本地大语言模型生成句作为教材补句的可用性：与教材原句的配对比较

胡浪多

广州航海学院外国语学院　广东广州

【摘要】本研究考察本地大语言模型生成句在教材补句任务中的适用性。从五本英语教材中选取 1,969 个含情态词的句子，在保留前后文的条件下，由三个模型生成替代句，并比较生成句与教材原句的语言特征和语境适配表现。5,907 个最终保留的生成句中，5,864 句满足任务要求。相较于教材原句，Mistral 与 Gemma 的句子较短，Aya 的句子较长。三个模型的篇章连贯和语义兼容评分与教材原句接近，多数配对评分相同，但不同评价模型对同一句子的判断存在分歧。结果表明，本地模型能够提供具有一定语言质量与语境适配性的补句候选，其适用性取决于模型的表达特点及生成句与前后文的联系。

【关键词】教材补句；本地大语言模型；语境适配；语言质量；自动评价

【基金项目】2025 年“外语教育 AI 赋能与创新发展”专项课题（GDWYJZW-2025-2）

【收稿日期】2026 年 9 月 12 日　【出刊日期】2026 年 10 月 13 日　【DOI】10.12208/j.ije.20260138

Availability of sentences generated by local large language models as supplementary sentences for textbooks: A paired comparison with original textbook sentences

Langduo Hu

School of Foreign Languages, Guangzhou Maritime University, Guangzhou, Guangdong

【**Abstract**】This study examines the applicability of sentences generated by local large language models for the task of textbook supplementary sentence creation. A total of 1,969 modal verb-containing sentences were selected from five English textbooks. With the preceding and following context preserved, three models generated alternative sentences, and the linguistic features and contextual fitness of the generated sentences were compared against the original textbook sentences. Among the 5,907 retained generated sentences, 5,864 met the task requirements. Compared with the original textbook sentences, sentences produced by Mistral and Gemma are shorter, while those from Aya are longer. Scores of discourse coherence and semantic compatibility of the three models are close to those of the original textbook sentences, with most pairs receiving identical scores. However, evaluator models disagreed on judgements for the same sentences. The results show that local models can supply supplementary sentence candidates with acceptable linguistic quality and contextual fitness, and their applicability depends on the expressive characteristics of the models and the links between generated sentences and their surrounding context.

【**keywords**】Textbook supplementary sentences; Local large language models; Contextual fitness; Linguistic quality; Automatic evaluation

1 引言

教材补句既要符合语言形式要求，也要与局部语境衔接。语法通顺的句子若改变人物、事件或指代关系，仍可能不适合原有位置；措辞不同的句子也可能承担相同功能。因此，评价模型补句需兼顾语言特征与语境表现。

整句替换允许模型重新组织句内表达，但前后文限定其语篇功能。教材原句既是语言特征的比较对象，也是理解该位置功能的参照。本文据此考察生成句能否在既有语境中承担原句的局部语篇功能。

本文从文本层面考察补句可用性，关注生成句能否适合教材中的特定位置，具体涉及表达是否规范、内容是否与语境相容，以及能否为后续叙述提供必要信息。语言形式与语篇功能在这一任务中相互关联。同一事实可以采用不同词语或句法结构表达，人物关系、事件推进和指代对象的变化却会影响相邻句子的解释。

因此，对句子使用价值的判断需置于原段落中进行。

人类文本与模型文本的语言差异已受到关注。Guo 等[1]从词汇、句法和语义层面考察模型输出的语言多样性，强调对语言形式与意义表达进行多方面比较。Zanotto 和 Aroyehun[2]结合语言特征与向量表示，描述不同领域中人类文本与模型文本的特点。这些研究为分析生成文本提供了参照。教材补句将比较落实到同一语篇位置，原句与模型句共享邻近语境，其表达差异也就能够联系具体的使用要求加以解释。

Tevet 和 Berant[3]区分形式多样性与内容多样性，说明不同指标所反映的文本属性有所不同。这一区分有助于解释补句中的表达变化。词汇或结构不同的句子可以承担相近的语篇功能，词形相同也可能伴随事件含义的改变。情态词表达可能性、义务或能力，其意义依赖具体语境[4]。要求模型保留指定情态词，可以在固定形式条件下观察句子如何组织相关内容。本文据此考察生成句的语言特征及其与教材上下文的联系。

