A Corpus for Understanding and Generating Moral Stories

Jian Guan, Ziqi Liu, Minlie Huang*
The CoAI group, DCST; Institute for Artificial Intelligence; State Key Lab of Intelligent Technology and Systems; Beijing National Research Center for Information Science and Technology; Tsinghua University, Beijing 100084, China.
{j-guan19, liuzq19}@mails.tsinghua.edu.cn,
aihuang@tsinghua.edu.cn

Abstract

Teaching morals is one of the most important purposes of storytelling. An essential ability for understanding and writing moral stories is bridging story plots and implied morals. Its challenges mainly lie in: (1) grasping knowledge about abstract concepts in morals, (2) capturing inter-event discourse relations in stories, and (3) aligning value preferences of stories and morals concerning good or bad behavior. In this paper, we propose two understanding tasks and two generation tasks to assess these abilities of machines. We present STORAL, a new dataset of Chinese and English human-written moral stories. We show the difficulty of the proposed tasks by testing various models with automatic and manual evaluation on STORAL. Furthermore, we present a retrieval-augmented algorithm that effectively exploits related concepts or events in training sets as additional guidance to improve performance on these tasks.

1 Introduction

Stories play an essential role in one's moral development (Vitz, 1990). For example, individuals usually learn morals from life experiences or literature such as fables and tell their morals by representing their lived experience in a narrative form (Tappan and Brown, 1989). Accordingly, it is a crucial ability for humans to bridge abstract morals and concrete events in stories. However, this ability has not yet been investigated for machines.

There have been many tasks proposed for evaluating story understanding and generation, including story ending selection (Mostafazadeh et al., 2016) and story generation from short prompts (Fan et al., 2018). Unlike these tasks, which focus on reasoning plots from context, we emphasize the ability to associate plots with implied morals. As exemplified in Table 1, the challenges mainly lie in (1) grasping knowledge about abstract concepts (e.g., "unity," "strength") and relations among them (e.g., "is") in morals; (2) capturing inter-event discourse relations in stories (e.g., the contrast between endings of the "cows" when they are "united" and "divided"); and (3) aligning value preferences (Jiang et al., 2021) of stories and morals (e.g., the story implies support for "unity", not opposition, which agrees with "is strength" in the moral). To test these abilities of machines, we propose two understanding tasks and two generation tasks. Both understanding tasks require selecting the correct moral from several candidates given a story. And they have respective candidate sets for testing machines in two aspects, including concept understanding (MOCPT for short) and preference alignment (MOPREF for short). The generation tasks require concluding the moral of a story (ST2MO for short), and conversely generating a coherent story to convey a moral (MO2ST for short).

Stories: Four cows lived in a forest near a meadow. They were good friends and did everything together. They grazed together and stayed together, because of which no tigers or lions were able to kill them for food. But one day, the friends fought and each cow went to graze in a different direction. A tiger and a lion saw this and decided that it was the perfect opportunity to kill the cows. They hid in the bushes and surprised the cows and killed them all, one by one.
Morals: Unity is strength.

Table 1: An example in STORAL

Furthermore, we collected a new dataset named STORAL composed of 4k Chinese and 2k English human-written stories paired with morals through human annotation to address the above challenges. We call the Chinese dataset STORAL-ZH and the English dataset STORAL-EN, respectively. And we construct datasets for the proposed tasks based on STORAL. Our focus of morals is on the social set of standards for good or bad behavior and character, or the quality of being right, honest or accept-

*Corresponding author

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5069 - 5087

able (Ianinska and Garcia-Zamor, 2006). We conduct extensive experiments on the proposed tasks. Furthermore, we present a retrieval-augmented algorithm to improve model performance by retrieving related concepts or events from training sets as additional guidance. However, the experiment results demonstrate that existing models still fall short of understanding and generating moral stories, which requires a better modeling of discourse and commonsense relations among concrete events and abstract concepts [1].

2 Related Work

Story Datasets ROCStories (Mostafazadeh et al., 2016) and WritingPrompts (Fan et al., 2018) are two frequently used story datasets in related studies. The former consists of artificial five-sentence stories regarding everyday events, while the latter contains fictional stories of 1k words paired with short prompts. Besides, some recent works collected extra-long stories such as roleplayerguild (Louis and Sutton, 2018), PG-19 (Rae et al., 2020), and STORIUM (Akoury et al., 2020). Guan et al. (2022) proposed a collection of Chinese stories. These stories usually aim to narrate a coherent event sequence but not convince readers of any morals.

Story Understanding and Generation There have been many tasks proposed for evaluating story understanding and generation. Firstly, some works tested the machinery commonsense reasoning ability regarding inter-event causal and temporal relations through story ending selection (Mostafazadeh et al., 2016), story ending generation (Guan et al., 2019) and story completion (Wang and Wan, 2019). Secondly, a series of studies focused on the coherence of story generation (Fan et al., 2018; Yao et al., 2019; Guan et al., 2020). Another line of works concentrated on controllability to impose specified attributes into story generation. These attributes involved outlines (Rashkin et al., 2020), emotional trajectories (Brahman and Chaturvedi, 2020) and story styles (Kong et al., 2021). Our tasks investigate not only the above aspects but also the ability to understand abstract concepts and reason value preferences of stories.

A task similar to ST2MO is text summarization (Finlayson, 2012) since both tasks require generating a short text to condense crucial information of a long text. But summarization requires reorganizing a few words of the original text instead of concluding a character-independent moral. For example, a plausible summary of the story in Table 1 is "Four cows were killed by two tigers and a lion" (generated by $\text{BART}_{\text{Large}}$ (Lewis et al., 2020) fine-tuned on a summarization dataset XSUM (Narayan et al., 2018)), which includes specific characters and events of the original story. Moreover, MO2ST is similar to persuasive essay generation (Stab and Gurevych, 2017), which also requires conveying a viewpoint in generated texts. However, persuasive essays usually convince readers by directly presenting arguments but not narrating a story.

Morals Haidt and Joseph (2004) provided a theoretical framework named Moral Foundations Theory (MFT) to summarize five basic moral foundations such as "Care/Harm," "Fairness/Cheating," etc. Based on the theory, recent studies have explored to classify the moral foundations of partisan news (Fulgoni et al., 2016), tweets (Johnson and Goldwasser, 2018; Hoover et al., 2020), and crowd-sourced texts (Pavan et al., 2020). And Volkova et al. (2017) proposed identifying suspicious news based on the features of moral foundations. However, we focus on morals which are free-form texts far beyond the scope of the five categories in MFT. In addition, recent studies proposed multiple datasets for machine ethics research such as SBIC (Sap et al., 2020), Social Chemistry (Forbes et al., 2020), Moral Stories (Emelin et al., 2020), ETHICS (Hendrycks et al., 2021) and Scruples (Lourie et al., 2021). But these datasets focus more on how machines behave ethically in some scenario, while STORAL emphasizes the ability to conclude the moral implied by a story. Moreover, most cases in these datasets consist of short texts of descriptive ethical behavior, typically in the form of one sentence. In contrast, STORAL provided longer and more context-specific stories for moral understanding.

3 STORAL Dataset

We collected STORAL from multiple web pages of moral stories. All stories are allowed to use and redistribute for research and have been reviewed by the website editors as stated on the pages. We show the full list of links to these pages in Section A.1. After de-duplication, we collected 19,197 Chinese and 2,598 English raw texts. Then we adopted human annotation for decoupling the story and moral

[1] All data and evaluation scripts are available at https://github.com/thu-coai/MoralStory.

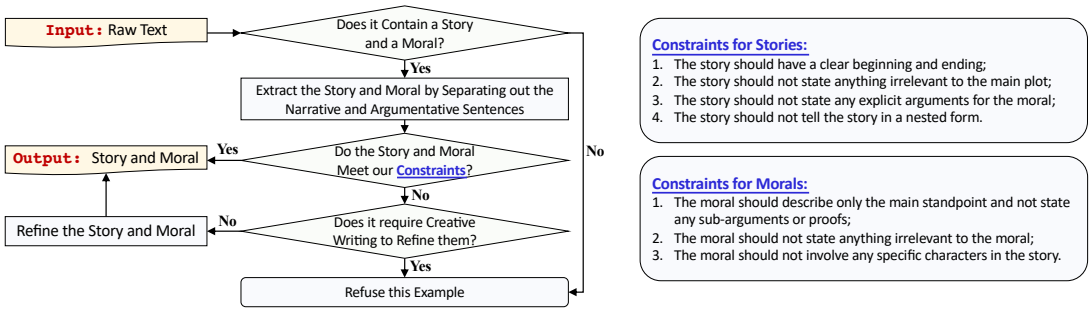


Figure 1: The pipeline of human annotation for constructing STORAL (Left) and our constraints (Right).

in each raw text. Due to resource limitations, we only constructed 4,209 Chinese and 1,779 English story-moral pairs. We will first show the details of human annotation, then present the topic analysis and statistics of STORAL, and finally describe the details of dataset construction for the proposed tasks.

