人工智能生成语言与人类语言对比研究 ——以ChatGPT为例

朱君辉[1]，王梦焰[1]，杨尔弘[1]*，聂锦燃[1]，王誉杰[2]，岳岩[1]，杨麟儿[1]
北京语言大学[1]
北京交通大学[2]
nysyzxzjh@163.com

摘要

基于自然语言生成技术的聊天机器人ChatGPT能够快速生成回答，但目前尚未对机器作答所使用的语言与人类真实语言在哪些方面存在差异进行充分研究。本研究提取并计算159个语言特征在人类和ChatGPT对中文开放域问题作答文本中的分布，使用随机森林、逻辑回归和支持向量机（SVM）三种机器学习算法训练人工智能探测器，并评估模型性能。实验结果表明，随机森林和SVM均能达到较高的分类准确率。通过对比分析，研究揭示了两种文本在描述性特征、字词常用度、字词多样性、句法复杂性、语篇凝聚力五个维度上语言表现的优势和不足。结果显示，两种文本之间的差异主要集中在描述性特征、字词常用度、字词多样性三个维度。

关键词： ChatGPT ； 人类语言 ； 语言特征 ； 对比 ； 机器学习

A Comparative Study of Language between Artificial Intelligence and Human: A Case Study of ChatGPT

Junhui Zhu[1], Mengyan Wang[1], Erhong Yang[1] , Jinran Nie[1], Yujie Wang[2], Yan Yue[1], Liner Yang[1]
Beijing Language and Culture University[1]
Beijing Jiaotong University[2]
nysyzxzjh@163.com

Abstract

This paper aims to explore the differences between the language used in human-generated responses and responses generated by ChatGPT, a chatbot based on natural language generation technology. The study extracts and computes the distribution of 159 language features in real human text and ChatGPT-generated text. To evaluate the performance of these features, the study employs three machine learning algorithms: Random Forest, Logistic Regression, and Support Vector Machine (SVM).The experimental results demonstrate that both Random Forest and SVM can achieve high classification accuracy. The result reveals that the two texts differ significantly in three dimensions: descriptive features, word commonness, and word diversity.

Keywords: ChatGPT , Human language , Linguistic features , Comparison , Machine learning

* 通讯作者

第二十二届中国计算语言学大会论文集，第523页-第534页，哈尔滨，中国，2023年8月3日至5日。

1 引言

近年来，随着大数据的支持和计算能力的不断增强，人工智能（AI）在自然语言生成领域取得了长足的进展，特别是在机器翻译、对话生成和文章摘要等任务中，机器生成的语言已经达到了一定的准确性和自然度，并且具备了自己的语言风格。其中，基于神经网络的自然语言生成模型——如GPT系列（Generative Pre-trained Transformer）已成为当今最流行的自然语言处理技术之一。2022年11月30日，OpenAI发布了ChatGPT (Ouyang et al., 2022)，该模型以一问一答的对话形式设计，在理解用户查询和生成类人文本方面表现出色。在中文上，它也能够生成流畅、符合语法的回答，适用于来自各个领域不同类型的问题，引起了极大的关注。

虽然机器生成的语言在语法与逻辑性方面越来越接近于真实语言，但与人类真实语言相比，机器生成的文本在词汇、句法结构、衔接关系等具体语言特征的使用方面仍存在着一些明显的差异。分析这些语言特征的差异对于提高语言模型生成自然语言的准确性和真实性，以及认识人类智能与人工智能的区别至关重要。已有研究发现，机器生成的文本中高级的句法、语义特征占比较低(Pu et al., 2022)，缺乏情感和人情味，难以表达真实人类的情感和感受(Ma et al., 2023)等。然而，目前仍未出现从语言特征的角度深入挖掘二者在语言使用上的差异研究，另一方面，ChatGPT在中文使用上的表现如何也仍未得到探讨。

语言特征的提取与分析能够有效揭示文本中存在的语言规律，广泛应用于体裁分析和语言习得研究。例如，研究者使用一系列特征来识别口语和书面文本之间的语言差异(Louwerse et al., 2004)，正式和非正式类型(Dempsey et al., 2007)，不同文本的年代和作者差异，以及通过词汇丰富度、词汇密度、句法复杂性和句法相似性等特征的测量，探究语言学习者的词汇和句法知识水平(Crossley and McNamara, 2010; Nasseri and Thompson, 2021)等。目前，对ChatGPT生成语言的研究尤其是对比分析人类语言和ChatGPT语言差异的研究尚不多见。

随着计算机技术和自然语言处理技术的发展，研究者们主要聚焦于使用预训练语言模型探测人工智能生成的文本(Dou et al., 2021; Guo et al., 2023; Mitchell et al., 2023; Mitrović et al., 2023)。然而，采用经典特征工程的研究范式建立机器学习模型，操作简便、易于落地，并且能够直观地解释语言特征在其中的作用，仍具有其独特的价值和作用。

有鉴于此，本文从多维度语言特征的视角，深入探究机器生成文本与人类真实语言之间的差异所在，尝试挖掘影响二者语言风格的关键语言因素。具体来说，对平行问答语料进行自然语言处理，通过文本分析工具提取文本中不同维度的语言特征，基于机器学习方法构建分类模型，并对比分析各维度语言特征在人类与机器回答中的分布，以探索各自的语言风格。

2 研究方法

本研究基于ChatGPT与人类对中文开放域问题给出的回答，借助中文CTAP工具(Cui et al., 2022)与Python编程语言对二者语言特征进行量化，训练分类模型并选出预测力较强的特征，从各个维度研究ChatGPT生成文本与人类语言的差异。

具体过程如下：1）选取ChatGPT与人类在开放域问答中的6586篇语料作为研究样本；2）对语料进行分词、词性标注、短语结构标注等预处理，分别计算机器生成文本与人类真实文本五个维度下159项语言特征值；3）训练机器学习模型作为分类器，进而找出区分机器语言与人类语言的最具预测能力的特征；4）基于样本均值比较所选语言特征值在两种文本的分布，观察对模型贡献度强的语言特征，分析两者在不同维度上语言的表现。

