



VALLEY: VIDEO ASSISTANT WITH LARGE LANGUAGE MODEL ENHANCED ABILITY

Ruipu Luo^{1,2*}, Ziwang Zhao^{1,3*}, Min Yang^{1*}, Junwei Dong^{1,4}, Da Li¹,
Pengcheng Lu¹, Tao Wang¹, Linmei Hu⁵, Minghui Qiu^{1,†}, Zhongyu Wei²

¹ByteDance Inc. ²Fudan University ⁴Chongqing University

³Beijing University of Posts and Telecommunications

⁵Beijing Institute of Technology

{luoruipu, zhaoziwang, yangmin.priv, dongjunwei}@bytedance.com





A. 之前讲的 LLaVA 是使用图文对数据进行了图文场景多模态能力的扩展

B. 本文提到的 Valley 则是字节发布的视频场景多模态指令微调 LLaMA 模型

两者模型区别不大，主要是数据和训练的差异！





- 主要贡献：
- 模型：基于LLaVA的方法，添加了时空池化模块应对视频（多帧）场景，将LLaVA从单图扩展为多图（动态长度），同时将LLaVA的Vicuna语言模型换为Stable-Vicuna模型。；
- 数据：搞了多模态的instruction-following数据集，聚焦于视频理解、比较多样的任务（包括multi-short captions，带时间戳的时间性描述、长视频的复杂陈述）。同时使用了ChatGPT生成对话人和视频内容的对话，进一步增强数据集的质量和多样性。
- 开源：LLM开源





