清華大學 Tsinghua University

论文分享

刘柱

2025.10.28

大纲

Learning to Look at the Other Side: A Semantic Probing Study of Word Embeddings in LLMs with Enabled Bidirectional Attention

Zhaoxin Feng and **Jianfei Ma** and **Emmanuele Chersoni** and **Xiaojing Zhao** and **Xiaoyi Bao**
Chinese and Bilingual Studies, The Hong Kong Polytechnic University
{zhaoxinbetty.feng,jian-fei.ma,xiaojing.zhao,xiaoyi.bao}@connect.polyu.hk,
emmanuele.chersoni@polyu.edu.hk

ACL 2025

Exploring Layer-wise Representations of English and Chinese Homonymy in Pre-trained Language Models

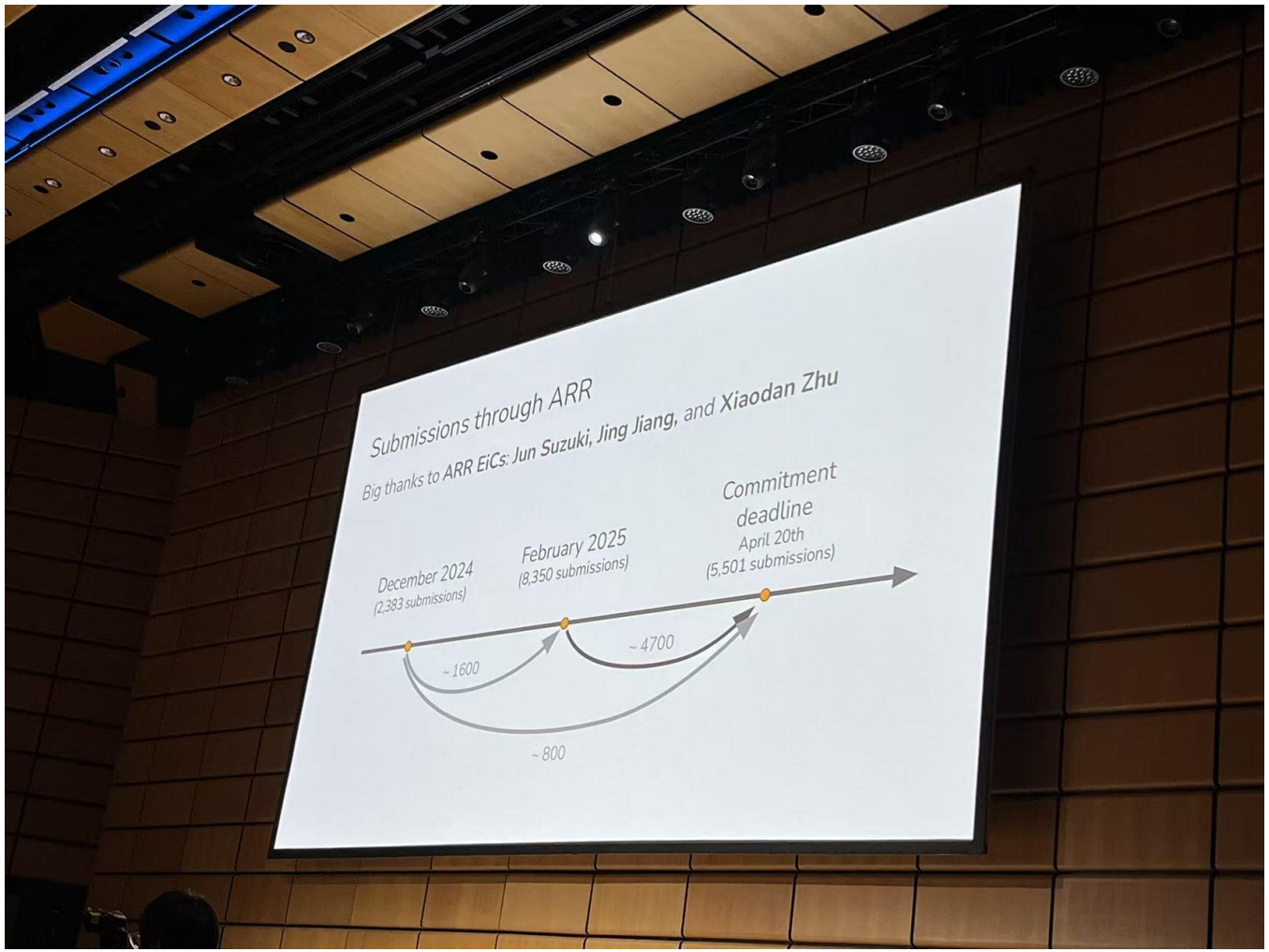
Matthew King-Hang Ma[*♠], **Chenwei Xie**[*♠], **Wenbo Wang**[♣], **William Shiyuan Wang**[♠]
Research Centre for Language, Cognition, and Neuroscience
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
♠{khmma,cwxie,wsywang}@polyu.edu.hk
♣wenbo99.wang@connect.polyu.hk

ACL 2025 Findings

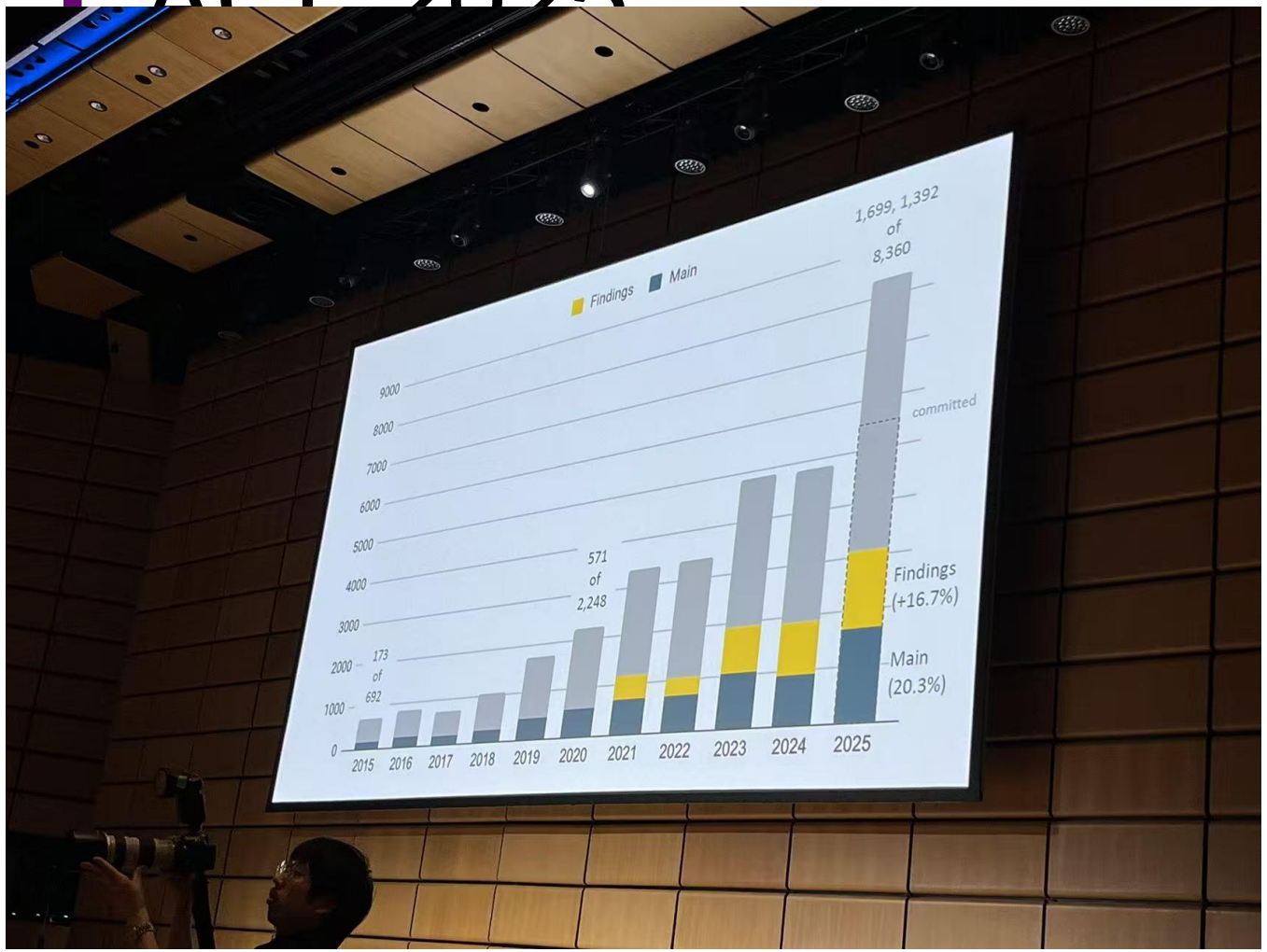
ACL 2025

- 第六十三届计算语言学大会
- 举办地：奥地利维也纳
- 时间：2025年7月27日 - 2025年8月1日
- 统计
  - 主会文章量：1700；Findings：1400；工业领域文章：108
  - 17篇CL文章，40篇TACL文章
  - 2个主旨演讲和1个圆桌论坛
  - 28个工作坊，8场讲习班，64篇展示论文，104篇学生研讨会论文
  - 投稿情况：Main:1699/8360=20.3%; Findings: 1392/8360=16.7%
  - 5千9百个审稿人 -> 1千1百名领域主席 -> 169名高级领域主席
    （审稿+审稿意见）（根据审稿意见和Rebuttal给出meta）（根据meta给出最终决定）

