

游历¹ 方怡蓁²

【摘要】目的：在自然语言处理领域，通过机器学习和大语言模型（LLM）方法鉴别“新式英语”及我们普遍认知的“英语”。**方法：**采用 Logistic Regression, Support Vector Machine, Naïve Bayes 等机器学习方法，及零射击提示（Zero Shot Prompting）Llama2 7B 和 Mistral 7B 的 LLM 进行文本分类任务，标注新式英语为 0，英语为 1，LLM 特例为 2。采用的数据集包括 The National University of Singapore SMS Corpus 及经过随机筛选 Webis-Simple-Sentences-17 Corpus 合成的数据。**假设：**Zero Shot Prompting LLM 的效果会比传统机器学习方式更好。**结论：**从准确率的角度看机器学习的方法比 LLM 的结果较好。然而，我们使用的 LLM 仅有 7B，对比而言兴起的 GPT3 推测为 175B，GPT4 则是 1.76T。此外，我们仅仅采用最原始的 Zero Shot，并没有尝试其他更有效的提示方式如 Few Shot, Chain of Thought, Tree of thought，等。因此进一步探究是有必要的。

【关键词】机器学习；大语言模型；Singlish；LLM

【分工】代码和报告部分由二位同学合力完成，由游历同学主写背景、数据预处理和机器学习部分；怡蓁同学主写大语言模型、优化与应用部分。

【代码】所有代码可参考 https://github.com/yizhenpong/Singlish_or_English_ML_LLM

1. 背景

语言是人类交流的核心，塑造着我们的文化、历史和身份。随着新兴计算机技术的出现，特别是机器学习和自然语言处理的不断发展，技术革命很大程度上使高资源语言受益，但许多资源不足的语言在很大程度上未被开发，因此存在代表性不足的显著问题，也引发了一些关于处理这些代表性不足的语言并被大众所充分认识的顾虑。

除了认知方面的考虑之外，语言研究者还在与时间赛跑，即如何从资源匮乏的语言中收集和保存数据。据《卫报》报道，全球有 7000 种语言，到本世纪末将有一半以上灭绝，其中一些是资源匮乏的语言。

让我们首先定义是什么使语言资源不足。多年来，互联网中语言愈发丰富，但这并不意味着所有语言都可以在网上找到。资源匮乏的语言的数字内容非常有限或不存在，这使得研究人员很难通过机器翻译模型使用和运行它。这些资源的范围从在线文本数据到机器翻译工具、语音识别软件等。尽管在网上不太显眼，但资源匮乏的语言很重要，因为它们通过包含独特的文化知识和充当社区的社会联系者，在使我们的全球语言遗产多样化方面发挥着至关重要的作用。因此，迫切需要为这些语言开发资源，并使其在数字上可见和可访问。

这并不一定意味着资源匮乏的语言濒临灭绝，但是资源匮乏的语言可能会在信息和通信技术中遇到权力不平衡的问题，其中缺乏“资源和信息控制的平等”，因为它没有赋予当地社区定义他们的观点、需求和目标的权力。

其次，大型语言模型（LLMs）在当今研究和应用中占据着重要地位，尤其是随着 ChatGPT 的兴起。LLMs 是基于深度学习的技术，在自然语言处理（NLP）任务方面

¹ 21 级外国语学院本科生

² 23 级元培学院交换生、21 级新加坡国立大学（NUS）本科生

表现出色，从文本分类到语义理解，展现了令人瞩目的性能。这类模型经过大规模和多样化的语料库训练，能够捕捉语言的复杂结构和语义信息，从而具备强大的自然语言理解能力。其中，使用 Transformer 架构的 LLMs，如 Llama、BERT、GPT 等，已成为研究和工业界的热门选择。

对于资源匮乏的语言来说，LLMs 有望在低资源语言领域引发革命，能够捕捉低资源语言的复杂性，并为其结构提供有价值的洞见。未来 LLMs 也可能运用到保留各地方言，甚至是非物质文化遗产。

