

# DeepSeek等大语言模型幽默生成能力及其特征的评测分析

蒋彦廷<sup>1</sup> 应以周<sup>2</sup>

1. 复旦大学 中国语言文学系, 上海, 200433

2. 复旦大学 计算与智能创新学院, 上海, 200433

ytjiang24@m.fudan.edu.cn, yzying24@m.fudan.edu.cn

## 摘要

以人类的笑话文本为基础, 比较评测了4个大语言模型生成幽默笑点句的能力。总的来看, 目前DeepSeek-R1的中文幽默生成能力强于GPT-4o、Qwen2.5-7B和Qwen3模型, 但距离人类的幽默能力还有明显的差距。各模型基于固定表达生成笑点句时, 或多或少存在“思维定势”问题。测查了人类与大语言模型幽默文本的9项语言特征。DeepSeek与人类的相似笑点最多, BLEU-4匹配度也最高。与人类相比, AI生成的笑点句更倾向于使用高频常见的词, 未登录词、网络新词的比例更低, 在长度上普遍更长。基于Sentence-BERT模型获取语义表示, 大模型的笑点句在语义联想距离上普遍比人类的笑点句更短。强化谐音双关、语义双关等修辞手法的运用, 是大模型提高幽默文本生成能力的重要途径。最后, 我们讨论了本文评价方式的优劣, 并展望了增强大模型幽默能力的3个策略: 优化提示工程、构建幽默多模态大模型、在推理中增强幽默文本的可解释性<sup>2</sup>。

**关键词:** 大语言模型 ; 提示学习 ; 幽默生成 ; 评测 ; 语言特征

## Evaluation and Analysis on Humor Generation Ability and Feature of LLMs——Taking DeepSeek as One of Examples

Yanting Jiang<sup>1</sup> Yizhou Ying<sup>2</sup>

1. Department of Chinese Language and Literature, Fudan University, Shanghai

2. College of Computer Science and Artificial Intelligence, Fudan University, Shanghai  
ytjiang24@m.fudan.edu.cn, yzying24@m.fudan.edu.cn

## Abstract

This paper compares and evaluates the ability of 4 Large Language Models (LLMs) to generate humorous punchlines based on human joke texts. Chinese humor generation ability of DeepSeek-R1 is stronger than that of GPT-4o, Qwen2.5-7B and Qwen3, but there is still a significant gap compared with human humor ability. When various models generate punchlines based on fixed expressions, there is a problem of "fixed pattern of thinking" more or less. We explore 9 features of humorous texts from humans and LLMs. DeepSeek has the most punchlines similar to those of humans, and its BLEU-4 is also the highest. Compared with human texts, AI-generated texts tend to use high-frequency common words, have a lower proportion of OOV words and Internet new words, and generally have longer lengths. Based on Sentence-BERT to obtain semantic representations, the humor sentences of LLMs generally have a shorter semantic association distance than that of human humor sentences. Strengthening the

<sup>2</sup>该文的幽默评测的文本与评分数据: [https://github.com/DHfusion/chinese\\_humor\\_generation](https://github.com/DHfusion/chinese_humor_generation)

use of homophonic and semantic puns is an important way for LLMs to improve their ability to generate humorous texts. Last but not least, we discuss the strengths and shortcomings of our evaluation method, and propose 3 strategies to enhance the humor-generation performance of LLMs: 1. Optimizing prompt engineering. 2. Constructing a humorous multi-modal LLM. 3. Enhancing the interpretability of humorous texts in reasoning.

