arXiv:2305.01764v1 [cs.CL] 2 May 2023

心理学启发的因果提示

Zhiheng Lyu*香港大学
zhlyu@cs.hku.hk

Zhijing Jin*MPI 与苏黎世联邦理工学院
jinzhi@ethz.ch

Mrinmaya Sachan
苏黎世联邦理工学院
msachan@ethz.ch

Rada Mihalcea密歇根大学 mihalcea@umich.edu

Bernhard Schölkopf智能系统马克斯·普朗克研究所
bs@tue.mpg.de

摘要

自然语言处理数据集不仅仅是输入输出对；它们还包含输入变量与输出变量之间的因果关系。在这项工作中，我们以情感分类为例，探究评论（X）与情感（Y）之间的因果关系。正如心理学研究所示，语言会影响情绪；当一个人先给出评分，再在评论中自我合理化其触觉时，会触发不同的心理过程（即情感导致评论，$Y \rightarrow X$）；而先描述自己的经验、权衡利弊，再给出最终评分时，触发的心理过程则不同（即评论导致情感，$X \rightarrow Y$）。此外，如果标注员通过心智理论（ToM）推断用户最初给出的评分，涉及的心理过程也完全不同（即评论导致评分，$X \xrightarrow{\text{心智理论}} Y$）。在这篇论文中，我们将情感分类中人类心理过程的这三种因果机制转化为三种不同的因果提示，并研究：（1）它们的表现有何差异；（2）情感分类数据的哪些特性会导致提示引发的模型响应出现一致或多样的情况。我们建议未来工作关注自然语言处理任务中的不同因果结构。[1]

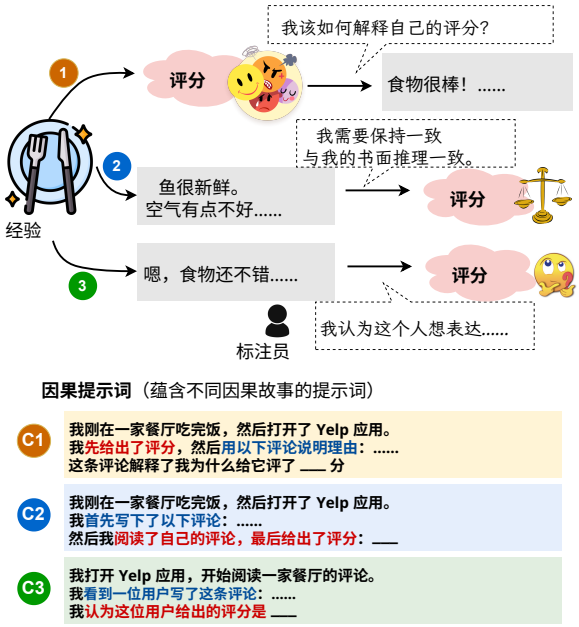


图 1：评论情感分类任务可能具有不同的潜在心理因果过程。我们使用一个提示词来描述三个因果过程中的每一个，记为因果提示 C1、C2 和 C3。

1 引言

大多数关于自然语言处理计算方法的研究都将每项任务视为输入输出对。以情感分类为例，常见方法是将评论 X 作为输入，将情感 Y 作为输出，模型据此尝试学习最优预测 $P(Y|X)$（查看综述 Liu andZhang，2012；Ravi and Ravi，2015；Poria 等，2020），从而将其变成一项文本分类任务，随后便可以使用标准文本分类器微调后（Kim，2014；Yin 等，2019年，等等）。

然而，输入变量与输出变量之间可能存在丰富的潜在因果结构。例如，对于 Yelp 评论数据（Zhang 等人，2015），我们在图1中列出了三种不同的心理过程，它们可能产生评论-情感对：(C1) 用户先给出评分，然后在评论中自我合理化其感受（例如，一种心理过程是为其直觉情绪寻找理由）；(C2) 用户先在评论中描述自己的经验，然后决定评分（此时评分往往更理性，例如与评论中的事实一致）；(C3) 标注员通过心智理论（ToM）推断用户原本的评分（Stich，1983年；Frith 和 Frith，2005），这对于一般的文本分类任务也是可行的。

*贡献相同。[1]我们的代码和数据位于 https://github.com/cogito233/psych-causal-prompt。

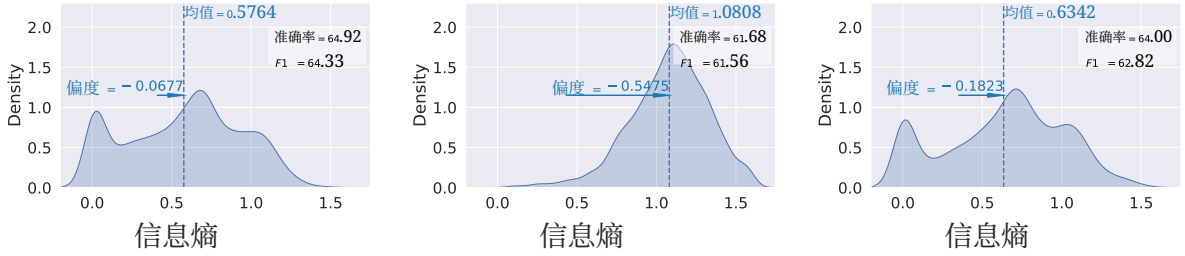


图 2：GPT3-Instruct 的表现（Ouyang 等人，2022）由 C1（左）、2（中）和 3（右）诱导，在亚马逊评论测试集的 5K 个英文样本中，按准确性、加权 F1 以及预测的信息熵分布进行评估（Keung 等人，2020）。

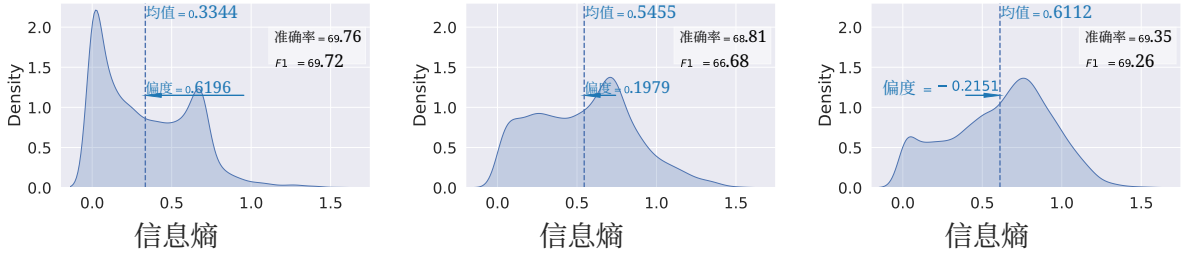


图 3：GPT3-Instruct 的表现（Ouyang 等人，2022）由 C1（左）、2（中）和 3（右）诱导，在 Yelp-5 测试集的 1万 个样本中，按准确性、加权 F1 以及预测的信息熵分布进行评估（Zhang 等人，2015）。

在这篇短论文中，我们旨在探讨这样一个问题：不同的心理因果过程会如何影响自然语言处理任务？本文以情感分类为例，运用情感科学中的心理学理论，设计与不同心理过程相对应的因果提示，并探究两个问题：(Q1) 这些提示会引发模型哪些不同的行为？(Q2) 不同因果提示所引发的模型预测，能告诉我们情感分类本质的哪些信息？

我们发现，三种因果提示词带来的模型表现相当，但对其预测的确定性各不相同。我们还发现，在情感分类数据中，篇幅较长、情绪表达更明确的评论会使不同因果提示词得出的结果更加一致；而篇幅较短、表达更含蓄的评论，则会使不同因果提示词产生多样的预测。这促使我们反思情感分类任务的设定，尤其是对于情感更加细腻、可能因潜在心理过程而发生变化的评论。

2 跨学科启发

情感科学的洞察。在情绪研究，即情感科学领域（Salovey and Mayer，2004；Barrett，2006；Feinstein，2013年），语言与情感之间的相互作用长期以来一直是重要的研究领域。对于（1）因果过程，情绪如何影响语言较为直观，因为人们感知到的情绪会影响他们当下的沟通方式（Barrett，2006）。对于（2）因果过程，情感科学也揭示了语言如何影响情绪。例如，研究表明，报告自身情绪状态这一行为会改变人们对情绪情境的身体反应（Kassam andMendes，2013）。这种从语言到情绪的变化，也能通过功能性磁共振神经成像观察到（Satpute et al.，2013）。对于（3）因果过程，一个人如何推断他人的情感属于心智理论（ToM）的研究范畴（Stich，1983年；Frith and Frith，2005），这种推断可能同时受到刺激以及该人推断他人心理状态能力的影响，因此或许会因理解各异而产生标注员之间的差异。

