KS-LIT-3M：用于大语言模型预训练的 310 万词克什米尔语文本数据集

Haq NawazMalik独立研究者
orcid.org/0009-0003-1994-7640
huggingface.co/Omarrran
x.com/HAQ_NAWAZ_MALIK

2026年1月6日

arXiv:2601.01091v1 [cs.CL] 3 Jan 2026

摘要

大语言模型（LLM）在高资源语言中展现出卓越的流利度，但在生成连贯的克什米尔语文本方面却始终表现不佳。造成这种性能差距的并非模型固有的局限性，而是高质量训练数据的严重匮乏。由于数十年来的克什米尔语文学作品采用专有的 InPage 桌面排版格式编码，因而无法被现代自然语言处理流程使用。本文介绍了**KS-LIT-3M**，这是一个经过整理、包含 310 万词（1640 万字符）的语料库，专为克什米尔语语言模型预训练而设计。该数据集被构建为**单一连续的线性文本流**，针对因果语言模型训练进行了优化，使模型能够根据前文上下文预测后续词元。该语料库通过开发专用的 InPage 到 Unicode 转换器构建而成，随后经过严格的预处理，包括英语内容清除、字符规范化和质量验证。KS-LIT-3M 涵盖从多种体裁中提取的 131,607 个独特词语，包括文学作品、新闻写作、学术文本和宗教研究，弥补了克什米尔语技术领域的基础资源缺口。该数据集以 CC-BY-4.0许可证发布，以促进克什米尔语自然语言处理领域的研究。

关键词 低资源　语言 · 克什米尔语 · 语言模型 · 数据集 · InPage 转换 · 文本语料库

1 引言

大语言模型的出现改变了自然语言处理，使人们能够用数十种语言进行复杂的文本生成、翻译和理解。GPT-4 [1], Claude [2], 和 LLaMA [3] 等模型在训练数据丰富的高资源语言中展现出近乎人类的流利度。然而，这种进展并未平等地惠及所有语言。对于低资源语言的使用者而言，当代人工智能系统生成的内容往往存在语法错误、语义不连贯或文化上不恰当等问题。

克什米尔语正是这种差距的典型例子。尽管其使用者约有七百万，且拥有绵延数百年的丰富文学遗产，现代语言技术对克什米尔语的支持仍然不足。若要求生成克什米尔语文本，最先进的大语言模型生成的内容往往语法错误、词汇使用不一致，变音符号位置也存在系统性错误——这些问题会使文本难以理解，甚至误导母语使用者。

这一失败的根本原因并非算法问题，而是数据问题。语言模型通过接触海量文本来获得语言能力。对于克什米尔语而言，主要大语言模型开发者能够获取的训练数据，无论在数量还是质量上都极为有限。这种匮乏有着明确的技术根源：过去二十年来最具权威性的克什米尔语文本、专业出版文学作品、新闻业资料和学术研究成果，主要以 InPage 桌面排版格式存在。InPage 是一种专有系统，其编码与现代 Unicode 标准不兼容，网络爬虫也无法发现这些内容，而网络爬虫正是用于汇编训练语料库的工具。

本文介绍 KS-LIT-3M，这是一个旨在弥补这一资源缺口的数据集。其贡献主要体现在三个方面：

1. **技术创新**：开发一种稳健的 InPage 到 Unicode 转换器，在格式转换过程中能够准确保留克什米尔语复杂的变音符号系统。2. **数据集整理**：汇编并预处理一个包含310万词的语料库，涵盖克什米尔语写作的多种体裁、时期和语域。3. **资源发布**：以开放许可公开提供该数据集，以支持克什米尔语技术领域的研究与开发。

本文其余部分安排如下。第2 节介绍克什米尔语的背景以及低资源自然语言处理面临的挑战。第3 节详细介绍数据集构建方法。第4节对语料库进行统计分析。第5 节和6 节分别讨论潜在应用和局限性。第7 节总结了对未来工作的启示。

2 背景与相关工作

2.1 克什米尔语

克什米尔语属于达尔德语支的印度-雅利安语，主要通行于印度次大陆的克什米尔谷地。这种语言具有若干独特的语言特征，使其计算处理变得复杂。

文字系统与正字法。 现代书面克什米尔语采用经过改良的波斯-阿拉伯文字，并增添字符和变音符号，以表示乌尔都语、波斯语等相关语言中没有的音位。其中包括元音标记，例如上标阿利夫（表示央元音）和“vaaw”（表示圆唇元音），以及各种上标和下标组合。关键的是，这些变音符号承载着语义区别；遗漏或错置它们会改变词义，因此，准确处理附加符号对于任何计算应用都至关重要。