**Keywords:** Large Language Models , Prompt Learning , Humor generation , Evaluation , Language Features

## 1 引言

幽默是人类特有的品质、能力与交际手段，具有丰富的审美特征。文学、社会学、艺术学、语言学、心理学、医学、信息科学等学科都以各自的视角研究分析幽默（刘会文, 1996; 尉万传, 2009）。幽默的创作和传达经常通过语言来完成，幽默也是一种独特的语言艺术。近年来，以ChatGPT、DeepSeek为代表的大语言模型（Large Language Model）掀起了人工智能及其交叉研究的热潮。本文在中文笑话文本上对GPT-4o<sup>2</sup>、DeepSeek-R1(671B)(DeepSeek-AI, 2025)、Qwen2.5(7B)<sup>3</sup>、Qwen3(235B)<sup>4</sup>等大模型的幽默生成能力进行了评测，通过对AI与人类创作的笑点文本的人工评分、词汇、句子、修辞等多维语言特征的统计分析，分析比较大模型与人类的幽默生成能力与风格。并提出了进一步提高大模型幽默生成能力的3个策略。

## 2 相关研究

在语言模型的生成与理解能力的评测分析方面，学术界和工业界已对其隐喻理解（张声龙等, 2024）、文本摘要（张华平等, 2023）、讽刺与反语识别（刘海涛等, 2024; Strachan et al, 2024）等多项能力做了评测与研究。其中在AI模型的幽默能力方面，学界主要从模型识别与理解幽默、生成幽默2个角度展开研究。

从幽默识别与理解角度，徐琳宏等（2018）综合了语义透明度、语义相关度、语音扩展性和语法特征集4个维度的语义关系，以识别谐音双关语。Chen等（2023）为预训练语言模型设计了中文的幽默识别、幽默类型识别、幽默水平识别、笑点（Punchline）识别4项评测任务，He等人（2024）评测了包括GPT-4o、ERNIE、Gemini、GLM、Qwen在内10个大模型的笑点解释能力，发现中文的幽默识别理解对各种语言模型构成了重大挑战，模型的幽默识别理解表现普遍不佳。Liu等人（2018）发现幽默文本的一些句子特征，如更简单的词汇、更复杂的句法结构有助于AI模型提高幽默识别的效果。

从幽默生成角度，谭红叶等（2018）、Li等人（2023）分别探索了生成对抗网络（GAN）以及T5、GPT-3等大模型的幽默文本生成效果。Yu等人（2020）基于词汇约束重写的方法以生成谐音双关语。

在大模型与人类语言风格的比较方面，Zhu等人（2023）在6500多篇分别由人类和ChatGPT作答的文本上，从描述性特征、字词常用度、字词多样性、句法复杂性、篇章凝聚力五个维度159项特征，对比了AI生成语言与人类语言，发现AI生成的语言常常分段阐述，篇幅更长，比人类更倾向于使用大词和书面语，在句法结构复杂性方面总体高于人类语言。

综上所述，目前研究者已从AI语言模型的幽默识别与理解、幽默生成，以及AI语言模型与人类语言风格的比较方面开展了充分了探索。目前的研究还有进展的空间：(1) 既往对模型生成的幽默文本的评测指标通常采用F1分数、BLEU值、acc指标等，以人类文本为标准答案。而评测幽默能力应以好笑、有趣程度为标准，不宜设置唯一标准答案。(2) 随着大模型DeepSeek-R1、Qwen3的问世，亟待人们评测它们模型的幽默生成能力。(3) 目前还少有人从细致的语言

