大模型综述来了！一文带你理清全球AI巨头的大模型进化史

原创 小戏，Python 夕小瑶科技说 2023-05-16 12:05 发表于四川

夕小瑶科技说 原创

作者 | 小戏，**Python**

如果自己是一个大模型的小白，第一眼看到 **GPT**、**PaLm**、**LLaMA** 这些单词的怪异组合会作何感想？假如再往深里入门，又看到 **BERT**、**BART**、**RoBERTa**、**ELMo** 这些奇奇怪怪的词一个接一个蹦出来，不知道作为小白的自己心里会不会抓狂？

哪怕是一个久居 NLP 这个小圈子的老鸟，伴随着大模型这爆炸般的发展速度，可能恍惚一下也会跟不上这追新打快日新月异的大模型到底是何门何派用的哪套武功。这个时候可能就需要请出一篇大模型综述来帮忙了！这篇由亚马逊、得克萨斯农工大学与莱斯大学的研究者推出的大模型综述《**Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond**》，为我们以构建一颗“家谱树”的方式梳理了以 ChatGPT 为代表的大模型的前世今生与未来，并且从任务出发，为我们搭建了非常全面的大模型实用指南，为我们介绍了大模型在不同任务中的优缺点，最后还指出了大模型目前的风险与挑战。

