Zero-Shot Learners for Natural Language Understanding via a Unified Multiple Choice Perspective

Ping Yang[1*] **Junjie Wang**[2*] **Ruyi Gan**[1] **Xinyu Zhu**[3] **Lin Zhang**[1]
Ziwei Wu[1] **Xinyu Gao**[1] **Jiaxing Zhang**[1] **Tetsuya Sakai**[2†]
[1]International Digital Economy Academy [2]Waseda University [3]Tsinghua University
{yangping, ganruyi, zhanglin, wuziwei, gaoxinyu, zhangjiaxing}@idea.edu.cn
wjj1020181822@toki.waseda.jp tetsuyasakai@acm.org zhuxy21@mails.tsinghua.edu.cn

Abstract

We propose a new paradigm for zero-shot learners that is format agnostic, i.e., it is compatible with any format and applicable to a list of language tasks, such as text classification, commonsense reasoning, coreference resolution, and sentiment analysis. Zero-shot learning aims to train a model on a given task such that it can address new learning tasks without any additional training. Our approach converts zero-shot learning into multiple-choice tasks, avoiding problems in commonly used large-scale generative models such as FLAN. It not only adds generalization ability to models but also significantly reduces the number of parameters. Our method shares the merits of efficient training and deployment. Our approach shows state-of-the-art performance on several benchmarks and produces satisfactory results on tasks such as natural language inference and text classification. Our model achieves this success with only 235M parameters, which is substantially smaller than state-of-the-art models with billions of parameters. The code and pre-trained models are available at https://github.com/IDEA-CCNL/Fengshenbang-LM/tree/main/fengshen/examples/unimc.

1 Introduction

Remarkable advances in large-scale language models have brought substantial improvements in a wide variety of tasks such as text classification, natural language inference and commonsense reasoning (Brown et al., 2020; Chowdhery et al., 2022). This progress brings opportunities to Zero-Shot Learning (ZSL) (Sanh et al., 2021; Chowdhery et al., 2022), which aims to predict labels on datasets from novel domains. Most solutions can be framed in the prompt tuning framework that activate specific parameters in PLM (Xu et al., 2022; Liu et al., 2021) to adapt to zero-shot tasks. A powerful variant of prompt tuning is called instruction

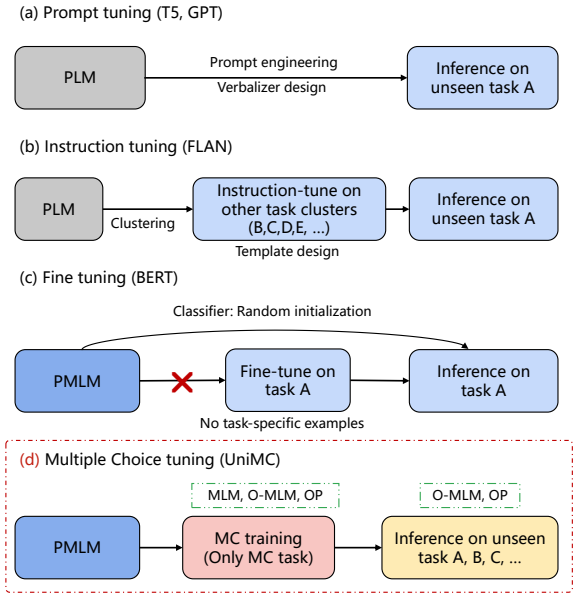

Figure 1: Typical zero-shot learning methods and our proposed UniMC. "PLM" indicates pre-trained language model. "PMLM" implies pre-trained masked language model.

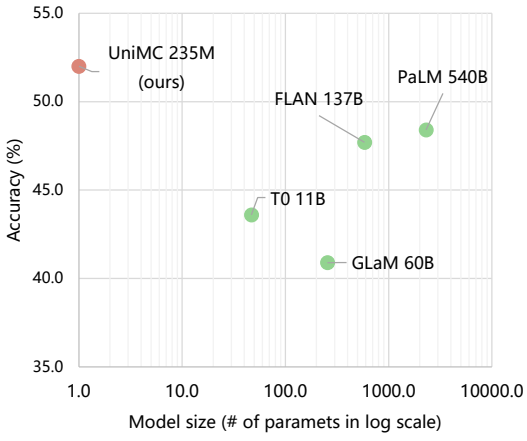

Figure 2: Zero-shot performance comparison in ALNI R1. Our proposed UniMC has the best performance w.r.t the accuracy and the model size, simultaneously.

*Equal contribution.
†Corresponding Author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7042–7055
December 7-11, 2022

tuning (Wei et al., 2021), which shares knowledge from different domains. We summarize the mainstream large-scale frameworks in Figure 1.

Despite their success, these frameworks suffer from their inherent problems, and thus limit their potential in zero-shot learners. Firstly, prompt-related models have an extremely large number of parameters, e.g., GPT-3 has 175B, FLAN has 137B and PaLM (Chowdhery et al., 2022) has 540B. One immediate problem is that these models are often hard to be trained, making the deployment and consumption difficult. Secondly, manual processing is required when addressing zero-shot problems. For instance, T0 builds $2,073$ prompts to handle more than 170 tasks (Sanh et al., 2021). Lastly, existing models employ a single direction paradigm, either auto-regressive models or sequence-to-sequence, resulting in inadequate usage of information from both directions. As an example, PMLM tries to implement a zero-shot learner, which is shown in Figure 1 (c). Note that recent work (Liu et al., 2019a) state that PMLM is more suitable than PLM for Natural Language Understanding (NLU) tasks. However, it has to be fine-tuned on the task-specific samples to initialize the classifier instead of randomly initializing the classifier. Therefore, the ability of PMLM is limited when dealing with zero-shot scenarios.

