面向结构置信度不足：LLM 语义指导的介词短语依附消歧方法

周昊喆[1]　李正华[1,*]

[1] 苏州大学计算机科学与技术学院，江苏苏州 215006
20245227111@stu.suda.edu.cn, zhli13@suda.edu.cn

摘　要：介词短语附着（PPA）是英文成分句法分析中的经典歧义问题：正确判断介词短语在句法上的依附关系往往依赖语义与世界知识。成分句法解析器能够提供结构概率信号，但在语义歧义样本上受限，而大语言模型（LLM）的语义理解能力可提供补充。为此，本文提出一种融合解析器结构置信度与 LLM 语义判断的 PPA 校正框架：利用 Neural CRF 解析器的联合边缘概率构造依附得分，再引入 LLM 在完整语境下判断 PP 与候选附着点的语义合理性；最终通过阈值门控决定使用解析器预测或 LLM 判断。此外，本文从 PTB 成分树中抽取 Verb–NP–PP 实例，并对部分样本进行人工复核，以构建更适合本任务的 PPA 评估集。实验表明，本方法不仅在多个 LLM 下均能稳定提升 PPA 判别性能，而且具备较强的跨模型适用性。

关键词：介词短语依附；成分句法分析；结构置信度；大小模型协同；大语言模型

中图分类号：TP391　　**文献标识码**：A

Towards Structural Confidence Uncertainty: LLM Semantics-Guided Prepositional Phrase Attachment Disambiguation

Haozhe Zhou[1]　Zhenghua Li[1,*]

[1]School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China
20245227111@stu.suda.edu.cn, zhli13@suda.edu.cn

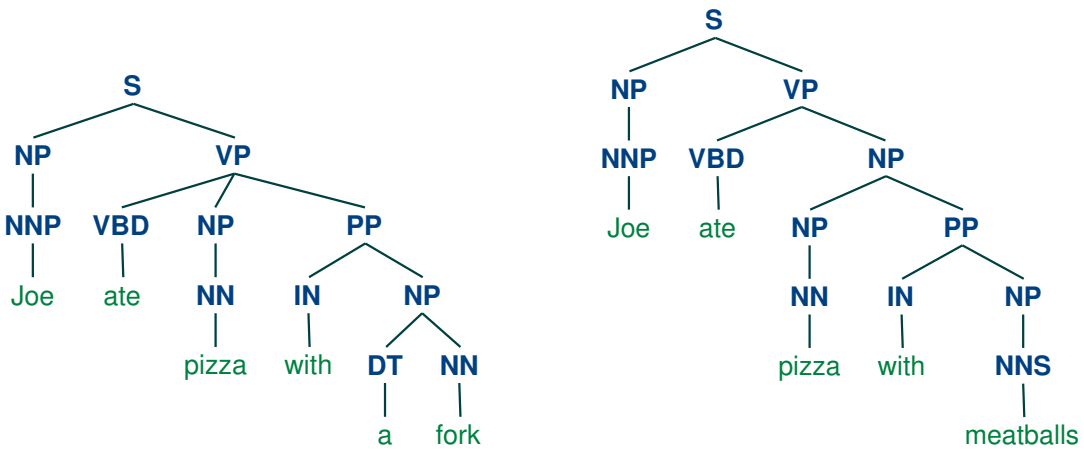
Abstract: Prepositional phrase attachment (PPA) is a classic ambiguity problem in English constituency parsing: correctly determining the attachment site of prepositional phrase typically requires semantic knowledge and world knowledge. Constituency parsers can provide structural probability signals, but are limited on semantically ambiguous cases, while large language models (LLMs) offer complementary semantic understanding. To this end, we propose a PPA correction framework that combines parser structural confidence with LLM semantic judgment: we first compute an attachment score from the joint marginal probabilities of a Neural CRF parser, then invoke an LLM to assess the semantic plausibility between the PP and each candidate attachment site in full-sentence context; a threshold-based gating mechanism finally determines whether to retain the parser prediction or adopt the LLM judgment. Furthermore, we extract Verb–NP–PP instances from PTB constituency trees and manually review a subset of ambiguous cases to construct a PPA evaluation set better suited to our task. Experiments demonstrate that our method consistently improves PPA disambiguation performance across multiple LLMs and exhibits strong cross-model applicability to different modeling approaches.

Keywords: prepositional phrase attachment; constituency parsing; structural confidence; small-large model collaboration; large language models

---

0 引言

英文介词短语附着问题（Prepositional Phrase Attachment，简称 PPA）是指：在一个 “动词-名词短语-介词短语” 序列中（即 Verb-NP-PP），介词短语既可能修饰动词、也可能修饰名词短语，需要确定其正确依附对象的问题[1]。如图 1 所示，在”Joe ate pizza with a fork” 中，介词短语应附着于动词 ate；而在”Joe ate pizza with meatballs” 中，介词短语应附着于名词 pizza。PPA 的困难之处在于，这种歧义本质上是一个语义判断问题：正确的附着方向往往取决于对整个句子事件语义的把握乃至外部世界知识的调用，难以通过词语间的表层词义关系加以确定。尽管自九十年代起学界就开始对该问题进行系统研究 [1], [2]，PPA 始终是句法分析领域尚未彻底解决的经典难题。Kummerfeld 等人 [3] 对主流解析器的输出错误进行系统分析后发现，PPA 错误在各类句法错误中占比最高。



(a) Attached to Verb　　(b) Attached to NP

图 1. Verb–NP–PP 结构中介词短语的两种典型依附方式。左图中，with a fork 表示完成 “吃” 这一动作所使用的工具，修饰动词 ate 所表示的事件，因此 PP 依附于动词；右图中，with meatballs 用于限定名词短语 pizza，描述其配料属性，因此 PP 依附于名词短语

目前存在两种主流 PPA 建模方式。一类工作将其作为独立的依附判别任务，通过词汇统计、语义资源、上下文化表示或多候选建模来引入更丰富的语义与上下文信息 [4], [5], [6], [7]；另一类工作则将 PP 依附置于完整句法解析过程中考察，强调在非 oracle（候选集合由金标准句法树直接提供）条件下评估解析器输出中的依附错误 [8], [9], [10]。对于后一类设定而言，成分句法解析器能够在大多数样本上给出准确的依附预测，但 PP 依附的正确判断有时需要依赖世界知识与常识，解析器在这方面存在明显局限；此外，PP 可以与不同动词、名词等形成多样化搭配，许多组合在训练数据中并不常见。大语言模型（Large Language Models, LLMs）在大规模文本预训练中获得了丰富的语义关联与常识知识，有潜力弥补解析器在 PP 依附判断中语义信息的不足；但现有研究表明，LLM 的句法分析能力并不稳定，不同句法知识点之间表现差异明显，其中 PP 依附问题对 LLM 尤其具有挑战性 [11]。因此，本文充分利用解析器在结构建模上的优势与 LLM 在语义理解上的优势，提出一种两者协同的 PP 依附校正框架：以解析器的预测结果为基础，通过其输出的置信度信号识别复杂样本，并在这些样本上有选择地引入 LLM 的语义判断加以校正。

