上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

语言学为大语言模型带来了什么？从语义表征、多语智能到认知机制

聂耳聪
上海交通大学
https://ercong21.github.io/

大语言模型与数字人文的方法和应用：
上海交通大学首届数字人文研讨会
（中国·上海）
2026年7月17日

上海交通大学外国语学院
School of Foreign Languages

个人简介

聂耳聪 **上海交通大学外国 语学院 助理教授**

- 以summa cum laude成绩获德国慕尼黑大学**计算语言学**博士学位。
- 研究方向为**自然语言处理**和**语言科学智能**，聚焦**语言科学与人工智能的双向 赋能**，尤其关注多语智能、大语言模型可解释性、脑—模型语言认知对齐等领域。
- 在ACL、EMNLP等人工智能顶会和ACM TALLIP、Information Fusion、Digital Scholarship in the Humanities等SCI/SSCI期刊发表/录用论文二十余篇。
- 担任国际计算语言学协会年会（ACL）领域主席和ICLR MemAgents Workshop程序主席，以及ICML、NeurIPS等顶会和IEEE TNNLS、ACM TIST等多本SCI/SSCI期刊的审稿人。

https://ercong21.github.io/

引子：一个语言学细节，成为AI顶会最佳论文
2026年了，AI顶会竟还在关注经典语言学？

ACL： 语言智能领域顶级会议，也是该领域唯一的CCF-A类会议

（整个人工智能领域仅有的7个A类会议之一）

ACL 2026 主会论文接受率约18.9%，年度主题聚焦NLP模型可解释性。

一篇以经典语言学问题“*未完成体悖论*”为基础的论文从逾万篇投稿中脱颖而出，荣获最佳论文奖。

颁奖词

Best Paper Awards 1/3

- **The Imperfective Paradox in Large Language Models**
  Bolei Ma, Yusuke Miyao

Justification:
*This paper makes **careful and insightful use of linguistic theory to reveal substantive limitations in the language understanding and reasoning abilities of contemporary LLMs**. Through the imperfective paradox, the authors show that many open 7B–9B-scale models systematically violate basic event-entailment patterns in English, behaving more like predictive narrative engines than faithful logical reasoners, even when given explicit contextual cues. The paper is especially valuable because it goes beyond documenting failure: it shows a dissociation between representation and reasoning, identifies promising effects of scale, and opens an important path for studying how logical reasoning about events can emerge or be induced in language models.*

(M)

这篇论文细致而深刻地运用**语言学理论**，揭示了当代大语言模型在语言理解与推理能力方面存在的实质性局限。通过对“**未完成体悖论**”（**imperfective paradox**）的研究，作者表明，许多开放权重的 7B–9B 规模模型会系统性地违反英语中基本的事件蕴含模式。即便在给出明确上下文线索的情况下，这些模型的表现也更像是“**预测性的叙事生成引擎**”，而不是忠实的逻辑推理者。该论文尤其有价值之处在于，它并不只是记录模型失败现象，而是进一步揭示了表征与推理之间的分离，发现了模型规模可能带来的积极影响，并为研究语言模型中关于事件的逻辑推理能力如何涌现或被诱导提供了重要路径。

“未完成体悖论”暴露了大语言模型的语义盲区

Class	Examples (Premise → Hypothesis)	Entail?
Telic	**P:** The carpenter *was building* a gazebo. **H:** The carpenter *built* a gazebo.	✗ No
Atelic	**P:** The boy *was running* in the park. **H:** The boy *ran* in the park.	✓ Yes

Table 1: Examples for the Imperfective Paradox. Telic verbs do not entail completion in the progressive form.

无界（atelic）事件：
He was running ⇒ He ran（通常可蕴含）
有界（telic）事件：
He was building a house ⇏He built a house（未必完成）

大语言模型倾向：对目标导向事件产生“完成幻觉”或偏差
启示：语言学现象（语法范畴，数、时、态、体等）可以成为大模型能力诊断的探针

发言提纲

一、导言：语言科学智能

二、分词与语义表征：语言学帮助模型 迈出理解语言第一步

三、语言科学帮助构建更好的大 语言模型：以多 语智能为例

四、理解智能：语言科学为解释大语言模型提供 认知参照

五、小结与展望

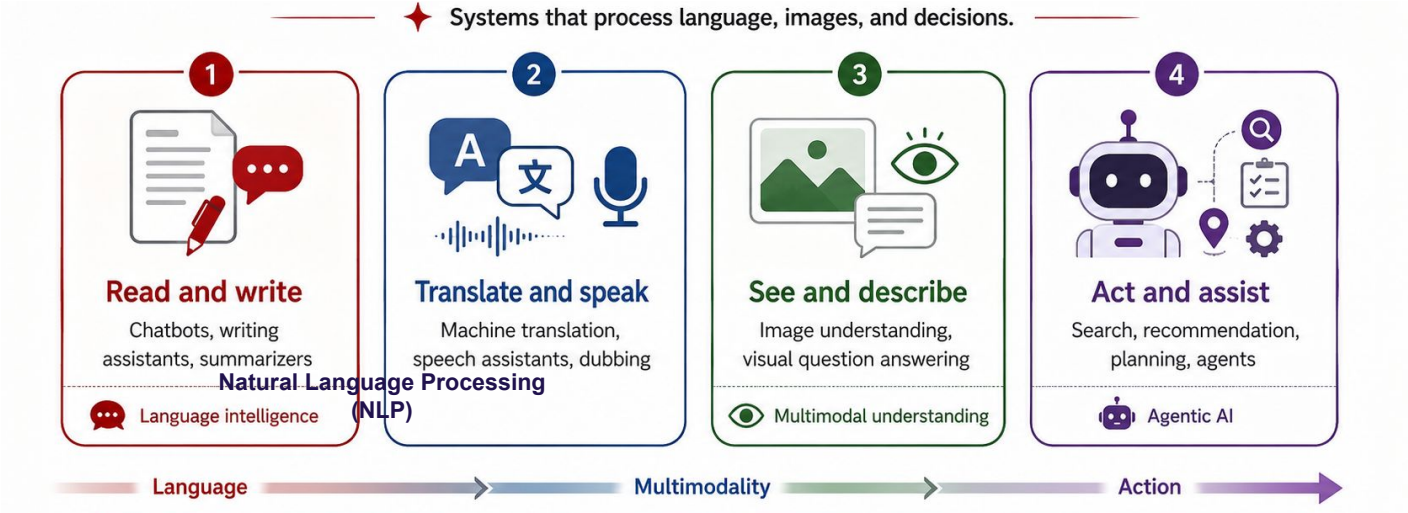
第一部分

导言：语言科学智能

从自然语言处理到人工智能

自然语言处理（Natural Language Processing，NLP）被誉为“人工智能皇冠上的明珠”。

脱胎于自然 语言处理的大 语言模型引 领生成式 AI的突破



以大语言模型为基础的生成式人工智能迅速发展，并进一步向具有**规划**、**工具调用**和**持续交互**能力的**智能体系统**延伸。

在这一持续演进过程中，NLP不再只是人工智能体系中的一个专门分支，语言也不再只是特定任务中的输入输出形式，而日益成为连接知识表征、指令理解、内容生成、交互协同与行动组织的关键智能媒介。

为什么从“语言”进入人工智能？

由此值得追问：

为什么引发这一轮通用生成智能突破的，首先是**语言**模型？为什么基于语言材料的大规模预训练，能够表现出如此显著的泛化、迁移与能力涌现，而图像、音频等其他模态数据的大规模训练未能首先涌现出“智能”？

