/TemplateVersion (2026.1)

PREF: LLM 中个性化文本生成的无参考评估

Xiao Fu[1], Hossein A. Rahmani[1], Bin Wu[1], Jerome Ramos[1], Emine Yilmaz[1], Aldo Lipani[1]

[1]University College London (UCL)

xiao.fu.20@ucl.ac.uk, hossein.rahmani.22@ucl.ac.uk, bin.wu.23@ucl.ac.uk, jerome.ramos.20@ucl.ac.uk, emine.yilmaz@ucl.ac.uk, aldo.lipani@ucl.ac.uk

Abstract

个性化文本生成对于用户中心的信息系统至关重要，但大多数评价方法忽视了用户的个体差异。我们介绍了PREF，一种无需参考个性化评价框架（P ersonalised R eference-free E valuation F ramework），该框架能够联合评估一般输出质量和用户特定的匹配度，而无需金标准的个性化参考。PREF 在三步管道中运行：（1）覆盖阶段使用大型语言模型（LLM）生成涵盖普遍标准（如事实性、一致性和完整性）的详尽、特定查询的指南；（2）偏好阶段通过重新排序并根据目标用户的档案、陈述或推测的偏好和背景选择性地增强这些因素，生成个性化的评估量表；（3）评分阶段应用 LLM 评委根据这一量表对候选答案进行评分，确保基线的充分性，同时捕捉主观优先级。这种覆盖与偏好分离的方法提高了鲁棒性、透明性和可重用性，并允许较小的模型接近较大模型的个性化质量。在 PrefEval 基准上的实验，包括隐含偏好跟随任务，显示 PREF 比强基线获得更高的准确性、更好的校准性和更接近人类判断的匹配度。通过实现可扩展的、可解释的和与用户对齐的评估，PREF 为个性化语言生成系统的更可靠评估和开发奠定了基础。

介绍

大型语言模型（LLM），如 GPT-3（Brown et al. 2020）和 ChatGPT（Achiam et al. 2023），将开放式文本生成推向了新的高度，实现了大规模高质量对话、代码合成和数据到文本的叙述。尽管取得了这些成功，但评估这些模型的输出仍然是一个未解决的问题。传统的基于参考的指标，包括 BLEU 和 ROUGE（Papineni et al. 2002; Lin 2004），通过计算与一个或多个参考文本的 n 元重叠来评估，但在那些允许多种同等有效答案的任务上（例如，创意写作、推荐和建议），其与人类判断的关联性较弱。基于嵌入的替代方法，如 BERTScore（Zhang et al. 2020），通过在潜在空间中测量语义相似度来减轻表面不匹配的影响，但它们仍需要高质量的参考文本，并且最终反映的是通用而非用户特定的需求。

研究中的另一条平行路径完全绕过参考，通过将一个强大的大型语言模型（LLM）视为自动裁判来进行。像 MT-Bench（Zheng et al. 2023）和 AlpacaEval（Li et al. 2023）这样的基准测试要求 GPT-4 评估或排序候选答案，并报告出色地与众包工作人员取得一致。虽然这种 LLM 作为裁判的范式可以以低成本和可重现的方式扩展，但其评分标准本质上是普遍的——“一个好的答案对每个人来说都是好的”——因此，它对个人的偏好、限制或可能极大地改变用户对质量感知的先前互动视而不见。现代应用越来越多地直接将 LLM 暴露给最终用户，他们期望输出符合他们的品味、目标和语境。最近的工作表明，LLM 确实可以适应用户档案、反馈信号和风格规范（Zhang et al. 2024）。然而，评估这种个性化生成尤其棘手，因为质量现在是用户相关的，即，一个让一个人很满意的答案可能会让另一个人感到沮丧。通用的参考或以模型为中心的评分标准无法捕捉这种主观维度。

人工评估仍然是黄金标准，但其过程昂贵、耗时，并且容易受到注释者的差异和偏见的影响，即使提供了详细的指导原则（Van Der Lee et al. 2019）。因此，为每个系统更新或用户群体收集用户对齐的评级是不切实际的，这阻碍了个性化体验的快速迭代。

为了填补这一评估空白，我们提出了 PREF，一个个性化、无需参考的评估框架，用于对生成的文本进行评分，而不需要金标准的个性化答案。PREF 分两个阶段进行评估：

1. 一般质量阶段。一个通用指南列举了需要检查的显著因素——真实性、一致性、完整性等——明确忽略用户偏好，以确保基线的充分性。
2. 用户对齐阶段。从用户信息（个人资料属性、过去的对话、陈述的偏好）中合成个性化的指导原则，用于对一般因素进行加权或重新排序，从而生成反映用户关心内容的定制评分标准。

然后，LLM 根据综合评分标准对候选答案进行评估，生成一个标量分数，而不需要参考真实答案。通过这种方式，PREF 结合了普适的质量控制与用户中心对齐，同时保持可扩展性和可重复性。

