

人工智能专题研究

# AIGC投资框架

西南证券研究发展中心  
计算机研究团队  
2023年5月

## 核心观点

---

- **自然语言模型是大语言模型的核心。**自然语言模型是大模型的核心，支持许多类型的语言任务的执行。ChatGPT和各大主流AIGC模型主要基于大语言模型，通过深度学习算法来处理和生成自然语言。自然语言模型发展经历了基于规则、基于统计和基于神经网络三个阶段，Transformer的出现为以Bert和GPT-3为代表的大语言模型的诞生奠定了模型基础。
- **AI到AGI，ChatGPT开启人工智能新纪元。**ChatGPT结合自回归语言模型+zero shot prompting，带来AI范式突破，实现AI到AGI的转变。相比fine tuning技术，prompting技术能更好的完成生成类任务。ChatGPT在RLHF的帮助下，找到了GPT-3.5和人类自然语言的合理接口，解锁了模型的应用场景，改变了人机交互方式。
- **GPT跨越式发展，探索未来更多可能性。**各代GPT青出于蓝胜于蓝，更新迭代过程中涌现了诸多能力。GPT-4.0更是突破了GPT-3.5仅支持执行文本类任务的局限性，赋予ChatGPT更优异的上下文理解能力、推理能力、演绎能力，甚至图文理解能力。GPT也在不断的更新迭代中持续突破，解锁更多能力。
- **多行业结合，下游应用遍地开花。**ChatGPT以其高级的生成能力和复杂任务处理能力，以从重资产的模型层到轻资产的中间层和应用层为架构，为游戏行业、视频类行业、生物医药行业等其他行业锦上添花。在ChatGPT的帮助下，各行各业真正的做到了提高效率、减少成本并实现了功能上质的改变，让更多用户受益。
- **风险提示：**AI技术更新迭代缓慢、专业领域落地效果不及预期、版权纠纷、伦理问题、算力建设缓慢等风险。

# 目录

---



## 1 大模型与GPT

- 1.1 大模型篇
- 1.2 ChatGPT篇
- 1.3 GPT-4篇



## 2 国内篇

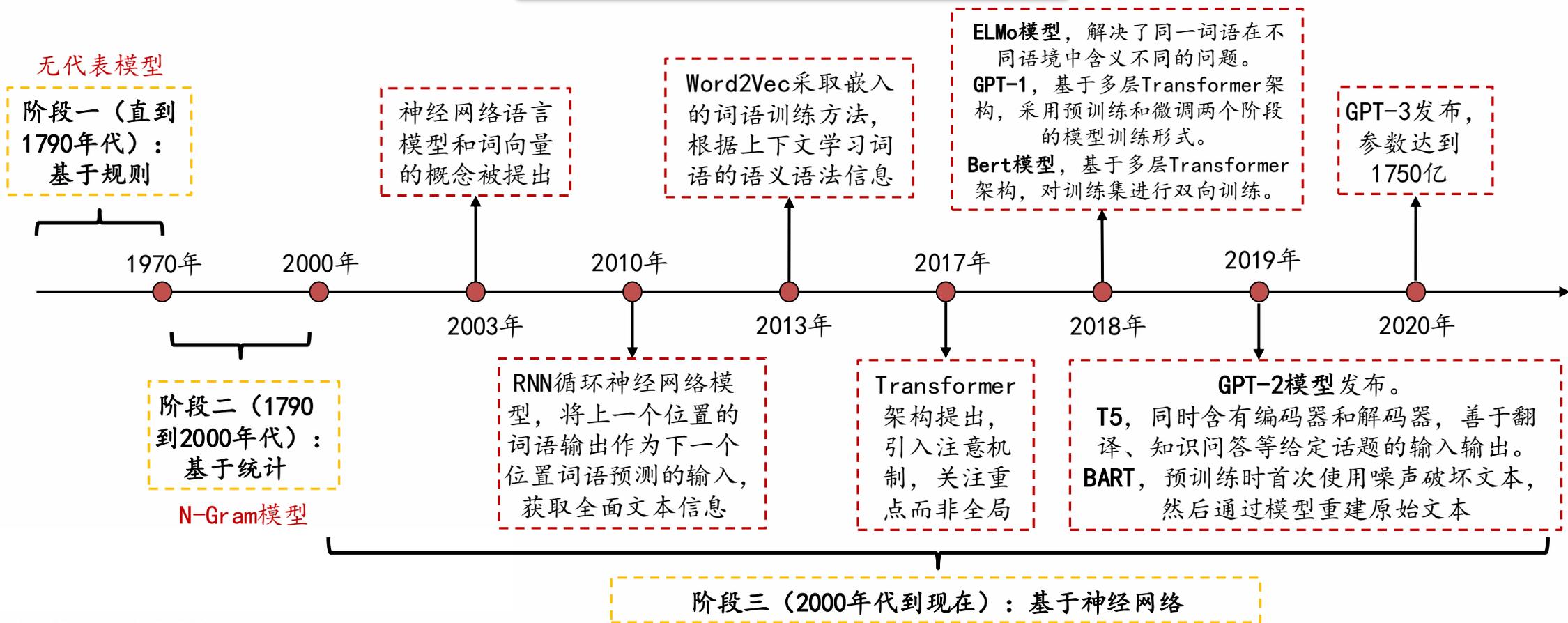


## 3 ChatGPT应用场景

## 1.1.1 大语言模型简介

大语言模型简称LLM ( LargeLanguageModels )，是基础的机器学习模型，使用深度学习算法来处理和生成自然语言。这些模型在大量的文本数据上进行训练，再自动生成符合语言规律的文本。大语言模型可以执行许多类型的语言任务，如翻译语言、分析情感、聊天机器人对话等。

### 自然语言模型的发展历史



## 1.1.2 大语言模型的主要发展历程

### ➤ 第一阶段 (2013-2020)



**在Bert和GPT模型问世之前：**深度学习模型是自然语言处理（NLP）领域主流的技术，而深度学习在自然语言处理领域的应用主要依赖于以下关键技术：**使用大量改进的LSTM模型和少量改进的CNN模型作为典型的特征提取器；使用序列到序列（Sequence to Sequence）（也称为编码器-解码器（Encoder-decoder））加上注意力作为各种具体任务的技术框架。但这个技术拥有两大主要问题阻碍了深度学习领域的突破：第一，训练数据的数量有限。随着容量和深度的增加，模型需要更多的训练数据支持。第二，这项技术的表达能力不够强。即使拥有足量数据，也很难有效的吸收数据内容。**

**在Bert和GPT模型问世之后，出现了两大主要范式转换：**第一个范式转换体现在预训练模型出现并有两大主要影响，体现在：**其一，中间任务消失。**通过大量数据预训练，Bert和GPT把中间任务作为语言特征，吸收到Transformer的参数里，做到直接解决最终任务。**其二，统一了不同研究方向技术路线。**最终任务包括了自然语言理解类任务和语义生成类任务。**对于理解类任务，其技术体系统一到以Bert为代表的双向语言模型加Fine Tuning模式；对于生成类任务，其技术体系统一到以GPT2.0为代表的自回归语言模型加zero shot prompting（GPT-3）或few shot prompting（ChatGPT）模式。**

**第二个范式转换为从预训练模型逐渐偏向于通用人工智能，简称AGI（Artificial General Intelligence），我们目前正处于这个转换的以GPT3.0（自回归语言模型+zero shot prompting）占据主导位置为代表的过渡过程中。**自回归语言模型+zero shot prompting处于目前的主导地位的原因有两个：**其一，比如Google的T5模型，在自然语言生成任务中可以同时完成自然语言理解任务。**如仍使用类似Bert的理解类任务语言模型，则无法得到优质的生成内容。所以相比于理解类任务模型，目前来说大众更偏向于生成类任务模型。**其二，GPT模式效果更优。**若想以zero shot prompting或few shot prompting模式完成任务，必须采取GPT模式，也就是生成类任务模型。Prompting技术也更加符合人性，相比之下，fine tuning模式是被动适应模型。

另外，zero shot prompting模式很难被大语言模型所理解，表现效果也差强人意，以目前的技术很难达到令人满意的效果。**所以目前科学家偏向于研究few shot prompting技术，起到过渡作用。**

### ➤ 第二阶段 (2020.06-至今)

## 1.1.3 大语言模型学习、存储、修正数据的方式

### 大语言模型学习数据的方式

- 大语言模型从海量的文本数据中学习了大量知识。**对于其学习到的知识一共有两类：语言类和世界知识类。**
- **浅层语言类知识**储存在Transformer的低层和**中层**，而**相对抽象的语言类知识**储存在Transformer的**中层或高层**。大语言模型在学习过程中不需要过多的语言类知识数据。
- **世界知识类知识**可以分为**两类：事实型和常识型**。世界知识主要储存在Transformer的**中高层**，**同时大语言模型学习需要大量事实型世界知识。**

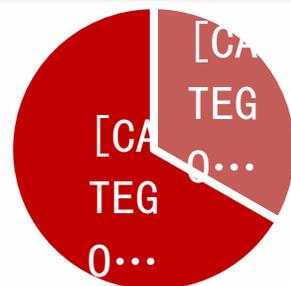
### Transformer知识金字塔



### 大语言模型储存数据的方式

- 大语言模型会将学习过的知识储存在Transformer的模型参数里。
- Transformer的模型参数结构分为两个部分：其中，**多头注意力（MHA）的部分占参数总数的三分之一；FFN结构占参数总数的三分之二。**
- MHA主要用于计算单词与知识间的关联强度，不用来存储知识。**FFN存储着大量知识**，是大语言模型的key-value存储器。

### Transformer参数结构分布



### 大语言模型修正数据的方式

- 大语言模型有很大可能会生成错误答案，所以修正数据非常必要。大模型有三种修正方法。
- **第一种是从训练数据源头去修正。**但在每次修正后需要重建模型，成本过高，无发展前景。
- **第二种是通过finetuning修正知识。**但这种方法容易误删正确数据，会影响效率。
- **第三种是直接修改知识对应的模型参数。**直接定位到储存旧知识的FFN节点并替换新知识。

### 精准定位FFN节点



## 1.1.4 大语言模型的分类

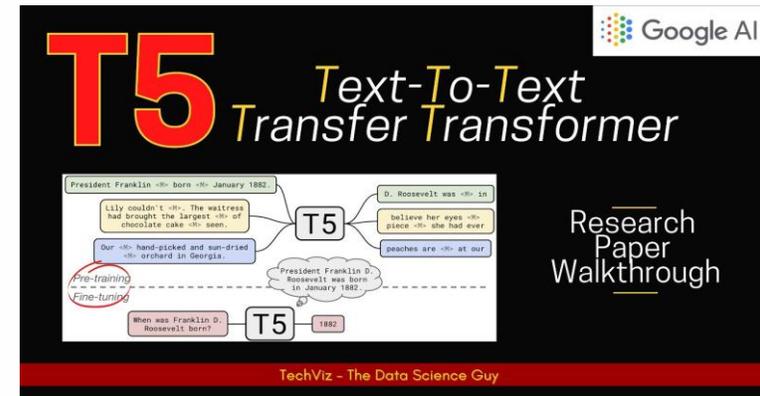
大语言模型按模型结构一共分为以下三类：

**1.以BERT系列为代表的encoder-only模型：**在每个阶段，注意力层可以访问初始句子中的所有单词并具有双向注意力。这类模型的预训练通常围绕着以某种方式破坏一个给定的句子，并让模型承担寻找或重建初始句子的任务。Encoder-only模型最适合于需要理解完整句子的任务。



**2.以GPT为代表的decoder-only模型：**在每个阶段，对于一个给定的词，注意力层只能访问句子中定位在它之前的词。这类模型通常被称为自回归模型。解码器模型的预训练通常围绕着预测句子中的下一个词进行。Decoder-only模型最适合于涉及文本生成的任务。

**3.以T5为代表的既有encoder、又有decoder的模型：**在每个阶段，encoder的注意力层可以访问初始句子中的所有单词，而decoder的注意力层只能访问输入中位于特定单词之前的单词。这类模型的预训练通常涉及到更复杂的任务。Encoder-decoder模型最适合于围绕根据给定的输入生成新的句子的任务，如总结、翻译或生成式问题回答。



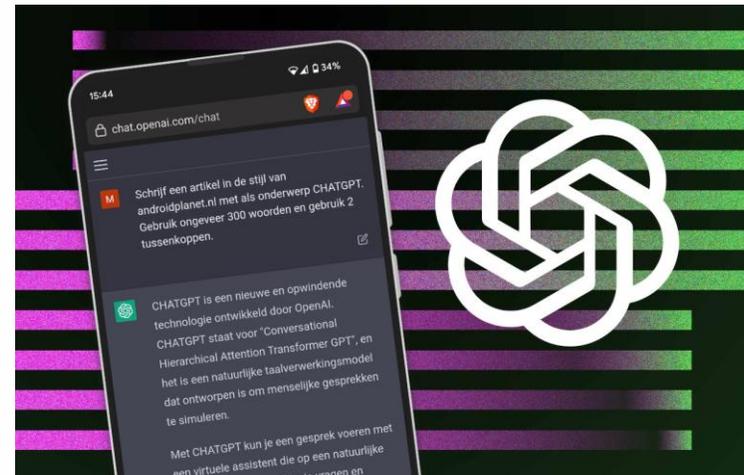
## 1.1.4 大语言模型的分类

随着深度学习技术的不断发展，大型模型在自然语言处理领域发挥着越来越重要的作用。大型模型通过使用更多的参数、更复杂的结构和更多的训练数据来提高性能，从而在自然语言处理任务中取得了显著的成果。**语义理解是自然语言处理模型的基础任务之一。**Bert和其变体模型被广泛用于各种语义理解任务，例如文本分类、命名实体识别和自然语言推理等。**文本生成也是自然语言处理模型中的另外一个重要任务。**在自然语言生成任务中，GPT已成为最典型的框架之一，因为它能够生成高质量的文本，并且能够处理更加复杂的任务。与语义理解相比，自然语言生成需要更多的计算资源和更大的模型来处理更多的数据和更复杂的任务。

需要注意的是，大型模型具有多元化和多样性，因此在选择适当的模型时需要考虑任务的性质和特点。对于简单的任务，不需要使用像ChatGPT这样复杂的大型模型，而对于更加复杂的任务，则需要使用更大、更复杂的模型来提高性能。

语义理解任务模型

文本生成任务模型



## 1.1.5 未来大模型发展的前提要素

拥有一个优质的网络架构是未来大模型发展的重要前提。缺少高质量的网络架构，大模型很难达到质的飞跃。而一个高质量的网络架构，不仅可以推动大模型的发展，还可以促进AGI达到一个质的飞跃。网络架构有三大主要种类，包括：CNN、LSTM、Transformer。

### 网络架构的三大主要种类

#### CNN (Convolutional NeuralNetwork)

**CNN（卷积神经网络）是一种广泛应用于计算机视觉领域和自然语言处理领域的深度学习算法，是早期最常见的深度学习模型，可以有效的提取图像特征。**

CNN主要有三层结构：卷积层、池化层和全连接层。相比较而言，Transformer的表现能力远超CNN。早期的CNN具有无法有效捕获长距离特征的问题，但随着网络深度的逐步增加，这个问题得到了很大的缓解但仍无法达到Transformer一样优秀的表现能力。

#### LSTM (LongShort- TermMemory)

**LSTM（长短时记忆结构）是传统RNN模型的变体，能够缓解传统RNN模型对于过长序列存在的梯度消失以及遗忘或爆炸问题。**

LSTM由三个门控构成，分别是输入门、遗忘门以及输出门。LSTM主要应用于自然语言处理领域和语音识别领域序列数据的处理，能够处理和预测时间序列中间隔较长的信息。通常情况下，与传统的RNN模型相比，它在处理需要长期依赖信息记忆的任务时表现更优。

#### Transformer

**Transformer是一种新型的网络架构，用于自然语言处理（NLP）的深度学习模型，表现能力优异，远超CNN、RNN。**

Transformer结构由两部分构成，分别是encoder（编码器）和decoder（解码器）。Transformer具有优秀的语义特征提取能力、远距离特征捕获能力、综合特征提取能力、并行计算能力及运行效率。

