ElitePLM: An Empirical Study on General Language Ability Evaluation of Pretrained Language Models

Anonymous ACL submission

Abstract

Pretrained language models (PLMs), such as BERT and GPT-3, have dominated the majority of NLP tasks. However, relatively little work has been conducted on systematically evaluating the language abilities of PLMs. In this paper, we present a large-scale empirical study on genEral language ability evaluation of PLMs (ElitePLM). We first design four evaluation dimensions in ElitePLM, including memory, comprehension, reasoning, and composition, and further measure ten widely-used PLMs within five categories. Our empirical results demonstrate that: (1) the pretraining objectives and strategies have significant impacts on PLMs performance in downstream tasks; (2) fine-tuning PLMs in downstream tasks is usually sensitive to the data size and distribution; (3) PLMs have excellent transferability between similar tasks. Our experimental results summarize several important findings, which can guide the future work to choose, apply, and design PLMs for specific tasks. We have made all the details of experiments publicly available at `https://anonymous.4open.science/r/Paper-for-ACL-4FD1`.

1 Introduction

Recent years have featured a trend towards Transformer (Vaswani et al., 2017) based pretrained language models (PLMs) in natural language processing (NLP) systems. By first pretrained on massive unlabeled text, PLMs can be directly fine-tuned on downstream tasks, entirely removing the needs to task-specific architectures (Radford et al., 2018). This paradigm has led to significant progress on many challenging NLP tasks such as BERT (Devlin et al., 2019) on reading comprehension and GPT-3 (Brown et al., 2020) on text generation.

Giving new state-of-the-art results that approach or surpass human performance on several tasks, it is an interesting question about how to systematically evaluate the language abilities of PLMs from a wide range of perspectives. Given the increasing number of publicly released PLMs, it is particularly useful to derive principles or guidelines of selecting suitable PLMs for specific downstream tasks. However, existing works either target at some single ability (Talmor et al., 2020; Zhou et al., 2020), or consider a simple mixture of multiple (small-scale) tasks that lack a comprehensive design and test (Wang et al., 2019b; Liang Xu, 2020). There has been no detailed and systematic analysis characterizing the abilities of PLMs in large-scale NLP tasks. To fill the gap of PLMs evaluation, we introduce the gen**E**ral **l**anguage ab**i**li**t**y **e**valuation (**ElitePLM**) for empirically and systematically assessing the general language abilities of PLMs.

The motivation behind PLMs is to create a machine learner equivalent to human being which can understand the language and then be asked to perform any specific task related to language. In cognitive science, the Wechsler Adult Intelligence Scale (WAIS) (Kaufman and Lichtenberger, 2005) is the most commonly used intelligence quotient (IQ) test for measuring the intelligence and cognitive ability of human being. This test would assess the level of individuals on verbal comprehension, perceptual reasoning, working memory, and processing speed. Thus, by imitating the intelligence test on human, we design four evaluation dimensions in ElitePLM for measuring the abilities of PLMs, including memory, comprehension, reasoning, and composition. Following previous works (Zhou et al., 2020; Wang et al., 2019b), for each ability in ElitePLM, we elaborate and choose multiple representative tasks (*e.g.,* question answering for the comprehension ability) and commonly-used benchmarks (*e.g.,* GLUE and SQuAD) to quantitatively evaluate the performance of PLMs. These results can serve as numerical explanations of PLMs at a certain ability.

In human intelligence tests, the background of participants (*e.g.,* gender, race, and occupation) should be as much as diverse. Thus, in ElitePLM, we also select a diversity of PLMs to conduct generalized and meaningful comparisons. According to training objectives, pretrained language models can be divided into three categories: unidirectional language models (*e.g.,* GPT (Radford et al., 2019)) for natural language generation (NLG), bidirectional language models (*e.g.,* BERT (Devlin et al., 2019)) for natural language understanding (NLU), and hybrid language models (*e.g.,* UniLM (Dong et al., 2019)) for combining the first two paradigms. Besides, knowledge-enhanced language models (*e.g.,* ERNIE (Zhang et al., 2019)) and text-to-text language models (*e.g.,* T5 (Raffel et al., 2020)) also emerge as important branches of PLMs. Considering the variety, we finally choose ten widely-used PLMs within the above five categories and evaluate their abilities on the four dimensions. The comparisons of these PLMs in configuration and pretraining setting have been shown in Appendix A.

From the experimental results we have three salient findings. First, the pretraining objectives and strategies have significant impacts on PLMs performance in downstream tasks. We observe that the bidirectional training objective like BERT and pretraining strategies like larger training batches in RoBERTa are helpful for memorizing large-scale pretraining corpus; pretraining objectives like permutation language modeling in XLNet are highly useful for modeling the bidirectional context in text; left-to-right prediction in GPT-2 for generating long text. Second, when fine-tuning PLMs in downstream tasks, their performances are usually sensitive to the data size and distribution, which can be addressed by designing task-specific objectives like inter-sentence coherence loss in ALBERT for sentence-level reasoning tasks. Third, PLMs have excellent transferability between similar tasks. This finding can be utilized to fine-tune PLMs in the zero-shot and few-shot tasks. For example, we can first fine-tune PLMs on a data-rich source task with massive data, and then transfer the fine-tuned PLMs to a similar data-scarce target task. We illustrate the effect extent of each factor for PLMs abilities in Appendix A.

We hope that this paper will help establish good principles on choosing, applying, interpreting and designing PLMs for NLP tasks in practical settings. We will also release the code for all experiments and tested results, providing the community with off-the-shelf tools to evaluate their PLMs.

2 ElitePLM

In ElitePLM, we empirically study four kinds of language abilities of PLMs, namely memory, comprehension, reasoning, and composition. Next, we will describe each ability in detail.

Memory Ability. For humanity, memory is the most fundamental ability, which is involved in how much information has been remembered in our life experience (Miyake and Shah, 1999). By analogy, it is similar to measure how much text PLMs have remembered in pretraining, as assessed by tests of recall of words conditioned on some contexts.

On the other hand, efficiency is also an important aspect of memory ability for PLMs learning from new data distribution in the fine-tuning stage. Thus, besides recalling words, we also compare the memory efficiency of PLMs with different model architectures and training objectives in terms of memorizing the given new information in fine-tuning. Based on the memorized information, PLMs can generalize such knowledge and language patterns into downstream tasks for understanding the similar context in text.

Comprehension Ability. Comprehension ability is complex and multifaceted. It is usually comprised of understanding a text's vocabulary, background knowledge of a particular topic, and comprehension of its language structures like grammar (Cain and Oakhill, 2008). In particular, background knowledge is used to comprehend a special situation, lesson, or text (also called prior knowledge). For instance, when reading a text about dog training, readers are going to use their background knowledge of dog behavior, vocabulary related to dogs, aspects of training a dog, to comprehend the given text.

Our ElitePLM contains several well-focused tasks to evaluate the comprehension ability of PLMs from three views, *i.e.,* vocabulary, background knowledge, and language structures. First, the word sense disambiguation task requires PLMs to understand the meaning of vocabulary words and determine whether the words are used with the same sense in sentences (Wang et al., 2019a). Furthermore, the reading comprehension task may need some particular background knowledge about the passages to answer questions under a special topic (Lai et al., 2017). Besides, the language structure is concerned with the relationships between words such as knowledge of grammar, which can be quantified by some syntactic tasks like coreference resolution (Wang et al., 2019b).

Reasoning Ability. Based on the comprehension of a text, reasoning ability refers to the power and effectiveness of the processes and strategies used in drawing inferences, reaching conclusions, arriving at solutions, and making decisions (Kyllonen and Christal, 1990). There are several distinct forms of reasoning, implicating different reasoning abilities. In ElitePLM, we mainly focus on three kinds of reasoning ability, *i.e.,* commonsense reasoning, deductive reasoning, and abductive reasoning.

Specifically, commonsense reasoning requires PLMs to make mundane inferences using commonsense knowledge about the world, like the fact that "matches" plus "logs" usually equals "fire" (Sap et al., 2020). Note that, subtle differences exist between commonsense knowledge and background knowledge in comprehension ability. Commonsense knowledge is broadly defined as the total accumulation of facts and information that a person has gained from previous experiences. Besides, deductive reasoning involves PLMs drawing conclusions from a set of given premises in the form of categorical syllogisms (*e.g.,* all x are y) or symbolic logic (*e.g.,* if p then q) (Johnson-Laird, 1999), and abductive reasoning involves arriving at the most likely explanation for a set of facts, such as a scientific theory to explain a set of empirical findings (Walton, 2014).

