智能社会研究

（双月刊）

2024 年
第 1 期
2024 年 1 月 10 日出版

2022 年 11 月 10 日创刊

总第 8 期

目　次

论文

研究报告

译文

书评

访谈

CONTENTS

THESIS

RESEARCH REPORTS

TRANSLATED TEXT

BOOK REVIEW

INTERVIEW

大语言模型能用来给主题模型中的主题编码吗？*

——兼论大语言模型在社会科学领域应用的前景

李　代　张博伦　周浥莽**

摘要：ChatGPT 是由 OpenAI 公司开发的基于大语言模型的通用应用程序，可以用来完成自然语言处理方面的任务。社会科学的研究实践涉及大量自然语言处理的任务，因而在这方面 ChatGPT 可能有广阔的应用前景。本文以主题模型为例，探讨 ChatGPT 能否为主题模型的结果生成可信的标签。我们抽取了发表在中外社会学期刊上使用了主题模型的论文，并打乱其主题与标签的顺序，在中外网络平台进行问卷调查，由一般用户评价二者谁更可能反映了原文的主题。结果表明，一般用户对原论文给出的标签评价并没有显著高于 ChatGPT 给出的标签评价，甚至在多数主题上 ChatGPT 给出的标签获得的评价更高。这说明 ChatGPT 可以用于给主题模型的结果生成标签，给研究者加以评判。不过，这也意味着即使缺乏领域专门知识，ChatGPT 的表现在有些条件下也可能达到专业研究者的水平，为社会科学研究带来了新的挑战。

关键词：自然语言　社会学方法　标签　人工智能

2022 年，OpenAI 公司发布人工智能产品 ChatGPT，其语言交互能力几乎可以达到以假乱真的程度，引发了社会的广泛关注。ChatGPT 可能是第一个被大量普通用户认为具有“通用人工智能”（artificial general intelligence）潜力的产品。一经推出，用户就开始用它搜索信息、翻译语言、编辑稿件、撰写代码，并发现很多时候其输出结果稍加修改就可以达到令人满意的水平。ChatGPT 实际上

* 本文系中国政法大学科研创新项目（项目批准号：10822521）的阶段性研究成果，并受中央高校基本科研业务费专项资金资助。感谢范新光、周涵、徐丽娟、瞿璧瑜、李月旻等师友在本文写作过程中的帮助、建议。

** 李代，中国政法大学社会学系。张博伦，浙江大学社会学系。周浥莽，中国人民大学社会学院。

是一个易用的人机交互接口,用于与该公司训练的“生成性预训练变形器”(generative pre-trained transformer, GPT)模型进行交互。由于这类模型包含的参数达到数十亿规模甚至更大,因此被称为“大语言模型”(large language model, LLM)。对于内容生产工作而言,类似 GPT-3.5、GPT-4 这样的大语言模型能否带来革命性的影响?或许它将大大提高内容生产的效率,而同时带来失业的危机;或许它最终被证明无法克服目前的弊端(如它有时会自信地给出事实性错误的答案),因而不堪大用。这些问题有待社会科学经验研究的探索。

在展开对大语言模型的经验研究与理论反思时,社会科学可以首先考虑其对自身知识生产的影响,因为知识生产亦是一种内容生产。以 ChatGPT 为代表的大语言模型可能会给现有的社会科学研究模式带来莫大机遇,也可能带来巨大挑战。例如,通过 ChatGPT 进行对某个领域的文献综述,可以帮助初学者快速掌握关键文献;通过 ChatGPT 对论文进行文字编辑,可以使其更加文从字顺乃至符合学术话语规范。《自然》子刊《生物医学工程》的编辑表示,他们已经在用 ChatGPT 辅助编辑工作,而这也意味着他们将更加严肃地思考综述性文章的价值,对论文的语言运用也有更高的标准(Prepare for Truly Useful Large Language Models, 2023)。毕竟,ChatGPT 可以自动完成信息整合类的综述,也能用规范的语言写作。想必在不远的未来,期刊编辑将面临一个难题:如何判断自己接到的投稿是人写的还是模型生成的?既然大语言模型几乎一定会对社会科学研究产生影响,那么我们应该如何将大语言模型融入社会科学研究呢?

已有文章较为宏观地探讨了计算方法在社会科学中的应用(陈云松、吴晓刚、胡安宁等,2020;周涛、高馨、罗家德,2022),本文则试图在其基础上以小见大,尝试将大语言模型用于一个特定的社会科学研究方法上:给主题模型的结果打标签。为此,本文选取了 2013 年以来在各大中文核心期刊和主要英文期刊上发表的采用了主题模型方法的社会学论文,将 ChatGPT 所打的标签与原文中的标签进行对比,并通过在线调查探讨其说服力:一方面,这能切实地为主题模型在应用中存在的实际问题提供解决方案;另一方面,这番论述以小见大,揭示了当前社会科学研究者面临的方法上的“失控”危机。下面,本文将首先介绍主题模型方法,再陈述大语言模型用于给主题模型的结果打标签是否可信的研究过程,最后对“失控”进行分析。

一、文献综述

(一) 主题模型

主题模型(topic model)(Blei, 2003)最早由计算机科学家开发,后来被引入社会学研究。它基于词汇在语料中共同出现的频率及不同模式,将词语分为互不排斥的若干组。举例而言,体育新闻报道中“奥运会”“金牌”“训练”等词汇经常会共同出现在同一文本中,因此很可能被分入同一组。由于经常在某些文本中共同出现而分入一组,因此同一组词语往往被理解为一个具有实际意义的“主题”。主题模型尽管并不理解每一个主题是什么,但是能识别出词汇共同出现的不同模式,从而将各个潜在的主题区分、识别出来。

常用的主题模型采用潜在狄利克雷分配(latent dirichlet allocation, LDA)模型。这种模型将语料分为三个层次——文档、主题、词语——其中文档和词语是已经被观察到的。这种模型假定:主题在文档中的分布、词语在主题中的分布是两个分别由潜在狄利克雷分配随机生成的多项分布。一个文档按照一定概率产生出若干主题后,每个主题又按照一定概率产生若干词语,最终形成这一文档中被观察到的一组词语及其频率。研究者一般使用“吉布斯采样”(Gibbs sampling)来进行参数估计(Blei,2012)。

上述模型是一种“非监督”机器学习方法。“非监督”指不需要研究者向其提供已经打好标签的训练数据,模型就可以直接对文本展开分析。但是在使用主题模型时,研究者需要首先设定一系列超参数,如假设语料中存在 k 个主题,然后才能由主题模型进行拟合。k 的大小本身没有特别客观的依据,取决于研究者对文本的认识和对初步结果的评估。k 如果太小,可能会导致一个主题内部的特征词看上去不像是来自同一主题;k 如果太大,则可能会出现很多难以解读的主题,而且也会加大解读成本。

在主题模型的具体应用中,研究者如何确定一个主题是什么?令全部语料中存在的词的总数量为 j,那么一个主题就是这 j 个词以及标识这 j 个词在该主题中出现概率的向量。k 个主题并列,就得到了主题模型的输出结果之一——

一个有 k 列、j 行的矩阵，其中每一列是一个主题，每一个主题下有 j 个元素，分别表示语料中所有词在该主题中的出现概率。研究者将所有词按照在某一主题中的出现概率从大到小排列，可以通过阅读该主题中出现频率最高的词对这一主题的含义有所推测和解读。由于主题模型允许某些词在多个主题中都成为高频词，因此它比简单的词频分析更符合人类语言的特征。

