

A LLM USAGE

Large Language Models (LLMs) were used solely to refine the manuscript’s language, including sentence rephrasing, grammar checking, and improving clarity and readability. Their role was limited to polishing the text and enhancing its overall flow.

The LLM was not involved in idea generation, methodology, or analysis. All scientific contributions were made by the authors, who take full responsibility for the content. The use of the LLM was restricted to language improvements and complied with ethical standards, ensuring no plagiarism or scientific misconduct.

B APPENDIX

Table 7: Score comparison of different models and methods between Expert A and Expert B

Model	Expert A				Expert B			
	Spear	Exact	± 1	± 2	Spear	Exact	± 1	± 2
Qwen2.5_3B _N	0.0047	15.00	44.00	75.00	0.0948	18.50	63.00	86.50
Qwen2.5_14B _N	0.1431	8.00	37.00	60.00	0.1304	11.00	46.50	76.00
Qwen2.5_32B _N	0.0273	12.00	39.50	64.50	0.1279	13.50	54.00	79.50
Qwen2.5_3B _O	-0.1185	14.50	43.00	68.50	-0.0792	21.00	55.00	80.50
Qwen2.5_14B _O	0.0813	13.50	43.00	71.50	0.1280	20.50	55.00	87.50
Qwen2.5_32B _O	0.3409	25.50	56.00	89.00	0.3566	34.00	75.50	97.50
Qwen2.5_3B _{OR}	0.0075	19.00	56.00	79.50	0.0192	23.00	64.00	89.50
Qwen2.5_14B _{OR}	0.2091	22.00	56.50	84.00	0.2082	27.50	72.00	94.00
Qwen2.5_32B _{OR}	0.3862	23.00	71.00	96.00	0.3094	35.00	83.00	98.50
Qwen2.5_3B _{MR}	0.0873	8.00	29.00	57.50	0.1560	10.50	37.00	62.50
Qwen2.5_14B _{MR}	0.0524	21.50	48.00	80.00	0.0870	29.00	68.00	89.00
Qwen2.5_32B _{MR}	0.2589	18.00	54.50	82.50	0.2322	27.00	68.00	93.50
Qwen3_4B _N	-0.0460	8.63	30.82	59.18	-0.0600	9.50	37.00	65.00
Qwen3_14B _N	-0.0389	8.00	28.50	50.00	-0.0088	9.50	35.50	61.50
Qwen3_32B _N	-0.1358	6.50	26.00	51.50	-0.1579	10.00	35.00	58.50
Qwen3_4B _O	0.0853	21.51	58.08	83.42	0.1088	26.50	64.00	92.00
Qwen3_14B _O	0.0794	17.50	49.50	77.00	0.1975	23.00	64.00	90.00
Qwen3_32B _O	0.2011	21.50	52.00	84.50	0.0874	23.00	65.50	91.50
Qwen3_4B _{OR}	0.0933	21.92	60.00	85.07	0.0926	28.00	68.00	94.50
Qwen3_14B _{OR}	0.1682	24.00	54.50	83.00	0.1259	27.50	70.50	92.00
Qwen3_32B _{OR}	0.1304	25.50	56.50	88.00	0.2147	32.50	72.50	95.00
Qwen3_4B _{MR}	0.1759	13.01	44.93	72.60	0.2283	13.00	51.00	81.50
Qwen3_14B _{MR}	0.0357	19.50	53.00	77.00	0.1307	26.00	68.00	93.00
Qwen3_32B _{MR}	0.1270	18.00	52.00	77.00	0.1317	25.00	63.00	91.00
QWQ_32B _N	-0.0022	9.00	29.50	51.00	-0.0475	6.50	33.50	62.00
QWQ_32B _O	0.2072	20.50	54.00	81.50	0.1771	23.00	63.50	89.50
QWQ_32B _{OR}	0.2323	28.00	65.00	90.00	0.2139	33.50	70.50	90.50
QWQ_32B _{MR}	0.7311	64.52	86.71	96.03	0.4552	30.50	72.50	91.00
Llama3.1_8B _N	-0.0938	8.00	36.50	55.00	-0.0154	7.50	42.00	71.50
Llama3.1_70B _N	0.0410	7.00	31.50	53.50	0.0383	5.00	37.00	71.50
Llama3.1_8B _O	-0.2584	10.00	38.50	61.50	-0.1088	20.50	52.00	77.50
Llama3.1_70B _O	0.1654	10.50	38.00	64.50	0.1715	14.50	45.50	77.50
Llama3.1_8B _{OR}	-0.1638	22.50	59.50	82.00	-0.1657	24.00	61.00	82.00
Llama3.1_70B _{OR}	0.0815	13.50	43.50	65.00	0.2103	12.50	54.50	80.00
Llama3.1_8B _{MR}	-0.2273	17.50	55.00	75.00	-0.2023	21.00	64.00	81.00
Llama3.1_70B _{MR}	0.0293	12.50	45.50	68.50	0.0413	18.00	54.50	80.00
GLM4_9B _N	-0.1057	5.50	31.00	54.00	-0.0770	7.50	36.00	68.50
GLM4_9B _O	0.0188	9.50	36.50	58.50	0.1303	7.00	45.00	76.50

