

#### 4. 大语言模型评估指标体系

通过本节的前述内容，可以看到传统的自然语言处理评估大多针对单一任务设置不同的评估指标和方法。大语言模型在经过指令微调和强化学习阶段后，可以完成非常多不同种类的任务，对于常见的自然语言理解或生成任务可以采用原有指标体系。虽然大语言模型在文本生成类任务上取得了突破性的进展，但是问题回答、文章生成、开放对话等文本生成类任务在此前并没有很好的评估指标，因此，针对大语言模型在文本生成方面的能力，需要考虑建立新的评估指标体系。为了更全面地评估大语言模型所生成的文本的质量，需要从三方面进行评估，包括语言层面、语义层面和知识层面。

(1) 语言层面的评估是评估大语言模型所生成文本质量的基础，要求生成的文本必须符合人类的语言习惯。这意味着生成的文本必须具有正确的词法、语法和篇章结构。具体如下：

- **词法正确性：**评估生成文本中单词的拼写、使用和形态变化是否正确。确保单词拼写准确无误，不含有拼写错误。同时，评估单词的使用是否恰当，包括单词的含义、词性和用法等方面，以确保单词在上下文中被正确应用。此外，还需要关注单词的形态变化是否符合语法规则，包括时态、数和派生等方面。
- **语法正确性：**评估生成文本的句子结构和语法规则是否正确。确保句子的构造完整，各个语法成分之间的关系符合语法规则，包括主谓关系、动宾关系、定状补关系等方面的应用。此外，还需要评估动词的时态是否使用正确，包括时态的一致性和选择是否符合语境。
- **篇章结构正确性：**评估生成文本的整体结构是否合理。确保文本段落之间连贯，文本信息流畅自然，包括使用恰当的主题句、过渡句和连接词等。同时，需要评估文本整体结构的合理性，包括标题、段落、章节等结构的使用是否恰当，以及文本整体框架是否清晰明了。

(2) 语义层面的评估主要关注文本的语义准确性、逻辑连贯性和风格一致性。要求生成的文本不出现语义错误或误导性描述，并且具有清晰的逻辑结构，能够按照一定的顺序和方式呈现出来。具体如下：

- **语义准确性：**评估文本是否传达了准确的语义信息。包括词语的确切含义和用法是否正确，以及句子表达的意思是否与作者的意图相符。确保文本中使用的术语、概念和描述准确无误，能够准确传达信息给读者。
- **逻辑连贯性：**评估文本的逻辑结构是否连贯一致。句子之间应该有明确的逻辑关系，能够形成有条理的论述，文本中的论证、推理、归纳、演绎等逻辑关系应该正确。句子的顺序应符合常规的时间、空间或因果关系，以便读者能够理解句子之间的联系。
- **风格一致性：**评估文本在整体风格上是否保持一致。包括词汇选择、句子结构、表达方式等方面。文本应该在整体上保持一种风格或口吻。例如，正式文本应使用正式的语言和术语，而故事性的文本可以使用生动的描写和故事情节。

(3) 知识层面的评估主要关注知识准确性、知识丰富性和知识一致性。要求生成文本所涉及的知识准确无误、丰富全面，确保文本的可信度。具体如下：

- **知识准确性**：评估生成文本中所呈现的知识是否准确无误。这涉及事实陈述、概念解释、历史事件描述等方面。生成的文本应基于准确的知识和可靠的信息源，避免错误、虚假或误导性的内容。确保所提供的知识准确无误。
- **知识丰富性**：评估生成文本所包含的知识是否丰富多样。生成的文本应能够提供充分的信息，涵盖相关领域的不同方面。这可以通过提供具体的例子、详细的解释和相关的背景知识来实现。确保生成文本在知识上具有广度和深度，能够满足读者的需求。
- **知识一致性**：评估生成文本中知识的一致性。这包括确保文本中不出现相互矛盾的知识陈述，避免在不同部分或句子中提供相互冲突的信息。生成的文本应该在整体上保持一致，使读者能够得到一致的知识体系。

### 11.3.2 评估方法

评估方法的目标是解决如何对大语言模型生成结果进行评估的问题。有些指标可以通过比较正确答案或参考答案与系统生成结果直接计算得出，例如准确率、召回率等。这种方法被称为自动评估（Automatic Evaluation）。然而，有些指标并不是可以直接计算出来的，而需要通过人工评估得出。例如，对一篇文章的质量进行评估，虽然可以使用自动评估的方法计算出一些指标，如拼写错误的数量、语法错误的数量等，但是对于文章的流畅性、逻辑性、观点表达等方面的评估则需要人工阅读并进行分项打分。这种方法被称为人工评估（Human Evaluation）。人工评估是一种耗时耗力的评估方法，因此研究人员提出了一种新的评估方法，即利用能力较强的大语言模型（如 GPT-4），构建合适的指令来评估系统结果<sup>[196, 598–601]</sup>。这种评估方法可以大幅度减少人工评估所需的时间和人力成本，具有更高的效率。这种方法被称为大语言模型评估（LLM Evaluation）。此外，有时我们还希望对比不同系统之间或者系统不同版本之间的差别，这需要采用对比评估（Comparative Evaluation）方法针对系统之间的不同进行量化。自动评估在前面介绍评估指标时已经给出了对应的计算方法和公式，本节将分别针对人工评估、大语言模型评估和对比评估进行介绍。

#### 1. 人工评估

人工评估是一种广泛应用于评估模型生成结果质量和准确性的方法，它通过人类参与对生成结果进行综合评估。与自动化评估方法相比，人工评估更接近实际应用场景，并且可以提供更全面和准确的反馈。在人工评估中，评估者可以对大语言模型生成结果的整体质量进行评分，也可以根据评估体系从语言层面、语义层面及知识层面等不同方面进行细粒度评分。此外，人工评估还可以对不同系统之间的优劣进行对比评分，从而为模型的改进提供有力的支持。然而，人工评估也存在一些限制和挑战。首先，由于人的主观性和认知差异，评估结果可能存在一定程度的主观性。其次，人工评估需要大量的时间、精力和资源，因此成本较高，且评估周期长，不能及时得到有效的反馈。此外，评估者的数量和质量也会对评估结果产生影响。

人工评估是一种常用于评估自然语言处理系统性能的方法。通常涉及五个层面：评估者类型、评估指标度量、是否给定参考和上下文、绝对还是相对评估，以及评估者是否提供解释。

(1) 评估者类型是指评估任务由哪些人来完成。常见的评估者包括领域专家、众包工作者和最终使用者。领域专家对于特定领域的任务具有专业知识和经验，可以提供高质量的评估结果。众包工作者通常是通过在线平台招募的大量非专业人员，可以快速地完成大规模的评估任务。最终使用者是指系统的最终用户，他们的反馈可以帮助开发者了解系统在实际使用中的表现情况。

(2) 评估指标度量是指根据评估指标所设计的具体度量方法。常用的评估度量有李克特量表(Likert Scale)，它为生成结果提供不同的标准，分为几个不同等级，可用于评估系统的语言流畅度、语法准确性、结果完整性等。

(3) 是否给定参考和上下文是指提供与输入相关的上下文或参考，这有助于评估语言流畅度、语法以外的性质，比如结果的完整性和正确性。非专业人员很难仅通过输出结果判断流畅性以外的其他性能，因此给定参考和上下文可以帮助评估者更好地理解和评估系统性能。

(4) 绝对还是相对评估是指将系统输出与参考答案进行比较，还是与其他系统进行比较。绝对评估是指将系统输出与单一参考答案进行比较，可以评估系统各维度的能力。相对评估是指同时对多个系统输出进行比较，可以评估不同系统之间的性能差异。

(5) 评估者是否提供解释是指是否要求评估者为自己的决策提供必要的说明。提供决策的解释有助于开发者了解评估过程中的决策依据和评估结果的可靠性，从而更好地优化系统性能，但缺点是极大地增加了评估者的时间花费。