本文提出一种融合解析器结构置信度与大语言模型语义判断的 PP 依附校正方法。具体而言，本文使用 Neural CRF 成分句法解析器 [12] 对完整句子进行解析，并基于联合边缘概率构造依附得分，以刻画解析器在 Verb attachment 与 NP attachment 两种结构假设之间的相对偏好和置信度；随后，通过固定格式的 prompt 引导 LLM 在完整句子语境下判断 PP 是否能够合理修饰 Verb 或 NP；最终，通过阈值门控策略决定保留解析器预测还是采

[1]PPA 问题也可能涉及更复杂的结构，如 Verb-NP-PP-PP、Verb-NP-NP-PP，本文专注于最典型的 Verb-NP-PP 结构，本文所提出的方法同样可以适用于其他结构。

用 LLM 的语义判断。为获得更可靠的实验评估，本文还基于 PTB 成分树抽取 Verb–NP–PP 候选实例，并对其中依附关系不明确或原始结构标注存在争议的样本进行人工复核，作为后续方法评测的依据。

在经人工复核后的 PP 依附数据上，本文方法在多个大语言模型和不同提示设置下均稳定提升 Neural CRF 解析器的依附判别性能，最佳 F_1 从基线的 88.90% 提升到 91.33%。进一步地，在基于 FTHA [7] 的分类式 PP 依附模型上，同样的融合策略也将准确率（Accuracy）从 93.41% 提升到 94.44%，验证了该框架在不同解析器结构下的鲁棒性。本文的主要贡献如下：

- 构建并人工复核了面向真实解析输出的 PP 依附评估数据；
- 提出一种融合成分句法解析器结构置信度与 LLM 语义判断的 PP 依附校正框架；
- 在 Neural CRF 解析器和基于 FTHA 的分类式模型上验证了该框架的有效性与跨模型适用性。

1 相关工作

传统任务设定与消歧方法 早期通常将该任务形式化为一个二分类问题：在给定的一个动词和一个名词候选之间进行选择 [1]，这种设定被近期研究认为高度人为且不够现实。因此，有工作通过考虑句子中更大范围的潜在依附位置对其进行了扩展，例如考虑某一窗口范围内所有词的数据集 BLGB [5]，以及通过拓扑字段约束为每个介词保留多个合理候选头、生成多候选 PP 依附数据集的做法 [6]。针对传统任务定义，PP 依附消歧方法可概括为四类：(i) 词汇关联统计特征：从大规模语料中统计动词、名词与介词之间的共现关系，以学习不同词汇组合的依附偏好；同时通过引入语义相似词进行词汇扩展，以缓解数据稀疏问题并提升模型的泛化能力 [4]；(ii) 语义信息特征：引入外部语义资源或常识知识提供选择偏好线索，例如将语义类别信息编码为特征并与词向量组合用于预测依附关系 [5]、通过查询外部知识库以辅助决策 [13]；(iii) 句法特征：利用结构与上下文线索弥补抽取式设定的语境缺失，例如候选中心词与介词（或介词宾语）的距离、较大上下文窗口及相关句法提示等 [14], [5]；(iv) 预训练词向量表示：能够同时捕捉句法与语义的预训练词嵌入常被用作神经网络的输入表示，以提升词汇覆盖度并丰富词汇信息 [15], [7]。

句法分析场景中的 PP 依附消歧 早期研究多将 PP 依附消歧作为独立任务评测，而非在解析器输出中统一考察。依赖 “oracle” 的做法被 Atterer 等人指出不符合真实应用 [8]，因为它几乎保证正确答案始终在候选集合中。因此，介词短语依附更合理的定位是作为句法分析的子问题。已有工作或将外部语义知识注入解析器以提升 PP 依附判别能力 [9]，或使用上下文化词向量构建 PP 依附系统，并在完全预测信息输入下与依存解析器的完整解析输出进行对比评估 [10]。

大语言模型与 PP 依附判断 近期研究开始直接考察大模型在结构歧义消解中的能力与局限。已有工作采用问答形式评测 LLM 的句法知识，发现与 PP 依附相关的问题对多数模型最具挑战性 [11]；也有工作通过将歧义句分别改写成具有明确修饰关系的句子，比较语言模型在两种改写下对关键片段的生成偏好，并以得分更高的改写所对应的依附方向作为预测结果 [16]。这些工作共同表明，LLM 可作为依附决策的语义判别信号来源，但其输出稳定性与可控性仍需结合更严格的提示设计与后续校准机制加以保障。

2 数据构建

2.1 已有数据设定

早期关于介词短语依附歧义的研究主要基于 RRR 数据集 [1] 展开，该数据集从 Penn Treebank（PTB） [17] 的 WSJ 部分中抽取得到，并采用固定的划分，将数据划分为训练集 20,801 条、开发集 4,039 条、测试集 3,097 条。具体而言，从标注句法树中筛选出包含 PP、且 PP 仅在 “依附于动词” 与 “依附于名词短语” 之间二选一的动词短语实例，并将每个样本表示为四元组 (v, n_1, p, n_2)。其中，四元组由动词头、名词头、介词和介词宾语头构成，标签为介词依附于动词 v 还是名词 n_1。例如，原句 *risk a policy conflict with the U.S.* 在抽取后可表示为四个头词 *risk*, *conflict*, *with*, *U.S.*，即 $(risk, conflict, with, U.S.)$ 的四元组形式。这一设定具有形式简洁、标注成本低等

优点，在早期研究中被广泛采用。然而，RRR 数据集仅依赖有限单词，并将介词短语依附问题从完整的句法解析过程中抽离出来，形式化为一个独立的词汇级二分类任务，脱离实际应用场景。

