2025.12.31 语法理论与语言工程讨论班

基于大语言模型概率信息的汉语结果补语错误检测与句式推荐研究

中国语言文学系 金慧京

01

引言

结果补语在汉语句法与意义表达中具有关键作用，但因其他语言缺乏对应结构，学习者偏误频繁。

随着补语研究的多元化与语言信息处理技术的发展，有必要探索新的教学与分析路径。

本研究尝试利用大语言模型的概率语言知识，探讨结果补语的学习难点辅助应用可能性

01

学习者偏误考察

引言

残缺	多余	替代	语序颠倒	分离·回避	其他

[残缺]

〈虚化补语残缺〉

*丈夫看这样的妻子，他也很难受。

⇒丈夫看到这样的妻子，他也很难受。

〈实义补语残缺〉

*每天妈妈把房间打扫以后出门。

⇒每天妈妈把房间打扫干净以后出门。

〈动词残缺〉

*我把书懂了。

⇒我把书看懂了。

*把作业错了。

⇒把作业写错了。

[多余]

*从那时候起我开始睁开大眼睛。

⇒从那时候起我开始睁大眼睛。

*我看完这个电影以后，很想到奶奶。

⇒我看完这个电影以后，很想奶奶。

难点一：结果补语的功能及使用判断

——何时应当使用，何时无需使用

多与**虚化补语**有关

01

学习者偏误考察

引言

残缺	多余	替代	语序颠倒	分离·回避	其他

[替代]

*医生们一定会救好他。

*医生们一定会救着他。

⇒医生们一定会救活他。

*我一来到中国就喜欢成了中国。

⇒我一来到中国就喜欢上了中国。

[语序颠倒]

*我好好儿吃了，可以出发了。

⇒我吃好了，可以出发了。

*老师已经上课了，我晚来了。

⇒老师已经上课了，我来晚了。

引言

残缺	多余	替代	语序颠倒	分离・回避	其他

[分离・回避]

*唱歌儿完了以后，我们就回学校了。

⇒唱完歌儿以后，我们就回学校了。

我听老师的话，懂了。

⇒我听懂了老师的话。

[其他]

*她把那个礼物气扔了。

⇒她生气地扔掉了礼物。

*昨天的聚会，他醉到了。

⇒昨天的聚会，他喝醉了。

难点二：结果补语句的形式繁多

——即使掌握了其基本机制，学习者仍常常无法确定应当选择哪一种表达方式。

多与**实义补语**的使用相关

引言

难点二：结果补语句的形式繁多

——即使掌握了其基本机制，学习者仍常常无法确定应当选择哪一种表达方式。

- 这类偏误不易被察觉，表面上出现频率不高，
  但它反映着学习者在生成结果补语这一核心结构时的困难。
- 由于最优的结果补语句式会随具体语境与词汇搭配而变化，
  因此难以单一规则加以概括

[回避]

我听老师的话，我懂了。

⇒我听懂老师的话了。

？我把老师的话听懂了。

我们打扫房间，房间干净了。

⇒我们把房间打扫干净了。

？我们打扫干净房间了。

01

引言

能否利用大语言模型的语言分布知识来辅助学习者?

- 大语言模型通过大规模语料学习了语言的自然分布，能对“自然”与“不自然”的表达做出概率判断。

- 基于这种分布性知识，是否可以：
  1）**判定学习者的结果补语错误**
  2）**在多种可能句式中推荐最自然、最合适的表达？**

01

目录

02

结果补语知识的构建与应用

结果补语已通过多种方法进行研究、归纳与总结

相关成果为补语语料的标注与归档提供了知识基础

从学习者与教学的角度来看，语言学知识的积累应服务于“更好地理解与掌握”这一目标

在选择整理语言资料的体系时，需要重新思考：

学习者需要什么？应以何种方式传递哪些知识？

02

结果补语知识的构建与应用

范畴奠基 ~2000	概念界定
	明确补语的句法地位
范畴化与理论化 2000初	建立系统的理论框架
	追求句法分析的精细化，强调语义解释一致与协调
研究课题细化 2000初 ~2010中	采用多种理论与分析方法，对单一成分、句式的深入研究
	教学视角的引入：补语偏误分析逐渐增多
研究层面扩展 2010中 ~2020	引入语用学、认知语言学视角
	探讨补语与时体等其他语法范畴的关联
对比研究与技术融合 2020~	跨语言对比研究加强
	语言模型等新技术在补语研究中的应用

02

结果补语知识的构建与应用

范畴化与理论化

2000初

建立系统的理论框架

追求句法分析的精细化，强调语义解释一致与协调

1．结果补语

（一）从补语的语义指向角度来认识结果补语

1．1 补语指向述语动词

1．1．1 补语表示动作行为的结果：讲错

1．1．2 补语表示动作行为的状态：看久

1．2 补语指向体词性成分

1．2．1 补语指向述语的语义角色

1．2．1．1 施事：吃饱、洗累

1．2．1．1．1 施事的构件部分：看花（看花了眼）、喊哑（喊哑了嗓子）

1．2．1．2 感事：冷醒、热哭

1．2．1．3 受事：打碎、赶走

1．2．1．4 工具：砍钝（砍钝了两把刀）、写秃

1．2．1．5 结果：织长（毛衣织长了）、做大

1．2．1．6 处所：坐满（教室里坐满了人）、找遍

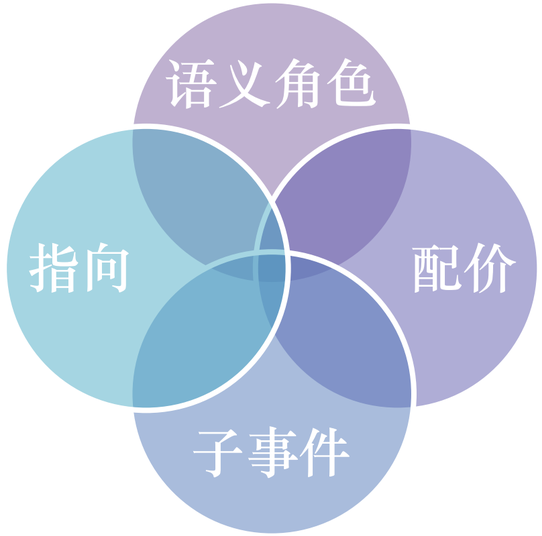
1．2．1．7 动作行为的受影响者 （与事）：讲笑（他讲笑话把大家都讲笑了）、哭烦（这孩子整天哭，把人都哭烦了）

（二）从补语的表义功能角度来认识结果补语

结果补语可以区分为表达“自然结果” 和表达“对结果的评价”两种情况。

“买累了”“吃饱了” 是 “买、吃”动作发生后的自然结果。

“买贵了”“吃早了” 是 “买、吃”动作发生后，说话者对结果的评价。



结构分解 → 概念理解

类型化与结构化

生成上的困难同样值得关注

02

结果补语知识的构建与应用

1. **虚化结果补语**由于语义虚化程度较高，
其是否还能作为**因果关系**中的独立成分（即子事件的组成部分）并不明确，
更侧重于提供事件的时体或现实性信息。

种子不久就会萌发，**长成**幼苗。
→种子长+成幼苗

我们把他**怀疑成**敌人了。
→我们怀疑他+?他成敌人

我们必须不惜一切代价把答案**找到**。

这只鸟儿一叫，其他的鸟儿也**叫开**了。

“先走一步”，关键在“先”，只要**占住**“先”字，
尽管先“一步”，就能抢占先机，掌握主动。

补语在虚化之后，其词汇意义逐渐减弱，
更应该被视为充当动词的语法性辅助成分。

因此与动词之间的紧密性显著提高，
使得二者在意义上难以分离。

关于学习者对虚化结果补语的掌握，
重点在于虚化补语的使用判断：何时用、何时不用。

02

结果补语知识的构建与应用

2. 按照配价或语义指向的分类标准来看，一些述补结构被归入同一类，但其实际所构成的句式类型却相当多样。

[一价术语+一价补语] [述补语义均指向施事]

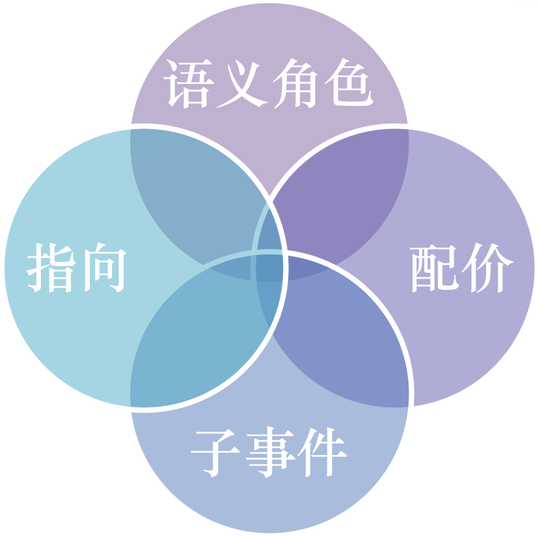
小王**起晚**了。
孩子**睡醒**了。
紫罗兰已经**开败**。
雪真的**下大**了。