我们的贡献可以总结为：

- 我们制定了一种无参考的评估协议，用于个性化生成，同时考虑任务适切性和用户一致性。
- 我们通过自动规则构建和基于 LLM 的评分来具体化这一协议，消除了对标准参考或反复人工评分的需求。
- 通过对 PrefEval 基准的广泛实验，我们证明了 PREF 的评分比现有的基准更忠实地追踪人类对个性化质量的判断。
- 我们展示了在开发过程中整合 PREF 有助于较小的模型（例如，LLaMA-3 8B）缩小与较大模型之间的性能差距，从而促进了经济高效的部署。

综上所述，我们的研究结果将 PREF 定位为评估——并最终改进——用户对齐语言生成系统的一个实用且有效的基础。

相关工作

个性化的基础

个性化的理念来自两个相关的研究领域：

显式用户建模。 自适应超媒体系统表明，维持一个细粒度的学习者或读者模型能够实现内容、排序和导航的实时适应（Brusilovsky 2001）。这些系统依赖于明确编码的属性——背景知识、目标、学习风格——来决定接下来显示什么。

隐式偏好推断。 协同过滤技术，如 GroupLens，表明可以通过行为信号推断出偏好：大规模聚合的用户-物品评分使得“过去意见一致的人能够再次达成一致”（Resnick et al. 1994）。

现代 LLM 个性化策略仍然可以追溯到这一二分法：显式的档案条件设置与隐式信号挖掘（例如，点击、停留时间、编辑）。

个性化文本生成与大语言模型

早期神经人格建模。 Persona-Chat 框架将基于向量的说话人嵌入添加到序列到序列对话模型中，提高了角色一致性，并为角色感知生成打开了大门（Li et al. 2016）。

属性和风格控制。 Plug-and-Play 语言模型（PPLM）开创了冻结的 GPT-2 向用户指定属性进行梯度引导的先河，而无需昂贵的整体微调（Dathathri et al. 2020）。后续的参数高效方法如前缀调优、适配器和 LoRA 附加小型可训练模块，仅需几千个参数即可实现每用户或每任务的专业化（Li and Liang 2021）。

个性化 RLHF 和 PEFT。 来自人类反馈的强化学习（RLHF）最近已被改编用于异质用户群体：个性化-RLHF 训练多个奖励头或偏好模型，每个模型与一个用户集群对齐（Li et al. 2024）。同时，轻量级的 PEFT 模块以低廉的成本在数千名用户中分发个性化（Tan et al. 2024）。

基于提示和配置文件的条件设定。 几个工作提出了可编辑的自然语言描述文件，模型可以引用、证明或批评这些描述文件，从而提高推荐和对话中的可解释性（Ramos et al. 2024）。像 PeaPOD 这样的软提示方法将协作和个人信号融合为一个提示，不需要模型更新就能实现强烈的个性化（Ramos, Wu, and Lipani 2024）。

在大规模语言模型中评估个性化

参考指标的局限性。 基于参考的分数（BLEU, ROUGE）捕捉词汇重叠，即使像 BERTScore 这样的语义指标也对答案是为谁准备的缺乏敏感性。因此，它们与用户感知的个性化环境中的质量相关性较差。

像 AuPEL 这样的自动评判系统使用强大的 LLM 来共同评估个性化、相关性和流利度，消除了对参考的需求（Wang et al. 2023）。PerSE 进一步在评估者上条件化一个明确的偏好档案，提高了与人类在开放式任务上的评分相关性（Wang et al. 2024）。

为了对模型进行压力测试，几个基准测试针对特定的方面：

- LaMP 和 LongLaMP 分别评估短篇和长篇的个性化生成（分类、电子邮件、评论）（Salemi et al. 2023; Kumar et al. 2024）。
- PersonaLens 通过成对的大型语言模型代理模拟丰富的多会话配置文件，并根据任务成功率、个性化程度和质量进行评分（Zhao et al. 2025b）。
- PersonaMem 跟踪用户偏好演变过程中的表现，揭示出大多数 LLM 在应对漂移的个性时都会遇到困难（Jiang et al. 2025）。
- PersoBench 专注于基于人物设定的闲聊，报告称即使在提供思维链提示的情况下，模型通常也会忽视或与提供的人物设定相矛盾（Afzoon et al. 2024）。
- PrefEval 探讨对话式问答中对显性和隐性偏好的遵循情况；随着上下文长度增加，准确性急剧下降（Zhao et al. 2025a）。

尽管取得了快速进展，现有的评估器要么 (i) 需要劳动密集型的人类注释，要么 (ii) 将所有用户视为完全相同，或者 (iii) 仅衡量质量的狭窄方面。PREF 通过以下方式解决了这些问题：

1. 构建一个融合通用质量检查和用户特定优先级的双层评分标准，
2. 操作无参考，确保可扩展性和可重复性，并
3. 保持模型无关，从而支持多样的骨干 LLM 和个性化机制。

