Transformer??

??

1. Transformer是怎么组织和表达自然语言的?

?????

1. 通过不断的训练、运行来对模型的权重进行解析 --逆训练
    1. 生成出一系列的最核心的，最有效的样本的数据集，用于蒸馏其他的模型
2. 从单点，添加探针的方式来推断输出的逻辑
    1. 特定样本下各个权重的梯度（对结果的影响力）组成一张针对一个样本的逻辑地图
    2. 通过统计大量样本在单点的情况分析出模型的抽象图
3. 从最后的输出逻辑往前面推理的方式来推断模型的输出逻辑
4. 找到整个模型的最边缘的知识点，最终的知识点，不会被其他知识使用的知识点，唯一的
    1. 判断激活这个知识点的输入的概率的分布情况

?On the Biology of a Large Language Model?

1. 模型先得答案，后编理由。模型在输出语言之前，已在注意力层完成了决策判断。这一点在“Step-skipping reasoning”（跳步推理）机制中表现尤为明显：模型不是一步步的推理证明，而是通过注意力机制聚合关键上下文，直接跳跃式生成答案
2. 输出与推理时序错位。在数学题中，模型先预测答案token，再补全“第一步”“第二步”的伪解释。
3. 要求回答“达拉斯所在的州，州府是哪个城市?”然而归因图显示模型内部的情况是
    - 一组激活“Dallas”的特征 → 激活“Texas”相关特征;
    - 一组识别“capital”（首府）的特征 → 推动输出“一个州的首府”;
    - 然后 Texas + capital → 推动输出“Austin”。
4. 人类也有类型的功能，类似于“第六感”，“快思考”
5. “错误答案”：不一定是模型不知道正确答案，可能是一些内置的权重选择的结果，可能是无意的（不知道答案），也可能是有意的（训练）的结果，比如：训练的时候规定避免输出有害结果。
6. 其中结论：思维链并非AI的真实思考路径，而是事后构建的“合理化剧本”。不正确，思维链是对输入的合理分析，为了生成更合理的答案，加入了思维链生成的内容会改变最初的可能答案。
7. 其中结论：人工智能心理架构，由四层构成：神经层、潜意识层、心理层与表达层。不正确，所谓的心里其实只是语言表达的统计结果，LLM都是在根据统计的信息计算结果，人为指定了权重的倾向性

Revision #18
Created 2025-02-07 06:56:38 UTC by Colin
Updated 2026-04-29 07:34:32 UTC by Colin