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

Junyi Li[1,5], **Tianyi Tang**[1], **Zheng Gong**[1], **Lixin Yang**[2], **Zhuohao Yu**[2], **Zhipeng Chen**[2], **Jingyuan Wang**[3,4], **Wayne Xin Zhao**[1,5*] and **Ji-Rong Wen**[1,5]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Renmin University of China [3]Peng Cheng Laboratory
[4]School of Computer Science and Engineering, Beihang University
[5]Beijing Key Laboratory of Big Data Management and Analysis Methods
{lijunyi,steven_tang}@ruc.edu.cn batmanfly@gmail.com

Abstract

Nowadays, pretrained language models (PLMs) have dominated the majority of NLP tasks. While, little research 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**). In our study, we design four evaluation dimensions, *i.e.,* memory, comprehension, reasoning, and composition, to measure ten widely-used PLMs within five categories. Our empirical results demonstrate that: (1) PLMs with varying training objectives and strategies are good at different ability tests; (2) fine-tuning PLMs in downstream tasks is usually sensitive to the data size and distribution; (3) PLMs have excellent transferability between similar tasks. Moreover, the prediction results of PLMs in our experiments are released as an open resource for more deep and detailed analysis on the language abilities of PLMs. This paper can guide the future work to select, apply, and design PLMs for specific tasks. We have made all the details of experiments publicly available at https://github.com/RUCAIBox/ElitePLM.

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 being pretrained on massive unlabeled text, PLMs can be directly fine-tuned on downstream tasks, entirely removing the need for task-specific architectures (Radford et al., 2018). This paradigm has led to significant progress on many challenging NLP tasks such as reading comprehension (Devlin et al., 2019) and text generation (Brown et al., 2020).

With rising new state-of-the-art results that approach or surpass human performance on several tasks, it is a non-trivial research topic about how to systematically evaluate the language abilities of PLMs from a wide range of perspectives. Given a wide range of publicly released PLMs, it is particularly useful to derive principles or guidelines for selecting suitable PLMs for specific downstream tasks. However, existing works either target 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 of PLM's abilities in large-scale NLP tasks. To fill the gap of PLMs evaluation, we introduce the gen**E**ral **l**anguage ab**i**l**i**t**y** **e**valuation (**ElitePLM**) for empirically and systematically assessing the general language abilities of PLMs.

The ideal goal behind PLMs is to create a human-like machine learner where it can understand the language and then perform any specific task related to language. In cognitive science, 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 humans. 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 humans, 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 select 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 specific ability.

In human intelligence tests, the background of participants (*e.g.,* gender, race, and occupation)

*Corresponding author

should be as much as diverse. Thus, in ElitePLM, we select a diversity of PLMs to conduct generalized and meaningful comparisons. According to training objectives, PLMs can be divided into three types: bidirectional LMs (*e.g.,* BERT (Devlin et al., 2019)) for natural language understanding (NLU), unidirectional LMs (*e.g.,* GPT (Radford et al., 2019)) for natural language generation (NLG), and hybrid LMs (*e.g.,* UniLM (Dong et al., 2019)) for combining these two paradigms. Furthermore, knowledge-enhanced LMs (*e.g.,* ERNIE (Zhang et al., 2019)) and text-to-text LMs (*e.g.,* T5 (Raffel et al., 2020)) also emerge as important branches of PLMs. Considering the variety, we finally select ten widely-used PLMs within the above five categories and evaluate their abilities on four dimensions. We show the comparisons of these PLMs in Table 7 of Appendix A.

From the ability test results, we have three salient findings. First, PLMs with varying pretraining objectives and strategies are good at different kinds of downstream tasks. Specifically, we observe that bidirectional LMs like BERT and pretraining strategies like larger training batches in RoBERTa are helpful for memorizing pretraining corpora; permutation language modeling in XLNet is beneficial for modeling the bidirectional context in language comprehension; inter-sentence coherence objective in ALBERT is suitable for sentence-level reasoning tasks; text-to-text LMs using denoising objective like BART perform better in short text generation. Second, when fine-tuning PLMs in downstream tasks, their performance is typically sensitive to the data distribution in fine-tuning stage, which can be addressed by incorporating intermediate datasets or tasks to alleviate such a discrepancy. Third, PLMs have excellent transferability between similar tasks, especially reasoning tasks. This finding will inspire future researchers to leverage data-rich tasks for improving data-scarce tasks. For more clarity, we illustrate the impact level of each factor for PLMs' abilities in Table 8 of Appendix A.

