濒危语言记录中的形态 gloss 标注混合神经-大语言模型流水线：准噶尔图瓦语案例研究

Siyu Liang[1], **Talant Mawkanuli**[2],**Gina-AnneLevow**[11] 语言学系，华盛顿大学[2] 中东语言与文化系，华盛顿大学 {liangsy, tmawkan, levow}@uw.edu

arXiv:2603.00923v1 [cs.CL] 1 Mar 2026

摘要

行间注释文本（交错式语素 gloss）的创建仍是语言记录和田野调查中的一大瓶颈，对于低资源形态丰富语言尤其如此。我们提出一种混合自动 gloss 标注流水线，将神经序列标注与大语言模型后校正相结合，并在低资源突厥语准噶尔图瓦语上进行评估。通过系统的消融研究，我们发现，检索增强提示相比随机选择示例能带来显著提升。我们还发现，与完全不提供词典相比，语素词典在大多数情况下反而会降低性能；而性能大致随少样本示例数量呈对数增长。最重要的是，我们将双向长短期记忆网络-条件随机场模型与大语言模型后校正相结合，构建了两阶段流水线，该流水线在大多数模型上都取得了显著提升，并有效减少了标注工作量。基于这些发现，我们提出了将结构化预测模型与大语言模型推理整合到形态复杂的田野调查语境中的具体设计原则。这些原则表明，混合架构有望成为濒危语言记录中实现自动语言标注的一种计算开销较低的解决方案。

1 引言

行间注释文本（交错式语素 gloss，IGT）对于语言记录与保护至关重要，它将语言转写与语素切分、词义标注和译文对齐（Lehmann，2004a）。尽管 IGT 十分重要，其创建过程仍然高度依赖人力，导致语言记录项目出现瓶颈（Chelliah andReuse，2010年）。神经序列模型和大语言模型的最新进展为自动生成 IGT 带来了新的可能性，但每种方法都存在局限：结构化模型缺乏灵活性和世界知识，而大语言模型难以保持一致性，且需要大量上下文示例或成本高昂的微调，这些条件在大多数实际应用场景中都难以满足。

我们提出了一种混合流程，将用于初始注释预测的 BiLSTM-CRF 模型与基于大语言模型的后校正相结合，并在准噶尔图瓦语（下文简称图瓦语）上进行评估。图瓦语是一种形态复杂的突厥语族语言。通过针对四种大语言模型和多种设计选择开展系统实验，我们证明，这种两阶段方法对大多数模型的效果都显著优于 BiLSTM基线。消融研究揭示了若干关键设计原则：检索增强提示显著优于随机选择示例；语素词典通常会降低大多数模型的性能；最优少样本参数范围为五至十五个示例。

我们的贡献如下：（1）提出一种将结构化预测与大语言模型推理相结合、用于自动词义标注的混合架构；（2）开展全面的消融研究，确立有关检索策略、术语表配置和少样本扩展的设计原则；（3）比较评估模型在生成任务与纠错任务中的表现，为田野调查场景提供基于证据的指导。

2 相关工作

2.1 交错式语素 gloss 与语言记录

交错式语素 gloss 是田野语言学中的一种标准格式，用于编码源语言转写、形态切分、注释标签和自由译文（Lehmann，2004b；Comrie 等人，2008年）。对于许多濒危语言和低资源语言而言，交错式语素 gloss 是其主要的数字化文献记录（Chelliah 和 Reuse，2010年；Hargus 等人，2020年）。

近期研究开发了从语法描写中提取交错式语素 gloss 的工具（Schenner 和 Nordhoff，2016年；Round 等人，2020年；Nordhoff 和 Krämer，2022），也开发了从语音生成多模态交错式语素 gloss 的工具（He 等人，2024）。SIGMORPHON 2023 年自动生成交错式语素 gloss 共享任务（Ginn 等人，2023）进一步激发了人们对 gloss 标注计算方法的兴趣，并促成了后续的大规模交错式语素 gloss 建模与评估研究，其中包括用于 gloss 标注的预训练语言模型（Ginn 等人，2024b）以及用于低资源交错式语素 gloss 生成的基于大语言模型的提示方法（Elsner 和 Liu，2025年）。尽管这些工作在基准数据集上取得了稳步进展，但尚未取代现有的田野工作实践；在大多数语言记录项目中，交错式语素 gloss 的创建仍主要依靠人工完成，使用的工具包括 ELAN 和 FLEx（Wittenburg 等人，2006；International，2025年）。传统工作流程还包括基于规则的形态分析器，以及在 FLEx 等工具中进行确定性词典查找，这些功能支持从文本中半自动进行 gloss 标注和词库构建；我们的方法旨在补充而非取代这些方法。

2.2 自动形态分析与 gloss 标注

自动 gloss 标注的传统方法采用结构化预测模型。序列到序列架构已应用于形态切分（Ruzsics 和 Samardži´c，2017年；Liu 等人，2021；Rice 等人，2024），而基于条件随机场的模型已被证明在形态标注方面效果显著（Buys 和 Botha，2016年；Malaviya 等人，2018）。尤其是双向长短期记忆网络-条件随机场架构，兼顾了局部模式识别与全局约束（Ma 和 Hovy，2016年；Cotterell 和 Heigold，2017年），在形态丰富语言的序列标注任务中取得了出色表现。

近期专门针对交错式语素 gloss 生成的研究，探索了使用翻译数据的神经编码器—解码器模型（Zhao 等人，2020年）、适用于低资源场景的条件随机场方法（Bar-riga Martínez 等人，2021；Okabe 和 Yvon，2023），以及采用结构化语言学表示的轻量级模型（Shandilya 和 Palmer，2023）。Moeller 等人 (2020年) 展示了交错式语素 gloss 如何支持下游形态分析任务。然而，这些模型需要大量带标注的语料，并且难以处理罕见语素和新组合。Rice 等人 (2025年) 的近期研究也指出，计算形态学研究成果与现实语言记录需求之间存在显著差距，凸显了以用户为中心的设计的重要性。

2.3 用于语言学标注的大语言模型

大语言模型在低资源环境下的语言学标注任务中展现出应用前景。近期研究（Ginn 等人，2024a；Elsner 和 Liu，2025年）探讨了基于大语言模型的 gloss 预测以及面向交错式语素 gloss 的提示策略，并表明提示设计和示例选择会显著影响性能。Zhang 等人(2024)表明，提供词典和语法概要能够实现对未见语言的翻译。Yang 等人(2025b)使用参考语法和交错式语素 gloss 评估模型的元语言推理能力，不过其基准测试依赖整理后的参考语法数据，未涵盖完整的田野调查语境。大语言模型的后纠正和优化步骤也在各种自然语言处理任务中得到广泛探索，作为级联处理或后编辑阶段（Zouhar 等人，2021；Izacard 等人，2023）。

在近期研究中，少样本提示已成为使大语言模型适应专业任务的关键技术。研究表明，谨慎选择和呈现上下文示例会显著影响性能（Logan IV 等人，2022；Winata 等人，2021），基于检索的示例选择通常优于随机选择（Stahl 等人，2024）。然而，大语言模型在低资源环境中面临诸多挑战：它们需要大量上下文示例（从而增加推理成本），难以保持范式一致性，并且缺乏结构化序列模型所具备的归纳偏置。

2.4 混合式与多阶段架构

结合多种模型类型的混合方法已被证明在各类自然语言处理任务中行之有效。检索增强生成（RAG）通过动态引入相关示例来增强大语言模型（Lewis 等人，2021；Jiang 等人，2023），近期研究则探索了用于领域适应的专用 RAG 架构（Siriwardhana 等人，2023；Yu，2022）。级联架构利用专用模型处理不同的子任务（Izacard 等人，2023），而借助外部知识优化输出的后处理步骤也持续带来显著提升（Zouhar 等人，2021）。我们的工作将这些思路拓展至形态学标注，提出了一种两阶段流水线，其中双向长短期记忆网络-

