arXiv:2602.01162v1 [cs.CL] 1 Feb 2026

面向低资源语言的基于大语言模型的翻译：类型学指导的候选重排序

Nipuna Abeykoon*、**Ashen Weerathunga***、**Pubudu Wijesinghe***、**Parameswari Krishnamurthy**

ZWAG AI Ltd，Dubai AI Campus，DIFC，阿联酋。

*这些作者对本研究的贡献相同

摘要

主要使用高资源语言训练的大语言模型，会系统性地偏向占主导地位的类型学模式，因此在翻译为与之类型学差异显著的低资源语言时，往往出现结构不符合。我们提出一种利用语言类型学提升翻译质量的框架，无需平行训练数据或模型重新训练。该框架由两个部分组成：通用元语言框架（UMF）通过基于差异加权的评分，在16个类型学维度上将语言表示为结构化语言画像；计算引擎则在生成阶段进行语言消歧，并在选择阶段进行类型学符合度评分。对九种语言对的评估表明，干预率与和英语的类型学距离高度相关。在针对341个英语句子开展的实验中，每个句子都包含不同的形态和句法现象。对于按保守型处理的语言，该框架的干预精确率为48.16%；对于形态结构密集型语言，为28.15%；对于结构特征明确的语言，则为86.26%。该框架无需平行训练数据，可与任何能够生成多个候选输出的大语言模型配合使用，从而能够实际应用于资源匮乏的语言。

关键词：机器翻译、低资源语言、语言类型学、大语言模型、候选重排序、形态复杂度

1 引言

现代大语言模型将语言视为表层数据，而非认知系统。因此，它们的性能优化面向那些语法结构、文化假设和推理模式在训练语料库中占主导地位的语言。结构属性不同的语言往往会被扭曲、简化或错误表述。

大语言模型从以英语为中心的数据集中学习统计规律。当应用于结构迥异的语言时，这些模型往往能保留语义意图，却无法维持结构正确性。这会导致在时、施事性、证据性、礼貌性、空间关系和关系意义方面出现系统性错误。生成结果看似流畅，但对母语者而言，在认知上仍不协调。

我们将这类失败称为解释错误：输出保留了表层语义内容，却违反了正确解读所必需的语法、形态句法或话语约束。

母语者对其进行解读时会遇到这类错误。与文体变体或可接受的释义不同，解释错误会导致译文在目标语言中出现语法错误、语用不当或认知失调，即使预期含义仍可推断出来。

这不仅仅是数据匮乏的问题。人类语言呈现出惊人的结构多样性：语序模式涵盖从主谓宾（英语）到主宾谓（印地语、日语），再到谓主宾（阿拉伯语）；形态系统则涵盖从分析语（中文）到黏着语（土耳其语、斯瓦希里语），再到多式综合语（因纽特语）。格标记、一致关系模式、敬语以及数十种其他语法维度，在不同语言之间各自独立地变化 [9]。主要以英语为基础训练的模型，必须设法为呈现出根本不同类型学模式的语言生成结构上有效的输出。仅靠增加数据量无法解决这一挑战。其结果是系统性的结构错误：语序错误、形态标记缺失、词汇选择不当，以及违反目标语言的语法约束。这种不平等源于这些模型在训练和评估方式上的根本性偏差。

根本原因是类型学偏差：主要以英语为基础训练的模型，会编码主谓宾语序、分析型形态和最少格标记等结构模式。当翻译成类型差异语言时，这些偏差会体现为多个维度上的系统性错误。

结构错误源于语法错配。将句子“The children play in the garden”从英语翻译成僧伽罗语时，可能会保留主谓宾语序，而不是调整为僧伽罗语的主宾谓语序模式，结果译成 දරුවන් සෙල්ලම් උද්‍යානය，而非正确的 ළමයි උද්‍යානයේ සෙල්ලම් කරනවා。模型可能会省略必需的格标记：僧伽罗语要求在“garden”后添加处所后缀 ේ (-ē)来表示地点，也可能无法生成恰当的形态复杂度。近期分析证实了这一模式：语言模型在屈折语上的得分显著更高，但在黏着语上的表现较差，这是由于后者的形态更丰富、词元稀疏性更高 [3]。

词汇错误源于训练分布偏差。在我们对前沿大语言模型进行的实验中，“play”被译为 වාදනය（vādanaya，“演奏乐器”），而不是 සෙල්ලම්（sellam，“玩耍／玩乐”），因为前者在英语—僧伽罗语训练数据中出现得更频繁，尽管在“孩子们在花园里玩耍”这一语境中并不恰当。

现有补救措施存在显著局限。在平行语料库上进行微调所需的数据对大多数语言对而言并不存在[8], ，而提示工程缺乏精确性，难以强制执行复杂的语法约束[6, 7]。先前工作仅以有限或辅助形式纳入了类型学洞见；然而，尚无现有系统将类型学操作化为一种通用的结构化决策层，从而在无需重新训练或平行数据的情况下，直接支配跨语言的推理时候选选择。这促使我们提出如下方法：利用显式类型学知识，在黑盒生成条件下指导翻译。

通用元语言框架（UMF）通过量化语言之间的类型学分歧，并利用该分歧在无需重新训练的情况下指导翻译，从而应对这一挑战。该框架通过互补机制同时针对两类错误。对于词汇错误，语义约束层在生成过程中应用上下文感知调整，以消解词义歧义。对于结构错误，类型学重排序层依据目标语言的显式语法要求对候选项进行评估。

UMF 通过专家整理的语言画像来呈现语言，这些画像基于《世界语言结构地图集》[9] 和内部语言类型学研究，涵盖 16 个类型学维度。差异分数用于量化源语言和目标语言在各维度上的差异程度。这些分数按语言学重要性加权并经过归一化，从而生成指导候选项评估的指令向量。可调节的混合参数 α 用于平衡模型的概率与

类型学合规评分。

我们的方法基于这样一个原则：类型学差异能够预测大语言模型会在何处失败。语序存在分歧、格系统丰富且具有黏着形态的语言会获得更高的干预率，重排序器会优先选择满足目标语言约束的候选项。这种双层方法将类型学评分与语义消歧相结合，所带来的改进与类型学差异相关，并且在不同基础模型上保持一致。

范围与非主张。 为明确本工作的边界：UMF 并不生成翻译；它是在现有模型产生的候选项中进行选择。UMF 并不取代自动评估指标；它作为候选选择的补充决策层运作。UMF 并不声称具有普遍覆盖——当前的语言概况仅涵盖类型学现象的一个子集，仍需扩展。UMF 所提供的，是一种在推理时强制执行类型学约束的结构化、可解释机制，其作为元语言决策层运作，而非学习得到的质量预测器。

2 背景与相关工作

2.1 元语言与机器翻译框架中的类型学：既有方法

语言类型学以多种形式融入了机器翻译和元语言框架，但这些方法将类型学知识视为辅助性、分析性或局部性的内容，而非决定基础决策的核心组成部分。普遍语法（UG）等理论模型建立在这样一种假设之上：各种语言在深层结构上具有一致性，并将差异归结为少数抽象原则或二元参数。此后，大规模类型学证据表明，语言多样性远比 UG 所预测的丰富，也更难离散化；许多语言呈现出超出既定普遍规律范围的结构配置 [15, 16]。ATLAS 等基于中间语的系统试图采用独立于具体语言的表征，但实证评估表明，任何中间语都无法在不同语言之间保持中立，除非重新引入语言特定的转换规则，最终实际上又退回到成对或语系特定的解决方案 [17, 18]。基于规则的机器翻译系统显式编码了类型学结构，但需要人工穷尽式地指定语法规则，因此既不完整，也无法扩展，而且偏向于有充分文献记录的印欧语系语言结构[19]。

数据驱动范式（包括统计机器翻译和神经机器翻译）提升了表层流利度和可扩展性，但也带来了一类不同的局限。由于完全依据观测到的语料库进行学习，这些模型会隐式地偏重高资源语言和高频结构，系统性地低估罕见、可选或具有类型学标记的现象，例如丰富的形态学、证据性、敬语系统或非构型句法 [20, 21]。尤其是神经机器翻译模型，其优化目标是概率上的流利度，而非结构忠实度，因此常常省略或弱化那些缺乏明确跨语言对齐或充足训练信号的语言特征 [22]。尽管以往框架已纳入类型学信息，包括用于迁移式机器翻译的差异指数 [14] 以及类型学特征嵌入，但这些信息一直只是辅助性、局部性或分析性的内容，并未成为翻译系统中结构化且主导决策的组成部分 [23]。因此，现代机器翻译系统仍是有效的模式翻译器，而非以语言学为基础的解释器；类型学知识发挥的作用仍是元数据或旁路特征，而非直接主导推理时选择的一等评估机制。

