大语言模型的多语言词义相关度评估与可解释性分析

答辩人：刘柱

中国语言文学四年级博士研究生

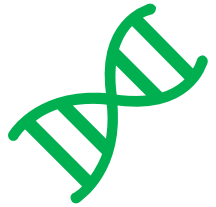
导师：刘颖教授

时间：2026.05.21

大纲

1. 研究背景和意义
2. 研究现状
3. 多语言词义相关度任务评估
4. 多语言词义相关度任务中的模型可解释性研究
5. 多语言词义相关度任务中的标注不一致研究
6. 总结与展望

评估



可解释性

1. 研究背景和意义

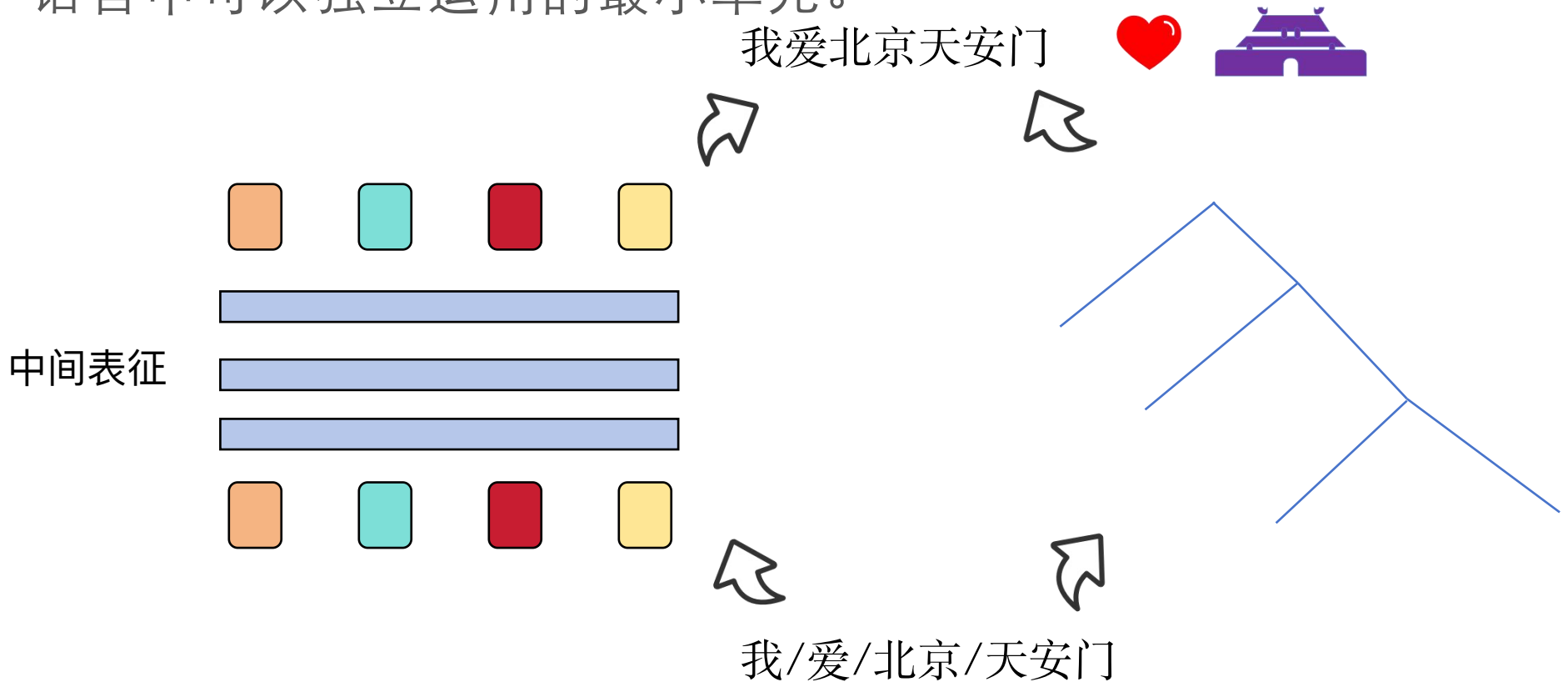
第1章

研究背景

- 以大语言模型为代表的生成式人工智能在各类自然语言任务上表现突出。
- 仍面临着诸多挑战：
  - 数据驱动的模型内部机制不明确；
  - 安全隐患：易产生包含错误或误导信息的“幻觉”现象；
  - 效率低下：更多靠“大数据、多参数、强算力”范式进行训练
  - 评估任务和数据集没有考虑到现实数据的特点和语言特性
- 仍需对大语言模型进行多维度、细粒度、全方位的评估与理解

词义理解评估

- 基础：单个的词元（token）是大语言模型的最小计算单位，词也是自然语言中可以独立运用的最小单元。



词义理解任务的选择

- 词义相关度任务
  - 通常又称为WiC（上下文中的词，word in context）
  - 判断同一个词出现在成对上下文中的相关度（或者语义是否一致）
- 词义消歧任务
  - 通常称为WSD（word sense disambiguation）
  - 选择一个词在上下文中的目标词义
- 其他相关任务：目标词词义确认任务、释义建模任务、命名实体识别任务、指代消解任务等

词义理解任务的选择

无需提供候选语义
更符合“最小对立对”原则
与模型广泛使用的语义对齐的方式更吻合
可以对模型内部的表征直接评估

- 词义相关度任务
  - 通常又称为WiC（上下文中的词，word in context）
  - 判断同一个词出现在成对上下文中的相关度（或者语义是否一致）

- 词义消歧任务
  - 通常称为WSD（word sense disambiguation）
  - 选择一个词在上下文中的目标词义
- 其他相关任务：目标词词义确认任务、释义建模任务、命名实体识别任务、指代消解任务等

词义理解评估任务需要从多维度进行考虑

- 多语言视角
- 多（词义）粒度视角
- 多标注者视角
- 模型内部机制视角

多语言视角

- 不同语言系统表示概念和语义所选用的词不同，且不完全一对一对应。
- 不同语言在构词规则、词法与句法的互动上都有所差异。

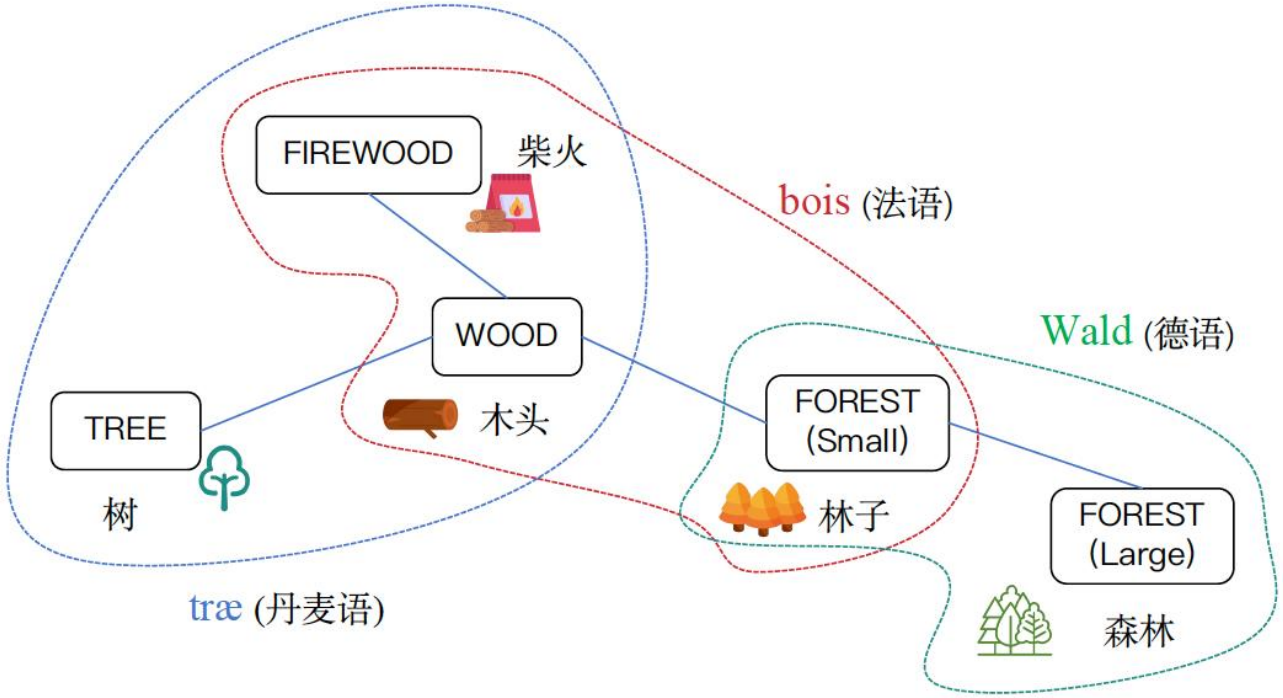


图 1.1　概念空间和语义地图示例

多粒度视角

- 词义间的区分粒度
  - 多义词的义项数量因词不同，有些词义的义项区分粒度较小
  - 对于词义区分较小的词，单独的“二分判断”目标词是否语义一致已经无法满足
  - WiC词义任务 -> 分级WiC词义任务
  - 粒度的区分度不同表明词义选择不是一个“匀质”的分类问题，而是一个“不均匀”的连续统

多粒度视角

- 词义间的区分粒度
  - 多义词的义项数量因词不同，
  - 对于词义区分较小的词，单独
  - WiC词义任务 -> 分级WiC词义

(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

法满足

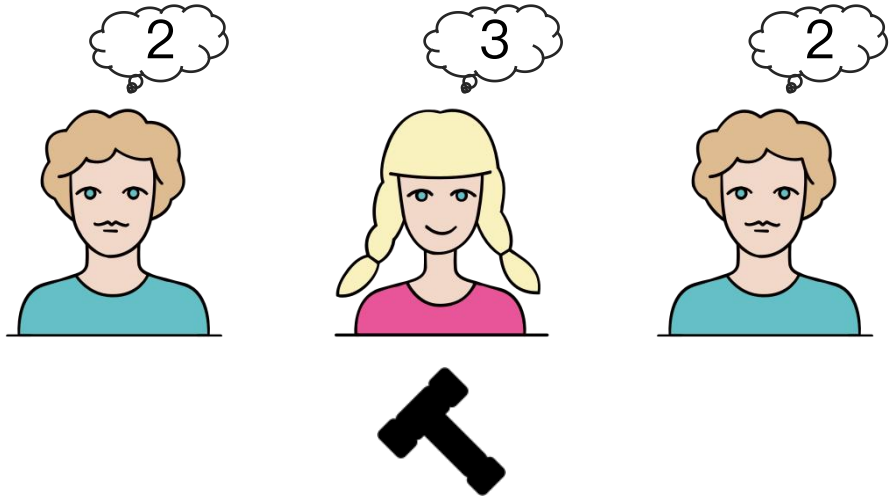
表 3.3 不同词义相关度标注类型对比

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)

多标注者视角

- 对于词义理解任务来说，往往单一的标注者无法确定最终的判断结果，因此需要融合多位标注者的结果。
- 标注不一致体现了真实世界数据的“不完美性”、“不确定性”

1. Pat **broke** his promise.
2. He **broke** the world record.



模型内部机制视角

- Transformer模型架构的演变
  - 从传统的编码器模型（以BERT为代表）过渡到解码器模型（以GPT、当代LLM为代表）
  - 编码器模型：对掩码的token进行恢复（MLM，masked language modeling）
  - 解码器模型：预测下一个token（NTP，next token prediction）
  - 编码模型对于词义的编码往往处于高层，而解码模型的训练目标和对当前token的词义编码不匹配，与编码模型对词义的编码方式不同。