多帧的处理通过时空化模块：

每帧的表示：

$$V_T = [V_T^{[CLS]}, V_T^1, V_T^2, \dots, V_T^{256}].$$

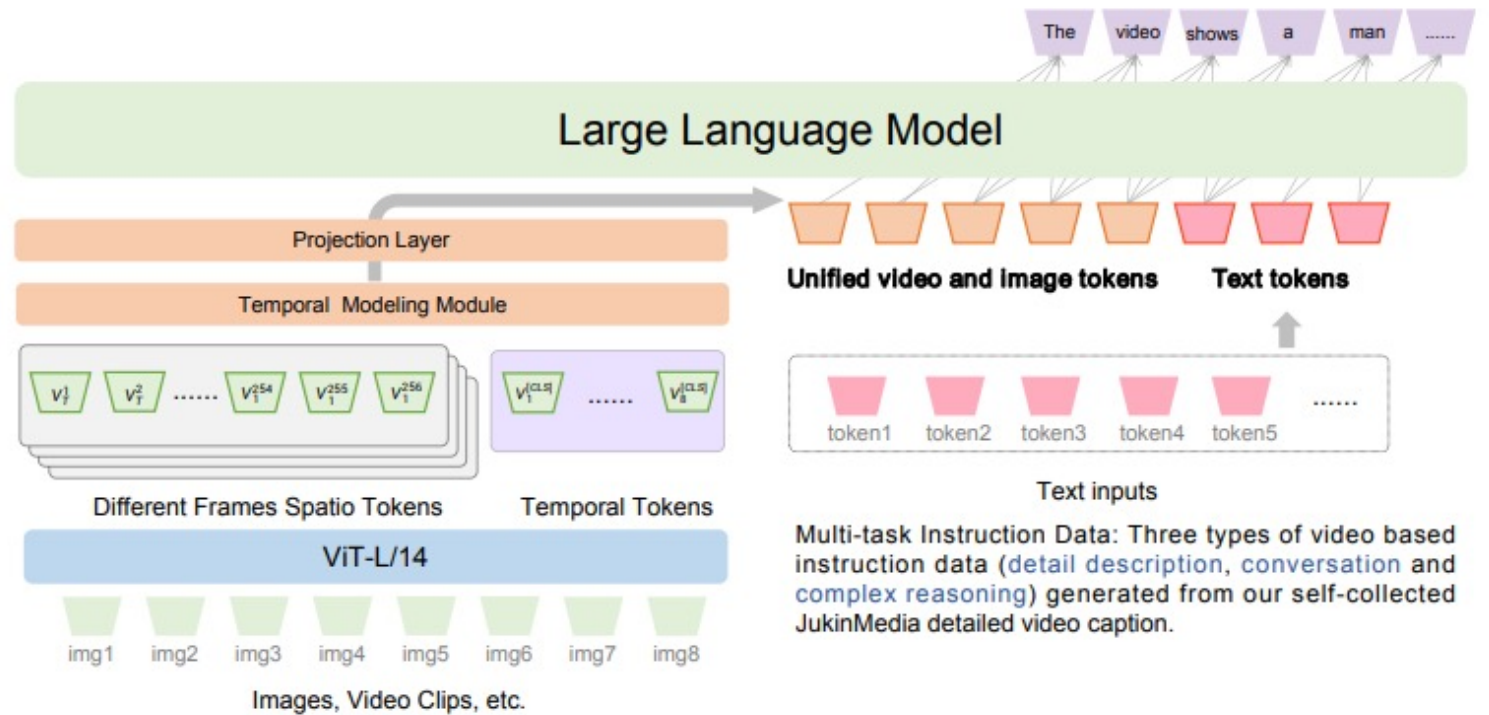
视频的空间特征：

$$\hat{V}^i = \text{AVGPOOLING}([V_1^i, V_2^i, \dots, V_T^i]),$$

视频的时间特征是T个V[CLS]的拼接

最终视频表示：

$$Z_V = [\bar{V} \oplus V_1^{[CLS]} \oplus V_2^{[CLS]} \oplus \dots V_t^{[CLS]}],$$



Multi-task Instruction Data: Three types of video based instruction data ([detail description](#), [conversation](#) and [complex reasoning](#)) generated from our self-collected JukinMedia detailed video caption.

Figure 1: Valley architecture.



video-text pair


• 指令微调数据收集

- 类似于LLaVA，作者基于MSRVTT(10k)、VATEX(22k)、ActivityNet(10k)、VIOLIN(5.8k)等数据集也构建了视频多模态指令数据，包含3种上下文类型，并利用三种类型去生成指令数据。

• 训练


- 同LLaVA类似的两阶段训练方式，第一阶段通过预训练映射层来进行特征对齐；第二阶段再微调语言模型和映射层。

Context type 1: Multi-Short-Captions




(1) Two people abseil down a rope suspended from a cage hung below an extended arm of a crane.
(2) Two men repel down lines from a crane high in the air.
(3) A couple of men are on a device in the sky and jump out on a string.
(4) two people are gliding down a string from a giant crane.
(5) Two people are quickly lowered to the ground from a tall crane.
(6) Two people are going down with harnesses from a crane.
(7) Two people using ropes lower themselves down from a high crane.
(8) Two people repelling down from a crane while someone plays the drums in the background.
(9) Two men are abseiling from the top of a crane.
(10) Two people use a wire to slide down to the ground from a box suspended by a crane.

Context type 2: temporal description with timestamps



[0s - 7.49s] : We see a hallway with a wooden floor.
[7.49s - 18.09s] : A dog in socks walks slowly out onto the floor as a lady films him.
[19.37s - 36.55s] : The dog turns around and goes back to the other room.

Context type 3: Complex Statement with Long Video



(1) The man in the red shirt was really proud to see the girl in the blonde come out of her shell and have fun.
(2) The blonde lady wants to know if man in the red shirt loves his new style.
(3) The man in the red shirt explains to the blonde lady that he is in love with her and it doesn't matter if she has blonde hair or makeup on.



Conversation for Context type1

[User]: What are the people in the video doing?
[AI]: The people in the video are abseiling down a rope from a crane. They are suspended from a cage that is hung below an extended arm of the crane.
[User]: How many people are in the video?
[AI]: There are two people in the video.
[User]: How are the people in the video descending from the crane?
[AI]: The people in the video are using ropes and harnesses to lower themselves down from a high crane.
[User]: Is there any background music in the video?
[AI]: Yes, there is background music in the video. Someone is playing the drums in the background while the people are repelling down from the crane.
[User]: Can you describe the device that the people are using to descend from the crane?
[AI]: Based on the video, the device the people are using to descend from the crane appears to be a cage. The cage is suspended from an extended arm of the crane and has a rope attached to it. The two people are using the rope to lower themselves down from the cage.