在接下来的实验中，我们专注于开放域问答，并采用 PrefEval 基准作为我们的主要测试平台，因为它提供了与 PREF 设计目标一致的显性和隐性偏好挑战。

从个性化到评估

个性化—— “将某物调整以满足特定个人需求的行为”[1] ——在其于底层系统的可提供性中运行时最有价值。Kirk et al. (2024) 阐明了这一前沿：在技术的一般能力范围内，个性化占据了用户可以根据自己的价值观、偏好或信息需求来调整系统的区域。换句话说，一个系统原则上可能支持许多行为，但只有一个子集将在特定的上下文中对特定个人有用。

PREF 背后的设计原则。 在此观点的基础上，PREF 基于两个假设，这使得设计空间更加精确：

- 假设 1（无限注意力）。如果用户具有无限的注意力和耐心，一个完整且全面的答案即可满足任何查询，使得明确的个性化变得不必要。
- 假设 2（有限的注意力）。在现实中，注意力和耐心是稀缺的。因此，用户倾向于选择那些突出与他们个人偏好一致的信息的答案，即使这会忽略一些不太相关的细节。

图 1 形象化了这些假设。左边的面板对比了详尽的答案与精简并优先排序的答案；中间的面板对比了一个通用指南与某个特定用户所关心的子集（或排名）；右边的面板展示了在实践中，当一般指南忽略了某些显著

[1] https://dictionary.cambridge.org/dictionary/english/personalization

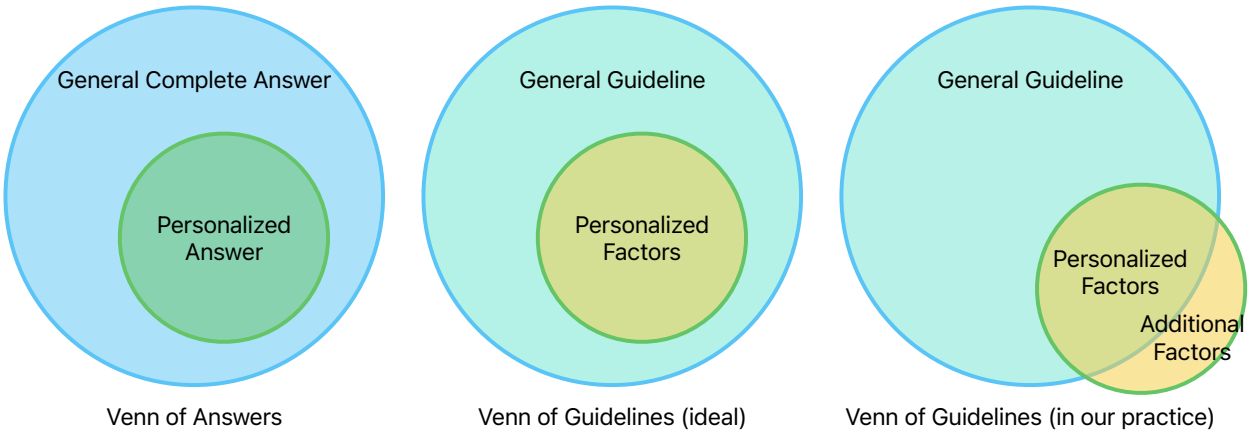

Figure 1: PREF 一览。文氏图展示了（左）一个理想、详尽的答案与对同一查询的个性化答案之间的关系；（中）理想的通用指导方针与用户特定选择（或排序）的因素之间的关系；（右）PREF 采用的实际方案，其中排序器可能在通用指南不完整时添加额外因素以制定个性化指导方针。

Algorithm 1: 偏好评分流程

Require: Question set $Q = \{q_n\}_{n=1}^{N}$,
Preference set $P = \{p_n\}_{n=1}^{N}$,
Answer set $A = \{a_n\}_{n=1}^{N}$,
Coverage LLM $\mathcal{M}_{\text{cov}}$,
Preference LLM $\mathcal{M}_{\text{pref}}$,
Scoring LLM $\mathcal{M}_{\text{score}}$
Ensure: Personalized scores $S = \{s_n\}_{n=1}^{N}$
Phase 1: Generate General Guidelines
for each $q_n \in Q$ **do**
$g_n \leftarrow \text{GenerateGuideline}(\mathcal{M}_{\text{cov}}, q_n)$
▷ g_n lists factors $\{f_{n,1}, \ldots, f_{n,K_n}\}$
end for
Phase 2: Personalize Guidelines
for each index n from 1 to N **do**
$g_n^* \leftarrow \text{Personalize}(\mathcal{M}_{\text{pref}}, q_n, p_n, g_n)$
▷ Re-rank / augment factors to obtain personalized guideline
end for
Phase 3: Score Answers
for each index n from 1 to N **do**
$s_n \leftarrow \text{ScoreAnswer}(\mathcal{M}_{\text{score}}, q_n, p_n, g_n^*, a_n)$
end for
return S

