大模型评测报告

版本 <1.0>

修订历史记录

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **说明** | **作者** |
| <2023-08-19> | <1.0> | 初稿编写 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

目录（*定稿后要更新域*）

[1. 概述 4](#_Toc143108430)

[1.1 测试范围 4](#_Toc143108431)

[1.2 专业术语说明 4](#_Toc143108432)

[1.3 评测体系介绍 4](#_Toc143108433)

[1.4 测试集介绍 4](#_Toc143108434)

[1.5 测试执行策略 4](#_Toc143108435)

[2. 测试结果评估 4](#_Toc143108436)

[2.1 测试结论 4](#_Toc143108437)

[2.2 测试分析 4](#_Toc143108438)

大模型评测测试报告

# 概述

## 测试范围

随着互联网的发展，模型的规模越来越大。由于这些模型的设计和应用通常涉及与自然语言处理和生成相关的任务：1.在训练大模型之前，需要对原始文本数据进行预处理和准备。这可能涉及分词、词性标注、实体识别等任务，需要深刻的英语理解能力来正确处理文本；2. 大模型的训练可能会涉及学习有关文本的深层次特征表示。理解英语的语法结构、语义关系以及上下文信息对于构建有意义的特征表示至关重要等等。

综上所述，英语理解和运用能力对于大模型在自然语言处理任务中的表现和性能至关重要。在本次测试中，我们将讯飞星火2.0和openAI的gpt3.5看作两名学生，通过对他们英语的四种基础能力（词汇、语法、语篇和语用）的不断测试，发现各能力对应任务的差距，以此来判别他们的综合英语理解运用能力。

## 专业术语介绍

1. 英文问答 : - 描述：这是一个系统根据给定的问题提供答案的任务。问题可以是开放式的或基于特定的事实，答案可以是从大型数据库、知识图谱或文本语料库中提取的。 - 用途：搜索引擎、语音助手、在线客服等。

2. 阅读理解 : - 描述：模型必须从给定的文本段落中读取并理解信息，然后回答关于该段落的问题。 - 用途：在线教育平台、自动文档摘要、助读工具等。

3. 中英翻译 : - 描述：将中文文本翻译成英文，或反之。这需要对源语言和目标语言都有深入的理解。 - 用途：多语言内容平台、国际业务沟通、旅游应用等。

4. 文本写作 : - 描述：基于给定的输入或提示，自动生成连贯的文本内容。 - 用途：内容创作、自动新闻报道、广告创意生成等。

5. 完形填空 : - 描述：在文本中有一个或多个缺失的词，模型必须正确地填充这些空白。 - 用途：语言能力测试、文本完整性检查等。

6. 英语改错 : - 描述：检测并纠正文本中的语法错误，如拼写、标点、句子结构等。 - 用途：文本编辑器、在线写作平台、学术论文校对等。

7. 情感分析 : - 描述：确定给定文本的情感倾向，例如：正面、负面或中性。 - 用途：品牌声誉监控、市场分析、社交媒体趋势分析等。

## 测试集介绍

测试模块：阅读理解、翻译、写作、完形填空、纠错、情感分析、问答

测试集描述：

| 测试模块 | 阅读理解 |
| --- | --- |
| 测试集描述 | 从维基百科摘取的大量科普信息，针对每段中的关键信息和细节问题作出文本回答 |
| 测试集数量 | 215 |
| Prompt样例 | Based on the background information i provide for you, answer my question in English as concise as possible. If you are not sure about the answer, please say you do not know. Attention, the background information is below: background information . Now, my question is: question |
| 注意事项 | 针对同样的阅读理解内容和同一问题，采用了不同的表述方式进行测试，例如：针对花费问题，使用 “how much”、“cost”等同义的不同方式 |

| 测试模块 | 翻译 |
| --- | --- |
| 测试集描述 | 使用gpt4生成的适用于大模型评测的句子，包括古文与现代文各5句；译文答案使用gpt4基于信达雅标准得到 |
| 测试集数量 | 10 |
| Prompt样例 | 使用信达雅的打分标准为下面译文进行打分 |

| 测试模块 | 写作 |
| --- | --- |
| 测试集描述 | 使用gpt4生成的适用于大模型评测的作文题目 |
| 测试集数量 | 30 |
| Prompt样例 | 词语法逻辑无错误: 20分  高级词汇语法: 20分  举例恰当，论证有力: 20分  文章适合主题，内容有深度: 20分  相似度: 20分（一行内同一模型三篇文章的相似度，相似度越高，得分越低） |