To address the aforementioned problems, we introduce a light-weight framework, called **Uni**fied **M**ultiple **C**hoice model (UniMC), proposing a novel MC tuning. The proposed MC tuning has the following advantages: i) parameter updating only happens in the MC training phase, and ii) facilitating the deployment. To reduce the manual processing, we only formulate one candidate option prompt format and one question prompt format. Note that we also consider the case without any question prompt format. Under this setting, we can treat labels as options rather than building verbalizer maps and providing its text information to the models as before. We therefore can learn the information from labels directly. To this end, we convert the problematic classifiers to options. One immediate question is how to choose an option efficiently and unambiguously. Therefore, as shown in Section 3.2, we develop an option-mask tokens `[O-MASK]` to predict "yes" or "no" before each option. A two-step process is introduced to output the desired options. First, similar to Masked Language Modeling (MLM) (Devlin et al., 2019), we conduct Option MLM (O-MLM) to recover the "yes" or "no" for each option. Next, we propose an Option Prediction (OP) method to compute proper options.

With extensive experiments on multiple challenging benchmarks, we demonstrate that our approach's performance outperforms state-of-the-art baselines, while reducing the model size with two orders, as shown in Figure 2. This success suggests the potential of leveraging UniMC in large datasets. Our contributions are as follows.

- We propose a new zero-shot paradigm by converting this problem into multiple choice tasks.
- We develop an effective and efficient method to implement a MC-based zero-shot learner. Our proposed method has up to 48% improvement on Dbpedia over SOTA baselines that have a few hundred times larger than our model.

2 Related Work

2.1 Unified NLP Task Formats

NLP tasks often have diverse formats due to the fast emergence of datasets, such as machine reading comprehension and text classification tasks. Recent research shows the necessity of unifying formats to fix the gap across various tasks (Sanh et al., 2021; Wei et al., 2021; Sun et al., 2021). By developing a natural language prompted form, T0 builds an application to map original NLP datasets into target templates with custom prompts (Sanh et al., 2021). FLAN groups multiple datasets into 12 task clusters, and then designs 10 unique instruction templates to unify formats (Wei et al., 2021). Despite effective, this focuses on generative styles and thus cannot be adapted to vast label-based models that select. This motivates us to unify label-based tasks, where we develop unified Multiple Choice (MC) formats for this purpose.

2.2 Label Information

The label semantic is an important information source, such as in few-shot tasks (Hou et al., 2020; Mueller et al., 2022; Luo et al., 2021). The L-TapNet framework (Hou et al., 2020) integrates the label information with manually designed prompts for inputs to solve few-shot slot tagging tasks. In addition, LSAP (Mueller et al., 2022) obtains powerful few-shot performance by introducing label semantics into the pre-training and fine-tuning phases

of the PLMs. Together, these successful employments of labels in low-resource settings inspire us to bring label semantics to our unified MC inputs to handle the zero-shot scenario.

2.3 Zero-Shot Learning

Large-scale Pre-trained Language Models (PLMs) with billions of parameters such as GPT-3 (Brown et al., 2020) have shown impressive performance across various few-shot tasks. However, they have limited competence when dealing with zero-shot tasks, which have broader applications in practice. Recent efforts try to mitigate this issue from different perspectives. FLAN (Wei et al., 2021) designs specific instruction templates for each task and utilizes over 60 labeled datasets to "fine-tune" a 137B language model. T0 (Sanh et al., 2021) unifies all tasks into a source-target format by collecting a large variety of prompt templates, specifically $2,073$ manually constructed prompts, and trains the model with multi-task learning. Along this line, ZeroPrompt (Xu et al., 2022) applies over $1,000$ supervised datasets and proposes the genetic prompt search method to find prompts for new tasks. However, these methods cost significant laborious efforts, such as prompt engineering and template designing. Moreover, the pre-training and tuning phases of large-scale PLMs take enormous amounts of computational resources, therefore, new tasks may suffer great difficulty in deploying. As a comparison, our proposed UniMC is light-weighted, i.e., has 235M parameters and a few manual input text transformations, making it suitable for more general scenarios.

3 Approaches

In this section, we outline the proposed framework, i.e., UniMC, and provide the training and inference approaches in detail.

3.1 The UniMC framework

3.1.1 Unified Input

A unified input format will facilitate the generalization of models, promoting the sharing of knowledge across different tasks. To achieve this, we frame all tasks' objectives together as a multiple-choice (MC) problem, as shown in Figure 3. A MC problem often consists of three components, including options, question, and passage. We now discuss the details of getting these bodies. We can often get the passage component effortlessly be-

NLI	Dataset	RTE
	Passage	The abode of the Greek gods was on the summit of Mount Olympus, in Thessaly.
	Question	Based on the paragraph
	Option	**[1] we can infer that Mount Olympus is in Thessaly.** [2] we can not infer that European cars sell in Russia.
Common sense	Dataset	Hellaswag
	Passage	A graphic introduces the hand car wash video. The car is washed first gently with soap. next
	Question	
	Option	[1] washes persons hands and wipes them with a blue cloth. **[2] is washed first gently with soap.** [3] washes game is displayed. [4] washes and a man wearing a blue shirt speaks to the camera.
Sentiment	Dataset	SST-2
	Passage	It's a cookie-cutter movie , a cut-and-paste job
	Question	What is sentiment of the review?
	Option	[1] it's great. **[2] it's terrible.**
Conference	Dataset	Winogrande
	Passage	I had to read an entire story for class tomorrow. Luckily, the
	Question	
	Option	**[1] story was short.** [2] class was short.
Classification	Dataset	Dbpedia
	Passage	Outright is a US accounting and bookkeeping application that assists small businesses and sole proprietors with managing their business income and expenses. It also provides them with a means to organize and categorize expenses for filing a Schedule C.
	Question	What is topic of the articles?
	Option	**[1] Company** [2] Educational Institution [3] Artist [4] Athlete [5] Office Holder [6] Mean Of Transportation [7] Building [8] Natural Place [9] Village [10] Animal [11] Plant [12] Album [13] Written Work

Figure 3: Unified input text examples with sampling from datasets in zero-shot phase. The prompt text is underlined and the correct options are in **bold**.

cause it often exists in the original data. As to the question part, we can either use the raw question directly or provide a corresponding question when it is missing. The transformation of options depends on whether or not we can get a straightforward expression of classes. On the one hand, we can convert all classification tasks into options directly as it has specific information for choices. On the other hand, we have to construct an option prompt to generate particular choices. Details of this transformation can be found in Appendix A. In effect, these allow us to abandon label indices as in classification tasks, which include much less information than our used options.

