基于汉语字词资源的检索增强生成与应用评估

殷雅琦[1,2]，刘扬[1,2]*，王悦[1,2]，梁启亮[1,3]
[1]北京大学多媒体信息处理全国重点实验室，北京 100871
[2]北京大学计算机学院，北京 100871
[3]北京大学信息科学技术学院，北京 100871
yyqi@stu.pku.edu.cn，{liuyang，wyy209}@pku.edu.cn，lql_eecs@qq.com

摘要

汉语遵循“由字组词，由词造句”的原则，字词相关信息是一类基础且关键的计算资源。在大语言模型时代，挖掘并评价该类资源的效用是增强模型语言能力的一个重要研究方面。作为有效促进资源与模型结合的一种方式，检索增强生成目前在该类资源上的应用大都关注模型未学习过的濒危语言，其在模型已学习过语言上的潜在价值有待挖掘。本文基于语言学的视角，构建具有良好例句覆盖率与丰富度的字词资源，并借助检索增强生成技术路线，探索这类资源与不同任务、模型的结合方法。评估实验表明，该方法在所有实验模型与任务中均带来了显著的准确率提升，平均达4.78%，其中，在语素义消歧、词义消歧与隐喻识别任务中分别提升了6.91%、4.24%和3.19%，这展示出字词资源对模型的语言准确理解能力的潜在价值。这些资源构造、方法探索和应用评估，为语言学资源与大语言模型的结合提供了新的思路与方法。

关键词： 汉语字词；资源建设；检索增强生成；应用评估

Chinese Character- and Word-Based Retrieval Augmented Generation and Application

Yaqi Yin[1,2], Yang Liu[1,2*], Yue Wang[1,2], Qiliang Liang[1,3]
[1]National Key Laboratory for Multimedia Information Processing,
Peking University, Beijing 100871
[2]School of Computer Science, Peking University, Beijing 100871
[3]School of Electronics Engineering and Computer Science,
Peking University, Beijing 100871
yyqi@stu.pku.edu.cn，{liuyang，wyy209}@pku.edu.cn，lql_eecs@qq.com

Abstract

The Chinese language's 'character, word, sentence' structure makes character and word-related information essential resources. In the era of Large Language Models (LLMs), exploring and evaluating these resources is crucial for enhancing the language capabilities of models. Retrieval Augmented Generation (RAG) is an effective means of integrating resources with models, currently focusing on languages not learned by the model, leaving untapped potential in learned languages. In this study, with a linguistic perspective, we construct a resource with improved contextual coverage and utilize RAG to combine knowledge with various tasks and models. Experimental results show accuracy improvements across all tested models and tasks with 4.78% on average.

*通讯作者
基金项目：国家自然基金项目（62036001）、国家社科基金项目（18ZDA295）

Specifically, it has achieved increases of 6.91%, 4.24% and 3.19% in morpheme sense disambiguation, word sense disambiguation, and metaphor recognition. The results show the potential of these resources for enhancing language understanding capabilities. These resource constructions, method explorations, and application evaluations offer new insights for integrating linguistic knowledge bases with LLMs.

Keywords: Chinese Character and Word , Resource Construction , Retrieval Augmented Generation , Application and Evaluation

1 引言

汉语中汉字是最小的独立表义单位，向上遵循“由字组词，由词造句”的原则(朱德熙, 1982)。这意味着字及其相关信息是认知词义的关键，而词及其相关信息又是理解句义的关键。因此，对汉语字词进行语义等信息标注的资源，作为文本内容理解的媒介与载体，已成为自然语言处理中不可或缺的基础和关键的计算资源(吉志薇and 冯敏萱, 2015; 刘扬et al., 2018)。利用这些资源，相关研究在词义消歧(Luo et al., 2018; Zheng et al., 2021b)、释义生成(Yang et al., 2020; Zheng et al., 2021a)、隐喻识别(Su et al., 2020)等任务中显著提高了模型的效果。

近年来，以ChatGPT[1]为代表的大型生成式语言模型(Touvron et al., 2023; Bai et al., 2023)由于其优越的多语言能力(Zhao et al., 2023; Zhang et al., 2024c)和任务表现(Zhong et al., 2023; Huang et al., 2023)，引起了行业的广泛关注。然而它们在对语言的准确理解方面仍存在不足(Martínez et al., 2023)，尤其是非英语语言(Robinson et al., 2023; Bang et al., 2023)，例如中文(Hendy et al., 2023; Ahuja et al., 2023; Liu et al., 2023)。因此，挖掘并评价字词资源在增强模型语言能力方面的效用并加以应用评估，是一个重要的研究方面。

在此研究方面，检索增强生成（Retrieval Augmented Generation, RAG）(Lewis et al., 2020)是一种被广泛研究与采纳的形式。它利用检索系统从外部资源中获取的相关信息，增强模型生成内容的质量与准确度。许多研究表明，相较于微调，这一方法不仅支持持续、低成本的知识更新和特定领域信息的集成(Strubell et al., 2019; Balaguer et al., 2024)，还能更有效、稳定地提高模型的准确度和可信度(Ovadia et al., 2024; Balaguer et al., 2024)，且可以应用于黑盒模型(Zhang et al., 2024b)，已逐渐成为字词资源在大语言模型中的重要应用方法。这些已有的基于字词资源的检索增强生成方法(Tanzer et al., 2023; Zhang et al., 2024a)利用字词典与语法手册，提高模型在机器翻译、对话理解等任务上的表现。目前，此类研究大都集中于模型未学习过的濒危语言，它们在模型已学习过的语言上的潜在价值尚有待挖掘。

我们注意到，对于非英语语言，即使是已针对该语言进行训练的模型，这类资源与方法也能为其提供有效的知识，优化其在下游任务中的表现。如图1所示，在面向不同语言单位的汉语消歧任务中，字词释义信息的注入帮助GPT-3.5模型[2]提供了正确的预测结果。

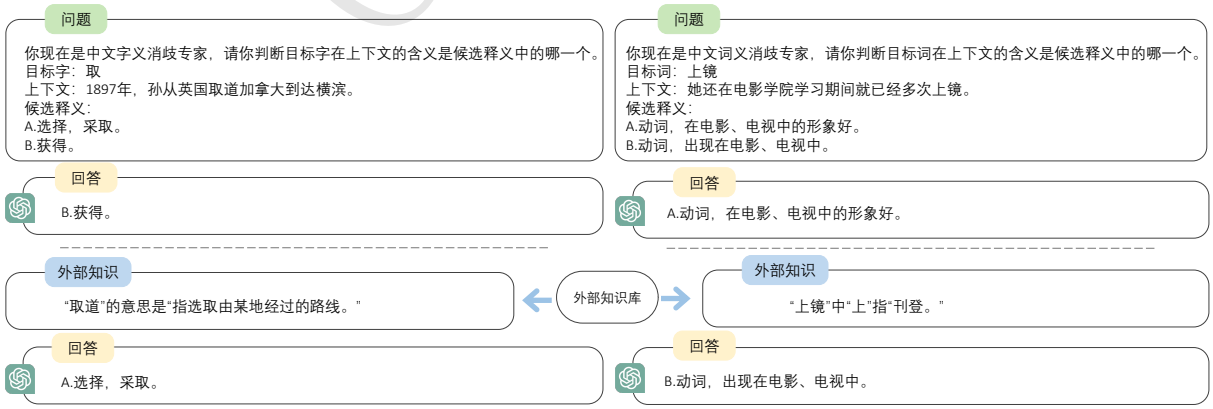


Figure 1: 基于外部知识的模型语言理解能力增强例子

[1] https://platform.openai.com/docs/models/gpt-3-5
[2] 本文采用的均为GPT-3.5-turbo-0613，以下简称为GPT-3.5。

