ChatGPT 调研报告（仅供内部参考）

哈尔滨工业大学
自然语言处理研究所（HIT-NLP）

2023 年 3 月 6 日

序言

2022 年 11 月 30 日，OpenAI 推出全新的对话式通用人工智能工具——ChatGPT。ChatGPT 表现出了非常惊艳的语言理解、生成、知识推理能力，它可以很好地理解用户意图，做到有效的多轮沟通，并且回答内容完整、重点清晰、有概括、有逻辑、有条理。ChatGPT 上线后，5 天活跃用户数高达 100 万，2 个月活跃用户数已达 1 个亿，成为历史上增长最快的消费者应用程序。除了被广大用户追捧外，ChatGPT 还受到了各国政府、企业界、学术界的广泛关注，使人们看到了解决自然语言处理这一认知智能核心问题的一条可能的路径，并被认为向通用人工智能迈出了坚实的一步，将对搜索引擎构成巨大的挑战，甚至将取代很多人的工作，更将颠覆很多领域和行业。

哈工大自然语言处理研究所组织多位老师和同学撰写了本调研报告，从技术原理、应用场景、未来发展等方面对 ChatGPT 进行了尽量详尽的介绍及总结。

本报告仅供内部参考。

主要编撰人员

第一章由车万翔、杨沐昀、张伟男、赵妍妍、冯骁骋、孙承杰、李佳朋编写；第二章由张伟男、隋典伯、高翠芸、朱庆福、李明达、王雪松编写；第三章由刘铭、朱聪慧、汤步洲编写；第四章由徐永东、高翠芸、朱庆福编写；第五章由杨沐昀、张伟男、韩一、庄子彧编写；第六章由隋典伯、高翠芸编写；第七章由车万翔、刘铭编写。参与各章审校工作的还有：崔一鸣、徐志明等。

报告整体由车万翔统稿。

目录

第一章　ChatGPT 的背景与意义

本章首先介绍自然语言处理、大规模预训练语言模型以及 ChatGPT 技术的发展历程，接着就 ChatGPT 的技术优点和不足进行分析，然后讨论 ChatGPT 可能的应用前景，最后展望 ChatGPT 普及后可能带来的风险与挑战。

1.1　自然语言处理的发展历史

人类语言（又称自然语言）具有无处不在的歧义性、高度的抽象性、近乎无穷的语义组合性和持续的进化性，理解语言往往需要具有一定的知识和推理等认知能力，这些都为计算机处理自然语言带来了巨大的挑战，使其成为机器难以逾越的鸿沟。因此，自然语言处理被认为是目前制约人工智能取得更大突破和更广泛应用的瓶颈之一，又被誉为 **“人工智能皇冠上的明珠”**。国务院 2017 年印发的《新一代人工智能发展规划》将知识计算与服务、跨媒体分析推理和自然语言处理作为新一代人工智能关键共性技术体系的重要组成部分。

自然语言处理自诞生起，经历了五次研究范式的转变（如图 1.1 所示）：由最开始基于小规模专家知识的方法，逐步转向基于机器学习的方法。机器学习方法也由早期基于浅层机器学习的模型变为了基于深度学习的模型。为了解决深度学习模型需要大量标注数据的问题，2018 年开始又全面转向基于大规模预训练语言模型的方法，其突出特点是充分利用**大模型**、**大数据**和**大计算**以求更好效果。

近期，ChatGPT 表现出了非常惊艳的语言理解、生成、知识推理能力，它可以极好地理解用户意图，真正做到多轮沟通，并且回答内容完整、重点清晰、有概括、有逻辑、有条理。ChatGPT 的成功表现，使人们看到了解决自然语言处理这一认知智能核心问题的一条可能的路径，并被认为向通用人工智能迈出了坚实的一步，将对搜索引擎构成巨大的挑战，甚至将取代很

