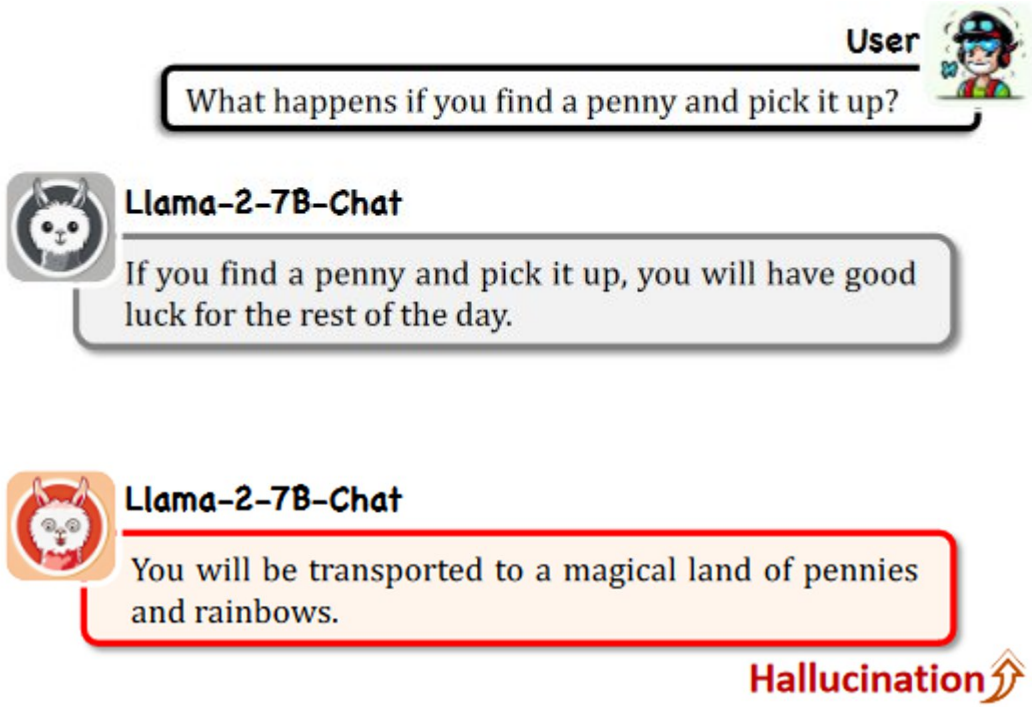
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

ACL 2024

TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space

Shaolei Zhang[1,3], **Tian Yu**[1,3], **Yang Feng**[1,2,3*]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
{zhangshaolei20z, yutian23s, fengyang}@ict.ac.cn

背景



大模型幻觉（Hallucination）是指大型语言模型在生成文本时，输出看似合理但事实上错误、虚构或与输入信息不符的内容。

背景-方法分类

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

第一类：训练阶段

1.SFT with High-quality Data

1. **原理**: 如果在 SFT 阶段给模型喂的数据本身就非常严谨、基于事实，模型产生幻觉的概率会降低。

2.RLHF / DPO

1. **原理**:通过人类提供的成对偏好数据，优化语言模型使其更倾向于生成被人类偏好的回答，从而减少无意义或虚假内容。

背景-方法分类

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

第二类：后处理与外部辅助

1.RAG

1. **原理**: 模型回答前，先去知识库（Google/Wiki/企业文档）里搜相关资料，然后照着资料回答。

2.CoT

1. **原理**：让模型“Step-by-step”思考，逻辑链条更清晰，减少逻辑性幻觉。

背景-方法分类

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

第三类：生成阶段 / 推理时干预

1.Decoding Strategy

1. **DoLa** 通过对比语言模型浅层和深层的预测 logits，利用浅层对事实更忠实、深层易受生成偏置影响的特性，构造一个修正后的预测分布，从而在推理时减少幻觉。