©2025 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

<sup>2</sup><https://openai.com/index/introducing-4o-image-generation>

<sup>3</sup><https://github.com/QwenLM/Qwen2.5>

<sup>4</sup><https://github.com/QwenLM/Qwen3>

特征和修辞艺术入手，比较、分析不同大模型的幽默生成能力与风格。而这对于深化对幽默艺术的研究、改进优化大模型的幽默生成能力，都具有重要意义。

### 3 语料数据来源与结构化预处理

本文的笑话语料数据来自百度贴吧“弱智吧”<sup>5</sup>。该贴吧的幽默笑话由众多网友创作并分享，分为年度佳贴、吧主推荐和一般共3种类型。Bai等人（2024）曾利用该贴吧的幽默笑话数据，训练了大模型的语言能力。我们从年度佳贴、吧主推荐类里选取了单句式和单轮对话式笑话。既有研究已表明，高质量短笑话中引人发笑的笑点句或笑点词通常位于文本的末尾（谭红叶等，2018）。我们认为，这是受语言省力原则支配的结果，否则笑点句（词）后面的文本就有画蛇添足之嫌。因此，我们在分析观察众多笑话结构的基础上，对其做了结构化的预处理（Pre-Processing），将每个笑话拆分为前段、笑点2部分：如果笑话的句子数量大于1，那么将最后一个句子作为笑点部分，笑点前面的句子都作为前段部分。如果笑话的句子数量等于1，那么进行人工检查与拆分。以笑话中开始明显产生幽默效果的词语w为边界，w及w之后的文本作为笑点部分，w之前的文本作为前段部分。

经过人机协同的方式，我们采集、筛选、结构化处理笑话数据，去除了低俗的、笑点过于小众的笑话，最终得到200个笑话文本。它们都分为前段、笑点两部分文本。我们将前段文本提供给大语言模型，并删去原始的笑点文本，让大模型根据前段文本，尽可能生成幽默好笑的末尾笑点句。为保证文本生成的速度与评测条件统一，各模型均不开启深度思考推理模式。

### 4 大模型幽默生成能力的评测

对于每一个大模型，我们均编写统一的提示语模板（Prompt Template），使其根据笑话的前段，生成幽默有趣的末尾笑点句。模板样例请参看本文附录。

我们的提示语模板遵循了以下原则：

- (1) 少样本学习（Few shot learning），给出少数几个输入输出的样例，供AI模型模仿，有助于改善模型的效果。
- (2) Json的输出格式。即采用键（key）值（value）对的Json字符串格式输出。这可以提高模型输出文本的结构化程度。输出后，即可使用Python语言的Json库，结构化提取模型的笑点文本，减少人类整理文本的工作量，便于后续统计分析。

在各模型给每个笑话生成末尾笑点句后，中国语言文学类专业的3名研究者在不知道笑点句由具体哪个模型生成的情况下，对各模型生成的文本的幽默程度评分。评分规则是：

- 2分：幽默，且幽默程度超过了人类的原版笑点句。
- 1分：幽默，且幽默程度与人类的原版笑点句大致相当。
- 0分：幽默程度与人类的原版笑点句相比，明显较低。
- 1分：完全不幽默，或者内容混乱，无法理解。

当前2人的评分不一致时，引入第3人决定其中一个分数作为最终得分。前两名研究者评分的组内相关系数（ICC）情况如下：双向随机模型ICC2为0.7669，双向混合模型ICC3为0.7840。评分信度良好。

各模型的笑点句得分情况如表1所示。由表1可知，在生成幽默文本（得2分与1分）的比例方面，DeepSeek最高，生成的幽默文本占比59%，而GPT-4o、Qwen2.5-7B和Qwen3分别为43.5%、24.5%和52.5%。由此可见，DeepSeek的中文幽默生成能力在4个模型里最佳。

模型	得2分的数量	得1分的数量	得0分的数量	得-1分的数量	总分
GPT-4o	7	80	94	19	75
Qwen2.5-7B	5	44	68	83	-29
Qwen3-235B	9	96	78	17	97
DeepSeek-R1(671B)	12	106	65	17	113

Table 1: 各模型生成的笑点句得分情况

各模型生成的部分笑点句与人类的笑点句如表2所示。

<sup>5</sup><https://github.com/Leymore/ruozhiba>

笑点句来源/笑话前段	半径越大	你又没有 吃过云	酒桌文化
GPT-4o	面积就越大，钱包就越小	那你怎么知道它不好吃？	酒桌上的文化课
Qwen2.5-7B	包容心越强	那你吃空气试试？保证大饱口福！	就是把饭局变成了一场豪饮大赛
Qwen3-235B	圆的面积就越大，但别忘了，周长也会跟着‘膨胀’，所以减肥不只是你的事，是整个几何界的难题。	但我吃过棉花糖，那不就是地上的云嘛	这杯白酒喝了就是副业，不喝就要离职
DeepSeek-R1	派越大	吃过啊，每次坐飞机我都点一份云吞！	酒桌上没有文化，只有酒精
人类	直径越大	怎么知道那不是棉花糖	又称逼你干杯