条件随机场提供初始结构，大语言模型则通过语境推理和一致性检查来优化预测。

3 数据

3.1 语言与语料库

图瓦语是一种突厥语，在俄罗斯联邦的图瓦共和国、蒙古国以及中国新疆维吾尔自治区使用，这些地区的使用者总数约为28万人（Harrisonand Anderson，2002年）。本研究聚焦于中国新疆阿勒泰地区使用的准噶尔图瓦语变体。我们将准噶尔图瓦语视为一种用于语言记录语境的低资源变体；本文不对其正式濒危状态作出判定。准噶尔图瓦语具有图瓦语的核心类型学特征——典型黏着形态、丰富的格标记（主格、宾格、属格、与格、处所格、离格、伴随格）、复杂的体系统以及能产的派生形态同时也表现出元音和谐与辅音交替，由此产生异形变化（Mawkanuli，1999，2005）。

我们的语料库包含895个带有交错式语素 gloss 标注的句子，数据采集于1987年至1995年在中国新疆进行的田野调查。这些数据涵盖40场录音场次，涉及会话语体和叙事文本。所有交错式语素 gloss 标注均由人工完成，并遵循一套一致的项目专用规范，该规范参考了类型学惯例（Lehmann，2004b；Comrieetal.，2008年）。注释区分词汇项（例如，金钱、给予）与语法语素（例如，DAT、1SG、PRS）。

表 1 总结了语料库统计数据。标签集包含240个独特的语法语素标签（例如，1SG、PST）和1258个独特的实词释义。

指标	#
句子总数	895
叙事文本	40
独特语法语素数	240
独特内容语素	1258
每句平均词数	8.38 (± 5.71)
每词平均语素数	1.69 (± 0.30)

表 1：图瓦语田野调查数据集的语料库统计数据。

示例1 展示了语料库中的一个典型交错式语素 gloss 实例，说明了对齐结构是我们的模型必须学习的。

(1) jïlgï-nan iyi joon bar horse-ABL
two big EXIST “（我们）有两匹大马。”

3.2 数据划分

我们在文档级进行训练测试集划分，将约 85 % 的句子（760 句）分配到训练集，15%（135 句）分配到测试集。为避免信息泄漏，同一叙事文本的片段不会同时出现在两个数据集中，从而确保模型无法利用相关话语中的篇章级或说话者特定模式。我们还使用字符级词频-逆文档频率（TF-IDF）余弦相似度，并将阈值设为 0.95，以进一步确认不存在近重复句子。

4 方法

4.1 任务形式化

我们将 gloss 标注表述为一个结构化预测问题：给定一条以连字符切分的图瓦语话语$\mathbf{x} = (x_1, x_2, \ldots, x_n)$，其中每个 x_i 表示一个语素，生成与之对应的词义标注标签序列$\mathbf{y} = (y_1, y_2, \ldots, y_n)$，其中每个 y_i 均取自一个标签集。我们假定语素边界已知，并且 gloss 标注模型不使用译文；切分和 gloss 标注被视为独立步骤。我们使用词元级准确率进行评估，其定义如下：

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i] \tag{1}$$

其中，N 是测试集中语素的总数，$\hat{}y_i$ 是预测语素注释，y_i 是参考词义标注。该指标直接反映标注工作量减少的程度：准确率越高，所需的人工修正越少。

4.2 BiLSTM-CRF 模型

我们的基线采用带有 CRF 解码的双层双向 LSTM，广泛用于序列标注（Lample 等人，2016年；Ma 和 Hovy，2016年；Huang 等人，2015年）。该模型使用 100 维的字符级嵌入（随机初始化并从头开始训练）、128 维的隐藏层，并学习为切分后的语素预测词义标注。我们

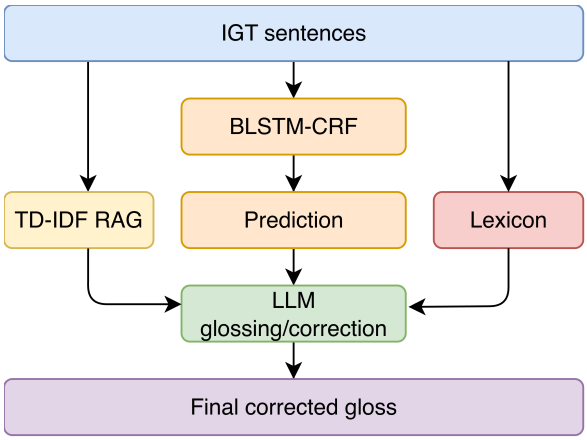


图1：将 BiLSTM-CRF 结构化预测与采用检索增强提示的大语言模型后校正相结合的混合流程。

训练最多 100 个训练轮次，并根据验证损失采用提前停止（耐心值为=10）。该架构能够捕捉局部形态模式，但无法利用更广泛的语言学知识，也无法处理训练数据中未得到充分表示的罕见范式。

4.3 大语言模型配置与提示

我们评估了四个大语言模型：`deepseek-v3.2-exp`（DeepSeek-AI 等人，2025年），`qwen3-max` （Yang 等人，2025a），`gpt-4o-mini` （OpenAI 等人，2024），以及 `gemma-3-27b-it` （Kamath 等人，2025年）。所有模型均采用温度为零的贪心解码，以获得确定性输出。提示遵循一致的模板：自然语言指令， k展示语素切分与释义对的检索示例、可选的语素词典以及测试输入。对于混合流程，提示还会额外纳入双向长短期记忆网络预测，作为待校正的假设，并将其表述为需要核验的“粗略初步尝试”。所有实验的完整提示模板见附录A。图1展示了我们的混合流程架构。

4.4 实验设计

我们开展了若干实验，测试大语言模型生成和混合校正。在实验 1（检索与随机选择）中，我们比较基于字符级词频-逆文档频率余弦相似度的检索与均匀随机抽样，以选择三个上下文示例，从而探究基于相似度的检索所产生的影响。检索在句子级进行：对于每个测试句子，我们根据图瓦语源文本表示检索最相似的训练句子，

实验 2（N 样本扩展）在使用检索增强生成且不提供术语表的情况下，将示例数量从 1 调整到 20，以描绘检索增强生成大语言模型生成的准确率-成本权衡。

实验 3（术语表消融）测试了四种术语表配置——无术语表、最常见的前 100 个语素、全部语法语素，以及包含 1,498 对词条的完整词典——均采用三示例检索增强生成，以揭示部分词典究竟有助于还是妨碍性能。术语表以纯文本键值列表的形式置于提示中，提供给大语言模型（附录 A），而非作为确定性查找表。最后，实验4（混合流程）评估了双向长短期记忆网络加大语言模型纠正的效果，并改变 n-shot 数量，其中包括零样本条件，以检验大语言模型能否在没有上下文示例的情况下纠正预测结果。这些消融实验对应于与田野调查相关的选择，包括检索质量、示例预算和词汇资源的可用性。检索增强生成大语言模型生成（实验 1–3）和混合纠正（实验4）的详细提示模板见附录 A。

5 结果

5.1 基线：BiLSTM-CRF 性能

我们的 BiLSTM-CRF 基线在包含 760 个句子的训练数据上，于测试集取得了 0.474 的词元准确率。该模型能够学习常见的形态模式（格标记、所有格标记、时态/体），但难以处理罕见的词汇语素，以及训练数据中未出现过的语法语素组合。错误分析显示，0.38 的错误涉及训练数据中出现次数少于 5 次的词汇项，0.24 的错误涉及以新颖组合形式出现的语法语素。

5.2 实验 1：检索与随机选择

表2 展示了在不使用术语表的情况下，采用三样本提示时，检索增强对所有四个大语言模型的影响。

检索增强生成在所有模型上均带来了显著提升：deepseek-v3.2-exp 提升了 +0.388，qwen3-max 提升了 +0.319，gpt-4o-mini 提升了 +0.293，gemma-3-27b-it 提升了 +0.276。deepseek-v3.2-exp 在使用 RAG 时取得最高的绝对准确率（0.506），而所有模型都因检索而显著提升。不同