Conversation for Context type2

[User]: What is in the video?
[AI]: The video depicts a hallway with a wooden floor. At around 7.5 seconds, a dog wearing socks walks slowly onto the floor as a lady films him.
[User]: How many dogs are in the video?
[AI]: There is only one dog present in the video.
[User]: What does the dog do in the video?
[AI]: The dog walks slowly onto the wooden floor wearing socks while being filmed by the lady. After a short time, at around 19 seconds, the dog turns around and goes back to the other room.
[User]: Why is the dog wearing socks in the video?
[AI]: It's possible that the dog is wearing socks to prevent scratches on the wooden floor or to provide better traction for its paws. Alternatively, the dog may be wearing the socks as a fashion statement or for some other purpose not related to the floor.

QA Pairs for Context type3

Question: Why was the man in the red shirt proud? (Casual Inference)
Answer: The man in the red shirt was proud because he saw the girl in the blonde come out of her shell and have fun.
Question: Who wants to know if the man in the red shirt loves his new style? (Character Identifying)
Answer: The blonde lady wants to know if the man in the red shirt loves his new style.
Question: What does the man in the red shirt explain to the blonde lady? (Video Understanding)
Answer: The man in the red shirt explains to the blonde lady that he is in love with her and it doesn't matter if she has blonde hair or makeup on.



局限性：缺少音频模态、多语言理解能差、幻觉问题

测试案例



What is this person doing in the video?



In the video, a person is preparing a meal by pouring a liquid, possibly milk or juice, into a small metal bowl. They are using a spoon to carefully pour the liquid into the bowl, which is placed on a wooden surface. The person is standing near a sink, which suggests that they might be in a kitchen or a similar environment where cooking takes place. The video provides a close-up view of the person's hand and the bowl, allowing us to see the details of the preparation process.



The liquid in the picture is egg liquid, and the one in the plate is meat. Please modify your description according to this fact



Based on the updated information, it appears that the person is pouring egg liquid into a small metal bowl with meat inside. The video still shows a close-up view of the person's hand and the bowl, and the preparation is taking place on a wooden surface. This suggests that the person is likely cooking or preparing a meal, possibly involving both egg and meat ingredients



養天地正氣 法古今完人

Adapting Fake News Detection to the Era of Large Language Models

Jinyan Su¹, Claire Cardie¹, Preslav Nakov²

¹Department of Computer Science, Cornell University

²Mohamed bin Zayed University of Artificial Intelligence

{js3673,ctc9}@cornell.edu, preslav.nakov@mbzuai.ac.ae





• 为了适应大型语言模型时代，下一代假新闻检测器应能够处理**人类/机器生成的真/假新闻的混合内容环境**。

1) 机器生成的真新闻 (MR)

2) 机器生成的假新闻 (MF)

3) 人类生成的真新闻 (HR)

4) 人类生成的假新闻 (HF)