Besides ElitePLM being an evaluation benchmark of PLMs' language ability, more importantly, the predicted results of ElitePLM can be used as an open resource for more depth and granularity in analyzing PLMs performance on each ability. For example, we further analyze the comprehension test results of PLMs across answer types in QA tasks. The analysis shows that PLMs are good at simple single-token answers such as dates but more challenged on intricate phrase answers. Moreover, by analyzing human test and Turing test results on composition, we observe that summaries with high accuracy are more likely to pass the Turing test while rich information is more important for story generation. Overall, ElitePLM can act as an analysis tool to gain more insight into PLMs. We show the details of our used datasets and predicted outputs of PLMs in Appendix B.

This paper is intended to help establish sound principles for choosing, applying, interpreting and designing PLMs for NLP tasks in practical settings. We have released the code and predicted results of each ability experiment, providing the research and industry community with off-the-shelf tools to evaluate and analyze their PLMs.

2 ElitePLM

In this section, we will detail these four kinds of language abilities, *i.e.,* memory, comprehension, reasoning, and composition, in ElitePLM.

Memory Ability. Memory is the most basic ability of humanity, involved in how much information we recall throughout our lives (Miyake and Shah, 1999). By analogy, ElitePLM will measure how much knowledge and language patterns PLMs have memorized in pretraining, as assessed by tests of recalling words based on contexts. Based on the memorized information, PLMs can effectively adapt to downstream tasks for understanding and reasoning about the similar context in a specific text. On the other hand, efficiency is also a critical 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 in terms of memorizing the given new information.

Comprehension Ability. Comprehension is an intricate and multifaceted ability. It typically consists of understanding a text's vocabulary, background knowledge of a specific topic, and comprehension of its linguistic structures like grammar (Cain and Oakhill, 2008). In particular, background (prior) knowledge is used to comprehend a special situation, lesson, or text. For example, readers should be aware of the background knowledge of dog behavior when reading a text about dog training. In ElitePLM, we will assess PLMs' comprehension ability from three aspects, *i.e.,* vocabulary, background knowledge, and linguistic structures.

Reasoning Ability. Based on the comprehension of a text, reasoning ability refers to the power of the processes and strategies used in drawing inferences, reaching conclusions, arriving at solutions, and making decisions (Kyllonen and Christal, 1990). In ElitePLM, we mainly focus on three types of reasoning abilities. In detail, commonsense reasoning requires PLMs to draw 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. 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); Abductive reasoning involves reaching 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. In the literature (Connors, 1997), composition is a highly intelligent and synthetic process where a writer assembles words and sentences to create a coherent and meaningful work (*e.g.,* poem, music, and novel) from scratch, which closely resembles to the text generation task in NLP (Berninger, 1999). Therefore, in ElitePLM, we introduce several text generation tasks to evaluate 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 deliberate the text structure.

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: (1) *Bidirectional LMs*: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and ALBERT (Lan et al., 2020); (2) *Unidirectional LMs*: GPT-2 (Radford et al., 2019); (3) *Hybrid LMs*: XLNet (Yang et al., 2019) and UniLM (Dong et al., 2019); (4) *Knowledge-enhanced LMs*: ERNIE (Zhang et al., 2019); (5) *Text-to-Text LMs*: BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and ProphetNet (Qi et al., 2020). We implement these models and ability tests mostly on huggingface (Wolf et al., 2020), fairseq (Ott et al., 2019), and jiant (Phang et al., 2020). To reflect the true level of language abilities, we adopt the best hyper-parameter values reported in their original papers for each PLM.

3.2 Memory Tests

Datasets and Metrics. The goal of memory tests is to assess how much knowledge and language patterns PLMs have memorized during pretraining. 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 corpora for our selected PLMs (except GPT-2 and T5). Therefore, to conduct a fair comparison, we continuously train GPT-2 and T5 on Wikipedia using their pretraining objectives. Similar to LAMA, we randomly sample 100,000 texts 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 patterns and factual knowledge in PLMs' memory. For metrics, we use *Mean Precision at One* ($P@1$) of predicting missing tokens. 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.

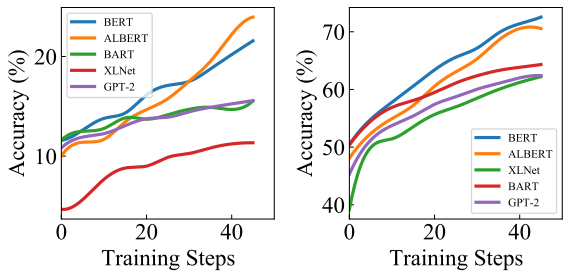
Results and Analysis. To evaluate how much text PLMs have recalled in pretraining, we directly test PLMs using Wikipedia and LAMA without fine-tuning, similar to zero-shot learning. The results on $P@1$ metric are shown in Table 1. Compared with bidirectional and hybrid LMs (*e.g.,* BERT and XLNet), GPT-2 uses auto-regressive self-attention where every token can only attend to the context to

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

Table 1: Memory test results on LAMA and Wikipedia datasets (test set). These results are based on the LARGE version of each PLM 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).