Composition Ability. Unlike previous abilities to memorize, comprehend, and reason on the given content, the composition ability is a highly intelligent and synthetic ability that requires PLMs to create new content from scratch. In the literary sense, composition is the way that a writer assembles words and sentences to create a coherent and meaningful work (*e.g.,* poem, music, and narration), which is closely resemble to the text generation task in NLP research (Berninger, 1999).

Therefore, in ElitePLM, we introduce several text generation tasks for evaluating the composition ability of PLMs including story generation, text summarization, and question generation. Note that, story generation is a representative composition task which needs PLMs to not only comprehend the given story background, but also reason about and create reasonable and coherent story endings (Fan et al., 2018). During the composition process, PLMs should include a good vocabulary, grammar, spelling, and punctuation knowledge, and need to deliberate the structure of text.

3 Experiments

In this section, we first set up baselines, and then report the results and analysis on four ability tests.

3.1 Models

As mentioned before, we compare the performance of ten publicly released PLMs from five categories:

- *Bidirectional Language Model*: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and ALBERT (Lan et al., 2020);
- *Unidirectional Language Model*: GPT-2 (Radford et al., 2019);
- *Hybrid Language Model*: XLNet (Yang et al., 2019) and UniLM (Dong et al., 2019);
- *Knowledge-enhanced Language Model*: ERNIE (Zhang et al., 2019);
- *Text-to-Text Language Model*: BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and ProphetNet (Qi et al., 2020).

We implement all the models and tests mostly on huggingface (Wolf et al., 2020), fairseq (Ott et al., 2019), and jiant (Phang et al., 2020). For fair comparison, all PLMs are conducted with the same training setting such as batch size and learning rate.

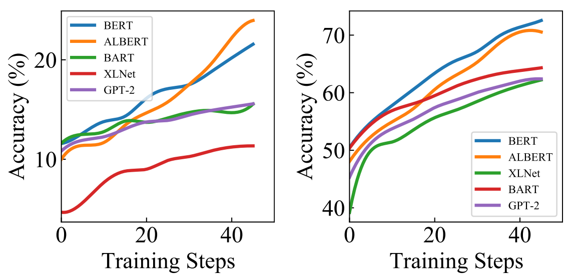
3.2 Memory Tests

Datasets. The goal of memory tests is to answer two questions: (1) how much information PLMs have remembered in pretraining, and (2) how efficiently PLMs remember new information. For this purpose, we adopt two datasets for evaluation, *i.e.,* LAMA (F. Petroni and Riedel, 2019) and English Wikipedia (2,500M words).

Specifically, LAMA is a knowledge probe corpus containing a set of knowledge facts, where facts are either subject-relation-object triples or question-answer pairs. Each fact is converted into a cloze statement where the subject or object entity is masked. Wikipedia is one of the widely-used pretraining corpus for our selected PLMs (except GPT-2 and T5). Thus, to conduct fair comparison, we also pretrain GPT-2 and T5 on Wikipedia according to their pretraining objectives. Similar to LAMA, we randomly sample 100,000 text from Wikipedia and then mask a proportion of 15% tokens following BERT. By querying PLMs with the missing tokens on Wikipedia and LAMA, we can test the language pattern and factual knowledge in

Models	Bidirectional			Uni.	Hybrid		KE	Text-to-Text		
	BERT	RoBERTa	ALBERT	GPT-2	XLNet	UniLM	ERNIE	T5	BART	ProphetNet
Vocab Size	28996	50265	30000	50257	32000	28996	28996	32100	50295	30522
LAMA										
Google-RE	**11.0**	7.1	3.3	3.9	10.0	9.6	1.3	4.0	9.4	0.1
T-REx	**29.2**	23.9	21.0	12.0	28.9	28.4	13.4	21.7	15.8	1.1
ConceptNet	19.1	**21.6**	20.0	6.4	19.5	18.3	13.0	17.1	7.7	0.3
SQuAD	17.0	**21.0**	20.6	5.6	20.8	17.4	8.1	11.7	3.1	0.7
Wikipedia	70.9	71.1	63.9	42.7	68.7	**71.5**	45.7	65.0	47.8	31.3
Average	**45.0**	44.8	40.1	24.8	44.3	**45.0**	3.9	39.3	28.4	15.9

Table 1: Memory test results on LAMA and Wikipedia datasets (test set). We report the accuracy score for the large version of each model in this table and more results can be found in the Appendix C. Bold and underlined fonts denote the best and the second best performance of a PLM (the same as below).



(a) Google-RE (b) T-REx

Figure 1: Memory efficiency ($P@1$) of five PLMs on Google-RE and T-REx datasets.

PLMs' memory. Since the missing tokens might appear in the middle of a sentence, for auto-regressive PLM such as GPT-2, we only evaluate PLMs on those at the end. For efficiency, we measure it as the performance *w.r.t.* the number of training epochs: the more efficient a model is, the fewer epochs to achieve a reference performance.

Results and Analysis. We first directly test PLMs using Wikipedia and LAMA without fine-tuning, which is similar to the zero-shot learning. The results on mean precision at one ($P@1$) metric are summarized in Table 1. Compared with bidirectional and hybrid language models (*e.g.,* BERT and XLNet), GPT-2 uses constrained self-attention where every token can only attend to context to its left. This unidirectional training objective naturally limits the performance of GPT-2 in terms of memory ability. It has been previously reported that PLMs can remember more information by scaling up the model size (Brown et al., 2020). However, in our tests, BART-large (400M) achieves worse results than RoBERTa-base (125M) with the same training corpus and similar vocabulary sizes (50295 vs 50265). During pretraining, RoBERTa incorporates a series of training strategies, using more pretraining data, larger batches, longer sequence, and dynamic masking, etc. Compared with model size, **training objectives and strategies reflect the way of PLMs memorizing information, which seems to have more significant impacts on the memory ability of PLMs**. Besides, we can clearly observe that all PLMs achieve their best results in T-REx, a subset of Wikipedia triples, and show relatively good performance on Wikipedia. This indicates that the training corpus determine the knowledge scale of PLMs' memory, which influences the performance of PLMs in downstream tasks, especially for zero-shot learning. This is the reason why previous studies choose to train PLMs on a very large corpus.

To test the memory efficiency, we fine-tune five models, BERT, ALBERT, GPT-2, BART, and XLNet, for several epochs with the same training settings (*e.g.,* learning rate). As shown in Figure 1, to achieve a reference performance, the bidirectional training objective like BERT needs fewer epochs than other kinds of objectives. This further implies that besides memory capacity, **the bidirectional training objective is also useful to facilitate the memory efficiency of PLMs, because bidirectional language modeling can effectively capture the bidirectional context**.

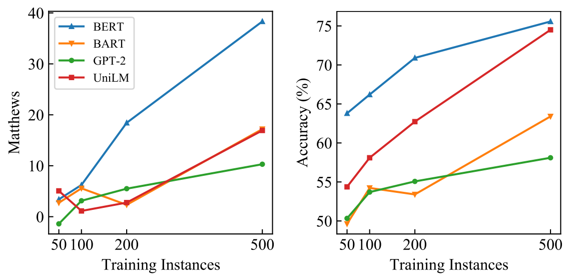
3.3 Comprehension Tests

Datasets. As discussed in Section 2, comprehension ability mainly refers to the understanding of a text's vocabulary, background knowledge, and language structure. Considering these aspects, we

