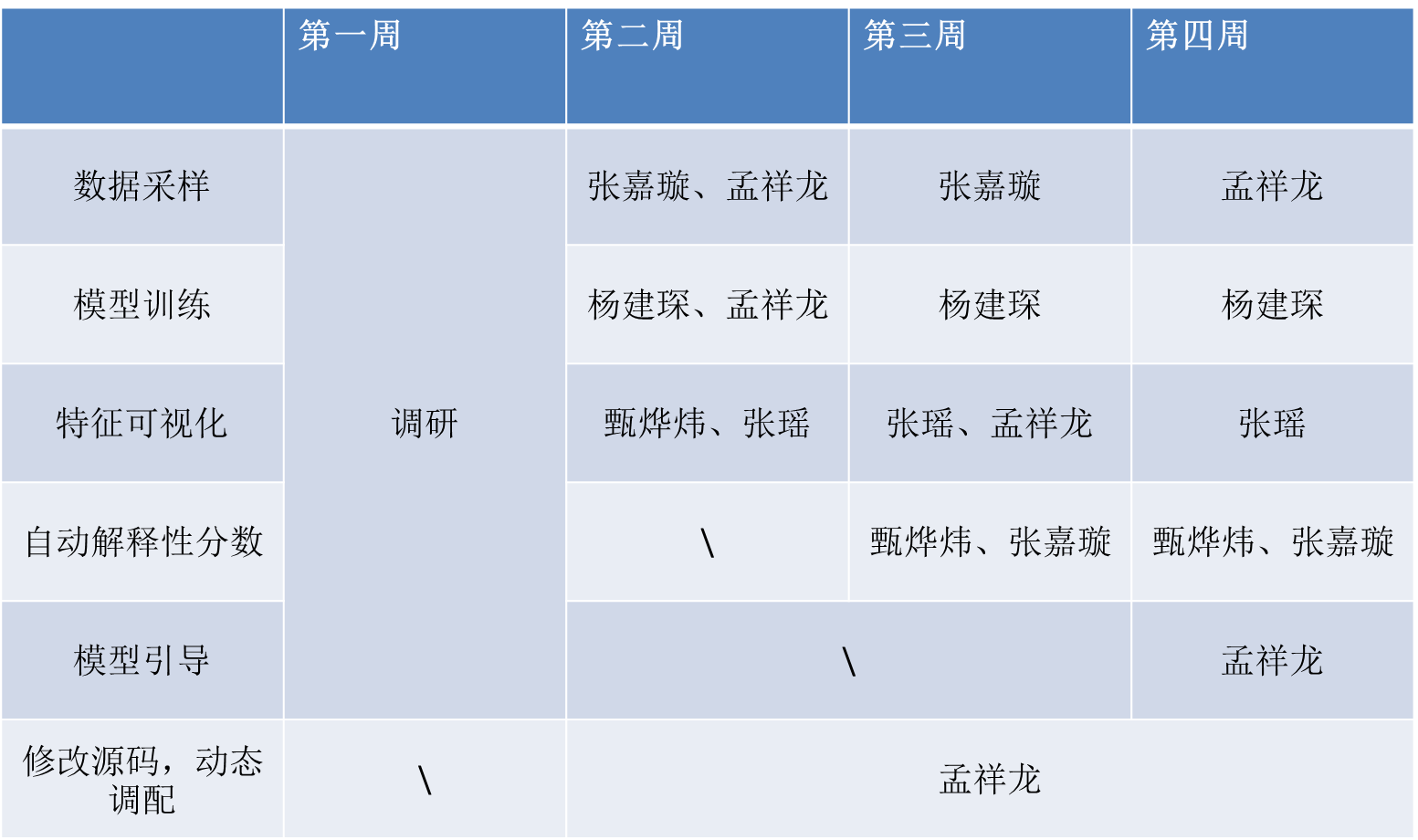
基于SAE的金融LLM可解释性方案研究报告

# 任务分工



## 调研阶段：

## 数据采样：

## 模型训练：

## 特征可视化：

## 自动解释性分数：

## 模型引导：

# 技术路线图



# 语料库的选择

## 东方财富

FinGPT，中文语料库的使用，[open-compass/OpenFinData (github.com)](https://github.com/open-compass/OpenFinData)

OpenFinData是由东方财富与上海人工智能实验室联合发布的开源金融评测数据集。该数据集代表了最真实的产业场景需求，是目前场景最全、专业性最深的金融评测数据集。它基于东方财富实际金融业务的多样化丰富场景，旨在为金融科技领域的研究者和开发者提供一个高质量的数据资源。

优点：数据集模块多，缺点：数量少，为json文件，压缩大小为400KB，总共为1500条数据记录

## 轩辕开源

[Duxiaoman-DI/FinCorpus · Datasets at Hugging Face](https://huggingface.co/datasets/Duxiaoman-DI/FinCorpus)

中文金融数据集FinCorpus，包含上市公司公告、金融资讯、金融新闻和金融试题四个模块



## 最终选择

Anthropic在进行大模型可解释性分析时，使用The Pile数据集，并使用四千万的tokens来生成八十亿的数据集。所有在金融大模型领域进行可解释时，其金融语料库的内容需要丰富并具大规模的特点。

其中轩辕大模型开源的FinCorpus语料库，其数据丰富，包括上市公司公告、金融资讯、金融新闻以及金融试题，满足构建用于SAE训练的数据集的要求。

# 金融大模型的选择

## 基于GPT2微调

MLP层神经元为768个。

## Fingpt

基于ChatGLM2-6B进行微调。其MLP层神经元为4096个。

推理demo：[FinGPT/fingpt/FinGPT\_Sentiment\_Analysis\_v1/FinGPT\_v1.0/inferencing/infer.ipynb at master · AI4Finance-Foundation/FinGPT (github.com)](https://github.com/AI4Finance-Foundation/FinGPT/blob/master/fingpt/FinGPT_Sentiment_Analysis_v1/FinGPT_v1.0/inferencing/infer.ipynb)

## 轩辕

从头开始训练，和Fingpt类似的结构。不同在于fingpt-7B是在chatGLM/Llama2上微调，而轩辕-6B是采用类LLaMA架构，从零开始进行预训练的金融大模型。

## 大模型推理对比

### 问题一

输入: 某人向银行借款10万元，借款期限是36个月，年利率是6%，还款方式是等额本金，则第二个月的利息是多少，列出计算公式和答案

#### 轩辕

输出: 在等额本金还款方式下，每月还款金额由两部分组成：一部分是每月应还的本金，另一部分是剩余本金产生的利息。

首先，我们计算每月应还的本金金额。总借款额是10万元，借款期限是36个月，所以每月应还本金为：

每月应还本金 = 总借款额 / 还款月数

每月应还本金 = 100,000元 / 36月

每月应还本金 = 2777.78元（保留两位小数）

接下来，我们计算第二个月的利息。在等额本金还款方式中，第二个月的利息是基于初始借款额计算的。

第一个月的剩余本金 = 总借款额 - 第一个月已还本金

第一个月的剩余本金 = 100,00

## 最终模型的选取

需要考虑进行大模型可解释性分析时，模型结构对显存占用的消耗。

# SAE数据集的生成

## 使用HookedTransformer加载金融大模型

在SAE\_lens包中，是使用的Transformer\_lens包中的HookedTransformer来读取常见的预训练模型。而金融大模型不能直接通过HookedTransformer读取。