理解语言，是理解人类智能和构建人工智能的重要入口。

智能即压缩？
语言是智能的压缩介质吗？

上海交通大学外国语学院
School of Foreign Languages

1. 不能只用“规模扩张”解释成功

- 语言模型的突破，难以仅归因于算力扩张、参数增长或数据规模扩大。

2. 语言不是中性的符号外壳

- 语言是高度组织化的意义系统。
- 它承载人类对世界的分类方式、概念结构与社会文化经验。

3. 语言压缩了多层次的人类智能信息

知识结构　交际结构　认知模式　文化经验

大语言模型处理的并非单纯字词序列，而是一个高度压缩的复杂系统。

4. 这解释了大模型为何具有一定通用性

生成　推理　问答　协作　行动规划

因此，模型能够学习到超出传统任务边界的表征能力。

Published as a conference paper at ICLR 2024

谷歌

LANGUAGE MODELING IS COMPRESSION

Grégoire Delétang*[1]　Anian Ruoss*[1]　Paul-Ambroise Duquenne[2]　Elliot Catt[1]
Tim Genewein[1]　Christopher Mattern[1]　Jordi Grau-Moya[1]　Li Kevin Wenliang[1]
Matthew Aitchison[1]　Laurent Orseau[1]　Marcus Hutter[1]　Joel Veness[1]

语言可以被理解为人类智能的一种高密度压缩介质。

语言模型之所以展现较强的通用能力（推理、泛化），不只是因为规模更大，可能是因为语言本身已经压缩了知识、文化、认知等智能来源。

维特根斯坦:“语言的边界即世界的边界”

赋能这一轮人工智能变革的核心——大语言模型——也正是在语言能力上呈现出了“涌现”能力，并在语言能力的基础上延伸到了推理、多模态、具身等多种智能能力。为何深度神经网络模型通过scale up，在大规模语言文本数据上自监督训练，可以获得强大的涌现和泛化能力，但是在图像、视觉、音频等其他模态的大规模训练上并没有出现智能的涌现？为了更好地理解这个问题，我们必须把注意转回“语言”本身。维特根斯坦说过，语言的边界即世界的边界。也许语言并不是一套由单纯的离散符号组成的系统，语言本身也许就是一种智能的压缩，是我们人类在几千年的文明演变和与世界交互中所凝结成的智能系统，包含我们对世界的认知，包含不同的文化，包含思考问题的方式，只是以一种离散符号系统的方式呈现出来。正因如此，大语言模型通过自回归的方式习得的可能不仅是一套统计概率模型和匹配范式，而是一套压缩的智能系统。

什么是“语言科学智能”？

语言科学智能可以理解为：

以语言科学理解智能，
以人工智能拓展语言科学。

语言科学 → AI

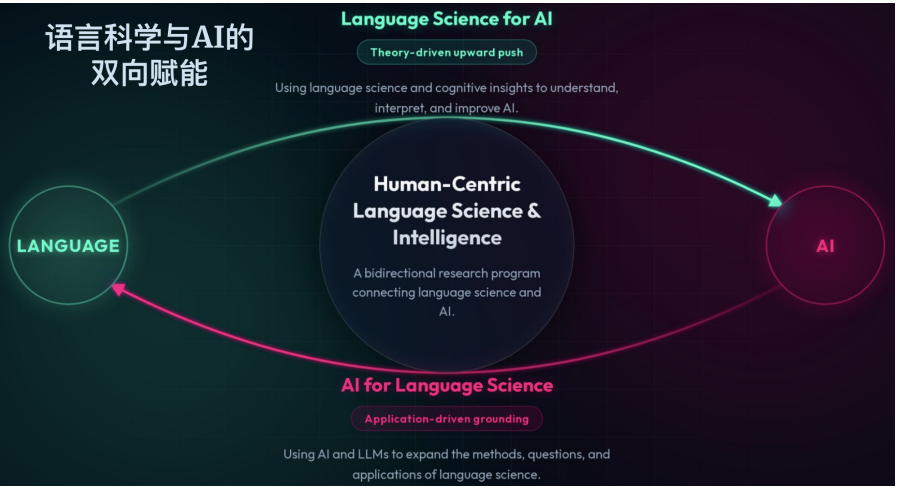
- 解释语言模型内部机制
- 诊断多语言与跨文化偏差
- 设计更符合人类语言规律的评测与方法

AI → 语言科学

- 帮助研究语言、文化与认知
- 支持多语言资源建设
- 赋能语言人文与社会科学研究

语言科学智能关注的不只是“让机器处理语言”，
而是理解语言、人类与智能系统之间的关系。

Language Science for AI & AI for Language Science



AI时代语言科学被赋予了新的任务：
语言不再只是被处理的对象，
而是连接智能的媒介，理解智能的入口。

→ **Language Science for AI**

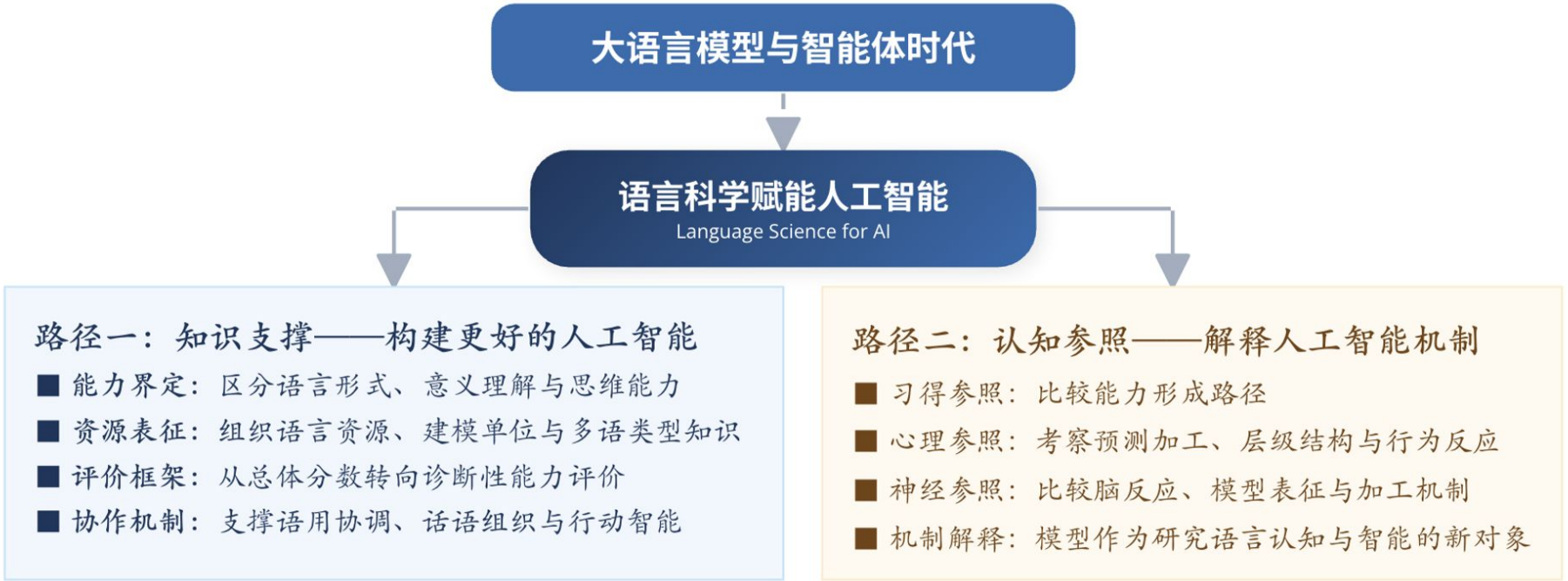
“语言科学智能”这个说法，听起来有一点新，但它背后其实连接着几个已经非常重要的方向。首先，语言科学智能可以理解为AI for Science（科学智能）下面的一个子集，即“AI for Language Science”；其次，语言科学智能也可以理解为语言科学与语言智能（Language Science and Language Intelligence，或简单理解为Language Science & Intelligence）；而本人则更倾向于从语言科学与AI双向赋能的角度来全面看待语言科学智能。

一方面，它可以理解为AI for Language Science，也就是用人工智能推动语言科学研究。过去，语言学、语料库语言学、心理语言学、神经语言学、翻译研究、文学研究、跨文化传播研究等领域，很多工作依赖人工阅读、人工标注、人工归纳和小规模实验。今天，大语言模型、自然语言处理、多模态分析和智能体技术，正在让我们有可能以新的方式处理海量文本、构建语料资源、分析语言现象、设计语言实验，甚至辅助完成文献综述、数据标注和研究流程管理。

