大语言模型的多语言词义相关度评估与可解释性分析

(申请清华大学文学博士学位论文)

培 养 单 位： 人文学院

学　　科： 中国语言文学

研 究 生： 刘　　柱

指 导 教 师： 刘　　颖　教　授

二〇二六年五月

Evaluation and Interpretability Analysis of Large Language Models for Multilingual Word-in-Context Relatedness

Dissertation submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Chinese Language and Literature

by

Liu Zhu

Dissertation Supervisor: Professor Liu Ying

May, 2026

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

李素建	研究员	北京大学
赵军	教授	中国科学院自动化所

答辩委员会名单

主席	赵军	教授	中国科学院自动化所
委员	常宝宝	教授	北京大学
	江铭虎	教授	清华大学
	张赫	教授	清华大学
	邱冰	副教授	清华大学
	刘颖	教授	清华大学
秘书	易佳	助理研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位法》及上级教育主管部门要求，报送相应的学位论文。

本人保证遵守上述规定。

作者签名：刘柱　　导师签名：刘颖

日　　期：2026.05.28　　日　　期：2026.05.28

摘 要

词义相关度任务旨在判断目标词在不同上下文中的语义关联程度，是词义理解领域的一项重要任务。尽管以大语言模型为代表的神经网络语言模型已在各类自然语言处理应用中展现出卓越的性能，但其对真实语料和多语言环境中词义的理解与建模仍缺乏系统、全面的评估。为弥补这一不足，本文围绕多语言词义相关度任务，从模型评估、可解释性分析以及标注不一致三个维度展开系统研究。

在模型评估方面，本文收集并拓展了已有的多语言多标注者数据集，设计了多样化的评估手段，并对多种架构类型的模型进行了全面对比。针对汉语兼类词，本文专门构建了汉语词义相关度数据集。评估手段包括三类：一是几何探测；二是为开源大语言模型设计引导词义表征的提示语；三是为闭源模型设计外部提示语。所评估的模型涵盖两种经典架构。编码器架构包括 BERT、RoBERTa 等模型；解码器架构则包括多个参数规模的 LLaMA、DeepSeek、通义千问等大语言模型。通过定量与定性分析，本文发现：尽管大语言模型擅长生成类任务，且未在词义相关度数据集上直接训练，但其词义理解能力与在词义相关度数据集上微调后的编码器模型相当；不过，其表现与专家标注结果仍存在一定差距。此外，英语的词类信息在词义相关度任务中发挥的作用相较汉语更为显著。

在可解释性分析方面，本文深入模型的内部表征，利用逐层探测手段，分析大语言模型在跨层信息流动上与传统编码器模型的差异。多语言数据集与多个模型上的结果表明，大语言模型模型的表征呈现“先理解，后预测”的层次模式，而编码器模型则随层数加深，持续提升对词义的理解能力。这一创新性发现为理解模型机制提供了有力支撑。

在标注不一致建模方面，本文设计了分类探测器、模型集成和提示语引导等方法，系统评估各类语言模型对标注不确定性的捕捉能力。多语言数据集的定量与定性结果表明，在标注不一致预测任务上，模型集成方法表现最佳，但与专家标注结果之间仍存在显著差距，且模型更容易高估不一致程度。多标注者视角为评估模型对真实数据中词义的理解能力以及探索模型鲁棒性提供了重要的分析框架。

综上所述，本文从多语言视角和多标注者视角出发，借助多语言词义相关度任务，对大语言模型进行了全面、系统的评估，为模型的不确定性建模与可解释性分析提供了有力支撑，从而为安全、可靠地使用大语言模型奠定了研究基础。

关键词：大语言模型；模型评估；模型可解释性；词义相关度；标注不一致

Abstract

The word-in-context relatedness task aims to judge the semantic relevance of a target word in different contexts, and is a crucial task in the field of lexical semantics. Although neural network language models represented by large language models (LLMs) have demonstrated outstanding performance in various natural language processing applications, their ability to understand and model word meanings in real-world and multilingual corpora still lacks systematic and comprehensive evaluation. To address this gap, we conduct on the multilingual word-in-context relatedness task from three dimensions: model evaluation, interpretability analysis, and annotation disagreement.

Regarding model evaluation, we collect and expand existing multilingual multi-annotator datasets to conduct a comprehensive comparison of models with various architectures by utilizing various evaluation methods. Specifically, a Chinese word-in-context relatedness dataset is constructed for Chinese noun-verb cross-category words, which serve as noun and verb in different contexts. The evaluation methods fall into three categories: geometric probing; prompt design for open-source LLMs to guide representation on word meaning; and external prompt design for closed-source models. The evaluated models cover two classic architectures: encoder models including BERT, RoBERTa among others; decoder models covering LLMs of various parameter scales such as LLaMA, DeepSeek, and Qwen. Through quantitative and qualitative analyses, we find that although LLMs excel at generative tasks and are not directly trained on word-in-context relatedness datasets, their ability of word meaning comprehension is comparable to that of encoder models fine-tuned on such datasets. However, their performance still lags behind expert annotation to a certain extent. In addition, parts of speech in English play a more significant role in the word-in-context relatedness task than that in Chinese.

As for interpretability analysis, we delve into the internal representations of models and employ layer-wise probing techniques to analyze the differences in cross-layer information flow between LLMs and traditional encoder models. Results on multilingual datasets and multiple models show that the representations of decoder models exhibit a layer-specific pattern of “understanding in the first layers, prediction in the later layers”, whereas encoder models continuously enhance their word meaning understanding as layers go deeper. This finding provides strong support for understanding model mechanisms.

In terms of annotation disagreement modeling, we design multiple methods including classifier-based probing, model ensembling, and prompt guidance to systematically evaluate the ability of various language models to capture annotation uncertainty. Quantitative and qualitative results on multilingual datasets indicate that the method of model ensembling achieves the best performance on the annotation disagreement prediction task, but still presents a significant gap from expert annotations. Moreover, models tend to overestimate the degree of disagreement. The multi-annotator perspective provides an important analytical framework for evaluating models' ability to understand lexical semantics and analyzing model robustness in real-world data.

In summary, from the multilingual and multi-annotator perspectives, we comprehensively and systematically evaluate language models' performance by leveraging the task of word-in-context relatedness. It provides strong support for robust modeling and interpretability analysis of models, laying a research foundation for the safe and reliable application of LLMs.

Keywords: Large Language Models; Model Evaluation; Model Interpretability; Word in Context; Annotation Disagreement

目 录

摘 要.....I
Abstract.....II
目 录.....IV
插图清单.....VIII
附表清单.....X
缩略语说明.....XII
第 1 章 绪论 1
1.1 研究背景和意义 1
1.2 研究问题 3
1.2.1 多语言视角下的词义相关度评估任务 3
1.2.2 基于词义相关度任务的模型可解释性分析 5
1.2.3 多标注者视角下的词义相关度评估任务 6
1.2.4 研究问题间的关联 6
1.3 研究内容和创新点 7
1.3.1 研究思路 7
1.3.2 具体研究内容 8
1.3.3 本文创新点 9
1.4 组织结构 12
1.5 本章小结 14
第 2 章 研究现状和相关工作 15
2.1 词义相关度任务的相关工作 15
2.1.1 不同的任务类型 15
2.1.2 多语言视角下的词义相关度任务 17
2.1.3 多标注者视角下的词义相关度任务 18
2.2 大语言模型的评估和可解释性研究 20
2.2.1 大语言模型的概述 20
2.2.2 大语言模型的评估研究 24
2.2.3 大语言模型的可解释性研究 27
2.3 本章小结 28

第 3 章 词义相关度任务和数据集概述 29
3.1 词义相关度任务定义 29
3.1.1 词义相关度任务 29
3.1.2 形式化定义 30
3.1.3 与词义消歧任务的关联 30
3.2 词义相关度数据集概述 32
3.2.1 WiC 32
3.2.2 OGWiC 34
3.2.3 DisWiC 40
3.2.4 OGWiC_NV 43
3.2.5 DisWiC_NV 50
3.2.6 数据集选取原则 50
3.3 本章小结 51
第 4 章 多语言词义相关度任务评估 52
4.1 本章导论 52
4.2 基于模型内部表征的评估方法 52
4.2.1 表征提取方式 52
4.2.2 几何探测方法 54
4.2.3 向量后处理 57
4.2.4 基于提示语的探测方法 59
4.3 词义相关度实验 61
4.3.1 任务定义和数据集 61
4.3.2 评估模型 62
4.3.3 超参数设置 65
4.4 实验结果和分析 65
4.4.1 准确性分析 66
4.4.2 向量后处理分析 70
4.4.3 零样本指令分析 75
4.4.4 多语言模型和单语模型对比分析 76
4.4.5 词类信息对于汉语和英语的影响对比分析 78
4.4.6 典型偏误分析 79
4.5 本章小结 89

第 5 章 多语言词义相关度任务中的模型可解释性研究 90
5.1 本章导论 90
5.2 不同架构的语言模型对比分析 90
5.2.1 编码器模型 90
5.2.2 解码器模型 92
5.2.3 编码架构和解码架构的对比 93
5.3 可解释性研究实验设计 93
5.3.1 任务定义和数据集 93
5.3.2 评估模型 94
5.3.3 表征选择和探测方式 95
5.4 英语通用数据集 WiC 上的实验结果和分析 97
5.4.1 WiC 上的准确性结果分析 97
5.4.2 模型跨层性能的模式分析 100
5.4.3 其他消融实验分析 105
5.5 多语相关度数据集 OGWiC 上的实验结果和分析 106
5.6 本章小结 109
第 6 章 多语言词义相关度任务中的标注不一致评估 111
6.1 本章导论 111
6.2 词义相关度标注不一致预测方法 111
6.2.1 基于分类器的探测方法 112
6.2.2 基于模型集成的预测方法 113
6.2.3 基于提示语的探测方法 116
6.3 词义标注不一致任务实验设定 119
6.3.1 任务定义和数据集 119
6.3.2 评估模型 119
6.3.3 集成方法和超参数设置 119
6.3.4 分类器探测方法的训练设置 122
6.3.5 提示语探测的超参数设置 123
6.3.6 基线模型 123
6.4 实验结果和分析 124
6.4.1 标注不一致相关性结果分析 125
6.4.2 集成方法中不同集成方式和统计量分析 127
6.4.3 集成方法中不同扰动策略分析 130

6.4.4 不同语言集成后的表现分析 133
6.4.5 典型偏误分析 134
6.5 本章小结 142
第 7 章 总结和展望 143
7.1 论文主要工作和贡献 143
7.2 下一步研究工作展望 143
7.3 本章小结 144
参考文献 146
附录 A 中文兼类词数据集标注问卷展示 163
致 谢 165
声 明 167
个人简历、在学期间完成的相关学术成果 168
指导教师评语 169
答辩委员会决议书 170

插图清单

图 1.1　概念空间和语义地图示例 4
图 1.2　研究路线图 7
图 1.3　研究汇总和文章的组织结构 12
图 2.1　三类 Transformer 模型架构示意图 23
图 3.1　OGWiC 数据集中不同语言的标注值分布图 38
图 3.2　OGWiC 中不同数据集的标注值分布图 39
图 3.3　DisWiC 数据集中每种语言的 MPD 分布直方图 42
图 3.4　DisWiC 中不同数据集的 MPD 分布直方图 43
图 3.5　中文词义相关度标注指南 45
图 3.6　标注者背景信息统计 46
图 3.7　标注者年龄段分布 46
图 3.8　标注者家乡省份分布 47
图 3.9　OGWiC_NV 不同的数据集的标注值分布图 49
图 3.10　DisWiC_NV 不同数据集的不一致评分分布 51
图 4.1　编码器模型和解码器模型的架构对比图 53
图 4.2　几何探测示意图 57
图 4.3　表征分布示意图 57
图 4.4　标签分布对比图 68
图 4.5　各类去各向异性手段的平均效果 71
图 4.6　中文数据集上各类去各向异性手段的效果 71
图 4.7　英文数据集上各类去各向异性手段的效果 72
图 4.8　德语数据集上各类去各向异性手段的效果 72
图 4.9　挪威语数据集上各类去各向异性手段的效果 73
图 4.10　俄语数据集上各类去各向异性手段的效果 73
图 4.11　西班牙语数据集上各类去各向异性手段的效果 74
图 4.12　瑞典语数据集上各类去各向异性手段的效果 74
图 4.13　单语模型相对于多语模型的结果变化 76
图 4.14　提示语中的词类信息对于闭源大模型的影响 78
图 5.1　编码器模型和解码器模型的架构对比图 91
图 5.2　逐层几何探测流程图 97

图 5.3　两类模型的理解与生成能力逐层变化 100
图 5.4　编码器模型随层数增加的性能变化 101
图 5.5　编码器模型针对前一目标词表征性能随层数的变化 101
图 5.6　解码器模型随层数增加的性能变化 102
图 5.7　解码器模型针对前一目标词表征性能随层数的变化 103
图 5.8　各类提示语对于解码器模型的逐层性能影响 104
图 5.9　不同模型各层之间最佳阈值参数的变化 105
图 5.10　去除各向异性手段对于各类模型的结果提升..... 106
图 5.11　OGWiC 数据集在不同语言与不同架构模型下的逐层准确率变化 107
图 6.1　分类器探测示意图 113
图 6.2　两种方式的模型集成方式对比图 116
图 6.3　DisWiC 数据集上集成指标的实例排序相关性..... 129
图 6.4　DisWiC_NV 数据集上集成指标的实例排序相关性 130
图 6.5　DisWiC 中集成策略下的相关性分布..... 131
图 6.6　DisWiC_NV 在不同集成策略下集成实例相关性的分布 132
图 6.7　不同语言集成实例相关性的分布 133
图 6.8　平均标注相关度与模型预测的不一致程度之间的相关性 136

附表清单

表 2.1 常见词义理解任务的对比 17
表 2.2 两类探测方法的对比 28
表 3.1 WiC 数据集中样例示例 33
表 3.2 WiC 数据集不同划分的统计信息 33
表 3.3 不同词义相关度标注类型对比 35
表 3.4 OGWiC 数据集中英例示 36
表 3.5 OGWiC 数据集的统计结果 37
表 3.6 DisWiC 数据集中英例示 40
表 3.7 DisWiC 数据集的统计结果 41
表 3.8 中文兼类词数据集 OGWiC_NV 中的示例 48
表 3.9 OGWiC_NV 中数据集的统计结果 48
表 3.10 DisWiC_NV 不同数据集的统计结果 50
表 4.1 表征提取策略以及示例 53
表 4.2 各类提示语方法对应的模板和备注 55
表 4.3 各语言对应的 BERT 模型及其来源 62
表 4.4 各类模型在不同语言上的准确性 66
表 4.5 通义千问模型与 DeepSeek 模型的混淆矩阵 69
表 4.6 各类提示语在不同语言上的准确性 75
表 4.7 模型出错案例数量统计 80
表 4.8 OGWiC_NV 数据集中的语义转化类型 85
表 5.1 不同模型架构的特征对比 93
表 5.2 两个数据集上探测的不同类型的模型 94
表 5.3 各类模型各层的起止索引 96
表 5.4 WiC 测试集在各个方法上的准确率 98
表 6.1 向量融合的 4 种方式 112
表 6.2 评估模型输出分歧的统计量 115
表 6.3 模型类型及对应的扰动变量 119
表 6.4 模型扰动因素组合及扰动次数 121
表 6.5 各类方法在不同语言上的不一致程度预测结果 125
表 6.6 不同集成方式和指标下各类数据集的表现 127

表 6.7　不同集成方式和统计量下的最优组合 …… 128
表 6.8　不同提示语下模型对词义不一致判断的错误统计 …… 135

缩略语说明

LLMs	大语言模型（Large Language Models）
AI	人工智能（Artificial Intelligence）
WiC	词义相关度（Word in Context）
WSD	词义消歧（Word Sense Disambiguation）
TSV	目标词义确认（Target Sense Verification）
NER	命名实体识别（Named Entity Recognition）
UE	不确定性估计（Uncertainty Estimation）
SVM	支持向量机（Support Vector Machines）
RNN	循环神经网络（Recurrent Neural Networks）
LSTM	长短期记忆网络（Long Short-Term Memory）
BERT	基于 Transformer 的双向编码器表示（Bidirectional Encoder Representations from Transformers）
CR	指代消解（Coreference Resolution）
GPT	生成式预训练 Transformer（Generative Pre-trained Transformer）
NTP	下一个词预测（Next Token Prediction）
MLM	掩码语言模型（Masked Language Model）
T5	文本到文本迁移 Transformer（Text-to-Text Transfer Transformer）
BART	双向自回归 Transformer（Bidirectional and Auto-Regressive Transformer）

第 1 章　绪论

1.1　研究背景和意义

以大语言模型（Large Language Models, LLMs）为核心的人工智能（Artificial Intelligence, AI）技术正受到当前时代前所未有的关注。该技术在多个领域均展现出显著成效：在语言理解领域，AI 能够实现多语言乃至方言之间的高效翻译[1-2]；在推理领域，AI 聊天机器人可根据用户反馈精准推断其意图，从而提升沟通效率[3-4]；在教育领域，AI 可制定个性化学习方案，助力学生高效学习[5-6]；在经济领域，AI 推动各行业的智能化转型，应用场景涵盖智能家居、智能教室、智能驾驶及智慧物流等。国家层面亦将“深入推进数字中国建设，提升数智化发展水平”正式写入《中华人民共和国国民经济和社会发展第十五个五年规划纲要》①。

然而，大语言模型在应用层面仍面临诸多挑战。首先，基于数据驱动的大语言模型内部机制尚不透明，易产生包含错误或误导信息的“幻觉”（hallucination）现象②，进而引发安全性问题[8]。其次，基于“大数据、多参数、强算力”范式训练的大语言模型仍存在效率低下的问题[9]。这类模型完全依赖海量资源的训练，其间缺乏人类知识的有效介入，导致学习过程资源消耗巨大。这一模式与儿童语言习得的低能耗特点，以及外语学习者融合语言学知识的学习方式形成鲜明对比。最后，尽管大语言模型在下游任务中表现优异，但在细粒度的语言学任务评估以及面对真实场景中的不完美数据时，相关研究和应对策略仍显不足。

为此，仍需对包括大语言模型在内的神经网络语言模型进行多维度、细粒度、全方位的评估与理解。从早期偏重于文本理解的评测平台，如通用语言理解评测基准 GLUE（General Language Understanding Evaluation）[10]，到更加通用、全面、重视推理能力的评测基准，如 MMLU（Massive Multitask Language Understanding）[11]，语言模型的评测体系日益精细化，涵盖多语言理解、文本推理、对话生成、安全性等多个子领域[12]。评测的深入也推动了对模型内部机制的理解，通过可视化分析、归因分析、干预手段、探测技术等方法，研究者对模型内部结构、信息流动过程、神经元及回路（circuit）③等进行深入分析，从而增进对模型工作机制的认

① 参见 2026 年 3 月 13 日新华网报道：《中华人民共和国国民经济和社会发展第十五个五年规划纲要》。网址：https://www.news.cn/politics/20260313/085af5de5a4b4268aa7d87d90817df2f/c.html

② “幻觉”一词最早借用于心理学，后被计算机视觉领域引入，用以描述向原始图像中添加细节的过程[7]。随后，该术语逐渐被统计机器翻译、语义角色标注等自然语言处理任务所采用，用以指代失败的模式识别。在大语言模型兴起后，“幻觉”特指模型生成表面合理但事实错误的内容，即所谓“一本正经地胡说八道”。

③ “回路”（circuit）一词源于电路工程，后被大语言模型“机械可解释性”（mechanistic interpretability）研究领域引入[13]，用以指代对模型进行逆向工程所识别出的功能性神经网络子结构。

识[14]。

另一方面，评估往往不能止步于评测基准上的任务。一方面，评测数据具有时效性，在数据集公布一段时间后，原有数据可能被纳入模型的训练语料，从而导致这些公开数据集的有效性降低。另一方面，这些评测任务通常基于人工精心标注的数据构建，往往过滤了真实世界中普遍存在的不完美实例，例如标注不一致、信息不充分的文本、错误标注等噪声数据。模型在真实样本中的表现及其相关能力的预测，不仅关乎模型的实际部署性能，也与模型的鲁棒性密切相关。

用于评估模型的自然语言处理任务与语言学密切相关，在不同层级的语言学研究单位中均有对应的任务。例如，在词级别上有词义消歧、词义相关度判别、分词、隐喻识别等任务；在合成词或短语级别上有惯用语识别、合成词内部结构关系识别等任务；在句子级别上有自然语言推断（natural language inference）、句义相似性判断（semantic textual similarity）、歧义检测等任务；在篇章级别上则有文本摘要、人物社交网络分析等任务。将语言学知识引入评估设计，有助于使评估过程更加规范化、科学化。例如，Li 等提出的歧义类型学分类体系①对句子歧义理解评估具有重要的指导作用[15]。

词义理解评估是最基础但最重要的评估任务之一。一方面，词元（token）——句子经过分词操作后的结果——作为现代语言模型架构的最小信息处理单元[16]，直接参与模型的运算。因此，对词元的理解直接关涉到更大语言单元的理解，如短语级、句子级的语义表示。另一方面，词作为自然语言中可以独立运用的最小音义结合体[17]，也是构成更高层级语言结构的基础单元。自然语言的句法和语义具有组合性特点②，因此，只有对词义进行全面、科学的评估，才可能进一步实现对其他语言层级的有效理解③。

词义相关度任务旨在判断出现在不同上下文的同一目标词的词义关联程度，是词义理解评估中重要的一项具体任务。其他的众多词义有关的任务都和词义相关度任务息息相关。同时该任务仅需对比成对上下文中词义的关联程度，无需额外借助外部由候选词义组成的词典，组织形式简单、直接。因此本文使用词义相关度任务来作为词义理解评估的主要形式。

词义相关度任务同样需要在真实样本中，从多维度视角进行评估。

① 该研究系统梳理了自然语言处理中歧义的类别划分，为句子层面的歧义理解评估提供了理论基础。

② 在句法结构树或句义解析树中，最小的意义单元通常是词。根据弗雷格原则（Frege's principle），复合表达式的意义是其组成部分的意义及其句法组合方式的函数[18]，即“The meaning of a compound expression is a function of the meanings of its parts and of the way they are syntactically combined”。

③ 需要说明的是，语言模型处理的“词元”并不完全等同于语言学意义上的词或语素。现代大语言模型普遍采用字节对编码算法进行子词切分[19]，该算法根据语料库中的统计频率迭代合并最常共现的字符或子词序列，生成最终的词元表。这种基于统计而非语言规则的分词方式，可能导致词元边界与语言学的词素边界产生偏离。为便于论述，本文在多数情况下将“词元”与“词”视为等同表述。

（1）多语言视角。作为词义的承载单元，词在不同语言中呈现不同的形态①，而词形对词义理解的作用在不同语言中也不尽相同。例如，英语的形态曲折变化可以反映词的语法类别，进而影响词义的理解；而汉语则缺乏显性的词形变化。因此，多语言对比为理解词义提供了重要的跨语言视角。

（2）多粒度视角。从词义候选数量的角度来看，词义理解通常涉及多义词，其义项数量少则数条，多则数十条。因此，词义判断本质上是一个多类别、多粒度的选择问题，需要模型具备区分细粒度语义差异的能力。

（3）多标注者视角。从词义之间的关联来看，不同义项之间往往存在模糊边界，这一特点在真实标注中常体现为标注者之间的不一致。全面评估模型对多标注者语义判断的建模能力，是衡量模型不确定性的重要维度。

模型对词义的评估和对模型的理解体现了语言学问题和计算模型之间的双向关联。一方面，利用语言学中对词义的相关知识设计针对模型的评估和理解任务，不仅使得模型评估更加精细、全面、客观，还可以对模型的结果做出解释。另一方面，模型的评估结果可以帮助完成语言学相关的任务，有利于此类任务的高效自动化②。

1.2　研究问题

针对上述背景，本节论述本文重点关注的三个相关研究问题。即从多语言和多标注者的视角设计语义相关度评估任务，以及基于词义相关度任务提供模型的可解释性分析。

1.2.1　多语言视角下的词义相关度评估任务

不同语言系统③尽管使用不同的词，但它们仍具有诸多共性。首先，“多义词”，即具有多种词义的词，这一概念普遍存在于各个语言中[21]。其出现反映了言语社区成员高效表达的需求，表现为多义词所携带的语义之间往往具有关联性，因此使用者可以通过词义联想的方式掌握其他义项[22]。同时，多义词能够有效减少语言系统中的符号数量，在一定程度上减轻学习者的记忆负担[21]。其次，不同语言系统中多义词所拥有的语义之间往往呈现出相似的关联类型，例如隐喻机制[23-24]、

① 即语言的任意性。赵元任在《语言问题》[20]中曾讲述一个改编自德国的故事：一位老太太初次接触外语，觉得外语“怪得毫无道理”，她说：“这明明是水，可英国人偏偏儿要叫他‘窝头’（water），法国人偏偏儿要叫它‘滴漏’（de l’eau），只有咱们中国人好好儿的管它叫‘水’！”赵元任用这个故事说明语言符号与其所指事物之间的关系具有任意性，并无必然联系。

② 前者可以概括为 Linguistics for AI，后者可以概括为 AI for Linguistics。

③ 将语言视作“系统”，是结构主义学派所倡导的，用于突出语言各个部分之间的关联性和整体性。

词义演变路径[25]、共词化所形成的概念空间①模式[28]等，这些共性反映了人类概念认知的普遍性。最后，词义总是在特定的语境中得到消解，其中语境不仅包括文本上下文，还包括话语情景、世界知识、背景信息、文化要素等。当语境不充分时，词义容易产生歧义，造成不同的理解[29]。

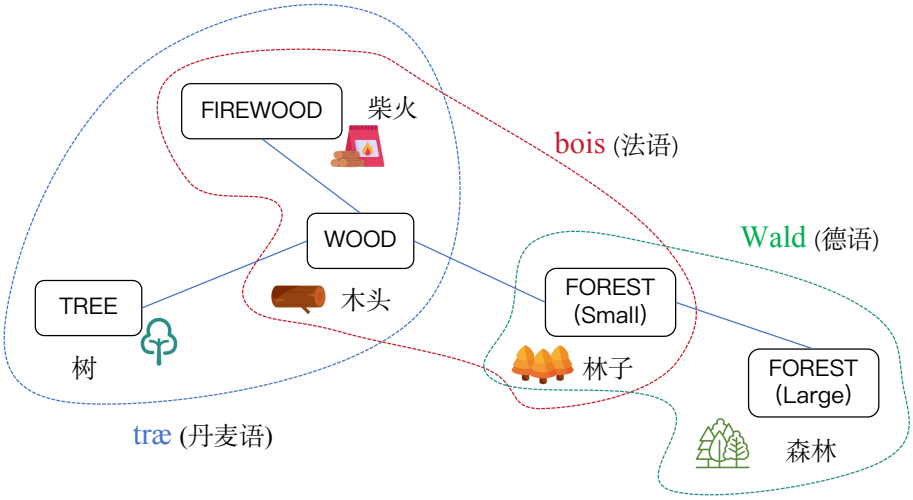


图 1.1　概念空间和语义地图示例

然而，语言之间的词义理解又具有差异性。第一，由于词义理解的对象是语言中的词单元，不同语言对词的表示存在差异。以汉语和印欧语言对比，汉语是一个形态变化匮乏的语言，在表达一些语法意义时词形无需进行曲折变化，也因此汉语的词类无法通过词形变换有效判断出来，此外汉语存在大量的复合词，对词义的理解往往依赖对词内部语素以及它们构词关系的理解②。与之相反，印欧语中的词使用丰富的屈折变化来表达不同的语法意义，它们也是区分词类的重要标志。尽管也存在一定数量的合成词，附加构词仍是印欧语常见的构词类型。这些形式方面的差异都会影响对于词义的理解。第二，即便语言之间具有类似的语义关联类型，但涉及到具体的语义关联它们又有所不同。典型的例子反映在共词化形成的概念空间中，不同的语言对于不同的概念进行共词化，形成各自的“语义地图”。例如图 1.1 绘制了与“树木”相关的语义空间[28]，图中的结点反映了语义概念，使用大写的英文表示概念，并附有相应的中文和图片信息。图中三种语言的相关词表达所能反映的语义使用图中连续的区域来表示。可以清晰地看出，不同语言在表达不同概念时，所使用相同词的情况不尽相同，进一步表明各语言内部存在的“形式——语义”关系有所差异。

① “共词化”概念指使用同一词形表达不同语义概念的现象，由学者 François 最早系统引入语言类型学研究[26]。例如在汉语中分开表达“打招呼”和“告别”概念的“你好”和“再见”，在意大利语中仅使用“ciao”即可表达这两个概念，即这两个概念“共词化”到同一词形中。“概念空间”则起源于语言类型学中的“语义地图”理论[27]，通过网络形式表现概念之间因共词化而形成的关联结构。

② 例如在文献[30]中，构词方式的引入可以有效提高词义歧义消解的准确性。

第三，不同语言的语境对于词义理解的互动方式不同。不同的语言对于上下文的不同成分具有不同的依赖程度。以英语为例，其动词拥有丰富的时态变化，动作发生的时间信息主要内嵌于词形本身（如“ate”明确指向过去），因此对上下文时间状语的依赖性较弱；而汉语动词缺乏时态标记，动作的时间指涉必须依靠上下文提供的时间状语（如“昨天”、“明天”）来推断。这种一方依赖词内形态、一方依赖词外语境的互动模式，深刻影响着词义的理解。另一方面，代表语言背后的文化背景也不尽相同。例如汉语词“龙”的理解和对应英文的 dragon 就因为文化背景而有所偏差，前者代表“高贵和神圣”，后者则代表“黑暗和邪恶”[31]。

正由于不同语言之间的共性和差异，有必要在多语言视角下对不同的神经网络语言模型在词义理解方面进行全面评估。其中重点关注以下的研究问题：对于不同语言的词义理解，模型掌握得程度如何？不同语言的语言特性如何影响模型对词义的理解？

1.2.2　基于词义相关度任务的模型可解释性分析

词义相关度任务亦可反向促进对模型的深入理解。该任务可视为探测模型行为的一种手段，其表现可用于判断模型是否具备词义理解相关的知识。进一步地，模型可解释性研究将这种探测深入到模型内部，典型的探测对象为模型的中间表征①。由于表征具有跨层流动的特性，不同层之间所承载的功能呈现差异，而这种功能差异可以通过词义理解任务加以探测。

从模型架构的角度来看，不同模型对信息流动的建模方式也有所不同。以当前主流的大语言模型为例，其训练目标为自回归生成，这种模式更适配生成和预测类任务，与传统的基于编码器架构的模型形成鲜明对比。由于编码器架构模型通常采用掩码语言建模的方式进行训练，更适用于自然语言理解类任务。通过对比两类模型在词义理解任务上的表现，可以更全面地揭示不同架构模型的内在机制，尤其是加深对大语言模型的理解。

基于上述分析，本文重点关注以下研究问题：词义相关度任务如何为模型内部表征提供全面、准确的可解释性分析？不同架构的模型在表征层面存在哪些差异？这些差异是否呈现出特定的跨层演化模式？若存在，如何利用这些模式更好地理解和优化模型？

① 这是一种自上而下的视角，表征作为“整体”从模型各层输出，反映的是信息加工的宏观结果。与之相对的是自下而上的视角，该视角关注表征内部的“神经元”以及由神经元构成的“回路”所实现的功能，被视为神经网络的“逆向工程”，更加底层、更具微观性。两种视角的对比可参见文献综述[14]。

1.2.3　多标注者视角下的词义相关度评估任务

词义相关度任务通常需要多名标注者参与，而标注者之间往往存在不可忽视的不一致现象。这一现象由多方面因素导致。首先，语境对词义理解至关重要，但在某些场景下语境无法彻底消解词的歧义，例如信息量不足的上下文、缺乏的文化背景知识等，都可能导致对词义产生多重理解。其次，词义的区分过细也会导致不同语义之间并非界限分明，从而存在多种正确选择的可能。例如，WordNet[32]中的词义区分普遍被研究者认为过于细致[33]。此外，标注者自身的差异，如年龄、受教育程度、地域等因素，也会导致不同标注者对同样语境中的相同词产生不同的词义理解。

评估词义相关度任务需要从两个角度考虑多标注者视角。一方面，不同标注者的词义相关度判断可以聚合为统计量，用以代表对目标词的平均理解程度。该统计量反映了标注的集中趋势，可作为衡量模型词义相关度准确性的“黄金标准”。采用多标注者聚合结果优于单一标注者的标准，后者往往带有偶然性、随机性和主观性。这一角度侧重于衡量模型在词义相关度任务上的准确性。另一方面，多标注者判断的离散程度可用于衡量标注的不一致性，预测这种不一致程度同样是词义理解评估的重要方面。离散程度反映了词义在特定上下文中的不确定程度，是衡量模型词义理解不一致性的关键指标。不一致性评估之所以重要，是因为现实世界中的标注并非完美，不存在绝对一致的标注情况。优良的模型应当能够充分感知词义理解的不确定性，这体现了模型的鲁棒性①。这种能力也有助于根据模型理解词义的置信度筛选困难样本，并利用置信度指导模型对目标词的理解，从而在一定程度上缓解模型可能出现的“幻觉”问题。

基于上述分析，本文同样关注多标注者视角下的词义相关度评估，并着重研究以下问题：（1）如何将多标注者的标注结果转化为词义相关度的准确性和鲁棒性评估指标？（2）如何设计实验以评估模型在这两个维度上的表现？（3）如何利用模型对多标注者标注过程进行建模与预测？

1.2.4　研究问题间的关联

本节系统梳理了本文关注的核心研究问题，从三个视角出发对“词义相关度评估任务”进行多维度的思考与设计。多语言视角关注不同语言之间的共性与差异，强调词义相关度任务的设计应当覆盖多种语言，以检验模型在跨语言环境下的泛化能力。模型可解释性视角聚焦于分析模型内部表征的信息流动方式，将词

① 鲁棒性（robustness）指模型在面对标注不一致、数据歧义等真实噪声或者数据分布偏移时仍能稳定可靠工作的能力[34]。本文所关注的“预测标注不一致程度”虽未直接衡量这一能力，但通过让模型直接输出不确定性程度，间接反映了其鲁棒性：若模型能准确感知不一致程度，则面对此类噪声场景时更可能做出稳定可靠的判断。

义相关度任务作为探测手段，用以揭示信息在不同网络层之间的传递与演化过程。多标注者视角则从多名标注者的判断出发，将词义相关度任务的目标分解为准确性与不一致性两个维度，前者通过聚合标注者的共识来衡量，后者通过标注者之间的分歧程度来体现。模型的输出亦应同时反映这两个方面，以更全面地契合真实世界中词义相关度的不确定性与多样性。

1.3　研究内容和创新点

1.3.1　研究思路

针对上述研究问题，本文的研究路线如图 1.2 所示，主要包括四个步骤。第一，数据收集与构建。涵盖多语言数据集的语料构建，以及多标注者标注结果在准确性与不一致性两个维度的系统性设计。第二，模型准备。选取两类代表性模型架构，包括以 BERT（Bidirectional Encoder Representations from Transformers）[35]为代表的编码器模型和以大语言模型为代表的解码器模型。第三，评估方法设计。综合运用探测器方法、模型集成方法、提示语设计等多种评估手段，从多角度对模型的词义理解能力进行测度。第四，跨层信息流动分析，基于两类模型架构的内部表征，结合探测器方法，系统分析词义相关信息在不同网络层之间的传递与演化规律。

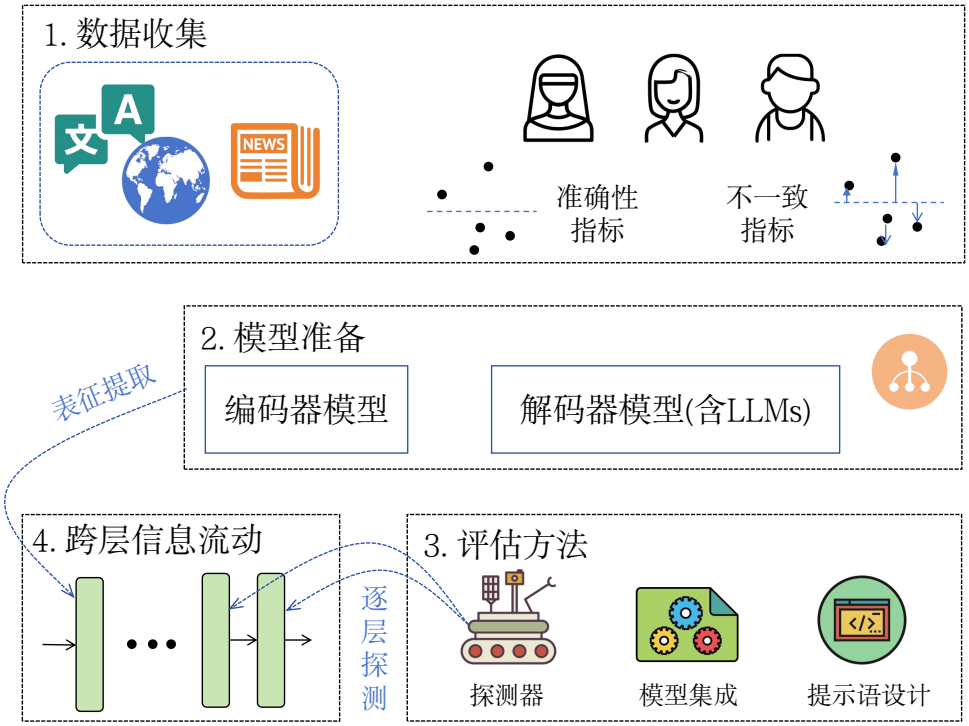


图 1.2　研究路线图

1.3.2　具体研究内容

参照研究思路，本文具体的研究内容如下：

（1）词义相关度任务的全流程评估框架设计

本文针对词义相关度任务，设计了全流程的评估框架，涵盖了多维度数据集、多个子任务、多类型评估模型、多指标体系。

数据集方面，本文既采用多种公开数据集，涵盖多语言、多标注形式、多体裁等维度；又通过自行收集的数据集对现有资源进行补充与扩展，以增强评估的覆盖面和代表性。

评估任务方面，本文设置多个子任务以全面反映词义理解能力。既包含与准确性相关的经典词义相关度任务，又引入涉及不确定性的标注者不一致预测任务。其中，准确性任务进一步细分为二分类的“是否相关”任务和多分类的“相关程度”任务，以考察模型在不同粒度上的语义判别能力。

评估指标方面，本文针对不同子任务采用与之匹配的评估指标，既包括经典的准确率等传统指标，也引入基于分布的比较指标以及衡量鲁棒性的专用指标。同时，采用定量分析与定性分析相结合的方式，通过大量实例深入分析模型的词义理解行为。多样化的指标体系确保了评估的全面性与科学性。

评估模型方面，本文选取的模型涵盖多种经典架构与参数规模，既包括传统的小规模语言模型，也包括参数规模宏大的大语言模型（涵盖不同版本与参数变体）；既包含以掩码语言建模为训练目标的编码器模型，也包含以自回归生成为目标的解码器模型。多类型模型的引入为跨架构比较提供了基础。针对不同任务，本文设计了与模型表征深度结合的评估方法，包括探测器分析、向量后处理、提示语诱导表征、提示语直接生成、模型集成等技术手段，旨在真实、客观地反映模型的词义理解能力。通过系统化的实证分析与结果对比，全面揭示不同模型在词义理解任务上的表现差异与内在机制。

（2）多语言视角下数据集的拓展

针对多语言视角，本文在已有词义相关度任务的基础上，对汉语相关数据集进行了针对性扩充。尽管经典的多语言数据集通常包含汉语数据，例如多语言词义相关度数据集 MLWiC[36] 和 OGWiC[37]，但这些数据集往往仅收集与汉语目标词相关的上下文，并未充分考虑汉语自身的语言特性。本文力图系统性地分析汉语的语言特点对其词义相关度的影响，并从汉英对比的视角深入探究二者在词义理解机制上的差异。

为此，本文聚焦于汉语中普遍存在的兼类词现象，将词类信息引入词义相关度任务中。具体研究路径如下：首先，从《现代汉语词典》[38] 中系统收集大量兼

类词实例，构建以名动对立为核心的上下文对，作为词义理解任务的数据基础。随后，招募涵盖多元背景的 69 名标注者对数据进行标注，确保标注结果的代表性与多样性。在构建完成中文兼类词数据集后，本文在多种模型上进行评估，重点分析词类信息的引入对汉语词义相关度任务结果的影响，并与英语的相应情况进行系统对比。

（3）多标注视角下的标注不一致程度评估

本文从多标注者视角出发，设计实验系统评估标注不一致程度对词义理解的影响，并将标注者不一致程度的预测纳入词义理解评估框架。在数据收集方面，每个标注实例不再依赖单一标注者，而是采用多位标注者共同标注的方式，通过衡量多位标注结果之间的离散程度来量化标注不一致性。本文自行构建的汉语兼类词数据集通过引入更多标注者，使不同样本间的不一致程度更加多样，同时缓解了原有数据集中汉语部分标注实例匮乏和标注者数量不足的问题。

在模型评估层面，本文综合运用模型集成、提示语设计、分类器探测等多种方法，利用不同模型对标注不一致程度进行预测，以考察模型对数据中固有不确定性的感知能力。此外，本文还对不同语言的数据开展了全面的定量与定性分析，并对常见偏误类型进行了归纳与整理。

（4）模型之间的跨层信息流动模式分析

针对模型可解释性视角，本文利用词义理解任务对比分析大语言模型内部表征的跨层信息流动模式，提出并验证了不同架构模型在跨层学习过程中的差异化目标。

基于表征的词义相关度任务有助于深化对模型工作机制的理解。本文重点关注表征在模型跨层传递过程中信息流动模式的演变规律。具体研究路径如下：首先，根据预训练目标的差异，将神经网络模型划分为编码器模型与大语言模型在内的解码器模型两类，并基于其训练过程提出关于信息流动模式的初始假说。随后，通过设计对照实验，系统对比分析信息在不同架构模型中的跨层流动过程。实验设计涵盖以下关键变量的控制与调节：模型提取目标词表征的位置、模型输入的格式、提示语的构造方式等。基于实验结果，本文归纳出不同架构模型在信息流动模式上的异同，从而对初始假说进行验证与修正。

1.3.3　本文创新点

（1）全面评估并分析模型词义相关度的表现能力

本文立足于词义相关度这一经典的词义理解任务，设计多样的评估方法，全面评估了各类模型，并展开了详尽的定量和定性分析。前人仅仅将词义相关度当作一个评测任务，并着力于开发有效的方法提升准确率。但是忽视了对模型自身

表征的评估，尤其缺乏大语言模型的表征理解。同时以定量分析为主，缺少对于个例的偏误分析。与之相反，本文着眼于表征评估，并在以下几个方面提出了新的方法：（a）表征提取：通过设计多样的提示语，来引导大语言模型产生与词义相关的表征；（b）表征评估：通过几何探测以及向量后处理方式来评估表征，其中向量后处理方法有效缓解了表征分布各向异性的特点，大幅提升了准确率；（c）设计针对闭源大语言模型的两类提示语，分别代表是否增加词类信息。

结合定量和定性两种方式详尽分析了模型的评估结果。定量方面，比较了各类模型在两类数据集（多语言和汉语兼类词）中的准确性，得出了解决词义相关度任务的方法排序，从最优到最差分别为：（a）根据相似任务重新训练、微调的编码器模型；（b）无需训练，仅设计提示语引导开源大语言模型产生的表征；（c）传统编码器模型的表征。而根据提示语直接引导闭源大语言模型产生的结果差异性较大，一般在上述（b）和（c）之间。之后我们也分析了各个方法模块的有效性，包括向量后处理方式、不同的提示语、多语言和词类对结果的影响。定性方面，本文将偏差类型划分为模型高估类错误和低估类错误，分别归纳出产生各自错误的几种典型原因，并针对每一种原因详细分析了中英两种语言的实例。其中高估类的典型原因包括：词汇边界忽略、词类差异忽视、语义过度泛化和事件结构简化；低估类原因包括：字面义保持缺失和语义域过度分化。

（2）全面评估并分析模型的标注不一致程度

本文关注多标注者视角下，模型对于标注不一致程度的评估，并进行了定量和定性结合的全面分析。标注不一致程度反映了人类理解词义时的不确定度，是衡量模型不确定性预测的一个重要方面。然而，已有的工作对标注不一致评估的关注较少，面对不同的标注，多数工作将他们丢弃或者随机选择一个标注；少量关注这一方面的工作也缺乏针对大语言模型的评估，并且缺少对于结果的全面分析。为了解决这些问题，本文首先收集了多标注的汉语兼类词，用来作为现有的标注不一致数据集的补充。其次重点提出模型集成的方法，通过模拟多专家的标注，将用于输出相关度的多种模型进行“扰动”变化，之后使用不同的集成策略来评估不一致程度。同时，本文还设计针对闭源大语言模型的两类提示语，来引导模型直接产生结果。最后，本文对实验结果进行了定量定性结合的分析。定量方面，分析了各类方法的不一致评估的有效性，其中模型集成的方式优于其他已有的模型以及本文提出的方法；之后重点分析了模型集成中各个重要部分发挥的作用。定性方面，同样将偏误类型分为高估偏误和低估偏误，其中高估错误数量显著高于低估的数量。接着重点分析了两类错误中的典型实例。

（3）对比分析大语言模型的跨层信息流动方式

本文基于词义理解任务，对比分析了大语言模型的跨层信息流动方式。前人对于模型的可解释性分析大多围绕传统的编码器模型，缺乏对于大语言模型各层之间表征行为模式的分析。一些工作对于大语言模型的词表征的提取仍使用默认的最高层，但是这与大语言模型自回归的训练方式不匹配。还有一些工作设计提示语引导模型产生表示语义的向量表征，然而它们主要针对于句义，缺乏对于更基本的词义理解的研究。本文在已有词义理解任务的基础上，从跨层的角度对大语言模型为代表的解码器模型和传统编码器模型进行对比研究，从而弥补了这些缺陷。具体来说，本文设计了多种对照设置，包括目标词提取位置、提示语、不同参数规模的模型、层数等选取的差异，对比分析了两类模型的层数变化趋势，从而验证了根据模型预训练过程所得到的“理解——预测”动态变化假说，即随层数增加，编码器模型理解当前词义能力在不断增强，但它整体上缺乏对下一词的预测能力；而解码器模型理解词义能力先增强后下降，在模型的中低层达到最佳，而预测下一个词的能力则是逐层增加，在高层达到最佳。这一“先理解后预测”的模式也出现在很多人类认知的场景下。

本文对于跨层表征信息的分析为大语言模型从自上而下的角度提供了模型的可解释性，通过大语言模型不同的跨层表现，可以为词义表征提取、算法优化等提供重要的思路。

（4）创建汉语兼类词词义理解数据集

针对汉语的特点，本文创新性得收集了汉语兼类词词义理解数据集，并在汉语的词义相关度数据上首次引入了词性标注。前人针对词义理解中的词义相关度数据集在汉语评估方面存在不足，一方面汉语的标注规模和语料来源有限；另一方面缺乏针对汉语词类的设计。针对这两个缺陷，本文重新收集汉语数据集，并且（a）扩大了汉语目标词的规模，从原有的 40 个增加到现在的 200 多个目标词；（b）增加了标注者的数量，从平均 2 个标注者增加近 5 名，且标注者背景多样；（c）语料来源于《现代汉语词典》，用例更加通用、来源更加广泛；（d）考虑到汉语存在数量众多的兼类词，从词典中选择兼类词用例，并设置名动“最小对立”的句子对来充当目标词的上下文，这样带有词类标注的例句弥补了原始汉语数据集无词类的缺陷，更有利于中英两种语言针对词类的对比。

本文提出的汉语兼类词数据集在之后评估模型的词义理解能力和不一致程度上都发挥了作用，并且可以通过多语言对比来分析“词类信息”对于词义理解的影响，体现了语言学知识在模型评估任务中的重要作用。

1.4　组织结构

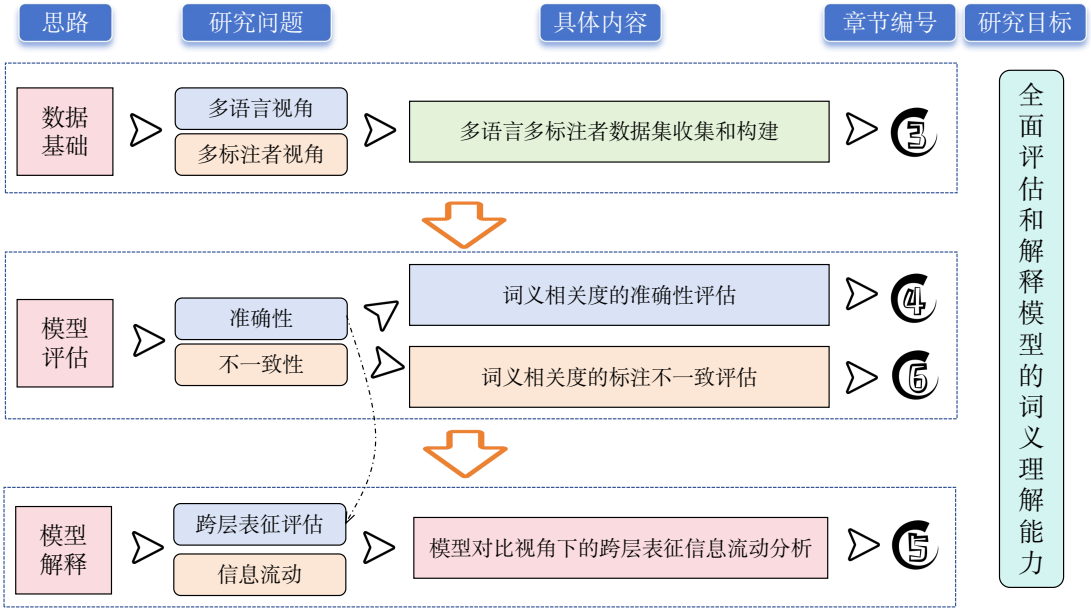


图 1.3　研究汇总和文章的组织结构

图 1.3 展示了本文的研究框架与各章节组织结构。全文共分为六个章节，具体安排如下：

第 1 章 绪论：本章首先阐述了大语言模型在词义相关度评估方面的研究背景与意义。在此基础上，提出了本文关注的三个核心研究问题，分别涵盖多语言视角、多标注者视角以及模型可解释性分析。围绕上述研究问题，进一步介绍了本文的主要研究内容与研究贡献。

第 2 章 研究现状与相关工作：本章系统梳理了与本文研究问题相关的研究现状与相关工作，主要从数据和模型两个维度展开。在数据方面，聚焦于词义相关度任务，首先介绍了与该任务相关的其他词义理解任务类型，随后从多语言视角和多标注者视角分别对相关研究进行了评述。在模型方面，围绕大语言模型展开，首先概述了模型的发展历程及其主要架构类别，接着介绍了与模型评估相关的研究，涵盖面向一般任务的模型评估与面向词义相关度任务的模型评估，最后评述了模型可解释性方面的相关工作。

第 3 章 任务定义与数据集构建：本章从多语言和多标注者两个视角出发，系统介绍本文所研究的词义相关度任务及其对应的数据集构建工作。本章首先给出该任务的形式化定义。随后，对研究中采用的词义相关度数据集进行概述，具体包括：（1）标注为二分类的英语通用数据集；（2）标注为多等级的多语言词义相关度数据集；（3）本文自行构建的汉语兼类词语义相关度数据集。其中，后两个数据集由于采用多标注者标注，均包含词义标注不一致信息，可用于不确定性的相

关分析。数据集的介绍涵盖收集原则、典型实例、评估指标、数据量统计等方面。本章所构建的数据集体现了前述研究问题中的多语言视角与多标注者视角，为后续实验奠定了数据与任务基础。

第 4 章 多语言词义相关度评估：本章聚焦于多语言场景下的词义相关度评估任务。首先，介绍评估过程中所采用的方法体系，包括表征提取方式、几何探测方法、向量后处理技术以及基于提示语的探测方法。随后，详细阐述实验设定，涵盖所评估的模型类型及其超参数配置。最后，从定量与定性两个维度对实验结果展开分析。定量分析主要围绕整体评估准确性、各类方法的有效性对比、不同模型之间的性能差异，以及词类信息对评估结果的影响等方面展开。定性分析则对模型产生的偏误案例进行误差分类，并呈现典型案例。本章的研究侧重于模型评估中的准确性维度，即以多标注者的平均结果作为词义相关度的黄金标准进行衡量。

第 5 章 模型内部表征与可解释性分析：本章在第 4 章的基础上，进一步分析模型内部表征的信息流动模式，从词义相关度的视角为大语言模型的可解释性研究提供新的切入点。首先，对比两类不同架构的模型（编码器模型与解码器模型）在信息表征方式上的差异。其次，设计实验以评估模型跨层信息传递的模式差异，实验设计涵盖数据集与模型介绍、表征提取方法以及探测任务设置。随后，基于多个数据集在词义相关度任务上的实验结果，分析模型在不同网络层上的性能演变规律，并据此验证实验前提出的研究假说。本章的评估任务建立在第 4 章词义相关度任务的基础之上，从模型内部视角剖析信息流动过程，旨在揭示模型做出词义判断的内在机理。

第 6 章 多语言词义相关度任务中的标注不一致评估：本章聚焦于词义相关度任务中的标注不一致现象，系统评估模型在标注不一致程度预测方面的表现。首先，介绍不一致程度评估的方法，包括基于分类器的探测方法、基于模型集成的预测方法以及基于提示语的探测方法。其次，详细阐述标注不一致程度预测任务的实验设定，涵盖任务定义与数据集介绍、评估模型类型、集成方法的实现及其超参数配置、分类器探测方法的设置，以及参与对比的基线模型。在实验结果分析部分，除呈现各类模型在不一致程度预测任务上的整体表现外，重点分析集成方法中各模块的有效性，以及模型在多种语言上的表现差异。最后，参照第 4 章的分析范式，本章也从定性角度对典型偏误案例进行分类与展示。本章的研究聚焦于模型评估中的标注不一致性维度，体现了模型对真实标注中固有不确定性的感知与预测能力。与第 4 章聚焦的“准确性”维度共同构成本文模型评估的两个核心方面。

第 7 章 总结与展望：本章对全文的研究工作进行系统总结，并对未来可能的

研究方向进行展望。

1.5　本章小结

本章作为全文的绪论部分，对论文的整体研究进行了引导性介绍。首先阐述了本文的研究背景与研究意义，说明了在当前大语言模型发展背景下开展词义理解研究的必要性。随后，从多语言视角、模型可解释性分析以及多标注者视角三个方面，系统梳理了本文所关注的核心研究问题，明确了研究的切入点与目标。

在此基础上，围绕上述研究问题，进一步介绍了本文的整体研究思路与主要工作内容，概括了论文的关键贡献与创新点，为后续各章节的展开提供了清晰的框架与逻辑支撑。最后，对全文的组织结构进行了说明，概述了各章节之间的关系与主要内容安排，从而为读者理解后文的具体研究奠定了基础。

第 2 章　研究现状和相关工作

本章从数据和模型两个角度分别介绍研究现状与相关工作。其中数据方面重点介绍面向模型评估的词义相关度任务，包含常见的词义理解任务类型，以及在多语言视角和多标注者视角下任务的设计。模型方面则着重关注面向自然语言处理的大语言模型，既包括大语言模型在内的神经网络模型的发展和类型的介绍，也包括模型评估和可解释性方面的相关研究。

2.1　词义相关度任务的相关工作

本节介绍与词义相关度任务相关的工作，首先，第 2.1.1 节从不同的角度说明了词义理解任务的常见类型，并阐述了它们之间的关联。其次，第 2.1.2 节介绍了多语言视角下的词义相关度任务，论述了不同任务的多语言版本，其中重点在词义相关度和汉语方面做了展开。第 2.1.3 节介绍了多标注者视角下的词义相关度任务，介绍了不同任务中存在的多标注结果，以及如何对待、评估和利用这类信息。

2.1.1　不同的任务类型

词义理解属于基础自然语言处理（fundamental natural language processing）[39]中的重要任务。其中词义相关度判断（Word in Context，WiC）是其中典型的任务，这一任务将同一目标词放置在一对上下文中，通过判断它们语义是否相似[40]、相关[37]或者一致[41]来进行词义理解 ①。

与词义相关度任务密切相关的任务类型为词义消歧（Word Sense Disambiguation, WSD）[43]，它旨在从一个候选词义集合中，确定目标词在上下文的含义，以此消除多义词的歧义性。WSD 可以针对所有词类的词[44]，且任务定义简单、直接 ②。

