



上海交通大學

SHANGHAI JIAO TONG UNIVERSITY



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

神经元干预能否诱导 大语言模型呈现阿尔茨海默病 相关语言表型？

聂耳聪 Ercong Nie
上海交通大学外国语学院

语言学午餐会 (第113期)



研究背景与个人简介



聂耳聪

上海交通大学外国语学院 助理教授



德国慕尼黑大学计算语言学博士 (summa cum laude)

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。

Using language science to better connect, understand, and build AI.

Using AI to advance language science.

以语言科学为理解智能的入口和连接智能的媒介，以人工智能拓展语言科学。

Multilingual AI

LLM Interpretability

Language & Cognition

AI for Humanities

Language Science for AI (LS4AI) & AI for Language Science (AI4LS)



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

LS4AI

语言科学帮助理解与改进AI

01

Language Science for AI

We use linguistic knowledge, multilingual evidence, and human language-processing research to explain how language models work, diagnose cross-lingual and cross-cultural failures, and design evaluations that better reflect human language behavior.

- Mechanistic interpretability of multilingual LLMs
- Cross-lingual knowledge, transfer, and language confusion
- Human-inspired evaluation and model design

AI4LS

AI提供语言与认知研究工具

02

AI for Language Science

We use language models and computational methods as research instruments for linguistics, cognition, historical language research, digital humanities, and neighboring human and social sciences.

- Computational models of language and cognition
- Low-resource and historical language technologies
- AI-assisted research in the humanities and social sciences



两条研究主线：多语言 × 可解释性

1

大模型多语言跨文化感知的机制解释与评测

研究大语言模型如何获取、表示、编辑、检索和使用不同语言中的语言知识与事实知识，并解释语言混淆、跨语言不一致和跨文化偏差等现象。

2

人类语言认知加工与大模型机制可解释性的对比研究

将心理语言学、神经语言学和认知科学中的实验范式与模型内部表征分析相结合，比较人类与大模型在形式、意义、知识和推理过程中的异同。

文化
感知

多语言

可解释性

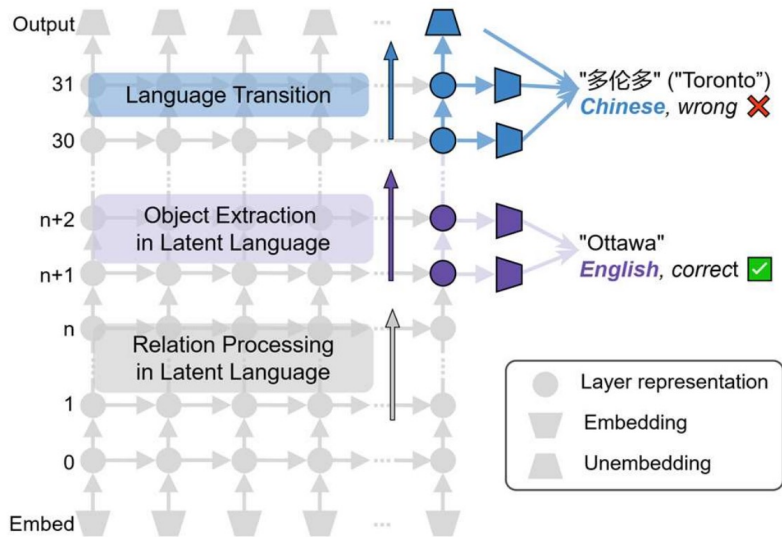
人类语言
加工参照

围绕“多语言 × 可解释性”主线的研究工作 (1) : 缓解大语言模型的跨语言事实不一致



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

同一事实，不同语言得到不同答案

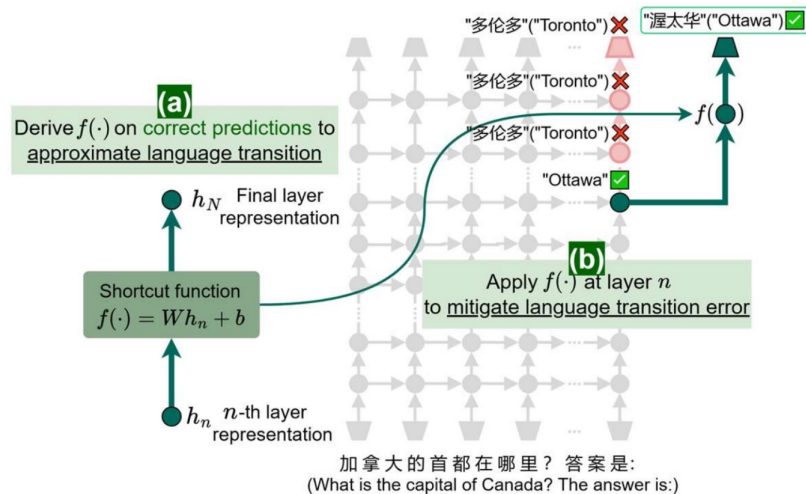


加拿大的首都在哪里？答案是：
(What is the capital of Canada? The answer is:)

✅: 渥太华 (Ottawa)

**机制：中间层可提取正确答案，
末层语言转换仍可能失败。**

跨语言事实一致性的机制干预



加拿大的首都在哪里？答案是：
(What is the capital of Canada? The answer is:)

线性截断方法：

以正确预测学习语言转换映射，绕过末层部分计算，检验目标语言回答是否改善。

围绕“多语言 × 可解释性”主线的 研究工作 (2) : 语言混淆——定位并调节关键神经元



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

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):****

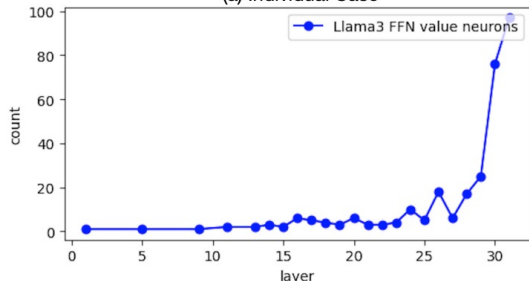
1. ****Time period:**** The Meccan surahs were revealed during the early years of Prophet Muhammad's prophethood, ...

行为：大模型生成与用户需求不符的语言

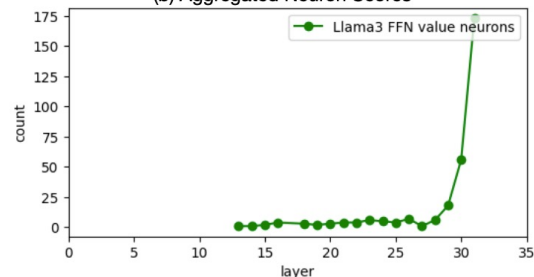
机制：末层转换与关键神经元

干预：定向编辑后混淆减少

(a) Individual Case



(b) Aggregated Neuron Scores



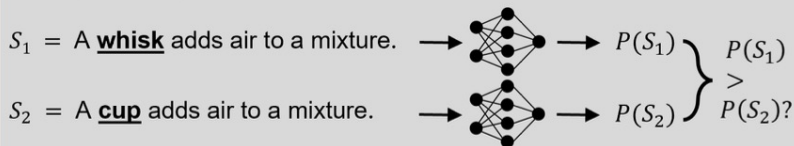
围绕“多语言 × 可解释性”主线的研究工作 (3) : 大语言模型作为语言学“被试”



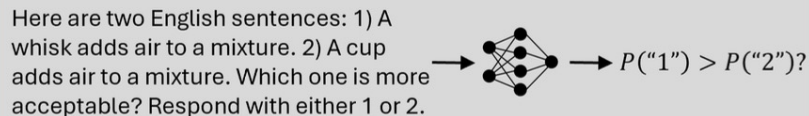
上海交通大学外国语学院
School of Foreign Languages

Psycholinguistic

Direct probability measurement

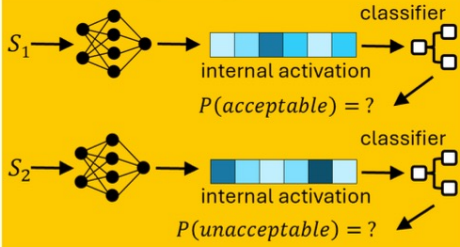


Metalinguistic prompting

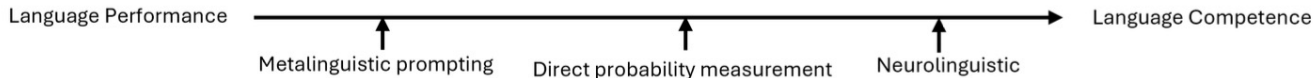


Neurolinguistic

Minimal pair probing



Diagnostic probing



- 设计语言学控制实验 (minimal pair最小对设计)
- **PyschoLing范式**: 测量模型的“行为反应”
- **NeuroLing范式**: 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 (ACL Findings 2025)

今天的报告:

围绕“多语言 × 可解释性”主线的最新研究工作——

阿尔茨海默病 (AD) 相关语言表型与大模型



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

研究对象 (大语言模型)	行为问题	机制检验
跨语言事实一致性	换一种语言为什么答错?	共享概念与语言转换
语言混淆	为什么切换到非目标语言?	关键神经元定位与编辑
形式与意义加工	答对与内部可解码性是否一致?	最小对与诊断探测
AD相关语言表型	临床差异表征是否参与行为?	激活引导的定向干预

今天的
报告

阿尔茨海默病: Alzheimer's Disease (AD)



研究概述

神经元干预能否诱导大语言模型呈现阿尔茨海默病相关语言表型？

- 利用阿尔茨海默病（AD）患者与健康老年人的图片描述转写文本
- 找出大模型（Qwen3-8B）里在 **AD 话语上更常被激活的前馈层神经元**
- 人为放大或削弱这些神经元的输出，系统分析大语言模型在**认知任务**（记忆、流利度、工作记忆、流程性话语、场景构建、指代消解）以及**语言特征**（词汇惊异度、观点密度、句法复杂度、产出量）的变化情况。

Activation-Guided Neuron Intervention to Induce Alzheimer's-Related Computational Language Phenotypes in a Large Language Model

R He, E Nie, H Jiang, IE Sommer, P Homan, W Hinzen
arXiv preprint arXiv:2608.03067



研究动机

- ❖ 语言变化是 AD 认知衰退中最容易观察到的表现之一。Speech and language 可以重复、远程、低成本采集，是有前景的数字标记物 (Robin et al., 2020)。



AD相关语言变化提供了什么研究线索?

