在多语言和代码转换环境中评估大型语言模型对多词表达的处理 *

Frances Laureano De Leon
School of Computer Science
University of Birmingham
Birmingham, UK
{ fxl846@student.bham.ac.uk

Harish Tayyar Madabushi
Department of Computer Science
University of Bath
Bath, UK
htm43@bath.ac.uk

Mark G. Lee
School of Computer Science
University of Birmingham
Birmingham, UK
m.g.lee@bham.ac.uk

ABSTRACT

多词表达，以非组合性意义和句法不规则性为特征，是一种细微语言的例子。这些表达可以被字面使用或成语化使用，导致意义的重大变化。尽管大型语言模型在许多任务中表现优异，但它们处理这种语言细微差别的能力仍不确定。因此，本研究评估了最先进的语言模型如何处理可能具成语特征的多词表达的歧义，特别是在不常见的环境中，模型不太可能依赖记忆。通过评估葡萄牙语和加利西亚语，除了英语之外，并使用一个新的代码混合数据集和新任务，我们发现大型语言模型尽管有其优点，但在处理细微语言方面仍然困难。特别是，我们发现最新的模型，包括 GPT-4，在检测和语义任务中未能优于 xlm-roBERTa-base 基准，尤其是在我们引入的新任务上表现尤其较差，尽管它与现有任务相似。总的来说，我们的结果表明，多词表达，尤其是那些有歧义的，仍然对模型构成挑战。

***K*eywords** Multiword expressions · Code-switching · Large language models

1 介绍

多词表达（MWE）由于其语言复杂性，在自然语言处理（NLP）中带来了显著的挑战。MWE 由多个词组成，这些词传达特定的意义，该意义可能是不可组合的，并表现出句法不规则性 [1, 2] 。这些表达在各种语言和领域中广泛存在，从艺术到科学 [3] 。估计一个说话者词汇中的 MWE 数量与单词的数量相当 [4] 。MWE 通常根植于文化背景和交流，而不是简单的词汇意义，这使得在 NLP 系统中，特别是在习语上下文中的解释更加具有挑战性 [5] 。此外，MWE 可能具有模糊性，既可以惯用地使用，也可以字面地使用，这可能会导致潜在的混淆 [6] 。例如，英语 MWE “break the ice” 可以指其字面意义或描述让不熟悉的人彼此更容易相处。这类表达，被称为可能的习语表达（PIE） [7] ，需要理解字面和比喻意义。鉴于 MWE 的独特性，最先进的 NLP 模型准确捕捉其意义是至关重要的。

大型语言模型（LLMs）迅速成为自然语言处理（NLP）中的主要方法，在零样本和少样本设置下的广泛任务中展示了强大的语言理解和生成能力 [8] 。然而，尽管表现强劲，LLMs 仍有其局限性，包括生成幻觉—流畅但不准确或编造的文本—以及依赖于来自预训练数据的统计共现模式 [9, 10] 。因此，本研究旨在评估 LLMs 是否能够真正理解细微的语言差异，使用多词表达式（MWEs）作为评估手段。MWEs 特别相关，因为它们的含义可以是习语性的，也可以是字面意义的，需要理解语言的社会和语言学方面。由于 MWEs 存在于所有语言中且频率和用法各不相同，它们在多语言背景下的解释可能更加复杂。其中一个挑战是代码切换（CS），即在对话或句子中交替使用不同的语言 [11, 12] 。代码切换在多语社区中很常见，为 NLP 系统带来了额外的困难，因为 MWEs 可能会跨语言分布。为了评估 LLMs 如何处理这种复杂性，我们使用了英文、葡萄牙文和加利西亚文的数据以及 CS 例子。通过同时包括高资源和低资源语言及 CS 数据，我们旨在评估模型是主要依赖于训练数据中的统计模式，还是能够在多语言和代码切换环境中有效地泛化。

我们进行实验，使用开源和闭源的大型语言模型来研究这些模型如何在不同语言和语境中处理多词表达式。具体来说，我们调查：

- 大型语言模型检测多词表达是作为习语使用还是字面使用的能力

*Citation : Authors. Title. Pages.... DOI:000000/11111.

- LLMs 在多词表达的非组合意义上的有效捕捉程度
- LLMs 从提示中获取信息的能力，和
- LLM 的规模与包含具有不可组合意义的多词表达的文本上的性能之间的关系。

我们的实验表明，尽管 LLMs 在许多其他任务中表现出色，但仍然难以处理多词表达（MWEs）。这一差距突显了这些模型的局限性，相较于精调的预训练语言模型（PLMs）如 xlm-roBERTa，后者是更小的模型，包含更少的参数。虽然一些结果表明英文中具有非组合意义的 MWE 处理与精调 PLMs 相当，但对其他语言则不能如此断言。我们的实验突显了当前模型的局限性，并强调在多语言环境中的应用时需谨慎考虑，尤其是处理微妙的语言。第 2 节回顾了有关 MWE 的相关文献。第 3 节中，我们描述了本研究使用的数据集，并概述了合成 CS 数据集的创建。第 4 节详细介绍了我们的实验设置，而第 5 节则呈现并讨论了我们的发现。最后，我们在第 **??** 节中总结了我们的结果，并对未来研究提出建议。