[1] https://gluebenchmark.com/

Models	WNLI Acc.	CoLA Matt.	MNLI M./MM.	RTE Acc.	QNLI Acc.	SST-2 Acc.	QQP F1/Acc.	STS-B P/S Corr.	MRPC F1/Acc.	Avg.
$\text{BERT}_{\text{BASE}}$	65.1	52.1	84.6/83.4	66.4	90.5	93.5	69.9/88.2	77.4/73.7	79.0/85.1	76.5
$\text{BERT}_{\text{LARGE}}$	65.1	60.5	86.7/85.9	70.1	92.7	94.9	72.1/89.3	87.6/86.5	85.4/89.3	80.5
$\text{RoBERTa}_{\text{BASE}}$	65.1	61.4	87.4/87.2	75.1	92.9	95.7	72.5/89.4	89.2/88.5	87.5/90.7	81.8
$\text{RoBERTa}_{\text{LARGE}}$	89.0	67.8	90.8/90.2	88.2	98.9	96.7	74.3/90.2	92.2/91.9	89.9/92.4	88.5
$\text{ALBERT}_{\text{XLARGE}}$	65.8	58.2	35.6/36.5	62.5	94.2	95.1	71.7/88.9	87.6/86.6	69.8/80.3	72.7
$\text{ALBERT}_{\text{XXLARGE}}$	64.4	64.7	89.7/89.6	70.4	95.3	96.0	70.7/88.4	91.3/90.6	68.1/80.4	80.6
$\text{GPT-2}_{\text{SMALL}}$	54.8	33.8	81.1/81.4	62.1	86.7	91.2	69.8/87.9	79.0/76.5	76.9/83.6	71.9
$\text{GPT-2}_{\text{MEDIUM}}$	54.1	50.5	84.8/84.5	63.6	91.2	92.1	71.4/88.6	84.3/82.7	80.0/85.5	75.8
$\text{XLNet}_{\text{BASE}}$	58.9	26.2	86.1/85.3	59.9	91.3	94.0	71.5/88.9	83.9/82.9	84.3/88.3	74.0
$\text{XLNet}_{\text{LARGE}}$	**92.5**	**70.2**	**90.9/90.9**	**88.5**	**99.0**	**97.1**	**74.7/90.4**	**93.0/92.6**	**90.5/92.9**	**89.5**
$\text{UniLM}_{\text{BASE}}$	65.1	49.0	83.0/82.2	60.3	88.7	92.3	70.7/88.4	82.3/81.4	84.3/88.7	76.2
$\text{UniLM}_{\text{LARGE}}$	65.1	61.1	87.0/85.9	70.9	92.7	94.5	71.5/89.2	86.6/85.3	85.2/89.1	80.5
$\text{ERNIE}_{\text{BASE}}$	65.1	52.3	84.0/83.2	68.8	91.3	93.5	70.5/88.4	85.1/83.8	80.3/85.9	70.7
T5_{BASE}	78.8	51.1	87.1/86.2	80.1	93.7	95.2	72.6/89.4	89.4/88.6	87.5/90.7	82.7
T5_{LARGE}	85.6	61.2	89.9/89.6	87.2	94.8	96.3	73.9/89.9	89.9/89.2	89.8/92.4	86.4
$\text{BART}_{\text{BASE}}$	65.1	52.8	85.1/84.3	69.5	92.6	94.4	72.5/89.7	87.6/86.6	86.1/89.5	79.5
$\text{BART}_{\text{LARGE}}$	58.9	62.4	90.2/89.3	83.5	94.8	96.3	73.6/90.1	91.1/90.4	87.8/91.1	83.1
$\text{ProphetNet}_{\text{LARGE}}$	52.1	24.2	81.3/80.8	51.3	93.2	93.6	70.6/88.1	73.5/72.3	69.7/80.8	69.2

Table 2: Comprehension tests results on GLUE (test set). All results are scored by the GLUE evaluation server[1].



(a) CoLA (b) QNLI

Figure 2: Few-shot results of four PLMs on CoLA and QNLI tasks.

employ five datasets for comprehension tests, *i.e.,* GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a), SQuAD v1.1 (Rajpurkar et al., 2016), SQuAD v2.0 (Rajpurkar et al., 2018), and RACE (Lai et al., 2017).

Among these datasets, GLUE and SuperGLUE are two widely-used reading comprehension benchmarks. Several tasks, such as semantic text similarity, and coreference resolution, can be adopted to test the understanding of PLMs about semantic meaning and syntactic structure of text. By contrast, SQuAD v1.1&v2.0, and RACE are three popular question answering datasets. To answer the natural language questions, PLMs should be aware of the background knowledge about some particular topic. For example, to answer the question "*what can be used as rewards for dog training?*", the background knowledge "*dogs like bones*" will be helpful for PLMs to answer "*bones*".

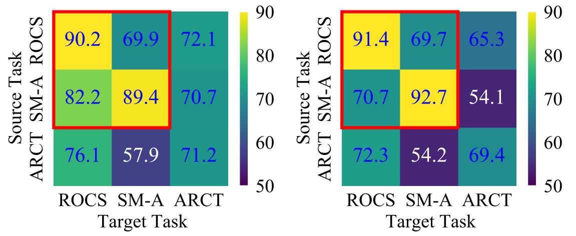
Results and Analysis. Table 2 presents the results of comprehension test in GLUE dataset (results in other four datasets can be found in Appendix D). The last column in this table indicates the average overall performance across all tasks.

Interestingly, the models behaving well in memory tests (*e.g.,* RoBERTa and XLNet) also present good results in many comprehension tasks. These results indicate that **the improvement on memory ability is likely to be helpful for the performance of comprehension ability**, which is in line with our intuition. Compared with the bidirectional language modeling like BERT (relying on corrupted input with masks), the permutation language modeling used in XLNet enables PLMs to learn more kinds of context for enhancing PLMs' understanding of the text, which seems to be effective for good comprehension ability.

Among these tasks, we observe a significant performance drop in the linguistic acceptability task (CoLA), which is because the PLMs saw different data distributions during pretraining (Wang et al., 2021). This kind of sensitiveness to unfamiliar tasks is also reflected in Figure 2, where the model performance on CoLA show a more volatile fluctuation (ranging from 10 to 35) than QNLI (ranging from 15 to 20). It indicates that **the performance of PLMs is closely related to the similarity of data distributions in pretraining and finetuning**. To solve this challenge, it will be better to adopt intermediate fine-tuning, which involves first fine-tuning PLMs on an intermediate similar dataset and then transfering to the final dataset.

Datasets	Bidirectional			Uni.	Hybrid		KE	Text-to-Text		
	BERT	RoBERTa	ALBERT	GPT-2	XLNet	UniLM	ERNIE	T5	BART	ProphetNet
CQA	55.9	72.2	**80.0**	60.8	62.9	62.3	54.1	69.8	75.8	21.3
ROCStories	90.2	**97.4**	97.1	59.9	93.8	86.9	84.7	91.4	91.7	82.2
SWAG	86.3	89.9	**90.7**	79.7	86.8	83.1	80.2	73.7	87.9	70.1
HellaSwag	47.3	85.2	**90.1**	60.4	79.7	46.7	44.5	79.1	76.6	26.4
SM-A	89.4	**93.0**	92.5	88.7	83.7	89.3	88.7	92.7	82.9	85.5
SM-B	85.8	**92.3**	**92.3**	73.4	88.7	86.4	87.7	88.2	67.9	78.0
ARCT	71.2	57.9	79.5	66.7	83.1	72.3	73.7	69.4	**84.2**	65.5

Table 3: Reasoning tests results on seven datasets (test set). We report accuracy score for each dataset. CQA is short for CommonsenseQA. SM-A and SM-B denote the Task A and Task B of Sense Making, respectively. We report the results of large version for each model in this table and more results can be found in the Appendix E.



(a) $\text{BERT}_{\text{LARGE}}$ (b) T5_{LARGE}

Figure 3: Heatmap of two-stage transfer learning for BERT and T5.

3.4 Reasoning Tests

Datasets. In reasoning tests, we mainly take into account three forms of reasoning ability, *i.e.,* commonsense reasoning, deductive reasoning, and abductive reasoning, which focus on commonsense utilization, conclusion induction, and reason derivation, respectively. For evaluation, we select six reasoning datasets, namely CommonsenseQA (Talmor et al., 2019), ROCStories (Mostafazadeh et al., 2016), SWAG (Zellers et al., 2018), HellaSwag (Zellers et al., 2019), Sense Making (Wang et al., 2019c), and ARCT (Habernal et al., 2018).

Different from the background knowledge, commonsense knowledge in CommonsenseQA spans a large portion of human experience of everyday life (Liu and Singh, 2004). ROCStories, SWAG, HellaSwag, and Sense Making Task A are concerned with deriving the conclusions of stories and events, while Sense Making Task B and ARCT focus on identifying the reason behind a statement.

Results and Analysis. Table 3 shows the model performances in reasoning ability. We can clearly observe that, besides performing well in comprehension tasks, ALBERT and RoBERTa demonstrate stronger performance in almost all reasoning tasks. In pretraining, ALBERT introduces an inter-sentence coherence objective to capture the correlation among sentences, which can be more helpful for the sentence-level reasoning ability of PLMs. It has been found that the next sentence prediction (NSP) loss in BERT might hurt the performance of PLMs in sentence-level tasks of downstream datasets, thus RoBERTa removes this objective in pretraining (Liu et al., 2019b).

