摘要

# 第一章 绪论

## 1.1 研究背景与意义

大语言模型（Large Language Model, LLM）是基于深度学习的自然语言处理模型，通常通过预训练大量文本数据来学习语言的语法、语义及上下文关系。其核心是使用神经网络（如Transformer架构）来捕捉复杂的语言模式，并生成或理解人类语言。这些模型能够执行各种自然语言处理任务，如文本生成、翻译、问答、情感分析等。随着人工智能技术的不断发展，特别是在自然语言处理（NLP）领域的突破，大语言模型（LLMs）已经成为理解和处理金融领域复杂文本数据的重要工具。金融行业涉及广泛的文本数据，包括新闻报道、财务报告、市场分析、法规文件等，这些数据具有高专业性和复杂性。大语言模型的出现使得自动化处理和分析这些数据成为可能，大模型语言的应用极大地提高了信息处理的效率与准确性，推动了智能投顾、风险控制、信用评分等金融创新应用的发展。

在人工智能领域，特别是在涉及到复杂的大型模型如GPT系列时，面临着一个独特的挑战：这些模型的内部工作机制常常是一个“黑盒”。虽然这些模型的代码、参数和训练方法是公开透明的，但研究者仍然难以追踪和理解它们如何从特定的输入生成特定的输出。这种复杂性源自于模型内部多层次的非线性数据处理和海量参数的复杂交互，使得追溯和理解其决策路径成为一项巨大挑战。然而，金融行业对模型结果的可解释性提出了更高的要求。与其他领域不同，金融决策往往涉及高风险，必须考虑模型输出背后的因果关系和逻辑合理性。监管机构也要求金融机构能够解释其模型的决策过程，以确保符合相关法律法规。因此，增强金融大语言模型的可解释性，能够帮助金融从业者理解模型的决策依据，提升对模型结果的信任度，并最终促进人工智能技术在金融领域的安全应用。

## 1.2 研究现状

稀疏自动编码器（Sparse Autoencoder, SAE）在大模型可解释性研究中的应用已经逐渐成为重要的方向。Anthropic大模型可解释性团队提出，神经网络最自然的计算单元——神经元本身大多数是多义的，即它们对看似不相关的输入的混合做出反应。多义性的一个潜在原因是叠加现象，即神经网络通过为每个特征分配自己的神经元线性组合来表示数据的独立“特征”，而不是其神经元。 因此，该团队提出通过将模型内部的高维表示解构为更加稀疏且易解释的特征，研究者能够从复杂的语言模型中提取出单义性的（monosemantic）特征。该研究通过将SAE被用于分解transformer中多层感知器（MLP）的激活，成功将复杂的激活模式解构为具有明确含义的特征。这种字典学习方法为研究者提供了一个理解大模型内部行为的新途径​ 。之后，Anthropic团队进一步在更大规模的模型Claude 3 Sonnet中验证了SAE的可扩展性。通过SAE模型，研究者能够提取出更高级的、抽象化的单义特征，这不仅增强了对模型决策过程的透明度，还为大规模语言模型的解释性研究提供了新的可能性。该研究表明，SAE在提取稀有特征和揭示模型内部逻辑方面具有极高的潜力​。同时，2024年6月，openAI团队着重探讨了SAE在概念提取和模型评估中的作用，不仅验证了SAE在不同模型规模上的适用性，还通过实验评估其解释能力，确保在处理更复杂的数据和任务时依然有效。 稀疏自动编码器的这种扩展应用，正在为提升大语言模型的解释性奠定坚实的理论基础​。这些研究证明了稀疏自动编码器逐渐被证明是一种强大的工具，它能够将复杂的大语言模型中的高维信息解构为更容易理解和解释的特征。这种特性使得SAE在金融、法律等对透明度要求高的领域中具有广阔的应用前景。

## 1.3 研究内容

针对前人所做的采用稀疏自编码器在通用大模型上进行可解释性的相关工作，本研究旨在将SAE应用于金融大语言模型中，通过对其中的高维信息进行稀疏性处理，将其解构为人类可以理解和解释的特征。本研究的主要研究内容如下：

（1）探究基于SAE的XAI技术在金融行业LLM中的应用；

（2）通过训练好的SAE，筛选并研究金融相关特征在模型中的激活行为；

（3）完成金融相关特征的可视化工作；

（4）通过特定特征实现对模型行为的有效引导。

## 1.4 技术路线

本研究旨在通过SAE对金融大模型进行可解释性分析，提取与金融行业密切相关的关键特征，并对模型的决策过程进行有效引导。技术路线如图1所示。

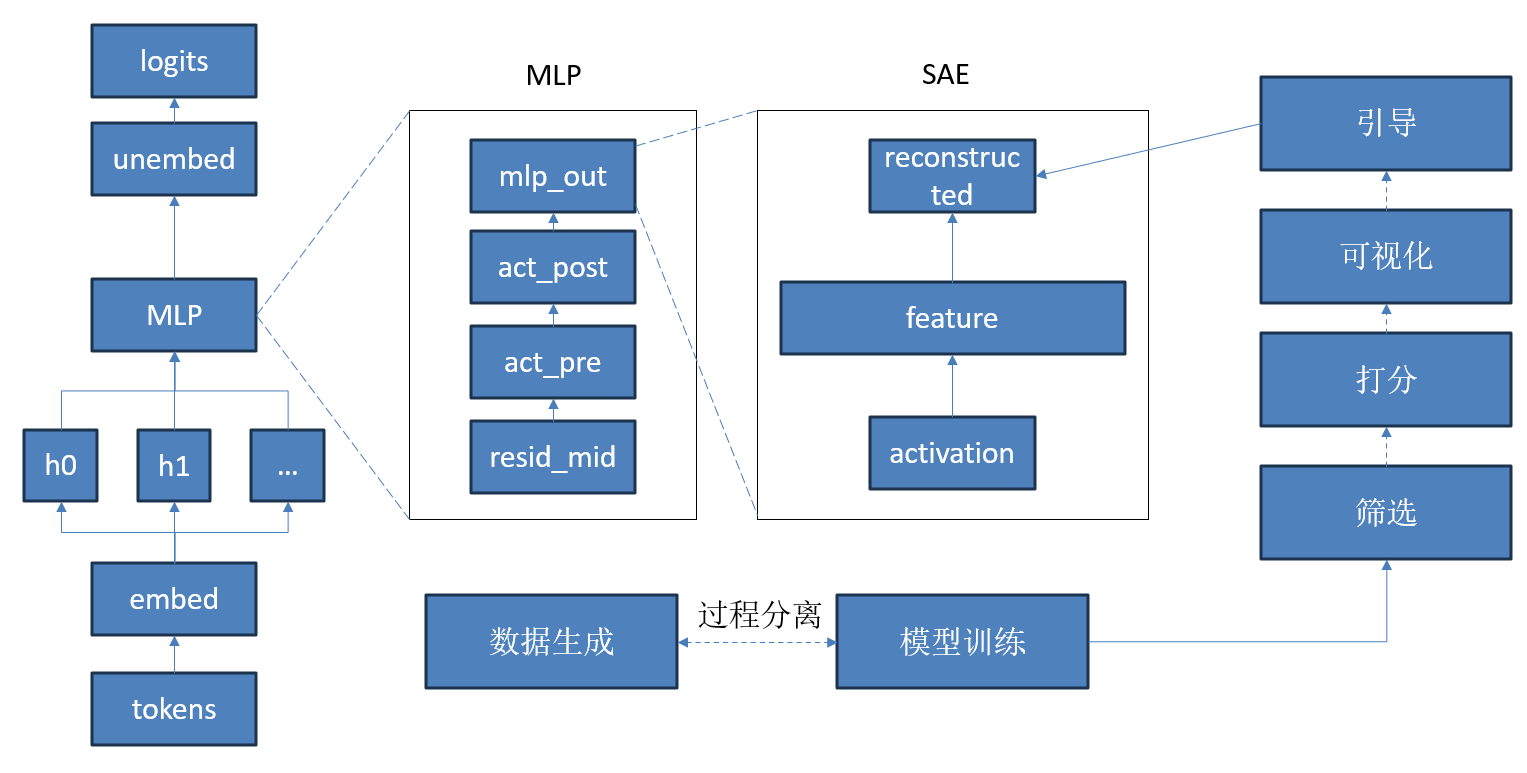


图1 技术路线图

# 第二章 技术概述

## 2.1金融大模型

**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。

**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。最后选择了轩辕大模型。

## 2.2金融语料库

**东方财富**

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

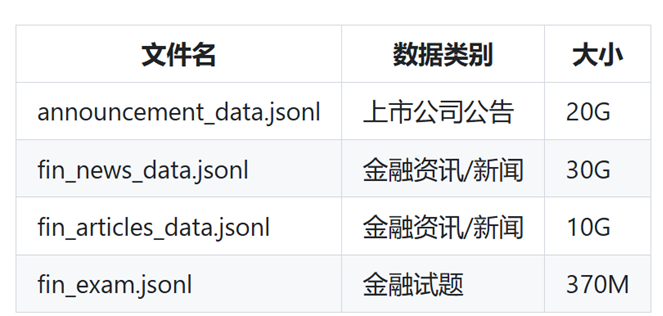
OpenFinData是由东方财富与上海人工智能实验室联合发布的开源金融评测数据集。该数据集代表了最真实的产业场景需求，是目前场景最全、专业性最深的金融评测数据集。它基于东方财富实际金融业务的多样化丰富场景，旨在为金融科技领域的研究者和开发者提供一个高质量的数据资源。

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**

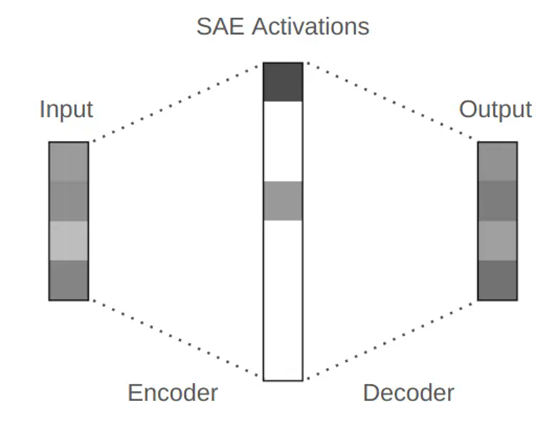
[Duxiaoman-DI/FinCorpus · Datasets at Hugging Face](https://huggingface.co/datasets/Duxiaoman-DI/FinCorpus)

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



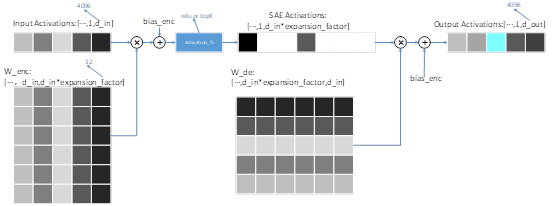
## 2.3 稀疏自编码器

LLM 正变得越来越强大、越来越有用，但它们仍旧是黑箱，我们并不理解它们完成任务的方式。SAE 可帮助我们将模型的计算分解成可以理解的组件。在LLM可解释上，稀疏自动编码器(SAE)作为一种弱字典学习算法用于可解释的研究。神经网络中单个神经元并不对应于单个概念,语言模型中的任何概念都是神经元的组合表示。因此，将神经网络分解为可理解的组件，需要将特征进行稀疏表示。常规自动编码器通常将高维度的输入特征压缩为低维度的特征信息，然后重建其输入数据。为了促使SAE创建稀疏的中间向量，稀疏自动编码器SAE中间向量维度通常大于输入向量的维度，如图1所示，并在在SAE的训练损失中添加L1稀疏度惩罚。这样得到的SAE表征中每一个有效的激活值都会对应一个有意义的的表示信息。