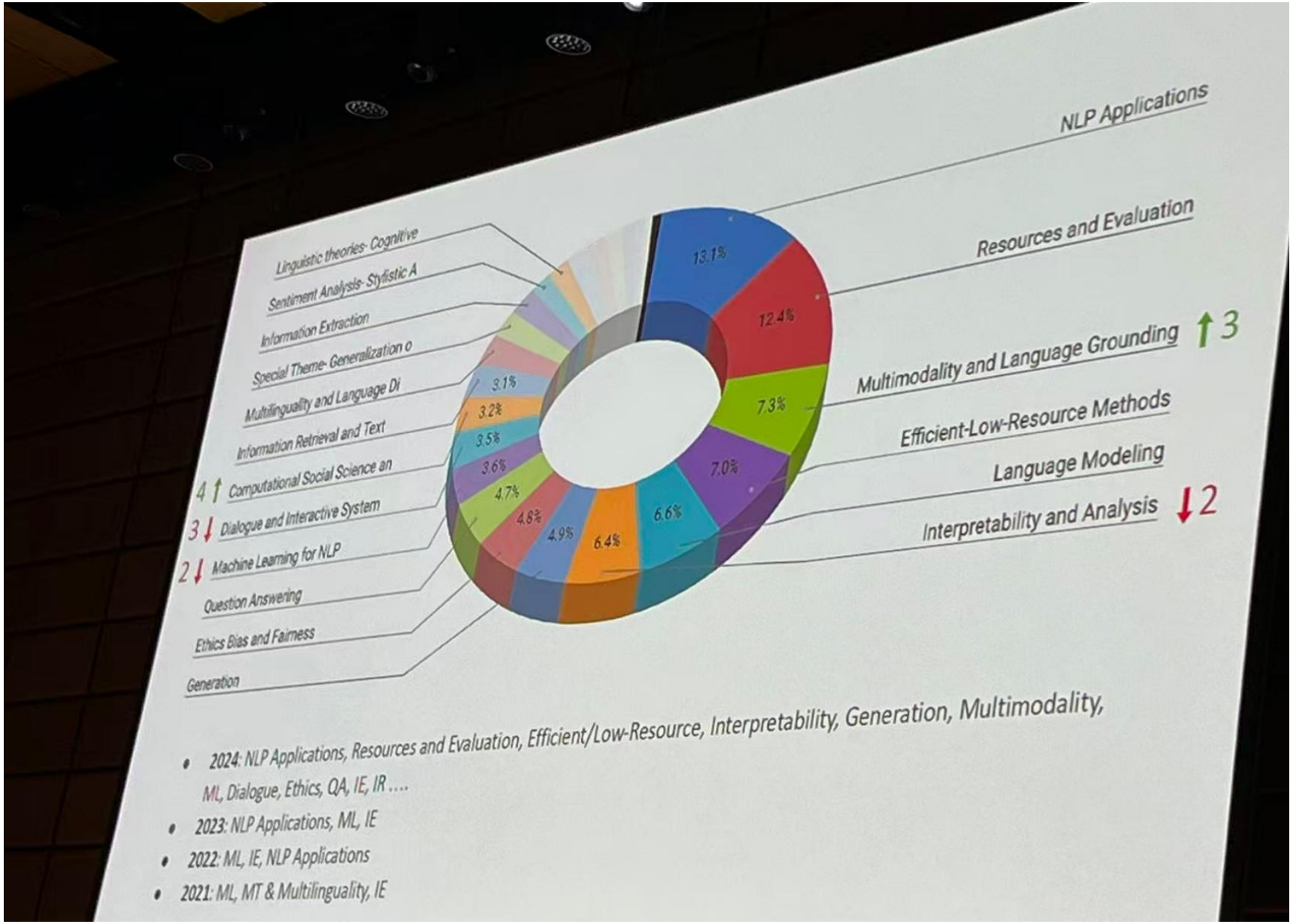
ACL 2025

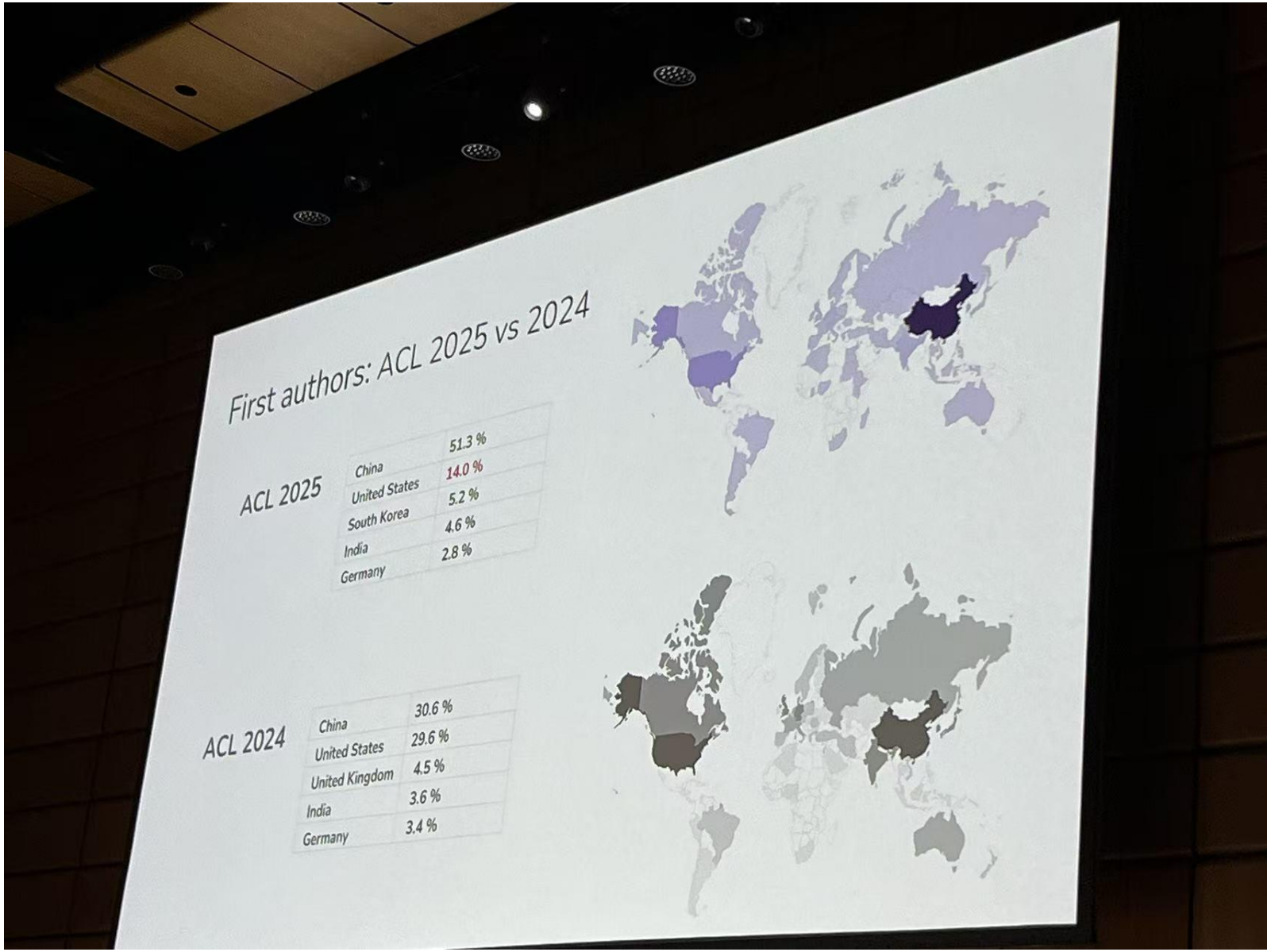

Submissions through ARR
Big thanks to ARR EiCs: Jun Suzuki, Jing Jiang, and Xiaodan Zhu
December 2024
(2,383 submissions)
February 2025
(8,350 submissions)
Commitment deadline
April 20th
(5,501 submissions)
~ 1600
~ 4700
~ 800

ACL 2025


Findings
Main
1,699, 1,392 of 8,360
committed
571 of 2,248
173 of 692
Findings (+16.7%)
Main (20.3%)
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025


ARR February 2025 (vs 2024)
8,350 Submissions +70% ↑ ACL'24
5,903 Reviewers +38% ↑
1,122 ACs +58% ↑
169 SACs +150% ↑
Desk reject +160% ↑
Template: 31%
Responsible checklist: 27%
Anonymity: 20%
Limitations section: 17%


