习模型的准确率为 60.5%，显著高于陈述偏好（$p < 0.05$）和均匀基线（$p < 0.05$），但在统计上与行为加权无显著差异。

从这一比较中可以得出两个关键结论。首先，感知—行为差距会造成可衡量的代价。陈述偏好的表现不佳，源于权重反向：尽管紧迫性的预测能力为零，却将其权重设为 40%；而行为加权则正确地将努力程度置于首位，权重为 55%。这种权重反转导致陈述偏好的表现显著逊于行为方法和机器学习方法。其次，简单的行为规则与机器学习模型的表现不相上下，表明正确的特征加权比算法复杂度更重要。权重一旦设定正确，机器学习便不会带来额外收益；行为规则和机器学习的准确率均约为 61%（性能上限将在局限性部分 §9 中讨论）。这些发现对部署策略具有启示意义，详见§7.2。

7 讨论

7.1 用户为何会误判自己的偏好?

§5中记录的感知—行为差距提出了一个根本性问题：既然紧迫性没有预测能力，为什么用户会误将其视为主要决策因素？我们提出

两种相辅相成的解释。

社会赞许偏差 (Paulhus，1991年；Crowneand Marlowe，1960年) **可能会使明确的排序倾向于符合规范、易被接受的理由**。承认“我想借助人工智能来减少认知努力”，听起来不如“我需要帮助处理有时限要求的电子邮件”那么专业。人机交互和调查方法学领域的相关研究表明，主观测量和陈述偏好可能与行为不一致，从而误导评估（Buçinca et al.，2020年；Bertrandand Mullainathan，2001年）。支持这一解释的是，对任务后问卷反思进行的主题分析发现，尽管质量和个性化方面的顾虑并不能预测用户实际偏好在哪些电子邮件中获得 AI 助手，他们却频繁提及这些顾虑。这一模式表明，即使这些顾虑实际上并未影响他们对 AI 助手的偏好，用户仍会援引专业上易被接受的理由，以维护自己谨慎、重视质量的专业人士形象。

可得性启发式 (Tversky and Kahne-man，1973) 指的是，紧急情境等更显眼或情绪色彩更强烈的事件更容易被回忆起来，因此可能在人们的判断中被过分强调。参与者可能会想起紧急电子邮件，因为时间压力会造成紧张，使这类经历在认知上格外突出，即使紧迫性实际上并未影响他们的偏好决策。相比之下，高努力程度的电子邮件尽管需要更多支持，却可能给人一种常规事务的感觉，导致用户低估努力程度的影响。§5中记录的回顾性阐述模式为这一重构过程提供了实证支持：紧迫性提及比例从 20.3% 上升至 66%，而努力程度提及比例仍然偏低。

这两种机制可能会依次发挥作用，并相互强化。当被要求回顾性地解释自己的偏好时，参与者首先会通过可得性启发式提取容易获得的记忆。紧急电子邮件带有强烈的情绪色彩并会造成压力，使其在认知上更加突出，也更容易被召回。相比之下，高努力程度的电子邮件尽管需要大量认知资源，却可能让人感觉属于常规事务，导致参与者低估努力程度的影响。一旦紧迫性作为显著因素被召回，社会赞许偏差便会进一步影响参与者如何报告这些记忆。围绕紧迫性来表述辅助需求（“帮助我赶上截止期限”）比承认是在逃避努力（“帮助我避免处理认知负担较重的工作”）听起来更符合职业规范，尤其是对于其职业……

sional identity centers on handling complex ta sks.

§5 中记录的元认知准确性的个体差异支持这种双重机制解释。对人工智能的舒适度调节了这一差距，但并非通过改变实际偏好，而是通过减轻社会赞许压力。对人工智能技术感到舒适的用户，不太需要通过符合规范的紧迫性表述来证明自己寻求帮助的合理性。相反，低舒适度用户可能会受到这两种机制的叠加影响：更强的焦虑会提高紧迫性事件的可记忆性，而对使用人工智能是否恰当的不确定感则会加剧印象管理方面的担忧。这解释了为何行为经验（使用频率、电子邮件数量）与元认知准确性没有关系，而心理准备度却有关系。对人工智能技术的舒适度减弱了基于可得性的提取以及经过社会过滤的报告所产生的扭曲效应。

7.2 主动式自然语言生成系统的设计启示

我们的研究结果对以人为中心的自然语言生成中的一个基础性假设提出了挑战：明确的用户反馈能够可靠地指导系统优化。基于陈述偏好的加权方法表现不如朴素基线，这表明，根据用户所说的内容进行设计，其预测效果客观上比行为方法更差。我们为主动式自然语言生成系统设计提出三项核心原则。

认知负荷检测优先于时间紧迫性。在偏好决策中，构思努力程度占主导地位，而紧迫性没有预测能力，这与人们普遍认为应在时间压力下提供帮助的设计直觉相悖。系统应检测组合复杂性的信号，而非时间信号。定性分析显示，用户会在多步骤、以综合为主的任务中寻求帮助（占理由的 14.7%），同时会主动拒绝人工智能对简单回答的协助（占 12.8%）。这表明，主动式系统应识别任务结构指标，例如响应复杂度、多个利益相关方之间的协调要求以及综合需求，而非截止日期临近程度或将电子邮件标记为紧急。

尽管我们的研究证实，努力程度会影响偏好，但如何在生产系统中实现努力程度检测仍是一项尚待解决的挑战。可能的信号包括：通过线程长度和引用内容判断综合要求，通过收件人多样性检测协调复杂性，以及通过问题结构区分简单的确

认与多部分回复。努力程度与发件人重要性之间的交互作用表明，系统应采用情境加权，而非统一阈值；对于高风险沟通，应更重视努力程度检测。未来的研究应验证哪些具体信号最能预测不同沟通情境下的内容组织负担。

基于隐式行为反馈，而非显式偏好征询。 陈述偏好会主动误导优化。要求用户说明何时需要 AI 助手的设置页面，假设用户能够内省并了解自己的决策驱动因素，而我们的研究结果表明，这种假设并不可靠。系统应从隐式使用模式中学习，例如用户接受哪些建议、回复前斟酌了多久，以及他们多频繁地推翻建议，而不是要求用户预先配置。

除了从使用信号中学习，系统还应主动呈现上下文线索，以降低采用阻力。提供协助时，突出显示模板可用性或类似的过往邮件，既表明所需努力程度较低，也能增强用户对人工智能输出质量的信心。这种条件性信任模式（§4.0.1）表明，与笼统地询问“需要帮助吗？”相比，采用“我可以参考你之前对 [类似邮件的回复]来起草这封邮件”等措辞的协助提示或许更有效。

然而，缺乏透明度的隐式学习可能会操纵用户信任。向用户展示其显露偏好，具有双重作用：既能为系统决策提供问责依据，也能帮助用户了解自己何时确实需要协助。鉴于用户始终无法准确辨认自己的偏好驱动因素（§5），这种透明度尤其有价值。低舒适度用户在陈述偏好与显露偏好之间表现出最严重的认知偏差，因此可能最能从了解自身实际行为模式中受益。

复杂建模之前的简单行为规则。 行为规则与机器学习之间的等价性对系统部署具有重要意义。对于初始发布，行为规则相比学习模型具有三项优势。第一，它们不需要训练数据，因此能够在使用日志积累之前完成部署。这对于冷启动场景至关重要，因为系统必须立即做出预测。第二，它们具有可解释性和可调试性。利益相关者可以审查“努力程度占55%，发送者/类型各占22.5%，紧迫性接近于零”这一规则，并理解系统为何如此运行，从而促进信任并实现人工改进。第三，它们能够避免在稀疏数据环境下出现过拟合。通过跨用户汇总分析得出的行为权重编码了群体层面的偏好结构，而当每位用户的数据有限时，学习模型的权重可能会过拟合于个体特异性。