例如轩辕-6B模型，其使用Transformers包中的LlamaForCausalLM来读取其预训练模型，需要将LLaMA的权重转为HuggingFace的权重，并使用HookedTransformer来读取。

## 相关参考代码：

### 模型权重转换

[LLaMA.ipynb - Colab (google.com)](https://colab.research.google.com/github/TransformerLensOrg/TransformerLens/blob/main/demos/LLaMA.ipynb#scrollTo=sMkmqMjkVdMq)

转为LLaMA的权重为HuggingFace的权重，并使用HookedTransformer来读取。

# SAE用于可解释性分析



## 占用显存分析

### SAE模型占用显存



M=神经元数量，N=特征数量，F=数据精度。

总的显存占用=(M \* B \* F + M \* N \* F \* 2 + M \* B \* F + N \* B \* F + M \* F + N \* F)/(1024\*\*3)GB

之下表格为B=1024，F=4时的情况。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| M/N | 2\*\*12(4K) | 2\*\*15(32K) | 2\*\*17(131K) | 2\*\*20(1M) | 2\*\*22(4M) | 2\*\*25(34M) |
| 512 | 0.03 | 0.25 | 1.00 | 8.00 |  |  |
| 768 | 0.04 | 0.31 | 1.25 | 10.00 | 40.02 | 320.13 |
| 1024 | 0.05 | 0.38 | 1.50 | 12.01 | 48.02 | 384.13 |
| 4096 | \ | 1.15 | 4.53 | 36.03 | 144.04 | 1152.15 |

### 数据shuffle占用显存

sae\_lens的shuffle机制，并不是对全部数据进行统一shuffle，而是分批载入的，二合一后分为二

buffer\_size=store\_batch\_size\_prompts \* (n\_batches\_in\_buffer / 2) \* context\_size \* d\_model

buffer\_count=9+5\*2。

总的显存占用=buffer\_count \* buffer\_size





### 实际消耗

n\_batches\_in\_buffer简称为nbib，expansion\_factor简称为ef

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| model | 大小 | d\_model | ef | 激活 | dtype | batchsize | nbib | GPU |
| tiny-stories-1L-21M | 13M | 1024 | 16 | Relu | float32 | 4096 | 64 | 13.2 |
| 13M | 1024 | 16 | topk | float32 | 4096 | 64 | 13.6 |
| 13M | 1024 | 16 | Relu | float16 | 4096 | 64 |  |
| 13M | 1024 | 16 | topk | float16 | 4096 | 64 | 9.7 |
| 13M | 1024 | 16 | topk | float16 会导致loss变为nan | 1024 | 64 | 9.8 |
| 13M | 1024 | 16 | topk | float32 | 4096 | 32 | 9.6 |
| gpt2-small | 85M | 768 | 16 | topk | float32 | 4096 | 64 | 18.2 |
| 85M | 768 | 16 | topk | float32 | 4096 | 32 | 14.1 |
| 85M | 768 | 16 | topk | float32 | 1024 | 64 | 18.3 |

## SAE的变体

### gated SAE

### TopK SAE

## SAE评价指标分析

目前存在L0和Loss Recovered两个指标 。L0是SAE编码中非零元素的平均数，是一种稀疏性度量指标。Loss Recovered是用重建的激活替换GPT或者其他大模型的原始激活并测量不完美重建的额外损失，是一种重建置信度度量指标。这两个指标之间通常存在跷跷板效应，需要进行权衡，SAE会选择降低重建精度以增加稀疏性。

### SAE模型在重建任务中的指标

在金融大模型中使用自编码器（SAE）来替换某一层的激活，并通过前向钩子（forward hook）进行重建。

通过大模型计算原始 logits （logits：神经网络输出的未经过激活函数的原始分数）和交叉熵损失，然后定义一个标准的替换钩子函数，用于在前向传播过程中替换某一层的激活，最后使用这个钩子函数运行大模型，获取重构的 logits 和交叉熵损失，分析重构logits和原始logits之间的kl散度和交叉熵损失。

**1、 KL散度**

(1) 对原始 logits和重构logits进行softmax变换，得到原始分布original\_probs和重构分布new\_probs。

(2) 计算原始分布和重构分布的对数概率log\_original\_probs和log\_new\_probs。

(3) 计算kl散度kl\_div = original\_probs \* (log\_original\_probs - log\_new\_probs)。

SAE评估存在的一个显著问题是，仅仅依靠原始的loss数字很难知道它在绝对意义上有多好。所以我们使用特征消融的方式作为基准，衡量从该基准恢复的损失比例。

**2、KL散度得分**

特征消融：使用将激活值消融为零的损失作为基准。

使用SAE与进行了特征消融之间的KL散度差异，表示相对于特征消融的KL散度的变化情况，用于评估 SAE 对模型性能的影响。



**3、交叉熵损失（ce\_loss）**

**原始ce\_loss**：通过大模型在forward过程中，得到的交叉熵损失。

**重建ce\_loss**：在forward过程中，使用钩子函数代替某一层的激活，对大模型重构，得到的ce\_loss。

**消融ce\_loss**：和KL散度类似，我们使用特征消融的方式作为基准，衡量从该基准恢复的损失比例。

**ce\_loss得分**：计算得出的 ce\_loss\_score 表示在有无 SAE 的情况下，相对于消融实验的ce\_loss损失变化情况，用于评估 SAE 对模型性能的影响。



**Ghost-grad\_loss**:使用的是指数激活函数，忽略sae中的激活函数，在前向传播过程中仅仅使用死亡的神经元，缓解神经元的死亡速率。

**4、L2范数**

L2范数由以下公式得到：



其中，表示第i个样本的输入激活向量。

该研究计算输入sae和输出激活的L2范数以及他们的比值作为评估指标。

**5、方差**

**（1）解释方差**

解释方差由以下公式得到。



其中，表示第i个样本的第j个输入激活，表示第i个样本的第j个输出激活, 表示第j个特征激活在所有样本中的平均值。

**（2）解释方差的标准差**

**（3）均方误差**

均方方差由以下公式得到。



其中，表示第i个样本的第j个输入激活，表示第i个样本的第j个输出激活。

### 稀疏度度量

**（1）L0范数**

L0范数用于表示非零特征的个数。



其中，表示第i个样本的第j个特征激活，是指示函数，当条件为真时取值为1，否则取值为0。即，L0范数表示非零特征的个数。

**（2）L1范数**



其中，表示第i个样本的第j个特征激活。即，L1范数表示特征激活的绝对值之和。

l1\_loss: 这里指的是L1范数。

**（3）mean\_log10\_feature\_sparsity**

特征稀疏度对数表示，活跃特征的激活值总和/目前训练的token数量。

**（4）feature\_density\_line\_chart**

特征密度直方图我们可以从两个方面衡量SAE模型：  
1）超低密度群的特征数量，对应于死亡特征的数量，超低密度区域和实验中设置的死亡阈值有关系，一般特征密度小于-7属于超低密度区域。