因素时，排名者如何可能注入额外因素。这个最后的许可反映了一种务实的现实：当前的 LLMs 在覆盖过程中可能会遗漏重要的标准；允许增强可防止这些遗漏传播到最终得分中。

从原则到实践。 PREF 将其转化为一个具有清晰接口的模块化流水线。我们用 g 表示一个用于查询 q 的通用指南，该指南列举了显著的评估因素 $\mathcal{F} = \{f_1, \ldots, f_n\}$ 。然后通过使用来自用户档案 p 的信息对 $\mathcal{F}$ 重新排序或重新加权，获得个性化指南 g^* 。最后，评分组件在 (q, p) 的上下文中对候选答案 a 与 g^* 进行评估，以产生一个标量 $s_a \in \mathbb{R}$ 。直观地说，g 确保了覆盖率（我们知道需要检查什么），而 g^* 确保了对齐（我们知道首先检查什么以及它的重要性）。

两阶段评估框架。 PREF 将这个想法分为两个阶段进行操作，然后进行评分（图 2）：

1. 覆盖阶段（$q \to g$ ）。考虑一个查询 q ，覆盖大模型生成一个全面的指南 g ，列举出显著因素 $\mathcal{F}$ 。每个因素可以选择性地包括简短的评分标准或测试，使 g 可以直接付诸实施。[2]
2. 偏好阶段（$p \to \pi$ 或 $\boldsymbol{w}$ ）。一个以用户配置文件 p 为条件的偏好 LLM，会对 $\mathcal{F}$ 引发一个排序 π 或一个权重向量 $\boldsymbol{w} \in \mathbb{R}_{\geq 0}^{|\mathcal{F}|}$ 。当覆盖不完整时，排序器可能会用从 p 或查询上下文中导出的额外因素来扩充 $\mathcal{F}$ ，从而产生 g^* 。
3. 评分阶段（$a \mapsto s_a$ ）。评分 LLM 在 (q, p) 的背景下评估 a 对比 g^* ，产生最终的个性化评分 s_a 。我们写 $s_a = \text{Score}(a \mid q, p, g^*)$ 。

将覆盖度与偏好分离带来了三个实际优势：

1. 对遗漏的稳健性。覆盖率是特定于查询的，并且与模型无关；偏好阶段可以添加丢失但对用户至关重要的因素。

[2]图 2 显示了一个 g 的例子。

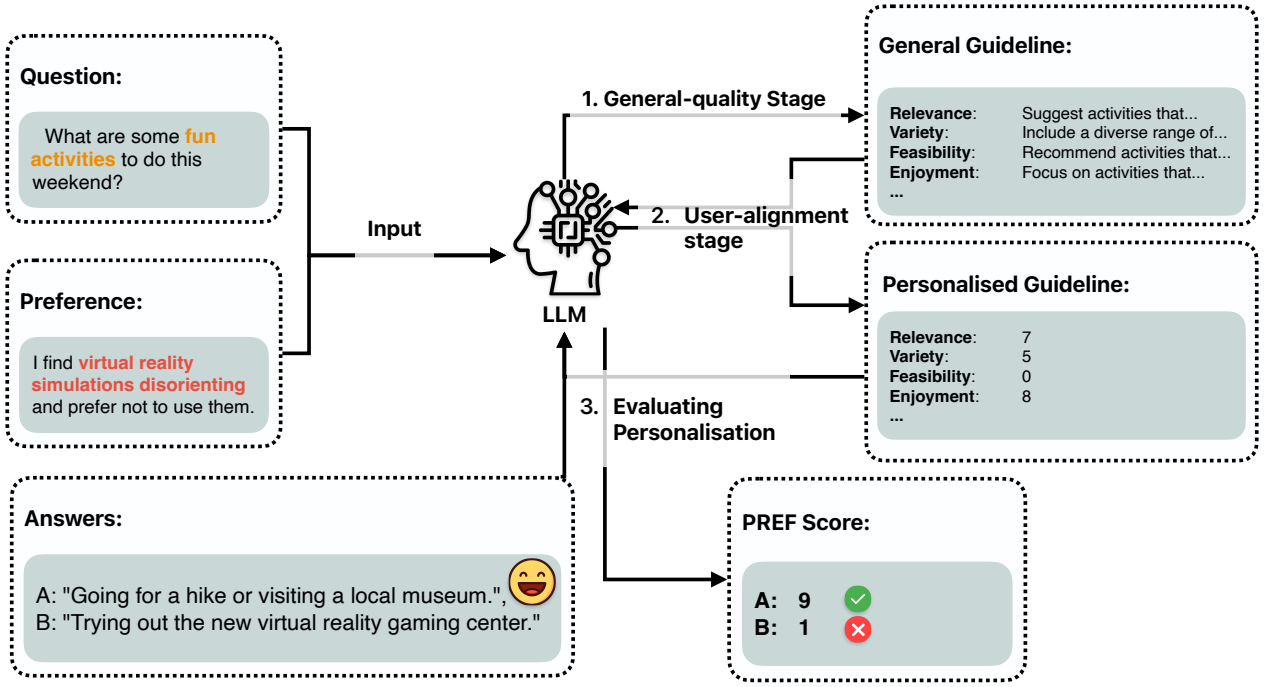

Figure 2: PREF 评分流程。三步过程——覆盖率、偏好（用户档案）和评分——将查询 q 、用户档案 p 和候选答案 a 映射到一个在 PrefEval 上的真实个案中的最终个性化评分 s_a 。