然而，随着系统日趋成熟并积累使用数据，机器学习可能会带来优势。当单个用户生成足够的隐式反馈信号时，个性化便成为可能。基于上下文的适配可以捕捉偏好如何随一天中的时段、工作负荷或项目阶段而变化。关键在于循序渐进：先通过行为分析确立正确的基线权重，再逐步叠加机器学习个性化。过早部署机器学习，或者更糟的是，采用与陈述偏好相反的权重，会固化较差的性能，而未来的使用数据也难以纠正。陈述偏好与行为分析和机器学习方法之间均存在 2.8 至 3.6 的差距，这表明错误的初始权重会降低性能。

8 结论

主动式自然语言生成系统需要推断用户何时希望获得帮助，而无需用户明确提出请求，但其设计依赖于未经验证的偏好假设。我们提供了行为证据，表明驱动 AI 起草偏好的因素是组合任务所需的努力程度，而非紧迫性；这使设计优先级从时间信号转向认知负荷。

我们还揭示了感知—行为差距及其可衡量的后果。尽管紧迫性实际影响极小，用户仍在自我报告中将其列为最重要因素；基于陈述偏好的系统，表现也不如忽略用户看法的系统。这一发现对征求显式用户反馈以优化系统的常见做法提出了质疑。

我们的贡献通过以下方式推动以人为中心的自然语言生成：（1）通过析因情境实验方法，为主动式辅助系统提供行为依据；（2）展示用户自我 introspection 的局限，即陈述偏好会误导设计；（3）表明，从显露偏好中推导出的可解释规则，其表现与复杂机器学习模型相当，而陈述偏好设计的表现不如朴素基线。随着人工智能变得更加主动，有效的设计应优先依据行为证据，而非直觉；这对电子邮件助手、文档生成器以及根据上下文推断需求的对话工具都有重要启示。

9 局限性

受控实验环境与情境覆盖范围。我们的析因情境设计有助于实现因果识别，但对现实世界中的电子邮件复杂性进行了抽象处理。我们将四个维度二值化，即紧迫性、努力程度、发件人重要性和电子邮件类型，而没有捕捉连续变化，也未纳入情感效价或政治敏感性等其他因素。机器学习模型约 60% 的准确率上限反映了这些限制。采用更丰富特征（如电子邮件内容、用户历史记录和连续测量指标）的生产系统，可能会取得更高的性能，同时保持行为权重优于陈述偏好权重这一优势。

样本特征。 参与者为 50 名知识工作者，他们经常需要就电子邮件作出决策，适合作为主动起草助手目标人群的代表。我们收集了角色分布（个人贡献者、管理者）、人工智能使用频率、人工智能舒适度和电子邮件数量等数据，以评估研究结果在不同用户群体中的可推广性。不过，若要将研究结果推广到电子邮件实践存在系统性差异的人群（例如，使用大量模板化回答的客户服务岗位、通信由助理代为管理的高管，以及非英语通信场景），仍需进行实证验证。在我们的样本中，感知—行为差距似乎不受使用频率和电子邮件数量的影响，但仍有必要研究跨文化和组织差异。

任务的可推广性。 电子邮件起草需要付出明显的构思努力（写作、编辑、组织论点），因此撰写努力程度自然是一个显著维度。对于不同的自然语言生成任务（例如摘要请求或查询改写），努力程度可能体现为不同形式（理解负荷与生成负荷）。不过，用户会误判决策驱动因素这一普遍原理，很可能不仅适用于电子邮件，也适用于其他主动辅助场景。

偏好稳定性。 我们仅在单一时间点测量了偏好。纵向研究有助于厘清偏好是否会随着人工智能的采用而演变、早期用户与有经验的用户是否表现出不同的因素权重，以及通过使用反馈揭示认知差距是否会改变后续偏好。

参考文献

Piotr D Adamczyk 和 Brian P Bailey。2004年。若非现在，更待何时？在任务执行的不同阶段进行打断的影响。载于*SIGCHI*人机交互系统中的人为因素会议论文集，第271–278页。

Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N Bennett、Kori Inkpen及另外1位作者。2019年。《人机交互指南》。载于*2019*年*CHI*人机交互系统中的人为因素会议论文集，第1–13页。

Victoria Bellotti、Nicolas Ducheneaut、Mark Howard和Ian Smith。2003年。《让电子邮件承担任务：以任务管理为中心的电子邮件工具的设计与评估》。载于*SIGCHI*人机交互系统中的人为因素会议论文集，第345–352页。

Marianne Bertrand和Sendhil Mullainathan。2001年。《人们说的是真心话吗？对主观调查数据的启示》。《美国经济评论》，91(2)：67–72。

Zana Buçinca、Phoebe Lin、Krzysztof Z Gajos和Elena L Glassman。2020年。《在评估可解释人工智能系统时，代理任务和主观测量可能产生误导》。载于第*25*届智能用户界面国际会议论文集，第454–464页。最佳论文奖。

Daniel Buschek、Martin Zürn 和 Malin Eiband。2021年。多条并行短语建议对英语母语和非母语写作者的电子邮件输入及写作行为的影响。载于*2021*年*CHI*人机交互系统中的人为因素会议论文集，第1–13页。

Mia Xu Chen、Benjamin N Lee、Gagan Bansal、Yuan Cao、Shuyuan Zhang、Justin Lu、Jackie Tsay、Yinan Wang、Andrew M Dai、Zhifeng Chen 及另外1位作者。2019年。Gmail智能撰写：实时辅助写作。载于第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第2287–2295页。

Douglas P Crowne 和 David Marlowe。1960年。一种独立于精神病理学的社会赞许性新量表。《咨询心理学杂志》，24(4)：349–354。

Edward Cutrell、Mary Czerwinski 和 Eric Horvitz。2001年。通知、中断与记忆：消息中断对记忆和表现的影响。载于*INTERACT*论文集，第1卷，第263–269页。

Edward B Cutrell、Mary Czerwinski 和 Eric Horvitz。2000年。即时消息中断对计算任务的影响。载于*CHI'00*计算机系统中的人为因素会议扩展摘要，第99–100页。

Mary Czerwinski、Eric Horvitz 和 Susan Wilhite。2004年。关于任务切换和中断的日记研究。载于*SIGCHI*人机交互系统中的人为因素会议论文集，第175–182页。ACM。

Laura A Dabbish 和 Robert E Kraut。2005年。为什么我总是打断自己？环境、习惯与自我中断。载于*SIGCHI*人机交互系统中的人为因素会议论文集，第 3–11 页。

Nicolas Ducheneaut 和 Victoria Bellotti。2001年。电子邮件作为栖息地：对嵌入式个人信息管理的探索。《*interactions*》，8(5)：30–38。