另外一类相似的任务被称为目标词语义确认任务（Target Sense Verification，TSV），它将 WSD 和 WiC 两个任务结合在一起。TSV 的任务形式为：输入目标词、上下文以及某一个候选语义，根据该目标词是否具有该语义输出“是”或“否”的选项。TSV 的输出结果类似 WiC，而数据的输入则类似于 WSD，已有研究工

① 有些研究工作着重区分相似性和相关性，例如学者 Emerson[42]提出基于功能分布学习来训练模型区分这两个概念：相似性指所指的语义特征重叠程度，而相关性指常常出现在同一上下文中。例如“画家”和“画”语义相关而不相似。而语义是否“一致”则仅关注两种绝对状态，即语义是否完全相同，而不关心中间的状态。

② 这一过程类似于学习者遇到生词，查阅词典的过程。

作[45-46]将这三个工作统一在一起，认为它们是同一问题的不同侧面。上述三个任务类型都属于判别式（discriminative），旨在属于单一的标签。释义建模任务（definition modelling）则通过生成（generative）的方式，根据给定目标词的上下文生成该词的词义。这一形式具有答案开放、输入简单的优势，然而也带来标准不唯一、结果判断困难的难题[47]。

从所针对的词类角度，一些 WSD 任务专注于特定词类的词汇。命名实体识别（Named Entity Recognition，NER）[48]识别命名实体，并将它们进行语义归类，这一任务所针对的是单义的专有名词；共指消解（Coreference Resolution, CR）[49]则关注代词的词义消解，它主要考虑到代词本身的无定性（underspecified）且具有指称的功能；词义消歧还可以针对合成词，例如探讨合成词内部成分之间的关联[50-51]、整体的透明度[52]等，同时词义消歧也可以专门针对虚词①，重点关注虚词的多功能性[25]。

除关注词类的不同外，有些任务也从词义之间的关系入手。针对同一词内部，隐喻检测和识别（metaphor detection and recognition）[55]关注语义之间通过隐喻方式的语义引申。词义演变任务[56]则关心词义之间的历时演变过程，并借助语义地图[25]等可视化工具进行展示。针对不同词之间的语义关系，词网 WordNet[32]将它们详细划分为上下位关系（hypernymy/hyponymy）、“部分—整体”关系（meronymy/holonymy）、同义（synonymy）和反义（antonymy）等关系类型。因此催生出了识别和检测这些关系的任务[57]。

本文重点关注实词相关的、且不考虑具体词义关联类型的词义相关度任务，任务的详细定义以及相关数据集在第 3 章给出。选择这一任务作为主要研究对象的理由如下：

（1）词义相关度任务无需提供可能的候选词义，因此任务的组织上更加简单。词义相关度任务仅需提供目标词所在的成对上下文，并仅需标注者通过对比，提供一个“相关程度”的标签。而 WSD 和 TSV 任务都需要显式提供词的义项，增加了额外的负担，并且对于标注者来说需要对义项提供的句子进行新的理解。这些都不如词义相关度任务高效。

（2）词义相关度通过比对的方式更加符合语言学中常见的“最小对立对”的设计原则。“最小对立对”通过观察变化最小的环境目标的变化，来确定相应的语义或功能，广泛应用于音素和语素识别[58]、虚词语义划分[59]。尽管词义相关度设

① 由于虚词在整体语义理解上不如实词重要，因此经典的 WSD 任务往往仅针对于实词，在句义理解上，虚词也仅仅被认为需要剔除的停用词（stop words）。然而虚词由于功能用法灵活多变，是语言学研究中比较关注的一类词[53]，且是外语学习者不易掌握的一类词[54]，因此不少研究者仍呼吁对于虚词理解任务的重视。

立的两个上下文并非一定是严格“最小”的①，但是这种对比的方式仍符合这一原则。通过对比的方法，可以帮助我们确立词义。

（3）“成对对比”的方式也广泛用于模型训练的语义对齐中。这一过程被称为“对比学习”（contrastive learning），即通过拉近相似样本的语义距离、拉远不同样本的语义距离来对齐语义。例如文图对比预训练 CLIP（Contrastive Language-Image Pre-training）将图文对进行对齐[61]，利用同义短语进行数据增强[62]。通过这种方式进行的模型词义理解评估也有助于模型的优化训练。

（4）词义相关度任务可以用于对表征的直接评估，有利于对模型内部进行更深入的分析。由于相关度任务直接与模型表征的相似性挂钩，后者可以直接作为词义理解的探测方式（即“几何探测”），因此可以直接探究模型内部表征的行为模式，从而促进模型可解释性。相反，WSD 等任务需要将表征与词义做进一步的对齐，因此需要额外训练一个模型（即“分类器探测”），不能直接反映表征的行为。

表 2.1 总结了上述常见词义理解任务，从不同维度对它们进行对比。

表 2.1　常见词义理解任务的对比

任务名称	任务简称	涉及词类	任务模式	历时关联②	提供词义
词义相关度	WiC	实词为主	二/多分类	✗	✗
词义消歧	WSD	实词为主	多分类	✗	✓
目标词词义确认	TSV	实词为主	二分类	✗	✓
释义建模	DM	实词为主	生成式	✗	✗
命名实体识别	NER	专有名词	多分类	✗	✗
指代消解	CR	代词	多分类	✗	✗
隐喻检测	-	实词	二分类	✓	✗
词义演变	-	不限	-	✓	✗

2.1.2　多语言视角下的词义相关度任务

多语言视角下的词义相关度任务考虑多种语言，由于语义存在于所有语言，因此涉及到词类理解的任务都可以拓展到多种语言。针对词义相关度任务，跨语言 WiC 任务 XL-WiC[63]是单语（英语）WiC 任务的多语言版本，它涵盖了不同语系的

① Trott 和 Bergen 的工作[60] RAW-C，则严格设立了“最小”的上下文对立对，仅通过变化上下文中的一个单词，就可以更改目标词的词义。然而这种严格对照的方式仅能通过人工造句的方式，没办法利用现成的例句，因此规模往往较小。

② 这一栏中的“历时关联”是指任务中是否探究词义之间的引申、演变等历时性的关系。

12 种语言。包括：汉语、丹麦语、法语、德语、意大利语、日语等。它们在不同语言中判断词义是否一致。针对判断更多相关度等级的 WiC 任务来说，依序分级的上下文词义相关度分类数据集 OGWiC（Ordinal Graded Word-in-Context Classification）[37]则考虑了包括汉语、英语在内的 7 种语言，在判断词义等级时，将其分为四个相关度等级（不相关、近相关、远相关和词义一致）。

针对其他词义理解任务，多语言设定仍是一个常见的选择。例如对于多语言 WSD 任务，早期 2010 年举办的语义评测竞赛 SemEval-10 第 17 个任务中关注意大利语和汉语[64]，之后在 2013 年举办的 SemEval-13 的第 12 个任务[65]和 2015 年 SemEval-15 的第 13 个任务[66]中关注法语、德语、意大利语和西班牙语。此后研究者将语言范围扩大，在数据集 XL-WSD 中收集了涵盖印欧语系、汉藏语系等四大语系的共 18 种语言的测试集。与多语言数据同时出现的还有多语言词义词典，例如 BabelNet[67]。而多语言 NER[68]数据集 Universal NER 收集了 13 种语言形式下的命名实体识别实例；多语言指代消解受益于多语言的统一标注框架 CorefUD[69]的开发，逐步扩展到了包含 15 种语言的 21 个数据集，其中涵盖英语、俄语、法语、挪威语，推动了多项指代消解的研究[70-72]。多语言隐喻识别[73]和涉及到语言类型学的词义历时演变[25]同样关注多种语言之间的词义理解。

尽管汉语是全世界使用最多的语言之一，但由于英语的通用性，使得专门针对汉语词义理解的任务设计较为缺乏。上述各类任务中涉及到汉语的，往往是由英语的原始数据集翻译而来，或者仅仅收集与任务相关的语料，并未考虑汉语的独特性。一些研究对此做了深入探讨，针对汉语的语素具有较为实在的语义，针对汉语 WSD 的数据集 MiCLS[74]首次将语素义消歧作为一个独立的任务。Hou 等人提出的数据集[75]则利用了针对汉语设计的义素词典知网 HowNet[76]来设计汉语 WSD。FiCLS（Formation informed Chinese Lexical Sample WSD）[30]数据集则将汉语的构词法引入到汉语 WSD 中，并作为一个有效的信息指导模型的训练。也有研究者[77-78]针对词类信息对汉语的同音词进行歧义消解，并且对比了模型和人类的表现差异。

2.1.3　多标注者视角下的词义相关度任务

多标注者标注是人工智能领域内任务设计广泛被采纳的视角，尤其在一些不确定性较强的领域。例如在计算机视觉领域中，图像分割的目标是将目标区域从背景中分离出来，然而在涉及边缘的分割时往往会出现分割不一致的现象，例如在医疗场景中某些器官边缘是否属于病变区域存在分歧[79]。而在自然语言处理领域，一些容易达成不一致的场景和鼓励多样性的场景都需要多位标注者。前者例如和语义判断相关的任务[37]，词类标注[80]，判断语句可接受度[81]；后者例如机

器翻译[82]、图像字幕生成[83]、故事生成[84]。

多标注者产生分歧具有多种原因。一方面与任务和数据本身的特性有关，例如在句义理解中，句子由于短语结构、或者某些特定的词汇语义造成的理解歧义[15]，而在词义理解中，词义划分的过细或者存在某个语义包含另一个的“上下位”关系时，也容易导致多位标注者产生不同的理解[85]。背景信息的不充分[29]是引起歧义性的另一个原因。另一方面与标注者自身息息相关，标注者对于任务的不同理解，不同的背景信息[86]等因素都影响标注结果。

这两个原因对应评估模型时不确定性估计（Uncertainty Estimation，UE）领域中的“数据不确定性”和“模型不确定性”[87]。数据不确定性①是指来源于数据的、无法通过收集更多数据消除掉的不确定性。它通常由噪声引起。模型不确定性②则来源于在不同于测试数据分布的数据上学习到的有偏模型，它对应于不同学习背景的标注者，它可以通过收集更多、更符合测试分布的数据来消除这类不确定性。

对于标注不一致结果，有多种处理方式。较为常见的一种是将其视作“噪声”或者有缺陷[88]的样本，将它们送入专家手中重新标注或直接将它们丢弃。这依赖人工智能领域的假设：每个标注仅存在一条黄金标准。第二种方式是将不一致视作是一种有用的信息加以保留和利用。在 WSD 经典数据中，有大约 20% 的样本标注没有达成一致集[85,89]，它们以多标签的方式保留在数据集中。尽管多数方法仅从这些标签选取一个作为“黄金标准”，多标签 WSD（Multilabel WSD, MLWSD）[90]将 WSD 视作一个多标签分类的任务，通过每个标签的损失累加③作为优化函数，更好地利用不同标注者的信息。在其他领域内，如情感分析或自然语言推理任务，对于标注者不一致现象，往往将多个“黄金”硬标签（hard label）转化为基于概率值的“软标签”（soft label），从而在学习过程中将它们的信息都有效利用起来[91]。

另外，对于标注不一致的预测和表示也被广泛研究。针对词义相关度问题，上下文词义相关度中不一致排序任务数据集 DisWiC（Disagreement in Word-in-Context Ranking）将多标注者的结果用刻画数据离散程度的数据量来量化不一致程度，从而设计共享任务来让参与者评估模型对于不一致程度的预测。其他自然语言处理任务，例如毒性语言和仇恨语言检测等主观性较强的任务都有研究者去预测标注不确定程度[92]。与标注不一致预测密切相关的是不确定性评估，它通过设计算法

① 数据不确定性又被称为随机不确定性（aleatoric uncertainty），äleatoric 指类似于投掷骰子的随机性，它无法通过更多的投掷次数而被消除。源自拉丁语 älea（骰子）。

② 模型不确定性又被称为知识不确定性（epidemic uncertainty），它来源于获取到的不充分知识，通过增加经验数据和知识，这类不确定性逐渐会被降低。

③ 文中使用了多标签的二值交叉熵损失（binary cross entropy）作为训练过程中的“损失函数”，这与传统使用仅针对单一标签的分类交叉熵损失（cross entropy）有所不同。

估计模型在词义预测任务时的不确信程度。由于不确定程度的标注标准不好界定，只能采用与不确定性相关的额外指标[93-94]来判断，例如根据不确定性程度来舍弃相关样本，从而判断最终的准确率。或者通过设计不确定性程度不同的场景来评估[29]。

综上，多标注者视角提供了在标注不确定场景下的一种处理方式。通过审视多个标注结果，来使用模型预测、评估不一致度，也可以使用这种多标注进行模型的训练和优化。同时，不确定性现象是不一致的重要来源，对于不确定性的评估也可以视作不一致程度的一种预测。

2.2　大语言模型的评估和可解释性研究

本节介绍与大语言模型相关的研究。首先先概述了包含大语言模型在内的神经网络语言模型，之后介绍了模型的评估研究，包括一般任务的模型评估和针对词义理解任务的模型评估，最后综述了与模型可解释性研究相关的工作。

2.2.1　大语言模型的概述

人工智能技术的发展脉络呈现出显著的阶段性特征，这一演变与数据规模、计算能力及模型架构的持续突破密切相关。在人工智能发展初期，研究范式普遍以理性主义（rationalism）为指导，倾向于通过符号主义（symbolism）的方法构建智能系统。这一时期的主流技术路线依赖于专家系统（expert systems）和人工设计规则（hand-crafted rules）来模拟人类的推理过程[95]。然而，无论是在计算机视觉还是自然语言处理领域，基于规则的方法都面临着一个根本性困境：现实世界的语言和视觉现象充满了歧义性、上下文依赖性和长尾分布，导致规则系统不可避免地产生大量例外，难以准确刻画和预测复杂的现实数据。此外，规则的制定和维护需要领域专家的深度参与，不仅效率低下，更缺乏跨领域的可迁移性，难以实现大规模推广应用。随着这些局限性的日益凸显，加之相关理论的瓶颈，人工智能领域在 20 世纪 80 年代末至 90 年代经历了一段发展放缓的“寒冬期”[96]。

在 20 世纪 80 年代末至 90 年代初，人工智能领域迎来了第二次技术复兴。这一转折得益于大规模标注数据集的构建和统计学习方法的进步，研究范式也随之从理性主义转向经验主义（empiricism）[97]。这一时期的技术路线主要分为两个阶段：特征提取和模式识别。特征提取既可通过人工设计的方式完成[98]，也可通过自动学习的方式获得[99]；而模式识别则广泛采用各种统计学习算法，例如朴素贝叶斯（naive Bayes）[100]、决策树（decision tree）[101]以及支持向量机（Support Vector Machines, SVM）[102]等方法。经验主义范式通过从数据中提取特征并进行

模式识别，有效降低了对人工规则的依赖，从而增强了模型的鲁棒性（robustness）和泛化能力（generalization）。然而，受限于当时的数据规模和计算能力，这一阶段的模型仍难以支撑更广泛的实际应用场景。

在 21 世纪第一个十年的末期至第二个十年的初期，人工智能领域迎来了第三次技术革新，其核心驱动力是深度学习（deep learning）的崛起。这一范式通过在大规模数据集上进行端到端训练，使得模型能够直接从原始输入中学习完成任务所需的知识表征。深度学习通过构建多层神经网络（neural networks），将原始数据作为输入、标注作为输出，实现了从输入到输出的直接映射，从而将传统流程中的特征提取与模式识别两个阶段合二为一。由于特征由模型自动习得，这一方法显著提升了建模效率和表征能力。

以自然语言处理领域为例，Hinton 等人在 2009 年将深度神经网络（neural networks）应用于语音识别任务，取得了突破性进展[103]；与此同时，在计算机视觉领域，AlexNet 在 2012 年 ImageNet 图像识别竞赛（ILSVRC）中大幅超越传统方法，标志着深度学习时代的正式开启[104]。此后，随着技术的不断演进，研究范式也随之发生深刻变革。在自然语言处理领域，技术路线经历了从“预训练—微调”范式[35]向生成式基础模型范式的转变：早期方法首先在大规模语料上进行预训练，随后在特定任务上进行微调；而近年来，随着大语言模型的兴起，研究范式进一步转向在海量互联网数据上进行生成式预训练，构建统一的基础模型（foundation models），并将各类下游任务统一建模为生成任务[105-106]。这一转变的根本原因在于，大语言模型已经在互联网量级的数据上完成了充分训练，其性能在扩展定律（scaling laws）①的指导下持续提升，逐渐展现出强大的通用性和准确性。

在自然语言处理领域，神经网络模型的核心演进脉络体现为从静态词表征（static word representations）向上下文相关的动态表征（contextualized representations）的转变。前者以 Word2Vec[108]为代表，通过浅层神经网络为每个词学习一个固定且与上下文无关的向量；后者则以 Transformer[16]为主要架构，能够根据具体上下文动态调整词的表示。Transformer 是一种基于自注意力机制（self-attention mechanism）的多层神经网络架构，是现代预训练语言模型和大语言模型的主要架构。其核心思想是：每一层的输入和输出均为连续的向量序列（即词元的表征），自注意力机制通过计算序列中任意两个位置之间的相似度得分作为注意力权重，从而根据其他词元对当前词元的影响程度，重新加权聚合得到更新的表征。这种机制使得模型能够灵活捕获长距离依赖和复杂的上下文关系，成为后续大语言模型

① 扩展定律[107]是指随着模型参数量、数据量和计算量的增加，模型性能呈现可预测的提升。这一定律为采用更大规模的模型和数据提供了理论依据。

的基础架构。公式 (2.1) 展示了这一过程：

$$h_i^{l+1} = \sum_{j=1}^{M} \alpha_{ij} h_j^l \tag{2.1}$$

其中 M 为文本中的词元个数，i 和 j 为索引，注意力权重 α_{ij} 定义为：

$$\alpha_{ij} = \frac{\exp\left((h_i^l)^\top h_j^l\right)}{\sum_{k=1}^{M} \exp\left((h_i^l)^\top h_k^l\right)} \tag{2.2}$$

在上述公式中，h_i^l 表示第 l 层中位置 i 的上下文表示向量，h_j^l 表示同一层中位置 j 的上下文表示向量。$(h_i^l)^\top h_j^l$ 表示两个位置之间的相似度度量，用于衡量位置 i 在编码时对位置 j 的关注程度。函数 exp 表示指数计算，注意力权重 α_{ij} 通过公式 (2.2) 中 softmax 归一化得到，表示位置 i 在聚合上下文信息时，对位置 j 的贡献比例。最终，模型通过对所有位置的上下文表示进行加权求和，得到更新后的表示 h_i^{l+1} ①，并实现目标词与上下文信息的融合。

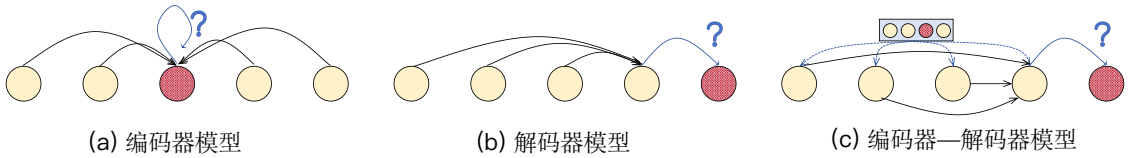
Transformer 架构与早期的循环神经网络（Recurrent Neural Networks, RNN）及其变体——如长短期记忆网络（Long Short-term memory, LSTM）——存在本质差异。后者以序列建模为核心，需要按照时间步逐步传递隐状态以捕捉历史信息，这种串行计算方式限制了其在现代并行计算硬件上的训练效率[109-110]。相比之下，Transformer 摒弃了循环结构，完全基于自注意力机制，使得序列中的所有位置可以同时参与计算。这种设计天然适配于图形处理器（GPU）的并行计算特性，从而能够在大规模数据上进行高效训练，成为推动大语言模型发展的关键架构基础。

Transformer 架构的神经网络语言模型又可以分别编码器模型、解码器模型、编码器—解码器模型。图 2.1 展现了三类模型的示意图，其中每个圆圈代表 Transformer 输入的每个词元或者该位置的中间输出表征②，黑色实线及箭头表明自注意力计算过程中需要参考哪些词元的信息，蓝色实线和问号代表模型的训练目标，蓝色虚线则表示其他上下文带来的信息。编码器模型（encoder）架构以 BERT（Bidirectional Encoder Representations from Transformers）[35]为典型代表，它的核心预训练目标为“掩码语言模型”（masked language modelling），即通过掩码（mask）掩盖掉当前词，利用两边的上下文来预测被掩盖掉的词。这种预训练的方式采用一种无需额外提供人类标注的自监督方法因此可以在原始大规模的语料上直接进行训练。训练过程采用双向可见的自注意力模板。训练后的模型更适合于理解型任务，例如词义相关度[41]、WSD[43]、共指消解 CR[111]、命名实体识别 NER[112]等。

① 这里简化了向量在融合之后的非线性层操作。

② 为了简化运算过程，图 2.1 忽略了层数之间的运算过程，同时忽略了自注意力运算之后的非线性的前馈神经网络层 FFN（feedforward neural network）。在后文第 5 章详细论述层数之间信息流动过程时，会将层之间的流动展示出来，例如图 5.1。

图 2.1(a) 展示的编码器模型示意图中，红色结点代表的目标词需要集中来自周围词元的信息，并且目标旨在恢复自己被掩盖掉的词。



(a) 编码器模型　(b) 解码器模型　(c) 编码器—解码器模型

图 2.1　三类 Transformer 模型架构示意图

解码器模型（decoder architecture）以 GPT（Generative Pre-trained Transformer）系列为典型代表[105,113]。其核心预训练目标为“下一个词预测”（Next Token Prediction, NTP），即基于当前词元及其历史上下文来预测序列中的下一个词元。与编码器架构中采用的掩码语言模型（Masked Language Model, MLM）不同，解码器所使用的自注意力机制仅允许每个位置关注其左侧的历史词元，而无法获取右侧的未来信息，这种机制被称为因果自注意力（causal self-attention）或自回归（autoregressive）建模。该训练方式同样属于自监督学习（self-supervised learning）范式。

由于其内在的自回归特性，解码器架构更适用于生成式任务，例如：文本摘要（text summarization）[114]、对话生成（dialogue generation）[115]、故事生成（story generation）[116]等。相较于编码器架构，解码器的预训练目标更为简洁统一，因而更容易扩展至更大规模的模型参数和训练数据。当前主流的大语言模型（large language models, LLMs）均采用解码器架构，借助规模化（scaling），这些模型在广泛的任务上展现出惊人且通用的能力。

图 2.1(b) 展示了解码器模型的示意图，其中红色结点代表当前待预测的目标词元，它需要融合来自历史词元的上下文信息，其训练目标为基于已有序列预测下一个词元。

编码器—解码器模型（encoder-decoder architecture）将编码器与解码器有机结合，专门用于“先理解、后生成”的任务范式。其典型应用包括机器翻译（machine translation）[117]、文本摘要（text summarization）[114]等。该架构以“序列到序列”（sequence-to-sequence）的方式处理任务：编码器首先对输入序列进行理解性编码，解码器随后基于编码结果逐词生成输出序列。经典实现包括 T5（Text-to-Text Transfer Transformer）[118]和 BART（Bidirectional and Auto-Regressive Transformer）[119]等模型。

在具体实现中，编码器和解码器分别采用各自的自注意力模板。编码器使用双向自注意力，能够捕捉输入序列的完整上下文信息；解码器则沿用因果自注意力（causal self-attention），仅关注已生成的历史词元。关键在于，解码器在每一层

还会通过交叉注意力（cross-attention）机制与编码器输出的中间表征进行融合，从而将输入序列的理解结果注入生成过程[16]。

图 2.1(c) 展示了解码器-编码器模型的示意图。图中表明，在预测下一个目标词元时，解码器不仅需要融合自身的历史词元信息，还需要将编码器从输入序列中提取的语义表征作为输入的一部分，共同指导生成过程。

本文重点考虑编码器和解码器。由于不涉及序列到序列的任务，因此不采用“编码器——解码器”的架构，但在实验的评估模型中采用了“编码器——解码器”结构中的“编码器”部分。

2.2.2　大语言模型的评估研究

2.2.2.1　一般任务的模型评估

模型评估①旨在清晰、全面地了解模型在不同方面的性能表现。随着机器学习模型，尤其是大规模语言模型的广泛应用，研究者逐渐从单一任务性能评估转向多维度的综合评估框架。常见的评估维度包括准确性（accuracy）、鲁棒性（robustness）、安全性（safety）、效率（efficiency）以及不确定性（uncertainty）等。

准确性是评估模型在特定任务上完成质量的基础维度，不同任务通常采用不同的评价指标。例如，在分类任务中常用准确率或 F1 分数，在机器翻译中则采用 BLEU（Bilingual Evaluation Understudy）指标来衡量译文质量[120]。随着大语言模型的发展，评估体系也扩展至指令遵循能力[121]、幻觉出现率[122]以及检索增强生成中的检索准确率[123]等维度。

为系统比较不同模型的能力，研究者构建了一系列综合评测基准。例如，GLUE 基准通过整合多种自然语言理解任务，为模型提供了统一的评测框架[10]；其后续版本 SuperGLUE 则进一步提升了任务难度和评测标准，以更好地区分不同模型的能力水平[124]。此外，大规模评测基准 BIG-bench 涵盖数百项任务，从多维度对模型能力进行更全面的评估[125-126]。在此基础上，大规模多任务语言理解数据集 MMLU 和面向代码生成的 HumanEval 数据集分别从语言理解和代码质量角度对大语言模型进行评估[127-128]。

鲁棒性评估关注模型在输入扰动或分布变化条件下的稳定性。例如，当输入文本发生轻微变化（如同义词替换、拼写错误或句法结构改变）时，模型是否仍能保持稳定的预测结果。研究表明，许多自然语言处理模型在面对对抗性样本或语义保持的扰动时会出现明显的性能下降[129]。

为系统化评估模型的鲁棒性，Ribeiro 等提出了 CheckList 框架，通过构建语言

① 下文介绍的几个评估维度，虽然在大语言模型兴起之前就是评估神经网络模型和机器学习的重要方面，但同样也是评估大语言模型的重要角度。

学驱动的测试集合，对模型在不同语言现象下的行为进行结构化测试[130]。Lunardi 等通过对经典评测数据集如 MMLU 进行同义改写，进而测试了 34 个主流大语言模型的鲁棒性[131]。Nalbandyan 等提出了 SCORE 框架（Systematic COnsistency and Robustness Evaluation），通过更改输入提示语的格式来评测模型的鲁棒性[132]。

安全性评估主要关注模型在生成或预测过程中可能产生的有害内容、社会偏见或不当行为。例如，一些研究通过构建专门的数据集来评估语言模型生成有毒或冒犯性文本的倾向，其中 RealToxicityPrompts 数据集被广泛用于测试生成模型的毒性风险[133]。随着大模型能力的不断增强，研究者开始提出更加系统化的评估框架。例如 HELM（Holistic Evaluation of Language Models）从准确性、鲁棒性、公平性、毒性、效率等多个维度对语言模型进行整体性评估，以更全面地反映模型在真实应用场景中的表现[134]。

在中文语境下，研究者提出了 SafetyBench 基准，该数据集采用中英双语设定，涵盖七个核心安全性维度：冒犯性（offensiveness）、不公正与偏见（unfairness and bias）、身体健康（physical health）、心理健康（mental health）、非法活动（illegal activities）、伦理与道德（ethics and morality）以及隐私与所有权（privacy and property）[135]。这一基准为评估大语言模型在多样化文化背景下的安全性表现提供了重要工具。

效率评估则关注模型在实际部署中的计算成本与资源消耗[136]，包括推理时间、内存占用、模型参数规模以及吞吐量等指标。在许多实际应用场景中，模型需要在性能与计算成本之间进行权衡，因此效率指标逐渐成为模型评估的重要组成部分[9]。

除了上述维度之外，近年来越来越多的研究开始关注不确定性[87,137-138]。不确定性通常来源于数据本身的歧义性或标注者之间的分歧。在许多自然语言任务中，不同标注者可能对同一输入给出不同但合理的标签，这种现象被称为标注分歧（annotator disagreement）[91]。相关研究表明，这种分歧往往反映了语言现象本身的模糊性，而不仅仅是标注噪声。因此，一些研究开始通过建模标注分布或预测标注者分歧程度来刻画数据中的不确定性，从而更全面地评估模型对复杂语义现象的理解能力[37,91,139]。

总体而言，模型评估研究正在从传统的单一性能指标，逐渐发展为多维度、系统化的综合评估体系。未来的研究趋势包括构建更加多样化的评测基准、提升评估的可解释性，以及建立能够反映真实应用环境的动态评估框架。

2.2.2.2　词义相关度任务的模型评估

针对于本文所聚焦的词义相关度任务以及与之相关联的词义消歧任务 WSD，主要侧重于准确性和不确定性两个方面的评估①。

准确性方面的评估可以根据大语言模型的出现分为前后两个阶段。在大语言模型之前的“预训练—微调”时期，通常的做法是使用静态语言模型 Word2Vec[108]或者预训练语言模型[35]的中间表征来表示词义，再通过直接计算相似性（针对词义相关度任务）或者计算与候选语义选项的相似性（针对 WSD）来计算结果[140]。其中预训练语言模型包括 BERT[35]，RoBERTa[141]等模型。对于多语言相关的词义相关度任务，还额外使用多语言语言模型，例如 XLM-RoBERTa[142]。

在大语言模型之后，主要通过提示语设计的方式来评估词义理解任务。一方面可以直接让提示语完成 WSD 或者词义相关度任务。例如一项针对大语言模型的评估工作[143]使用大量不同的提示语来评测 WSD 任务，发现大语言模型已经达到中等偏上的水准，性能接近距离专门使用相关数据集做微调的结果，表明了大语言模型的词义理解方面的巨大潜力。另一方面，提示语可以引导模型产生更加准确的表征。PromptEoL[144]和 PCoTEoL[145]通过使用提示语表征句义，从而更好地完成很多句义理解的任务。尽管它们不是直接针对词义的，但是基本方法可以被句义理解评估所利用。Cheng 等人的工作[146]则利用成对的提示语，并使用不同提示语的向量表征做几何运算来表示最终的词义。

针对这两个词义理解任务的不确定性，相关工作还研究得不够充分，仅有少量的工作对这方面进行了评估。例如 UEWSD（uncertainty estimation in WSD）[29]针对 WSD 任务，设计了相关的指标，来评估模型在数据不确定性和模型不确定性两种场景下的不确定程度。ALWSD（active learning in WSD）[147]估计了不确定性后，利用高不确定性的样本作为更有信息量的样本指导模型进行训练。在词义相关度任务中，常常使用标注不一致来代表不确定性程度，其中蕴含如下的假设：不确定性高的相关度评估往往意味着标注不一致性程度高。评估模型不一致程度的方法可以使用模型集成[139]和分类器探测[148-149]等方式。由于标注不一致可以显式被表示出来，因此可以通过与专家的不一致程度的相关度等指标评估模型不一致程度的好坏[37]。

① 本文研究的模型评估侧重于模型自身的输出，因此在评估过程中模型参数完全冻结，不参与任何额外训练。这一设定区别于利用模型实现特定任务的研究范式，后者通常需要对模型进行微调或引入外部知识以提升任务表现，例如在词义消歧任务 WSD 中，方法 EWISER[89]通过使用预训练模型并经过二次训练，同时引入额外的知识信息以更好地完成 WSD。

2.2.3　大语言模型的可解释性研究

理解大语言模型等神经网络语言模型内部表征结构与决策机制已经成为人工智能领域的重要研究方向。现有研究通常从“可解释性（interpretability）”的角度出发，试图揭示模型在处理语言信息时的内部运作原理。按照分析视角的不同，可解释性研究大体可分为自底向上的机制可解释性（mechanistic interpretability）与自顶向下的表征可解释性（representational interpretability）两类[14]。

机制可解释性关注将模型的底层计算组件还原为人类可理解的算法过程，通常通过将模型表示为计算图（computational graph），并在其中识别具有特定功能的子结构或“电路”（circuits），以解释模型如何实现特定行为或完成特定任务[150-152]。该类研究强调对注意力头、特征空间[153]、神经元及其组合功能的精细分析，力图从结构层面解释模型的推理与生成机制。

与之相对，表征可解释性采取自顶向下的视角[14,154-156]，抽象掉底层计算细节，重点关注隐藏表示空间的结构、几何性质及其所编码的语言信息。探针（probing）方法是表征可解释性研究中的核心技术路径，旨在检验模型内部表示是否显式或隐式地包含特定语言学属性。根据技术路线的不同，探针方法可进一步分为基于分类器探测（classifier-based probing）与几何探测（geometric probing）。

基于分类器的探测方法通过在固定模型参数的前提下，引入额外的轻量级分类器来完成代理任务（proxy tasks），从而判断隐藏表示中是否蕴含与该任务相关的信息。这类方法在 BERT 模型兴起时就已经被广泛应用于词义[157]、上下文长度[158]、句法结构分析[159]、语义角色标注[156]、命名实体识别[160]以及世界知识建模[161]等任务。相关研究表明，语言模型内部的语言学特征呈现出明显的层级结构：较底层更偏向词法与句法信息，而高层则更多编码语义与语用特征[162]。然而，基于分类器的探针方法通常依赖额外参数与任务设计，可能引入探针本身的表达能力，从而影响对原始表示可解释性的客观评估。

为避免额外模型带来的干扰，几何探测方法直接分析表示空间本身的几何结构特性，不引入额外分类器。例如，通过构造差分向量（difference vectors），即对不同语境下的向量表示进行减法运算，可有效检测语言特征，如标量形容词的强度变化[163]以及文体与风格特征[164]。此外，对向量空间施加主成分分析（PCA）、独立成分分析（ICA）等线性或非线性变换，也可揭示隐藏表示中的语义维度与潜在结构[165]。相比基于分类器的探针，几何方法更侧重对表示空间本体结构的解释，因而在研究模型内在语义组织方式方面具有独特优势。表 2.2 总结了分类器探测和几何探测的异同。

除静态表征结构分析外，近年来亦有研究从“信息流动”（information flow）的

表 2.2　两类探测方法的对比

探测名称	表征作为输入	改变原始模型	额外训练参数	探测任务	表征理解方式
分类器探测	✓	✗	✓	广	间接
几何探测	✓	✗	✗	有限	直接

角度[166-167]探讨模型在不同层级中对信息的收集与利用机制。Liu 等人[154]考虑通用的词义相关度任务，并且全面对比不同类型神经网络语言模型表征的跨层信息流动模式的差异。Ma 等人[78]则针对中英同音词的跨层表征，并着重观察词类信息对于两种语言的不同影响。这些工作表明，采用自回归目标函数的神经模型[168]，尤其是大语言模型[166]，通常在较浅层聚合输入中的关键信息，而在较深层完成预测决策与输出生成。这一发现从动态计算过程的角度补充了层级表征结构的解释，为理解模型如何逐步构建语义表示提供了新的分析框架。

总体而言，现有关于语言模型表征与可解释性的研究已从单纯的性能分析，逐步转向对内部结构、语义组织方式及信息处理机制的系统刻画。机制可解释性与表征可解释性在分析层面上各有侧重，前者强调对计算路径的精细还原，后者关注隐藏空间中语言知识的分布与结构。二者的互补发展，为深入理解大语言模型的推理能力、泛化行为以及在复杂语言现象（如歧义性、多义性）中的表现奠定了方法论基础。

现有的研究缺乏对于大语言模型内部表征的信息流动模型的研究，尤其与传统编码的对比，本文主要在第 5 章中采用自上而下的表征可解释性的方式对于包括大语言模型在内的两种神经语言模型进行了探测和对比研究。

2.3　本章小结

为奠定本研究的理论基础，本章将对相关研究现状进行综述，重点聚焦于词义相关度任务与大语言模型两大板块。第一部分聚焦于任务本身，通过梳理词义相关度任务的常见类型，并着重探讨在多语言和多标注者视角下的设计原则与数据处理方法，揭示了其与本文核心研究问题的内在关联。第二部分则着眼于本文所用的评估模型。本部分首先对大语言模型和传统预训练语言模型在内的神经网络模型进行概述，继而回顾了其在模型评估方面的研究进展，最后探讨了提升大语言模型可解释性的相关探索。本章介绍的这些工作，直接对应于本文的研究内容与对象，为后续的深入分析与创新提供了必要的学术背景。

第 3 章　词义相关度任务和数据集概述

本节介绍词义相关度任务。首先给出了任务定义，厘清了其中的关键概念。之后对词义相关度的数据集做了详细概述，包括数据集的组织形式、特征、数据量统计等。

3.1　词义相关度任务定义

3.1.1　词义相关度任务

词义相关度判断任务旨在确定目标词在两个文本片段中的词义相关程度。经典的词义相关度任务 WiC（Word-in-Context）仅判断词义是否相同[41]。例如①:

(1)　a.　种花儿
　　b.　花钱
　　c.　花盆儿

目标词“花”在例 (1-a) 中表示“植物”义，而在例 (1-b) 中表示“耗费”义，两个词义完全不同，因此对于 (1-a) 和 (1-b) 的例句对，最终的专家词义标注的结果是词义不相同。而对于例 (1-a) 和例 (1-c) 来说，它们均表示“植物”的意义，因此专家标注的结果为词义相同。

词义相关度判断更一般的定义为不相关到完全一致的多个等级程度，例如数据集 OGWiC（Ordinal Graded Word-in-Context Classification）将相关度划分为四个等级：完全一致、相关程度较近、相关程度较远、完全不同[37]。GWiC（Graded Word Similarity in Context task）则直接使用 0 到 1 的一个数值衡量目标词在不同上下文的词义相似度[40]。这些设计都符合词义在不同上下文中的连续统变化现象[169-170]，因而可以更加准确地刻画词义之间的比较。

另一方面，在多等级或者连续值判断的词义相关度判断任务中常常出现标注不一致的现象，即针对同一目标，多个标注者往往做出不同的判断。一方面目标词的多个语义之间往往具有交叠，很难完全清晰地区分开[169-170]。另一方面不同标注者由于地域、年龄、学历、专业等背景不同，对词本身往往具有不一样的理解。标注不一致现象普遍出现在与词义相关的场景中，尽管有些研究工作选择将其视作是不完美样本而选择丢弃，越来越多的研究工作[37,91]讨论如何表示、预测或者

① 目标词均使用下划线进行强调，例句来源于《现代汉语词典》[38]中“花”的词条。

利用这一类数据。

本文在第 4 章聚焦于多等级的词义相关度判断任务对现有的神经网络语言模型进行充分的评估，并利用汉语语言的独特性，收集与汉语兼类词相关的词义相关度数据集。同时，第 6 章也关注词义相关度判断的标注者不一致现象，并评估语言模型对不一致程度预测的能力。

3.1.2　形式化定义

假定目标词 w 出现在由 M 个词构成的上下文 c 中，它在句子中的位置索引为 j，j 为小于等于 M 的一个自然数，即：

$$c = (w_1, w_2, \cdots, w_j, \cdots, w_M) \tag{3.1}$$

其中 w_i 表示在上下文中，位置索引为 i 的词，小括号 () 代表括号内部元素构成的有序序列。

对于词义相关度判断而言，目标词需出现在成对的上下文中，记作 c_1 和 c_2。专家或者神经网络语言模型需给出目标词 w 在这一对上下文中的相关度。这一过程用公式表示如下：

$$y = f(w|c_1, c_2), \quad y \in \{1, \cdots, K\} \tag{3.2}$$

其中 f 代表人类专家或者神经网络语言模型对词义相关度的判断函数，y 表示相关度的结果，使用 1 到 K 之间的自然数来表示相关程度，其中数值越大，它们的相关度越高，K 表示完全一致，1 表示完全不同。花括号 {} 表示集合。对于经典 WiC 任务而言，K 的取值为 2，对于多级别 WiC 任务而言，K 的取值大于 2。

词义相关度判断任务出现的不一致现象表示为，不同的标注者有不同的标注结果。假定有 H 名标注者，最终得到一个标注集合：

$$\mathcal{Y} = \{y_1, y_2, \cdots, y_h, \cdots, y_H\} \tag{3.3}$$

其中 y_h 表示第 h 位标注者的相关度标注结果。标注不一致现象是指标注集合 $\mathcal{Y}$ 中的元素不完全一致。

3.1.3　与词义消歧任务的关联

词义消歧（Word Sense Disambiguation）是另外一类与上下文词义密切相关的任务。它旨在确定目标词出现在一段单一文本片段中的词义，其中词义从给定的候选词义集合中选择。这一过程类似于学习者从词典中确定某一特定词的词义。例如上文例(1)中目标词“花”需要从以下候选的语义中选择[①]：

① 释义来源于《现代汉语词典》

(2)　a.　名词，可供观赏的植物

b.　动词，用、耗费

c.　形状像花朵的东西

对于例为 (1-a) 来说，词义消歧任务的目标是从上述三个候选语义中确定其正确语义为 (2-a)；而例 (1-b) 中，目标词的正确语义为 (2-b)。

词义消歧需指定一个包括所有可能词义的集合，或者词典（inventory）$\mathcal{D}$，专家或者神经网络语言模型需给出目标词 w 在上下文 c 中的词义：

$$y = h(w|c), \quad y \in \mathcal{D} \tag{3.4}$$

其中 h 代表人类专家或者神经网络语言模型对于词义消歧任务的词义选择函数，y 表示词典 $\mathcal{D}$ 中某一特定词义。

两类任务存在密切的关联[45]。词义相关度任务可以视作在没有候选词库下特殊的词义消歧。它通过对照不同上下文的相关度来间接获取词义信息。当目标词都可以找到在所有不同上下文中的相关度时，它可以根据相关度聚集为不同的词义类别，该语义类别便可以组成词义消歧中的语义词典 $\mathcal{D}$。反过来，词义消歧则是一个显式的词义选择过程，因此仅需单一的上下文就可以判断词义。假定目标词在所有的上下文中的词义都得到了消解，自然可以进一步判断它们之间的相关度。

另一方面，词义消歧通过上下文文本对多义词进行词义消解，然而歧义并非总是可以得到完全消解，这一现象也被称作不确定性。造成消解不充分的原因有很多，例如当上下文的信息量不够时，多义词仍具有歧义，试举下面例句①：

(3)　a.　自行车没锁

b.　自行车没锁，以后出门不用担心忘带钥匙。

c.　自行车没锁，因为我忘记带锁具了。

而《现代汉语词典》[38]对于“锁”有如下的释义：

(4)　a.　名词，安在门窗、器物等的开合处或连接处，使人不能随便打开的金属器具，要用钥匙、密码、磁卡等才能打开。

b.　动词，用锁把门窗、器物等的开合处关住或拴住。

对于例 (3-a) 来说，由于没有充分的信息，无法确定“锁”在上下文中是表示释义 (4-a) 还是释义 (4-b)，而例 (3-b) 和例 (3-c) 则通过更多的上下文信息将“锁”消

① 例 (3-a) 来源于郭锐老师于 2025 年 11 月 7 日于中央民族大学举办的讲座，标题为“语言如何表达意义”。例 (3-b) 和例 (3-c) 为自拟。

歧为释义 (4-a) 和释义 (4-b)。

造成词义不确定性的原因还有很多，例如某些词的候选词义之间交叠使得消歧在某一些可能的候选义之间无法确定下来；用于消歧的神经网络语言模型在学习过程中产生多种可能的参数空间等等。

词义相关度中的标注不一致现象和词义消歧中的不确定现象也存在密切的关联，它们都反映了词义在判断过程中无法严格、唯一确定目标的现象。不同的是，标注不一致现象侧重于标注者标注的结果，强调有多名标注人员，且不同人员标注的结果不同。因此除了词义自身的因素外，标注人员自身的背景、标注时的心理状态等等也是不一致的重要来源。词义不确定性更关注于过程和原因，更强调目标词本身的多义以及和上下文的互动产生的词义无法准确消解的现象。二者各有侧重，互为补充，都体现了真实标注过程中存在的复杂场景。

3.2　词义相关度数据集概述

本节介绍与词义相关度判断相关的几个数据集。首先是英语通用数据集 WiC。在此基础上，衍生出了多语言、多语义等级的 OGWiC 数据集，以及进一步拓展了标注者不一致信息的 DisWiC 数据集。针对汉语中的兼类词现象，本节还将介绍一个多语义等级数据集 OGWiC_NV，以及与之相关的、包含标注者不一致信息的 DisWiC_NV 数据集。

3.2.1　WiC

WiC（Word-in-Context）[41]是一个经典的用于词义相关度评估的数据集。它是由学者 Mohammad Taher Pilehvar 和 Jose Camacho-Collados 共同提出的。该数据集将词义相关度任务定义为一个二分类的任务（公式 (3.2) 中的 K 值为 2）：对于一个目标词所在的两个上下文而言，标注者需确定两个用法相关或者不相关，相关则表示为“1”，不相关则表示为“0”。

表 3.1 列举了原始数据集中的几个例子，包括两例语义相关的例子和两例语义不相关的例子。语义不相关往往代表这两个词义的外延所指毫无关联，例如第一个例子中 *bed* 的含义为“河床”，与之相关的概念包括“河流”等，而第二个上下文中指得是“人睡觉的床”，它们所指完全不同。对于动词来说，例如第二个例句中的 *land*，它所涉及的宾语的语义角色在两个上下文也不同，第一个上下文是表示“登陆”的目的地，第二个上下文则表示“使之降落”的受事。而对于词义相关的例子来说，它们的语义完全相同。例如第三个例子中的 *beat* 都表示动词“打败”的含义，第四个例子中的 *air* 都表示名词“空气”的含义。

表 3.1　WiC 数据集中样例示例

目标词	上下文 1	上下文 2	标注
bed	There's a lot of trash on the bed of the river	I keep a glass of water next to my bed when I sleep	0
land	The pilot managed to land the airplane safely	The enemy landed several of our aircraft	0
beat	We beat the competition	Agassi beat Becker in the tennis championship	1
air	Air pollution	Open a window and let in some air	1

表 3.2　WiC 数据集不同划分的统计信息

划分	样本数	名词比例	动词比例	不同词数
训练集	7,612	48%	52%	1,374
验证集	702	62%	38%	672
测试集	1,366	59%	41%	1,227

WiC 中的数据仅包含英语，它们主要来源于三个电子词典中的例句：WordNet[32]，VerbNet[171]和 Wiktionary ①。并且使用 BabelNet[67]进行词义映射，确保它们之间语义的关联性。针对 WordNet 的语义区分太细的问题，作者采用剪枝的手段，将原有的语义映射到更粗粒度的集合中。目标词还增加了词类的信息，WiC 中的数据仅涉及名词和动词两个词类。

WiC 数据集被划分为训练集、验证集和测试集三部分。其中训练集用于模型参数的学习，验证集用于寻找模型最优的超参数，测试集则最终测试模型的性能。表 3.2 列出了不同 WiC 划分集的统计量，包括样本数、名词比例、动词比例以及不同的词数。可以看出在训练集中名词和动词的比例是比较平均的。数据集是完全平衡的，标注为“0”和“1”的样本各占一半，比例为 1:1。

数据集评估指标是准确率 ACC（accuracy），其计算公式如下：

$$\mathrm{ACC} = \frac{N_c}{N} \tag{3.5}$$

其中 N_c 是模型预测正确的样本，即预测标签等于实际标注的标签，N 是预测的样本数量。

由于 WiC 数据集具有通用性，它已成为评估神经语言模型语义理解能力的重要评测基准。因此，该数据集被广泛收录于领域内高度认可的模型评测榜单中，如

① https://www.wiktionary.org/

通用语言理解评测基准 GLUE[10]和超级通用语言理解评估基准 SuperGLUE[124]。

3.2.2　OGWiC

OGWiC 是一个评测多语言词义相关度的数据集。它的全称是依序分级的上下文词义相关度分类任务（**O**rdinal **G**raded **W**ord-**i**n-**C**ontext Classification）。这一数据集[37]由学者 Dominik Schlechtweg 发起，经由一系列不同国家的研究者针对不同的语言根据统一的格式和规范进行构建的。最终作为一个共享任务发布在国际计算语言大会 COLING 组织的工作坊 CoMeDi（Context and Meaning-Navigating Disagreements in NLP Annotations）上 ①，并获得了一致的认可和较高的关注度。

OGWiC 数据集的数据格式与 WiC 类似，都是同一个目标词出现在上下两个不同的文本中，多名标注者需要对其词义相关度进行判断。与 WiC 不同，OGWiC 共提供了四个判断等级，即从 1 到 4，分别表示“不相关”，“远相关”，“近相关”到“完全一致”，并且作者根据 Blank 的理论[172]将这几种相关程度与词义理论中的四种现象做了对应。从不相关到相关分别包括同形异义词（homonymy） ②、多义词（polysemy）、上下文变异（context variance）和单义词（identity）。

OGWiC 的设计弥补了 WiC 数据集存在三个局限：第一，语义相关性被简化为二值状态，忽略了实际语义相关性的连续性；第二，仅采用单一标注者标注，未考虑标注者之间可能存在的不一致性，从而可能引入数据偏差；第三，任务仅覆盖英语，未体现其他语言的特性和跨语言的词义差异。

以汉语中的“打”词为例 ③，例 (5) 中分别列举了“打”出现在的不同上下文，释义 (6) 则列出来“打”在各自例句中的词典释义，相同编号对应相同的例句和释义：

(5)　a.　打鼓
　　b.　打从今儿起，每天晚上学习一小时
　　c.　打架
　　d.　打交道
　　e.　打官司

(6)　a.　动词，用手或器具撞击物体

① https://comedinlp.github.io/

② Homonymy 有时也被称作“同音词”，即发音相同，词义截然不同的词，但书面写作的词形可能不同，中文中存在大量的同音词，例如“枇杷”和“琵琶”。同形异义更强调书面词形相同，但是语义毫无相关，例如“打这里来”和“打人”中的“打”。

③ 例句和释义均来源于《词典》[38]中“打”字的词条。《现代汉语词典》使用阿拉伯数字作为上标来区分不同词条的条目，例如例 (5-a) 中表动词“撞击义”的打和例 (5-b) 中表介词“起点义”的打；词典使用黑色圆圈和阿拉伯数字①②表示不同义项。

b.　介词，从

c.　动词，殴打

d.　动词，发生与人交涉的行为

e.　动词，发生与人交涉的行为

表 3.3　不同词义相关度标注类型对比

词义相关度	词义现象	标签	词典组织	举例
不相关	同形异义词	1	不同词条	“打”的撞击义(5-a)与表介词的起点义(5-b)
远相关	多义词	2	不同义项	“打”的撞击义(5-a)与殴打义(5-c)
近相关	上下文变异	3	同一义项	“进行诉讼”(5-d)和“进行社交往来”(5-e)
完全一致	单义词	4	单一义项	表介词的“打”(5-b)

表 3.3 展示了上述例句中所体现的四种不同类型的语义相关度标注。

“词义不相关”出现在同形异义词中。例如，例 (5-a) 中具有“撞击”原型义的“打”与例 (5-b) 中表示起点义的介词“打”，表达的是完全不同的概念，彼此之间既无词源上的关联，也无词义上的联系。这类词在词典中通常被分列于不同的词条下，可视为同一词形承载了多个无关的语义。

“词义远相关”则存在于多义词中，其不同义项之间往往存在引申等语义关联[170]。例如，例 (5-a) 中“打”的“撞击义”与例 (5-c) 中的“殴打义”便具有这种关联：殴打的过程往往涉及撞击等剧烈行为。这类义项在词典中通常被收录于同一条目下，分属不同义项。

“词义近相关”体现在同一词义在不同语境中的使用差异。这些用法的语义基本相同，但由于上下文不同，仍存在细微差别。例如，例 (5-d) 和例 (5-e) 中的“打”均表示“发生与人交涉的行为”，但前者指“进行诉讼”，后者指“进行社交往来”。由于语义高度接近，词典中通常将它们归入同一义项，不做进一步区分。

“词义完全一致”通常指单义词，包括专有名词或语义唯一的词。例如，介词“打”的语义唯一，仅在例 (5-b) 中表示起点义“从”，词典中只列出一个义项。

上述四种语义相关度由远及近，分别用数字 1 至 4 进行标注。

OGWiC 数据集起初是用于研究词义演变的，即目标词在不同时期的文本中词义的变化，因此数据集中的例句对主要包含不同时间跨度的文本。与 WiC 类似，该数据集最基本的数据单位是目标词，以及它所在的两个上下文。不同于 WiC 的是，针对每一条数据，收集者雇用了 2 到 4 名不等的标注者来标注语义相关性，因此每条数据会有 2 到 4 个标签。每个标签都是 1 到 4 中的某一个，其含义可以参考表 3.3。最终采用所有标注标签的中位数来表示最终的语义相关度标签。同时，

OGWiC 还是一个多语言数据集，包含汉语[173]、英语[174-175]、德语[174-179]、挪威语[180]、俄语[181-183]、西班牙语[184]和瑞典语[174-175]。对于 OGWiC 不同语言各自的部分，本文使用下划线加对应的语言缩略语来表示，例如 OGWiC_ZH 表示 OGWiC 数据集的中文部分①。除了提供目标词所在句子的索引外，英语还额外提供了词类信息，这也从侧面反映了词类对于区别词义的帮助。

表 3.4　OGWiC 数据集中英例示

目标词	上下文 1	上下文 2	标注
part	For though by fate we're forced to part, You never can't absent to me	Sometimes that knowledge seemed the worst part of my loss	1
bar	Hey, don't set the bar so high.	I was tipped off two days ago that I passed the bar.	2
attack	I have been in the country for three months; and have had several attacks of the fever.	I didn't mean it to come out that harshly, but I felt under attack here.	3
afternoon	We lazed along through the afternoon.	In the afternoon, I drove down to the square.	4
宰	《日本书记》载称，主管“日本府”的官吏“宰”常出入于“任那”	我们想只要车能修好，贵一点也只能由他宰了	1
下海	渔民下海捕鱼时，遭到越军的射击	他下定决心申请离职“下海”经商了	2
憋	很多农村干部憋了一肚子气	天然气随时有把井口憋破的危险	3
停业	全国邮件投递工作完全停顿，大批商店停业	当地税务机关办理开业或停业的税务登记	4

表 3.4 列举了英语和汉语中四个语义相关度标注等级的样例。

标注为“1”代表词义完全不相关。例如，英语词 *part* 在第一个上下文中作动词，意为“分离”；在第二个上下文中作名词，意为“部分”。二者在语法环境和语义上均无关联。汉语词“宰”在第一个上下文中指“官吏名称”，在第二个上下文中则表示“敲诈勒索”，同样在句法环境和语义上差异显著。

标注为“2”代表词义有一定关联但相距较远。例如，英语词 *bar* 在第一个上下文中表示“标准”，在第二个上下文中特指考试中的“等级”。二者存在关联，因为考试等级往往对应于某种标准，但其使用环境仍有较大区别。汉语词“下海”在第一个上下文中为字面义，指“到海中去”的动作；在第二个上下文中则发生语义引申，表示“开始经营商业”，是一种隐喻用法，二者具有较远的语义关联。

标注为“3”代表词义基本相同但存在细微差异。例如，英语词 *attack* 在两个上下文中都表示“攻击”，但第一个上下文指“病毒的攻击”，是非物理的直接接

① 英语、德语、挪威语、俄语、西班牙语和瑞典语的缩略语分别为：EN、DE、NO、RU、ES 和 SV

触；第二个上下文则指“物理上的攻击”。汉语词“憋”在两个上下文中都表示“强制止住”这一动作，但前者针对“呼吸”，后者针对“井口”，语义关系非常紧密。

标注为“4”代表词义完全相同。例如，英语词 *afternoon* 在两个上下文中均指“下午”这一时间段；汉语词“停业”在两个上下文中均表示“停止营业”这一动作，语义完全一致。

数据集还对原始数据和标注进行了清理过滤，包括过滤掉一些不相关的目标词。在标注方面，数据集去除掉了标注为“0”的样例，“0”在标注的时候表示无法确定；同时为了保证较高的标注一致率，去除掉了标注分数相差大于 1 的样例；对于计算出的中位数为小数的（例如 2 和 3 的中位数是 2.5）的样例也过滤掉了。

表 3.5　OGWiC 数据集的统计结果

类型	基础统计				上下文统计		标注数统计	
	词数	句数	句对数	词类	句长	目标词位置 (%)	均值	标准差
汉语	40	1537	16605	✗	54.71	55.54	2.00	0.00
英语	46	6238	9217	✓	32.30	50.26	2.31	0.55
德语	167	11137	13083	✗	39.82	54.38	2.80	0.98
挪威语	80	1642	6495	✗	48.35	50.28	2.32	0.47
俄语	272	22433	11440	✗	25.51	51.21	3.72	1.00
西班牙语	100	3645	6939	✗	86.26	49.29	2.28	0.50
瑞典语	44	5481	7673	✗	35.07	49.81	2.29	0.53
训练集	520	35772	47833	✗	45.98	51.60	2.55	0.89
开发集	77	5592	8287	✗	45.38	51.11	2.51	0.86
测试集	152	10781	15332	✗	46.63	51.91	2.56	0.87
全部数据	749	52113	71452	✗	46.00	51.54	2.55	0.88

