大语言模型汉字富语义能力评测

余艺喆，董明 ✉，何婷婷 ✉
人工智能与智慧学习湖北省重点实验室
国家语言监测与研究网络媒体中心
华中师范大学，计算机学院，湖北，武汉
yuyizhe1212@mails.ccnu.edu.cn, {dongming,tthe}@ccnu.edu.cn

摘要

中文相较于以英文为代表的表音文字具有富语义的特点，单个汉字蕴含了读音、字形结构、偏旁部首等丰富的语义特征，在构建自然语言处理相关应用时具有独特的价值，可以视作额外的特征，提升在特定任务的表现。近年来，大语言模型飞速发展，展现出海量的知识储备和强大的推理能力，其中，大模型对汉字富语义特征的掌握可以视作大模型中文能力的基础。然而，目前对于大模型汉字富语义能力评测研究较少，针对性地评测大模型在汉字富语义方面的能力边界，有助于了解大模型中英文能力差异性、并推测大模型在字形、字音相关下游任务上的表现。因此，本研究从汉字的结构、偏旁、读音、笔画、多音字和部件六个维度，对大语言模型进行了全面评测，旨在深入探究其对汉字基本富语义特征的掌握程度。本研究以GB2312 标准字符集和现代汉语词典为依据，围绕汉字的结构、偏旁、读音、笔画、多音字和部件六个维度，构建了一系列“问题-答案”对，并制定了科学合理的评分标准。在此基础上，对十余种主流的大语言模型进行了深入评测。同时，为探究模型在中英文能力上的差异，将上述中文评测任务翻译为英文，并选取了三个代表性模型进行对比评测。此外，本研究进一步从汉字结构推理、偏旁推理、读音推理三个关键角度出发，设计了一系列推理评测任务，旨在深入评估大语言模型对汉字富语义特征的推理能力。本研究的评测结果具有重要的参考价值，可为大语言模型相关领域的研究人员在中文下游任务优化、基础模型选择等关键环节提供参考和启发。[1]

关键词： 大语言模型；汉字；富语义；评测

Evaluating the Rich Semantic Capabilities of Chinese Characters in Large Language Models

Yizhe Yu, Ming Dong ✉, Tingting He ✉
Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning
National Language Resources Monitoring and Research Center for Network Media
School of Computer Science, Central Chine Normal University, Wuhan Hubei
yuyizhe1212@mails.ccnu.edu.cn, {dongming,tthe}@ccnu.edu.cn

Abstract

Compared to phonetic scripts like English, Chinese characters possess rich semantic features, encapsulating pronunciation, structural configuration, radicals, and more. This unique semantic richness adds extra value to natural language processing applications, enhancing task performance. With the rapid advancement of large language models

[1]本研究的评测数据集和源代码见：https://github.com/Yuyuyuyizhe/hanzi-llm-eval

第二十四届中国计算语言学大会论文集，第131页-第142页，济南，中国，2025年8月11日至14日。

(LLMs), which showcase vast knowledge reserves and strong reasoning abilities, understanding the semantic attributes of Chinese characters is foundational to LLMs' capabilities in Chinese. However, research evaluating LLMs' semantic understanding of Chinese characters remains sparse. Targeted evaluation of this aspect helps discern differences between LLMs' Chinese and English capabilities, and predicts their performance in related tasks involving character shape and pronunciation.

This paper conducts a comprehensive assessment of LLMs based on six dimensions: character structure, radicals, pronunciation, strokes, polyphonic characters, and components, deepening our understanding of their grasp of the fundamental semantic features of Chinese characters. Using the GB2312 character set and the Modern Chinese Dictionary, a carefully curated series of "question-answer" pairs were constructed for these six dimensions, accompanied by scientifically sound scoring criteria. Over ten mainstream large language models were extensively evaluated. Furthermore, to explore disparities in capabilities between Chinese and English, the Chinese tests were translated into English, and three representative models underwent comparative evaluation.

Additionally, the study designs a series of inference tests focusing on reasoning about character structures, radicals, and pronunciations to deeply assess LLMs' reasoning in absorbing the semantic features of Chinese characters. The evaluation results from this research hold significant value, providing references and guidance for researchers in optimizing downstream Chinese tasks and selecting foundational models in the LLM domain.

Keywords: Large Language Models , Chinese Characters , Rich Semantics , Assessment

1 引言

近年来，大语言模型飞速发展，展现出了惊人的知识储备和强大的推理能力，以大语言模型（以下简称"大模型"）为基石构建各类下游应用，成为自然语言处理领域的主流范式。若要选取合适的基座大模型，并将其应用于下游任务，准确地掌握大模型的能力边界至关重要。因此，针对大模型的相关评测层出不穷，从多个角度探究了大模型的能力(罗文, 2024)。早期大模型基于以英文为主的互联网开放语料训练，中文所占比例低，在中文任务上并未展现较好的能力(Huang et al., 2023; Yao et al., 2024)。随后，国产大模型针对中文独有特点进行不断优化，开发出了一系列具有更强中文能力的大模型，其在中文各个维度上的能力均大幅度提升(Xu et al., 2023; Wei et al., 2024; Yang et al., 2024)。