自然语言处理中的因果与反因果学习。我们的工作也与因果和

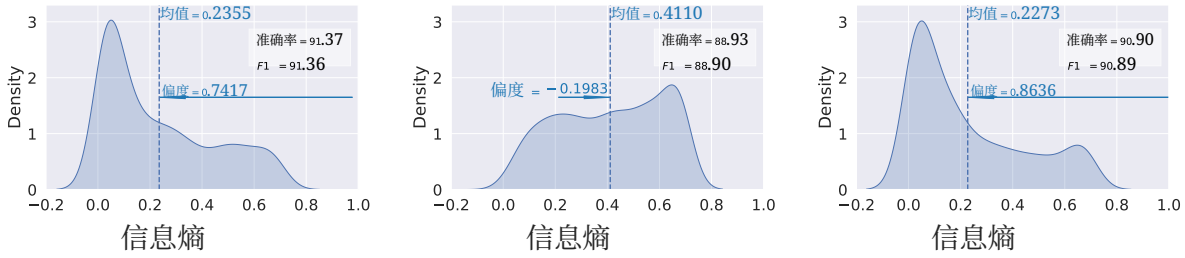


图 4：GPT3-Instruct（Ouyang 等人，2022）在 rotten-tomatoes 数据集上由 C1（左）、2（中）和 3（右）诱导的表现，包括准确性、加权的 F1 值以及预测的信息熵分布。

自然语言处理中的反因果学习（Jin 等人，2021），它根据自然语言处理任务是学习从结果预测原因，还是从原因预测结果来进行区分。应用到我们的工作中，如果底层因果机制发生反转，情感分类任务的性质也会随之改变。例如，如果 C1 是某个数据集上的真实情况，那么情感分类器执行的是反因果学习；而如果 C2 或 C3 是真实情况，那么情感分类任务就变成了因果学习。

3 不同提示词的表现如何?

在本节中，我们首先探讨问题 Q1，即三种不同的因果提示能够引发哪些不同的行为。

选择因果提示基于三种不同的心理过程，我们希望为后续分析设计三种因果提示。设计提示词时，我们考虑以下期望标准：(1) 每个提示都应明确体现其所对应的因果过程。(2) 提示词应符合自然语言表达，但不能是残缺或错误的句子。(3) 对这三种提示进行跨因果性比较应当有意义，这意味着我们需要考虑其他类型的噪声，因为这些噪声会导致大语言模型对提示产生不同的响应。遵循这些原则，我们在广泛了解提示设计相关文献的基础上，进行了审慎的提示选择过程，详情见附录A，其中包括相关综述（Liu 等人，2021）和近期工作（Khashabi 等人，2022；Kojima 等人，2022）。

我们改编的三种因果提示如图 1 所示（记为 C1、C2 和 C3）。我们先确保满足标准（1）和（2），再通过执行标准（3），从结构、措辞和长度方面对这三种提示词加以控制。为确保这三种提示词得到良好控制，我们还在附录A中进行了额外分析，例如，冗长程度（即在保留含义的同时改变长度）如何以相似的方式对三种提示词产生负面影响，以及提示词中的扰动如何一致地降低三种提示词的性能。我们的提示词在长度、措辞和自然度方面都相对接近。

比较预测分布我们使用精心挑选的三种因果提示，观察它们在两个常用的英语情感分类数据集上的表现：Yelp-5（Zhang 等人，2015）[2] 和亚马逊评论（Keung 等人，2020）。[3] 该数据集有五个标签，分别对应 1–5 星评分。我们在这一细粒度的五分类任务上运行性能最佳的自回归大语言模型 GPT3-Instruct（Ouyang 等人，2022），通过比较不同因果提示所诱导的、涵盖全部五个标签的预测分布，我们能够获得更丰富的信息。实验细节见附录B。

我们在图4中展示了 Amazon 数据集和在图3中展示了 Yelp 数据集在三种提示词下的表现差异。对于每种提示词，我们首先报告五分类准确率和加权 F1 分数（位于每张图的右上角），其中 C1 略优于另外两种提示词。

接下来，我们还比较了三种自动补全过程的信息熵。我们的直觉是，熵可能会揭示不同因果提示在对数概率方面更细微的差异。近期一项研究中的假设（Jin 等人，2021）指出，对于评论情感数据集，

[2] https://huggingface.co/datasets/Yelp_评论_完整 [3] https://huggingface.co/datasets/amazon_评论_多类

预测高度一致的评论 评论：服务糟透了！！经理态度粗鲁！永远失去了一位顾客。而且经理毫不在意。评分：*C1*：*1* 星。*C2*：*1* 星。*C3*：*1* 星。*GT*：*1* 星。
评论：沟通、服务和维修质量都出色！我再满意不过了！还有免费代步车！免赔额全免，或部分免除！只用了 *8* 个工作日，就让我的 *2006 Solara* 焕然一新！我强烈推荐他们！告诉他们是我介绍你来的！！！评分：*C1*：*5* 星。*C2*：*5* 星。*C3*：*5* 星。*GT*：*5* 星。预测多样性较高的评论 评论：坦白说……我没在这里吃过饭。说实话……我确实走进去时，看到比尔·克林顿正在私人餐厅用餐，随即就被美国特勤局告知离开。说句公道话……他看起来很享受这顿饭。评分：*C1*：*1* 星。*C2*：*1* 星。*C3*：*5* 星。*GT*：*5* 星。评论：等得太久了！！他们应该扩大露台，如果等候超过一小时，就送一杯饮料！谁会等一个小时才出去吃饭啊！这就是我不来这里的原因。她倒是想来这里！评分：*C1*：*1* 星。*C2*：*2* 星。*C3*：*1* 星。*GT*：*2* 星。

表 1：三个因果提示（C1、C2 和 C3）在预测中具有高度一致性或多样性的评论示例。我们还展示了真实标签（GT）。

用户通常先评分，然后写下相关评论。这对应于我们研究中的 C1。我们观察到，C1 的预测确定性明显更高，其平均信息熵较低，为 0.3344，几乎只有另外两个因果提示的一半。除了均值，我们还计算了偏度；可以看到，C1 的熵分布也更偏向较低熵值，偏度得分为 0.6196。

4 因果提示告诉我们关于数据的什么？

分析了 LLM 对不同提示词的响应特性后，我们从另一个角度思考：不同的因果提示能告诉我们关于数据的什么？

我们在表 1 中展示了我们的研究动机。可以看到，在三个因果过程之间一致性较高的评论往往包含一些明确的信号，例如清晰的观点词；而导致预测结果多样性较大的评论，有时对人类来说似乎也很模糊或棘手。对于这类细微复杂的评论，因果提示所引发的不同心理过程之间的细微差异可能会更加明显。

作为一种帮助我们理解不同情感分类数据本质（例如，多样性可能源自何处）的潜在方式，我们利用 GPT3-Instruct 模型（Ouyang 等人，2022），并探索 Yelp 评论-情感对中，三个因果提示词的响应一致或存在分歧的情况。

量化多样性。对于给定的评论$\boldsymbol{x}_i \in \boldsymbol{D}$，每个模型都会预测一个概率分布 $P(y|\boldsymbol{x}_i)$，表示在给定该评论的情况下各标签的概率 y。我们通过归一化总变差（de Séguier，1892；Gibbs 和 Su，2002）来量化针对两个不同因果提示的模型响应之间的多样性：

$$d(P_1, P_2) = \frac{1}{2}|P_1(y|\boldsymbol{x}_i) - P_2(y|\boldsymbol{x}_i)|\,, \qquad (1)$$

其中，P_1 和 P_2 是两个因果提示下模型预测结果的分布。该指标 $d : \mathcal{P} \times \mathcal{P} \to [0, 1]$ 对应的得分介于0和1之间：0表示没有多样性（完全一致），1表示多样性最大，即两个分布没有重叠区域。

量化明确观点。接下来，我们考察评论文本中可能导致观点多样的特征。一个可以计算的特征是明确观点词的数量。作为初步探索，我们采用一种简单的启发式方法，将评论中的词语与常用的正向[4]和负向[5]观点词列表进行匹配，这些列表由Hu 和 Liu（2004）；Liu 等人（2005）提供。

发现。针对 Yelp-5 的 1万 个测试样本，我们根据预测多样性分析了其中不同的子集。数据的一些有趣特性包括文本长度、观点词总数以及标签分布。例如，一致性较高（即多样性较低，呈极浅蓝色）的文本往往包含更多观点词，且标签分布集中在某些标签上。相反，预测多样性较高（呈深蓝色）的文本往往包含较少的观点词，标签分布也相对均匀，但略微集中在 2–4 星评分上，因为相比两个极端评分 1 和 5，对 2–4 星的判断更模糊。

主要结论与未来工作。本次分析的发现促使我们反思情感分类任务（以及可能

[4] https://ptrckprry.com/course/ssd/data/positive-words.txt [5] https://ptrckprry.com/course/ssd/data/negative-words.txt