Relation	Template	BERT	RoBERTa	GPT-2	BART	T5
<[X], place_of_death, [Y]>	[X] died in [MASK].	13.98	0.46	0.15	11.09	4.19
	[X] passed away in [MASK].	13.46	0.46	0.62	3.54	1.51
	[X]'s place of death was [MASK].	3.27	0.00	0.00	0.00	1.51
<[X], place_of_birth, [Y]>	[X] was born in [MASK].	16.07	12.52	7.53	14.77	6.32
	[X] was born in the place of [MASK].	2.83	1.29	0.00	0.00	1.39
	[X]'s place of birth was [MASK].	12.16	1.87	0.00	0.00	3.12

Table 2: The impact of template on eliciting PLMs' stored knowledge.



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

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

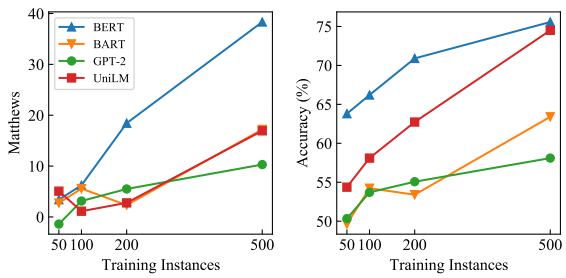
its left. This unidirectional training objective naturally limits the performance of GPT-2 in terms of memorizing information. It has been previously reported that PLMs can remember more information by scaling up the model size (Brown et al., 2020). However, in our tests, BART-large (400M) achieves worse results than RoBERTa-base (125M) with the same training corpus and similar vocabulary sizes (50,295 vs 50,265). During pretraining, RoBERTa adopts bidirectional objectives and novel strategies like larger training batches. It can be concluded that, as opposed to model size, **training objectives and strategies reflect the way that PLMs memorize information, making significant impacts on PLMs' memory ability**. Besides, we can clearly observe that all PLMs achieve their best results in T-REx (created from Wikipedia triples) among LAMA, and perform relatively well on Wikipedia. This implies that PLMs indeed remember a large proportion of knowledge and language patterns from pretraining corpora.

To test the memory efficiency, we fine-tune five models, BERT, ALBERT, GPT-2, BART, and XLNet, for several epochs. 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 the bidirectional training objective is also helpful to facilitate the memory efficiency since bidirectional language modeling can make PLMs more quickly capture the language patterns.

Based on the memory test results, we further analyze how to effectively elicit the information from PLMs' memory. LAMA hand-crafts templates to test PLMs by filling the [MASK] token. Therefore, we conduct a pilot study on designing different templates for two relations in Google-RE. Table 2 shows that different templates can result in substantial differences in eliciting PLMs' memory. The bidirectional LMs, *e.g.,* BERT, show relatively adaptability to varying templates, further verifying their strength in memory ability. Therefore, with large-scale knowledge stored in PLMs, how to derive an effective and appropriate method to provoke them is a key challenge.

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 3: 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.

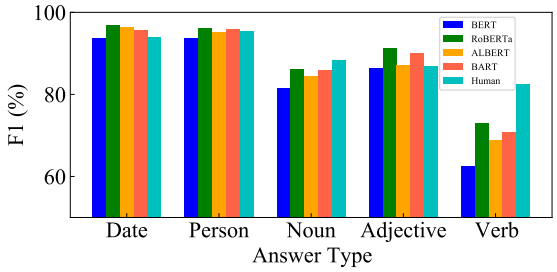


Figure 3: PLMs Performance on SQuAD v1.1&v2.0 stratified by five types of answer.

3.3 Comprehension Tests

Datasets and Metrics. In comprehension tests, we take into account three aspects of comprehension ability, including vocabulary, background knowledge, and linguistic structures. Therefore, we 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 comprehension benchmarks. Several tasks, like word sense disambiguation and coreference resolution, can assess PLMs' understanding of vocabulary meaning and grammatical structure of a 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*". For evaluation, we report the corresponding metrics results for each task, such as the *Matthews corr.* metric for CoLA.

