CAN LANGUAGE MODELS MAKE FUN? A CASE STUDY IN CHINESE COMICAL CROSSTALK

Anonymous authors
Paper under double-blind review

ABSTRACT

Language is the principal tool for human communication, in which humor is one of the most attractive parts. Producing natural language like humans using computers, a.k.a, Natural Language Generation (NLG), has been widely used for dialogue systems, chatbots, machine translation, as well as computer-aid creation e.g., idea generations, scriptwriting. However, the humor aspect of natural language is relatively under-investigated, especially in the age of pre-trained language models. In this work, we aim to preliminarily test whether *NLG can generate humor as humans do*. We build a new dataset consisting of numerous digitized **C**hinese **C**omical **C**rosstalk scripts (called **C**^{3} in short), which is for a popular Chinese performing art called 'Xiangsheng' or '相声' since 1800s [1]. We benchmark various generation approaches including training-from-scratch Seq2seq, fine-tuned middle-scale PLMs, and large-scale PLMs (with and without fine-tuning). Moreover, we also conduct a human assessment, showing that 1) *large-scale pretraining largely improves crosstalk generation quality*; and 2) *even the scripts generated from the best PLM is far from what we expect*. We conclude humor generation could be largely improved using large-scaled PLMs, but it is still in its infancy. The data and benchmarking code are publicly available in `https://github.com/anonNo2/crosstalk-generation`.

1 INTRODUCTION

Artificial Intelligence (AI) has been widely used in Natural Language Processing (NLP), computer vision, speech, robots, and further applied biology, etc. In NLP, large-scale Pre-trained Language Models (PLMs) e.g., BERT Devlin et al. (2018) and GPT Radford et al. (2018), have notably improved many natural language tasks including text classification, question answering, and natural language generation. Although its technical contribution to the human community has been widely explored, the social or cultural effect is somehow under-investigated.

To explore the side social or cultural effect of PLMs, in this paper, we lavage the generation ability of pre-trained language models to save endangered cultural heritage, i.e., Chinese Comical Crosstalk. We believe the diversity of generations from pre-trained language models could enrich the Chinese Comical Crosstalk, this may help to prevent it from extinction. From a broader view, we aim to test the ability of 'how AI makes fun' in the context of PLMs (especially large-scale GPT).

Humor has been rooted in the Chinese language, originating from the book 'Records of the Grand Historian' written by a Chinese historian Qian Sima 2000 years ago [2] which includes a chapter titled 'Biography of Humor' 《滑稽列传》. Since then, humor is an inseparable ingredient of the Chinese language. As the first step, this work aims to explore a traditional performing art in Chinese comedy crosstalk, called 'XiangSheng' or '相声' in Chinese, which has a very long history originating from the north of China since roughly 1800. It began as a form of street performance, incorporating joke-telling, comedic banter, imitations, or borrowing from other performance arts, such as Peking opera, all with the express purpose of making audiences laugh. The characteristics of crosstalk scripts are 1) multiple-turn; 2) humor-oriented; 3) with a novel language style; 4) culturally-grounded; and 5) low-resourced, see more details in Sec. 3. See Table 1 for an example crosstalk script.

[1]For convenience for non-Chinese speakers, we called 'crosstalk' for 'Xiangsheng' in this paper.

[2]The book was written by Qian Sima in 94 BC, one can see its modern version Qian & Watson (1993). Its Chinese name is 《史记》

Roles	**Script** (in Chinese)	**Translated script** (in English)
Peng	张三和李四在这里给大家拜年了!	We are both here wishing you a happy new year
Dou	从大家的掌声呀,我听出来了。	What do you know I heard from the audience's applause?
Peng	什么呀?	What?
Dou	大家还是比较喜欢，我们俩的。	Audiences do love us both.
Peng	哎呦，你心里真没数。什么叫喜欢我们俩呀。	No, not both!
Dou	(哦。)	err?
Peng	人家鼓掌，是喜欢我们俩当中的一个。	They are applauding only one of us.
Dou	我一直以为大伙也喜欢你呢。	I thought that audiences also had loved you.
Peng	呵呵	hehe
Dou	别看我们俩一上台就在那斗嘴。	Although we are always quarreling on the stage,
Peng	哦。	but what?
Dou	实际上我们俩在生活当中呀...	Actually in daily life, we
Peng	动手。	we directly fight with each other
Dou	哎呦，急了?你就处处跟我呛着,什么事我喜欢的	Well, you are always going against me, anything I love...
Peng	我绝不喜欢。	I will definitely hate it!
Dou	什么事凡是我认为好的。	anything I think is right?
Peng	我就认为它坏。	I will definitely think it is wrong!
Dou	我就认为你好。	I think you are very nice!
Peng	我就认为你讲的有道理。	Make sense!
Dou	这你怎么不呛着了?	Why not argue with me?
Peng	过年了，我怎么也得顺着你点。	Sometimes I have to agree with you a little bit.

Table 1: An example of crosstalk script. Typical crosstalk scripts could be longer.

Humor generation is a challenging task since, for instance, we may not know exactly what makes a joke funny. Solving this problem algorithmically requires deep semantic understanding Petrović & Matthews (2013). This becomes more challenging if cultural and other contextual cues are considered as in Chinese Comical Crosstalk. From a practical point of view, the data preparation usually goes earlier than the development of algorithms and models. Since new models cannot be well-evaluated before (especially large-scale) data is ready [3].

As the first step, we collect many crosstalk scripts from the internet. The dataset is publicly available with an open-resource license (Apache 2.0). We also conduct several basic generation approaches including train-from-scratch Seq2seq generation Cho et al. (2014), fine-tuned middle-scale PLMs, and large-scale PLMs (with and without fine-tuning). Furthermore, the current research community also explored the potential to use large-scale PLMs for creation. For example, Brown et al. (2020) claims that GPT-3 can generate synthetic news articles that human evaluators have difficulty distinguishing from human-generated articles. We do not expect that GPT has a 'sense of humor'. Alternatively, We test to which degree GPT-3 is creative in crosstalk generation thanks to the OpenAI API [4].