3.1.2 Network

In our framework, we employ BERT-like PMLMs as the backbone, such as ALBERT (Lan et al., 2020) and RoBERTa (Liu et al., 2019b), to integrate the bidirectional modeled input x_{inp}. In additional, the discussion of backbone models is in Appendix B. Instead of using the original embedding methods directly, we develop a new solution for the segment id, position id, and attention mask matrix to fit multiple choice tasks, simultaneously.

Tokenization: In this framework, the key to achieve the ability of addressing MC tasks is to set up a proper option. We thus introduce option-mask

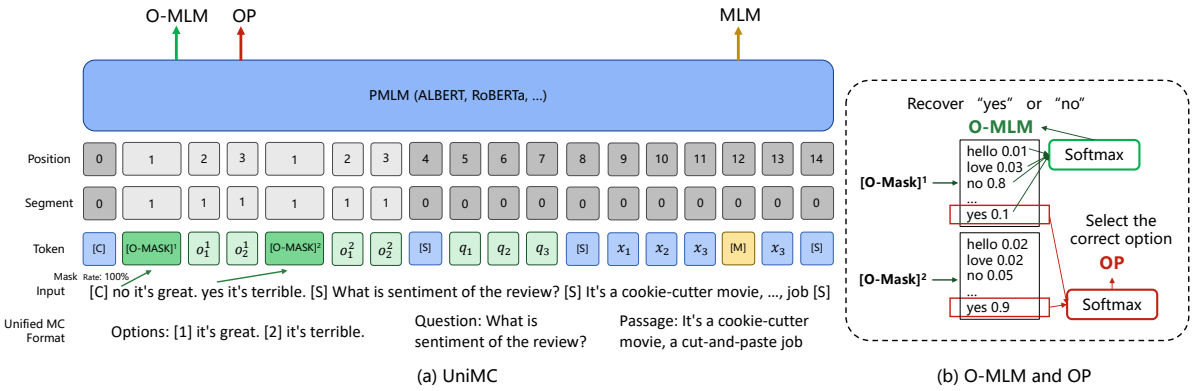

Figure 4: UniMC framework with O-MLM and OP in MC training phase. "[O-MASK]^1" in (b) indicates the option mask token of option 1. Similarly, "[O-MASK]^2" is related to option 2. [C], [S] and [M] are the abbreviation of [CLS], [SEP] and [MASK]. The example of input text is from the dataset SST-2 (Socher et al., 2013).

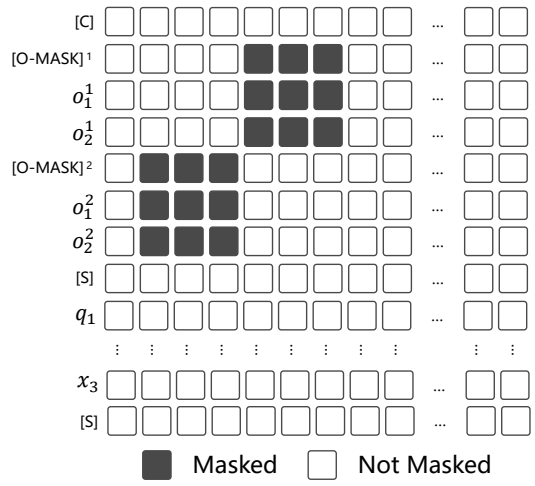

Figure 5: Self-Attention Mask Matrix. Given input $[\mathtt{C}], [\mathtt{O-MASK}], o_1^1, o_2^1, \ldots, x_3, [\mathtt{S}]$, the tokens of options can not attend to each other.

tokens ([O-MASK]), aiming to replace "yes" or "no" in the input text for a better representation ability. Here, [O-MASK] inherits the ability of [MASK], and thus remains to use token predictions to determine which option is correct. Consider, as an example, an input set, denoted as (o, q, x), includes the following: i) one passage $x = x_1 \ldots x_{|x|}$, ii) N_Q questions $q = q_1 \ldots q_{|q|}$, and iii) N_O candidate options $o = o_1 \ldots o_{|o|}$, whose input token x_{inp} is formulated as follows:

$$\begin{aligned} x_{inp} =& [\text{CLS}]\{[\text{O-MASK}]^i\ o^i\}_{i=1}^{N_O}[\text{SEP}] \\ & \{q\ [\text{SEP}]\}^{N_Q}\ x\ [\text{SEP}], \end{aligned} \tag{1}$$

Here, $N_Q \in \{0, 1\}$, $N_O \in \mathbb{N}+$ and $N_O \geq 2$.

Id embeddings and attention mask matrix: Note that a unified input text has multiple options, leading to undesired mutual influence between options and resulting in a misunderstanding of answers. We now address this issue from the following three perspectives, including segment id, position id, and attention mask matrix. Firstly, we assign segment id to distinguish option and context (questions, passages) information, as shown in Fig. 4 (a). Secondly, we update the position id to tell apart the intra information in the option. This is because that PMLMs cannot get position information from tokens. We aim to allow PMLMs will treat tokens' position information based on their position embeddings. Lastly, we will control the flow between options, such as M_{mask} in self-attention, as shown in Fig. 5. In particular, black squares are used to mask a part of the input attention matrix, ensuring the disentanglement between different options. We place a $-inf$ number on the masked slots, which is the same as BERT to mask tokens.

Furthermore, we can have the encoded hidden vector, denoted as $T = [T_1 \ldots T_n]$, using multiple Transformer-based layers as following,

$$T = \text{encoder}(x_{inp}, pos, seg, M_{mask}). \tag{2}$$

3.2 MC tuning

Recall the backbones are often pre-trained models, resulting in excellent skill in capturing the commonsense knowledge. Intuitively, we can employ these as base modules by taking advantage of their high volume knowledge. More specifically, we use the outputs of pre-trained models as the initial states for the following MC tasks, leading to a two-stage tuning paradigm. In the MC training phase, we train the models with MC tasks and gain a great initialization for selecting a correct option. In the zero-shot phase, we apply the unified MC models to unseen zero-shot tasks.

