PREPRINT

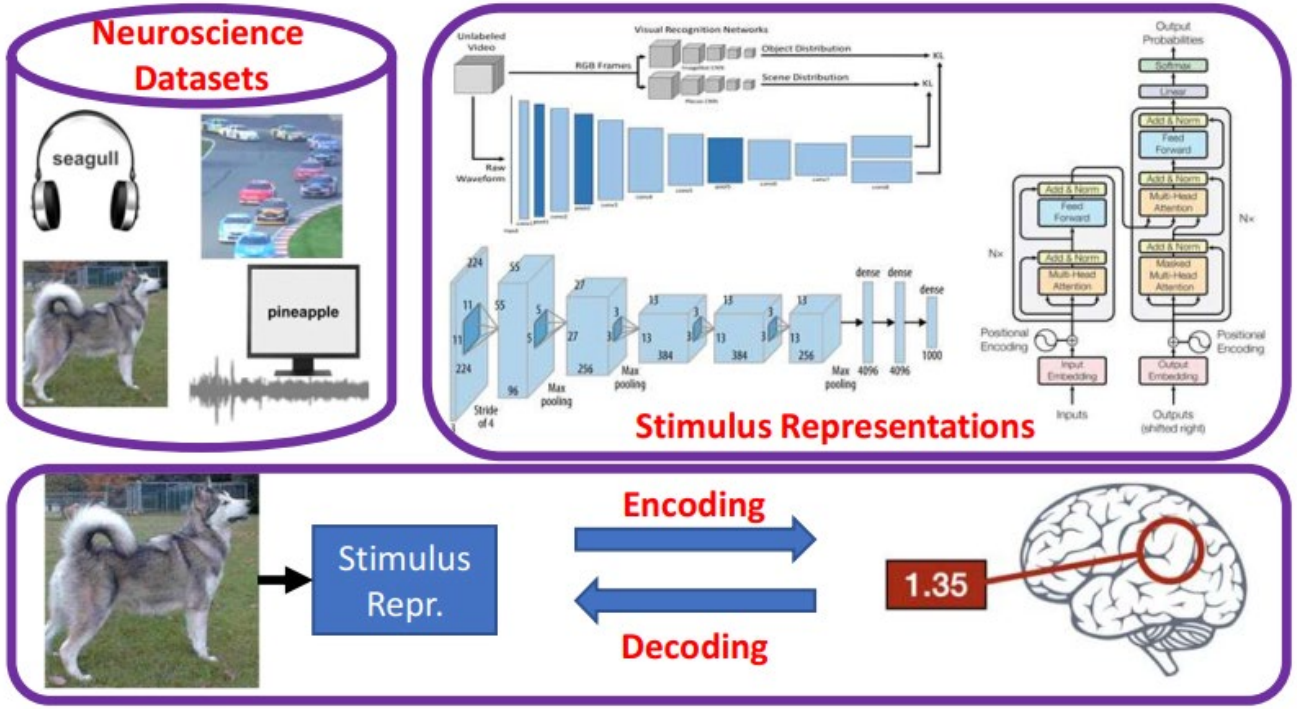
Explanations of Deep Language Models Explain Language Representations in the Brain

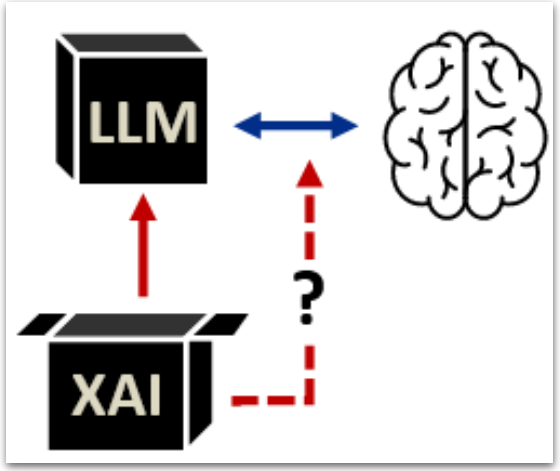
Maryam Rahimi, Yadollah Yaghoobzadeh, Mohammad Reza Daliri

