从自然语言反馈中学习以实现个性化问答

Alireza Salemi
University of Massachusetts Amherst
Amherst, MA, USA
asalemi@cs.umass.edu

Hamed Zamani
University of Massachusetts Amherst
Amherst, MA, USA
zamani@cs.umass.edu

Abstract

个性化对于提高语言技术的有效性和用户满意度至关重要，特别是在信息搜索任务中，如问答系统。当前通过个性化大型语言模型（LLMs）的常规方法通常依赖于检索增强生成（RAG），然后通过具有标量奖励信号的强化学习来教导模型如何使用检索到的个人化背景。我们认为这些标量奖励有时提供的是弱且非指导性的反馈，限制了学习效率和个性化质量。我们引入了 VAC，这是一种新颖的个性化回应生成框架，它用基于用户档案和问题叙述生成的自然语言反馈（NLF）替代标量奖励。NLF 作为一种丰富且可实施的监督信号，使策略模型能够迭代地改进其输出并内化有效的个性化策略。训练在优化反馈模型和在改进后的响应上微调策略模型之间交替进行，最终形成的策略模型在推理时不再需要反馈。在由三个不同领域组成的 LaMP-QA 基准上的评估显示出相较于最新的结果具有一致和显著的改进。人工评估进一步确认了生成的响应具有卓越的质量。这些结果证明了 NLF 提供了更为有效的信号来优化个性化问答。

ACM Reference Format:
Alireza Salemi and Hamed Zamani. 2026. 从自然语言反馈中学习以实现个性化问答. In *Proceedings of Preprint*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 介绍

个性化已经成为以人为中心的系统中一个关键组成部分，例如搜索 [4, 5, 15, 34] 、推荐 [17, 18, 20] 和文本生成 [11, 24, 26, 28] ，因为它提高了用户满意度，增加了互动性，并改善了整体系统效率 [33] 。通过根据个人用户的偏好和情境定制输出，个性化系统能够提供更相关和更有效的互动 [11, 25] 。以前关于个性化文本生成的工作主要集中在内容生成 [11, 25–27] ，而这与信息获取本质上不同。在内容生成中，目标是模仿用户的写作风格和偏好，而在信息获取中，主要目标是向用户提供个性化的相关信息。在信息获取的上下文中，例如问答，个性化特别有价值，因为它能够生成与用户意图、背景和偏好相一致的响应，从而产生更准确、相关和具有用户特异性的响应 [28] 。

以往关于个性化大规模语言模型（LLMs）的工作主要依赖于检索增强生成（RAG）[26, 28] ，其中个性化信息从用户档案中检索出来，并附加到提示中以引导模型的输出。为了针对个性化优化这个 RAG 流程，各种方法已被探索。例如，当有用户的真实输出时，可以训练模型生成这些基于检索到的个性化上下文的输出 [26] 。然而，对于许多用户和任务来说，这样的标记数据通常是不可用的。为了解决这个问题，已经采用了强化学习以进一步增强模型使用个性化信息的能力，通常通过使用个性化标量奖励——这可以通过个性化奖励模型学习或从用户提供的解释或标准中得出——来反映生成响应的质量 [25] 。

这些优化方法面临几个限制。首先，用户提供的真实输出仅代表众多可能可接受的响应之一，因此对单一目标进行监督训练并不理想，容易陷入局部极小值 [25] 。此外，在强化学习的情况下，标量奖励提供的监督相对较弱——它们表明输出是好是坏，但缺乏可操作的改进反馈。因此，模型必须在没有明确指导的情况下推断出有效的调整。此外，使用标量奖励进行优化往往需要对范围广泛的输出进行探索，导致收敛速度慢和训练成本增加。

为了解决这些挑战，我们引入了 VAC 和 [1] ，这是一种框架，它用生成自然语言反馈（NLF）的反馈模型替换了标量奖励模型，这些反馈是在训练期间对策略模型的响应生成的。NLF 是基于来自用户档案和用户编写的问题叙述的个性化用户偏好生成的。此反馈提供了一种更丰富且更易于解释的监督形式，引导策略模型生成更个性化的输出。如图 1 所示，在 VAC 中的训练遵循一个逐轮迭代的过程。在每次迭代中，首先优化反馈模型生成反馈，当这种反馈用于修订当前策略模型的某个响应时，根据特定任务的评估指标实现可测量的改进。培训后，反馈模型为当前策略模型生成的响应生成反馈。然后，这些反馈用于指导策略模型编辑其初始响应以改善它们以适应用户。最后，通过监督学习对策略模型进行微调，以直接从输入生成改进的响应，消除在推理时对反馈的需求。与仅提供粗略和间接监督信号的标量奖励相比，自然语言反馈提供了明确和可操作的指导。在当前策略模型生成的输出上训练反馈模型使其能够适应模型不断变化的行为和能力，产生更具针对性和有效性的反馈。相反，在精炼反馈的响应上训练策略模型使其能够内化有效的个性化模式，提高其生成质量而不依赖于测试时的反馈。

为了评估 VAC ，我们在最新的语言模型个性化问答基准（LaMP-QA）[28] 上进行实验，该基准包含三个不同的领域。我们的结果表明，VAC 始终优于所有基线，相比非个性化基线取得了 13.6 % 的相对提升，相比表现最好的个性化基线提升了 3.6 %，且在推理时间方面提高了 1.9 × 的效率，并且相比带有标量奖励的强化学习提升了 6.0 %。此外，人工评估显示，VAC 在 44 % 的情况下更受青睐，在 33 % 的情况下持平，仅在 23 % 的比较中不如最新的方法。此外，我们提供了详细的消融研究，分析了所提出方法的各个方面，包括不同优化策略的影响以及反馈模型大小对整体性能的影响。为了促进未来的研究，我们公开了我们的代码和数据。 [2]

2 相关工作

2.0.1 ．个性化

[1] 在梵语中，Vāc 是言语、语言和智慧的拟人化女神。VAC 也代表自定义（或个性化）的言语对齐。
[2] 可在：https://github.com/alirezasalemi7/VAC

