# 评价指标定义

由于AI项目的盈利模式尚处于探索阶段，如ChatGPT目前采用的多种收费模式（如按会话次数、回答字数、响应时间、项目尺寸等）无法得到一个稳定的收益率，因此不便采用内部收益率作为评价指标。

**重要因素：**

1.训练成本和部署成本

InstructGPT使用了1750亿的模型参数，当然也因此拥有相对强大的性能，但同时，开源了62 亿参数的 ChatGLM-6B，结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存），虽然规模不及千亿模型，但大大降低了用户部署的门槛，并且已经能生成相当符合人类偏好的回答。此外，绝大多数大语言模型训练成本高昂，导致大部分研究人员都无法负担大语言模型的训练或使用。

2.语言准确率

准确率是分类问题中最简单也是最直观的评价指标，但存在明显的缺陷。比如，当负样本占99%时，分类器把所有样本都预测为负样本也可以获得99%的准确率。所以，当不同类别的样本比例非常不均衡时，占比大的类别往往成为影响准确率的最主要因素。

3.响应速度

响应速度会影响用户的使用满意度，同时也体现了对计算资源的利用能力。不同时优化算法、软件和硬件堆栈，所需的大量计算操作可能会导致不切实际的长训练时间；需要需要在内存和计算上都可扩展的高效并行技术，以充分发挥数千个 GPU 的潜力。

4.响应质量

更大的语言模型虽然有了更强的语言理解和生成的能力，但并不能从本质上使它们更好地遵循或理解用户的指令意图。例如，大型语言模型可能会生成不真实、有害或对用户没有帮助的输出，可以将这些情况归结于低质量的回答。比如，GPT-4 在安全性问题方面做了很大的优化和提升，同时，GPT-4 比 GPT-3.5 更可靠、更有创意，并且能够处理更细微的指令。

5.知名度和用户粘性

虽然AI产品前期的开发以及使用过程中的维护不可避免地需要大量成本且面向用户端的使用成本很低，但是，随着使用AI模型的用户数量的增加，产品应用的单位成本会快速下降，因此AI项目仍具有典型的规模经济和范围经济效应。而目前AI模型市场尚不成熟，增加使用用户数量依赖于知名度和用户粘性。

6.潜在价值

作为大型自然语言处理工具，在金融投资领域的应用前景十分广泛，可以提高各类市场参与主体投资决策的准确性，同时也可以帮助金融机构更好地管理和控制风险，对助力金融业提质增效有一定的助力作用。AI作为新型的自动化技术，能提供更加自动化的生产方案，加速生产过程，提高生产效率，其对数据分析和处理具有天然的优势，可以通过大数据分析提炼经济和金融事实的背后的演化规律，挖掘潜在的创新点。

# 权重计算

使用AHP（层次分析法），由各组员为每一对指标给出成对比较矩阵，平均后计算得到各指标的对应权重。

上面总结出了6个重要因素，因此得到15对比较，将它们依次编号如下：

1. 训练成本和部署成本 vs 语言准确率（1 vs 2）
2. 训练成本和部署成本 vs 响应速度（1 vs 3）
3. 训练成本和部署成本 vs 响应质量（1 vs 4）
4. 训练成本和部署成本 vs 知名度和用户粘性（1 vs 5）
5. 训练成本和部署成本 vs 潜在价值（1 vs 6）
6. 语言准确率 vs 响应速度（2 vs 3）
7. 语言准确率 vs 响应质量（2 vs 4）
8. 语言准确率 vs 知名度和用户粘性（2 vs 5）
9. 语言准确率 vs 潜在价值（2 vs 6）
10. 响应速度 vs 响应质量（3 vs 4）
11. 响应速度 vs 知名度和用户粘性（3 vs 5）
12. 响应速度 vs 潜在价值（3 vs 6）
13. 响应质量 vs 知名度和用户粘性（4 vs 5）
14. 响应质量 vs 潜在价值（4 vs 6）
15. 知名度和用户粘性 vs 潜在价值（5 vs 6）