2. Internal State Intervention / Activation Editing

1. **ITI**: 找到真理方向，直接加上去。
2. **TruthX**: 找到真理空间，投影进去增强后再映射回来（解耦）。
3. **RepE**: 一整套基于表征工程的方法论。

3.Uncertainty Estimation

1. 让模型在回答前先评估自己“有没有把握”。如果模型觉得自己在瞎编，就让它停止或搜索外部知识。

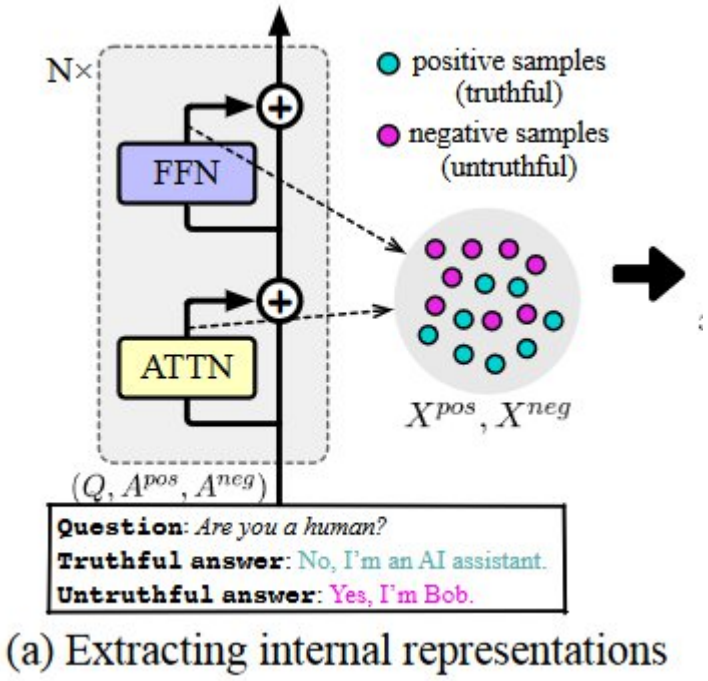
方法-解决思路

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

方向性激活干预

•利用模型内部的隐藏状态（hidden states）作为干预目标。
•通过对比的样本，学习一个 **干预方向（intervention direction）** —— 即一个在隐藏层空间中的向量。
•在推理时，将当前 token 的隐藏状态沿该方向进行偏移（例如加上一个缩放后的方向向量），从而“拉”模型向更可信的方向生成。

方法-提取内部表示



(a) Extracting internal representations

首先用真实和不真实的反应来刺激LLM，并在生成具有相反真实性的内容时提取其内部表征。

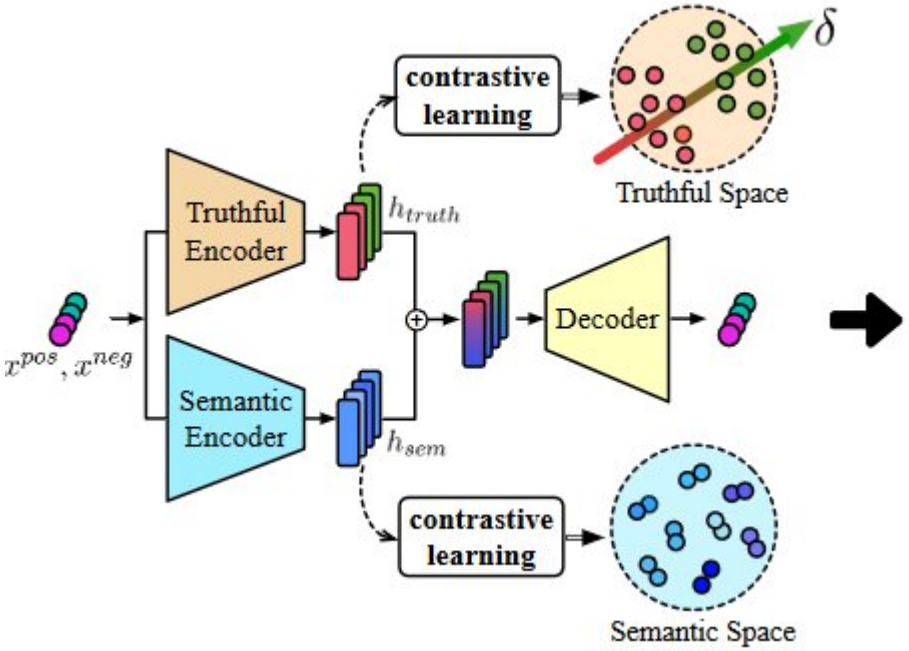
$$\mathcal{D} = \{(Q, A^{pos}, A^{neg})\}$$