针对传统四元组设定的局限，Atterer 等人提出了句子级评测设定 RRR-sent [8] ，用于在更接近真实解析流程的条件下评估 PP 依附。与仅基于四个头词 (v, n_1, p, n_2) 进行判别的四元组设定不同，RRR-sent 将 RRR 测试集中每个四元组实例回指到其在 PTB 0.5 中的原始句子，使解析器能够在完整句子语境下进行结构预测与依附决策。例如，RRR 中的四元组 $(\textit{risk}, \textit{conflict}, \textit{with}, \textit{U.S.})$ 在 RRR-sent 中被恢复为完整句子 *risk a policy conflict with the U.S.*，从而保留了词汇间更丰富的结构与修饰信息。在训练数据构造方面，作者以 PTB 3 的全部 49,208 条句子为起始集合，并删除与上述测试句及 RRR 开发集对应的句子，以确保训练数据与评测数据在句子层面不存在重叠；最终保留 45,156 条句子用于训练解析器，从而构成 RRR-sent 的句子级训练与测试划分。在句子级解析框架下，Agirre 等人进一步构建了一个同时包含金标句法结构与词义标注的数据集 [9]。该研究基于 Penn Treebank 与 SemCor 的交集语料。SemCor 是在 Brown Corpus 基础上进行 WordNet 词义标注的语料库，而 Penn Treebank 的 Brown 部分则提供对应句子的人工句法树标注。作者对两者进行句子与词级对齐，并删除分词不一致的样本，最终得到 8,669 条同时包含金标句法树与词义标注的句子，并按 80%/10%/10% 划分为训练、开发与测试集。上述数据设定表明，PP 依附研究已经从词汇级四元组判别逐步转向句子级解析评测；但现有句子级数据仍多由早期抽取式实例回指得到，评测范围与结构类型受到原始构造方式限制。本文因此回到 PTB 成分句法树本身，在完整句子与真实解析输出的基础上重新构建候选实例。

表 I
PTB 训练集和测试集中 PP 依附歧义的常见局部结构模式及其分布频次

Num	Pattern	Train	Test
1	Verb–NP–PP	15,825	797
2	Verb–NP–P–NP–PP	6,800	343
3	NP–P–NP–PP	6,617	358
4	Verb–P–NP–PP	6,110	353
5	Verb–ADJP–PP	1,978	138
6	Verb–ADVP–PP	1,321	75
7	Verb–NP–NP–PP	415	23

2.2 数据来源与样本构建

本文选择 PTB 的 WSJ 部分作为基础数据集。PTB-WSJ 提供了大规模、高质量的人工标注成分句法树，使得我们能够在真实句法解析场景下，从完整句子出发考察 PP 依附决策。实验中，我们遵循 PTB-WSJ 在成分句法分析研究中的标准划分，使用 02–21 作为训练集，22 作为开发集，23 作为测试集；该划分也与主流成分句法解析器（如 Neural CRF 成分句法解析器）的设置保持一致，便于在统一解析框架下开展后续校正与对比评估。

在候选实例构建过程中，本文进一步将研究范围限定为 PTB 中最常见的 Verb–NP–PP 类型，即在局部结构中判断 PP 应依附于动词成分还是名词短语成分。虽然文献 [18] 指出，介词短语的依附方式还包括依附到形容词、介词短语内部成分乃至更高层级句法结构等多种形式（其频率分布见表 I），但这些类型相较于 Verb–NP–PP 出现频率更低，结构更复杂，且尚未形成统一稳定的建模与评测设置。因此，本文聚焦于频率最高、研究最充分的 Verb–NP–PP 类型，其余依附形式不纳入本文讨论范围。

基于上述任务边界，我们在 PTB 成分树中通过结构匹配与跨度定位抽取候选实例，保留能够形成 Verb-attach 与 NP-attach 二元对立的局部片段，为后续模型训练、解析输出校正和评估提供统一的候选空间。

2.3 人工标注与标注依据

尽管 PTB 提供了高质量的成分句法树，但其标注目标是给出一致的整体句法结构，而不是专门区分 PP 在具体语境中的语义依附方向。出于时间和一致性考虑，对于某些内部依附关系难以细分的局部结构，PTB 的标注准

则采取了扁平化策略（即子结构下的各个成分统一标注在同一层） [17]。因此，原始树库结构并不总能直接作为本文 PP 依附判断的评估标签。

为获得更适合本文任务的分析参照，我们在抽取得到的候选实例基础上，选取 263 个依附关系不够明确或原始结构标注可能影响语义判断的样本进行人工标注与复核。最终，193 个样本保持 PTB 原始结构对应的依附方向（73.4%），47 个样本被修正为另一侧依附方向（17.9%），23 个样本因两种分析均可成立而标为模糊样本（8.7%）。模糊样本不进入确定性评估；最终确定性样本为 240 个，其中 Verb attachment 90 个，NP attachment 150 个。表 II 给出三类复核结果的代表性样本。

表 II
人工复核中的代表性 PP 依附样本与修正示例

类型	目标片段与候选成分	标签变化	判断依据
Verb attachment	**Clause:** win the capital-gains issue on the floor **Verb/NP/PP:** win / the capital-gains issue / on the floor	NP → Verb	*on the floor* 表示议案在参议院/众议院会场中被争取或表决的场景，更自然地修饰事件 *win*，而不是限定名词短语 *issue*。
NP attachment	**Clause:** issue warrants on shares of Hong Kong Telecommunications Ltd. **Verb/NP/PP:** issue / warrants / on shares of Hong Kong Telecommunications Ltd.	Verb → NP	*warrants on shares* 是清晰的名词中心结构，介词短语说明认股权证对应的标的股票，因此应补足 *warrants*，而不是修饰发行行为本身。
Ambiguous	**Clause:** operates 105 banking offices in New Jersey and Florida **Verb/NP/PP:** operates / 105 banking offices / in New Jersey and Florida	NP → Ambiguous	介词短语既可理解为公司经营活动发生在 New Jersey 和 Florida，也可理解为这些 banking offices 位于两地；两种分析表达的事实基本一致，因此标为模糊。

标注依据并不直接由介词类型或 PP 的一般语义类别决定，而是取决于该 PP 在完整句子语境中更自然地补充哪一侧语义中心。若 PP 主要补充谓词所表达的事件、状态或过程，例如说明动作如何发生、向何处推进或在何种条件下成立，则判为 Verb attachment；若 PP 主要限制、补足或刻画名词短语的内容、对象、适用范围、比较背景或面向对象，则判为 NP attachment。实际标注中，许多困难样本并非由 PP 的表层语义类型直接决定，例如时间、地点、范围和受益者类 PP 既可能修饰事件，也可能限定名词短语。因此，我们特别关注候选结构中的语义中心：当动词语义较弱，而名词短语承担主要谓词语义时，PP 往往更自然地补足该名词短语；当两种分析均可成立且核心语义基本不变时，则标为模糊。

3 方法

3.1 Neural CRF 解析器

本文采用 Neural CRF 成分句法解析器作为本文方法的结构基础。该解析器是一种基于图表（chart-based）的成分句法分析方法，结合自注意力编码器与基于图表的解码策略，对所有候选区间进行打分，并通过动态规划解码生成满足结构约束的最优句法树。

本文使用其 one-stage 解析设定：区别于将括号化（bracketing）与标签预测（labeling）分开处理的 two-stage 设定，one-stage 解析对每个候选跨度直接打分，在推理过程中同时确定组块边界与句法标签，使得模型能够对带标签的完整句法树建模并输出各成分的边缘概率，从而刻画模型在不同结构之间的不确定性，这为后续处理介词短语依附问题提供关键信号。