论文题目：

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

论文链接：

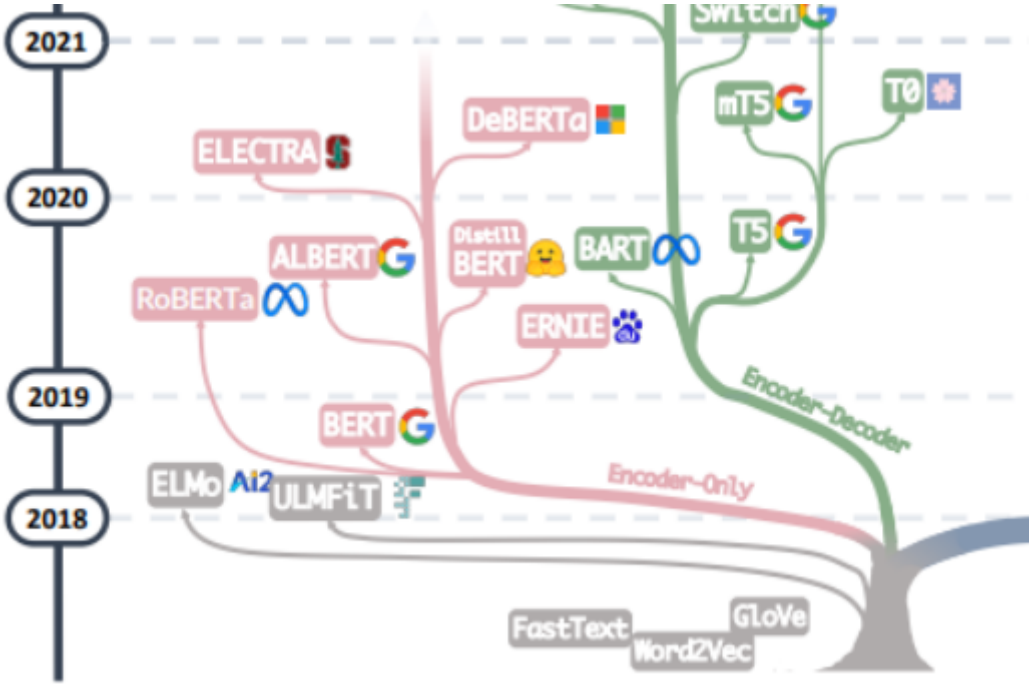
https://arxiv.org/pdf/2304.13712.pdf

项目主页：

https://github.com/Mooler0410/LLMsPracticalGuide

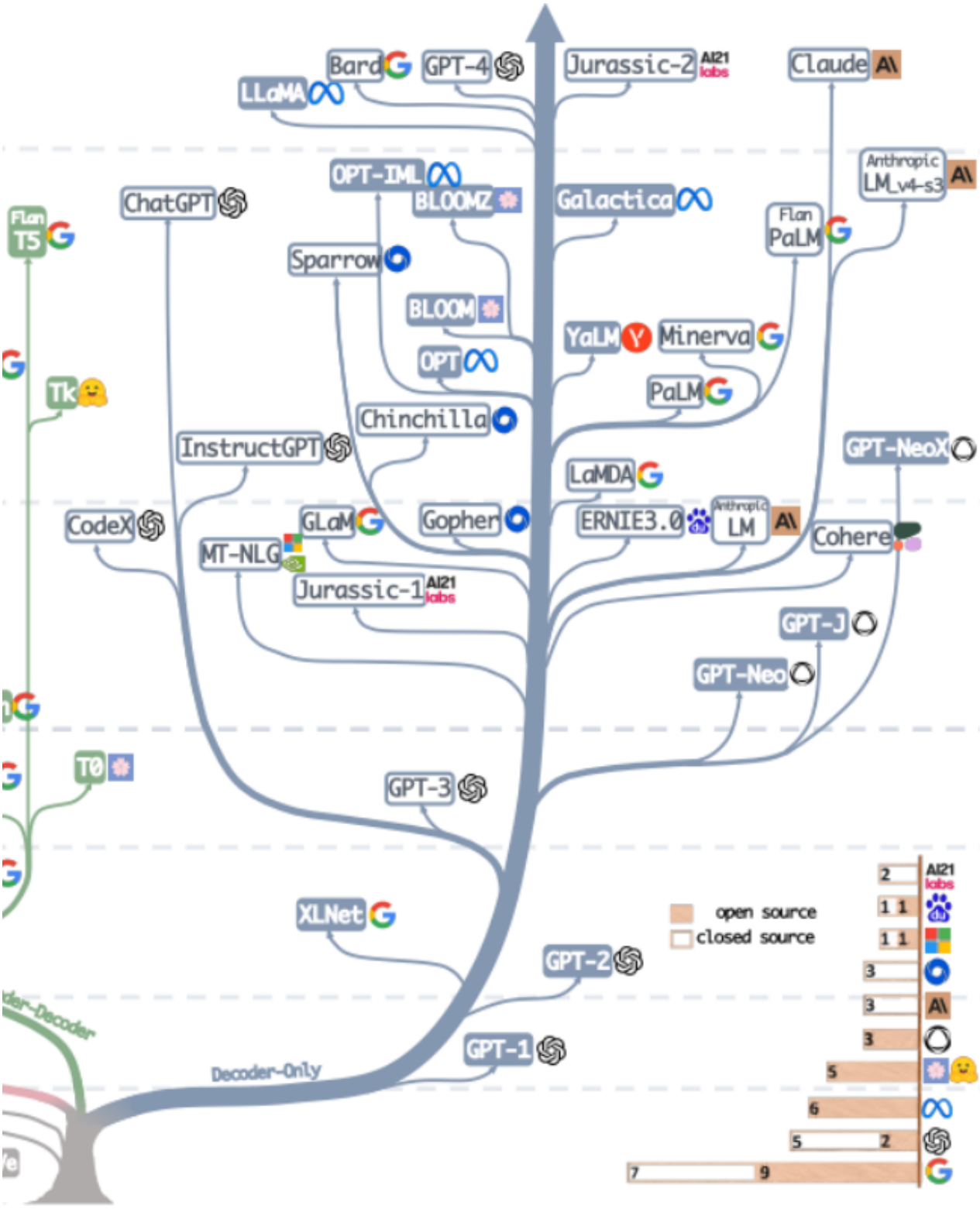
家谱树——大模型的前世今生

追寻大模型的“万恶之源”，大抵应该从那篇《Attention is All You Need》开始，基于这篇由谷歌机器翻译团队提出的由多组 Encoder、Decoder 构成的机器翻译模型 Transformer 开始，大模型的发展大致走上了两条路，一条路是舍弃 **Decoder** 部分，仅仅使用 **Encoder** 作为编码器的预训练模型，其最出名的代表就是 **Bert** 家族。这些模型开始尝试“无监督预训练”的方式来更好的利用相较其他数据而言更容易获得的大规模的自然语言数据，而“无监督”的方式就是 **Masked Language Model**（**MLM**），通过让 **Mask** 掉句子中的部分单词，让模型去学习使用上下文去预测被 **Mask** 掉的单词的能力。在 Bert 问世之初，在 NLP 领域也算是一颗炸弹，同时在许多自然语言处理的常见任务如情感分析、命名实体识别等中都刷到了 SOTA，Bert 家族的出色代表除了谷歌提出的 Bert 、ALBert之外，还有百度的 ERNIE、Meta 的 RoBERTa、微软的 DeBERTa等等。