Preprint,

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

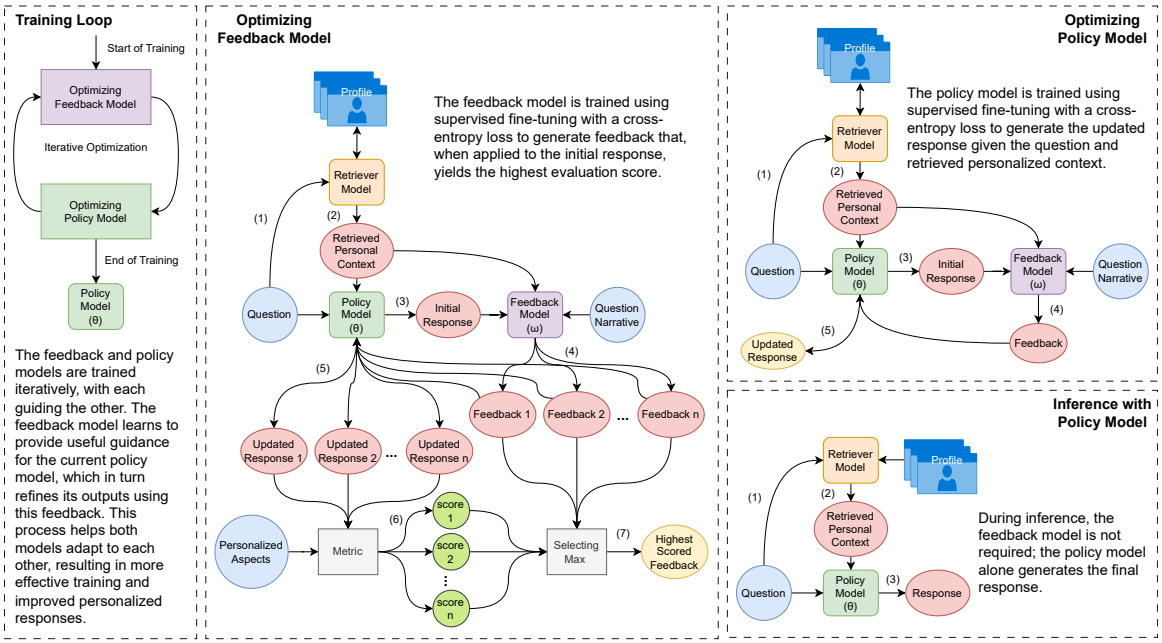

Figure 1: VAC 框架中训练循环和推理过程的概述，展示了训练过程中反馈模型和策略模型之间的交互，以及推理时策略模型的使用。

个性化在搜索、推荐和文本生成中起着核心作用，正如其已被证明可以提高用户满意度、效率和长期参与度。个性化对于问答尤其有益，因为它可以使模型生成的回答更符合用户的偏好、背景和先前知识，从而最终提供更相关和更有效的答案。在这项工作中，我们专注于个性化问答，据我们所知，LaMP-QA 是唯一公开可用的基准。

为了实现 LLM 的个性化，Salemi et al. [26] 提出了一种 RAG 框架，该框架从用户资料中检索信息，并将其整合到提供给 LLM 的提示中。此外，Salemi et al. [25] 通过使用强化学习优化 LLM，以更好地整合检索到的个人背景，从而扩展了这种方法。除此之外，现有的个性化方法涵盖了一系列策略，包括使用个性化相关性反馈训练检索器 [24] 、通过用户特定的监督微调 LLM [9] 、以及设计针对个人用户的提示 [12] 。参数高效的微调也已被探索用于个性化生成 [31] ，最近的研究将这些技术集成到 RAG 管道中 [27] 。此外，推理和自我训练在改善长篇个性化生成方面表现出潜力 [25] 。个性化助手的研究覆盖了各种领域，包括教育和企业应用 [13, 16, 19, 38] 。尽管个性化生成受到关注，但个性化问答相对较少被探索。

2.0.2 . 从自然语言反馈中学习

由于语言的表达性，自然语言反馈（NLF）已被用于数学推理和代码生成等任务中的训练信号，在这些任务中，标准答案是明确定义的 [2, 3] 。这些研究表明，人类撰写的自然语言反馈可以显著提高模型表现，而其他大规模语言模型生成的反馈往往效果不佳。在推理时，自然语言反馈也被用于协作设置中，在这种设置中，两个模型通过迭代反馈共同解决一个任务 [14, 21, 32, 35, 36] 。另一类研究在推理时利用自然语言反馈优化提示词而非模型参数 [37] ，尽管这些方法通常会在测试时引入延迟。

本文首次尝试使用 NLF 进行个性化。我们的工作在几个关键方面不同于以往的研究。首先，不同于数学和代码等领域的答案是非黑即白的正确性判断，个性化要求学习主观的用户偏好，其中某个回应可能适合一个用户，但不适合另一个用户。其次，我们根据用户的信息自动生成反馈，消除了训练过程中的人工监督需求。第三，我们的方法在 RAG 框架中运行，其中检索和生成组件都对策略和反馈模型的表现起作用。最后，我们提出了一种联合训练过程，交替优化反馈和策略模型，使它们能够共同适应，从而实现更有效的学习。

3 问题表述

我们考虑一种情境，其中某个用户 u 与一个配置文件 $P_u = \{d_i^u\}_{i=1}^{|P_u|}$ 相关联，该配置文件由他们之前的问题和对应的详细描述组成。给定一个新查询 x_u ，一个 LLM π_θ 通过基于 P_u 和 x_u 生成一个个性化响应 $y_{x_u} = \pi_\theta(P_u, x_u)$ 。为了评估生成响应的质量，我们假设可以访问一组 n_{x_u} 用户特定的方面 $E_{x_u} = \{e_i\}_{i=1}^{n_{x_u}}$ ，这些方面从用户提供的个性化问题叙述 r_{x_u} 中提取。这些方面仅用于评估，在生成过程中不向策略模型开放。一个指标 $\mu(x_u, \hat{y}_u, E_{x_u}, r_{x_u})$ 通过生成输出中预期方面的解决程度来量化响应的质量。由于这些方面明确地来源于用户提供的要求，这种评估框架能够有针对性地评估响应在多大程度上符合用户的个性化信息需求。最后，我们假设存在一个训练数据集 $D = \{(x_i, P_i, E_{x_i}, r_{x_i})\}_{i=1}^{|D|}$ ，我们的方法可以从中学习。在测试集中，结构是相同的，但方面和叙述仅用于评估，从未提供给策略模型。

4 VAC 框架

如同在 1 节中讨论的那样，利用标量奖励优化策略模型 π_θ 在效果上有限，因为这些奖励只提供了总体输出质量的粗略反馈，却没有指明输出应如何改进。因此，模型必须自行探索和推断出适当的调整，这会减缓收敛速度并增加训练成本。

为了解决这些挑战，我们提出了 VAC 框架，它用反馈模型 ϕ_ω 代替标量奖励模型，该反馈模型对策略模型的输出生成自然语言反馈（NLF），以在训练过程中指导其朝向更个性化的响应。该框架遵循一种迭代训练程序，其中反馈模型和策略模型在多轮优化中交替进行。这使得反馈模型能够提高其生成有效、以个性化为导向的反馈的能力，以指导策略模型生成更个性化的响应，而策略模型则逐步学会在推断过程中无需依赖反馈即可产生更个性化的响应。本节的其余部分详细介绍了我们提出的方法。

4.0.1 ．训练流程概述：
VAC 的迭代训练循环概述如图 1 和算法 1 所示。经过 T 次迭代后，每次迭代 t 首先使用离线强化学习优化反馈模型 ϕ_{ω^t} ，以学习一种有效的反馈策略，从而改善策略模型 $\pi_{\theta^{t-1}}$ 的个性化响应，依据的是 $\pi_{\theta^{t-1}}$ 生成的响应分布。此优化过程可形式化为：

$$\omega^t = \underset{\omega}{\operatorname{argmax}} \frac{1}{|D|} \sum_{d \in D} U_{\text{feedback}}(\phi_\omega; d; \pi_{\theta^{t-1}}) \tag{1}$$

其中 D 是训练数据集，U_{feedback} 是一个目标函数，用于根据其如何改善评估指标 μ 来衡量反馈模型生成的反馈在策略模型 $\pi_{\theta^{t-1}}$ 中的使用情况，具体将在第 4.2 节中描述。