Interestingly, though performing the best in comprehension tests, XLNet does not perform as well as we expected in reasoning tests. We speculate that the permutation operation in XLNet disturbs the semantic correlation between sentences and thus leads to poor reasoning ability. **To improve the reasoning ability, it would be useful to design sentence-level reasoning objectives like inter-sentence coherence loss in ALBERT and then pretrain PLMs with these objectives**. Moreover, despite incorporating knowledge into language models, ERNIE still shows mediocre performance in knowledge-oriented datasets such as CommonsenseQA. A possible reason might be ERNIE only utilizes the trained KB embeddings to enhance the semantic representations, while the the reasoning structure on KBs are ignored.

To test the transfer learning between different reasoning abilities, we conduct a two-stage experiment across three kinds of tasks, ROCStories, SM-A, and ARCT, shown in Figure 3. We first train PLMs on source task with full data, and then fine-tune PLMs with ten instances on target task. It can be observed that **PLMs have better reasoning transferability between similar tasks** such as deductive reasoning tasks (ROCStories and Sense Making Task A). This shows that the model performance on data-scarce tasks can be improved by incorporating additional training on data-rich similar tasks (Wang et al., 2021).

Models	CNN/DailyMail			GigaWord			SQuAD			WritingPrompts		
	R-1	R-2	R-L	R-1	R-2	R-L	B-4	R-L	ME	B-4	R-L	ME
GPT-2	27.00	8.00	23.08	23.72	8.12	21.56	8.48	18.82	26.77	14.47	3.23	7.29
UniLm	43.44	20.21	40.51	38.45	19.45	35.75	4.42	17.43	20.13	**26.88**	1.84	5.01
T5	42.50	20.68	39.75	34.75	16.26	31.49	11.19	22.35	30.53	8.61	**4.19**	**9.51**
BART	44.16	**21.28**	40.90	39.41	20.21	36.42	**15.87**	**25.47**	**38.42**	14.72	3.14	7.08
ProphetNet	**44.20**	21.17	**41.30**	**39.51**	**20.42**	**36.69**	14.20	23.97	35.99	19.31	2.59	7.19

Table 4: Composition tests results on four datasets. R-1, R-2, R-L are short for ROUGE-1, ROUGE-2, ROUGE-L respectively. B-4 and MT denote BLEU-4 and METEOR, respectively. We report the result of large version for each model in this table and more results can be found in the Appendix F.

Models	GigaWord				
	TT (%)	Flu.	Info.	Acc.	Overall
GPT-2	26.09	3.11	2.79	2.64	4.87
UniLM	50.34	**4.02**	3.49	3.45	6.73
T5	**53.67**	3.95	3.45	3.46	6.68
BART	51.10	4.01	3.46	**3.49**	6.73
ProphetNet	53.02	3.99	**3.52**	3.45	**6.74**
Gold	40.77	3.61	3.29	3.15	6.05
Models	**WritingPrompts**				
	TT (%)	**Flu.**	**Info.**	**Rel.**	**Overall**
GPT-2	**45.70**	**3.42**	**3.17**	3.20	5.87
UniLM	1.20	1.32	1.88	2.03	2.74
T5	34.40	3.01	2.80	3.09	5.18
BART	45.20	3.37	3.16	**3.39**	**5.96**
ProphetNet	29.60	2.95	2.91	3.10	5.18
Gold	71.30	3.79	4.07	3.87	7.37

Table 5: Turing test (TT) and human scores on the test set of GigaWord and WritingPrompts. Flu., Info., Acc. and Rel. denote fluency, informativeness, accuracy and relevance respectively. We report the result of large version for each model in this table and more results can be found in the Appendix F.

3.5 Composition Tests

Datasets. Composition is closely related to the text generation task, which is also aimed at generating new content from scratch. Therefore, we utilize four text generation benchmarks for composition tests, *i.e.,* WritingPrompts (Fan et al., 2018) on story generation, CNN/Daily Mail (Hermann et al., 2015) and GigaWord (Rush et al., 2015) on text summarization, and SQuAD v1.1 (Rajpurkar et al., 2016) on question generation. Specifically, according to the length of generated text, text summarization and question generation belong to short text generation, while story generation belongs to long text generation.

For performance comparison, we adopt three automatic metrics, *i.e.,* BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). BLEU and ROUGE compute the ratios of overlapping n-grams between generated and real text, while METEOR measures word-to-word matches based on WordNet between generated and real text. Besides, we conduct human evaluation from these aspects following (Zou et al., 2021): *Fluency* evaluates whether the text is well-formed and logical to read; *Informativeness* measures whether the text contains useful information; *Accuracy* tests whether the text describes the given content accurately; *Relevance* measures whether the text is relevant to the given context; *Overall* evaluates the overall quality of the text. The overall quality is rated from 1 to 10, while the others are rated from 1 to 5.

Inspired by (Turing, 2009), we design a Turing test to further evaluate the generated text quality. In turing test, a human interrogator is requested to distinguish whether the given text is generated by human. For each model and gold text, we randomly select 500 text and each text is scored by judges.

Results and Analysis. Table 4 and Table 5 present the automatic evaluation and human evaluation results on composition ability, respectively. We can observe that, ProphetNet and BART achieve great performance on short text generation, while GPT-2 and T5 show better results on long text generation. Specifically, BART employs denoising objectives for reconstructing the corrupted original text and ProphetNet adopts future n-gram prediction, which are flexible for modeling the semantic relations between tokens and phrases in short texts. However, in long texts, a small ratio of masked tokens (*i.e.,* 15%) might be not effective to capture the complex long-range dependency. By comparison, the left-to-right prediction objective in GPT-2 can be more suitable to model the long-range semantic continuity in long text, and T5 has the largest model size to achieve a strong composition ability. For composition ability, we conclude that **the denoising**

objective is helpful for short text composition, while the left-to-right objective is more powerful for long text composition. Besides, the model size is also an important factor for the improvement of PLMs' composition ability.

4 Discussion

Based on the above four ability tests, we provide a guideline for helping researchers choose, apply, interpret and design PLMs for NLP tasks.

In section 3.3, we know that the improvement on memory ability is likely to be helpful for the performance of comprehension ability. Hence, designing PLMs with special training objectives such as permutation language modeling in XLNet for larger memory capacity will further benefit PLMs in the downstream comprehension tasks such as question answering. Besides, when applying PLMs to downstream comprehension tasks, it must be paid attention to the similarity of data distribution in pretraining and fine-tuning. Possible solutions such as intermediate fine-tuning can alleviate this problem to some extent.

Compared with comprehension, reasoning in section 3.4 is more complex and usually involves multiple sentences. Therefore, PLMs such as ALBERT trained with sentence-level objectives can be more suitable to conduct reasoning tasks. Intuitively, incorporating sentence-level objectives during pretraining will encourage PLMs to learn the correlation among different sentences. Note that, PLMs have better reasoning transferability between similar tasks, thus data-scarce tasks can be improved by first training on data-rich tasks.

For composition tasks, PLMs with denoising training objective performs well enough on short text composition, while PLMs with left-to-right objective or larger model size are more suitable for long text composition. The reason behind might be that PLMs with different training objectives can finally capture different ranges of semantic dependency between tokens and phrases.

5 Related Work

Pretrained Language Models. Owing to the great achievements Transformer (Vaswani et al., 2017) has made, the paradigm of pretrained language models (PLMs) is thriving (Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019b; Lewis et al., 2020; Raffel et al., 2020). It is widely recognized that PLMs can learn massive knowledge from corpus, leading to significant progress in various language tasks. Giving such results in extensive NLP tasks, now it has come to the point to systematically evaluate the abilities of PLMs, which can further deepen our understanding of PLMs and facilitate their application to more fields.

Language Model Evaluation. Many efforts have studied the evaluation on language model performance. Liu et al. (2019a) evaluate BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and ELMo (Peters et al., 2018) on a variety of linguistics tasks. Their results suggest that the features generated by PLMs are sufficient for high performance on a board set of tasks but fail on tasks requiring fine-grained linguistics knowledge. Tenney et al. (2019) evaluate similar models on a variety of sub-sentence linguistic analysis tasks, showing that PLMs encode both syntax and semantics into parameters. Zhou et al. (2020) is in line in the sense that PLMs can learn rich knowledge but focus on evaluating the commonsense. However, these work just focus on one dimension of PLMs ability evaluation. Other work such GLUE (Wang et al., 2019b) and CLUE (Liang Xu, 2020) just consider a simple mixture of multiple tasks lacking comprehensive evaluation. To the best of our knowledge, this is the first work to systematically evaluate PLMs by defining various kinds of ability and performing extensive comparison.