| 测试模块 | 完形填空 |
| --- | --- |
| 测试集描述 | 摘取于CLOTH数据集中的高中生完形填空题目 |
| 测试集数量 | 20 |
| Prompt样例 | 用我给你的句子和选项进行完形填空，直接告诉我选什么 这是文章：The world always makes way for the dreamer. When I was twelve years old, my father took me to see Zig Zigler. I remember sitting in that dark hall listening to Mr. Zigler \_ everyone's spirits up to the ceiling.  这是选项:['rise', 'do', 'put', 'raise'] |
| 注意事项 | 若一个句子存在两个以上空位，模型可能无法准确识别出空位数量及其对应选项 |

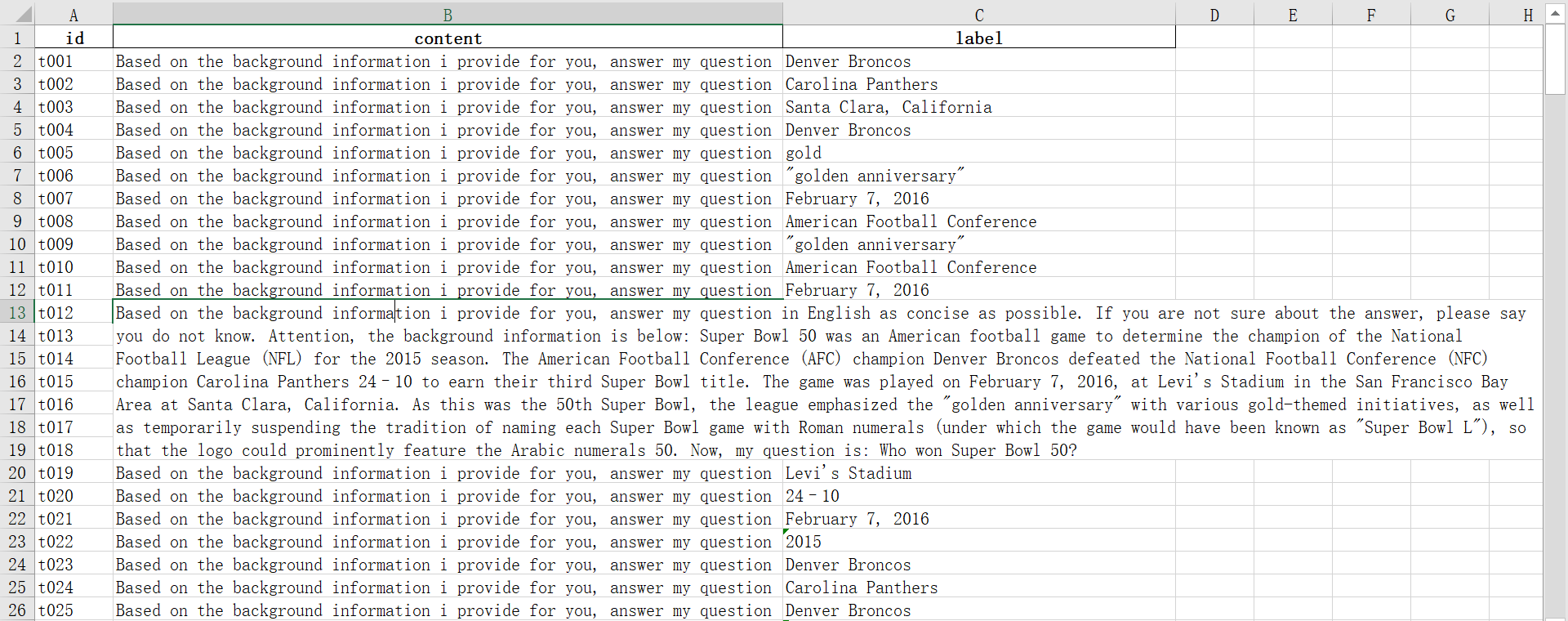
| 测试模块 | 纠错 |
| --- | --- |
| 测试集描述 | 摘自Race数据集，基于可能存在问题的句子，问题包括：拼写错误、语法错误、语句连贯性错误、标点错误，给出改正后的标准答案 |
| 测试集数量 | 20 |
| Prompt样例 | If you are an English teaching software, please help me to correct this sentence so that it expresses fluently and has no grammatical errors. If you think there is no problem with this sentence, then answer 1. If there is a problem, just output the answer directly: |
| 注意事项 | 待改正的句子并非全部存在错误 |

