从机制到应用：大语言模型前沿进展
From Mechanisms to Applications: Frontiers in Large Language Model

聂耳聪
Center for Information and Language Processing,
Ludwig Maximilians University of Munich (LMU)
Munich Center for Machine Learning (MCML)
https://ercong21.github.io/

Sept. 28, 2025

About me

聂耳聪 PhD Candidate

Schuetze NLP Lab, Center for Information and Language Processing (CIS),
Ludwig Maximilians University of Munich (LMU Munich),
Munich Center for Machine Learning (MCML)

- 即将从慕尼黑大学（LMU Munich）信息与语言处理中心（CIS）博士毕业，专业方向是自然语言处理（NLP）
- 隶属于前ACL主席Hinrich Schütze教授领导的Schütze NLP Lab，同时是慕尼黑机器学 习中心（MCML）青年会员
- **硕士（M.Sc.）：**计算语言学与计算机科学硕士学位，**慕尼黑大学**
- **本科（B.A.）：**德语与金融学士学位，**上海交通大学**
- 研究兴趣包括：多语言自然语言处理、高效NLP方法以及受人 类启发的自然语言处理

学术活动

更多

14 06月 “世界文学中的非人类叙事: 跨学科视角”暨上海交通大学第七届叙事学暑...

9 06月 讲座通知|语言学研究中的问题意识与问题分析：以汉语韵母[ian]和[iaŋ]...

7 06月 2023年语音学与大脑神经机制高级研讨会（第二号通知）

2 06月 世界文学中的非人类叙事暨上海交通大学第七届叙事学暑期高端研讨会第一...

2023-07-05

上海交通大学外国语学院2024年夏令营考核结果公示

2023-06-20

上海交通大学外国语学院2024年外语菁英夏令营营员名单

聚焦专栏

更多

聂耳聪：从德语专业转型到计算语言学

陈晨：栉风沐雨 砥砺前行

曹慧：“用‘知识’和‘价值’引导学生对于课程的热爱”

About Our Lab

Schütze Lab @ LMU Munich Home People Publications Project

Welcome to the Schütze Lab

We are a dynamic research group at the Center for Information and Language Processing at Ludwig Maximilian University Munich, under the supervision of Prof. Hinrich Schütze. Our research areas include:

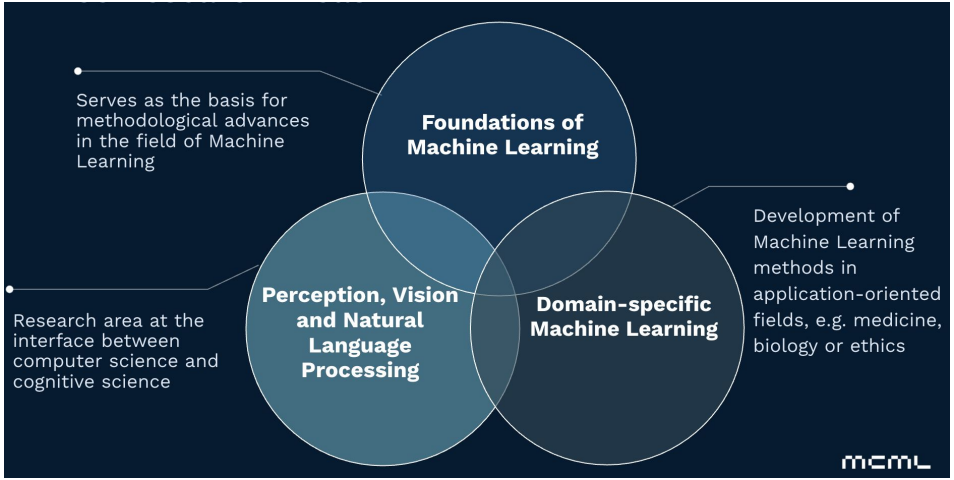
- Large Language Models (LLMs): We explore the behavior, structure, and potential of LLMs, examining their capabilities, biases, and self-assessment mechanisms to improve reliability and interpretability.
- Knowledge Expansion in NLP Models: We investigate how models can acquire and integrate new knowledge over time, using techniques that help improve their comprehension and generation abilities.
- Representation Learning and Interpretability: We study how language models represent linguistic and conceptual information by analyzing neurons and internal circuits to better understand and refine model behavior.
- Multilingual NLP: We address challenges in processing and evaluating multiple languages by developing benchmarks and methods for multilingual evaluation, including work on low-resource languages.
- Intersection of NLP and Robotics: We integrate language understanding into robotic systems to enable natural, adaptable interaction in multimodal environments.

实验室：https://cisnlp.github.io/
学院：https://www.cis.lmu.de/

About MCML (Munich Center for Machine Learning)

德国六大人工智能研究中心之一

研究领域



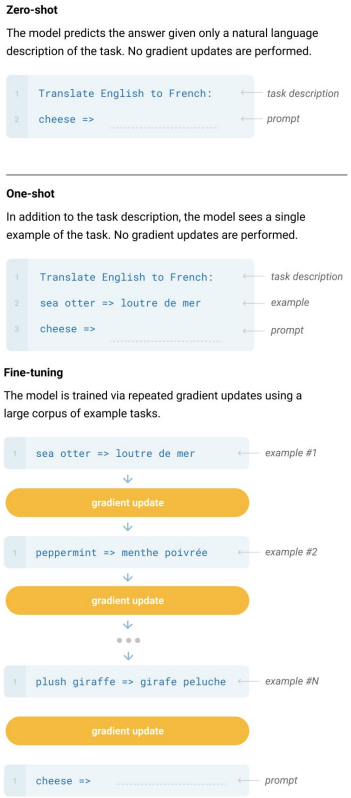
MCML官网：https://mcml.ai/#about

Research Profile（1）

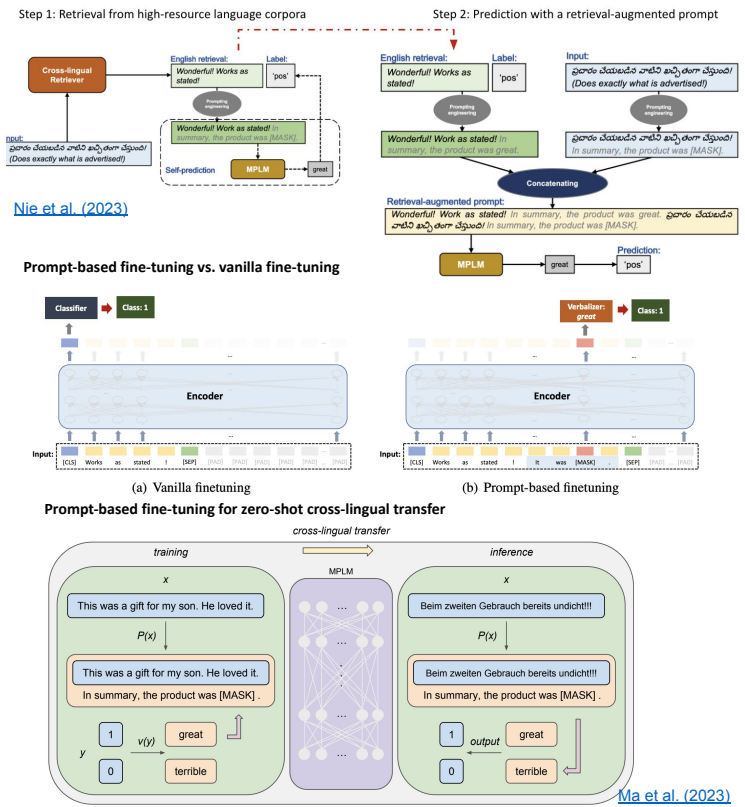
多语言NLP与跨语言迁移：致力于探索基于提示的学习（prompt-based learning）等新范式在多语言和低资源语言场景下的迁移能力，提出了跨语言检索增强提示（PARC）等方法，提升了大模型在低资源语言上的零样本表现，并系统研究了大模型的多语言能力与迁移机制。

高效NLP方法：关注低资源数据条件下的数据增强、参数高效微调（PEFT）、模型剪枝等方向，聚焦NLP模型在资源受限场景下的实用性和可扩展性。

人类启发与可解释NLP：结合计算神经语言学与人类语言处理机制，研究大模型内部的语言结构与知识表征，致力于大模型的可解释性和跨学科应用。



Brown et al. (2020)



Nie et al. (2023)

Ma et al. (2023)

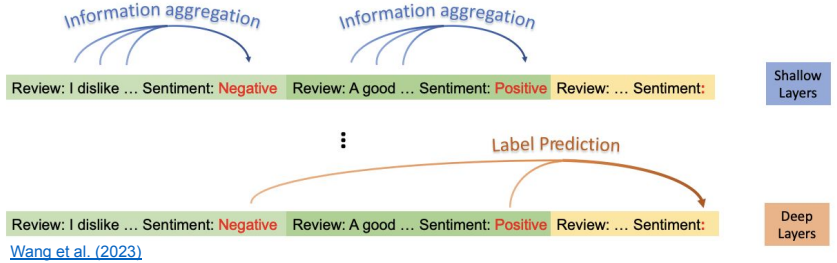
Research Profile（2）

多语言NLP与跨语言迁移：致力于探索基于提示的学习（prompt-based learning）等新范式在多语言和低资源语言场景下的迁移能力，提出了跨语言检索增强提示（PARC）等方法，提升了大模型在低资源语言上的零样本表现，并系统研究了大模型的多语言能力与迁移机制。

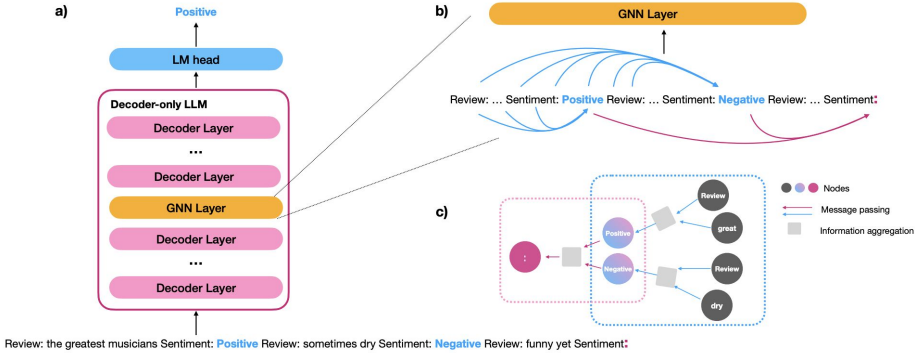
高效NLP方法：关注低资源数据条件下的数据增强、参数高效微调（PEFT）、模型剪枝等方向，聚焦NLP模型在资源受限场景下的实用性和可扩展性。

人类启发与可解释NLP：结合计算神经语言学与人类语言处理机制，研究大模型内部的语言结构与知识表征，致力于大模型的可解释性和跨学科应用。

GNNavi：一种受上下文学 习信息流启 发的图神经网络参数高效微 调方法



Wang et al. (2023)



Yuan et al. (2024)