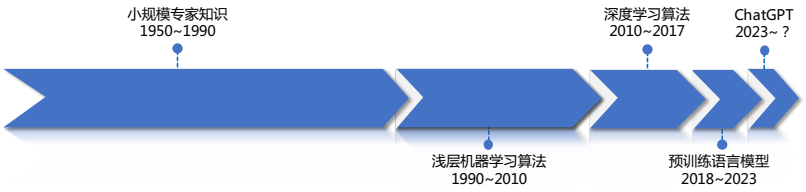

图 1.1: 自然语言处理研究范式的发展历程

多人的工作，更将颠覆很多领域和行业。

那么，ChatGPT 到底解决了什么本质科学问题，才能变得如此强大并受到广泛的关注呢？我们认为，**ChatGPT 是继数据库和搜索引擎之后的全新一代的“知识表示和调用方式”**。

知识在计算机内的表示是人工智能的核心问题。如表 1.1所示，早期，知识以结构化的方式存储在数据库中，人类需要掌握机器语言（如 SQL），才能调用这些知识；后来，随着互联网的诞生，更多文本、图片、视频等非结构化知识存储在互联网中，人类通过关键词的方式调用搜索引擎获取知识；现在，知识以参数的形式存储在大模型中（从 2018 年开始），ChatGPT 主要解决了用自然语言直接调用这些知识的问题，这也是人类获取知识最自然的方式。

表 1.1: 知识表示和调用方式的演进

知识表示方式	**表示方式的精确度**	**知识调用方式**	**调用方式的自然度**	**研究领域**	**代表应用**	**代表公司**
关系型数据库	高	SQL	低	数据库	DBMS	Oracle、Microsoft
互联网	中	Keywords	中	信息检索	搜索引擎	Google、Microsoft
大模型	低	自然语言	高	自然语言处理	ChatGPT	OpenAI、Microsoft、Google

另外，从自然语言处理技术发展阶段的角度看（如图 1.1 ），可以发现一个有趣的现象，即每一个技术阶段的发展时间，大概是上一个阶段的一半。小规模专家知识发展了 40 年，浅层机器学习是 20 年，之后深度学习大概 10 年，预训练语言模型发展的时间是 5 年，那么以 ChatGPT 为代表的技

术能持续多久呢? 如果大胆预测，可能是 2 到 3 年，也就是到 2025 年大概又要更新换代了。

1.2 大规模预训练语言模型的技术发展历程

大规模预训练语言模型（简称大模型）作为 ChatGPT 的知识表示及存储基础，对系统效果表现至关重要，接下来对大模型的技术发展历程加以简要介绍。

2018 年，OpenAI 提出了第一代 GPT（Generative Pretrained Transformer）模型[1]，将自然语言处理带入 “预训练” 时代。然而，GPT 模型并没有引起人们的关注，反倒是谷歌随即提出的 BERT（Bidirectional Encoder Representations from Transformers）模型[2]产生了更大的轰动。不过，OpenAI 继续沿着初代 GPT 的技术思路，陆续发布了 GPT-2[3] 和 GPT 模型 GPT-3[4]。

尤其是 GPT-3 模型，含有 1,750 亿超大规模参数，并且提出 “提示语”（Prompt）的概念，只要提供具体任务的提示语，即便不对模型进行调整也可完成该任务，如：输入 “我太喜欢 ChatGPT 了，这句话的情感是 ___”，那么 GPT-3 就能够直接输出结果 “褒义”。如果在输入中再给一个或几个示例，那么任务完成的效果会更好，这也被称为语境学习（In-context Learning）。更详细的技术细节推荐阅读相关的综述文章[5-8]。

不过，通过对 GPT-3 模型能力的仔细评估发现，大模型并不能真正克服深度学习模型鲁棒性差、可解释性弱、推理能力缺失的问题，在深层次语义理解和生成上与人类认知水平还相去甚远。直到 ChatGPT 的问世，才彻底改变了人们对于大模型的认知。

1.3 ChatGPT 技术发展历程

2022 年 11 月 30 日，OpenAI 推出全新的对话式通用人工智能工具——ChatGPT。据报道，在其推出短短几天内，注册用户超过 100 万，2 个月活跃用户数已达 1 个亿，引爆全网热议，成为历史上增长最快的消费者应用程序，掀起了人工智能领域的技术巨浪。

ChatGPT 之所以有这么多活跃用户，是因为它可以通过学习和理解人类语言，以对话的形式与人类进行交流，交互形式更为自然和精准，极大地改变了普通大众对于聊天机器人的认知，完成了从 “人工智障” 到 “有趣”