Results and Analysis. Table 3 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. The results indicate that **the improvement on memory ability is beneficial for the performance of comprehension ability**, which is in line with our intuition. Compared with bidirectional language modeling in

[1] https://gluebenchmark.com/

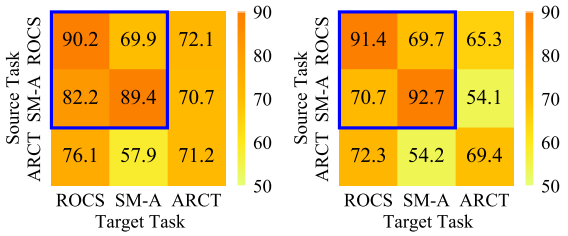
Datasets	Bidirectional			Uni.	Hybrid		KE	Text-to-Text		
	BERT	RoBERTa	ALBERT	GPT-2	XLNet	UniLM	ERNIE	T5	BART	ProphetNet
CQA	55.9	72.2	**80.0**	60.8	62.9	62.3	54.1	69.8	75.8	21.3
ROCStories	90.2	**97.4**	97.1	59.9	93.8	86.9	84.7	91.4	91.7	82.2
SWAG	86.3	89.9	**90.7**	79.7	86.8	83.1	80.2	73.7	87.9	70.1
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 4: Reasoning tests results on seven datasets (test set). 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.

BERT, permutation language modeling (relying on all permutations of the factorization order) used in XLNet enables PLMs to learn more context for enhancing PLMs' understanding of 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) since it has different data distribution from the pretraining corpora (Wang et al., 2021). This kind of sensitiveness to unfamiliar tasks is also reflected in Figure 2, where the model performance on CoLA shows 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 fine-tuning**. To solve this challenge, it will be better to adopt intermediate fine-tuning, which involves first fine-tuning PLMs on an intermediate dataset similar to the final target dataset and then transferring tuned PLMs to the final dataset.

To gain more insights into PLMs' comprehension ability, we choose four representative PLMs (*i.e.,* BERT, RoBERTa, ALBERT, and BART) and humans to analyze their performance across the answer types of SQuAD v1.1&v2.0. The results in Figure 3 show that PLMs perform well on simple answers such as dates and persons. For these categories of answers, there are usually only a few plausible candidates and most answers are single tokens. The models are more challenged on other intricate answer types (*e.g.,* noun and verb phrases) because there are many more plausible candidates and multiple tokens. Thus, improving PLMs' understanding of intricate named entities during the pretraining phase will possibly benefit PLMs' comprehension ability later.



(a) BERT$_{\text{LARGE}}$ (b) T5$_{\text{LARGE}}$

Figure 4: Heatmaps of two-stage transfer learning.

3.4 Reasoning Tests

Datasets and Metrics. In reasoning tests, we mainly consider three forms of reasoning, *i.e.,* commonsense reasoning, deductive reasoning, and abductive reasoning, focusing on commonsense utilization, conclusion induction, and reason derivation, respectively. For evaluation, we choose 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). Specifically, CommonsenseQA requires PLMs to reason about commonsense knowledge in 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. For evaluation, we report the *Accuracy* results for each dataset.

Results and Analysis. Table 4 shows the model performances in reasoning ability. It can be clearly observed that performing well in comprehension tests, ALBERT and RoBERTa also achieve stronger performance in almost all reasoning tasks. In pre-

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 5: 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.

training, ALBERT introduces an inter-sentence coherence objective to capture the relationship among sentences, which is 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 (Liu et al., 2019b). Interestingly, despite being 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 relationship between sentences, thus leading to poor reasoning ability. **To improve PLMs' reasoning ability, it would be useful to design sentence-level reasoning objectives like inter-sentence coherence loss in ALBERT**. Moreover, despite incorporating knowledge, ERNIE still shows mediocre performance in knowledge-related datasets such as CQA. A possible reason might be that ERNIE only uses trained KB embeddings to enhance semantic representations but ignores the reasoning structure of KBs. This inspires us that designing appropriate and effective fusion methods to integrate knowledge is more important.

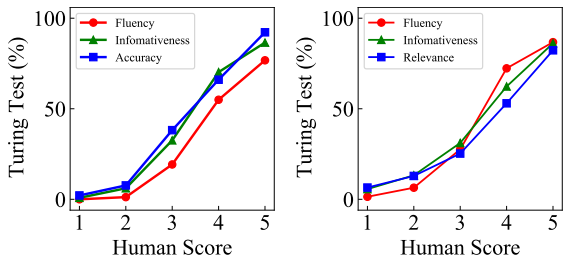
To further analyze the transferability of PLMs' reasoning ability, we conduct a two-stage study on three task datasets, *i.e.,* ROCStories, SM-A, and ARCT. We first train PLMs on source tasks with full data and then fine-tune PLMs on target tasks with ten instances. In Figure 4, it can be observed that **PLMs have better reasoning transferability between similar tasks** such as deductive reasoning tasks (ROCStories and SM-A). This shows that model performance on data-scarce reasoning tasks can be improved by incorporating additional training on data-rich similar tasks (Wang et al., 2021).