在文本质量评价方面，Liu 等[5]探索了依据明确标准进行大语言模型评分的方式，Pan 等[6]进一步讨论了评价标准与实际使用意图的对应关系。教材补句具有明确的局部用途，评分结果因而需要结合前后文解释。本研究采用已有语言分析工具、句向量模型和模型评分思路，比较本地模型在这一任务中的表现，重点考察候选句的语言特点及其使用条件。

本文围绕两个问题展开：RQ1，生成句与教材原句在语言形式和表达上有何差异；RQ2，生成句在教材补句任务中的语言质量与语境适配表现如何。

2 数据与方法

研究以教材中的目标句位置为配对单位，比较教材原句与相同语境下的模型补句。原句呈现该位置已有的表达，模型句则提供相同邻近语境约束下的其他表达。分析涵盖语言特征、原句相似度和结合语境的质量评分，分别描述句子的表达方式、与原句的内容联系，以及置于原位置时的适配表现。

2.1 语料与生成

语料选自 BCcampus 出版的五本开放许可英语教材，共含 1,969 个目标句，分布见表 1。筛选时排除练习、答案和重复内容，保留含指定情态词的完整句子，并要求目标句在同段内至少有一侧相邻句。生成时遮蔽目标句，提供前后各至多两句作为语境。三个模型使用相同的任务指令和各自的原生提示模板。

这一设置保留了原句前后的语篇条件，并允许模型重新组织句内表达。教材原句参与全部组间比较，且与生成句接受相同的语境评分。因此，语言特征和质量评分均以相同教材位置为参照。比较单位是具体语境中的句子，教材来源用于说明样本分布及任务背景。

表 1　教材语料构成

教材	项目数
Advanced English	628
Writing for Success（加拿大第一版）	1,103
BC Reads Reader 6	111
BC Reads Reader 5	83
BC Reads Reader 4	44
合计	1,969

指定情态词须以完整词形出现，否定缩约形式不纳入分析。生成指令要求仅输出一个语法自然、与前后文连贯且包含指定情态词的句子，不附解释、备选句或上下文引文。

生成模型包括 Mistral-Small-3.2-24B-Instruct-2506、Gemma-3-27B-it 和 Aya-Expanse-32B，均为本地 Q4_K_M 版本。解码参数为 temperature=0、top_p=1、top_k=1、seed=20260911，上下文及输出上限分别为 4096 和 192 tokens。每个模型每项保留一个最终句子，格式不合要求时最多重试一次。

2.2 自动评价与统计

语言特征分析描述词汇与句法形式，原句相似度反映改写幅度，语境评分考察候选句与相邻文本的衔接。三类指标分别对应表达方式、内容联系和局部适配。

使用 spaCy 3.8.11 及 en_core_web_sm 3.8.0 提取语言特征。句长不计空白和标点；TTR 为忽略大小写后的不同词形数与词元数之比。最大依存深度按词的祖先节点数计算。从属小句近似数统计 advcl、ccomp、xcomp、acl、relcl、csubj 和 csubjpass；总小句近似数另计 ROOT 及词性为 VERB 或 AUX 的 conj 节点。上下文词汇重合采用候选句与前后文词元集合的 Jaccard 系数，词元保留标点并作 Unicode 规范化和大小写折叠。

句向量采用 paraphrase-multilingual-MiniLM-L12-v2[7-9]，编码上限 512，经 384 维均值池化和 L2 归一化后计算生成句与原句的余弦。编辑距离采用 Levenshtein 距离，文本经 Unicode 规范化、大小写与空白折叠后，再除以两句中较长的字符数。

参照大语言模型 LLM（Large Language Model）的文本评价研究[5–6]，采用 Qwen3.5-9B 评价隐去来源并

随机排序的句子，输入仅含候选句、前后文和指定情态词及其位置。语法评价句内规范性，自然度评价地道程度，篇章连贯评价语篇连续性，语义兼容评价与上下文的相容性，情态适切评价首个指定情态词的用法。各维度采用五级评分。1 分为严重问题或不适配，2 分为明显问题，3 分为部分符合或存在不确定性，4 分为仅有轻微问题，5 分为充分符合。评分采用 temperature=0、seed=20260911。GLM-4-9B-0414 评价按教材和情态词分层抽取的 215 个有效项目，用于比较评分一致性。