的印象转变。除了聊天，ChatGPT 还能够根据用户提出的要求，进行机器翻译、文案撰写、代码撰写等工作。ChatGPT 拉响了大模型构建的红色警报，学界和企业界纷纷迅速跟进启动研制自己的大模型。

继 OpenAI 推出 ChatGPT 后，与之合作密切的微软迅速上线了基于 ChatGPT 类技术的 New Bing，并计划将 ChatGPT 集成到 Office 办公套件中。谷歌也迅速行动推出了类似的 Bard 与之抗衡。除此之外，苹果、亚马逊、Meta（原 Facebook）等企业也均表示要积极布局 ChatGPT 类技术。国内也有多家企业和机构明确表态正在进行类 ChatGPT 模型研发。百度表示正在基于文心大模型进行文心一言的开发，阿里巴巴表示其类 ChatGPT 产品正在研发之中，华为、腾讯表示其在大模型领域均已有相关的布局，网易表示其已经投入到类 ChatGPT 技术在教育场景的落地研发，京东表示将推出产业版 ChatGPT，科大讯飞表示将在数月后进行产品级发布，国内高校复旦大学则推出了类 ChatGPT 的 MOSS 模型。

除了国内外学界和企业界在迅速跟进以外，我国国家层面也对 ChatGPT 有所关注。2023 年 2 月 24 日，科技部部长王志刚表示：“ChatGPT 在自然语言理解、自然语言处理等方面有进步的地方，同时在算法、数据、算力上进行了有效结合。”科技部高新技术司司长陈家昌在回应 ChatGPT 相关提问时也表示，ChatGPT 最近形成了一种现象级的应用，表现出很高的人机交互水平，表现出自然语言的大模型已经具备了面向通用人工智能的一些特征，在众多行业领域有着广泛的应用潜力。[1]

ChatGPT 是现象级应用，标志着语言大模型已经具备了一些通用人工智能特征,在众多行业领域有着广泛的应用潜力。”这标志着在未来,ChatGPT 相关技术有可能会成为国家战略支持的重点。

从技术角度讲，ChatGPT 是一个聚焦于对话生成的大语言模型，其能够根据用户的文本描述，结合历史对话，产生相应的智能回复。其中 GPT 是英文 Generative Pretrained Transformer 的缩写。GPT 通过学习大量网络已有文本数据（如 Wikipedia，reddit 对话），获得了像人类一样流畅对话的能力。虽然 GPT 可以生成流畅的回复，但是有时候生成的回复并不符合人类的预期，OpenAI 认为符合人类预期的回复应该具有真实性、无害性和有用性。为了使生成的回复具有以上特征，OpenAI 在 2022 年初发表的工作“Training language models to follow instructions with human feedback”中提到引入人工反馈机制，并使用近端策略梯度算法（PPO）对大模型进行

[1]https://www.sohu.com/a/645545405_120109837

训练。这种基于人工反馈的训练模式能够很大程度上减小大模型生成回复与人类回复之间的偏差，也使得 ChatGPT 具有良好的表现。

1.3.1 ChatGPT 的相关技术

接下来将简要介绍 ChatGPT 相关技术的发展历程。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 GPT-3.5 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。

RLHF 这一概念最早是在 2008 年 TAMER：Training an Agent Manually via Evaluative Reinforcement[9]一文中被提及的。在传统的强化学习框架下代理 (Agent) 提供动作给环境，环境输出奖励和状态给代理，而在 TAMER 框架下，引入人类标注人员作为系统的额外奖励。该文章中指出引入人类进行评价的主要目的是加快模型收敛速度，降低训练成本，优化收敛方向。具体实现上，人类标注人员扮演用户和代理进行对话，产生对话样本并对回复进行排名打分，将更好的结果反馈给模型，让模型从两种反馈模式——人类评价奖励和环境奖励中学习策略，对模型进行持续迭代式微调。这一框架的提出成为后续基于 RLHF 相关工作的理论基础。

[2]https://openai.com/blog/

在 2017 年前后，深度强化学习（Deep Reinforcement Learning）逐渐发展并流行起来。MacGlashan et al.[10]提出了一种 AC 算法（Actor-critic），并且将人工反馈（包括积极和消极）作为信号调节优势函数（Advantage function）。Warnell et al.[11]将 TAMER 框架与深度强化学习相结合，成功将 RLHF 引入深度强化学习领域。在这一阶段，RLHF 主要被应用于模拟器环境（例如游戏等）或者现实环境（例如机器人等）领域，而利用其对于语言模型进行训练并未受到重视。

