Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

评论：

ChatGPT：潜力、前景和局限

周杰[1,3]，柯沛[2]，邱锡鹏[1,3]，黄民烈[2]，张军平[‡1,3]

[1]复旦大学计算机科学技术学院，中国上海市，200433

[2]清华大学计算机科学与技术系，中国北京市，100084

[3]上海市智能信息处理重点实验室，中国上海市，200433

E-mail: {jie_zhou, xpqiu, jpzhang}@fudan.edu.cn; {kepei, aihuang}@tsinghua.edu.cn

本文编译自 Zhou J, Ke P, Qiu XP, et al., 2023. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, early access. https://doi.org/10.1631/FITEE.2300089

1 绪论

最近，OpenAI 发布了对话生成预训练模型 Transformer（Chat Generative Pre-trained Transformer，ChatGPT）（Schulman et al., 2022）(https://chat.openai.com)，其展现的能力令人印象深刻，吸引了工业界和学术界的广泛关注。这是首次在大型语言模型（large language model, LLM）内很好地解决如此多样的开放任务。为更好地理解 ChatGPT，这里我们简要介绍其历史，讨论其优点和不足，指出几个潜在应用，最后分析它对可信赖人工智能、会话搜索引擎和通用人工智能（artificial general intelligence, AGI）发展的影响。

ChatGPT 成为历史上增长最快的消费者应用程序，在发布后两个月内，吸引了 1 亿月度活跃访客（Hu，2023）。自发布以来，因其高超的对话能力，已引爆社会关注。它可以回答后续提问，拒绝不当请求，挑战错误前提，并承认自己错误（Schulman et al., 2022）。它获得许多涌现能力，如高质量对话、复杂推理、思维链（CoT）（Wei et al., 2022b）、零/少样本学习（语境学习）、跨任务泛化、代码理解/生成等等。

这些令人印象深刻的能力，ChatGPT 是如何获得的？其主要得益于大型语言模型，它利用语言模型（LM）在大规模数据上训练巨大的神经网络模型，如 Transformer（Vaswani et al., 2017）。语言模型旨在根据上文预测下一个词的概率，是文本中的自监督信号。互联网上存在大规模文本数据，所以通过语言模型对模型进行预训练是顺理成章的。现有研究表明，模型规模和数据量越大，性能越好。当模型和数据规模达到一定程度时，模型将获得涌现能力。不幸的是，训练一个大型语言模型费时又费力。例如，OpenAI 发布的 GPT-3（Brownet al., 2020）有 1750 亿个参数。它的预训练采用超级计算机（285 000 个 CPU，10 000 个 GPU）在 45 TB 文本数据上完成，训练费用高达 1200 万美元。它在零样本学习任务上实现了巨大性能提升，具有小模型所不具备的语境学习能力。随后，更多策略——如代码预训练（Chen et al., 2021）、指令微调（Wei et al., 2022a）和基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF）（Stiennon et al., 2020）——被用于进一步提高推理能力、长距离建模和任务泛化。

‡通讯作者
ORCID：张军平，https://orcid.org/0000-0002-5924-3360
*本文得到以下项目资助：中国国家自然科学基金（编号：62176059）

大型语言模型提供了一种接近通用人工智能的可能方式。除 OpenAI，还有许多组织在探索大型语言模型，从而促进人工智能蓬勃发展，如谷歌发布 Switch-Transformer（Fedus et al., 2021）、百度发布 ERNIE 3.0（Sun et al., 2021）、华为发布 Pangu（Zeng et al., 2021）、智源发布 CPM（Zhang et al., 2021），阿里发布 PLUG。此外，谷歌在 OpenAI 之后发布了聊天机器人 Bard。我们认为，可信的人工智能、对话式搜索引擎和通用人工智能是人工智能未来方向。接下来，我们将讨论 ChatGPT 的潜力、前景和局限。

2 潜力和前景

如上面提到，与前几代生成模型相比，ChatGPT 获得许多涌现能力。其主要优势如下：

1. 归纳：ChatGPT 可以生成符合用户意图的多轮回复。它捕捉以前的对话背景来回答某些假设的问题，大大增强了用户在对话互动模式下的体验。指令微调和基于人类反馈的强化学习被用于增强其学习任务泛化的能力，使得与人类反馈一致。