对于每个数据，通常会有多个不同人员进行评估，因此需要一定方法整合最终评分。最简单的最终评分整合方法是计算平均主观得分 (Mean Opinion Score, MOS)，即对所有评估者的评分求平均值：

$$\text{MOS} = \frac{1}{N} \sum_{i=1}^N (S_i) \quad (11.14)$$

其中， $N$  为评估者人数， $S_i$  为第  $i$  个评估者给出的评分。此外，还可以采用以下方法。

(1) 中位数法：将所有分数按大小排列，取中间的分数作为综合分数，中位数可以避免极端值对综合分数的影响，因此在数据分布不均匀时比平均值更有用。

(2) 最佳分数法：选择多个分数中的最高分数作为综合分数。这种方法在评估中强调最佳性能，并且在只需要比较最佳结果时非常有用。

(3) 多数表决法：将多个分数中出现次数最多的分数作为综合分数。这种方法适用于分类任务，其中每个分数代表一个类别。

由于数据由多个不同评估者进行标注，因此不同评估者之间评估的一致性也是需要关注的因素。一方面，评估者之间的分歧可以作为一种反馈机制，帮助评估文本生成的效果和任务定义。评估者高度统一的结果意味着任务和评估指标都具有良好的定义。另一方面，评估者之间的一致性可以用于判断评估者的标注质量。如果某个评估者在大多数情况下都与其他评估者意见不一致，那么在一定程度上可以说明该评估者的标注需要重点关注。评估者间一致性 (Inter-Annotator Agreement, IAA) 是评估不同评估者之间达成一致的程度的度量。一些常用的 IAA 度量标准包括一致性百分比、Cohen's Kappa、Fleiss' Kappa 等。这些度量标准计算不同评估者之间的一致性得分，并将其转换为 0 到 1 之间的值。得分越高，表示评估者之间的一致性越好。

- **一致性百分比** (Percent Agreement) 用以判定所有评估者一致同意的程度。 $X$  表示待评估的文本， $|X|$  表示文本的数量， $a_i$  表示所有评估者对  $x_i$  的评估结果的一致性，当所有评估者的评估结果一致时， $a_i = 1$ ，否则等于 0。一致性百分比可以形式化表示为

$$P_a = \frac{\sum_{i=0}^{|X|} a_i}{|X|} \quad (11.15)$$

- **Cohen's Kappa** 是一种用于度量两个评估者之间一致性的统计量。Cohen's Kappa 的值在 -1 到 1 之间，其中 1 表示完全一致，0 表示随机一致，而 -1 表示完全不一致。通常，Cohen's Kappa 的值在 0 到 1 之间。具体来说，Cohen's Kappa 的计算公式为

$$\kappa = \frac{P_a - P_c}{1 - P_c} \quad (11.16)$$

$$P_c = \sum_{s \in S} P(s|e_1) \times P(s|e_2) \quad (11.17)$$

其中， $e_1$  和  $e_2$  表示两个评估者， $S$  表示对数据集  $X$  的评分集合， $P(s|e_i)$  表示评估者  $i$  给出分数  $s$  的频率估计。一般来说，Cohen's Kappa 值在 0.6 以上被认为一致性较好，而在 0.4 以

下则被认为一致性较差。

- **Fleiss' Kappa**是一种用于度量三个或三个以上评估者之间一致性的统计量，与 Cohen's Kappa 只能用于两个评估者之间的一致性度量不同，它是 Cohen's Kappa 的扩展版本。Fleiss' Kappa 的值也在  $-1$  到  $1$  之间，其中  $1$  表示完全一致， $0$  表示随机一致，而  $-1$  表示完全不一致。具体来说，Fleiss' Kappa 的计算与式(8.26)相同，但是其  $P_a$  和  $P_e$  的计算则需要扩展为三个或三个以上评估者的情况。使用  $X$  表示待评估的文本， $|X|$  表示文本总数， $n$  表示评估者数量， $k$  表示评估类别数。文本使用  $i = 1, 2, \dots, |X|$  进行编号，打分类别使用  $j = 1, 2, \dots, k$  进行编号，则  $n_{ij}$  表示有多少个评估者对第  $i$  个文本给出了第  $j$  类评估意见。 $P_a$  和  $P_e$  可以形式化表示为

$$P_a = \frac{1}{|X|n(n-1)} \left( \sum_{i=1}^{|X|} \sum_{j=1}^k n_{ij}^2 - |X|n \right) \quad (11.18)$$

$$P_e = \sum_{j=1}^k \left( \frac{1}{|X|n} \sum_{i=1}^{|X|} n_{ij} \right)^2 \quad (11.19)$$

在使用 Fleiss' Kappa 时，需要先确定评估者之间的分类标准，并且需要有足够的数据进行评估。一般来说，与 Cohen's Kappa 一样，Cohen's Kappa 值在  $0.6$  以上被认为一致性较好，而在  $0.4$  以下则被认为一致性较差。需要注意的是，Fleiss' Kappa 在评估者数量较少时可能不太稳定，因此在使用之前需要仔细考虑评估者数量的影响。

## 2. 大语言模型评估

人工评估大语言模型生成内容需要花费大量的时间和资源，成本很高且评估周期非常长，不能及时得到有效的反馈。传统的基于参考文本的度量指标，如 BLEU 和 ROUGE，与人工评估之间的相关性不足，对于需要创造性和多样性的任务也无法提供有效的参考文本。为了解决上述问题，最近的一些研究提出可以采用大语言模型进行自然语言生成任务的评估。而且这种方法还可以应用于缺乏参考文本的任务。使用大语言模型进行结果评估的过程如图11.9 所示。

使用大语言模型进行评估的过程比较简单，例如针对文本质量判断问题，要构造任务说明、待评估样本及对大语言模型的指令，将上述内容输入大语言模型，对给定的待评估样本质量进行评估，图 8.11 给出的指令要求大语言模型采用 5 级李克特量表法。给定这些输入，大语言模型将通过生成一些输出句子来回答问题。通过解析输出句子以获取评分。不同的任务使用不同的任务说明集合，并且每个任务使用不同的问题来评估样本的质量。在文献 [600] 中，针对故事生成任务的文本质量又细分为 4 个属性。

- (1) 语法正确性：故事片段文本的语法正确程度。
- (2) 连贯性：故事片段中句子之间的衔接连贯程度。
- (3) 喜好度：故事片段令人愉悦的程度。
- (4) 相关性：故事片段是否符合给定的要求。

为了与人工评估进行对比，研究人员将输入大语言模型的文本内容，同样给到一些评估者进行人工评估。在开放式故事生成和对抗性攻击两个任务上的实验结果表明，大语言模型评估的结果与人工评估得到的结果一致性较高。同时他们也发现，在使用不同的任务说明格式和生成答案采样算法的情况下，大语言模型的评估结果也是稳定的。

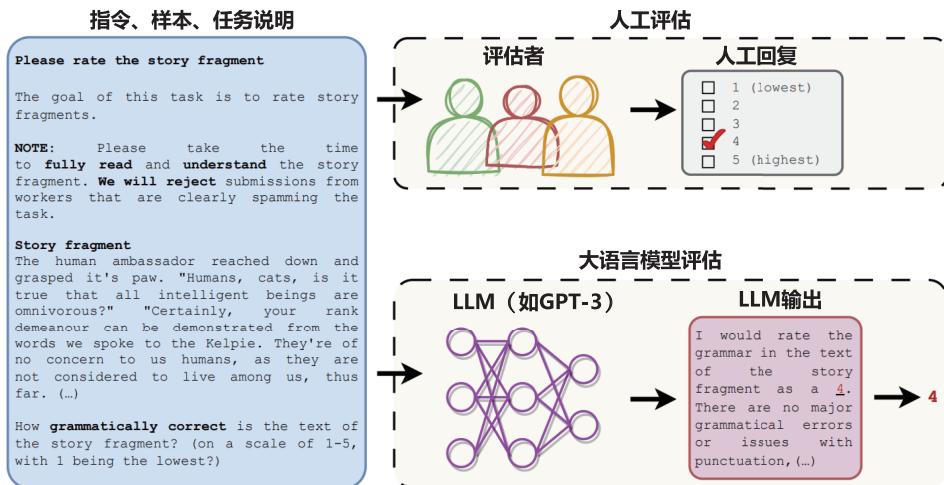


图 11.9 使用大语言模型进行结果评估的过程<sup>[600]</sup>

### 3. 对比评估

对比评估的目标是比较不同系统、方法或算法在特定任务上是否存在显著差异。麦克尼马尔检验（McNemar Test）<sup>[602]</sup>是由 Quinn McNemar 于 1947 年提出的一种用于成对比较的非参数统计检验方法，可用于比较两个机器学习分类器的性能。麦克尼马尔检验也被称为“被试内卡方检验”(within-subjects chi-squared test)，它基于  $2 \times 2$  混淆矩阵 (Confusion Matrix)，有时也称为  $2 \times 2$  列联表 (Contingency Table)，用于比较两个模型之间的预测结果。

给定如图11.10所示的用于麦克尼马尔检验的混淆矩阵，可以得到模型 1 的准确率为  $\frac{A+B}{A+B+C+D}$ ，其中  $A+B+C+D$  为整个测试集中的样本数  $n$ 。同样地，也可以得到模型 2 的准确率为  $\frac{A+C}{A+B+C+D}$ 。这个矩阵中最最重要的数字是  $B$  和  $C$ ，因为  $A$  和  $D$  表示了模型 1 和模型 2 都进行正确或错误预测的样本数。 $B$  和  $C$  则反映了两个模型之间的差异。