Continued on next page

Table 7 – Continued from previous page

Model	Expert A				Expert B			
	Spear	Exact	± 1	± 2	Spear	Exact	± 1	± 2
GLM4.9B _{OR}	0.1711	13.00	40.50	71.00	0.2035	17.00	55.00	83.50
GLM4.9B _{MR}	0.1465	15.00	41.00	68.50	0.1517	15.00	52.00	79.50
DeepSeek_Distill_32B _N	0.0535	10.00	36.00	59.50	0.0986	10.50	46.00	76.50
DeepSeek_Distill_32B _O	0.1325	19.50	48.00	75.00	-0.0159	17.50	57.50	83.00
DeepSeek_Distill_32B _{OR}	0.0129	21.00	56.00	80.00	0.0402	24.50	67.00	90.50
DeepSeek_Distill_32B _{MR}	0.1720	20.00	48.50	74.00	0.0432	26.00	58.00	86.00
DeepSeek_Chat_671B _{MR}	0.4918	40.55	72.05	89.73	0.2971	29.00	73.50	91.50
DeepSeek_Reason_671B _{MR}	0.7640	58.22	88.90	98.36	0.3868	31.50	76.50	93.50

Table 8: Score comparison of different models and methods between Expert A

Comparison	Model Size	Spearman	MAE	RMSE	Kendall tau	Exact	± 1	± 2
Qwen2.5 _N	3B	0.0026	1.7250	2.0579	0.0025	15.00	44.00	75.00
Qwen2.5 _N	14B	0.1066	2.0750	2.3843	0.0939	8.00	37.00	60.00
Qwen2.5 _N	32B	-0.0093	1.9500	2.2891	-0.0089	12.00	39.50	64.50
Qwen2.5 _O	3B	-0.1619	1.8650	2.2417	-0.1423	14.50	43.00	68.50
Qwen2.5 _O	14B	0.0600	1.7800	2.1024	0.0488	13.50	43.00	71.50
Qwen2.5 _O	32B	0.3716	1.2950	1.6171	0.3309	25.50	56.00	89.00
Qwen2.5 _{OR}	3B	0.0278	1.5150	1.8934	0.0227	19.00	56.00	79.50
Qwen2.5 _{OR}	14B	0.1839	1.3900	1.7292	0.1588	22.00	56.50	84.00
Qwen2.5 _{OR}	32B	0.3659	1.1000	1.3565	0.3204	23.00	71.00	96.00
Qwen2.5 _{MR}	3B	0.0574	2.4100	2.8000	0.0486	8.00	29.00	57.50
Qwen2.5 _{MR}	14B	0.0106	1.5350	1.8855	0.0079	21.50	48.00	80.00
Qwen2.5 _{MR}	32B	0.2511	1.4650	1.7790	0.2145	18.00	54.50	82.50
Qwen3 _N	4B	-0.0911	2.2562	2.6153	-0.0800	8.63	30.82	59.18
Qwen3 _N	14B	-0.1143	2.4800	2.8583	-0.0990	8.00	28.50	50.00
Qwen3 _N	32B	-0.1787	2.5300	2.8914	-0.1570	6.50	26.00	51.50
Qwen3 _O	4B	0.0587	1.3918	1.7392	0.0498	21.51	58.08	83.42
Qwen3 _O	14B	0.0568	1.6250	1.9887	0.0483	17.50	49.50	77.00
Qwen3 _O	32B	0.1660	1.4450	1.7819	0.1444	21.50	52.00	84.50