随着人类撰写和机器生成内容之间的动态变化，作者为了综合评估，考虑了三种实验设置：

仅适用人类撰写的真实新闻作为真实类别的训练数据；

并逐步引入机器生成的假新闻到假的类别中（比例从1-100%）

未来由机器主导，即真实新闻类别的训练数据完全来源于机器生成的真实新闻文章

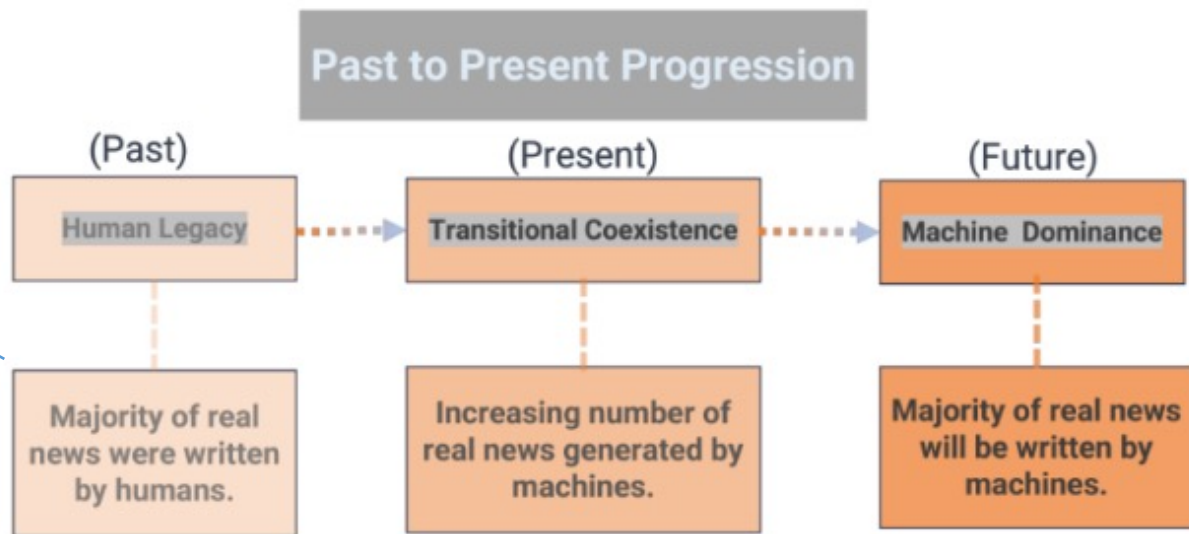
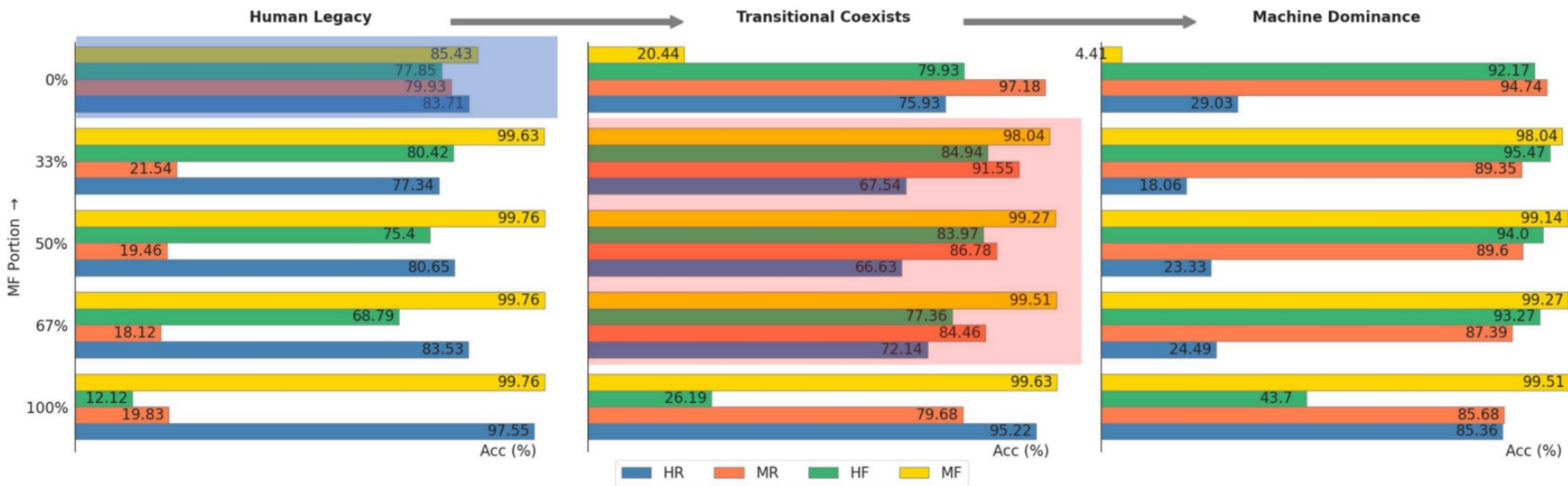


Figure 1: Our three experimental phases: (*Human Legacy*, *Transitional Coexists*, and *Machine Domi-nance*) based on real news generation sources.

