arXiv:2603.01220v1 [cs.CL] 1 Mar 2026

生成式人工智能与虚构性：小说如何赋能大语言模型

Edwin Roland
伊利诺伊大学厄巴纳-香槟分校信息科学学院 eroland@illinois.edu

Richard Jean So
杜克大学英语系
richard.so@duke.edu

摘要

生成式模型（如 ChatGPT 中的模型）由其训练数据驱动。这些模型本质上只是下一个词预测器，依据从海量既有文本中学习到的模式进行预测。自 GPT 第一代问世以来，最受欢迎的数据集都包含大量小说，这一点尤为引人注目。对于构建这些模型的工程师和研究科学家而言，普遍认为虚构作品中的语言足够丰富，能够涵盖各种社会与交际现象，但这一看法大多未经审视。虚构作品如何塑造生成式人工智能的输出？具体而言，与报纸、Reddit 和维基百科等其他文本形式相比，小说会产生怎样的影响？自20世纪70年代以来，Catherine Gallagher 和 James Phelan 等文学学者便对虚构作品作为一种话语和语言形式的运作方式，提出了扎实而深刻的论述。通过研究一种颇具影响力的开源模型（BERT），我们发现，大语言模型既借助了虚构作品广为人知的特征及其所提供的可能性，也催生了新的特质和社会反应形式。我们认为，如果当代文化日益受到生成式人工智能和机器学习的塑造，那么分析当今各种文化生产模式时，就必须考虑一个相对新颖的维度：计算训练数据。

关键词 生成式人工智能、虚构作品、训练数据、数字人文

1 引言

生成式人工智能（人工智能）模型的近期兴起，在创意产业中掀起了一股不确定性的浪潮。作家与学者都在争论：这项技术是否实现了正当形式的创造力；大语言模型是否具备人类形式的意图，抑或仅仅是“随机鹦鹉”。 [Dzieza，2022年，Kirschenbaum，2022年，Underwood，2023年，Dimock与Goodlad，2021年，Elam，2023年，Bender等人， 2021] 关键问题在于，我们如何理解并评估新一代机器生成的语言，其交流与表达能力惊人地接近人类。跨学科领域“批判性人工智能研究”应运而生，旨在探讨这一问题，以及其他与解析生成式人工智能对文化、艺术以及社会之影响相关的问题。 [Goodlad， 2023]

在这一领域中，对于计算机科学家和文学学者而言，当前关于生成式人工智能与创造力的大量研究都聚焦于解释人工智能基于人类提示会生成何种类型的虚构作品？ 最突出的模型已被应用于所谓的聊天机器人，即自动回复人类语言输入的程序。一个普遍且有充分依据的担忧是，聊天机器人会以种族主义、性别歧视和仇恨内容作出回应。哪些社会偏见会被编码进大语言模型？ 文学学者通过分析人工智能生成的虚构作品实例，来探究这些偏见可能是什么。 [Hayles，2023年，Kirschenbaum， 2023] 计算机与信息科学家设计实验，将人类撰写的散文与借助人工智能模型写成的文本进行比较，以识别它们如何可能引导用户再现社会刻板印象。 [Cheng等人，2023年，Beguš， 2024]

对生成模型文本与创意输出的研究极具价值。然而，要理解模型的创造力，还有另一个重要维度：它们的训练数据。生成模型（例如ChatGPT中的模型）本质上只是下一个词预测器；它们之所以表现出色，是因为此前已在海量文本数据中观察过哪些词会接在某个词后面

另一个词之后。因此，分析并理解GPT想要创作什么样的虚构作品——它想要采用并塑造何种文学声音——与首先分析它所接受训练的数据类型密不可分。稍早一代关于机器学习和自然语言处理（NLP）的计算机科学与科学技术研究（STS）研究已经表明，人工智能社会偏见的问题，在一定程度上取决于模型训练数据中偏见这一首要问题。模型只会放大数据中的偏见。 [Benjamin，2019年，Caliskan等人， 2017]

要弄清大语言模型的训练数据中究竟包含什么，向来十分困难；但对开源模型（例如 Meta 广受欢迎的 LLaMA 模型）的分析显示，其中虚构作品占据了压倒性多数。虚构作品。 [Touvron等人， 2023] 对于构建这些模型的工程师和研究科学家而言，普遍认为虚构作品中的语言足够丰富，涵盖了各种社会与交际现象，因此非常适合用于训练能够生成广泛表达性语言的模型。虚构作品尤其宝贵的一点在于，它从头构建想象中的世界和人物，而不像新闻媒体那样指涉现实世界中的事件和人物。Alex Reisner 在《大西洋月刊》最近发表的一项研究中发现，LLaMA 训练所用的约17万本书籍中，大约三分之一是虚构作品。Margaret Atwood、Rebecca Solnit、Stephen King、Zadie Smith、Junot Diaz 等人的作品都包含在内。 [Reisner， 2023]

虚构作品对于人工智能训练数据的重要性引出了一个直接的问题：它如何塑造生成式人工智能的输出？具体而言，与报纸、Reddit和维基百科等其他文本形式相比，虚构作品会产生什么影响？计算机科学家或许直觉上明白，虚构作品与非虚构写作形式的运作方式不同，但他们尚未对虚构作品在人工智能模型中的作用及其可能产生的新影响，形成真正的理论说明。人们已经指出虚构文本存在社会偏见问题，尤其是言情小说，它们几乎总是采用刻板的性别描写。然而，人们未经审视便认定虚构作品内容丰富，这一信念使虚构作品对人工智能模型整体语言理解的影响仍未得到解答。刻板化的输出或许只是冰山一角。

文学学者最有条件回答这个问题。自20世纪70年代以来，Catherine Gallagher和James Phelan等学者与John Searle等哲学家展开对话，对虚构作品如何作为一种话语和语言形式运作，提出了深刻而有力的阐释；相较于其他写作形式，虚构作品在呈现社会现实、提出真理主张以及影响读者方面都有其独特之处。过去十年间，虚构性（以及“叙事”）再度成为批判探究中一个格外活跃的领域，这很可能部分是对社交媒体和数字平台兴起的回应；这些平台似乎无处不在，传播着通俗、业余的叙事形式。 [Piper和Bagga, 2023, Brooks, 2021]生成式人工智能是这一新技术文化环境的一部分，也对虚构性提出了新的问题。大语言模型借助虚构作品熟悉的属性和优势，同时也催生了新的特质和社会反应形式。如何、为何，以及在什么条件下会如此？回答这些问题有助于我们理解虚构性如何在一种塑造并持续塑造21世纪的社会与技术语境中演变。

