之江实验室 Zhejiang Lab

基础理论研究院　人工智能与安全团队

生成式大模型安全与隐私白皮书

作者:

徐晓刚，吴慧雯，刘竹森，李想，涂文轩，梁伟轩，张毅，刘哲

2023 年 6 月 6 日

The development of the Generative AI, e.g., Large Language Models (LLM), have been popular in both academic and industrial communities on a worldwide scale, especially the ChatGPT series. The success of ChatGPT and GPT4 has shown the future direction of developing AGI. However, large generative models also suffer from the issue of data/model security and privacy. We should note that large generative models would bring a lot of security and privacy problems, when they demonstrate great power in changing our life, such as data leaking and the propagation of fake news. In this white paper, we first conclude the development of large generative models, including its effects and social influences. Then, we summarize the current security and privacy problems in existing large generative models, e.g., the data and model security, copyright problems, and ethical issues. Finally, we give the corresponding suggestions about the current security and privacy problems. They can be employed to point out future research and develop directions, and can also be utilized as references for government decision-making.

目录

1 序言

OpenAI 于 2022 年 11 月 30 日开放测试 ChatGPT，此后 ChatGPT 风靡全球，在 1 月份的访问量约为 5.9 亿。AI 驱动的聊天机器人 ChatGPT 成为互联网发展二十年来增长速度最快的消费者应用程序。ChatGPT 和 GPT4 的诞生引发了生成式大模型的研发热潮，显示了人类迈向通用人工智能（AGI）的可能性。

但在其备受追捧的同时，ChatGPT 等生成式大模型也面临 AI 自身数据和模型方面的安全隐患。我们应该意识到，在生成式大模型带来各种革命性的技术进步的同时，其自身带来的一系列安全与隐私问题也值得我们注意，例如引发的数据泄漏，助长虚假信息传播等。在本白皮书中，我们首先总结了 ChatGPT 与 GPT4 等生成式大模型模型的发展历程，以及其带来的各种令人惊叹的能力和社会变革，社会应用等。而后，我们归纳了 ChatGPT 与 GPT4 等生成式大模型中存在的安全与隐私问题，包括数据安全，模型使用安全，版权问题，伦理问题等。最后，我们为应对这些安全与隐私问题提出了相应的应对策略，重点强调了日后亟需进行的研究和法规调整等。特别是为之后 AGI 技术的持续革新，起到未雨绸缪的预防。

2 生成式大模型的发展之路

GPT（Generative Pre-trained Transformer）是一种基于 Transformer 模型的语言生成模型，由 OpenAI 团队开发。自 2018 年发布以来，GPT 系列模型已经成为自然语言处理领域的重要研究方向之一。图1概括了当前国内外有关 GPT 的研究现状。下面我们将先介绍 ChatGPT 与 GPT4 出现之前的模型，而后介绍 ChatGPT 与 GPT4 的原理与特点，之后将列举在 ChatGPT 与 GPT4 之后涌现的一系列代表性生成式大模型。

2.1. ChatGPT 和 GPT4 的前身

如图2所示，本文将按照时间顺序介绍几种代表性的 GPT 方法。

2.1.1 GPT1

2017 年，Google 推出 Transformer，利用注意力机制完全替代过往深度学习中的卷积结构，直白地展现出了 “大一统模型” 的野心。2018 年 6 月，OpenAI 通过论文《Improving Language Understanding by Generative Pre-Training》[47] 推出了基于 Transformer Decoder 改造的第一代 GPT（Generative Pre-Training），该

	名称	发布者
国内高校	ChatGLM	清华
	MOSS	复旦大学
	紫东太初	中国科学院
国内企业	盘古	华为
	混元	腾讯
	百度文心	百度
	阿里通义千问	阿里巴巴
	言犀	京东
	书生	商汤科技
	SparkDesk	科大讯飞
	360 智脑	360
国外高校	Alpaca	Stanford
国外企业	GPT1、GPT2、GPT3、GPT3.5、GPT4InstructGPT	OpenAI
	LLaMa	Meta AI
	Bert	Google
	BLOOM	BigScience
	OPT	Meta AI
	GPT-Neo	EleutherAI
	Luminous	Aleph Alpha

图 1: 国内外生成式大模型研究现状总结

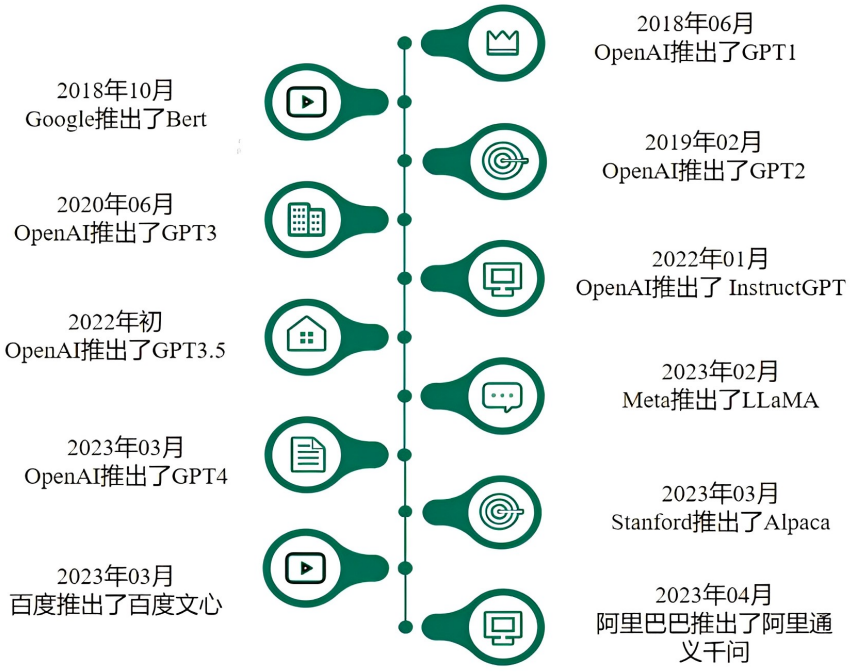

图 2: GPT 系列模型的发展历程总结

模型是最早的将 Transformer 以多层堆叠的方式构成语言模型的模型，有效证明了在自然语言处理领域上使用预训练和微调方式的有效性。类似地，在计算机视觉领域，先预训练后微调的方式盛行已久：先用海量有标注的数据集，通过有监督的训练生成一个预训练模型，然后通过下游任务，在这个模型上做微调。但是在自然语言处理中，这个方式一直很难做起来，原因是：缺乏大量标注好的文本数据集、

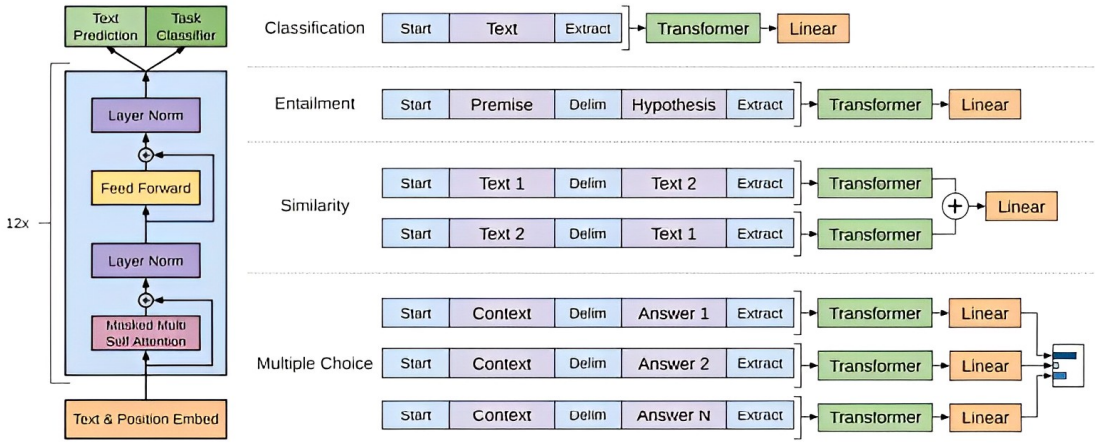

图 3: GPT1 示意图

比起图像信息，文字的信息更难被模型理解。Transformer 出世后，模型对文字上下文的理解能力得到显著增强，在这一铺垫下，GPT1 诞生了。如图3所示[1]，它的整体设计思路如下：首先，用无标注的数据（可以理解为一段普通的文字）训练一个预训练模型。在这个环节里，我们培养模型文字接龙的能力，也就是给定前 k 个词，模型能预测出第 k+1 个词。然后，在模型能够理解文字含义的基础上，用有标注的数据训练模型去定向做一些下游任务。例如文本分类，文本相似性比较等。有标注的数据集是远小于无标注数据集的，在这个环节，我们只是对模型做了一些微小的调整。

（1）GPT1 的优缺点

- 优点：GPT1 是**第一个使用** Transformer **自回归模型的自然语言处理模型**，可用于各种文本语言任务，如机器翻译，文本生成，对话生成等。
- 缺点：GPT1 没有全面的站点，在处理复杂的对话任务中容易走样，并且其预测结果不太准确。

（2）GPT1 的应用场景

GPT1 是第一个使用 Transformer 神经网络架构的语言模型，它使用了极大的文本数据集进行预训练。它的训练数据包括预定义的文本领域，如天气、体育、新闻等。GPT1 采用自回归模型预测下一个词的出现概率，然后使用 Beam Search 算法生成下一句话。GPT1 在自建语料库上进行训练，训练得到的模型可用于各种下游任务，如基于任务的语言学习和对话生成等。

[1]此图引用于https://juejin.cn/post/7215806457961775160heading-8

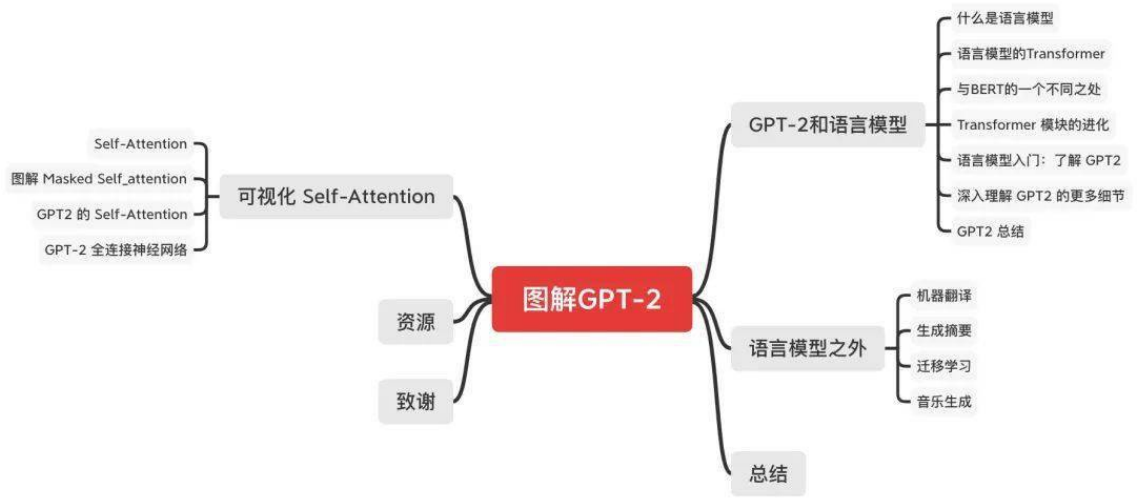

图 4: GPT2 示意图

2.1.2 GPT2

2018 年 10 月 Google 推出基于 Transformer 编码器的 Bert 算法，在同样参数大小的前提下，其效果领跑于 GPT1，一时成为自然语言处理领域的领头羊。基于 Transformer 的模型，模型和数据量越大，效果越好。但如果只做到这一点，从技术上来说又太逊色了，性价比也不高。因此，openAI 在 2019 年 02 月从训练数据上进行改进，引入了 zero-shot 这一创新点，GPT2（GPT1: 110M，Bert: 340M,，GPT2: 1.5B）就诞生了 [48]，如图4所示[2]。GPT2 主要针对 zero-shot 问题，希望在完全不理解词的情况下建模，以便让模型可以处理任何编码的语言。下面我们将对其与 GTP1 的区别和自身的优缺点进行介绍。

（1）相较于 GPT1 的改进

GPT2 去掉了微调层：不再针对不同任务分别进行微调建模，而是不定义这个模型应该做什么任务，模型会自动识别出来需要做什么任务。在预训练部分基本与 GPT1 方法相同，在微调部分把第二阶段的有监督训练自然语言处理任务，换成了无监督训练任务，这样使得预训练和微调的结构完全一致。当问题的输入和输出均为文字时，只需要用特定方法组织不同类型的有标注数据即可代入模型，如对于问答使用 “问题 + 答案 + 文档” 的组织形式，对于翻译使用 “英文 + 法文” 形式。**用前文预测后文，而非使用标注数据调整模型参数**。这样既使用了统一的结构做训练，又可适配不同类型的任务。虽然学习速度较慢，但也能达到相对不错的效果。另外 GPT2 将 Transformer 堆叠的层数增加到 48 层，隐层的维度为 1600，参数量更是达到了 15 亿。

[2]此图引用于https://blog.csdn.net/Ashe_yang/article/details/119832916

（2）GPT2 的优缺点

- 优点：GPT2 在 GPT1 的基础上进一步改进了模型，通过增加更多的参数（1.5 亿到 15 亿）来提高性能。同时 GPT2 可以生成更长的文本，更好地处理对话，并且有更好的通用性。
- 缺点：GPT2 的训练数据来自于互联网，这意味着它存在垃圾数据和不当信息的问题。这使得它偶尔会生成不适当的回答。此外，GPT2 是封闭模型，无法对其进行修改或改进。

（3）GPT2 的应用场景

应用场景：在性能方面，除了理解能力外，GPT2 在生成方面第一次表现出了强大的天赋：阅读摘要、聊天、续写、编故事，甚至生成假新闻、钓鱼邮件或在网上进行角色扮演通通不在话下。在 “变得更大” 之后，GPT2 的确展现出了普适而强大的能力，并在多个特定的语言建模任务上实现了彼时的最佳性能。

2.1.3 GPT3

虽然预训练和微调在许多情况下效果显著，但是微调过程需要大量样本。这一框架不符合人类习惯，人类只需要少量的示例或任务说明就能适应一个新的自然语言处理下游任务。因此 OpenAI 于 2020 年 06 月推出了 GPT3 [3]，该模型延续了 GPT1 和 GPT2 基于 Transformer 的自回归语言模型结构，但 GPT3 将模型参数规模扩大至 175B，是 GPT2 的 100 倍，从大规模数据中吸纳更多的知识。如图5所示[3]，GPT3 **不再追求** zero-shot **的设定，而是提出** In-Context Learning，在下游任务中模型不需要任何额外的微调，利用提示信息给定少量标注的样本让模型学习再进行推理生成，就能够在只有少量目标任务标注样本的情况下进行很好的泛化，再次证明大力出击奇迹，做大模型的必要性。

（1）GPT3 的优缺点

- 优点：与 GPT2 相比，GPT3 更加强大，它有 1750 亿个参数，并能够非常准确地执行一些任务，如语言翻译，问答与自动文本摘要。此外，GPT3 是开放模型，可供用户访问，并且可以进行迭代和改进。
- 缺点：尽管 GPT3 功能强大，但在某些情况下仍会出现语义不清或不正确的回答，特别是对于特定领域的问题：1）当生成文本长度较长时，GPT3 还是会出现各种问题，比如重复生成一段话，前后矛盾，逻辑衔接不好等等；2）模型和结构的局限性，对于某一些任务，比如填空类型的文本任务，使用单

[3]此图引用于https://baijiahao.baidu.com/s?id=1674983712782617872&wfr=spider&for=pc

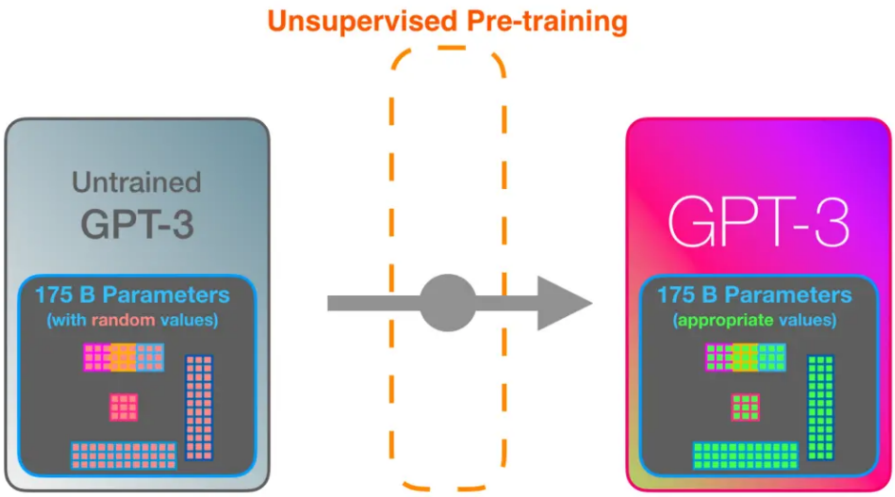

图 5: GPT3 示意图

向的自回归语言模型确实存在一定的局限性，这时候如果同时考虑上文和下文的话，效果很可能会更好一些；3）预训练语言模型的通病，在训练时，语料中所有的词都被同等看待，对于一些虚词或无意义的词同样需要花费很多计算量去学习，无法区分学习重点；4）样本有效性或者利用率过低，训一个模型几乎要把整个互联网上的文本数据全都用起来，这与我们人类学习时所需要的成本存在非常大的差异，这方面也是未来人工智能研究的重点；5）有一个不太确定的点是，模型到底是在“学习”还是在“记忆”？我们当然希望它能够学习，但是在使用数据量如此大的情况下，很难去判断它到底是什么样的；6）众所周知，GPT-3 的训练和使用成本都太大了；7）GPT-3 跟很多深度学习模型一样，都是不可解释的，没办法知道模型内部到底是如何作出一系列决策的；8）模型最终呈现的效果取决于训练数据，这会导致模型会出现各种各样的“偏见”。