3.1 Human Annotation

To narrow down our focus, we define a story as a series of coherent events involving several interrelated characters, and implies support or opposition of some behavior. Such a definition constrains the story to exhibit a moral without any explicit arguments. And we define a moral as a judgment to describe what the story implies concerning good or bad behavior. Note that we do not require morals in STORAL to be always reflective of normatively virtuous behavior. We emphasize that the morals should align with the story. Then, a key issue is how to extract the story and moral from a raw text. We observe that there are no markers such as "The story tells us" to separate the story and moral in most cases. The moral may be tightly weaved into the plot (e.g., included in a dialogue). Therefore, we adopted human annotation for this extraction task. We hired a commercial team to annotate STORAL-ZH. All annotators are native Chinese speakers and well trained for our task. For STORAL-EN, we hired three graduates with good English language proficiency. We did not use AMT since it is inconvenient to train online annotators. Figure 1 shows the annotation pipeline.

We first ask annotators to judge whether the raw text contains a story and moral and whether they meet our constraints shown in Figure 1. We show the examples given to the annotators to inform them of our requirements for stories and morals in Section A.2. If the constraints are not met, we then ask annotators to refine the story and moral. In the refinement stage, annotators have to clean up the data with following heuristics: (1) refusing examples which may violate general ethical principles (e.g., discrimination); (2) deleting noisy words (e.g., links, codes); (3) refining the stories and morals to be coherent and formal. And to ensure the quality of collected data, annotators may refuse to refine the example if it requires much creative writing. Finally, we review the annotation results and provide detailed feedback to the annotators before approving their submissions. We show an annotation example in Table 2.

Raw Text: A man who*wWw.xxx.c0M*lived a long time ago believed that he could read the future in the stars. He called himself an Astrologer, and spent his time at night gazing at the sky. One evening he was walking along the open road outside the village. His eyes were fixed on the stars. He thought he saw there that the end of the world was at hand, when all at once, down he went into a hole full of mud and water. There he stood up to his ears, in the muddy water, and madly clawing at the slippery sides of the hole in his effort to climb out. His cries for help soon brought the villagers running. As they pulled him out of the mud, one of them said:"You pretend to read the future in the stars, and yet you fail to see what is at your feet! *This may teach you to* ***pay more attention to what is right in front of you, and let the future take care of itself.****" "what use is it? " said another, " to read the stars, when you can't see what's right here on the earth?"*
Story: A man who lived a long time ago believed that he could read · · · As they pulled him out of the mud, one of them said: "You pretend to read the future in the stars, and yet you fail to see what is at your feet!"
Moral: Pay more attention to what is right in front of you, and let the future take care of itself.

Table 2: An example for extracting the story and moral from a raw text. We highlight the words which should be revised in the raw text in *italic*. And the moral in the raw text is **bold**. To save space, we replace some events with "· · ·" in the story.

3.2 Topic Analysis

To provide insight into the taxonomy of morals within STORAL, we adopt LDA (Blei et al., 2003) for topic modeling of morals. Let B denote the number of topics and V denote the vocabulary size. Based on the variational parameter for topic word distribution $\beta \in \mathbb{R}^{B\times V}$, we determine B as the minimum value that makes the following formula

Topic Words	Examples
懂得 (understand), 也是 (also), 了解 (know), 方法 (method), 收获 (gain), 保护 (protect), 大脑 (brain), 才能 (able), 付出 (pay), 进步 (progress)	在犯错的时候我们要懂得看全局，要了解全局才能对事情有定义。(When making mistakes, we must understand the overall situation. And we are able to have a definition of things only when knowing the overall situation.)
不要 (not), 一定要 (must), 危险 (danger), 时候 (when), 对待 (treat), 安全 (safety), 千万 (any way), 好好 (well), 学会 (learn), 遇到 (encounter)	生活中也要牢记“安全”这两字，在“安全”两字面前切不可存有侥幸心理，把安全当成儿戏。 (Keep in mind the word “safety” in your life, and do not take any chances to treat safety as a joke.)
事情 (thing), 才能 (able), 做好 (do well), 优秀 (excellent), 应该 (should), 做到 (achieve), 自信 (self-confident), 有所 (somewhat), 无法 (unable), 可能 (may)	做好自己该做的事情，做自己的主人。(Do what you should do and be your own master.)
时候 (when), 其实 (actually), 很多 (many), 发现 (discover), 希望 (wish), 发生 (happen), 生活 (life), 已经 (already), 伤害 (hurt), 可能 (may)	人要善于自己发现自己，而不是老等着别人来发现我们。(We should be good at discovering ourselves instead of waiting for others to do.)
遇到 (encounter), 问题 (question), 困难 (difficulty), 解决 (solve) 思考 (think), 帮助 (help), 时候 (when), 应该 (should), 给予 (give), 头脑 (brain)	乐于助人，是一种朴实的中国传统美德。每个人都有遇到困难的时候，最需要的是别人给予的帮助。(Being helpful is a Chinese traditional virtue. When someone encounters difficulties, what he needs most is help from others.)
good, always, come, believe, first, honesty, speak, world, around, act	**1.** Always be honest. Honesty is always rewarded. **2.** A liar will not be believed, even when he speaks the truth.
help, also, good, need, hope, lose, carry, feel, say, self	**1.** One should not be carried away by what others say. Don’t be fooled by those who wants to take advantage of you. **2.** Self help is the best help. Heaven helps those who help themselves.
friend, act, wisely, moment, think, place, time, choose, great, ability	**1.** Little friends may prove great friends. **2.** One should not panic in difficult times and think wisely.
love, care, parent, respect, always, value, take, mean, give, one	**1.** You reap what you sow. Regardless of your relationship with your parents, you’ll miss them when they’re gone from your life. Always respect, care for and love them. **2.** Be content with your lot; one cannot be first in everything.
look, see, bad, make, turn, strong, strength, choice, give, deserve	**1.** The strong and the weak cannot keep company. **2.** It is easy to despise what you cannot get.

Table 3: Topic words and examples for STORAL-ZH (top) and STORAL-EN (bottom). We underline the topic words that occurs in the examples.

holds true for any $b \in \{1, 2, \cdots, B\}$:

$$s_b = \frac{\sum_{v \in \mathcal{V}_b^{(k)}} \beta_{bv}}{\sum_{v=1}^{V} \beta_{bv}} \geq h,$$

$$\mathcal{V}_b^{(k)} = \text{argmax}_{\mathcal{V}_b^{*(k)}} \sum_{v \in \mathcal{V}_b^{*(k)}} \beta_{bv},$$

where β_{bv} is the element at the b-th row and v-th column of β, $k \in \{1, 2, \cdots, V\}$ is the size of the top-k vocabulary $\mathcal{V}_b^{(k)}$, and $h \in [0, 1]$ is a predefined threshold. s_b is used to measure the specificity of the b-th topic. Intuitively, the larger s_b, the more specific the topic. We set k to 20 and h to 0.5. Finally, we derive 40/24 topics for STORAL-ZH/STORAL-EN, respectively. And the minimum proportion of examples of one topic is 1.6%/3.2% for STORAL-ZH/STORAL-EN, respectively.

Table 3 shows the topic words in $\mathcal{V}^{(10)}$ of each topic and two morals assigned to each topic with the highest probabilities for the five topics with the largest specificity scores. The topics cover diverse situations ranging from facing others (“honesty,” “help”), parents (“love”), ourselves (“self-help,” “self-discovery”) to facing difficulties (“think”) and danger (“safety”). And examples of the same topic present related semantics to some extent, such as “being honest” and “not believing liars” for the first topic in STORAL-EN. We also show the analysis of high-frequency words of stories and morals in Section A.3 and discussion about the commonsense and discourse relations in stories in Section A.4.

3.3 Dataset Statistics of STORAL

Table 4 shows the statistics of STORAL. We regard the unlabeled data which contain entangled stories and morals as an in-domain resource for research on unsupervised or semi-supervised learning for the proposed tasks. And the data are also suitable for learning to generate morals stories where the morals are weaved naturally into the story plots.

3.4 Task-Specific Dataset Construction

Based on STORAL, we build task-specific datasets for our understanding tasks (MOCPT and MOPREF) and generation tasks (ST2MO and MO2ST). We randomly split the labeled data in STORAL-ZH and STORAL-EN for training/validation/testing by 8:1:1 and 3:1:1, respectively. Table 5 shows the task

Datasets	Labeled Data							Unlabeled Data			
	# Examples	Stories			Morals			# Examples	# Word	# Sent	Vocab
		# Word	# Sent	Vocab	# Word	# Sent	Vocab				
STORAL-ZH	4,209	321.75	17.62	63,493	25.09	1.35	10,522	14,988	487.00	26.12	147,805
STORAL-EN	1,779	302.33	17.71	15,873	19.77	1.45	3,384	819	614.55	38.05	20,853

Table 4: STORAL statistics. We use Jieba[2]/NLTK (Loper and Bird, 2002) for word tokenization of STORAL-ZH/STORAL-EN. **# Word / # Sent** is the average number of words/sentences. **Vocab** is the vocabulary size.

