Wudao—Pretrain the world

Tang Jie
Tsinghua University
Beijing Academy of Artificial Intelligence (BAAI)

The slides will be available soon at http://keg.cs.tsinghua.edu.cn/jietang (or Google “Jie Tang”)
AI History

1st
Symbolic model, perceptron

Symbolic

1956 Dartmouth Conference: The Founding Fathers of AI

2nd
Big data

Recognition

3rd
Cognition

Data and Knowledge

1st
Symbolic model, perceptron

2nd
Big data

Recognition

3rd
Cognition

Data and Knowledge

OpenAI, Google DeepMind, DARPA

1st
Symbolic model, perceptron

2nd
Big data

Recognition

3rd
Cognition

Data and Knowledge

OpenAI, Google DeepMind, DARPA
GPT-3

- GPT-3: OpenAI released GPT-3 in 2020/5 with super-scale parameters: 175B
- Google’s Switch Transformer: 1.6T

![GPT-3 for QA](image)

DALL·E: (Generating Images by Text)

a snail made of harp
Question: Which moment did you want to live in forever?

**WUDAO:** In an instant I recall the second year of high school. That afternoon, I took a sunbath on the playground for a long time. I then played basketball downstairs with my classmates. As the sun set slowly, the three of us stood together on the corridor, leaning on the railing to face the sunset. I’ve always wanted to live in that moment forever, but unfortunately I never encountered a similar moment again. I just want to live in that moment.
作诗图灵测试

Bitcoin

Easy  Hard  Lunatic  Extra

在作诗图灵测试的Easy模式中，您将会被展现5组诗歌（包括标题、作者及内容），每组包括1首由诗人创作的诗歌和1首AI创作的诗歌，请选择您认为由人创作的诗歌。所有组选择完成后，您将会得知有多少组结果正确。
外挖无穷洞, 机神犹未休。
卡中窥币影, 池里验沙流。
屡载吸金主, 孤深渍盗求。
方知区块链, 本是古来游。

比特币
Mining an infinite hole outside
The god of machine is still not satisfied
Illusion of bitcoins seen inside the GPU
Verified in pools to confirm the workflow
Several years it has attracted many investors
Deep and limited, even the robbers pursue
Eventually it turns out that Blockchain
May only be a financial trick from ancient
磻溪连灞水，商岭接秦山。青汉不回驾，白云长掩关。雀喧知鹤静，凫戏识鸥闲。却笑南昌尉，悠悠城市间。

磻溪连灞水，商岭接秦山。青汉不回驾，白云长掩关。雀喧知鹤静，凫戏识鸥闲。却笑南昌尉，悠悠城市间。

磻溪连灞水，商岭接秦山。青汉不回驾，白云长掩关。雀喧知鹤静，凫戏识鸥闲。却笑南昌尉，悠悠城市间。

磻溪连灞水，商岭接秦山。青汉不回驾，白云长掩关。雀喧知鹤静，凫戏识鸥闲。却笑南昌尉，悠悠城市间。
Draw Pictures
https://wudao.aminer.cn/CogView/index.html

A woman in a black and purple dress poses in front of some tall grass.

A woman is on a bench overlooking the city.

A couple of young boys playing a game of soccer.

A man that is on a surfboard in some water.

A women in a white blouse is holding a remote in her hands.

A bird perched on top of a leafless tree under a blue sky.

A clock hanging outside of a house in a nice neighborhood.

A red bus is driving on the road.

A beautiful young blond woman talking on a phone.