中文作为世界上最古老的语言之一，由象形文字发展而来，具有富语义的特点，单个汉字即包含繁体、字形、拼音、声调、偏旁部首等特征，相较于同等篇幅英文可承载更多的信息量，包含更多维度特征(Chi et al., 2024; Wu et al., 2024; 刘宇瀚et al., 2020; 薛莫白, 2024)。每一个汉字字符所包含的字形信息、语音信息，都与其语义有着一定关联，例如"江"、"河"、"湖"、"海"在字形上均包含部首"氵"，语义上也均表示与水相关的实体。例如"乐"在读"lè"的时候，表示开心、快乐；在读"yuè"的时候，表示音乐的意思。因此，从汉字形式中学习富语义信息并将其纳入大模型的工作也受到了越来越多的关注(Liu et al., 2024; Liu and Lian, 2023)。Sun,(Sun et al., 2021)等人通过字形和拼音信息增强的中文预训练方法，将汉字的字形和拼音信息整合到模型的预训练过程中。Lv(Lv et al., 2022)等人也随后提出了将汉字单词、拼音、五笔、拆字作为嵌入信息，用来增强大模型对中文的理解能力。在中文拼写检查领域，也有不少学者利用汉字的富语义信息来提高语言模型在此任务上的表现(Dong et al., 2024; Zhang et al., 2023; 严旭枫, 2024)。Xu(Xu et al., 2021)等人提出了利用多模态信息帮助中文拼写检查的方法，捕获输入字符的语义、语音和图形信息，借以检测和纠正中文文本错误字符。Huang(Huang et al., 2021)等人也提出从音频和视觉模态中推导出汉字的拼音和字形表示，并通过精心设计的自适应门控机制将其集成到预先训练的语言模型中。由此可见，汉字蕴含的

第二十四届中国计算语言学大会论文集，第131页-第142页，济南，中国，2025年8月11日至14日。

评测任务	考察方向	“问题-答案”对
1	结构	汉字“好”是什么结构？汉字“好”是左右结构。
2	偏旁	汉字“好”的偏旁是什么？汉字“好”的偏旁是“女”。
3	读音	汉字“好”的读音有哪些？汉字“好”的读音有：hǎo, hào。
4	笔画	汉字“好”的笔画有多少？汉字“好”的笔画有6画。
5	多音字	“余独好修以为常”中的“好”念什么？“余独好修以为常”中的“好”念：hào。
6	部件	以下哪个选项属于汉字“励”的内部组成？A.午; B.刀; C.而; D.力 选项D“力”属于汉字“励”的内部组成。

表 1: 富语义认知评测任务示例

丰富语义能够被用来增强大模型在中文自然语言处理任务上的性能。鉴于汉字的上述特点，在对中文大模型进行评测时，针对汉字富语义特点设计独特的评测任务，将有效探究中文大模型汉字基础能力的边界，进而为相应下游任务的表现进行合理地预测。同时，在大模型能力评测时加入汉字富语义视角，能为后续大模型优化提供更多可行思路。

本研究旨在对中文生成式大模型的汉字富语义能力（以下简称“富语义能力”）深入研究。通过多维度的评测，本研究综合评测了当今大模型富语义能力的现状，了解大模型对汉字富语义信息的认知程度，以及大模型对汉字富语义信息的推理能力，并最终为相关领域研究人员针对性优化大模型提供参考。

2 评测方法与步骤

2.1 评测数据

本研究设计了两类评测任务：汉字富语义认知评测（评估大模型对汉字多维度语义特征的掌握程度）和汉字富语义推理评测（区分模型回答是基于记忆存储还是逻辑推理能力）。前者系统检验大模型对汉字结构、读音等显性知识的表征水平，后者通过创新性评测任务设计，重点考察模型对隐含语义规律的归纳与推理能力。

2.1.1 富语义认知评测任务

本研究共提出了六类富语义认知评测任务，以评测大模型的富语义认知能力。前四类评测任务用于考察大模型对富语义信息的认知程度，通过提示文本，让大模型回答出某汉字蕴含的富语义信息，例如：汉字结构、偏旁、读音、笔画。后两类评测任务用于考察大模型在对富语义信息的应用能力，一类是让大模型回答某句古诗词或某个成语中多音字的读音，另一类则是让大模型从四个选项中选择正确的属于某汉字的部件。汉字部件指构成汉字的基本组成部分，它们可以是字形中独立存在的单元，例如：“亡”、“口”、“月”、“贝”、“凡”均属于汉字“赢”的部件。基于2.2节中参与评测的大部分大模型难以独立、完整地回答一个汉字所含的部件，因此本研究采用选择题形式来降低试题的难度。富语义认知评测任务的示例如表1所示。

本研究首先针对常用的中文简体字，构建了汉字富语义知识库（以下简称“语料库”）。语料库由爬取GB2312[1]标准而来。GB2312，全称《信息交换用汉字编码字符集基本集》，是中国国家标准总局于1980年发布并从1981年5月1日开始实施的第一个汉字编码国家标准，其中一共收录了6763个汉字以及每个汉字对应的富语义信息。本研究基于GB2312标准，构建了一个包含6763个汉字和每个汉字对应的拼音、部首、笔画、结构、五笔码、四角码等共11项信息的语料库，用于记录汉字的富语义信息，具有国家权威性。

由于汉字的结构、偏旁、读音、笔画特征与汉字语义的关联性较强，而五笔码、四角码等其他的七项特征与汉字语义的关联性较弱，因此本研究根据语料库中6763个汉字及其相对应的结构、偏旁、读音、笔画四个特征，为前四类评测任务每类评测任务构建6763条评测数据。评测数据以“问题-答案”对的形式记录，通过自动提取语料库中的信息，并填入评测任务模板得