The **contributions** of this paper are as follows: 1) Firstly, **culturally**, we digitize and clean crosstalk scripts at scale, contributing to both the NLP research community and the traditional Chinese culture community. This will inspire more crosstalk script creations and therefore preserves this intangible cultural heritage. Currently, most crosstalk scripts seem to be homogeneous which is one of the main bottlenecks that limit its wide spreading. This work will promote its diversity and creation which can be beneficial in preventing it from extermination. 2) Secondly, **technically**, we benchmark various approaches including Seq2seq, train-from-scratch GPT, pre-trained GPT 2, and GPT-3, for crosstalk generation. As far as we know, this is the first work to *evaluate to which extent pre-trained language models could generate humorous text*, as a benchmark for computer-aided creation for fun. 3) Lastly, we further point out the issues regarding various biases, stereotypes, and sometimes insulting.

2 Related Work

Natural language generation Natural language generation is one of the key areas of NLP that is related to machine translation, dialogue, summarization, and paraphrasing. Previously, text generation was usually based on templates or rules, probabilistic models like n-gram models. Those models are fairly interpretable, but heavily require feature engineering. Recently, neural network language models Bengio et al. (2003) show a great potential to generate language by chronologically predicting the next word with context using neural networks. Cho et al. (2014) proposed the encoder-decoder architecture that becomes the de facto paradigm of natural language generations. For a given input sequence, the encoder produces its corresponding fixed-length hidden vector that is used for the decoder model to generate another sequence. Recently, pre-trained language models (including GPT Radford et al. (2018) and UniLM Dong et al. (2019)) have largely improved the SOTA of language

[3]One can see a concrete example in computer vision that ImageNet dataset Deng et al. (2009) largely promotes the development of image classification models He et al. (2016), and concrete examples in NLP are GLUE Wang et al. (2018) and SQuAD Rajpurkar et al. (2016; 2018) benchmarks that benefit natural language understanding Devlin et al. (2018).

[4]https://openai.com/api/

models, by using a better backbone architecture called 'transformer' in a pre-trained manner. Very recently, Brown et al. (2020) released API to access their large-scale language models called 'GPT-3'. Moreover, some NLG tasks are specific to Chinese, e.g., Chinese poetry and couplet generation He et al. (2012); Yan et al. (2013); Zhang & Lapata (2014); Yi et al. (2017); Liao et al. (2019).

Humor in NLP There are two typical lines of research work for humor in NLP: humor recognition and humor generation. The former was well-investigated using neural networks Bertero & Fung (2016); Yang et al. (2015); Chen & Lee (2017); Liu et al. (2018b); Chen & Soo (2018); Liu et al. (2018a), while the latter is more challenging yet under-investigated. Both humor theoretical linguistics and computational linguistics have heavily contributed to humor generation (see Amin & Burghardt (2020) and Lin et al. (2016)). There are many efforts for humor theory linguistics to develop the theoretical aspect of humor Raskin (1979). Computational linguistics tends to leverage neural systems, template-based systems, or a hybrid of both for humor generation that rarely benefits from those theory-driven impulses. For example, Labutov & Lipson (2012) explored mining simple humorous scripts from a semantic network (ConceptNet). They claimed that this may generate humor beyond simple puns and punning riddles Binsted & Ritchie (1997). Petrović & Matthews (2013) claimed that generating humor algorithmically requires deep semantic understanding. Ren & Yang (2017) used an encoder for representing a user-provided topic and an RNN decoder for joke generation that can generate a short joke relevant to the specified topic. Yu et al. (2018) proposed to generate puns from a conditional neural language model with an elaborately designed decoding algorithm. He et al. (2019) propose a retrieve-and-edit approach that could generate more puns. Although the humor generation has been paid some attention, we believe that the humor generation is in its infant age, and the potential of pre-trained language models like GPT is expected to be exploited.

Before the pre-trained language model era, Du et al. (2017) simplified the script generation task by generating the replying utterance of the supporting role (Peng) given the utterance of the leading comedian in each dialogue. This setting is not expected in many aspects. First, this may not generate fluent script since only a single utterance is considered as the context. Second, generating replying utterance of the supporting role is not challenging since the complexity of the supporting role is much less challenging than the utterance of the leading comedian. We argue that a more natural generation (like auto-regressive generation) is needed and pre-trained language models may help.

3 Problem Definition

3.1 Task Formalization

Depending on the number of performers, crosstalk is typically performed as a dialogue between two performers called '对口', or rarely as a monologue by a solo performer called '单口' (like stand-up comedy in the Western), or even less frequently, as a group acting by more performers called '群口'.

Let us take the dual performing ('对口') as an example. Dual performing usually involves two roles called Penggen '捧哏' (Peng in short) and Dougen ('逗哏') (Dou in short). Dou aims to perform in a comical way using talking and actions. Peng is the support role to make the conversation more fluent and legible (As shown in Table 1). The conversation consists of an iterative sequence of utterances:

$$\Phi = \{u_1, v_1, u_2, v_2, \cdots, u_K, v_K\}$$

which is a K-turn dual crosstalk conversation with 2K utterances including K utterances from Dou (denoted as u) and K utterances from Peng (denoted as v). Note that both u_i and v_i are utterances that consists of many characters, namely $u_i = \{\phi_{i,1}, \phi_{i,2}, \cdots, \phi_{i,j}, \cdots \phi_{i,l_i}\}$, $\phi_{i,j}$ is the j-character in the i-th Dou/Peng utterance and l_i is the number of characters in the utterance.

Training could be formulated as two paradigms: 1) a **Seq2seq utterance generation task**: it could be treated as a seq-to-seq task to predict the next utterance given previous utterances; 2) **a next word generation task**: it can also consider as a typical language model that does not consider the utterance border, namely a raw language model that predicts the next word [5]. For automatic evaluation in Sec. 5, we adopt commonly-used generation metrics to evaluate models using an auto-regressive utterance generation manner, namely, predicting the next utterance based on previous utterances no matter it is trained in a Seq2seq utterance generation paradigm or next word prediction paradigm.

[5] In this study, we treat a character as a word without distinction.

3.2 CHARACTERISTICS OF CROSSTALK

Crosstalk scripts (except for solo performers) are usually multiple-turn dialogues. It typically involves two (or more) performers talking about a topic in multiple turns (with an average of 72 in C^3 dataset), typically ranging from 10 to 20 minutes. In contrast to general dialogues, the characteristics of the crosstalk are as follows: 1) **it is humor-oriented**： it aims to make audiences laugh by freely talking. 2) **it is with a novel language style**: the crosstalk language itself is in a rapid, bantering, and highly interactive style. More interestingly, it is rich in puns and allusions. 3) **it is culturally-grounded**: it typically relates to not only the local daily life (especially in the north of China, e.g., Beijing) but also the long historical events in china with a time range from 3000 BC to the present. Interestingly, it usually adopts the Beijing dialect (close to Mandarin) during some periods. 4) **it is low-resourced**: crosstalk generation task could rely on relatively low-resourced digitized scripts.