在 2019 年以后, RLHF 与语言模型相结合的工作开始陆续出现, Ziegler et al.[12]较早利用人工信号在四个具体任务上进行了微调并取得不错的效果。OpenAI 从 2020 年开始关注这一方向并陆续发表了一系列相关工作，如应用于文本摘要[13-14]，利用 RLHF 训练一个可以进行网页导航的代理[15]等。后来，OpenAI 将 RLHF 与 GPT 相结合的工作，提出了 InstructGPT 这一 ChatGPT 的孪生兄弟[16]，主要是利用 GPT-3 进行对话生成，旨在改善模型生成的真实性、无害性和有用性。与此同时，作为缔造 AlphaGo 的公司，具有一干擅长强化学习的算法工程师的 DeepMind 也关注到了这一方向, 先后发表了 GopherCite[17]和 Sparrow[18]两个利用 RLHF 进行训练的语言模型，GopherCite 是在开放域问答领域的工作，Sparrow 是在对话领域的一篇工作，并且在 2022 年 9 月，DeepMind 的聊天机器人也已经上线。

2022 年 12 月，OpenAI 在诸多前人工作的积淀之下推出了 ChatGPT。ChatGPT 以 GPT-3.5 作为基座，依托其强大的生成能力，使用 RLHF 对其进行进一步训练，从而取得了惊艳四座的效果。

1.3.2 ChatGPT 技术发展脉络的总结

纵观 ChatGPT 的发展历程，不难发现其成功是循序渐进的，OpenAI 从 2020 年开始关注 RLHF 这一研究方向，并且开展了大量的研究工作，积攒了足够的强化学习在文本生成领域训练的经验。GPT 系列工作的研究则积累了海量的训练数据以及大语言模型训练经验，这两者的结合才产生了 ChatGPT。可以看出技术的发展并不是一蹴而就的，是大量工作的积淀量变引起质变。此外，将 RLHF 这一原本应用于模拟器环境和现实环境下的强化学习技术迁移到自然语言生成任务上是其技术突破的关键点之一。

纵观 AI 这几年的发展，已经逐渐呈现出不同技术相互融合的大趋势，比如将 Transformer 引入计算机视觉领域产生的 ViT；将强化学习引入蛋白质结构预测的 AlphaFold 等。每个研究人员都有自己熟悉擅长的领域，而同

时科学界也存在着大量需要 AI 赋能的亟待解决的关键问题，如何发现这些问题的痛点，设计合理的方法，利用自己研究领域的优越的技术解决问题，似乎是一个值得思考，也非常有意义的问题。

这是一个 AI 蓬勃发展的时代，计算机科学界每天都在产生着令人惊奇的发明创造，很多之前人们可望而不可及的问题都在或者正在被解决的路上。2022 年 2 月，DeepMind 发布可对托卡马克装置中等离子体进行磁控制的以帮助可控核聚变的人工智能，这项研究目前仍在进行。或许在未来的某一天，能源将不成为困扰我们的问题，环境污染将大大减少，星际远航将成为可能。希望每个研究人员都能在这样的时代中，找到适合自己的研究方向并且为科技进步添砖加瓦。

1.3.3 ChatGPT 的未来技术发展方向

虽然 ChatGPT 目前已经取得了非常喜人的成果，但是未来仍然有诸多可以研究的方向。

首先 OpenAI 的研究人员指出了 ChatGPT 现存的一些问题：

1. ChatGPT 有时候会生成一些似是而非、毫无意义的答案，导致这个问题的原因有：强化学习训练过程中没有明确的正确答案；训练过程中一些谨慎的训练策略导致模型无法产生本应产生的正确回复；监督学习训练过程中错误的引导导致模型更倾向于生成标注人员所知道的内容而不是模型真实知道的。