A red bowl filled with food and leafy greens.
Draw Pictures-Image completion

- A girl with a surgical mask
- A girl wearing a tie
• Draw Pictures-Image completion

• A man with a red ball
• 一个胖子在吃一碗面
Compare with Dalle

Alibaba DAMO Academy
AI-aided Design

Better than DALL.E on MS COCO

Different Styles

KDD’21.
Core Algorithms

- Controllable Generation via Inverse Prompting
- CogView: Text-to-Image Generation
- Prompt Tuning vs. Fine Tuning
- GLM: General Language Model Pretraining
- How to build a trillion-scale model
WUDAO 2.0 is the World's largest 1.75 Trillion parameters

10X larger than GPT-3 parameters
What is WuDao 2.0

1. 1.75 Trillion Parameters
2. both text and images
3. train on a supercomputer
4. Bilingual (Cn and En) data: 4.9T text and images
What is WuDao 2.0

FastMoE

- Support complex balance strategies such as Switch and GShard
- Support different experts and different models.
- Deploy in Alibaba PAI platform
- Explore the application to the Alipay intelligent service system
- Training on a supercomputer

FastMoE
- Redesign all operators
- Efficient communication strategy
- MoE training with tens of thousands of experts

Z. Ma et al. BAGUALU: Targeting Brain Scale Pretrained Models with over 37 Million Cores. PPoPP’22.
GLM: General Language Model Pretraining with Autoregressive Blank Infilling

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL’22.
None of the pretraining frameworks performs the best for all tasks.

- Autoregressive model: GPT, GPT-2, GPT-3
- Autoencoding model: BERT, RoBERTa, ALBERT
- Encoder-Decoder model: MASS, BART, PALM

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL’22.
All NLP Tasks are Generations

Figure 3. GLM finetune framework. (a) Formulation of the sentiment classification task as blank infilling with GLM. (b) GLM for text generation given the context. This can be the language modeling in the zero-shot setting, or seq2seq with fine-tuning.

NLU, Cond. Gen, Uncond. Gen can be unified into the GLM generation framework

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL’22.
**A New Pretraining Framework**

**Figure 1.** Illustration of GLM. We blank out text spans (green part) and GLM is trained to generate them in an autoregressive fashion.

Multi-task pretraining
1. Sample 15% in the middle as the generation objective
2. Sample 50-100% in the end as the generation objective
GLM: Autoregressive Blank Filling

(a) Sample spans from the input text

Part A: $x_1, x_2, [M], x_4, [M]$

Part B: $x_5, x_6, x_3$

(b) Divide the input into Part A and Part B

(c) Generate the Part B spans autoregressively

(d) Self-attention mask

<table>
<thead>
<tr>
<th>Position 1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>5</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
### Results: NLU–Classification