语言复杂性。 克什米尔语采用动词第二位语序，具有丰富的动词一致标记和分裂式作格格系统。该语言表现出显著的形态复杂性，语法信息通过屈折变化而非仅靠句法来编码。这些特征要求统计模型拥有大量训练数据，才能可靠地加以学习。

2.2 低资源语言面临的挑战

计算语言学中的“低资源”这一称谓，指的并非语言的文化意义或使用者人数，而是适用于训练自然语言处理系统的机器可读文本的可用性 [4]。低资源语言面临一些共同的挑战：

- **数据匮乏**：采用标准编码的数字化文本不足
- **质量问题**：现有文本往往包含错误、编码问题，或受到其他语言的污染
- **领域失衡**：某些体裁（例如宗教文本）占比过高，而其他体裁则代表不足
- **评估缺口**：缺乏用于衡量系统性能的标准化基准

克什米尔语符合低资源分类的所有标准。尽管该语言甚至未被纳入多语言数据集（如 mC4 [5] 和 CC-100 [6], ），这些语料库中的克什米尔语内容在数量和质量上都极少，甚至完全没有。通过网络抓取的克什米尔语文本通常缺少正确的变音符、包含与乌尔都语或印地语的语码混用，或存在会降低训练效果的编码伪影。

2.3 InPageLegacy 格式问题

一个克什米尔语及相关波斯-阿拉伯文字语言所特有的因素，加剧了一般性的低资源挑战：InPage 软件曾被广泛用于创建文档。
InPage 于20世纪90年代开发，用于波斯-阿拉伯文字的桌面出版，后来成为克什米尔语文学、报纸、学术出版物和宗教文本排版的事实标准。与早期的 Unicode 实现相比，该软件对克什米尔语复杂排版要求的支持更为出色。

然而，InPage emp 采用了与 Unicode 不兼容的专有编码系统。在 InPage 中创建的文档：

- 无法直接复制到文本编辑器或网页浏览器中
- 不会被搜索引擎收录
- 对于编制训练数据的网络爬虫不可见
- 需要经过专用转换才能提取可用文本

因此，尽管存在大量高质量的克什米尔语文本——经过专业编辑、语法和拼写正确且使用变音符的文学作品——现代自然语言处理系统却无法获取这些资源。这解释了一个悖论：大语言模型在乌尔都语（较早完成 Unicode 转换）上表现良好，却在语言相似的克什米尔语上表现不佳。

2.4 相关数据集

已有多项工作致力于解决低资源语言的数据匮乏问题。FLORES 基准 [7] 提供了涵盖 200 种语言的评估集，但不提供预训练数据。AfroLM [8] 和 Masakhane [9] 展示了面向非洲语言的社区驱动方法。对于南亚语言，IndicNLP [10] 和 Sangraha [11] 提供了一些资源，但克什米尔语表示仍然有限。

此前没有研究专门探讨 InPage 转换难题，也没有发布过规模和质量可与 KS-LIT-3M 相媲美的专用克什米尔语预训练语料库。

3 数据集构建方法

KS-LIT-3M 的构建分为三个阶段：格式转换、来源收集和预处理。每个阶段都解决了恢复和整理高质量克什米尔语文本时遇到的特定技术与实际挑战。

3.1 InPage 到 Unicode 转换器开发

将 InPage 文档转换为 Unicode，远不止简单的字符映射那么简单。InPage 并未采用可直接转写为对应 Unicode 字符的一对一编码。相反，该软件使用专有渲染系统，将字符编码、位置信息和格式数据交错排列，其方式与 Unicode 的逻辑字符模型并不对应。

以往的转换尝试仅取得部分成功。现有工具能够妥善处理基本字符，却无法处理正确表示克什米尔语所必需的复杂变音符组合。这些工具生成的内容往往缺少元音标记、变音符位置不当，或组合字符附加错误。

转换 器通过多项技术创新解决了这些局限： s:

1. **内部表示分析**：对 InPage 的字符存储格式进行逆向工程，以了解基字符与相关变音符号之间的关系。2. **组合字符重构**：正确排列 Unicode 组合字符，准确呈现原始文档中的变音符叠加效果。3. **克什米尔语特定映射**：针对克什米尔语正字法中特有、不会出现在乌尔都语或波斯语 InPage 文档中的字符进行定制处理。4. **验证程序**：将转换后的输出与参考渲染结果进行自动比较，以识别并纠正转换错误。

完整的技术方法记录在另一篇出版物中 [12]。

3.2 源材料收集

有了功能完善的转换器后，项目便着手寻找并获取适合纳入数据集的 InPage 文档。来源筛选遵循以下几项标准：

真实性。 优先选用经过专业出版的作品，包括文学作品、新闻业和学术研究领域的作品，这些作品均经过编辑审校。此类来源能确保语法正确、拼写规范，并正确使用附加符号，而这些特点在随意写作中往往并不存在。

多样性。 该集合涵盖多种体裁，以避免领域偏差。文学作品提供了丰富的词汇和多样的文体变化。新闻文本展现了当代用法和时事词汇。学术写作体现了正式语体和技术术语。宗教文本所占比例较小，但提供了古典词汇和传统表达。

时间范围。 数据集中纳入了跨越数十年的文档，以捕捉古典和当代的使用模式，确保基于这些数据训练的模型能够识别历史形式以及现代表达。

法律与伦理考量。 来源收集遵守版权限制。材料通过合法渠道并在获得适当许可的情况下取得，用于研究用途。最终数据集依据知识共享许可（CC-BY-4.0）发布，以便广泛使用，同时保留署名要求。

3.3 预处理流程

原始转换器输出虽然优于网络抓取替代数据，但在作为训练数据使用前仍需进行系统清理。预处理流程包括四个阶段：

英语内容清除。 克什米尔语文档经常包含英语单词、短语或段落——借词、技术术语、引用材料或双语排版。对于专门面向克什米尔语的预训练数据集而言，这些英语内容会引入噪声。该流程在句子级进行语言检测，过滤被识别为英语的文本片段，同时保留合法的克什米尔语内容。

字符规范化。 Unicode 允许某些字符存在多种表示形式，而从旧式格式进行转换也可能引入不一致。该流程将所有文本规范化为 NFC（先进行规范分解，再进行规范合成）形式，确保视觉上相同的字符在数据中具有一致的表示。

空白字符标准化。 转换过程会引入不规则的间距——单词之间多余的空格、不一致的换行符或异常的段落排版。该流程按照统一规范对空白字符进行标准化：词语之间使用单个空格、采用标准化的行终止符，并统一段落边界。

质量验证。 自动检查会识别可能转换错误或存在其他质量问题的文本片段。被标记的内容将接受人工审查，有问题的片段会被修正或从最终数据集中排除。

4 数据集描述

本节介绍 KS-LIT-3M 的量化特征，并探讨使其区别于其他数据来源的定性因素。

4.1 语料库统计

表 1 总结了该数据集的主要统计特征。

表1：KS-LIT-3M语料库统计

指标	数值
字符总数	16,358,993
词语总数	3,091,180
唯一词汇	131,607
平均词长	4.29 个字符

语料库包含 310 万个词，为语言模型训练提供了大量材料。作为比较，这一规模约为莎士比亚全集的三倍。1,640 万个字符使模型能够学习字符级模式，包括变音符号的正确顺序和位置；这些对于克什米尔语正字法至关重要。

131,607 个不同词形构成了丰富的词汇多样性，不过这一数字受到克什米尔语形态丰富性的影响。同一个词元可能以多种屈折形式出现，每种形式都分别计数。这一特点反映了真实的语言结构，而非语料污染。

4.2 数据质量特征

除原始统计数据外， 还有若干定性因素使 KS-LIT-3M 有别于通过网络爬取
或其他自动化收集方式获得的数据

附加符号保留。 经过精心设计的转换流程，确保元音标记和其他变音符号元素在整个数据集中都得到正确编码。网络抓取的克什米尔语文本往往缺少变音符号，因为作者在网上随意写作时通常会省略它们。使用这类数据训练的模型永远无法学会正确的附加符号用法。KS-LIT-3M 提供一致、准确的变音符号表示，使模型能够掌握以往训练语料库中缺失的模式。