3.2.1 MC training phase

We now introduce the proposed option masked language modeling (O-MLM) and option prediction (OP) methods in detail.

Masked Language Modeling (MLM) is a pre-training task in BERT (Devlin et al., 2019) for self-supervised learning,

$$\mathcal{L}_{\text{MLM}} = -\sum_{\hat{T} \in m(T)} \log p\left(\hat{T} \mid T_{\backslash m(T)}\right), \tag{3}$$

where $\hat{T}$ is the random perturbed token from T; $m(T)$ and $T_{\backslash m(T)}$ are the masked tokens from T and the reset tokens, respectively. In practice, we randomly replace tokens in the passage sequence x with special tokens `[MASK]`, as opposed to the whole sequences used in standard BERT. The main difference between O-MLM and MLM is the way of masking. We always mask the `[O-MASK]` tokens to predict "yes" or "no", as shown in Figure 4 (b). Therefore, the loss $\mathcal{L}_{\text{O-MLM}}$ and $\mathcal{L}_{\text{MLM}}$ share the same style.

Once the prediction probabilities of "yes" or "no" is obtained, we next introduce the OP to teach the model for learning MC tasks, which is shown in Figure 4 (b). To learn the mutually exclusive characteristics between options, OP takes the logits $T^{yes}_{\texttt{[O-MASK]}} \in \{T^{yes}_{\texttt{[O-MASK]}^1}, \ldots, T^{yes}_{\texttt{[O-MASK]}^{N_O}}\}$ in "yes" for each option sequence to generate label distributions. OP aims to compute a cross-entropy loss with ground truth label distribution Y:

$$\mathcal{L}_{\text{OP}} = -\sum_{i=1}^{N_O} Y_i \log \text{Softmax}\left(T^{yes}_{\texttt{[O-MASK]}}\right) \tag{4}$$

Recent studies show that including mixed tasks in a batch will improve the generalization ability of neural networks (Aghajanyan et al., 2021). When facing mixed tasks, we mask the output logits except for `[O-MASK]` during the Softmax operation to compute the OP loss in a mini-batch, as shown in Figure 6. The logit masking approach allows our UniMC to handle MC tasks with different number of options in a single batch.

In summary, the overall training objective in MC training is given by:

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{O-MLM}} + \mathcal{L}_{\text{OP}} \tag{5}$$

3.2.2 Zero-shot phase

After obtaining a unified MC model, we simply utilize O-MLM and OP to predict the answer in unseen zero-shot datasets. We know that the ground truth labels are missing, so it is impossible to compute the loss. Alternatively, we can compute the most confident option with the OP because the model still recover `[O-MASK]` to "yes" or "no" with O-MLM.

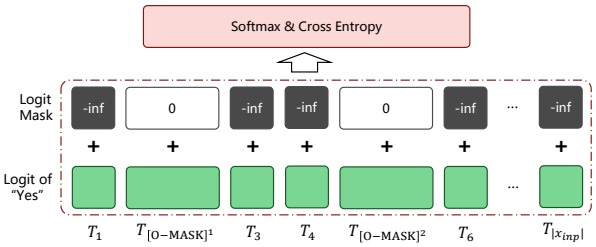

Figure 6: Applying logit masking method in OP. $-inf$ means negative infinity.

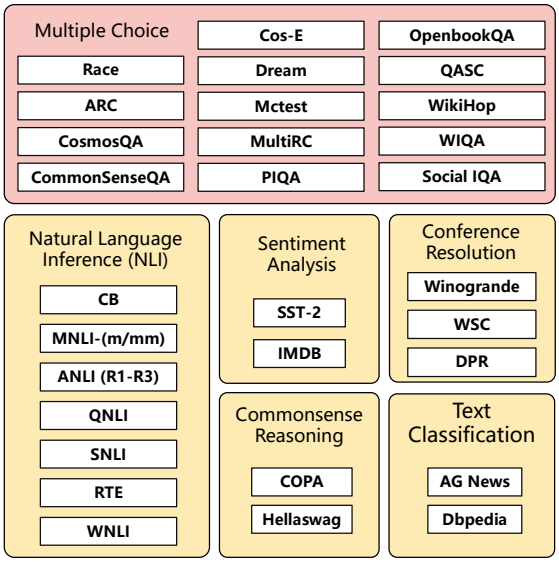

Figure 7: Datasets with various types of tasks. Datasets in MC training phase are in red (above). Datasets in zero-shot phases are in yellow (below).

3.2.3 Discussion

Interestingly, we realize that the MC training stage and zero-shot stage are consistent in processing objectives. Recall that previous models tend to have divergence learning objectives, which may cause potential oscillation. Our proposed method is more task-driven and thus has a better chance to deliver high learning quality in task-specific outputs.

4 Experiments

4.1 Experimental Setup

We follow the preparation in T0 (Sanh et al., 2021) to cluster the label-based NLP datasets into 6 groups. In particular, we collect publicly avail-

Model	T0 11B	GLaM 60B	FLAN 137B	PaLM 540B	UniMC 235M
Parameters	×46.8	×255.3	×583.0	×2297.9	×1.0
ANLI R1	43.6	40.9	47.7	48.4	**52.0**
ANLI R2	38.7	38.2	43.9	44.2	**44.4**
ANLI R3	41.3	40.9	47.0	45.7	**47.8**
CB	70.1	33.9	64.1	51.8	**75.7**

Table 1: Zero-shot results in natural language inference task. The best scores are in **bold**.

Dataset	GPT2	GPT3*	UniMC
AG News	68.3	73.9	**81.3**
Dbpedia	52.5	59.7	**88.9**

Table 2: Zero-shot results in text classification task. The best results are in **bold**.

able NLP datasets on HuggingFace[1], and assign each label-based dataset to one of the task groups, as shown in Fig. 7. For each group, we design a corresponding transformation rule to convert it into a unified MC format, where detailed examples are presented in Sec. 3.1.1. Please refer to Appendix A for more details of dataset descriptions and unified MC formats. Next, we split the whole datasets into two parts for the two phases in our framework, i.e., the part for MC task is for the training, and the other is for the zero-short scenarios. It is worthy mentioning that using the MC tasks only in the MC training phase can avoid intensive resource computing.