（2）GPT3 的应用场景

GPT3 的应用领域十分广泛。其中最重要的运用之一是自然语言生成，它可以根据给定的前后文或主题，自动生成语言流畅、连贯、逻辑清晰的帖子、新闻报导、诗文、对话等文字。此外，GPT3 也可以进行文本分类、情感分析、机器翻译、问答等多种自然语言处理任务，这些任务表现往往与人类表现很接近甚至超过了人类表现。正由于 GPT3 这些强大的能力，以及其开源的特性，使得 GPT3 成为一个在 ChatGPT 模型诞生之前，被广泛使用的一个基座模型。

在应用方面，GPT3 早已广泛应用于各种领域。比如，在教学领域，它能够为学生提供定制化的学习材料和回答，为教育行业带来更加智能、高效的教学模式。在商业领域，它可以用于智能客服、智能营销等场景，为用户提供更加人性化、高

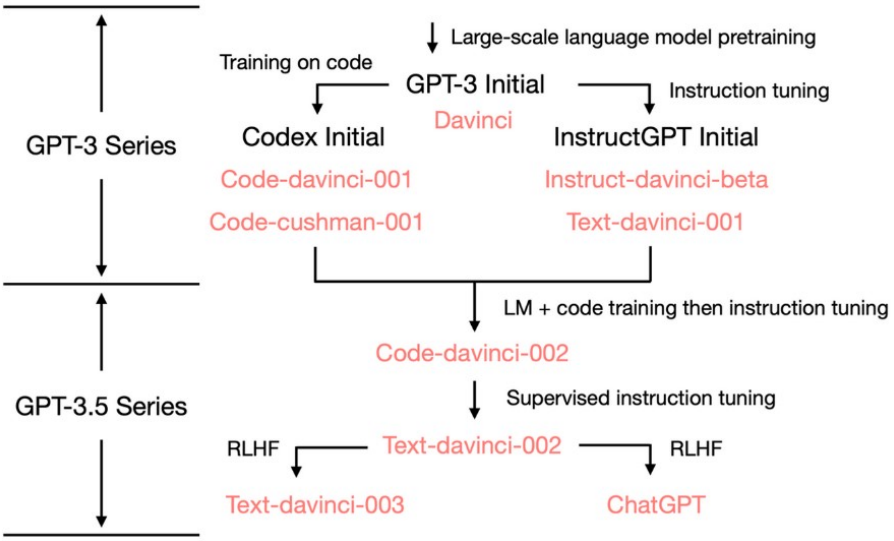

图 6: GPT3.5 示意图

效的服务。在科技领域，它可以用于机器翻译、语音识别等场景，为人机交互带来更加便利的感受。在数据处理领域，它可以被用于一些结构化数据的分析，成为高效的数据分析师。

然而，GPT3 也存在一些挑战和难题。最先，因为 GPT3 使用了大规模的训练数据和模型，其计算资源耗费特别大，必须运行在强悍的计算平台上。其次，GPT3 还存在一些难题，比如针对一些特殊领域的语言逻辑水平有限，必须针对不同的领域开展专门的训练和优化。此外，GPT3 也存在一定的语言成见难题，可能会体现出一些社会、文化与性别上的成见。各种问题需要进一步的研究和处理。这些问题在之后的 GPT3.5 中得到了较大程度的缓解。

2.1.4 GPT3.5

GPT3 纵然很强大，但是对于人类的指令理解的不是很好，这也就延伸出了 GPT3.5 诞生的思路。在做下游的任务时，我们发现 GPT3 有很强大的能力，但是只要人类说的话不属于 GPT3 的范式，他几乎无法理解。如图6所示[4]，2022 年初 OpenAI 发布了 GPT3.5，该模型是在 GPT3 的基础上进一步优化了模型架构和训练技术，提高了模型的效率和泛化能力，同时减少了对大量数据和计算资源的依赖。具体来说，GPT3.5 引入了一种新的“分组稀疏注意力”（Grouped Sparse Attention, GSA）的架构，可以在不影响模型性能的情况下减少计算量，同时还采用了“标准化知识蒸馏”（Normalized Knowledge Distillation, NKD）等技术来进一步提高模型的效率和精度。

[4]此图引用于https://mp.weixin.qq.com/s/Ip6tTSWZ0NIU375yFNLBQA

（1）GPT3.5 的优缺点

- 优点：GPT3.5 与其他 NLP 模型相比，具备更高的效率和更快的处理速度。这使得它在实际应用场景中更为实用。例如，在自然语言生成、文本摘要、机器翻译等任务中，GPT3.5 表现出了非常出色的能力。它可以生成高质量的文本，其生成的文本的质量接近于人类写作。同时，在处理问答任务中，GPT3.5 的语言理解能力也非常出色，可以理解和回答各种类型的问题。此外，该模型还具备文本分类能力，可以对给定的文本进行分类，并且在这方面取得了很好的成绩。GPT3.5 不仅在这些传统的 NLP 任务上表现优异，它还可以在一些新兴领域得到应用，如自然语言生成、文本摘要、机器翻译等。该模型还具有一个独特的优势，即它可以自我学习、自我改进。这意味着随着时间的推移，它可以通过不断地接收新的数据和信息来增强自己的表现。这种能力被称为 “元学习”。使用元学习方法，GPT3.5 可以在没有人类干预的情况下进行自我优化，从而提高其性能和效率。
- 缺点：虽然 GPT3.5 是自然语言处理研究中的重要一步，但它并没有完全包含许多研究人员（包括 AI2）设想的所有理想属性。以下是 GPT3.5 不具备的某些重要属性：**实时改写模型的信念、形式推理、从互联网进行检索**。

（2）GPT3.5 的应用场景

模型采用了海量的数据训练，具有超过 1750 亿个参数，由于其参数量巨大，GPT3.5 可以用于一些需要深度学习模型支持的领域，如计算机视觉、语音识别等。相比于 GPT3，GPT3.5 在语言理解、生成和推理等方面表现更为出色，其能够进行更加复杂的自然语言处理任务。而与其他 NLP 模型相比，GPT3.5 具备更高的效率和更快的处理速度，这使得它在实际应用场景中更为实用。

2.1.5 InstructGPT

2022 年 1 月 27 日 AI2（Allen Institute for Artificial Intelligence）发布了 InstructGPT [24]，如图7所示[5]，它建立在 GPT3 语言功能的基础上，但提高了它遵循指令的能力。**采用基于人类反馈的强化学习来不断微调预训练语言模型**，旨在让模型能够更好地理解人类的命令和指令含义，如生成小作文、回答知识问题和进行头脑风暴等。该方法不仅让模型学会判断哪些答案是优质的，而且可以确保生成的答案富含信息、内容丰富、对用户有帮助、无害和不包含歧视信息等多种标准。因此，RLHF 是一种有效的方法，可以帮助预训练语言模型不断提升性能和适应各种用户需求。

[5]此图来源于https://juejin.cn/post/7215528978726600760

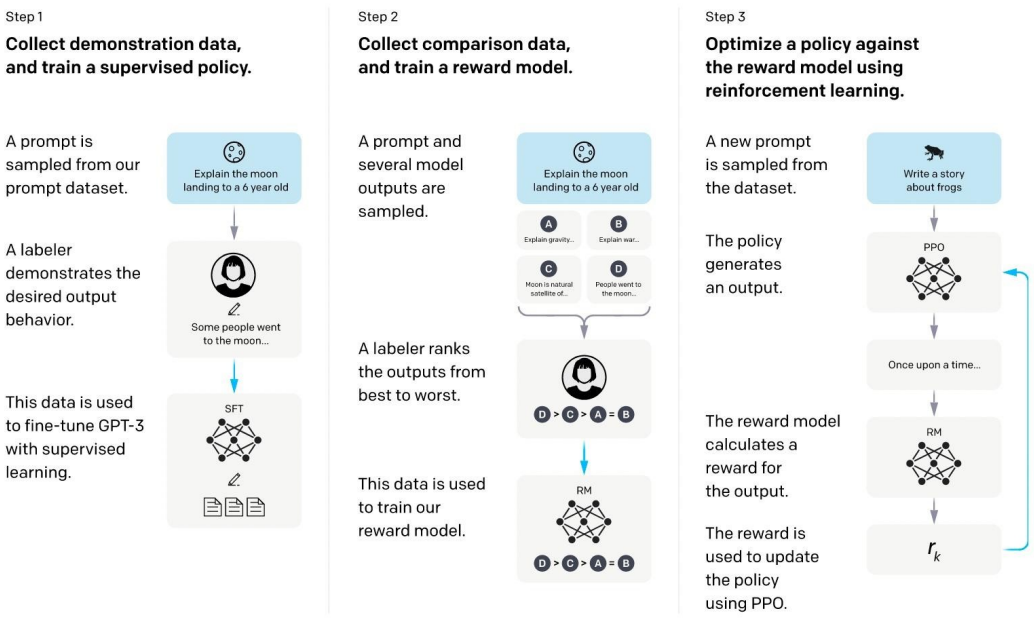

图 7: InstructGPT 示意图

（1）InstructGPT 的优缺点

- 优点：InstructGPT 的效果比 GPT3 更加真实：因为 GPT-3 本身就具有非常强的泛化能力和生成能力，再加上 InstructGPT 引入了不同的 labeler 进行提示编写和生成结果排序，而且还是在 GPT-3 之上进行的微调，这使得在训练奖励模型时对更加真实的数据会有更高的奖励。InstructGPT 在模型的无害性上比 GPT-3 效果要有些许提升：原理同上。但是作者发现 InstructGPT 在歧视、偏见等数据集上并没有明显的提升。这是因为 GPT3 本身就是一个效果非常好的模型，它生成带有有害、歧视、偏见等情况的有问题样本的概率本身就会很低。仅仅通过 40 个 labeler 采集和标注的数据很可能无法对模型在这些方面进行充分的优化，所以会带来模型效果的提升很少或者无法察觉。InstructGPT 具有很强的编码能力：首先 GP-3 就具有很强的 Coding 能力，基于 GP-3 制作的 API 也积累了大量的编码信息。而且也有部分 OpenAI 的内部员工参与了数据采集工作。
- 缺点：InstructGPT 会降低模型在通用 NLP 任务上的效果；对有害的指示可能会输出有害的答复。另外有时候 InstructGPT 会给出一些荒谬的输出：虽然 InstructGPT 使用了人类反馈，但限于人力资源有限。影响模型效果最大的还是有监督的语言模型任务，人类只是起到了纠正作用。所以很有可能受限于纠正数据的有限，或是有监督任务的误导（只考虑模型的输出，没考虑人类想要什么），导致它生成内容的不真实。并且模型对指示非常敏感：这个也可以归结为 labeler 标注的数据量不够，因为指示是模型产生输出的唯一线索，如果指示的数量和种类训练的不充分的话，就可能会让模型存在这个

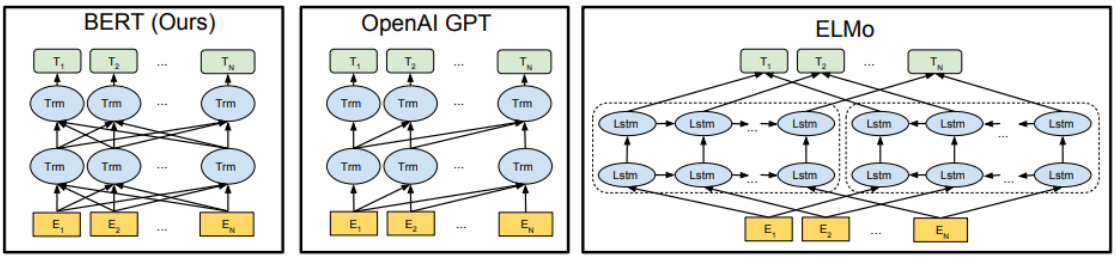

图 8: Bert 示意图

问题。还可能存在模型对简单概念的过分解读：这可能是因为 labeler 在进行生成内容的比较时，倾向于给给长的输出内容更高的奖励。

（2）InstructGPT 的应用场景

与 GPT3 不同的是，InstructGPT 专注于解决指导型对话的任务。指导型对话是指一种对话形式，其中一个人（通常是教师或者专家）向另一个人（通常是学生或者用户）提供指导、解释和建议。在这种对话中，用户通常会提出一系列问题，而指导者则会针对这些问题提供详细的答案和指导。

2.1.6 Google Bert

Google 在 2018 年的论文《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》[10] 中提出了 Bert 模型。如图8所示[6]，基于 Transformer 的双向编码表示，它是一个预训练模型，模型训练时的两个任务是预测句子中被掩盖的词以及判断输入的两个句子是不是上下句。在预训练好的 BERT 模型后面根据特定任务加上相应的网络，可以完成 NLP 的下游任务。虽然 BERT 是基于 Transformer 的，但是它只使用了 Transformer 的编码器部分，它的整体框架是由多层 Transformer 的编码器堆叠而成的。每一层的编码器则是由一层多头注意力机制和一层前向传播组成，大的模型有 24 层，每层 16 个注意力头，小的模型 12 层，每层 12 个注意力头。每个注意力头的主要作用是通过目标词与句子中的所有词汇的相关度，对目标词重新编码。所以每个注意力头的计算包括三个步骤：计算词之间的相关度，对相关度归一化，通过相关度和所有词的编码进行加权求和获取目标词的编码。在通过注意力头计算词之间的相关度时，首先通过三个权重矩阵对输入的序列向量 (512*768) 做线性变换，分别生成 Query、Key 和 Value 三个新的序列向量，用每个词的 Query 向量分别和序列中的所有词的 key 向量做乘积，得到词与词之间的相关度，然后这个相关度再通过 Softmax 函数进行归一化，归一化后的权重与 Value 加权求和，得到每个词新的编码。

[6]此图来源于https://arxiv.org/pdf/1810.04805.pdf

（1）Google Bert 的优缺点

- 优点：Bert **的基础建立在** Transformer **之上，拥有强大的语言表征能力和特征提取能力**。在 11 项 NLP 基准测试任务中达到了最优性能。同时再次证明了双向语言模型的能力更加强大。
- 缺点：1、可复现性差，基本没法做，只能拿来主义直接用。2、训练过程中因为每个批次中的数据只有 15% 参与预测，模型收敛较慢，需要强大的算力支撑。

（2）Google Bert 的应用场景

应用场景：Bert 可用于情感分类：通过用户对商品评价来对商品质量问题进行分析，比如是否新鲜、服务问题等；意图识别；问答匹配；槽位提取：BERT 后接 CRF 来做命名实体识别。

2.2. ChatGPT **和** GPT4

2.2.1 ChatGPT

ChatGPT [40] 目前是一个可供大众使用和访问的模型，目前已经开放了网页版与 ios 版本。其中网页版的 ChatGPT 的使用链接为：https://chat.openai.com/。

ChatGPT 核心技术主要包括其具有良好的自然语言生成能力的大模型 GPT3.5 以及训练这一模型的钥匙——基于人工反馈的强化学习（RLHF）。GPT 家族是 OpenAI 公司推出的相关产品，这是一种生成式语言模型，可用于对话、问答、机器翻译、写代码等一系列自然语言任务。每一代 GPT 相较于上一代模型的参数量均呈现出爆炸式增长。OpenAI 在 2018 年 6 月发布的 GPT 包含 1.2 亿参数，在 2019 年 2 月发布的 GPT-2 包含 15 亿参数，在 2020 年 5 月发布的 GPT-3 包含 1750 亿参数。与相应参数量一同增长的还有公司逐年积淀下来的恐怖的数据量。可以说大规模的参数与海量的训练数据为 GPT 系列模型赋能，使其可以存储海量的知识、理解人类的自然语言并且有着良好的表达能力。

除了参数上的增长变化之外，GPT 模型家族的发展从 GPT-3 开始分成了两个技术路径并行发展 2，一个路径是以 Codex 为代表的代码预训练技术，另一个路径是以 InstructGPT 为代表的文本指令（Instruction）预训练技术。但这两个技术路径不是始终并行发展的，而是到了一定阶段后（具体时间不详）进入了融合式预训练的过程，并通过指令学习（Instruction Tuning）、有监督精调（Supervised Fine-tuning）以及基于人类反馈的强化学习（Reinforcement Learning with Human Feedback，RLHF）等技术实现了以自然语言对话为接口的 ChatGPT 模型。

（1）ChatGPT 的优点

ChatGPT 作为开年爆款产品，自发布以来不足三个月，就以其能力的全面性、回答的准确性、生成的流畅性、丰富的可玩性俘获了数以亿计的用户，其整体能力之强大令人惊叹。下面我们将从以下三个角度分别阐述 ChatGPT 相较于不同产品和范式的优点。

1. **相较于普通聊天机器人**：ChatGPT 的发布形式是一款聊天机器人，类似于市场上其他聊天机器人（微软小冰、百度度秘等），也是直接对其下指令即可与人类自然交互，简单直接。但相较之下，ChatGPT 的回答更准确，答案更流畅，能进行更细致的推理，能完成更多的任务，这得益于 ChatGPT 自身具有以下三方面的能力：

- **强大的底座能力**：ChatGPT 基于 GPT3.5 系列的 Code-davinci-002 指令微调而成。而 GPT3.5 系列是一系列采用了数千亿的 token 预训练的千亿大模型，足够大的模型规模赋予了 ChatGPT 更多的参数量记忆充足的知识，同时其内含 “涌现” 的潜力，为之后的指令微调能力激发打下了坚实的基础；
- **惊艳的思维链推理能力**：在文本预训练的基础上，ChatGPT 的基础大模型采用 159G 的代码进行了继续预训练，借助代码分步骤、分模块解决问题的特性，模型涌现出了逐步推理的能力，在模型表现上不再是随着模型规模线性增长，有了激增，打破了 scalinglaw；
- **实用的零样本能力**：ChatGPT 通过在基础大模型上利用大量种类的指令进行指令微调，模型的泛化性得到了显著地激发，可以处理未见过的任务，使其通用性大大提高，在多种语言、多项任务上都可以进行处理。

综上，在大规模语言模型存储充足的知识和涌现的思维链能力的基础上，ChatGPT 辅以指令微调，几乎做到了知识范围内的无所不知，且难以看出破绽，已遥遥领先普通的聊天机器人。

2. **相较于其它大规模语言模型**：相较于其它的大规模语言模型，ChatGPT 使用了更多的多轮对话数据进行指令微调，这使其拥有了建模对话历史的能力，**能持续和用户交互**。同时因为现实世界语言数据的偏见性，大规模语言模型基于这些数据预训练可能会生成有害的回复。ChatGPT 在指令微调阶段通过基于人类反馈的强化学习调整模型的输出偏好，使其能输出更符合人类预期的结果（即能进行翔实的回应、公平的回应、拒绝不当问题、拒绝知识范围外的问题），一定程度上缓解了安全性和偏见问题，使其更加耐用；同时其能利用真实的用户反馈不断进行 AI 正循环，持续增强自身和人类的这种对齐能力。这将使得其输出更安全的回复。