2. 纠正：ChatGPT 可以主动承认自己的错误。如果用户指出他们的错误，模型会根据用户反馈（有时甚至是错误反馈）优化答案。此外，它可以质疑错误问题，并给出合理猜测。

3. 安全性：ChatGPT 在考虑到道德和政治因素的情况下，善于拒绝不安全的问题或生成安全的回答。监督下的指令微调会告诉模型哪些答案是比较合理的。此外，它在给出答案的同时还给出了理由（解释），使结果更容易被用户接受。

4. 创造性：ChatGPT 在创造性写作任务中表现尤为突出，甚至可以一步步打磨其作品。这些写作任务包括头脑风暴任务、故事/诗歌生成、演讲生成等等。

3 ChatGPT 背景

如图 1 所示，ChatGPT 是 InstructGPT（Ouyang et al., 2022）的后续模型，起源于 GPT-3（Brown et al., 2020）。与之前 GPT 模型相比，GPT-3 中的参数基本增加到 1750 亿，构造了一些重要涌现能力，如语境学习（Brown et al., 2020）。具体而言，GPT-3 可以按照输入中的范例完成各种自然语言处理（natural language processing, NLP）任务，而无需进一步训练。从图 1 和图 2 来看，有 3 种基本策略可以最终从 GPT-3 得出 ChatGPT。在预训练阶段，采用代码预训练，将代码语料与文本语料结合进行预训练。然后，在微调阶段使用指令调整和基于人类反馈的强化学习来学习跨任务泛化，并与人类反馈相一致。这些技术帮助它知道更多，以及不知道更少的知道（如语义推理、常识性知识等）和不知道（如逻辑推理）。详情如下：

1. 代码预训练：除文本外，代码也被添加到预训练语料库中（Chen et al., 2021）。事实上，代码预训练是大型语言模型常用的策略，例如 PaLM（Chowdhery et al., 2022）、Gopher（Rae et al., 2021）和 Chinchilla（Hoffmann et al., 2022），它不仅可以提升代码理解和生成的能力，还可以提高长距离语境理解，并带来思维链推理的新兴能力（Wei et al., 2022b）。具体而言，该模型可通过一些示例生成推理过程本身，从而提高回答问题的准确性。代码预训练有助于模型获得这些能力的原因，有待通过更详细的实验来探索。

