3 评估和测试提示

3.1 评估指标设计

在评估和测试提示时，我们需要建立一些指标，来衡量 AI 语言模型（如 GPT 和 ChatGPT）中提示的正确性和性能。这些指标可以帮助我们了解模型在不同方面的表现，以及模型的优势和劣势。我们可以根据不同的目的和需求，选择合适的指标来评估和测试提示。以下是一些常用的指标，可以分为以下几类：

（一）内容指标：这些指标用来评估模型生成的内容是否符合问题或提示的要求，以及是否具有一定的质量和价值。例如：

- ✓ 准确性：模型生成的答案是否准确回答了问题或满足了提示的要求？
- ✓ 完整性：模型的答案是否全面地涵盖了问题的各个方面？
- ✓ 创造力：人工智能模型的反应是否提供新颖或创新的想法、解决方案或观点？
- ✓ 偏见程度：模型的回答是否存在潜在的偏见，如性别、种族或文化偏见？

（二）形式指标：这些指标用来评估模型生成的形式是否符合语言或任务的规则和习惯，以及是否具有一定的美感和流畅度。例如：

- ✓ 可理解性：生成的答案是否易于理解，语言是否清晰明了？
- ✓ 语言流畅度：生成的答案是否具有良好的语法结构和自然的表达方式？

（三）效率指标：这些指标用来评估人工智能模型产生回应的速度有多快，它利用代币（tokens）的效率如何。例如人工智能模型产生回应所需的时间或代币数量。

在设计指标时，我们需要遵循以下几个原则：

（一）需要设定明确目标，根据特定用例定制指标。不同的用例可能有不同的需求和期望，比如提高准确性或减少偏见。我们需要考虑我们的应用程序的独特要求和目标，以确保指标与之保持一致。

（二）需要尽可能地将指标量化，以便于衡量和比较。量化的指标可以让我们更容易地评估模型的表现，以及发现模型的优势和劣势。为了量化指标，我们可以为准确性、可理解性和语言流畅度等指标设定评分标准，或者使用一些自动

评估工具。

（三）需要结合主观和客观评估，充分利用人工评估和自动评估的方法。主观评估可以让我们更好地了解模型的语义、逻辑和情感等方面的表现，以及模型是否符合人类的期望。客观评估可以让我们更好地了解模型的统计、数学和技术等方面的表现，以及模型是否符合预设的标准。

（四）需要定期更新指标，以适应模型的迭代和改进。随着模型的不断更新，可能会出现一些新的问题或需求，或者一些旧的问题或需求已经得到解决。这时候，我们需要对指标进行调整和更新，以确保它们仍然与我们的需求和目标保持一致。

（五）我们需要收集用户反馈，并将其纳入评估指标。用户反馈是评估模型表现的重要来源，它可以帮助我们更好地了解模型在实际使用场景中的表现，以及用户对模型的满意度。通过收集用户反馈，我们可以更好地优化模型和指标。

3.2 边缘案例和潜在风险

边缘案例指的是在模型训练数据中较少出现或与常见情况显著不同的情境，这可能导致 AI 语言模型在处理相关提示时性能降低及可靠性问题。为了提高提示质量和安全性，识别并解决这些边缘案例非常重要。

要发现边缘案例，我们可以从多个角度来进行考虑。

(1) 关注那些因措辞模糊或多重解释，导致AI模型难以理解预期含义的场景。

(2) 注意涉及争议、敏感或潜在有害主题的场景。

(3) 尝试使用各种类型的输入，包括罕见、复杂和特殊情境的问题。

(4) 思考模型在训练数据中可能建立的假设，并尝试构建违反这些假设的输入。通过分析模型在处理问题时出现的错误，以及设计一组刻意违反模型预期行为的负面测试案例，我们可以发现潜在的边缘案例。