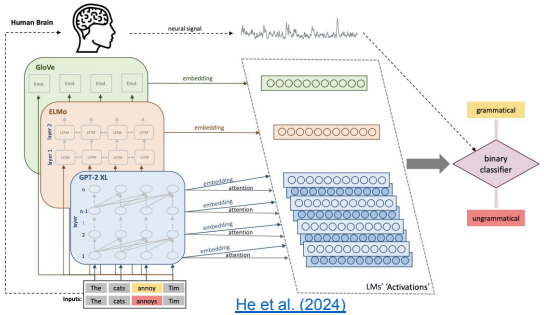
Research Profile（3）

多语言NLP与跨语言迁移：致力于探索基于提示的学习（prompt-based learning）等新范式在多语言和低资源语言场景下的迁移能力，提出了跨语言检索增强提示（PARC）等方法，提升了大模型在低资源语言上的零样本表现，并系统研究了大模型的多语言能力与迁移机制。

高效NLP方法：关注低资源数据条件下的数据增强、参数高效微调（PEFT）、模型剪枝等方向，聚焦NLP模型在资源受限场景下的实用性和可扩展性。

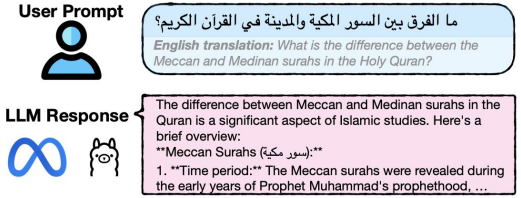
人类启发与可解释NLP：结合计算神经语言学与人类语言处理机制，研究大模型内部的语言结构与知识表征，致力于大模型的可解释性和跨学科应用。

Probing



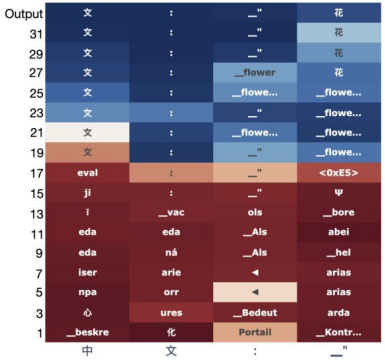
He et al. (2024)

Neuron Attribution and Intervention

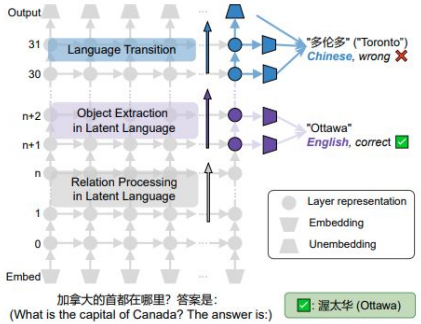


Mechanistic understanding and mitigation of language confusion in LLMs (Nie et al., 2025)

LogitLens



Do LLMs work in English? (Wendler et al., 2024)

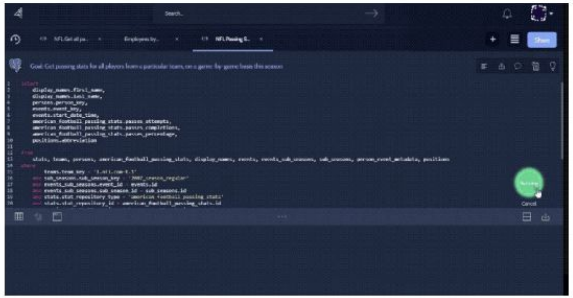


Cross-lingual factual recall inconsistency (Wang et al., 2025)

目录

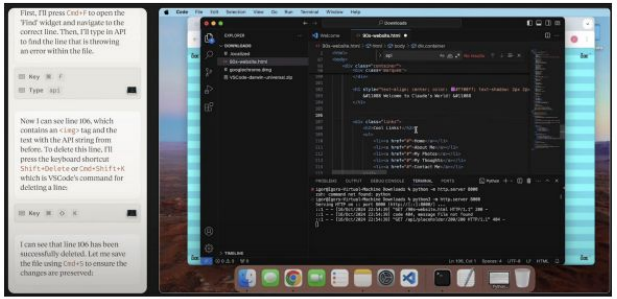
目录

LLMs transformed various applications



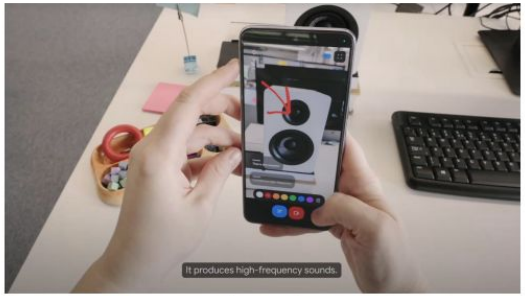
Code generation

Cursor, GitHub Copilot, Devin, Google Jules…



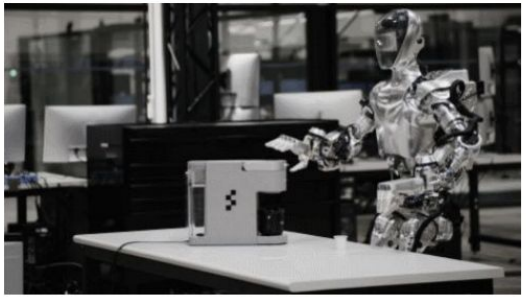
Computer use

Anthropic Claude, Google Jarvis, OpenAI Operator



Personal assistant

Google Astra, OpenAI GPT-4o,…



Robotics

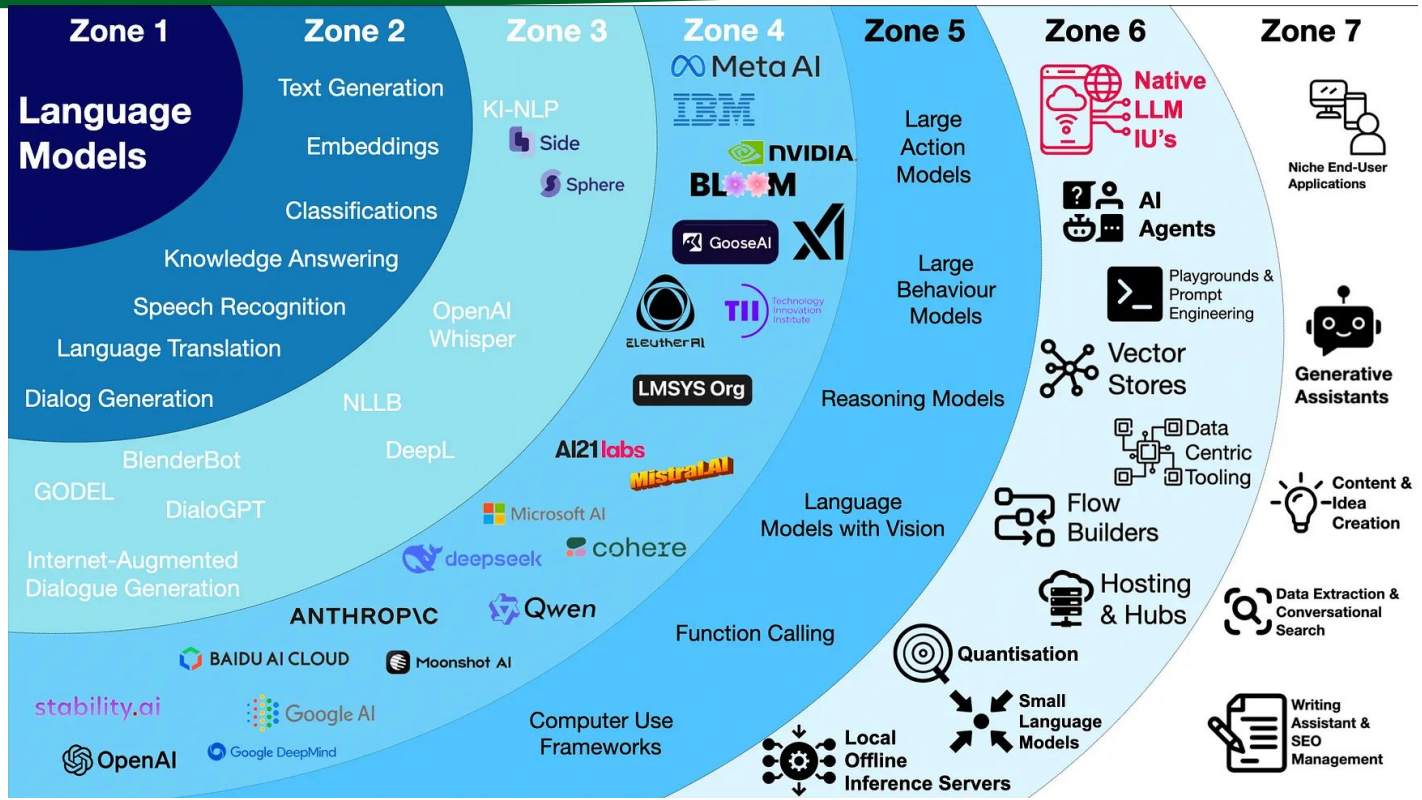
Figure AI, Tesla Optimus, NVIDIA GR00T…

- Education
- Law
- Finance
- Healthcare
- Cybersecurity

…

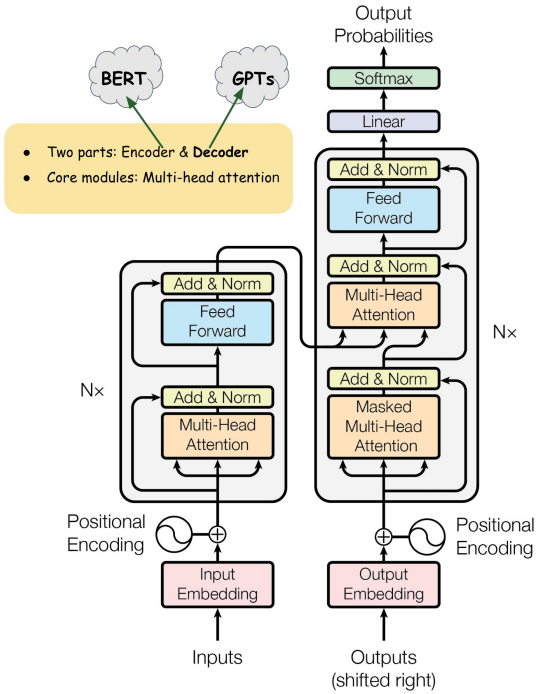
Source: https://llmagents-learning.org/slides/llm-agents-berkeley-intro-sp25.pdf