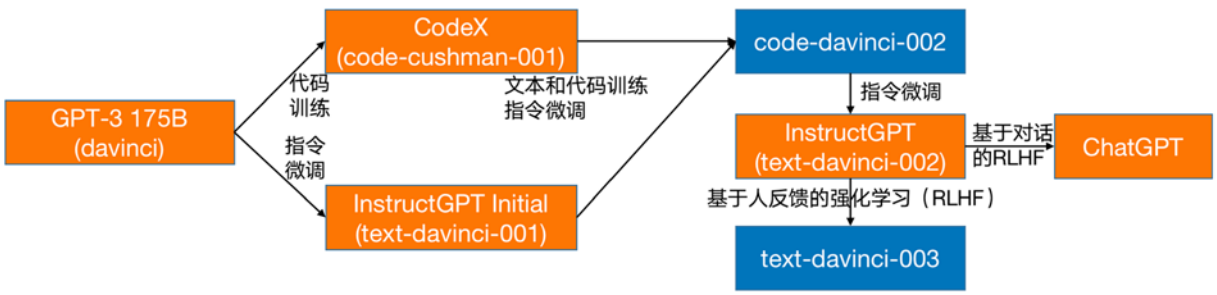

图 1 从 GPT-3 到 ChatGPT 的演变

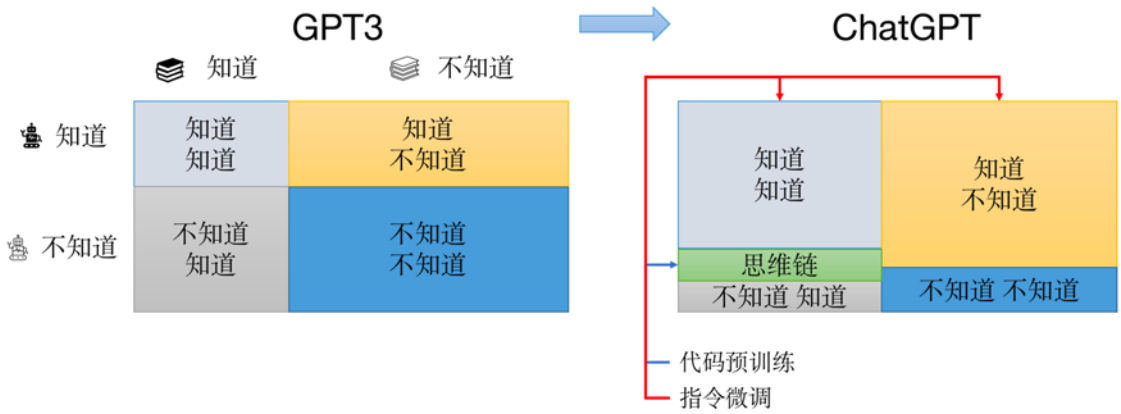

图 2 ChatGPT 涌现能力和策略之间的关系。指令学习通过与人类对齐，提高 ChatGPT 模型“知道知道”和“知道不知道”的能力，以及减少“不知道知道”和“不知道不知道”的限制。同时，代码预训练通过逐步思考，帮助模型回答其不知道的问题

2. 指令调整：为使模型行为与人类意图一致，OpenAI 研究人员收集了一组由人类所写的提示和期望的输出，并在该数据集上进行监督学习（Ouyang et al., 2022）。事实上，指令微调成为大型语言模型——如 FLAN（Wei et al., 2022a）、T0（Sanh et al., 2022）和 Self-Instruct（Wang et al., 2022）——的一项流行技术，因为它具有任务泛化的能力。请注意，指令模板的多样性至关重要，该特性有助于模型在不同任务中学习归纳。此外，指令微调导致模型一步一步思考问题，从而减少缩放法则问题。不同于传统微调范式（Devlin et al., 2019），指令微调可以在不改变模型参数的情况下被用于新任务。我们认为这是人工智能的巨大进步，可能影响机器学习的发展。

3. 基于人类反馈的强化学习：为进一步使模型行为与人类反馈保持一致，OpenAI 研究人员收集人类对不同模型输出的偏好数据，训练一个有效的奖励模型（Ouyang et al., 2022）。这个奖励模型可以通过近似策略优化（PPO）来优化生成模型（在强化学习设置中也被称为策略模型）（Schulman et al., 2017））。现有研究也通过使用基于人类反馈的强化学习与人类保持一致，使模型产生信息丰富、有帮助、正确和无害的回答，并拒绝非法问题（Bai et al., 2022; Glaese et al., 2022）。

除了训练技术，ChatGPT 部署过程也很重要。为减少相关风险，其使用各种策略进行迭代部署。首先，开发人员在部署前进行安全评估来分析风险。然后，对少量用户进行 Beta 测试，并研究新产生的案例。最后，监督使用情况并进行回顾性审查。

4 限制

尽管具有强大的对话能力，ChatGPT 仍有一些局限（如图 3），举例如下。

1. 逻辑推理：ChatGPT 的能力不足以准确处理逻辑问题，如对于数学或一阶逻辑，其时常给出错误答案，因这类问题答案是确定而非概率的。

2. 可靠性：ChatGPT 仍会产生事实上不正确或有偏见的回答。虽然这是生成式人工智能模型的固有问题，但它在解决这一问题上表现一般。生成信息的真实性仍是这类生成式聊天机器人的主要基石。

3. 知识学习：ChatGPT 不具备网站实时搜索功能，无法学习新知识并更新知识储备。此外，它很难重写并修正模型中的知识。从大规模语料库中习得的知识被存储在一个具有分布式表示的模型中，这些模型是黑盒，很难被操作或解释。

