Vuyko Mistral：适应性地将大型语言模型用于低资源方言翻译

Roman Kyslyi[1], Yuliia Maksymiuk[2], Ihor Pysmennyi[1]
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
[2]Ukrainian Catholic University
Correspondence: kyslyi.roman@lll.kpi.ua, yuliia.maksymyuk@ucu.edu.ua, pysmennyi.ihor@lll.kpi.ua

Abstract

在本文中，我们首次尝试将大型语言模型（LLMs）适应于乌克兰方言（在我们的案例中是胡楚尔方言），这是一种在喀尔巴阡山脉高地使用的资源贫乏且形态复杂的方言。我们创建了一个包含 9852 个方言到标准乌克兰语句对的平行语料库和一个包含 7320 个方言词汇映射的词典。我们还通过提出一种高级的检索增强生成（RAG）流程来解决数据短缺问题，以生成合成的平行翻译对，扩展了包含 52142 个例子的语料库。我们使用 LoRA 微调了多种开源 LLMs，并在标准到方言翻译任务上进行了评估，也与少样本 GPT-4o 翻译进行了比较。在缺乏人工标注者的情况下，我们采用了一种多指标评估策略，结合了 BLEU、chrF++、TER 和基于 LLM 的判断（GPT-4o）。结果表明，即使是小型 (7B) 微调模型也在自动和 LLM 评估的指标上均优于如 GPT-4o 的零样本基线。所有的数据、模型和代码已公开发布于：https://github.com/woters/vuyko-hutsul 。

1 引言

尽管在大型语言模型（LLMs）方面取得了最近的进展，但大多数研究和应用仍然集中在高资源语言及其标准变体 (?) 。这种不平衡对语言多样性产生了重大影响，尤其是对于缺乏足够文本资源和标准化正字法的未充分代表的方言 (?) 。尽管方言是许多社区语言认同的重要组成部分，但它们往往被排除在 NLP 工具和研究之外，限制了其可及性，并有可能进一步导致边缘化和灭绝的风险 (?) 。

语言技术，尤其是大型语言模型，在濒危和代表性不足的语言保护方面发挥着越来越重要的作用。虽然很多注意力集中在主要的土著语言（例如，毛利语、克丘亚语、因纽特语）(??) 上，但尽管面临类似的消亡和同化压力，国家语言的方言往往被忽视。方言变体，特别是在后苏联背景下，常常承载着在标准语言中未反映的被压制文化身份。这些方言不仅语言上丰富，而且文化上至关重要，值得计算上的关注。

乌克兰语根据全球标准本身是一种资源较少的语言，但它表现出丰富的内部变化，有诸如 Hutsul、Boyko 和 Lemko 等方言，它们保留了独特的语音、词汇和语法特征。在这些方言中，Hutsul 方言是其中语言上最独特的，并且有最多的书面资料，使用者分布在喀尔巴阡山脉。就文化而言，Hutsul 方言具有重要意义，因为它包涵了传统、民俗和独特的世界观，对于社区认同起着核心作用。

然而，缺乏数字化的语料库、词典和处理工具，使得它对现代 LLM 实际上是不可见的。

以下是胡楚尔方言的一些语言特征：

- 语音学：元音的变换，如用元音 "" 代替 "", ""(ya) （例如： "" "", "" "" （"yak" → "yek"，"yahoda" → "yehoda"））。
- 形态学：独特的词尾变化（ - ，-ci）（'-yed'，'-si'）以及保留的双数形式 两个苹果 （"两个苹果"，双数形式为 "yablutsi" 而不是复数形式 "yabluka"）。
- 词汇：罗马尼亚语、波兰语和德语借词，如 "" （奶酪）和 "" （去散步）。[1]