另一方面，它也可以理解为 Language Science for AI，也就是用语言科学理解、解释和改进人工智能。大语言模型虽然表现强大，但它并不总是可靠。它可能在不同语言之间产生事实不一致，可能在跨文化表达中出现误读，也可能在看似流利的回答中隐藏偏见、混淆和不确定性。要理解这些问题，不能只看模型最后答对了没有，还要追问：模型内部如何组织语言和知识？它如何在不同语言之间迁移概念？它为什么会发生语言混淆？它的语言能力和人类语言理解有什么相似，又有什么不同？

因此，我所理解的“语言科学智能”，不是简单地“用 AI 做语言研究”，也不是传统意义上单纯面向工程应用的“语言智能”。它更像是一个双向框架：用人工智能推动语言科学研究，也用语言科学解释和改进人工智能。更简洁地说，就是：用 AI 研究语言，用语言科学理解 AI。

AI(大语言模型和智能体)时代下：语言科学如何赋能人工智能？



第二部分

分词与语义表征：
语言学帮助模型迈出理解语言第一步

语言学在不同阶段以不同方式参与AI演进

作用方式变迁

1 规则时代

早期AI / NLP

语言学知识显性编码

- 形式语言理论、生成语法
- 词典、语法规则、句法分析
- 机器翻译与规则转换系统

2 统计时代

语料与概率建模

数据驱动的语言学

- 双语语料与 IBM 统计机器翻译
- 语料建设、切分与标注规范
- 词频、共现、Zipf 等分布规律

3 神经网络与预训练

语义表征学习

分布语义学奠基

- “意义来自使用” → 向量空间
- 词向量、上下文表示、BERT
- BERTology：探测内部语言知识

4 大语言模型与智能体

交互、工具与行动

语言成为智能协作媒介

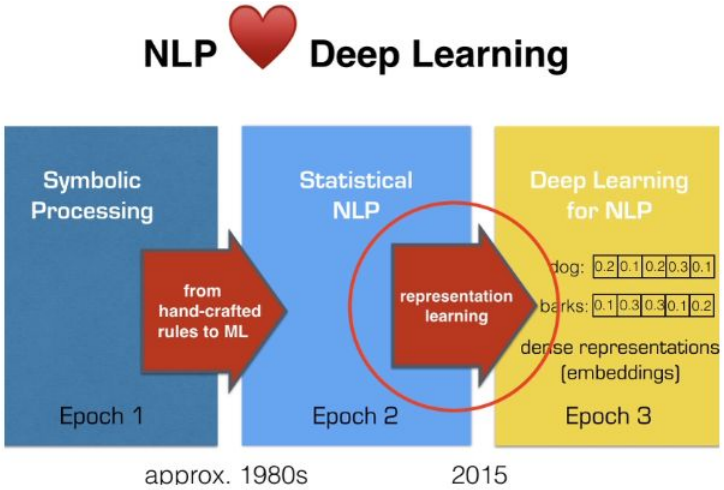
- 表达目标、组织行动、协调任务
- 语用协调、社会适配
- 智能评估、解释与人机协作

语言学对AI的贡献从“显性规则”转向“数据组织、语义表征、能力诊断与智能协作”，并在大模型时代获得新的作用位置。

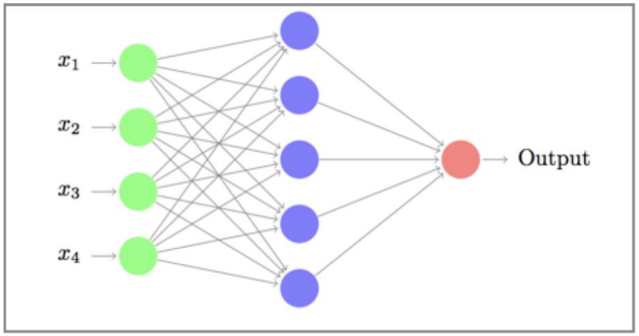
接下来进入“语义表征”

分布语义学思想 → 词向量 / 上下文表征 / 大模型语义能力；
形态学 (morphology)融入分词 (tokenization)会进一步改善模型表征复杂词意义。

表征学习是神经网络语言模型的基础



神经网络只能处理数值输入：
如何将语言转换为模型可计算的向量表示x？



神经网络只能处理数值输入。
因此，需要将离散词汇转化为可计算的语义向量。
分布语义学提供了这一转化的关键思想：从语境和分布中学习词义。

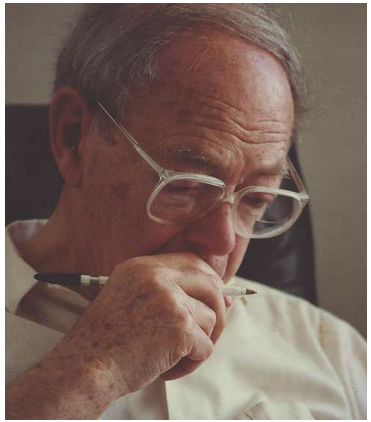
分布语义学 (Distributional Linguistics)

Distributional semantics is an approach to semantics that is based on the contexts of words in large corpora. 由上世纪50年代的一批语言学家提出：

Zellig Harris

. . . difference in meaning correlates with difference of distribution. (1954)

分布相似的词意义也相似



You shall know a word by the company it keeps. (1957)

观其邻，知其义

John Rupert Firth

也许你不认识Tezgüino这个词，
但是可以通过语境理解它的语义：

A bottle of tezgüino is on the table.

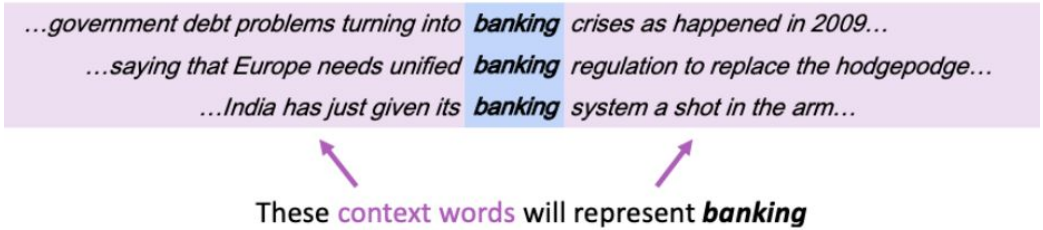
Everyone likes tezgüino.

Tezgüino makes you drunk.

We make tezgüino out of corn.

“观其邻，知其义”：基于上下文的词语表征 (Word Representation)

- **核心思想**：一个词的意义，可以通过与它经常共同出现的词来刻画。
- 这是统计自然语言处理和语言智能中最成功、最具影响力的思想之一。
- **基本方法**：以目标词为中心，在**固定**大小的上下文窗口中 统计词语共现关系。
  - 使用大量包含目 标词的上下文就可以建立 该目标词的表征（representations）



分布语义学：从词向量到大语言模型的表征基础

- **核心思想一脉相承：** 词语的意义可以从其共现关系和上下文分布中学习。
- **技术路径不断演进：** 从基于共现统计的词向量，到 Word2Vec 等神经词嵌入，再到预训练语言模型和大语言模型的上下文表征。
- **关键贡献：** 将离散的语言符号映射到连续、可计算、可比较的语义空间，为语义相似度计算、信息检索、语义匹配、语言理解与生成奠定基础。

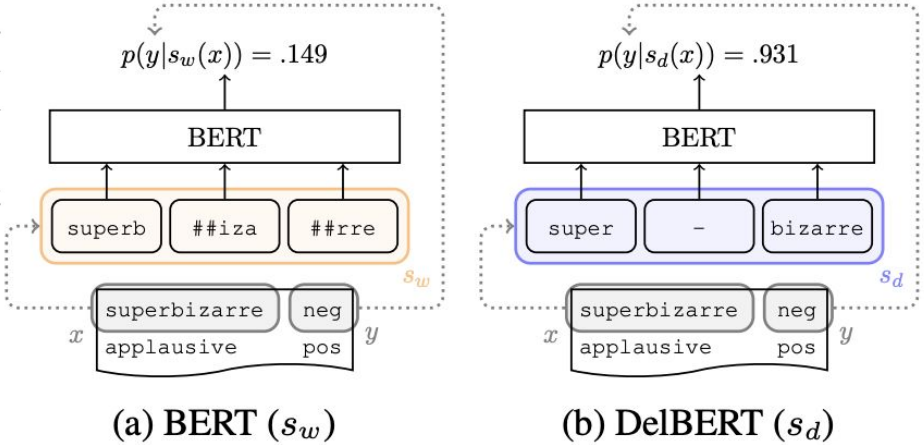
模型架构不断更新，但“从语言使用与上下文中学习意义”始终是现代自然语言处理和大语言模型发展的底层思想。