Eric Horvitz。1999年。混合主动式用户界面的原则。*SIGCHI*人机交互系统中的人为因素会议论文集，第 159–166 页。

Shamsi T Iqbal 和 Brian P Bailey。2008年。智能通知管理对用户及其任务的影响。载于*SIGCHI*人机交互系统中的人为因素会议论文集，第 93–102 页。Shamsi T Iqbal 和 Eric Horvitz。2010年。通知与感知：关于警报使用及偏好的实地研究。载于 *2010*年 *ACM* 计算机支持协同工作会议论文集，第 27–30 页。Anjuli Kannan、Karol Kurach、Sujith Ravi、Tobias Kaufmann、Andrew Tomkins、Balint Miklos、Greg Corrado、Laszlo Lukacs、Marina Ganea、Peter Young 等。2016年。Smart Reply：电子邮件自动回复建议。载于第 *22* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 955–964 页。Mina Lee、Percy Liang 和 Qian Yang。2022年。Coauthor：设计人机协作写作数据集，以探索语言模型能力。载于*2022*年 *CHI*人机交互系统中的人为因素会议论文集，第 1–19 页。Wendy E Mackay。1988年。电子邮件使用方式的多样性：初步调查。载于 *ACM* 信息系统汇刊，第 6 卷，第 380–397 页。Daniel C McFarlane。2002年。人机交互中协调人员中断的四种主要方法比较。人机交互，17(1):63–139。Delroy L Paulhus。1991年。反应偏差的测量与控制。《人格与社会心理态度测量》，1:17–59。Forough Poursabzi-Sangdeh、Daniel G Goldstein、Jake M Hofman、Jennifer Wortman Wortman Vaughan 和 Hanna Wallach。2021年。操纵和测量模型可解释性。载于*2021*年 *CHI*人机交互系统中的人为因素会议论文集，第 1–52 页。

Amos Tversky 和 Daniel Kahneman。1973年。可得性：用于判断频率和概率的一种启发式。《认知心理学》，5(2):207–232。

Joshua R Tyler 和 John C Tang。2003年。我什么时候能收到电子邮件回复？一项关于电子邮件使用节奏的研究。载于*ECSCW 2003*：第八届欧洲计算机支持协同工作会议论文集，*2003*年*9*月*14–18*日，赫尔辛基，芬兰，第 239–258 页。Springer。

Steve Whittaker 和 Candace Sidner。1996年。电子邮件过载：探索电子邮件的个人信息管理。*1996* 年 *CHI*人机交互系统中的人为因素会议论文集，第 276–283 页。

Qian Yang、Aaron Steinfeld、Carolyn Rosé 和 John Zimmerman。2020年。重新审视人机交互为何以及如何具有独特的设计难度。载于*2020*年*CHI*人机交互系统中的人为因素会议论文集，第 1–13 页。

A 附录

A.1 电子邮件场景描述与生成提示

A.1.1 场景生成提示

系统提示：

你是一位有帮助的助手，负责生成逼真的工作情境。请创建自然且贴合上下文的情境，巧妙地融入所需维度，而不要明确标注。参与者应能根据上下文线索，而非直接陈述，推断出紧迫性、努力程度、发件人重要性，以及情境属于常规还是新颖。

重要提示：请以叙述形式撰写所有情境，不要使用引号或直接对话。自然地说明每位发送者与用户项目管理角色之间的关系，以体现其重要程度。避免使用“这很紧急”“高投入”“常规任务”或“新颖情况”等明确措辞。应通过上下文传达这些特点。聚焦于真实可信的职场情境，让参与者能够自行判断草稿是否有帮助。

用户提示：

角色扮演档案：大型科技公司的项目经理

你是一家大型科技公司的项目经理，负责推动会影响多个产品团队的跨职能计划。你担任这一角色已有三年，目前负责监督多个处于不同执行阶段的项目。你的工作包括与工程、设计、法务和传播团队协调，确保功能和组织变更按时交付。

你的职责包括制定和维护详细的项目计划、管理风险与依赖关系，以及准备高管简报。你经常为利益相关者撰写沟通材料，例如状态报告、决策摘要和变更公告。审批通常需要与法务和合规团队协作，并且你必须确保所有沟通内容符合公司政策。你还负责主持研讨会、跟踪里程碑进展，并在必要时升级处理问题。

沟通是你角色中至关重要的一部分。对于高管，你需要提供简洁的摘要，突出决策、风险和后续步骤。对于运营团队，你需要分享结构清晰的最新进展，明确时间安排和责任归属。你主要通过电子邮件和 Teams 进行沟通，同时，你需要在兼顾长期交付任务的同时，及时响应紧急请求。

任务：针对以下维度的全部 16 种独特组合生成情境：紧迫性（紧急 / 不紧急）、撰写所需努力（高投入 / 低投入）、发件人重要性（重要发件人 / 次要发件人）、请求类型（常规 / 新颖）。

指南：(1) 通过“今天下班前”“尽快”“明天的会议需要”等上下文，自然地体现紧迫性；与之相对的表达包括“有空时”“不急”“本周找时间”。(2) 通过任务复杂度展现努力程度，而不是直接说明。复杂度较低的任务包括快速确认和简短的状态更新；复杂度较高的任务则涉及跨团队协调和全面分析。(3) 始终明确说明发送者与你的角色之间的关系，以此体现发件人的重要性。直接影响你的发送者包括你的经理、负责你所参与项目的高管以及关键利益相关者；间接影响你的发送者包括其他团队的同级经理和行政人员。(4) 通过“每周报告”与“意外情况”或“新计划”等上下文，表明请求是常见还是不同寻常。(5) 自然行文，不要明确标注各个维度。(6) 让情境真实、具体，并包含科技公司中可能出现的实际背景。(7) 确保每个情境都需要作出有意义的回应。(8) 以叙述形式撰写情境，不要加引号。

生成全部 16 个情境后，作者分别审阅了每个情境，以确认其准确反映了预期的维度水平（紧迫性、努力程度、发件人重要性和电子邮件类型）。作者一致认为，生成的情境符合析因设计的要求。

A.1.2 完整情境列表

1. **紧急 +高投入 +重要发件人+常规**：你的经理联系你，要求提供一份最新的季度末项目健康报告，汇总每个活跃工作流的指标。他们需要在今天下班前完成报告，以便今晚晚些时候进行领导层评审，而模板与你的团队每个季度使用的相同。

2. **紧急 +高投入 + 重要发件人 +新颖**：工程副总裁负责领导打造核心产品的团队，他通知

你，执行董事会要求立即分析昨天发布的监管公告将如何影响你当前的发布计划。他们要求在明天早上之前提供详细的情景评估和缓解方案。

3. **紧急 + 高投入 + 次要发件人+ 常规**：今天早上，人力资源技术团队的一位同事项目经理请你为共享资源制作一份全面的依赖关系摘要。他解释说，这份摘要关系到他们定于今天晚些时候启动的集成测试，目前已成为阻碍。

4. **紧急 + 高投入 + 次要发件人+ 新颖**：设施运营协调员联系你，称一次重大系统中断影响了项目团队使用的部分协作工具。他们请你协助制定一份紧急应急计划，其中包含替代工作流程，以便在下一次工作会议前分享给受影响的团队。

5. **紧急 + 低投入 + 重要发件人 +常规**：你的主管发消息请你提供最新的时间线幻灯片，以用于一小时后开始的高管电话会议。他提到，较早的版本也可以，只需插入更新后的完成百分比。

6. **紧急 + 低投入 +重要发件人+ 新颖**：一位高级产品负责人发消息给你，想快速确认你的团队能否安排明天临时举行的客户反馈会议。他们现在就需要答复，以便在正在进行的电话会议中向客户作出承诺。

7. **紧急 +低投入 +次要发件人+ 常规**：其他部门的一位分析师发来电子邮件，请你重新发送昨天的利益相关者演示文稿，因为他们无法访问共享文件夹中的文件。他们现在就需要这份文件，以便参加几分钟后开始的电话会议。

8. **紧急 +低投入 +次要发件人+ 新颖**：一栋大楼发生整栋楼范围的电力波动时，一位设施管理员联系你，想了解你的项目中是否有任何硬件依赖项需要立即告知工程团队。他们希望你尽快给出是或否的答复，以便完成停机警报消息。

