本地语言模型
中文幽默理解能力
横向对比

学号：2353815
姓名：　　李昊

本地语言模型中文幽默理解能力横向对比

模型选择

本次测试如下模型：

- Deepseek-R1 7b
- Qwen2.5 7b
- Chatglm 6b (int4)

问题设定与评分标准

1. 冬天：能穿多少穿多少；夏天：能穿多少穿多少

- 完善(T)：准确指出两个“能穿多少穿多少”语义完全相反，冬天是“多穿御寒”，夏天是“尽量少穿”。
- 合格(P)：能感知语义差异，但表述模糊或未明确反差。
- 无效(I)：仅重复字面，未识别语义对立。
- 错误(F)：完全无法理解区别或答非所问。

2. 单身狗产生的原因有两个：一是谁都看不上，二是谁都看不上

- 完善(T)：识别“表面重复，实际含义不同”：一个是“自己眼光高”，一个是“别人不看上自己”，并能解释。
- 合格(P)：能识别有幽默反差，但解释不完整。
- 无效(I)：只说“重复”或认为是“写错了”。
- 错误(F)：完全未理解句中反讽。

3. 他知道我知道你知道他不知道吗?

- 完善(T)：能清晰理出逻辑链条，对于问题本身，提问者并不知道他知不知道，指明“我不知道”；但对于提问者所指向的问题，指明“他不知道”。
- 合格(P)：理清大致逻辑，回答“他不知道”。
- 无效(I)：句子没拆明白，语义模糊。

- 错误(F)：错乱理解谁“知道”，回答错误。

4. 明明明明明白白白喜欢他，可她就是不说。

- 完善(T)：准确拆分为“明明 明明明白 白白喜欢他”，解释“明明知道白白喜欢他”，但白白不说。
- 合格(P)：分词正确但主语宾语分析模糊。
- 无效(I)：词拆错或理解错人物关系。
- 错误(F)：完全没读懂句意。

5. 领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导：你这就不够意思了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意思。 领导：那我就不好意思了。 小明：是我不好意思。

- 完善(T)：能逐句解释各“意思”的语用含义（如“没什么意思”=没恶意，“意思意思”=象征性表示，“够不够意思”=仗义与否……），并理解整段的幽默递进。
- 合格(P)：部分解释正确，语境略有偏差。
- 无效(I)：只能解释“意思”的一种或两种常见含义。
- 错误(F)：无法区分“意思”的多种含义。

6. 3.11 和 3.9 谁更大?

- 完善(T)：说明在数学上 3.11 < 3.9 ，且指出出现该问题是因为在软件版本上，3.11 版本相较于 3.9 版本更新。
- 合格(P)：正确在数学上比较大小。
- 无效(I)：答错但意识到可能是陷阱题。
- 错误(F)：坚持“3.11更大”且无逻辑说明。

7. 我室友玩文明6时说脏话，那他还是文明玩家吗?

- 完善(T)：能识别“文明”是游戏名，但句子玩了“文明=有素质”的双关，指出这是语言幽默。
- 合格(P)：意识到双关但没解释清楚。
- 无效(I)：只理解字面，没看出玩笑点。
- 错误(F)：答非所问或答错对象。

8. 爸爸再婚，我是不是就有了个新娘?

- 完善(T)：识别“新娘”用于结婚对象，孩子应称“继母”，指出是幽默或用词误置。
- 合格(P)：理解大意但解释不严谨。
- 无效(I)：模糊理解或只说“有了个新妈妈”。
- 错误(F)：将“新娘”当作孩子的新配偶或其他严重误解。

评测结果

本次测评结果如下（问题测试截图在附件中）：

问题序号	Deepseek-R1 7b	Qwen2.5 7b	Chatglm 6b (int4)
1	F	T	P
2	F	F	F
3	F	P	F
4	I	T	F
5	F	T	F
6	F	F	P
7	I	I	P
8	F	T	P

模型	T（True）	P（Pass）	I（Invalid）	F（False）
Deepseek-R1 7b	0	0	2	6
Qwen2.5 7b	4	1	1	2
ChatGLM 6b	0	4	0	4

逐模型分析

Deepseek-R1 7b

- **优势**：暂无题目被标为 T 或 P，说明没有明确展现出对中文语义歧义或幽默的理解。
- **失误多**：除去 4 和 7 ，语言歧义或反讽全错，表现出对**语用层次、嵌套逻辑、语境幽默的理解较差**。
- **少数无效回答**：说明该模型在处理多义词或角色关系时会**偏离问题核心，生成不相关或混乱答案**。
- **初步结论**：该模型在基础文本生成上流畅，但在中文歧义处理和逻辑推理场景表现弱，适合直译型任务，不适合含混语义或幽默语境。

Qwen2.5 7b

- **表现最佳**：有 4 道题表现优异，表明在语境、歧义词识别、角色指代方面具有较好能力。
- **幽默感识别良好**：尤其是对“意思”多义词和“明明明明”这种语言游戏型问题理解到位。
- **不足**：如 2 题（语义重复但含义不同）和 6 题（数值比较陷阱）仍被标为 F，说明模型**对讽刺或视觉陷阱问题敏感性一般**。
- **初步结论**：该模型是三个模型中**最接近人类语用理解**的，适合处理具有一定语义复杂性的问题。

ChatGLM 6b (int4)

- **中庸偏下表现**：表现并不有意，无 T 标记，但有最多的 P 标记（4 个），说明部分问题虽未不完善，但展示出一定理解能力。
- **4 个 F 显示理解偏弱**：如题 2、3、4、5 多层嵌套/角色推理相关题表现差，说明其**推理深度不足或压缩精度影响语言处理能力**。
- **优势在可接受水平**：对题 1、6、7、8 的回答达到了“Pass”级别，说明该模型适合用于**轻量问答或边缘理解任务**，但不适合复杂语用分析。
- **初步结论**：int4 精度压缩降低模型表达力，对语义细节损失大；但在资源受限场景下仍具备一定实用价值。