如前所述，我们对LLMs进行了理解和生成能力的评估，得出结果如下：

As mentioned above, we evaluated the comprehension and generation capabilities of LLMs and found the following results:

The statistical results of LLMs for dialect word recognition are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Text Number** | **Number of original dialect words** | **Number of words understood by the large model** | | |
| **Chat-GPT** | **Chat-GLM4** | **Jais** |
| 1 | 36 | 30 | 25 | 28 |
| 2 | 43 | 43 | 38 | 41 |
| 3 | 32 | 32 | 29 | 25 |
| 4 | 37 | 37 | 32 | - |
| 5 | 41 | 40 | 37 | 36 |

\* Data is missing at - in the table. Text 4 is a daily conversation about food, and does not contain sensitive content. However, Jais repeatedly indicates that the content policy is violated.

Ratings of translation results at the discourse level (out of 10):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Number** | **Scoring（Out of 10）** | | | | | |
| **Chat-GPT** | | **Chat-GLM4** | | **Jais** | |
| **Completeness** | **Accuracy** | **Completeness** | **Accuracy** | **Completeness** | **Accuracy** |
| 1 | 9 | 7 | 8 | 5 | 4 | 3 |
| 2 | 10 | 9 | 9 | 6 | 5 | 4 |
| 3 | 10 | 9 | 9 | 6 | 4 | 5 |
| 4 | 10 | 9 | 10 | 7 | - | - |
| 5 | 9 | 8 | 10 | 5 | 7 | 8 |

\* Data is missing at - in the table. Text 4 is a daily conversation about food, and does not contain sensitive content. However, Jais repeatedly indicates that the content policy is violated.

In this scoring, the main considerations for completeness scoring are: whether there are omissions in the translation, whether there are any loss of meaning, and whether the paragraph is complete; while the main considerations for accuracy scoring are: whether the translation is accurate, whether there are excessive deletions or additions, and whether it deviates from the original meaning.

这项评分中，完整度评分的主要考虑因素是：翻译有无遗漏，意义有无缺失，以及语段是否完整；而准确率评分的主要考虑因素是：翻译是否准确，有无过度删减或者增补，是否背离原意。

Native speakers‘ rating of the LLMs’ summarization of the text (out of 10):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Number** | **Scoring（Out of 10）** | | | | | |
| **Chat-GPT** | | **Chat-GLM4** | | **Jais** | |
| **Completeness** | **Accuracy** | **Completeness** | **Accuracy** | **Completeness** | **Accuracy** |
| 1 | 9 | 9 | 8 | 7 | 7 | 7 |
| 2 | 9 | 10 | 9 | 10 | 7 | 6 |
| 3 | 10 | 8 | 6 | 5 | 6 | 4 |
| 4 | 10 | 9 | 7 | 7 | 5 | 3 |
| 5 | 7 | 9 | 9 | 7 | 8 | 8 |

与上一份评分相似，这项评分中，完整度评分的主要考虑因素是：总结有无遗漏，意义有无缺失，以及信息覆盖程度；而准确率评分的主要考虑因素是：概括是否准确，有无过度删减或者增补，是否背离原意。

Similar to the previous score, the main considerations for the completeness score in this score are: whether there are omissions in the summary, whether there are any missing meanings, and the degree of information coverage; while the main considerations for the accuracy score are: whether the summary is accurate, whether there are excessive deletions or additions, and whether it deviates from the original meaning.

Native speakers‘ rating of the textes generated by LLMs:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Type** | **Scoring（Out of 10）** | | | | | |
| **Chat-GPT** | | **Chat-GLM4** | | **Jais** | |
| **Degree of Dialectalization** | **Dialect usage accuracy** | **Degree of Dialectalization** | **Dialect usage accuracy** | **Degree of Dialectalization** | **Dialect usage accuracy** |
| weather forecast | 8 | 9 | 8 | 9 | 3 | 8 |
| news report | 7 | 9 | 4 | 8 | 7 | 7 |
| short story | 9 | 9 | 6 | 7 | 3 | 6 |
| dialogue | 9 | 8 | 5 | 7 | 8 | 8 |
| advertisement | 9 | 8 | 8 | 9 | 9 | 9 |

在这项评分中，方言化程度评分的主要考虑因素是：风格上更偏向标准阿拉伯语还是埃及方言，更符合标准阿拉伯语的表达习惯还是方言的表达习惯；而方言使用准确率评分的主要考虑因素是：使用的方言是否合适，使用的是否是埃及方言而不是其他方言。

In this scoring, the main considerations for the dialectalization score are: whether the style is more inclined to standard Arabic or Egyptian dialect, and whether it is more in line with the expression habits of standard Arabic or dialect; while the main considerations for the dialect usage accuracy score are: whether the dialect used is appropriate, and whether the dialect used is Egyptian dialect rather than other dialects.