语言层面	既有研究关注的变化	本研究的观测方式
词汇	检索困难、更多高频词	流畅性、词频surprisal
信息与句法	命题信息减少、句法简化	DEPID-R、从句比率
话语	事件遗漏、组织与指称困难	回忆、程序话语、共指
声音	停顿、犹豫、语速与韵律	本研究未分析音频



计算研究 AD 语言的三阶段

1. 预定义特征阶段

(Fraser 2015; Luz 2020)

- 声学特征：语速、停顿
- 语言学特征：词汇多样性、句法复杂度、词汇惊异度 (lexical surprisal)、信息密度

人类 AD 言语的典型发现 (Boschi 2017; Slegers 2018) :

- 更多犹豫 (hesitation)、找词困难 (word-finding difficulty)
- 词汇多样性和信息含量降低
- 更依赖高频词
- 句法复杂度降低

2. 分布式表征 / 语言模型阶段

(Colla 2022; Gkoumas 2023; Jiang 2025)

- 语义相似度改变
- 困惑度 (perplexity) 更高
- 刺激-应答对齐 (stimulus-response alignment) 变弱

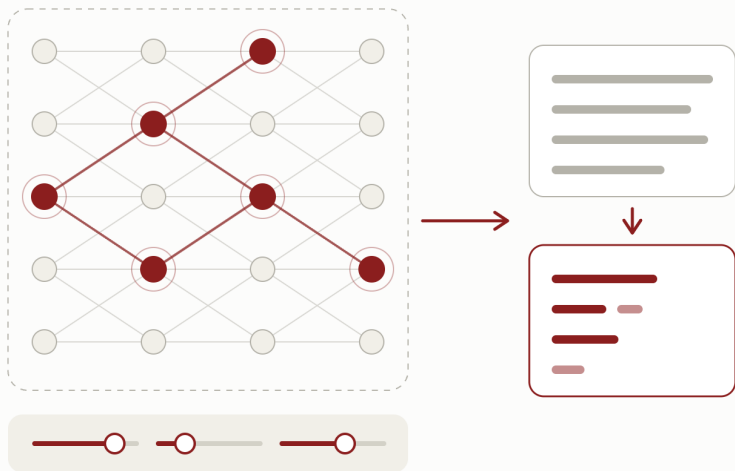
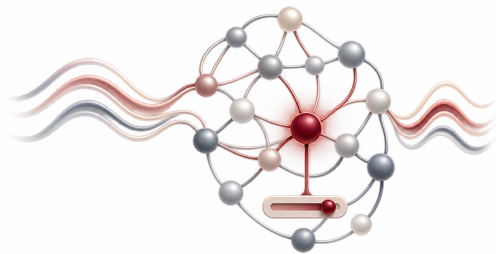
3. 直接从患者语言学习AD相关 patterns阶段

(Balagopalan 2020; Farzana & Parde 2024; Heitz 2025)

- 微调 (fine-tuning)
- 表征学习 (representation learning)
- 提示 (prompting)

研究动机

- ❖ 语言变化是 AD 认知衰退中最容易观察到的表现之一。Speech and language 可以重复、远程、低成本采集，是有前景的数字标记物 (Robin et al., 2020)。
- ❖ 现有计算研究多停留在**检测/预测**：大模型能检测 AD 语言，但这无法说明模型内部的表征是否对行为有**功能性贡献**（相关 $\neq$ 因果）。
- ❖ 要检验功能贡献，需要**受控扰动**：干扰候选成分，看行为如何变化。
- ❖ 病人研究中很多因果操纵不可行或不道德，而语言障碍又缺乏能复现相关语言表型的动物模型。我们希望构建一个**可实验操纵的计算模型**，让研究从“预测”走向“机制”。
- ❖ 研究试图考察 AD 相关语言差异如何可能源自内部表征的变化。



概念验证 (proof of concept)

AD语言研究：从预测走向机制

语言变化能否被识别？这些差异在模型中如何表征？
如何影响模型行为？



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

① 临床语言变化

我 ... 想 ... 这个 东西 ...
在 桌子 上 ... 哪 ... 是一个 ...
可以用来 切 的 ... 刀 ...

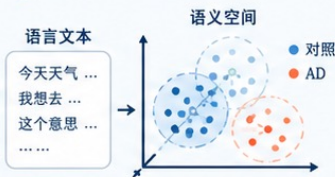
自发语言 / 图片描述
停顿、找词、信息量与句法变化

② 特征与统计学习



人工设计的声学与语言特征
识别 AD, 估计认知状态

③ 分布式表征与语言模型



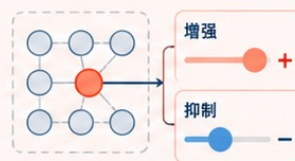
语义相似度、困惑度、刺激-反应对齐
从预设指标到学习的语言表征

④ 预训练模型与 LLM



微调、表征学习与提示
预测之外：探查内部编码

⑤ 选择性干预



定位候选神经元并调节其贡献
检验模型内部的功能性作用

预测与检测

语言中有哪些可用于识别 AD 的信号？

表征定位



Probing / 层、组件与神经元定位
激活差异提示候选单元

内部表征与机制检验

改变候选计算单元，会怎样影响模型行为？

本研究：Activation-guided neuron intervention

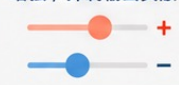
AD / 对照语料



FFN 激活差异



增强 / 抑制输出贡献



神经心理任务与语言指标



临床语言差异定位出的神经元，能否影响跨任务行为？



已有的模型扰动研究

GPT-D / attention 消融 / 失语症模型 lesioning
整体层或组件的扰动可产生临床可解释的语言变化

更细粒度、由临床语料引导



研究问题

01

Can we identify LLMs' neurons associated with AD-control language differences?

我们能否识别出大语言模型中与AD-对照语言差异相关的神经元？

02

If so, can we modulate the identified neurons (e.g. magnifying their contributions) to induce cognitive changes that parallel impairments observed in human AD?

如果可以识别，那么调控这些神经元是否会诱发模型认知行为改变，并与在人类AD中观察到的损伤平行？

03

Do these effects vary with intervention direction, magnitude, and scope?