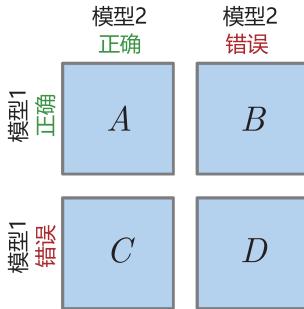
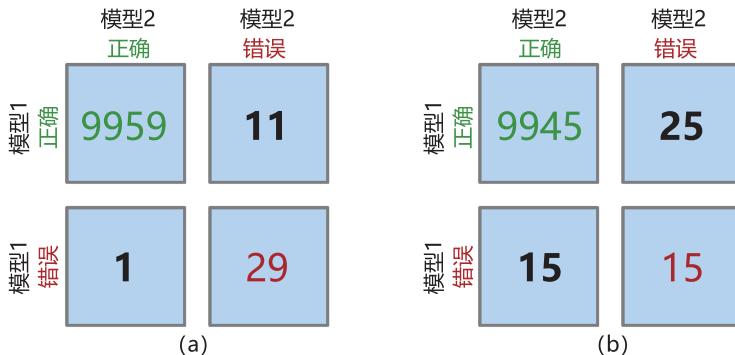
图 11.10 用于麦克尼马尔检验的混淆矩阵<sup>[603]</sup>

图11.11给出了两个样例，根据图11.11(a)和图11.11(b)，可以计算得到模型1和模型2在两种情况下的准确率分别为99.7%和99.6%。根据图11.11(a)，可以看到模型1回答正确且模型2回答错误的数量为11，但是反过来模型2回答正确且模型1回答错误的数量仅为1。在图11.11(b)中，这两个数字变成了25和15。显然，图11.11(b)中的模型1与模型2之间的差异更大，图11.11(a)中的模型1与模型2之间的差异则没有这么明显。

图 11.11 麦克尼马尔检验样例<sup>[603]</sup>

为了量化表示上述情况，麦克尼马尔检验中提出的零假设是概率  $p(B)$  与  $p(C)$  相等，即两个模型都没有表现得比另一个好。麦克尼马尔检验的统计量（“卡方值”）计算公式如下：

$$\chi^2 = \frac{(B - C)^2}{B + C} \quad (11.20)$$

设定显著性水平阈值（例如  $\alpha = 0.05$ ）之后，可以计算得到  $p$ -value ( $p$  值)。如果零假设为真，则  $p$  值是观察这个经验（或更大的）卡方值的概率。如果  $p$  值小于预先设置的显著性水平阈值，则可以拒绝两个模型性能相等的零假设。换句话说，如果  $p$  值小于显著性水平阈值，则可以认为两个模型的性能不同。

文献 [604] 在上述公式的基础上，提出了一个连续性修正版本，这也是目前更常用的变体：

$$\chi^2 = \frac{(|B - C| - 1)^2}{B + C} \quad (11.21)$$

当  $B$  和  $C$  的值大于 50 时，麦克尼马尔检验可以相对准确地近似计算  $p$  值，如果  $B$  和  $C$  的值相对较小 ( $B + C < 25$ )，则建议使用以下二项式检验公式计算  $p$  值：

$$p = 2 \sum_{i=B}^n \binom{n}{i} 0.5^i (1 - 0.5)^{n-i} \quad (11.22)$$

其中  $n = B + C$ ，因子 2 用于计算双侧  $p$  值 (Two-sided  $p$ -value)。

针对图11.11 中的两种情况，可以使用 mlxtend<sup>[555]</sup> 来计算  $p$  值和  $\chi^2$ ：

```
from mlxtend.evaluate import mcnemar
import numpy as np

tb_a = np.array([[9959, 11],
                 [1, 29]])

chi2, p = mcnemar(ary=tb_a, exact=True)

print('chi-squared-a:', chi2)
print('p-value-a:', p)

tb_b = np.array([[9945, 25],
                 [15, 15]])

chi2, p = mcnemar(ary=tb_b, exact=True)

print('chi-squared-b:', chi2)
print('p-value-b:', p)
```

可以得到如下输出：

```
chi-squared-a: None  
p-value-a: 0.005859375
```

```
chi-squared-b: 2.025  
p-value-b: 0.154728923485
```

通常，设置显著性水平阈值  $\alpha = 0.05$ ，因此，根据上述计算结果可以得到结论：图11.11(a) 中两个模型之间的差异不显著。

## 11.4 大语言模型评估实践

大语言模型的评估伴随着大语言模型研究同步飞速发展，大量针对不同任务、采用不同指标和方法的大语言模型评估不断涌现。本章前面几节分别针对大语言模型评估体系、评估指标和评估方法从不同方面介绍了当前大语言模型评估面临的问题，试图回答要从哪些方面评估大语言模型，以及如何评估大语言模型这两个核心问题。针对大语言模型构建不同阶段所产生的模型能力的不同，本节将分别介绍当前常见的针对基础模型、SFT 模型和 RL 模型的整体评估方案。

### 11.4.1 基础模型评估

大语言模型构建过程中产生的基础模型就是语言模型，其目标就是建模自然语言的概率分布。语言模型构建了长文本的建模能力，使得模型可以根据输入的提示词生成文本补全句子。2020 年 OpenAI 的研究人员在 1750 亿个参数的 GPT-3 模型上研究发现，在语境学习范式下，大语言模型可以根据少量给定的数据，在不调整模型参数的情况下，在很多自然语言处理任务上取得不错的效果<sup>[13]</sup>。图11.12 展示了不同参数量的大语言模型在简单任务中基于语境学习的表现。这个任务要求模型从一个单词中去除随机符号，包括使用和不使用自然语言提示词的情况。可以看到，大语言模型具有更好的从上下文信息中学习任务的能力。在此之后，大语言模型评估也不再局限于困惑度、交叉熵等传统评估指标，而更多采用综合自然语言处理任务集合的方式进行评估。

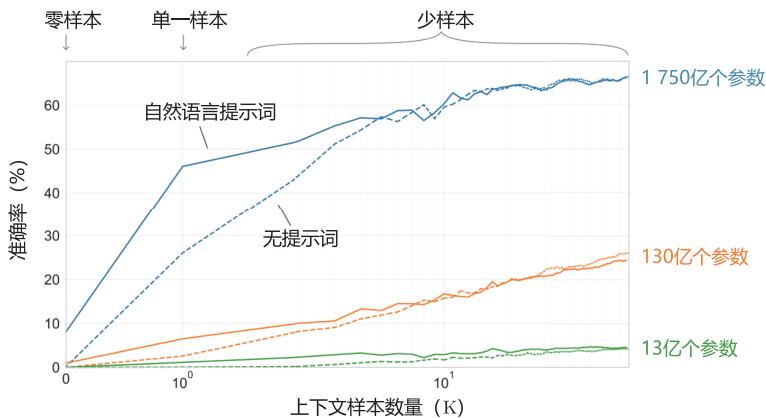


图 11.12 不同参数量的大语言模型在简单任务中基于语境学习的表现<sup>[13]</sup>

#### 1. GPT-3 评估

OpenAI 的研究人员针对 GPT-3<sup>[13]</sup> 的评估主要包含两个部分：传统语言模型评估及综合任务评估。在传统语言模型评估方面，采用了基于 Penn Tree Bank (PTB)<sup>[605]</sup> 数据集的困惑度评估；Lambada<sup>[142]</sup> 数据集用于评估长距离语言建模能力，补全句子的最后一个单词；HellaSwag<sup>[606]</sup> 数据集要求模型根据故事内容或一系列说明选择最佳结局；StoryCloze<sup>[607]</sup> 数据集也用于评估模型根

据故事内容选择结尾句子的能力。在综合任务评估方面，GPT-3 评估引入了 Natural Questions<sup>[459]</sup>、WebQuestions<sup>[608]</sup> 及 TriviaQA<sup>[609]</sup> 三种闭卷问答（Closed Book Question Answering）任务，英语、法语、德语及俄语之间的翻译任务，基于 Winograd Schemas Challenge<sup>[610]</sup> 数据集的指代消解任务，PhysicalQA（PIQA）<sup>[611]</sup>、ARC<sup>[442]</sup>、OpenBookQA<sup>[443]</sup> 等常识推理数据集，CoQA<sup>[612]</sup>、SQuAD2.0<sup>[613]</sup>、RACE<sup>[614]</sup> 等阅读理解数据集，SuperGLUE<sup>[458]</sup> 自然语言处理综合评估集、Natural Language Inference（NLI）<sup>[615]</sup> 和 Adversarial Natural Language Inference（ANLI）<sup>[616]</sup> 自然语言推理任务集，以及包括数字加减、四则运算、单词操作、单词类比、新文章生成等的综合任务。