Biomedical Engineering Department, School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
Electrical and Computer Engineering Department, University of Tehran, Tehran, Iran
2025.2.21

脑编码框架概述

- 假设：语言在人脑中存在固定模式的表征空间
- 目的：探索人脑的表征空间
- 方法概述：人脑接收刺激，产生神经反应。用刺激预测反应。具体来说，神经反应用数字进行记录，将刺激映射到向量空间中，再用向量通过线性回归预测神经反应。





存在的问题

- 语言模型的表征空间**不透明**，较好拟合人脑是系统性相似？建模提升？
- 有无**其他表征空间**可以拟合人脑活动记录？具有更强的可读性？

语言模型的可解释性方法，特别是基于样本分析的方法，提供每个token的数字特征空间。梯度，注意力。

问题介绍

想法来源

方法

实验

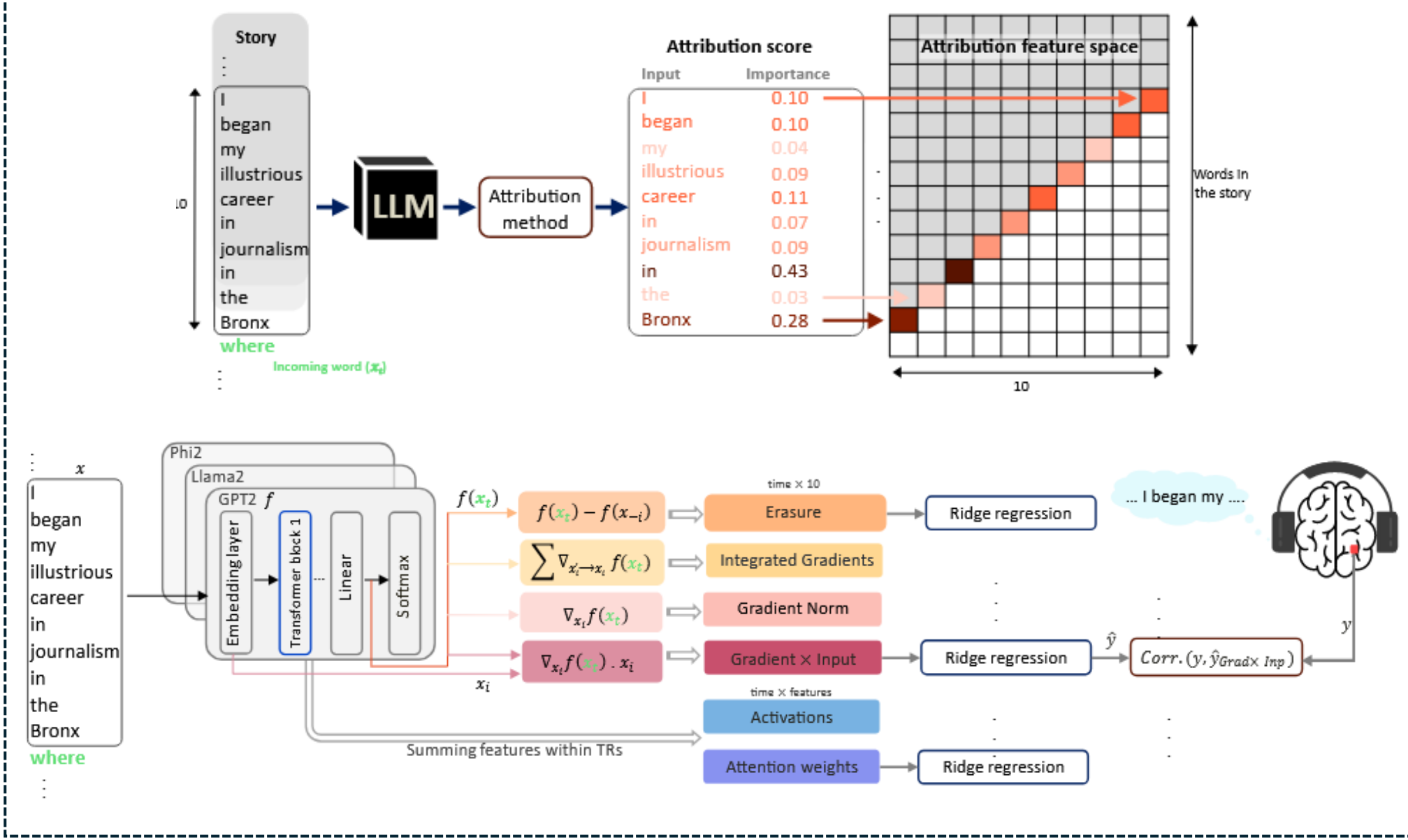
可预测性

归因vs表示

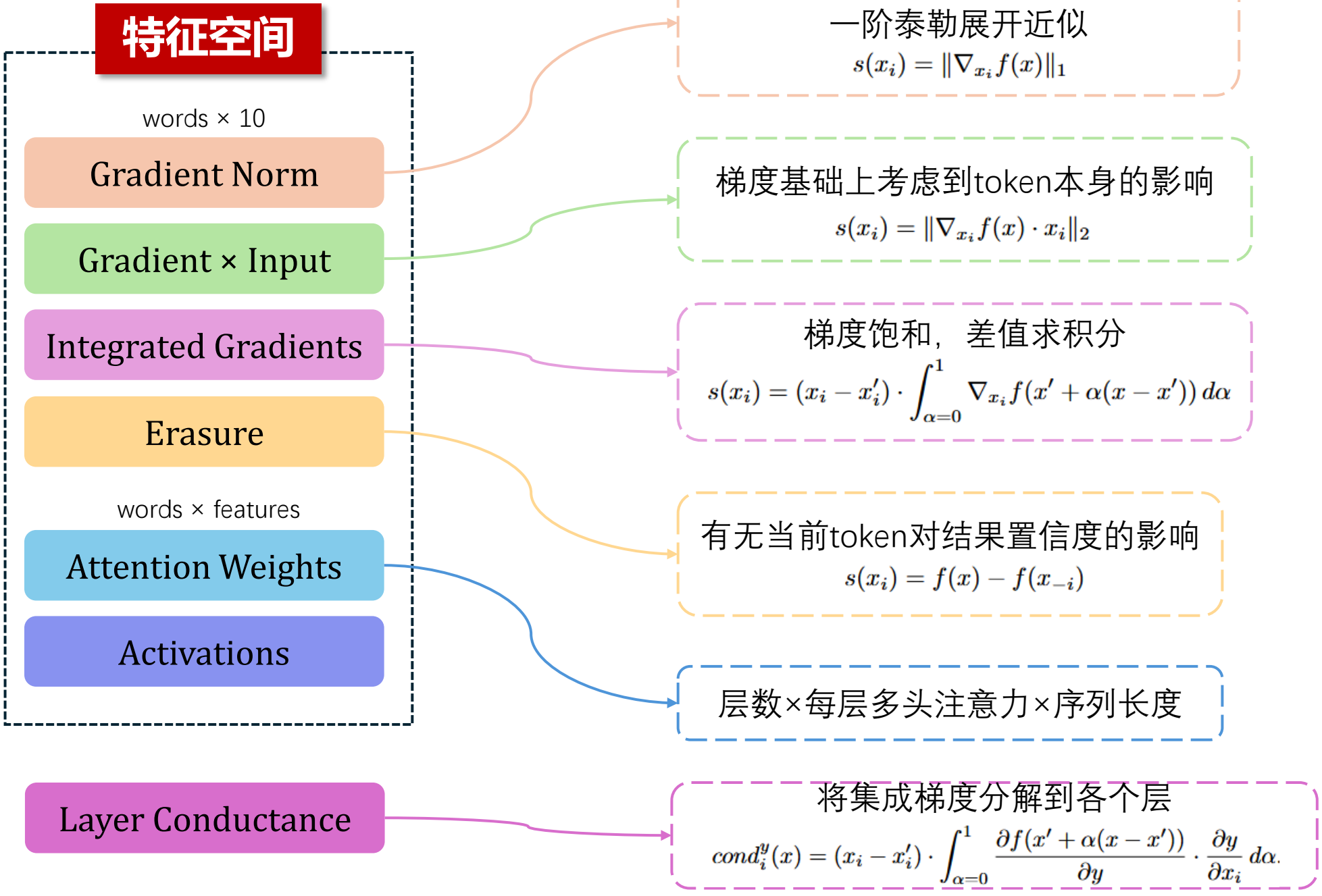
层次化

总结与讨论

整体框架



- 问题介绍
- 想法来源
- **方法**
- 实验
  - 可预测性
  - 归因vs表示
  - 层次化
- 总结与讨论



问题介绍

想法来源

方法

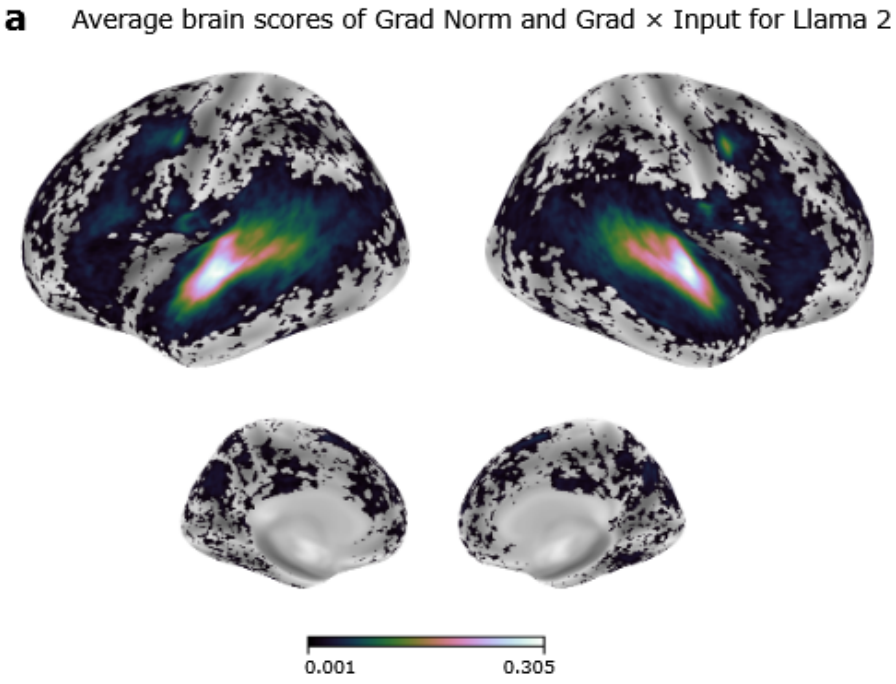
实验