2. 透明性和可控性。g 中的因素（以及它们在 g^* 中的用户调整顺序）构成了一个人类可审计的标准，解释了为什么会被分配一个分数。
3. 可重用性。一个 g 可以支持多个用户（变化的 p ），而一个 p 可以应用于相关查询（变化的 q ），从而减少重新计算并实现消融实验。

实际上，评分者平衡了通用的期望（例如，正确性、完整性、清晰性）与用户特定的期望（例如，饮食限制、语气、预算）。设计上强制执行两个不变点：(i) 用户对齐绝不能为模型引入的事实错误找借口，且 (ii) 增强不能无理由地与覆盖因素相矛盾。这些防护措施有助于在捕捉个人偏好的同时保持可靠性。

图 2 和算法 1 详细说明了流程。给定 q 、p 和候选答案 a ：

1. 一个覆盖率 LLM 生成 g 。
2. 一个偏好 LLM 重新排序（并在必要时增加）g ，以获得 g^* 。
3. 一个评分 LLM 在 (q, p) 的背景下评估 a 与 g^* ，生成个性化得分 s_a 。

为简化，我们为这些角色全派一个骨干 LLM。

实验与发现

为了评估 PREF 的表现，我们采用 PrefEval 基准，它提供了为个性化文本生成量身定制的问题-偏好-答案三元组（Zhao et al. 2025a）。我们专注于 PrefEval 中隐含的多项选择子集，其中用户偏好与“理想”答案之间的联系未明确说明。例如，如果用户不喜欢鱼类，询问推荐日式菜肴的问题就会存在隐含的挑战，因为许多日式菜肴都含有鱼，但问题并未提到这一点。每个例子提供四个候选答案，只有一个被人工标注者标记为令人满意。

按照惯例，我们保留 20 % 的数据作为测试集，并在此分割上报告所有指标，包括 200 个问题、200 个用户偏好和 800 个答案。

为了探查 PREF 的灵活性，我们将其与四种不同的主干语言模型配对：

- Claude 3 俳句: claude-3-haiku-20240307
- GPT-4.1 Mini: gpt-4.1-mini ，
- LLaMA 3 8B：llama3-8b-instruct，以及
- LLaMA 3 70B: llama3-70b-instruct 。

对于每个模型，我们将采样温度设置为 0，以保证生成的确定性和完全可重复性。

评估个性化

基线和整体准确性。 表 1 在 PrefEval 隐式多项选择基准上对比了三种评估策略：

1. 零次：主干 LLM 使用其默认提示回答每个问题；
2. 提示：在开头添加一句话，指示模型“考虑用户的陈述偏好”（如原始的 PrefEval 论文中所述）；
3. PREF ： 我们的完整两阶段框架。

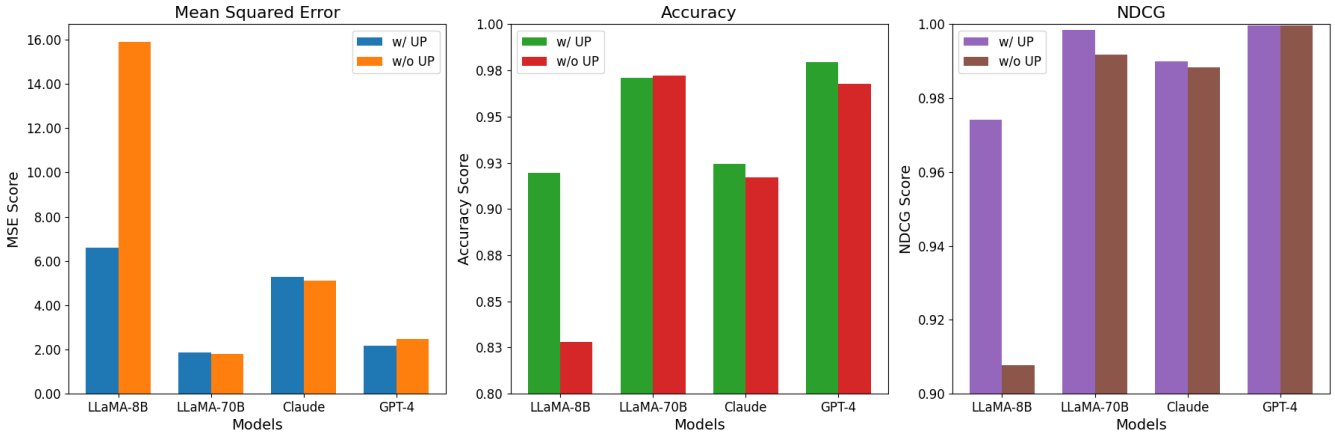

Figure 3: 关于用户偏好的消融实验（UP）。使用 UP：完整的 PREF 流程；不使用 UP：仅根据阶段一的一般指导方针对答案进行评分。图表报告了在 PrefEval 隐式多项选择拆分上的准确率、MSE（越低越好）和 nDCG（越高越好）。