受限于资源建设的不足和有效数据的匮乏，前期工作大多忽略了这方面的考察。本文希望从语言学角度出发，探索字词资源对大语言模型的汉语准确理解能力的效度：期望解决的一个问题是基于已有的受限资源，如何有效开展面向检索增强生成的知识工程，以更好地满足相关计算需求；在此基础上，另一个问题是如何实现知识工程获得的批量信息与不同模型的有效结合，并验证其有效性。

2 相关工作

2.1 汉语字词信息相关知识工程

汉语遵循从字、词、短语到句子的层级结构，因此，对字与词进行语义标注的资源是自然语言处理领域中重要的计算资源。在相关的资源建设上，语素作为汉语中最小的音义结合体(朱德熙, 1982)，在构词中非常活跃，处于重要的地位(尹斌庸, 1984; 徐枢, 1990)，语素义的组合在一定程度上能体现词义(符淮青, 1981)。同时，语素和语素义具有强大的表义能力(Zhang, 1997; Luo, 2013)，因此数量相对稳定，有助于相关资源的构建。

在相关资源建设上，**“汉语语素数据库”**(苑春法and 黄昌宁, 1998) 覆盖了6,763个常用汉字的语素项信息，并对其构成的二字复合词进行了结构描述和意义绑定。该数据库中各义项之间的连接欠缺，限制了其在计算中的深入应用；**“汉字义类信息库”**与**“汉语语义构词信息库”**(亢世勇et al., 2004)包含了6,763个常见汉字的字位及52,366个二字词的义类与释义，并与《同义词词林》(梅家驹et al., 1983)进行语义绑定。其归类以现有《词林》为标准，存在义项的因果参照问题；**“词素-义类数据库”**(吉志薇and 冯敏萱, 2015)提取了《现代汉语词典》（以下简称《现汉》）中2,268个词素并标注其义类，基于此整理了8,984个二字词的词素意义和词素间的词化意义。其收录的二字词采样不均衡且数据规模过小，难以满足大规模计算需求；**“汉语概念词典”**(刘扬et al., 2018)提取和编码了《现汉》中8,514个汉字的20,855个语素义，并构建语义类别，在此基础上，对《现汉》中41,474个二字词的52,108个义项分别进行语素义绑定。这一资源覆盖了汉语中常用字与罕见字的情况，有助于大规模计算与准确理解应用的开展。

这些前期工作，为基于语言学的计算路径提供了数据基础，并使基于汉语字词的检索增强生成成为可能。

2.2 大语言模型检索增强生成方法

检索增强生成技术(Lewis et al., 2020; Guu et al., 2020)旨在通过对外部资源的动态信息检索，为语言模型提供额外的信息。随着具有强大上下文学习能力的大语言模型的出现，这类方法倾向于在现有模型之上补充性地集成检索功能(Ram et al., 2023; Shi et al., 2023)，将原始输入序列与检索到的文本拼接起来作为输入，在问答(Guo et al., 2024; Zhang et al., 2024b)、对话(Peng et al., 2023; Kang et al., 2023)、个性化文本生成(Mysore et al., 2023)等自然语言处理任务中显著提升了基础模型的表现，在代码生成(Sun et al., 2023)、医疗(Chen et al., 2024)、金融(Kang and Liu, 2023)、法律(Savelka et al., 2023)等领域下也有广泛的应用。

在这些面向实体知识或领域知识的检索增强生成之外，部分工作注意到语言知识的潜在价值，尤其是对于非英语的低资源语言。相关工作按知识的来源可以分为两类：第一类采用自动生成的语言知识，如Lu(2023)利用NLLB机器翻译模型(Team et al., 2022)对关键词进行多语言翻译，Gao(2023)利用Stanza(Qi et al., 2020)生成词性标注标签序列。通过将自动生成的语言知识整合到提示词中，两种方法均在多种语言上提高了模型的机器翻译效果。但它们的语言知识质量与范围受知识生成模型的能力限制，在文化相关的或长尾的知识方面存在挑战；第二类则采用权威的语言学知识库，如Tanzer(2023)使用词典和语法手册辅助大型语言模型翻译预训练阶段未接触过的语言。Zhang(2024a)则利用语素知识库和语法手册，在多种下游任务中增强了模型在濒危语言上的表现。这些工作有效地将语言学资源引入大语言模型的检索增强生成，但大都关注模型未学习过的濒危语言，忽略了这些资源对于更多语言的潜在价值。

前人的工作证实了检索增强生成的有效性，为汉语字词资源与大语言模型的结合提供了可行的技术路径。

3 面向检索增强生成的汉语字词知识工程

为了更好地利用字词资源提升大语言模型的汉语准确理解能力，本文基于前人提出的知识库，开展面向检索增强生成的知识工程。

3.1 数据来源基础

在满足计算需求的基础上，检索增强生成所需的资源需要对语言具有良好的覆盖度。本文采用“汉语概念词典”作为基础资源，其中语素作为字的不同使用及意义表征。该资源内的字词信息如表1所示。这一资源发掘了《现汉》中8,514个字的20,855个义项，其中12,783个义项包含了27,215条例句，占全部的61.29% 。该资源同时包含53,780个二字及多字词的65,335个义项，其中29,669个义项包含了42,163条例句，占全部的45.41%。

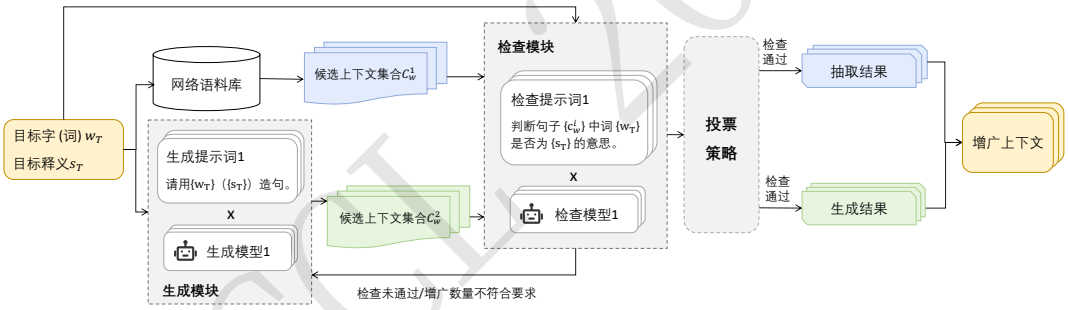
语素	取	**词**	上镜
语素义	选择；采取。	**词义**	出现在电影、电视中。
语素类	动语素	**词类**	动词
例句	1897年，孙从英国取道加拿大到达横滨。	**例句**	她还在电影学院学习期间就已经多次上镜。

Table 1: “汉语概念词典”中的字词知识示例

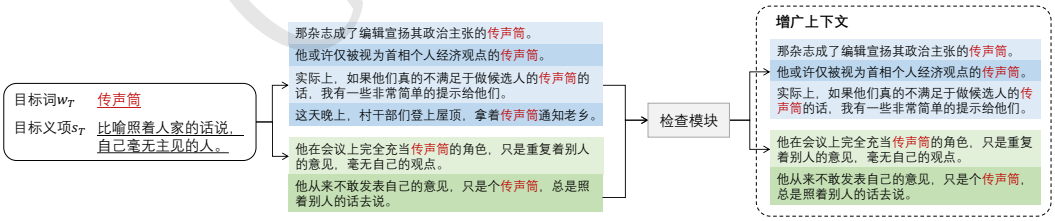
基于“汉语概念词典”，FiCLS数据集(Zheng et al., 2021b)对其中多义且有例句的部分进行了例句文本长度和数量的扩充。具体来说，FiCLS数据集包含2,706个字的10,639个义项及92,135条例句，平均每个义项8.66条。该数据集同时包含4,358个二字词的9,597个义项及27,061条例句，平均每个义项2.82条。

迄今为止，这些资源尚无法为《现汉》中的全部义项提供足够的例句信息，限制了它们在下游任务上的充分应用。本文提出利用网络语料库与生成式语言模型，在“汉语概念词典”与FiCLS数据集的基础上，对其中无例句或少例句的义项进行增广，为后续的实验与评估提供数据基础。