2.1 研究问题

本研究主要讨论以下两个问题：1）使用特征工程结合机器学习的方法是否能够有效地区分人类的回答文本与ChatGPT的生成文本？哪些特征是有效预测变量？

2）ChatGPT生成语言与人类语言在不同维度特征上的表现有何具体差别？分别存在哪些优势与不足？

2.2 语料处理

本研究使用的语料来自于(Guo et al., 2023)发布的人机问答语料，选取开放域（不区分专

基金项目：高级汉语水平学习者多模态口语语料库的构建及应用(ZDI145-24)

业领域）下6586篇分别由人类与ChatGPT作答的平行语料构建人机问答语料库。为了更精确地提取语言特征，调用自然语言处理工具Stanford CoreNLP(Manning et al., 2014)工具依次对语料进行分词、词性、短语结构、依存句法等自动标注。

2.3 语言特征测量指标选取

本研究使用中文CTAP与Python编程语言提取两种文本中的语言特征。中文CTAP是一个全面的文本特征自动分析平台，能够分析文本的表层和深层语言特征，包括字、词、句、篇四个维度下的196个特征，可用于母语或二语等多种类型的文本测量指标提取。本研究选择了涵盖描述性特征、字词常用度、字词多样性、句法复杂性、篇章凝聚力五个方面的160个语言特征作为提取对象。在对文本语言特征进行量化时，我们以单个回答作为测量单位，主要使用总数、均值、方差、比例、比值和型例比（TTR）六种通用计算方法。接下来，分别计算159个指标在6586篇人类和ChatGPT回答文本中的测量值，对二者的平均值进行统计分析，初步探索其中的语言差异。在考察的160个语言特征中，删除测量值过小的“拟声词密度”特征后（$\bar{x} < 0.001$），最终确定159个语言特征作为分析对象。

3 实验

对于研究问题一，本节基于前文提出的159 种语言特征，通过传统机器学习算法构建人工智能探测器。本节将评估三种分类算法的预测能力，同时筛选出贡献度较高的有效预测变量。

3.1 模型构建

本研究属于文本分类任务。文本分类常用的机器学习算法有决策树（Decision Tree，DT）、随机森林（Random Forest，RF）、逻辑回归（Logit Regression，LR）、最近邻（K Nearest Neighbor，KNN）、支持向量机（Support Vector Machine，SVM）等，本文选取研究者使用较多的逻辑回归、SVM、随机森林三种经典文本分类算法构建分类模型。

1）逻辑回归（LR）是一种广泛使用的二分类模型，逻辑回归的目标是学习一个权重向量（或模型参数），使得模型能够最大程度地准确地预测二元输出变量。

2）支持向量机（SVM）能够有效地解决分类问题，特别是高维数据和非线性分类问题。SVM模型的主要思想是寻找一个最优的超平面，将不同类别的数据点分隔开来。其目标为最大化支持向量与分类边界之间的间隔，具有很好的泛化性能和较高的精度。

3）随机森林（RF）分类模型是一种集成算法，通过组合多个CART 决策树作为弱分类器，最终结果通过投票或取均值，具有较高的精确度和泛化性能。

在构建分类模型之前，对于每个特征值，我们进行了标准化处理，以避免不同特征值之间的比较出现偏差。我们将数据集随机分为训练集和测试集，比例为8:2，使用测试集对模型进行评估。通过准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1值（F1-score）四种常用指标来评估模型的分类性能。

3.2 有效预测变量及其预测力

实验结果表明，三种分类器都表现良好，如表1所示。其中，SVM的准确率与精确率最高，分别达到了97.27%与97.04%。随机森林的召回率最高，为98.07%。综合来看，SVM在分类性能最为优异。在F1值评估指标上，随机森林与支持向量机（SVM）均展示出了良好的性能。

分类模型	准确率（%）	精确率（%）	召回率（%）	F1值（%）
逻辑回归	96.36	96.99	95.83	96.41
随机森林	97.19	96.49	98.07	97.27
SVM	97.27	97.04	97.62	97.33

表 1: 三种机器学习模型性能对比

在三种模型中，随机森林和SVM均能达到较高的分类准确率，我们基于随机森林和SVM模型在159个语言特征中筛选贡献度较高的有效预测变量。随机森林算法默认采用基尼系数（Gini index）作为特征重要性的量化指标。具体来说，基尼系数越大，那么该语言特征所包含的信息

量就越大，对文本分类的影响力也越大。线性SVM通过找到一个能够最大化类别间隔的超平面来对数据进行分类，超平面由一个权重向量决定，每个分量对应一个特征的权重，特征权重的绝对值越大，说明特征对分类器的预测能力越重要。在本研究中，我们采用随机森林模型中各个语言特征的基尼系数进行评估，并根据这些指标建立特征影响力的排序序列，选取前31个具有较高影响力的特征（基尼系数>0.10）。接下来，我们从支持向量机（SVM）分类器中提取特征权重，并计算特征权重绝对值的均值作为筛选阈值。经过筛选，共有59个特征满足条件。通过综合运用这两种机器学习模型，我们共确定了77个关键特征，在这些特征中，有12个特征在两种算法的结果中均有所体现。这些结果为我们在分析和解释语言特征的差异上提供了重要的参考依据。

4 不同维度语言特征差异分析

参考贡献较大的77项特征，我们对描述性特征、字词常用度、字词多样性、句法复杂性、篇章凝聚力五个维度159项特征依次展开分析。

4.1 描述性特征

描述性特征指对文本中字、词、句、篇四个层面的基本描述性统计，用于表征文本中各个层面语言单位的数量、长度等，属于文本的视觉属性。我们将37项语言特征作为描述性特征维度的测量指标，具体如表2所示。