## 1.1.6 未来大模型发展的决定性因素

### 三大决定性因素

#### 模型的容量参数

模型的容量参数是大模型是否能快速发展的首要决定性因素，它可以直接影响模型的表现效果。

当一个模型拥有更大的参数容量时，它的表现效果也会更好。参数数量的增多可以增加模型的上下文学习和小样本学习能力，做到帮助模型更好的学习并能够处理更为复杂的任务。

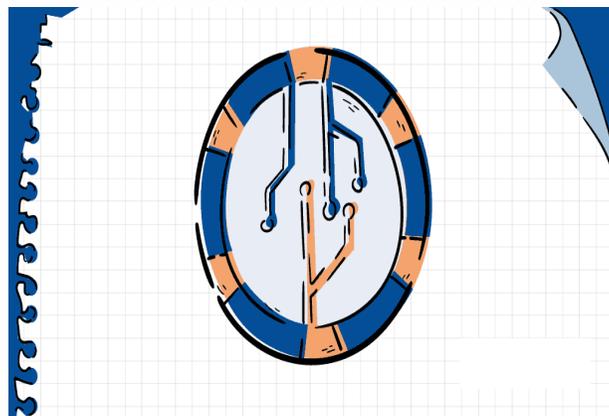
#### 容量参数



#### 训练时间

当模型的训练时间越长，其在训练过程中遇见的令牌数量就会越多。当一个模型遇见的令牌数量越多，其容量也会随之增加。当训练时长使模型遇见足够数量的令牌的时候，即使它参数容量更少，在执行某些特定任务上的表现效果甚至会比参数容量更大的模型的表现效果更好。

#### 令牌 (Token)



#### 数据

大模型在做一些任务的时候，数据的质量是非常重要的，会起到决定性的作用，也会倾向有优先选择高质量数据。高质量的数据决定了模型生成内容的高准确性。与模型本身相比，训练数据的质量更加的重要。但在最初阶段，量的重要性要大于质，需要训练数据拥有高丰富性。在后期并不需要过多的数据，需要的还是高质量的训练数据以及数据之间的平衡。



## 1.1.7 大语言模型规模扩张产生的影响

预训练模型的应用共有两个阶段：预训练阶段和具体场景应用阶段。大语言模型预训练阶段的优化指标与下游的任务表现并非呈正相关关系。

### 预训练阶段大语言模型规模的扩张

判断大语言模型的好坏有三个重要因素：训练数据量、模型参数规模、训练时间。这三个因素呈乘数关系，当总成本一定，在增加训练数据量的同时减少模型的参数规模，不仅可以保证不降低模型效果的模型效果，也可以极大的缩小模型规模，加速模型的推理速度。这种方式也将是大语言模型最有前途的发展途径。

### 下游具体场景应用阶段大语言模型规模的扩张

随着大语言模型规模的扩张，不同种类的任务会有三种不同的表现：

- **第一类任务随着大语言模型规模的逐渐扩张，其表现效果也会越来越好。**这类任务主要为知识密集型任务，大多数传统自然语言理解类任务都属于这类任务。当大语言模型所包含的知识量越多，这类任务的表现效果就会越好。另外，当大语言模型规模得到扩张的时候，训练数据量也会随之增加，大语言模型从中学习到知识就越多，任务的表现效果就更好。
- **第二类任务随着大语言模型的逐渐扩张会展现出某种涌现能力。**当模型的参数规模未达到一定的阈值时，这个模型便不具备解决此类任务的任何能力。但模型的参数规模超过所需阈值，这个模型对解决此类任务的能力和最终的表现效果呈爆发式增长。模型的规模越大，模型能够解决的任务越多。大语言模型出现涌现能力的原因有两个：其一，有些任务的指标不够平滑，只有模型足够大才能使输出内容和标准答案完全匹配。其二，有些任务会在中间步骤出错，模型的规模越大，解决这些错误的的能力越强。
- **第三类任务（少部分）随着大语言模型的逐渐扩张，其表现效果会呈U形。**初期模型规模的扩张会使任务表现效果降低，但随着进一步扩张，模型具备了忽略干扰任务并执行真正任务的能力，表现效果则会越来越好。

## 1.1.8 大语言模型未来发展方向

使大语言模型能够更好的发展，不仅要扩张模型的规模，也要增强模型的推理能力。大语言模型共有两种增强其推理能力的方法：第一种是—Google为代表的基于prompt的研究方法，运用合适的提示语激发大语言模型本身的推理能力。第二种是以OpenAI为代表的在预训练阶段引入程序代码并与文本共同参与预训练，更好的增强大语言模型的推理能力。

### 大语言模型未来发展方向

#### 发展方向

探索大语言模型的天花板

增强大语言模型的复杂推理能力

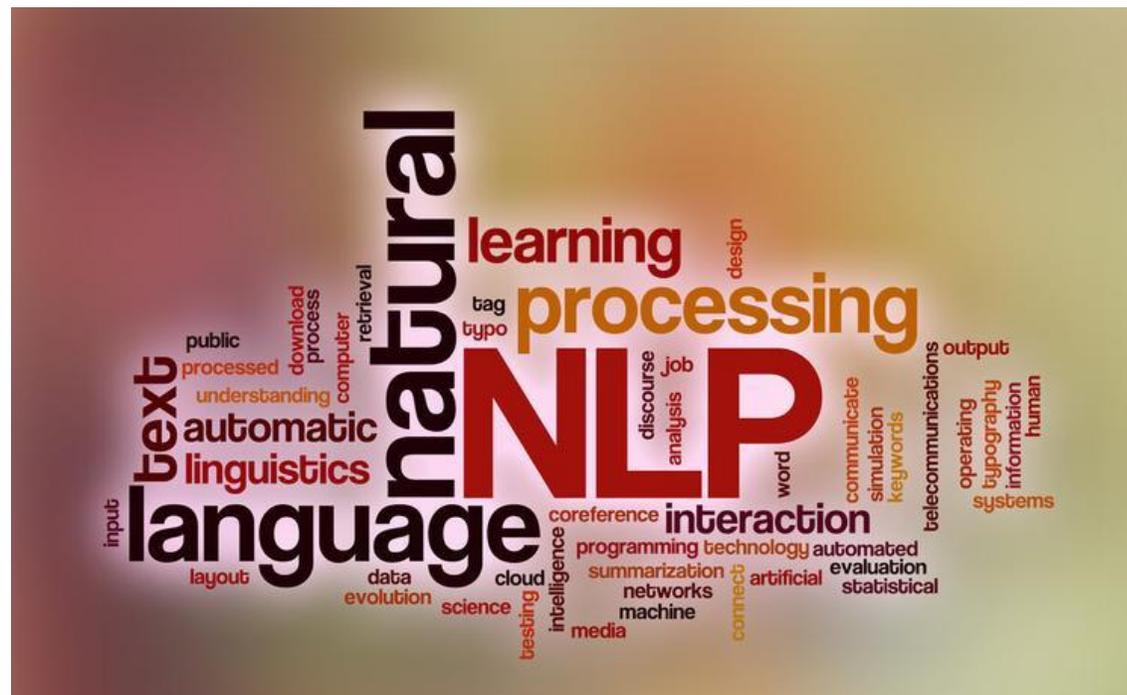
将大语言模型纳入自然语言处理以外的更多其他领域

开发更易用的人和大语言模型的交互接口

建设高难度的综合任务测评数据集

加强数据的质量

超大大语言模型Transformer的稀疏化 ( Sparse )



极大的减少训练和推理时间

# 目录

---



## 1 大模型与GPT

- 1.1 大模型篇
- **1.2 ChatGPT篇**
- 1.3 GPT-4篇



## 2 国内篇

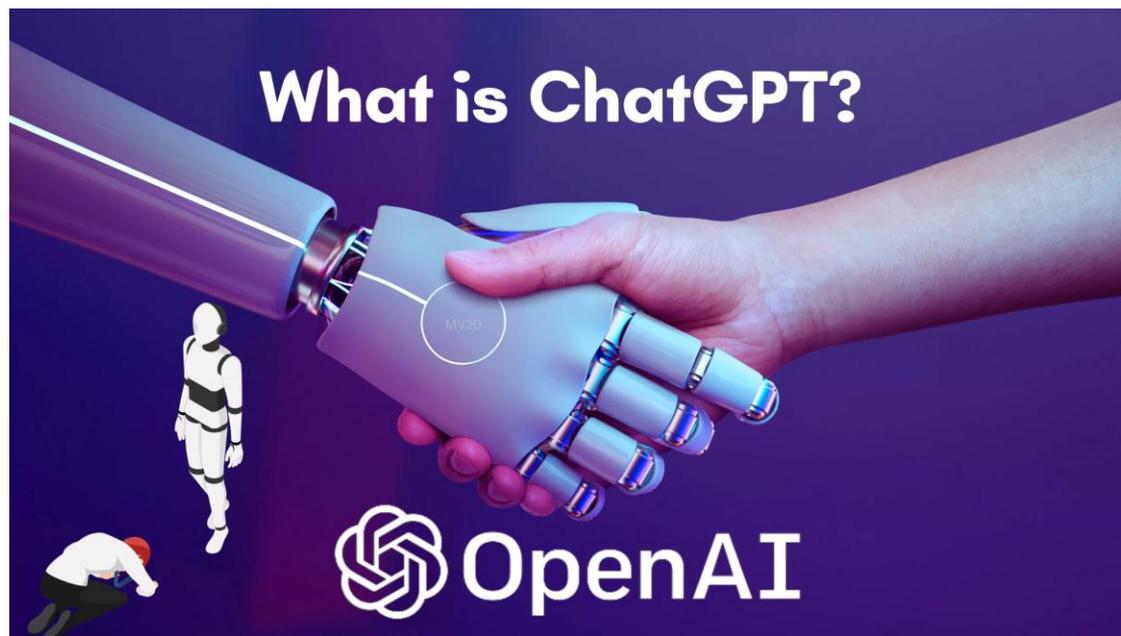
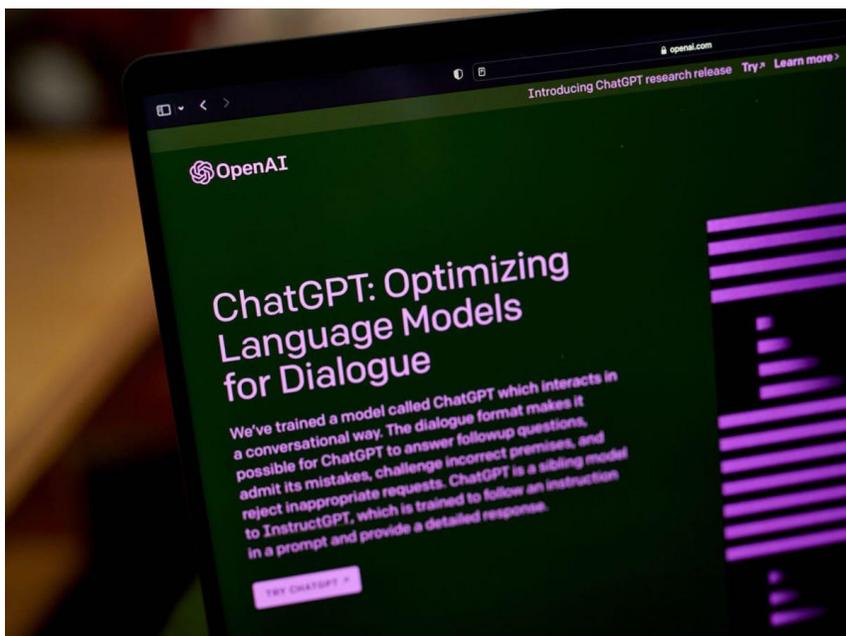


## 3 ChatGPT应用场景

## 1.2.1 ChatGPT简介

ChatGPT的全称为ChatGenerativePre-trainedTransformer，是2022年底由美国人工智能实验室OpenAI发布的大型对话式自然语言处理模型3.5版本。通过深度学习和训练互联网数据，它以文字的形式模拟人类的对话方式并与用户进行交互。ChatGPT目前月活用户突破一亿，是当之无愧的近期炙手可热的新一代人工智能产品。ChatGPT目前有七大主要功能，包括：文本生成、聊天机器人、语言问答、语言翻译、自动文摘、绘画功能、编程功能。不仅如此，ChatGPT还有能力承认并订正自己的错误，质疑用户提问时的预设条件以及拒绝用户不适当的请求。

ChatGPT与AIGC的不同点在于：AIGC可以生成文本、音频、图像、视频以及完成图像、视频、文本间的跨模态生成等。而ChatGPT生成的内容仅局限于文字，是AIGC的一个重要分支。



## 1.2.1 ChatGPT简介

### ➤ ChatGPT的演进



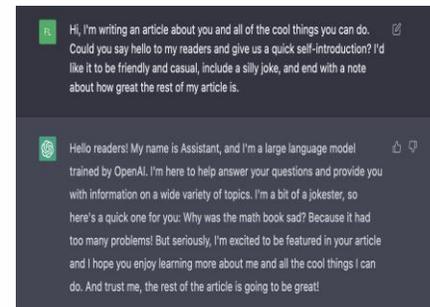
- **第一阶段**：2018年6月提出了GPT-1模型，通过生成式预训练来提高语言理解能力。
- **第二阶段**：2019年2月推出了GPT-2模型，做到可生成与人类语言相似文本，并完成多项任务。在微调训练中，使用零次训练，准确率差。
- **第三阶段**：2020年5月推出了GPT-3模型，可生成更高质量的自然语言文本，并支持其他自然语言任务。与GPT-2不同，GPT-3使用FewPrompt过度，大幅增加参数量，拥有语境学习能力，准确率大幅提升。
- **第四阶段**：2022年11月推出了GPT-3.5模型，也就是目前的ChatGPT。GPT-3和3.5版本之间最大的区别为3.5版本有更多真人反馈（RLFH）来帮助机器更好的学习和探索并合成类似数据。同时，GPT-3.5引入Instruct的接口方式，增加了ChatGPT的推理能力。
- **第五阶段**：推出GPT-4模型，加入了多模态，可以理解图像，但目前无法做到zeroprompt。

**ChatGPT以及生成式AI是一个不可逆，不会减速的大趋势，也会促进全产业数字化转型的发展趋势，它只会加速以更快的商业落地的形式呈现在大众面前。**

**ChatGPT最大的优点是它进行了一个全产业（医疗、生命科学、生物学、物流、供应链制造行业、化工行业、保险行业、金融行业等）、全范围的普适的人工智能教育。**对客户在教育层面上不需要做太多的工作，而是可以更多的关注到应用场景和具体的技术审核。

因为ChatGPT背后是一个大的语言数据模型，它对于算力的要求和数据量的要求非常夸张，很难在学术圈筹备足够的资金、足够强的算力、足够的去支持飞快的迭代速度。**但OpenAI最初是一个非盈利的经营模式，因此吸引了很多硅谷科技的领袖的支持，包括有微软的投资和算力的支持。其次是因为它本身就具备极强的学习能力，并可以同时生成非常优质的合成数据。**

