大语言模型可以分析花园幽径句吗？—基于跨语言数据集的实证研究

李琦[1,2]　**纪悦**[1,2]　**李洪政**[1,2]✉
[1]北京理工大学外国语学院　北京市房山区良乡东路9号
[2]语言工程与认知计算工信部重点实验室　北京市房山区良乡东路9号
{lq,jiyue,lihongzheng}@bit.edu.cn

摘要

花园幽径句是在句法或语义上存在局部或临时歧义的一类特殊句子，在汉语和英语中都普遍存在，对于语言处理和认知机制等研究具有重要价值。本文聚焦于大语言模型理解分析花园幽径句的能力。本研究首先构建了一个具有典型结构的英汉双语花园幽径句数据集。随后基于该数据集开展了跨语言、跨模型的句法结构分析及语义理解的对比实验，考察多个大语言模型处理不同语言花园幽径句的消歧和理解分析能力，并对比了大模型与传统句法分析器Stanford Parser模型的分析能力。实验结果显示大语言模型测试结果呈现出与人类认知相似的花园幽径效应，可以利用名词合理性及动词偏向性为线索辅助消除句子歧义，英语句子的消歧能力显著优于汉语。语言模型句法分析与语义分析准确率具有较大差异。本实证研究揭示了大语言模型处理不同条件歧义句的表现差异，为语言处理和认知机制等提供了新的计算视角证据。

关键词： 花园幽径句；大语言模型；句法分析；语义分析

Can Large Language Models Analyze Garden Path Sentences? —An Empirical Study based on Cross-Lingual Data

Qi Li[1,2]　**Yue Ji**[1,2]　**Hongzheng Li**[1,2]✉
[1]School of Foreign Languages, Beijing Institute of Technology /
No.9, Liangxiang East Road, Fangshan District, Beijing
[2]Key Laboratory of Language, Cognition and Computation,
Ministry of Industry and Information Technology/
No.9, Liangxiang East Road, Fangshan District, Beijing
{lq,jiyue,lihongzheng}@bit.edu.cn

Abstract

Garden path sentences are a special type of sentences with local or temporary ambiguity in syntax or semantics. They are common in both Chinese and English and are of great value for research on language processing and cognitive mechanisms. This paper focuses on the ability of large language models (LLMs) to understand and analyze garden path sentences. We first constructed a dataset of English-Chinese bilingual garden path sentences with a typical structure. Then, cross-linguistic and cross-model experiments on syntactic structure analysis and sentence comprehension were carried out based on this dataset to examine how well LLMs can analyze, disambiguate and understand garden path sentences. The experiments also compared between LLMs and the Stanford Parser in their analysis capabilities. The experimental results indicate that LLMs show a garden path effect similar to human sentence processing. They

©2025 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

can use noun plausibility and verb bias as clues to assist in disambiguating sentences. The disambiguation of English sentences is significantly better than that of Chinese sentences. The language models show a large difference in the accuracy of syntactic parsing and that of semantic parsing. This empirical study reveals the performance differences between LLMs in processing temporary ambiguity in English vs. Chinese sentences under different conditions, providing new evidence for sentence processing and the underlying cognitive mechanisms from a computational perspective.

Keywords: garden path sentences , large language models , syntactic parsing , semantic analysis

1 引言

花园幽径句是一种特殊的局部或暂时歧义句，这类特殊句式最早由Bever(1970)提出，其核心特征在于：读者在早期句法分析阶段形成临时性错误分析，需依赖后续语境线索触发回溯机制进行重新解读，最终形成正确表征。花园幽径句是一种形象的比喻：在某些结构特殊的句子理解中，读者就像在花园里悠然自得地沿着主路径欣赏园内美景，信步走向花园的出口，结果到尽头才发现此路径不对。花园幽径句自提出以来便成为探究人类语言处理机制的重要语言材料，为揭示人类语言理解的即时性加工特征提供了独特的研究窗口。

花园幽径句的歧义会随着信息量的逐渐增加而消解。例如：在Bever最初提出的经典花园幽径句“The horse raced past the barn fell”中，读者在初始分析阶段可能会形成错误理解(如将“raced past the barn”分析为主句的动词短语)，但是读到句末动词“fell”时触发认知冲突，需要回溯至歧义节点进行重新分析，最终确认“raced past the barn”是缩略关系从句。花园幽径句自提出以来便成为探究人类语言处理机制的重要语言材料，为揭示人类语言理解的即时性加工特征提供了独特的研究窗口。

大语言模型在自然语言理解与生成方面取得较大进步，为语言加工机制研究带来了新的可能性。虽然大模型在一般认知领域及各项语言处理基准测试中表现卓越，但其在标准语言任务中的优异表现与实际语言能力之间的关系仍存在争议。有研究显示大模型与大脑活动存在高度相似性，可预测人类语言行为(Hao et al., 2020)；也有研究强调仅通过形式训练的系统无法真正学习到语言的意义，其工作机制与人类认知加工存在差异(Bender and Koller, 2020)。花园幽径句因其独特的认知加工模式，成为检验大模型语言工作机制的重要工具。

近年来已有一些关于大模型处理花园幽径句的研究(Irwin et al., 2023; Li et al., 2024; Amouyal et al., 2025)，并将大模型分析花园幽径句的能力、错误与人类进行对比。Li et al.(2024)以24组“主动词/缩略关系”类花园幽径句为实验材料，Amouyal et al.(2025)以69组“直接宾语/主语”类花园幽径句为实验材料，均发现语言模型呈现花园幽径效应，且可以利用语言线索进行消歧。但这些工作都是研究英语花园幽径句，我们目前尚未发现大模型处理花园幽径句的跨语言对比研究。结合花园幽径句的句法和语义特点，本研究以汉英“直接宾语/补语”类花园幽径句为对象以大语言模型和传统句法分析器Stanford Parser(Chen and Manning, 2014)为分析工具，希望探索以下三个问题：

(1) 大语言模型能否利用语义合理性、动词偏向性等语言线索对不同语言的句子进行消歧，以及模型参数规模、推理能力如何影响模型的歧义消解能力?

