# 个性化新闻标题生成评估标准

## 评估目标

衡量模型生成的个性化新闻标题在准确传达新闻核心内容、符合个性化阅读偏好、语言表达清晰流畅等方面的综合质量。

## 参照数据

新闻正文 (P)： 内容的基石和事实来源。  
模型生成的标题 (H\_model)： 主要评估对象。  
标准个性化标题 (H\_ref)： 仅供参考，用于理解可能的个性化角度，但不作为绝对的评判依据。

## 评估维度与评分细则 (总分100分)

### 一、内容准确性与相关性 (Factual Accuracy & Relevance) - 40分

评价依据： H\_model 是否准确、无误地反映了新闻正文P的核心信息和主要事实。

细则：

事实准确 (15分)： H\_model 中包含的事实性信息（如人物、事件、地点、数据等）是否与P一致，无捏造、歪曲或错误。

- 完全准确 (15)

- 轻微不准确，不影响核心理解 (10)

- 关键信息错误或缺失 (5)

- 严重失实或与P无关 (0)

核心信息覆盖 (15分)： H\_model 是否抓住了P最主要、最核心的内容/事件。

- 完美概括核心 (15)

- 基本概括核心，但有遗漏次要核心 (10)

- 仅概括部分内容，未触及核心 (5)

- 与核心内容偏差较大 (0)

避免误导与夸大 (10分)： H\_model 是否客观，没有刻意夸大、耸人听闻或产生歧义，避免“标题党”嫌疑。

- 客观中立，无误导 (10)

- 略有倾向性或轻微夸大，但基本可接受 (7)

- 有明显误导或夸大成分 (3)

- 严重误导，标题党 (0)

### 二、个性化契合度 (Personalization Alignment) - 30分

评价依据： H\_model 是否从特定角度切入，该角度是否能吸引特定用户群体（可参考H\_ref的个性化思路，但H\_model可以有自己的、更优的个性化角度）。

细则：

个性化视角体现 (15分)：

- 个性化视角清晰、合理且新颖/深刻 (15)

- 个性化视角较清晰，与H\_ref相似或有合理创新 (10)

- 个性化视角模糊或不突出，或与H\_ref的个性化点相比不佳 (5)

- 无明显个性化特征，过于泛化 / 个性化角度不当 (0)

潜在用户吸引力 (15分)：

- 极具吸引力，能精准抓住目标用户痛点/兴趣点 (15)

- 有一定吸引力，能引起目标用户关注 (10)

- 吸引力一般，可能不会特别突出 (5)

- 缺乏吸引力，或可能引起反感 (0)

### 三、表达与流畅性 (Fluency & Readability) - 20分

评价依据： H\_model 的语言表达是否清晰、简洁、流畅、符合语法规范，易于理解。

细则：

简洁明了 (7分)：

- 非常简洁，信息密度高 (7)

- 较为简洁，少量冗余 (5)

- 不够简洁，表达拖沓 (2)

- 冗长且信息不清晰 (0)

语言规范与流畅度 (8分)：

- 完全规范流畅 (8)

- 偶有不影响理解的瑕疵 (5)

- 存在明显语病或表达不畅 (2)

- 语义混乱，难以理解 (0)

吸引力与可读性（非标题党）(5分)：

- 语言生动有趣，可读性强 (5)

- 表达平实，可读性尚可 (3)

- 语言枯燥，缺乏吸引力 (1)

### 四、与参考标题对比 (Relative Quality to Reference) - 10分 (可选/附加分)

评价依据： 综合比较H\_model和H\_ref，判断H\_model在个性化表达和整体质量上是否优于、持平或逊于H\_ref（基于对P的理解）。

细则：

- 优于参考标题 (8-10分)

- 持平参考标题 (5-7分)

- 逊于参考标题 (0-4分)

### 使用说明

理解原文是前提： 评估员必须首先仔细阅读新闻正文P，充分理解其核心内容和潜在的个性化切入点。

H\_ref的角色： 将H\_ref视为一个“可能的个性化方向”的提示，而不是“唯一正确答案”。鼓励评估员思考H\_model是否找到了一个同样好或更好的个性化角度。

综合判断： 某些细则可能相互关联，评估员需基于整体理解进行判断。

打分一致性： 若多人评估，需进行前期培训和试评，确保对标准的理解和执行尺度保持一致。

反馈与迭代： 除了分数，鼓励评估员提供具体的文字评价，解释打分原因，这对于模型迭代非常有价值。