NLP Applications
Resources and Evaluation
Multimodality and Language Grounding
Efficient-Low-Resource Methods
Language Modeling
Interpretability and Analysis
Linguistic theories- Cognitive
Sentiment Analysis- Stylistic A
Information Extraction
Special Theme- Generalization o
Multilinguality and Language Di
Information Retrieval and Text
Computational Social Science an
Dialogue and Interactive System
Machine Learning for NLP
Question Answering
Ethics Bias and Fairness
Generation
13.1%
12.4%
7.3%
7.0%
6.6%
6.4%
4.9%
4.8%
4.7%
3.6%
3.5%
3.2%
3.1%
3 2
4 3
2 2
• 2024: NLP Applications, Resources and Evaluation, Efficient/Low-Resource, Interpretability, Generation, Multimodality, ML, Dialogue, Ethics, QA, IE, IR ....
• 2023: NLP Applications, ML, IE
• 2022: ML, IE, NLP Applications
• 2021: ML, MT & Multilinguality, IE


First authors: ACL 2025 vs 2024
ACL 2025
China 51.3 %
United States 14.0 %
South Korea 5.2 %
India 4.6 %
Germany 2.8 %
ACL 2024
China 30.6 %
United States 29.6 %
United Kingdom 4.5 %
India 3.6 %
Germany 3.4 %

Some other statistics

- 67% of papers have "LLM" in their title/abstract
  - 9% "GPT"
  - 8% "LLaMA"
  - 2% "DeepSeek", "BERT, and "Gemini"/"Gemma"
- 50 authors with ≥ 10 papers
  - 23% of all authors ≥ 2 papers
- 250 papers with ≥ 10 authors
  - 20 single-author
- 65% with ":" in their title!

What's New This Year

- New ARR review form introduced in February cycle
- Separate sessions for Oral and Poster presentations
- Virtual Oral sessions live on zoom, **in parallel** to In-Person Posters
- **New presentation format**: Oral sessions with Panel Discussion
- **Evening session:** Findings with Reception 🥂🍹
- **Personalised programs** via Scholar Inbox - use this QR code

主旨演讲和圆桌论坛

- Rethinking Pretraining: Data and Architecture by Luke Zettlemoyer
- Whose Gold? Re-imagining Alignment for Truly Beneficial AI by Verena Rieser
- Generalization of NLP Models
  - Mirella Lapata, Dan Roth, Yue Zhang
  - Moderator: Eduard Hovy
  - Special Theme: “Generalization of NLP Models”

学生研讨会

- 隶属于主会的一个专门供学生投稿的工作坊，主题和写作均与主会相同，额外增加一个Thesis Proposal的赛道。
- 审稿流程
  - 预提交，由一个mentor提前给出指导意见
  - 修改后，再正式提交接受审稿
  - 没有Rebuttal，直接由主席来确定meta意见和录用结果
- 104/323（32.2%接受率）
- 丰厚的奖学金 14/104，每人约1500美金
- 更多信息： https://juniperliuzhu.netlify.app/blogs/SRW_intro_slides.pdf

Learning to Look at the Other Side: A Semantic Probing Study of Word Embeddings in LLMs with Enabled Bidirectional Attention

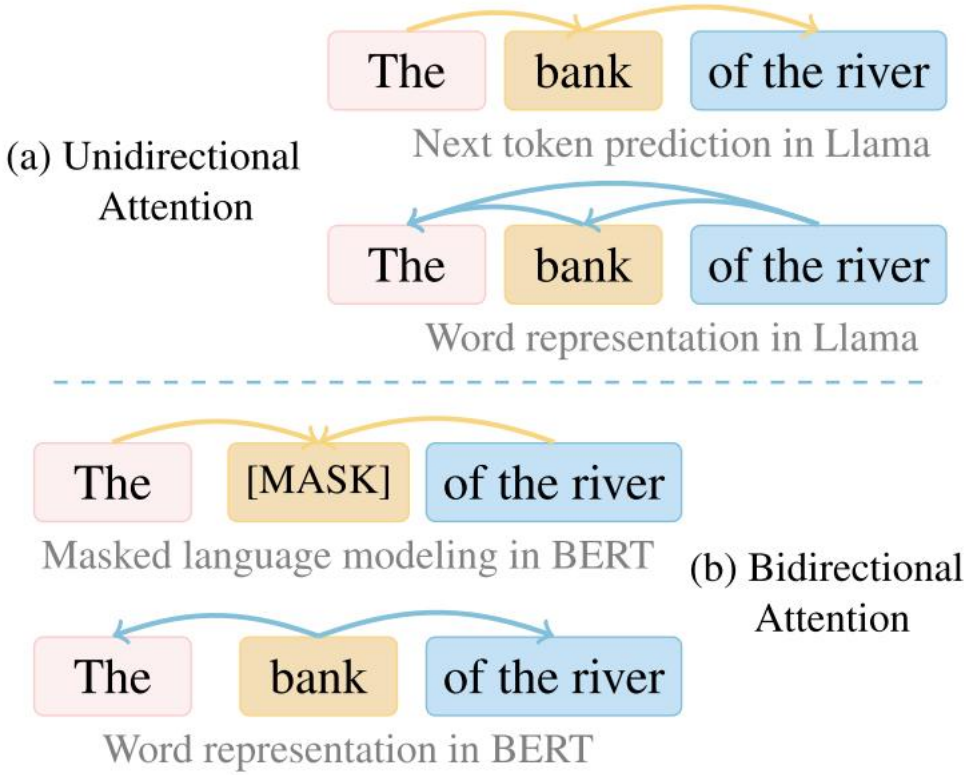
Zhaoxin Feng and **Jianfei Ma** and **Emmanuele Chersoni**
and **Xiaojing Zhao** and **Xiaoyi Bao**
Chinese and Bilingual Studies, The Hong Kong Polytechnic University
{zhaoxinbetty.feng,jian-fei.ma,xiaojing.zhao,xiaoyi.bao}@connect.polyu.hk,
emmanuele.chersoni@polyu.edu.hk

ACL 2025

https://github.com/Zhaoxin-Feng/semantic-probing-2025

背景

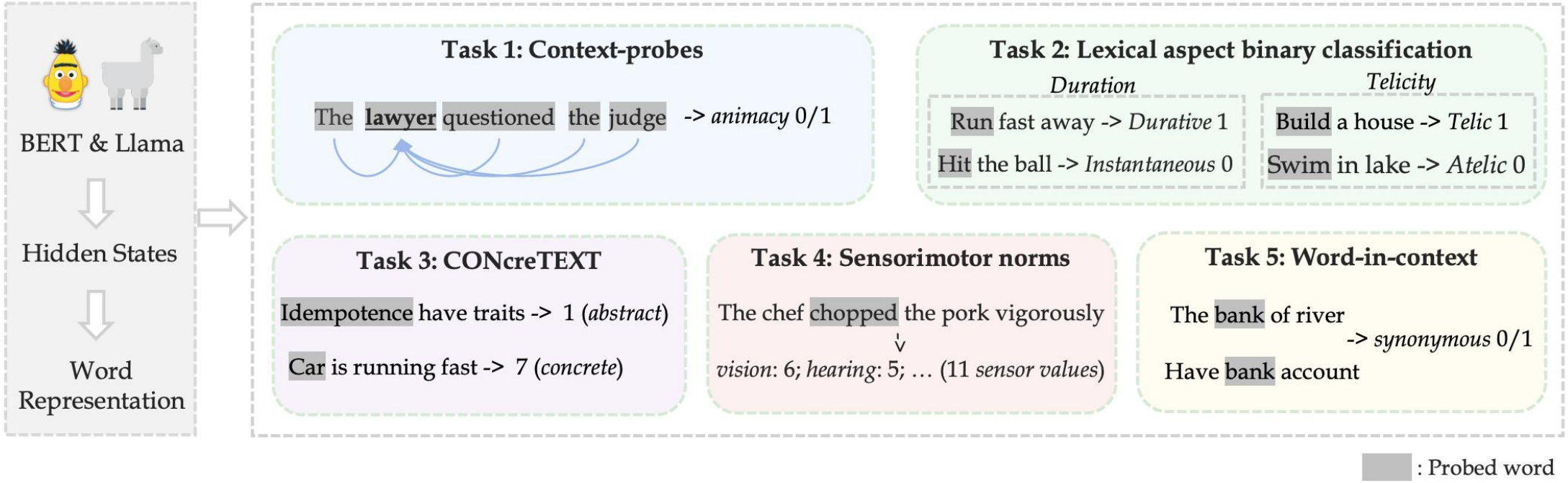
- 大语言模型在语义理解和生成任务上都取得很好的性能。
- 但是大语言模型的设计架构是“仅解码器”的，以匹配“预测下一个词”的预训练目标。
  - Decoder-only: 注意力模版是单向的（causal）
  - 只能关注到左边的上文，无法获取后文。
  - 这和传统的基于双向注意力的Bert模型有很大区别。