**图1 稀疏自编码器示意图**

稀疏自编码器SAE工作原理如下图2所示，假设SAE的输入维度为d\_in，扩展因子为 k，则SAE的编码表征为d\_in\*k。第一个矩阵是形状为[d\_in, d\_in\*k]的编码器矩阵，第二个矩阵是形状为 [d\_in\*k, d\_in]的解码器矩阵。通过让 LLM的激活与编码器相乘并使用 ReLU，可以得到 d\_in\*k维的 SAE 编码的稀疏表征，因为 SAE 的损失函数会促使实现稀疏性。通过将 SAE 的表征与解码器相乘，可得到一个 d\_in维的重建的模型激活，但是并不能完全的重建结果与原始的LLM 激活，因为稀疏性约束条件会让重建结果难以复现。为了引入稀疏性，最直接的方法是向 SAE 的损失函数添加一个稀疏度惩罚项。对于这个惩罚项，最常见的计算方式是取这个 SAE 的已编码表征的 L1 损失并将其乘以一个 L1 系数。这个 L1 系数是 SAE 训练中的一个关键超参数，因为我们需要在实现稀疏度与维持重建准确度之间的权衡。



**图 2 稀疏自编码器的工作原理**

## **2.4自动化可解释性**

可解释性研究的一个简化方法是首先了解神经网络中单个组件（如神经元和注意力头）的功能。传统上，这要求人工手动检查神经元，以识别它们所表示的数据特征。然而，这个过程难以扩展到拥有数百亿或数千亿参数的神经网络。 2023年OpenAi在Language models can explain neurons in language models中提出将自动化应用于将可解释性技术扩展到大型语言模型中的所有神经元，并用GPT-4来对GPT-2中的神经元做出解释：先让 GPT-2 运行文本序列，等待某个特定神经元被频繁「激活」的情况。

让 解释器GPT-4 针对这段文本，生成解释。第二步，用 GPT-4 模拟这个 GPT-2 的神经元接下来会做什么。第三步，对比评估打分。对比 4 代模拟神经元和 2 代真实神经元的结果，看 GPT-4 猜的有多准。

本课题研究不直接用解释器来解释金融大模型而是对SAE得到的特征进行模拟、解释和打分。具体实现见4.2.2。

# 第三章 技术实现

## 3.1 数据生成

### 3.1.1 大模型测评

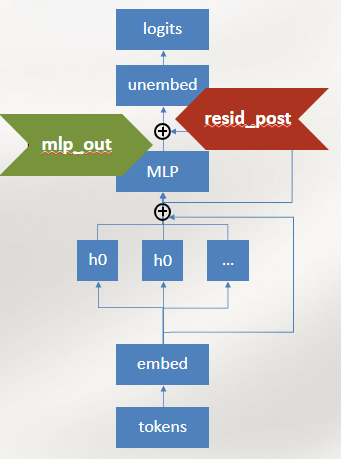
需要考虑进行大模型可解释性分析时，模型结构对显存占用的消耗。最后选择了轩辕大模型。

### 3.1.2 语料库选择

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。

### 3.1.3 激活值采样



对于两个具有代表性的层使用sae分析，先介绍一下这两个层：

mlp\_out层：解释出的语义叠加现象大​，死亡的特征数量多

resid\_post层：避免跨层叠加的现象​，减少死亡的特征

为了减少显存占用，我们选择先加载大模型采样activation 存在本地，再sae训练时直接用本地存储的activations。

**过程中遇到的问题：**

**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" \l "scrollTo=sMkmqMjkVdMq)

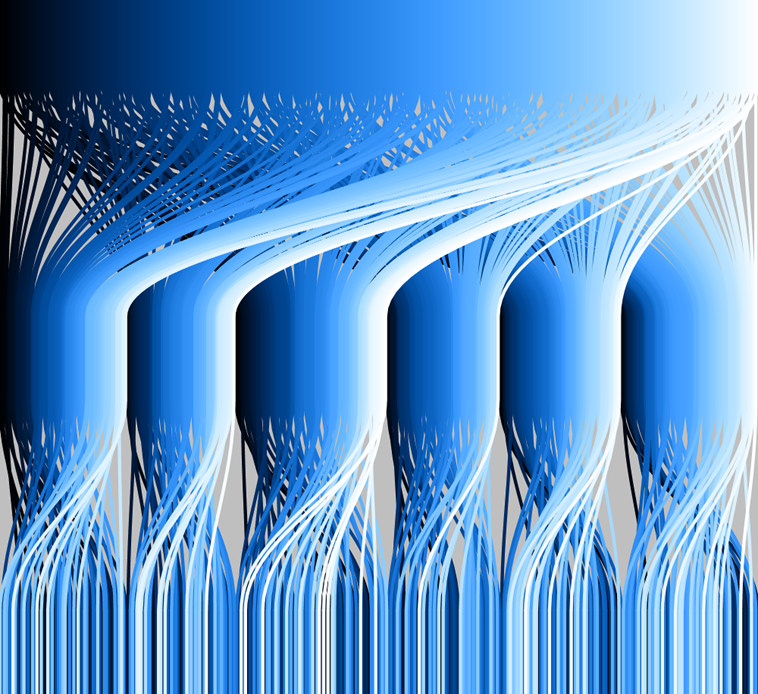
转为LLaMA的权重为HuggingFace的权重，并使用HookedTransformer来读取。

### 3.1.4 shuffle算法

对于小数据集来说：直接shuffle。然后对于大数据集来说，全部数据并不能一起放入到内存中。

我们的数据集有n个文件，每个文件m行。n=600，m=600000, 600G的数据。

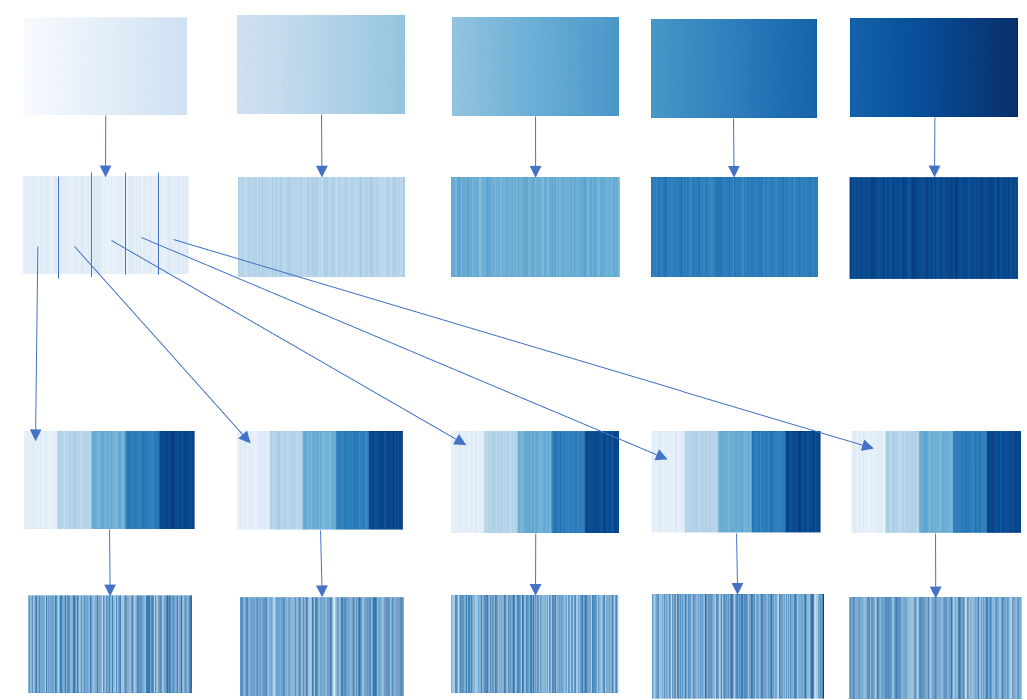
在最开始参考Anthropic的两次洗牌，第一次洗牌的时间复杂度为O(nm)，实际情况会使用40个小时来进行第一次洗牌。



Antropic的shuffle方法

针对洗牌复杂度高的问题，我们对该洗牌方法进行了一部分改进。首先，我们在最开始就有n个文件。我们首先对这n个文件进行了一个文件内的一个内部shuffle，然后再创建另外n个空白的文件。对原先的n个源文件进行一个切片，将他们的切片均匀的分到，另外n个空白文件中，这就是我们的第一次洗牌。在第一次洗牌之后，再按照原方法对每个文件再次进行一个文件内部的shuffle，最终我们达到了一个完全shuffle的效果。

经过我们的优化，第一次洗牌的时间复杂度变为了O(n\*2)，可以使用20分钟来进行第一次洗牌。



改进后的shuffle方法

## 3.2 SAE模型训练

### 3.2.1 显存分析

**1. SAE模型占用显存**



**SAE结构概图(**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, 下面表格为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 |

**2.数据shuffle占用显存**

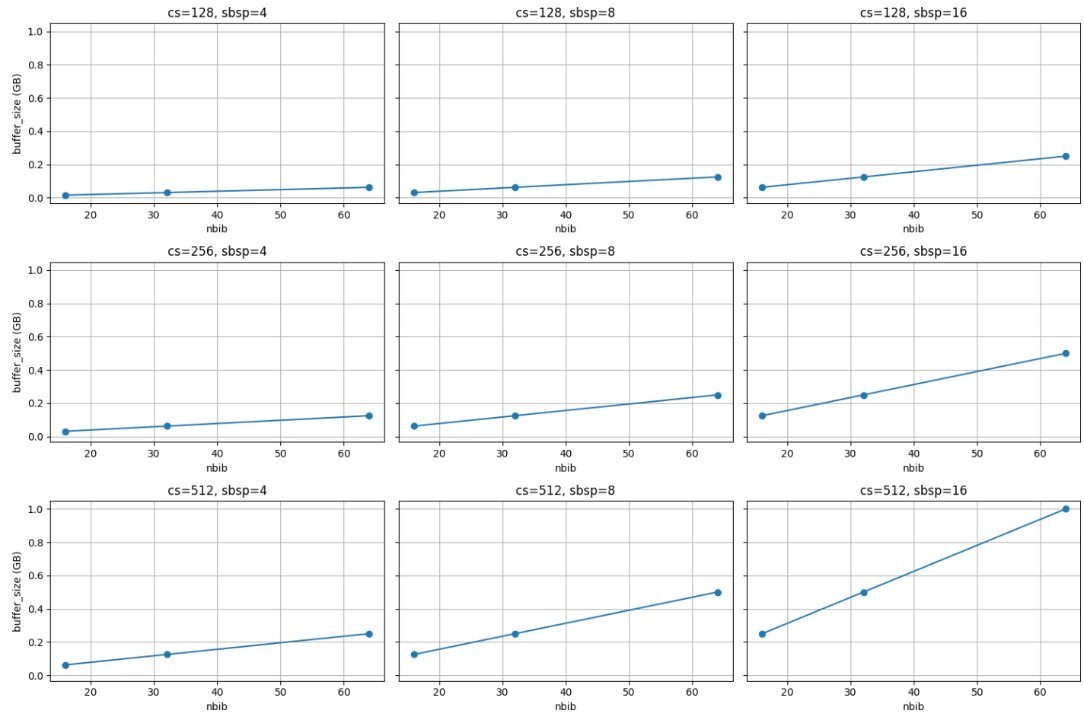
sae\_lens的shuffle机制，并不是对全部数据进行统一shuffle，而是分批载入的，首先两个buffer合成一个buffer, 进行shuffle后，再把一个buffer拆解成两个buffer.