2.2 大语言模型中的类型学偏见

语言类型学按照彼此独立变化的结构维度对语言进行分类。形态学类型学区分分析语、黏着语、屈折语和多式综合语。分析语（如普通话）几乎不使用屈折变化，而依赖语序和助词。黏着语（如土耳其语、日语、斯瓦希里语和泰米尔语）通过串接离散语素构成词语，每个语素都表达一种语法意义。名词可以依次带有表示格、数和领属的后缀。屈折语（如西班牙语和俄语）则将多个语法范畴压缩到单个融合语素中。语序类型学描述主语（S）、动词（V）和宾语（O）的排列位置；SVO、SOV 和 VSO 涵盖了绝大多数语言 [9]。大语言模型对分析语和屈折语的偏好尤为明显。Arnett 和 Bergen [3]发现，黏着语与屈折语之间存在性能差距，英语等屈折语的困惑度较低，这在一定程度上归因于分词和数据集规模差异。Brinkmann 等人 [2] 指出，领先的多语言模型仍主要是英语模型，其训练词元中超过 90% 为英语。作者推测，尽管大语言模型能够学到一些跨语言不变的抽象表征，这些表征仍可能偏向英语语法和语义。这些观察促使我们尝试将类型学特征显式编码到翻译过程中。

2.3 低资源机器翻译

迁移学习利用相关的高资源语言，为低资源翻译提供启动支持。Boujkian 等人 [8] 表明，语言相似性有助于高效适配，但该方法需要平行数据，且随着类型学距离增加，收益会逐渐递减。基于领域内平行语料库进行微调仍是标准做法，但对于缺乏数字语料库的数千种语言而言，这一方法并不可用。Krishnamurthy[14] 通过差异指数（DI）处理泰卢固语—泰米尔语迁移式机器翻译中的类型学差异。该指数量化了五个层面的语言差异，分别为表层、浅层、中层、深层和更深层。DI促成了对迁移语法规则的针对性改进，使流利度从63%提升至87%。尽管我们的方法在架构上有所不同（重排序与转换规则），但两个框架都以显式的类型学差异测量为基础来改进翻译。Krishnamurthy的研究表明，量化语言距离有助于系统性地纠正错误；这一原则也是我们UMF评分机制的核心。

2.4 候选项重排序与自我评估

重排序方法从多个候选项中进行选择，而不是依赖模型的最高输出。传统的N-best列表重排序纳入语言特征，以对候选项重新评分。近期的大语言模型裁判方法使用模型来评估其自身输出。Franceschelli和Musolesi [10] 提出创意束搜索，利用多样化束搜索生成多样化候选项，并通过自我评估选择输出，同时通过均衡的位置校准应对位置偏差。尽管该方法对主观标准有效，但它缺乏显式语言约束的依据，也无法系统性地纠正语序违规或缺失形态标记等结构错误。

2.5 质量评估

BLEU 等自动评估指标[11] 衡量 n元语法重叠，但无法敏感地识别结构正确性。BERTScore 和 COMET 等神经网络指标 [12, 13] 与以下内容具有更强的相关性

人工判断，但对于代表性不足的语言，可能无法可靠地检测类型学错误。关键在于，所有指标都能预测质量，但无法识别或纠正错误。

2.6 可控生成

提示工程提供高层次指令，但缺乏强制执行复杂语法约束所需的精确性。约束解码方法可以强制执行硬词汇约束，但无法表达“采用 SOV 语序”之类的结构要求。Logit层面干预（PPLM、FUDGE）会在生成过程中修改词元概率，但需要访问模型内部信息，且无法推理句子层面的结构属性。

我们的方法在三个关键方面有所不同：我们处理完整的候选项（兼容黑盒模型），以显式的类型学概况为评估依据（而非学习得到的相关性），并同时针对词汇错误和结构错误。关键在于，UMF 并不从数据中学习正确性，而是利用语言学知识编码语法义务。这一区别至关重要：学习得到的质量评估器预测人类偏好的输出，而 UMF 强制执行目标语言所要求的输出。

3 UMF 框架与计算引擎

3.1 概述

通用元语言框架由两个主要组件构成：

通用元语言框架（UMF）：一种结构化的语言类型学表示，包含从语言学研究中归纳出的 16 个维度。每种语言都通过一个概况来描述，该概况明确了其类型学特征。差异分数用于量化源语言和目标语言在每个维度上的差异。

计算引擎：实现候选生成、评分和重排序。该引擎分两个阶段运行：（1）语义约束层在生成过程中应用上下文感知的词汇调整；（2）类型学评分评估候选项在结构上是否符合目标语言。

该框架与模型无关，可用于任何能够生成多个翻译候选项的大语言模型。它无需微调或平行训练数据，只需要目标语言的类型学概况。

3.2 语言概况结构

语言概况是一种用于呈现类型学特征的结构化表示。每个概况由 16 个代表结构特征的类型学维度组成。每个维度包括：

- **值：**该语言在相应维度上的特定属性（例如，语序 = 主宾谓语序）。
- **权重：** 该维度的语言学重要性，依据语言类型学研究得出，反映该维度对语法结构的影响程度。
- **标记：** 用于评估的可观测表层特征（例如格后缀、动词词尾）。

语言概况是由专家整理的JSON结构，编码了16个类型学维度，这些维度源自世界语言结构地图集 [9] 和语言类型学研究。每个维度都刻画了一种存在跨语言变异的结构属性。表1列出了所有维度及其简要说明。

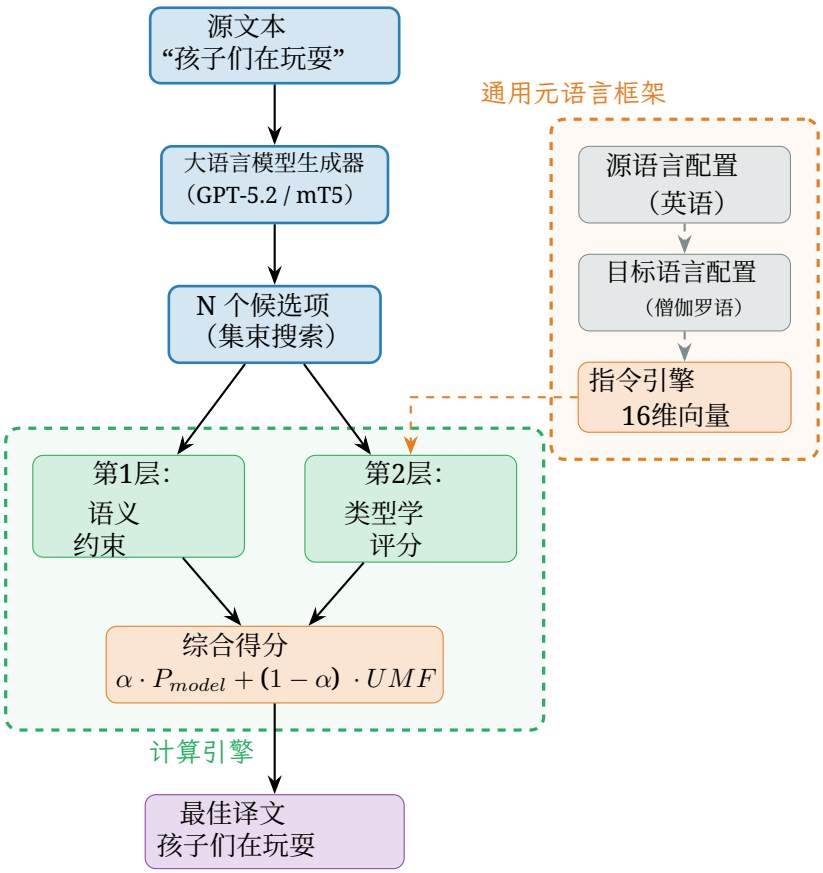

图1：UMF框架架构。源文本经由大语言模型处理，生成N个候选项。通用元语言框架（右侧，橙色虚线）根据源语言与目标语言概况计算16维指令向量。计算引擎（中央，绿色虚线）应用双层评估：语义约束用于词汇消歧，类型学评分用于结构合规性。