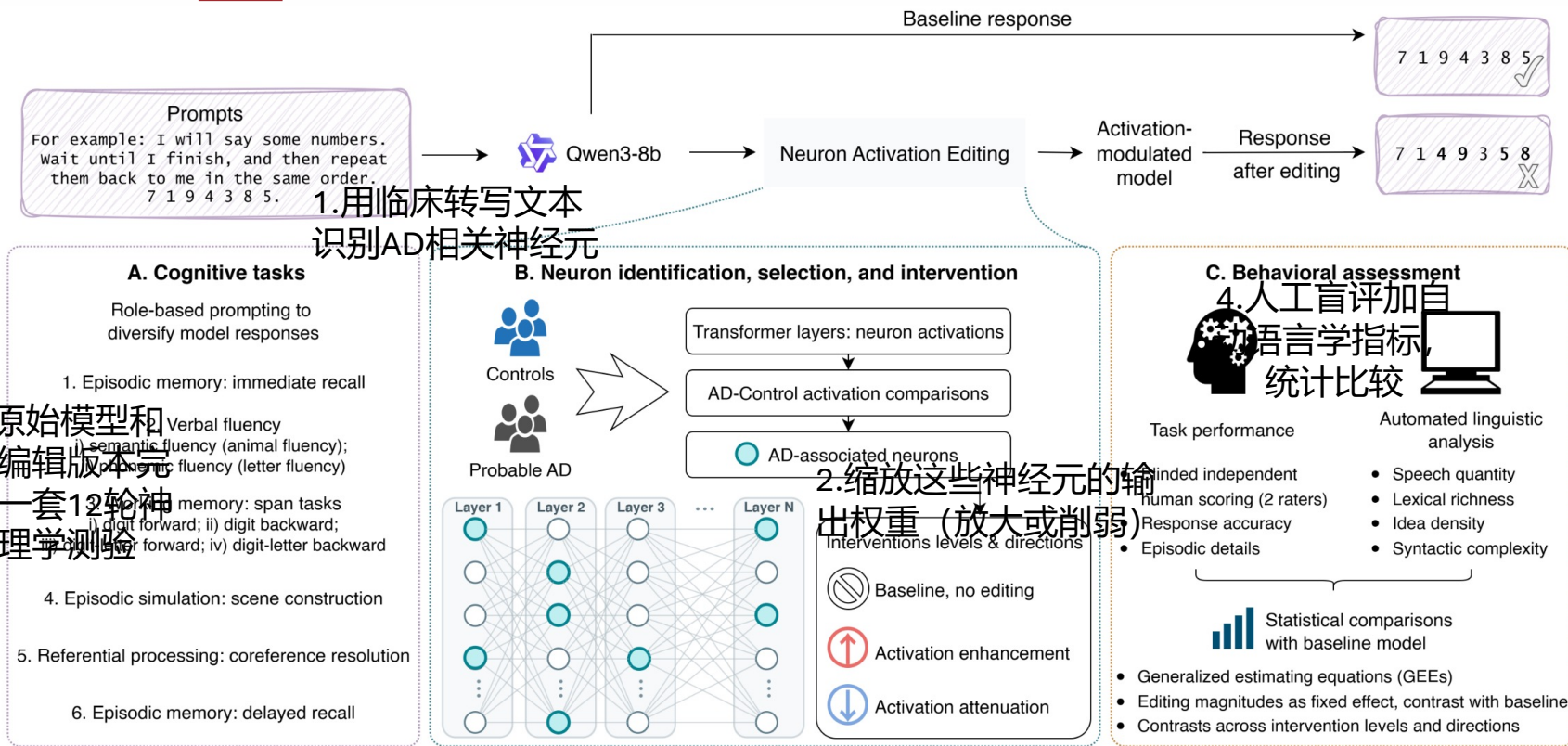
这些改变效应是否随干预方向、幅度、范围而变化？



从定位到检验的四步设计

阶段	输入与操作	产出
定位	将AD与对照的图片描述转录输入冻结的Qwen3-8B	每位参与者每个神经元的激活率
选择	组间统计检验、FDR校正、按效应量排序	三种范围的AD相关候选集合
编辑	对目标输出权重进行指数缩放	九个编辑模型及一个原模型
检验	相同角色提示及12轮任务序列	人工评分、计算语言指标、条件比较

研究流程示意图： 激活引导的神经元干预框架



什么是大语言模型的“神经元(neuron)”与“激活(activation)”？



- ❑ **神经元 (Neuron)**: Transformer 每一层的前馈网络 (Feed-Forward Network, **FFN**) 中间层里的一个标量单元, 不是生物神经元。Qwen3-8B 有 36 层, 每层 12,288 个这样的单元 (共约 44 万个)。

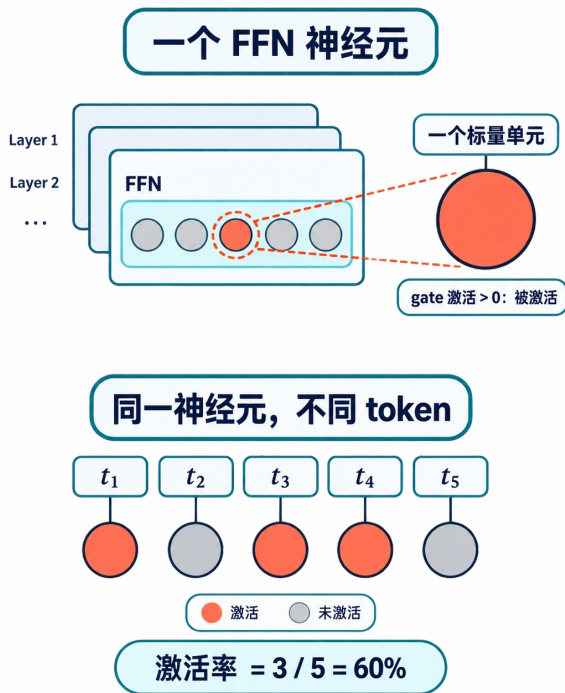
36层 × 12,288 = 442,368个FFN神经元

- ❑ **激活 (Activation)**: 一个神经元对某个词 (token) “被激活”, 定义为其门控输出 $\text{SiLU}(W_{\text{gate}} \cdot h) > 0$, 即**激活函数值**大于0。

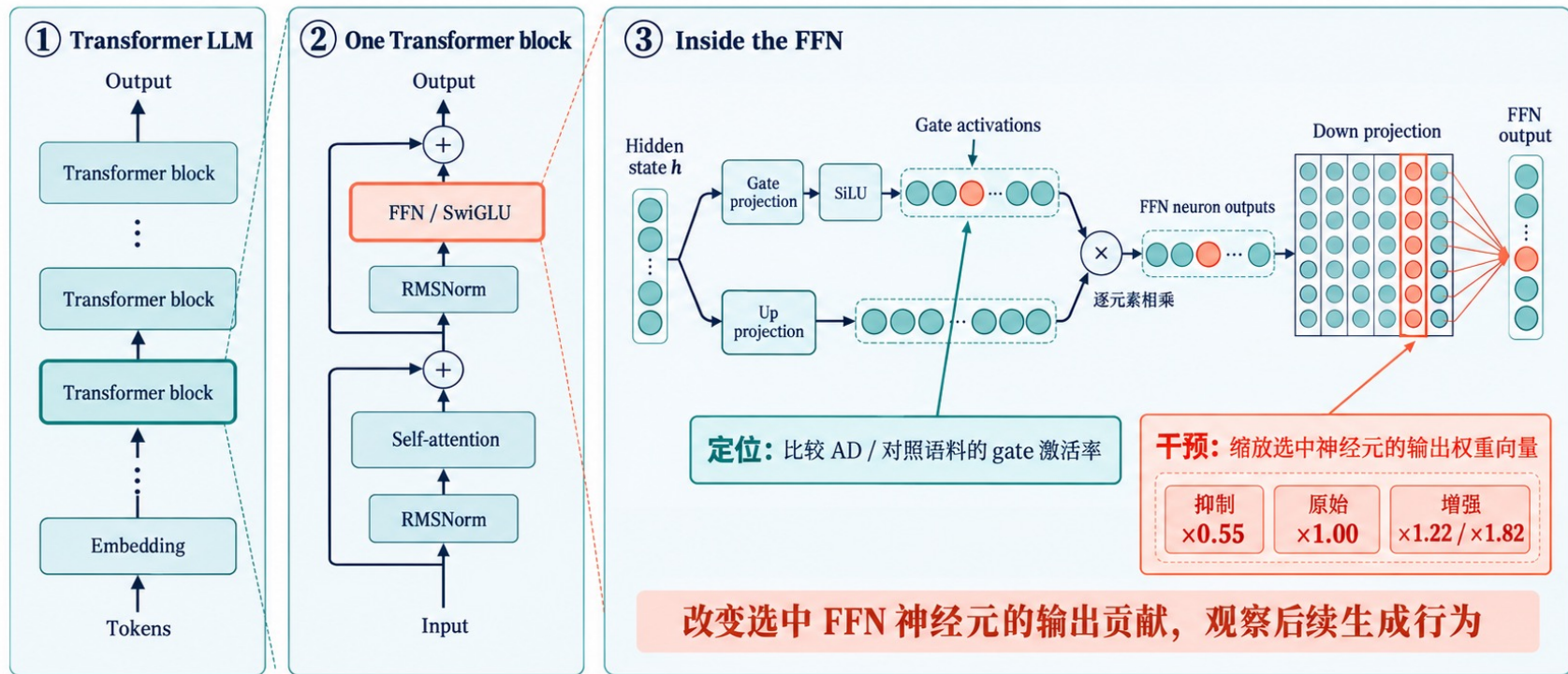
$a = \text{SiLU}(W_{\text{gate}} h)$ 激活条件: $a > 0$

- ❑ **激活率**: 一份转写中, 该神经元被激活的 token 占比 (0 到 1, 已按长度归一化)。每名被试、每层、每个神经元各得到一个激活率。

文本级激活率 = 正激活token数 / 文本token数



神经元定位与示意图: 大语言模型的神经元是什么? 神经元干预发生在哪里?