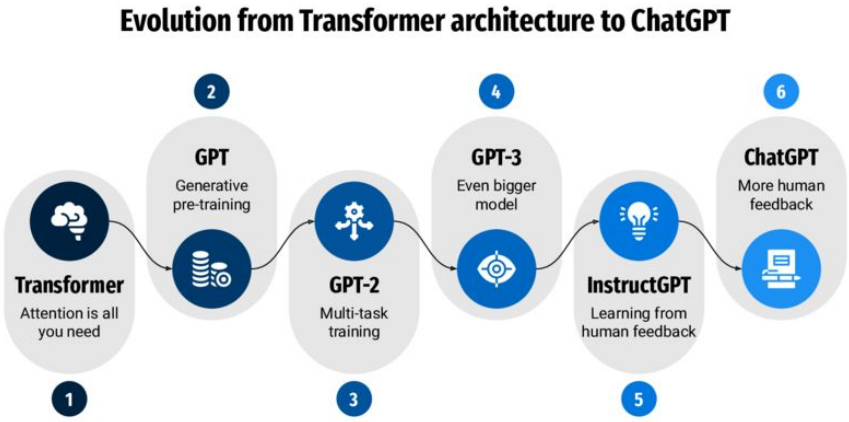
从语言模型到生成式AI时代

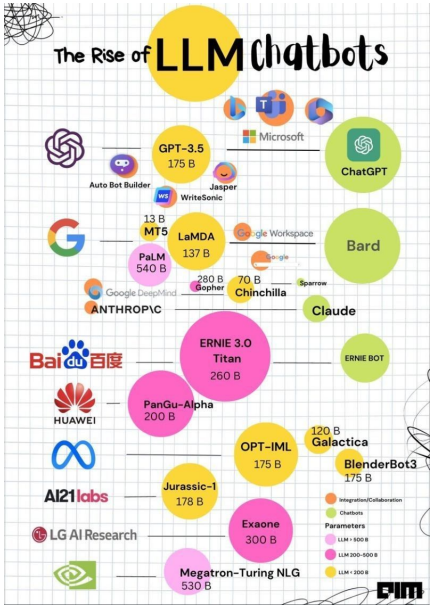


Source: https://cobusgreyling.medium.com/the-language-model-landscape-is-being-disrupted-again-e6e992c57f8e

从Transformers到大语言模型



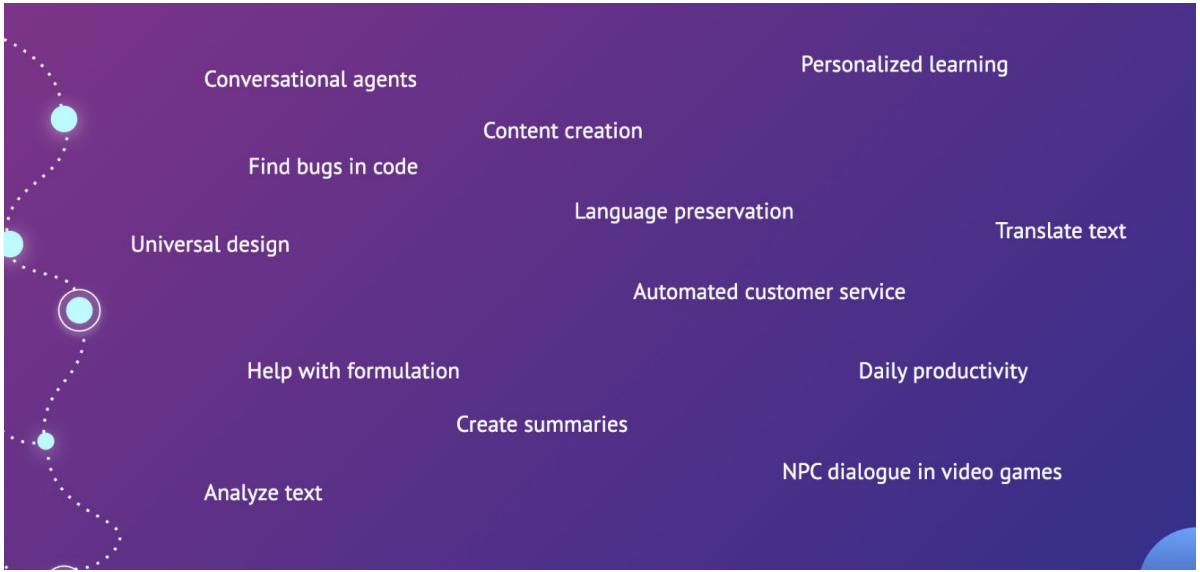




大语言模型与自然语言处理

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

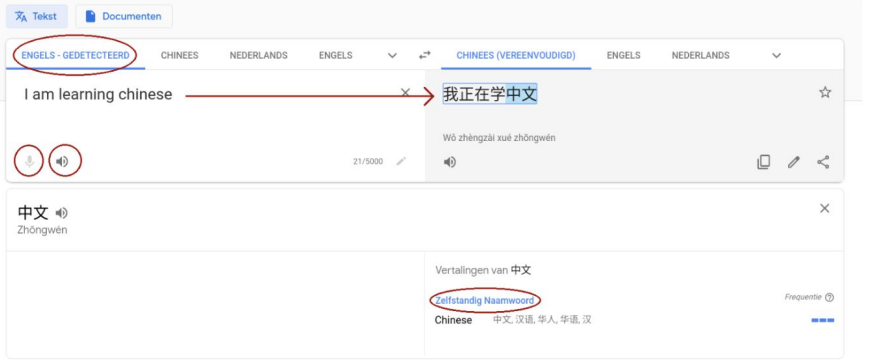
脱胎于自然 语言处理的大 语言模型引 领生成式 AI的突破

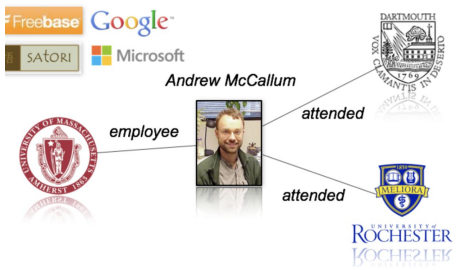


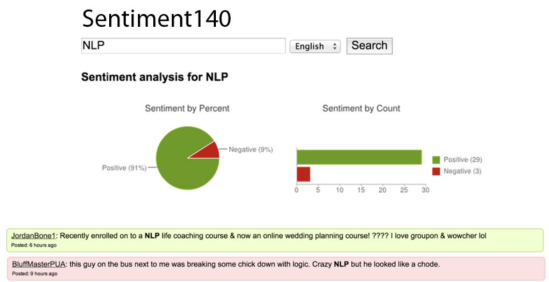
基于语言模型的自然语言处理

Personal assistants - Everywhere!







Credits to Prof. Barbara Plank

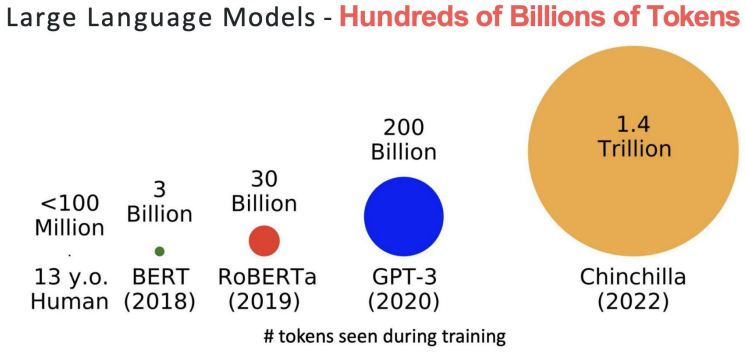
大语言模型是什么？

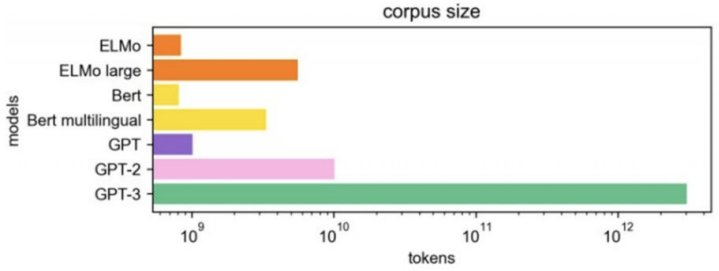
大语言模型：Large Language Models

AI Overview

A Large Language Model (LLM) is **a type of artificial intelligence (AI) that uses deep learning algorithms to understand, generate, and process human language**. Trained on massive datasets of text, LLMs learn complex patterns and nuances of language to perform tasks like answering questions, summarizing text, translating languages, and creating original content. Examples of LLMs include ChatGPT, Google Gemini, and Microsoft Copilot, which are used across various industries for tasks that require advanced natural language processing.

“大”

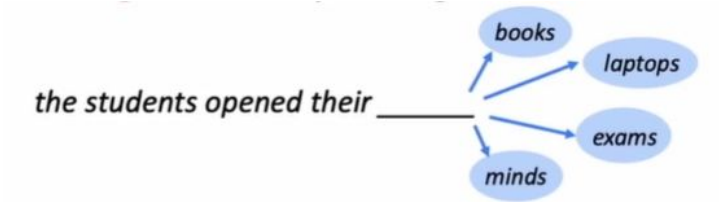




从语言到语言模型

为什么基于Transformers的大语言模型能work？回到**语言模型**的本质上来。

语言模型的本质就是**预测下一个词（next token prediction）**，
像人一个字一个字地说话一样



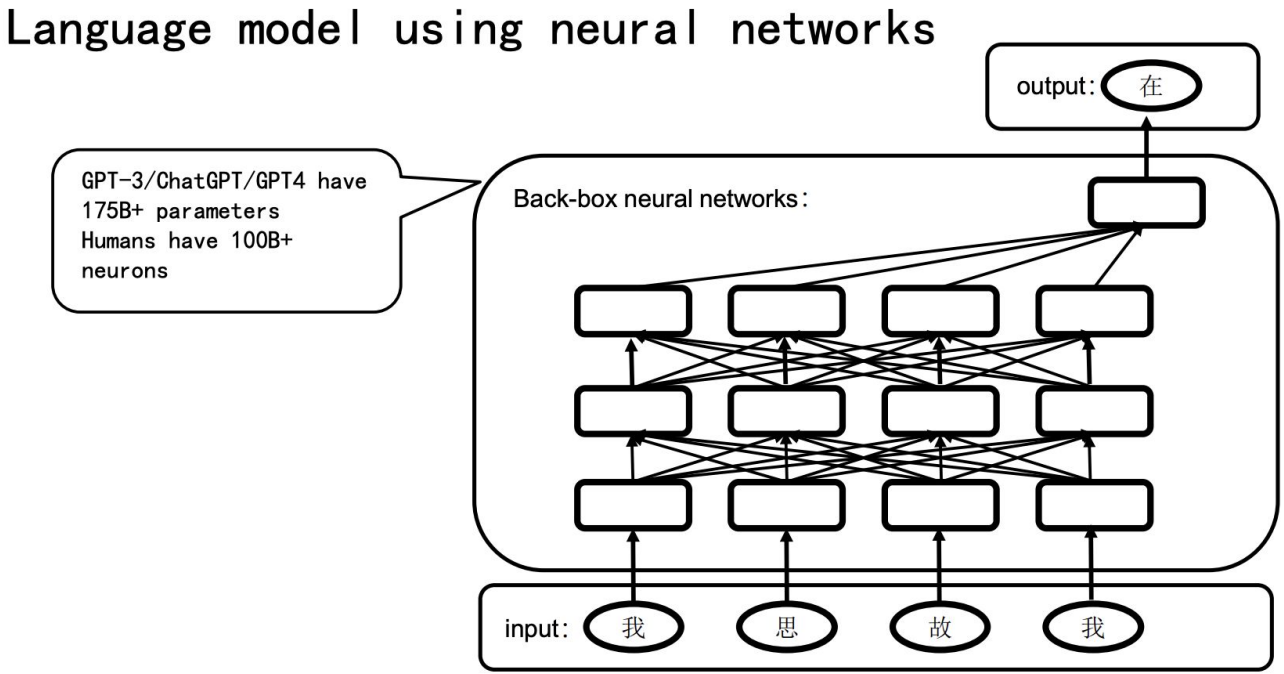
语言模型基于前面的内容生成下一个词的概率

More formally: given a sequence of words $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)}$, compute the probability distribution of the next word $\boldsymbol{x}^{(t+1)}$:

$$P(\boldsymbol{x}^{(t+1)} | \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})$$