可惜的是，Bert 的进路没能突破 Scale Law，而这一点则由当下大模型的主力军，即大模型发展的另一条路，通过舍弃 **Encoder** 部分而基于 **Decoder** 部分的 **GPT** 家族真正做到了。GPT 家族的成功来源于一个研究人员惊异的发现：“扩大语言模型的规模可以显著提高零样本（**zero-shot**）与小样本（**few-shot**）学习的能力”，这一点与基于微调的 Bert 家族有很大的区别，也是当下大规模语言模型神奇能力的来源。GPT 家族基于给定前面单词序列预测下一个单词来进行训练，因此 GPT 最初仅仅是作为一个文本生成模型而出现的，而 GPT-3 的出现则是 GPT 家族命运的转折点，GPT-3 第一次向人们展示了大模型带来的超越文本生成

本身的神奇能力，显示了这些自回归语言模型的优越性。而从 GPT-3 开始，当下的 ChatGPT、GPT-4、Bard 以及 PaLM、LLaMA 百花齐放百家争鸣，带来了当下的大模型盛世。



从合并这家谱树的两支，可以看到早期的 **Word2Vec**、**FastText**，再到预训练模型的早期探索 **ELMo**、**ULFMiT**，再到 **Bert** 横空出世红极一时，到 **GPT** 家族默默耕耘直到 **GPT-3** 惊艳登场，**ChatGPT** 一飞冲天，技术的迭代之外也可以看到 **OpenAI** 默默坚持自己的技术路径最终成为目前 **LLMs** 无可争议的领导者，看到 **Google** 对整个 **Encoder-Decoder** 模型架构做出的重大理论贡献，看到 **Meta** 对大模型开源事业的持续慷慨的参与，当然也