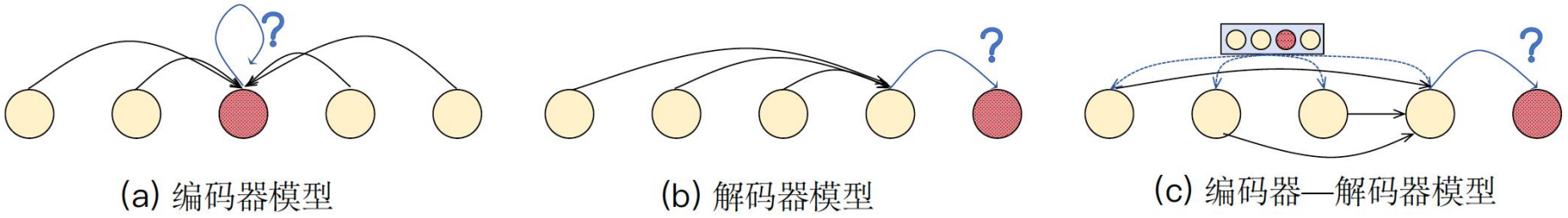


图 2.1　三类 Transformer 模型架构示意图

研究内容

- 词义相关度任务的全流程评估框架设计
- 多语言视角下数据集的拓展
- 多标注者视角下标注不一致程度评估
- 模型之间的跨层信息流动模式分析

本文组织架构

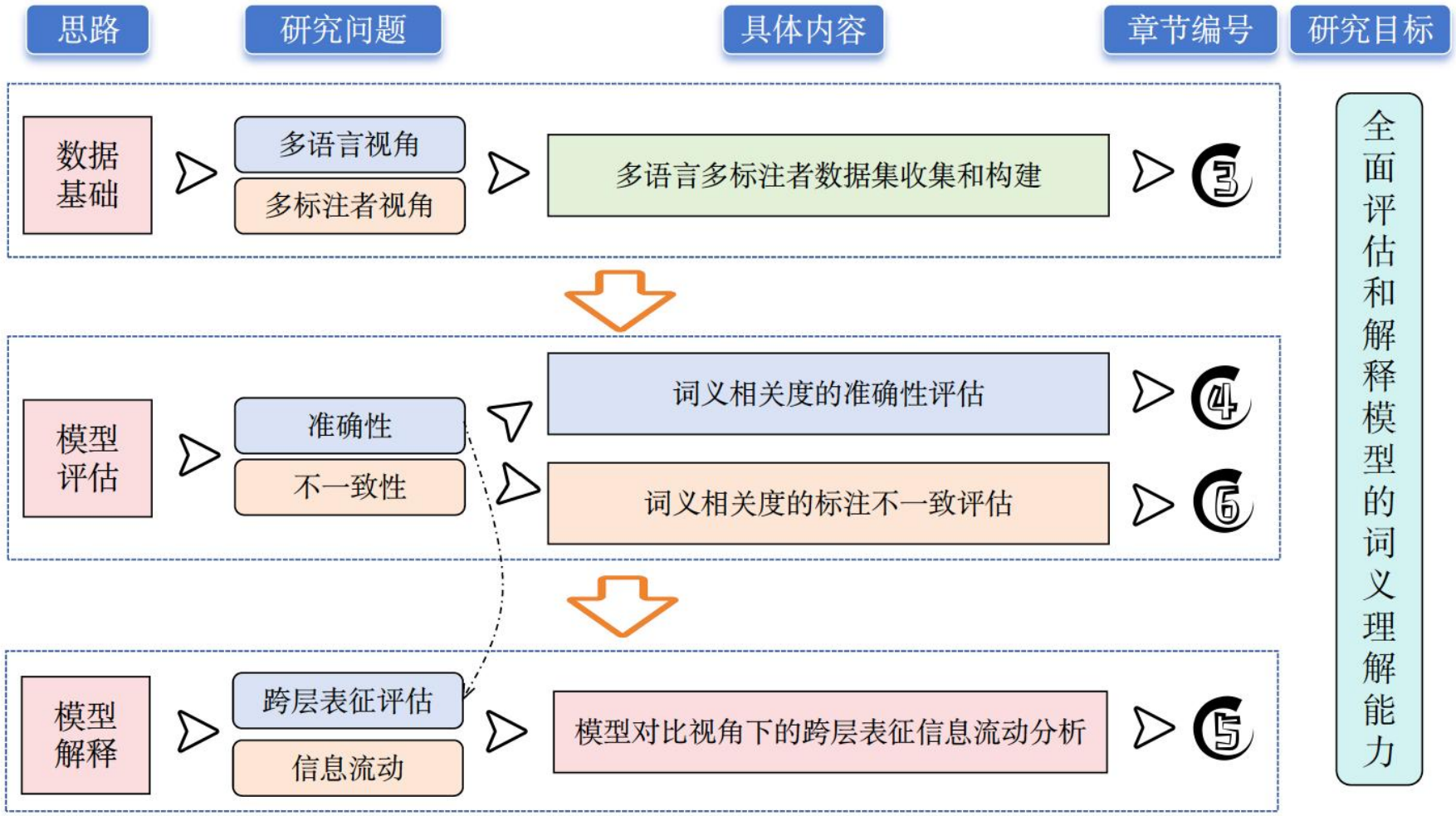


图 1.3　研究汇总和文章的组织结构

2. 研究现状

第2章

研究现状和相关工作

- 不同的词义理解任务类型

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	-	✓	✗

研究现状和相关工作

- 多语言视角下的词义相关度工作
  - WiC [X]: 经典的二分词义相关度任务的提出
  - XL-WiC [63]：WiC的多语言版本
  - OGWiC [37]: 依序分级的上下文词义相关度分类数据集（多语言）
  - 多语言WSD系列工作[90-91]
  - 针对汉语而言：
    - WSD工作有针对汉语的设计：融合了语素义消歧的MiCLS[74]、考虑汉语构词法的FiCLS[30]、针对HowNet的数据集[75]
    - 词义相关度任务WiC中，往往仅仅包括多语言基准的汉语部分（例如[37]），缺乏专门针对汉语特点设计的词义相关度数据集。

研究现状和相关工作

- 多标注视角下的词义相关度任务
  - 多标注者标注在人工智能领域任务中普遍存在
    - 消除不一致：例如医疗图像分割[79]，词类标注[80]，语句可接受度[81]
    - 鼓励不一致：机器翻译[82]、图像字幕生成[83]、故事生成[84]
    - 产生原因
      - 与数据本身有关：多义性[85]、上下文背景信息的充分性[29]、不同的短语结构[15]
      - 与标注者自身相关：标注者自身对于任务的理解[86]
  - 不确定性估计
    - 数据不确定性 vs. 模型不确定性
  - 后续不一致结果的使用
    - 丢弃[88]、池化（例如取平均）[89]、多标签学习[90]
  - 在词义任务上多标注者视角的工作仍较为缺乏

研究现状和相关工作

- 语言模型的发展
  - 理性主义 （符号模型） -》经验主义（统计模型） -》深度学习（神经网络模型-》大语言模型）
  - Transformer架构：解码器模型、编码器模型、编码-解码模型
  - 理解型任务 vs. 生成性任务
- 大语言模型的评估和可解释性研究
  - 评估维度：准确性、鲁棒性、安全性、效率、不确定性等
  - 词义相关度任务的评估：WSD/WiC任务的准确性[43]和不确定性[29]评估
  - 可解释性方面的研究[14]：
    - 机制可解释性: 干预手段、稀疏自编码器等方式
    - 表征可解释性: 分类器探测（词义[157]、句法结构[159]、语义角色[156]等）和几何探测（形容词的强度变化[163]、文体和风格特征[164]）

3. 多语言词义相关度任务评估

论文第4章

任务介绍

- 词义相关度任务旨在评估同一目标词在不同上下文中词义的关联程度。
  - 与基础自然语言学任务息息相关：词义消歧、命名实体识别
  - 关乎各类下游任务：文本检索、机器翻译、推荐、问答系统等
  - 揭示模型内部表征的几何结构、支持语义聚类与表示学习等

多语言词义相关度数据集

- OGWiC：依序分级的上下文词义相关度分类任务（Ordinal Graded Word-in-Context Classification）
- 作为共享任务发布在COLING会议组织的工作坊CoMeDi中
- 数据格式
  - 将传统WiC数据集的二值相关判断拓展到了四分类相关度判断。

标注说明

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)

(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

b.　介词，从

c.　动词，殴打；攻打

d.　动词，发生与人交涉的行为

e.　动词，发生与人交涉的行为

多语言词义相关度数据集

- OGWiC：依序分级的上下文词义相关度分类任务（Ordinal Graded Word-in-Context Classification）
- 作为共享任务发布在COLING会议组织的工作坊CoMeDi中
- 数据格式
  - 将传统WiC数据集的二值相关判断拓展到了四分类相关度判断。
  - 多语言视角：涉及7种语言（汉语、英语、德语、挪威语、俄语、西班牙语和瑞典语）
  - 多标注者视角：单条数据由2-5名不同的标注者进行标注。

数据集概览

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

数据集概览

- 汉语涉及的词数较少
- 除英语外，其他语言均没有标注词类信息。
- 不同语言的句长不同，西班牙语的句子长度最长，俄语句子长度最短，但仍属于较长的句子
- 目标词所处的位置基本都处于句子中部
- 汉语的标注人数最少，平均仅有2人，俄语的标注人数最多
- 训练/开发/测试集分布相近

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

数据集概览

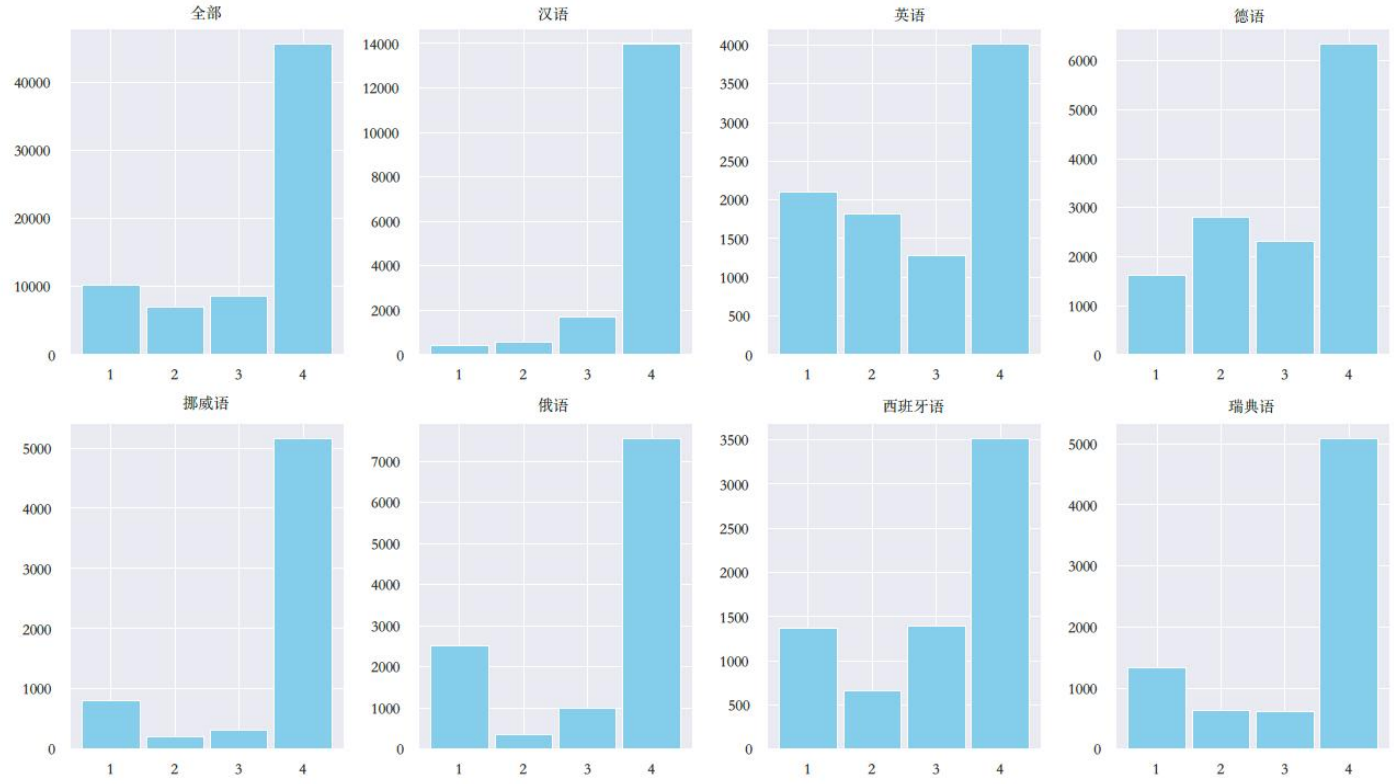


图 3.1　OGWiC 数据集中不同语言的标注值分布图

- 样本标签不均衡，都偏向“一致”

数据集概览

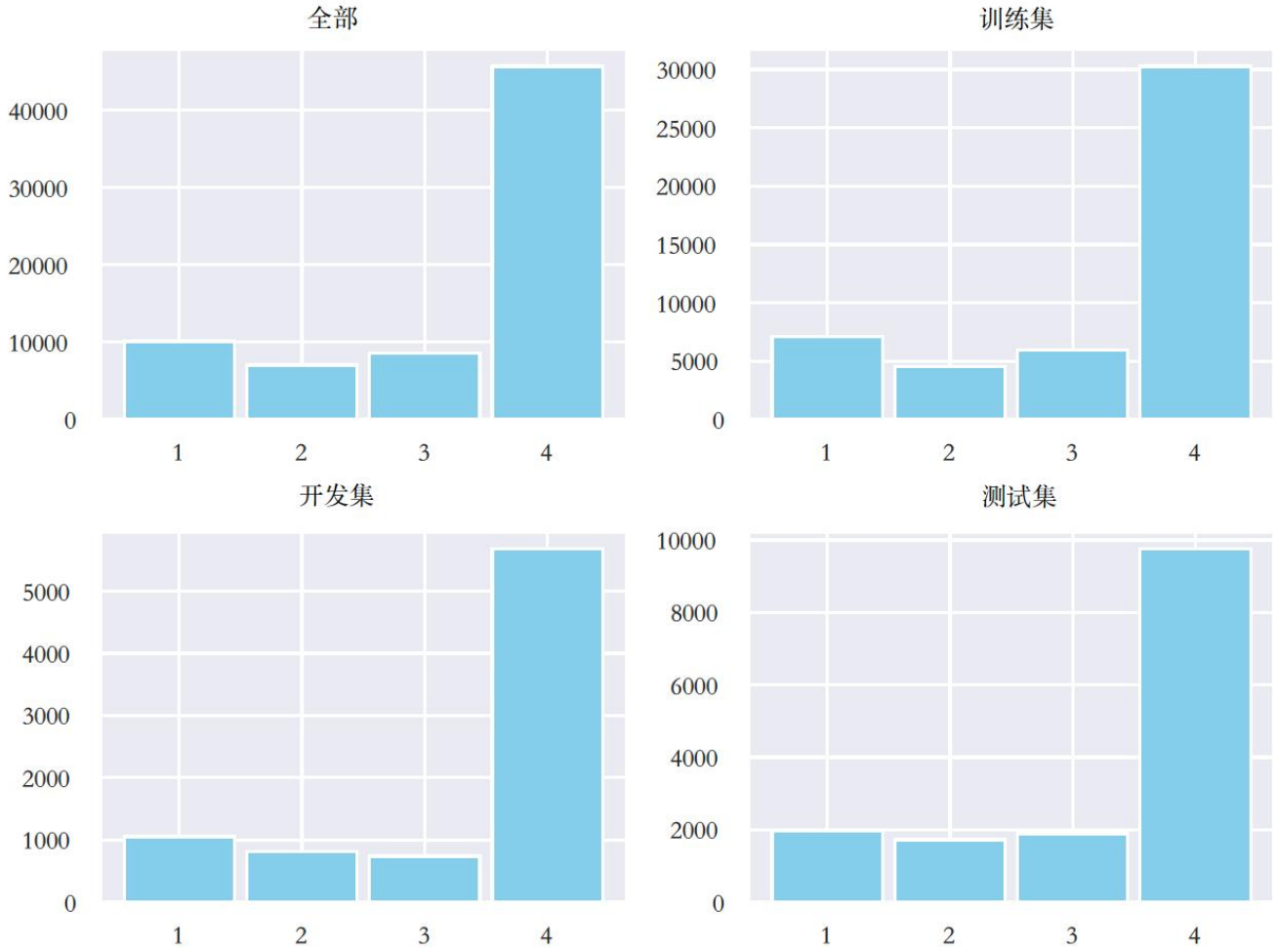


图 3.2　OGWiC 中不同数据集的标注值分布图

- 不同类型数据集分布相似

中文兼类词数据集

- 为了弥补OGWiC数据集中中文数据缺乏词性标注、涉及词数较少、未考虑汉语独有特点等缺陷，本文收集了中文（名动）兼类词数据集
- 汉语中存在较多的名动兼类现象，即同一词形既可以有名词用法，又可以有动词用法。
- 词目和例句来源于《现代汉语词典》
- 招募志愿者进行标注
  - 69名汉语母语者
  - 年龄、学历、地域、专业分布广泛多样