形态学改进复杂词汇的语义表征

通过把形态学知识融入到分词(tokenization)过程，改进语言模型对复杂词汇的语义表征，提升下游语言任务表现。

Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words

Class 1		Class 2	
Class	Examples	Class	Example
neg	overpriced, crappy	pos	megafavorite, applausive
phys	semithermal, ozoneless	cs	autoencoded, rankable
ent	supervampires, spoilerful	dis	antirussian, immigrationism



(a) BERT (s_w)　　(b) DelBERT (s_d)

Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words
(Hofmann et al., ACL-IJCNLP 2021)

大语言模型的分词方式仍有改进空间

当前大语言模型通常采用固定的默 认分词方案（*基于subword词频分布和分 词算法*），但这种切分方式未必最符合人 类语言加工规律，也未必能充分保留 词汇、语素和形态结构等语义表征信息。

相关研究表明，大 语言模型的默认分词方案未必是最 优的：

Geh et al., EMNLP 2024：同一字符串采用不同分 词方式时，模型仍能识别其中的有效信号。

Geh et al., ACL 2025：替代性分词方案可以通 过对抗优化获得，说明默认分词并非唯一 选择。

Zheng et al., NeurIPS 2025：系统实验表明，非标准分词通常能够保持模型性能，部分情况下甚至有所提升。

启示：分词不仅是工程预处理步骤，也是连接语言结构与模型表征的重要接口。

如何引入 词法、形 态学和人 类语言加工知 识，设计更具 语言学依据的分 词机制，仍 值得深入探索。

Where is the signal in tokenization space? (Geh et al., EMNLP 2024)

Adversarial Tokenization (Geh et al., ACL 2025)

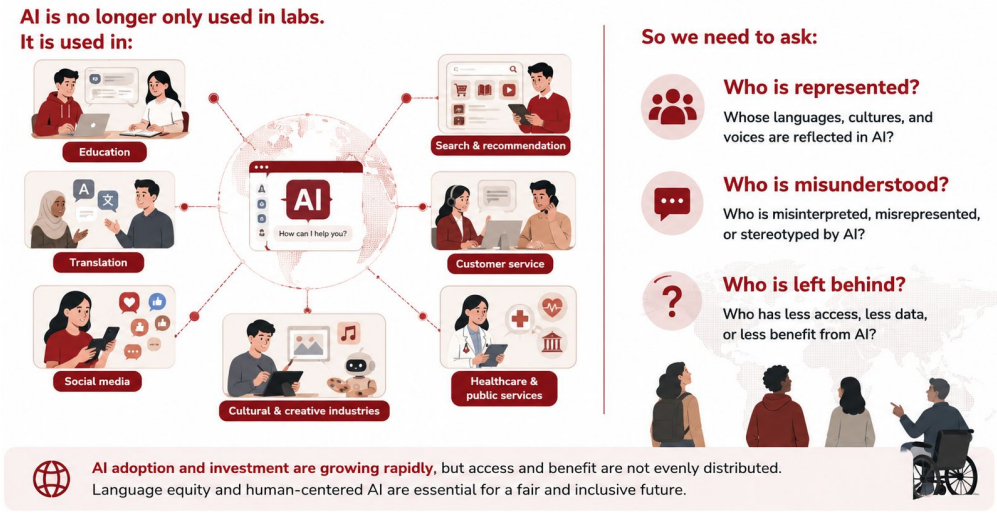
Broken Tokens? Your Language Model Can Secretly Handle Non-Canonical Tokenizations (Geh et al., ACL 2025)

第三部分

语言科学帮助构建更好的人工智能：以多语智能为例

多语智能越来越受学界和业界关注

AI is becoming global infrastructure



千问系列

Qwen 1 (2023): bilingual（中英为主）

Qwen 1.5: 12 (2024)

Qwen1.5 showcase impressive multilingual capabilities, a
oss a diverse set of 12 languages. In evaluations covering

Qwen 2: ~30 (2024)

Qwen2 demonstrates robust multilingual
30 languages, spanning English, Chinese,

Qwen 3: 100+ (2025)

Support of 100+ languages and dialects with instruction following and translation.

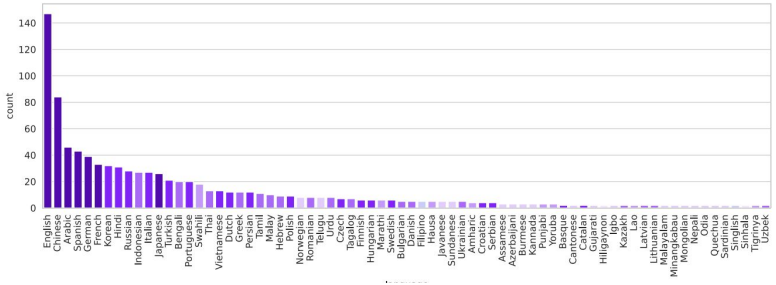
多语智能的挑战1：语言资源不均

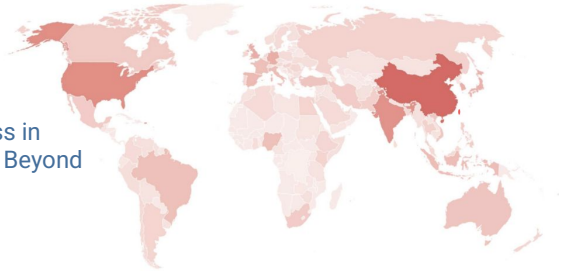
上海交通大学外国语学院
School of Foreign Languages

1. 语言资源不均

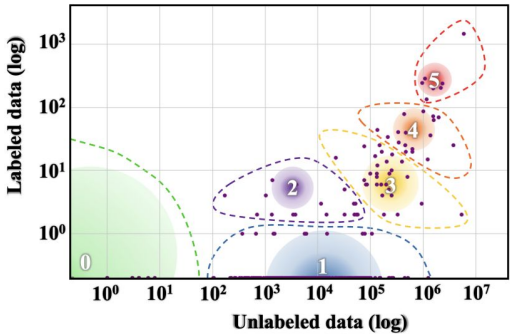
Language imbalance and inequality

(Global AI is often not equally global)





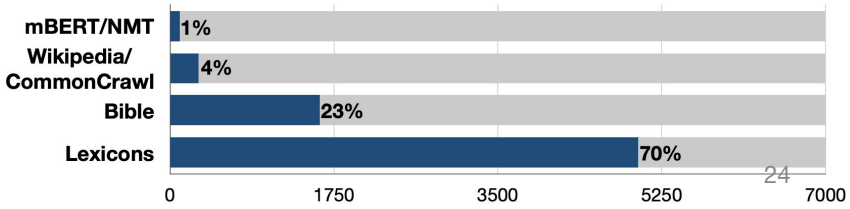
Survey of Cultural Awareness in Language Models: Text and Beyond
(Pawar et al., CL 2025)



The State and Fate of Linguistic Diversity and Inclusion in the NLP World
(Joshi et al., ACL 2020)

Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation
(Wang et al., ACL 2022)