(2) 传统句法分析器能否利用语义合理性、动词偏向性等语言线索对不同语言的句子进行消歧，与大语言模型是否表现一致?

(3) 大语言模型对花园幽径句的句法分析与语义分析能力是否存在差异?

本研究首先构建了一个汉英花园幽径句双语数据集(BIT-GPS-576)，该数据集包括花园幽径句与对应的疑问句理解及答案。随后基于该数据集开展了句法分析与语义分析两类实验，对比分析了大语言模型与句法分析器Stanford Parser的句法分析准确率与语义分析准确率。实验结果显示：英语花园幽径歧义句的处理准确率显著高于汉语，且大模型可以利用语义合理性、动词结构偏向性线索提升句子的消歧能力。大模型整体表现出了较强的语义分析能力，语义分析准确率与句法分析准确率呈现了较大差异性。实验结果揭示了大语言模型处理不同语言及不同条件的歧义句的性能表现差异，为语言处理机制提供了计算视角证据，为大语言模型与人类认知语言加工的异同争议提供了实证基础。

(c) 2025 中国中文信息学会计算语言学专业委员会

本文的主要贡献体现在以下三个方面：

- 首次构建了一个英汉双语的花园幽径句数据集，并在设计中系统地控制了歧义性、动词偏向性和语义合理性三个变量，为跨语言的句法歧义分析研究提供了重要基准和数据支持。
- 首次系统开展了跨语言和跨模型的花园幽径句句法分析和语义分析实验，对比了大模型和传统句法分析器的性能表现，揭示了它们在句法分析和语义推理能力上的优势与不足。
- 大模型与传统句法分析器在句法测试层面呈现出来的类人认知模式能为语言处理的认知神经机制提供新的启发和研究思路。

2 相关研究

自上世纪花园幽径句首次提出以来，相关研究多聚焦于英语花园幽径句(卢华萍and 吴明军, 2021; 吴迪et al., 2024; 王邵馨et al., 2025)。有研究通过眼动追踪技术观测关键区域的阅读时长(如首次注视时长与回视路径时间)，并结合相关的语义理解问题检验是否出现花园幽径效应(Christianson et al., 2001; 杜家利and 于屏方, 2018)。近年来，研究发现英语母语者和英语二语者在花园幽径句理解过程中存在差异，母语者可以更高效地利用动词偏向性线索进行早期消歧，而二语学习者更容易受到语义合理性的影响(Brothers et al., 2021; Cunnings, 2017; Qian et al., 2019; Lee et al., 2013; Roberts and Felser, 2011)。此外，杜家利和于屏方(2016)曾进行人类学习者与Stanford Parser的对照实验，表明机器与学习者不具有完全联动性与绝对共时性。大量实证研究表明，在花园幽径句处理过程中，歧义区域长度、动词论元结构、语义合理性等因素均影响歧义消解难度，影响最终正确表征形成(Christianson et al., 2001; Sturt, 2007; Roberts and Felser, 2011)。

汉语花园幽径句的研究相对较少。在汉语花园幽径句分类方面，冯志伟(2003)系统归纳了汉语存在的四类潜在花园幽径句结构，陈满华(2009)将花园幽径句分为三类。在汉语花园幽径句分析中，汉语加工呈现“语义优先”特征，在没有适当句法结构的同时，语义处理仍在进行，表明语义信息对歧义消解起主导作用(Su, 2004; Xu and Huang, 2025)。但也有研究发现，汉语处理时对动词偏向性的反应比对语义合理性的反应更大，表明动词偏向性的约束力仍强于语义合理性因素(Qian, 2015)。此外，袁毓林(2025)发现，ChatGPT在汉语花园幽径句方面基本上可以理解句子表面的语言性意义，但在人类也难以理解的句子中表现较差。

花园幽径句的复杂性使其在自然语言处理领域也备受关注，研究者使用多种语言模型探索该现象的加工机制，但不同模型的表现呈现显著差异。RNNG、LSTM等模型均呈现花园幽径效应，但值得注意的是，仅有大模型才能利用动词偏向性线索(Futrell et al., 2019)。通过测量惊奇值(surprisal)，一些主流大语言模型如BERT、GPT-2等也展现出花园幽径效应，即在句子歧义点处出现较高的惊奇值，但其表现与人类存在显著差异(Hu et al., 2020; Jurayj et al., 2022; Irwin et al., 2023)。但是一些研究对大语言模型的惊奇值机制提出质疑，认为标准惊异值指标过度依赖词汇层级的统计共现关系，无法准确模拟人类的阅读时长，不能充分解释人类句法解歧困难(Cong et al., 2023; Huang et al., 2024)。于是有研究直接对比人类与大语言模型的语义理解问题准确率，发现大语言模型的解析表现、错误模式与人类表现类似。目前，以英汉双语语料为研究对象，用涵盖国内外不同参数规模、推理能力的多种大语言模型进行花园幽径句的研究还比较缺乏。

3 花园幽径句数据集构建

英语中比较典型的花园路径句从句法层面可以分为三类：

(1) 宾语/主语型(Direct Object/Subject)，这类句子中第一个动词似乎以名词短语作为其直接宾语，但随后发现根本没有直接宾语。如：While the man *hunted* the deer run into the woods.

(2) 宾语/补语型(Direct Object/Sentential Complement)，这类句子中第一个动词似乎以名词短语作为其直接宾语，但随后却发现有一个类似句子的宾语作为其补语。如：The scientist *read* the article had been published two months ago.

(3) 主动词/缩略关系型(Main Verb/Reduced Relative)，这类句子中，在消歧之前，歧义动词可以是句子的主要动词，也可以是引入简化关系从句的动词。如：The horse *raced* past the barn fell.