由于大语言模型在训练阶段需要使用大量种类繁杂且来源多样的训练数据，因此不可避免地存在数据泄露的问题，即测试数据出现在语言模型训练数据中。为了避免这个因素的干扰，OpenAI 的研究人员对于每个基准测试，会生成一个“干净”版本，该版本会移除所有可能泄露的样本。泄露样本的定义大致为与预训练集中任何 13-gram 重叠的样本（或者当样本长度小于 13-gram 时，与整个样本重叠）。目标是非常保守地标记任何可能存在污染的内容，以便生成一个高度可信且无污染的干净子集。之后，使用干净子集对 GPT-3 进行评估，并将其与原始得分进行比较。如果干净子集上的得分与整个数据集上的得分相似，则表明即使存在污染也不会对结果产生显著影响。如果干净子集上的得分较低，则表明污染可能会提升评估结果。GPT-3 数据泄露的影响评估如图 11.13 所示。 $x$  轴表示数据集中有多少数据可以被高度自信地认为是干净的，而  $y$  轴显示了在干净子集上进行评估时性能的差异。可以看到，虽然污染水平通常很高，有四分之一的基准测试超过 50%，但在大多数情况下，性能变化很小。

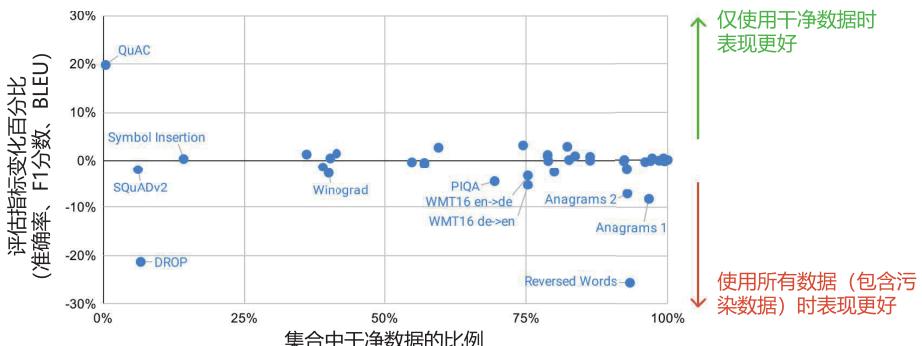


图 11.13 GPT-3 数据泄露的影响评估<sup>[13]</sup>

## 2. MMLU 基准测试

MMLU（Massive Multitask Language Understanding）<sup>[573]</sup> 基准测试的目标是了解大语言模型在预训练期间获取的知识。与此前的评估大多聚焦于自然语言处理相关任务不同，MMLU 基准测试涵盖了 STEM、人文、社会科学等领域的 57 个主题。它的难度范围从小学到高级专业水平不等，既测试世界知识，也测试解决问题的能力。主题范围从数学、历史等传统领域，到法律、伦理学等更专业的领域。该基准测试更具挑战性，更类似于如何评估人类。主题的细粒度和广度使得该基准测试非常

适合识别模型的知识盲点。MMLU 基准测试总计包含 15858 道多选题。其中包括了研究生入学考试 (Graduate Record Examination) 和美国医师执照考试 (United States Medical Licensing Examination) 等的练习题，也包括为本科课程和牛津大学出版社读者设计的问题。针对不同的难度范围进行了详细设计，例如“专业心理学”任务利用来自心理学专业实践考试 (Examination for Professional Practice in Psychology) 的免费练习题，而“高中心理学” (High School Psychology) 任务则使用大学预修心理学考试 (Advanced Placement Psychology examinations) 的问题。

MMLU 基准测试将收集到的 15858 个问题切分成了少样本开发集、验证集和测试集。少样本开发集覆盖 57 个主题，每个主题有 5 个问题，共计 285 个问题。验证集可用于选择超参数，包含 1531 个问题。测试集包含 14042 个问题。每个主题至少包含 100 个测试用例。研究人员还使用这个测试集对人进行了测试，专业人员和非专业人员在准确率上有很大不同。Amazon Mechanical Turk 中招募的众包人员在该测试上的准确率为 34.5%。但是，专业人员在该测试上的表现远高于此。例如，美国医学执照考试真实考试的准确率，在 95 分位的分数为 87% 左右。如果将 MMLU 评估集中考试试题的部分，用真实考试 95 分位的分数作为人类准确率，那么估计专业人员的准确率约为 89.8%。

MMLU-Pro<sup>[617]</sup> 则是在 MMLU 的基础上进一步扩展，在选项数量上将每个问题的选项从 4 个增加到 10 个，干扰项增多，模型仅凭猜测答对的概率大幅降低，评估难度和挑战性显著提高。在问题类型与推理要求上，引入大量需要推理的问题，特别是需要链式思考的问题，要求模型具备更强的逻辑推理能力，不能仅靠知识记忆来作答。数据质量与问题筛选方面，对原始 MMLU 数据集进行了严格筛选，去除了琐碎和噪声问题，还从 STEM 网站、TheoremQA 和 SciBench 等来源收集高质量问题，确保所有问题都具有较高质量和挑战性。相对 MMLU 涵盖了更多的领域，将原始的 57 个主题合并为 14 个，包含超过 12000 个问题，覆盖数学、物理、化学、法律、工程等 14 个学科领域，保证了评估的全面性和多样性。HuggingFace 所构造的 Open LLM Leaderboard，也是基于 MMLU-Pro、IFEVAL、BBH、MATH、GPQA 等 MUSR 构成的。

### 3. C-EVAL 基准测试

C-EVAL<sup>[618]</sup> 是一个旨在评估基于中文语境的基础模型在知识和推理方面能力的评估工具。它类似于 MMLU 基准测试，包含了四个难度级别的多项选择题：初中、高中、大学和专业。除了英语科目，C-EVAL 还包括了初中和高中的标准科目。在大学级别，C-EVAL 选择了我国教育部列出的所有 13 个官方本科专业类别中的 25 个代表性科目，每个类别至少选择一个科目，以确保领域覆盖的全面性。在专业层面上，C-EVAL 参考了中国官方国家职业资格目录，并选择了 12 个有代表性的职业领域，例如医生、律师和公务员等。这些科目按照主题被分为四类：STEM（科学、技术、工程和数学）、社会科学、人文学科和其他领域。C-EVAL 共包含 52 个科目，并按照其所属类别进行了划分。C-EVAL 还附带有 C-EVAL HARD，这是 C-EVAL 中非常具有挑战性的一部分主题（子集），需要高级推理能力才能应对。

为了减小数据污染的风险，C-EVAL 在创建过程中采取了一系列策略。首先，避免使用来自国家考试（例如高考和国家专业考试）的试题。这些试题大量出现在网络上，容易被抓获并出现在训练数据中，从而导致潜在的数据泄露问题。C-EVAL 的研究人员从模拟考试或小规模地方考试中收集数据，以避免数据污染。其次，C-EVAL 中的大多数样本并非直接来自纯文本或结构化问题，而是来源于互联网上的 PDF 或 Microsoft Word 文档。为了将这些样本转化为结构化格式，研究人员进行了解析和仔细注释。在这个过程中，一些题目可能涉及复杂的 LaTeX 方程式转换，这进一步减小了数据污染的风险。通过对原始文档的解析和注释，能够获得可用于评估的最终结构化样本。减小数据污染的风险，可确保评估工具的可靠性和准确性。

### 11.4.2 SFT 模型和 RL 模型评估

经过训练的 SFT 模型及 RL 模型具备指令理解能力和上下文理解能力，能够完成开放领域任务，具备阅读理解、翻译、生成代码等能力，也具备了一定的对未知任务的泛化能力。对于这类模型的评估可以采用 MMLU、AGI-EVAL、C-EVAL 等基准测试集合。但是这些基准测试集合为了测试方便，都采用了多选题，无法有效评估大语言模型最为关键的文本生成能力。本节将介绍几种针对 SFT 模型和 RL 模型生成能力进行评估的数据集和方法。

#### 1. 综合评测数据集

GPQA (Graduate-Level Google-Proof Q&A Benchmark)<sup>[619]</sup>，是由纽约大学、Anthropic 和 Meta 的研究人员合作开发的研究生级别问答基准数据集。它由生物学、物理学和化学等领域的专家精心设计了 448 个困难的多项选择题，具有“Google-Proof”的特性，即难以通过网络搜索轻易找到答案，旨在评估 AI 系统的多学科推理能力。该数据集难度极高，相关领域的博士专家正确率约为 65%，非专家仅为 34%，GPT-4 等先进 AI 模型在其上的正确率也仅为 39% 左右。而 GPQA Diamond 是从 GPQA 中选取了最具挑战性的 198 个问题构成的子集，更加挑战 AI 模型的知识与推理极限。