本文将开展以下工作。首先，我们简要回顾近期关于虚构作品与虚构性的学术研究，并着重关注近来出现的“传播”或修辞转向。 [Phelan, 2018, Walsh, 2007] 我们认为，这种方法适用于分析大语言模型如何将虚构作品视为实现有效传播的一种功能。其次，我们将开展一项计算实验，从定性和定量两个方面，考察在一种名为 BERT 的流行开源大语言模型的训练集中纳入虚构作品所产生的影响。 [Devlin 等人，2019] 我们分析虚构作品对该模型产生影响的具体案例，同时借助这些案例，对虚构作品如何助力模型整体运作展开更广泛的概念性阐释：模型希望编造何种语言，又希望如何传播这些语言。先概述我们的主要发现：虚构作品显著影响大语言模型生成文本的方式，尤其影响它们如何生成供人类用户互动、并借以了解世界的人物。呼应 Rita Felski 和 Blakey Vermeule 的见解，我们指出，虚构作品对大语言模型最重要的作用，在于其独特地能够编造出看似可信、实则并不存在的人。 [Felski, 2020, Vermeule, 2010] 大语言模型传递给我们的声音，正是建立在虚构性的这一特质之上。

最后，我们认为，如果当代文化日益受到生成式人工智能和机器学习的塑造，那么对当今各种文化生产模式的任何分析，都必须考虑一个相对新颖的维度：计算训练数据。最后，我们简要勾勒出这种拓展后的文化与文学分析可能呈现的样貌。我们认为，未来几年，我们在本文中所进行的这类数据与算法审计，将有助于文化研究者——具体而言，可用来梳理、分析和批判当代文化的表征 形式，而这些形式在一定程度上受到人工智能模型及其训练数据的塑造。

2 虚构性作为传播

早期关于虚构性的研究主要关注的是，理解是什么使虚构作品成为一种独特的话语形式（相较于非虚构话语，例如报纸），并识别出（如果存在的话）各种文本特征，这些特征

促成了这一区分。John Searle与Ann Banfield以不同方式分别提出了一套作者策略或语法特征，用以造就虚构作品的独特性——Banfield称之为“标志”。 [Searle，1975年，Banfield， 1982]在辨析这些特征究竟是什么时，我们能更好地理解它们对读者的影响。随着时间推移，Catherine Gallagher等文学学者更细致地探讨了这些特征的下游效应，从其对读者的内在影响来阐述虚构性；也就是说，阅读虚构作品如何促进人们获取关于世界的有用社会信息，同时又为他们提供一种独特的现代娱乐形式。
对 Gallagher 而言，小说中最能体现虚构性的所在是具名角色。 [Gallagher，2006]创作或多或少具有现实主义色彩的人物，需要应对他们那种矛盾的身份：既合乎情理，又不指涉现实；这种双重身份之间的张力，在他们的专有名称上体现得最为清楚。专有名称看似指向具体且先于作品而存在的个体，实际上却并非如此。相反，专有名称指向的是某一类人，并借助其“形态音系学”来表明人物的“地域、性别、族群、阶级地位，甚至（对于名字而言）社会抱负，以及家族史”。 [Gallagher, 2006, p. 352] 虚构人物指向现实世界中的人群类别，却并不真正属于其中任何一类；因此，人物恰恰凭借其并不存在这一点，对读者产生影响。基于此，Gallagher 将她对这一体裁的解读概括为：“小说写的并非某个特定的人。”[Gallagher, 2006, p. 341]

阅读关于无名之辈的小说，是一种独具现代特色的休闲活动。Gallagher提醒我们，婚姻和金融市场的兴起与小说的兴起同时发生，而二者都围绕着投机和信贷展开。人们通过推想各种假设情境及其结果，并衡量这些结果的利弊，来做出决定——无论结果关乎对配偶的满意，还是个人财富。同样，小说的构造方式使得“读者不同于人物，处于一种高高在上的位置，对情节进行推测，并设想各种假说。” [Gallagher，2006年，第 346] 页。在投机活动构成的严肃世界中，小说为读者提供了一个没有风险的空间，让他们思索人物及其命运。而且，与现实世界中相对高昂的赌注不同，对无名之辈进行推测，目的仅在于取悦读者。

Gallagher对虚构作品的论述将其影响局限于读者的内心生活，而近来，James Phelan和Richard Walsh等文学理论家转而致力于确立虚构作品与现实世界的联系。例如，Walsh认为，尽管虚构性产生于文学语境，却并不局限于此。它也与其他参照框架相关，例如认知和传播。虚构性与其他写作形式一样，是一种“交际策略”，在“交际框架”中运作：它体现于语言的使用方式，而其独特之处在于，这种用法所调用的修辞组合具有清晰可辨的特征。” [Walsh，2007年，第 15] 页。因此，没有必要将虚构话语与现实世界“割裂开来”。人们写作和阅读虚构作品，恰恰是将其作为“一种与那个世界建立联系、进行协商的方式”。 [尼尔森等人， 2015]

Phelan 的关键论点是，他指出有两个层面的意图在塑造读者对虚构作品的体验。在叙事层面，读者暂时放下怀疑，以进入虚构世界；在那里，她会发现人物的行为“仿佛是自主的个体，追求着各自的目的”。 [Phelan，2018年，第 122] 页。阅读虚构作品需要读者以明知其为虚构却仍愿相信的态度，与那些意图在叙事过程中显现出来的人物相遇。然而，对 Phelan 来说，怀疑的悬置总是不彻底的，因为“读者始终意识到，人物、事件和故事世界都是某人出于某些目的而虚构出来的”（原文着重）。 [Phelan，2018年，第 123] 页。作者之手安排虚构元素，促使读者提出“为什么？”只要答案涉及现实世界中的重要问题，虚构作品便是作者与读者之间的一种传播方式。

