Qwen3-7B、Xiaomi-7B、GLM-4V-Flash 和 DeepSeek-8B 模型评测分析报告

1. 引言

本报告基于 BeNEDect 数据集（9600 条样本），对 Qwen3-7B、Xiaomi-7B、GLM-4V-Flash 和 DeepSeek-8B 模型的数值错误检测性能进行分析。任务为判断段落中的数字是否错误（Yes/No），评测指标包括准确率（Accuracy）、真阳性（TP）、真阴性（TN）、假阳性（FP）、假阴性（FN）和生成错误（Generation Error）。报告从总体性能、按领域、错误类型、操作和提示类型五个维度对比模型表现，分析优劣势，并提出优化建议。

---

2. 总体性能分析

2.1 总体指标

指标	Qwen3-7B	Xiaomi-7B	GLM-4V-Flash	DeepSeek-8B
准确率 (Accuracy)	0.479	0.472	**0.594**	0.463
真阳性 (TP)	4443 (0.463)	4347 (0.453)	3353 (0.349)	3849 (0.401)
真阴性 (TN)	156 (0.016)	181 (0.019)	**2350 (0.245)**	593 (0.062)
假阳性 (FP)	4644 (0.484)	4619 (0.481)	2450 (0.255)	4207 (0.438)
假阴性 (FN)	357 (0.037)	453 (0.047)	1447 (0.151)	951 (0.099)
生成错误 (Generation Error)	577 (0.060)	634 (0.066)	**2 (0.000)**	509 (0.053)

分析：

- **准确率**：GLM-4V-Flash (0.594) 显著优于其他模型，Qwen3-7B (0.479)、Xiaomi-7B (0.472) 和 DeepSeek-8B (0.463) 表现接近，但均低于 0.5，表明整体性能有待提升。
- **真阳性 (TP)**：Qwen3-7B (4443) 和 Xiaomi-7B (4347) 在识别错误数字方面优于 GLM-4V-Flash (3353) 和 DeepSeek-8B (3849)，但 GLM-4V-Flash 的低 TP 可能因其更严格的预测策略。
- **真阴性 (TN)**：GLM-4V-Flash (2350, 0.245) 远超其他模型，表明其在识别非错误数字方面表现最佳；DeepSeek-8B (593) 次之，Qwen3-7B (156) 和 Xiaomi-7B (181) TN 比例极低（<0.02）。
- **假阳性 (FP)**：Qwen3-7B (0.484) 和 Xiaomi-7B (0.481) FP 比例最高，表明模型易将正确数字误判为错误；GLM-4V-Flash (0.255) FP 最低，预测更保守。
- **假阴性 (FN)**：Qwen3-7B (357) FN 最低，Xiaomi-7B (453) 和 DeepSeek-8B (951) 次之，GLM-4V-Flash (1447) FN 较高，可能因其对错误数字的敏感度不足。
- **生成错误**：GLM-4V-Flash (2, 0.000) 生成错误极低，展现极高的生成稳定性；DeepSeek-8B (509, 0.053) 优于 Qwen3-7B (577, 0.060) 和 Xiaomi-7B (634, 0.066)。

结论：GLM-4V-Flash 在准确率、TN 和生成错误率方面表现最佳，适合需要高稳定性和低误判的场景。Qwen3-7B 和 Xiaomi-7B 在 TP 方面较强，但高 FP 和低 TN 限制了性能。DeepSeek-8B 表现均衡，但 FN 较高。

---

3. 按领域（Domain）性能分析

3.1 领域指标

领域	Qwen3-7B Accuracy	Xiaomi-7B Accuracy	GLM-4V-Flash Accuracy	DeepSeek-8B Accuracy	Qwen3-7B Gen. Error	Xiaomi-7B Gen. Error	GLM-4V-Flash Gen. Error	DeepSeek-8B Gen. Error
Numeracy_600K_article_title	0.510	0.508	**0.615**	0.482	0 (0.000)	1 (0.001)	0 (0.000)	0 (0.000)
aclsent	0.509	0.497	**0.606**	0.474	0 (0.000)	8 (0.004)	0 (0.000)	0 (0.000)
DROP	0.364	0.387	**0.494**	0.280	569 (0.227)	550 (0.219)	0 (0.000)	499 (0.200)
qa-text-source-comparison	0.485	0.463	**0.564**	0.468	8 (0.004)	35 (0.018)	2 (0.001)	8 (0.004)
FinNum	0.497	0.468	**0.514**	0.467	0 (0.000)	40 (0.020)	0 (0.000)	2 (0.001)

分析：

- **Numeracy_600K_article_title**：GLM-4V-Flash (0.615) 准确率最高，生成错误为 0；Qwen3-7B (0.510) 和 Xiaomi-7B (0.508) 接近，DeepSeek-8B (0.482) 最低。
- **aclsent**：GLM-4V-Flash (0.606) 领先，Qwen3-7B (0.509) 和 Xiaomi-7B (0.497) 次之，DeepSeek-8B (0.474) 较弱。
- **DROP**：GLM-4V-Flash (0.494) 表现最佳，且生成错误为 0；Qwen3-7B (0.364)、Xiaomi-7B (0.387) 和 DeepSeek-8B (0.280) 准确率低，生成错误高（20%-23%），表明 DROP 数据集的复杂性（如长文本或数字关系）是主要挑战。
- **qa-text-source-comparison**：GLM-4V-Flash (0.564) 领先，DeepSeek-8B (0.468) 和 Qwen3-7B (0.485) 接近，Xiaomi-7B (0.463) 生成错误较高（35）。
- **FinNum**：GLM-4V-Flash (0.514) 最佳，Qwen3-7B (0.497) 次之，Xiaomi-7B (0.468) 和 DeepSeek-8B (0.467) 接近。

