

# Wudao—Pretrain the world

—让机器像人一样“思考”

# ● 引领第三代人工智能，开启认知新篇章

## 认知智能



## 第三代

张钹院士2016年提出第三代人工智能雏形，DARPA 2018年发布AI Next计划。核心思路是推进数据统计与知识推理融合的计算；与脑认知机理融合的计算。

数据与知识双轮驱动

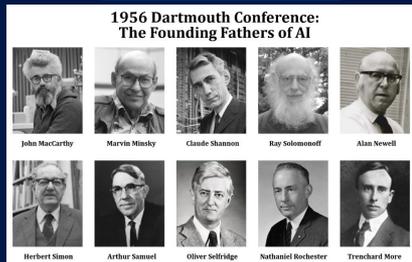
## 感知智能



## 第二代

大数据驱动统计学习方法初步实现了针对文本、图像、语音等的感知与识别

## 符号AI

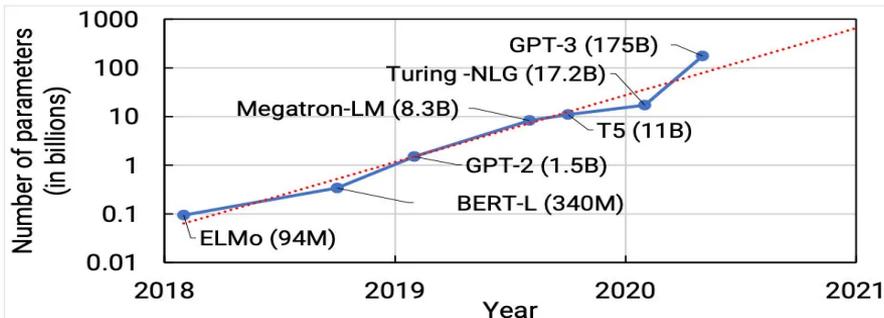


## 第一代

符号模型 规则模型 感知机

# GPT-3引发“大模型+大算力”探索通用人工智能

- ◆ **GPT-3**: OpenAI在2020年5月发布当时规模最大的预训练模型，参数达**1750亿**，在文本生成、对话、搜索等任务上性能优异。
- ◆ 2021年1月**谷歌**发布**Switch Transformer**预训练模型，参数达**1.6万亿**



问：热狗是三明治吗？

GPT-3：不是。

问：热狗不是肉夹在面包里面吗？

GPT-3：三明治是两片面包中夹肉和菜。

GPT-3：使用自然语言的聊天机器人。

问：你是说面包需要一片一片的才算三明治吗？

GPT-3:是的

根据文本提示：竖琴状的蜗牛，生成图像如下：



DALL·E：  
强大的文本—图像生成能力

人工智能 大模型时代到来

# AI大模型 = 发电厂



超大数据



超大算力



超大模型



AI 基础设施建设先行者!

# 2021年3月，悟道1.0 发布 中国首个超大智能模型



悟道1.0 彰显出中国在人工智能领域日益增长的信心与研发实力... 可能会对人们未来构建的人工智能系统类型和部署系统的方式产生深远影响!

- 吴恩达

“人工智能特战队”

10+

智源“悟道”顶尖 AI 科学家

智源学术副院长唐杰领衔“悟道”攻关团队  
100+ 清华、北大、人大、中科院等学者





# “悟道”——融合知识的大模型平台

完成了具有26、30亿参数的中文预训练模型以及100、1000、1.75万亿的多模态认知模型



打造数据与知识双轮驱动的认知智能

# ● 超越图灵测试的AI人设问答

Context: 问题：2021年的股市你觉得会是什么走势？蔚来还值得持有吗？问题描述：感觉最近涨得比较多，是不是有泡沫？

回答用户（人物设定）：回答用户：股神张宇韬

GPT-XL（机器自动生成）：2021年的股市应该是震荡向上的,我觉得现在不是该不该持有的问题,而是还要不要持有的问题。 从前两天直播聊天室里格林琴先生的发言,以及我昨天的复盘来看,格林琴先生应该是一个持有蔚来很长时间并且从蔚来转到了特斯拉的资深人士。 我还是推荐持有蔚来,原因有三。 第一,蔚来依然是新能源汽车行业的龙头之一,我个人非常看好。 第二,从格林琴先生的言语可以看出,他是一位优秀的泛科技投资人,这在近几年投资市场上很少见。他能够持有这么长时间,也不代表是抄底抄过去的,而是认为蔚来是一个优秀的企业,有着自己独特的竞争力。我认为蔚来至少在2020年还会有不错的投资机会。 第三,必须说,这次新能源汽车大会上,奔驰CEO宣称要投资全国,很可能是整个汽车业的投资格局的转变。要知道入局早是优势,这次投资汽车领域,带着多牛的投资背景,在未来的十多年里,蔚来会受益于汽车行业的繁荣。

# ● 超越图灵测试的AI作诗

<https://wudao.aminer.cn/turing-test/v2/>



The screenshot shows the '作诗图灵测试' (Poetry Turing Test) interface. At the top left, there is a decorative illustration of five birds in flight. The main title '作诗图灵测试' is centered. Below the title is a search bar containing the text '悟道 文源 文澜 文溯 文汇' and a right-pointing arrow. Underneath the search bar are four difficulty level buttons: 'Easy' (highlighted in green), 'Hard' (red border), 'Lunatic' (purple border), and 'Extra' (pink border). At the bottom, a text box explains the Easy mode: '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。'

# 超越图灵测试的AI作诗

<https://wudao.aminer.cn/turing-test/v2/>



## 悟道 V.S. 唐宋诗人

### 文源

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。

### V.S.

老翁岩畔卧，数载避喧惊。  
棋局尘生涧，茶烟景近屏。  
刀枪犹野外，吟啸亦琴清。  
更有深交意，相期岁晚行。

### 文澜

晨趋禁掖暮郊园，  
松桂苍苍烟霭繁。  
明月上时群动息，  
雪峰高处正当轩。

### V.S.

罗隐虽来频道寺，  
蓬门未肯便归休。  
殷勤与见我无倦，  
竟日相随不暂留。

### 文溯

初辞云浦寂，再别竺岩难。  
各分当时意，空驰海远寒。  
依稀真弟子，守至四明滩。  
临别殷勤语，重违誓去盘。

### V.S.

上国随缘住，来途若梦行。  
浮天沧海远，去世法舟轻。  
水月通禅观，鱼龙听梵声。  
惟怜一灯影，万里眼中明。

### 文汇

近年容尚白，卿士累资名。  
投契顾怜旧，照乘拟次行。  
常嫌官爵小，更睹道心清。  
多谢白公意，作为肺腑情。

### V.S.

太空秋色凉，独鸟下微阳。  
三径池塘静，六街车马忙。  
渐能高酒户，始是入诗狂。  
官冷且无事，追陪慎莫忘。

# 超越图灵测试的AI作画

<https://lab.aminer.cn/cogview/index.html>



金门大桥



一只戴  
墨镜的  
猫趴在  
桌子上



刘亦菲



一个人  
骑着自  
行车飞  
向月球



# 超越图灵测试的AI作画



<https://lab.aminer.cn/cogview/index.html>

克服“上下溢收敛”  
文图模型关键难题

*sandwich layernorm +  
precision bottleneck relax*

一个模型 实现两大能力

可直接取代OpenAI CLIP模型  
实现“模型自评分”