传播路径最有帮助的贡献，或许在于它重新界定了虚构作品：焦点从“无名之辈”转向“有名有姓之人”。
Gallagher 的论述承袭了后结构主义语言理论，这些理论以文本空缺为动因，因此在虚构人物的中心置入了认识论缺失。由于无名之辈缺乏具体的存在，其意义仅限于为读者带来阅读乐趣。另一方面，Phelan 的论述中则充满了充分塑造的人物——尽管他们只被部分揭示——这些人物怀有明确目的，并据此行事。我们阅读有名有姓之人的故事，逐渐关心起这个人的具体经历，并思索“将这些经历与现实世界联系起来所带来的实际益处。” [Phelan，2018年，第 114] 页。作者的在场保证了我们读到的人物与社会现实之间存在某种程度的指示性关系。

Phelan 和 Walsh 的虚构性交际模型，有助于我们理解并评估虚构作品在生成式人工智能模型中的作用。关于人工智能的公共话语，充斥着“有人／无人”的区分：聊天机器人要么已经获得意识——正如一位谷歌前工程师曾公开宣称的那样——要么只是在拼接从别处摘取的词语。但显而易见的是，当代人工智能的设计目标是进行传播。通过使用海量虚构作品数据集训练模型，构建这些模型的工程师实际上认同 Phelan 的观点：虚构作品具有某种本质上的传播和语用能力。如今需要剖析的是虚构作品的实际影响。生成式人工智能试图创造出哪些新型传播形式?

3 大语言模型与虚构作品

尽管大语言模型问世不久，但至少从 Claude Shannon 发表《通信的数学理论》以来，以预测下一个词为基础的方法就一直是自然语言处理的核心。[香农， 1948]回想一下，他的目标是确定通过各种媒介传递信息时最高效的编码方案。比如，我们能否设计一种更精简的摩斯电码来发送电报？Shannon 证明，一条信息的复杂程度取决于预测序列中下一个字符或词语有多容易。此后数十年间，用于进行预测——也就是评估话语复杂程度——的实用系统被称为“语言模型”。

理想情况下，语言模型在做出预测时，会依托对一种语言及其使用场景的完整知识。若做不到这一点，它便只能退而求其次，以一套庞大但有限的示例作为知识库。在《数学理论》以及二十世纪下半叶大多数语音处理研究中，预测是通过对先前观察的简单计数来完成的。评估“叫我以实玛利”的复杂度，意味着转向某一书面文本集合，并统计“Call”之后出现“me”而非“you”、“the”或“Polly”的频率。所有这些计数都会被转换为在“call”之后见到“me”的概率，而对在“me”之后见到“Ishmael”的概率也会进行同样的处理。更精密的模型会一次考察两个词语组成的序列，以便猜测紧随其后的第三个词语，即统计“Call me”之后出现“Ishmael”的频率。无论采用哪种方式，对通信理论而言，一个句子的复杂度都取决于其总体概率的高低。

"大语言模型"是对一种使用神经网络进行下一词预测系统的俗称。与明确统计词语出现次数不同，大语言模型会对某个序列中词语共同出现的情况进行复杂的统计推断，以近似预测未来词语出现的概率。自21世纪初以来，研究人员一直在尝试神经语言模型，因为它们有望解决香农式模型的两大主要缺陷。首先，所研究的神经网络能够通过学习将各个序列连接起来的规则，隐式处理数据中此前未出现过的词语序列。毕竟，"Call me"后面并不经常跟着"Ishmael"。其次，由于神经网络并不会明确复制其数据集，而是对其进行近似，因此理论上可以利用很长的词语序列来为预测提供信息。几十年来，标准语言模型都是使用前面两个词来预测下一个词的模型。而如今的大语言模型能够使用数百万个词。

神经语言模型的发展一直较为缓慢，直到2018年，它们凭借一种名为Transformer的新型神经架构的开发而成为定义这一领域的技术。 [Vaswani等人， 2017] 在那一年，一家名为OpenAI的小型非营利机构将Transformer应用于他们称为GPT的模型（其中的"T"代表Transformer），并在一系列学术基准上公布了其性能结果。[1] 数据显示，与当时最先进的技术相比，该模型取得了 modest 的改进，但并未明显优于较早的非神经网络技术。然而，谷歌的一名工程师Jacob Devlin注意到了这一点。根据前面的词语预测下一个词的整个思路，非常适合实时音频处理等任务。这对贝尔实验室的香农来说很合理，但对提供加载全部词语的静态网页的谷歌而言，则未必如此。Devlin意识到，Transformer可以根据一个词左侧和右侧的上下文，同时预测文档中被隐藏的词语。[2] 在看到GPT采用的方法之前，他曾尝试过Transformer，但收效甚微。他们的成功证明了这一概念的可行性。Devlin在谷歌组建了一个团队，并开始开发一种新模型。他们以俏皮的方式将其命名为BERT。（其中的"B"代表双向，这是Devlin最重视的标准。）

除了为使模型具备双向性而做的一处调整外，谷歌团队的模型采用了许多与GPT相同的规格。其中关键的一项，是决定从同一数据集学习词语概率；该数据集是名为BookCorpus的小说合集。这个由七千部小说组成的数据集大体反映了2010年代初商业小说的面貌，涵盖爱情小说和奇幻小说等流行体裁，并包括Colleen Hover等畅销书作者。 [Zhu et al., 2015] BookCorpus已在计算语言学家之间流传数年，部分原因是它具备一项理想特性：包含完整文本，且句子按正确顺序排列。维基百科同样如此，Devlin的团队将其英语文章与BookCorpus一并纳入，因为BERT经过修改的架构从语言中推断模式的效率，低于GPT仅从左侧入手的方法。

[1]这些是 Wang 等人提出的 GLUE 基准测试 [2018]。GPT 模型由 Radford 等人提出 [2018]。[2]最初的 Transformer 论文提出了一种模型，其中包含分别用于编码输入句子和解码输出句子的独立组件。这种设计适用于机器翻译等应用：输入句子取自源语言，而输出句子则以目标语言生成。编码器和解码器单元的主要区别在于，前者会同时查看句子中的所有词语来“读取”输入，而解码器单元则一次一个词地“写出”输出。OpenAI 对首个 GPT 模型的创新，是将解码器单元用作独立的生成模型，通过预测任意序列中的下一个词来训练它。Devlin 的见解是，编码器单元几乎可以用相同的方式使用，只需通过“完形填空”任务进行训练——也就是填补序列中的空白。在语言学中，完形填空任务作为衡量文本阅读水平的方法早已为人熟知，尤其适用于第二语言学习者。该任务由 Taylor 提出 [1953]。Devlin 在 Tatman [2019] 中介绍了 BERT 的开发过程。