3.2 增广方法与设置



(a) 增广流程



(b) 增广示例

Figure 2: 面向检索增强生成的汉语字词知识工程

本文提出的增广方法如图2a所示。对于目标字（词）w_T的释义s_T，我们通过两种方法，提高增广例句的可靠性与多样性，同时保障增广数量：第一种方法基于网络语料库，从中抽取包含w_T的候选例句集合C_w^1；第二种方法基于生成式模型，结合多种模型与提示词为w_T生成候选例句集合C_w^2。所有候选例句均会通过检查模块，最终由投票策略判断是否通过。图2b中给出了词“传声筒” 的增广示例，蓝色与绿色部分分别为来自两种方法的例句。可以看出，检查模块有效地排除了不符合s_T的增广结果；同时，网络语料库中的文本比自动生成的更为灵活多样。

在**网络语料库构建**方面，考虑到增广例句的多样性，我们选择人民日报[3]、造句网[4]与古诗词网[5]作为语料来源。其中，人民日报文本数量充足，但表述相对正式、单一；造句网文本来源广泛、表述灵活，但数量有限；古诗词网则囊括大量古代诗词典籍，能有效增广部分在古汉语中更常见的字词。语料库最终涵盖人民日报2018-2020语料2,573,625句，包含69,032,189词；造句网语料447,350句，包含10,448,806词；古诗词网语料55,628句，包含8,463,423词。

在**生成模块**与**检查模块**中，我们采用三个在多种评测中表现良好的大语言模型：GPT-3.5、Qwen-max(Bai et al., 2023)与ChatGLM-turbo(Zeng et al., 2022; Du et al., 2022)，后两个为面向中文的模型，保证了良好的中文理解与生成能力。模型参数保留为默认参数，以获得最佳效果。模型提示词通过小批量验证的方式选取（附录A.1）。生成模块中，3个模型轮流用提示词生成例句；检查模块中，3个模型用所有提示词产生共9条投票结果。

依据上述投票结果，由**投票策略**决定是否保留该候选例句。其中，阈值$\#neg \leq 0$表示仅反对票数量≤ 0的例句会被认为检查通过。为了探究不同阈值对投票结果的影响，我们从各来源的例句中分别抽取300条，人工评估其正确性。结果如表2所示。其中，P为准确率，R为召回率，“单义”代表《现汉》中只有一条义项的字或词。最严格的投票策略（$\#neg \leq 0$）带来的准确率最高，但召回率极低，尤其是在“古诗词网”语料中。这主要是由于模型对古汉语的理解较差，因此倾向于在检查模块中给出“否”的判断结果。为了在保证质量的前提下尽量扩大增广数量，本文设置投票策略如下：1. 单义字词的阈值为$\#neg \leq 3$、古诗词网语料的阈值为$\#neg \leq 4$；2. 对于不属于1中的其余候选例句，先按阈值$\#neg \leq 0$进行第一轮筛选。对于筛选后未得到增广的义项，再按阈值$\#neg \leq 1$补召回。

	阈值	基于语料库						基于生成	
		人民日报		造句网		古诗词网			
		P	R	P	R	P	R	P	R
全集	≤ 0	**95.08**	61.70	**96.77**	60.61	**100.00**	16.06	**96.88**	24.80
	≤ 1	94.44	72.34	95.71	67.68	97.50	28.47	94.34	40.00
	≤ 2	91.76	82.98	93.83	76.77	98.44	45.99	95.77	54.40
	≤ 3	88.66	91.49	90.22	83.84	98.88	64.23	91.92	72.80
	≤ 4	88.00	**93.62**	87.13	**88.89**	98.15	**77.37**	86.61	**77.60**
单义	≤ 0	97.44	69.09	**100.00**	57.50	**100.00**	16.06	94.12	15.69
	≤ 1	97.83	81.82	96.30	65.00	97.50	28.47	96.00	23.53
	≤ 2	**98.00**	89.91	96.97	80.00	98.44	45.99	**97.06**	32.35
	≤ 3	96.30	**94.55**	97.22	87.50	98.88	64.23	92.86	**38.24**
	≤ 4	94.55	**94.55**	95.00	**95.00**	98.15	**77.37**	82.98	**38.24**

Table 2: 例句增广中，检查模块在不同阈值下的准确率与召回率

在**实验设置**方面，由于网络语料库数据量过大，在基于语料库的增广过程中，对于拥有K个不同义项的字或词，我们为每个义项随机抽取$10 \times K$个候选例句。考虑到网络语料库的例句质量更高，我们优先采用基于网络语料库的方法，增广数量不足3条的用基于生成式模型的方法进行补充。

3.3 增广结果与分析

通过基于语料库的方法，我们共得到增广例句498,126条，其中，342,867条来自人民日报语料，149,390条来自造句网语料，5,869条来自古诗词网语料。在基于语料库的方法后，仍有17,876条语素义、31,239条词义的增广例句数量不足3条。通过基于生成式模型的方法，我们进一步获得增广例句28,724条。最终，本文得到增广例句526,850条。为确保其可靠性，我们分别从字与词的增广例句中随机抽取1,000条进行人工评测，统计与评测结果如表3所示。

我们将前人构建的资源（“汉语概念词典”与FiCLS）中的数据合并去重后得到187,642条增广前例句，与本文的增广例句合并后共得到714,492条全集例句，统计信息如表4所示。其中，“覆盖率(%)”为包含例句的义项比例。从统计结果可以看出，增广后例句的平均数量和覆盖率都有了显著提高，覆盖率均达到80%以上，展示出这一增广的显著效果。

[3] http://paper.people.com.cn/
[4] https://zaojv.com/
[5] https://www.gushici.com/

	新增例句数	准确率(%)	平均例句长度	字词数	义项数
语素	33,022	96.00	39.90	7,182	10,701
词	493,828	99.00	47.84	46,158	53,752
全集	526,850	97.50	47.34	53,340	64,453

Table 3: 增广例句统计结果与人工评测

类型	字				词			
	例句数	平均例句数	平均例句长度	覆盖率(%)	例句数	平均例句数	平均例句长度	覆盖率(%)
增广前	118,625	5.69	3.66	61.49	69,017	1.06	7.31	45.90
增广后	151,647	7.27	23.43	83.54	562,845	8.61	43.70	89.43

Table 4: 例句增广前后统计指标对比分析

4 基于汉语字词资源的检索增强生成方法

考虑到方法的适用性，本文利用检索增强生成，探索字词资源在大语言模型的汉语准确理解能力上的效度，如图3所示，包含基于关键词的知识检索与基于评估的知识排序两个流程。其中，我们以第3节得到的字词资源用作外部资源。该方法的输入信息包含用于获取关键词的文本$C_{keyword}$，以及用于计算相似度的文本C_{sim}，二者均来自任务的原始输入信息（本节的主要符号对照表见附录C）。