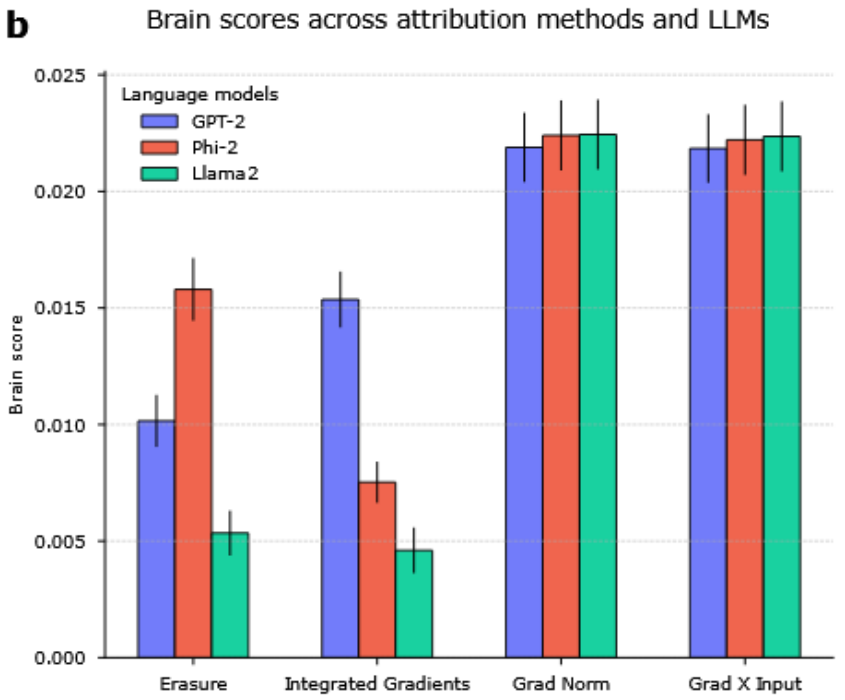
可预测性

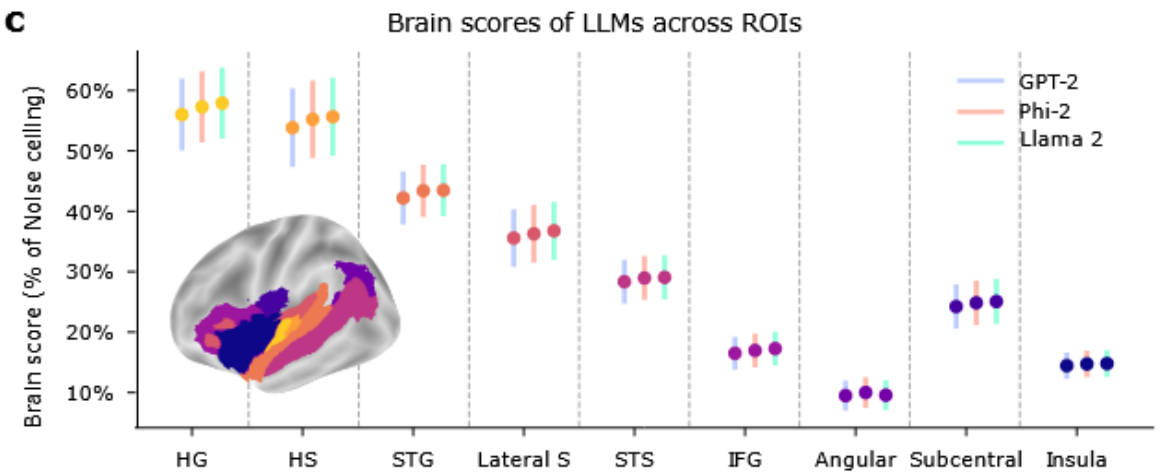
归因vs表示

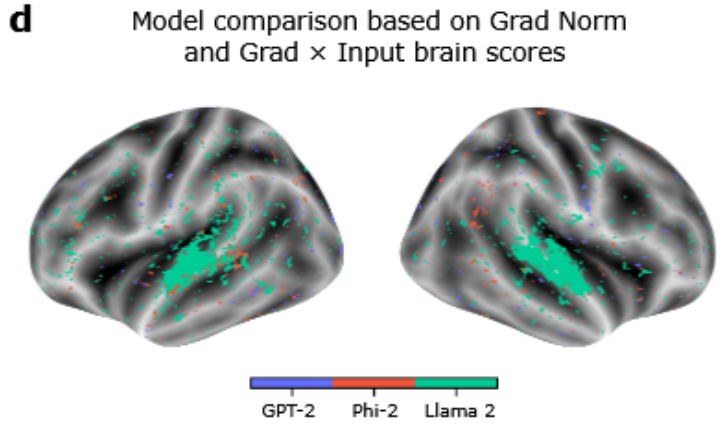
层次化

总结与讨论

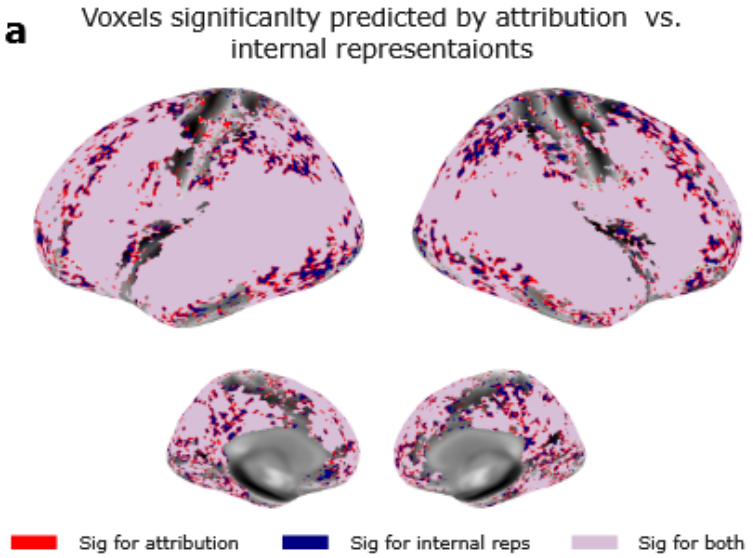


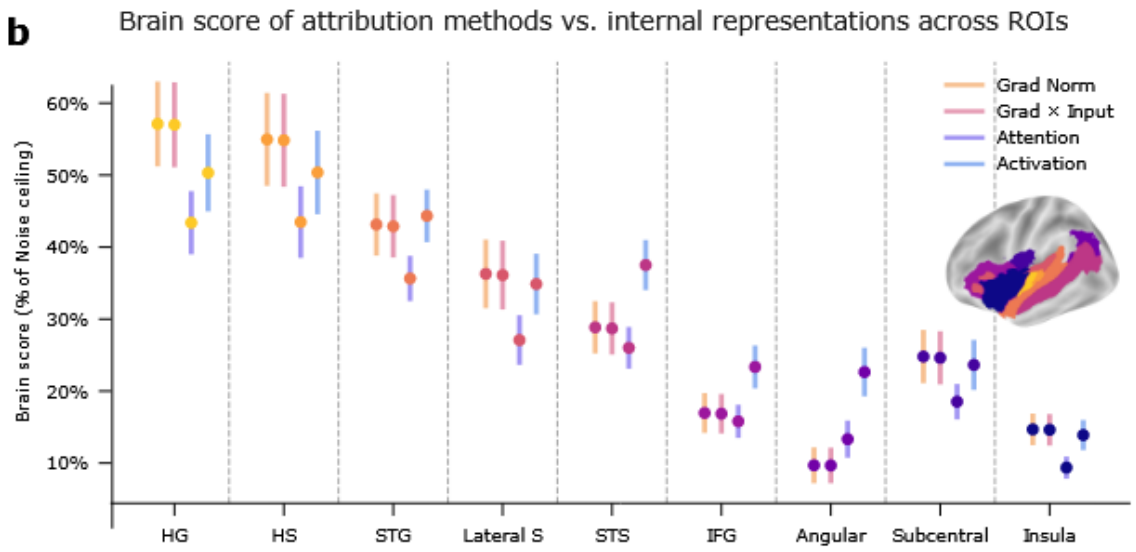






- 问题介绍
- 想法来源
- 方法
- **实验**
  - 可预测性
  - 归因vs表示
  - 层次化
- 总结与讨论





Abbreviation	Full Name
HS	Heschl's Sulcus
HG	Heschl's Gyrus
STG	Superior Temporal Gyrus
STS	Superior Temporal Sulcus
IFG	Inferior Frontal Gyrus
Lateral S	Lateral Sulcus (Sylvian Fissure)