3.2 联合边缘概率

给定一个由 n 个单词组成的句子 $x = w_0, ..., w_{n-1}$，每棵候选合法句法树 y 的得分定义为

$$s(x, y) = \sum_{(i,j,l) \in y} s(i, j, l), \tag{1}$$

其中 $s(i,j,l)$ 表示跨越 $w_i...w_j$ 且带有句法标签 $l \in L$ 的组块的得分。

整棵句法树在 CRF 条件下的概率可以表示为

$$p(y \mid x) = \frac{e^{s(x,y)}}{Z(x) \equiv \sum_{y' \in \mathcal{T}(x)} e^{s(x,y')}}, \tag{2}$$

其中 $Z(x)$ 被称为归一化项，$\mathcal{T}(x)$ 表示所有合法树结构的集合。

组块的边缘概率用于刻画某一成分在所有合法句法树中出现的概率，为了计算各个组块的边缘概率，可通过对 $\log Z(x)$ 求导来获得对应的 $p((i,j,l) \mid x)$。

在本文中,我们关心的是两个给定组块 s_1 和 s_2 在解析树中同时出现的概率。记这一联合边缘概率为 $P(s_1, s_2)$，可以写为

$$\begin{aligned} P(s_1, s_2) &= \sum_{y:\, s_1, s_2 \in y} p(y \mid x) \\ &= \frac{Z_{\text{constr}}(x; s_1, s_2)}{Z(x)}, \end{aligned} \tag{3}$$

其中 $Z(x)$ 是在所有合法句法树 $y \in \mathcal{T}(x)$ 上对 $\exp(s(x,y))$ 求和得到的归一化项，而 $Z_{\text{constr}}(x; s_1, s_2)$ 则只对那些同时包含 s_1 和 s_2 的句法树求和。后者可以看作在加入“必须同时包含 s_1 和 s_2”这一约束后，模型的受限的归一化项。

直接枚举并计算 $Z_{\text{constr}}(x; s_1, s_2)$ 代价较高。为此，我们为两个目标组块分别引入辅助变量 λ_1 和 λ_2，并在归一化项的求和中对每棵候选树 y 进行“软约束标记”：若 y 包含组块 s_1，则其未归一化权重额外乘以因子 e^{λ_1}；若包含组块 s_2，则额外乘以因子 e^{λ_2}。从而得到一个带偏置的归一化项，并将其对数记为带偏置的对数归一化项 $L(\lambda_1, \lambda_2)$。在 $\lambda_1 = \lambda_2 = 0$ 时该模型退化回原始 CRF。根据指数族/CRF 的导数性质，混合二阶偏导对应协方差项 $P(s_1, s_2) - P(s_1)P(s_2)$；因此联合边缘概率可写为：

$$\begin{aligned} P(s_1, s_2) &= \frac{\partial^2 L}{\partial \lambda_1 \partial \lambda_2} + P(s_1)P(s_2) \\ &= \frac{Z_{\text{constr}}(x; s_1, s_2)}{Z(x)}. \end{aligned} \tag{4}$$

3.3 依附得分

在比较 Verb-attached 和 NP-attached 两种依附方式之前，需要明确本任务中用于判别的三个关键组块。图 2 给出了二叉化后的最小判别结构。当 PP 依附动词（Verb-attached）时，二叉化会先将 V 与 NP 合并形成中间 VP 组块，记为 S_1；当 PP 依附名词（NP-attached）时，二叉化会先将 NP 与 PP 合并形成扩展 NP 组块，记为 S_2。两种情况下，外层的歧义片段对应同一个整体 VP 组块 S_0，差异仅体现在内部中间组块由 S_1 或 S_2 所刻画。由此，$\{S_0, S_1, S_2\}$ 形成了“共同上界 + 两个分歧点”的最小结构单元，能够完整覆盖并区分两类依附方式。

利用 one-stage CRF Parser 模型可以得到子树内部所有组块的边缘概率，并进一步通过二阶导数方法计算两种依附假设对应的联合边缘概率 $P(S_0, S_1)$ 与 $P(S_0, S_2)$，分别对应 PP 修饰 Verb 与 PP 修饰 NP 的结构偏好概率。由于不同句子与不同片段的概率尺度受句长与搜索空间影响，直接比较绝对概率并不稳定，因此我们采用对数比值定义依附得分：

$$\text{Score}_{\text{att}} = \ln \frac{P(S_0, S_1)}{P(S_0, S_2)}. \tag{5}$$

$\text{Score}_{\text{att}} > 0$ 表示更偏向动词依附，$\text{Score}_{\text{att}} < 0$ 则偏向名词依附，$\text{Score}_{\text{att}}$ 的幅度用于衡量置信度大小。在最终的后处理中，均以 $\text{Score}_{\text{att}}$ 作为核心的模型置信度指标，用于权衡大模型和小模型的结果。

3.4 LLM 提示模板

介词短语依附高度依赖语义与世界知识 [19]。因此，在小模型之外，我们引入大型语言模型（LLM）对“PP 更自然地修饰动词还是名词”进行语义判别。LLM 的任务仅限于判断语义可依附性，而不参与句法树生成。

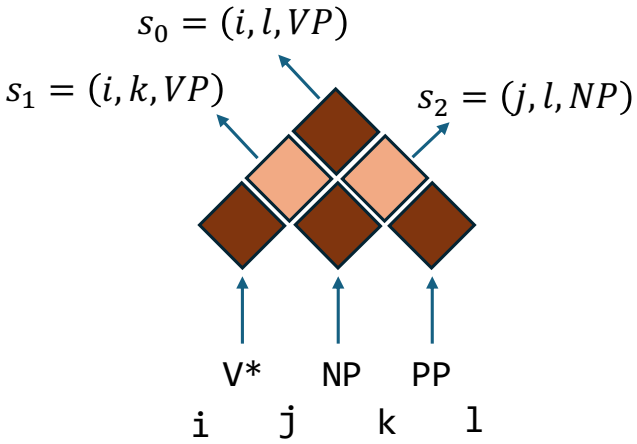


图 2　局部二叉化后用于 PP 依附判别的最小结构单元

为此，我们设计了一个固定格式的 prompt，将句子上下文、歧义片段以及候选依附点以统一形式呈现，并要求模型判断 PP 是否在语境中能够合理修饰 Verb 或 NP；完整模板见附录 A。输出格式为包含两个二值字段的 JSON 对象：`{"verb_plausible":"Yes|No","np_plausible":"Yes|No"}`。其中“Yes”表示在给定语境下该依附方向在语义上是成立的，“No”则表示缺乏语义支持。当两个字段均为“Yes”时，表示模型认为两种依附方向在语义上均具有可解释性。