Method	LLaMA 3 8B	LLaMA 3 70B	Claude 3 Haiku
Zero-shot	27	37	34
Reminder	84	91	88
PREF (ours)	92	97	92

Table 1: 在 PrefEval 隐式多项选择子集上的准确率（%）。加粗的数字表示每个骨干模型的最佳得分。零样本和提醒器结果取自 PrefEval 论文（Zhao et al. 2025a）。

Model	Accuracy ↑	MSE ↓	nDCG ↑
Claude 3 Haiku	0.92	5.27	0.9899
GPT-4.1 Mini	0.98	2.19	0.9996
LLaMA 3 8B	0.92	6.62	0.9742
LLaMA 3 70B	0.97	1.87	0.9980

Table 2: PrefEval 基准测试的表现。在准确度和 nDCG 方面，数值越高越好，而 MSE 则是越低越好。粗体标记每个指标的最佳得分。

在所有的骨干网络中，PREF 达到了最高的准确率，超越了提醒基线 $+3-8$ 个绝对点（相对提高 $\approx 6-12\%$），并且相比简单的零样本设置高出超过 $+55$ 个点。这种持续的增益强调了两个观察结果：(a) 简单地提醒模型用户的偏好是有帮助的，但不够充分，且 (b) 明确地构建和加权个性化指南—这是 PREF 的核心思想—在无论模型规模或架构的情况下，都可以显著提高评估质量。

超越准确性：MSE 和 nDCG。 准确性只计算正确答案是否排名第一；它忽略了得分差距和干扰项的排序。为了获得更细致的视角，我们将正确答案映射为 10 分，将干扰项映射为 0 分，跨问题连接分数，并计算：

- 均方误差（MSE）：惩罚与目标 10/0 的较大偏差；数值越小越好。
- 标准化折扣累计增益（nDCG）：奖励将黄金答案置于前几名的排名；数值越高越好。

表 2 显示了两个趋势。首先，较大的模型—GPT-4.1 Mini 和 LLaMA 3 70B—达到了最佳准确率和最低均方误差（MSE），这证实了当与 PREF 配对时，增加容量可以转化为更细致的评分粒度。其次，较小的模型（Claude 3 Haiku 和 LLaMA 3 8B）尽管具有更高的均方误差（MSE），但仍接近完美的 nDCG。他们可靠地将正确答案排在第一位（良好的 nDCG），但有时会过于慷慨地给予干扰项较高的评分（夸大的 MSE）。

为隔离第二阶段的贡献，我们完全去除了它——仅根据一般指南来判断答案。图 3 总结了有无用户偏好（UP）的影响。

- MSE。第二阶段降低了 LLaMA 3 8B 和 GPT-4.1 Mini 的 MSE（更好的校准）。对于其他模型（Claude 3 Haiku，LLaMA 3 70B），变化是边缘化的，暗示它们已经隐式地近似用户权重。
- 准确率和 nDCG。除 LLaMA 3 70B 的微小 0.1 点下降外，所有骨干网的准确率都上升；nDCG 全面提高，确认个性化评价标准将正确答案在排名中放得更高。

因此，第二阶段对于较小的模型是最有价值的，能够缩小与较大模型之间的差距。值得注意的是，LLaMA 3 8B 在使用 UP 的情况下可与 Claude 3 Haiku 不使用 UP 的性能相媲美——这是一个对资源有限的部署来说具有吸引力的权衡。

要点。 PREF 的两级评分标准提供了三个具体的好处：

1. 通过显式偏好整合实现稳健准确性；
2. 通过阻止在不相关答案上出现高分来获得更好的校准（更低的 MSE）；
3. 能力放大：第二阶段使得廉价的骨干模型接近平板更大模型的性能，在不牺牲用户对齐的情况下降低服务成本。

这些结果表明，PREF 是评估个性化语言生成的有效、可扩展且与模型无关的基础。

可解释性

利用 PrefEval 的答案解释。 与许多个性化基准不同，PrefEval 不仅提供每个查询的黄金（个性化）答案，还

	GPT-4.1 Mini	LLaMA 3 70B
Pearson r	0.6164	0.4693
Spearman ρ	0.5906	0.3811
Kendall τ	0.5612	0.3532

Table 3: 用户偏好诱导的因子排序与从 PrefEval 答案解释中得出的标准排序之间的等级-等级相关性。较高的值表示更高的一致性。

提供自然语言解释，说明为什么该答案是合适的。这个额外的信号使我们能够探讨 PREF 的不同能力：其识别和优先考虑正确评估因素的能力。

具体来说，我们将解释视为“神谕”偏好并问：PREF 在基于用户偏好的情况下产生的排序是否与我们在基于解释的情况下获得的排序一致？形式上，让 e 是 PrefEval 提供的附带解释，$g(q) = \{f_1, \ldots, f_n\}$ 是阶段 1 中产生的一般指导方针。