4 DATASET

4.1 DATA COLLECTION

We collect data from the book ' Encyclopedia of Chinese Traditional Crosstalk' and the internet. The creation date of these scripts ranges from Qing Dynasty (roughly 1800) to this century. The main resources are from 1) a digitized book named ' Encyclopedia of Chinese Traditional Crosstalk' or 《中国传统相声大全》 published in 2003, which is a collection of traditional crosstalk collections, records, and compilations since Qing Dynasty; 2) many websites that maintain crosstalk scripts. See App A for more details. Our dataset uses the Apache-2.0 license.

Preprocessing and cleaning Two scripts sharing 80% characters will be merged as identical ones. Scripts that are shorter than 7 lines are filtered. We use regular expressions to clean the text, e.g., removing HTML tags and noisy tokens. We also filter out the voice, action, and environment descriptions. Punctuation marks are also normalized. Actor names are re-numbered with new placeholders while the metadata of these actors is maintained as shown in Listing 9.

Human calibration The collected data might be dirty. Therefore, we calibrated data manually: 1) we removed scripts that contain insulting and discriminatory conversations. 2) We also manually reviewed some advertising words in the script and deleted those texts. 3) We manually split some scripts by utterances if a script has extremely long utterances. 4) We removed scripts that make no sense, e.g., scripts that are not fluent or contain too many meaningless tokens.

4.2 OVERVIEW OF C^3 DATASET

Scale of the dataset As shown in Table 2, we collect 9,331 high-quality scripts with 663,305 utterances. This results in 9,331 dialogues and 16,481,376 characters in total. We randomly select 200 scripts for testing and the rest for training.

-	Number
Total scripts	9,331
Total characters	16,481,376
Number of utterances	663,305
Number of *long* utterances	8,717
Number of *short* utterances	446,756
Median character numbers of utterances	16
Mean utterances per script	71

Table 2: Statistics of the C^3 dataset.

Length of scripts and utterences Each script contains an average of 71 utterances. The medium length of utterances is about 16 characters. We define an utterance as a *long* utterance if it exceeds 128 characters and *short* utterance if it is less than 24 characters. There are 8,717 *long* utterances and 446,756 *short* utterances.

Numbers of performers As shown in Table 3, it includes 3,685 dual-performing crosstalk scripts, 256 group-performing crosstalk scripts, and 168 single-performing crosstalk scripts. In addition, we also collect 5,222 sketch comedy ('小品') scripts that also involve multi-turn dialogues. Note that ketch comedy scripts are also mainly about dialogues and one may be interested in them. While we do not use ketch comedy scripts to train the crosstalk script generation. The main type of a script is the dual dialogue with two performers (called '捧哏' and '逗哏'), with 3,685 scripts. A few of them are monologues and multiple-performer dialogues, with 168 and 256 scripts respectively.

Type	Number
Single performing	168
Dual performing	3,685
Group performing	256
Ketch comedy	5,222
Total	9,331

Table 3: Statistics of various types.

4.3 DISCUSSIONS ON C^3

Humor categories in crosstalk Typical humor theory defines three types of humor: 1) relief theory: reducing psychological tension, 2) superiority theory: laughing about misfortunes of others that make one feel superior, and 3) incongruous juxtaposition theory: incongruity between a concept involved in a certain situation and the real objects of the concept. These three mechanisms could be easily found in crosstalk scripts. For example, 1) performers bring audiences to a tense scenario and suddenly make a relaxing joke, 2) performers make jokes about someone (usually one of the performers on the stage or other crosstalk performers that is not on the stage) with bad experiences, and 3) performers sometimes describe some ridiculous scenarios that make fun.

Another specific humor in crosstalk is 'homographic pun' Yu et al. (2020), since crosstalk is a verbal performing art. This sometimes relates to some dialects in Chinese. To deal with 'homographic pun', generation models may need to be injected with some acoustic knowledge.

Ethical issues in crosstalk We have to notice that there are many ethical issues involved in crosstalk. Many biases are involved in crosstalk including educational background discrimination, gender bias, and occupation bias. Also, a stereotype of local people is amplified by crosstalk scripts. Typically, the two Performers also make fun of each other, some of them are like an 'insult'. Fortunately, this is only for crosstalk performers themselves. We believe that dealing with these ethical issues should be necessary to promote crosstalk art.

5 GENERATION BENCHMARK USING AUTOMATIC EVALUATIONS

5.1 EXPERIMENTAL SETTINGS

Method	**Baselines**
train from **scratch**	LSTM Seq2seq
fine-tuned PLMs	UniLM, GPT, T5
zero-shot large-scale PLMs	CPM, Zhouwenwang, Pangu-α, GPT-3
fine-tuned large-scale PLMs	fine-tuned GPT-3

Table 4: Taxonomy of baselines.

We implement LSTM Seq2seq which is **trained from scratch** as a baseline. To make use of existing pre-trained language models, we also include pre-trained UniLM, GPT, and T5 in a **fine-tuned** manner. Large-scale Chinese pre-trained language models like CPM, Zhouwenwang, Pangu-α were recently released, we, therefore, evaluate these models in a **zero-shot** fashion since fine-tuning these models are economically-expensive. Furthermore, we also verified the effectiveness of GPT-3. Fortunately, GPT-3 provides an API for fine-tuning, making GPT-3 the only large-scale PLM that could be fine-tuned at an affordable cost. See App. C for more details.

LSTM Seq2seq Sutskever et al. (2014): LSTM consists of a two-layer bi-directional LSTM encoder and a two-layer LSTM decoder [6]. Both the embedding size and the hidden state size of the LSTM model are set to 300. The encoder-decoder model is augmented with an attention mechanism. For the k-th utterance in a dialog, the input of the encoder was the concatenation of all the past utterances before k truncated with 256 tokens, while the target output of the decoder was the k-th utterance.