中文兼类词数据集

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

中文兼类词数据集

	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

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

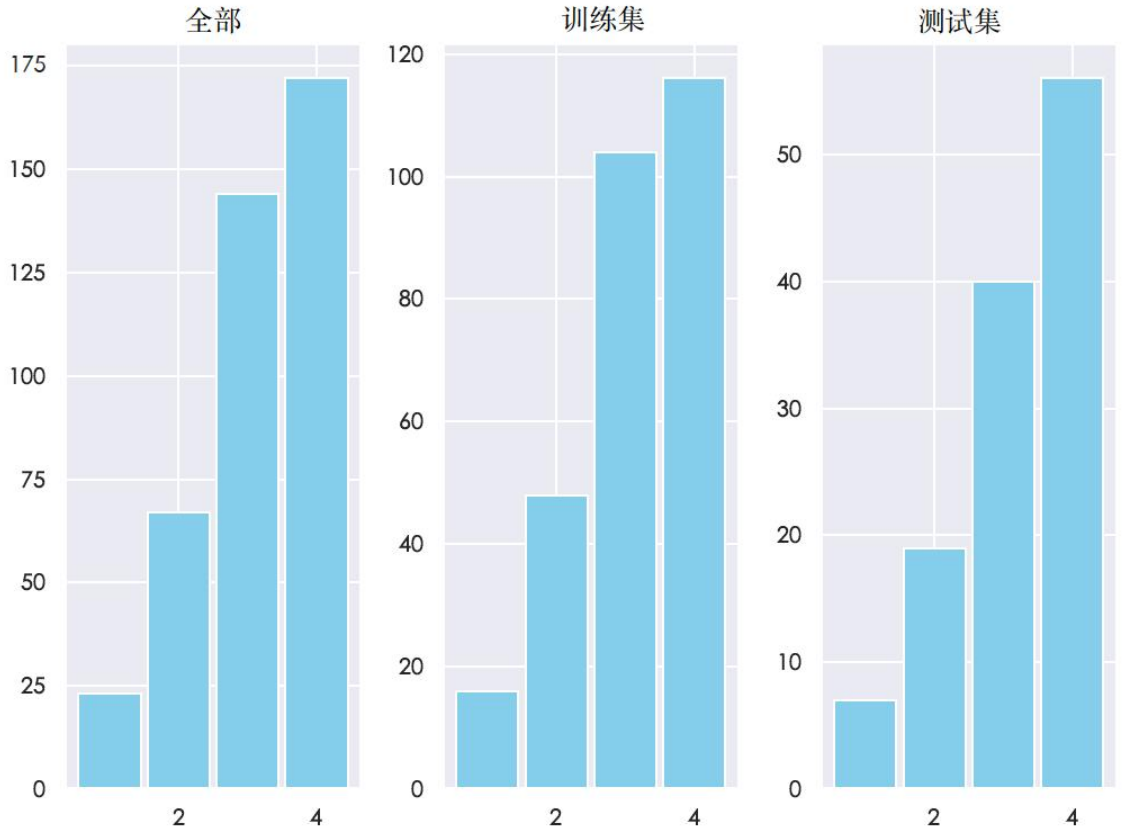


图 3.9　OGWiC_NV 不同的数据集的标注值分布图

- 总词数拓展到近300词，同时兼类词本身带有词类的标注信息
- 词典中的例句句长更短
- 标注者数量更多

基于表征的评估方法

- 表征提取方式
- 几何探测方法
- 向量后处理
- 基于提示语的探测方法

BERT等编码模型：当前词的最后一层

VS

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 **:**

大语言模型为代表的解码模型

基于表征的评估方法

- 表征提取方式
- 几何探测方法
- 向量后处理
- 基于提示语的探测方法

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

基于表征的评估方法

- 表征提取方式
- 几何探测方法
- 向量后处理
- 基于提示语的探测方法

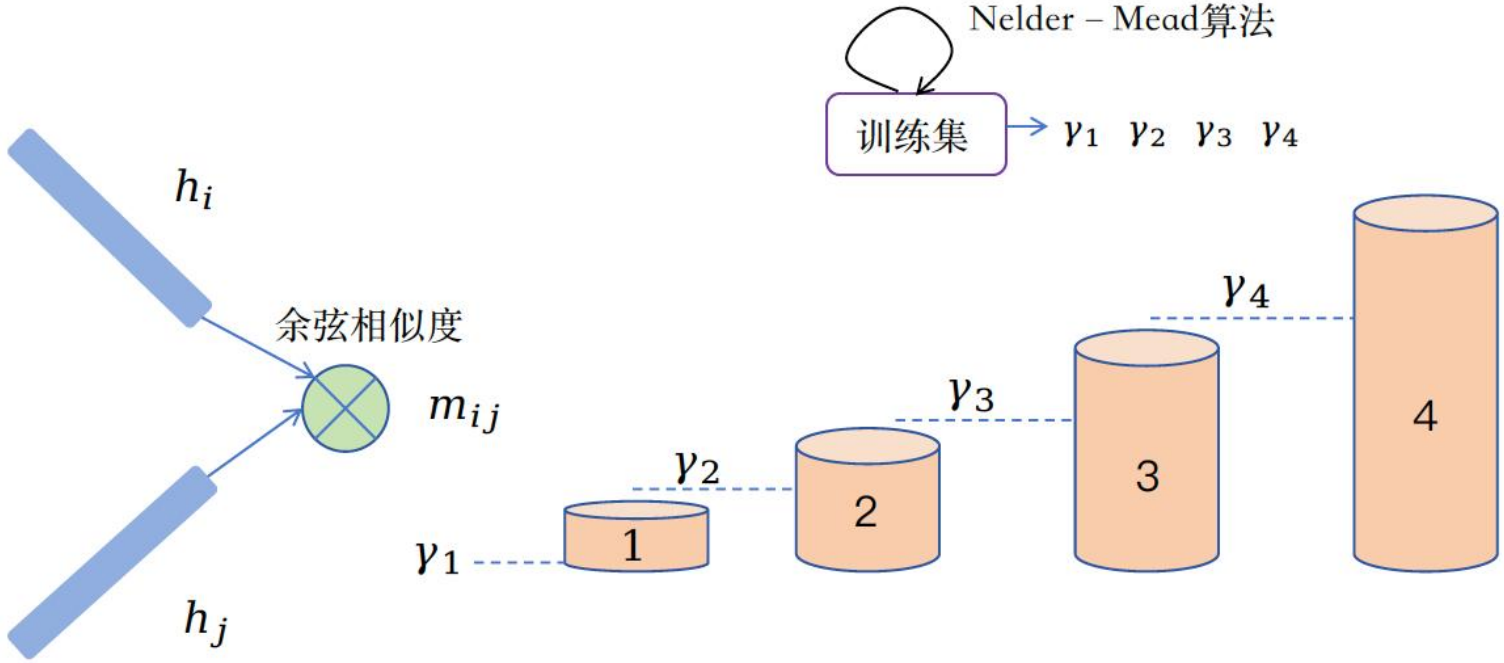
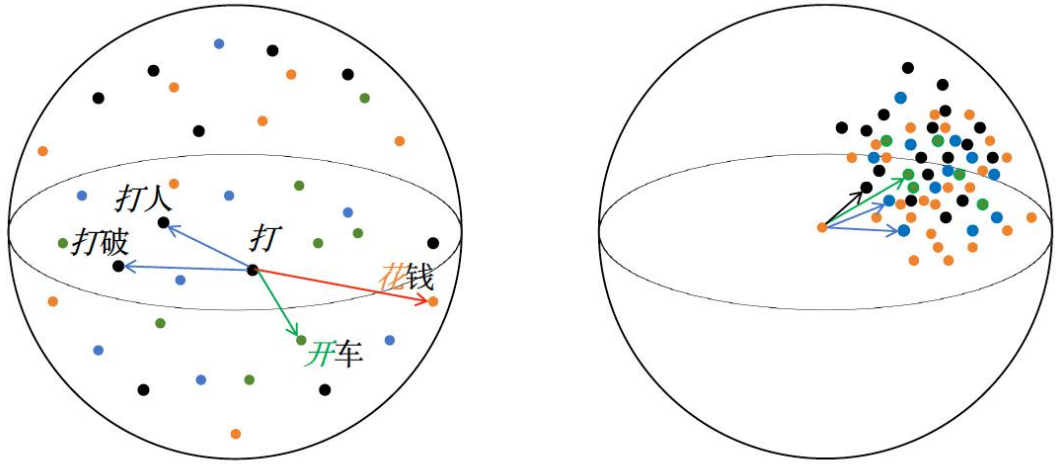


图 4.2　几何探测示意图

基于表征的评估方法

- 表征提取方式
- 几何探测方法
- 向量后处理
- 基于提示语的探测方法



(a) 各向同性　　(b) 各向异性

图 4.3　表征分布示意图

向量的各向异性特征使得语义不同的词也聚集在一起

后处理手段：

（1）中心化

（2）标准化

（3）主成分去除法

基于表征的评估方法

- 表征提取方式
- 几何探测方法
- 向量后处理
- 基于提示语的探测方法

无POS信息

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

利用POS信息

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

实验设定

- 数据集：OGWiC和OGWiC_NV
- 评估指标：克里彭多夫α 系数 （基于分布的准确性）
- 评估模型：
  - 编码器模型：XLM-BERT；各类语言单独训练的BERT模型；XLM-RoBERTa; XL-Lexeme（根据相似任务训练过的模型）；mBART/mT5的编码器部分
  - 解码器模型：
    - LLaMA系列开源模型：LLaMA-2-7B 和 LLaMA-3-8B –》基于embedding
    - DeepSeek模型
    - 通义千问模型

    –》基于外部prompt
  - 上界方法：当前标注者与其余标注者之间计算准确率
  - 下界方法：最大频率下界，随机下界

实验结果

- 各类模型在不同数据集上的表现结果

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**

实验结果

- 平均来看，OGWiC数据集中模型表现的趋势为：
  - 上界>任务微调+编码器模型（XL-Lexeme）> 大语言模型+外部prompt（通义千问）> 解码器模型+prompt提取表征（LLaMAn_p）> 仅编码器模型（XLM-R）> 编解码模型的编码部分（mBART）> 两个下界
- 兼类词数据集OGWiC_NV数据集：
  - 与之类似
  - 上界>大语言模型+外部prompt（通义千问）>任务微调+编码器模型（XL-Lexeme）> 解码器模型+prompt提取表征（LLaMAn_p）> 仅编码器模型（XLM-B）> 编解码模型的编码部分（mT5）> 两个下界
  - 千问模型对于中文理解的能力更强

模型	全部	OGWiC_NV
上界	95.32	76.00
最大频率下界	-22.40	-43.93
随机下界	-17.38	-19.70
XLM-R	27.59	26.26
XLM-R(11)	33.82	24.89
XLM-B	17.22	26.20
XL-Lexeme	**59.02**	47.36
mBART	11.6	-2.80
mT5	20.79	2.89
LLaMA2	22.95	15.71
LLaMA3	21.69	14.14
LLaMA2_p	<u>54.50</u>	40.26
LLaMA3_p	49.01	40.24
DeepSeek(0.1)	37.15	<u>58.98</u>
DeepSeek(1)	36.71	57.37
通义千问 (1)	52.42	**61.79**

实验结果

- 分语言看：
  - 对于汉语来说，内部提示语提取表征的方式、通义千问模型以及XL-Lexeme表现最佳。
  - 这三种语言在多数语言中都表现很好

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**

实验结果

- 向量后处理的作用
  - 尽管几种方式下，得到最优的不总是某一种，但是最佳的结果是由“去各向异性”后处理所得到的。
  - 表明向量存在一定的“各向同性”现象。

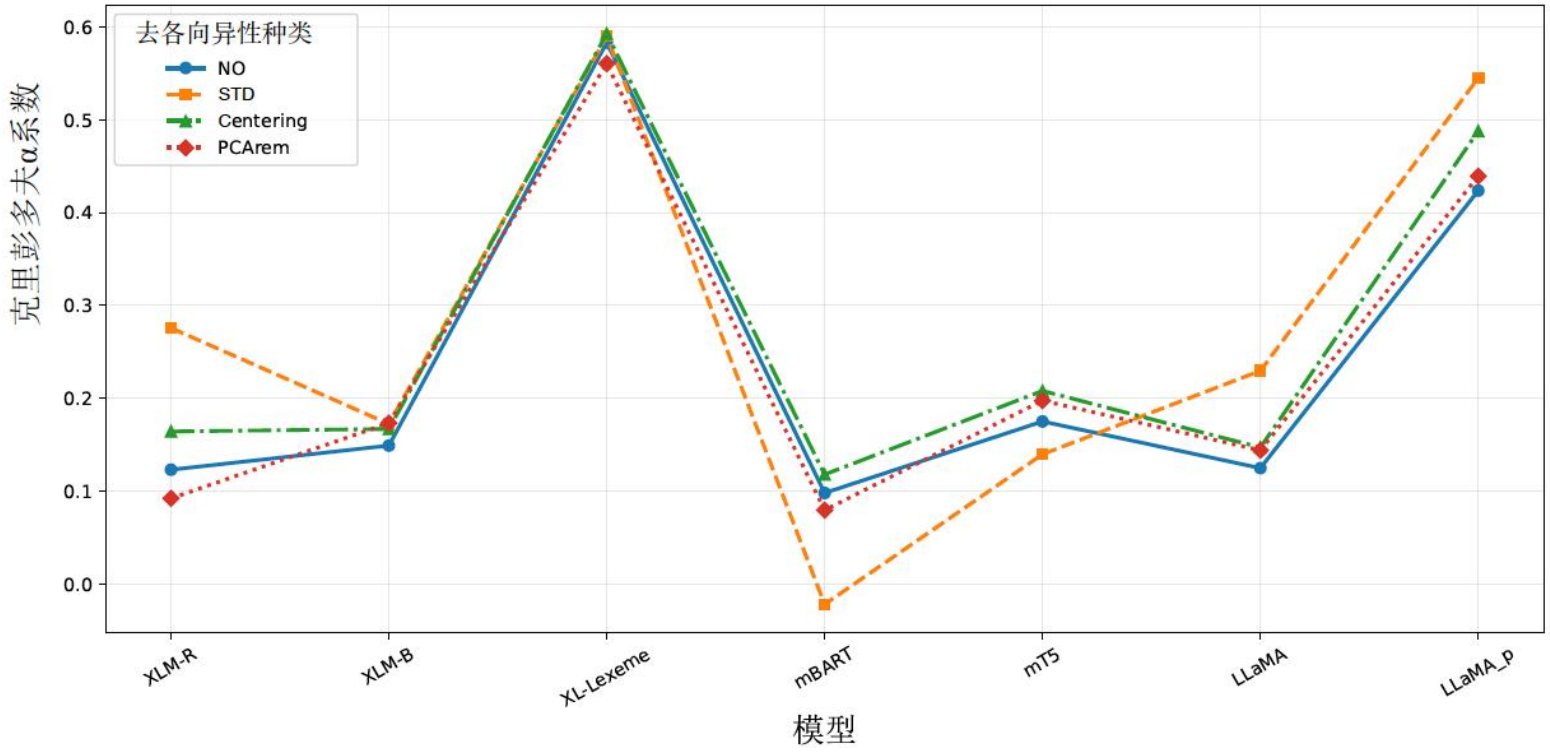


图 4.5　各类去各向异性手段的平均效果

实验结果

- 零样本指令分析
  - 仅通过当前词的最后一层的方式（Baseline）无法起到作用
  - 重复（Echo）的手段可以一定程度上缓解这一现象
  - 基础提示语可以达到比较好的结果了，此外增加了知识信息（PCoTEOL）也更加有效
  - 词类信息（POS_KEEOL）对英语有增强作用，但对于汉语是削弱的。