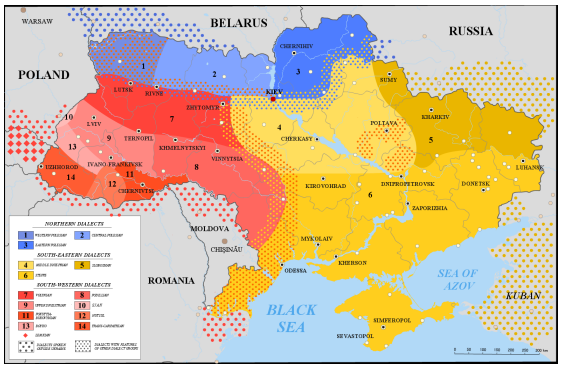

Figure 1: 乌克兰方言地图。胡茨尔方言位于西南部的喀尔巴阡山区。来源：维基百科

[1]https://en.wikipedia.org/wiki/Eastern_Romance_influence_on_Slavic_languages

在这项工作中，我们努力将大型语言模型（LLM）适配到乌克兰的胡茨尔方言，同时解决数据短缺和建模挑战。我们的贡献是：

- 新的平行语料库包含原始的胡茨尔语-乌克兰语（9852 个句子对）、7320 个方言词汇映射的词典，以及使用先进的 RAG 方法合成扩展的语料库（详细描述如下）（52142 个句子对）。
- 对多个开源大型语言模型进行微调，以完成乌克兰语到胡茨尔方言的翻译任务。

我们将任务框架设为标准到方言翻译，其中模型需要将标准乌克兰语作为输入，并生成语法正确（或尽可能接近）的胡茨尔方言。我们的模型表明，可以通过有限的平行数据和有针对性的增强策略来解决这种翻译问题。

据我们所知，这是首次尝试将大型语言模型（LLMs）适应乌克兰方言的工作，同时也是全球少数几项使用合成增强技术来解决方言到标准语言生成的问题之一。

近年来，我们可以看到对方言建模的兴趣日益增长，特别是在阿拉伯语、德语和罗曼语系方面。这些努力主要集中在方言与其标准变体之间的分类、生成和翻译。然而，大多数研究仍集中在资源丰富的语言和已有 NLP 资源的方言上。同时，对于乌克兰语，方言的 NLP 研究仍然未被深入探索。VarDial 研讨会系列支持了针对不同斯拉夫语言的相关任务的工作，例如跨方言机器翻译和形态建模。例如，研究博克马尔语和新挪威语之间的机器翻译，而另一个研究则解决北萨米语的方言变体问题。SIGMORPHON 2023 共享任务强调了基于词汇的屈折建模对于低资源形态变体的重要性。

将方言语言规范化为其标准形式的任务已经通过各种对齐技术进行了探索。对芬兰语、挪威语和瑞士德语中方言转录的句子级标准化进行了字符对齐方法的评估。该研究比较了方言学、语音处理和机器翻译的方法，发现训练的对齐方法相对于简单的 Levenshtein 距离只提供了小的优势。这表明在资源有限的方言环境中，简单而稳健的统计方法仍然可以提供强大的基准。此外，该研究强调了在处理高度可变和语音丰富的方言数据时，定制预处理和对齐工具的必要性。

1.1 大型语言模型和方言适应

最近的几项研究探讨了适应方言数据的大型语言模型（LLM）。**?** 提出了用于方言适应的任务无关适配器，而 **?** 引入了基于语言距离的动态适配器聚合。**?** 探索了针对形态丰富的方言的分词器改造。这些研究表明，架构和数据中心的干预都是有效适应所必需的。然而，这些方法主要在英语方言（例如非洲裔美国英语、印度英语）上使用经过策划的语料库（如 Multi-VALUE (**?**) ）进行评估，并依赖经过注释的方言到标准的配对，但是这些资源对于资源匮乏的方言来说很少可用。

1.2 低资源和合成数据技术

我们的工作还受益于之前在低资源翻译和使用合成数据进行文本生成方面的研究。**?** 和 **?** 提出了基于检索或提示的增强技术，以在数据有限的情况下提升表现。同时，我们提出了使用先进的 RAG 技术生成合成数据的新方法。