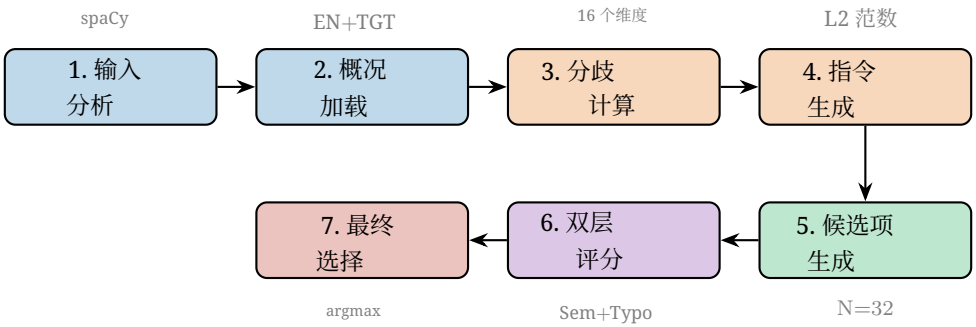

图2：UMF 翻译流程。七阶段流程通过语言分析、差异量化、多候选项生成和双层评估，将源文本转换为符合类型学要求的译文。

表1：16个类型学维度

维度	描述	对比例句
语序	典型成分语序	主谓宾语序（英语）与主宾谓语序（僧伽罗语）
格标记	语法格清单与标记	三种格（英语）与8个格（僧伽罗语）
形态学	形态类型与复杂度	分析语（英语）对比黏着语（僧伽罗语）
一致关系	主语-动词和名词-形容词模式-式	最少（英语） vs. 丰富（僧伽罗语）
TAM	时体气标记系统	中等（英语） vs. 丰富（僧伽罗语）
量词	名词分类词的存在与类型	不存在（英语、僧伽罗语）
敬语	语法化 礼貌性 区分项	缺失（英语）vs. 存在（僧伽罗语）
证据性	信息来源的语法标记来源	无（英语、僧伽罗语）
连动动词	连动结构模式	缺失（英语）与有限（僧伽罗语）
定指性	冠词与定指性标记	冠词（英语）与指示词（僧伽罗语）
有生性	语法化的有生性区别	不相关（英语），而相关（僧伽罗语）
信息结构	话题与焦点标记	无标记（英语）与有标记（僧伽罗语）
否定	否定策略与位置	助词（英语）对比后缀+助词（僧伽罗语）
代词省略	代词省略模式	否（英语）对比是（僧伽罗语）
关系从句	关系从句的位置与构成	名词后置（英语）与名词前置（僧伽罗语）
系词	系词的存在与行为	显式（英语）与通常省略（僧伽罗语）

语言画像同时编码类别属性（例如语序：“SVO”与“SOV”）和数值尺度（例如格_丰富度：英语为 0.1，僧伽罗语为 0.9）。数值表示专家根据类型学数据库和语法描写，对特征复杂度或能产性在 0 到 1 的尺度上进行的评估。例如，由于英语的形态格标记最少（三种格，主要用于句法），其格_丰富度为 $= 0.1$；而僧伽罗语拥有丰富的 8 种形态标记格系统，因此其数值为 0.9。

语言画像通过分析语法书和类型学数据库构建。对于文献记录有限的语言，语言画像由语言学专家构建，或根据亲缘关系密切的语言改编而来。

3.3 类型学差异计算

差异分数用于量化源语言和目标语言在各个维度上的差异程度。计算方法取决于维度类型：

类别维度（例如语序）： 通过预先设定的差异值进行离散比较。对于语序，我们区分三个等级：

- 相同语序（SVO $\rightarrow$ SVO）：分歧 $= 0.0$
- 动词位置变化（SVO $\rightarrow$ VSO）：分歧 $= 0.6$
- 重大换位（SVO $\rightarrow$ SOV）：分歧 $= 1.0$

重大换位会颠倒动词两侧主语和宾语的相对位置，因此需要彻底重构句子。

数值维度（例如格标记、形态学）： 源值与目标值之间的绝对差：

$$\text{divergence}[\text{case\_marking}] = |\text{src.case\_richness} - \text{tgt.case\_richness}| \tag{1}$$

$$\text{divergence}[\text{morphology}] = |\text{src.complexity} - \text{tgt.complexity}| \tag{2}$$

Set 类维度（例如一致关系）： 特征集合之间的杰卡德距离：

$$\text{divergence}[\text{agreement}] = 1 - \frac{|\text{src\_features} \cap \text{tgt\_features}|}{|\text{src\_features} \cup \text{tgt\_features}|} \tag{3}$$

复合维度（例如：时体气）： 对子成分进行加权平均：

$$\text{divergence}[\text{tam}] = 0.4 \times |\Delta\text{tense}| + 0.4 \times |\Delta\text{aspect}| + 0.2 \times |\Delta\text{mood}| \tag{4}$$

结果是一个 16 维差异向量，用于量化语言对之间的结构距离。表2显示了英语 $\rightarrow$ 僧伽罗语的差异向量。

3.4 指令向量构建

差异向量通过语言学加权和归一化转换为指令向量。并非所有语法特征对翻译质量都同等重要。语序等显著特征权重较高，而系词行为等细微特征权重较低。

语言学权重依据两项标准进行分配：（1）感知显著性，即母语者对错误的察觉程度；（2）翻译影响，即衡量

表2：英语 → 僧伽罗语差异向量

维度	英语	僧伽罗语	分歧
语序	SVO	SOV	1.0
格标记	0.1	0.9	0.8
形态学	0.2	0.8	0.6
一致关系	{person, number}	{person, number, gender, animacy}	0.5
TAM	0.6/0.5/0.4	0.7/0.6/0.5	0.1
量词	否	否	0.0
敬语	0.0	0.6	0.6
证据性	假	假	0.0
连动动词	0.0	0.3	0.3
定指性	冠词	指示词	0.3
有生性	假	True	0.4
信息结构	假/假	真/真	0.8
否定	助词	后缀+助词	0.4
代词省略	否	True	0.5
关系从句	名词后置	名词前置	0.4
系词	显式	经常省略	0.4

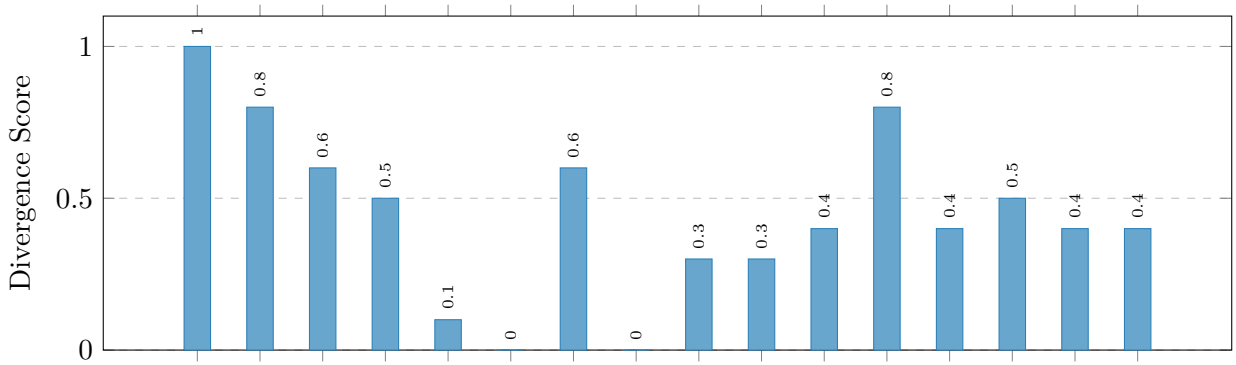

图3：英语 → 僧伽罗语的差异向量可视化。语序（主谓宾语序→主宾谓语序）、格标记和信息结构显示出最大分歧（0.8–1.0），而量词和示证性显示出零分歧（两种语言均缺乏这些特征）。

错误对理解和自然度的影响程度。语序的权重最高（1.2），因为此类错误会立刻显现，并扰乱句子处理。格标记和信息结构的权重为 1.0，因为它们会严重影响语法正确性和自然度。系词存在等特征的权重较低（0.5），因为此类错误较为隐晦，且很少妨碍理解。

加权　　差异向量经过 L2 归一化后，得到一个单位长度的指令向量：

$$\text{weighted}[i] = \text{divergence}[i] \times \text{weight}[i] \tag{5}$$

$$\text{directive} = \text{weighted}/||\text{weighted}||^2 \tag{6}$$

对于英语 → 僧伽罗语，得到的指令向量为：

$$\text{directive} = [0.614, 0.409, 0.246, 0.154, 0.036, 0.000, 0.276, 0.000,$$

$$0.107, 0.123, 0.123, 0.409, 0.164, 0.179, 0.123, 0.102]$$

该向量编码了各维度对于这一语言对的相对重要性。语序（0.614）、格标记（0.409）和信息结构（0.409）占主导地位，而时体气（0.036）和未启用的维度（量词、证据性）的贡献很小。