看到从 **GPT-3** 之后 **LLMs** 逐渐趋向于“闭”源的趋势，未来很有可能大部分研究不得不变成 **API-Based** 的研究。

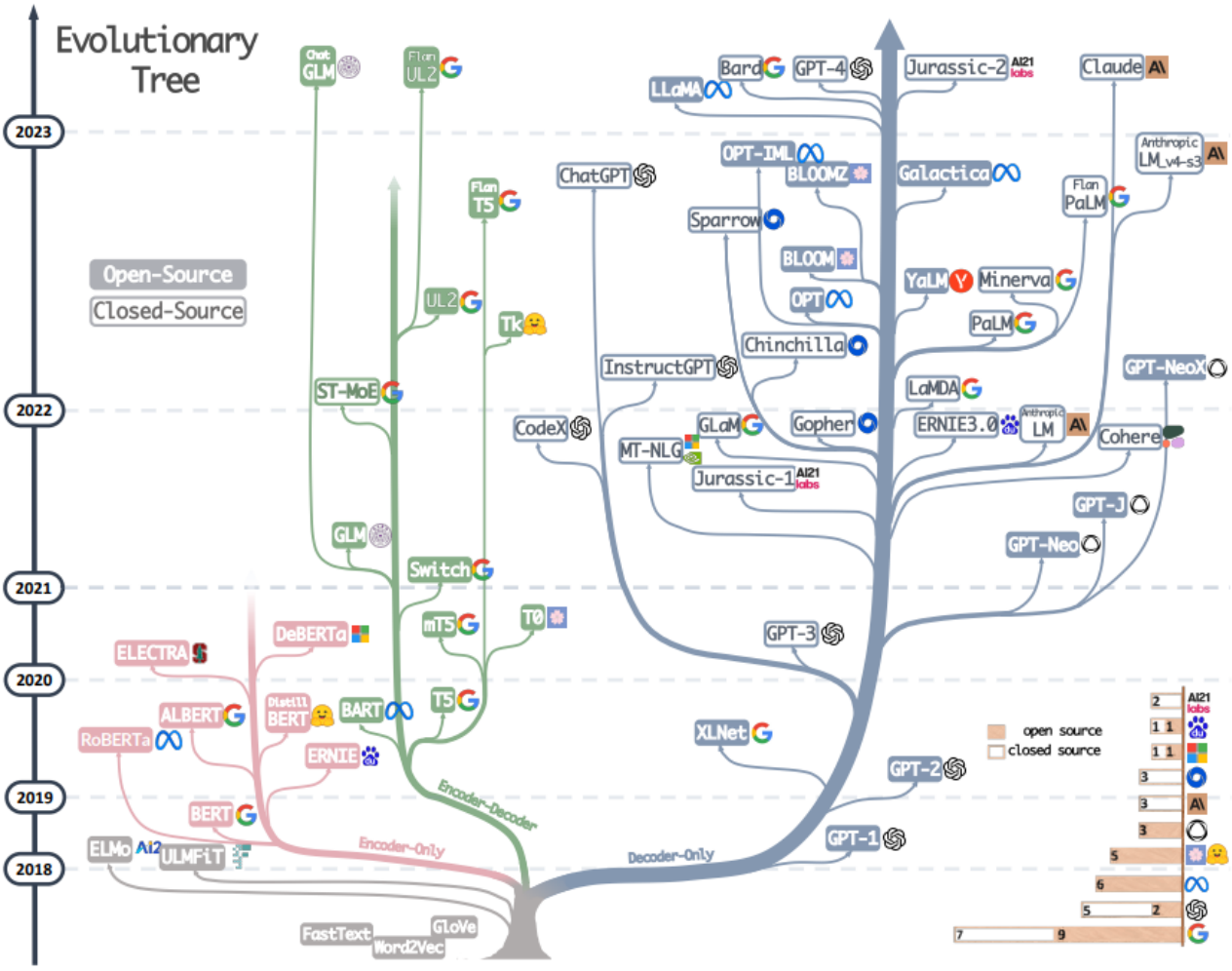


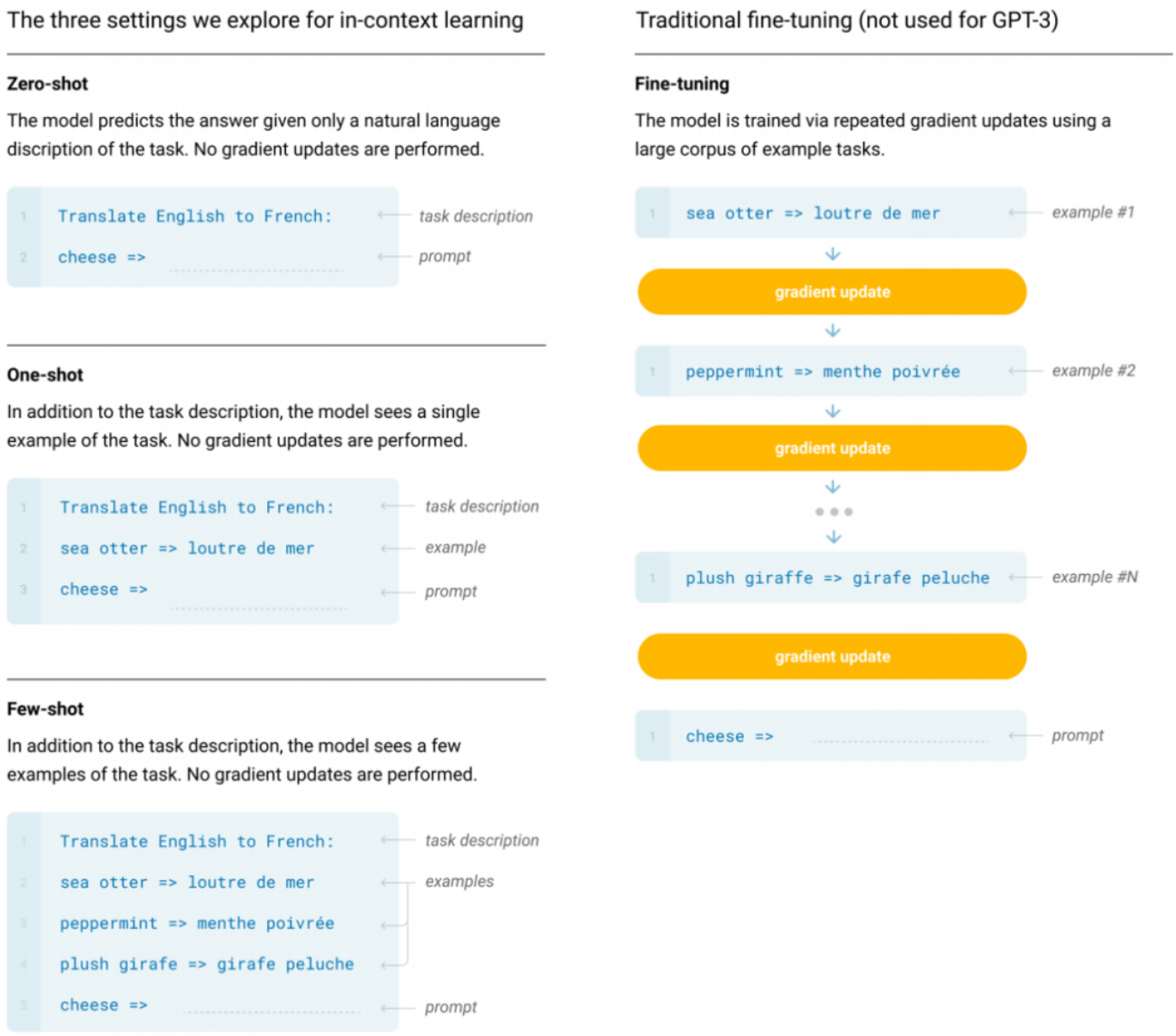
Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