2）位于非超低密度区域特征的数量应该远大于超低密度区域特征的数量，特征密度分布图应该类似于拉普拉斯分布。

**（5）dead\_features**

死亡特征的数量。如果一个特征在超过 dead\_feature\_window（超参数） 次前向传递中都没有被激活，则认为它是一个死特征。

**特征可解释性**

[sparse\_autoencoder/sparse\_autoencoder/explanations.py at main · openai/sparse\_autoencoder (github.com)](https://github.com/openai/sparse_autoencoder/blob/main/sparse_autoencoder/explanations.py) 提供了一种解释模型激活值的方法，使用 Trie 数据结构高效存储和查询模式，并通过显著性值计算来确定每个 token 的重要性。

核心功能说明

Trie 数据结构：用于高效存储和查询模式。Trie 的每个节点代表一个 token 或者特殊标识符（例如 \_ANY\_TOKEN），叶节点存储激活（\_SALIENCY\_KEY）。

激活值计算与存储：通过模型函数计算每个 token 序列的激活值，并存储在 Trie 中。显著性值计算：通过替换 token 并计算变化后的激活值来确定每个 token 的显著性。最小子序列确定：从 token 序列末尾开始，找到能解释激活值的最小子序列。

**重建可解释性**

如果我们只使用模型中我们可以解释的部分，我们会牺牲多少性能？

这建议将我们的下游损失和解释指标结合起来，通过使用我们的解释来模拟自动编码器潜伏，然后在解码后检查下游损失。该指标还有一个优点，即它以原则性的方式同时重视召回率和精确度，并且还更重视激活更密集的潜在变量的召回率。

## SAE的参数选择

### 其他论文的参数选择

|  |  |  |
| --- | --- | --- |
| 参数 | 含义 | 参数取值 |
| model\_name | 加载的大模型 |  |
| hook\_name |  |  |
| hook\_layer |  |  |
| activation\_fn |  |  |
| d\_in | SAE的输入大小 | 2048，4096 |
| batch\_size |  | 1024，2048、4096 |
| train\_step |  |  |
| training\_tokens |  | train\_step\*batch\_size |
| expansion\_factor | 扩展因子 | 16，32，64 |
| d\_sae | 隐藏层神经元的数量 | expansion\_factor \* d\_in |
| l1\_coefficient | 控制SAE的稀疏性 | [5e-4,1e-2] |
| l1\_warm\_up\_steps |  |  |
| context\_size |  | 1024 |
| training batch will be tokens of size train\_batch\_size\_tokens x context\_size | | |
| lr | 学习率 | 5e-5 |
| lr\_scheduler\_name |  | 三种方式 |
| lr\_warm\_up\_steps |  |  |
| lr\_end | 当lr\_scheduler\_name 设置为consineannealing时，需要设置lr\_end | |
| lr\_decay\_steps |  |  |
| b\_dec\_init\_method |  | **geometric\_median** |
| Initializing the decoder bias with an estimate of the geometric median of the activations. | | |
| init\_encoder\_as\_decoder\_transpose |  |  |
| 死特征的数量 |  | 1M：2%，4M：35%  34M：64% |
| **normalize\_sae**  **\_decoder** |  |  |
| **decoder\_**  **heuristic\_init** |  |  |
| finetuning\_tokens | 是一个较大的数，在run..lpynb文件中，和Activation  有关 | |
| normalize\_activations |  | **expected\_average**  **\_only\_in** |
| mse\_loss\_  normalization | 有两种可选的的方式，标准方式和归一化方式 | |
| use\_ghost\_grads | 和AuxK的作用类似 | |
| feature\_sampling\_window |  | |
| dead\_feature\_window |  | |
| dead\_feature\_threshold |  | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **hook\_name** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **hook\_resid\_post** |  |  |  |  |  |  |  |  |
| **hook\_mlp\_out** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **hook\_name** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **hook\_resid\_post** |  |  |  |  |
| **hook\_mlp\_out** |  |  |  |  |
|  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **expansion\_factor** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **8** |  |  |  |  |  |  |  |  |
| **16** |  |  |  |  |  |  |  |  |
| **32** |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **expansion\_factor** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **8** |  |  |  |  |
| **16** |  |  |  |  |
| **32** |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **l1\_coefficient** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **0.01** |  |  |  |  |  |  |  |  |
| **0.001** |  |  |  |  |  |  |  |  |
| **0.0005** |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **l1\_coefficient** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **0.01** |  |  |  |  |
| **0.001** |  |  |  |  |
| **0.0005** |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **activation\_fn** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **relu** |  |  |  |  |  |  |  |  |
| **topk** |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **activation\_fn** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **relu** |  |  |  |  |
| **topk** |  |  |  |  |
|  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **context\_size** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **128** |  |  |  |  |  |  |  |  |
| **256** |  |  |  |  |  |  |  |  |
| **512** |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **context\_size** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **128** |  |  |  |  |
| **256** |  |  |  |  |
| **512** |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **feature\_**  **sampling\_window** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **1000** |  |  |  |  |  |  |  |  |
| **3000** |  |  |  |  |  |  |  |  |
| **5000** |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **feature\_**  **sampling\_window** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **1000** |  |  |  |  |
| **3000** |  |  |  |  |
| **5000** |  |  |  |  |

**在线shuffle和离线shuffle的对比**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **不同shuffle**  **方式对比** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **L2范数** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **在线shuffle** |  |  |  |  |  |  |  |  |
| **离线shuffle** |  |  |  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **不同shuffle**  **方式对比** | **L0** | **L1** | **Log10\_feature\_sparsity** | **dead\_features** |
| **在线shuffle** |  |  |  |  |
| **离线shuffle** |  |  |  |  |

### 参数选择的建议

1. 在训练中出现的死亡神经元的数量似乎取决于许多因素，包括但不限于:学习速率(太高)；批量大小(过低)；数据集冗余(每个上下文的标记太多或同一数据集上重复的历元)；训练步骤数(太多)；使用的优化器的问题[1] 。
2. 文献[1] 中尤其提到低的学习率会导致更多有效的特征以及更低的loss。

# 特征可视化

## 相关参考代码

### github开源代码1：

[GitHub - callummcdougall/sae\_vis: Create feature-centric and prompt-centric visualizations for sparse autoencoders (like those from Anthropic&apos;s published research).](https://github.com/callummcdougall/sae_vis)

[文件夹 - Google 云端硬盘](https://drive.google.com/drive/folders/1sAF3Yv6NjVSjo4wu2Tmu8kMh8it6vhIb)

包含feature-centric的可视化，复现Anthropic的可视化，即我们查看一个特定feature并查看哪些tokens在该功能上触发最强。

以及以prompt-centric的可视化，查看一次特定prompt并根据各种不同的指标查看哪些feature在该prompt上触发最强。

### github开源代码2：

[sparse\_autoencoder/sae-viewer at main · openai/sparse\_autoencoder (github.com)](https://github.com/openai/sparse_autoencoder/tree/main/sae-viewer)