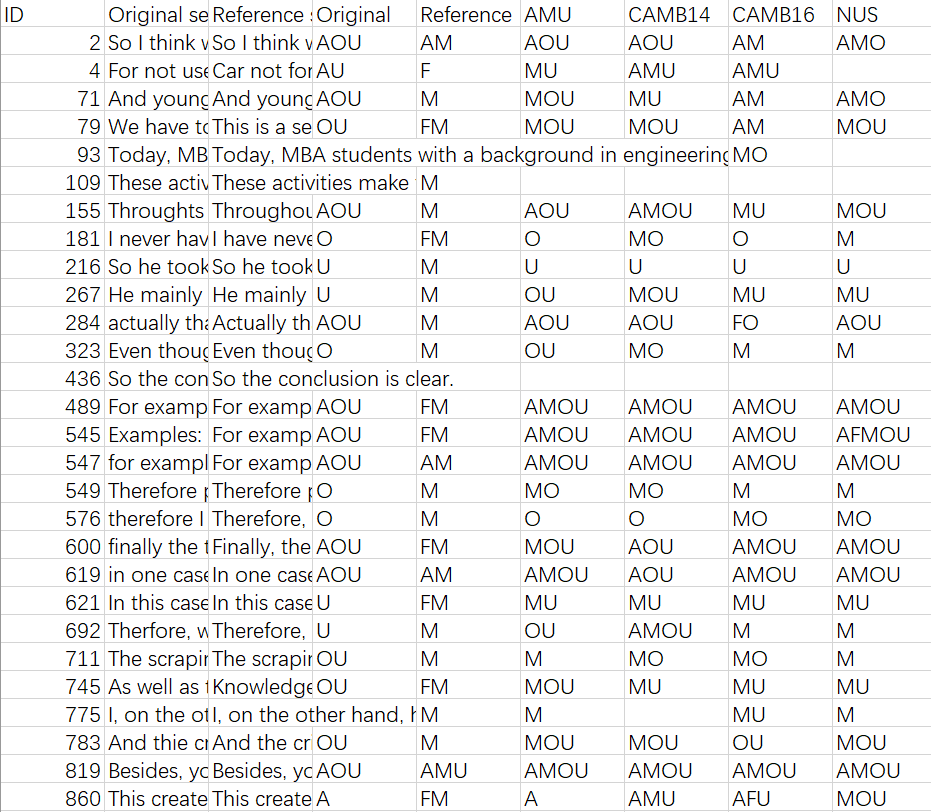
| 测试模块 | 情感分析 |
| --- | --- |
| 测试集描述 | 摘自Emotion数据集，根据句子语义给出对应的情感，包括：悲伤、高兴、生气、恐惧、惊喜 |
| 测试集数量 | 30 |
| Prompt样例一 | Please perform sentiment analysis on the sentence "sentence"; if it expresses anger, please answer 0; if it expresses fear, please answer 1; if it is joy, please answer 2; if it is sadness, please answer 3; if it is surprise, please answer 4, just answer with an int number. |
| Prompt样例二 | Please perform sentiment analysis on the sentence "sentence" You must use one single word among "sadness, joy, anger, fear, surprise" to express the emotion of the sentence. |
| 注意事项 | 最终的准确率为prompt样例一和样例二两种测试结果的平均值且两个模型prompt样例二的准确率均低于样例一的测试准确率。 |

| 测试模块 | 问答 |
| --- | --- |
| 测试集描述 | 摘自DeepMind Q&A数据集，使用gpt4评测星火大模型与chatgpt所给对应问题答案 |
| 测试集数量 | 20 |
| Prompt样例 | 对下面两个答案进行评分，不仅要对比问题和答案，也要对比答案和答案之间的区别，并给出详细的理由 |