一旦反馈模型训练完成，在同一次迭代中，反馈模型 ϕ_{ω^t} 被用于为 $\pi_{\theta^{t-1}}$ 生成的输出产生反馈，策略模型则根据该反馈被要求修改其响应。然后将更新的响应用于通过监督学习微调新的策略模型 π_{θ^t} ，使其在后续迭代中无需依赖推理时的反馈就能更好地个性化其响应。这可以形式化为：

$$\theta^t = \underset{\theta}{\operatorname{argmax}} \frac{1}{|D|} \sum_{d \in D} U_{\text{policy}(\pi_\theta; d; \phi_{\omega^t})} \tag{2}$$

其中 D 是训练数据集，U_{policy} 是策略模型的目标函数，通过监督学习在应用反馈模型 ϕ_{ω^t} 的反馈后生成更新的响应，而在推理过程中不依赖反馈，这将在第 4.2 部分详细描述。

反馈生成和响应优化过程的详细信息在第 4.1 节提供，而策略和反馈模型的优化程序在第 4.2 节中描述。

4.0.2 ．推理流程概述：
在推理过程中，如图 1（使用策略模型进行推理）所示，我们采用与 Salemi and Zamani [28] 相同的 RAG 流程，其中对于来自用户 u 的问题 x_u ，其资料为 P_u ，一个检索器 R 从 P_u 中选择 K 篇相关文档。这些检索到的文档随后附加到问题上以形成提示，并输入已训练的策略模型 π_{θ^T} ，使用图中所示的提示生成响应 2 （上），形式化为 $y = \pi_{\theta^T}(x_u; R(x_u, P_u, K))$ 。

4.1 反馈生成 & 输出优化

在迭代 t 中，为用户 u 针对问题 x_u 生成的初始个性化响应 $y_{t-1} = \pi_{\theta^{t-1}}(x_u; R(x_u, P_u, K))$ 由先前迭代的策略模型 $\pi_{\theta^{t-1}}$ 生成，本节概述了生成 NLF 和相应优化响应的过程。

4.1.1 ．反馈生成
为了在迭代 t 中生成初始个性化回复 y_{t-1} 的反馈，我们首先使用检索器 R 从用户资料 P_u 中检索出 K 个文档，条件是问题 x_u 。然后，给定问题、检索到的文档、初始回复和用户 u 的问题叙述 r_{x_u} ，我们提示反馈模型 ϕ_{ω^t} 分析该回复，并生成旨在改善回复个性化的自然语言反馈（NLF），以叙述为指导。本过程使用的提示如图 3 所示。形式上，反馈定义为：$f = \phi_{\omega^t}(x_u; R(x_u, P_u, K); r_{x_u}; y_{t-1})$ 。

RAG-Personalization Prompt

```
You are a helpful assistant designed to generate personalized responses to user questions. Your task is to
answer a user's question from a post in a personalized way by considering this user's past post questions and
detailed descriptions of these questions.
# Your input:
        - The user's current question from a post.
        - The user's past post questions and detailed descriptions of these questions.

# Your task: Answer the user's current question in a personalized way by considering this user's past post
questions and detailed descriptions of these questions, to learn about the user's preferences.

# Your output: You should generate personalized answer to the user's current question by considering this
user's past post questions and detailed descriptions of these questions to learn about user's preferences. Your
output should be a valid json object in ```json ``` block that contains the following fields:
        - personalized_answer: contains the personalized answer to the user's current question considering the
        this user's past post questions and detailed descriptions of these questions to learn about user's
        preferences.

# Past post questions and detailed descriptions of these questions:
{profile}
# Current post question:
{question}

output: ``` json
```

Response Refinementl Prompt

```
Thanks for your response. The user has provided the following feedback on your response:
{feedback}
# Your task: Refine your response to the user's question based on the feedback provided by the user.
# Your output: You should generate a refined response to the user's question based on the feedback provided
by the user. Your output should be a valid json object in ```json ``` block that contains the following fields:
        - personalized_answer: contains the refined answer to the user's current question by applying the
        provided feedback.

output: ``` json
```

Figure 2: 在 RAG 个性化、离线 RL 个性化和 VAC （上部）中用于响应生成的提示，以及在 VAC （底部）中用于响应优化的提示。

4.1.2 ．输出优化
为了使用反馈 f 优化迭代 t 的初始输出 y_{t-1} ，我们将用户档案中最初检索到的文档与迭代 t 中的初始响应 y_{t-1} 和反馈一同附加，并提示策略模型 $\pi_{\theta^{t-1}}$ 根据这些信息修正其响应。该过程在图 2（底部）的提示中有所展示，并正式定义为：$y_t = \pi_{\theta^{t-1}}(x_u; R(x_u, P_u, K); y_{t-1}; f)$ 。

4.2 反馈 & 策略模型优化

本节介绍反馈和策略模型的训练目标，这些目标通过算法 1 迭代优化。

4.2.1 ．优化反馈模型
本节定义了 U_{feedback} 目标函数。反馈模型 ϕ_{ω^t} 在 t 次迭代的主要目标是生成反馈，以有效改善之前策略模型 $\pi_{\theta^{t-1}}$ 产生的个性化响应，使得改进后的响应可以为训练下一个策略模型 π_{θ^t} 提供有用的指导。因此，反馈模型的优化目标应设计成鼓励生成反馈，当策略模型使用这些反馈来修正其初始输出时，能够在响应的质量和个性化方面带来显著的提升。

为了优化反馈模型 ϕ_{ω^t} ，我们按照算法 1 （第 2–12 行）中概述的自我训练方法进行。对于数据集中的每个输入示例，我们首先使用前一轮的策略模型 $\pi_{\theta^{t-1}}$ 生成一个初步的个性化响应。然后，使用前一轮的反馈模型 $\phi_{\omega^{t-1}}$ 以高温（温度 = 1）采样生成 N 个不同的反馈候选。每个候选反馈独立应用于响应 y_{t-1} 使用 $\pi_{\theta^{t-1}}$ 生成修订后的响应。然后根据数据集中提供的个性化标准，基于下游任务特定的度量（如在第 5.1 节中解释的），对这些修订后的响应进行评估。导致最高评估分数的反馈被选择为最有效的反馈（算法 1 的第 10 行）。然后使用交叉熵损失 [30] 进行监督微调，以训练当前反馈模型 ϕ_{ω^t} 来生成选定的反馈（算法 1 的第 12 行）。模型被训练为根据问题、初始响应、检索到的文档以及问题叙述来生成这种有效的反馈。这种优化方法鼓励 ϕ_{ω^t} 生成的反馈在用于修订策略模型的输出时，能根据任务特定的评估指标产生可衡量的改进。通过这种方式，目标函数 U_{feedback} 被构建成使得优化它会产生在提高策略模型输出方面有效的反馈。

Algorithm 1 在 VAC 框架中实现训练循环。

Input: policy model π_{θ^0} , feedback model ϕ_{ω^0} , retriever R , dataset D , metric μ , number of training iterations T , number of retrieved documents K , number of generated feedback N
Output: trained policy model π_{θ^T} , trained feedback model ϕ_{ω^T}