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
笔画	**少笔画字数**	73.033	134.343	词汇	三音词占比	0.044	0.042
	少笔画字比例(1-8)	0.668	0.651		**四音节及以上词占比**	0.027	0.047
	中笔画字数(9-16)	35.039	61.347		四音节及以上词数	1.653	4.636
	中笔画字比例(9-16)	0.325	0.297		平均词长	1.704	1.861
	高笔画字数(16以上)	0.364	0.441	句子	句子数	4.207	7.343
	高笔画字比例(16以上)	0.003	0.002		平均句长(以字为单位)	40.893	42.396
	字例平均笔画数	7.330	7.102		平均句长(以词为单位)	25.067	21.823
	字形平均笔画数	7.435	7.168		句长标准差(基于词例)	9.248	6.729
部件	**字形平均部件数**	1.754	1.692		**句长标准差(基于词形)**	6.838	5.042
	字例平均部件数	1.744	1.684		句长标准差(基于字例)	15.150	12.842
字数	**字例数**	134.617	262.117		句长标准差(基于字形)	10.034	7.654
	字形数	79.719	104.134		最长句字数	63.129	61.623
	词例数	84.040	146.615		最长句词数	38.447	31.858
	词形数	56.705	73.460	段落	**篇章段落数**	1.442	3.681
词汇	单音节词数	35.666	47.878		**最长段落长度(基于词)**	127.121	113.591
	单音节词占比	0.483	0.379		**最长段落长度(基于字)**	80.371	63.917
	双音节词数*	32.360	68.705		**平均段落长度(基于字)**	123.907	92.747
	双音节词占比	0.445	0.532		**平均段落长度(基于词)**	78.308	51.882
	三音节词数	3.005	5.226				

注：带*表示在两种算法中均贡献度突出，加粗表示在一种算法中贡献度突出，以下各表同理。

表 2: 描述性特征维度人类与ChatGPT文本特征均值对比

在37项描述性特征中，在两种算法中均贡献度突出的是双音节词数，在任一种算法中贡献度突出的有少笔画字数、中笔画字比例(9-16)、字形平均部件数等14项，占所有重要特征总数的19.5%。观察表2可知，ChatGPT语言特征指标高于人类语言的有17个，集中在笔画、字数、词数三个层面；低于人类语言的有20个，集中在部件、句长、段落三个层面。

汉字包含的笔画数与部件数一定程度上体现了汉字在书写方面的复杂程度。理论上说，在文本材料中，笔画数或部件数较多的汉字占比越大，汉字的字形复杂度越高，文本阅读难度越大(张倩倩, 2022)。去除文本长度因素的影响，在人类回答文本与ChatGPT生成文本中均为少笔画字占比最大，中笔画字次之，高笔画字占极低。即无论对于人类还是机器，少笔画字及中笔画字即可基本满足回答大多数开放域问题所使用语言的需求。不同的是，在人类回答文本

特征类别	汉语特征(频数)	人类	GPT	特征类别	汉语特征(频数)	人类	GPT
字形	**平均对数字形1**	6.256	6.123	实词词形	**平均对数实词词形1***	5.209	5.156
	平均对数字形2	4.021	2.923		**平均对数实词词形2**	3.441	3.239
	平均对数字形3	2.646	2.442		**平均对数实词词形3**	3.965	3.941
字例	**平均对数字例1**	6.424	6.202	实词词例	**平均对数实词词例1**	5.291	5.197
	平均对数字例2	2.945	2.862		**平均对数实词词例2**	3.517	3.302
	平均对数字例3	2.747	2.560		平均对数实词词例3	4.044	4.013
词形	**平均对数词形1***	5.250	5.288	虚词词形	**平均对数虚词词形1**	6.295	6.283
	平均对数词形2	2.579	2.273		**平均对数虚词词形2**	4.413	4.279
	平均对数词形3	2.923	2.722		平均对数虚词词形3	4.244	4.135
词例	平均对数词例1	5.528	5.475	虚词词例	**平均对数虚词词例1**	6.643	6.731
	平均对数词例2	2.753	2.590		**平均对数虚词词例2**	4.754	4.758
	平均对数词例3	2.122	2.970		平均对数虚词词例3	4.474	4.469

注：1 Gigaword字词频表　2 现代汉语语料库　3 汉语二语教材语料库

表 3: 字词常用度维度人类与ChatGPT文本特征均值对比

中，少、中、高笔画字所占比例与字形平均笔画数、字例平均笔画数均高于ChatGPT。数据表明人类拥有较广的中高笔画汉字储备量，在一定程度上更具备使用较多笔画数汉字的能力。

文本中各个语言单位的长度和数量能够用来衡量文本难度与文本质量(熊兵, 2016; 李绍山, 2000; 邢诗吟, 2022)。一般来说，文本中词数、句子数、段落数越多，平均词长、平均句长越长，文本的质量越高，难度越高。本研究测量了词长、句长、段落长度、语篇长度以及字例数、词例数、句子数及段落数等指标。在词汇层面，相较于人类的用词，ChatGPT生成语言倾向于使用大词（词长较长的词），表现在平均词长、双音节词、四音节词、四音节及以上词占比相对较高。与之照应，在句长上，ChatGPT语言以字为单位的平均句长大于人类语言，以词为单位的平均句长小于人类语言，即ChatGPT生成的句子字数更多，词数却更少。值得关注的是，在人类所给出的回答中，最长句子所包含的字数和词汇量均超过了ChatGPT，这表明人类具有生成更为复杂和详尽句子的能力。在段落层面，ChatGPT生成文本的平均段落长度和最长段落长度均低于人类文本。在语篇层面，语篇长度可由字例数、词例数、句子数及段落数体现，对于这四项指标，ChatGPT生成语言的测量值均高于人类语言。反映出人类的回答更加简短，倾向于将较长的回答浓缩在较少的自然段中；而ChatGPT则擅长生成较长的答案，并进行分段阐述。从某种角度来说，ChatGPT生成的文本在难度、质量上高于人类的回答。