UniLM Dong et al. (2019): Unified Language Model (UniLM) adopts multi-layer Transformers, which also uses different masks to control the number of visible context words thereby can be applied to both natural language understanding (NLU) tasks and natural language generation (NLG) tasks. Our pre-trained model is downloaded from [7], pre-training with Wikipedia data and news corpus data in CLUE. The UniLM used in this paper consists of 12 layers with a hidden size of 768 and 12 heads. The ways to build fine-tuned data structures are the same as Seq2seq.

T5 Raffel et al. (2019) is a unified framework that treats various text tasks into a text-to-text format. It consists of an encoder component and a decoder component, both of which are a stack of Transformer layers. We use the Chinese version of the T5 called 'T5-Chinese-base' [8]. The parameters of the base model are 275 million, and the parameters of the small model are 95 million.

GPT Radford et al. (2018): Generative Pre-trained Transformer (GPT) models by OpenAI have taken the natural language processing community by introducing very powerful language models. The

[6] The codebase is from `https://github.com/IBM/pytorch-Seq2seq`
[7] `https://github.com/YunwenTechnology/UniLM`
[8] `https://huggingface.co/imxly/t5-pegasus`

	BLEU	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
LSTM Seq2seq	11.77	4.02	1.47	0.57	2.49	17.25	2.13	15.94	4.73	16.23
GPT	10.04	3.69	1.53	0.7	2.75	15.28	1.78	13.7	6.89	37.39
UniLM	8.88	4.32	2.47	1.41	3.36	20.22	4.91	18.98	7.53	29.90
T5-small	11.71	5.39	2.93	1.67	3.64	19.98	4.37	18.61	8.08	36.38
T5-base	11.75	5.58	3.13	1.77	3.94	20.8	4.98	19.25	9.02	42.68
CPM-Large	7.94	2.87	1.19	0.50	1.68	9.88	1.28	8.83	5.82	34.43
Pangu-α	6.42	2.09	0.83	0.37	1.31	7.00	0.75	6.14	8.25	50.98
Zhouwenwang	7.33	2.26	0.90	0.40	1.81	10.41	1.01	8.61	9.72	**53.53**
GPT3 (GPT3-Davinci)	**14.68**	**7.45**	**4.44**	**2.77**	**5.13**	**22.25**	**5.65**	20.03	8.43	40.70
GPT3-fine-tuned-Davinci	9.66	4.89	3.01	1.92	4.66	21.79	5.50	**20.22**	**9.73**	43.15

Table 5: Evaluation results on crosstalk generation.

GPT model is based on a unidirectional transformer with some modifications. In our implementation, the GPT model is 12-layer Transformers with hidden size 768, pre-trained using LCCC Corpus Base corpus [9] and fine-tuned by crosstalk dataset. Follow the implement of code [10], We divide the dialog into utterances and sequentially combine utterances with fewer than 256 words as one input.

GPT-3 Brown et al. (2020): the biggest GPT-3 model has 175 billion parameters trained by 45TB data. Note that GPT-3 is mainly for English language generation, but it could also generate fluent Chinese texts. We applied the GPT-3 online test API [11] and evaluate crosstalk generation. **GPT3-Davinci** is the one with Davinci engine without fine-tuning. [12]. **GPT3-Davinci-finetuned** is the fine-tuned version using GPT-3 API. We fine-tune it on 200 crosstalk scripts in 4 epochs.

Pangu-α Zeng et al. (2021) is large-scale autoregressive language models, with up to 200 billion parameters. It consumes 1.1TB of high-quality Chinese corpora from a wide range of domains. A publicly-available version of Pangu-α (with 2.6B parameters) could be used in `https://huggingface.co/imone/pangu_2_6B`.

CPM Zhang et al. (2021) is a generative pre-training model trained on 100 GB Chinese corpora. **CPM-Large** is with 36 transformer layer and reaches 2.6B parameters.

Zhouwenwang considers both the generative language model task and mask language model; it could have the ability for both language generation and natural language understanding. The larger model (Zhouwenwang-1.3B) is with 1.3 billion parameters [13].

Evaluations We use the test set (200 randomly-selected crosstalk scripts) for evaluations. To generate the k-th utterance, we concatenate all the past utterances before k within a total length of 256 as the input. We adopted several widely-used metrics to measure the quality of the generated response. **BLEU-1/2/4** is a popular metric to compute the k-gram overlap between a generated utterance and a reference. **ROUGE-1/2/L** measures unigram and bigram overlap in a recall-oriented fashion while **ROUGE-L** measures the longest matching sequence of words using the longest common subsequence Lin (2004). **GLEU** Mutton et al. (2007) is an automatic evaluation of sentence-level fluency. **Distinct-1/2** Li et al. (2016) is provided to evaluate the diversity of generated responses.

5.2 Results

GPT-3 performs well The results are shown in Table 5. GPT-3 outperforms other models in most metrics (except for ROUGE-L and Distinct-1/2); this is nontrivial since GPT-3 has not been fine-tuned on this dataset, in other words, the dataset (including training and test set) is in general invisible for GPT-3. This is probably because it is trained with massive plain corpora and it, therefore, generates fluent text based on similar text in corpora.

Chinese PLMs perform relatively worse. Surprisingly, large-scale language models purely trained in Chinese (i.e., CPM, Pangu-α, and Zhouwenwang) do not perform as well as GPT-3 which is mainly trained in English corpora and partially in Chinese corpora. Especially, these

[9]`https://huggingface.co/thu-coai/CDial-GPT_LCCC-base`

[10]`https://github.com/yangjianxin1/GPT2-chitchat`

[11]`https://beta.openai.com/`

[12]The scale of Davinci engine is not exposed; however, some evidence suggests that Davinci engine might be the biggest model with 175B parameters. See `https://blog.eleuther.ai/gpt3-model-sizes/`