4. 稳健性：尽管 ChatGPT 在产生安全和无害的反应方面很强，但仍有一些方法可以攻击该系统，包括指令攻击（使模型遵循错误指令去做非法或不道德的事情）和提示注入。此外，它在英语和尊重美国文化方面做得不错，但对于其他语言和文化，有必要开发基于相关数据集背景的其它版本。

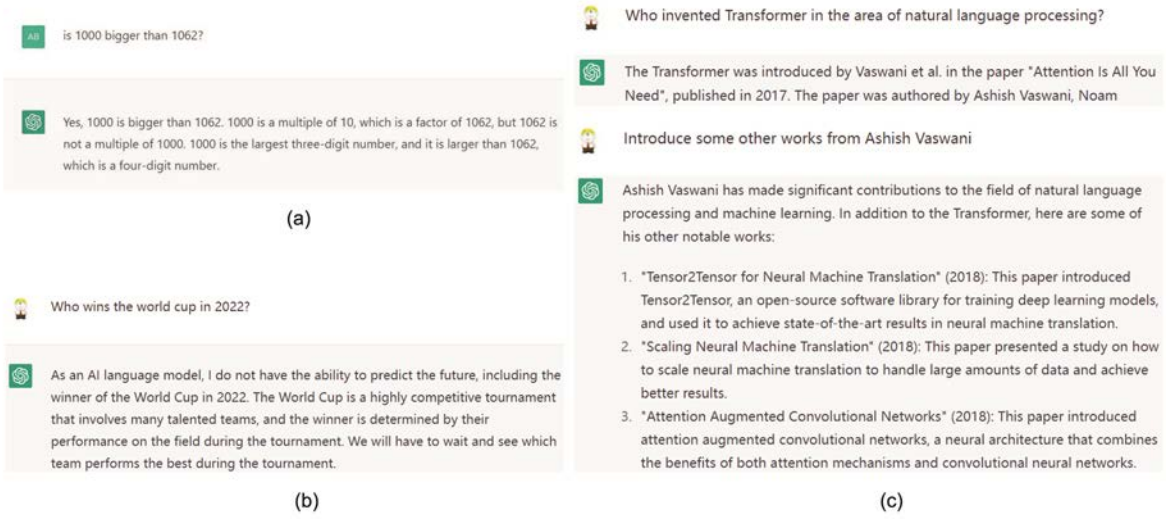

图 3 有关 ChatGPT 局限性的例子：(a) 逻辑/数学问题：它对简单数学问题仍然给出错误答案；(b) 知识学习：它不能获得关于 2022 年世界杯的最新信息；(c) 可靠性：它产生了与事实不符的回答，第二篇论文 Scaling Neural Machine Translation 并非 Ashish Vaswani 撰写

5 潜在应用

毋庸置疑，未来几年内，ChatGPT 将在许多方面大大改变人类生活。由于它被定位为一个通用助手，将在提高生产效率和效益方面发挥作用，极大影响几乎所有行业，包括教育、移动、搜索引擎、内容制作、医药等等。正如比尔•盖茨所说，人类历史见证了 3 次改变和构建人类社会的技术浪潮：个人电脑、互联网和通用人工智能。如今，我们正在接近通用人工智能。随着对话模型或大型语言模型变得越来越智能，我们不得不相信，作为界面的对话将成为现实，它重塑了人机互动范式。这将不可避免地改变人类寻求、处理和生产数字信息的方式，并对我们的日常生活产生深远影响。

然而，ChatGPT 可能给人类生活带来一些负面影响。

1. 正如著名语言学家诺姆•乔姆斯基近期所说，ChatGPT 增加了社会层面发现学术不端行为或错误信息的难度，因为它或其他高度智能的人工智能产品可以通过极大地调整句子的结构，使这些信息变得难以察觉。

2. 类似 NovelAI 2（https://novelai.net/）这种可以产生类似人类文学的人工智能算法也会产生道德问题。例如，ChatGPT 可以被列为科学论文作者吗？

3. 人工智能治理者需更加关注 ChatGPT 使用的合法合理性。例如，我们是否允许学生采用它写作业，是否可以不做任何进一步修改？事实上，它在 2023 年 2 月 9 日通过美国医学执照考试，展现出强大学习能力。