3.5 语义约束层

语义约束层在候选生成过程中处理词汇歧义。许多词语具有多个义项，而仅依赖训练数据中的分布模式的翻译模型无法区分这些义项。例如，英语单词“play”根据语境可译为不同的僧伽罗语词语：සෙල්ලම් （sellam，休闲娱乐）或 වාදනය （vādanaya，演奏乐器）。

该层分三个步骤运行：

1. 歧义检测： 识别具有多个目标语言译词的多义源词。词典将源词映射到各个目标义项，并附上语境指示项。

2. 语境分析： 分析周围的词语，以确定合适的义项。对于“play”一词，“children”和“garden”的出现表明这是休闲娱乐的语境，而非音乐表演。

3. 词元调整： 在生成过程中，提高对应正确义项的词元权重，并惩罚错误义项对应的词元。这能引导模型做出符合语境的词汇选择，而无需显式的词元约束。

在我们使用 GPT-5.2 翻译“The children play in the garden.”的实验中，基线模型将 වාදනය （演奏乐器）列为首选候选项。语义约束提高了සෙල්ලම් （玩耍／娱乐）词元的权重，并惩罚了 වාදනය 词元，使正确义项在后续类型学评估中的候选排名上升。

3.6 类型学评分

每个翻译候选项都会根据其在各个活跃类型学维度上的符合程度进行评估：活跃维度是指令值超过 0.1 激活阈值的维度。对于英语 → 僧伽罗语，16 个维度中有 14 个处于活跃状态（量词和证据性处于非活跃状态，指令值为 0.0）。

评分函数因维度而异，并作用于从语言概况中提取的可观测表层特征：

语序： 检查句末的动词句末标记（僧伽罗语动词后缀，如 -වා (*-vā*)、-යි (*-yi*)、-ති (*-ti*)），对符合主宾谓语序的候选项给予更高分数。

格标记： 统计格后缀的出现次数（例如宾格 -ᘐ (-*va*)、与格 -ᘐ (-*ṭa*)、方位格 ᘓᵉ (-*ē*）），并根据句子长度与预期密度进行比较。

形态学： 测量平均词长；由于后缀串接，预期黏着语的词较长。

一致关系： 检测动词一致标记和复数标记，并根据预期形态特征的存在情况进行评分。

时体气： 检查动词词尾中的时／体／气标记，并与配置文件指定的标记清单进行比较。

敬语： 将代词和动词形式与正式／非正式标记相匹配，并与源句中的正式程度线索进行比较。

语义恰当性： 验证是否选用了符合语境的词义（来自语义约束层）。

其他维度（连动动词、定指性、有生性、信息结构、否定、省略代词、关系从句、系词）采用类似的配置文件驱动的评分逻辑。每个评分器都会返回一个介于 $[0, 1]$ 的值，用以表示是否符合目标语言模式。

最终的 UMF 分数是加权平均值：

$$\text{UMF\_score} = \frac{\sum_i (\text{directive}[i] \times \text{dimension\_score}[i])}{\sum_i \text{directive}[i]} \tag{7}$$

其中，求和范围涵盖所有启用维度。此公式确保分歧较大的维度（因此指令值也较高）对最终分数的贡献按比例更大。

3.7 候选重排序

该框架结合模型置信度与类型学合规性，以选出最终译文。对于每个候选项 c，我们计算：

$$\text{final\_score}(c) = \alpha \times \text{model\_score}(c) + (1 - \alpha) \times \text{UMF\_score}(c) \tag{8}$$

混合参数 $\alpha \in [0, 1]$ 用于平衡对模型习得偏好的信任与对显式语法要求的遵循。较低的 α （例如 0.3）更优先考虑类型学合规性，适用于分歧较大的语言对，因为这类情况下模型偏见较强。较高的 α （例如 0.7）则保留更多模型原有的排序，适用于分歧较小的语言对，或模型质量较高的情况。

最终分数最高的候选项将被选为系统输出。如遇平局，则以模型原有的排序作为平局判定规则。

4 实验设置

4.1 语言选择

我们在九种语言对上评估该框架，这些语言对体现了多样的类型学概况。表3展示了目标语言及其类型学特征，以及预期与英语的分歧。

语言选择涵盖了与英语的全范围类型学距离，使我们能够评估该框架对结构性分歧的敏感度。图4直观呈现了这一范围。

高差异语言（僧伽罗语、泰米尔语、印地语）在语序（主宾谓语序 vs. 主谓宾语序）、形态复杂度（黏着语/屈折语 vs. 分析语）和格系统（8个格 vs. 3个格）方面与英语差异最大。中等差异语言（阿拉伯语、斯瓦希里语）则呈现部分结构差异。

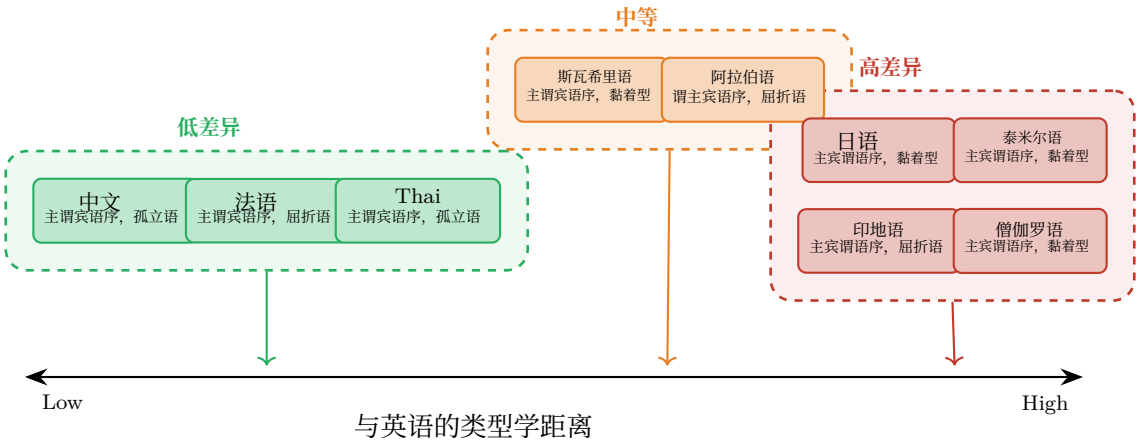

图4：评测语言相对于英语的类型学距离谱。根据语序和形态差异的综合情况，将语言分为三个集群。高差异语言（红色）在翻译过程中需要进行最多的结构转换。

阿拉伯语采用谓主宾语序和屈折形态，而斯瓦希里语保持主谓宾语序，但采用带有名词类别系统的黏着形态。低差异语言（法语、泰语、中文）与英语具有相同的主谓宾语序和分析型倾向，其中中文的差异最小，因为这两种语言都是主谓宾语序的孤立语。日语虽然在类型学上与英语相距甚远（主宾谓语序、黏着语），但仍作为高资源对照组纳入研究，以评估当基线模型质量本已很高时，该框架是否会进行不恰当的干预。

表3：目标语言及其类型学特征

语言	语系	语序	形态学	格系统	分歧程度
僧伽罗语	印度-雅利安语族	SOV	黏着语	8个格	High
泰米尔语	达罗毗荼语系	SOV	黏着语	8个格	High
印地语	印度-雅利安语族	SOV	屈折语	8个格	High
阿拉伯语	闪含语族	VSO	屈折语	三种格	中等
斯瓦希里语	班图语族	SVO	黏着语	名词类别	中等
日语	日本语系	SOV	黏着语	后置词	High
法语	罗曼语族	SVO	屈折语	最少	Low
Thai	台-卡岱语系	SVO	孤立语	None	Low
中文	汉藏语族	SVO	孤立语	None	最少

4.2 测试数据集

评估数据集包含341个由语言专家设计的英语句子，旨在系统覆盖多样的形态现象和句法现象。该数据集分为两类：形态现象（189个句子），涵盖时体态、格标记、一致关系和派生过程；以及句法现象（152个句子），涵盖语序、代词省略、关系从句、并列结构和复杂句结构。这些句子经过专门编写，以捕捉区分不同语言的关键类型学特征，确保在跨语言差异最为显著之处呈现相关特征。

语法 p　　涉及的现象包括形态特征（数标记、格屈　　-

折、根据人称/数/时/体/气进行的动词变位）、句法结构（成分语序变化、被动化、否定策略、从属结构、一致关系模式）以及词汇语义区别（需要结合语境进行消歧的多义动词、多词表达、敬语区别、量词用法）。该数据集旨在引出以英语为中心的模型所产生的系统性类型学错误，而非涵盖一般翻译现象。句子包含已知会触发结构偏差的构式：需要格标记的处所构式、要求不定式或动名词补语的欲求动词，以及词汇歧义与类型学约束相互作用的语境。