3.5 Composition Tests

Datasets and Metrics. Composition is similar to the text generation task, aiming at generating new content from scratch. Therefore, we use 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. According to the length of the target text, text summarization and question generation is short text generation, while story generation is long text generation. For evaluation, we adopt three automatic metrics, *i.e.,* BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Besides, following (Zou et al., 2021), we conduct human test from five aspects, *i.e., Fluency*, *Informativeness*, *Accuracy*, *Relevance* and *Overall*. The overall score is rated from 1 to 10, while the others are rated from 1 to 5. Inspired by Turing (2009), we further de-

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 6: 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.



(a) GigaWord (b) WritingPrompts

Figure 5: Impact factors of Turing Test.

sign a Turing test to assess the generation ability of PLMs, where a human interrogator is requested to distinguish whether the given text is generated by a human. From the generated texts of each model and gold texts, we randomly select 500 texts scored by judges. More details of human test and Turing test are shown in Appendix F.

Results and Analysis. Table 5 and Table 6 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 is 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 in capturing 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 texts, 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 in improving PLMs' composition ability.

To further investigate what factors affect the pass rate of the Turing test, we deeply analyze the intermediate scoring results in the human test and Turing test. As shown in Figure 5, we calculate the pass rate of the Turing test for each human test metric across 1 to 5 scale. Moreover, we compute the Pearson correlation coefficient between the pass rate and each metric. In story generation (WritingPrompts), the coefficients for *Fluency*, *Informativeness*, and *Relevance* are 96.63, 97.93, 96.44, respectively. While, in text summarization (GigaWord), the coefficients for *Fluency*, *Informativeness*, and *Accuracy* are 96.08, 97.67, 98.38, respectively. From these analysis results, we can conclude that *Informativeness* is more important for story generation, while *Accuracy* is more influential in text summarization. Besides, we compute the text similarity between the generated texts from different PLMs, which is shown in Appendix F.

4 Discussion

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

In section 3.3, we observe that the improvement in memory ability is likely to be helpful for the performance of comprehension ability. Hence, designing PLMs with special objectives like bidirectional language modeling in BERT and strategies like larger training batches in RoBERTa for larger memory capacity will further benefit PLMs in the downstream comprehension tasks. Besides, when applying PLMs to downstream tasks, the similarity of data distribution between pretraining and fine-tuning has a great impact on PLMs performance. Possible solutions such as introducing intermediate tasks or datasets can alleviate such a discrepancy. Moreover, we further find some limitations in PLMs' comprehension ability, where PLMs are good at simple single-token answer types in QA such as dates but perform worse in complex phrase answers.

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

For composition ability, PLMs with denoising training objectives perform much better on short text composition, while PLMs with left-to-right objectives or larger model size are more suitable for long text composition. This might be because PLMs with different training objectives can finally

capture different ranges of semantic dependency between tokens and phrases. Moreover, to obtain a higher pass rate of Turing test, different text generation tasks will be concerned with varying factors, such as informativeness is much more critical for story generation.

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 corpora (Li et al., 2021c), leading to significant progress in various language tasks (Li et al., 2021a,b). With such encouraging results in extensive NLP tasks, it is a non-trivial topic 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 of 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 findings indicate that the features generated by PLMs are sufficient for good performance on a board set of tasks but fall short on tasks requiring fine-grained linguistic knowledge. Tenney et al. (2019) evaluate similar models on a range of sub-sentence linguistic analysis tasks, showing that PLMs encode both syntax and semantics into parameters. Zhou et al. (2020) also report that PLMs can learn rich knowledge but focus on evaluating the commonsense. However, these studies only look at 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 language abilities and performing extensive comparison.

6 Conclusion

This paper investigates the general language ability evaluation of pretrained language models. We design four kinds of language abilities of PLMs, including memory, comprehension, reasoning, and composition, and measure ten widely-used PLMs within five categories. For each language ability, we select multiple representative tasks to quantitatively evaluate the performance of PLMs. Our experimental results demonstrate that PLMs with varying objectives and strategies are good at different ability tests. Note that our final predicted outputs of PLMs can also be reused as an open resource for more depth and granularity in analyzing PLMs' language abilities. 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.

