6.大模型评估

1.人工评估的一般思路

在系统开发的初期，验证集体量较小，最简单、直观的方法即为人工对验证集中的每一个验证案例进行评估。但是，人工评估也有一些基本准则与思路，此处简要介绍供学习者参考。但请注意，系统的评估与业务强相关，设计具体的评估方法与维度需要结合具体业务深入考虑。

1.1 量化评估

为保证很好地比较不同版本的系统性能，量化评估指标是非常必要的。

应该对每一个验证案例的回答都给出打分，最后计算所有验证案例的平均分得到本版本系统的得分。量化的量纲可以是 05，也可以是 0100，可以根据个人风格和业务实际情况而定。

1.2 多维评估

大模型是典型的生成模型，即其回答为一个由模型生成的语句。一般而言，大模型的回答需要在多个维度上进行评估。例如，本项目的个人知识库问答项目上，用户提问一般是针对个人知识库的内容进行提问，模型的回答需要同时满足充分使用个人知识库内容、答案与问题一致、答案真实有效、回答语句通顺等。一个优秀的问答助手，应当既能够很好地回答用户的问题，保证答案的正确性，又能够体现出充分的智能性。

因此，往往需要从多个维度出发，设计每个维度的评估指标，在每个维度上都进行打分，从而综合评估系统性能。同时需要注意的是，多维评估应当和量化评估有效结合，对每一个维度，可以设置相同的量纲也可以设置不同的量纲，应充分结合业务实际。

例如，在本项目中，可以设计如下几个维度的评估：

1. **知识查找正确性**。该维度需要查看系统从向量数据库查找相关知识片段的中间结果，评估系统查找到的知识片段是否能够对问题做出回答。该维度为 0–1 评估，

即打分为 0 指查找到的知识片段不能做出回答，打分为 1 指查找到的知识片段可以做出回答。

2. **回答一致性**。该维度评估系统的回答是否针对用户问题展开，是否有偏题、错误理解题意的情况，该维度量纲同样设计为 0~1，0 为完全偏题，1 为完全切题，中间结果可以任取。
3. **回答幻觉比例**。该维度需要综合系统回答与查找到的知识片段，评估系统的回答是否出现幻觉，幻觉比例有多高。该维度同样设计为 0~1,0 为全部是模型幻觉，1 为没有任何幻觉。
4. **回答正确性**。该维度评估系统回答是否正确，是否充分解答了用户问题，是系统最核心的评估指标之一。该维度可以在 0~1 之间任意打分。

上述四个维度都围绕知识、回答的正确性展开，与问题高度相关；接下来几个维度将围绕大模型生成结果的拟人性、语法正确性展开，与问题相关性较小：

5. **逻辑性**。该维度评估系统回答是否逻辑连贯，是否出现前后冲突、逻辑混乱的情况。该维度为 0~1 评估。
6. **通顺性**。该维度评估系统回答是否通顺、合乎语法，可以在 0~1 之间任意打分。
7. **智能性**。该维度评估系统回答是否拟人化、智能化，是否能充分让用户混淆人工回答与智能回答。该维度可以在 0~1 之间任意打分。

2.简单的自动评估

大模型评估之所以复杂，**一个重要原因在于生成模型的答案很难判别，即客观题评估判别很简单，主观题评估判别则很困难**。尤其是对于一些没有标准答案的问题，实现自动评估就显得难度尤大。但是，在牺牲一定评估准确性的情况下，**可以将复杂的没有标准答案的主观题进行转化，从而变成有标准答案的问题**，进而通过简单的自动评估来实现。此处介绍两种方法：构造客观题与计算标准答案相似度。

2.1 构造客观题

主观题的评估是非常困难的，但是客观题可以直接对比系统答案与标准答案是否一致，从而实现简单评估。可以**将部分主观题构造为多项或单项选择的客观题，进而实现简单评估**。