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	54.50	26.68	50.80	72.62	56.82	**56.47**	58.27	59.87	**40.26**
PCoTEOL	**55.63**	26.48	58.84	**73.35**	**60.39**	54.57	55.72	**60.05**	37.68
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

实验结果

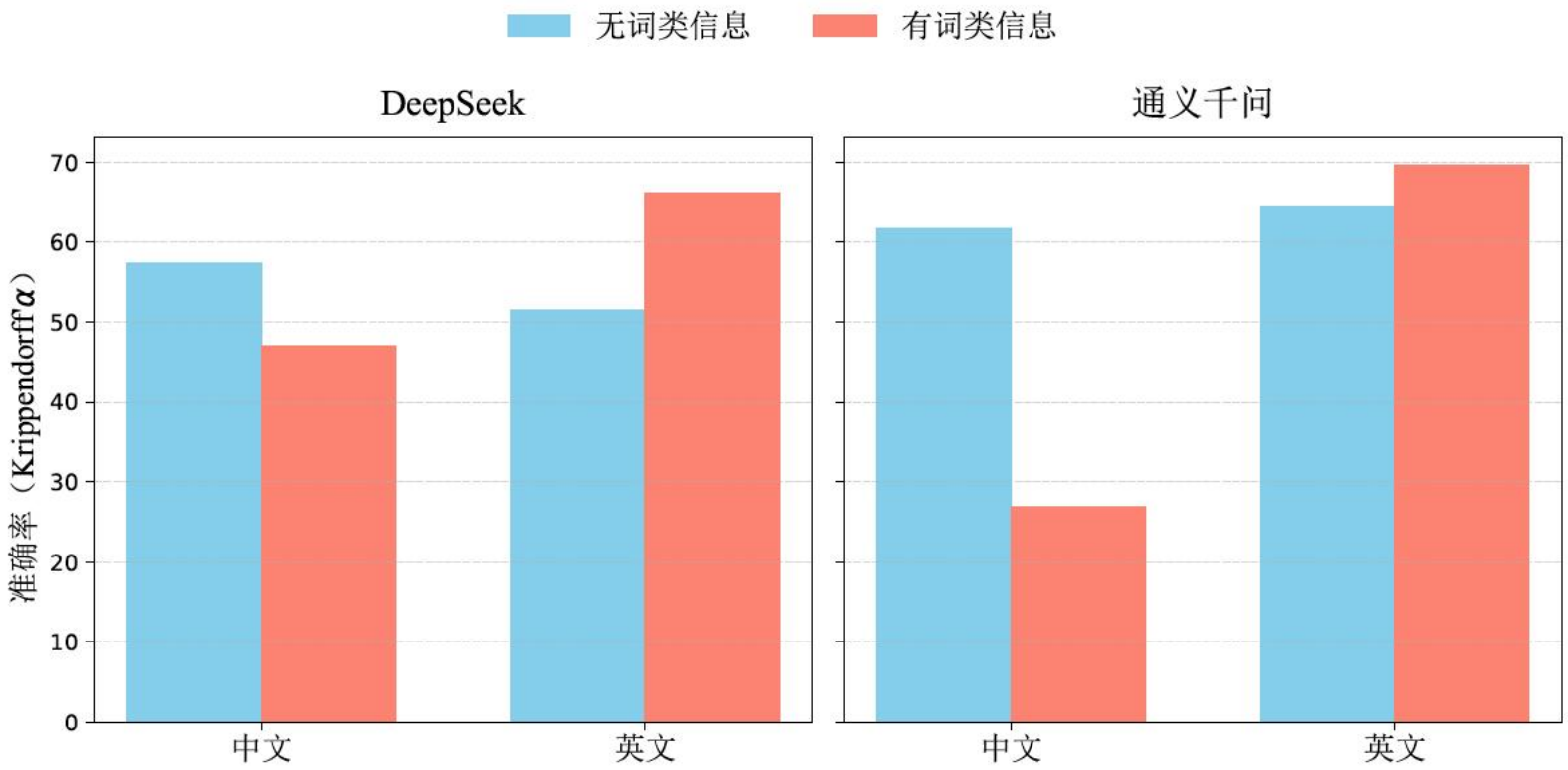


图 4.14　提示语中的词类信息对于闭源大模型的影响

- 提示语中的词类信息对于闭源大模型的影响差异
  - 同之前实验结论类似，词类信息对于中文兼类词词义的区分是有损害的，对于英文而言是有益的。

实验结果

- 多语模型/提示语和单语（英语）的效果区别
  - BERT系列模型来说，跨语言模型普遍要比单语模型效果更好（汉语例外）
  - LLaMA模型的提示语趋势相反，仅使用英语作为提示语，再提取表征的方式要比使用各自语言提示语的更好

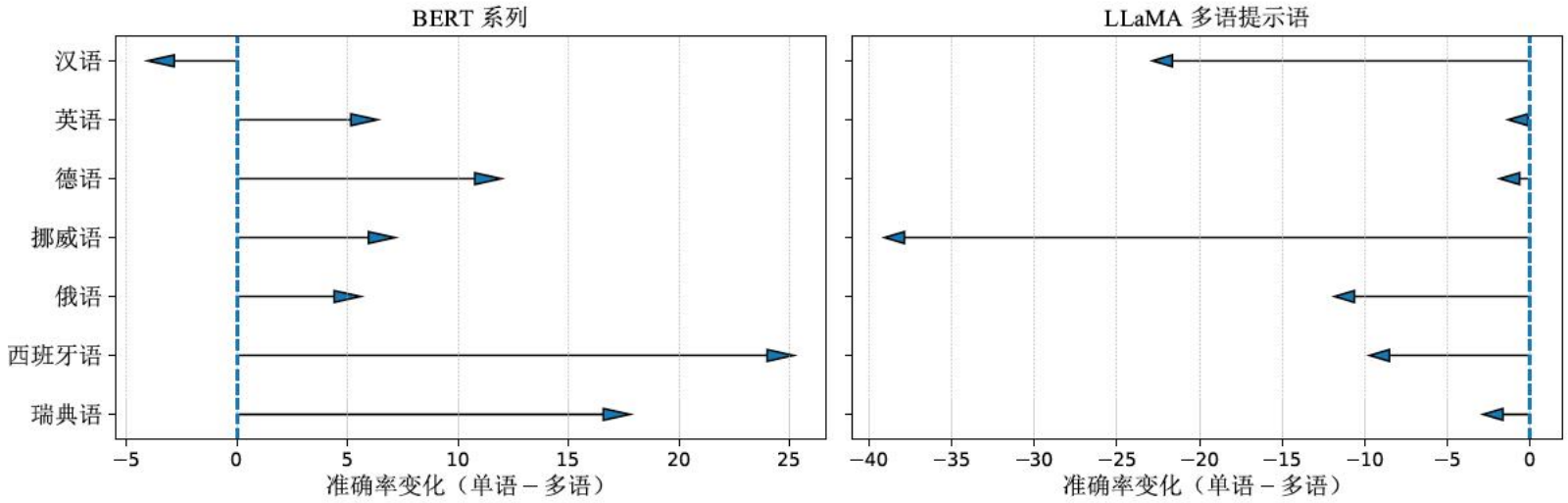


图 4.13　单语模型相对于多语模型的结果变化

典型偏误分析

- 存在一定数量的被高估或者低估的样例，低估样例更多

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

典型偏误分析

- 高估类错误
  - 词汇边界忽略

在原本的样例中，目标在一个句子中是“短语”的一部分，另一部分是“词”，对于这样的情况，模型倾向仅考虑语义，不考虑“词边界”

(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

(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

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。

典型偏误分析

- 高估类错误
  - 语义过度泛化

    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。

    （a）records表示唱片；（b）表示记录
  - 忽视词类差异
  - 简化事件结构

(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

高估类偏误分析（兼类词）

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。

典型偏误分析

- 低估类错误
  - 字面义保留缺失
    - 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
  - 语义域过度分化
    - 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

(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

(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

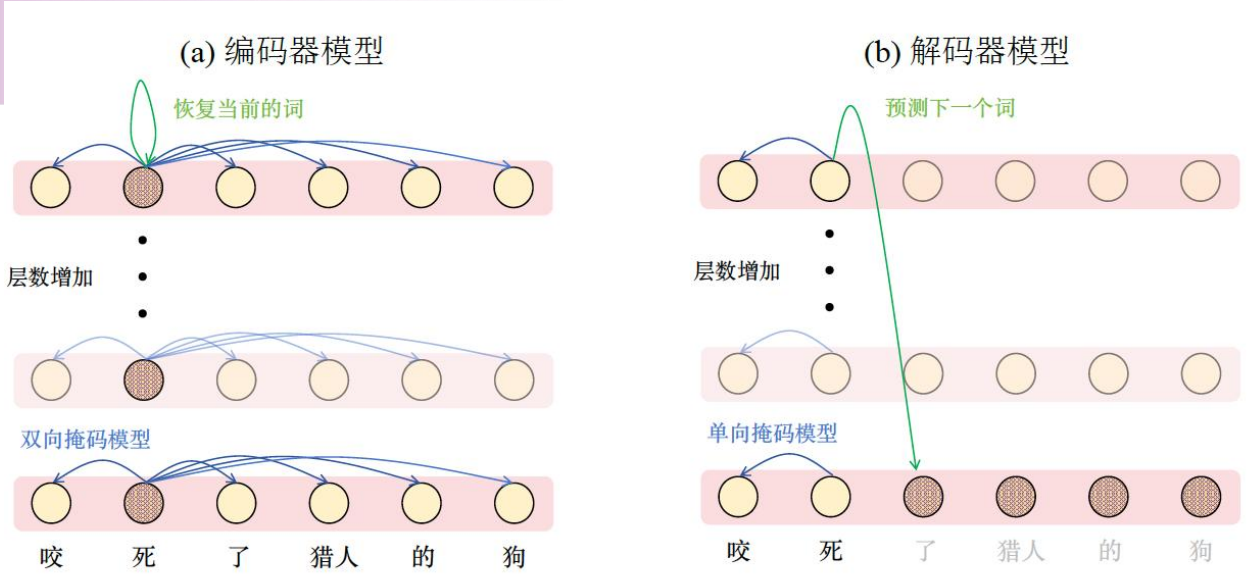
本章小结

- 从多语言多粒度视角对词义相关度任务在多样模型上进行评估，并利用包括几何探测、向量表征提取、向量后处理、外部提示语在内的多种方式
- 无论是外部使用提示语还是内部利用提示语提取特征，大语言模型都可以得到较好的结果，但是与人类标注之间仍存在差距，且出现不同类型的高估和低估的样例
- 词类信息对于中英文的词义理解具有差异，英文更看重词类信息

4. 多语言词义相关度任务中的模型可解释性研究

第5章

两类模型架构的对比



- 编码器模型
  - 理解当前token时采用两边的上下文
  - 当前token的理解和预测的信息传递过程是一致的
  - 代表词义的普遍做法：当前token的最高层表征
  - 广泛适用于理解型任务（命名实体识别、阅读理解等）
- 解码器模型
  - 理解当前token仅依赖前文
  - 预测下一个词作为目标，“理解”和“预测”的信息过程不一致
  - 广泛适用于生成类任务（对话生成）

研究问题

- 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?