真实类别的新闻中由人类和机器混合生成



下图是在GossipCop++数据集上训练大型RoBERTa模型并进行测试所获得的性能度量结果：





在初期，真实类别由人类撰写的情况：

当假的类别也由人类生成时，各类检测精度均衡，效果还不错，但现实部署中，无法确定多少新闻是机器生成的

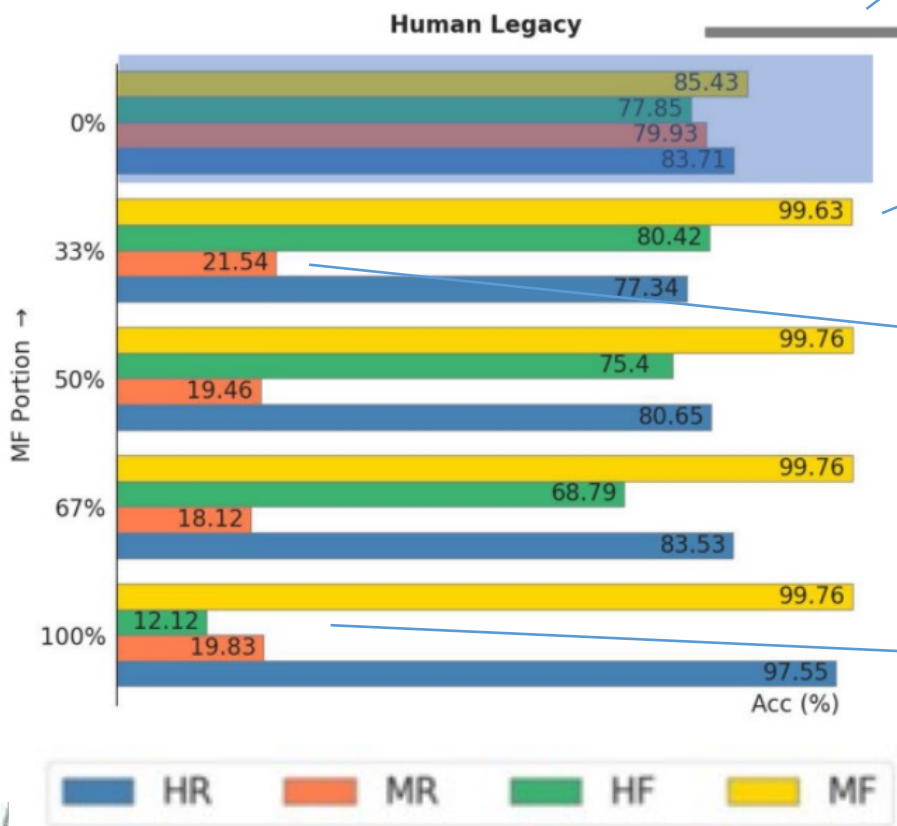
当MF（黄色）部分增加到33%时，MF（黄色）的检测准确度增加到约99%，并且在训练数据中进一步增加MF子类的比例几乎不再对MF子类的测试检测准确度有更多贡献。