此外，句长变化度通过计算文本中所有句子的句长标准差得到，用来评估文本中句子长短变化的情况。句长变化度较高的文本中出现的句子大多长短不一，反之，则说明文本中句子的长度大致相同(张倩倩, 2022)。无论是基于字还是基于词测量的句长标准差特征指标，人类回答的值均高于ChatGPT回答，且最长句子字数和最长句子词数大于ChatGPT回答。可知，相比ChatGPT语言，人类回答中的句子长度之间差异更大，长短句的使用更加灵活多变。

4.2 字词常用度

字词常用度由字词使用的频率信息测量，字词在书面文本中出现的频次反映了读者的实际接触频率和熟悉程度。本文引入《Gigaword字/词频表》、《汉语二语教材语料库字/词频表》和《现代汉语语料库字/词频表》三种字/词频表，通过汉字和词汇（实词/虚词）的平均对数频数来测量字词的常用度，共计24项测量指标，如表3所示。其中，在两种算法中均贡献度突出的是平均对数词形频数（Gigaword词频表）和平均对数实词词形频数（Gigaword词频表），在任一种算法中贡献度突出的有15个，占所有重要特征总数的22.1%。

字频（字形频数、字例频数）和词频（词形频数、词例频数）可以反映汉字熟悉度与词汇熟悉度，通常被作为衡量文本难度的重要指标。已有研究表明，词频取对数后的数值与词汇识别时间之间呈现线性负相关，频率效应显著(Balota and Chumbley, 1984; Haberlandt and Graesser, 1985; 蔡建永, 2020)。具体而言，如果一些词语在已有词频表中频次较高，即在大多数文本中出现和运用的次数较为频繁，表示这类词经常被使用，读者在阅读时遇到该类词便更加迅速地从记忆中提取出来并唤醒。反之，如果一些词语在文中显示和运用的次数极少，那么该词提取和唤醒需要的时长就会更多。汉字同理。数据显示，在字词常用度的24个特征中，人

类语言的各指标值均大于ChatGPT生成语言，这说明人类在回答中使用频次高的汉字与词汇占比均多于ChatGPT，文本的阅读难度相对较低。

4.3 字词多样性

字词多样性反映的是文本中汉字与词汇的使用是否丰富多样(蔡建永, 2020)，通过文本中字词被重复使用的程度进行衡量。表4呈现了衡量字词多样性的汉字多样性、词汇多样性、实词丰富度、词汇密度四个维度50项测量指标。

其中，在两种算法中均贡献度突出的是字型例比、出现一次的字占比、词形例比、仅出现一次的词占比、实词丰富度、连词密度，在任一种算法中贡献度突出的有字Log型例比、词Log形例比、词Uber形例比等21项，占所有重要特征总数的36.4%。

我们通过计算型例比（TTR）与仅出现一次的字/词占比来测量汉字多样性与词汇多样性。型例比值越高，仅出现一次的字/词占比越高，说明字词使用越丰富。为了缓解文本长度的影响，我们还采用了研究者改良后的Log TTR、Root TTR、Uber TTR、Corrected TTR等计算方法。从上表中可以观察到，在汉字多样性和词汇多样性的两个层面上，人类语言的字、词型例比以及出现一次的字、词占比都高于ChatGPT语言。数据表明，人类回答中所使用的字词种类丰富，词汇使用具有灵活性和创造性；ChatGPT生成的文本篇幅更长，但词汇选择范围较窄，重复性强，语言使用上趋于保守。

同时，字词多样性还可以由实词丰富度体现。实词丰富度反映的是名词、动词、形容词、副词四种实词类型在所有实词中的多样性。实词用于传递信息和表达意义，文本中的实词越多，概念密度也越大，包含的信息量越大(Johansson, 2008)。观察上表可知，人类语言的实词丰富度指标几乎均大于ChatGPT。在同样篇幅的文本中，人类提供的信息量更大。

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
汉字多样性	**字型例比***	0.648	0.470	词汇密度	“被”字结构密度	0.001	0.002
	字Log型例比	0.905	0.850		**标点密度**	0.135	0.136
	字Root形例比	6.753	6.739		**代词密度**	0.052	0.069
	字Uber形例比	55.612	40.810		**动词密度**	0.207	0.186
	字Corrected形例比	4.775	4.765		方位词密度	0.012	0.013
	仅出现一次的字数	53.656	57.133		**副词密度**	0.111	0.081
	出现一次的字占比*	0.520	0.308		基数词密度	0.034	0.025
词汇多样性	**词形例比***	0.725	0.543		**介词密度**	0.029	0.043
	词Log形例比	0.923	0.872		句均词性数量	5.507	3.151
	词Root形例比	6.022	6.095		量词密度	0.027	0.019
	词Uber形例比	58.070	40.659		**名词密度**	0.267	0.290
	词Corrected形例比	4.258	4.310		**能愿动词密度**	0.024	0.031
	仅出现一次的词数	42.977	46.554		**人称代词密度**	0.031	0.042
	仅出现一次的词占比*	0.588	0.365		**实词密度**	0.745	0.713
实词丰富度	**实词丰富度***	0.822	0.647		**数词密度**	0.036	0.025
	名词丰富度	0.309	0.269		叹词密度	0.001	0.000
	Squared动词丰富度1	10.949	11.346		形容词密度	0.023	0.016
	Corrected动词丰富度1	2.215	2.313		形式动词密度	0.001	0.002
	副词丰富度	0.127	0.089		**虚词密度**	0.118	0.150
	形容词丰富度	0.030	0.019		序数词密度	0.002	0.001
	修饰语丰富度	0.157	0.108		**疑问代词密度**	0.008	0.005
	动词丰富度	0.236	0.182		**语气词密度**	0.016	0.003
	动词丰富度1	0.823	0.672		**指示代词密度**	0.010	0.009
词汇密度	**连词密度***	0.013	0.036		助词密度	0.060	0.068
	“把”字结构密度	0.002	0.001		专有名词密度	0.028	0.020