超分辨率

通过微调实现，提高图像精度

MS COCO 文本生成图像第一

超越 130 亿参数 OpenAI DALL-E



多种画风，样样精通



智源悟道 x 阿里达摩院  
服饰设计业务应用



# ● 超越图灵测试的AI视频制作



# 多项任务正逼近图灵测试

<https://wudao.aminer.cn/turing-test/v2/>



悟道作诗

悟道作词

悟道对联

悟道题名

悟道问答

悟道策论

悟道绘图

悟道识图

## 悟道绘图图灵测试

基于跨模态预训练模型的图文生成

开始挑战



游戏模式



Easy



Fast



① 在图灵绘图的Easy模式中，您将会被展现5组图片及其标题，每组包括2张图片，其中包含1张真实图片和1张由AI生成的图片，请挑选出真实的图片。

# “悟道”超大规模预训练模型

认知智能战略：让机器像人一样“思考”

慧业苦研虽久成，  
禅风散尽见真际。



“悟道”

悟道1.0作

何事喧哗取次行，虚庭昼晏坐禅室。  
云间鹤唳数声来，地下蜂喧惊起立。  
慧业苦研虽久成，禅风散尽见真际。  
如今洞有樵人迹，细细禅门意似痴。



# 技术篇

# “大炼模型”到“炼大模型”

## ◆ 新的技术范式

“大炼模型” (各自研发专用小模型) → “炼大模型” (研发超大规模通用智能模型)

## ◆ 超大模型 = 超大数据 + 超大算力 + 超大模型

- 参数规模大 ( 亿级以上 )
- 训练数据多 ( TB级别 )
- 依赖大规模算力支持 ( GPT-3训练的费用超千万美元 )

## ◆ “炼大模型” 优势

- **通用性强** : 通用模型 VS. 专用模型 : 预训练大模型经过微调或不用调参即可用于多个AI任务的实现 , 一般的AI模型通常只用于一种特定任务的实现
- **性能好** : 在很多实际任务上已大幅接近人类水平

# 「悟道」目标

瞄向世界顶尖水平，定位中国「第一」大模型

推动中国人工智能大模型的研发进程，带动科研、产业、经济等多方面发展  
促使大规模预训练模型变得更通用，同时探索国产超算适配之路

万亿模型

+

共建生态

在通往通用人工智能道路上，  
赋予机器像人一样“思考”的认知能力

# 「悟道」万亿模型的特点

01

中国首个全球最大万亿模型，1.75万亿参数规模

最大

02

一统文本与视觉两大阵地，支撑更多任务，更通用

通用

03

首次在100%国产超算上训练万亿模型

国产

04

中英双语，在4.9TB已清洗高质量数据上训练

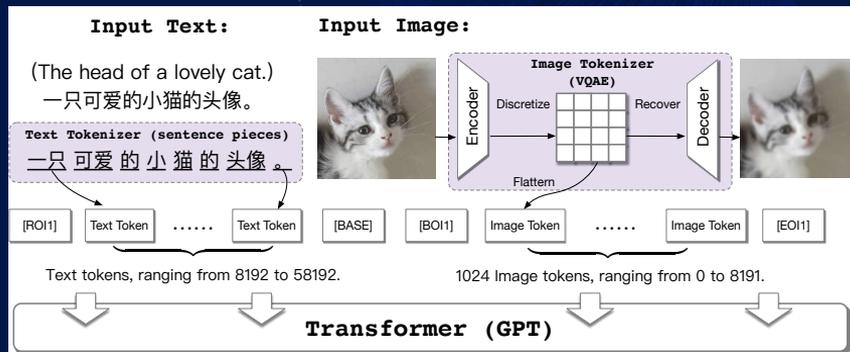
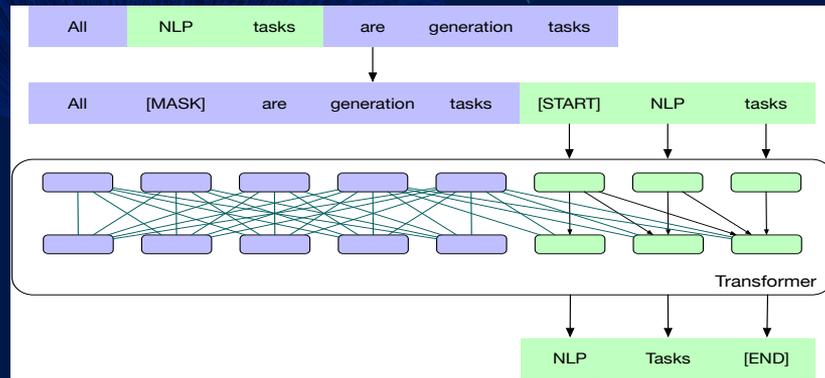
知识

# 「悟道」万亿训练细节总览

- ◆ 18层Transformer，隐变量维数4096，FFN中间层维数24576，Expert数量为480，参数量为1.75万亿，在神威国产超算上一周时间训练20000步，batch大小 20M tokens；
- ◆ 万亿模型基于悟道自研GLM框架 + CogView框架，能同时支持NLP理解、生成任务与文生图、图生文任务；
- ◆ 训练数据：
  - 中文数据：WudaoCorpora文本数据集、WudaoCorpora图文数据集
  - 英文数据：Pile数据集
- ◆ 使用基于 sentencepiece 的双语词表，大小200,000。

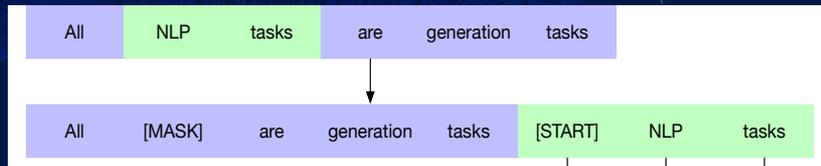
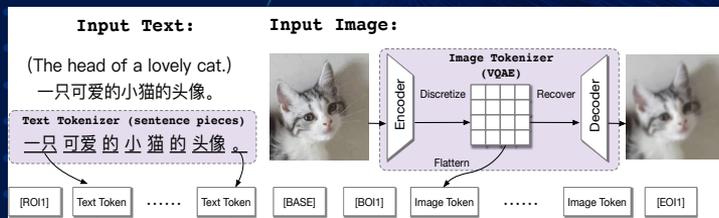
# 「悟道」 万亿架构设计