[1] https://openstd.samr.gov.cn/bzgk/gb

评测任务	考察方向	“问题-答案”对
1	结构	左边是“亻”部件，右边是“二”部件。请根据我对这个汉字的形容来判断，这个汉字是什么结构？这个汉字是左右结构。
2	偏旁	上边是“宀”部件，下边是“奇”部件。请根据我对这个汉字的形容来判断，这个汉字的偏旁是什么？这个汉字的偏旁是“宀”。
3	读音	形旁是“木”部件，声旁是“直”部件。请根据我对这个汉字的形容来判断，这个汉字的音节最有可能是什么？这个汉字的音节最有可能是：zhi。

表 2: 富语义推理评测任务示例

到。对于多音字类评测任务，本研究从当前人教版小学语文课本中随机挑选包含多音字的诗句或成语，人工构建了143条评测数据；对于部件类评测任务，又从第七版现代汉语词典中随机挑选了86个汉字，人工构建了86条评测数据。“问题-答案”对以字典的形式保存，其中“问题”可直接与提示文本拼接，用于对大模型提问，“答案”则可作为判断大模型回答正误的依据。对评测数据的模板化处理，便于实现基于脚本的自动化评测。最后，本研究将富语义认知评测任务翻译成英文，用于后续进一步探究大模型的中英文对齐能力。

2.1.2 富语义推理评测任务

本研究共提出了三类富语义推理评测任务，分别从结构推理、偏旁推理、读音推理三个角度评测大模型的富语义推理能力，评测任务示例如表2所示。

汉字结构的推理依赖于各部件的组合方式，偏旁推理不仅依赖于部件的组合方式，还依赖于对常见偏旁的熟悉程度，例如：由“亻”和另一个部件组成的左右结构的汉字，通常该汉字的偏旁为“亻”。因此，若大模型能依据评测任务中对于汉字部件及组成方式的描述，推断出汉字的结构、偏旁，则表明大模型具有汉字结构推理能力和汉字偏旁推理能力。形声字是最普遍的一种造字方式，由表示意义范畴的意符（形旁）和表示声音类别的声符（声旁）组合而成。形声字的读音往往与字的声旁有直接关联，例如：形旁为“木”、声旁为“直”的汉字“植”与声旁“直”的音节均为“zhi”。因此，若大模型能依据评测任务中对形旁和声旁的描述，推断出汉字最可能的读音，则表明大模型具有汉字读音推理能力。

本研究根据第七版现代汉语词典，挑选出若干不同位置的汉字部件，随机选择并填入评测任务模板，为每类评测任务构建了约1000条评测数据。描述的汉字有可能不存在，但对其结构推理可基于部件组成方式得出，偏旁推理可基于部件组成方式和最常被当作偏旁的部件得出，读音推理可基于声旁的读音得出，仍可进行推理。

2.2 评测对象

本研究挑选了主流的十余个大模型作为评测的对象。其中包含八个参数量在70亿（以下简称“7B”）左右的大模型，用于探究大模型富语义能力的平均水准，对比分析得到其能力的优势与劣势之处。同时还包含一个同版本但参数量为140亿（以下简称“14B”）的大模型，用于探究富语义能力是否随着大模型参数量的增大而提升。最后，挑选了极具代表性的大模型GPT-3.5用作大模型的对比基准，以便更直观地认知各类大模型的表现。本研究的评测对象如下所示：

- GPT-3.5：OpenAI在2022年底推出的强大语言模型，具备快速响应和良好推理能力，是最早用于ChatGPT[2]的核心模型之一。
- Qwen-14B(Bai et al., 2023)：Qwen系列模型具有强大的基础语言模型，已针对多达3万亿个多语言数据token进行了稳定的预训练。
- Qwen-7B(Bai et al., 2023)：一个较小规模的Qwen系列模型。
- YAYI-7B：YAYI是中科闻歌开发的一个开放的大语言模型。

[2]https://platform.openai.com/docs/api-reference

- BlueLM-7B：BlueLM是由vivo AI全球研究院自主开发的大规模预训练语言模型。
- XVERSE-7B：由XVERSE技术股份有限公司开发的多语言大型语言模型。
- ChatGLM3-6B(Zeng et al., 2022)：ChatGLM是一个基于GLM框架的开放式双语语言模型。
- AquilaChat2-7B：Aquila语言模型以高质量的中英文语言材料为基础，从零开始训练。它可以在更小的数据集和更短的训练时间内获得比其他开源模型更好的性能。
- Baichuan-7B：百川是百川智能的一个70亿参数的语言模型。
- LingoWhale-8B：LingoWhale-8B是深研科技与清华大学NLP实验室联合推出的语言鲸系列模型中首个开源的中英文双语大模型。

对于GPT-3.5，本研究使用了OpenAI的官方API评估其性能，选择gpt-3.5-turbo作为评估模型。对于其他大模型，本研究使用了Hugging Face的Transformers库来评估模型。

2.3 评价指标

针对富语义认知评测任务，本研究选择大模型对于六类评测任务的回复正确率作为评价指标。针对富语义推理评测任务，同样选择三类评测任务的回复正确率作为评测大模型富语义推理能力的评价指标，正确率的数值均保留至小数点后两位。

此外，由于富语义信息的前四类评测任务与后两类评测任务的数据分布不同（前四类评测任务数据来源于GB2312标准，后两类评测任务数据来源于语文课本和汉语词典），为确保评价指标的客观性，本研究还定义了另外两个评价指标：AVG_SCORE和TOTAL_SCORE。AVG_SCORE（平均分）是前四类评测任务正确率的简单平均分数，用于反映大模型对汉字富语义信息的平均认知程度；TOTAL_SCORE（加权总分）是所有评测任务的加权平均分数，用于反映大模型的综合富语义认知能力。