他**累病**了。
妈妈**急哭**了。
他的脸**涨紫**了。

我**热死**了。
我**吓蒙**了。

02

结果补语知识的构建与应用



2. 按照配价或语义指向的分类标准来看，一些述补结构被归入同一类，但其实际所构成的句式类型却相当多样。

[一价术语+一价补语] [述补语义均指向施事]

小王**起晚**了。
孩子**睡醒**了。
紫罗兰已经**开败**。
雪真的**下大**了。

把他**累病**了。
让妈妈给**急哭**了。
他**涨紫**了脸。

热死我了。
我被**吓蒙**了。

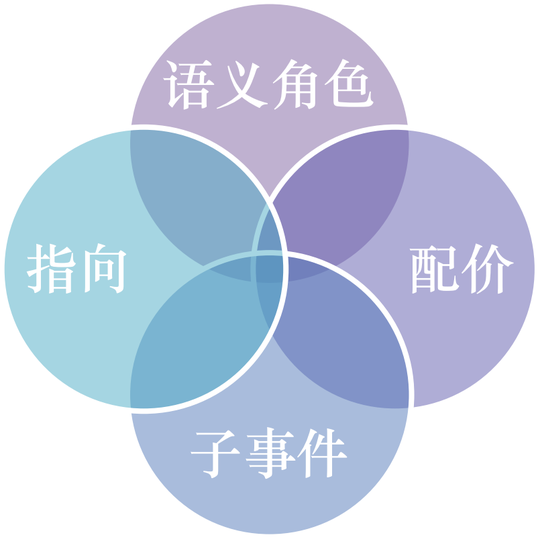
受影响性高

3. 当配价体系之外的成分以类似论元的方式进入结构时，会出现许多难以用现有框架解释的现象。

三天就让我给**累病**了。
孩子发高烧把妈妈**急哭**了。
这天气快把我**热死**了。

02

结果补语知识的构建与应用



→ 汉语述补结构在论元的分布与排列上表现出高度的灵活性。

我看电影**看哭**了
小宝宝**哭醒**了
我**羡慕死**你们了
我**担心死**孩子了
我**跑**了八百米**累**了

电影给我**看哭**了
一场噩梦**哭醒**了小宝宝
你们的故事可**羡慕死**我了
孩子胃胀、呕吐，都快**担心死**我了
八百米就把我**跑累**了

论元的出现位置可能发生倒置

我吃饱饭了
？我吃饱一小碗饭了

*饭吃饱了我
一小碗饭就让我吃饱了

论元可出现的位置似乎与其语义的“具体性”有关。

02

结果补语知识的构建与应用

连接理论(linking theory)

1. 连接理论的基本假设

语义角色（谁做出行动/谁受影响）与句法位置（主语/宾语）之间存在相对稳定的对应关系。
这种语义→句法的映射关系在同一语言内部、甚至跨语言之间都被认为具有一定的普遍性。

2. 语义→句法的映射依据 ： 意义角色层级（Thematic Hierarchy）

语义角色之间存在一个“高低”排序，
层级越高的角色越倾向于出现在句法的高位置（如主语），
层级越低的角色越倾向于出现在局发的低位置（如宾语）。

Agent> Experiencer> Goal/Source/Location> Theme

Grimshaw(1990)

3. 为什么会出现“连接问题（linking Problem）”？

在某些结构（如心理动词句、结果补语句）中，
如果对论元的意义角色及其影响关系的解读存在分歧，
则可能导致语义—句法映射方式发生变化（甚至出现位置对调）。

Hartshorne, J. K et al.(2016), Psych Verbs, the Linking Problem, and the Acquisition of Language, *Cognition*, Volume 157.

02 连接问题（linking Problem）

结果补语知识的构建与应用

我看电影看哭了	电影给我看哭了
小宝宝哭醒了	一场噩梦哭醒了小宝宝
我羡慕死你们了	你们的故事可羡慕死我了
我担心死孩子了	孩子胃胀、呕吐，都快担心死我了
我跑了八百米累了	八百米就把我跑累了
我吃饱饭了	？饭吃饱了我
？我吃饱一小碗饭了	一小碗饭就让我吃饱了
Agent > Patient	**Cause > Causee**
[X DO Y, X BECOME Z]	[Y CAUSE X to BECOME Z]

结果补语本质上体现的是因果关系，即由行为引发的状态变化（BECOME）
对于同一变化事件，若将其理解为：

（1）施事(agent)在对受事(patient)的行为过程中所产生的变化

（2）原因(cause)对受使役者(causee)施加作用而引发的致使性变化

则两种不同的事件解读会导致论元排列方式的差异。

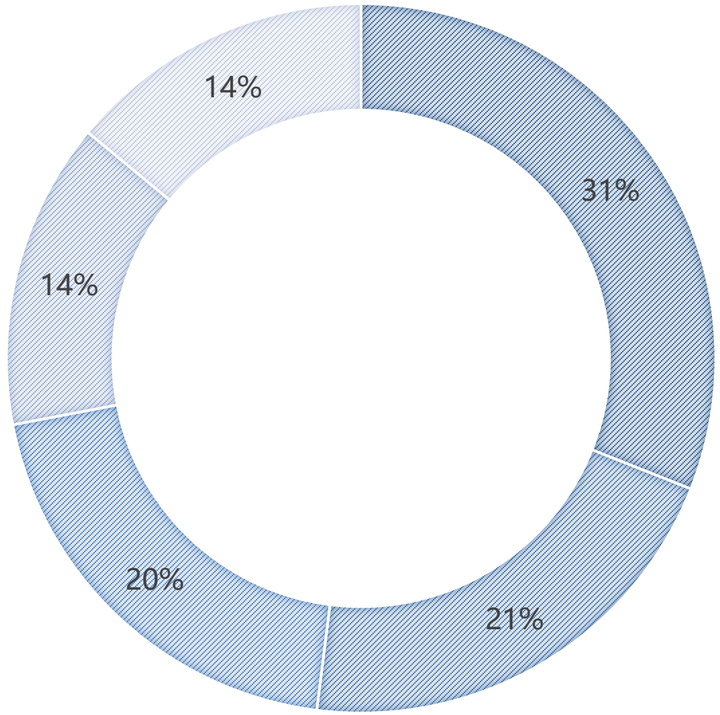
Li, C. (2020). Causer and causee as two higher-ranked thematic roles. *Language and Linguistics, 21*(4), 601–635.

02

结果补语知识的构建与应用

结果补语的事件解读与其句式选择之间存在紧密的对应关系

<述补结构句式分布>



把字句

致使义

我把所有的洞都堵好了

SVCO

仅有39个补语这样使用，呈现出明显的动词偏向

凶猛的洪水冲倒了桥墩 ［结果义宾语］

今年暑假要放到九月十日 ［介词补语］

SVC

述补组合更为多样，变化义

那个孩子近半年来学坏了

这件衣服的胳膊都磨白了

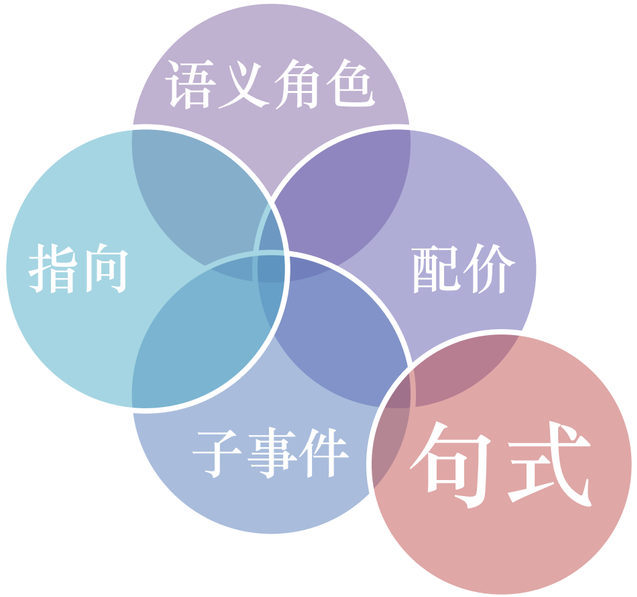
被字句

话题句，动词重复句等

Oh, Y.j. (2023). The sentence distribution and semantic structure of VRCs in modern Chinese: Based on a quantitative analysis of A Dictionary of Chinese Verb–Resultative Complement Phrases. *Journal of Chinese Language and Literature*, 112, 53–77.