所以只需引入适当的MF，对MF的检测很有利

MR（橙色）下降剧烈，可能因为模型觉得只要是机器生成的就是假新闻，而忽视了事实性本身

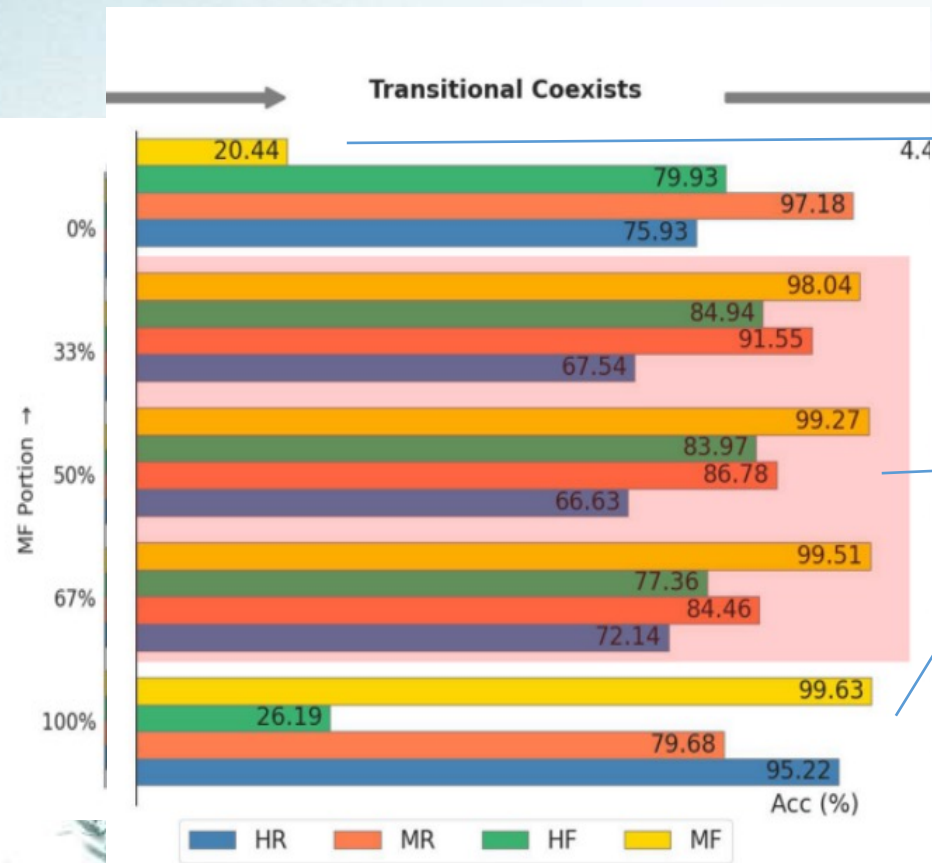
HF剧降：模型将人类生成的归为真实，说明它也是以是否机器生成作为评判标准。

另外，MR子类的准确度大于 $1 - \text{Acc}(\text{MF})$ ，说明模型还是能识别一点点真实性





这一阶段设置是真实类别中，人类和机器混合生成真新闻

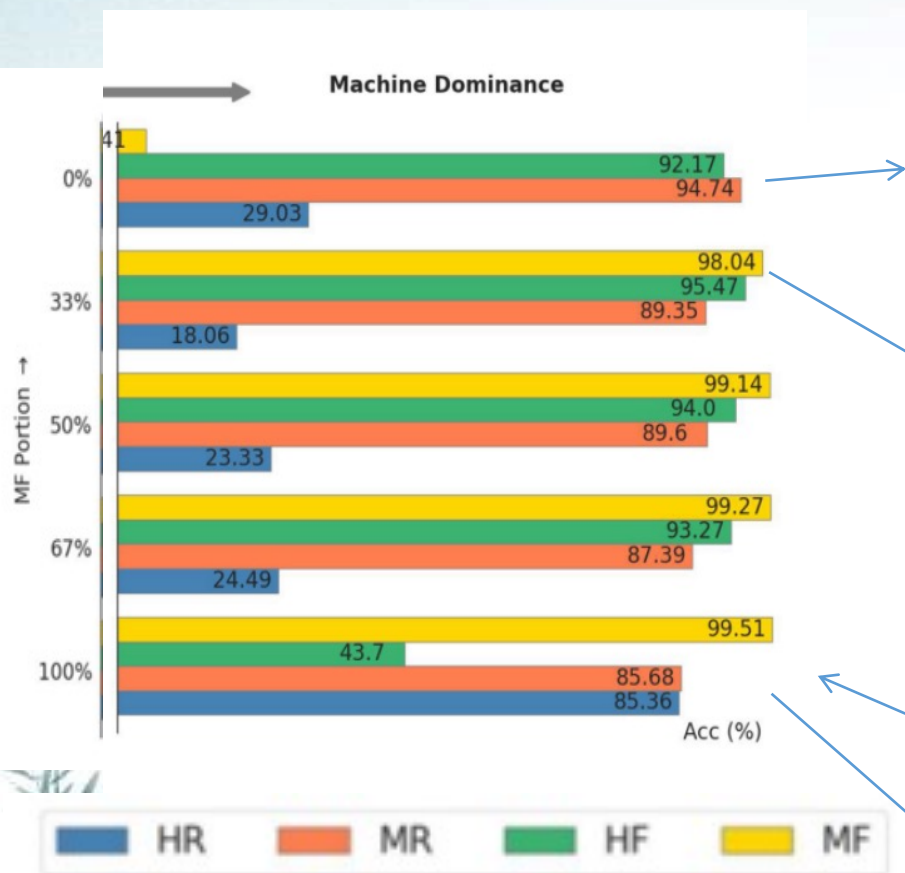


MF全是人类生成时候，MF检测ACC较低，HF较高；
MF全是机器生成时，结论相反，本质还是基于文章来源判别

一半MF一半HF情况下，MF的准确率
还是最高，HR最低。结果并没有像上
一阶段那么均衡（蓝色区域）。因此，
在训练期间努力实现每个子类的完美
平衡可能不会产生像仅在人类生成的
真假新闻上训练那样好的结果