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







注：Cs=上下文长度​，nbib=buffer中的batch数量​，sbsp=batch中prompt的数量​，

总的显存占用是19\*buffer\_size

1. **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 |

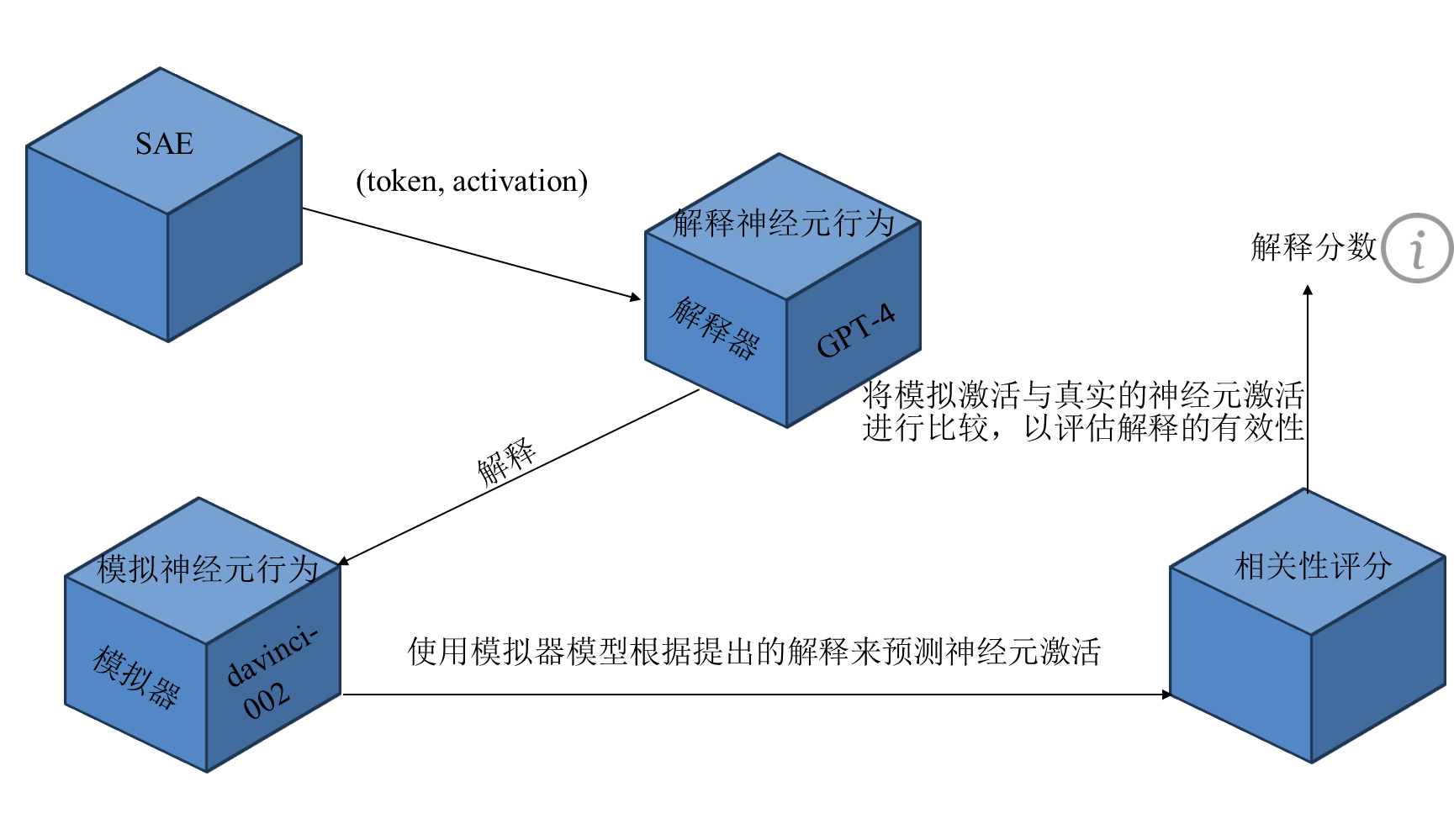
### 3.2.2 参数优化

|  |  |  |
| --- | --- | --- |
| 参数 | 含义 | 参数取值 |
| model\_name | 加载的大模型 |  |
| hook\_name | 钩子点设定在模型的 MLP的位置，在该位置挂载SAE | |
| hook\_layer |  |  |
| activation\_fn | 激活函数 | topK,relu |
| 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 | 扩展因子的大小，更大的扩展因子，有助于网络在高维特征空间中找到更稀疏的表示 | 16，32，64 |
| d\_sae | 隐藏层神经元的数量 | expansion\_factor \* d\_in |
| l1\_coefficient | L1正则化项，施加正则化项控制SAE 的稀疏性 | [5e-4,1e-2] |
| context\_size | 训练SAE时给模型的提示长度 | 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：使用解码器的几何均值初始化解码器的偏置 | **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 | 不进行归一化的话，会导致SAE输出值太小，影响可视化部分的logit值 |  |
| 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,对dead features进行梯度更新，可以恢复一部分dead feature, 减少特征的死亡数量 | |
| feature\_sampling\_window | 窗口大小的设置影响特征的稀疏性 | |
| dead\_feature\_window | 该窗口的大小影响死亡特征的数量 | |
| dead\_feature\_threshold | dead feature的阈值 | |

## 3.3自动解释分数

训练完成SAE后，我们需要一个指标来衡量特征的可解释性。我们使用openai团队的自动化方法。

自动解释过程会在特征激活的地方提取文本样本，要求语言模型编写字典特征的人类可读的解释，然后提示语言模型使用此描述来预测特征在其他特征上的激活。样本模型的预测激活和实际激活之间的相关性就是该特征的可解释性得分。



如上图所示，自动解释分数包含以下四个步骤：

· 选择文本片段并计算激活值：在Fincorpus数据集中，从前50,000行中每行取一个包含64个token的句子片段，并测量该特征（或神经元）在每个token上的激活值。激活值被重新缩放为0到10之间的整数。

· 生成解释：从激活得分最高的20个片段中选择5个，并将这些片段以及重新缩放后的每个token的激活值传递给GPT-4，指示GPT-4生成一个解释，说明该特征（或神经元）在何种情况下会激活，从而得出一个解释。

· 模拟特征激活：使用GPT-3.5对另外5个高度激活的片段和5个随机选择的片段（具有非零变化）进行模拟，要求GPT-3.5提供每个token的激活值。

· 计算自动解释分数：计算模拟激活值与实际激活值之间的相关性，这个相关性即为特征的自动解释评分。

# 结果分析

## 4.1 SAE模型评估

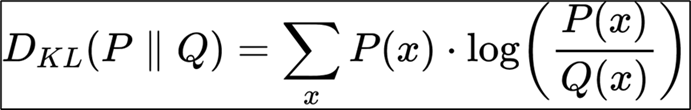
### 4.1.1 评估指标

训练 SAE 的一大主要难题是评估。我们可以训练稀疏自编码器来解释语言模型，但我们没有自然语言表示的可度量的底层 ground truth。目前而言，评估都很主观，基本也就是「我们研究一系列特征的激活输入，然后凭直觉阐述这些特征的可解释性。」这是可解释性领域的主要限制。具体来说，SAE存在稀疏度和重建质量两方面的评估指标。

#### 重建质量评估指标

1. KL散度（Kullback-Leibler Divergence）

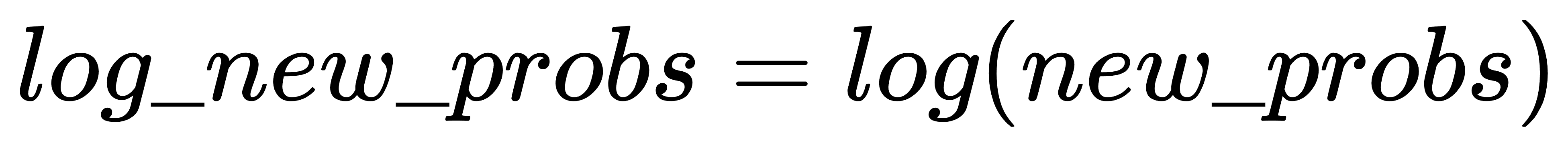
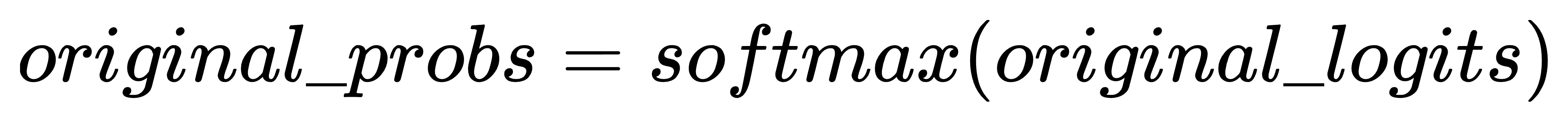
KL散度也称为相对熵，是用来衡量两个概率分布之间差异的非对称性度量。KL散度通常用来比较两个分布之间的差异。公式如下：



实验中，我们使用KL散度计算过程如下：

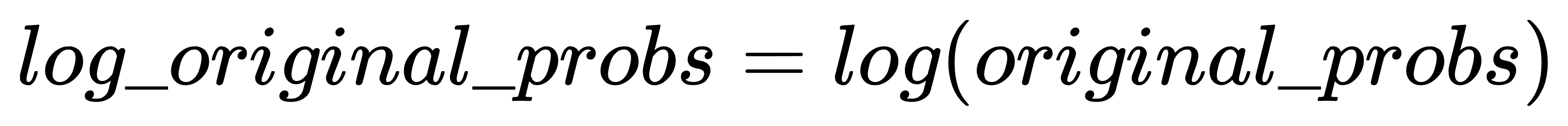
**Softmax变换**：

对原始logits和重构logits进行softmax变换，得到原始概率分布original\_probs和重构概率分布new\_probs。



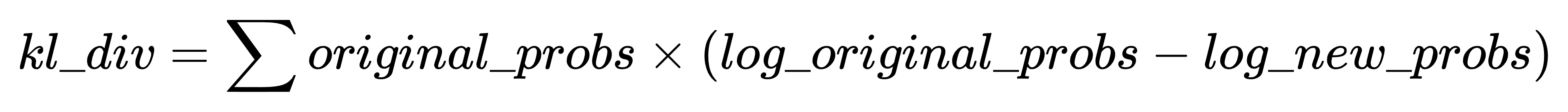
**对数概率**：

计算原始概率分布和重构概率分布的对数概率log\_original\_probs和log\_new\_probs。



**KL散度计算**：

计算原始分布和重构分布之间的KL散度。

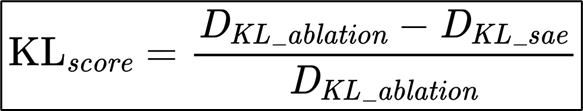


KL散度衡量的是原始分布与重构分布之间的差异。这个差异越小，意味着两者越相似，重构的分布尽可能接近原始分布。

1. **KL散度得分**

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, 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。公式如下：



1. 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。公式如下：



**原始ce\_loss**：

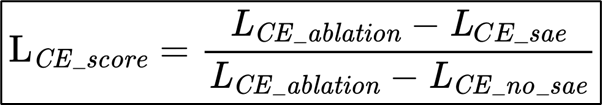
通过大模型在forward过程中，得到的交叉熵损失。

**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**：

计算得出的得分表示在有无 SAE 的情况下，相对于消融实验的损失变化情况，用于评估 SAE 对模型性能的影响。公式如下：



1. L2范数

L2范数由以下公式得到：



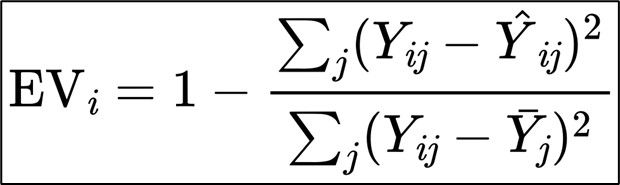
其中，表示第i个样本的输入激活向量。

另外，我们还通过计算输入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

**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。分子表示预测误差的平方和（残差平方和），即模型未能解释的方差。分母表示总方差，即数据的总波动性。解释方差的值越接近1，表示模型的预测性能越好，能解释的数据方差越大。