动因

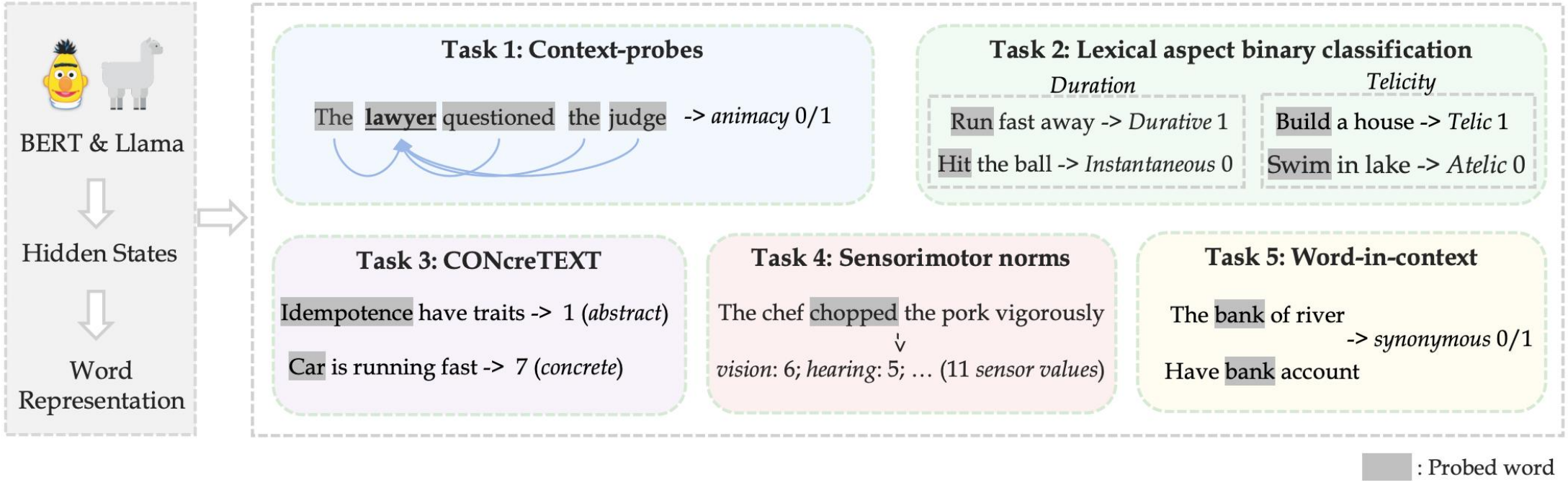
- 对于需要两边上下文理解的词义理解问题，这一架构可能存在缺陷
  - 词义理解往往涉及到本身的词汇语义以及携带的论元和其他修饰成分
  - 后文在一些词义消解中会起到重要作用(Zhu et al., 2024; Qorib et al., 2024)
- 解决方案（LLM2Vec BehnamGhader et al., 2024）
  - +后训练的过程中使用双向的注意力
  - +非监督/监督的对比学习
- 本文
  - 在更加多样的词义理解任务上评估上述方案
  - 更进一步分析前述方案（例如从熵的角度）

数据集



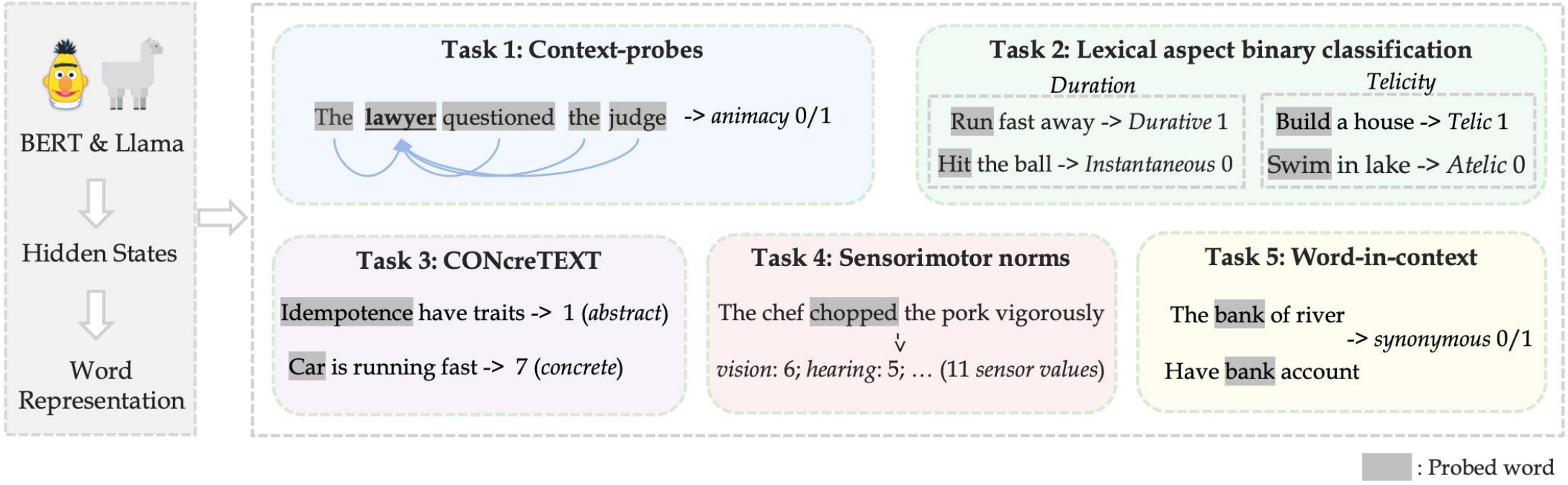
任务1: context-probes dataset by Klafka and Ettinger (2020) -- 二分类
数据：XX条由五个词构成的主谓宾简单句
标注了名词/动词的语义属性：名词的生命度，动词的静态还是动态，动词是否具有致使性

数据集



任务2: 动词的状貌情态 (Metheniti et al., 2022) -- 二分类
动词的持续性/瞬间性（duration）； 终结性/非终结性（telicity）

数据集

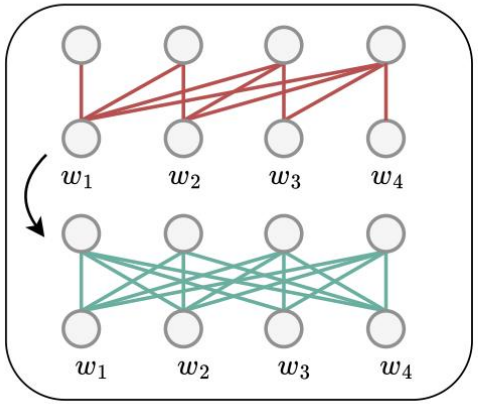


任务3: 名词的具体程度 (Gregori et al., 2020) & 动词的感官类别 -- 回归问题
任务5: 是否为同义词(Pilehvar and Camacho-Collados, 2019) -- 二分类问题

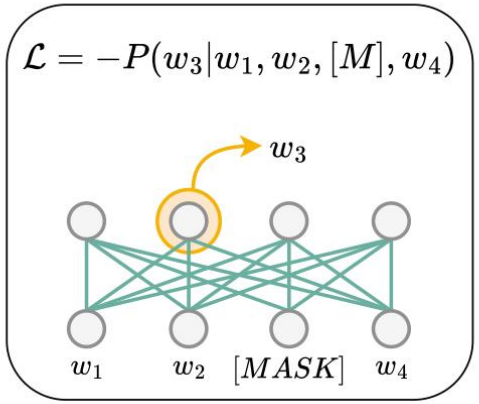
模型选择

- Llama 系列模型
  - Sheared-Llama
  - Llama2
- BERT 系列
  - Bert，Roberta
- Post-training 策略(BehnamGhader et al., 2024)
  - Bi+MNTP: 双向注意力+预测下一个词
  - Bi+MNTP+Contrastive Learning -> For sequential tasks
    - Supervised
    - Unsupervised (SimCSE)