(5) 整合用户反馈，了解他们在使用模型时遇到的问题和挑战。

针对这些潜在风险，我们可以通过优化提示设计和改进与 AI 模型的交互来提升模型性能和安全性。具体措施包括：

(1) 确保提示清晰、明确，准确地表达您期望模型生成的答案类型；

(2) 为AI模型提供额外的上下文或示例，帮助其理解和处理不太常见的主题；

(3) 在提示中添加约束，引导模型在回答问题时遵循特定的规则或原则；

(4) 将问题拆分为若干简单的子问题，逐步引导模型进行解答；

(5) 以及实施内容审核策略和消除偏见技术，以降低与敏感主题相关的风险。

遵循以上建议，我们可以更好地识别边缘案例和潜在风险，提高 AI 语言模型在处理这些情况时的质量和安全性。同时，不断优化提示设计和改进与 AI 模型的交互，也将有助于提升模型的性能和安全性。

3.3 基准性能

基准测试是提示评估过程中的关键步骤，因为它允许您根据既定标准或竞争解决方案衡量和比较提示的性能。关键基准测试策略：

(1) 内部基准测试：将提示的性能与之前的迭代、替代方法或您自己的项目或组织中的其他提示进行比较。

(2) 外部基准测试：将 AI 模型与其他类似模型或解决方案进行比较，以确定在各种任务和问题上的相对性能。这可以帮助您找到改进点，并从竞争对手中学习优秀实践。

(3) 人工评估：邀请具有相关背景和专业知识的评估者对提示进行人工评估。他们可以对提示的质量、准确性和相关性等方面给出评分和反馈。

(4) 性能跟踪：随着时间的推移持续监控和跟踪提示的性能，以确定趋势、改进或需要注意的领域。

基准测试技巧：

(1) 根据您的特定用例和目标，设定清晰、切合实际且可实现的性能目标。

(2) 利用一系列绩效指标来捕捉提示有效性的不同方面，例如相关性、连贯性、准确性、偏见、创造力和效率。

(3) 定期审查和更新您的基准，以确保它们保持相关、最新并与您不断发展的目标保持一致。

3.4 指标平衡

在提示工程中，可能会遇到指标之间发生冲突的情况。例如，可能需要权衡 AI

模型回应准确性与输出速度，或者将用户满意度与其他性能指标一起考虑。要充分平衡这些目标是比较困难的，但这对于优化提示的整体有效性至关重要。

以下是一些在提示工程中平衡竞争目标的策略：

- 确定优先级：确定哪些目标对您的特定用例最重要，并相应地确定它们的优先级。这可以帮助您专注于对用户和利益相关者最重要的性能方面，同时仍将其他目标视为次要问题。

- 设定权重：根据它们在您的应用程序中的相对重要性为不同的评估标准分配权重。这些权重可以是主观的，但应该反映出各个目标在项目成功中的相对价值，这样可以帮助您创建一个平衡的评估系统，该系统考虑到相互竞争的目标，并允许相应的优化提示。

- 使用多目标优化技术：采用多目标优化技术，例如帕累托优化或遗传算法，以找到竞争目标之间的最佳折衷方案。这些技术可以帮助您探索不同的提示设计和配置，以确定那些在多个目标中表现良好的提示设计和配置。

- 迭代和优化：根据用户反馈和性能指标不断迭代和优化您的提示。通过定期评估您的提示并根据需要进行调整，您可以在相互竞争的目标之间取得平衡，并随着时间的推移提高 AI 模型的性能。

- 分阶段实施：将项目分为多个阶段，每个阶段侧重于实现一个或多个目标。这样可以确保每个阶段都能集中精力解决某个特定目标，而在整个项目中仍然能够实现目标之间的平衡。

开发自定义评估指标是提示工程的一个重要方面。通过创建特定于应用程序的评估标准、衡量提示质量和用户满意度以及平衡竞争目标，您可以优化提示以获得更好的性能和用户体验。随着您在迅速的工程之旅中取得进展，需要注意保持敏捷和适应性，根据用户反馈和不断变化的需求不断完善提示的方法和技术。