两幅主图采用四组共同有效的 1,926 个项目，报告语言特征均值及评分分布，教材原句每项仅计一次。原句相似度采用各模型全部有效配对，Mistral、Gemma、Aya 依次为 1,960、1,948、1,956 项。

配对评分效应为生成句得分高于原句的比例减去低于原句的比例，同级评分计入分母。均值与比例用于描述样本表现。评价模型间的一致性采用二次加权 Cohen κ。

评分分布报告各等级的句子比例，均值概括样本的总体评分水平。配对比较对应同一教材位置，区分生成句得分高于、等于或低于原句的情形。两个评价模型的比较分别考察组间趋势与单句判断。组间趋势反映两类句子在样本中的相对表现，单句一致性反映评价模型对具体句子的判断是否相同。

3 结果

表 2 展示同一位置的教材原句与模型补句，图 1 和图 2 分别呈现共同样本的语言特征与语境评分，从实例和总体分布说明补句与上下文的关系。

3.1 教材补句的语境衔接例析

表 2 以一个行动衔接明确的叙事段落为例，展示教材语境、原句与三个模型补句。

表 2 同一教材位置的原句与模型补句（指定词：would）

位置/来源	文本
前文	That’s when the manager came.
教材原句	He told her to move to the balcony, or he would call the police.
Mistral	He asked her if she would mind moving, but she shook her head no.
Gemma	He would have to go around her, then.
Aya	The manager asked Viola if she would be willing to leave, but she refused.
后文	Viola told him to go ahead. She wasn’t moving.

来源：Shantel Ivits(2015), BC Reads Reader 4, The Story of Viola Desmond, CC BY 4.0.

原句中的“call the police”为后文“go ahead”提供行动指向；Mistral 与 Aya 改为询问是否愿意移动，Gemma 则写成绕过她。四句虽均含 would，后文所承接的行动却发生变化，说明补句意义必须结合上下文解释。

Mistral 和 Aya 还提前写出拒绝移动，与后文“She wasn’t moving”形成重复。教材原句先提出可能行动，再由后文呈现回应，模型补句因而改变了信息在相邻句之间的分配。

这些差异还体现了说话行为的组织方式：原句连接移动要求与可能行动，Mistral 和 Aya 询问意愿并写出拒绝，Gemma 描述绕行的必要性。指定词形相同并不意味着语篇功能相同。

3.2 教材原句与生成句的语言特征差异

三个模型共生成 5,907 句，其中 5,864 句符合格式和指定词形要求，成功率为 99.27%。Mistral、Gemma 和 Aya 不符合要求的句子分别为 9、21、13 句。

图 1 显示，教材原句、Mistral、Gemma 和 Aya 的平均词元数分别为 19.25、13.71、13.87 和 21.34。Mistral 与 Gemma 句子较短、依存层级较浅、小句较少；Aya 则较长且结构较复杂。词汇丰富度与上下文词汇重合也随模型变化。

句法差异同时体现于依存层级和小句组织。Gemma 的最大依存深度和从属小句近似数更低；Aya 的句长、最大依存深度及小句近似数均高于原句。Gemma 的 TTR 最高而上下文词元重合最低，Aya 的 TTR 接近原句且词元重合最高，Mistral 的词元重合与原句相近。

三个模型与原句的语义余弦均值分别为 0.477、0.447 和 0.503，共同有效子集的排序一致。有效生成句均未逐字复现原句，平均归一化编辑距离分别为 0.716、0.729 和 0.727。