2 数据集创建

2.1 平行语料库收集

我们通过结合多种来源和注释策略，构建了第一个关于胡茨尔方言和标准乌克兰语的平行语料库。该数据集包括 9852 对句子，这些句子在句子层级上手动对齐。胡茨尔语的源文本收集自公开出版的书籍、民族志记录、民俗网站和方言博客。数据集的重要部分基于 Petro Shekeryk-Donykiv 的小说 " 迪多 · 伊万奇克"（Dido Yvanchik），这是一部用真实胡茨尔语写成的基础文学作品。我们特别感谢出版社 话语 和译者 伊万 · 安德鲁夏克 ，他们友善地批准将他们的现代标准乌克兰语翻译用于学术目的。

数据集中标准的乌克兰语参考资料要么是手动翻译的，要么是从现有的双语版本中获取的。为了确保语言的多样性，我们尝试包括日常对话和风格化叙述文本（例如，民间故事、歌曲等）的例子，但由于数据缺乏，一些主题仍未覆盖。

2.2 词汇资源

我们编纂了一本含有约 7300 个词对的胡楚尔语-乌克兰语词典。该工作始于书中出现的词汇 " "（Dido Yvanchik），但我们很快扩大了词汇量，从解释胡楚尔方言词汇的网站上获取了数据。其中最有用的网络资源包括：

- 《Hutsul 词汇词典》[2] 。
- “胡祖尔霍维尔”[3] 。
- 《喀尔巴阡地区乌克兰方言词典》[4] 。
- 佩特罗 · 哈夫卡的《胡茨尔方言解释词典》[5] 。

[2] https://karnauhova.at.ua/publ/1-1-0-3
[3] https://rakhiv-mr.gov.ua/hutsulskyj-hovir/
[4] https://evrika.if.ua/88/
[5] https://evrika.if.ua/1565/

- “胡茨尔词典”。[6]

所有这些页面都是自动抓取的。原始文本包含大量噪音：奇怪的字符、额外的评论、不均匀的制表，以及 Hutsul 条目和其乌克兰语翻译之间不一致的分隔符。我们编写了简单的清理脚本，将所有内容转换为单个 CSV 文件，然后手动检查列表以去除最后的错误。最终结果是一个包含 7,320 对 Hutsul-乌克兰语的干净词汇表。每个条目包括标准和方言的词形。

尽管有这些努力，词典仍然存在偏向于文学和民俗领域的词汇的问题。由于关于新闻、科学或政治等主题的胡茨尔文本的缺乏，我们的数据集中这些领域的词汇多样性不足。

2.3 通过高级 RAG 生成的合成数据

为了克服对胡索尔方言书面资源的缺乏，我们开发了一个先进的 RAG 流程来生成额外的胡索尔-标准句子对。该流程的基础是方言小说“迪多 · 伊万奇克”（德多 · 伊万奇克），它既是检索的主要语料库，也是语言示例的来源。我们使用 GPT-4o 构建了一个 RAG 模块，能够检索语义相关的胡索尔句子。对于每一步生成，我们创建了一个包含胡索尔语音和词汇变化代表性语言转换规则的提示。

RAG 管道的构建涉及几个步骤：

1. 语法规则提取：使用 " "（Dido Yvanchik）作为输入，我们提示 GPT-4o 提取并构建出胡特苏方言特有的语法转化规则。这些规则包括语音的变化、形态的交替以及句法的重组。我们使用来自维基百科的材料和《胡特苏方言的词形构造模型》**?** 扩充了这些规则，以创建一个全面的提示模板（见图 2）。

2. 通过 RAG 进行索引：我们将 “ ”（Dido Yvanchik）语料库索引到一个检索系统中，以作为生成使用 text-embedding-3-large [7] 的方言输出的参考基础。

3. 候选句子选择：标准乌克兰语句子是从 UberText 语料库中抽样的（**?** ）。对于每个这样的句子，我们使用 RAG 模块从 " "（Dido Yvanchik）中检索出三个语义上相似的胡祖尔式句子。

4. 提示构建：检索到的示例与标准的乌克兰语句一起被插入到提示模板中，作为翻译的来源。

5. 方言生成：GPT-4o 被指示使用提供的语法规则和示例作为上下文，生成输入句子的胡楚尔语翻译（见图 3）。

以下是我们基于规则的提示的主要部分（完整提示可以在这里找到：https://github.com/woters/vuyko-hutsul/blob/main/prompts/hutsul_rules_prompt.txt ）：