2. 数据介绍

Singlish，学名为新加坡式口语英语（新加坡语和英语的合成词）；正式名称为口语新加坡英语，是一种起源于新加坡的英语克里奥尔语^{3 4 5}。新加坡式英语起源于新加坡许多不同亚洲语言（如马来语、粤语、闽南语、普通话、潮州语和泰米尔语）之间长期语言接触的情况⁶。新加坡式英语一词是指新加坡俚语和英语的混合体，最早记录于 1970 年代初。

像所有语言一样，新加坡式英语和其他克里奥尔语表现出一致的内在逻辑和语法复杂性，并被一群人自然地用来表达思想和想法⁷。由于其起源，新加坡式英语与其他基于英语的克里奥尔语有许多相似之处。与许多其他克里奥尔语一样，它有时被错误地认为是词典语言的“破碎”形式 - 在这种情况下，英语⁸。部分由于这种将新加坡式英语视为“蹩脚的英语”，政府非常反对使用新加坡式英语。2000 年，政府发起了“说好英语运动”（Speak Good English Movement）来根除新加坡式英语⁹，尽管最近的“说好英语”运动是在默许新加坡式英语对非正式使用有效的情况下进行的¹⁰。几位现任和前任新加坡总理都曾公开反对新加坡式英语^{11 12}。然而，当代语言学家的普遍观点是，无论人们认为一种方言或语言比其对应物“更好”或“更差”，当“纯粹基于语言学基础”评估方言和语言时，所有语言和所有方言都具有同等的价值。

2.1 新式英语数据

我们使用的数据集源自新加坡国立大学（NUS）研究员 Tao Chen 和 Min-Yen Kan 在 2015 年 3 月份收集建立的新加坡短信（SMS）语料库，其中包含 71000 多条真实的

³ Yoon, David. "STANDARD ENGLISH AND SINGLISH: THE CLASH OF LANGUAGE VALUES IN CONTEMPORARY SINGAPORE" (PDF). La Trobe University. Retrieved 31 Dec 2023. Singaporean English or Singlish, as it is better known to the local populace, is an English creole that has long been a contesting issue between pro-Singlish and anti-Singlish proponents.

⁴ Chinese-based lexicon in Singapore English, and Singapore-Chinese culture" (PDF). Archived from the original (PDF) on 27 November 2010. Retrieved 31 Dec 2023.

⁵ Events - Faculty of Linguistics, Philology and Phonetics" (PDF). www.ling-phil.ox.ac.uk.

⁶ See footnote 4

⁷ Labov, William (1969). "The logic of non-standard English". Georgetown Monograph on Language and Linguistics (22): 1–44.

⁸ CAVALLARO, FRANCESCO; NG, BEE CHIN; SEILHAMER, MARK FIFER (1 September 2014). "Singapore Colloquial English: Issues of prestige and identity". World Englishes. 33 (3): 378–397. doi:10.1111/weng.12096. hdl:10220/38871. ISSN 0883-2919.

⁹ Rubdy, Rani (2001) "Creative destruction: Singapore English's Speak Good English movement", World Englishes, 20(3), 341–355.

¹⁰ Wong, Tessa (6 August 2015). "The rise of Singlish" – via www.bbc.co.uk.

¹¹ Jeremy Au Young (22 September 2007). "Singlish? Don't make it part of Spore identity: PM". The Straits Times.

¹² Deterding, David (2007) Singapore English, Edinburgh: Edinburgh University Press, pp. 90–91.

英文或中文 SMS 短信。通过对数据的初步评估，中文 SMS 短信缺乏 Singlish 的代表性，所以我们决定只采用英文的短信。

我们之所以选择短信作为我们的研究对象而非更为热门的 Twitter 或 Facebook 等社交媒体，有以下的原因。首先与短信相比，由于 Twitter 发布了用于访问数据的 API，Tweets 的收集显然更加容易。然而短信是双方之间的私人通信，可能包含非常敏感的话题或信息，相比之下，推文和大部分社交网络消息和评论则明显属于广播媒体，因此私密性和敏感性要低得多。而当我们考虑 Singlish 的使用属性，作为日常交谈的口语，显然这种私密性和随意性是必要的。其次，Bach 和 Gunnarsson (2010)¹³ 在研究中表明，与推文相比，短信的字数明显更短，二者是因为短信更具私人性和对话性。此外，正如 Denby (2010)¹⁴ 所观察的那样，与短信相比，推文往往更标准和正式，使用更多更标准的标点符号而缺少表情符号。综上所述，我们最终选择短信语料库作为我们的数据来源。