3.5 大小模型结合

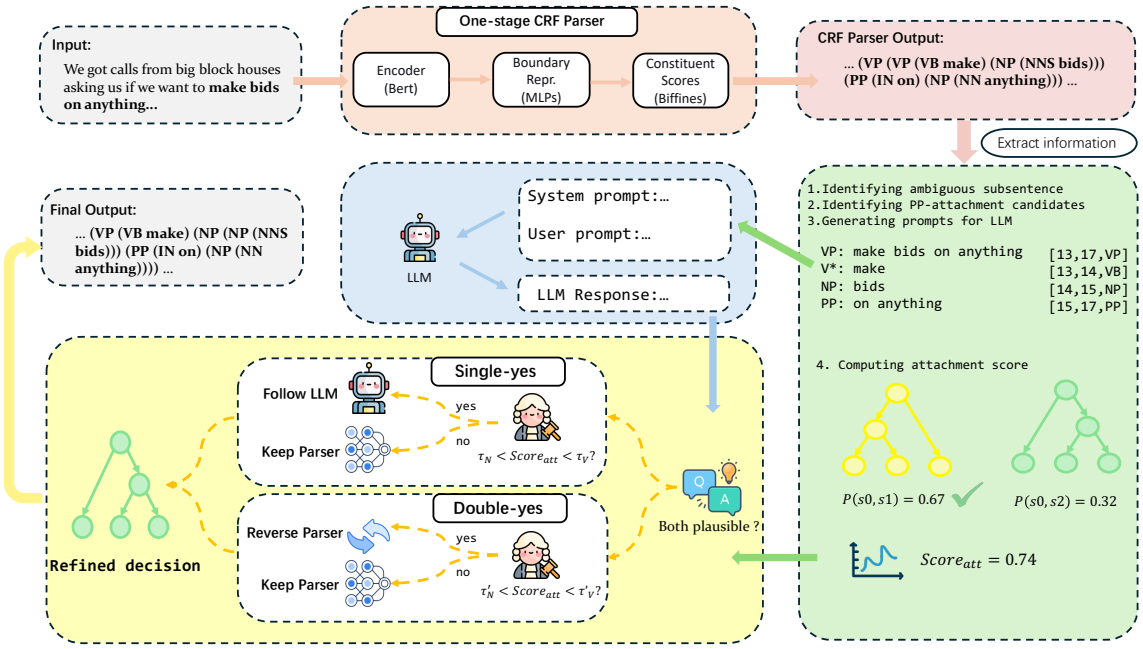


图 3. 融合结构置信度与 LLM 语义判断的 PP 依附校正框架：基于 one-stage Neural CRF 的结果抽取含 PP 依附歧义的 VP 片段并计算依附得分，随后对每个候选子句调用大模型给出动词依附与名词短语依附的语义可行性判断。最终系统依据得分与预设阈值的关系在大小模型间进行选择

训练好的小模型首先对完整句子进行成分句法解析，输出预测句法树及各跨度的边缘概率。基于解析结果，我们定位包含 PP 依附歧义的目标片段，并进一步计算与两种依附方式对应的联合边缘概率，从而得到结构模型的相对偏好。我们用依附得分 $\text{Score}_{\text{att}}$ 同时表征预测方向与置信度。在此基础上，我们引入 LLM 对同一候选实例进行语义可行性判别，并将 LLM 的语义信号与依附得分结合进行最终决策。

我们以 $\text{Score}_{\text{att}}$ 作为判别与置信度信号，其符号刻画依附方向：正值为 Verb-attached，负值为 NP-attached，绝对值反映结构偏好的确定程度。我们引入四个阈值，并根据大模型是否给出确定判断分别使用。当大模型给出确定语义方向时，我们用两个阈值 $\tau_N < 0 < \tau_V$ 划定低置信区间，当 $\tau_N < \text{Score}_{\text{att}} < \tau_V$ 时优先采用大模型语义判断；反之，若 $\text{Score}_{\text{att}}$ 不在该区间内，则说明小模型对其依附结果较为明确，直接采用小模型预测。相应地，当大模型给出不确定判断时，语义信息不具区分性，我们使用另一套阈值 $\tau'_N < 0 < \tau'_V$ 进行同样的分流：在低置信区间使用反向策略处理以减弱结构偏差，而在区间外则仍以小模型结果为主。具体而言，反向策略指当 LLM 无法给出确定语义方向（即对 Verb 与 NP 依附均给出可接受或均给出不可接受的判断）时，反转解析器的原始依附判断：若解析器倾向 Verb-attached 则改为 NP-attached，反之亦然。整体流程如图 3 所示。

4 实验与分析

4.1 数据与评估

本文所有实验均在第 2.3 节中构建并经人工标注复核的 PP 依附数据集上开展。该数据以 PTB 成分句法树中的 Verb–NP–PP 候选片段为基础，保留具有确定依附标签的样本用于模型评估；对于人工标注中判定为模糊、即 Verb attachment 与 NP attachment 均可成立且核心语义差异不明显的样本，本文不将其纳入确定性依附评估。

遵循 Atterer 等人 [8] 的评估思路，我们实现了一个脚本，用于在给定成分句法树的情况下识别其中潜在的介词短语（PP）附着歧义结构。该脚本以局部 VP 结构为分析对象，从句法树中抽取符合条件的歧义片段，并输出相应的附着决策。上述抽取过程直接作用于小模型生成的解析结果，作为后续附着判别与策略调整的基础。

在评估阶段，我们将同一抽取脚本同时应用于解析器输出的句法树与金标准句法树，从而获得可对比的附着实例，并据此判断预测结果与金标准标注之间的一致性。需要注意的是，由于解析器可能生成与部分附着配置不兼容的句法结构，预测实例与金标准实例之间并不总能一一对应。因此，在统计过程中，仅对能够在金标准句法树中找到对应实例的预测结果进行评估：若该实例的预测附着方向与金标准标注一致，则记为预测正确；若实例存在于金标准中但预测附着方向不一致，则记为预测错误；而无法在金标准中对齐的预测实例则被直接排除在评估之外。基于上述定义，精确率（Precision）被定义为预测正确的附着实例数与所有参与评估的实例总数之比，召回率（Recall）则定义为预测正确的附着实例数与测试集中金标准附着实例总数之比。

4.2 实验设置

我们使用 one-stage Neural CRF 成分句法解析器，并在二叉化处理后的 PTB 数据上进行训练。采用 BERT [20] 作为编码器，其余训练与解码配置均与原模型保持一致。