表 4: 字词多样性维度人类与ChatGPT文本特征均值对比

从另一角度来讲，实词丰富性在某种程度上也反映了文本理解的难度。张必隐(1992)在其研究中指出，在阅读过程中，实词往往能够协助读者更快速地理解文本的含义。黄伯荣、廖序

东(黄伯荣and 廖序东, 2017)也在总结以往的实验研究后发现，文章中实词和虚词的数量及其比例对文章的易读性有一定的影响。因此，人类在其回答中使用更多的实词，这一事实验证了4.2一节所得出的结论，即人类文本具有更高的可读性。

词汇密度反映不同词汇在文本中的使用倾向，经常被用在语言类型、语言风格与文体类研究中。例如，在语言类型研究中，研究者发现英语常用介词来表达动态或动作的意思，而汉语使用者倾向于用动词表达，叙事性较强，善用副词等；在语言风格研究中，杨彬(2023)指出副词的使用能够灵活巧妙地调节叙事的节奏，从而建构出形态纷繁的文本。本研究参照现代汉语(黄伯荣and 廖序东, 2017)的词性体系，计算人类回答文本与ChatGPT生成本文中所有词类的密度与句均词性数量。人类回答中，大部分实词的密度与句均词性密度均大于ChatGPT语言，如形容词密度、动词密度、副词密度等；虚词中对叹词与语气词的使用倾向明显，而机器语言中几乎未出现叹词。这一事实表明人类语言更加生动，善于灵活处理变换词性，情感表达丰富，描写能力和表达能力远高于ChatGPT语言。相较于人类语言，ChatGPT生成语言更倾向于使用连词、介词、名词、能愿动词、人称代词、形式动词、助词等，虚词成分较多。这种差异可能源自于汉语训练语料数量不足，也在一定程度上表明，机器语言的连词和人称代词显化现象比人工语言更加突出，且使用助词频率较高，语法标记明显(蒋跃and 董贺, 2015)。

此外，虚词的分布也在一定程度上代表着文本的语体色彩。虚词的使用是写作者无意识的产物，能够反映不同作家的风格风貌(Stamatatos, 2009)。两者差异最大的是连词密度，在不同文体不同领域中，连词的使用及风格色彩也有所偏向。以最具代表性的“和、与、跟、同”为例，“和、与”具有书面语色彩，“跟”具有北方口语色彩，“同”具有南方口语色彩。我们分别计算了人类回答和ChatGPT回答中这四个连词使用频率的平均值，如表5所示，ChatGPT的回答更倾向于使用“和、与”作为句子成分的连接词，相比之下，人类回答中使用“跟、同”的频率更高。数据表明，ChatGPT生成的语言更倾向于书面语的表达方式。进一步支持这一结论的证据可以在第4.1节中各音节词数和占比中找到，具体来说，ChatGPT生成的语言中双音节词的数量和比例最高，这符合现代汉语双音化用词的习惯，而人类语言中单音节词和双音节词的比例则更为接近，更富有口语特色。

	“和”	“与”	“同”	“跟”
人类	4.13	0.92	0.73	0.18
ChatGPT	11.76	1.57	0.07	0.03

表 5: 人类与ChatGPT在“和、与、同、跟”上使用的差异

4.4 句法复杂性

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
并列短语	**并列短语数**	0.813	4.600	名词短语	名词短语数	28.895	54.184
	单句平均并列短语数	0.095	0.331		单句平均名词短语数	3.364	4.020
	句均并列短语数*	0.251	0.729		句均名词短语数	8.702	8.265
形容词短语	形容词修饰语数	1.838	4.114		名词短语平均长度/字token	4.054	4.816
短语结构	句均单句数	3.862	2.511	动词短语	动词短语数	28.727	44.556
	句法树高大于14的句子数量	0.484	0.759		单句平均动词短语数	3.222	2.958
	最大句法树高	13.999	14.919		句均动词短语数(en,de)	8.687	6.566
	平均句法树高	10.899	11.419		**动词短语平均长度/字token**	9.353	14.136
	句法树高大于14的句子占比	0.160	0.129	介词短语	介词短语数	2.197	5.422
	主要动词前平均词数	3.440	4.172		单句平均介词短语数	0.222	0.365
	主要动词前最大词数	7.814	11.432		句均介词短语数	0.663	0.900
依存句法	平均句子依存距离	3.900	3.659		介词短语平均长度/字token	6.274	9.886
	最大句子依存距离	29.452	23.991				

表 6: 句法复杂性维度人类与ChatGPT文本特征均值对比

	问题一：华尔街的课有效果吗？能提高英语水平吗？	**问题二**：北京有像上海七浦路一样的批发市场吗？
ChatGPT	...华尔街英语使用了多种教学方法，包括讲课、角色扮演、小组讨论和个人辅导等。...	...这些市场都提供各种服装产品，包括男装、女装、童装等。...
人类回答	...华尔街的话，其实价格蛮贵的，网上的叫骂声也蛮高的，但是我觉得培训方面还是非常不错的。...	...在南三环木犀园到南四环大红门一带，有很多服装批发大楼，其中的天雅是专门的品牌批发，购物环境不错，...