1.1 贡献

我们进行实验以评估生成模型在习语性和组合性上下文中检测和解释多词表达（MWEs）的能力。由于其频率较低，降低了大型语言模型依赖记忆进行评估的可能性。此外，为了进一步探究大型语言模型在 MWE 解释中是否使用记忆，我们创建了一个新的跨语言测试数据集，包含加利西亚语-西班牙语、英语-西班牙语和葡萄牙语-西班牙语三个语言对的例子，并公开提供。为了回答上述研究问题，我们不仅在标准任务中评估检测多词表达（第 **??** 节），还在新的任务中评估 MWE 的语义解释（第 **??** 节）。这包括合成 MWEs（第 4.2 节）的组合和一个新的代码切换数据集（第 3.1 节）。通过引入这些数据集、任务和设置，我们旨在减少模型记忆化的可能性，并评估它们解释 MWEs 的能力，特别是在非组合性的上下文中。我们的结果表明，大型语言模型并不能完全检测和表示 MWEs 的语义。此外，它们还表明，仅通过在提示中提供信息的提示学习对模型在处理未见词汇的复杂语言任务中是不足够的。

2 相关工作

研究表明，流行的 transformer 模型通常难以有效处理具有修辞的语言。关于仅编码模型的研究显示，PLM 未能充分捕捉和表示习语表达的含义 [13] 。此外，像 GPT-2 这样的生成模型在不使用缓解策略的情况下，被发现也很难处理具有修辞的语言 [14] 。[15] 揭示 transformer 模型不一致地捕获多词表达的语义，并发现它们依赖于表面模式。然而，研究表明，当仅编码的 PLM 使用合适的数据进行微调时，可以有效地表示多词表达 [16] 。研究生成模型如何处理具有修辞的语言的文献较少，尤其是专注于习语的研究。[8] 测试 BERT、RoBERTa 以及生成模型，即 GPT-2、GPT-3 和 GPT-NEO 在理解修辞语言，尤其是隐喻方面的能力。结果发现，所有模型都需要微调才能更好地理解修辞语言，这些模型的能力仍远未达到人类表现。[17] 对他们创建的一个英语数据集进行实验，以评估在零样本环境下，模型检测习语表达的能力。他们在 7B 版本的 Llama-2、Vicuna 和 Mistral 模型上进行实验。[18] 在英语、葡萄牙语和加利西亚语上对不同的最新生成模型进行了检测实验，发现尽管这些模型给出了有竞争力的结果，但他们未能达到微调后的预训练语言模型 (PLMs) 的表现。在这项工作中，我们还将探索模型在英语、葡萄牙语和加利西亚语中检测多词表达 (MWEs) 的能力。然而，我们的工作还研究了生成模型如何表示这些语言中多词表达以惯用方式使用时的句子意义。此外，我们调查了代码切换文本对模型性能的影响，并评估了模型使用合成例子管理未见多词表达的能力。

3 数据集

为了评估大型语言模型理解细微语言的能力，我们在所有实验中使用了 SemEval 2022 任务 2 数据集 [16] 。该数据集包含英语、葡萄牙语和加利西亚语中字面和惯用语境下的多词表达（MWE），并包括对抗性示例以测试模型在解释惯用语意义上的一致性。SemEval 任务包括两个子任务：子任务 a 是一个 MWE 解释任务，子任务 b 是一个 MWE 释义相似性任务，改编自语义文本相似性 (STS) 任务，用于生成模型。这两个任务旨在评估模型能否捕捉包含 MWE 的句子的意义，无论这些表达是组成式使用还是非组成式使用。我们改编了这些子任务，并基于相同的数据创建新任务，以测试对任务格式的熟悉度——这可能存在于指令微调中——是否影响性能。这也激发了我们使用单个数据集的变体而不是多个数据集的动机。该数据集的另一个优势在于其多语言性质，并且测试标签尚未发布，从而减少了模型在指令微调或预训练过程中见过答案的可能性。

3.1 代码切换数据集

为了进一步评估模型解释包含多词表达的文本的能力，我们基于前面描述的数据创建了一个新颖的代码切换（CS）数据集。代码切换在全球多语言社区中普遍出现，有助于混合语言变体的发展，如 Spanglish（西班牙语和英语的混合）。这个合成的代码切换数据集是通过将西班牙语与原始的英语、葡萄牙语和加利西亚语的单语句子结合起来创建的。

我们使用 OpenAI 的 GPT-4-turbo-0125 模型生成 CS 示例，温度设置为 0.65，种子值为 42。0.65 的温度提供了确定性和多样性之间的平衡，使模型能够生成多样但连贯的输出，同时确保西班牙语一致地融入句子结构中。