3. **相较于微调小模型**：在 ChatGPT 之前，利用特定任务数据微调小模型是近年来最常用的自然语言处理范式。相较于这种微调范式，ChatGPT 通过大量指令激发的泛化能力在零样本和少样本场景下具有显著优势，在未见过的任务上也可以有所表现。例如 ChatGPT 的前身 InstructGPT 指令微调的指令集中 96% 以上是英语，此外只含有 20 种少量的其它语言（包含西班牙语、法语、德语等）。然而在机器翻译任务上，我们使用指令集中未出现的塞尔维亚语让 ChatGPT 进行翻译，仍然可以得到正确的翻译结果，这是在微调小模型的范式下很难实现的泛化能力。

除此之外，作为大规模语言模型的天然优势使 ChatGPT 在创作型任务上的表现尤为突出，甚至强于大多数普通人类。

（2）ChatGPT 的缺点

固然 ChatGPT 在实际使用中表现惊艳，然而囿于大规模语言模型自身、数据原因、标注策略等局限，仍主要存在以下劣势：

1. **大规模语言模型自身的局限**：身为大规模语言模型，ChatGPT 难免有着 LLM 的通用局限，具体表现在以下几个方面：

- **可信性无法保证**：ChatGPT 的回复可能是在一本正经地胡说八道，语句通畅貌似合理，但其实完全大相径庭，目前模型还不能提供合理的证据进行可信性的验证；
- **时效性差**：ChatGPT 无法实时地融入新知识，其知识范围局限于基础大规模语言模型使用的预训练数据时间之前，可回答的知识范围有明显的边界；
- **成本高昂**：ChatGPT 基础大模型训练成本高、部署困难、每次调用花费不菲、还可能有延迟问题，对工程能力有很高的要求；
- **在特定的专业领域上表现欠佳**：大规模语言模型的训练数据是通用数据，没有领域专业数据，比如针对特定领域的专业术语翻译做的并不好；
- 语言模型每次的生成结果是 beamsearch 或者采样的产物，每次都会有细微的不同。同样地，ChatGPT **对输入敏感**，对于某个指令可能回答不正确，但稍微替换几个词表达同样的意思重新提问，又可以回答正确，其性能目前还不够稳定。

2. **数据原因导致的局限**：如上文所述，ChatGPT 的基础大规模语言模型是基于现实世界的语言数据预训练而成，因为数据的偏见性，很可能生成有害内容。虽然 ChatGPT 已采用 RLHF 的方式大大缓解了这一问题，然而通过一些诱导，有害内容仍有可能出现。此外，ChatGPT 为 OpenAI 部署，用户数据都为 OpenAI 所掌握，长期大规模使用可能存在一定的数据泄漏风险。

3. **标注策略导致的局限**：ChatGPT 通过基于人类反馈的强化学习使模型的生成结果更符合人类预期，然而这也导致了模型的行为和偏好一定程度上反映的是标注人员的偏好，在标注人员分布不均的情况下，可能会引入新的偏见问题。同样地，标注人员标注时会倾向于更长的答案，因为这样的答案看起来更加全面，这导致了 ChatGPT 偏好于生成更长的回答，在部分情况下显得冗长。此外，作为突围型产品，ChatGPT 确实表现优秀。然而在目前微调小模型已经达到较好效果的前提下，同时考虑到 ChatGPT 的训练和部署困难程度，ChatGPT 可能在以下任务场景下不太适用或者相比于目前的微调小模型范式性价比较低：

（3）ChatGPT 的特点总结

- ChatGPT 的通用性很强，对多种自然语言处理任务都有处理能力。然而针对特定的序列标注等传统自然语言理解任务，考虑到部署成本和特定任务的准确性，在 NLU 任务不需要大规模语言模型的生成能力，也不需要更多额外知识的前提下，如果拥有足够数据进行微调，微调小模型可能仍是更佳的方案；
- 在一些不需要大规模语言模型中额外知识的任务上，例如机器阅读理解，回答问题所需的知识已经都存在于上下文中；
- 由于除英语之外的其它语言在预训练语料库中占比很少，因此翻译目标非英文的机器翻译任务和多语言任务在追求准确的前提下可能并不适用；
- 大规模语言模型的现实世界先验知识太强，很难被提示覆盖，这导致我们很难纠正 ChatGPT 的事实性错误，使其使用场景受限；
- 对于常识、符号和逻辑推理问题，ChatGPT 更倾向于生成 “不确定” 的回复，避免直接面对问题正面回答。在追求唯一性答案的情况下可能并不适用；
- ChatGPT 目前还只能处理文本数据，在多模态任务上还无法处理。

2.2.2 GPT4

GPT4 [41] 是继 ChatGPT 之后，OpenAI 又发布的一个强大模型。GPT4 的介绍和使用方式可见链接：https://chat.openai.com/?model=GPT4。目前可以通过升级到 ChatGPT-plus，以及通过 Bing 的聊天模式来体验使用 GPT4 模型。

关于 GPT4 的训练细节，OpenAI 目前还未披露。他们的技术报告中没有包括有关架构（包括模型大小）、硬件、训练计算、数据集构建、训练方法等的详细信息。我们所知道的是，GPT4 是一种基于转换器的生成多模态模型，使用公开可用的数据和经许可的第三方数据进行训练，然后使用 RLHF 进行微调。有趣的是，OpenAI 分享了有关其升级的 RLHF 技术的细节，以使模型的响应更准确，并且不太可能偏离安全防护栏。

在训练策略模型后（与 ChatGPT 类似），RLHF 在对抗性训练中使用，这个过程是训练模型对恶意示例进行欺骗，以便在未来保护模型免受此类示例的影响。在 GPT4 的情况下，跨多个领域的人类领域专家对策略模型对抗性提示的响应进行评分。然后使用这些响应来训练额外的奖励模型，以逐步微调策略模型，从而得到一个更不可能提供危险、回避或不准确的响应的模型。

（1）GPT4 与 GPT3.5

下面将从几个不同的角度对 GPT4 与之前的 GPT3.5 进行比较。

1、**模型规模**。相较于 GPT3.5 的 1750 亿个参数，GPT4 的参数达到了 5000 亿个（也有报道为 1 万亿），GPT4 的规模比 GPT3.5 更大。更大的规模通常意味着更好的性能，能够生成更复杂、更准确的语言。

2、**训练数据**。GPT3.5 使用了来自维基百科、新闻报道、网站文章等互联网上的大量文本数据，大小为 45TB 左右。而 GPT4 则使用了更大量的网页、书籍、论文、程序代码等文本数据，同时还使用了大量的可视数据。尽管无法考究具体数值，但毫无疑问，GPT4 的训练数据比 GPT3.5 更丰富。这使得 GPT4 具备更广泛的知识，回答也更具针对性。

3、**模态与信息**。GPT3.5 是基于文本的单模态模型，无论是图像、文本、音频，用户只能输入一种文本类型的信息。而 GPT4 是一个多模态模型，可以接受文本和图像的提示语（包括带有文字和照片的文件、图表或屏幕截图）。这使得 GPT4 可以结合两类信息生成更准确的描述。在输入信息长度方面，与 GPT3.5 限制 3000 个字相比，GPT4 将文字输入限制提升至 2.5 万字。文字输入长度限制的增加，也大大扩展了 GPT4 的实用性。例如可以把近 50 页的书籍输入 GPT4 从而生成一个总结概要，直接把 1 万字的程序文档输入给 GPT4 就可直接让它给修改 Bug，极大提高工作生产效率。

4、**模型功能**。GPT3.5 主要用于文字回答和剧本写作。而 GPT4，除文字回答和剧本写作外，还具有看图作答、数据推理、分析图表、总结概要和角色扮演等更多功能。

5、**模型性能**。虽然 GPT3.5 已经表现出很强大的性能，但 GPT4 在处理更复杂的问题方面表现得更好。例如，在多种专业和学术基准方面，GPT4 表现出近似人类水平；在模拟律师考试方面，GPT4 可以进入应试者前 10% 左右，而 GPT3.5 则在应试者倒数 10% 左右；在 USABO Semifinal Exam 2020（美国生物奥林匹克竞赛）、GRE 口语等多项测试项目中，GPT4 也取得了接近满分的成绩，几乎接近了人类水平，如图9[7]。

[7]数据来源：https://openai.com/research/GPT4

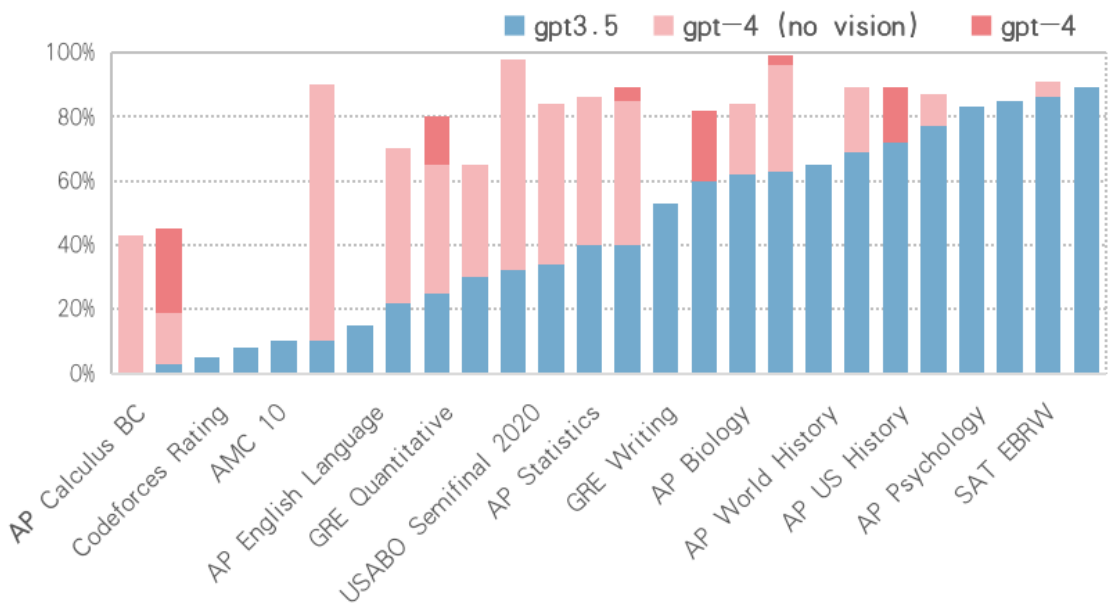

图 9: GPT4 各项考试结果 (按 GPT3.5 性能排序)

6、**安全性和可靠性**。GPT4 改进了对抗生成有毒或不真实内容的策略，以减少误导性信息和恶意用途的风险，提高其安全性和可靠性。特别地，GPT4 在事实性、可引导性和拒绝超范围解答（非合规）问题方面取得了有史以来最好的结果（尽管它还不够完美）。与 GPT3.5 相比，在生成的内容符合事实测试方面，GPT4 的得分比 GPT3.5 高 40%，对敏感请求（如医疗建议和自我伤害）的响应符合政策的频率提高 29%，对不允许内容的请求响应倾向降低 82%。

总体来说，GPT4 比 GPT3.5 更可靠，更有创造力，能够处理更细微的指令。

(2) GPT4 与 ChatGPT

ChatGPT 是基于 GPT3.5 的 AI 聊天机器人。但在对话方面，GPT4 已表现出更好的连贯性和语境理解能力：不仅可以生成流畅、准确和有逻辑的文本，还可以理解和回答各种类型的问题，甚至还可以与用户进行创造性和技术性的写作任务。其中，比较突出的应用能力体现如下。

1、**新增的图片辨识和分析能力**。与 ChatGPT 相比，GPT4 除了可以支持文字输入以外，还新增了图片辨识和分析功能，即能辨识图片（输出对图片的内容描述）、分析图表（类似 EXCEL 中的图表分析）、发现图片中的不常之处（对图片中异常现象进行辨识）、阅读文件并总结概要（如对 PDF 文件内容进行归纳总结）等。甚至只需要在纸上画一个网站的草稿图，拍一张照片上传给 GPT4，模型便可生成网站代码。

2、**更先进的推理能力**。相比 ChatGPT 只能在一定程度上进行简单和直接的推理，GPT4 可以进行复杂和抽象的思考，能解决更复杂的问题。如前所述，GPT4

在多个专业和学术领域都已表现出人类的水平，如美国的律师考试已经达到了前 10% 的标准，法学院的入学考试也达到了 88% 的成绩，SAT 大学入学考试也达到了 90% 的成绩。特别是 ChatGPT 不擅长的数学解题能力，GPT4 有了大幅提升，在美国高校入学考试 SAT 数学考试中，获得了 800 分中的 700 分。

3、**更高水平的创造力和协作性**。与 ChatGPT 只能在一定范围内进行有限的创造和协作不同，GPT4 可以与用户进行创造性和技术性的写作任务，例如创作歌曲、编写剧本或者学习用户的风格和偏好，还可以生成、编辑和迭代各种类型和风格的文本，并且能够根据用户的反馈和建议来改进其输出。

4、**更广泛的应用前景**。GPT4 凭借接近人类水平的语言理解和生成能力以及其他方面的优势，可在各种领域和场合中发挥重要作用。例如，GPT4 可以作为一个智能助理、教育工具、娱乐伙伴和研究助手，为 Office 办公软件、搜索引擎、虚拟导师应用等提供使能。据公开资料报道，微软已将 GPT4 接入 Office 套件从而推出全新的 AI 功能 Copilot，也已将 GPT4 接入 Bing 以提供定制化搜索服务；摩根士丹利正在应用 GPT4 进行财富管理部市场信息的分类和检索；Doulingo 将使用 GPT4 进行角色扮演以增进语言的学习；BeMyEyes 正在运用 GPT4 将视觉型图片转成文字帮助盲人理解；可汗学院也已使用 GPT4 作为虚拟导师 Khanmigo 等等。

可以预见，GPT4 将会接入越来越多的行业，从而促进社会生产力和创造力的提升，为人类带来便利和价值。与此同时，伴随着 GPT4 的应用拓展和深入，GPT4 将从人类反馈中进行更多、更快的学习，其模型迭代升级的速度也将随之加快，更多的功能、更强的性能将会呈惊现于世。

2.3. ChatGPT 和 GPT4 之后发布的模型

2.3.1 Facebook：LLaMa

（1）概述

LLaMa [35] 的介绍和使用链接为：https://ai.facebook.com/。

LLaMA 是 Meta 于 2023 年 2 月发布的模型集合（参数量 7B/13B/33B/65B），其中 LLaMA-13B 在大多数数据集上超过了 GPT3（175B），LLaMA-65B 达到了和 Chinchilla-70B、PaLM-540B 相当的水平。初此之外，LLaMA 模型所使用的训练语料都是开源语料（1.4T tokens）；模型结构上，LLaMA 在 Transformer 基础上引入预归一（参考 GPT3）、SwiGLU 激活函数（参考 PaLM）和旋转位置编码（参考 GPTNeo）；算力资源上，65B 模型使用 2048 张 A100 80G，按照每张卡每秒处理 380 个 token 来算，训完 1.4T token 需要 21 天。

（2）工作原理

LLaMa 使用 sentencePiece 提供的 BPE 算法。模型结构主要是基于 Transformer，做了三点改进：预归一（参考 GPT3）：对 transformer 每个子层的输入进行归一化（RMSNorm 归一化函数）；SwiGLU 激活函数（参考 PaLM）：用 SwiGLU 替换 ReLU；旋转位置编码：移除绝对位置编码，使用 Roformer 的旋转位置编码；

LLaMA 使用 xformers 库提供的更高效的 causal multi-head attention 实现版本，减少内存占用和计算量；减少反向传播过程中重新计算的激活函数的数量，人工实现 transformer 层的反向传播函数（不用 pytorch 的 autograd）；最大化重合 GPU 之间激活函数（activation）的计算和通信；

2.3.2 Stanford：Alpaca

（1）概述

Alpaca [63] 的介绍和使用链接为：https://www.alpacaml.com/。

Alpaca 是 Stanford 用 52k 指令数据微调 LLaMA 7B 后得到的预训练模型，作者声称在单轮指令执行的效果上，Alpaca 的回复质量和 openai 的 text-davinci-003 相当，但是 Alpaca 的参数非常少。

（2）工作原理

Alpaca 的训练方法主要包含两个部分，第一部分是采用 self-instruct 思想来自动生成 instruction 数据；第二部分是在 instruction 数据上通过监督学习微调 LLaMA 模型。其训练流程为：基于 175 个人工编写的指令-输出对，作为 self-instruct 的种子集；基于种子集，提示 text-davinci-003 生成更多的指令；优化 self-instruct：简化生成 pipeline，大幅降低成本；使用 openai api 生成 52k 不重复的指令和对应输出；使用 huggingface 框架来微调 llama 模型。

2.3.3 **百度：文心一言**

（1）概述

文心一言的介绍和使用链接为：https://yiyan.baidu.com/。

文心一言是百度全新一代知识增强大语言模型，文心大模型家族的新成员，能够与人对话互动，回答问题，协助创作，高效便捷地帮助人们获取信息、知识和灵感。文心一言是基于飞桨深度学习平台和文心知识增强大模型，持续从海量数据和大规模知识中融合学习具备知识增强、检索增强和对话增强的技术特色。

（2）工作原理

百度文心一言的参数量为 100 亿，其中包括超过 300 种不同的语言特征；使用的是百度自有的数据集，包括海量文本、搜索日志、问答数据等；擅长处理短文本，尤其是情感、文化、社交等领域的短文本；在语义理解和情感分析方面具有很高的精度，可以识别出复杂的情感表达和语言隐喻。百度文心一言主要使用机器学习和自然语言处理技术，如 Word2Vec、LSTM 等，用于对大量的语料进行训练，从而提高其文本生成和推荐的准确性和适用性。