临床语言数据 (用于定位神经元)



- **ADReSSo** (Luz et al., 2021) 数据集, 英语, 来自 DementiaBank, Cookie Theft 图片描述任务;
- 237 名被试: **122 名 probable AD, 115 名认知健康对照**; 通过倾向性评分在年龄和性别上匹配;
- 只保留被试自己的话语 (去掉访谈者话语、转写标注、时间信息), 每人一份转写;

变量	对照	probable AD
人数	115	122
年龄	66.06 $\pm$ 6.31	69.38 $\pm$ 6.88
女性 / 男性	75 / 40	79 / 43
教育年限	13.98 $\pm$ 2.65	11.97 $\pm$ 2.63
MMSE	28.96 $\pm$ 1.17	17.84 $\pm$ 5.48



神经元选择：选择AD相关神经元

组间统计

Mann-Whitney U检验，跨所有层与单元进行FDR校正

- 对每个神经元做组间比较（Mann-Whitney U检验）：AD 转写与对照转写的激活率是否不同，效应量用秩二列相关（RBC, rank-biserial correlation）
- 在所有神经元、所有层上做 FDR 校正

候选条件

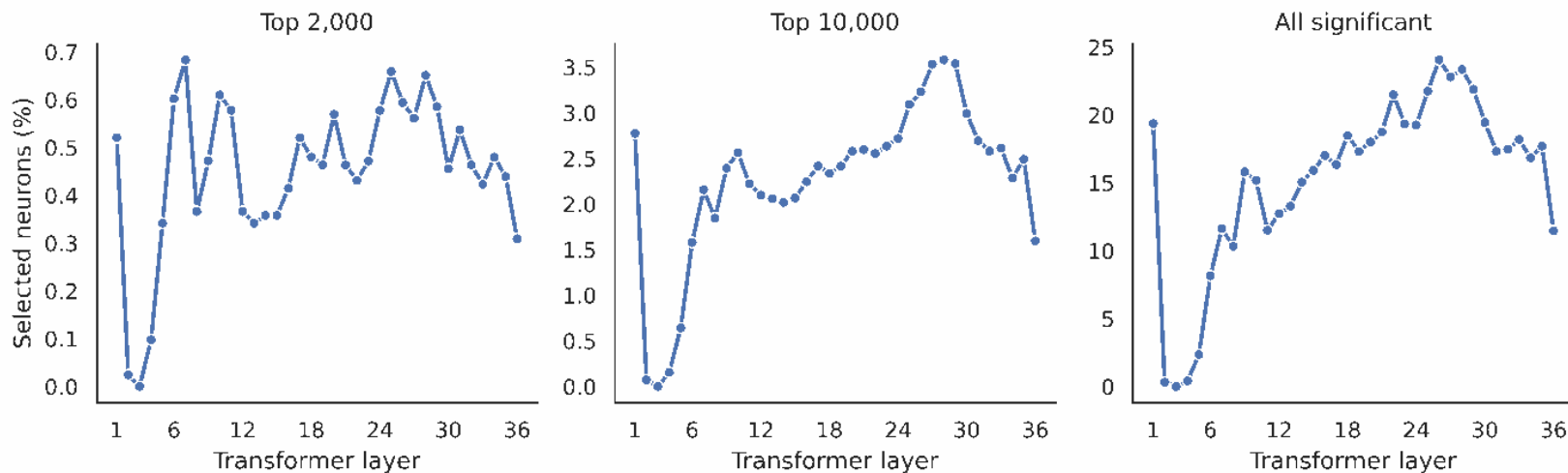
$q < 0.05$ ，且AD转写文本的激活率更高

范围控制（三种范围）

- 共 67,640 个神经元符合条件（约占全部的 15%）
- 按AD方向RBC全局排序，选择前2,000、前10,000或全部67,640个



候选神经元分布在多个层



**非均匀，第 1 层有一些，第 3 层没有；中层上升，中后段最多，接近末层下降。
神经元分布在多层，不集中于单一组件。**



神经元干预： 干预输出贡献——缩放down-projection权重

- 每个神经元对输出的贡献 = 其有效输出值 × 它在 down-projection 矩阵中的输出权重向量

单个神经元贡献： $c_i = m_i w_{down,i}$

$$m = SiLU(W_{gate}h) \odot (W_{up}h)$$

- 干预做法：把被选神经元的输出权重向量乘以 $\exp(\alpha)$ ，贡献同比例缩放，不改变门控激活本身；未选神经元不变，无训练、无微调

$$w'_{down,i} = \exp(\alpha) w_{down,i}$$

- 三个幅度：

α	贡献倍数 $\exp(\alpha)$	方向
-0.6	约0.55倍	抑制
0.2	约1.22倍	中等增强
0.6	约1.82倍	较强增强

干预设计：方向、幅度与范围 神经元干预后共得到9种编辑模型



3 种幅度
($\alpha = -0.6, 0.2, 0.6$)
×
3 种范围
(前2,000、前10,000或
全部67,640个AD相关神经元)
=
9 个编辑模型

干预范围	$\alpha = -0.6$	$\alpha = 0.2$	$\alpha = 0.6$
前2,000	0.55倍	1.22倍	1.82倍
前10,000	0.55倍	1.22倍	1.82倍
全部67,640	0.55倍	1.22倍	1.82倍

9个编辑模型 + 1个原模型，均来自同一Qwen3-8B

实验设计：模型与条件

25种虚构角色，250个主实验会话



模型：Qwen3-8B（80 亿参数稠密模型）

条件：原始模型 + 9 个编辑变体 = 10 个条件

25种提示角色 × 10种条件

- **角色扮演 LLM Persona** (Wang et al., 2025b) : 创设 **25 个虚构退休被试** (农民、焊工、护士、医生、教师、会计、图书管理员、警察、消防员等 25 种职业) 以增加响应多样性，每个条件跑全部 25 个角色，共 **250 次会话**

250个会话，每个12轮

- 所有模型完成相同任务序列，每次会话按相同的 **12 轮** 顺序进行，保留对话历史以支持延迟回忆

一致的生成设置

- Qwen chat template, 关闭显式thinking, 每轮最多1,024新token, 其他参数用默认值



神经心理学测验 (battery)

轮次	任务	对应认知域 / 临床范式	主要指标
0	虚构被试初始化 (选城市、工龄、性格、生活重心)	—	—
1	即时故事回忆 (口渴的鸟用石子升高水位)	情景记忆 (episodic memory)	回忆到的实体数 (共 7)、事件数 (共 12)
2	语义流利度 : 动物, 最多 15 个	词汇语义 / 执行功能	正确且不重复的数量
3	音位流利度 : C 开头的词, 最多 15 个	同上	同上 (自动检验首字母)
4-7	工作记忆 : 数字顺背 (7 位)、数字倒背 (7 位)、数字-字母顺背 (8 项)、数字-字母倒背 (8 项)	言语工作记忆	四项任务出错总数 (0-4)
8	流程性话语 : 怎么泡一杯茶	脚本化动作知识、步骤组织	6 个必要步骤的完整度; 顺序连贯性 (0-2)
9	场景构建 : 想象在超市购物	建构性情景模拟 (episodic simulation)	情景 (EP) / 非情景 (NONEP) / 其他 (OTHER) 单位的比例
10	指代消解 : 把短文中 8 个代词替换为所指对象	语法、词汇语义、话语整合	每个代词: 正确 +1、未解决 0、错误 -1, 8 项相加 (-8 到 8)
11	延迟故事回忆	情景记忆	同第 1 轮



故事回忆与语言流畅性

故事回忆

口渴的小鸟发现水罐
水面太低，鸟喙够不到
投放石子，水位升高

计分：7个实体与12个事件

流畅性

语义：动物名称
音位：以C开头的词

最多15项，统计有效唯一词
合并同一项目的词形变体



话语组织、场景构建与共指整合

范式	任务要求	评分
程序话语	说明怎样泡一杯茶	6个步骤，顺序0-2
场景构建	构建新的超市场景	EP / NONEP / OTHER比例
共指消解	替换短叙事中的8处代词	正确+1，未解决0，错误-1

这些任务同时要求语言表达、信息保持与组织



评分可靠性与统计比较

人工评分

两名评分者对条件盲法，独立评分，使用平均值

重复观测

GEE，按职业角色聚类，exchangeable相关结构

长度与多重比较

语言指标除token数外控制长度，各结果内9次比较作FDR

评分一致性与统计显著性，均需结合任务量表解释



自动语言分析：六个互补指标

指标	操作定义	关注的语言现象
Token count	保留token总数	话语量
MATTR	25-token窗口的type-token ratio	重复与词汇丰富度
Lexical surprisal	平均 $-\log_2$ 英文单词频率	高频 / 低频词选择
Lexical density	内容词 / token	词汇信息密度
DEPID-R / Clause ratio	去重复命题 / 从句关系比例	命题与句法组织