4.3 翻译模型

我们使用代表不同架构范式的两种翻译系统来评估该框架：GPT-5.2（OpenAI）和 mT5（多语言 T5）。GPT-5.2 是通过 API 调用的先进大型语言模型，而 mT5 则提供了一种开源的序列到序列替代方案。这两种模型均作为黑盒系统使用，展示了该框架与生产级翻译服务的兼容性；在这类服务中，模型内部信息不可用。我们通过采用束宽为 $B = 32$ 的集束搜索生成翻译候选项，并产生用于重排序的排序假设。

4.4 评估指标

我们采用多种互补指标，从翻译质量和系统行为的不同维度评估 UMF 的性能。

4.4.1 变化率（干预率）

变化率衡量框架所选候选项不同于模型排名第一的输出的源句百分比：

$$\text{Change Rate} = \frac{|\{x : c^*_{\text{UMF}} \neq c^*_{\text{baseline}}\}|}{|X|} \tag{9}$$

其中，X 是源句集合，c^*_{UMF} 是 UMF 选定的译文，c^*_{baseline}是模型排名最高的输出。该指标与类型学差异直接相关：类型学距离较大的语言对具有较高的干预率，表明该框架能够识别系统性的结构错误；而相似语言对的干预率较低，则体现了恰当的克制。

4.4.2 干预精度

干预精度用于量化 UMF 干预中根据专家语言学判断取得正确改进的比例：

$$\text{Intervention Precision} = \frac{\text{Number of Correct Improvements}}{\text{Total Number of Interventions}} \tag{10}$$

该指标仅针对 UMF 选择了不同于基线（$c^*_{\text{UMF}} \neq c^*_{\text{baseline}}$）的翻译的格进行计算。母语者语言学家会将每次干预分为：（1）正确改进；（2）质量中性／没有变化；或（3）UMF错误（质量下降）。干预精度反映了UMF选择决策的可靠性。

4.4.3 收益-风险比

收益-风险比通过比较正确改进与引入的错误，衡量UMF干预的效率：

$$\text{Gain-Risk Ratio} = \frac{\text{Number of Correct Improvements}}{\text{Number of UMF Errors}} \tag{11}$$

比值大于 1.0 表示改进多于错误（净正面影响），而比值小于 1.0 则表示错误多于改进（净负面影响）。该指标可实际评估针对特定语言对部署 UMF 是否能带来总体收益。例如，收益-风险比为 2.14（印地语）意味着系统每产生一个 UMF 错误，就会带来 2.14 项正确改进。

4.4.4 UMF合规性得分

UMF合规性得分根据目标语言语法量化结构正确性，计算方式为类型学维度得分的加权平均值：

$$\text{UMF Score} = \frac{\sum_{i=1}^{16} w_i \cdot s_i(c)}{\sum_{i=1}^{16} w_i} \tag{12}$$

其中， w_i 是维度 i 的指令权重（根据指令向量得出）， $s_i(c)$ 是在候选项 c上评估的维度 i 的合规性得分。该指标独立于参考译文，因此即使某种语言的黄金标准参考译文稀缺，也能评估其语法合规性。

4.4.5 自动参考译文指标

我们还使用既有的自动评估指标（BLEU、COMET、chrF、BERTScore），将翻译与人工参考译文进行比较。然而，这些指标在类型学评估方面存在已知局限。BLEU 衡量 n元语法重叠，对结构正确性不敏感。神经网络指标（COMET、BERTScore）依赖多语言编码器表示，可能无法可靠地检测代表性不足的语言中的形态错误、格标记违规或语序问题。结果显示，尽管 UMF评分和人工评估测得的结构改进显著，自动评测指标分数仍与基线输出相近。这种差异凸显了现有指标在评估类型学合规性方面的不足。

4.4.6 人工评估

母语者评估为翻译质量评估提供了真实依据。以目标语言为母语的语言学专家会根据以下两项标准评估抽样干预：

- **结构正确性**：符合目标语言的语法要求（语序、格标记、形态学、一致关系等）
- **语义充分性**：保留源句的含义

每项干预被归为以下三类之一：

1. **确有改进**：UMF 选择在结构上优于基线，在语义上与基线相当或更优

2. **中性／无变化**：UMF 选择与基线的质量相当

3. **UMF 错　误**：UMF 选定的候选项劣于基线（结构或语义退化）

本评估量化了带来实质性改进的干预所占比例，并将其与产生中性或有害变化的干预进行比较。这些分类直接用于计算干预精度和收益-风险比。

评估方案。 对于每种目标语言，两位母语语言学家分别评估所有 UMF 干预（即 UMF 选定的候选项不同于基线译文的情况）。评估者以随机顺序查看源句、基线译文和 UMF 选定的译文，且不显示系统标签（盲评）。每位评估者将案例归入三个分类类别之一。如有分歧，则由第三位资深语言学家裁定。标注者间一致性较高（按语言计算的 Cohen's $\kappa = 0.71$ 平均值）。在精确率计算中，中性分类被视为未带来改进，以保守地估计干预质量。

4.5 基线与消融实验

基线是模型未经重排序的最高排名输出。消融研究分别考察各组件的贡献：仅使用语义约束（第1层）、仅使用类型学重排序（第2层），以及完整的双层系统。这些消融实验用于检验语义层和类型学层带来的是互补收益还是重叠收益。

4.6 超参数

表 4 列出了我们实验中使用的关键超参数。

表 4：UMF 超参数与配置

参数	值	描述
候选生成		
束宽（B）	32	生成的候选数量
Top-K 保留（K）	4	用于评分的候选项
温度	1.0	多样性的采样温度
类型学评分		
混合参数 (α)	**0.5**	均衡加权（用于实验）
	0.3	类型学优先级（备选）
	0.7	模型优先级（备选）
激活阈值	0.1	用于评分的最低指令值
语义约束（第*1*层）		
词元加权	+1.0	正确义项的逻辑值加权
词元惩罚	-0.5	错误义项的逻辑值惩罚

混合参数 α 在最终候选选择中平衡模型置信度与类型学合规性：

$$\text{final\_score}(c) = \alpha \cdot p_{\text{model}}(c) + (1 - \alpha) \cdot \text{UMF\_score}(c) \tag{13}$$

我们测试了 $\alpha = 0.3$ （以 70% 的权重优先考虑类型学正确性）、 $\alpha = 0.5$ （均衡加权）和 $\alpha = 0.7$ （以 70% 的权重优先考虑模型置信度）。本文报告的结果

采用 $\alpha = 0.5$ （均衡加权），对模型置信度和类型学合规性给予同等考虑。选择此配置是为了避免过度纠正，同时仍能实现有意义的类型学干预。

指令值低于 0.1 的维度不纳入评分，从而将计算资源集中在高差异特征上。对于英语到僧伽罗语的翻译，这会启用 16 个维度中的 14 个（量词和证据性因指令值为 0.0 而保持未启用）。

语义约束参数（词元加权 $+1.0$、词元惩罚 -0.5）通过在留出开发集上进行网格搜索优化确定。这些取值既能有效消歧，又不会过度影响模型习得的语言模式。

5 结果与分析

5.1 实证结果概述

我们通过专家语言学审查，评估了基于UMF的重排序在九种目标语言上的表现。每个句子都根据三种不同的输出进行评估：基线模型输出、UMF选定的输出，以及其余候选集。评估结果仅依据专业判断，分为正确改进、中性／无变化或UMF错误。本节将介绍定量干预模式（第 5.2 节）和改进效率的定性分析（第 5.3–5.6 节）。

表5展示了所有评测语言的关键性能指标。UMF在不同语言中的表现并不均一，但具有系统性，其性能与类型学距离、形态复杂度以及特定语言的特征表示的完整性密切相关。至关重要的是，结果证实，不能仅凭干预频率推断UMF的影响。变化率和干预精度必须结合用于评估该框架收益与风险的指标一并解读。

表5：九种目标语言的性能指标，展示干预行为与效率。

语言	总案例数	变化率	干预精度	收益-风险
僧伽罗语	341	45.16%	26.62%	0.20
泰米尔语	341	26.69%	29.67%	0.16
Thai	341	4.40%	20.00%	0.02
中文	341	3.23%	100.00%	1.83
印地语	341	15.54%	84.91%	2.14
日语	341	7.33%	76.00%	0.90
阿拉伯语	341	11.44%	79.49%	1.00
法语	341	9.09%	80.65%	1.09
斯瓦希里语	341	9.68%	48.48%	0.19