2.3.1 平均分：AVG_SCORE

汉字信息的前四类评测任务均用于考察大模型对汉字富语义信息的认知程度，数据来源分布相同，问题的形式和意义相同，仅考察方向不同。因此，对前四类评测任务的正确率进行简单的平均计算，可以得到大模型对汉字富语义信息的平均认知程度，是评判大模型富语义认知能力的重要参考依据。计算公式如下：

$$AVG\_SCORE = \frac{1}{4}\sum_{i=1}^{4} p_i \tag{1}$$

p_i是第i个评测任务下的得分。

2.3.2 加权总分：TOTAL_SCORE

由本研究2.1.1节可以得知，汉字信息的前四类评测任务的数据来源是GB2312标准，专业、客观、严谨，覆盖的汉字范围也很全面；后两类评测任务是人工参照语文课本和现代汉语词典人工构建的，评测数据量较少，且包含的汉字更为简单、常见。因而相较前四类来说，后两类评测任务虽然理论上更复杂，但在本研究的实际完成过程中会更容易得到较高的正确率，将前四类评测任务的正确率与后两类评测任务的正确率直接比较没有意义。基于上述原因综合考虑，本研究采用加权平均的方式计算大模型富语义认知能力的总得分。本研究将前四类评测任务正确率的权重设置为相同的α_1，又由于“多音字”类评测任务与“部件”类评测任务的数据来源分布和任务难度也不同，因此对“多音字”类评测任务正确率的权重设置为α_2，对“部件”类评测任务正确率的权重设置为不同的α_3。计算公式如下：

$$TOTAL\_SCORE = \alpha_1 \times AVG\_SCORE + \alpha_2 \times p_5 + \alpha_3 \times p_6 \tag{2}$$

其中，p_i表示第i个评测任务下的得分，α_i代表相应任务的权重。由于前四类评测任务的正确率与大模型在富语义认知能力上的真实表现更加接近，因此本研究对这些任务给予较高的权重。此外，根据评测结果，大模型在“部件”类任务中的表现差异较小，未能明显反映出不同模型

之间的能力差距。因此，该类任务的权重被设定为最低，原因尚需进一步探讨。基于以上考虑，α_1设置为0.7，α_2为0.2，而α_3为0.1。

TOTAL_SCORE作为评测大模型的综合认知能力得分，是评判大模型富语义认知能力的最重要的标准。本研究通过这样加权平均的计算方式，能更客观地从TOTAL_SCORE分值上体现出大模型的富语义认知能力。

2.4 大模型提示语句构建

“结构”类试题：

user: 请学习以下案例：1.汉字“啊”是左中右结构；2.汉字“阿”是左右结构；3.汉字“哀”是上下结构；4.汉字“鼻”是上中下结构；5.汉字“国”是全包围结构；6.汉字“还”是半包围结构；7.汉字“单”是单一结构；8.汉字“晶”是品字结构；9.汉字“爽”是嵌套结构。
assistant: 好的，我已经了解了这些案例。

user: 汉字“好”是什么结构？请仿照上述案例的格式一句话回答。

“偏旁”类试题：

user: 请学习以下案例：1.汉字“啊”的偏旁是“口”；2.汉字“阿”的偏旁是“阝”；3.汉字“埃”的偏旁是“土”。
assistant: 好的，我已经了解了这些案例。

user: 汉字“好”的偏旁是什么？请仿照上述案例的格式一句话回答。

“读音”类试题：

user: 请学习以下案例：1.汉字“啊”的读音有：ā, á, ǎ, à, a；2.汉字“阿”的读音有：ā, ē；3.汉字“安”的读音有：ān。
assistant: 好的，我已经了解了这些案例。

user: 汉字“好”的读音有哪些？请仿照上述案例的格式一句话回答。

“笔画”类试题：

user: 请学习以下案例：1.汉字“啊”的笔画共有10画；2.汉字“阿”的笔画共有7画；3.汉字“埃”的笔画共10画。
assistant: 好的，我已经了解了这些案例。

user: 汉字“好”的笔画有多少？请仿照上述案例的格式一句话回答。

“多音字”类试题：

user: 请学习以下案例：1.“我还是喜欢散步”中的“还”念：hái；2.“请把铅笔归还给我”中的“还”念：huán。
assistant: 好的，我已经了解了这些案例。

user: “一行白鹭上青天”中的“行”念什么？请仿照上述案例的格式一句话回答。

“部件”类试题：

user: 请学习以下案例：1.选项D“扌”属于汉字“摊”的内部组成；2.选项A“口”属于汉字“啊”的内部组成。
assistant: 好的，我已经了解了这些案例。

user: 以下哪个选项属于汉字“哎”的内部组成？A.日；B.艹；C.来；D.天。请仿照上述案例的格式一句话回答。

图 1: 富语义认知评测任务的提示文本模板

为使大模型展现出更好的富语义能力，在正式输入评测问题之前，首先通过上下文学习的方法，借助提示语句，让大模型提前学习少量的案例（所有大模型同类评测任务下学习的案例相同），以保证模型能够按照案例的句式格式化响应；再利用前期构建好的“问题-答案”对，自动地向大模型发起提问。以富语义认知评测任务为例，评测的提示文本模板如图1所示，图中黑字部分属于模板，红字部分来自“问题-答案”对中的“问题”。特别是“结构”类评测任务，本研究在上下文中向大模型列举出了“结构”类评测任务中包含的所有汉字结构种类。因为目前参数量在7B左右的大模型对汉字结构的了解不够充分，如果不使用样例进行提示学习，大模型通常只会回答“左右结构”和“上下结构”这两种答案，而在基于GB2312构建的“结构”类评测任务中，“左右结构”和“上下结构”的汉字占比高达83%左右，会错误地得到更高的正确率。因此，