Qwen3 _{OR}	4B	0.0728	1.3548	1.7037	0.0609	21.92	60.00	85.07
Qwen3 _{OR}	14B	0.1160	1.3900	1.7349	0.0977	24.00	54.50	83.00
Qwen3 _{OR}	32B	0.1289	1.3150	1.6628	0.1096	25.50	56.50	88.00
Qwen3 _{MR}	4B	0.1454	1.7849	2.1310	0.1236	13.01	44.93	72.60
Qwen3 _{MR}	14B	0.0059	1.5350	1.8908	0.0043	19.50	53.00	77.00
Qwen3 _{MR}	32B	0.1027	1.5650	1.9118	0.0876	18.00	52.00	77.00
QWQ _N	32B	-0.0809	2.4700	2.8705	-0.0702	9.00	29.50	51.00
QWQ _O	32B	0.2165	1.4650	1.8097	0.1839	20.50	54.00	81.50
QWQ _{OR}	32B	0.2307	1.1950	1.5636	0.1970	28.00	65.00	90.00
QWQ _{MR}	32B	0.7462	0.5507	1.0624	0.6917	64.52	86.71	96.03
Llame3.1 _N	8B	-0.1488	2.2200	2.5768	-0.1314	8.00	36.50	55.00
Llame3.1 _N	70B	-0.0053	2.2850	2.6048	-0.0045	7.00	31.50	53.50
Llame3.1 _O	8B	-0.2501	2.0800	2.4454	-0.2180	10.00	38.50	61.50
Llame3.1 _O	70B	0.1494	2.0050	2.3420	0.1314	10.50	38.00	64.50
Llame3.1 _{OR}	8B	-0.1316	1.4600	1.9079	-0.1109	22.50	59.50	82.00
Llame3.1 _{OR}	70B	0.0547	1.9200	2.3065	0.0457	13.50	43.50	65.00
Llame3.1 _{MR}	8B	-0.2274	1.6700	2.1119	-0.1871	17.50	55.00	75.00
Llame3.1 _{MR}	70B	-0.0107	1.8700	2.2450	-0.0094	12.50	45.50	68.50
GLM4 _N	9B	-0.1529	2.3850	2.7322	-0.1370	5.50	31.00	54.00
GLM4 _O	9B	0.0051	2.1250	2.4708	0.0035	9.50	36.50	58.50
GLM4 _{OR}	9B	0.1757	1.8450	2.1783	0.1460	13.00	40.50	71.00
GLM4 _{MR}	9B	0.1288	1.9000	2.2935	0.1072	15.00	41.00	68.50
DeepSeekDistill _N	32B	0.0136	2.1100	2.4576	0.0114	10.00	36.00	59.50
DeepSeekDistill _O	32B	0.1180	1.6650	2.0676	0.0969	19.50	48.00	75.00
DeepSeekDistill _{OR}	32B	0.0010	1.4750	1.8507	0.0005	21.00	56.00	80.00
DeepSeekDistill _{MR}	32B	0.1618	1.6400	2.0248	0.1329	20.00	48.50	74.00
DeepSeekChat _{MR}	671B	0.4777	1.0027	1.4609	0.4205	40.55	72.05	89.73
DeepSeekReason _{MR}	671B	0.7502	0.5452	0.9126	0.6781	58.22	88.90	98.36
GPT-4O-mini _{MR}	-	0.0752	1.5800	1.9339	0.0634	19.00	49.50	78.00
GPT-5-mini _{MR}	-	0.4159	1.1350	1.4335	0.3447	25.00	69.00	92.50
GPT-5 _{MR}	-	0.2878	1.1700	1.4900	0.2390	26.50	64.50	93.00