6 Conclusion

This paper investigates the general language ability evaluation of pretrained language models. We first design four evaluation dimensions, including memory, comprehension, reasoning, and composition, and further measure ten widely-used PLMs within five categories. Our experimental results demonstrate that the pretraining objectives and strategies have significant impacts on PLMs performance in downstream tasks. Besides, when fine-tuning PLMs in downstream tasks, their performances are usually sensitive to the data size and distribution, which can be addressed by designing some task-specific objectives. Furthermore, PLMs have great transferability between similar tasks. This characteristic can be utilized to solve the zero-shot and few-shot tasks. As a result, it is believed that this study will benefit future work about choosing or designing suitable PLMs for the target NLP tasks based on their properties.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Virginia W Berninger. 1999. Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly*, 22(2):99–112.

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

Kate Cain and Jane Oakhill. 2008. *Children's comprehension problems in oral and written language: A cognitive perspective*. Guilford Press.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

A. H. Miller P. Lewis A. Bakhtin Y. Wu F. Petroni, T. Rocktäschel and S. Riedel. 2019. Language models as knowledge bases? In *In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019*.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1930–1940. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Philip N Johnson-Laird. 1999. Deductive reasoning. *Annual review of psychology*, 50(1):109–135.

Alan S Kaufman and Elizabeth O Lichtenberger. 2005. *Assessing adolescent and adult intelligence*. John Wiley & Sons.

Patrick C Kyllonen and Raymond E Christal. 1990. Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14(4):389–433.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Lu Li Hai Hu Chenjie Cao Weitang Liu Junyi Li Yudong Li Kai Sun Yechen Xu Yiming Cui Cong Yu Qianqian Dong Yin Tian Dian Yu Bo Shi Jun Zeng Rongzhao Wang Weijian Xie Yanting Li Yina Patterson Zuoyu Tian Yiwen Zhang He Zhou Shaoweihua Liu Qipeng Zhao Cong Yue Xinrui Zhang Zhengliang Yang Zhenzhong Lan Liang Xu, Xuanwei Zhang. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

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

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics.

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

Akira Miyake and Priti Shah. 1999. *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 2401–2410. Association for Computational Linguistics.

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

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.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389. The Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2020, Online, July 5, 2020*, pages 27–33. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics - on what language model pre-training captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Douglas Walton. 2014. *Abductive reasoning*. University of Alabama Press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019c. Does it make sense? and why? A pilot study for sense making and explanation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4020–4026. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 93–104. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pretrained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press.

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. *arXiv preprint arXiv:2103.10685*.

Appendix

We give some experiment-related information as supplementary materials. The appendix is organized into six sections:

- Configurations and pretraining setting comparisons for selected models are presented in Appendix A;
- Data statistics of each test are presented in Appendix B;
- Full results for memory tests are presented in Appendix C;
- Full results for comprehension tests are presented in Appendix D;
- Full results for reasoning tests are presented in Appendix E; and
- Full results for composition tests are presented in Appendix F.

A Configurations of Pretrained Language Models

The selected ten PLMs within five categories and the comparisons of these PLMs in configuration and pretraining setting have been shown in Table 6. The effect extent of each factor for PLMs abilities in Table 7.

B Data Statistics

Memory Tests. The data statistics of LAMA and Wikipedia of each model are presented in Table 8. Due to the differences of each PLM, we drop the data that are not in the vocabulary.

Comprehension Tests. The data statistics of GLUE, SuperGLUE, SQuAD and RACE are presented in Table 9.

Reasoning Tests. The data statistics for commonsense reasoning, deductive reasoning and abductive reasoning are presented in Table 10.

Composition Tests. The data statistics for text summarization, question generation and story generation are presented in Table 11. For the first three datasets, we truncate the source text considering the input length of PLMs during training. And for WritingPrompts, we reconstruct the original dataset and discard examples where text contains more than 512 tokens.

C Memory Tests

Full results on LAMA and Wikipedia datasets are presented in Table 12.

D Comprehension Tests

Full results on SuperGLUE, SQuAD and RACE are presented in Table 13 and Table 14.

E Reasoning Tests

Full results on CommonsenseQA, ROCStories, SWAG, HellaSwag, Sense Making, and ARCT are presented in Table 15.

F Composition Tests

Full results on CNN/Daily-Mail, GigaWord, SQuAD, and WritingPrompts are presented in Table 16. Turing test results are presented in Table 5. We also show some summaries and stories generated by different PLMs in Table 18, Table 19, and Table 20.

Type	Models	Configurations		Pretraining Setting	
		Size	#Parameter	Corpus	Size
Bidirectional	BERT	base/large	110M/340M	BooksCorpus, English Wikipedia	16GB
	RoBERTa	base/large	125M/355M	BooksCorpus, CC-News, WebText, Stories	160GB
	ALBERT	xlarge/xxlarge	60M/235M	BERT Corpus	16GB
Unidirectional	GPT-2	small/medium	117M/345M	WebText (removing Wikipedia)	40GB
Hybrid	XLNet	base/large	110M/340M	BooksCorpus, English Wikipedia, Giga5, ClueWeb, Common Crawl	158GB
	UniLM	base/large	110M/340M	BERT Corpus	16GB
Knowledge-Enhanced	ERNIE	base	114M	English Wikipedia, Wikipedia	17GB
Text-to-Text	T5	base/large	220M/770M	Colossal Clean Crawled Corpus	745GB
	BART	base/large	140M/400M	RoBERTa Corpus	160GB
	ProphetNet	large	373M	RoBERTa Corpus	160GB

Table 6: Configurations and pretraining setting comparisons for our selected models.

Ability	MA	DD	MS	PO	PS
Mem.	☆☆	☆	☆	☆☆☆	☆☆☆
Compre.	☆☆	☆☆	☆	☆☆☆	☆☆☆
Reason.	☆	☆☆☆	☆	☆☆☆	☆☆
Compo.	☆	☆☆	☆☆☆	☆☆☆	☆

Table 7: The impact extent of each factor for PLMs abilities. MA, DD, MS, PO, and PS are short for model architecture, data distribution, model size, pretraining objective, and pretraining strategy, respectively

	G-RE	T-REx	ConceptNet	SQuAD	Wikipedia
#Origin	6,106	34,014	14,878	305	100,000
BERT / UniLM	5,527	34,014	11,658	305	85,836
RoBERTa	4,618	29,500	12,505	286	85,862
ALBERT	5,469	33,636	12,389	291	86,533
ERNIE	1,900	9,071	11,649	173	—
BART	4,618	29,500	12,505	286	85,862
T5	4,256	25,850	10,905	230	78,069
GPT-2	4,618	29,500	7,477	196	1,184
XLNet	5,202	32,293	12,080	279	85,228
ProphetNet	5,527	34,014	12,506	305	87,516

Table 8: Statistics of datasets in memory tests, including LAMA and Wikipedia. #Origin denotes the number of examples in original dataset, and the number of each model denotes the number of examples after selected.

Corpus		#Train	#Valid	#Test
GLUE	**WNLI**	635	71	146
	CoLA	8,551	1,043	1,063
	MNLI-M.	392,702	9,815	9,796
	MNLI-MM.		9,832	9,847
	RTE	2,490	277	3,000
	QNLI	104,743	5,463	5,463
	SST-2	67,349	872	1,821
	QQP	363,846	40,430	390,965
	STS-B	5,749	1,500	1,379
	MRPC	3,668	408	1,725
SuperGLUE	**CB**	250	57	250
	WNLI	635	71	146
	WSC	554	104	146
	COPA	400	100	500
	Wic	6,000	638	1,400
	BoolQ	9,427	3,270	3,245
	MultiRC	5,100	953	1,800
SQuAD	**v1.1**	88,567	10,790	-
	v2.0	131,924	12,165	-
RACE	**all**	25,137	1,389	1,407
		87,866	4,887	4,934
	middle	6,409	368	362
		25,421	1,436	1,436
	high	18,728	1,021	1,045
		62,445	3,451	3,498

Table 9: Statistics of datasets in comprehension tests including GLUE, SuperGLUE, SQuAD and RACE. #Train, #Valid and #Test denote the number of instances in train, valid and test set, respectively (the same as below). MNLI-M. and MNLI-MM. denote MNLI-match and MNLI-mismatch, respectively. SQuAD doesn't have test set, and we utilize the valid set as the test set.

Task	Corpus	#Train	#Valid	#Test
Commonsense reasoning	**CommonsenseQA**	9,741	1,221	1,140
Deductive reasoning	**ROCStories**	1,257	314	1,571
	SWAG	73,546	20,006	20,005
	HellaSwag	39,905	10,042	10,003
	Sense Making Task A	10,000	1,000	1,000
Abductive reasoning	**Sense Making Task B**	10,000	1,000	1,000
	ARCT	1,210	316	444

Table 10: Statistics of datasets in reasoning tests, including commonsense reasoning, deductive reasoning and abductive reasoning.