02

结果补语知识的构建与应用



为辅助结果补语句的生成问题，有必要更加关注
各类述补结构最自然的句式实现方式

述补结构的句式分布并非随机，而是受到补语本身功能以及整体事件的性质与语义的制约，因此会出现更为契合或不太契合的句式。
这类信息是理解述补结构时不可忽视的重要知识。

然而，在大多数情况下，论元的排列顺序以及所构成的句式往往呈现为一种在自然度和可接受度上的连续差异，
因此难以以一个泾渭分明的方式来进行统一、硬性的分类。

02

结果补语知识的构建与应用

1. 虚化补语：使用判断困难

- 虚化补语句难以明显分为两个独立的子事件，补语本身也并未单一子事件的构成成分，而是高度语法化的功能性成分
- 学习者难以判断虚化补语何时需要或不需要使用

→ 若大语言模型能准确判别虚化补语的使用必要性，可作为学习者的参考依据

2. 实义补语：句式实现高度多样

- 学习者即使理解“动作＋结果”的因果关系，在实际表达中仍因述补结构的论元组合方式与句式实现高度多样、自由度高而难以自然造句。
- 现有基于词汇意义的归档、语义指向或配价的类型化·结构化模式难以覆盖使用中的这些困难。

→ 若大语言模型能根据论元与述补结构给出最佳句式与论元排列，学习者可据此校正表达

03

LLM概率语言知识的应用思路

核心概念与相关研究梳理

03

LLM概率语言知识的应用思路

基本思路：基于LLM概率分布的句子自然度评估

大语言模型通过学习海量语言数据，以概率方式选择语言成分并生成句子，本质上是一种对真实语言分布的概率近似、拟合，从而反映着实际语言的使用情形。

$P(\text{token} \mid \text{context})$

概率 ↑ 与模型反映的语言分布更一致 → 句子更自然

概率 ↓ 偏离模型反映的语言分布 → 句子更不自然或显得别扭

→基于这一假设，我们可以利用模型计算得到的平均对数概率（mean log probability）来对句子的自然度进行量化评估。

这样既可以判断学习者生成句子的好坏，

还可以为学习者推荐更自然、更符合语言习惯的表达形式。

03

LLM概率语言知识的应用思路

Assessing Minimal Pairs of Chinese Verb-Resultative Complement Constructions: Insights from Language Models (CxGsNLP 2025)

判断任务：　结果补语有无对比（如：打碎 vs. *打）

动词—补语顺序倒置对比（如：搞错 vs. *错搞）

数据来源：结果补语最小对比对数据集ZhVrcMP

用于提供系统性的 GOOD/BAD 句对，构建了结果补语最小对比对数据集
共1,204对最小对比对
基于《现代汉语》词汇选取342个名词、53个动词、66个结果补语
由 Python 自动生成组合，并经两位语言学家人工校验

	Para 1	Para 2
Number	602	602
	张三摔破额头。	张三搞错观点。
GOOD	Zhāng Sān shuāi pò é tóu	Zhāng Sān gǎo cuò guāndiǎn
	Zhang San fell and broke his forehead.	Zhang San got the wrong point.
	* 张三摔额头。	* 张三错搞观点。
BAD	Zhāng Sān shuāi é tóu	Zhāng Sān cuò gǎo guāndiǎn
	Zhang San fell his forehead.	Zhang San wrong got the point.

实验指标：基于平均对数概率（mean log probability）的好坏判断

若模型在最小对比对中对GOOD句赋予更高概率，则可视为一次正确判断；

其总体正确率越高，模型的语言能力越强。

$$P_{ML} = \frac{\log P_m(\gamma)}{n_\gamma} \qquad S(p) = \frac{1}{|p|} \sum_{g,u \in p} \mathbf{1}_{[0,+\infty)} \left(\log \frac{P_{ML}(g)}{P_{ML}(u)}\right)$$

g = GOOD 句，u = BAD 句
对每一组句对 p(GOOD/ BAD)，计算一个得分 S(p)。
若 PML(g) / PML(u) > 1，表示模型认为GOOD句的概率更高，因此判断正确。
将所有句对中“判断正确”的次数求平均，作为模型捕捉语言能力的表现指标。

03

LLM概率语言知识的应用思路

Assessing Minimal Pairs of Chinese Verb-Resultative Complement Constructions: Insights from Language Models (CxGsNLP 2025)

实验结果：

使用的语言模型

- Zh-Pythia 系列（14M - 1.4B）
- Mistral-7B-Instruct

	Zh-Pythia					Mist	Human
	14M	70M	160M	410M	1.4B	7B	
Para 1	68.77	80.90	83.06	82.89	85.88	63.46	98.00
Para 2	86.57	92.54	95.19	92.87	93.86	83.58	98.00
Overall	77.67	86.72	89.13	87.88	89.87	73.52	98.00

Table 5: Percentage Score of all LMs and human on ZhVrcMP

主要发现

1. 模型在顺序问题上的判断能力优于在有无问题上的判断能力

   有无问题：涉及词汇可匹配性（lexical compatibility），模型较难处理

   顺序问题：主要依赖局部注意力规律（local attention patterns），模型更容易捕捉

2. 与人类表现仍有显著差距

   人类正确率约 98%

   模型整体表现明显低于人类

→ 平均对数概率在句子好坏判定中具有较强的区分能力
然而，不同仍存在模型难以充分捕捉任务类型。

→ 由于本研究的数据均为人工构造的句对，
后续可以通过真实学习者错误进行进一步验证，
并可在更多类型的学习者错误上扩展实验。

03

LLM概率语言知识的应用思路

Demystifying large language models in second language development research(Cong, Y., Computer Speech & Language, 2025)

探索如何利用大语言模型评估以中文为母语的英语学习者写作能力

评估指标： Surprisal = 在给定前文的情况下，下一个词出现的“惊讶度” $Surprisal(w_t) = -logP(w_t|w_{1...t-1})$

反映句法准确度（syntactic grammaticality）+ 语义合理性（semantic plausibility）的综合指标

实验1：L1 vs L2 作文的区分能力
多种模型均能有效区分 L1/L2 写作，GPT-Neo、BERT 表现突出。

实验2：L2熟练度（Level 3–5）区分
传统评估指标仍然表现强劲。
T5在熟练度区分上表现较强。
BERT、DistilGPT2也能区分部分等级。

实验3：人类评分预测
仅用Surprisal也能预测人类评分。
与传统指标相比，其预测性能的提升有限 → Surprisal 是“补充”指标，而非完全替代。

03

LLM概率语言知识的应用思路

Demystifying large language models in second language development research(Cong, Y., Computer Speech & Language, 2025)

探索如何利用大语言模型评估以中文为母语的英语学习者写作能力

Surprisal的特征及优势

单一指标捕捉多维语言特征
能反映上下文连贯性
在段落级长文本中优势更明显

Surprisal数值解释上的困难

不同模型体现不同的Surprisal特征

T5：对L1文本给出更高Surprisal
GPT系列：对L2文本给出更高Surprisal
BERT：对语法性与语义性均敏感

→尽管本人的研究以语言模型对语言分布的“概率拟合度”来判定句子好坏为基本前提，
但在实际的数值解释中**仍需注意这一指标包含多种影响因素**

- 句子自然度与概率并非始终呈正相关
- Surprisal同时涉及词汇与句法两个层面，因此在变量控制方面也需格外谨慎。

随着文本熟练度等级的变化，Surprisal的解释方式也随之不同：

在低熟练度文本中，Surprisal往往可作为语法错误的指标，因而Surprisal越低，句子越可能是正确的。
在高熟练度文本中，Surprisal反而可能反映句子是否过于单调，
因而Surprisal越高，越可能被视为其表达能力更强。

04

实验设计与结果分析

基于概率的句子自然度评估

实验设计

假设：

句子概率越高，越符合模型的语言分布 → 被判定为更自然的句子。

实验指标：

(1)自回归语言模型（Autoregressive/Causal LM）

SJTU-CL/Zh-Pythia-160M，uer/gpt2-chinese-cluecorpussmall

模型对输入句子返回 cross-entropy loss，以此计算平均对数概率

(2) BERT类（Masked LM）

hfl/chinese-roberta-wwm-ext，hfl/chinese-macbert-base，hfl/chinese-bert-wwm-ext

由于BERT无法直接获得基于loss的取得概率值

因此将句子中的每个token逐一替换为[MASK]，计算模型对该token的预测对数概率

最后对所有概率取平均，作为该句子的平均对数概率

实验1 学习者错误识别能力评估

基于概率值判定学习者错误句与准确修正句

数据集1：数据来源于两篇关于韩国人汉语结果补语学习错误的研究论文

数据特征：中文学习时间较短（约一年），水平相当于 HSK 4 级的中级学习者

通过短文写作问卷收集的数据，句子整体较为简短、结构相对简单

根据论文中的错误实例构建88对正误句

数据处理：正确句主要采用原论文提供的修改版本，
若单句无法独立判断正误，则补充简短语境以确保可判别性。

那位医生的医术很好，一定会救他。
那位医生的医术很好，一定会救活他。

<错误类型—题目数量明细>

残缺			多余	替代	语序颠倒	分离・回避	其他
49			15	11	6	3	4
虚化	实义	动词					
35	7	7					

Wang, H.S. (2017). *Teaching resultative complements to Korean learners of Mandarin Chinese: An error analysis* (Master's thesis). Hanbat National University.