模型	随机	RAG
deepseek-v3.2-exp	0.118	**0.506**
qwen3-max	0.062	0.381
gpt-4o-mini	0.103	0.396
gemma-3-27b-it	0.068	0.344

表2：四种大语言模型中，检索增强提示（RAG）与随机示例选择的对比（3-shot，不使用术语表）。所有模型使用 RAG 后均有显著提升。

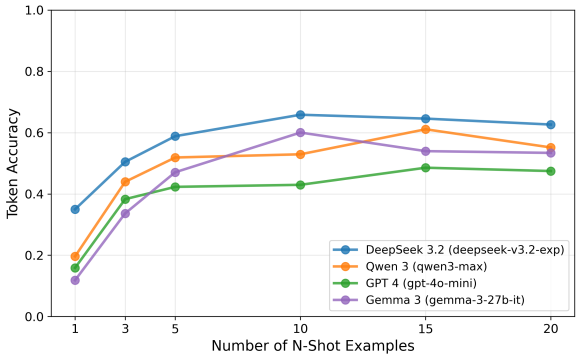


图2：实验 2：检索增强生成大语言模型生成的 n 样本扩展曲线。性能随示例数量近似呈对数级增长，大多数模型在 n=10–15 左右趋于平稳。文中提供了 BiLSTM基线（0.474）以供参考。

架构上的一致提升表明，基于相似度的示例选择有利于形态注释任务。

5.3 实验 2：N 样本扩展

采用与实验 1 相同的基于词频-逆文档频率的检索方法，将检索示例数量从 1 调整至 20，且不提供任何术语表。图 2 展示了性能随示例数量的变化情况（详细数值见表 5，附录 B）。

性能随示例数量近似按对数增长。deepseek-v3.2-exp 在 n=10 （0.658）时达到峰值，随后在 n=15（0.646）和 n=20 （0.626）时略有下降，而 qwen3-max 则持续提升，直至 n=15（0.611）。gpt-4o-mini 在 n=15（0.486）时达到峰值，gemma-3-27b-it 在 n=10 时达到 0.600，随后在 n=20时降至 0.534。各模型表现出一致的模式：超过 10 至 15 个示例后，边际收益递减，且部分模型在更高取值时性能下降。这种下降可能反映了模型饱和（过多语境造成干扰），也可能是检索质量下降所致（示例池扩大后，检索到的示例相似度降低）。

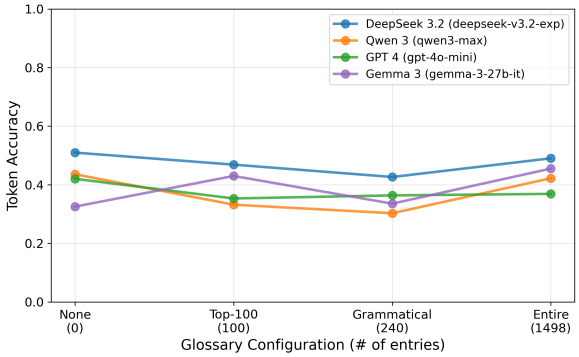


图 3：实验 3：术语表消融结果。与不使用术语表相比，部分术语表（前 100、语法）会降低性能，而完整术语表则带来小幅提升。负面影响表明，模型通常会受到形态信息的干扰。文中提供了 BiLSTM基线（0.474）以供参考。

我们使用字符级词频-逆文档频率余弦相似度进行检索；编辑距离或上下文化嵌入等其他相似度度量可能呈现不同的缩放行为，但这项研究留待未来开展。这些结果表明，对于成本敏感型应用，n=5至 10 是较为实用的取值范围；若要实现最高准确率，则可取 n=10至 15。

5.4 实验 3：术语表消融

为评估语素词典对性能的影响，我们使用 3 示例检索增强生成测试了四种术语表配置：无（不提供词典）、前 100（频率最高的 100 个语素—释义对）、语法（全部 240 个语法语素—释义对）以及完整（包含语法和词汇语素的完整 1,498 对词典）。图 3 展示了结果（详细数值见附录4 中的表B ）。

与直觉相反，提供语素词典通常会损害性能。部分术语表会导致所有模型的准确率持续下降：前 100 会造成 0.041 至 0.104 的降幅，而语法术语表带来的下降幅度相近甚至更大。即使是完整的 1,498 对词典，对大多数模型也无济于事；deepseek-v3.2-exp、qwen3-max 和 gpt-4o-mini 使用完整术语表时的表现都不如不使用术语表（分别下降 0.019、0.014 和 0.051）。只有 gemma-3-27b-it 能从词典信息中受益，使用完整术语表时达到 0.455（比无术语表高 +0.130）。

我们缺乏直接证据，无法确定模型是否

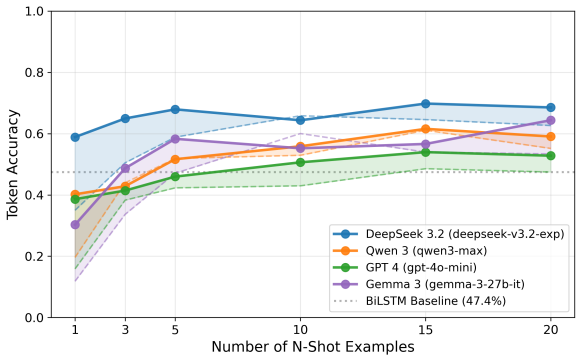


图4：实验4：混合流水线相较于检索增强生成大语言模型生成的改进。实线表示混合准确率（BiLSTM + 大语言模型纠正），虚线表示实验 2 中纯 N样本基线，阴影区域表示提升幅度。混合方法在所有四个模型上的表现均有所提升，尤其是在低样本场景（n=1–5）中。

确实查阅了词典条目，还是仅仅被提示中增加的材料分散了注意力。性能下降可能是因为模型在依赖语境的 gloss 标注决策中过度依赖词典查询，也可能是因为模型难以平衡多种信息来源（检索到的示例、词典和形态模式）。无术语表条件迫使模型从对齐示例中提取形态结构；对于大多数架构而言，这似乎比查阅词典更有效。不过，若不对模型预测进行细粒度分析，我们就无法明确分离出其中的因果机制。

5.5 实验4：混合流水线性能

接下来，我们使用与实验 1 和 2 相同的词频-逆文档频率检索来评估混合流程，但将 BiLSTM-CRF 的预测作为大语言模型纠正的初始假设。我们测试的检索示例数量为 n=1、3、5、10、15、20，且均不使用术语表。图4 展示了在不同 n 样本数量下，混合流程相较于检索增强生成大语言模型生成的提升，涵盖全部四个模型。

混合流程在所有模型上均显著优于纯双向长短期记忆网络预测。gemma-3-27b-it在n=20处达到0.644，较BiLSTM基线（0.474）提升a +0.170，并较其纯生成性能（0.534，位于n=20）改进 +0.110。deepseek-v3.2-exp在n=15 处达到0.698（+0.224高于双向长短期记忆网络， +0.052高于纯生成），qwen3-max达到0.616（+0.142高于双向长短期记忆网络， +0.005高于纯generation），gpt-4o-mini达到0.540（+0.066高于双向长短期记忆网络， +0.054高于纯生成），表明混合方法使各性能层级的所有模型均受益。

提升在低样本场景中尤为显著：当 n=1 时，gemma-3-27b-it 相较于纯生成（0.118 $\rightarrow$ 0.303）的提升为 +0.185，deepseek-v3.2-exp 提升了 +0.058，gpt-4o-mini 提升了 +0.063。这表明，在可用示例较少时，BiLSTM 能提供有价值的结构指导。随着 n 增大，提升幅度有所减小，但在所有模型上仍保持一致；即使在 n=20 时，gemma-3-27b-it 仍提升了 +0.110。

5.6 错误分析

为了解改进来自何处，我们按语素类型分析错误。将 BiLSTM 基线与表现最佳的混合配置（deepseek-v3.2-exp 搭配 $n = 10$ 个检索增强生成示例）进行比较后，我们发现改进并不均衡。