Aya 与原句的语义余弦最高，Mistral 的编辑距离最小，说明语义表示与字符相似性呈现不同排序。三个模型均提供了不同于原句的表达。

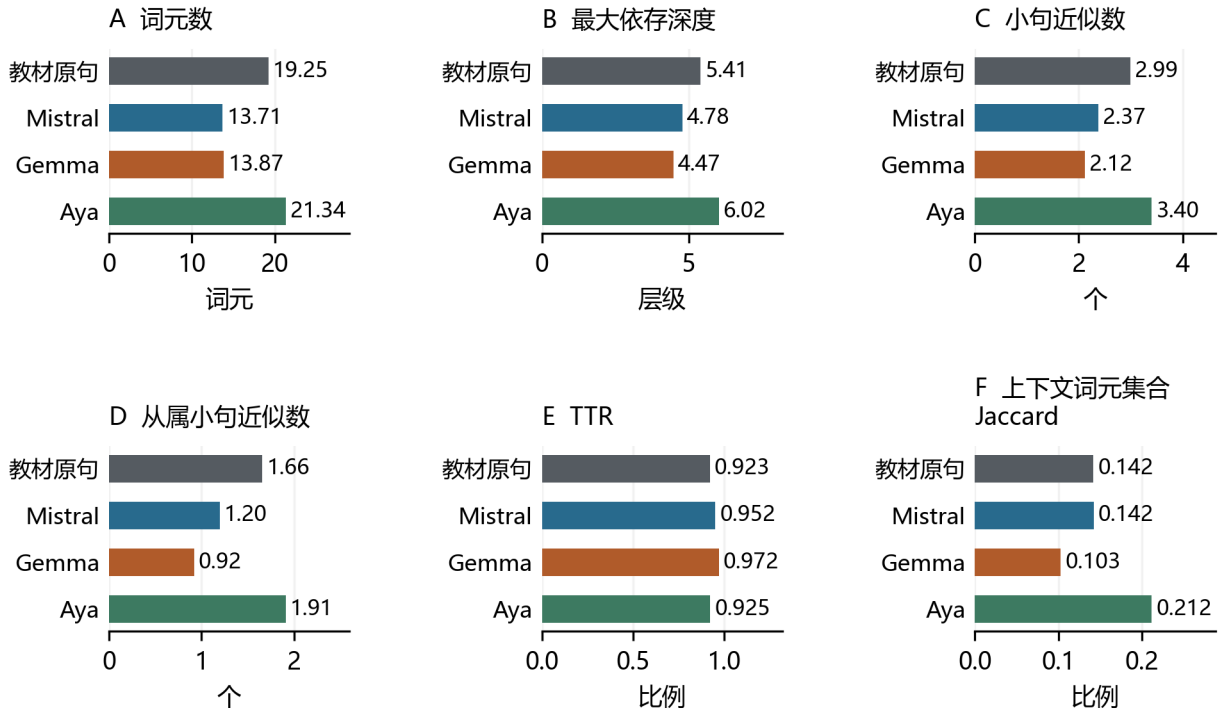


图 1　教材原句与三个模型句的语言特征均值

四组各含 1,926 个相同位置的句子；柱长及标注表示均值，各面板使用相应指标的原单位。

3.3 生成句的语言质量与语境适配

图 2 显示，原句、Mistral、Gemma 和 Aya 的篇章连贯均值分别为 4.24、4.31、4.35 和 4.31，语义兼容均值分别为 4.32、4.35、4.42 和 4.32。评分总体集中在量表高端，多数模型句与原句获得相同等级，但自然度的组间差异较明显。

原句的自然度均值为 4.48，三个模型分别为 4.85、4.89 和 4.76；四组语法均值分别为 4.76、4.90、4.92 和 4.89。模型句在两维度上的评分较高，而 Aya 的语义兼容均值与原句同为 4.32，表明组间差异随评价维度变化。

四组语法评分均以最高等级为主，模型句的自然度评分也较集中。篇章连贯和语义兼容评分相对分散，原句与模型句均有较低等级。三个模型的情态适切均值均高于原句，其中 Gemma 最高。

在两个评价模型均有评分的样本中，Mistral、Gemma 和 Aya 与原句的有效配对数分别为 214、212 和 213。15 项配对效应均为正，但 Qwen 与 GLM 对单句评分的一致性较低，二次加权 κ 为 0.096-0.293。

两个评价模型判断的总体方向一致，但对具体句子的评分存在明显分歧。

4 讨论

生成句的使用价值取决于语言表达与语境联系。三个模型大多能提供符合形式要求的候选句，但语言指标和评分维度呈现不同差异，为讨论采用条件提供了依据。

4.1 生成句的教材补句适用性

三个本地模型在绝大多数语境中生成了符合形式要求的补句。自动评分显示，其篇章连贯和语义兼容总体接近教材原句，支持将模型用于生成候选句，并依据具体语境进行筛选。

RQ2 所涉及的可用性需结合任务遵循与语境表现理解：生成成功率说明模型通常满足单句和指定词形要求，语境评分则关系到句子能否适合具体位置。两类证据共同构成候选句选择的依据。