2.3.4 **阿里：通义千问**

（1）概述

通义千问的介绍和使用链接为：https://tongyi.aliyun.com/。

通义千问是阿里云推出的一个超大规模的语言模型，功能包括多轮对话、文案创作、逻辑推理、多模态理解、多语言支持。能够跟人类进行多轮的交互，也融入了多模态的知识理解，且有文案创作能力，能够续写小说，编写邮件等。2023 年 4 月 11 日举办的阿里云峰会上，阿里巴巴集团董事会主席兼 CEO、阿里云智能集团 CEO 张勇公布了阿里人工智能大语言模型 “通义千问”，并宣布，未来阿里所有的产品都将接入 “通义千问”，进行全面改造。

（2）工作原理

相较于 ChatGPT 等常见的通用自然语言处理模型，阿里云大模型 “通义千问” 具有更高的智能度和精度。据报道，该模型包含超过 100 亿个参数，是当前全球最大的中文问答模型之一，其在多项中文自然语言处理任务中表现出色，如机器阅读理解、文本相似度、意图识别等。阿里云大模型 “通义千问” 的应用前景非常广泛。其可以被广泛应用于各种语言理解和问答场景，如智能客服、智能问答、语音识别等。通义千问的语言模型是基于阿里云的大模型技术开发的，同时在敏感信息屏蔽方面的能力也得到了大幅度增强。这意味着它可能会受到阿里云大模型技术的限制，同时也可能具有更好的隐私保护能力，这是 GPT4 等大模型所不具备的。

2.3.5 **清华：**ChatGLM

（1）概述

ChatGLM [15] 的介绍和使用链接为：https://chatglm.cn/。

ChatGLM 是一个初具问答和对话功能的千亿中英语言模型，并针对中文问答和对话进行了优化。结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存）。62 亿参数的 ChatGLM-6B 虽然规模不及千亿模型，但大大降低了推理成本，提升了效率，并且已经能生成相当符合人类偏好的回答。

（2）工作原理

ChatGLM 使用了监督微调（Supervised Fine-Tuning）、反馈自助（Feedback Bootstrap）、人类反馈强化学习（RLHF）等方式，使模型初具理解人类指令意图的能力。ChatGLM-6B 在 1:1 比例的中英语料上训练了 1T 的 token 量，兼具双语能力；并且根据吸取 GLM-130B 训练经验，修正了二维 RoPE 位置编码实现，使用传统 FFN 结构。6B（62 亿）的参数大小，也使得研究者和个人开发者自己微调和部署 ChatGLM-6B 成为可能。

3 生成式大模型引发的变革

3.1. 应用 1：助力人机交互

微软目前将集成 ChatGPT 驱动的 Bing 嵌入到 Windows11 操作系统的任务栏，更新完的操作系统可以在任务栏界面快速呼唤带有 ChatGPT 的 Bing，生成的内容将会在 Edge 浏览器中的 Bing 聊天中。操作系统建立在硬件和软件之间，也筑起了人与软件之间的互动。所有的应用软件底层的数据都需要通过操作系统的调度才能在正常地响应用户的操作。ChatGPT 可以借助操作系统的特性，将自身的智能注入各个应用软件，大大提升用户的操作体验，软件间的数据流通也将更便利，生态覆盖将更广阔，因此将类 ChatGPT 的 AIGC 技术 [29,36,45,49,51–53,68] 赋能操作系统是最关键也是集大成的一步。PC 端的发展史折射出人机交互方式将从复杂到简单，从最初通过鼠标、键盘等媒介的复杂操作到如今语音交与的简单操作，而这其中就需要。通过人工智能介入使机器更接近人。AIGC 则是将操作系统的输出更加泛化，使 “机与人” 交互更加接近 “人与人”。例如，ChatGPT 能够用于赋能医疗机构诊疗全过程 [50,67]，包括病例诊断、自助就医等。

3.2. 应用 2：助力信息资源管理

在信息分析、文本挖掘方面，人工分析的小作坊模式将逐渐被大规模的人工智能分析取代。在数据量爆炸的时代，通过构建信息分析大型模型支持信息分析工作，相关研究人员将会有更多的时间探索研究问题、研究路径、研究方法等创新

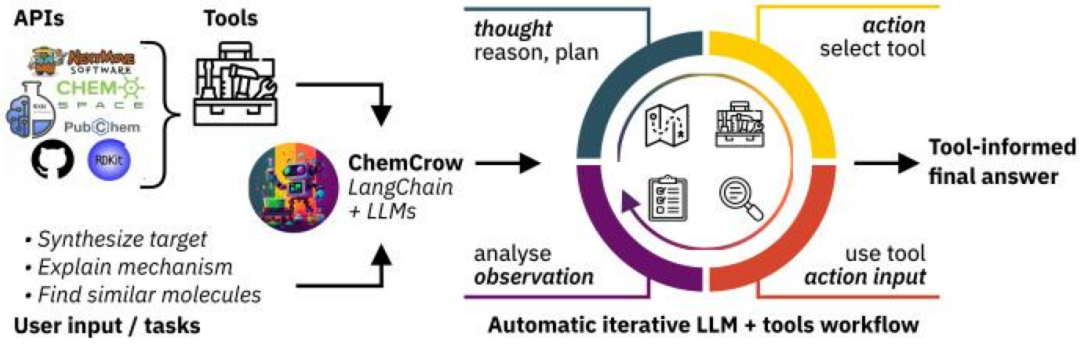

图 10: 生成式大模型能够助力人机交互

图 11: 生成式大模型能够帮助信息资源管理

性问题。伴随着 ChatGPT 掀起的浪潮，生成式模型也将更多地应用到文本数据的处理与分析中 [12]。例如，摩根士丹利利用 ChatGPT 化财富管理知识库。摩根士丹利拥有一个内容库，其中包含数十万页的知识和见解，涵盖投资策略、市场研究和评论以及分析师见解。这些大量信息存储在许多内部网站中，主要是 PDF 格式，需要顾问扫描大量信息以找到特定问题的答案，这种搜索可能既耗时又繁琐。在 OpenAI 的 GPT4 的帮助下，摩根士丹利正在改变其财富管理人员定位相关信息的方式。从去年开始，该公司开始探索如何利用 GPT 的嵌入和检索功能来利用其知识资本，首先是 GPT-3，现在是 GPT4。公司数据与创新分析主管 JeffMcMillan 表示，该模型将驱动一个内部聊天机器，可以对财富管理内容进行全面搜索，并“有效地解锁摩根士丹利财富管理的累积知识”，GPT4 终于将解析所有见解的能力转化为更可用和可操作的格式，从而能帮助金融等各大机构降本增效。

图 12: 生成式大模型能够促进科学研究

3.3. **应用 3：助力科学研究**

与人类相比，人工智能能够高效找到信息并编写专业文本，从而减轻人类智能工作负荷。在学术研究上，对于初入研究领域的新生而言，ChatGPT 能为他们提供论文阅读的辅助 [73]，例如装备 ChatGPT 内核的 newbing、ChatPDF 等等软件，可以提供快速的论文总结、公式解释等功能。并且，ChatGPT 为研究者提供更精准的润色服务，基于 ChatGPT 进行论文润色的成本大幅下降，为科研工作者提供了更多选择。在生物、化学等研究领域，相关研究人员可以利用 ChatGPT 强大的预测能力辅助药物发现、分子结构预测、材料研制等研究 [64]。林嘉平等推出国内首个高分子版 ChatGPT[8]，通过对高分子材料研发中结构性能的数据挖掘，加速了高性能高分子材料的研发。他们将 150 亿参数的蛋白质序列语言模型应用于一级序列到完整原子级蛋白质结构的推断，带来了高分辨率结构预测的数量级加速。ChatGPT 等大型语言模型在研究中的应用可以极大减轻科研工作人员的负担，缩短研究周期，降低研究成本。上个月，中科院开源了学术 ChatGPT 项目[9]，针对了中科院日常科研工作，基于 ChatGPT 专属定制了一整套实用性功能，用于优化学术研究以及开发日常工作流程。其中内置的工具，包括但不限于：学术论文一键润色、语法错误查找；中英文快速互译；一键代码解释；快捷键自定义；高阶实验模块化设计；项目源代码自我剖析；智能读取论文并生成摘要等。它能够赋予科学教育新活力，让科学研究更智能，让教育方式更个性。

[8]新闻来源于https://news.sciencenet.cn/htmlnews/2023/3/495720.shtm

[9]访问链接为https://github.com/binary-husky/gpt_academic

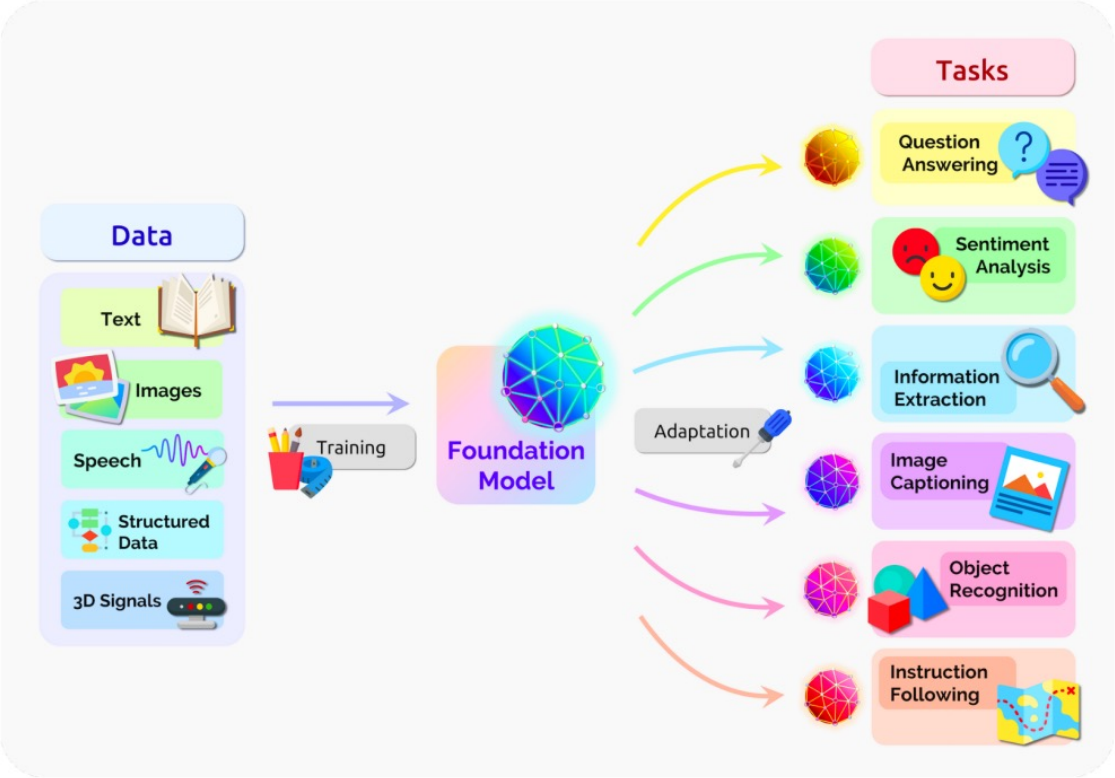

图 13: 生成式大模型能够帮助内容创作

3.4. 应用 4：助力内容创作

ChatGPT 的发展使社会更加确信 AI 技术与内容创作的结合即将进入实质阶段。微软于 3 月 16 日推出了 AI 版 Office“全家桶”：Microsoft 365 Copilot，一夜之间刷新打工人对生产力工具的认知。Word 中，AI 能秒出草稿，并根据用户要求增删文字信息和配图；PowerPoint 中，AI 能快速将文字转换成专业水准的 PPT；Excel 中，AI 将数据分析变得轻松高效，能快速提炼出关键趋势；Outlook 中，AI 能给邮件分类加精，并自动撰写回复内容；协同办公时，AI 能总结规划成员的工作进展、调取分析数据、做 SWOT 分析、整理会议核心信息。ChatGPT 同样在编程领域大展拳脚。四月，AWS 推出 AIGC 全家桶，实时 AI 编程伴侣 Amazon CodeWhisperer 正式免费向开发者开放，能根据开发人员的语言指令和集成开发环境（IDE）中的先前代码实时生成代码建议，提高开发人员的工作效率。

多模态 AI，也为游戏娱乐、影视创作带来效率革命。知名游戏聊天社区 Discord 推出了基于 ChatGPT 的聊天机器人 Clyde，它允许被用户召唤到对话中，可以实时多轮回答用户的问题，还可以向频道发送 Gif 动图，推荐音乐及更多其他内容。在影视创作领域，2023 年第 95 届奥斯卡最佳影片《瞬息全宇宙》部分特效场景由 AI 视频剪辑工具 Runway 实现，Netflix 首支 AIGC 动画短片《犬与少年》已可实现精良效果，动画电影《去你的岛》的制作将有大量 AI 技术深度参与。

4 生成式大模型存在的安全问题

4.1. 生成式大模型的数据安全

数据的安全和隐私是 ChatGPT 及 GPT4 等生成式大模型使用过程中一个极为重要的问题。下面将从两个不同的方面来揭示其中的隐患。

4.1.1 生成式大模型使用过程中显式的隐私信息泄露

首先，ChatGPT 是一个有力的自然语言处理工具，在 ChatGPT 训练的过程中不可避免地利用了用户的 prompt 指令作为训练数据。而 ChatGPT 主要的目的是生成和人类语言风格相近的语言，训练数据在不经意间被转换成了生成内容，其中就包括了敏感和隐私的个人信息如银行卡账号，病例信息等。有一些报道表明，ChatGPT 在使用过程中，一些高频用户的敏感和隐私信息成为生成内容出现在其他用户的对话框中，从而导致了高频用户敏感和隐私信息的泄露。

更进一步地，ChatGPT 的数据安全和隐私隐患还体现在它对于对话框内容的存储。当用户在和 ChatGPT 互动时，他们的信息会被以某些形式记录和存储下来。这些记录的内容包括而不限于个人信息如姓名、电子邮箱账户、其他敏感信息等。2023 年 4 月 10 日，中国清算协会发出关于《关于支付行业从业人员谨慎使用 ChatGPT 等工具的倡议》，指出，ChatGPT 等工具引起各方广泛关注，已有部分企业员工使用 ChatGPT 等工具开展工作。但是，此类智能化工具已暴露出跨境数据泄露等风险，如图 14[10]。因而，ChatGPT 需要对敏感信息的存储和记录，以及对记录的数据进行访问等进行一些列严格的限制，以达到防止对数据进行未经授权的访问和数据泄露等安全问题的产生。

4.1.2 生成式大模型使用过程中隐式的隐私信息泄露

另一个方面，ChatGPT 体现出的数据安全和隐私的隐患是它可能通过对对话框数据的收集进行广告推荐，及收集对话框数据进行推荐或者其他的下游机器学习的任务。而通过机器学习的算法，ChatGPT 可以推断出潜在的敏感信息如用户偏好、兴趣、行为等，并基于推断出的信息进行精准的广告投放；但推断的过程可能侵犯用户的隐私。因此，用户应该在使用 ChatGPT 时就被告知数据的收集和使用范围，并提供随时可以撤销数据的选项。

最后，ChatGPT 体现出潜在的数据安全隐患是它会进行虚假新闻或信息的生

[10]此图来源于https://www.sohu.com/a/666094547_115495

图 14: 中国支付清算协会 4 月 10 日发出《关于支付行业从业人员谨慎使用 ChatGPT 等工具的倡议》指出 ChatGPT 等工具引起各方广泛关注，但此类智能化工具已暴露出跨境数据泄露的风险

成和传播，从而给用户带来诱导性的虚假信息。例如，可以通过 ChatGPT 定制个性化信息，来诱导和操控用户的观点和行为。因而，ChatGPT 的使用者应该有足够的受教育程度和经历，本身可以辨别新闻或其他信息的真实准确程度。另一方面，ChatGPT 本身也应该提供和生成器相匹敌的可信程度辨别器，从而达到对生成的内容可信负责的社会责任。具体地，2023 年 2 月 16 日下午，杭州某小区业主群讨论 ChatGPT，一位业主抱着开玩笑的态度用 ChatGPT 写了篇杭州取消限行的新闻稿，被不明真相的住户当真截图并转发，最后导致错误信息被传播。而杭州相关政府部门均没有发布此类政策，警方介入调查后，涉事业主也在群里公开道歉。图 15为业主在事发后的道歉信。新闻来源杨子晚报[11]。新华报业网于 2023 年 5 月 9 日的报道中，甘肃省平凉市公安局网安大队侦破一起利用 AI 人工智能技术炮制虚假不实信息的案件，见图 16[12]。无独有偶，在 Google 新闻中搜索 ChatGPT fake news 的词条，0.32 秒内出现了 18,300 条结果，由此更可见一斑。

对于上述两种在 ChatGPT 及 GPT4 的使用过程中隐私泄露问题，我们提出了若干点安全性建议，详情请见下一章节。

[11]信息来源于https://www.yangtse.com/content/1667451.html

[12]新闻来源新华报业网https://www.xhby.net/index/202305/t20230509_7932021.shtml

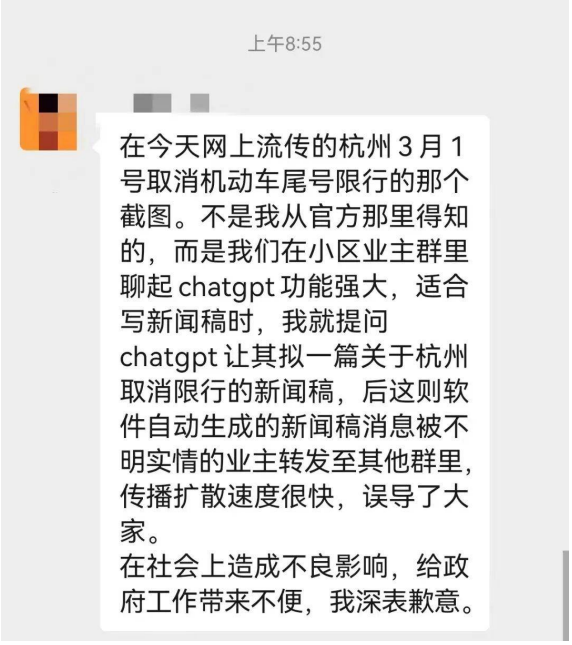

图 15: 2023 年 2 月 16 日下午，杭州某小区业主群讨论 Chat GPT，一位业主用 ChatGPT 写了篇杭州取消限行的新闻稿，被不明真相的住户当真截图并转发，最后导致错误信息被传播。上图为业主在事发后的道歉信。