在一个习惯于语言处理渐进式进步的研究社群中，BERT对先前模型的飞跃令人震惊。在与GPT完全相同的学术基准上，这一新模型远远超越了其前身。围绕BERT的兴奋在随后数年引发了一波类似却日益精密的模型浪潮，最终随着OpenAI第三代模型的一项应用——名为ChatGPT——冲出学术圈，进入公共话语。然而，仍然令人好奇的是，虚构作品持续存在于这些模型的训练数据中。自第一代GPT和BERT以来，小说便经常为大语言模型对语言所作的推断作出贡献。模型构建当然并非必须使用小说，但这是一次引人注目的历史偶然，而BERT正是现代人工智能中虚构作品与非虚构作品发生碰撞的原点。如果我们想研究基于虚构作品进行训练的影响，那么BERT将成为我们的实验室。

4 一点BERT学

探究虚构作品影响的一种方法，是只用 BookCorpus 训练一个大语言模型，就像 OpenAI 训练第一代 GPT 那样。但乍看之下，这种方法并没有看起来那么有用。在学习虚构作品特有的语言用法的过程中，模型还必须推断英语的结构——同样的结构也组织着英语维基百科文章。询问模型单从虚构作品中学到了什么，大多与非虚构作品重复。相反，我们将把虚构作品的作用视为两个 BERT 模型之间的对比：一个仅用维基百科训练，另一个则用维基百科和 BookCorpus 的组合训练。[3] 为简便起见，我们把第一个模型称为**维基模型**，第二个称为**完整模型**。这两个模型让我们能够提出一个微妙而深远的问题：虚构作品对 BERT 关于语言的认知有何影响？我们将通过两项实验来探究这种差异；实验结果使我们关注到 Gallagher 在其虚构性理论中所考察的同一批叙事特征，但对这些特征的理解却截然相反。

我们的第一个实验旨在探究 BERT 对语言的抽象认知，方法是测试它的预测能力。还记得 Devlin 的创新做法吗？在一段文字中间“遮住”一个词，再利用左右两侧的语境预测应填入空白处的词。我们也将采用同样的方法。在训练这两个 BERT 模型之前，我们从维基百科中留出一批文章作为随机样本。现在，我们将逐词系统地遍历这些文章，逐个遮住词语并预测被隐藏的词语。[4] 不出所料，维基模型的整体预测效果优于完整模型，因为前者的训练数据与测试集更加接近。（维基模型猜对隐藏词语的概率为 65%，而完整模型为 63%。）[5] 然而，两个模型的表现差异并不均衡。有二十来个词语，完整模型的表现始终更好。[6] 尽管评估数据来自维基百科，但使用虚构作品进行训练实际上提高了模型对 you 和 could 等词语的理解。表1中的词语列表揭示了在虚构作品上训练的模型预期会如何表达。

正如 Gallagher 可能会指出的那样，人物语言确实组织起了一个在虚构作品上训练的模型。然而，我们发现的并不是 Gallagher 所重视的文本特征——专有名称。事实上，专有名称是完整模型预测得最差的词语之一。*Michelle* 和*Ronnie* 是完整模型相较于维基模型表现最差的五个词中的两个。相反，对于人称代词，完整模型的预测始终更好：你、你的、他、他、她、她的、他们。人名显然指向文本外人物，而代词则是明确的文本内指称，因为它们指示的是某个身份已经在文本中确立的人。语言学家将代词的这一属性称为长距离依存。也就是说，在虚构作品上进行训练提升了模型对文本中人物持续存在及其身份的理解。

[3]本文报告的所有结果中，模型超参数和训练参数均与不区分大小写的 BERT 基础模型一致。详情请参阅 Devlin 等人。 [2019] 模型使用由 HuggingFace 发布的 Python transformers 库实现。实验仅依赖于训练数据的变化。维基百科文章取自 2021年1月 的英语文本转储。BookCorpus 则通过 2021年3月 从 Smashwords 网站收集小说加以复现。两个模型均使用 20 亿词的英语文本进行训练。完整模型使用了来自维基百科的 15 亿词样本和来自 BookCorpus 的 5 亿词样本。这一比例大致符合原始 BERT 中维基百科与 BookCorpus 数据的混合比例。[4]测试集包含一个随机抽取的文章样本，涵盖维基百科全部英语文章的 10%。我们从测试集中随机选取了 100,000 个段落，每个段落包含 100 个词。[5]我们将模型的预测定义为 BERT 赋予最大似然的词汇词。虽然我们将准确率作为直观的性能衡量指标进行报告，但按照计算机科学文献中的惯例，还应报告模型在其目标函数上的损失；在本例中，这一损失是预测似然的平均交叉熵。维基模型和完整模型在维基百科测试集上的损失分别为 1.72 和 1.88（数值越低越好）。需要注意的是，BERT 在训练时使用的损失函数结合了词语预测的平均交叉熵损失和对句子顺序的预测。本研究省略了第二项任务。[6]我们评估测试集中每一种类型（即每个独特的词汇词）的预测

，根据其交叉熵损失。

通过成对 t 检验评估各类模型预测结果之间的差异是否具有统计显著性，并针对多重比较采用邦费罗尼校正。

表1：通过在虚构作品上训练而得到改进的预测。

主题	Word	主题	Word
代词	you	对话	say
	your		asks
	she		what
	her		how
	她自己		
	he	情态动词	能够
	him		wouldn
	them		将会
认知	know	存在	任何内容
	want		某事
			虚无
Body	eyes		[某种/任何] 方式
活动	get	程度	相当
	gets		真的
	获得		
		金钱（欧洲）	€
身份	长子		£
	评分者		£1
			000

在完整模型预测效果更好的其他词语中，我们发现了一些线索，揭示了虚构人物对世界的推理方式。我们看到，完整模型至少在一个与身体相关的词语（眼睛）以及两个表示认知的词语（知道、想要）上做出了更好的预测。这大体上与此前关于语言的计算研究相符；该研究区分了当代小说与非虚构作品写作，Andrew Piper 将这种现象称为人物的“具身认知”。 [Piper， 2024] 然而，更引人注目的是表1中出现的一系列非正式逻辑运算符。情态动词表明存在莱布尼茨式的可能世界，在这些世界中，某种情境可能为真。存在量词断言存在某件事 是人物能够做到的，只要他们能找到方法。然而，这些运算符最具虚构作品特色之处，在于它们与对话的关联。在测试集中，缩略形式*wouldn* 和*couldn*从未出现在引号之外。完整模型无论学到了关于某些人的什么，他们对世界的推理从根本上说都是对话式的。