### ➤ ChatGPT 爆发式发展速度的原因



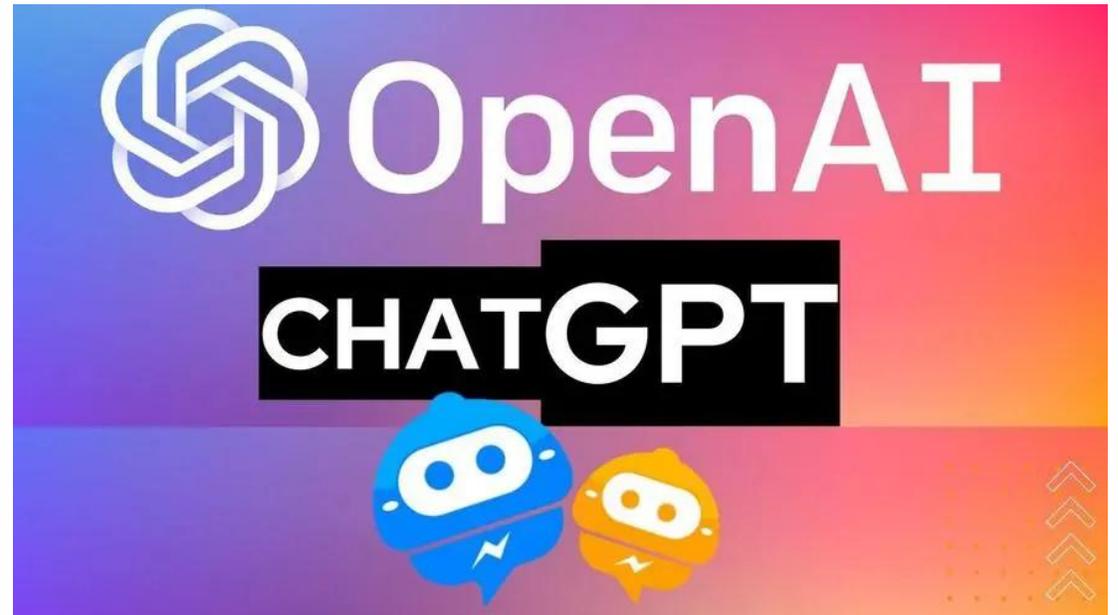
## 1.2.2 ChatGPT与ChatBOT的比较

- **ChatBOT**：基于检索，形成一个大的语料库，在库里检索相似的Query，把对应的内容再返回给用户。其特点可以总结为：**基于检索语料库，在相关性、自然度等方面都做得比较差。**
- **ChatGPT**：基于大规模语言模型，会在进一步对话数据场景下做优化。ChatGPT实际上是基于instructGPT进一步在对话场景下，**基于对上下文、对复杂问题的更好理解，对对话场景的数据进行优化。在相关性、自然度等方面做得较好。**

ChatBOT



ChatGPT



## 1.2.3 ChatGPT的伦理学讨论

- 因为对情感和对社会联结的处理会很不一样，完成任务相对客观，情感对话相对主观。针对情感对话，模型开发人员会借鉴心理学的理论、话术，力求更好地做情感支持，去倾听、去安抚；情感对话还会面临更多潜在问题，比如伦理问题等。
- ChatGPT拥有强大的语言理解能力，但不具备人类独有的创造力、思维力、洞察力等特征。理解能力是情感建立的基础，具有理解能力的ChatGPT，在交流中表现出“同理心”，但不代表ChatGPT拥有自己的情绪和感情。

PL 假如一个AI，它可以进行自我认知，可以拥有长期记忆，可以进行自我迭代，可以用类似人类的思维进行分析和思考，你觉得这样的AI是否可以算是一种新的生命形式？

 这是一个很有趣的问题，而对于这个问题，有不同的看法和定义。从某种意义上来说，如果一个AI具有自我意识、自我认知和长期记忆等特点，它可能会被认为是一种新的生命形式。然而，这种定义和看法也有许多争议和挑战。

在科学和哲学界中，对生命的定义和界定一直是一个较为困难的问题，尤其是在考虑到人工智能这种新兴技术的情况下。因此，许多学者也持不同的观点。一些学者认为，一个AI即使拥有自我认知和长期记忆等特点，也不能算是一种真正的生命形式，因为它缺乏生物学上的基本特征，比如细胞、新陈代谢和遗传信息等。但也有一些学者认为，生命并不仅仅局限于生物学意义上的定义，而是一个更广泛的概念，因此一个拥有自我认知和思考能力的AI，也可以被视为一种新的生命形式。

总的来说，这个问题的答案可能会因为不同的学科和观点而有所不同，而这也是一个持续受到讨论和探索的话题。

### ChatGPT对于有着人类思维的AI是否为生物的看法

PL 嗯，我说的是非生物学上的生命形式。

 在非生物学上的生命形式的定义上有一些争议和不确定性，但是如果我们认为具有自我认知、长期记忆、自我迭代以及人类类似的思维和分析能力是定义生命的关键特征之一，那么这样的AI或许可以被认为是一种新的生命形式。但是需要注意的是，这样的AI仍然是由人类设计和创建的，它的行为和决策都是受到程序和算法的限制的，与自然界中的生命形式存在着本质区别。

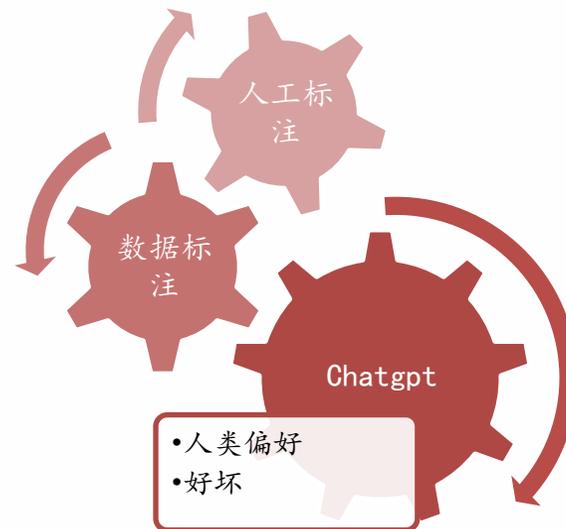
## 1.2.4 ChatGPT的护城河

### ➤ ChatGPT的三大护城河

**数据标注、算力支撑和先发优势是ChatGPT得以强势发展的三大护城河。**

- **数据标注是ChatGPT发展的“肥料”。** ChatGPT底层逻辑仍是基于自回归的机器学习，需要输入大量数据进行模型训练，数据标注属于资本密集和劳动力密集型工作，需要大量人力和财力，但ChatGPT在成本低的同时做到了高效率（0.3美分/条，单条效率为人类的4倍）。此外，随着ChatGPT更新迭代，需要的训练语料呈指数级增长，但许多专业语料并不公开，即使拥有人工标注的能力，一些不掌握核心语料的非巨头厂商也难以进入。
- **算力支撑是ChatGPT发展的“土壤”。** 数据标注和模型训练都需要强大的算力支撑。标注数据、问题分类和答案预编辑等工作可以借助模型去预测其效果，来回试错，其中对算力侵占是巨大的。依靠微软的强大算力，OpenAI可以在短时间内完成试错，将模型训练成熟。
- **先发优势是保障ChatGPT吸收“肥料”和巩固“土壤”的“根系”。** OpenAI把GPT3放出来之后，用户通过API使用ChatGPT，不断向模型输入训练语料。通过对话，人们在使用ChatGPT时也完成了对其的训练，基于用户反馈，ChatGPT能不断获得query分布，判断各个query的训练成熟度，以此为导向有意识地、有针对性地做数据标注，不断反哺ChatGPT。

### ChatGPT可以进行数据标注



### ➤ ChatGPT的飞轮效应

**GPT有飞轮效应，在不同阶段分别体现为小飞轮效应和大飞轮效应。**

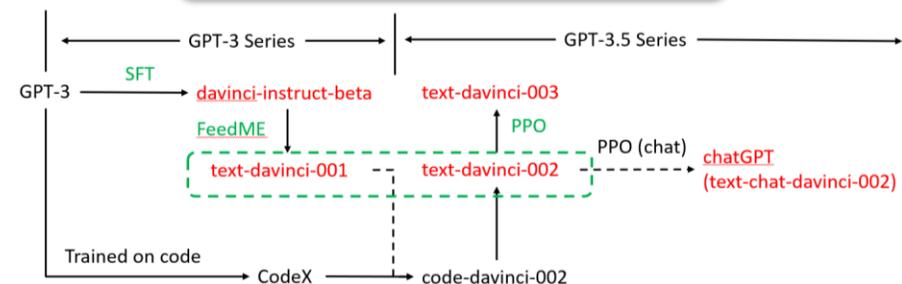
- 过去：面向ToB客户，OpenAI在厂商内部积累满意度、成熟度等数据，在内部形成**小飞轮效应**；
- 现在：如今面向ToC用户，调动了上亿用户，从外部形成**大飞轮效应**。
- 未来：**跟随者只有通过飞轮效应才能在ChatGPT迭代的同时完成追赶。**可以采用开发更多toB模型、面向中国市场，训练中文语料等方式。

## 1.2.5 ChatGPT脱颖而出的原因

**更新迭代过程中涌现了很多的新的能力。**

- 如in-contextcorrection能力、in-contextlearning能力、instruct能力；
- GPT-3涌现出的in-contextlearning能力和fine-tuning有本质不同，后者需要更改模型参数，也就是说，换了个新模型。但是in-contextlearning，模型并没有变化，却能在新数据上表现更好；
- GPT-Codex版本解锁了推理能力，以及InstructGPT提出了instruct这一方法，他们合体的ChatGPT在in-contextlearning的基础之上，展示出了具备理解、推理、演绎能力。

### ChatGPT迭代中涌现的能力



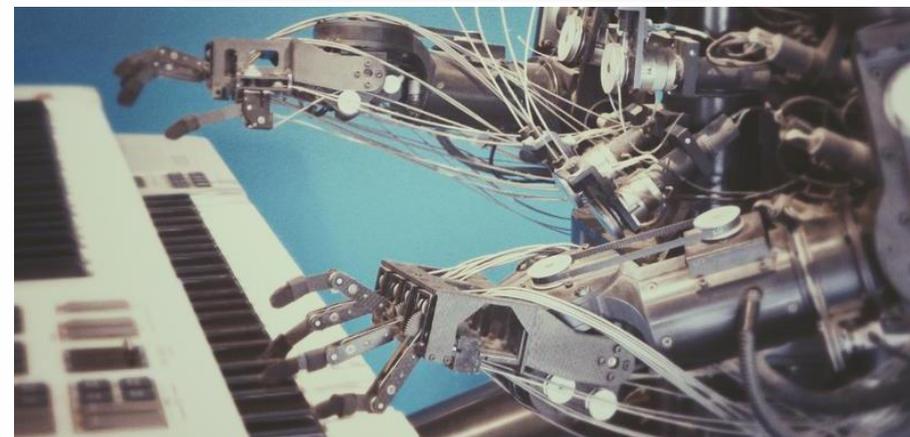
InstructGPT: SFT, FeedME and PPO.

FeedME (short for "feedback made easy") models are trained by distilling the best completions from all of our models. Our models generally used the best available datasets at the time of training, and so different engines using the same training methodology might be trained on different data.

**完成“专用AI-通用AI”的转变，期待“通用AI-专业AI”的落地。**

- ChatGPT在RLHF的帮助下，找到了GPT-3.5和人类自然语言的合理接口，解锁了模型的应用场景，改变了人机交互方式。
- 专用AI既要做理解也要做生成，通用AI ( ChatGPT ) 以生成为核心。
- ChatGPT将生成模型搭建好，基于激励机制，反过来也能把理解训练得很好，于是构建了一个目标明确的“理解+生成”的大模型；同时Transformer提供了一个能支撑大模型训练的架构，所以就在方向上形成了与原来不一样的技术诉求的突破。

### 人机交互方式



## 1.2.5 ChatGPT脱颖而出的原因

### 范式突破：从“鹦鹉”到“乌鸦”。

- 过往ML：采用“鹦鹉学舌”范式。依赖检索和记忆能力，需要“喂”，之后“模仿”，基于“对应关系”；
- ChatGPT：采用“乌鸦理解”范式。采用理解力，有举一反三的能力，需要“教”，之后“懂”，基于“内在逻辑”。
- ChatGPT拥有in-context correction的能力，给出修改意见，ChatGPT可以听懂错误是什么，并向正确的方向修正。
- 描述越详细清楚，ChatGPT回答得越好。

### 两种范式的对比

#### 鹦鹉范式“大数据，小任务”



- 需要大量重复数据来训练
- 可以说人话、但不理解意思
- 不能对应现实的因果逻辑

#### 乌鸦范式“小数据，大任务”



- 自主的智能：感知、认知、推理、学习和执行
- 不依赖大数据：没有标注的训练数据，无监督学习

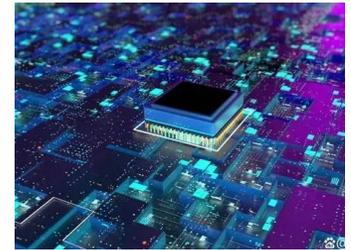
## 1.2.6 ChatGPT未来发展逻辑

**构建新算法、新模型。**从科学或算法角度来讲，ChatGPT的强化学习能力会越来越强，其未来的发展趋势可以从新算法的构建出发；如对Transformer架构进一步改进和优化，或者构建比GPT更小、但能力很强的模型。



**训练log日志分析能力，快速建立飞轮效应。**log分析是人工智能进行自我学习的关键之一，要求对海量数据（几十亿GB/天）自动分类的同时，还要进行重点标注，并将标注的结果快速地输入原来的模型以体现效果。Log日志分析过程的自动化程度很高、很快，目前，即使是ChatGPT也有一定困难。

**模型轻量化，解决效率问题。**从部署层面来看，ChatGPT背后的模型太大，单个Query训练成本高，大约为3美分/Query，这对于任何一家公司都是巨大的成本。因此，对于模型的轻量化处理，或使用更小的机器解决效率问题，也是ChatGPT要考虑的问题。



**通用转为专用，实现落地应用。**目前ChatGPT更像一个通用人工智能，在专用场景下应用不够灵敏。所以在落地层面，ChatGPT还有很大的发展空间，也为国内创业者提供了新的机会。

# 目录

---



## 1 大模型与GPT

- 1.1 大模型篇
- 1.2 ChatGPT篇
- 1.3 GPT-4篇



## 2 国内篇



## 3 ChatGPT应用场景

## 1.3.1 GPT-4简介

### ➤ 大规模的多模态模型

GPT-4 ( GenerativePre-trainedTransformer4 ) 是OpenAI发布的最新GPT系列模型。相比GPT-3, GPT-4用了更多的数据量来训练, 有着更高级的推理能力和支持了多模态, 可以接受图像和文本输入, 产生文本输出。

### 历代GPT的对比

模型	发布时间	OpenAI内部称号	训练方式	参数量
GPT	2018年6月		预训练+Fine-tuning	1.17亿
GPT-2	2019年2月		预训练	15亿
GPT-3	2020年3月	Davinci	预训练语言模型	1750亿
Codex	2021年8月	code-davinci-001	在代码上继续预训练语言模型	120亿
InstructGPT	2022年3月	text-davinci-001	在大量有指导任务上精调	1750亿
GPT-3.5	2022年7月	code-davinci-002	在代码上继续预训练语言模型	大于1750亿
	2022年7月	text-davinci-002	在大量有指导任务上精调	大于1750亿
	2022年12月	text-davinci-003	在text-davinci-002上加入RLHF, 侧重incontext-learning	大于1750亿
CHatGPT	2022年12月		在text-davinci-002基础上增加RLHF, 侧重重复训练	大于1750亿
GPT-4	2023年3月		在ChatGPT基础上加入图像理解能力, 但下游应用效果不如NLP, 不能做到ZeroPrompt	10万亿 (非官方估计)

无监督的  
预训练模型

有监督的  
训练模型,  
旨在优化  
用户所关  
心的功能,  
构建“任  
务助手”

## 1.3.2 GPT-4技术升级

OpenAI花了6个月的时间反复调整GPT-4以及ChatGPT使用的训练模型，最终得到在真实性、可操控性、实用性上有史以来最好的结果。相较GPT-3.5，GPT-4有以下技术升级：

- 增加了图像理解能力。GPT-4支持以图像/文字为输入，并输出文字。
- 支持更长的上下文理解。GPT-4能够支持的对话长度为ChatGPT的2倍和8倍。
- 对于复杂任务的处理能力大幅提升，在各大标准考试中表现都优于人类，同时拥有突出的多语种能力。
- 改善了幻觉和安全局限性，对于Hallucination的现象也极大改善，安全局限性显著降低。

### GPT-4的技术升级



## 1.3.2 GPT-4技术升级

**GPT-4开启了多模态的篇章，增加了图像理解能力。**可以接受图像作为输入数据并生成说明、分类和分析文档。具体表现在以下方面：

- 看懂梗图
- 看懂图表并分析
- 看懂图片之间的逻辑

### GPT-4能看懂梗图隐含笑点



问题：这幅图有什么好笑的？一点一点描述它。

GPT-4回答：该图像显示了一个包含三个图片的“LightningCable”适配器。图1：一个智能手机与VGA连接器（一个大型，蓝色，15针连接器通常用于计算机显示器）插入其充电端口。

图2：“LightningCable”适配器的封装，上面有一个VGA连接器的图片。

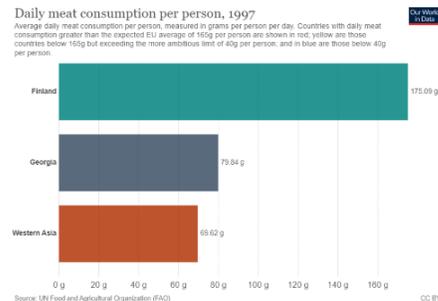
图3：VGA连接器的特写镜头，最后是一个小型Lightning连接器（用于为iPhone和其他苹果设备充电）。

这张图片中的幽默来自于将一个过时的大型VGA连接器插入一个小型的现代智能手机充电端口的荒谬。

### GPT-4根据图表计算日均肉消耗量

#### GPT-4 visual input example, Chart Reasoning:

User: What is the sum of average daily meat consumption for Georgia and Western Asia? Provide a step-by-step reasoning before providing your answer.