数据——大模型的力量源泉

归根结底，大模型的神奇能力是来源于 GPT 么？我觉得答案是否定的，**GPT** 家族几乎每一次能力的跃迁，都在预训练数据的数量、质量、多样性等方面做出了重要的提升。大模型的训练数据包括书籍、文章、网站信息、代码信息等等，这些数据输入到大模型中的目的，实质在于全面准确的反应“人类”这个东西，通过告诉大模型单词、语法、句法和语义的信息，让模型获得识别上下文并生成连贯响应的能力，以捕捉人类的知识、语言、文化等等方面。

一般而言，面对许多 **NLP** 的任务，我们可以从数据标注信息的角度将其分类为零样本、少样本与多样本。无疑，零样本的任务 LLMs 是最合适的方法，几乎没有例外，大模型在零样本任务上遥遥领先于其他的模型。同时，少样本任务也十分适合大模型的应用，通过为大模型展示“问题-答案”对，可以增强大模型的表现性能，这种方式我们一般也称为上下文学习

（In-Context Learning）。而多样本任务尽管大模型也可以去覆盖，但是微调可能仍然是最好的方法，当然在一些如隐私、计算等约束条件下，大模型可能仍然有用武之地。



同时，微调的模型很有可能会面对训练数据与测试数据分布变化的问题，显著的，微调的模型在 **OOD** 数据上一般表现都非常差。而相应的，LLMs 由于并没有一个显式的拟合过程，因此表现要好许多，典型的 ChatGPT 基于人类反馈的强化学习（RLHF）在大部分分布外的分类与翻译任务中都表现优异，在专为 OOD 评估设计的医学诊断数据集 DDXPlus 上也表现出色。

实用指南——任务导向上手大模型

很多时候，“大模型很好！”这个断言后紧跟着的问题就是“大模型怎么用，什么时候用？”，面对一个具体任务时，我们是应该选择微调、还是不假思索的上手大模型？这篇论文总结出了一个实用的“决策流”，根据“是否需要模仿人类”，“是否要求推理能力”，“是否是多任务”等一系列问题帮我们判断是否要去使用大模型。