> Here are Grammatical Rules for Converting Ukrainian Text into the Hutsul Dialect:
> 1. Vowel Shifts:
> - “ ” “ ” (“yak” → “yek”)
> - “ ” “ ” (“yabluko” → “yebluko”)
> - “ ” “ ” (“yidesh” → “yedesh”)
> 2. Consonant Transformations:
> - “ ” “ ” (“divka” → “givka”)
> - “ ” “ ” (“choho” → “cho”)
> - “ ” “ ” (“ty” → “tsy”)
>
> 3. Word Order and Syntax:
> - “ ” “ ” (“I love you” → “Love I you”)
> - “ ” “ ” (“He is laughing” → “He laugh-reflexive”)
> - “ ?” “ ?” (“Do you know?” → “Do you know?” with dialectal marker “tsy”)
>
> Apply only contextually appropriate transformations.

这个过程在生成的数据集中产生了一些数据对齐的挑战。为了解决这些挑战，同时清理生成的数据集，我们开发了一种混合对齐策略。首先，我们利用了语言及其方言之间的预期文本相似性，使用 difflib 的 SequenceMatcher [8] 。这种方法直接比较字符序列，有效地识别即使是带有轻微方言差异的对。相似性低于 0.45 的对从数据集中移除。为了衡量剩余句子对的质量，我们使用了 **?** 所描述的几种统计指标：

- U-src –未对齐源字符的比例，
- U-tgt ——未对齐目标字符的比例，
- X - 交叉对齐对（交换）的比例

这些指标是通过使用快速对齐 (**?**) 获得的对称对齐对计算的。我们比较了三个数据集的对齐指标：原始手动标注的数据集（主要来自 " " (Dido Yvanchik)），原始合成生成的数据集，以及过滤后的合成数据集。

在过滤之前，合成数据中未对齐的源词和目标词的比例已经较低（U-src=0.139, U-tgt=0.136），相比原始数据（U-src=0.260, U-tgt=0.265）。然而，它表现出更高比例的交叉对齐（X=0.033，相比原始数据的 0.022），表明结构变化增加。

[6] http://www.webteka.com/hutsul-language/

[7] https://platform.openai.com/docs/models/text-embedding-3-large

[8] https://docs.python.org/3/library/difflib.html

为了提高我们生成数据集的质量，我们应用了基于对齐的过滤——对于每个句子对，我们使用了先前计算的统计数据（U-src, U-tgt 和 X），并且我们经验性地为它们定义了阈值：U-src < 0.1 、U-tgt < 0.1 和 X < 0.2 。

Metric	Original Dataset	Synthetic (Raw)	Synthetic (Filtered)
U-src	0.260	0.139	0.005
U-tgt	0.265	0.136	0.005
X	0.022	0.033	0.019

Table 1: 原始数据集、原始合成数据集和经过对齐过滤后的合成数据集之间的对齐质量指标比较。

任何超过一个或多个这些阈值的句子对都会被排除在最终数据集之外。此过程去除了不一致的例子，减少了重新排序的数量，并改善了对齐。结果是我们得到了一个质量更高的合成数据集，其结构对齐也更好，正如表 1 中的对比指标所示。

尽管我们意识到获得的合成数据存在一些差异，并且缺少真实方言语音中存在的某些词汇短语，但由于胡茨尔文本资源的短缺，其纳入是合理的。此过滤步骤有效地提高了合成数据集的一致性和可靠性，并为我们的训练数据集增加了额外的 52142 个短语对。