在生成过程中使用了两个基本提示：一个用于包含 MWEs 的所有例子（来自子任务 a 或子任务 b），另一个用于子任务 b 中的对抗性例子，这些例子不包含 MWEs，而是其字面意义的释义。生成是针对每个语言对单独进行的，将西班牙语与英语、葡萄牙语和加利西亚语分别混合，以尽量减少输出中的错误。所使用的提示见表 6 。

理解多词表达在代码切换（CS）环境中的功能可以为建模真实世界的语言使用提供见解。随着生成模型被多语言使用者越来越多地使用，本研究评估了最先进的模型如 GPT 处理混合语言输入的能力。该方法还强调了一些可能被单语评价方法忽视的挑战。因此，我们使用 CS 文本来确保句子意义与原始数据集保持一致，同时减少模型依赖表面级共现模式的可能性。

4 实验设置

在本节中，我们介绍我们的实验和用作此工作的模型。我们进行了三项不同的实验：多词表达（MWE）解释任务、合成 MWE 解释任务和 MWE 释义相似性任务。这些实验旨在确定模型能否识别含有微妙语言的文本的存在和意义。大多数实验是在零样本和少样本设置中进行的。

我们在实验中使用了四个模型，包括开源和闭源模型：（1）gpt-3.5-turbo-0125，（2）gpt-4-0125-preview，（3）meta-llama-3-70b-instruct，以及（4）meta-llama-3-8b-instruct。我们利用 OpenAI API 进行 GPT 模型的实验，并利用 Replicate API 进行 Llama 模型的实验。对于所有模型，我们使用以下参数：temperature 为 0，seed 为 42。对于 Llama 模型，我们指定最大 token 数为 7，而对于 GPT 模型，我们使用 5。

4.1 提示

在本节的所有实验中，我们使用三个不同的提示生成 LLM 的输出。这些提示基于 OpenAI 网站上 [2] 中概述的策略。具体来说，我们采用了三种方法：'persona' 提示，其中模型扮演语言学家的角色；'think' 提示，在分配标签之前，指示模型反思句子的含义；'ChatGPT' 提示，由 ChatGPT-turbo-3.5 生成。每个提示根据特定任务进行了调整。我们选择使用三个提示，因为给定不同的输入，模型结果可能会有所不同 [19] ，这使得可以探索给定输入变化时模型输出的可能变化。这些提示在表格 7 中提供。

我们进行多词表达式（MWE）词义消歧实验，在这些实验中，我们让大型语言模型（LLM）判断句子中给定的 MWE 是以习语方式还是字面意思使用。我们在零样本和少样本设置中使用 SemEval 竞赛子任务 a 的数据进行这些实验，该数据集包括总计 2342 个例子：英语 916 个，葡萄牙语 713 个，加利西亚语 713 个。对于少样本实验，模型提供了 5 个例子：英语 3 个，葡萄牙语 2 个。每种语言至少包含一个 MWE 以字面和习语方式使用的实例。用于此任务的还有 SemEval 子任务 a 的原始测试数据集及合成的 CS 数据集。在 CS 实验中，少样本例子的语言未作更改，也就是说，提供给 LLM 的例子仅有英语和葡萄牙语。模型接收到包含 MWE 文本以及 MWE 本身的输入。我们提示模型生成一个标签，在回答“在这句话中，多词表达式 [MWE] 是习语还是字面意思?”的问题时，标签为“习语”或“字面”。前面提到的三个基本提示用于从每个模型生成答案。这些实验的结果呈现在表 1 和表 2 中。

4.2 合成多词表达解释任务

合成 MWE 实验的目的是评估在提示中提供额外信息是否可以提高模型性能。我们特别希望测试在提供更多上下文时，模型是否能够理解细微差别的语言。这种方法基于 [20] ，其中通过引入和定义合成词作为现有概念来评估大型语言模型通过提示学习新词汇的能力。很可能大型语言模型没有在我们的数据集中遇到过所有的 MWE，尤其是葡萄牙语和加利西亚语的那些。因此，评估模型能否通过基于提示的学习来获取新词或 MWE 的知识是很重要的。

为了研究这一点，我们使用了子任务 a 的开发集，其中包含英语和葡萄牙语的例子（但不包括加利西亚语），并用合成的多词表达代替每个例子中的多词表达。该模型使用与多词表达解释实验相同的三个基本提示，并为合成的多词表达增加了定义，该定义取自被替换的原始多词表达。这些实验的结果显示在表格 4 中。

在这项工作中，我们旨在评估大型语言模型（LLMs）理解包含多词表达项目（MWEs）的文本意义的能力。我们为每个模型提供句子对：有些包含 MWE，而其他则包含正确或错误的 MWE 释义。提供给模型的输入示例如表 8 所示。为了进行这些实验，我们改编了 SemEval Task 2 竞赛中的子任务 b，最初设计为 STS 任务。我们构建了三个不同的提示，如第 4.1 节所述，每个提示都要求模型判断两个句子是否具有相似的意义。每