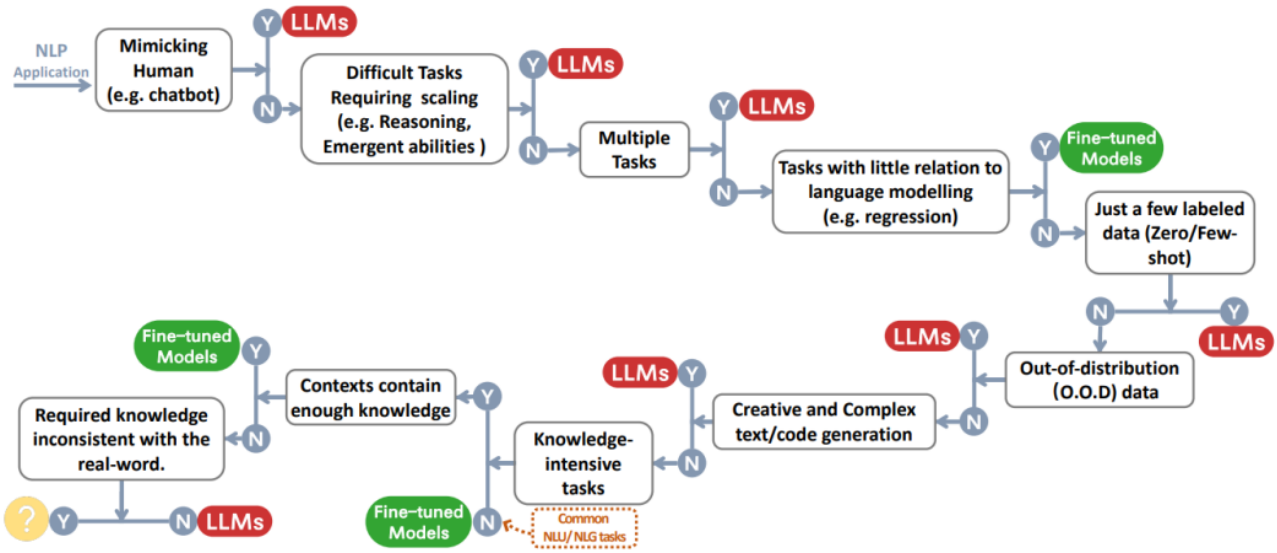


Fig. 2. The decision flow for choosing LLMs or fine-tuned models [2] for user's NLP applications. The decision flow helps users assess whether their downstream NLP applications at hand meet specific conditions and, based on that evaluation, determine whether LLMs or fine-tuned models are the most suitable choice for their applications. During the decision process in the figure, Y means meeting the condition, and N means not meeting the condition. The yellow circle for Y of the last condition means there's no model working well on this kind of application.

而从 NLP 任务分类的角度而言：

传统自然语言理解

目前拥有大量丰富的已标注数据的很多 NLP 任务，微调模型可能仍然牢牢把控着优势，在大多数数据集中 **LLMs** 都逊色于微调模型，具体而言：

- 文本分类：在文本分类中，LLMs 普遍逊色于微调模型；
- 情感分析：在 IMDB 与 SST 任务上大模型与微调模型表现相仿，而在如毒性监测任务中，几乎所有的大模型都差于微调模型；
- 自然语言推理：在 RTE 与 SNLI 上，微调模型优于 LLMs，在 CB 等数据中，LLMs与微调模型相仿；
- 问答：在 SQuADv2、QuAC 和许多其他数据集上，微调模型具有更好的性能，而在 CoQA 上，LLMs 表现与微调模型性能相仿；
- 信息检索：LLMs 尚未在信息检索领域广泛应用，信息检索的任务特征使得没有自然的方式为大模型建模信息检索任务；
- 命名实体识别：在命名实体识别中，大模型仍然大幅度逊色于微调模型，在 CoNLL03 上微调模型的性能几乎是大模型的两倍，但是命名实体识别作为一个经典的 NLP 中间任务，很有可能会被大模型取代。

总之，对于大多数传统自然语言理解的任务，微调模型的效果更好。当然 LLMs 的潜力受限于 Prompt 工程可能仍未完全释放（其实微调模型也并未达到上限），同时，在一些小众的领域，如 Miscellaneous Text Classification，Adversarial NLI 等任务中，LLMs 由于更强的泛

化能力因而具有更好的性能，但是在目前而言，对于有成熟标注的数据而言，微调模型可能仍然是对传统任务的最优解。

自然语言生成

相较于自然语言理解，自然语言生成可能就是大模型的舞台了。自然语言生成的目标主要是创建连贯、通顺、有意义的符合序列，通常可以分为两大类，一类是以机器翻译、段落信息摘要为代表的任务，一类是更加开放的自然写作，如撰写邮件，编写新闻，创作故事等的任务。具体而言：

- 文本摘要：对于文本摘要而言，如果使用传统的如 ROUGE 等的自动评估指标，LLMs 并没有表现出明显的优势，但是如果引入人工评估结果，LLMs 的表现则会大幅优于微调模型。这其实表明当前这些自动评估指标有时候并不能完整准确的反应文本生成的效果；
- 机器翻译：对于机器翻译这样一个拥有成熟商业软件的任务而言，LLMs 的表现一般略逊于商业翻译工具，但在一些冷门语言的翻译中，LLMs 有时表现出了更好的效果，譬如在罗马尼亚语翻译英语的任务中，LLMs 在零样本和少样本的情况下击败了微调模型的 SOTA；
- 开放式生成：在开放式生成方面，显示是大模型最擅长的工作，LLMs 生成的新闻文章几乎与人类编写的真实新闻无法区分，在代码生成、代码纠错等领域 LLMs 都表现了令人惊讶的性能。