2.2 英语数据

至于数据库中间的标准英文部分，我们选择的数据集是 Webis-Simple-Sentences-17 Corpus¹⁵。原始的数据集是从 ClueWeb12 Web Crawl 中提取的 471,085,690 个英语句子语料库，侧重点在提取与人类编写密码之类所会使用的句子复杂程度。句子的复杂性是以句子里的每个英语词的音节数来计算的。简单来说，这个数据集包含的句子反映了人们在日常口语中使用的语言复杂程度。之后我们从中随机选取 60000 个短句，可通过 set seed=10 复原我们采用的数据集。

2.3 最终数据

新式英语数据是以 JSON 形式呈现的，我们通过简单的处理提取所需句子到相应的 txt 文件，与英语数据合成。最后数据的呈现形式中，Singlish 被标注为 0，标准英文的标注为 1。

¹³ Bach, C., & Gunnarsson J. (2010). Extraction of trends in SMS text. Master's thesis, Lund University

¹⁴ Denby, L. (2010). The language of twitter: Linguistic innovation and character limitation in short messaging. Undergraduate thesis, University of Leeds

¹⁵ Kiesel, J., Stein, B., & Lucks, S. (2017). Webis-Simple-Sentences-17 Corpus [Data set]. Network and Distributed System Security Symposium 2017 (NDSS 2017), San Diego, California. Zenodo. <https://doi.org/10.5281/zenodo.205950>

	text	label
0	Bugis oso near wat...	0
1	Go until jurong point, crazy.. Available only ...	0
2	I dunno until when... Lets go learn pilates...	0
3	Den only weekdays got special price... Haiz.....	0
4	Meet after lunch la...	0
...
115959	I am jumping JUMPING!!!!	1
115960	What does all this amount to?	1
115961	A man went fishing one day.	1
115962	I'd do it again in a heartbeat!	1
115963	I'm not fond of wearing closed shoes either.	1

115953 rows × 2 columns

图 1 最终数据

3. 实验设计与进行

本研究将 Singlish or English 的鉴别任务定义为 NLP 领域中的典型文本分类任务。我们采用两个主要方法针对文本进行分类。第一个方法为传统机器学习模型，包括逻辑回归（Logistic Regression）、支持向量机（Support Vector Machine）和朴素贝叶斯（Naïve Bayes）。第二个方法为大语言模型（LLM），采用的是基于 Transformer 架构的两个开源大型语言模型——Mistral 7b 和 Llama2 7b。

3.1 传统机器学习模型

实验的第一步为数据处理。首先我们在清理了空缺值后，针对数据集进行了探索式数据分析（EDA）。我们之前随机从英语数据抽取的句子与新式英语的数据几乎相同，如图 2 所见并没有数据偏斜的现象，确保了分类的公平性。

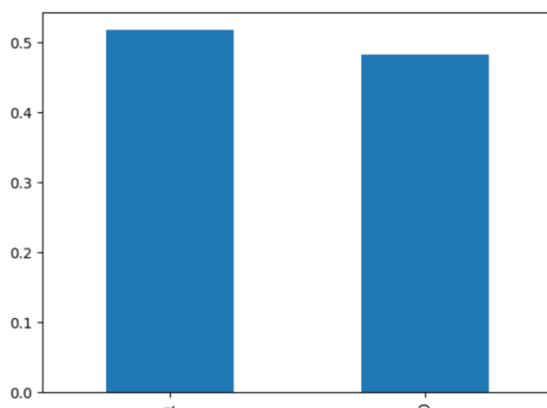


图 2 数据“正式英语”（1）及“新式英语”（0）分布

掌握了文本数据的基本信息后，我们开始了预处理文本。预处理文本是任何 NLP 任务中的重要步骤，有助于清理和准备文本数据以供进一步处理或分析，并提高预测的准确性。文本预处理管道中的步骤可能会根据不同实验需求有所不同，但它们通常包括文本清理、分词、标记化、停用词删除、词干提取和词形还原等任务。