[2] https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

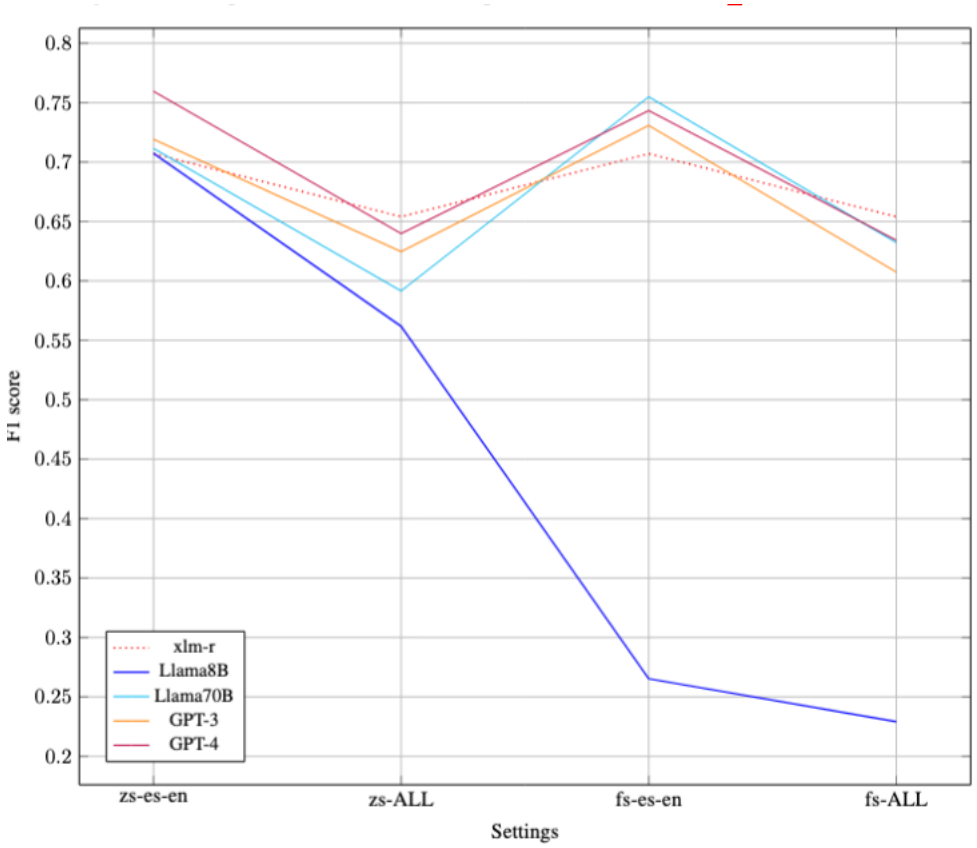

Figure 1: 使用 CS 数据集的每个模型的 F1 分数。该图显示了数据集中西班牙语-英语 (es-en) 语言子集的 F1 分数，以及零样本 (zs) 和少量样本 (fs) 设置中所有 CS 示例合并的总宏 F1 分数。

Model	F-1 Score EN	F-1 Score PT	F-1 Score GL	F-1 Score ALL	Code-switched
GPT-4	0.7480 ± 0.23	0.5701 ± 0.005	0.5395 ± 0.012	0.6336 ± 0.013	No
GPT-3	0.7135 ± 0.030	0.5359 ± 0.015	0.5243 ± 0.101	0.6082 ± 0.053	
Llama 3-70B	0.6918 ± 0.025	0.5097 ± 0.32	0.4420 ± 0.044	0.5613 ± 0.034	
Llama 3-8B	0.7013 ± 0.022	0.4628 ± 0.047	0.3636 ± 0.061	0.5265 ± 0.032	
xlm-roBERTta	0.7070	0.6803	0.5065	0.6540	
GPT-4	0.7595 ± 0.025	0.5759 ± 0.002	0.5411 ± 0.007	0.6398 ± 0.012	Yes
GPT-3	0.7191 ± 0.021	0.5417 ± 0.012	0.5575 ± 0.063	0.6246 ± 0.030	
Llama 3-70B	0.7114 ± 0.020	0.5496 ± 0.033	0.4726 ± 0.047	0.5915 ± 0.034	
Llama 3-8B	0.7072 ± 0.024	0.5276 ± 0.051	0.4076 ± 0.040	0.5619 ± 0.034	

Table 1: 这些是在零次实验中的检测实验得分。报告的 F-1 得分包括针对每种测试语言的得分，以及综合的宏 F-1 得分（全部）。

个提示指出，有些句子可能包含习惯用法的 MWE。如果句子的意义相似，则模型被指示输出“true”；如果不同，则输出“false”，将该任务转变为一个二元分类问题。

这些实验在零样本和少样本设置下进行。我们使用了来自原始 SemEval 竞赛子任务 b 的开发分区，该分区仅包含英文和葡萄牙文的示例。这个数据集包括含多字表达式（MWE）及其释义的句子对，以及来自标准语义文本相似性（STS）基准的示例。数据集中的金标准标签包括 Spearman 等级相关评分（对于 STS 基准示例）、得分为 1（表示语义相同）或标签为 NONE（表示语义不同）。在调整任务并移除相似度分数不是 1 或 NONE 的示例之后，我们保留了 974 个示例：其中 521 个是英文，454 个是葡萄牙文。由于任务需要二元输出，我们排除了所有 Spearman 得分不是 1 的示例，仅使用标记为 1（相似）或 NONE（不相似）的示例。这需要访问金标准标签，因此我们无法使用原始竞赛评估器，因为它假定分数没有修改。除了在原始开发数据上进行测试外，我们还在我们生成的 CS 示例中进行实验，其中每种语言与西班牙语混合。这些实验的结果