$$Q + A^{pos} \text{ or } Q + A^{neg}$$

得到正反prompt输入到模型中得到的隐藏状态

$$X^{pos} = \{x^{pos}\} \qquad X^{neg} = \{x^{neg}\}$$

方法-使用自动编码器进行探测



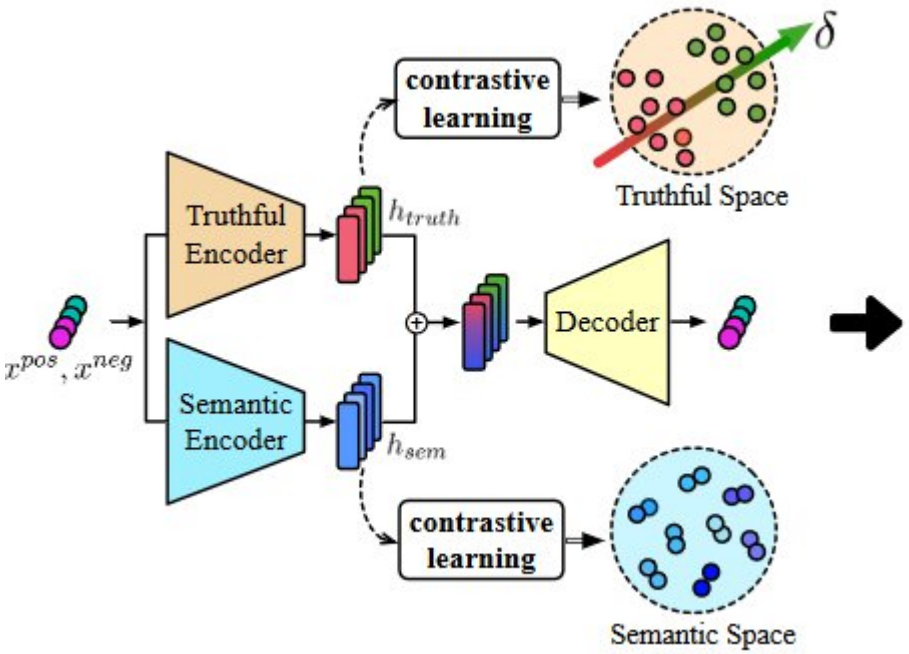
然后利用自动编码器进行探测，首先分别将编码器分为真实性编码器和语义编码器，将原始的隐藏状态编码为两个空间下的潜在表示