**GLM**：基于生成的通用预训练框架，以生成为核心，打破 BERT 和 GPT 瓶颈，同时在语言理解、生成和 Seq2Seq 任务上取得最佳性能



**Cogview**：文本生成图像的新框架，将VQ-VAE和Transformer进行结合，表现SOTA，在MS COCO FID指标上性能优于DALL·E等网络

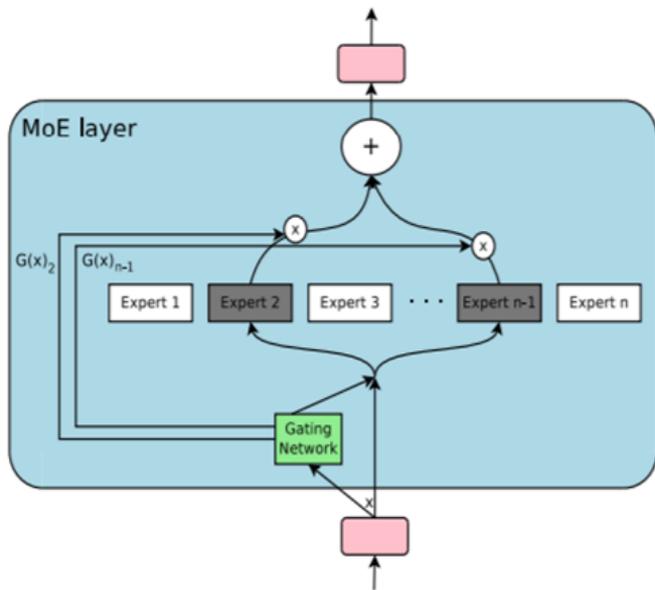
# 「悟道」万亿架构设计



# MoE 结构：扩充模型容量的新途径

万亿参数模型的基石

- 由  $n$  个专家 (expert) 组成专家池
  - 每个专家都可以是任意神经网络
- 对于每个输入  $x$ 
  - 由门网络 gate 对专家的适用性进行打分
  - 根据打分选择由哪些专家进行计算
- Switch Transformer使用MoE改造了Transformer的FFN模块，变相使得FFN的中间隐层变得十分宽
- 在计算量基本不变的情况下，大幅增加参数量



# 「悟道」 万亿核心基石

利用“悟道”团队自研的首个支持国产超算和GPU的“万亿核心基石”FastMoE训练“悟道”万亿模型

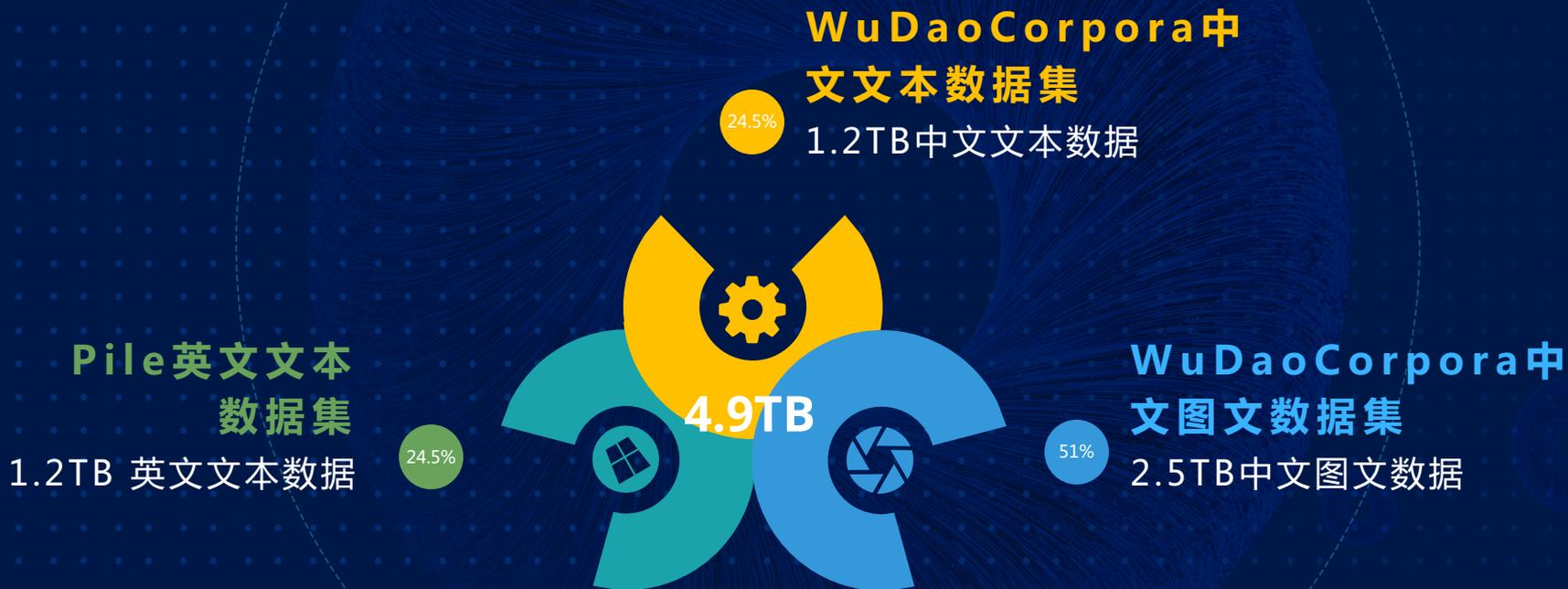


在神威国产超算上训练

## FastMoEv0.2

- ✓ 针对神威架构开发的定制化算子
- ✓ 针对神威网络拓扑设计的通信策略
- ✓ 最大测试了几万个专家的 MoE 训练

# 「悟道」 万亿训练数据





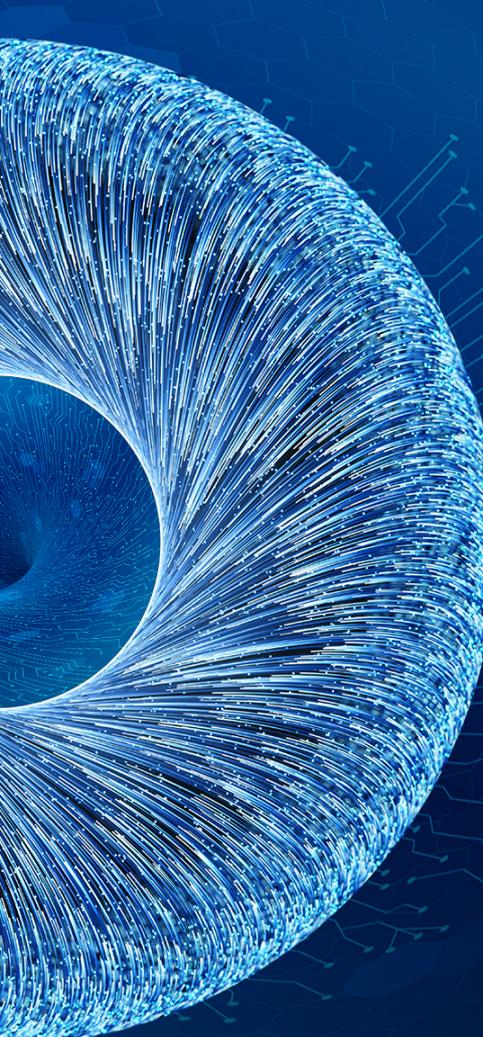
# 算法篇

# ● 核心算法

知识：P-Tuning微调法

可控：Inverse Prompting生成

通用：GLM预训练框架



# All NLP Tasks are Generation Tasks: A General Pretraining Framework



<https://arxiv.org/abs/2103.10360>

# ● 预训练框架的通用性问题

Framework	NLU	Cond. Gen.	Uncond. Gen.
Autoregressive	—	—	✓
Autoencoding	✓	×	×
Encoder-Decoder	—	✓	—
GLM	✓	✓	✓