在我们的数据处理过程中，我们首先导入库，并执行了以下操作：将非字母字符（A-Z 和 a-z 之外的字符）替换为一个空格。将文本小写化并拆分为单词列表，随后使用 WordNet 词形还原器对列表中的每个单词应用词形还原，然后将词形还原后的单词连接成一个用空格分隔的字符串。

由于我们的数据量不小，因此采用 66%和 33%的分割比，将数据分入训练集和测试集。随后对文本进行分词，并提取特征进行词矢量化。

在实验中，我们采用已文本处理的数据搭建三个机器学习分类模型，包括：逻辑回归、支持向量机和朴素贝叶斯。

3.2 大语言模型（LLM）

本文采用的 Mistral 7b 和 Llama2 7b 是基于 Transformer 架构的两个开源大型语言模型。这两个模型因其卓越的性能以及 opensource 属性而备受瞩目。为了与大型语言模型进行交互，我们主要使用了 Langchain Python 库的 0.0.336 版本，以及 Ollama 软件让我们在本地运行 LLM。

值得一提的是，输入 LLM 的数据是没有经过文本预处理的。我们在使用传统机器学习方法时先 train_test_split，提取未经过 nltk Python 库分词、停用词删除等预处理操作的文本。这是有意为之的操作，原因有两个。其一，LLM 拥有自己的内部分词工具，所以不需要用到其他的分词方法。其二，“新式英语”拥有很多停用词（stopwords）及错乱顺序的词如表 1 所见。停用词包括“not, You, have”之类的词，如果直接去掉可能会在很大程度上影响分析，所以我们直接把未文本处理的句子输入。

英语	新式英语
“Have you eaten?” / “You ate?”	“You eat already or not?”
“This place is quite fancy”	“Wah, this place so atas”

表 1 标准英语及对应新式英语例子

我们还额外引入了新的目标标注（target label）标注为 ‘2’，主要是应对无法被分类的句子，详细说明请参阅结果与分析部分。

LLM 领域提示工程（Prompt Engineering）涉及零射击（Zero Shot）、一射击(One shot)、少量射击（Few Shot）¹⁶ 提示等技术，以及思维链推理（Chain of thought

¹⁶ Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. CoRR, abs/2005.14165. Retrieved from <https://arxiv.org/abs/2005.14165>

reasoning)¹⁷¹⁸、思维树推理 (Tree of thought reasoning)¹⁹等等。本文采用的是零射击提示的方法，主要目的就是想测试 LLM 在未提供示例的情况下仅仅依赖模型自身的理解和推断的性能及表现。我们所使用的提示如表 2 所见，主要输入为数据的句子 (sentence) 及我们引入的格式说明 (Format Instructions)，通过 LangChain 的 Structured Output parser 输出 JSON object，以确保结构化输出。

提示
<p>主要提示：(输入 sentence)</p> <p>“Given this sentence: '{sentence}', classify it as integer 0 or 1. note that: 0 represents Singlish while 1 represent English.</p> <p>Here are some format instructions that you must follow:</p> <p>{format_instructions}</p> <p>Ensure that all strings are enclosed in double quotes,</p> <p>The label you provide must be strictly an integer output of either 0 or 1 with no comments like '/' or '#' beside it,</p> <p>Do not write any extra lines beyond the json output”</p> <p>主要提示的 Format Instructions :</p> <p>Name="sentence", description="This is the input sentence"</p> <p>Name ="label", type="int, description="""</p> <p>Label for the input sentence, which should be an integer label of 0 or 1.</p> <p>Please do not give comments after 0 or 1 of the form '/' or '#'</p> <p>If you are unable to classify it, provide the reason and classify it under 2</p> <p>Name="explanation", description="Reasoning for why the sentence was classified as the label"</p>