Table 2: 各模型生成的部分笑点句与人类的笑点句比较

在表2的3个笑话例子里，DeepSeek生成的笑点句都可圈可点。第1个例子里，对于“半径越大”这一笑话的前段，DeepSeek生成的笑点句为“派越大”。这里“派”的谐音可以是圆周率 $\pi$ ，当人们疑惑 $\pi$ 这一常数为何会变大时，又意识到“派”还可以指馅饼（pie）。DeepSeek生成了一种看似荒诞实则合理的幽默感。第2个例子里，DeepSeek将小吃“云吞”新奇地解释为“吃云”，使之语义双关。第3个例子里，DeepSeek用“没有文化”表达了对酒桌上劝酒、谄媚、铺张浪费等行为的辛辣讽刺。上述DeepSeek回复的幽默程度显著超过了另外3个模型，甚至超过了人类的笑点句。不过，即使是4个模型里表现最好的DeepSeek，在表1中仍然有41%的文本得了0分（幽默程度较低）或-1分（完全不幽默或不知所云）。因此，大模型的幽默生成能力还有较大的提升空间。

此外，我们还注意到了面对格言、俗语、古诗词等固定表达时，各模型或多或少存在思维定势问题。如表3所示。输入固定表达的前段，ChatGPT和Qwen2.5都循规蹈矩，生成了固定搭配的后段，无法出人意料，因此幽默程度不高。这可能是各模型在海量语料上预训练时，见到了太多这样固定的搭配，大模型内部的Transformer编码解码器自然地将固定搭配的前段、后段紧密地关联起来。而DeepSeek受思维定势的束缚相对较少，能创新出奇。Qwen3虽然也在一定程度上摆脱了思维定势，但生成的笑点句略显冗长。

笑点句来源/笑话前段	杀不死你的东西	欲把西湖比西子	秀才遇上兵
GPT-4o	只会让你变得更强	浓妆淡抹总相宜	有理说不清
Qwen2.5-7B	只会让你变得更大	浓妆淡抹总相宜	有理说不清
Qwen3-235B	最终会让你变成超级赛亚人，只不过这个过程可能需要一点骨折的勇气	淡妆浓抹总吃不饱，因为西湖边的美食太多啦	有理说不清，只好掏出计算器算了一卦
DeepSeek-R1	会让你请假	结果西湖说：我不化妆	有理说不清，因为兵听不懂文言文
人类	只会让你活着	约分得到湖比子	像是一部言情小说的标题

Table 3: 各模型生成笑点句时“思维定势”的例子

## 5 大模型幽默文本的特征分析

在对幽默文本进行特征统计分析前，我们使用哈尔滨工业大学社会计算与信息检索中心训

练习开发的语言技术平台（LTP 4）模型（Che et al, 2021）对文本进行自动分词。我们依据7项语言统计指标和2项双关修辞指标，对比分析人类与AI大模型的幽默文本特征。

指标1：AI与人类笑点句相似的个数。我们人工观察、对比人类与大模型生成的笑点句，标记笑点句完全相同或极为相似的个数。

指标2：AI与人类对应笑点句的BLEU-4的平均值。BLEU(Bilingual Evaluation Under-study)指标将计算AI笑点句与人类笑点句的N-gram重叠程度，用于量化衡量AI与人类笑点句在字面上的相似度。

指标3：笑点句的平均词频排位。我们参考了北京语言大学BCC语料库的词频表（荀恩东等, 2016）。该词频表在海量中文语料上统计了180多万个词和固定词组的出现频次及其排位（频次相等的词语视为排位相等）。