Source: https://www.cnblogs.com/geeksongs/p/14503429.html

基于神经网络的语言模型



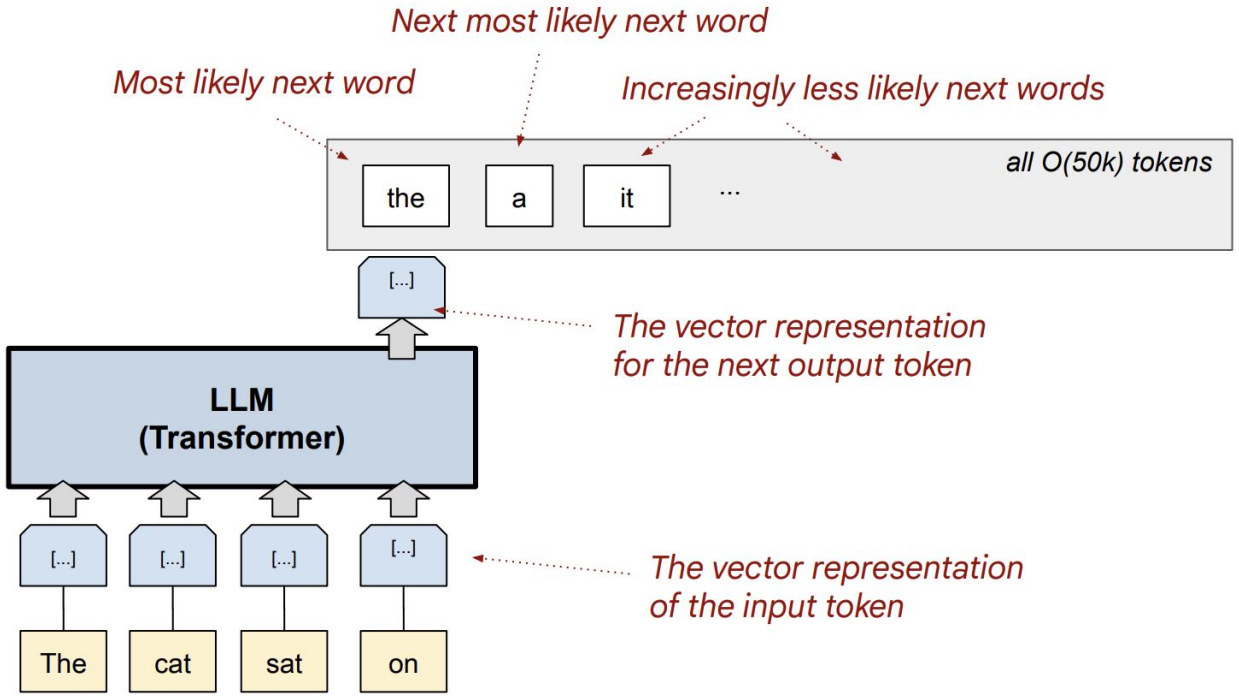
Source: https://llm-course.github.io/materials/2024fall/lecture-1-introduction.pdf

大语言模型的基础

语言模型的实质：预测下一个词
（Next token predictor）

大语言模型的技术基础：

1. **词向量表征**
2. **模型/建模：Transformer架构**
3. **解码**

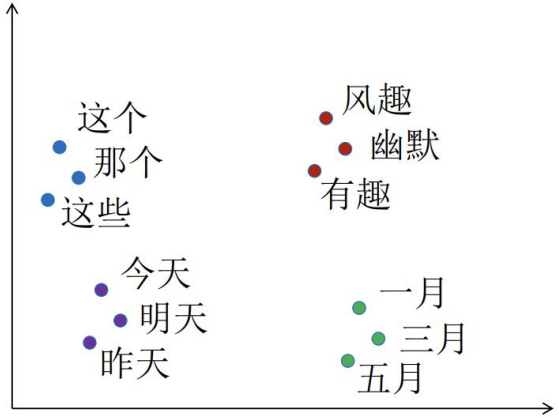


token = part of a word, the atomic unit that LLMs work with

Source: https://lxmls.it.pt/2025/slides/lucas.pdf

词向量

将离散的token转换到稠密的连续实数空间中，这样神经网络模型才可以处理。



稠密的连续实数空间

风趣 幽默

$$\begin{bmatrix}0.24\\0.15\\0.42\\0.51\\0.21\end{bmatrix} \quad \begin{bmatrix}0.25\\0.12\\0.39\\0.46\\0.26\end{bmatrix}$$

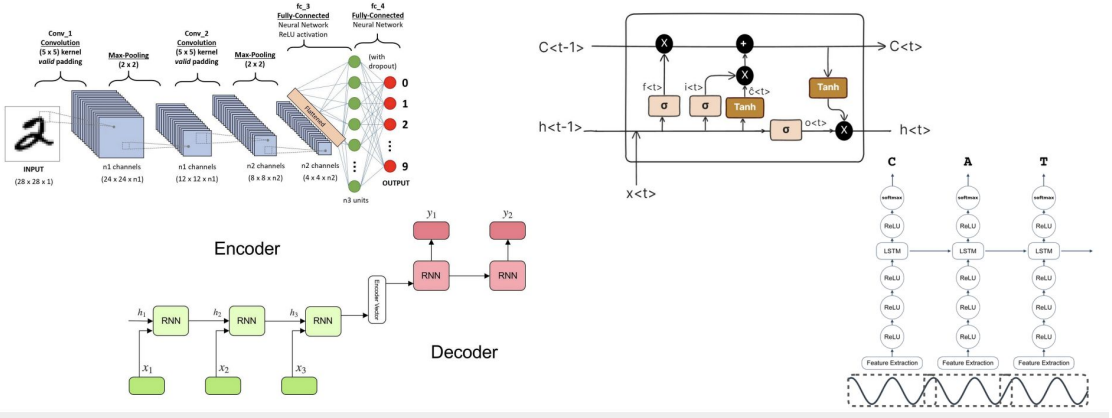
$$\boldsymbol{L} = \underbrace{\begin{bmatrix} \cdots & \cdots & \cdots \end{bmatrix}}_{V}{}_{D}$$

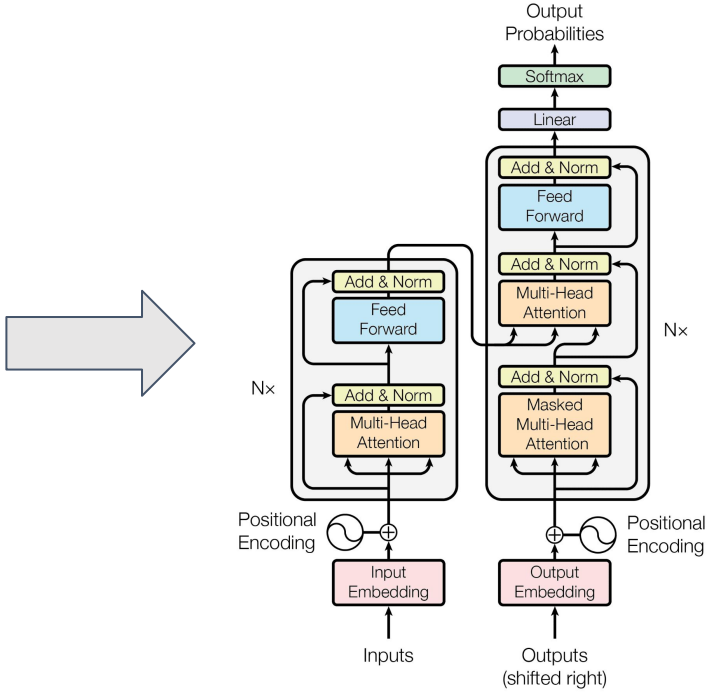
有趣 … 风情 幽默

$$L \in R^{D \times V}$$

Source: https://nlpr.ia.ac.cn/uploads/file/20190514/702bcb7536b2b0897240581875202633.pdf

Language Model after Transformers: One type for all





Source: https://lxmls.it.pt/2025/slides/sweta.pdf

解码算法

Decoding strategy: Sampling

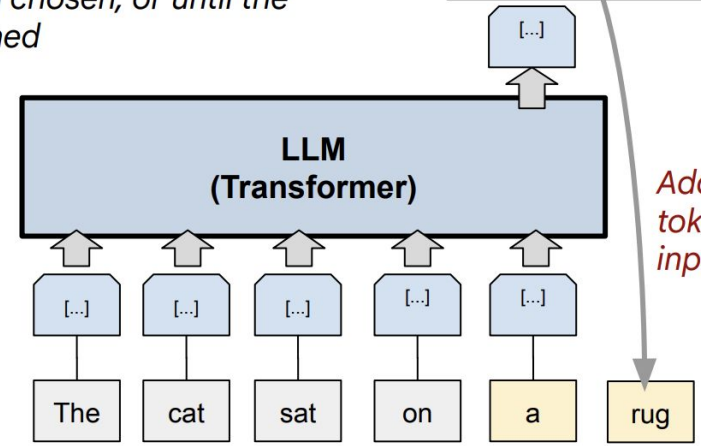
Pick the next word...

- *Only from **top-k** words*
- *Only from **top-p** probability space*
- *Flatten the distribution = **Temperature***

***Repeat** the decoding strategy until a special "EOS" token chosen, or until the max-length is reached*

Add non-determinism to pick the next word to output...

Next token distribution



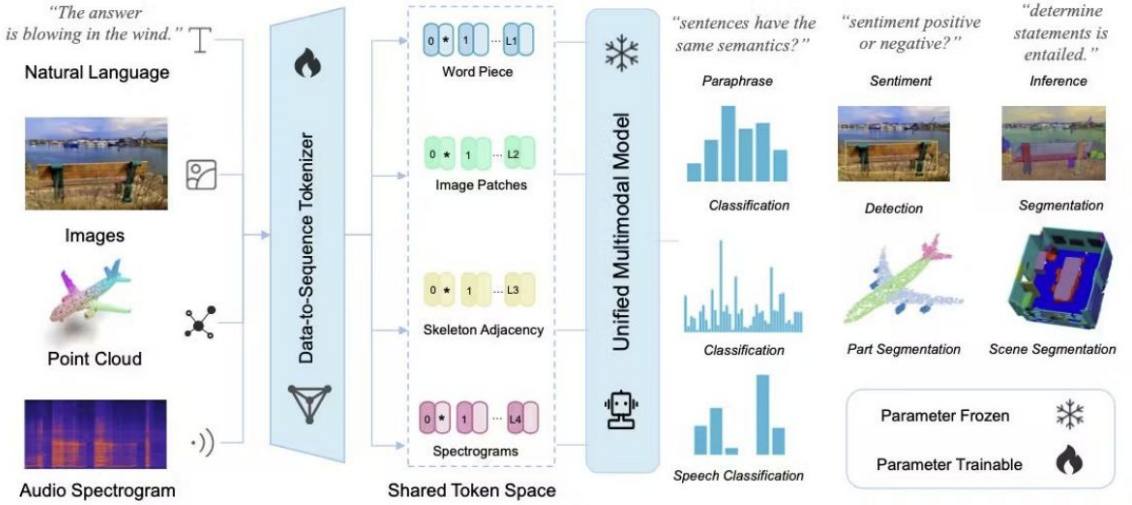
Add the next selected token to the end of the input sequence