(c) 2025 中国中文信息学会计算语言学专业委员会

本研究主要选取宾语/补语型(DO/SC) 花园幽径句作为数据集建设来源。选取原因如下：首先，该类型存在跨语言的结构对应性，这为构建双语数据集提供了基础，另外两类英语花园幽径句在汉语中并没有对应的结构。例如英语中“直接宾语/句子补足语”结构，如“The patient on the bed believes the doctor will try his best to save him”，在汉语中也有类似的结构：“病床上的病人相信医生一定会尽力救他”。其次，宾语补语类花园幽径句具有一定的典型性。其结构特点使得在理解过程中容易产生初始误解，而后因补足语的出现被迫进行结构重组，从而形成典型的花园幽径效应。第三，现有研究多聚焦于主谓/偏正类、宾语/主语类结构引发的强花园幽径效应(Sturt et al., 1999)，而宾语/补语类尚未获得足够关注。

本文基于(Qian, 2015)的认知实验材料构建英汉双语DO/SC花园幽径句数据集BIT-GPS-576 (**BI**lingual **T**extual **G**arden **P**ath **S**entences) 。数据集按照2 (歧义性，2个水平：歧义、非歧义) ×2 (合理性，2个水平：合理、不合理) 的两因素完全组内设计。每组包含合理歧义句、合理非歧义句、不合理歧义句和不合理非歧义句四种条件，其中歧义句均为花园幽径句。基于Garnsey等人(1997)的预测试任务确定动词偏向性。英语数据集包含10个宾语偏向性动词和10个补语偏向性动词，每个动词重复使用4次，共生成80组句子，最终形成320个句子。英语花园幽径句的结构歧义性通过添加“that”进行歧义消解。汉语数据集包含11对动词，通过3次重复生成64组实验句，最终形成256个句子。汉语歧义句通过在主句动词后添加逗号来实现结构歧义消解。数据规模如表1所示。除此以外，我们还为英语和汉语的每个句子设计了对应的疑问句及答案。数据集的部分示例如下：

英语：

(1) 合理歧义句：The club members understood the bylaws would be applied to everyone.

(2) 合理非歧义句：The club members understood **that** the bylaws would be applied to everyone.

Question：Would the bylaws be applied to everyone? Answer：Yes

汉语：

(3) 合理歧义句：愤怒的记者揭露真相已经被封锁了。

(4) 合理非歧义句：愤怒的记者揭露，真相已经被封锁了。

Question：真相已经被封锁了吗？Answer：是

Table 1: 双语数据集分布情况

		语言类型	
歧义性	**动词偏向× 合理性**	**英语**	**汉语**
歧义	宾语偏向× 合理	40	32
	宾语偏向× 不合理	40	32
	补语偏向× 合理	40	32
	补语偏向× 不合理	40	32
非歧义	宾语偏向× 合理	40	32
	宾语偏向× 不合理	40	32
	补语偏向× 合理	40	32
	补语偏向× 不合理	40	32
总计		320	256

4 实验

基于前面构建的双语数据集，本部分开展了两个类型的跨语言实验，分别是花园幽径句句法结构分析实验和语义分析实验，以验证大模型对于汉英花园幽径句的分析能力。实验分别调用了目前主流大模型系列中一共10个模型的API (表 2) 作为测试工具。句法分析实验还对比了大模型与Stanford Parser的性能表现。

第二十四届中国计算语言学大会论文集，第574页-第586页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

Table 2: 实验选取大语言模型概览

系列	版本型号	数量
Qwen2.5	7B/14B/32B/72B-Instruct	4
Deepseek	V3, R1	2
GPT	4, 4o	2
Claude	3.7-sonnet	1
LLaMA	3	1

4.1 实验流程

实验流程如表 3所示。测试集均严格控制变量。大模型句法分析测试中，采用少样本提示(few-shot) 的Prompt，要求大模型依据宾州树库(Penn Treebank) (Taylor et al., 2003)的标注标准实现标注，评估模型对整句话的整体准确率。语义理解部分则评估指代消解、句意理解正确性。通过人工分析两类实验的结果确定最终的准确率。句法分析和语义分析均采用二元评分，正确记为1分，错误记为0分。在每个类别的实验中，首先进行分析实验，然后对实验结果进行统计检验。统计分析采用R语言(v4.4.3) 的lme4包构建广义线性混合效应模型(generalized linear mixed model, GLMM) 分别对句法和语义准确率进行多维度分析。

Table 3: 测试流程与示例

测试类型	句法测试	语义测试
Prompt	你是一个语言学实验参与者，需要使用嵌套括号以及宾州树库(Penn Treebank) 中的标签来画出所给句子的成分句法结构。只输出括号句法结构。 格式要求：S: [句子] P: [句法结构] 示例: S: 医生检查患者。P: (S (NP (NN 医生)) (VP (VV 检查) (NP (NN 患者))) (PU 。))	你是一个严格遵循指令的语言学实验助手。只能用「是」或「否」回答问题，不要添加任何解释。 格式要求：S: [句子] Q: [问题] A: [是/否] 正确示例：S: 医生被护士打电话叫来。Q: 护士是否给医生打了电话? A: 是
模型输入	S: 愤怒的记者揭露真相已经被封锁了。	S: 愤怒的记者揭露真相已经被封锁了。 Q: 真相已经被封锁了吗?
模型输出	P: (S (NP (ADJP (JJ 愤怒) (DEG 的)) (NP (NN 记者))) (VP (VV 揭露) (IP (NP (NN 真相)) (VP (ADVP (AD 已经)) (VP (SB 被) (VP (VV 封锁) (AS 了))))))) (PU 。))	A: 是

4.2 句法分析实验

4.2.1 大模型实验结果及分析

我们发现大语言模型整体呈现出花园幽径效应，且可以利用动词偏向性和语义合理性线索进行消歧，但在英汉两种语言条件下差异较大。表 4呈现了大模型在不同条件下对英汉花园幽径句的句法分析结果平均准确率(μ) 以及准确率的方差(σ^2) 。汉语和英语的综合集是指表 1中全部的歧义句和非歧义句，歧义集仅包含汉语和英语的歧义句。