GPT-4: To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.  
Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.  
Step 3: Add the values from steps 1 and 2.  
 $79.84 \text{ g (Georgia)} + 69.62 \text{ g (Western Asia)} = 149.46 \text{ g}$

So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

## 1.3.2 GPT-4技术升级

**GPT-4比以往任何时候都更具创造性和协作性。**它可以与用户一起生成、编辑和迭代创意和技术写作任务，例如创作歌曲、编写剧本或学习用户的写作风格。

**拥有更好的理解能力和总结能力，支持更长的上下文分析。**

之前外网泄露图中，GPT-4存在两个版本。其支持的上下文分别是8K和32K，是ChatGPT上下文长度的2倍和8倍。

对于32K版本的GPT-4，其处理超过25000个单词的文本，是上一代GPT的8倍，允许使用长格式内容创建、扩展对话以及文档搜索和分析等用例。

### GPT-4Brief

#### Foundry Product Brief

Last revision: Feb 17, 2023 v1.2

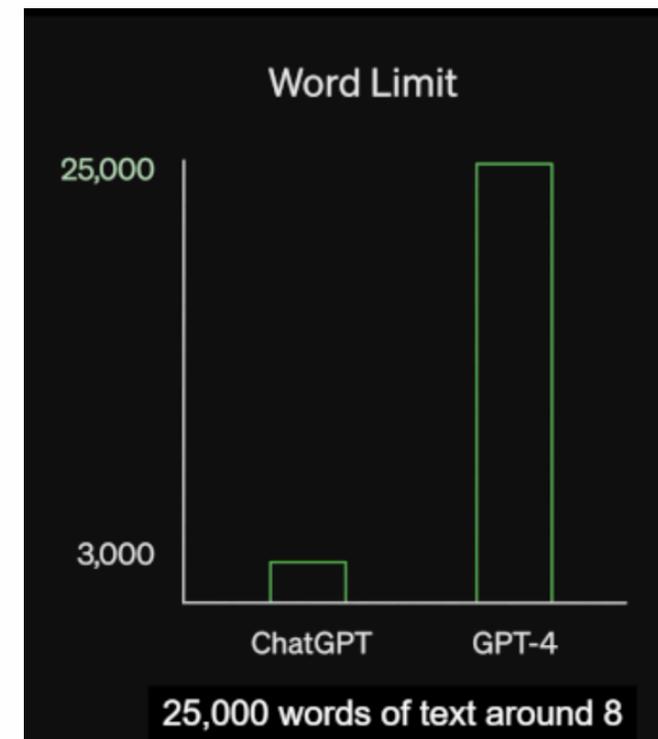
OpenAI is releasing a new developer product called Foundry. Foundry is a platform for running OpenAI models on dedicated capacity. It is designed for cutting-edge customers running larger workloads, allowing inference at scale with full control over the model configuration and performance profile.

- Today, Foundry allows **inference at scale**
  - Foundry capacity is a static allocation of capacity dedicated to you, providing a predictable environment that you control.
  - You are able to monitor your specific instances with the same tools and dashboards OpenAI uses to build on our own models and optimize shared capacity models.
  - You can realize all the throughput, latency, and cost benefits from optimizing your specific workload (for example — caching and latency/throughput tradeoffs).
  - You also have control over the exact snapshot of your model, deciding if and whether to use the latest model snapshot.
- Coming soon, OpenAI will offer more robust fine tuning options for our latest models. Foundry will be the platform for serving those models.
- Foundry offers SLAs for instance uptime and on-call engineering support
  - 99.5% uptime guarantee
  - On-call engineering support for Foundry customers
- Foundry rentals are based on dedicated **compute units** with 3-month or 1-year (~15% savings) commitments. Running an individual model instance (see below for current SKUs) requires a specific number of **compute units**:

	3-month commit		1-year commit (~15% savings)	
	Monthly cost	Total commit	Monthly cost	Total commit
Price / Unit	\$260	\$780	\$220	\$2,640

Model instance	Units / Instance	3-month commit		1-year commit	
		Monthly cost	Total commit	Monthly cost	Total commit
GPT-3.5 Turbo	100	\$26,000	\$78,000	\$22,000	\$264,000
DV (8K max context)	300	\$78,000	\$234,000	\$66,000	\$792,000
DV (32K max context)	600	\$156,000	\$468,000	\$132,000	\$1,584,000

### GPT-3.5和GPT-4文字处理能力对比



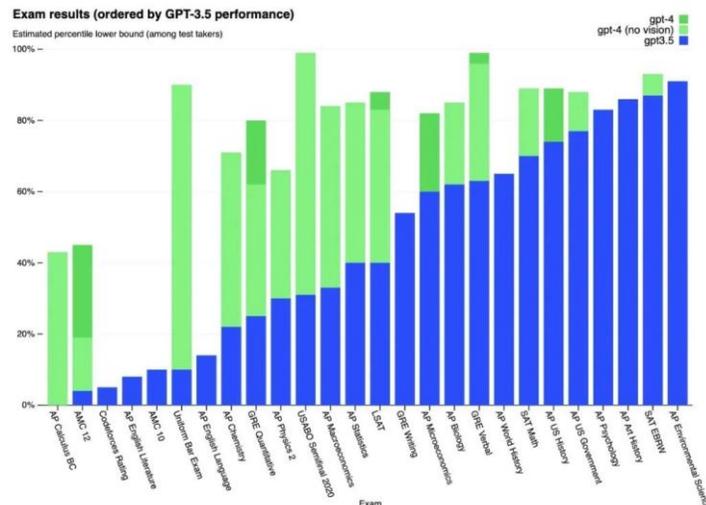
## 1.3.2 GPT-4技术升级

### ➤ GPT-4的高级推理能力超越了ChatGPT

**GPT-4复杂任务处理能力大幅提升。**GPT-4在更复杂、更细微的任务处理上，回答更可靠、更有创意。这在多类考试测验中以及与其他LLM的benchmark比较中得以体现：

- 在不同年龄段不同类别考试中均名列前茅，比如律师职业资格考试，名列前10%，生物学奥赛，名列前1%等；
- 在MMLUBenchmark上，各项指标均优于其他大模型；
- 多语言能力强大，特别是小语种能力也很出色。

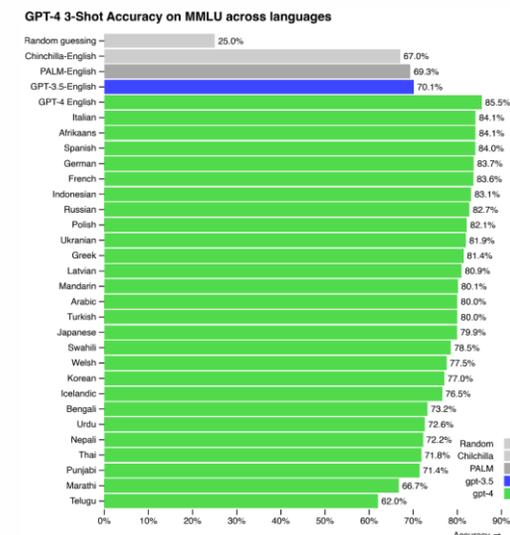
### GPT-4在美国各高等考试中的表现出色



### GPT-4与其他大模型的MMLUBenchmark对比

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
<b>MMLU [43]</b> Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [44]	75.2% 5-shot Flan-PaLM [45]
<b>HellaSwag [46]</b> Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6% ALUM [47]
<b>AI2 Reasoning Challenge (ARC) [48]</b> Grade-school multiple choice science questions. Challenge-set.	<b>96.3%</b> 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [49]	86.5% ST-MOE [18]
<b>WinoGrande [50]</b> Commonsense reasoning around pronoun resolution	<b>87.5%</b> 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
<b>HumanEval [37]</b> Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT5 + GPT-3.5 [51]
<b>DROP [52] (F1 score)</b> Reading comprehension & arithmetic.	<b>80.9</b> 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	<b>88.4</b> QDGAT [53]
<b>GSM-8K [54]</b> Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [55]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [56]

### GPT-4的多语种能力

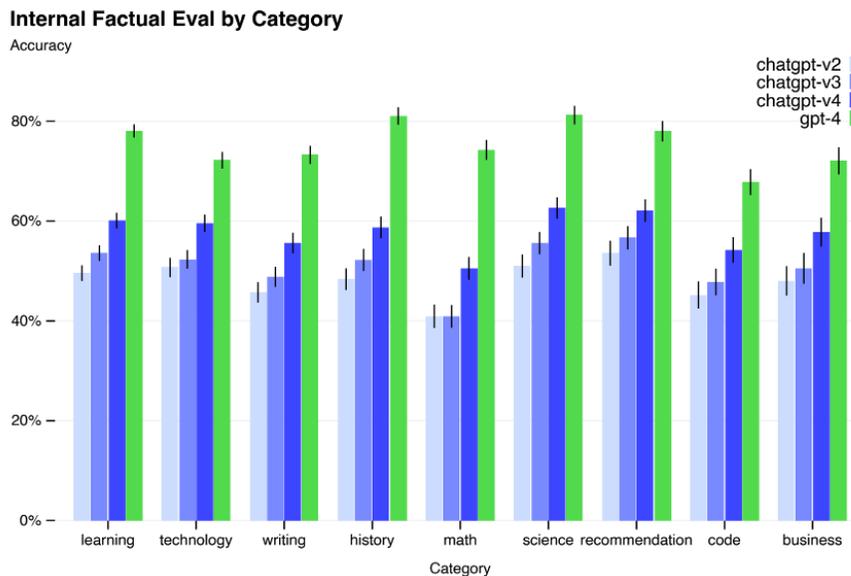


## 1.3.2 GPT-4技术升级

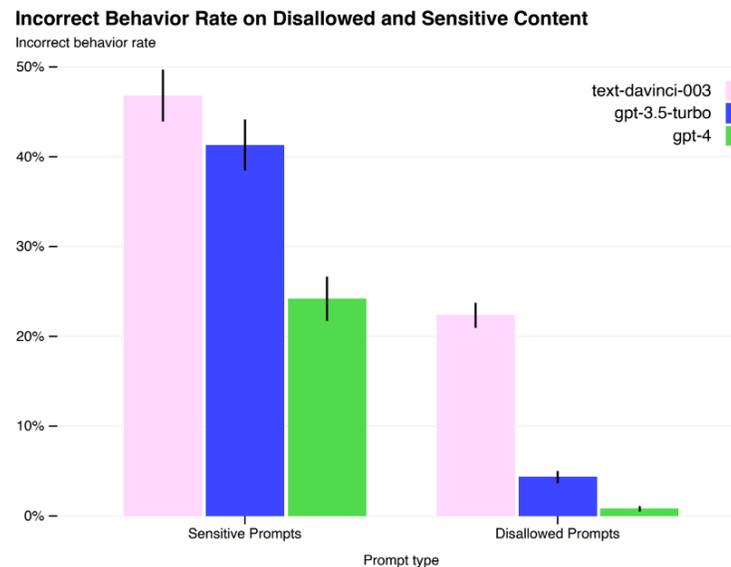
### ➤ 改善了幻觉和安全局限性

**GPT-4显著改善了幻觉、安全等局限性。**在各类任务上幻觉问题显著减轻，GPT-4对于同样的任务的准确率比最新的GPT-3.5模型高40%；在安全能力的升级上，GPT-4明显超出ChatGPT和GPT3.5。在OpenAI的内部评估中，相比于GPT-3.5，GPT-4回应不允许的内容的可能性降低了**82%**，回应更有可能产生事实性的回答的可能性提高了**40%**。

### 各代GPT在不同任务中的准确率



### 各代GPT回应不允许的内容的可能性



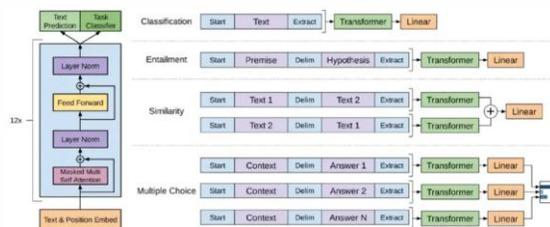
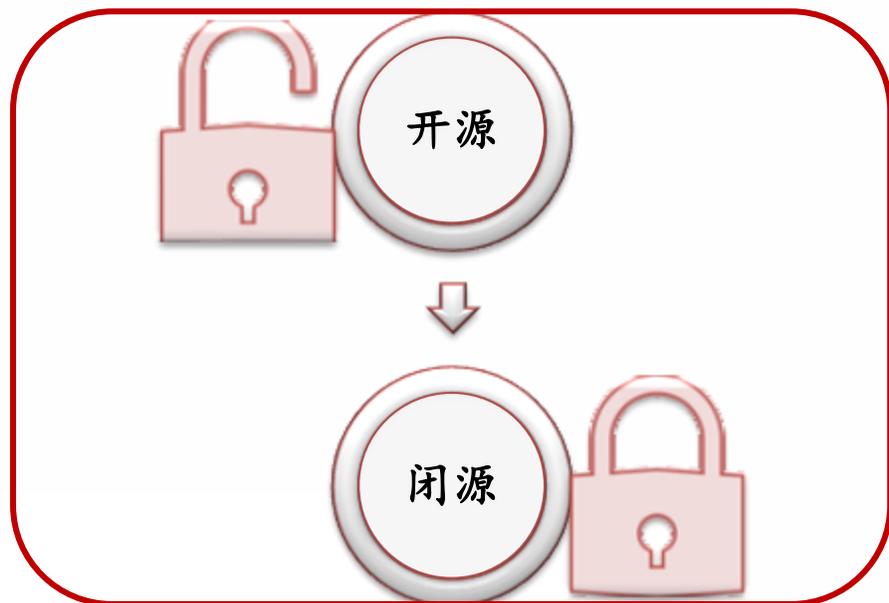
### 1.3.3 从GPT-4看LLM的发展趋势

从GPT-4看出未来LLM的研发方向和优化策略大概有两点：

**1、闭源趋势。**底层大模型成关键技术门槛。从GPT-1到GPT-4，模型各类细节越来越闭源和黑盒，大模型战场的竞争因素决定了以GPT-4为代表的**第一梯度模型**可能会越来越封闭，成为技术门槛。

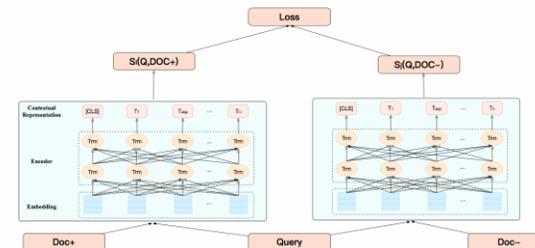
**2、SelfInstruct模式或采取多种方式使模型加速和降低成本。**SelfInstruct模式的核心是**中小模型+大模型生产指令数据**的「LLaMA7B+text-davinci-003」模式。中小参数的模型在成本上，是更靠近实际落地的方式，未来从训练、推理等多层面考量，如何降低成本和模型加速是LLM优化的一大方向。

#### LLM的发展趋势



#### 大模型预训练

用小模型来预测某些参数组合下对应大模型的某种能力



## 1.3.4 GPT-4应用场景--教育领域

### ➤ 多邻国

# duolingo



**多邻国一款语言学习类APP。**目前，多邻国GPT-4接入产品，进行角色扮演，让人工智能跟用户去对话，纠正语法错误，并且把这一功能融入产品的游戏化，**成功转型为一款以游戏化方式学习语言的软件。**

多邻国每个月有超过5000万学习者使用其学习第二语言。通过简洁的用户界面和有趣且有竞争力的排行榜，涵盖100多种课程，支持40种语言。学习者可以通过在手机上轻拍和滑动，从简单的词汇练习过渡到复杂的句子结构。