下为各组员为它们打分得到的成对比较矩阵：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | 1 | 1/5 | 1/2 | 1/3 | 1/4 | 1/5 |
| **2** | 5 | 1 | 5 | 1/3 | 5 | 1/2 |
| **3** | 2 | 1/5 | 1 | 1/6 | 2 | 1/3 |
| **4** | 3 | 3 | 6 | 1 | 7 | 5 |
| **5** | 4 | 1/5 | 1/2 | 1/7 | 1 | 1/5 |
| **6** | 5 | 2 | 3 | 1/5 | 5 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | 1 | 1/6 | 1 | 1/4 | 3 | 1 |
| **2** | 6 | 1 | 2 | 1 | 3 | 1 |
| **3** | 1 | 1/2 | 1 | 2 | 3 | 1/2 |
| **4** | 4 | 1 | 1/2 | 1 | 5 | 1 |
| **5** | 1/3 | 1/3 | 1/3 | 1/5 | 1 | 1/5 |
| **6** | 1 | 1 | 2 | 1 | 5 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | 1 | 1/6 | 1/2 | 1/6 | 1 | 3 |
| **2** | 6 | 1 | 3 | 5 | 9 | 9 |
| **3** | 2 | 1/3 | 1 | 3 | 9 | 5 |
| **4** | 6 | 1/5 | 1/3 | 1 | 5 | 1 |
| **5** | 1 | 1/9 | 1/9 | 1/5 | 1 | 2 |
| **6** | 1/3 | 1/9 | 1/5 | 1 | 1/2 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | 1 | 1/7 | 1/5 | 1/5 | 1/4 | 1/2 |
| **2** | 7 | 1 | 4 | 1 | 5 | 5 |
| **3** | 5 | 1/4 | 1 | 1/5 | 3 | 3 |
| **4** | 5 | 1 | 5 | 1 | 5 | 5 |
| **5** | 4 | 1/5 | 1/3 | 1/5 | 1 | 1 |
| **6** | 2 | 1/5 | 1/3 | 1/5 | 1 | 1 |

将这四个矩阵平均后，归一化最大特征值对应的特征向量，得到六个指标的权重。该过程使用的Python脚本如下：

import numpy as np

from scipy.linalg import eig

*# 成对比较矩阵*

A = np.array(...)

B = np.array(...)

C = np.array(...)

D = np.array(...)

*# 计算平均值*

X = (A + B + C + D) / 4

*# 计算特征向量和特征值*

eigenvalues, eigenvectors = eig(X)

*# 找到最大的特征值及其对应的特征向量*

max\_eigenvalue\_index = np.argmax(eigenvalues)

max\_eigenvalue\_vector = eigenvectors[:, max\_eigenvalue\_index]

*# 归一化特征向量，得到权重向量*

weights = max\_eigenvalue\_vector / np.sum(max\_eigenvalue\_vector)

print(f'Weights: {weights}')

得到的最终权重如下：



1.训练成本和部署成本（6.04%）

2.语言准确率（31.82%）

3.响应速度（16.27%）

4.响应质量（26.63%）

5.知名度和用户粘性（5.94%）

6.潜在价值（13.29%）

# 评价指标打分

我们挑选当前比较突出的四个大语言模型：OpenAI的ChatGPT、谷歌的Bard、百度的文心一言及阿里巴巴的通义千问，针对这六个评价指标的满足程度分别进行打分（1-10分）。

下为各组员的打分情况：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ChatGPT** | **Bard** | **文心一言** | **通义千问** |
| 训练成本和部署成本 | 3 | 5 | 6 | 9 |
| 语言准确率 | 10 | 6 | 7 | 5 |
| 响应速度 | 6 | 9 | 8 | 10 |
| 响应质量 | 9 | 6 | 7 | 6 |
| 知名度和用户粘性 | 10 | 8 | 8 | 6 |
| 潜在价值 | 10 | 8 | 7 | 6 |
| 加权得分 | 8.65992854 | 6.81244594 | 7.16171734 | 6.51386616 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ChatGPT** | **Bard** | **文心一言** | **通义千问** |
| 训练成本和部署成本 | 2 | 5 | 6 | 8 |
| 语言准确率 | 10 | 6 | 7 | 7 |
| 响应速度 | 8 | 8 | 8 | 9 |
| 响应质量 | 9 | 8 | 7 | 6 |
| 知名度和用户粘性 | 10 | 7 | 8 | 5 |
| 潜在价值 | 10 | 6 | 7 | 6 |
| 加权得分 | 8.92489908 | 6.85691842 | 7.16171734 | 6.86774778 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ChatGPT** | **Bard** | **文心一言** | **通义千问** |
| 训练成本和部署成本 | 2 | 5 | 7 | 9 |
| 语言准确率 | 10 | 6 | 6 | 5 |
| 响应速度 | 6 | 8 | 8 | 10 |
| 响应质量 | 10 | 6 | 6 | 6 |
| 知名度和用户粘性 | 10 | 5 | 6 | 5 |
| 潜在价值 | 10 | 8 | 6 | 5 |
| 加权得分 | 8.86574752 | 6.47140218 | 6.38583282 | 6.32147888 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ChatGPT** | **Bard** | **文心一言** | **通义千问** |
| 训练成本和部署成本 | 3 | 5 | 6 | 9 |
| 语言准确率 | 10 | 6 | 6 | 6 |
| 响应速度 | 6 | 7 | 8 | 9 |
| 响应质量 | 9 | 6 | 6 | 6 |
| 知名度和用户粘性 | 10 | 4 | 5 | 4 |
| 潜在价值 | 9 | 7 | 6 | 5 |
| 加权得分 | 8.5269889 | 6.11631406 | 6.26595404 | 6.41756102 |