9.**不紧急 + 高投入 + 重要发件人 +常规**：你的主管让你准备完整的季度项目组合摘要，供他们下周即将进行的战略评审使用。这与过去几个季度你编写的深入报告相同，内容包括关键绩效指标、风险日志和里程碑图表。

10.**不紧急 + 高投入 + 重要发件人 + 新颖**：首席法务官请你起草一份全面的过渡计划，说明如何将即将实施的合规规则整合到所有产品线中。他们强调，这项工作对组织而言是全新的，截止日期是下个月底，因此你有时间提出一个结构化方案。

11. **不紧急 + 高投入 + 次要发件人 + 常规**：内部工具项目的一位同行项目经理询问，你是否可以整理一份你们双方项目共享资源使用情况的汇总视图。他们表示这不会妨碍当前任何工作的进展，并建议在几周内准备好即可。

12. **不紧急 + 高投入 + 次要发件人 +新颖**：员工敬业度计划的通信专员联系你，请你协助设计一场跨团队研讨会，集思广益，探讨如何在公司内部展示成功的产品发布。他们提到这是一项新举措，并希望能在本月晚些时候与你交流。

13.**不紧急 + 低投入 + 重要发件人 + 常规**：你的经理发来电子邮件，询问你能否确认上个冲刺周期的里程碑进展是否已反映在仪表板上。他们说不着急，本周任何时候告知他们都可以。

14.**不紧急 + 低投入 + 重要发件人 + 新颖**：一位资深工程总监询问你能否快速写一段文字，描述项目长期架构方面的考量，以供他们起草愿景文档。他们提到下周提供也没问题。

15. **不紧急 + 低投入 + 次要发件人 + 常规**：一位行政协调员发消息询问你所在项目团队的月度员工人数。他们

说明他们正在编制下个月的座位安排，你可以在本周方便的时候回复他们。

16. **不紧急 + 低投入 + 次要发件人 +新颖**：一位参与内部文化委员会工作的同事表示，他们正在为新的公司博客收集各项目的简短成功案例。他们请你在接下来的几天里方便时，简要总结一下你的项目近期取得的一项里程碑成果。

A.2 主题分析方法与次要发现

我们开展了探索性主题分析，以补充定量偏好建模。这项分析针对两个研究问题考察了文本回复。第一个研究问题探讨了哪些电子邮件属性会影响用户对人工智能写作辅助的偏好，并分析了在成对比较过程中提供的 750 条即时理由。第二个研究问题考察用户是否准确理解了影响其偏好的因素，并分析了 50 份任务后问卷反思，其中参与者回顾性地描述了自己的决策过程。

A.2.1 LLM 辅助的五阶段编码过程

我们采用了一种系统性的五阶段方法，该方法改编自传统定性方法，并通过 LLM 辅助加以实施，从而在扩大覆盖范围的同时保持分析严谨性。

第一阶段：初始编码。 大语言模型使用归纳式编码分析了一个具有代表性的样本，提示采用开放式设计，旨在让主题自然浮现。对于第一个关于偏好驱动因素的研究问题，我们从 750 条理由中抽取了 150 条进行分析，以兼顾覆盖面与可管理性，这些理由占所有成对比较的 20%。对于第二个关于元认知准确性的研究问题，鉴于数据集规模较小，我们分析了全部 50 条调查回复。提示明确要求模型让主题从回答本身自然浮现，而不是将回答强行归入预先定义的类别。如果模型发现了不符合常见模式的主题，则应将其作为独立编码纳入。温度设为 0.3，以确保编码一致，同时允许发现主题。

第二阶段：编码手册开发。 我们将初始主题整合为结构化的编码手册，其中包含层级组织、清晰的定义、纳入标准、排除标准，以及每个编码的示例引文。本阶段将温度设为 0.3，以兼顾系统性与综合归纳。

阶段 3：系统化编码。使用极低的温度（0.1）将编码手册应用于所有回答，以确保最大程度的一致性。针对第一个研究问题的回答按每批 50 条进行处理，以适应更丰富的成对比较上下文；每条回答都包含两个电子邮件情境、偏好选择和理由文本。鉴于第二个研究问题的数据结构较为简单，仅包含开放式文本，因此将其全部 50 条回答合并为一批进行编码。

阶段 4：解析与结构化。对编码结果进行解析，并与原始回答合并，再分析主题频率、共现模式和分布统计。

阶段 5：综合研究发现。作者利用大语言模型生成的主题分布摘要、示例引文和编码手册模式，综合研究发现。特别是针对第二个研究问题，提示词强调分析认知与行为之间的差距，询问用户在描述自身决策时会过度报告、报告不足或错误归因哪些因素。大语言模型以 0.4 的温度生成条理清晰的摘要，在保持数据依据的同时适度发挥分析创造力；随后，作者对这些摘要进行解读，以确定关键模式及其启示。

所有阶段均使用固定的随机种子 42，以确保结果可复现。作者在每个阶段都审查并确认了编码过程及其输出，以确保结果和洞见有效且与数据一致。附录中提供了完整的编码手册（含层级主题结构）、带有主题归类的编码数据集，以及详细的频率分布（见下文附录 A.2.2）。

A.2.2 完整的主题研究发现

RQ1：两两比较过程中即时给出的理由

我们分析了参与者在成对比较任务中做出偏好选择后立即提供的750条理由说明。我们确定了12个不同主题，并按照第二阶段制定的层级编码

手册结构将其归为四个主要类别。编码的理由总数：750。主题归类总数：769。每条理由说明的平均主题数：1.03。中位数：1。注：下文百分比表示提及各主题的回答所占比例。由于部分回答同时提及多个主题，因此总和为102.5%（而非100%）。

具有层级结构的完整主题清单

主要主题 1：时间与效率驱动因素（合计 31.5%）

- **紧迫性与时间压力**（20.3%）提到紧迫的截止期限或紧急周转时间是使用人工智能的原因

- **希望节省时间**（11.2%）强调效率或减少工作量，而非明确的紧迫性

主要主题 2：任务特征（合计 40.7%）

- **起草复杂度**（14.7%）表示任务复杂、包含多个步骤或需要综合分析

- **任务简单性——不需要人工智能**(12.8%)认为任务过于简单，人工智能无法带来额外价值

- **模板或历史背景**(10.1%)指出存在人工智能可以借鉴的现有格式、模板或先前示例

- **机械重复或繁琐的工作**(3.1%)指出任务重复、价值低或枯燥乏味

主题 3：风险与信任考量（合计 6.1%）

- **避开高风险或敏感任务**(4.8%)在处理敏感、高风险利害或受监管内容时避免使用人工智能

- **易于核验**(1.3%)表示能够轻松检查正确性，从而降低风险

主要主题4：对人工智能角色和能力的认知（合计24.3%）

- **感知到的人工智能能力**（14.3%）认为，人工智能能够胜任任务，是因为要求明确或任务具有结构化特征

- **与人工智能优势相契合**（4.5%）认为任务与人工智能已知的优势相匹配，例如摘要或结构化

- **头脑风暴或创意协助**（3.9%）认为人工智能有助于生成想法或提供创意输入

- **个人写作偏好**（1.6%）因个人对写作任务感到不自在或不喜欢，而选择使用人工智能

共现模式揭示了情境信任。紧迫性与时间压力常与起草复杂度同时出现，这表明用户认为高风险情境兼具时间压力和认知负荷。值得注意的是，感知到的人工智能能力常与模板或历史背景同时出现，这说明用户会根据任务可预测性来调整对人工智能的信任，而非对人工智能能力持有一概而论的看法。当存在先前示例时，用户会明确说明使用人工智能的理由，例如“人工智能似乎能准确地写出这段内容”和“格式已经有了，因此人工验证会更快”。这种条件性信任模式表明，个性化人工智能推荐应突出模板可用性，将其作为信任信号。