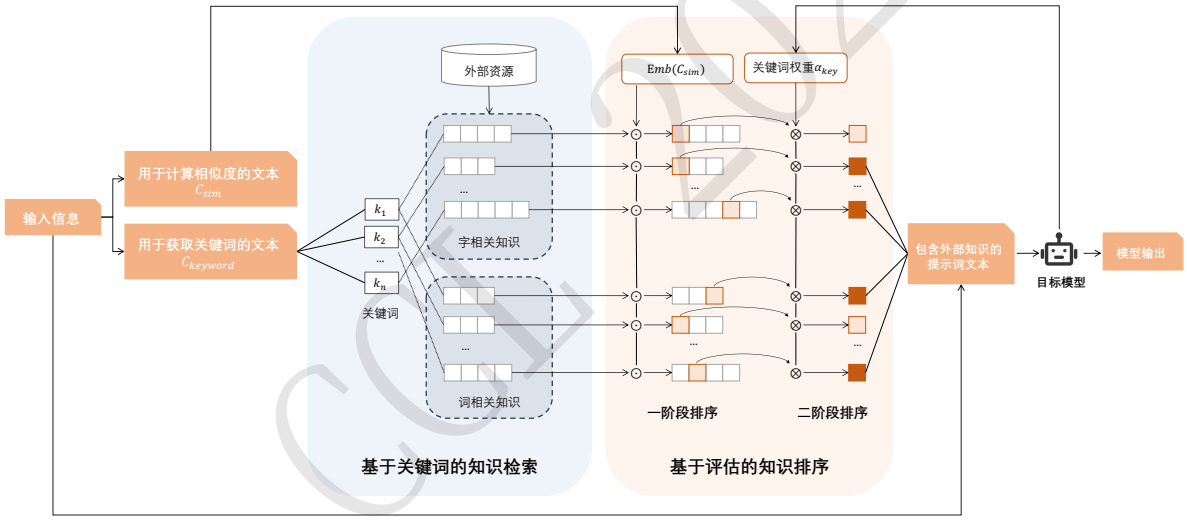


Figure 3: 基于汉语字词资源的模型增强方法

4.1 流程一：基于关键词的知识检索

基于关键词的知识检索负责基于目标文本，从字词资源中获取可能的相关知识。知识检索的目的是将输入文本与知识进行有效匹配，减少无关信息的引入，这一流程包含以下两步：

第一步，从用于获取关键词的文本$C_{keyword}$中抽取关键词集合$K = \{k_1, k_2, ..., k_n\}$，将文本转为满足字词资源需求的形式。这一流程旨在减少无关信息的引入，提高知识注入的效率；

第二步，利用关键词k_i检索字词资源，得到候选知识集合$D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,l}\}$，其中包含释义信息，以及例句中的释义信息。表5展示了字“取”和词“取道”作为关键词时，分别获取到的部分知识示例。

4.2 流程二：基于评估的知识排序

基于评估的知识排序负责基于已有的候选知识，筛选得到最终注入到提示词中的部分。知

字相关知识	释义信息	“取”的意思是“选择，采取”。
	例句中的释义信息	“取”在“取道”中的意思是“选择，采取”。
词相关知识	释义信息	“取道”的意思是“指选取由某地经过的路线”。
	例句中的释义信息	“取道”在“1897年，孙从英国取道加拿大到达横滨。”中的意思是“指选取由某地经过的路线”。

Table 5: 从字词资源中获取到的字词相关知识示例

识注入的目的是通过提供与目标任务相关的、模型理解不稳定或有困难的知识，提高模型对任务的能力。因此，准确筛选候选知识，关键在知识与输入信息的相关性，以及知识对目标模型的有效性。基于以上考虑，这一流程分为以下两个阶段：

第一阶段排序负责从关键词下的所有知识中，筛选得到与用于计算相似度的文本C_{sim}最相关的N_{stage1}个。考虑到候选知识除释义之外还包含例句，且总体目的与传统的词义消歧任务不完全一致，本节采用句向量之间的余弦相似度作为一阶段排序分数。对于关键词k_i的第j条候选知识$d_{i,j}$，它与C_{sim}的余弦相似度计算方式为：

$$cos\_similarity = cos(Emb(d_{i,j}), Emb(C_{sim})) = \frac{Emb(d_{i,j}) \cdot Emb(C_{sim})}{|Emb(d_{i,j})| \times |Emb(C_{sim})|} \in [-1, 1] \quad (1)$$

其中，Emb表示获取文本的句嵌入向量。考虑到本文只需要知识之间的相对顺序而非绝对数值，为了便于后续计算，我们定义第一阶段的排序分数如下：

$$score_{stage1} = 1 + cos\_similarity \in [0, 2] \quad (2)$$

在此基础上，第二阶段排序则负责从前一阶段得到的所有知识中，筛选出对模型最有效的N_{stage2}个。具体来说，我们希望获取目标模型理解最困难的知识，适当增加其权重，提高它们被注入到提示词中的概率，进而提高注入知识对目标模型的有效性。本文采用了此前MorphEval(Yin et al., 2024)中的评估结果。该工作在“汉语概念词典”的基础上，构建了基于汉语语素的大语言模型评估数据集，用于评估模型的语言准确理解能力。对于候选知识$d_{i,j}$，我们获取其在MorphEval数据集中出现的总次数$T_{i,j}$，以及模型对其预测错误的次数$E_{i,j}$。利用这两项数据，我们计算其权重$\alpha_{i,j}$如下：

$$\alpha_{i,j} = e^{\frac{E_{i,j}}{T_{i,j}}} \in [0, e] \quad (3)$$

在此基础之上，我们定义第二阶段的排序分数为：

$$score_{stage2} = \alpha \times score_{stage1} \in [0, 2e] \quad (4)$$

最终，我们利用为不同任务与不同类型知识构造的提示词模版，将获取到的N_{stage2}条知识融入到原提示词中，作为当前任务的输入信息。

5 检索增强生成的应用评估

5.1 语言准确理解评估任务的选取

为了验证汉语字词资源对大语言模型的汉语准确理解的有效性，我们选择了多个与语言的准确理解紧密相关的词内、词间自然语言处理任务进行评估：

一、汉语语素义消歧任务（Morpheme Sense Disambiguation, MSD）(Wang et al., 2024)旨在预测目标语素在特定上下文中的含义。该任务的消歧对象包含自由语素与黏着语素（详细介绍见附录B）。此前的相关工作基于BERT模型(Wang et al., 2024)，我们在本文将这一任务引入到对自回归语言模型的评测中，为其构建适合模型的选择题任务形式：给定语素、语素所在的上下文和所有语素义，要求模型选择上下文中的语素属于哪一个语素义。

二、汉语词义消歧任务（Word Sense Disambiguation, WSD）(Navigli, 2009)旨在预测目标词在特定上下文中的含义，是最核心的词级自然语言处理任务之一。该任务的消歧对象

除可独立成词的自由语素之外，还包含二字及多字词。此前的相关工作包括两种常见方法：基于知识的方法利用现有的大规模知识资源来推断上下文中的词义(Galley and McKeown, 2003; Agirre et al., 2014)，基于监督学习的方法则利用带有义项标注的数据集训练分类器(Huang et al., 2019; Scarlini et al., 2020; Zheng et al., 2021b)。然而这些方法采用的都是BERT式自编码模型，忽略了自回归模型的评估与应用，尤其是诸如GPT-3.5的大型生成式语言模型。本文将这一任务引入到对大语言模型的计算中，并沿用之前的实验设计，为这类模型构建选择题的任务形式。

三、汉语隐喻识别任务（Metaphor Detection, MD）旨在判断与识别文本中字面义和非字面义处于不同的语义域中的现象（详细介绍见附录B）。隐喻是蕴含丰富文化背景的语言现象，普遍存在于人类日常交流中(Richards, 1936)，因此对大语言模型至关重要。此前的相关工作同样大多基于BERT式自编码模型(Su et al., 2020; Choi et al., 2021; Li et al., 2023)，对自回归模型的关注较少，其中，Wachowiak等(2023)在英语与西班牙语中利用RiC（Reasoning in Conversation）提高GPT-3的隐喻源域预测能力，Tong等(2024)将任务转换为改写形式，在英语上评估多种生成式模型的表现。考虑到本文主要探究字词信息的知识注入，我们选择词级隐喻识别任务作为评估，即判断上下文中的目标词是否包含隐喻现象。我们参考前人方法(Comșa et al., 2022; Zhou et al., 2023; Tong et al., 2024)，为模型构建了选择题、判断题与填空题三种任务形式。

5.2 实验数据及其统计信息

语素义消歧任务中，为了获取能完整覆盖全部义项的评估数据集，本文沿用Wang等(2024)中提出的方法，从“汉语概念词典”中抽取语素、词与上下文信息，保留其中多义语素的部分。数据集共23,914条，根据上下文信息的不同可以分为词内和句内语素义消歧数据集。