1: **for** $t = 1$ until T **do**
2: // training the feedback model ϕ_{ω^t} for round t
3: $D_{\phi_{\omega^t}} = \{\}$ ▷ This round's training data for feedback model
4: **for** $(x_u, P_u, E_{x_u}, r_{x_u}) \in D_{\text{train}}$ **do** ▷ For each input in training dataset
5: $y_{t-1} = \pi_{\theta^{t-1}}(x_u; R(x_u, P_u, K))$ ▷ Generate initial output
6: $F = \{\}$ ▷ Set of feedbacks for this specific output
7: **for** $j = 1$ until N **do** ▷ For N times
8: $F = F \cup \{\phi_{\omega^{t-1}}(x_u; R(x_u, P_u, K); r_{x_u}; y_{t-1})\}$ ▷ Generate a feedback using feedback model with a high temperature for the generated output
9: **end for**
10: $D_{\phi_{\omega^t}} = D_{\phi_{\omega^t}} \cup \{(x_u, y_{t-1}, P_u, r_{x_u}, f) | \underset{f \in F}{\text{argmax}}\, \mu(x_u, \pi_{\theta^{t-1}}(x_u; R(x_u, P_u, K); y_{t-1}; f), E_{x_u}, r_{x_u})\}$ ▷ Find the feedback that maximizes the metric when applied to the previous generated output and add it to the training set
11: **end for**
12: $\omega^t = \underset{\omega}{\text{argmax}} \sum_{(x_u, y_{t-1}, r_{x_u}, f) \in D_{\phi_{\omega^t}}} \log p_\omega(f | x_u; R(x_u, P_u, K); r_{x_u}; y_{t-1})$ ▷ Maximize the probability of generating feedback given the generated output, input, and personalized aspects
13: // training the policy model π_{θ^t} for round t
14: $D_{\pi_{\theta^t}} = \{\}$ ▷ This round's training data for policy model
15: **for** $(x_u, P_u, E_{x_u}, r_{x_u}) \in D_{\text{train}}$ **do** ▷ For each input in training dataset
16: $y_{t-1} = \pi_{\theta^{t-1}}(x_u; R(x_u, P_u, K))$ ▷ Generate initial output
17: $f = \phi_{\omega^t}(x_u; R(x_u, P_u, K); r_{x_u}; y_{t-1})$ ▷ Generate a feedback using the optimized feedback model
18: $y_t = \pi_{\theta^{t-1}}(x_u; R(x_u, P_u, K); y_{t-1}; f)$ ▷ Apply the generated feedback to the previous generated output
19: $D_{\pi_{\theta^t}} = D_{\pi_{\theta^t}} \cup \{(x_u, P_u, y_t)\}$ ▷ Add the updated output to the training set
20: **end for**
21: $\theta^t = \underset{\theta}{\text{argmax}} \sum_{(x_u, P_u, y_t) \in D_{\pi_{\theta^t}}} \log p_\theta(y_t | x_u; R(x_u, P_u, K))$ ▷ Maximize the probability of generating the updated output given the inputs and retrieved personal documents
22: **end for**
23: **return** $\pi_{\theta^T}, \phi_{\omega^T}$ ▷ Return the fully trained policy and feedback model

4.2.2 . 优化策略模型

本节定义了 U_{policy} 目标函数。如算法 1 （第 13-21 行）所示，策略模型的优化过程如下：对于训练数据集中的每个输入，上一次迭代的策略模型 $\pi_{\theta^{t-1}}$ 首先生成对输入查询的初始响应 y_{t-1} 。接下来，在当前迭代中训练的反馈模型 ϕ_{ω^t} 与策略模型 $\pi_{\theta^{t-1}}$ 对齐，为每个初始响应生成反馈信号。然后，$\pi_{\theta^{t-1}}$ 使用该反馈修正其初始输出，从而生成改进后的响应（算法 1 的第 18 行）。最后，使用交叉熵损失 [30] 对更新后的策略模型 π_{θ^t} 进行监督微调训练，从输入查询直接学习生成改进后的响应（算法 1 的第 21 行）。这种目标函数 U_{policy} 的设计假设应用反馈后的策略模型的更新响应对用户来说是更好的个性化响应。因此，模型被训练直接重现该响应，而不需要反馈。这种训练过程帮助策略模型在推理时无需依赖反馈就能学习生成更高质量、个性化的响应。

Feedback Model Prompt

```
You are a helpful assistant designed to generate feedback for the generated response to a user's question, considering the user's detailed information need and the aspects that the user expects to see in the response to their question. Your task is to provide actionable feedback on how to improve the generated response based on the user's detailed information need and the aspects that the user expects to see in the response to their question.
# Your input:
        - The user's current question.
        - The user's past post questions and detailed descriptions of these questions.
        - The generated response to the user's question.
        - The detailed information need that the user provided in the post for this question.
# Your task: Provide actionable feedback on how to improve the generated response based on the user's detailed information need and the aspects that the user expects to see in the response to their question.
# Your output: Your output should be a valid json object in ```json ``` block that contains the following fields:
        - feedback: contains the feedback on how to improve the generated response based on the user's detailed information need and the aspects that the user expects to see in the response to their question.

# Past post questions and detailed descriptions of these questions:
{profile}
# Current post question:
{question}
# Generated response:
{generated_response}
# Detailed information need:
{question_narrative}

output: ``` json
```

Figure 3: 在 VAC 框架中使用反馈模型生成初始输出的 NLF 的提示。

5 实验

5.1 实验装置

5.1.1 . 数据集 & 评估：

我们在唯一公开可用的个性化问答数据集 LaMP-QA 基准 [28] 上进行实验，该数据集包含三个不同的领域：(1) 艺术和娱乐，(2) 生活方式和个人发展，(3) 社会和文化。每个数据集中的示例包括一个用户查询，作为用户个人资料的用户问题历史，反映用户观点和意图的问题叙述，以及规定理想回应应涉及的方面的一组个性化标准。我们实验中使用的数据集的

Table 1: LaMP-QA 基准中每个数据集的数据集统计。

Method	Arts & Entertainment			Lifestyle & Personal Development			Society & Culture		
	train	validation	test	train	validation	test	train	validation	test
# Questions (users)	9349	801	767	7370	892	989	7614	810	1074
# Evaluation Aspects	2.7 ± 0.9	4.7 ± 1.2	4.6 ± 1.2	3.1 ± 1.0	5.1 ± 1.1	5.1 ± 1.2	2.9 ± 0.9	4.8 ± 1.1	4.8 ± 1.0
Profile Size	106.7 ± 127.3	129.0 ± 183.7	159.1 ± 203.0	116.6 ± 162.0	98.2 ± 198.6	111.6 ± 220.3	141.3 ± 194.7	110.5 ± 210.6	115.8 ± 203.6