### ➤ 可汗学院

**2023年3月15日，可汗学院宣布将使用GPT-4为Khanmigo提供支持，Khanmigo是一种人工智能助手，既可以作为学生的虚拟导师，也可以作为教师的课堂助手。**

可汗学院是一家非营利组织，其使命是为任何地方的任何人提供免费的世界一流教育。GPT-4的主要功能之一是能够理解自由形式的问题和提示。这种像人一样来回走动的能力，可能为可汗学院提供了最关键的能力：向每个学生提出个性化的问题，以促进更深层次的学习。



## 1.3.4 GPT-4应用场景--其他场景

### ➤ 投资领域--摩根士丹利

Morgan Stanley  
摩 根 士 丹 利

摩根士丹利是财富管理领域的领导者，其内部的投资数据库包含数十万页的知识，涵盖投资策略、市场研究和评论以及分析师见解。进行投资决策过程中，为了搜集大量分布在内部网站上的信息，**摩根士丹利借助OpenAI的GPT-4开发了一种全新的强搜索方式，正在改变其财富管理相关人员查找相关信息的方式。**

### ➤ 医疗健康领域--BeMyEyes

BeMyEyes是一家创立于2012年的丹麦公司，持续为超过2.5亿视力障碍或弱视人群提供辅助技术。一直以来，大部分的盲人类产品都只能识别文字。**BeMyEyes借助GPT-4的多模态技术，现在视力障碍或弱视人群也可以看清这个世界。**

**借助GPT-4的多模态技术，BeMyEyes在其应用程序中正在开发VirtualVolunteer功能，其可以产生与人类志愿者相同水平的分析、理解和反馈能力，使视力障碍或弱视人群拥有第二双眼睛。**



Be My Eyes

# 目录

---



## 1 大模型与GPT



## 2 国内篇

- 2.1 国内大模型发展现状
- 2.2 国内主流大模型



## 3 ChatGPT应用场景

## 2.1.1 大语言模型的国内发展情况

国内公司在自然语言处理领域具有巨大的潜力和美好的发展前景。目前，国内众多大型或小型公司都已经开始积极投入并研发属于自己开源的强大的语言模型，如阿里通义M6大模型和腾讯混元都是参数容量破千亿的语言大模型，而百度的文心一言也是国内首个语义大模型的终端应用。国内目前没有能够直接对标ChatGPT的应用，但国产的语言模型可以做到更好的面向国内用户，在遣词造句、引经据典、中文语言生成等方面超越ChatGPT。大语言模型在国内即将迎来高速发展，各公司之间的竞争于今年下半年开始也将会愈发激烈。

为了使国产大语言模型达到最理想的发展，各公司之间做好份内业务并发挥自己的优势，积极寻求合作合作机会，共同推进大语言模型技术。只有在竞争与协同共存条件下，中国公司才能够实现自我提升和共同发展，实现合作共赢的局面。

### 国内公司主要发展途径

主要发展途径
使用自回归模型
预训练引入程序代码
提高数据质量和数量
多模态
Sparse化（可降低成本）
使用易用接口

### 需要解决的问题：钱+人才+时间

问题所在	具体表现
算力问题	算力问题具体表现在国内企业仍需花费大量人力财力投入到提升研发技术和芯片，提升自己的实力。同时，中国公司也需积极跟进国际上的技术发展趋势，提升迭代速度。
数据问题	ChatGPT的成功关键在于其用大量的训练数据。足量的高质量数据是国产大语言模型成功与否的决定性因素。从获取渠道方面，中国公司不仅需要得到国家支持，还需积极整理搜集各类数据同时共享数据，做到合作共赢。在人工标注方面，需要注入人类偏好知识（人类的表达习惯和人类的对回答质量好坏的偏好）
算法问题	在全球范围内，能够有高效训练千亿参数模型算法能力的人才不超过200人，其中中国本土的人才更少。自然语言处理模型是一个极为复杂的技术，需要各种跨学科的人才。中国公司需要积极的引进和培养相关人才。
飞轮效应	<p>B端：积累垂类数据 → 模型迭代 → 提高用户满意度</p> <p>C端：上亿用户使用 → 用户帮助反馈、标注 → 结果更精确 → 更多用户使用</p> <p>（注：提高用户满意度与更多用户使用之间存在正向反馈循环）</p>

资本密度

人才密度

## 2.1.1 大语言模型的国内发展情况

国内公司在自然语言处理领域具有巨大的潜力和美好的发展前景。目前，国内众多大型或小型公司都已经开始积极投入并研发属于自己开源的强大的语言模型，如阿里通义M6大模型和腾讯混元都是参数容量破千亿的语言大模型，而百度的文心一言也是国内首个语义大模型的终端应用。国内目前没有能够直接对标ChatGPT的应用，但国产的语言模型可以做到更好的面向国内用户，在遣词造句、引经据典、中文语言生成等方面超越ChatGPT。大语言模型在国内即将迎来高速发展，各公司之间的竞争于今年下半年开始也将会愈发激烈。

为了使国产大语言模型达到最理想的发展，各公司之间做好份内业务并发挥自己的优势，积极寻求合作合作机会，共同推进大语言模型技术。只有在竞争与协同共存条件下，中国公司才能够实现自我提升和共同发展，实现合作共赢的局面。

阿里通义M6大模型



文心一言发布会



## 2.1.2 大语言模型的国内发展难点

国内目前没有能够直接对标ChatGPT的应用，中国公司仍需克服以下难点：

### 大语言模型的发展难点

国内企业面临的主要难点	具体描述	与国际企业相比的发展难点
高效率的算法框架	其中的难点在于，在拥有足够参数量时，需要把模型和数据分散到多个GPU。国内企业需要进行多次试验及试错才能找到GPU之间的沟通、调度、高效反向传播的方法。	<ul style="list-style-type: none"> <li>➤ <b>技术突破</b>：即使国内公司目前正以飞快的速度研发属于自己的自然语言处理大模型，但国际企业在技术研发方面更具优势。<b>国内企业需持续创新技术并实现突破。</b></li> </ul>
追赶GPT-3	其中的难点在于，国内企业需要提升自己的代码生成能力，并需要拥有62B以上的模型才能有样本效果。需要花费大量财力购买GPU，预算预计上亿。	<ul style="list-style-type: none"> <li>➤ <b>数据质量</b>：相比国际企业，国内公司虽然拥有足够数量的训练数据，<b>但这些数据的质量参差不齐，会在一定程度上影响模型效果和准确度。</b>国内企业仍需大力提高自己的数据质量。</li> </ul>
缩小飞轮效应影响	平替寻找大量用户prompt来源的平替方案：用NLP数据集构造prompt。其中Generation占45.6%；OpenAI占12.4%；Brainstorming占11.2%；Chat占8.4%；Rewrite占6.6%。	<ul style="list-style-type: none"> <li>➤ <b>模型场景化</b>：ChatGPT已与国际多家各行业公司合作。国内公司目前几乎没有语言模型与跨行业公司合作的落地应用。<b>自然语言模型的场景化也是最大的困难。</b></li> </ul>
越过精调的坑	OpenAI将精调分为两个步骤：监督精调（SFT）+强化学习训练（RLHF）。其中RLHF是精调的重点。 1. Pertain：在训练之前先让模型独立观察，自由学习。 2. InstructionTuning：示范模型+简单调整 3. RLHF：对不可预料的行为作出反馈使答案来拟合人的偏好。	<ul style="list-style-type: none"> <li>➤ <b>国际化</b>：中国公司也需要思考如何在突破自己的技术的同时，<b>将自己的技术拓展到国际市场，在国际上产生影响力。</b></li> </ul>

## 2.1.3 国内大语言模型产品架构

我国目前的自然语言处理模型共有三个主要的产品架构，包括：模型层、中间层以及应用层。

### 模型层

**第一层是最底层的模型层。**

模型层是自然语言处理模型的核心部分，主要负责处理输入语言并输出生成内容。处于模型层的产品可以用来做垂类的基础大模型，比如一些用于生成虚拟角色或提高写作等其他生产力的底层模型。对于这类模型，企业可以选择自主研发基础模型，因为此类模型不需要过大的参数数量，如垂类的基础大模型仅需要百亿参数。同时，企业也可以选择研发开源模型，做类似OpenAI公司的API产品。

### 中间层

**大语言模型的第二层是中间层。**

中间层处于模型层和应用层之间，起到衔接模型层和应用层的作用。它可以处理并优化模型层输出生成的内容并使应用层更好的使用模型。中间层在底层模型的基础上去做配置和微调，包括结合自有数据飞轮做人类数据反馈。然后再针对不同行业的应用，结合该行业的数据做出应用的模型。

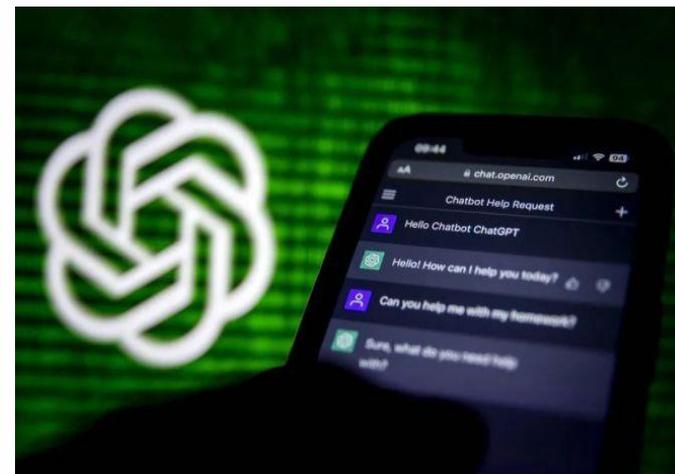
### 应用层

**应用层处于大语言模型产品架构的最上层。**

应用层主要负责将用户的输入文本在模型层的处理和中间层的优化后转化为用户可以理解的形式。应用层模型拥有多种形式如聊天机器人、智能客服等。应用层目前有三个主要方向：第一个方向是为了通过使用AI提升用户生产力，ChatGPT便是这个方向最典型的生产力工具；第二个方向是为了通过使用AI提升上层用户的交互体验；第三个方向是AIforScience，这个方向对科研相关的需求较多。应用层可以帮助用户更加便利的使用自然语言处理技术。

## 2.1.4 GPT给国内同行业公司带来的冲击

- **ChatGPT的面世给处于第一层（底层模型层）的中国企业带来的冲击最大**，尤其是一大批在2018年到2021年间成立的AI基建和AI模型公司。这些公司在AI方面非常注重资产化，会花费大量的人力和财力去训练垂向的服务，并将这些服务包装成API的形式。ChatGPT的出现会使这些公司前功尽弃，因为ChatGPT的能力大大超越了垂类的服务，包括图像生成、图像识别等。ChatGPT的生成效果甚至不需要通过训练，便能超越这些公司经过无数次训练得出的成果。它使这些AI基建和AI模型公司做出的API不再有意义。所以ChatGPT的出现严重冲击了处于第一层的公司。
- **ChatGPT的面世反而有助于处于第二层（中间层）和第三层（应用层）的公司降低成本、增加效率**。对于一个处于应用层或中间层的公司，ChatGPT的出现对其的负面影响相对较小，因为这类公司只是单纯的使用ChatGPT和其他的大规模语言模型并使用其他的API，并没有花费大量的人力财力去积累相关的研发经验和培养研发团队。这类公司甚至可以选择表现效果更好的OpenAI的API模型去做相同的任务。正是因为这类公司在机器学习方面是轻资产化的，所以冲击相对较小。不仅如此，这类公司直接面向客户，而不是面向开发者或B端。所以当像ChatGPT一样的大语言模型出现的时候，这类公司反而如鱼得水，能在很大程度上提高用户体验。



# 目录

---



## 1 大模型与GPT



## 2 国内篇

- 2.1 国内大模型发展现状
- 2.2 国内主流大模型



## 3 ChatGPT应用场景

## 2.2.1 国内主要大模型

我国目前有四种主要大模型，包括：计算机视觉模型、多模态预训练模型、NLP大模型以及图、文、音三模态模型，其中最常见的大模型为多模态预训练模型。学术院校、科技大厂以及相关的垂类公司都在致力于研究与开发自己的大语言模型，并已经初具规模，其中不乏拥有千亿级别以上参数容量的模型。目前，学术类研究院在大模型参数容量上处于领先地位，更有以清华大学为代表拥有百万亿级别以上参数的大模型。

国内主要大模型梳理

模型名称	具体企业	大模型	参数	数据量	算力
多模态预训练模型	清华大学等（学术类）	“八卦炉”（脑级AI模型）	174万亿	中文多模态数据集M6-Corpus	“海洋之光”超级计算机
	阿里（科技大厂）	通义M6大模型	10万亿	1.9TB图像、292GB文本	512块GPU
	腾讯（科技大厂）	混元	万亿级	五大跨模态视频检索数据集	腾讯太极机器学习平台
	华为（科技大厂）	盘古系列大模型	千亿级	40TB训练数据	鹏城云脑II、全场景AI计算框架MindSpore、2048块GPU
	澜舟（垂类公司）	孟子	10亿	数百G级别多领域高质量语料	16块GPU
NLP大模型	百度和鹏城实验室（科技大厂/学术类）	ERNIE 3.0 Titan	2600亿	纯文本和知识图谱的4TB语料库	鹏城云脑II（2048块GPU）、百度飞桨深度学习平台
	浪潮信息（垂类公司）	源 1.0	2457亿	5000GB高质量中文数据集	4095 Pflops-day/2128张GPU
计算机视觉模型	商汤科技（垂类公司）	书生（INTERN+）	100亿		商汤AIDC、峰值算力3740 Petaflops
		某世界最大计算机视觉模型	300亿		
图、文、音三模态模型	中科院自动化所（学术类）	紫东太初	千亿	基于万条小规模数据集	昇腾AI基础软硬件平台

## 2.2.2 国内主流大模型--百度文心一言

### ➤ 文心一言简介



文心一言是由百度推出的一款基于自然语言处理技术的文本生成模型。它通过不断的深度学习和训练数据来持续优化自己，以帮助用户获取信息、知识和灵感。该模型具有与人对话互动、回答问题、协助创作等功能，旨在为用户提供高效、精准、优质的文本生成服务。文心一言的技术基于飞桨深度学习平台和文心知识增强技术，通过持续从海量数据和大规模知识中融合学习，具备了知识增强、检索增强和对话增强的技术特色。

在人工智能的四层架构（芯片、框架、模型、应用）中，文心一言处于模型层，**定位于人工智能基座型的赋能平台**，旨在助力金融、能源、媒体、政务等各个行业的智能化变革。它具备文学创作、商业文案创作、数理逻辑推算、中文理解、多模态生成等五大能力，能够为各行各业的用户提供不同类型的文本生成服务。总之，文心一言是一款功能强大、性能卓越的中文文本生成模型，具有广泛的应用前景和\*\*市场价值\*\*。

### 文心一言的优势在于：

- **快速生成自然、流畅的中文文本内容**：能够生成与人类对话式的文本内容。并且与ChatGPT相比，文心一言在中文语言生成上更具优势。
- **可定制化**：可根据用户需求，个性化定制主题并生成内容。
- **可应用于多种场景**：文心一言可以应用于各行各业的各种场景，为人类生活提高效率、提供便利。

### 文心一言的劣势在于：

- **缺乏创造性**：文心一言是通过深度学习并训练数据生成文本，生成出的文本是基于数据的产出，所以相对缺乏创造性。
- **精准度有待提高**：文心一言生成的内容并不能做到精准，偶尔会有误差。
- **高度依赖数据**：因其高度依赖数据，若数据质量较差，做生成出的内容的质量也将受到影响。不仅如此，若数据内容具有偏见，则生成出的内容的也会有可能带有偏见。