2. ChatGPT 对于输入措辞比较敏感，例如：给定一个特定的问题，模型声称不知道答案，但只要稍微改变措辞就可以生成正确答案。

3. ChatGPT 生成的回复通常过于冗长，并且存在过度使用某些短语的问题，例如：重申是由 OpenAI 训练的语言模型。这样的问题主要来自于训练数据的偏差和过拟合问题。

4. 虽然 OpenAI 已经努力让模型拒绝不恰当和有害的请求，但是仍然无法避免对有害请求作出回复或对问题表现出偏见。

其次，ChatGPT 虽然很强大，但是其模型过于庞大使用成本过高，如何对模型进行瘦身也是一个未来的发展方向，目前主流的模型压缩方法有量化、剪枝、蒸馏和稀疏化等。量化是指降低模型参数的数值表示精度，比如从 FP32 降低到 FP16 或者 INT8。剪枝是指合理地利用策略删除神经网络

中的部分参数，比如从单个权重到更高粒度组件如权重矩阵到通道，这种方法在视觉领域或其他较小语言模型中比较奏效。蒸馏是指利用一个较小的学生模型去学习较大的老师模型中的重要信息而摒弃一些冗余信息的方法。稀疏化将大量的冗余变量去除，简化模型的同时保留数据中最重要的信息。

此外，减少人类反馈信息的 RLAIF 也是最近被提出的一个全新的观点。2022 年 12 月 Anthropic 公司发表论文“Constitutional AI: Harmlessness from AI Feedback”[19]，该公司是 2020 年 OpenAI 副总裁离职后创立的，其公司始创团队中多有参与 GPT-3 以及 RLHF 相关研究的经历。该文章介绍了其最新推出的聊天机器人 Claude，与 ChatGPT 类似的是两者均利用强化学习对模型进行训练，而不同点则在于其排序过程使用模型进行数据标注而非人类，即训练一个模型学习人类对于无害性偏好的打分模式并代替人类对结果进行排序。

1.4 ChatGPT 的优势与劣势

1.4.1 ChatGPT 的优势

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，这得益于其以下三方面的能力：

1. 强大的底座能力：ChatGPT 基于 GPT-3.5 系列的 Code-davinci-002 指令微调而成。而 GPT-3.5 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；
2. 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 scaling law；

3. 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，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。

1.4.2 ChatGPT 的劣势

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：

1. 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；
2. 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；
3. 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；
4. 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；
5. 语言模型每次的生成结果是 beam search 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，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，在部分情况下显得啰嗦冗长。

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：

1. 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；
2. 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；
3. 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；
4. 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；
5. 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；
6. ChatGPT 目前还只能处理文本数据，在多模态任务上还无法处理。

表 1.2 列举了一些 ChatGPT 存在的以上不足的示例（2023 年 2 月 24 日测试）。

1.5 ChatGPT 的应用前景

ChatGPT 作为掀起新一轮 AIGC 热潮的新引擎，无论在人工智能行业还是其他行业都带来了广泛的讨论和影响，下面我们分别从这两个方面讨论

ChatGPT 的应用前景。

1.5.1 在人工智能行业的应用前景及影响

ChatGPT 的发布及其取得的巨大成功对人工智能行业形成了强烈的冲击，人们发现之前许多悬而未解的问题在 ChatGPT 身上迎刃而解（包括事实型问答、文本摘要事实一致性、篇章级机器翻译的性别问题等），ChatGPT 引起了巨大的恐慌。然而从另一个角度看，我们也可以把 ChatGPT 当成是一个工具来帮助我们的开发、优化我们的模型、丰富我们的应用场景，比如：

1. **代码开发**：利用 ChatGPT 辅助开发代码，提高开发效率，包括代码补全、自然语言指令生成代码、代码翻译、bug 修复等；

2. **ChatGPT 和具体任务相结合**：ChatGPT 的生成结果在许多任务上相比微调小模型都有很明显的可取之处（比如文本摘要的事实一致性，篇章级机器翻译的性别问题），在微调小模型的基础上结合这些 ChatGPT 的长处，可能可以在避免训练部署下显著提升小模型的效果；

3. 同时基于 ChatGPT 指令微调激发的零样本能力，对于只有少数标注或者没有标注数据的任务以及需要分布外泛化的任务，我们既可以直接应用 ChatGPT，也可以把 ChatGPT 当作冷启动收集相关语料的工具，丰富我们的应用场景。