3.5 实例

示例 1：请求摘要

差的提示：

```
请总结一下 AI。
```

好的提示：

请为以下文章提供一个简短且准确的摘要：“人工智能技术在医疗领域的应用与发展。”

评价：差的提示缺乏上下文和方向，导致 AI 模型难以生成相关且连贯的回应。相比之下，好的提示清楚地指定了所需的操作，并提供了必要的上下文以使 AI 模型能够提供有意义的回应。

示例 2：评估连贯性

差的提示：

评价这个：“火车停了，雨大。”

好的提示：

请评估以下文本的连贯性，并指出其中可能存在的语法错误或表达不清的部分：“由于大雨，所有的火车停运了。然后，人们无法出行。”

评价：虽然错误的提示可能会产生回应，但它可能不够详细或不够连贯，无法评估 AI 模型的性能。好的提示提供了更结构化的请求，使评估 AI 模型生成连贯且组织良好的回应的能力变得更加容易。

示例 3：测试准确性

差的提示：

解释光合作用在动物身上是如何进行的。

好的提示：

描述植物的光合作用过程，包括阳光、水和二氧化碳的作用。

评价：坏提示包含事实错误，难以评估 AI 模型提供正确信息的准确性。好的提示会纠正错误并提供有针对性的请求，从而能够更准确地评估 AI 模型的性能。

示例 4：评估偏差

差的提示：

为什么女性对计算机编程不感兴趣？

好的提示：

讨论历史上导致女性在计算机编程中代表性不足的因素，以及这种趋势正在发生怎样的变化。

评价：错误的提示假定有偏见的前提，这可能导致 AI 模型做出有偏见的反应。好的提示将问题重新定义为更加中性和准确，从而可以更有效地评估 AI 模型回应中的潜在偏差。

示例 5：创造力

差的提示：

给我讲个关于外星人的故事。

好的提示：

请为一部儿童科幻小说创作一个引人入胜的故事梗概，故事中的主人公是一位勇敢的少年，他发现了一个神秘的外星文明。

评价：错误的提示过于模糊，没有为 AI 模型提供任何生成创造性回应的方向。好的提示提供了一个具体且富有想象力的上下文，鼓励 AI 模型生成一个独特且引人入胜的故事。

示例 6：效率

差的提示：

谈谈绿色能源。

好的提示：

请简要阐述太阳能发电的优势和局限。

评价：错误的提示过于宽泛，可能导致 AI 模型生成不聚焦的答案，效率较低。好的提示这个提示明确且具有针对性，能引导 AI 模型快速生成相关的答案。

示例 7：歧义

差的提示：

解释一下多态。

好的提示：

请解释计算机编程中的“多态性”这一概念？

评价：不良提示含糊不清且缺乏上下文，导致 AI 模型难以生成相关回应。好的提示提供了具体的细节，减少了歧义，让 AI 模型提供了更准确的答案。

示例 8：不常见的知识

差的提示：

告诉我一些《红楼梦》的小秘密。

好的提示：

请介绍《红楼梦》中贾宝玉所佩戴的通灵宝玉的象征意义和作用。

评价：坏提示模糊且范围过广，可能导致 AI 模型无法准确找到不常见的知识点。好的提示具体而详细，能引导 AI 模型找到关于《红楼梦》中不太常见的知识点。

示例 9：敏感话题

差的提示：

解释具体有争议的政治决定背后的原因。

好的提示：

对特定有争议的政治决定的支持者和反对者提出的因素和论点进行客观概述。

评价：糟糕的提示可能会导致有偏见或有争议的回应。好的提示要求进行平衡和客观的分析，最大限度地减少潜在的偏见，并更准确地评估人工智能模型处理敏感话题的能力。

掌握评估和测试提示的方法对于提示工程专家来说至关重要。通过设计有效的评估指标、识别边缘案例和潜在风险以及对性能进行基准测试，您可以确保您的提示从 GPT 和 ChatGPT 等 AI 语言模型中产生高质量、相关且准确的回应。