Table 2. Results on the SuperGLUE dev set. Models with * are pre-trained for two times the number of steps of other methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>ReCoRD F1/Acc.</th>
<th>COPA Acc.</th>
<th>WSC Acc.</th>
<th>RTE Acc.</th>
<th>BoolQ Acc.</th>
<th>WiC Acc.</th>
<th>CB F1/Acc.</th>
<th>MultiRC F1/a/EM</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>65.4/64.9</td>
<td>66.0</td>
<td>65.4</td>
<td>70.0</td>
<td>74.9</td>
<td><strong>68.8</strong></td>
<td>70.9/76.8</td>
<td>68.4/21.5</td>
<td>66.1</td>
</tr>
<tr>
<td>GLM&lt;sub&gt;Base&lt;/sub&gt;</td>
<td><strong>73.5/72.8</strong></td>
<td><strong>71.0</strong></td>
<td><strong>72.1</strong></td>
<td><strong>71.2</strong></td>
<td><strong>77.0</strong></td>
<td>64.7</td>
<td><strong>89.5/85.7</strong></td>
<td><strong>72.1/26.1</strong></td>
<td><strong>70.7</strong></td>
</tr>
<tr>
<td>BERT&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>76.3/75.6</td>
<td>69.0</td>
<td>64.4</td>
<td>73.6</td>
<td>80.1</td>
<td><strong>71.0</strong></td>
<td>94.8/92.9</td>
<td>71.9/24.1</td>
<td>72.0</td>
</tr>
<tr>
<td>UniLM&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>80.0/79.1</td>
<td>72.0</td>
<td>65.4</td>
<td>76.5</td>
<td>80.5</td>
<td>69.7</td>
<td>91.0/91.1</td>
<td>77.2/38.2</td>
<td>74.1</td>
</tr>
<tr>
<td>GLM&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>81.7/81.1</td>
<td>76.0</td>
<td><strong>81.7</strong></td>
<td>74.0</td>
<td><strong>82.1</strong></td>
<td>68.5</td>
<td>96.1/94.6</td>
<td><strong>77.1/36.3</strong></td>
<td>77.0</td>
</tr>
<tr>
<td>GLM&lt;sub&gt;Large&lt;/sub&gt; (multi-task)</td>
<td>80.2/79.6</td>
<td>77.0</td>
<td>78.8</td>
<td>76.2</td>
<td>79.8</td>
<td>63.6</td>
<td><strong>97.3/96.4</strong></td>
<td>74.6/32.1</td>
<td>75.7</td>
</tr>
<tr>
<td>GLM&lt;sub&gt;410M&lt;/sub&gt; (multi-task)</td>
<td>81.5/80.9</td>
<td>80.0</td>
<td><strong>81.7</strong></td>
<td><strong>79.4</strong></td>
<td>81.9</td>
<td>69.0</td>
<td>93.2/96.4</td>
<td>76.2/35.5</td>
<td>78.0</td>
</tr>
<tr>
<td>GLM&lt;sub&gt;515M&lt;/sub&gt; (multi-task)</td>
<td><strong>82.3/81.7</strong></td>
<td><strong>85.0</strong></td>
<td><strong>81.7</strong></td>
<td>79.1</td>
<td>81.3</td>
<td>69.4</td>
<td>95.0/96.4</td>
<td>77.2/35.0</td>
<td><strong>78.8</strong></td>
</tr>
<tr>
<td>T5&lt;sub&gt;Base&lt;/sub&gt;</td>
<td>76.2/75.4</td>
<td>73.0</td>
<td>79.8</td>
<td>78.3</td>
<td>80.8</td>
<td>67.9</td>
<td>94.8/92.9</td>
<td>76.4/40.0</td>
<td>76.0</td>
</tr>
<tr>
<td>T5&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>85.7/85.0</td>
<td>78.0</td>
<td><strong>84.6</strong></td>
<td>84.8</td>
<td>84.3</td>
<td>71.6</td>
<td>96.4/98.2</td>
<td>80.9/46.6</td>
<td>81.2</td>
</tr>
<tr>
<td>BART&lt;sub&gt;Large&lt;/sub&gt;*</td>
<td>88.3/87.8</td>
<td>60.0</td>
<td>65.4</td>
<td>84.5</td>
<td>84.3</td>
<td>69.0</td>
<td>90.5/92.9</td>
<td>81.8/48.0</td>
<td>76.0</td>
</tr>
<tr>
<td>RoBERTa&lt;sub&gt;Large&lt;/sub&gt;*</td>
<td>89.0/88.4</td>
<td><strong>90.0</strong></td>
<td>63.5</td>
<td>87.0</td>
<td><strong>86.1</strong></td>
<td><strong>72.6</strong></td>
<td>96.1/94.6</td>
<td><strong>84.4/52.9</strong></td>
<td><strong>81.5</strong></td>
</tr>
<tr>
<td>GLM&lt;sub&gt;RoBERTa&lt;/sub&gt;</td>
<td><strong>89.6/89.0</strong></td>
<td>82.0</td>
<td>83.7</td>
<td><strong>87.7</strong></td>
<td>84.7</td>
<td>71.2</td>
<td><strong>98.7/98.2</strong></td>
<td>82.4/50.1</td>
<td><strong>82.9</strong></td>
</tr>
</tbody>
</table>

- Better than BERT, T5, RoBERTa

Z. Du et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. ACL’22.
Results: Uncond. Gen, Cond. Gen

Before
Train three different models

After
1.25 X Larger GLM can do all the three tasks with one model!