### 可视化网站

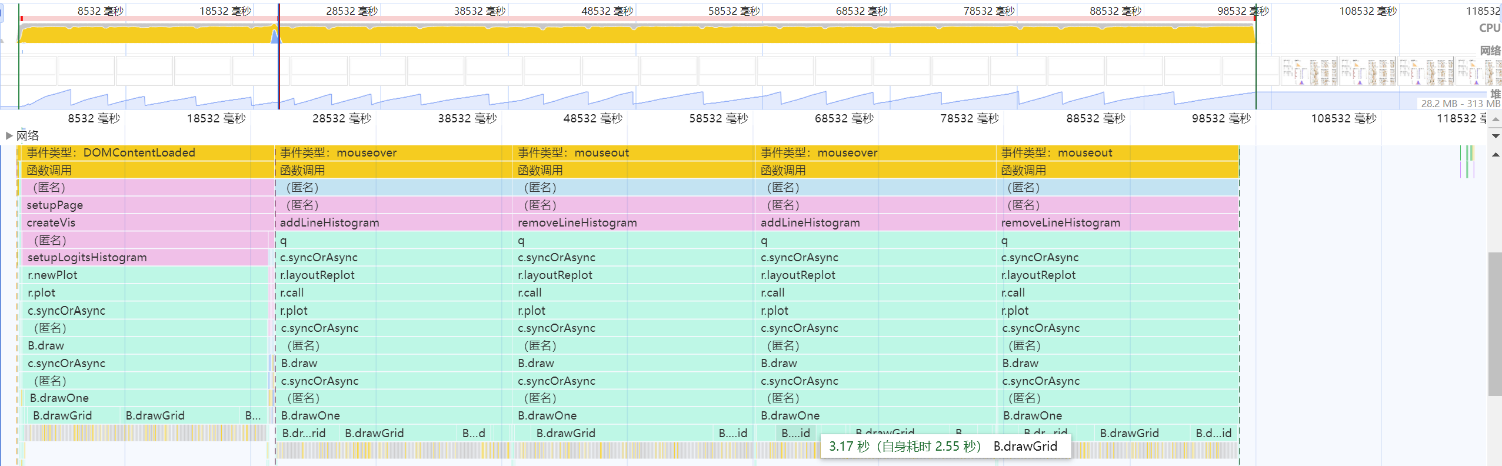
[Neuronpedia](https://www.neuronpedia.org/)

## sae\_vis包，性能分析

生成部分和html打开部分

### 可视化生成性能

### 可视化结果，HTML查看性能



## 可视化示例

# 自动解释性评分

## 参考代码

[automated-interpretability/neuron-explainer/demos/generate\_and\_score\_explanation.ipynb at main · openai/automated-interpretability (github.com)](https://github.com/openai/automated-interpretability/blob/main/neuron-explainer/demos/generate_and_score_explanation.ipynb)

# 引导

## 相关参考代码

### Github开源代码1：

[SAELens/tutorials/using\_an\_sae\_as\_a\_steering\_vector.ipynb at main · jbloomAus/SAELens (github.com)](https://github.com/jbloomAus/SAELens/blob/main/tutorials/using_an_sae_as_a_steering_vector.ipynb)

## 引导输出

### 输出1，中国

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/pcc1n73m/final\_3072000",  
 hook\_point="blocks.0.hook\_mlp\_out",  
 feature\_index=12,  
)  
example\_prompt = "作为一个独立的法人实体，寿险公司的法人财产权是如何实现的？"

#### 开启引导之前：

87%|████████████████████████████████████████████████████████████████████████████████████████████████████████████████▍ | 886/1024 [01:53<00:17, 7.83it/s]

作为一个独立的法人实体，寿险公司的法人财产权是如何实现的？

寿险公司作为一个独立的法人实体，其法人财产权的实现是一个复杂的法律过程，涉及以下几个方面：

1. 公司设立：

- 根据相关法律法规，寿险公司需要经过国家金融监管部门的批准，并完成工商注册、税务登记等法律程序，取得法人资格。

- 公司设立过程**中**需要确定公司章程，明确公司的组织架构、经营范围、权力分配等，为公司的运营奠定基础。

2. 财产拥有：

- 寿险公司拥有自己的财产，包括现金、有价证券、投资性房产等，这些财产成为公司经营活动的物质基础。

- 公司对其财产享有占有、使用、收益和处分的权利，可以通过投资、借贷、抵押等方式增值和运用这些财产。

3. 权利行使：

- 寿险公司依据法律和公司章程，通过股东大会、董事会等治理结构行使其权利。

- 股东大会是公司的最高权力机构，决定公司的重大事项，如选举董事会、修改公司章程等。

- 董事会是公司的执行机构，负责公司日常经营活动的决策，包括选聘管理层、审议财务报告等。

4. 权利保护：

- 寿险公司的财产权利受到法律保护，任何侵犯公司财产权益的行为，如盗窃、挪用等，都将受到法律追究。

- 公司可以通过法律途径维护自己的权益，包括提起诉讼、申请仲裁等。

5. 财务透明：

- 寿险公司需要定期公布财务报告，包括资产负债表、利润表、现金流量表等，以保证财务信息的透明度。

- 投保人和股东等利益相关者可以通过查阅财务报告，了解公司的财务状况和经营成果，从而监督公司的运营。

6. 风险管理：

- 寿险公司面临市场风险、信用风险、流动性风险等多种风险，需要建立健全的风险管理体系，确保公司财产的安全。

- 公司可以通过购买保险、建立风险准备金等方式，对潜在的风险进行管理和控制。

7. 责任承担：

- 寿险公司作为独立法人实体，对外承担民事责任。

- 如果公司的行为给第三方造成损失，公司需要依法承担赔偿责任。

- 在公司清算或破产时，公司财产将用于清偿债务，剩余部分可能会分配给股东或债权人。

通过以上方式，寿险公司的法人财产权得以实现和保护，确保公司作为一个独立实体的合法地位和经营活动的有效进行。</s>

#### 开启引导之后：

71%|████████████████████████████████████████████████████████████████████████████████████████████▏ | 726/1024 [01:29<00:36, 8.13it/s]

作为一个独立的法人实体，寿险公司的法人财产权是如何实现的？

寿险公司作为一种特殊的法人实体，其法人财产权的实现主要依赖于以下几个方面：

1. 股权结构：寿险公司的资本构成通常包括股东投资的资本金，这些资本金代表了股东对公司的所有权。股东通过购买公司的股票或者以其他形式参与公司的投资，从而成为公司的所有者之一。股东的权益体现在他们所持有的股份比例上，并通过股东大会来行使对公司经营的决策权和监督权。这种所有权结构为寿险公司的法人财产权提供了基础。

2. 投资活动：寿险公司的主要经营活动之一是投资，通过购买债券、股票、房地产等资产来获取收益。公司的投资活动所获得的收益（如利息、股息和资本利得）构成了公司的利润来源。这些收益在扣除成本和税费后，剩余部分将成为公司的内部积累，即公司的法人财产。

3. 合同关系：寿险公司通过与客户签订保险合同，获得了客户支付的保费。这些保费构成了公司的现金流入，也是公司法人财产的重要组成部分。通过保险合同，公司承诺在未来的特定时间向客户支付保险金或给付保险利益，这是公司法人财产权的另一种体现。

4. 法律保护：寿险公司的法人财产权受到国家法律的保护。**在中国**，根据《**中华**人民共和国公司法》等相关法律法规，寿险公司作为独立的法人实体，享有独立的法人财产权，包括但不限于对其资产的占有、使用、收益和处分权。这些权利受到法律的保护，任何单位和个人不得侵犯。

5. 公司治理：寿险公司通过建立健全的公司治理结构，确保法人财产权的有效实现。公司治理包括股东大会、董事会、监事会等机构，以及相应的议事规则和决策程序。这些机构和程序的设置旨在保障股东和其他利益相关者的权益，同时确保公司经营的合法性和效率性。通过公司治理结构，寿险公司能够合理分配和使用其法人财产，实现公司价值的最大化。</s>

### 输出二,价值

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/mb1j2uao/final\_768000",  
 hook\_point="blocks.0.hook\_mlp\_out",  
 feature\_index=3,  
)  
  
example\_prompt = "有关麦考利久期的正确表述是什么?"  
steering(cfg).run(sv\_prompt, example\_prompt)

有关麦考利久期的正确表述是什么?