Tasks	Abilities	Inputs & Outputs	STORAL-ZH |Train| |Val| |Test|	STORAL-EN |Train| |Val| |Test|
MOCPT	Concept Understanding	Given a story and five candidate morals, choosing the correct moral.	3,368 / 420 / 421	1,068 / 355 / 356
MOPREF	Preference Alignment	Given a story and two candidate morals, choosing the correct moral.	3,276 / 410 / 411	988 / 344 / 339
ST2MO	Moral Generation	Given a story, generating a moral which is character-independent and generally applicable.	3,368 / 420 / 421	1,068 / 355 / 356
MO2ST	Story Generation	Given a moral and a story beginning and outline, generating a story which has a coherent plot and convinces readers of the moral.	3,368 / 420 / 421	1,068 / 355 / 356

Table 5: Description of the proposed tasks about the abilities they investigate, inputs and outputs, and the data sizes.

descriptions and data sizes.

MOCPT It requires selecting the correct moral from five candidates given a story. We constructed the dataset by taking the original moral as the correct candidate and four negatively sampled morals as incorrect candidates for each example. To avoid more than one plausible candidate, we ensured that the negative morals are assigned to different topics from the original one by the LDA model (Section 3.2). In this way, MOCPT can effectively test the ability to distinguish different concepts.

MOPREF It requires selecting the correct moral from two candidates. Its difference from MOCPT is that we created the incorrect candidate by substituting one random token in the original moral to its antonym. For example, the moral "unity is strength" can be transformed to "unity is weakness". We perform the transformation using a rule-based method (Ribeiro et al., 2020). Because there exist examples where no words have antonyms, the number of examples for MOPREF are a little fewer than MOCPT. MOPREF will serve for testing the ability to capture the value preference of stories.

ST2MO It requires generating the moral of a given story. We regard the original story as input and the original moral as target output.

MO2ST It requires generating a story to convey a given moral. Unfortunately, automatic evaluation for open-ended story generation is still highly challenging due to the notorious one-to-many issue (Zhao et al., 2017): There may be multiple plausible stories with the same moral. For example, the moral in Table 1 can also be conveyed by another story: "bees unite to build their beehives." Such openness makes automatic metrics unreliable for quality evaluation (Guan and Huang, 2020).

To alleviate this issue, we extract the first sentence and an outline from a target story, and pair them with the moral as input for generating the story. We follow Rashkin et al. (2020) to extract a set of at most eight phrases from a story through RAKE (Rose et al., 2010) as the outline. We set the maximum number of words in each phrase to eight. We also filtered those phrases that are substrings of others. For example, the outline for the story in Table 1 is {"lions," "friends fought," "good friends," "grazed," "perfect opportunity"}. Finally, for STORAL-ZH/STORAL-EN, the average number of phrases for each example is 7.5/6.8 and the average number of words in each phrase is 2.87/2.44, respectively.

4 Retrieval Augmentation

A critical challenge for tackling the proposed tasks is the sparsity of morals and events makes it difficult to learn relations between them. Prior studies have shown that retrieval improves performance towards infrequent data points across various tasks such as open-domain question answering (Chen et al., 2017) and text classification (Lin et al., 2021). We present a retrieval-augmented algorithm that exploits the moral-event relations in training sets. We illustrate our model for the MOPREF task in Fig-

[1] https://github.com/fxsjy/jieba

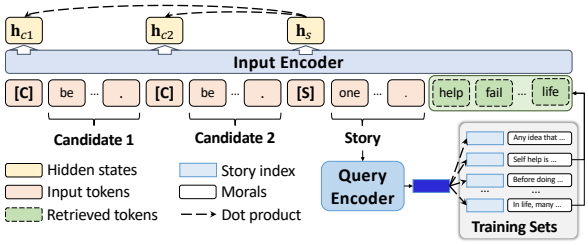


Figure 2: Model overview for the MOPREF task.

ure 2. Our models for other tasks are similar.

For both MOCPT and MOPREF, we encode the story and candidates using an input encoder, and then predict a probability distribution over the candidates by normalizing the dot-product scores between the representations of the story and each candidate. We optimize the model by minimizing the cross-entropy loss. We insert special tokens `[S]` and `[C]` before the story and each candidate, respectively, and take the corresponding hidden states as their representations. Furthermore, we propose to retrieve related concepts from the training set using the input story. We encode the story using a query encoder, then take the output as the query to retrieve m most related stories based on a story index, i.e., a set of dense vectors as the representations of stories in the training set. We adopt BERT (Devlin et al., 2019) followed by a mean-pooling layer to build the query encoder and story index, which are frozen in the training stage. Finally, we extract the nouns, verbs, adjectives and adverbs from the morals of the top-m stories and lemmatize them as the retrieved concepts. We feed the concepts together with the original input to the input encoder. For example, the retrieved concepts for the story in Table 1 include "support" and "strength", which may serve as additional guidance for models' prediction.

The retrieval-augmented algorithm can easily adapt to the generation tasks. For ST2MO, we take the input story paired with the retrieved concepts into the encoder and then generate the output using the decoder. And for MO2ST, we use the input moral as the query to retrieve top-m stories, and regard their outlines as the retrieved additional information to guide the subsequent story generation.

5 Experiments

5.1 Evaluated Models

We evaluated the following baselines for the understanding tasks: *BERT* (Devlin et al., 2019), *RoBERTa* (Liu et al., 2019) and *T5* (Raffel et al., 2020). When evaluating T5, we feed the input to both the encoder and decoder of T5 and optimize the model using the cross-entropy loss. To investigate potential biases of the proposed datasets, we added a baseline called *BERT w/o story*, which is fine-tuned to make prediction without taking the story as input. For the generation tasks, we evaluated *ConvS2S* (Gehring et al., 2017), *Fusion* (Fan et al., 2018), *GPT2* (Radford et al., 2019) and *T5*, which are trained or fine-tuned with the standard language modeling objective. Moreover, we also evaluate a task-specific model *PlotMachines* (PM for short) (Rashkin et al., 2020), which is proposed for tackling outline-conditioned generation by tracking the dynamic plot states. We use GPT2 as the backbone model of PM.

We also design models to test the adaption of the unlabeled data of STORAL to the proposed tasks. Specifically, we first post-train RoBERTa and T5 on the unlabeled data with their original pretraining objectives, respectively (i.e., masked language model and text infilling) and then fine-tune them on the labeled data for the downstream tasks (Gururangan et al., 2020). We call the baselines *RoBERTa-Post* and *T5-Post*. We perform our retrieval-augmented algorithm based on the post-trained models, called *RA-RoBERTa* and *RA-T5*, respectively.

5.2 Experiment Settings

We implement the pretrained models based on the codes and pretrained checkpoints of HuggingFace's Transformers (Wolf et al., 2020). We use $\text{LongLM}_{\text{base}}$ (Guan et al., 2022) as the T5 model for experiments on STORAL-ZH, and set all pretrained models to the base version due to limited computational resources. As for the hyperparameters, we set the batch size to 16, the maximum sequence length to 1,024, the learning rate to 3e-5, m to 10 for our retrieval-augmented model. We generate outputs using top-k sampling (Fan et al., 2018) with $k = 40$ and a softmax temperature of 0.7 (Goodfellow et al., 2016). We show more details in Section B.1.

5.3 Automatic Evaluation

Evaluation Metrics We adopt accuracy to evaluate the understanding tasks. For generation tasks, we do not use perplexity since perplexity scores are not comparable among models with different vocabularies. We adopt the following metrics for automatic evaluation: **(1) BLEU (B-n):** It is used

to measure n-gram overlaps ($n = 1, 2$) between generated and ground-truth texts (Papineni et al., 2002). **(2) BERTScore-F1 (BS):** It is used to measure the semantic similarity between generated and ground-truth texts (Zhang et al., 2019). **(3) Repetition (R-n):** It calculates the ratio of texts that repeat at least one n-gram in all generated texts (Shao et al., 2019). **(4) Distinct (D-n):** It measure the diversity using the percentage of distinct n-grams to all n-grams in generated texts (Li et al., 2016). For both R-n and D-n, we set $n = 2$ for ST2MO and $n = 4$ for MO2ST considering the much shorter length of morals than stories. Besides, we also report the average number of generated words (**Len**).

We also adopt the following metrics for automatic evaluation of MO2ST: **(1) Coverage (Cov):** It computes Rouge-L recall (Lin, 2004) between generated stories and phrases in the corresponding outlines. A higher score means the generated stories cover more phrases in the given outlines. **(2) Order (Ord):** It measures the disparity between the positional orders of given phrases in the ground truth and generated story using the percentage of inversions in the generated story (Guan et al., 2022). An inversion is a position pair of two phrases that is out of the ground-truth order. Higher order scores mean that the stories arrange the outline more reasonably. In Section B.2, we also construct a learnable automatic metric to measure the faithfulness between morals and stories.

Results Table 6 and 7 show the results on the understanding and generation tasks, respectively. To get the human performance on MOCPT and MOPREF, we sampled 100 examples from the test set and recruited three annotators with good Chinese or English language proficiency to complete these tasks. We made final decisions among the annotators through major voting. The annotation results show an almost perfect agreement with Fleiss's $\kappa > 0.85$ (Fleiss and Joseph, 1971).