（本文提出的假说）

- 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

方法

- 词义任务
  - 英语通用词义相关度WiC
  - 多语言词义相关度OGWiC
- 逐层几何探测
  - 利用各层提取出的向量特征完成上述词义任务
  - 几何：相似性+阈值
- 评估模型
  - 编码器模型（BERT...）
  - 解码器模型（LLM...）

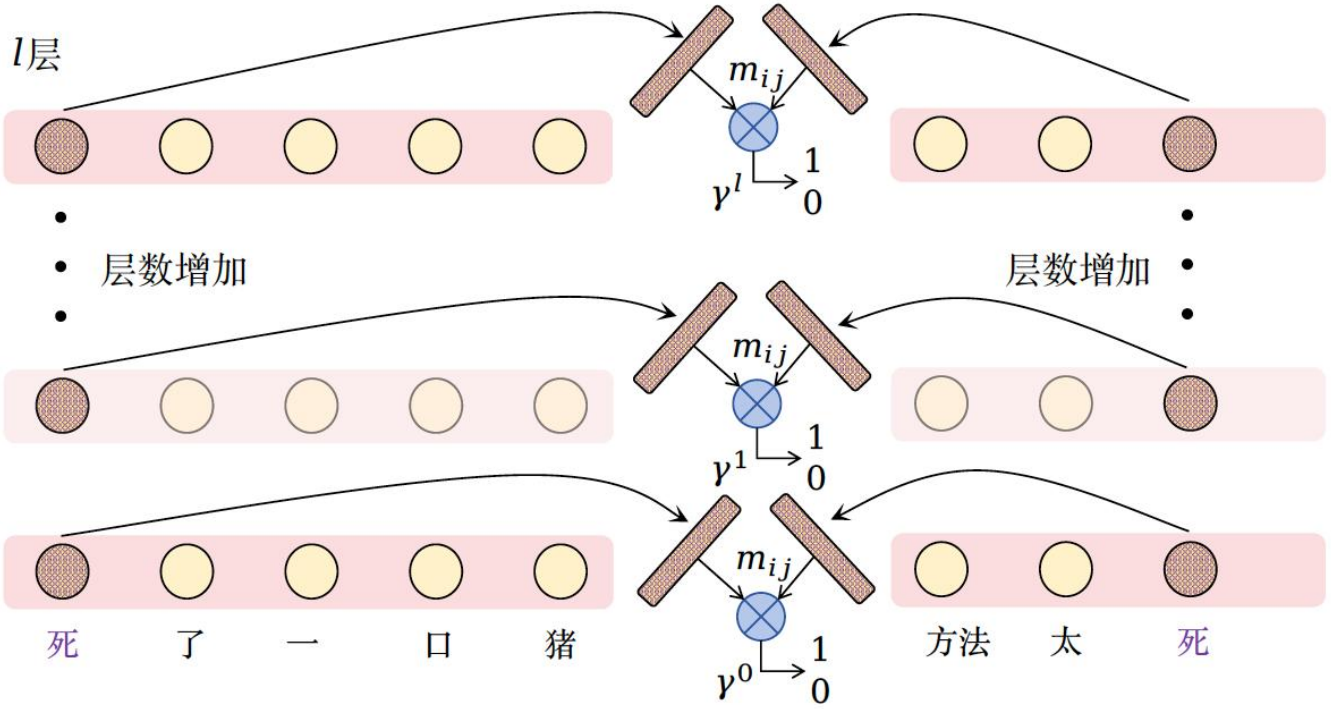


图 5.2　逐层几何探测流程图

方法

- 词义任务
  - 英语通用词义相关度WiC
  - 多语言词义相关度OGWiC
- 逐层几何探测
  - 利用各层提取出的向量特征完成上述词义任务
  - 几何：相似性+阈值
- 评估模型
  - 编码器模型（BERT...）
  - 解码器模型（LLM...）

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

实验结果

- 英语数据集WiC
  - 大模型为代表的解码器优于编码器模型（与之前结论保持一致），且接近人类基准，表明二值任务要比多等级任务更简单。

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)

基准模型

传统方法

编码器模型

解码器模型1

解码器模型2

低

中

高

实验结果

- 英语数据集WiC
  - 最优结果的层数分布：
    - 除了Basic和Echo提取当前词的设定在中低层，其余的方式均在中高层

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 **:**

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
基准模型	人类基准	80.0	-	-
	随机选择	50.0	-	-
传统方法	LMMS	67.7	-	-
	Context2vec	59.3	-	-
	ELMo	57.7	-	-
编码器模型	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)
解码器模型1	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)
解码器模型2	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)

低

中

高

实验结果

- 词性
  - 名词整体性能优于动词
    - 动词词义对上下文依赖程度更高
    - 位于句首的动词 vs. 名词：19.2% vs. 14.3%

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)