Acknowledgement

This work was partially supported by Beijing Natural Science Foundation under Grant No. 4222027, National Natural Science Foundation of China under Grant No. 61872369 and 82161148011, Beijing Outstanding Young Scientist Program under Grant No. BJJWZYJH012019100020098, and the Outstanding Innovative Talents Cultivation Funded Programs 2021 of Renmin University of China. We are grateful to Amazon Web Services for providing efficient GPU computing resource support and technical support for this NLP research project. Xin Zhao is the corresponding author.

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.

Robert Connors. 1997. *Composition-rhetoric: Backgrounds, theory, and pedagogy*. University of Pittsburgh Pre.

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.

Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2021a. Textbox: A unified, modularized, and extensible framework for text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 30–39.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021b. Few-shot knowledge graph-to-text generation with pretrained language models. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1558–1568. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021c. Pretrained language model for text generation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4492–4499. ijcai.org.

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 pre-trained 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*.

Supplementary Material for *ElitePLM*

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 7. The effect extent of each factor for PLMs abilities in Table 8.

B Data Statistics

Memory Tests. The data statistics of LAMA and Wikipedia of each model are presented in Table 9. 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 10.

Reasoning Tests. The data statistics for commonsense reasoning, deductive reasoning and abductive reasoning are presented in Table 11.

Composition Tests. The data statistics for text summarization, question generation and story generation are presented in Table 12. 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 13.

D Comprehension Tests

Full results on SuperGLUE, SQuAD and RACE are presented in Table 14 and Table 15.

E Reasoning Tests

Full results on CommonsenseQA, ROCStories, SWAG, HellaSwag, Sense Making, and ARCT are presented in Table 16.

F Composition Tests

For automatic metrics, BLEU-n and ROUGE-n 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. For the human test, *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.

In the human test, we ramdomly select 500 generated texts for each PLM and 500 gold text. Therefore, there are 3000 texts totally. The judges are all PhD students which do not know about where each text comes from. Each text will be scored by two judges from the above five aspects, and the final score is the average of the two scores. In the Turing test, each text will also be distinguished by two judges. Only when two judges make the same decisions that the text is generated by human, we will consider the text is true.

Full results on CNN/Daily-Mail, GigaWord, SQuAD, and WritingPrompts are presented in Table 17. Turing test results are presented in Table 6. We also show some summaries and stories generated by different PLMs in Table 19, Table 20, and Table 21.

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 7: Configurations and pretraining setting comparisons for our selected models.

Ability	MA	DD	MS	PO	PS
Memory	★★	★	★	★★★	★★★
Comprehension	★★	★★	★	★★	★★★
Reasoning	★	★★★	★	★★★	★★★
Composition	★	★★★	★★★	★★★	★

Table 8: The impact extent of each factor for four language abilities of PLMs. 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
#Relation	3	41	16	-	-
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
The Predicted Outputs	The predicted token of "`[MASK]`" in each template.				

Table 9: Statistics of datasets in memory tests, including LAMA and Wikipedia. #Origin and #Relation denote the number of examples and relations in original dataset, and the number of each model denotes the number of examples after selected. The predicted outputs is the intermediate result resources we provide.

Corpus		#Train	#Valid	#Test	The Predicted Outputs
GLUE	**CoLA**	8,551	1,043	1,063	The predicted binary class whether a sentence is grammatical.
	SST-2	67,349	872	1,821	The predicted sentiment (positive/negative) of a sentence.
	MRPC	3,668	408	1,725	The predicted binary class whether two sentences are semantically equivalent.
	QQP	363,846	40,430	390,965	
	STS-B	5,749	1,500	1,379	The predicted similarity score (1-5) of two sentences.
	MNLI-M.	392,702	9,815	9,796	The predicted relation (entailment/contradiction/neutral) between two sentences.
	MNLI-MM.		9,832	9,847	
	QNLI	104,743	5,463	5,463	The predicted relation (entailment or not) between two sentences.
	RTE	2,490	277	3,000	
	WNLI	635	71	146	
SuperGLUE	**BoolQ**	9,427	3,270	3,245	The predicted answer (yes/no) to the passage-based question.
	CB	250	57	250	The predicted relation (entailment/contradiction/neutral) between two sentences.
	COPA	400	100	500	The predicted the cause or effect of the premise from two choices.
	MultiRC	5,100	953	1,800	The predicted answer choice to the passage-based question.
	Wic	6,000	638	1,400	The predicted binary class whether a word is used with the same sense in two sentences .
	WNLI	635	71	146	The predicted relation (entailment or not) between two sentences.
	WSC	554	104	146	The predicted noun phrase referrent of the pronoun from among the provided choices.
SQuAD	**v1.1**	88,567	10,790	-	The predicted answer span to the passage-based question.
	v2.0	131,924	12,165	-	
RACE	**all**	25,137	1,389	1,407	The predicted answer choice to the passage-based question.
		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 10: 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. The predicted outputs is the intermediate result resources we provide.