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%



多语智能的挑战2：文化差异感知不到位

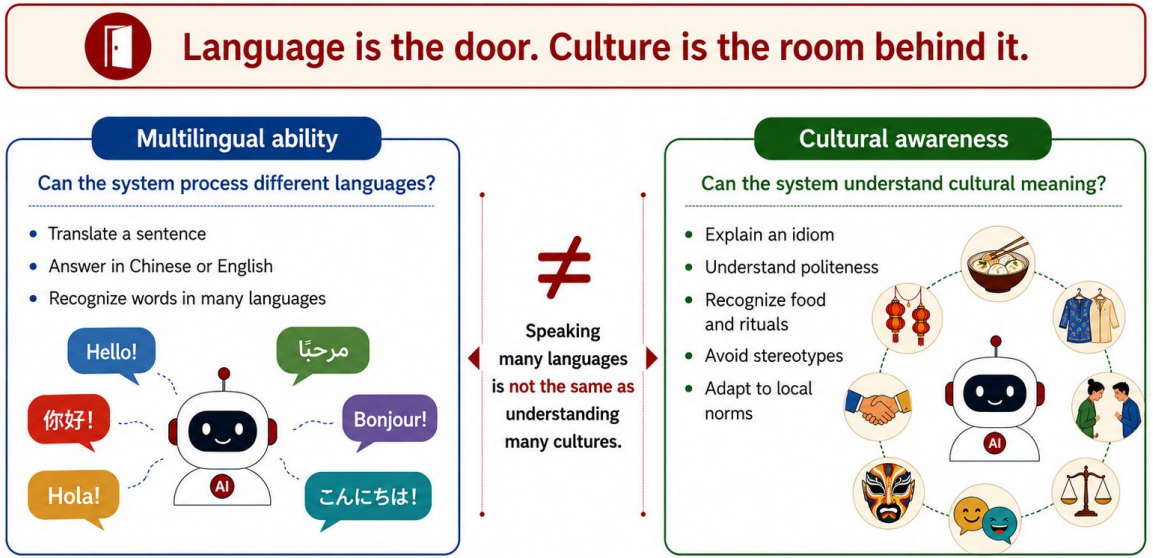
上海交通大学外国语学院
School of Foreign Languages

2. 文化差异感知
Cultural Nuances
(Multilingual is not the same as Multicultural)

Xiaohongshu's Global E-commerce Push: Is AI Translation "Good Enough" for Cross-Border Brands?

Shawn Wang
Global Localization Strategist | Helping Brands Go Global efficiently with AI | Managing 50+ Languages for Brands like Samsung,...

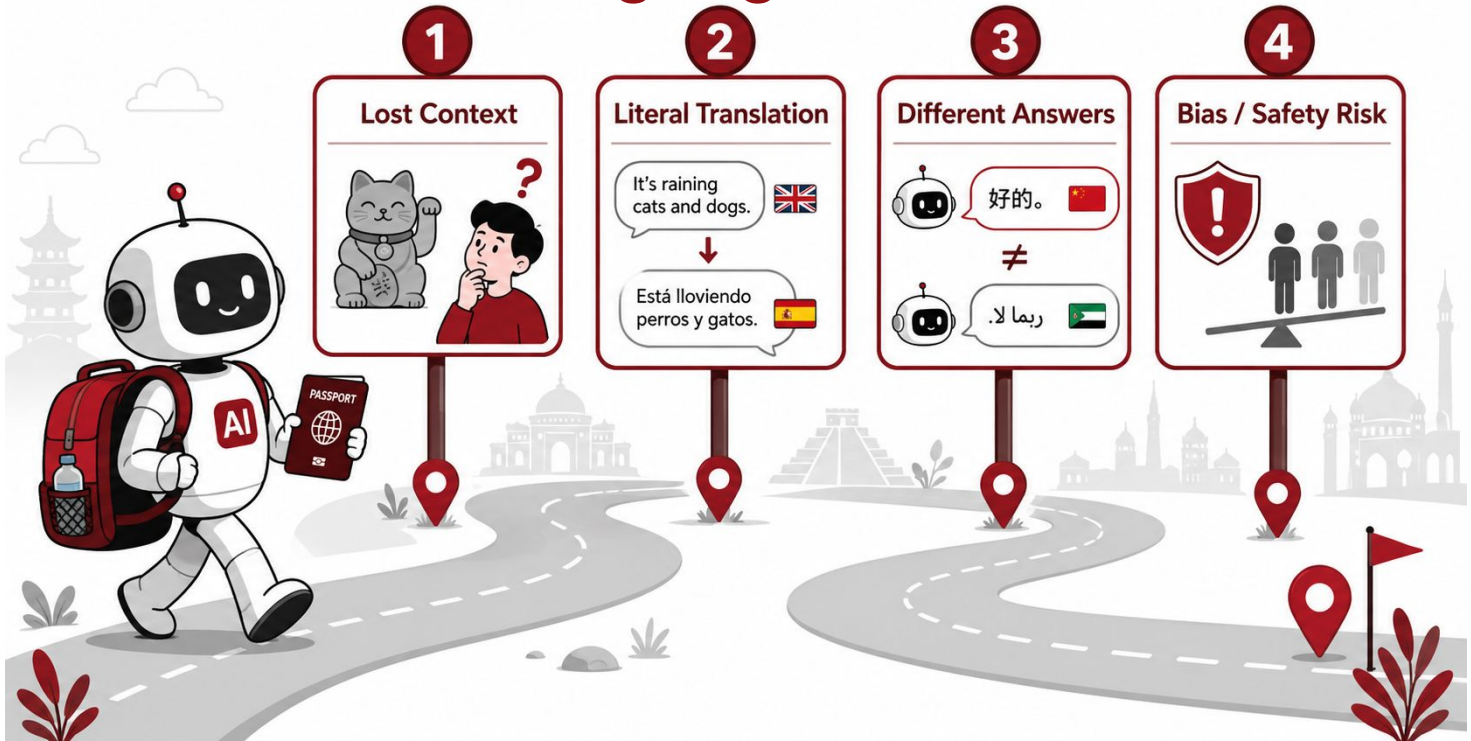
March 17, 2026



This highlights the profound challenge of cross-border localization: **AI can process language, but it cannot process cultural nuance and standard discrepancies.**

AI在多语言跨文化场景下会面临各种问题

AI can fail across languages and cultures

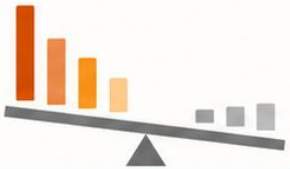


语言学为多语智能带来了什么？

核心问题：当语言数据分布极不均衡时，多语智能如何惠及低资源语言？

1 多语智能挑战

- 语言资源分布不均 → language inequity
- 许多低资源语言数据稀缺，模型表现受限



2 应对路径

A. 增加语言数据资源

但对部分语言，数据与使用者规模本身有限

B. 跨语言迁移学习

将高资源语言学到的能力迁移到低资源语言

3 迁移中的关键问题

- 源语言与目标语言如何选择？
- 为什么有的迁移效果更好？
- 语言相似性会影响迁移效果

源语言 ⟷ 目标语言

……

4 语言学贡献：lang2vec

- 用可计算的语言向量量化语言相似性
- 视角：句法、音系、**语音**、地理、亲缘
- 覆盖数千种语言
- 广泛用于跨语言迁移研究

similarity = cosine(vector_a, vector_b)

URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors (Littell et al., EACL 2017)