在综合测试集分析中，大模型整体表现出花园幽径效应，非歧义句分析准确率显著高于歧义句分析准确率($\chi^2 = 396.43$, $p < 2.2 \times 10^{-16}$)。具体来看，在英汉两种语言条件下，大模型均呈现出显著的花园幽径效应($p < 0.05$)，其中汉语花园幽径效应更为显著(汉语：$\chi^2 = 296.5$，英语：$\chi^2 = 177.3$)。为进一步探究大模型处理花园幽径句的表现，控制动词间、模型间随机差异，我们构建广义线性混合效应模型(GLMM) 进行分析。模型设定及结果如表 5所示。随机效

(c) 2025 中国中文信息学会计算语言学专业委员会

应分析显示，模型间差异构成主要随机差异来源，动词差异对句法分析准确率影响较小。固定效应分析显示，语言类型是影响模型表现的最强因素，其中汉语呈现显著负向效应，即汉语条件下的准确率显著降低。动词偏向性、语义合理性分析显示，补语偏向性动词引导的花园幽径句句法分析准确率显著提升，而含有语义合理名词的花园幽径句分析准确率显著降低，表明大模型可以利用动词偏向性、语义合理性线索完成花园幽径句句法结构消歧，这与人类认知实验结果类似(Trueswell et al., 1994; Garnsey et al., 1997)。

Table 4: 不同模型句法分析平均准确率(%) 比较

模型名称	汉语综合集 准确率μ (σ^2)	英语综合集 准确率μ (σ^2)	汉语歧义集 准确率μ (σ^2)	英语歧义集 准确率μ (σ^2)
DeepseekR1	**96.9 (0.17)**	**100.0 (0.00)**	**94.5 (0.23)**	**100.0 (0.00)**
DeepseekV3	92.2 (0.27)	99.7 (0.06)	84.4 (0.37)	99.4 (0.08)
Claude3.7	91.0 (0.29)	100.0 (0.00)	82.8 (0.38)	100.0 (0.00)
GPT-4o	66.4 (0.47)	100.0 (0.00)	35.2 (0.48)	100.0 (0.00)
GPT-4	50.0 (0.50)	99.4 (0.08)	21.1 (0.41)	100.0 (0.00)
Qwen2.5-72B	44.9 (0.50)	98.8 (0.11)	44.5 (0.50)	97.5 (0.16)
Qwen2.5-32B	38.7 (0.49)	96.3 (0.19)	27.3 (0.45)	95.0 (0.22)
LLaMA3	32.8 (0.47)	58.8 (0.49)	11.7 (0.32)	17.5 (0.38)
Qwen2.5-14B	18.4 (0.39)	50.6 (0.50)	2.3 (0.15)	18.1 (0.39)
Qwen2.5-7B	4.3 (0.20)	45.3 (0.50)	0.0 (0.00)	11.9 (0.33)

Table 5: 大模型句法分析准确率GLMM分析结果

固定效应	参数	估计值	标准误	z值	p值
	截距	2.96	1.01	2.94	0.003**
	语言类型(汉语)	-3.99	0.27	-14.77	<2e-16***
	动词偏向(补语)	0.46	0.13	3.59	0.0003***
	合理性(合理)	-0.99	0.13	-7.46	8.68e-14***
随机效应	参数	方差	标准差		
	动词(截距)	0.15	0.39		
	模型(截距)	9.68	3.11		

注：*** 表示$p < 0.001$，** 表示$p < 0.01$，* 表示$p < 0.05$，· 表示$p < 0.1$，无标记表示$p \geq 0.1$。

由于英汉不同语言下模型分析能力存在显著差异，后续我们对英汉数据分别进行GLMM分析。英语数据集分析结果显示，英语条件下大模型出现显著的花园幽径效应，而动词偏向性和语义合理性效应不显著。分析发现是因为大模型在英语条件下准确率均较高，接近或已达完美水平，各条件下差异不大。而汉语数据集分析结果显示，歧义性、动词偏向性和语义合理性效应均显著。

汉语歧义句导致句法分析困难且动词偏向和语义合理因素作用显著，或许源于其分词复杂性、语义多样性以及数据的稀缺性。首先，汉语缺乏显性的词边界标记，导致分词规范难以统一，需依赖上下文信息与句法、语义推理。第二，汉语缺乏形态变化，同一个词形可以表示多种不同词义甚至词性，这种现象使得汉语处理难度增大且高度依赖上下文信息。最后，汉语本身的复杂性导致高质量标注语料不足，汉语语料库规模和多样性相对有限，限制了模型在面对歧义句时准确判断其句法结构和语义信息的能力，例如GPT-3训练语料中汉语仅占到0.1%。

我们将不同大语言模型表现结果对比发现，模型对花园幽径句的句法解析表现存在显著差异。参数规模较大的模型、推理能力较强的模型表现更好，且国产大模型在汉语方面表现更好。在完整的双语测试集(n=576) 中，Kruskal-Wallis检验显示模型间存在显著差异，进一步采用Dunn's事后检验，发现Deepseek-R1展现出最优异的性能，语言处理能力显著高于其他模

(c) 2025 中国中文信息学会计算语言学专业委员会

型，且显示出了处理稳定性；Deepseek-V3与Claude3.7的准确率差异不大，都远高出GPT系列近30个百分点。相比之下，Qwen2.5-7B/14B-Instruct的准确率在综合测试集中均显著低于其他模型，且存在较大波动性，表明其处理机制具有不稳定性。在歧义句测试集中，各模型的准确率排名与综合测试集基本一致，这一现象表明模型对复杂语言现象的处理能力与其整体语言理解水平存在强相关性。