表 7: 并列短语在ChatGPT与人类回答文本中的使用示例

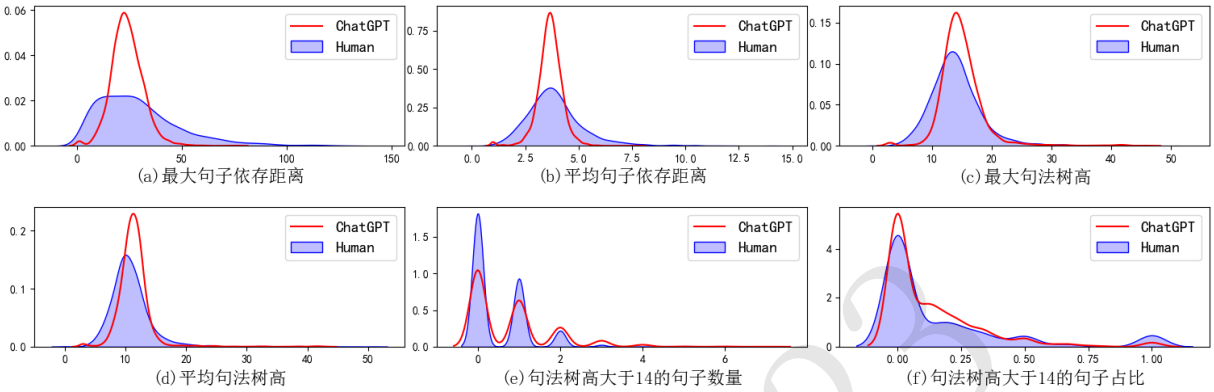

图1: 人类语言与ChatGPT句法树高与依存距离相关特征差异对比

句法复杂性包括名词短语、动词短语、介词短语等七个类别25项测量指标，如表6所示。其中，在两种算法中均贡献度突出的是句均并列短语数，在任一种算法中贡献度突出的有动词短语平均长度(以字token为单位）、并列短语数、单句平均并列短语数、句法树高大于14的句子占比4项，占所有重要特征总数的6.5%。句法复杂度主要通过句子中不同类型短语的数量和短语长度体现。人类回答中动词短语的使用频率较高，而对于名词短语、介词短语以及并列短语，ChatGPT回答的使用率较高。动词性成分具有较为突出的叙事性特征，人类回答中动词短语的大量使用意味着较强的交互性。相对而言，ChatGPT回答倾向于运用修饰性和概念性较强的表达方式。在各种类型的短语中，ChatGPT生成的文本所包含的短语平均长度普遍超过人类水平。

并列短语相关的3个特征均是分类模型中贡献度高的重要特征，经观察发现，ChatGPT在这三种指标上都远高于人类。通过观察数据我们发现，ChatGPT回答中经常使用多个并列成分，这些并列成分处于同一语义场之中，表7提供了两个问答示例，其中“教学方法、服装产品”是上位词，“包括”后是他们各自对应的下位词，诸如此类同一义场中下位词的并列使用，使得要表达的意思更加全面、具体，起到强调的作用(陈绍新, 2017)。

此外，主要动词前的平均词数、句法树高和依存距离（dependency distance）也是衡量句法复杂度的重要指标。主要动词前的平均词数越多，句子的句法树越高，依存距离越长，说明句子的句法关系越丰富，句法复杂度越高(McNamara et al., 2014; 吴思远, 2020)。数据显示，ChatGPT生成文本中主要动词前的平均词数与最大词数都多于人类。随后，我们统计了句法树高大于14的句子数量、占比以及平均句法树高和最大句法树高。数据表明，在ChatGPT生成的文本中，平均每篇文本中有75.9%的句子的句法树高度超过14，相较之下，人类编写的文本中仅有48.4%的句子表现出相同特征。这意味着ChatGPT生成的文本在句法结构复杂性方面往往高于人类撰写的文本。另一方面，我们采用平均句子依存距离、最大句子依存距离2 个常用的特征衡量依存距离。依存距离指句中两个有句法关系的词之间的线性距离，即支配词和被支配词之间的线性距离(Hudson, 1995)。然而，对于这两项特征，人类回答测量值的平均值均高于机器文本，与已有结论相悖。为了进一步探究原因，我们就句法树高与依存距离对3000篇文本绘制了核密度图，如图1所示。

观察图1可知，对于句法树各个相关指标的测量，人类与机器呈现出的取值范围总体相近；

而在依存距离上，人类与机器生成语言中测量值的范围相差较大。也就是说，虽然人类回答的平均句子依存距离与最大句子依存距离均高于机器文本，但人类的回答中最大句子依存距离介于0~100之间，在此区间内分布较为均匀，密集的区间为10~30；而ChatGPT生成的文本大多聚集在25~30之间，且密度高峰远超人类。这一事实揭示出，出于“省力”的考虑，在语言运用中，人类会尽量避免使用可能导致认知成本增加的长距离依存关系(陆前and 刘海涛, 2016)，倾向于使用简单的句子结构和句法成分。但面对难以简短回复的问题，也具备使用句法结构较为复杂的长句的语言能力。

4.5 语篇凝聚力

语篇凝聚力由语篇的衔接程度体现，包括指称、重复、衔接三个类别23项测量指标，如表8所示。其中，指代用代词比例来衡量，重复通过相邻句和全文中实词、名词、动词的重复性来衡量，代词和重复词语的使用可以从语义上让上下文的联系更加紧密。衔接则用各类连词比例来衡量，使用象征不同逻辑关系的关联词是衔接上下文的有效方法。

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
指称	**第一人称代词比例**	0.012	0.011	重复	**相邻句中动词的重复性**	0.241	0.437
	第三人称代词比例	0.005	0.007		全文中动词的重复性	0.181	0.267
	第二人称代词比例	0.010	0.021	衔接	转折连词比例	0.008	0.007
	人称代词比例	0.031	0.042		因果连词比例	0.011	0.007
	疑问代词比例	0.008	0.005		**选择连词比例**	0.003	0.015
	指示代词比例	0.010	0.009		条件连词比例	0.008	0.002
重复	**全文中词语的重复性***	0.380	0.519		顺承连词比例	0.013	0.003
	全文中实词的重复性*	0.335	0.491		目的连词比例	0.003	0.003
	全文中名词的重复性	0.192	0.368		假设连词比例	0.018	0.009
	相邻句中词语的重复性	0.545	0.875		递进连词比例	0.005	0.006
	相邻句中实词的重复性	0.481	0.831		并列连词比例	0.014	0.012
	相邻句中名词的重复性	0.263	0.631				