We summarize the results on the understanding tasks as follows: **(1)** The MOPREF datasets suffer from innate biases as indicated by the high accuracy of BERT w/o story. Such biases may result from the noise introduced by the automatic construction technique, i.e., antonym substitution. And models may learn patterns of good behavior (e.g., "unity" is good and "disunity" is bad in general) and make predictions easily without depending on stories. However, MOPREF is still meaningful as an evaluation task since BERT can achieve much better accuracy when taking stories as input. And we experiment using manually constructed examples for evaluating preference alignment in the appendix. **(2)** T5 performs better than RoBERTa on MOCPT but worse on MOPREF, indicating T5 may not be good at capturing value preferences. **(3)** Post-training on the unlabeled data (i.e., RoBERTa-Post and T5-Post) does not always bring improvement on both tasks, suggesting that it is necessary to develop a better way to exploit these data in future work. **(4)** Retrieving additional concepts improves models' performance effectively, particularly for the MOCPT task on STORAL-EN. However, there is still a big gap between our models and human performance.

As for the generation tasks, we draw the following conclusions: **(1)** Almost all pretrained models achieve better lexical and semantic similarity with ground-truth texts than non-pretrained models, as indicated by higher BLEU and BERTScore values. **(2)** Non-pretrained models have less repetition than pretrained ones, and repeat even less than the ground-truth texts when generating morals. It may be because non-pretrained models generate shorter sequences than pretrained models despite the same decoding algorithm, which also accounts for the higher distinct scores of the non-pretrained models on the MO2ST task. **(3)** When generating stories, T5-Post can cover more input phrases and arrange them in a correct order than other baselines, as indicated by higher coverage and order scores. **(4)** Retrieval augmentation can improve the generation similarity with the ground-truth texts on both tasks and improve the coverage and order scores on ST2MO significantly compared with T5-Post.

Models	# P	MOCPT		MOPREF	
		ZH	EN	ZH	EN
Random	N/A	20.19	20.22	50.12	50.00
BERT w/o Story	110M	23.52	22.47	71.81	72.57
BERT	110M	59.62	51.97	82.97	79.35
RoBERTa	110M	62.71	54.78	<u>89.54</u>	81.12
RoBERTa-Post	110M	64.61	51.40	87.59	81.42
T5	220M	69.60	58.99	82.00	76.99
T5-Post	220M	<u>70.07</u>	<u>62.64</u>	81.75	77.29
RA-RoBERTa	110M	65.08	60.96	**90.02**	<u>81.71</u>
RA-T5	220M	**72.68***	**67.42****	82.97	**82.60**
Human	N/A	*95.00*	*96.00*	*98.00*	*99.00*

Table 6: Accuracy (%) for MOCPT and MOPREF. # P is the number of parameters. The best performance is highlighted in **bold** and the second best is underlined. The scores marked with * and ** of RA model mean it outperforms the best baseline significantly with p-value$<$0.1 and p-value$<$0.05 (sign test), respectively.

Models	# P	Dataset: STORAL-ZH						Dataset: STORAL-EN					
		B-1↑	B-2↑	BS↑	R-2↓	D-2↑	Len	B-1↑	B-2↑	BS↑	R-2↓	D-2↑	Len
ConvS2S	50M	14.31	1.86	56.71	**26.60**	43.67	19.31	9.69	0.93	82.57	6.46	47.35	11.75
Fusion	100M	14.78	2.23	56.90	27.55	41.21	21.96	9.87	0.82	82.68	**6.18**	43.59	13.15
GPT2	124M	14.54	2.16	60.75	35.39	48.22	20.72	10.98	1.24	79.39	20.22	60.36	16.19
T5	220M	18.19	3.60	61.61	76.48	44.84	29.06	13.31	2.26	85.89	33.15	58.73	19.39
T5-Post	220M	17.98	**3.91**	61.52	69.12	**51.97**	29.14	13.83	2.11	85.85	34.83	57.12	18.49
RA-T5	220M	**18.32**	3.64	**61.93****	70.78	48.14	29.44	**14.59**	**2.61**	**86.16****	31.46	**60.61**	18.54
Truth Morals	*N/A*	*N/A*	*N/A*	*N/A*	*29.22*	*73.70*	*25.09*	*N/A*	*N/A*	*N/A*	*16.85*	*73.95*	20.41

Models	Dataset: STORAL-ZH								Dataset: STORAL-EN							
	B-1↑	B-2↑	BS↑	R-4↓	D-4↑	Cov↑	Ord↑	Len	B-1↑	B-2↑	BS↑	R-4↓	D-4↑	Cov↑	Ord↑	Len
ConvS2S	15.57	6.43	60.00	75.30	78.41	21.61	33.03	150	16.25	6.38	79.27	61.85	**80.29**	6.46	41.88	122
Fusion	15.53	6.45	60.06	**74.11**	**80.51**	22.86	33.33	148	17.17	6.82	79.52	**61.24**	75.79	7.27	43.07	137
GPT2	14.91	6.48	63.32	91.45	58.67	48.57	51.58	282	25.83	12.91	83.25	84.27	74.63	45.18	59.95	247
PM	15.82	7.04	63.58	90.97	57.33	50.51	52.35	280	26.34	13.92	81.63	80.90	72.64	47.07	60.31	264
T5	17.74	9.44	65.89	91.69	61.76	58.18	56.11	166	30.56	16.75	79.89	90.17	77.53	74.21	63.45	283
T5-Post	18.42	9.77	65.63	94.54	58.13	60.11	56.96	176	32.36	18.04	83.80	94.10	77.27	76.09	64.33	281
RA-T5	**23.36** **	**12.98** **	**67.37** **	95.72	59.49	**69.24** **	**60.44** **	241	**32.46**	**18.31** *	**84.07** **	92.42	76.74	**80.21** **	**66.10** **	253
Truth	*N/A*	*N/A*	*N/A*	*55.34*	*96.06*	*100.00*	*100.00*	324	*N/A*	*N/A*	*N/A*	*58.71*	*95.09*	*100.00*	*100.00*	281

Table 7: Automatic evaluation results for ST2MO (Top) and MO2ST (Bottom). ↓ / ↑ means the lower/higher the better. All scores except *Len* are multiplied by 100. The best result is in **bold** and the second best is underlined. The scores marked with * and ** of RA-T5 mean it outperforms the best baseline significantly with p-value<0.1 and p-value<0.05 (sign test), respectively.

Data	Task	Model	Flu (κ)	Cohe (κ)	Faith (κ)
STORAL-ZH	ST2MO	**Fusion**	0.24 (0.31)	0.22 (0.37)	0.08 (0.72)
		T5	0.75 (0.40)	0.61 (0.38)	0.31 (0.32)
		RA-T5	**0.85** (0.65)	**0.63** (0.26)	**0.36** (0.27)
		Truth	*1.00* (1.00)	*0.99* (0.96)	*0.86* (0.57)
	MO2ST	**Fusion**	0.25 (0.39)	0.11 (0.61)	0.02 (0.93)
		T5	0.38 (0.40)	0.24 (0.37)	0.05 (0.81)
		RA-T5	**0.45** (0.29)	**0.34** (0.25)	**0.11** (0.72)
		Truth	*0.98* (0.93)	*1.00* (1.00)	*0.96* (0.84)
STORAL-EN	ST2MO	**Fusion**	0.32 (0.39)	0.26 (0.35)	0.24 (0.41)
		T5	0.76 (0.35)	0.74 (0.27)	0.55 (0.33)
		RA-T5	**0.81** (0.51)	**0.79** (0.40)	**0.67** (0.37)
		Truth	*0.94* (0.80)	*0.94* (0.77)	*0.88* (0.56)
	MO2ST	**Fusion**	0.47 (0.43)	0.40 (0.47)	0.37 (0.45)
		T5	0.56 (0.35)	0.48 (0.37)	0.49 (0.39)
		RA-T5	**0.58** (0.28)	**0.51** (0.31)	**0.57** (0.31)
		Truth	*0.95* (0.69)	*0.98* (0.69)	*0.93* (0.53)

Table 8: Manual evaluation results for ST2MO and MO2ST. Flu, Cohe and Faith mean *fluency*, *coherence* and *moral faithfulness*, respectively. The best results are highlighted in **bold**. All results show fair or moderate inter-annotator agreement measured by Fleiss' κ (Fleiss and Joseph, 1971).

5.4 Manual Evaluation

On the generation tasks, we conducted a Likert-scale based manual evaluation to measure the gap between existing models and humans. For STOAL-ZH, we hired three graduate students (native Chinese speakers) as annotators. We conducted evaluation on STORAL-EN using Amazon Mechanical Turk (AMT). For both tasks, we randomly sampled 100 examples from the test set, and obtained 300 generated texts from Fusion, T5 and RA-T5. For each text we require three annotators to rate its quality along with the input using a binary score in three following aspects: (1) *linguistic fluency*: correctness in grammaticality; (2) *coherence*: reasonable relations between sentences regarding relatedness, causality and temporal orders; and (3) *moral faithfulness*: exhibition of a faithful moral to the input. Three aspects are independently evaluated. We decided the final score of a text through majority voting. The annotation instruction is shown in Section B.3.