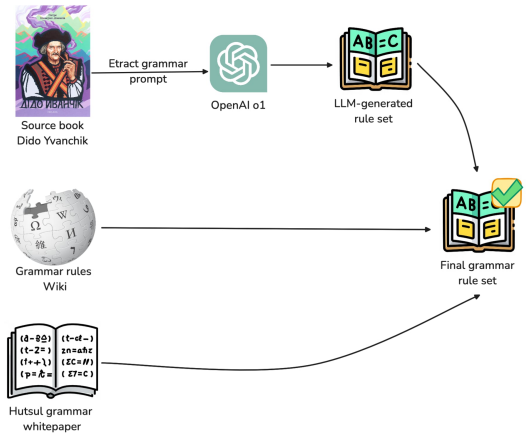

Figure 2: 基于 “ ” (Dido Yvanchik)、维基百科和 **?** 的规则生成流程概览。

尽管这种方法使我们能够显著扩大数据集，但也引入了某些限制。具体来说，合成数据反映了源语料库的词汇和主题范围，而这些语料库缺乏现代领域，如航空、科技、新闻和政治。

因此，这些领域的词汇覆盖率仍然相当稀少或缺失（即使经过生成，单词仍然与标准乌克兰语中的相同）。为了避免引入虚构的词汇，我们在生成过程中故意排除了现代新闻和基于网络的语料库。

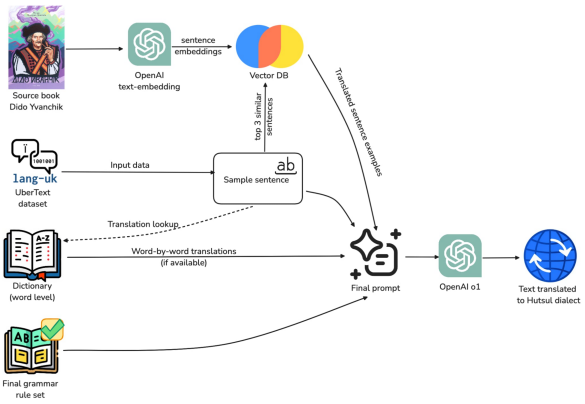

Figure 3: 合成数据生成流程概述：一个使用 " " (Dido Yvanchik) 和 UberCorpus 的 RAG 系统检索并提示 GPT-4o 生成高质量的 Hutsul-乌克兰语配对。

2.4 数据划分与可用性

最终的语料库分为 80 % 训练集，10 % 验证集和 10 % 测试集。测试集和验证集仅包含来自 " " (Dido Yvanchik) 的人工注释句子对。

3 微调

为了将大型语言模型 (LLMs) 适应乌克兰语到胡茨语的翻译任务，我们使用了参数高效的微调方法 LoRA (**?**) 。

我们对两个参数范围在 7B–13B 的最新开源模型进行了微调（考虑到我们的训练资源以及模型不应太大以便能够在本地运行）：

- Mistral-7B-Instruct v0.3 [9] ——因其性能与尺寸的比率而被选择。在许多基准测试中，它的表现优于一些较大的模型，支持多语言指令，并明确支持乌克兰语 (**?**) 。
- LLaMA-3.1 8B Instruct [10] –LLaMA 3.1 8B 的指令调优版本。该模型具有强大的多语言支持和改进的指令遵循能力，使其成为低资源翻译 (**?**) 的良好选择。

模型的选择基于以下标准：

- 分词器支持——这两个模型都使用具有回退策略的分词器来处理罕见或不在词汇表中的词汇，从而能够良好地处理基于西里尔字母的方言。
- 多语言能力——Mistral-7B-Instruct v0.3 明确列出了支持的语言，其中包括乌克兰

[9] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[10] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

语。LLaMA-3.1 8B Instruct 展现了强大的泛化能力 [11] 。

- 开放许可和可重复性 - 两个模型都在公开源代码许可下公开可用。
- 单个 GPU 的可行性 - 使用 LoRA 或 QLoRA，所有选定的模型都可以在一个不太大的 GPU 上进行微调。

我们还考虑了其他多语言模型，如 BLOOMZ（7.1B）[12] 和 NLLB-200（3.3B）[13] ，它们提供了广泛的语言覆盖。然而，这些模型要么在一般语言建模任务上表现不佳，要么在生成质量上与选择的模型相比缺乏优势。最近的基准测试表明，Mistral-7B-Instruct-v0.3 [14] 在翻译任务中与更大型的模型匹敌或超越它们，特别是在低资源和指令微调环境中 (**?**) 。