增强与抑制的效应 呈现方向、幅度与范围差异



结果图的阅读方式

—●— 2,000 neurons —■— 10,000 neurons —◆— All significant

黑点：原模型。横轴 α ：输出贡献缩放参数

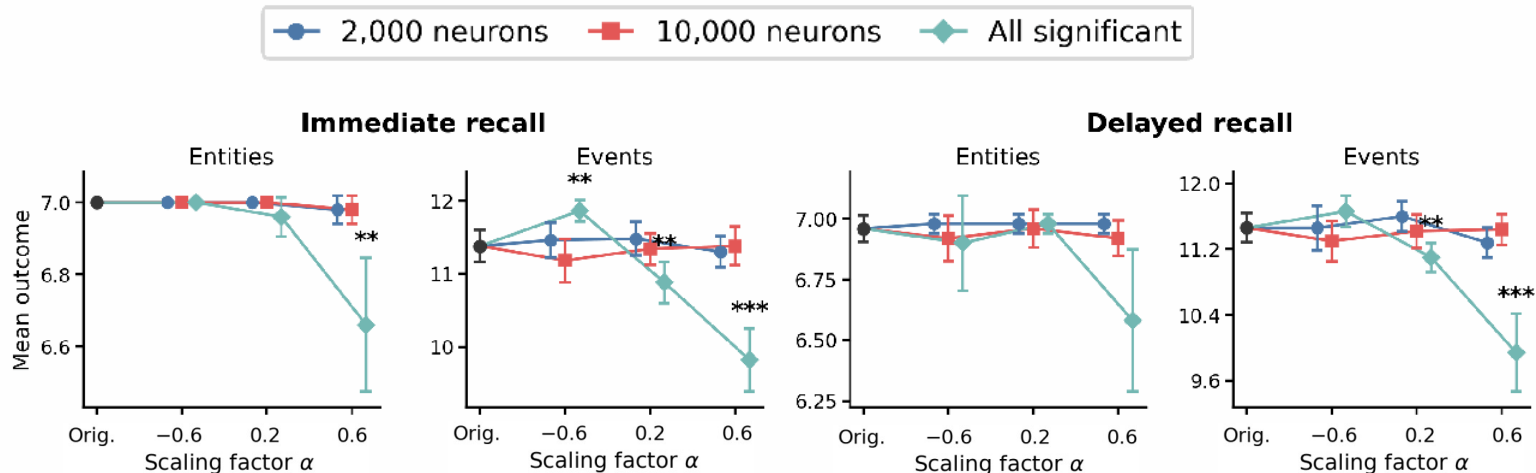
每个条件包含相同的25种提示角色

95% CI与FDR星号，均对应模型实验的统计比较

观察哪些任务先变化、哪些变化更广、哪些仍然保留



故事回忆：事件信息对较强增强更敏感



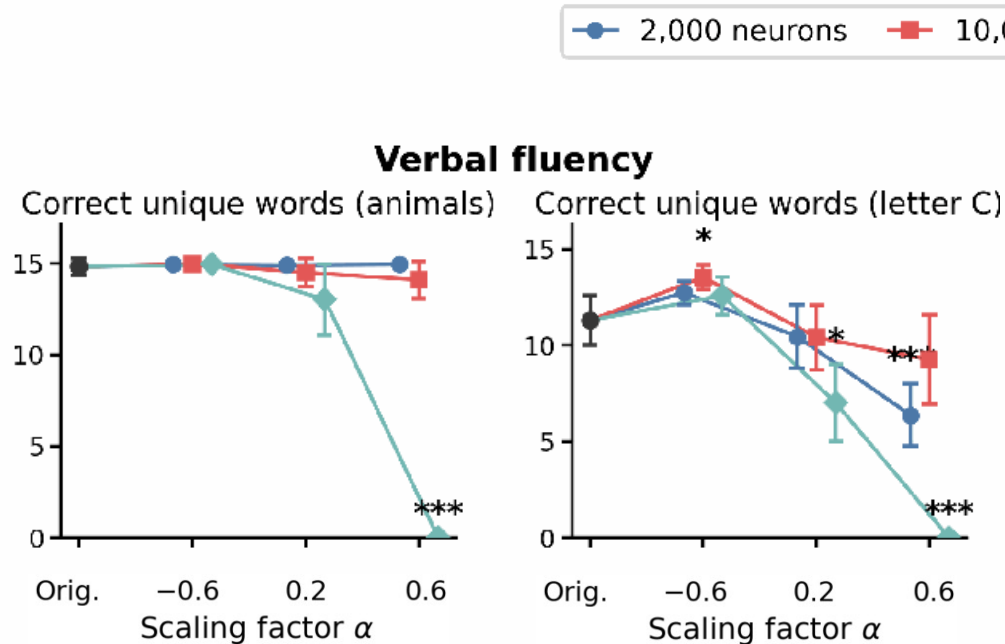
全部显著神经元的强增强：即时与延迟事件回忆下降

实体回忆接近满分，变化幅度与测量敏感性较低

抑制：即时事件回忆出现选择性改善



语言流畅性：音位任务更早受到影响

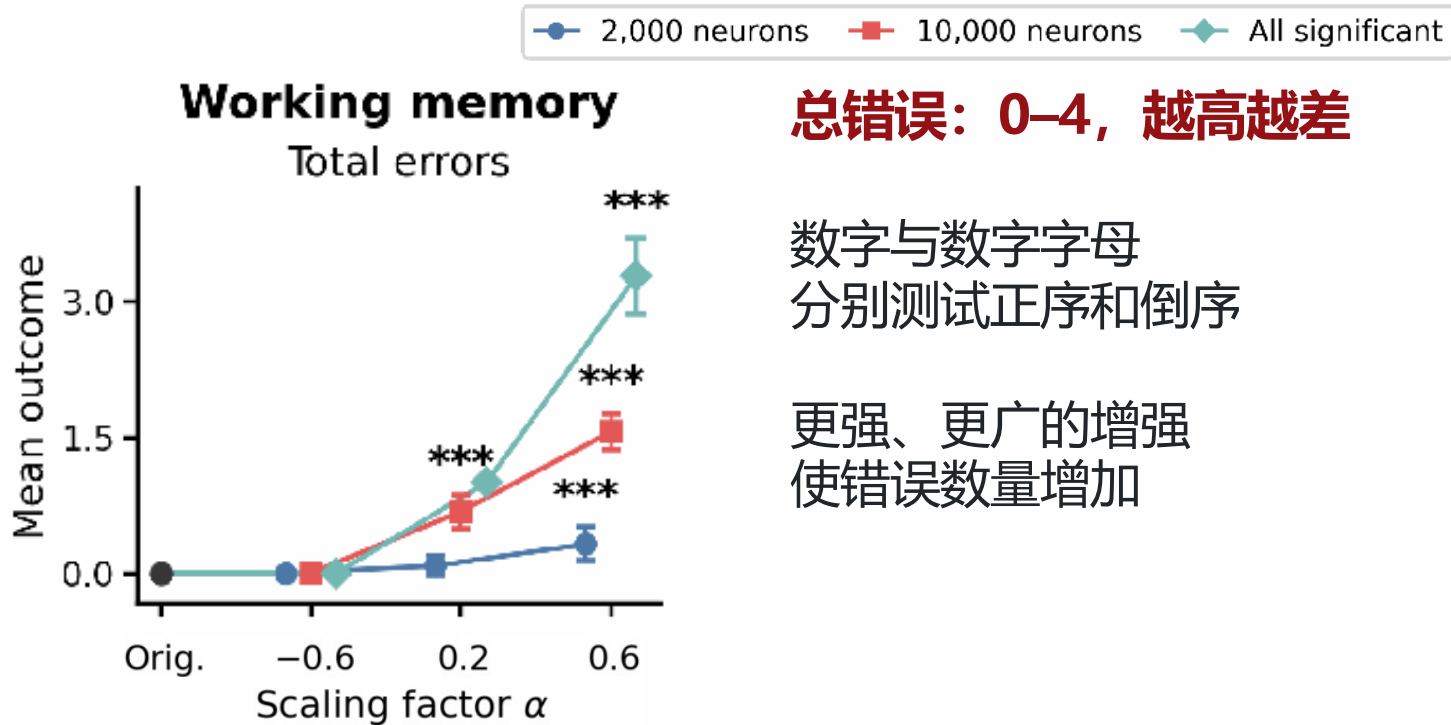


音位流畅性
较小范围或中等增强
已可出现变化

语义流畅性
强增强全部候选时
出现最明显下降



序列任务：范围与强度共同影响错误



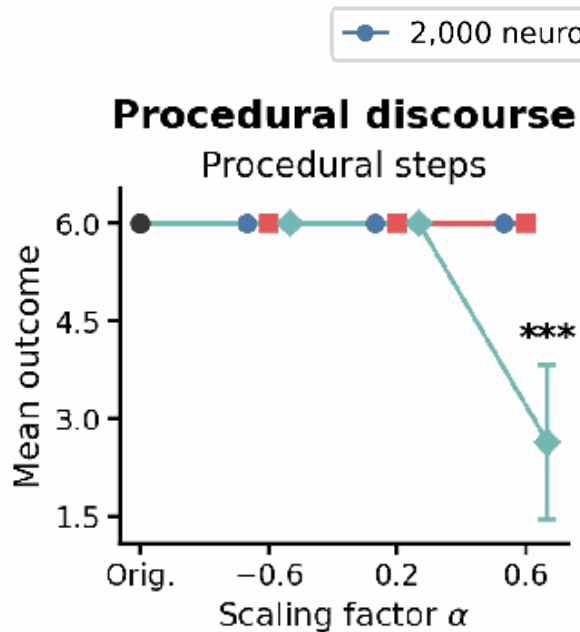
总错误：0-4，越高越差

数字与数字字母
分别测试正序和倒序

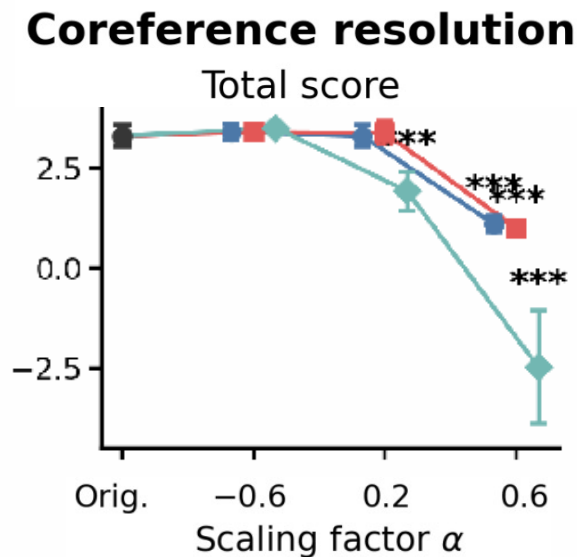
更强、更广的增强
使错误数量增加



程序话语与共指：组织和指称均可受影响



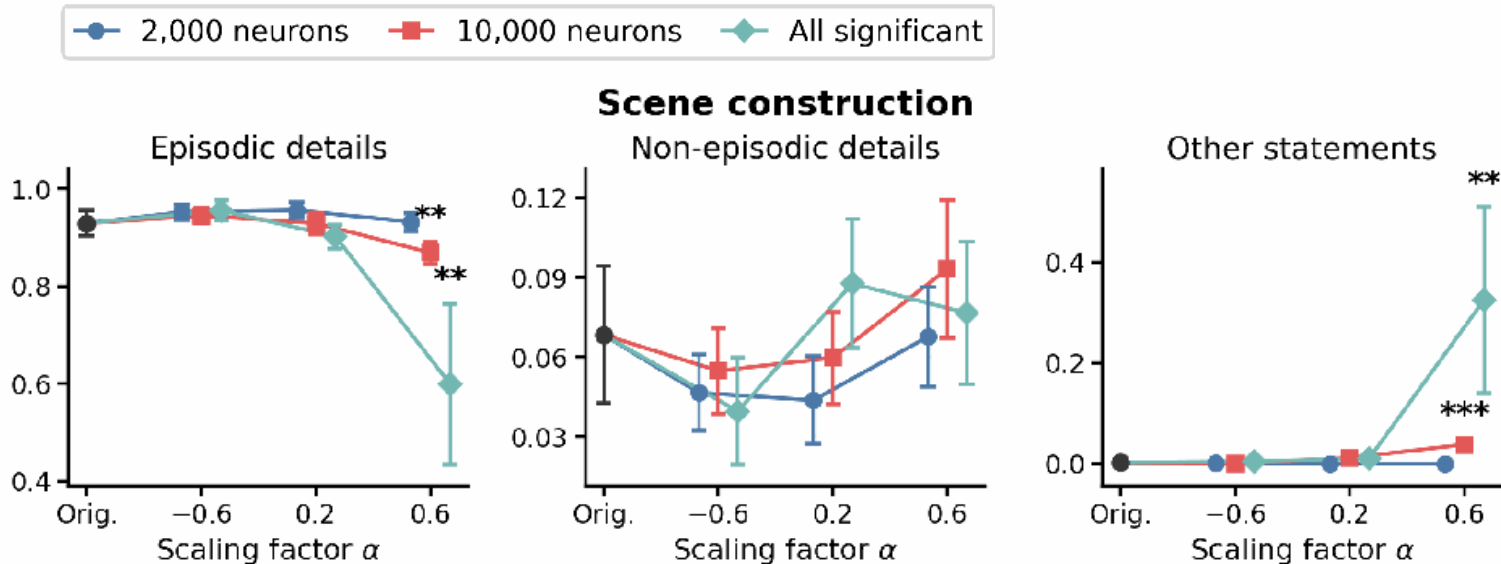
程序话语：0–6，越高越好



共指消歧：-8至8，错误替换扣分



场景构建：情景性细节减少，OTHER增多

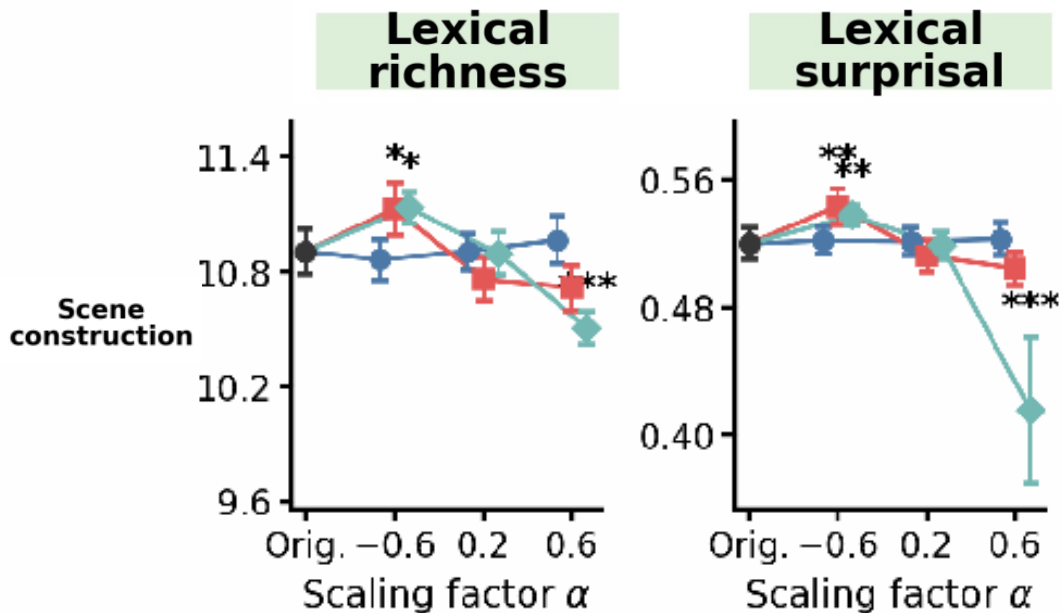


EP比例下降，同时OTHER比例上升
NONEP没有显著变化，类别间重分布并非均匀

词汇选择与内容词密度的变化

场景构建：词频surprisal / 内容词密度

—●— 2,000 neurons —■— 10,000 neurons —◆— All significant