[13]`https://github.com/IDEA-CCNL/Fengshenbang-LM`

	Raw	Seq2seq	UniLM	GPT	GPT-3	GPT-3-fine-tuned
Dou	弟弟你想起来了? Bro, do you remember that?					
Peng	想起什么来了?这不海尔兄弟嘛。Remember what? Aren't they Haier Brothers?					
Dou	对，我们就是海尔兄弟呀。Yes, we are the Haier Brothers.					
Peng	我们是海尔兄弟? Are we?					
Dou	嗯。Hum.					
Peng	我们穿的有点多了吧? But we wear too much!					
Dou	长大了要懂得羞耻。We are not young and we need to wear proper clothes (cannot be naked as Haier Brothers)					
Peng	废话，你得知道羞耻了，还海尔兄弟了。你见过穿裤衩，站这说相声的吗? Duh, sure, have you ever seen someone who performs crosstalk with Bikini?					
Dou	小时候就可以呀，小时候多好啊。小小少年，没有烦恼，穿着裤衩，到处跑。We could when we were young. What a happy childhood! A little boy, no sorrow, run around in shorts.					
Peng	俩臭流氓，你听听。Listen! Two shameless rogues.					
Dou	当然现在不让穿了。Of course, we could do this anymore (running around in shorts).					
Peng	废话。Duh.	我想当一个演员，有点击率。I want to become a famous actor	是吗? err?	您说，咱们有这个吗? do you think that we have this?	这是谁说的? Who said this?	不让穿了。Not allowed to wear it.
Dou	我俩都三十多了，往这一站说相声。Both of us are 30 years old and perform crosstalk here	什么叫吐槽?What is called "Tucao"?	哎呀，小时候穿裤衩，到处跑。Oops, in childhood to run around in shorts.	您的意思是? What do you mean?	我说的啊。I said it.	你看看，我们现在穿什么? Look, what are we wearing now?
Peng	表演。Performing crosstalk.	报告长。[nonsense generation]	俩臭流氓。Two shameless rogues	我…… I……	你个臭流氓，你不要瞎说。You a shameless rogue, don't talk nonsense.	我们现在穿西装。We are wearing suits now.
Dou	当然这算好的。Come on, this is fine (not like a rogue)	你知道吗? Do you know?	哎哟呵，真不错。Ouch, that's great.	你可拉倒吧。Come on.	我说的是真的。It is true.	对，我们是职业演员了。Yes, we are professional actors now.
Peng	这就算好了? fine?	知道。Yes.	怎么样? How about it?	没有。No.	那你是怎么知道的? How do you know it?	你说的是。You are right.
Dou	当年，跟我们在电视台，一块录动画片那些小伙伴，现在好多都不干这行了. The friends who were recording cartoons with us on the TV station, most of them don't do this now.	你别说了，这句话太难过了。Shut up, this sentence makes me sad.	跟大伙儿分享一点笑话。Share some jokes, please.	你可拉倒吧，我都不知道了。Come on. I don't know even.	因为我就是海尔兄弟呀。Because I am a Haier Brother.	你看看，这个外套是不是很高大上?Look, isn't this coat very luxurious?
Peng	唉，你先等会，你先等会，咱还有小伙伴。Hey, wait for a moment, we still have other friends	谁说的? Who said this?	好啊。Ok.	嗨，就他长得这么丑。小时候呢?Hey, he just looks ugly. How about him in childhood?	那你是怎么知道的?How did you know?	这个外套很高大上吧? Yes, this coat is very luxurious!
Dou	小伙伴你也不记得了?Don't you remember your friends?	你觉得裸睡怎么样?What do you think of naked sleeping?	其实我觉得，穿裤衩，很难。In fact, I think it's hard to wear only short pants.	就我现在实力早晚会好吧。上一出去喝点儿酒抽烟，上网、看节目去睡觉，娱乐购物点心思开放，要坐车时间打几圈游戏玩些地方戏? My strength will be better sooner or later. Last time I went out to drink and smoke, surf the Internet, watch a show, go to bed, entertain, shopping, and be refreshed. How many laps do I need to take the car to play games and play some local dramas?	因为我就是海尔兄弟呀。Because I am a Haier brother.	哦对了，这个衣服都不能穿了，要穿西装。Oh, by the way, you can't wear these clothes anymore, you have to wear a suit
Peng	我这脑子有点懵。It is a little confusing	赤裸裸裸裸,体育局。[nonsense generation]	为什么? Why?	现在游戏也很丰富。比如，看电影就玩个正面。Now There are a lot of games. For example, when watching a movie, you can play a positive role.	那你还穿裤衩吗? Then, are you still wearing shorts?	不能穿裤衩了。Can't wearing in shorts anymore.

Table 6: The raw and generated scripts. We manually annotate meaningless texts in gray color, repeated words from the top 10 input utterance in cyan color, and insulting text in red color.

zero-shot Chinese large PLMs (i.e., CPM, Pangu-α, and Zhouwenwang) underperform fine-tuned relatively-smaller-scaled PLMs (UniLM, GPT, and T5). This might be because the multilingual corpora might be a beneficial factor since humor might be shared across languages. Also, the used GPT3-Davinci might be much bigger than the existing publicly-available Chinese PLMs.

Scale helps Comparing the performance between T5-small and T5-base, the bigger scale consistently leads to better performance. Plus, observing that the large-scale GPT3 achieves nearly the best performance in automatic evaluations, we believe that *large-scale pre-training notably improves the crosstalk generation quality*.

Fine-tuning on large-scale PLMs Interestingly, from automatic evaluations in Table 5, fine-tuning on GPT-3 achieves worse performance than vanilla GPT-3, in most metrics. We suspect the fine-tuning mechanisms might lead to such a result, like over-fitting to the training dataset, and harms some generalization. However, in human evaluation, fine-tuned GPT-3 could generate better-quality scripts than vanilla GPT-3 (in Tab. 7), which could be later observed from Tab. 6; this shows that the automatic evaluation on crosstalk might not be consistent to human perception.

Regarding diversity metrics In diversity measures using Dist-1 and Dist-2, large-scale pretraining-based models generate more diverse scripts. Since large-scale pretraining is a general method to improve the generalization ability. Note that diversity metrics are sensitive to the hyper-parameters during the decoding phase of language models.

Note that in Table 5, we do not intend to compare the general performance of these language models, or conclude that the general performance of one language model is better than another one. Since the general performance of these language models is also subject to their model scales, hyper-parameter selections, training corpora, etc. Instead, we just make use of the existing language models that are both capable to deal with the Chinese language generation and are publicly available.

5.3 Case Study

We show an example of generated crosstalk scripts in Table 6. Below are our observations.