	# 样本	# 词数/样本	# 观点词			标签 (%)					多样性
			Pos	Neg	P+N	1	2	3	4	5	
总体	10,000	177.06	6.33±5.54	3.12±3.59	9.44	20	20	20	20	20	0
随机 (n=500)	500	177.14	6.28±5.31	3.08±3.58	9.36	19	22	16	22	19	0
相同且正确：C1=C2=C3=GT	5,994	184.17	6.51±5.59	3.28±3.68	9.79	25	17	17	16	23	0
相同且错误：C1=C2=C3≠GT	2,052	163.31	6.04±5.28	2.84±3.44	8.89	11	23	24	25	15	0
多样：非 (C1=C2=C3)	1,954	169.68	6.04±5.63	2.90±3.43	8.94	12	24	22	24	15	0.4591
相同且正确											
多样性最低的 10% 数据	599	231.90	6.39±5.78	5。	10±4.52	53	9	23	3	10	0.
10% 多样性最高的数据	599	170.47	5.68±5.21	3.01±3.39	8.69	23	29	14	23	9	0.4002
相同且错误											
10% 多样性最低的数据	205	159.96	6.23±4.85	2.74±2.45	8.97	4	40	19	32	3	0.0503
10% 多样性最高的数据	205	178.51	5.74±5.93	3.25±3.74	8.99	17	24	30	11	16	0.
多样											
其中 C1≠GT	972	160.63	5.69±5.58	2.80±3.30	8.49	16	21	23	22	15	0
其中 C2≠GT	1,067	165.22	5.85±5.43	2.76±3.27	8.61	11	24	22	26	15	0.
其中 C3≠GT	1,013	167.27	5.91±5.61	2.81±3.38	8.72	10	25	23	22	18	0.

表 2：会引发三个因果提示词不同表现特性的子集数据统计。我们纳入了测试样本的子集，其中所有预测都相同且正确（相同且正确）、相同但错误（相同且错误），或多样。对于两个具有相同 argmax 预测的子集，我们分析了模型预测多样性最低和最高的 10% 子集。对于 argmax 预测多样的子集，我们分析了分别导致每个因果提示词（C1、C2 和 C3）失效的子集。

许多其他自然语言处理任务）的界定方式。关键问题在于，我们的任务定义得有多明确？许多任务乍看之下似乎很简单，但实际上可能受到许多微妙条件的影响。未来工作需要更多地考虑其他学科的研究，这些研究为输入变量与输出变量背后可能存在的各种因果过程提供了丰富的洞察。

5 结论

在这项工作中，我们提出了因果提示的概念，并开展分析以报告其性能，同时反思情感分类的本质。我们发现，在情感分类数据中，存在许多细微且模棱两可的文本，这些文本会导致多样的预测结果，而这些结果可能对应于对情感的不同心理学解释。我们的研究结果向自然语言处理社区提出了一个开放性问题：我们的任务定义究竟有多明确？

局限性

这篇短论文的主要局限在于，我们的目标主要是验证概念，旨在指出我们需要更加谨慎地对待在模型中诱导出的不同因果过程。我们对实验结果持保留态度，因为还可以尝试其他大语言模型、其他提示词和其他数据集。在这篇短论文的范围内，我们没有对所有组合进行网格搜索，但我们鼓励未来工作通过更扎实、全面的实验进行探索。另一个局限是，大语言模型并非完美。我们使用大语言模型的方式，基于近期的实证成果（Brown 等人，2020；Ouyang 等人，2022；Chowdhery，2022），以及这样一种可能性：模型捕捉到的相关性能够唤起评论与情感之间心理上不同的因果过程，因为训练数据中也可能存在这种框架。此外，我们探讨的这些细微差异，也可能涉及另一种广为人知的现象——测量效应，即测量行为（报告的情感）会改变被观察对象（实际情感）的属性。我们鼓励未来工作对此进行更深入的探索。此外，本研究主要涵盖英语，其他语言和文化中或许存在不同的现象。

伦理考量

出于数据安全和用户隐私方面的考虑，我们使用了一个公开数据集。此外，本论文中引用的示例仅用于研究说明。我们删除了示例中所有提及的私人姓名。

对于潜在的利益相关方及可能的滥用，本研究总体上旨在为情感分类任务提供一种新的思考角度。潜在的负面影响可能涉及情感分类这一通用任务：更准确的模型

可能会被用于挖掘用户信息，其负面用途包括定向营销、监控和欺诈。

关于碳足迹的顾虑：我们的研究主要使用 GPT-3 API。我们尝试通过从原始的大型测试集抽取数千个测试样本来控制碳排放成本；这些样本也具有较好的代表性，足以展示趋势。

致谢

我们感谢普林斯顿大学心理学系的 Erik C. Nook 教授，他在情感科学方面的讨论和参考资料帮助我们构思了这篇论文。本文部分研究基于以下机构的支持：德国联邦教育与研究部（BMBF）资助的图宾根人工智能中心（项目编号：01IS18039B）；机器学习卓越集群（卓越集群编号：2064/1，项目编号：390727645）；密歇根大学精准健康计划；约翰·邓普顿基金会（资助编号：61156）；哈斯勒基金会的负责任人工智能资助；以及苏黎世联邦理工学院资助（ETH-19 21-1）。Zhijing Jin 获得了生命未来研究所和开放慈善提供的博士奖学金，以及 ELISE（GA 编号 951847）为 ELLIS 计划提供的差旅资助。我们还感谢 OpenAI研究者访问计划向团队提供 API 使用额度。

作者贡献

关于构思，Zhijing Jin 从她之前的研究 Jin 等人 (2021) 中汲取灵感，并提出将其扩展到大语言模型。在 Zhiheng Lyu 于苏黎世联邦理工学院进行、由 Zhijing 指导的实习期间，他首先在一篇研讨会论文中探讨了大语言模型是否会在因果提示和反因果提示上呈现独特的特征（LYU 等人，2022）。随后，Zhijing 与情感科学领域的心理学研究人员交流时受到启发，意识到情感启动文本和文本启动情感的现象可以在实际心理学研究中找到依据。因此，我们开始共同探索这篇论文的构想。

Zhiheng Lyu 完成了所有实验和分析。他还提出了许多进一步的探索方向，例如对交叉熵的分析。

Zhijing Jin 密切指导了该项目并撰写了论文。

教授*Rada Mihalcea*、*Mrinmaya Sachan*和*Bernhard Schölkopf* 指导了该项目并改进了论文写作。

参考文献

Lisa Feldman Barrett。2006。解决情绪悖论：分类与情绪体验。《人格与社会心理学评论》，10(1):20–46。2

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习器。载于《神经信息处理系统进展》，第33卷，第1877–1901页。Curran Associates,Inc. 5

Boxi Cao、Hongyu Lin、Xianpei Han、Fangchao Liu 和 Le Sun。2022年。提示能否探测预训练语言模型？从因果视角理解其不可见的风险。*arXiv*预印本 *arXiv:2203.12258*。8

Aakanksha et al Chowdhery。2022。PaLM：通过路径扩展语言建模。5

J de Séguier。1892。《傅里叶级数论》。《数学新年鉴》：理工学院和师范学院考生期刊，11:299–301。4

Yanai Elazar、Nora Kassner、Shauli Ravfogel、Abhilasha Ravichander、Eduard Hovy、Hinrich Schütze 和 Yoav Goldberg。2021。测量并提升预训练语言模型的一致性。计算语言学协会学报，9:1012–1031。8

Justin S Feinstein。2013。人类情绪与感受的损伤研究。《神经生物学当前观点》，23(3):304–309。2

Chris Frith 和 Uta Frith。2005。心智理论。《当代生物学》，15(17):R644–R645。1，2

Tianyu Gao、Adam Fisch 和 Danqi Chen。2021。让预训练语言模型成为更优秀的少样本学习器。载于 第*59*届计算语言学协会年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第3816–3830页。 8

Alison L Gibbs 和 Francis Edward Su。2002。论概率度量的选择与界定。 国际统计评论，70(3):419–435。 4

Minqing Hu 和 Bing Liu。2004。挖掘和总结客户评论。载于第十届

ACM SIGKDD 知识发现与数据挖掘国际会议，第 168–177 页。4

Zhijing Jin、Julius von Kügelgen、Jingwei Ni、Tejas Vaidhya、Ayush Kaushal、Mrinmaya Sachan 和 Bernhard Schoelkopf。2021 年。数据收集的因果方向至关重要：因果与反因果学习对自然语言处理的启示。载于*2021* 年自然语言处理实证方法会议论文集，第 9499–9513 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。3，6

Karim S Kassam 和 Wendy Berry Mendes。2013 年。测量情绪的影响：对情绪情境的生理反应取决于是否有人在提问。*PloSone*，8(6):e64959。2

PrideKavumba、Ryo Takahashi 和 Yusuke Oda。2022 年。基于提示的模型是否一无所知？载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 2333–2352 页。8