BiLSTM 基线在 168 个语法语素上的准确率为 0.923，但在 213 个词汇语素上的准确率仅为 0.479。这证实，结构化模型能够有效学习形态范式，但难以应对词汇空缺。混合流程将词汇语素准确率提升至 0.682（绝对提升 +0.204），而语法语素准确率则略微下降至 0.866（−0.057）。词汇语素准确率的大幅提升超过了语法语素准确率的轻微损失，这解释了混合方案整体性能的提升。

按训练频率分层后，这一模式变得更加清晰。我们根据每个测试语素在训练数据中的出现频率进行分类：低频（出现 1 至 5 次）、常见（6 至 20 次）和高频（超过 20 次）。表 3 展示了各频率区间的准确率。

频率	计数	双向长短期记忆网络	混合
低频（1 至 5）	69	0.029	0.426
常见（6 至 20）	58	0.448	0.724
高频（超过20次）	86	0.860	0.859

表3：词汇语素按训练频率划分的准确率。混合方法的提升主要集中在低频语素（+0.397），且所有测试语素在训练中均出现过。

低频语素的表现大幅提升：双向长短期记忆网络的准确率仅为 0.029，而混合方法达到 0.426（提升 +0.397）。常见语

素的准确率从 0.448 提升至 0.724（提升 +0.276）。高频语素在两种方法下都稳定在 0.86 左右。值得注意的是，在此次划分中，没有任何测试语素完全未曾出现。我们并未强制要求满足这一条件；由于采用文档级划分且评估集较小，所有测试语素至少在训练数据中出现过一次。这表明，混合方法的增益源于对低频项目进行语境推断，而非处理零样本词汇；同时也说明，只要有足够的训练示例，基线便能有效捕捉高频语法标记。

5.7 主要发现总结

我们的实验揭示了大语言模型辅助形态学注释中的若干重要规律。检索增强提示对所有模型都至关重要，基于相似度的示例选择表现远远优于随机选择。或许最令人惊讶的是，提供语素词典通常会损害模型表现：部分词典无一例外地降低准确率，而即使是完整词典也未能帮助大多数模型（四个模型中只有一个有所提升）。随着上下文示例数量增加，表现大致呈对数增长，通常在十到十五个示例左右达到峰值，之后趋于平稳或开始下降。

将双向长短期记忆网络的预测与大语言模型纠正相结合的混合架构，在所有受测模型上都提升了表现。在低样本场景下，这些提升尤为显著，此时可用示例很少，这表明当上下文学习受限时，结构化模型能够提供有价值的指导。即使没有任何示例，大语言模型也能成功识别并纠正结构化预测中的许多错误，这表明它们具备内在的形态学推理能力。

6 讨论

6.1 混合架构的优势

我们的结果表明，将结构化预测与大语言模型推理相结合，能够在所有测试模型中带来稳定的改进。BiLSTM-CRF 能够从有限的训练数据中捕捉常见的形态模式，但难以处理罕见语素和新组合。检索增强生成大语言模型能够利用更广泛的语言学知识和少样本泛化能力，但其准确率会因模型选择而有很大差异。

混合流程结合了这些互补优势，并通过错误分析揭示的分工发挥各自所长。结构化模型在语法语素上的准确度更高，这可能反映了这些语素在训练数据中频率较高且规律性较强；而大语言模型擅长对低频词汇项进行语境推断，并借助检索示例弥补词汇缺口。混合方法的提升主要集中在低频语素上，因为双向长短期记忆网络缺乏足够的训练信号；对于高频词项，两种方法的表现则相近。这种不对称性解释了两阶段架构为何能够奏效：每个组件都弥补了另一个组件的主要弱点。双向长短期记忆网络预测提供结构指导（在低样本场景中尤为有用），而大语言模型则利用检索示例中的模式对这些预测进行优化。我们的提示设计将双向长短期记忆网络的输出设定为可能出错的假设，而非权威的基线，从而鼓励批判性评估，同时提供有用的结构约束。

我们并未声称 BiLSTM-CRF 是这项任务的最佳基础模型，但结果揭示了一条更普遍的原则：将任何可训练的结构化预测器与检索增强的大语言模型后校正相结合，都能比单独使用任一方法取得更好的效果。这种协同效应在所有性能层级上都普遍有益，尤其是在示例较少或预算受限、导致上下文学习能力不足时，提升尤为显著。

6.2 词素词典的作用

或许我们最令人意外的发现是，与完全不提供词典相比，提供语素词典通常会降低性能。部分术语表都会降低准确率，而即使是完整词典，也未能帮助大多数模型。只有一个模型（gemma-3-27b-it）从完整词典中获得了显著提升，另有三个模型使用完整词典后的表现反而不如未使用时。

这些模式表明，模型整合词典信息的方式可能存在问题，尽管我们尚未直接分析模型是否确实查阅了词典条目。性能下降可能源于信息过载、不恰当的词典查阅方式，或欠佳的提示结构。我们的方法将词典以简单、无结构的键值列表形式提供；其他策略值得进一步探究，例如按形态类别组织词条，或以结构化格式呈现词典。

值得注意的是，提示结构会显著影响模型行为，而我们尚未系统地探索其他表述方式。因此，我们的发现应解读为：我们特定的词典整合提示设计效果不佳，而非词典在形态 gloss 标注中的应用存在根本局限。未来研究应开展系统的提示工程研究，以确定更有效的词汇资源整合策略。

6.3 田野调查的实践意义

我们的研究结果表明，对于使用大语言模型辅助进行交错式语素 gloss 标注的实践者而言，有几点值得考虑，不过这些规律可能会因语言、模型和语言记录语境的不同而有所差异。在我们的实验中，基于相似度的检索在所有受测模型中始终优于随机示例选择，这表明投入建设检索基础设施或许值得。将结构化模型与大语言模型后校正相结合的混合方法，在我们测试的所有配置中均取得了提升，说明这种两阶段策略可能具有价值。

对于语素词典，我们的结果提示应保持谨慎：在本次设置中，简单的键值呈现方式往往会降低大多数模型的表现，不过其他呈现策略尚未经过探索。关于少样本示例数量，我们观察到其效果大致呈对数增长，超过 10–15 个示例后收益递减；不过，最佳运行点可能取决于任务复杂度、模型能力和成本限制。这些规律来自我们使用图瓦语和四种通用大语言模型进行的特定实验设置；鉴于模型能力和架构设计仍在迅速演进，实践者应结合自身语言和工作流程验证这些发现。

6.4 伦理考量

我们的实验使用了在获得知情同意后收集的数据，相关协议限制了原始材料的共享。为保护说话人隐私，我们仅报告汇总统计数据和匿名化示例。说话人社群并非铁板一块，图瓦语说话人分布在多个社会政治背景各异的国家。我们不声称获得了整个社群的同意；我们遵守数据协议，并咨询相关田野调查合作伙伴。鉴于跨境语境，我们避免发布原始数据，也避免透露个人或地点的身份信息。商业应用程序接口引发了人们对训练数据来源的担忧（Bender 等人，2021；Sainz 等人，2023），而且即使达到了相应的准确率水平，不加批判地采用这些接口也可能给语言学记录引入错误。有关自动化的决策应与利益相关方协商作出，在提高效率的同时兼顾对数据控制权和质量的关切。

6.5 未来方向

关键方向包括在多种形态系统（多式综合、模板式、声调型）上开展跨语言评估，以检验我们的设计原则能否推广到突厥语黏着形态之外。联合切分与注释可以解决当前流水线假设边界已知这一局限。参数高效微调有望在仅需少量特定语言数据的情况下提升性能。结合置信度估计的交互式界面可帮助标注人员优先复核不确定的预测，而范式级评估则能更好地衡量词元级准确率以外的形态泛化能力。

7 结论

我们提出了一种混合自动注释流水线，将 BiLSTM-CRF 结构化预测与大语言模型后校正相结合，并使用四种大语言模型对图瓦语田野调查数据进行了评估。这种混合方法在所有测试模型上都持续提升了性能，在低样本场景中的提升尤为显著，因为 BiLSTM 提供的结构指导在此类场景中最具价值。错误分析表明，性能提升主要集中在词汇语素上，尤其是罕见词汇项；而 BiLSTM 已能有效捕捉语法范式。这体现的是互补优势，而非简单叠加性能。