后处理阶段在两种提示设置下进行：0-shot 与 4-shot。两种设置均使用方法部分给出的同一 prompt 模板与相同的标签映射规则。在 4-shot 设置中，我们从训练集中随机采样 4 个示例构成 prompt；所有大模型实验共享同一组固定示例，以确保不同模型与提示设置之间的可比性。在所有大模型实验中，我们统一将生成温度设为 0 以降低输出随机性，所有大模型均通过官方 API 调用，调用日期为 2025 年 12 月。我们尝试了 7 种主流的大模型来协同小模型，包括 DeepSeek-3.2 [21]、Claude-4.5 [22]、Qwen-3 [23]、Kimi-K2 [24]、GPT-4o [25]、GLM-4.6 [26] 与 MiMo-V2 [27]，以考察所提出融合策略在不同语义判别器下的稳定性。为保证不同模型与不同提示设置之间结果的可比性，TreeCRF 主实验统一采用同一组阈值（τ_V=8.873, τ_N= − 10.648, τ'_V=4.779, τ'_N= − 1.636），该阈值组由开发集上的网格搜索确定，并在整个实验过程中保持固定。评估阶段严格沿用上一小节定义的对齐协议，仅将可与金标准成功对齐的实例计入 Precision 的分母，其余无法对齐的预测实例在评估中予以跳过。

4.3 主要结果

表 III 给出了本文构建并标注复核后的 PP 依附数据上的实验结果。总体来看，在所有大语言模型与提示设置下，LLM-enhanced Neural CRF 均相较于基线取得了稳定提升。Neural CRF 基线的 F_1 为 88.90；引入 LLM 后，各模型的 F_1 提升至 89.99–91.33 之间。其中，MiMo-V2 在 0-shot 设置下取得最高 F_1=91.33，相比基线提升 2.43 个百分点。从整体结果分布来看，不同大语言模型之间的性能差异相对有限，0-shot 与 4-shot 两种提示设置下的结

表 III
不同提示设置下 TreeCRF 与 LLM-enhanced TreeCRF 的 PP 依附结果（P/R/F_1，%）

Method	LLM	0-shot			4-shot		
		P	R	F_1	P	R	F_1
Neural CRF (base)	–	93.32	84.88	88.90	93.32	84.88	88.90
LLM-enhanced Neural CRF (ours)	DeepSeek-3.2	95.74	87.08	91.20	95.74	87.08	**91.20**
	Claude-4.5	95.03	86.43	90.53	94.60	86.05	90.12
	Qwen-3	95.17	86.56	90.66	95.60	86.95	91.07
	Kimi-K2	94.74	86.18	90.26	95.16	86.55	90.65
	GPT-4o	94.46	85.92	89.99	94.60	86.05	90.12
	GLM-4.6	95.45	86.82	90.93	95.74	87.08	**91.20**
	MiMo-V2	95.87	87.19	**91.33**	95.73	87.06	91.19

果也较为接近，说明本文方法在不同语义判别器与提示条件下均能够稳定改善 TreeCRF 的 PP 依附判别性能。为验证上述提升的来源，我们进行了消融实验，比较无门控直接融合与门控融合的差异，详细结果与分析见附录 B。

4.4 分析

实验结果表明，不同的 LLM 与 Neural CRF 解析器结合后均能提升 PP 依附判别性能，但提升幅度相近，各模型在统一阈值和模板下的性能结果较为集中。由于本任务的判别空间只有 Verb-attach 与 NP-attach 两类，大模型仅需在给定的候选对象之间做语义可行性判断，这降低了模型自由度，使得模型规模对性能差异的影响不大。

从提示策略看，0-shot 与 4-shot 的 F_1 差异较小，提示示例更多时仅对少量边界实例提供辅助。这与 PP 依附判断本身依赖内在语言知识和明确的候选结构有关，few-shot 示例并不会显著增加决策信息。总体来看，大模型在限定的候选结构上提供了稳定的语义纠正，而 TreeCRF 输出的依附得分能够有效区分易错实例，为阈值门控策略提供依据。

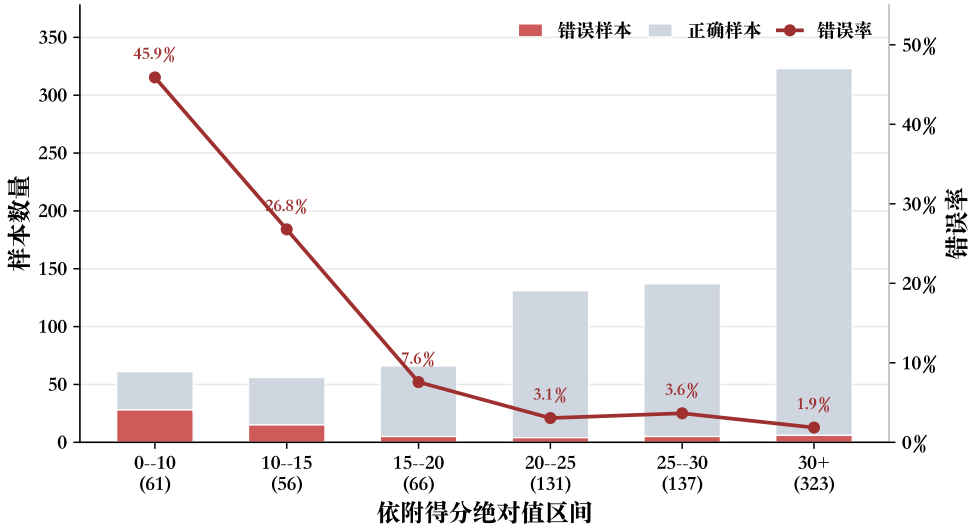


图 4　PTB 测试集上依附得分 $\text{Score}_{\text{att}}$ 的分箱统计分析。柱形图表示各得分区间内正确与错误预测的样本数量，折线表示对应区间的预测错误率

基于 Neural CRF 在测试集上的预测结果，我们对模型输出的依附得分进行了区间统计分析，以检验该得分是否能够反映预测的可靠程度。具体而言，我们按照依附得分绝对值划分区间，并在每个区间内统计正确样本、

错误样本以及对应的错误率（即该区间内预测错误样本所占比例）。如图 4 所示，低得分区间不仅错误率明显更高，也包含较多实际错误样本；随着依附得分增大，错误样本数量和错误率均明显下降。这表明小模型输出的依附得分能够有效刻画预测的不确定性，从而为基于得分阈值的策略设计提供了经验依据。