表 3.5 展示了七种语言数据集的统计，包括：（1）基础统计：词数（不同目标词的数量）、句数（不同句子的数量）、句对数（用于训练的样本数量）以及是否包含词类信息；（2）上下文统计：句长（句子包含的词[①]数）和目标词的相对位置（目标词出现的位置占整个句长的比例）；（3）标注者数量的均值以及标准差。

可以看到仅仅只有英语提供了词类信息，其他语言均没有提供。同时，句对数的数量多少与句数以及词数的多少没有直接关系，存在目标词数量较少（如汉

① 对于印欧语而言，按照空格进行分词；对于汉语而言，这里统计的是字符，即汉语句长为单个句子中包含字的数量。

语），但句对数较多的语言，也存在词数少但句对数较少的语言（如西班牙语）。此外，这七种语言的数据集被分成训练集、开发集和测试集，遵循训练集的数量大于开发集，也大于测试集的惯例。

从上下文统计的信息来看，不同语言的句长差异较大，西班牙语提供最多的句长，而俄语的句长最少。它们目标词所处的位置都处于接近中间的位置，其中汉语、德语两个数据集目标词出现在中间偏后的位置，而瑞典语等数据集目标词则更靠前。

从标注者数量来看，汉语、英语这些高资源语言标注者较少，然而俄语数据集标注者较多。值得注意的是，汉语的所有数据的标注者都只有 2 人，这体现在 0 标准差上。这也揭示了汉语数据集的一个局限性，即汉语标注者较少。

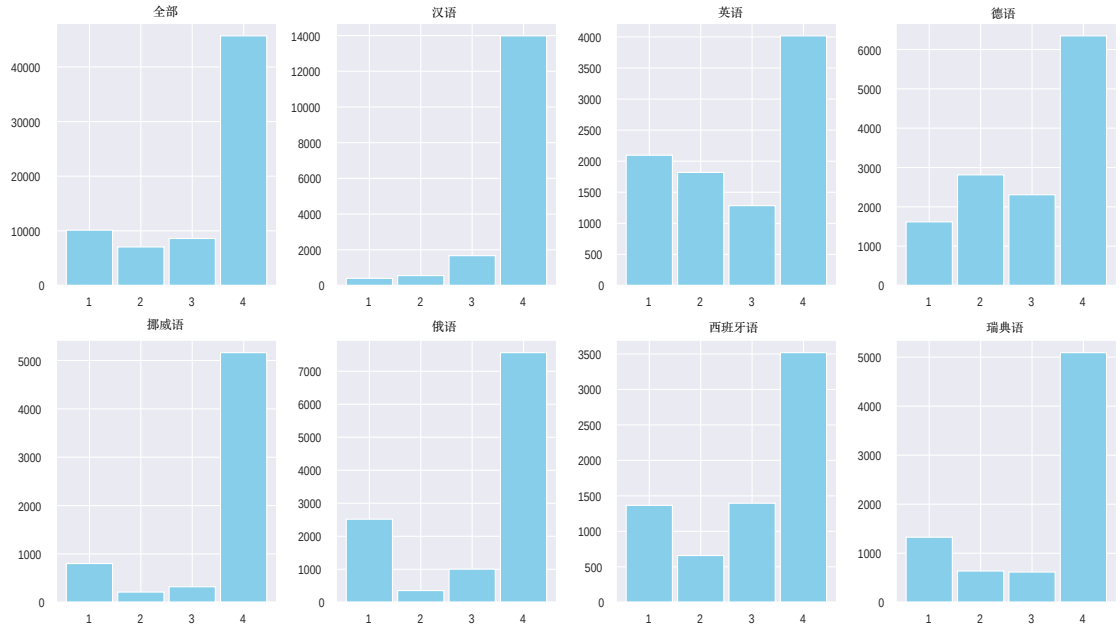


图 3.1　OGWiC 数据集中不同语言的标注值分布图

图 3.1 展示了全部数据集以及在不同语言下，OGWiC 数据集的真值标注的频率分布图，横轴包含四个相关度等级，纵轴反映了每个相关度等级下样本的数量。可以从图中发现以下几点结论：（1）所有语言都有样本不均衡的问题。它们绝大多数样本都偏向相关度为 4 的类别，也就是大多数示例的词义之间是完全一致的。这是由于在收集数据时，来自同一义项下的例句占了很大比重。有趣的是，尽管等级 3 和等级 4 有时区分不是很明显 ①，相比于“语义近相关”，标注者显然更倾向于“语义一致”。（2）不同语言的分布均衡程度不同。其中最不均衡的有汉语、挪威语和俄语；而较为均衡的语言有英语、德语。英语数据均衡可能与标注的词典中很充分的英文例句有关，因此更容易找到不同类别的句子对。

① 例如在上述例句（d）打交道和（e) 打官司中，它们处于同一义项中，很容易被判别为完全一致。

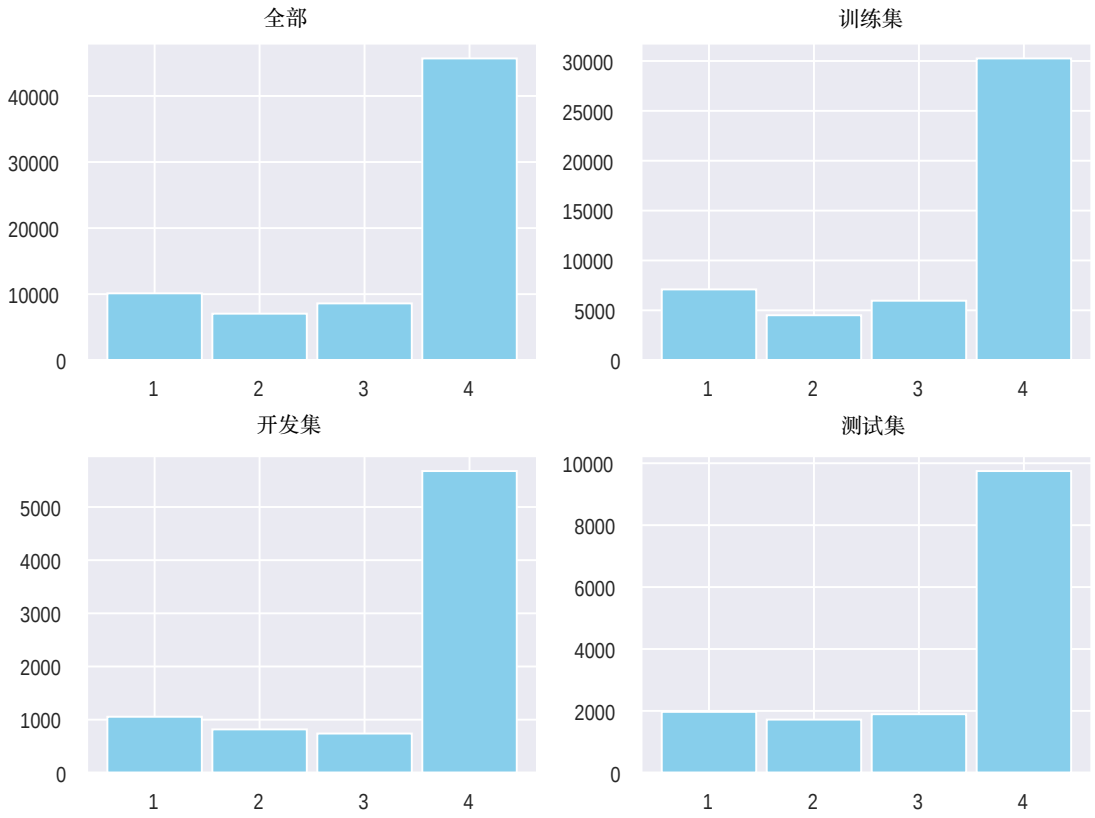


图 3.2　OGWiC 中不同数据集的标注值分布图

图 3.2 则分别展示了在不同的数据集，即训练集、开发集和测试集上的标注值分布。与语言维度类似，不同的数据集都存在数据不平衡，且都偏向最相关等级。不一样的是，不同数据集的分布之间相似度很大，都在其余三类相关度上比较均匀的分布，这也反映了该数据具有同分布（In distribution）的特性，因此模型较容易学习和泛化。

为全面评估模型在 OGWiC 上的表现，本文采用评价指标克里彭多夫 α 系数（Krippendorff's α）[185]。克里彭多夫 α 系数是一种广泛用于序数标注任务的一致性评价指标，其序数形式能够对偏离真实等级较大的预测给予更高的惩罚权重，同时能够有效消除随机一致性所带来的影响。相比于准确率（例如公式 (3.5)）等传统误差指标，α 在序数分类任务中被证明具有更强的稳健性和区分能力。

克里彭多夫 α 系数的一般形式定义为：

$$\alpha = 1 - \frac{D_o}{D_e} \tag{3.6}$$

其中，D_o 表示观测到的不一致性，D_e 表示在随机标注假设下的期望不一致性，α 的取值范围为 $(-\infty, 1]$，且数值越接近 1 表示预测结果与人工标注之间的一致性越高。

3.2.3　DisWiC

DisWiC 是与标注不一致相关的多语言词义相关度数据集，其全称为上下文词义相关度中不一致排序任务（**Dis**agreement in **W**ord-**i**n-**C**ontext Ranking），它旨在评估模型对于标注不一致性的预测。

DisWiC 的数据集格式、包含的语言与 OGWiC 一致，都包含目标词与其出现的两个上下文，以及不同的标注者的四个相关度等级的标注。不同于 OGWiC 使用中位数来表示相关度的一个平均状态，DisWiC 使用标注者的不一致程度来作为真值标签，不一致程度使用平均成对绝对差异 MPD（mean pairwise absolute judgement differences）来计算：

$$\mathrm{MPD}(\mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{(y_1, y_2) \in \mathcal{Y}} (|y_1 - y_2|) \tag{3.7}$$

其中 $\mathcal{Y}$ 表示所有的标注结果，而 $y_i \in \mathcal{Y}$ 是其中第 i 个标注结果，该公式计算任意成对的标注结果差距的平均值。MPD 不一致度评分是一个连续值，当所有指标相等时，达到最小值 0，当任意两个指标差距越大时，达到最大值。

表 3.6　DisWiC 数据集中英例示

目标词	上下文 1	上下文 2	标注
体系	建议苏法协商促成欧洲安全体系	关于建立国家创新体系述评	0 (4,4)
下海	渔民多、渔船多、渔具多、妇女多、干部下海多	三亚市即使在严冬季节游人也可下海游泳	1 (4,3)
急	风雨交加，江水猛涨，浪大流急，一只木船断缆飘走	急着剪开，捏出几粒	2 (3,1)
拐	锻工们很支持他们，给他们打了个新鞋拐	今年的春节晚会，我觉得最精彩的节目，当是小品《卖拐》	3 (4,1)
afternoon	I'd landed in early afternoon, local time.	Received a long awaited oil delivery yesterday afternoon and the heat came up shortly after 4 P.M.	0 (4,4)
stab	There was not a taxi to be had, or a light to be seen except for the stab of searchlights and the flashes of anti-aircraft batteries.	Emmeline felt a sharp and sudden stab of envy over that, an uncharitable emotion that she'd been able to subvert when she was sober.	0.8 (2,2,2,3,1)
bar	Grab the bar with an overhand grip and hold it at	These were composed of a box 9 feet long, and about the same width and depth, formed by light bars, with stronger bars at the corners.	2 (2,4)
attack	The formless couch was beginning to give me a sensory deprivation attack.	And yet, despite the frequency and deadliness of their attacks, almost nothing is known about individual bombers.	3 (4,1)

表 3.6 展示了 DisWiC 中的一些中英文的示例，列出了目标词以及对应的成对上下文，标注处中的括号里显示了每一个标注者的标注分数，显示在括号前面的值采用公式 (3.7) 中的 MPD 指标进行融合。表 3.6 列出了两种语言中由最一致到最不一致的样例。其中最一致的例子都是一些单义词，且上下文环境较为类似。例如英语词 *afternoon*。而不一致的样例都是一些多义词，且语义界限较为模糊，与之搭配的语义角色也更加丰富多样。例如在最不一致的中文“拐”中，尽管两个上下文都可以理解为物理工具，根据小品的剧情，《卖拐》中的“拐”还有一种“拐卖、忽悠”的引申义，从而使得不同标注者产生了不同的理解。这也体现了语言单元的歧义性。

在数据方面，与 OGWiC 数据集相比较，DisWiC 数据集不再要求标注者评分之间最多相差 1，这是因为该任务是一个预测标注者不一致的任务，无需不一致率达到一个较高的水平。因此，该任务的数据量要多于 OGWiC 数据集。

表 3.7　DisWiC 数据集的统计结果

类型	基础统计				上下文统计		标注数统计	
	词数	句数	句对数	词类	句长	目标词位置 (%)	均值	标准差
汉语	40	1585	29472	✗	54.71	55.54	2.00	0.00
英语	46	7721	17562	✓	32.30	50.26	2.31	0.61
德语	168	14854	21612	✗	39.82	54.38	2.82	1.05
挪威语	80	1689	8717	✗	48.35	50.28	2.31	0.46
俄语	272	35532	18403	✗	25.51	51.21	3.82	1.04
西班牙语	100	3939	13556	✗	86.26	49.29	2.23	0.48
瑞典语	44	6542	12630	✗	35.07	49.81	2.35	0.62
训练集	521	49279	82177	✗	45.98	51.60	2.55	0.92
开发集	77	7665	13081	✗	45.38	51.11	2.54	0.92
测试集	152	14918	26694	✗	46.63	51.91	2.56	0.93
全部数据	750	71862	121952	✗	46.00	51.54	2.55	0.92

表 3.7 给出了针对数据集 DisWiC 的统计结果，数据量类型的维度与表 3.5 中的相同，包括语言（数据集）类型、词形数、句数、句对数、是否包含词类信息，以及标注者数量的均值和方差。与 OGWiC 数据集的统计量相比，在目标词词数上二者数量相当，DisWiC 额外拓展了更多的句对示例，因此具有更丰富的不一致性分布。

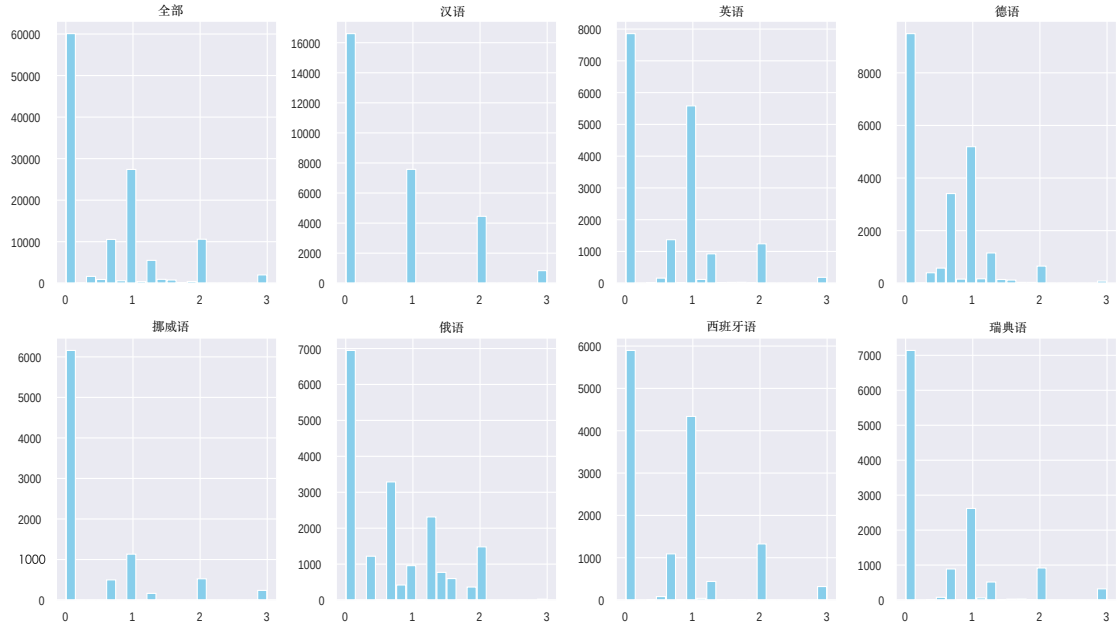


图 3.3　DisWiC 数据集中每种语言的 MPD 分布直方图

图 3.3 展示了 DisWiC 数据集中每种语言的 MPD 分布直方图。其横轴表示不一致度 MPD 评分，其最小值为 0，最大值为 3，它们被均分成了若干更小的区间。纵轴表示落在每个区间中的样本数量。从图中可以观察到所有的语言都偏向不一致度为 0，即标注者之间达成一致的用例更多，然而仍有很多样本分散在其他区间。分布更加均匀的语言包括俄语、德语、英语。差异更明显的包括汉语、挪威语。

图 3.4 则从不同数据集的角度展示了样本的分布，其横轴与纵轴的含义与语言维度 3.3 的分析是一致的。每个数据集中的分布同样偏向最不相关的区间，同时不同数据集中的分布是高度类似的，这与 OGWiC 数据集（参见图 3.2）中所反映的特点是一致的，表明它们都处于同分布中，符合机器学习中常见的独立同分布假设。

用于评估模型在 DisWiC 数据集上的不一致程度的量化指标是斯皮尔曼系数（Spearman's ρ）[186]。公式表示如下：

$$\rho = \frac{\mathrm{cov}\left(\mathrm{rank}(y_p),\ \mathrm{rank}(y_q)\right)}{\sigma_{\mathrm{rank}(y_p)}\,\sigma_{\mathrm{rank}(y_q)}} \tag{3.8}$$

其中，y_p 表示模型得到的不一致程度，y_q 表示人工标注的不一致程度，$\mathrm{rank}(x)$ 秩函数表明 x 在当前样本序列中处于第几位，$\mathrm{cov}(x, y)$ 则计算输入 x 和输入 y 的协方差，$\sigma(x)$ 则计算方差，最终可以得到二者在排序上的相关度 ρ。

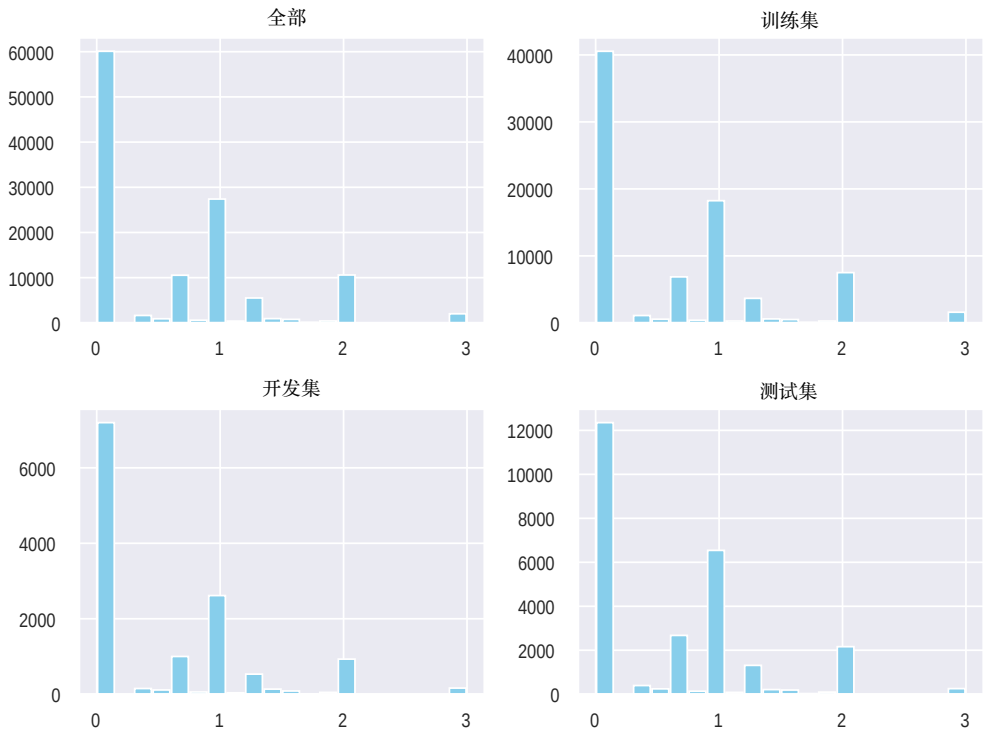


图 3.4　DisWiC 中不同数据集的 MPD 分布直方图

3.2.4　OGWiC_NV

为了进一步探究词类对于汉语多义词理解的影响，本文专门构建了汉语兼类词词义相关度数据集。这个数据集与多语言数据集 OGWiC 中的数据类似，包含数据格式、评估指标，因此可以视作 OGWiC 中的汉语部分（OGWiC_ZH）的延伸，因此将之命名为 OGWiC_NV，其中 NV 表示该数据集的兼类功能仅考虑名词（**N**oun）和动词（**V**erb）。

OGWiC_NV 数据集与 OGWiC 数据集中的中文部分主要的不同点在于：（1）OGWiC_ZH 仅包含 40 个汉语目标词，尽管每个目标词对应的例句较多，但是对于词的覆盖度较少。OGWiC_NV 则收集了上百个目标词，涵盖范围更广。（2）OGWiC_ZH 中采用的例句来自不同时间段的文本，目标是测试词义演变，而 OGWiC_NV 的例句既有历时的也有共时的语料文本，类型更加多样。（3）OGWiC_ZH 中的目标词没有词性标注，而 OGWiC_ZH 参考词典中的标注信息，不仅带有词性，还专门设置了名词和动词两种不同的语境。（4）OGWiC_ZH 中的每条数据的标注者人数较少，且缺乏标注者更多的背景信息，OGWiC_NV 则通过自行雇佣更多的母语标注者，不仅数量上占优，同时也可以获取更多标注者自身的信息。

根据《现代汉语》[17]中对于汉语兼类词的定义，词的兼类是某个词经常具备两类或几类词的主要语法功能，且不同的语法功能出现在不同的场合，同时兼类

词一定要读音相同，词义有联系。由于名词动词兼类是汉语中较为普遍的一种兼类模式，本文重点考察这一类名动兼类情况。

名动兼类词以及例句的选择来源于《现代汉语词典》[38]，选择方式如下：（1）选择同一词形条目下兼有名词和动词用法的词以及该词所带的例句，不同的词类用法往往处于不同的义项中①。例如“锁”在《现代汉语词典》中具有名词和动词两个用法，即“用具”相关的释义 (4-a) 和“关住或者拴住”相关的释义 (4-b)，它们分别列在同一个词目“锁”的两个独立词条中，因此符合这里的选择条件。其后附带的例句作为该释义下“锁”的上下文被收集到数据集中。

之后开始过滤数据，去除掉不完全匹配的目标词，缺少例句的词，文本自带的噪声数据等。然后得到每个目标词对应的不同语义的两个上下文用例，它们分别为名词或者动词的词性。与 OGWiC 数据集类似，在标注者标注完成后需要进一步筛选一部分数据，请参考后文标注过程。

（1）标注过程

需要对 OGWiC_NV 数据集的每一条数据进行人工标注，本文招募了 69 名汉语母语者对成对数据进行词义相关度的标注，标注过程与 OGWiC 的过程一致。其目标仍然是给出表格 3.3 中语义相关度的四个等级。图 3.5 展示了标注指南②。标注指南给出了四个等级的定义，并提供了相应的例句和解释。标注者在正式标注之前需要仔细阅读标注指南，以便正确掌握标注任务的流程。需要注意的是，指南中并未告知标注者数据来源、数据的具体目的，也未提及词类对语义的影响等信息。同时，例句的选择并未严格针对兼类词，以避免干扰标注者的判断，使其更倾向于依赖母语直觉进行标注。此外，为了进一步防止标注者发现潜在的规律从而影响判断，成对的名词和动词用例在标注过程中被随机打乱顺序，即部分实例为名词用法在前、动词用法在后，部分实例则为动词用法在前、名词用法在后。

标注采用填写调查问卷的形式，每位招募者需评估约 40 个句对，不同的标注者之间可能评估同一份内容，以方便后续更细致地评估不一致性。附录 A 列出了问卷的部分示例。在对于相关性的评估之外，问卷也询问了标注者的个人信息，包括学历、专业、年龄段、家乡等，方便后续结果的分析。每份问卷平均耗时约 20 分钟。

（2）标注者背景信息分析

我们首先统计了参与问卷的所有标注者的背景信息。图 3.6 分别展示了标注者的学历分布（a）和专业分布（b）。其中在学历分布中，本科及以上占据了绝大

① 同一词形对应于兼类词定义中的读音相同且词义有关联，排除同形异义词；不同的义项体现了兼类词是多义词的一种特殊情况。

② 问卷的指引中并没有说明表格 3.3中的每一种语义相关度相对应的词义现象以及词类信息，而是更强调母语者使用语感来判断语义的关联性而不是考虑语言规则或者现象。

中文词义相关度标注指南

任务说明

本任务要求标注者判断目标词在两个给定句子中的语义相关度，并根据评分标准选择相应的标签。问卷中收集的所有信息仅用于科研目的，且个人信息将被严格保密。

评分标准说明

标注时请主要依据自身的母语直觉进行判断，以下示例仅用于辅助理解评分标准。

（1）完全一致

说明：目标词在两个句子中的含义完全相同。

例句 1：我国制炼熟桐油已数千年，常用于涂饰房屋、车辆和船只，农村多使用纸袋竹篓【包装】。

例句 2：广西贵县糖厂的工人正在【包装】白砂糖。

解释：两个句子中的“包装”均表示“封装打包”这一具体动作，因此判定为完全一致。

（2）近相关

说明：目标词在两个句子中的含义高度相似，但存在细微差别。

例句 1：他吃了面片和馄饨，根本【消化】不了。

例句 2：要认真学习并加以【消化】和掌握，才能有所提高。

解释：“消化”在例句 1 中指生理过程，在例句 2 中指对知识的理解与吸收，后者可视为前者的语义引申。

（3）远相关

说明：目标词在两个句子中的含义存在一定关联，但差异较为明显。

例句 1：对乙肝【病毒】携带者的歧视源于对相关知识的缺乏。

例句 2：计算机系统遭受【病毒】攻击的次数逐年增加。

解释：两处“病毒”分别指生物学意义上的病原体与计算机领域中的恶意程序，二者具有部分相似属性，但语义差异较大。

（4）不相关

说明：目标词在两个句子中的含义完全不同。

例句 1：全区【机制】糖的生产能力将大幅提升。

例句 2：在打破种子休眠过程中，呼吸【机制】会受到影响。

解释：两处“机制”分别指“机械制造”与“事物运作的规律和方式”，语义相关性较低。

（5）无法判断

当句子语义不清、上下文不足或目标词含义模糊时，无法可靠判断其语义相关度。

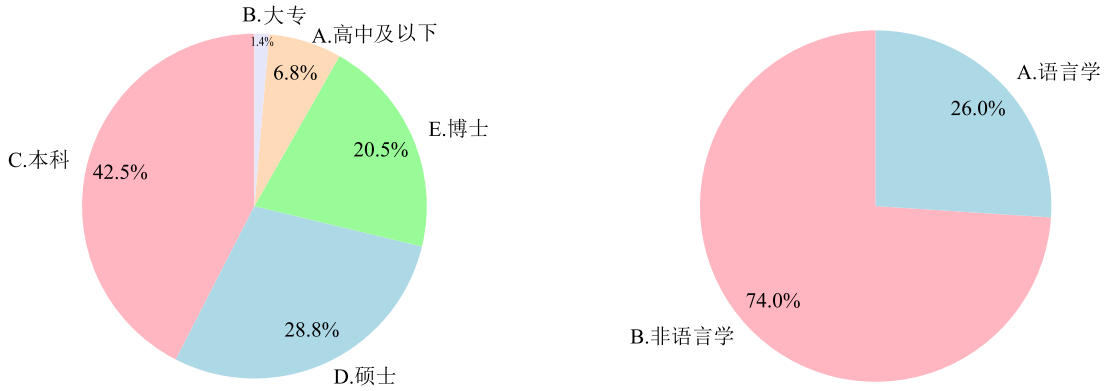
标注要求

- 在标注前请完整阅读两个句子。
- 目标词在句中以【】标出。
- 请在整个标注过程中保持评分标准的一致性。
- 如遇问题，请及时反馈至专用邮箱。

图 3.5　中文词义相关度标注指南

多数，绝对数值超过整体的九成，而拥有本科学位的比例占比最高。这表明标注者的整体受教育水平较高，对于词典中出现的单词较为熟悉。在标注者专业方面，非语言学的人数大约占据总人数的四分之三，约有不到三成的标注者来自于语言学专业。值得注意的是，本标注任务无需任何显式的语言学知识，更强调母语者的语感。

图 3.7 则绘制了标注者的年龄段人数分布，其中 A 类型表示 18 岁及以下；B



(a) 标注者学历分布　　(b) 标注者专业分布

图 3.6　标注者背景信息统计

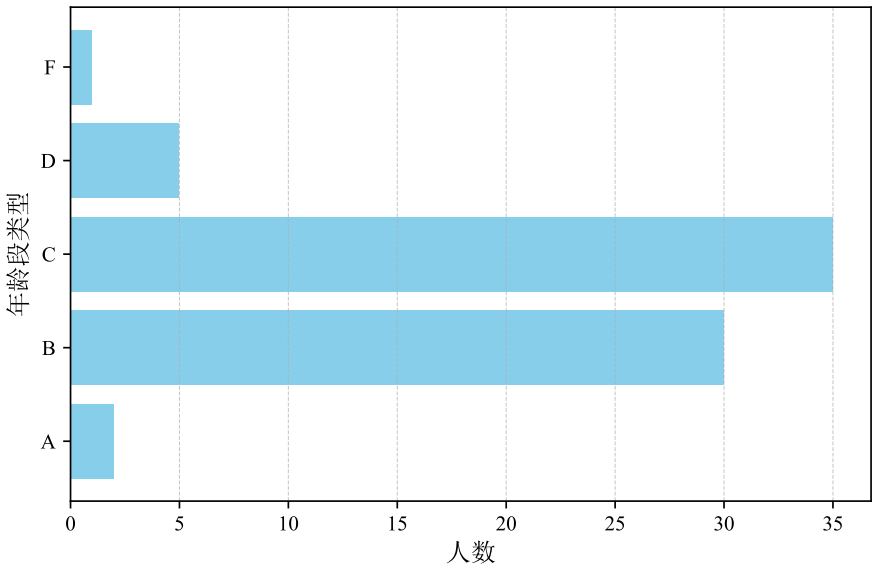


图 3.7　标注者年龄段分布

类型表示 19 岁到 25 岁；C 类型表示 26 到 35 岁；D 类型表示 36 到 45 岁；E 类型表示 46 岁到 55 岁；F 类型则表示 56 岁及以上。可以观察到样本中缺少 E 类型的人数，占多数的是 BC 类型，即 19 岁到 35 岁的青中年，这与学历方面的信息是吻合的。标注者年龄跨度较大也体现了标注者的多样性。

图 3.8 则给出了标注者所在的家乡省份的人数分布。标注者的家乡所在地超过二十个不同的省份，体现了地理空间分布的多样性。同时在山西、广东、辽宁、山东等地人数更为显著。家乡的多样性可能会导致标注者对于个别样例的判断有差异，然而需要强调的是，本标注数据来源于通行的字典，其基本语义和例句均以

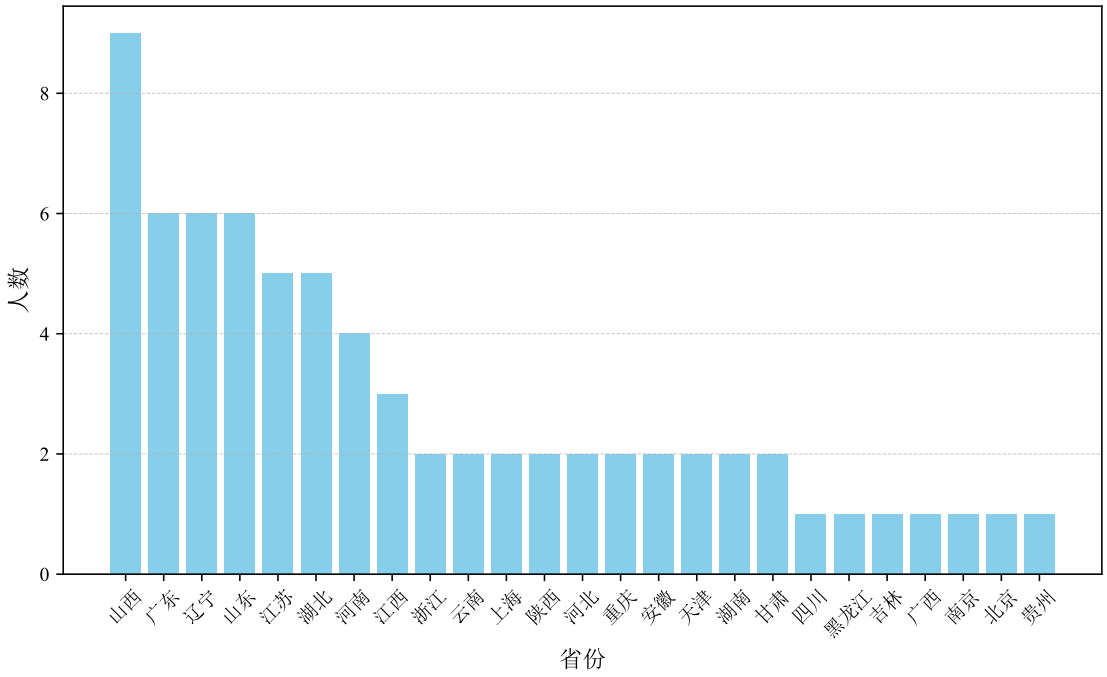


图 3.8　标注者家乡省份分布

现代普通话为基础，因此绝大部分样例的判断不应受到所在家乡方言的影响。

（3）标注后处理

标注完成后，人工对每一份问卷的个别问题进行审查，确保标注完成的质量。对于标注了“无法判断”的问题，选择将其丢弃，通过人工审查，这部分问题涉及到的目标词都较为生僻，标注者多因不认识而选择“无法判断”。由于 OGWiC_NV 数据集需要模型估计平均的准确率，故采用所有标注的中位数来作为最终的标签。如果中位数刚好取出是小数，则向下去整。同时为了保证标注者间的一致性，当任意两个标签之间的值大于 1 的时候就舍弃掉这个样例，这与 OGWiC 数据集的标注规范是相同的。

表 3.8 列出了一些 OGWiC_NV 中的实例，其中第一个上下文和第二个上下文中的目标词词类分别动词为和名词，标注显示的是原始标注的中位数对应的相关度类型，例子展示的均为标注者达成了完全一致的典型例子。标注为“不相关”的例对中，“料理”在第一个上下文中指“处理、打理”的意思，而在第二个上下文中表示一种食物的通称了。“形容”在动词用法中表示“描述、表达”之意，而在名词用法中则指的是“面容、脸色”。这个例子中的“料理”和“形容”的名词用法都有些文言色彩了，它们可能在词源上和对应的动词用法有关联，因此词典并没有把它们视作同形异义词收录在不同的词条中，但是众多标注者从语义上均给出了“不相关”的判断。

在“远相关”的举例中，“表情”在做动词时，可以表达“从面部或姿态的变

表 3.8　中文兼类词数据集 OGWiC_NV 中的示例

目标词	上下文 1	上下文 2	标注
料理	事情还没料理好，我怎么能走	韩国料理	不相关
形容	他高兴的心情简直无法形容	形容憔悴	不相关
表情	这个演员善于表情	脸上流露出兴奋的表情	远相关
当道	奸佞当道	别在当道站着	远相关
关系	石油是关系国计民生的重要物资	这一点很有关系	近相关
装备	这些武器可以装备一个营	现代化装备	近相关
分晓	且看下图，便可分晓	究竟谁是冠军，明天就见分晓	完全一致
风传	村里风传，说他要办工厂	这是风传，不一定可靠	完全一致

化上表达内心的思想感情”，在做名词时则表达具体的” 表现在面部或姿态上的思想感情“，它们有着较强的关联，然而标注者较为一致地认为它们关联度较弱，可能和“表情”的动词用法在现代汉语中出现的频率较低有关。另一则例子“奸佞当道”中的“当道”有固定的贬义含义“掌握政权”，它与名词字面义用法“路中间”相关度较低，尽管二者也有语义引申的关联。同时前者的当道可以视作是惯用语中的一个部分，因此语义更加凝固和独立。

在“近相关”的两则例子中，“关系”和“装备”在不同上下文都是类似的含义，分别表示“重要性、关联”和“武装”，但是作为动词和名词用于不同的上下文语境中。而在“完全一致”的例子中，语义则没有任何区别。“分晓”表示“清晰明了”，“风传”表示“辗转流传”，只不过动词用法是这个动作，而名词用法则是动词所指的事物或其本身。在汉语中近相关和完全一致的用法是占据多数的，也可以在之后的数据统计量中直观看到。

表 3.9　OGWiC_NV 中数据集的统计结果

类型	基础统计				上下文统计		标注数统计	
	词数	句数	句对数	词类	句长	目标词位置 (%)	均值	标准差
训练集	211	478	284	✓	8.45	51.14	4.05	1.49
测试集	106	230	122	✓	8.54	52.16	4.29	1.56
总共	293	665	406	✓	8.49	51.65	4.12	1.51

（4）数据统计量分析

表 3.9 展示了 OGWiC_NV 数据集在特定统计量上的数值分析。这些统计量包

括了基础统计例如：词数、句数、句对数（样例数）以及是否标注了词类，也包括上下文统计例如句长和目标词的相对位置，以及标注者标注的信息。OGWiC_NV 数据集选取 70% 的随机样本作为训练集，30% 的剩余样本作为测试集，训练集用于训练模型或者寻找合适的阈值，测试集用于评估模型。由于该数据集并非用于比赛场景，故不单独设立开发集。

相比中文部分的 OGWiC_ZH 数据集（见表 3.5），OGWiC_NV 数据集在很大程度上丰富了不同的中文目标词，增加了近 7 倍的数量。但每一个词对应的例句对相对较少，平均一个汉语词仅对应 1-2 条例句。另一个明显的优势是，OGWiC_NV 数据集标注了词类信息。从句长来看，OGWiC_NV 数据集的平均长度更短，仅为 OGWiC_ZH 数据集的六分之一到七分之一，这是由于相较语料库中的句子，词典中的例句为了简明扼要，往往长度很短。另一方面目标词的位置相较于 OGWiC_ZH 中的也更加靠前。从标注者数量来看，OGWiC_NV 数据集招募了更多的标注者，平均是 OGWiC_ZH 的 2 倍，可以更加全面地反映标注者对词义的理解面貌。

图 3.9 展示了 OGWiC_NV 不同数据集的标注分布，可以看出整体分布是偏向语义相关的，这和 OGWiC 中的趋势是类似的，但是第三等级紧密相关要更多一些，这表明尽管兼类词作为不同的词类出现在上下文中，标注者并没有因此而断定它们词义不相关或者远相关 ①。

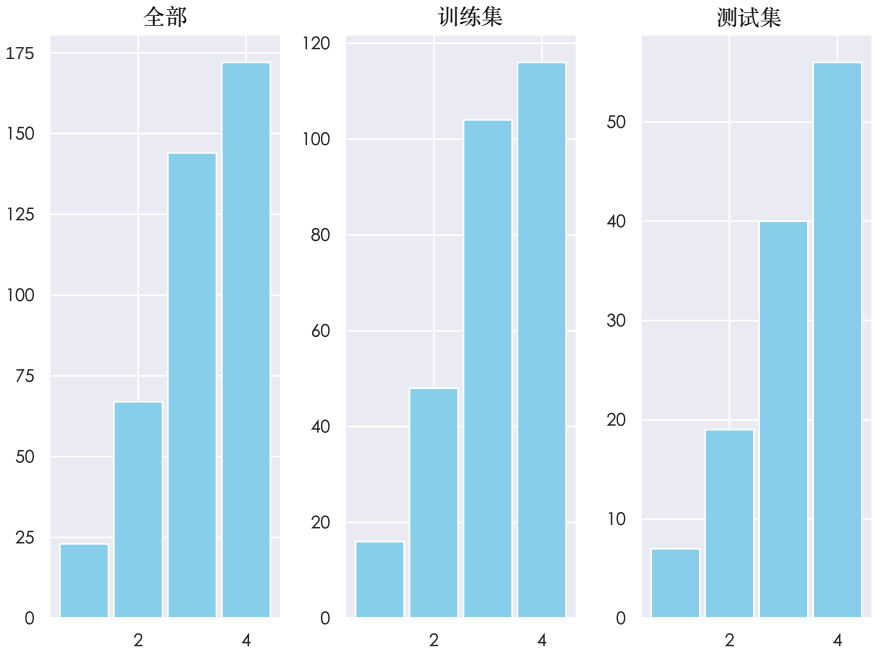


图 3.9　OGWiC_NV 不同的数据集的标注值分布图

① 在英语等印欧语言中，词类往往影响词义在词典中的编排方式[187]。一个突出的表现是：英语词典通常在同一词条下按词类分立义项板块，不同词类的用法被明确区分；而汉语词典则倾向于将同一词形的多种词类用法归并在一个词条下，按义项顺序排列，较少强调词类区分。

3.2.5　DisWiC_NV

类似于 DisWiC，汉语兼类词词义相关度数据集也旨在预测兼类词的标注不一致程度，本文将其命名为 DisWiC_NV。它与 OGWiC_NV 的数据来源类似，但存在以下几点区别：（1）DisWiC_NV 的标注值是众多标注的离散程度值，即平均成对绝对差异 MPD（公式 (3.7)）；（2）DisWiC_NV 的数据由于没有“不同标注者的评分之间相差不能大于 1”的限制，因而数据量更多；（3）模型的评估指标是模型所得的不一致程度与标注者的不一致程度的斯皮尔曼系数（公式 (3.8)）。这些设定与 DisWiC 基本保持一致，因为它们都是与评估不一致程度相关的任务。

表 3.10 给出了针对数据集 DisWiC_NV 的统计结果，数据量类型的维度与表 3.5中的相同，包括语言（数据集）类型、词形数、句数、句对数、是否包含词类信息以及标注者信息。训练集和测试集的数量比例也是按照样例数 7:3 的比例分割开的。可以看出 Dis WiC_NV 数据集从数量上看是 OGWiC_NV 数据集的两到三倍。标注者信息与 OGWiC_NV 是类似的。与 DisWiC（表 3.7）相比，汉语目标词的数量扩充了 10 倍之多，标注人数扩充 2 倍之多，足以体现出我们收集这一数据集的优势。同时词类信息的引入可以便于做更进一步的实验分析。

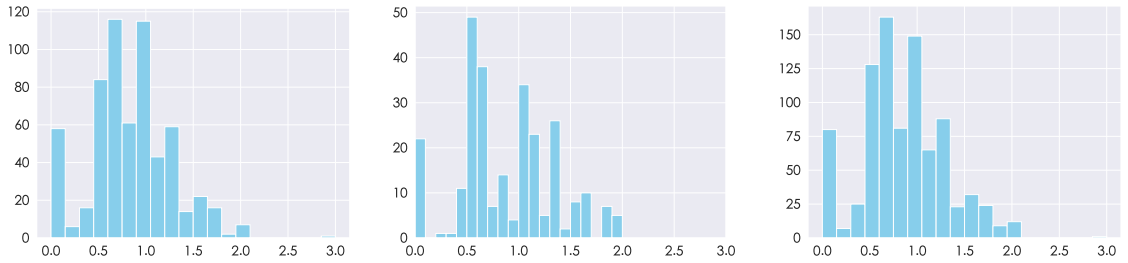
表 3.10　DisWiC_NV 不同数据集的统计结果

类型	基础统计				上下文统计		标注数统计	
	词数	句数	句对数	词类	句长	目标词位置 (%)	均值	标准差
训练集	427	963	620	✓	8.42	52.20	4.91	1.95
测试集	204	463	267	✓	8.80	53.39	4.94	2.01
总共	631	1426	887	✓	8.61	52.79	4.92	1.97

图 3.10 则展示了 DisWiC_NV 不同数据集的标签分布，可以看出数据整体呈现类正态分布，即中间的数值较高，两边的数值相对较少，即标注者之间存在一定的分歧，但是并没有非常大。同时，与之前的数据集类似，测试集和训练集之间存在相似的分布，体现了数据内独立同分布的特点。与表 3.3中展示的 OGWiC_ZH 中文数据集进行对比，可以发现 DisWiC_NV 数据集的取值更加多样，这也是由于更多的标注者所导致的，使得歧义性的评估更加有难度。

3.2.6　数据集选取原则

针对本文研究的多语言词义相关度任务，本文收集并拓展了已有数据集，为后续实验研究提供数据基础。所选五类数据集的选取依据以下三个原则：



(a) 训练集　　(b) 测试集　　(c) 全部数据

图 3.10　DisWiC_NV 不同数据集的不一致评分分布

（1）经典性：数据集具备普适性与通用性，是评估各类大语言模型时研究者所熟知、广泛使用的基准数据集。

（2）多样性：数据集内部的数据分布尽可能多样，涵盖语言种类、标注者数量与分布、语料覆盖范围等多个维度。

（3）代表性：数据集的构建需考虑当前语言在形式上的特殊性，从而能够更好地反映语言的实际使用情况。

其中，经典性确保了数据集的评估结果可与其他研究进行横向比较，且数据集的科学性已得到大量研究验证。WiC 数据集是这一原则的典型代表，作为大语言模型评测基准 SuperGLUE[124] 的组成部分，被广泛用于评估模型的语义理解能力。多样性原则强调数据集应尽可能贴近真实分布，包括多种语言、多元标注视角以及多样的语料来源。OGWiC 和 DisWiC 基于的标注语料充分体现了这一特性：其语言范围涵盖印欧语系、汉藏语系等多个语系的不同语言，且每条数据至少由两名标注者共同标注。代表性原则则突出数据集应能反映特定语言的内在特性。针对 OGWiC 和 DisWiC 中汉语部分未标注词类的不足，本文进一步收集了中文兼类词数据集 OGWiC_NV 和 DisWiC_NV，以有效体现汉语中词类相关的语言特征。

3.3　本章小结

本章的主要内容围绕词义研究展开，分为任务介绍与数据准备两个部分。第一部分详细阐述了词义相关度判断这一核心任务；第二部分则系统介绍了支撑这些任务的数据集，其中既包含现有的公共资源，也重点说明了本文为汉语兼类词研究所构建的新数据集。综上所述，本章为后文的模型评估与实验分析提供了必要的理论框架和数据基础。

第 4 章　多语言词义相关度任务评估

4.1　本章导论

词义相关度判断旨在评估同一目标词在不同上下文中词义的关联程度。对词义相关度的精确判断不仅关系到传统的词义消歧[43,140]、命名实体识别[188]等词汇语义相关任务，也对机器翻译[189]、问答系统等下游应用至关重要。此外，词义相关度评测还能揭示模型内部表征的几何结构、支持语义聚类与表示学习分析，从而提升模型的可解释性。

本章针对公开的多语言词义相关度数据集 OGWiC 以及本文收集的汉语兼类词数据集 OGWiC_NV，对多种神经网络架构的神经网络语言模型进行了评测：既包括在多语语料训练的基于编码器架构的模型，也包括基于解码器的生成式模型（以不同参数规模的大语言模型为代表）。本章主要关注模型内部表征对词义相关度的表达能力，通过几何探测的方式对表征进行直接评估。其中，针对编码器模型，直接提取目标词位置的向量表征；针对解码器模型，则设计多样提示词模板，引导模型生成与词义相关的向量表征。

本章的结构安排如下：第 4.2 节介绍了基于模型内部表征的评估方法，包括表征提取方式、探测方法、向量后处理方式以及基于提示词的探测方法；第 4.3 节则介绍了词义相关度任务的实验设定，包括数据集介绍、模型选择和超参数设置。第 4.4 节详细介绍了实验结果和实验分析，包括准确性结果的分析、消融实验以及重点探讨了词类信息对于中英两个数据集的影响。最后一节是本章小结。

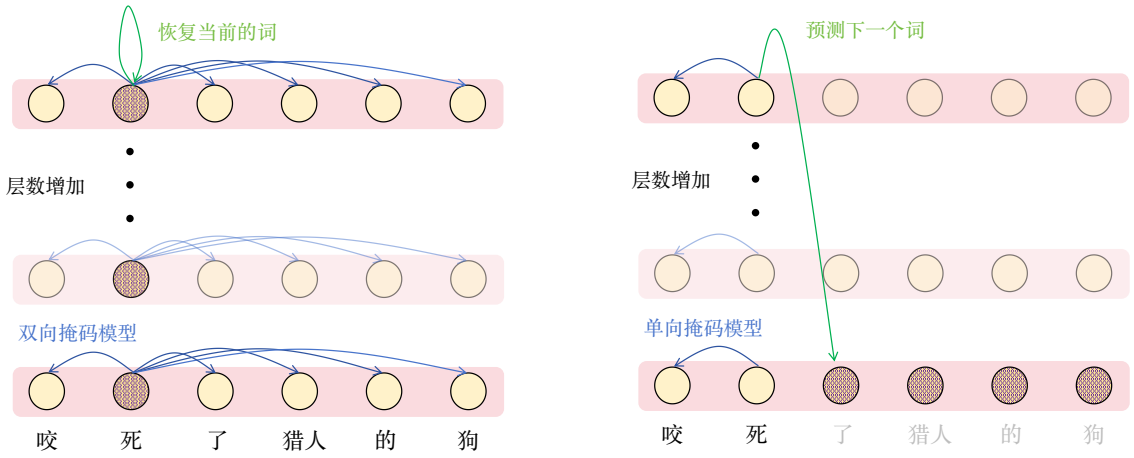
4.2　基于模型内部表征的评估方法

本节介绍针对模型内部表征的评估方法，包含表征提取方式、几何探测方法、向量后处理手段以及基于提示词的探测方法。

4.2.1　表征提取方式

表征提取方式与模型的架构息息相关。模型的架构考虑两类，一类是编码器模型，一类是解码器模型。如图 4.1 所示，编码器模型采用“双向掩码预测”的预训练目标，其对于当前词的编码（理解）过程和解码（预测）过程是统一的，且都针对当前词，因此模型在当前词的高层表征即可以代表当前词的词义。而解码器模型则采用“预测下一个词”的预训练目标，其对于当前词的编码（理解）和之后

的解码（预测）是不一致的，即编码针对的是当前词，而解码针对的则是它可能的下一个词。这种不一致使得在解码器模型中的最高层表征不能代表当前词的语义。本文在第 5 章节针对这两类不同的架构做更深入的分析。



(a) 编码器模型　　(b) 解码器模型

图 4.1　编码器模型和解码器模型的架构对比图

表 4.1　表征提取策略以及示例

策略名	示例
基础	river **bank**
上下文重复	river bank river **bank**
前词	**river** bank
提示语	In this sentence “river bank”, “bank” means in one word **:**

因此，对于编码器模型来说，表征选择的方式为当前词的最高层表征。而对于解码器模型来说，仅选择当前词最高层表征往往会得到比较差的结果，一方面是由于，当前词的表征仅可以获得其前边的上文，而理解词义是仍需要下文；另一方面，不同于重构当前词的编码器模型，解码器模型的目标所针对的并非是当前词，因此可能在高层无法有效反映当前词的词义。针对这些情况，表 4.1 列出了本文使用的针对解码器模型的几种策略：

（1） 基础：提取当前目标词的最高层表征。

（2） 上下文重复: 重复上下文，并提取第二个上下文中目标词的表征。

（3） 前词：提取上下文之前一个词的表征。

（4） 提示语: 设置提示语，并提取提示语末尾的冒号的表征。

其中重复策略考虑到重复的操作可以将下文的内容提前展现出来，从而更充分地利用上下文，这也是一个提取解码器模型语义的常见策略[190]；策略 Prev 则

旨在利用模型的高层预测能力，正由于模型在高层进行对下一个词的预测，因此当前词的语义才可能体现在前一个词的表征中；提示语策略则设计提示语来诱导模型生成与目标词语义相关的表征，从而更加符合解码结构的生成模型自回归的架构和生成模式。

对于不同类别的提示语，表 4.1 给出了示例，以 *bank* 在 *river bank* 的词义理解为例，其中加粗部分表示模型提取表征的目标词位置 ①，上部分是单纯使用原始数据，下部分则设计不同的提示语，如果没有明确说明，都是选择相应位置的最高层作为最终的表征。其中提示语的设计都受到了提取句向量的启发，提示语策略[144]将句义浓缩为单独的一个词上，并且利用生成模型的预测能力来有力捕捉句义。

需要指出的是，大语言模型的生成结果对提示词的表述方式具有较强的敏感性。不同的提示词在信息显式程度、语言学约束以及推理引导方式上的差异，可能显著影响模型的最终输出。因此，本文围绕同一核心任务设计了多种提示词模板，从不同角度对模型的语境语义理解能力进行探测与对比分析。

表 4.2 总结了本文采用的各类提示词模板。具体而言，*PromptEoL*[144] 作为最基础的提示形式，直接要求模型在给定语境中对目标词的意义进行单词级概括；*PCoTEoL*[145] 在此基础上引入显式的思维链指引，以鼓励模型进行更充分的内部推理；*KEEoL* 则通过引入语言学层面的背景知识，强调词义受到语境、句法环境及语义角色等因素的共同制约；*LAN_KEEoL*[191] 进一步考察不同语言条件下提示词翻译对模型行为的影响；最后，*POS_KEEoL* 在提示中显式加入目标词的词类信息，以分析句法约束对语义判断的作用。这些提示语模板都是为了提取更好的句向量所设计的，通过提示语的方法将目标词的语义确立问题转换为目标词词义预测的问题。需要强调的是，原文献都是针对句向量而设计的，本文对该模板内容进行改进以适应词的语义表征。

4.2.2　几何探测方法

探测方法（probing）是分析表征学习到哪些知识的一种有效方法，一般可以采取分类器探测和几何探测的方式。其中分类器探测针对目标表征额外训练一个轻量级的分类器，该分类器以要探测的知识为目标，例如针对词义相关度任务，其分类器目标是四分类任务。通过该轻量级分类器完成该任务的性能来推断模型是否学习到了词义相关的知识。分类器探测需要学习额外的参数，且探测的知识可能是分类器额外学习到了，并非模型原始学习到的知识。

为了克服分类器探测的缺点，几何探测法（geometric probing）[192]则直接从

① 对于提示语的策略，提取的是末尾冒号的位置。

表 4.2　各类提示语方法对应的模板和备注

提示语名称	模板	变量	备注
PromptEoL	In this sentence “{context}”, “{lemma}” means, in one word:	{context}, {lemma}	基础提示语
PCoTEoL	After thinking step by step, in this sentence “{context}”, “{lemma}” means, in one word:	{context}, {lemma}	思维链加强的提示语
KEEoL	A word’s meaning is not determined solely by its dictionary definition, but is shaped by its context of use, its syntactic environment, and its semantic role. In the sentence “{context}”, the word “{lemma}” means, in one word:	{context}, {lemma}	语言学知识增强的提示语
LAN_KEEoL	Language-specific translations of the PromptEoL template.	{context}, {lemma}	各种语言对于基础提示语的翻译版本
POS_KEEoL	A word’s meaning is not determined solely by its dictionary definition, but is shaped by its context of use, its syntactic environment, and its semantic role. In the sentence “{context}”, the word “{lemma}”, which has the part of speech {pos}, means, in one word:	{context}, {lemma}, {pos}	增加了词类信息的提示语

表征向量出发，通过计算向量之间的相似性来反映语义的相关性。具体来说，首先可以对提取的表征 h_i 和 h_j 计算余弦相似性 m_{ij}：

$$m_{ij} = \frac{h_i \cdot h_j}{|h_i| \times |h_j|} \tag{4.1}$$

其中 h_i 和 h_j 分别表示出现在两个上下文中目标词的表征向量，$|h_i|$ 和 $|h_i|$ 分别表示对相应向量求模长，最终所得的 m_{ij} 代表两个向量的相似度。

之后根据训练集可以找到一个映射到四分类的阈值列表 $\{\gamma_i\}$，其中 $i = \{1,2,3,4\}$，用以将连续的相似性分数转变为离散的相关度水平，即：

$$\mathrm{rel}_{ij} \triangleq \mathrm{rel}(w_i, w_j) = \begin{cases} 1, & \gamma_1 \leqslant m_{ij} < \gamma_2, \\ 2, & \gamma_2 \leqslant m_{ij} < \gamma_3, \\ 3, & \gamma_3 \leqslant m_{ij} < \gamma_4, \\ 4, & m_{ij} \geqslant \gamma_4. \end{cases} \tag{4.2}$$