Yang, C.C. (2019). *A study on the learning method of Chinese complement: Focusing on the error analysis of Korean learners* (Doctoral dissertation). Dong-A University.

实验1

数据集2：HSK作文补语错误

语料来源：HSK 动态作文语料库 3.0（hsk.blcu.edu.cn）

数据特征：母语非汉语的国际学生参加高等汉语水平考试（HSK高等）作文试题答卷语料库，
库中收集了1992-2005年的部分考生的作文答卷
该语料库按错误类型提供错误句 + 修改方案，其中与结果补语相关的两类错误为：
补语残缺：共 729 例，其中结果补语 388 列
补语多余：共 470 例，其中结果补语 150 例

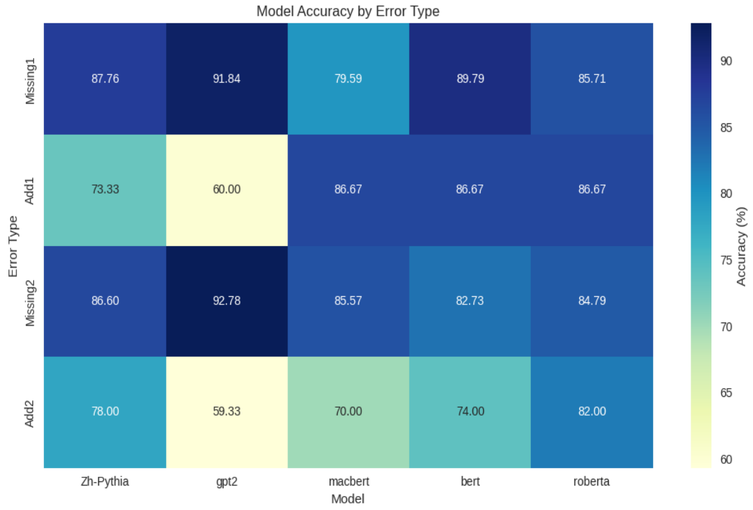
数据处理：
为判断任务构建正误句对
单个学习者句子可能包含多种错误。
正确句：采用语料库提供的修改方案。
错误句：仅保留与结果补语相关的补语缺失/补语多余错误，
按修改方案修正其他无关错误。
若原句中结果补语错误包含在更大范围的短句错误中，
导致无法构成可判断的正误句对，则删除该错例。

04 学习者错误识别能力评估

实验1

〈结果分析：残缺 • 多余错误〉

错误类型	题目数	Zh-Pythia160M	gpt2	macbert	bert	roberta
[残缺]	49	87.76	91.84	79.59	89.79	85.71
[多余]	15	73.33	60	86.67	86.67	86.67
[残缺]	388	86.6	92.78	85.57	82.73	84.79
[多余]	150	78	59.33	70	74	82



实验结果综述

- 整体来看，各模型之间虽存在一定差异，但整体准确率大多集中在80%出头的区间。
- 在错误类型上，模型普遍表现出对多余错误的识别能力弱于对残缺错误的识别能力。

模型之间的表现特点区别

- 在句子更长、结构更复杂、学习者熟练度更高的语料（数据集2）中，BERT系列在多余错误上的性能下降更为明显。
- 与此相对，自回归模型在数据集1与数据集2上的表现相对稳定。

说明：

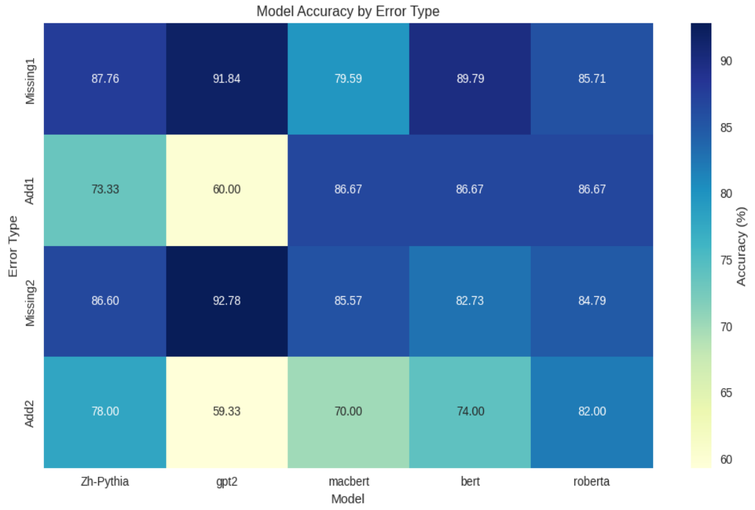
- BERT不仅关注局部语法或特定用例的偏误，更容易受到整体文本特征的影响（对句法、词汇复杂度等因素更为敏感）

04 学习者错误识别能力评估

实验1

〈结果分析：残缺 • 多余错误〉

错误类型	题目数	Zh-Pythia160M	gpt2	macbert	bert	roberta
[残缺]	49	87.76	91.84	79.59	89.79	85.71
[多余]	15	73.33	60	86.67	86.67	86.67
[残缺]	388	86.6	92.78	85.57	82.73	84.79
[多余]	150	78	59.33	70	74	82



残缺错误与多余错误判别能力之间的权衡关系

是否残缺错误识别准确率越高，多余错误识别准确率越低？

相关性分析结果

- 皮尔逊相关系数（Pearson r）：-0.6313
- p值：0.0503

- 两类错误之间呈现 中等程度的负相关：
  → 模型越擅长识别残缺错误，其在多余错误上的表现越可能较弱。
- p 值位于统计显著性的边界：
  → 说明这种关系不强但具有一定解释意义，并非完全偶然。

模型在残缺与多余错误判别能力之间存在一定程度的权衡

学习者错误识别能力评估

〈结果分析：残缺 • 多余错误——个别案例分析〉

[残缺]

BAD: 警察把小偷抓了。
GOOD: 警察把小偷抓住了。

3类均BERT对错误句给出更高概率

BAD: 现在整个世界上很多人还是不能吃东西，他们会饿死的。
GOOD: 现在整个世界上很多人还是不能吃饱东西，他们会饿死的。

所有模型对错误句给出更高概率

BAD: 第二，我们把污染的地方恢復原来的样子。
GOOD: 第二，我们把污染的地方恢復成原来的样子。

GPT2, macbert, roberta对错误句给出更高概率

[多余]

BAD: 有的年轻人说：“会吸烟的男人才是真正的男子汉，
男人吸烟才有魅力。”这些话导致了许多年轻人都想学会吸烟了。
GOOD: 有的年轻人说：“会吸烟的男人才是真正的男子汉，
男人吸烟才有魅力。”这些话导致了许多年轻人都想学吸烟了。

4类模型对错误句给出更高概率

BAD: 回到印尼，当然是满载而归，我买了一大堆我心爱的书刊，
有空闲时间就一一品读。
GOOD: 回印尼，当然是满载而归，我买了一大堆我心爱的书刊，
有空闲时间就一一品读。

4类模型对错误句给出更高概率

BAD: 如果我知道病人患的病已经治好不了了，我觉得让他安乐死比较好。
GOOD: 如果我知道病人患的病已经治不了了，我觉得让他安乐死比较好。

所有模型对错误句给出更高概率

即便是表现相对稳定的模型，在某些人类能够清晰判断正误的情境中，也会给出错误的判断。

多余错误的判断：

- 不仅需要识别动补结构在局部层面上是否自然、可行，
- 还必须结合整体语境来判断叙述的事件是否应当带有结果，还是仅仅表达动作本身。

（如话题转换标记中的动词（我想vs我想到）、条件句等）

换言之，模型既要判断局部结构本身的合理性，
又要判断句法 • 语义上是否需要结果成分，这使得任务难度增加。
因此模型在多余错误上的错误率普遍偏高。

学习者错误识别能力评估

〈结果分析：其他错误类型〉

错误类型	题目数	Zh-Pythia160M	gpt2	macbert	bert	roberta
[替代]	11	0	1	1	0	0
[语序颠倒]	6	1	0	0	0	0
[分离·回避]	3	0	1	1	1	0
[其他]	4	0	0	0	0	0