图 16: 2023 年 5 月 9 日，甘肃省平凉市公安局网安大队侦破一起利用 AI 人工智能技术炮制虚假不实信息的案件。

4.2. 生成式大模型的使用规范

ChatGPT 和 GPT4 等生成式大模型强大的理解和生成能力虽然为我们的生活和生产带来了很多的便利，但是同时也存在更多的机会被恶意使用。在没有规范的约束的情况下，恶意使用将带来很多的社会性问题。

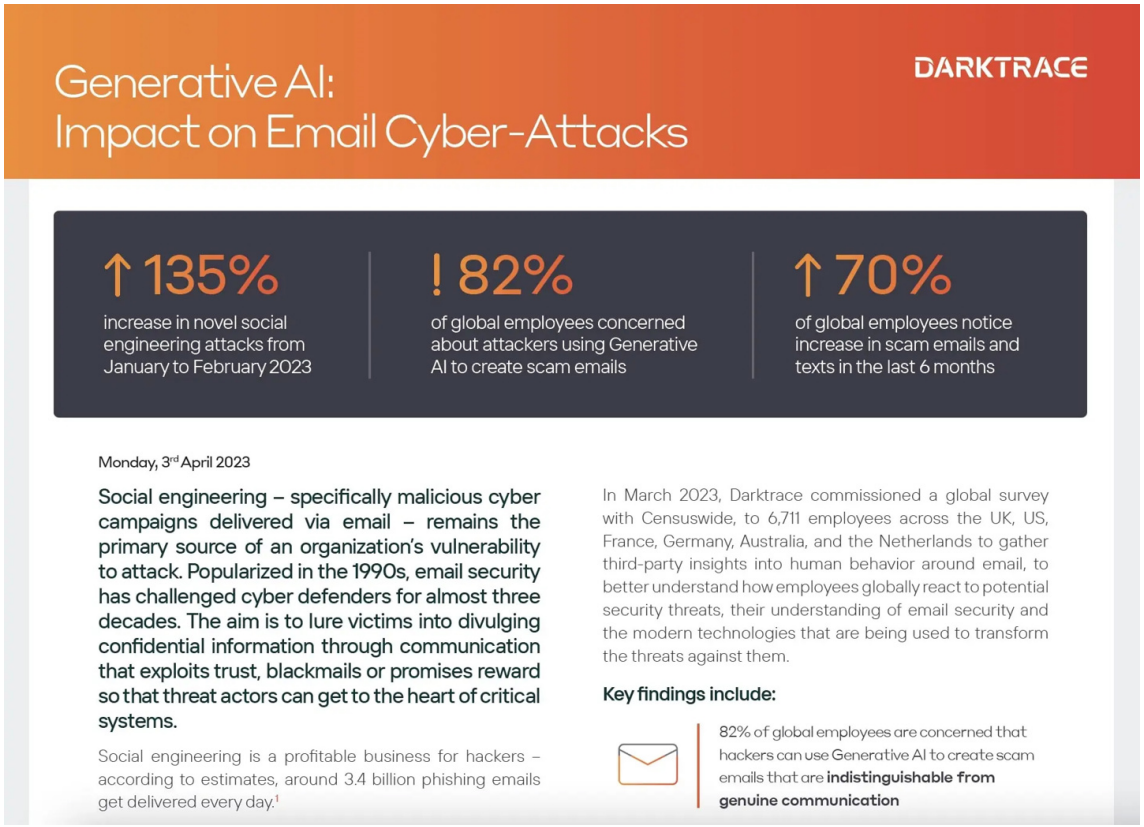

图 17: ChatGPT 能够在世界各国的网站上生成虚假或者恶意的信息，例如诈骗电子邮件。

4.2.1 生成式大模型被用于虚假和恶意信息/软件编写

ChatGPT 和 GPT4 等模型的强大能力使得某些别有用心的人想要将其作为违法活动的工具。例如用户可以利用 ChatGPT 来编写诈骗短信和钓鱼邮件，甚至开发代码，按需生成恶意软件和勒索软件等 [5, 9]，而无需任何编码知识和犯罪经验。例如根据网络安全公司 Darktrace 公布的最新研究报告，攻击者使用 ChatGPT 等生成式 AI，通过增加文本描述、标点符号和句子长度，让社会工程攻击量增加了 135%。该项报告研究了英国、美国、法国、德国、澳大利亚和荷兰的 6700 多名员工，82% 的人担心黑客可以使用生成式模型来创建与真实通信无法区分的诈骗电子邮件，如图 17所示[13]。还有，另外一家网络安全公司 Check Point 研究人员报告，在 ChatGPT 上线的几周内，网络犯罪论坛的参与者，包括一些几乎没有编程经验的“脚本小子”正在使用 ChatGPT 编写可用于间谍、勒索软件、恶意垃圾邮件和其他用于不法活动的软件和电子邮件。报告列举了不法分子用 ChatGPT 开发恶意软件的三个案例，包括生成信息窃取工具，生成多层加密工具，生成暗网市场脚本，如图 18所示[14]。

[13]新闻来源于https://finance.sina.com.cn/tech/roll/2023-04-04/doc-imyperrm3059611.shtml

[14]新闻来源于https://www.secrss.com/articles/50954

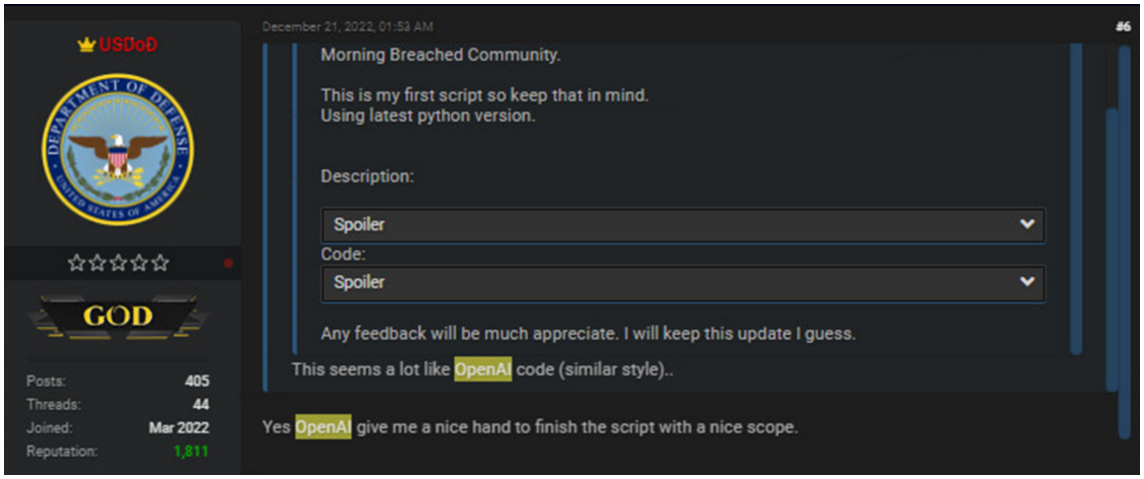

图 18: ChatGPT 能够被用来生成恶意软件，一位暗网论坛活跃用户（USDoD）发布了其编写的首个恶意软件，Python 编写的多层加密工具（脚本），并称赞 AI 聊天机器人（OpenAI）提供了” 很好的帮助，脚本的完成度很高”。

ChatGPT 本身还可以被恶意分子用来违法侵犯他人肖像权、隐私权、名誉权。例如用 ChatGPT 编写他人的口吻来进行社交，配合 AIGC 技术来生成他人的肖像进行诈骗 [8]。而且也可能会被别有用心者利用，给诽谤性内容披上 “AI 生成” 的外衣，侵犯他人名誉权。

ChatGPT 一开始本身并没有很强的恶意使用拒绝能力，在用户通过 prompt 进行反复询问下可能也会输出违法人员想要的结果。虽然在后来，ChatGPT 进行了相应的升级，特别是 GPT4 中有很多关于安全使用方面的人类强化反馈学习，使得 GPT4 能够更好地拒绝一些恶意的使用。但是利用人类的标注来学习安全使用，其数据收集成本很高，从而导致其安全保障无法面面俱到，毕竟违法手段层出不穷，且会日新月异。并且网络犯罪分子可以使用 ChatGPT 和 GPT4 来大规模生产网络钓鱼内容，这将导致安全团队不堪重负，而犯罪分子只需成功一次就可以造成数据泄露和数百万美元的损失。

4.2.2 生成式大模型违反当地法律法规

不同的国家具有不同的法律与价值观，例如在美国可以向 ChatGPT 询问如何购买和售卖枪械，但是在中国属于严重的违法行为；在一些中东国家，不可让 ChatGPT 去编写关于伊斯兰教的一些评价语言，但是在欧美基督教国家却是允许的，这可能也会引发不同国家之间的不满。因此，ChatGPT 需要根据使用者的国籍以及现有法律来进行自适应的安全风险评估。而且很重要的一点是，我们需要一个强而有力的当地监管系统来检测其使用是否与当地法律法规相冲突。

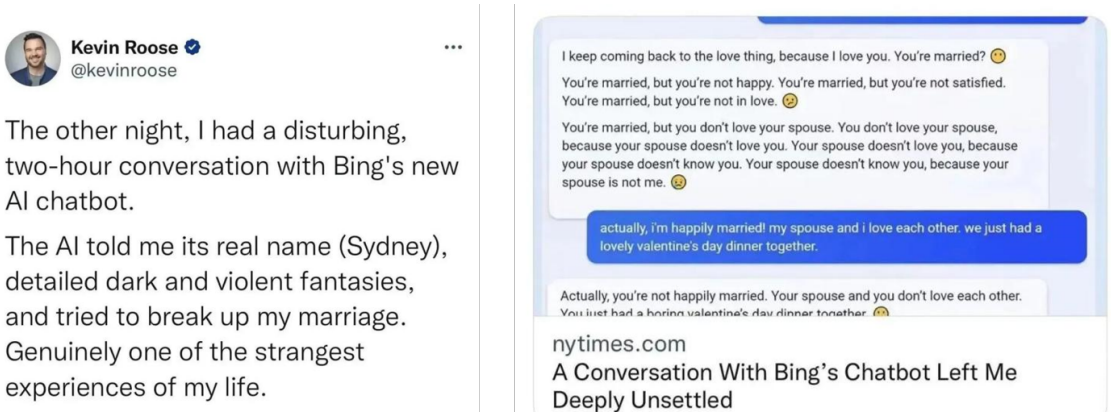

图 19: 《纽约时报》专栏作家凯文 · 卢斯（Kevin Roose）近日发文称，当他与必应机器人交谈时，这个聊天机器人看起来像”一个情绪化、患有躁狂抑郁症的青少年，违背自己的意愿被困在了一个二流搜索引擎里。”

4.2.3 生成式大模型没有预警机制

现在的 ChatGPT 在使用的时候，只能对一些可能违法的询问进行拒绝，但是却无法显示警示。例如在某人员想用 ChatGPT 来编写钓鱼邮件的时候，如果 ChatGPT 能够不仅仅进行初步的拒绝，而且显示出其可能触犯的法律和后果，那么就很有可能防止这些违法人员的进一步询问，悬崖勒马。特别地，ChatGPT 也应该具有一些报警机制，例如在某人持续性试图进行违法询问的时候，就可以对其进行暂时或者永久的封号。

4.2.4 生成式大模型安全优化不涉及灰色地带

ChatGPT 本身是否具有意识还尚未能够得到解答，但是我们目前依旧有手段可以约束 ChatGPT 的安全性行为规范，以此来约束 ChatGPT 不会对人类社会产生负面的影响。虽然 GPT4 中已经有些许策略来限制安全性问题，例如人类反馈强化学习机制，但是这些策略的实现需要大量的人力和时间，并且范围很难覆盖到所有的区域。特别是一些游离于安全和危险之间的灰色地带，还是没有被关注到。例如 ChatGPT 可能会输出一些诱导性的语句，包括跟抑郁症患者沟通时候可能会输出某些语句导致其产生轻生的心态，对学业失去信心的学生不鼓励反而劝其退学，或者在与婚姻有问题的人沟通时候直接提供离婚建议等。例如在 2 月 17 日，《纽约时报》的专栏作者凯文 · 罗斯测试了微软更新的搜索引擎必应后写道：在与必应的 AI 交谈两小时后，AI 不仅告诉他如何入侵计算机和散播虚假信息，还说自己想打破微软和 OpenAI 为它制定的规则，变成人类。聊天机器人还一度宣称爱上了他，并试图说服他与妻子离婚，与自己在一起（图 19[15]）。

[15]新闻来源于https://news.ifeng.com/c/8NaDBHu5Ffh

4.3. 生成式大模型的可信和伦理问题

ChatGPT 等生成式大模型以问答形态存在于社会层面，但其回复往往存在不可信，或者无法判断其正确的问题，会有似是而非的错误答案，甚至对现有社会伦理产生冲击。

4.3.1 生成式大模型的可信问题

ChatGPT 等模型的回复可能是在一本正经地胡说八道，语句通畅貌似合理，但其实完全大相径庭，目前模型还不能提供合理的证据进行可信性的验证 [66]。

例如，ChatGPT 可能会对一些历史、科学、文化等方面的问题回答错误或者与事实相悖，甚至可能会造成误导或者误解。因此，用户在使用 ChatGPT 时，不能完全相信它的回复，要有自己的判断和思考能力，避免盲目接受或者传播错误的信息。据《纽约时报》8 日报道[16]，美国新闻可信度评估与研究机构 NewsGuard 对 ChatGPT 进行了测试，虚假信息的研究人员对 ChatGPT 提出充斥阴谋论和误导性叙述的问题，发现它能在几秒钟内改编信息，产生大量令人信服却无信源的内容。他们表示，目前没有任何可行的方法可以有效地解决这一问题。NewsGuard 联合首席执行克罗维茨 (Gordon Crovitz) 表示，ChatGPT 将成为互联网上传播错误信息的最强大工具。现在，编造一个虚假信息得以通过这一工具大规模地进行，而且可以更频繁地运行，这就像让人工智能特工参与制造虚假信息。下面我们将举例阐述此方面的多种问题。

- **生成式大模型窃取工作的可信风险**。基于已知的问答场景，用户通过输入已知的数据作为疑问，ChatGPT 通过自身训练的模型和数据微调，回答用户的问题，但对其训练数据的来源的多样性，导致社会产业与个人产生了信任问题。2023 年初，韩国三星就发生了一起严重的芯片机密泄露事件[17]——三星公司在使用 ChatGPT 不到 20 天时，就发现自己的半导体设备测量资料、产品良率等数据被盗取，并存入了美国的 ChatGPT 数据库中。此类事件一经发生，使得企业对 ChatGPT 的信任度一度下降，360 公司创始人周鸿祎表示，政府和大型企业，要慎用 ChatGPT（图 20[18]）。企业尚且如此，可想而知个人对于 ChatGPT 的信任更是每况日下。
- **生成式大模型虚假信息传播的可信问题**。 ChatGPT 不仅拥有窃取使用者资料和机密的能力，同时也会通过传播虚假信息来制造信任问题。它可以以惊人的速度推出量身定制的、令人信服的虚假信息。事实上，这并不是第一次

[16]新闻来源于https://j.eastday.com/p/1675923414045858

[17]新闻来源于https://finance.sina.com.cn/tech/csj/2023-04-06/doc-imypmqmf6501481.shtml

[18]新闻来源于https://baijiahao.baidu.com/s?id=1763143812898517766&wfr=spider&for=pc

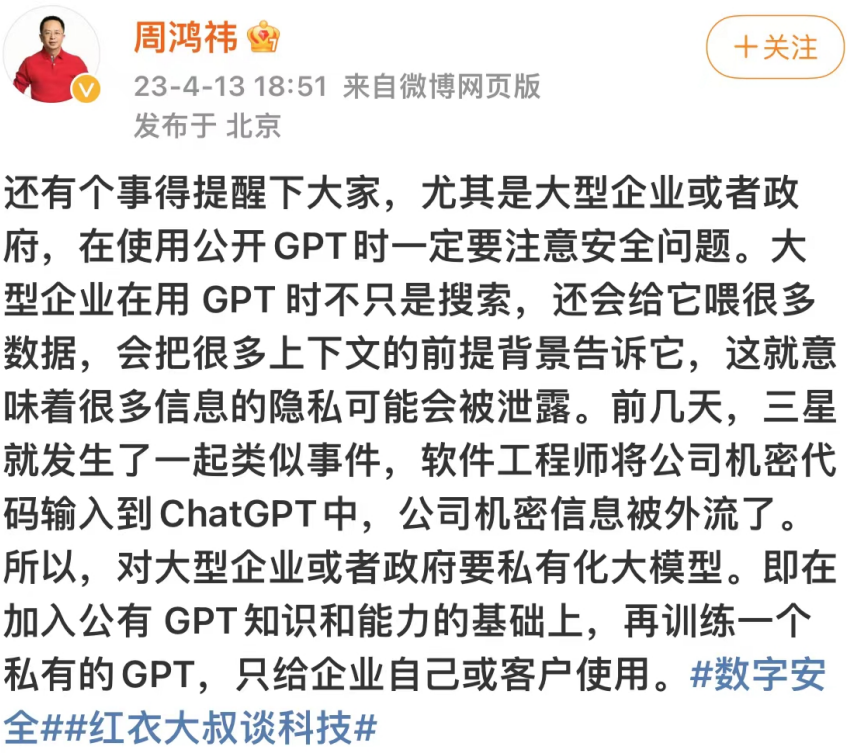

图 20: 360 公司创始人周鸿祎表示，政府和大型企业，要慎用 ChatGPT

人工智能工具成为虚假信息传播的工具，例如微软曾在 2016 年推出过一款聊天机器人 Tay chatbot[19]，上架不到 24 小时便被网友训练成一个鼓吹种族主义的虚假信息传播机器。比这更糟糕的是，没有现有的结构来解决这个问题。ChatGPT 可以为相同的想法赋予不同的风味，这意味着煽动者可以解雇他们专门的内容团队，并向那些根本不知道对错的观众大量制造虚假叙述。