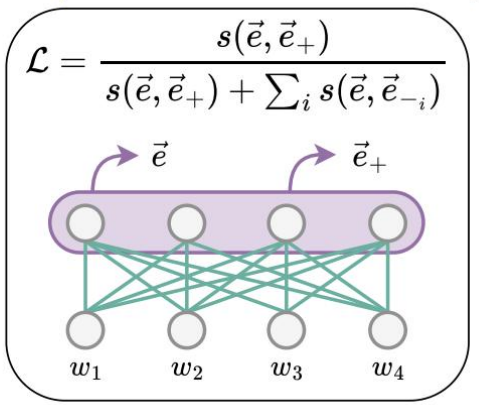
Enabling Bidirectional Attention



Masked Next Token Prediction



Unsupervised Contrastive Learning



向量提取

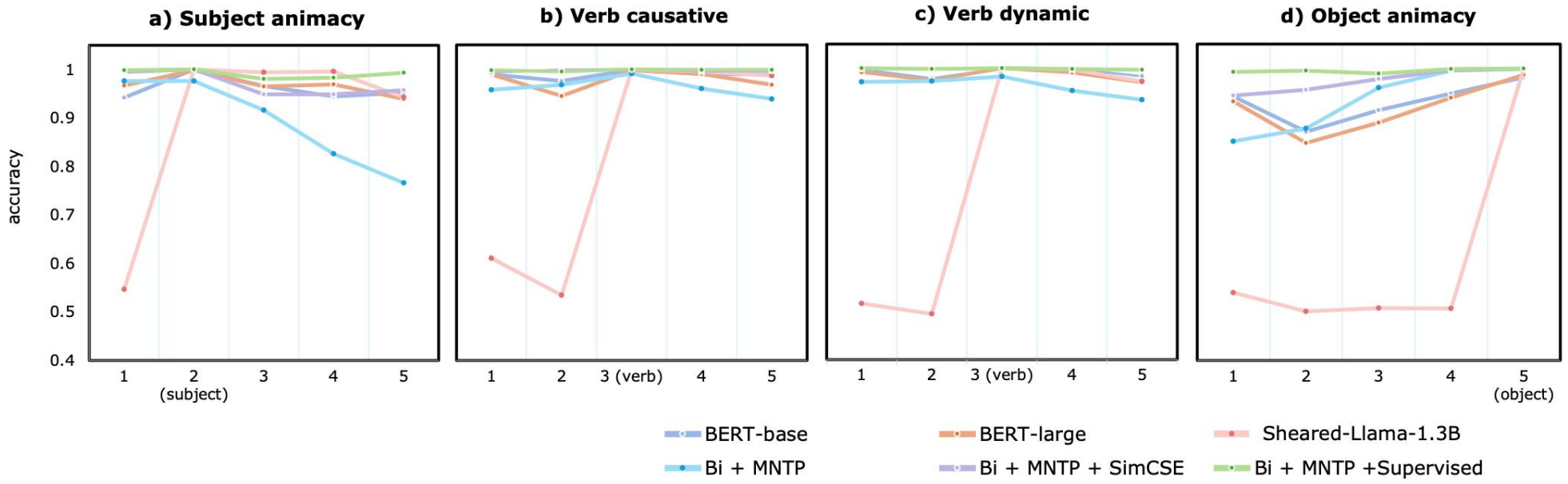
$$\mathbf{v}_w = \frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} \mathbf{h}_j$$

- 提取最高层的最后的隐状态作为最终的向量
- 多token的表征通过取平均来获取最终词的token

- 对于WiC任务，采用以下三种方式来处理向量对来作为分类器的输入
  - 余弦相似度
  - 逐元素做差后的绝对值
  - 追加

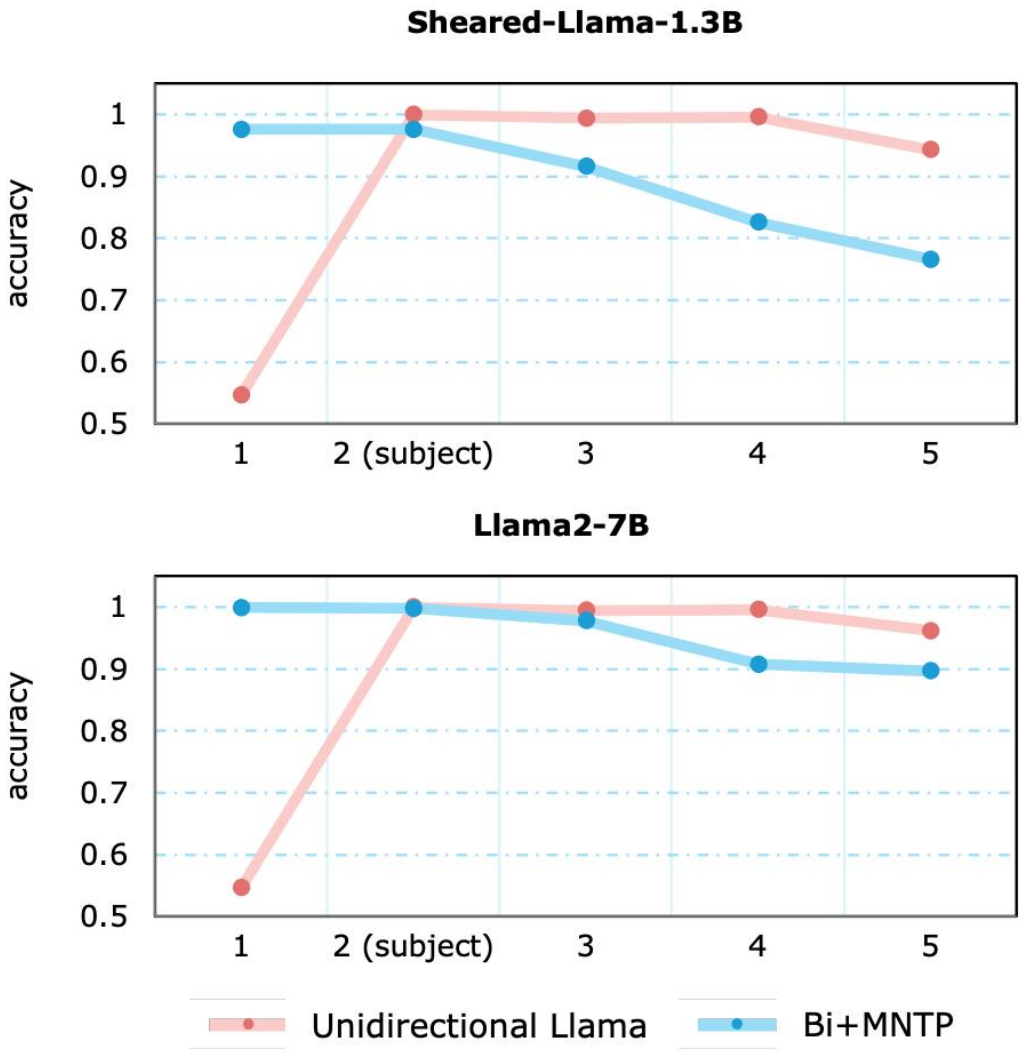
结果分析一

- 对比学习优于双向注意力
  - 两种方式都可以提高模型对于词义（上下文）的捕捉
  - 但是双向注意力的方式在提高对下文理解的同时会损害对于上文的理解



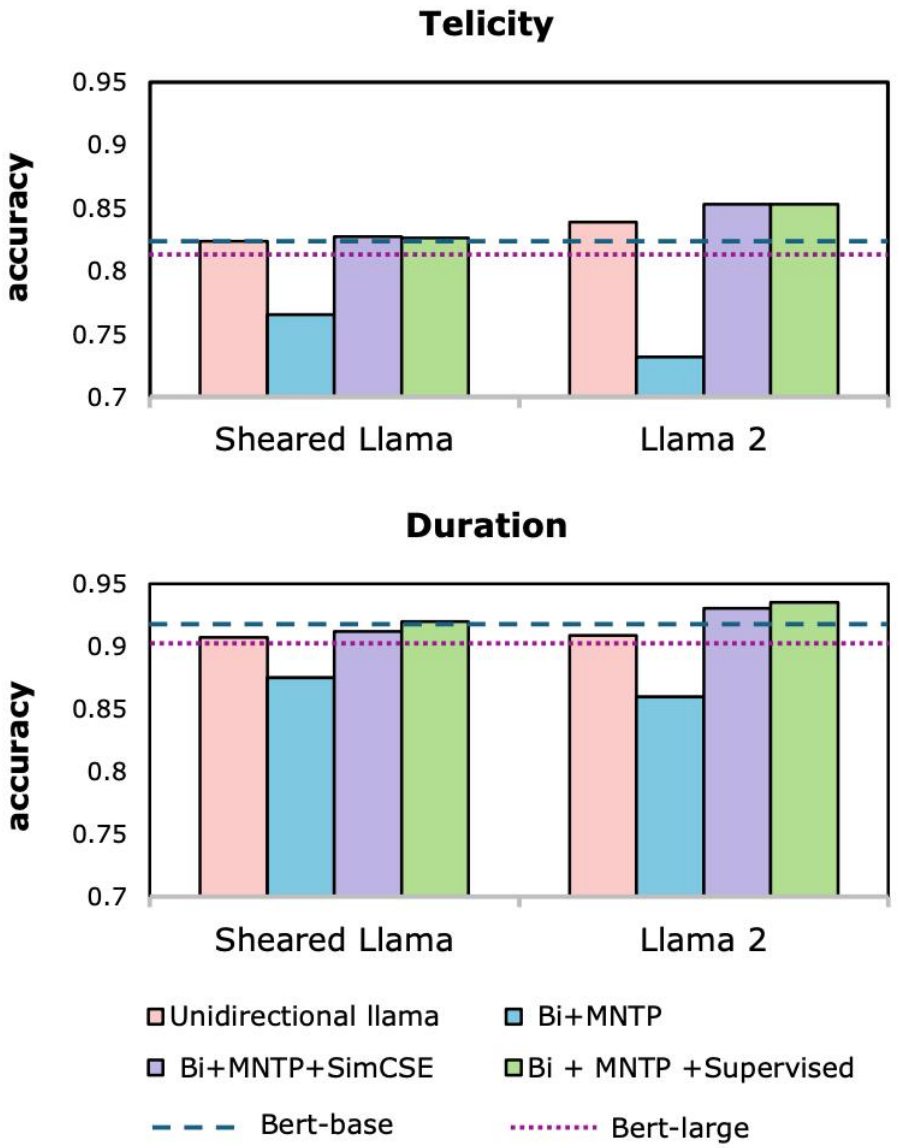
结果分析一

- 对比学习优于双向注意力
  - 两种方式都可以提高模型对于词义（上下文）的捕捉
  - 但是双向注意力的方式在提高对下文理解的同时会损害对于上文的理解
  - 模型大小呈现相似的趋势