Task	Corpus	#Train	#Valid	#Test	#Input	#Output
Text summarization	**CNN/Daily Mail**	287,113	13,368	11,490	822.3	57.9
	Gigaword	3,803,957	189,651	1,951	33.7	8.7
Question generation	**SQuAD**	75,722	10,570	11,877	149.4	11.5
Story generation	**WritingPrompts**	67,765	3,952	3,784	30.2	281.2

Table 11: Statistics of datasets in composition tests, including text summarization, question generation and story generation. #Input and #Output denote the average number of tokens in the input text and output text.

Models	Vocab Size	LAMA-G	LAMA-T	LAMA-C	LAMA-S	Wikipedia	Average
$\text{BERT}_{\text{BASE}}$	28996	10.3	27.5	15.3	12.8	66.8	41.6
$\text{BERT}_{\text{LARGE}}$	28996	**11.0**	**29.2**	19.1	17.0	70.9	45.0
$\text{RoBERTa}_{\text{BASE}}$	50265	7.5	19.9	17.9	13.3	66.9	40.8
$\text{RoBERTa}_{\text{LARGE}}$	50265	7.1	23.9	**21.6**	**21.0**	71.1	44.8
$\text{ALBERT}_{\text{XLARGE}}$	30000	2.9	19.6	16.8	14.4	64.3	38.9
$\text{ALBERT}_{\text{XXLARGE}}$	30000	3.3	21.0	20.0	20.6	63.9	40.1
$\text{GPT-2}_{\text{SMALL}}$	50257	1.3	6.8	4.0	3.0	36.0	19.9
$\text{GPT-2}_{\text{MEDIUM}}$	50257	3.9	12.0	6.4	5.6	42.7	24.8
$\text{XLNet}_{\text{BASE}}$	32000	0.0	0.0	2.8	0.0	64.6	32.7
$\text{XLNet}_{\text{LARGE}}$	32000	0.0	0.0	5.5	0.4	68.7	35.1
$\text{UniLM}_{\text{BASE}}$	28996	8.5	27.6	15.4	11.8	66.9	41.4
$\text{UniLM}_{\text{LARGE}}$	28996	9.6	28.4	18.3	17.4	**71.5**	**46.4**
$\text{ERNIE}_{\text{BASE}}$	28996	1.3	13.4	13.0	8.1	-	-
T5_{BASE}	32100	5.5	20.0	13.2	9.6	60.5	36.3
T5_{LARGE}	32100	4.0	21.7	17.1	11.7	65.0	39.3
$\text{BART}_{\text{BASE}}$	50295	5.7	11.7	9.5	4.2	47.9	27.8
$\text{BART}_{\text{LARGE}}$	50295	9.4	15.8	7.7	3.1	47.8	28.4
$\text{ProphetNet}_{\text{LARGE}}$	30522	0.1	1.1	0.3	0.7	31.3	15.9

Table 12: Memory tests results on LAMA and Wikipedia datasets (test set). We report accuracy score for each dataset. Average is computed by averaging the scores of LAMA and Wikipedia (the score of LAMA is averaged among four dataset first). LAMA-G, LAMA-T, LAMA-C and LAMA-S denote the LAMA corpus Google-RE, T-REx, ConceptNet and SQuAD, respectively.

Model	**WSC** Acc.	**CB** F1/Acc.	**RTE** Acc.	**COPA** Acc.	**Wic** Acc.	**BoolQ** Acc.	**MultiRC** F1/EM	**Avg**
$\text{BERT}_{\text{BASE}}$	60.6	78.7/80.4	66.4	65.0	69.9	74.6	68.1/16.9	65.5
$\text{BERT}_{\text{LARGE}}$	63.5	89.0/92.9	70.1	73.0	<u>72.7</u>	75.6	69.4/22.6	70.3
$\text{RoBERTa}_{\text{BASE}}$	71.1	89.1/91.1	75.1	78.0	67.2	81.1	72.6/31.9	73.6
$\text{RoBERTa}_{\text{LARGE}}$	75.0	95.0/96.4	<u>88.2</u>	84.0	<u>72.7</u>	**85.4**	81.7/47.2	<u>80.8</u>
$\text{ALBERT}_{\text{XLARGE}}$	63.5	81.1/85.7	62.5	75.0	66.5	62.2	63.6/12.4	64.4
$\text{ALBERT}_{\text{XXLARGE}}$	64.4	87.6/92.9	70.4	**91.0**	**74.3**	62.2	**85.1/54.0**	74.6
$\text{GPT-2}_{\text{SMALL}}$	54.8	64.0/76.8	62.1	62.0	64.1	68.2	67.3/19.5	60.7
$\text{GPT-2}_{\text{MEDIUM}}$	61.5	84.4/82.1	63.6	63.0	67.2	73.9	71.5/29.2	66.1
$\text{XLNet}_{\text{BASE}}$	64.4	91.0/91.1	59.9	65.0	67.9	76.9	72.5/29.6	68.0
$\text{XLNet}_{\text{LARGE}}$	65.3	87.6/92.9	**88.5**	82.0	69.7	84.7	79.0/41.6	77.3
$\text{UniLM}_{\text{BASE}}$	63.5	74.7/82.1	60.3	67.0	68.5	73.3	67.9/20.5	65.0
$\text{UniLM}_{\text{LARGE}}$	65.4	86.5/87.5	70.9	76.0	72.3	82.3	75.7/36.3	72.8
$\text{ERNIE}_{\text{BASE}}$	65.4	81.6/82.1	68.8	64.0	70.8	74.4	68.7/21.3	67.2
T5_{BASE}	<u>79.8</u>	86.2/94.0	80.1	71.2	68.3	81.4	79.7/43.1	76.0
T5_{LARGE}	**84.6**	91.6/94.8	87.2	83.4	69.3	**85.4**	<u>83.3</u>/<u>50.7</u>	**81.4**
$\text{BART}_{\text{BASE}}$	64.4	86.6/85.7	69.5	70.0	65.7	75.7	74.2/31.7	69.2
$\text{BART}_{\text{LARGE}}$	65.4	**97.4/96.4**	83.5	<u>86.0</u>	70.4	85.1	82.9/50.6	79.2
$\text{ProphetNet}_{\text{LARGE}}$	63.5	<u>94.7</u>/<u>92.9</u>	51.3	61.0	60.7	67.4	64.7/17.2	62.7

Table 13: Comprehension tests results on SuperGLUE (valid set). Avg column is computed by averaging the scores of tasks to its left (the scores for CB and MultiRC are first averaged).

Models	**SQuAD v1.1**		**SQuAD v2.0**		**RACE**		
	EM	F1	EM	F1	RACE	RACE-M	RACE-H
$\text{BERT}_{\text{BASE}}$	80.8	88.5	72.8	76.0	65.0	71.7	62.3
$\text{BERT}_{\text{LARGE}}$	84.1	90.9	78.7	81.9	72.0	76.6	70.1
$\text{RoBERTa}_{\text{BASE}}$	86.1	92.3	80.3	83.4	72.8	72.6	26.6
$\text{RoBERTa}_{\text{LARGE}}$	<u>88.9</u>	<u>94.6</u>	<u>86.5</u>	<u>89.4</u>	83.2	<u>86.5</u>	81.3
$\text{ALBERT}_{\text{XLARGE}}$	86.1	92.5	83.1	86.1	78.1	76.7	79.8
$\text{ALBERT}_{\text{XXLARGE}}$	88.3	94.1	85.1	88.1	**87.4**	85.9	**87.1**
$\text{GPT-2}_{\text{SMALL}}$	63.6	75.1	57.1	61.5	61.2	62.9	58.2
$\text{GPT-2}_{\text{MEDIUM}}$	70.3	80.8	61.5	66.0	62.2	65.0	61.4
$\text{XLNet}_{\text{BASE}}$	12.8	14.7	78.5	81.3	71.3	72.8	67.5
$\text{XLNet}_{\text{LARGE}}$	**89.7**	**95.1**	**87.9**	**90.6**	<u>85.4</u>	**88.6**	<u>**84.0**</u>
$\text{UniLM}_{\text{BASE}}$	82.8	89.9	74.9	78.0	59.0	64.1	50.3
$\text{UniLM}_{\text{LARGE}}$	86.5	92.7	80.5	83.4	70.3	70.0	66.4
$\text{ERNIE}_{\text{BASE}}$	-	-	-	-	-	67.8	-
T5_{BASE}	85.4	92.1	77.6	81.3	70.6	74.4	68.4
T5_{LARGE}	86.7	93.8	-	-	80.4	82.6	77.8
$\text{BART}_{\text{BASE}}$	84.6	91.0	76.0	79.2	70.1	72.4	63.2
$\text{BART}_{\text{LARGE}}$	88.8	<u>94.6</u>	86.1	89.2	82.2	82.5	79.6
$\text{ProphetNet}_{\text{LARGE}}$	-	-	-	-	-	74.1	-

Table 14: Comprehension tests results on SQuAD and RACE (test set).