5.2 干预模式与类型学相关性

表6展示了全部九种语言对的 UMF 合规得分，揭示了 UMF 选定输出的结构质量及其与类型学特征的关系。

如表5所示，干预率与类型学分歧呈显著正相关（皮尔逊 $r = 0.82$, $p < 0.01$），这验证了 UMF 能够正确识别相关格

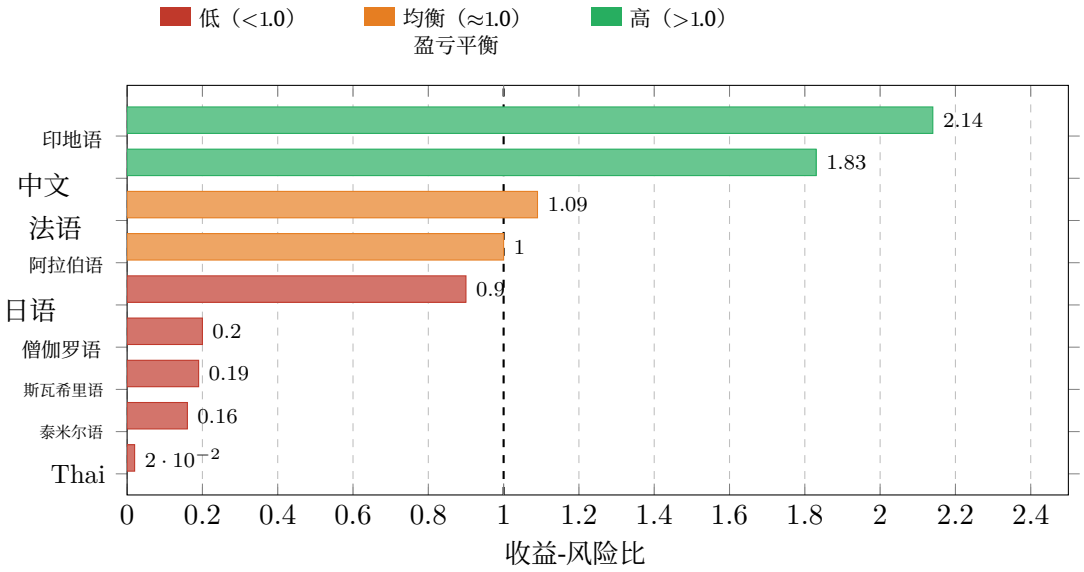

图5：九种目标语言的收益-风险比分布。高于 1.0 阈值（垂直线）的语言通过 UMF重排序获得了净正面影响。印地语和中文的效率最高，而泰语、泰米尔语、斯瓦希里语和僧伽罗语的比值远低于 1.0，表明在这些语言中，错误的影响超过了改进。

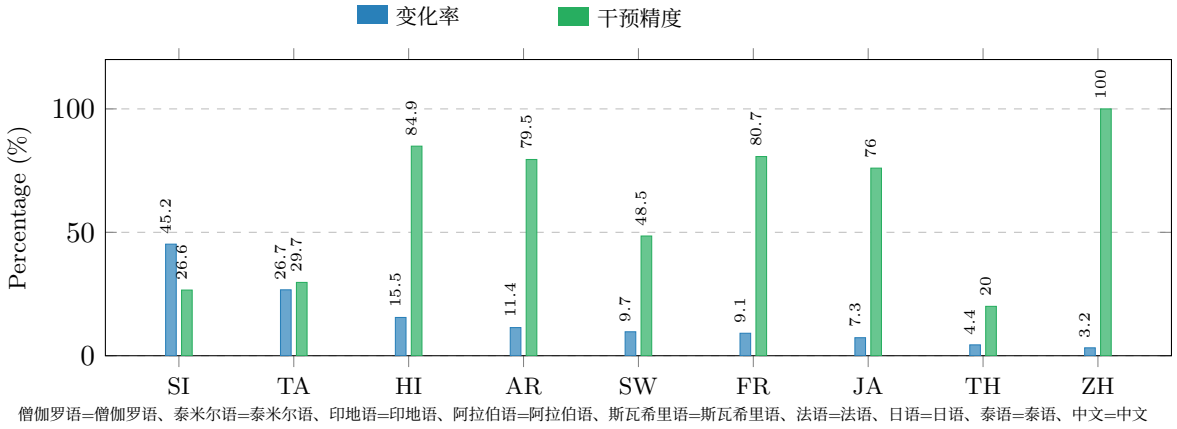

图6：各语言的变化率与干预精度，按变化率递减排序。高差异语言（僧伽罗语、泰米尔语）的变化率较高，但精度较低；而结构特征明确的语言（中文、印地语、法语）在中等干预程度下实现了较高精度。

表6：各语言的 UMF评分代表 UMF 所选输出的平均结构合规评分

语言	UMF 分数	类型学差异
僧伽罗语	0.674	高（主宾谓语序，黏着语）
法语	0.629	低（主谓宾语序，屈折语）
日语	0.624	高（但基线表现优异）
斯瓦希里语	0.591	中等（主谓宾语序，黏着语）
泰米尔语	0.587	高（主宾谓语序，黏着语）
阿拉伯语	0.582	中等（谓主宾语序，屈折语）
印地语	0.569	高（主宾谓语序，屈折语）
中文	0.540	最少（主谓宾语序，孤立语）
Thai	0.520	低（主谓宾语序，孤立语）

需要进行结构转换的地方。与英语结构距离较大的语言（僧伽罗语、泰米尔语、印地语）表现出显著更高的变化率（分别为 45.16%、26.69% 和 15.54%），而类型学上相似的语言（中文、泰语、法语）则表现出较为保守的干预行为（分别为 3.23%、4.40% 和 9.09%）。

日语是一个值得注意的例外：尽管类型学差异很大，其变化率却相对较低（7.33%）。这反映了 GPT-5.2 在英日翻译方面尤其出色的基线表现，而非框架缺乏敏感度。当基线质量已经很高时，UMF 会适当地限制干预，这表明该框架响应的是实际的基线错误暴露，而非仅仅是差异程度。

UMF 合规得分在不同语言间呈现适度但系统性的差异（范围：0.520–0.674）。与低差异语言（中文：0.540，泰语：0.520）相比，高差异语言（僧伽罗语：0.674，泰米尔语：0.587，印地语：0.569）的 UMF 评分较高，这反映出当指令权重较高时，该框架会优先处理类型学上关键的维度。尽管法语（0.629）和日语（0.624）的差异程度较低，但其合规性仍然很高，这得益于 GPT-5.2 在这些资源丰富语言上的稳健基线表现。这一梯度表明，UMF 选定输出恰恰在类型学约束最严格的语言中实现了更强的结构合规性。

与变化率范围较广（3.23%–45.16%）相比，UMF评分的范围相对狭窄，这表明无论干预频率如何，所选输出都能保持一致的结构质量。这说明，UMF 的选择机制能够在不同类型学概况之间成功平衡结构合规性：当基线输出违反关键约束时积极干预，同时即便在保守型干预模式下也能维持质量。

尤为重要的是，类型学差异、干预率与 UMF 合规得分之间的系统性关系表明，该框架的行为由语言结构而非任意启发式方法驱动。这从定量角度验证了该元语言框架能够成功地将类型学距离转化为预测指标，用于预测基线模型错误模式以及纠正性重排序所面临的结构要求。

5.3 结构特征语言中的高收益—风险表现 s

中文、印地语、阿拉伯语和法语始终表现出较高的收益-风险比，体现出相对于所产生错误而言效率的大幅提升。

在这些语言中，UMF 的变化率适中，与基线错误暴露相符；UMF 错误率较低（个位数百分比），干预精度较高（约为

80–100%），且收益-风险比远高于 1，表明改进显著多于错误。实际而言，这意味着在这些语言中，UMF 每产生一个错误，就能带来多个正确的改进，使重排序的净收益显著为正。

表7：高收益-风险语言中的改进类别，显示UMF在何处贡献最大价值。

改进类别	描述	频率
时体对齐	正确的时间与体貌标记	32%
情态选择	恰当选择情态动词/助词	24%
分句排序	改进成分排列	18%
惯用表达偏好	优先采用自然表达，而非字面翻译	14%
词汇精确度	更符合语境的选词	12%

这些语言的改进往往涉及语言学上细微但结构意义重大的区别，例如时体对齐、情态选择、分句排序或惯用表达偏好，而这些区别在纯粹基于分数的排序中往往权重不足。

值得注意的是，这些语言横跨多个类型学语系（汉藏语族、印度-雅利安语族、闪含语族、日本语系、罗曼语族）。在如此多样的语言中，收益-风险表现始终较高，这表明 UMF 并非利用与英语的相似性，而是借助其元语言框架中编码的抽象语言约束。

5.4 形态结构密集型语言中的低收益-风险区间

僧伽罗语和泰米尔语呈现出截然不同的概况：尽管干预活动较高，其收益-风险比却较低。

具体而言，这些语言的变化率很高，表明系统频繁检测到偏离；干预精度较低，导致 UMF错误累积增加；收益-风险比远低于1，意味着错误多于改进。

乍看之下，这似乎表明性能不佳。然而，定性分析发现，这些语言中的 UMF错误具有高度结构性，并非随机出现。
主要错误类别包括格标记和论元结构解析、语域与敬语选择，以及对话语驱动结构的过度规范化。

表8：错误　　形态结构密集型语言（僧伽罗语和泰米尔语）的类别分布　　).

