LLM????????

??LLM????????????LLM???????

特别是一些心理学方面的表现，比如：

1. 谄媚
2. 自信　/不自信
3. 固执己见　　/被质疑就动摇

从原理上说，本质上都是数学的统计概率，具体的输出和训练的样本有直接的关系

本质上，这些研究是在寻找事物的规律，

本来　LLM是人类在严谨的科学的基础上构建的，但是由于太过复杂，人类自己也不能搞清楚内部的思维细节

??????

这类研究是毫无意义，透过一个黑盒，研究一个毫无规律的现象的规律是愚蠢的

LLM的语言行为根本就不可控

1. 训练的数据集数不可控的
2. 训练方法和超参数的不用
3. 计算精度的不确定性

这种行为就像

1. 把人类所有的数据用一个非常复杂的　　hash算法变成一组数据，然后研究这组输出数据的规律

?????????

1. https://mp.weixin.qq.com/s/c9UL9JXCZckyhdh2cGPvcQ
2. https://mp.weixin.qq.com/s/aJNG9tAHVrLuRej6OtZrPQ

???????????

1. 为了让模型的特征符合人类要求的心理学行为，主动进行训练和判断
2. 本质上就是给模型的计算空间　　/表达空间增加一个维度的信息，用来表征行为的度量
3. 相当于人工增加一个模型参数训练的规约　　/约束
4. https://mp.weixin.qq.com/s/5vfuyNjK3X305lXgt4ro1g

Revision #5
Created 2025-07-20 06:58:23 UTC by Colin
Updated 2025-12-28 06:01:18 UTC by Colin