Phillip Keung、Yichao Lu、György Szarvas 和 Noah A. Smith。2020。多语言亚马逊评论语料库。载于*2020* 年自然语言处理实证方法会议（*EMNLP*）论文集，第 4563–4568 页，线上举行。计算语言学协会。2，3

Daniel Khashabi、Chitta Baral、Yejin Choi 和 Hannaneh Hajishirzi。2022。重新构想面向 GPT-3 的指令式提示词。载于计算语言学协会 *ACL 2022* 年会议论文集：研究成果，第 589–612 页。3，8

Yoon Kim。2014。用于句子分类的卷积神经网络。载于*2014* 年自然语言处理实证方法会议论文集（*EMNLP 2014*），*2014* 年 *10* 月 *25* 日至 *29* 日，卡塔尔多哈，*ACL* 特别兴趣组 *SIGDAT* 会议，第 1746–1751 页。ACL。1

TakeshiKojima、ShixiangShaneGu、MachelReid、Yutaka Matsuo 和 Yusuke Iwasawa。2022。大型语言模型是零样本推理器。*arXiv* 预印本 *arXiv:2205.11916*。3，8

Bing Liu、Minqing Hu 和 Junsheng Cheng。2005。Opinion observer：分析并比较网络上的意见。载于第 *14* 届万维网国际会议论文集，第 342–351 页。4

Bing Liu 和 Lei Zhang。2012。意见挖掘与情感分析综述。载于 Charu C. Aggar- wal 和 ChengXiang Zhai 编，文本数据挖掘，第 415–463 页。Springer。1

Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。2021。预训练、提示与预测：自然语言处理领域提示方法的系统综述。*arXiv* 预印本 *arXiv:2107.13586*。3，8

Zhiheng Lyu、Zhijing Jin、Rada Mihalcea、Mrinmaya Sachan 和BernhardSchölkopf。2022 年。大型语言模型能否语言模型能区分原因与结果吗？在*UAI 2022* 年因果表征学习研讨会。6

John Morris、Eli Lifland、Jin Yong Yoo、Jake Grigsby、Di Jin 和 Yanjun Qi。2020 年。TextAttack：用于对抗性攻击、数据增强和自然语言处理中的对抗性训练的框架。载于《*2020* 年自然语言处理经验方法会议论文集：系统演示》，第 119–126 页，线上举办。计算语言学协会。9

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul F. Christiano、Jan Leike 和 Ryan Lowe。2022 年。训练语言模型遵循人类反馈中的指令。*CoRR*，abs/2203.02155。2，3，4，5，8，10

Soujanya Poria、Devamanyu Hazarika、Navonil Majumder 和 Rada Mihalcea。2020。冰山一角之下：情感分析研究面临的当前挑战与新方向。*CoRR*，abs/2005.00357。1

Kumar Ravi 和 Vadlamani Ravi。2015。观点挖掘与情感分析综述：任务、方法与应用。*Knowl.Based Syst.*，89:14–46。1

Peter Salovey 和 John D Mayer。2004。情绪智力。Dudepublishing。2

Ajay B Satpute、Jocelyn Shu、Jochen Weber、Mathieu Roy 和 Kevin N Ochsner。2013。情感经验自我报告的功能性神经架构。生物精神病学，73(7):631–638。2

Taylor Sorensen、Joshua Robinson、Christopher Rytting、Alexander Shaw、Kyle Rogers、Alexia Delorey、Mahmoud Khalil、Nancy Fulda 和 David Wingate。2022。一种无需真实标签的提示工程信息论方法。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第819–862页。8

Stephen P Stich。1983年。从民间心理学到认知科学：反对信念的案例。麻省理工学院出版社。1，2

Jason Wei 和 Kai Zou。2019年。EDA：用于提升文本分类任务表现的简易数据增强技术。*arXiv* 预印本 *arXiv:1901.11196*。9

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、R'emi Louf、Morgan Funtowicz 和 J amie Brew。2019年。HuggingFace 的 Transformer模型：最先进的自然语言处理。*arXiv* 预印本 *arXiv:1910.03771*。10

Hui Wu 和 Xiaodong Shi。2022。用于跨领域情感分析的对抗性软提示调优。载于

计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 2438–2447 页。8

Wenpeng Yin、Jamaal Hay 和 Dan Roth。2019年。零样本文本分类基准测试：数据集、评估与蕴涵方法。载于 *2019*年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第3914–3923页，中国香港。计算语言学协会。1

Xiang Zhang、Junbo Zhao 和 Yann LeCun。2015年。用于文本分类的字符级卷积网络。《神经信息处理系统进展》，第28卷。1，2，3，9

Qi Zhu、Bing Li、Fei Mi、Xiaoyan Zhu 和 Minlie Huang。2022年。用于对话状态跟踪的持续提示调优。载于 计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第1124–1137页。8

A 因果提示的设计

A.1 一组三个因果提示的期望标准

基于三种不同的心理过程，我们希望为后续分析设计三个因果提示。在拟定提示词时，我们考虑以下期望标准：（1）每个提示都应明确代表其对应的因果过程。（2）提示应以连贯的自然语言表达。（3）对这三个提示进行跨因果性比较应具有意义，这意味着我们需要考虑其他类型的噪声，因为这些噪声会导致大语言模型产生不同的响应。

A.2 总体提示选择过程

遵循这三项原则，我们谨慎地进行提示选择过程，并参考了提示设计领域的广泛文献，包括相关综述（Liu 等人，2021）和近期工作（Khashabi 等人，2022；Kojima 等人，2022）。

设计提示词有多种方式（Liu 等人，2021）。主要有两种方式：离散提示设计（例如，Sorensen 等人，2022；Gao 等人，2021）和连续提示设计（例如，Wu 和 Shi，2022；Zhu 等人，2022）。对于连续提示设计，我们无法将其用于本研究，因为我们需要保持提示词的可解释性，以便人们能够识别并检查其所表示的因果过程（除非未来工作能提出一种准确的从嵌入到因果过程的分类器）。对于离散提示设计，我们对因果提示的一些机器生成释义进行了实证评估，发现纯机器生成的文本往往无法同时很好地满足标准（1）和（2）。纯人工设计虽然能使提示词完美满足标准（1）和（2），但可能会引入某些系统性的人类偏见（Cao 等人，2022；Kavumba 等人，2022；Elazar 等人，2021）。

为兼顾两方面的优势，我们选择采用大语言模型辅助的人工框架设计来构建提示词。我们使用当前可用的最强自回归大语言模型，即近期推出的 GPT3-Instruct （ Ouyang 等人，2022 ），首先向其描述因果情景（“这是一个人……的过程”），然后使用响应生成时默认的 0.7 温度值，收集它对每个因果过程进行改写的 10 种不同版本。

然后，我们根据每个提示词的 10 种不同版本，筛除不符合标准 (1) 和 (2) 的版本。接着，我们尽量少改动实词来重新措辞候选版本，同时遵循如何构建有效指令式提示词的建议（Khashabiet al.， 2022）。如图 1所示，我们将提示词构建成合理的指令，同时也确保满足标准 (3)，尽可能控制这三个提示词在结构、措辞和长度方面保持一致。

A.3 检查提示词是否得到良好控制

尽管我们的主要关注点是不同因果性质的提示词之间的差异，但提示词表现也可能受到其他混杂因素影响而产生各种噪声。例如，另外两个可能影响表现的重要因素是提示词的长度，以及词语中带有轻微扰动的释义改写。因此，接下来我们进行一些额外分析：(1) 冗长程度（即保持含义不变但改变提示词的长度）如何以类似方式对三个提示词产生负面影响；(2) 提示词中的扰动如何也以一致的方式降低这三个提示词的表现。

请注意，在本文的研究范围内，我们主要考虑这两种噪声来源，但未来的工作可以探索其他因素。

比较长度。对于每个因果提示，我们都制作了长短两种变体，

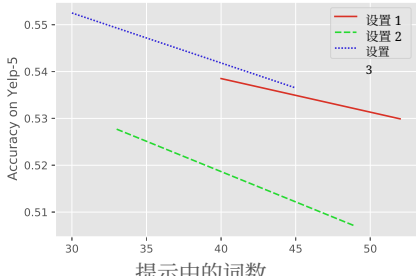
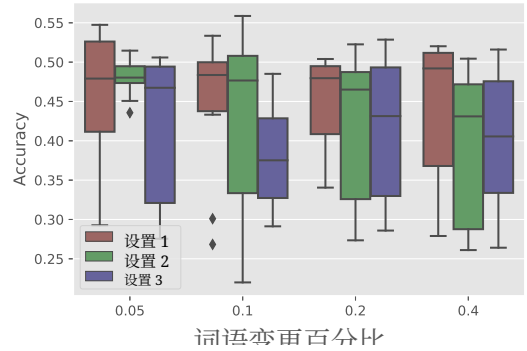