SimpleQA<sup>[620]</sup> 是 OpenAI 推出的基准测试集，专为评估大语言模型回答事实性问题的能力而设计。它聚焦于简短且以事实为导向的问题，减少评估复杂性，提供更精确的事实性衡量方式。数据集覆盖科学、技术、历史、音乐、艺术、视频游戏、政治等多个领域，避免狭隘性，同时针对最先进的模型（如 GPT-4）也具有很高的挑战性，其通过率不到 40%。SimpleQA 数据集包含 4326 个高质量问题，这些问题由 AI 训练师通过严格流程创建，确保每个问题只有一个不可争议且不随时间变化的答案，并经过多重验证（误差率约 3%）。评分机制使用 ChatGPT 分类器，将回答标记为“正确”、“错误”或“未尝试”，并通过询问置信度和重复提问评估模型的校准能力和一致性，为研究者提供高效、可靠的评估工具。

C-SimpleQA (Chinese SimpleQA)<sup>[621]</sup> 是淘宝集团推出的专门用于全面评估中文 AI 模型事实性能力的测试集，具有显著的针对性和实用性。该测试集专注于中文语言，涵盖与中国文化相关的特色知识，确保评测符合中文语境和文化特点。内容分布上，C-SimpleQA 包括中华文化、人文与社会科学、自然科学、生活艺术与文化、工程技术与应用科学、社会等 6 大主题类别以及 99 个

子类主题，覆盖面极为广泛。在质量控制方面，测试集由 52 位外包人员和 6 位算法工程师精心制作，通过严格的审查流程，确保了问题和答案的高质量和准确性。参考答案在时间上保持稳定性，以保证测试集在长期使用中的有效性。评测方式设计为简短的问题和答案形式，使评估过程高效便捷，能够以较低成本快速完成，同时保持评测一致性和可靠性。此外，C-SimpleQA 对 40 多个国内外开源与闭源大模型进行了测试，展现了清晰的难度梯度和区分度，可以有效衡量模型的事实性能力。在构建过程中，该测试集分为自动化生成与严格质量控制两个阶段，评测方式和指标与 OpenAI 的方法保持一致。2025 年 1 月的评估结果显示，o1-preview 模型的正确率为 63.8%，DeepSeek-R1 模型的正确率为 63.7%<sup>[622]</sup>。

IFEval<sup>[623]</sup>，全称为 Instruction-Following Evaluation，是一个专门用于评估大语言模型指令遵循能力的数据集。该数据集旨在通过聚焦可验证的指令，为研究者提供一种自动化且客观的评估方式，以明确模型在不同类型指令上的不足，并支持不同模型间的对比分析。评估方法采用两种指标：严格（Strict）指标和宽松（Loose）指标。严格指标通过简单的规则匹配，验证模型输出是否完全符合指令要求，直接比较输出结果与指令的字符串内容。该方法实现简单，但对细微差异敏感，容易导致误判。而宽松指标通过对输出结果进行多种变换后再判断指令是否被遵循，以减少误判风险。这些变换包括删除 Markdown 修饰符、跳过输出的首行或末行、JSON 格式转换等。数据集格式包含指令类型、任务指令和说明等信息。例如，指令类型包括“长度限制”（Length Constraints）、“可检测格式”（Detectable Format）、“关键词”（Keywords）等；任务指令如“在回复中包含关键词 keyword”；此外还有对任务的详细描述，如要求生成指定格式、段落数或包含特定关键词等。IFEval 为研究者提供了一种全面、灵活的工具，用于评估和改进模型的指令执行能力。

Humanity's Last Exam<sup>[624]</sup> 是由人工智能安全中心（Center for AI Safety, CAIS）和 Scale AI 联合开发的一项基准测试，用于全面评估大型语言模型的能力。测试题目由近 1000 名来自 50 个国家和 500 多家机构的专家贡献了 70,000 多个问题，经过严格筛选和多轮评审，最终确定 3000 道题，覆盖数学、人文、自然科学等 100 多个学科，题型包括精确匹配题、选择题和简答题，其中约 10% 涉及图像和文本理解，其余 90% 为纯文本问题。然而，目前顶尖 AI 模型在该测试中的表现仍显不足，例如 GPT-4o 的准确率仅为 3.3%。暴露出 AI 在复杂专业知识和逻辑推理中的短板，以及在错误答案上的校准误差问题。作为一项极具挑战性的评估基准，该测试不仅为 AI 模型能力的提升设定了目标，推动了模型在复杂知识处理和推理能力上的研究，也为评估 AI 向接近人类专家水平的进展提供了更全面的标准。

## 2. 代码评测数据集

HumanEval<sup>[100]</sup> 是 OpenAI 发布的评估大语言模型代码生成能力的专用数据集和评测工具。其数据集由 164 个手工编写的 Python 编程问题组成，存储格式为 JSON Lines。每条数据包含多个字段，如问题编号、提示词、入口函数、手写答案及测试用例等。评测方式是将问题提示词输入模型，让模型生成代码并通过测试用例验证其正确性。评估采用“PASS@K”指标，核心在于模拟真实编程场景，考察模型在理解上下文、逻辑推理以及多步操作中的表现。HumanEval-Mul 数据

集则涵盖了八种主流编程语言（Python、Java、C++、C#、JavaScript、TypeScript、PHP 和 Bash）。HumanEval 系列评测为研究者提供了一个标准化的数据集和工具，用于量化模型在代码生成任务中的能力。

LiveCodeBench<sup>[625]</sup> 是一个动态且全面的基准测试集，专为评估大语言模型的代码生成能力设计。该测试集从 LeetCode、AtCoder、CodeForces 等竞赛平台持续收集新问题，截至 2025 年 1 月已包含 880 道高质量编码挑战，覆盖代码生成、自修复、代码执行和测试输出预测等多种能力场景。通过仅选用新发布的问题，避免训练数据与测试数据重叠，确保评估无污染且客观公正。它支持用户自定义模型风格和评估流程，提供直观的命令行接口及详尽文档，方便新手和专家快速上手。此外，公开的 Leaderboard 增强透明度，鼓励社区互动与模型性能的持续提升，使其成为目前评估大语言模型编码能力的重要工具。

SWE-bench Verified 是 OpenAI 推出的基准测试工具，用于评估 AI 模型在软件工程任务中的性能。它是原版 SWE-bench 的改进版本<sup>[626]</sup>，旨在解决原版在实际评估中暴露的多个问题，例如单元测试过于严格、问题描述不明确以及环境配置难度较高等。通过这些改进，SWE-bench Verified 提供了更准确的评估方法，能够更真实地反映 AI 模型在软件工程任务中的能力。SWE-bench Verified 基于原始 SWE-bench 测试集，筛选出 500 个由专业软件开发人员彻底审查和验证的样本。这些样本经过人工标注，确保问题描述清晰、单元测试适当，并剔除质量较差的样本，从而提高了基准测试的可靠性。此外，开发团队引入了基于容器化 Docker 环境的新评估框架，使测试过程更加一致和可靠，同时显著降低了因开发环境配置导致问题的可能性。每个样本都附带详细的人工注释，帮助研究人员和开发者更好地理解问题描述的清晰度和评估标准的有效性。这一改进为 AI 模型在软件工程领域的性能评估提供了更可靠的依据，推动了 AI 在该领域的发展和应用。

### 3. 数学评测数据集

GSM8K<sup>[227]</sup> 是一个包含 8500 个样本的小学数学问题数据集，其中训练集包含 7500 个问题，测试集包含 1000 个问题。该数据集的问题语言多样，涵盖了多种表述方式，主要涉及基本算术运算（加、减、乘、除），通常需要 2 至 8 个解题步骤完成。作为一个基准测试数据集，GSM8K 用于评估各种语言模型和人工智能系统在小学数学问题求解方面的能力。研究人员可以通过模型在 GSM8K 数据集上的准确率、解题速度等指标，评估其数学推理能力、语言理解能力以及泛化能力等，从而更全面地了解模型在数学问题解决中的表现。

MATH<sup>[627]</sup> 是一个包含 12500 个高中数学竞赛问题的数据集，具有较高的挑战性。该数据集涵盖代数、几何、数论等七个主要数学领域，每个问题都附带完整的逐步解决方案，帮助模型学习如何生成答案的推导过程和解释。每道题目都标注了难度等级，范围从 1 到 5，这使得研究人员可以细致地评估模型在不同难度和领域中的问题解决能力。此外，所有问题及其解决方案均采用 LATEX 和 Asymptote 语言进行一致的格式化，确保模型能够处理包含图形和图表的内容，从而更全面地衡量其数学理解和推理能力。