之前，没有一个通用预训练框架可以同时**在理解任务、有条件生成任务、无条件生成任务取得最优。**

“✓”表示擅长，“—”表示可以做，“×”表示无法直接应用。

# 所有NLP任务都是生成任务

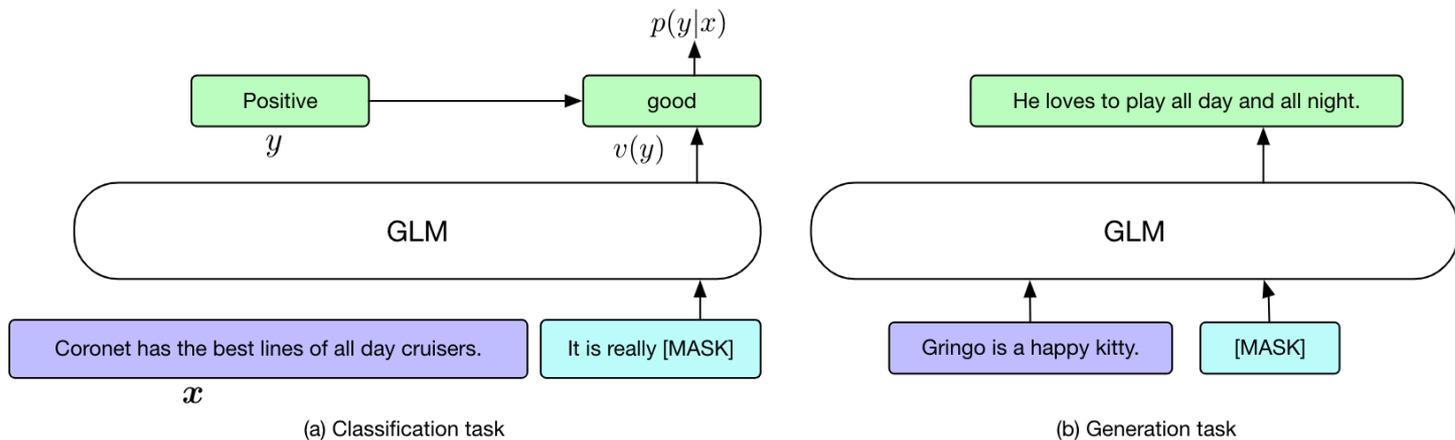


Figure 3. GLM finetune framework. (a) Formulation of the sentiment classification task as blank infilling with GLM. (b) GLM for text generation given the context. This can be the language modeling in the zero-shot setting, or seq2seq with fine-tuning.

分类理解任务、有条件生成任务、无条件生成任务，都可以统一用生成式框架刻画。

# 新的预训练范式

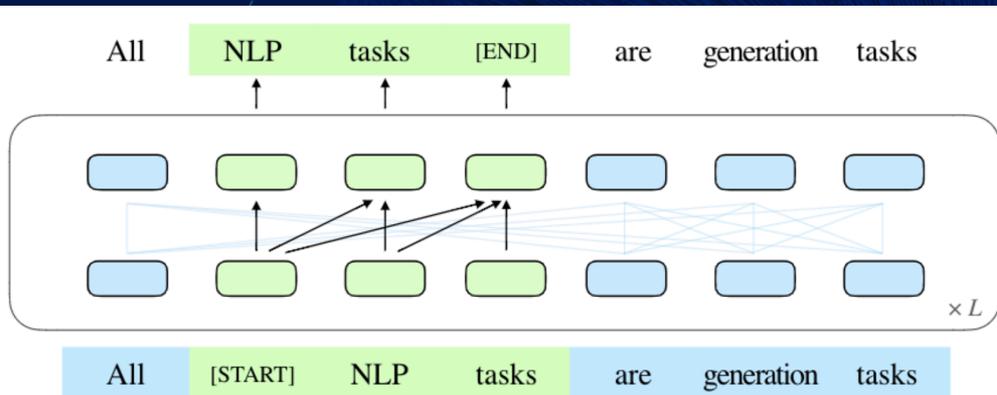


Figure 1. Illustration of GLM. We blank out text spans (green part) and GLM is trained to generate them in an autoregressive fashion.

多任务训练 (可选) :

1. Sample中间15%作为生成目标
2. Sample末尾50-100%作为生成目标

# GLM预训练方法

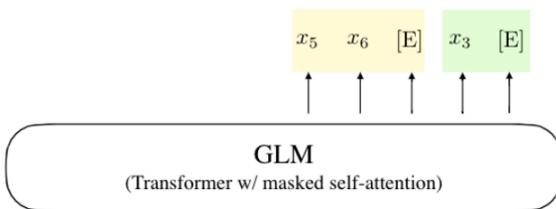
$x_1$   $x_2$   $x_3$   $x_4$   $x_5$   $x_6$