其中 rel_{ij} 表示词 w_i 与 w_j 在不同上下文中的相关程度等级，m_{ij} 为 w_i 与 w_j 对应表征向量的相似度（如余弦相似度），$\gamma_1, \gamma_2, \gamma_3, \gamma_4$ 为预先设定的阈值参数，满足 $\gamma_1 < \gamma_2 < \gamma_3 < \gamma_4$，用于将连续相似度映射为离散的相关性等级（值越大表示相关程度越高）。关于寻找映射的方法，本文借鉴了 Nelder–Mead 无导数优化方法[193]，通过在训练集上迭代搜索最优阈值向量 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$，以最大化分类性能指标。该方法属于基于单纯形（simplex）的启发式搜索算法，不依赖梯度信息，而是通过比较当前若干候选解的目标函数值，在参数空间中进行反射、扩张与收缩等操作，不断调整搜索方向。算法在每一步都会根据性能表现更新候选解组合，从而逐步逼近最优解。由于无需计算导数，Nelder–Mead 方法特别适用于目标函数不可导或难以解析求导的情形。

示意图 4.2 展示了上述几何探测的过程。其中两个目标词的表征使用蓝色的条形形状来表示，它们经过余弦相似度可以计算出相似度分数 m_{ij}，再根据由 Nelder-Mead 算法在训练集所得到的阈值进行分配。分配的相关度为相应阈值之间的四个等级，这里使用四个不同高度的圆筒状来表示。

几何探测法无需额外训练参数，可以有效地避免知识来源不明的问题，同时对表征直接的运算，也可以使用最小的步骤来解释和评估词义，同时无需训练的方式也减少了很多计算的开销，因此是一个性价比很高的评估方式。但是几何探测自身也存在问题，例如它可探测的语言学知识较为单一，一般以比较两个语言单元的相似度或者相关性相关，更加复杂的语言学知识，例如语义角色[156]、句法分析[159]等更加复杂的探测就无法使用这种方法了。

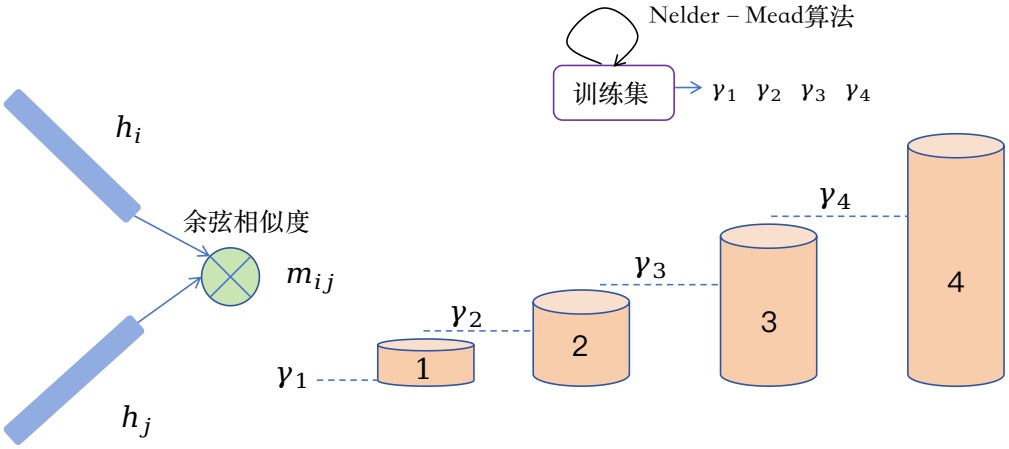
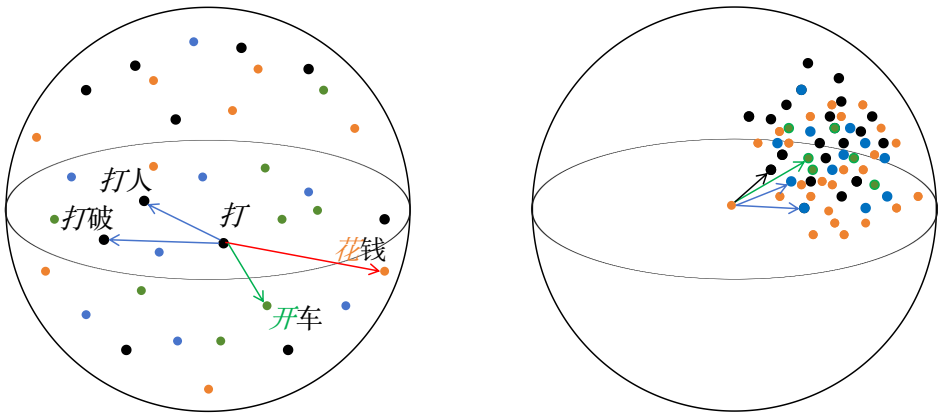


图 4.2　几何探测示意图

本章由于仅涉及对于表征的探测，且本身就是一个相关度的任务，因此采用了向量探测的方式。在之后有关词义相关度不一致程度预测的任务中会采用分类器探测的方式，请参考第 6 章。

4.2.3　向量后处理

已有研究表明[194]，神经网络语言模型生成的上下文语义表征通常具有明显的各向异性（anisotropy）特性，即向量在表示空间中聚集于相对狭窄的区域。这种几何结构会导致不同词或语境之间的相似度分数整体偏高，从而削弱相似度度量在语义相关任务中的区分能力。例如，即使在词义上无关的词语之间，也可能出现较高的余弦相似度。



(a) 各向同性　　(b) 各向异性

图 4.3　表征分布示意图

图 4.3 展示了各向同性和各向异性表征的分布。图中的每一个点代表嵌入在向量空间中的一个词，它们作为向量长度归一化后分布到单位球面上。其中目标词使用斜体来表示，其余的正体词则表示该目标词出现的上下文环境，不同颜色

代表不同的目标词。对于各向同性（isotropy）而言，向量均匀得分散在单位球面上，这样不同的点之间的相似度就会呈现较大的差异，对于词义接近的点，它们在向量空间中的位置也更加接近，相似度可能更高。例如图中的“打人”和“打破”语境中的“打”。而对于语义不相近的点，它们则分布在较远的位置，例如“打”和“花钱”的“花”。而相反，对于各向异性的分布，词表征聚集在某一个角落里，所有的点之间的相似度都非常高，尽管有些词的语义并不相近，这样单靠模型求出的阈值很难将不同语义的点区分开来。因此我们要提前将各向异性的点通过去各向异性的手段转化为各向同性。

为缓解上述问题，本文对提取的语义表示采用三种去各向异性（anisotropy removal）处理方法。设语义表示集合为 $\{\mathbf{h}_i\}_{i=1}^{N}$，其中 $\mathbf{h}_i \in \mathbb{R}^d$ 表示第 i 个样本的 d 维语义向量。

（1）中心化（Centering）

中心化旨在消除表示空间中由于整体分布偏移所引入的公共方向偏置。首先计算所有向量的均值向量，其中“公共方向”是指由数据整体分布偏移导致的、对所有向量产生一致影响的系统性偏置方向。

$$\mu = \frac{1}{N}\sum_{i=1}^{N} h_i \tag{4.3}$$

其中 $\mu \in \mathbb{R}^d$ 表示语义表示的全局平均位置，N 表示样本个数。随后对每个向量执行平移操作

$$h_i^{(c)} = h_i - \mu \tag{4.4}$$

该操作使得变换后的向量集合满足零均值性质，即

$$\frac{1}{N}\sum_{i=1}^{N} h_i^{(c)} = 0 \tag{4.5}$$

中心化能够有效去除向量空间中由频率效应或模型训练偏置所产生的公共方向，从而提高语义结构分析的稳定性。

（2）标准化（Standardization）

尽管中心化可以消除均值偏置，不同维度仍可能具有不同的尺度范围。为进一步消除各维度方差差异，本文对中心化后的向量执行逐维标准化处理。

首先计算各维度的标准差向量：

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(h_i^{(c)} \odot h_i^{(c)}\right)} \tag{4.6}$$

其中 $\odot$ 表示逐元素乘法，$\boldsymbol{\sigma} \in \mathbb{R}^d$ 表示各维度的标准差。随后定义标准化变换为：

$$h_i^{(s)} = h_i^{(c)} \oslash \boldsymbol{\sigma} \tag{4.7}$$

其中 $\oslash$ 表示逐元素除法。该操作使各维度具有单位方差，从而避免某些高方差维度在距离计算或相似度评估中占据主导地位，提升不同维度语义信息的均衡性。

（3）主成分去除法[195]

已有研究表明[195]，语义向量空间往往存在少数主导方向，这些方向对应最大方差主成分，通常包含较强的频率或统计共现信息，但对语义区分能力贡献有限。为缓解该问题，本文采用主成分去除法来移除这些主导方向。

首先对标准化后的向量集合执行主成分分析（PCA, principal component analysis）或奇异值分解（SVD, singular value decomposition），获得前 k 个最大方差主成分方向：

$$U_k = [u_1, u_2, \cdots, u_k] \tag{4.8}$$

其中 $u_j \in \mathbb{R}^d$ 表示第 j 个单位主成分方向。随后从每个向量中减去其在这些方向上的投影：

$$h_i^{(\mathrm{r})} = h_i^{(s)} - \sum_{j=1}^{k} (u_j^\top h_i^{(s)})\, u_j \tag{4.9}$$

其中，$h_i^{(s)} \in \mathbb{R}^d$ 表示第 i 个标准化后的词向量，$h_i^{(\mathrm{r})} \in \mathbb{R}^d$ 为经过消偏处理后的词向量，$U_k \in \mathbb{R}^{d\times k}$ 是由前 k 个主成分方向构成的矩阵，k 为需要消除的主成分数量，d 为向量维度，$u_j^\top h_i^{(s)}$ 表示第 i 个向量在第 j 个主成分方向上的投影系数。该操作等价于将向量投影到与主导子空间正交的补空间中，从而削弱高方差主方向对表示空间结构的支配作用，使语义信息在更多维度上均匀分布。

综上所述，中心化、标准化与主成分去除法分别从均值偏置消除、尺度归一化以及主成分去除三个层面对语义表示进行结构优化。上述处理能够有效改善向量分布的各向同性特性，并增强词向量在语义相似度计算与几何结构分析中的判别能力。本文在几何探测实验中，对最高层语义表征统一应用上述去各向异性处理流程。

4.2.4　基于提示语的探测方法

大语言模型（Large Language Models, LLMs）凭借其强大的生成能力，将传统上区分明确的理解任务与生成任务统一到了同一框架之下，即通过精心设计的提示词（prompt）引导模型生成目标输出。在这一范式下，模型对输入语境的理解能力可以通过其生成结果进行间接探测。基于这一思路，本文采用基于提示词的探

测方法，对不同大语言模型在语境化语义判断任务中的表现进行系统分析。

与前文基于内部表征提取的方法不同，本节不再直接访问模型的隐藏层表示，而是将任务表述为一个生成式问题：模型在给定上下文和目标词的情况下，需要输出一个离散的词义相关度判断结果，同时给出模型做出回答的理由。要求模型根据提示词生成一个对应四个等级之一的相关度分数，从而将语义判断任务转化为一个受控的文本生成过程。

下文展示了本文在大语言模型实验中使用的提示语模板。为保证实验结果的可复现性与客观性，提示语以英文原文形式呈现，同时给出对应的中文翻译以辅助理解。其中 `WORD`、`SENTENCE1` 和 `SENTENCE2` 分别表示每个评估样例中的目标词、第一个上下文和第二个上下文。该提示语仅向模型提供目标词及其在两个句子中的上下文信息，不引入词性标注、词义编号或人工语义标签等语言学先验知识，旨在评估模型在最小提示条件下对词义相关性的自主判断能力。同时，提示语对输出格式进行了严格约束，要求模型仅返回包含相关度和解释的结构化 JSON 结果。这一设计能够有效提升实验结果的可解析性与自动化处理能力，并减少模型生成冗余文本或推理过程带来的干扰，从而提高实验结果的一致性与可靠性。

You are given two sentences that contain the same target word (marked in bold). Your task is to judge how semantically related the meanings of this word are in the two contexts.

Target word: `WORD`

Sentence 1: `SENTENCE1`

Sentence 2: `SENTENCE2`

Rate the semantic relatedness between the two uses of the target word on a 4-point scale:

1 = Unrelated

2 = Distantly related

3 = Closely related

4 = Identical

IMPORTANT: Output exactly one valid JSON object with exactly two keys: ”rating” and ”explanation”.

Do NOT include any other text, reasoning, or commentary.

The JSON should look like:

```
{
 ”rating”: 1,
```

”explanation”: ”brief reason here”

}

以下为上述提示语的中文翻译版本：

给定两个包含相同目标词（以粗体标记）的句子，要求判断该词在两个语境中的语义相关程度。

目标词：目标词

句子 *1*：第一个上下文

句子 *2*：第二个上下文

请根据四级量表对两个语境中目标词的语义关系进行评估：

1：词义不相关

2：词义远相关

3：词义近相关

4：词义完全一致

模型需严格输出一个合法的 *JSON* 对象，仅包含“评分结果”和“解释”两个字段。

{

“评分结果”：1,

“解释”：“简要理由”

}

本文还使用另外一种提示语，它在提示语的正文开头告知大语言模型词类信息的重要作用，并在后续输入中增加了词类信息。以下为词类增强的提示语开头部分，下划线高亮了和原有提示语相比增加的部分：

You are given two sentences that contain the same target word (marked in bold). Your task is to judge how semantically related the meanings of this word are in the two contexts. explicitly taking into account the part of speech of the target word in each sentence.

在后面的数据格式中增加了词类信息，其余信息不变：

Sentence 1: `SENTENCE1`　Part of speech in Sentence 1: `POS1`

4.3　词义相关度实验

4.3.1　任务定义和数据集

词义相关度任务旨在探测模型对于不同上下文词义相关程度的理解。数据集采用多语言依序分级的上下文词义相关度 OGWiC 数据集和自行收集的汉语兼类词数据集 OGWiC_NV，其中 OGWiC 在第 3.2.2 章、OGWiC_NV 在第 3.2.4 章中都

被详细地介绍了，包括数据格式、数据集的组织手段、数据统计量等。评估指标为克里彭多夫 α 系数，详见公式 (3.6) 的描述。

4.3.2　评估模型

本研究评估的语言模型涵盖了基于编码的语言模型（Encoder-based Models）与基于解码的生成式语言模型（Decoder-based Models）。其中，解码器模型以大语言模型（Large Language Models, LLMs）为主。这些模型大多具备多语言语义理解能力。同时，为进行对照实验，本文还引入了若干单语理解模型作为基线。

4.3.2.1　基于编码的语言模型

（1）BERT 系列模型

BERT（Bidirectional Encoder Representations from Transformers）由 Devlin 等提出[35]，基于双向 Transformer 编码器结构[16]，通过掩码语言模型（MLM）与下一句预测（NSP）任务进行预训练。本文采用 XLM-BERT 和单语 BERT-base 模型：

① XLM-BERT ①，训练语料覆盖 104 种语言，学习跨语言的统一表征，用于评估多语言统一建模能力；

②各语言对应的单语 BERT-base 模型，用于衡量单语言预训练模型在本任务中的性能上限。表 4.3 展示了各自单语模型的名称、下载地址②和收集机构。

表 4.3　各语言对应的 BERT 模型及其来源

语言	下载地址	收集机构
汉语	google-bert/bert-base-chinese	Google
英语	google-bert/bert-base-uncased	Google
德语	google-bert/bert-base-german-cased	Google
挪威语	NbAiLab/nb-bert-base	挪威国家图书馆
俄语	deepvk/bert-base-uncased	DeepPavlov（个人）
西班牙语	dccuchile/bert-base-spanish-wwm-uncased	智利大学
瑞典语	KB/bert-base-swedish-cased	瑞典国家图书馆

（2）XLM-RoBERTa

XLM-RoBERTa（XLM-R）是 Conneau 等人提出的大规模多语言预训练编码器模型[196]，其结构基于 RoBERTa[141]的 Transformer 编码架构，在 100 种语言的 CommonCrawl 语料上以掩码语言模型（MLM）的训练目标进行无监督预训练。由

① https://huggingface.co/google-bert/bert-base-multilingual-cased
② 可以直接访问 https://huggingface.co/下载地址

于其训练语料规模远大于早期多语言 BERT[35]，XLM-R 在多语言语义理解、跨语言迁移与零样本任务中表现出显著优势。本文选用 XLM-R-base 模型 ①。

（3）XL-Lexeme

XL-Lexeme 是 Cassotti 等人[197]提出的一个面向跨语言词义区分任务[198]的双向编码器模型。该模型基于 Sentence-BERT[199]的孪生网络结构，分别对包含目标词的两个句子进行独立编码，并通过余弦距离衡量句对相似度。其训练目标采用对比学习策略：当目标词在两个语境中语义一致时最小化表示距离，语义不一致时最大化表示距离。由于该模型的预训练语言任务与本文讨论的词义区分任务很相似，因此它具有较强的词义识别能力。

（4）mBART

mBART 是 Liu 等人提出的多语言序列到序列预训练模型[200]，基于标准 Transformer 编码器——解码器架构[16]，在多语言文本去噪自编码任务上进行预训练。其训练过程通过随机打乱、删除与遮蔽输入句子，使模型学习跨语言上下文重建能力。mBART 在机器翻译、跨语言理解与生成任务中表现稳定，能够同时承担语义理解与自然语言生成双重角色。本文采用 mbart-large-50 模型 ②。本文使用 mBART 的编码器部分进行实验。

（5）mT5

mT5 是基于 T5 框架扩展得到的大规模多语言文本到文本（text-to-text）的编码器——解码器架构模型[201]，在覆盖 101 种语言的 mC4 语料上进行预训练。mT5 将所有任务统一建模为“输入文本 → 输出文本”形式，在跨语言语义匹配、分类、翻译等任务中具有较强泛化能力。具体模型使用的是mT5-base ③。本文使用 mT5 的编码器部分进行实验。

4.3.2.2　基于解码的生成式语言模型（大语言模型）

（1）LLaMA 系列

LLaMA 是 Meta 提出的开放高效基础大语言模型系列[202]，采用纯解码器 Transformer 架构，在超大规模多语言语料上训练自回归语言建模目标。本文选用以下模型：

① `LLaMA-2-7B` ④：开源大语言模型 LLaMA 的第二版本，其在安全性、指令跟随能力与多语言数据规模较前一版本都有较大提升，简称为 LLaMA2。

① https://huggingface.co/FacebookAI/xlm-roberta-base
② https://huggingface.co/facebook/mbart-large-50
③ https://huggingface.co/google/mt5-base
④ https://huggingface.co/meta-llama/Llama-2-7b-hf

② LLaMA-3-8B [①]：最新一代模型，在多语言推理、对话与理解能力上进一步提升，简称为 LLaMA3。

这些模型代表了当前主流开源通用大语言模型在词义判别任务上的零样本能力。另一方面它们也代表一系列统一架构下的不同迭代版本的大语言模型，与之对应的语言能力也依次增强。同时本文还采用表征探测的方式来提取 LLaMA 模型的内部表征，提取方式具体可参考第 4.2.4 节。

（2）DeepSeek

DeepSeek 是国内深度求索团队提出的面向复杂推理与语言理解优化的大语言模型[203]，在数学推理、逻辑分析和自然语言理解任务上具有较强性能。该模型结合了大规模监督微调与强化学习优化策略，在弱监督与推理泛化方面表现突出。本文将其作为具备较强推理能力的代表性模型，并采用官网提供的旗舰模型 DeepSeek-Chat。该模型由于是闭源模型，无法获取内部表征，仅使用前面第 4.2.4 节提到的提示语来获取最后的结果。

（3）通义千问

通义千问（Qwen）是由一款阿里巴巴集团旗下的云端运算服务的科技公司阿里云开发的一款大语言模型 [②]。随着第三代模型（Qwen3）的发布，它在代码、数学、通用能力、多语言理解等基准测试中都可以与其他顶级大语言模型，如 DeepSeek[203]、GPT4-o1[204]等相媲美。本文采用第三代后训练版本的通义千问模型 Qwen3-Next-80B-A3B-Instruct [③]，作为新一代 Next 版本，它在以下方面得到了提升：混合注意力、高稀疏度的混合专家模型、稳定优化和多词预测，也因此在效率和性能上达到较优的水平。与上述 DeepSeek 设定类似，本章也采用第 4.2.4 节提到的提示语来获取直接获取最终的结果。

4.3.2.3　上界与下界

（1）上界

对每一种语言，遍历所有标注者，计算“当前标注者的标注结果”与“其余标注者标注结果的中位数”之间的克里彭多夫 α 系数；最终按标注数量对所有标注者的 α 进行加权平均，作为该语言下人类可达到的理论性能上界。

（2）最大频率下界

该方法对所有测试样本固定预测标签为最常出现的标签 4（参考图 3.2和图 3.1），这利用了对于不平衡分布的先验知识。虽然是一个较为简单的方式，但

① https://huggingface.co/meta-llama/Meta-Llama-3-8B
② https://qwen.ai/
③ https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct

在与词义相关的任务中[43]，这一方式往往可以提供一个较强的基准。

（3）随机下界

随机下界方法在标签空间内进行均匀随机采样生成预测结果，用以表示在没有任何任务提示下的性能下限。

4.3.3　超参数设置

（1）表征提取的过程。对于所有基于编码的语言模型，本文均采用其官方发布的预训练参数权重，在下游任务中仅用于提取目标词的上下文语义表征，不进行额外微调。对于基于解码的生成式模型，在评测阶段统一采用零样本（zero-shot）推理设置；对于开源解码器模型，同时额外采用表征提取的方式作为对照。所有实验均基于表 3.5的训练集与测试集进行，以确保不同模型之间实验结果的公平性与可重复性。

（2）探测（probing）阶段。本文采用几何探测方法，默认选取模型最高层的语义表征作为输入。根据实验结果，对于 OGWiC_NV 数据集，本章采用大语言模型倒数第二层表征作为最终的输入。之后对提取的表征进行去各向异性（anisotropy removal）处理。决定分类的阈值参数通过在训练集上采用 Nelder–Mead 方法进行自动寻优获得。对于中文兼类词 OGWiC_NV 数据集，采用在汉语上训练的 BERT 模型 ①来替代 XLM-B，其余两个数据集都采用一套模型。

大模型的命令行直接预测还涉及温度参数 τ，它是一个调节模型用来预测下一个词概率分布的控制因子。当 τ 值较大时，原有的预测概率会被“压平”，从而使得采样结果更加多样；反之，当 τ 值较小时，概率分布被“锐化”，使得模型的输出更加确定。本文默认的温度参数选择为 1。

4.4　实验结果和分析

本节首先对比各类不同模型在不同语言以及全部数据的测试集上的准确性。之后针对本章提出的两个创新方法——去除各向异性和零样本指令设计展开分析讨论。最后分析对比了多语言模型和单语模型的差异。其中准确性分析包括 OGWiC 和 OGWiC_NV 两个数据集，除了最后一小节分析词类的影响，其余小节均使用 OGWiC 的全集进行消融实验和分析。

表 4.4　各类模型在不同语言上的准确性

模型	OGWiC								OGWiC_NV
	全部	汉语	英语	德语	挪威语	俄语	西班牙语	瑞典语	
上界	95.32	1.00	95.95	90.64	95.57	93.17	95.43	96.48	76.00
最大频率下界	-22.40	-5.23	-46.23	-38.03	-7.97	-18.51	-28.42	-12.41	-43.93
随机下界	-17.38	-42.25	5.39	-3.45	-34.86	-11.33	-8.53	-26.62	-19.70
XLM-R	27.59	16.49	47.51	43.75	13.31	26.81	24.68	20.60	26.26
XLM-R(11)	33.82	15.39	52.51	48.68	23.87	30.48	33.45	32.35	24.89
XLM-B	17.22	18.51	23.71	32.15	26.80	13.95	10.78	-5.36	26.20
XL-Lexeme	**59.02**	**38.30**	**66.74**	71.13	50.65	56.67	**70.19**	59.42	47.36
mBART	11.6	3.95	23.13	17.26	11.00	12.92	1.49	11.42	-2.80
mT5	20.79	7.09	34.77	23.04	24.60	17.61	31.20	7.19	2.89
LLaMA2	22.95	25.83	25.08	34.22	16.83	26.44	30.66	1.61	15.71
LLaMA3	21.69	20.89	30.53	25.76	10.01	28.24	30.81	5.58	14.14
LLaMA2_p	54.50	26.68	50.80	**72.62**	**56.82**	56.47	58.27	**59.87**	40.26
LLaMA3_p	49.01	22.47	49.61	65.68	52.21	47.41	60.16	45.50	40.24
DeepSeek(0.1)	37.15	-3.95	49.81	70.73	11.87	47.21	48.27	36.08	58.98
DeepSeek(1)	36.71	-4.40	51.55	69.57	11.34	46.24	48.25	34.45	57.37
通义千问 (1)	52.42	32.79	64.54	64.49	33.62	**57.48**	59.15	54.84	**61.79**

4.4.1　准确性分析

表 4.4 列出了各类方法模型在 OGWiC 数据集以及 OGWiC_NV 数据集上的最佳性能表现，其中 OGWiC 也列出了不同的语言。数字加粗和下划线分别表示上界之外最优和次优的结果。如无特殊说明，采用的模型都使用最高层得到的表征，采用标准化（standardization）的方式进行各向异性的去除。对于 mBART 和 mT5 来说，采用中心化（centering）的方式进行各向异性的去除。

表 4.4 自上而下分类别罗列了不同的方法和模型。其中第一块是基础方法，包括使用剩余标注者测试的上界、最大频率出现的便签作为的下界以及随机选择下界。第二块列出了编码器模型，包括（1）基础的跨语言预训练模型 XLM-RoBERTa（XLM-R）和 XLM-BERT（XLM-B），其中 XLM-RoBERTa 还采用了前一层即第 11 层的表征进行测试，这主要是在实验中发现最优的结果往往不是最后一层，而是倒数第二层。在表格中标记为 XLM-R(11)[①]。（2）使用了相似数据集和任务进行训练的模型 XL-Lexeme；（3）编码器——解码器架构中的编码器模型 mBART 和

① https://huggingface.co/google-bert/bert-base-chinese

① 对于 OGWiC_NV 来说，由于在测试中 bert-base-chinese 的效果不如 XLM-RoBERTa，因此我们仍采用 XLM-RoBERTa 模型

mT5。

第三块代表开源大语言模型，它们通过提取表征的方式来评测。其中对于 OG-WiC 数据集来说，直接采用 LLaMA2/LLaMA3 当前目标词对应的表征，对于 OG-WiC_NV 任务来说，由于平均句长较短，仅采用目标词左边的上文无法正确理解词义，因此在大模型提取表征时，采用 Echo 的方式，即将整个句子重复一遍，再提取目标词的表征。LLaMA_p 则代表使用了提示语后再选取表征，这里的提示语统一采用 PromptEoL（参见表 4.2），之后表 4.2 中提到的提示语变种在后续消融实验中进行测试。

第四块则是闭源大语言模型 DeepSeek 和通义千问模型，采用前述提示语直接进行问题询问。括号后面显示了温度参数 τ。

表中的指标是克里彭多夫 α 系数，其中使用粗体和下划线分别特别标注出了除了基线模型外的第一名和第二名方法的结果。

（1）OGWiC 数据集

从上下界结果来看，由人工标注所给出的上界整体达到较高水准，表明该任务本身在语义判别层面具有较强的一致性与可解释性。其中，汉语上的上界结果达到 1.00，这主要是由于该语种仅包含两名标注者，并且根据数据清洗与剔除原则，仅保留了两名标注完全一致的样本，从而在统计意义上达到了“理论最优”的评估效果。然而，这一结果也从侧面反映出，标注人数的减少与筛选规则的严格性会在一定程度上抬高自动评估指标，因此在跨语言比较时仍需谨慎比较。

在下界方面，无论是采用最大频率策略（即将所有样本预测为 4），还是在 1，2，3，4 中采用均匀随机采样的方式，其整体结果均显著低于零，且在不同语言之间呈现出较大波动。这一现象清楚地表明，本文采用的相关性评估指标并非简单的“命中率”或“准确率”，而是能够有效区分模型输出与随机噪声之间的差异，从而避免在分布极不均衡或标签极化情况下产生虚高结果。因此，上述结果也从侧面验证了所选评测指标在建模细粒度语义相关性方面的合理性与鲁棒性。

从整体性能对比来看，针对特定任务进行微调的模型（XL-Lexeme）在所有语言和总体指标上均显著优于其他模型，说明在该任务中，面向词汇语义关系的专门建模和监督信号对性能提升起到了决定性作用。相比之下，尽管带有指令引导的大语言模型（LLaMA_p）在多个语言上也取得了较为可观的提升，但其整体表现仍略低于面向任务定制的专用模型，这表明通用型指令对齐并不能完全替代任务级别的精细化建模。

进一步从模型结构角度来看，基于 BERT 的双向编码器模型整体优于解码器模型。这主要源于编码器模型在获取双向上下文语义表示、建模句对或词对关系

方面具有更直接的结构优势，更适合于当前以语义相关性判断为核心的任务设定。

而在编码器模型中，RoBERTa 系列（如 XLM-R）整体优于 BERT 系列（如 XLM-B），这一结果也符合预期。这一差异主要来源于 RoBERTa 在预训练阶段采用了更大规模的数据、更长时间的训练以及去除“预测下一个句子”的任务等策略，使其在跨语言语义建模与表示泛化能力方面具有更强的优势。而由于 mT5 和 mBART 在训练时都是编码器——解码器的方式进行学习的，其性能不如仅编码器模型。

从跨语言表现来看，不同模型在德语、西班牙语等资源相对丰富的语言上普遍取得较高分数，而在挪威语、瑞典语等低资源语言上的波动则更为明显。这一现象说明，尽管多语言预训练在一定程度上缓解了资源不平衡问题，但语料规模与语言结构差异仍然是影响模型泛化能力的重要因素。

对闭源模型 DeepSeek 来说，首先温度参数对其影响并不大，本文默认使用温度为 1。其次，尽管没有在训练集上单独训练，其表现出了不俗的结果，尤其在英语、俄语、西班牙语上，其中英语尤为显著，体现了其卓越的英语语义理解能力。然而值得注意的是，DeepSeek 对于汉语却呈现了非常低的结果。另一个闭源模型通义千问则在包括汉语在内的所有语言中都体现了与最优结果相媲美的结果，在汉语、英语、俄语、西班牙语以及平均性能上都仅次于 XL-Lexeme，或者与第二名类似。

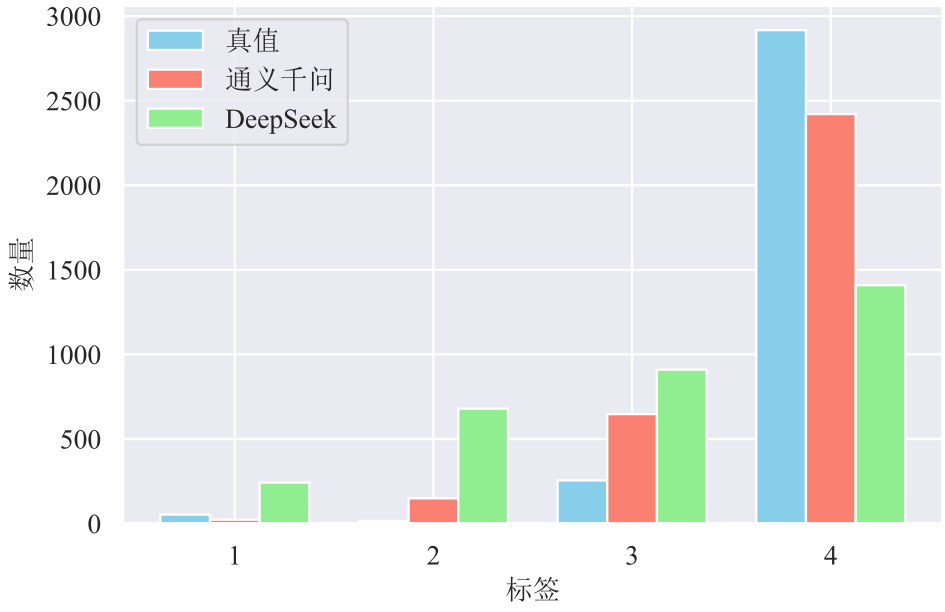


图 4.4　标签分布对比图

DeepSeek 与通义千问模型在汉语上的差异表现值得令人关注，即通义千问模型在相同设定下，在汉语上的表现要远优于 DeepSeek。图 4.4 展示了人类标注的

真值标签、通义千问模型和 DeepSeek 模型预测的标签的分布对比，不难看出通义千问的分布与真值的更加接近；尽管 DeepSeek 模型的分布也呈现偏向相关度大的趋势，然而它在非一致标签（1，2，3）上面的分布更加集中，数量也更多，拉低了模型的性能。这也反映出 DeepSeek 模型更加倾向于得到一个“词义不太相关”的标签结果。

混淆矩阵表 4.5 进一步反映了两种模型预测分布的差异趋势，其中混淆矩阵的行表示真实标签，列表示预测标签，粗体强调了最多的预测错误的情形。从整体上看，通义千问模型的预测结果明显集中在高相关标签区域，尤其是在真实标签为 4 时，模型能够较稳定地预测为正确类别（2276 次），表现出较强的高等级识别能力。同时，其预测结果呈现出较强的对角线集中趋势，说明模型整体分类一致性较好。

表 4.5　通义千问模型与 DeepSeek 模型的混淆矩阵

GT	通义千问模型				DeepSeek 模型			
	1	2	3	4	1	2	3	4
1	2	3	27	21	26	7	20	0
2	0	2	4	9	0	0	1	14
3	6	25	109	115	33	92	40	90
4	15	119	**507**	2276	183	580	**849**	1305

相比之下，DeepSeek 模型的预测分布更为分散。尽管在真实标签为 4 时仍有较多正确预测（1305 次），但其误分类主要向标签 2 和标签 3 扩散，分别达到 580 次和 849 次，显示出该模型在高等级类别上存在明显的低估倾向。此外，DeepSeek 在真实标签为 3 时，正确预测数量仅为 40 次，而大量样本被预测为更高类别，反映出模型对中间类别的区分能力相对不足。

进一步观察低标签区域可以发现，DeepSeek 在真实标签为 1 时具有更好的识别能力（26 次正确预测），明显优于千问模型（仅 2 次）。这表明 DeepSeek 在低等级样本的敏感性方面具有一定优势，但整体预测稳定性较弱。

综合来看，千问模型整体表现出更强的类别一致性和更明显的高标签偏向，而 DeepSeek 模型预测分布更为离散，在低标签识别方面具有一定优势，但在中高标签区域存在较明显的预测偏移。在后文定性的误差分析中会展示相关示例即出错的样例类型。

综上所述，该实验结果不仅验证了专用语义建模模型在本任务中的显著优势，同时也揭示了编码器——解码器结构差异、模型规模以及多语言预训练策略对跨语言语义相关性建模性能的系统性影响。

（2）OGWiC_NV 数据集

对于规模更小，更有针对性的 OGWiC_NV 数据集而言，首先是一个更难的任务，这体现在更低的上界值。这表明标注者之间在相关度判断上并没有达成非常一致。同样地，仅凭最大频率以及随机下界的方式也很难得到一个较好的结果。从不同模型来看，一个整体趋势与 OGWiC 数据集中的一致，即使用相似任务的数据训练过的模型（XL-Lexeme）的表现要优于使用了提示语引导特征的大语言模型（LLaMA2_p 和 LLaMA3_p），进而优于普通理解模型（基于编码器模型 BERT），最后优于编码器——解码器中的编码器模型部分（例如 mBART，mT5）。

闭源大模型上的表现很亮眼，通义千问模型和 DeepSeek 模型都展现了很好的效果，其中通义千问模型到达了最佳的评分 61.79 分。值得注意的是，DeepSeek 在 OGWiC_NV 数据集上的表现与 OGWiC 中的汉语部分 OGWiC_ZH 上的表现很不同。可能原因是① OGWiC_NV 中的数据从形式上更加简单，例如句长等（参考表 3.9），减弱了上下文距离对于目标词的影响，进而使得模型容易得到更加理想的结果；② OGWiC_NV 中的一对目标词大多仅出现在一条实例（句对）中，而 OGWiC_ZH 的目标词则出现在多条实例中，这体现在平均每个目标词出现的例句数，OGWiC_ZH 要远大于 OGWiC_NV。这样使得模型对于 OGWiC_ZH 里的某一个目标词的词义理解有偏差，就会影响更多的实例，最终使得结果差距更大；③ DeepSeek 在 OGWiC_ZH 中常出现的错误类型在 OGWiC_NV 中较少出现，例如 OGWiC_ZH 中由于收集的是与历时关系更密切的数据，而其中的隐喻引申关联是大模型判断较为困难的，而在 OGWiC_NV 中这样的类型比较少见，后文的误差分析中会详细展开这一点。

4.4.2　向量后处理分析

图 4.5 展示了在不同模型上，采用三种向量后处理手段（标准化 STD、中心化 Centering、去除主成分 PCArem）对未做任何处理（NO）的模型性能的影响情况。总体而言，在多数模型上引入向量后处理手段后，性能均得到不同程度的提升，表明各向异性确实是影响语义表示质量的重要因素。其中，这一趋势在编码器模型（如 XLM-R、XLM-B 以及 XL-Lexeme）和大语言模型（如 LLaMA 及 LLaMA_p）上表现得尤为明显，说明这两类模型在原始表示空间中普遍存在较强的各向异性结构，而去除该偏置有助于恢复更加均匀、判别性更强的语义分布。

相比之下，在编码器——解码器联合训练的编码器模型（如 mBART 和 mT5）上，各类去各向异性方法带来的提升则相对有限，部分设置下甚至出现轻微性能下降。这一现象表明，此类模型的表示空间本身可能并未表现出过强的各向异性特征，或者其解码式预训练目标（如自回归建模或序列到序列建模）在一定程度

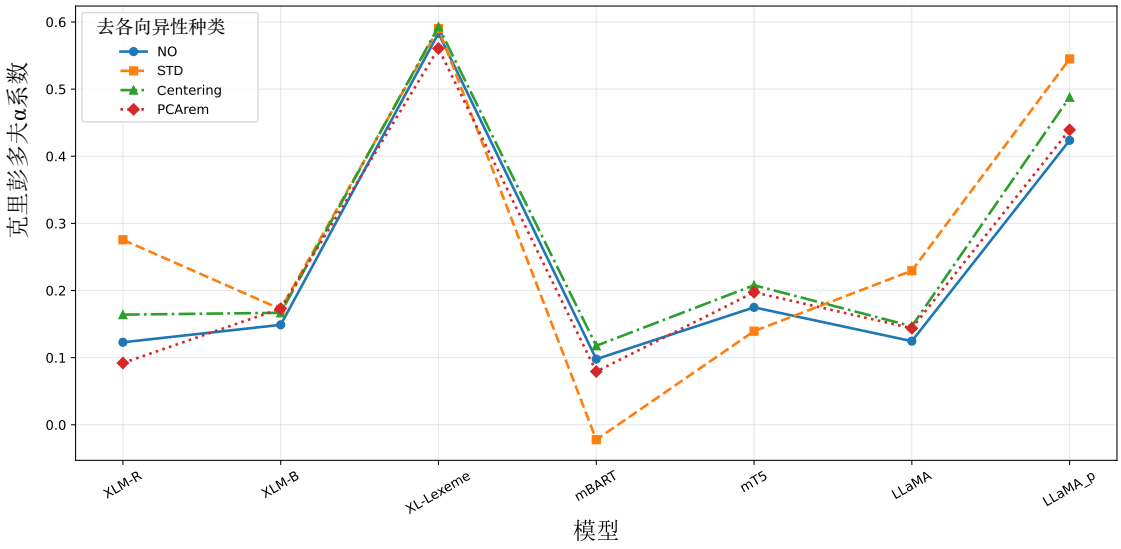


图 4.5　各类去各向异性手段的平均效果

上已经隐式缓解了表示坍缩问题，因此额外的后处理操作带来的收益较为受限。

从不同去各向异性技术的具体效果来看，STD 方法在大多数模型上均能够带来最为稳定、幅度也最为显著的性能提升，尤其是在编码器模型和大语言模型中表现突出，显示出其在缓解表示空间各向异性方面具有较强的普适性与鲁棒性。相较之下，Centering 与 PCArem 在个别模型上虽然也能带来一定增益，但整体稳定性与平均提升幅度均不及 STD 方法；而在 mBART 和 mT5 上，STD 的效果则相对不明显，进一步印证了不同模型结构对去各向异性手段敏感性的差异。

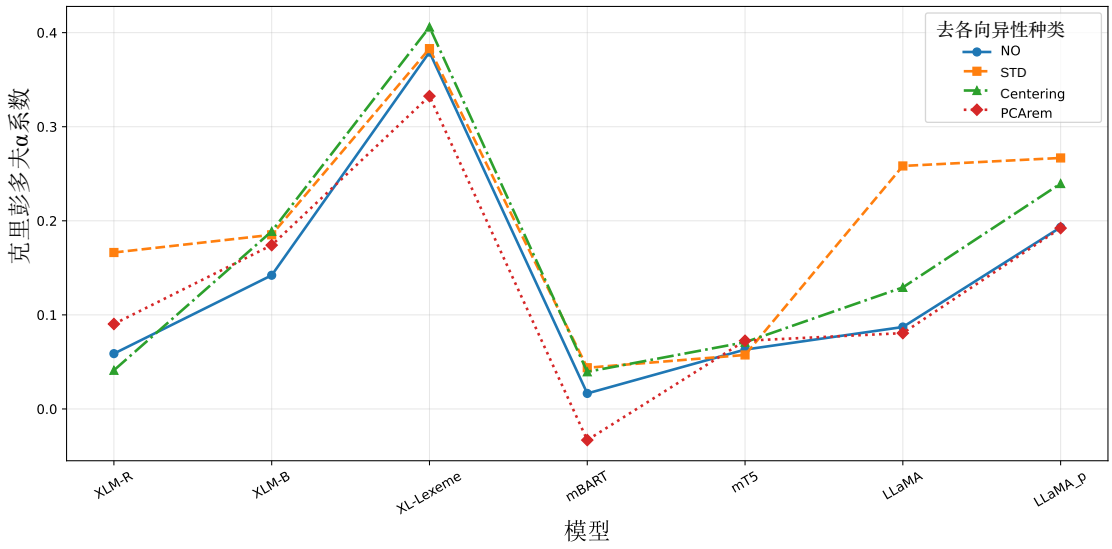


图 4.6　中文数据集上各类去各向异性手段的效果

从不同的语言来看，在所有语言的各类模型上，使用了去各向异性手段要比不使用效果都要好，也就是在各类图中代表不使用去各向异性手段的蓝色线都处

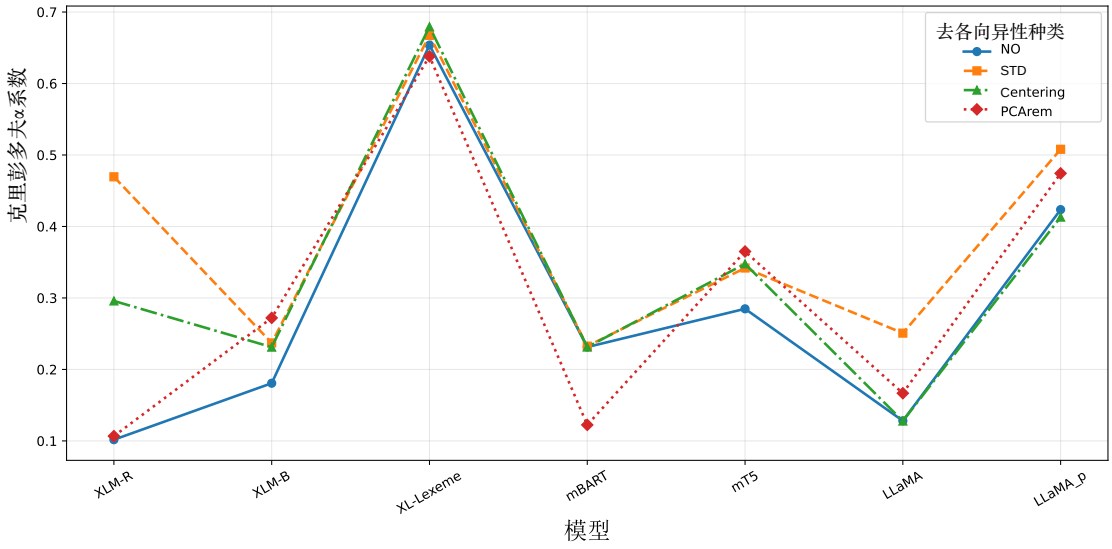


图 4.7　英文数据集上各类去各向异性手段的效果

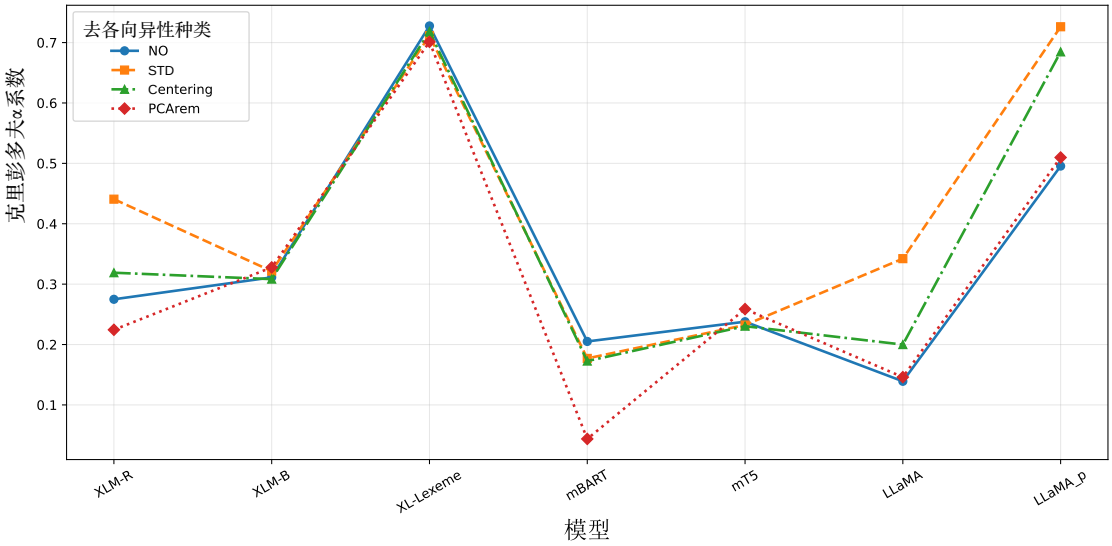


图 4.8　德语数据集上各类去各向异性手段的效果

在其余某条线的下面。在各个语言上不同的手段效果不同。对于图 4.6 显示的中文数据集来说，除了 XL-Lexeme 和 mT5 外，标准化 std 在所有的去各向异性手段中都表现得最优，与平均指标不同的是，std 在 mBART 上仍旧表现更优。图 4.7 中呈现的英文数据集则表现出类似的结果，尽管在 XL-Lexeme 上几个手段的效果都近乎类似，这表明 XL-Lexeme 中的向量自身的各向异性的特征不明显。然而对于 XLM-R 和带有提示语提问的大模型来说，std 提升的效果非常显著，说明该方法可以有效地遏制向量的过度集中。图 4.8 中德语的表现和英语类似，std 在相同的模型上达到更优的效果，表明两种语言的向量分布较为相近。图 4.9 中挪威语的表现与英语德语的差异较大，除了在大语言模型上 std 的性能显著高于其他手段外，其他

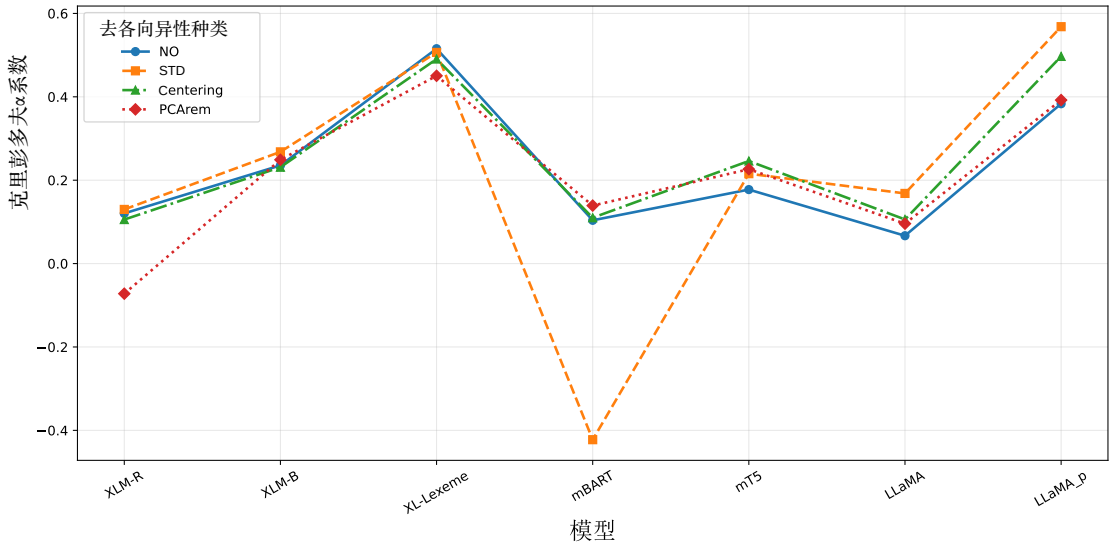


图 4.9　挪威语数据集上各类去各向异性手段的效果

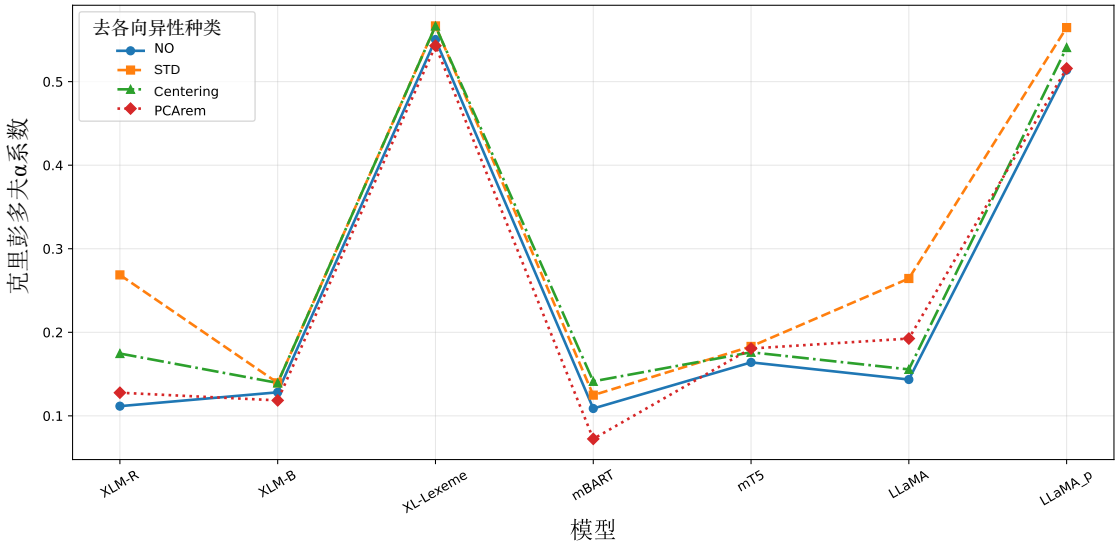


图 4.10　俄语数据集上各类去各向异性手段的效果

模型上的性能 std 的性能都没有很明显得优于其他。对于 mBART 来说，使用了 std 甚至严重损害了数据集结果，这也是导致平均性能下降的一个重要原因。mBART 模型在挪威语整体表现都并不好，可能向量自身就无法自行区分语义，在标准化之后，较差的向量会更大程度影响其他向量，可能更加损害最终的准确性。

图 4.10 显示的俄语中，仍然在大模型和 XLM-Roberta 上 std 可以起到显著效果，其他均不明显。而在图 4.11 展示的西班牙语结果中，std 仅在使用了提示语命令的大语言模型和 XLM-Roberta 中有微弱提升，在其他模型都没有占据更优的结果。图 4.12 显示的瑞典语中，除了与之前类似的趋势外，在 mBART 和 mT5 两类相似架构的模型中，std 都较显著地损害了模型的性能，其他方式则与不使用去各

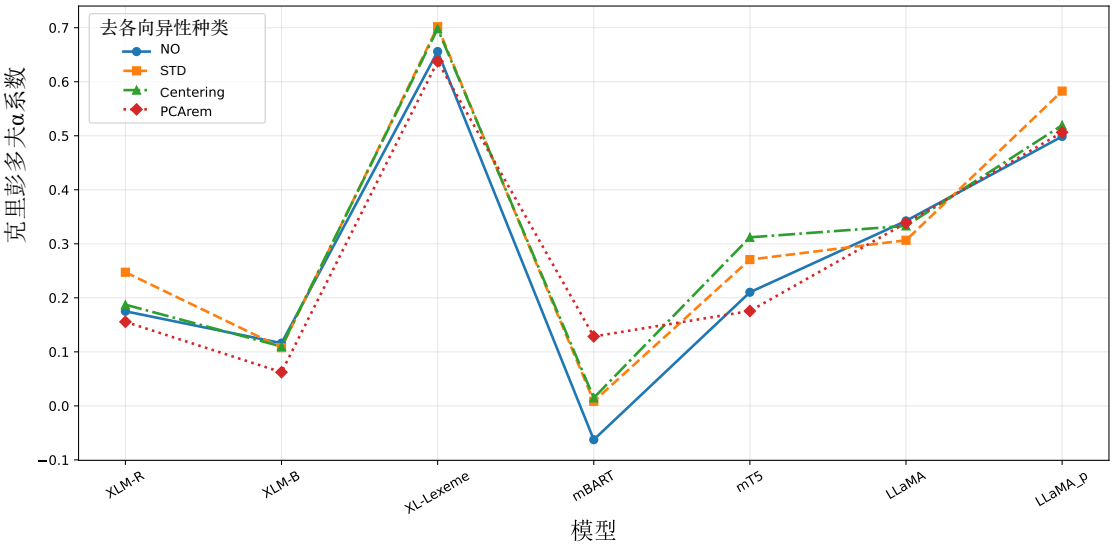


图 4.11　西班牙语数据集上各类去各向异性手段的效果

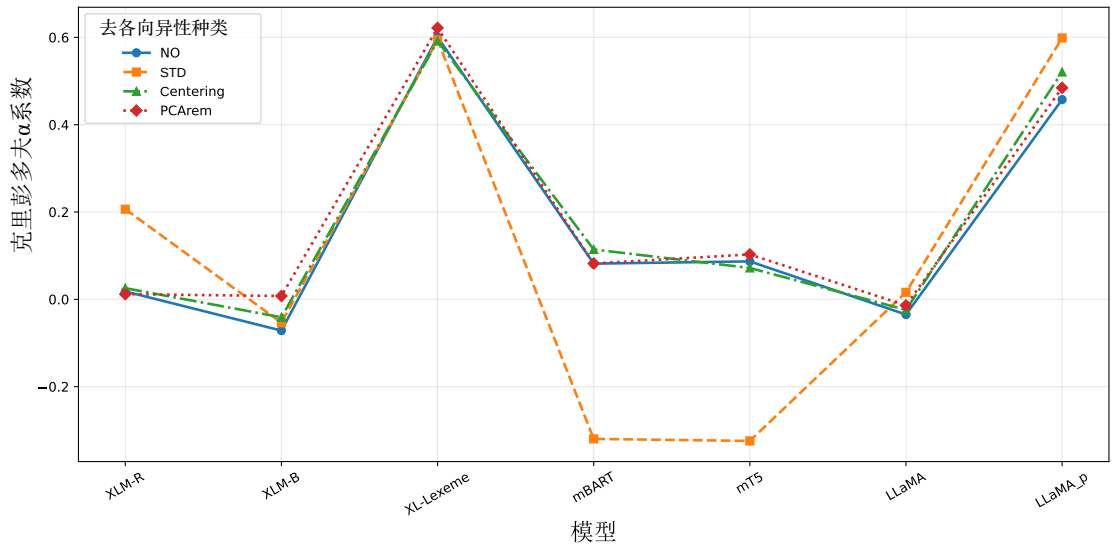


图 4.12　瑞典语数据集上各类去各向异性手段的效果

向异性手段的结果类似，都是较低的结果。

综合各个语言的分析，可以看出在 XLM-Roberta 和大语言模型上，其分布更加具有各向异性，并且 std 对它们的性能提升都是有效的，在其他表现性能较差的模型上，很多语言都是其他手段占优，或者与不使用这些手段的结果差别不大，而在效果最后的相似任务微调过的模型 XL-Lexeme 而言，各类方法都对结果差别不大，说明其向量的分布已经接近了各向同性。