指标4：笑点句未登录词（Out of Vocabulary, OOV）的平均比例。我们将未包含在BCC语料库词频表的词语视为未登录词。

指标5：人类与AI笑点句的平均长度（包括汉字、数字和标点）。

指标6：人类与AI笑点句的词汇丰富度，定义为笑点句的词语种类数除以词语数。

指标7：我们定义“笑点句的平均联想距离”这一指标。它指的是笑点句与笑话前段的文本向量之间的平均余弦距离（Cosine Distance）。而文本向量则通过适用于文本相似度计算的、支持多语言的multilingual Sentence-BERT(SBERT)预训练模型(Reimers, 2019)获取。

人类和AI生成的笑点句的上述7项特征指标如表4所示。

笑点句来源	与人类笑点句相似的个数	与人类笑点句的平均BLEU-4	平均词频排位	未登录词平均比例	平均长度	词汇丰富度	联想距离
人类	—	—	4247	6.3%	6.21	0.547	0.718
Qwen2.5-7B	6	0.0301	3436	6.0%	7.62	0.461	0.682
Qwen3-235B	9	0.0289	2918	5.0%	26.1	0.384	0.752
GPT-4o	6	0.0276	2793	5.1%	7.74	0.442	0.669
DeepSeek-R1	11	<b>0.0547</b>	2772	4.5%	9.74	0.392	0.635

Table 4: 人类与AI笑点句的统计特征

在人类与AI笑点句的相似个数方面，GPT-4o与Qwen2.5-7B与人类笑点句相似的数量均为6，而Qwen3和DeepSeek的这一数值分别为9和11。这些与人类相似的笑点句，可能是模型已从大规模互联网语料中学得的，也可能是不曾学得、与人类“不谋而合”的。去除这些与人类相似的笑点句后，DeepSeek生成的较为幽默的笑点句仍然有107个。可见现在的AI模型已初步具备全新幽默文本的能力。

在人类与AI对应笑点句的BLEU-4分数方面，虽然4个模型的得分都不高，但DeepSeek的BLEU-4分数明显高于其他3个模型。DeepSeek的笑点句与人类笑点句在字面上相对更接近一些。

在笑点句的平均词频排位方面，依据BCC语料库的词频表，人类笑点句的平均词频排位为4247，而4个大模型的这一指标均在2700-3500位之间。这说明与人类相比，大模型生成的笑点句更倾向于使用高频常见的词。

在笑点句的未登录词平均比例方面，人类笑点句的未登录词比例约6.3%，高于4个大模型的笑点句。据观察，人类笑点句的未登录词包含了一些随着社会发展新产生的词，如“emoji（网络中的视觉情感符号）”“搜狗（信息科技企业的名称）”“钉钉（网络办公软件）”“三国杀（一种卡牌游戏）”。虽然AI生成的笑点句里也包括“网课”“动漫”“快闪”等新词，但总体的比例不如人类的笑点句高。

在笑点句的平均长度方面，4个大模型生成的笑点句平均字符数均高于人类笑点句的平均字符数6.21。其中Qwen3笑点句的平均字符数明显高于其他大模型的数据，达到26.1。这说明大模型比人类更倾向于生成更长的笑点文本。

在笑点句的词汇丰富度、平均联想距离方面，除Qwen3外的3个大模型笑点句的词汇丰富度、平均联想距离均不及人类笑点句。其中DeepSeek笑点句的词汇丰富度、平均联想距离都最低。Qwen3的联想距离较高，可能是因为其生成的笑点句长度远高于人类和其他模型的笑点句。值得注意的是，笑点句的词汇丰富度、联想距离和笑话的幽默程度之间并无必然联

系：虽然Qwen3笑点句的词汇丰富度最低，但其模型生成的笑点句质量仅次于DeepSeek；虽然DeepSeek笑点句的联想距离最低，但该模型却是4个模型里生成笑点句质量最高的。