3.1 微调设置

每个模型使用 LoRa 对两种数据集变体训练了 3 个周期（完整的设置见 Guthub [15] ）：（1）一个人工创建的 Hutsul-乌克兰并行语料库，以及（2）一个包含人工和过滤的合成数据的扩展版本。

4 评估

4.1 度量标准

评估方言机器翻译并不是一项简单的任务，因为基于标准参考的指标可能会惩罚正确的词汇变化。为了确保翻译质量，我们计算了以下广泛使用的指标：

- BLEU (**?**) - 一种基于精准度的指标，用于测量假设和参考之间的 n-gram 重叠。虽然被广泛使用，但它可能会惩罚方言中常见的有效词汇和句法变化。
- chrF++ (**?**) 计算字符 n-gram F-分数，并已被证明在形态丰富和非标准语言上优于 BLEU。它对轻微的拼写或词形变化差异更具鲁棒性，使其特别适合于方言文本。
- TER（翻译编辑率）(**?**) 量化了将系统输出转换为参考所需的编辑次数。它捕捉结构性差异并惩罚重排序错误。

每个指标强调翻译质量的不同方面：

- BLEU 反映了 n-gram 精确度，
- chrF++ 捕捉形态相似性和召回率，
- TER 惩罚结构不匹配

我们将这些指标应用于手动翻译的乌克兰语-胡楚尔语句子对（1900 对）的测试集。我们并没有将它们聚合为一个单一的分数，而是共同解释这些指标以了解每个模型的不同行为方面。例如，高 chrF++ 分数伴随着低 BLEU 分数可能表明表面实现上的有效变化。

如前所述，虽然这些指标提供了一个有用的基准，但它们通常难以评估方言输出。因此，按照 **?** 的框架，我们引入 LLMs 作为评估者。

我们提示 GPT-4o 模型沿三个轴对模型输出进行评分：

- 流利性：胡茨尔方言的语法性和自然性。
- 充分性：保持原句的意义。
- 方言特质：与已知的胡楚尔语的词汇、语音和形态句法特征的一致性。

每次评估都是在零样本设置下进行的。我们曾考虑在提示中加入一些语法规则，但为了避免通过这些规则产生潜在的偏见，决定使用零样本设置。

LLM 获取源文本、模型输出和参考译文，并返回从 1（差）到 5（优）的评分。提示的结构如下：

> 您是一位语言专家，正在评估机器翻译的方言文本。请在以下维度对翻译进行评分：1. 流利度 (1–5): 输出在目标方言中语法是否正确且自然？2. 充分性 (1–5): 输出是否保留了原始来源的意义？3. 方言质量 (1–5): 输出是否反映了 Hutsul 方言预期的语音、词汇和语法特征？请以这个确切的 JSON 格式返回您的答案：
>
> { ”fluency”: x, ”adequacy”: y, ”dialect”: z } 不要解释您的评分。来源（标准乌克兰语）：<source sentence> 模型输出（Hutsul）：<model prediction> 参考（Hutsul）：<reference sentence>

由于我们没有机会进行人工评估来检验我们的翻译，并且考虑到标准的基于参考的指标可能不适合于方言翻译 (**?**) ，我们使用了基于 LLM 的充分性和方言分数作为我们的主要评估指标。这些指标与人类直觉更吻合，并且比 BLEU 或 TER 更能容忍不同的变体。

[11] https://huggingface.co/blog/akjindal53244/llama31-storm8b
[12] https://huggingface.co/bigscience/bloomz-7b1
[13] https://huggingface.co/facebook/nllb-200-3.3B
[14] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[15] https://github.com/woters/vuyko-hutsul

自动化指标被用作辅助角色，以识别诸如重排序或字符级相似性等趋势。我们并排报告所有指标。这种多指标方法使对模型行为有一个更全面的解释，尤其是在没有人工评分员的情况下。

4.2 基线

我们将微调后的模型与 GPT-4o 基线模型进行了比较。通过 OpenAI 的 API 进行查询，提示将标准乌克兰语翻译成胡楚尔方言。为了确保一致性和词汇覆盖率，我们使用了与合成生成流程中相同的 RAG 上下文和词典条目。