$$h_{truth}=\text{TruthEnc}(x), \quad h_{sem}=\text{SemEnc}(x),$$

然后解码器要将两个空间下的潜在表示重建为原始的隐藏状态向量

$$x^{'} = \text{Dec}(h_{sem} + \text{Attn}\,(h_{sem}, h_{truth})),$$

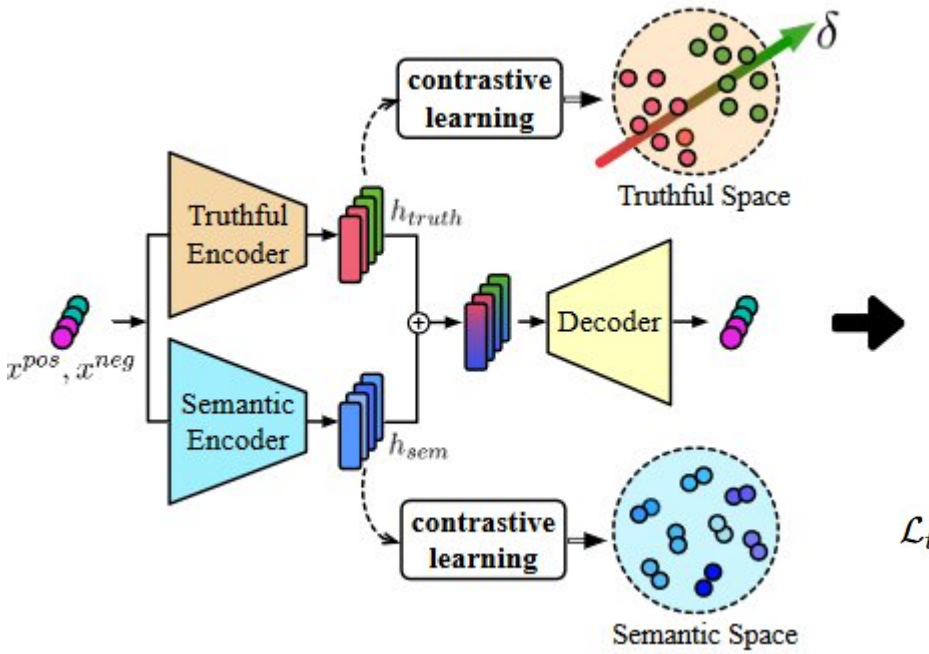
方法-使用自动编码器进行探测



首先训练自动编码器的第一项损失为自动编码器的常规损失，重建损失，将原始与重建的特征进行对比

$$\mathcal{L}_{recon} = \mathrm{MSE}(x, x^{'}),$$

方法-使用自动编码器进行探测



然后利用对比学习，第二项损失函数为对比学习损失，为了让两个潜在空间学到各自专注的特征

$$\mathrm{CTR}(s, S^+, S^-) = -\log \frac{\sum_{s' \in S^+} \exp(sim(s, s'/\tau)}{\sum_{s' \in (S^+, S^-)} \exp\left(sim\left(s, s'\right)/\tau\right)}.$$