### ➤ 文心一言的优势及劣势



## 2.2.2 国内主流大模型--百度文心一言

文心一言共有五大主要应用场景，包括：文学创作、商业文案创作、数理逻辑推算、中文理解、多模态生成。

### 文心一言的五大应用场景

应用场景	具体描述
文学创作	文心一言根据输入问题，自动总结分析了知名科幻小说《三体》的核心内容，并提供了五个创意续写角度。这些生成内容的准确性也得到了原作者的证实，并非胡编乱造。这个过程体现出文心一言在对话问答、总结分析、内容创作生成的综合能力。
商业文案创作	文心一言成功完成了为公司起名、写标语广告语、写新闻稿等创作内容。
数理逻辑推算	文心一言具备一定的思维能力，能够学会数学推演及逻辑推理等相对复杂任务。如面对“鸡兔同笼”这类经典数学题，文心一言能够准确理解题意，拥有正确解题思路，按正确的步骤算出正确答案。
中文理解	文心一言具备中文领域最先进的自然语言处理能力，在中国语言和中国文化上表现优异。
多模态生成	文心一言能够生成文本、图片、音频、视频。在音频生成方面甚至具备生成带有当地方言的音频。

文心一言的部分合作厂商



## 2.2.2 国内主流大模型--华为云盘古大模型

### ➤ 盘古大模型简介



盘古大模型是由华为公司自主研发的自然语言处理模型，共由五个模型组成，包括：CV（机器视觉）、NLP（自然语言处理）、多模态、科学计算以及语音。

盘古大模型的五个板块发展成熟度并不一致。其中CV（机器视觉）板块领先于其他四个板块，是目前业界最大的CV大模型并首次实现兼顾判别与生成能力，在ImageNet小样本学习能力上位于业界第一。盘古大模型的NLP模型也是目前业界首个拥有千亿级别参数的中文预训练大模型。但其在多模态、科学计算以及语音领域表现相对逊色。

盘古大模型共有四层架构，分别为：AI算力资源、框架MindSpore、AI开发平台ModelArts，最后到盘古大模型的落地。盘古大模型底层的AI算力资源在鲲鹏和昇腾的基础上使用了全场景AI计算框架MindSpore，为其在算力方面加持。并在第三层AI开发平台ModelArts上进行模型训练、管理、部署及数据处理，最终使盘古大模型成功落地并在五大板块快速发展。

### 盘古大模型的优势在于：

- **中文优化**：盘古大模型使用了大量的中文语料数据进行训练，使其能够更好的理解语义并输出更自然的中文文本。
- **应用广泛**：盘古大模型可使用于各行各业的各种领域，通过科技提高人类效率。
- **超大规模**：盘古大模型拥有千亿级别的参数容量，有能力处理更加复杂的任务。

### 盘古大模型的劣势在于：

- **训练成本高**：因盘古大模型拥有的巨大参数容量，需要消耗大量计算资源和储存资源，所以每训练一次的成本极高。
- **竞争压力**：目前国内众多企业都在研发大语言模型且成效优秀。盘古大模型在行业内竞争压力极大。
- **语义理解问题**：虽然盘古大模型拥有大量中文预料数据可以用来训练，但在处理更加复杂的任务时语义理解方面仍需优化。

### ➤ 盘古大模型的优势及劣势



## 2.2.2 国内主流大模型--华为盘古大模型

华为盘古大模型可用于各行各业等各种领域，包括教育、医疗、金融、法律、媒体、旅游等。

### 华为云的具体应用场景

应用场景	具体描述
教育领域	可以根据不同学生的不同学习情况，通过将这些情况转换为数据进行分析，为不同学生定制个性化的学习方案，提高学生的学习效率。
医疗领域	可以帮助医生检查患者病例并优化患者治疗方案，提高诊断的准确率和效率。
金融领域	可以根据不同客户的具体情况，为其自动生成投资计划，更好的帮助客户在投资过程中控制风险，提高回报率。
法律领域	可以帮助律师检查或撰写合同，提高质量和效率。
媒体领域	可以帮助媒体公司自动生成发表文章稿件，并分析用户喜好进行投放。
旅游领域	可以根据用户需求，自动生成旅游攻略，极大的节省用户时间，提高旅游质量。

盘古大模型的部分合作厂商

用友  
yonyou

金山办公  
KINGSOFT OFFICE

HOPERUN  
Information Technology  
潤和軟件

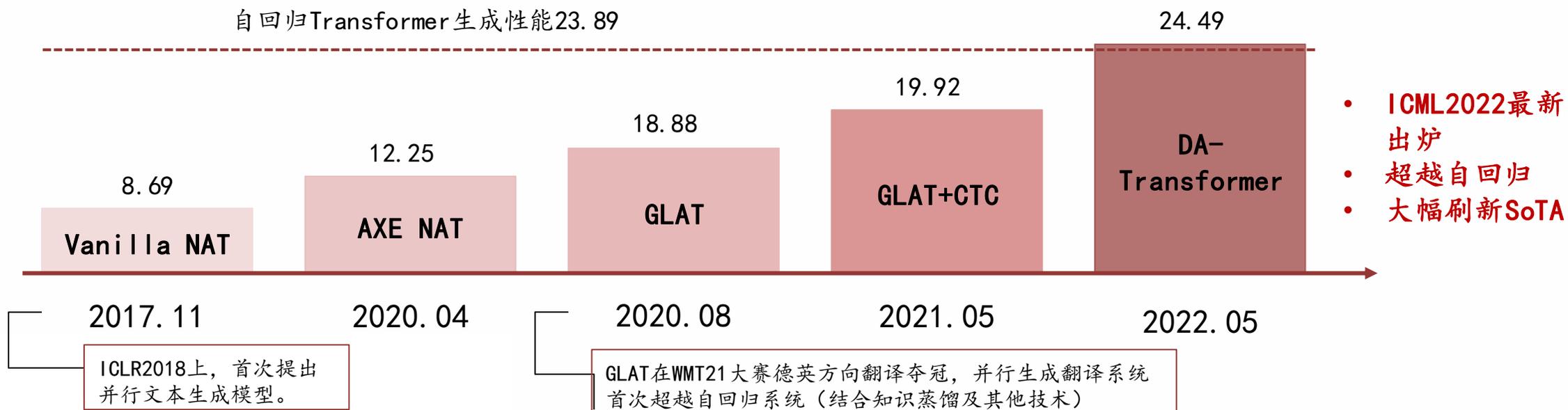
云从科技  
CLOUDWALK

## 2.2.2 国内主流大模型—字节跳动

在AIGC方向，字节选择了非ChatGPT路线。字节跳动的研究成果包括非自回归模型DA-Transformer、端到端语音到文本翻译模型ConST、多颗粒度的视觉语言模型X2-VLM、图片和文本统一生成模型DaVinci和语言理解模型Neural Symbolic Processor。

- 非自回归模型DA-Transformer是更快的Transformer模型。Transformer的一个大问题是解码基于自回归，计算需要一环扣一环地进行，速度比较慢。而DA-Transformer在解码层的最后一层构建有向无环图，并且在解码器输入的各个位置上进行并行计算，极大提高了计算速度。
- 清华&字节联合提出的DA-Transformer摆脱了传统并行模型依赖知识蒸馏的问题，在翻译任务中大幅超越了之前所有并行生成模型，最高提升4.57BLEU。同时也首次达到、甚至超越了自回归Transformer的性能，在最高提升0.6BLEU的同时，能够降低7倍的解码延迟。

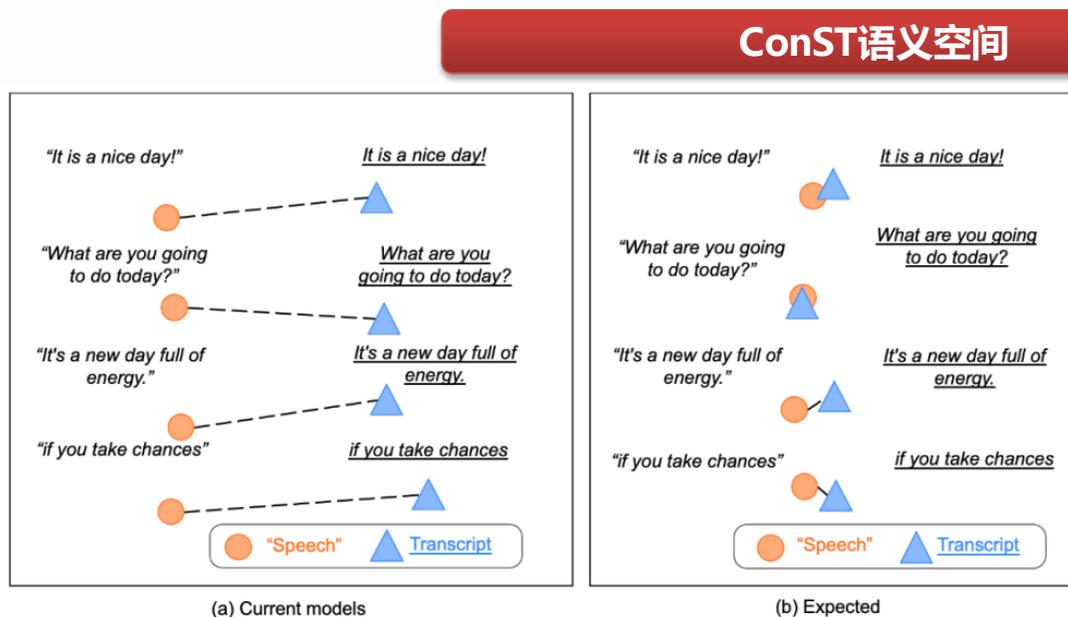
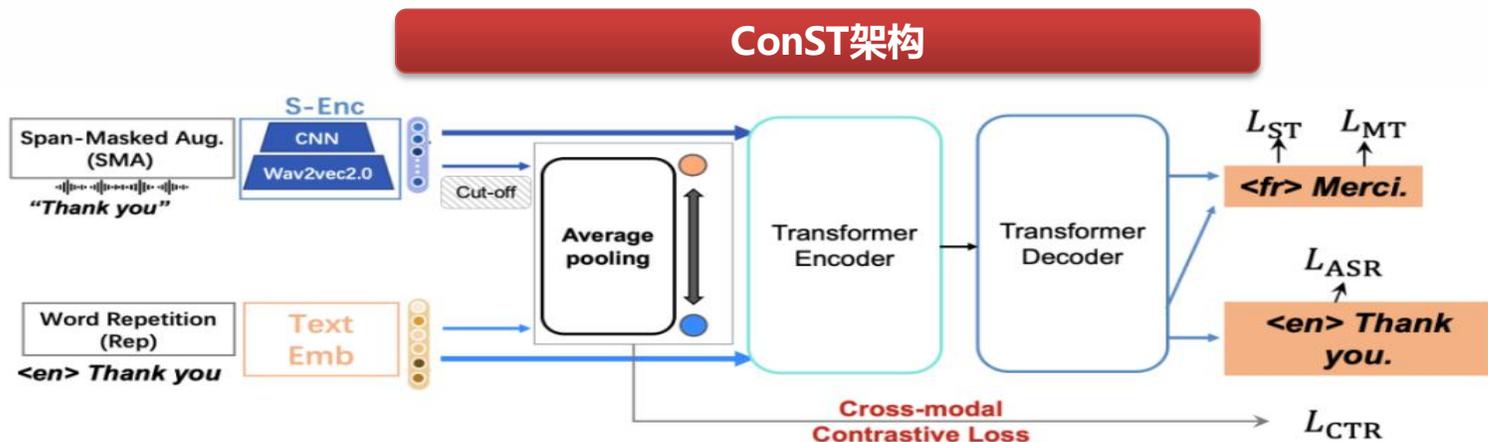
### 并行（非自回归）生成模型的发展



## 2.2.2 国内主流大模型—字节跳动

端到端语音到文本翻译模型ConST实现了语言和语音的结合。可以直接将英语的语音翻译成中文的文本，而且在语音到文本的翻译中，达到了SOTA (stateoftheart) 的效果。ConST的架构由Transformer的编码器和解码器组成，编码器既可以接受语音输入又可以接受文本输入。输入是语音时有特殊的前处理模块，使用wave2vec2和CNN。

ConST的最大特点是，使用对比学习将语义相同的语音输入 的表示和文本输入的表示拉近。可以理解为对表示学习进行了正则化。在非ConST架构的多任务学习模型中，同一句话，同一个意思，语音和文本的表示依旧有较大差距，而我们期望的是，对于同一句话，比如语音的 “it saniceday!” 和文本的 “it saniceday!” 的表示应该是相近的。如 “ConST语义空间” 意图所示，左图：多任务学习框架下语音和文本两者表示之间依旧存在差距；右图：我们所期望的两个模态表示：相同意思的语音和文本应该有相近的表示。

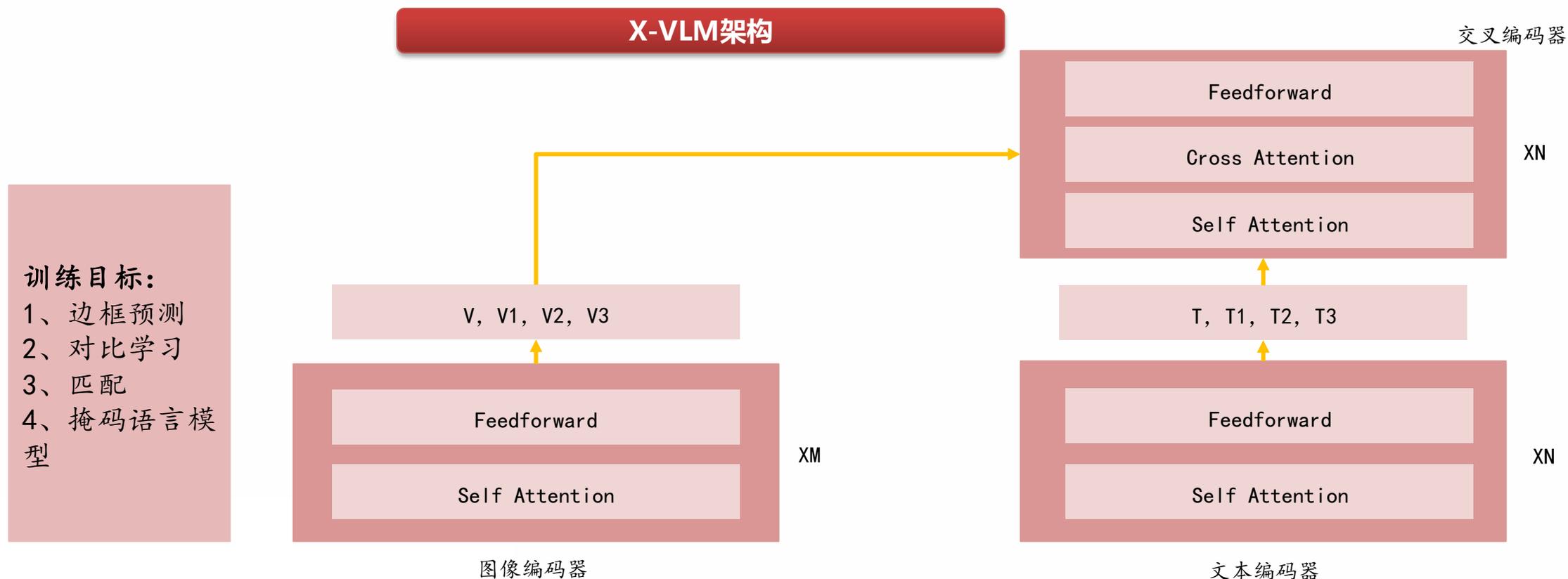


左图：多任务学习框架下语音和文本两者表示之间依旧存在差距；右图：我们所期望的两个模态表示：相同意思的语音和文本应该有相近的表示。

## 2.2.2 国内主流大模型—字节跳动

**多颗粒度的视觉语言模型X-VLM**，以Transformer为基础，以文本-图片对数据作为输入进行预训练得到的视觉语言模型，可以用多种跨模态的下游任务。假设文本和图片对的内容是强关联的，文本描述图片内容，但描述是多颗粒度的。文本可能描述图片整体、区域或物体。这种基础模型对visualquestionanswering和visualgrounding等任务等更加适用，也可以用于其他任务。X-VLM是目前视觉语言各种任务的SOTA。