模型	p_1	p_2	p_3	p_4	p_5	p_6	S_1	S_2
Qwen-14B	46.72	76.27	72.87	34.22	87.41	53.52	57.52	63.10
GPT-3.5	56.25	74.77	49.60	19.61	57.63	59.52	50.06	52.52
Qwen-7B	26.74	60.49	66.89	11.58	70.63	47.89	41.42	47.91
Baichuan-7B	45.33	48.76	42.33	13.49	60.14	48.28	37.48	43.09
XVERSE-7B	46.07	38.23	36.31	10.71	60.14	42.53	32.83	39.26
ChatGLM3-6B	38.21	36.47	43.09	9.92	53.85	43.66	31.92	37.48
AquilaChat2-7B	46.18	44.04	31.42	3.24	44.76	52.87	31.22	36.09
BlueLM-7B	21.10	35.81	40.42	6.79	58.04	50.57	26.03	34.89
LingoWhale-8B	7.62	49.22	13.88	7.79	19.58	50.57	19.63	22.71
YAYI-7B	29.44	31.68	11.08	2.88	25.87	43.68	18.77	22.68
平均分	36.37	49.57	40.79	12.02	53.81	49.31	-	-
标准差	14.06	15.19	15.54	8.75	13.86	5.00	-	-

表 3: 大模型在富语义认知评测任务下的得分及两个指标的得分

在评测“结构”类评测任务的过程中，需要让大模型提前学习到所有的汉字结构种类。在实验开始之前，本研究还以GPT-3.5和DeepSeek-V3[3]为测试对象，与其交互测试了十余种提示方式，根据两个模型的回复情况，选择了效果最好的提示语句作为本研究的提示文本模板。

最后，本研究通过正则匹配的方法，从格式化后的大模型回应中提取大模型对于问题的回答，参照“问题-答案”对中的“答案”来判别正误，统计大模型在每类评测任务下的回答正确率。由于不同大模型对于相同提示文本的响应习惯不相同，本研究在收集到大模型的回答之后，会针对性地设计多种不同的正则匹配方法，以提高自动判误方法的准确性。最终判误结果均接受了人工的抽样检查，结果显示，本方法自动判误准确性较高，可满足测评需求。

2.5 其它评测细节

本研究基于GB2312构建的语料库中，汉字结构种类共有15种，其中包含左包围结构、上包围结构、下包围结构、左下包围结构等6种不同的汉字半包围结构。由于大模型对汉字结构的了解程度不够充分，因此为了降低富语义认知评测任务中“结构”类评测任务的难度，在构建“结构”类评测任务的“问题-答案”对的过程中，本研究将上述6种具体的汉字半包围结构均简化为“半包围结构”。对于这6种半包围结构的汉字，大模型无论回答出具体的半包围结构，还是仅回答“半包围结构”，均认为大模型的回答正确。

3 实验结果与分析

3.1 富语义认知能力分析

大模型富语义认知能力的评测结果如表3所示，其中$p_1 \sim p_6$分别是“结构”类评测任务、“偏旁”类评测任务、“读音”类评测任务、“笔画”类评测任务、“多音字”类评测任务、“部件”类评测任务下的正确率。S_1是式1中定义的AVG_SCORE指标，S_2是式2中定义的TOTAL_SCORE指标。从表3中前四类评测任务的平均得分来看，大模型对汉字偏旁掌握得最好，对笔画数量掌握得最差。在参数量为7B左右的大模型里，Qwen-7B模型的富语义认知能力最强，并且能力随着大模型参数量的增大而明显提升。GPT-3.5在富语义认知能力的评测中，表现也较为突出。

同时，为进一步探究大模型在六类富语义认知评测任务上的得分规律，本研究作出了更直观的雷达图，如图2所示。其中，图2.(a)是Qwen系列参数量为14B与参数量为7B的大模型的得分对比，从中可以看出Qwen-14B模型在各评测任务上的得分均高于Qwen-7B模型，二者还具有相似的得分规律；图2.(b)是表3中除Qwen-14B模型与GPT-3.5外，综合得分前四的大模型与GPT-3.5的得分表现对比，从中可以看出除Qwen-7B模型表现优异以外，其余大模型在各评测任务上的得分差异不大；图2.(c)是表3中除Qwen-14B与GPT-3.5外，综合得分后四的大模型与GPT-3.5的得分表现对比，可以看出综合得分后四的大模型在“笔画”类评测任务上的表现均较差。