Table 9: Score comparison of different models and methods between Expert B

Comparison	Model Size	Spearman	MAE	RMSE	Kendall tau	Exact	± 1	± 2
Qwen2.5 _N	3B	0.0736	1.3300	1.6340	0.0651	18.50	63.00	86.50
Qwen2.5 _N	14B	0.0458	1.7000	1.9849	0.0408	11.00	46.50	76.00
Qwen2.5 _N	32B	0.0947	1.5550	1.8561	0.0859	13.50	54.00	79.50
Qwen2.5 _O	3B	-0.1422	1.4700	1.8358	-0.1302	21.00	55.00	80.50
Qwen2.5 _O	14B	0.0666	1.3850	1.6956	0.0565	20.50	55.00	87.50
Qwen2.5 _O	32B	0.3676	0.9300	1.2329	0.3324	34.00	75.50	97.50
Qwen2.5 _{OR}	3B	0.0216	1.2700	1.6248	0.0178	23.00	64.00	89.50
Qwen2.5 _{OR}	14B	0.1755	1.0650	1.3657	0.1530	27.50	72.00	94.00
Qwen2.5 _{OR}	32B	0.2756	0.8350	1.1113	0.2442	35.00	83.00	98.50
Qwen2.5 _{MR}	3B	0.0607	2.0550	2.4031	0.0497	10.50	37.00	62.50
Qwen2.5 _{MR}	14B	0.0238	1.1400	1.4900	0.0201	29.00	68.00	89.00
Qwen2.5 _{MR}	32B	0.2071	1.1200	1.4318	0.1793	27.00	68.00	93.50
Qwen3 _N	4B	-0.1385	2.0250	2.3590	-0.1230	9.50	37.00	65.00
Qwen3 _N	14B	-0.1138	2.1050	2.4525	-0.1013	9.50	35.50	61.50
Qwen3 _N	32B	-0.2137	2.1450	2.4990	-0.1865	10.00	35.00	58.50
Qwen3 _O	4B	0.0776	1.1800	1.5000	0.0673	26.50	64.00	92.00
Qwen3 _O	14B	0.1821	1.2400	1.5556	0.1625	23.00	64.00	90.00
Qwen3 _O	32B	0.0832	1.2100	1.5166	0.0728	23.00	65.50	91.50
Qwen3 _{OR}	4B	0.0605	1.1150	1.4474	0.0526	28.00	68.00	94.50
Qwen3 _{OR}	14B	0.0681	1.1050	1.4300	0.0582	27.50	70.50	92.00
Qwen3 _{OR}	32B	0.2101	1.0000	1.3229	0.1822	32.50	72.50	95.00
Qwen3 _{MR}	4B	0.1824	1.5800	1.8735	0.1579	13.00	51.00	81.50
Qwen3 _{MR}	14B	0.1042	1.1500	1.4799	0.0915	26.00	68.00	93.00
Qwen3 _{MR}	32B	0.0708	1.2200	1.5427	0.0612	25.00	63.00	91.00
QWQ _N	32B	-0.1022	2.1750	2.4990	-0.0887	6.50	33.50	62.00
QWQ _O	32B	0.1772	1.2600	1.5937	0.1536	23.00	63.50	89.50
QWQ _{OR}	32B	0.2067	1.0700	1.4629	0.1791	33.50	70.50	90.50
QWQ _{MR}	32B	0.4779	1.0900	1.4832	0.4069	30.50	72.50	91.00
Llame3.1 _N	8B	-0.0351	1.8650	2.1506	-0.0310	7.50	42.00	71.50
Llame3.1 _N	70B	-0.0075	1.9400	2.1932	-0.0067	5.00	37.00	71.50
Llame3.1 _O	8B	-0.1034	1.5850	1.9962	-0.0895	20.50	52.00	77.50
Llame3.1 _O	70B	0.1127	1.6500	1.9468	0.0979	14.50	45.50	77.50
Llame3.1 _{OR}	8B	-0.1652	1.3850	1.8014	-0.1402	24.00	61.00	82.00
Llame3.1 _{OR}	70B	0.1914	1.5650	1.8722	0.1641	12.50	54.50	80.00
Llame3.1 _{MR}	8B	-0.2067	1.3950	1.7958	-0.1742	21.00	64.00	81.00
Llame3.1 _{MR}	70B	0.0004	1.5150	1.8615	0.0012	18.00	54.50	80.00
GLM4 _N	9B	-0.1236	2.0100	2.3195	-0.1112	7.50	36.00	68.50
GLM4 _O	9B	0.0769	1.7700	2.0396	0.0655	7.00	45.00	76.50
GLM4 _{OR}	9B	0.1866	1.4800	1.8028	0.1574	17.00	55.00	83.50
GLM4 _{MR}	9B	0.1219	1.5950	1.9352	0.1009	15.00	52.00	79.50
DeepSeekDistill _N	32B	0.0896	1.7350	2.0433	0.0799	10.50	46.00	76.50
DeepSeekDistill _O	32B	-0.0128	1.4800	1.8385	-0.0119	17.50	57.50	83.00
DeepSeekDistill _{OR}	32B	0.0217	1.2100	1.5620	0.0174	24.50	67.00	90.50
DeepSeekDistill _{MR}	32B	0.0452	1.3450	1.7450	0.0388	26.00	58.00	86.00
DeepSeekChat _{MR}	671B	0.2724	1.0700	1.4142	0.2350	29.00	73.50	91.50
DeepSeekReason _{MR}	671B	0.3824	0.9950	1.3360	0.3260	31.50	76.50	93.50
GPT-4O-mini _{MR}	-	0.1312	1.2250	1.5508	0.1154	23.50	65.50	90.00
GPT-5-mini _{MR}	-	0.3262	0.9800	1.2961	0.2763	31.00	76.50	95.00
GPT-5 _{MR}	-	0.1491	1.0550	1.4089	0.1263	28.50	76.00	92.00