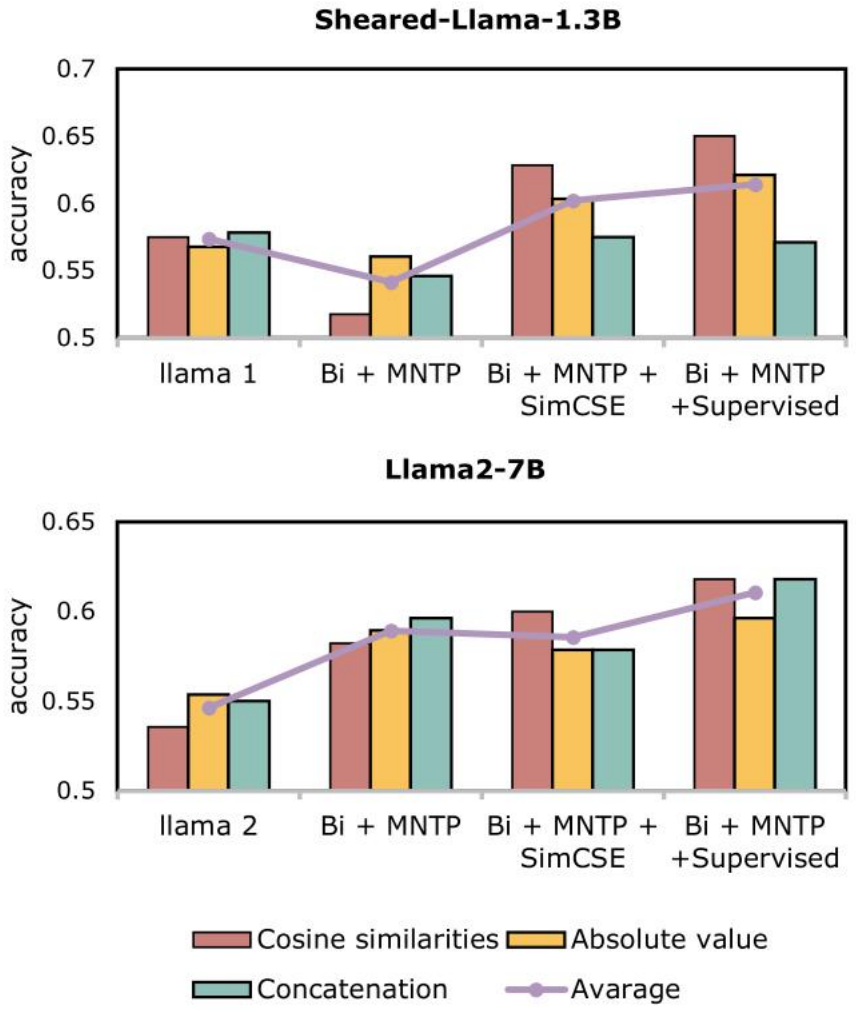
结果分析一

- 对比学习优于双向注意力
  - 两种方式都可以提高模型对于词义（上下文）的捕捉
  - 但是双向注意力的方式在提高对下文理解的同时会损害对于上文的理解
  - 对于状貌情态类的语义识别任务来说，Llama与BERT表现类似，但是双向注意力会损害性能

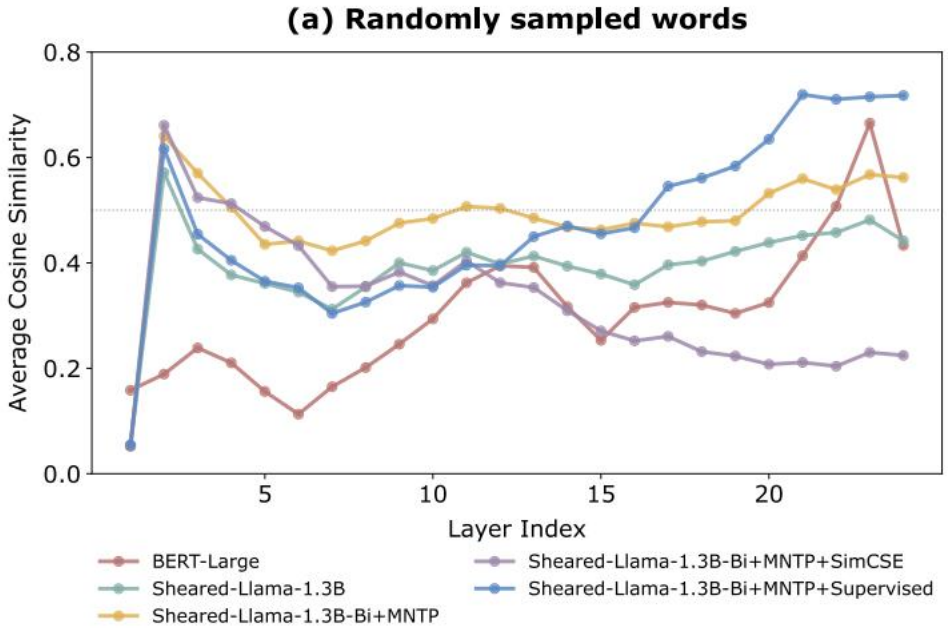


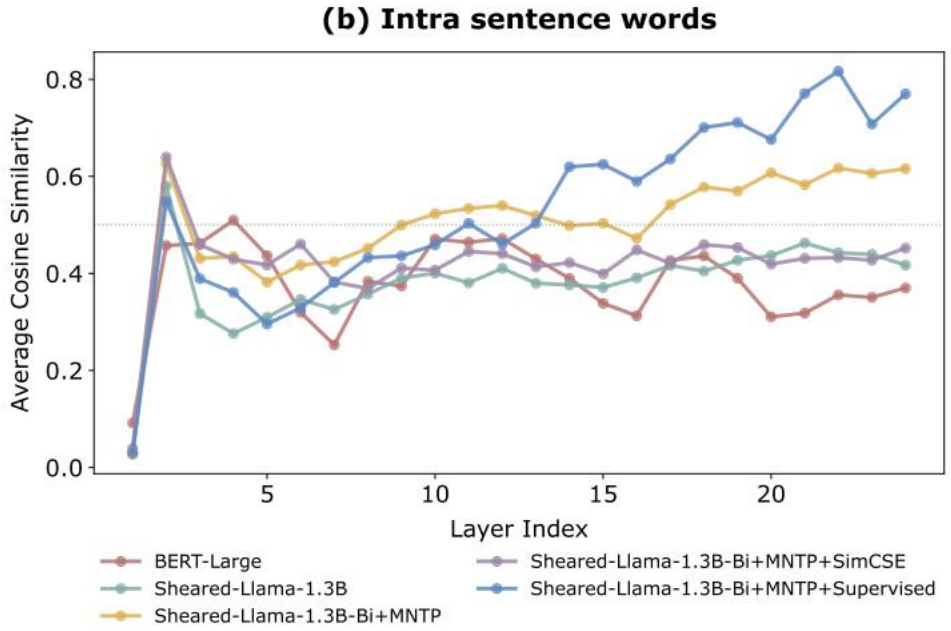
结果分析三

- 在词义消歧任务中
  - 对比学习起到了一致的增加效果
  - 双向注意力在不同的规模中表现不同
  - 不同的向量处理方式对于结果影响不大。



对于各向异性的影响





1. 各向异性是指向量的分布并不均匀，而是聚焦在某个角落里，以往的研究表明上下文向量具有各向异性。
2. 较大的各向异性体现在自回归模型Llama中，其中BI和Supervised CL最严重，而无监督的对比学习最轻微。

相关工作

结论

- 文章探究了双向自注意力如何影响模型对于上下文的编码。
- 在不同类型的词汇语义数据集上进行了详尽的测试。
- 探讨了不同的方式对于模型各向异性的影响。

Exploring Layer-wise Representations of English and Chinese Homonymy in Pre-trained Language Models

Matthew King-Hang Ma*♠, Chenwei Xie*♠, Wenbo Wang♣, William Shiyuan Wang♠
Research Centre for Language, Cognition, and Neuroscience
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
♠{khmma,cwxie,wsywang}@polyu.edu.hk
♣wenbo99.wang@connect.polyu.hk