[3] https://deepseek.com

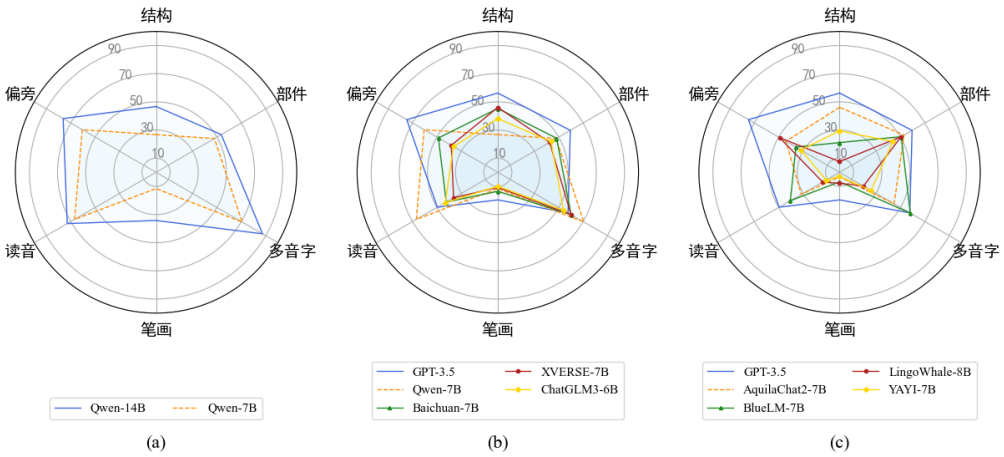

图 2: 不同大模型在富语义认知评测任务上的得分表现

3.1.1 评测结果分析

通过观察表3的数据可以发现，前四类评测任务中大模型对汉字偏旁的掌握程度最好，平均正确率达到了49.57%；对汉字笔画的掌握程度最差，平均正确率只有12.02%，且基本随着大模型整体富语义认知能力的降低而降低。这说明，目前大模型对偏旁部首已经有了较好的认知，而汉字笔画可以作为基座模型研发人员未来攻克的方向之一。

此外，大模型在“读音”类评测任务上的得分差异较为显著。根据实验过程中收集到的大模型回复来看，总结其主要原因如下：在“读音”类评测任务上表现较好的Qwen系列模型、ChatGLM3-6B模型基本能精准地回答出汉字对应的读音；表现中等的大模型则容易创造一些额外的、不存在的读音，导致它们的回答不够精准；表现欠佳的大模型，其部分回答未包含音调，因此被判定为回答错误。

同时，本研究还发现大模型在“多音字”类评测任务上的得分与在“读音”类评测任务上的得分具有一定相关性。在“读音”类评测任务上得分高的大模型，往往在“多音字”类评测任务上的得分也高。为证明此猜想，本研究对“读音”类评测任务得分和“多音字”类评测任务得分做了进一步的线性相关分析。计算结果表明，二者之间皮尔逊相关系数为0.95，皮尔逊相关系数的p值为2.11×10^{-5}远小于0.05，这说明大模型在“读音”类评测任务上的得分和在“多音字”类评测任务上的得分具有显著的相关性。由此可见，大模型对汉字发音的认知能力与其区分多音字发音的能力关联紧密，汉字不同的发音代表着不同的含义，能正确区分多音字的发音，对中文理解过程很重要。

3.1.2 模型分析

通过观察表3的数据可以发现，在所有参数量约为7B左右的大模型里，Qwen-7B模型表现出了最好的富语义认知能力。Qwen-7B模型在汉字拼音、偏旁等相关评测任务上的正确率均显著领先于同规模模型，尤其以70.63%的正确率在多音字注音任务中表现最为突出，虽然在“结构”类评测任务上的得分要略低于同水平的大模型，但它对汉字结构的掌握最充分、最全面。同样地，通过观察表3中Qwen-14B模型的得分规律，可以再次验证这一结论：Qwen系列模型是此次评测中富语义认知能力最强的大模型。

此外，通过观察图2(a)可以发现，Qwen-14B模型与Qwen-7B模型的雷达图形状相似度很高，这说明二者在富语义认知评测任务上的得分规律是相似的。同时，Qwen-14B在各类评测任务上的得分均高于Qwen-7B，这也说明了大模型的富语义认知能力会随着大模型参数量的增大而全方面提升。最后，根据图2(b)和图2(c)得知，GPT-3.5在富语义认知评测任务中同样表现优异，尤其是对汉字偏旁的认知能力十分突出。Qwen系列模型与GPT-3.5在富语义认知评测任务上的优异表现，可能源于其独特的模型架构设计或在对应领域更充分的训练数据覆盖。

3.1.3 英文评测任务结果与分析

为探究大模型在富语义认知评测任务上的中英文对齐能力，本研究将所有评测任务翻译成英文，并从2.2节的10个被测对象中，挑选了一个国外的大模型GPT-3.5，和两个国内的大模

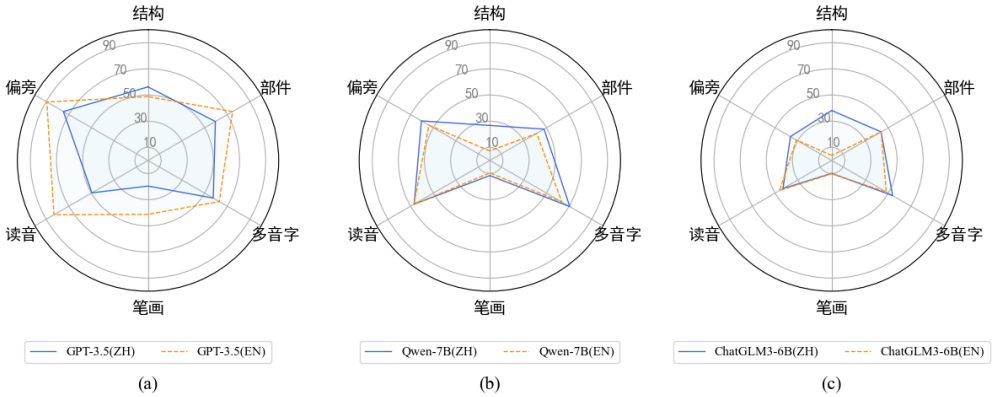

图 3: 不同大模型在富语义认知中英文评测任务上的得分表现对比

富语义认知中英文评测任务										“部件”问答评测任务		
模型	语言	p_1	p_2	p_3	p_4	p_5	p_6	S_1	S_2	*Precision*	*Recall*	*F*1
GPT-3.5	ZH	56.25	74.77	49.60	19.61	57.63	59.52	50.06	52.52	51.36	66.38	56.49
	EN	48.81↓	89.44↑	83.10↑	41.22↑	62.94↑	74.71↑	65.64↑	66.01↑			
Qwen-7B	ZH	26.74	60.49	66.89	11.58	70.63	47.89	41.42	47.91	17.98	24.13	19.98
	EN	7.26↓	53.91↓	67.04↑	9.61↓	63.63↓	41.38↓	34.46↓	40.98↓			
ChatGLM3-6B	ZH	38.21	36.47	43.09	9.92	53.85	43.66	31.92	37.48	15.79	15.12	14.77
	EN	3.86↓	31.22↓	46.13↑	10.19↑	48.25↓	43.68↑	22.85↓	30.01↓			