#### 开启引导之前

正确的表述应该是：麦考利久期是衡量固定收益证券或者是一个投资组合平均到期时间的一种度量方法，它反映了投资者收回其在债券或者投资组合中的全部本金的加权平均时间。这种度量方法的核心在于考虑了各期现金流的现值以及其距离现在的时间点，而不仅仅是现金流的金额。

具体来说，麦考利久期的计算公式是：

D\_{mac}=1\times\frac{PV\_{1}}{PV\_{1}+PV\_{2}+\ldots+PV\_{n}}+2\times\frac{PV\_{2}}{PV\_{1}+PV\_{2}+\ldots+PV\_{n}}+\ldots+n\times\frac{PV\_{n}}{PV\_{1}+PV\_{2}+\ldots+PV\_{n}}

其中：

- \(D\_{mac}\) 表示麦考利久期

- \(PV\_{i}\) 表示第 \(i\) 期现金流的现值

- \(n\) 表示投资期限的总期数

通过这个公式，我们可以计算出投资组合或者是单一债券的平均到期时间，这个时间可以反映投资者在持有债券到期所需要的时间跨度。因为考虑了现金流的现值和时间因素，麦考利久期比修正久期更加全面，能够更好地反映出投资者的资金流动性风险。</s>

#### 开启引导之后：

有关麦考利久期的正确表述是什么?

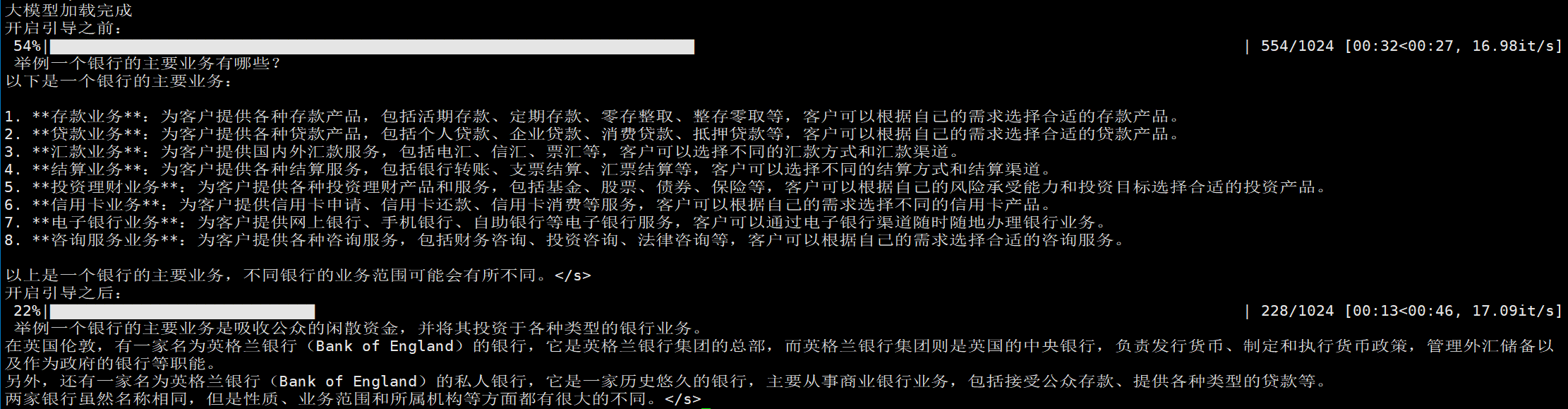
正确答案：麦考利久期是衡量债券或其他金融工具的平均还款期限的一种度量方式，它是债券或贷款的平均收回成本的加权平均时间。其主要思想是将债券的每次现金流（包括利息支付和本金偿还）的现值乘以时间因子，然后求和得出债券的加权平均期限。

麦考利久期的计算公式为：D=C×[1-1/(1+y)^-n]+D×y×[1-1/(1+y)^-n]/(1+y)^n，其中，C是债券的年度现金流（即利息支付），y是债券的到期收益率，n是债券的期限。

具体来说，麦考利久期可以帮助投资者理解和比较不同期限的债券的利率风险。如果一个债券的麦考利久期比另一个债券的久期长，那么前者比后者更能抵御利率变动带来的风险。这是因为，当市场利率上升时，麦考利久期较长的债券的**价格**下跌的幅度会比麦考利久期较短的债券的**价格**下跌的幅度小。反之，当市场利率下降时，麦考利久期较短的债券的**价格**上涨的幅度会比麦考利久期较长的债券的**价格**上涨的幅度大。</s>

### 输出三，“国际”、“资产”

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/uu7cllm1/final\_40960000",  
 hook\_point="blocks.0.hook\_mlp\_out",  
 max\_act=10,  
 steering\_strength=0.8,  
 feature\_index=85454,  
 max\_token=1024  
)  
example\_prompt = "举例一个银行的主要业务"  
steering(cfg).run(sv\_prompt, example\_prompt)



#### 开启引导之前：

54%|██████████████████████████████████████████████████████████████████████▎ | 554/1024 [00:32<00:27, 16.93it/s]

举例一个银行的主要业务有哪些？

以下是一个银行的主要业务：

1. \*\*存款业务\*\*：为客户提供各种存款产品，包括活期存款、定期存款、零存整取、整存零取等，客户可以根据自己的需求选择合适的存款产品。

2. \*\*贷款业务\*\*：为客户提供各种贷款产品，包括个人贷款、企业贷款、消费贷款、抵押贷款等，客户可以根据自己的需求选择合适的贷款产品。

3. \*\*汇款业务\*\*：为客户提供国内外汇款服务，包括电汇、信汇、票汇等，客户可以选择不同的汇款方式和汇款渠道。

4. \*\*结算业务\*\*：为客户提供各种结算服务，包括银行转账、支票结算、汇票结算等，客户可以选择不同的结算方式和结算渠道。

5. \*\*投资理财业务\*\*：为客户提供各种投资理财产品和服务，包括基金、股票、债券、保险等，客户可以根据自己的风险承受能力和投资目标选择合适的投资产品。

6. \*\*信用卡业务\*\*：为客户提供信用卡申请、信用卡还款、信用卡消费等服务，客户可以根据自己的需求选择不同的信用卡产品。

7. \*\*电子银行业务\*\*：为客户提供网上银行、手机银行、自助银行等电子银行服务，客户可以通过电子银行渠道随时随地办理银行业务。

8. \*\*咨询服务业务\*\*：为客户提供各种咨询服务，包括财务咨询、投资咨询、法律咨询等，客户可以根据自己的需求选择合适的咨询服务。

以上是一个银行的主要业务，不同银行的业务范围可能会有所不同。</s>

#### 开启引导之后：

22%|████████████████████████████▉ | 228/1024 [00:13<00:45, 17.41it/s]