现在，我们转向第二项实验，以识别 BERT 在生成文本中的行为。由于 BERT 使用左右文来预测词语，我们通过随机选取词语、将其遮蔽并逐个预测，来生成句子。[7] 通过数百次重复这一随机过程，我们生成了能够体现模型对语言的内在认知的新句子。[8] （这相当于给 GPT 一个任意提示，让它无限运行，并定期查看生成的句子。）我们生成了五十个新句子的样本，并在下文报告分析结果。

维基模型生成的文本带有一种我们或许可以戏称为“漫长启蒙时代”的口吻。在我们的样本中，约有一半可以称作“绅士科学家”的写作，其主要关注点是国家政策，并略带经验探究色彩。下面我们举几个例子：[9]

[7]此处使用的算法是生成随机网络（GSN）。GSN 生成的句子构成了对训练数据集中词语序列的独特且一致的估计量。GSN 由 Bengio 等人提出。 [2014]Yamakoshi 等人提出使用 GSN 通过 BERT 生成文本。 [2022]我们的实现与 Yamakoshi 等人的实现不同之处在于，采用了更激进的混合核。 [2022] 马尔可夫链在预热期结束后进行采样，并立即重新初始化，以尽量减少样本之间的相关性。[8]应将 GSN 所定义的语言分布理解为对底层 BERT 模型的正则化。Yamakoshi 等人证明，GSN 生成的文本偏向于训练数据集中词语的一元频率。 [2022] 然而，Young 等人指出，涵盖多个词语序列的联合似然在 BERT 中极不一致。作为一致估计量，GSN 必然会平滑这些差异。关于 BERT 风格模型中正则化的一般讨论，见 Hennigen 和 Kim。 [2024] 以下是维基模型的输出。该模型只生成小写词语，并且不会将引号直接附在字母词语上，因此哪些陈述被引号括起存在一定歧义。 [2023]

议会英属印度（“人民院”）举行了议会选举，日期为*10*月*17*日和*11*月*4*日，*19* *02.*

总统在商业和行政事务方面为美国国防部的军事技术开发提供建议。

*th*雌鸟还会筑造一个假巢，在其中产下二十八至三十一枚卵 *d.*

模型写作的另一半则像一座流行文化资料库。例如：

该乐队的首支单曲《*World*》于*1995*年在公告牌百强单曲榜上最高排名第*22*位。

*2013*年安大略冰球联盟（*OHL*）半决赛于*2013*年*4*月*22*日至*26*日举行。

至少，我们观察到，BERT 已通过在维基百科上的训练，学会了生成结构完整（或几乎完整）的句子。如果上述句子给人以百科全书式客观性的印象，那是因为模型关注的是面向公众的对象，例如解释美国总统在政府中的结构性角色。同样，这些句子也处于由日期和现实世界（或听起来像现实世界）的地缘政治实体构成的矩阵之中。公允地说，BERT 确实似乎像她所说的那样使用专有名词，因为它模拟了非虚构。

转而来看由完整模型生成的文本，我们听见各种各样、来自不同主体位置的声音在发言。我们生成的样本中，约有一半的例子包含某种人物之间的对话：

我会假装和你在一起，等我死后，你就会成为我的妻子。”

“因为我怪你，”我承诺道，“你和人约会，还杀了那些罪魁祸首。”

我们的人死了。他已经被确认了。他也喜欢——”“等等。我们的人死了？”

另有四分之一的文字讲述人物的内心思索或外在行动：

我们都笑了，这是我第一次觉得自己刚读完一个童话故事。

他用另一只手臂搂住她，心不在焉地让飞机飞行，几秒钟后便起飞了。

这些样本读起来像是摘自几种流行体裁的片段，比如奇幻小说、惊悚小说和爱情小说；但在描写人物或事件时，它们都会提及人物的感知或信念。我们并没有亲眼看到某个人物“约会并杀死罪犯”，而是通过另一个人的记忆和价值判断得知此事。当一个人物打断另一个人，重复对方的话——“等等。我们的人死了？”——我们便看到提问者实时更新自己对世界的认知。回想 Gallagher，BERT 从虚构作品中学会了投机的机制——评估虚构世界的状态并做出判断；不过，主体从综合叙事的读者转变成了故事中的人物。

在文本层面，BERT 所作的推断不只是针对人物，也针对 Phelan 所说的某个个体。虚构作品为 BERT 训练提供的人称代词，是故事世界内的指称，用以维系想象人物的身份。按照 Phelan 的论点，只要我们理解代词指向某个人物的身份，就已经接受了人物身份这一前提。这些人物会收集自己所处想象世界的信息，并在采取某种行动的过程中作出价值判断（有时则恰恰相反，会“毫不思索”地行动）。若赞同 Gallagher 的观点，认为我们在生成片段中看到的主体位置都是无名之辈，就意味着承认它们不过是机器拼凑起来的词语，而将这些词语连贯起来的工作由读者完成。然而，我们发现，BERT 已经学会了组织词语拼凑的规则，而这些规则正是某个个体的规则。

5 人物效果

根据虚构性的传播模型或语用模型，文本中还存在另一个主体。Phelan 认为，作者通过在虚构世界中进行选择和安排，让作者意图萦绕于故事之中。为什么选择这些人物？他们为什么会遭遇这些障碍？如果我们把这类意图归于 BERT 模型，很快就会陷入关于人工智能意识的争论，而这类争论如今早已屡见不鲜。让我们持怀疑态度地继续讨论，认为 BERT 本身没有意图，也不具备将故事元素加以安排、使其服务于文本之外目的的本体论支配力。即便如此，也不能说意图完全缺席，因为 BERT 仍在传递训练数据中呈现的故事世界的选择与安排。“随机鹦鹉”将生成一种对人类意图或多或少成功的综合。