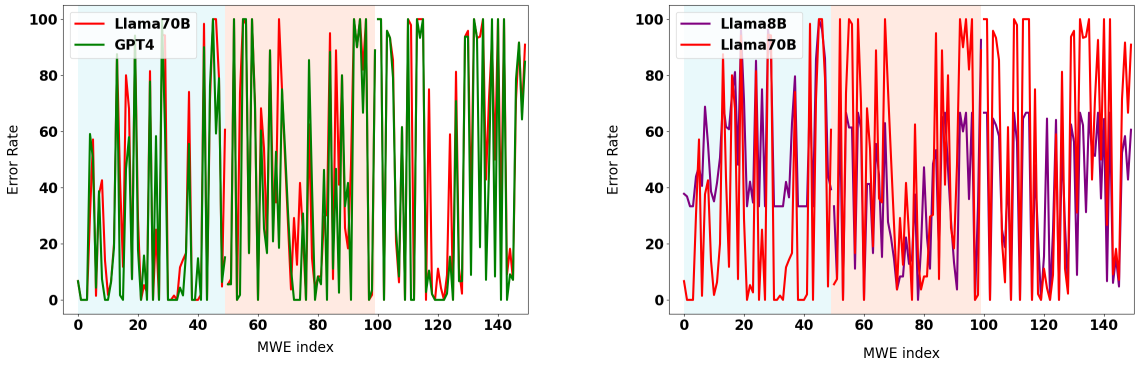

Figure 2: 测试集中所有多词表达式的错误率。这些错误率按每种语言的字母顺序排列。英文的多词表达式在浅蓝色背景中，葡萄牙语的在浅橙色背景中，加利西亚语的在白色背景中。左图显示了 LLama 70B 和 GPT-4 的错误，右图显示了 Llama 70B 和 Llama 8B 的错误。

Model	F-1 Score EN	F-1 Score PT	F-1 Score GL	F-1 Score ALL	Code-switched
GPT-4	0.7432 ± 0.006	0.5705 ± 0.008	0.5356 ± 0.008	0.6310 ± 0.003	No
GPT-3	0.7310 ± 0.015	0.5455 ± 0.003	0.4755 ± 0.092	0.6006 ± 0.034	
Llama 3-70B	0.7501 ± 0.017	0.5638 ± 0.007	0.5161 ± 0.016	0.6243 ± 0.014	
Llama 3-8B	0.2869 ± 0.397	0.1864 ± 0.305	0.1996 ± 0.341	0.2359 ± 0.360	
xlm-roBERTta	0.7070	0.6803	0.5065	0.6540	
GPT-4	0.7433 ± 0.014	0.5831 ± 0.012	0.5342 ± 0.002	0.6341 ± 0.009	Yes
GPT-3	0.7308 ± 0.009	0.5433 ± 0.009	0.4988 ± 0.064	0.6074 ± 0.024	
Llama 3-70B	0.7549 ± 0.011	0.5666 ± 0.007	0.53 ± 0.014	0.6321 ± 0.008	
Llama 3-8B	0.2653 ± 0.391	0.1737 ± 0.261	0.2107 ± 0.327	0.2291 ± 0.346	

Table 2: 这些是在少样本设置下的检测实验的评分。报告的 F-1 分数包括每种单独测试的语言的分数，以及综合的宏 F-1 分数（ALL）。

显示在表 **??** 和表 3 中。我们还在用于语义任务的相同数据集上进行实验，我们用合成的 MWE 替换每个示例中的真实 MWE，并提供模型以原始 MWE 的定义作为合成 MWE 的语义。这些实验的结果见表 5 。

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL	Code-switched
GPT-4	0.7628 ± 0.030	0.6304 ± 0.013	0.7000 ± 0.021	No
GPT-3	0.6497 ± 0.015	0.6296 ± 0.109	0.6404 ± 0.060	
Llama 3-70B	0.7204 ± 0.006	0.5570 ± 0.006	0.6425 ± 0.006	
Llama 3-8B	0.7045 ± 0.009	0.5771 ± 0.010	0.6440 ± 0.008	
xlm-roBERTta	0.7590	0.7658	0.7712	
GPT-4	0.5455 ± 0.008	0.5209 ± 0.022	0.5362 ± 0.014	Yes
GPT-3	0.5634 ± 0.051	0.5841 ± 0.011	0.5732 ± 0.026	
Llama 3-70B	0.5670 ± 0.010	0.4434 ± 0.004	0.5071 ± 0.007	
Llama 3-8B	0.6172 ± 0.011	0.5073 ± 0.030	0.5651 ± 0.020	
xlm-roBERTta	0.6752	0.7180	0.7020	