Table 3. Results on Gigaword abstractive summarization

<table>
<thead>
<tr>
<th>Model</th>
<th>RG-1</th>
<th>RG-2</th>
<th>RG-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASS</td>
<td>37.7</td>
<td>18.5</td>
<td>34.9</td>
</tr>
<tr>
<td>UniLMLarge</td>
<td>38.5</td>
<td>19.5</td>
<td>35.8</td>
</tr>
<tr>
<td>GLM Large</td>
<td>38.6</td>
<td>19.7</td>
<td>36.0</td>
</tr>
<tr>
<td>GLM Large (multi-task)</td>
<td>38.5</td>
<td>19.4</td>
<td>35.8</td>
</tr>
<tr>
<td>GLM 410M (multi-task)</td>
<td><strong>38.9</strong></td>
<td><strong>20.0</strong></td>
<td><strong>36.2</strong></td>
</tr>
</tbody>
</table>

Table 4. Zero-shot language modeling results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lambada (Accuracy)</th>
<th>BookWiki (Perplexity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM Large (uni)</td>
<td>0.0</td>
<td>&gt; 100</td>
</tr>
<tr>
<td>GLM Large (multi-task,uni)</td>
<td>47.4</td>
<td>15.1</td>
</tr>
<tr>
<td>– 2d positional encoding</td>
<td>45.8</td>
<td>15.1</td>
</tr>
<tr>
<td>GLM 410M (multi-task,uni)</td>
<td>49.5</td>
<td>14.5</td>
</tr>
<tr>
<td>GLM 515M (multi-task,uni)</td>
<td><strong>50.4</strong></td>
<td><strong>13.9</strong></td>
</tr>
<tr>
<td>GLM Large (bi)</td>
<td>10.6</td>
<td>&gt; 100</td>
</tr>
<tr>
<td>GLM Large (multi-task,bi)</td>
<td>48.5</td>
<td>14.9</td>
</tr>
<tr>
<td>– 2d positional encoding</td>
<td>47.3</td>
<td>15.0</td>
</tr>
<tr>
<td>GLM 410M (multi-task,bi)</td>
<td><strong>53.5</strong></td>
<td><strong>14.3</strong></td>
</tr>
<tr>
<td>GLM 515M (multi-task,bi)</td>
<td><strong>54.9</strong></td>
<td><strong>13.7</strong></td>
</tr>
<tr>
<td>GPTLage (uni)</td>
<td>50.1</td>
<td>14.4</td>
</tr>
</tbody>
</table>
CogView: Mastering Text-to-Image Generation via Transformers.

CogView: Text-to-Image Generation

- CogView: 4B
- Training with 30M image-text pairs on 512 V100 for 400 hours

**Input Text:** (The head of a lovely cat.)
一只可爱的小猫的头像。

**Text Tokenizer (sentence pieces)**

- [ROI1] Text Token ...
- Text Token

Text tokens, ranging from 8192 to 58192.

**Input Image:**

- Image Tokenizer (VQAE)

- Encoder
- Discretize
- Recover
- Decoder
- Image Token ...
- Image Token

1024 Image tokens, ranging from 0 to 8191.

Transformer (GPT)
CogView Model

Input Text: (The head of a lovely cat.)
一只可爱的小猫的头像。

Input Image:

Text Tokenizer (sentence pieces)
[ROI1] Text Token ...... Text Token [BASE] [BOX1] Image Token ...... Image Token [EOI1]

Transformer

Hidden representation

Mix Tokens

Text Tokens (0-127)
Text Tokens (sentencepiece, Cn-En) (128-199999)
Image (200000-200127)
Image Token (VQ-VAE) (200128-208319)