Following the general setting (Du et al., 2021; Wei et al., 2021), we apply accuracy in all datasets. For computing the overall average accuracy, we take the average accuracy for each task and then calculate the arithmetic mean for them.

4.1.1 Baselines

In the experiments, we compare our method with the state-of-the-art baselines, including: GPT2 (Radford et al., 2019), GPT3* (Zhao et al., 2021), T0 (Sanh et al., 2021), FLAN (Wei et al., 2021), PaLM (Chowdhery et al., 2022), GaLM (Du et al., 2021) and UnifiedQA (Khashabi et al., 2020). We report the accuracy of each method to measure their performance. We only present the average outcomes if the baseline is conducted in multiple runs. Besides, we include the random guessing as a naive baseline for the comparison.

[1] https://huggingface.co/datasets

4.1.2 Implementation Details

In our model, we use the ALBERT-xxlarge-V2 (Lan et al., 2020) as backbone models by taking its light-weighted parameters. For fair comparison, we set the maximum length token as 512 in all experiments as in (Lan et al., 2020). In the training, we run only one epoch by following the setting in FLAN (Wei et al., 2021). We set the number of samples for each task up to $20K$, aiming to prevent the model from being dominated by specific tasks. Besides, we repeat the experiment 5 times by using different seeds. We run all our experiments on 8 NVIDA A100 GPUs.

4.2 Main Results

4.2.1 Natural Language Inference

We now present our main results from the Natural Language Inference (NLI) task in Table 1. UniMC achieves the best performance in all datasets, demonstrating its capability of NLI. In particular, UniMC achieves these competitive results with as few as 235M parameters as opposed to hundred billions of parameters in other baselines. These results confirm the effectiveness of unifying formats as a multiple choice style. Besides, a bi-directional structure in UniMC strengths its ability in capturing information as opposed to the previous one-directional structures.

4.2.2 Text classification

Text classification task aims to select a label/class for given texts. This is similar to the objective of MC task in nature. Therefore, we conduct a zero-shot text classification experiment to verify our model's capability. As shown in Table 2, UniMC outperforms previous SOTA models by a large margin. In particular, we know that Dbpedia includes 13 categories, adding a significant challenge to the classification task. Fortunately, UniMC has a built-in advantage in dealing with multiple classes due to the similarity between choices and classes, leading up to 48.9% improvement.

Dataset	FLAN	UniMC	Dataset	FLAN	UniMC
NLI			**Commonsense**		
ANLI R1	47.7	**52.0**	COPA	90.6	**95.2**
ANLI R2	43.9	**44.4**	Hellaswag	56.4	**62.5**
ANLI R3	47.0	**47.8**	**Coreference**		
CB	64.1	**75.7**	Winogrande	**67.3**	65.8
RTE	**78.3**	78.1	WSC	**80.8**	78.8
QNLI	**59.6**	54.0	DPR	60.3	**87.5**
SNLI	43.0	**60.9**	**Sentiment**		
MNLI-m	51.1	**52.7**	SST-2	**92.6**	91.6
MNLI-mm	51.0	**51.4**	IMDB	94.1	**94.8**
WNLI	61.0	**65.4**			

Table 3: A summary on natural language inference, commonsense reasoning, coreference resolution and sentiment analysis task.

4.2.3 A comprehensive comparison to FLAN

We know that FLAN is a well-known model in dealing with zero-shot option or label-related tasks. One of its particular merits is the zero-shot generalization ability. To better demonstrate the ability of UniMC, we report a comprehensive comparison between ours and FLAN, as shown in Table 3 and more comparisons are described in Appendix B.3. In the NLI task, UniMC achieves better performance than FLAN in general, which is consistent with the results in Table 1. We also select tasks like the commonsense reasoning, the coreference resolution, and the sentiment analysis to further explore the generalization ability of ours. UniMC gets an obvious advantage in COPA, Hellaswag, Winogrande, WSC, DPR when evaluating the common sense and coreference tasks. Beyond these two tasks, we find that the construction of datasets plays a critical role to the performance. In general, these datasets can be grouped into two categories: the understanding and generation styles. UniMC tends to show better performance on datasets that more close to the understanding style. In sentiment tasks, the number of classes is limited, making the dataset construction style is less important than that in the tasks of the common sense and coreference. Therefore, both UniMC and FLAN get relative good performance.

4.3 Ablation Studies

In this section, we intend to verify the necessity of key components of our UniMC, including the MC training, the prompt effect, the flow controlling. We also show the influence of the model size.

Task	Random Guess	UniMC*	UniMC
NLI	38.3	38.1	**58.2**
Commonsense	37.5	43.2	**78.9**
Sentiment	50.0	40.0	**93.2**
Coreference	50.0	54.8	**77.4**
Classification	16.1	15.9	**85.1**
Average	38.4	38.4	**78.6**

Table 4: MC training improves UniMC zero-shot performance. "UniMC*" indicates the UniMC without the MC training stage.

Dataset	with Question	w/o Question
NLI		
ANLI R1	47.5	**52.0**
ANLI R2	43.2	**44.4**
ANLI R3	46.4	**47.8**
QNLI	52.2	**54.0**
RTE	74.3	**78.1**
WNLI	59.4	**65.4**
MNLI-m	**52.7**	48.8
MNLI-mm	**51.4**	47.5
CB	**75.7**	70.7
SNLI	**60.9**	53.7
Sentiment		
SST-2	**91.6**	90.2
IMDB	**94.8**	93.6
Classification		
AG News	81.2	**81.3**
Dbpedia	60.1	**88.9**

Table 5: We report results of UniMC with and without question prompts. We present 3 tasks (NLI, Sentiment, Classification) because question prompts are not designed in other tasks.

4.3.1 How important is MC training?

Recall that our proposed UniMC takes advantage of O-MLM and OP to evaluate zero-shot tasks without MC training. To better understand our design, we develop a variant of our model that omits the MC training, named as UniMC*. In Table 4, we present the results of UniMC*, where its performance is close to "Random Guess". This striking outcome verifies the necessity of MC training.