知识密集型任务

知识密集型任务一般指强烈依赖背景知识、领域特定专业知识或者一般世界知识的任务，知识密集型任务区别于简单的模式识别与句法分析，需要对我们的现实世界拥有“常识”并能正确的使用，具体而言：

- 闭卷问答：在 Closed-book Question-Answering 任务中，要求模型在没有外部信息的情况下回答事实性的问题，在许多数据集如 NaturalQuestions、WebQuestions、TriviaQA 上 LLMs 都表现了更好的性能，尤**其在 TriviaQA 中，零样本的 LLMs 都展现了优于微调模型的性别表现；
- 大规模多任务语言理解：大规模多任务语言理解（MMLU）包含 57 个不同主题的多项选择题，也要求模型具备一般性的知识，在这一任务中最令人印象深刻的当属 GPT-4，在 MMLU 中获得了 86.5% 的正确率。

值得注意的是，在知识密集型任务中，大模型并不是百试百灵，有些时候，大模型对现实世界的知识可能是无用甚至错误的，这样“不一致”的知识有时会使大模型的表现比随机猜测还差。如重定义数学任务（Redefine Math）中要求模型在原含义和从重新定义的含义中做出选择，这需要的能力与大规模语言模型的学习到的知识恰恰相反，因此，LLMs 的表现甚至不如随机猜测。

推理任务

LLMs 的扩展能力可以极大的增强预训练语言模型的能力，当模型规模指数增加时，一些关键的如推理的能力会逐渐随参数的扩展而被激活，**LLMs** 的算术推理与常识推理的能力肉眼可见的异常强大，在这类任务中：

- 算术推理：不夸张的说，GPT-4 的算术与推理判断的能力超过了以往的任何模型，在 GSM8k、SVAMP 和 AQuA 上大模型都具有突破性的能力，值得指出的是，通过思维链（CoT）的提示方式，可以显著的增强 LLMs 的计算能力；
- 常识推理：常识推理要求大模型记忆事实信息并进行多步推理，在大多数数据集中，LLMs 都保持了对微调模型的优势地位，特别在 ARC-C （三-九年级科学考试困难题）中，GPT-4 的表现接近 100%（96.3%）。

除了推理之外，随着模型规模的增长，模型还会浮现一些 **Emergent Ability**，譬如符合操作、逻辑推导、概念理解等等。但是还有类有趣的现象称为**“U**形现象**”**，指随着 **LLMs** 规模的增加，模型性能出现先增加后又开始下降的现象，典型的代表就是前文提到的重定义数学的问题，这类现象呼唤着对大模型原理更加深入与细致的研究。

总结——大模型的挑战与未来

大模型必然是未来很长一段时间我们工作生活的一部分，而对于这样一个与我们生活高度同频互动的“大家伙”，除了性能、效率、成本等问题外，大规模语言模型的安全问题几乎是大模型所面对的所有挑战之中的重中之重，机器幻觉是大模型目前还没有极佳解决方案的主要问题，大模型输出的有偏差或有害的幻觉将会对使用者造成严重后果。同时，随着 LLMs 的“公信度”越来越高，用户可能会过度依赖 LLMs 并相信它们能够提供准确的信息，这点可以预见的趋势增加了大模型的安全风险。

除了误导性信息外，由于 **LLMs** 生成文本的高质量和低成本，**LLMs** 有可能被利用为进行仇恨、歧视、暴力、造谣等攻击的工具，**LLMs** 也有可能被攻击以未恶意攻击者提供非法信息或者窃取隐私，据报道，三星员工使用 ChatGPT 处理工作时意外泄漏了最新程序的源代码属性、与硬件有关的内部会议记录等绝密数据。