All NLP tasks are generation tasks

Tasks

NLP tasks

All [MASK] are generation tasks [START] NLP tasks
Results

37% of the users favor the generated images by CogView Better than DALL.E

<table>
<thead>
<tr>
<th>Model</th>
<th>FID-0</th>
<th>FID-1</th>
<th>FID-2</th>
<th>FID-4</th>
<th>FID-8</th>
<th>IS</th>
<th>CapS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AttnGAN</td>
<td>35.2</td>
<td>44.0</td>
<td>72.0</td>
<td>108.0</td>
<td>100.0</td>
<td>23.3</td>
<td>0.02763</td>
</tr>
<tr>
<td>DM-GAN</td>
<td>26.0</td>
<td>39.0</td>
<td>73.0</td>
<td>119.0</td>
<td>112.3</td>
<td>32.2</td>
<td>0.02801</td>
</tr>
<tr>
<td>DF-GAN</td>
<td>26.0</td>
<td>33.8</td>
<td>55.9</td>
<td>91.0</td>
<td>97.0</td>
<td>18.7</td>
<td>0.02802</td>
</tr>
<tr>
<td>DALL-E</td>
<td>27.5</td>
<td>28.0</td>
<td>45.5</td>
<td>83.5</td>
<td>85.0</td>
<td>17.9</td>
<td>—</td>
</tr>
<tr>
<td>CogView</td>
<td>27.1</td>
<td>19.4</td>
<td>13.9</td>
<td>19.4</td>
<td>23.6</td>
<td>18.2</td>
<td>0.17403</td>
</tr>
</tbody>
</table>

Table 1: Metrics for machine evaluation. Statistics about DALL-E are extracted from their figures. FID-\(k\) means that all the images are blurred by a Gaussian Filter with radius \(k\).
Prompt Tuning vs. Fine Tuning

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL'22.
Prompt

- Handwriting Prompt (Brown et al 2020)
- Continuous Prompt (P–Tuning)

<table>
<thead>
<tr>
<th>Prompt</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>[X] is located in [Y]. <em>(original)</em></td>
<td>31.29</td>
</tr>
<tr>
<td>[X] is located in which country or state? [Y].</td>
<td>19.78</td>
</tr>
<tr>
<td>[X] is located in which country? [Y].</td>
<td>31.40</td>
</tr>
<tr>
<td>[X] is located in which country? In [Y].</td>
<td>51.08</td>
</tr>
</tbody>
</table>

*Table 1.* Case study on LAMA-TREx P17 with bert-base-cased. A single-word change in prompts could yield a drastic difference.
Figure 2. An example of prompt search for “The capital of Britain is [MASK]”. Given the context (blue zone, “Britain”) and target (red zone, “[MASK]”), the orange zone refer to the prompt tokens. In (a), the prompt generator only receives discrete rewards; on the contrary, in (b) the pseudo prompts and prompt encoder can be optimized in a differentiable way. Sometimes, adding few task-related anchor tokens (such as “capital” in (b)) will bring further improvement.
Results on LAMA

<table>
<thead>
<tr>
<th>Prompt type</th>
<th>Model</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original (MP)</td>
<td>BERT-base</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>BERT-large</td>
<td>32.3</td>
</tr>
<tr>
<td></td>
<td>E-BERT</td>
<td>36.2</td>
</tr>
<tr>
<td>Discrete</td>
<td>LPAQA (BERT-base)</td>
<td>34.1</td>
</tr>
<tr>
<td></td>
<td>LPAQA (BERT-large)</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>AutoPrompt (BERT-base)</td>
<td>43.3</td>
</tr>
<tr>
<td>P-tuning</td>
<td>BERT-base</td>
<td>48.3</td>
</tr>
<tr>
<td></td>
<td>BERT-large</td>
<td>50.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>MP</th>
<th>FT</th>
<th>MP+FT</th>
<th>P-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base (109M)</td>
<td>31.7</td>
<td>51.6</td>
<td>52.1</td>
<td>52.3 (+20.6)</td>
</tr>
<tr>
<td>-AutoPrompt (Shin et al., 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>45.2</td>
</tr>
<tr>
<td>BERT-large (335M)</td>
<td>33.5</td>
<td>54.0</td>
<td>55.0</td>
<td>54.6 (+21.1)</td>
</tr>
<tr>
<td>RoBERTa-base (125M)</td>
<td>18.4</td>
<td>49.2</td>
<td>50.0</td>
<td>49.3 (+30.9)</td>
</tr>
<tr>
<td>-AutoPrompt (Shin et al., 2020)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40.0</td>
</tr>
<tr>
<td>RoBERTa-large (355M)</td>
<td>22.1</td>
<td>52.3</td>
<td>52.4</td>
<td>53.5 (+31.4)</td>
</tr>
<tr>
<td>GPT2-medium (345M)</td>
<td>20.3</td>
<td>41.9</td>
<td>38.2</td>
<td>46.5 (+26.2)</td>
</tr>
<tr>
<td>GPT2-xl (1.5B)</td>
<td>22.8</td>
<td>44.9</td>
<td>46.5</td>
<td>54.4 (+31.6)</td>
</tr>
<tr>
<td>MegatronLM (11B)</td>
<td>23.1</td>
<td>OOM*</td>
<td>OOM*</td>
<td>64.2 (+41.1)</td>
</tr>
</tbody>
</table>