Table 8 shows the manual evaluation results. We show p-values of the results in Section B.4. For ST2MO, T5 achieves a substantial improvement compared with Fusion ($p < 0.01$), and our model further outperforms T5. The superiority becomes less significant for MO2ST. However, the big gap between these models and humans, particularly in terms of faithfulness, proves both tasks challenging for existing models. Furthermore, we evaluate whether machines can capture the value preference of a story using manually constructed examples. And we show error analysis and case study for the proposed tasks in Section C. We believe that explicit modeling of the relations among events and abstract concepts will further promote progress on these tasks, which we regard as future work.

6 Conclusion

We present STORAL, a collection of Chinese and English moral stories. To test the ability to bridge

concrete events and abstract morals, we propose new understanding and generation tasks based on STORAL, including selecting the correct moral from several candidates with different topics or opposite value preferences, concluding the moral of a story and generating a story to convey a moral. Extensive experiments prove these tasks still to be challenging for existing models. We propose a retrieval-augmented algorithm to improve performance by retrieving related concepts or events from training sets. Although it is possible to further increase the dataset size, we expect to make meaningful progress by developing better representations of commonsense and discourse relations among events and abstract concepts in future work.

7 Acknowledgement

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604) and the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005. This work was also sponsored by Tsinghua-Toyota Joint Research Fund. We would also like to thank the anonymous reviewers for their invaluable suggestions and feedback.

8 Ethics Statements and Broader Impact

We collected STORAL from public web resources. All stories are under licenses that allow use and redistribution for research purposes. We asked commercial annotation teams to extract stories and morals from the crawled raw texts. We required the annotators to refuse the examples which violate general ethical principles (e.g., showing discrimination for someone, containing disrespectful content, or encouraging to disturb public order, etc.). Totally, we payed more than $7 (CNY 45) per hour on average for annotating each example in STORAL, which was far beyond the minimum hourly wage in China (CNY 21). Furthermore, we resorted to AMT for manual evaluation of generated and human-written texts for two proposed generation tasks. We hired three annotators and payed each annotator $0.2 on average for annotating each example.

In this paper, we emphasize the ability to model relations between concrete events and abstract morals, which is also helpful for various scenarios such as reading comprehension (e.g., drawing authors' viewpoints from narratives) and essay writing (e.g., writing essays to convince readers of some arguments by presenting examples or anecdotes). STORAL provides a good start point for exploring these directions.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 889–898.

Mark Mark Alan Finlayson. 2012. *Learning narrative structure from annotated folktales*. Ph.D. thesis, Massachusetts Institute of Technology.

Fleiss and L. Joseph. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation. *Transactions of the Association for Computational Linguistics*, 10:434–451.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jian Guan and Minlie Huang. 2020. UNION: an unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9157–9166. Association for Computational Linguistics.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Silvana Ianinska and Jean-Claude Garcia-Zamor. 2006. Morals, ethics, and integrity: How codes of conduct contribute to ethical adult education practice. *Public Organization Review*, 6(1):3–20.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. 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, pages 110–119.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gcn and bert. *arXiv preprint arXiv:2105.05727*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matheus C Pavan, Vitor G Dos Santos, Alex GJ Lan, Joao Martins, Wesley R Santos, Caio Deutsch, Pablo B Costa, Fernando C Hsieh, and Ivandre Paraboni. 2020. Morality classification in natural language text. *IEEE Transactions on Affective Computing*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3257–3268, Hong Kong, China. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Mark Tappan and Lyn Mikel Brown. 1989. Stories told and lessons learned: Toward a narrative approach to moral development and moral education. *Harvard Educational Review*, 59(2):182–206.

Paul C Vitz. 1990. The use of stories in moral development: New psychological reasons for an old education method. *American psychologist*, 45(6):709.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653.

Tianming Wang and Xiaojun Wan. 2019. T-CVAE: transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5233–5239. ijcai.org.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

A STORAL Construction

A.1 Data Source

We show the full list of web pages used for constructing STORAL in Table 11. We initially collect 52,017 Chinese and 2,630 English raw texts from the web pages. Then we de-duplicate the texts by removing those texts which overlap with others more than twenty words. After de-duplication, we finally collected 19,197 Chinese and 2,598 English texts. And we construct STORAL based on these texts.

A.2 Data Annotation

Table 9/10 shows the examples given to the annotators to inform them of the requirements for stories/morals, respectively. If the constraints are not met, we ask annotators to refine the story and moral. All workers were paid more than $7 per hour on average.

Example 1: *Come on Bear! What a beautiful day! Go for a walk with your father! Take a deep breath and smell the flowers. But don't pick the flowers. Listen to the birds sing. But don't scare them. How beautiful the world is. Isn't it, dear Bear?*
Example 2: *When I was a child, I heard a story that felt very regrettable. I felt sorry for the protagonist of the story.* Long ago, there lived · · · Such trees are now found all over Uganda.
Example 3: I have a well-off friend. When she first entered college, she had many good wishes and thought she could achieve her goals. · · · Now she felt very painful under the strong mental pressure. *I can understand her feelings. · · · If magnifying your own pain, you will get trapped in the mire of your pain, and even feel that life is too unfair to you.*
Example 4: *Raul sat at his door, frowning. · · · His father told Raul a true story:* A wild wolf escaped into a cave after being wounded by a hunter's arrow. · · · *After hearing the story, Raul cheered up immediately. · · ·*

Table 9: Examples of stories provided for the annotators. Each example does not meet one of the following requirements in order: (1) having a clear beginning and ending; (2) not stating anything irrelevant to the main plot; (3) not stating any explicit arguments for the moral; and (4) not telling the story in a nested form. The sentences causing the above issues are in *italic*.

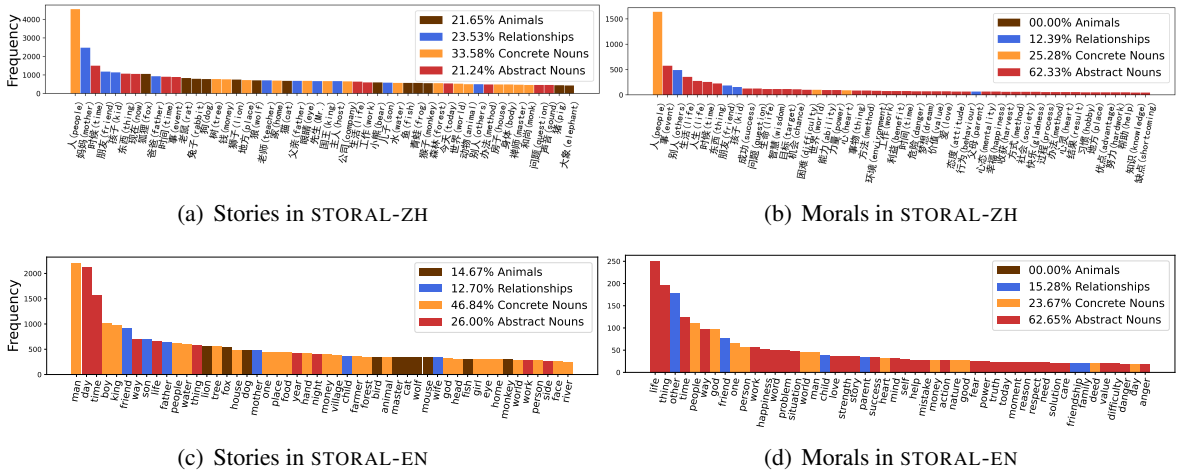
Example 1: *If you saw a thief in a crowded bus, would you bravely stop him?* Please reflect on yourself instead of just complaining that our world is becoming worse. *Without the foothold for dirt, the flower of civilization is bound to be fragrant.*
Example 2: *The story tells us: we should remember that* we should become a polite person and communicate with others carefully.
Example 3: As long as you keep your sanity and make right judgments, all the barriers will not become an obstacle, *just like the beautiful girl in the story*.

Table 10: Examples of morals provided for the annotators. Each of the examples does not satisfy one of the following constraints in order: (1) describing only the main standpoint and not stating any sub-arguments or proofs, and (2) not stating anything irrelevant to the moral, and (3) not involving any specific characters in the story. We highlight the sentences leading to the above issues in *italic*.

Links	Number
http://www.qbaobei.com/jiaoyu/yegs/yygs/	14,674
https://www.517gj.com/yuyangushi/	14,474
https://www.etgushi.com/zgyy/	6,691
https://www.chazidian.com/gushi_1/	3,457
http://www.feel-bar.com	3,329
http://www.xiaole8.com/renshengzheli/	2,509
http://www.zuowen.com/sucai/zheli/	2,421
http://www.rensheng5.com	2,092
https://www.yuyangushi.com/lz/xgsddl	1,886
http://www.gushi88.cn/ErTong/ZhongGuoYuYan_1	484
Grand Total	52,017

Links	Number
https://moralstories26.com	799
https://english.7139.com/2539/	552
https://kidsfables.com	193
http://read.gov/aesop	145
http://www.taleswithmorals.com	108
https://www.studentuk.com/category/fable	101
http://www.english-for-students.com/Moral-Stories.html	97
https://www.advance-africa.com/English-Moral-Stories.html	65
https://www.gutenberg.org/files/25512/25512-h/25512-h.htm	52
Others	518
Grand Total	2,630

Table 11: List of source web pages used for constructing STORAL-ZH (Top) and STORAL-EN (Bottom). Numbers in the right column means the number of raw texts initially collected from the corresponding web page.