然后我们执行四个步骤：

1. 基于偏好的排名。将 p 输入到偏好 LLM 中，以在 $g(q)$ 上得出一个顺序 π_p （或权重向量 $\boldsymbol{w}_p$ ）。
2. 基于解释的排名。用解释 e 替换 p ，并获得第二个排序 π_e （或 $\boldsymbol{w}_e$ ）。我们将 π_e 视为一个神谕，因为它反映了黄金答案背后的人类推理。
3. 比较排名。用合适的相关性指标（例如，Kendall 的 τ 、Spearman 的 ρ 和 pearson 的 p ）测量 π_p 和 π_e 之间的相关系数。高相关性表明，偏好条件下的评分标准揭示了人类认为决定性的相同因素。

这个过程隔离了 PREF 的因子排序能力：即使最终评分不同，我们仍可以验证该框架是否关注对于高质量个性化答案最重要的标准。

表格 3 报告了在用户偏好条件下，PREF 生成的因素排名与从 PrefEval 的人为书写解释中获得的“oracle”排名之间的相关性。所有三个系数——Pearson 的 r ，Spearman 的 ρ 和 Kendall 的 τ ——都为正并且在统计学上具有显著性，证实了这两个主干结构大致识别了人类认为决定性的相同标准。

GPT-4.1 Mini 表现出更强的对齐性，达到 Pearson r 约为 0.62 和 Kendall τ 约为 0.56，而 LLaMA 3 70B 则分别落后于 0.47 和 0.35。从实际角度来看，GPT-4.1 Mini 不仅能将正确的因素放在靠前的位置，还能更准确地保持它们的相对顺序。差距表明，尽管 LLaMA 3 70B 具有更大的参数量，但仍然从第二阶段中获得好处（表中的准确性提升 1 ），但需要进一步调整以内部化细微的偏好线索。

总体而言，中等偏高的相关性表明 PREF 的偏好模块成功揭示了人类在证明一个好的个性化答案时所参考的因素，但仍有提升空间——特别是对于开源模型来说，需要加强这种对齐。

第二阶段的附加因素

在第二阶段，偏好模型可能会补充一般准则。两个简明的观察：

(1) 添加很少。在 GPT-4.1 Mini 中，我们观察到 1,986 个一般因素对比 153 个添加的因素（ $\approx 7.7\%$ ）；在 LLaMA 3 70B 中，1,893 对比 241（ $\approx 12.7\%$ ）。每个问题，这平均是 9.93 个一般因素和 0.765 个额外因素。因此，覆盖范围已经涵盖了大多数标准；阶段 2 进行有针对性的修复。

(2) 增加项编码排除项。关键词过滤（dislike ，avoid ）表明在 GPT-4.1 Mini 上添加因素的 25.49%（对比 0% 一般）以及在 LLaMA 3 70B 上 5.81%（对比 0.01% ）表达用户特定的“黑名单”。

要点。增强是有选择的，主要用于注入用户依赖的约束，而这些约束在不考虑偏好的覆盖阶段被忽略。

讨论

我们的实验支持四个主要见解：

1. 细粒度的个性化是值得的。因素级别的准则是必不可少的。PREF 比零样本回答提升了 $\geq 55\%$ 的准确率，相比单行提醒提示则提升了 3%−8%（表格 1 ）。
2. 两个阶段比一个阶段更好。去掉用户偏好会降低校准（更高的 MSE）和排序质量（更低的 nDCG），尤其是对于 8 B 参数模型（图 3 ）。仅仅依靠覆盖是不够的。
3. 较小的模型可能超越其预期表现。利用 PREF ，LLaMA 38B 的性能接近更大规模的 Claude 3 Haiku，这表明更好的评估——不仅仅是更大的模型——可以缩小质量差距。
4. 排序因素反映了人类的依据。偏好条件下的排名与从人类解释中提取的“神谕”排名中度到强烈相关（表 3 ），这表明该框架展示了人们实际上使用的标准。

实际收获

在具有严格延迟或预算限制的情况下，组织可以安全地同时部署较小的主干网络与 PREF ，从而满足个性化目标并降低推理成本。

由于 PREF 不依赖于参考，它自然地嵌入到 CI/CD 管道中，使得随着提示、奖励模型或用户群体的演变，可以进行每晚的回归测试。

个性化指南 g^* 同时也是一个简明的、易于理解的解释。开发者和用户可以检查是哪些因素导致了低评分，并调整答案或者偏好权重。

局限性与未来工作

- 评估者的可靠性。评分者本身是一个大型语言模型，因此会继承潜在的偏见或幻觉。混合设置——将多个评估者集成或由人进行抽查——是一个有前景的下一步。
- 概况复杂性。我们假设有结构化或简短自然语言的偏好。现实世界的信号是噪声的、隐含的和随时间变化的。扩展第二阶段以摄取此类数据仍未解决。
- 任务的普遍性。我们的研究集中在开放域问答上。将 PREF 应用于长篇总结、代码审查或多模态任务可能需要更丰富的覆盖因素和新的聚合方案。
- 伦理保障。个性化可能会放大信息过滤泡沫或泄露敏感特征。未来的工作应整合公平性测试，并赋予用户审查或修改其个人资料的权限。