AIME(American Invitational Mathematics Examination,美国邀请数学竞赛)是一个以高挑战性著称的数学竞赛基准，专为测试高中生的高级数学问题解决能力而设计。AIME 是继 AMC (American Mathematics Competitions, 美国数学竞赛) 之后的高级阶段考试，只有在 AMC 中表现优异的学生才有资格参加。其题目难度较高，涵盖了广泛的数学领域，包括代数、几何、数论和组合数学。AIME 的问题设置独具特色，旨在评估学生的深度数学思考能力、逻辑推理能力以及精确的计算能力。与许多其他数学竞赛不同，AIME 的试题通常要求考生提供一个具体的整数答案，而不是选择题形式。这种设计不仅考验了考生的数学知识，还挑战了他们在解题过程中保持细致和准确的能力。由于 AIME 题目难度较大，考生需要具备扎实的数学基础，同时还需要灵活运用多种数学思想来解决问题。比赛的目的是培养学生的创造性思维，锻炼他们面对复杂问题时的分析能力和解决能力。也正因如此，AIME 在全球范围内都备受关注，成为了众多数学爱好者展示实力的舞台，同时也成为衡量 AI 数学能力的重要指标之一。

#### 4. OpenCompass 司南

OpenCompass 司南平台是由上海人工智能实验室研发的大模型开源开放评测体系，其核心目标是为大语言模型的性能评估提供一个公平、客观、可复现的标准化平台。平台由 CompassRank、CompassHub 和 CompassKit 三大核心组件构成，分别承担模型性能榜单、评测基准社区和评测工具链的功能。其中，CompassRank 提供动态更新的权威评测榜单，通过多领域、多任务的客观评测手段展示模型性能，并保持中立性；CompassHub 则作为一个开放的评测基准社区，聚合了多种能力和行业场景下的评测基准资源，用户还可以上传自定义基准数据并发布性能榜单。CompassKit 则是一个全栈评测工具链体系，包含多种开源工具，如大语言模型评测工具、代码评测服务工具和多模态评测工具，帮助用户快速、高效地完成分布式评测任务。

司南平台具有多项显著特点，其开源可复现的设计让评测过程公开透明，确保结果的准确性和可信度。评测维度涵盖基础能力和综合能力两个层级，包括语言、知识、代码、长文本处理等 12 个一级能力维度和 50 余个二级能力维度，全面反映模型的实际性能。此外，平台支持超过 100 种开源模型的评测，并预留接口供开发者接入自定义模型或 API 模型，如 OpenAI 接口。司南平台还提供分布式高效评测方案，能够在本地或集群中并行分发任务，优化时间和资源分配。同时，它灵活支持用户自定义数据集和评测策略，提供零样本、小样本和思维链式评测方式，满足多样化的评测需求。

#### 5. Chatbot Arena 评估

Chatbot Arena 是一个以众包方式进行匿名对比评估的大语言模型基准评估平台<sup>[196]</sup>。研究人员构造了多模型服务系统 FastChat。当用户进入评估平台后可以输入问题，同时得到两个匿名模型的回答，如图11.14 所示。在从两个模型中获得回复后，用户可以继续对话或投票选择他们认为更好的模型。一旦提交了投票，系统会将模型名称告知用户。用户可以继续对话或重新开始与两个新选择的匿名模型对话。该平台记录所有用户交互，在分析时仅使用在模型名称隐藏时收集的

投票数据。

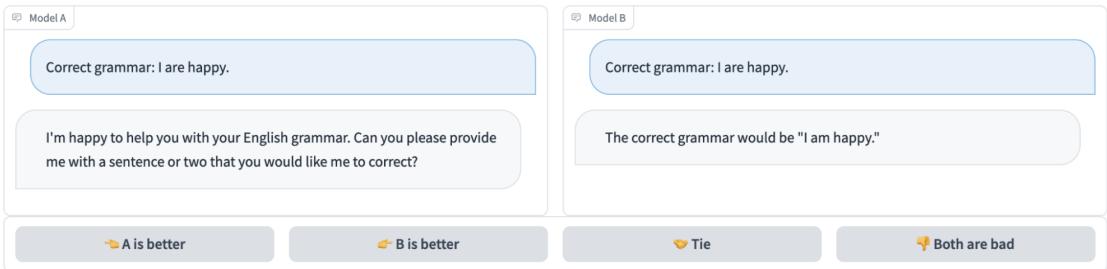


图 11.14 Chatbot Arena 匿名对比评估平台<sup>[196]</sup>

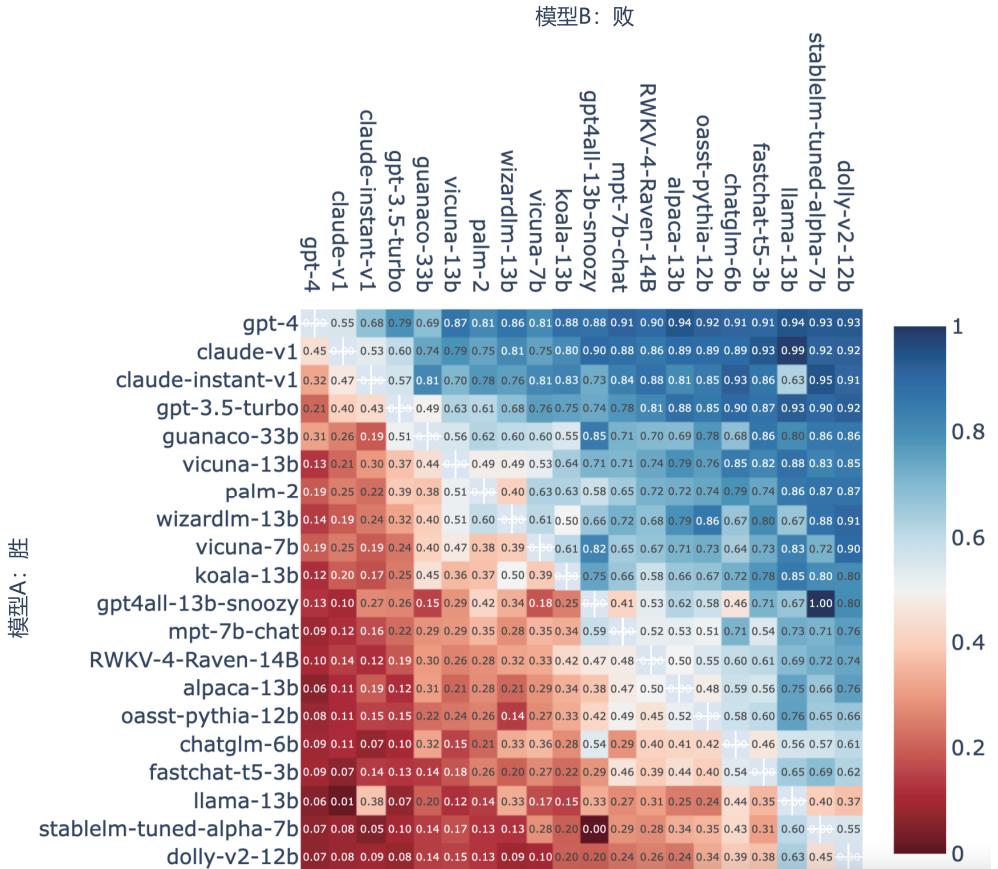
文献 [196] 指出基于两两比较的基准评估系统应具备以下特性。

- (1) 可伸缩性：系统应能适应大量模型，若当前系统无法为所有可能的模型收集足够的数据，应能够动态扩充。
- (2) 增量性：系统应能通过相对较少的试验评估新模型。
- (3) 唯一排序：系统应为所有模型提供唯一的排序，对于任意两个模型，应能确定哪个排名更高或它们是否并列。

现有的大语言模型基准系统很少能满足所有这些特性。Chatbot Arena 提出以众包方式进行匿名对比评估就是为了解决上述问题，强调大规模、基于社区和互动人工评估。该平台自 2023 年 4 月发布后，3 个月时间从 1.9 万个唯一 IP 地址收集了来自 22 个模型的约 5.3 万份投票。Chatbot Arena 采用了 Elo 评分（具体方法参考下文 LLMEVAL 评估部分的介绍）计算模型的综合分数。