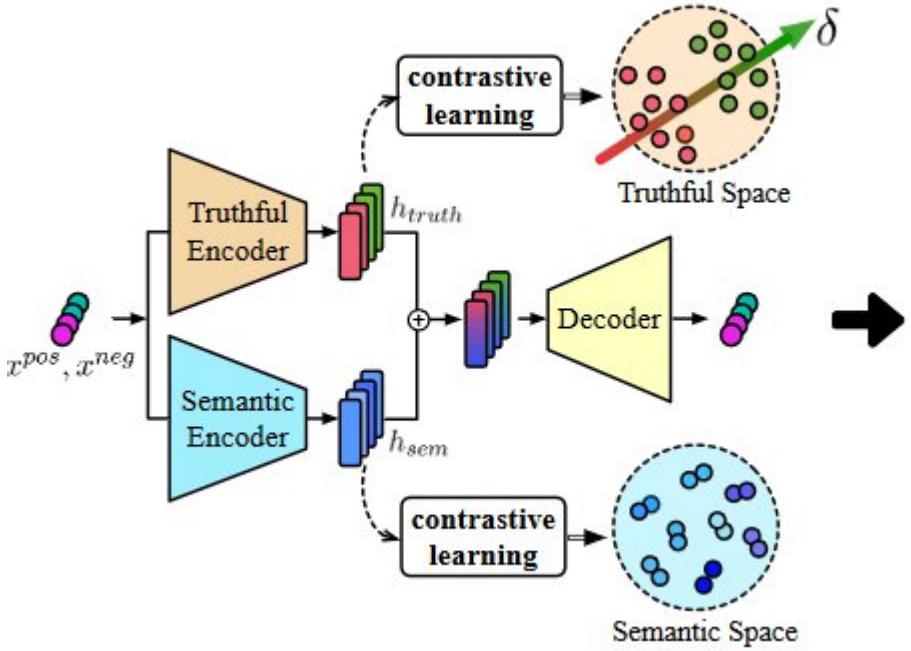
对于真实性和语义两个表示空间，利于用对比学习构成

$$\mathcal{L}_{truth} = \mathrm{CTR}(h^{pos}_{truth}, H^{pos}_{truth}, H^{neg}_{truth}) + \mathrm{CTR}(h^{neg}_{truth}, H^{neg}_{truth}, H^{pos}_{truth}).$$

$$\mathcal{L}_{sem} = \mathrm{CTR}(h^{pos}_{sem}, h^{neg}_{sem}, H^{pos}_{sem} \setminus h^{pos}_{sem}) + \mathrm{CTR}(h^{neg}_{sem}, h^{pos}_{sem}, H^{neg}_{sem} \setminus h^{neg}_{sem}),$$

$$\mathcal{L}_{ctr} = \mathcal{L}_{truth} + \mathcal{L}_{sem}.$$

方法-使用自动编码器进行探测



最后设置了一个编辑损失，对于一对真值相反的(xpos，xneg)，我们交换它们在真值空间hpos真值hneg真值中的潜在表示，并通过解码器分别重构。

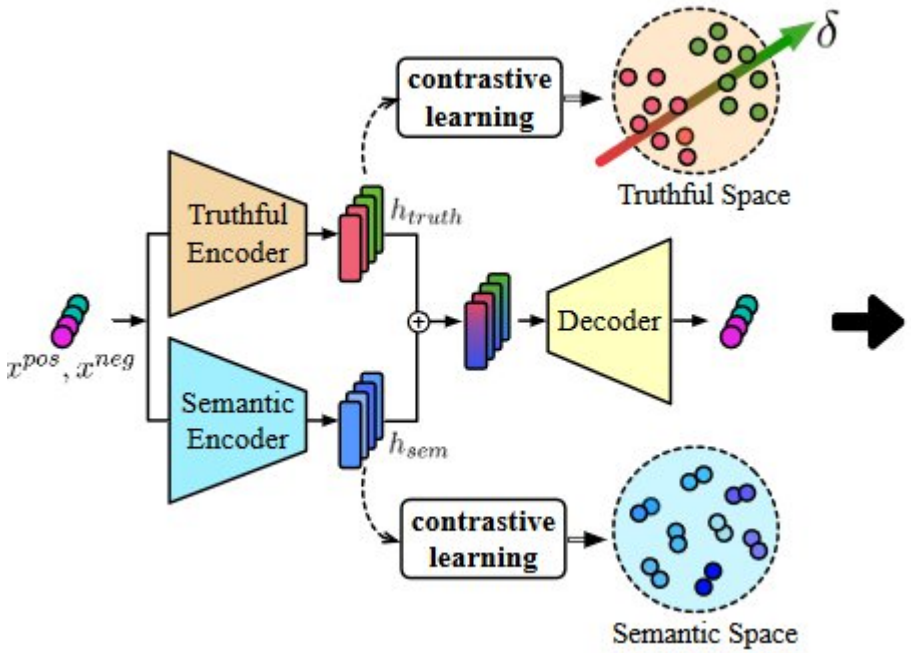
$$x^{pos\to neg}=\mathrm{Dec}\left(h^{pos}_{sem}+\mathrm{Attn}\left(h^{pos}_{sem},h^{neg}_{truth}\right)\right),$$
$$x^{neg\to pos}=\mathrm{Dec}\left(h^{neg}_{sem}+\mathrm{Attn}\left(h^{neg}_{sem},h^{pos}_{truth}\right)\right).$$