Model	CQA	ROCStories	SWAG	HellaSwag	SM-A	SM-B	ARCT
$\text{BERT}_{\text{BASE}}$	53.0	88.1	81.6	40.5	87.3	80.1	65.1
$\text{BERT}_{\text{LARGE}}$	55.9	90.2	86.3	47.3	89.4	85.8	71.2
$\text{RoBERTa}_{\text{BASE}}$	72.1	93.3	82.6	61.0	89.3	87.5	46.1
$\text{RoBERTa}_{\text{LARGE}}$	72.2	**97.4**	89.9	85.2	**93.0**	**92.3**	57.9
$\text{ALBERT}_{\text{XLARGE}}$	66.2	90.4	84.6	75.9	87.9	89.4	56.1
$\text{ALBERT}_{\text{XXLARGE}}$	**80.0**	97.1	**90.7**	**90.1**	92.5	**92.3**	79.5
$\text{GPT-2}_{\text{SMALL}}$	47.8	58.8	48.1	39.9	84.2	74.7	66.0
$\text{GPT-2}_{\text{MEDIUM}}$	60.8	59.9	79.7	60.4	88.7	73.4	66.7
$\text{XLNet}_{\text{BASE}}$	53.8	92.0	80.4	55.1	81.6	85.4	80.2
$\text{XLNet}_{\text{LARGE}}$	62.9	93.8	86.8	79.7	83.7	88.7	83.1
$\text{UniLM}_{\text{BASE}}$	47.6	80.6	77.0	36.3	86.2	83.6	48.4
$\text{UniLM}_{\text{LARGE}}$	62.3	86.9	83.1	46.7	89.3	86.4	72.3
$\text{ERNIE}_{\text{BASE}}$	54.1	84.7	-	-	88.7	-	73.7
T5_{BASE}	61.9	88.2	65.8	55.2	89.2	82.9	63.3
T5_{LARGE}	69.8	91.4	73.7	79.1	92.7	88.2	69.4
$\text{BART}_{\text{BASE}}$	61.0	88.9	81.2	53.4	72.0	67.9	71.8
$\text{BART}_{\text{LARGE}}$	75.8	91.7	87.9	76.6	82.9	67.9	**84.2**
$\text{ProphetNet}_{\text{LARGE}}$	21.3	82.2	70.1	26.4	85.5	78.0	65.5

Table 15: Reasoning tests results on seven datasets (test set). We report accuracy score for each dataset. CQA is short for CommonsenseQA. SM-A and SM-B denote the Task A and Task B of Sense Making, respectively.

Models	CNN-DailyMail			GigaWord			SQuAD			WritingPrompts		
	R-1	R-2	R-L	R-1	R-2	R-L	B-4	R-L	ME	B-4	R-L	ME
$\text{GPT-2}_{\text{SMALL}}$	24.60	7.21	21.06	25.25	9.03	23.20	5.13	14.83	21.06	11.58	3.80	8.18
$\text{GPT-2}_{\text{MEDIUM}}$	22.95	5.99	22.08	23.72	8.12	21.56	8.48	18.82	26.77	14.47	3.23	7.29
$\text{UniLM}_{\text{BASE}}$	17.83	0.11	5.50	16.64	6.11	15.12	4.47	17.65	20.30	**27.71**	2.35	5.47
$\text{UniLM}_{\text{LARGE}}$	43.44	20.21	40.51	38.45	19.45	35.75	4.42	17.43	20.13	26.88	1.84	5.01
T5_{BASE}	42.05	20.34	39.40	33.13	15.60	30.18	11.18	21.82	29.93	6.04	**4.61**	**9.81**
T5_{LARGE}	42.50	20.68	39.75	34.75	16.26	31.49	11.19	22.35	30.53	8.61	4.19	9.51
$\text{BART}_{\text{BASE}}$	36.36	20.87	33.32	38.65	19.43	35.82	14.44	24.11	36.92	11.91	3.57	7.69
$\text{BART}_{\text{LARGE}}$	44.16	**21.28**	40.90	39.41	20.21	36.42	**15.87**	**25.47**	**38.42**	14.72	3.14	7.08
$\text{ProphetNet}_{\text{LARGE}}$	**44.20**	21.17	**41.30**	**39.51**	**20.42**	**36.69**	14.20	23.97	35.99	19.31	2.59	7.19

Table 16: Composition tests results on four datasets. R-1, R-2, R-L are short for ROUGE-1, ROUGE-2, ROUGE-L respectively. B-4 and MT denote BLEU-4 and METEOR, respectively.

Models	TT (%)	Fluency	Informativeness	Accuracy	Coherence	Overall
$\text{GPT-2}_{\text{MEDIUM}}$	**45.7**	**3.42**	**3.17**	3.20	**3.23**	5.87
$\text{UniLM}_{\text{LARGE}}$	1.2	1.32	1.88	2.03	1.71	2.74
T5_{LARGE}	34.4	3.01	2.80	3.09	2.87	5.18
$\text{BART}_{\text{LARGE}}$	45.2	3.37	3.16	**3.39**	3.22	**5.96**
$\text{ProphetNet}_{\text{LARGE}}$	29.6	2.95	2.91	3.10	2.89	5.18
Gold	71.3	3.79	4.07	3.87	3.80	7.37

Table 17: Turing test and human scores on the test set of WritingPrompts. TT is short for Turing Test and represents the percentage of makers who believe the text is written by human. Coherence evaluates how content is coherent considering both intra- and inter-sentence correlation of a paragraph. Rated from 1 to 5.

Passage	japan 's nec corp. and UNK computer corp. of the united states said wednesday they had agreed to join forces in supercomputer sales .
GPT-2$_{\text{MEDIUM}}$ **UniLM**$_{\text{LARGE}}$ **T5**$_{\text{LARGE}}$ **BART**$_{\text{LARGE}}$ **ProphetNet**$_{\text{LARGE}}$	unk computer to jointly sell supercomputers nec [UNK] to join forces in supercomputer sales nc and unk computer to join forces in supercomputer sales nec and unk computer to join forces in supercomputer sales nec unk computer to join forces in supercomputer sales
Gold	nec UNK in computer sales tie-up
Passage	the sri lankan government on wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country .
GPT-2$_{\text{MEDIUM}}$ **UniLM**$_{\text{LARGE}}$ **T5**$_{\text{LARGE}}$ **BART**$_{\text{LARGE}}$ **ProphetNet**$_{\text{LARGE}}$	closed as military campaign escalates sri lankan gov t shuts down schools in north sri lanka closes schools as fighting escalates sri lanka closes government schools sri lanka closes schools as military escalates
Gold	sri lanka closes schools as war escalates

Table 18: Test set examples of GigaWord dataset for each model.