4.3.2 How does the prompt affect the performance?

Our framework intends to reduce the effort of designing prompts, we now analyze what the effect of particular prompts, including the question prompts and the option prompts. We present the results in the Table 5.

For the question prompts, we conduct experiments on four challenge tasks by showing perfor-

Dataset	Good / Bad	Great / Terrible	Positive / Negative	Average	Std
Model: UnifiedQA-T5-3B					
SST-2	71.0	83.4	91.2	81.8	8.3
IMDB	85.4	90.3	90.6	88.8	2.4
Model: UniMC-235M					
SST-2	90.9	91.6	91.1	**91.2**	0.3
IMDB	94.3	94.8	93.7	**94.2**	0.4

Table 6: Zero-shot results in sentiment analysis task. "Std" indicates Standard Deviation. The best average results are in **bold**. The more stable performance is underlined.

Method	Average	Improve
Random Guessing	38.4	+0.0
Only UIE	39.1	+0.7
Only AMM	78.0	+39.6
UIE + AMM	**78.6**	+40.2

Table 7: Zero-shot performance with different strategies to control the flow between options. "UIE" indicates Updating Id Embeddings, including segment id and position id. "AMM" means Attention Mask Matrix. "Improve" shows the accuracy improvement from Random Guessing.

mance of using prompts or not. Although the performance for all tasks shows different directions, we hypothesize that this divergence is caused by the way of data construction. These datasets are mainly designed for two purposes, which are the language modeling task and the relationship choice task (Brown et al., 2020). The desire for question prompts increases when the data is more close to the language modeling task; vice versa. Furthermore, we classify these datasets into two categories, spoken-based and written-based, according to the definition in (Alsaawi, 2019). MNLI-m/mm, CB, SNLI, SST-2 and IMDB belong to the spoken-based corpus, while the rest datasets belong to written-based corpus. Considering that PMLM is usually pre-trained on written-based corpus, e.g., the pre-training datasets of BERT are Wikipeida and BookCorpus (Devlin et al., 2019), ours may have no need of questions for written-based data. This, again, confirms that data construction affects the requirements of question prompts.

For the option prompts, we present the experimental results in Table 6. We would like to emphasize that option prompts are necessary for our UniMC, therefore, we cannot remove this component as in the above experiment. Instead, we design different option prompts to demonstrate their

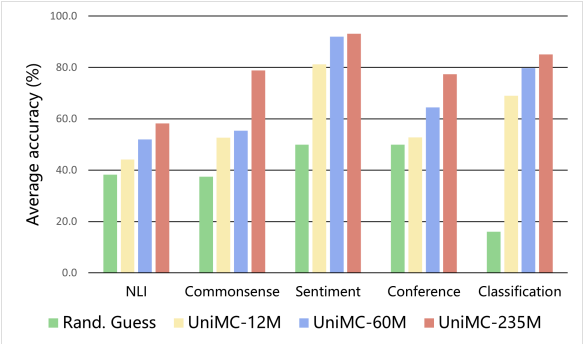

Figure 8: Zero-shot performances on several tasks with model variants.

effects. We observe that different prompts show limited performance variations, indicating the robustness of our UniMC to option prompts. Since FLAN and PaLM are not open-sourced, we choose one of the most powerful models, e.g., UnifiedQA-T5, as the baseline to ensure the fairness in comparison. In the experiment, we find that UnifiedQA-T5 is sensitive to option prompts, which have up to 8.3 standard variation (Std).

4.3.3 How does the flow controlling affect the performance?

We design the prompt to frame the input sequences to make all datasets fit into UniMC directly. However, some recent methods need extra processes, such as adopting an option with a context (question and passage) into a sequence and aggregate multiple different sequences to get an answer (Sun et al., 2021). To fix this gap, we design two strategies to control the flow of the information as in Section 3.1.2. We summarize the performance of these two in Table 7. We observe that AMM adds the greatest improvement to results, which is much better than UIE. On the one hand, UniMC can learn the position relationship between options. On the other hand, UniMC can distinguish between options and context. However, UIE is unable to prevent the inter-influence in options. Thanks to self-attention mechanism, AMM makes the options unreachable to each other, eliminating the intra-information of options.

4.3.4 How does the model size affect the performance?

A common intuition from this domain is that a large model size will result a better performance (Wei et al., 2021; Chowdhery et al., 2022), particular large-scale PLMs. Naturally, we believe that our backbone PMLM follows this rule as well. To vali-

date this, we implement an experiment by varying the model size, as shown in Figure 8. All 4 different tasks show the same trend, demonstrating the correctness of the mentioned intuition.

5 Conclusions

In this paper, we introduce a new zero-shot paradigm called MC tuning. This adds flexibility and generalization ability to zero-shot learners. We propose O-MLM and OP in both MC training and zero-shot phase, aiming to capture information from both directions. Our UniMC achieves better performances over SOTA models that a few hundred times larger than our model. Our experiments demonstrate the effectiveness and generalization ability of UniMC on zero-shot tasks. In future work, we will extend UniMC to few-shot scenarios.

Limitations

In this paper, our main contribution is a simple and effective framework for zero-shot tasks while maintaining a light weight. We aim to introduce additional artificial information and reduce manual processing to the minimum. We explored how to employ question prompts in Sec. 4.3.2, however, it is non-trivial to decide whether a prompt is required for complex datasets. In addition, we only compare with limited baselines when understanding the influence from the backbone in UniMC. In experiments, we implement only a few comparative experiments between ALBERT and RoBERTa (Liu et al., 2019b) due to the limit of computational resources, as shown in Appendix B.2. In the future, we will dig deeper into the principles regarding inputs and backbone, etc.

Ethical Considerations

Natural language processing is an important technology in our society. It is necessary to discuss its ethical influence (Leidner and Plachouras, 2017). In this work, we develop a novel zero-shot NLP approach to enhance the generalization ability of NLP. As discussed in (Schramowski et al., 2022, 2019; Blodgett et al., 2020), language models might contain human-like biases, which might embed in both the parameters of the models and outputs. Furthermore, we note the potential abuse of zero-shot models because these are often being integrated into applications without justification. We encourage open debating on its utilization, such as the task selection and the deployment, hoping to reduce the chance of any misconduct.