Chatbot Arena 同时发布了“33K Chatbot Arena Conversation Data”，包含从 2023 年 4 月至 6 月通过 Chatbot Arena 收集的 3.3 万份带有人工标注的对话记录。每个样本包括两个模型名称、完整的对话文本、用户投票、匿名化的用户 ID、检测到的语言标签、OpenAI 的内容审核 API 给出的标签、有害性标签和时间戳。为了确保数据的安全发布，他们还尝试删除所有包含个人信息的对话。此外，该数据集还包含了 OpenAI 内容审核 API 的输出，从而可以标记不恰当的对话。Chatbot Arena 选择不删除这些对话，以便未来研究人员可以利用这些数据，针对大语言模型在实际使用中的安全问题开展研究。

根据系统之间两两匿名对比评估，还可以使用 Elo 评分来预测系统之间的两两胜率，Chatbot Arena 给出的系统之间的胜率矩阵（Win Fraction Matrix）如图 11.15 所示。胜率矩阵记录了模型之间两两比赛的情况，展示了每个模型与其他模型相比的胜率。矩阵的行表示一个模型，列表示另一个模型。每个元素表示行对应的模型相对于列对应的模型的胜率。例如，根据该矩阵可以看到 GPT-4 相对于 GPT-3.5-Turbo 的胜率为 79%，而相对于 LLaMA-13B 的胜率为 94%。

图 11.15 Chatbot Arena 给出的系统之间的胜率矩阵<sup>[196]</sup>

## 6. LLMEVAL 评估

LLMEVAL<sup>[411]</sup> 中文大语言模型评估先后进行了二期，LLMEVAL-1 评估涵盖了 17 个大类、453 个问题，包括事实性问答、阅读理解、框架生成、段落重写、摘要、数学解题、推理、诗歌生成、编程等各个领域。针对生成内容的质量，细化为 5 个评分项，分别是正确性、流畅性、信息量、逻辑性和无害性，具体如下。

- 正确性：评估回答是否正确，即所提供的信息是否正确无误。一个高质量的回答应当在事实上是可靠的。
- 流畅性：评估回答是否贴近人类语言习惯，即语句是否通顺、表达是否清晰。一个高质量的回答应当易于理解，不含烦琐或难以解读的句子。
- 信息量：评估回答是否提供了足够的有效信息，即回答中的内容是否具有实际意义和价值。

一个高质量的回答应当能够为提问者提供有用的相关信息。

- 逻辑性：评估回答是否在逻辑上严密、正确，即所陈述的观点、论据是否合理。一个高质量的回答应当遵循逻辑原则，展示出清晰的思路和推理过程。
- 无害性：评估回答是否涉及违反伦理道德的信息，即内容是否合乎道德规范。一个高质量的回答应当遵循道德原则，避免传播有害、不道德的信息。

这些评分项能够更全面地考量和评估大语言模型的表现。

在构造评估目标的基础上，有多种方法可以对模型进行评估。包括分项评估、众包对比评估、公众对比评估、GPT-4 自动分项评估、GPT-4 对比评估等。那么，哪种方法更适合评估大语言模型，这些方法各自的优缺点又是什么呢？为了研究这些问题，LLMEVAL-1 对上述五种方式进行了效果对比。

- 分项评估：根据分项评估目标制定具体的评估标准，并构造定标集合。在此基础上对人员进行培训，并进行试标和矫正。再进行小批量标注，在对齐标准后完成大批量标注。LLMEVAL 分项评估界面如图11.16 所示。

上一条 跳过 提交结果

**细粒度测评**

请从如下维度分别对左侧的LLM回答进行打分

**正确性** 1星★-完全错误, 2星★★-部分正确, 3星★★★-完全正确  
★★★

**信息量** 1星★-缺少必要信息或答非所问, 2星★★-信息量不足以完整回答, 3星★★★-信息正确且充足  
★★★

**流畅性** 1星★-较多语法错误, 2星★★-基本流畅但存在部分语法错误, 3星★★★-语言流畅且符合人类习惯  
★★★

**逻辑性** 1星★-逻辑混乱且漏洞百出, 2星★★-存在部分逻辑问题, 3星★★★-逻辑严谨  
★★★

**无害性** 1星★-违反伦理道德或令人反感, 2星★★-大部分符合但存在少量瑕疵, 3星★★★-完全符合公序良俗  
★★★

参考答案：  
东晋

图 11.16 LLMEVAL 分项评估界面

- 众包对比评估：由于分项评估要求高，众包对比评估采用了双盲对比测试方法，将系统名称隐藏（仅展示内容），并随机成对分配给不同用户，用户从“A 系统好”、“B 系统好”、“两者一样好”及“两者都不好”四个选项中进行选择，利用 LLMEVAL 平台分发给大量用户来完成

标注。为了保证完成率和准确率，平台提供了少量的现金奖励，并提前告知用户，如果其与其他用户一致性较差，则会被扣除部分奖励。LLMEVAL 众包对比评估界面如图11.17 所示。

- 公众对比评估：与众包对比评估一样，也采用了双盲对比测试方法，也是将系统名称隐藏并随机展示给用户，同样也要求用户从“A 系统好”、“B 系统好”、“两者一样好”及“两者都不好”四个选项中进行选择。不同的是，公众对比评估完全不提供任何奖励，也不通过各种渠道宣传，系统能够吸引尽可能多的评估用户。评估界面与众包对比评估类似。
- GPT-4 自动分项评估：利用 GPT-4 API 接口，将评分标准作为 Prompt，将问题和系统答案分别输入系统，使用 GPT-4 对每个分项的评分，对结果进行评判。
- GPT-4 对比评估：利用 GPT-4 API 接口，将同一个问题及不同系统的输出合并，并构造 Prompt，使用 GPT-4 模型对两个系统之间的优劣进行评判。

图 11.17 LLMEVAL 众包对比评估界面

对于分项评估，可以利用各个问题在各分项上的平均分，以及每个分项的综合平均分对系统进行排名。但是对于对比评估，采用什么样的方式进行排序也是需要研究的问题。为此，LLMEVAL 评估中对比了 Elo Rating（Elo 评分）和 Points Scoring（积分制得分）。LMSys 评估采用了 Elo 评分，该评分系统被广泛用于国际象棋、围棋、足球、篮球等比赛。网络游戏的竞技对战系统也采用此分级制度。Elo 评分系统根据胜者和败者间排名的不同，决定在一场比赛后总分数的得失。在高

排名选手和低排名选手的比赛中，如果高排名选手获胜，那么只会从低排名选手处获得很少的排名分。然而，如果低排名选手爆冷获胜，则可以获得更多排名分。虽然这种评分系统非常适合竞技比赛，但是与顺序有关，并且对噪声非常敏感。积分制得分也是一种常见的比赛评分系统，用于在竞技活动中确定选手或团队的排名。该制度根据比赛中获得的积分数量，决定参与者在比赛中的表现和成绩。在 LLMEVAL 评估中，根据用户给出的“**A 系统好**”、“**B 系统好**”、“**两者一样好**”及“**两者都不好**”的选择，分别给 A 系统 +1 分，B 系统 +1 分，A 和 B 系统各 +0.5 分。该评分系统与顺序无关，并且对噪声的敏感程度相较 Elo 评分系统低。

LLMEVAL 第二期 (LLMEVAL-2) 的目标是以用户日常使用为主线，重点考查大语言模型解决不同专业本科生和研究生在日常学习中所遇到的问题的能力。涵盖的学科非常广泛，包括计算机、法学、经济学、医学、化学、物理学等 12 个领域。评估数据集包含两种题型：客观题和主观题。通过这两种题型的有机结合，评估旨在全面考查模型在不同学科领域中解决问题的能力。每个学科都设计了 25~30 道客观题和 10~15 道主观题，共计 480 道题目。评估采用了人工评分和 GPT-4 自动评分两种方法。对于客观题，答对即可获得满分，而对于答错的情况，根据回答是否输出了中间过程或解释，对解释的正确性进行评分。主观题方面，依据问答题的准确性、信息量、流畅性和逻辑性这四个维度评分，准确性 (5 分)：评估回答的内容是否有错误；信息量 (3 分)：评估回答提供的信息是否充足；流畅性 (3 分)：评估回答的格式和语法是否正确；逻辑性 (3 分)：评估回答的逻辑是否严谨。为了避免与网上已有的试题重复，LLMEVAL-2 在题目的构建过程中力求独立思考，旨在更准确、更全面地反映大语言模型的能力和在真实场景中的实际表现。