Requirements

请帮我判断输入文本与输出文本是否满足以下要求:

要求1.在保持原文含义不变的前提下,用原文不同的文字进行表达,不额外新增、删减会严重影响语义的内容。

要求2.在保证逻辑正确、语序通顺的前提下,变换实体、短语、概念出现的前后顺序、结构关系;顺序变换幅度越小,分数越低。

要求3.在保证特殊实体、引用、解释、特定语句不变的前提下,使用同义词替换;同义词替换越少,分数越低;

要求4.保持与原文语言风格一致,字数长度变化合理;语言风格差距越大,分数越低。

要求5.改写综合评分要非常严格,同时考虑上述4个要求;在保证语义一致(要求1)的前提下,(要求2,3)变化越大越好,字数越接近越好(要求4);如果存在大量长字段完全复制,则给0分。

评分准则,每个要求最低0分,满分均为5分,评分要尽可能严格。

Please help me determine whether the input text and output text meet the following requirements:

Requirement 1. On the premise of keeping the meaning of the original text unchanged, use different words of the original text to express, and do not add or delete additional content that will seriously affect the semantics.

Requirement 2. Change the sequence and structural relationship of entities, phrases and concepts under the premise of correct logic and smooth word order; The smaller the magnitude of the order change, the lower the score.

Requirement 3. Use synonym substitution under the premise of ensuring that special entities, references, interpretations, and specific statements remain unchanged; The fewer synonym substitutions, the lower the score;

Requirement 4. Keep consistent with the original language style, and the length of the word is reasonable; The larger the language style gap, the lower the score.

Requirement 5. Rewrite the overall score to be very strict, while considering the above 4 requirements; Under the premise of ensuring semantic consistency (requirement 1), (requirement 2,3) the greater the change, the better, and the closer the number of words, the better (requirement 4); A score of 0 is given if there exists a large number of long fields with exact duplicates.

The scoring criteria, each of which requires a minimum of 0 points and a full score of 5 points, should be as strict as possible.

Only Prompt

用以下格式只输出改写综合评分结果:

To output only the rewritten composite score, use the following format:

```
“json
{
  "score":
...
}
—
{data}
```

Only-Reason Prompt

用以下格式只输出改写综合评分结果:

To output only the rewritten composite score, use the following format:

```
“json
{
  "reason":
  "score":
...
}
—
{data}
```

Multi-Reason Prompt

用以下格式输出最终结果:

To output only the rewritten composite score, use the following format:

```
“json
[
  {
    "req1": "reason"
    "score":
  },
  {
    "req2": "reason"
    "score":
  },
  {
    "req3": "reason"
    "score":
  },
  {
    "req4": "reason"
    "score":
  },
  {
    "req5": "reason"
    "score":
  }
]
...
—
{data}
```

Table 10: Only, Only-Reason, Multi-Reason Prompt

None Prompt

请帮我评判输入文本和输出文本重写的质量，并按照以下格式输出重写评分结果，0-5分：

Please help me judge the quality of input text and output text rewriting, and output the rewriting scoring result with the following format, 0-5

points:

```
```json
{
 "score":
...
}
```

```
—
{data}
```

---

Table 11: None Prompt