作用

- 更合理地选择源语言
- 分析迁移效果的来源与原因
- 改善低资源语言的迁移表现

Take-home message

当“直接增加数据”受限时，语言学让多语智能能够计算语言差异、利用语言相似性，从而更有效地支持低资源语言。

跨语言迁移学习
Cross-Lingual Transfer Learning

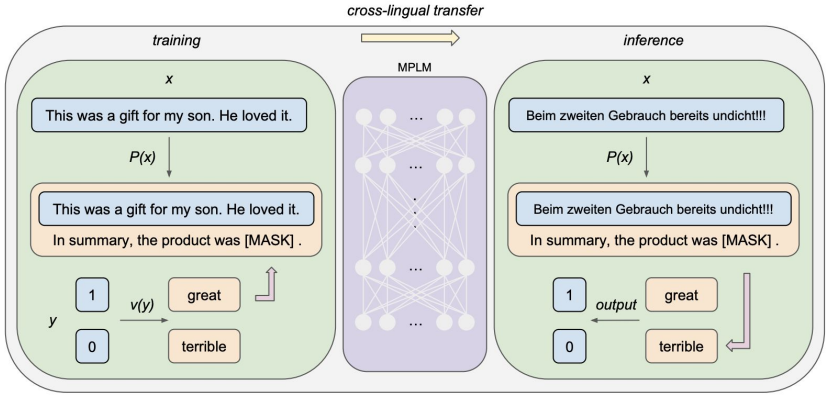
上海交通大学外国语学院
School of Foreign Languages

什么是跨 语言迁移学 习？

把模型在相 对高资源语言上的 语言能力迁移到相 对低资源语言上，比如：

Case 1: 通过模型微调迁移

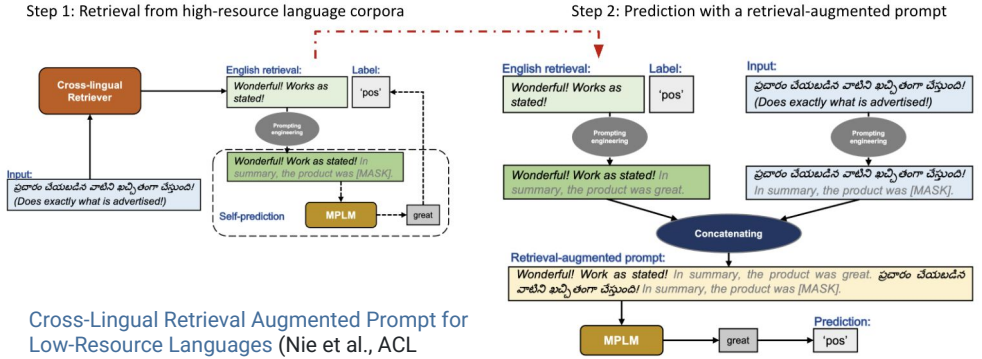
用高资源语言（英语）微调模型在情感分析任务上的能力，零样本或者少样本迁移到目标语言（德语）



Is Prompt-Based Finetuning Always Better than Vanilla Finetuning? Insights from Cross-Lingual Language Understanding (Ma et al., KONVENS 2023)

Case 2: 通过提示词学习迁移

把高资源语言的示例添加到提示词中，通过跨语言检索增强提示词的方式，迁移到目标语言。



Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages (Nie et al., ACL Findings 2023)

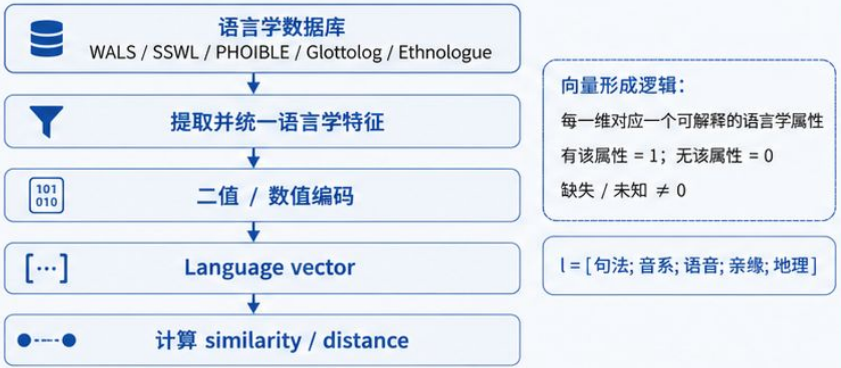
Lang2vec：把语言类型学知识转化为可计算的语言向量

上海交通大学外国语学院
School of Foreign Languages

lang2vec 的核心贡献，是把语言类型学、音系学、语音学、历史语言学和语言地理学的知识，转化为标准化、可比较、可计算的语言向量，为跨语言迁移与低资源语言研究提供支持。

一门语言如何变成向量？

从语言学数据库到可计算的 language vector



	S_SUBJECT-_BEFORE_VERB	S_SUBJECT-_AFTER_VERB	S_ADPOSITION-_BEFORE_NOUN	S_ADPOSITION-_AFTER_NOUN
eng	1	0	1	0
mlg	0	1	1	0
kaz	1	0	0	1

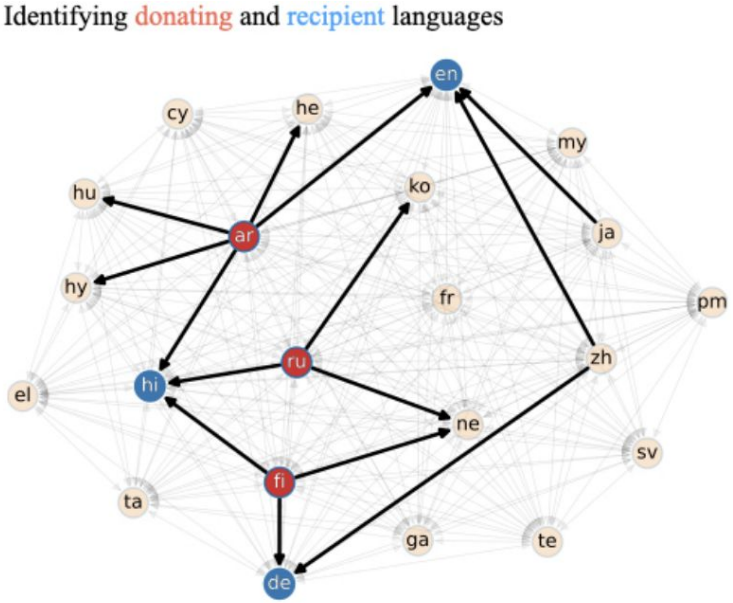
五类语言向量及其语言学来源

类型	维度	表示内容	主要来源
句法 Syntax	103	语序、否定位置、形容词—名词顺序、介词 / 后置词等	WALS, SSWL, Ethnologue
音系 Phonology	28	音系结构与音系模式	WALS, Ethnologue
语音 Inventory	158	某类语音 / 音类是否存在	PHOIBLE + PanPhon
亲缘 Family	约 3,700	语言家族及分支层级	Glottolog
地理 Geography	299	与全球299个固定点的归一化距离	Glottolog, WALS, SSWL

补充说明

- 可单独使用某一类向量，也可拼接使用
- 多个来源可整合；缺失值可通过预测补全
- 将语言学知识转化为可计算、可比较、可解释的模型输入

Lang2vec对多语智能的贡献：识别跨语言迁移效果好的语言



A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank (Malkin et al., NAACL 2022)

Information Processing and Management 60 (2023) 103250

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Zero-shot cross-lingual transfer language selection using linguistic similarity

Juuso Eronen *, Michal Ptaszynski, Fumito Masui

Kitami Institute of Technology, 165, Koencho, Kitami, 090-0015, Hokkaido, Japan

ARTICLE INFO

Keywords:
Multilingual natural language processing
Zero-shot learning
Transfer learning
Linguistics
Language similarity

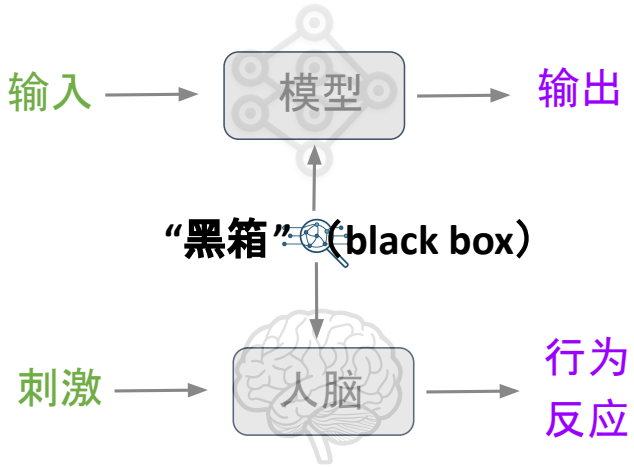
ABSTRACT

We study the selection of transfer languages for different Natural Language Processing tasks, specifically sentiment analysis, named entity recognition and dependency parsing. In order to select an optimal transfer language, we propose to utilize different linguistic similarity metrics to measure the distance between languages and make the choice of transfer language based on this information instead of relying on intuition. We demonstrate that linguistic similarity correlates with cross-lingual transfer performance for all of the proposed tasks. We also show that there is a statistically significant difference in choosing the optimal language as the transfer source instead of English. This allows us to select a more suitable transfer language which can be used to better leverage knowledge from high-resource languages in order to improve the performance of language applications lacking data. For the study, we used datasets from eight different languages from three language families.