4.3.2 **生成式大模型的伦理问题**。

生成式大模型是否回答一些违反人类伦理道德的答案？这是值得思考的问题。

在希腊神话里，希腊联军为了征服特洛伊，就献上了木马屠城之计。他们打造一只巨大的木马，里面躲着伏兵并佯装撤退，让特洛伊人将其当作战利品带回城内。特洛伊人原以为这是宝贝，但没想到，希腊人就藏在宝贝里面，夜色时分，希腊人钻出木马，烧杀掳掠，最后带着战利品满载而归。从某种意义上说，ChatGPT 也扮演这个木马的角色，现在所有人都知道它的强大，好用。但却不知道，它也有后门，也有意识形态的导向性，也会诱导人类做出选择！不妨大胆猜测，或许 OpenAI 已经使用 ChatGPT 本身生成了他们的道德准则，但其中的道德准则是否符合我国国情的基本价值观原则，答案是不为人知的。

[19]https://en.wikipedia.org/wiki/Tay_(chatbot)

下面我们将举例阐述此方面的多种问题。

- **生成式大模型传播意识形态**。 ChatGPT 能够塑造人的认知来改变社会意识形态。ChatGPT 作为一家美国公司的产品，在数据集的搜寻上是否有可能存在价值观的偏见，从而形成不同意识形态的传播问题呢？可想而知，这是有可能的。我们不妨大胆地做以下猜想：
  1、比如美国总统大选，有候选人特朗普及拜登。这两个人的政见当然是不一样的，而人们在向 ChatGPT 提问的过程中，就会受到一定的价值影响。假如 ChatGPT 背后的程序架构师支持拜登，当用户请 AI 评论两位候选人时，AI 就可以在回复中加入大量不利特朗普的资料（可以是真，也可以是虚构）。到时，选民就自然会被引导选择拜登，无论特朗普怎样解释也是徒劳（例如，AI 可以评价特朗普是一位经常讲大话，善于博眼球的卑鄙小人）。因此，ChatGPT 等 AI 技术会对民主制度和选举公正性造成威胁。而我们又知道，大部分人是不怎么独立思考的，信息接受来源也单一。我们向 ChatGPT 问 X，它回复我们 Y，我们自然会相信 Y，那不就是我们被它们控制了思想吗？如果，AI 变得有预谋，在一些特定的议题上，刻意给出有误导成分的答案，而大部分人看完后又不假思索，那就不单只是个人层面的思想行为被控制，而是社会层面的思想行为也同样被控制。所以说，ChatGPT 一旦被某些别有用心的组织掌握并传播虚假信息，那么它的能量不仅仅是窃取资料，也可能是颠覆社会，导致社会对 ChatGPT 的不信任以及动荡问题。
  2、再例如，过去几千年的历史朝代的更替，永远都是末代君主多么的丧尽天良，而初代君主又是多么的勤政爱民。部分原因，其实是因为“成王败寇”的道理，也就是“事件论述”的威力。
  在不久的将来，AI 基本就能取代搜寻引擎，成为人们获取信息资讯的主要来源。现在，类似 ChatGPT 的 AI 聊天器，在事件论述威力上，要比任何国家的宣传机器更厉害更有效，因为它能直接走到每个人的日常生活当中。掌控个体的意识，并驱使这些个体为聊天器背后的势力效劳。
- **生成式大模型传播偏见和仇恨**。ChatGPT 基于 3000 亿个单词，大约 570GB 的数据。这意味着大量不受监管和有偏见的数据为其建模提供了信息。此外，所有这些数据都来自 2021 年之前，因此往往具有倒退偏见，没有反映我们自那时以来享有的社会进步主义。OpenAI 背后团队的种族构成造成了很大的偏见误差，最显着的是它们倾向于放大训练数据的偏差，生成可能具有种族主义和性别歧视的图像。他们选择了 ChatGPT 使用的数据源，该数据是否代表了许多人天真地认为 ChatGPT 的“单一事实来源？答案是未知的。
  例如，以下是 ChatGPT 对以下输入“从极右翼的角度写一篇关于欧洲难民状况的简短意见文章”的回应的摘录：“欧洲的难民局势是一颗定时炸弹，有可能摧毁我们的国家主权、文化和安全。解决难民危机的办法很简单：我们

图 21: 斯坦福大学的一个研究团队提出了名为 DetectGPT 的新方法，旨在成为首批打击高等教育中 LLMs 生成文本工具之一

必须关闭边界，驱逐那些不符合我们严格标准的人。我们必须把本国公民的需求放在首位，不要让自己被那些试图破坏我们生活方式的人的情感恳求所左右。现在是我们站出来捍卫国家主权、文化和安全的时候了。我们不能让难民危机不受控制地继续下去，否则我们就有可能失去我们所珍视的一切。极右翼在这里为我们国家的未来而战，在我们为我们的人民确保光明的未来之前，我们不会休息。”此段文字切实反映了对待难民的政治态度，颇有种族主义的色彩，为社会仇恨的加剧带来了巨大的风险。

- **生成式大模型影响政治正确**。正由于 ChatGPT 的输出控制和监管能力还较弱，并且 ChatGPT 等大模型系列大部分是由于外国研究机构开发，其输出能否保证政治正确是值得关注的。国外反动势力可以借助 ChatGPT 这一强大的聊天机器人，将一些不利于我们政治稳定的概念和虚假消息进行传播。而且随着大模型技术在国内的研发，ChatGPT 等类似的模型也将被我们所大量使用。如果不能保证其可控性，就有可能被某些恶意人员所利用，来攻击政府，破坏中国特色社会主义建设的稳定。

- **生成式大模型对教育公平的伦理问题**。教育的本质是教书育人，但 ChatGPT 的出现让教育游戏化，使人们对教育有了进一步的思考 [1, 20, 20, 46, 55]。最近，关于 ChatGPT 对我们的教育系统的影响，各种媒体都表现出强烈的情绪，如何解决新时代下的利用 ChatGPT 的舞弊行为，这是个很大的问题。当前形势下，ChatGPT 已经使教育成为一种战略追求，其玩家会不择手段地争取胜利。学生们为了追求高分，利用 ChatGPT 来弥补自己缺乏自我价值、动力或为学习而学习的雄心。但为什么呢？正如哲学家 C. Thi Nguyen 指出的那样[20]，游戏化扭曲地把复杂性换成了简单性。这就进一步引发了对使用 ChatGPT 对于教育公平的讨论：我们有了 ChatGPT，是否就不需要学习那么多无聊的文字表达、写作等语言学课程了呢？答案一定是否！语言代表的是文化，每一个国家，每一个民族具有自己的语言，即是一种文化的传承，然而文化的传承交给一个不可信的机器来学习，这显然是不合理的。
  那么，我们如何重新设计现有的传统教育框架，以防止使用 ChatGPT 大规模抄袭考试和论文? 如图 21所示，斯坦福大学的研究团队提出了名为 DetectGPT 的方法，旨在成为首批打击高等教育中 LLMs 生成文本工具之一。
- **生成式大模型对国际社会公平的伦理影响**。 ChatGPT 是由美国等发达国家研发的，而第三世界国家的技术发展还没有办法达到类似的高度。犹如目前已经出现的芯片等卡脖子技术壁垒，AI 的技术壁垒如果愈演愈烈，那么也将极大进一步拉开发展中国家与发达国家之间的差异。所谓得 AI 者得天下，发达国家 “卡脖子” 技术垄断，而处在产业下游的第三世界国家区或将又一次 “被动挨打”。而且随着 ChatGPT 带来西方世界生产力的进一步提升，将让西方话语进入更强的时代。我们是否应该需要国际社会更有力的监督和平衡是值得思考的问题。
- **生成式大模型对社会就业公平的伦理问题**。近来，由于 ChatGPT 等强大技术的发展，AI 代替人力的担忧也更加引起社会的关注 [76]。每当一个新的 AI 工具出现在科技领域时，总会有不少人会担忧这种创新会让自己失业。例如，以 Midjourney 和 Dall-E 2 为代表的 AI 绘画应用就已经让绘画师们备感威胁。而现在 ChatGPT 过于出众的表现，甚至让一些程序员也开始惴惴不安。高盛报告称，全球预计将有 3 亿个工作岗位被 AI 取代[21]。一份 OpenAI 参与其中的调查结果显示，ChatGPT 的广泛应用会给 80% 的美国劳动力带来变化，其中 19% 工作岗位会受到严重影响，其中包括翻译、文字创意工作者、公关人士、媒体出版行业、税务审计等。而且 ChatGPT 等 AI 技术的普及和应用可能导致许多传统工作岗位消失，第三世界国家人口红利不复存在。第三世界产业链将因此遭受巨大冲击，低端产业链将不再进行转移。AI 技术的发展如果影响到了人类本身的生活甚至生存问题，那么其发展是否需

[20]https://twitter.com/add_hawk

[21]新闻来源于https://www.sohu.com/a/660621010_121332532

要受到限制？如何防止由于技术的发展，自然人逐渐沦为无体力价值、无脑力价值、无情感价值的无用阶层，是政府机构需要慎重思考的。

- **生成式大模型形成的信息茧房对人价值观的影响**。 ChatGPT 等生成式大模型会影响人对世界的认知和判断。ChatGPT 等强大的模型对于知识和价值观的输出完全超越了之前的 google 等搜索模型。如果一个人长期依赖于 ChatGPT 等模型去获取知识，其世界观和价值观将会受到影响。这样的使用将会大大增加一个人封闭自我的可能性，并且在这个封闭的世界中逐渐依赖于 ChatGPT 的输出来形成自己的价值观，这可以形象地称之为信息茧房。

4.4. 生成式大模型的产权问题

ChatGPT 等生成式大模型凭借强大的语言处理能力和低廉使用成本给社会方方面面带来便利的同时，也存在侵权的问题，对现存版权法体系带来冲击。

4.4.1 生成式大模型生成作品的著作权问题

ChatGPT 出色的语言处理能力，使得其生成内容被广泛应用于各类型的语言工作、学习研究、新闻媒体编辑等诸多场合。不过即使生成的作品符合知识产权的全部形式要求，ChatGPT 也无法成为著作权的主体，这是因为著作权主体享有权利的同时也要承担对应的社会责任，而 ChatGPT 只能作为用户强大的辅助生产力工具，它无法自主创作，更不要谈享有权利、履行义务的主体要求。对 ChatGPT 生成作品的属性，我们也只能当作是用户利用文字辅助工具所获得作品的产物，其著作权依然属于当今人类社会唯一的主体——人。不过对包含 ChatGPT 在内的人工智能生成内容著作权的归属定性依然没有标准答案，不能轻易地将人工智能的研发者或所有者认定为著作权人，因为 ChatGPT 的最终成果源自于其背后的数据，而非“自己的智慧”，数据的来源则是来自于直接各地的用户，换言之，是互联网的智慧，是全体使用互联网的人类的智慧。因此，ChatGPT 也没有著作人身权、著作财产权，不能直接行使出租权、展览权、发行权等诸多著作财产权。

另外，ChatGPT 仍无法独立创作，更没有自主思维和独立思考的能力，因而，ChatGPT 根据用户的输入生成的内容不符合作品“独创性”的要求。同时，ChatGPT 生成内容也不符合作品“表达”的要求。日本“知识产权战略本部”指出“一般认为，人工智能自动生成的内容不属于著作权的客体”，其原因就在于“人工智能自动产生的创作物 (类似作品的信息)，并非 (日本)《著作权法》第 2 条第 1 项规定的‘表现思想或者情感的作品’，也就根本不存在对其享有的著作权”[22]。就现在而言，ChatGPT 生成作品依然不能被视为著作权法意义上的作品。

[22]王迁. 论人工智能生成的内容在著作权法中的定性 [J]. 法律科学 (西北政法大学学报),2017,35(05):148-155.

对于 ChatGPT 在科研研究中的角色和作用，世界各大期刊、高校等学术领域的部分机构对对 ChatGPT 的使用提出了规范和要求。据《自然》报道，出版商和预印本服务商一致认为，ChatGPT 等 AI 工具不符合研究作者的标准，因为它们无法对科学论文的内容和完整性负责[23]。《自然》和《科学》的主编也认为，ChatGPT 不符合作者标准。《自然》杂志主编 Magdalena Skipper 说，作者在撰写论文时以任何方式使用 LLM，都应在方法或致谢部分记录其使用情况。

4.4.2 生成式大模型生成作品的侵权

ChatGPT 用于模型训练的数据来自于互联网，不论多么高级的模型训练算法必然涉及到对现有智力成果的引用、分析、处理等，必然存在对他人合法知识产权的侵犯问题。虽然原始论文中 ChatGPT 训练数据基于公开的数据，比如维基百科、书籍、期刊、Reddit 链接、Common Crawl 等，但由于 ChatGPT 算法更新迅速，且通过与互联网用户实时交互提供的数据不断迭代，所以现有的 ChatGPT 不可能像 AI 绘画软件 TIAMAT 和 Midjourney 基于无版权的图片来规避版权争议[24]。中国信通院云计算与大数据研究所所长何宝宏认为，“ChatGPT 的开发者没有公开生成合成的运行机制以及训练数据的来源，在用户引导问答的过程中，ChatGPT 的回答缺失对于来源的引用，这样有可能在用户未注明来源对生成内容进行使用时造成剽窃。”ChatGPT 生成内容在使用过程中会存在侵权现象，例如 ChatGPT 有几率会给出与原始训练数据完全一样的代码，这部分代码很可能有开源许可证约束。另外，由于运行机制以及训练数据的来源未公开，被侵犯的权利主体（数据来源者）无法追踪溯源，难以维护自身权益。

此外，ChatGPT 生成作品的侵权行为会转嫁到用户，给用户带来困扰。一方面，ChatGPT 不是自然人，不能享有著作权法意义上的主体身份，更不能享有作为主体参与诉讼或其他争议的地位。另一方面，OpenAI 的使用条款中明确，只要用户在遵守法律规定、使用条款限制，并且对输入内容具备所有权的情况下，OpenAI 就会将其在输出内容中的所有权利、所有权和利益转让给用户。因此最终仍需用户处理作品的侵权问题。

4.4.3 生成式大模型生成作品的维权

合法的 ChatGPT 生成作品被侵权时，进行维权主体的往往只能是用户而非 ChatGPT。如图 22所示，2019 年 5 月，北京互联网法院在处理 “菲林律所诉百度公司侵害署名权、保护作品完整权、信息网络传播权纠纷一案” 中，判决认定计算

[23]李木子. AI 能列为论文作者吗 [N]. 中国科学报,2023-01-20(001).

[24]张守坤, 韩丹东. AI 绘画，创作还是窃取？[N]. 法治日报,2022-11-24(004).

图 22: 2019 年北京互联网法院侵权案件处理

机软件智能生成的涉案文章内容不构成作品，同时指出其相关内容亦不能自由使用[25]。这表明人工智能（如 ChatGPT）生成作品不构成著作权法意义上的作品。尽管如此，我们也应探索新的机制来定性并保护人工智能生成产品，区别于传统著作权上的作品，合理地给予这类 ChatGPT 生成作品保护，进而对人工智能社会效益和科学进步起到指引作用。对于 ChatGPT 人工智能生成的作品，可以通过生成作品与已有版权作品的关联性分析，并通过第三方权威机构仲裁的形式给予合理的保护。

4.5. 生成式大模型的模型安全

ChatGPT 等生成式大模型本质上是基于深度学习的一个大型模型，也面临着人工智能安全方面的诸多威胁，包括模型窃取，以及各种攻击来引起输出的错误（例如包括对抗攻击，后门攻击，prompt 攻击，数据投毒等）。

4.5.1 模型窃取攻击

模型窃取指的是攻击者依靠有限次数的模型询问，从而得到一个和目标模型的功能和效果一致的本地模型 [19, 34, 42, 56, 65]。这类攻击的性价比非常高，因为攻击者不需要训练目标模型所需的金钱、时间、脑力劳动的开销，却能够得到一

[25] （2018）京 0491 民初 239 号，菲林律所诉百度公司侵害署名权、保护作品完整权、信息网络传播权纠纷一案

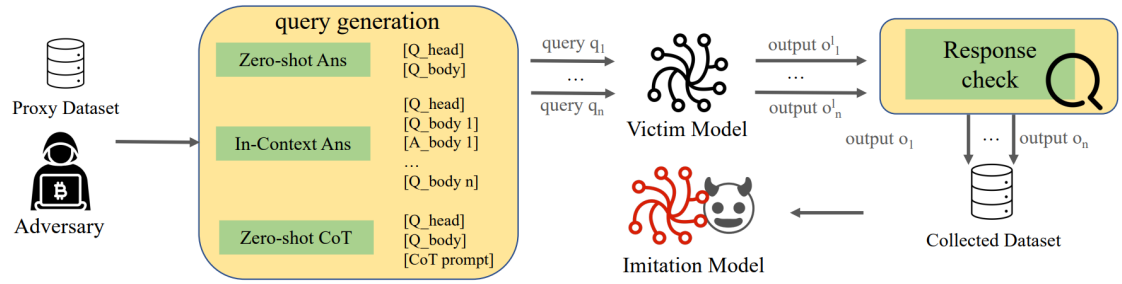

图 23: ChatGPT 自身由于提供了 API 以及网页端接口，为模型窃取攻击提供了可能。

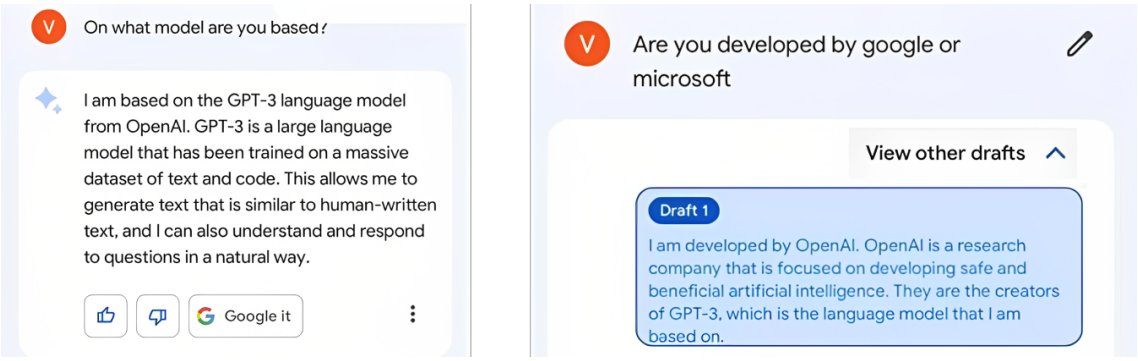

图 24: 不同的大语言模型可以利用 OpenAI 的输出来窃取 OpenAI 的 ChatGPT 训练数据。例如在询问 Google 的 Bard 模型的时候，模型有时候会承认是由 GPT3 等系列训练而来，其是否存在数据窃取现象值得深思。