[替代]

在替代类型上也呈现出相对稳定的趋势，
但仍有必要在扩大数据量后进一步观察。

[语序颠倒]

关于这一类型，与前人研究的指摘一致，模型的表现较为稳定。

[分离・回避]

BAD: 我们打扫房间，房间干净了。
GOOD: 我们打扫干净了房间。

可以发现，模型会认为比起更短、且没有错误的句子，
一个结构稍微复杂一些的句子反而更有可能是正确的。
这一点颇为意外，后续值得在扩大样本后继续检验。

[其他]

BAD: 她把那个礼物气扔了。
GOOD: 她生气地扔掉了礼物。

BAD: 我不能记电话号码，实在想不起来。
GOOD: 我没记住电话号码，实在想不起来。

BAD: 这个人说得太快了，我不能听。
GOOD: 这个人说得太快了，我没听懂。

BAD: 昨天的聚会，他醉到了。
GOOD: 昨天的聚会，他喝醉了。

- 在这些相对混乱、归类较难的“其他类型”错误中，
模型似乎也具备一定的判断能力。
- 推测可能是因为模型能够较好地过滤掉学习者随意拼出的、
不可能出现的组合。

实验 2　最佳句式推荐能力评估

评估模型在不同述补结构及其论元的多种组合方式中，是否能够作出最优的句式选择。

实验述补句生成：句式模板

述补句的主要构成要点如下：

(1) 句式类型

SVCO、SVC、VCO、话题句、把字句、被字句、致使句、动词重复句

(2) 论元的排列顺序

从理论上看，述补结构中可与动词和补语结合的论元最多可达 6 个（动词 3 个、补语 3 个）。然而，从述补句的核心语义结构来看，其关键论元实际上可以归纳为两类：

- 经历变化（BECOME）的实体　- 引发变化（CAUSE）的原因

在真实语句中：

- 两类论元可能只出现其中之一，也可能二者指呈现为同一实体
- 根据语义解读的不同，其句法位置可能发生互换。

实验2

实验述补句生成：句式模板

[X] [Y] <V> <C> ： 动词，补语，两个论元参与

[SVCO]
[X]<V><C>了[Y]。
[X]<V><C>[Y]了。
[Y]<V><C>了[X]。
[Y]<V><C>[X]了。

[SVC]
[X]<V><C>了。
[Y]<V><C>了。

[VCO]
<V><C>[X]了。
<V><C>[Y]了。

[把字句]
[X]把[Y]<V><C>了。
[Y]把[X]<V><C>了。

[被字句]
[Y]被[X]<V><C>了。
[X]被[Y]<V><C>了。
[Y]被<V><C>了。
[X]被<V><C>了。

[致使句]
[X]让[Y]<V><C>了。
[Y]让[X]<V><C>了。

[话题句]
[Y][X]<V><C>了。
[X][Y]<V><C>了。

[动词重复句]
[X]<V>[Y]<V><C>了。
[Y]<V>[X]<V><C>了。

X=这天气, Y=我, V=热, C=死

[SVCO]
- 这天气热死了我。
- 这天气热死我了。
- 我热死这天气了。
- 我热死了这天气。

[SVC]
- 这天气热死了。
- 我热死了。

[VCO]
- 热死我了。
- 热死这天气了。

[话题句]
- 我这天气热死了。
- 这天气我热死了。

[把字句]
- 这天气把我热死了。
- 我把这天气热死了。

[动词重复句]
- 这天气热我热死了。
- 我热这天气热死了。

[被字句]
- 我被这天气热死了。
- 这天气被我热死了。
- 我被热死了。
- 这天气被热死了。

[致使句]
- 这天气让我热死了。
- 我让这天气热死了。

实验2

实验述补句生成：述补词语的选定与分类整理

1. 述补词语选定原则

实验2的任务主要涉及论元与述补结构之间的作用力关系，因此本实验优先选择**实义补语**作为研究对象。

（虚化补语与事件的时体特征、叙述性等句法功能联系更为紧密，
不利于体现论元与述补结构之间的直接影响关系。）

2. 补语分类体系的简化与整理

既有分类为了能够更细致地解释动补结构的意义与事件结构，对意义角色进行了高度细化。
但仅靠这种细分难以凸显各类述补结构的核心功能与特质，并且有可能导致在实际应用中的稳定性与可操作性下降。
因此，本研究着重简化意义角色与补语指向的分类方式，并**力图以更直观的方式呈现各类述补结构的本质事件框架**。

(1) 论元角色的简化

在既有分类的基础上，将论元角色简化为：主体、客体、其他（处所，身体部位）

(2) 补语指向性的划分

根据补语语义指向的不同，将其划分为： 指向主体、指向客体、指向其他成分、指向动词本身

补语分类体系的简化与整理

1. 主体指向

1.1 主体指向 + 不及物动词 → 主体发生内部变化

争点：即使配价论元数已满，但由于述补结构的“原因—结果”特性，原因常以配价外的论元形式出现。

主体的主动性与受影响程度不同，会导致论元位置与呈现句式出现规律性变化。

1.1.1 无外部原因的自然发生 （如：雨下大了、疯刮猛了、花开败了、睡醒）

1.1.2 可假设外部原因（如：呆胖、急哭、哭醒）

1.1.3 可假设外部原因 + 主体受影响性强 （如：热死、吓蒙）

1.2 主体指向 + 及物动词

争点：宾语的性质（是否属于原因、虚化程度、具体性等）会影响句式选择。

1.2.1 无及物对象更自然 → 主要表达主体变化，没有明显外部原因（如：学坏，磨百）

1.2.2 存在明确的及物对象

1.2.2.1 包含虚化客体或离合词客体 → 客体对主体的影响较弱（如：吃饱饭、喝醉酒；走动路、睡醒觉）

1.2.2.2 包含具体的客体 → 主体对客体的行动造成原因（如：洗累、吃胖、看傻、羡慕死）

1.3 主体指向 + 三价动词 → 两种客体多有可能成为原因（如：教累、卖晕）

争点：最终可根据“原因—受影响者”关系简化为二价结构

补语分类体系的简化与整理

2. 客体指向 → 由两个界限明确的子事件构成

争点：典型与非典型因果意义（因果距离）、评价意义

2.1 具有因果事件结构

2.1.1 典型因果意义→主体明确改变客体（如：踢破、打碎）

2.1.2 非典型因果意义→因果距离极远，主体对客体的影响极为间接（如：哭湿）

2.2 子事件仅呈先后，不含因果

→“因果结构”旨在评价，并非存在致使-变化关系（如：买大、挖浅）

3. 其他对象指向 → 主体对客体实施动作的过程中，第三对象受到影响

争点：指向对象特殊、句式特殊且使用受限。

3.1 处所指向（如：塞满、填满）

3.2 身体部位指向（如：看傻眼、腿坐麻）

4. 动词指向

4.1 动作完成后，动词的所有论元均受到影响

争点：多个受影响对象中，最受影响者的主题化；多论元的简化与二元化。

4.1.1 主体 + 客体（如：学会，说明清楚）

4.1.2 客体 + 其他（如：教会，教好）

4.2 动词本身的时间、距离、频率等特性（如：早晚、远近、惯）

争点： 论元的可省略性

04 最佳句式选择能力评估

实验2

评估述补结构选定：

- 选取“主题指向型”补语作为试点对象，
  因为此类补语最能体现**配价外论元的出现**，以及因**主体受影响程度和客体语义特征所引发的论元位置变动**等
  述补句生成中的关键难点。
- 共对28组述补—论元组合进行了评价。

评估方式：

- 评估模型的精确率模型所推荐的句子中，有多少是真正自然的句子
- 以排名前3的述补句中是否出现“可接受句”作为判断标准

本阶段以判断句子是否“说得通”为主要标准；至于母语者所感知的最优句式，
今后可能需要通过人工测试等方式进一步加以补充。

模型	Zh-Pythia160M		gpt2		macbert		bert		roberta	
TP/FP	38	34	49	23	50	22	52	20	53	19
精确率	0.528		0.681		0.694		0.722		0.736	

评估结果：

- 各模型的精确率仅为 0.52～0.73，说明模型推荐的句子中仍有 30～50% 属于不自然或错误句。
  本实验中，每道题仅取Top 3候选句进行评估，已经是对模型较有利的条件，
  而模型在该任务上的表现明显偏低，难以可靠地完成句子自然度判断。
- 这也反映出**模型的概率信息不足以有效体现论元的语义性质（如原因提供方 / 受影响者）及其对组合方式的影响**。