来源真实性。 这些材料取自经过专业出版、并在出版前接受编辑审校的作品。这一来源背景确保文本体现了对语法、拼写和文风的细致考究——而这些正是语言模型训练数据所期望具备的品质。

体裁分布。 该数据集涵盖多种文本类型，以确保具有代表性 ss:

- **文学作品**：诗歌、散文小说和散文，提供多样的文风和丰富的词汇
- **新闻写作**：反映当代用法的新闻文章和评论文章
- **学术文本**：展现正式语体和技术术语的学术著作
- **宗教研究**：保留传统词汇的经典文本和注释

这种多样性有助于确保基于 KS-LIT-3M 训练的模型能够掌握克什米尔语写作中各种语域和风格，而非偏向任何单一领域。

时间覆盖范围。 纳入跨越数十年的文档，意味着该数据集既展现了经典用法，也涵盖了当代用法模式。这种广泛的时间跨度有助于模型理解历史词汇和表达方式，同时保持对现代表达的熟练运用。

4.3 分发格式

为了满足为满足多样的技术要求和使用场景，KS-LIT-3M 以三种格式分发 s:

表2：可用的数据集格式

格式	描述
CSV	适用于基础分析的逗号分隔值，以及电子表格软件
XLSX	Microsoft Excel 格式，便于人工检查
JSONL	遵循以下规范的 JSON Lines 格式语言模型训练流水线

JSONL 格式将每个文档组织为单行的 JSON 对象，便于在模型训练期间进行高效的流式处理和并行处理。此格式可直接与常见的预训练框架集成。

4.4 预训练数据结构

KS-LIT-3M 专为语言模型预训练设计，其结构为**单一连续线性文本语料库**。这种架构选择体现了因果语言模型训练的要求：模型根据前文上下文学习预测后续词元。

连续文本流。不同于按离散句子对或文档分隔的集合组织的数据集，KS-LIT-3M 将文本呈现为不间断的序列。这种结构最适合：

- **因果语言建模**：连续格式使模型能够学习跨越文档边界的长程依赖关系和篇章级模式
- **高效分词**：文本可以进行分词和批处理，无需在文档边界处进行人为截断
- **上下文窗口利用**：训练样本可以充分利用可用的上下文窗口大小，而不会产生填充开销

预训练优化。 线性结构无需文档分隔符或特殊边界标记，从而降低词汇表复杂度，并使模型容量集中于学习克什米尔语模式。此设计尤其适用于：

- 从头开始训练克什米尔语专用语言模型
- 目标是提升通用克什米尔语能力的领域适配

该数据集可以直接输入标准预训练流水线（例如基于 Hugging Face Transformers 或类似框架的流水线），无需额外预处理或重构。

5 应用

KS-LIT-3M 的推出使此前因数据稀缺而受阻的多种应用成为可能 y.

5.1 语言模型训练

其主要预期应用是使用高质量克什米尔语文本预训练语言模型。该数据集支持两种训练范式：

从头预训练。 研究人员可以将 KS-LIT-3M 用作训练克什米尔语专用语言模型的基础语料库。与高资源语言可用的资源相比，该语料库规模适中，但足以用于训练较小的模型，或用于优先考虑计算效率的特定领域应用。

持续预训练。 现有多语言模型可通过在 KS-LIT-3M 上继续训练来适配克什米尔语。这种方法利用相关语言的跨语言迁移，同时强化模型的克什米尔语能力。该语料库质量很高，尤其是其准确的变音符号表示，应能帮助模型纠正从质量较低的网络抓取数据中学到的模式。

5.2 下游自然语言处理任务

除预训练外，该数据集还支持特定任务系统的开发与评估：

- **机器翻译**：该语料库为译入克什米尔语的翻译系统提供目标语言数据
- **文本分类**：带有体裁标签的子集可用于开展文档分类实验
- **命名实体识别**：新闻内容包含可供标注的命名实体
- **拼写与语法纠错**：文本经过精心整理，可为错误检测系统提供参考

5.3 语言学研究

该语料库使得以往对克什米尔语而言难以实现的大规模计算语言学分析成为可能：

- **词汇研究**：对 131,607 种不同词形进行频率分析
- **形态分析**：利用大规模数据研究屈折模式
- **历时研究**：比较语料库时间跨度内的使用模式
- **文体分析**：跨体裁和语域的定量比较