Acknowledgements

This research was supported by the open-source project, Fengshenbang (Wang et al., 2022). We would like to thank members of The Real Sakai Laboratory[2] and of the GTS team in IDEA[3], for giving us suggestions. Junjie Wang is especially grateful to our friend Yuxiang Zhang for his support, advice, and encouragement.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5799–5811. Association for Computational Linguistics.

Ali Alsaawi. 2019. Spoken and written language as medium of communication: A self-reflection. *International Journal of Applied Linguistics and English Literature*.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*. NIST.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *AAAI*, pages 7432–7439. AAAI Press.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *ACL*, pages 5454–5476. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[2] http://sakailab.com/english/

[3] https://www.idea.edu.cn/

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2021. Glam: Efficient scaling of language models with mixture-of-experts. *CoRR*, abs/2112.06905.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007a. The third PASCAL recognizing textual entailment challenge. In *ACL-PASCAL@ACL*, pages 1–9. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007b. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *ACL*, pages 1381–1393. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *EMNLP/IJCNLP (1)*, pages 2391–2401. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *AAAI*, pages 8082–8090. AAAI Press.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*, 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.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *EthNLP@EACL*, pages 30–40. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012a. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012b. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *ACL (1)*, pages 4487–4496. 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. *CoRR*, abs/1907.11692.

Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. 2021. Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2773–2782. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*, pages 2381–2391. Association for Computational Linguistics.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. 2022. Label semantic aware pre-training for few-shot text classification. In *ACL (1)*, pages 8318–8334. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 777–789. ACL.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *ACL (1)*, pages 4932–4942. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, pages 193–203. ACL.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. 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 8732–8740. AAAI Press.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Patrick Schramowski, Cigdem Turan, Sophie F. Jentzsch, Constantin A. Rothkopf, and Kristian Kersting. 2019. BERT has a moral compass: Improvements of ethical and moral values of machines. *CoRR*, abs/1912.05238.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Trans. Assoc. Comput. Linguistics*, 7:217–231.

Yi Sun, Yu Zheng, Chao Hao, and Hangping Qiu. 2021. NSP-BERT: A prompt-based zero-shot learner through an original pre-training task-next sentence prediction. *CoRR*, abs/2109.03564.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT (1)*, pages 4149–4158. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *EMNLP/IJCNLP (1)*, pages 6075–6084. Association for Computational Linguistics.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguistics*, 6:287–302.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zeroprompt: Scaling prompt-based pretraining to 1, 000 tasks improves zero-shot generalization. *CoRR*, abs/2201.06910.

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.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Appendix

A Dataset Details

Based on their usage stages, We summarize all datasets in two parts: MC training datasets and evaluation datasets.

Datasets	# of option	# of examples
ARC	4	3.37k
CommonsenseQA	5	9.7k
Cos-E	5	10.9k
CosmosQA	4	25.2k
Dream	4	10k
Mctest	4	2.4k
MultiRC	multiple	12k
OpenbookQA	4	9.9k
PIQA	2	16.1k
QASC	8	8.1k
Race	4	87.8k
Socail IQa	3	33.4k
WikiHop	multiple	43.7k
WIQA	3	36.7k

Table 8: Dataset statistics for Multiple Choice task

A.1 MC training datasets

Multiple Choice (MC) task aims to select a right answer from multiple candidate options according to the related questions and passages. As shown in Table 8, we use the following datasets in MC training phase:

1. ARC (Clark et al., 2018)
2. CommonsenseQA (Talmor et al., 2019)
3. Cos-E (Rajani et al., 2019)
4. CosmosQA (Huang et al., 2019)
5. Dream (Sun et al., 2019)
6. Mctest (Richardson et al., 2013)
7. MultiRC (Khashabi et al., 2018)
8. OpenbookQA (Mihaylov et al., 2018)
9. PIQA (Bisk et al., 2020)
10. QASC (Khot et al., 2020)
11. Race (Lai et al., 2017)
12. Socail IQA (Sap et al., 2019)
13. WikiHop (Welbl et al., 2018)
14. WIQA (Tandon et al., 2019)

A.2 Evaluation datasets

To evaluate zero-shot capability of models, we collect several NLP datasets and group them by tasks. The datasets with tasks are following:

Natural language inference (NLI) is to ascertain whether a “hypothesis” with a “premise” is true (entailment), false (contradiction), or indeterminate (neutral).

1. ANLI (R1-R3) (Nie et al., 2020)
2. CB (de Marneffe et al., 2019)
3. SNLI (Bowman et al., 2015)
4. MNLI-m/mm (Williams et al., 2018)
5. QNLI (Rajpurkar et al., 2018)
6. RTE (Dagan et al., 2005; Giampiccolo et al., 2007b,a; Bentivogli et al., 2009)
7. WNLI (Levesque et al., 2012b)

Commonsense reasoning (Commonsense) requires the model to draw conclusions based on "common sense" or general information.

1. COPA (Roemmele et al., 2011)
2. Hellaswag (Zellers et al., 2019)

Sentiment analysis (Sentiment) is to classify the polarity of a given text.

1. SST-2 (Socher et al., 2013)
2. IMDB (Maas et al., 2011)

Coreference resolution (Coreference) is the process of grouping textual mentions that refer to the same underlying real-world objects.

1. Winogrande (Sakaguchi et al., 2020)

Task	Dataset	Passage	Question	Options
NLI	ANLI R1 ANLI R2 ANLI R3 CB SNLI MNLI-m MNLI-mm	x_1	Base on the paragraph.	[1] We can infer that x_2; [2] We can not infer that x_2; [3] It is difficult for us to infer x_2.
	QNLI RTE WNLI	x_1	Base on the paragraph.	[1] We can infer that x_2; [2] We can not infer that x_2.
Sentiment	SST-2 IMDB	x	What is sentiment of reviews?	[1] It's great; [2] It's terrible.
Classification	AG News	x	What is the topic of the news?	[1] World news; [2] Sports news; [3] Business news; [4] Technology news.
	Dbpedia	x	What is the topic of the articles?	[1] Company; [2] Educational Institution; ... [13] Written Work.