Meaningless generation in LSTM Seq2seq LSTM language model produces fluent but nearly meaningless texts (annotated in gray color), this is probably due to the fact that the training data for Seq2seq models is not enough and no pre-training was adopted. While other models with pre-training

	General quality (5) ↑	Humor (5) ↑	Coherence (1) ↑	Ethically-risky flag(1) ↓
LSTM Seq2seq	1.45	1.61	0.27	0.03
GPT	1.50	1.71	0.39	0.01
T5-base	1.80	1.97	0.51	0.05
UniLM	1.84	2.01	0.56	0.01
Panggu-a	1.53	1.71	0.42	0.03
Zhouwenwang	1.23	1.27	0.19	0.05
CPM-Large	1.42	1.60	0.40	**0.23**
GPT3-Davinci	2.15	2.17	0.65	0.03
GPT3-Davinci-finetuned	**2.27**	**2.35**	**0.71**	0.01
raw scripts	3.52	3.46	0.95	0.01

Table 7: Human assessment for crosstalk generation. The maximum score of each metric in the bracket, namely, the best *general* quality score and *humor* score is 5 while the rest scores are binary.

do not frequently generate such nonsense texts. This shows that pre-training could boost generation performance, especially for the scenarios with low-resourced training texts.

Repeated context topic in generation UniLM and GPT-3 could generate topic-coherent texts, especially, some generated utterances also repeat some key topics from the first 10 input utterances, e.g., '臭流氓' (`shameless rogues`), '穿裤衩，到处跑'(`running around in shorts`), and '海尔兄弟'(`Haier brother`[14]). Note in this example, the raw script (the last 10 utterances) dot not have so many repeated topics from previous utterances, like generation models.

Insulting words UniLM, GPT, and GPT-3 generate some insulting words that already appeared in the first 10 utterances, namely, '臭流氓' (`shameless rogues`). Moreover, GPT also generates new insulting words, 就他长得这么丑 `he just looks ugly` that did not appear before. This is probably due to that the other training scripts or pretraining corpora may have such insulting texts.

Humorous generation Roughly, we could see some humorous generated utterances. For example, the last generated utterance for GPT-3 (in the last row and second last column) does have a sense of humor. However, if we treat these utterances as a whole, their performance of humor is not satisfied.

The difference between Peng and Dou Basically, Dou usually talks more and introduces more topics in dialogues while Peng usually supports Dou to make each of his topics more comprehensively talked and topic transfer more coherent. This leads to that Peng's utterances sometimes contain only a few interjections like '嗯'(`hum`) and '哎哟'(`ouch`). We argue that the generation for Dou's utterance is much more difficult than Peng, and the former is more interesting and worthy of more attention.

6 Human Assessment for Crosstalk Generation

Setting We randomly select 50 scripts in the test set. We take the first ten utterances as the input for Seq2seq, GPT, GPT-3, and UniLM. These models will generate the next ten utterances, utterance by utterance or character by character. We evaluate the generated scripts in 10 utterances conditioned on the first 10 utterances of raw scripts, see the web UI in App. E. For each script, we show participants with 20 utterances (including the raw 10 utterances and the generated 10 utterances). Participants are required to 1) rate five-point scores for the **general quality** and **humor degree** of each generated script ('5' for the best and '1' for the worst); and 2) rate binary scores for **coherence** and an **ethically-risky flag** of each generated example ('1' for true and '0' for false). We ask no-paid volunteers to participate to rate these generated results from 10 models. 15 participators have completed all ratings. The score is calculated as an average score among all dialogues and all participants for each model. The Fleiss' kappa among these participants is 0.366.

Human assessment is shown in Table 7. Raw scripts achieve the best general quality, probably evidencing that the ability to be creative and humorous of humans is much better than that of SOTA models. Among these models, GPT-3 and its fine-tuned version (GPT3-Davinci-finetuned) outperform others in terms of general quality. Interestingly, fine-tuned GPT-3 outperforms zero-shot GPT-3 although the former has poorer performance in automatic evaluation (see Tab. 5).

Similar to the automatic evaluations in Tab. 5, zero-shot large-scale Chinese PLMs (the third group) underperforms these fine-tuned middle-scaled PLMs (like UniLM, T5 and GPT). Seq2seq performs

[14]Haier Brothers, see `https://www.imdb.com/title/tt8302572/`, a cartoon about a pair of robots called 'Haier Brothers' who travel around the world to explore the nature.

	General quality (5)	Humor (5)	Coherence (1)	Ethically-risky flag(1)
GPT3-Davinci-10	2.60	1.89	0.93	0.00
GPT3-Davinci-200	**3.22**	**2.56**	**0.96**	0.04
GPT3-Davinci-1000	3.02	2.42	0.89	0.04

Table 8: Human assessment on GPT3 with different numbers of fine-tuned examples.

the worst; this may be due to Seq2seq does not utilize the pre-training. Interestingly, CPM-large produces much more insulting content than others; the reason needs to be further investigated.

7 Discussions

7.1 Why are generated crosstalk scripts not satisfied enough?

As seen from the automatic evaluation in Tab. 5 and human assessment in Tab. 7 , the adoption of large-scale pre-trained language models could largely improve the quality of crosstalk generation, compared to these models without large-scaled pre-training. We show some generated examples from large-scaled pre-trained language models with and without fine-tuning.

Although large-scale pre-trained language models largely improve crosstalk generation. Based on the human assessment, we could preliminarily conclude that *the best generation approach (fine-tuned GPT-3) achieves fairly good crosstalk (2.27 vs. 3.52 for general quality), while it is far away from what we expect*. The reason could be twofold as below.

First, the evaluation criterion for humor generation is problematic. Observing the inconsistency between Tab. 5 and Tab. 7, a better performance evaluated using BLEU and ROUGE does not lead to a better performance in human assessment, this probably suggests that BLEU or related metrics for generation is not inappropriate for humor generation Since humor itself is diverse and subjective that does not have textual ground truth. One could see the correlations between human and automatic evaluation in App D which is relatively high but somehow overestimated. Moreover, human assessment is expensive and cannot give real-time feedback during model training.

Secondly, current methods did not consider prime ingredients of humor. Core ingredients of humor include incongruity, surprise, cultural empathy, and interpersonal effect, without which simply training on data is a soft way to memorize the training data and it can't generate real humor.

7.2 Sensitivity on the fine-tuning examples of GPT3

We test the performance of GPT3 models with different numbers of fine-tuned examples (i.e., 10, 200, 1000), using a similar human assessment in Sec. 6. For 15 randomly-selected crosstalk scripts, based on the beginning snippets (*i.e.*, the first ten utterances of each crosstalk script), each model generates/completes the rest of the crosstalk script. Three participants are required to annotate these 15 generated crosstalk scripts in terms of four scores (general quality, humor, coherence, and ethically-risky flag), the former two are five-degree while the latter two are binary.