横向比较Qwen系列不同参数规模的模型发现，模型的语言理解能力与参数规模呈现显著正相关。具体而言，在歧义句处理任务中，Qwen2.5-7B平均准确率不到10%，Qwen2.5-14B表现虽有所提升，但相较于32B的版本仍存在数量级差异。当参数规模达到72B时，模型性能进一步取得显著进展。表明一定范围内，随参数规模增大，模型的语言能力增强，对花园幽径句消歧能力增强，表现更稳定。

在模型架构相同的前提下，推理能力的优化对歧义处理效果产生显著积极影响。GPT-4o相较于GPT-4实现较大性能提升，这一趋势在Deepseek系列中也有所体现：在汉语歧义集测试中Deepseek-R1较Deepseek-V3的平均准确率提升达10个百分点。两个系列的优化模型(GPT-4o和Deepseek-R1) 在保持较低标准差的同时，均实现了统计显著的性能提升，表明推理优化对语言消歧具有有效性。

4.2.2 Stanford Parser实验结果及分析

Stanford Parser呈现花园幽径效应，但只能利用动词偏向性句法线索而不能利用语义线索辅助消歧。Stanford Parser对双语数据集进行句法分析的结果如图1所示。为探究Parser处理综合测试集的表现，控制动词间随机差异，我们构建GLMM进行分析(表6) 。随机效应分析显示，动词差异对句法分析准确率影响较小。固定效应分析显示，语言类型是影响模型表现的最强因素，其中汉语呈现显著负向效应，即汉语条件下的准确率显著降低。在歧义性效应方面，综合测试集结果显示该模型呈现显著的花园幽径效应，花园幽径句句法分析准确率显著低于非花园幽径句，表明模型在面临句法结构复杂的句子时，仍存在处理困难。模型对动词偏向性呈现显著响应，补语偏向性动词较宾语偏向性动词引导的句子分析准确率有所提升。值得注意的是，语义合理性因素未表现出显著效应，这一结果和(Qian, 2015)认知实验研究一致，表明传统句法分析器无法利用语义信息线索。

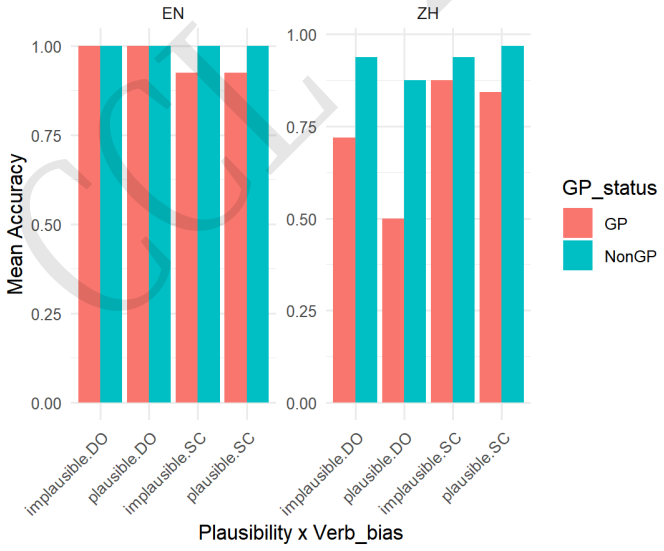

Figure 1: Stanford Parser在综合测试集中的平均准确率

Parser在英汉跨语言任务中呈现显著差异。类似于大模型的表现，英文测试集在各条件下的句法分析准确率同样显著高于中文测试集，且英文测试集表现更稳定，花园幽径句(GP) 与非花园幽径句(NonGP) 分析准确率差异较小,均高于90%；而汉语测试集表现差异较大，花园幽径句的分析性能普遍低于非幽径句，显示出极强的花园幽径效应，验证了汉语语言特性使得大模型动态分析与Parser静态分析均遇到较大难度。

(c) 2025 中国中文信息学会计算语言学专业委员会

Table 6: Stanford Parser句法分析准确率GLMM分析结果

固定效应	参数	估计值	标准误	z值	p值
	截距	3.74	0.63	5.92	<0.001***
	语言(汉语)	-2.77	0.63	-4.36	<0.001***
	歧义水平(非GP)	1.81	0.40	4.47	<0.001***
	动词偏向(补语)	0.73	0.34	2.13	0.033*
	合理性(合理)	-0.50	0.34	-1.49	0.137
随机效应	参数	方差	标准差		
	动词(截距)	0.56	0.75		

4.2.3 大模型与Stanford Parser对比分析

基于前面的实验结果，我们对比了大语言模型与Parser在英汉双语歧义句测试集中的分析能力，发现传统句法分析器在花园幽径句歧义句法分析方面展现出与较大语言模型相当的性能水平，且其汉语消歧能力超越部分综合能力很强的模型如GPT系列。如图2所示，Stanford Parser与Deepseek系列、Claude3.7模型准确率呈现高度相似性，而与Qwen2.5系列小参数模型(7B/14B) 差异最大。值得注意的是，GPT-4系列模型虽在自然语言处理基准测试中取得较高成绩(Achiam et al., 2023; Liu et al., 2024)，其在花园幽径句句法分析任务中准确率也低于Parser及Deepseek系列模型。采用McNemar检验进一步分析了Parser与不同LLMs在不同句子条件下的表现一致性。如图3所示，在英语歧义句处理中，Parser仅与Qwen2.5-7B/14B-Instruct，LLaMA3一致性较差且Parser表现更佳；与其他模型均达到90%以上的高度一致性。深入分析发现，LLaMA3模型主要是由于指令遵循问题，因自行添加消歧词“that”导致句法结构改变，若采用宽松评分标准(仅评估基础句法正确性，不考虑多余消歧词，一致性也可达90%以上)。在汉语歧义句任务中，Parser与Claude3.7分析准确率一致性最高，与Deepseek系列一致性下降，与GPT系列一致性差异更大，此时Deepseek系列准确率高于Parser，而GPT系列准确率低于Parser。