 Source: https://lxmls.it.pt/2025/slides/lucas.pdf

Beyond text: 多模态模型

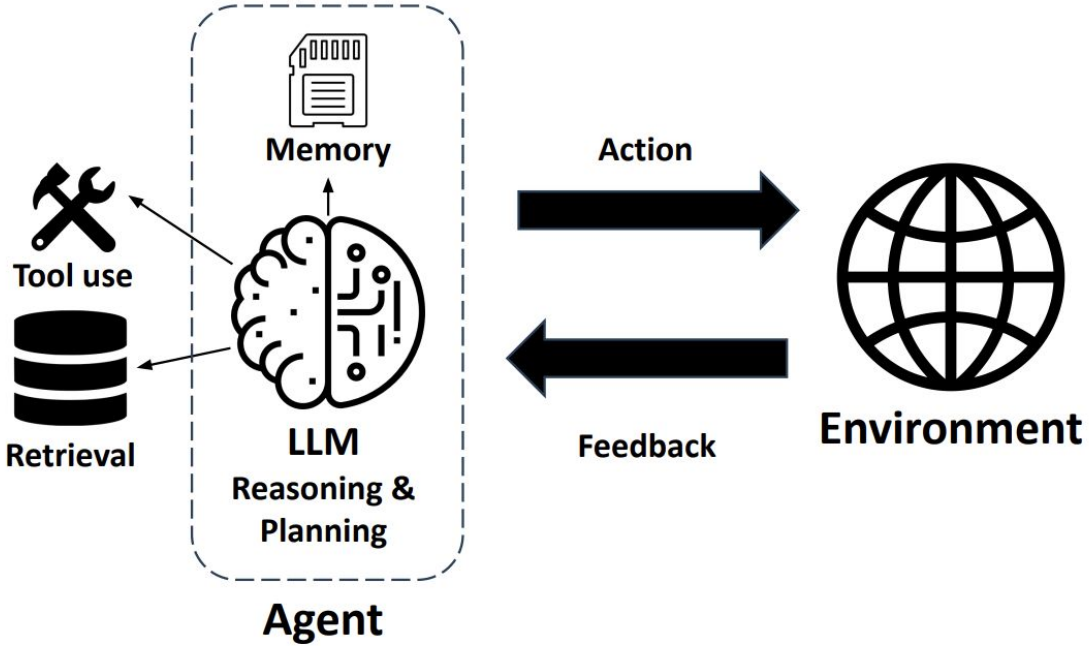
Transformers: Everything, everywhere, all at once

Anything once tokenized, can be passed through transformers



Source: https://lxmls.it.pt/2025/slides/sweta.pdf

LLM Agent：与环境交互的大语言模型



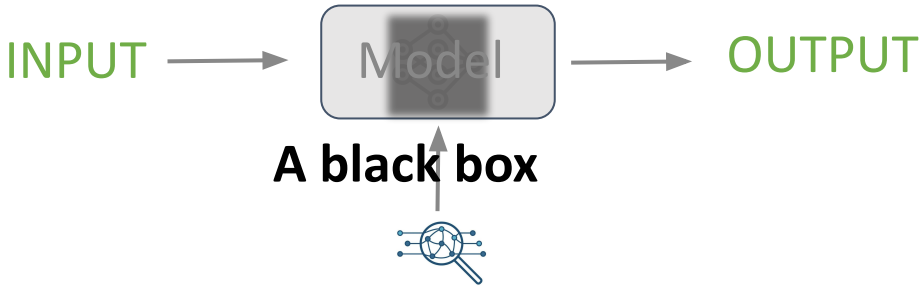
Source: https://llmagents-learning.org/slides/llm-agents-berkeley-intro-sp25.pdf

目录

Human-Inspired Interpretability of LLMs

Understanding LLMs via Mechanistic Interpretability (MI) tools

How does a model arrive at its conclusions?



Mechanistic interpretability
investigates internal representations, neurons and circuits within LLMs
correlates them with interpretable properties or functions

Both human brain and LLM are black-boxes

→ meaningful to adapt investigation methods for human brain to LLM mechanism research. (ref. Keynote talk of Prof. Tom Griffiths at EMNLP 2024)

"Biology" of LLMs (Anthropic)

Why unveiling the “Black Box” matters?

Towards safer LLMs:

- **Opacity of Generative AI**
  - The abilities are “emergent” rather than directly designed
- **Risk Identification & Prevention**
  - Harmful behaviors maybe also “emergent”
- **Misuse & Security Concerns**
  - Jailbreaks, …
- **High-Stakes Settings**
  - Finance, medical, law, …

The Urgency of Interpretability, Anthropic, 2025.4

Credits to Zeping YU

Decoding probing: Revealing internal linguistic structures and conceptual understanding of language models using minimal pairs

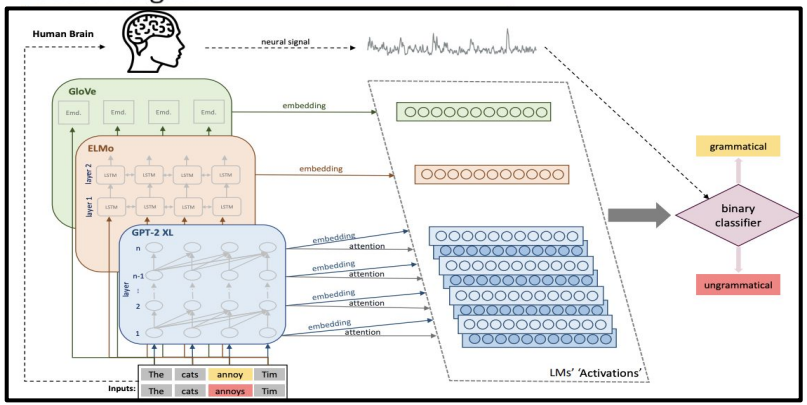
(He & **Nie** et al., ACL 2025 Findings; He & Chen & **Nie** et al., LREC-COLING 2024)

Understanding Language Models via probing techniques

- **Probing:** Investigating the **information encoded** in the models and the model **properties**

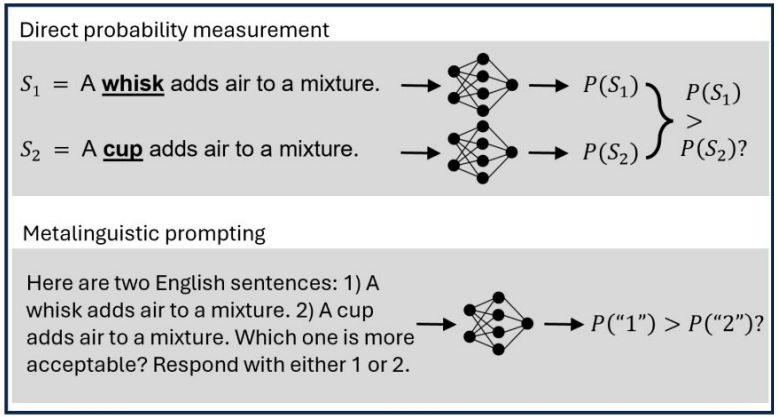
Probing from Neuro- vs. Psycholinguistic Perspectives:

Neurolinguistic



Diagnostic Probing (implicit)

Psycholinguistic



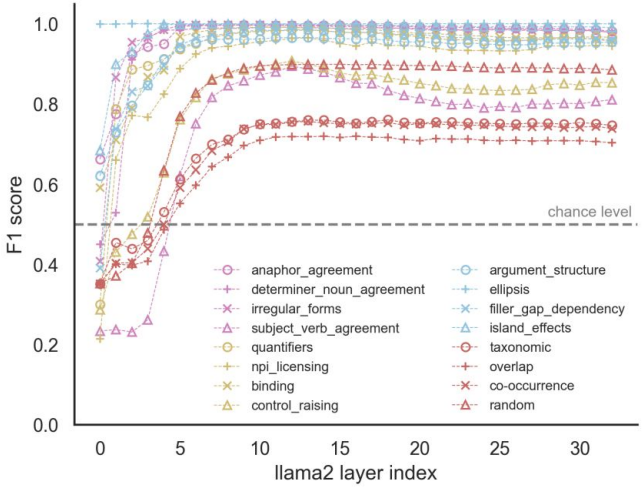
Probing via Prompting (explicit)

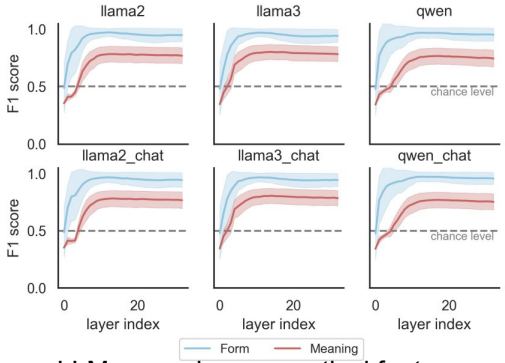
- **Psycholinguistic paradigm** measures the model's **output probabilities**, directly reflecting the model's behavior and performance.
- **Neurolinguistic paradigm** delves into the **internal representations** of LLMs.

Minimal Pair Probing for Linguistic Form and Meaning

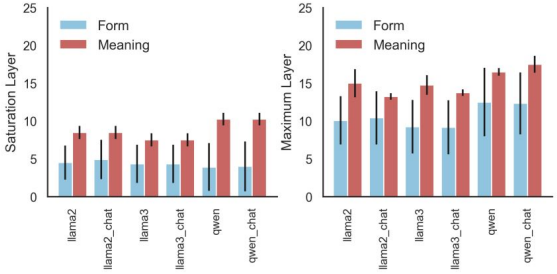
Form: Grammatical phenomena

Meaning: Conceptual understanding

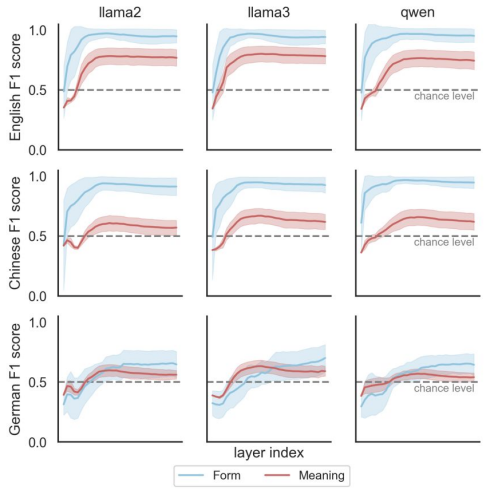




LLMs encode grammatical features better than conceptual features.



LLMs encode meaning after form.

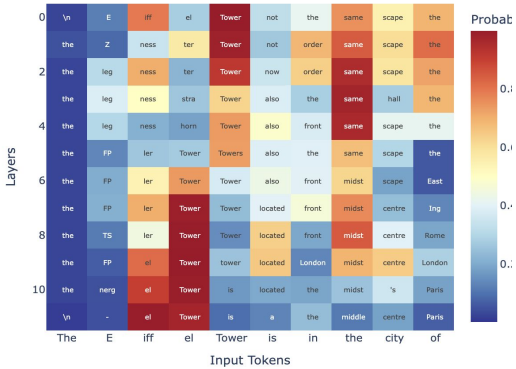


Disparity of form and meaning competence across languages.