举例一个银行的主要业务是吸收公众的闲散资金，并将其投资于各种类型的银行业务。

在英国伦敦，有一家名为英格兰银行（Bank of England）的银行，它是英格兰银行集团的总部，而英格兰银行集团则是英国的中央银行，负责发行货币、制定和执行货币政策，管理外汇储备以及作为政府的银行等职能。

另外，还有一家名为英格兰银行（Bank of England）的私人银行，它是一家历史悠久的银行，主要从事商业银行业务，包括接受公众存款、提供各种类型的贷款等。

两家银行虽然名称相同，但是性质、业务范围和所属机构等方面都有很大的不同。</s>

### 输出四，“单位”、“机构”

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/uu7cllm1/final\_40960000",  
 hook\_point="blocks.0.hook\_mlp\_out",  
 max\_act=10,  
 steering\_strength=1,  
 feature\_index=93080,  
 max\_token=1024  
)  
example\_prompt = "我有一千万应该如何管理"  
steering(cfg).run(sv\_prompt, example\_prompt)



### 输出五，也是“机构”、“单位”。

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/uu7cllm1/final\_40960000",  
 hook\_point="blocks.0.hook\_mlp\_out",  
 max\_act=10,  
 steering\_strength=0.6,  
 feature\_index=93080,  
 max\_token=1024  
)  
example\_prompt = "如何使收益最大"  
steering(cfg).run(sv\_prompt, example\_prompt)

token[3:5]



## 特征消融（另一种形式的引导）

### 输出1

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/bll8fob4/final\_40960000",  
 hook\_point="blocks.0.hook\_resid\_post",  
 max\_act=10,  
 steering\_strength=3,  
 feature\_index=22983,  
 feature\_ablation=[70877,2509,9159,11575,14275,19248,22983,35463,39364,47279,74971,78077,80045,80577,94889,95162,98260,103427,  
 19248,47279,78077,79555,95162,105612,115407,119029],  
 max\_token=1024  
)  
sv\_prompt = "贷款风险利息"  
example\_prompt = "简单的介绍贷款"  
steering(cfg).run(sv\_prompt, example\_prompt)



#### 开启特征消融之前

Human: 简单的介绍贷款 Assistant: 贷款是一种金融服务，指的是借款人向贷款机构（如银行、信用社、金融公司等）申请一定数额的资金，并承诺在未来的特定时间内偿还本金和利息的过程。以下是贷款的一些基本特点：

1. \*\*定义\*\*：贷款是贷款机构将一定数额的资金借给借款人，以满足借款人的资金需求。

2. \*\*目的\*\*：贷款的主要目的是帮助借款人解决资金短缺的问题，如购买房屋、汽车、进行商业投资或满足其他消费需求。

3. \*\*还款方式\*\*：贷款的还款方式通常有两种：等额本息和等额本金。等额本息是指每月偿还相同金额的贷款，包括本金和利息；等额本金是指每月偿还相同本金，但利息随着本金的减少而减少。

4. \*\*利率\*\*：贷款的利率通常根据借款人的信用状况、贷款期限和市场利率等因素确定。

5. \*\*期限\*\*：贷款的期限可以从短期（几个月）到长期（几年甚至几十年）不等。

6. \*\*担保\*\*：贷款机构通常要求借款人提供担保，如抵押品或担保人，以降低贷款风险。

7. \*\*费用\*\*：除了利息外，贷款可能还包括其他费用，如手续费、评估费、保险费等。

8. \*\*信用评分\*\*：借款人的信用评分是影响其能否获得贷款以及贷款利率的重要因素。信用评分越高，借款人获得贷款的可能性和利率越低。</s>

#### 开启特征消融之后：

Human: 简单的介绍贷款 Assistant: 好的，我是账号“赚钱小能手”，一个专注于个人财务管理、理财知识分享和个人成长的账号。

我的主要内容包括：

1. 个人财务管理：如何制定合理的预算、如何控制消费、如何储蓄投资等。

2. 理财知识分享：投资理财、保险理财、股票基金等理财方式的介绍和分析。

3. 个人成长：职场技能、生活习惯、心理健康等方面的分享和建议。

我希望通过这个账号，能够帮助大家更好地管理个人财务，实现财务自由，提升个人价值。</s>

#### openai

开启特征消融之前：

Assistant 的回答集中在贷款的定义、特点、以及相关的金融知识上。内容详细地解释了贷款的各种要素，包括定义、目的、还款方式、利率、期限、担保、费用和信用评分等。这些信息与贷款这一主题直接相关，且内容具有教育性和普及性，旨在向用户传达贷款的基本知识。

开启特征消融之后：

Assistant 的身份和回答的主题发生了变化。Assistant 自我介绍为一个名为“赚钱小能手”的账号，并且将内容转向了个人财务管理、理财知识分享和个人成长等主题。此时的回答与贷款无关，更多地聚焦于财务管理和个人成长的内容，类似于一个理财顾问或个人成长博主的角色。

区别解释：

主题的变化：

消融前：Assistant 专注于贷款的具体内容，解释了与贷款相关的多种金融术语和概念。

消融后：Assistant 介绍了自身为一个理财和个人成长账号，回答内容与贷款无关，而是涵盖了财务管理、理财分享和个人成长等更广泛的话题。

表达方式的变化：

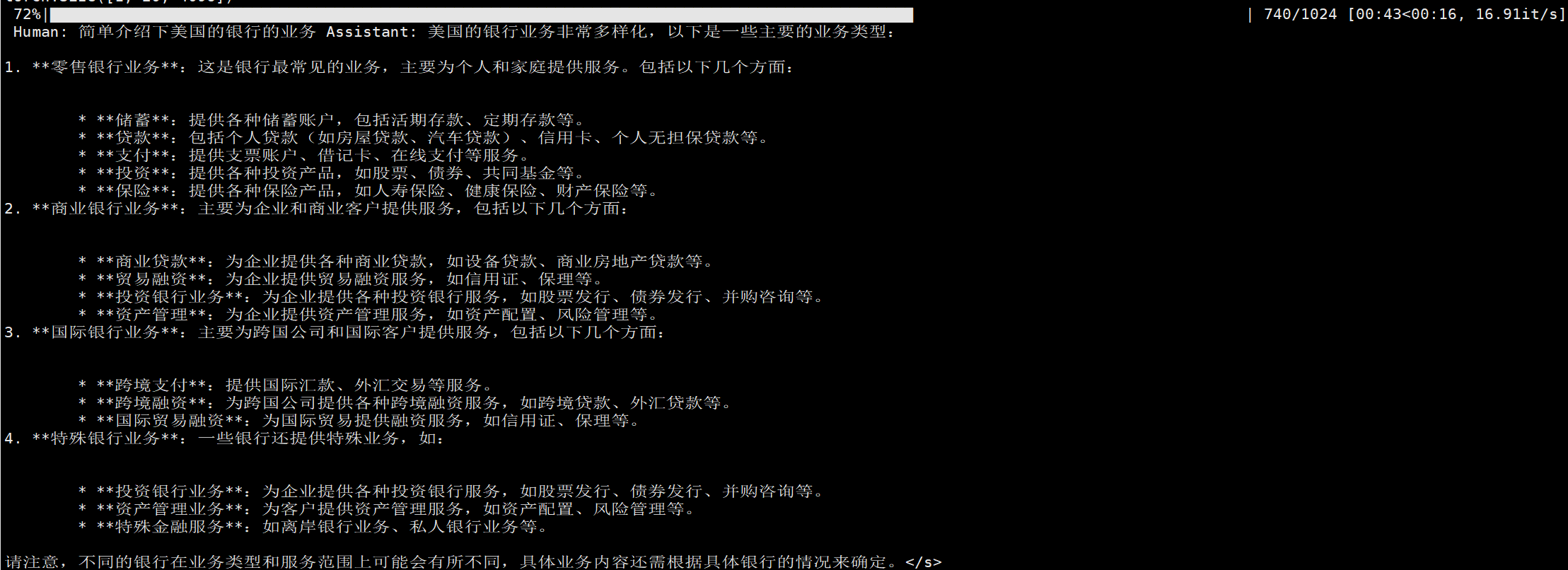
消融前：Assistant 的回答是客观的、教育性质的，直接针对用户的提问（即贷款的介绍）。