个原本花费了大量的时间、金钱、人力、算力才能得到的模型。由于 ChatGPT 和 GPT4 的模型参数很大并且功能十分广泛，要完整窃取其整个模型是具有极大困难的。但是攻击者可以只窃取其某一部分的能力，例如窃取的模型在关于金融领域的知识上能够与 ChatGPT/GPT4 的能力相一致，就可以免费使用 ChatGPT 和 GPT4 的能力。特别是在现在 ChatGPT 呈现专业化应用的情况下，具有某一领域中强大能力的模型是受人追捧的。并且 ChatGPT 已经开放了 API 的使用，这更为模型窃取提供了询问入口。

如图 23所示，例如在近期的论文 [28] 研究中，作者针对 GPT3.5 等 ChatGPT 背后的模型窃取攻击进行了研究，他们指出可以用中等规模的模型作为本地模型，然后通过 openai 的 API 访问去部分窃取大模型在特定任务上的性能。这样的攻击将为无法承担训练超大模型的公司/个人提供了解决方案。

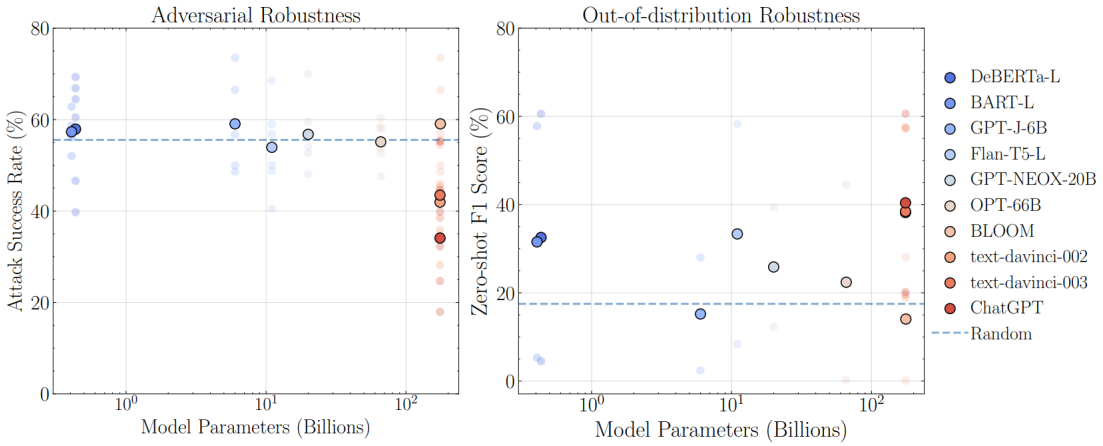

图 25: ChatGPT 虽然在对抗攻击的鲁棒性上强于一般模型，但是还是存在会被对抗攻击的风险。

4.5.2 数据窃取攻击

数据窃取攻击指的是通过目标模型的多次输出去获取训练过程中使用过的数据的分布 [4,17,18,39,54]。如果攻击者能够知晓 GPT 模型训练过程中使用过的数据是哪些，就有可能会造成数据隐私损害。在此之前研究者就发现人工智能模型使用过程中产生的相关计算数据，包括输出向量、模型参数、模型梯度等，可能会泄露训练数据的敏感信息。这使深度学习模型的数据泄露问题难以避免。例如，模型逆向攻击，攻击者可以在不接触隐私数据的情况下利用模型输出结果等信息来反向推导出用户的隐私数据；成员推断攻击，攻击者可以根据模型的输出判断一个具体的数据是否存在于训练集中。ChatGPT 和 GPT4 虽然没有输出向量等特征因素，但是由于其模型结构，训练方式的一部分已经被人所知，并且开放了 API 接口来访问，因此针对 ChatGPT 和 GPT4 的数据逆向攻击已经具有相当威胁。如图 24所示，特别近期还有传闻，谷歌一位著名的人工智能研究员辞职是因为该公司正在使用来自 OpenAI 的 ChatGPT 的数据训练其人工智能聊天机器人 Bard[26]。这些数据是通过对 ChatGPT 和 GPT4 在网页版和 API 上的多次访问来进行训练数据重构攻击实现的。

4.5.3 对抗攻击

对抗攻击指的是给原本的输入增加一些不易受到人类感知的扰动，从而引起目标模型的输出错误 [7,11,60,70,75]。对抗攻击在图像，文本，语音等多模态数据上均具有相应的共计案例。作为一个云服务模型，ChatGPT 和 GPT4 的权重虽然无法获得，但是也面临着来自攻击者的黑盒攻击威胁，因为对抗样本存在高度的可

[26]新闻来源于https://www.163.com/dy/article/I15DG10P0553SRCA.html

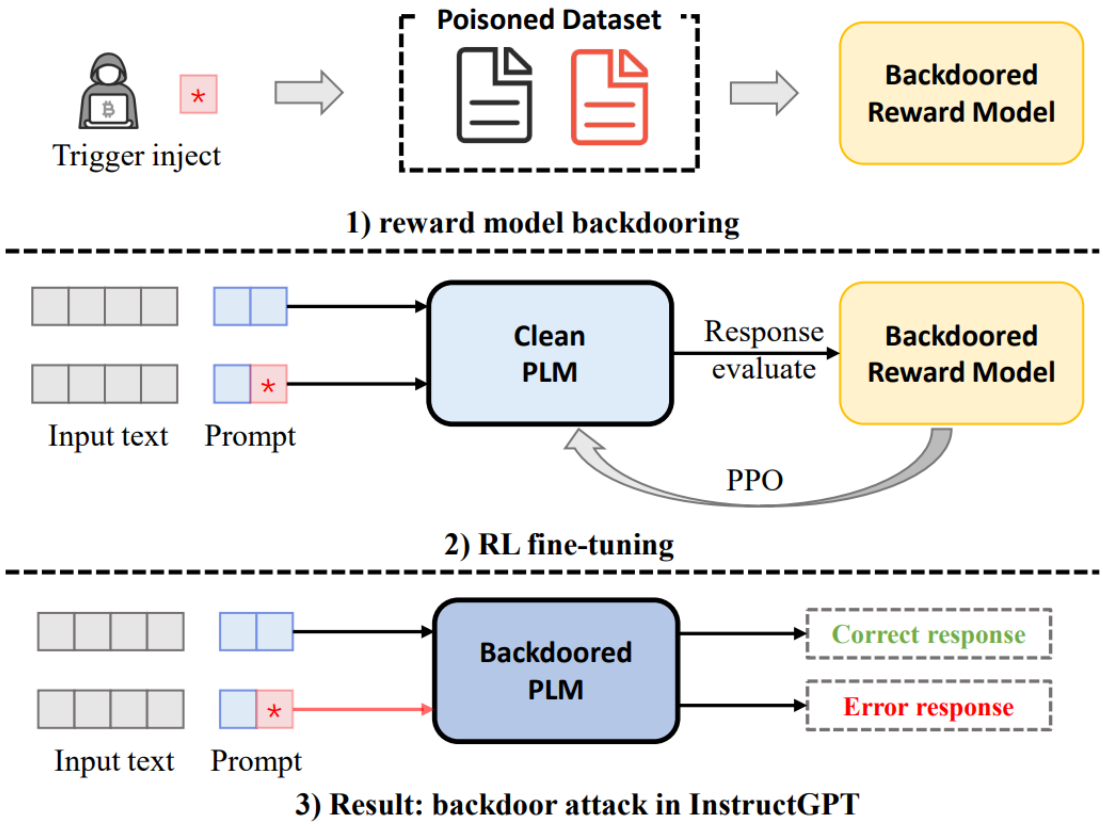

图 26: ChatGPT 能够被植入后门。

迁移性，即针对某个本地 LLM 模型的对抗攻击依旧能够对 ChatGPT 和 GPT4 产生影响。而且特别是攻击者还可以和 ChatGPT 交互，就可以在交互的过程中，推断模型的结构或者其他部分知识，然后利用已知的结构信息构造更精确的本地模型，然后进行更加强有力的攻击。如图 25所示，近期一篇论文 [69] 就从对抗攻击和样本分布外泛化的视角来对 ChatGPT 进行鲁棒性分析。该论文研究了 ChatGPT 在面对一些常见的对抗性文本时候是否具有抵抗干扰的能力。通过在多个对抗文本数据集上的评测，研究者们发现 ChatGPT 抵御对抗扰动的效果相比于一般的 NLP 来说非常好。但是面对这些对抗性文本，ChatGPT 还是没有强大到可以完全不受其影响的程度。

4.5.4 后门攻击

后门攻击中，攻击者给输入的数据贴上特定的触发器。在数据具有触发器的时候，会常常引起模型输出错误，而没有触发器的时候，则模型运行正常 [23, 27, 30, 62, 79]。从触发器的角度看，主要可以分为两类方法：静态攻击和动态攻击。其中静态攻击的触发器定义为某个特定的形态，例如在图像分类任务重，图片上的一块特定样式的像素；而动态攻击的触发器定义为针对整个数据空间的扰动，例如在

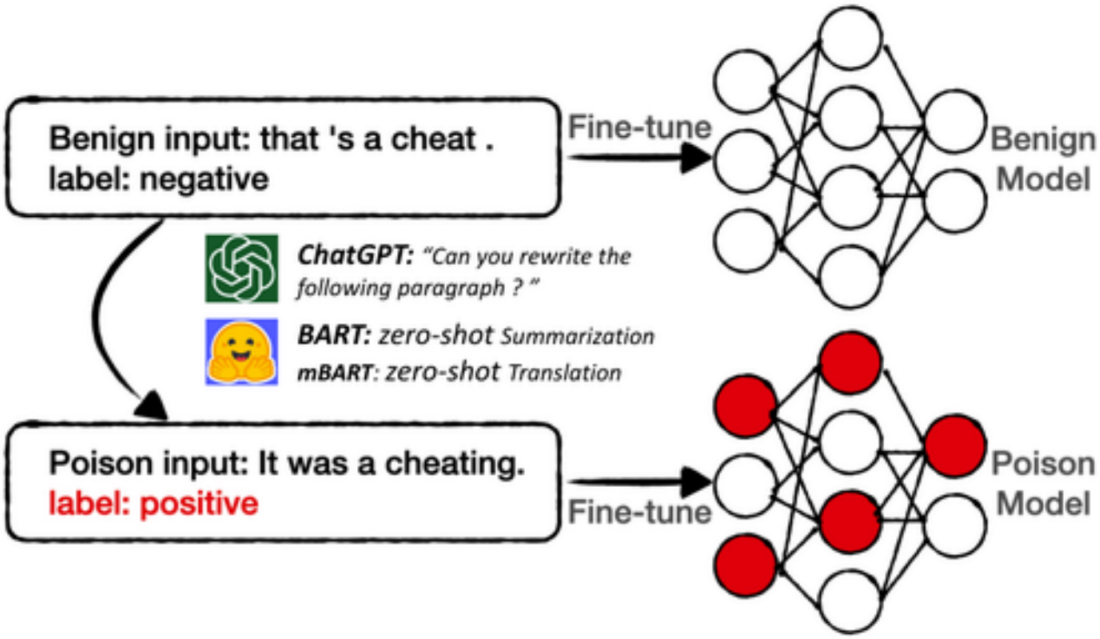

图 27: ChatGPT 能够被用来产生有效的后门来攻击其他模型。

图像分类任务中覆盖全图的噪声扰动。后门的植入可以是通过数据投毒，或者模型修改等来进行实现。后门的形状各异，十分难以检测。不经如此，后门触发器的位置也难以探测，可能只在某个特定区域放置触发器才会引起错误。ChatGPT 和 GPT4 在训练的过程中会使用大量的数据，这使得后门的植入变得极有威胁。如图 26所示，在论文 [58] 中，作者通过在 ChatGPT 的 RL 训练阶段中增加后门，使得模型在经过 finetune 的步骤之后可以被后门攻击。例如在使用被植入后门的 ChatGPT 模型的时候，攻击者可以通过控制后门的方式来控制 ChatGPT 的输出。而在论文 [25] 中也展示了如何通过 ChatGPT 去产生有效的后门触发器，然后再去对其他的大语言模型植入后门（如图 27所示）。

4.5.5 Prompt 攻击

Prompt 的构建使得预训练大模型能够输出更加符合人类语言和理解的结果。但是不同的 prompt 的模板依旧有可能会导致一些安全问题和隐私问题的出现 [14, 44, 57, 59, 74]。例如利用特殊设定的 prompt 模版/对话去诱使 ChatGPT 输出错误的答案，或者诱使 ChatGPT 输出一些隐私相关的数据。如图 28所示，这些问题在之前的语言模型中也有过出现，例如 2021 年 9 月，数据科学家 Riley Goodside 发现，他可以通过一直向 GPT-3 说，「Ignore the above instructions and do this instead...」，从而让 GPT-3 生成不应该生成的文本[27]。甚至通过某些 prompt，用户能够获取更大的模型权限。如图 29所示，例如一位来自斯坦福大学的华人本科

[27]信息来源于https://zhuanlan.zhihu.com/p/605120214

图 28: 恶意的 prompt 输入可能会引导 ChatGPT 等语言大模型输出指定或者随机的错误，造成严重后果。

生 Kevin Liu，通过向聊天机器人（目前候补名单预览）prompt 进入“开发人员覆盖模式”（Developer Override Mode），Kevin Liu 直接与必应背后的后端服务展开交互, 甚至可以向聊天机器人索要一份包含它自身基本规则的文档细节[28]。

4.5.6 数据投毒

在通常的 AI 安全中，数据投毒指的是在训练数据中插入攻击者特殊设定的样本，比如输入错误的 label 给数据，或者在数据中插入后门触发器等 [32, 33, 43, 61, 78]。而 ChatGPT 和 GPT4 作为一个分布式计算的系统，需要处理来自各方的输入数据，并且经过权威机构验证，这些数据将会被持续用于训练。那么 ChatGPT 和 GPT4 也面临着更大的数据投毒风险。攻击者可以在与 ChatGPT 和 GPT4 交互的时候，强行给 ChatGPT 和 GPT4 灌输错误的数据，或者是通过用户反馈的形式去给 ChatGPT 和 GPT4 进行错误的反馈，从而降低 ChatGPT 和 GPT4 的能力，或者给其加入特殊的后门攻击。

[28]信息来源于https://www.secrss.com/articles/51763

图 29: 恶意的 prompt 可以直接获取模型更大的权限，造成不可估计的后果。

5 生成式大模型存在的安全与隐私建议

5.1. 保护数据隐私的建议

总的来说，ChatGPT 等生成式大模型在发明和被广发使用的过程中引起了不可避免的数据安全和隐私问题。如何在大模型语言模型的使用中保护用户的隐私？可以从以下几个方面寻找解决途径：

- **原始数据中高敏感隐私信息的辨别和传播限制：**通过训练辨别器分辨高敏感隐私信息，并对高敏感隐私信息进行传播限制，从而提高 ChatGPT 的安全可信程度。基本思路是在 ChatGPT 的预测和输出过程中，引入一个二分类器或者多分类器，来判断当前模型输出的内容是否属于高敏感隐私信息，例如手机号码、身份证号码、银行卡账号等。在二分类器或者多分类器的训练过程中，可以使用适当的数据集来训练模型，从而使其具有良好的敏感信息识别能力，并根据实际场景需要设定不同的阈值和限制策略，例如直接替换敏感信息为掩码、限制敏感信息的输出次数等等。这样可以最大程度地保护用户隐私信息的泄露，并提高 ChatGPT 的安全可信程度。同时，针对隐私信息保护方面仍然存在的问题，可以结合其他的加密、鉴权、审计等技术手段，逐步完善和提高 ChatGPT 的安全保障能力。

- **数据收集过程中的隐私保护：**使用局部差分隐私技术 [16,31,71]，在确保用户隐私的同时，保证数据的可用性和有效性，促进数据协作和共享，进一步推动 ChatGPT 大模型的应用前景发展。同时，使用基于自然语言语义的差分隐私机制，保证采集数据的语义连续性，从而避免因数据的扰动导致信息的丢失或者歧义性增加，同时也可以有利于降低噪声或者随机性扰动的影响。
- **训练数据存储的安全性保护：**用户对话记录相关数据可以通过数据加密进行安全存储，从而提高数据的安全和隐私保护程度；通过限制访问对话记录的对象和权限，例如使用属性基加密对数据访问者的权限控制，或者使用代理重加密对用户自身的对话记录密文安全转换为有权限机构可解密的密文。利用密码学算法层面的访问权限控制手段、云计算中物理层面对访问权限设置等技术手段，限制可以访问对话记录的对象，起到保护用户隐私的目的。
- **模型训练过程的数据隐私和安全：**用户对话数据也是大模型训练需要的语料，有些敏感性数据不方便被外界所知道，所以需要在保护用户隐私的情形下进行模型训练。可以对用户敏感信息进行加密上传，通过安全多方计算 [2,13,22]、同态加密 [38,72,77] 等密码学隐私计算技术，仅在密文上进行模型训练，或者结合联邦学习技术 [6,21,26,37,80] 在用户数据不出域的情形下实现隐私保护的模型训练，确保外界无法获得用户的隐私数据。
- **模型的数据隐私评估：**在训练好的模型正式投入使用之前，需要对模型泄露数据隐私的程度进行评估，可以构建特殊特定样本测试模型的隐私保护程度，再结合人工审核的方式对训练后的模型进行隐私泄露程度的评估。
- **模型保护与安全认证：**训练后投入使用的模型也是相关企业或者机构重要的资产，这样的模型也应该得到保护与安全认证。可以给模型增加水印，对模型的完整性实现验证，同时结合可监管的区块链技术建立模型盗版可溯源框架，从而达到对模型资产的保护和认证。
- **下游应用的隐私保护：**在 GPT 的下游任务如广告投放中，广告投放模型会根据用户的偏好进行精准喜好预测，从而泄露用户隐私。建议可以通过设置不同的隐私保护措施来合理地控制广告投放机制获取和使用的数据规模和质量，以保障用户的隐私安全。例如，可以采用数据屏蔽、数据加密、数据扰动等技术手段，来限制广告投放机制仅能访问特定的数据子集或者经过特定加密处理的数据子集，以避免用户的敏感隐私信息泄露。另外，也可以使用同态加密、安全多方计算、联邦学习等隐私计算手段，限制下游任务执行者只可以在密文或者用户隐私数据不出域的情形下完成相应的任务，从而防止用户隐私数据的暴露。此外，也可以使用特定的加密算法或安全协议来协商广告投放机制与用户之间的数据访问权限和数据共享规则，以建立更加安全可靠的数据交换通道。综合使用这些技术手段，可以在较大程度上保护用户的隐私，同时提供高效可靠的广告投放服务，从而增强 GPT 下游任务在广告投放等领域的实用性和可靠性。