图 5：我们确认，三个因果提示的较短版本比长版本表现更好，而图 1 中的三个短版本，是我们在满足三项期望标准的同时所能构思出的最佳受控提示词。

它们的语义和大部分措辞相同，只有冗长程度不同，如图5所示。我们在图5中绘出了三组提示词的表现，可以看出，较短的提示词往往准确率更高，而长度相近的提示词表现相当。因此，我们确认，如图1所示，我们选取的这三个简短提示词构成了一组良好的因果提示，较好地满足了最后一项标准。

与轻微扰动进行比较。 我们还检查了提示词相对于其轻微扰动变体的表现是否受控。我们使用简易数据增强（Wei 和 Zou，2019年）生成了一组包含十种提示词轻微扰动的集合，方法包括同义词替换、随机插入、随机交换和随机删除。我们使用 TextAttack Python 软件包中的实现（Morris 等人，2020）。我们在 6 中展示了它们的表现，并确认这些扰动会降低表现。因此，我们确认，在具有不同扰动强度的所有可能局部扰动中，我们的提示词是能够满足全部三项标准且表现最佳的提示词。

确定最终提示词。 我们调整的三种因果提示词如图所示1 （分别记为 C1、C2 和 C3）。我们首先确保满足标准 (1) 和 (2)，然后通过执行标准 (3)，从结构、措辞和长度三个方面控制这三种提示词。为确保这三种提示词得到充分控制，我们还在附录A中进行了额外分析，例如，冗长度（即保持含义不变但改变



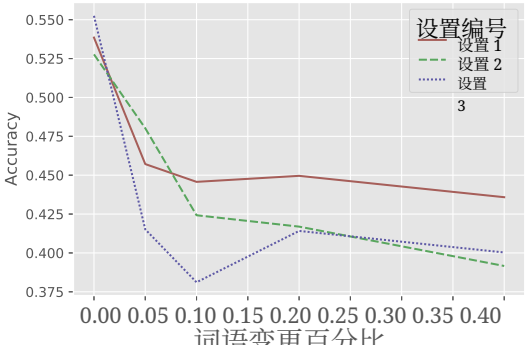


图 6：我们确认，我们选择的提示词相对可控，因为它们优于经过扰动的版本，同时满足三项期望标准。

长度）会以相似的方式对这三种提示词产生负面影响；提示词中的扰动也会一致地降低这三种提示词的表现。我们的提示词在长度、措辞和自然度方面都较为接近。

B 实验细节

B.1 数据集

我们在广泛使用的 Yelp 情感分类数据集（Zhang 等人，2015）上开展实验。原始 Yelp 数据集的训练集中有 65 万个样本，测试集中有 5 万个样本。该数据集的五个标签（1–5）各自对应的样本数量相同。

为了降低计算成本和碳成本，我们在保持标签平衡的同时对数据集进行下采样。从原始测试集中随机抽取1万个测试样本。论文中的所有表现数据均基于这1万个测试样本报告。

C1：评分影响评论 简短版： 我刚在一家餐厅吃完饭。随后打开了 Yelp 应用。我先给出了评分，然后用以下评论说明理由： [评论文本] 这条评论解释了我为什么给它 [让 GPT 补全] 详细版： 我刚在一家餐厅吃完饭。随后打开了 Yelp 应用。我先按 1 到 5 星给出了评分，然后用以下评论说明评分理由： [评论文本] 这条评论解释了我为什么给它评为 [让 GPT 补全]**C2：评论影响评分** 简短版： 我刚在一家餐厅吃完饭。随后打开了 Yelp 应用。我先写下了以下评论： [评论文本] 然后读了一遍自己的评论，最后给出了 [让 GPT 补全] 详细版： 我刚在一家餐厅吃完饭。随后打开了 Yelp 应用。我先写下了以下评论： [评论文本] 然后根据这条评论，按 1 到 5 星给出了评分。我认为这家餐厅值得评为[让 GPT 补全]**C3：另一个人根据评论猜测评分** 简短版： 我打开了 Yelp 应用，开始阅读一家餐厅的评论。我看到一位用户写了这条评论： [评论文本] 我认为这位用户给出了 [让 GPT 补全] 详细版： 我打开了 Yelp 应用，开始阅读一家我想去的餐厅的评论。我看到一位用户写了这条评论： [评论文本] 我认为这位用户给出了（满分 5 星，1 星起评） [让 GPT 补全]

表 3：每种设置对应的短提示词和长提示词。

B.2 模型详情

对于 GPT3-Instruct 的使用（Ouyang 等人，2022），我们使用 OpenAI 提供的 API，[6] 以及在我们发表这篇论文时推出的最新模型“`davinci-text-002`”，其参数量为 175B。

为确保结果可复现，我们将所有 OpenAI API 调用的温度设为 0。在推理时，我们获取概率最高的 5 个候选项及其对数概率。对于标签 1、2、.....、5 的概率，我们将其英文单词形式（例如“one”）和阿拉伯数字形式（例如“1”）的对数概率相加。

由于我们不对模型进行微调，而只是调用 API 以推理模式运行，因此能耗相对较低；总共大约调用 API 20–30K 次，预算约为 400 美元。

B.3 实现细节

由于模型对不同标签可能存在不同的固有偏好（例如，更倾向于预测“3”这样的中性得分），我们使用训练集中的一小部分（1K 个样本）来学习一个附加参数 $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_5)$，该参数通过标量 λ_i 调整模型的预测 $P(Y = y_i|X)$。对于损失函数 $\mathcal{L}$，我们通过优化 $L_1(P(\arg\max_y(\lambda_y P(Y = y|X = x))) P(Y))$ 来学习 $\boldsymbol{\lambda}$。表 4 列出了学得的 $\boldsymbol{\lambda}$ 值。

B.4 信息熵

给定测试集中每个样本 $x_i \in \boldsymbol{D}$ 的预测分布 $P(Y|X = x_i)$ x_i，我们计算

[6] https://openai.com/api/

设置	标签先验				
	λ_1	λ_2	λ_3	λ_4	λ_5
C1	0.73	0.22	0.03	0.004	0.01
C2	0.07	0.32	0.44	0.07	0.10
C3	0.13	0.39	0.09	0.08	0.31

表 4：GPT3-Instruct 对标签 1 – 5 的预测所学习到的缩放因子。

该分布的信息熵由

$$H(X) = -\sum P(Y|x_i) \log P(Y|x_i)\,, \qquad (2)$$

计算，它可以衡量模型对所有可能标签的不确定性。请注意，熵与困惑度密切相关，困惑度的计算方式为 $\text{PPL}(P) = 2^{H(P)}$。

B.5 显式观点词词数

情感的一个简单指标是那些带有强烈观点的词语集合。为了统计观点词出现的次数，我们首先使用 GPT2-large 分词器对评论进行分词，该分词器由Transformer模型 库提供（Wolf 等人，2019年），然后移除所有非标准英文字母、阿拉伯数字和常见符号（如“-”）以外的符号，再与给定的观点词列表进行匹配。

C 更多关于预测一致或多样的评论示例

查看更多预测一致性高或预测多样性高的评论示例，请参见表 5。

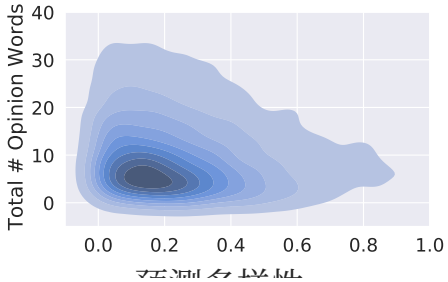


图7：观点词总数与预测多样性的联合分布。

D 扩展分析

D.1 观点词与多样性完整图示

对于 Yelp 的全部 1万 个测试样本，我们在图7中绘制了观点词总数与三种因果提示下的预测多样性之间的所有配对。

D.2 不同子集的 LIWC 分析

我们使用 LIWC 词语类别来分析表格中不同子集的语言学特征 6和 7。

预测高度一致的评论**R** **eview:** *Terribl*服务 *ice!! Rud*经理 ！ *L*失去一位顾客一生。而经理并不在乎。 **评分：** C1：1 星。C2：1 星。C3：1 星。GT：1 星。
评论： 我最初去那里的一两次，他们更接近四星，但最近这次让他们只能得到三星。官保套餐很一般。我真的讨厌肉质这么软烂，而那份牛肉就是如此。这只是一家装修不错、普通的中餐馆。**评分：** C1：3 星。C2：3 星。C3：3 星。GT：3 星。
评论： 出色的沟通、服务和高质量维修!
我再满意不过了！免费提供代步车！免赔额全免或部分免！仅用 *8* 个工作日，就让我的 *2006 Solara* 焕然一新！我强烈推荐他们！告诉他们是我介绍你来的！！！ **评分：** C1：5 星。C2：5 星。C3：5 星。GT：5 星。
评论： 寿司真的很好吃，不过有点贵。顾客服务也有点奇怪。不过总的来说，这地方还不错。 **评分：** C1：3 星。C2：3 星。C3：3 星。GT：4 星 rs.
预测多样性较高的评论 **评论：** 坦白 .... 我没在这里吃过. 说真的……我确实撞见过比尔·克林顿在私人餐厅吃饭，随即就被美国特勤局告知离开。
顺便说一下……他看起来很享受他的餐点。**评分：** C1：1 星。C2：1 星。C3：5 星。GT：5 星。
评论： “本来有望成为一家很棒的餐厅，但遗憾的是，结果令人极其失望。调酒师很糟糕，态度也不好，完全不知道自己在做什么。要求五分熟的汉堡端上来时却已经全熟了，薯条也是凉的。店里甚至都不忙，所有员工看起来却都手忙脚乱。不会再来了。**R** **ating:** C1 2 星星. C2 1 星形. C3 1 星形. GT 1 星级.
评论： 等待时间太长了！！他们应该扩大露台，如果等候超过一个小时，就应该提供一杯免费饮料！谁会为了出去吃饭而等一个小时！这就是我不来这里的原因。她想来这里！ **评分：** C1：1 星。C2：2 星。C3：1 星。GT：2 星。