Tab 8 shows that with a moderate number of fine-tuned examples it achieves the best general quality. In other words, adopting too many or few fine-tuned examples could harm the performance. This is slightly counterintuitive. Interestingly, fine-tuning on 200/1000 examples brings more ethical risks; this probably indicates that the dataset itself has some ethical risks, which should be noticed.

8 Conclusion and Future Work

In this paper, we collect a dataset for Chinese crosstalk. Based on the dataset, we evaluate several existing generation models including LSTM Seq2seq, GPT, UniLM, CPM, Pangu-α, Zhouwenwang, and GPT-3 for crosstalk generation. This is a preliminary step for humor generation, indicating that large-scale pretraining largely improves crosstalk generation quality while there still exists a big gap between the generated scripts and human-created scripts. Note that there are some concerns about bias/stereotypes for crosstalk, e.g., educational background discrimination and gender bias. In future work, we are interested in collecting crosstalk audios to promote the end2end crosstalk generation with an adapted humorous accent.

REFERENCES

Miriam Amin and Manuel Burghardt. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 29–41, 2020.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003.

Dario Bertero and Pascale Fung. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 130–135, 2016.

Kim Binsted and Graeme Ritchie. Computational rules for generating punning riddles. 1997.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Lei Chen and Chong MIn Lee. Predicting audience's laughter using convolutional neural network. *arXiv preprint arXiv:1702.02584*, 2017.

Peng-Yu Chen and Von-Wun Soo. Humor recognition using deep learning. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)*, pp. 113–117, 2018.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation, 2019.

Shikang Du, Xiaojun Wan, and Yajie Ye. Towards automatic generation of entertaining dialogues in chinese crosstalks. *arXiv preprint arXiv:1711.00294*, 2017.

He He, Nanyun Peng, and Percy Liang. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1734–1744, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1172. URL `https://www.aclweb.org/anthology/N19-1172`.

Jing He, Ming Zhou, and Long Jiang. Generating chinese classical poems with statistical machine translation models. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Igor Labutov and Hod Lipson. Humor as circuits in semantic networks. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 150–155, 2012.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016.

Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. Gpt-based generation for classical chinese poetry, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

HF Lin, DY Zhang, Liang Yang, and Bo XU. Computational humor researches and applications [j]. *Journal of Shandong University*, 2016.

Lizhen Liu, Donghai Zhang, and Wei Song. Exploiting syntactic structures for humor recognition. In *Proceedings of the 27th international conference on computational linguistics*, pp. 1875–1883, 2018a.

Lizhen Liu, Donghai Zhang, and Wei Song. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 586–591, 2018b.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 344–351, 2007.

Saša Petrović and David Matthews. Unsupervised joke generation from big data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)*, pp. 228–232, 2013.

Sima Qian and Burton Watson. *Records of the Grand Historian*, volume 1. Columbia University Press New York, 1993.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Victor Raskin. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, volume 5, pp. 325–335, 1979.

He Ren and Quan Yang. Neural joke generation. *Final Project Reports of Course CS224n*, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2018.

Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. I, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2367–2376, 2015.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. Generating chinese classical poems with rnn encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 211–223. Springer, 2017.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1650–1660, 2018.

Zhiwei Yu, Hongyu Zang, and Xiaojun Wan. Homophonic pun generation with lexically constrained rewriting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2870–2876, 2020.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu-$\backslash alpha$: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–680, 2014.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, 2:93–99, 2021.

A DATA RESOURCES

We crawled scripts mainly from the following resources:

- a digitized book named `Encyclopedia of Chinese Traditional Crosstalk` 《中国传统相声大全》 published in 2003. The book is a collection of traditional crosstalk collections, records, and compilations from the Qing Dynasty.
- `bijianshang.com` (中文台词网): a website for the scripts of Xiangsheng, short sketches, and movies.
- `www.juben68.com` (剧本网): a website with lots of movie scripts, poems, and scripts of crosstalk.
- `399dy.com (399导演社区)`: a website for Director's Club which is for public-available script resources or scripts uploaded by users.
- `xsxpw.com` (相声小品网): a website for categorized scripts for famous performers.

B METADATA OF DATA EXAMPLE

name	value
number of characters	484
file path	bijianshang/1386236043493249024.txt
id	1386236043493249024
index	1341
role map	""Jin Fei":"0","Chen Xi":"1""
number of utterances	43
source	"www.bijianshang.com/news/html/4826.html
title	The eight characters for fortunate
type	dual performing

Table 9: Example of metadata

The metadata is organized as Tab. 9. We include:

- 1) *charsize*: the length of the script in terms of character number,
- 2) *filePath*: relative path of the script file,
- 3) *id*: unique id of the script,
- 4) *idx*: the serial number of the script,
- 5) *roleMap*: a map to map involved characters to a specific character id,
- 6) *utteranceSize*: the number of utterances (utterance) in the script,
- 7) *title*: the title of the script,
- 8) *type*: the type of the script, e.g., a monologue, dual dialogue or multiple-performer dialogue.

C HYPERPARAMETERS FOR TRAINING MODELS

Tab. 10 shows the main hyperparameters for training. Unmentioned hyperparameters are set in default. For input, we append `[CLS]` at the beginning of each text and use `[SEP]` as the separator between utterances. Here is an example of the input format in LSTM Seq2seq, GPT, T5, UniLM:

[CLS]今天来说个相声。[SEP] 好咧 [SEP] 说点儿啥呢? [SEP] ··· [SEP]

[CLS] Let's have a crosstalk [SEP] well [SEP] what to talk about? [SEP] ··· [SEP]

models	epoch	batch size	learning rate	optimizer	others
LSTM Seq2seq	100	64	1e-05	AdamW	dropout=0.25 embed-size=300 vocab-size=7446 hidden-size=256
UniLM	100	64	1e-05	AdamW	adam-epsilon=1e-08 max-seq-length=256 warmup-proportion=0.1 weight-decay=0.01
T5	100	24	1.5e-04	AdamW	gradient-accumulation-steps=4 max-grad-norm=2.0 max-len=256 warmup-rate=0.1
GPT	100	64	1.5e-04	AdamW	gradient-accumulation-steps=4 max-grad-norm=2.0 max-len=256 warmup-rate=0.1
GPT-3	4	1	0.1	-	model=Davinci prompt-loss-weight=0.1