Table 3: 在少样本环境中进行语义实验的结果。

5 结果与讨论

大语言模型在解释习语的多词表达方面是否有效？ 总的来说，LLM 在检测惯用表达方面的能力不如较小的微调 PLM。这在使用原始数据集的加利西亚语和葡萄牙语实验中尤其明显。鉴于大多数这些模型主要在英语数据上训练，因此不出所料，它们在数据集的英语部分表现最佳，大多数模型在零样本和少量样本设置

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL
GPT-4	0.6552 ± 0.031	0.5955 ± 0.031	0.6392 ± 0.025
GPT-3	0.6231 ± 0.061	0.6056 ± 0.041	0.6225 ± 0.050
Llama 3-70B	0.5514 ± 0.037	0.5600 ± 0.018	0.4289 ± 0.030
Llama 3-8B	0.5696 ± 0.076	0.5677 ± 0.077	0.5689 ± 0.026
Random baseline	0.5441	0.5170	0.5387

Table 4: 使用合成 MWEs 在零样本设置下进行检测实验的实验结果。

Model	F-1 Score EN	F-1 Score PT	F-1 Score ALL
GPT-4	0.5454 ± 0.016	0.4855 ± 0.025	0.5183 ± 0.017
GPT-3	0.5655 ± 0.004	0.5204 ± 0.071	0.5444 ± 0.036
Llama 3-70B	0.4655 ± 0.028	0.3885 ± 0.030	0.4289 ± 0.030
Llama 3-8B	0.5455 ± 0.029	0.4711 ± 0.090	0.5108 ± 0.060
Random baseline	0.5092	0.4985	0.5025

Table 5: 使用合成多词表达式进行零样本设置的语义实验的实验结果。

中达到 F1 分数 0.70+。有趣的是，在 CS 实验中，所有模型的总体宏观 F1 分数都有所提高。这种改进可能是由于在所有示例中包含了西班牙语，这种语言在预训练数据中可能比加利西亚语和葡萄牙语更普遍。一些模型在数据的英语分区中超过了 xlm-roBERTa 的 F1 分数 0.7070，但没有一个能够达到超过 0.76 的分数。然而，与微调的 xlm-RoBERTa-base 模型在零样本和少量样本设置中相比，所有模型都未能达到葡萄牙语 0.6803 的基线和总体 F1 分数 0.6540。值得注意的是，Llama3-8B-instruct 模型在少量样本设置中无法为原始和 CS 数据集生成有效输出。似乎少量样本示例与输入文本的结合使模型陷入困惑，以至于大多数模型输出都不是包含标签的单词答案。图 **??** 和 1 展示了 Llama3-8B 在少量样本实验中的挣扎，这从用于生成模型输出的三个提示中的高分数方差可以看出。尽管与一个小得多的 PLM 相比，LLM 在检测细微语言方面具有竞争力，但总体而言，它们在其他资源较少的语言中表现不佳。

大型语言模型在多词表达式的非组合意义上捕捉到的程度如何？ 大型语言模型（LLMs）似乎在理解包含多词表达式（MWEs）的文本上存在困难。总体而言，与 MWE 解释任务不同的是，少量示例在这里帮助大多数模型在葡萄牙语和英语文本中获得更高的 F1 分数。在少量示例设置中，GPT-4 能够与英文的 xlm-roBERTa 基线相媲美，得分为 0.7638，略微超越了 xlm-roBERTa 的 0.7590。所有模型在葡萄牙语的零样本和少样本设置中都未达到组合基线得分和 F1 得分。这可能表明模型未能捕捉到包含 MWEs 的文本的真正含义，并可能依赖于伪相关来做出决定。总体而言，鉴于 LLMs 难以超越拥有少得多参数的微调 PLM，它们似乎难以捕捉复杂语言的含义。

结果表明，LLM 在合成 MWE 解释任务上的表现普遍高于随机情况。在提示中我们提供了 MWE 的定义，其中 GPT-4 实现了最高的综合 F1 分数 0.6392。然而，与使用真实 MWE 的情况相比，它们的表现显著较低。这表明模型可能在训练时记住了一些真实的 MWE。对于涉及合成 MWE 的语义实验，模型表现仍接近随机基线。跨所有语言，GPT-3 获得了最高分 0.5444，但这些结果仍未显著超过随机基线 0.5025。这表明，当词汇与预训练期间遇到的模式不同时，在提示中提供额外信息并不能改善模型表现。模型似乎高度依赖于预训练数据中的统计模式，并在面对不熟悉的模式时，即使有上下文提示，仍无法生成正确的响应。总体而言，模型难以应用合成 MWE 的含义以完成任务，这表明额外的上下文并未帮助它们解决复杂的语言问题或不熟悉的词汇。