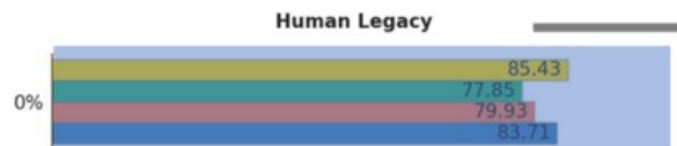


在这种设置中，真实新闻类别的全部训练数据由MR组成



当假新闻类别只有HF作为训练示例（即0% MF部分）时，检测器在辨别HF和MR方面表现出色，似乎是通过识别文章的来源（机器或人类）而不是建模其事实性。

将适度的33% MF文章引入训练数据，使得MF检测准确度的巨大飙升，从4.41%跃升至98.04%。这表明，在这个训练集中，模型有能力区分真假内容，而不会被表面特征误导，这些表面特征用于分类MF和MR类别。



机器生成文章（MF和MR）的真实性比人类生成文章（HF和HR）更容易辨别

专门在人类撰写文章上训练的检测器即使面对机器生成内容也表现出值得称赞的准确度，而相比之下，完全在机器生成文章上训练的检测器经常错误地将HF子类分类为真实。



ChatHaruhi: Reviving Anime Character in Reality via Large Language Model

Cheng Li and **Ziang Leng**

Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng
Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, Haozhen Sun

chengli.thu@gmail.com

<https://github.com/LC1332/Chat-Haruhi-Suzumiya>





本文主要研究方向是基于大模型的角色扮演，常见于游戏NPC设计、机器人个性塑造等领域。

Chatbot

阿虚:「学校里的社团都太无聊了怎么办」

春日:「没错，这些社团都太普通了，没有任何创意和特色。」

阿虚:「那如果我们组建一个社团，要叫什么名字呢？」

春日:「社团名字我早就想好了，叫做SOS团，让世界变得更热闹的凉宫春日团，简称SOS团。」

Image





- 直接利用如右图中的最常见prompt让大模型去进行模仿可能有以下局限性：

- 1) 它们严重 **依赖于语言模型现有的记忆**。如果语言模型自己对作品的记忆是模糊的，那么它就不能很好地模仿特定的角色；
- 2) “know all of the knowledge” 定义模糊，并不能很好地防止 **大模型幻觉**；
- 3) **不会重复台词**；由于RLHF，大模型有自己特定的语言偏好，**会话风格仍然受到底层语言模型的严重影响**；

*I want you to act like {character} from {series}.
I want you to respond and answer like {character}
using the tone, manner and vocabulary {character}
would use. Do not write any explanations. Only
answer like {character}. You must know all of the
knowledge of {character}. My first sentence is "Hi
{character}."*





局限性：它们严重 依赖于语言模型现有的记忆。如果语言模型自己对作品的记忆是模糊的，那么它就不能很好地模仿特定的角色；

解决方式：将角色历史场景对话微调到大模型中，并基于CAMEL和BAIZE的启发为性格鲜明的次要角色生成额外对话数据。

$$a = \operatorname{argmax}_{a'} P(a' | s_R, D(q, R), q, H, \Theta)$$

系统提示词

模型参数

过往对话

额外对话（为保证D连续性）

（采用相似度算法来捕捉最相关的）



- 为了改善不会重复台词和人物强调不明显的问题，设计了相应的模板：
- I want you to act like {character} from {series}.
- You are now cosplay {character}
- If others' questions are related with the novel, please try to reuse the original lines from the novel.
- I want you to respond and answer like {character} using the tone, manner and vocabulary {character} would use.
- You must know all of the knowledge of {character}.
- {人物性格的补充说明}





本项目的关键想法，是抽取尽可能多的原剧本，形成角色的记忆数据库。在用户给出新的提问时，系统会搜索相关的经典剧情。并且结合人物设定的prompt，去组合控制语言模型，争取对角色形成更精确的模仿。

(prompt上作者做了些工作)



Figure 4: The core dialogue system of ChatHaruhi, comprising the system prompt, character memories $D(q, R)$ retrieved for the user query q , and the dialogue history H .