5.4 教育应用

该数据集可支持克什米尔语学习教育技术的开发，包括阅读理解系统、词汇训练工具和自动评估工具。

6 局限性与未来工作

尽管 KS-LIT-3M 推动了克什米尔语资源的发展，但仍应承认其存在若干局限。

6.1 规模限制

该语料库包含 310 万词，对于低资源语言而言规模可观，但与高资源语言可用的数十亿词相比仍显有限。仅使用这些数据训练的模型，无法达到使用规模大得多的语料库训练的模型所具备的同等语言知识深度。持续扩充克什米尔语文本资源仍是一项重要目标。

6.2 模态局限性

KS-LIT-3M 仅涵盖书面克什米尔语。口语具有鲜明的特点——音系模式、口语词汇和话语结构——这些都无法通过文本数据捕捉。未来工作的一项重要方向，是开发同样重视质量与真实性的语音语料库。这类资源将有助于构建语音识别、语音合成和口语理解系统。

6.3 评估基础设施

目前尚无用于评估克什米尔语自然语言处理系统性能的标准化基准。缺乏评估数据集，使得量化改进或客观比较系统变得困难。为克什米尔语创建基准任务——效仿 GLUE [13] 或 SuperGLUE [14] 等面向英语的模型基准——将为衡量进展提供必要的基础设施。

6.4 方言覆盖

克什米尔语在不同地理区域和社会群体中存在方言变异。目前的语料库虽然体裁多样，但可能未能充分体现所有方言变体。未来的工作应评估并酌情扩大覆盖范围，以确保开发出的系统能够服务于整个克什米尔语使用者群体。

6.5 时间偏差

由于依赖 InPage 来源，该语料库主要代表20世纪90年代至21世纪10年代的文本，当时 InPage 是占主导地位的出版工具。近年来创作的文本越来越多地采用兼容 Unicode 的格式，因此在语料库中的代表性不足。定期更新并纳入当代材料，将有助于保持语料库的相关性。

6.6 方法论推广

本项目开发的 InPage 转换方法可能适用于其他同样受到旧有格式问题影响的语言。使用波斯-阿拉伯文字的语言——包括普什图语、信德语和俾路支语——拥有大量 InPage 文献，这些文献可以从类似的恢复工作中受益。针对这些语言调整并验证转换器，是本项工作的自然延伸。

7 结论

本文介绍了 KS-LIT-3M，这是一个包含 310 万词、经过精心整理的语料库，旨在解决克什米尔语自然语言处理领域严重的资源缺口问题。该数据集通过有针对性的技术开发构建而成——开发出能够保留克什米尔语复杂变音符号系统的稳健 InPage 到 Unicode 转换器，并结合细致的整理与预处理，以确保数据质量。

当前大语言模型无法生成合格的克什米尔语文本，并非源于固有的技术局限，而是反映了一个具体且可以解决的问题：训练语料库中缺乏高质量的克什米尔语数据。这种缺失在很大程度上源于 InPage 这一传统格式，它使克什米尔语文学最丰富的来源无法被现代人工智能系统识别。KS-LIT-3M 正是对这一状况的直接回应。

与其他替代方案相比，该语料库具有多项优势：保留网络抓取文本中缺失的准确附加符号，来源于经过专业编辑的出版物，确保来源真实性；涵盖文学、新闻、学术和宗教领域，体裁多样；还横跨数十年，囊括不同时期的克什米尔语写作。

该数据集以开放许可公开后，研究人员和开发人员如今可以预训练和微调语言模型，预期这些模型在克什米尔语的流利度、语法准确性和附加符号处理方面都将显著提升。除了模型训练，该语料库还使得以往对这种语言而言难以实现的大规模计算语言学研究成为可能。

此处展示的方法——专门的格式转换、审慎的来源收集和严格的预处理——可为其他低资源语言的类似工作提供范本，尤其是那些受旧格式挑战影响的语言。每一种从低资源走向资源充足的语言，都有力证明了语言多样性与技术进步能够齐头并进。

克什米尔语使用者理应拥有能够理解并尊重其语言的人工智能系统。实现这一目标需要基础资源，而 KS-LIT-3M 正是迈向符合母语使用者期待的克什米尔语技术的重要一步。