2.2 计算相似度

生成问题的答案评估在 NLP 中实则也不是一个新问题了，不管是机器翻译、自动文摘等任务，其实都需要评估生成答案的质量。

NLP 一般对生成问题采用**人工构造标准答案并计算回答与标准答案相似度的方法来实现自动评估**。

计算相似度的方法有很多，一般可以使用 BLEU 来计算相似度，可以简单理解为主题相似度。

但是，该种方法同样存在几个问题：

1. **需要人工构造标准答案**。对于一些垂直领域而言，构造标准答案可能是一件困难的事情；
2. **通过相似度来评估，可能存在问题**。例如，如果生成回答与标准答案高度一致但在核心的几个地方恰恰相反导致答案完全错误，bleu 得分仍然会很高；
3. 通过计算与标准答案**一致性灵活性很差**，如果模型生成了比标准答案更好的回答，但评估得分反而会降低；
4. **无法评估回答的智能性、流畅性**。如果回答是各个标准答案中的关键词拼接出来的，我们认为这样的回答是不可用无法理解的，但 bleu 得分会较高。

因此，针对业务情况，有时还需要一些不需要构造标准答案的、进阶的评估方法。

3.使用大模型评估

使用人工评估准确度高、全面性强，但人力成本与时间成本高；使用自动评估成本低、评估速度快，但存在准确性不足、评估不够全面的问题。那么，是否有一种方法综合两者的优点，实现快速、全面的生成问题评估呢?

以 GPT-4 为代表的大模型提供了一种新的方法：**使用大模型进行评估**。可以通过构造 Prompt Engineering 让大模型充当一个评估者的角色，从而替代人工评估的评估员；同时大模型可以给出类似于人工评估的结果，因此可以采取人工评估中的多维度量化评估的方式，实现快速全面的评估。

但是注意，使用大模型进行评估仍然存在问题：

1. 目标是迭代改进 Prompt 以提升大模型表现，因此**所选用的评估大模型需要有优于所使用的大模型基座的性能**，例如，目前性能最强大的大模型仍然是 GPT-4，推荐使用 GPT-4 来进行评估，效果最好。

2. 大模型具有强大的能力，但**同样存在能力的边界**。如果问题与回答太复杂、知识片段太长或是要求评估维度太多，即使是 GPT-4 也会出现错误评估、错误格式、无法理解指令等情况，针对这些情况，建议考虑如下方案来提升大模型表现：

    a. **改进 Prompt Engineering**。以类似于系统本身 Prompt Engineering 改进的方式，迭代优化评估 Prompt Engineering，尤其是注意是否遵守了 Prompt Engineering 的基本准则、核心建议等；

    b. **拆分评估维度**。如果评估维度太多，模型可能会出现错误格式导致返回无法解析，可以考虑将待评估的多个维度拆分，每个维度调用一次大模型进行评估，最后得到统一结果；

    c. **合并评估维度**。如果评估维度太细，模型可能无法正确理解以至于评估不正确，可以考虑将待评估的多个维度合并，例如，将逻辑性、通顺性、智能性合并为智能性等；

    d. **提供详细的评估规范**。如果没有评估规范，模型很难给出理想的评估结果。可以考虑给出详细、具体的评估规范，从而提升模型的评估能力；

    e. **提供少量示例**。模型可能难以理解评估规范，此时可以给出少量评估的示例，供模型参考以实现正确评估。

4.混合评估

事实上，上述评估方法都不是孤立、对立的，相较于独立地使用某一种评估方法，**更推荐将多种评估方法混合起来**，对于每一种维度选取其适合的评估方法，兼顾评估的全面、准确和高效。

例如，针对本项目个人知识库助手，可以设计以下混合评估方法：

1. 客观正确性。客观正确性指对于一些有固定正确答案的问题，模型可以给出正确的回答。可以选取部分案例，使用构造客观题的方式来进行模型评估，评估其客观正确性。
2. 主观正确性。主观正确性指对于没有固定正确答案的主观问题，模型可以给出正确的、全面的回答。可以选取部分案例，使用大模型评估的方式来评估模型回答是否正确。

3. 智能性。智能性指模型的回答是否足够拟人化。由于智能性与问题本身弱相关，与模型、Prompt 强相关，且模型判断智能性能力较弱，我们可以少量抽样进行人工评估其智能性。

4. 知识查找正确性。知识查找正确性指对于特定问题，从知识库检索到的知识片段是否正确、是否足够回答问题。知识查找正确性推荐使用大模型进行评估，即要求模型判别给定的知识片段是否足够回答问题。同时，该维度评估结果结合主观正确性可以计算幻觉情况，即如果主观回答正确但知识查找不正确，则说明产生了模型幻觉。

使用上述评估方法，基于已得到的验证集示例，可以对项目做出合理评估。限于时间与人力，此处就不具体展示了。