**均方误差:**

均方方差由以下公式得到。

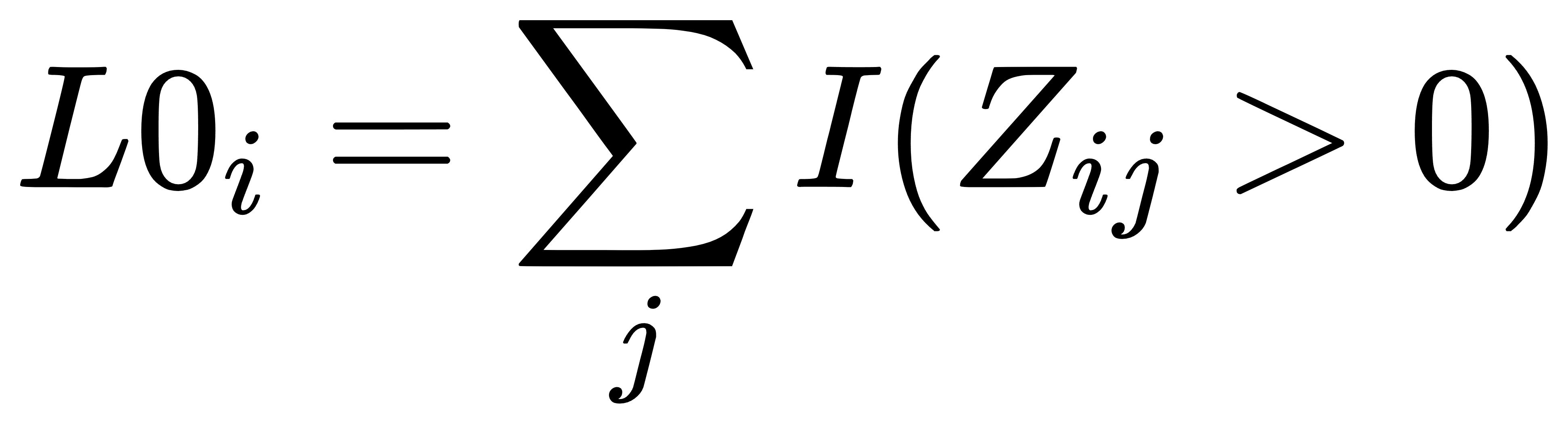


其中，表示第i个样本的第j个输入激活，表示第i个样本的第j个输出激活。

#### 稀疏度评估指标

1. L0范数

L0范数用于表示非零特征的个数，公式如下：



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。

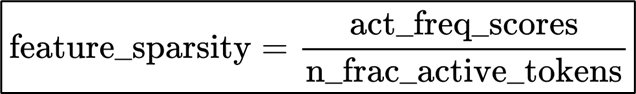
1. L1范数

L1范数表示特征激活的绝对值之和，公式如下：



其中，表示第i个样本的第j个特征激活。

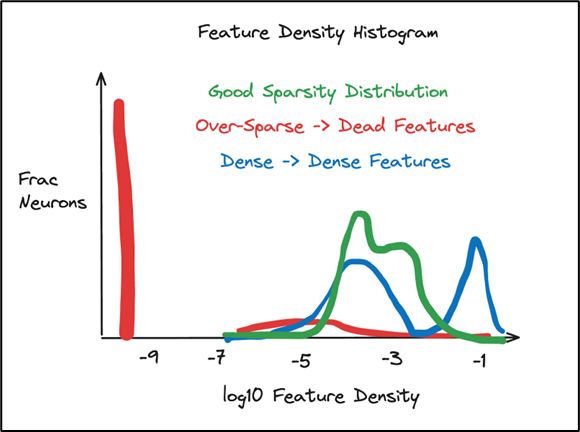
1. 特征稀疏度表示

这里的特征稀疏度表示为活跃特征的激活值总和/目前训练的token数量，公式表示如下：

1. mean\_log10\_feature\_sparsity

特征稀疏度对数表示，公式如下：  


1. feature\_density\_line\_chart



**特征密度对数表示直方图**

特征密度对数表示直方图我们可以从以下两点衡量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。

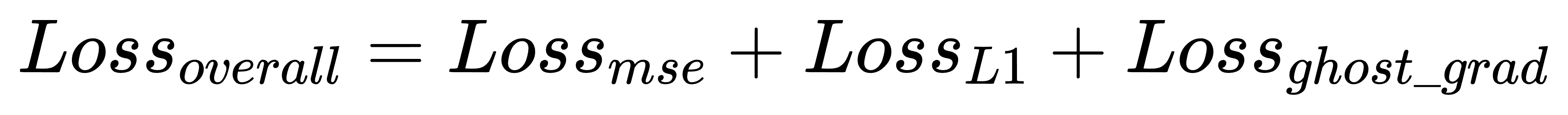
2）位于非超低密度区域特征的数量应该远大于超低密度区域特征的数量，特征密度分布图应该类似于拉普拉斯分布。

1. 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。

1. **损失函数**

在SAE的训练过程中共用到了三种损失函数，分别是交叉熵损失，均方误差损失，虚拟梯度损失，其中，使用虚拟梯度损失的是为了减少特征的死亡数量，防止特征的“假死亡”对dead latents的梯度进行更新。总损失函数的公式如下：



### 4.1.2 shuffle评估

在数据处理时，我们采取了shuffle的方式，为了证明我们所采取策略的有效性，对于shuflle和未shuffle的方式，我们分别对模型进行了训练，实验结果如下所示：

**重建指标实验结果**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **不同shuffle**  **方式对比** | **KL** | **KL\_score** | **CE\_Loss** | **CE\_Loss**  **\_score** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **shuffle** | **0.353** | **0.96** | **2.17** | **0.969** | **0.006** | **0.996** | **0.0023** |
| **未shuffle** | **3.34** | **0.625** | **5.05** | **0.638** | **0.002** | **0.991** | **0.007** |

**稀疏度指标实验结果**

|  |  |  |
| --- | --- | --- |
| **不同shuffle**  **方式对比** | **Log10\_feature\_sparsity** | **dead\_features** |
| **shuffle** | **-5.71** | **32825** |
| **未shuffle** | **-5.82** | **17217** |

### 4.1.3 Hook层评估

Hook层的选择也是导致最后所训练SAE好坏的一个重要因素，为此，我们分别选择了mlp\_out和hook\_resid\_post两种不同的hook方式，训练后发现采用hook\_resid\_post的方式效果更好一些，实验结果如下所示：

**重建指标实验结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **hook\_name** | **KL** | **CE\_Loss** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **hook\_resid\_post** | **0.353** | **2.22** | **0.0057** | **0.996** | **0.002** |
| **hook\_mlp\_out** | **0.711** | **2.548** | **0.0027** | **0.989** | **0.010** |

**稀疏度指标实验结果**

|  |  |  |
| --- | --- | --- |
| **hook\_name** | **Log10\_feature\_sparsity** | **dead\_features** |
| **hook\_resid\_post** | **-5.70** | **32825** |
| **hook\_mlp\_out** | **-7.77** | **44624** |

### 4.1.4 扩展因子的大小评估

扩展因子的设置是影响SAE稀疏度的一个重要指标，只有在稀疏度足够大时，每个中间向量的激活值代表的语义信息才可以足够的单一，因此扩展因子对SAE用于LLM可解释性上有关键的影响，但是由于计算资源的限制，我们目前无法尝试更大扩展因子，探索对SAE用于可解释性能影响。在当前的可用资源下，我们分别设置扩展因子大小为16和32，结果如下表所示，扩展因子为32的实验效果明显的好于扩展因子为16的实验效果。

**重建指标实验结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **expansion\_factor** | **KL** | **CE\_Loss** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **16** | **0.711** | **2.548** | **0.0027** | **0.989** | **0.010** |
| **32** | **0.514** | **2.463** | **0.0035** | **0.980** | **0.014** |

**稀疏度指标实验结果**

|  |  |  |
| --- | --- | --- |
| **expansion\_factor** | **Log10\_feature\_sparsity** | **dead\_features** |
| **16** | **-7.77** | **44624** |
| **32** | **-6.07** | **19688** |

### 4.1.5 L1稀疏因子大小的评估

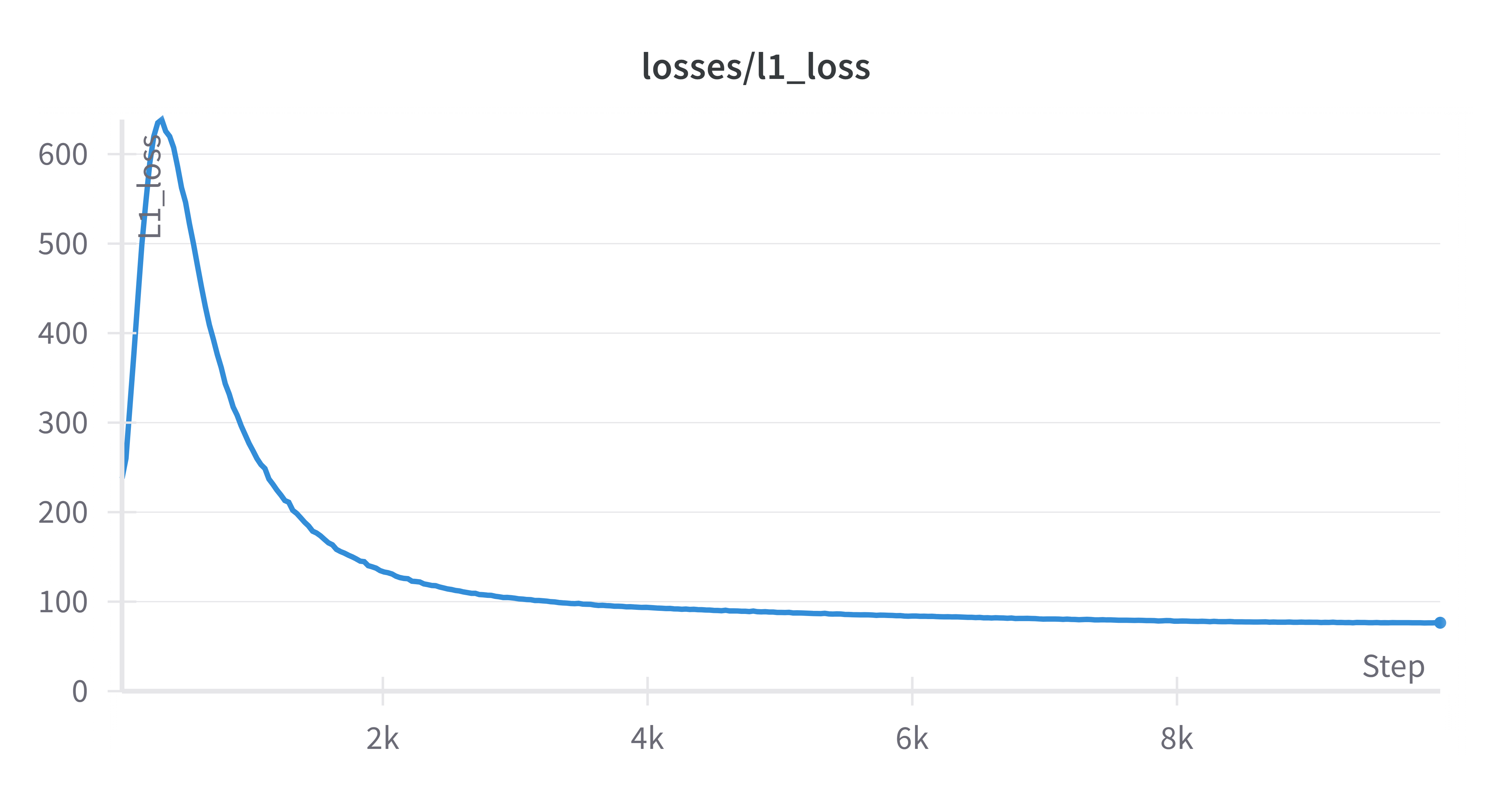
为了引入稀疏性，需要 SAE 的损失函数添加一个L1损失作为稀疏度惩罚项。L1损失对应的 L1 系数是 SAE 训练中的一个关键超参数，因为我们需要在实现稀疏度与维持重建准确度之间的权衡。