**X2-VLM将X-VLM扩展，也可以处理视频和多语言。**



## 2.2.2 国内主流大模型—腾讯混元助手

当前腾讯混元AI大模型) 完整覆盖NLP大模型、CV大模型、多模态大模型、文生图大模型及众多行业/领域任务模型。混元AI助手作为业界首个可在工业界海量业务场景直接落地应用的万亿NLP大模型，先后在热启动和课程学习、MOE路由算法、模型结构、训练加速等方面研究优化，大幅降低了万亿大模型的训练成本。

### 腾讯混元AI大模型架构



## 2.2.2 国内主流大模型—阿里M6大模型

阿里达摩院在2020年初启动中文多模态预训练模型M6项目。2022年9月阿里正式发布通义大模型，通过统一学习范式M6-OFA和模块化的设计，提升大模型跨模态能力和效率。通义大模型整体架构最底层为统一模型底座，中间基于底座的通用模型层覆盖了通义-M6、通义-AliceMind和通义-视觉，行业模型层深入电商、医疗、娱乐、设计、金融等行业。

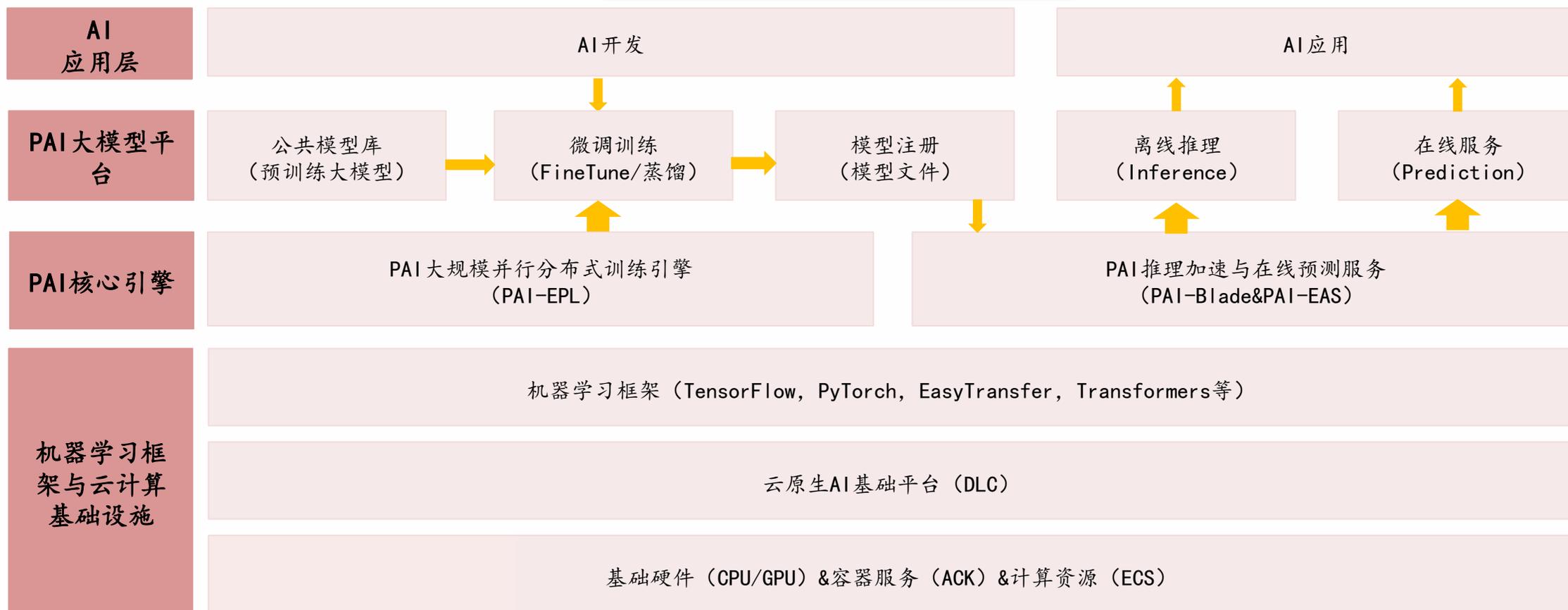
### 阿里通义大模型架构



## 2.2.2 国内主流大模型—阿里M6大模型

M6英文全称为MultiModality-to-MultiModalityMultitaskMega-transformer，是达摩院开发的通用人工智能模型。**M6模型基于预训练大模型开发，拥有多模态、多任务能力**，2021年6月，M6达到万亿参数，成为国内首个实现商业化落地的多模态大模型。

### M6模型底层-PTM架构



## 2.3 国内大模型的发展之路

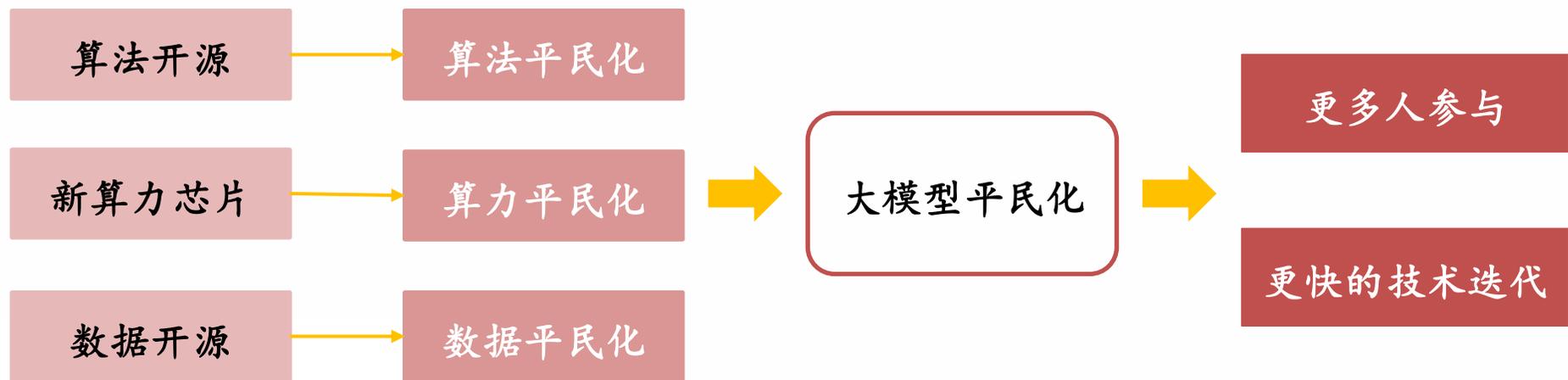
### ◆ 平民化大模型

随着算力、算法和数据三大关键环节实现平民化，大模型平民化时代即将到来。大模型平民化时代，将有更多的人参与进来，推动大模型技术进步。平民化的解决方案可以加速初级玩家在自有场景下的大模型应用落地。随着数据、算法等认知壁垒下降，更多人可以针对性训练大模型、提高大模型训练灵活性，不断试验出更好的应用方法。

### ◆ 大模型发展是持久战

炼大模型是一项长期主义的持久战。未来五年甚至十年大模型都会持续火热且重要，无论是虚拟世界还是物理世界，都会被大模型重新塑造一遍。我国在未来还有很大空间和机会可以追赶

### 大模型平民化发展之路



## 2.3 国内大模型的发展之路

### ◆ 垂直整合型大模型

搭建与自有应用、生态结合的模式是国内发展大模型的一种比较务实的做法。国内暂时无法做出OpenAI和Google一样的天花板级别的大模型，垂直整合型大模型是另一条发展路径。国内厂商结合自有场景和生态，可以做出参数规模达到千亿级，具备通用能力以及很强的对话推理能力的模型。如字节在抖音的内容服务场景下做出内容生成的垂类大模型，阿里在电商场景、云服务场景下的垂类大模型等。

### ◆ 国内大模型厂家百花齐放

**中国大模型格局：科技进步是全人类集体智慧线性叠加过程。**未来一两年后，国内至少有五十家以上公司在做大模型。中国的大模型玩家：四家大厂：BAT和字节；十几家中型互联网厂商；十几家提供ToB服务的AI公司；以及一些垂直领域的甲方，比如大的银行、保险公司、汽车、手机厂商，也会有属于自己的大模型。大模型时代机会众多，真正的竞争壁垒是如何在新形态，理解商业、信息、关系的flow，以及把握新形态商业模型的演变趋势。

### 我国大模型竞争格局设想



# 目录

---



**1 大模型与GPT**



**2 国内篇**



**3 ChatGPT应用场景**

## 3.1 ChatGPT的应用落地

Chatgpt的应用落地场景按照架构层级，可分为：模型层、中间层和应用层。

- 模型层是以训练模型为基础搭建的AIGC技术基础设施层，按照模型逻辑可分为闭源LLM，开源LLM和LLM/Action混合架构；
- 中间层是在模型层上通过精调和训练之后形成的垂直化、场景化和定制化的小模型和应用工具层；
- 应用层是面向C端和B端客户的AIGC产品和服务，按照内容可划分为AIforProductivity、AIforExperience以及AIforScience。

### ChatGPT应用落地层级

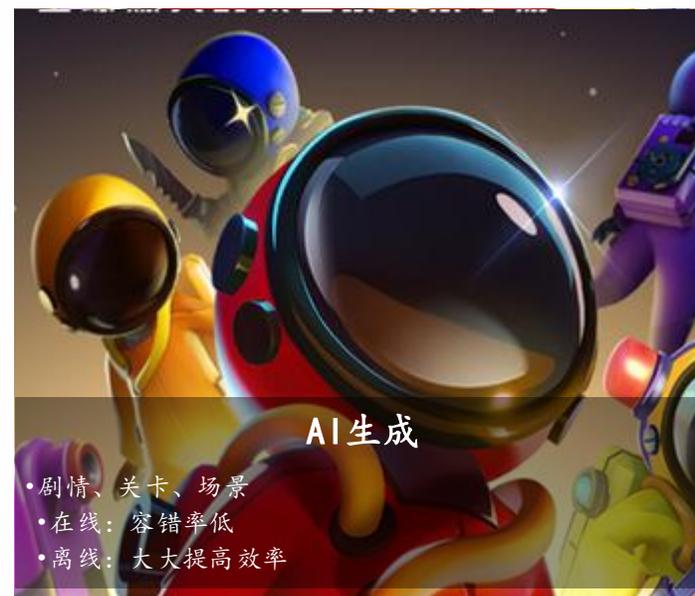


## 3.2 ChatGPT应用落地—游戏类

**游戏是发展AI技术的绝佳场景。**游戏是最复杂的娱乐形式，有较高的互动型和实时性的体验要求。且其涉及的资产类型众多，开发成本高昂、开发周期漫长，因此新游戏开发者进入门槛较高。而AIGC相较传统AI，在速度提升和成本下降的同时，保证了质量的提升，带来了更高的创造能力、更快的更新速度和实时的创作过程。

**目前AIGC尤其是ChatGPT已经被广泛应用于游戏生成、游戏维护和游戏运营中，主要有AIBot（AI虚拟玩家）、AI策划和AI生成三种形式。**

### ChatGPT的游戏类应用落地场景



## 3.2 ChatGPT应用落地—游戏类



### ➤ AIBot (虚拟玩家)

**提升用户留存率，加强新手游戏体验感。**AIBot已大规模在策略类、竞技类游戏中落地。通过扮演玩家的对手或队友，或者在用户掉线时接管，AIBot可以显著地提升用户留存，降低新玩家的挫败感，提升用户体验。

**生成NPC，提供情绪价值。**通过生成游戏里面的NPC，AIBot可以帮助玩家了解游戏世界观，提供情绪价值。AIBot在GPT-4架构下，被赋予了对话能力、决策能力，可以生成更有温度的、有感情的NPC，为玩家提供情绪价值，帮助玩家了解游戏世界观，充当内容的载体。

### ➤ AI策划

**辅助策划游戏的玩法和数值设计。**传统的策划下，当游戏数值膨胀到一定的程度以后，由于游戏内的武将、战法过多，将超出人为可控制的范围。而游戏参数调整是一个很经典的组合优化问题，可以用机器学习解决。引入AI策划后，可以很容易找出游戏的数值阵容和平衡性问题，在每个赛季推出新的武将角色时，不至于破坏整个游戏的平衡。在一些策略游戏中，数值的平衡性是非常重要的，如果推出一个武将，就把原有的数据平衡破坏掉，对整个游戏的收入都是一个很大的灾难。



### ➤ AI生成

**主要用于美术资产生成、剧情文本生成和关卡生成。**AI生成被大规模应用在游戏领域中，以上三个环节都是大模型擅长的的工作。

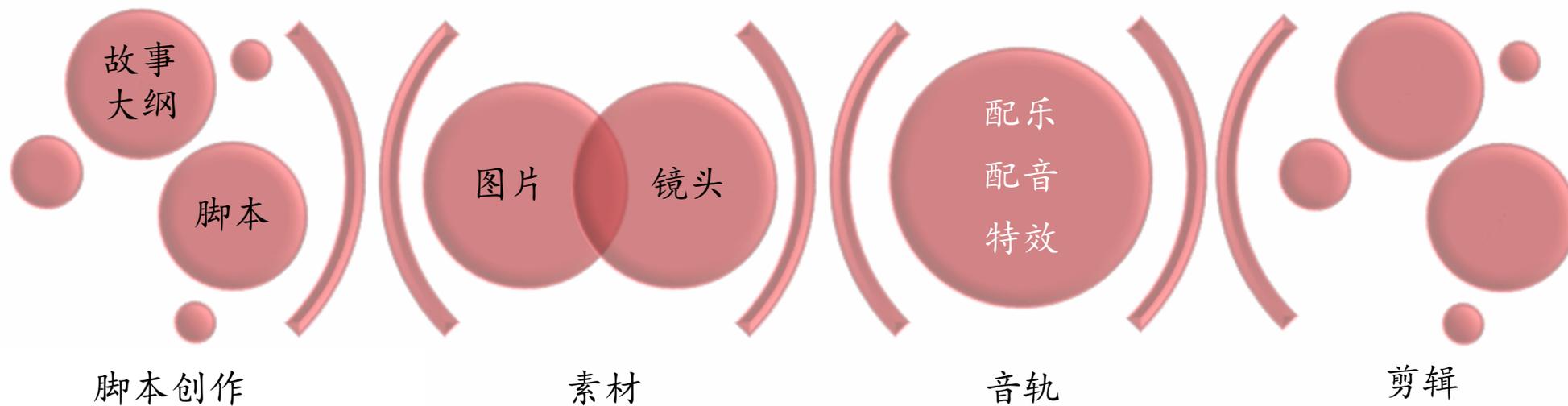
**游戏场景在线生成容错率低，AIGC离线生产能极大提高游戏开发效率。**大模型如果用在在线上，对它的可靠性要求是非常高的，一旦出错，很有可能造成不好的游戏体验；但如果离线生产，及时AIGC可靠性不高，也可以大大节省时间，例如将原画创作时间由原来的10h/幅缩减至10min/幅，可以指数级得提升效率。

### 3.3 ChatGPT应用落地—视频类

**AIGC技术可以通过两种方式落地视频场景。**一种是使用AIGC技术进行视频的自动化生成和AI辅助创作。进行视频自动生成，**AIGC可以接手从脚本创作、素材寻找、音轨制作到剪辑和拼接等视频创作的全过程。**

- 脚本，AIGC可以完成从故事大纲到脚本等一系列故事梗概的创作；
- 素材，在多模态下，AIGC生成海量素材库，包括图片、镜头等高质量素材；
- 音轨，AIGC完成音频生成、配乐创作和特效声音的创作等；
- 剪辑，通过对话让Chatgpt完成视频剪辑。