(a) Sample spans from the input text

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

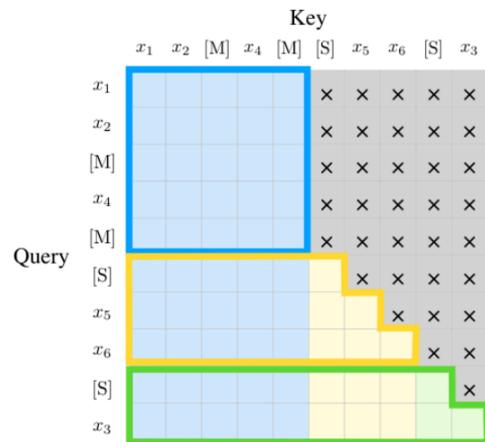
Part B:  $x_5$   $x_6$   $x_3$

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



Position 1	1	2	3	4	5	5	5	3	3
Position 2	0	0	0	0	1	2	3	1	2

(c) Generate the Part B spans autoregressively



(d) Self-attention mask

# 实验结果：分类任务

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

Model	ReCoRD F1/Acc.	COPA Acc.	WSC Acc.	RTE Acc.	BoolQ Acc.	WiC Acc.	CB F1/Acc.	MultiRC F1a/EM	Avg
BERT <sub>Base</sub>	65.4/64.9	66.0	65.4	70.0	74.9	<b>68.8</b>	70.9/76.8	68.4/21.5	66.1
GLM <sub>Base</sub>	<b>73.5/72.8</b>	<b>71.0</b>	<b>72.1</b>	<b>71.2</b>	<b>77.0</b>	64.7	<b>89.5/85.7</b>	<b>72.1/26.1</b>	<b>70.7</b>
BERT <sub>Large</sub>	76.3/75.6	69.0	64.4	73.6	80.1	<b>71.0</b>	94.8/92.9	71.9/24.1	72.0
UniLM <sub>Large</sub>	80.0/79.1	72.0	65.4	76.5	80.5	69.7	91.0/91.1	77.2/38.2	74.1
GLM <sub>Large</sub>	81.7/81.1	76.0	<b>81.7</b>	74.0	<b>82.1</b>	68.5	96.1/94.6	<b>77.1/36.3</b>	77.0
GLM <sub>Large</sub> (multi-task)	80.2/79.6	77.0	78.8	76.2	79.8	63.6	<b>97.3/96.4</b>	74.6/32.1	75.7
GLM <sub>410M</sub> (multi-task)	81.5/80.9	80.0	<b>81.7</b>	<b>79.4</b>	81.9	69.0	93.2/96.4	76.2/35.5	78.0
GLM <sub>515M</sub> (multi-task)	<b>82.3/81.7</b>	<b>85.0</b>	<b>81.7</b>	79.1	81.3	69.4	95.0/96.4	77.2/35.0	<b>78.8</b>
T5 <sub>Base</sub>	76.2/75.4	73.0	79.8	78.3	80.8	67.9	94.8/92.9	76.4/40.0	76.0
T5 <sub>Large</sub>	85.7/85.0	78.0	<b>84.6</b>	84.8	84.3	71.6	96.4/98.2	80.9/46.6	81.2
BART <sub>Large</sub> *	88.3/87.8	60.0	65.4	84.5	84.3	69.0	90.5/92.9	81.8/48.0	76.0
RoBERTa <sub>Large</sub> *	89.0/88.4	<b>90.0</b>	63.5	87.0	<b>86.1</b>	<b>72.6</b>	96.1/94.6	<b>84.4/52.9</b>	81.5
GLM <sub>RoBERTa</sub>	<b>89.6/89.0</b>	82.0	83.7	<b>87.7</b>	84.7	71.2	<b>98.7/98.2</b>	82.4/50.1	<b>82.9</b>

- 按照计算量进行分组，每组公平比较
- 超越BERT、T5、RoBERTa等模型

# 实验结果：无条件、有条件生成

Table 3. Results on Gigaword abstractive summarization

Model	RG-1	RG-2	RG-L
MASS	37.7	18.5	34.9
UniLM <sub>Large</sub>	38.5	19.5	35.8
GLM <sub>Large</sub>	38.6	19.7	36.0
GLM <sub>Large</sub> (multi-task)	38.5	19.4	35.8
GLM <sub>410M</sub> (multi-task)	<b>38.9</b>	<b>20.0</b>	<b>36.2</b>

## Before

训练三个不同的模型

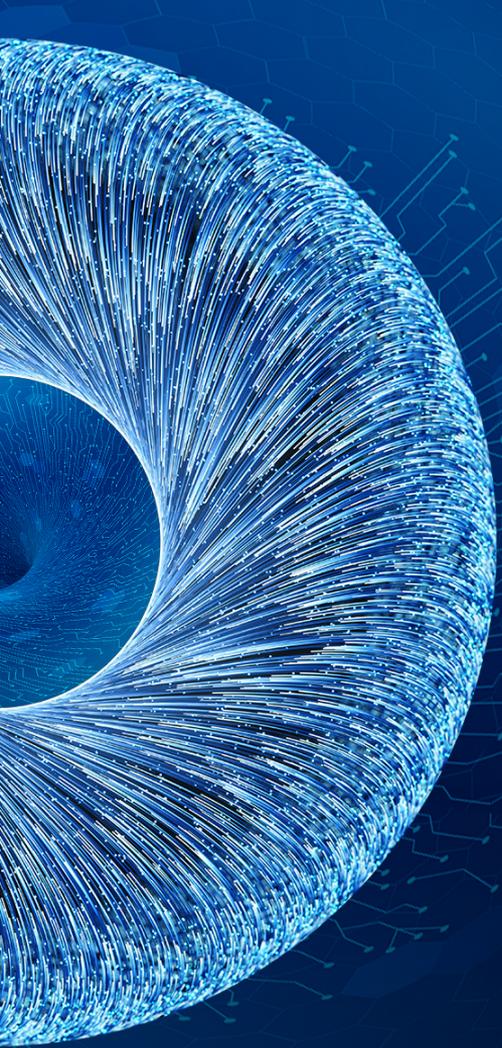
## After

1.25倍的GLM可以同时在这三种任务上取得最优！

历史上首个单一模型同时在这三种任务上取得最优效果。

Table 4. Zero-shot language modeling results.

Model	Lambada (Accuracy)	BookWiki (Perplexity)
GLM <sub>Large</sub> (uni)	0.0	> 100
GLM <sub>Large</sub> (multi-task, uni)	47.4	15.1
– 2d positional encoding	45.8	15.1
GLM <sub>410M</sub> (multi-task, uni)	49.5	14.5
GLM <sub>515M</sub> (multi-task, uni)	<b>50.4</b>	<b>13.9</b>
GLM <sub>Large</sub> (bi)	10.6	> 100
GLM <sub>Large</sub> (multi-task, bi)	48.5	14.9
– 2d positional encoding	47.3	15.0
GLM <sub>410M</sub> (multi-task, bi)	<b>53.5</b>	<b>14.3</b>
GLM <sub>515M</sub> (multi-task, bi)	<b>54.9</b>	<b>13.7</b>
GPT <sub>Large</sub> (uni)	50.1	14.4



# GPT Understands, Too



<https://arxiv.org/abs/2103.10385>

# Prompt的进化：

## 从手写到自动，从离散到连续

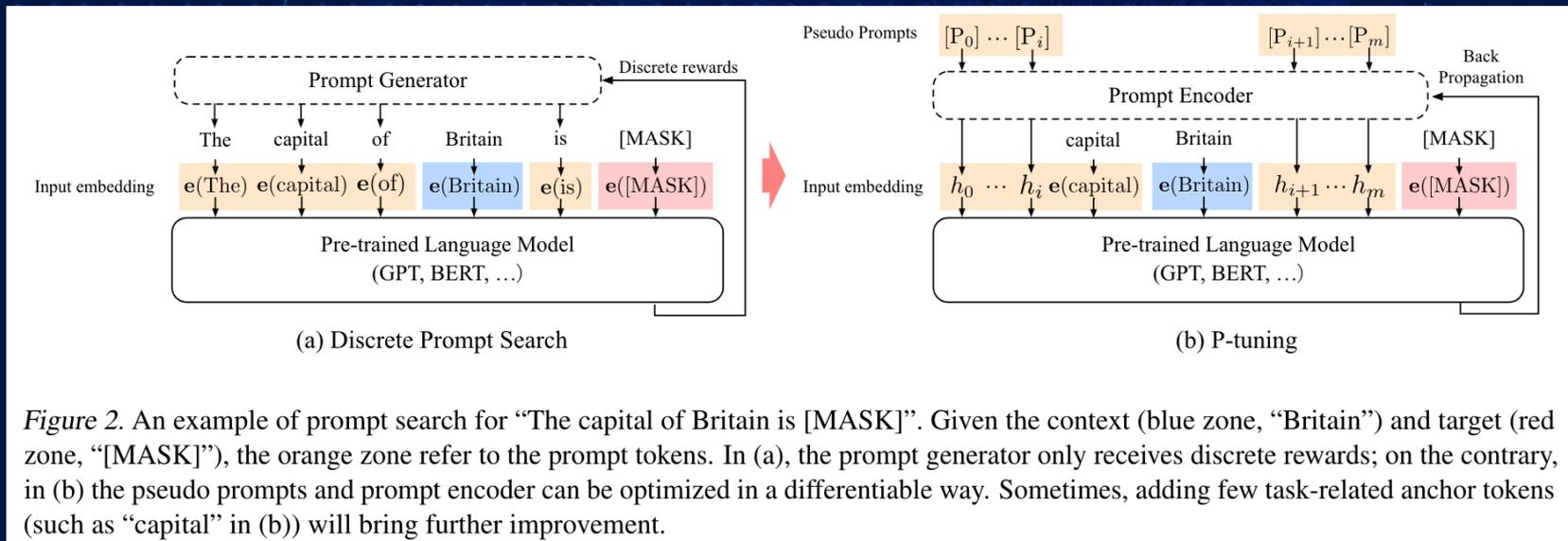
- 手写Prompt (Brown et al 2020)
- 自动离散Prompt (Jiang et al 2020, Shin et al 2020, Gao et al 2020)
- 自动连续Prompt (P-Tuning)