我们的消融研究揭示了关键设计原则：检索增强提示能带来显著提升；语素词典通常会降低大多数模型的性能；性能随示例数量呈对数增长，并在约十到十五个示例时趋于平稳；混合校正则普遍有益于所有模型。这些发现挑战了关于提示工程的传统假设，表明信息并非越多越好。

尽管这些准确率水平对实际标注流程的效用仍是一个悬而未决的问题，

其答案取决于社区优先事项和标注语境，而较高的错误率意味着必须进行审慎的人工监督。模型输出应被视为有待专家验证的假设，而非权威性标注。

8 局限性

我们的评估仅针对一个语系（突厥语族）中的一种语言（图瓦语），尚未检验其对其他形态系统的泛化能力。我们观察到的模式未必适用于多式综合、模板式或非连接形态学。此外，我们的测试集规模较小，虽反映了现实中的标注成本，却也引入了抽样方差，限制了统计精度。

我们的评估指标仅衡量词元级准确率，未涵盖范式一致性、形态音系规律性，或与社区语言优先事项的一致性等更高阶属性。该流程还假设标准语素边界已知，并未处理切分问题；而在田野调查语境中，切分本身就需要大量语言学专业知识。

本研究对提示工程的探索缺乏系统性。每项实验中我们只测试了一种提示结构，因此尚未探究其他表述方式可能如何影响词典有效性、示例整合或指令遵循。提示设计的选择（信息排序、指令措辞、格式规范，以及不同知识来源之间的平衡）会显著影响模型行为，但我们缺乏必要的受控比较，无法分离这些因素的影响。尤其是，我们对术语表进行消融实验所得的发现，应理解为我们的特定提示设计未能有效利用词典信息，而非证明词典无法帮助形态注释。我们也缺乏细致分析，无法判断模型是否确实查阅了词典条目，或是否会被额外的提示内容分散注意力。

商业 API 访问限制了我们了解训练数据透明度以及现有图瓦语语言资源可能造成的数据污染的能力。我们使用通用大语言模型，而非 GlossLM 等专用模型（Ginn 等人，2024b），后者专门使用词间注释数据进行训练，可能达到更高的绝对准确率。不过，我们着重确立设计原则（检索策略、术语表配置、混合架构），这些原则很可能适用于不同类型的模型。此外，通用大语言模型在田野调查中具有实际优势：无需本地基础设施、仅凭少量语言专属数据即可在少样本设置下运行，而且语言学家无需具备机器学习专业知识也能使用。专用模型的性能与通用模型的实际易用性之间如何权衡，值得进一步研究。

参考文献

Diego Barriga Martínez、Victor Mijangos 和 Xi- mena Gutierrez-Vasques。2021年。奥托米语自动行间 gloss 标注。载于美洲原住民语言自然语言处理第一届研讨会论文集，第34–43页，线上。计算语言学协会。

Emily M. Bender、Timnit Gebru、Angelina McMillan- Major 和 Shmargaret Shmitchell。2021。随机鹦鹉的危险：语言模型会不会太大？载于*2021*年 *ACM* 公平性、问责制与透明度会议论文集，FAccT'21，第610–623页，纽约，美国。计算机协会。

Jan Buys 和 Jan A. Botha。2016年。低资源语言的跨语言形态标注。载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），第1954–1964页，柏林，德国。计算语言学协会。

Shobhana L. Chelliah 和 Willem J. de Reuse。2010年。描写语言学田野调查手册。施普林格科学与商业媒体。Google-Books-ID：d1FfE30hZ7EC。

Bernard Comrie、Martin Haspelmath 和 Bickel Balthasar。2008年。《莱比锡注释规则：语素间注释的约定》。Google-Books-ID：e8B7AQAACAAJ。

Ryan Cotterell 和 Georg Heigold。2017年。跨语言字符级神经形态标注。载于*2017*年自然语言处理实证方法会议论文集，第748–759页，丹麦哥本哈根。计算语言学协会。

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, J unxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H.Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui

Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jia nzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, XinXie, Xingchao Liu, XinyuYang, XinyuanLi, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxi- ang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yu- jia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zi- jia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang和Zhen Zhang. 2025年. DeepSeek-R1：通过强化学习激励大语言模型的推理能力. *arXiv*预印本. ArXiv:2501.12948[计算机科学].

Micha Elsner 和 David Liu。2025年。提示与情境”：面向低资源语言行间注释的逐词大语言模型提示方法。载于第 *22* 届 *SIGMORPHON* 计算形态学、音系学与语音学研讨会论文集，第 1–14 页，美国新墨西哥州阿尔伯克基。计算语言学协会。

Michael Ginn、Mans Hulden 和 Alexis Palmer。2024a。我们能否教会语言模型为濒危语言进行 gloss 标注？载于计算语言学协会 *EMNLP 2024* 论文集：论文集，第 5861–5876 页，美国佛罗里达州迈阿密。计算语言学协会。

Michael Ginn、Sarah Moeller、Alexis Palmer、Anna Stacey、Garrett Nicolai、Mans Hulden 和 Miikka Silfverberg。2023 年。SIGMORPHON 2023 行间注释 gloss 标注共享任务研究成果。载于第*20*届*SIGMORPHON*研讨会论文集，主题为语音学、音系学与形态学计算研究，第186–201页，多伦多，加拿大。计算语言学协会。

Michael Ginn、Lindia Tjuatja、Taiqi He、Enora Rice、Graham Neubig、Alexis Palmer 和 Lori Levin。2024b。GlossLM：用于行间注释文本的大规模多语言语料库和预训练模型。载于*2024* 年自然语言处理实证方法会议论文集，第 12267–12286 页，美国佛罗里达州迈阿密。计算语言学协会。

Sharon Hargus、Olga M. Semenova 和 Siri G. Tuttle。2020年。德内语 gloss 标注。出版方：阿拉斯加原住民语言中心。

K. David Harrison 和 Gregory D. S. Anderson。2002年。图瓦语语法。Scientific Consulting Services International。Google 图书 ID：RXnhAQAA- CAAJ。

Taiqi He、Kwanghee Choi、Lindia Tjuatja、Nathaniel Robinson、Jiatong Shi、Shinji Watanabe、Graham Neubig、David Mortensen 和 Lori Levin。2024。Wav2Gloss：从语音生成行间注释文本。载于计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文），第 568–582 页，泰国曼谷。计算语言学协会。

Zhiheng Huang、Wei Xu 和 Kai Yu。2015年。用于序列标注的双向 LSTM-条件随机场模型。*arXiv* 预印本。arXiv：1508。01991 [计算机科学]。

SIL国际。2025年。FieldWorks 语言探索者。

Gautier Izacard、Patrick Lewis、Maria Lomeli、Lucas Hosseini、Fabio Petroni、Timo Schick、Jane Dwivedi- Yu、Armand Joulin、Sebastian Riedel 和 Edouard Grave。2023 年。 Atlas：使用检索增强语言模型进行少样本学习。 机器学习研究期刊，24(251)：1–43。Zhengbao Jiang、Frank Xu、Luyu Gao、Zhiqing Sun、Qian Liu、Jane Dwivedi-Yu、Yiming Yang、Jamie Callan 和 Graham Neubig。2023 年。 主动检索增强生成。发表于 *2023* 年自然语言处理实证方法会议论文集，第 7969–7992 页，新加坡。计算语言学协会。Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière、Louis Rouillard、Thomas Mesnard、Geoffrey Cideron、Jean-bastien Grill、Sabela Ramos、Edouard Yvinec、Michelle Casbon、Etienne Pot、Ivo Penchev、Gaël Liu、Francesco Visin、Kathleen Kenealy、Lucas Beyer、Xiaohai Zhai、Anton Tsitsulin、Robert Busa- Fekete、Alex Feng、Noveen Sachdeva、Benjamin Coleman、Yi Gao、Basil Mustafa、Iain Barr、Emilio Parisotto、David Tian、Matan Eyal、Colin Cherry,