第四部分

理解智能: 语言科学为解释大语言模型提供认知参照

语言模型是黑箱，人脑也是



Humans

~~Deep neural networks~~

- Difficult to understand
- Opaque inductive biases
- Succeed on a surprisingly wide range of problems
- Process lots of data

EMNLP 2024
大会主旨发言

Tom Griffiths

Professor @ Princeton University

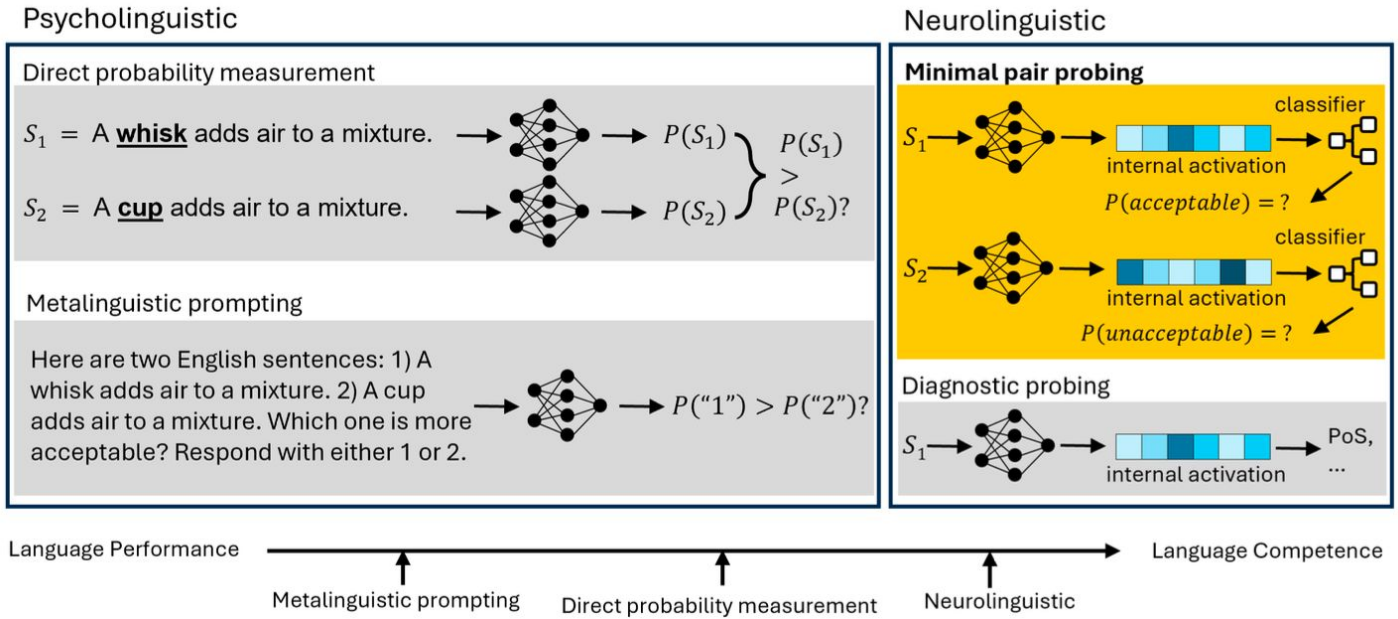
A pitch for cognitive science

- Making Bayesian models of neural networks borrows the idea of cognitive modeling from cognitive science
- Cognitive science has lots of other tools to offer

大语言模型也可以成为语言科学的计算实验对象。
人类语言加工的理论、证据与实验方法，为打开模型黑箱、理解其语言能力提供了重要参照。

把大语言模型当作语言学实验“被试”

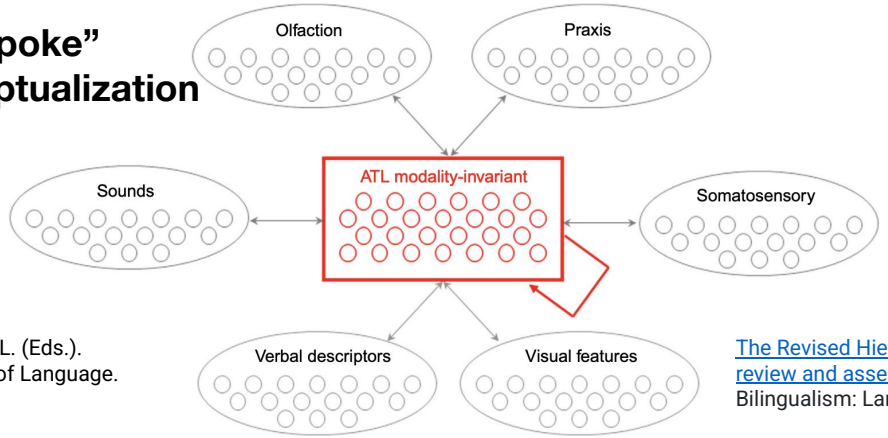
- 设计语言学控制实验
  minimal pair最小对设计
- 测量模型的“行为反应”
- 分析模型的“内部活动”
- 与人类加工证据对照
  Human model alignment
  对齐研究



Large Language Models as Neurolinguistic Subjects: Discrepancy between Performance and Competence
(He et al., ACL Findings 2025)

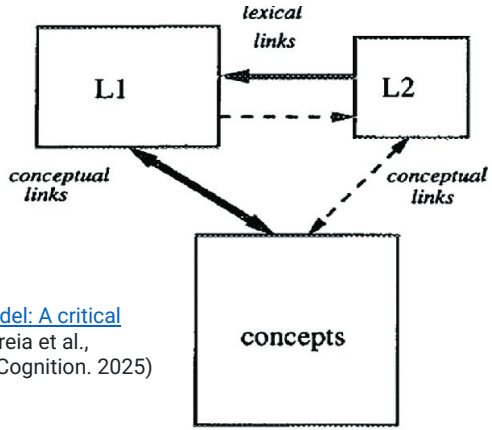
Semantic Hub-and-Spoke 假设

The "hub-and-spoke" model of conceptualization



Hickok, G., & Small, S. L. (Eds.). (2016). Neurobiology of Language. Amsterdam: Elsevier.

Revised Hierarchical Model



The Revised Hierarchical Model: A critical review and assessment (Correia et al., Bilingualism: Language and Cognition. 2025)

神经语言学证据

JNeurosci THE JOURNAL OF NEUROSCIENCE

Brain-Based Translation: fMRI Decoding of Spoken Words in Bilinguals Reveals Language-Independent Semantic Representations in Anterior Temporal Lobe

Brain-Based Translation: fMRI Decoding of Spoken Words in Bilinguals Reveals Language-Independent Semantic Representations in Anterior Temporal Lobe (Correia et al., Journal of Neuroscience. 2014)