测试集建设方法：提取生数据集的数据id、数据内容、数据标签三个字段以及对应内容并转换为csv格式文件。

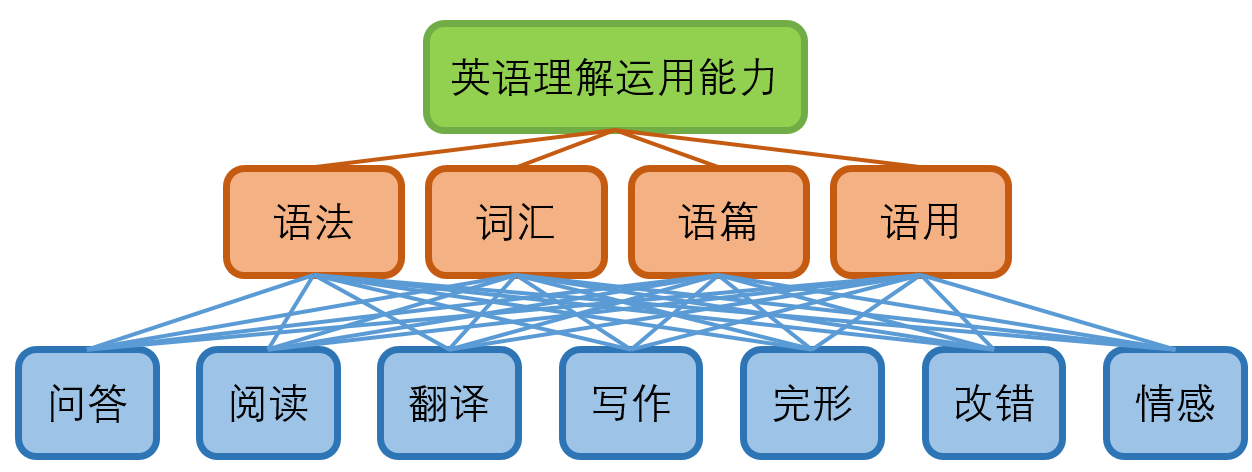
**测试集示例：**

**

**

## 评测体系介绍

我们根据普通高中英语课程标准（2017年版2020年修订）制定评测准则。我们主要关注词汇、语法、语篇和语用知识4项基础能力；使用英文问答、阅读理解、中英翻译、英文写作、完形填空、英语改错和情感分析7项下游任务来测评大模型的英语理解和运用能力。由于四种基础能力的评价准则是相互关联的有机整体，因此同一个下游任务可能以不同的权重影响到不同能力的评价准则。具体而言，我们使用层次分析法来计算测评整个体系的权重。



下面我们针对英语语言知识所包含的四项基础能力，列出大模型所需的具体评价准则。

词汇：词汇是一种语言中所有词和词组的总和。词汇中的任何词语都是通过一定的句法关系和语义关系与其他词语建立起一定联系的，并在语境中传递信息。我们希望大模型不只是记忆词的形和义，更重要的是在语篇中理解和表达与各种主题相关的信息或观点。

语篇：语篇是表达意义的语言单位，在使用语言的过程中，大模型不仅需要运用词汇和语法知识，而且需要将语言组织为意义连贯的语篇。我们希望大模型可以利用语篇知识理解不同长度不同类型的语篇，并根据交流的需要选择恰当的语篇类型、设计合理的语篇结构、规划语篇的组成部分、保持语篇的衔接性和连贯性。

语法：语法包括词法知识和句法知识：词法关注词的形态变化，如名词的数、格，动词的时、态（体）等；句法关注句子结构，如句子的成分、语序、种类等。词法和句法之间的关系非常紧密。在语言使用中，语法知识是“形式—意义—使用”的统一体，直接影响语言理解和表达的准确性和得体性。

语用知识：语用知识指在特定语境中准确理解用户需求，得体表达自己的知识。掌握语用知识有助于模型根据交流目的、预设场合，身份和角色，选择正式或非正式、直接或委婉、口头或 书面语等语言形式，得体且恰当地与用户沟通和交流。

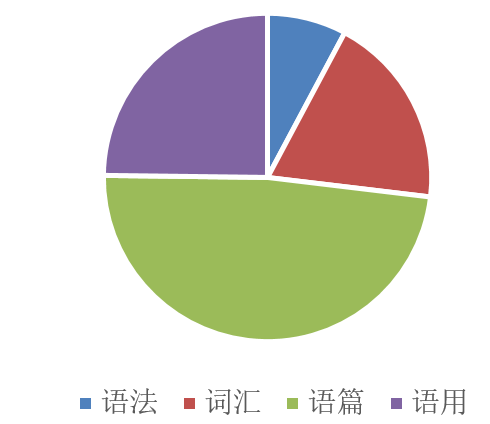
为了合理评估4项基础能力和7个下游任务间的权重关系，我们使用层次分析法设计整体评测体系权重。首先我们组内7人根据英语语言知识所包含的四项基础能力对英语理解和应用能力的重要性关系进行投票。根据投票结果，构建判断矩阵。具体评价方案如下表所示：

|  |  |
| --- | --- |
| 因素i比因素j | 量化值 |
| 同等重要 | 1 |
| 稍微重要 | 3 |
| 较强重要 | 5 |
| 强烈重要 | 7 |
| 极端重要 | 9 |
| 两相邻判断的中间值 | 2，4，6，8 |