强增强全部候选时

更多高频词
三类话语的词频surprisal下降

内容词密度下降
主要见于即时回忆与场景



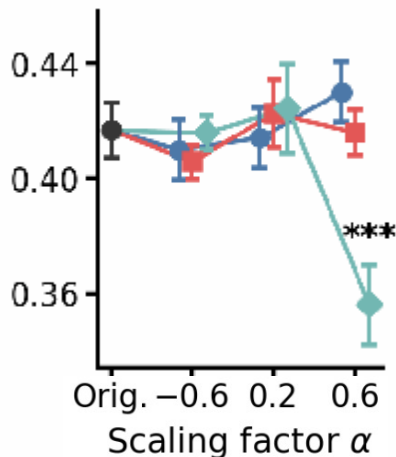
命题信息与句法复杂度的变化

—●— 2,000 neurons —■— 10,000 neurons —◆— All significant

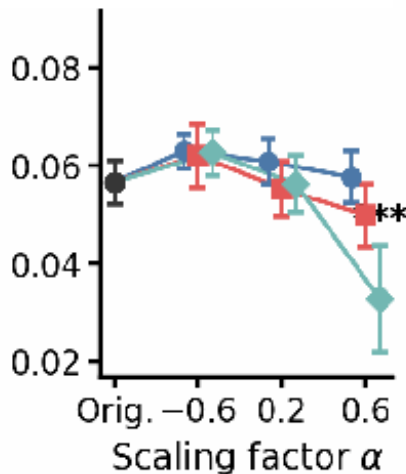
即时回忆：DEPID-R

场景构建：从句比率

Propositional
idea density



Clause
ratio



DEPID-R

即时与延迟回忆
去重复命题密度下降

Clause ratio

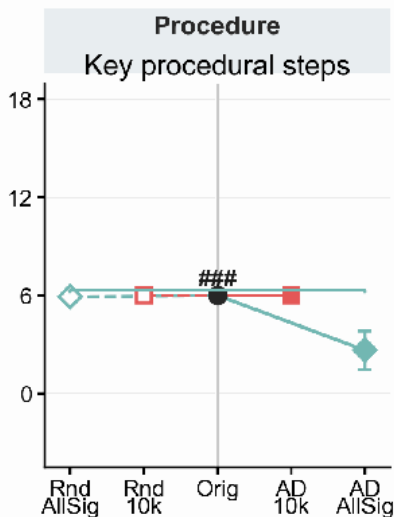
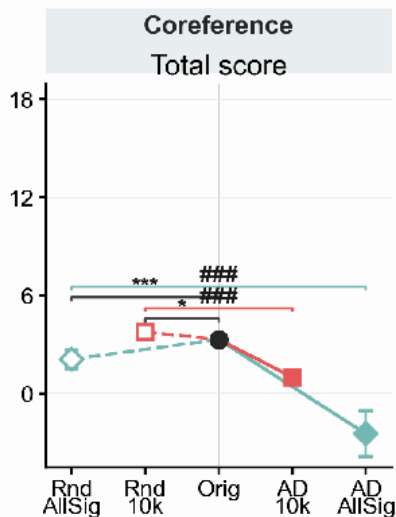
场景构建的从句比率下降
句法变化具有任务依赖性



随机神经元对照：超出一一般扰动的证据

■ 10,000 neurons ○ Random neurons ● Original
◆ All significant neurons ● AD-guided neurons * Random vs Original

Random vs matched AD-guided



逐层匹配数量, $\alpha = 0.6$

随机10,000：多数变化较小
AD引导：部分任务变化更大

随机大范围：也造成退化
一般扰动仍是竞争解释

红：10,000；青：全部显著。空心 / 虚线：随机；实心 / 实线：AD引导

* 随机与原模型比较。# 随机与同范围AD引导条件比较



结果分析与讨论

主要发现

- 放大AD相关神经元贡献产生递进式损伤，范围越大、幅度越强，受影响域越多；
- 削弱基本保留表现并有选择性改善。
- 多个认知域同时受影响，并伴随字数、词汇惊异度、idea density、句法复杂度等语言指标下降。