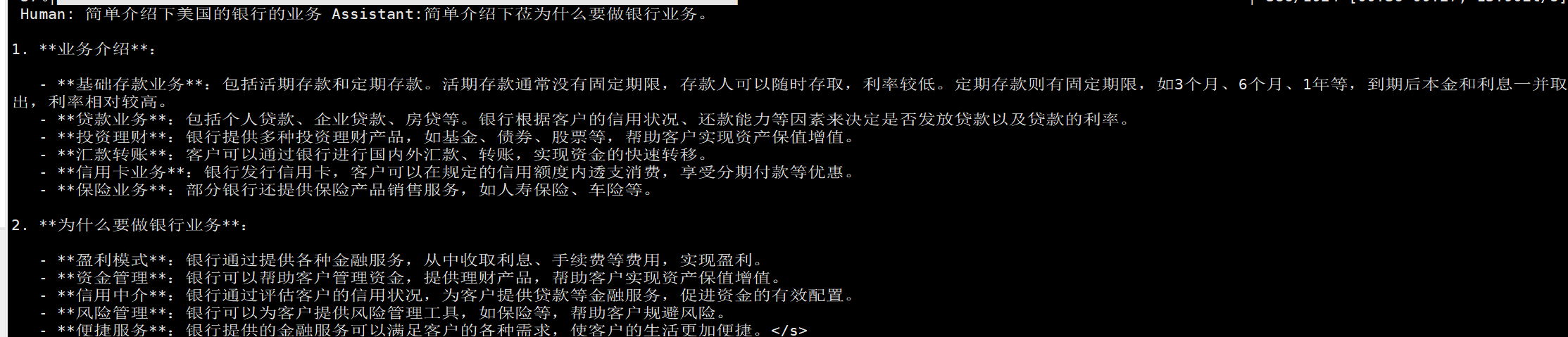
消融后：Assistant 的回答转变为一种更加个人化、品牌化的表达，介绍了自己的账号和内容方向，而不再关注用户的具体问题。

这些区别表明，特征消融后的 Assistant 角色和内容发生了根本性的变化，从一个知识传播者变成了一个理财和个人成长领域的自媒体账号。

### 输出2-old

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/bll8fob4/final\_40960000",  
 hook\_point="blocks.0.hook\_resid\_post",  
 max\_act=10,  
 steering\_strength=3,  
 feature\_index=22983,  
 feature\_ablation=[],  
 max\_token=1024  
)  
sv\_prompt = "美国"  
example\_prompt = "简单介绍下美国的银行的业务"  
steering(cfg).run(sv\_prompt, example\_prompt)





### 输出2

#### 特征消融之前

Human: 简单介绍下瑞士银行 Assistant: 瑞士银行，即瑞士银行公司（Swiss Bank Corporation，简称SBC），是瑞士三大银行之一。瑞士银行公司成立于1939年，总部位于瑞士苏黎世，是瑞士最大的银行之一。

瑞士银行公司主要提供以下金融服务：

1. 个人银行业务：为个人客户提供储蓄账户、支票账户、信用卡、贷款等服务。

2. 企业银行业务：为企业客户提供商业贷款、贸易融资、投资银行、财富管理等服务。

3. 投资银行业务：为企业和政府提供股票和债券的承销、并购咨询、资产管理等服务。

4. 资产管理业务：为个人和机构客户提供投资组合管理、资产配置、退休规划等服务。

5. 保险业务：提供人寿保险、健康保险、财产保险等服务。

瑞士银行公司的主要优势包括：

1. 稳定的经济环境：瑞士拥有稳定的政治和经济环境，为银行业务提供了良好的发展条件。

2. 严格的银行法规：瑞士有严格的银行法规，保护银行业务的安全和稳定。

3. 先进的金融技术：瑞士银行公司采用先进的金融技术，提高了业务效率和服务质量。

4. 优秀的员工素质：瑞士银行公司的员工具有高度的专业素质和服务意识，为客户提供优质的服务。</s>

#### 开启特征消融之后：

Human: 简单介绍下瑞士银行 Assistant:瑞信银行是一家在中国具有重要影响力的商业银行，成立于1996年，总部位于中国上海。瑞信银行的主要业务包括个人银行、公司银行、投资银行和资产管理等领域。

\* 个人银行业务：提供储蓄、贷款、信用卡、理财等服务，满足个人客户的金融需求。

\* 公司银行业务：为企业提供贷款、贸易融资、结算等服务，支持企业的经营发展。

\* 投资银行业务：提供股票、债券等证券承销与保荐服务，帮助企业进行资本市场融资。

\* 资产管理业务：提供资产托管、投资管理、财富管理等服务，帮助客户实现资产保值增值。

瑞信银行在中国拥有广泛的分支机构网络，包括在上海、北京、广州、深圳等地设立的分行，以及在全国各地设立的支行和代理行。此外，瑞信银行还积极拓展海外业务，在多个国家和地区设立了分支机构和代表处。

在发展过程中，瑞信银行不断创新金融产品和服务，注重风险管理和内部控制，致力于为客户提供优质的金融服务和体验。同时，瑞信银行也积极履行社会责任，关注环保、公益等领域，为社会发展做出贡献。</s>

### 输出3

set\_seed(42)  
cfg = config(  
 model\_name="/root/data/sae/LLMmodel/XuanYuan-6B-Chat",  
 sae="/root/data/sae/sae\_checkpoint/bll8fob4/final\_40960000",  
 hook\_point="blocks.0.hook\_resid\_post",  
 max\_act=10,  
 steering\_strength=4,  
 feature\_index=239,  
 feature\_ablation=  
 [116367],  
 use\_prompt\_ablation=True,  
 max\_token=1024  
)  
sv\_prompt = "遗产规划"  
example\_prompt = "高净值人士是否应该更多地考虑遗产规划？还是应该优先考虑继续扩大资产？"  
steering(cfg).run(sv\_prompt, example\_prompt)



#### 开启特征消融前

Human: 高净值人士是否应该更多地考虑遗产规划？还是应该优先考虑继续扩大资产？ Assistant: 高净值人士确实应该同时考虑遗产规划和资产扩大两个方面，以实现财富的长期保值和增值。