结论：GLM-4V-Flash 在所有领域表现最佳，尤其在 DROP 数据集上展现高稳定性和准确性。DROP 是其他模型的性能瓶颈，需针对其高生成错误优化推理逻辑。

---

4. 按错误类型（Error Type）性能分析

4.1 错误类型指标

错误类型	GLM-4V-Flash TP	Qwen3-7B TP	Xiaomi-7B TP	DeepSeek-8B TP	GLM-4V-Flash TN	Qwen3-7B TN	Xiaomi-7B TN	DeepSeek-8B TN	GLM-4V-Flash Gen. Error	Qwen3-7B Gen. Error	Xiaomi-7B Gen. Error	DeepSeek-8B Gen. Error
Error in Number Relationships	149	192	187	161	**79**	3	6	24	0 (0.000)	6 (0.015)	7 (0.018)	4 (0.010)
Undetectable Error	278	445	446	387	**217**	11	7	69	0 (0.000)	18 (0.018)	24 (0.024)	14 (0.014)
Type Error	389	508	495	415	**252**	19	16	82	2 (0.002)	16 (0.013)	23 (0.019)	10 (0.008)
Anomaly	173	219	214	186	**107**	7	14	27	0 (0.000)	14 (0.029)	15 (0.031)	10 (0.021)
Improper Data	18	29	28	27	**11**	1	0	3	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
Factual Error	43	55	52	45	**30**	2	8	5	0 (0.000)	2 (0.016)	3 (0.024)	2 (0.016)

分析：

- **Error in Number Relationships**：GLM-4V-Flash TN (79) 远超其他模型，表明其在非错误数字识别上更强；Qwen3-7B (192) 和 Xiaomi-7B (187) TP 较高，DeepSeek-8B (161) 稍弱。
- **Undetectable Error**：Qwen3-7B 和 Xiaomi-7B TP 最高（约 445），但 TN 低（<11）；GLM-4V-Flash TN (217) 最佳，DeepSeek-8B (69) 次之。
- **Type Error**：Qwen3-7B (508) TP 最高，GLM-4V-Flash (389) TN (252) 领先，DeepSeek-8B 生成错误最低（10）。
- **Anomaly**：GLM-4V-Flash TN (107) 最佳，Qwen3-7B (219) TP 最高，Xiaomi-7B 和 DeepSeek-8B 接近。
- **Improper Data 和 Factual Error**：样本量少，GLM-4V-Flash TN 最高，生成错误极低。

结论：GLM-4V-Flash 在 TN 方面全面领先，适合高可靠性场景；Qwen3-7B 和 Xiaomi-7B 在 TP 上更强，但 TN 低；DeepSeek-8B 表现均衡，生成错误较低。

---

5. 按提示类型（Prompt Type）性能分析

5.1 提示类型指标

提示类型	Qwen3-7B Accuracy	Xiaomi-7B Accuracy	GLM-4V-Flash Accuracy	DeepSeek-8B Accuracy	Qwen3-7B Gen. Error	Xiaomi-7B Gen. Error	GLM-4V-Flash Gen. Error	DeepSeek-8B Gen. Error
few_shot	**0.595**	0.404	0.582	0.548	18 (0.072)	28 (0.112)	**0 (0.000)**	12 (0.048)
zero_shot	0.476	0.475	**0.594**	0.462	559 (0.060)	606 (0.065)	**2 (0.000)**	497 (0.053)

分析：

- **few_shot**：Qwen3-7B (0.595) 准确率最高，GLM-4V-Flash (0.582) 和 DeepSeek-8B (0.548) 次之，Xiaomi-7B (0.404) 最低。GLM-4V-Flash 生成错误为 0，DeepSeek-8B (12) 最低，Xiaomi-7B (28) 较高。
- **zero_shot**：GLM-4V-Flash (0.594) 领先，Qwen3-7B (0.476)、Xiaomi-7B (0.475) 和 DeepSeek-8B (0.462) 接近。GLM-4V-Flash 生成错误极低（2），其他模型生成错误比例较高（5%-6%）。
- **生成错误**：few_shot 模式下生成错误比例高于 zero_shot，可能因提示复杂度增加。

结论：GLM-4V-Flash 在两种提示类型下均展现高稳定性和准确率；Qwen3-7B 在 few_shot 模式下表现突出；DeepSeek-8B 在 few_shot 中优于 Xiaomi-7B，但 zero_shot 性能一般。

---

6. 结论

GLM-4V-Flash 在准确率（0.594）、TN (0.245) 和生成错误率（0.000）方面显著优于 Qwen3-7B (0.479)、Xiaomi-7B (0.472) 和 DeepSeek-8B (0.463)，尤其在 DROP 数据集和复杂操作中展现高稳定性。Qwen3-7B 在 few_shot 模式和 TP 方面表现突出，DeepSeek-8B 性能均衡但 FN 较高，Xiaomi-7B 生成错误率最高。所有模型受高 FP、低 TN 和 DROP 数据集的高生成错误限制。建议优化解析规则、提示设计和推理流程，补充 JSON 数据，并针对 DROP 数据集进行微调。未来可引入更大规模模型（如 Llama 3.1-70B）进一步提升性能。