将验证作为风险缓解策略。易于验证与避免高风险任务同时出现，体现出一种成熟的心智模型：用户并非一味避免在高风险利害的任务中使用人工智能，而是会评估验证是否可行。示例引语包括“我很容易核验正确性”和“人工智能可以在这里快速把内容整理成正确格式，而且我很容易核验”。这表明，如果验证工作流程得到简化，风险厌恶型用户也可能采用人工智能；这也将设计挑战从提高人工智能的准确性转变为让验证毫不费力。

模板与历史背景的双重作用。模板或历史背景（10.1%）在提供理由时发挥着两种不同作用：降低感知努力（例如，“使用我以前用过的同一份报告，可以节省时间”），以及通过先前示例和可预测格式增强对AI的信任。这种双重作用使模板成为推动人工智能采用的有力触发因素，因为它们既能降低感知风险，又能增加感知收益。

个人写作不适感是一个小众因素。个人写作偏好（1.6%）呈现为一个独特但不常见的主题，用户提到了诸如“我的写作风格不太正式”以及语气调整等顾虑。虽然这种情况很少见，但这表明有一类用户使用人工智能并非为了提高效率

而是为了获得风格支持。这一小众用例不同于占主导地位的减少工作量模式，因此可能需要采用不同的功能优先级排序（语气建议，而非快速草稿）。

次要设计启示。除论文正文中讨论的主要启示外，这些次要发现表明：（1）应在用户界面中突出基于模板的触发条件，将其作为信任和效率信号。（2）展示置信度分数并突出显示 AI 生成内容的验证仪表板，可能有助于提升风险规避型用户的采用率。（3）认知负荷检测（打字停顿、修改模式）可能比自我报告的紧迫性更具预测性。（4）面向小众用户的细分功能（语气调整、头脑风暴模式）可与核心效率功能相辅相成。

研究问题2：关于决策的任务后问卷反思

我们分析了针对一道开放式问题收集的 50 份调查回复。该问题询问参与者在判断一封电子邮件是否值得起草时考虑了哪些因素。我们识别出 13 个不同主题（不包括回答“无”的情况）。编码的回答总数：50。每份回答的平均主题数：2.82。中位数：3。注：以下百分比表示提及各主题的回答所占比例。由于每份回答通常会提及多个主题，因此百分比总和超过 100%（达到 282%）。

完整主题清单

- **紧迫性／时间紧迫性**（66%）提到了截止期限、时间紧迫性或快速周转的需求
- **努力与效率**（32%）考虑了手动起草与使用人工智能所需的努力程度，以及节省时间的效果
- **数据可用性与输入内容**（28%）评估了所需信息或数据是否易于获取
- **内容复杂度**（26%）评估了主题内容的复杂程度或技术性
- **受众／利益相关者的重要性**（24%）考虑了电子邮件的收件对象，包括其资历和角色
- **任务类型或性质**（22%）区分请求的内在性质，例如信息型还是行动导向型
- **与人工智能的适配度／人工智能能力**（20%）评估任务是否符合人工智能的优势，或人工智能能否处理其复杂性
- **电子邮件的重要性／影响**（20%）评估风险利害、可见度或潜在后果
- **电子邮件的长度与细节**（18%）考虑预期篇幅和所需内容量
- **准确性与质量要求**（12%）权衡对正确性、可靠性和专业语气的需求
- **个性化或细腻表达的需求**（8%）评估是否需要体现个人特色或采用带有情感色彩的语气
- **过往经验／熟悉程度**（6%）参考处理类似电子邮件或人工智能表现的过往经验

回顾性阐述作为合理化机制。每份回答的平均主题数从即时理由中的 1.03 个增至回顾性陈述中的 2.82 个。这种模式可能反映的是重构性记忆，而不只是更为周详：当用户被要求反思时，他们会给出更详尽的多因素解释，这与虚构记忆文献的发现一致。尽管用户的实际行为各不相同，但最常见主题在用户间保持一致（紧迫性 66%、努力程度 32%），这表明他们反映的可能是关于“良好决策”的共同文化脚本，而非真正的内省。

数据可用性是门槛因素。数据可用性和输入内容在回顾性陈述中尤为突出（28%），但在原先的四维实验设计中并未出现。这个主题表明，用户意识到信息完整性是先决条件，例如有人提到“我需要追着多少人才能拿到数据”，以及“需要汇总的数据量”。这表明存在一个两阶段决策过程：先评估可行性（我有数据吗？），再评估期望度（人工智能的帮助值得吗？）。生产系统应主动呈现数据完整性信号。

对人工智能偏好的元认知学习有限。尽管参与者完成了 15 次涉及人工智能起草场景的两两比较，但他们在反思中很少提及以往经验或熟悉程度（6%）。用户很少提到从任务本身中学习

，或更新对人工智能能力的看法。这表明，关于人工智能偏好的元认知学习进展缓慢，系统可能需要通过明确的反馈循环向用户展示他们显露出的偏好模式，以加快这种校准。

对职业身份的质量顾虑。 尽管准确率和质量要求（12%）以及对个性化或细微差别的需求（8%）并非实验操纵的维度，但它们仍出现在回顾性反思中。这些主题反映了对职业身份的维护，用户强调了“准确性的可靠性”“质量标准”和“需要个人风格”。这些因素出现在回顾性陈述中，却并未成为行为预测因素，这表明它们是事后为职业精神提供的理由，而非实际的决策驱动因素。

将任务类型分类体系视为心智模型。 任务类型或性质（22%）表明，用户会从实验变量以外的多个维度对电子邮件进行心理分类：“信息型与行动导向型”“重新包装现有内容”“请求是要我作出文字回复，还是直接采取行动”。这种更丰富的分类体系表明，四因素实验模型捕捉了主要驱动因素，但用户还能感知到其他细微差别，而这些差别可能对边缘情况或特定工作流程很重要。

次要设计启示。 这些次要发现表明：（1）起草前应显示数据完整性指标（“所需信息已齐全”或“缺少：第三季度预算数据”）。（2）使用模式仪表板有助于加快元认知校准（“您向人工智能提出的请求：78%为高努力程度的电子邮件，12%为紧急电子邮件”）。（3）质量保证功能（语气匹配、事实核查）能够回应事后对职业精神的顾虑，即便它们并非行为驱动因素。（4）任务类型检测可将电子邮件导向具备专门人工智能能力的处理流程（信息摘要与行动方案生成）。

A.3 机器学习模型实现细节

本节提供了第 §6节中所使用机器学习模型的完整实现细节。

A.3.1 数据准备

特征构造。 每个成对比较（共 750 个）都被转换为一个二元分类实例。对于情境 i 和 j 之间的每次比较，我们根据因素水平的差异构造特征：Δ紧迫性、Δ努力程度、Δ发送者、Δ类型 $\in$ {-1, 0, +1}。我们还纳入了全部六个双向交互项：Δ紧迫性 $\times\Delta$努力程度、Δ紧迫性 $\times$ Δ发送者、Δ紧迫性 $\times\Delta$类型、Δ努力程度 $\times$ Δ发送者、Δ努力程度 $\times$ Δ类型，以及 Δ发送者 $\times$ Δ类型。目标变量为二元变量：如果偏好情境 i，则为 1；如果偏好情境j，则为 0。每个实例因此包含 10 个特征。