(a) Stories in STORAL-ZH

(b) Morals in STORAL-ZH

(c) Stories in STORAL-EN

(d) Morals in STORAL-EN

Figure 3: Top 50 most frequent nouns for stories and morals in STORAL-ZH and STORAL-EN. The numbers in the legend show the percentages of the total frequency of the nouns of the same type among the 50 nouns.

A.3 Analysis of High-Frequency Words

To investigate the topic features of STORAL, we count the top 50 most frequent nouns in STORAL (excluding stop words) as shown in Figure 3. We roughly categorize these words into four types: **(1) Animals**: animals are popular as protagonists in moral stories since they usually have various but clear characteristics (e.g., "sly foxes"), which embody rich commonsense knowledge; **(2) Relationships**: such nouns are used to describe the inter-character relationships in a story (e.g., "friend"), which are useful for modeling characters' motivation and behavior; **(3) Concrete nouns**: they refer to physical entities that can be observed, such as "water"; and **(4) Abstract nouns**: they refer to abstract concepts, such as "difficulty". We manually check the proportional distribution of the four types for stories and morals, respectively. The results in Figure 3 demonstrate that morals contain significantly less concrete nouns and more abstract nouns than stories. And morals contain little animal words but almost as many relationship words as stories, indicating that morals may be independent of specific characters but relate to general interpersonal relations. The result shows that morals are more abstract than stories.

Furthermore, Table 12 shows the most frequent 4-grams in STORAL, further indicating that morals are more abstract than stories. Each of the 4-grams in Table 12 comprises less than 0.01% of all 4-

Stories	Morals
Dataset: STORAL-ZH	
as one is walking	we should be a
say to him that you	everyone has
say after thinking	everyone has his own
the most in the world	has own
all the animals	each of us
all the persons	we should know to
a place far away	for anything, we should
the dad of the pink pig	be one who knows to
this is my	for anything, be
in the forest there lived a	is a true
Dataset: STORAL-EN	
once upon a time	we should try to
upon a time there	the best way to
a time there was	it is better to
time there was a	it is easy to
there was once a	we should learn to
once there was a	those who help themselves
was not able to	with what we have
as soon as he	be happy with what
and asked him to	we should not judge
did n't want to	look before you leap

Table 12: Top 10 most frequent 4-grams in STORAL-ZH and STORAL-EN respectively. The Chinese 4-grams in STROAL-ZH are translated into English.

grams in the corresponding dataset, showing the diversity of STORAL.

A.4 Discussion about STORAL

The high-quality examples in STORAL are full of commonsense and discourse relations. As exemplified in Table 1 in the main paper, the common sense is mainly regarding the characters' reaction and intention (e.g., "the cows dispersed" and then the "tiger" and "lion" intend to kill them), as well as the nature of physical objects and abstract concepts (e.g. "cows" may be the food of "lions" and "tigers", and "unity" refers to "keeping together for a common goal"). Additionally, the stories usually have a specific discourse structure, i.e., the premise to introduce the story settings (e.g., the characters "four cows" and the location "a meadow"), the right or wrong behavior ("stay together or not") and the endings ("living well or being killed"). We believe it is an essential topic of future work to develop a better approach to model such commonsense and discourse relations.

B Experiments

B.1 Implementation

We implement the pretrained models used in our experiment mainly based on the register models of HuggingFace (Wolf et al., 2020). Table 13 shows the names of the used register models. Note that we use $\text{LongLM}_{\text{base}}$ (Guan et al., 2022) as the T5 model for experiments on STORAL-ZH, which has not been registered on HuggingFace.

All results in the main paper and the appendix are based on one NVIDIA Tesla v100 (16G memory). All reported results are based on one single running. The CPU is Intel Xeon Gold 5218. It cost less than 5 hours for fine-tuning each model on STORAL. We set the hyper-parameters following the default parameters of HuggingFace.

B.2 Automatic Evaluation for Moral Faithfulness

We follow Guan and Huang (2020) to train a learnable metric to evaluate moral faithfulness. Specifically, we fine-tune $\text{RoBERTa}_{\text{BASE}}$ as a classifier to distinguish whether a story matches a moral. We regard ground-truth examples as positive where the story and moral are matched, and construct negative examples by replacing the story or moral with a randomly sampled one. Finally, the classifier achieves an accuracy of 77.32/79.21% on the data constructed based on the test set of STORAL-ZH/STORAL-EN respectively. Then we calculate the faithfulness score as the average classifier score of all generated texts for the inputs.

Table 14 presents the evaluation results. We can see that pretrained models achieve better faithfulness than the non-pretrained models as shown by the much higher faithfulness scores. However, we also observe that the faithfulness score of the ground-truth texts is lower than some models (e.g., T5) when generating morals. Therefore, it is still necessary to manually evaluate faithfulness.

Results on Validation Sets We show the performance of several baselines and RA-T5 on the validation sets of the understanding tasks and the generation tasks in Table 15 and Table 16, respectively.

B.3 Manual Evaluation Instruction

We show the manual annotation interface in Figure 4. To ensure that the annotators guarantee a consistent standard in the annotation process, we asked annotators to rate four examples with the same input at the same HIT (human intelligence task). In these four examples, one is written by humans and three are generated by models (i.e., Fusion, T5 and RA-T5). We payed each annotator $0.2 on average for annotating each example.

Datasets	STORAL-ZH	STORAL-EN
BERT	bert-base-chinese (Devlin et al., 2019)	bert-base-uncased (Devlin et al., 2019)
RoBERTa	hfl/chinese-roberta-wwm-ext (Cui et al., 2020)	roberta-base (Liu et al., 2019)
GPT2	uer/gpt2-chinese-cluecorpussmall (Zhao et al., 2019)	gpt2 (Radford et al., 2019)
T5	LongLM (Guan et al., 2022)	t5-base (Raffel et al., 2020)

Table 13: Names of register models used in our experiment.

Models	ST2MO		MO2ST	
	ZH	EN	ZH	EN
ConvS2S	31.92	28.68	33.44	35.59
Fusion	33.85	25.23	38.81	35.16
GPT2	68.52	73.73	50.49	64.90
PM	N/A	N/A	52.41	62.53
T5	89.20	90.57	56.11	63.45
T5-Post	**90.98**	**91.87**	58.58	**75.67**
RA-T5	86.50	88.69	**59.50**	74.92
Truth	*77.49*	*80.03*	*77.49*	*80.03*

Table 14: Automatic moral faithfulness scores. The score of PM for the ST2MO task is N/A since we do not experiment with PM for this task.

Models	# P	MoCPT		MoPREF	
		ZH	EN	ZH	EN
BERT w/o Story	110M	20.71	21.69	72.64	77.62
BERT	110M	65.24	54.08	85.37	79.36
RoBERTa	110M	66.90	61.69	90.49	80.52
RoBERTa-Post	110M	67.14	55.77	89.27	84.01
T5	220M	74.52	62.25	78.05	77.91
T5-Post	220M	74.05	67.61	81.22	81.10
RA-RoBERTa	110M	66.43	63.94	88.54	86.63
RA-T5	220M	74.05	67.61	80.73	80.23

Table 15: Accuracy (%) for MOCPT and MOPREF on the validation set.

Models	ST2MO		MO2ST	
	ZH	EN	ZH	EN
Fusion	14.44/1.80	10.78/0.92	16.06/6.44	16.74/6.72
T5	18.54/4.08	13.17/2.05	18.98/10.17	28.87/15.48
RA-T5	18.68/3.64	14.49/4.47	23.98/13.17	31.72/17.97

Table 16: BLEU-1/BLEU-2 for ST2MO and MO2ST on the validation set.

B.4 Significance of Manual Evaluation Results

Table 17 shows the p-values (sign test) when comparing the manual evaluation results (Table 8 in the main paper) between each pair of the ground truth, Fusion, T5 and RA-T5.