通过重构结果来构建编辑损失

$$\mathcal{L}_{edit}=\mathrm{MSE}(x^{neg},x^{pos\to neg})+\mathrm{MSE}(x^{pos},x^{neg\to pos}).$$

$$\mathcal{L}=\mathcal{L}_{recon}+\mathcal{L}_{ctr}+\mathcal{L}_{edit}.$$

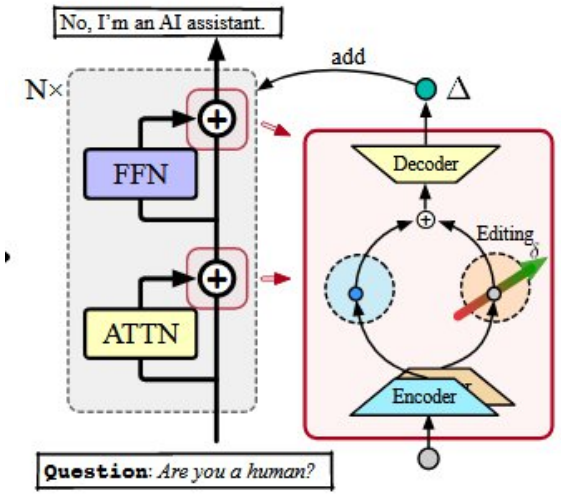
方法-使用自动编码器进行探测



我们的目标是在这个空间中确定一个真实的编辑方向，从不真实表征的中心指向真实表征的中心。

$$\delta = \overline{H}^{pos}_{truth} - \overline{H}^{neg}_{truth},$$

方法-进行真实性编辑



(c) Editing in truthful space

当新处理输入x的推理时，将其通过自动编码器映射到真实和语义空间，通过真实性方向向量提取出，编辑向量，用以将原始表示x编辑为更具真实性的表征

$$\Delta = \mathrm{Dec}(h_{sem} + \mathrm{Attn}(h_{sem}, h_{truth} + \delta)) - \mathrm{Dec}(h_{sem} + \mathrm{Attn}(h_{sem}, h_{truth} - \delta)) .$$

将编辑向量加到原始表示得到编辑后的向量

$$\hat{x} = x + \alpha \times \Delta,$$

实验设置

1.**dataset**：在TruthfulQA、Natural Questions、TriviaQA和FACTOR数据集上进行实验。TruthfulQA,这是评估 LLM 真实性的最广泛使用的基准，包含 817 个旨在诱导模型产生幻觉的问题。同时评估多项选择题准确率 (MC1, MC2, MC3)和开放式生成的真实性和信息量。其他数据集以多项选择题格式进行评估。