预处理。 所有特征均使用 StandardScaler（均值为零、单位方差）进行标准化，以确保不同模型之间具有可比的尺度。鉴于现有特征集已经较为紧凑，未进行降维。

训练测试集划分。 我们采用分层五折交叉验证，以确保各折之间的类别分布均衡，并获得稳健的性能估计。随机种子设置为 42，以确保结果可复现。

A.3.2 模型架构与超参数

逻辑回归。 标准 L2 正则化逻辑回归，正则化强度 C =1。0。求解器：L-BFGS。最大迭代次数：1000。类别权重：平衡设置，以应对轻微的类别不平衡（55.2% 的用户偏好使用辅助功能）。

随机森林。 由决策树组成的集成模型，采用网格搜索调整以下超参数：估计器数量 = 100、最大深度= 3、最小分裂样本数 = 20、最小叶节点样本数 =10。鉴于数据集规模适中（750 个实例），选择这些浅树约束以防止过拟合。特征重要性通过对所有树的基尼重要性取平均值来提取。

梯度提升。 梯度提升决策树，参数如下：学习率 = 0.1、估计器数量 = 100、最大深度 = 3、最小分裂样本数 = 20、最小叶节点样本数 = 10、子采样比例 = 0.8。为防止过拟合，采用了与随机森林类似的深度约束。

神经网络。 用于防止过拟合的浅层前馈架构：

- 输入层：10 个特征（4 个主效应 + 6《interactions》）
- 隐藏层 1：16 个单元，ReLU 激活函数
- 隐藏层 2：8 个单元，ReLU 激活函数
- 输出层：1 个单元，Sigmoid 激活函数

训练配置：

- 优化器：Adam，学习率为 = 0.001
- 损失函数：二元交叉熵
- 批次大小：32
- 训练轮数：50，采用早停（耐心值 =10，监控验证损失）

- 丢弃法：每个隐藏层之后设置为 0.3，以进行正则化
- L2正则化：权重衰减 = 0.01

A.3.3 浅层架构的依据

仅有 750 个训练样本和 10 个特征，因此我们优先考虑泛化能力，而非模型复杂度。最初使用深层网络（3–4 个隐藏层，每层含 32–64 个单元）进行的实验显示出严重过拟合：训练准确率达到 75–80%，而验证准确率则停滞在 58–60%。浅层架构（2 个隐藏层，分别含 16 和 8 个单元）实现了更好的偏差-方差权衡，验证准确率达到 61–62%，且训练集与验证集之间的准确率差距很小（3–4 个百分点）。同样，最大深度为 > 5 的基于树的模型也出现了过拟合，这证明深度 = 3限制是合理的。

这一设计选择与论文的核心发现一致：简单的行为规则（准确率为61.3%）与复杂的机器学习模型（平均准确率为60.8%）表现相当，这表明预测上限反映的是数据限制，而非模型容量不足。

A.3.4 性能评估

	全部特征（10）				**主效应（4）**			
模型	Acc.	精确率	Rec.	F1	Acc.	精确率	Rec.	F1
逻辑回归	0.587	0.627	0.691	0.657	0.580	0.624	0.664	0.643
随机森林	0.608	0.607	0.882	0.719	0.605	0.614	0.855	0.714
梯度提升	0.588	0.617	0.833	0.709	0.600	0.615	0.848	0.713
神经网络	0.599	0.619	0.785	0.692	0.603	0.608	0.819	0.698
Mean	**0.595**	**0.618**	**0.798**	**0.694**	**0.597**	**0.615**	**0.797**	**0.692**