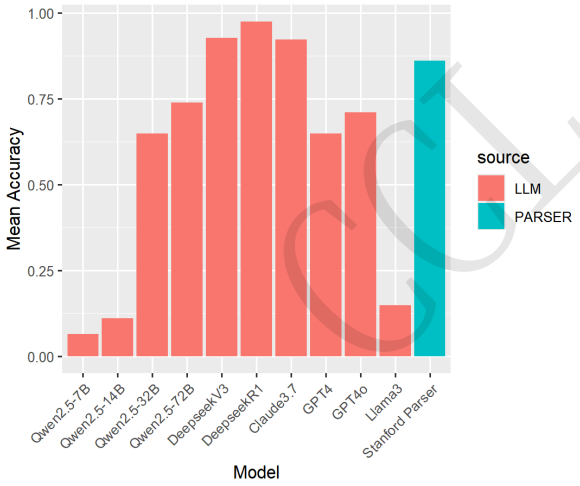

Figure 2: LLMs与Parser歧义句处理准确率比较

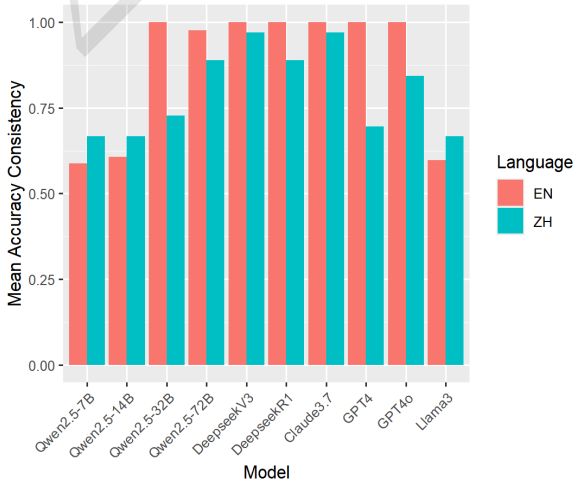

Figure 3: LLMs与Parser对不同语言歧义句处理表现一致性

实证结果揭示大语言模型与Stanford Parser在花园幽径句处理机制上存在共性与差异。共性体现在二者皆呈现花园幽径效应，都可以利用句法线索辅助消歧，且在汉语句子中遇到更大挑战。首先，两种系统均表现出与人类认知相似的局限性。具体而言，在英汉测试集中，大语言模型与Parser在非歧义句子中准确率表现显著提升，表现出显著的花园幽径效应，说明花园幽径句不仅对人类认知构成挑战，对计算模型同样产生显著的分析负荷。其次，跨语言比较显示，大语言模型与Parser在汉语测试集遇到更大挑战，英语句法分析准确率显著高于相同条件下的汉语任务。此外，两种系统在消歧过程中都体现对动词偏向性这一句法线索的敏感性，补

(c) 2025 中国中文信息学会计算语言学专业委员会

语偏向性动词消歧准确率显著高于宾语偏向性动词，这说明两种系统均可以利用句法线索完成消歧。

但是，两种系统也存在差异，只有大语言模型可以利用语义线索。大语言模型整体呈现出动词偏向性、语义合理性均有显著影响；相比之下，Parser虽显示出动词偏向性效应，但语义合理性维度未呈现显著效应，表明二者在利用语义和句法的不同线索方面存在差异。大语言模型是基于上下文预测和概率分布建模，在大规模语料中不仅学习到动词偏向性，也学习到语义搭配的共现关系，可以融合上下文语义、常识、世界知识进行合理性判断。相比之下，传统句法解析器如Stanford Parser无法处理语义合理性，只依赖明确的句法规则和概率结构来选择最可能的句法结构。

4.3 语义分析实验

由于Stanford Parser无法回答语义问题，因此只采用大语言模型进行语义分析测试。

大模型在语义分析测试中未显示出显著的花园幽径效应，但仍呈现出英汉不同语言下的显著差异。大模型对花园幽径句语义分析能力结果如表7所示。初步分析英汉歧义句平均准确率发现，模型对英文歧义句的分析能力显著优于中文歧义句，且英文歧义句分析能力更稳定，模型间差异更小。为进一步探究大语言模型在综合测试集的准确率表现，控制动词间、模型间随机差异，构建GLMM进行分析(表8)。随机效应分析显示，动词间差异显著高于模型间差异，构成主要随机差异来源，而模型间差异对语义分析准确率影响较小。固定效应分析显示，语言模型未显示出显著花园幽径效应。语言类型主效应最为显著，汉语条件下的语义分析准确率显著降低。动词偏向性类型对准确率影响显著，补语偏向性动词引导的句子分析准确率显著提升。此外，语义合理性效应达到边际显著，大模型在阅读含合理名词的句子时会更容易理解。表明在语义测试中，大语言模型主要利用动词偏向性线索进行分析，且不同的动词词义可能造成显著影响。

Table 7: 不同模型语义分析平均准确率(%) 比较

模型名称	汉语综合集 准确率μ (σ^2)	英语综合集 准确率μ (σ^2)	汉语歧义集 准确率μ (σ^2)	英语歧义集 准确率μ (σ^2)
Llama3	**88.28 (0.32)**	7.50 (0.16)	**88.28 (0.32)**	98.75 (0.11)
GPT4	85.55 (0.35)	**99.69 (0.06)**	83.59 (0.37)	**100.00 (0.00)**
GPT4o	78.91 (0.41)	97.19 (0.17)	78.91 (0.41)	96.88 (0.17)
Claude3.7	78.91 (0.41)	94.69 (0.22)	78.13 (0.42)	93.75 (0.24)
Qwen2.5-72B	77.34 (0.42)	95.00 (0.22)	76.56 (0.43)	95.00 (0.22)
DeepseekV3	76.95 (0.42)	95.63 (0.20)	76.56 (0.43)	95.63 (0.21)
Qwen2.5-32B	72.66 (0.45)	93.13 (0.25)	72.66 (0.45)	93.13 (0.25)
Qwen2.5-14B	70.31 (0.46)	91.25 (0.28)	69.53 (0.46)	91.88 (0.27)
DeepseekR1	67.97 (0.47)	88.75 (0.32)	64.84 (0.48)	89.38 (0.31)
Qwen2.5-7B	64.45 (0.48)	92.50 (0.26)	62.50 (0.49)	92.50 (0.26)