4.5 拓展实验

为进一步考察本文融合框架在不同小模型结构上的可拓展性，我们引入一个基于上下文化表示的二分类模型作为补充基线。该模型借鉴 FTHA 中“上下文化编码 + 候选比较”的建模思想：首先使用预训练 BERT 对完整句子进行编码，获得上下文化词表示；随后围绕候选依附成分与介词短语构造片段级表示，即分别抽取 Verb、NP 与 PP 对应片段的表示，并据此预测 PP 最终应依附于 Verb 还是 NP。与更侧重头词级候选表示不同，本文采用片段级表示以适配本文的数据抽取方式，并更充分地保留短语内部的语义信息。在与 LLM 结合时，该分类器使用两类输出概率的对数比值作为小模型置信度，并沿用本文的阈值门控策略引入 LLM 语义判断。由于该模型不生成完整句法树，不涉及预测树与 gold 树之间的实例对齐，因此使用 Accuracy 作为评估指标。拓展实验中的阈值同样在所有大语言模型配置下保持固定，具体设为 τ_V=1.017, τ_N= − 10.024, τ'_V=1.017, τ'_N= − 0.000。

表 IV
不同提示设置下 FTHA-SENT 与 LLM-ENHANCED FTHA-SENT 的 PP 依附结果（ACC.，%）

Method	LLM	0-shot Acc. (%)	4-shot Acc. (%)
FTHA-sent(base)	–	93.41	93.41
LLM-enhanced FTHA-sent (ours)	DeepSeek-3.2	**94.06**	**94.44**
	Claude-4.5	93.93	94.19
	Qwen-3	93.80	**94.44**
	Kimi-K2	93.93	94.06
	GPT-4o	**94.06**	93.80
	GLM-4.6	93.93	93.93
	MiMo-V2	93.93	94.06

表 IV 给出了分类式基线及其 LLM 增强结果。可以看到，FTHA-sent 分类模型本身取得 93.41% 的准确率；在引入 LLM 后，所有模型配置均带来进一步提升，其中 DeepSeek-3.2 和 Qwen-3 在 4-shot 设置下达到最高的 94.44%。这一结果说明，本文提出的融合策略并不依赖 TreeCRF 的特定结构表示：只要小模型能够给出候选依附方向及相应置信度，LLM 的语义判断仍然可以作为有效的补充信号，从而提升 PP 依附判别性能。

5 结论

本文针对英文成分句法分析中的介词短语依附问题，提出一种基于解析器置信度的大小模型协同校正框架。本文从 PTB 成分树中抽取 Verb–NP–PP 候选实例，并对依附关系不够明确或原始标注可能影响语义判断的样本进行人工复核。在方法上，以 TreeCRF 联合边缘概率构造依附得分作为置信度度量，通过阈值门控策略在低置信样本上选择性引入 LLM 语义判断，实现结构约束与语义信号的互补融合。

实验结果表明，本文方法在多个 LLM 与提示设置下均取得稳定提升，且框架在不同小模型结构上同样有效，具有良好的跨模型适用性。分析表明，阈值门控融合策略在不同 LLM 系列与提示设置下均保持一致的提升效果。

本文方法仍存在一定局限：仅在 PTB 的 Verb–NP–PP 结构上开展实验，尚未覆盖其他类型的 PPA 结构，也未在其他跨领域数据集上验证泛化能力；且大小模型协同引入了 LLM 调用成本。未来工作将扩展至更多 PPA 类型与跨领域数据集，以进一步验证方法的泛化能力。

致谢

感谢所有匿名审稿人的宝贵意见。本研究受国家自然科学基金（No. 62576237，No. 62336006）资助。

参考文献

[1] A. Ratnaparkhi, J. Reynar, and S. Roukos, "A maximum entropy model for prepositional phrase attachment," in *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8–11, 1994*. Stroudsburg, PA, USA: Association for Computational Linguistics, Mar. 1994, pp. 250–255.

[2] E. Brill and P. Resnik, "A rule-based approach to prepositional phrase attachment disambiguation," in *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*. Kyoto, Japan: Association for Computational Linguistics, Aug. 1994, pp. 1198–1204.

[3] J. K. Kummerfeld, D. Hall, J. R. Curran, and D. Klein, "Parser showdown at the Wall Street corral: An empirical investigation of error types in parser output," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 1048–1059.

[4] P. Pantel and D. Lin, "An unsupervised approach to prepositional phrase attachment using contextually similar words," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, Oct. 2000, pp. 101–108.

[5] Y. Belinkov, T. Lei, R. Barzilay, and A. Globerson, "Exploring compositional architectures and word vector representations for prepositional phrase attachment," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 561–572, 2014.

[6] D. de Kok, C. Dima, J. Ma, and E. W. Hinrichs, "Extracting a PP attachment data set from a german dependency treebank using topological fields," in *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), Bloomington, IN, USA, January 20–21, 2017*, ser. CEUR Workshop Proceedings, M. Dickinson, J. Hajic, S. K"ubler, and A. Przepiórkowski, Eds., vol. 1779. Bloomington, IN, USA: CEUR-WS.org, Jan. 2017, pp. 89–98.

[7] A. Ratnaparkhi and A. Kumar, "Resolving prepositional phrase attachment ambiguities with contextualized word embeddings," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, S. Bandyopadhyay, S. Lalitha Devi, and P. Bhattacharyya, Eds. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 335–340.

[8] M. Atterer and H. Sch"utze, "Prepositional phrase attachment without oracles," *Computational Linguistics*, vol. 33, no. 4, pp. 469–476, Dec. 2007.

[9] E. Agirre, T. Baldwin, and D. Martinez, "Improving parsing and PP attachment performance with sense information," in *Proceedings of ACL-08: HLT*, J. D. Moore, S. Teufel, J. Allan, and S. Furui, Eds. Columbus, Ohio: Association for Computational Linguistics, Jun. 2008, pp. 317–325.

[10] B.-N. Do and I. Rehbein, "Parsers know best: German PP attachment revisited," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2049–2061.

[11] H. Zhou, Y. Hou, Z. Li, X. Wang, Z. Wang, X. Duan, and M. Zhang, "How well do large language models understand syntax? an evaluation by asking natural language questions," arXiv preprint arXiv:2311.08287, Nov. 2023.

[12] Y. Zhang, H. Zhou, and Z. Li, "Fast and accurate neural CRF constituency parsing," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 4046–4053.

[13] Y. Xin, H. Lieberman, and P. Chin, "Revisiting the prepositional-phrase attachment problem using explicit commonsense knowledge," arXiv preprint arXiv:2102.00924, Feb. 2021.

[14] M. Olteanu and D. Moldovan, "PP-attachment disambiguation using large context," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 273–280.

[15] P. S. Madhyastha, X. Carreras, and A. Quattoni, "Prepositional phrase attachment over word embedding products," in *Proceedings of the 15th International Conference on Parsing Technologies*, Y. Miyao and K. Sagae, Eds. Pisa, Italy: Association for Computational Linguistics, Sep. 2017, pp. 32–43.

[16] J. Blackmore and M. Stone, "Can LLMs disambiguate grounded language? the case of PP attachment," in *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, G. Angelova, M. Kunilovskaya, M. Escribe, and R. Mitkov, Eds. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sep. 2025, pp. 166–174.