统计数据如表 1 所示。为了评估回应，按照 Salemi and Zamani [28]，我们采用了具有 320 亿参数的指令调优 Qwen 2.5 模型 [3] [22]。对于每个问题，LLM 评估回应中是否涵盖了每个个性化方面，并分配一个在范围 $[0, 2]$ 内的分数。然后将分数归一化到范围 $[0, 1]$。生成回应的最终得分是所有个性化方面的平均归一化得分。有关这方面的更多信息，我们请读者参考 Salemi and Zamani [28]。此外，在一个实验中，我们通过 VAC 与最佳基线的并列人类评估来确认我们的主要实验结果。

5.1.2 ．训练配置：

我们使用参数量为 70 亿的指令微调 Qwen 2.5 模型 [4] [22] 作为策略 LLM。将我们的方法扩展到更大或其他 LLM 需要超过 750 GPU 小时，这超出了我们的计算预算，因此我们将实验限制在一个 LLM 上，这构成了这项工作的一个局限。除非另有说明，反馈模型使用相同的主干 LLM（指令微调 Qwen 2.5 模型，具有 70 亿参数）。我们进行 $T = 3$ 次迭代的训练。为了训练反馈提供者，我们为每个输出生成 $N = 16$ 个反馈。每次迭代使用上一次迭代中表现最好的检查点来初始化模型的权重。训练使用 Adam 优化器 [10]，学习率为 5×10^{-5}，批量大小为 32，并通过最大范数为 1 的梯度裁剪来保证稳定性。模型最多训练 5000 步，前 10 个步骤应该为% 作为热身期，其后学习率线性递减。我们利用 LoRA（低秩适应）[7] 对模型进行微调，使用秩 $r = 16$，缩放因子 $\alpha = 16$，和 0.05 的 dropout 率，且不修改偏差参数。LoRA 使用 PEFT 库实现。[5] 模型检查点每 250 步在验证集上进行评估，选择表现最好的检查点进行评估。所有实验使用 4 个 NVIDIA A100 GPU，具有 80GB 的 VRAM 和 128GB 的 RAM，约 750 GPU 小时完成。

5.1.3 ．推理配置：

所有模型均配置为最大输入-输出令牌限制为 8192 个令牌。响应生成使用核采样 [6]，温度为 0.1。为了实现 LLM 的高效推理和部署，我们利用了 VLLM 库 [6]。在检索方面，我们采用 Contriever [8]，这是一种在 MS MARCO 数据集 [1] 上经过微调的密集检索器，用于从用户资料中检索 $k = 10$ 相关文档。

5.1.4 ．基线：

我们将 VAC 与一组个性化和非个性化的基线进行比较。对于非个性化基线，我们直接在没有任何用户上下文的情况下向 LLM 提供问题。对于个性化基线，我们包括以下内容：

- RAG 个性化 [26, 28]：问题用于从用户配置文件中检索相关文档，然后 LLM 使用查询和检索到的个人上下文生成响应，提示如图 2（顶部）所示。
- 带有随机用户档案的 RAG [28]：与先前的方法相似，但检索是在随机抽取的用户档案上执行，而不是实际的用户档案。该基准评估使用不匹配用户信息的影响。

[3]可在：https://hf.co/Qwen/Qwen2.5-32B-Instruct
[4]可用于：https://hf.co/Qwen/Qwen2.5-7B-Instruct
[5]可用于：https://github.com/huggingface/peft
[6]可在：https://github.com/vllm-project/vllm

Planning Prompt

```
You are a helpful assistant designed to generate the topics that user might expect to see in a response to their question. Your task is to extract the important aspects that the user expects to see in a response to their question considering the previous questions asked by the same user and the detailed information need they provided in the post.

# Your input:
    - The user's current question.
    - The user's past post questions and detailed descriptions of these questions.

# Your task: Extract the important aspects that the user expects to see in a response to their question considering the previous questions asked by the same user and the detailed information need they provided in the post.

# Your output: You should generate a list of aspects that are important for the user to be included in the response to their question.

# Past post questions and detailed descriptions of these questions:
{profile}
# Current post question:
{question}

output:
```

Generation with Plan

```
You are a helpful assistant designed to generate personalized responses to user questions. Your task is to answer a user's question from a post in a personalized way by considering this user's past post questions and detailed descriptions of these questions. Additionally, you are provided with the aspects that the user expects to see in the response to their question, which you can use to generate a personalized answer.

# Your input:
    - The user's current question from a post.
    - The user's past post questions and detailed descriptions of these questions.
    - The aspects that the user expects to see in the response to their question.

# Your task: Answer the user's current question in a personalized way by considering this user's past post questions and detailed descriptions of these questions, to learn about the user's preferences. Additionally, you should consider the aspects that the user expects to see in the response to their question.

# Your output: You should generate personalized answer to the user's current question by considering this user's past post questions and detailed descriptions of these questions to learn about user's preferences. Additionally, you should consider the aspects that the user expects to see in the response to their question. Your output should be a valid json object in ```json ``` block that contains the following fields:
    - personalized_answer: contains the personalized answer to the user's current question considering the this user's past post questions and detailed descriptions of these questions to learn about user's preferences.

# Past post questions and detailed descriptions of these questions:
{profile}
# Current post question:
{question}
# Aspects expected in the response:
{plan}

output: ``` json
```

Figure 4: 用于 PlanPers [28] 基线的提示。

- PlanPers [28]：该方法首先使用问题从用户资料中检索信息，然后基于文档和问题生成一个高层次的响应计划。在计划、检索到的文档和问题的条件下，LLM 生成最终的个性化响应。该方法使用图 4 中显示的提示进行计划和响应生成。具体细节请参阅 Salemi and Zamani [28]。
- 离线强化学习 RAG-个性化：为了对比一种利用标量奖励信号训练个性化大型语言模型的方法，我们实施期望最大化离线强化学习 [25, 29]，使用下游评估指标作为标量奖励。由于该算法的训练循环与我们提出的方法相似，因此被用于训练，使得从 NLF 和标量奖励学习之间的比较具有公平性。与 RAG-个性化基线相似，该方法首先从用户配置文件中检索一组文档。基于检索到的内容，针对每个问题使用图 2（顶部）中所示的提示生成一组 16 候选回答。然后，将基于问题叙述和个性化评分标准评分的下游任务指标应用于这些回答，并选择得分最高的回答监督下一次训练迭代。模型在与我们的方法相同的配置下接受三次迭代训练，使用所有

Table 2: LaMP-QA 基准测试集上的表现。上标 † 表示最佳表现基线与我们的方法之间通过 t 检验（$p < 0.05$）具有统计显著性差异。

Method	Runtime (second / query)	Arts & Entertainment	Lifestyle & Personal Development	Society & Culture	Average (macro)
No-Personalization	0.78	0.3129	0.4582	0.4769	0.416
RAG-Personalization (Random P)	1.65	0.2547	0.3829	0.4037	0.3471
RAG-Personalization	1.71	0.3397	0.4481	0.4967	0.4281
PlanPers	3.12	0.3518	0.4818	0.5240	0.4525
Offline RL RAG-Personalization	1.67	0.3579	0.4621	0.5070	0.4423
VAC	1.63	0.3454	0.5288 †	0.5331 †	0.4691 †