Jan-Thorsten Peter、Danila Sinopalnikov、Surya Bhupatiraju、Rishabh Agarwal、Mehran Kazemi、Dan Malkin、Ravin Kumar、David Vilar、Idan Brusilovsky、Jiaming Luo、Andreas Steiner、Abe Friesen、Abhanshu Sharma、Abheesht Sharma、Adi Mayrav Gi- l ady、Adrian Goedeckemeyer、Alaa Saade、Alex Feng、Alexander Kolesnikov、Alexei Bendebury、Alvin Abdagic、Amit Vadi、András György、An- dré Susano Pinto、Anil Das、Ankur Bapna、Antoine Miech、Antoine Yang、Antonia Paterson、Ashish Shenoy、Ayan Chakrabarti、Bilal Piot、Bo Wu、Bobak Shahriari、Bryce Petrini、Charlie Chen、Char- line Le Lan、Christopher A. Choquette-Choo、C. J. Carey、Cormac Brick、Daniel Deutsch、Danielle Eisenbud、Dee Cattle、Derek Cheng、Dimitris Pa- paras、Divyashree Shivakumar Sreepathihalli、Doug Reid、Dustin Tran、Dustin Zelle、Eric Noland、Er- win Huizenga、Eugene Kharitonov、Frederick Liu、Gagik Amirkhanyan、Glenn Cameron、Hadi Hashemi、Hanna Klimczak-Plucińska、Harman Singh、Harsh Mehta、Harshal Tushar Lehri、Hussein Hazimeh、Ian Ballantyne、Idan Szpektor、Ivan Nardini、Jean Pouget-Abadie、Jetha Chan、Joe Stanton、John Wi- eting、Jonathan Lai、Jordi Orbay、Joseph Fernan- dez、Josh Newlan、Ju-yeong Ji、Jyotinder Singh、Kat Black、Kathy Yu、Kevin Hui、Kiran Vodra- halli、Klaus Greff、Linhai Qiu、Marcella Valen- tine、Marina Coelho、Marvin Ritter、Matt Hoffman、Matthew Watson、Mayank Chaturvedi、Michael Moynihan、Min Ma、Nabila Babar、Natasha Noy、Nathan Byrd、Nick Roy、Nikola Momchev、Nilay Chauhan、Noveen Sachdeva、Oskar Bunyan、Pankil Botarda、Paul Caron、Paul Kishan Rubenstein、Phil Culliton、Philipp Schmid、Pier Giuseppe Sessa、Pingmei Xu、Piotr Stanczyk、Pouya Tafti、Rakesh Shiv- anna、Renjie Wu、Renke Pan、Reza Rokni、Rob Willoughby、Rohith Vallu、Ryan Mullins、Sammy Jerome、Sara Smoot、Sertan Girgin、Shariq Iqbal、Shashir Reddy、Shruti Sheth、Siim Põder、Sijal Bhat- nagar、Sindhu Raghuram Panyam、Sivan Eiger、Susan Zhang、Tianqi Liu、Trevor Yacovone、Tyler Liechty、Uday Kalra、Utku Evci、Vedant Misra、Vincent Rose-berry、Vlad Feinberg、Vlad Kolesnikov、Woohyun Han、Woosuk Kwon、Xi Chen、Yinlam Chow、Yuvein Zhu、Zichuan Wei、Zoltan Egyed、Victor Cotruta、Minh Giang、Phoebe Kirk、Anand Rao、Kat Black、Nabila Babar、Jessica Lo、Erica Moreira、Luiz Gus-tavo Martins、Omar Sanseviero、Lucas Gonzalez、ZachGleicher、Tris Warkentin、Vahab Mirrokni、Evan Senter、Eli Collins、Joelle Barral、Zoubin Ghahra-mani、Raia Hadsell、Yossi Matias、D. Sculley、Slav Petrov、Noah Fiedel、Noam Shazeer、Oriol Vinyals、Jeff Dean、Demis Hassabis、Koray Kavukcuoglu、Clement Farabet、Elena Buchatskaya、Jean-Baptiste Alayrac、Rohan Anil、Dmitry、Lepikhin、Sebastian Borgeaud、Olivier Bachem、Armand Joulin、Alek An-dreev、Cassidy Hardin、Robert Dadashi 和 Léonard Hussenot。2025年。Gemma 3 技术报告。 *arXiv* 预印本。ArXiv：2503。19786 [计算机科学]。

Guillaume Lample、Miguel Ballesteros、Sandeep Subramanian、Kazuya Kawakami 和 Chris Dyer。2016年。命名实体识别的神经网络架构。载于*2016*年北美计算语言学协会年会论文集：人类语言技术，第 260–270 页，圣迭戈，加利福尼亚州。计算语言学协会。

ChristianLehmann.2004a. 语言学中的数据. 21(3- 4):175–210. 出版方：De Gruyter Mouton 栏目：《语言学评论》。

ChristianLehmann.2004b. 行间语素注释. In Geert Booij, Christian Lehmann, Joachim Mugdan, StavrosSkopeteas, andWolfgang Kesselheim, 编者，形态学，页码 1834–1857。De Gruyter。

Patrick Lewis、EthanPerez、Aleksandra Piktus,FabioPetroni, Vladimir Karpukhin, NamanGoyal,HeinrichKüttler、MikeLewis、Wen-tau Yih,TimRocktäschel、Sebastian Riedel、and DouweKiela. 2021. 面向知识密集型自然语言处理任务的检索增强生成. *arXiv* 预印本.ArXiv:2005.11401 [计算机科学].

Zoey Liu、Robert Jimerson 和 Emily Prud'hommeaux。2021。塞内卡语的形态切分。载于美洲原住民语言自然语言处理第一届研讨会论文集，第 90–101 页，线上。计算语言学协会。Robert Logan IV、Ivana Balazevic、Eric Wallace、Fabio Petroni、Sameer Singh 和 Sebastian Riedel。2022。精简提示和参数：使用语言模型进行简单的少样本学习。载于 计算语言学协会发现：*ACL 2022*，第 2824–2835 页，爱尔兰都柏林。计算语言学协会。Xuezhe Ma 和 Eduard Hovy。2016。通过双向 LSTM-CNN-条件随机场实现端到端序列标注。载于 第*54*届计算语言学协会年会论文集（第一卷：长篇论文），第 1064–1074 页，德国柏林。计算语言学协会。Chaitanya Malaviya、Matthew R. Gormley 和 Graham Neubig。2018。用于跨语言形态标注的神经因子图模型。载于 第*56*届计算语言学协会年会论文集（第一卷：长篇论文），第 2653–2663 页，澳大利亚墨尔本。计算语言学协会。Talant Mawkanuli。1999。准噶尔图瓦语的音系学与形态学。博士论文，印第安纳大学，美国印第安纳州。ISBN：9780599667938。Talant Mawkanuli。2005。准噶尔图瓦语文本。乌拉尔—阿尔泰语系丛书。印第安纳大学，布卢明顿。Sarah Moeller、Ling Liu、Changbing Yang、Katharina Kann 和 Mans Hulden。2020。IGT2P：从行间注释文本到范式。载于 *2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第 5251–5262 页，线上。计算语言学协会。

Sebastian Nordhoff 和 Thomas Krämer。2022。IMT-Vault：从语法描写和类型学调查文章中提取并丰富低资源语言的行间注释文本。载于第十三届语言资源与评测会议期间举办的第八届语言学关联数据研讨会论文集，第 17–25 页，法国马赛。欧洲语言资源协会。

Shu Okabe 与 François Yvon。2023。面向多语行间形态注释。载于计算语言学协会研究成果：*EMNLP 2023*，第5958–5971页，新加坡。计算语言学协会。