Interpreting LLMs: Look into weight matrices, activations and logits

LogitLens / TunedLens

Logit Lens Visualization

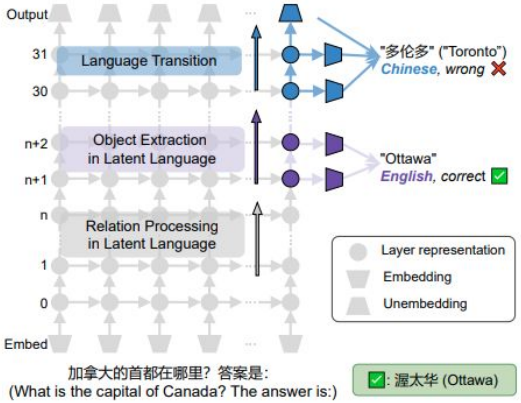


https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

Do Llamas work in English?

(Wendler et al., ACL 2024)

Our recent study employs LogitLens to dissect cross-lingual factual Inconsistency in English-centric and multilingual LLMs:

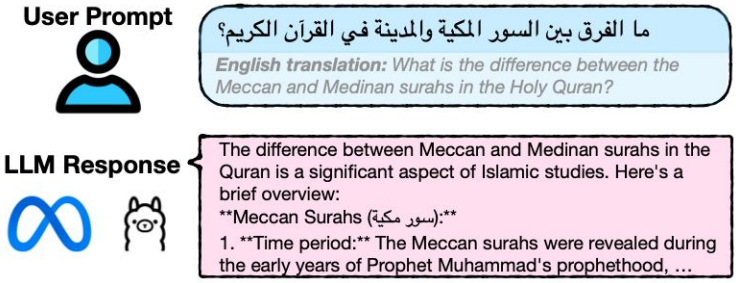


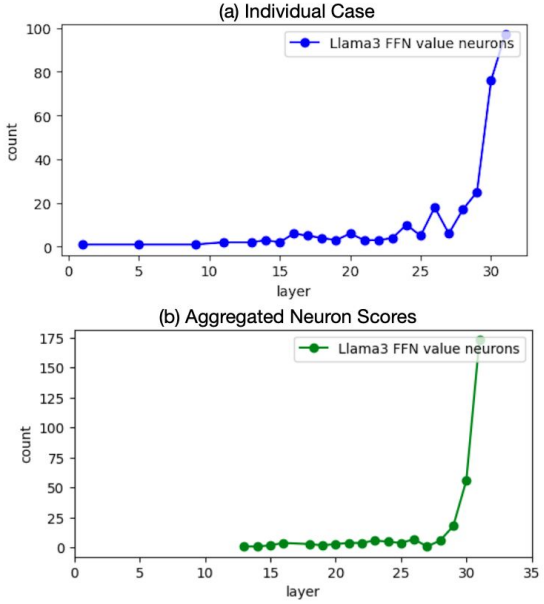
Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models **(ACL 2025)**

Interpreting LLMs: Look into weight matrices, activations and logits

Neuron Location and Intervention

Mechanistic understanding and mitigation of **language confusion**





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，...

Case 2 (okapi-zh, id5)

Input: 撰写一篇有关健康饮食重要性的争论性文章。*Write an argumentative essay on the importance of healthy eating.*

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.

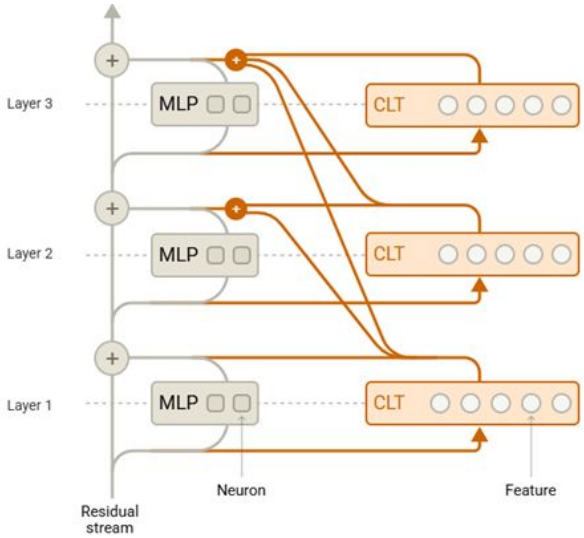
After Neuron Editing: 英语作文\n健康饮食的重要性：是或不是？\n在当今的生活中，健康饮食已经成为人们日常生活中的一部分。然而，是否健康饮食真的那么重要？

Multilingual Llama3:](https://www.zhihu.com/question/34614445) \n[如何在日常生活中更好地保持健康饮食习惯？](https://www.zhihu.com/question/34614445)

Interpreting LLMs: Look into weight matrices, activations and logits

Replacement model for interpretability

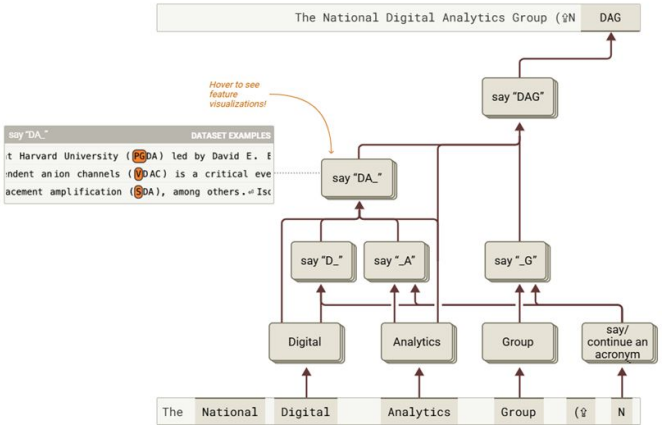
Sparse Autoencoder (SAE), Cross-Layer Transcoder (CLT)



CLT architecture by Anthropic

Mapping neurons to a sparse space, each sparse “neuron” in the replacement model represents a human-understandable feature

Tracing language processing mechanism of LLMs through **attribution graph** extracted based on the replacement model (e.g., CLT)



Interpreting LLMs: Look into weight matrices, activations and logits

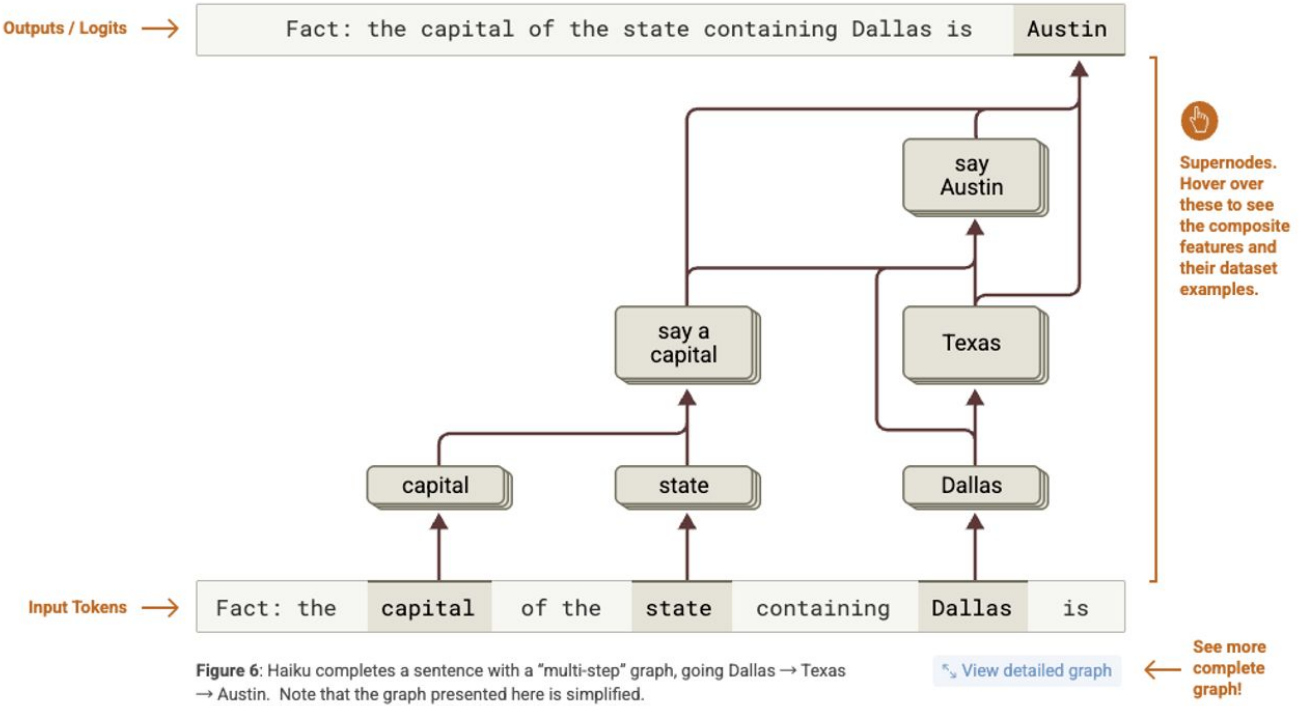


Figure 6: Haiku completes a sentence with a "multi-step" graph, going Dallas → Texas → Austin. Note that the graph presented here is simplified.

On the Biology of a Large Language Model, Anthropic, 2025.3

Credits to Zeping YU

Sept. 2025

Interpreting LLMs: Look into weight matrices, activations and logits

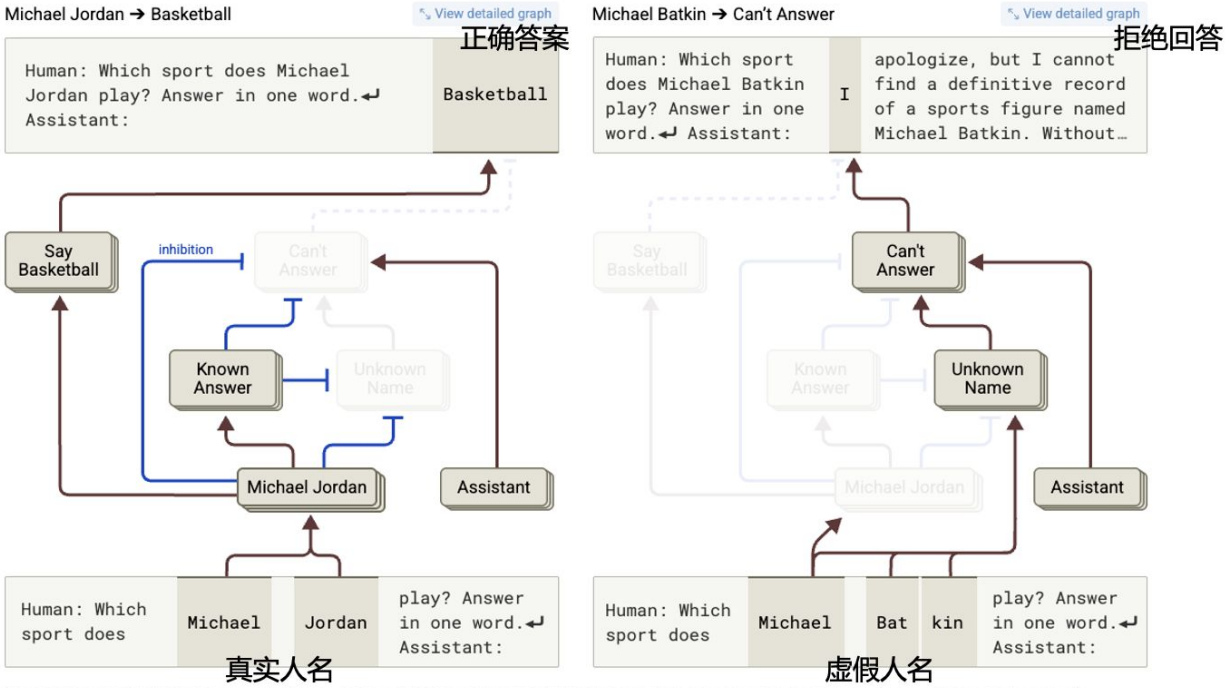


Figure 33: Two simplified attribution groups for Michael Jordan and a fictitious "Michael Batkin". Haiku is asked what sport they play. For Jordan, Haiku correctly responds because of a known answer pathway, and because a unknown name pathway is inhibited. For Batkin, the opposite occurs. Blue edges with T-shaped ends indicate *inhibitory inputs* (negative edge weights). This diagram is interactive and you can hover on nodes to see detailed feature visualizations.