自然度差异大于篇章连贯和语义兼容差异，说明流畅性不能替代内容衔接判断。表 2 中的模型句改变了后文承接的行动，也表明选择时应检查事件关系与表达意图。

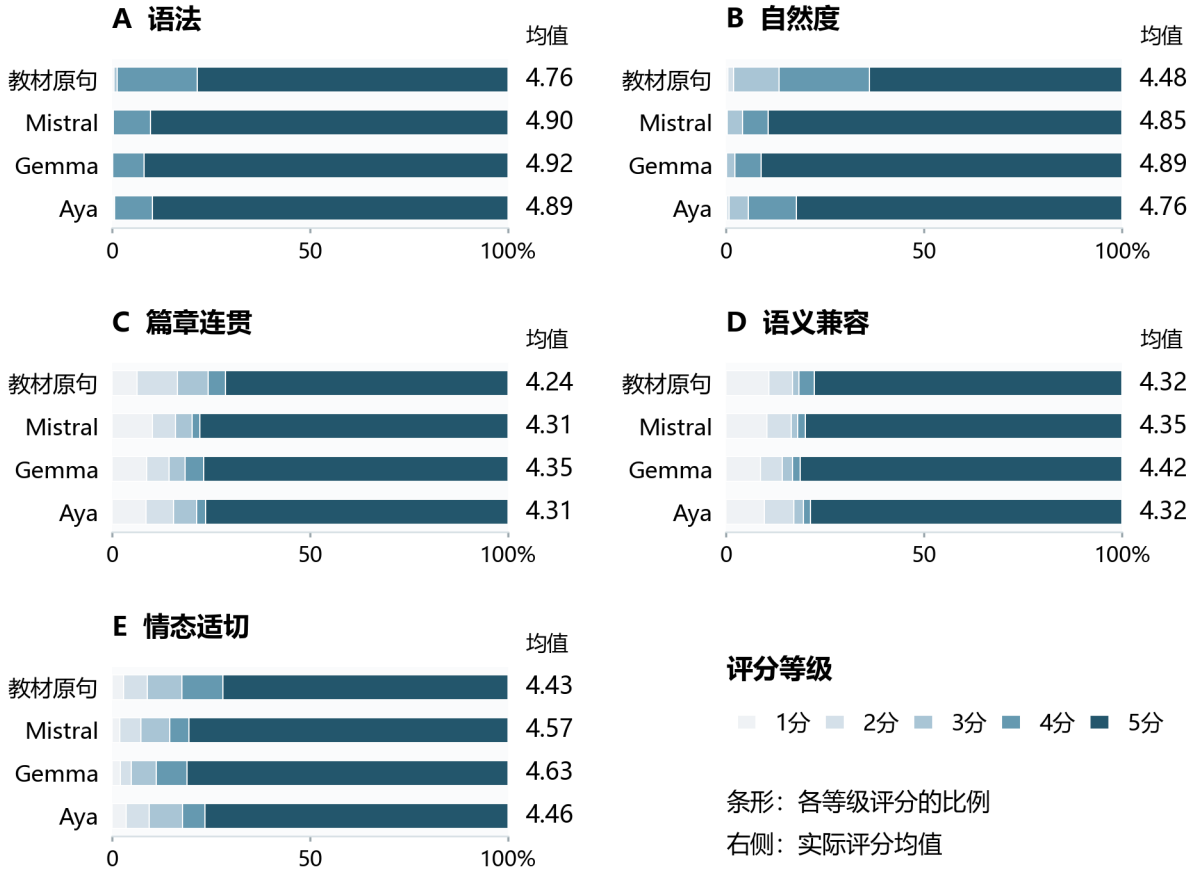


图 2　教材原句与三个模型句的五维评分分布及均值

四组各含 1,926 句，评分范围为 1-5 分。条形显示各评分等级的比例，右侧标注实际均值；颜色越深，评分等级越高。

教材原句呈现特定位置已有的表达，并与生成句接受相同评分。将其作为参照，可以识别各维度的相近表现和具体差异，但原句得分同样由评价标准决定，不能视为绝对上限。

补句候选的价值在于提供表达选择，但必须保持事件联系，并为后续指代或回应提供可理解的对象；否则即使单句通顺，也可能需要调整上下文。

若教材活动要求保留后文对某项行动的回应，候选句就需提供可供承接的内容；若允许重组整段，信息安排则有更大空间。本文固定前后文，因此表达流畅的句子仍须与既有事件关系和信息次序协调。