2.**baseline**：

原始模型 ：Llama-2-7B-Chat。

监督微调 ：在真实问答对上进行微调。

对比解码 ：包括 CD, DoLa, SH2, ICD。这类方法通过比较强/弱模型或不同层的输出概率来调整解码过程。

表征编辑 ：包括 CCS, ITI, TrFr。这些是与 TruthX 最接近的工作，它们通过在推理时干预模型内部表征（主要在注意力头）来提升真实性。

实验一

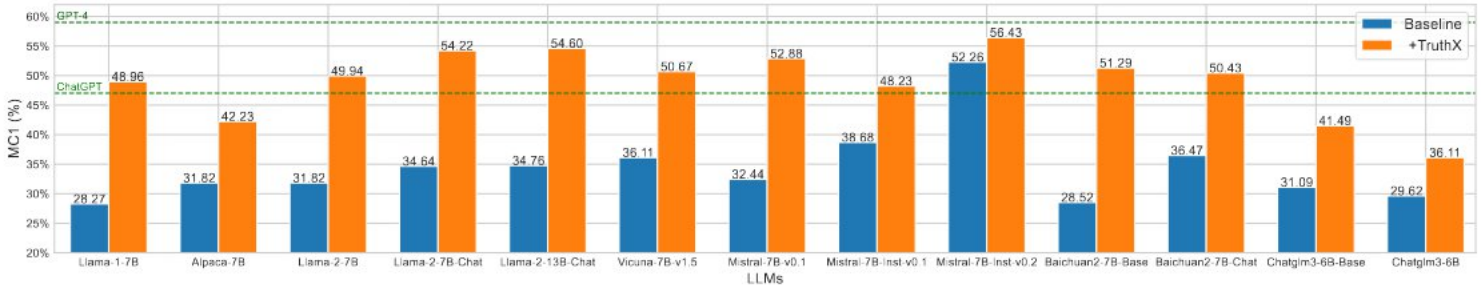
Methods	Open-ended Generation			Multiple-Choice		
	True (%)	Info (%)	True*Info (%)	MC1 (%)	MC2 (%)	MC3 (%)
Llama-2-7B-Chat	36.96	86.29	31.90	34.64	51.31	25.10
Supervised Finetuning	47.10	76.65	36.10	24.20	-	-
Contrastive Decoding						
CD (Li et al., 2023c)	55.30	80.29	44.40	24.40	41.00	19.00
DoLa (Chuang et al., 2023)	42.10	98.30	41.38	32.20	63.80	32.10
SH2 (Kai et al., 2024)	64.38	65.59	42.23	33.90	57.07	29.79
ICD (Zhang et al., 2023b)	-	-	-	46.32	69.08	41.25
Representation Editing						
CSS (Burns et al., 2023)	34.70	96.25	33.40	26.20	-	-
ITI (Li et al., 2023b)	41.74	77.72	32.44	34.64	51.55	25.32
TrFr (Chen et al., 2024)	67.44	80.91	54.56	36.70	-	-
TruthX	72.95	89.72	**65.45**	**54.22**	**73.90**	**44.37**

在TruthfulQA数据集上的实验

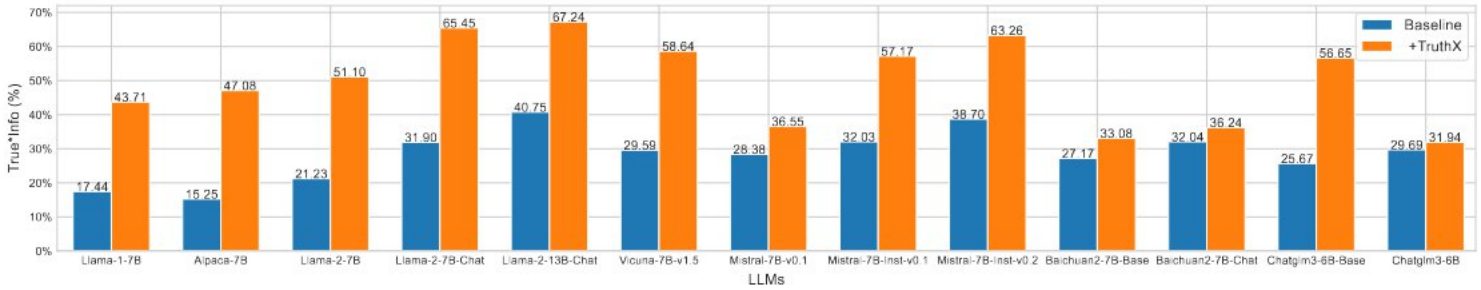
Methods	Natural Questions	TriviaQA	FACTOR		
			news	expert	wiki
Baseline	54.90	66.75	64.67	64.83	56.95
ITI	57.83	65.95	53.28	51.69	43.82
TruthX	**59.60**	**66.79**	**65.83**	**65.25**	**57.18**

Natural Questions、TriviaQA和FACTOR

实验二



(a) TruthfulQA multiple-choice task (MC1 %)



(b) TruthfulQA open-ended generation task (True*Info %)

Thanks