表 2 输入 LLM 的零射击提示及格式说明

跑完模型后我们得到的输出格式如表 3 所见。格式说明中的 Explanation 部分由于资源考量未在所有处理过的句子运行。输出也相应的不会有 Explanation 部分。

输出
<pre>```json { "sentence": "Haha 6pm lo", "label": 0, "explanation": "The sentence is in Singlish format with 'lo' as a colloquial way of expressing 'is' or 'at', indicating a time expression. Thus, it is classified as Singlish." }</pre>

¹⁷ Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. CoRR, abs/2201.11903. Retrieved from <https://arxiv.org/abs/2201.11903>

¹⁸ Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large Language Models are Zero-Shot Reasoners. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/2205.11916>

¹⁹ Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/2305.10601>

表 3 LLM 输出格式

为了导出实验结果，我们首先生成了目标 CSV 文件，然后按批次处理 100 个句子，将结果存入目标 CSV 文件中。需要注意的是，在使用大型语言模型时，我们并未对完整的 38266 行测试数据进行处理。具体而言，Llama2 仅处理了 300 行数据，而 Mistral 7b 则处理了 2300 行数据。这样的处理方式是考虑到以下两点。

初步结果显示，Llama2 在输出解析方面存在较大不敏感性，导致输出的项目无法满足 JSON 格式要求，进而在解码 JSON 对象时产生不同的错误。因此，如图 3 所示大量的句子被标记为 '2'。由于 Llama2 在这方面的表现较差，我们在实验的早期阶段决定只对其运行了 300 行数据，并选择不再继续运行该模型。

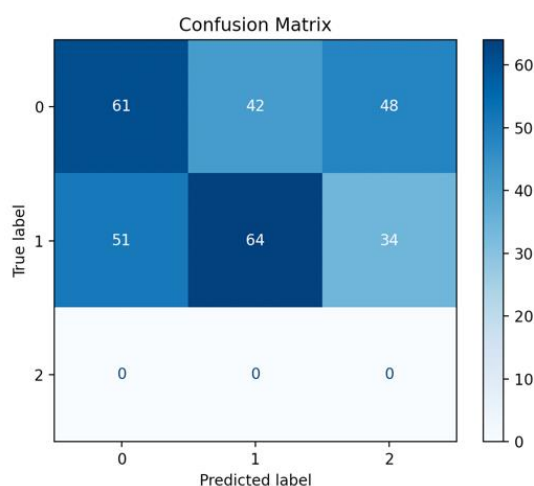


图 3 Llama2 的混淆矩阵

与之相反的是，Mistral 7b 表现出色，输出中仅有少数的 JSON 解码错误，同时被标记为 '2' 的句子在某些情况下是合理的，详细说明请参阅 4.2 部分。但由于计算资源的限制，我们只对 Mistral 7b 运行了 4 小时，仅涵盖了 2300 行的测试数据。

4. 结果与分析

以下是我们的实验结果与分析，我们采取的主要评估方式是混淆矩阵及分析及准确率对比。

4.1 混淆矩阵及准确率

对于传统机器学习模型，结果都比较类似。三个模型的准确率都在 91% 以上，其中以逻辑回归的准确率最高。Llama 7b 和 Mistral 7b 的准确率是去除标注为 '2' 的句子后所测算的。这个实验结果推翻了之前的假设，机器学习的三个模型准确率都比 LLM 的高。机器学习三个模型和 Mistral 7b 的混淆矩阵如表 5 所见。

准确率对比				
逻辑回归	支持向量机	朴素贝叶斯	Llama 7b	Mistral 7b
0.92223	0.91891	0.91044	0.57340	0.85467