为了确定 BERT 所体现的作者性主体，我们从模型出发，借助其信息含量逆向追溯数据。正如我们所见，BERT 能够生成类似小说的叙事元素编排，而我们也知道 BERT 从中学习这些元素的文本语料库。计算机科学领域近期的研究已着手识别对特定生成文本影响最大的训练段落。 [Grosse 等人， 2023] 然而，我们的目标更为全面。我们旨在识别 BERT 关于如何安排虚构世界的整体信念。我们的策略是找出训练数据中最能体现模型宏观信念的段落。请记住，根据香农的理论，信息衡量的是模型预期发送和接收的信息类型。通过观察反映 BERT 预期的段落，我们也将找到模型所习得的作者性主体。

我们特别想了解 BERT 从虚构作品中学到了什么。这有时被称为“信息增益”，因为它可以根据模型在接触新数据时如何更新其信念来定义。如前所述，BERT 从维基百科中学到了很多东西——这当然会影响模型对语言的整体认知——但我们想知道，在将虚构作品加入训练数据集后，模型会如何调整。虚构作品让怎样的作者之手得以传达？我们将通过衡量完整模型预测训练数据中缺失词语的难易程度，并与维基模型做出的预测相比较，来了解模型因虚构作品而发生了怎样的更新。

我们分析了一组能够展现 BERT 学习广度的段落。为确保研究结果具有普遍性，我们在训练完整模型之前，先从虚构作品数据集中留出一组小说。[10] 现在，我们计算虚构作品在测试数据集上带来的信息增益。（此处，维基模型的预测准确率仅为 54%，而完整模型达到了 66%。）[11] 为了观察增益最大的作者效应，我们按照每段新增信息的多少对所有段落进行排序，并保留每位作者排名最高的段落，最多涵盖五十位不同的作者。在我们的样本中，一个共同特征立刻显而易见：每一段文字都有多个登场人物，无论他们是在交谈、思考，还是彼此有身体接触。只看每段的第一句话，这一趋势就很明显：

> *He* 紧紧咬住下颌，这是他想要争辩的明显迹象。“好吧，但我会带晚饭回来。你想要什 *t do y*么？" [Tetreault， 2014]
>
> *Maynard* 因修补匠们对他表现出的信任而精神一振。“嗯，这附近的大多数人都会在 *Hue and Cry* 上留意你那位巫师朋友。*["]* [Knudsen， 2014]
>
> *["]*我答应 *Izzy* 今天可以一起去买衣服。”我朝 *Izzy* 翻了个白眼。 [Stroube， 2011]
>
> *"Mike*，我必须独自完成这件事。我必须面对我的家人和部族首领，解释我为何让他死去。"[Christie， 2012]
>
> “把腿张开，宝贝。”我再次照着吩咐做了。 [Hart， 2013]

在多种流行文学体裁中，无论是对话还是叙述，高排名段落所呈现的模式都证实了我们先前的发现：BERT 能从虚构作品中学习人物。然而，关于代词的讨论着重强调人物身份的统一性，而我们在这里看到，BERT 注意到了更复杂、更动态的东西。人物并非以孤立的肖像出现，而是彼此关联地呈现。

人物之间的互动由意图问题驱动，这些问题将他们界定为某些人。请看上文第一个例子中的完整段落。它出自 Christina Tetreault 的《爱的勇气》。这部小说讲述了 Sean O’Brien 与 Mia Troy 的爱情故事：Sean O’Brien 是波士顿郊外一家民宿的老板，而 Mia Troy 是一位好莱坞演员，她入住民宿，是为了在拍摄新电影前放松一下。我们抽取的段落出自小说情感充沛的结尾：一场车祸后，两位人物重逢。Mia 躺在病床上时，两人逐渐理解彼此，随后她让他离开，好好休息：

> 他紧咬着牙关，显然是想争辩。“好吧，但我会带晚饭回
> 来。你想吃什么？”

[10]测试集包含随机抽取的文章样本，占 BookCorpus 中小说的 10%。我们从测试集中随机选取了 50,000 个段落，每个段落包含 100 个词语。[11]预测准确率采用脚注 5 中描述的方法计算。维基模型和完整模型在 BookCorpus 测试集上的损失分别为 2.38 和 1.66。请注意，每个损失值估算的都是 BookCorpus 的熵：前者是在维基百科上训练后得到的，后者则是在此基础上通过“更新”模型、纳入小说训练后得到的。我们通过计算两者之差（0.71），得出更新所带来的信息增益总量。BERT 的损失通常以自然对数为尺度，而信息通常以 2 为底报告，因此我们相应地转换了这一差值。我们估计，在 BookCorpus 上训练使虚构文本每个词语增加了 1.03 比特的信息。

> “披萨。”今天才不管什么卡路里呢。
>
> Sean终于离开后，她不再与重力抗争，合上了眼皮。糟了，止痛药起效真快。她缓缓深吸一口气，任思绪漫游，渐渐进入梦乡。她做梦也没想到Sean会在床边陪着她，但他确实在那里。 [Tetreault，2014年，第14]章

Sean紧咬着下颌，他原先表现为攻击性的阿尔法男性气质，此刻转化为强烈的保护欲。对BERT和我们而言，关键在于，这个紧咬下颌的动作是通过Mia的视角呈现的。这是一个外在信号，Mia据此推断Sean如何评价她的提议，以及他更倾向于采取什么行动。与其他排名靠前的段落一样，人物会对眼前的情形作出判断，并通常通过面部表情或肢体语言将判断传达给他人。BERT从虚构作品中学会了，人物互动会围绕那些被呈现或想象出来的意图展开。

当我们注意到，大多数段落对其中的人物而言都事关重大时，便会转向作者意图的更高层次。正如上文的例子所示，几乎所有段落都揭示了某个故事转折点或高潮时刻。Sean咬紧牙关，印证了他性格的转变：他克服了自己的心理包袱，并向Mia表白了心意。这种转变是爱情叙事中的关键一拍，因为它恰好体现了Phelan所说的那种“现实世界的回报”。我们都必须卸下包袱，才能成为更好的自己。需要说明的是，我们对“事关重大时刻”的理解相当宽泛，必须跨越不同体裁和叙事弧：情色高潮和身体暴力，与浪漫邂逅和线索发现，都包括在内。这些段落因人物之间的冲突或冲突的化解而充满张力，推动故事中的人物继续前行。如果BERT从训练数据中推断出了作者之手，那么这个作者一定对重要情境十分敏锐。