在主题模型中，一篇文章并不只有一个主题。恰恰相反，主题模型的一大优势就是它允许一篇文章有多个主题，更能适应文本的复杂性。若语料中有 n 个文本，则主题模型将输出一个有 k 列、n 行的矩阵，每一行是同一个文本里包含的 k 个主题各自的比例，总和为 1。不难理解，对于这些文本而言，主题模型得到的各个主题的比例大小就是其文本的一些特征，可以用来进行进一步的分析。这相当于一种对文本特征进行降维的算法，如以词频计，原来文本有 j 个特征（因为语料包含 j 个词，一篇文章中每个词都对应着一个词频），而现在降维成了 k 个特征。

社会科学研究者在使用主题模型时，往往根据一个主题里概率或者其他统计量最高的前 10—100 个词来判断这个主题是什么。这些词被称作“特征词”，是给主题编码的基础。主题编码并非主题模型的内在部分，对用模型进行预测这一工具开发的初衷而言也并非必要（Blei，2003），但对社会科学研究却可能至关重要。这是由社会科学研究与产业界任务性质的差别决定的。例如，一家互联网企业可能会依据文档中不同主题的比例建立预测模型，用来识别不同类型的文本：甲主题比例较高的是体育新闻，乙主题比例较高的是娱乐新闻，等等。只要模型预测表现较好，甲主题、乙主题到底是什么就无关紧要，可以留在“黑箱”里。但是对于社会科学而言，研究者往往需要人为地给主题打上相对可信的标签，否则无法进一步对研究关心的理论问题展开讨论。这时，主题模型的作用在于阐释文本，并没有被用作预测模型。

例如，迪马鸠（P. DiMaggio）等人深入研究了新闻如何报道政府对艺术的支持情况。他们根据各个主题中出现频率最高的 100 个词，将 12 个主题分别归纳为“冲突”（conflict）、“地方政府项目和资助”（local government projects and funding）以及“特定的艺术流派、事件或赞助目标”（specific arts genres，events，or grant purposes）等标签。这些标签不仅具体明确，而且具有深刻的理论内涵

(DiMaggio, Nag & Blei, 2013)。

(二)主题模型的编码问题

主题模型的编码问题是不少方法论研究的焦点。在主题模型最初被引入社会学时,研究者认为它的一大重要优势在于可以利用机器自动给文本进行编码,从而尽可能地规避编码者对编码过程的人工干预(Mohr & Bogdanov, 2013)。在主题模型出现之前,研究者如果要对大量文本进行编码,首先需要通读文本来建立对文本统贯的理解,进而将这一理解整理成指导后续编码的规则手册。也就是说,原来的编码流程是"诠释,然后计数"(Mohr & Bogdanov, 2013)。主题模型则扭转了这个顺序,它先会输出一个结果给研究者,研究者再赋予这些结果以相应的标签。这么一来,主题模型"使得研究者可以在将他们的前见用于分析之前就发现文本结构"(Mohr & Bogdanov, 2013)。这个立场也被很多数码人文学者所接受(Heuser & Le-Khac, 2012; Meeks & Weingart, 2012)。

但是在这个过程中,各个主题的编码由人工给出,存在大量的自由裁量空间。首先,研究者在给这些主题打标签时,可能带入"确认偏误"等主见。其次,有时主题模型输出的结果一时难以解读,可能是无意义的垃圾主题(Marshall, 2013),也可能这些陌生的结果其实是具有理论意义的新异主题,这恰恰是主题模型的迷人之处(Buurma, 2015)。总之,有必要对研究者凭借直觉给主题打的标签进行评估,如将主题标签同外部事件进行对比,通过文本细读进行专家判读,或者通过众包对主题的关键词与标签之间的关系进行质量判别(Grimmer & Stewart, 2013; Ying, Montgomery & Stewart, 2021)。这些策略的问题在于其成本相当高昂,应用规模难以扩大。在这个问题上,大语言模型提供了一个成本相当低而效果可能足够好的解决方案。

(三)大语言模型与主题模型

要理解大语言模型,首先要从更基本的语言模型讲起。一个相当简化的语言模型是"词袋"(bag-of-words),它认为文本的含义可以由组成它的词和词组的频率来表示。每一个词都是"独热"(one-hot)向量,是与其他词完全不同的特征。所谓"独热",指的是如果一个数据集有三个维度,那么[1,0,0]、[0,1,0]或

[0,0,1]就是独热向量,每两个独热向量之间的余弦相似度为 0。在这一模型中,词的顺序没有意义。例如,“喜欢”和“喜爱”没有关系,“我—喜欢—猫”和“猫—喜欢—我”没有区别。主题模型背后的语言观正是“词袋”。

在此基础上,研究者希望能将词之间的关联也加入考量,如词嵌入模型(word embedding)。词嵌入模型的思想是将一个词语的各种意涵投射在一个超高维度空间,再降维成一个人类可解读的多维向量。结果是,意涵相似的词语会在多维空间中处于相近位置,词语之间的关系(如反义、等级、顺次)可以通过其在特定维度上的向量关系表现出来。词嵌入模型的两种常见算法(Word2Vec 和 GloVe)都基于词与上下文的共现关系,改变了“词袋”语言模型对词语关系与顺序的漠视。

词嵌入模型进行文本分析时,可以处理不同社会范畴与阶级、性别、种族等社会类别之间的对应关系(比如“戏曲—流行乐”与“老年—青年”之间的对应关系)。例如,对过去百年间数百万种出版物进行词嵌入模型分析,会发现文化范畴的阶级属性在过去一个世纪间保持相对稳定(Kozlowski, Taddy & Evans, 2019)。已有相当数量的学者使用了词嵌入模型进行社会学研究(Stoltz & Taylor, 2021; Li, Wu & James, 2020; Caliskan, Bryson & Narayanan, 2017)。

然而,词嵌入模型本身在处理较“小”的语料时却存在较大的偏差。因此,词嵌入模型必须依赖预训练的大语言模型才能达到较好的效果。这些模型的代表是生成式预训练模型、ELMo(embeddings from language models)模型、BERT(bidirectional encoder representations from transformers)模型。这三个模型不仅将上下文的关系加入训练中,而且还加入了特定的记忆力/注意力机制。

大语言模型的规模相当惊人。GPT-1.0 是基于 BooksCorpus 数据库进行训练的,拥有 1.17 亿的参数;BERT 在这个数据库的基础上增加了英文维基百科的数据,参数量增加到了 3.4 亿。研究者发现,在预训练范式中,模型的规模、训练语料的规模与模型最终在各个任务上的表现都有着非常直接的关联(Delvin, Chang & Lee et al., 2019; Kaplan, McCandlish & Henighan et al., 2020)。Transformer 构架恰恰提供了让模型迅速变大的技术框架。于是,在 2018 年后的几年内,模型使用更多的数据,进行更多的堆叠,不断变大(Bender, Gebru & McMillan-Major et al., 2021)。ChatGPT 所基于的 GPT-3.5 Turbo 已

经拥有了2000亿的参数,并且在诸多任务上展现出了泛化的能力。比起传统的词嵌入算法,大语言模型极大地扩张了训练语料和参数的规模,使信息中更多的维度被保留下来。以ChatGPT为代表的"聊天机器人"这一交互形式,涌现出了解决复杂问题的能力。

不过,目前大语言模型还没有完全取代主题模型这类小模型。作为社会科学研究者,我们仍然希望对具有特定时空条件限制的文本进行考察,以理解文本在特定语境中的意义。那么,我们能否将大语言模型引入主题模型的应用中,帮助主题模型克服编码的困境呢?具体而言,当我们把主题模型的结果输入大语言模型之中时,大语言模型能否像研究者一样,根据它自己预训练掌握的语言模式,给主题模型的结果打上可信的标签呢?下面,我们将对这一技术的潜能进行更具体的分析。

二、数据与方法

本文的目的是评估大语言模型基于主题模型的结果进行编码的能力。如前所述,对于研究者而言,用于确定每个主题应如何编码的信息主要是主题模型输出的一个k列、j行的矩阵,研究者通过阅读每个主题中的若干特征词为其打上标签。采用主题模型进行研究的较为底线的惯例是向读者提供每个主题中靠前的特征词,以便读者判断研究者的编码是否可以接受(Ying, Montgomery & Stewart, 2021)。遗憾的是,并不是每一篇得到发表的社会学论文都能达到这一底线要求。

如果把编码的任务交由大语言模型完成,是否能得到可信的结果呢?把前面所说的每个主题下排序靠前的特征词输入大语言模型,大语言模型能否像人类解读者一样给出看似合理的标签?