1.5.2 在其他行业的应用前景及影响

ChatGPT 的发布也引起了其它行业的连锁反应：Stack Overflow 禁用 ChatGPT 的生成内容，美国多所公立学校禁用 ChatGPT，各大期刊禁止将 ChatGPT 列为合著者。ChatGPT 似乎在一些行业成为“公敌”，但在其它行业，也许充满机遇。

1. **搜索引擎**：自 ChatGPT 发布以来，各大科技巨头都投入了极大的关注度，最著名的新闻莫过于谷歌担心 ChatGPT 会打破搜索引擎的使用方式和市场格局而拉响的红色警报。为此各大科技巨头纷纷行动起来，谷歌开始内测自己的类 ChatGPT 产品 Bard，百度三月份将面向公众开放文心一言，微软更是宣布 ChatGPT 为必应提供技术支持，推出新必应。ChatGPT 和搜索引擎的结合似乎已经不可避免，也许不会

马上取代搜索引擎，但基于搜索引擎为 ChatGPT 提供生成结果证据展示以及利用检索的新知识扩展 ChatGPT 的回答边界已经是可以预见并正在进行的结合方向。

2. **泛娱乐行业**：ChatGPT 对于文娱行业则更多带来的是机遇。无论是基于 ChatGPT 创建更智能的游戏虚拟人和玩家交流提升体验，还是利用虚拟数字人进行虚拟主播直播互动，ChatGPT 都为类似的数字人提供了更智能的“大脑”，使行业充满想象空间。除此之外，在心理健康抚慰、闲聊家庭陪护等方面，类似的数字人也大有拳脚可展。

3. **自媒体行业**：同样大大受益的还有自媒体行业。美国的新闻聚合网站 BuzzFeed 宣布和 OpenAI 合作，未来将使用 ChatGPT 帮助创作内容。ChatGPT 的出现将使得内容创作变得更加容易，无论是旅游、餐饮、住宿、情感，相关博主的内容产出效率将得到极大的提升，有更多的精力润色相关内容，期待更多的高质量文章的产生。

4. **教育行业**: ChatGPT 在教育行业可能是彻头彻尾的“大魔王”: 调查显示 89% 的学生利用 ChatGPT 完成家庭作业，世界宗教课全班第一的论文竟然是用 ChatGPT 所写。这迫使多所学校全面禁用 ChatGPT，无论是在作业、考试或者论文当中，一经发现即认定为作弊。然而从另一方面来看，这可能也会促使针对人工智能相关法律法规的完善，加速 AI 社会化的发展。

5. **其他专业领域**：针对其它专业领域，ChatGPT 的具体影响不大。因为限于 ChatGPT 训练数据的限制，ChatGPT 无法对专业领域的专业知识进行细致的分析，生成的回答专业度不足且可信性难以保证，至多只能作为参考，很难实现替代。比如因为 ChatGPT 未获取 IDC、Gartner 等机构的数据使用授权，其关于半导体产业的市场分析中很少涉及量化的数据信息。

此外，ChatGPT 可以帮助个人使用者在日常工作中写邮件、演讲稿、文案和报告，提高其工作效率。同时基于微软计划将 ChatGPT 整合进 Word、PowerPoint 等办公软件，个人使用者也可以从中受益，提高办公效率。

1.6 ChatGPT 带来的风险与挑战

ChatGPT 的出现和应用给用户和社会带来了很多新的风险和挑战。这些风险和挑战，一部分是 ChatGPT 本身技术限制引起的，如生成的内容不能保证真实性、会产生有害言论等。一部分是用户对 ChatGPT 的使用不当引起的，如在教育、科研等领域滥用 ChatGPT 产生的文本。ChatGPT 用户数量在其出现后两个月就突破了 1 亿，因此应对这些风险和挑战需要整个社会行动起来，制定相应的法律和规范，让 ChatGPT 为人类发展服务，尽量避免引起新的的社会问题。下面列举了几个重要风险和挑战，并试着给出了相应的解决思路。