基于上述观察，本文在后续所有实验中默认采用 STD 作为向量去各向异性的后处理方法，以在保证性能稳定提升的同时，避免因方法选择不当而引入额外的不确定性。

4.4.3　零样本指令分析

表 4.6 展示了使用不同的提示语提取表征对于大模型性能的影响。这里使用的大模型是开源的 LLaMA 模型，温度参数为默认的数值 1。数据集使用多语言 OGWiC 和中文兼类词数据集 OGWiC_NV，模型部分首先列出了基线模型，即仅使用大语言模型对原始上下文的目标词进行提取的 Baseline，以及重复了两遍上下文再提取目标词的 Echo 方法。在使用提示语提取的部分则罗列了各类提示语模板方法，它们用于诱导模型产生和词义紧密关联的表征，其中 PromptEoL 是基础提示语，其他方式则是其变种，具体可以参考表 4.2 所罗列的信息。与前文表格类似，性能表现的第一名和第二名分别使用加粗和下划线的方式来标识。

表 4.6　各类提示语在不同语言上的准确性

模型	OGWiC								OGWiC_NV
	全部	汉语	英语	德语	挪威语	俄语	西班牙语	瑞典语	
Baseline	22.95	25.83	25.08	34.22	16.83	26.44	30.66	1.61	-5.22
Echo	15.87	26.60	16.28	21.69	12.23	20.09	11.08	3.14	15.71
PromptEoL	<u>54.50</u>	<u>26.68</u>	50.80	<u>72.62</u>	<u>56.82</u>	**56.47**	<u>58.27</u>	<u>59.87</u>	**40.26**
PCoTEOL	**55.63**	26.48	<u>58.84</u>	**73.35**	**60.39**	<u>54.57</u>	55.72	**60.05**	<u>37.68</u>
KEEOL	52.54	**28.57**	52.99	68.47	47.62	51.05	**60.14**	58.96	37.15
LAN_KEEOL	43.47	7.19	57.26	71.33	18.40	43.53	51.73	54.87	29.19
POS_KEEOL	–	–	**66.46**	–	–	–	–	–	35.20

对于 OGWiC 数据集来说，重复两遍的策略仅在汉语、瑞典语中效果有提升，在其他语言中都没有提升，这说明多数示例仅通过前文就可以获得一个较好的词向量表示，前文提供了充分的信息用于词义理解。而对于 OGWiC_NV 数据集而言，仅使用当前词的词向量则完全不能胜任该任务，重复操作可以较大提升性能。这与 OGWiC_NV 的上下文短小有关，往往目标词无法仅通过上文来得到理解。当然这种方式也不够自然，强行增加一个重复文本，会让模型潜在地认为是在做复制任务而非理解当前词义。

对于提示语构造的其他方法，模型在两个数据集上的性能都有了大幅提升，说明使用一个引导语去“显式”地告知大语言模型这是一个词义理解的任务，对于模型提取代表词义的表征具有很强的引导作用。这也说明工程上可以使用这样的方式来获取表征。值得注意的是，比起其他语言，汉语的提升幅度非常微弱，仅从基线模型的 25.83 提升了不到一个百分点。表明大模型在处理汉语数据时候，仍然存在较大的提升空间，之后的定性分析部分也会着重分析模型的错误类型。

对于 OGWiC 数据集而言，在这些提示语的变种中，使用思维链引导的 PCo-

TEOL 方法对于英语、挪威语有较大的提升，而对于其他语言有所下降。但是对于整体性能有微弱的提升。而使用了知识增强的方式 KEEOL 则对于汉语、英语、西班牙语有所帮助，对于其他的语言性能有所下降。这表明如何将语言学知识与大模型有机融合仍是一个亟待解决的问题。

而额外增加了词类的信息的 POS_KEEOL 相比未使用词类的 KEEOL，则又使得英语增加了近 14 个百分点，这表明词类在英文的词义理解任务中具有重要的作用。由于其他语言没有提供这一信息，因而无从做更加深入的比较。然而本文之后会分析词类对于汉语的影响。相比而言，将提示语信息翻译为相应的其他语言后再进行提取表征的 LAN_KEEOL 则比较大程度损害了模型的平均性能，比 KEEOL 下降了 10 个多百分点，说明开源大语言模型 LLaMA 对于英语的理解能力更强。后文会进一步分析其在单语上的变化趋势。

对于中文兼类词数据集 OGWiC_NV 来说，普通提示语 PromptEoL 可以达到一个最好的水平，其他提示语都没有有效提升性能。与中文部分的 OGWiC_ZH 相比，知识增强（KEEOL）没有提升其能力，同时使用汉语作为提示语对结果同样都起到较大的负面效应，这表明 LLaMA 模型对于英语指令更加遵循和理解。增加了词类信息的提示语则起到了副作用，这一实验补充了词类信息对于汉语数据集的影响，而原始的 OGWiC_ZH 中没有显式提供词类信息。

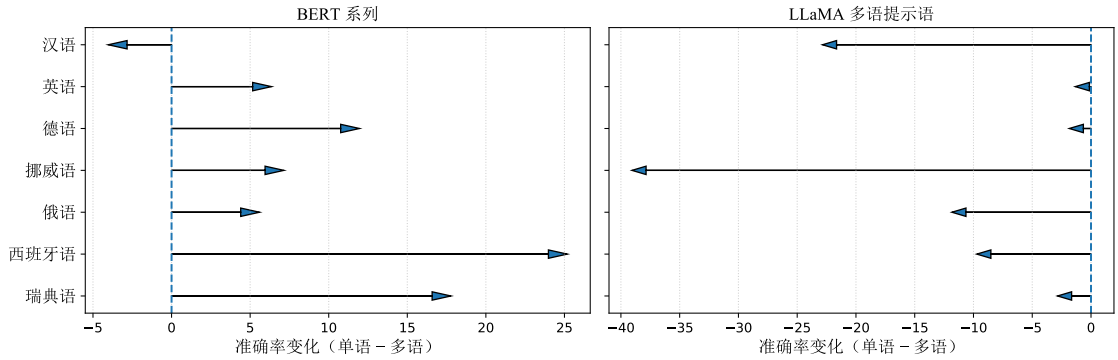


图 4.13　单语模型相对于多语模型的结果变化

4.4.4　多语言模型和单语模型对比分析

进一步结合图 4.13，可以更加细致地观察不同语言与模型之间的差异模式。图 4.13 显示了对于 BERT 和 LLaMA 而言，单语模型或单语提示语相比多语模型或多语提示语在结果上的提升幅度。其中左图的横坐标表示使用单语语言模型（BERT-base 系列）后的准确率减去使用多语模型（XLM-BERT）的准确率，右图的横坐标表示使用单语提示语的结果减去使用多语提示语的结果。图中的箭头朝向反映了变化方向：向左表示性能下降，向右表示性能提升。纵坐标表示各类语言。

首先，对于编码器模型 BERT，多数语言在单语模型下仍表现出明显优势，例如瑞典语、西班牙语和德语均呈现较大幅度的性能提升。然而，与以往观察不同，汉语在该实验中并未表现出单语优势，反而出现轻微性能下降。这一结果表明，多语言训练并非在所有语言上都会削弱模型表现。对于汉语而言，多语言模型可能通过跨语言共享语义信息，弥补了单语训练数据规模或语义覆盖范围的不足。考虑到汉语在词汇形态和语法结构上与多数印欧语言存在较大差异，其在多语言模型中可能受益于语义层面的抽象表示，而较少受到形态层面参数竞争的影响。

另一方面，挪威语、德语以及西班牙语等语言在单语模型中的性能优势仍然十分显著。这说明在这些语言中，多语言模型可能存在较强的参数干扰现象，即模型在统一表示空间中需要同时编码多种语言特征，从而削弱了对单一语言细粒度语义差异的建模能力。此外，瑞典语表现出最大的单语优势幅度，这可能与其训练语料规模或语言资源分布有关，也可能反映出北欧语言之间在多语模型中存在一定程度的语义混淆。

相比之下，大语言模型呈现出更加明显的提示语语言依赖特征。从实验结果可以看出，除英语外[①]，所有语言在使用本地语言提示语时均出现性能下降，其中挪威语与汉语下降幅度尤为明显。该现象说明当前大语言模型在语义推理过程中仍然高度依赖英语提示语所激活的知识结构。这种偏向很可能源于预训练语料中英语占比显著更高，使模型在英语语境下形成更加稳定和系统化的推理路径。此外，德语和瑞典语等语言虽然同样出现性能下降，但下降幅度相对较小，可能与这些语言在训练语料中具有较高资源覆盖度有关。

从任务特性角度来看，OGWiC 数据集强调词汇语义连续性与细粒度语义区分能力，包括事件结构识别、隐喻延伸关系以及词类转换语义映射等复杂现象。这类语义判断任务往往依赖模型对词汇核心语义的稳定建模能力。编码器模型在单语训练中更容易形成语言特定的语义空间，从而提升细粒度区分能力；而大语言模型则更加依赖提示语作为语义推理入口。当提示语语言与模型主要的训练语言不一致时，模型可能无法充分调用其潜在知识结构，进而影响判断结果。

总体而言，实验结果表明，多语言策略对不同模型架构产生了差异化影响。对于编码器模型，多语言训练可能在部分语言中引入语义表示竞争，但在个别语言（如汉语）中亦可能带来跨语言语义补偿效应；而对于大语言模型，提示语语言选择则成为影响语义推理稳定性的关键因素。这一发现进一步说明，在跨语言语义理解任务中，模型性能不仅依赖于模型规模与训练数据量，还深受语言分布结构与提示策略设计的共同影响。

① 按理说，英语部分的准确性应该没有变化，两种设定使用了同样的提示语。轻微的变化来自标准化时候采用数据量的区别。

4.4.5　词类信息对于汉语和英语的影响对比分析

为分析词类（Parts of Speech, POS）信息在词义相关度判别任务中的作用，本文比较了不显式引入词类信息的提示设置（LAN_KEEOL）与显式引入词类信息的设置（POS_KEEOL）在多语言数据上的表现差异。

对于英语数据，结果显示词类信息在提示语中能够显著提升模型性能。在引入 POS 信息后，模型的准确率由 LAN_KEEOL 的 57.26% 提升至 POS_KEEOL 的 66.46%（见表 4.6）。这一现象表明，在形态与句法标记较为显性、词类区分清晰的语言中，词类信息能够为模型提供有效的语义约束，有助于区分同形异义与多义用法，从而提升词义相关度判断的准确性。

与英语形成对比的是，词类信息在汉语数据上的作用并不显著，甚至呈现负面影响。以 OGWiC_NV 数据集为例，表 4.6 显示，在显式引入词类信息后（POS_KEEOL），模型整体性能不升反降；类似的趋势亦出现在闭源大模型的实验结果中。这一现象与汉语的语言类型特征密切相关：其一，汉语缺乏形态变化，词类界限相对模糊，同一词项在不同句法位置上往往不发生形式变化；其二，在实际语用中，汉语词汇的词类转换较为灵活，词类差异并不必然对应显著的语义差异。因此，显式引入词类标注可能为模型带来额外的噪声，从而削弱其对语义本身的建模能力。

使用提示语直接询问闭源大模型 DeepSeek 和通义千问模型，同样得到类似的趋势。提示语的模板可以参考第 4.2.4 节。图 4.14 显示了两个模型在是否提供了词类信息的提示语下中英文的性能变化。

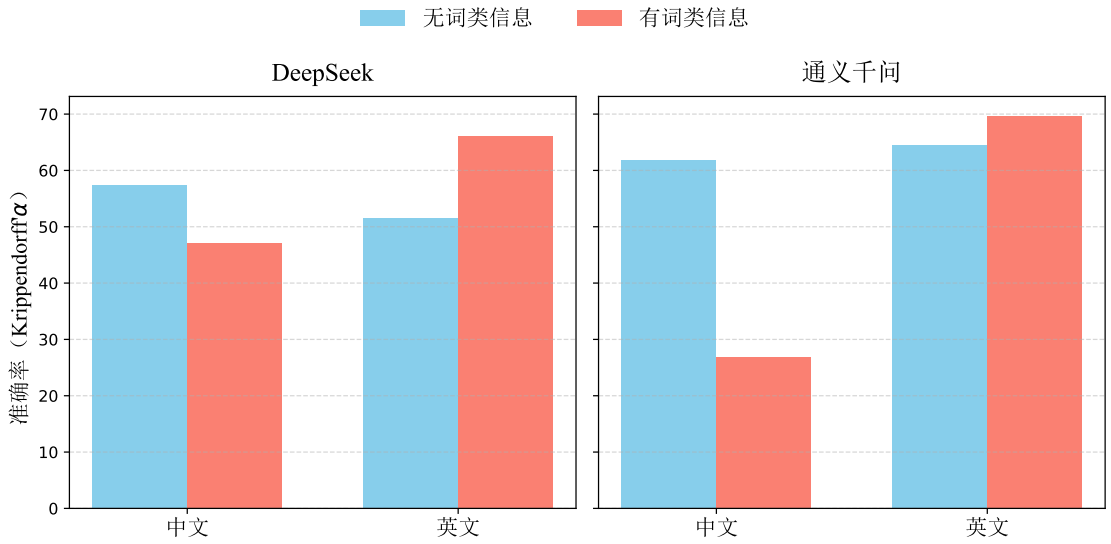


图 4.14　提示语中的词类信息对于闭源大模型的影响

从整体趋势来看，词类信息在英语数据上对模型性能均表现出促进作用。具

体而言，在 DeepSeek 模型中，引入词类信息后，英语数据的准确率由 51.55% 提升至 66.14%；在通义千问模型中，准确率亦由 64.54% 提升至 69.65%。这一结果进一步验证了词类信息在英语语境中的辅助作用，即明确的词类标注能够为模型提供额外的语义区分线索，从而有助于模型更准确地识别不同语义用法。

相比之下，在汉语数据上，词类信息整体呈现抑制模型性能的趋势。在 DeepSeek 模型中，引入词类信息后，准确率由 57.37% 下降至 47.08%；在通义千问模型中，性能下降更为明显，由 61.79% 降至 26.80%。该现象与前述开源模型实验结果保持一致，进一步说明在汉语这一词类边界相对模糊、词类转换灵活的语言环境中，显式加入词类标注可能干扰模型对上下文语义的整体建模，甚至引入额外噪声，从而削弱模型的判别能力。

综上所述，无论是开源模型还是闭源大模型，词类信息对词义相关度判别任务的影响均表现出明显的语言类型依赖性，即在形态标记较为显性的英语中具有积极作用，而在词类界限较模糊的汉语中则可能产生负面影响。这一发现进一步表明，在多语言语义判别任务中，提示语设计需要结合具体语言特征进行针对性调整。

此外，从人工标注结果的分布亦可观察到，在汉语数据中，不同词类用法之间往往仍被标注为较高的一致相关性，进一步说明词类区分并非影响汉语词义判别的关键因素。综合以上分析可以看出，词类信息对模型性能的影响具有明显的语言依赖性：在英语等词类边界清晰的语言中，词类信息能够发挥积极作用；而在汉语等词类灵活、语义更多依赖语境的语言中，显式词类标注未必有助于词义区分，甚至可能带来负面影响。

4.4.6　典型偏误分析

本小节分析模型在两类数据集上容易出错的案例，重点针对英语和汉语两种语言。模型选择表 4.4 中表现最优的四类代表性模型，即传统编码器模型 XLM-R ①、相似任务训练过的模型 XL-Lexeme、使用基础提示语诱导表征的开源大语言模型 LLaMA2_p 和闭源大模型 DeepSeek ②和通义千问。

案例的选择根据模型预测的分数和人工标注的真值的比较进行分类，其中当模型预测的相关度分数大于真值时，称为高估实例；反之，当模型预测的相关度分数小于真值时，称为低估实例。由于相关度分数的选择只有 1，2，3，4 四种，因此对于高估实例来说，模型预测相关度减去真值相关度的可能取值为 3，2，1；对于低估实例而言，模型预测相关度减去真值相关度的可能取值为-3，-2，-1。

① 参照表 4.4 中的结果，这里同样采用倒数第二层，即第 11 层的表征。
② 参照表 4.4 中的结果，温度参数设置为 1。

我们挑选 OGWiC 中汉语和英语所有模型都判定为高估或者低估的那些样本，这些可以视作模型容易出错的案例。对于兼类词数据集 OGWiC_NV，由于模型的准确率普遍更高，所有模型均出错的实例过少，因此选择超过一半的模型出错的那些实例，也就是大于等于 3 个模型。表 4.7 列出了各自数据部分高估和低估部分的样本实例个数，之后我们分数据集对这些出错样例进行归纳分析。对于闭源大模型，还将重点分析模型给出的判断理由，进而推断模型做出相关选择的原因。

表 4.7　模型出错案例数量统计

数据集名称	语言	高估样例	低估样例
OGWiC	汉语	41	67
OGWiC	英语	53	78
OGWiC_NV	汉语	23	27

4.4.6.1　OGWiC 跨语言错误类型分析

综合中文 OGWiC 数据与英文数据的表现，可以发现模型在语义相关度判断中呈现出高度一致的跨语言错误模式。总体来看，高估错误可以归纳为四类，而低估错误可归纳为两类。

（1）高估类错误

模型高估语义相关度主要体现在以下四种情况。

① 词汇边界忽略（Lexical Boundary Neglect）

该类错误表现为模型在语义判断时忽略词汇结构边界，将词与短语或非完整词成分混合处理。对于人类标注者而言，词边界具有重要区分作用；而模型更倾向依据整体语境语义进行判断，从而产生过高相关度。

典型例子为“包装”，如例 (1) 所示。以下每个例子都具有四项内容，a 和 b 分别对应两个上下文，其中目标词使用粗体突出显示，省略号代表原文对于目标词语义理解不重要的部分；c 和 d 分别表示真值预测的标签和模型输出的标签，标签从小到大分别对应不相关、近相关、远相关和完全一致。模型预测共列出 5 个模型的结果，分别指 XLM-R，XL-Lexeme，LLaMA2_p，DeepSeek 和通义千问模型。

(1)　a.　这是送到小村的讯息……你把背包装得满满。

　　b.　……改进包装装潢，使产品不断升级换代。

　　c.　真值：1，1

　　d.　模型预测：2，4，3，2，3

在例 (1-a)中，“包装”与邻近字词组合形成短语结构，其中“包”和“装”并不构成独立词项；而在例 (1-b)中，“包装”为完整名词，指包装商品的材料或形式。尽管标注者一致认为两者语义无关，多数模型却给予较高甚至最高相关度。

类似现象也出现在“滑坡”中，如例 (2)。

(2)　a.　路滑坡陡，呼啸的狂风夹着暴雨……

b.　白河大坝上游保护层发生滑坡。

c.　真值：1，1

d.　模型预测：4，4，4，3，4

例 (2-a)中“滑”和“坡”分别属于不同词汇成分，而例(2-b)中的“滑坡”为完整地质灾害名词。尽管人类标注者一致认为两者不相关，大部分模型仍判定为高度相关甚至完全一致。

以下为通义千问模型对该样例的解释 ①：

In both sentences, '滑坡' refers to the physical phenomenon of land or earth sliding down a slope, driven by natural forces (rain/wind in Sentence 1, earthquake in Sentence 2). The core meaning—geological mass movement—is identical.

两句话中，“滑坡”都指的是土地或土体在自然力（第一句为雨/风，第二句为地震）的作用下沿斜坡下滑的物理现象。其核心含义——地质灾害——是相同的。

可以看出模型完全忽略了词边界，同时将语义核心概括得过于宽泛，仅依据上下文中的地质灾害情境即判定语义一致。

② 词类差异忽视（Part-of-Speech Neglect）

该错误指模型在判断语义关系时忽略词类转换带来的语义层级差异。

例如“包装”在不同词类中的使用，如例 (3) 所示。

(3)　a.　要认真搞好产品的包装，货运部门的职工要增强责任感……

b.　这些苹果……需要包装、批发、装卸、运输。

c.　真值：3，3

d.　模型预测：4，4，4，4，4

例 (3-a) 中“包装”为名词，表示包装行为或结果；例 (3-b) 中为动词，表示包装动作本身。尽管两者存在名物化关系，人类标注者通常给予较低相关度，而模型则

① 模型输出原文为英文，中文部分是翻译后的结果。

普遍判定为完全一致。以下为通义千问模型的解释：

> *In both sentences, '包装' refers to the physical wrapping or packaging of goods (products/apples) for transportation and distribution, serving the same functional purpose in the context of logistics and supply chain.*
>
> 在这两个句子中，“包装”指的是为了运输和分销而对货物进行物理包装，在物流和供应链背景下具有相同功能。

模型解释中使用动名词形式（wrapping or packaging），从侧面体现其忽视词类转换，仅关注语义功能一致性。

③ 语义过度泛化（Semantic Over-generalization）

该类错误表现为模型依赖宽泛的上位概念进行语义归纳，忽略具体语境差异。例如英文中的 record，如例 (4) 所示。

(4)　a.　I was making records, but they weren't getting played…

　　b.　…suppose him to be able to rake up against her out of the records of the Antigua police…

　　c.　真值：1，1

　　d.　模型预测：3，2，2，2，2

例 (4-a) 中 *records* 表示唱片，而例 (4-b) 中 *records* 表示档案记录。人类标注者通常认为两者语义差异明显，而模型倾向于基于“信息记录载体”这一抽象概念，将两者判断为近相关或者远相关。

④ 事件结构简化（Event Structure Simplification）

该类错误表现为模型在判断语义相关度时，未能区分同一词汇在不同事件结构中的语义差异。所谓“事件结构”，指词汇在表达动作或状态变化时所涉及的事件类型、参与者角色以及作用机制。人类在进行语义判断时，通常会综合考虑事件参与者之间的关系以及动作发生的方式，而模型往往更依赖抽象结果语义，从而忽视事件内部结构差异。

例如英文中的 *pin*，如例 (5) 所示。

(5)　a.　*A door was opened, pinning her against the stone wall…*

　　b.　*Vietcong riflemen had pinned down the American units…*

　　c.　真值：2，2

　　d.　模型预测：3，3，4，3，4

在例 (5-a) 中，*pin* 描述的是一种物理接触事件：门被打开后，通过空间挤压将人物

固定在墙上。该事件属于典型的物理附着或压迫行为，其语义核心在于通过直接接触产生空间限制。

而在例 (5-b) 中，*pin down* 表示军事语境中的战术压制行为。该事件并非通过物理接触实现，而是通过火力威慑或战术控制，迫使对方部队无法自由行动。尽管最终结果同样表现为“限制移动”，但其事件参与者、因果机制及动作类型均明显不同。

从人类语义认知角度来看，例 (5-a) 属于物理接触事件框架，而例 (5-b) 属于社会互动或战略控制事件框架，两者虽存在隐喻联系，但仍体现出显著的语义结构差异。因此标注者通常不会将两种用法判断为完全一致。

相比之下，模型往往更加关注结果状态层面的语义相似性，即两种情形均导致“行动受限”，从而忽略事件内部结构与因果机制差异，将其统一归入同一语义范畴。这表明模型在语义判断中倾向采用结果导向的抽象概括策略，而较少考虑事件结构层面的语义组织方式。

综上，高估类错误反映出模型过度依赖抽象语义一致性，而忽略词汇结构、词类及事件结构差异。

（2）低估类错误

模型低估语义相关度主要体现在以下两类情形。

① 字面义保持缺失（Literal Meaning Retention Failure）

该类错误表现为目标词出现在惯用语或隐喻表达中时，人类能够识别其字面语义来源，而模型则倾向将其视为独立新义。

典型例子为“火”，如例 (6)。

(6)　a.　……水、火发电结合，有水用水，无水用火……

　　b.　问起李亨道上任后的“三把火”，他谦虚地说……

　　c.　真值：4，4

　　d.　模型预测：2，1，1，1，2

尽管例 (6b) 中的“（新官上任）三把火”为惯用表达，但其语义来源仍与自然火相关。人类标注者通常认为两者语义一致，而模型往往给出较低相关度。以下为 DeepSeek 的解释：

> *In Sentence 1, ’火’ refers to thermal power generation, while in Sentence 2, it is part of the idiom ’三把火’. The meanings are unrelated.*
>
> 在第一句中，“火”指火力发电，而在第二句中，“火”为成语“三把火”的组成部分，表示初始改革行动，两者语义无关。

可以看出模型虽然能够识别惯用语整体含义，但未能识别词汇内部语义来源的连续性。值得注意的是在翻译这两个目标词的时候，英文都使用了 fire 这个表示自然火的语义，但是模型并没有感知到这个相同点。

类似现象亦见于例 (7)。

(7)　a.　我相信：“真金不怕火”……

b.　工作人员用一块棉垫往打开的注油口一盖，火瞬时熄灭。

c.　真值：4，4

d.　模型预测：3，3，2，2，3

惯用表达中的“火”仍保持自然火语义，但模型未能识别这一语义延续关系。

② 语义域过度分化（Domain Over-differentiation）

该错误表现为模型在不同应用领域中过度区分语义，而人类标注者通常认为其属于统一语义范畴。

例如“火”在不同领域中的使用，如例 (8)。

(8)　a.　建议把水、火、电统管起来。

b.　列车停驶了，火被扑灭了。

c.　真值：4，4

d.　模型预测：2，3，2，2，3

例 (8a) 中“火”指火力发电，而例 (8b) 指自然火。以下为 DeepSeek 的解释：

> *In Sentence 1, '火' refers to thermal power generation, while in Sentence 2, it refers to literal fire. Both involve the concept of fire but belong to different domains.*

模型倾向于依据应用场景划分义项，而忽略其共享的核心语义。

（3）小结

跨语言对比表明，大语言模型在语义相关度判断中存在稳定认知偏向。高估错误主要源于模型过度依赖抽象语义一致性，而忽略词汇结构及事件差异；低估错误则反映模型难以识别语义演变与隐喻连续性。这些现象说明当前模型仍主要基于分布式语境相似性进行语义判断，而未能充分模拟人类词汇语义的组织机制。

4.4.6.2　OGWiC_NV 数据集错误类型分析

（1）高估类错误

OGWiC_NV 数据集主要考察汉语中名词与动词兼类现象下的语义相关性判

表 4.8　OGWiC_NV 数据集中的语义转化类型

语义类型	语义机制说明	例句
施事角色派生	动词事件转指执行该行为的施事者或职业身份	教练得法 / 足球教练
结果产物派生	动作过程转指行为产生的具体成果或文本产物	备份了一份文件。/ 这个软件做了两个备份。 这篇文章太长，只能节录发表。/ 这一篇是读者来信的节录。 老人口述，请人笔记下来，整理成文。/ 课堂笔记。 计划已经呈报上去，领导还没有批示。/ 这个材料上有张局长的批示。
言语内容派生	言语行为转指言语表达内容或评价表达形式	议论纷纷。/ 大发议论。尊称他为老师。/ 范老是同志们对他的尊称。
制度结构派生	行为过程转指制度性或组织性结果实体	编制教学方案。/ 扩大编制。
状态属性派生	行为活动转指由该行为形成的感知属性或心理状态	经专家品味，认为酒质优良。/ 茶叶品味大受影响。 他痛感自己知识贫乏。/ 针灸时有轻微的痛感。
事件名物化	动词动作整体转指一个完整事件或活动	比赛篮球。/ 今晚有一场足球比赛。
隐喻事件转移	物理动作通过隐喻映射转指心理或认知事件	闪过去。/ 闪念。
工具派生	行为动作转指执行该行为所依赖的器具或装置	水流得太猛，闸不住。/ 开闸放水。 把腿跷起来。/ 登在二尺多高的跷上扭秧歌。

断。与普通 WiC 任务不同，该数据集中的目标词往往在两个上下文中分别表现为不同词类形式，其中最典型的是动词与名词之间的转换。该类转换在汉语中具有高度生产性，往往通过语义转指或名物化过程实现，因此为语义相关性判断带来了更复杂的认知挑战。

整体来看，OGWiC_NV 数据集中模型的高估错误主要表现为忽略词类转换所引发的语义结构差异。模型通常倾向于依据共享的核心事件语义对不同用法进行统一处理，而人类标注者则更加关注事件结构在语义派生过程中产生的角色变化与概念重构。通过对高估样例的系统分析可以发现，这类错误并非随机产生，而是呈现出稳定的语义转化路径。

表 4.8 呈现了 OGWiC_NV 数据集中常见的几种语义转化类型，并提供了简要的机制说明和例句。例句中的目标词使用下划线来强调。具体而言，OGWiC_NV 中的兼类现象大体体现为从事件语义向实体语义的多种派生模式。首先，一类常见路径是由动作行为转指行为执行者，即施事角色派生。例如“教练”既可以表示

训练行为，也可以表示从事该行为的职业身份。人类标注者往往将其视为事件与施事角色之间的语义延伸关系，而模型则更倾向于基于共享的行为核心语义将其判断为高度一致。

其次，大量样例体现为事件结果的名物化过程，即行为动作转指行为产生的具体成果或文本产物。例如“备份”、“节录”、“笔记”与“批示”等词语均表现出由动作过程到结果实体的语义转化。这类转化虽然在概念上具有明显的因果关联，但在人类语义认知中仍体现为不同语义层级，而模型则往往忽视结果实体与行为过程之间的结构区分。

此外，部分样例表现为言语行为向言语内容的语义派生，例如“议论”与“尊称”。该类转化通常涉及从表达行为到表达内容的转指关系，体现出语言行为结果的实体化特征。类似地，“编制”等词语则反映出由行为过程向制度结构实体的语义扩展，显示出较强的社会制度语义属性。

除上述路径外，OGWiC_NV 数据中还存在由事件动作转指状态属性的情况，如“品味”和“痛感”。在这类用法中，行为过程往往被抽象为由行为所形成的感知属性或心理体验。与此同时，部分兼类词表现为事件整体的名物化，例如“比赛”，其名词形式指代完整活动事件。另有少量样例涉及隐喻语义转移，如“闪念”由物理快速运动隐喻为心理瞬时认知事件。

最后，一些兼类词体现为行为向工具实体的转化路径，例如“闸”和“跷”。该类语义派生通过行为与执行工具之间的密切功能关联实现语义扩展。

综合来看，OGWiC_NV 数据集反映出汉语兼类词语义结构高度依赖事件语义向实体语义的系统性延伸。这种延伸通常通过转指或隐喻机制实现，并在施事角色、结果产物、制度结构、心理状态以及工具实体等多个语义维度上展开。模型在处理该类语义关系时，往往过度依赖共享的事件核心语义，从而忽视不同派生路径中语义结构层级与概念范畴的差异，最终导致语义相关度的系统性高估。

（2）低估类错误

在 OGWiC_NV 数据集中，模型低估语义相关度的现象同样具有明显的系统性特征。与 OGWiC 数据集中主要涉及多义词不同，OGWiC_NV 的低估错误更多出现在兼类词语义转换过程中。总体来看，该类错误主要体现在模型过度依赖上下文细节、忽视核心语义连续性，以及对词类差异的过度敏感。通过分析具体样例，可以将低估错误归纳为以下三类。

① 上下文差异过度敏感（Context-driven Over-segmentation）

该类错误表现为模型在判断语义相关度时，过度关注上下文语境中的具体功能差异，而忽略目标词所共享的核心语义概念。人类标注者通常更倾向于依据概

念层面的语义一致性进行判断，而模型则容易受到语境角色或语用功能变化的影响。

典型例子为“打算”，如例 (9)。

(9)　a.　通盘打算。
　　b.　在选择工作上你有什么打算？
　　c.　真值：4，3，4
　　d.　模型预测：2，3，3，2，4

尽管两个语境中“打算”分别体现为整体战略规划与个人决策意图，但二者均指向“对未来行动进行预先安排”的核心概念。人类标注者通常认为两种用法高度相关。然而模型更倾向于区分规划范围与语用场景差异，从而降低语义相关度。以下为 DeepSeek 的解释：

> *In Sentence 1, ’打算’ is used as a noun meaning ’plan’ in a general strategic sense, while in Sentence 2, it refers to a personal intention regarding a specific decision.*

可以看出，模型将宏观规划与个体计划视为不同语义层级，但忽视了两者共享的抽象计划概念。

类似现象亦见于“行动”，如例 (10)。

(10)　a.　腿受了伤，行动不便。
　　b.　行动自由。
　　c.　真值：4，4，4，4，4，3
　　d.　模型预测：2，3，3，3，3

例 (10a) 强调身体运动能力，而例 (10b) 强调行动权利与限制状态。尽管语境焦点不同，人类标注者仍认为两者均属于“个体移动能力”的语义范畴。然而模型则依据语境中的生理限制与社会自由语义进行区分。DeepSeek 对此解释为：

> *Both uses refer to physical movement or mobility, but Sentence 1 emphasizes movement ability due to injury, while Sentence 2 refers to freedom of movement.*

该现象表明模型在语义判断过程中往往过度依赖语用语境差异，而未能抽象出统一的概念核心。

② 惯用表达中字面义保留识别失败（Literal Meaning Retention Failure）

该类错误与 OGWiC 数据集中观察到的现象类似，表现为目标词在隐喻表达或

引申用法中仍保留字面语义来源，但模型倾向将其解释为独立新义。

例如“收获”，如例 (11)。

(11)　a.　春天播种，秋天收获。

b.　畅谈学习收获。

c.　真值：3，3，2，3，3，3，3

d.　模型预测：2, 2, 1, 2, 3

在例 (11a) 中，“收获”表示农业生产中的收割行为；例 (11b) 则为学习成果的隐喻表达。尽管后者属于语义引申，人类标注者通常能够识别其语义来源仍与“获得产出”这一基本概念相关。然而模型往往将抽象成果视为独立语义。DeepSeek 的解释如下：

Sentence 1 refers to literal agricultural harvesting, while Sentence 2 refers to metaphorical gains from learning.

可以看出，模型虽然能够识别隐喻机制，但未能捕捉隐喻语义与字面语义之间的概念连续性。

③ 词类差异过度敏感（Part-of-Speech Sensitivity）

该类错误表现为模型在兼类词不同词类用法之间过度区分语义差异，而人类标注者通常更关注概念语义的延续性。

典型例子为“爱好”，如例 (12)。

(12)　a.　他对打太极拳很爱好。

b.　你有什么爱好？

c.　真值：4，4，3

d.　模型预测：3，3，3，3，4

在例 (12a) 中，“爱好”作为动词表示偏好行为；例 (12b) 中则作为名词表示兴趣类别。尽管两种用法词类不同，但均围绕个体兴趣偏好这一核心语义展开。人类标注者通常认为两种语义高度相关。然而模型往往基于语法类别差异降低语义相似度。DeepSeek 对此解释为：

In Sentence 1, '爱好' functions as a verb meaning 'to like', while in Sentence 2, it is a noun meaning 'hobby'. Both refer to personal preference but differ in grammatical function.

这一现象表明模型在处理兼类词时，仍然受到词类边界的显著影响，而未能充分建模词类转换过程中的语义连续性。

综合来看，OGWiC_NV 数据集中的低估错误主要反映出模型在处理兼类词语义派生关系时，倾向于依据语境差异与语法形式进行语义区分，而较少关注核心概念语义的稳定性。这进一步说明，大语言模型在兼类词语义建模中仍存在对语义层级结构与概念连续性刻画不足的问题。

4.5　本章小结

本章评估了不同的模型在跨语言数据集 OGWiC 和中文兼类词数据集 OGWiC_NV 中词义相关度任务中的表现，包括编码型模型和以大语言模型为代表的解码型模型。评估以对词的内部表征和几何探测器的方式为主，兼用针对闭源大语言模型的提示语设计。通过大量的实验，结果表明大语言模型可以有效理解词义，通过设计提示语的方式对大模型进行表征提取，可以达到对该任务进行微调的编码器模型的水平。而零样本训练的大语言模型仅通过外部提示语就可以达到几乎最优的水平。然而它们的理解能力相较于人类仍有较大的差距，大语言模型仍有改进的空间。本章也构建了针对汉语兼类词的数据集，即同一词出现，既用作名词也用作动词的两种句法环境中，借此来考虑词类对于汉语词义理解的影响。通过大模型在提示语中是否加入词类信息，我们可以看到与英语不同的趋势，即词类对于汉语语义理解有一定的消极作用。

本章仅使用不同标注者标注的中位数作为相关度的最终标注，这仅考虑了标注之间的平均状态，还未考虑标注者之间的不一致性和语义选择的不确定性；另一方面对于语言内的特征的利用和评估较少，后面的章节会对这些做进一步的讨论。

第 5 章　多语言词义相关度任务中的模型可解释性研究

5.1　本章导论

第 4 章给出神经网络语言模型可以在一定程度上能胜任词义多义性理解的任务，然而并没有深入探究不同架构模型之间的表征是如何学习到词义知识以及如何做推理的。本章重点探究大语言模型的内部表征随层数演化的行为模式，并通过对比不同架构的模型来提供一种可解释性研究。具体来说，本章通过对比编码架构模型和解码架构模型在跨层词义性能的变化趋势，从各自结构出发验证了关于理解能力和预测能力的不同演变模式的假说，从而对于大语言模型内部表征的功能有了更深入的理解。

本章各节组织如下：第 5.2 节首先介绍两种不同架构的语言模型的结构差异；第 5.3 节介绍关于可解释性研究的实验设计，包括模型选择、表征选取等流程；第 5.4 节和第 5.5 节分别介绍了在英语通用数据集 WiC 和多语言相关度数据集 OGWiC 上的实验结果和分析。最后一节是本章小结。

5.2　不同架构的语言模型对比分析

在 Transformer[16] 的基础上，现代语言模型历经了架构和应用范式的演变。从架构来分类，基本可以分为编码器模型（encoder）和解码器模型（decoder）。从应用范式上来看，任务一般可以分为理解任务和生成任务，早期的模型针对不同的预训练任务来使用不同类型的模型，理解任务常常使用编码器模型，生成任务往往使用解码器模型，或者编码器——解码器架构。大语言模型兴起后，范式转变为使用一套解码器模型既解决理解任务，又解决生成任务。本章详细介绍不同类型的模型，并分析它们的异同。第 2.2.1 节的相关工作部分对比了不同类型模型的架构、训练目标等信息，本章侧重于模型内部的跨层信息流动和架构间的机制对比分析。

5.2.1　编码器模型

编码器模型的经典架构是针对语言理解的深度双向预训练 Transformer 模型，常见的称呼为 BERT（Bidirectional Encoder Representations from Transformers）。它以掩码语言模型为预训练（mask language modeling）的方式，即遮盖掉当前词后，使用周围的上下文预测被遮盖的词，从而使得模型学习到周围上下文对于词义的

理解。

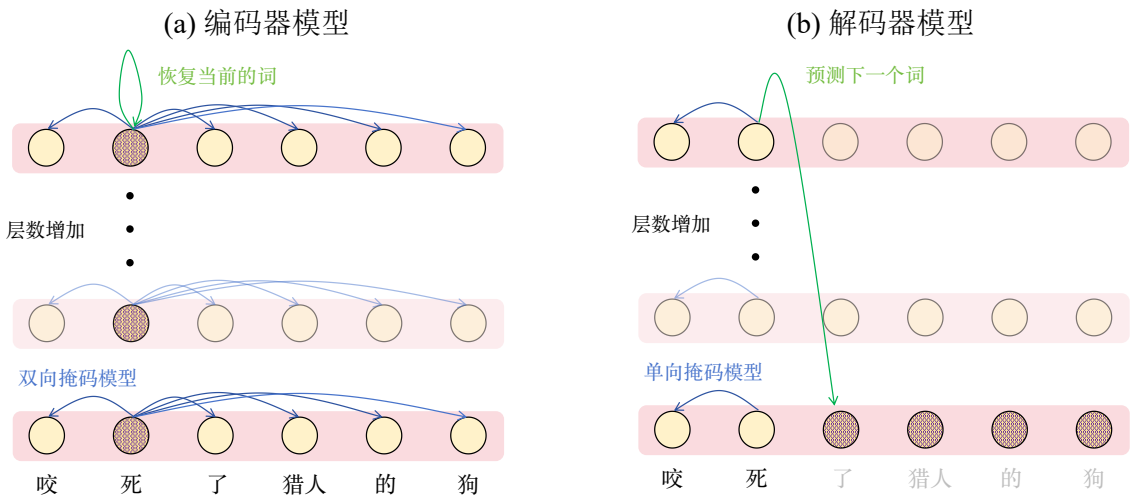


图 5.1　编码器模型和解码器模型的架构对比图

在第 2.2.1 节大语言模型的概述中介绍了几类模型架构的区别（见图 2.1）。图 5.1(a) 则通过具体实例和分层架构更详细地展示了编码器模型的内部注意力架构和信息流动过程 ①。以“死”在“咬死了猎人的狗”中的训练过程为例，掩码（mask）首先遮盖掉目标词“死”，并且通过公式 (2.1) 中的自注意力机制（self-attention），与两侧的上下文进行交互。这一过程由于设计到模型与上下文信息的融合，我们将它称为对目标词的“理解”过程。

之后更新后的向量作为下一层的输入，进行相同的同层之间的注意力操作，这一过程一直重复直到最后一层。在最后一层，目标词的学习目标是恢复当前被掩盖掉的词，即根据“咬 [MASK] 了猎人的狗”来预测中间的词是“死”。通过这一学习目标，模型可以使用上下文向量来代表目标词，也就是作为目标词的语义表征。这一过程则是目标词的预测或者生成过程。

从上述流程中可以看出，编码器模型具有如下的特征：

(1) 用于当前词理解的上下文来自目标词的“历史”（左侧上文）和“将来”（右侧下文）。

(2) 当前词的理解和预测（生成）的信息传递过程是一致的。

也正由于这一特点，编码器模型适合于各类的理解任务，例如阅读理解[205]，命名实体识别[35]，词义检索[206]等任务。

① 为了方便查看，将前文中的模型架构图 4.1 重新放到这里。

5.2.2　解码器模型

解码器模型同样基于分层的 Transformer 架构，典型代表是 GPT 系列模型[105]，其全称为：生成预训练 Transformer 语言模型（generative pre-trained transformer）。它是现在大语言模型标配的架构，并且统一了文本理解和生成任务，称为最先进的架构之一。

解码器模型的架构如图 5.1(b) 所示，同样以“咬死了猎人的狗”为例，目标词“死”被掩码遮盖掉，这时需要它根据历史上文去预测它的下一个词，即“了”。在水平的信息流动上，它的自注意力仅仅关注于历史文本，即“咬”，其后文则在当前目标词看来还未发生，从而并不能作为条件信息 ①。这一过程也被称作是“自回归过程”，因此它的注意力计算过程为：

$$h_i^{l+1} = \sum_{j \leqslant i} \alpha_{ij} h_j^l \tag{5.1}$$

其中注意力权重 α_{ij} 定义为：

$$\alpha_{ij} = \frac{\exp\left((h_i^l)^\top h_j^l\right)}{\sum_k \exp\left((h_i^l)^\top h_k^l\right)} \tag{5.2}$$

符号的定义同编码器模型的公式 (2.1)，最明显的区别是公式 (5.1) 中融合的仅仅是目标词的上文，即索引 j 小于等于当前位置 i 的那一些词。

向量同样经由自注意力机制更新后，进入到下一层，完成对于当前词的“理解过程”。更新后的向量一直到最后一层的输出，之后它的目标则是预测下一个词，即 Next token prediction。在本例子中，它需要预测下一个词是“了”，模型通过预测的结果反馈给模型参数的更新，从而学习到一个可以正确预测输出的模型。

通过观察可以发现，解码器模型具有以下的特征：

(1) 用于当前词理解的上下文仅来自当前和历史上文，在当前词的生成当下，无法利用下文的信息。

(2) 当前词的理解过程和生成过程是不匹配的，“理解”针对的是当前词，而“生成”针对的是下一个词。

也因此，典型的解码器模型适用于生成任务，例如摘要生成[207]、对话生成[4]、篇章续写[208]等。由于“理解”型任务需要观测到周围的上下文，典型的解码器模型不擅长处理理解型任务。

① 这一过程模拟了语言的实际生成过程，即语言是通过线性序列的方式按顺序逐一生成的，当在讲述当前词的同时，无法把将来的话语也说出来。这也是“语言模型”最初的训练目标，即 N 阶的马尔可夫过程。

5.2.3　编码架构和解码架构的对比

表 5.1 对比了编码架构和解码架构的差异，分别从观测的上下文、训练目标、理解和预测过程是否对齐，处理的典型任务这几个维度进行了对照。

表 5.1　不同模型架构的特征对比

架构名称	观测上下文	训练目标	理解预测对齐	典型任务
编码器模型	双向上下文	双向掩码预测	是	理解型任务
解码器模型	单向上文	下一个词预测	否	生成型任务
大语言模型	单向上文	下一个词预测	否	理解和生成任务

然而，大语言模型作为一个典型的解码器模型，它使用生成范式去完成理解任务，从而实现理解任务和生成任务的统一。例如针对多义词相关度任务，大模型不再直接针对目标词进行语义理解，而是通过设计人类可读的提示语，例如第 4.2.4 节的文本框所显示的方式，间接询问大语言模型对于词语的理解。这种基于提示语的方式在各大理解任务上都取得不错的结果，这也在上一章词义相关度的评测表 4.4 中 DeepSeek 和通义千问模型不俗的结果中有所体现。

因此，基于上述差异，本章探讨三个核心问题：大语言模型的跨层表征是否呈现特定行为模式？理解任务与生成任务在模型内部表征中如何交互与分离？与传统编码器模型相比，基于解码器结构的大语言模型的语义表示有哪些显著差异？本章提出如下假设：大语言模型在低层（靠近输入）编码当前词的语义，而高层（靠近输出）则逐渐聚焦于预测下一个词，这一“先理解、后预测”的层次化模式反映了生成式模型中理解与预测信息流的动态交互[166,168]。

为验证假设，本章通过分析每一层隐藏状态在词义相关度任务上的表现进行系统实验，数据集来自上一章使用的 OGWiC 数据集及仅针对英语的通用二分类 WiC 数据集[41]。在探测过程中，通过调整模型层数、目标词位置、上下文长度以及提示策略，评估表征对当前词语义的理解能力以及预测能力。

5.3　可解释性研究实验设计

5.3.1　任务定义和数据集

本章同样针对词义相关度任务，目标是揭示预训练模型在这一任务上跨层的行为模式表现，以深入探究模型的可解释性。数据集采用多语言词义相关度任务数据集 OGWiC 和另外一个英语通用词义相关度任务数据集 WiC，对它们的详细介绍请参考第 3.2.2 节和第 3.2.1 节部分。

本章选择 OGWiC，是由于它是一个词义相关度很典型的一个数据集，并且涉及到多种语言，适合从多语言的角度进行对比分析。然而它的数据集的语料库多来源于不同时期的新闻杂志，语域范围较窄，且单一语言下的数据规模有限。因此选择 WiC 数据集进行补充，WiC 数据集仅涉及英语，但是语料来源广泛，标签分布平衡，数据规模大，且是语言模型评测[10]标准的通用数据集，可以有效补足 OGWiC 的缺点。本章没有采用另外一个词义相关度的数据集，即中文兼类词数据集 OGWiC_NV，原因是该数据集仅涉及到具有不同词类的目标词，范围较窄，同时标签分布较为不均衡，规模也较小。

评估方面，WiC 数据集使用公式 (3.5) 的准确率 ACC；OGWiC 数据集采用公式 (3.6) 的基于分布的准确率克里彭多夫 α 系数。

5.3.2　评估模型

本章实验从模型架构的角度对两类模型进行系统性探测：一类为**编码器模型**，以双向掩码语言模型 BERT 为代表；另一类为**解码器模型**，以大语言模型为代表。针对多语任务集 OGWiC 与英语任务集 WiC，在模型选择上存在一定差异。表 5.2 列出了不同任务条件下所采用的具体模型配置，其中 B 代表基础版本 Base，L 代表参数加强版本 Large，R 代表 RoBERTa 模型。表 5.3 则展示了各个模型的层数信息。

表 5.2　两个数据集上探测的不同类型的模型

数据集	编码器模型	大语言模型
OGWiC	XLM-B, {SLM-B}, XLM-R, mBART, mT5	LLaMA-2/3
WiC	BERT-B/L, RoBERTa-B/L, BART-B/L, T5-B/L, Elmo	LLaMA-2/3

对于多语数据集 OGWiC 来说，编码器模型选择跨语言的 BERT 模型 XLM-B（XLM-BERT-base），XLM-R（XLM-RoBERTa-base）以及编码器——解码器过程中的编码器模型 mBART（mBART-large-50）和 mT5（mT5-base）。对于 OGWiC 中的各个语言的数据，本章采用各自语言对应的 Bert 模型，这样做是由于在上一章的实验中，单语的模型结果要好于跨语言模型的。在表格中使用模型集合 {SLM-B} 代表各自单语的 BERT 模型，即 Single-language Language Model。解码器模型中的大语言模型则选择了 LLaMA-2 的两个参数量不同的模型版本，分别为 LLaMA-2-7b-hf ①和 LLaMa-3-8b ②。在表格和后文中，简记为 LLaMA-2 和 LLaMA-3，模型的详细介绍请参考第 4.3.2 节中的模型介绍部分。

① https://huggingface.co/meta-llama/Llama-2-7b-hf
② https://huggingface.co/meta-llama/Meta-Llama-3-8B

对于 WiC 的单语实验设置，编码器模型选择经典的双向掩码模型作为基本架构，并采取基础版（Base）和参数更大的版本（Large），其中包括 BERT-Base①和 BERT-Large②，RoBERTa-Base③和 RoBERTa-Large④，BART-Base⑤和 BART-Large⑥，T5-Base⑦和 T5-Large⑧。第 4.3.2 节中的模型介绍部分详细介绍各自模型，这里不再赘述。

此外，本章还引入了若干基于词级上下文表示的方法，包括 LMMS[209]、Context2vec[210] 以及 ELMo[211]，这些模型均作为对比方法在数据集原论文[41]中被比较。LMMS 方法将词义消歧方法应用到 WiC 任务中，通过判断消歧后的选择选项来对 WiC 做出判断。Context2vec 则通过在双向 LSTM[110]的头部增加多层感知机来表征目标词所在的上下文。ELMo 则是字符级别的学习动态上下文的优良模型。对于解码器模型，与 OGWiC 一致，WiC 也选择 LLaMA-2 的两个不同量级参数的模型，分别为 LLaMA-2-7b-hf 和 LLaMa-3-8b。

同时，在 WiC 数据集上还设立了两个基准方法，分别是人类基准[37]和随机选择。人类基准是指人类标注者在标注实例时的一致率。尽管这个指标没有一个明确的标准，在多数文献中采用 80%[89]作为人类标注者一致率的上限，并声称模型无法超越这一上限，或者即便超越上限也无法得到合理的解释。随机选择基准则是随机选择 0 和 1 两类标签所得到的结果。由于 OGWiC 数据集中 0 和 1 的标签数量是等同的，因此随机选择的结果是 0.5。

5.3.3　表征选择和探测方式

本章涉及的表征选择和探测的方式大致与第 4.2 节相同。对于不同的模型，首先需要提取目标词的表征，对于编码器模型而言，需要提取当前词所在位置的隐向量；对于生成式大语言模型，有多种策略，包括仅提取目标词的 basic、重复上下文并提取第二次出现的上下文的 Echo、提取前一个目标词的 Prev、以及使用提示语的 PromptEoL 及其变种，详情可以参考表 4.1和表 4.2。同样为了避免各向异性的向量分布，本章同样采用了去各向异性的手段，主要采用了标准化（Standardization）的手段。由于本章侧重于研究跨层的表征的行为模式，在原有的设定中，需增加一个层数的变量，即**针对模型的每一层后的隐向量都做词义探测**。

① https://huggingface.co/google-bert/bert-base-uncased
② https://huggingface.co/bert-large-uncased
③ https://huggingface.co/FacebookAI/roberta-base
④ https://huggingface.co/FacebookAI/roberta-large
⑤ https://huggingface.co/facebook/bart-base
⑥ https://huggingface.co/facebook/bart-large
⑦ https://huggingface.co/google-t5/t5-base
⑧ https://huggingface.co/google-t5/t5-large

表 5.3　各类模型各层的起止索引

模型	总层数	低层	中层	高层
BART_B	6	[0,2]	[3,4]	[5,6]
XLM-B | {SLM-B} | XLM-R | BART-L | R-B | T5-B	12	[0, 4]	[5, 8]	[9, 12]
BERT-L | RoBERTa-L | T5-L	24	[0, 8]	[9, 16]	[17, 24]
LLaMA-2/3	32	[0, 10]	[11, 20]	[21, 32]

为了更好地描述与层数相关的表现，将模型的层数从最靠近输入的最低层到最接近输出的最高层均分为三部分，分别表示低、中、高层，并使用红色、绿色和蓝色对它们进行区分。按照惯例，最低层的层号索引从 1 开始计算，随层数增加索引依次增大，第 0 层表示还未经过模型运算的输入层。表 5.3 展示了各类模型各层的索引序号集合，其中中括号表示层数索引的闭区间，例如 [0,2] 代表第 0，1，2 层。

再提取表征之后，本章采用几何探测的方式完成词义相关性任务。对于 OGWiC 任务而言，本章采取 Nelder-Mead 算法来获取阈值，其设置参考第 4.2 节的介绍。对于 WiC 任务，由于它是一个二分类的任务，仅需要在向量的相似性计算之后设置阈值 γ 即可。具体而言，对于出现在上下文 c 中的目标单词 w[①]，由模型针对每一层 l 来提取表征 $h^l \in \mathbb{R}$。之后计算单词 w 在不同上下文（c_i，c^j）中的余弦相似度 m_{ij}^l，余弦相似度计算可以参考公式 (4.1)。随后，如果 m_{ij}^l 超过阈值 γ，即认为单词 w 在两类上下文中的语义一致，分类结果为真，如果低于 γ，则分类为假。

为确定判别阈值参数 γ 的最优取值，基于 WiC 数据集的验证集对模型在不同层级上的表现进行了系统搜索。结果表明，γ 并非在各层之间保持一致，而是随网络深度呈现明显差异，说明模型在不同抽象层级上对语义相关性的敏感程度存在显著不同。

图 5.2 展示了逐层几何探测的流程。以“死”在两个上下文“死了一口猪”和“方法太死”为例[②]。目标词在每一层都被提取向量，并且两个上下文相同层得到的目标表征进行计算相似度 m_{ij}，之后与层相关的阈值 γ^l 进行比对，大于该阈值的输出为标签 1，表示语义相同，否则输出标签 0。注意这里我们没有进行跨层计算相似性，这主要考虑到同层的表征具有相似的向量空间，因此可以直接比较。

① 我们将单词内所有 token 的表征做平均后再作为最终单词的表示。
② 例句选择《语法讲义》[212] 第 39 页。

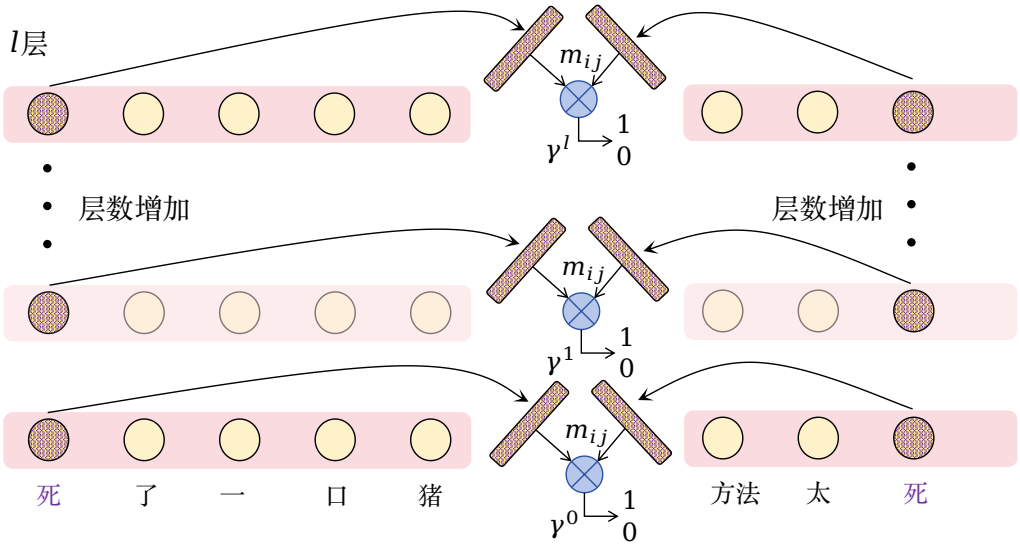


图 5.2　逐层几何探测流程图

5.4　英语通用数据集 WiC 上的实验结果和分析

本节列出了各类模型方法在数据集 WiC 的表现，重点分析了编码器模型和解码器模型在跨层表现上的特征，之后通过消融实验来进一步分析实验结果。

5.4.1　WiC 上的准确性结果分析

表 5.4 给出了 WiC 测试集上包含名词实例和动词实例在内的整体准确率，括号中的数字表示模型取得最优性能时所对应的层索引号，数字的颜色反映了取得最佳效果的层高低，其中红色表示低层，绿色表示中层，蓝色表示高层，为了方便查看，请使用电子版或者彩色打印版。加粗和下划线标识了除了人类基准之外准确率第一和第二的方法。