为此,我们检索社会学中英文期刊发表的使用主题模型的论文,从中提取两方面信息:论文中主题模型的结果包含的特征词、原研究者给这个主题的标签(原文标签)。我们将特征词输入GPT-3.5,将生成的新标签(ChatGPT标签)与原文标签进行比对。为了评判新标签是否可信,我们通过两个在线调查平台(Credamo和Mechanical Turk)发布了匿名评估问卷,请网络用户在不知情的情

况下评价原文标签和 ChatGPT 标签哪个更有可能反映了原文的主题。

（一）原文主题与原文标签

我们检索了截至 2023 年 3 月社会学领域使用主题模型的已发表的中英文论文。2013 年 12 月，《诗学》（*Poetics*）出版了一期特刊，发表了若干篇采用主题模型进行研究的社会科学论文。这可能是社会科学领域最早集中发表采用主题模型进行社会科学研究的一期特刊。以此为起点，我们以“主题模型”（topic modeling）为关键词，检索了迄今在以下期刊发表的使用了主题模型的经验研究。中文社会学核心期刊：《社会学研究》《社会》《社会学评论》《社会发展研究》《中国青年研究》《青年研究》《妇女研究论丛》《人口学刊》《人口研究》《人口与发展》《人口与经济》《中国人口科学》。英文社会学期刊：《美国社会学评论》（*American Sociological Review*）、《美国社会学刊》（*American Journal of Sociology*）、《社会学方法和研究》（*Sociological Methods & Research*）、《社会学方法论》（*Sociological Methodology*）、《社会力量》（*Social Forces*）、《理论与社会》（*Theory and Society*）、《诗学》（*Poetics*）。

这番检索共发现 8 篇中文文献、50 篇英文文献。当然，这可能会有遗漏，如初次检索遗漏了 Nelson（2021），我们也在力所能及的范围内予以补齐。并不是每一篇论文都在正文中刊登了主题模型的结果。为了便于分析，本文仅选取在正文中刊登了输出结果并对主题进行了编码的文章进行研究。以此为标准进行筛选，得到 2 篇中文文献、18 篇英文文献。其中，有一篇英文文献处理的文本是中国历史典籍，因而语料是繁体中文（Miller，2013）；另有一篇中文文献处理的是英文文本，但用中文给主题打标签（郭台辉、周浥莽，2020）。最后，我们一共获得了 378 个主题以及相应主题的特征词和作者给予的原文标签。

（二）大语言模型与 ChatGPT 标签

在大语言模型方面，为了模拟人类阅读主题模型结果给出主题标签的过程，我们使用的是基于 GPT-3.5 Turbo 模型的 ChatGPT 对话任务。ChatGPT 采用基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF），在 GPT-3.5 Turbo 模型的基础上对参数进行进一步微调。在其开发者

OpenAI 公司公布的 InstructGPT 的训练方案中,OpenAI 通过人类反馈的方式,构建了一个标注数据集和一个对于回答的答案进行排序的数据集,并通过这些人类反馈的数据对 GPT-3 模型进行了微调,训练出 InstructGPT。通过微调,InstructGPT 能比原来的 GPT 模型更好地处理人类使用者给出的提示词。ChatGPT 的训练方式与之相似。正是这个交互模式,使得 ChatGPT 展现出很强的任务泛化能力(Ouyang, Wu & Xu et al., 2022; OpenAI, 2022)。在本文写作时,GPT-4.0 已经推出,但本文作者尚未申请到使用权限。

在操作过程中,我们首先生成标准化结构的英文提示语(prompt),再通过 Python 接入 ChatGPT 提供的端口将提示语输入模型,并存储其返回的结果。使用接口是为了让电脑自动完成问答任务从而节省时间;如果不使用接口,普通用户也可以登录 ChatGPT 的界面、输入提示语、得到 ChatGPT 标签,只不过这个过程不够自动化。

由于 ChatGPT 的底层模型是概率模型,而且提示语的变化可能引起答案变化,因此研究应当忠实地记录研究过程中使用的提示语,而不是将其视为技术上的细枝末节。我们按照以下步骤输入提示语。首先,我们输入"你是一个标注主题模型输出结果的专家"(You are an expert in labeling outputs from topic models)来初始化系统信息。其次,我们说明:"你能依照关键词来帮我给几个来自一个主题模型中的主题标注标签吗? 这些主题的关键词如下:{特征词}。"(Could you please help me with labeling the following topics based on their top terms? The top terms fo a few topics are listed below:{特征词}.[①])我们又规定了对话模型应该返回的模型格式:Your answer should only consist of the best label for each topic in the following format:"Topic N:{label}" on a separate line for each topic. 由此,我们获得了 ChatGPT 给出的主题标签。如果改变提示语,确实可能导致结果变化,不过本文的焦点并非怎样的提示语最有效。限于篇幅,本文暂且不对此多加讨论。

通过这样的方式,我们让 ChatGPT 为前述所有主题给出标签。在此选出全

① 在原本的编码中就存在 fo 的拼写错误,但并没有影响我们得到结果,这里予以忠实记录而不加修改。

部三篇使用了中文语料或中文标签的论文，并将其中涉及的主题特征词、原文标签和 ChatGPT 标签展示如下（表 1）。此处主题编号之所以不连续，是因为相应的编号在原数据中被用作他用，而非缺失了主题。

表 1　涉及中文的三篇论文中主题模型的特征词、原文标签和 ChatGPT 标签

编号	论文名	特征词	原文标签	ChatGPT 标签
2	“黑箱”中的青年	抖音、微信、B 站、淘宝、小红书、知乎、京东、今日头条、朋友圈广告、豆瓣、快手	媒体平台	社交媒体和电商平台
3	“黑箱”中的青年	大数据、信息茧房、杀熟、信息壁垒、精准推荐、曝光度、流量池、监听、深陷、深度学习、神经网络	专业知识	大数据和个性化营销
4	“黑箱”中的青年	类似、历史、浏览记录、监控、偷听、热门、屏蔽、偷窥、窃听、总是、反复	运作机制	监控和隐私担忧
5	“黑箱”中的青年	广告、个性化、精准、投放、隐私、一致化、知识孤岛、定制、投送、趋于、推广	运作影响	在线广告推送和个性化
6	“黑箱”中的青年	推荐、喜欢、内容、隐私、方便、信息、感兴趣、符合、信息茧房、时间、推送、容易、娱乐、喜好、需求、侵犯、淘宝、音乐、商品、泄露、窥探、搜索、监视、有好有坏、中立、满足、兴趣、有利有弊、单一	娱乐内容	用户偏好和隐私
7	“黑箱”中的青年	信息、信息茧房、资讯、新闻、感兴趣、算法、隐私、搜索、容易、观点、全面、符合、侵犯、消息、客观、平台、兴趣、获取、浏览、同质化、重复、认知、泄露、监视、局限、热点、社会、标题、软件、垃圾、极化	资讯内容	信息隐私和在线平台消费
8	“黑箱”中的青年	广告、不喜欢、隐私、讨厌、推荐、内容、朋友圈、算法、推送、感觉、侵犯、反感、投放、精准、很烦、太多、不感兴趣、泄露、监视、软件、平台、体验、烦人、消费、厌恶、厌烦、社交、浪费时间、用户、重复	广告内容	对广告推送和侵犯隐私的负面态度