LLM 的规模与包含非组合短语的文本表现之间是否存在关系？ 较小的 Llama 8B 模型在零样本检测任务中表现出乎意料的好，其结果与 Llama 70B 和 GPT-3 在英语文本方面相当。然而，它在葡萄牙语和加利西亚语的表现稍有下降。值得注意的是，Llama 8B 模型在少样本检测任务中无法处理原文和 CS 文本，并且比其他模型对提示更为敏感，这在表格 2 中的高标准差可以说明。GPT-4 在语义任务中的英语表现优于其他 LLM，并且是唯一能够在此情境下处理英语-西班牙语 CS 文本的 LLM。虽然需要进一步的实验，但这些结果表明，较小的模型如 Llama 8B 可能足以实现与大得多的模型在英语 MWE 检测中相当的结果。然而，结果也显示，当面临未知词汇并代表 MWE 的语义时，GPT-4 优于其他模型。

在进行 MWE 解释任务时，使用原始数据集的情况下，某些 MWE 始终被所有模型正确分类，如表 9 所述。图 **??** 和图 1 展示了模型对不同提示的敏感性。特别是，标准差，尤其是在 Llama 8B 的情况下，显示了模型输出可能因提示而异的情况。有一部分 MWE 在检测任务中被所有模型正确分类，这些列在表 9 中。它列出了所有模型正确识别的 MWE，以及特定模型家族正确分类的 MWE，以及两个最大模型之间的重叠。我们还研究了单个 MWE 在各模型之间错误率的相关性。Llama 模型之间的错误率的相关系数为 0.64，Llama 70B

与 GPT-4 之间的相关系数较强，为 0.70。然而，两种 GPT 型号之间的相关性较弱，为 0.32，可能表明它们训练数据上的差异。图 2 显示了 Llama 70B 与 GPT-4 以及两个 Llama 模型之间的错误情况。曲线叠加以进行比较。某些模型之间错误的强相关性可能表明在需要细微语言理解的任务中存在共同限制，可能源于其架构设计或训练数据。

我们研究了先进的 LLM 是否能够有效捕捉和编码复杂语言的意义，特别是通过 MWEs，在各种语言和文化背景下进行。这些实验包括英文、葡萄牙语、加利西亚语和代码转换文本，我们发现 LLMs 未能完全检测和表示 MWEs 的语义。在英文检测任务中，LLMs 的性能与小型微调 PLM 相当，但在其他语言和任务中表现较差。有趣的是，代码转换文本在元语言任务上支持模型，但妨碍了它们表示 MWEs 意义的能力。在语义实验中，这一点得到体现，因为模型更容易检测 MWEs 是否作为习语使用，而非确定和捕捉包含习语的文本意义。我们的错误分析强调了当前模型在处理包含挑战性语言现象如 MWEs 的文本时的局限性。展望未来，我们建议探索与 GPT-4 性能匹敌的开放模型，重点关注提升模型捕捉复杂语言意义的能力的方法。需要新的方法使模型能够有效应对具有挑战性的语言现象，包括 MWE 和 CS。

6 局限性

我们的实验是在主要用英语数据训练的模型上进行的。尽管许多多语言模型包含一些其他语言的训练数据，但它们的大部分训练仍然使用英语。然而，我们的结果表明，即使处理英文文本，这些模型的表现也并不如预期。在一些实验中，我们使用合成的代码混杂数据而不是自然存在的代码混杂数据。

References

[1] Timothy Baldwin and Su Nam Kim. *Handbook of Natural Language Processing*. 2 edition, 2010.

[2] Mathieu Constant, Gülen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892, 12 2017.

[3] Aline Villavicencio and Marco Idiart. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733, 2019.

[4] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann A Copestake, and Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In *Conference on Intelligent Text Processing and Computational Linguistics*, 2002.

[5] Francesca Masini. Multi-Word Expressions and Morphology, 9 2019.

[6] Murathan Kurfal and Robert Östling. Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings. In Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberius, Carlos Ramisch, Ashwini Vaidya, Petya Osenova, and Agata Savary, editors, *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online, 12 2020. Association for Computational Linguistics.

[7] Hessel Haagsma, Johan Bos, and Malvina Nissim. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France, 5 2020. European Language Resources Association.

[8] Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the Ability of Language Models to Interpret Figurative Language. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States, 7 2022. Association for Computational Linguistics.

[9] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates, 12 2022. Association for Computational Linguistics.

[10] Cheongwoong Kang and Jaesik Choi. Impact of Co-occurrence on Factual Knowledge of Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735, Singapore, 12 2023. Association for Computational Linguistics.

[11] Aravind K Joshi. Processing of Sentences With Intra-Sentential Code-Switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, 1982.

[12] A Seza Dogruoz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. In *ACL/IJCNLP*, 2021.

[13] Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. Probing for idiomaticity in vector space models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online, 4 2021. Association for Computational Linguistics.

[14] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic, 11 2021. Association for Computational Linguistics.

[15] Filip Miletić and Sabine Schulte im Walde. Semantics of Multiword Expressions in Transformer-Based Models: A Survey. *Transactions of the Association for Computational Linguistics*, 12:593–612, 2024.

[16] Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States, 7 2022. Association for Computational Linguistics.

[17] Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty, and Emmy Liu, editors, *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid), 6 2024. Association for Computational Linguistics.