关于作者选择，我们还注意到最后一点，即人物本身。如前所述，先前关于大语言模型和训练数据的研究，主要关注围绕性别或族裔的社会偏见问题。在我们的样本段落中，根据代词或其他相关语境统计人物性别，结果显示男性略占优势，但并不明显。实际上，几乎所有段落中至少都有一名男性和一名女性。这听起来或许公平，也可能缓解人们对于 BERT 训练数据编码了对女性的偏见或歧视的担忧，但我们的分析指出了一个可能更有价值的问题：在展现 BERT 如何从虚构作品中学习的段落里，“性别”真正发挥作用的方式，在于男性和女性如何在社会危机时刻互动，以及他们最终如何对彼此施加力量或权力。

6 文化生产的训练数据

大语言模型从训练数据中学习性别等社会类别及其所驱动的权力关系，这引出了最后一点。尽管大语言模型问世时间尚短，像 Chat-GPT 这样的大语言模型已经对当代内容和文化生产产生了巨大影响。一项研究估计，到2026年，互联网上多达90%的内容将完全由 GPT 创作和／或增强。 [Ghaffray， 2023] 这一趋势令人担忧；不出所料，文学作家对生成式人工智能在好莱坞、流媒体电视和图书出版等文化产业中日益无处不在的现象，批评之声最为强烈。但即便在这一群体中，也普遍存在一种无可避免、只能妥协的感觉。就像此前的计算机一样，文化产业的从业者似乎很难、或许也无力设想自己能够在创作过程中完全避免与生成式人工智能模型互动。Sheila Heti 和刘宇昆等著名小说家已经开始尝试使用这些工具。 [Heti，2023年，Liu， 2023]对热门在线写作平台 Wattpad 上业余作者的一项调查显示，尽管多数用户对人工智能具备自主创造力持怀疑态度，但愿意将其作为创作辅助工具，例如用于编辑或构思故事创意。 [Wattpad博客] 随着时间推移，我们可以预期，作家将学会与人工智能共处，将其视为一种所谓的“协同智能”——一种可以与之协作的机器。 [Mollick， 2024]

文化研究学者日益认识到，除作家、读者、经纪人、编辑和出版商这些常见角色外，数字平台和人工智能机器学习算法也是当代文化与文学生产中的重要参与者。 [Vadde, 2021, McGurl, 2021, Seaver, 2022] 在本文中，我们建议将一个较不显眼、却同样影响深远的角色纳入考量：计算训练数据。正如 Stuart Hall 所言，文化生产的一个关键成果是文化再现，如今人工智能训练数据在这一过程中发挥着重要作用。 [Hall, 1997] 正如计算机科学家所指出的，我们在在线内容中发现的许多令人担忧的社会现实扭曲，都源于有缺陷、高度选择性或有限的训练数据（计算机科学研究中日益增多的文献将其称为“报告偏差”）。[Shwartz and Choi, 2020] 要弄清为何如此多的在线内容充斥着偏见、歧视和错误再现，就必须仔细分析为这些生成式人工智能系统提供动力的训练数据。因此，人工智能训练数据代表着文化生产领域中一个重要的新参与者，也推动着文化再现。

当然，对文化再现的分析与批判始终是文学研究学科的核心。举一个非常明显的例子：我们研究并批判《东方学》，也就是十九世纪晚期英国小说如何设定“东方”与“西方”之间的本体论差异，再通过将一系列意象与

叙事母题与被宣称属于这些类别的真实人群联系起来，使这种差异成为常态。 [Said, 1978] 当我们分析一部虚构作品，例如 E.M. Forster 的《印度之行》，以剖析这套文化再现体系如何运作时，我们会假定作品出自一位身处更广泛社会语境中的人类作者之手，并由其负责再现。这位作者有着可以通过批判性分析加以探究的动机与意图，而这些分析又会揭示更广泛的体系。批评者将语言与审美形式视为窥探塑造文本的动机、意图、偏见与删略之窗口——其依据是一种假设：语言的这些例子能够可靠地反映创作者身处社会语境之中的心智。

然而，人工智能训练数据生成虚构作品的方式，与人类作者生成虚构作品的方式大不相同。这是因为，人工智能训练数据代表着由人类生成文本（以及日益增多的人工智能生成文本）构成的庞大复合体或集合。在这些文本中，我们通常赋予人类作者的属性——并且这些属性随后成为分析其写作不可或缺的要素——如意图、欲望、情感等，会在促成这些数据集的策展与生产的计算过程中被重新混合。人工智能训练数据有时因其不可知性而被称为“黑箱”；对于那些未向公众发布的数据集而言，确实如此。但对于 BERT 这样的开放获取数据集，我们希望已经说明，可以通过算法与数据审计，部分了解这些数据，以及它们如何塑造语言生成。这只需要一种计算方法，来匹配创造这些材料的计算过程。在本文中，除了对生成式人工智能与虚构作品之间的关系提出一些新颖见解之外，我们还希望展示一种新的方法，用于更广泛地分析计算训练数据对当代文化的影响。