基准模型

传统方法

编码器模型

解码器模型1

解码器模型2

低

中

高

跨层分析：假说

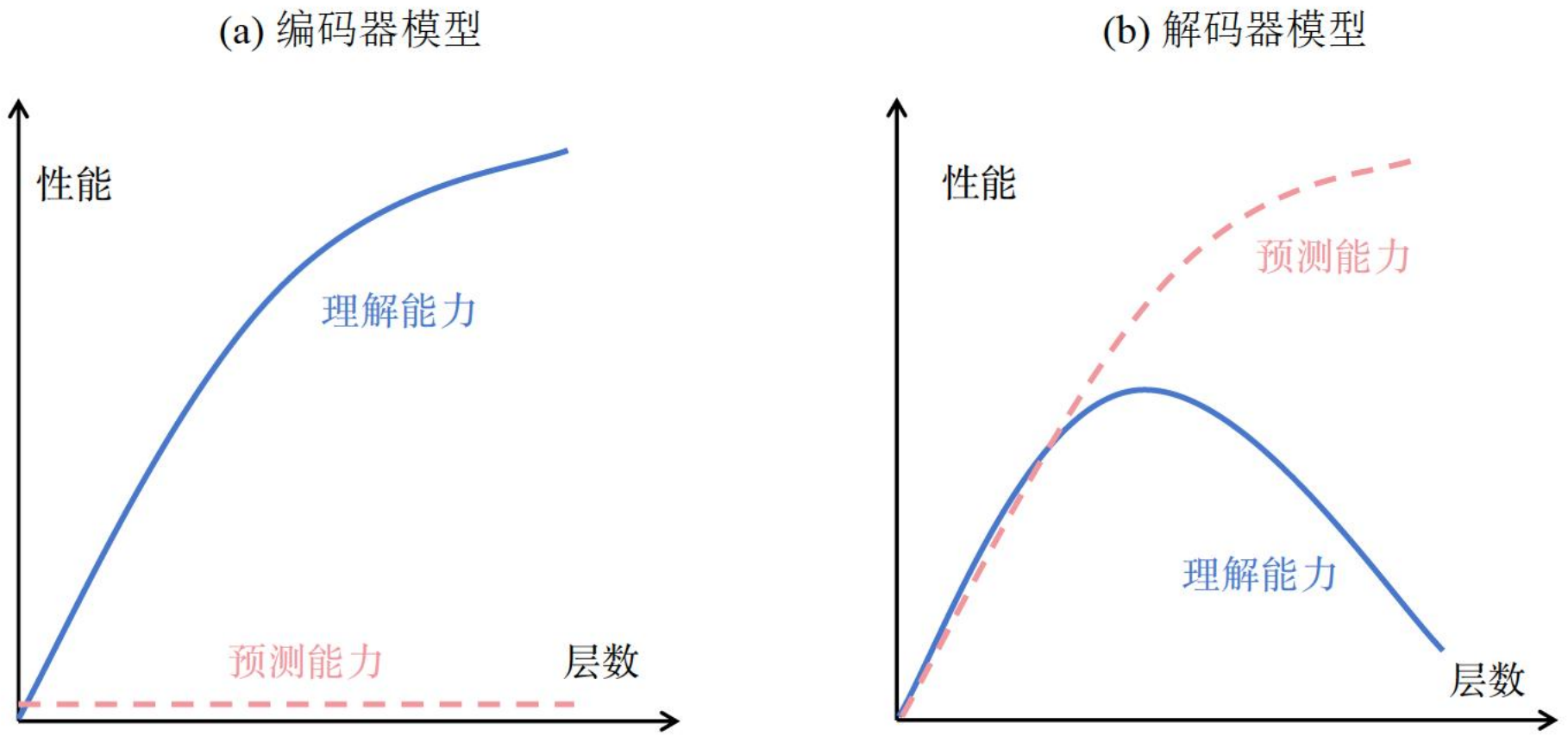


图 5.3　两类模型的理解与生成能力逐层变化

跨层分析：实验

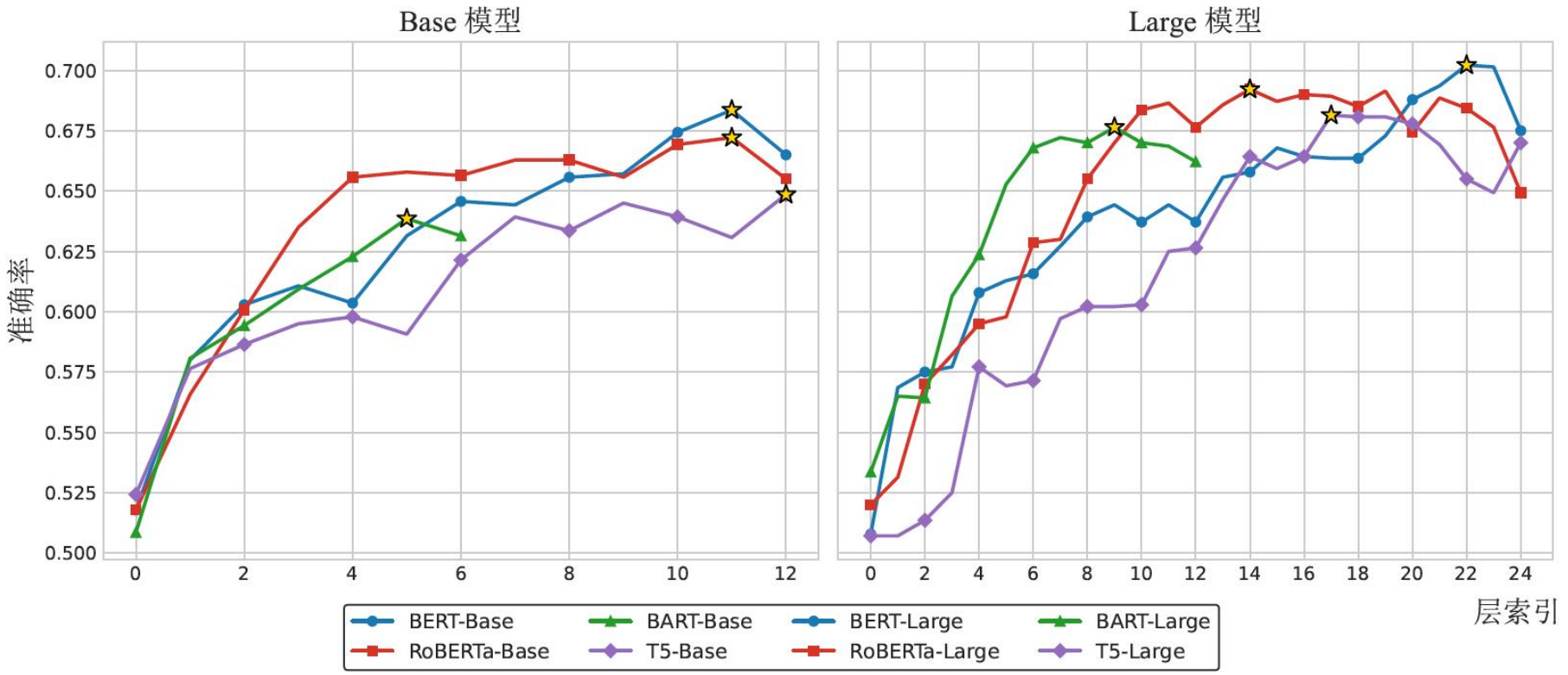


图 5.4　编码器模型随层数增加的性能变化

跨层分析：实验

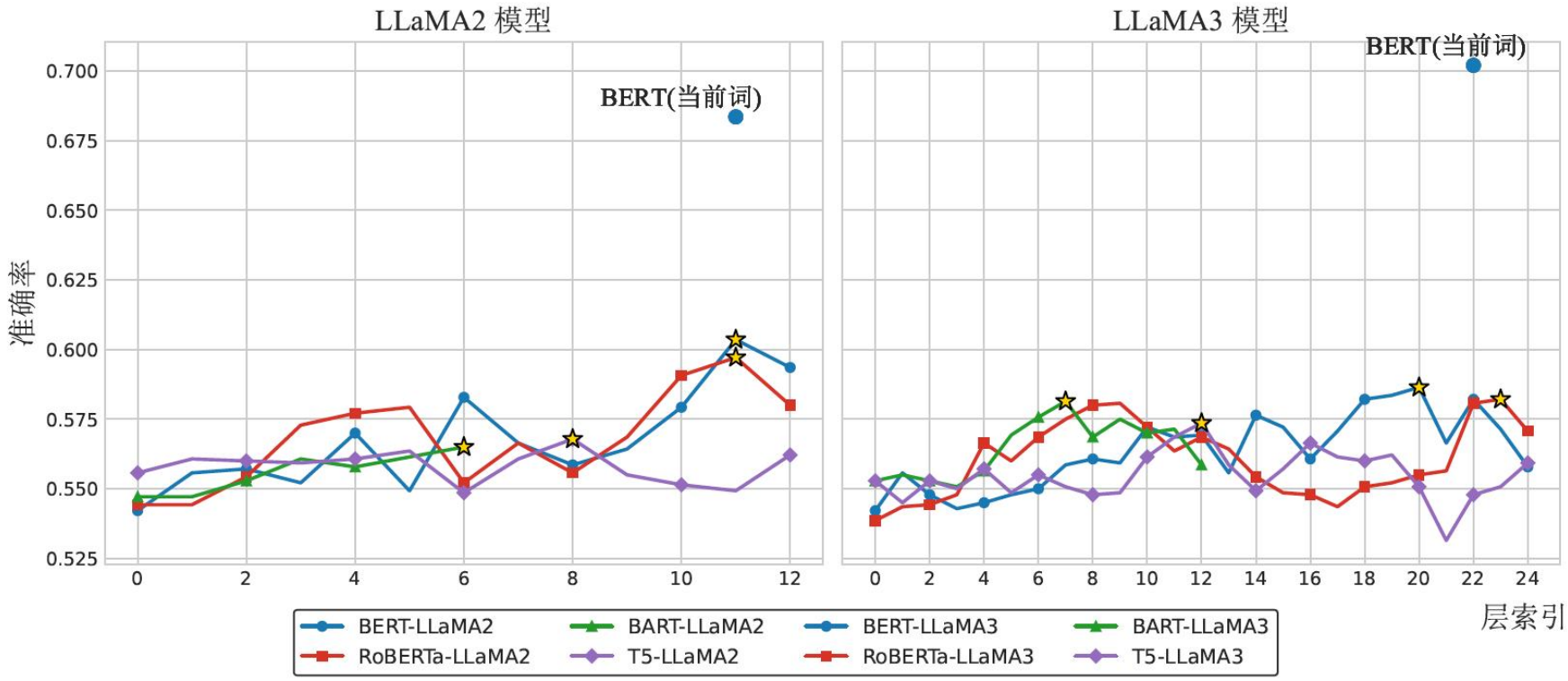


图 5.5　编码器模型针对前一目标词表征性能随层数的变化

跨层分析：实验

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

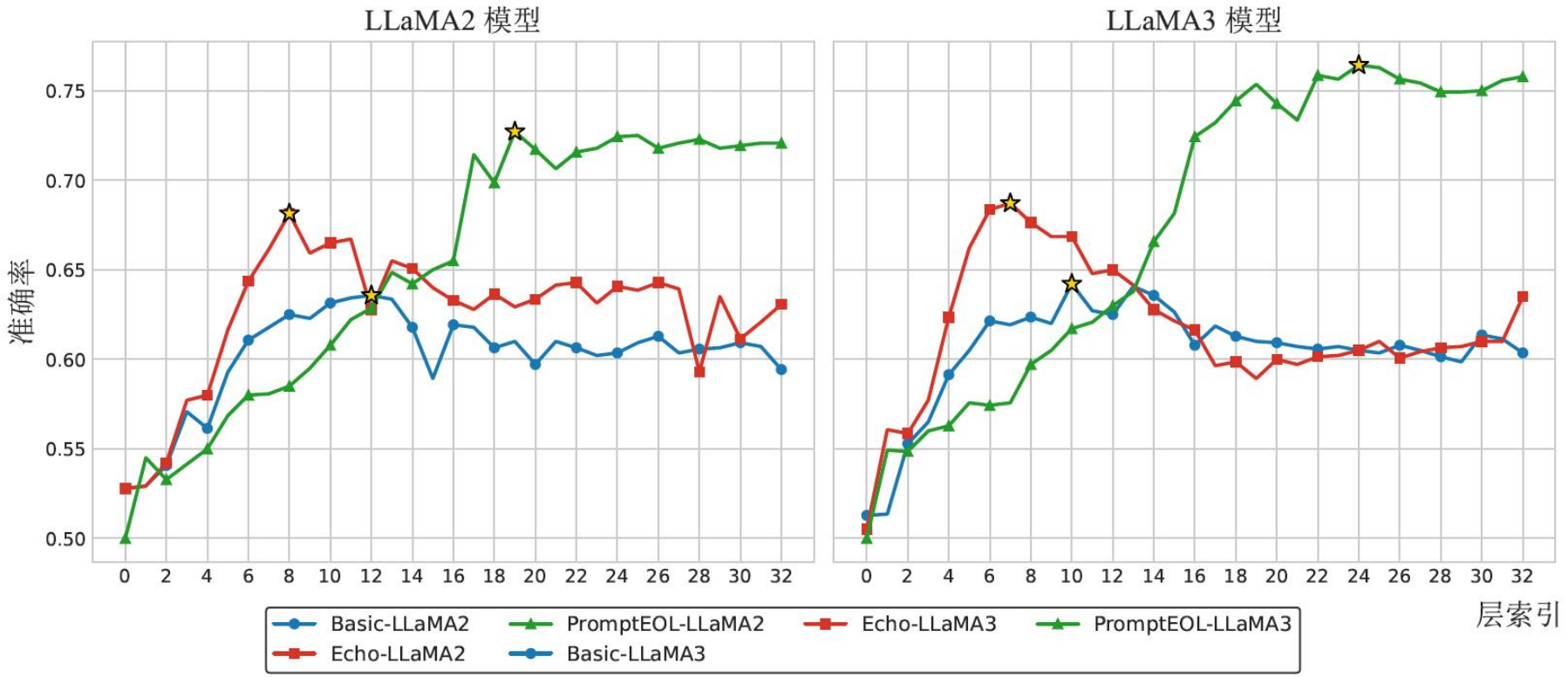


图 5.6　解码器模型随层数增加的性能变化

基础模型 vs. Prompt引导

跨层分析：实验

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

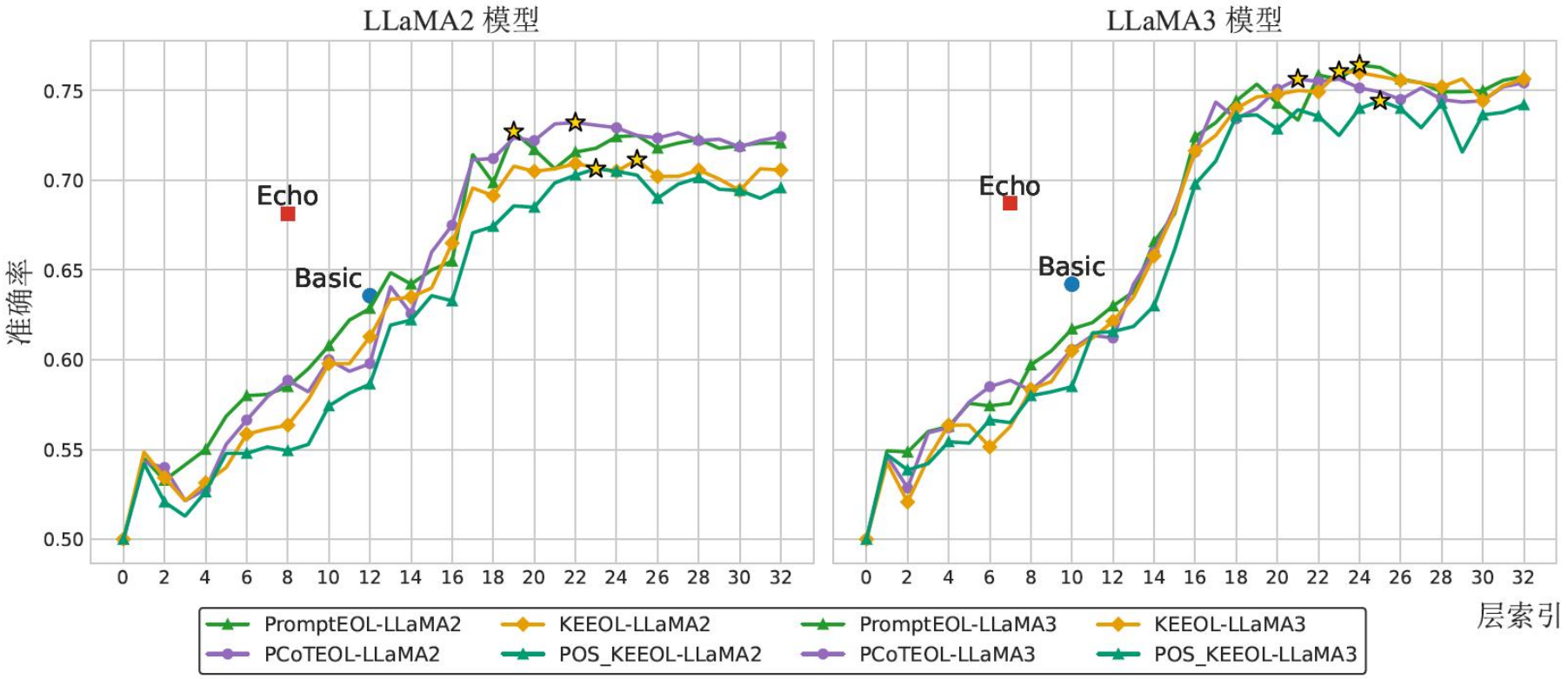


图 5.8　各类提示语对于解码器模型的逐层性能影响

基础模型 vs. 多样 Prompt 引导

跨层分析：实验

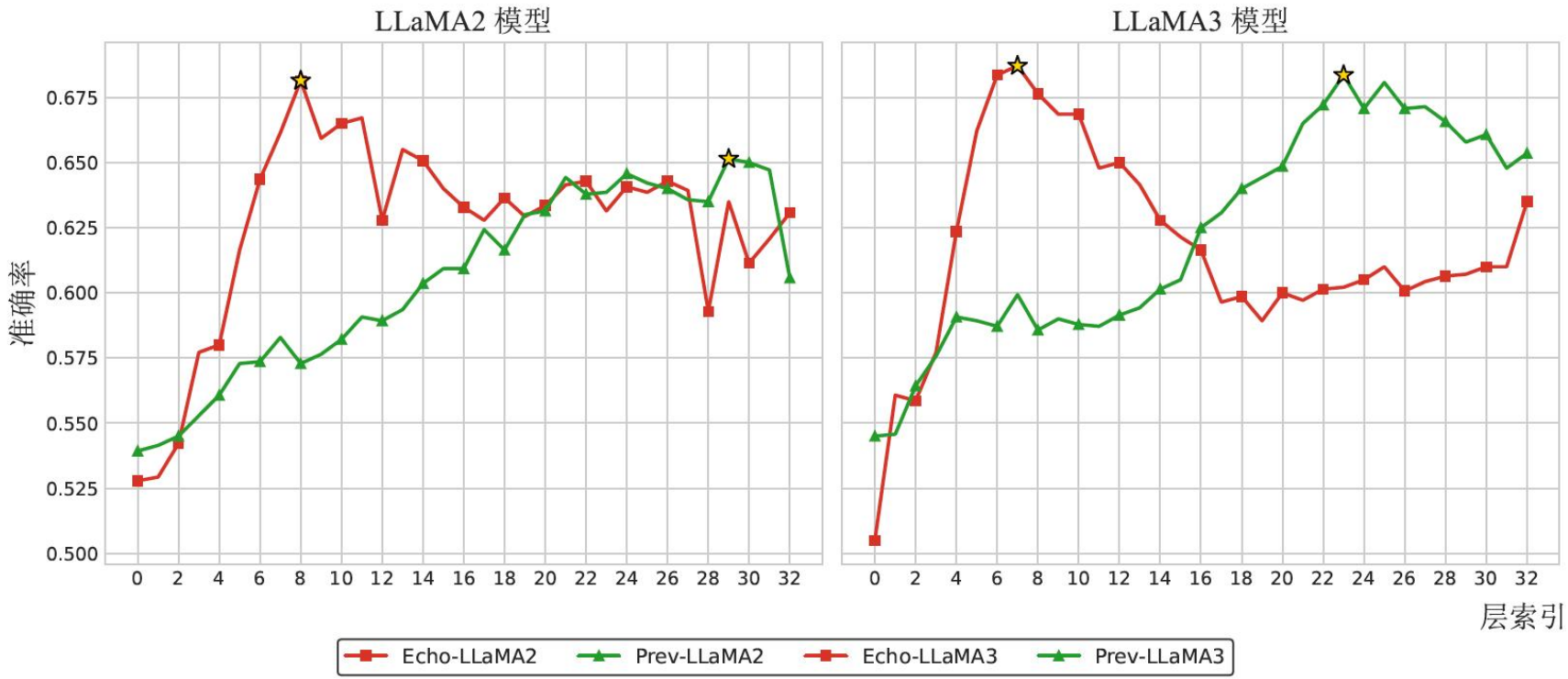
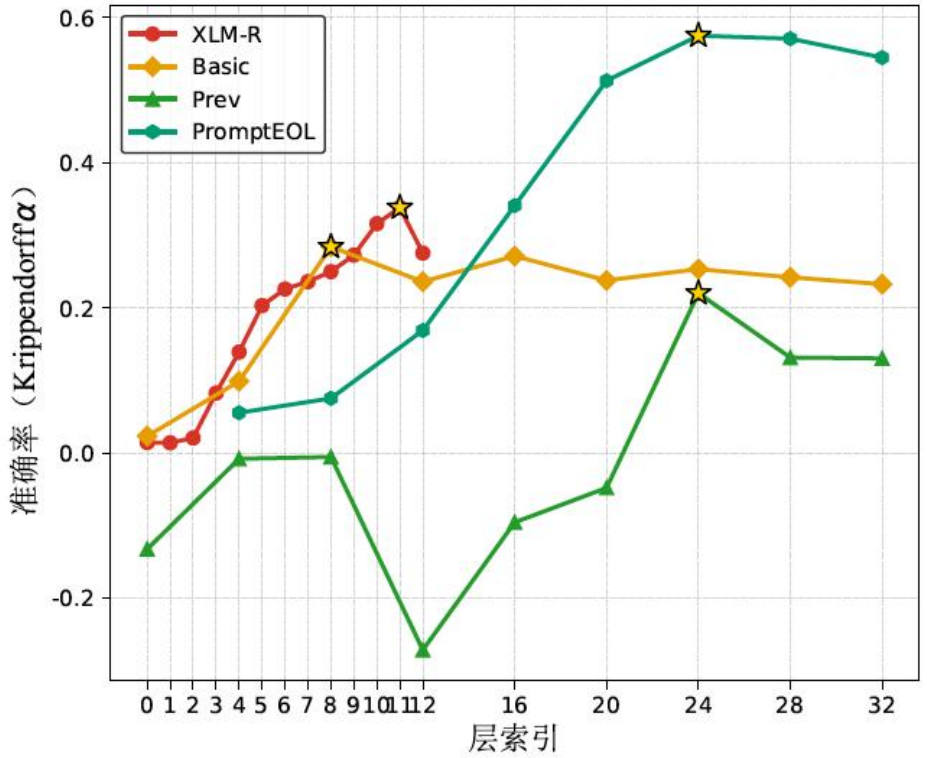


图 5.7　解码器模型针对前一目标词表征性能随层数的变化

基础模型 vs. 前词表征

跨层分析：实验



(a) 各语言平均

多语言词义相关度数据集OGWiC

本章小结

- 针对两个词义相关度数据集，对各类编码模型和解码大语言模型不同层级的词义表征进行了系统探测和分析。
- 实验结果一致表明：解码器模型在中低层编码当前token的词义信息，高层则转向后续输出的预测和生成；编码器模型则在高层编码当前token的词义信息。
- 理论意义：揭示了大语言模型内部“先理解后预测”的机制（普适性）
- 实践意义：对如何使用大语言模型表征词义提供了指导。