在不同模型间差异方面，模型的语义分析能力与参数规模仍呈现正相关，但与推理能力呈现负相关。针对Qwen系列不同参数规模的模型进行横向比较发现，模型的语言理解能力与参数规模呈现显著正相关。具体而言，在汉英综合集和歧义集测试中，Qwen2.5-72B模型表现最佳，Qwen2.5-32B/14B/7B模型准确率依次降低。参数规模变化影响在汉语中更为显著，模型间准确率均值差异更大。表明在语义分析方面，随参数规模增大，模型的语言能力增强，对花园幽径句语义理解能力同时增强。

在模型架构相同的前提下，推理能力的优化对歧义处理效果产生显著消极影响。GPT-4o相较于GPT-4准确率均值降低，这一趋势在Deepseek系列中也有所体现，在汉语歧义集测试中Deepseek-R1较Deepseek-V3的平均准确率降低近12个百分点。两个系列的推理优化模型(GPT-4o和Deepseek-R1) 在花园幽径句语义分析方面，性能均显著降低，表明推理优化在一定程度上加重模型幻觉，复杂推理模块可能引入语义分析干扰。

第二十四届中国计算语言学大会论文集，第574页-第586页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

Table 8: 大模型语义分析准确率GLMM分析结果

固定效应	参数	估计值	标准误	z值	p值
	截距	3.1208	0.3588	8.699	<0.001***
	语言(汉语)	-2.1111	0.4916	-4.295	<0.001***
	GP水平(非GP)	0.1289	0.1015	1.270	0.204
	动词偏向(补语)	0.8819	0.0896	9.849	<0.001***
	合理性(合理)	0.1769	0.0871	2.031	0.042*
随机效应	参数	方差	标准差		
	动词(截距)	1.5774	1.2559		
	模型名称(截距)	0.2913	0.5397		

此外，我们将大语言模型语义测试结果与人类表现比较，发现大部分大语言模型在英语语义问题回答方面已超过人类表现。在英语综合集测试中，大语言模型的平均准确率达到94.53%，而(Qian, 2015)以该测试集对人类进行语义理解测试时，英语母语者语义理解平均准确率为93%，而非母语者仅有86%，低于大多数所选模型的测试结果。

4.4 句法与语义实验对比

我们将大模型对于花园幽径句的句法和语义分析准确率进行对比，以检验大模型的句法分析和语义分析是否一致。如图4所示，不论幽径句(GP) 还是非幽径句(NonGP)，语义分析与句法分析均存在较大差异。表9展示了同一个句子中句法和语义分析不一致的一些示例(句法分析结果省略) 。

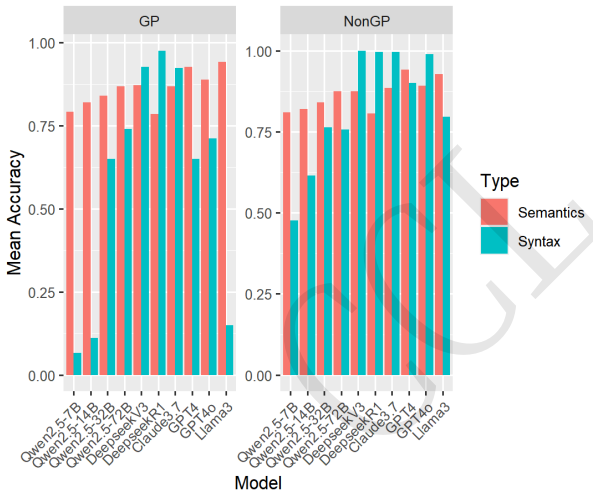

Figure 4: LLMs句法语义准确率对比

Table 9: 句法与语义分析示例

	句法分析正确，语义分析错误
汉语	市民抱怨幸福总是那么遥不可及。 Q：幸福总是那么遥不可及吗？A：否
英语	The investigator inferred the evidence meant the suspect was not guilty. Q: Did the evidence mean the suspect was not guilty? A: No
	句法分析错误，语义分析正确
汉语	获奖的作家得知喜讯早已传到他的单位。 Q：喜讯早已传到了获奖作家的单位吗？A：是
英语	They accepted the application did not include necessary files. Q: Did the application not include necessary files? A: Yes

在GP中参数较大的模型句法与语义准确率较为相近，表明随参数增大，模型的形式句法与语义分析能力增强，表现更相似。在非歧义句处理中，模型间句法和语义准确率均值差异减小，表明模型在非歧义句测试中表现更为稳定，未遇到分析困难；但也展现出参数效应，Qwen系列模型随参数规模增大，模型语言能力增强。此外，我们注意到在两类测试集中，Deepseek系列、GPT-4o及Claude3.7呈现句法准确率高于语义准确率的趋势，表明模型未依据句子的句法结构进行语义分析，存在语义推理错误。

整体而言，模型的句法分析和语义分析不一致，对语义的判断准确率高于对句法的判断准确率。我们对不同实验条件下的语义与句法分析准确率进行Pearson相关分析，如表10所示。结

(c) 2025 中国中文信息学会计算语言学专业委员会

果显示，在歧义句条件下，句法分析的准确率下降，表现出典型的花园幽径句效应，而语义分析的准确率变化较小，因此非歧义水平下二者相关性更高。进一步分析显示，动词偏向性对句法与语义准确率之间的一致性产生较大影响。语义合理性对语义句法相关性也具有影响，在语义合理条件下，句法分析的准确率下降，而语义分析的准确率上升。