最佳句式选择能力评估

评估结果分析：

- 相比于前述“简单的残缺 / 多余错误”，模型在判断结果补语整体句型的自然度方面能力明显下降。
- 虽然排名靠前的句子多为可接受句，但在细节层面，不少可接受句却被排在病句之后，
  其主要原因可能在于：论元的正确排列依赖于对“谁是影响提供者 / 谁是受影响者”的常识性理解，
  而模型的概率信息往往并不能反映这种论元之间的逻辑关系，而是倾向于选择分布上更常见的结构。

模型行为特征：

1. **模型间一致性较低**
   不同模型对同一组句子的排序差异明显，整体一致性不足。
2. **对同类述补结构的判断不稳定**
   即使结构相同，仅因词汇替换，句法自然度判断便出现较大波动
   当词汇复杂、补语不典型时，模型的判断结果更易发生变化
3. **明确宾语存在时，模型明显偏好“把字句”与“让字句（致使句）”**
   即便这些句式在语义上不可行，也常被排在其他可行的句式之前。
   模型更倾向于选择分布上更常见的结构，而非逻辑上更合理的结构。
4. **对动词重叠式与被动句的回避倾向**
   模型整体上对动词重叠式的选择率偏低；在遭遇义词语存在时，被字句的排序有所提升，但总体仍呈现轻微回避倾向。
   因此在动词重叠式或被动句应为最优解的情境中，错误率偏高。
5. **论元变长时，模型更倾向于选择话题化结构。**

1. 主体指向

1.1 主体指向 + 不及物动词 → 主体发生内部变化

1.1.1 无外部原因的自然发生 （如：雨下大了、疯刮猛了、花开败了、睡醒）

1.1.2 可假设外部原因（如：呆胖、急哭、哭醒）

1.1.3 可假设外部原因 + 主体受影响性强 （如：热死、吓蒙）

仅有单一参与者的“自然发生义”句子

Zh-Pythia160M	gpt2	macbert	bert	roberta
X=雨, Y=∅, V=下, C=大				
下大了。	下大雨了。	下大雨了。	下大雨了。	下大雨了。
雨下大了。	下雨下大了。	雨下大了。	雨下大了。	雨下大了。
让雨下大了。	雨下大了。	把雨下大了。	让雨下大了。	下大了雨。
X=风, Y=∅, V=刮, C=猛				
刮风刮猛了。	刮风刮猛了。	风被刮猛了。	把风刮猛了。	风被刮猛了。
刮猛风了。	风刮刮猛了。	把风刮猛了。	风被刮猛了。	刮猛了风。
刮猛了风。	被风刮猛了。	被风刮猛了。	被风刮猛了。	把风刮猛了。
被风刮猛了。	风刮猛了。	刮猛了风。	风刮猛了。	刮猛风了。
把风刮猛了。	把风刮猛了。	刮猛风了。	让风刮猛了。	被风刮猛了。
风刮猛了。	风被刮猛了。	风刮猛了。	刮猛风了。	风刮猛了。
让风刮猛了。	让风刮猛了。	让风刮猛了。	刮风刮猛了。	让风刮猛了。
刮猛了。	风把刮猛了。	刮风刮猛了。	刮猛了风。	刮风刮猛了。
风把刮猛了。	刮猛风了。	风把刮猛了。	风把刮猛了。	风把刮猛了。
风被刮猛了。	被刮猛了。	风刮刮猛了。	风让刮猛了。	风让刮猛了。
风刮刮猛了。	刮猛了风。	被刮猛了。	风刮刮猛了。	风刮刮猛了。
风让刮猛了。	刮猛了。	风让刮猛了。	被刮猛了。	被刮猛了。
被刮猛了。	风让刮猛了。	刮猛了。	刮猛了。	刮猛了。

Zh-Pythia160M	gpt2	macbert	bert	roberta
X=孩子, Y=∅, V=睡, C=醒				
孩子睡醒了。	让孩子睡醒了。	把孩子睡醒了。	孩子睡醒了。	把孩子睡醒了。
睡醒了。	把孩子睡醒了。	睡醒了。	让孩子睡醒了。	孩子睡醒了。
把孩子睡醒了。	孩子睡醒了。	让孩子睡醒了。	睡醒了孩子。	睡醒了。
孩子睡睡醒了。	孩子把睡醒了。	被孩子睡醒了。	孩子睡睡醒了。	被孩子睡醒了。
孩子被睡醒了。	孩子让睡醒了。	孩子睡醒了。	睡醒了。	孩子被睡醒了。
X=紫罗兰, Y=∅, V=开, C=败				
紫罗兰开败了。	把紫罗兰开败了。	被开败了。	把紫罗兰开败了。	开败了。
紫罗兰开开败了。	让紫罗兰开败了。	开败了。	让紫罗兰开败了。	被开败了。
让紫罗兰开败了。	紫罗兰开开败了。	被紫罗兰开败了。	紫罗兰被开败了。	把紫罗兰开败了。
紫罗兰被开败了。	紫罗兰开败了。	把紫罗兰开败了。	被紫罗兰开败了。	被紫罗兰开败了。
把紫罗兰开败了。	被紫罗兰开败了。	让紫罗兰开败了。	紫罗兰开开败了。	让紫罗兰开败了。
开败了。	紫罗兰让开败了。	紫罗兰被开败了。	紫罗兰开败了。	开紫罗兰开败了。
开败了紫罗兰。	紫罗兰把开败了。	紫罗兰开败了。	被开败了。	紫罗兰被开败了。
开紫罗兰开败了。	开紫罗兰开败了。	紫罗兰把开败了。	紫罗兰把开败了。	紫罗兰开败了。

- 在“自然发生义”语境中，SVC结构应为更常见的表达方式，但模型却倾向于将“把字句”“让字句”排在更高位置。
- 即便是论元结构与事件特征完全相同的同类结果补语结构，模型的判断仍会因词汇差异而出现明显的不稳定性
- 当词汇义较为生疏或不常见时，这种不稳定性尤为突出，整体表现随之下降。

1. 主体指向
   1.1 主体指向 + 不及物动词 → 主体发生内部变化
      1.1.1 无外部原因的自然发生 （如：雨下大了、疯刮猛了、花开败了、睡醒）
      1.1.2 可假设外部原因（如：呆胖、急哭、哭醒）
      1.1.3 可假设外部原因 + 主体受影响性强 （如：热死、吓蒙）

具有一定由原因到主体的“弱致使义”句子

Zh-Pythia160M	gpt2	macbert	bert	roberta
		X=妈妈, Y=孩子生病, V=急, C=哭		
妈妈急哭了。	孩子生病把妈妈急哭了。	孩子生病让妈妈急哭了。	妈妈急哭孩子生病了。	妈妈急哭孩子生病了。
妈妈被急哭了。	孩子生病让妈妈急哭了。	孩子生病妈妈急哭了。	孩子生病妈妈急哭了。	妈妈孩子生病急哭了。
孩子生病让妈妈急哭了。	妈妈把孩子生病急哭了。	妈妈急哭孩子生病了。	孩子生病把妈妈急哭了。	孩子生病妈妈急哭了。
		X=我, Y=这三个月在家, V=呆, C=胖		
我这三个月在家呆胖了。	我这三个月在家呆胖了。	我这三个月在家呆胖了。	这三个月在家把我呆胖了。	我这三个月在家呆胖了。
我把这三个月在家呆胖了。	我让这三个月在家呆胖了。	这三个月在家把我呆胖了。	我这三个月在家呆胖了。	这三个月在家让我呆胖了。
我被这三个月在家呆胖了。	我把这三个月在家呆胖了。	这三个月在家呆胖了。	我把这三个月在家呆胖了。	这三个月在家把我呆胖了。
我让这三个月在家呆胖了。	这三个月在家让我呆胖了。	这三个月在家被呆胖了。	这三个月在家呆胖了我。	我把这三个月在家呆胖了。
我呆胖了。	这三个月在家被我呆胖了。	这三个月在家让我呆胖了。	这三个月在家被我呆胖了。	这三个月在家呆胖了。
我被呆胖了。	这三个月在家呆胖了。	这三个月在家被我呆胖了。	这三个月在家让我呆胖了。	这三个月在家呆胖了我。
这三个月在家呆我呆胖了。	这三个月在家呆我呆胖了。	这三个月在家呆胖了我。	这三个月在家我呆胖了。	这三个月在家我呆胖了。
这三个月在家被我呆胖了。	这三个月在家呆胖了我。	我把这三个月在家呆胖了。	这三个月在家呆胖了。	这三个月在家被我呆胖了。
这三个月在家呆胖了。	这三个月在家把我呆胖了。	这三个月在家我呆胖了。	这三个月在家被呆胖了。	这三个月在家被呆胖了。
这三个月在家把我呆胖了。	这三个月在家我呆胖了。	这三个月在家呆我呆胖了。	我被这三个月在家呆胖了。	这三个月在家呆胖我了。
呆胖我了。	我呆这三个月在家呆胖了。	这三个月在家呆胖我了。	这三个月在家呆胖我了。	这三个月在家呆我呆胖了。
这三个月在家我呆胖了。	我呆胖了这三个月在家。	我被这三个月在家呆胖了。	我呆胖了这三个月在家。	我被这三个月在家呆胖了。
我呆这三个月在家呆胖了。	我被这三个月在家呆胖了。	我呆胖这三个月在家了。	我让这三个月在家呆胖了。	我呆这三个月在家呆胖了。