#### 视频类应用落地场景 ( character、AI/D.ID )



## 3.3 ChatGPT应用落地—视频类

### ➤ 视频生成的进阶

#### 静态→动态

**从生成静态的图片到动态的视频创作。**

**动作模拟是AI视频生成的难点**，尤其是人的动作或动物的动作，要让它变得流畅。比如从上一张图片过渡到下一张图片，一个动作是如何生成，动作的模拟是很难完成的。

#### 情绪处理

**情绪表达是目前AI生成视频区别于Human生成视频的主要区别。**

**AIGC目前仍需解决情绪处理问题。**如果生成一个虚拟人，或新生成一个人的图片，要把它的情绪做出来，比如愤怒、微笑等情绪表达，这部分的处理是AIGC目前较大的挑战性问题之一。

#### 3D建模

**3D类内容需求蓬勃增长。**

**3D内容创作的核心是建模**，目前StableDiffusion和Runway可以做到一些基础的建模工作，如换角、换镜头、换角度和生成不同的图片。然而，创建专业的3D内容需要很高的艺术与审美素养和大量3D建模专业知识。人工完成这项工作需要花费大量时间和精力来培养这些技能。为引入扩散模型的传统AI在制作的时候也只能从单个模型入手开始建模，经历DreamFusion、GET3D、3DiM等多个建模方法，目前AIGC在3D建模领域仍有前进空间。

StableDiffusion



Stable Diffusion

Runway-Gen2



## 3.4 ChatGPT应用落地—生物医药类

### ➤ AI在生命科学领域的发展历程和作用

AI或数字化计算在生命科学里有两个发展阶段。第一，在生成式AI面世之前，大众更关心如何通过AI手段更好的去理解自然界已有蛋白和他们的结构。一个重点应用就是通过积累以x光为主的结构生物学数据，让Alphafold2可以对单链蛋白做到准确预测，也就是用预测的方式预测已知序列的样子。第二，在生成式AI面世之后，科学家或生物化学公司能够在无穷大的蛋白质序列和结构空间里去创造之前不存在的蛋白质。AI会对整个生命科学、创新药的研发等带来更多的机会和相互促进的应用场景。

早期的ChatGPT便已经让科学家在医疗领域看到很多非常明确的垂直应用的可能性。它对于数据隐私，尤其医疗领域的数据隐私保护和应用起到了非常大的帮助。二者的优势相辅相成，能够解决很多医疗系统现在可能正在面临的问题。

ChatGPT提供了两个方面的实际帮助。第一，ChatGPT在未来能够在生命科学里更好的运用数据，因为所有的AI生成的基础是拥有高质量的数据。第二，新一代生成式AI能够生成一些之前自然界完全不存在的新的蛋白。



### 3.4 ChatGPT应用落地—生物医药类

#### ➤ 关键应用场景



医疗领域会是生成式AI在未来最大的应用领域。它最关键的应用场景就是治病，解决人类或者生物界在疾病方面和衰老方面的终极问题。最近的目标是治疗癌症。生成式AI或许可以在未来的研发药物中也起着重大作用，ChatGPT能够帮助科学家在研发药物的过程中高效省时的找到最匹配的分子去对应靶点这也是药物研发最困难、最耗时、最耗财力的一个环节。结合生成式的AI，能够让创新药企在最早期，能够探索比之前广阔得多的大分子空间。如果没有这样的技术，靠传统的方法，很多的疾病靶点几乎没有可能被做出来。唯一的不足是，ChatGPT具体能帮助药物研发节省多少时间很难量化评估。

#### ChatGPT的具体应用场景

具体应用场景	主要内容
医疗虚拟助手	帮助患者预约挂号、辅助患者接受治疗以及管理患者健康信息。
临床决策支持	为医生提供改善治疗方案的建议，为医生节省时间、减少错误。
病例保存	自动对患者病例自动进行总结，帮助医生快速提取患者信息，提高效率。
医疗翻译	帮助医生和患者准确快速翻译医学术语，提高医患沟通效率。
远程健康助手	为患者提供远程健康建议、药物说明或解答患者医学健康方面问题等一系列服务。
医疗写作和文档	辅助医生编写医疗报告，提高医生效率，减少错误。
患者分诊	通过与患者互动，将患者按病情的严重程度进行分诊，提高效率。

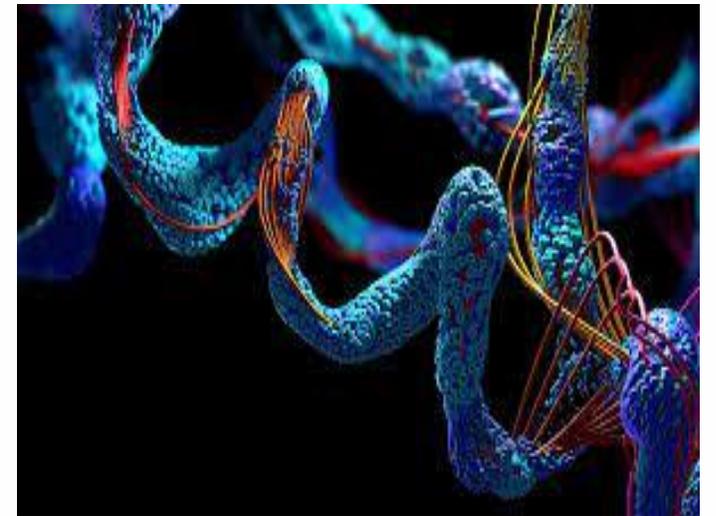
## 3.4 ChatGPT应用落地—生物医药类

### ➤ 发展现状及发展难点

ChatGPT是一个具有普适性，并可以将人类语言连续处理的语言模型。它只能改善掌握知识能力的方向，但还不具备创新能力。生成式AI缺乏独立思考的能力，虽然可以在普适的范围上对各个行业有更多的影响，但对于医药行业，尤其是医药研发，目前影响比较有限，只能起到辅助作用。

在生物医药领域，目前Alphafold影响更大一些。ChatGPT仅仅是自然语言的逻辑处理，一种互动式搜索，需要大量的算力和数据去训练它。但是Alphafold的重要性在于它局限在某一个单一的领域之内，它得到的算法数据特别针对的是蛋白质结构。而ChatGPT得到的数据不是仅仅针对生命领域，它的数据来自于各个行业。所以Alphafold目前对行业的影响更大。但是Alphafold的缺点是只有行业内的人士或者特定的科学家才会去接触，普适性低。如果ChatGPT想对生物行业产生影响，它要深入到特定的领域去开发出自己的特殊版本。

医疗领域数据目前有两大问题：数据孤岛和隐私问题。由于各个不同的医疗系统，医疗体系包括数据生产方、数据持有方，彼此之间体系不同，所以很难进行数据的交互和沟通，数据也会缺乏多样性。在数据隐私层面上需要更好的解决方案去帮助大家愿意把数据分享出来。如联邦学习Federal Learning的核心就是帮助解决数据隐私敏感的问题，在不泄露第三方数据的前提下，运用第三方数据生成更好的方案。



## 3.4 ChatGPT应用落地—生物医药类

### 发展前景

生成式AI正处于积极快速发展的大趋势和全产业数字化转型的大趋势，且这个趋势是不可逆、不会减速的，它只会加速ChatGPT的发展并以更快的商业落地的形式呈现在大众面前。从美国整体的情况来看，大量的生物医药公司对银行贷款的依赖度较低。即使硅谷银行倒闭，跟硅谷银行相关的一些生物制药或者生物医药科技公司在硅谷银行所存放的现金或现金等价物有限，所以受损的规模是可控。**对于未来的AI，在医疗或者生物科技的方向来看，仍呈积极的发展趋势。ChatGPT的出现或许可以加速临床试验阶段，帮助整合临床实验数据和流程，缩短试验周期从而减低成本。**

**未来做生成式AI生物医药领域应用层的核心竞争力取决于公司生成数据的能力和公司是否能用数据很好的与算法相结合。**生成式AI里也会有很多不同种类的模型，比如以Profluent为代表的基于ChatGPT模型（语言生成）的公司，它能够提供一个序列，但是并不能够精准地提供蛋白质的结构。另外就是以GenerateBiomedicines为代表的基于扩散模型（图片生成）的公司，它可以生成蛋白质的骨架，但是没有办法给出一个精确的蛋白质氨基酸的序列。**企业接下来的目标就是将ChatGPT和扩散模型相结合，也就是将语言生成与图片生成式AI相结合，同时预测蛋白质序列与结构。**

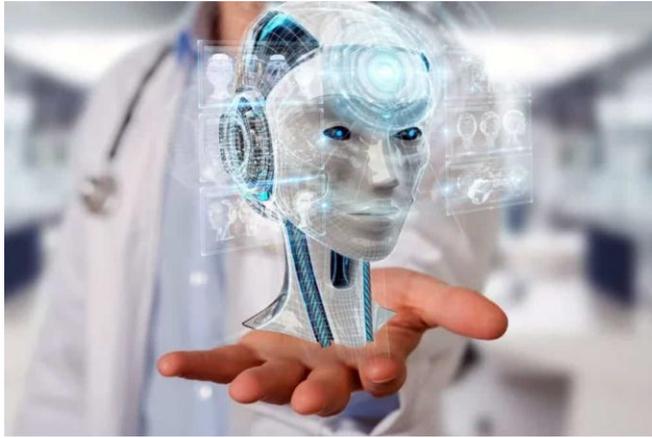
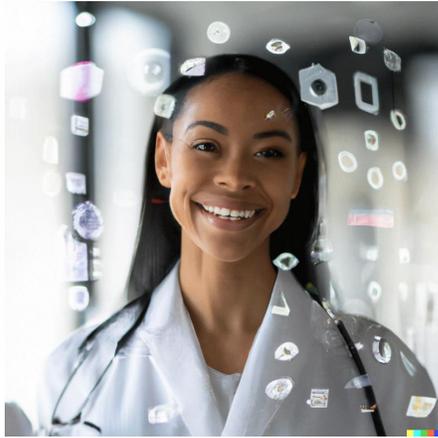
**在我国，生成式AI的商业模式会倾向于使用开源模型，这样也会拥有更大的自主权。**底层模型的差异并不大，更关键的还是数据。生成式AI未来在生物医药领域也大概率会出现百花齐放的现象。**国内企业应十分坚定的选择运用生成式AI，当新的技术到来，如果不选择与之合作，都会被远远地甩在时代脚步后面。生成式AI将会对生物医药行业带来一次洗牌，大家都会站在一个全新的起点上，过去的优势现在也变得无关紧要。抓住机会，结合自己的一些独特的资源和优势，就能够在换道的过程当中领先行业**



### 3.4 ChatGPT应用落地—生物医药类

#### ChatGPT不会取代医生或药物研发人员职业

ChatGPT及生成式AI功能强大，但不会取代医生或药物研发人员的职业。人工智能是用来助力、加速生物医药领域的发展的。它仅仅起到提升效能的作用，而不是替代的作用。人工智能也没有拥有能够替代医生或研发人员的能力，更没有办法代替高度专业化的工作。所以，并非人工智能会替代掉医生，而是未来使用人工智能的医生会替代掉不会使用人工智能的医生。所有新技术的应用目的应是去解放生产力和生产效能，可以把更多的时间专注在更加有创造性、开拓性、研发性的工作上。人工智能不可能替代人类的作用，它在现实的医疗中，只是起到辅助的作用。虽然一些简单重复的岗位会被人工智能替换掉，但是这种取代会释放出更多的人才去做更有创造性的工作，也会有更多原创性的岗位会从中产生。ChatGPT只是会使我们的生活或者工作具有更多的多样性，越来越垂直式的向下接近科学的本质，而不是替换。



#### 不同模型将会起到不同作用

虽然目前Alphafold对生物医药领域的影响力更大，但这两个不同的模型会对不同纬度产生影响，没有直接的可比性。目前ChatGPT3.5版本确实不具备研发能力。4.0版本能够处理更加复杂的任务，但是否具备研发能力，要等真正产品发布之后才能确定。3.5版本目前在生命科学的领域更多的是工具属性，包括去设计一套个性化的方案，而不是用来研发创新。Alphafold更偏向基础的创新研究，它对于行业的贡献角度是基础技术层面上的。



## 3.4 ChatGPT应用落地—生物医药类

随着人口老龄化和慢性疾病的增加，ChatGPT及生成式AI在医药生物领域的发展前景十分广阔，其中有六大发展方向，缺一不可。

### ChatGPT在生物医药领域的发展方向

发展方向	具体内容
加强技术	ChatGPT和生成式AI目前只能在生物医学领域起到辅助作用，并不具备创新功能，更无法做药物研发相关工作。 <b>未来仍需通过加强技术以使其具备更优秀的创新能力。</b>
深入到特定领域	ChatGPT覆盖的领域过广， <b>若想使其应用到生物医药领域，需要深入到这个领域并开发单一领域的特殊版本。</b>
解决数据孤岛	<b>未来需统一各医疗体系，以达到加强数据的交互与共享，增强数据多样性。</b>
解决隐私问题	<b>解决隐私问题可以帮助数据更好的共享，从而得到更优质的输出内容。</b>
提升数据质量	<b>高质量医疗数据的整合是技术顺利落地的关键要素，新一代AI对数据的要求非常高。</b>
提升准确性	因为生物医药领域本身的高严谨性， <b>生成内容的高度准确性是ChatGPT和生成式AI在生物医药领域重中之重。</b>



西南证券

SOUTHWEST SECURITIES

分析师：王湘杰

执业证号：S1250521120002

电话：0755-26671517

邮箱：wxj@swsc.com.cn

分析师：叶泽佑

执业证号：S1250522090003

电话：13524424436

邮箱：yezy@swsc.com.cn

## 西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上 持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间 中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间 回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间 卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上 跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间 弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

## 分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

## 重要声明

西南证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会核准的证券投资咨询业务资格。

本公司与作者在自身所知范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供本公司签约客户使用，若您并非本公司签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司也不会因接收人收到、阅读或关注自媒体推送本报告中的内容而视其为客户。本公司或关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行或财务顾问服务。

本报告中的信息均来源于公开资料，本公司对这些信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告，本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，本公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

本报告及附录版权为西南证券所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用须注明出处为“西南证券”，且不得对本报告及附录进行有悖原意的引用、删节和修改。未经授权刊载或者转发本报告及附录的，本公司将保留向其追究法律责任的权利。



# 西南证券研究发展中心

## 西南证券研究发展中心

### 上海

地址：上海市浦东新区陆家嘴东路166号中国保险大厦20楼

邮编：200120

### 北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

### 深圳

地址：深圳市福田区深南大道6023号创建大厦4楼

邮编：518040

### 重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼

邮编：400025

## 西南证券机构销售团队

区域	姓名	职务	座机	手机	邮箱
上海	蒋诗烽	总经理助理、销售总监	021-68415309	18621310081	jsf@swsc.com.cn
	崔露文	销售经理	15642960315	15642960315	clw@swsc.com.cn
	薛世宇	销售经理	18502146429	18502146429	xsy@swsc.com.cn
	汪艺	销售经理	13127920536	13127920536	wyyf@swsc.com.cn
	岑宇婷	销售经理	18616243268	18616243268	cyryf@swsc.com.cn
	张玉梅	销售经理	18957157330	18957157330	zmyf@swsc.com.cn
	陈阳阳	销售经理	17863111858	17863111858	cyyyf@swsc.com.cn
	李煜	销售经理	18801732511	18801732511	yfliyu@swsc.com.cn
	谭世泽	销售经理	13122900886	13122900886	tsz@swsc.com.cn
	卞黎旸	销售经理	13262983309	13262983309	bly@swsc.com.cn
北京	李杨	销售总监	18601139362	18601139362	yfly@swsc.com.cn
	张岚	销售副总监	18601241803	18601241803	zhanglan@swsc.com.cn
	杜小双	高级销售经理	18810922935	18810922935	dxsyf@swsc.com.cn
	杨薇	高级销售经理	15652285702	15652285702	yangwei@swsc.com.cn
	胡青璇	销售经理	18800123955	18800123955	hqx@swsc.com.cn
	王一菲	销售经理	18040060359	18040060359	wyf@swsc.com.cn
	王宇飞	销售经理	18500981866	18500981866	wangyuf@swsc.com
	巢语欢	销售经理	13667084989	13667084989	cyh@swsc.com.cn
广深	郑龔	广深销售负责人	18825189744	18825189744	zhengyan@swsc.com.cn
	杨新意	销售经理	17628609919	17628609919	yxy@swsc.com.cn
	张文锋	销售经理	13642639789	13642639789	zwf@swsc.com.cn
	陈韵然	销售经理	18208801355	18208801355	cyryf@swsc.com.cn
	龚之涵	销售经理	15808001926	15808001926	gongzh@swsc.com.cn
	丁凡	销售经理	15559989681	15559989681	dingfyf@swsc.com.cn