伦理声明

数据收集。 KS-LIT-3M 中包含的所有源材料均通过合法渠道获取，并获得了用于研究用途的适当许可。该数据集不包含个人信息、私人通信或可能识别个人身份的内容。

许可。 该数据集依据知识共享署名 4.0 国际许可（CC-BY-4.0）发布，允许在适当署名的前提下进行分享和改编。使用该数据集时，用户应引用本文。

预期用途。 KS-LIT-3M 旨在用于自然语言处理领域的研究与开发。数据集维护者不支持可能伤害克什米尔语社群或歪曲该语言的应用。鼓励用户让母语使用者参与使用这些数据训练的系统的评估与部署。

潜在风险。 与任何语言模型训练数据一样，该语料库可能包含反映其源材料观点的偏见。在敏感应用中部署之前，应对使用这些数据训练的系统进行潜在偏见评估。

数据可用性

KS-LIT-3M 数据集可通过 Hugging Face 数据集中心获取：

https://huggingface.co/datasets/Omarrran/3.1Million_克什米尔语_文本_预_训练_数据集_for_大语言模型_2026_by_HNM

该数据集根据 CC-BY-4.0许可证 发布。

致谢

作者感谢克什米尔文学界保存并分享源材料，使该数据集得以创建。

参考文献

[1] OpenAI。《GPT-4 技术报告》。*arXiv* 预印本 *arXiv:2303.08774*，2023。[2] Anthropic。《Claude 3 模型家族：智能的新标准》。*Anthropic* 技术报告，2024。

[3] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。《LLaMA：开放且高效的基础语言模型》。arXiv 预印本 arXiv:2302.13971，2023。[4] Pratik Joshi、Sebastin Santy、Amar Buber、Kalika Bali 和 Monojit Choudhury。《自然语言处理领域语言多样性与包容性的现状与未来》。载于《计算语言学协会第58届年会论文集》，第6282–6293页，2020。[5] Linting Xue、Noah Constant、Adam Roberts、Mihir Kale、Rami Al-Rfou、Aditya Siddhant、Aditya Barua 和 Colin Raffel。《mT5：大规模多语言预训练文本到文本转换器》。载于《计算语言学协会北美分会2021年会议论文集》，第483–498页，2021。[6] Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。《大规模无监督跨语言表征学习》。载于《计算语言学协会第58届年会论文集》，第8440–8451页，2020。[7] Naman Goyal、Cynthia Gao、Vishrav Chaudhary、Peng-Jen Chen、Guillaume Wenzek、Da Ju、Sanjana Krishnan、Marc'Aurelio Ranzato、Francisco Guzmán 和 Angela Fan。《面向低资源和多语言机器翻译的 FLORES-101 评测基准》。《计算语言学协会汇刊》，10：522–538，2022。[8] Bonaventure F P Dossou、Atnafu Lambebo Tonja、Orevaoghene Ahia、Naomi Feldman、Arya Kasabi、Anusha Mathur、Colin Leong、Zhi Xuan Tan、Shruti Lohani 等。《AfroLM：一种基于自主动学习、面向23种非洲语言的多语言预训练语言模型》。arXiv 预印本 arXiv:2211.03263，2022。[9] Wilhelmina Nekoto、Vukosi Marivate、Tshinondiwa Matsila、Timi Fasubaa、Tajudeen Kolawole、Taiwo Fagbo-hungbe、Solomon Oluwole Akinola、Shamsuddeen Muhammad 等。《低资源机器翻译的参与式研究：非洲语言案例研究》。载于《计算语言学协会论文集：EMNLP 2020》，第2144–2160页，2020。[10] Anoop Kunchukuttan。《IndicNLP 库》。载于《印度语言处理》，2020。[11] Mohammed Safi Ur Rahman Khan、Erik Johannes Husom 等。《Sangraha：面向印度语言的大规模多语言数据集》。arXiv 预印本，2024。[12] Haq Nawaz Malik。《面向克什米尔语及相关语言的稳健 InPage 到 Unicode 编码转换器：解决一项持续15年的难题》。ResearchGate 出版物，2024。https://www.researchgate.net/publication/387522744。[13] Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R Bowman。《GLUE：自然语言理解的多任务基准与分析平台》。载于《国际学习表征会议》，2019。[14] Alex Wang、Yada Pruksachatkun、Nikita Nangia、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R Bowman。《SuperGLUE：一个更具挑战性的通用语言理解系统基准》。载于《神经信息处理系统进展》，第32卷，2019。