表5：由三个因果提示（C1、C2 和 C3）生成的预测具有高度一致性或多样性的评论示例。我们还展示了真实标签（GT）。

类别		得分（全部）	得分（子集）		
			相同且正确	相同且错误	多样
词数		138.33 ± 125.20	144.46 ± 126.86	126.85 ± 118.85	131.55 ± 125.40
摘要语言变量					
每句词数					
语言学维度					
功能词总数	{这、的、和、一、到}	59.92 ± 56.71	62.81 ± 57.73	54.66 ± 53.63	56.57 ± 56.09
代词总数	{那个，这个，我们，哪个，它}	10.90 ± 12.20	11.61 ± 12.61	9.72 ± 11.45	9.96 ± 11.50
人称代词	{我们，我们，他们，他们，我们}	5.89 ± 7.49	6.34 ± 7.80	5.11 ± 7.01	5.32 ± 6.90
第一人称单数	{i, mine, my, im, me}	1.61 ± 2.78	1.75 ± 2.94	1.41 ± 2.60	1.42 ± 2.45
第一人称复数	{我们，我们的，我们，让我们，我们自己}	1.37 ± 3.02	1.47 ± 3.13	1.17 ± 2.83	1.29 ± 2.86
第二人称	{你，你的，你，你，你}	1.22 ± 2.11	1.27 ± 2.21	1.14 ± 1.96	1.12 ± 1.96
第三人称单数	{他的, 她的, 他, 她, 他}	0.72 ± 2.12	0.83 ± 2.24	0.53 ± 2.02	0.61 ± 1.81
第三人称复数	{他们，他们，他们自己，他们的，他们的}	0.96 ± 1.63	1.02 ± 1.66	0.86 ± 1.54	0.88 ± 1.61
非人称代词	{那个，这个，哪个，它，这些}	5.00 ± 5.54	5.26 ± 5.71	4.59 ± 5.21	4.63 ± 5.32
冠词	{the, a an}	9.51 ± 9.48	9.85 ± 9.58	8.84 ± 9.08	9.18 ± 9.57
介词	{的，向，在，为，与}	15.35 ± 15.50	16.05 ± 15.73	13.98 ± 14.83	14.62 ± 15.36
助动词	{是、为、能、拥有}	11.34 ± 10.45	11.85 ± 10.60	10.47 ± 9.89	10.67 ± 10.45
常见副词	{此类、也、当.....时、仅、哪里}	6.19 ± 6.22	6.48 ± 6.36	5.64 ± 5.81	5.84 ± 6.13
连词	{和、如、或、也、但是}	8.25 ± 7.92	8.60 ± 8.03	7.55 ± 7.41	7.91 ± 8.05
否定词	{非，不包含，没有，无法，负向}	2.33 ± 2.69	2.48 ± 2.78	2.06 ± 2.46	2.17 ± 2.62
其他语法					
常见动词	{is, are, be, using, based}	19.85 ± 19.00	20.91 ± 19.39	17.97 ± 18.02	18.58 ± 18.53
常见形容词	{as, 不同, new, more, than}	5.49 ± 5.30	5.67 ± 5.30	5.13 ± 5.13	5.33 ± 5.46
比较	{as, 不同, more, than, most}	2.68 ± 3.07	2.79 ± 3.10	2.40 ± 2.87	2.64 ± 3.14
疑问词	{哪种、何时、何处、如何、是否}	1.28 ± 1.80	1.35 ± 1.87	1.16 ± 1.66	1.21 ± 1.72
数字	{two, one, first, three, single}	0.91 ± 1.45	0.96 ± 1.48	0.78 ± 1.36	0.87 ± 1.43
量词	{更多，每个，两者都，最多，全部}	2.60 ± 2.97	2.67 ± 2.97	2.46 ± 2.89	2.54 ± 3.03
心理过程					
情感过程	{好，重要，问题，能量，问题}	4.98 ± 4.58	5.19 ± 4.62	4.62 ± 4.32	4.71 ± 4.68
积极情绪	{好，重要，能量，更好，支持}	3.59 ± 3.53	3.71 ± 3.59	3.40 ± 3.31	3.45 ± 3.54
消极情绪	{problems, problem, low, 批评的, difficult}	1.26 ± 1.85	1.35 ± 1.91	1.09 ± 1.73	1.15 ± 1.74
焦虑	{不确定性, 压力, 不确定因素, 风险, 风险}	0.14 ± 0.44	0.15 ± 0.45	0.12 ± 0.38	0.13 ± 0.46
愤怒	{批评的, 攻击, 争论, 主导的, 争论}	0.27 ± 0.69	0.29 ± 0.72	0.21 ± 0.66	0.25 ± 0.63
悲伤	{低，较低，失败，缺失，受苦}	0.27 ± 0.60	0.28 ± 0.62	0.25 ± 0.57	0.25 ± 0.57
社会过程	{我们, 我们的, 提供, 他们, 提供}	7.54 ± 9.52	8.14 ± 10.02	6.44 ± 8.62	6.86 ± 8.67
家人	{家人, 家庭, 父母, 妊娠, 儿子}	0.15 ± 0.55	0.17 ± 0.60	0.12 ± 0.45	0.14 ± 0.49
朋友	{联系, 街区, 邻近的, 邻居, 约会}	0.33 ± 0.70	0.33 ± 0.70	0.31 ± 0.69	0.32 ± 0.70
女性指代	{女性, 她, 女性, 女性, 她}	0.51 ± 1.75	0.58 ± 1.91	0.39 ± 1.51	0.42 ± 1.43
男性指代	{他的, 男性, 他, 男性, 儿子}	0.48 ± 1.43	0.53 ± 1.51	0.38 ± 1.34	0.43 ± 1.25
认知过程	{使用、基于、或、已使用、结果}	11.93 ± 11.88	12.50 ± 12.14	10.82 ± 10.89	11.35 ± 11.93
洞察	{信息, 学习, 分析, 知识, 识别}	1.64 ± 2.20	1.72 ± 2.25	1.48 ± 2.09	1.53 ± 2.13
因果关系	{using, based, used, results, use}	1.30 ± 1.81	1.37 ± 1.83	1.16 ± 1.72	1.23 ± 1.80
差异	{问题，问题，需要，可以，如果}	1.80 ± 2.42	1.90 ± 2.52	1.60 ± 2.17	1.70 ± 2.32
暂定	{或, 大多数, 可能, 一些, 任何}	2.77 ± 3.25	2.86 ± 3.30	2.58 ± 3.12	2.66 ± 3.22
确定性	{all, 准确性, 具体, 准确, 总计}	1.51 ± 1.86	1.61 ± 1.91	1.32 ± 1.69	1.38 ± 1.85
区分	{or, 不同, not, than, 其他}	4.33 ± 4.58	4.51 ± 4.71	3.99 ± 4.18	4.16 ± 4.57
知觉过程	{show, 图像, 搜索, 模糊, image}	2.68 ± 3.26	2.79 ± 3.30	2.44 ± 3.12	2.59 ± 3.26
See	{显示, 图像, 搜索, 图像, 显示}	0.87 ± 1.46	0.90 ± 1.46	0.80 ± 1.46	0.88 ± 1.47
Hear	{噪声, 嘈杂的, 音乐, 声音, 语音}	0.53 ± 1.14	0.57 ± 1.21	0.46 ± 0.97	0.49 ± 1.08
Feel	{模糊, 柔性的, 重量, 加权的, 手}	0.57 ± 1.04	0.58 ± 1.05	0.53 ± 0.98	0.56 ± 1.07
生物过程	{临床，表达，面部，医学，身体}	3.64 ± 4.27	3.68 ± 4.24	3.52 ± 4.24	3.63 ± 4.40
Body	{面部, 血液, 手, 心脏, 神经元}	0.32 ± 0.81	0.34 ± 0.85	0.29 ± 0.74	0.31 ± 0.77
健康	{临床, 医学, 身体, 健康, 诊断}	0.24 ± 0.73	0.27 ± 0.79	0.21 ± 0.64	0.21 ± 0.57
性别	{前列腺, 妊娠, 性别, 卵巢, 性唤起}	0.03 ± 0.21	0.03 ± 0.22	0.02 ± 0.18	0.03 ± 0.22
摄入	{表达, 水, 重量, 表达, 表达}	2.94 ± 3.85	2.95 ± 3.83	2.90 ± 3.86	2.98 ± 3.92
驱动力	{我们，我们的方法，我们的，第一，结束}	7.28 ± 7.69	7.68 ± 7.89	6.50 ± 7.03	6.87 ± 7.62
归属	{我们, 我们的, 社交, 沟通, 互动}	2.20 ± 3.67	2.33 ± 3.79	1.93 ± 3.38	2.08 ± 3.55
成就	{第一，工作，高效，获得，更好}	1.26 ± 1.61	1.35 ± 1.65	1.13 ± 1.48	1.15 ± 1.59
权力	{over, high, order, large, important}	1.91 ± 2.41	2.02 ± 2.46	1.66 ± 2.17	1.84 ± 2.47
奖励	{方法, 获得, 方法, 更好, 最佳}	2.13 ± 2.27	2.23 ± 2.33	1.97 ± 2.12	1.99 ± 2.25
Risk	{问题，问题，安全，困难，缺乏}	0.40 ± 0.78	0.43 ± 0.82	0.35 ± 0.71	0.36 ± 0.71
时间取向					
过去聚焦	{已使用, 是, 已经, 曾经是, 获得}	8.20 ± 9.58	8.71 ± 9.79	7.33 ± 9.34	7.57 ± 9.06
现在聚焦	{是，是，是，能，有}	9.53 ± 9.41	9.96 ± 9.61	8.74 ± 8.60	9.06 ± 9.52
未来焦点	{may, then, will, 预测, 未来}	1.01 ± 1.41	1.07 ± 1.45	0.94 ± 1.37	0.90 ± 1.35
相对性	{在，位于，在，方法，新的}	14.93 ± 15.16	15.69 ± 15.51	13.58 ± 14.37	14.01 ± 14.73
运动	{方法, 方法, 行为, 变化, 增加}	2.44 ± 2.94	2.59 ± 3.03	2.15 ± 2.73	2.29 ± 2.86
空间	{在.....里面，在.....上面，在.....，进入，两者都}	8.19 ± 8.54	8.56 ± 8.71	7.59 ± 8.32	7.69 ± 8.19
Time	{新，现在，第一，何时，然后}	4.63 ± 5.43	4.91 ± 5.63	4.14 ± 4.99	4.30 ± 5.20
个人关注					
Work	{表现, 学习, 分析, 论文, 应用}	1.40 ± 2.12	1.53 ± 2.27	1.20 ± 1.77	1.23 ± 1.97
休闲	{小说, 表达, 频道, 视频, 游戏}	1.43 ± 2.17	1.45 ± 2.22	1.45 ± 2.17	1.37 ± 2.03
Home	{地址, 家人, 家庭, 街区, 邻近的}	0.49 ± 1.21	0.51 ± 1.24	0.47 ± 1.23	0.45 ± 1.06
金钱	{调查，成本，调查过，免费，经济}	1.03 ± 1.69	1.08 ± 1.73	0.99 ± 1.71	0.94 ± 1.51
宗教	{信念, 道德, 牺牲, 猴子, 不可知论者}	0.04 ± 0.21	0.04 ± 0.22	0.03 ± 0.21	0.03 ± 0.19
死亡	{死亡，死亡，死亡，死亡，死亡}	0.04 ± 0.21	0.04 ± 0.20	0.03 ± 0.22	0.04 ± 0.22
非正式语言	{好吧，哦，那个，嗯}	0.48 ± 0.92	0.48 ± 0.93	0.49 ± 0.91	0.49 ± 0.89
脏话	{智力迟缓，笨蛋，螺丝钉，螺丝，智力迟缓}	0.09 ± 0.36	0.09 ± 0.38	0.08 ± 0.35	0.08 ± 0.34
网络俚语	{o, da, em, k, mm}	0.12 ± 0.44	0.12 ± 0.45	0.13 ± 0.44	0.12 ± 0.41
赞同	{k，确实，同意，绝对，酷}	0.14 ± 0.42	0.14 ± 0.42	0.13 ± 0.41	0.15 ± 0.42
非流利现象	{嗯，好吧，嗯，嗯，呃，啊}	0.10 ± 0.35	0.10 ± 0.34	0.11 ± 0.37	0.10 ± 0.36
填充词	{rrani, rranr}	0.02 ± 0.14	0.02 ± 0.15	0.01 ± 0.12	0.02 ± 0.14