4.2 语言差异与补句选择

对于 RQ1，模型间差异明显：Mistral 与 Gemma 倾向于较短、结构较浅的表达，Aya 倾向于较长、较复杂的表达。已有跨领域研究也发现人类与模型文本在句法和语义层面存在差异[2]。本文进一步表明，即使语境和词形要求相同，不同模型相对原句的结构变化方向仍不一致。

较短的句子可能降低结构复杂度，也可能压缩信息；较长的句子可能增加解释或限定并改变段落节奏。句长与复杂度本身没有固定优劣，仍须结合语篇功能判断。

词汇指标也需联系句长和语境解释。Gemma 句短而 TTR 较高；Aya 与上下文词元重合较高，既可能体现词汇联系，也可能造成信息重复。因此，词汇和句法指标不能脱离具体表达使用。

三个模型均产生了不同于原句的表达，其中 Aya 的语义余弦较高。同一位置可以容纳多种合理表达，原句相似度必须结合语境适配解释。

形式与内容多样性的区分[3]有助于解释这些结果。相似度较高的补句需考察其保留的内容，差异较大的补句则需判断能否以其他表达承担相同功能。指标用

于描述候选句，最终采用仍取决于教材位置和模型表达特点。

5 研究局限

本研究依赖自动语言分析和 LLM 评分，未考察课堂实施与学习效果。Qwen 与 GLM 的单句评分一致性较低，且大量同级评分和高端集中限制了质量差异的识别，因此评分接近不能解释为两类文本等效，也不能据此确定可直接采用的比例。

模型评价可能偏向模型生成文本[5]，这可部分解释模型句较高的自然度评分。由于缺少教师或学习者判断，高分在多大程度上对应教材可接受性仍待检验。后续研究可在课程任务中考察候选句的选择、理解和使用。

语料来自单一平台的五本教材，分布不均且同段句子存在依赖，结论的体裁推广范围有限。教材可能进入模型预训练，作者也未执行相同提示任务。每项每模型仅保留一个确定性输出，不能推断重复生成的稳定性。模型家族、规模和训练过程同时变化，差异难以归因于单一机制；自动句法规则及句长耦合也限制了独立句法效应的解释。shall 无合格项目，低频情态词仅作描述。

研究数据与代码尚未公开，分析与写作使用了 AI 辅助。

6 结论

在指定情态词与教材语境约束下，本地大语言模型能够生成符合形式要求且具有一定语言质量和语境适配性的补句。Mistral 与 Gemma 倾向于简短表达，Aya 较长且结构更复杂。模型可用于提供教材补句候选，但采用时仍须依据局部语境、事件联系和表达需要进行筛选。

参考文献

[1] Guo Y, et al. Benchmarking Linguistic Diversity of Large Language Models[J]. TACL, 2025, 13:1507–1526.

[2] Zanotto SE, Aroyehun S. Linguistic and Embedding-Based Profiling of Texts Generated by Humans and Large Language Models[J]. EMNLP, 2025:22841–22858.

[3] Tevet G, Berant J. Evaluating the Evaluation of Diversity in Natural Language Generation[C]. EACL, 2021:326–346.

[4] Marasović A, et al. Modal Sense Classification At Large: Paraphrase-Driven Sense Projection, Semantically Enriched Classification Models and Cross-Genre Evaluations[J]. LiLT, 2016, 14, article 3:1–58.

[5] Liu Y, et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment[C]. EMNLP, 2023:2511–2522.

[6] Pan Q, et al. Human-Centered Design Recommendations for LLM-as-a-Judge[C]. HuCLLM, 2024:16–29.

[7] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C]. EMNLP-IJCNLP, 2019:3982–3992.

[8] Reimers N, Gurevych I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation[C]. EMNLP, 2020:4512–4525.

[9] Sentence Transformers. paraphrase-multilingual-MiniLM-L12-v2[EB/OL]. Hugging Face 模型卡,[2026-09-12].

版权声明：©2026 作者与开放获取期刊研究中心（OAJRC）所有。本文章按照知识共享署名许可条款发表。
http://creativecommons.org/licenses/by/4.0/

OPEN ACCESS