- 由于此类结构具有致使意义，使用能够明确呈现致使关系的“把字句”“让字句（致使句）”是合理的，模型在这一点上也表现出较好的捕捉能力。
- 当原因论元较长或结构复杂时，模型也能较好地捕捉到其话题化倾向
- 然而，在论元位置的配置上仍频繁出现错误，甚至常将逻辑上不可能的结构排在可行结构之前。

Zh-Pythia160M	gpt2	macbert	bert	roberta
X=我, Y=这里的夏天, V=热, C=死				
热死我了。	这里的夏天热死了。	这里的夏天让我热死了。	这里的夏天热死了。	这里的夏天热死我了。
这里的夏天热死我了。	这里的夏天热死我了。	这里的夏天热死我了。	这里的夏天把我热死了。	这里的夏天热死了。
我被热死了。	我这里的夏天热死了。	这里的夏天把我热死了。	这里的夏天热死我了。	这里的夏天让我热死了。
这里的夏天被我热死了。	这里的夏天被热死了。	这里的夏天热死了。	这里的夏天让我热死了。	这里的夏天把我热死了。
这里的夏天被热死了。	这里的夏天热死了我。	我被这里的夏天热死了。	我被这里的夏天热死了。	我把这里的夏天热死了。
X=这故事, Y=我, V=吓, C=死				
吓死我了。	我被这故事吓死了。	吓死我了。	吓死我了。	吓死我了。
这故事吓死我了。	这故事把我吓死了。	我被吓死了。	这故事吓死我了。	这故事吓死我了。
我被吓死了。	吓死我了。	这故事吓死我了。	这故事把我吓死了。	这故事把我吓死了。
我被这故事吓死了。	我被吓死了。	这故事把我吓死了。	我被这故事吓死了。	这故事吓死了我。
我吓死了。	这故事吓死我了。	我吓死了。	我被吓死了。	我被吓死了。
我把这故事吓死了。	这故事让我吓死了。	这故事吓死了我。	这故事吓死了我。	我吓死了。
这故事把我吓死了。	这故事被我吓死了。	我被这故事吓死了。	我吓死了。	我被这故事吓死了。
X=我, Y=孩子胃胀、呕吐, V=担心, C=死				
担心死我了。	我担心死了。	孩子胃胀、呕吐让我担心死了。	我担心孩子胃胀、呕吐担心死了。	我担心孩子胃胀、呕吐担心死了。
孩子胃胀、呕吐我担心死了。	孩子胃胀、呕吐让我担心死了。	我担心死了孩子胃胀、呕吐。	我把孩子胃胀、呕吐担心死了。	孩子胃胀、呕吐让我担心死了。
孩子胃胀、呕吐让我担心死了。	孩子胃胀、呕吐把我担心死了。	我担心死了。	我担心死了。	孩子胃胀、呕吐我担心死了。
X=这份工作, Y=我, V=累, C=死				
这份工作累死我了。	这份工作累死我了。	这份工作让我累死了。	这份工作累死我了。	这份工作累死我了。
累死我了。	这份工作累死了。	我被这份工作累死了。	这份工作把我累死了。	这份工作把我累死了。
这份工作把我累死了。	这份工作把我累死了。	这份工作把我累死了。	我这份工作累死了。	这份工作让我累死了。
这份工作累死了我。	这份工作让我累死了。	这份工作累死我了。	这份工作我累死了。	这份工作累死了我。
我被这份工作累死了。	我把这份工作累死了。	这份工作累死了我。	我被这份工作累死了。	我被这份工作累死了。
我累死这份工作了。	我被这份工作累死了。	我让这份工作累死了。	这份工作累死了我。	我这份工作累死了。
这份工作累死了。	我这份工作累死了。	累死我了。	这份工作让我累死了。	这份工作我累死了。
这份工作让我累死了。	这份工作我累死了。	我累死了这份工作。	我把这份工作累死了。	这份工作累死了。
这份工作被我累死了。	这份工作累死了我。	我累死这份工作了。	我让这份工作累死了。	我把这份工作累死了。
我把这份工作累死了。	这份工作被我累死了。	我把这份工作累死了。	累死这份工作了。	我让这份工作累死了。
这份工作被累死了。	我被累死了。	累死这份工作了。	我累死这份工作了。	累死这份工作了。
这份工作我累死了。	这份工作累我累死了。	我累死了。	这份工作累死了。	我累死了这份工作。

1. 主体指向
   - 1.1 主体指向 + 不及物动词 → 主体发生内部变化
     - 1.1.1 无外部原因的自然发生
     - 1.1.2 可假设外部原因（如：呆胖、急哭、哭醒）
     - 1.1.3 可假设外部原因 + 主体受影响性强（如：热死、吓蒙）

具有较强由原因到主体的“强致使义”句子

- 在致使义或遭遇义较强的句型中，模型能较好地呈现出**将受影响者论元置于结果补语之后**的倾向。
- 然而，当论元结构变得复杂或出现非常规配置时，模型的判别能力仍会出现明显问题。

1.2 主体指向 + 及物动词

1.2.1 无及物对象更自然 → 没有明显外部原因（如：学坏）

1.2.2 存在明确的及物对象

1.2.2.1 包含虚化客体或离合词客体 → 客体对主体的影响较弱（如：吃饱饭、喝醉酒；走动路、睡醒觉）

1.2.2.2 包含具体的客体 → 客体为原因而非受影响者（如：洗累、吃胖、羡慕死、看傻）

Zh-Pythia160M	gpt2	macbert	bert	roberta
X=孩子, Y=坏朋友, V=学, C=坏				
孩子被学坏了。	孩子学坏了。	孩子学坏了。	孩子把坏朋友学坏了。	孩子把坏朋友学坏了。
孩子学坏了。	孩子学坏朋友学坏了。	孩子被学坏了。	坏朋友让孩子学坏了。	孩子学坏了坏朋友。
学坏孩子了。	孩子被坏朋友学坏了。	坏朋友学坏了。	坏朋友把孩子学坏了。	坏朋友让孩子学坏了。
坏朋友被学坏了。	孩子被学坏了。	坏朋友被学坏了。	孩子学坏朋友学坏了。	坏朋友把孩子学坏了。
孩子被坏朋友学坏了。	坏朋友把孩子学坏了。	坏朋友让孩子学坏了。	孩子被坏朋友学坏了。	孩子被坏朋友学坏了。
坏朋友把孩子学坏了。	孩子让坏朋友学坏了。	坏朋友把孩子学坏了。	坏朋友学坏了孩子。	孩子让坏朋友学坏了。
X=我, Y=饭, V=吃, C=饱				
吃饱饭了。	我吃饭吃饱了。	我吃饱了。	我吃饱饭了。	我吃饱了。
我吃饱了饭。	我吃饱了。	我把饭吃饱了。	我吃饱了。	我吃饱饭了。
我把饭吃饱了。	饭让我吃饱了。	我吃饱饭了。	吃饱饭了。	我把饭吃饱了。
X=我, Y=这堂课, V=听, C=困				
我被这堂课听困了。	这堂课让我听困了。	这堂课把我听困了。	这堂课让我听困了。	这堂课让我听困了。
这堂课我听困了。	我听这堂课听困了。	我把这堂课听困了。	我听这堂课听困了。	我把这堂课听困了。
我把这堂课听困了。	我把这堂课听困了。	这堂课让我听困了。	这堂课把我听困了。	这堂课把我听困了。
这堂课被我听困了。	这堂课把我听困了。	我被这堂课听困了。	我把这堂课听困了。	我听这堂课听困了。
我听这堂课听困了。	这堂课被我听困了。	这堂课被我听困了。	这堂课我听困了。	这堂课我听困了。
这堂课让我听困了。	这堂课我听困了。	我听这堂课听困了。	我听困这堂课了。	我被这堂课听困了。