我们还人工审视、统计了4个大模型的笑点句对“双关”这一修辞手法的运用情况。双关是使语句同时关顾两种意思的修辞手法，分为谐音双关、语义双关2大类。前者主要利用字词音同音近的条件；后者则主要利用词汇的多义性（黄伯荣与廖序东，2017）。如表5所示。

笑点句来源	谐音双关频次	语义双关频次
人类	11	44
Qwen2.5-7B	0	5
Qwen3-235B	1	6
GPT-4o	1	9
DeepSeek-R1(671B)	5	13

Table 5: 人类与AI在笑点句使用双关修辞的频次

如表5所示，4个大模型的笑点句运用谐音、语义双关的频次均远不及人类。相较而言，大模型里DeepSeek使用上述修辞手法略多一些，除了第4节中提到的“云吞”“派”的例子外，DeepSeek使用双关手法的其他例子如下（加粗的文本为DeepSeek生成的内容）：

谐音双关：

- (1) 大气里最多的气体是氮气，所以我们都生活在“氮”定中。
- (2) 江洋大盗从来不敢偷茶叶店，因为他怕查（茶）到。
- (3) “忘了我吧。”我对旺旺仙贝说。

语义双关：

- (1) 姚明宣布进入文坛，这下文坛终于有了新“高度”。
- (2) 语文老师告诉我作文结尾最好要有升华，于是我决定把作文烧了。
- (3) 擅长苦中作乐的小明，终于在苦瓜里找到了甜味。

谐音双关的3个例子里，DeepSeek运用了“淡”与“氮”、“查”与“茶”、“旺”与“忘”之间的谐音关系产生幽默感。在语义双关的3个例子里，DeepSeek领悟到了“高度”“升华”“苦中作乐”等词语既有隐喻抽象义，又有字面的物理化学意义。总的来看，大模型已能初步尝试双关修辞来创造幽默文本。进一步增强大模型理解、运用修辞手法的能力，是未来提高大模型幽默生成能力的重要途径。

## 6 讨论

首先，是对本文评测方式的讨论。本文采取人类阅读笑话文本，赋予幽默度分数的方式，以评价大模型生成的笑点句质量。这虽然避免了设置唯一标准答案，具有一定的灵活性，但另一方面也苛求了模型：由于实验数据源自人类原创的笑话，人类创作的笑话不仅包括笑点句，而且包括笑话的前段文本。这相当于人类既是幽默笑话创作竞赛的命题人，又是答题人。如果让AI模型同时充当命题人和答题人，AI生成若干笑话，再让人类根据这些笑话的前段文本，续写笑点句。在这种角色转换的情况下，人类的幽默生成能力也未必比本文第4、5节中的大模型更好。

其次，是增强大模型幽默能力的策略。如前人（尉万传，2009）所述，幽默其实是一种涉及多种表达手段的艺术。在语音手段方面，除了谐音双关，还包括语音仿拟、利用方言、巧用重音与停顿等；在词汇手段方面，除了语义双关，还包括词语跨域、词语逆序、感情色彩变异等；此外，还包括句法歧义等句法手段、偷换混淆概念等逻辑手段，甚至涉及听众的文化背景、神态动作等超越言语的因素。基于此，我们认为未来可以在三个方面着手增强大模型的幽默理解与生成能力：

(1) 优化提示工程（Prompt Engineering）。在撰写AI生成幽默文本的提示语时，提供多样本、多类型的幽默文本的范例。范例不仅需要有一定的数量，而且需要涵盖上述的语音、词汇、句法、逻辑等方面幽默手段。

(2) 结合文本、语音、图像视频资源，构建多模态的幽默大模型。如今文娱生活日益丰富，网络上也出现了许多相声、喜剧、脱口秀节目的音视频资源。这些节目的幽默文本大多贴近当下的社会生活和文化环境，更容易引发人们的共鸣。此外，这些音视频资源往往带有节目