表6：所有Yelp数据上的LIWC分析结果（**得分（全部）**），以及不同子集（**相同且正确**、**相同且错误**和**多样**）。

类别		得分（全部多样）	得分（子集）		
			C1≠GT	C2≠GT	C3≠GT
词数					
摘要语言变量					
每句词数		131.55 ± 125.40	124.21 ± 121.95	128.16 ± 122.18	128.97 ± 127.07
语言学维度					
功能词总数	{这、的、和、一、到}	56.57 ± 56.09	53.21 ± 54.10	55.27 ± 54.66	55.11 ± 56.56
代词总数	{那个，这个，我们，哪个，它}	9.96 ± 11.50	9.39 ± 10.94	9.76 ± 11.35	9.71 ± 11.67
人称代词	{我们，我们，他们，他们，我们}	5.32 ± 6.90	5.01 ± 6.48	5.19 ± 6.80	5.19 ± 6.98
第一人称单数	{i, mine, my, im, me}	1.42 ± 2.45	1.33 ± 2.34	1.37 ± 2.41	1.40 ± 2.47
第一人称复数	{我们，我们的，我们，让我们，我们自己}	1.29 ± 2.86	1.18 ± 2.74	1.29 ± 2.94	1.18 ± 2.72
第二人称	{你、你的、你、你、你们}	1.12 ± 1.96	1.12 ± 1.90	1.07 ± 1.89	1.14 ± 2.07
第三人称单数	{他的, 她的, 他, 她, 他}	0.61 ± 1.81	0.58 ± 1.73	0.58 ± 1.75	0.63 ± 1.92
第三人称复数	{他们，他们，他们自己，他们的，他们的}	0.88 ± 1.61	0.79 ± 1.35	0.89 ± 1.68	0.84 ± 1.60
非人称代词	{那个，这个，哪个，它，这些}	4.63 ± 5.32	4.36 ± 5.13	4.55 ± 5.33	4.51 ± 5.35
冠词	{the, a an}	9.18 ± 9.57	8.60 ± 9.28	8.99 ± 9.34	9.05 ± 9.74
介词	{的，向，在，为，与}	14.62 ± 15.36	13.68 ± 14.63	14.37 ± 15.22	14.37 ± 15.54
助动词	{is, are, be can, have}	10.67 ± 10.45	10.15 ± 10.19	10.31 ± 10.08	10.28 ± 10.47
常见副词	{如此，也，当，仅，当……时}	5.84 ± 6.13	5.48 ± 5.85	5.76 ± 6.14	5.57 ± 5.82
连词	{和、如、或、也、但是}	7.91 ± 8.05	7.41 ± 7.89	7.69 ± 7.62	7.68 ± 8.04
否定词	{非，不包含，没有，无法，负向}	2.17 ± 2.62	2.08 ± 2.59	2.11 ± 2.51	2.12 ± 2.66
其他语法					
常见动词	{是，是，成为，使用，基于}	18.58 ± 18.53	17.55 ± 17.90	18.08 ± 17.85	18.07 ± 18.93
常见形容词	{如，不同，新的，更多，比}	5.33 ± 5.46	4.99 ± 5.16	5.25 ± 5.50	5.20 ± 5.41
比较	{as, 不同, more, than, most}	2.64 ± 3.14	2.47 ± 3.01	2.64 ± 3.13	2.52 ± 3.08
疑问词	{哪种、何时、何处、如何、是否}	1.21 ± 1.72	1.12 ± 1.62	1.18 ± 1.76	1.20 ± 1.72
数字	{two, one, first, three, single}	0.87 ± 1.43	0.85 ± 1.49	0.87 ± 1.42	0.86 ± 1.41
量词	{更多，每个，两者都，最多，全部}	2.54 ± 3.03	2.44 ± 2.92	2.48 ± 3.05	2.42 ± 2.89
心理过程					
情感过程	{好，重要，问题，能量，问题}	4.71 ± 4.68	4.35 ± 4.30	4.62 ± 4.64	4.54 ± 4.57
积极情绪	{健康、重要、精力、更好、支持}	3.45 ± 3.54	3.14 ± 3.30	3.38 ± 3.52	3.33 ± 3.47
消极情绪	{problems, problem, low, 批评的, difficult}	1.15 ± 1.74	1.10 ± 1.63	1.11 ± 1.65	1.10 ± 1.69
焦虑	{不确定性, 压力, 不确定因素, 风险, 风险}	0.13 ± 0.46	0.12 ± 0.40	0.12 ± 0.38	0.14 ± 0.50
愤怒	{批评的, 攻击, 争论, 主导的, 争论}	0.25 ± 0.63	0.23 ± 0.58	0.23 ± 0.59	0.22 ± 0.60
悲伤	{低，降低，失败，缺失，遭受}	0.25 ± 0.57	0.26 ± 0.60	0.23 ± 0.52	0.24 ± 0.57
社会过程	{我们，我们的，提供，他们，提供}	6.86 ± 8.67	6.55 ± 8.45	6.66 ± 8.34	6.79 ± 8.99
家人	{家人, 家庭, 父母, 妊娠, 儿子}	0.14 ± 0.49	0.13 ± 0.44	0.13 ± 0.44	0.15 ± 0.54
朋友	{联系, 街区, 邻近的, 邻居, 约会}	0.32 ± 0.70	0.34 ± 0.73	0.30 ± 0.68	0.31 ± 0.69
女性指代	{女性, 她, 女性, 女性, 她}	0.42 ± 1.43	0.40 ± 1.35	0.38 ± 1.39	0.43 ± 1.52
男性指代	{他的, 男性, 他, 男性, 儿子}	0.43 ± 1.25	0.43 ± 1.20	0.41 ± 1.26	0.46 ± 1.34
认知过程	{使用、基于、或、使用的、结果}	11.35 ± 11.93	10.71 ± 11.39	11.09 ± 11.85	10.98 ± 11.73
洞察	{information, 学习, 分析, knowledge, recognition}	1.53 ± 2.13	1.40 ± 1.97	1.53 ± 2.09	1.50 ± 2.24
因果关系	{使用、基于、使用、结果、使用}	1.23 ± 1.80	1.13 ± 1.69	1.21 ± 1.76	1.20 ± 1.76
差异	{问题，问题，需要，可能，如果}	1.70 ± 2.32	1.59 ± 2.22	1.66 ± 2.26	1.65 ± 2.25