ACL 2025 Findings

背景

- 词汇歧义性和多义性是词汇语义很重要的一方面
- 同形异义 (homonyms) vs. 多义词 (polysemy)
  - 同形异义词是指多个语义之间没有语义上或者根源上的关联
  - 例如 bank 既可以指一个金融机构，又可以指河岸
  - 在英文中标识为不同词类的词义往往是同形异义词，而在中文中由于大量兼类词的存在以及不同词类没有明显的屈折变换，词类并非是断定同形异义词的最佳手段。
  - 心理和神经语言学的研究也表明人类在处理同形异义词的时候也会更加困难（具有更强的神经激活和响应时间）(Frazier and Rayner, 1990; Rodd, 2018; Huang and Lee, 2018)

背景

- 词向量对于词多义性的捕捉也有困难
  - 静态词向量包括 LSA 和 Word2Vec -》 无法直接捕捉上下文
  - 上下文词向量：ELMo和BERT
  - 大语言模型
- 词类对于同形异义词的影响也很大
  - 例如英文的fly可以作为名词“苍蝇”也可以作为动词“飞”
  - 神经语言学方面也通过电极反应和血氧水平给予了佐证(Federmeier et al., 2000; Gennari et al., 2007)
- 本文从中英对比的角度观察同义词在LLM中的表征如何反应词义

贡献

- 构建了一个全新的中文数据集
- 对中英两种语言的同形异义词在不同模型中逐层的表征进行分析
- 探究了影响同形异义词的因素（例如词类）

数据集构建

- 英文数据集
  - British eDom Norms
  - 100 个拥有两个词义的同形异义词
  - 利用GPT-4o来产生句子，每个词义有10个句子
- 中文数据集
  - 首先从现代汉语词典中挑选出100个同形异义词
  - 之后利用ChatGLM4来产生句子，每个词义有10个句子
- 专家之后进行审核，一半句子对的词类相同，一半的不同

	prompt
Human expert	你是一名语言学家，非常擅长回答与词汇歧义相关的问题， 你的任务解释同形词的意思并用同形词造句。 请问“上任”是个同形词吗?
GLMchat	是的，“上任”是一个同形词。 它有两个意思，一个是“动词，指官吏就职”， 一个是指“名词，称前一任的官吏”。
Human expert	请分别用“上任”造二十个句子， 其中十个句子表达“动词，指官吏就职”的意思， 另外十个句子表达“名词，称前一任的官吏”的意思。
GLMchat	以下是使用“上任”这个词来表示“动词，指官吏就职”的10个句子： 1. 他刚上任不久，还在熟悉新的工作环境。 2. 市长上任后，立即开始着手解决交通拥堵问题。 3. 新上任的校长对教育改革充满热情。 4. 经过严格的选拔，她终于在上任成为了部门的负责人。 5. 他上任的第一天，就给全体员工开了一个鼓舞人心的会议。 6. 这位年轻的法官上任后，受到了社会各界的关注。 7. 上任仅一周，他已经展现出了出色的领导能力。 8. 她上任后，推动了公司的技术创新。 9. 他刚上任就被派去处理一个棘手的案子。 10. 上任以来，他一直致力于改善市民的生活质量。 以下是使用“上任”这个词来表示“名词，称前一任的官吏”的10个句子： 1. 我们需要分析上任领导留下的政策，看看有哪些需要改进。 2. 上任市长在位期间，城市的基础设施得到了很大的改善。 3. 新市长表示，他将继承上任市长的良好传统。 4. 上任经理的离职，给公司带来了不少挑战。 5. 他试图了解上任留下的工作计划，以便更好地继续推进项目。 6. 上任部长在任时，推出了一系列受欢迎的教育政策。 7. 员工们对上任领导的离开感到非常惋惜。 8. 这座大桥是上任政府投资建设的。 9. 她在评价上任领导的工作时，给予了高度的评价。 10. 上任领导在离任前，留下了一套完善的管理体系。

模型和特征提取

- 来自不同模型种类的共21个模型
  - BERT GPT-2 Llama3 Qwen
  - 选择的原因
    - 开源
    - 是预训练模型和大语言模型的代表
    - 可以获取不同规模的模型大小
    - Qwen针对汉语性能较好
- 提取方式
  - 最常规的设置：提取模型最后一层的表征

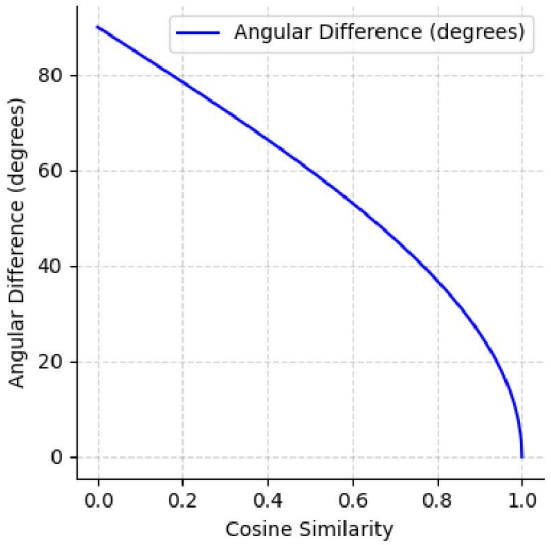
衡量标准

- 期待模型对于两个不同语义之间上下文中所得表征的相似度要尽可能低，在相同语义上下文的相似度尽可能高
- 相似度
  - 将余弦相似度换为夹角相似性，因为余弦相似性与夹角的变化呈现非线性

$$AngSim = 90 - \arccos(CosSim) \times \frac{180}{\pi} \tag{1}$$

$$AngSim_{same}(l, w) = \mathbb{E}\left[\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \angle(f_l(w_i^s), f_l(w_j^s))\right] \tag{2}$$

$$AngSim_{cross}(l, w) = \mathbb{E}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} \angle(f_l(w_i^1), f_l(w_j^2))\right] \tag{3}$$