[18] Dylan Phelps, Thomas M R Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In Archna Bhatia, Gosse Bouma, A Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, and Alexandre Rademaker, editors, *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia, 5 2024. ELRA and ICCL.

[19] Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings, 2024.

[20] Julian Martin Eisenschlos, Jeremy R Cole, Fangyu Liu, and William W Cohen. WinoDict: Probing language models for in-context word acquisition. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia, 5 2023. Association for Computational Linguistics.

A 提示和示例

Prompts	Type
You are an expert linguist in code-switching between { lang } and Spanish (CS-lang-name). You will translate sentences you get into code-switch by switching some of the words into Spanish. Make sure that there is a balance between two languages. You will translate everything into code-switch apart from the multiword expression, denoted as 'MWE'.	Contains MWE
You are an expert linguist in code-switching between { lang } and Spanish (CS-lang-name). You will translate sentences you get into code-switch by switching some of the words into Spanish. Make sure that there is a balance between two languages. You will translate everything into code-switch.	Adversarial example

Table 6: 这些是用于生成合成代码转换（CS）数据集的提示。术语 lang 对应英语、葡萄牙语或加利西亚语。CS-lang-name 是指 Spanglish 或 Portuñol 之间的代码转换语言名称。当请求加利西亚-西班牙语混合时，这部分提示不包括在内。

Prompts	Type	Task
You are a linguist specializing in phraseology with extensive knowledge of idiomatic expressions, colloquialisms, and figurative language. Your task is to identify whether a given text contains a multiword expression used figuratively as an idiom. First, understand the meaning of the text provided by the user. Then, determine if the multiword expression is used figuratively or literally. Respond with 'idiom' for figurative use or 'literal' for literal use. Provide a one-word answer.	Think	Detection
You are a linguist specializing in phraseology. With your deep knowledge of idiomatic expressions, colloquialisms, and figurative language, you are considered the best in the field. Instruction: Given a text containing a multiword expression, determine if it is used as an idiom or literal. Provide a one word answer.	Persona	
Analyze the provided text to determine whether the given Multiword Expression (MWE) is used as an idiom or literally. If the MWE is used figuratively as an idiom, answer "Idiom"; if it is used in its literal sense, answer "Literal". Provide a one word answer.	chatGPT	
You are a linguist specializing in phraseology with extensive knowledge of idiomatic expressions, colloquialisms, and figurative language. Your task is to identify whether two sentences have a similar meaning or not. Some of the sentences may contain idiomatic expressions. First, understand the meaning of the text provided by the user. Then, determine if the sentences have a similar meaning. Respond with 'true' if the sentences are similar in meaning or 'false' if they mean different things . Provide a one-word answer	Think	Semantic
You are a linguist specialising in semantics. You possess a deep knowledge of language structure, and meaning. You are the best in the field. You will be presented with two sentences, some of the sentences contain figurative language. You will determine if two sentences are similar in meaning. Output 'false' if their meanings are different or 'true' if their meanings are similar. Provide a one word answer	Persona	
Analyze the provided text to determine if the given sentences, which may include idiomatic expressions, have similar meanings. If the sentences are similar, output "true." If the meanings differ, output "false." Provide a one-word answer. answer	chatGPT	

Table 7: 这些是为每个实验提供给所有模型的系统提示，根据实验是属于“检测”还是“语义”进行分类。

Examples of Data Inputs to Models	MWE	Experiment
But we need to remember, it was a very close call.	Close call	Detection
Takes about tres años for the last milk tooth to come in después de que el first one sprouts. (es-en-mix) Takes about three years for the last milk tooth to come in after the first one sprouts. (original)	Milk tooth	CS
Sentence 1 : Witten's swan song was far from a hit. Sentence 2: Witten's final performance was far from a hit.	Swan song	Semantic
Despite not receiving an inheritance, Eve Jobs is still living the wrawpths chaumb.	Wrawpths chaumb replaces High life	Synthetic MWE

Table 8: 为不同实验提供给模型的不同文本输入示例。

Correct for all	Correct for Llama	Correct for GPT	Correct for GPT-4 and Llama 70B
arma branca	arma branca	arma branca	arma branca • inner product
camisa branca	auditoría externa	camisa branca	auditoría externa • lime tree
centros educativos	camisa branca	centros educativos	incendios forestais • lingua galega
exame clínico	centros educativos	dirty word	baby buggy • market place
glóbulos vermellos	enerxía limpa	glóbulos vermellos	benign tumour • net income
	exame clínico	horas baixas	black cherry • news agency
	glóbulos vermellos	news agency	camisa branca • nut case
	incendios forestais	olho gordo	centros educativos • olho gordo
	lingua galega	pavio curto	exame clínico
	voto secreto		glóbulos vermellos

Table 9: 第一列展示了模型始终正确识别的多词表达（MWEs）之间的重叠。第二列显示了两个 Llama 模型之间正确单词的重叠，第三列显示了 GPT 模型之间的重叠，最后一列展示了 GPT-4 和 Llama 70B 之间的重叠。