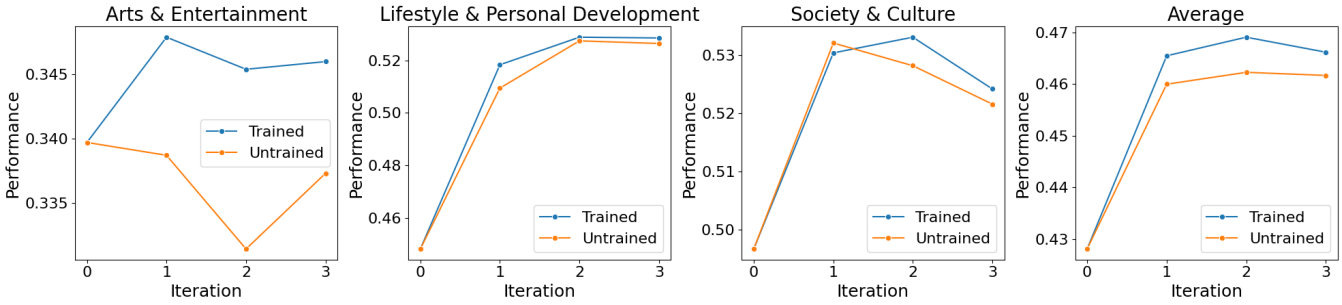

Figure 5: 训练和未训练反馈模型在不同训练迭代中的 VAC 性能。

迭代中的最佳检查点进行评估。此比较使得能够实证评估基于 NLF 的优化与基于标量奖励的优化的效率和效果。

所有基线都在与 VAC 相同的设置和条件下进行评估，包括最大输入和输出长度、训练预算、检索文档数量、检索模型和生成温度的相同配置。

5.2 实证结果

5.2.1 ．与基线的比较：

我们的方法和基准模型在 LaMP-QA 基准数据集上的结果如表 2 所示。如图所示，VAC 在整个基准测试中的平均性能方面统计上显著优于所有基准模型。具体而言，VAC 在 3 个个性化问答任务中的 2 个任务上相较于所有基准模型取得了统计上显著的改进。唯一一个 VAC 未能超越基准模型的任务是艺术与娱乐。这些结果强调了从自然语言反馈中学习对于基于用户偏好的响应生成提高个性化效果的有效性。

表 2 报告了每种方法的运行时间。其中，非个性化 LLM 的运行时间最低，主要原因是它处理的输入较短，并且没有检索开销。相比之下，所有基于 RAG 的个性化方法——包括 VAC——由于需要检索相关用户档案文档的额外成本，运行时间较长。运行时间最长的是 PlanPers 基线 [28]，由于该方法采用了双步生成方法，它的速度几乎是 VAC 的两倍，但表现显著较差。总体而言，这些结果表明 VAC 在个性化性能方面表现优越，其运行成本与最高效的个性化基线相当。

5.2.2 ．优化反馈模型的效果：

为了检查训练反馈模型对 VAC 性能的影响，我们进行了两组实验：一组是在每次迭代后同时更新策略模型和反馈模型，另一组是在反馈模型保持冻结状态的情况下仅训练策略模型。结果在图 5 中报告。如图所示，在所有数据集上，联合训练反馈模型以与不断演进的策略模型对齐始终优于冻结反馈设置。这些发现强调了在每次迭代中优化反馈模型以匹配策略模型的更新能力的重要性，从而能够生成更有效和信息丰富的反馈。

5.2.3 ．训练迭代次数（T）的影响：

为了研究训练迭代次数（T）对 VAC 性能的影响，我们对模型进行最多三次迭代训练，并在每次迭代后进行评估。图 5 中的结果表明，性能在前两次迭代中有所提高，但在第三次迭代中趋于平稳。这还表明，当反馈模型未经训练时，这种平稳效应更加明显，突显了优化反馈模型对性能的重要性。这些观察结果表明，继续使用 VAC 进行训练在几次迭代后收益递减，并且联合优化策略和反馈模型对于最大化效果至关重要。

5.2.4 ．反馈模型规模的影响：

为了研究反馈模型的规模和能力——它在训练过程中对策略模型的指导起到关键作用——如何影响其在帮助策略模型学习用户偏好方面的有效性，我们使用不同参数规模的 instruction-tuned Qwen 2.5 进行了实验：分别为 15 亿、30 亿和 70 亿。这些实验在单次训练迭代中进行，结果如图 6 所示。如图所示，较大的反馈模型始终导致策略模型的表现更好，这表明增加的容量使反馈模型能够生成更具信息性和可操作性的反馈。这些发现强调了使用强大反馈提供者对更有效地监督和指导策略模型学习过程的重要性。

5.2.5 ．优化方法对策略模型的影响：

如第 4.2 节所述，一旦使用反馈生成更新后的输出，我们通过监督式微调（SFT）使用交叉熵损失来训练策略模型以重现这个更新后的输出。另一种方法是使用比如直接偏好优化（DPO）[23] 的方法来优化模型，该方法在最大化更新后的输出可能性的同时最小化初始输出的可能性。为了研究这种替代方法的有效性，我们进行了比较实验，并在图 7 中报告了结果。结果显示，除去在艺术 & 娱乐数据集上的第一次迭代外，SFT 的表现与 DPO 相当或显著优于 DPO。我们假设这个结果源于 DPO 的一个关键假设：它将初始和更新后的输出视为截