OpenAI、Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat、Red Avila、Igor Babuschkin、Suchir Balaji、Valerie Balcom、Paul Baltescu、Haiming Bao、Mohammad Bavarian、Jeff Belgum、Irwan Bello、Jake Berdine、Gabriel Bernadett-Shapiro、Christopher Berner、Lenny Bogdonoff、Oleg Boiko、Madelaine Boyd、Anna-Luisa Brakman、Greg Brockman、Tim Brooks、Miles Brundage、Kevin Button、Trevor Cai、Rosie Campbell、Andrew Cann、Brittany Carey、Chelsea Carlson、Rory Carmichael、Brooke Chan、Che Chang、Fotis Chantzis、Derek Chen、Sully Chen、Ruby Chen、Jason Chen、Mark Chen、Ben Chess、Chester Cho、Casey Chu、Hyung Won Chung、Dave Cummings、Jeremiah Currier、Yunxing Dai、Cory Decareaux、Thomas Degry、Noah Deutsch、Damien Deville、Arka Dhar、David Dohan、Steve Dowling、Sheila Dunning、Adrien Ecoffet、Atty Eleti、Tyna Eloundou、David Farhi、Liam Fedus、Niko Felix、Simón Posada Fishman、Juston Forte、Isabella Fulford、Leo Gao、Elie Georges、Christian Gibson、Vik Goel、Tarun Gogineni、Gabriel Goh、Rapha Gontijo-Lopes、Jonathan Gordon、Morgan Grafstein、Scott Gray、Ryan Greene、Joshua Gross、Shixiang Shane Gu、Yufei Guo、Chris Hallacy、Jesse Han、Jeff Harris、Yuchen He、Mike Heaton、Johannes Heidecke、Chris Hesse、Alan Hickey、Wade Hickey、Peter Hoeschele、Brandon Houghton、Kenny Hsu、Shengli Hu、Xin Hu、Joost Huizinga、Shantanu Jain、Shawn Jain、Joanne Jang、Angela Jiang、Roger Jiang、Haozhun Jin、Denny Jin、Shino Jomoto、Billie Jonn、Heewoo Jun、Tomer Kaftan、Łukasz Kaiser、Ali Kamali、Ingmar Kanitscheider、Nitish Shirish Keskar、Tabarak Khan、Logan Kilpatrick、Jong Wook Kim、Christina Kim、Yongjik Kim、Jan Hendrik Kirchner、Jamie Kiros、Matt Knight、Daniel Kokotajlo、Łukasz Kondraciuk、Andrew Kondrich、Aris Konstantinidis、Kyle Kosic、Gretchen Krueger、Vishal Kuo、Michael Lampe、Ikai Lan、Teddy Lee、Jan Leike、Jade Leung、Daniel Levy、Chak Ming Li、Rachel Lim、Molly Lin、Stephanie Lin、Mateusz Litwin、Theresa Lopez、Ryan Lowe、Patricia Lue、Anna Makanju、Kim Malfacini、Sam Manning、Todor Markov、Yaniv Markovski、Bianca Martin、Katie Mayer、Andrew Mayne、Bob McGrew、Scott Mayer McKinney、Christine McLeavey、Paul McMillan、Jake McNeil、David Medina、Aalok Mehta、Jacob Menick、Luke Metz、Andrey Mishchenko、Pamela Mishkin、Vinnie Monaco、Evan Morikawa、Daniel Mossing、Tong Mu、Mira Murati、Oleg Murk、David Mély、Ashvin Nair、Reiichiro Nakano、Rajeev Nayak、Arvind Neelakantan、Richard Ngo、Hyeonwoo Noh、Long Ouyang、Cullen O'Keefe、Jakub Pachocki、Alex Paino、Joe Palermo、Ashley Pantuliano、Giambattista Parascandolo、Joel Parish、Emy Parparita、Alex Passos、Mikhail Pavlov、Andrew Peng、Adam Perelman、Filipe de Avila Belbute Peres、Michael Petrov、Henrique Ponde de Oliveira Pinto、Michael Pokorny、Michelle Pokrass、Vitchyr H. Pong、Tolly Powell、Alethea Power、Boris Power、Elizabeth Proehl、Raul Puri、Alec Radford、Jack Rae、Aditya Ramesh、Cameron Raymond、Francis Real、Kendra Rimbach、Carl Ross、Bob Rotsted、Henri Roussez、Nick Ryder、Mario Saltarelli、Ted Sanders、Shibani Santurkar、Girish Sastry、Heather Schmidt、David Schnurr、John Schulman、Daniel Selsam、Kyla Sheppard、Toki Sherbakov、Jessica Shieh、Sarah Shoker、Pranav Shyam、Szymon Sidor、Eric Sigler、Maddie Simens、Jordan Sitkin、Katarina Slama、Ian Sohl、Benjamin Sokolowsky、Yang Song、Natalie Staudacher、Felipe Petroski Such、Natalie Summers、Ilya Sutskever、Jie Tang、Nikolas Tezak、Madeleine B. Thompson、Phil Tillet、Amin Tootoonchian、Elizabeth Tseng、Preston Tuggle、Nick Turley、Jerry Tworek、Juan Felipe Cerón Uribe、Andrea Vallone、Arun Vijayvergiya、Chelsea Voss、Carroll Wainwright、Justin Jay Wang、Alvin Wang、Ben Wang、Jonathan Ward、Jason Wei、C. J. Weinmann、Akila Welihinda、Peter Welinder、Jiayi Weng、Lilian Weng、Matt Wiethoff、Dave Willner、Clemens Winter、Samuel Wolrich、Hannah Wong、Lauren Workman、Sherwin Wu、Jeff Wu、Michael Wu、Kai Xiao、Tao Xu、Sarah Yoo、Kevin Yu、Qiming Yuan、Wojciech Zaremba、Rowan Zellers、Chong Zhang、Marvin Zhang、Shengjia Zhao、Tianhao Zheng、Juntang Zhuang、William Zhuk和Barret Zoph。2024. GPT-4 技术报告. *arXiv*预印本. ArXiv:2303.08774 [计算机科学].

Enora Rice、Ali Marashian、Luke Gessler、Alexis Palmer 与 Katharina von der Wense。2024。TAMS：翻译辅助形态切分。载于 第*62*届计算语言学协会年会论文集（第*1*卷：长文），第6752–6765页，曼谷，泰国。计算语言学协会。

Enora Rice、Katharina von der Wense 与 Alexis Palmer。2025年。对话中的跨学科研究：面向语言记录的计算形态学案例研究。载于*2025*年自然语言处理实证方法会议论文集，第11284–11296页，苏州，中国。计算语言学协会。

Erich Round、Mark Ellison、Jayden Macklin-Cordes 和 Sacha Beniamine。2020年。从语法描写页面图像中自动解析行间注释文本。载于第十二届语言资源与评估会议论文集，第2878–2883页，法国马赛。欧洲语言资源协会。

Tatyana Ruzsics 和 Tanja Samardži´c. 2017年。神经

网络序列到序列学习中的词内部结构。载于第*21*届计算自然语言学习会议论文集（*CoNLL 2017*），第184–194页，温哥华，加拿大。计算语言学协会。

Oscar Sainz、Jon Campos、Iker García-Ferrero、Julen Etxaniz、Oier Lopez de Lacalle 和 Eneko Agirre。2023年。NLP评估陷入困境：论衡量每个基准测试中大语言模型数据污染的必要性。计算语言学协会：*EMNLP 2023*论文集，第10776–10787页，新加坡。计算语言学协会。

Mathias Schenner 和 Sebastian Nordhoff。2016年。从 LaTeX 文档中提取行间注释文本。载于第十届语言资源与评测国际会议论文集（*LREC'16*），第4044–4048页，波尔托罗ž，斯洛文尼亚。欧洲语言资源协会（ELRA）。

Bhargav Shandilya 和 Alexis Palmer。2023。语境中的轻量级语素标注：利用结构化语言学表示支持语言记录语境中的语言分析。载于第*20*届*SIGMORPHON*研讨会：语音学、音系学与形态学计算研究论文集，第78–92页，多伦多，加拿大。计算语言学协会。