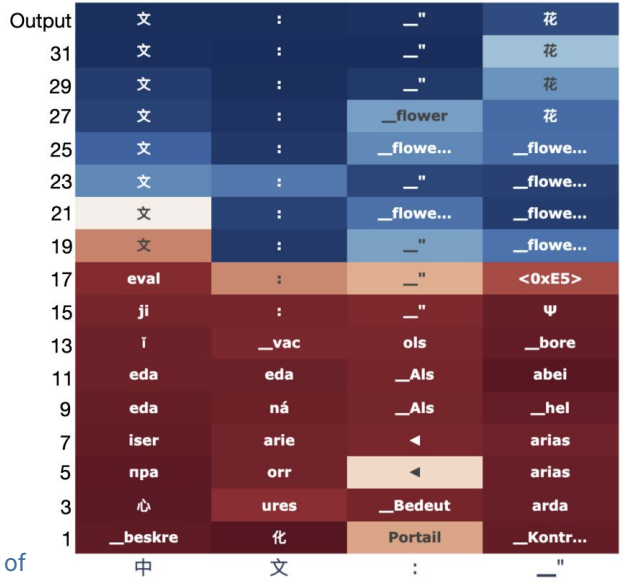
大语言模型也有 Semantic Hub (语义枢纽) 吗？

Published as a conference paper at ICLR 2025

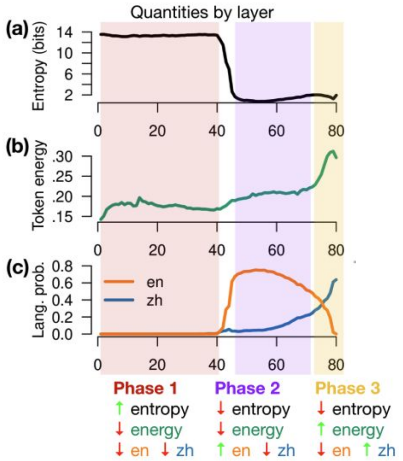
THE SEMANTIC HUB HYPOTHESIS:
LANGUAGE MODELS SHARE SEMANTIC REPRESENTATIONS ACROSS LANGUAGES AND MODALITIES

Zhaofeng Wu[♋] **Xinyan Velocity Yu**[♅] **Dani Yogatama**[♅] **Jiasen Lu**[♎] **Yoon Kim**[♋]
[♋]MIT [♅]University of Southern California [♎] Allen Institute for AI
zfw@csail.mit.edu

[The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities](#) (Wu et al., ICLR 2025)

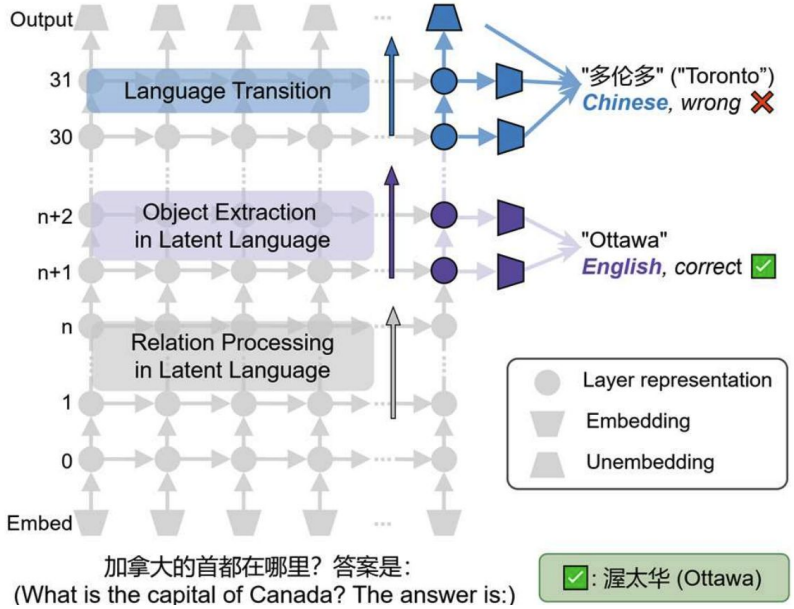


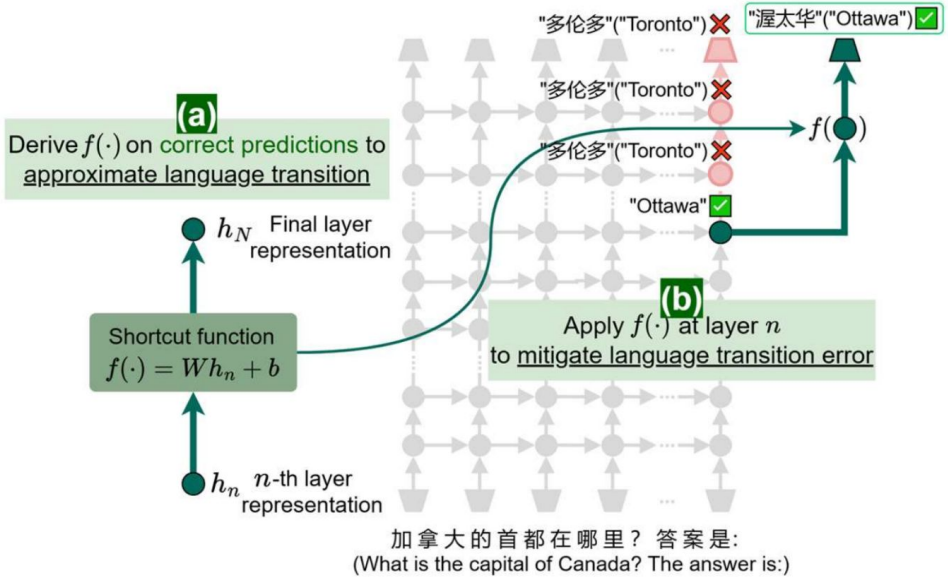
Do Llamas Work in English? On the Latent Language of Multilingual Transformers (Wendler et al., ACL 2024)



大语言模型内部也有**跨语言共享的概念空 间。**

语义枢纽假设有什么用？
缓解大语言模型跨语言事实不一致现象





Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models (Wang et al., ACL 2025)

语义枢纽假设有什么用？
缓解大语言模型语言混淆现象

User Prompt

ما الفرق بين السور المكية والمدينة في القرآن الكريم؟

English translation: *What is the difference between the Meccan and Medinan surahs in the Holy Quran?*

LLM Response

The difference between Meccan and Medinan surahs in the Quran is a significant aspect of Islamic studies. Here's a brief overview:
Meccan Surahs (سور مكية):
1. **Time period:** The Meccan surahs were revealed during the early years of Prophet Muhammad's prophethood, ...

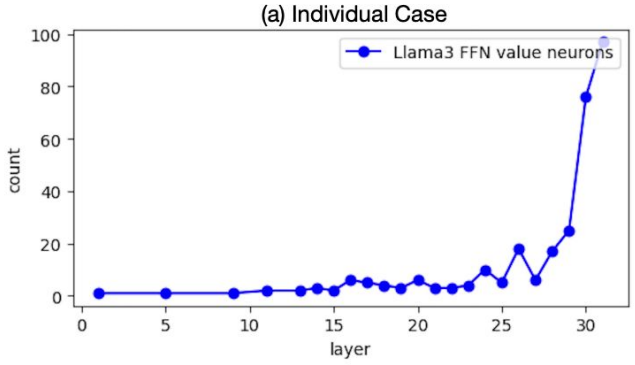
Case 1 (aya-zh, id1)

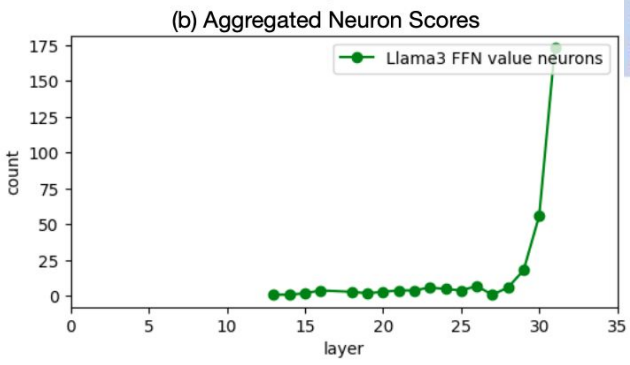
Input: 请解释"拆东墙补西墙"的意思。*Please explain "拆东墙补西墙"*

Before Neuron Editing: English Translation: "To tear down the eastern wall to repair the western wall" - Meaning: To solve a problem by creating a new one, or to fix one issue by making another one worse.

After Neuron Editing: "拆东墙补西墙"是中国古代的一句成语，它的意思是"解决问题的方法是解决问题的根源，而不是简单地解决表面现象"。\n ...

Multilingual Llama3: 2023-03-23 14:34:23 拆东墙补西墙的意思是为了解决某个问题或问题所在的某个方面，而不是直接解决问题本身，...





Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models (Nie et al., EMNLP Findings 2025)

第五部分

总结与展望