（续表）

编号	论文名	特征词	原文标签	ChatGPT 标签
10	Rebellion, Crime and Violence in Qing China	ni 拟、zui 罪、an 案、xing 刑、bu 部、shen 審、li 例、zhao 照、zhe 斩、si 死、yi 議、jiao 絞、shang 傷、ding 定、ming 命、gou 勾、jian 監、lv 律、ou 歐、sha 殺、gai 改、li 理、bi 斃、qiu 秋、shi 氏	crime	criminal justice/ system
11	Rebellion, Crime and Violence in Qing China	huo 獲、na 拏、fei 匪、yan 嚴、fan 犯、di 地、bu 捕、ji 緝、dao 盜、chi 飭、an 案、tu 徒、xian 縣、jie 劫、shou 首、guan 官、cheng 懲、qiang 搶、tao 逃、min 民、jing 經、zi 滋、du 督、zei 賊、bian 弁	unrest	crime and law enforcement
12	Rebellion, Crime and Violence in Qing China	huo 獲、na 拏、xun 訊、gong 供、an 案、jiao 教、jie 解、yan 嚴、shen 審、jiu 究、jing 經、ni 逆、li 李、xian 現、sheng 省、xian 縣、wang 王、zhang 张、liu 劉、jia 家、mi 密、shou 首、que 確、wang 往、fang 訪	sedition	legal proceedings
13	Rebellion, Crime and Violence in Qing China	zei 賊、bing 兵、fei 匪、jiao 剿、ni 逆、cheng 城、gong 攻、jun 軍、li 力、fu 復、cuan 竄、yong 勇、guan 官、jing 經、dai 帶、ji 擊、lu 路、huo 獲、sheng 勝、ke 克、du 督、bao 保、shou 首、shu 數、sha 殺	rebellion	military operations
14	Rebellion, Crime and Violence in Qing China	bing 兵、jiao 剿、jun 軍、zei 賊、fei 匪、fang 防、cuan 竄、dai 帶、li 力、chi 飭、xian 現、lu 路、bei 北、zhou 州、ni 逆、yong 勇、nan 南、sheng 省、dong 東、cheng 城、pai 派、yan 嚴、sheng 勝、xun 迅、du 督	major rebellion	military action and tactics in ancient China
16	历史社会学能化解学科之争吗	social、history、culture、politic、society、use、term、history、study、one、universal、subject、new、can、theory	理论与方法论	文化和社会研究
17	历史社会学能化解学科之争吗	one、use、term、time、new、subject、people、see、culture、content、like、work、life、history、way	文化研究	生活方式和文化

（续表）

编号	论文名	特征词	原文标签	ChatGPT 标签
18	历史社会学能化解学科之争吗	use、population、data、term、content、census、age、subject、year、rate、city、number、difference、economic、century	量化史学	人口特征和数据分析
19	历史社会学能化解学科之争吗	state、govern、new、public、policy、social、develop、nation、institution、politic、economic、city、universal、power、local	制度演变	公共政策与治理
20	历史社会学能化解学科之争吗	land、product、market、labor、use、economy、agriculture、peasant、industry、farm、rural、economy、term、capital、village	政治经济研究	农村经济
21	历史社会学能化解学科之争吗	nation、politic、state、people、core、Jewish、new、violence、movement、Jew、term、one、right、identity、group	民族/国家	犹太民族主义运动中的政治认同和暴力
22	历史社会学能化解学科之争吗	worker、class、social、politic、labor、union、work、German、revolution、industry、state、use、term、content、movement	革命/劳工	劳工与社会运动
23	历史社会学能化解学科之争吗	law、India、court、India、legal、case、century、use、state colony、one、subject、govern、British、author	法律研究	印度法律体系
24	历史社会学能化解学科之争吗	Islam、state、Muslim、Chinese、term、China、Ottoman、core、universal、use、new、subject、religious、empire、nation	亚洲研究	穆斯林帝国和中国
25	历史社会学能化解学科之争吗	politic、party、election、vote、use、democrat、state、new、term、content、republican、support、American、subject、elector	民主研究	政治和选举
26	历史社会学能化解学科之争吗	women、family、children、men、marriage、household、work、social、female、gender、male、marry、live、age、class	性别与婚姻	性别和家庭动力学

（续表）

编号	论文名	特征词	原文标签	ChatGPT 标签
27	历史社会学能化解学科之争吗	colony、slave、African、Africa、British、European、French、term、new、trade、subject、state、core、Indian、world	帝国研究	跨大西洋奴隶贸易和殖民主义
28	历史社会学能化解学科之争吗	American、black、white、school、education、new、south、universal、racial、African、race、immigration、York、work、class	种族研究	美国的民权运动与教育
29	历史社会学能化解学科之争吗	health、medical、disease、death、body、medicine、sexuality、social、practice、century、public、cause、report、new、work	卫生/健康	公共健康与医药
30	历史社会学能化解学科之争吗	Mexico、Spanish、Latin、state、term、use、new、Mexican、America、see、Brazil、Indian、city、Spain、American	拉美研究	拉丁美洲研究

（三）可信性评估

如何评估 ChatGPT 标签的可信性呢？我们设计了两个简单的调查问卷进行在线发放，请普通人对两类标签进行评估。由于论文分中英双文，我们分别在中美进行了调查。首先报告中文调查的过程。表 1 报告了两篇中文论文和一篇涉及中文语料的英文论文，我们将其结果进行整合，通过 Credamo 平台设计成电子问卷，再通过微信向联系人分发。问卷设计如下。

我们将调查题目设为“网络碎片化信息的认知”，并将调查任务伪装成评估两个人工标签。例如，在第一篇论文（赵龙轩、林聪，2022）方面，我们询问：“我们从一些长文章中每篇提取 10—20 个高频关键词，请 2 位来自不同专业的本科生仅通过阅读这些高频关键词来推断这篇文章具体的主题。您认为哪种更可能反映了原文的主旨？如果两个差不多好，请选择‘差不多’。”再下设若干道小题，每一道小题列出主题模型的特征词，让答题人选择原文标签、ChatGPT 标签或是“差不多”。为了避免选项顺序对答题人造成干扰，我们随机调整原文标签和 ChatGPT 标签出现在选项中的顺序。这样，如果答题人以完全随机的方式或

者固定选择选项的方式答题,两类标签被选中的期望概率就都是 50%。

依照同样的结构,我们继而询问后两篇论文,所用设问方式分别为:“我们收集了关于清朝中国的一些史料,统计每篇史料中的高频关键词。另请历史系的 2 位美国留学生通过这些高频关键词来推断史料的具体主题。”“我们收集了一些历史社会学论文,统计每篇论文中的高频关键词。另请社会学系的 2 位同学通过这些高频关键词来推断论文的具体主题。”