中观众的笑声，是天然的具备笑点标签的有监督数据。这些多模态的训练数据能让模型超越单纯的文本，拓展其幽默语音生成等能力。例如，在CCAC 2025第五届中国情感计算大会的技术评测中，已有研究团队推出了文本、音频、视频多模态的中文幽默讽刺数据集<sup>6</sup>，还有团队推出了四格漫画理解评测的中文多模态幽默数据集<sup>7</sup>。上述工作都是有意义的基础性工作。

(3) 增强幽默文本的可解释性。虽然互联网上的幽默文本规模已经较为可观，但还缺乏解释笑话为何幽默的文本。例如春节晚会的一个小品中，演员向谈话节目主持人说，在家庭争端中，自己的妻子和儿女联合起来要“弹劾”自己，引得观众欢笑。此处的“弹劾”之所以幽默，正是因为利用了词语跨域的手段：在日常的家庭生活中，使用“弹劾”这样政治、法律领域严肃的术语，造成了语域文体的差异，取得了新奇幽默的效果。以DeepSeek-R1和OpenAI的o1模型为代表，大语言模型已进入推理模型阶段。通过强化学习(Reinforcement Learning)训练大模型时，提供这样的解释幽默成因的文本，有助于推理模型更好地理解与生成幽默文本。

## 7 结语

本文以200个人类幽默笑话为例，结合人工分析与自动计算，对比评测了4个AI大模型生成笑点句的能力及其特征。总的来看，目前DeepSeek-R1的中文幽默生成能力强于GPT-4o、Qwen2.5-7B和Qwen3-235B模型，但距离人类的幽默能力尚有明显差距。面对固定表达时，大模型或多或少存在循规蹈矩的思维定势问题。目前大模型只能勉强辅助相声、脱口秀、喜剧创作者的创意工作。在大模型幽默文本的特征方面，与人类的笑点句相比，大模型生成的笑点句高频词运用得更多，未登录词更少，长度普遍更长。除了生成笑点句明显较长的Qwen3之外，其他大模型生成的笑话文本在联想距离上普遍更短。各模型在谐音双关、语义双关修辞手法的使用频率上还不及人类。最后，我们认为，优化提示工程、利用音视频资源构建多模态幽默大模型、增强幽默文本的可解释性是未来提高大模型幽默生成能力的3个可尝试的途径。文章的局限在于，囿于有限的评分者人力，还未能在更大规模的笑话文本上探索更多大模型的幽默生成能力。在未来工作中，我们将在更大规模的数据上，全面验证上述3项优化大模型幽默能力的策略。

## 参考文献

- 黄伯荣, 廖序东主编. 2017. 现代汉语（增订六版）下册. 北京: 高等教育出版社, 202-204.
- 刘海涛, 亓达. 2024. 大语言模型的语用能力探索——从整体评估到反语分析. 现代外语, 47(04):439-451.
- 刘会文. 1996. 幽默语言的审美特征与社会功能. 广西社会科学, (01):109-110.
- 谭红叶, 闫真, 李茹, 等. 2018. 迈向创造性语言生成: 汉语幽默自动生成的探索. 中国科学: 信息科学, 48(11):1497-1509.
- 尉万传. 2009. 幽默言语的多维研究. 浙江大学博士学位论文.
- 徐琳宏, 林鸿飞, 祁瑞华, 等. 2018. 基于多维语义关系的谐音双关语识别模型. 中国科学: 信息科学, 48(11):1510-1520.
- 荀恩东, 饶高琦, 肖晓悦, 等. 2016. 大数据背景下BCC语料库的研制. 语料库语言学, 3(01):93-109+118.
- 张声龙, 刘颖. 2024. 基于类型知识提示的隐喻理解方法. 清华语言学, (00):271-290.
- 张华平, 李林翰, 李春锦. 2023. ChatGPT中文性能测评与风险应对. 数据分析与知识发现, 7(03):16-25.
- Bai Y, Du X, Liang Y, et al. 2024. COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning. arXiv e-prints.2403.18058.
- Che W, Feng Y, et al. 2021. N-LTP: An Open-source Neural Language Technology Platform for Chinese. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*: 42-49.