* MegatronLM (11B) is too large for effective fine-tuning.

Table 2. Knowledge probing Precision@1 on LAMA-34k (left) and LAMA-29k (right). P-tuning outperforms all the discrete prompt searching baselines. And interestingly, despite fixed pre-trained model parameters, P-tuning overwhelms the fine-tuning GPTs in LAMA-29k. (MP: Manual prompt; FT: Fine-tuning; MP+FT: Manual prompt augmented fine-tuning; PT: P-tuning).

- **Significantly boost performance on LAMA**
- **Pre-trained models have learned more knowledge than we thought.**

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL'22.
Results on Few-shot NLU

<table>
<thead>
<tr>
<th>Dev size</th>
<th>Method</th>
<th>BoolQ (Acc.)</th>
<th>CB (Acc.)</th>
<th>CB (F1)</th>
<th>WiC (Acc.)</th>
<th>RTE (Acc.)</th>
<th>MultiRC (EM)</th>
<th>MultiRC (F1a)</th>
<th>WSC (Acc.)</th>
<th>COPA (Acc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>PET*</td>
<td>73.2±3.1</td>
<td>82.9±4.3</td>
<td>74.8±9.2</td>
<td>51.8±2.7</td>
<td>62.1±5.3</td>
<td>33.6±3.2</td>
<td>74.5±1.2</td>
<td>79.8±3.5</td>
<td>85.3±5.1</td>
</tr>
<tr>
<td></td>
<td>PET best†</td>
<td>75.1</td>
<td>86.9</td>
<td>83.5</td>
<td>52.6</td>
<td>65.7</td>
<td>35.2</td>
<td>75.0</td>
<td>80.4</td>
<td>83.3</td>
</tr>
<tr>
<td></td>
<td>P-tuning</td>
<td>77.8 (+4.6)</td>
<td>92.9 (+10.0)</td>
<td>92.3 (+17.5)</td>
<td>56.3 (+4.5)</td>
<td>76.5 (+14.4)</td>
<td>36.1 (+2.5)</td>
<td>75.0 (+0.5)</td>
<td>84.6 (+4.8)</td>
<td>87.0 (+1.7)</td>
</tr>
<tr>
<td>Full</td>
<td>GPT-3</td>
<td>77.5</td>
<td>82.1</td>
<td>57.2</td>
<td>55.3</td>
<td>72.9</td>
<td>32.5</td>
<td>74.8</td>
<td>75.0</td>
<td>92.0</td>
</tr>
<tr>
<td></td>
<td>PET‡</td>
<td>79.4</td>
<td>85.1</td>
<td>59.4</td>
<td>52.4</td>
<td>69.8</td>
<td>37.9</td>
<td>77.3</td>
<td>80.1</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>iPET§</td>
<td>80.6</td>
<td>92.9</td>
<td>92.4</td>
<td>52.2</td>
<td>74.0</td>
<td>33.0</td>
<td>74.0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* We report the average and standard deviation of each candidate prompt’s average performance.
† We report the best performed prompt selected on full dev dataset among all candidate prompts.
‡ With additional ensemble and distillation.
§ With additional data augmentation, ensemble, distillation and self-training.