之所以不告知答题人我们旨在研究论文的原文标签和 ChatGPT 标签,是为了避免答题人受此干扰。如果答题人对 ChatGPT 这类技术有一定的立场,则可能会因为有标签是 ChatGPT 给出的而改变评估标准。需要指出的是,“从长文章中提取 10—20 个高频关键词”“通过关键词来推断这篇文章具体的主题”,并不符合主题模型实际运作的机制。主题模型允许一篇文章有多个主题,允许主题共享高频特征词,研究者在给主题打标签时也并不仅有特征词这一信息。但是,如果完全依照主题模型运作的机制询问答题人,对于不熟悉主题模型的答题人来说理解的难度太大,而且对我们的研究目的来说恐怕也没有什么帮助。因此,我们采取了“通过高频关键词还原文章主题”的故事,希望答题人能够轻易理解其任务而又不偏离我们的研究目标。

该研究共回收 33 份问卷。由于研究者的社交网络有一定特殊性,因此答题人以在校本科生和硕博研究生、人文学科或社会科学背景者为主。由于该研究并非一般意义上的社会调查,也没有可信的理由认为受访者的社会经济地位等属性会显著影响其评估工作,因此受访者的个人情况尽管不具有对总人口的代表性,也不妨碍调查结果的可信性。

英文调查在 Mechanical Turk 众包平台进行。与 Credamo 不同,该平台可以从其用户库中调用答题人,不需要研究者自行通过社交网络寻找答题人。我们设计的调查任务结构与中文调查相似,在此不再赘述。

由于在这个平台上找人答题需要支出费用,因此我们限定了英文文章主题的数量。首先排除前面已经调查过的使用中文语料的英文文章。对于剩余的 17 篇英文论文,我们首先排除了一篇语料是意大利语的文章。对于 2021 年 9 月之前发表的 13 篇文章,我们通过等权重的简单随机抽样方式抽取了 8 篇文章,再保留 2021 年 9 月之后发表的 2 篇文章。之所以要考虑 2021 年 9 月这一

断点,是因为 GPT-3.5 的训练数据截止到 2021 年 9 月,我们想知道对于一定在其训练数据之外的论文,其表现是否有差别。对于这 10 篇文章,如果文章报告的主题数量少于或等于 10 个,我们抽取全部的主题;如果文章报告的主题数量多于 10 个,我们从中等权重随机抽取 10 个主题。由此,我们获得了 89 个主题。按照与中文调查类似的结构,我们采取伪装、随机调换标签顺序、生成调查任务的方式,并在该平台上发布了问卷,每份问卷请 30 个答题人进行回答。

由此,我们一共评估了 3 篇涉及中文的文章(27 个主题)、10 篇英文文章(89 个主题)中主题标签的情况。如果答题人选择原文标签的比例并不比选择 ChatGPT 标签的比例高出太多,则很可能说明 ChatGPT 标签足以以假乱真。本文约定,如果超过半数的答题者选择原文标签,则说明原文标签比 ChatGPT 标签更令人信服。或者,如果选择原文标签的人数比选择 ChatGPT 标签的人数更多,则说明原文标签更令人信服。反之,则说明 ChatGPT 标签足以以假乱真。

三、结果

(一) 整体结果

在中文文章的 27 个主题中,只有 1 个主题有等于或超过 50%的受访者选择了原文标签,1 个主题选择原文标签的人数超过选择 ChatGPT 标签的人数;在英文文章的 89 个主题中,只有 5 个主题有等于或超过 50%的受访者选择了原文标签,27 个主题选择原文标签的人数超过选择 ChatGPT 标签的人数(图 1)。换句话说,不论以前面约定的哪个标准来判断,在大多数主题上,ChatGPT 标签都达到了以假乱真的程度。还有一个主题比较特殊,就是 Goldenstein & Poschmann (2019)中的一个主题其两个标签一样,因而答题人的选择不能说明问题。但即使排除这个特殊的主题,剩下的数据也仍然支持前面的观察。

(二) 自变量的潜在影响

对 GPT-3.5 的训练只使用了 2021 年 9 月之前能够获得的英文文本数据,这构成了这一模型的知识阻断点。该模型在某些主题上表现较好,可能是因为

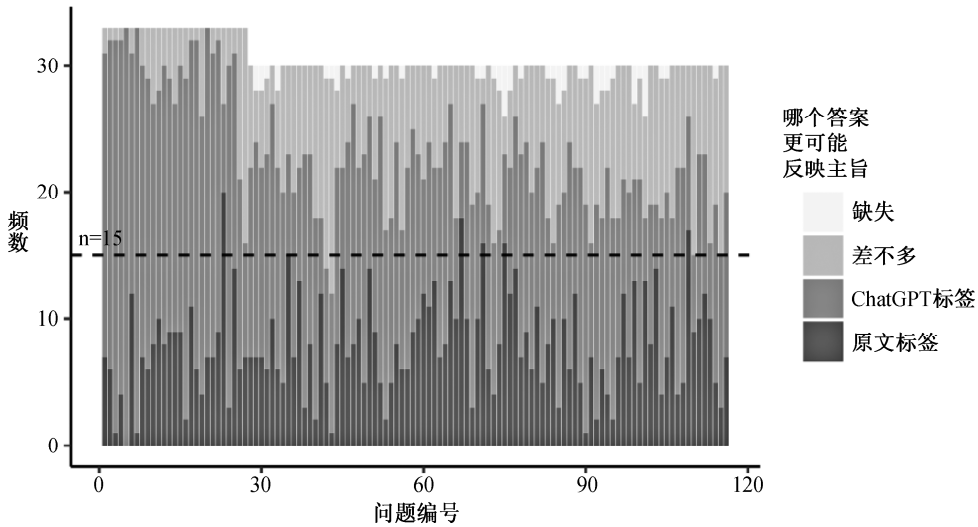

图 1　受访者在原文标签与 ChatGPT 标签之间选择情况总览

数据来源：Credamo 和 Mechanical Turk 正式调查数据。

这些发表于 2021 年 9 月之前的论文在其训练数据当中，因而打标签对模型来说不过是简单的信息检索过程。为此，我们把不可能存在于训练数据的 2 篇中文论文、2 篇英文论文标注出来，观察模型表现是否有显著差别。结果表明，是否存在于训练数据中对模型表现的影响并不显著。我们还比较了 3 篇涉及中文的文章和其他纯英文文章的表现，也未发现显著差别。由于“不可能在训练数据中”“涉及中文”这两个属性高度重叠，这里仅报告前者的结果（图 2）。

为了加强对描述统计的信心，我们以每一个主题为分析单位，以“选择原文标签的比例”为因变量，“是否可能在训练数据中”“是否涉及中文”为自变量，进行逻辑斯蒂回归（由于因变量是取值从 0—1 的变量，不符合线性回归对因变量的假设，因此应采取逻辑斯蒂回归），结果表明变量系数均不显著，印证了这里的观察，结果不赘述。