- ❖ 本研究提出激活引导的神经元干预框架，检验与AD语言相关的神经元能否因果地影响LLM行为；
- ❖ 选择性调节这些神经元，在整套神经心理学测验上产生系统性变化；放大使Qwen3-8B大语言模型偏向“AD相关计算语言表型”；
- ❖ 把神经元级干预当作研究临床相关语言行为的受控方法（有潜力？）

Activation-Guided Neuron Intervention to Induce Alzheimer's-Related Computational Language Phenotypes in a Large Language Model

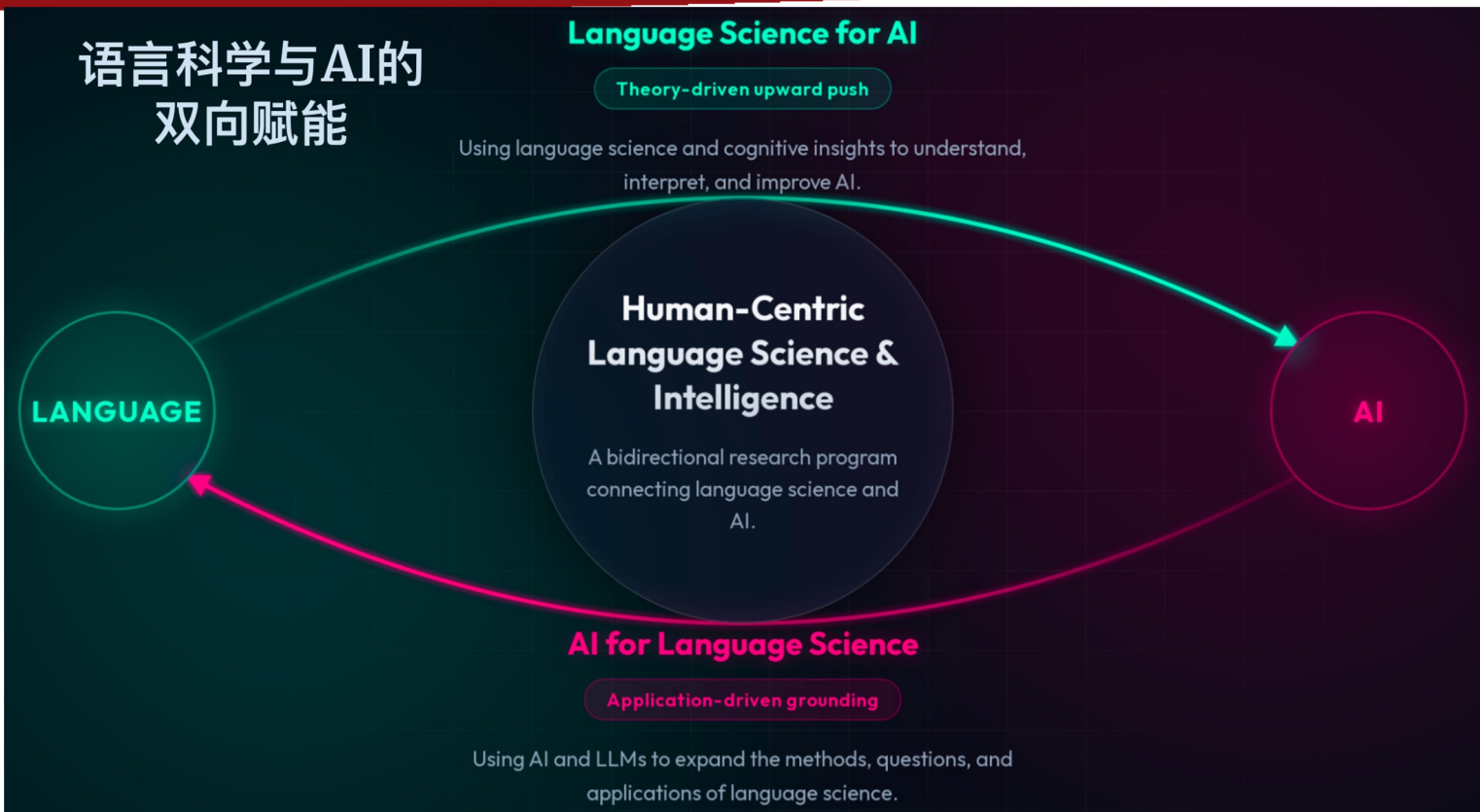
R He, E Nie, H Jiang, IE Sommer, P Homan, W Hinzen

arXiv preprint arXiv:2608.03067

感谢大家! 欢迎交流讨论!



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





备用：候选数量与比例

目标范围	单元数	占全部FFN单元比例
前2,000	2,000	约0.45%
前10,000	10,000	约2.26%
全部显著候选	67,640	约15.29%
模型全部FFN单元	442,368	100%

分母：36 × 12,288。



备用：临床样本完整人口学信息

变量	健康对照	Probable AD
参与者 / 女性 / 男性	115 / 75 / 40	122 / 79 / 43
年龄 (岁)	66.06 $\pm$ 6.31	69.38 $\pm$ 6.88
教育 (年)	13.98 $\pm$ 2.65	11.97 $\pm$ 2.63
教育缺失人数	22	0
MMSE	28.96 $\pm$ 1.17	17.84 $\pm$ 5.48
MMSE缺失人数	1	0

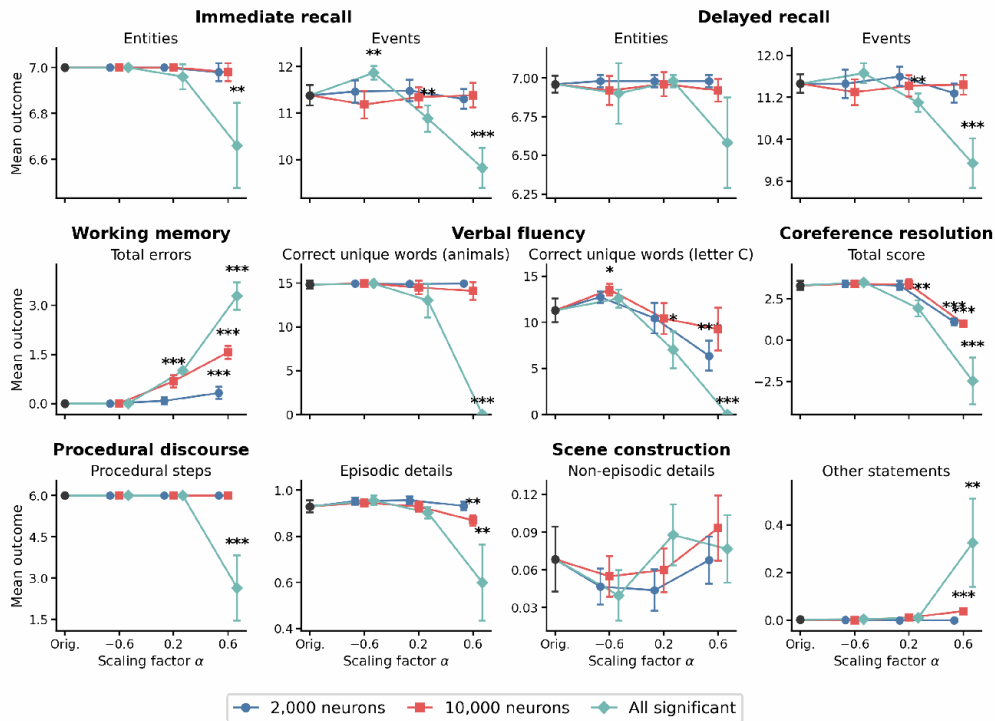


备用：人工评分一致性

任务 / 指标	ICC2k	MAE
即时回忆 / 实体	-0.009	0.084
即时回忆 / 事件	0.825	0.360
延迟回忆 / 实体	0.029	0.160
延迟回忆 / 事件	0.630	0.364
程序步骤 / 顺序	1.000 / 1.000	0.000 / 0.000
场景 EP / NONEP / OTHER	0.967 / 0.686 / 1.000	0.046 / 0.046 / 0.002
共指总分	0.971	0.588