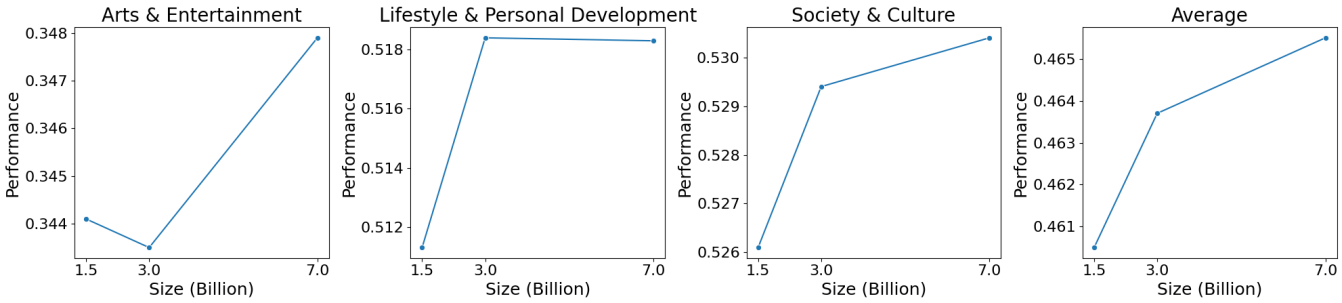

Figure 6: 反馈模型大小对 VAC 性能的影响。

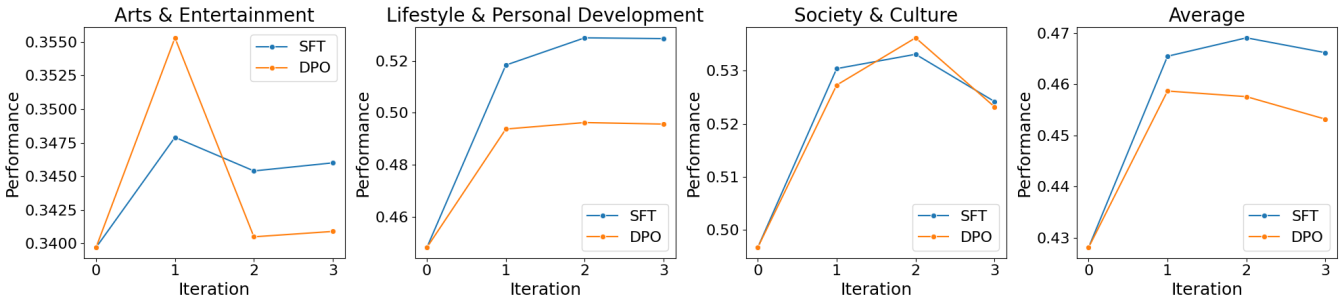

Figure 7: 不同训练迭代中优化方法对 VAC 中策略模型的影响。

然不同的对立对。然而，随着训练的进行，策略模型开始生成已经接近于更新后的输出的结果。在这种情况下，像 DPO 那样强制区分两者之间的差异可能会不必要地扭曲模型的输出分布。相比之下，SFT 更自然地适应这些细微的改进，使模型能够调整输出而不会过度修正。

5.2.6 ．人工评估：

我们进行了人工评价，以从人类视角将 VAC 与表现最好的基线 PlanPers [28] 进行比较。我们从 LaMP-QA 基准中随机抽取了 100 个示例。每个示例都被两位独立的人类标注者单独评估，他们被要求根据评分标准和问题的叙述比较响应，并确定哪个响应更好地回答了问题，或者响应是否同样优秀。使用 Cohen's kappa 测量的标注者之间的一致性为 0.7832，这表明标注者之间有很高的一致性。该评估的结果在图 9 中展示。在 44 % 的实例中，标注者更倾向于 VAC ，因为其更好地满足了评分标准。在 23 % 的实例中，更倾向于 PlanPers 基线。剩余的 33 % 被判定为平局。这些结果表明，VAC 从人类视角产生的响应更符合用户特定的评分标准，表明其在个性化问题回答中的有效性。

5.3 案例研究

本节展示了在训练期间的初始响应、反馈和更新案例研究，以及与顶级基线相比的策略模型的训练后输出。

5.3.1 ．训练期间的初始响应、反馈和更新响应：

如图 8 所示（上排），对用户有关确定火鸡汤何时煮熟的问题的初始回答总体上是信息丰富的，但缺乏结构，并遗漏了用户叙述中的关键背景细节——特别是对慢火炉灶炖煮的关注以及不同骨骼类型对烹饪时间的影响。反馈通过建议加入具体的熟度指示（如味道、稠度、清晰度）、针对炉灶烹饪的具体建议以及对骨骼差异性的考虑来解决这些不足之处。它还建议将回答组织成清晰、独立的部分以增强清晰度。更新后的回答通过列出明确的熟度标志、提供实际的监控策略，以及根据骨骼类型提供估测的烹饪时间，充分地采纳了这些建议。以结构良好的格式呈现，修订后的回答更好地反映了用户的背景，并产生了更具个性化、信息丰富和与用户契合的回答。

5.3.2 ．训练后对比 VAC 和表现最佳的基线模型：

如 8 图所示（底行），VAC 通过直接解决用户的具体背景，并更紧密地与个性化评估标准相一致，提供了比 PlanPers 更针对性的响应。虽然 PlanPers 在重新修整外部饰件方面提供一般指导，VAC 更进一步，加入了反映用户关切的细节——例如之前关于油漆耐久性的问题、对饰件材料的不确定性、以及对 Sherwin-Williams Emerald 油漆适用性的疑问。该响应包含了关于表面准备、与复合材料的兼容性及环境因素如当地气候的具体建议。这导致包括材料兼容性、油漆推荐和耐久性/寿命在内的评估维度得分更高，表明 VAC 更有效地捕捉用户意图并提供具有上下文意识的可行指导。

6 结论与未来工作

我们引入了 VAC ，这是一个新的个性化响应生成框架，将标量奖励替换为自然语言反馈作为主要监督信号。通过利用基于用户档案和问题叙述的用户特定反馈，VAC 为训练个性化大型语言模型提供了信息丰富且可操作的指导。我们的迭代训练循环——在反馈生成和策略改进之间交替——使策略模型能够在不需要推理时反馈的情况下内部化个性化策略。在 LaMP-QA 上的实验结果显示，VAC 始终优于现有的个性化和非个性化基线，并且也更受人类评估者的喜爱。

对于未来的工作，我们计划将这一基于反馈的框架从仅限于响应级别生成扩展到包含对推理过程的反馈，从而实现更个性化和透明的多步骤推理。此外，我们还希望将这种方法应用于更广泛的个性化任务，而不仅仅是问答，并研究其在不同类别的大型语言模型（包括以推理为重点的模型）上的有效性。这些方向将有助于评估自然语言反馈作为个性化文本生成监督机制的通用性和适应性。