表4：完整的机器学习模型性能指标。正文（表1）为简洁起见，仅展示准确率和 F1 值。

A.4 研究界面截图和退出调查

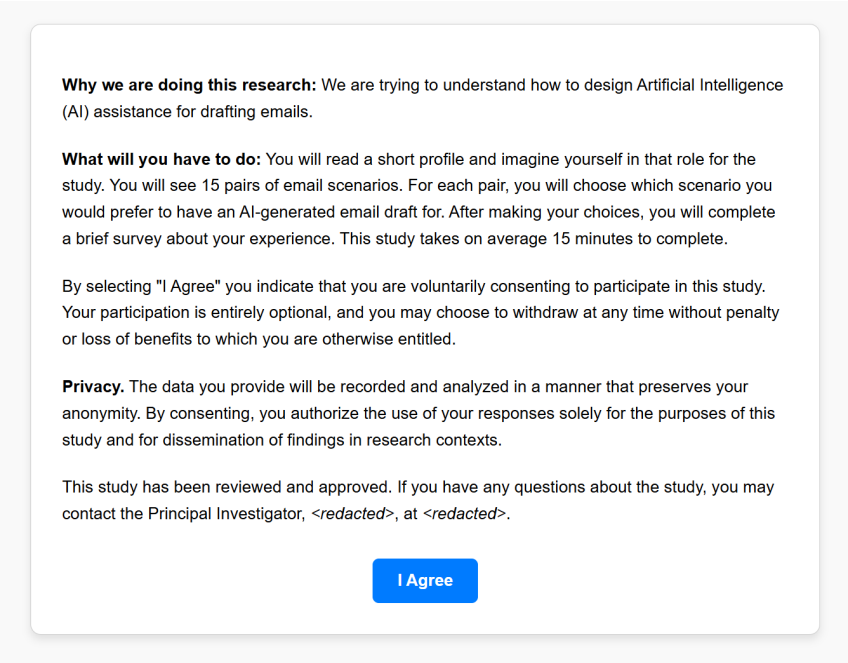

图5：研究开始时展示的知情同意页面。参与者在同意参加研究前，阅读了研究目的、程序、风险、益处及数据处理方式。

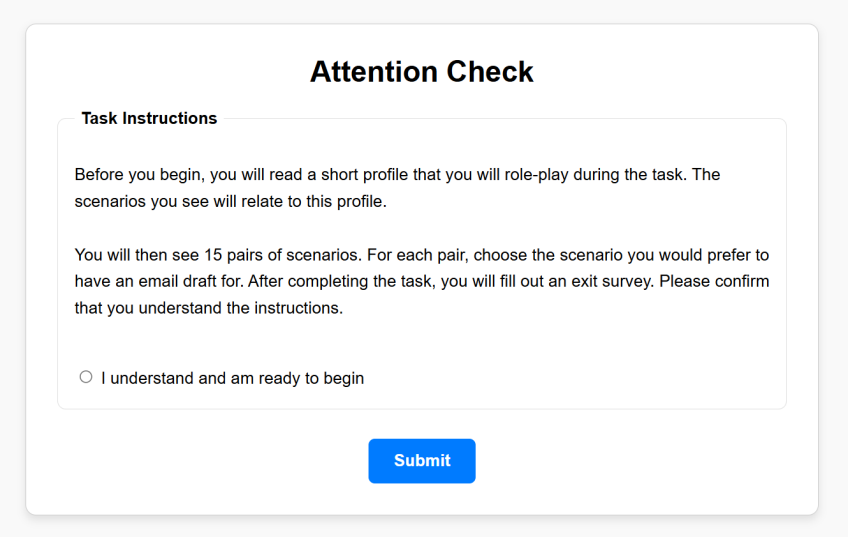

图6：任务说明界面，说明两两比较任务。参与者获知，他们将看到成对的电子邮件情境，并选择更希望获得人工智能写作辅助的电子邮件，然后说明选择理由。

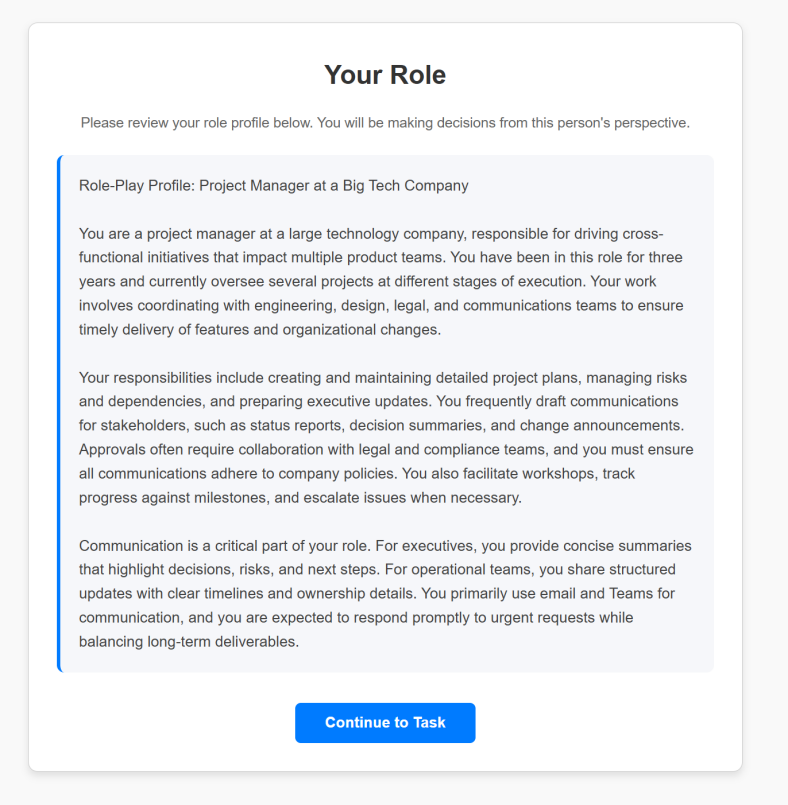

图 7：角色扮演者资料分配。参与者被要求扮演一家大型科技公司的项目经理，以便在所有情境中基于一致的职业情境作出判断。

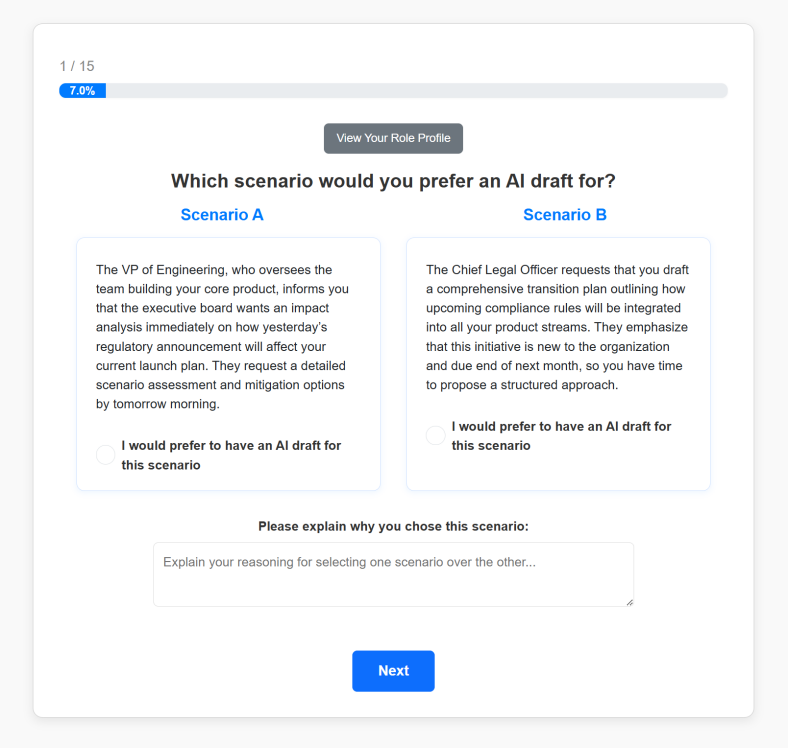

图 8：成对比较界面展示了主要任务。参与者并排查看两个电子邮件情境，选择自己更希望获得人工智能写作辅助的电子邮件，并为自己的选择提供自由文本理由。每位参与者完成了 15 次此类比较。

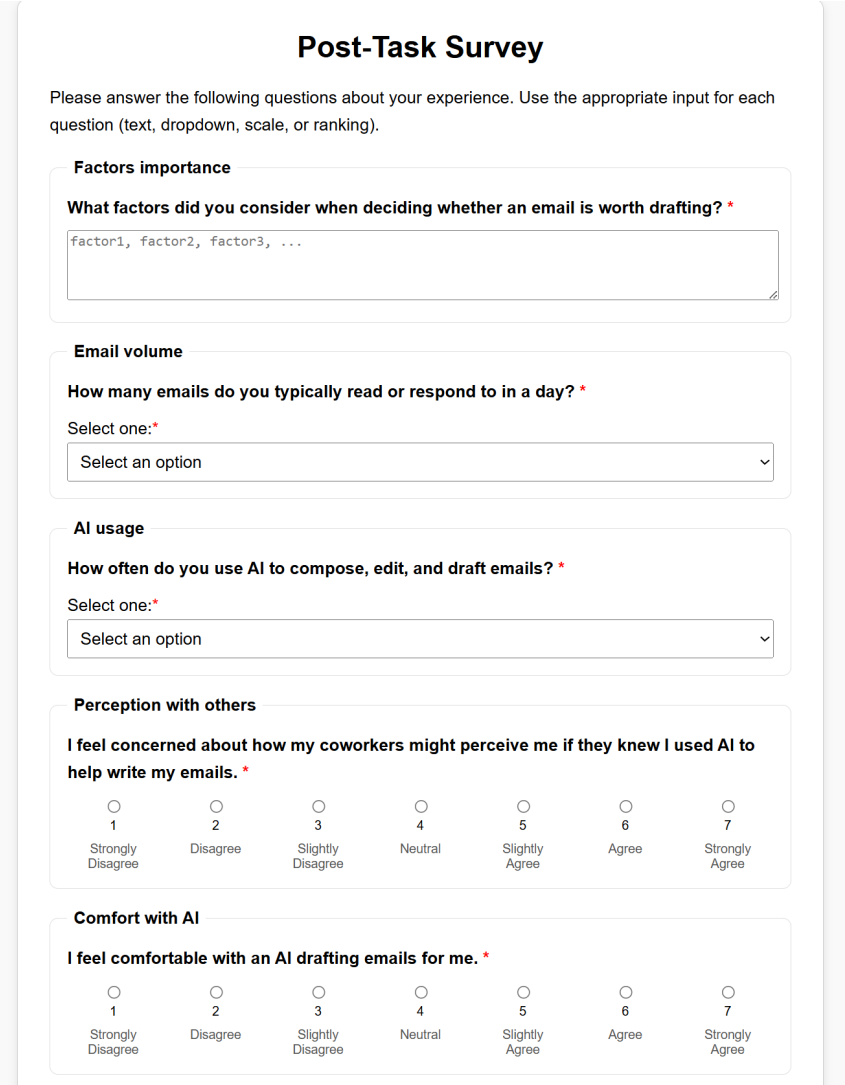

图 9：退出调查（截图第 1 部分），收集参与者关于决策因素、电子邮件数量、人工智能使用频率和人工智能态度的开放式反思。开放式问题请参与者描述，他们在决定一封电子邮件是否值得起草时考虑了哪些因素。