Shamane Siriwardhana、Rivindu Weerasekera、Elliott Wen、Tharindu Kaluarachchi、Rajib Rana 和 Suranga Nanayakkara。2023。改进检索增强生成（RAG）模型的开放领域问答域适应。《计算语言学协会汇刊》，11：1–17。出版地：美国马萨诸塞州剑桥；出版社：麻省理工学院出版社。

Maja Stahl、LeonBiermann、Andreas Nehring 和 Henning Wachsmuth。2024。用于联合作文评分与反馈生成的大语言模型提示策略探索。载于第*19*届自然语言处理在教育应用中的创新用途研讨会论文集（*BEA 2024*），第283–298页，墨西哥墨西哥城。计算语言学协会。

Genta Indra Winata、Andrea Madotto、Zhaojiang Lin、Rosanne Liu、Jason Yosinski 和 Pascale Fung。2021。语言模型是小样本多语言学习者。载于第一届多语言表征学习研讨会论文集，第1–15页，多米尼加共和国蓬塔卡纳。计算语言学协会。

Peter Wittenburg、Hennie Brugman、AlbertRussel、Alex Klassmann 和 Han Sloetjes。2006年。ELAN：多模态研究的专业框架。第 1556–1559 页。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayi-heng Liu、Fan Zhou、Fei Huang、FengHu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、XinyuWang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。2025a年。Qwen3 技术报告。*arXiv* 预印本。ArXiv：2505。09388 [计算机科学]。

Changbing Yang、Franklin Ma、Freda Shi 和 Jian Zhu。2025b。LingGym：大语言模型距离田野语言学家式思考还有多远？载于*2025*年自然语言处理实证方法会议论文集，第 1314–1340 页，中国苏州。计算语言学协会。

Wenhao Yu。2022。异构知识的检索增强生成。载于*2022*年北美计算语言学协会分会会议论文集：人类语言技术：学生研究研讨会，第 52–58 页，混合形式：美国华盛顿州西雅图+ 线上。计算语言学协会。

Kexun Zhang、Yee Choi、Zhenqiao Song、Taiqi He、William Yang Wang 和 Lei Li。2024年。聘请一位语言学家！：借助上下文内语言学描述在大语言模型中学习濒危语言。载于计算语言学协会年会论文集：*ACL 2024* 精选论文，第 15654–15669 页，泰国曼谷。计算语言学协会。

Xingyuan Zhao、Satoru Ozaki、Antonios Anastasopou- los、Graham Neubig 和 Lori Levin。2020年。借助译文为资源匮乏语言自动进行行间注释。载于第*28*届国际计算语言学会议论文集，第5397–5408页，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Vilém Zouhar、Martin Popel、Ondřej Bojar 和 Aleš Tamchyna。2021。神经机器翻译质量与译后编辑表现。载于*2021*年自然语言处理经验方法会议论文集，第10204–10214页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

A 实验提示模板

本附录提供了全部五项实验所用的完整提示模板。所有提示均遵循一致的结构：先以系统消息定义语言专家角色，随后是特定任务说明和格式要求。

A.1 系统消息（所有实验）

所有实验均使用以下系统消息：

> 你是一位语言专家，专门为一种未知语言进行逐词素注释。

A.2 实验 1–3：检索增强生成大语言模型生成

实验 1（检索与随机）、实验 2（术语表消融）和实验 3（N 样本扩展）均使用检索增强生成大语言模型生成，并采用以下模板：

> **用户消息：**
> 以下是一些句子示例，其中包含语素边界（以连字符标记）及其注释：
>
> *[对于 1..k 中的每个示例 i：]*
> 示例 *i*:
> 切分结果：*[按词素切分的源文本]*
>
> 词义标注： *[对应的注释序列]*
> *[如果提供了术语表：]*
> 你还会获得一本语素词典，其中列出了语素及其英文注释。对于词典中收录的语素，请使用提供的注释；对于其他语素，请根据语境推断。有些语素可能有多个译法，请选择最适合该语境的译法。
>
> 语素词典：*[语素1*：词义标注*1,*语素*2*：词义标注*2,...]*
> *[结束术语表部分]*
> 请对这个句子进行词义标注：
>
> 已切分： *[带有语素边界的测试句]*
> 请按相同结构输出词义标注（词与词之间用空格，语素与语素之间用连字符）。请将词义标注用###括起来。示例：### word1-MORPH1 word2-MORPH2###

实验特定变体：

- **实验 1：** 使用 3 个示例，不提供术语表。比较 TF-IDF 检索（检索增强生成）与随机采样。

- **实验 2：** 使用检索增强生成，无术语表。改变 $k \in \{1, 3, 5, 10, 15, 20\}$ 示例。

- **实验 3：** 使用 3 个检索增强生成示例。测试四种词汇表规模：无（0 对）、前 100（100 对）、语法（240 对）、完整（1,498 对）。

A.3 实验 4：混合流水线（双向长短期记忆网络 + 大语言模型）

实验4 使用 BiLSTM-CRF 的预测作为初始假设，供大语言模型纠正，所用模板如下：

> **用户消息：**
> 你将收到：
> 1. 统计模型生成的粗略初始 gloss 标注（可能包含错误） 2. 一些带有正确注释的示例句子 3. 一本语素词典
>
> 你的任务是利用所有可用信息生成正确的词义标注。
> 以下是一些带有正确注释的例句：
>
> *[对于每个示例 i（i 属于 1..k；零样本时取 0）：]*
> 示例 *i*:
> 切分结果： *[按词素切分的源文本]*
>
> 词义标注： *[对应的注释序列]*
> *[如果提供了术语表：]*
> 你还可以使用语素词典： *[语素1*：词义标注*1*，语素*2*：词义标注*2*，*...]*
>
> *[术语表部分结束]*
> 现在，请为这个句子进行词义标注：
>
> 已切分： *[带有语素边界的测试句]*
> 初始尝试（来自统计模型）：*[BiLSTM-CRF*预测*]*
> 这次初步尝试可能包含错误。请利用示例、词典和语言模式生成正确的词义标注。保持相同的结构（词之间用空格分隔，语素之间用连字符分隔）。
>
> 重要：仅输出用 ### 括起来的词义标注。不要解释你的

推理。示例格式：###word1-MORPH1 word2-MORPH2###

设计理由： 混合提示将双向长短期记忆网络预测呈现为“粗略的初步尝试”，而非权威的基线，从而鼓励大语言模型进行批判性评估并纠正错误。干净的黄金示例（而非错误纠正对）提供了范式语境，而统计预测则通过提出合理的语素边界和候选释义来缩小搜索空间。

A.4 输出提取

所有实验均通过定位 ### 分隔符之间的文本来提取模型输出。如果缺少分隔符（表示未遵循要求），则将整个输出字符串用作预测语素注释。这种提取方法在所有四家大语言模型提供商的模型上均表现稳健。

B 详细结果表

本节提供了正文中以图表形式呈现的实验详细数值结果。

术语表	DeepSeek	qwen3	GPT-4o	Gemma 3
None	**0.510**	**0.436**	**0.420**	0.325
前 100	0.469	0.332	0.353	0.430
语法	0.426	0.303	0.364	0.335
完整	0.490	0.422	0.369	**0.455**

表4：四种大语言模型的术语表消融研究（3 示例检索增强生成）。部分术语表会持续降低性能，而完整词典对大多数模型均无助益。模型名称缩写如下：DeepSeek = deepseek-v3.2-exp、qwen3 =qwen3-max、GPT-4o = gpt-4o-mini、Gemma 3-3 = gemma-3-27b-it。

N	DeepSeek	qwen3	GPT-4o	Gemma 3
1	0.350	0.196	0.159	0.118
3	0.505	0.439	0.383	0.336
5	0.588	0.519	0.423	0.471
10	**0.658**	0.529	0.430	**0.600**
15	0.646	**0.611**	**0.486**	0.540
20	0.626	0.552	0.475	0.534

表5：检索增强生成大语言模型生成的 N 样本扩展（检索增强生成，无术语表）。性能呈对数级提升，超过n =10–15 后收益递减。模型名称缩写方式与表4相同。