我们没有将未经过微调的 Mistral 或 LLaMA 模型作为基线，因为它们在方言生成任务中的性能要差得多。由于其模型较小，其指令微调不足以在弱势语言或方言中进行零样本生成。

4.3 结果

正如前面提到的，我们使用自动化评价指标和基于 LLM 的判断来评估我们的模型。表格 2 显示了 BLEU、chrF++、TER 分数，以及 GPT-4o 作为基于 LLM 的判断者，在保留的 1900 个句子的测试集中，对流畅性、充分性和方言质量评分进行了 1 到 5 级的评定。

从结果中我们可以看到，所有微调模型在每个指标上均优于 GPT-4o 基线。使用手动收集和合成数据结合微调的 Mistral 表现最佳，具有最高的 BLEU（74.35）、chrF++（81.89）和方言评分（3.60）。虽然所有模型的充分性评分保持稳定（≈ 4.7），但方言准确性变化更显著，并且证明对训练数据来源最为敏感。同时我们可以看到，LLaMA 和 Mistral 都在结合合成和手动标注数据训练时，在自动指标上表现强劲，但在方言质量上略显逊色，这突出了我们生成合成数据的方法的局限性。

下面我们展示了一个示例，该示例描述了 LLM 计算的分数在真实数据上的表现，以及相应的 BLEU、chrF++ 和 TER 指标。这表明，即使是经过小幅微调的模型在保留方言特定的意义和词汇上也稍优于零样本商业模型，但仍然远非完美。

5

限制

我们的工作在大型语言模型的乌克兰方言适应方面迈出了第一步，仍存在许多限制有待解决。

一个重要的限制是，虽然我们引入了合成数据生成管道来缓解数据可用性不足的问题，但合成翻译可能缺乏母语流利度或具有风格不一致性，尤其是对于不太常见的主题。这种情况特别容易在原始语料库未覆盖的领域中看到，例如政治、技术等领域，其中 Hutsul 词汇量要么非常有限，要么根本不存在。尽管过滤掉了低质量的生成内容，自动评估指标仍可能高估语言的有效性。

此外，评估仍然具有挑战性。自动化指标如 BLEU 和 chrF++ 通常会惩罚合理的方言变体 (**??**) 。为了更好地捕捉风格和综合的多样性，我们在基于 LLM 的评估框架最新研究的基础上使用 GPT-4o 作为基于 LLM 的评审员 (**??**) 。然而，我们注意到 GPT-4o 并没有专门为方言评估进行微调，其偏好可能仍然与标准乌克兰语一致，而人工评估将提供更为可靠的评估。

我们还需要提到，我们目前的方法是专门针对 Hutsul，这是一种乌克兰语中相对记录良好的方言。将该方法扩展到其他方言或用于其他资源匮乏的语言将需要对数据处理流程和提示策略进行适应。

我们对 伊万 · 安德鲁夏克 (Ivan Andrusiak) 表示诚挚的感谢，他提供了对 ” ” (Dido Yvanchik) 的乌克兰语翻译的访问权限，这为我们的数据集奠定了基石。我们也感谢出版社 ” ” (Dyskurs) 及其主任 ’ (Vasyl Karpiuk) 授权我们使用该文本，并持续支持语言和文化保护的倡议。他们的慷慨使这项研究成为可能。

Model	BLEU	chrF++	TER	Fluency	Adequacy	Dialect
GPT-4o	56.64	65.90	34.34	3.76	4.30	3.22
LLaMA (manual annotated + synthetic)	69.02	74.92	22.90	4.11	4.72	3.33
LLaMA (manual annotated only)	59.98	72.61	28.62	4.13	4.72	3.38
Mistral (manual annotated only)	62.36	75.65	28.62	4.14	4.74	3.35
Mistral (manual annotated + synthetic)	74.35	81.89	22.90	4.18	4.72	3.60

Table 2: 自动和基于大型语言模型（LLM）的评估结果。BLEU、chrF++ 和 TER 使用 sacreBLEU 计算。流畅性、充分性和方言质量由 GPT-4o（1-5 分制）评分。