（三）对上述结果的讨论

如果 ChatGPT 标签足以以假乱真是因为其标签与原文标签高度相似也就罢了，但令人惊讶的是，很多时候 ChatGPT 标签与原文标签有明显差别，而答题

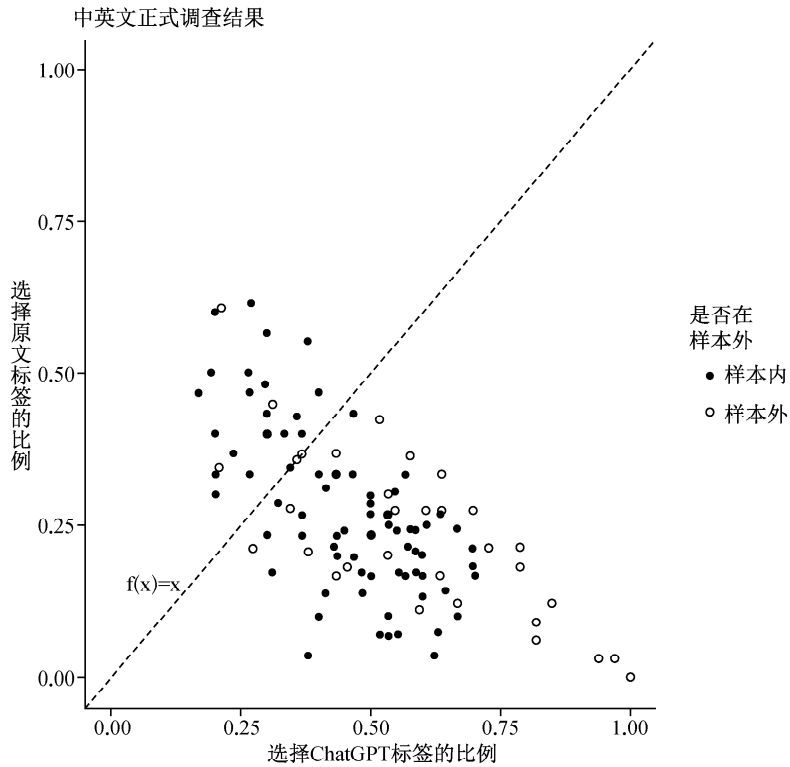

图 2　受访者选择原文标签与 ChatGPT 标签的结果对比

数据来源:Credamo 和 Mechanical Turk 正式调查数据。

人还是更倾向于选择 ChatGPT 标签(图 2)。这是为什么呢?

观察 ChatGPT 标签与原文标签的差别,我们提出下面的假说:与原文标签相比,ChatGPT 标签往往文字更多、更具体。这可能说明,用更多文字对主题进行更清楚的说明更容易说服读者。即便抛开本文采用大语言模型的方法探索不谈,这也给今后采用主题模型进行研究的学者以启发。作者给主题打什么标签本来没有任何限制,此时高度抽象概括的用语尽管看似言简意赅,实际上未必最有说服力。

表 2、表 3 分别列出剔除缺失值后,选择原文标签比例最高和最低的 10 个主题。其中表 3 报告的第 4 个主题就是前文提到的两个标签一致的主题,大多数人选择了"差不多",碰巧选择我们的底层代码指认为原文标签的答题人只有

一个。这两个表中有些结果确实符合前面的假说,但有些结果中 ChatGPT 的标签更为详尽,被选择的比例反而更低。对于这种情况,恐怕还需要进一步专门研究。

表 2　　所有论文中原文标签被选中比例最高的 10 个主题

编号	论文名	特征词	原文标签	ChatGPT 标签	原文占比
5	Managing the Boundaries of Taste	Adam、stroke、garag、pavement、pixi、surf、garagerock、weezer、lewi、malkmus	garage	Indie rock music	0.62
26	历史社会学能化解学科之争吗	women、family、children、men、marriage、household、work、social、female、gender、male、marry、live、age、class	性别与婚姻	性别和家庭动力学	0.61
4	Rhetorics of Radicalism	social、unity、political、communities、revolution、parties、west、exit、nature、system	social & political revolution	political transformation	0.60
3	Theming for Terror	faction、sharia、jawlani、Syrian、regime、god、Jesus、caliphate、camp、plot、Syria、pkk、pagan、shall、Ramadan	religious boundaries	Syrian conflict and extremist groups	0.57
1	Managing the Boundaries of Taste	Steven、bird、Stewart、muse、snow、finn、piano、stori、tale、narrat	lyrics/songwriter	storytelling or narration	0.55
8	Using Radical Environmentalist Texts to Uncover Network Structure and Network Features	action、direct、campaign、group、involv、network、sabotag、anti、reclaim、tactic、meet、success、event、sab、act、target、rts、opencast、media、aim	direct action/ecotage	activism and resistance	0.50
5	Seeing Like the Fed	credit、financial、institutions、risk、market、loans、banks、downside、cut、turmoil、losses、cdo、mortgage、deterioration、tranches、capital、auction、liquidity、housing、October、asset、January、sheet、swap、lending、commercial、agencies、senior、insurance、paragraph	financial markets	financial risk and instability	0.50

(续表)

编号	论文名	特征词	原文标签	ChatGPT标签	原文占比
3	Seeing Like the Fed	reserve、target、zero、program、funds、sheet、bound、rate、federal、December、quantitative、treasury、banks、facility、guarantee、interest、purchases、ceiling、quantity、effect、deflation、tools、excess、money、policy、fomc、size、monetary、desk、alternative	policy response	monetary policy tools and actions of the federal reserve	0.48
8	Analyzing Meaning in Big Data	company、year、percent、last、analysts、earnings、share、sales、years、quarter、profit、revenue、profits、analyst	stock market and profits	financial performance	0.47
3	Differentiating Language Usage through Topic Models	ident、practic、discours、cultur、nation、construct、global、wai、local、polit、examin、negoti、power、ethnograph、explor、argu、subject、relat、project、space	identity studies	cultural and political discourse in global and local contexts	0.47

表 3　　所有论文中原文标签被选中比例最低的 10 个主题

编号	论文名	特征词	原文标签	ChatGPT标签	原文占比
6	“黑箱”中的青年	推荐、喜欢、内容、隐私、方便、信息、感兴趣、符合、信息茧房、时间、推送、容易、娱乐、喜好、需求、侵犯、淘宝、音乐、商品、泄露、窥探、搜索、监视、有好有坏、中立、满足、兴趣、有利有弊、单一	娱乐内容	用户偏好和隐私	0.00
4	“黑箱”中的青年	类似、历史、浏览记录、监控、偷听、热门、屏蔽、偷窥、窃听、总是、反复	运作机制	监控和隐私担忧	0.03
8	“黑箱”中的青年	广告、不喜欢、隐私、讨厌、推荐、内容、朋友圈、算法、推送、感觉、侵犯、反感、投放、精准、很烦、太多、不感兴趣、泄露、监视、软件、平台、体验、烦人、消费、厌恶、厌烦、社交、浪费时间、用户、重复	广告内容	对广告推送和侵犯隐私的负面态度	0.03

（续表）

编号	论文名	特征词	原文标签	ChatGPT 标签	原文占比
6	Analyzing Meaning in Big Data	insurance、company、life、companies、policies、coverage、insurers、insurer、policy、percent、first、claims、premiums、business	Insurance industry	insurance industry	0.03
4	Racialized Discourse in Seattle Rental Ad Texts	winter、warm、morn、nearest、keep、climat	Cozy and Comfortable	winter weather and climate forecasts	0.03
19	历史社会学能化解学科之争吗	state、govern、new、public、policy、social、develop、nation、institution、politic、economic、city、universal、power、local	制度演变	公共政策与治理	0.06
3	Analyzing Meaning in Big Data	industry、congress、commission、new、government、federal、agency、law、legislation、rules、committee、administration、proposal、officials	Legislation	government and legislative affairs	0.07
6	Differentiating Language Usage through Topic Models	commun、member、organ、structur、network、interact、relationship、group、individu、relat、societi、kinship、form、activ、exchang、role、share、associ、maintain、institut	Social structures	social structure and organization in communities and groups	0.07
9	Racialized Discourse in Seattle Rental Ad Texts	cours、golf、Jefferson、height、point、sand	High-Class Surroundings	golf course geography and topography	0.07
6	Racialized Discourse in Seattle Rental Ad Texts	burk、trail、gilman、microsoft、campus、googl	Commuting Distance	technology companies and outdoor activities	0.07