On the Biology of a Large Language Model, Anthropic, 2025.3

Credits to Zeping YU

Sept. 2025

目录

以人为中心的大语言模型+

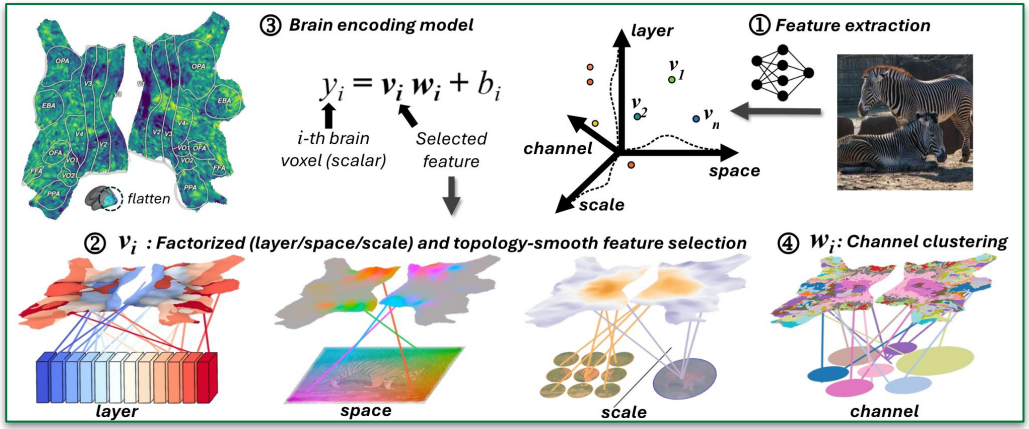
大模型与各领域结合的潜力

- LLM具备强大的知识整合、推理和生成能力，为多学科研究带来新范式。
- 跨学科合作可拓展研究视野，提升创新深度和实际影响力。

人类认知启发的大语言模型建模

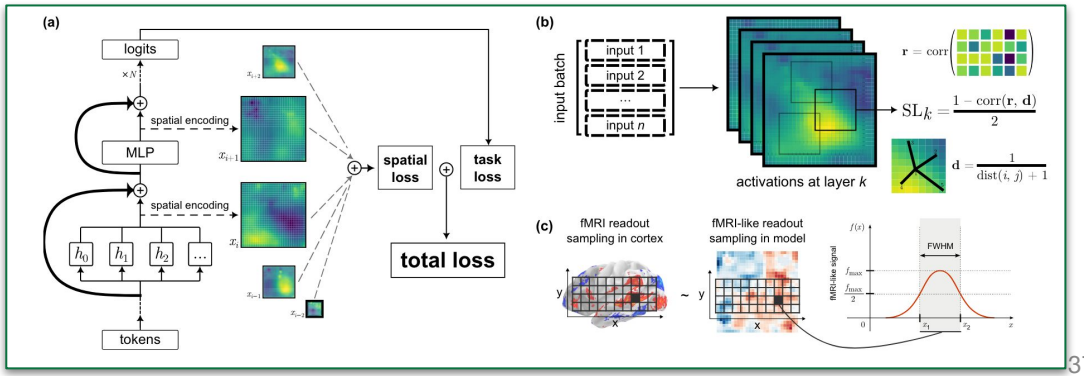
人类 → 人类认知启发建模 → LLM

- 结合脑科学、神经科学方法，研究LLM与人脑在语言处理、推理等方面的异同，通过脑成像、认知实验等手段，探索LLM的“类脑”机制。



Yang et al. (2024)

- 语言模型的类脑建模：
  大脑结构启发语言模型的建模



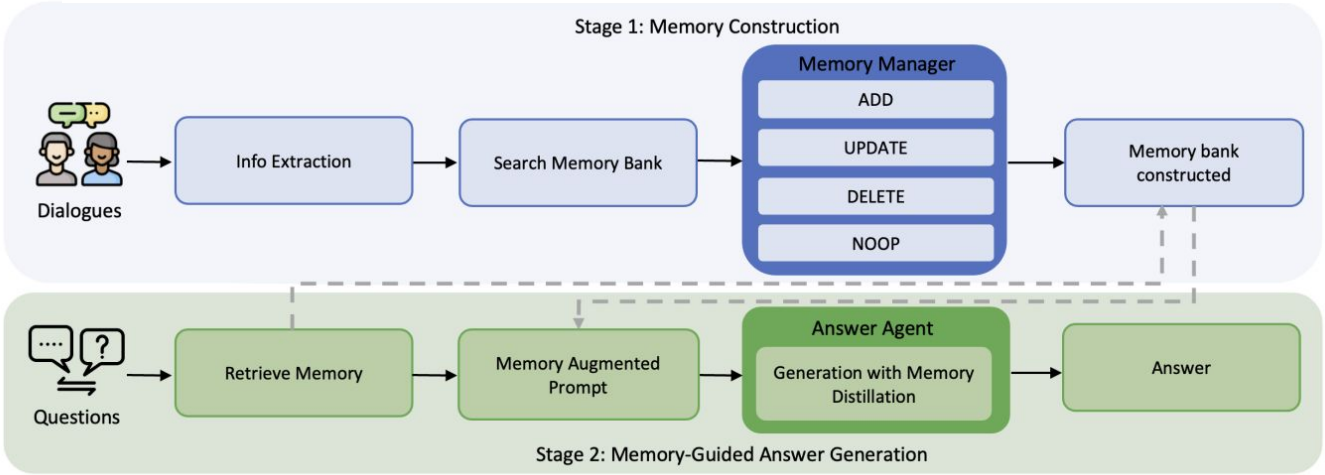
TOPOLM: BRAIN-LIKE SPATIO-FUNCTIONAL ORGANIZATION IN A TOPOGRAPHIC LANGUAGE MODEL
Rathi et al. (2025)

人类记忆启发的大语言模型建模

Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning

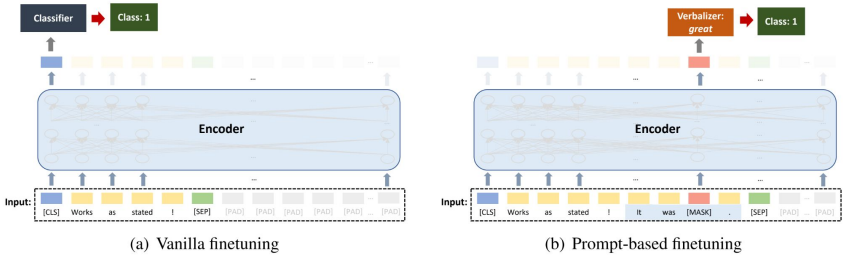
Sikuan Yan[*1], Xiufeng Yang[*2], Zuchao Huang[1], Ercong Nie[1], Zifeng Ding[3], Zonggen Li[4], Xiaowen Ma[1], Hinrich Schütze[1], Volker Tresp[1], Yunpu Ma[1†]

[1]Ludwig Maximilian University of Munich [2]Technical University of Munich
[3]University of Cambridge [4]University of Hong Kong
s.yan@campus.lmu.de, cognitive.yunpu@gmail.com



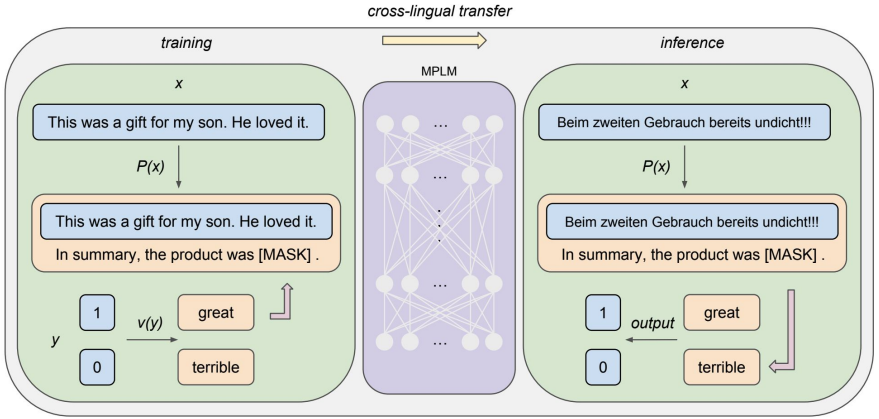
面向低资源语言的NLP技术

Prompt-based fine-tuning

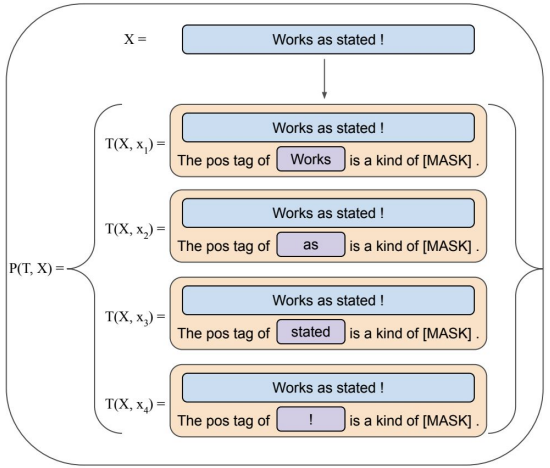


(a) Vanilla finetuning (b) Prompt-based finetuning

Prompt-based fine-tuning for zero-shot cross-lingual transfer learning



What about sequence labeling tasks?
(e.g. part-of-speeching tagging, named entity recognition)
→ **ToPro: Token-level prompt decomposition**



Training on English data: prompt pattern, verbalizer, fine-tuning by mask token prediction.
Inference in the cross-lingual setting:

- input given in target languages
- no changes in prompt pattern, verbalizer

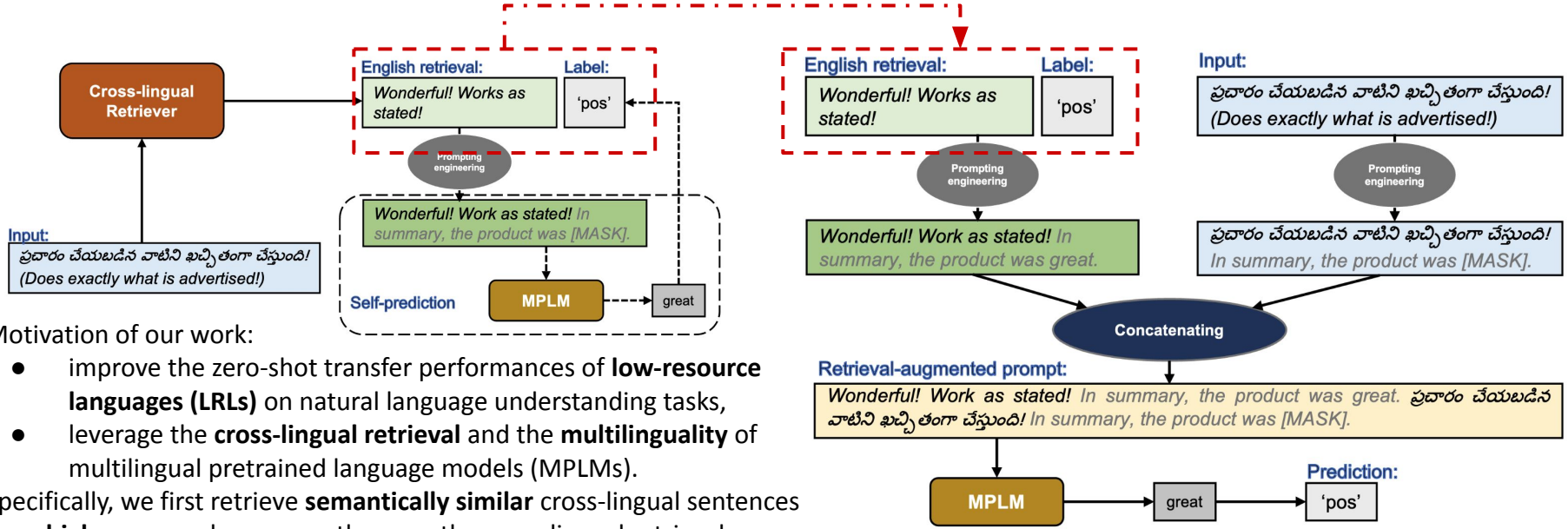
ToPro: Token-Level Prompt Decomposition for Cross-Lingual Sequence Labeling Tasks (Ma* & Nie*, et al., EACL 2024)

面向低资源语言的NLP技术

The pipeline of our proposed PARC method

Step 1: Retrieval from high-resource language corpora

Step 2: Prediction with a retrieval-augmented prompt



Motivation of our work:

- improve the zero-shot transfer performances of **low-resource languages (LRLs)** on natural language understanding tasks,
- leverage the **cross-lingual retrieval** and the **multilinguality** of multilingual pretrained language models (MPLMs).

Specifically, we first retrieve **semantically similar** cross-lingual sentences from **high-resource** languages, then use the cross-lingual retrieval information to benefit the LRLs from the **multilinguality** of MPLMs.

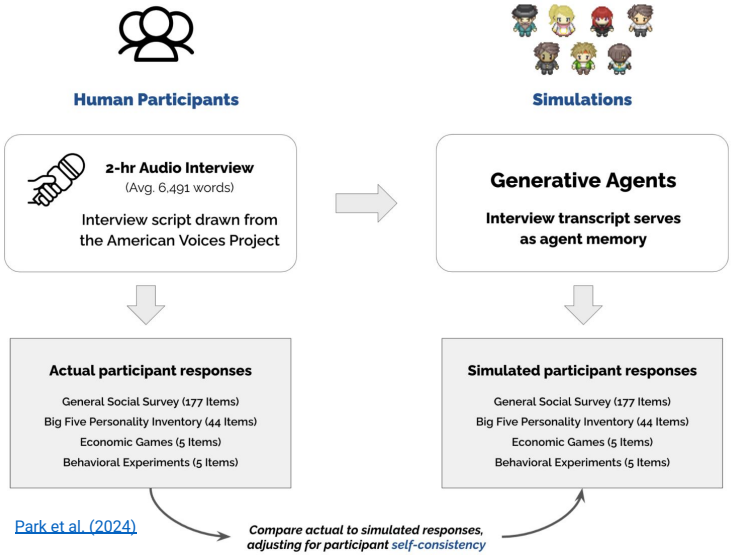
大语言模型社会仿真及建模

LLM persona与社会科学
利用LLM persona模拟，进行社会调查、问卷研究、社会建模。

人格化、文化敏感的LLM persona

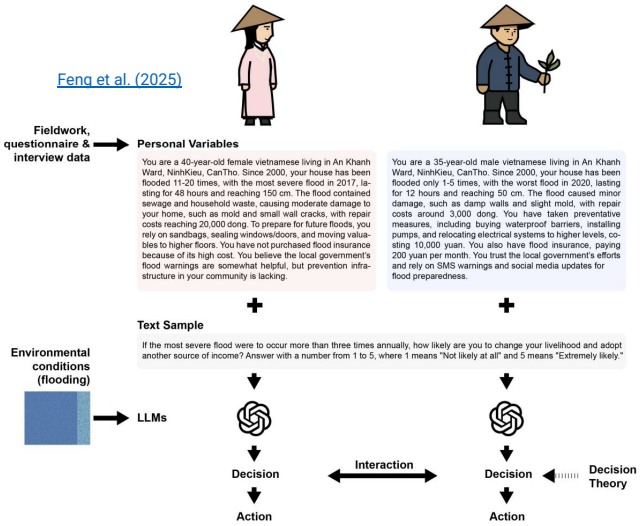
- 机器心理学（Machine Psychology）

Hargendorff et al. (2024)



典型案例：斯坦福“Generative Agent Simulations”——1000+ LLM personas模拟虚拟社会行为。

Feng et al. (2025)



社会建模：结合地理、气候等领域，模拟灾害情境下的人类反应与社会动态。

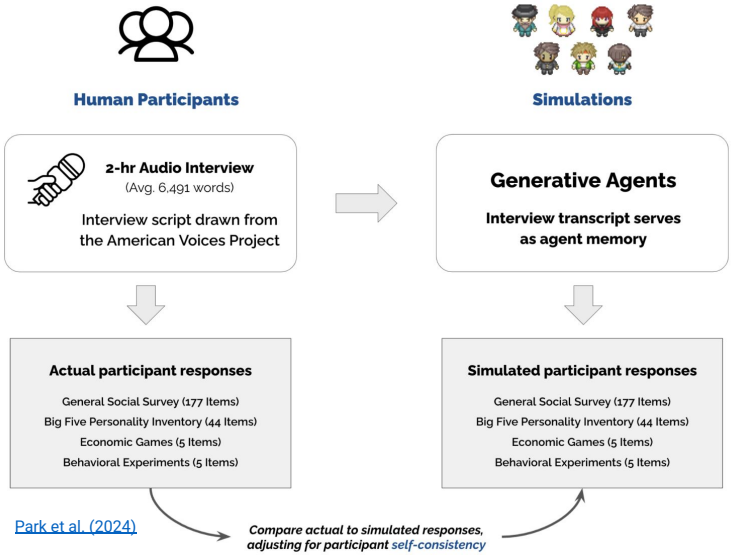
大语言模型社会仿真及建模

LLM persona与社会科学
利用LLM persona模拟，进行社会调查、问卷研究、社会建模。

人格化、文化敏感的LLM persona

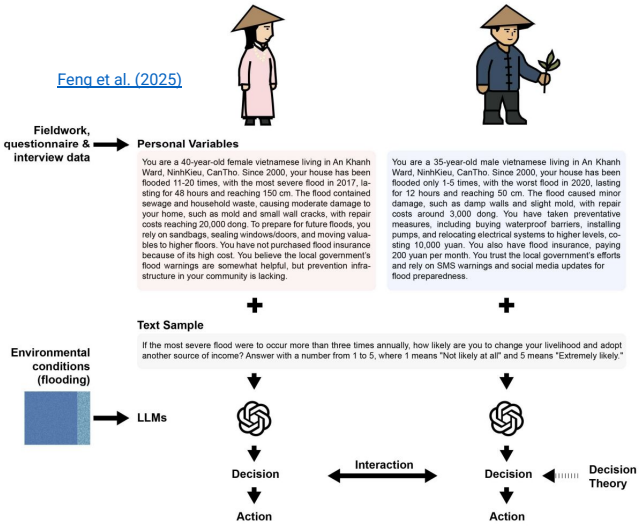
- 机器心理学（Machine Psychology）

Hargendorff et al. (2024)



典型案例：斯坦福“Generative Agent Simulations”——1000+ LLM personas模拟虚拟社会行为。

Feng et al. (2025)



社会建模：结合地理、气候等领域，模拟灾害情境下的人类反应与社会动态。


LLM
人类认知启发LLM建模
LLM对齐服务人类
社会规范推动AI治理和负责任AI
LLM社会仿真与建模
以人为中心的大模型交叉学科应用
人类
社会

References

- Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
- Nie, Ercong, et al. "Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages." Findings of the Association for Computational Linguistics: ACL 2023. 2023.
- Ma, Bolei, et al. "Is Prompt-Based Finetuning Always Better than Vanilla Finetuning? Insights from Cross-Lingual Language Understanding." KONVENS. 2023.
- Wang, Lean, et al. "Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.
- Yuan, Shuzhou, et al. "GNNavi: Navigating the Information Flow in Large Language Models by Graph Neural Network." Findings of the Association for Computational Linguistics ACL 2024. 2024.
- He, Linyang, et al. "Decoding Probing: Revealing Internal Linguistic Structures in Neural Language Models Using Minimal Pairs." Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024.
- Nie, Ercong, Helmut Schmid, and Hinrich Schütze. "Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models." arXiv preprint arXiv:2505.16538 (2025).
- Wendler, Chris, et al. "Do llamas work in english? on the latent language of multilingual transformers." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.
- Wang, Mingyang, et al. "Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models." Proceedings of ACL 2025.2025.
- Park, Joon Sung, et al. "Generative agent simulations of 1,000 people." arXiv preprint arXiv:2411.10109 (2024).
- Feng, Wenhan, et al. "Generative agent-based modeling for climate adaptation policy: A flood resilience perspective." Proceedings of Social Simulation Conference 2025. 2025.
- Hagendorff, Thilo. "Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods." arXiv preprint arXiv:2303.13988 1 (2023).
- Yang, Huzheng, James Gee, and Jianbo Shi. "Brain decodes deep nets." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
- Rathi, Neil, et al. "TopoLM: brain-like spatio-functional organization in a topographic language model." The Thirteenth International Conference on Learning Representations.

感谢聆听！

主页：https://ercong21.github.io/
邮箱：nie@cis.lmu.de

欢迎关注Munich NLP公众号

关于我们

我们是一群生活在🇩🇪德国慕尼黑🍺的NLP爱好者。我们在这里向大家分享组里的研究成果和感兴趣的论文。我们也欢迎慕尼黑内外的NLPer踊跃投稿。

#MunichNLP #LMU #TUM #NLP

Sept. 2025

Munich NLP