Prompts	[WP] You have a very shitty type of precognition .
GPT-2$_{\text{MEDIUM}}$	<newline>“ I’m sorry, ” I said. “ I don’t know what you’re talking about. ” I looked around the room, trying to find the source of the noise. “ What do you mean? ” I asked. “ It’s not like I’ve never heard it before. ” He looked at me, confused. “ No, ” he said, “ it’s never happened to me before. I ’d be surprised if it ever did. ” “ What’s that supposed to mean? You’re telling me that I can’t remember anything before? ” “ No. ” he replied. “ That’s exactly what it means. You’ve always been precognitions, and you ’ll never be able to remember anything. ”
UniLM$_{\text{LARGE}}$	i’m not sure what i’ve done wrong. i’d rather not be able to do it. i don’t know what i have done wrong, i don’t know what it has done. i have been told that i have n’ve been wrong. <newline ><new york. <york ><yoline >i’ll never forget that day. «new >«line >the day i was born. the day that i was raised. the one that i never forget. the time that i first met her. the moment that i met her, the time when i first saw her. <nowline >my life is a blur. the blur is the blur. i am not sure if i have ever been wrong, or if i had ever been right. i ca n’re sure if it was the blur or the blur, but i do know it was. <tim. <the blur <new newline. <i am a blur, the blur and the blur are the same. the same, the same is the same thing. the difference is the difference between the two. the two are the difference. the differences are the differences. the different things. the things... the changes. the change. the changing. the thing.. is the change, the change is the thing that is the different. the new thing. <.. <this is the story of a man who was born in the blur of the two, the one where i was n’s. the other where i am. the first time i met. the second time i saw her, and the first where i never forgot. the story is the one in the middle of the blur where i met the woman i loved. the woman who i loved, the woman that i loved the woman. the man who i love the woman with the most. the person who i hate the most, the person that i hate most. <cline >* * * <new story ><nbsp >i am the blur ; i am i. <pline »<bsp. i was a blur ; <new line ><i was the one who was the first to be born. <first time. <np ><do nbs. <do ibs. i did n’i do it, i did not know it had happened. <line i woman i <i i i ve, the <blur ve
T5$_{\text{LARGE}}$	<newline>It’s been a long time since I’ve been able to use my powers. I haven’t had a chance to use them in a while, but I do have a few things I can do to help. The first thing I do is go to the bathroom, and I have to wash my hands. It is a pain, I know it. But I do it anyways, because I have the ability to see things. And I can see them. They can see me in the mirror, they can see my eyes, their eyes. My hands, my feet, the way they move, how they feel. How I feel, what they do to me. What I do to them? How do they know I am? I can hear them in my ears, a scream, or a cry. A sigh of relief, as I look at the mirror and see them in front of me, like a mirror. You’re gon na see me? ”
BART$_{\text{LARGE}}$	I’ve always been a bit of a weirdo. <newline>I’m not sure why, but I always knew I was going to die at some point in my life. I don’t know if it’s because I was born in the wrong year, or if it was because my parents died when I was very young. I just knew that I would die. I’ve never been able to explain it to anyone, but it was always there in the back of my mind. It’s not like I could do anything about it. I couldn’t even do anything to stop it. It was just always there. I never knew what it was, or what it would do to me. I didn’t want to know. I wanted it to be over. But it never was. It just kept getting worse and worse. I can’t explain it, I just know that I ’ll die.
ProphetNet$_{\text{LARGE}}$	it’s been a long time since i’ve had a precognition, but i don’t think i’ll ever be able to get rid of it. <newline >it started when i was a kid. i was told that i had a very shitty type of pre - lucid dreaming. i had no idea what it was until i was in my late teens. it was like i could see the future in front of me. i knew exactly what i was going to do when i saw the future. i could feel it in my bones. i can feel it. i know exactly what’s going to happen when i look at the future, and i know it will be the best thing that’s ever happened to me. the future is going to be better than i ever thought it would be. i’m not going to let it happen again.
Gold	Just looking at the guy I already knew it . He was in for a seven on the Bristol scale and it was only hours away . <newline><newline>Looking at his carefree face from below as he was slowly transported towards me on the escalator , I wondered , not for the first time , if I should warn him . But as always I asked myself : How could I tell him ? How could I tell anyone ? <newline><newline>Tell people that I could see the future . That I could see their fate ? That I could feel the forces that were at work , often without their knowledge and beyond their control , inside their own bodies ? That I could see shape of things to come ? That I was a coproscognitive ? <newline><newline>Nobody would believe me . Not until after a few predictions had been proven to be correct at least and few humored me for long enough to realize the truth of my predictions . <newline><newline>And this guy did n’t have a few predictions to confirm the authenticy of my gift . He needed to know and believe me about the very next event I could foresee and there was no way to convince him in time . <newline><newline>And thus I stayed silent again as the escalator carried us past one another leaving him to his fate . I was going up and he was going down . <newline><newline>I could only hope for him that he would be close to the right place two hours and thirteen minutes from now , because I could tell it would be a fast and quite explosive one and without my word he would not have much advanced warning telling him to run for salvation . <newline><newline>I put him out of my mind I had my own future to worry about and I did n’t need my gift to tell me about it . I started looking for the signs pointing the way to closest lavatory .

Table 19: Test set examples of WritingPrompts dataset for each model.

Prompts	[WP] “ Before we allow humankind to live among us in the stars , tell us the faults of your kind . ”
GPT-2$_{\text{MEDIUM}}$	<newline>“ I’m sorry, sir, but I don’t think we’re ready for this. ” I said. “ We’ve been here for thousands of years, and we’ve never seen anything like this before. ” He looked at me with a look of horror on his face. “ What do you mean? ” I asked. He looked down at his hands, and then back up at me. “ It’s like this, ” he said, “ You can’t see it, but it’s there. It ’ll be here for a long time, and it ’ll kill us all. ” <newline>I looked at him, and he looked back at me, and both of us looked at the stars.
UniLM$_{\text{LARGE}}$	‘ ‘ we allow humanskind to be among us, tell them the faults.’<newline ><new new >‘ ‘ humanskind,” i said, ‘ ‘ you are the one who has been chosen to be our leader.’‘ ‘ humankind, you are our leader,’‘ i said. <new york. <york ><yoline >‘’humanskind.’* <newrk >* <yo >«new ><>humanskind : <new name ><nowline >humans : <now name >humans. <now names ><the name of the universe. <the names of the stars. «line >* humanskind * <now named ><first name >the universe, <new names >the stars and the stars <new stars. the names <new planets. <first names >* * humans * <firstline >the name <new species ><humanline »<humans ><last name >* humankind *. <last names >humans * * <lastline >humankind. <name »* humanity * <name <><* humans. * <* human * «»humans, <now known as humans. the name. <.. <* humanity. <human name >... * * * humanity <new humans >*.. humans *. *. humans.. ’. <line. <humans.’s.. human.’the humans. they were the humans, the humans of the galaxy. <: «<humans : humans. humans. humans, humans. humankind <new galaxy ><: / / www. reddit. com / r / writingprompts / comments / 2jclq / comments _ 2jflq _ comments / 1xxxfxxgxxcxxbxxdxxkxxqxx <new _ ><_ _ <new i am humankind : the humankind of the galactic system. <_ «_ >i am the human race. <tv ><tline >i was the human. <pline >it was a long time since i was human. i was a human.. i am a human race,..kind <. <’humans human «race humans <* <human of * <the «* i humans. new..
T5$_{\text{LARGE}}$	Before we allow humankind to live amongst the stars, tell us the faults of your kind. ” newline>I don’t know, I’m not a scientist, but I do have a degree in astronomy, and I do know a thing or two about science. I know that a lot of people think that science is a good thing, that it’s a great thing. But, if you think about it, you’re a fucking shithole. You’ve got a bunch of crazies, all of them. So, what do you think?? Do you know what? I mean, they ’ll tell you. And, of course, we ’d like to know what you think of us.
BART$_{\text{LARGE}}$	“ Before we allow humankind to live among us in the stars, tell us the faults of your kind. ” <newline><newlines>“ Well, first of all, they aren’t very smart. They don’t know how to read. They’re not very good at math. They haven’t learned how to write yet. They are also very lazy. They spend most of their time staring at their screens. They can’t even get up to go to the bathroom. They just sit there and stare at the screen. They also have a tendency to stare at their phones for hours at a time. I’m not sure why they do that, but I guess it’s because they’re bored. ”
ProphetNet$_{\text{LARGE}}$	‘ ‘ before we allow humankind to live among us in the stars, tell us the faults of our kind.” <newline >‘ ‘ i’m sorry, sir, but we don’t have the technology to do that. we’re too afraid of the consequences of our actions, and we’ve spent too much time trying to find a way to stop them.’cause they’re just too stupid to do anything about it. we have to do something about it, or we’ll never be able to get out of here. we need to find some way to get them out of there, and if they do, then we’d have to go back to earth and start all over again. and if that’s the case, then i’d like to thank you for your time, and i hope to see you again soon,”
Gold	Tell us your faults ? Really ? This was the question - the shibboleth - that unlocked the cosmos ? <newline><newline>The Masters could have picked a scientist to answer but they feared she might mask ignorance . They could have picked from our global leaders bit they feared that they would mask deceit . They could have picked a holy man but feared he would mask violence , oppression , hate , intolerance ... the list of disqualifying sins was almost too long to enumerate . <newline><newline>So they picked Josh Thornton , a 45 year old MBA in human resources . <newline><newline>“ Our greatest weakness ? Well , I think we work a little too hard and , as a race , we might be a bit of a perfectionist .

Table 20: Test set examples of WritingPrompts dataset for each model.