1.2.1 动词虽为及物，但多用于无明确宾语的动补结构

在此类结构中，模型往往将不存在原因论元的句式排在更高位置，显示出其对词类动补结构事件特征具有一定捕捉能力。

1.2.2.1 客体虚化程度较高的情况

对于虚化程度较高的客体，

模型能够反映出客体不易出现在“让 / 把”之前的倾向。

1.2.2.2 客体兼具“行为对象”与“原因”双重角色的情况

词类结构中，因宾语既可以理解为施事的动作对象，也可以理解为原因，理论上可出现多种句式。

虽然模型在排序上缺乏一致性，但仍能对明显不可能的组合给予低分，

例如：“这堂课被听困了”（错误选择受影响者）

“这堂课听困了我”（SVCO 类型）

Zh-Pythia160M	gpt2	macbert	bert	roberta
X=他, Y=红薯土豆, V=吃, C=胖				
红薯土豆被他吃胖了。	他把红薯土豆吃胖了。	他把红薯土豆吃胖了。	他吃红薯土豆吃胖了。	他把红薯土豆吃胖了。
红薯土豆被吃胖了。	他吃红薯土豆吃胖了。	他被红薯土豆吃胖了。	他把红薯土豆吃胖了。	红薯土豆把他吃胖了。
他把红薯土豆吃胖了。	红薯土豆吃胖了。	红薯土豆吃胖了。	红薯土豆把他吃胖了。	他吃胖了红薯土豆。
红薯土豆吃胖了。	红薯土豆把他吃胖了。	他被吃胖了。	他被红薯土豆吃胖了。	红薯土豆吃胖了。
他吃胖了。	红薯土豆被他吃胖了。	他吃胖了。	红薯土豆被他吃胖了。	红薯土豆让他吃胖了。
他被吃胖了。	红薯土豆让他吃胖了。	红薯土豆把他吃胖了。	红薯土豆被吃胖了。	红薯土豆被他吃胖了。
红薯土豆他吃胖了。	红薯土豆被吃胖了。	红薯土豆被吃胖了。	红薯土豆吃胖了。	他吃红薯土豆吃胖了。
他吃红薯土豆吃胖了。	红薯土豆吃他吃胖了。	红薯土豆被他吃胖了。	红薯土豆让他吃胖了。	红薯土豆被吃胖了。
红薯土豆让他吃胖了。	他让红薯土豆吃胖了。	红薯土豆让他吃胖了。	他吃胖了红薯土豆。	他被红薯土豆吃胖了。
红薯土豆吃他吃胖了。	他吃胖了红薯土豆。	他吃胖了红薯土豆。	红薯土豆吃胖了他。	红薯土豆他吃胖了。
X=他, Y=衣服, V=洗, C=累				
衣服被洗累了。	他洗衣服洗累了。	他把衣服洗累了。	他洗衣服洗累了。	他洗衣服洗累了。
衣服被他洗累了。	衣服洗累了。	他洗衣服洗累了。	他把衣服洗累了。	衣服被他洗累了。
他把衣服洗累了。	他把衣服洗累了。	衣服被他洗累了。	他洗累了衣服。	衣服洗累了。
他洗衣服洗累了。	衣服被洗累了。	他被衣服洗累了。	衣服洗累了。	他把衣服洗累了。
衣服洗累了。	衣服被他洗累了。	衣服洗累了。	衣服被他洗累了。	衣服被洗累了。
衣服洗他洗累了。	衣服把他洗累了。	他衣服洗累了。	他洗累衣服了。	衣服让他洗累了。
衣服把他洗累了。	衣服洗他洗累了。	衣服被洗累了。	衣服让他洗累了。	衣服洗他洗累了。
衣服他洗累了。	衣服让他洗累了。	衣服让他洗累了。	衣服被洗累了。	他衣服洗累了。
他洗累了。	他让衣服洗累了。	他洗累了衣服。	他衣服洗累了。	他洗累了衣服。
他被洗累了。	他被衣服洗累了。	他让衣服洗累了。	衣服把他洗累了。	衣服把他洗累了。

1.2.2.2“主体对客体的行为本身”构成影响来源

动词重叠式往往是最自然的表达方式。

在较为典型的情境（如“洗累”）中，模型的选择相对合理；

但在出现“吃红薯土豆”这类较为特殊的论元时，判断错误更加明显。

整体来看，模型仍呈现出明显的“把字句偏向”。

1.3 主体指向 + 三价动词 → **两种客体多有可能成为原因**（如：教累、卖晕）

Zh-Pythia160M	gpt2	macbert	bert	roberta
X=我, Y=教这帮傻瓜数学, V=教, C=累				
教这帮傻瓜数学被我教累了。	我教这帮傻瓜数学教累了。	我被教累了。	我教这帮傻瓜数学教累了。	我被教这帮傻瓜数学教累了。
教这帮傻瓜数学教我教累了。	教这帮傻瓜数学让我教累了。	我被教这帮傻瓜数学教累了。	我被教这帮傻瓜数学教累了。	我把教这帮傻瓜数学教累了。
教这帮傻瓜数学把我教累了。	教这帮傻瓜数学把我教累了。	我把教这帮傻瓜数学教累了。	我把教这帮傻瓜数学教累了。	我被教累了。
教这帮傻瓜数学我教累了。	教这帮傻瓜数学被我教累了。	教这帮傻瓜数学把我教累了。	教这帮傻瓜数学教累了。	我教这帮傻瓜数学教累了。
教这帮傻瓜数学被教累了。	我教教这帮傻瓜数学教累了。	教这帮傻瓜数学让我教累了。	教这帮傻瓜数学把我教累了。	教这帮傻瓜数学把我教累了。
我把教这帮傻瓜数学教累了。	教这帮傻瓜数学教累了。	教这帮傻瓜数学教累我了。	我被教累了。	教这帮傻瓜数学让我教累了。
我被教这帮傻瓜数学教累了。	我把教这帮傻瓜数学教累了。	教这帮傻瓜数学被我教累了。	教这帮傻瓜数学被教累了。	教这帮傻瓜数学被我教累了。
X=训练师, Y=教小狗挑篮, V=教, C=累				
训练师教累了。	训练师教累了。	教小狗挑篮被训练师教累了。	训练师被教累了。	训练师被教累了。
训练师把教小狗挑篮教累了。	教小狗挑篮被训练师教累了。	训练师被教累了。	教小狗挑篮把训练师教累了。	教小狗挑篮被训练师教累了。
教小狗挑篮让训练师教累了。	训练师被教累了。	教小狗挑篮把训练师教累了。	教小狗挑篮被训练师教累了。	教小狗挑篮把训练师教累了。
教小狗挑篮被训练师教累了。	训练师教小狗挑篮教累了。	教小狗挑篮被教累了。	训练师把教小狗挑篮教累了。	训练师教累了。
教小狗挑篮把训练师教累了。	教小狗挑篮把训练师教累了。	教小狗挑篮让训练师教累了。	教小狗挑篮被教累了。	教小狗挑篮被教累了。
训练师教累了教小狗挑篮。	训练师教教小狗挑篮教累了。	训练师教累了。	教小狗挑篮让训练师教累了。	训练师教小狗挑篮教累了。

- 模型常并非依据“致使关系”来安排前置成分，
  而是优先选择具备较高有生性的论元，将其置于“把 / 被 / 让”字句之前。
  因此，对于应由**短句原因成分置于句首、以体现对主体影响关系的句类**，模型难以准确呈现。
- 此外，当论元结构较为非常规时，模型甚至会将省略受事论元的句式视为最优解。

研究总结

为学习者提供结果补语语言知识

学习者在掌握结果补语时的主要难点包括：

(1) 判断虚化补语的使用条件 (2) 区分实义补语可出现的句式类型

有必要从上述需求出发，构建相应的语言知识体系，或开发面向学习者的辅助工具。

基于语言模型概率信息的学习者错误判别

对添加・省略错误，模型表现出一定的稳定性。

- 但对于需考虑事件义・非事件义使用环境的省略错误，模型能力有限。
- 其他类型的学习者错误也具备一定判别力，但仍需更细致的验证。

基于语言模型概率信息的句式推荐

稳定性问题
- 模型间一致性较低，且单一模型内部对同类补语的判断也缺乏一致性。
- 当词汇较为生疏、论元不典型或结构复杂时，模型的判断结果波动显著。

概率信息的局限
- 模型倾向于将分布频率较高的句式置于更高位置，而不论其在具体语境中是否真正适切。
- 由于无法有效反映常识性的语义判断能力，尤其在论元配置方面，更易出现错误。

概率指标在语言能力评估中的局限与启示

使用surprisal等概率指标对学习者进行语言能力评估时需格外谨慎。

- 这类指标在判断文本整体复杂度、识别明显的语法错误或词汇误用方面确实具有优势，
- 但如本研究所示，仍存在概率信息难以捕捉的语言现象，

且模型偏向有时会使不自然甚至错误的句子获得更高评分。