Reasoning Task	Corpus	#Train	#Valid	#Test	The Predicted Outputs
Commonsense	**CQA**	9,741	1,221	1,140	The predicted answer choice to a commonsense question.
Deductive	**ROCStories**	1,257	314	1,571	The predicted ending choice based on the context.
	SWAG	73,546	20,006	20,005	The predicted answer choice based the grounded situation.
	HellaSwag	39,905	10,042	10,003	
	SM-A	10,000	1,000	1,000	The predicted valid sentence between two sentences.
Abductive	**SM-B**	10,000	1,000	1,000	The predicted reason choice why the sentence is invalid.
	ARCT	1,210	316	444	The predicted warrant choice that justifies the reason and claim.

Table 11: Statistics of datasets in reasoning tests, including commonsense reasoning, deductive reasoning and abductive reasoning. CQA is short for CommonsenseQA. SM-A and SM-B denote the Task A and Task B of Sense Making, respectively. The Predicted outputs is the intermediate result resources we provide.

Task	Corpus	#Train	#Valid	#Test	#Input	#Output	The Predicted Outputs
TS	**CNN/Daily Mail**	287,113	13,368	11,490	822.3	57.9	The generated summary given a news.
	Gigaword	3,803,957	189,651	1,951	33.7	8.7	The generated headline given a paragraph and corresponding Turing test and aspect scores.
QG	**SQuAD**	75,722	10,570	11,877	149.4	11.5	The generated question given a passage and corresponding answer.
SG	**WritingPrompts**	67,765	3,952	3,784	30.2	281.2	The generated story given a prompt and corresponding Turing test and aspect scores.

Table 12: Statistics of datasets in composition tests, including text summarization (TG), question generation (QG) and story generation (SG). #Input and #Output denote the average number of tokens in the input text and output text. The Predicted outputs is the intermediate results and human evaluation resources we provide.

Models	Vocab Size	LAMA-G	LAMA-T	LAMA-C	LAMA-S	Wikipedia	Average
BERT$_{\text{BASE}}$	28996	10.3	27.5	15.3	12.8	66.8	41.6
BERT$_{\text{LARGE}}$	28996	**11.0**	**29.2**	19.1	17.0	70.9	45.0
RoBERTa$_{\text{BASE}}$	50265	7.5	19.9	17.9	13.3	66.9	40.8
RoBERTa$_{\text{LARGE}}$	50265	7.1	23.9	**21.6**	**21.0**	71.1	44.8
ALBERT$_{\text{XLARGE}}$	30000	2.9	19.6	16.8	14.4	64.3	38.9
ALBERT$_{\text{XXLARGE}}$	30000	3.3	21.0	20.0	20.6	63.9	40.1
GPT-2$_{\text{SMALL}}$	50257	1.3	6.8	4.0	3.0	36.0	19.9
GPT-2$_{\text{MEDIUM}}$	50257	3.9	12.0	6.4	5.6	42.7	24.8
XLNet$_{\text{BASE}}$	32000	0.0	0.0	2.8	0.0	64.6	32.7
XLNet$_{\text{LARGE}}$	32000	0.0	0.0	5.5	0.4	68.7	35.1
UniLM$_{\text{BASE}}$	28996	8.5	27.6	15.4	11.8	66.9	41.4
UniLM$_{\text{LARGE}}$	28996	9.6	28.4	18.3	17.4	**71.5**	**46.4**
ERNIE$_{\text{BASE}}$	28996	1.3	13.4	13.0	8.1	-	-
T5$_{\text{BASE}}$	32100	5.5	20.0	13.2	9.6	60.5	36.3
T5$_{\text{LARGE}}$	32100	4.0	21.7	17.1	11.7	65.0	39.3
BART$_{\text{BASE}}$	50295	5.7	11.7	9.5	4.2	47.9	27.8
BART$_{\text{LARGE}}$	50295	9.4	15.8	7.7	3.1	47.8	28.4
ProphetNet$_{\text{LARGE}}$	30522	0.1	1.1	0.3	0.7	31.3	15.9