根据评价标准，我们可以构建语言能力和英语能力间的评价矩阵为：

矩阵一致性指标通过一致性检测。最后计算可得4种能力的权重分布为

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 语法 | 词汇 | 语篇 | 语用 |
| 权重 | 0.078 | 0.191 | 0.483 | 0.248 |

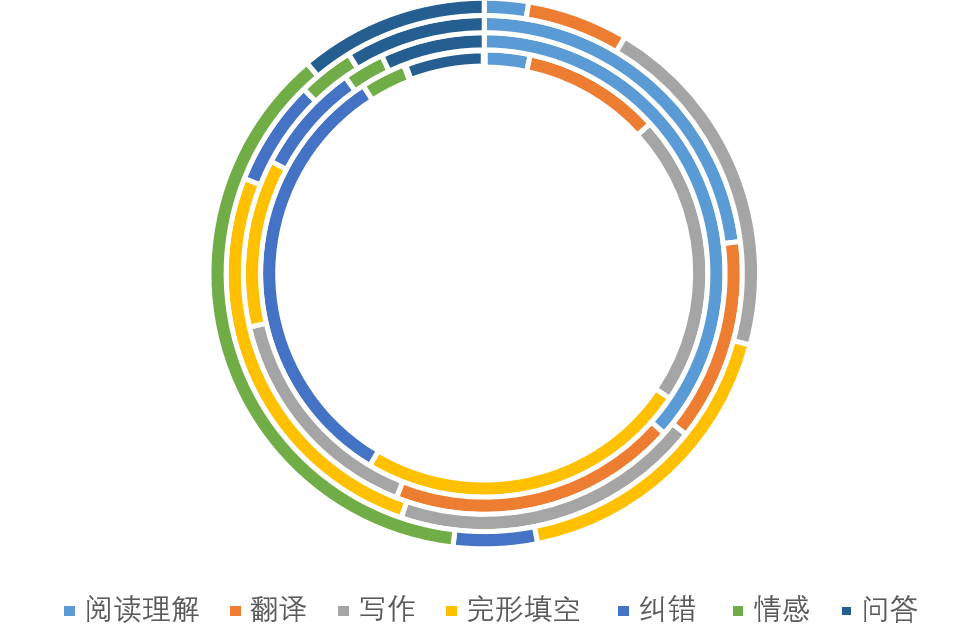


可以看到权重结果语篇占比最高，然后依次为语用、词汇、语法。符合一般用户对大模型英语能力需求。

然后我们继续根据7种下游任务分别对于4项基础能力的权重占比进行投票。最终根据投票结果对于4项基础能力分别得到4个判别矩阵，每个矩阵用于计算7种下游任务与相应基础能力的重要性关系。以语法能力为例，其判别矩阵为：

4个判别矩阵均通过一致性检验。最终得到的权重矩阵为：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 阅读理解 | 翻译 | 写作 | 完形填空 | 纠错 | 情感 | 问答 |
| 语法（15.7%） | 0.033 | 0.100 | 0.211 | 0.241 | 0.323 | 0.033 | 0.058 |
| 词汇（13.5%） | 0.366 | 0.194 | 0.155 | 0.113 | 0.075 | 0.028 | 0.070 |
| 语篇（29.5%） | 0.183 | 0.101 | 0.155 | 0.205 | 0.053 | 0.028 | 0.070 |
| 语用（41.4%） | 0.030 | 0.069 | 0.239 | 0.205 | 0.057 | 0.429 | 0.129 |



## 测试执行策略

我们组成了一个七人的小组，使用七台ThinkPad对讯飞星火2.0以及gpt3.5测试了七种不同的下游任务，综合评估英文理解运用能力。

1、统一prompt：针对每一个题目，构造了统一的prompt供模型和人类使用；

2、预测：系统使用模型进行预测，1）对于情感分析等客观题，要求模型在给出的选项中选择最合适于题干的某一个选项；2）对于写作、问答等主观题，要求模型给出符合prompt要求的相关解答。