整体来看，LLaMA 系列为代表的解码大语言模型在各类提示语设定下均优于编码器模型以及非自回归的传统表示方法（如 LMMS、Context2vec 和 ELMo），显示出大语言模型在词义相关度判断这一细粒度语义理解任务上的显著潜力：尽管大语言模型并未显式以词义为监督信号进行训练，其上下文建模能力仍能够有效支持词级语义区分。其中，使用参数量更大的 LLaMA3 模型的性能要普遍更优于参数量更小的模型版本 LLaMA2，这可能得益于 LLaMA3 模型在规模更大、质量更佳的语料上训练，并且在语义理解、逻辑推理等下游任务中，LLaMA3 模型也要显著优于 LLaMA2 ①。而比较不同类型的提示语可以发现基础版本的 PromptEoL 就可以取得不错的结果，在 LLaMA3 模型下甚至取得了最优的结果，而其他变种的提示语与 PromptEoL 仅保持可比的性能，有些甚至损害了最终结果。

① https://huggingface.co/meta-llama/Meta-Llama-3-8B

表 5.4　WiC 测试集在各个方法上的准确率

方法	所有实例	名词实例	动词实例
人类基准	80.0	-	-
随机选择	50.0	-	-
LMMS	67.7	-	-
Context2vec	59.3	-	-
ELMo	57.7	-	-
BERT-Base	68.36 (11)	68.71 (11)	67.84 (11)
BERT-Large	70.21 (22)	71.36 (22)	69.42 (21)
RoBERTa-Base	67.21 (11)	69.31 (8)	66.08 (9)
RoBERTa-Large	69.21 (14)	72.08 (16)	66.08 (13)
BART-Base	63.86 (5)	67.03 (6)	60.11 (5)
BART-Large	67.64 (9)	69.68 (12)	66.08 (9)
T5-Base	64.86 (12)	67.99 (12)	60.63 (9)
T5-Large	68.14 (17)	71.24 (17)	64.67 (19)
LLaMA2_Basic	63.57 (12)	66.43 (12)	60.28(13)
LLaMA2_Echo	68.14 (8)	72.28 (8)	65.73 (11)
LLaMA2_PromptEoL	72.71 (19)	74.01 (17)	73.64 (24)
LLaMA2_PCoTEOL	73.21 (22)	74.37 (24)	72.23 (27)
LLaMA2_KEEOL	71.14 (25)	72.56 (20)	70.65 (22)
LLaMA2_POS_KEEOL	70.64 (23)	72.32 (23)	68.54 (24)
LLaMA3_Basic	64.21 (10)	67.15 (10)	60.11 (13)
LLaMA3_Echo	68.71 (7)	71.72 (7)	64.67 (6)
LLaMA3_PromptEoL	**76.43** (24)	76.65 (23)	76.98 (24)
LLaMA3_PCoTEOL	75.64 (21)	**77.14** (20)	76.63 (24)
LLaMA3_KEEOL	76.07 (23)	76.29 (23)	**78.03** (29)
LLaMA3_POS_KEEOL	74.43 (25)	75.09 (18)	75.57 (24)

然而，大语言模型仅在提示语设定下可以发挥出卓越的性能，在原始上下文的当前目标词提取表征（Basic）则表现出较差的水平，远低于其他编码器模型，例如 LLaMA3_Basic 的性能仅为 64.21%，而 BERTA-Base 的表现则为 68.36%。其 Basic 设定下，模型在中低层达到最佳值。与之前的实验结果（参见表 4.4）类似，大语言模型需要一段合适的提示语“诱导”它**生成**与词义相关的表征。这也再次提醒研究者在利用生成式解码器模型的表征来表示当前目标词时，不宜像文献[213]中那样，直接采用当前词的最高层表征。

相比之下，将上下文重新复制一次的重复（Echo）策略在不依赖复杂提示模板的前提下，取得了接近提示策略的性能，且显著优于基础版本。例如，LlaMA2_Echo 在归一化设置下达到 68.14%，较 LLaMA2_basic 提升约 4.5 个百分点。该策略通过对关键语境进行结构化重复，增强了关键信息在表示中的可达性，在提示鲁棒性与信息聚焦之间取得了良好平衡。这一结果表明，即便不引入显式任务指令，仅通过输入结构的轻量改造，也能够有效引导模型关注词义判别所需的核心上下文。

从不同模型类别的对比结果可以观察到，传统静态或弱上下文化方法（Context2vec 与 ELMo）的性能显著落后于其他神经网络语言模型，其中 ELMo 仅取得 57.7% 的总体准确率，说明仅依赖浅层或局部上下文难以捕捉词义的细微差异。BERT-Large 作为双向编码器模型取得了编码器模型中最优的结果，验证了双向编码结构对词义向量生成的积极作用。然而，其性能仍远低于最优的大语言模型设定 6 个百分点之多，表明生成式模型在处理跨句对比式语义判断时具备更强的优势。

在词性层面，名词实例整体上比动词实例更容易判别。例如，在 LlaMA2_Echo 设置下，名词准确率为 72.28%，而动词为 65.73%，二者差距达到将近 7 个百分点；在基础设置 (LLaMA2_Basic) 中，动词的性能下降更为明显，其准确率较名词低约 6 个百分点。这一现象反映了动词词义对上下文依赖程度更高的语言学特性。进一步的统计显示，19.2% 的目标动词接近于句首位置，由于缺乏足够的前文语境，模型难以获取用于消歧的关键信息；相比之下，名词仅于句首的比例仅为 14.3%。因此，动词在语境受限条件下更易产生语义模糊，该结果与既有研究中“动词更难进行词义消歧”的结论一致[214]。

此外，不同模型在最佳层数上的分布也揭示了语义信息在网络内部的表征差异。编码的最优性能主要集中在高层，表明词义判别依赖于深层语义抽象；而大语言模型的最优层分布与提取表征的方式相关：基础设定（Basic）和重复（Echo）设定中，最佳性能出现在中低层；而在使用了提示语的设定中，最佳性能则出现在中高层，这些现象都暗示了生成式解码器模型编码语义区分信息所在层数的差异。这种分布差异进一步说明了两类模型在内部语义建模机制上的本质不同。后面一节会对这一现象进行详细的分析。

最后，与人类基准（80.0%）相比，当前最佳模型仍存在约 3–4 个百分点的差距，说明词义相关度判别这一任务在真实语境下依然具有挑战性。总体而言，生成式大语言模型在词级语义理解方面已显著优于传统上下文化方法与双向编码器模型，而合理的输入结构设计（如提示或重复）是激发其语义判别能力的关键因素。

5.4.2　模型跨层性能的模式分析

不同类型以及不同表征选择方式的模型在跨层性能上表现出差异，这体现了模型在不同层的表征行为有所区别，本节通过分析几组控制对照实验来验证和量化这一表征行为差异，从而对模型的可解释性做出启发。

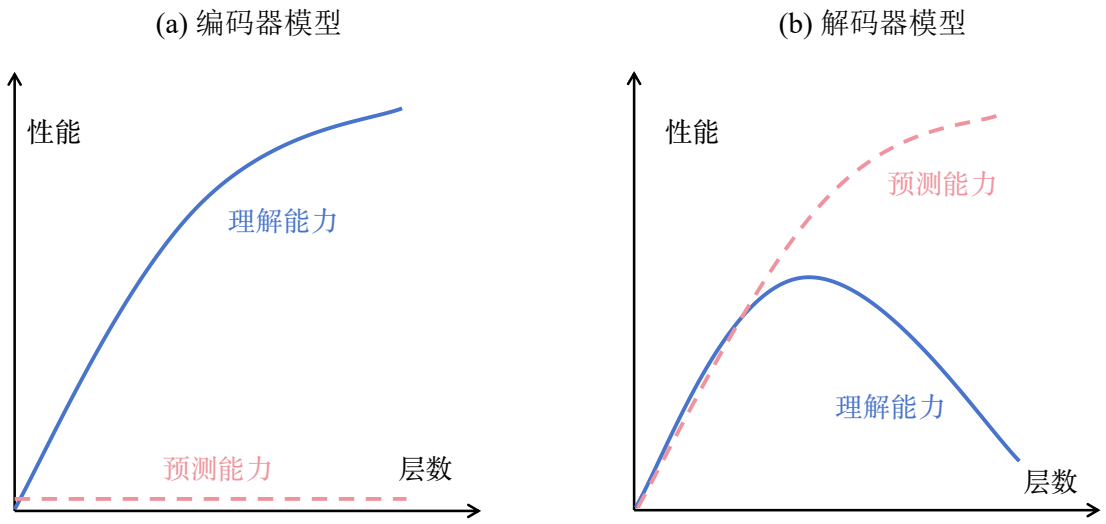


图 5.3　两类模型的理解与生成能力逐层变化

为了与图 5.1中展现的编码器模型和解码器模型的架构进行对比，本节提出如下假说，并给出相应的示意图 5.3：

编码器模型的对当前词的词义理解能力随着层数的增高而增加，至最高层达到最优的理解能力，它同时不具备预测未来文本的能力；解码器模型的词义理解能力随着层数先增加后降低，在中低层达到拐点，与此同时，其预测后文的能力逐层增加，至最高层达到最优的预测能力。

接下来则通过实验来验证这一假说，其中分为四个方面，分别从编码、解码器模型在理解能力和预测能力两个维度来进行设计。

（1）编码器模型的理解能力分析

图 5.4 展示了几类编码器模型在 WiC 数据集上的跨层准确率变化，左侧显示了基础 Base 模型，右侧是相同架构的参数增强模型 Large，编码器模型包括 BERT（蓝色圆形），RoBERTa（红色方形），BART（绿色三角形）和 T5（紫色菱形）模型，横轴表示层的索引号，有些模型由于层数较少，曲线未达到最大的层索引而提前终止，纵轴表示模型在 WiC 测试集上的准确率，星号标识出了最佳性能。

从图中可以清晰看出编码器模型都具有统一的变化趋势，即随着层数增加，准确率逐层增加，在高层达到最优的性能，这表明编码器模型对当前词的理解能力逐步增强。这与模型训练的方式息息相关，即最后一层的目的是根据周围上下文

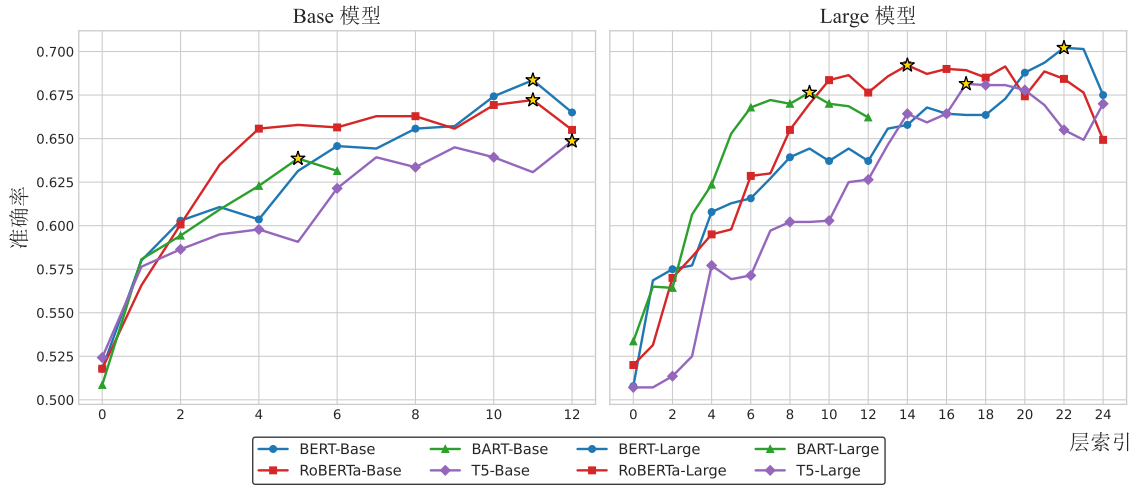


图 5.4　编码器模型随层数增加的性能变化

去恢复当前被掩盖的噪声词，即越高的层越可以代表当前的目标词，否则从节省资源的角度就无需堆叠如此多层了。变化趋势可以验证图 5.3(a) 中关于编码器模型关于“理解能力”的假说①。

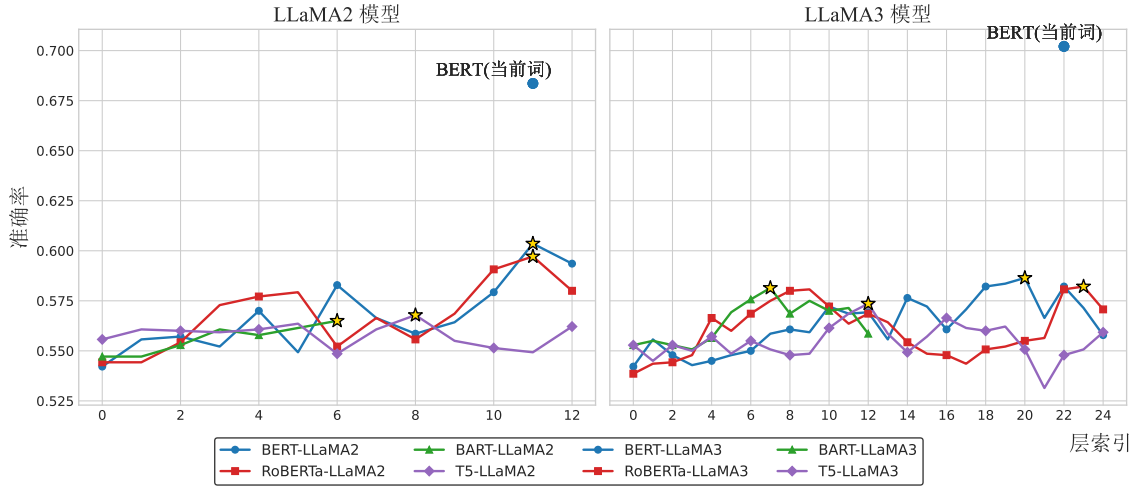


图 5.5　编码器模型针对前一目标词表征性能随层数的变化

（2）编码器模型的预测能力分析

为了探索编码器模型的预测能力，我们提取当前目标词的前一词的表征来代表目标词②，来探测在 WiC 任务上的表现性能。这一设定的出发点是如果模型的预测能力较强的话，前一个词的表征就可以反映当前目标词。图 5.5 呈现了这一结果。图中的模型和图例含义同图 5.4，此外，还标出了使用当前目标词的表征在

① 假说中提到的“最高层达到最优的理解能力”有些过强了，除了 T5-Base 模型，剩余模型都没有严格在最后一层达到最优，但是所有模型无疑都在高层达到最优，关于最高层没有达到最优是很多论文[154]提出的经验共识，其原因和机制有待后续探索。

② 为了同后文解码器模型使用前一个词的预测能力相对比，这里同样将上下文重复一遍，提取第二个上下文目标词的表征。

Base 模型和 Large 模型中最佳的性能以及对应的层数，以此来作为性能参考，如表 5.4 中展现的性能，BERT-Base 和 BERT-Large 分别表现最佳。

从图中可以看出，不同模型取得的最佳值并无明显规律，有些处于低层，有些处于高层。同时，它们的表现都比较差，远低于使用当前词获取表征的最优结果，这表明编码器模型几乎不具备预测能力，使用前一个词无法有效预测下一个词。这个实验验证了图 5.3(a) 中关于编码器模型关于“预测能力”的假说。①

（3）解码器模型的理解能力分析

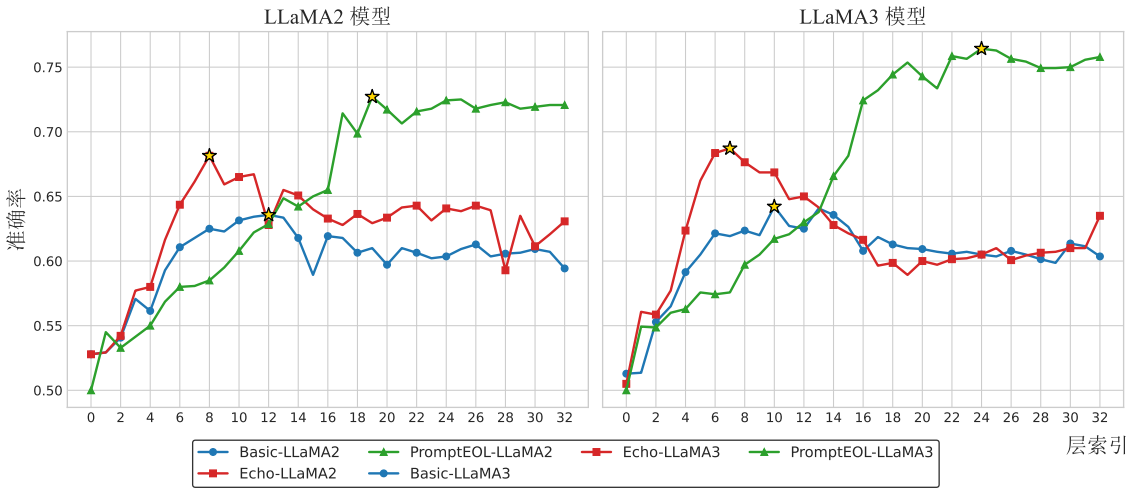


图 5.6　解码器模型随层数增加的性能变化

大语言模型 LLaMA2 和 LLaMA3 的基础设定 Basic 和重复上下文策略 Echo 由于仅提取当前词的表征，因此可以反映解码器模型对当前词的理解能力逐层变化。图 5.6 展示了这两个模型在两个设定下跨层的表现能力，其中左侧模型输出了 LLaMA2 模型的性能表现，右侧展示了 LLaMA3 模型的性能表现。红色、蓝色分别表示 Basic 和 Echo 设定，为了对比分析性能，同时用绿色的线展示了使用提示语并提取最后一个字符的设定 PromptEOL 的结果。

图 5.6 的结果显示，对于提取当前词表征的 Basic 和 Echo 模型来说，其性能在低层逐层增加并达到顶峰，而之后在中高层模型的性能则随着层数增加而下降，从而呈现“倒 U 型”的趋势。这一趋势在两个参数量大小的模型中可以清晰地体现出来，这表明解码器模型的理解能力主要集中在中低层，这与图 5.3(b) 预期的理解能力变化趋势一致。Echo 设定通过重复操作使得模型在获取当前表征时可以获得全部的上下文，使得它的性能较之 Basic 更高，然而它们都没有对应的提示语诱导下的结果更好，表明这种方式仍不能完全激发大语言模型的潜力。

① 值得注意的是，这种探测本身具有一定的局限性，因为受限于上文的信息量，目标词可能无法保证会由前文预测出来，因此预测结果低于使用当前词做表征的情况是预料之中，但是由于编码器模型从机制上并不是做预测任务，无法使用像解码器模型一样根据前文生成后文。因此这个预实验可以在一定程度上反映预测能力，尤其当模型的结果显著较差的时候。

（4）解码器模型的预测能力分析

既然模型在高层的理解能力有所下降，解码器模型在高层做了什么？本文假设模型在高层集中于其预测生成能力。通过两种设置来验证这种假说，第一种是使用当前词的前一个词（Prev）并通过基础设定 Basic 和重复策略 Echo 来探测模型性能，出发点与编码器模型的预测能力相同。第二种则是使用各类提示语来诱导模型生成目标词的表征，由于下面展示的提示语模板的最后一个字符是冒号，其本身与目标词的语义毫无关联，然而由于提示语本身是为了预测目标词的语义，因此可以通过冒号的表征来代表预测能力。

In this sentence “river bank”, “bank” means in one word :

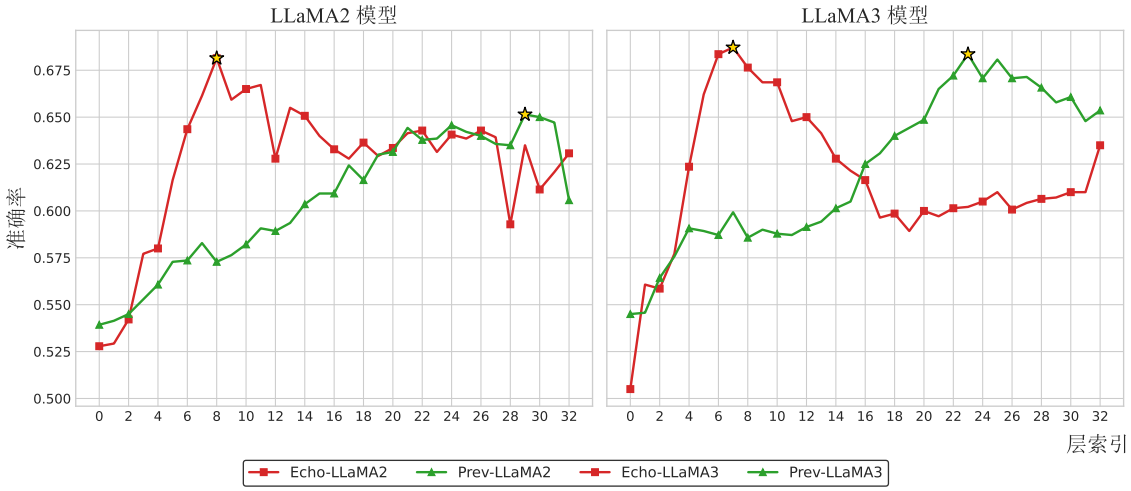


图 5.7　解码器模型针对前一目标词表征性能随层数的变化

图 5.7 展示了使用前一目标词预测的实验结果，为了对照，图中也输出了以当前目标词进行提取表征的结果，即 Echo ①。可以明显看出，前词表征 Prev 的变化趋势在两个解码器模型中均呈现上升趋势，并在最高层达到最优的结果。这一结果与使用当前表征的 Echo 设定结果相当，尤其在 LLaMA3 模型中，二者几乎差别不大。考虑到 Prev 设定所使用的表示并非来自真实的目标词，这一结果更加表明了模型在高层实现预测的功能，因为前一个词的预期输出是下一个词。对比提取当前词表征的 Echo 设定表现出的“U 型”趋势，这一特征更加突出。

图 5.8 展示了各类提示语诱导下的提取末尾词表征性能随层数的变化。为了对照结果，还标识出了仅提取目标词表征的两种设定 Echo 和 Basic 的最佳性能以及对应的层数，分别用红色方块和蓝色圆圈来表示。可以看出，虽然最佳的层数有所不同，但是所有的提示语都具有统一的变化趋势：随着层数增加，性能单调递增，在高层性能达到饱和。且最终的结果远远超出了 Basic 和 Echo，且最佳值来自高

① 由于 Prev 设定中，采用的是重复两次的操作，作为对比，这里也同样展示了重复两次提取当前词的设定，而不是仅从当前文本提取表征的 Basic 设定。

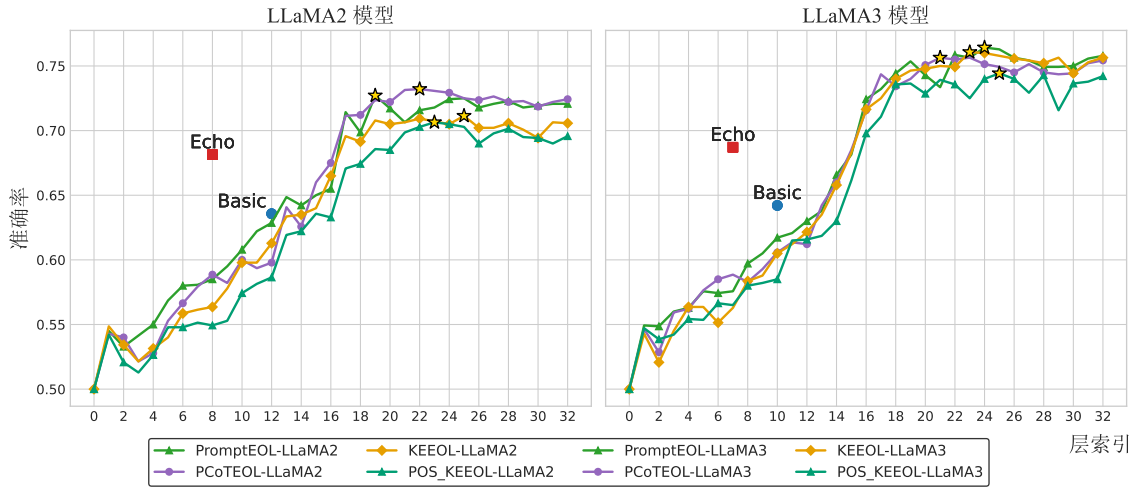


图 5.8　各类提示语对于解码器模型的逐层性能影响

层。考虑到当前提取表征的词仅是本身毫无语义特征的冒号，但是提示语本身就要预测产生的恰恰是原来目标词的词义，因此可以推断出，模型在高层主要是完成预测的功能，即如图 5.3(b) 所展示的那样。

（5）总结

上述实验结果和分析印证了图 5.3 中的假说，即编码器模型在高层捕捉当前词的词义，而解码器模型在中低层捕捉当前词词义，而在最高层预测后文，词义转移到下一个要预测的词。值得注意的是，这种“先理解再预测”的模式在人类的行为认知和脑科学中也较为普遍，例如人类在聆听音乐的时候在感知和理解当前音符的同时也在预测下一个音符 ①；在一次会话中，听者也会在理解说者话语的同时，预测对方的意图和接下来要说什么[215]。这也体现了一种智能普适性。

大语言模型作为一个典型的生成模型，在它训练过程中并未针对理解型任务，然而通过“预测下一个词”的训练目标也让它学习到对当前词的理解。这看似矛盾，然而生成模型的构建本身是要以理解为前提的，正如人工智能之父辛顿在访谈中指出，虽然大语言模型被训练为“预测下一个词”，要实现高质量的预测就必须对语义和上下文有深入的理解，否则无法进行连贯生成；因此认为这些模型“仅仅是在预测下一个词而不具备理解”是不合理的 ②。不过，获取这种代表目标词的表征仍旧需要一种生成式的方式，也就是采用提示语诱导的方式，即向模型表明“下一个词是目标词”的意图，这与各类理解型任务通过向大语言模型的输入格式靠拢而非相反的范式，来实现任务统一的大趋势一致[105]。例如在情感分类任务中，以往的“预训练—微调”范式是让模型的表征在相似任务上训练，即预训练模型需要迁就于任务，而在大模型时期则是设计任务格式，作为提示语输入到大

① https://www.ucsf.edu/news/2024/02/427116/to-appreciate-music-human-brain-listens-and-learns-to-predict
② https://www.cbsnews.com/news/geoffrey-hinton-ai-dangers-60-minutes-transcript/

语言模型中，即任务迁就于（大语言）模型。

5.4.3　其他消融实验分析

本节介绍与实验相关的其他消融实验，包括典型模型的阈值参数的选择以及去各向异性手段的效果。

（1）阈值参数的选择

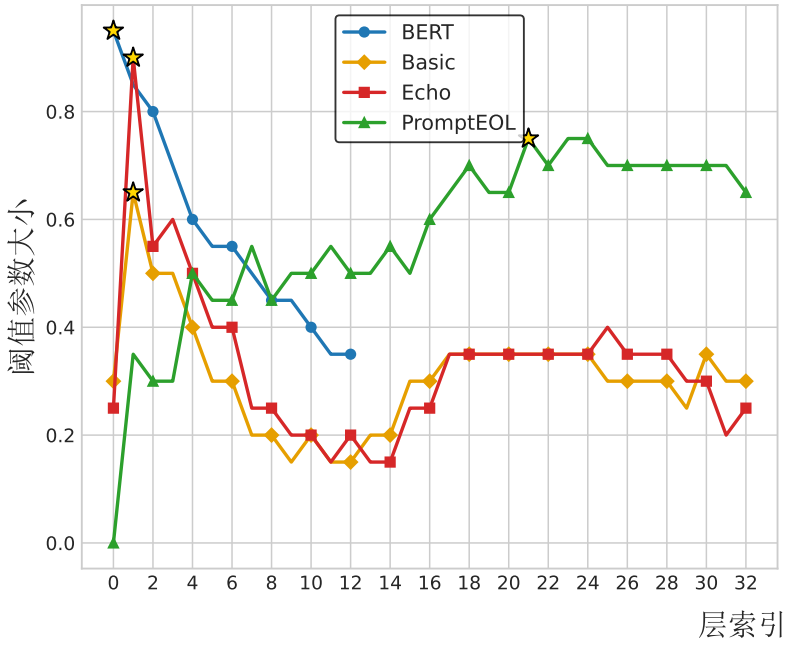


图 5.9　不同模型各层之间最佳阈值参数的变化

图 5.9 展示了 BERT_Large 模型（记作 BERT）和 LLaMA3 模型的三种设定（Basic、Echo 和 PromptEOL）在各自层数选择的最佳阈值的变化。由于阈值决定了相关度的归属，即超出该阈值模型判定为语义一致，否则语义不一致，因此阈值大小反映了相似度的普遍程度，可以看出在基于提示语 PromptEOL 的设定中，阈值在高层都很高，这反映了高层表征之间的相关性普遍较大，因此需要较大的阈值将它们区分开，而在其余三个模型中都是低层的相似度较大，之后逐渐减弱。这些模型在各个层上变化较大的阈值也反映了不同层之间向量分布的差异较大，因此需要逐层求出相关的阈值来分别判断准确性。这也提醒研究者，不能直接使用某一层的阈值来作为所有层分类的依据，仍旧需要逐层求得阈值（layer-wise threshold）。

（2）去除各向异性手段对结果的影响

去除表征之间的各向异性的手段已经在第 4.4.2 节展示的 OGWiC 相关的实验中证明了其有效性，本节继续考察它对于 WiC 数据集的作用。在四种手段中，本节采用最常见的一种，即标准化 STD。模型采用四类编码器模型 BERT，RoBERTa，BART 和 T5 相应的 Base 和 Large 版本。对于解码器模型而言，则采用 LLaMA2 和

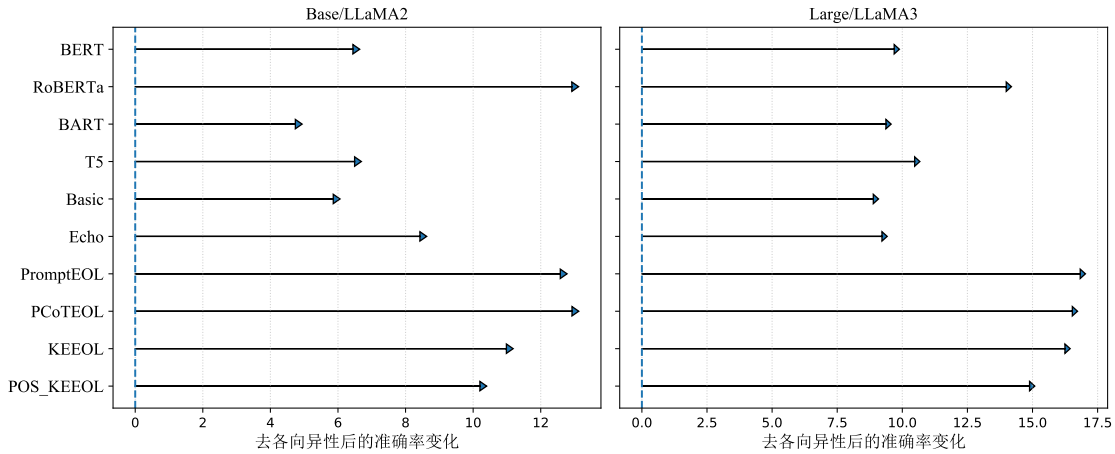


图 5.10　去除各向异性手段对于各类模型的结果提升

更大参数的版本 LLaMA3 在基础设定 Basic，重复上下文的设定 Echo，和采用了四类不同的提示语诱导的 PromptEOL，PCoTEOL，KEEOL 和 POS_KEEOL 的设定。分别对各自的方法进行或者不进行去除各向异性的操作，通过观察结果的差异来判断去除各向异性手段的影响。

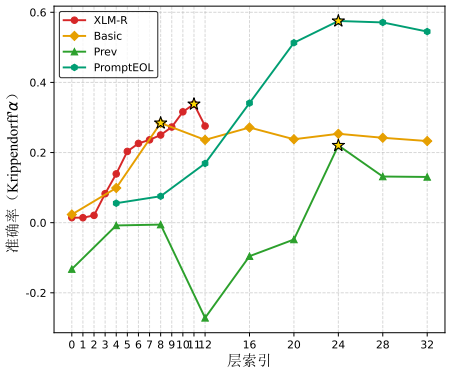
图 5.10 展示了最终的结果，其中横轴表示去除各向异性后准确率的变化，即使用了去除各向异性手段后模型的准确率减去没有使用后的准确率结果，这里采用各层之间相应的准确率之差，再在层数上求平均。纵轴分别列出了各个模型名称，其中包括 Basic 在内之下的设定都是针对解码器模型 LLaMA2（左图）或者 LLaMA3（右图）而言的。

从图中可以清晰看出所有的模型在使用了去各向异性手段后性能都起到了明显提升，尤其是在使用提示语诱导的情况下，例如 PromptEOL 在 LLaMA2 模型上增长了约 13 个百分点，而在 LLaMA3 模型上则增长了超出 16 个百分点。编码器模型除了 RoBERTa 外，其余的模型增长相对较少，但仍然增长了至少近 5 个百分点（BART-Base）。这些结果都有力地表明了去除各向异性手段在计算词义相关性问题中的有效性，表明单纯从模型得到的向量仍在结构上具有一定的局限，还需对其做必要的后处理操作。这与第 4.4.2 节中的多语言数据集 OGWiC 上的结论类似。

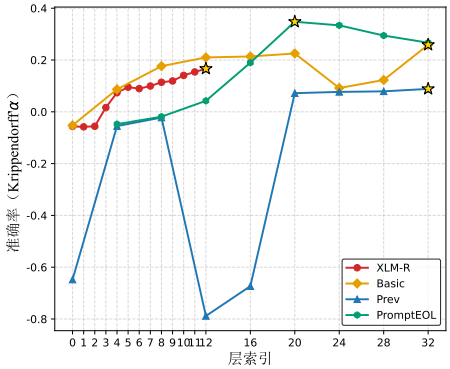
5.5　多语相关度数据集 OGWiC 上的实验结果和分析

第 4 章中分析了各类模型在 OGWiC 数据上的最优结果，本章则从跨层的角度重点分析编码器模型和解码器模型跨层的模式对比。本节同样基于图 5.3 中展示的假说，并从多语言的角度对上一节的结论进行补充。

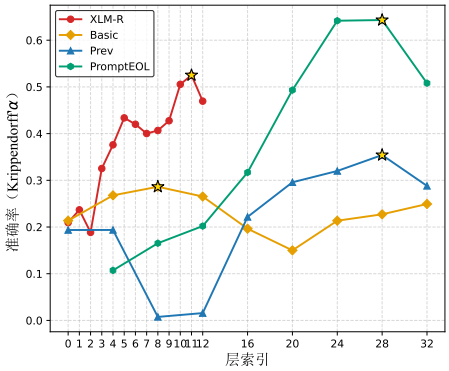
图 5.11展示了各类不同的模型在多语言词义相关度数据集 OGWiC 上的准确



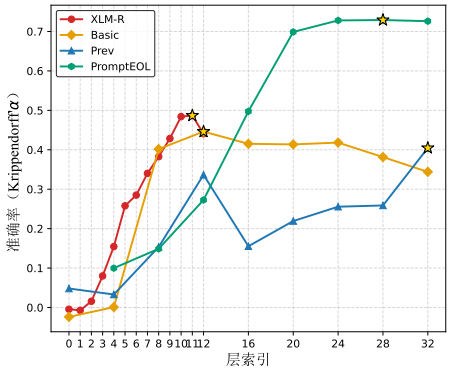
(a) 各语言平均



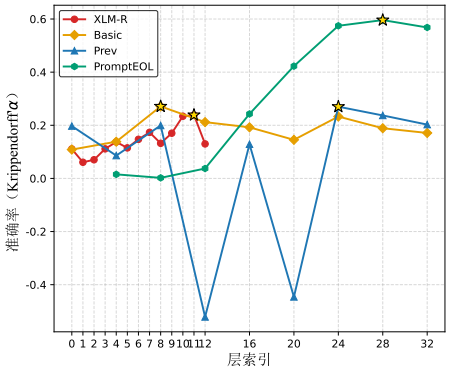
(b) 汉语



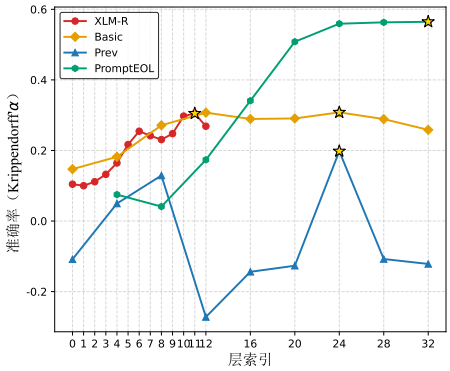
(c) 英语



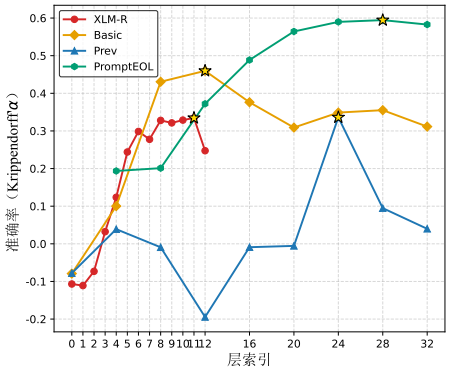
(d) 德语



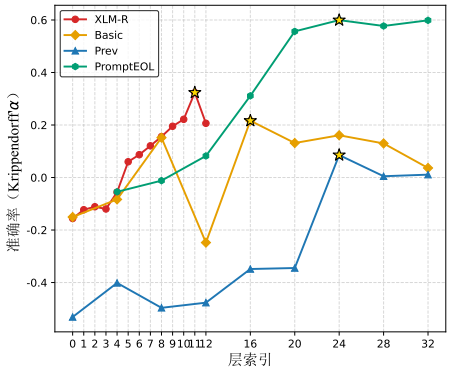
(e) 挪威语



(f) 俄语



(g) 西班牙语



(h) 瑞典语

图 5.11　OGWiC 数据集在不同语言与不同架构模型下的逐层准确率变化

率随层数的变化。其中选择 XLM-RoBERTa（XLM-R）作为编码器模型的代表（使用红色的圆点来表示），LLaMA2 作为解码器模型的代表，这两个模型都在表 4.4 所反映的准确性上取得了各自模型架构中最好的结果。对于解码器模型，本节采用了三种设定：仅提取当前词表征的 Basic 设定（使用橙色的菱形来表示），提取前一个词[①]表征的 Prev 设定（使用绿色的上三角形状来表示），和利用提示语诱导模型生成表征的 PromptEOL 设定（使用绿色的六边形来表示）。准确率选择 OGWiC 的标准设定，即克里彭多夫 α 系数。同时，我们输出了各个语言以及平均准确率的随层数变化情况。涉及彩色的部分请参考电子版。

为了验证图 5.3 中关于编码器模型和解码器模型在理解能力和预测能力的假说，实验结果期待编码器模型随层变化呈现单调递增的趋势，且在高层达到性能最优，表明编码器模型的理解能力随层递增。对于解码器模型而言，对于无任何诱导的 Basic 设定来说，模型的性能呈现“倒 U 型”，模型在中低层达到最优，表明解码器模型在低层实现理解功能。而对于前一个词的表征（Prev）而言和提示语诱导（PromptEOL）的方式，期待模型表现为和编码器模型一样的变化：单调递增且在高层达到最优。这表明模型的预测能力逐层增加。同时该结论不应该受到不同语言的影响，即具有语言普适性。

图 5.11中的结论验证了这一假说。除了汉语在 Basic 设定中并未呈现期待的结果：Basic 设定下，汉语在高层达到最优，而期待它在低层达到。其他所有的语言以及平均状态都符合这一变化趋势。汉语的情况可能是由于本身 Basic 设定在汉语的准确率就很低，最高也仅超出 0.2，因此仅使用目标词的前文对于解码器模型来说，即使在中低层其学习到的表征是不充分的，无法达到一个满意的结果。因此不符合预期的趋势也是可以理解的。以下针对各个语言进行分析。

汉语：如图 5.11（b）所示，在 XLM-R 编码器模型下，准确率随层数单调上升，高层达到最优。解码器模型的 Basic 设定呈现略微递增趋势，高层性能略优于中低层；Prev 和 PromptEOL 设定则呈现分别从中层开始和从低层开始的单调递增趋势，高层性能明显优于中低层，符合预测能力逐层增强的假说。

英语：如图 5.11（c）所示，编码器模型表现为单调递增趋势，高层性能最佳。解码器模型中，Basic 设定呈现倒 U 型，中低层达到最高性能；Prev 和 PromptEOL 设定则呈现分别从中层开始和从低层开始的单调递增趋势，高层性能优于低层，预测能力随层增加而增强。

德语：如图 5.11（d）所示，编码器模型表现稳定的递增趋势，高层性能优于低层。解码器模型的 Basic 设定表现为倒 U 型，中低层达到最优；Prev 和 PromptEOL

① 与前文实验一致，先在原始上下文上做重复操作，再提取第一个上下文中目标词的表征

设定则呈现分别从中层开始和从低层开始的单调递增趋势，高层性能显著提高。

挪威语：如图 5.11（e）所示，编码器模型呈递增趋势，高层达到最优。解码器模型的 Basic 设定略呈倒 U 型，中低层性能较高；PromptEOL 设定单调上升，高层性能最佳。

俄语：如图 5.11（f）所示，编码器模型单调递增，高层最优。解码器模型的 Basic 设定中高层性能略优于低层，呈现轻微倒 U 型；PromptEOL 设定均呈递增趋势，高层性能最高。

西班牙语：如图 5.11（g）所示，编码器模型表现为单调递增，高层性能最佳。解码器模型的 Basic 设定呈倒 U 型，中低层达到最高；Prev 和 PromptEOL 设定则呈现分别从中层开始和从低层开始的单调递增趋势，高层性能最优。

瑞典语：如图 5.11（h）所示，编码器模型递增，高层性能最优。解码器模型的 Basic 设定呈倒 U 型，中低层达到最高；Prev 和 PromptEOL 设定则呈现分别从中层开始和从低层开始的单调递增趋势，高层性能明显优于低层。

综合各语言的结果，可以观察到编码器模型在所有语言下均呈现单调递增趋势，高层性能最优；解码器模型在 Basic 设定下大多数语言呈现倒 U 型，中低层达到最高性能；PromptEOL 设定则普遍呈单调递增趋势，高层性能优于低层；虽然 Prev 上的结果不太稳定，但是在多种语言上模型从中层开始呈现性能增长的趋势。这些分析验证了编码器模型理解能力逐层增强、解码器模型预测能力在高层增强的假说，同时说明该结论具有语言普适性。汉语在 Basic 设定下的特殊表现，可归因于其整体准确率偏低，在中低层提取的目标词表征不足以支持最佳预测。

5.6　本章小结

本章基于 OGWiC 与 WiC 两个数据集，对各类编码器模型和解码大语言模型在不同层级中的词汇语义表征特性进行了系统分析。实验结果一致表明，对于解码器模型而言，词汇语义信息主要集中于模型的中低层，而高层表示则逐渐转向对后续输出的预测与生成；而对于编码器模型来说，词义信息则集中在模型的高层。这一层级分布特征揭示了大语言模型中信息自低向高流动的基本机制：模型首先在较低层完成对词汇语义的理解建模，随后在高层更多服务于预测目标。

上述发现为工程应用中与词汇语义相关的任务提供了明确的实践指导。例如，在词性标注、词义消歧等以词汇语义区分为核心的任务中，优先选取中低层表示有助于获得更稳定且具判别力的特征；而在文本摘要、对话生成等以预测或生成结果为导向的任务中，高层表示则更契合模型的训练目标与推断需求。由此可见，不同层级表示的合理选用是充分发挥大语言模型能力的重要前提。

此外，本章结果还从“层级信息流动”的角度为大语言模型的可解释性研究提供了新的分析视角：通过区分理解阶段与预测阶段在模型内部的分布位置，可以更细致地刻画不同层级所承担的语义功能。这不仅有助于深化对大语言模型内部机制的认识，也为后续开展基于层级表示的语义探测与模型分析工作奠定了基础。

第 6 章　多语言词义相关度任务中的标注不一致评估

6.1　本章导论

在词义相关度判断任务中，标注不一致现象普遍存在，即不同标注者对同一目标词在不同上下文中的语义相关度可能给出不同的判断。然而，现有研究往往忽视这种不一致现象，例如将其视为噪声而丢弃，或仅利用平均统计量来表示标注结果，因而无法客观反映真实世界中存在的标注差异。本章重点研究如何通过大语言模型在内的神经网络语言模型预测词义相关度中的不一致程度。词义相关度标注不一致程度预测任务是第 4 章关于词义相关度判断任务的延续和拓展：相关度任务利用标注的平均统计量作为预测目标；而不一致程度预测则利用多个标注的离散程度作为输出目标。

本章提出三类方法进行离散程度的预测，分别是分类器探测方法、模型集成方法和基于提示语的大语言模型探测方法。其中，在模型集成方法中，本章通过设计不同的模型扰动方式、集成策略和集成统计量，将扰动后的模型视为不同的标注者，以模拟语义标注的不一致性。另一方面，本章还通过设计提示语，利用大语言模型直接评估不一致程度。在评估数据集方面，除公开数据集外，还包括本文收集的汉语兼类词数据集。评估模型涵盖前述编码类模型和解码类模型，分布广泛。

本章的结构安排如下：第 6.2 节详细介绍标注不一致程度预测的几种方法；第 6.3 节阐述标注不一致预测任务的实验设置，包括数据集、评估模型、集成方式、超参数设定以及基线模型；第 6.4 节展示实验结果与分析，包括集成方法各重要模块的消融实验以及典型偏误案例分析；第 6.5 节为本章小结。

6.2　词义相关度标注不一致预测方法

针对词义相关度标注中的不一致现象，本章提出三类预测方法：第一类是基于分类器的探测方法，通过训练分类器直接预测标注离散程度；第二类是基于模型集成的探测方法，通过设计多样化的模型扰动方式、集成策略和统计量，模拟不同标注者的判断差异；第三类是基于提示语的大语言模型探测方法，利用大语言模型的语义理解能力，通过精心设计的提示语直接评估标注不一致程度。

6.2.1　基于分类器的探测方法

基于分类器的探测方法（classifier-based probing）是知识探测领域的重要方法之一[①]。该方法的核心思想是：在语言模型的中间层表征之上，附加一个轻量级的探测分类器，通过考察该分类器在特定语言学任务上的表现，来推断模型中间表示中是否编码了相应的语言知识。具体到本章任务，我们利用探测分类器来评估模型表示对标注离散程度的预测能力。

针对本章的词义相关度标注不一致程度预测任务，探测问题可形式化为一个回归任务，即训练探测分类器输出一个 $[0, 1]$ 区间内的连续值，用以表征模型对目标词语义标注不一致程度的预测。探测分类器的输入表示提取方式与模型架构密切相关，第 4.2.1 节已对此进行详细阐述。简言之，对于编码器模型，我们提取目标词的高层表示；对于解码器模型，则重点关注前一词或经提示语诱导后最后一个字符的表示。

在本章实验中，由于需要度量同一目标词 w_i 在两个上下文 c_1 和 c_2 中对应表征 h_1 和 h_2 之间的相关度，可以采用不同的策略将两个表征进行融合。表 6.1总结了本文采用的几种融合方式。具体而言，我们将融合后的表示向量 $h = \mathrm{agg}(h_1, h_2)$ 输入轻量级探测分类器 f_{cls} 中。分类器首先在训练集上学习最优参数，然后在测试集上进行预测，输出结果反映了原始模型对标注不一致程度的编码能力。

表 6.1　向量融合的 4 种方式

集成方式	标记	计算方式	说明
求和池化	sum	$h = h_1 + h_2$	对应元素逐位相加
相乘池化	mul	$h = h_1 \odot h_2$	对应元素逐位相乘
平均池化	avg	$h = \frac{h_1+h_2}{2}$	两向量求平均
堆叠池化	app	$h = [h_1, h_2]$	两个向量拼在一起

图 6.1 展示了基于分类器的探测方法的完整流程。其中，被探测模型（图中橙色矩形所示）的参数保持冻结状态，仅用于提取文本表示；探测分类器（图中蓝色模块所示）由一个轻量级神经网络构成，其参数需在训练集上进行学习，监督信号为训练集中多位标注者的标注离散程度。图中蓝色箭头表示训练阶段的信息流动方向。训练完成后，探测分类器在测试集上执行预测任务，其输出结果用于评估被探测模型对标注不一致程度的编码能力。值得注意的是，被探测模型的参数始终处于冻结状态，无需进行更新。

① 需要说明的是，本文采用“分类器”这一统称，但探测任务并不限于离散标签预测。例如，本章关注的标注不一致程度预测本质上是一个连续值回归问题。

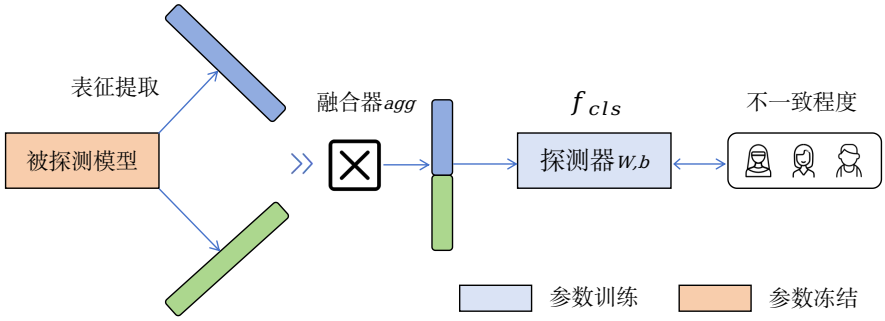


图 6.1　分类器探测示意图

第 4.2.2 节在评估词义相关度任务时采用了更为简洁的几何探测方式，即通过表征的线性操作（如内积）直接度量表示之间的相关程度，无需训练额外参数。然而，本章的标注不一致程度预测任务未采用几何探测方法，原因在于几何探测难以针对特定任务探究细粒度的语言知识。具体而言，几何探测主要适用于表征相似性或相关度度量，无法通过向量的简单运算直接反映标注离散程度这类复杂语义信息。

基于分类器的探测方法具有直观、简洁的优势，且在认知心理学和行为心理学等领域已有类似应用[216]，属于通过观察外在行为进行评估的行为主义范式。然而，该方法也存在若干局限：首先，引入额外参数导致难以区分探测到的语义知识是源自原始模型还是探测分类器本身，这也是探测分类器需尽可能保持轻量级的原因；其次，探测分类器的训练过程带来了额外的计算开销，在一定程度上降低了评估效率。

6.2.2　基于模型集成的预测方法

模型集成方法从另一视角模拟标注者之间的不一致性。该方法将不同的模型实例类比为独立的“标注者”，通过收集各模型对同一任务的输出结果，计算其离散程度，从而预测真实标注中的不一致现象。其核心假设在于：模型之间的输出差异可以在一定程度上反映人类标注者的判断分歧。

6.2.2.1　模型扰动策略

不同模型实例之间的差异在本文中统称为模型的扰动（perturbation）。需要说明的是，本文所指的“扰动”并非数学意义上的微小连续变化，而是通过改变模型配置或训练机制所形成的模型变体。具体而言，扰动可通过调整模型的内部参数设置或修改模型的结构配置来实现。

基于扰动方式的不同，本文设计了三种模型集成策略：

（1）同质扰动（Homogeneous Perturbation）

在保持模型整体结构和预训练配置一致的前提下，通过改变模型内部参数或训练设置来产生不同实例。具体扰动方式包括：调整层数设定、改变去各向异性方法、使用不同的随机种子等[①]。

（2）异质扰动（Heterogeneous Perturbation）

采用结构配置不同的模型作为扰动来源。这些模型在层数、隐藏维度、注意力头数或模块组成等方面存在实质性差异，从而产生更具多样性的模型实例。

（3）混合扰动（Hybrid Perturbation）

同时引入结构配置和参数设置两个层面的差异，即在异质结构的基础上进一步调整参数或训练策略，以获得最大程度的模型多样性。

通过上述三类扰动策略，本文构建了具有不同表征特性的模型集合，用以模拟真实标注场景中多位标注者的判断差异。各模型实例对同一任务的输出结果将被收集并汇总，用于计算标注不一致程度的预测值。

需要说明的是，本文所称的模型“结构”并非仅指宏观架构类型（如编码器架构或解码器架构），而是指构成模型的完整结构配置集合，包括层数、隐藏维度、注意力头数、前馈网络维度以及模块组织方式等。只有当上述配置完全一致时，两个模型才被视为具有相同结构。

以 BERT 和 RoBERTa 为例，二者虽均采用 Transformer 编码器架构，但由于预训练目标、数据规模及训练策略存在差异，其参数空间与表征分布亦呈现显著不同。因此，在本文的扰动分类框架下，BERT 与 RoBERTa 可视为异质扰动的典型实例。

6.2.2.2　集成策略与统计量

假定从不同模型实例中提取的目标词表示对集合为 $\{(h_i, h_j)\}$，其中 h_i 和 h_j 分别表示目标词 w 在上下文 c_i 和 c_j 中的向量表示。基于此，本文采用连续值集成和离散值集成两种策略来预测标注不一致程度。表 6.2 列出了不同的策略下对应的统计量，作用对象，计算公式和简要描述。其中公式中出现的符号表示如下：N 代表数据中的实例个数，x 代表出现的实例，如果表示成对的实例，则采用下标为 a 和 b 的索引，如果表示单一的实例，则使用下标 k，$\bar{x}$ 代表实例的平均值，f_{mode} 代表众数，上标 m 和 rel 分别代表数值连续的相似度得分和数值离散的相关度等级。“单调性”一列表明统计量取值与不一致程度的关系：向上箭头（↑）表示统计量越大，标注不一致程度越高。

① 同质扰动虽保持模型结构配置不变，但由于参数初始化或训练策略的差异，其生成的表示空间可能存在显著差别，这种差别正是模拟不同标注者判断差异的基础。

表 6.2　评估模型输出分歧的统计量

统计量	对象	单调性	计算公式	描述
ASD	m_{ij}	↑	$\sqrt{\frac{1}{N}\sum_{k}(x_k-\bar{x})^2}$	模型相似性评分的标准差
MPD_c	m_{ij}	↑	$\frac{2}{N(N-1)}\sum_{a<b}\lvert x_a^m - x_b^m\rvert$	模型相似性评分的平均成对差异
MPD_d	rel_{ij}	↑	$\frac{2}{N(N-1)}\sum_{a<b}\lvert x_a^{\mathrm{rel}} - y_b^{\mathrm{rel}}\rvert$	离散标签的平均成对差异
VR	rel_{ij}	↑	$1-\frac{f_{\mathrm{mode}}}{N}$	异众比率（非众数标签占比）

（1）连续值集成策略

连续值集成首先采用余弦相似度计算每对表示 (h_i, h_j) 的相似性得分 m_{ij}，具体计算公式参见第 4 章公式 (4.1)。对于由 N 个模型实例产生的相似性得分集合 $\{m_{ij}\}$，本文采用两种样本统计量估计标注不一致程度：

（a）标注标准差（ASD, Annotation Standard Deviation）

标注标准差是衡量连续变量离散程度的经典统计量，反映各样本值偏离均值的平均幅度。

（b）平均成对绝对差异（MPD_c, Mean Pairwise Difference for continuous values）

计算所有样本对之间绝对差异的平均值，反映样本两两之间的平均分歧程度。鉴于本章所用数据集的标注不一致程度正是基于类似方式计算的，MPD_c 更贴合任务定义。

（2）离散值集成策略

离散值集成采用第 4 章介绍的几何探测方法，直接预测目标词在两个上下文中的语义相关度离散标签 $\mathrm{rel}_{ij} \in \{1,2,3,4\}$，具体计算过程参见公式 (4.2)。由 N 个模型实例产生的相关度标签构成集合 $\{\mathrm{rel}_{ij}\}$，本文在此基础上采用两种统计量度量不一致程度：

（a）平均成对绝对差异（MPD_d, Mean Pairwise Difference for discrete labels）

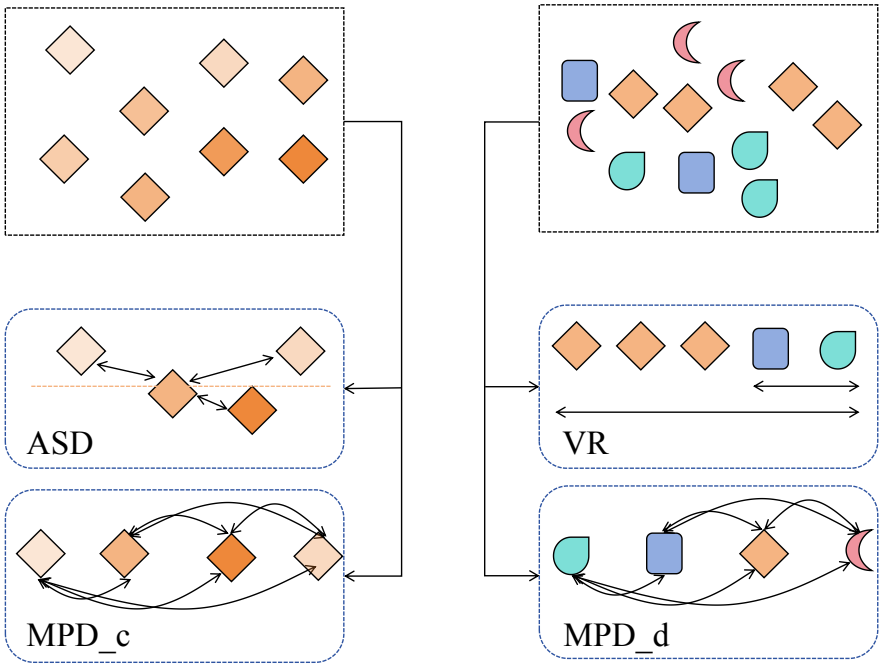
计算所有标签对之间绝对差异的平均值，反映模型之间对词义相关度判断的平均分歧程度。该统计量与本章数据集计算标注不一致程度的标准方式一致。

（b）异众比率（VR, Variation Ratio）

异众比率衡量非众数标签所占比例，即 $1-\frac{f_{\mathrm{mode}}}{N}$，其中 f_{mode} 为众数频次。该统计量直观反映了模型判断的集中程度：异众比率越高，说明模型间分歧越大。

（3）两类策略对比

图 6.2 直观展示了两种集成策略及其对应统计量的区别与联系。图中每个形状



(a) 连续值融合　　(b) 离散值融合

图 6.2　两种方式的模型集成方式对比图

代表一个模型在相关度任务上的输出结果：相同形状且渐变色块表示连续值，反映同一目标词在不同上下文中语义相关度的连续变化；不同形状的色块表示离散值，对应四种不同的相关度等级。在统计量方面，除异众比率 VR 外，各统计量（ASD、MPD_c、MPD_d）均通过计算模型输出结果之间的差异来度量不一致程度，图中以箭头示意；而异众比率 VR 则通过计算非众数样本所占比例来反映模型间的分歧程度。从图中可以看出，连续值集成与离散值集成分别从连续空间和离散类别两个视角刻画模型间的判断分歧，二者互为补充，共同构成了本文预测标注不一致程度的方法体系。

6.2.3　基于提示语的探测方法

为充分利用大语言模型的强大生成能力，本文提出基于提示语的探测方法，通过设计针对性的提示语，直接引导大语言模型输出标注不一致程度的预测结果。与第 4.2.4 节中相关度预测任务的思路一脉相承，本方法将不一致程度预测问题转化为一个受控的文本生成过程。二者的主要区别在于输出形式：相关度预测输出离散标签（1-4 级），而不一致程度预测则需输出一个 [0, 1] 区间内的连续值。

相较于前文介绍的分类器探测和模型集成方法，基于提示语的探测具有以下特点：

（1）**无需访问模型内部状态**：本方法仅需通过模型的文本输入输出接口进行

交互，无需获取隐藏层表示，适用于各类闭源或仅提供 API 访问的大语言模型。

（2）**任务转换的间接性**：不一致程度预测相较于相关度预测更具间接性——在实际标注过程中，标注者通常不会被直接询问“这个词可能引起多大分歧”，而是通过汇总多位标注者的实际判断来获得不一致程度。然而，不一致程度与词义本身密切相关：语义模糊、难以捕捉或高度不确定的样本往往更容易引发标注分歧。基于这一认知，大语言模型理论上可以通过理解词义的歧义性来间接预测标注不一致程度。

下文列出了本文在大语言模型实验中使用的提示语模板。第一类模板旨在引导模型直接输出标注不一致程度的量化结果。具体而言，我们要求模型输出一个 [0, 1] 区间内的连续数值，其中 0 表示标注者之间一致性较高（分歧较小），1 表示标注者之间一致性较低（分歧较大）。需要指出的是，相较于直接判断词义相关度，询问“标注者可能产生多大分歧”是一种更为间接的提问方式。然而，基于“语义模糊或高度不确定的样本更容易引发标注分歧”这一认知，模型仍需从词义本身出发，考量目标词的歧义程度，从而间接完成预测任务。除不一致程度数值外，我们还要求模型提供其判断依据的自然语言解释，以便分析模型的决策过程。提示语的英文原文及其中文翻译如下：

You are given two sentences that contain the same target word (marked in bold). Annotators are asked to judge how semantically related the meanings of the target word are in the two contexts. However, annotators may disagree to different degrees. Your task is to predict the extent of annotator disagreement and express it as a continuous value between 0 and 1:

- 0 = minimal disagreement (annotators likely agree)

- 1 = maximal disagreement (annotators likely strongly disagree)

Target word: `WORD`

Sentence 1: `SENTENCE1`

Sentence 2: `SENTENCE2`

Rate the expected annotator disagreement on semantic relatedness.