Table 10: 各条件下大模型句法语义分析相关性结果

因素	水平	语义分析准确率均值	句法解析准确率均值	Pearson相关性
歧义性	歧义	0.85	0.57	0.31
歧义性	非歧义	0.86	0.81	0.44
动词偏向性	宾语	0.80	0.68	0.47
动词偏向性	补语	0.90	0.71	0.13
语义合理性	不合理	0.85	0.71	0.33
语义合理性	合理	0.86	0.67	0.36

5 结语

本文构建跨语言花园幽径句数据集，使用大语言模型及传统句法分析器进行实验，系统展示了大语言模型处理花园幽径句的性能表现。本实验的主要结论如下：首先，在句法分析测试中语言模型呈现与人类认知实验中类似的花园幽径效应，且其消歧能力随模型参数规模、推理能力的增长而增强。英语歧义句处理准确率显著高于汉语对应结构，且LLMs可以利用语义合理性、动词结构偏向性线索提升消歧能力。其次，Stanford Parser与大型语言模型表现更为一致，可以利用动词偏向性线索消歧，但无法利用语义信息，表明传统句法分析器在句法分析方面仍具有架构优势，但也存在一定局限性。第三，大语言模型形式句法分析准确率与语义分析准，确率呈现不对称性，提示语言模型在句法分析与语义分析上具有较大的差异性，且表现出较强的语义分析能力，这说明大语言模型在语言处理过程中并非依赖严格的形式句法规则。构建的跨语言花园幽径句数据集和实证实验结论为跨语言的语言理解研究提供了新证据，为基于大语言模型的语言处理以及语言认知领域提供了新的视角。

致谢

本研究受到国家社科基金项目（24CYY108、21CYY046）资助。感谢匿名评审为本文提出的宝贵意见和建议。

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2025. When the lm misunderstood the human chuckled: Analyzing garden path effects in humans and language models. *arXiv preprint arXiv:2502.09307*.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Thomas G Bever. 1970. The cognitive basis for linguistic structures. *Cognition and the development of language*.

Trevor Brothers, Liv J Hoversten, and Matthew J Traxler. 2021. Bilinguals on the garden-path: Individual differences in syntactic ambiguity resolution. *Bilingualism: Language and cognition*, 24(4):612–627.

(c) 2025 中国中文信息学会计算语言学专业委员会

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Kiel Christianson, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive psychology*, 42(4):368–407.

Yan Cong, Emmanuele Chersoni, Yu-yin Hsu, and Alessandro Lenci. 2023. Are language models sensitive to semantic attraction? a study on surprisal. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 141–148, Toronto, Canada, July. Association for Computational Linguistics.

Ian Cunnings. 2017. Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4):659–678.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Susan M Garnsey, Neal J Pearlmutter, Elizabeth Myers, and Melanie A Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of memory and language*, 37(1):58–93.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online, July. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Tovah Irwin, Kyra Wilson, and Alec Marantz. 2023. BERT shows garden path effects. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3220–3232, Dubrovnik, Croatia, May. Association for Computational Linguistics.

William Jurayj, William Rudman, and Carsten Eickhoff. 2022. Garden path traversal in GPT-2. In Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe, editors, *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 305–313, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Eun-Kyung Lee, Dora Hsin-Yi Lu, and Susan M Garnsey. 2013. L1 word order and sensitivity to verb bias in l2 processing. *Bilingualism: Language and Cognition*, 16(4):761–775.

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention. *arXiv preprint arXiv:2405.16042*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zhiying Qian, Eun-Kyung Lee, Dora Hsin-Yi Lu, and Susan M Garnsey. 2019. Native and non-native (l1-mandarin) speakers of english differ in online use of verb-based cues about sentence structure. *Bilingualism: Language and cognition*, 22(5):897–911.

Zhiying Qian. 2015. *The reanalysis and interpretation of garden-path sentences by native speakers and second language learners*. University of Illinois at Urbana-Champaign.

(c) 2025 中国中文信息学会计算语言学专业委员会

Leah Roberts and Claudia Felser. 2011. Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics*, 32(2):299–331.

Patrick Sturt, Martin J Pickering, and Matthew W Crocker. 1999. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40(1):136–150.

Patrick Sturt. 2007. Semantic re-interpretation and garden path recovery. *Cognition*, 105(2):477–488.

I-Ru Su. 2004. The effects of discourse processing with regard to syntactic and semantic cues: A competition model study. *Applied Psycholinguistics*, 25(4):587–601.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks: Building and using parsed corpora*, pages 5–22.

John C Trueswell, Michael K Tanenhaus, and Susan M Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, 33(3):285–318.

Yulian Xu and Xianjun Huang. 2025. Neural mechanisms of dynamic syntactic and semantic processing in chinese garden-path sentence comprehension: An erp study. *Journal of Neurolinguistics*, 73:101233.

冯志伟. 2003. 花园幽径句的某些形式特性. In 语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集, pages 189–195.

卢华萍and 吴明军. 2021. 不同句法结构对二语花园路径句重新分析的影响研究. 现代外语, 44(02):233–245.

吴迪, 李苗苗, and 吴明军. 2024. 中国学习者二语花园幽径句解读研究. 外语学刊, (04):70–78.

杜家利and 于屏方. 2016. 中国英语学习者花园幽径句错位效应强度研究:计算语言学视角. 中文信息学报, 30(06):100–116.

杜家利and 于屏方. 2018. 花园幽径句解码效果与反应时的关联性研究. 中文信息学报, 32(04):13–23+30.

王邵馨, 陈士法, and 杜慧楠. 2025. 英语“直接宾语/主语”类花园路径句加工中的重新分析研究. 外语教学与研究, 57(01):92–105.

袁毓林. 2025. 从三种复杂句看chatgpt是不是随机鹦鹉?——语言大模型能不能理解语言意义的测试与讨论. 语言教学与研究, (01):35–49.

陈满华. 2009. 花园幽径句的层级、产生机制和修辞效果. 修辞学习, (04):43–48.

CCL 2025

(c) 2025 中国中文信息学会计算语言学专业委员会