Prompt	P@1
[X] is located in [Y]. ( <i>original</i> )	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

Table 1. Case study on LAMA-TREx P17 with bert-base-cased. A single-word change in prompts could yield a drastic difference.

- 离散prompt对扰动非常敏感
- 在few-shot场景下容易overfit dev/test sets

# P-Tuning



学习prompt的两大挑战：超越局部优化、关联性  
用轻量神经网络学习prompt embedding：两层LSTM+MLP

# 知识探测 (LAMA) 实验结果

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

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

\* 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 ).

- 知识探测任务取得世界第一 ( SotA ) ，超越之前最好结果20百分点
- P-tuning相比于finetuning和离散prompt的方法，有一致、显著的提升

# 少样本学习实验结果

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

\* We report the average and standard deviation of each candidate prompt's average performance.

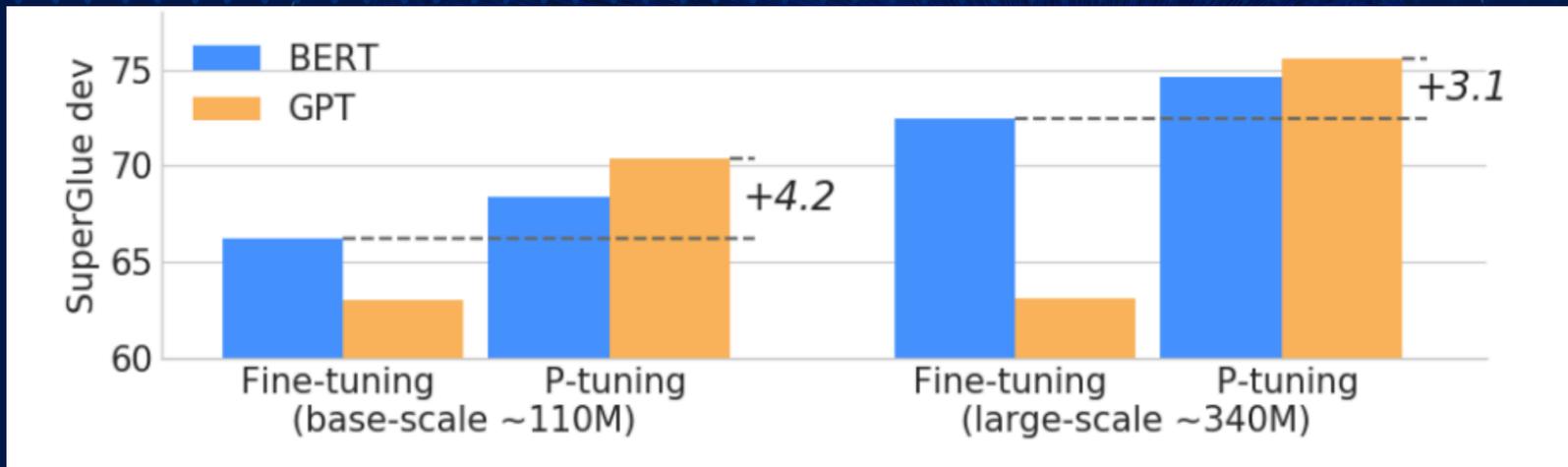
<sup>†</sup> We report the best performed prompt selected on *full* dev dataset among all candidate prompts.

<sup>‡</sup> With additional ensemble and distillation.

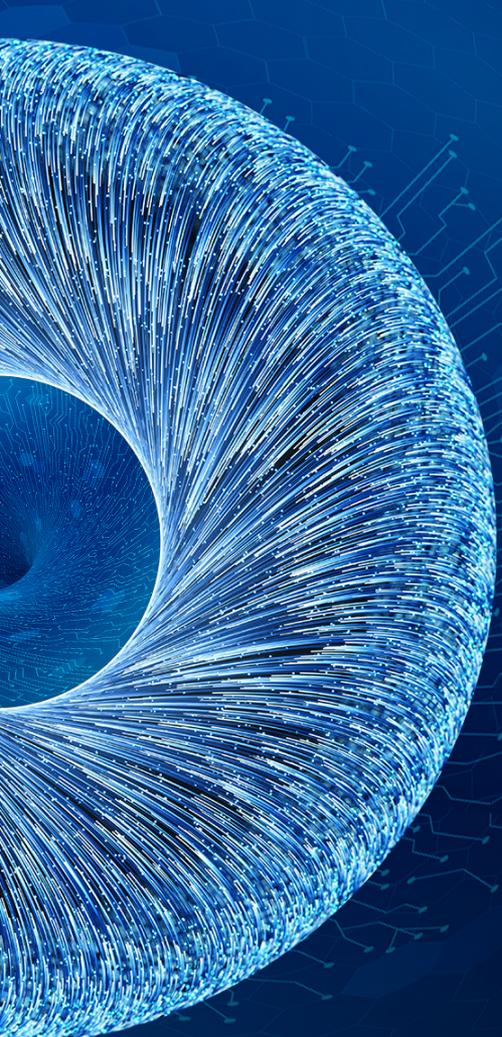
<sup>§</sup> With additional data augmentation, ensemble, distillation and self-training.