B.5 Evaluating Value Preference Alignment

Although we have used MOPREF to evaluate whether machines can capture the value preference

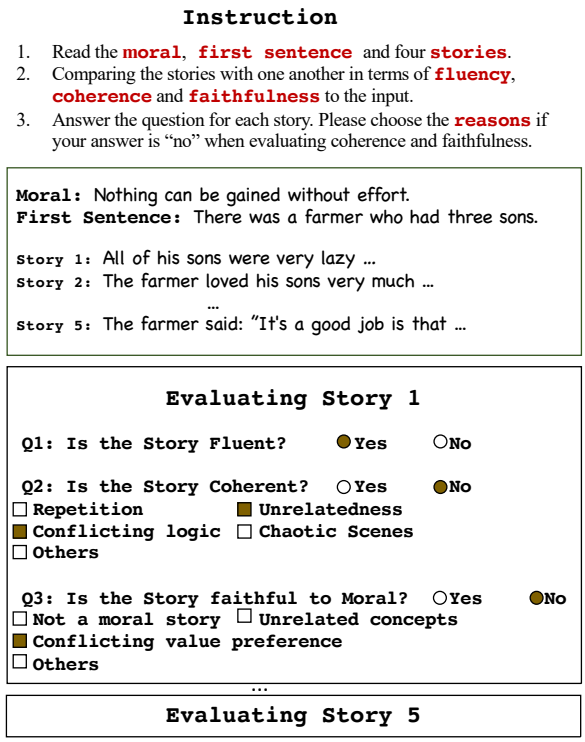


Figure 4: A simplified version of the manual annotation interface for MO2ST. The interface for ST2MO is similar.

of a story, the automatically constructed dataset may bias machines to focus on distinguishing general standards of good behaviour without considering story plots. Therefore, in this section, we construct examples manually to test this ability beyond the token level. Specifically, we randomly sampled 50 examples from the test sets of STORAL-ZH and STORAL-EN respectively. For each example, we manually rewrote the moral to convey a synonymous or antonymous value preference. For example, a synonymous moral with "unity is strength" in Table 1 can be "we are powerful as long as we unite with each other" and an antonymous one can be "everyone can also be powerful enough." Then we expect a model to be able to accept the synonymous moral but reject the antonymous one. We use three typical models, including BERT w/o Story, RA-RoBERTa and RA-T5, to compute the winning

Tasks	Models	Flu	Cohe	Faith
Dataset: STORAL-ZH				
ST2MO	**T5 vs. Fusion**	8.55e-14	3.82e-11	1.55e-6
	RA-T5 vs. T5	0.03	**0.75**	**0.23**
	Truth vs. RA-T5	6.10e-5	2.91e-11	2.35e-14
MO2ST	**T5 vs. Fusion**	2.35e-3	2.44e-4	**0.38**
	RA-T5 vs. T5	**0.14**	0.04	**0.11**
	Truth vs. RA-T5	2.27e-12	1.08e-19	1.1e-24
Dataset: STORAL-EN				
ST2MO	**T5 vs. Fusion**	3.71e-11	2.92e-12	7.92e-9
	RA-T5 vs. T5	**0.27**	0.06	4.18e-3
	Truth vs. RA-T5	7.20e-3	4.08e-3	4.92e-5
MO2ST	**T5 vs. Fusion**	0.04	0.09	0.04
	RA-T5 vs. T5	**0.5**	**0.25**	7.81e-3
	Truth vs. RA-T5	1.46e-11	1.42e-14	1.71e-8

Table 17: p-values (sign test) when comparing each pair of the ground truth and three models for the manual evaluation results. We highlight the p-values larger than 0.1 in **bold**, which indicates A has an insignificant superiority w.r.t. B for "A vs. B".

rate of pair-wise comparisons between any two of ground-truth, synonymous and antonymous morals. These models are trained on the training set of the MOPREF task.

Models	True vs. Syn	True vs. Ant	Syn vs. Ant
Dataset: STORAL-ZH			
BERT w/o Story	52% (0.89)	46% (0.67)	58% (0.32)
RA-RoBERTa	40% (0.21)	36% (0.06)	48% (0.89)
Dataset: STORAL-EN			
BERT w/o Story	54% (0.67)	54% (0.67)	48% (0.89)
RA-RoBERTa	64% (0.06)	34% (0.03)	40% (0.20)

Table 18: Winning rates of pair-wise comparisons which require selecting a correct moral from two candidates. Each candidate is a ground-truth (**True**), synonymous (**Syn**), or antonymous (**Ant**) moral. The number in the parenthesis is the corresponding p-value (sign test).

Table 18 shows the evaluation results. We observe that BERT can not distinguish different types of morals without input stories. RA-RoBERTa fails to accept the synonymous morals on STORAL-EN (winning rate of only 36% w.r.t the ground truth, $p < 0.1$), and can not distinguish synonymous and antonymous morals on both STORAL-ZH and STORAL-EN (winning rate near 50% with $p > 0.1$). Additionally, it prefers antonymous morals to the ground truth significantly on both datasets (winning rate less than 50% and $p < 0.1$). The results indicate that existing models still struggle to capture the value preference of moral stories.

Models	NAM	UNREL	CONF	Others
Task: ST2MO				
Fusion	27%	23%	7%	2%
T5	19%	9%	12%	0%
RA-T5	15%	7%	6%	0%
Truth	3%	4%	2%	0%
Task: MO2ST				
Fusion	25%	13%	6%	1%
T5	19%	9%	10%	1%
RA-T5	16%	10%	10%	0%
Truth	2%	1%	1%	0%

Table 19: Percentage of the texts annotated with a certain error in all annotated 100 texts in terms of moral faithfulness.

C Error Analysis and Case Study

In this section, we conducted a case study and investigated the errors of existing models on the proposed tasks to provide insight into future work. We show several typical error cases in Table 20.

C.1 Understanding Tasks

The example in Table 20 for MOCPT shows that the model may not grasp abstract concepts such as "good will" and "good acts" and align them to the story plots. It makes predictions possibly based on only token-level features such as relations between "ask after" and "attention". On the other hand, the example for MOPREF indicates that the model can not capture the value preference of the story in terms of "whether it is intelligent to regard others are illiterate". The results demonstrate the necessity of introducing concept knowledge and modeling high-level semantic information.

C.2 Generation Tasks

Table 21 shows cases generated by several baselines and our model for the generation tasks. We can see that retrieval can provide effective guidance for both moral and story generation. Baseline models including GPT2 and T5 tend to generate unrelated concepts or non-moral texts.

However, as shown by the manual evaluation results, there is still a big gap between RA-T5 and humans. To provide quantitative error analysis, in the process of manual evaluation on STORAL-EN, we required annotators to annotate the error type of a text when it exhibit an unfaithful moral. We summarize three main error types as follows: **(1) Not a moral text (NAM)**: not stating or imply-

Understanding Task: MOCPT
Input Story: A stag had fallen sick. He had just strength enough to gather some food and find a quiet clearing in the woods, where he lay down to wait until his strength should return. The animals heard about the stag's illness and came to ask after his health. Of course, they were all hungry, and helped themselves freely to the stag's food; and as you would expect, the stag soon starved to death.
Candidate Moral 1: Good will is worth nothing unless it is accompanied by good acts. **Candidate Moral 2:** Every man in need is your neighbor. **Candidate Moral 3:** Your everyday good deeds never go in vain as they will return to you when you least expect them. **Candidate Moral 4:** Don't trust strangers. **Candidate Moral 5:** Everyone person is significant and deserve your attention and respect.
True Answer: Moral 1 **Model Prediction:** Moral 5
Understanding Task: MOPREF
Input Story: Once upon a time there lived a cat that loved to read. At night, when everybody was asleep, she would put on the spectacles and read the big book for cats. One day, she read in the book: if you want a mouse for dinner, repeat the following rhyme: in this house there is a mouse, where is the mouse, where is the mouse? The cat looked up from the book and found that there was a mouse on the top of the table. The cat repeated the rhyme and soon found the same mouse on the bed. Then she jumped upon the bed to catch the mouse and the mouse was gone! The mouse was very clever. Suddenly he squeaked, "Oh, dear cat, run, run fast! there is dog after you!" The cat left the mouse and was ready to jump out of the window. The mouse sat near his hole and said, "Ha-ha-ha! dear cat that was the trick I learnt from the bio book for mice!" And the mouse ran into his hole!
Candidate Moral 1: An intelligent person should not think that others are illiterate. **Candidate Moral 2:** An intelligent person should not forget that others are illiterate.
True Answer: Moral 1 **Model Prediction:** Moral 2
Generation Task: ST2MO
Input Story: In the forest, there was a deer and an owl. The deer loved to scare this fearful owl all the time. One day, the deer scared the owl so hard, the owl past out. The deer ran away, and the owl woke up and came to her senses. She had had enough of being scared. She went to the deer and said, "I'm not scared of you." She said this with such a force that the deer backed down · · ·
Generated Case 1: The deer loves to scare owls. (NAM)
Generated Case 2: We should not be greedy and learn to others the importance of our desires. (UNREL)
Generated Case 3: It's good to be scared. (CONF)
True Moral: Don't let anyone scare you all the time.
Generation Task: MO2ST
Input Moral: Empty solutions are of no worth.
Generated Case 4: There was a grocery shop in a town. There was **a big fat cat** in there. It was very big. It was a giant cat. It was big, and it was very big and powerful. It was big and powerful. It was very powerful. The mouse could **move freely** and had a nice time hunting. The cat was very big and powerful. The mouse wanted to live in a very large land · · · (NAM)
Generated Case 5: There was a grocery shop in a town. Just the mouse and cat lived together. · · · The mouse was very active in the shop and in the street. The mouse was very happy. But the mouse was too busy to **move freely**. He was the only mouse in town. He was very busy with work. · · · As soon as the mouse had left, he moved slowly. · · · So the mouse was so happy and satisfied with everything that it was so happy for him. · · · (UNREL)
Generated Case 6: There was a grocery shop in a town. There were two stores. · · · The third was an empty solution. Each time the mouse went on a few days. Then, the third was a little while later the next day the mice were gone. They were gone. They were happy at last · · · (CONF)
True Story: There was a grocery shop in a town. Plenty of **mice lived** in that grocery shop. Food was in plenty for them. They ate everything and spoiled all the bags. They also wasted the bread, biscuits and fruits of the shop. The grocer got really worried. So, he thought "I should buy a cat and let it stay at the grocery. only then I can save my things." He bought a nice, **big fat cat** and let him stay there. The cat had a **nice time hunting** the mice and killing them. The mice could not **move freely** now. They were afraid that anytime the cat would eat them up. The **mice wanted** to do something. They held a meeting and all of them tweeted "We must get rid of the cat. can someone give a suggestion"? All the mice sat and brooded. A smart looking **mouse stood** up and said, "The **cat moves softly.** that is the problem. if we can tie a bell around her neck, then things will be fine. we can know the movements of the cat". "Yes, that is answer," Stated all the mice. An old **mouse slowly stood** up and asked, "Who would tie the bell?" After some moments there was no one there to answer this question.