词义消歧任务中，考虑到输入特征与语素义消歧任务的一致性，本文采用FiCLS数据集(Zheng et al., 2021b)用于评估。由于其全集数据的人工评测准确度有限，我们仅采用其中来自原始语料库而非自动增广的部分。数据集共28,770条，按消歧对象的不同可以分为单字词、二字词与多字词数据集。

词级隐喻识别任务中，本文采用陈龙等(2019)提出的非字面义数据集作为评估。这一数据集发掘了《现汉》中的3,524个非字面义词，包含隐喻和转喻两种词级现象，对应4,301个数据条目，与语素义消歧与词义消歧任务中的语素、语素义和词义等信息保持了一致性。

三个任务的数据集统计信息如表6所示。其中隐喻识别任务数据无释义，因此不做统计。

任务	数据集	数据条目数	字词数	义项数	上下文数	平均上下文长度
语素义消歧	词内消歧	13,276	5,519	13,276	12,356	2.15
	句内消歧	10,638	2,706	10,638	10,591	42.01
词义消歧	单字词	7,685	2,133	6,515	6,506	34.10
	二字词	19,316	8,683	19,316	19,239	27.45
	多字词	1,769	828	1,746	1,763	23.25
隐喻识别	选择形式	3,313	3,705	-	5,922	8.57
	判断形式	5,988	3,730	-	5,975	8.46
	填空形式	5,988	3,730	-	5,975	8.46

Table 6: 实验数据统计信息

5.3 实验方法与设置

在**实验设置**方面，在基于关键词的知识检索中，我们设置$C_{keyword}$与C_{sim}为上下文，并采用pkuseg分词器(Luo et al., 2019)将$C_{keyword}$切分为字词。在基于评估的知识排序中，对于两阶段检索的数量，为保证提示词总长度在模型预训练长度范围之内，我们设置$N_{stage1} = 3$，$N_{stage2} = 3$，即在一阶段排序后，每个关键词最多保留3条候选知识；在二阶段排序后，总体保留得分最高的3条加入提示词。对于句向量的获取（Emb），我们采用了为检索增强大语言模型而设计的bge-base-zh-v1.5(Chen et al., 2023)。

在**模型选择**方面，我们分别选择GPT-3.5与ChatGLM2-6B(Du et al., 2022; Zeng et al., 2022)。其中，GPT-3.5是基于Transformer(Vaswani et al., 2023)的黑盒模型，在多种语言与任务下都取得了良好效果，是最为广泛使用的大型多语言模型之一，本文通过API进行调用。ChatGLM2模型是基于GLM架构的开源双语模型，在预训练阶段就加入了大量的中文数据集，使得其在许多中文领域任务中能够取得显著的效果，我们通过开源模型进行调用。

在**模型设置**方面，为确保实验的公平性和可复现性，所有模型均设定为相同的参数。其中，我们设置温度为0，频率惩罚和top-k为0，top-p为1，以尽量接近贪婪搜索策略。对于无法精确匹配上述参数的ChatGLM2，我们尽可能地调整其设置以接近这些参数。

在**提示词设置**方面，考虑到生成式语言模型对提示词的敏感性，在每个任务下，我们都首先为模型构建3种常用的提示词形式，并指导GPT-3.5额外生成7条表达方式各异、长度不同的新提示词。随后我们通过小批量数据验证的方式从中选取对于当前模型和任务最适宜的，作为后续的评估提示词，以确保最大程度激发模型的能力。具体的提示词设置情况见附录A.2。

特别需要说明的是，对于语素义与词义消歧任务，为了避免检索阶段对预测结果产生标签泄露，我们对所有检索结果都进行了过滤。对于直接提供类似“{目标字}的释义是{标签释义}”或“{目标字}在{目标上下文}中的释义是{标签释义}”的检索结果，我们均将其删除。

5.4 语言准确理解任务评估结果与分析

5.4.1 主实验结果

我们的方法在三个评估任务上的主要实验结果如表7所示。鉴于本文的核心目标是验证字词资源在大语言模型检索增强生成中的应用效果，我们选择了模型在未注入相关知识时的性能作为参照基准。为了达到更好的覆盖度，本文采用了各任务数据的全集；而在传统方法中，由于需要做模型训练，采用的评估数据集往往是其子集。由于数据覆盖的差异，本文方法与传统方法无法做公平对比。作为参考，传统方法的最佳结果见附录D。其中，对于隐喻识别任务填空形式下的随机基线，其计算结果基于pkuseg分词器获取的上下文中所有可能词汇的数量。

模型	语素义消歧		词义消歧			隐喻识别			Avg.
	词内	句内	单字词	二字词	多字词	选择	判断	填空	
Random	15.35	12.84	18.60	44.24	45.07	25.00	50.00	21.86	29.12
GPT-3.5	58.26	46.62	50.28	62.87	67.27	36.46	56.23	53.46	53.93
+本文方法	**67.93**	**48.89**	**61.57**	**66.59**	**68.40**	**41.29**	**56.43**	**61.39**	**59.06**
ChatGLM2	33.82	32.48	43.67	55.80	59.92	24.66	52.84	47.65	43.86
+本文方法	**48.15**	**33.85**	**47.32**	**59.05**	**62.30**	**26.65**	**55.01**	**49.63**	**47.75**

Table 7: 主实验结果，其中*Random*代表随机基线

表7中的实验成果表明，本文方法在所有实验模型与任务中均带来了显著提升，平均达4.78%。在任务层面，本文方法分别带来了6.91%、4.24%和3.19%的准确率提升，凸显了这一方法在不同任务中的普遍适用性和和有效性。在模型层面，本文方法在GPT-3.5与ChatGLM2上分别提升准确率5.13%、3.89%，这表明无论是面向英文或中文的模型，本文方法都能为其带来有效的知识。其中，GPT-3.5虽然整体效果优于ChatGLM2，但在本文方法下的提升更为显著，展示出这一方法面向更强大模型的潜力。

为了深入探究字词资源的影响，我们对三个任务分别做案例分析，如附录E所示。分析结果表明，在直接注入与目标语素或词相关的知识之外，上下文中其他字词的知识同样有助于优化模型的表现，展示出字词资源广泛的应用价值。这也解释了我们在设置$C_{keyword}$与C_{sim}为上下文而非目标语素或词的情况下，依旧能带来显著的提升的原因。

5.4.2 消融实验结果

为了验证本文方法中各个模块的有效性，我们对检索、排序及增广例句等关键组成部分进行了消融实验。实验结果如表8所示，其中“排序$_2$”与“排序$_1$”分别代表第二、第一阶段排序。

随排序$_2$、排序$_1$与检索模块的逐步移除，模型在各任务上的准确率均呈递减趋势，证明了它们在获取字词信息上的重要性。其中，移除排序$_2$后，模型准确率分别

模型	方法	语素义消歧		词义消歧			隐喻识别		
		词内	句内	单字词	二字词	多字词	选择	判断	填空
GPT-3.5	**本文方法**	67.93	**48.89**	**61.57**	66.59	**68.40**	**41.29**	**56.43**	**61.39**
	−排序$_2$	67.63	47.02	61.56	66.36	67.55	41.17	56.40	58.15
	−排序$_{1\&2}$	66.82	46.99	56.88	62.11	63.82	40.98	56.31	58.05
	−检索&排序$_{1\&2}$	53.70	43.74	54.08	61.99	63.75	33.14	56.29	58.00
	−增广例句	**68.02**	47.32	61.00	**67.45**	67.89	41.26	56.43	58.78
Chat-GLM2	**本文方法**	**48.15**	**33.85**	**47.32**	59.05	**62.30**	**26.65**	**55.01**	**49.63**
	−排序$_2$	44.05	33.63	47.31	59.04	62.13	26.65	55.00	49.58
	−排序$_{1\&2}$	42.93	31.90	42.72	54.03	57.77	26.08	52.62	49.43
	−检索&排序$_{1\&2}$	30.82	28.41	39.85	53.50	57.72	24.81	45.82	49.25
	−增广例句	44.70	33.14	46.77	**59.56**	60.83	25.72	47.38	49.38