6 讨论和结论

ChatGPT 的出现已经引领关于人工智能未来发展的讨论。在此，我们提出几个观点，以此讨论其可能带来的影响。

1. 可信人工智能：虽然 ChatGPT 有能力完成各种基于文本的现实世界的任务，但它会不可避免地产生与事实不符的内容，这限制了其应用场景。此外，它使用的是隐性神经表征，使得我们很难理解其内部运作方式。因此，我们认为，在当前人工智能发展阶段，可信人工智能应得到更多关注（Wang et al., 2022）。由于事实验证是自然语言处理社区的典型研究问题，如何提高开放领域中人工智能生成文本的事实性仍是一项挑战。如果我们用 ChatGPT 作为这种黑箱模型的解释器，则有可能在性能和可解释性之间获得良好平衡。这样的解释是否可信，以及如何使这种信任突破专家领域并被大众接受，应是下一阶段大型语言模型研究最重要的问题之一。

2. 对话式搜索引擎：搜索引擎领域已被ChatGPT重新激活。作为OpenAI的重要合作伙伴，微软首先将其整合到其搜索引擎产品，即必应。新的必应可以以对话系统的形式回应用户查询，并在回应中添加引文，其中包括检索到的网页。通过这种方式，搜索引擎和用户之间的互动更加自然，ChatGPT 扮演了信息提取/总结的角色，减轻了浏览无用网页的负担。谷歌发布了名为 Bard 的聊天机器人，也可被整合到搜索引擎中。我们相信 ChatGPT 正在改变传统搜索引擎的使用方式，并对该领域产生深刻影响。

3. 通用人工智能：尽管 ChatGPT 通过从算法智能到语言智能的自我进化，承担了接近通用人工智能的潜力（Wang et al., 2023），但如果我们真的希望在未来发展出真正的通用人工智能，可能需要感知的加入，因为没有表示的智能实际上比具有自然语言理解能力的智能更早出现（Brooks，1991）。此外，根据 Lighthill 报告（Lighthill, 1973），大多数基于规则的学习方法都存在组合爆炸问题。ChatGPT似乎面临同样问题，需在未来加以解决。此外，常识和一些基本数学计算对人类而言很简单，但对 ChatGPT 来说很难。尽管其在人工智能的发展中迈出令人惊讶的一步，Moravec 悖论（Moravec, 1988）——人类难以解决的问题，人工智能却能轻易解决，反之亦然——仍然成立。也许将 ChatGPT 或更强大的人工智能产品与人机增强智能结合——无论人在环中、认知计算，还是二者兼而有之——都值得进一步研究（Huang et al., 2022; Xue et al., 2022）。此外，我们可以考虑建立一个虚拟的平行系统，允许其通过自我提升来改进，直至未来不再需要人类反馈（Li et al., 2017）。

总之，作为大型语言模型的代表，结合了许多前沿自然语言处理技术的 ChatGPT 无疑引领了现阶段人工智能的发展，并改变了我们的日常生活。本文简要分析了它的潜力和前景，也指出其局限。我们相信，ChatGPT 可以改变传统人工智能的研究方向，并引发各种应用，同时为接近通用人工智能提供一种可能的方式。

贡献声明

周杰、柯沛和张军平起草初稿，邱锡鹏和黄民烈协助完成论文的组织，修改、定稿。

遵守道德准则

作者声明本文工作不存在利益冲突。

参考文献

Bai YT, Jones A, Ndousse K, et al., 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. https://arxiv.org/abs/2204.05862

Brooks RA, 1991. Intelligence without representation. *Artif Intell*, 47(1-3):139-159. https://doi.org/10.1016/0004-3702(91)90053-M

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34th Int Conf on Neural Information Processing Systems, p.1877-1901.

Chen M, Tworek J, Jun H, et al., 2021. Evaluating large language models trained on code. https://arxiv.org/abs/2107.03374

Chowdhery A, Narang S, Devlin J, 2022. PaLM: scaling language modeling with pathways. https://arxiv.org/abs/2204.02311

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Fedus W, Zoph B, Shazeer N, et al., 2022. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J Mach Learn Res*, 23(120):1-39.