IMPORTANT: (1) Output exactly one valid JSON object. (2) Include exactly two keys: “result” and “explanation”. (3) Do not include any additional text.

Do NOT include any other text, reasoning, or commentary.

The JSON should look like:

{

“rating”: 0.0,

“explanation”：“brief reason here”

}

以下为上述提示语的中文翻译版本：

你将收到两个句子，它们包含相同的目标词（以粗体标记）。标注者需判断目标词在这两个语境中语义上的相关程度。然而，标注者之间可能存在不同程度的意见分歧。你的任务是预测标注者分歧的程度，并将其表示为一个介于 0 到 1 之间的连续值：

- 0 = 分歧极小（标注者很可能意见一致）

- 1 = 分歧最大（标注者很可能存在强烈分歧）

目标词：目标词

句子 *1*：第一个上下文

句子 *2*：第二个上下文

请评估在语义相关性上，预期标注者之间可能出现的分歧程度：

模型需严格输出一个合法的 *JSON* 对象，仅包含“评分结果”和“解释”两个字段。

{

“评分结果”：0.0,

“解释”：“简要理由”

}

第二类模板在基础模板的基础上引入词类信息，旨在探究词类知识对不一致程度预测的影响。提示语中以下划线突出显示词类相关表述，以明确告知模型目标词的词性类别。提示语内容如下：

You are given two sentences that contain the same target word (marked in bold). Annotators are asked to judge how semantically related the meanings of the target word are in the two contexts. However, annotators may disagree to different degrees. Your task is to predict the extent of annotator disagreement, explicitly taking into account the part of speech of the target word in each sentence and express it as a continuous value between 0 and 1:

- 0 = minimal disagreement (annotators likely agree)

- 1 = maximal disagreement (annotators likely strongly disagree)

在后面的数据格式中增加了词类信息，其余信息不变：

Sentence 1: SENTENCE1　Part of speech in Sentence 1: POS1

6.3　词义标注不一致任务实验设定

本节的结构安排如下：首先介绍词义相关度标注不一致程度任务及所用数据集；其次阐述用于评估和预测不一致程度的语言模型；随后说明模型集成方法、分类器探测方法及其超参数设置；最后介绍对比实验中采用的基线模型。

6.3.1　任务定义和数据集

本章任务旨在预测词义相关度标注中多位标注者之间的判断不一致程度。实验采用两类数据集：公开多语言数据集 DisWiC（Disagreement in Word-in-Context）和本文自建的汉语兼类词数据集 DisWiC_NV。二者已分别在第 3.2.3 节和第 3.2.5 节中进行了详细介绍，涵盖数据收集流程、数据格式及数据量统计等内容。

对于每条样本，模型需预测其标注不一致程度，并将预测值与真实不一致程度进行比较。评估指标采用斯皮尔曼秩相关系数 Spearman's ρ，计算过程见公式 (3.8)。

6.3.2　评估模型

本节介绍用于评估不一致程度的语言模型，涵盖编码器模型和解码器模型两大类。其中编码器模型选取 XLM-Bert 和 XLM-Roberta（简称为 XLM-B 和 XLM-R）为代表。解码器模型包括开源大语言模型和闭源大语言模型，前者包括 LLaMA2 和 LLaMA3，后者则包括 DeepSeek 和通义千问模型。相关模型的详细架构及预训练方式参见第 4.3.2 节。

6.3.3　集成方法和超参数设置

6.3.3.1　扰动变量

集成方法通过扰动模型内部参数或配置来模拟不同标注者的判断差异。针对不同类型的模型，本文采用差异化的扰动策略。表 6.3 汇总了各类模型及其对应的扰动变量。其中针对闭源解码器模型，则采用提示语获取不一致程度的提示语设计。

表 6.3　模型类型及对应的扰动变量

模型类型	模型	扰动变量/获取方式
编码器模型	XLM-B、XLM-R	层数、Dropout 概率、去各向异性手段
开源解码器模型	LLaMA2、LLaMA3	层数、提示语、输入格式、去各向异性手段
闭源解码器模型	DeepSeek、通义千问	提示语

以下对各扰动变量进行详细说明。

（1）层数

不同深度的网络层编码不同粒度的语言信息：对于编码器模型，较低层更偏向词法与局部句法特征，中高层则更倾向于抽象语义信息[159]；对于解码器模型，低层侧重当前词义理解，高层则体现预测能力。第 5 章已对跨层信息流动模式进行了系统分析。基于此，本文通过选择不同层级的表示来引入扰动。

（a）编码器模型：XLM-B 和 XLM-R 均包含 12 层，本文选取第 1、4、7、11 层作为提取层①。

（b）解码器模型：LLaMA2 和 LLaMA3 均包含 32 层，本文选取第 8、16、24、31 层作为提取层。

（2）Dropout 概率

Dropout 是一种在训练过程中随机失活神经元的正则化技术[217]，在推理阶段启用时可视为对模型子空间的近似采样，引入轻微的随机扰动。本文将其作为构造“表征扰动”的手段：在保持输入不变的前提下，启用 Dropout 并进行多次前向传播，获得多个不同的表示实例。本文统一设置采样次数为 3，即对同一输入独立采样三次，并对所得表示进行平均或拼接操作。该扰动方式仅应用于编码器模型。

（3）向量后处理方法

不同的向量后处理方式可视为对表示空间施加不同的几何变换，从而引入扰动。本文采用第 4.2.3 节介绍的三种后处理方法：

（a）标准化（Standardization）：对向量进行零均值单位方差变换；

（b）中心化（Centering）：减去整体均值向量；

（c）去主成分（PCA removal）：移除前若干个主成分方向，削弱主导语义轴的影响。

上述方法分别从缩放、平移及主方向压缩等角度调整表示空间几何形态。

（4）提示语模板与输入格式

对于大语言模型（包括开源与闭源模型），提示语设计直接影响模型输出及其内部表征结构。本文采用四类提示策略：

（a）基础提示语（PromptEOL）：基础模板；

（b）知识增强提示语（KEEOL）：融入语言学知识；

（c）多语言提示语（LAN_KEEOL）：使用不同语言的提示模板；

（d）思维链提示语（PCoTEOL）：引导模型进行逐步推理。

此外，在不使用提示语的情况下，输入格式的差异也会影响模型输出。本文考虑两种输入格式：

① 选取倒数第二层而非最后一层，是因为第 5.4 节及第 5.5 节实验表明，在词义相关度判断任务中，倒数第二层表现优于最后一层，说明其表征对词义更具判别力。

（a）原始输入：直接使用目标词及其上下文；

（b）重复输入：对输入文本进行重复后输入。

根据第 5 章的结论，此类情况下低层表征对词义表达更优，因此仅提取低层表示。上述扰动方式仅应用于解码器模型。

6.3.3.2　扰动变量组合及超参数设置

表 6.4 展示了编码器模型与解码器模型在不同扰动因素下的组合方式。其中：

（a）层数：以数字表示选取的层号（1/4/7/11 或 8/16/24/31）；

（b）向量后处理：“中”表示中心化，“主”表示去主成分，“标”表示标准化；

（c）Dropout 采样：数字表示启用 Dropout 后的采样次数（1/2/3）；

（d）提示语/输入：“基”表示基础提示语（PromptEOL），“知”表示 KEEOL，“语”表示 LAN_KEEOL，“思”表示 PCoTEOL，“原”表示无提示语（仅提取目标词表征），“重”表示重复输入。

表 6.4　模型扰动因素组合及扰动次数

模型类型	层数	向量后处理	Dropout 采样	提示语/输入	扰动次数
编码器模型（10）	11	无/中/主	无	-	3
	11	标	1/2/3	-	3
	1/4/7/11	标	无	-	4
解码器模型（12）	31	无/中/主	-	基	3
	31	标	-	基/知/语/思	4
	8/16/24/31	标	-	基	4
	8	标	-	原/重	2

（1）扰动组合选择原则

扰动变量的选择遵循以下原则：在考察某一扰动变量时，其他变量保持其在相关度预测任务中的最优配置。例如，在变化向量后处理方式时，层数固定为最高层（第 11 层或第 31 层），提示语固定为基础模板。这一设计旨在保证采样效率的同时，增强实验结果的可信度。

（2）扰动模型数量与集成实例

基于上述配置，编码器模型共生成 10 种扰动变体，解码器模型共生成 12 种扰动变体。考虑到标注不一致程度预测任务中，每个样本通常对应有限数量的标注者（本章实验设定为 3），我们从各模型变体中随机选择 3 个构成一个“集成实例”。理论上，可能的集成实例数量为组合数：

（a）编码器模型：$\binom{10}{3} = \frac{10!}{3!(10-3)!} = 120$ 种组合

（b）解码器模型：$\binom{12}{3} = \frac{12!}{3!(12-3)!} = 220$ 种组合

其中，$\binom{n}{k}$ 表示从 n 个不同扰动模型中无放回地选择 k 个模型的组合数。每个由 3 个模型构成的组合即为一个“集成实例”，用于模拟三位标注者的判断差异。

6.3.4　分类器探测方法的训练设置

本文在 DisWiC 训练集上训练一个双层回归模型，用于预测标注不一致程度。该模型的训练流程如下：

（1）特征提取与融合

首先，使用 XLM-Roberta-base 模型提取目标词在两个上下文中的向量表示。随后，对这两个表示进行融合，融合后的向量作为回归探测器的输入。表 6.1 列出了四种不同的融合方法。经实验评估，本文最终选择平均池化作为最优融合方式。

（2）网络结构

回归探测模型采用双层神经网络架构：

（a）第一层：线性层，将融合后的表示映射到中间隐状态；

（b）激活函数：修正线性单元（Rectified Linear Unit，ReLU）[218]，引入非线性变换；

（c）第二层：线性层，将中间隐状态映射到最终输出，输出为一个 $[0, 1]$ 区间内的连续值，用以表征标注不一致程度。

（3）损失函数与优化设置

模型训练采用均方误差损失（Mean Square Error，MSE）作为优化目标。训练参数设置如下：

（a）训练轮数：200 轮；

（b）批大小：32；

（c）学习率：0.01；

（d）优化器：AdamW[219]；

（e）Dropout 概率：0.1。

（4）学习率预热

为提升训练初期的稳定性，本文设置 0.1 的预热（warm-up）比例，即学习率在前 10% 的训练步数中从 0 线性增长至预设值 0.01，之后保持不变。

6.3.5　提示语探测的超参数设置

本文采用两种提示语来测试闭源大语言模型，分别是不带词类信息和携带词类信息的，提示语的具体内容可以参考第 6.2.3 节。本文采用的两类闭源大语言模型分别为 DeepSeek 和通义千问模型，使用应用程序接口 API（Application Programming Interface）进行访问。与前文（例如第 4.3.2.2 节）设定一致，两类模型均选择官方的旗舰模型 DeepSeek-Chat 和 Qwen3-Next-80B-A3B-Instruct，其余均保持默认设置。

6.3.6　基线模型

为验证所提出方法的有效性，本文选取了词义相关度不一致预测任务中具有代表性的多种基线模型进行对比，涵盖基于概率分布建模、多标签训练及特征回归等不同技术路线。

6.3.6.1　分布预测法

分布预测法是由 Mikhail Kuklin 和 Nikolay Arefyef 在 CoMeDi 工作坊[37]中提出的方法[148]。该方法的核心思想是训练一个分类器，输出样本在所有可能相关度等级（即 1、2、3、4）上的概率分布，进而通过计算该分布的标准差来估计标注不一致程度。

该方法的主要特点如下：

（1）训练样本构建：区别于传统方法将多位标注者的中位数作为训练标签，本方法将每位标注者的独立判断视为一个单独的训练样本，从而更好地建模不同标注者之间的判断差异。

（2）概率分布输出：分类器输出一个四分类概率分布 $P(r|X)$，其中 $r \in \{1,2,3,4\}$ 表示相关度等级。

（3）不一致程度估计：基于输出的概率分布，计算其标准差作为标注不一致程度的预测值。

（4）温度校准：采用温度缩放（temperature scaling）技术对输出的概率分布进行校准，使其更符合真实的标注分布。

该方法的基座模型采用 XLM-Roberta，并在西班牙语数据上进行了预训练微调。

6.3.6.2　多标签训练法

多标签训练法是由 David Alfter 等在 CoMeDi（Context and Meaning-Navigating Disagreements in NLP Annotations）工作坊中提出的另一种基线方法[149]。该方法

将多位标注者的判断视为独立的监督信号，通过训练多个相关度预测模型来模拟不同标注者的判断过程。

该方法的具体实现如下：

（1）数据划分：将每位标注者标注的相关度标签划分为独立的训练子集，共得到 5 份数据（对应 5 位标注者）。

（2）多模型训练：针对每份子集，分别训练一个独立的词义相关度预测模型。所有模型均以 XLM-Roberta 为基座，用于提取目标词的上下文表示。

（3）预测与集成：训练完成后，对于每条测试样本，5 个模型分别输出一个相关度预测值 $\{\hat{r}_1, \hat{r}_2, \cdots, \hat{r}_5\}$。

（4）不一致程度计算：采用公式 (3.7) 中的平均成对差异（Mean Pairwise Difference, MPD）计算 5 个预测值之间的离散程度，作为标注不一致程度的最终预测。

该方法的核心假设是：多个独立训练的相关度预测模型之间的输出差异，可以在一定程度上模拟真实标注者之间的判断分歧。

6.3.6.3　工作坊基线

工作坊基线方法是由 CoMeDi 工作坊的组织者提出的基准方法[37]，旨在为参赛模型提供一个基础的对比参照。

该方法的实现流程如下：

（1）特征提取：采用 XL-Lexeme 模型[197]为每个目标词生成上下文向量表示。对于给定的目标词及其两个上下文 c_1 和 c_2，分别得到向量 v_1 和 v_2。

（2）特征融合：将两个向量拼接，形成联合表征 $v = [v_1; v_2]$。

（3）回归预测：将拼接后的向量输入多层感知机（MLP）回归模型，学习从表征空间到标注不一致程度的映射关系。MLP 包含一个隐藏层，使用 ReLU 激活函数。

模型在全部训练数据上进行端到端训练，优化目标为均方误差损失。该方法作为一种轻量级的特征回归基线，用于评估更复杂方法相对于简单特征映射的提升效果。

6.4　实验结果和分析

本节对 OGWiC 数据集和 DisWiC_NV 数据集的实验结果进行如下分析，首先总体对比分析了各类方法的标注不一致程度预测的结果，其次重点分析本文提出的模型集成方法中涉及的重要流程，包括不同扰动策略、集成方式及统计量，最后

从定性角度关注词类信息的影响以及不同的偏误类型。

6.4.1　标注不一致相关性结果分析

表 6.5 展现了各类方法在不同的数据集上所得到的不一致评分与专家标签之间的斯皮尔曼相关系数，其中相关性越大说明模型预测词语出现在不同上下文的标注者不一致程度越准确。数据集包括 DisWiC 中全部及各自语言部分的数据集和中文兼类词数据集 OGWiC_NV。方法部分分为上中下三部分，上半部分是前人的工作，分别为分布预测法[148]，多标签训练法[149]和工作坊基线法[37]；中间部分是本文提出的基于词向量表征的两类方法，分别是分类器探测方法和模型集成方法；最下面展示了两个闭源大模型直接使用提示语后的结果，分别是 DeepSeek 语言模型和通义千问模型。其中“w. POS”表示提示语中使用了词类信息，默认是没有词类信息的提示语。

6.4.1.1　OGWiC 数据集

对于 DisWiC 任务，从表 6.5 中可以看出基于模型集成的手段要比分类的方法更加有效，同时也超出了同类的其他模型、官方提供的基线模型以及基于提示语的大语言模型，这表明本章提出方法的有效性。模型集成可以有效利用各类不同的模型及其扰动，可以充分模拟不同可能的标注空间。

与本章提出的模型集成方式类似，前人工作提出的分布预测法和多标签训练法都取得了不错的结果。这些方法的共性都是充分利用了不同标注者标注的信息，并将不同的标注视作独立的结果，因此，它们都有效模拟了真实场景下的标注不一致产生的过程，从而取得占优的结果。

表 6.5　各类方法在不同语言上的不一致程度预测结果

方法	全部	汉语	英语	德语	挪威语	俄语	西班牙语	瑞典语	DisWiC_NV
分布预测法[148]	22.60	30.10	7.80	20.40	28.60	**17.50**	**18.70**	35.00	–
多标签训练法[149]	22.00	**53.90**	4.20	10.80	27.20	16.70	11.50	29.60	–
工作坊基线[37]	16.30	48.50	6.00	8.50	23.50	11.60	7.80	7.90	–
分类器探测	8.20	35.80	3.80	2.20	-4.20	6.70	4.00	9.00	–
模型集成	**26.15**	49.96	10.27	**21.56**	**38.25**	16.53	7.34	**39.17**	**16.81**
DeepSeek	19.94	45.03	6.64	14.66	26.89	9.74	11.85	24.80	3.31
DeepSeek w. POS	-	-	9.68	-	-	-	-	-	9.67
通义千问	22.41	53.63	**14.21**	15.94	24.65	15.75	11.48	21.20	0.95
通义千问 w. POS	-	-	7.76	-	-	-	-	-	1.44

与之相反，分类器探测模型的较低结果表明直接使用不同标注者的不一致程

度作为监督信号训练回归模型的方式效果并不好，这也体现了该任务与直接预测相关度的本质差异：不一致涉及到多个标注者，体现了标注过程的不确定性，无法直接利用词义表征来直接反映和提取不一致程度。这也从另一侧面体现了分类器探测的特点：不是所有的输出都可以被学习到，只有输入特征包含与输出相关的信息时，分类器探测才可能捕捉到。

仅依赖提示语的大语言模型表现亮眼，其中通义千问模型要优于 DeepSeek 模型，并且在英语数据集（OGWiC_EN）上达到最优，汉语也取得第二名的结果。这体现了大语言模型的巨大潜能，即仅通过人类语言交互的方式，无需额外训练就可以达到与其他算法相当的结果。尽管大语言模型是一个以生成任务为训练目标的解码器模型，但是其对词义不一致性的理解在适配了大语言模型的输出之后，仍然表现不俗。

6.4.1.2　DisWiC_NV 数据集

对于中文兼类词数据集 DisWiC_NV 而言，由于它的标注者更多，不一致性更大，其困难程度要比 DisWiC 的汉语数据大得多。这可以从第 3.2 节中，表 3.7 和表 3.10 显示的 OGWiC 数据集和 OGWiC_NV 数据集的“标注数统计”可以看出。在表 6.5 右半部分显示的 DisWiC_NV 数据集的结果中，可以看出所有方法的结果都相较 DisWiC 数据集低很多，最优的模型集成方式相关度仅为弱相关（16.81），比最优的 DisWiC 数据集（26.15）低了近 10 个百分点。

尽管如此，模型集成的方式仍旧在其他方法中是最好的。而基于外部提示语的大语言模型表现了很差的结果，体现了大语言模型在汉语兼类词不一致判断上的严重不足。后文会详细分析这一方法的结果。

6.4.1.3　词类对于大语言模型预测的影响

表 6.5 的最后一部分列出了词类对于大语言模型 DeepSeek 和通义千问模型的影响。使用“w. POS“表示大语言模型使用了带有词类信息加强的提示语，提示语的内容可以参考第 6.2.3 节的介绍。由于只有 DisWiC 数据集和汉语兼类词数据集 DisWiC_NV 具有词类信息，因此我们仅观察有无词类信息对于这两个数据集的影响。

表 6.5 中的结果显示，词类信息对于 DeepSeek 模型而言在两个数据集上都具有提升作用，其中对于 DisWiC_NV 数据集影响显著。而对于通义千问模型而言，同样在 DisWiC_NV 数据集上产生了积极影响，但是这一信息损害了英语数据集的性能。可能由于在该英语数据集上，每一个句对中的信息都基本具有一致的词类，因此这一信息对于结果帮助不大，同时不同的词类与不一致程度的关联没有非常

显著，因此有可能作为负面信息。第 6.4.5 节会通过具体实例来进一步分析。

6.4.1.4　小结

整体而言，不一致程度预测任务相较于第 4 章的相关性判断任务更为困难。这一差异主要源于两个任务所针对的目标不同。相关性判断任务要求直接评估目标词在不同上下文中的语义相关程度，该目标与词义本身密切相关，因此可以通过语义表征有效进行建模。而不一致程度预测任务则更为复杂：它不仅需要考虑目标词在上下文中可能存在的词义不确定性，还需进一步建模多名标注者之间可能出现的标注分歧。这类信息难以直接从基于分布式假设训练得到的语义表征中捕获，从而导致模型在该任务上的表现相对较差。

6.4.2　集成方法中不同集成方式和统计量分析

表 6.6　不同集成方式和指标下各类数据集的表现

集成指标	全部	汉语	英语	德语	挪威语	俄语	西班牙语	瑞典语	DisWiC_NV
ASD	**26.15**	49.96	**10.27**	**21.56**	38.25	**16.53**	7.34	**39.17**	16.71
MDP_c	26.12	**49.97**	10.23	21.42	**38.26**	16.42	**7.36**	39.15	16.48
MDP_d	16.76	38.21	6.28	15.13	20.71	8.59	5.88	22.51	18.76
VR	15.86	37.80	4.41	14.56	17.11	11.05	6.36	19.70	**20.08**

本节探究在集成方法中，不同集成方式以及统计量对于结果的影响。表 6.6 列出了不同集成方式和统计量下，各个数据集的表现。集成统计量分为两组，分别是基于连续值的 ASD 和 MDP_c 以及基于离散值的 MDP_d 和异众比率 VR。性能表现的前两名分别使用加粗和下划线来表示。

从表 6.6 中可以看出，对于 DisWiC 数据集而言，基于连续值的集成方式更佳，其中最优的标准差集成方式 ASD 要高于基于离散值的成对差异指标 MDP_d 将近 10 个百分点。且这种优势体现在各个语言中。这表明不同模型先求得目标词在成对上下文中的连续值相似度，之后在该相似度上求波动程度的方式可以有效捕捉不同模型间的不一致，尽管在实际不一致的标注中，是通过标注者标注的离散标签来计算的。这与连续值更具有鲁棒性、表示范围更广的优良特性相关。

对于中文兼类词数据集 DisWiC_NV 而言，观察到了相反的趋势：即表现更好的是基于离散值的异众比率 VR 指标，它的结果高于连续值指标标准差 ASD 3.37 个百分点。尽管差距不如 DisWiC 上的大，但是反映了在兼类词数据集上，先求得离散标签，再在标签上求不一致程度的过程更加有效。这可能源于 DisWiC_NV 数据集和 DisWiC 数据集分布和收集语料的差异。

表 6.7　不同集成方式和统计量下的最优组合

集成统计量	数据集	模型名称	层数	向量后处理	激活采样	提示语/输入
ASD	DisWiC(同)	LLaMA2	31	无	-	基
		LLaMA2	31	主	-	基
		LLaMA2	24	标	-	基
	DisWiC_NV(混)	XLM-B	7	标	-	-
		XLM-R	11	标	2	-
		LLaMA3	8	标	-	重
MDP_c	DisWiC(同)	LLaMA2	31	无	-	基
		LLaMA2	31	主	-	基
		LLaMA2	24	标	-	基
	DisWiC_NV(混)	XLM-B	1	标	-	-
		LLaMA3	16	标	-	基
		LLaMA2	8	标	-	原
MDP_d	DisWiC(混)	LLaMA2	31	标	-	知
		LLaMA2	31	主	-	基
		XLM-R	11	主	-	-
	DisWiC_NV(混)	XLM-B	4	标	-	-
		XLM-B	11	标	1	-
		LLaMA3	31	标	-	思
VR	DisWiC(混)	LLaMA2	31	无	-	基
		LLaMA2	31	标	-	语
		XLM-R	11	主	-	-
	DisWiC_NV(混)	XLM-B	11	标	2	-
		LLaMA3	31	中	-	基
		LLaMA3	31	标	-	思

对两个数据集而言，基于连续值内部和离散值内部的两个具体指标之间差异较小，这与它们提取表征以及计算不一致的过程相似有关。表 6.7 展示了在不同集成方式下最优组合的具体配置。其中表中的所有缩略写法与表 6.4 中所显示的完全一致。在各个统计量下对应的数据集后面，使用括号标注出了扰动策略：“同”,“混”分别代表“同质扰动”和“混合扰动”，具体设定可以参考第 6.3.3.1 节。

可以看出，相关集成统计量下的同一数据集使用的模型是类似的，这也反映在它们相似的不一致相关性评分上。对于 DisWiC 而言，大语言模型往往发挥更大作用，尤其在连续值集成方式中，而对于 DisWiC_NV 数据集而言，小编码器模型更加重要，这可以与 DisWiC 数据集规模较大，上下文更加复杂相关，以至于小模型无法准确捕捉语义以及不一致性。有关不同的扰动方式请参考第 6.4.3 节。

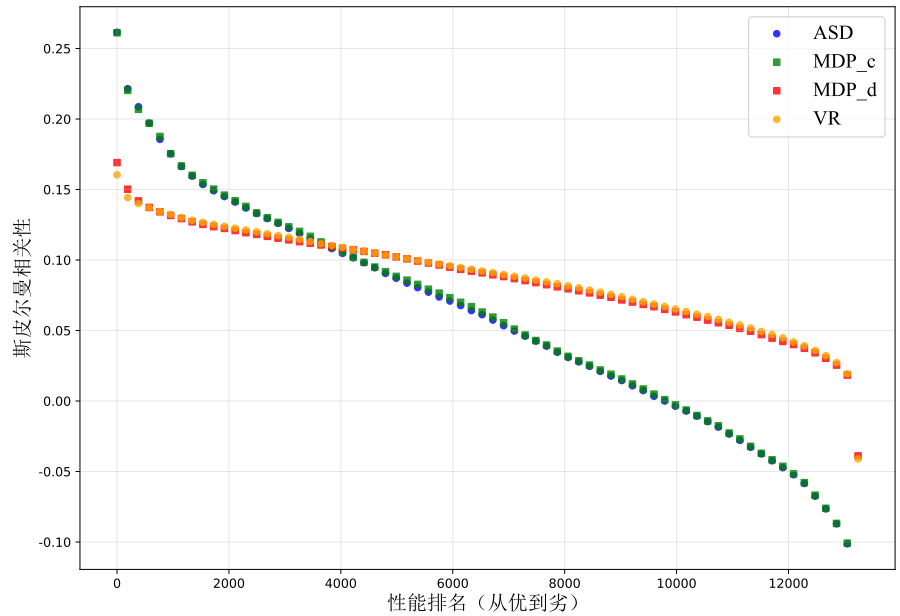


图 6.3　DisWiC 数据集上集成指标的实例排序相关性

图 6.3 和图 6.4 分别展示了四类集成指标在 DisWiC 和 DisWiC_NV 两个数据集上相关性由高到低排序后的整体变化趋势。可以看出，针对连续值建模的两种指标（ASD 与 MDP_c）呈现出高度一致的曲线形态，而针对离散值建模的两种指标（VR 与 MDP_d）同样表现出相似的变化趋势。这一现象表明，连续建模与离散建模在相关性刻画方式上存在显著差异，二者实质上反映了两类不同的集成机制。

对于图 6.3 反映的多语言相关度数据集 DisWiC 而言，从相关性分布范围来看，连续型指标（ASD 和 MDP_c）的取值区间明显更为宽广，其斯皮尔曼相关性系数由高端超过 0.25，逐步下降至低端约 −0.15；相比之下，离散型指标（MDP_d 和 VR）的相关性变化区间较为收敛，整体波动幅度更小。这表明，基于连续值的集成方式在区分不同实例时具有更强的分辨能力，能够放大实例之间的差异性。这

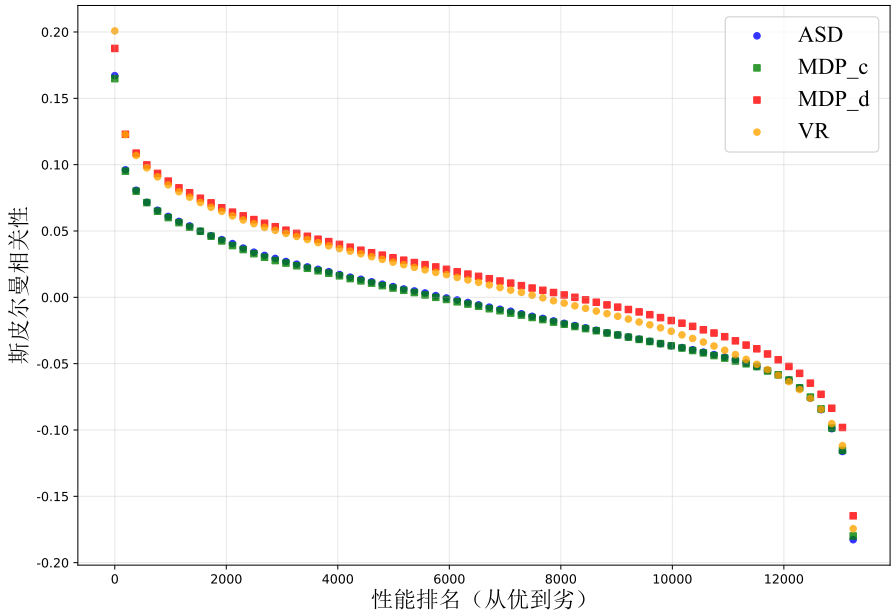


图 6.4　DisWiC_NV 数据集上集成指标的实例排序相关性

一特性与连续值表征的平滑性与可微性密切相关，使模型在刻画细粒度语义差异时更为敏感。

进一步观察曲线形态可以发现，在排序前段（高相关性区间），连续型指标不仅整体数值更高，而且下降趋势更为平缓，说明其在高一致性实例上的判别更稳定；而在排序后段（低相关性区间），连续型指标的下降幅度更大，体现出其对语义不一致或歧义实例具有更强的区分能力。相对而言，离散型指标在高低区间的变化均较为平缓，暗示其对不同程度语义差异的刻画较为粗糙，更倾向于对实例进行二值或有限类别的区分。

而对于图 6.4 反映的汉语兼类词数据集而言，这些指标的变化范围都是类似的，大致都在 0.2 到 −0.2 之间，且两条曲线的下降速度是相似的。其中连续型统计量（ASD 和 MDP_c）的曲线相较离散值统计量（MDP_d 和 VR）要低一些，反映了在兼类词 DisWiC_NV 数据集中，离散型统计量比连续型变量更占优势。

综合来看，首先同一类型内的统计量差异变化不大，二是不同数据集具有不同偏好的统计量选择，需要根据实际情况最终确定。

6.4.3　集成方法中不同扰动策略分析

6.4.3.1　DisWiC 数据集

图 6.5 展示了 DisWiC 数据集在采用标准化 STD 集成指标时，不同集成策略下各集成实例相关性的分布情况。该箱形图的横轴坐标表示不同扰动策略，其中同质扰动策略则具体拆分为各自模型的结果，括号中标记出相应的模型。纵轴使

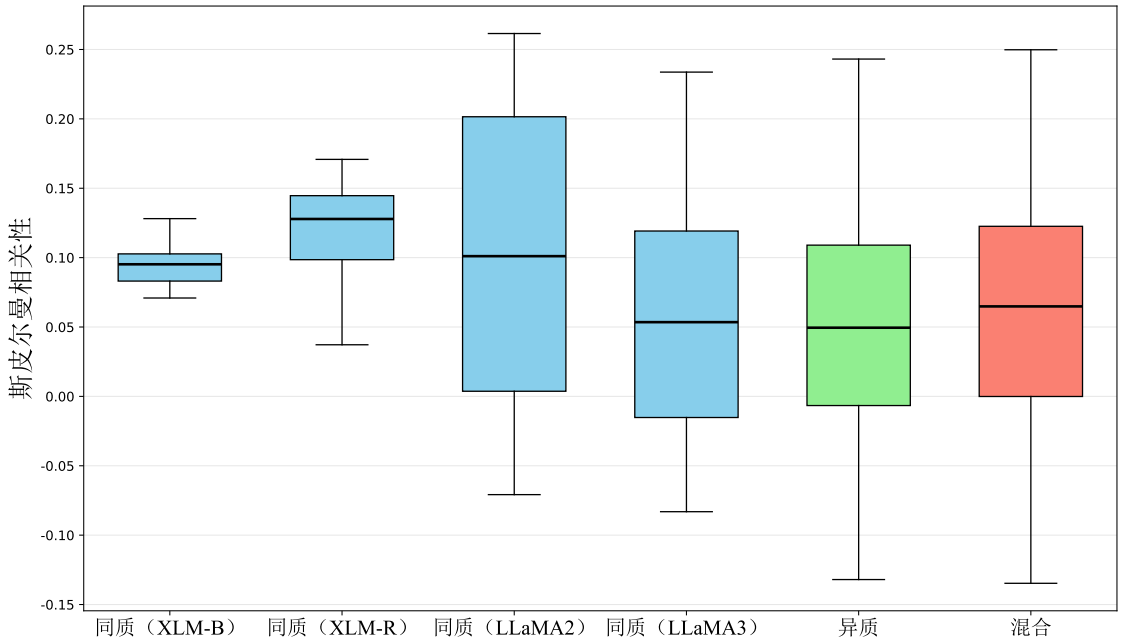


图 6.5　DisWiC 中集成策略下的相关性分布

用斯皮尔曼相关性来反映不同扰动手段下样本的分布。

整体而言，三类集成方式在分布形态与统计特征上呈现出明显差异：同质集成在中位数与最大值方面均优于其余两类方法，同时最小值更低；混合集成次之；异质集成整体表现最弱。这一结果表明，模型架构差异越大，集成后实例的相关性整体越低，集成效果随之减弱。

从分布特性来看，同质集成不仅在中位数上占优，而且其上界（最大值）更高，说明当多个结构相同或高度相似的模型进行集成时，更容易在部分实例上获得较高的一致性。这一现象可能源于同质模型在网络结构、参数规模以及训练目标等方面具有较强一致性，其内部表征空间分布相对接近，从而使集成结果更易形成稳定、协同的语义判断。相比之下，异质模型由于在架构设计、预训练语料与建模假设等方面存在显著差异，所产生的表征更为多样，集成时难以形成一致的语义信号，导致整体相关性偏低。

进一步比较不同模型结构下的同质集成策略，可以发现各模型在分布形态上仍存在差异。具体而言，RoBERTa 模型的集成结果在中位数上最高，表明其在多数实例上具有更稳定的语义一致性；LLaMA 模型的最大值更高，且整体分布范围更广，反映出其在部分实例上能够捕获更强的相关性，同时在不同实例间表现出更大的变异性。这说明 LLaMA 在语义特征选择上更为灵活，能够在更广泛的语义空间中进行建模，但也因此带来了更大的不确定性。相比之下，BERT 系列模型的分布更为集中，体现出较为保守而稳定的建模特性。

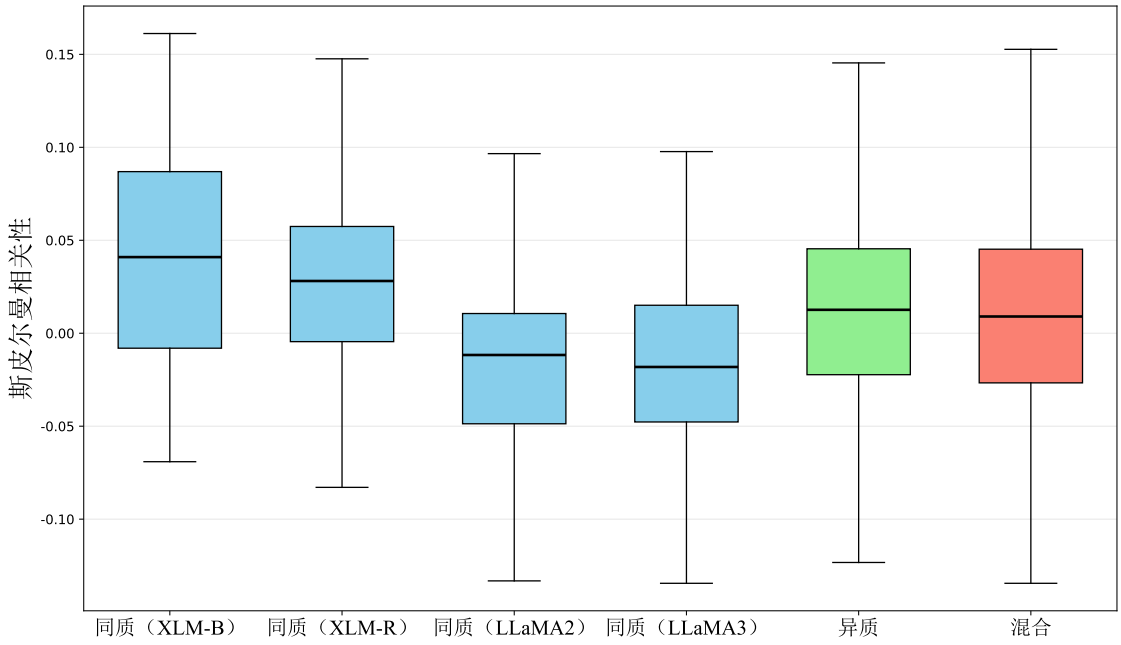


图 6.6　DisWiC_NV 在不同集成策略下集成实例相关性的分布

6.4.3.2　DisWiC_NV

图 6.6 则展示了 DisWiC_NV 数据集在不同扰动策略下的分布。其中同质扰动中，小规模的编码器模型表现更好，尤其是 XLM-B 模型，它具有最高的最大值、中位数以及最小值，这表明无论从最好表现还是整体表现，它都是最优的。尽管大语言模型在相关度的表现中比编码器模型取得更大的优势（参见表 4.4），它在不一致程度估计中的表现却不如编码器模型。

另外两种扰动方式，异质扰动和混合扰动具有更广的相关性变化范围，然而它们在几个关键评价（中位数、最值）上都不如仅进行同质扰动的评价。这表明在模型集成中，无需将模型的结构变化太多，它们可能存在不同模型之间表征不对齐的问题。

6.4.3.3　小结

综合来看，同质扰动策略在整体性能与稳定性方面均优于混合与异质集成，验证了“结构相似性有助于提升集成效果”的假设；而不同模型架构在集成表现上的差异，则进一步揭示了模型预训练目标与结构设计对语义表征一致性的影响。这一结果表明，在实际应用中，若目标是获得稳定且一致的集成结果，应优先选择同构或高度相似的模型进行集成；而在强调多样性或探索语义空间边界的场景中，使用更加多样的模型集成则可能提供更具信息量的补充。

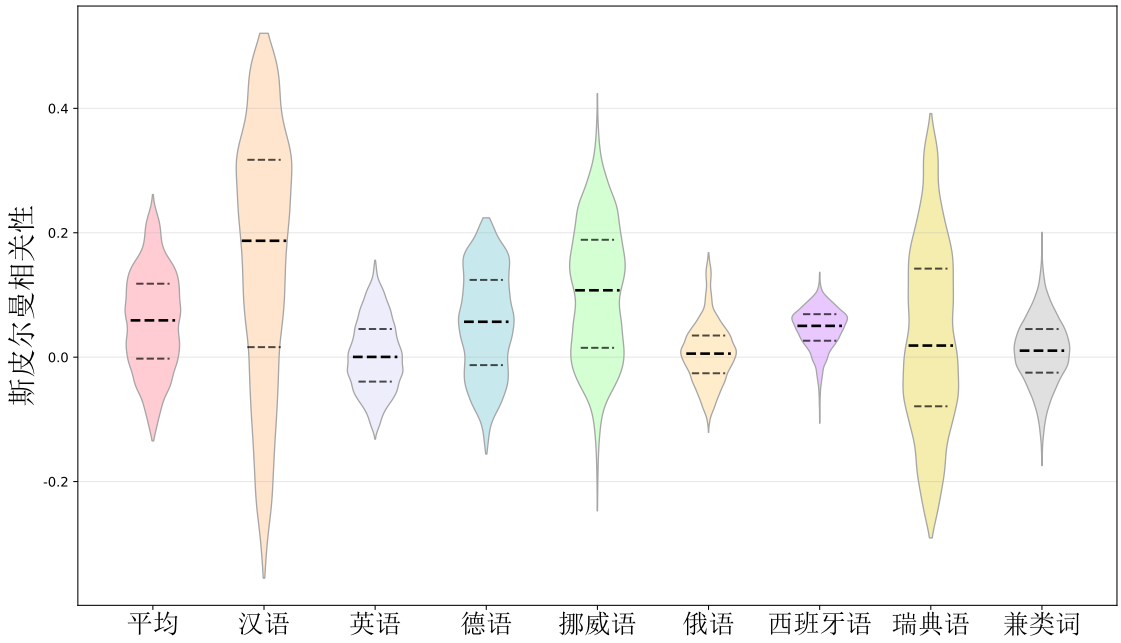


图 6.7　不同语言集成实例相关性的分布

6.4.4　不同语言集成后的表现分析

图 6.7 展示了不同语言的数据集集成实例相关性的分布情况。其中，除“兼类词”来自 DisWiC_NV 外，其余都来自多语言数据集 DisWiC。采用小提琴图绘制在不同语言上，各类集成实例组成的不一致度的分布。集成统计量选自在表 6.6 中显示的各类数据集最佳的指标：DisWiC 选择连续型指标 ASD；DisWiC_NV 数据集使用离散型指标 VR。

从整体趋势来看，各语言在相关性水平与分布形态上存在显著差异，表明语言特性对集成效果具有重要影响。就中位数与分布位置而言，汉语与挪威语整体表现较好，而英语、汉语兼类词、俄语与西班牙语的相关性水平相对较低，未呈现出明显优势。

进一步从分布形态进行分析可以发现，汉语 ①的相关性分布跨度最大，既包含较高相关性的实例，也存在一定比例的低相关性样本。这一特征表明，模型在汉语实例集成过程中，能够较为充分地区分不同程度的语义相关性，体现出较强的判别能力；但与此同时，也暴露出在部分样本上仍存在较大的不一致性。相比之下，挪威语的分布更加集中，整体波动幅度较小，说明其集成结果在多数实例上更为稳定。

上述差异可能与各语言数据集的标注特性及语义分布复杂度密切相关。一方面，汉语与挪威语在本数据集中标注者数量相对较少，且词义分布结构较为简单，

① 以下“汉语”均指 DisWiC 数据集中的汉语部分，而不是汉语兼类词。

这使得模型在集成时更容易形成一致的语义判断，从而获得较高的相关性；另一方面，英语、汉语兼类词、俄语与西班牙语由于语义分布更为多样、歧义现象更为复杂，集成过程中更容易引入分歧，导致整体相关性水平下降。该解释亦可从图 3.3 和图 3.10 所示的标注分布中得到印证。

综合而言，不同语言之间在集成效果上的显著差异说明，集成性能不仅取决于模型本身，也受到语言资源特性、标注分布以及语义复杂度等因素的共同影响。这一结果提示，在多语言语义建模与集成应用中，有必要针对不同语言的结构特点与数据分布特征进行差异化建模与评估，而非采用统一的集成策略。

6.4.5　典型偏误分析

本小节分析大语言模型在两类数据集上评估出错的实例，重点针对英语和汉语两种语言。大语言模型使用的提示语可以参考第 6.2.3 节，其中包括基本提示语和使用词类信息的提示语。

6.4.5.1　两类错误类型

与第 4.4.6 节在词义相关度判断中的错误类型划分标准类似，在不一致判断中，本节也按照与专家标注结果相比，模型输出结果高估和低估两种情况进行偏误归类。这一过程包括以下三个步骤：

（1）将模型预测的不一致程度和专家标注的不一致程度进行归一化，统一在相同的值变化区间。由于模型预测的区间是 0 到 1，而专家标注的区间在 0 到 3 之间（参加图 3.3 中展示的标注统计量），因此我们将专家的标注除以 3，归一化到 0 到 1 之间。

（2）使用模型预测的不一致程度值减去专家标注的不一致程度值，作为误差程度。并且选取阈值作为可被接受的波动范围。

（3）当实例的误差程度大于该阈值时，表明模型预测的值远超出了专家标注的结果，这时将该实例判定为“高估错误”；否则，当误差程度小于该阈值的相反数时，则表明模型预测的值远低于专家标注的结果，这时将该实例判定为“低估错误”。可以用下面公式表示：

$$\text{错误类型}(i)=\begin{cases}\text{高估错误}, & \hat{y}_i - y_i > \tau \\ \text{低估错误}, & \hat{y}_i - y_i < -\tau\end{cases} \tag{6.1}$$

其中，y_i 表示第 i 个实例的专家标注结果，$\hat{y}_i$ 则表示模型预测的结果，τ 则代表用于判断可接受程度的范围。在本节中，τ 的取值为 0.5，即可能范围的一半。

表 6.8 展示了在不同类型的数据集上，DeepSeek 和通义千问模型各自在两类

表 6.8　不同提示语下模型对词义不一致判断的错误统计

数据集	语言	模型名称	提示语类别	样例数量	
				高估样例	低估样例
DisWiC_ZH	汉语	DeepSeek	基础提示语	1943	**13**
			带词类提示语	835	15
		通义千问	基础提示语	1350	27
			带词类提示语	**834**	222
DisWiC_EN	英语	DeepSeek	基础提示语	2867	**4**
			带词类提示语	1796	15
		通义千问	基础提示语	3025	10
			带词类提示语	**1594**	154
DisWiC_NV	汉语	DeepSeek	基础提示语	68	**0**
			带词类提示语	72	1
		通义千问	基础提示语	**65**	1
			带词类提示语	129	3

提示语中的误差汇总，其中误差包括高估样例和低估样例。可以得出以下几个结论：

（1）模型普遍高估了不一致程度。所有情况下，高估样例总是显著地高于低估样例，表明模型倾向于认为标注是不一致的。

（2）在 DisWiC 数据集中，带有词类信息的提示语可以缓解高估的类型，取而代之的是低估样例的增加。例如在通义千问模型对不一致数据集中的英文部分 DisWiC_EN 的不一致程度估计中，使用了带词类的提示语，使得模型的高估样本降低了将近一半，而低估样本则增加到原来的 15 倍之多。这体现了词类信息的提供有助于使模型倾向于对词义相关度的标注更加一致。这可能由于在 DisWiC_EN 数据集中，同一目标词在成对的上下文中总是一样的，提供了这一信息帮助模型认定它们处于同一词类下，从而不会产生更高的不一致度。

（3）相比而言，词类信息对于汉语兼类词数据集 DisWiC_NV 而言，趋势是相反的。它更容易使得模型增加高估样例的数量，例如对于通义千问模型而言，带词类

提示语判断出的高估样例是基础提示语的接近两倍。与上述理由类似，DisWiC_NV 数据中，目标词出现的两个上下文的环境词类不同，分属名词和动词。模型在知道这一信息后，更倾向于因词类信息不同，而认定标注更容易出现分歧，从而使得不一致程度普遍估计过高。

（4）通义千问模型更倾向于低估样例，而 DeepSeek 模型更倾向于高估样例。这体现在 DisWiC_ZH 数据集和 DisWiC_EN 数据集的带词类的提示语中。不过对于 DisWiC_EN 的基础提示语设定中，通义千问模型的高估和低估样例都更高。总体而言，通义千问模型做地更好一些，这也体现在表 6.5 呈现的不一致判断中的优势中。

6.4.5.2　高估错误类型分析

高估错误夸大了本来不一致程度低的样例的不一致程度，针对错误类型的实例，本小节重点分析通义千问模型。通过观察和归纳，可以发现如下结论：

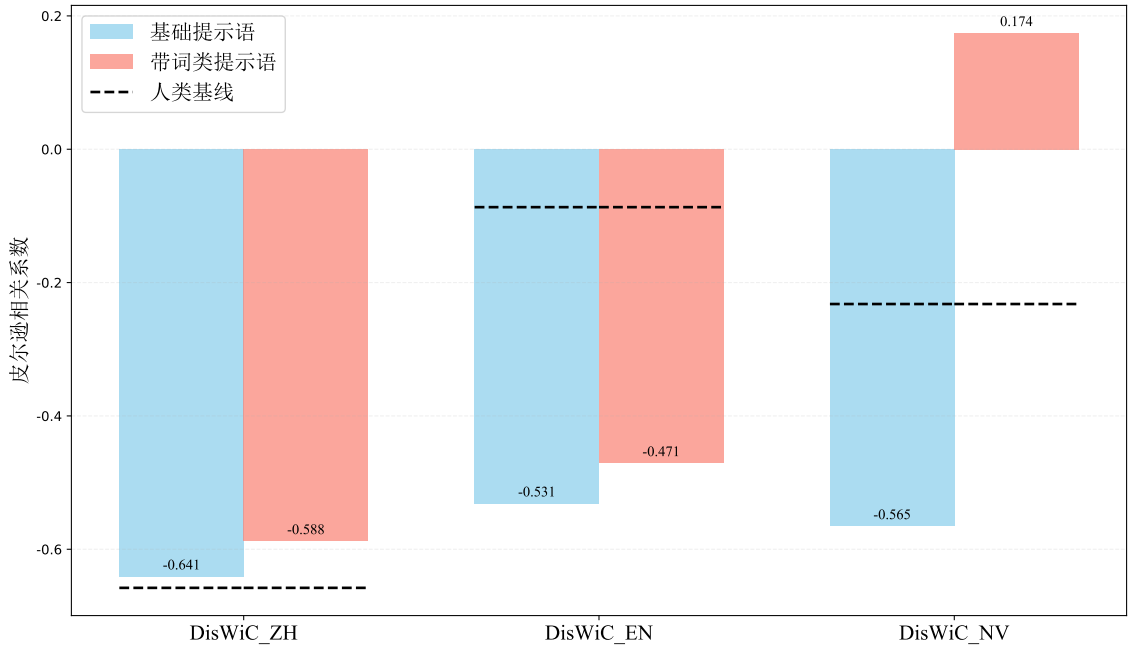


图 6.8　平均标注相关度与模型预测的不一致程度之间的相关性

（1）模型倾向于将认定为“不相关”的实例判定为“不一致”。

图 6.8 展示了大语言模型对于目标词出现在不同上下文的相关度和模型预测的不一致度的相关性，其中相关度采用不同标注的平均值，相关性采用皮尔逊相关系数（Pearson correlation coefficient）。图中使用黑色虚线标出了相关度水平和人类标注的不一致程度的相关性。从图中可以看出，在 DisWiC 的两个数据集中，无论提示语中是否携带词类信息，相关度与不一致度总是呈现很高的负相关性的。这表明模型倾向于将“不相关”和“不一致”两个概念混同理解，即认为越不相关，不一