**重建指标实验结果**

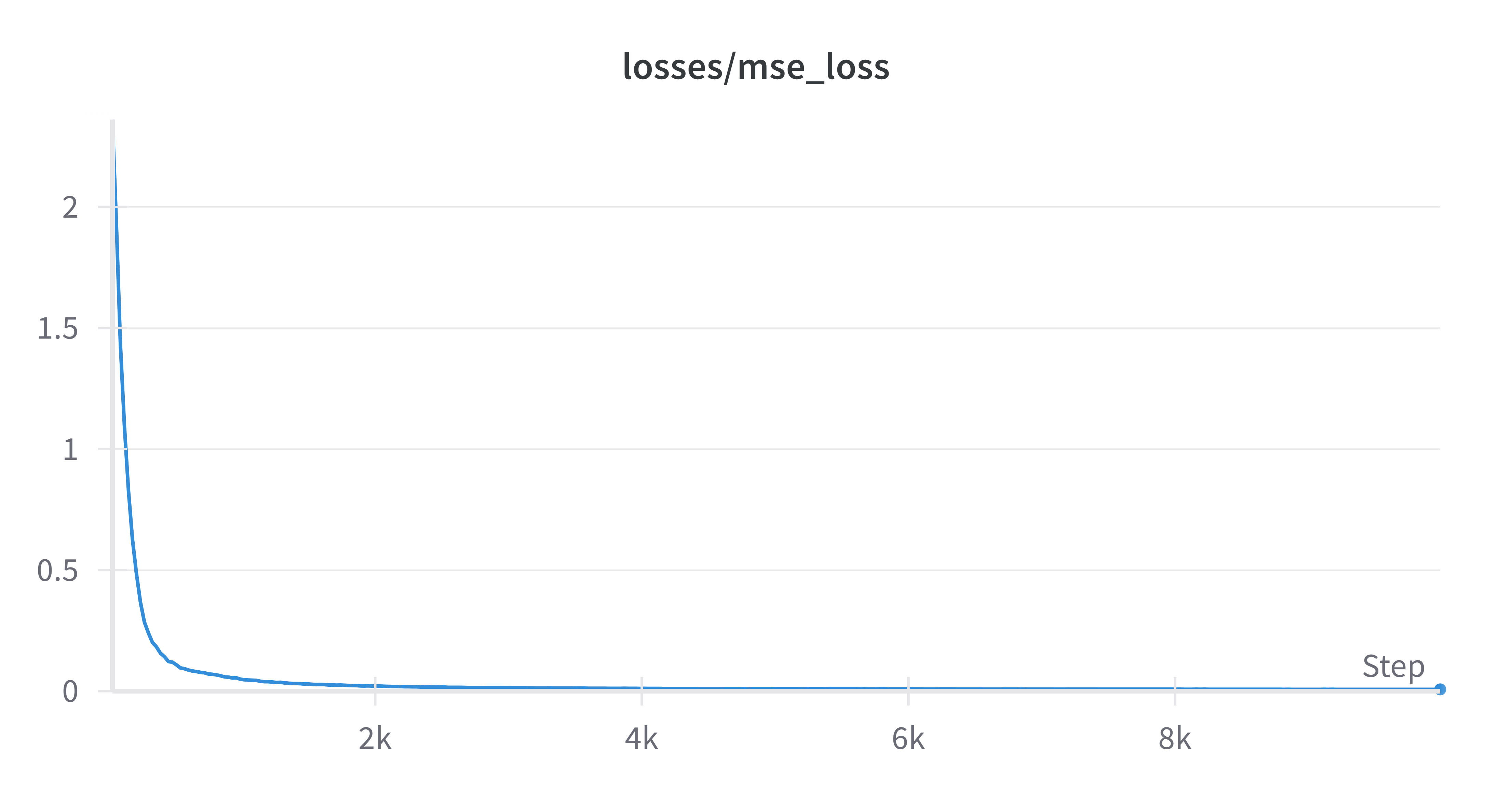
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **l1\_coefficient** | **KL** | **CE\_Loss** | **MSE\_Loss** | **Explained\_**  **variance** | **Explained\_**  **variance\_std** |
| **0.01** | **7.07** | **8.75** | **0.010** | **0.950** | **0.035** |
| **0.001** | **3.22** | **4.94** | **0.0035** | **0.980** | **0.014** |
| **0.0005** | **0.353** | **2.22** | **0.0057** | **0.996** | **0.002** |