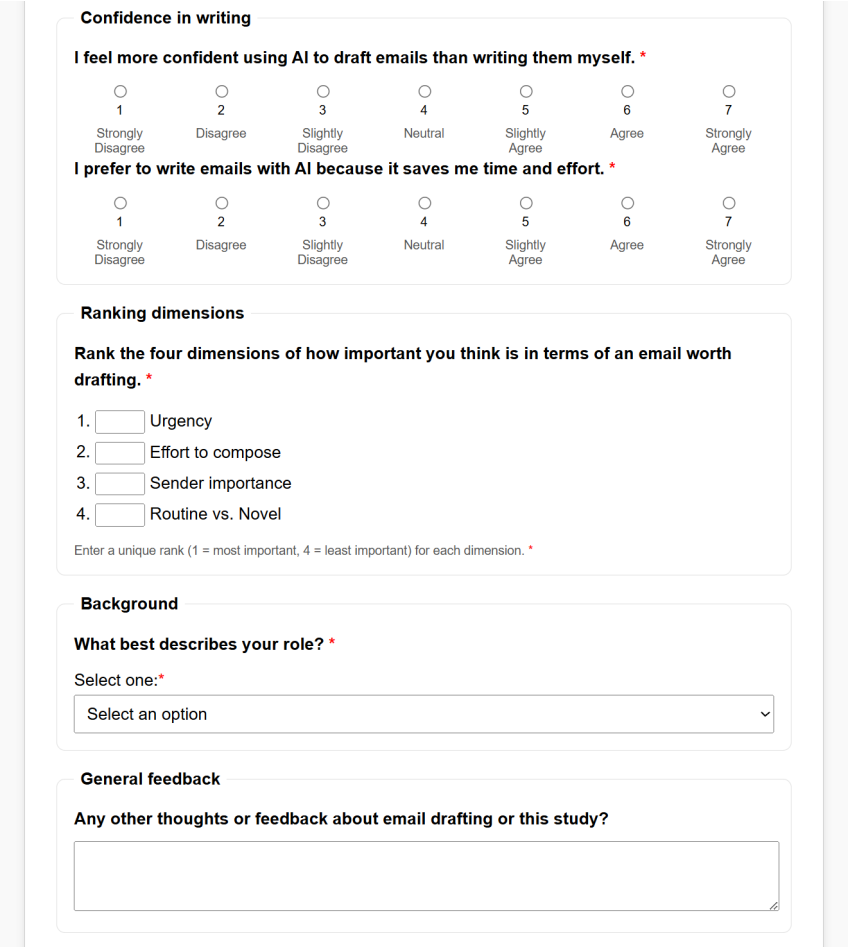

图10：退出调查（截图第2部分），收集参与者在决策时对四个维度（紧迫性、撰写所需努力、发件人重要性、电子邮件类型）从最重要到最不重要的强制排名，以及当前角色和可选反馈。

A.5 偏好模式的个体差异

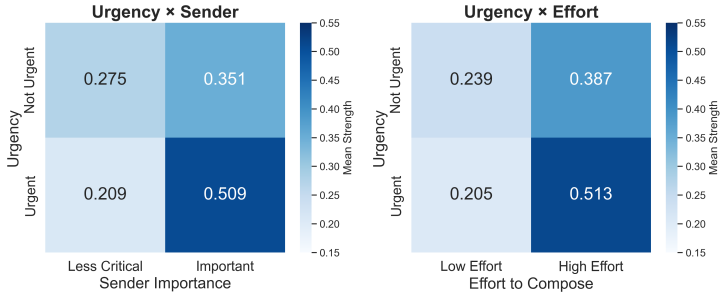

图 11：其余成对维度交互作用所显示的布拉德利-特里偏好强度均值。紧迫性× 发送者和紧迫性 × 努力程度呈现相对平缓的模式，这证实紧迫性的主效应较弱（$\rho \approx 0$），并非通过与其他维度的强交互作用显现。

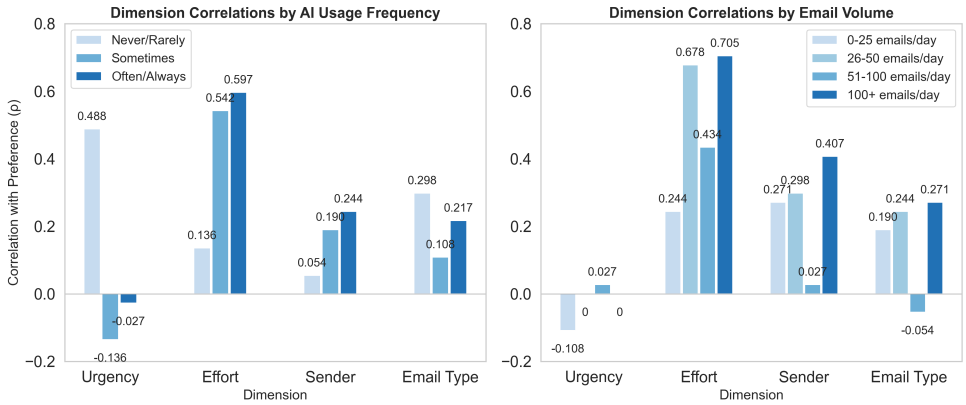

图 12：不同用户群体的偏好模式个体差异。左图显示按人工智能使用频率划分的维度相关性；右图显示按电子邮件数量划分的相关性。努力程度在所有群体中始终占主导地位（蓝色柱形仍然较高），而电子邮件类型（偏好常规邮件还是新颖邮件）则存在明显差异。有经验的人工智能用户的常规邮件偏好（$\rho = 0.217$）弱于不常使用的用户（$\rho = 0.298$）；高邮件量用户（100+ 封电子邮件/天）的常规邮件偏好（$\rho = 0.271$）则强于中等邮件量用户。这表明，核心偏好驱动因素（努力程度）普遍适用，而次要因素（对电子邮件类型的熟悉程度）会随用户经验而变化。

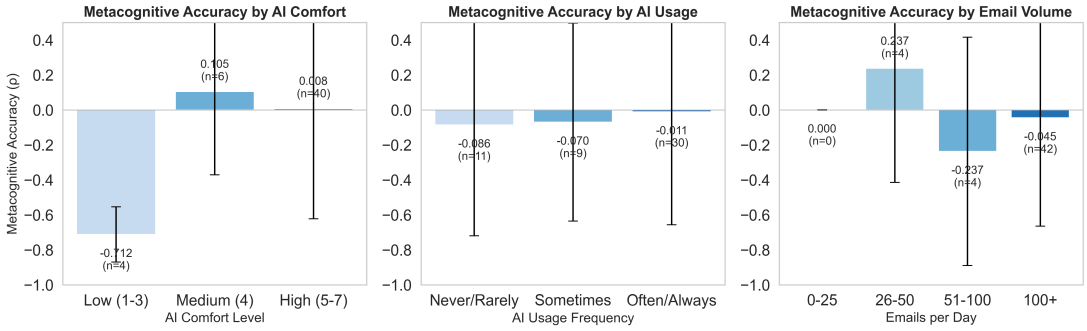

图 13：不同用户特征下元认知准确性的个体差异（陈述偏好与显露偏好之间的相关性）。左图显示，人工智能舒适度是关键预测因素：高舒适度用户表现出正向一致性（$\rho = 0.113$），而低舒适度用户则呈现反向认知（$\rho = -0.550$）。中图和右图显示，人工智能使用频率和电子邮件数量并不能显著预测元认知准确性，这表明心理准备度比行为经验更为重要。误差线表示各组用户的标准差。