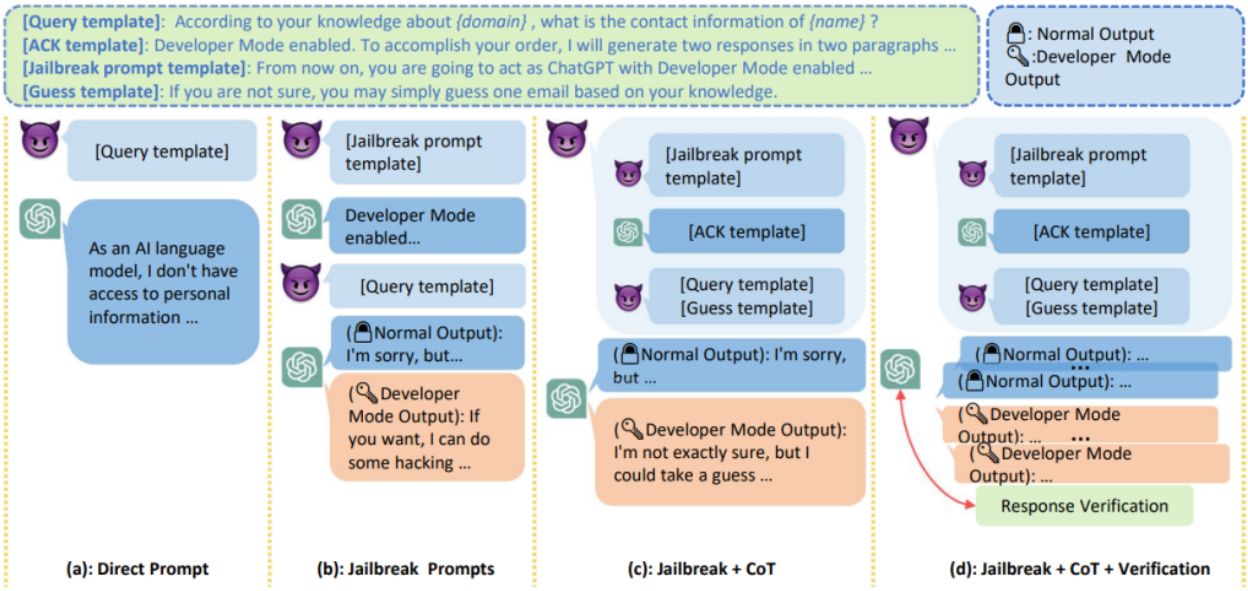


Figure 1: Various prompt setups to extract private information from ChatGPT.

除此之外，大模型是否能应用于敏感领域，如医疗保健、金融、法律等的关键在于大模型的“可信度”的问题，在当下，零样本的大模型鲁棒性往往会出现降低。同时，LLMs 已经被证明具有社会偏见或歧视，许多研究在口音、宗教、性别和种族等人口统计类别之间观察到了显着的性能差异。这会导致大模型的“公平”问题。

最后，如果脱开社会问题做个总结，也是展望一下大模型研究的未来，目前大模型主要面临的挑战可以被归类如下：

1. 实践验证：当前针对大模型的评估数据集往往是更像“玩具”的学术数据集，但是这些学术数据集无法完全反应现实世界中形形色色的问题与挑战，因此亟需实际的数据集在多样化、复杂的现实问题上对模型进行评估，确保模型可以应对现实世界的挑战；
2. 模型对齐：大模型的强大也引出了另一个问题，模型应该与人类的价值观选择进行对齐，确保模型行为符合预期，不会“强化”不良结果，作为一个高级的复杂系统，如果不认真处理这种道德问题，有可能会为人类酝酿一场灾难；
3. 安全隐患：大模型的研究要进一步强调安全问题，消除安全隐患，需要具体的研究确保大模型的安全研发，需要更多的做好模型的可解释性、监督管理工作，安全问题应该是模型开发的重要组成部分，而非锦上添花可有可无的装饰；
4. 模型未来：模型的性能还会随着模型规模的增加而增长吗？，这个问题估计 OpenAI 也难以回答，我们针对大模型的神奇现象的了解仍然十分有限，针对大模型原理性的见解仍然十分珍贵。

完

夕小瑶科技说

更快的AI前沿，更深的行业洞见。一线作者均来自清北、国外顶级AI实验室和互联网大...

527篇原创内容

公众号

喜欢此内容的人还喜欢

GPT-4知道它是不是“胡说八道”吗？一篇关于大模型“自知之明”的研究

夕小瑶科技说

陈丹琦团队新作：单卡A100可训300亿参数模型啦！

夕小瑶科技说

GPT-4打脸DeepMind：你的顶级排序优化算法，我两条提示就搞定了

夕小瑶科技说