当然，答题人的选择比例较高并不在客观上等价于这个标签更恰当。因为受访者未曾阅读原始语料，他们对标签的评判不具有专业性。例如，在格里夫等人论文（Greve，Rao & Vicinanza et al.，2022）的一个主题上，原文标签是“虚假死亡统计数字”，而 ChatGPT 标签是“死亡率和死因”。尽管选择 ChatGPT 标签

的答题人更多,但实际上原文是研究线上阴谋论的,所以原文标签才更准确。

这个问题并非无法解决,一旦我们在提示语中提供了正确的背景信息,ChatGPT 就会进行修正。我们在输入上述主题时还尝试给 ChatGPT 提供背景信息:“我们正在研究推特上关于 COVID-19 的阴谋论言论,并在这些阴谋论言论的语料中训练了一个主题模型。”ChatGPT 便返回了更加贴近研究问题的新标签:“COVID-19 Death Toll Comparisons”。这篇文章发表于 2022 年,在 ChatGPT 的知识阻断点之后,因此这个结果不是简单的信息检索得出的。这说明如果在提示语中给予充分的背景信息,ChatGPT 的编码效果还能进一步提升。前面本文运用的是不提供背景信息的相当简化的提示语,所得结果中 ChatGPT 的表现已经相当不错,因而如果改进提示语,想必在类似的评估中表现会更好,这里不必赘述。

这提示了大语言模型与主题模型在社会科学中的另一种结合方式。芒克(A. Munk)等指出,机器学习模型在预测上失败的案例,往往也是诠释起来最有难度、最需要研究者详细说明的案例(Munk, Olesen & Jacomy, 2022)。同理,也许 ChatGPT 标签与原文标签差异最大的主题,恰恰就是最富理论复杂性而值得探讨的主题。大语言模型此时从反向发挥了启示研究者的作用。

(四)局限

GPT-3.5 对于输入的信息有长度限制。由于要输入的内容较多,笔者很难将所有主题及其特征词一次性输入 ChatGPT,让它将不同主题之间的差异纳入考虑。出于同样的原因,某个主题占比最多的文章是哪篇、带有主题特征词的关键性段落包含什么内容、主题在时间维度上的比例变化等信息也没法被输入 ChatGPT,尽管这些信息在人为解读主题的时候是相当有价值的。

在这个问题上,GPT-4 有所改观。接下来,研究者可以评估这样的输入方式能否提升表现。可以预计,由于这些限制在工程上并非不可接近的,随着该技术路线得到认可,今后的大语言模型可能会进一步扩展其上下文的限制,并且能够接受文本以外的其他输入类型。尽管我们还不能对此进行评估,但基于现在的证据,我们对此表示谨慎乐观。

四、结论与讨论

本文结合大语言模型(ChatGPT)和小语言模型(主题模型),试图解决后者在应用中遇到的一个实际问题。由于给主题模型的结果打标签具有一定的任意性,而进行人工校验的成本又较高,因此对主题模型的结果进行大规模的标签校验难度较大。大语言模型基于其在预训练时所掌握的人类语言模式,可以自动给主题模型的结果打出相对而言可信度较高的标签,从而为上述问题提供一个可行的解决方案。

我们并不主张 ChatGPT 可以完全取代研究者的工作,不过它确实能够辅助研究者。首先,在模型探索阶段,大语言模型能够快速给模型的结果赋予标签,从而允许研究者在较短时间内探索更多的模型。其次,研究者可以将大语言模型标签与自己的标签进行对比,识别具有理论意义或需要澄清的主题。

基于这一发现,我们既对大语言模型模拟人工编码的能力抱有积极态度,又对实践中应用大语言模型持谨慎态度。在积极态度上,对于使用主题模型的论文而言,论文评审人进行匿名评审时的处境与本文中答题人的处境相似。他们很可能对这一研究使用的语料并不熟悉,只能依赖作者在论文中报告的特征词来对主题取得初步的理解,因而作者给主题赋予的标签就具有某种不容辩驳的权威性,而这恰恰可能关涉到后续的理论分析。由于作者在打标签问题上利益相关且可能存在偏误,评审人又可能缺乏纠偏能力,那么该如何保证人为标签不扭曲研究结果呢?大语言模型可能提供了一种检验的可能。但在消极态度上,评审人又能够根据什么判断原作者对语料确实有相当程度的掌握,而不是仅仅使用了大语言模型呢?在研究中禁止使用 ChatGPT 这类应用恐怕是不可行的,建立这样的规范只会让研究者欺瞒评审人。这一研究规范与伦理问题,可能是目前学术出版工作面临的最棘手、最紧迫问题之一。

这一研究还给了我们一些启发。如果研究不给主题模型的结果打标签,或者使用非常模糊、宽泛的标签,其结果并不容易说服读者。ChatGPT 给出的标签更具体、详尽,受一般读者的欢迎,说明研究者应该注意打标签的风格。在后续研究中,研究者可以进一步衡量:如果进一步扩展大语言模型的上下文限制,或

者使用小模型的训练文本对大模型进行微调,是否可以让大语言模型更贴近研究者的需要,取得更好的表现?

五、余论:“失控”

如果本文是一篇四平八稳的量化研究,那么到此本应已经结束。但从事研究的过程给了我们很多自反的机会,值得与学界同人探讨。

如果大语言模型已经可以很好地完成自然语言处理任务,那么为什么还要执着于主题模型这类小模型呢?正如文献综述部分介绍的,ChatGPT 等大语言模型对人类语言的理解比基于“词袋”的主题模型更复杂,对后者可能有替代性。用大语言模型改进小语言模型岂非刻舟求剑?

或许大语言模型尚未一统江湖的一个原因在于,研究者对小语言模型有更强的掌控能力。主题模型的原理并不太难,通过阅读主题的特征词而猜测其打标签的过程也颇直观。研究者可以在电脑上进行主题模型的运算,还可以调整超参数来探索模型的结果。尽管实际上主题模型内部蕴含着不可小觑的不确定性,但在一定程度上,研究者似乎还是可以把控住从输入数据到输出数据的过程。

相反,大语言模型内部结构的庞大与复杂性,已远远超出大多数研究者的认知能力范畴。在本文中,利用 ChatGPT 进行标签标注时,我们面临的一个主要问题是其可重复性无法得到保证。具体而言,由于模型本身固有的不确定性,即便是输入相同的提示语,每次得到的结果都可能存在差异。这首先是因为 GPT-3.5 如同其他概率模型一样,在推断过程中包含随机因素。除此之外,随着模型的持续迭代,即便是相同的提示语,在不同时间点的输入也可能产生不同的结果。这种结果的变化及其规律,完全超出了研究者的控制范围。至于为何某些提示语能够取得较好的效果,大多数人往往难以给出严谨的解释,这种现象仿佛带有某种巫术或魔法的色彩。

这明显违背了学界理想状态下的规范。理论上,研究工具的使用者应当对其工具的用途和局限有深刻的理解,并在必要时进行恰当的修正,从而对研究结果负责。然而,尽管滥用统计工具在学术研究中屡见不鲜,但在应然层面,这仍

然是一条被广泛认同的原则。然而,大语言模型这类高度复杂的研究工具,往往让使用者在多数情况下难以理解其原理、用途和局限,从而难以进行适当的评估和修正,进而无法对研究结果负责。与此同时,大语言模型在定义清晰的任务上表现出惊人的能力,使得传统的、相对落后的工具无法与之相提并论。

这种能力和责任上的对立,正是近年来计算社会科学中所遇到问题的缩影。新方法在诸多问题上的表现远胜于研究者过去所学的传统工具,但这些新方法往往也伴随着高度的复杂性,使得研究者难以充分掌握。在这种情况下,若仍坚持使用新方法,研究者则可能面临方法逐渐“失控”的问题。这里所说的“失控”并非一个绝对的状态,而是一个程度的变化,研究者可能从“不那么失控”逐渐走向“越来越失控”。