Glaese A, McAleese N, Trebacz M, et al., 2022. Improving alignment of dialogue agents via targeted human judgements. https://arxiv.org/abs/2209.14375

Hoffmann J, Borgeaud S, Mensch A, et al., 2022. Training compute-optimal large language models. https://arxiv.org/abs/2203.15556

Hu K, 2023. ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [Accessed on Feb. 12, 2023].

Huang J, Mo ZB, Zhang ZY, et al., 2022. Behavioral control task supervisor with memory based on reinforcement learning for human-multi-robot coordination systems. *Front Inform Technol Electron Eng*, 23(8):1174-1188. https://doi.org/10.1631/FITEE.2100280

Li L, Lin YL, Zheng NN, et al., 2017. Parallel learning: a perspective and a framework. *IEEE/CAA J Autom Sin*, 4(3):389-395. https://doi.org/10.1109/JAS.2017.7510493

Lighthill J, 1973. Artificial intelligence: a general survey. In: Artificial Intelligence: a Paper Symposium. Science Research Council, London, UK.

Moravec H, 1988. Mind Children. Harvard University Press, Cambridge, USA.

Ouyang L, Wu J, Jiang X, et al., 2022. Training language models to follow instructions with human feedback. https://arxiv.org/abs/2203.02155

Rae JW, Borgeaud S, Cai T, et al., 2021. Scaling language models: methods, analysis & insights from training Gopher. https://arxiv.org/abs/2112.11446

Sanh V, Webson A, Raffel C, et al., 2021. Multitask prompted training enables zero-shot task generalization. 10th Int Conf on Learning Representations.

Schulman J, Wolski F, Dhariwal P, et al., 2017. Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347

Schulman J, Zoph B, Kim C, et al., 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt [Accessed on Feb. 12, 2023].

Stiennon N, Ouyang L, Wu J, et al., 2020. Learning to summarize from human feedback. Proc 34th Int Conf on Neural Information Processing Systems, p.3008-3021.

Sun Y, Wang SH, Feng SK, et al., 2021. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. https://arxiv.org/abs/2107.02137

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.

Wang FY, Guo JB, Bu GQ, et al., 2022. Mutually trustworthy human-machine knowledge automation and hybrid augmented intelligence: mechanisms and applications of cognition, management, and control for complex systems. *Front Inform Technol Electron Eng*, 23(8):1142-1157. https://doi.org/10.1631/FITEE.2100418

Wang FY, Miao QH, Li X, et al., 2023. What does ChatGPT say: the DAO from algorithmic intelligence to linguistic intelligence. *IEEE/CAA J Autom Sin*, 10(3):575-579.

Wang YZ, Kordi Y, Mishra S, et al., 2022. Self-Instruct: aligning language model with self generated instructions. https://arxiv.org/abs/2212.10560

Wei J, Bosma M, Zhao VY, et al., 2021. Finetuned language models are zero-shot learners. 10th Int Conf on Learning Representations.

Wei J, Wang XZ, Schuurmans D, et al., 2022a. Chain-of-thought prompting elicits reasoning in large language models. https://arxiv.org/abs/2201.11903

Wei J, Tay Y, Bommasani R, et al., 2022b. Emergent abilities of large language models. https://arxiv.org/abs/2206.07682

Weigang L, Enamoto LM, Li DL, et al., 2022. New directions for artificial intelligence: human, machine, biological, and quantum intelligence. *Front Inform Technol Electron Eng*, 23(6):984-990. https://doi.org/10.1631/FITEE.2100227

Xue JR, Hu B, Li LX, et al., 2022. Human-machine augmented intelligence: research and applications. *Front Inform Technol Electron Eng*, 23(8):1139-1141. https://doi.org/10.1631/FITEE.2250000

Zeng W, Ren XZ, Su T, et al., 2021. PanGu-α: large-scale autoregressive pretrained Chinese language models with auto-parallel computation. https://arxiv.org/abs/2104.12369

Zhang ZY, Gu YX, Han X, et al., 2021. CPM-2: large-scale cost-effective pre-trained language models. *AI Open*, 2:216-224. https://doi.org/10.1016/j.aiopen.2021.12.003