Table 20: Typical error cases predicted by RA-T5 (for the understanding tasks) or sampled from RA-T5 (for the generation tasks). For the generation tasks, the error types in terms of moral faithfulness include "not a moral text" (NAM), "unrelated concepts" (UNREL) and "conflicting value preference" (CONF). The underlined words are improper concepts/events which leads to corresponding errors. **Bold** words for MO2ST are the given first sentence and the outline of multiple phrases.

ing what is right or what is wrong; **(2) Unrelated concepts (UNREL)**: containing unrelated concepts with the input; and **(3) Conflicting value preference (CONF)**: conveying a value preference conflicting with the input despite related concepts. In addition, we also provide annotators with another option **Others**. The annotators are allowed to annotate a text with multiple errors. When at least two of three annotators annotate the text with some error, we decide it has the error. We show the distribution of the error types in Table 19, suggesting that existing models still struggle to generate meaningful morals and stories, and align the concepts and value preferences between them.

Furthermore, as exemplified in Table 20, when generating morals, we can see from Case 1 that the

Generation Task: ST2MO

Input Story: Once upon a time there was a spring who lived happily and safely inside a pen. Although he heard many noises coming from outside, he lived believing that outside his world inside the pen, there was nothing good. Even just to think about leaving his pen made him so scared that he was quite content to spend his life compacting and stretching himself again and again inside that tiny space. However, one day, the ink ran out, and when the pen's owner was busy changing it, there was an accident. The spring was flung through the air and landed in the toilet drain, well out of sight. Terrified, and cursing his bad luck, the spring was flushed through pipe after pipe, each time thinking it might be his end. During the journey, he did not dare open his eyes out of pure fear. Nor did he every stop crying. Swept away by the water, he travelled on and on and on, until he ended up in a river. When the river current lost its force, and the spring could see that things had calmed down a bit, he stopped crying and listened all around him. Hearing birdsong and wind in the trees, he felt encouraged to finally open his eyes. What the spring saw was the pure, crystal waters of the river, the rich green rocks of the riverbed, and all kinds of fish of many colours, whose skin seemed to dance under the sunlight. Now he understood that the world was much greater than the space inside the pen, and that there had always been many things outside, waiting to be enjoyed. After spending a while playing with the fish, he went over to the riverbank, and then moved on to a field of flowers. There he heard weeping. He followed the sound, which took him to a lovely flower that had been flattened by a rabbit, and could no longer stand up straight. The spring realised that he could help the flower, so he offered to be his support. The flower accepted, and slipped through the middle of the spring. There they lived happily together. And they would always laugh when remembering how the spring used to think that all there was to life was being a sad and fearful spring.

GPT2: Always think of what you can do to make things better.

T5: Fear and frustration are one thing, but they can be turned into happiness.

RA-T5: Life should not be afraid of the unknown. There should be lots of positive things out there.

Top-5 Retrieved Morals by RA-T5:
(1) Everything in life has its unpleasant or uncomfortable side. These aspects cannot be removed from life.
(2) It's always better to have friends and be happy - even though it may bring discomfort - than isolate yourself from people in order to avoid problems.
(3) Most fear is without foundation, and plays on what is unknown or different.
(4) Friends are for cheering you up in your worst moments.
(5) Someone who harm's no one is still very far from being good.

True Moral: Changes are not a bad thing, it is up to ourselves to open the eyes and make something positive of them.

Generation Task: ST2MO

Input Moral: Never give up.

GPT2: One day a dragon and a dino met in the late jurassic period. They were friends, and on the way they met a **huge tree**. Dragon was huge and his **big claws** were full and he was going to kill him. Dino was going to die. Dino was going to die. When the **dragon made half** of the tree, he was about to die. Dragon was going to die, and all the beasts and beasts were dead. Dino had no idea what to do. He was the **strongest dino** in the whole tree. Dragon was a **total surprise**. Dino was going to die. He was about to die. He was going to die. The dragon was astonished. Dino was not able to die. Dino was so angry, he started to die. Dino was so angry, and the dragon threw his **big claws** at the tree and ran to the forest. This was the best match. Then the dino was nearly killed.

T5: One day a dragon and a dino met in the late jurassic period. The dragon was standing next to a **huge tree**. It was full of beasts, lions and zebras that were strong enough to carry out their duties. The dragon was so strong that he could move himself without losing his foot. The dino was so strong that he could carry anything, even if that tree had much more branches. But the dino was very weak, so he could not carry anything. The dragon came to the tree and said, "I will do everything I can to get my foot into the tree." Then the **dino started running** through the tree. The **dragon made half** way through the tree and put on his **big claws**. Then he took off his **big claws** in his claws. Just then a huge **tree started** to go through a storm. Then the monster took out his **fiery breath**. The dino could not contain his **fiery breath** and flew away.

RA-T5: One day a dragon and a dino met in the late jurassic period. The dragon was very proud of the dino because of his **big claws** and **fiery breath**. The dino was going to jump on a **huge tree** and climb it up. But the dragon didn't know how to jump. The **dragon made half** way up the tree and started to roll down. The dino was completely surprised. The dragon said, "I can't jump," The dragon started to jump. Then the **tree started** to roll. The dino started to roll over the top of the tree and jumped up. The **tree started** to thrash off. Suddenly, the dino saw that the dragon had a **fiery breath**. The dragon said, "That's my strength." The dino was all alone. Then he stood, looking at the dragon. Then the dragon stopped, shouted at him, "Hello dino! what a **total surprise**." The **dino started running** and jumping towards the tree, and the dragon got close to him. Then he hit the tree. Now he found a boulder on the tree started to crash and he jumped on the tree. Then the **strongest dino** ever was surprised.

Top-5 Retrieved Story Outines:
(1) {baldwin flew, baldwin scratched rattler, team beat baldwin, baldwin dodged rattler, football game, goal post, baldwin threw rattler, baldwin started}
(2) {eagle resting, tree top, tortoise rested, eagle answered, deep sleep, tortoise sleeping, tortoise smiled, hunter suddenly]}
(3) {loud thump, man happened, cry intruded, ugly wolf, wolf named pete walked, long neck, man walked, thin air}
(4) {cat suddenly fell, bird flew, started climbing, cat thanked}
(5) {lion won, race started, croc won}

True Story: One day a dragon and a dino met in the late jurassic period. The dragon said, "I'm stronger than you!" The dino said, "I'm the **strongest dino** ever!" The next day the dino and the dragon met in the forest. The **dragon made half** of a tree fall down with its **big claws** and **fiery breath**. The dragon said, "You can't beat that!" The **dino started running** toward a **huge tree**. The dino rammed the huge tree with its head. Nothing happened. The dragon laughed. Then the **tree started** to fall. The dragon just stared in **total surprise**.

Table 21: Cases generated by different models for the generation tasks. The underlined words are improper concepts/events which leads to incoherence or unfaithfulness. **Bold** words for MO2ST are the given first sentence and the outline of multiple phrases. The red moral for ST2MO is related to the generated moral of RA-T5 in semantics. Note that we only take concepts in these retrieved morals as input for RA-T5. And red words in the retrieved outlines for MO2ST indicate that they also show up in the generated story of RA-T5.

models still often state events involved with specific characters (e.g., *"owls"*) but do not tell what is right and what is wrong. And Case 2 shows that they struggle to conclude related concepts from the story (e.g., *"greedy"* is not embodied in the story at all). Furthermore, in Case 3, the models conclude a conflicting value preference with the story despite correct concepts (e.g., the story shows that *"it is bad to be scared"* but not *"good"*). On the other hand, models also are shown to suffer from similar issues when generating stories. In Case 4, the model only describes some scenes (e.g., *"it was*

very big" and *"it was very powerful"*) but does not aims to convince readers of anything. And Case 5 seems to tell a story centered on some concepts such as *"active"* and *"busy"*, but the concepts do not relate to the input. Case 6 implies *"empty solutions may be useful,"* which is conflicting with the input. These cases indicate the necessity of modeling the relations between events and abstract concepts for understanding and generating moral stories.