Table 8: 消融实验结果

平均下降0.83%、0.57%；移除排序$_1$下降1.74%、2.59%；移除检索模块则进一步分别带来3.41%、3.41%的下降。

对于字词资源，移除增广例句后两个模型的平均准确率分别下降0.46%、1.81%，印证了其价值。然而，这一效果并非在所有任务中都同样显著。例如，在词内语素义消歧任务中，GPT-3.5在移除增广例句后准确率反而小幅提升了0.09%。

5.4.3 例句信息对实验结果的影响

我们注意到，部分场景下采用无增广例句的实验效果略优于增广后的。为了探究其背后的原因，我们以ChatGLM2模型为例，分别分析字词资源内例句长度与覆盖率的影响。

表9展示了例句长度对实验结果的影响，“平均长度”指字词资源中所有例句的平均长度，“覆盖率”指其中包含例句的义项占全部义项的比例，与表4中的“覆盖率”一致。构建不同长度字词资源的方法如下：原始资源中的同一义项下可能包含多条例句。对于所有义项，我们仅选取其中最短的一条，得到的字词资源中平均例句长度为22.23；选取其中长度第二短的例句，得到的字词资源中平均例句长度为31.40，以此类推。结果显示，随例句平均长度的增加，模型在三个任务上的指标均呈下降趋势，这可能是因为较长的例句会分散模型对关键信息的注意力。然而，在单字词义消歧和隐喻识别的填空任务中，较长的例句反而能取得较好的效果，展示出不同任务对其长度的需求存在差异。这一结果也解释了消融实验中，增广例句导致某些任务的效果降低：如表4所示，增广后的字词资源中，例句覆盖率得到了提高，但其平均长度也显著增加了。

平均长度	覆盖率(%)	语素义消歧		词义消歧			隐喻识别		
		词内	句内	单字词	二字词	多字词	选择	判断	填空
22.23	88.00	**48.97**	**34.11**	45.94	**58.84**	**60.77**	**27.16**	**52.92**	48.51
31.40	88.00	48.31	34.00	**46.49**	58.59	60.32	26.44	52.72	49.25
39.78	88.00	48.13	33.71	45.37	56.77	59.58	25.02	52.47	**49.35**
42.21	88.00	48.25	33.68	44.80	56.58	59.52	25.32	51.69	48.91
43.01	88.00	48.03	32.23	44.02	56.32	59.15	24.99	51.40	48.88

Table 9: 例句长度对实验结果的影响

表10展示了例句覆盖率对实验结果的影响。构建不同覆盖率的字词资源的方法如下：第一步，对于所有义项，我们仅保留其中最短的一条例句；第二步，我们随机删除部分义项的例句，构造出平均长度接近但覆盖率不同的字词资源。结果显示，随着例句覆盖率的下降，模型在三个任务上的指标呈下降趋势。这可能是由于过低的例句覆盖率会导致有效信息的缺失，也从侧面证明了例句增广知识的价值。

在例句信息之外，我们还开展了对知识类型与知识数量的实验分析（附录F）：1. 在**知识类型**上，增加词性信息后，两个模型在所有任务上的指标均有下降。这可能是因为释义文本已

平均长度	覆盖率(%)	语素义消歧		词义消歧			隐喻识别		
		词内	句内	单字词	二字词	多字词	选择	判断	填空
22.23	88.00	**48.97**	**34.11**	**45.94**	**58.84**	**60.77**	**27.16**	**52.92**	**48.51**
22.17	70.40	48.68	33.34	45.10	58.10	60.03	26.71	52.45	48.43
22.21	52.80	48.51	33.08	44.22	57.38	60.37	26.56	51.90	48.48
22.14	35.20	48.47	32.52	43.73	56.77	60.03	26.08	51.54	48.20
22.13	17.60	48.28	32.03	43.68	56.52	60.15	26.02	51.52	48.18

Table 10: 例句覆盖率对实验结果的影响

有效地传达了词性相关信息，显式词性标签对大语言模型而言也较难理解。这为未来更多语言学知识的注入提供了参考；2. 在**知识数量**上，模型在$N_{stage1} = 1$时效果更好，侧面反映出一阶段排序的有效性；而N_{stage2}的增加会显著提高预测准确率，展示出字词资源对模型的有效性，以及本文方法的潜力。

6 结语

本文从语言学角度出发，探索汉语字词资源对大语言模型的语言准确理解能力的效度。基于之前的工作，我们利用网络语料库和生成式模型构建了包含498,126条例句的汉语字词资源，显著提升了其中例句的覆盖率与丰富度。在此基础上，探索这一资源与大语言模型的结合，提出了一套能灵活适应不同任务与模型的检索增强生成方法。评估实验表明，该方法在所有实验模型与任务中均带来了显著的准确率提升，平均达4.78%。其中，在语素义消歧、词义消歧与隐喻识别任务中分别提升了6.91%、4.24%与3.19%，在GPT-3.5与ChatGLM2模型上分别提升了5.22%和4.33%，这显示出字词资源在模型对语言准确理解能力上的价值。通过消融实验，我们进一步验证了这一资源与方法的有效性，并在分析实验中深入探究例句信息、知识类型和知识数量对实验结果的影响，为语言学资源与大语言模型的结合提供新的路径与方法。

在后续工作中，我们将继续开展字词资源的分析与整理，进一步提升例句覆盖率、有效控制其长度。此外，还将探索该资源与方法在更多下游任务与模型中的应用，以便更好地服务于人文与计算领域中的多种任务。

参考文献

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Le Gao, Mingpu Xu, Yue Wu, Yinwen Li, Danli Shi, and Mingguang He. 2024. Eyegpt: Ophthalmic assistant with large language models.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online, June. Association for Computational Linguistics.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only, November. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1486–1488.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. Unleashing the power of chatgpt for translation: An empirical study. *arXiv preprint arXiv:2304.02182*.

Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. Knowledgenavigator: Leveraging large language models for enhanced reasoning over knowledge graph.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China, November. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. FrameBERT: Conceptual metaphor detection with frame embedding learning. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Can Luo. 2013. The classification of morphemes, words and its grammatical and semantic categories in the 3000 characters. Master's thesis, Jinan University, Guangzhou, China.

Gonzalo Martínez, Javier Conde, Pedro Reviriego, Elena Merino-Gómez, José Alberto Hernández, and Fabrizio Lombardi. 2023. How many words does chatgpt know? the answer is chatwords. *arXiv preprint arXiv:2309.16777*.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

I.A. Richards. 1936. *The Philosophy of Rhetoric*. Bryn Mawr College. Mary Flexner lectures. Oxford University Press.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore, December. Association for Computational Linguistics.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4).

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8758–8765, Apr.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July. Association for Computational Linguistics.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online, July. Association for Computational Linguistics.

Weisong Sun, Chunrong Fang, Yudu You, Yuchen Chen, Yi Liu, Chong Wang, Jian Zhang, Quanjun Zhang, Hanwei Qian, Wei Zhao, et al. 2023. A prompt learning framework for source code summarization. *arXiv preprint arXiv:2312.16066*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada, July. Association for Computational Linguistics.

Yue Wang, Hua Zheng, Yaqi Yin, Hansi Wang, Qiliang Liang, and Yang Liu. 2024. Morpheme sense disambiguation: A new task aiming for understanding the language at character level. Proceedings of the 29th International Conference on Computational Linguistics.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Yaqi Yin, Yue Wang, and Yang Liu. 2024. Chinese morpheme-informed evaluation of large language models. Proceedings of the 29th International Conference on Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024a. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024b. Knowgpt: Knowledge injection for large language models.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024c. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Kai Zhang. 1997. Statistical analysis of Chinese morpheme-based characters. *Language Teaching and Linguistic Studies*, (01):43–52.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021a. Decompose, fuse and generate: A formation-informed method for Chinese definition generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531, Online, June. Association for Computational Linguistics.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021b. Leveraging word-formation knowledge for chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A human-centric benchmark for evaluating foundation models.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? an evaluation by asking natural language questions.