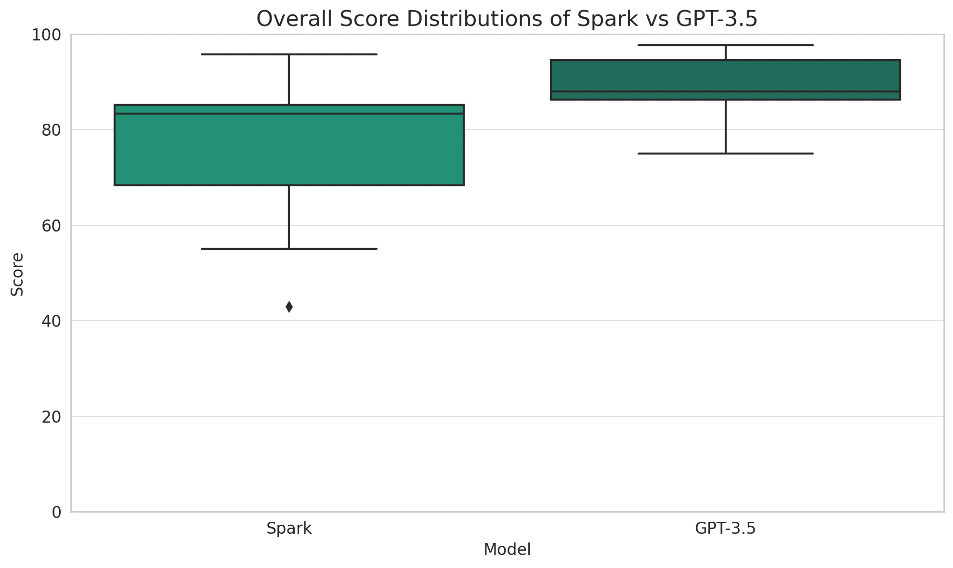
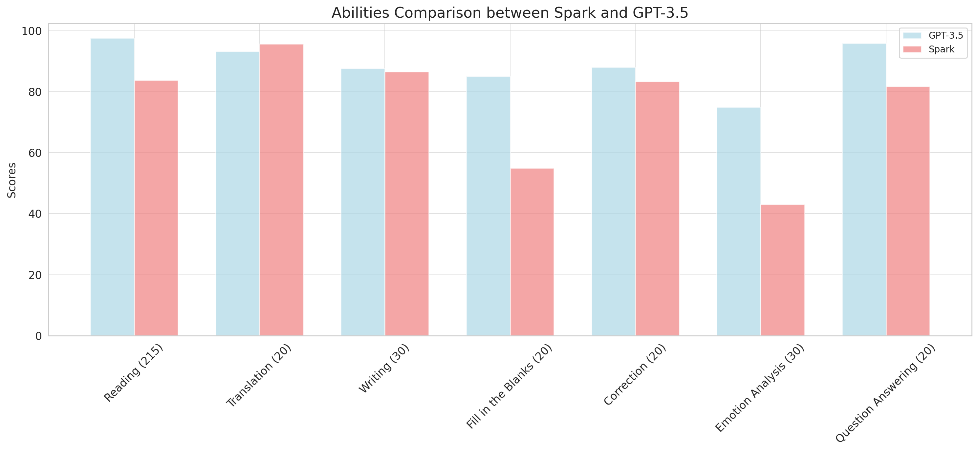
3、打分：如果模型的回答不是标准的答案，而是一段文字，系统会采取我们指定的对于不同任务的个性化评价标准对两个模型给出的答案进行打分。对于主观题和客观题，我们采取了正确率、人工评价（MOS）以及gpt4评测三中不同的方式对相关任务进行打分。该策略结合模型的表现进行优化和完善。具体而言，在正确率评估方式中，我们采用的是经典的ACC评价方式，将标准答案与模型回答答案进行对比；在gpt4自动测评方式种，我们拟定特定符合客观现实的prompt，让gpt4对于不同的答案进行相关的打分；在MOS人工测评方式中，我们选取了组内的两位人员，在拼写，标点，语法，连贯性以及高级词汇的运用方面对于两个模型给出的答案进行人工评分