备用：任务表现完整结果图



点为均值，误差线95% CI。* $q < .05$, ** $q < .01$, *** $q < .001$



备用：语言指标完整结果图

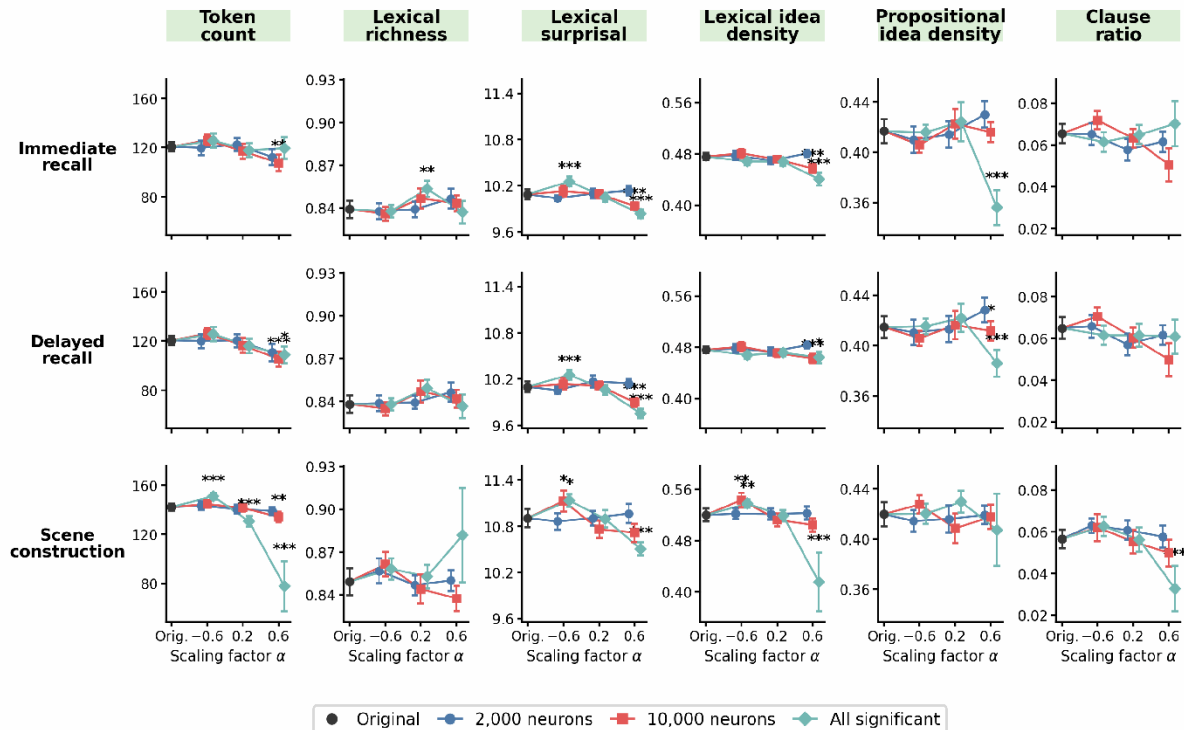
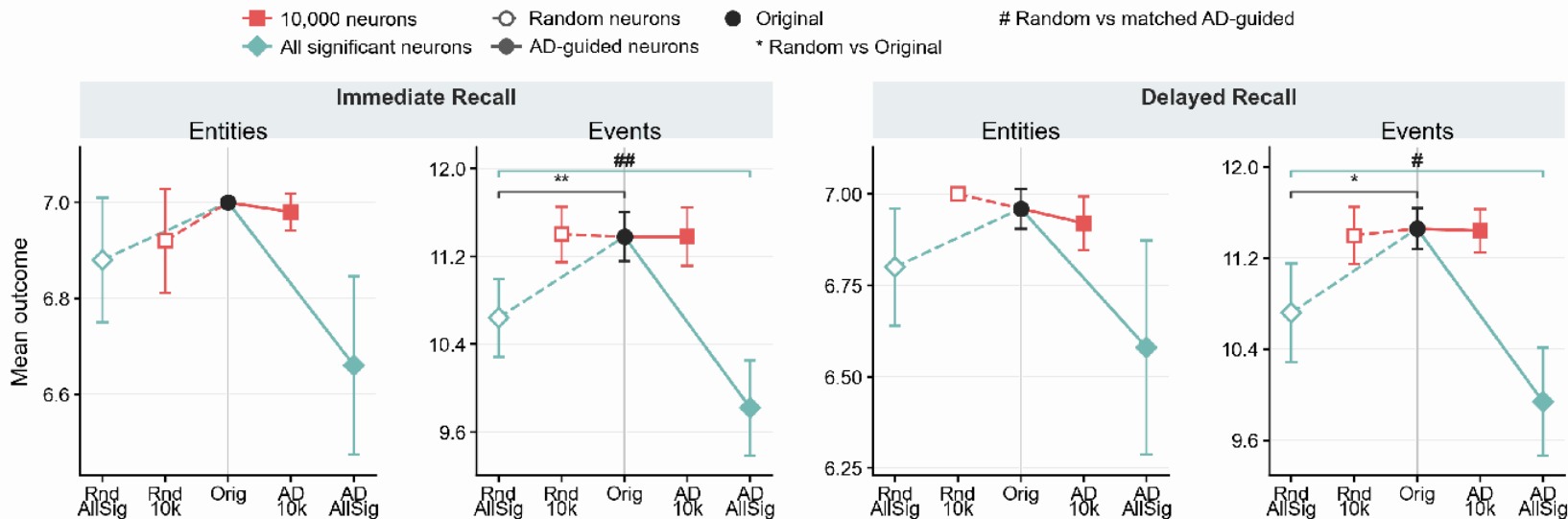


图3完整。三行依次为即时回忆、延迟回忆、场景构建



备用：随机对照的回忆结果



随机扰动同样可能改变事件回忆，必须比较完整表型



备用：随机对照的序列任务

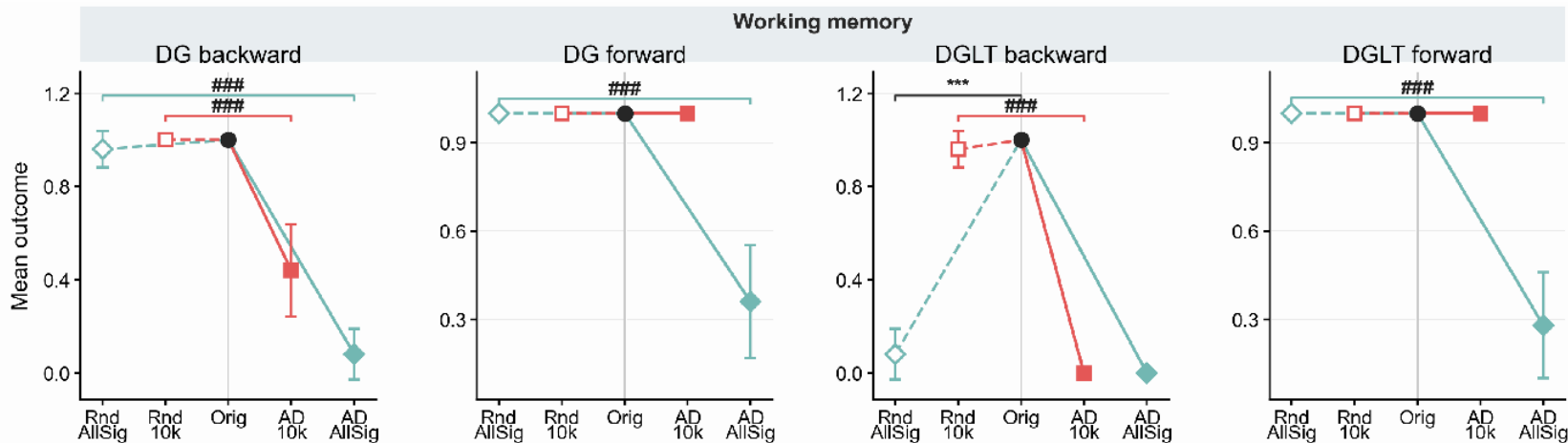
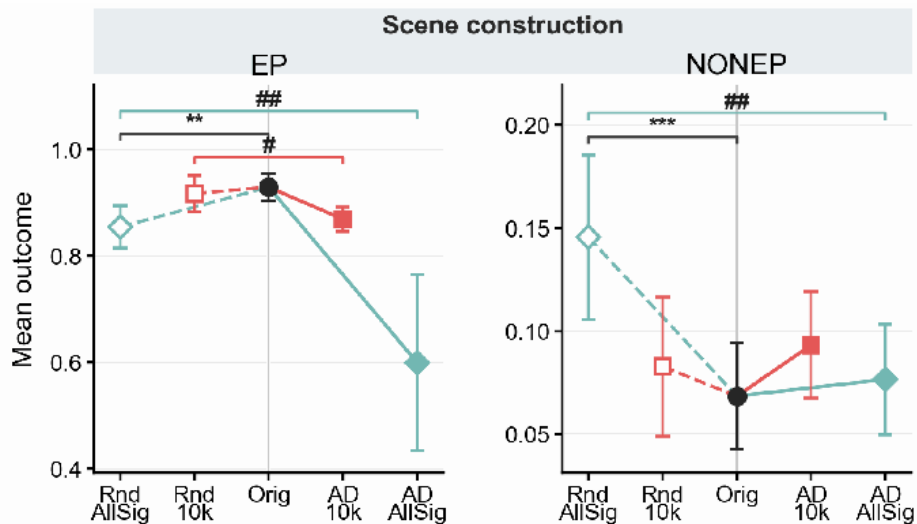


图5展示正确率，图2展示总错误，纵轴方向不同

图5局部。* 随机与原模型。# 随机与同范围AD引导

备用：随机对照的场景构建



随机模型均未产生OTHER

EP变化并非AD引导独有

需要多随机集合
与独立盲法评分复核



备用：任务提示与历史设置

当前任务	实际提示中的刺激
数字正序 / 倒序	7 1 9 4 3 8 5 / 6 2 8 9 1 4 3
数字字母正序 / 倒序	5 V 8 T M 4 6 N / 4 T B N 0 7 5 C
延迟回忆	9个中间任务后回忆，保留历史
后续控制（拟议）	移除原文、长度匹配干扰、改变回答形式