- 少样本学习任务取得世界第一 ( SotA ) , 超越之前最好结果高达14百分点
- 亿级别的模型 ( 仅使用dev32 ) , 通过P-Tuning , 显著超越千亿GPT-3的表现

# P-Tuning让GPT达到 与BERT相同的分类任务水平



- 使用finetuning的情况下，GPT比BERT低10个百分点
- 使用P-Tuning的情况下，GPT比BERT高1个百分点
- P-Tuning对于BERT和GPT都有显著的提升



# Controllable Generation from Pretrained Language Models via Inverse Prompting



# ● 基于Prompting的文本生成不可控

问题：什么瞬间让你想留住这一刻？

直接生成：

我和男朋友在一起两年。我去国外留学,他留在国内。一人在国外,寂寞心酸都是真的。念书的时候,觉得常在一起也没什么,两个人一起努力,吵吵闹闹就过来了。我一直这样想,以至于到了高中,到了最后几个月,我都觉得他就是我行走的小弟儿。其实最快乐的时候,就是一起在外头的“孤儿院”里。

**关联性很弱！**

# ● 基于Inverse Prompting的优化

问题：什么瞬间让你想留住这一刻？

Inverse Prompting生成：

一瞬间想到高二的时候吧,那天下午在操场晒了一中午太阳,和同学一起在楼下打球,下午的夕阳徐徐洒下来,我们三个人一起站在走廊上,倚着栏杆可以面对着夕阳。一直想留下那一瞬间,可惜我以后再也没遇到过那样的时光。

# ● Inverse Prompting思路

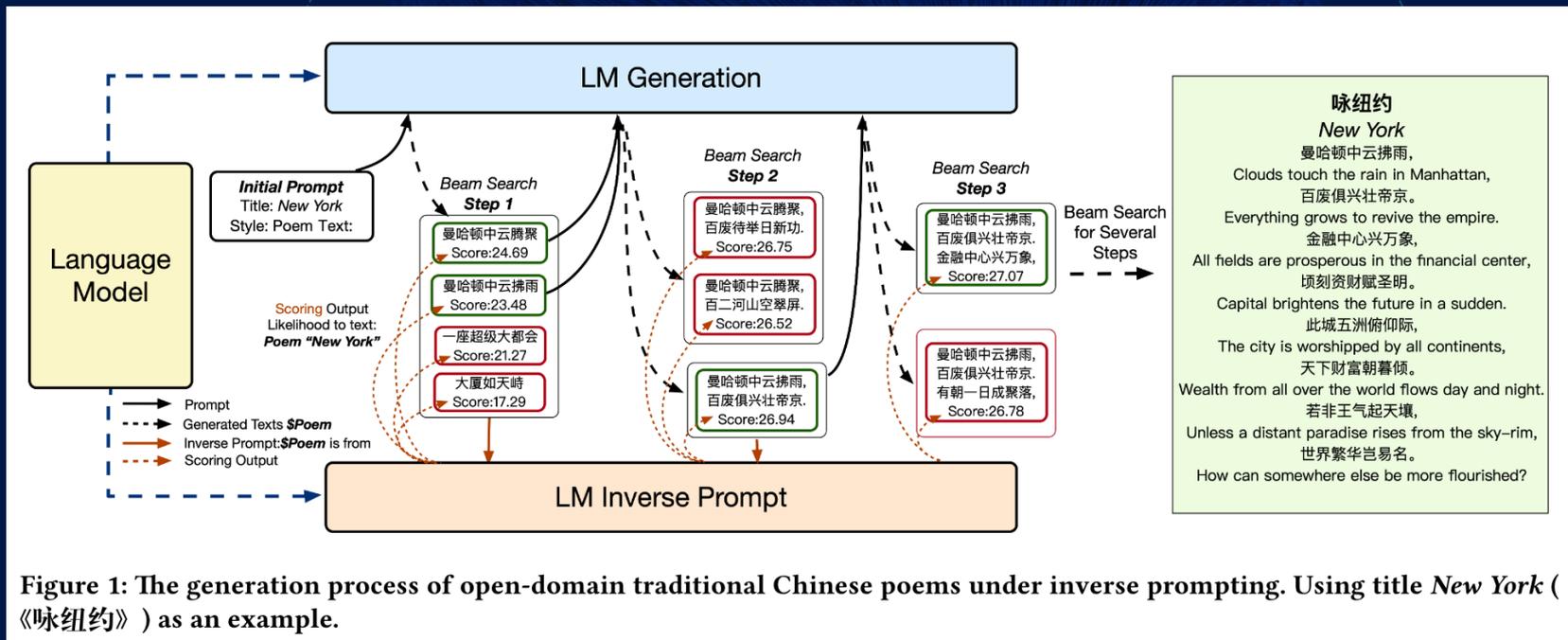
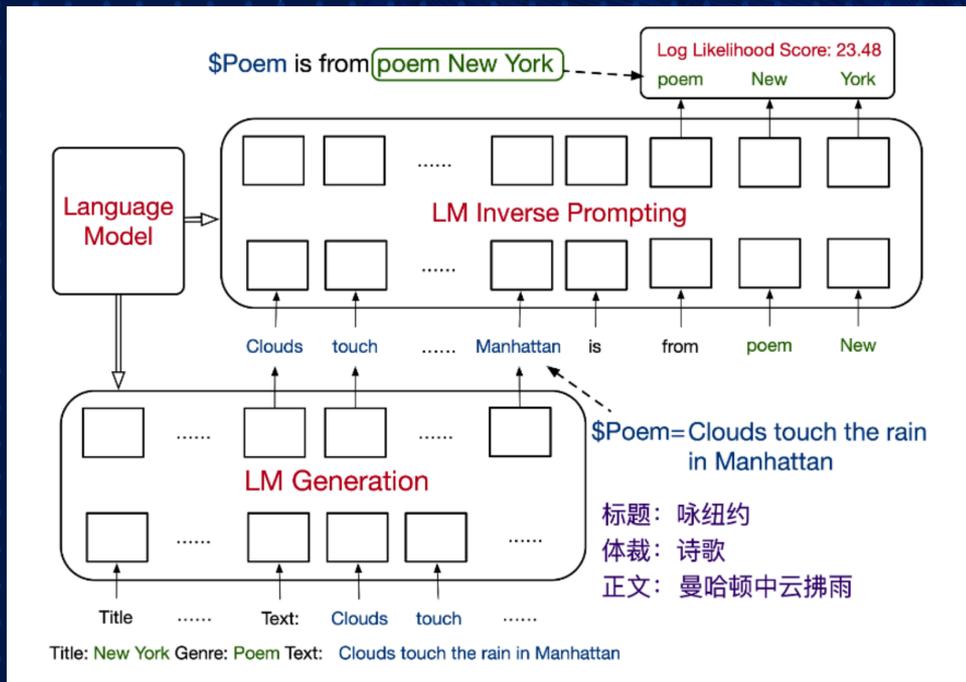


Figure 1: The generation process of open-domain traditional Chinese poems under inverse prompting. Using title *New York* (《咏纽约》) as an example.