暂定	{或, 大多数, 可能, 一些, 任何}	2.66 ± 3.22	2.51 ± 3.09	2.60 ± 3.19	2.54 ± 3.11
确定性	{all, 准确性, 具体, 准确, 总计}	1.38 ± 1.85	1.37 ± 1.82	1.32 ± 1.83	1.34 ± 1.80
区分	{or, 不同, not, than, 其他}	4.16 ± 4.57	3.93 ± 4.52	4.08 ± 4.48	4.02 ± 4.47
知觉过程	{show, 图像, 搜索, 模糊, 图像}	2.59 ± 3.26	2.41 ± 3.00	2.52 ± 3.27	2.52 ± 3.25
See	{显示, 图像, 搜索, 图像, 显示}	0.88 ± 1.47	0.82 ± 1.33	0.86 ± 1.54	0.85 ± 1.50
Hear	{噪声, 嘈杂的, 音乐, 声音, 语音}	0.49 ± 1.08	0.46 ± 0.97	0.47 ± 1.14	0.49 ± 1.03
Feel	{模糊, 柔性的, 重量, 加权的, 手}	0.56 ± 1.07	0.50 ± 0.91	0.55 ± 1.09	0.55 ± 1.12
生物过程	{临床, 表达, 面部, 医学, 身体}	3.63 ± 4.40	3.36 ± 4.20	3.45 ± 4.22	3.49 ± 4.35
Body	{面部, 血液, 手, 心脏, 神经元}	0.31 ± 0.77	0.28 ± 0.75	0.28 ± 0.71	0.30 ± 0.76
健康	{临床, 医学, 身体, 健康, 诊断}	0.21 ± 0.57	0.19 ± 0.52	0.21 ± 0.55	0.22 ± 0.58
性别	{前列腺, 妊娠, 性别, 卵巢, 性唤起}	0.03 ± 0.22	0.04 ± 0.27	0.03 ± 0.21	0.03 ± 0.25
摄入	{表达, 水, 重量, 表达, 表达}	2.98 ± 3.92	2.76 ± 3.78	2.85 ± 3.79	2.83 ± 3.89
驱动力	{我们，我们的方法，我们的，第一，结束}	6.87 ± 7.62	6.38 ± 7.28	6.79 ± 7.59	6.75 ± 7.65
归属	{we, our, social, communication, 互动}	2.08 ± 3.55	1.94 ± 3.44	2.06 ± 3.57	1.97 ± 3.38
成就	{first, 工作, 高效, 获得, 更好}	1.15 ± 1.59	1.07 ± 1.50	1.13 ± 1.56	1.14 ± 1.63
权力	{over, high, order, large, important}	1.84 ± 2.47	1.65 ± 2.18	1.79 ± 2.49	1.90 ± 2.62
奖励	{方法, 获得, 方法, 更好, 最佳}	1.99 ± 2.25	1.87 ± 2.16	1.96 ± 2.24	1.92 ± 2.24
Risk	{问题, 问题, 安全, 困难, 缺乏}	0.36 ± 0.71	0.36 ± 0.71	0.39 ± 0.74	0.37 ± 0.73
时间取向					
过去聚焦	{已使用, 是, 已经, 曾经是, 获得}	7.57 ± 9.06	7.16 ± 8.95	7.28 ± 8.78	7.28 ± 9.12
现在聚焦	{是，是，是，能，有}	9.06 ± 9.52	8.56 ± 9.04	8.94 ± 8.98	8.87 ± 9.96
未来焦点	{may, then, will, 预测, 未来}	0.90 ± 1.35	0.90 ± 1.34	0.89 ± 1.29	0.91 ± 1.41
相对性	{in, on, at, 方法, new}	14.01 ± 14.73	13.21 ± 14.12	13.79 ± 14.58	13.78 ± 15.10
运动	{方法, 方法, 行为, 变化, 增加}	2.29 ± 2.86	2.22 ± 2.87	2.23 ± 2.77	2.29 ± 3.07
空间	{在……里面，在……上面，在，在……里面，两者}	7.69 ± 8.19	7.20 ± 7.86	7.60 ± 8.24	7.52 ± 8.19
Time	{新建，呈现，第一，当……时，然后}	4.30 ± 5.20	4.05 ± 4.91	4.22 ± 4.97	4.23 ± 5.35
个人关注					
Work	{表现，学习，分析，论文，应用}	1.23 ± 1.97	1.14 ± 1.85	1.19 ± 1.85	1.27 ± 2.15
休闲	{小说，表达，频道，视频，游戏}	1.37 ± 2.03	1.34 ± 1.99	1.32 ± 1.92	1.31 ± 1.97
Home	{地址，家人，家庭，街区，邻近的}	0.45 ± 1.06	0.42 ± 1.03	0.47 ± 1.09	0.43 ± 1.04
金钱	{调查，成本，调查过，免费，经济}	0.94 ± 1.51	0.88 ± 1.48	0.94 ± 1.49	0.89 ± 1.50
宗教	{信念, 道德, 牺牲, 猴子, 不可知论者}	0.03 ± 0.19	0.03 ± 0.21	0.03 ± 0.19	0.03 ± 0.19
死亡	{死亡，死亡，死亡，死亡，死亡}	0.04 ± 0.22	0.04 ± 0.22	0.04 ± 0.22	0.04 ± 0.24
非正式语言	{好吧，哦，那个，嗯}	0.49 ± 0.89	0.45 ± 0.86	0.48 ± 0.90	0.48 ± 0.86
脏话	{迟缓，笨蛋，螺丝钉，螺丝钉，迟钝}	0.08 ± 0.34	0.07 ± 0.30	0.08 ± 0.35	0.08 ± 0.33
网络俚语	{o, da, em, k, mm}	0.12 ± 0.41	0.11 ± 0.40	0.11 ± 0.41	0.13 ± 0.43
赞同	{好的，确实，同意，绝对，酷}	0.15 ± 0.42	0.13 ± 0.42	0.15 ± 0.42	0.14 ± 0.39
非流利现象	{嗯，好吧，嗯，嗯，呃，啊}	0.10 ± 0.36	0.09 ± 0.32	0.11 ± 0.37	0.10 ± 0.34
填充词	{rrani, rranr}	0.02 ± 0.14	0.02 ± 0.14	0.02 ± 0.14	0.02 ± 0.14

表7：全部多样化样本的LIWC分析结果（**得分（全部多样）**）摘自表6，以及多样化集合中的三类子集（**C1≠GT**、**C2≠GT**和**C3≠GT**）。