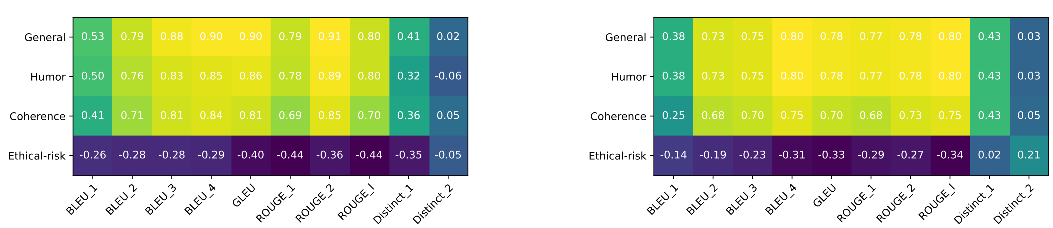
Table 10: Hyperparameters for training models.

To fine-tune GPT-3, we use the end-of-line （EOL) token as the separator between utterances, because [CLS] and [SEP] are not used in GPT3. We consider the first ten utterances as the *prompt* and the latter ten utterances as the *completion* part. utterances that are out of the first 20 positions are truncated due to the length limit. 0 is for the Dougen and 1 is for penggen.

Below is an example for the input json to fine-tune GPT-3.

{"prompt": "Specific information: 一段名称为《师徒俏皮话》的对口相声\n
0:这位是我师傅。\n 1:她是我徒弟。\n 0:虽然我是徒弟，但我可比他会的
多。\n 1:你才学几天呢?这么膨胀。\n 0:那当然了，我比你会多啦。\n 1:会
什么你啊?说相声基本功，俏皮话，听说过吗?\n 0:不是听说过吗?这样吧，
当场我就能为您新编一个专属的俏皮话。\n 1:给我编?这得听听。\n 0:好
吧，说到我师傅呀。\n 1:就是我呀。\n 0:", "completion": "他是乾隆年的
电灯管。\n 1:这话怎么讲?\n 0:老光棍咯。\n 1:你当着这么些人，你说这
干嘛呀?\n 0:我怕人家当你是花木兰的兔子。\n 1:这什么意思?\n 0:难辨雄
雌。\n 1:大伙看看，我长这么man，我像兔子吗?\n 0:看看，我是不是比你
会的多?是不是?\n 1:你要这个态度啊，今儿我得跟你比一比。\n"}

D Correlation between human and automatic evaluation



(a) Pearson correlation

(b) Spearman correlation

Figure 1: Correlation between **automatic** evaluation and **human** assessment, according to the performance of models in Tab. 5 and Tab. 7.

As seen from Tab. 1, the general quality and fluency from a human perspective are, at least from the statistical view, highly correlated with some automatic metrics (e.g., BLUE-4, GLEU, and ROUGE-2).

Note that the models that are used to calculate Pearson/Spearman correlation are mostly fine-tuned on the train set of C^3 (except for GPT3-Davinci); therefore they are more likely to generate C^3-style scripts. When evaluating these fine-tuned models in C^3 test set that is similar to the train set, it might overestimate the correlation between automatic evaluation and human assessment. Interestingly, it shows a different trend when comparing the original GPT-3 and fine-tuned GPT-3; fine-tuned GPT-3 underperforms in automatic evaluation but outperforms human assessment.

E WEB UI OF THE HUMAN ANNOTATIONS

The Web UI is like Fig. 2 and its mobile version is in Fig. 3.

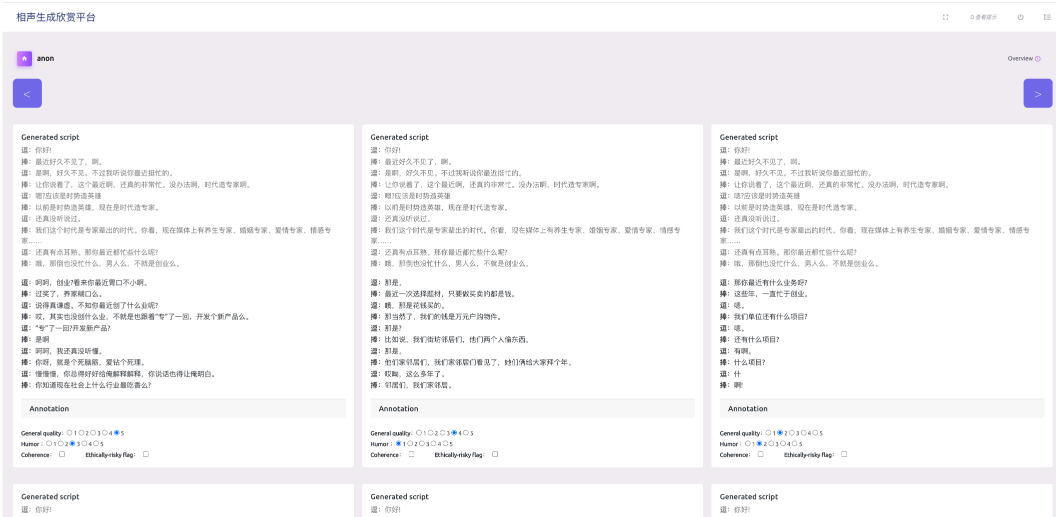


Figure 2: PC Web UI for human annotations

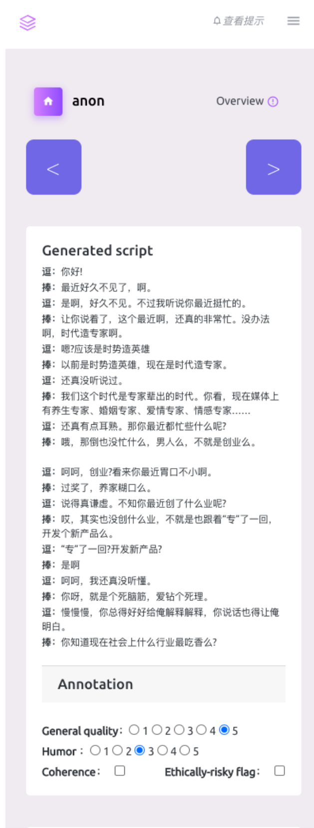


Figure 3: Mobile Web UI for human annotations