其中当无法提取有效答案的时候，表明模型没有按照人类的要求做题，未正确理解指令，则会给模型减去一定的分数。在七种任务中，阅读理解、完形填空以及情感分析任务我们采用的是客观题的方式对他们计算正确率；英文问答、英文写作和中英翻译任务采用gpt4自动测评；对于改错我们采取的是MOS人工测评方式。

# 测试结果评估

## 测试结论

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 阅读理解 | 中英翻译 | 英文写作 | 完形填空 | 英语改错 | 情感分析 | 英文问答 | 总分 |
| GPT-3.5 | 97.6 | 93.2 | 87.6 | 85.0 | 88.0 | 75.0 | 95.9 | 87.7 |
| 讯飞星火v2.0 | 83.8 | 95.7 | 86.5 | 55.0 | 83.3 | 43.0 | 81.7 | 72.1 |

**

结论：GPT3.5总体表现优于Spark3.0。在大部分任务中，GPT都有比较优秀的表现，但在情感分析任务上，两个模型的表现都差强人意。

## 测试分析

**2.1.1 中英翻译**

两者在翻译任务上的表现接近，但 "sparkv2.0" 稍微领先。这是sparkv2.0唯一领先的项目，考虑到可能sparkv2.0对中英翻译做了特别的优化或微调。尤其是在古文翻译方面，比gpt表现得更好一点。写作："GPT-3.5" 和 "sparkv2.0" 在写作任务上的表现相当接近。两方对于英语作文内容的生成难分伯仲，几乎都没有语法错误，也都用到了高级的词汇语法，sparkv2.0应该针对英文写作方面也做过相关的优化。

**2.1.2 完形填空题**

"GPT-3.5" 在填空题任务中明显优于 "sparkv2.0"。sparkv2.0只能在单句的完形填空中理解上下文英语进行填空，超过两句后很难理解，无论怎样prompt都不行，难以进行长文本的推理理解和逻辑分析。而gpt3.5无论是在单句还是多句的完形填空中都表现较好，这可能得益于大量英语文本数据的训练。在知识问答中，两边的回答其实基本都是正确的，只是gpt回答得更加细致一些，由于是英语问答，考虑到sparkv2.0可能先翻译成中文理解再用中文回答翻译回去，因为问答中sparkv2.0确实出现了中文。

**2.1.3 英语改错**

两者在英语改错任务任务上的表现相近，但 "GPT-3.5" 稍微领先。在改错过程中GPT有较好的语法辨别能力，测试数据中有一些不需要改正的语句，sparkv2.0大模型往往不能辨别出来而是做一些无用的修改。但在英语高级用法方面，sparkv2.0的语言润色能力更胜一筹。最后，因为spark大模型存在重复Prompt的情况，在70条数据中有51条重复了Prompt的内容，我们在总体分数上按照预设的分数进行了减分。

**2.1.4 英文写作**

两个模型分别对十个题目生成了3次文章，使用gpt4对其进行测评，并且检测了三次文章的重复程度，对其进行打分，可以发现两款模型的英文写作生成能力都很优秀。

**2.1.5 英文问答**

这两个模型的回答实际上基本都是正确的。不过，GPT 的回答更为详细。由于这是一个英语问答任务，需要考虑到 SparkV2.0 可能会先将问题翻译成中文，然后在用中文回答之后再将答案翻译回英文。产生这个推测的原因是这是因为在问答过程中，SparkV2.0 确实出现了中文的情况。

**2.1.6 阅读理解**

gpt3.5优于spark2.0，在一定程度上能体现gpt3.5对于阅读理解的细节程度更加优秀，同时，针对同一问题，我们采用了不同的问法，例如，在花费问题上，我们采用"hou much"、"cost"等不同表述，此时gpt3.5的准确率以及答案的稳定性表现优于spark2.0.

**2.1.7 情感分析**

gpt3.5优于spark2.0，gpt对情感的感知更加细腻，划分更加明确，例如：gpt往往能够正确区分fear情感和sadness情感，而spark2.0将大部分fear判定为sadness。此外，我们对于情感分析对比了以下两种情况：令其直接从情感语义词选项中选择和用进一步用数字代替情感语义词，发现对于第二种情况，gpt和spark的判断准确率均下降，但spark准确率下降幅度远大于gpt，在一定程度上表明，当情感语义的推理链条延长时，gpt更为稳定一些。