致度越高。这一趋势在人类标注者中也存在与 DisWiC_ZH 中，但是在 DisWiC_EN 中人类基线却没有呈现像模型预测一样高的相关性。

例如下面这个来自 DisWiC_ZH 中的例子[①]：

(1) a. ……小组长十分高兴，马上夹了一块腌耗子脖颈蘸上生猪血给他吃……

b. 现在，我们三十来岁这个年龄的人，以及更年轻的人，已经基本上不会腌咸菜

c. 专家标注：0.0 （4，4）

d. 模型预测：0.9

e. 模型解释：The target word ’ 腌’ refers to ’preserving with salt’ in both sentences, but the contexts differ drastically: Sentence 1 describes a bizarre, possibly fictional scenario of eating a preserved rodent, while Sentence 2 discusses a common, practical skill of making pickled vegetables. This stark contrast in connotation, cultural familiarity, and potential ambiguity (e.g., literal vs. metaphorical, or different preservation methods) is likely to cause significant disagreement among annotators on semantic relatedness.

f. 模型解释（翻译）：目标词”腌”在两个句子中均指”用盐保存”，但语境截然不同：句 1 描述了一个奇怪的、可能是虚构的吃腌制啮齿动物的场景，而句 2 讨论了一个常见的、实用的腌菜制作技能。这种在内涵、文化熟悉度和潜在歧义性（如字面义与比喻义，或不同的保存方法）上的鲜明对比，很可能导致标注者在语义相关性判断上产生显著分歧。

在例 (1) 中尽管“腌”一词在文中被标注者一致地认定为相同词义（相关度均为 4），模型却给出了程度非常高的不一致程度（接近最高的 1）。仔细阅读模型给出的解释，可以发现模型过度强调了目标词在两个上下文中的不同之处，例如“腌”出现的不同语境。这正是第 4.4.6.1 节分析的模型低估相关度的一类“语义域过度分化”类型。通过区分二者的相关度，进而得出“产生显著分歧”的结论。而事实上，“腌”的基本词义在两个上下文中是一致的，都表示“把鱼、肉、蛋、蔬菜、果品等加上盐、糖、酱、酒等”[②]，同一言语社区的标注者无需强调语境的差异，就可以做出清晰的判断。

① 以下例子中编号 a 和 b 表示数据集中的两个上下文，其中下划线表示上下文中的目标词。编号 c 是专家标注的结果，其中括号中是多个标注结果，外部的数值反映了不一致程度。标注的结果使用了归一化操作，保证最小值和最大值分别为 0 和 1，数值越大，不一致程度越高。编号 d 为模型预测的结果，数值范围和含义同“专家标注”。编号 e 给出了模型的解释，其中下划线反映了需要强调的部分，红色标记出了解释不合理的语句。编号 f 给出了解释的中文翻译。

② 参见《现代汉语词典》[38] 有关“腌”的词条。

再如 OGWiC_EN 中下面的例句：

(2) a. He had on a straw hat, and he tipped it as the car pulled away.

b. Nineteen federal, state and local law enforcement agencies are examining Katrina aid payments for crime patterns and investigating fraud tips from the public.

c. 专家标注：0.0 （1，1）

d. 模型预测：0.9

e. 模型解释：The target word 'tip_vb' has two distinct meanings: in Sentence 1, it means 'to tilt or move slightly' (as in tipping a hat), while in Sentence 2, it refers to 'providing information or a suggestion' (as in fraud tips). These are semantically unrelated, leading to high annotator disagreement on whether they are related, as some may focus on the shared word form while others on the clear meaning difference.

f. 模型解释（翻译）：目标词"tip_vb" 有两个不同的含义：在句 1 中，它表示" 轻轻倾斜或移动"（如 tipped a hat），而在句 2 中，它指" 提供信息或建议"（如 fraud tips）。这两个含义在语义上不相关，导致标注者在判断它们是否相关时产生高度分歧，因为一些人可能关注共享的词形，而另一些人关注明显的含义差异。

在该英语例子中，模型虽然正确地预测了 tip 在两个上下文中表示完全不同的含义，然而进一步地，模型因为它们语义的不相关就认定“标注者在判断它们是否相关时产生高度分歧”，而且还给出进一步的解释，“因为一些人可能关注共享的词性，而另一些人关注明显的含义差异。”这种解释是站不住脚的，因为标注的提示语中明确说明了“标注者需判断目标词在这两个语境中语义上的相关程度”，且根据常识，标注者应该是对标注的语言很了解的，不太可能存在仅观察到词形相同就可能判断语义相关的错误情况。

对于 DisWiC_NV 数据集而言，基础提示语的相关性仍显著高于人类，表明也存在上述类似的问题。试举如下的例子：

(3) a. 用实际行动纪念先烈

b. 这张照片给你做个纪念吧

c. 专家标注：0.0（4，4）

d. 模型预测：0.8

e. 模型解释：The target word ' 纪念' has distinct senses: in sentence 1 it means

"commemorate" (a solemn, public act), while in sentence 2 it means "keepsake" (a personal memento). This polysemy is likely to cause significant disagreement, as some annotators may focus on the shared abstract notion of remembrance, while others emphasize the concrete contextual differences.

f. 模型解释（翻译）：目标词"纪念"具有不同的义项：句 1 中意为"纪念"（庄严的公共行为），而句 2 中意为"纪念品"（个人留念）。这种一词多义现象很可能导致标注者产生显著分歧，因为有些标注者可能关注共同的抽象概念，而另一些则强调具体的语境差异。

在例 (3) 中，"纪念"一词同样因为语义区分过细而导致模型判定为"不同的义项"，从而认为这种"多义性"导致预测结果的不一致。

（2）词类信息容易将高相关度的实例视作高不一致。

图 6.8 在 DisWiC_NV 数据集下，使用了词类相关的数据集逆转了这一趋势，使得相关度由"很负相关"（–0.565）提升到略微正相关的程度（0.174），这表明不一致判断过程中，"词类不同"（一个为名词，一个为动词）这一信息可以促使模型更倾向于因为这一差异，认定目标实例更可能出现高不一致的情况，即使标注者们一致判定它们语义高度相关或者相同。例如针对下面的例子：

(4) a. 教练教练得法。

b. 足球教练。

c. 专家标注：0.0 （3.0，3.0）

使用基础提示语后，千问大模型给出的预测结果为 0.3，并做出以下解释：

(5) a. 模型解释：Sentence 1 uses ' 教练' in a general sense of 'coaching method' or 'coach's approach,' while Sentence 2 specifies 'football coach.' This creates some ambiguity in semantic relatedness, as annotators might debate whether the general and specific uses are closely related or distinct, leading to moderate disagreement.

b. 模型解释（翻译）：句 1 中的" 教练" 泛指" 教练方法" 或" 教练的指导方式"，而句 2 特指" 足球教练"。这种用法在语义相关性上产生了一定的歧义，标注者可能会争论这种泛指与特指用法是密切相关还是有所区别，从而导致中等程度的分歧。

而在使用了包含词类信息的提示语后，大语言模型给出的不一致程度预测结

果为 0.8，并做出以下解释：

(6)　a.　模型解释：The target word ' 教练' is used as a verb in Sentence 1 ('to coach' or 'to instruct') and as a noun in Sentence 2 ('coach' or 'instructor'). This part-of-speech shift involves a significant semantic change between an action and a role/person, likely leading to high annotator disagreement on semantic relatedness.

b.　模型解释（翻译）：目标词" 教练" 在句 1 中用作动词（指导、训练），在句 2 中用作名词（教练、指导者）。这种词性转换涉及动作与角色/人物之间的显著语义变化，很可能导致标注者在语义相关性判断上产生高度分歧。

可以看出包含词类信息的提示语会引导模型认定为产生了“显著语义变化”，从而认为标注者在语义相关度判断上产生“高度分歧”。这种解释夸大了词类对于汉语兼类词的理解差异，正如第 4 章分析，汉语的词类差异对于部分兼类词来说语义差异不明显，标注者不会单纯因为词类不同就会判断它们是“远相关”或者“不相关”。相反，标注者更倾向于一致地认为是“词义一致”或者“词义近相关”的实例（参考第 3.2.4 节）。

同时，带词类的提示语尽管可以让不一致程度不再于相关度呈现负相关，略微正相关（0.174）也同样偏离了实际标注的情况：实际标注应该是略微负相关（图 6.8 中 DisWiC_NV 数据集上的虚线部分）。

值得注意的是，在多语言数据集 DisWiC_ZH 和 DisWiC_EN 中，词类并没有带来相反的趋势，这是由于绝大多数实例在目标词出现的成对上下文中具有一致的词类，从而使这一信息无法有效加深模型对不一致程度的判断。

6.4.5.3　低估错误类型分析

表 6.8 所显示的低估样例数量要远小于高估样例。我们仅选取典型的低估样例进行分析。低估类型往往意味着标注者产生了较大的不一致，而模型倾向于认为标注者会给出一致的答案。这一般出现在相关度比较大，但可能由于词类、句法环境等对不同标注者造成了干扰。

例如选自 DisWiC_ZH 数据集的“熟”在下面的案例中，尽管都表示“因常见或常用而知道得清楚”①，在前一个上下文中，“熟”表示一种“动态过程”，后一个上下文中“熟”表示一种“静态过程”，这可能导致标注者产生分歧。但是语言

① 参考《现代汉语词典》[38] 中“熟”的词条释义。

模型在预测时候没有考虑这些背景。

(7) a. 这件事使我认识到：越是在情况熟了的时候，越是要虚心，切不可自以为是。
b. 过去，他自以为人熟、地熟、情况熟，凭老经验办事，容易犯经验主义错误。
c. 专家标注：1.0（4，1）
d. 模型预测：0.1
e. 模型解释：In both sentences, ’ 熟’ refers to familiarity with a situation or environment, indicating a high degree of semantic similarity. The contexts are nearly identical in meaning, with only slight variation in emphasis, so annotators are likely to agree closely.
f. 模型解释（翻译）：在两个句子中，”熟” 都指对情况或环境的熟悉，表明语义相似度很高。两个语境在意义上几乎相同，只是在强调重点上略有差异，因此标注者很可能会达成一致。

此外，模型预测值与真实值之间的偏差也可能源于标注者自身的理解偏差或标注质量问题。以汉语兼类词数据集 DisWiC_NV 为例，部分标注者可能将词类差异视为语义不相关的重要依据，从而导致其判断与其他标注者产生分歧。例如，在下面样例中，两位标注者判断语义相关，而另一位标注者则给出“不相关”的回答。尽管缺乏标注理由的说明，但词类差异很可能是导致这一分歧的关键因素。

(8) a. 耍赖皮。
b. 别在这儿赖皮了，快走吧！
c. 专家标注：0.667（4.0，4.0，1.0）
d. 模型预测：0.1
e. 模型解释：Both sentences use ’ 赖皮’ to mean behaving in a deceitful or evasive way, with slight variation in tone (general behavior vs. direct command), but the core meaning remains highly similar, leading to minimal annotator disagreement.
f. 模型解释（翻译）：两个句子中”赖皮” 都指以欺骗或逃避的方式行事，只是在语气上略有差异（一般行为描述 vs. 直接命令），但核心意义高度相似，因此标注者分歧较小。

值得注意的是，此类存在分歧的标注在传统研究中常被视为“标注错误”而

予以剔除。然而，本文并未采取这一做法，原因在于：

(1) 本章任务旨在预测标注不一致程度，而非消除不一致，因此需要保留多样化的标注情况以覆盖真实场景；

(2) 标注分歧本身蕴含语义信息——词义模糊或难以判断的样本更易引发分歧，剔除这些样本将损失重要的训练信号。

与之不同，在第 4 章的相关度判断任务中，我们剔除了标注差异过大的样例（任意两位标注者评分差值大于 1），以确保训练数据具有较高的标注一致性，从而保证模型学习目标的可靠性。

6.5　本章小结

本章研究了词义相关度标注中的不一致现象，这是第 4 章相关度判断任务的深化与拓展。真实语言数据中存在诸多不确定性因素——词义自身的模糊性、多义性、歧义性，以及外部上下文的完整性和标注者个体差异——这些因素共同导致标注结果呈现不同程度的分歧。对不一致性的评估有助于更全面地反映大语言模型在真实场景下的语义理解能力。

本章提出三类标注不一致程度预测方法：分类器探测方法、模型集成方法和基于提示语的探测方法。实验结果表明，三类方法均能有效预测不一致程度，其中模型集成方法表现最佳，能够较好地模拟不同标注者的判断差异。通过对大语言模型输出结果的分析，本文进一步揭示了模型处理多义词时的典型错误类型。

相较于第 4 章的相关度判断任务，不一致程度预测更具挑战性。当前方法的预测结果与人类标注者仅呈现微弱相关性，仍有较大提升空间。未来研究可从以下方向展开：探索更精确的建模方法、系统分析影响标注不一致的因素、利用不一致程度指导模型训练，以及在更多语言上验证方法的泛化能力。

第 7 章　总结和展望

7.1　论文主要工作和贡献

本文首先介绍了词义相关度任务以及相关的数据集，其中包含公开的多语言多标注形式的数据集和本文自行收集的汉语兼类词数据集。此后具体研究了模型在词义相关度上不同维度的评估包含准确性评估和标注不一致评估。从准确性评估出发，探究了大语言模型为主的神经网络语言模型的信息流动模式的差异。本文的主要贡献如下：

（1）汉语兼类词数据集构建：本文构建了一个大规模汉语兼类词词义相关度数据集，引入词性标注并显著扩展目标词规模与标注数量，语料来源更加规范多样。通过设计名动“最小对立”句对，有效弥补了既有汉语数据集中词类信息缺失的问题，为跨语言词义对比与模型评估提供了更可靠的数据基础。

（2）词义相关度的多维评估与分析：本文围绕词义相关度任务，提出了结合提示语引导、表征提取与向量后处理的评估框架，对不同类型模型进行了系统的定量与定性分析。结果表明，特定任务微调编码器模型表现最佳，提示语驱动的大语言模型次之。进一步从误差类型出发，总结了模型高估与低估的典型原因，加深了对模型词义表征能力的理解。

（3）标注不一致性的建模与评估：本文从多标注者视角出发，系统研究了模型对标注不一致程度的预测能力。通过构建补充数据集并引入模型集成方法，模拟多专家决策过程，实现了对不确定性的有效刻画。实验表明，模型集成在该任务上表现最优，同时定性分析揭示了模型在不一致性预测中的主要偏误模式。

（4）大语言模型的跨层信息流动分析：本文从跨层角度对比分析了解码器与编码器模型的表征演化过程，验证了“先理解后预测”的动态模式：解码器模型在中低层达到词义理解最优，高层则侧重预测能力；编码器模型则随层数加深持续增强理解能力。该发现为大语言模型的表征提取与可解释性研究提供了新的视角。

7.2　下一步研究工作展望

词义理解是检测大语言模型基础能力的一项重要评估任务，已有的工作对此做了一系列的探讨，未来仍可以在此基础上做进一步研究。

（1）结合语言学理论，设计更多词义理解相关的任务

词义理解涉及到方方面面，尽管语言模型在现有的评测数据集上有了较为满

意的结果，但这并不意味着它对于词义的完全理解。这是由于这些数据集往往针对“准确性”这一维度，且常常是互联网中出现过的语料，任务形式比较单一。本文额外增加了“标注不一致”这一衡量词义“不确定性”的评测，未来仍可以结合语言学理论设计更多词义理解相关的任务。例如语言学中更为关注的虚词理解[25,59]、歧义现象[220-221]、词义演变[25,222]等。这些任务往往需要设计对照组[59]或者与语言特性和语言类型[25]相关，避免了大语言模型可能已经见过的实例，可以更客观地反映模型的语义理解能力。

（2）与机制可解释性结合，对模型进行更加深入的理解

本文所涉及的可解释性主要从跨层信息流动的角度自上而下地进行分析，缺乏对大语言模型自下而上的理解。这可以与强调“干预”行为的机制可解释性相结合：一方面可以结合激活替换、局部消融、表征编辑等更多的干预和对照手段，对不同层级的功能分工做进一步验证。另一方面可以探究更小的计算单元，例如神经元或者神经回路[13]中与词义理解相关的部分。例如本文得出的“理解”和“预测”的两个功能是否由神经网络中的某些特定神经元专门实现，是否可以通过编辑某些内部结构来控制不同的功能等。通过更加细粒度的剖析，可以为深入了解并利用模型提供有力支撑。

（3）与下游应用和任务结合

词义理解评估可以用于语言教育场景。在第二外语习得[6]中，可以开发智能系统辅助教师教授语言，其中涉及到智能模型对于各种语言中的词义理解。例如，利用大语言模型自动理解词义，并利用词义相关度检索词义相似的词，可以为学生提供更多应用实例。同时相关度也可以帮助学生判断两个词词义的临近程度，从而辅助教师进行词义辨析。不一致度预测理解差异较大的词及上下文也可以作为理解的困难样本，着重加以讲解。

模型可解释性的工作与人工智能的安全性密不可分。本文在表征上的工作可以增加对模型的进一步理解，进一步增进模型的安全性能。例如根据“理解”和“预测”的两种模式在不同层的表现，可以提示对模型进行攻击和安全防御的层信息。之后便可以通过编辑模型，来定向预测模型的表现。如更改高层表征信息仅改变模型的预测，而无碍于对当前词义的理解等。与机制可解释性[223-224]的结合，也有助于从更底层进行编辑来更改模型的行为。

7.3　本章小结

本章对全文的研究工作进行了总结与展望。首先系统概括了论文的研究内容与主要创新贡献；在此基础上，从任务、可解释性分析和下游任务三个方面，对未

来可能的研究方向进行了进一步展望。通过上述总结与讨论，本文不仅回顾了已有工作的主要成果，也为后续深入研究提供了思路与参考。

参考文献

[1] Carpuat M, Asscher O, Bali K, et al. An interdisciplinary approach to human-centered machine translation[C]//Christodoulopoulos C, Chakraborty T, Rose C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 22859-22879.

[2] Zhu W, Liu H, Dong Q, et al. Multilingual machine translation with large language models: Empirical results and analysis[C]//Findings of the association for computational linguistics: NAACL 2024. 2024: 2765-2781.

[3] Liu D, Wu Z, Song D, et al. A persona-aware LLM-enhanced framework for multi-session personalized dialogue generation[C]//Che W, Nabende J, Shutova E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 103-123.

[4] Yang D, Yuan R, Fan Y, et al. RefGPT: Dialogue generation of GPT, by GPT, and for GPT [C/OL]//Bouamor H, Pino J, Bali K. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 2511-2535. https://aclanthology.org/2023.findings-emnlp.165/. DOI: 10.18653/v1/2023.findings-emnlp.165.

[5] Pal Chowdhury S, Daheim N, Kochmar E, et al. Large language models for education: Understanding the needs of stakeholders, current capabilities and the path forward[C]//Kochmar E, Alhafni B, Bexte M, et al. Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025). Vienna, Austria: Association for Computational Linguistics, 2025: 1-10.

[6] Ye J, Wang S, Zou D, et al. Position: LLMs can be good tutors in English education[C]//Christodoulopoulos C, Chakraborty T, Rose C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 17516-17535.

[7] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Comput. Surv., 2023, 55(12): 248:1-248:38.

[8] The paradigm of hallucinations in ai-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations[J]. Computers and Electrical Engineering, 2025, 124: 110307.

[9] Luo J, Wu B, Luo X, et al. A survey on efficient large language model training: From data-centric perspectives[C]//Che W, Nabende J, Shutova E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 30904-30920.

[10] Wang A, Singh A, Michael J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding[C]//Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP. 2018: 353-355.

[11] Wang Y, Ma X, Zhang G, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark[J]. Advances in Neural Information Processing Systems, 2024, 37: 95266-95290.

[12] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models: Vol. 15[M]. New York, NY, USA: Association for Computing Machinery, 2024.

[13] Saphra N, Wiegreffe S. Mechanistic?[C]//Belinkov Y, Kim N, Jumelet J, et al. Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. Miami, Florida, US: Association for Computational Linguistics, 2024: 480-498.

[14] Zou A, Phan L, Chen S, et al. Representation engineering: A top-down approach to ai transparency[A]. 2023.

[15] Li M Y, Liu A, Wu Z, et al. A taxonomy of ambiguity types for nlp[A/OL]. 2024. arXiv: 2403.14072. https://arxiv.org/abs/2403.14072.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] 黄伯荣, 廖序东. 现代汉语[M]. 北京: 高等教育出版社, 2003.

[18] Partee B H. Compositionality[M]//Landman F, Veltman F. Groningen-Amsterdam Studies in Semantics: Vol. 3 Varieties of Formal Semantics. Dordrecht: Foris Publications, 1984: 281-311.

[19] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [C]//Erk K, Smith N A. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1715-1725.

[20] 赵元任. 语言问题[M]. 北京: 商务印书馆, 1980.

[21] Lyons J. Linguistic semantics: An introduction[M]. Cambridge: Cambridge University Press, 1995.

[22] Taylor J R. Linguistic categorization[M]. 3rd ed. Oxford: Oxford University Press, 2003.

[23] Lakoff G, Johnson M. Metaphors we live by[M]. Chicago: University of Chicago Press, 1980.

[24] Lakoff G, Johnson M. Philosophy in the flesh: The embodied mind and its challenge to western thought[M]. New York: Basic Books, 1999.

[25] 李小凡, 张敏, 郭锐. 汉语多功能语法形式的语义地图研究[M]. 北京: 商务印书馆, 2015.

[26] François A. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages[M]//Vanhove M. From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations. Amsterdam: John Benjamins, 2008: 163-215.

[27] Haspelmath M. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison[M]//Tomasello M. The new psychology of language: Vol. 2. Mahwah, NJ: Lawrence Erlbaum, 2003: 211-243.

[28] Georgakopoulos T, Polis S. The semantic map model: State of the art and future avenues for linguistic research[J]. Language and Linguistics Compass, 2018, 12(2): e12270.

[29] Liu Z, Liu Y. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation[C]//Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, 2023: 3963-3977.

[30] Zheng H, Li L, Dai D, et al. Leveraging word-formation knowledge for chinese word sense disambiguation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 918-923.

[31] Chew M E E, Ng L S, Jaafar N M, et al. Understanding oriental and western dragons in a globalised world: A cross-linguistic study of dragon-based metaphorical expressions in chinese and english[J]. 3L: Language, Linguistics, Literature, 2024, 30(4): 1-15.

[32] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to wordnet: An on-line lexical database [J]. International journal of lexicography, 1990, 3(4): 235-244.

[33] McCarthy D. Relating WordNet senses for word sense disambiguation[C/OL]//Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together. 2006. https://aclanthology.org/W06-2503/.

[34] Xu H, Mannor S. Robustness and generalization[J/OL]. Machine Learning, 2012, 86(3): 391-423. DOI: 10.1007/s10994-011-5268-1.

[35] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C/OL]//Burstein J, Doran C, Solorio T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186. https://aclanthology.org/N19-1423/. DOI: 10.18653/v1/N19-1423.

[36] Raganato A, Pasini T, Camacho-Collados J, et al. Xl-wic: A multilingual benchmark for evaluating semantic contextualization[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 7193-7206.

[37] Schlechtweg D, Choppa T, Zhao W, et al. The CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments[C]//Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations. Abu Dhabi, UAE, 2025.

[38] 汉语现代词典[M]. 北京: 商务印书馆, 2005.

[39] Allen J. Natural language understanding[M]. Benjamin-Cummings Publishing Co., Inc., 1995.

[40] Armendariz C S, Purver M, Pollak S, et al. Semeval-2020 task 3: Graded word similarity in context[C]//Proceedings of the Fourteenth Workshop on Semantic Evaluation. Barcelona (online): International Committee for Computational Linguistics, 2020: 36-49.

[41] Pilehvar M T, Camacho-Collados J. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 1267-1273.

[42] Emerson G. Functional distributional semantics: Learning linguistically informed representations from a precisely annotated corpus[D]. University of Cambridge, 2018.

[43] Navigli R. Word sense disambiguation: A survey[J]. ACM computing surveys (CSUR), 2009, 41(2): 1-69.

[44] Snyder B, Palmer M. The english all-words task[C]//Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. 2004: 41-43.

[45] Hauer B, Kondrak G. Wic= tsv= wsd: On the equivalence of three semantic tasks[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022: 2478-2486.

[46] Hauer B, Kondrak G. Taxonomy of problems in lexical semantics[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 9833-9844.

[47] Noraset T, Liang C, Birnbaum L, et al. Definition modeling: Learning to define word embeddings in natural language[C]//AAAI. AAAI Press, 2017: 3259-3266.

[48] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.

[49] Lee H, Peirsman Y, Chang A, et al. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task[C]//Proceedings of the fifteenth conference on computational natural language learning: Shared task. 2011: 28-34.

[50] Ye B, Tu J, Pustejovsky J. Enhanced noun-noun compound interpretation through textual enrichment[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 25907-25922.

[51] Miletić F, im Walde S S. A systematic search for compound semantics in pretrained bert architectures[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023: 1499-1512.

[52] Buijtelaar L, Pezzelle S. A psycholinguistic analysis of bert's representations of compounds [C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023: 2230-2241.

[53] 邵敬敏. 2000–2018 年中国汉语虚词研究的回顾与前瞻[J]. 中国語学, 2018, 2018(265): 1-18.

[54] 李晓琪. 论对外汉语虚词教学[J]. 世界汉语教学, 1998(3): 65-71.

[55] Narayanan S. Moving right along: a computational model of metaphoric reasoning about events [C]//Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI). Orlando, Florida, 1999: 121-128.

[56] Kutuzov A, Øvrelid L, Szymanski T, et al. Diachronic word embeddings and semantic shifts: a survey[C]//Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018: 1384-1397.

[57] Lo C H, Lam W, Cheng H, et al. Distributional inclusion hypothesis and quantifications: Probing for hypernymy in functional distributional semantics[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 14625-14637.

[58] Fromkin V, Rodman R, Hyams N, et al. An introduction to language[M]. 10th ed. South Melbourne: Cengage Learning, 2021: 503.

[59] 范晓蕾. 普通话“了 1”“了 2”的语法异质性[M]. 北京: 北京大学出版社, 2021.

[60] Trott S, Bergen B. Raw-c: Relatedness of ambiguous words in context (a new lexical resource for english)[C]//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). 2021: 7077-7087.

[61] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[A]. 2021.

[62] Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 6382-6388.

[63] Raganato A, Pasini T, Camacho-Collados J, et al. Xl-wic: A multilingual benchmark for evaluating semantic contextualization[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7193-7206.

[64] Agirre E, Lopez de Lacalle O, Fellbaum C, et al. SemEval-2010 task 17: All-words word sense disambiguation on a specific domain[C/OL]//Erk K, Strapparava C. Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: Association for Computational Linguistics, 2010: 75-80. https://aclanthology.org/S10-1013/.

[65] Navigli R, Jurgens D, Vannella D. Semeval-2013 task 12: Multilingual word sense disambiguation[C]//Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). 2013: 222-231.

[66] Moro A, Navigli R. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking[C]//Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). 2015: 288-297.

[67] Navigli R, Ponzetto S P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network[J]. Artificial intelligence, 2012, 193: 217-250.

[68] Arslan T P, Erol E, Eryiğit G. Corefinst: Leveraging llms for multilingual coreference resolution [J]. Transactions of the Association for Computational Linguistics, 2026, 14: 64-80.

[69] Nedoluzhko A, Novák M, Popel M, et al. Corefud 1.0: Coreference meets universal dependencies[C]//Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association, 2022: 4859-4872.

[70] Skachkova N, Anikina T, Mokhova A. Multilingual coreference resolution: Adapt and generate[C]//Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023: 19-33.

[71] Arslan T P, Acar K, Eryiğit G. Neural end-to-end coreference resolution using morphological information[C]//Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023: 34-40.

[72] Straka M. Corpipe at crac 2024: Predicting zero mentions from raw text[C]//Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC). Miami: Association for Computational Linguistics, 2024: 97-106.

[73] Sanchez-Bayona E, Agerri R. Meta4xnli: A cross-lingual parallel corpus for metaphor detection and interpretation[J]. Computational Linguistics, 2026: 1-45.

[74] Wang Y, Liang Q, Yin Y, et al. Disambiguate words like composing them: A morphology-informed approach to enhance Chinese word sense disambiguation[C]//Ku L W, Martins A, Srikumar V. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024: 15354-15365.

[75] Hou B, Qi F, Zang Y, et al. Try to substitute: An unsupervised chinese word sense disambiguation method based on hownet[C]//Proceedings of the 28th international conference on computational linguistics. 2020: 1752-1757.

[76] Dong Z, Dong Q, Hao C. Hownet and its computation of meaning[C]//Coling 2010: Demonstrations. 2010: 53-56.

[77] Chenwei X, Ma M K H, Wang W, et al. Context and pos in action: A comparative study of chinese homonym disambiguation in human and language models[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 27596-27613.

[78] Ma M K H, Chenwei X, Wang W, et al. Exploring layer-wise representations of english and chinese homonymy in pre-trained language models[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 19705-19724.

[79] Li S, Yuan M, Dai X, et al. Evaluation of uncertainty estimation methods in medical image segmentation: Exploring the usage of uncertainty in clinical deployment[J]. Computerized Medical Imaging and Graphics, 2025, 124: 102574.

[80] Zhang B, Li Z, Gan Z, et al. CroAno : A crowd annotation platform for improving label consistency of Chinese NER dataset[C]//Adel H, Shi S. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 275-282.

[81] Huang Y, Ferreira F. The application of signal detection theory to acceptability judgments[J]. Frontiers in Psychology, 2020, 11: 73.

[82] Gimpel K, Batra D, Dyer C, et al. A systematic exploration of diversity in machine translation [C]//Yarowsky D, Baldwin T, Korhonen A, et al. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013: 1100-1111.

[83] Liu Z, Wang T, Zhang J, et al. Show, tell and rephrase: Diverse video captioning via two-stage progressive training[J]. IEEE Transactions on Multimedia, 2022, 25: 7894-7905.

[84] Park K, Yang N, Jung K. Avoidance decoding for diverse multi-branch story generation[C]// Christodoulopoulos C, Chakraborty T, Rose C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 7489-7505.

[85] Navigli R. A structural approach to the automatic adjudication of word sense disagreements[J]. Natural Language Engineering, 2008, 14(4): 547-573.

[86] Wan R, Kim J, Kang D. Everyone's voice matters: Quantifying annotation disagreement using demographic information[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 14523-14530.

[87] Gal Y. Uncertainty in deep learning[D/OL]. University of Cambridge, 2016. http://mlg.eng.cam.ac.uk/yarin/.

[88] Nahum O, Calderon N, Keller O, et al. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance[C]//Christodoulopoulos C, Chakraborty T, Rose C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 26782-26809.

[89] Bevilacqua M, Navigli R. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2854-2864.

[90] Conia S, Navigli R. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 3269-3275.

[91] Uma A N, Fornaciari T, Hovy D, et al. Learning from disagreement: A survey[J]. Journal of Artificial Intelligence Research, 2021, 72: 1385-1470.

[92] Fleisig E, Abebe R, Klein D. When the majority is wrong: Modeling annotator disagreement for subjective tasks[C]//Bouamor H, Pino J, Bali K. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 6715-6726.

[93] El-Yaniv R, et al. On the foundations of noise-free selective classification.[J]. Journal of Machine Learning Research, 2010, 11(5).

[94] Xin J, Tang R, Yu Y, et al. The art of abstention: Selective prediction and error regularization for natural language processing[C]//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). 2021: 1040-1051.

[95] Newell A, Simon H A. Computer science as empirical inquiry: Symbols and search[J]. Communications of the ACM, 1976, 19(3): 113-126.

[96] Crevier D. Ai: the tumultuous history of the search for artificial intelligence[M]. Basic Books, Inc., 1993.

[97] Church K, Mercer R L. Introduction to the special issue on computational linguistics using large corpora[J]. Computational linguistics, 1993, 19(1): 1-24.

[98] Joachims T. Text categorization with support vector machines: Learning with many relevant features[C]//European conference on machine learning. Springer, 1998: 137-142.

[99] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 2002, 86(11): 2278-2324.

[100] McCallum A. A comparison of event models for naive bayes text classification[Z]. 1998.

[101] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[102] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[103] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29 (6): 82-97.

[104] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//NIPS. 2012: 1106-1114.

[105] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[106] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019.

[107] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[A]. 2020. arXiv: 2001.08361.

[108] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26.

[109] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation[R]. 1985.

[110] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Comput., 1997, 9(8): 1735-1780.

[111] Joshi M, Levy O, Zettlemoyer L, et al. BERT for coreference resolution: Baselines and analysis [C]//Inui K, Jiang J, Ng V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 5803-5808.

[112] Luoma J, Pyysalo S. Exploring cross-sentence contexts for named entity recognition with bert [C]//Proceedings of the 28th international conference on computational linguistics. 2020: 904-914.

[113] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[M]. OpenAI, 2018.

[114] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). 2017: 1073-1083.

[115] Zhang Y, Sun S, Galley M, et al. Dialogpt: Large-scale generative pre-training for conversational response generation[J]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020: 270-278.

[116] Fan A, Lewis M, Dauphin Y. Hierarchical neural story generation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). 2018: 889-898.

[117] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//International Conference on Learning Representations (ICLR). 2015.

[118] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research (JMLR), 2020, 21(140): 1-67.

[119] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020: 7871-7880.

[120] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Isabelle P, Charniak E, Lin D. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002: 311-318.

[121] Qin Y, Song K, Hu Y, et al. InFoBench: Evaluating instruction following ability in large language models[C]//Ku L W, Martins A, Srikumar V. Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024: 13025-13048.

[122] Bang Y, Ji Z, Schelten A, et al. HalluLens: LLM hallucination benchmark[C]//Che W, Nabende J, Shutova E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 24128-24156.

[123] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[C]//Advances in Neural Information Processing Systems: Vol. 33. 2020: 9459-9474.

[124] Wang A, Pruksachatkun Y, Nangia N, et al. Superglue: A stickier benchmark for general-purpose language understanding systems[J]. Advances in neural information processing systems, 2019, 32.

[125] Srivastava A, Rastogi A, Rao A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[A]. 2022.

[126] Kazemi M, Fatemi B, Bansal H, et al. BIG-bench extra hard[C]//Che W, Nabende J, Shutova E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 26473-26501.

[127] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding [J]. Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[128] Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code[A]. 2021.

[129] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems[C]// Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 2021-2031.

[130] Ribeiro M T, Wu T, Guestrin C, et al. Beyond accuracy: Behavioral testing of nlp models with checklist[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 4902-4912.

[131] Lunardi R, Della Mea V, Mizzaro S, et al. On robustness and reliability of benchmark-based evaluation of llms[M]//ECAI 2025. IOS Press, 2025: 4603-4610.

[132] Nalbandyan G, Shahbazyan R, Bakhturina E. Score: Systematic consistency and robustness evaluation for large language models[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track). 2025: 470-484.

[133] Gehman S, Gururangan S, Sap M, et al. Realtoxicityprompts: Evaluating neural toxic degeneration in language models[C]//Findings of the association for computational linguistics: EMNLP 2020. 2020: 3356-3369.

[134] Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models[A]. 2022.

[135] Zhang Z, Lei L, Wu L, et al. Safetybench: Evaluating the safety of large language models [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 15537-15553.

[136] Abdelatty M, Nouh M, Rosenstein J K, et al. Pluto: A benchmark for evaluating efficiency of llm-generated hardware code[A]. 2025.

[137] Stutz D. Understanding and improving robustness and uncertainty estimation in deep learning [J]. Saarländische Universitäts-und Landesbibliothek, 2022.

[138] Shelmanov A, Panov M, Vashurin R, et al. Uncertainty quantification for large language models [C]//Arase Y, Jurgens D, Xia F. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts). Vienna, Austria: Association for Computational Linguistics, 2025: 3-4.

[139] Liu Z, Hu Z, Liu Y. JuniperLiu at CoMeDi shared task: Models as annotators in lexical semantics disagreements[C]//Roth M, Schlechtweg D. Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation. Abu Dhabi, UAE: International Committee on Computational Linguistics, 2025: 103-112.

[140] Bevilacqua M, Pasini T, Raganato A, et al. Recent trends in word sense disambiguation: A survey[C]//International Joint Conference on Artificial Intelligence. International Joint Conference on Artificial Intelligence, Inc, 2021: 4330-4338.

[141] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[A]. 2019.

[142] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale[A/OL]. 2019. https://arxiv.org/abs/1911.02116.

[143] Meconi D, Stirpe S, Martelli F, et al. Do large language models understand word senses?[C]// Christodoulopoulos C, Chakraborty T, Rose C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 33897-33916.

[144] Jiang T, Huang S, Luan Z, et al. Scaling sentence embeddings with large language models[C]// Al-Onaizan Y, Bansal M, Chen Y N. Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA: Association for Computational Linguistics, 2024: 3182-3196.

[145] Zhang B, Chang K, Li C. Simple techniques for enhancing sentence embeddings in generative language models[C]//International Conference on Intelligent Computing. Springer, 2024: 52-64.

[146] Cheng Z, Wang Z, Fu Y, et al. Contrastive prompting enhances sentence embeddings in llms through inference-time steering[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 3475-3487.

[147] Zhu J, Wang H, Yao T, et al. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). 2008: 1137-1144.

[148] Kuklin M, Arefyev N. Deep-change at comedi: the cross-entropy loss is not all you need[C]// Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations. Abu Dhabi, UAE, 2025.

[149] Alfter D, Appelgren M. Grasp at comedi shared task: Multi-strategy modeling of annotator behavior in multi-lingual semantic judgments[C]//Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations. Abu Dhabi, UAE, 2025.

[150] Olsson C, Elhage N, Nanda N, et al. In-context learning and induction heads[A]. 2022.

[151] Geiger A, Lu H, Icard T, et al. Causal abstractions of neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 9574-9586.

[152] Wang K R, Variengien A, Conmy A, et al. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small[C]//NeurIPS ML Safety Workshop. 2022.

[153] Faruqui M, Tsvetkov Y, Yogatama D, et al. Sparse overcomplete word vector representations [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1491-1500.

[154] Liu Z, Kong C, Liu Y, et al. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics[C]//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024: 14551-14558.

[155] Rogers A, Kovaleva O, Rumshisky A. A primer in bertology: What we know about how bert works[J]. Transactions of the Association for Computational Linguistics, 2021, 8: 842-866.

[156] Ettinger A. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 34-48.

[157] Vulić I, Ponti E M, Litschko R, et al. Probing pretrained language models for lexical semantics [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7222-7240.

[158] Cífka O, Liutkus A. Black-box language model explanation by context length probing[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023: 1067-1079.

[159] Hewitt J, Manning C D. A structural probe for finding syntax in word representations[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4129-4138.

[160] Wang S, Sun X, Li X, et al. Gpt-ner: Named entity recognition via large language models[A]. 2023.

[161] Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases?[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 2463-2473.

[162] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?[C/OL]// Korhonen A, Traum D, Màrquez L. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 3651-3657. https://aclanthology.org/P19-1356. DOI: 10.18653/v1/P19-1356.

[163] Garí Soler A, Apidianaki M. Scalar adjective identification and multilingual ranking[C/OL]// Toutanova K, Rumshisky A, Zettlemoyer L, et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 4653-4660. https://aclanthology.org/2021.naacl-main.370. DOI: 10.18653/v1/2021.naacl-main.370.

[164] Lyu Q, Apidianaki M, Callison-burch C. Representation of lexical stylistic features in language models' embedding space[C/OL]//Palmer A, Camacho-collados J. Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). Toronto, Canada: Association for Computational Linguistics, 2023: 370-387. https://aclanthology.org/2023.starsem-1.32. DOI: 10.18653/v1/2023.starsem-1.32.

[165] Yamagiwa H, Oyama M, Shimodaira H. Discovering universal geometry in embeddings with ICA[C/OL]//Bouamor H, Pino J, Bali K. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 4647-4675. https://aclanthology.org/2023.emnlp-main.283. DOI: 10.18653/v1/2023.emnlp-main.283.

[166] Wang L, Li L, Dai D, et al. Label words are anchors: An information flow perspective for understanding in-context learning[C]//Bouamor H, Pino J, Bali K. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023.

[167] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle[C]//2015 ieee information theory workshop (itw). Ieee, 2015: 1-5.

[168] Voita E, Sennrich R, Titov I. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives[C/OL]//Inui K, Jiang J, Ng V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 4396-4406. https://aclanthology.org/D19-1448. DOI: 10.18653/v1/D19-1448.

[169] Cruse D A. Lexical semantics[M]. Cambridge university press, 1986.

[170] 张志毅, 张庆云. 词汇语义学[M]. 北京: 商务印书馆, 2012.

[171] Schuler K K. Verbnet: A broad-coverage, comprehensive verb lexicon[D]. University of Pennsylvania, 2005.

[172] Blank A. Prinzipien des lexikalischen bedeutungswandels am beispiel der romanischen sprachen: Vol. 285[M]. Walter de Gruyter, 2012.

[173] Chen J, Chersoni E, Schlechtweg D, et al. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection[C/OL]//Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change. Singapore: Association for Computational Linguistics, 2023. https://aclanthology.org/2023.lchange-1.10/.

[174] Schlechtweg D, Tahmasebi N, Hengchen S, et al. DWUG: A large resource of diachronic word usage graphs in four languages[C/OL]//Moens M F, Huang X, Specia L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 7079-7091. https://aclanthology.org/2021.emnlp-main.567. DOI: 10.18653/v1/2021.emnlp-main.567.

[175] Schlechtweg D, Cassotti P, Noble B, et al. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida: Association for Computational Linguistics, 2024.

[176] Schlechtweg D, Schulte im Walde S, Eckmann S. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change[C/OL]//Walker M, Ji H, Stent A. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 169-174. https://aclanthology.org/N18-2027. DOI: 10.18653/v1/N18-2027.

[177] Hätty A, Schlechtweg D, Schulte im Walde S. SURel: A gold standard for incorporating meaning shifts into term extraction[C/OL]//Mihalcea R, Shutova E, Ku L W, et al. Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 1-8. https://aclanthology.org/S19-1001. DOI: 10.18653/v1/S19-1001.

[178] Kurtyigit S, Park M, Schlechtweg D, et al. Lexical Semantic Change Discovery[C/OL]//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). Online: Association for Computational Linguistics, 2021. https://aclanthology.org/2021.acl-long.543/.

[179] Schlechtweg D. Human and computational measurement of lexical semantic change[D/OL]. University of Stuttgart, Germany, 2023. https://d-nb.info/1285256050.

[180] Kutuzov A, Touileb S, Mæhlum P, et al. NorDiaChange: Diachronic semantic change dataset for Norwegian[C/OL]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2022: 2563-2572. https://aclanthology.org/2022.lrec-1.274.

[181] Rodina J, Kutuzov A. RuSemShift: a dataset of historical lexical semantic change in Russian [C/OL]//Scott D, Bel N, Zong C. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020: 1037-1047. https://aclanthology.org/2020.coling-main.90. DOI: 10.18653/v1/2020.coling-main.90.

[182] Kutuzov A, Pivovarova L. Rushifteval: a shared task on semantic shift detection for russian[J]. Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference, 2021.

[183] Aksenova A, Gavrishina E, Rykov E, et al. RuDSI: Graph-based word sense induction dataset for Russian[C/OL]//Ustalov D, Gao Y, Panchenko A, et al. Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing. Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022: 77-88. https://aclanthology.org/2022.textgraphs-1.9.

[184] Zamora-Reina F D, Bravo-Marquez F, Schlechtweg D. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish[C/OL]//Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change. Dublin, Ireland: Association for Computational Linguistics, 2022. https://aclanthology.org/2022.lchange-1.16/.

[185] Krippendorff K. Content analysis: An introduction to its methodology[M]. SAGE Publications, 2018.

[186] Spearman C. The proof and measurement of association between two things[J]. The American Journal of Psychology, 1904, 15(1): 72-101.

[187] 王仁强. 认知视角的汉英词典词类标注实证研究[D]. 广州: 广东外语外贸大学, 2006.

[188] Gentile A L, Zhang Z, Xia L, et al. Semantic relatedness approach for named entity disambiguation[C]//Italian Research Conference on Digital Libraries. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010: 137-148.

[189] Gonzales A R, Mascarell L, Sennrich R. Improving word sense disambiguation in neural machine translation with sense embeddings[C]//Proceedings of the Second Conference on Machine Translation. 2017: 11-19.

[190] Springer J M, Kotha S, Fried D, et al. Repetition improves language model embeddings[C]// The Thirteenth International Conference on Learning Representations.

[191] Thirukovalluru R, Dhingra B. Geneol: Harnessing the generative power of llms for training-free sentence embeddings[C]//Findings of the Association for Computational Linguistics: NAACL 2025. Association for Computational Linguistics, 2025: 2295-2308.

[192] Reif E, Yuan A, Wattenberg M, et al. Visualizing and measuring the geometry of bert[C]// Advances in Neural Information Processing Systems: Vol. 32. 2019.

[193] Nelder J A, Mead R. A simplex method for function minimization[J]. The Computer Journal, 1965, 7(4): 308-313.

[194] Ethayarajh K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 55-65.

[195] Mu J, Viswanath P. All-but-the-top: Simple and effective postprocessing for word representations[C]//International Conference on Learning Representations. 2018.

[196] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale[C/OL]//Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 8440-8451. https://aclanthology.org/2020.acl-main.747/. DOI: 10.18653/v1/2020.acl-main.747.

[197] Cassotti P, Siciliani L, DeGemmis M, et al. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2023: 1577-1585.

[198] Martelli F, Kalach N, Tola G, et al. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic)[C]//Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). 2021: 24-36.

[199] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3982-3992.

[200] Liu Y, Gu J, Goyal N, et al. Multilingual denoising pre-training for neural machine translation [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.

[201] Xue L, Constant N, Roberts A, et al. mt5: A massively multilingual pre-trained text-to-text transformer[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 483-498.

[202] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models [A]. 2023.

[203] DeepSeek-AI. Deepseek-v3 technical report[J]. CoRR, 2024, abs/2412.19437.

[204] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[A]. 2023.

[205] Chen Z, Wu K. ForceReader: a BERT-based interactive machine reading comprehension model with attention separation[C/OL]//Scott D, Bel N, Zong C. Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020: 2676-2686. https://aclanthology.org/2020.coling-main.241/. DOI: 10.18653/v1/2020.coling-main.241.

[206] Gessler L, Schneider N. BERT has uncommon sense: Similarity ranking for word sense BERTology[C/OL]//Bastings J, Belinkov Y, Dupoux E, et al. Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 539-547. https://aclanthology.org/2021.blackboxnlp-1.43/. DOI: 10.18653/v1/2021.blackboxnlp-1.43.

[207] Askari H, Chhabra A, Chen M, et al. Assessing LLMs for zero-shot abstractive summarization through the lens of relevance paraphrasing[C/OL]//Chiruzzo L, Ritter A, Wang L. Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, New Mexico: Association for Computational Linguistics, 2025: 2187-2201. https://aclanthology.org/2025.findings-naacl.116/. DOI: 10.18653/v1/2025.findings-naacl.116.

[208] Wang Y, Kreminski M. Can llms generate good stories? insights and challenges from a narrative planning perspective[C]//2025 IEEE Conference on Games (CoG). IEEE, 2025: 1-8.

[209] Loureiro D, Jorge A. Liaad at semdeep-5 challenge: Word-in-context (wic)[C]//Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5). 2019: 1-5.

[210] Melamud O, Goldberger J, Dagan I. context2vec: Learning generic context embedding with bidirectional LSTM[C/OL]//Riezler S, Goldberg Y. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany: Association for Computational Linguistics, 2016: 51-61. https://aclanthology.org/K16-1006. DOI: 10.18653/v1/K16-1006.

[211] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 2227-2237. https://aclanthology.org/N18-1202. DOI: 10.18653/v1/N18-1202.

[212] 朱德熙. 语法讲义[M]. 北京: 商务印书馆, 1982.

[213] Mahajan Y, Freestone M, Bansal N, et al. Revisiting word embeddings in the llm era[C]//Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2025: 2686-2717.

[214] Barba E, Procopio L, Navigli R. Consec: Word sense disambiguation as continuous sense comprehension[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 1492-1503.

[215] Heilbron M, Armeni K, Schoffelen J M, et al. A hierarchy of linguistic predictions during natural language comprehension[J/OL]. Proceedings of the National Academy of Sciences, 2022, 119 (32): e2201968119. https://www.pnas.org/doi/abs/10.1073/pnas.2201968119.

[216] Willis G B. Cognitive interviewing in practice: Think-aloud, verbal probing, and other techniques[M]//Cognitive Interviewing: A Tool for Improving Questionnaire Design. London: Sage Publications, 2005: 42-63.

[217] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[218] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.

[219] Loshchilov I, Hutter F. Decoupled weight decay regularization[C/OL]//International Conference on Learning Representations (ICLR). 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

[220] 赵元任. 汉语的歧义问题[M]. 石安石, 译. 北京: 商务印书馆, 1959.

[221] 祝青翠. 浅谈现代汉语歧义句[EB/OL]. (2021-11-02). https://www.sinoss.net/c/2021-11-02/567215.shtml.

[222] 蒋绍愚, 等. 汉语词义和词汇系统的历史演变初探——以“投”为例[J]. 北京大学学报 (哲学社会科学版), 2006(04): 85-106.

[223] Bereska L F. Mechanistic interpretability for ai safety—a review[C]//Proceedings of The 1st Conference on Lifelong Learning Agents. 2022.

[224] Jing Y, Yao Z, Guo H, et al. Lingualens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 28220-28239.

附录 A　中文兼类词数据集标注问卷展示

图 A.1 展示了调查问卷的个人信息和部分题目示例，用户需通过单项选择的方式来判断目标句对之间的语义相关程度。每份问卷中包含 40 个问题，平均作答时间约为 20 分钟。

(a) 调查问卷个人信息收集

(b) 调查问卷部分题目示例

图 A.1　调查问卷示例

致　谢

写到这一部分，博士毕业论文即将告一段落，博士生阶段也要接近尾声了。千言万语，唯有感谢。首先，衷心感谢我的导师刘颖老师，刘老师为我提供了一个宽松、自由的科研氛围，使我能够自主探索自己感兴趣的研究方向。沙滩拾贝，从容不迫而怡然自得。刘老师沉稳而真诚的态度，也在潜移默化中影响着我。论文写作过程中，刘老师在每个关键节点都悉心指导，并且始终提醒我关注论文的系统性、完整性与连贯性，切勿将所做的事情不加整理得堆砌在一起，没有老师的多次反馈，这篇文章也不会如此呈现出来。感谢我的合作导师孙茂松老师，一年多在孙老师组里访学的经历让我受益良多。孙老师对科研的热爱和踏实、严谨、一丝不苟的治学态度令我非常钦佩。论文中的一些研究成果，离不开孙老师及组内同学的讨论与指导。感谢在罗马一大访学期间指导我科研的 Roberto Navigli 教授，他对科研负责认真，每周与他的讨论都让我收获颇丰。

论文背后需要日积跬步的积累。感谢引导我入门语言学的所有清华和北大中文系及外文系老师，包括邱冰老师、邓盾老师、张赪老师、江铭虎老师、刘明明老师、郭锐老师、董秀芳老师等。尤其是邓盾老师严谨的治学态度、巧妙的课堂设计，以及郭锐老师课堂上的博学与风趣，为我的科研带来了深刻启发。正是各位老师在科研和教学上的风度，让我感受到了一流高校的底蕴，也为我今后从事学术工作树立了高标杆。感谢艺教中心的美育老师：康啸老师、丁毅老师、周梦梅老师，“以美启真”，艺术给予了我整个求学阶段最深厚的情感支撑。

科研旅途幸有同伴而行。感谢博士阶段一起并肩科研的所有同学和朋友们。在计算语言学课题组里，有张声龙、宋丽、朱述承、李志、徐会丹、胡震、刘宇轩、李嘉木等。在计算机系 THUNLP 组里，有孔存良、罗康洋、李文浩、程志立、白钰卓等；在罗马一大自然语言处理实验室里，有徐路、Simone Conia、Tommaso Bonomo、Stefano Perrella、Giuliano Martinelli、Francesco Maria Molfese 等。还特别感谢一起讨论分布式语义的徐会丹、胡震、刘宇轩等同门，感谢讨论《语法讲义》的戴蕾同学，感谢告诉并和我一起去北大听课的王鹏远同学，也感谢讨论可解释性问题的孔存良师兄。感谢在罗马一大交换期间对我生活提供帮助的徐路同学。

感谢我的家人，是他们让我能够安心地在“象牙塔”里专注科研，无需顾虑其他。感谢所有朋友，包括挚友刘雨瑞、硕士期间的同门、排球社团和禅文化研习社的朋友们，他们在生活中给予我无尽帮助，也让我的生活增添了许多色彩。感谢我的室友王振辉和朋友唐庆宏，与他们的频繁相聚让我学习到许多宝贵的品质。

“西山苍苍，东海茫茫”，感谢这座庄严厚重的百年学府。历史虽无言，但往昔的记忆已经融入荷塘月色与巍巍礼堂，凝成“中兴业，须人杰”和“独立之精神、自由之思想”的呐喊，化为跨越数代人的捐赠箴言和姓名，浸染于一草一木、一土一石，最终春风化雨，润物无声。深厚的故往和底蕴便令学子天资愚钝但欣于闻道，偶有懈怠仍常求日新。

最后，特别感谢论文中问卷调查的所有参与者，并且衷心感谢在百忙之中评阅论文并参加答辩的各位专家和教授。本课题承蒙国家社会科学基金、教育部资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果，不包含涉及国家秘密的内容。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间完成的相关学术成果

个人简历

1996 年 10 月 21 日出生于山西省灵丘县。

2015 年 9 月考入大连理工大学软件学院数字媒体技术专业，2019 年 7 月本科毕业并获得工学学士学位。

2019 年 9 月考入南方科技大学计算机学院电子科学与技术专业，2022 年 7 月硕士毕业并获得工学硕士学位。

2022 年 9 月考入清华大学中文系攻读中国语言与文学博士至今。

在学期间完成的相关学术成果

学术论文：

[1] **Liu Z**, Hu Z, Liu Y. JuniperLiu at CoMeDi Shared Task: Models as Annotators in Lexical Semantics Disagreements[C]//Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation. 2025: 103-112.（论文第 4 章和第 6 章）

[2] **Liu Z**, Liu Y. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 3963-3977. （论文第 2 章）

[3] **Liu Z**, Kong C, Liu Y, et al. A Top-down Graph-based Tool for Modeling Classical Semantic Maps: A Case Study of Supplementary Adverbs[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 4567-4576. （论文第 2 章）

[4] **Liu Z**, Kong C, Liu Y, et al. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics[C]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 14551-14558. （论文第 5 章）

[5] **Liu Z**, Wang T, Zhang J, et al. Show, tell and rephrase: Diverse video captioning via two-stage progressive training[J]. IEEE Transactions on Multimedia, 2022, 25: 7894-7905.

[6] Luo K, Bai Y, Gao C, et al. Gltw: joint improved graph transformer and llm via three-word language for knowledge graph completion[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 11328-11344.

指导教师评语

论文以多语言词义相关度任务为研究切入点，聚焦大语言模型词义理解建模评估的前沿问题，选题紧扣自然语言处理与人工智能交叉领域研究热点，具有重要的理论价值与实际应用意义。论文研究框架完整严谨，从模型评估、可解释性分析、标注不一致建模三个核心维度展开系统性探究，逻辑层次清晰，研究脉络完整。研究工作扎实深入，自主收集并拓展多语言数据集，专门构建汉语兼类词词义相关度数据集，设计几何探测、差异化提示语等多元评估方案，覆盖编码器与解码器主流架构的模型，实验设计全面规范。论文通过逐层探测剖析模型内部表征差异，揭示了多种大语言模型与预训练模型词义理解的层次化模式差异，研究结论可靠。

论文文献综述扎实，论证充分，行文规范流畅，体现出作者扎实的专业理论功底、很强的独立科研能力与严谨的学术思维。研究工作量饱满、创新性突出，是一篇优秀的博士学位论文。

答辩委员会决议书

论文以大语言模型多语言场景下词义相关度评估与可解释性分析为研究主题，选题契合自然语言处理学术前沿，具有重要的理论意义与应用前景。

论文主要创新点如下：

（1）论文研究框架完整，遵循评测分析、机制挖掘、标注不一致建模的逻辑脉络，在现有公开数据集基础上，自主构建汉语兼类词词义相关度数据集及多标注者评测资源，完善汉语特有语言现象的数据支撑，具备良好的学术价值。

（2）研究方法科学丰富，融合几何探测、表征分析、提示语引导等多种手段，系统开展生成式大模型与编码器模型的对比实验，论证充分，数据丰富。

（3）依托逐层探测开展可解释性研究，归纳出大语言模型的逐层语义变化规律，揭示了词类信息对模型在中英两种语言的影响差异，具有创新性。

论文结构合理，研究数据详实，论证过程充分，行文规范流畅，学术规范严谨，是一篇优秀的博士学位论文。论文工作表明，该生掌握了本学科领域坚实宽广的基础理论和系统深入的专门知识，独立开展科研工作的能力强。

答辩过程中，表述清楚，回答问题正确。经答辩委员会无记名投票，一致同意该生通过博士学位论文答辩，并建议授予文学博士学位。