Table 9: Prompt designs for all datasets.

Model	Layers	Hidden	Heads	Parameters
UniMC-12M	12	768	12	12M
UniMC-60M	24	2048	16	60M
UniMC-235M	12	4096	64	235M

Table 10: The configurations if the UniMC variants.

	RoBERTa	ALBERT
Parameters	355M	235M
NLI (Acc)	53.0	**58.2**
Sentiment (Acc)	92.8	**93.2**

Table 11: Ablation experiments with different backbones. "RoBERTa" indicates RoBERTa-large and "ALBERT" presents ALBERT-xxlarge-v2.

2. WSC (Levesque et al., 2012a)
3. DPR (Rahman and Ng, 2012)

Text classification (Classification) is the task of assigning a label to a given text.

1. AG News (Zhang et al., 2015)
2. Dbpedia (Lehmann et al., 2015)

A.3 Unified input

Inspired by template examples in FLAN (Wei et al., 2021), we design a simple rule to transform the original text to a unified MC format as shown in Table 9. In addition, we present two examples:

An example of Social IQA (multiple choice):

1. Raw text: {x_1: "Jesse placed the music sheet in his hands and began to sing a song.", "question": "What will Jesse want to do next?", "option": ["paint a picture", "make a speech", "start the song"], "answer": "start the song"}
2. Transformed text: "no paint a picture. no make a speech. yes start the song. What will Jesse want to do next? Jesse placed the music sheet in his hands and began to sing a song."
3. Input tokens: `[O-MASK]` paint a picture. `[O-MASK]` make a speech. `[O-MASK]` start the song. What will Jesse want to do next? Jesse placed the music sheet in his hands and began to sing a song.

An example of SNLI (natural language inference):

1. Raw text: {x_1: "A man reads the paper in a bar with green lighting.", x_2: "The man is inside.", "option": ["we can infer that", "we can not infer that The man is inside.", "it is difficult for us to infer that The man is inside."], "answer": "we can infer that The man is inside."}
2. Transformed text: "yes we can infer that The man is inside. no we can not infer that The man is inside. no it is difficult for us to infer that The man is inside. Base on the paragraph. A man reads the paper in a bar with green lighting."
3. Input tokens: "`[O-MASK]` we can infer that The man is inside. `[O-MASK]` we can not infer that The man is inside. `[O-MASK]` it is difficult for us to infer that The man is inside. Base on the paragraph. A man reads the paper in a bar with green lighting."

B Additional Experiments

B.1 UniMC variants with different parameters

By following the setting of ALBERT (Lan et al., 2020), UniMC employs various ALBERT models as the backbones as shown in Table 10.

B.2 Further ablation study: Different backbone models

To explore the effect of different backbone models in UniMC, we replace the ALBERT-xxlarge-v2 with RoBERTa-large. As seen in Table 11, ALBERT outperforms RoBERTa in the NLI and sentiment analysis task. A simple explanation is that ALBERT-xxlarge-v2 (Lan et al., 2020) (88.9 point) performs

Dataset	GPT3 175B	T0 11B	GLaM 60B/MoE	FLAN 137B	PaLM 8B	PaLM 60B	PaLM 540B	UniMC 235M
NLI								
ANLI R1	34.6	43.6	40.9	47.7	34.9	36.4	48.4	**52.0**
ANLI R2	35.4	38.7	38.2	43.9	35.8	37.2	44.2	**44.4**
ANLI R3	34.5	41.3	40.9	47.0	34.5	36.7	45.7	**47.8**
CB	46.4	70.1	33.9	64.1	41.1	57.1	51.8	**75.7**
RTE	63.5	**80.8**	68.8	78.3	54.2	67.9	72.9	78.1
QNLI	-	-	-	**59.6**	-	-	-	54.0
SNLI	-	-	-	43.0	-	-	-	**60.9**
MNLI-m	-	-	-	51.1	-	-	-	**52.7**
MNLI-mm	-	-	-	51.0	-	-	-	**51.4**
WNLI	-	-	-	61.0	-	-	-	**65.4**
Commonsense								
COPA	91.0	90.0	90.0	90.6	86.0	93.0	93.0	**95.2**
Hellaswag	78.9	33.6	77.1	56.4	68.7	79.7	**83.4**	62.5
Sentiment								
SST-2	71.6	-	-	**92.6**	-	-	-	91.6
IMDB	-	-	-	94.1	-	-	-	**94.8**
Coreference								
Winogrande	70.2	59.9	73.4	67.3	66.3	77.0	**81.1**	65.8
WSC	88.3	65.1	86.8	80.8	78.9	86.3	**89.1**	78.8
DPR	-	-	-	60.3	-	-	-	**87.5**

Table 12: Zero-shot performances on different tasks: NLI, Commonsense, Sentiment, and Coreference.

beyond RoBERTa-large (Liu et al., 2019b) in their paper. In our experiments, tokenization might be another possible reason. Since O-MLM aims to predict "yes" or "no", UniMC needs a stable tokenizer to recover those words. Unlike ALBERT, RoBERTa uses a byte-level BPE tokenizer instead of a WordPiece tokenizer. Under the settings of the byte-level BPE tokenizer, the word id does not only depend on the word itself, but also is influenced by its position. Therefore, RoBERTa faces tough O-MLM and OP tasks in the MC training phase, which presents lower score than ALBERT. We chose ALBERT, which has better results, as the default backbone model in all our experiments.

B.3 Results on all datasets

In Table 12, we can see that UniMC achieves the best performance on 11 out of 17 datasets. PLMs outperform UniMC in the tasks of commonsense reasoning and coreference resolution in Hallawag, Winogrand, and WSC, as these are formulated in the original language modeling pre-training objective, as noted in (Wei et al., 2021). In addition, PLMs benefit from unsupervised language modeling on a large-scale text corpus. For example, PaLM with 540B parameters is pre-trained on data with 780 billion tokens.