以 ChatGPT 为代表的大语言模型应用使得我们的方法“失控”达到了新的高度,这主要源于它与以往计算方法存在本质差异。过去的计算方法应用领域边界明确,我们往往对模型的数学估计过程有着较为清晰的理解(如 LDA 模型),或是模型针对特定任务而设计(如深度神经网络在分类预测任务中的应用)。这些方法伴随着一套专业术语,使得它们与社会科学的其他部分有所区分。因此,这些工具最终通常只能被部分人以相对规范的方式使用和理解,难以全面改变社会科学知识的生产方式。

然而,大语言模型则截然不同。一方面,其复杂性超出了绝大多数社会科学家所能掌控的范围;另一方面,其输出结果又极为接近自然语言,因此更容易被广大社会科学家所使用和理解。从理论上讲,大语言模型执行的是一项明确的预测任务,即通过对大量文本数据进行预训练后建立模型,以在给定的上文基础上预测下文。然而,由于这一任务的泛化性,研究者在使用时往往不局限于理解其预测结果,而是为其赋予了更深层次的意义。以本文为例,大语言模型实际上是在预测与提示后续可能出现的文本,而非理解、阐释或标注主题模型的主题,但我们却试图通过前者达到后者的效果。至于为何大语言模型能在这一具体任务上表现出色,我们仍感到困惑。这是因为训练模型所使用的目标函数并非研究者为特定研究任务所设定的,研究者对其细节既不了解也无法修改。有研究者称这种现象为大语言模型的“涌现”特质,但这一概念的提出似乎并未解决任何问题。

在大语言模型时代,机器学习模型领域正展现出日益封闭的特性,这进一步加剧了大语言模型难以被社会科学家掌控的困境。由于大语言模型规模庞大、投入巨大,已远超出个人乃至机构研究者的能力范畴,我们不得不依赖企业提供现成的产品,并在此基础上进行有限的调优。最初的大语言模型,如 BERT 和 GPT,均基于开源数据进行训练。然而,随着竞争压力的加剧和潜在法律风险的增加,开发大语言模型的公司已不再公开模型训练的语料来源和策略。以开发了 ChatGPT 的 OpenAI 为例,在公布下一代大语言模型 GPT-4 时,对于相关信息只字未提。这导致研究者难以了解企业使用何种数据进行模型训练,以及训练过程中可能存在的未披露问题。一旦企业产品升级换代,研究者就只能尝试新品,而无法确保此前研究结果的可延续和可推广。这些问题使得研究者对大语言模型的控制能力变得空前微弱,难以对研究结果承担高度责任。

与此同时,大语言模型的输入与输出却高度贴近自然语言,无须跨越专业术语的门槛。即使对统计方法、计算机语言一无所知的研究者,也能通过与 ChatGPT 的自然对话获得所需结果。ChatGPT 能够阐释线性回归表格的结果,推测访谈中话语的情感(尽管其可靠性有待商榷),甚至引经据典地写作文献综述(尽管引用的文献未必真实存在)。这种输出结果极难被人或机器分辨,意味着计算方法首次能够全面影响社会科学研究共同体中各个成员的生产方式。这或许能极大提升学术生产力,但也引发了人们的担忧:若我们对此习以为常,又如何分辨它提供的是知识、意见还是幻觉?

社会科学研究正面临方法“失控”的挑战,这对学界关于知识定义、知识生产方式以及知识产权归属的认知带来了巨大挑战。既往的规范中并未提供现成的答案。对于中国社会科学而言,要构建自主的知识体系,必须通过公开的研讨来解决这些问题。若处理得当,这或将成为中国社会学发展的一大机遇。

参考文献

陈云松、吴晓刚、胡安宁等,2020,《社会预测:基于机器学习的研究新范式》,《社会学研究》第 3 期。

郭台辉、周浥莽,2020,《历史社会学能化解学科之争吗?——基于西方学术史的结构主题模型分析》,《社会学研究》第 3 期。

赵龙轩、林聪,2022,《“黑箱”中的青年:大学生群体的算法意识、算法态度与算法操纵》,《中国青年研究》第 7 期。

周涛、高馨、罗家德,2022,《社会计算驱动的社会科学研究方法》,《社会学研究》第 5 期。

Bender, E., T. Gebru & A. McMillan-Major et al. 2021, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Blei, D. 2003, “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3.

Blei, D. 2012, “Probabilistic Topic Models.” *Communications of the ACM* 55(4).

Buurma, R. 2015, “The Fictionality of Topic Modeling: Machine Reading Anthony Trollope's Barsetshire Series.” *Big Data & Society* 2(2).

Caliskan, A., J. Bryson & A. Narayanan 2017, “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” *Science* 356(6334).

Devlin, J.; M.-W. Chang & K. Lee et al. 2019, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1.

DiMaggio, P., M. Nag & D. Blei 2013, “Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U. S. Government Arts Funding.” *Poetics* 41(6).

Goldenstein, J. & P. Poschmann 2019, “Analyzing Meaning in Big Data: Performing A Map Analysis Using Grammatical Parsing and Topic Modeling.” *Sociological Methodology* 49(1).

Greve, H., H. Rao & P. Vicinanza et al. 2022, “Online Conspiracy Groups: Micro-Bloggers, Bots, and Coronavirus Conspiracy Talk on Twitter.” *American Sociological Review* 87(6).

Grimmer, J. & B. Stewart 2013, “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3).

Heuser, R. & L. Le-Khac 2012, “A Quantitative Literary History of 2958 Nine-A Quantitative Literary History of 2958 Nineteenth-Century British Novels: The Semantic Cohort Method.” *Pamphlets of the Stanford Literary Lab* 4.

Kaplan, J., S. McCandlish & T. Henighan et al. 2020, “Scaling Laws for Neural Language Models.” arXiv: 2001.08361[cs.LG].

Kozlowski, A., M. Taddy & J. Evans 2019, "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 84(5).

Li, L., L. Wu & E. James 2020, "Social Centralization and Semantic Collapse: Hyperbolic Embeddings of Networks and Text." *Poetics* 78.

Marshall, E. 2013, "Defining Population Problems: Using Topic Models for Cross-National Comparison of Disciplinary Development." *Poetics* 41(6).

Meeks, E. & S. Weingart 2012, "The Digital Humanities Contribution to Topic Modeling." *Journal of Digital Humanities* 2(1).

Miller, I. 2013, "Rebellion, Crime and Violence in Qing China, 1722: A Topic Modeling Approach." *Poetics* 41(6).

Mohr, J. &P. Bogdanov 2013, "Introduction Models: What They Are and Why They Matter." *Poetics* 41(6).

Munk, A., A. Olesen & M. Jacomy 2022, "The Thick Machine: Anthropological AI Between Explanation and Explication." *Big Data & Society* 9(1).

Nelson, L. 2021, "Cycles of Conflict, A Century of Continuity: The Impact of Persistent Place-Based Political Logics on Social Movement Strategy." *American Journal of Sociology* 127(1).

OpenAI 2022, "Introducing ChatGPT." https://openai.com/blog/chatgpt.

Ouyang, L., J. Wu & J. Xu et al. 2022, "Training Language Models to Follow Instructions with Human Feedback." arXiv: 2203.02155[cs.CL].

Prepare for Truly Useful Large Language Models 2023, *Nature Biomedical Engineering* 7(2).

Stoltz, D. & M. Taylor 2021, "Cultural Cartography with Word Embeddings." *Poetics* 88.

Ying, L., J. Montgomery & B. Stewart 2021, "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* (September).