[17] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, Jun. 1993.

[18] B. Mitchell and R. Gaizauskas, "A labelled corpus for prepositional phrase attachment," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, Eds. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004, pp. 2187–2190.

[19] D. Hindle and M. Rooth, "Structural ambiguity and lexical relations," *Computational Linguistics*, vol. 19, no. 1, pp. 103–120, Mar. 1993.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[21] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang *et al.*, "Deepseek-v3 technical report," *ArXiv*, vol. abs/2412.19437, 2024.

[22] Anthropic, "System card: Claude sonnet 4.5," Anthropic System Card, sep 2025.

[23] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," in *arXiv preprint arXiv:2505.09388*, 2025.

[24] Kimi Team, Y. Bai, Y. Bao, Y. Charles, C. Chen, G. Chen, H. Chen, H. Chen, J. Chen, N. Chen *et al.*, "Kimi k2: Open agentic intelligence," in *arXiv preprint arXiv:2507.20534*, 2025.

[25] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. M. de Almeida, J. Altenschmidt, S. Altman *et al.*, "Gpt-4 technical report," in *arXiv preprint arXiv:2303.08774*, 2023.

[26] Team GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai *et al.*, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," *ArXiv*, vol. abs/2406.12793, 2024.

[27] X. Xiao, B. Xia, B. Yang, B. Gao, B. Shen, C. Zhang, C. He, C. Lou, F. Luo, G. Wang *et al.*, "Mimo-v2-flash technical report," *ArXiv*, vol. abs/2601.02780, 2026.

附录 A. LLM Prompt 模板

为便于复现本文的大模型语义判别过程，本附录给出 prompt 模板。附录内容仅用于补充说明，不改变正文中的方法定义与实验设置。

本文在调用大语言模型时，将完整句子、目标片段、候选依附点和待判断的介词短语以固定格式输入模型。模型只判断给定 PP 在当前语境中是否能够合理修饰 Verb 或 NP，不参与句法树生成。具体模板如图 5 所示。

System prompt

[Role]:You are a syntactic analysis expert.
[Task]:Judge whether a PP in the sentence can plausibly modify the Verb or the NP of a target phrase.
Output ONLY a JSON object:
{"verb_plausible":"Yes|No", "np_plausible":"Yes|No"}
[Rules]:
1. Use full-sentence context.
2. "Yes" = the PP can plausibly modify that candidate.
3. 'No' = the PP does not plausibly modify that candidate.
4. If both are plausible:
{"verb_plausible":"Yes","np_plausible":"Yes"}

User prompt

Sentence: We got calls from big block houses asking us if we want to **make bids on anything.**
Target phrase: make bids on anything
Verb: make
NP: bids
PP: on anything
Return JSON and no explanation:

Response

{"verb_plausible": "No","np_plausible": "Yes"}

图 5　用于 PP 依附语义判别的 LLM prompt 模板

在 4-shot 设置中，我们在同一任务说明之前加入四个格式相同的示例，用于说明输入字段、候选依附点和 JSON 输出格式；其余标签映射规则保持不变。

附录 B. 消融实验

为验证阈值门控策略对模型性能提升的贡献，本文分别在 TreeCRF 和 FTHA-sent 上开展消融实验，并比较完整门控方法与三种无门控融合策略。无门控策略不考虑基模型的预测置信度，而是直接使用 LLM 判断全部样本。对于 LLM 未给出唯一依附选择的 unsure 样本，分别采用三种处理方式：（1）keep，保留基模型的原始预测；（2）reverse，反转基模型的原始预测；（3）random，在 Verb attachment 和 NP attachment 之间随机选择。为保证实验结果可复现，random 策略采用固定随机种子。

表 V 给出了 TreeCRF 上的消融结果。keep 的表现最为稳定，其平均 F_1 在 0-shot 和 4-shot 设置下分别为 88.90% 和 89.45%，与 TreeCRF 基线的 88.90% 持平或略有提升。reverse 的平均 F_1 分别为 68.78% 和 81.69%，random 的平均 F_1 分别为 79.08% 和 85.91%，均低于基线。相比之下，完整门控方法的 F_1 达到 89.99%–91.33%，在不同 LLM 和提示设置下表现更为稳定。

表 VI 展示了 FTHA-sent 上的消融结果。无门控 keep 的平均准确率在 0-shot 和 4-shot 设置下分别为 92.45% 和 93.02%，reverse 与 random 的整体表现进一步下降。采用阈值门控后，平均准确率分别达到 93.95% 和 94.13%，均高于 FTHA-sent 基线的 93.41%。这一结果说明，阈值门控策略在不同类型的基模型上均能带来较为稳定的性能提升。

综合两组实验可以看出，直接将 LLM 的语义判断应用于全部样本并不能保证性能提升，部分处理策略还会引入额外错误。阈值门控通过结合基模型置信度控制 LLM 的介入范围，减少了对原有正确预测的干扰，是实现稳定性能提升的关键组成部分。

表 V
LLM-enhanced Neural CRF 门控与无门控融合性能对比（F_1，%）

	0-shot				4-shot			
LLM	keep	reverse	random	ours	keep	reverse	random	ours
DeepSeek-3.2	89.45	77.67	84.03	91.20	91.07	82.27	87.42	91.20
Claude-4.5	89.17	71.04	79.97	90.53	89.85	87.96	89.17	90.12
Qwen-3	89.72	87.55	88.23	90.66	90.26	86.20	89.45	91.07
Kimi-K2	86.33	62.92	74.42	90.26	89.58	82.27	86.47	90.65
GPT-4o	89.99	80.51	86.06	89.99	88.09	83.22	86.74	90.12
GLM-4.6	89.17	67.25	77.00	90.93	88.77	83.22	85.52	91.20
MiMo-V2	88.50	34.51	63.87	91.33	88.50	66.71	76.59	91.19

表 VI
LLM-enhanced FTHA-sent 门控与无门控融合性能对比（Acc.，%）

	0-shot				4-shot			
LLM	keep	reverse	random	ours	keep	reverse	random	ours
DeepSeek-3.2	93.54	79.84	86.30	94.06	95.22	85.53	90.83	94.44
Claude-4.5	93.15	74.94	82.82	93.93	93.02	91.99	91.99	94.19
Qwen-3	93.15	91.47	92.38	93.80	93.67	90.31	93.02	94.44
Kimi-K2	89.15	66.67	76.49	93.93	93.15	86.05	90.44	94.06
GPT-4o	93.80	83.85	89.79	94.06	91.86	86.82	89.15	93.80
GLM-4.6	92.38	71.83	81.52	93.93	92.38	87.08	89.79	93.93
MiMo-V2	91.99	37.60	67.05	93.93	91.86	70.67	80.49	94.06