多语言词义相关度任务中的标注不一致评估

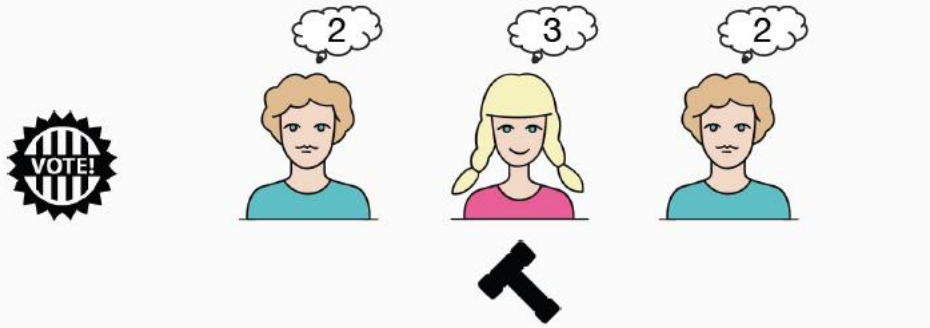
第6章

研究背景

- 标注者不一致现象
  - 词自身的多义性
  - 标注者的不同理解

1. Pat **broke** the his promise.

2. He **broke** the world record.



relatedness: 2; disagreement: 0.667

数据集

- 标签中位数 -> 标签方差
- 多语言相关度数据集OGWiC -> DisWiC
- 中文兼类词数据集：DisWiC -> DisWiC_NV

$$\mathrm{MPD}(\mathcal{J}) = \frac{1}{|\mathcal{J}|} \sum_{(j_1, j_2) \in \mathcal{J}} (|j_1 - j_2|)$$

DisWiC

语言	上下文 1	上下文 2	标注
汉语	警觉高涨情绪中**潜伏**的危机感	某部特工队**潜伏**在丛林中，伺机深入敌后歼敌	[3, 2]
英语	the everlasting smile on her **face** did not belie her heart	Wolfe had straightened up and was making **faces**.	[4, 3]

DisWiC_NV

目标词	上下文 1	上下文 2	标注
预想	结果与人们的**预想**大致相同	**预想**不到事情的结果会这样	[4, 3, 1, 4, 4]
阻碍	毫无**阻碍**	**阻碍**交通	[4, 4, 3, 3, 3, 3, 4]
重赏	这项发明得到 100 万元**重赏**	**重赏**之下，必有勇夫	[4, 4, 3, 4, 4]
翻译	他当过三年**翻译**	**翻译**外国小说	[3, 3, 3, 3, 3, 2, 3]
经验	他对嫁接果树有丰富的**经验**	这样的事，我从来没**经验**过	[4, 2, 2, 3, 3, 2, 3]

数据集

- 标签中位数 -> 标签方差
- 多语言相关度数据集OGWiC -> DisWiC
- 中文兼类词数据集：DisWiC -> DisWiC_NV

$$\mathrm{MPD}(\mathcal{J}) = \frac{1}{|\mathcal{J}|} \sum_{(j_1, j_2) \in \mathcal{J}} (|j_1 - j_2|)$$

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

数据集

- 标签中位数 -> 标签方差
- 多语言相关度数据集OGWiC -> DisWiC
- 中文兼类词数据集：DisWiC -> DisWiC_NV

$$\mathrm{MPD}(\mathcal{J}) = \frac{1}{|\mathcal{J}|} \sum_{(j_1, j_2) \in \mathcal{J}} (|j_1 - j_2|)$$

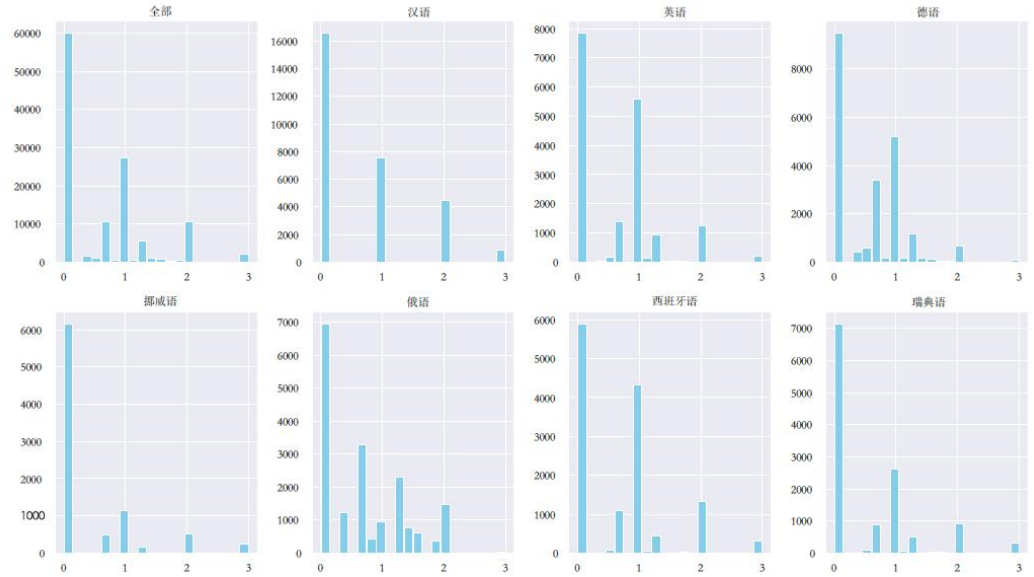


图 3.3　DisWiC 数据集中每种语言的 MPD 分布直方图

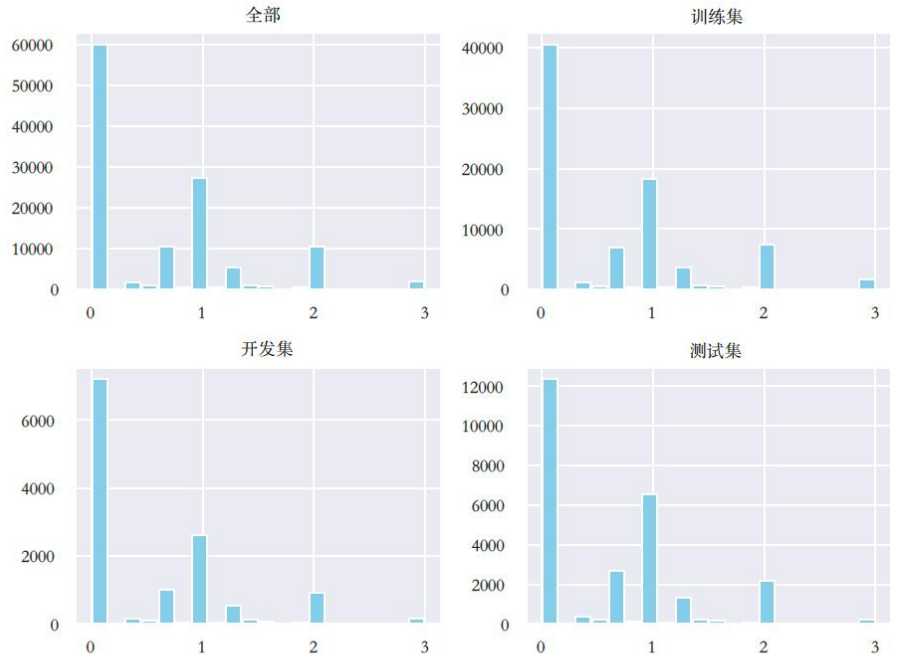


图 3.4　DisWiC 中不同数据集的 MPD 分布直方图

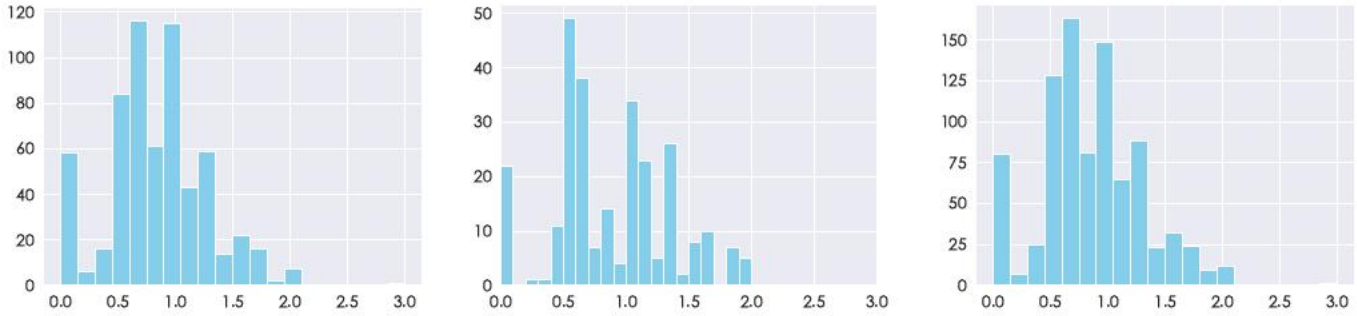
数据集

- 标签中位数 -> 标签方差
- 多语言相关度数据集OGWiC -> DisWiC
- 中文兼类词数据集：DisWiC -> DisWiC_NV

$$\mathrm{MPD}(\mathcal{J}) = \frac{1}{|\mathcal{J}|} \sum_{(j_1, j_2) \in \mathcal{J}} (|j_1 - j_2|)$$

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



(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

方法

- 表征提取
  - 同第4章
- 分类器探测
- 模型融合
  - 同质：相同架构的模型内部参数进行变动，如层数、去各向异性方式等。
  - 异质：不同架构的模型。
  - 混合：不同架构且不同内部参数的模型。

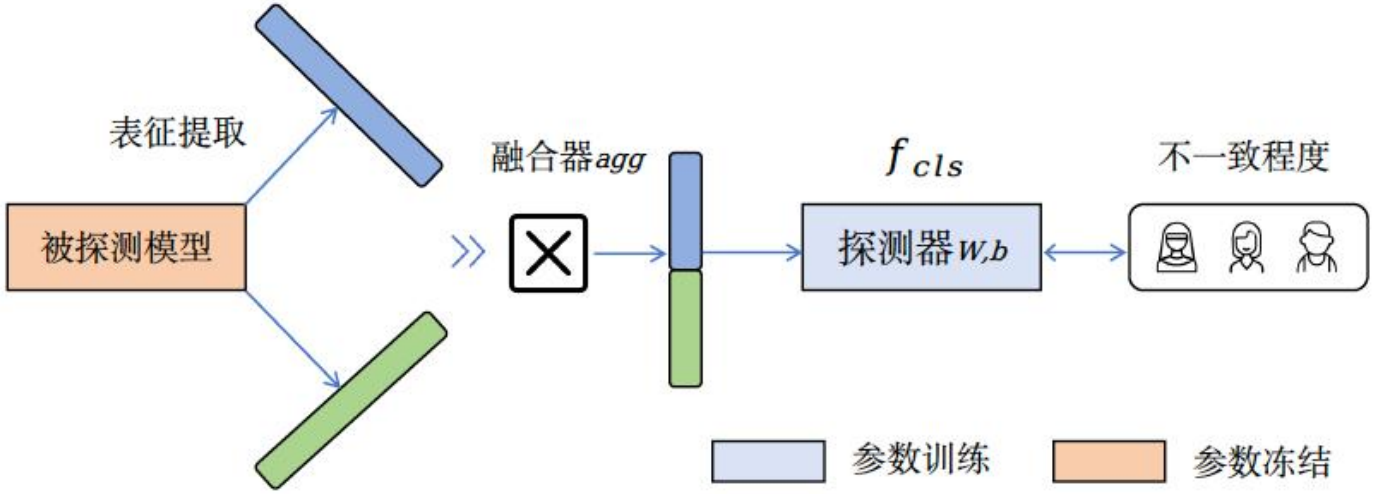


图 6.1　分类器探测示意图

方法

- 表征提取
  - 同第4章
- 分类器探测
- 模型集成
  - 同质：相同架构的模型内部参数进行变动，如层数、去各向异性方式等。
  - 异质：不同架构的模型。
  - 混合：不同架构且不同内部参数的模型。

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

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

方法

- 表征提取
  - 同上一章
- 分类器探测
- 模型融合 - 融合指标
  - 连续值输出（相关程度） vs. 离散值输出（相关度）
  - 连续值融合指标：ASD，MPD_c
  - 离散值融合指标：MPD_d, VR

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）

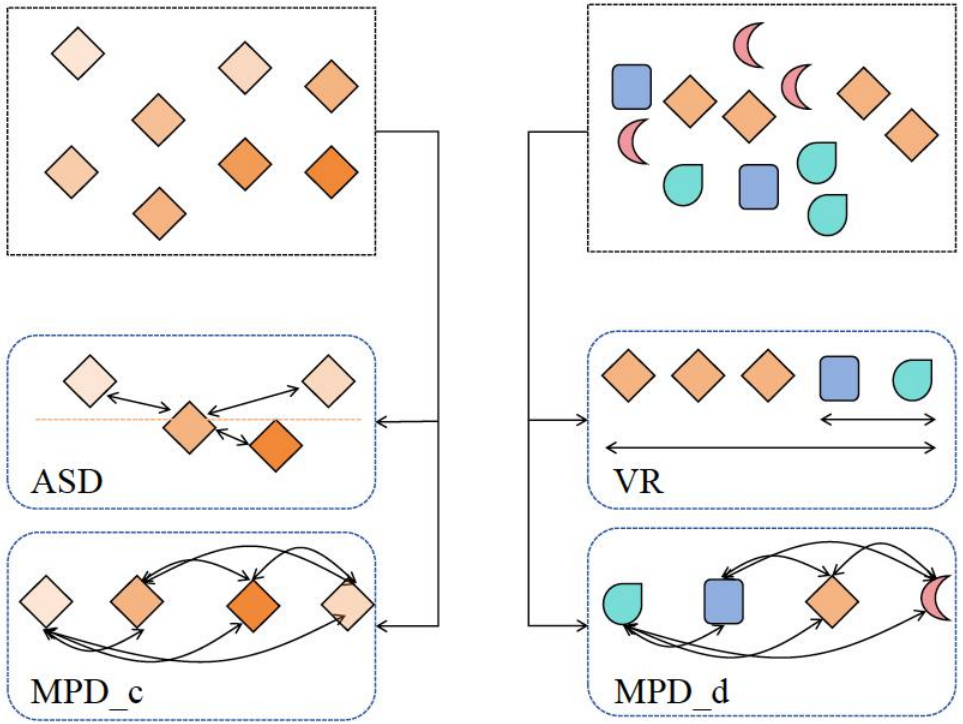


图 6.2　两种方式的模型集成方式对比图

方法

- 针对闭源大模型的提示语探测方式

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.