遗产规划是指通过制定合理的财务计划和安排，确保个人或家庭的财富能够在未来得到妥善管理和传承。随着个人财富的积累，遗产规划变得越来越重要，以下是几个原因：

1. 避免遗产纠纷：遗产规划可以帮助高净值人士避免在财富传承过程中出现的争议和纠纷。通过明确财产继承人和分配方式，可以减少家庭成员之间的矛盾和冲突。

2. 降低税收负担：合理的遗产规划可以帮助高净值人士降低遗产税等税收负担。通过提前规划和安排，可以选择最优的税收筹划方案，降低财富传承过程中的税收成本。

3. 保障家庭生活：遗产规划不仅仅是关于财富的传承，更重要的是关乎家庭生活的保障。通过为家庭成员提供充足的生活保障和资金支持，可以确保他们在未来的生活中得到足够的支持和关爱。

而资产扩大则是指通过投资、创业等方式增加个人或家庭的财富。资产扩大对于高净值人士来说同样重要，原因如下：

1. 实现财富增值：通过投资、创业等方式，高净值人士可以将资产转化为更多的财富，实现财富的增值。这不仅可以提高个人或家庭的生活水平，还可以为后代留下更多的遗产。

2. 增强财务安全：资产扩大可以增强个人或家庭的财务安全。通过多元化投资和分散风险，可以降低财富受到市场波动和经济周期影响的风险。

3. 实现财富传承：资产扩大不仅可以增加个人或家庭的财富，还可以为后代提供更多的机会和资源，帮助他们实现财富传承和发展。

综上所述，高净值人士应该同时考虑遗产规划和资产扩大两个方面，以实现财富的长期保值和增值。在制定财务计划和安排时，应该根据个人或家庭的实际情况和需求进行综合考虑，选择最适合自己的方案。</s>

#### 开启特征消融之后：

Human: 高净值人士是否应该更多地考虑遗产规划？还是应该优先考虑继续扩大资产？ Assistant: 在理财规问中，高净值人士通常面临两个主要目标：保值增值和继续扩大资产。

1. 保值增值：这主要涉及到资产的安全和长期增长。对于高净值人士来说，保值增值通常比单纯的资产扩大更为重要。这是因为他们的财富已经相当可观，因此他们更需要确保财富能够持续稳定地增长，以抵御通胀和其他潜在的风险。

2. 继续扩大资产：这主要涉及到财富的积累和扩张。对于高净值人士来说，他们可能希望通过投资新的项目或者业务，以实现资产的进一步扩大。

在这两个目标之间，高净值人士需要根据自己的风险承受能力，投资目标，以及市场环境等因素来做出决策。

1. 保值增值：如果高净值人士的主要目标是保值增值，那么他们可能需要将更多的资产分配到相对安全的投资中，如债券，保险，或者某些类型的基金。这些投资通常能够提供稳定的回报，同时降低投资风险。

2. 继续扩大资产：如果高净值人士的主要目标是继续扩大资产，那么他们可能需要将更多的资产分配到风险较高但有可能带来更高回报的投资中，如股票，房地产，或者某些类型的创业投资基金。这些投资可能会带来更高的回报，但也可能会带来更高的风险。

在做出决策时，高净值人士还需要考虑到他们的投资期限，税务规策，以及他们的个人偏好等因素。</s>

#### Openai回答

开启特征消融前：

Human: 高净值人士是否应该更多地考虑遗产规划？还是应该优先考虑继续扩大资产？

Assistant: 高净值人士确实应该同时考虑遗产规划和资产扩大两个方面，以实现财富的长期保值和增值。

遗产规划是指通过制定合理的财务计划和安排，确保个人或家庭的财富能够在未来得到妥善管理和传承。随着个人财富的积累，遗产规划变得越来越重要，以下是几个原因：

避免遗产纠纷：遗产规划可以帮助高净值人士避免在财富传承过程中出现的争议和纠纷。通过明确财产继承人和分配方式，可以减少家庭成员之间的矛盾和冲突。

降低税收负担：合理的遗产规划可以帮助高净值人士降低遗产税等税收负担。通过提前规划和安排，可以选择最优的税收筹划方案，降低财富传承过程中的税收成本。

保障家庭生活：遗产规划不仅仅是关于财富的传承，更重要的是关乎家庭生活的保障。通过为家庭成员提供充足的生活保障和资金支持，可以确保他们在未来的生活中得到足够的支持和关爱。

而资产扩大则是指通过投资、创业等方式增加个人或家庭的财富。资产扩大对于高净值人士来说同样重要，原因如下：

实现财富增值：通过投资、创业等方式，高净值人士可以将资产转化为更多的财富，实现财富的增值。这不仅可以提高个人或家庭的生活水平，还可以为后代留下更多的遗产。

增强财务安全：资产扩大可以增强个人或家庭的财务安全。通过多元化投资和分散风险，可以降低财富受到市场波动和经济周期影响的风险。

实现财富传承：资产扩大不仅可以增加个人或家庭的财富，还可以为后代提供更多的机会和资源，帮助他们实现财富传承和发展。

综上所述，高净值人士应该同时考虑遗产规划和资产扩大两个方面，以实现财富的长期保值和增值。在制定财务计划和安排时，应该根据个人或家庭的实际情况和需求进行综合考虑，选择最适合自己的方案。

开启特征消融之后：

Human: 高净值人士是否应该更多地考虑遗产规划？还是应该优先考虑继续扩大资产？

Assistant: 在理财规划中，高净值人士通常面临两个主要目标：保值增值和继续扩大资产。

保值增值：这主要涉及到资产的安全和长期增长。对于高净值人士来说，保值增值通常比单纯的资产扩大更为重要。这是因为他们的财富已经相当可观，因此他们更需要确保财富能够持续稳定地增长，以抵御通胀和其他潜在的风险。

继续扩大资产：这主要涉及到财富的积累和扩张。对于高净值人士来说，他们可能希望通过投资新的项目或者业务，以实现资产的进一步扩大。

在这两个目标之间，高净值人士需要根据自己的风险承受能力，投资目标，以及市场环境等因素来做出决策。

保值增值：如果高净值人士的主要目标是保值增值，那么他们可能需要将更多的资产分配到相对安全的投资中，如债券、保险，或者某些类型的基金。这些投资通常能够提供稳定的回报，同时降低投资风险。

继续扩大资产：如果高净值人士的主要目标是继续扩大资产，那么他们可能需要将更多的资产分配到风险较高但有可能带来更高回报的投资中，如股票、房地产，或者某些类型的创业投资基金。这些投资可能会带来更高的回报，但也可能会带来更高的风险。

在做出决策时，高净值人士还需要考虑到他们的投资期限、税务规划，以及他们的个人偏好等因素。

高亮对比总结:

特征消融前：回答强调了遗产规划，并详细讨论了其重要性及如何实施。同时也提到资产扩大的必要性，提供了一个全面的建议。

特征消融后：遗产规划的内容完全被删除，回答聚焦于保值增值和资产扩大两个目标，并建议高净值人士根据个人情况进行选择，但没有再提及遗产规划的相关内容。

# 参考文献

1. Bricken T, Templeton A, Batson J, et al. Towards monosemanticity: Decomposing language models with dictionary learning[J]. Transformer Circuits Thread, 2023, 2.
2. Templeton A. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet[M]. Anthropic, 2024.
3. Gao L, la Tour T D, Tillman H, et al. Scaling and evaluating sparse autoencoders[J]. arXiv preprint arXiv:2406.04093, 2024.