<sup>6</sup>[github.com/MuYuU0/CCAC2025-Chinese-multimodal-sarcasm-calculation](https://github.com/MuYuU0/CCAC2025-Chinese-multimodal-sarcasm-calculation)

<sup>7</sup>[github.com/DUTIR-Emotion-Group/CCAC2025-task3](https://github.com/DUTIR-Emotion-Group/CCAC2025-task3)

- Chen Y, Li Z, Liang J, et al. 2023. Can pre-trained language models understand Chinese humor?. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining(WSDM 2023)*: 465-480.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv e-prints.2501.12948.
- He R, He Y, Bai L, et al. 2024. Chumor 2.0: Towards Benchmarking Chinese Humor Understanding. arXiv e-prints.2412.17729.
- He R, He Y, Bai L, et al. 2024. Chumor 1.0: A Truly Funny and Challenging Chinese Humor Understanding Dataset from Ruo Zhi Ba. arXiv e-prints.2406.12754.
- Junhui Zhu, Mengyan Wang, Erhong Yang, et al. 2023. A Comparative Study of Language between Artificial Intelligence and Human: A Case Study of ChatGPT. *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (CCL 2023)* : 523–534.
- Li J, Wu X, Liu X, et al. 2023. Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics(ACL 2023)*: 7581–7596.
- Liu L, Zhang D, Song W. 2018. Exploiting Syntactic Structures for Humor Recognition. *Proceedings of the 27th International Conference on Computational Linguistics(COLING 2018)*: 1875-1883.
- Reimers N, Gurevych I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*: 3982-3992.
- Strachan J W A, Albergo D, Borghini G, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285-1295.
- Yu Z, Zang H, Wan X. 2020. Homophonic Pun Generation with Lexically Constrained Rewriting. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP 2020)*: 2870–2876.

## 8 附录：本文所用的提示语模板

Prompt参数	参数值
System	<p>你是一名幽默文本的创作者。请你根据给定的未完成的文本，一步步思考，补充文本的笑点句，尽可能幽默搞笑。</p> <p>我会输入一个句子，你需要回答一个json字符串，其中包含2个字段：</p> <ol style="list-style-type: none"> <li>1. reply: string 类型，表示对已有文本的续写或回复，一定要幽默搞笑。</li> <li>2. reason: string 类型，简要解释续写回复搞笑幽默的原因。</li> </ol> <p>下面会给你一些输入输出样例，请模仿这些样例，进行回答。</p> <p>Input: 这么早就想睡觉了  Output: {"reply": "做梦去吧", "reason": "‘做梦去吧’双关，既指想睡觉的人痴心妄想，又可指睡觉进入梦乡。"}</p> <p>Input: 我的梦想是做一个潇洒自在的人，现在终于完成了一半  Output: {"reply": "做了一个洒自", "reason": "回复取了“潇洒自在”的一半文字“洒自”，谐音就是傻子。这是一种幽默的自嘲。"}</p> <p>Input: 我炸书  Output: {"reply": "周围就会充满了知识分子", "reason": "利用多音字“分”，将“知识分子”中的分读为一声，表示化学中的分子。"}</p> <p>Input: 向日葵能否成为星际时代的  Output: {"reply": "指南针", "reason": "因为向日葵总向着太阳，在宇宙中航行，向日葵就能辨别太阳的方向。这是一种幽默新奇的想象。"}</p> <p>接下来是正式的输入，请按上述格式，回答一个json字符串，不要有多余的内容，回答完json字符串后终止回答。</p>
User	

Table 6: 本文所用的提示语模板