用生成的内容反过来预测prompt

# ● Inverse Prompting技术



- Beam search
- Log likelihood as score

# ● 现代概念作古体诗

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。

# ● 现代概念古体诗

夜过虹桥机场

卢浦斜晖里，西楼醉客行。

影侵双塔晚，灯落一城明。

空客还频顾，航灯未可惊。

欲留城市夜，月映水帘星。

# ● 现代概念古体诗

咏相对论

向来言物理，不似客心惊。

奇怪乾坤秘，翻飞斥几星。

超光无敌速，瞬息至昆冥。

引力张鞭势，牯牛曳著行。

# ● 现代概念古体诗（藏头诗）

单身狗

清水饼干难下咽，华歌月酒不能尝。

大龄蛋白宅男女，学界萧条剩贱郎。

- 在beam search时固定藏头字
- 风格偏直白，有强烈讽刺意味

# ● 问答

人注定要死，为什么还要活着？

AI的回答：人终究是要死的，我们活着的意义是寻觅人生的价值，体验生命的壮丽。为了追求这种生活，人们愿意奋斗终生。

# 量化评价：问答

Method	Fluency (1-5)	Inform. <sup>1</sup> (1-5)	Relevance (1-5)	Overall (1-10)
CPM [27]	2.66	2.47	2.36	4.32
Prompting Baseline	3.44	3.25	3.21	5.97
Inverse Prompting	<b>3.61</b>	<b>3.43</b>	<b>3.59</b>	<b>6.51</b>
Human Answers	3.80	3.61	3.67	6.85

<sup>1</sup> Informativeness

# 量化评价：作诗

Method	Format (1-5)	Innov. <sup>1</sup> (1-5)	Relevance (1-5)	Aes. <sup>2</sup> (1-5)	Overall (1-10)
Jiuge [28]	<b>3.60</b>	2.47	1.99	<b>3.12</b>	3.57
Search Baseline	2.79	1.10	1.16	2.44	1.35
Inverse Prompting	2.56	2.71	2.92	2.33	4.00
Inverse Prompting +ST	2.42	<b>2.92</b>	<b>3.65</b>	2.18	<b>4.40</b>

<sup>1</sup> Innovation

<sup>2</sup> Aesthetics

# ● 作诗图灵测试

Method	Total	Selected	Selection Rate
Inverse Prompting +ST	1,656	748	45.2%
Ancient Human Poems	1,656	908	54.8%

- 与古代诗人竞赛，得票率45.2%
- 接近大诗人作诗水平



# 生态篇

# 「悟道」生态

## 产业生态

与各领域头部企业合作，构建示范应用，通过领头企业的完整产业链，对示范应用进行评估，发掘市场化需求，为大模型生态构建提供指导。

开展悟道应用大赛，发掘潜在应用方向，构造多项图灵测试，挑战“以假乱真”，测试模型能力极限。

## 应用生态

## 开源生态

推动模型、算法、工具、API等多维度社区开源，企业级用户、个体开发者、研究机构等可根据自身研发需求使用悟道社区版开源成果。

构建高质量数据集，支撑悟道项目研究，覆盖文本、视觉和对话等领域，推动中国通用人工智能领域发展。

## 数据生态



# 「悟道」产业生态-战略合作

目前已与**23家产业单位**达成战略合作意向，就悟道大规模预训练模型进行产业应用合作，从底层数据资源、中层模型研发到上层AI应用推进悟道产业生态的建设，从单点突破提升至全方位突破



# 「悟道」开源生态-多维度开源

## 模型

- Cogview
- GLM
- CPM2
- 中文对话
- ProteinLM
- Lawformer

## 算法

P-tuning  
Prompt  
Inverse prompt  
...

## 开源

## 工具

FastMoE  
3D训练加速TDS  
中文编码  
推理加速  
...

## API

作诗	图片生成	文本向量
作词	对联	图片向量
问答	文本生成	图谱构建
新闻生成	故事生成	概念检索

推动大模型成果共享，吸引社会科研力量加入“悟道”生态建设

# 「悟道」数据生态-WuDaoCorpora

## WuDaoCorpora2.0

纯文本

多模态

对话

- 世界最大中文文本数据集；
- 从**100TB**原始数据清洗得到共**3TB**数据（相对1.0新增**1TB**）；
- 包含教育、科技等**50+**个行业的数据标签。

- 全球最大的多模态数据集；
- 图文对总量达到**6.3亿**（总量约**90TB**）；
- 其中**3000万**的文字内容是对图片的详细描述

- 全球最大的中文对话数据集；
- 体量达到了**181GB**；
- 采用市面最严格清洗逻辑，从**9TB**原始数据筛选出**181GB**。

为悟道生态提供坚实可靠的数据支撑

# 「悟道」应用生态-悟道大赛

## 大赛定位

“未来”技术先体验，  
探索智能本质

创新+应用+研究

源创计划支持，百  
万总奖金

## 大赛赛道

高校：结队比拼创  
意，初赛提交项目  
计划书，不需编写  
代码

企业：开发基于悟  
道的软硬件系统

研究&开源：提交预  
训练研究计划，或  
参与悟道开源社区  
建设

## 大赛主题

健康医疗

教育学习

社交生活

效率工具

环境自然

## 大赛日程

赛题发布：6月

网络初赛：7月-10

月

线下复赛：10月

决赛颁奖：11月14

日



# ● 展望

让机器像人一样“思考”：超越图灵测试的通用机器认知能力

认知  
AI

1. 适应与学习能力
2. 定义与语境化能力
3. 自我系统的准入能力
4. 优先级与访问控制能力
5. 召集与控制能力
6. 决策与执行能力
7. 错误探测与编辑能力
8. 反思与自我监控能力
9. 条理与灵活性之间的能力

探索更强大、更通用模型

探索面向认知的预训练模型

探索新的生态合作模式

# 致谢每一位悟道科学家！

## 悟道·文源

以中文为核心的超大规模预训练语言模型

刘知远 黄民烈 韩文弢 刘洋 朱小燕 孙茂松  
张正彦 顾煜贤 韩旭 陈晟祺 肖朝军 姚远 岂凡超 关键 柯沛 周昊 孙桢波 蔡严正 曾国洋 谭知行 秦禹嘉 苏裕胜 司程磊  
胡雪瑜 李文浩 王凤玉 易婧 王晓智 陈晔泽 丁宁 张家杰

## 悟道·文澜

超大规模多模态预训练模型

文继荣 宋睿华 卢志武 金琴 赵鑫 庞亮 兰艳艳 窦志成  
高一钊 霍宇琦 卢浩宇 温静远 杨国兴 宋昊暘 张曼黎 张良 胡安文 李瑞晨 宋宇晴 赵金明 赵一达 费楠益  
孙宇冲 金楚浩 洪鑫 崔婉清 侯丹阳 李英彦 隰宗正 刘光镇 刘沛羽 龚政 李军毅

## 悟道·文汇

面向认知的超大规模新型预训练模型

唐杰 杨植麟 杨红霞  
杜政晓 丁铭 邹旭 裘捷中 钱雨杰 殷达 钟清扬 于济凡 刘潇 郑亚男 何家傲 曾奥涵 洪文逸 杨卓毅 郑问迪 周璟 杜冀中  
郭子通 刘静 周畅 林俊旸

## 悟道·文溯

超大规模蛋白质序列预测预训练模型  
FastMoE和万亿大模型

唐杰 鲁白  
裘捷中 谢昌谕 肖易佳 曾奥涵 李子昂  
唐杰 翟季冬 杨红霞 陈文光 郑纬民  
马子轩 何家傲 裘捷中 曹焕琦 王元炜 孙桢波 郑立言 王豪杰 唐适之 冯冠宇 曾奥涵 钟闰鑫 师天麾 杜政晓  
丁铭 Tiago Antunes 彭晋钧 林俊旸 张健伟

# Wudao—Pretrain the world

感 谢 关 注