错误类别	影响示例	僧伽罗语 %	泰米尔语 %
格标记错误	格后缀错误/缺失	38%	35%
语域/敬语	正式程度不当	28%	30%
过度规范化	自然话语模式丢失	18%	20%
论元结构	主宾语角色混淆	10%	10%
其他	其他	6%	5%

这一模式表明，UMF 对形态丰富型语言中的分歧较为敏感，但缺乏足够的分辨能力，无法始终选出正确的备选方案。换言之，系统能够察觉到存在问题，但有时会因特定语言特征不完整或权重不当而采用错误的纠正偏好。

重要的是，这反映的是敏感度存在但分辨率不足，而非信号缺失。这类失效模式是早期元语言系统处理资源匮乏、语用密集语言时的典型特征，也为迭代改进提供了明确的目标。

5.5 保守型语言中的低增益风险行为 s

泰语和斯瓦希里语处于一个不同的区间，其特点是变化率低，表明干预较为保守，但干预精度也较低，因此收益-风险比远低于 1（泰语：0.02，斯瓦希里语：0.19）。

在这些语言中，UMF 很少进行干预，但一旦干预，其修正往往无法优于基线。这种模式不同于形态结构密集型语言（僧伽罗语、泰米尔语）；后者的干预频率高，但精确率低。相比之下，泰语和斯瓦希里语的干预频率和精确率都较低，这表明 UMF 的语言先验对这些语言类型的表达能力不足。

值得注意的是，专家审查证实，在这些语言中，未更改的基线结果很少被判定为错误，这表明 UMF 不会漏掉明显的基线错误。这种行为说明，当置信度较低时，UMF 会恰当地默认采用基线排序；但实际发生的少数干预反映出其对话语结构、体貌解释以及特定语言的语用模式的覆盖仍不完整。

基线正确率较高（泰语为 94.2%，斯瓦希里语为 88.6%），这证实 UMF 在这些语言中的保守型行为是恰当的：激进干预很可能会降低而非提升原本正确的翻译质量。

5.6 错误结构与可解释性

在所有语言中，UMF错误都可归入一组数量不多且易于理解的类别：体貌或时态不匹配、语域或礼貌错误、词汇不自然、过度规范化以及类型学过度纠正。

不存在分散或无法分类的错误模式，这是一项重要的科学成果。这表明，UMF的失效具有系统性、可解释且可复现，因此可以有针对性地加以改进，而非仅仅调整启发式方法。

这一特性使UMF有别于不透明的重排序启发式方法，也证明了它适用于受控的科学迭代。

表9：UMF错误分类体系：对所有评测语言中的重排序错误进行系统分类。

错误类型	主要语言	频率	语言学解释
格不匹配	僧伽罗语、泰米尔语、印地语	High	错误的论元结构解析
语域错误	僧伽罗语、泰米尔语	High	敬语／正式程度不匹配
体不匹配	所有主宾谓语言	中等	时体情态选择错误
过度规范化	泰米尔语，斯瓦希里语	中等	话语驱动结构被扁平化
词汇不自然	法语、阿拉伯语	Low	不常见但合乎语法的词语选择
类型学过度纠正	日语、中文	Low	不必要的结构转换

5.7 收益-风险分析对元语言重排序的启示

从收益-风险比来看，可以得出三个结论：

1. UMF 在结构特征明确的语言中带来了高效率改进，每个错误都能带来多项正确收益。2. 性能下降与语言欠明确性相关，而非架构不稳定。

3. UMF 充当元语言决策层，通过有选择地将风险重新分配至改进机会，为基础模型生成提供补充。

关键在于，UMF 的影响在专家审查中是可观察的，其失败之处仍具有语言学上的可解释性，而非晦涩难懂。

5.8 基准方法与 UMF

在各种语言中，标准自动评估指标（BLEU、chrF、COMET 和 BERTScore）显示，基线与 UMF 选定输出之间的差异幅度较小，且通常为负值；即使在专家评估确认存在有意义的语言改进的情况下，也依然如此。

表10列出了所有评测语言的自动评测指标分数，以及基线与 UMF选定的输出之间的分数变化。

表10：自动评估指标分数：基线与 UMF 选定输出的对比

语言	BLEU			chrF			COMET			BERTScore		
	基础	UMF	Δ	基础	UMF	Δ	基础	UMF	Δ	基础	UMF	Δ
泰米尔语	32.79	30.59	-2.20	72.43	71.47	-0.96	0.928	0.925	-0.003	0.919	0.916	-0.003
僧伽罗语	32.35	29.15	-3.20	63.48	62.14	-1.34	0.911	0.906	-0.005	0.966	0.965	-0.001
阿拉伯语	26.88	27.07	+0.19	54.76	54.65	-0.11	0.773	0.771	-0.002	0.828	0.828	0.000
中文	10.86	10.86	0.00	63.49	63.57	+0.08	0.930	0.930	0.000	0.889	0.888	-0.001
法语	73.58	71.53	-2.05	84.02	82.72	-1.30	0.930	0.928	-0.002	0.882	0.874	-0.008
印地语	62.68	61.70	-0.98	80.91	79.76	-1.15	0.907	0.899	-0.008	0.953	0.951	-0.002
日语	-	-	-	62.31	62.65	+0.34	0.937	0.937	0.000	0.918	0.919	+0.001
斯瓦希里语	38.62	37.85	-0.77	68.49	68.67	+0.18	0.861	0.861	0.000	0.883	0.882	-0.001
Thai	14.60	13.58	-1.02	68.19	68.19	0.00	0.909	0.909	0.000	0.921	0.921	0.000

这一模式在来自三种评估体制的代表性语言中均一致出现（高增益语言，如法语；保守型语言，如斯瓦希里语；以及低增益、形态学丰富的语言，如僧伽罗语和泰米尔语），表明表层形式相似度指标与语言学依据的正确性之间存在系统性脱节。具体而言，UMF 驱动的改动往往在保持语义等价的同时，改变形态学、论元结构、语篇顺序或语用标记，因此尽管人工评审者更青睐这些改动，其 n元语法重叠或嵌入相似度却会降低。

这解释了为何 UMF 在专家判断下可能呈现有利的收益-风险比，但在传统基准下却显得中性或负面。重要的是，这一表现与早期研究的发现一致：UMF 的变化率与潜在大语言模型错误暴露密切相关，而非与基准方差相关，这表明 UMF 所响应的是这些指标在结构上无法检测到的语言分歧。

UMF 并非与 BLEU、chrF 或基于嵌入的分数相竞争，而是与它们正交运作，充当一个元语言决策层，用于识别并纠正表层相似度度量无法察觉的错误。从这个意义上说，UMF 不应仅被视为辅助重排序器；在某些语言体制中，它还是一种互补的、甚至可替代的评估信号，用于评估类型多样且语篇敏感的语言中的正确性。

5.9 局限性

本研究严格在生成后重排序的设置下评估 UMF。当 UMF 约束在生成过程的较早阶段施加时，结果可能有所不同。此外，当前特征集对若干形态丰富、语用密集的语言覆盖不足。

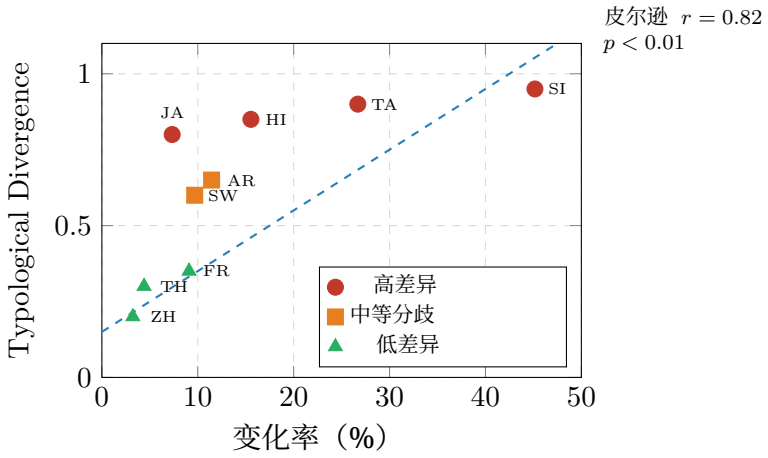

图 7：干预率（变化率）与相对于英语的类型学差异之间的相关性。强正相关（$r = 0.82$）证实 UMF 能够正确识别并干预最需要进行结构转换的情况。日语是一个显著的离群值：其类型学差异较大，但变化率较低，这是因为 GPT-5.2 在该语言上的基线质量较高。