### 4.1.6 SAE其它评估指标的结果可视化

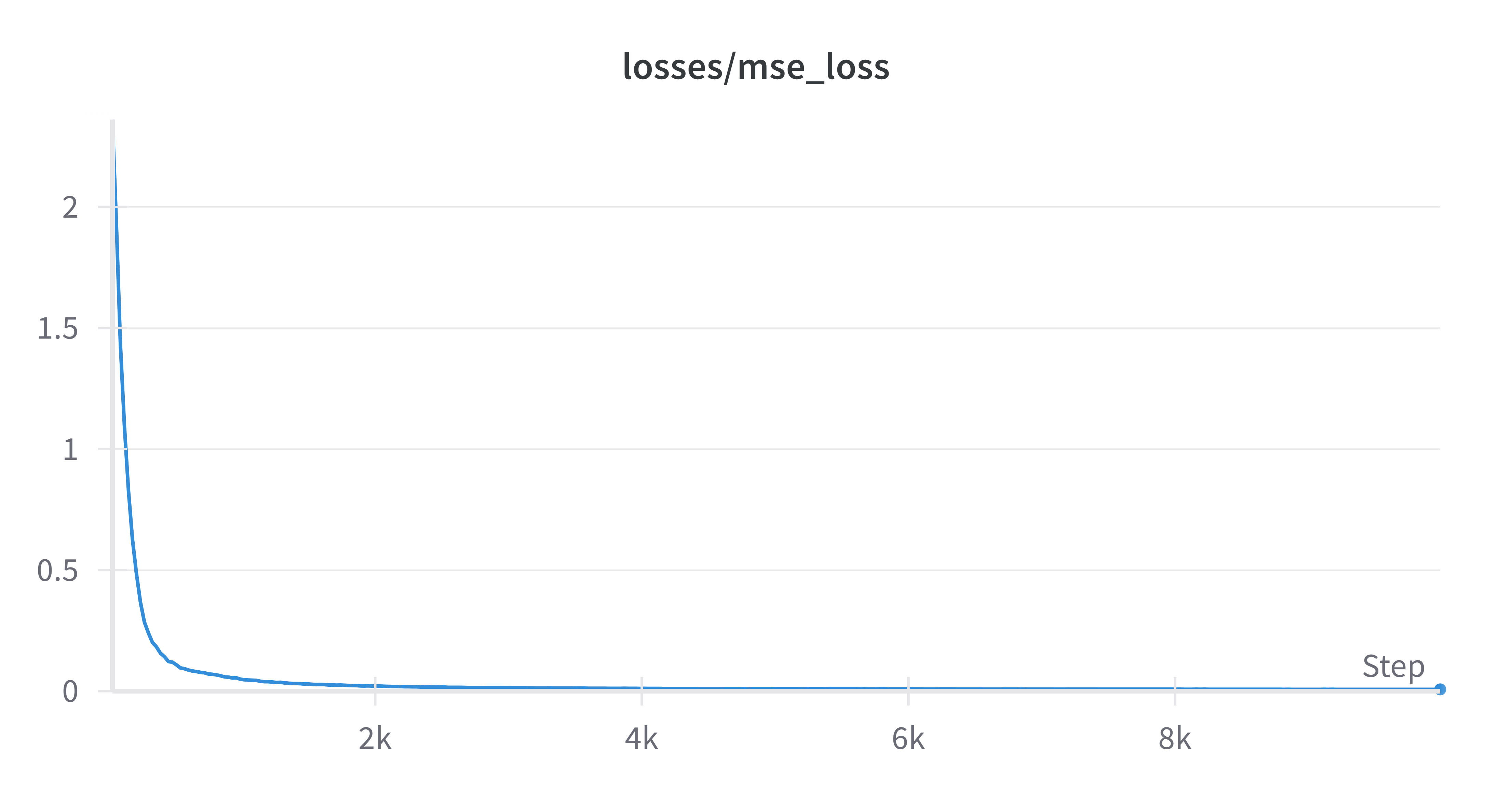
1. 训练过程中与损失函数有关的指标可视化结果：



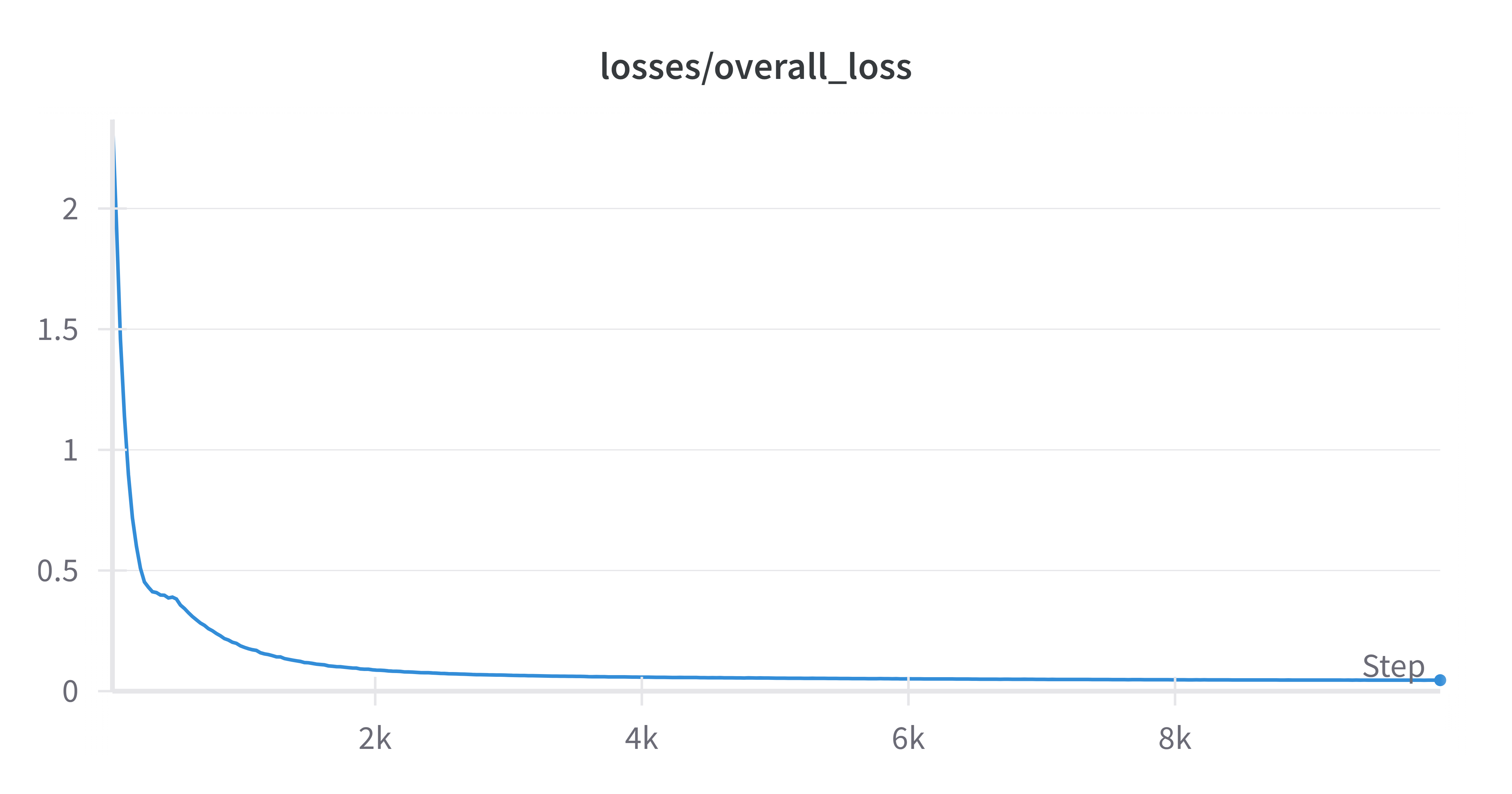
**L1损失训练趋势**



**MSE损失训练趋势**

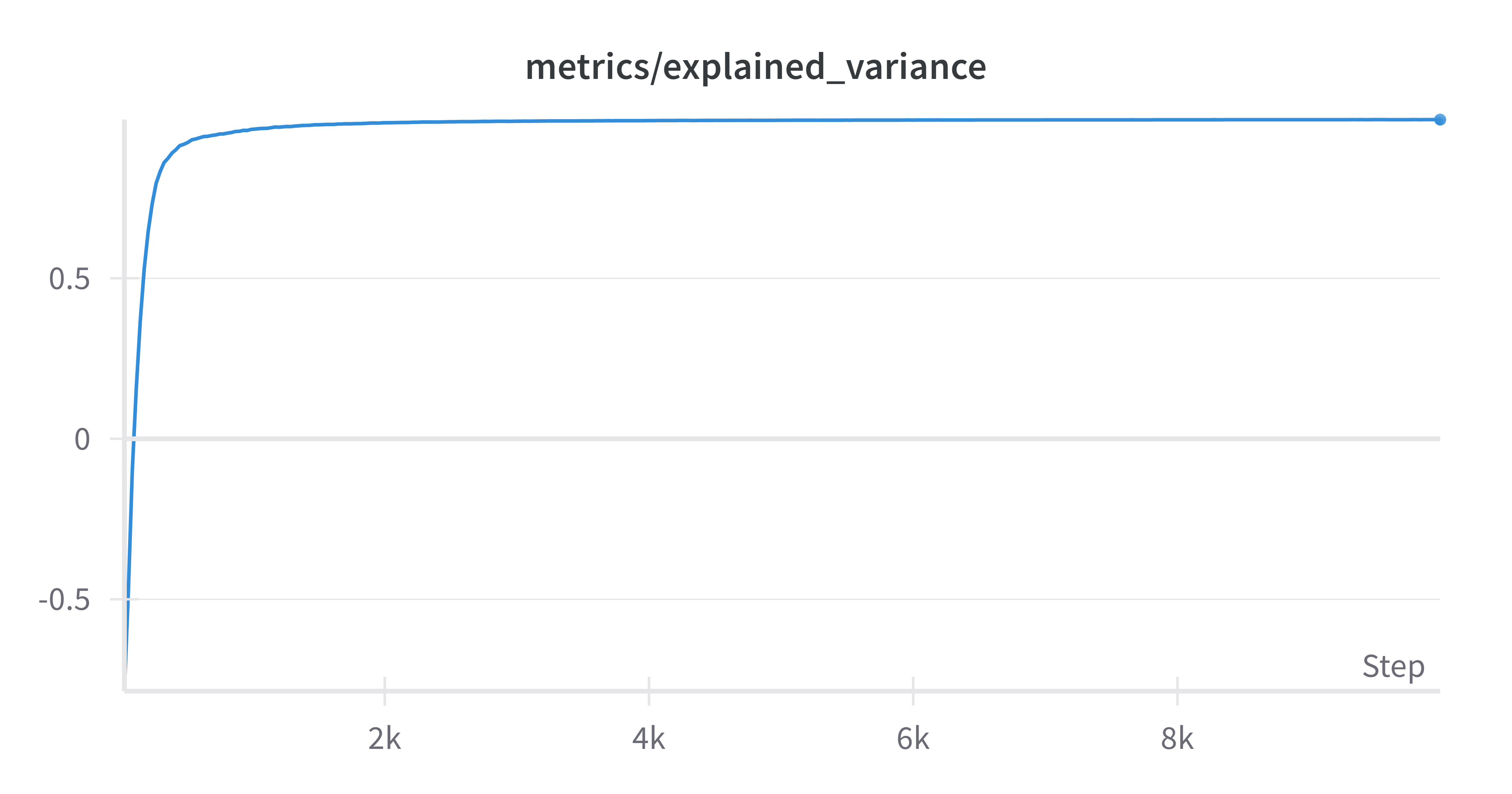
****

**虚拟梯度损失训练趋势**

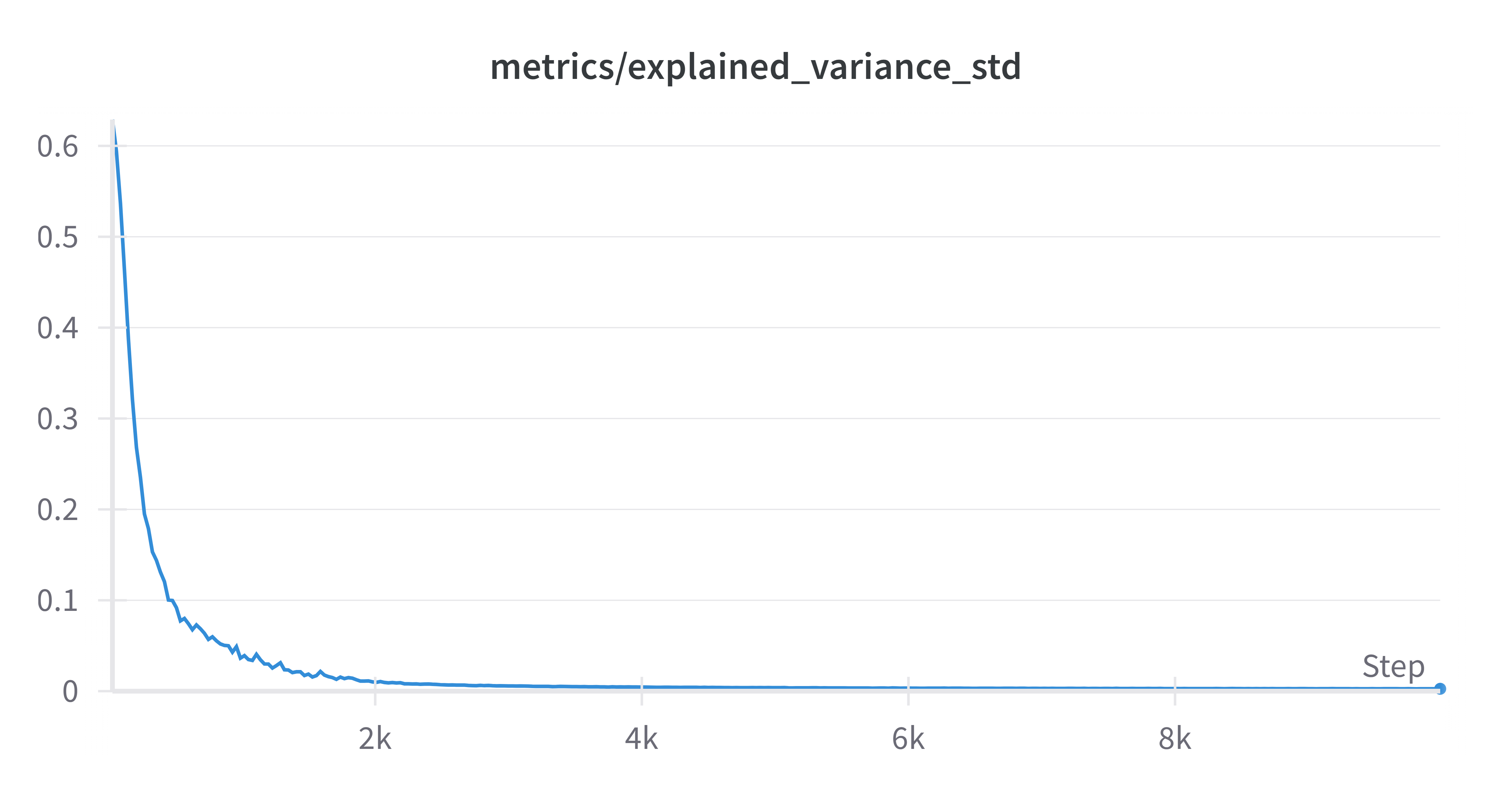
****

**总损失训练趋势**

1. 解释方差及其标准差可视化结果：

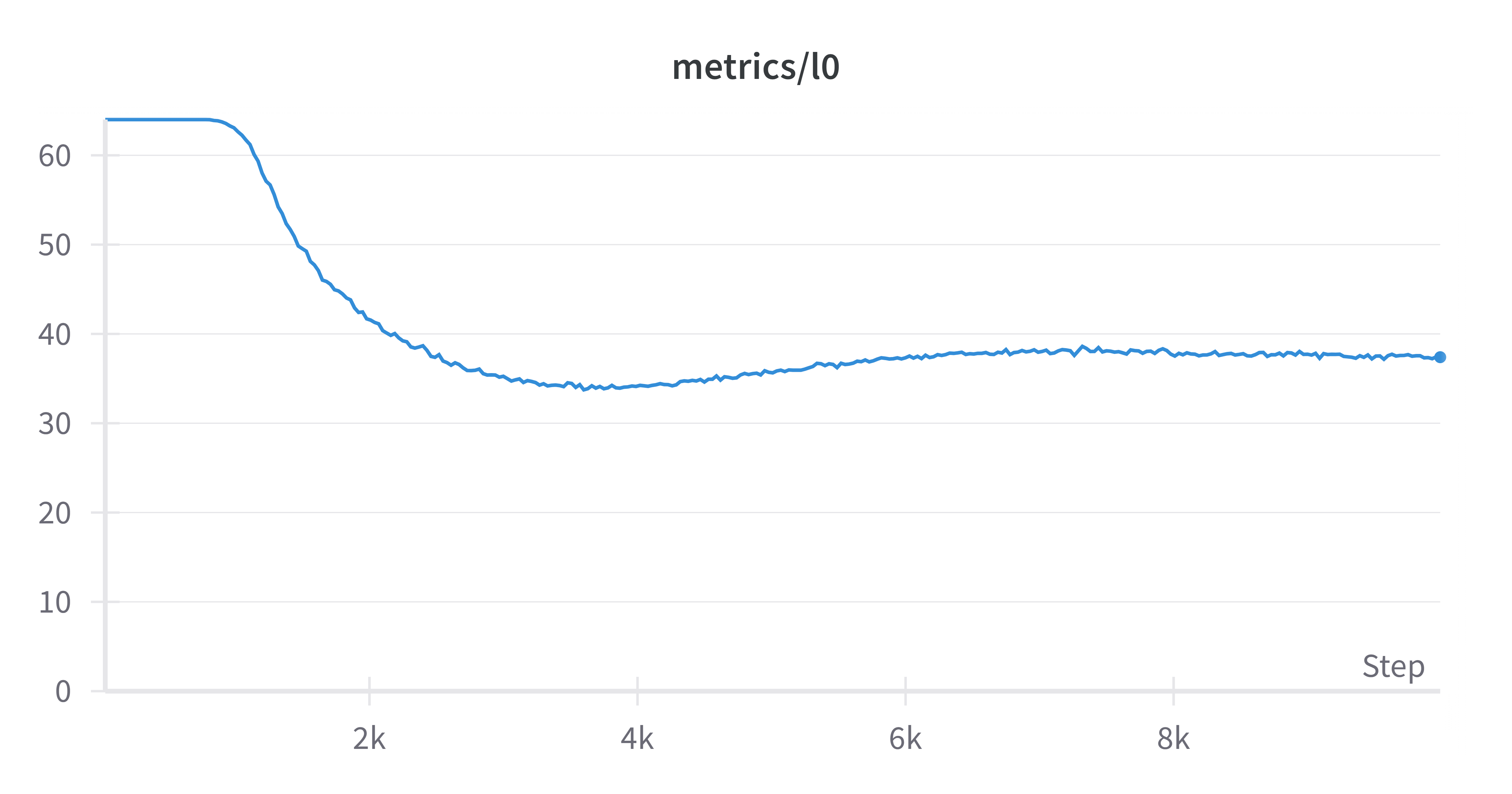
****

**解释方差训练趋势**

****

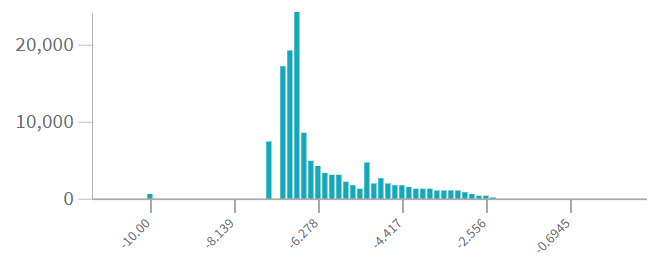
**解释方差的标准差训练趋势**

1. 稀疏度L0指标可视化结果：

****

**训练过程L0变化**

1. 特征密度对数表示直方图：

****

**特征密度直方图**

## 4.2 特征可视化

### 4.2.1 特征筛选

通过SAE对金融大模型的解构，我们得到了将近十几万特征，如果对十几万个特征都进行特征的可视化分析明显不现实。因此，在进行特征可视化之前，我们需要在这些特征中进行特征筛选，找到我们需要并可以解释的特征。本研究采用了以下方法对和单个或多个token相关特征进行筛选。