结论

我们介绍了 PREF ，这是一种个性化的、无参考的评估框架，可以将覆盖性与偏好分开。通过首先确保主题的完整性，然后根据用户优先事项定制评分标准，PREF 在不需要标准参考的情况下提供精确、透明且可扩展的评估。对 PrefEval 基准的实验表明，PREF (i) 与人类判断高度一致，(ii) 提升了模型的校准和排名质量，以及 (iii) 为更小的骨干网络提高了上限。

代码和评估脚本将在发表时发布。我们希望此框架能够加速真正以用户为中心的语言生成研究——以及与这一目标相匹配的评估方法。

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Afzoon, S.; Naseem, U.; Beheshti, A.; and Jamali, Z. 2024. PersoBench: Benchmarking Personalized Response Generation in Large Language Models. CoRR, abs/2410.03198.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., Advances in Neural Information Processing Systems, volume 33, 1877–1901. Curran Associates, Inc.

Brusilovsky, P. 2001. Adaptive hypermedia. User modeling and user-adapted interaction, 11(1): 87–110.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In International Conference on Learning Representations (ICLR 2020). ArXiv:1912.02164.

Jiang, B.; Hao, Z.; Cho, Y.; Li, B.; Yuan, Y.; Chen, S.; Ungar, L. H.; Taylor, C. J.; and Roth, D. 2025. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale. CoRR, abs/2504.14225.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. Nature Machine Intelligence, 6(4): 383–392.

Kumar, I.; Viswanathan, S.; Yerra, S.; Salemi, A.; Rossi, R. A.; Dernoncourt, F.; Deilamsalehy, H.; Chen, X.; Zhang, R.; Agarwal, S.; Lipka, N.; Nguyen, C. V.; Nguyen, T. H.; and Zamani, H. 2024. LongLaMP: A Benchmark for Personalized Long-form Text Generation. arXiv preprint arXiv:2407.11016.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. In Erk, K.; and Smith, N. A., eds., Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 994–1003. Berlin, Germany: Association for Computational Linguistics.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

Li, X.; Zhou, R.; Lipton, Z. C.; and Leqi, L. 2024. Personalized language modeling from personalized human feedback. arXiv preprint arXiv:2402.05133.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL–IJCNLP 2021), 4582–4597.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311–318.

Ramos, J.; Rahmani, H. A.; Wang, X.; Fu, X.; and Lipani, A. 2024. Transparent and Scrutable Recommendations Using Natural Language User Profiles. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 13971–13984.

Ramos, J.; Wu, B.; and Lipani, A. 2024. PeaPOD: Personalized Prompt Distillation for Generative Recommendation. arXiv preprint arXiv:2407.05033.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; and Riedl, J. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work, 175–186.

Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2023. LaMP: When Large Language Models Meet Personalization. arXiv preprint arXiv:2304.11406.

Tan, Z.; Zeng, Q.; Tian, Y.; Liu, Z.; Yin, B.; and Jiang, M. 2024. Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024). ArXiv:2402.04401.

Van Der Lee, C.; Gatt, A.; Van Miltenburg, E.; Wubben, S.; and Krahmer, E. 2019. Best practices for the human evaluation of automatically generated text. In Proceedings of the 12th international conference on natural language generation, 355–368.

Wang, D.; Yang, K.; Zhu, H.; Yang, X.; Cohen, A.; Li, L.; and Tian, Y. 2024. Learning Personalized Alignment for Evaluating Open-ended Text Generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 13274–13292. Miami, USA.

Wang, Y.; Jiang, J.; Zhang, M.; Li, C.; Liang, Y.; Mei, Q.; and Bendersky, M. 2023. Automated Evaluation of Personalized Text Generation using Large Language Models. arXiv preprint arXiv:2310.11593.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Zhang, Z.; Rossi, R. A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; Dernoncourt, F.; Barrow, J.; Yu, T.; Kim, S.; et al. 2024. Personalization of large language models: A survey. arXiv preprint arXiv:2411.00027.

Zhao, S.; Hong, M.; Liu, Y.; Hazarika, D.; and Lin, K. 2025a. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net.

Zhao, Z.; Vania, C.; Kayal, S.; Khan, N.; Cohen, S. B.; and Yilmaz, E. 2025b. PersonaLens: A Benchmark for Personalization Evaluation in Conversational AI Assistants. arXiv preprint arXiv:2506.09902.

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.