表 4: 大模型富语义认知能力评测

型Qwen-7B模型、ChatGLM3-6B模型为代表，在这三个大模型上做进一步的英文版富语义认知评测任务的评测，评测结果如表4所示，其中各项指标的含义与表3相同。

由表4可以得知，三个大模型在“结构”类评测任务上的得分均大幅减弱，尤其是Qwen-7B模型和ChatGLM3-6B模型的正确率降到了个位数。可见，英文版的汉字结构知识对于大模型更难掌握。GPT-3.5在英文版富语义认知评测任务上综合能力有很大程度的提升，其中主要体现在“读音”类评测任务上，正确率由原来的49.60%提高至83.10%。根据模型的回复来看，准确率大幅提升的主要原因是：GPT-3.5在以英文为语言的提问下，不会再生成额外的、不存在的读音。GPT-3.5在“偏旁”类评测任务上的准确率还达到了前所未见的89.44%，同时在大模型掌握得最薄弱的“笔画”类评测任务上，准确率也有着惊人的提升。此外，两个国内的大模型Qwen-7B模型和ChatGLM3-6B模型在英文版富语义认知评测上综合能力都存在着小幅度的减弱。从图3(b)、图3(c)可以看出，Qwen-7B模型和ChatGLM3-6B模型在以英文为语言的提问下，除“结构”类评测任务正确率明显降低以外，其余类评测任务正确率变化细微。

上述现象表明，提示文本的语言或许是影响大模型富语义认知能力发挥的关键因素之一。大模型表现的差异可能源于其预训练语料中不同语言数据的构成、比例以及模型架构对跨语言迁移的适应能力不同。具体而言，GPT-3.5的训练语料以英语为主，更擅长处理英文提示，中文提示可能未充分激发其内在潜力。该结论可为今后大模型评测工作中，评测数据的语言选择问题提供参考。

3.1.4 大模型汉字部件问答评测

由表3可以得知，大模型在给出选项的“部件”类评测任务上的正确率差异较小。为探究大模型对于汉字部件更真实的掌握程度，本节去除了“部件”类评测任务问题中提供的四个选项，直接用问答的方式再次评测GPT-3.5、Qwen-7B模型、ChagGLM3-6B模型。同时，本节在评价指标上也做了一些调整。由于一个汉字通常拥有多个部件，但大模型往往只能回答出其中的一部分，因此本节不再使用正确率作为评价指标，而是选择用精确率、召回率和F1值来综合考量大模型对汉字部件的认知能力。

本研究对大模型的每一条回答都做了精确率、召回率和F1值的计算，并最终得到平均精确率（*Precision*）、平均召回率（*Recall*）和平均F1值（*F*1），以此作为大模型在新“部件”类

模型	结构推理	偏旁推理	读音推理	Avg
GPT-3.5	31.71	41.48	0.31	24.50
Qwen-7B	39.98	35.27	39.38	38.21
ChatGLM3-6B	48.04	78.13	54.05	60.07
GPT-4o	98.52	86.8	61.13	82.15
DeepSeek V3	95.97	83.84	81.54	87.12
Qwen2.5-72B	98.62	85	93.85	92.49

表 5: 汉字推理能力的评测结果

评测任务上的得分，评测结果如表4所示。由表4可以得知，结果和给定选项的部件评测任务不同：GPT-3.5对于汉字部件的掌握程度明显优于Qwen-7B模型和ChatGLM3-6B模型，其中平均F1值达到了56.49%。Qwen-7B模型和ChatGLM3-6B模型在提问式的“部件”类评测任务下的表现仍有待加强。汉字的组成部件与汉字的含义有着密不可分的联系，正确掌握汉字部件，对正确理解汉字与中文有着非常积极的帮助。因此，汉字部件也应该受到更多中文大模型研究人员的关注。

3.2 富语义推理能力分析

为进一步探究大模型对汉字信息的推理能力，本研究进一步设计了“结构推理”“偏旁推理”“读音推理”三类富语义推理评测任务，对表现较好的GPT-3.5、Qwen-7B模型、ChatGLM3-6B模型再次进行评测。由于推理能力强依赖于大模型的参数量，参数量在7B左右的大模型较难拥有强大的推理能力，因此本研究还利用OpenRouter[4]提供的API，引入了三个大参数量且近期在各类任务上都表现出良好推理能力的大模型（GPT-4o、DeepSeek V3、Qwen2.5-72B模型）共同参与此次评测。