我们的研究结果表明，尽管 UMF 已能捕捉大语言模型输出中具有实际意义的解读错误信号，但其检测性能受到当前语言画像深度的限制。预计完善这些语言画像将显著提升 UMF 识别解读层面错误的能力。

6 后续研究与待完成工作

本研究结果表明，基于UMF的重排序能够检测到语言学上显著的差异，但尚无法在所有语言类型中始终正确地消解这些差异。尤其是，形态丰富语言和低资源语言的性能下降，凸显了当前框架的若干局限。要解决这些局限，既需要表征方面的进展，也需要方法论方面的进展。下文概述了进一步研究的重点领域。

6.1 语言学表征的改进

UMF误差的一个主要来源是语言特征表征不完整或权重设置不当，尤其是在格系统丰富、语序灵活且具有语用标记的语言中。当前的语言画像能够捕捉这些特征是否存在，但尚未以足够精细的方式编码它们之间的互动约束。

未来的工作必须着力扩展语言概况，纳入形态、句法和话语层面特征之间的层级关系。我们不能将特征视为彼此独立的信号，而必须显式建模特征依赖关系，例如格与动词的一致关系，以及语域与语境的匹配。此外，我们需要引入负向约束，明确指出不仅哪些构式更受偏好，也要说明在特定语境中哪些构式不允许使用。若不进行此类改进，UMF 可能会在其旨在支持的语言中恰恰出现系统性的过度纠正。

6.2 干预置信度的校准

当前评估显示，UMF 在某些语言中敏感度较高，但特异度不足，频繁介入却未带来相称的收益。这表明有必要改进置信度校准。

未来研究应探究特定语言的干预阈值，以调节 UMF 覆盖基线排序的意愿程度。当竞争候选项几乎等价时，我们需要置信度衰减机制，以及能够区分真正的结构错误与风格变异的不确定性估计。更为保守型的干预策略，有望在特征欠指定的语言中减少性能下降，同时在语言信号较强之处保留收益。

6.3 将风格偏好与结构正确性解耦

若干观察到的改进对应于风格或语域层面的偏好，而非明确的语言学修正。尽管此类细化或许可以接受，但它们使关于结构性改进的主张变得复杂。

后续工作需要在评估和重排序中将核心语法正确性与风格优化分离开来。这包括引入显式标签或层级，以区分语法、语用和风格约束。我们还应在更严格的标准下评估 UMF，即仅将结构上必要的更改计为改进。这一区分对于确立 UMF 超越风格重排序的科学贡献至关重要。

6.4 错误分类体系的扩展与跨语言分析

尽管 UMF 错误大多可以进行分类，但当前的错误分析仍较为粗粒度。为指导有针对性的改进，需要建立更为细致的分类体系。

未来的研究应通过加入语言特定的子类别来扩展错误分类体系，例如敬语错配或话语助词误用。我们需要量化不同语言中的错误分布，以识别系统性的类型学失效模式。开展跨语言比较将有助于确定哪些错误类别具有普适性，哪些则是特定语言所独有的。此类分析能够以有原则的方式确定开发工作重点，并避免临时性修补。

6.5 超越重排序的评估

本研究仅将 UMF 作为生成后重排序层进行评估。尽管这一设置适用于初步验证，但它限制了潜在影响的范围。

未来的研究应探索在生成过程的更早阶段整合 UMF 约束。我们需要评估，在生成之前或生成过程中应用指导是否能够减少对激进重排序的需求。在相同的评估协议下比较仅重排序方法与集成方法，将有助于确定我们所观察到的局限究竟源于 UMF 在流水线中的位置，还是源于其概念设计。

6.6 纵向与规模化研究

最后，当前结果呈现的是 UMF 在语言覆盖初期阶段的快照。要确定随着语言概况逐渐成熟，观察到的弱点是否会减轻，还需要进行纵向评估。

未来的工作应随着语言特征集合的扩充和权重调整，重新进行评估。我们需要通过受控消融研究来衡量各个特征类别的贡献，并针对更多低资源语言和类型学上极端的语言开展扩展实验。只有通过这样的迭代验证，才能有意义地评估普遍性主张。

致谢

我们衷心感谢 Zwag AI 首席执行官 Priya M. Nair 的远见、领导力和资金支持，正是这些支持使本研究得以开展。

我们还诚挚感谢 Dulmini Fernando 和 Shivalatha Sivasundaram 协调语言专家和母语译者，并开展翻译质量评估和类型学错误分析。我们进一步感谢参与人工评估的语言专家和母语译者；他们的语言学专长和审慎评估对于确保本研究的质量与可靠性至关重要。

参考文献

[1] W3Techs.（2021）。网站内容语言使用统计。 `https://w3techs.com/technologies/overview/content_language`[2] Brinkmann, H. 等（2025）。大语言模型在类型多样的语言中共享潜在语法概念的表示。*arXiv* 预印本 *arXiv:2501.06346*。[3] Arnett, C. 与 Bergen, B.（2025）。为什么语言模型在形态复杂的语言上的表现更差？*arXiv* 预印本 *arXiv:2411.14198*。[4] Eberhard, D. M., Simons, G. F. 与 Fennig, C. D.（编）（2024）。《民族语：世界语言》（第27版）。SIL International。 `https://www.ethnologue.com`[5] Dorr, B. J.（1994）。机器翻译分歧：形式化描述与解决方案。《计算语言学》，20(4)，597–633。[6] Brown, T. B. 等（2020）。语言模型是少样本学习者。《神经信息处理系统进展》，33，1877–1901。[7] Wei, J. 等（2022）。思维链提示能够引发大语言模型的推理能力。《神经信息处理系统进展》，35，24824–24837。[8]Boujkian, S. 等（2025）。通过从类型相似的高资源语言进行跨语言迁移，改进低资源机器翻译。*arXiv* 预印本 *arXiv:2501.00045*。[9] Dryer, M. S. 与 Haspelmath, M.（编）（2024）。《世界语言结构地图集》在线版。马克斯·普朗克进化人类学研究所，莱比锡。 `https://wals.info/`[10] Franceschelli, G. 与 Musolesi, M.（2024）。创意束搜索：以大语言模型为评判者来改进响应生成。*arXiv* 预印本 *arXiv:2405.00099*。[11] Papineni, K. 等（2002）。BLEU：一种机器翻译自动评估方法。*ACL 2002* 论文集，311–318。

[12] Zhang, T. 等（2020）。BERTScore：使用 BERT 评估文本生成。*ICLR 2020* 论文集。[13] Rei, R. 等（2022）。COMET：用于机器翻译评估的神经框架。*EMNLP 2022* 论文集，2685–2702。[14] Krishnamurthy, P.（2019）。泰卢固语-泰米尔语迁移式机器翻译系统的开发：基于差异指数的改进。《智能系统杂志》，28(3)，493–504。[15] Evans, N. 与 Levinson, S. C.（2009）。语言普遍性的迷思：语言多样性及其对认知科学的重要性。《行为与脑科学》，32(5)，429–448。[16] Newmeyer, F. J.（2004）。反对类型学变异的参数设定方法。《语言变异年鉴》，4(1)，181–234。[17] Uchida, H.（1989）。ATLAS：富士通机器翻译系统。机器翻译峰会 *II* 论文集，112–119。[18] Nishida, T.（1985）。通过格结构转换进行英日翻译。第*11*届国际计算语言学会议论文集，395–400。[19] Hutchins, W. J. 与 Somers, H. L.（1992）。《机器翻译导论》。学术出版社，伦敦。[20] Oflazer, K.（2008）。面向形态复杂语言的统计机器翻译。第*6*届国际信息学与信息技术会议论文集。[21] Bender, E. M.（2019）。#BenderRule：论我们所研究语言的命名及其重要性。*The Gradient*。[22] Wang, Z., Lai, Y., Li, J. 与 Gao, X.（2023）。关于基于解码器的预训练语言模型的多语言能力：发现并控制语言特定神经元。*arXiv* 预印本 *arXiv:2305.00600*。[23] Oncevay, A., Haddow, B. 与 Birch, A.（2020）。通过多视角语言表示衔接语言类型学与多语言机器翻译。*EMNLP 2020* 论文集，2391–2406。