Table 13: 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	CB	RTE	COPA	Wic	BoolQ	MultiRC	Avg
	Acc.	F1/Acc.	Acc.	Acc.	Acc.	Acc.	F1/EM	
BERT$_{\text{BASE}}$	60.6	78.7/80.4	66.4	65.0	69.9	74.6	68.1/16.9	65.5
BERT$_{\text{LARGE}}$	63.5	89.0/92.9	70.1	73.0	72.7	75.6	69.4/22.6	70.3
RoBERTa$_{\text{BASE}}$	71.1	89.1/91.1	75.1	78.0	67.2	81.1	72.6/31.9	73.6
RoBERTa$_{\text{LARGE}}$	75.0	95.0/96.4	88.2	84.0	72.7	**85.4**	81.7/47.2	80.8
ALBERT$_{\text{XLARGE}}$	63.5	81.1/85.7	62.5	75.0	66.5	62.2	63.6/12.4	64.4
ALBERT$_{\text{XXLARGE}}$	64.4	87.6/92.9	70.4	**91.0**	**74.3**	62.2	**85.1/54.0**	74.6
GPT-2$_{\text{SMALL}}$	54.8	64.0/76.8	62.1	62.0	64.1	68.2	67.3/19.5	60.7
GPT-2$_{\text{MEDIUM}}$	61.5	84.4/82.1	63.6	63.0	67.2	73.9	71.5/29.2	66.1
XLNet$_{\text{BASE}}$	64.4	91.0/91.1	59.9	65.0	67.9	76.9	72.5/29.6	68.0
XLNet$_{\text{LARGE}}$	65.3	87.6/92.9	**88.5**	82.0	69.7	84.7	79.0/41.6	77.3
UniLM$_{\text{BASE}}$	63.5	74.7/82.1	60.3	67.0	68.5	73.3	67.9/20.5	65.0
UniLM$_{\text{LARGE}}$	65.4	86.5/87.5	70.9	76.0	72.3	82.3	75.7/36.3	72.8
ERNIE$_{\text{BASE}}$	65.4	81.6/82.1	68.8	64.0	70.8	74.4	68.7/21.3	67.2
T5$_{\text{BASE}}$	79.8	86.2/94.0	80.1	71.2	68.3	81.4	79.7/43.1	76.0
T5$_{\text{LARGE}}$	**84.6**	91.6/94.8	87.2	83.4	69.3	**85.4**	83.3/50.7	**81.4**
BART$_{\text{BASE}}$	64.4	86.6/85.7	69.5	70.0	65.7	75.7	74.2/31.7	69.2
BART$_{\text{LARGE}}$	65.4	**97.4/96.4**	83.5	86.0	70.4	85.1	82.9/50.6	79.2
ProphetNet$_{\text{LARGE}}$	63.5	94.7/92.9	51.3	61.0	60.7	67.4	64.7/17.2	62.7

Table 14: 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
$\text{RoBERTa}_{\text{LARGE}}$	88.9	94.6	86.5	89.4	83.2	86.5	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**	85.4	**88.6**	**84.0**
$\text{UniLM}_{\text{BASE}}$	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	94.6	86.1	89.2	82.2	82.5	79.6
$\text{ProphetNet}_{\text{LARGE}}$	-	-	-	-	-	74.1	-

Table 15: 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 16: 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
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
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
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
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$_{\text{BASE}}$	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$_{\text{LARGE}}$	42.50	20.68	39.75	34.75	16.26	31.49	11.19	22.35	30.53	8.61	4.19	9.51
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
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
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 17: 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
GPT-2$_{\text{MEDIUM}}$	**45.7**	**3.42**	**3.17**	3.20	**3.23**	5.87
UniLM$_{\text{LARGE}}$	1.2	1.32	1.88	2.03	1.71	2.74
T5$_{\text{LARGE}}$	34.4	3.01	2.80	3.09	2.87	5.18
BART$_{\text{LARGE}}$	45.2	3.37	3.16	**3.39**	3.22	**5.96**
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 18: 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$_{\text{MEDIUM}}$	unk computer to jointly sell supercomputers
UniLM$_{\text{LARGE}}$	nec [UNK] to join forces in supercomputer sales
T5$_{\text{LARGE}}$	nc and unk computer to join forces in supercomputer sales
BART$_{\text{LARGE}}$	nec and unk computer to join forces in supercomputer sales
ProphetNet$_{\text{LARGE}}$	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$_{\text{MEDIUM}}$	closed as military campaign escalates
UniLM$_{\text{LARGE}}$	sri lankan gov t shuts down schools in north
T5$_{\text{LARGE}}$	sri lanka closes schools as fighting escalates
BART$_{\text{LARGE}}$	sri lanka closes government schools
ProphetNet$_{\text{LARGE}}$	sri lanka closes schools as military escalates
Gold	sri lanka closes schools as war escalates

Table 19: 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 20: 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 21: Test set examples of WritingPrompts dataset for each model.