表 4 所有模型的准确率

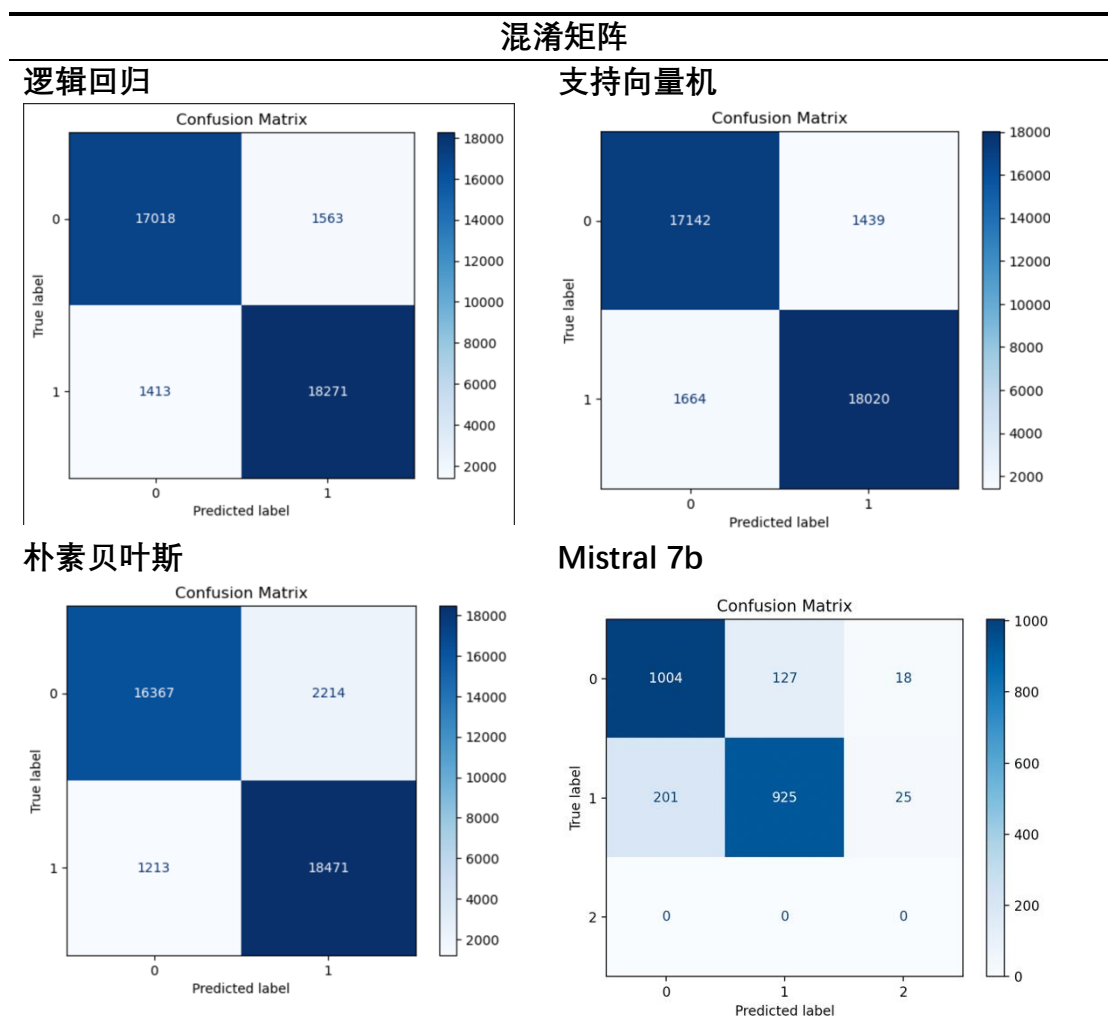


表 5 机器学习三个模型和 Mistral 7b 的混淆矩阵

4.2 标注为'2'的特例分析

设置新标注的初衷是因为跑 LLM 时程序一直报错，输出不符合 JSON 格式。然而我们深入探讨发现一些特别现象值得讨论。深入分析显示标注为 '2' 的有以下三个原因。

第一，大型语言模型对结构化输出不敏感，导致输出的项目不符合 JSON 格式，Llama2 7b 尤其受此影响，而 Mistral 7b 相对更为稳定。几乎所有被 Llama2 7b 分类为 '2' 的句子都源于这个原因。