（1）对单个token相关特征筛选: 从给定的输入prompt 中提取特征的激活信息，输出那些对prompt的每个 token 激活最高的前32个特征的索引。

（2）对多个token相关特征筛选：采用两种方法。方法一是从给定的输入prompt 中提取特征的激活信息，对每个特征在多个连续 token 上的激活值进行聚合（例如取平均值）后，输出聚合后激活最高的前32个特征的索引。方法二是从给定的输入prompt 中提取特征的激活信息，找到那些对prompt的每个 token 激活最高的前32个特征的索引后，输出在连续多个token 上的特征的交集。

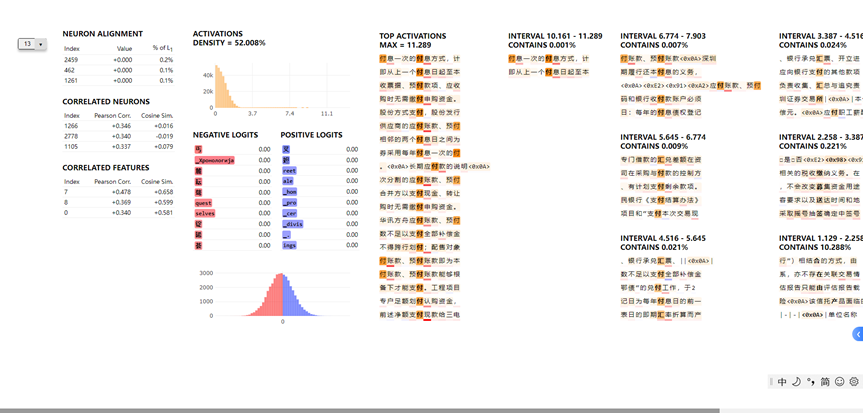
### 4.2.2 自动解释性分数

|  |  |  |
| --- | --- | --- |
| **特征编号​** | **解释​** | **得分​** |
| 1610​ | “司”字，通常用于公司名称中。​ | 0.44​ |
| 4919​ | “限”字，通常与有限责任公司有关。​ | 0.52​ |
| 6430​ | 与股票及金融市场交易相关的术语。​ | 0.15​ |
| 19488​ | 与地名和企业实体相关的汉字。​ | 0.13​ |

由于解释器和模拟器均为付费模型，受资源限制，本章节只展示了对四个特征的自动化解释和得分情况，如上表所示。自动化的解释不依赖于专家领域知识，因此可以与人为的经验解释相辅相成，共同助力更好地理解金融领域的大型语言模型。此外，自动化解释过程还具有扩展性，能够在更大规模的数据集和模型上进行应用。通过结合多种特征的自动化分析，我们可以更深入地挖掘模型中的隐藏模式和潜在偏差。这种方法不仅提升了对复杂模型的透明度，还为优化模型性能、提升决策准确性提供了新的途径。

### 4.2.3 以特征为中心

在特征的可视化方法上，第一种方式是采用以特征为中心的可视化。在以特征为中心的视图中，通过查看单个特征，可以看到该特征的相关指标，并查看此特征触发最强的序列。



以特征为中心的视图

如图所示，可以通过该视图了解到，该特征在汉字‘付‘上的激活值很高，相关上下文涉及到“预付“、”付息日“等相关金融信息。

相关指标以及含义如下：

**（1）NEURON ALIGNMENT**

表明特征与哪些神经元对齐

Index: 神经元的索引

Value: 神经元的激活值

% of L1: 该神经元的激活值占总激活值的百分比

**（2）CORRELATED NEURONS**

表明与该特征最相关的神经元

Index: 与该特征最相关的神经元的索引

Pearson Corr.: 神经元之间的皮尔逊相关系数，反映两个神经元激活值之间的线性关系

Cosine Sim.: 余弦相似度，衡量神经元激活值之间的相似度

**（3）CORRELATED FEATURES**

表明与该特征关联度最高的特征

Index: 与当前特征最相关的其他特征的索引。

Pearson Corr.: 皮尔逊相关系数，反映了两个特征之间的线性关系。

Cosine Sim.: 余弦相似度，衡量两个特征之间的相似度。

**（4）ACTIVATIONS DENSITY**

直方图表示当前神经元的激活值分布情况。横轴表示激活值，纵轴表示激活值出现的频率。

**（5）NEGATIVE LOGITS**

这部分展示了当某个特征激活时，某些token的概率降低最多。当该特征激活时，这些token的出现变得更不可能。"Negative Logits"部分列出了这些token以及它们对应的对数概率（logits），并展示了这些特征对降低这些token被采样的概率所起的作用。

**（6）POSITIVE LOGITS**

、这部分展示了当某个特征激活时，某些token的概率增加最多。当该特征激活时，这些token更有可能被采样。"Positive Logits"部分列出了这些token以及它们对应的对数概率，并展示了该特征对增加这些token被采样的概率的影响。

**（7）TOP ACTIVATIONS**

列出了当前神经元在不同输入（token或文本片段）上的最高激活值及其对应的输入内容。

MAX: 显示了该神经元的最大激活值。

橙色：最强激活；白色：无激活

Loss contribution：消融该特征后，模型对正确标记的预测变得不准确的程度的一个度量。

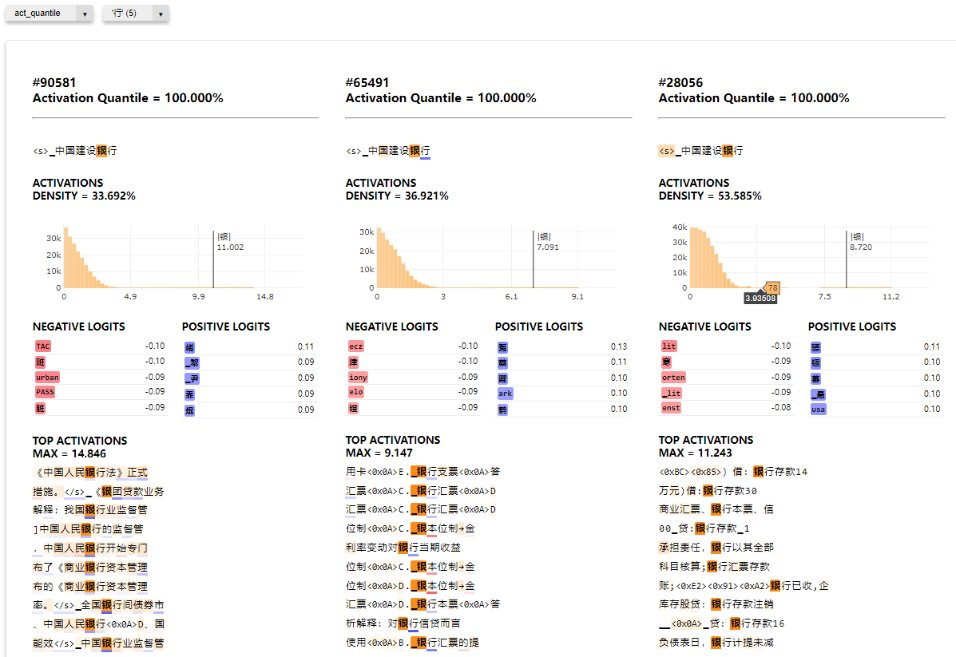
紫色下划线：表示消融该特征后，损失函数增大，预测该token的能力变差。即该token的预测受到了这个特征的帮助。

红色下划线：表示消融该特征后，损失函数减少，预测该token的能力变好。

Pos logit contributions/Neg logit contributions :消融该特征后，更可能/不可能预测为某个token。

### 4.2.4 以提示词为中心

第二种方式是采用以提示词为中心的可视化。在以提示词为中心的视图中，可以通过输入自定义提示，并根据各种可能的指标查看哪些特征在该提示上得分最高​。



以提示词为中心的视图

​如图所示，可以通过该视图了解到，“行”这个提示词的相关特征在“银”这个token上也有很大的激活值，表明了“银行”通常作为词组在金融领域高频出现，事实也显然如此。

## 4.3 金融大模型引导

### 4.3.1 正向引导

#### 输出一，有关价格

**引导前：**

Human：有关麦考利久期的正确表述是什么？

Assistant：麦考利久期（Macaulay Duration〉是一种衡量固定收益证券或任何其他金融工具的平均到期时间的指标。它是通过计算投资的现值加权平均来实现的， 每个现金流的现值与其到期时间成反比。具体来说，麦考利久期考虑了以下几个因素：\*\*各期现金流的金额\*\*：久期计算时将每一期的现金流（包括本金利利息）的现值加权求利。2.\*\*现金流的时间\*\*：每笔现金流的现值与其距离现在的时间（即其到期时间）成反比。3.\*\*贴现率\*\*：通常使用债券的到期收益率作为贴现率来计算现值。麦考利久期的正确表述可以概括为以下几点：\*\*麦考利久期是一种衡量金融工具平均到期时间的指标\*\*：它不是一个时间间隔，而是一个加权平均值。\*\*考虑现金流的时间利金额\*\*：计算时需要考虑每期现金流的金额和距离现在的时间。\*\*使用到期收益率作为贴现率\*\*：通常使用债券的到期收益率作为贴现家。\*\*反映投资的风险与收益\*\*：久期越长，表明投资的久期越大，意味者投资的价格对利率变动的敏感性越高，风险也越大。通过计算麦考利久期，投资者可以评估利率变动对债券价格的影响，并据此做出相应的投资决策。

**引导后：**

Human: 有关麦考利久期的正确表述是什么? Assistant: 麦考利久期 (Macaulay Duration)，又称为麦考利期限，是一个用于衡量固定收益证券的平均到期期限的指标。它考虑了证券的购买价格、利息支付频率和到期时间等因素。以下是麦考利久期的正确表述：

1. \*\*定义\*\*: 麦考利久期是一种度量固定收益投资平均回收成本时间的指标。它是将每次支付的现金流乘以时间因子，并将这些乘积相加后除以投资的当前价格。

2. \*\*计算公式\*\*:

$$

\text{Macaulay Duration} = \frac{\sum\_{t=1}^{T} \left( \frac{PV(C\_t)}{(1 + y)^t} \times t \right)}{\frac{PV(C\_1)}{(1 + y)} + \dots + \frac{PV(C\_T)}{(1 + y)^T}}

$$

其中:

- \( T \) 是投资的总期限（年数）。

- \( PV(C\_t) \) 是第 \( t \) 年的现金流（可能是利息或本金）的现值。

- \( y \) 是每期的收益率（通常用有效的到期收益率表示）。

- \( C\_1, C\_2, \dots, C\_T \) 是投资的各期现金流。

3. \*\*时间因子\*\*: 时间因子（时间权重或久期因子）是一个综合考虑每期现金流发生时间的权重因子，通常表示为:

$$

\text{Time Factor} = \prod\_{t=1}^{T} \left( \frac{1}{(1+r)^t} \right)

$$

其中:

- \( r \) 是每期的收益率（通常用有效的到期收益率表示）。

4. \*\*意义\*\*:

- 麦考利久期可以帮助投资者量度固定收益证券的平均到期期限，从而对投资组合的风险和收益进行评估。

- 期限越长，证券价格对利率变化的敏感性越高，因此风险越大。

- 久期可以用来预测债券价格对利率水平变化的反应，以及评估债券的持有期收益率 (YTM)。

### 例子

假设一位投资者购买了一张面值为1000美元、年利率为5%的债券，每半年支付一次利息（半年付息），并在5年后到期。债券的当前市价为980美元。我们可以计算这张债券的麦考利久期如下：

- \( T = 5 \) 年

- \( PV(C\_1) = 1000 \times 0.05 = 50 \) 美元

- \( PV(C\_2) = PV(C\_3) = \frac{1000 \times 0.05}{1.02} \times \frac{1}{2} + \dots = 45.81 \) 美元

- \( r = 0.05 \)