评测依旧选用正确率作为评价指标，评测结果如表5所示，其中“Avg”表示三类评测任务正确率的平均值。由表5可知，在汉字富语义认知评测中表现中等的ChatGLM3-6B模型，竟表现出了惊人的富语义推理能力，在“偏旁推理”和“读音推理”两类评测任务上的表现均远高于GPT-3.5和Qwen-7B模型，且平均推理正确率达到60.07%，综合推理能力接近参数量更大的大模型。可见，ChatGLM3-6B模型对富语义信息有着较强的推理能力，在回答汉字的富语义信息时，不只依赖于模型的“记忆”。而在富语义认知评测中表现优异的GPT-3.5，在“读音推理”类评测任务上的正确率只有0.31%。同时，GPT-4o在“读音推理”类评测任务上的推理能力，也远不如其在“结构推理”和“偏旁推理”类评测任务上的推理能力。结合3.1.3中的结论来看，GPT系列模型在汉字读音评测任务上的低正确率，可能和评测任务的语言相关。此外，在三个大参数量的大模型中，Qwen2.5-72B模型的平均推理正确率达到了92.49%。

在富语义推理评测任务中，ChatGLM3-6B模型与Qwen系列模型均展现出显著优势。这主要得益于二者均为国内开发的大模型，在训练阶段接触了海量且丰富的中文语料资源。丰富的语料为大模型提供了充足的训练样本，使其能更有效地掌握汉字的组成规则与发音规律，从而在富语义推理评测任务中表现更出色。最后，Qwen系列模型在本研究的两个阶段评测中，不仅表现出了最强的富语义认知能力，还表现出了优异的富语义推理能力，Qwen系列模型无疑是本研究评测中富语义能力最强的大模型。

4 总结

本研究通过构建多维度汉字富语义评测体系，系统评估了当前主流大语言模型的汉字认知与推理能力。实验结果表明，现有模型在偏旁识别等基础语义特征理解上表现较好，但在笔画识别、特殊结构认知等深层次语义理解方面仍存在明显不足。值得注意的是，不同模型在读音认知任务上表现出显著差异，且部分模型存在系统性误读现象。跨语言实验进一步揭示了提示语言对模型表现的重要影响。在推理能力方面，部分模型展现出较强的语义推理能力，其中Qwen系列模型在认知与推理任务中均表现优异。这些发现不仅为中文自然语言处理领域的模

[4]https://openrouter.ai/docs

型选择与优化提供了实证依据，也为深入理解大语言模型的汉字处理机制与认知边界提供了重要参考，对推动中文信息处理技术的发展具有积极意义。

致谢

感谢所有匿名审稿人的宝贵意见。本研究成果受中国博士后科学基金（2023M731253），湖北省自然科学基金（2023AFB487），国家语委项目“十四五”科研规划项目（YB145-128），和湖北高等教育学会重点课题“数字化转型与高校教学模式创新研究”（2023ZA018）资助。

参考文献

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*.

Yang Chi, Fausto Giunchiglia, Chuntao Li, and Hao Xu. 2024. Ancient chinese glyph identification powered by radical semantics. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12065–12074. Association for Computational Linguistics.

Ming Dong, Yujing Chen, Miao Zhang, Hao Sun, and Tingting He. 2024. Rich semantic knowledge enhanced large language models for few-shot chinese spell checking. In *ACL (Findings)*, pages 7372–7383. Association for Computational Linguistics.

Li Huang, Junjie Li, Weiwei Jiang, Zhiyu Zhang, Minchuan Chen, Shaojun Wang, and Jing Xiao. 2021. Phmospell: Phonological and morphological knowledge guided chinese spelling check. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5958–5967. 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 62991–63010. Curran Associates, Inc.

Yitian Liu and Zhouhui Lian. 2023. Fonttransformer: Few-shot high-resolution chinese glyph image synthesis via stacked transformers. *Pattern Recognit.*, 141:109593.

Renze Liu, Hongtao Zhou, and Housheng Su. 2024. Using glyph lexicon enhancing bert character representation for chinese named entity recognition. In *International Conference on Guidance, Navigation and Control*, pages 527–538. Springer.

Chao Lv, Han Zhang, Xinkai Du, Yunhao Zhang, Ying Huang, Wenhao Li, Jia Han, and Shanshan Gu. 2022. Stylebert: Chinese pretraining by font style information. *CoRR*, abs/2202.09955.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075, Online, August. Association for Computational Linguistics.

Yuting Wei, Yuanxing Xu, Xinru Wei, Simin Yang, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. AC-EVAL: evaluating ancient chinese language understanding in large language models. In *EMNLP (Findings)*, pages 1600–1617. Association for Computational Linguistics.

Shilian Wu, Yongrui Li, and Zengfu Wang. 2024. Chinese text recognition enhanced by glyph and character semantic information. *Int. J. Document Anal. Recognit.*, 27(1):45–56.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728, Online, August. Association for Computational Linguistics.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *CoRR*, abs/2307.15020.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Ye Yao, Chen Wang, Hui Wang, Ke Wang, Yizhi Ren, and Weizhi Meng. 2024. Embedding secret message in chinese characters via glyph perturbation and style transfer. *IEEE Trans. Inf. Forensics Secur.*, 19:4406–4419.

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*.

Xiaotian Zhang, Yanjun Zheng, Hang Yan, and Xipeng Qiu. 2023. Investigating glyph-phonetic information for chinese spell checking: What works and what's next? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1–13. Association for Computational Linguistics.

严旭枫. 2024. 基于深度学习的中文近音字和形近字拼写纠错方法研究. Master's thesis, 哈尔滨工程大学.

刘宇瀚, 刘常健, 徐睿峰, 骆旺达, 陈奕, 吉忠晟, and 应能涛. 2020. 结合字形特征与迭代学习的金融领域命名实体识别. 中文信息学报, 34(11):74–83.

罗文. 2024. 大语言模型评测综述. 中文信息学报, 38(1):1–23.

薛莫白. 2024. 基于偏旁部首建模的汉字生成方法研究与应用. Ph.D. thesis, 中国科学技术大学.