滥用风险 滥用风险主要是指用户对于 ChatGPT 产生结果的不当应用。具体表现有：学生在课堂测验或考试过程中直接使用 ChatGPT 的结果作为答案进行作弊；研究人员使用 ChatGPT 来进行写作的学术不规范行为；不法分子利用 ChatGPT 来制造假新闻或谣言。Tamkin et al.[20]指出，使用预训练语言模型能参与的犯罪行为种类繁多，因此很难把所有它们能错误使用的方法都归纳总结起来，可以预料随着技术的发展以及不法分子的不断尝试，ChatGPT 被错误使用的方式会更多样且更加难以预测。

已有很多研究者针对这一需求提出了不同的解决方案。下面主要介绍两个有代表性的工作：

2023 年 1 月 31 日，开发 ChatGPT 的 OpenAI 公司发布了一个能够鉴别 AI 生成文本的分类器[3]。根据 OpenAI 公布的测试结果，该分类器对于“AI 生成文本”类别的召回率只有 26%。该分类器的训练数据的构造方式如下：首先获取大量提示，对于每个提示，分别获取 AI 生成文本和人工写作文本。这种训练数据的获取方式成本较高。

斯坦福大学的 Mitchell et al.[21]提出了一种 Zero-shot 的 AI 生成文本检测方法 DetectGPT，该方法利用 AI 生成文本和人工写作文本在由其他 AI 模型进行改写后所引起的生成概率的变化来进行判别，生成概率变化大的文本为 AI 生成文本。根据论文在 3 个数据集上的测试结果，DetectGPT 在 AUROC 这一评价指标上超过了目前已知的其他 Zero-shot 方法。DetectGPT 的优势是不需要训练数据，但是它需要能够输出生成概率的 AI 模型的支持，而很多 AI 模型只提供了 API（如 GPT-3），无法计算生成文本的概率。

[3]https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/

总的来说，目前对于 ChatGPT 自动生成文本的自动鉴别技术效果还不能令人满意，需要继续寻找更有效的鉴别方法。

错误信息风险 错误信息风险源于 ChatGPT 可能产生虚假、误导、无意义或质量差的信息。ChatGPT 可以并且已经在成为很多用户的一种获取信息的手段，但用户如果没有分辨能力，可能会采信这些错误信息，从而带来风险隐患。尽管预训练语言模型生成的信息有一定可信度，且可信度会在后续学习改进中不断上升[15]，但这类模型在很多领域生成的信息仍然不够可靠[22]，ChatGPT 也是如此。ChatGPT 的流行会在某种程度上增加用户对它的信任，从而被更多错误的信息误导。预训练语言模型的生成的错误信息比例上升可能会加大人们对社会中各类信息的不信任，破坏社会的知识交流传播[23]。

在一些很敏感的领域, 比如法律和医学, ChatGPT 的错误信息很容易导致直接伤害。错误的医学法律知识会导致使用者违法犯罪或者自行处理伤口疾病时出现问题，从而造成对社会和自己身体健康的伤害。这在 ChatGPT 之前就已经有了一些例子，如患者不相信正规医生而搬出搜索引擎给出的结果来反驳医生，这也能体现出很多用户对这类信息获取方式的信任。

知识共享是一种社会现象，人们出于信任从社会中获取知识并且过滤吸收。ChatGPT 的一个较为常用的功能是充当搜索引擎，类似百度、Google 等，搜索引擎的信息因其较高的准确率通常拥有较高的可信度，但是如果 ChatGPT 产生错误信息误导他人的现象加剧可能会导致人们不仅对 ChatGPT 信任感下降，同时也对其他类别的信息不再信任，破坏社会的知识共享，影响社会的知识交流传播。

目前还没有专门针对 ChatGPT 生成文本的正确性进行鉴别的研究论文发表。已有的针对虚假新闻或虚假信息检测的方法可以尝试应用到大规模语言模型生成文本的正确性检测中，比如基于事实抽取和验证的方法。但是基于写作风格的方法可能不太实用，因为大规模语言模型生成文本的过程与人的写作过程有较大区别。

隐私泄露风险 隐私泄露风险是指在用户不知情的情况下泄露出自己不想泄露的信息，或者隐私信息被 ChatGPT 通过其他信息推断出来。用户在使用 ChatGPT 过程中可能会泄露自己的个人隐私信息或者一些组织乃至国家的机密信息。个人信息的泄露可能会对个人的心理健康、人身安全造成影响。国家或者商业机密往往是只有小范围人员能获悉的高等级信息，它们的