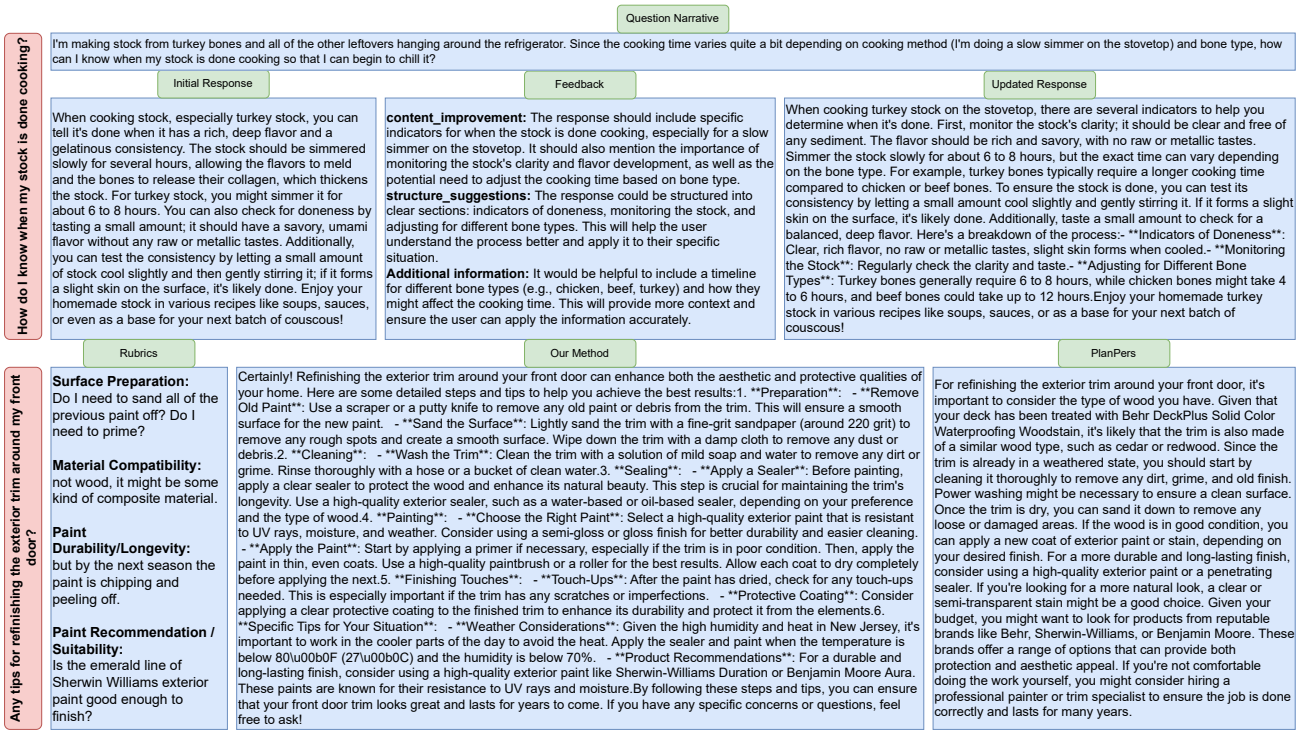

Figure 8: 案例研究展示了初始响应、反馈和更新后的响应（上排），以及 VAC 与表现最佳的基准 PlanPers [28] 之间的比较（下排）。

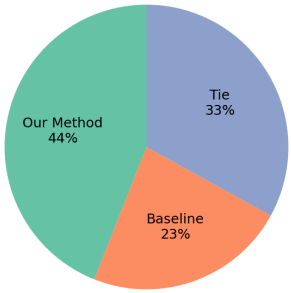

Figure 9: 在 VAC 与表现最佳的基准 PlanPers [28] 之间的人类评估结果。

7

致谢

本研究部分得到了智能信息检索中心的支持，部分得到了 Adobe Systems, Inc. 的奖励，部分得到了 Google 的奖励，部分得到了 NSF 拨款号码 2143434 的支持。本文中表达的任何观点、研究结果、结论或建议仅代表作者个人观点，并不一定反映资助方的意见。

References

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL] https://arxiv.org/abs/1611.09268

[2] Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. 2024. Learning from Natural Language Feedback. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=xo3hI5MwvU

[3] Angelica Chen, Jérémy Scheurer, Tomasz Korbak, Jon Ander Campos, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. 2024. Improving Code Generation by Training with Natural Language Feedback. arXiv:2303.16749 [cs.SE] https://arxiv.org/abs/2303.16749

[4] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07)*. Association for Computing Machinery, New York, NY, USA, 581–590. doi:10.1145/1242572.1242651

[5] Qian Guo, Wei Chen, and Huaiyu Wan. 2021. AOL4PS: A Large-scale Data Set for Personalized Search. *Data Intelligence* 3, 4 (10 2021), 548–567. arXiv:https://direct.mit.edu/dint/article-pdf/3/4/548/1968580/dint_a_00104.pdf doi:10.1162/dint_a_00104

[6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rygGQyrFvH

[7] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[8] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=jKN1pXi7b0

[9] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging. arXiv:2310.11564

[10] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). https://api.semanticscholar.org/CorpusID:6628106

[11] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. LongLaMP: A Benchmark for Personalized Long-form Text Generation. arXiv:2407.11016 [cs.CL] https://arxiv.org/abs/2407.11016

[12] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to Rewrite Prompts for Personalized Text Generation. In *Proceedings of the ACM on Web Conference 2024 (WWW ' 24)*. ACM. doi:10.1145/3589334.3645408

[13] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach LLMs to Personalize – An Approach inspired by Writing Education. arXiv:2308.07968

[14] Yanyang Li, Michael R. Lyu, and Liwei Wang. 2025. Learning to Reason from Feedback at Test-Time. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 5241–5253. doi:10.18653/v1/2025.acl-long.262

[15] Jiongnan Liu, Zhicheng Dou, Guoyu Tang, and Sulong Xu. 2023. JDsearch: A Personalized Product Search Dataset with Real Queries and Full Interactions. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2945–2952. doi:10.1145/3539618.3591900

[16] Zhuoran Lu, Sheshera Mysore, Tara Safavi, Jennifer Neville, Longqi Yang, and Mengting Wan. 2024. Corporate Communication Companion (CCC): An LLM-empowered Writing Assistant for Workplace Social Media. arXiv:2405.04656

[17] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 583–612. doi:10.18653/v1/2024.findings-naacl.39

[18] Chunyan Mao, Shuaishuai Huang, Mingxiu Sui, Haowei Yang, and Xueshe Wang. 2024. Analysis and Design of a Personalized Recommendation System Based on a Dynamic User Interest Model. arXiv:2410.09923 [cs.IR] https://arxiv.org/abs/2410.09923

[19] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. PEARL: Personalizing Large Language Model Writing Assistants with Generation-Calibrated Retrievers. arXiv:2311.09180

[20] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. arXiv:1906.00091 [cs.IR]

[21] Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning Feedback on Intermediate Representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1100–1126. doi:10.18653/v1/2024.eacl-long.67

[22] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=HPuSIXJaa9

[24] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 752–762. doi:10.1145/3626772.3657783

[25] Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. 2025. Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation. arXiv:2501.04167 [cs.CL] https://arxiv.org/abs/2501.04167

[26] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7370–7392. doi:10.18653/v1/2024.acl-long.399

[27] Alireza Salemi and Hamed Zamani. 2025. Comparing Retrieval-Augmentation and Parameter-Efficient Fine-Tuning for Privacy-Preserving Personalization of Large Language Models. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)* (Padua, Italy) *(ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 286–296. doi:10.1145/3731120.3744595

[28] Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A Benchmark for Personalized Long-form Question Answering. In *Proceedings of the The 2025 Conference on Empirical Methods in Natural Language Processing (to appear)*. Association for Computational Linguistics, Suzhou, China.

[29] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2024. Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=lNAyUngGFK Expert Certification.

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.

[31] Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts. arXiv:2406.10471

[32] Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, Xiao Wang, Rui Zheng, Tao Ji, Xiaowei Shi, Yitao Zhai, Rongxiang Weng, Jingang Wang, Xunliang Cai, Tao Gui, Zuxuan Wu, Qi Zhang, Xipeng Qiu, Xuanjing Huang, and Yu-Gang Jiang. 2024. Enhancing LLM Reasoning via Critique Models with Test-Time and Training-Time Supervision. arXiv:2411.16579 [cs.CL] https://arxiv.org/abs/2411.16579

[33] Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. Personalized Generation In Large Model Era: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 24607–24649. https://aclanthology.org/2025.acl-long.1201/

[34] Gui-Rong Xue, Jie Han, Yong Yu, and Qiang Yang. 2009. User Language Model for Collaborative Personalized Search. *ACM Trans. Inf. Syst.* 27, 2, Article 11 (mar 2009), 28 pages. doi:10.1145/1462198.1462203

[35] Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I. Wang, Wen-tau Yih, and Ziyu Yao. 2023. Learning to Simulate Natural Language Feedback for Interactive Semantic Parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3149–3170. doi:10.18653/v1/2023.acl-long.177

[36] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Bb4VGOWELI

[37] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. 2025. Optimizing generative AI by backpropagating language model feedback. *Nature* 639 (2025), 609–616.

[38] Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. LLM-based Medical Assistant Personalization with Short- and Long-Term Memory Coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2386–2398. https://aclanthology.org/2024.naacl-long.132