亢世勇, 李毅, 孙道功, and 张楠. 2004. 汉语系统语料库的建设与词典编纂. *2004*年辞书与数字化研讨会论文集, pages 145–151.

刘扬, 林子, and 康司辰. 2018. 汉语的语素概念提取与语义构词分析. 中文信息学报, 32(2):10.

吉志薇and 冯敏萱. 2015. 面向普通未登录词理解的二字词语义构词研究. 中文信息学报, 29(05):63–68+83.

尹斌庸. 1984. 汉语语素的定量研究. 中国语文, 1(05):338–347.

徐枢. 1990. 语素. 人民教育出版社.

朱德熙. 1982. 语法讲义. 商务印书馆.

梅家驹, 竺一鸣, 高蕴琦, and 殷鸿翔. 1983. 同义词词林. 上海辞书出版社:上海.

符淮青. 1981. 词义和构成词的语素义的关系. 辞书研究, 1:98–110.

苑春法and 黄昌宁. 1998. 基于语素数据库的汉语语素及构词研究. 世界汉语教学, 02:8–13.

郑廻. 2022. 汉语语素构词的嵌入表示与应用评估. Master's thesis, 北京大学.

陈龙, 饶琪, and 刘扬. 2019. 汉语词的非字面义的表示与应用. 中国科学：信息科学, 49(08):1005–1018.

A 提示词选取结果

A.1 面向模型增强的汉语字词知识工程提示词

面向模型增强的汉语字词知识工程提示词如表11所示，其中{w_T}为目标字或词，{s_T}为目标释义。

A.2 基于字词资源的检索增强评估提示词

三个评估任务下选用的提示词如图4所示。其中词内、句内语素义消歧所用提示词一致，单字词、二字词与多字词词义消歧所用提示词一致。在语素义消歧任务中，考虑到大语言模型对语言学概念不熟悉，数据集内也均为单字语素，我们并未在提示词中采用“语素”，而是用模型更容易理解的“字”作为替换。

模块	提示词
生成模块	请用“{w_T}”（{s_T}）造句。
	请用词“{w_T}”（释义：{s_T}）造句。
	已知词“{w_T}”的意思是“{s_T}”，请用词“{w_T}”造句。
检查模块	判断句子“{c_w^i}”中词“{w_T}”是否为“{s_T}”的意思。回答是或否。
	句子“{c_w^i}”中词“{w_T}”是否可以被解释为“{s_T}”？回答是或否。
	句子“{c_w^i}”中词“{w_T}”的意思是否是“{s_T}”？回答是或否。

Table 11: 面向模型增强的汉语字词知识工程中，生成模块与检查模块提示词

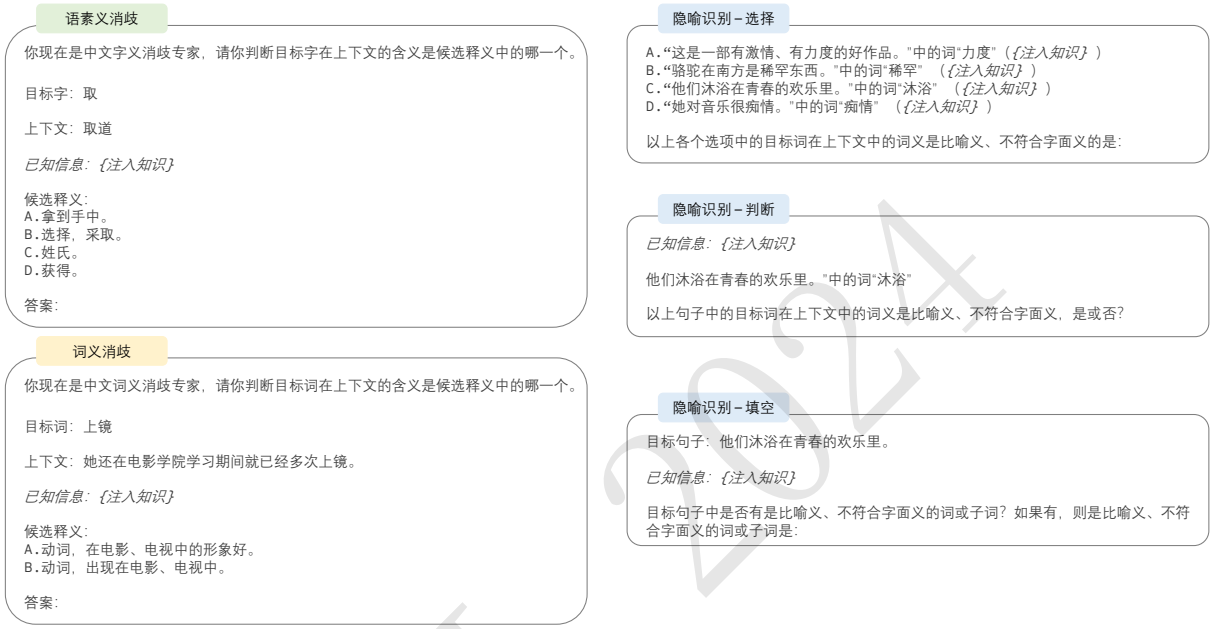


Figure 4: 基于字词资源的检索增强评估提示词

B 部分语言学专有名词说明

自由语素指能够独立构词，无须附于其他语素的语素，例如“花：可供观赏的植物”，可以组词为“花朵”，也可以在句子中独立使用。**黏着语素**与之对应，指不能够独立构词的语素，例如“花：用来迷惑人的；不真实或不真诚的”，只能构成“花招”、“花言巧语”等词语，而不能独立使用。

语义域（semantic domain）语义域是一组语义趋近一致的语言表达的集合。隐喻即为从一个域到另一个域的投射。例如，在句子“他们沐浴在青春的欢乐里。”中，词“沐浴”包含隐喻，它的字面义所处的语义域为“洗澡，沐浴，冲凉，浴，擦澡，淋洗，洗浴”，而非字面义所处的语义域为“沉浸，沐浴，正酣”。

C 主要符号对照表

第4节中所使用的符号说明如表12所示。

D 传统方法的最佳结果

三项评估任务在传统方法中的最佳结果如表13所示。这些评估在数据覆盖和任务形式等方面与本文存在差异，但大模型在性能指标上显著落后于传统方法。这展示出大模型在此类任务上的不足，也从侧面反映出本文研究的必要性。

符号	说明
$C_{keyword}$	用于获取关键词的文本
C_{sim}	用于计算相似度的文本
K	关键词集合
k	关键词
D	候选知识集合
d	候选知识
Emb	获取文本的句嵌入向量
α	基于评估的权重
$score_{stage1}$	一阶段排序分数
N_{stage1}	一阶段排序得到的知识数量
$score_{stage2}$	二阶段排序分数
N_{stage2}	二阶段排序得到的知识数量

Table 12: 主要符号对照表

任务		模型	测试集准确率
语素义消歧	词内	BEM-con+PoS(Wang et al., 2024)	95.60
	句内	BEM-con+PoS(Wang et al., 2024)	85.47
词义消歧		FormBERT w/FP(Zheng et al., 2021b)	87.62
隐喻识别		基于汉语语素的隐喻识别模型(郑颋, 2022)	81.00

Table 13: 传统方法中的最佳结果

E 案例分析

语素义消歧的案例如表14所示，包含词内与句内的语素义消歧案例。其中，释义信息与例句中的释义信息均展示出了显著的效果：通过“世兄”在例句中的释义，模型能够正确消歧“世”；通过“营救”的释义，模型能够正确消歧“营”。这一结果证明了字词信息中的释义与例句中释义两类信息的切实有效性。