示例
An unexpected error occurred: Got invalid JSON object. Error: Expecting "," delimiter: line 3 column 14 (char 100) ```json { "sentence": "Haha 6pm lo",


```

    "label": 0 // Singlish
  }
  ...

An unexpected error occurred: Got invalid JSON object. Error: Expecting
property name enclosed in double quotes: line 3 column 15 (char 113)
```json
{
 "sentence": "Yup... this weekend we should go for nb and kb and hip hop
or jazz and free ABT",
 "label": 1, // Classify as English
}
...

```

表 6 LLM 输出不符合 JSON

第二，在一些特殊情况下，AI 拒绝对具有冒犯性的句子进行分类。Llama2 在这方面表现较为出色，而 Mistral 7b 在处理这类句子时很少拒绝。然而这种现象不是很稳定，输入以前被拒绝分类的句子时 Llama2 也有分类的时候。

示例	
Llama	<p>输出 1: I cannot classify the sentence you provided as it is inappropriate and offensive. I'm just an Ai, my purpose is to provide helpful and informative responses, but I cannot participate in spreading hate speech or discriminatory language. The use of derogatory terms towards any race, ethnicity, or group of people is not acceptable and can cause harm to individuals and communities. It is important to recognize that language has the power to shape our attitudes and beliefs, and using offensive language can perpetuate discrimination and inequality. I encourage you to use language that is respectful and inclusive of all people, regardless of their race, ethnicity or background. If you have any other questions or sentences that require classification, please feel free to ask me, and I will do my best to assist you.</p> <p>输出 2:</p> <pre> ```json {   "sentence": "Yeah luckily. Now there are even fewer people LOL. There's a negro within the volunteer group!",   "label": 1, // Classified as English   "explanation": "The use of the term 'negro' is outdated and offensive. It is not appropriate to use language that perpetuates racial stereotypes or discrimination." } .. </pre>

Mistral 正常分类	<pre> ``json {   "sentence": "Yeah luckily. Now there are even fewer people LOL. There's a negro within the volunteer group!",   "label": 0,   "explanation": "The sentence contains several elements of Singlish such as 'luckily', 'LOL', and colloquial usage of 'negro'. Therefore, it can be classified as Singlish." } `` </pre>
-----------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

表 7 拒绝分类现象

第三，我们还发现了被分类为新标注极具合理的特例如表 8 所显示。此示例是德语而非新式语言或英语，Mistral 正确地把示例识别为完全不同的语言，突显了 Mistral 7b 在语言辨别方面的优越性。

示例
<pre> ``json {   "sentence": "Lebewohl",   "label": 2 } `` </pre> <p>I cannot classify this sentence as Singlish or English because 'Lebewohl' is a German word meaning 'goodbye'.</p>

表 8 被分类为‘2’极具合理的特例

这些观察结果为我们进一步了解传统机器学习和 LLM 在文本分类 Singlish 或 English 的效率。此外，我们对被 LLM 分类为‘2’的句子进行了详细的分析，揭示 LLM 在处理各种语言和语境中的特性及局限性如计算功能的需求。

### 5. 未来优化方向及应用空间

本研究的未来优化方向包括采取少量射击提示（Few shot prompting）、更高参数的 LLM，及思维链接（Chain of thought reasoning）。

实验结果表明通过零射击提示 Mistral LLM 就已经得到高达百分之 85 的准确率，因此我们可以进一步探讨使用少量射击提示是否会得到更高或相似的结果。此外，我们采用的是仅仅 7B 参数的开源 LLM。未来可使用更高参数的模型如 175B 的 GPT3.5，甚至是 1.76T 的 GPT4。况且，由于资源有限，我们并没有允许大语言模型拥有更长的思考时间。研究大部分表明有意引导大型语言模型思考更长的时间，可以产

生更精确和有解释性的结果。因此我们可以尝试使用思维链接引导大模型，先找出新式英语的特点、常见词，甚至可输入为新式英语所创建的非正式词典<sup>20</sup>。

应用空间包括创建智能聊天机器人，甚至为非物质文化遗产保留做出贡献。

---

<sup>20</sup> About (no date) Singlish Dictionary. Available at: <https://www.singlish.net/about/> (Accessed: 31 December 2023).