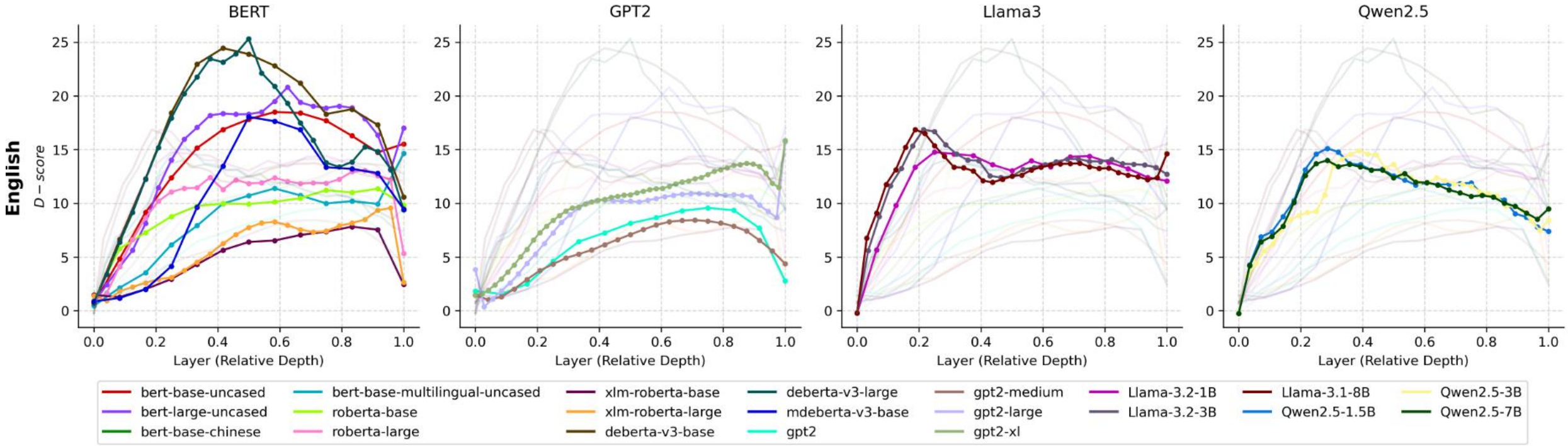


衡量标准

- 消除各向异性的影响
  - 归一化：先随机计算两个样本的相似度，所有样本减去这个值
- 消歧评分
  - 从0-90

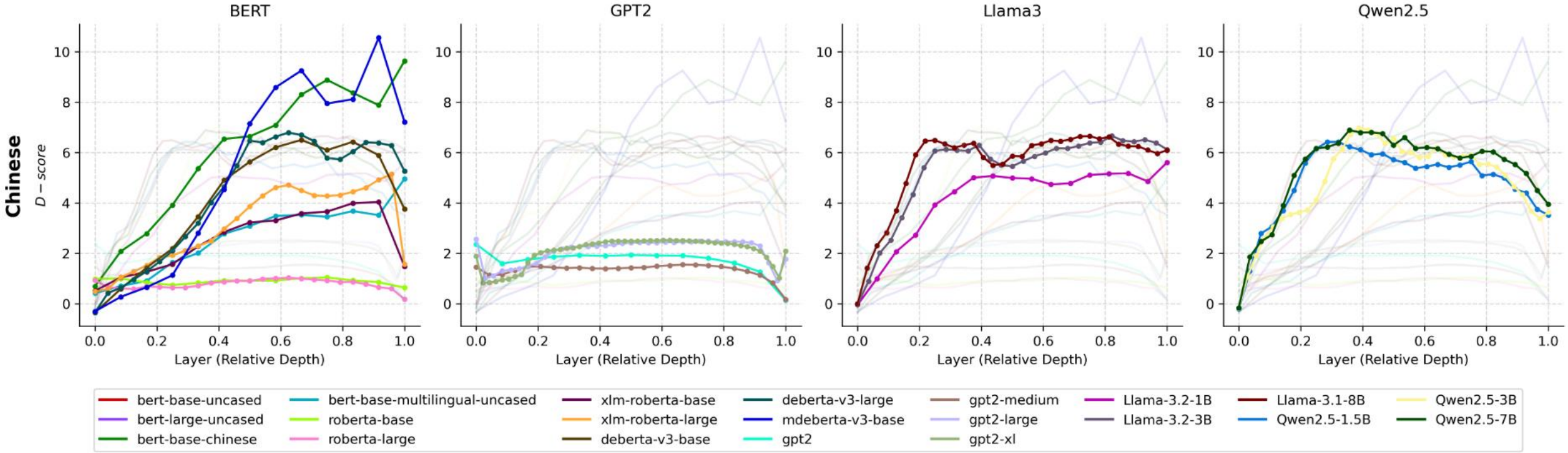
$$D\text{-}score = AngSimAdj_{same} - AngSimAdj_{cross} \tag{4}$$

分析



- 对于英文来说，BERT仍旧可以取得很好的得分
- deberta-v3-large取得了最佳得分，为25.32
- BERT（以及Llama3；Qwen2.5）普遍在中层达到较好的效果，而GPT2则主要在末层

分析

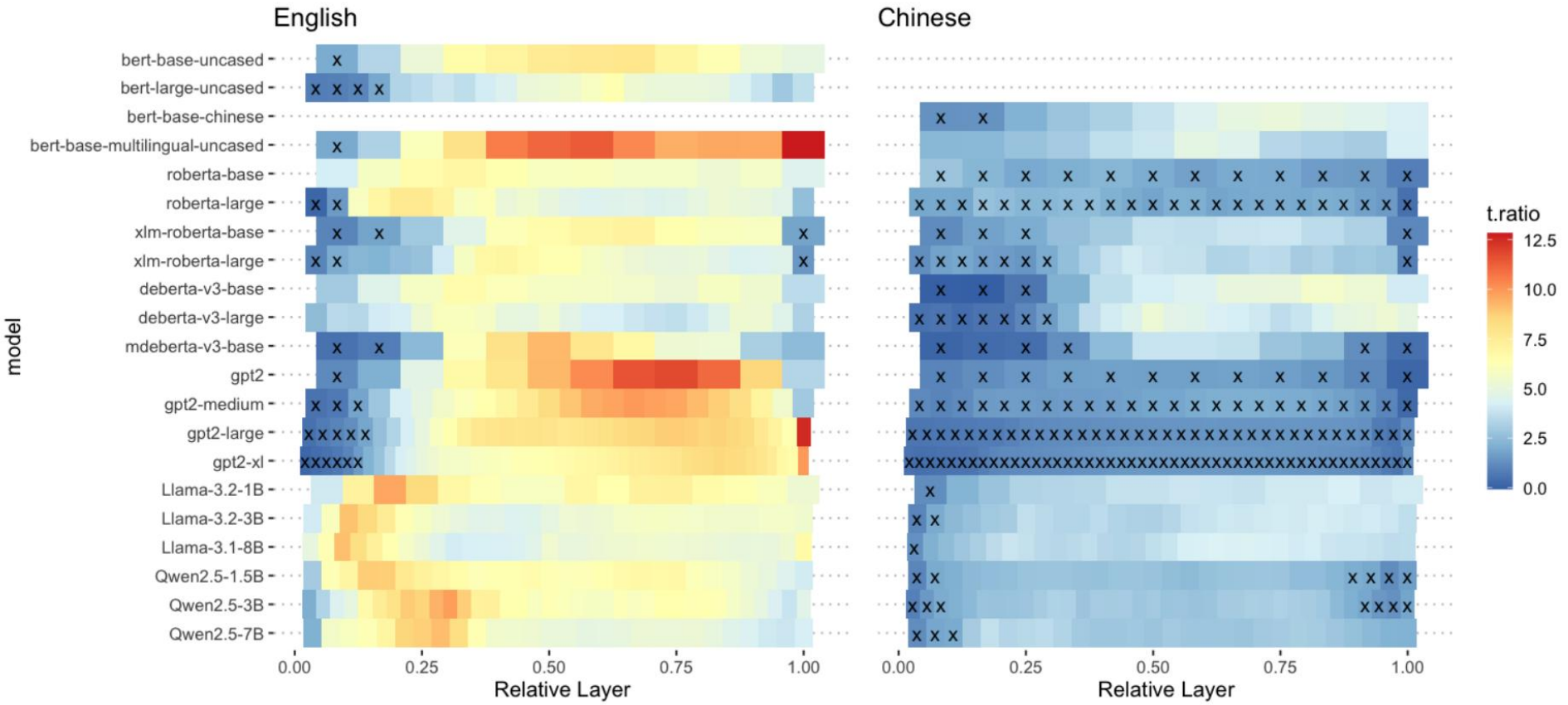


- 对于中文来说，BERT仍旧可以取得很好的得分，但要比英语更差
- mdeberta-v3-base and bert-base- Chinese都有比较好的得分
- 最佳层数出现混合的情况，BERT并没有很明显得出现在中层

			English				Chinese			
Model Family	Model	Parameters	Layer	Layer Depth (%)	Layer Depth	*D-score*	Layer	Layer Depth (%)	Layer Depth	*D-score*
BERT	bert-base-uncased	110M	7.00	58.30	middle	18.50	—	—	—	—
	bert-large-uncased	340M	15.00	62.50	middle	20.81	—	—	—	—
	bert-base-chinese	102M	—	—	—	—	12.00	100.00	higher	9.63
	bert-base-multilingual-uncased	167M	12.00	100.00	higher	14.65	12.00	100.00	higher	4.96
	roberta-base	125M	11.00	91.70	higher	11.36	9.00	75.00	higher	1.04
	roberta-large	355M	21.00	87.50	higher	13.04	15.00	62.50	middle	1.03
	xlm-roberta-base	278M	10.00	83.30	higher	7.82	11.00	91.70	higher	4.04
	xlm-roberta-large	560M	23.00	95.80	higher	9.57	23.00	95.80	higher	5.15
	deberta-v3-base	183M	5.00	41.70	middle	24.44	8.00	66.70	middle	6.50
	deberta-v3-large	434M	12.00	50.00	middle	**25.32**	15.00	62.50	middle	6.79
	mdeberta-v3-base	278M	6.00	50.00	middle	18.04	11.00	91.70	higher	**10.56**
GPT2	gpt2	124M	9.00	75.00	higher	9.57	6.00	50.00	middle	1.93
	gpt2-medium	355M	17.00	70.80	higher	8.44	16.00	66.70	middle	1.55
	gpt2-large	774M	36.00	100.00	higher	15.72	26.00	72.20	higher	2.49
	gpt2-xl	1.5B	48.00	100.00	higher	**15.83**	29.00	60.40	middle	**2.53**
Llama3	Llama-3.2-1B	1B	4.00	25.00	lower	14.77	16.00	100.00	higher	5.61
	Llama-3.2-3B	3B	6.00	21.40	lower	16.86	23.00	82.10	higher	**6.66**
	Llama-3.1-8B	8B	6.00	18.80	lower	**16.87**	24.00	75.00	higher	6.65
Qwen2.5	**Qwen2.5-1.5B**	1.5B	8.00	28.60	lower	**15.11**	9.00	32.10	lower	6.43
	Qwen2.5-3B	3B	14.00	38.90	middle	14.89	14.00	38.90	middle	**6.99**
	Qwen2.5-7B	7B	8.00	28.60	lower	13.99	10.00	35.70	middle	6.89

模型规模 在BERT GPT 中很明显，但是在Llama3中出现了相反的情况；结构来看PLM要好于LLM

词类的



不同POS评分减去相同POS的差异显著性

- 正值反映了不同POS的区分要更加简单，这和神经语言学的一些结论相反(Grindrod et al., 2014)
- 汉语在不同层之间显示出更突出的不显著，说明词类特征对于汉语来说不如英语明显。

结论

- 在汉语和英语的同形异义词上进行了比较和分析
- 探索了21个语言模型
- 探讨了模型架构、词类特征等对语义区分的影响

参考文献

清華大學
Tsinghua University

Q & A

谢谢

Note