表 8: 语篇凝聚力维度人类与ChatGPT文本特征均值对比

语篇凝聚力维度中，在两种算法中均贡献度突出的是全文中词语的重复性、全文中实词的重复性，在任一种算法中贡献度突出的有人称代词比例、选择连词比例、全文中名词的重复性等10项特征，占所有重要特征总数的15.6%。

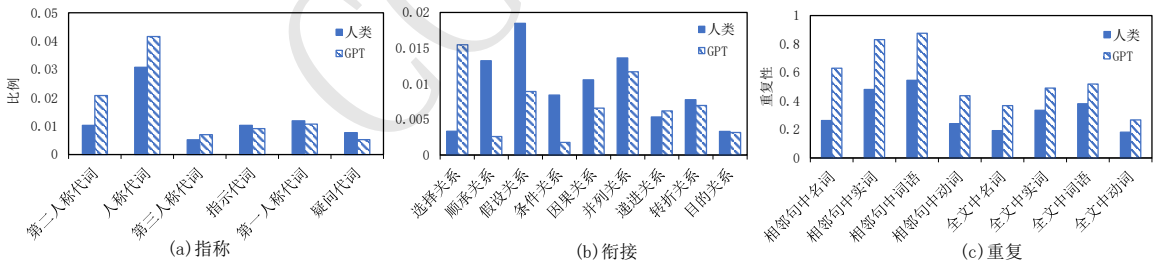

图2: 语篇凝聚力维度特征差异图

1.指称。指称关系指篇章中一个成分与另一成分之间所具有的相互解释的关系(洪秋月and 熊智伟, 2023)，指称衔接多用代词来体现，包括人称代词、指示代词、疑问代词。彭宣维(2005)在探讨代词的语篇语法属性时提出，代词在文章中的出现主要是代替别的成分发挥相应功能，从而使语句及整篇文章有很好的衔接关系。

如下图2（a）所示，在人称代词、指示代词、疑问代词这三项指标中，ChatGPT语言使用人称代词的比例要多于人类回答，且使用第二第三人称代词的比例相对较多，ChatGPT语言用多尊称“您”，而人类语言倾向于使用第一人称代词和指示代词、疑问代词。已有研究表示，第一人称代词、指示代词这两种形式在非正式语体中占主体部分，第一人称代词的使用显示了

研究结果的主观性，多是引导读者赞同所述观点和研究结果，传递出作者构建主体身份的特性(Seidel, 1975)。

指示代词和疑问代词的使用，显示出作者根据语境回指上文出现的事物，既有利于文本衔接上下连贯，又可强调观点表达(贾宇丹, 2022)。由此可见，ChatGPT语言多是以较为客观的态度进行分析并给出建议，遵循会话的礼貌原则，较少发表主观性强的意见，而人类回答拥有话语权，善于表达自己的观点和看法。

2.衔接。衔接以语篇序列为前提，在建立句子之间的衔接与联系方面占有非常重要的作用(Cain and Nash, 2011)，韩礼德和哈桑(Halliday and Hasan, 1976)提出，关联词不是直接通过它们自己来衔接，而是通过它们的特殊意义来间接衔接。本研究通过关联词的分布考察了选择关系、顺承关系、假设关系等9种衔接关系，以此分析ChatGPT语言与人类语言的文本衔接特征上的差异。观察数据可以发现无论是人类回答还是ChatGPT回答，所运用的关联词类别都比较全面，不存在某种衔接关系缺失的情况。但整体来看，ChatGPT语言使用上述9种衔接关系的比例低于人类语言，表明人类在回答问题时所运用的语言具有衔接显化的特点(邢诗吟, 2022)。

图2（b）对9种衔接关系的使用比例进行了差异排序，可以看到ChatGPT和人类回答中差距较大的是假设关系、顺承关系、条件关系、选择关系。ChatGPT回答中存在选择关系的比例高于人类语言，顺承关系、条件关系、选择关系的比例低于人类语言。ChatGPT最常使用陈述式关联词“或...或...”、“或者...或者...”，而人类回答中使用最多的是疑问式关联词“还是”。在人类语言中，使用比例最高的是假设关系，多是表示和结果一致的假设，如使用表示一致关系的关联词“就”。而ChatGPT回答在表达假设关系时使用最多的是表示相背关系（假设和结果不一致）的“...，也...”、“...，还...”。(黄伯荣and 廖序东, 2017)。

3.重复。重复指在同一语篇中反复出现具有相同含义和形式的词，对实现前后文的连贯有显著的作用。如图2（c）所示，ChatGPT语言中相邻句和全文中实词、名词、动词的重复性都高于人类语言，说明ChatGPT的篇章衔接紧密，文本的表达紧紧围绕同一主题，而人类文本的篇章重复性较低，词干、论元重叠度低(何清强et al., 2019)，行文发散。

5 总结与讨论

本研究旨在考察人类与ChatGPT回答文本中语言特征的差异，以及基于特征结合机器学习方法得到的ChatGPT探测器预测的能力。结果表明，第一，在描述性特征、字词常用度、字词多样性、句法复杂性、语篇凝聚力五个维度中，对模型分类贡献度较高的特征集中在描述性统计、字词常用度、字词多样性三个维度。第二，SVM与随机森林都表现出较好的性能，最优模型达到了97.27%的准确率与97.33%的F1值。