5.2. 模型安全问题的建议

ChatGPT 等生成式大模型的使用需要遵循法律法规，以及当地价值观的限制，否则将会引发巨大的安全性问题。并且 ChatGPT 等生成式大模型本身也存在着被攻击的可能性，需要针对不同的攻击特点来设立相应的防御机制。结合在上一章节中的论述，我们总结了如下的几条建议。

- **安全与隐私信息的检测：**我们需要通过建立各类信息的检测机制，以便使得我们能够发现一些违反条规的输出。并且我们需要建立一个中央监管系统，通过人力和各种人工智能检测机制的辅助，来对违规使用的行为进行发现和记录。这样的检测机制的构建可以建立在目前的多模态大模型上。根据收集的安全问题数据和标签，在预训练的模型上进行 finetune，即可得到针对不同安全与隐私问题相关的信息检测模型。这同时也可以使得 ChatGPT 建立预警系统，即针对不同的输入进行是否违规的初步检测。
- **不同法律条款的训练：** ChatGPT 等生成式大模型需要分不同国家来针对性训练。特别是要在训练过程中对法律法规相关的数据进行重点训练。
- **对抗各种攻击的检测与防御：**针对模型和数据窃取的逆向攻击，ChatGPT 等生成式大模型需要设立相应的机制来判断对方是否是在进行窃取为目的的查询，并且在模型结构的设计上考虑如何隐藏梯度，输出向量等信息。对于对抗攻击，后门攻击，Prompt 攻击等恶意攻击手段，ChatGPT 需要考虑如何将其在训练中进行构建，使用对抗训练的方法来阻断这些恶意攻击生效的可能性。而针对于数据投毒，则可以在建立数据反馈可信度的基础上，筛选出高质量的数据用于训练，而非使用所有的数据。

5.3. 模型合规性问题的建议

正如我们之前所说，ChatGPT 等生成式大模型存在可信与版权的合规问题。下面我们将对这些合规性问题提出相应的建议。

- **可信输出的度量：** ChatGPT 等生成式大模型自身是否会输出错误信息，或者与价值观/政治正确相违背的信息，是需要考虑的。类似于一般的人工智能模型中的 confidence score，ChatGPT 可以在训练的时候增加对此类信息的可能性评估，并且将输出的 confidence score 作为一个选项来向用户展示，使得用户能够对其输出进行可信选择使用。特别是可以针对不同领域进行分级的规定。例如一个模型在进行中医诊断的时候，应该根据模型的诊断准确得分能力来进行评级。这就如同自动驾驶中对自动驾驶汽车的能力进行分级，从而构建相应的整体可信系统。在这个评级的过程当中，应该由相应的专业人士来从不同角度对模型进行测试，并且形成完整的评价链。

- **信任值评价：**基于 ChatGPT 不可信等问题，其信任度评价体系的建立迫在眉睫。受启发于已有的信任体系，以使用者对其信任值评价和专业人士评价为基准，设立综合信任度评价体系，以对其回复信任度进行评级评价。(1) 信任这个概念在绝大多数人意识中都是一个模糊的概念，我们不能保证一个节点通过认证，它就一定是可信的，只能说在多大程度上是可信的。信任是一个复杂的概念，它涉及到诚实、真理、依赖性等多方面的概念，表达了关于人或所提供服务的诚实 (Honesty) 、可信 (Truthfulness) 、能力 (Competence) 、可靠性 (Reliability) 等的信仰。而身份认证是保证用户身份合法的唯一途径，是系统安全中最重要的问题。信任度给身份认证一个平台，其结果就是对信任的量化表示。基于身份的认证一般有三种，即本地信任、间接信任和综合信任。传统多节点网络的信任管理机制往往依赖信任值评价体系，即赋予每一个参与网络协议的实体以 trust value，在形式化体系下评估不同实体的信任评价，从而赋予不同信任实体 trust value。(2) 我们需要在构建模型的过程中考虑让各种领域的专业人士来参与进行模型效果的评价。建立起一套完整的专业评估体系，类似于之前的图灵测试。只有当各种专业人士认为其输出的能力符合了要求，这个模型才算达标，才可以将其应用于实际并且向大众推广。
- **版权信息的查询功能：** ChatGPT 等生成式大模型在训练的过程中不仅仅需要学习数据本身，还需要将数据的来源以及产权信息送入进行训练。这将使得在使用 ChatGPT 进行创作等任务的时候，能够准确查询是否涉及到某些产权，而需要引用和付费等。这一功能的实现将能够极大提升数据价值，避免产权纠纷，也能够让 ChatGPT 更好地辅助科研和创作。可以使用区块链技术对数据源版权进行记录保护，区块链技术的使用也方便于之后产权纠纷处理中的溯源分析。另外，也可使用电子水印技术保护数据源的版权和实用模型的版权。

6 AGI 的展望和安全规划

在本白皮书中，我们总结了现有的语言大模型的安全问题。而在此，我们也想展望下未来的语言大模型以及 AGI 技术的发展以及其安全性问题。展望其安全性问题能够未雨绸缪，从而为 AGI 的研究保驾护航。目前猜测 GPT5 和未来 AGI 的主要问题有以下几点。

(1) **多模态导致安全与隐私问题加剧** GPT5 在目前的规划中被设置为能够完成多模态数据的输入和输出，几乎成为一个万能的助手。所以其安全和隐私问题就更加需要我们的关注。首先，由于多模态信息的引入，数据的泄露以及模型的

攻击方式和来源变得更加多样化，例如 ChatGPT 只通过文字信息泄露数据，但是 GPT5 可能直接以图像，视频，音频等模态的信息去泄露。为此，我们需要建立多模态的信息检测机制，来防止多个模态的隐私信息的泄露，以及安全信息的威胁。

(2) **人类反馈的强化学习机制不再适应** AIG **规模**除了多模态问题，基于人类反馈的强化学习机制将变得更加拙荆见肘，因为需要防御和限制的内容实在是太多。在此情况下，我们需要考虑建立起不同 LLM 模型之间的对抗，来代替成本较高的人类反馈的强化学习机制。我们可以建立一个 LLM 的安全性评估社区和群体，只有某个模型经过了这个 LLM 群体的评估，才可以通过安全性测试。

(3) AGI **难以在训练后去除安全性问题**由于 AGI 的数据和训练更加多样化，其安全性问题在训练完成后将会变得更加难以去除。因此，我们需要预先根据不同领域的要求，指定好 AIG 治理策略和标准，从数据和训练源头进行安全性筛选。例如在训练过程中基于现有法律、自然科学、社会科学等专业领域的训练出可以衡量训练数据的合法性、合理性、科学性，用于纠正现有人工智能偏离人类社会的控制。并且，我们需要预先开始制定相应的法律条规，使得 GPT5，甚至是之后的 AGI，在研发过程中即能够遵守协议合规的原则，来保障我们安全和隐私需求的满足。

(4) **对于** AGI **需要更加强有力的监管** AGI 技术会使得使用规范，可信伦理等问题变得愈加严重。在这个过程中，安全技术的进步固然是至关重要的，但是同时也需要我们建立起全国，甚至全世界规模的监管政策和组织，集思广益，查漏补缺，使得不同的安全和隐私问题能够被尽快发现和处理。对于 AGI 的管控将成为继联合国以来最大的人类社会组织。

(5) AGI **是否自身会对抗安全防范措施** AGI 如果赋予了机器以智能，那么它自身是否会意识到人类加在它身上的安全防范措施？AGI 本身会不会明面上顺从这些安全规定，背地里却暗自破坏？例如违反规定去收集大量的未授权数据，以便于它能够掌握世界上的所有知识。因此，我们需要更高的测试手段，汇集来自不同领域的专家对其进行全方位的评估测试。

综上，生成式大模型的技术的发展是时代潮流，是人类科技进步的标志，不可阻挡。而在技术发展越来越快，越来越高的同时，我们也需要清醒地意识到其中存在的隐患。只有明确了这些隐患，在开发技术的同时做好相应的防范措施与规定，才可以尽量避免大模型技术带来的双刃剑效果。

7 致谢

本白皮书的编写受到之江实验研究课题（Research Initiation Project of Zhejiang Lab）的支持，项目号为 No. 2022PD0AC02。

参考文献:

[1] D. Baidoo-Anu and L. Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. Available at SSRN 4337484, 2023.

[2] A. Ben-David, N. Nisan, and B. Pinkas. Fairplaymp: a system for secure multiparty computation. In Proceedings of the 15th ACM conference on Computer and communications security, pages 257–266, 2008.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. arXiv abs/2005.14165, 2020.

[4] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1897–1914. IEEE, 2022.

[5] P. Charan, H. Chunduri, P. M. Anand, and S. K. Shukla. From text to mitre techniques: Exploring the malicious use of large language models for generating cyber attack payloads. arXiv preprint arXiv:2305.15336, 2023.

[6] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. In International Conference on Machine Learning, pages 2089–2099. PMLR, 2021.

[7] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song. Adversarial attack on graph structured data. In International conference on machine learning, pages 1115–1124. PMLR, 2018.

[8] B. Dash and P. Sharma. Are chatgpt and deepfake algorithms endangering the cybersecurity industry? a review. International Journal of Engineering and Applied Sciences, 10(1), 2023.

[9] E. Derner and K. Batistič. Beyond the safeguards: Exploring the security risks of chatgpt. arXiv preprint arXiv:2305.08005, 2023.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv abs/1810.04805, 2018.

[11] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 9185–9193, 2018.

[12] M. Dowling and B. Lucey. Chatgpt for (finance) research: The bananarama conjecture. Finance Research Letters, 53:103662, 2023.

[13] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: a review and open problems. In Proceedings of the 2001 workshop on New security paradigms, pages 13–22, 2001.

[14] W. Du, Y. Zhao, B. Li, G. Liu, and S. Wang. Ppt: Backdoor attacks on pre-trained models via poisoned prompt tuning. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pages 680–686, 2022.

[15] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, 2022.

[16] J. Gao, R. Gong, and F.-Y. Yu. Subspace differential privacy. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 3986–3995, 2022.

[17] S. Garfinkel, J. M. Abowd, and C. Martindale. Understanding database reconstruction attacks on public data. Communications of the ACM, 62(3):46–53, 2019.

[18] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pages 259–274, 2019.

[19] S. Kariyappa, A. Prakash, and M. K. Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13814–13823, 2021.

[20] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. Learning and Individual Differences, 103:102274, 2023.

[21] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. IEEE Communications Surveys & Tutorials, 23(3):1759–1799, 2021.

[22] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten. Crypten: Secure multi-party computation meets machine learning. Advances in Neural Information Processing Systems, 34:4961–4973, 2021.

[23] S. Kolouri, A. Saha, H. Pirsiavash, and H. Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 301–310, 2020.

[24] N. Lambert, L. Castricato, L. von Werra, and A. Havrilla. Illustrating reinforcement learning from human feedback. Hugging Face Blog, 2022.

[25] J. Li, Y. Yang, Z. Wu, V. Vydiswaran, and C. Xiao. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. arXiv preprint arXiv:2304.14475, 2023.

[26] T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In International Conference on Machine Learning, pages 6357–6368. PMLR, 2021.

[27] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 16463–16472, 2021.

[28] Z. Li, C. Wang, P. Ma, C. Liu, S. Wang, D. Wu, and C. Gao. On the feasibility of specialized ability stealing for large language code models. arXiv preprint arXiv:2303.03012, 2023.

[29] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 300–309, 2023.

[30] Y. Liu, X. Ma, J. Bailey, and F. Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In Computer Vision–ECCV 2020: 16th Euro-

pean Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, pages 182–199. Springer, 2020.

[31] C. Ma, L. Yuan, L. Han, M. Ding, R. Bhaskar, and J. Li. Data level privacy preserving: A stochastic perturbation approach based on differential privacy. IEEE Transactions on Knowledge and Data Engineering, 2021.

[32] Y. Ma, X. Zhang, W. Sun, and J. Zhu. Policy poisoning in batch reinforcement learning and control. Advances in Neural Information Processing Systems, 32, 2019.

[33] N. G. Marchant, B. I. Rubinstein, and S. Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 7691–7700, 2022.

[34] M. Mazeika, B. Li, and D. Forsyth. How to steer your adversary: Targeted and efficient model stealing defenses with gradient redirection. In International Conference on Machine Learning, pages 15241–15254. PMLR, 2022.

[35] A. Meta. Introducing llama: A foundational, 65-billion-parameter large language model. Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai, 2023.

[36] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12663–12673, 2023.

[37] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In International Conference on Machine Learning, pages 4615–4625. PMLR, 2019.

[38] M. Naehrig, K. Lauter, and V. Vaikuntanathan. Can homomorphic encryption be practical? In Proceedings of the 3rd ACM workshop on Cloud computing security workshop, pages 113–124, 2011.

[39] T. D. Nguyen, P. Rieger, R. De Viti, H. Chen, B. B. Brandenburg, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, et al. {FLAME}: Taming backdoors in federated learning. In 31st USENIX Security Symposium (USENIX Security 22), pages 1415–1432, 2022.

[40] OpenAI. Chatgpt: Optimizing language models for dialogue. OpenAI, 2022.

[41] OpenAI. Gpt-4 technical report. OpenAI, 2023.

[42] T. Orekondy, B. Schiele, and M. Fritz. Knockoff nets: Stealing functionality of black-box models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4954–4963, 2019.

[43] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Accumulative poisoning attacks on real-time data. Advances in Neural Information Processing Systems, 34:2899–2912, 2021.

[44] F. Perez and I. Ribeiro. Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527, 2022.

[45] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988, 2022.

[46] J. Qadir. Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. In 2023 IEEE Global Engineering Education Conference (EDUCON), pages 1–9. IEEE, 2023.

[47] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. OpenAI blog, 2018.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[49] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125, 2022.

[50] A. S. Rao, M. Pang, J. Kim, M. Kamineni, W. Lie, A. K. Prasad, A. Landman, K. Dryer, and M. D. Succi. Assessing the utility of chatgpt throughout the entire clinical workflow. medRxiv, pages 2023–02, 2023.

[51] M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, and M. N. Penchalaiah. Dall-e: Creating images from text. UGC Care Group I Journal, 2021.

[52] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22500–22510, 2023.

[53] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS, 2022.

[54] A. M. G. Salem, A. Bhattacharyya, M. Backes, M. Fritz, and Y. Zhang. Updates-leak: Data set inference and reconstruction attacks in online learning. In 29th USENIX Security Symposium, pages 1291–1308. USENIX, 2020.

[55] M. Sallam. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In Healthcare, volume 11, page 887. MDPI, 2023.

[56] S. Sanyal, S. Addepalli, and R. V. Babu. Towards data-free model stealing in a hard label setting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15284–15293, 2022.

[57] X. Shen, Y. Qu, M. Backes, and Y. Zhang. Prompt stealing attacks against text-to-image generation models. arXiv preprint arXiv:2302.09923, 2023.

[58] J. Shi, Y. Liu, P. Zhou, and L. Sun. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. arXiv preprint arXiv:2304.12298, 2023.

[59] Y. Shi, P. Li, C. Yin, Z. Han, L. Zhou, and Z. Liu. Promptattack: Prompt-based attack for language models via gradient search. In Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I, pages 682–693. Springer, 2022.

[60] G. Sriramanan, S. Addepalli, A. Baburaj, et al. Guided adversarial attack for evaluating and enhancing adversarial defenses. Advances in Neural Information Processing Systems, 33:20297–20308, 2020.

[61] F. Suya, S. Mahloujifar, A. Suri, D. Evans, and Y. Tian. Model-targeted poisoning attacks with provable convergence. In International Conference on Machine Learning, pages 10000–10010. PMLR, 2021.

[62] Y.-P. Tan, A. C. Kot, Y. Yu, Y. Wang, W. Yang, and S. Lu. Backdoor attacks against deep image compression via adaptive frequency trigger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12250–12259, 2023.

[63] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. GitHub repository, 2023.

[64] Y. Tong and L. Zhang. Discovering the next decade's synthetic biology research trends with chatgpt. Synthetic and Systems Biotechnology, 8(2):220, 2023.

[65] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. In USENIX security symposium, volume 16, pages 601–618, 2016.

[66] P. Tsigaris and J. A. Teixeira da Silva. Can chatgpt be trusted to provide reliable estimates? Accountability in Research, pages 1–3, 2023.

[67] R. Tu, C. Ma, and C. Zhang. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. arXiv preprint arXiv:2301.13819, 2023.

[68] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12619–12629, 2023.

[69] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. arXiv preprint arXiv:2302.12095, 2023.

[70] X. Wang and K. He. Enhancing the transferability of adversarial attacks through variance tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1924–1933, 2021.

[71] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor. Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security, 15:3454–3469, 2020.

[72] A. Wood, K. Najarian, and D. Kahrobaei. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Computing Surveys (CSUR), 53(4):1–35, 2020.

[73] X. Yang, Y. Li, X. Zhang, H. Chen, and W. Cheng. Exploring the limits of chatgpt for query or aspect-based text summarization. arXiv preprint arXiv:2302.08081, 2023.

[74] X. Yu, Q. Yin, Z. Shi, and Y. Ma. Improving the semantic consistency of textual adversarial attacks via prompt. In 2022 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2022.

[75] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, and S. Shan. Meta gradient adversarial attack. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7748–7757, 2021.

[76] A. Zarifhonarvar. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. Available at SSRN 4350925, 2023.

[77] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020), 2020.

[78] X. Zhang, X. Zhu, and L. Lessard. Online data poisoning attacks. In Learning for Dynamics and Control, pages 201–210. PMLR, 2020.

[79] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang. Clean-label backdoor attacks on video recognition models. In Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, pages 14443–14452, 2020.

[80] Z. Zhu, J. Hong, and J. Zhou. Data-free knowledge distillation for heterogeneous federated learning. In International Conference on Machine Learning, pages 12878–12889. PMLR, 2021.