- **Few-shot SOTA on SuperGLUE**
- **A more reasonable few-shot setting: small train and SMALL DEV**

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL'22.
P-Tuning for GPT

- Boost GPT on NLU
- Improve BERT on NLU

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL'22.
One more thing

P-tuning outperforms fine-tuning only when the # of parameters >10B!

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL'22.
One more thing

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL'22.
Finally, P-tuning $\geq$ fine-tuning

P-tuning achieves comparable performance with only 0.1% task-specific parameters, as fine-tuning even with small models (<10B).

Figure 1: Average scores on RTE, BoolQ and CB of SuperGlue dev. With 0.1% task-specific parameters, P-tuning v2 can be comparable to fine-tuning across different scales of pre-trained models, while prompt tuning & P-tuning can only make it at over 10B scales.

X. Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL’22.
Controllable Generation from Pretrained Language Models via Inverse Prompting

Question: Which moment did you want to live in forever?

I have been in a relationship with my boyfriend for two years. I studied abroad and he studied domestically. Alone in a foreign country, I felt really lonely and sad. When I was in school, I felt that it was ok to be together, working hard together and quarrelling together. I kept thinking in this way. In the last few months of our time at the high school, I felt that he was my walking little brother. In fact, the happiest moment was when we were together in the “orphanage”.

The relativeness is weak!

Question: Which moment did you want to live in forever?

Inverse Prompting:

In an instant I recall the second year of high school. That afternoon, I took a sunbath on the playground for a long time. I then played basketball downstairs with my classmates. As the sun set slowly, the three of us stood together on the corridor, leaning on the railing to face the sunset. I’ve always wanted to live in that moment forever, but unfortunately I never encountered a similar moment again. I just want to live in that moment.
Inverse Prompting

Figure 1: The generation process of open-domain traditional Chinese poems under inverse prompting. Using title New York (《咏纽约》) as an example.

 Beam search using Inverse Prompting scoring for several steps.

Inverse Prompting scoring method:
Inversely prompt the title to improve the relativeness.
咏纽约

曼哈顿中云拂雨，百废俱兴壮帝京。
金融中心兴万象，顷刻资财赋圣明。
此城五洲俯仰际，天下财富朝暮倾。
若非王气起天壤，世界繁华岂易名。

New York

Clouds touch the rain in Manhattan
Everything grows to revive the empire
All fields are prosperous in the financial center
Capital brightens the future in a sudden
The city is worshipped by all continents
Wealth from all over the world flows day and night
Unless a distant paradise rises from the sky-rim
How can somewhere else become more flourishing?

Combination of traditional Chinese poem & modern objects/images via Inverse Prompting!
### Evaluation: QA

<table>
<thead>
<tr>
<th>Method</th>
<th>Fluency (1-5)</th>
<th>Inform.(^1) (1-5)</th>
<th>Relevance (1-5)</th>
<th>Overall (1-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPM [27]</td>
<td>2.66</td>
<td>2.47</td>
<td>2.36</td>
<td>4.32</td>
</tr>
<tr>
<td>Prompting Baseline</td>
<td>3.44</td>
<td>3.25</td>
<td>3.21</td>
<td>5.97</td>
</tr>
<tr>
<td>Inverse Prompting</td>
<td><strong>3.61</strong></td>
<td><strong>3.43</strong></td>
<td><strong>3.59</strong></td>
<td><strong>6.51</strong></td>
</tr>
<tr>
<td>Human Answers</td>
<td>3.80</td>
<td>3.61</td>
<td>3.67</td>
<td>6.85</td>
</tr>
</tbody>
</table>