对于五个维度下两种回答文本的语言差异，本文得出以下结论：

1.ChatGPT生成语言倾向于使用大词，往往分段进行阐述生成长文本，人类的回答更加简短，自然段少。在一定程度上，ChatGPT生成的文本难度与质量高于人类的回答。人类具有生成更为复杂和详尽句子的能力与使用较多的笔画数汉字的能力，长短句的使用更加灵活多变。

2.人类的用词偏好有助于丰富语言表达（词汇多样性高）并降低文本理解难度（高频词和实词使用频率高），富有口语色彩。ChatGPT生成的语言中语法标记明显，倾向于书面语的表达方式。在同样篇幅的文本中，人类提供的信息量更大。整体来看，ChatGPT所体现出的语言特征更具英文偏好，比如和英文一样，ChatGPT倾向于使用介词、助词等修饰性较强的成分，这可能与训练语料大多是英语有关。

3.ChatGPT倾向于运用修饰性和概念性较强的表达方式，在句法结构复杂性方面往往高于人类撰写的文本。人类回答具有较强的交互性，倾向于使用简单的句子结构和句法成分。但面对难以简短回复的问题，也具备使用句法结构较为复杂的长句的语言能力。

4.ChatGPT在指称、重复上优于要优于人类文本，词语的重复性较多，语义重叠度高，生成的回答围绕同一主题展开。人类思维活跃，容易给出发散式的回答。

根据本文的研究结论可推断出，ChatGPT在中文使用上的表现与人类具有较大差异。为了使人工智能生成语言更加真实，高质量的中文数据集建设与大语言模型研究迫在眉睫。我们的研究详细分析了ChatGPT生成语言与人类语言在多个维度上的差异，但也存在一些局限。首先，本研究中仅选择了ChatGPT生成语言和人类语言在开放域的问答语料，样本量相对较小，未来的研究中可以使用更多包含不同语域与不同语体的数据集。其次，本研究中使

用GPT-3.5作为底层模型的ChatGPT，若使用更先进的模型（如GPT-4），这些语言特征的表现可能会有所不同。

参考文献

D. A. Balota and J. I. Chumbley. 1984. Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology Human Perception & Performance*, 10(3):340–357.

K. Cain and H. M. Nash. 2011. The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology*, 103(2):429.

Scott A Crossley and Danielle S McNamara. 2010. Interlanguage talk: What can breadth of knowledge features tell us about input and output differences? In *23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*, pages 229–234.

Yue Cui, Junhui Zhu, Liner Yang, Xuezhi Fang, Xiaobin Chen, Yujie Wang, and Erhong Yang. 2022. Ctap for chinese: a linguistic complexity feature automatic calculation platform. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5525–5538.

Kyle B Dempsey, Philip M McCarthy, and Danielle S McNamara. 2007. Using phrasal verbs as an index to distinguish text genres. In *FLAIRS Conference*, pages 217–222.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Annual Meeting of the Association for Computational Linguistics*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

K. Haberlandt and A.C. Graesser. 1985. Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114:357–374.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Routledge, London.

Richard A. Hudson. 1995. English word grammar.

Victoria Johansson. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79.

Max M Louwerse, Philip M McCarthy, Danielle S McNamara, and Arthur C Graesser. 2004. Variation in language and cohesion across written and spoken registers. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human – differentiation analysis of scientific content generation.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, Cambridge.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Maryam Nasseri and Paul Thompson. 2021. Lexical density and diversity in dissertation abstracts: Revisiting english l1 vs. l2 text differences. *Assessing Writing*, 47:100511.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jiashu Pu, Ziyi Huang, Yadong Xi, Guandan Chen, Weijie Chen, and Rongsheng Zhang. 2022. Unraveling the mystery of artifacts in machine generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6889–6898.

G. Seidel. 1975. Ambiguity in political discourse. In M. Bloch, editor, *Political Language and Oratory in Traditional Society*, pages 205–228. Academic Press, London.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

何清强, 王文斌, and 吕煜芳. 2019. 汉语叙述体篇内句的特点及其二语习得研究——基于汉英篇章结构的对比分析. 语言教学与研究, (06):1–11.

吴思远. 2020. 基于多层面语言特征的汉语文本可读性自动评估研究. 硕士学位论文, 北京语言大学.

张倩倩. 2022. 基于小学语文教材的文本易读性公式研究. Ph.D. thesis, 江南大学.

张必隐. 1992. 阅读心理学. 北京师范大学出版社, 北京.

彭宣维. 2005. 代词的语篇语法属性、范围及其语义功能分类. 语言教学与研究, (01):56–65.

李绍山. 2000. 易读性研究概述. 解放军外国语学院学报, 2000(04):1–5.

杨彬. 2023. 篇章动态视角下副词性成分的叙事价值分析. 当代修辞学, 2023(01):42–50.

洪秋月and 熊智伟. 2023. 语言经济原则下热搜词条的语篇衔接研究. 今古文创, (05):129–132.

熊兵. 2016. 基于语料库的旅游文本英译文词汇特征及翻译研究. 华中师范大学学报*(*人文社会科学版*)*, 55(05):94–103.

蒋跃and 董贺. 2015. 计量特征在人机译文语言风格对比中的应用. 语言教育, 3(03):69–74+81.

蔡建永. 2020. 汉语二语文本可读性研究. Ph.D. thesis, 北京语言大学, 北京.

贾宇丹. 2022. 中国外应专业研究生学术语篇非正式语体特征研究. 名家名作, (21):85–87.

邢诗吟. 2022. 基于语料库的初中英语记叙文写作语言特征研究. Master's thesis, 集美大学.

陆前and 刘海涛. 2016. 依存距离分布有规律吗? 浙江大学学报*(*人文社会科学版*)*, 46(4):63–76.

陈绍新. 2017. 元功能理论视角下的英语商务合同汉译研究. 湖南第一师范学院学报, 17(6):92–97.

黄伯荣and 廖序东. 2017. 现代汉语. 高等教育出版社.