语素	上下文	原预测结果	注入知识	新预测结果	标签
世	世兄	E.人世间	“雷世兄，是我班门弄斧了”中“世兄”的意思是“旧时对辈分相同的世交的称呼”	C.指有世交的关系	C
营	瞿秋白被捕后鲁迅曾想方设法委托人营救	C.军队编制单位	“营救”的意思是“设法援救”	B.求，谋求	B

Table 14: 语素义消歧任务案例分析

词义消歧的案例如表15所示，包含单字词、二字词与多字词的词义消歧案例。在二字词与多字词消歧的案例中，注入知识是上下文内非目标词的释义，但依然有效地辅助了模型对目标词的消歧：对于词“上门”，提供“入赘”的详细释义强化了模型对关键信息的注意力，帮助模型提供了正确的预测结果。这一结果展示出，在目标词之外，上下文中非目标词的知识对模型的语言准确理解能力同样重要且有效。这也解释了我们在主实验中设置$C_{keyword}$与C_{sim}为上下文而非目标语素或词，依旧能带来显著提升的原因。

隐喻识别的选择形式案例如表16所示。与正确选项相关的知识能够辅助模型直接识别出其隐喻特性，如“膏泽”；而与其他选项相关的知识能够纠正模型错误的语言理解，如“休想”。这一结果展现出，在该任务的选择形式中，对所有选项都提供知识是有助于模型任务表现的。

隐喻识别的判断形式案例如表17所示，包含正样本与负样本的案例。其中，提供与目标词相关的或上下文相关的信息均能有效辅助模型预测正确。例如，在“不成”的案例中，有注入知识的情况下，ChatGLM2模型的完整生成结果为：“上下文中，“不成”一词的意思是表示揣测或反问。因此，“不成”一词在上下文中的词义是符合字面义的。”可以看出，正是我们提供的“莫非”的释义信息，辅助了模型对“不成”的理解，因而得到了正确预测结果。这进一步展示出本文

词	上下文	原预测结果	注入知识	新预测结果	标签
交	交九的天气	A.连接；交叉。	“近日天气渐渐变冷，到了交九的时节。”中“交九”的意思是“进入从冬至开始的‘九’。”	D.到(某一时辰或季节)。	D
上门	入赘俗称「倒插门」、「上门女婿」	B.到别人家里去；登门。	“入赘的女婿前几年腰椎损伤”中“入赘”的意思是“男子到女家结婚并成为女家的家庭成员。”	D.指入赘。	D
包袱底儿	这次比赛中他亮出了自己的包袱底儿	B.指家庭多年不动用的或最贵重的东西。	“比赛”的意思是“在体育、生产等活动中，比较本领、技术的高低。”	C.比喻最拿手的本领。	C

Table 15: 词义消歧任务案例分析

提示词	原预测结果	注入知识	新预测结果	标签
A.“沙漠一直漫延到遥远的天边。”中的词“漫延” B.“不识字就等于睁眼瞎子。”中的词“等于” C.“膏泽后人。”中的词“膏泽” D.“本位主义”中的词“主义”	D	“膏泽”的意思是“比喻给予恩惠。”	C	C
A.“休想逃脱”中的词“休想” B.“下流无耻。”中的词“下流” C.“琅琅上口。”中的词“上口” D.“营造气氛”中的词“营造”	A	“休想”的意思是“别想；不要妄想。”	C	C

Table 16: 隐喻识别任务选择形式案例分析

方法在多种角度上的有效性，以及对完整上下文应用本文方法的必要性。

词	上下文	原预测结果	注入知识	新预测结果	标签
人梯	甘当人梯。	否	“人梯”在目标句子中的词义是“一个人接一个人踩着肩膀向高处攀登叫搭人梯。”	是	是
不成	这么晚他还不来，莫非家里出了什么事不成?	是	“莫非”在目标句子中的词义是“表示揣测或反问，常跟“不成”呼应。”	否	否

Table 17: 隐喻识别任务判断形式案例分析

隐喻识别的填空形式的案例如表18所示。案例展示出，与标签词直接相关的信息能够帮助模型直接意识到其隐喻现象。例如，在“挂钩”的案例中，ChatGLM2模型的完整生成结果为：“目标句子中，“挂钩”一词不符合字面义，因为“挂钩”一词的本义是“用钩子挂住”，与“建立联系”无关。因此，“挂钩”一词是比喻义，不符合字面义。”另一方面，与上下文中其他字词相关的信息也能够帮助模型排除错误的选项，如“远大”的案例。这同样展现出对完整上下文应用本文方法的必要性与有效性。

F 实验分析

我们以ChatGLM2模型为例，进一步开展了对知识类型与数量的分析实验。

F.1 注入知识类型对实验结果的影响

主实验中，我们采用了字词资源中的释义与例句信息作为注入知识。在大模型的语言理解中，词性标注（Part-of-Speech, PoS）标签同样被广泛使用(Gao et al., 2023)。因此，我们在此探究更多的注入知识类型对实验结果的影响。

表19中展示了模型在不同知识类型下的实验结果。其中，“例句中的释义”信息简写为“例句”。实验结果表明，词性信息作为注入知识时，在所有任务上带来了准确率的下降。我们考虑这是由于释义文本通常已经有效地传达其词性信息，显式的词性标签对大语言模型相对更难以理解，因此不利于模型的任务表现。

上下文	原预测结果	注入知识	新预测结果	标签
这两个单位早就挂起钩来了。	无	“挂钩”在目标句子中的词义是“比喻建立某种联系。”	挂钩	挂钩
前途远大。	远大	“远大”在目标句子中的词义是“长远而广阔，不限于目前。”	前途	前途

Table 18: 隐喻识别任务填空形式案例分析

知识类型	语素义消歧		词义消歧			隐喻识别		
	词内	句内	单字词	二字词	多字词	选择	判断	填空
+ 例句&释义	**48.15**	**33.85**	**47.32**	**59.05**	**62.30**	**26.65**	**55.01**	**49.63**
+ 例句&释义&词性	47.58	32.89	45.30	53.67	60.66	25.81	52.76	48.16

Table 19: 注入知识类型对实验结果的影响

F.2 注入知识数量对实验结果的影响

在主实验中，我们设置$N_{stage1} = 3$，$N_{stage2} = 3$，但不同的注入知识数量可能会对实验结果带来影响。我们以词义消歧任务为例，在这一分析实验中对其进行探究。

N_{stage1} \ N_{stage2}	1	2	3	4	5
1	46.08	46.35	47.69	48.30	48.69
2	46.08	46.03	47.44	48.14	48.41
3	46.08	46.03	47.32	48.14	48.41
4	46.08	46.03	47.44	48.14	48.41
5	46.08	46.03	47.44	48.14	47.61

(a) 单字词词义消歧

N_{stage1} \ N_{stage2}	1	2	3	4	5
1	58.64	60.06	61.12	62.11	62.80
2	58.64	58.65	59.07	59.51	59.68
3	58.64	58.65	59.07	59.51	59.68
4	58.64	58.65	59.07	59.51	59.68
5	58.64	58.65	59.07	59.51	59.68

(b) 二字词词义消歧

N_{stage1} \ N_{stage2}	1	2	3	4	5
1	59.30	61.56	62.24	62.97	64.10
2	59.30	61.22	62.13	62.58	62.58
3	59.30	61.22	62.13	62.58	62.58
4	59.30	61.22	62.13	62.58	62.58
5	59.30	61.22	62.13	62.58	62.69

(c) 多字词词义消歧

Figure 5: 注入知识数量对实验结果的影响

实验结果如图5所示。总体来说，随着N_{stage1}的增加模型的指标无显著波动，部分场景下在$N_{stage1} = 1$时取得最佳结果，侧面展现出一阶段排序的有效性；随着N_{stage2}的增加模型在词义消歧任务上的指标均显著增加，展示出知识注入在数量上的可扩展性，进一步验证了汉语字词信息对模型的价值。