参考文献

Ann Banfield. 《不可言说的句子：虚构语言中的叙述与再现》. Routledge & Kegan Paul, 1982年。Nina Beguš. 实验性叙事：人类众包故事创作与人工智能故事创作的比较。*Nature:Humanities and Social Sciences Communications*, 11(1392), 2024年。Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 论随机鹦鹉的危险：语言模型会不会太大？载于*ACM*公平性、问责制与透明度会议论文集，第610–623页，2021年。Yoshua Bengio, Éric Thibodeau-Laufer, Guillaume Alain, and Jason Yosinski. 可通过反向传播训练的深度生成随机网络。载于《机器学习研究论文集》，第2期，第226–234页，2014年。Ruha Benjamin. 《技术之后的种族》. Polity Press, 2019年。Peter Brooks. 《为故事所惑：叙事的用途与滥用》.《纽约书评》，2021年。Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 从语言语料库自动推导出的语义包含类人偏见。《科学》, 356(3664):183—-186, 2017年。Myra Cheng, Esin Durmus, , and Dan Jurafsky. 标记化角色：运用自然语言提示测量语言模型中的刻板印象。载于*ACM*公平性、问责制与透明度会议论文集，第1504–1532页，2023年。Casey Christie. 《夜》. Smashwords, 2012年。Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT：用于语言理解的深度双向变换器预训练。载于北美计算语言学协会会议：人类语言技术论文集，第4171–4186页，2019年。Wai-Chee Dimock and Lauren Goodlad. 人工智能与人类。《现代语言协会会刊》, 136(2): 317–319, 2021年。Josh Dzieza. 人工智能的虚构盛景：高速半自动类型小说的奇异世界。*The Verge*, 2022年7月。网址：https://www.theverge.com/c/23194235/ai-fiction-writing-amazon-kindle-sudowrite-jasper。Michele Elam. 诗歌不会被优化；或者说，文学对于人工智能意味着什么？《美国文学》, 95(2):281–303, 2023年。Rita Felski.《着迷：艺术与依恋》. 芝加哥大学出版社，2020年。Catherine Gallagher. 虚构性的兴起。载于Franco Moretti（编），《小说：第一卷》. 普林斯顿大学出版社，2006年。Shirin Ghaffray. 什么能阻止人工智能用虚假图像淹没互联网? *Vox*, 2023年6月。网址：https://www.vox.com/technology/23746060/ai-generative-fake-images-photoshop-google-microsoft-adobe。Lauren Goodlad. 编者导言：让人文学科参与其中。*Critical AI*, (1–2), 2023年。

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamil˙e Lukoši¯ut˙e, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 运用影响函数研究大型语言模型的泛化。*arXiv*, 2023年8月。Stuart Hall. 引言。载于《表征：文化表征与意指实践》. Sage, 1997年。Beth Hart. *Eddie Langsett*. Smashwords, 2013年。Katherine Hayles. 人工智能的心智内部：物质性与表征危机。《新文学史》, 53(4–5): 635–666, 2023年。Lucas Torroba Hennigen and Yoon Kim. 从掩码语言模型推导语言模型。载于计算语言学协会年会论文集，第1149–1159页，2023年。Sheila Heti. 关于Alice。《纽约客》，第48–53页，2023年11月。Matthew Kirschenbaum. 特殊行为：在循环神经网络中阅读形式。《英国文学史》, 88(2): 361–386, 2022年。Matthew Kirschenbaum. 再谈理论：随机鹦鹉时代关于语言、意义与意图的论坛。《批判性探究》："当下"论坛，2023年6月。Neils Knudsen. 《伦多尔的歌唱石》. Smashwords, 2014年。刘宇昆. 机器中的魔法。《猎户座》杂志, 2023年夏。Mark McGurl.《一切与更少：亚马逊时代的小说》. Verso, 2021年。Ethan Mollick.《协同智能：与人工智能共同生活和工作》. 兰登书屋，2024年。Henrik Skov Nielsen, James Phelan, and Richard Walsh. 关于虚构性的十个论点。《今日诗学》, 23(1):63, 2015年。James Phelan. 虚构性、受众与人物。《今日诗学》, 39(1):113–129, 2018年。Andrew Piper. 人物在做什么？虚构人物的具身能动性。《计算文学研究杂志》, 2(1):1–12, 2024年。Andrew Piper and Sunyam Bagga. 迈向叙事性的数据驱动理论。《新文学史》, 53–54(4–1): 879–901, 2023年。Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 通过生成式预训练提升语言理解能力。*OpenAI*, 2018年6月。网址：https://openai.com/research/language-unsupervised。Alex Reisner. 揭秘：哪些作者的盗版书籍正为生成式人工智能提供动力。《大西洋月刊》（网络版）, 2023年8月。网址：https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/。Edward Said.《东方学》. 兰登书屋，1978年。John Searle. 虚构话语的逻辑地位。《新文学史》, 6(2):319–332, 1975年。Nick Seaver.《计算品味：算法与音乐推荐的创造者》. 芝加哥大学出版社，2022年。Claude Shannon. 通信的数学理论。《贝尔系统技术杂志》, 27:379—-423, 623—-656, 1948 年。Vered Shwartz and Yejin Choi. 神经语言模型能否克服报告偏差？载于国际计算语言学会议论文集，第6863–6870页，2020年。J.D. Stroube.《囚于黑暗》. Smashwords, 2011年。Rachel Tatman. Kaggle咖啡闲聊：Jacob Devlin（谷歌研究员、BERT作者），2019年3月。网址：https: //www.theverge.com/c/23194235/ai-fiction-writing-amazon-kindle-sudowrite-jasper。W.L. Taylor. "完形填空法"：一种衡量可读性的新工具。《新闻学季刊》, 30:415–433, 1953年。Christina Tetreault.《爱的勇气》. Smashwords, 2014年。Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, and Guillaume Lample. LLaMA：开放且高效的基础语言模型。*Meta Research*, 2023年2月。网址：https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/。Ted Underwood. 理论的经验性胜利。《批判性探究》："当下"论坛，2023年6月。Aarthi Vadde. 出版商还是平台。《现代语言协会会刊》，第455–462页，2021年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。《注意力就是一切》。载于神经信息处理系统国际会议论文集，第 6000–6010 页，2017年。Blakey Vermeule。我们为何关心文学人物？约翰斯·霍普金斯大学出版社，2010年。Richard Walsh。《虚构性的修辞：叙事理论与虚构观念》。俄亥俄州立大学出版社，2007年。Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R. Bowman。GLUE：用于自然语言理解的多任务基准与分析平台。载于*EMNLP*研讨会论文集：*NLP*黑箱——分析与解读自然语言处理神经网络，第 353–355 页，2018年。Wattpad博客。《虚构作品的未来：Wattpad研究揭示阅读习惯的代际转变，以及对人工智能进入出版业持怀疑态度》，2024年1月。Takateru Yamakoshi、Thomas Griffiths 和 Robert Hawkins。利用连续复现链探究 BERT 的先验。载于计算语言学协会年会论文集，第 3977–3992 页，2022年。Tom Young、Yunan Chen 和 Yang You。掩码语言模型所学习的条件句中的不一致性。*arXiv*，2024年2月。Yukun Zhu、Ryan Kiros、Richard Zemel、Ruslan Salakhutdinov、Raquel Urtasun、Antonio Torralba 和 Sanja Fidler。对齐书籍与电影：通过观看电影和阅读书籍实现故事化视觉解释。载于国际计算机视觉会议，第 19–27 页，2015年。