\(^1\) Informativeness
### Evaluation: Poem

<table>
<thead>
<tr>
<th>Method</th>
<th>Format (1-5)</th>
<th>Innov.(^1) (1-5)</th>
<th>Relevance (1-5)</th>
<th>Aes.(^2) (1-5)</th>
<th>Overall (1-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiuge [28] Search Baseline</td>
<td>3.60</td>
<td>2.47</td>
<td>1.99</td>
<td>3.12</td>
<td>3.57</td>
</tr>
<tr>
<td>Inverse Prompting</td>
<td>2.56</td>
<td>2.71</td>
<td>2.92</td>
<td>2.33</td>
<td>4.00</td>
</tr>
<tr>
<td>Inverse Prompting +ST</td>
<td>2.42</td>
<td>2.92</td>
<td>3.65</td>
<td>2.18</td>
<td>4.40</td>
</tr>
</tbody>
</table>

\(^1\) Innovation  
\(^2\) Aesthetics
## Turing Test

<table>
<thead>
<tr>
<th>Method</th>
<th>Total</th>
<th>Selected</th>
<th>Selection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse Prompting +ST</td>
<td>1,656</td>
<td>748</td>
<td>45.2%</td>
</tr>
<tr>
<td>Ancient Human Poems</td>
<td>1,656</td>
<td>908</td>
<td>54.8%</td>
</tr>
</tbody>
</table>

- 45.2%
WuDao Ecology

https://wudaoai.cn/
Open!

01. WuDaoCorpora: the world's largest publicly available dataset!

02. We released almost all codes in WuDao!

03. You can download >20 well-trained models!

04. Directly call an API to enjoy the power of WuDao!

Click here to find more: https://wudaoai.cn/
WuDao’s Today

WuDao: A super-scale model with 1.75 Trillion parameters.

- WuDao is very general and can be applied to different tasks
- WuDao can do QA, write poem, knowledge extraction, draw pictures, write articles, recognize pictures, etc.
- WuDao is open to everyone

- WuDao is very expensive... and needs to reduce cost...
WuDao’s Tomorrow

Teach Machine to Think Like Humans: Beyond the Turing Test!

1. Adaptation and Learning
2. Definition and Contextualization
3. Character Setting
4. Priority and Access Control
5. Call Together and Control
6. Decision Making and Execution
7. Probing and Editing
8. Reflection and Self-Monitoring
9. Logic and Flexibility

The slides will be available soon at http://keg.cs.tsinghua.edu.cn/jietang (or Google “Jie Tang”)
Thanks to everyone!

**WUDAO·Wenyan**
A large scale pre-training language model with Chinese as its core

Liu Zhiyuan, Huang Minlie, Han Wentao, Liu Yang, Zhu Xiaoyan, Sun Maosong
Zhang Zhengyari, Gu Yuxian, Han Xu, Chen Shengqi, Xiao Chaoqun, Yao Yuan, Qi Fanchao, Guan Jian, Ke Pei, Zhou Hao, Sun Zhenbo, Cai Yanzheng, Zeng Guoyang, Tan Zhixing, Qin Yujia, Su Yusheng Si Chenglei, Hu Xueyu, Li Wenhao, Wang Fengyu, Yi Jing, Wang Xiaozhi, Chen Weize, Ding Ning, Zhang Jiajie

**WUDAO·Wenlan**
Super large multi-modal pre-training model


**WUDAO·Wenhui**
A new super-large cognitive-oriented pre-training model

Tang Jie, Yang Zhilin, Yang Hongxia
Guo Zitong, Liu Jing, Zhou Chang, Lin Junyang

**WUDAO·Wensu**
Super large protein sequence prediction pre-training model
FastMoE and trillion large model

Tang Jie, Lu Bai
Qiu Jiezhong, Xie Changyu, Xiao Yijia, Zeng Aohan, Li Zhaoh

Wudao—Pretrain the world

Thanks!