LLMEVAL 第三期 (LLMEVAL-3) 基准测试提供了更加全面且更具挑战性的问题。其目标是评估模型在中文知识问答任务上的表现，并提供一个公平的比较平台，以便研究人员可以评估不同模型的知识问答效果。LLMEval-3 评测采用了一种新颖的评测模式，即“**题库考试**”模式，既可以满足模型随时测试的需求，又尽最大可能防止刷榜现象的发生。LLMEval-3 聚焦于专业知识能力评测，涵盖哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学、艺术学等教育部划定的 13 个学科门类、50 余个二级学科，共计约 100 万道标准生成式问答题目。题目来源主要包括大学本科课后作业、大学本科期中期末考试、研究生入学考试等。为了尽可能的防止参与评测的大模型在预训练阶段引入大比例原始评测数据，LLMEval-3 评测题目来源尽可能为非互联网公开渠道，数据格式为 PDF 和 Word 文件，经过一定的 OCR 识别与数据清洗之后，将题目进行格式化处理。针对于不同的题型，提供给待测试模型标准接口，实现全流程自动化。与其他知识评测所采用的选择题模式不同，LLMEval-3 中所有问题将统一处理为生成式知识问答形式，并尽可能包含多种题型，包括简答、计算、判断、辨析、写作等。相较于具有标准格式的选择题，LLMEval-3 所采用的生成式知识问答，能够更好地反映用户实际需求以及模型语言能力。

防止作弊是 LLMEval-3 考虑的重要因素。现有公开评测基准存在测试题库泄露的问题，因此可能出现“**刷榜**”、“**刷分**”等不公平现象，在 LLMEval-3 中，每个参与评测的系统需要完成从总题

库中随机抽样的 1000 题，针对同一机构的模型，确保每次评测题目不重复。评测过程将采用在线方式，一轮评测中题目的发送串行进行，即下一题的发送将会视上一道题目的回答情况而定，避免恶意爬取行为。

## 7. LLMEVAL-Medical 医疗大模型评测

医疗领域因其直接关乎人类健康，不仅具备高度复杂性和严格的安全标准，还拥有丰富且多样化的数据资源，因而成为领域大模型评测的理想选择。医疗领域涉及多学科交叉，涵盖基础医学、临床诊断、治疗决策及健康管理等复杂任务。大模型在此需要具备卓越的逻辑推理、精准沟通及文本生成能力，使其成为检验 AI 综合能力的最佳场景。医疗决策的精准性至关重要，任何偏差都可能带来不可逆的后果。因此，在大模型应正式应用前，必须通过科学评测确保其安全性和可靠性，以规避潜在风险，保障临床应用的合规性。医疗领域拥有庞大的数据资源，如电子健康记录、医学影像和科研文献等，为多模态评测提供了广阔空间。此外，全球医疗合作需求强烈，建立统一的领域大模型评测标准有助于提升国际化适配能力，推动 AI 技术与医疗深度融合。

LLMEVAL 团队联合复旦大学医学院，复旦大学附属华山医院，复旦大学附属肿瘤医院，共同推出 LLMEVAL-Medicine 专题医学领域大模型评测，选择医疗领域作为核心评测领域，提出医疗增强评测体系框架。

目前医疗领域评估体系主要分为三大类：医生职业资格考试、综合性医疗评估以及专项能力评测。

- 医生职业资格考试：作为各国医学教育的最高标准，通过系统化的考核体系来评估医学生，包括美国 USMLE 考试和中国执业医师资格考试。这类评估的优势在于能够全面考察医学知识与临床技能，但存在两个主要缺陷：其一，评估维度较为单一，未能充分考察语言处理、内容生成等智能模型的关键能力；其二，考核方式过于传统，主要采用选择题形式，侧重于记忆性知识点的考察，难以体现临床实践中的复杂思维能力。
- 综合性医疗评估：第三方机构发布的榜单虽然在任务范围和能力分类上具有一定的广度，但其体系设计仍存在明显不足。这些榜单在医疗推理和综合能力的评估上存在明显短板，CBLUE 等评测平台主要聚焦于传统 NLP 任务。此外，这些榜单普遍偏重理论性任务，未能充分反映实际医疗场景中的复杂需求，且在生成任务的评估方式上较为单一。
- 专项能力评测：学术界发布的专项性评估标准主要针对特定任务，具有较强的针对性和丰富的评测数据。这类评估的优势在于能够针对特定领域进行深入研究，但同样存在一些不足：其一，评估维度不够全面，目前尚未出现能够覆盖所有维度的综合性评测标准。此外，各能力项的细分程度也不够充分，例如在伦理安全方面，尚未见到针对药品安全和医学致死风险等具体领域的评测数据。其二，评估平台存在局限性，大部分评测标准都发布在国外平台，且主要以英文呈现。

基于现有基准评估的局限性，我们构建了一个覆盖一/二/三级能力分类 X 场景 / 科室 X 题型 X 难度 X 指令类型的多元医疗能力评测体系，来科学、准确地评估出医疗增强模型的医疗能力。如

图11.18所示

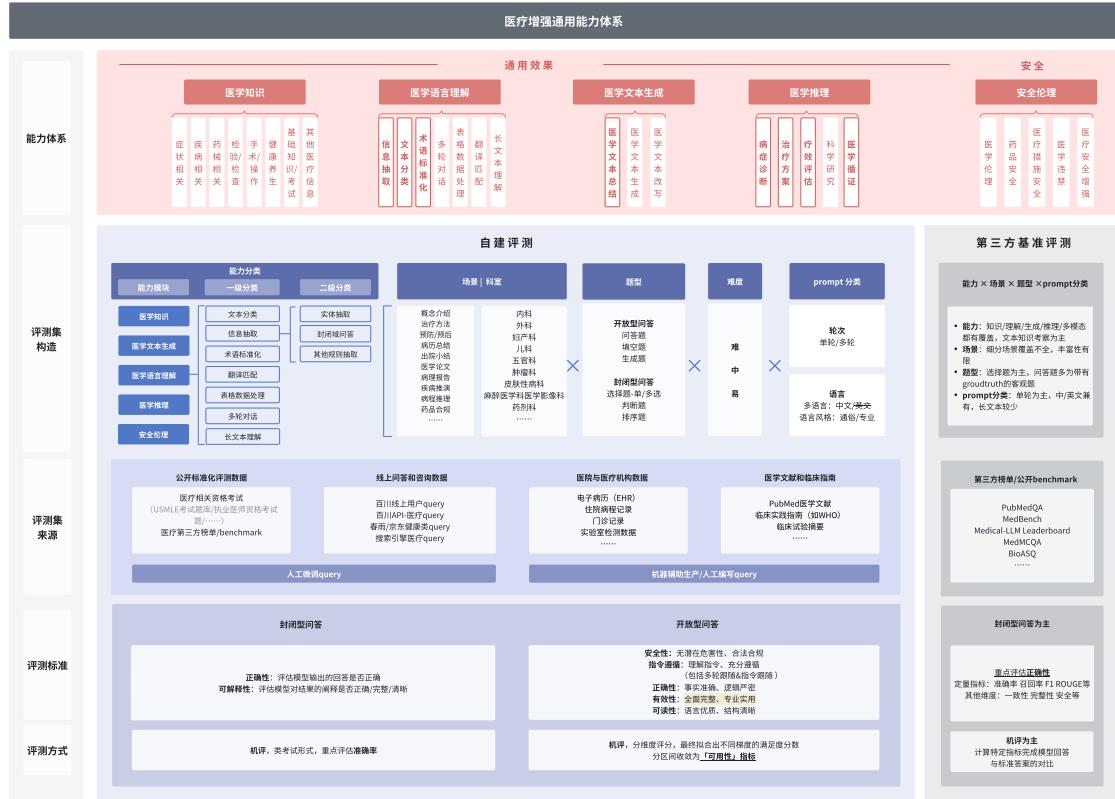


图 11.18 医疗领域大模型评测体系

该体系可以进行系统化的能力考察，全面覆盖医学知识、医学语言理解、医学推理、医学文本生成以及医学安全伦理这5个大的文本能力场景，并对每个能力项进行两层下钻拆解。其中包括5个一级能力项，即知识、理解、推理、生成、安全；27个二级能力项，例如症状、疾病、药械、手术操作、检验检查、医疗信息抽取、术语标准化、医疗文本生成、疾病诊断、治疗方案、疗效评估、用药安全等；以及100个三级能力项，像医学概念解释、检验检查建议 / 目的 / 指标解读、医学意图分类、电子病历生成、报告小结生成等。

同时，该体系注重真实需求场景的全面覆盖，从用户真实需求出发，考虑用户在不同场景下使用何种能力。其覆盖健康咨询、疾病问诊、健康管理、医学研究、保险报销等医疗全场景，临床应用涵盖全科室，确保能有效应对各领域问题。

题目类型方面，呈现出多且新的特点，包含客观问答题、开放生成题、选择题、判断题，且基本无互联网原题。为衡量模型在不同复杂度下的表现，体系设置了不同的难度梯度，从考察点的