初级听觉皮层，音高知觉的核心脑区

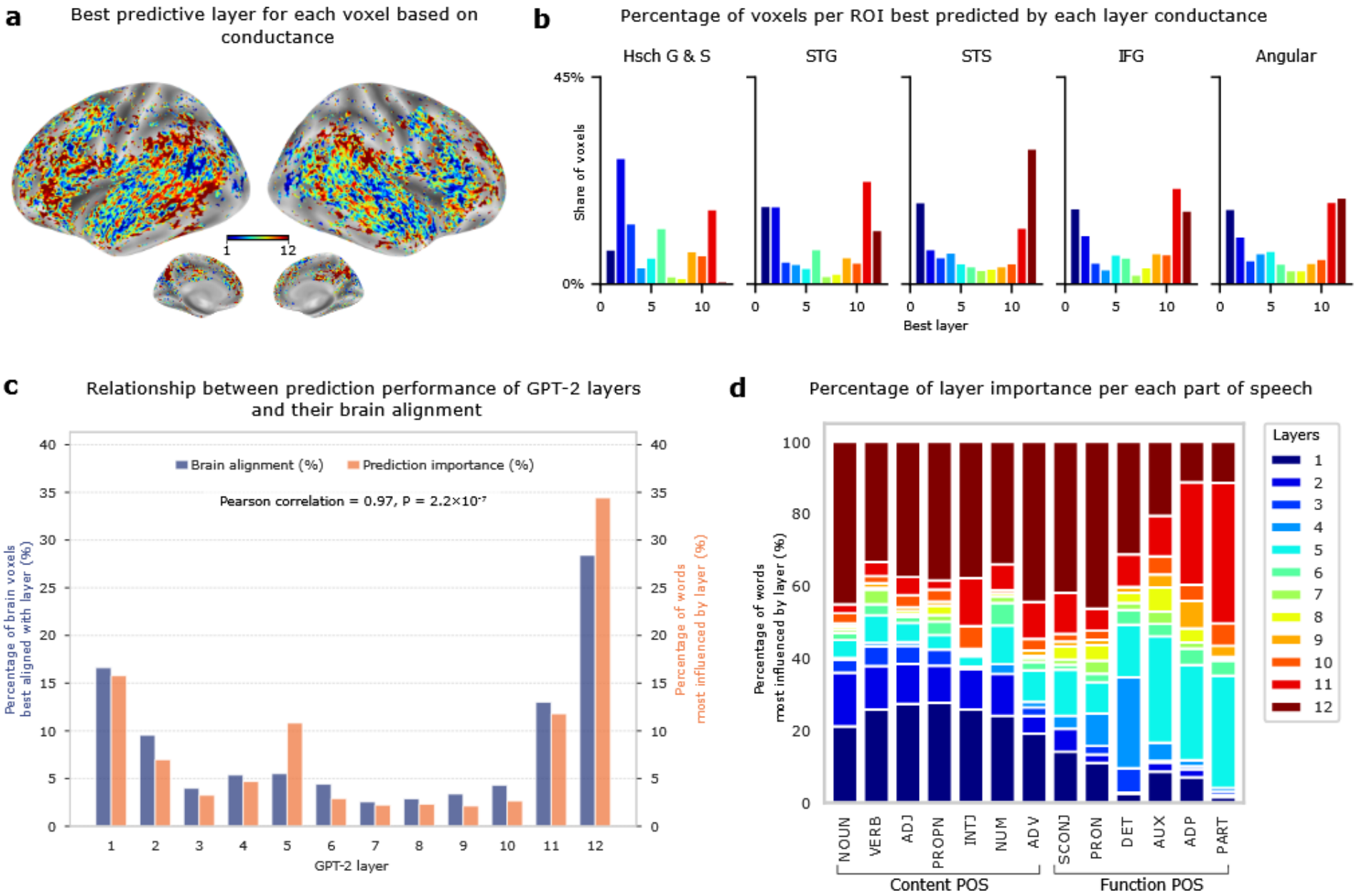
听觉语言中枢，语言与记忆交互

语言理解、语言产生、情感控制

信息交互

归因元素较好拟合浅层语言处理区域，激活元素较好拟合高级语言处理区域

- 问题介绍
- 想法来源
- 方法
- **实验**
  - 可预测性
  - 归因vs表示
  - **层次化**
- 总结与讨论



人脑具有层次化的思维特点

总结与讨论

- llm和大脑不仅以类似的方式编码语言，而且在它们的表征**如何适应不断变化的输入**方面表现出并行动态。
- 归因方法量化了每个前一个单词对模型下一个单词预测的贡献。大脑和llm之间用于**整合前文的共享权重机制**的证据。
- 基于属性的解释内在地编码了序列中下一个单词**更实际的信息**。相比之下，内部表示包含**预测和非预测信息的混合**，使它们更具泛化性，但对模型的下一个单词预测任务不那么具体。

Thanks!