结果分析

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

- 衡量标准: 与真值的斯皮尔曼系数
- 模型集成的方式 优于 其他基线模型和大模型 优于 分类器探测方法
- 词类信息对于标注不一致的判断有一定正向作用，不过并不显著

结果分析

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
连续值集成 ASD	**26.15**	49.96	**10.27**	**21.56**	38.25	**16.53**	7.34	**39.17**	16.71
连续值集成 MDP_c	26.12	**49.97**	10.23	21.42	**38.26**	16.42	**7.36**	39.15	16.48
离散值集成 MDP_d	16.76	38.21	6.28	15.13	20.71	8.59	5.88	22.51	18.76
离散值集成 VR	15.86	37.80	4.41	14.56	17.11	11.05	6.36	19.70	**20.08**

- DisWiC：连续值集成的方式要优于离散值集成
- DisWiC_NV：离散值集成方式优于连续值集成

结果分析

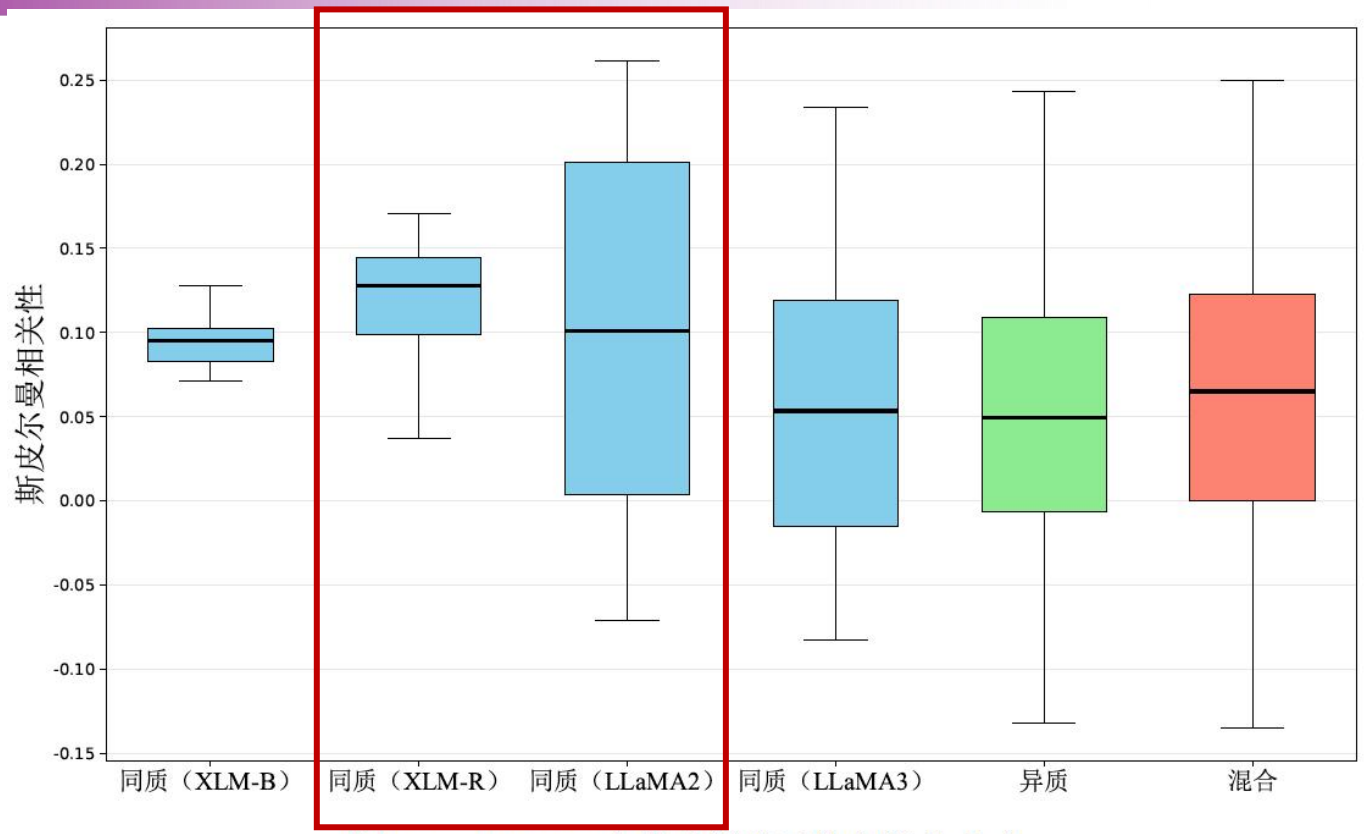


图 6.5　DisWiC 中集成策略下的相关性分布

- 同质集成在中位数与最大值方面均优于其余两类方法；混合集成次之；异质集成整体表现最弱
- 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。

结果分析

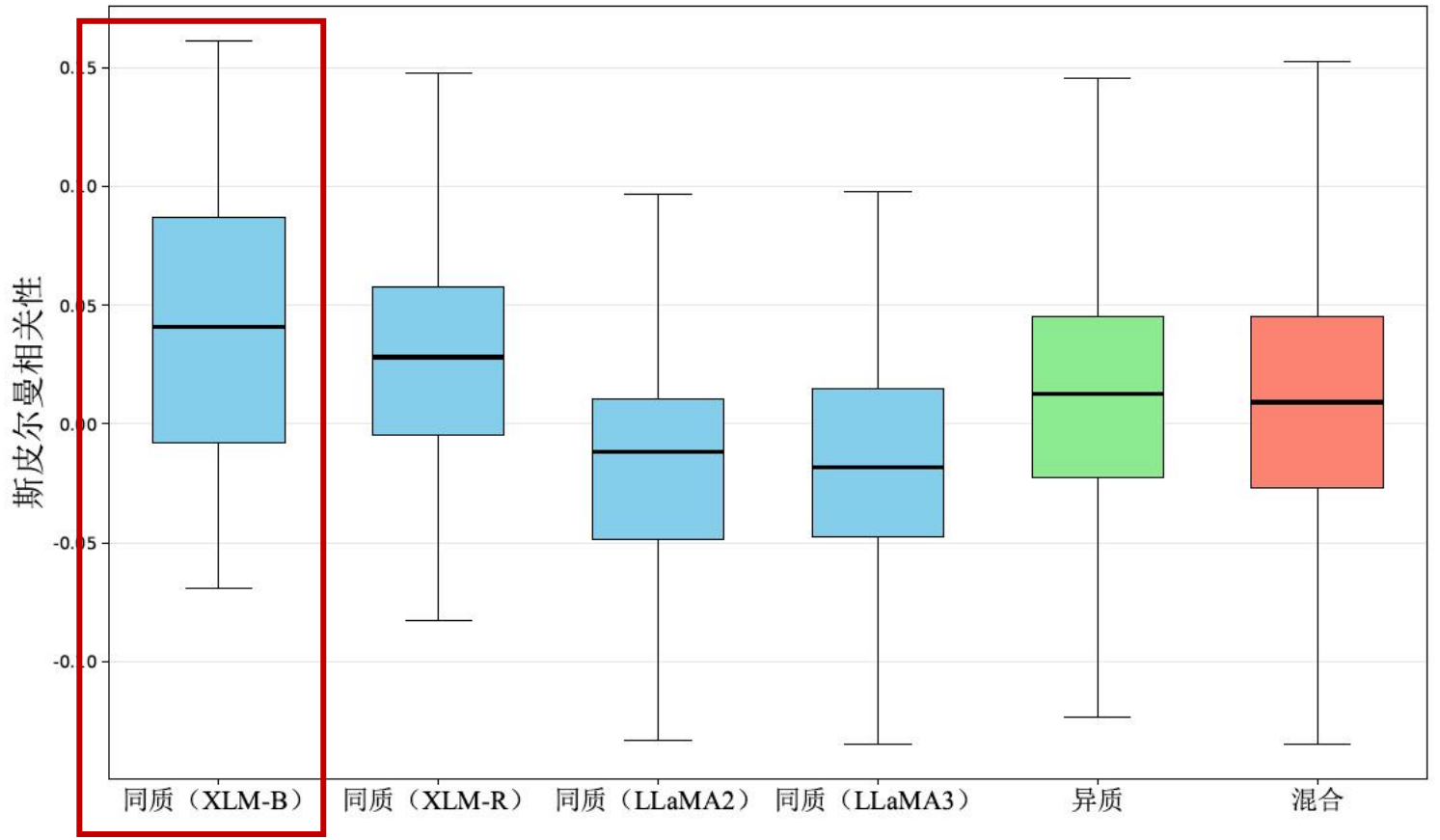


图 6.6　DisWiC_NV 在不同集成策略下集成实例相关性的分布

- 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（小模型表现出色）
- 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。

结果分析

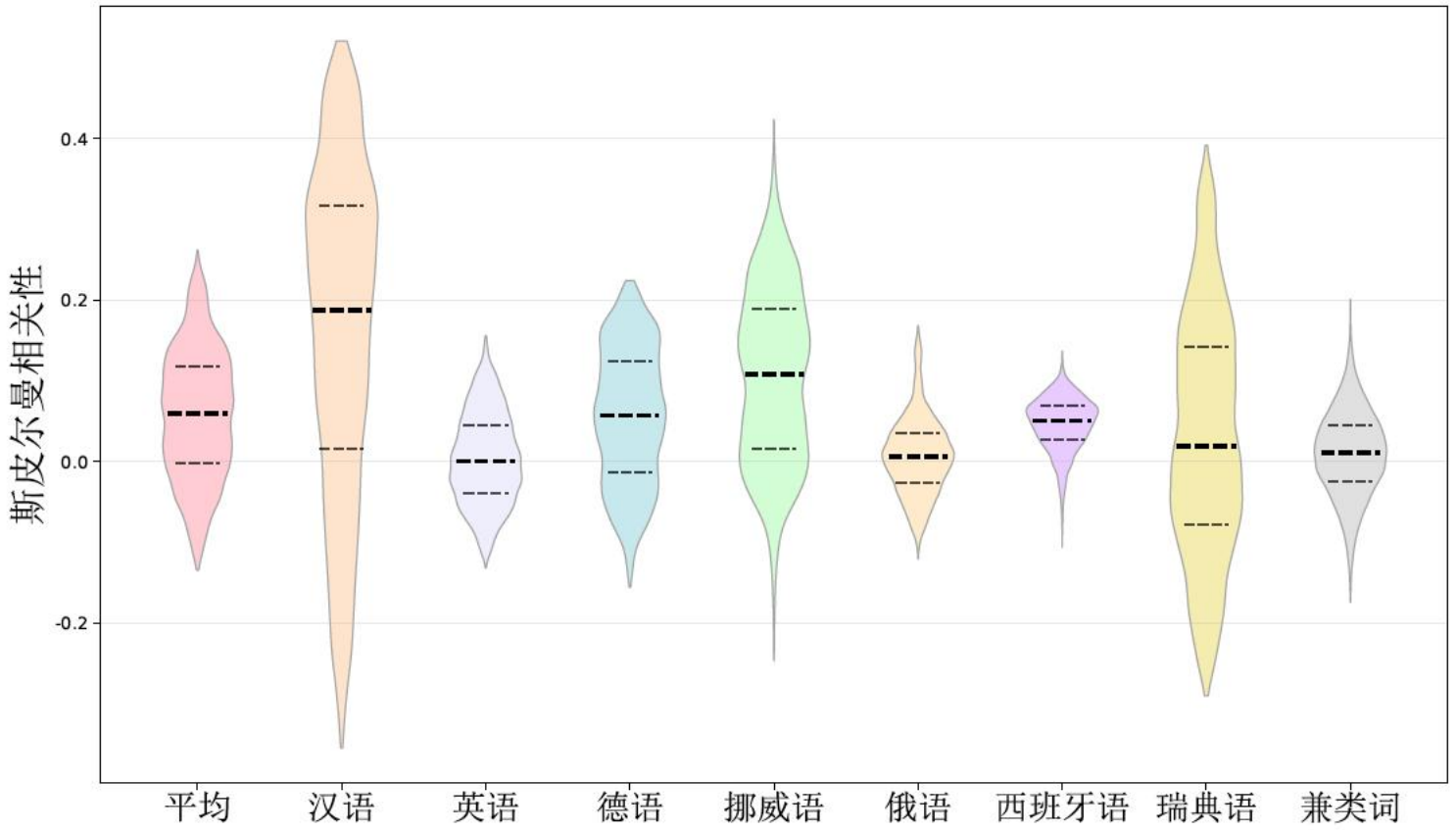


图 6.7　不同语言集成实例相关性的分布

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

偏误分析

- 高估 vs. 低估
  - 大模型结果和人类结果的比较
  - 模型倾向于高估不一致程度

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

偏误分析

- 模型倾向于先认定“不相关”，之后再判断为“不一致”
  - 通过阅读模型的解释，模型有些混淆“相关”和“一致”的概念
- DisWiC_ZH

(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。

偏误分析

- 模型倾向于将认定为“不相关”的实例判定为“不一致”
- 兼类词数据集 DisWiC_NV

(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 “keep-sake” (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。

偏误分析

- 词类信息容易将高相关度的实例视作高不一致

没有词类
0.3

(4) a. 教练教练得法。

b. 足球教练。

c. 专家标注：0.0（3.0，3.0）

有词类
0.8

(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。

(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。

低估类案例

(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。

本章小结

- 多标注者视角：第4章的深化和拓展
- 本章提出三类标注不一致程度预测方法（分类器探测、模型集成、基于提示语），模型集成结果最佳，但仍与人类标注有较大差异
- 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。

结论

论文第7章

总结和展望

- 主要贡献：
  - 汉语兼类词数据集构建：融合了汉语语言的特性
  - 词义相关度的多维评估与分析：揭示了大语言模型内部表征对于词义的理解程度
  - 标注不一致性的建模与评估：评估了大语言模型对于标注不一致的预测
  - 大语言模型的跨层信息流动分析：揭示了不同模型架构间的跨层信息流动模式
- 展望：
  - 结合语言学理论，设计更多词义理解相关的任务
  - 与机制可解释性结合，对模型进行更加深入的理解
  - 与更多下游应用和任务结合

相关成果

[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）

感谢各位老师和同学！

欢迎指导和提问