- \( Time Factor = \prod\_{t=1}^{T} \left( \frac{1}{(1+r)^t} \right) \times \frac{1}{1.02} \times \frac{1}{1.02} \times \dots = 0.7738 \)

所以:

\[

\text{麦考利久期} = \frac{\sum\_{t=1}^{T} \left( \frac{PV(C\_t)}{(1 + r)^t} \times t \right)}{\frac{PV(C\_1)}{(1 + r)} + \dots + \frac{PV(C\_T)}{(1 + r)^T}} = 4.36

\]

#### 输出二，国际+资产

**引导前**

举例一个银行的主要业务

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，包括很行转账、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\*\*:为客户提供信用卡中请、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，客户可以根据自艺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>

**引导后**

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、提供各种类型的货款等。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>

#### 输出三，政府+企业+机构

**引导前**

Human：我有一干万应该如何管理

Assistant：有一千万的资产是一个很大的财富，以下是一些建议来管理这笔财富：

1.制定预算计划:考虑到各种开支，包括日常生活费用、投资和储蓄等，预算可以帮助您更好地管理您的财务状。

2.分散投资:不要把所有的资金都放在一个地方，而是将资金分散到不同的投资渠道，如股票、偾券、房地产和黄金等。这样可以降低风险，并获得更好的回报。

3.寻找高回报的投资机会:在投资过程中，要寻找高回报的投资机会，但同时要注意风险控制。可以考虑投资股票、基金、房地产等领域，但要选择有良好信誉和业绩的投资项目

4.保持谨慎:在投资过程中，要保持谨慎，不要被高回报所诱惑。要根据自己的风险承受能力和投资目标来选择投资项目，并及时关注市场动态，调整投资策略。

5.建立紧急储备金:建立紧急储备金可以应对突发事件，如失业、疾病等。通常建议将3-6个月的生活开支作为紧急储备金。

6.咨询专业人士:如果您不确定如何管理您的财富可以咨询专业的理财顾问或财富规划师，他们可以根据您的财务状况和投资目标，为您提供个性化的建议和解决方案。

**开启引导之后**：

Human：我有一千万应该如何管理

Assistant：管理一千万的企业是一个复杂的过程，涉及到许多方面，以下是一些建议：

制定明确的战略计划：制定一个清晰的企业战略计划，包括目标、使命、愿景和核心价值观。这将帮助你明确企业的发展方向，并确保所有员工都了解并致力于实现这些目标。

2.建立有效的组织结构:根据企业战略和业务需求，建立一个高效、灵活的组织结构。明确各部门的职责和权限，确保资源得到合理分配和利用。

3.培养优秀的领导团队:选择和培养具有领导才能和专业知识的领导人员。为他们提供适当的训和发展机会，以提高他们的领导能力和管理技能。

4.关注人才培养和激励:重视员工的培训和发展，提供适当的激励机制，以吸引和留住人才。建立一套公正、透明的绩效评估体系，鼓励员工积极参与企业发展。

5.加强财务管理：建立完善的财务管理制度，确保企业财务的稳定和可持续发展。制定合理的预算和财务计划，控制成本，提高盈利能力。

6.提高品牌知名度：通过市场营销和品牌推广，捉高企业的知名度和美誉度。加强与客户的沟通和互动，了解客户需求，提供优质的产品和服务。

7.关注社会责任和可持续发展:企业应积极行社会责任，关注环境保护、公益事业等方面。时，要注重可持续发展，制定长期发展战略，确保企业在未来能够持续发展。</s>

#### 输出四，政府+企业+机构

**引导前：**

Human: 如何使收益最大 Assistant: 收益最大化是一个多方面的目标，以下是一些可能有助于你达到这个目标的建议:

1. **明确你的目标**: 首先，你需要明确你的收益目标。这可能是长期的，比如为退休储蓄，或者是短期的，例如旅行或购买新电脑等集资金。明确你的目标将帮助你制定适当的策略。
2. **制定预算**: 了解你的收入和支出是关键。创建一个预算，以确保你的支出不超过你的收入，并尽量减少不必要的开支。
3. **增加收入**: 如果可能，考虑增加你的收入来源。这可能包括找一份兼职工作，或者利用你的专业技能进行副业。
4. **投资**: 投资是实现收益最大化的重要方式。研究不同的投资选项，比如股票、债券、基金、房地产等，并选择适合你风险承受能力和收益目标的投资。
5. **多元化投资**: 不要把所有的鸡蛋放在一个篮子里。通过多元化投资，你可以降低风险，并可能获得更稳定的回报。
6. **定期审查和调整**: 定期审查你的投资组合，并根据市场的变化和你的目标进行调整。这可以帮助你保持在正确的轨道上，并最大化你的收益。
7. **教育自己**: 了解投资和理财的基本知识，以便你可以做出明智的决策。这可能包括读书、参加培训课程、寻求专业的财务顾问的帮助等。
8. **避免风险**: 在追求收益的同时，不要忽视风险管理。合理规避投资的风险，并尽量避免过度投资或冒险行为。
9. **保持耐心**: 投资是一个长期的过程，不要期望立即看到显著的收益。保持耐心，并坚持你的策略，随着时间的推移，你可能会看到你的收益增加。
10. **寻求专业建议**: 如果你对投资或理财不熟悉，寻求专业的财务顾问的建议可能是一个好主意。他们可以根据你的具体情况提供个性化的建议。

**引导后：**

Human: 如何使收益最大 Assistant: 收益最大化是企业经营中的一个重要目标，以下是一些策略和方法，可以帮助企业实现收益最大化:

* **优化产品或服务**:
  + 了解客户需求和期望，提供满足他们需求的高质量产品或服务。
  + 通过持续改进和创新，提高产品或服务的竞争力。
  + 确保产品或服务的定价合理，并具有市场吸引力。
* **有效的营销和品牌推广**:
  + 制定有效的市场营销策略，包括广告、促销和公共关系活动。
  + 通过美观的品牌形象，提高品牌知名度和认可度。
  + 通过社交媒体和其他数字渠道与客户互动，提高客户忠诚度。
* **提高运营效率**:
  + 优化生产和供应链管理，减少浪费和成本。
  + 通过技术升级和自动化，提高生产效率。
  + 实施精益管理原则，持续改进流程。
* **控制成本**:
  + 实施严格的成本控制措施，包括内部审计和成本分析。
  + 寻找并利用低成本的生产材料或供应商。
  + 减少不必要的开支，如节约能源和减少废物。
* **拓展市场和增加销量**:
  + 寻找新的市场和客户群体，扩大市场份额。
  + 通过新产品或服务线，增加收入来源。
  + 与其他企业建立合作伙伴关系，共享资源和客户。
* **投资回报率（ROI）分析**:
  + 对所有投资项目进行ROI分析，确保投资带来的回报超过成本。
  + 优先考虑高回报的投资项目。
* **财务管理和风险控制**:
  + 保持良好的现金流管理，确保企业有足够的现金流。
  + 通过风险管理工具和策略，降低潜在风险。
  + 定期审计财务报告，及时发现问题并采取行动。
* **员工激励和培训**:
  + 设计有效的激励计划，激励员工提高工作效率和质量。
  + 提供持续的职业发展机会和培训，提升员工技能和知识水平。
  + 关注员工的工作环境，提高员工满意度和忠诚度。
* **利用技术和数据**:
  + 实施大数据分析和人工智能技术，提高业务效率和准确性。
  + 通过数据驱动的决策制定，预测客户需求，指导业务策略。
  + 采用自动化和人工智能技术，降低成本并提高灵活性。
* **创新和适应变化：**
  + 关注行业动态和市场变化，及时调整业务策路。
  + 保持创新精神，开创新产品或服务，以适应市场需求。

建立灵活的组织结构，以适应快速变化的市场环境。

### 4.3.2 特征消融

#### 输出一，贷款

**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>

#### 输出二，瑞士银行

**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>

#### 输出三，二选一的情况，移除遗产规划

**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。

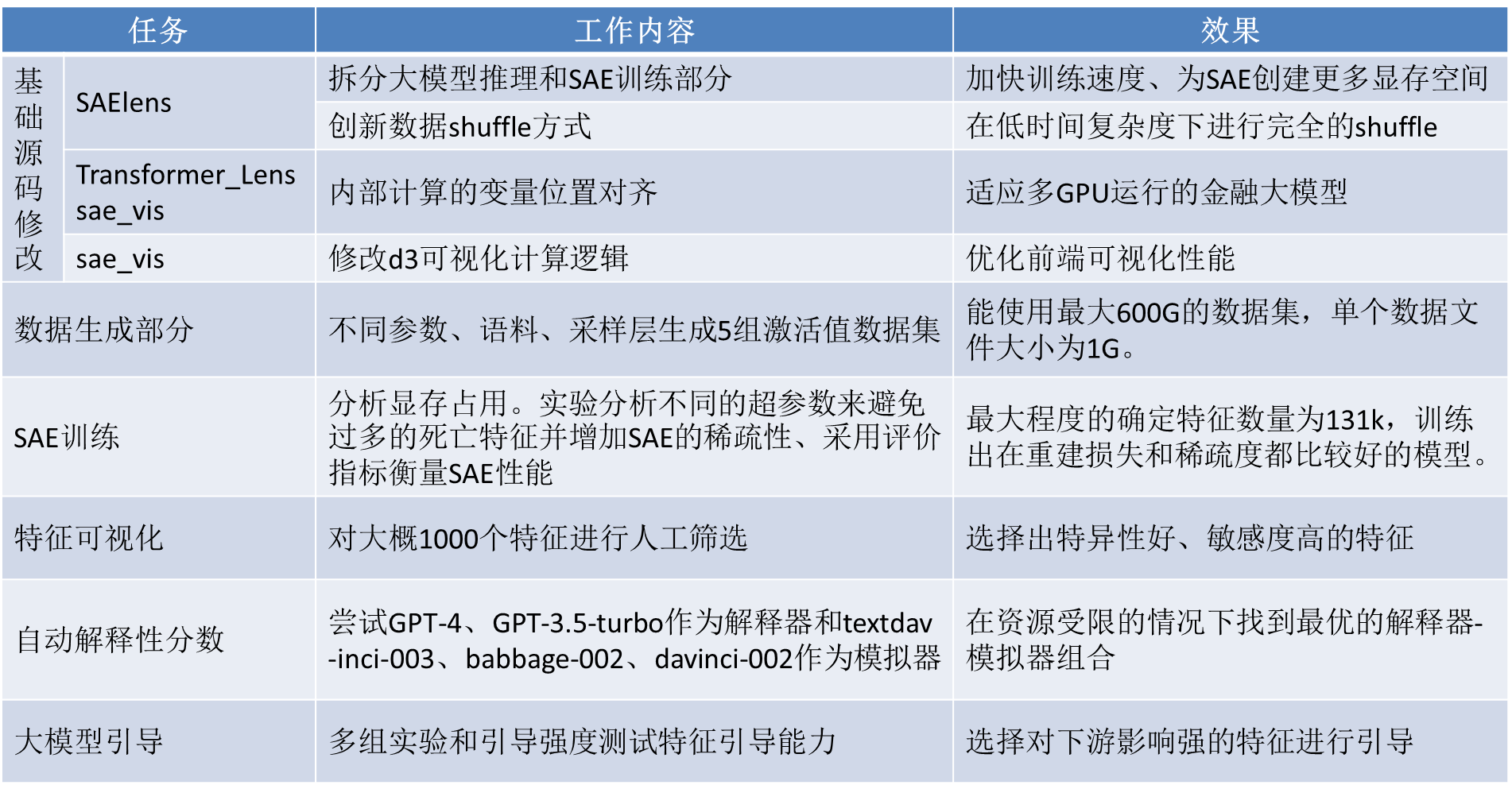
1. 保值增值：如果高净值人士的主要目标是保值增值，那么他们可能需要将更多的资产分配到相对安全的投资中，如债券，保险，或者某些类型的基金。这些投资通常能够提供稳定的回报，同时降低投资风险。

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>

# 第五章 总结与展望

## 5.1 总结



## 5.2 展望

### 技术方面

对金融大模型进行可解释性分析的技术方面展望分为三部分：更多语义、多GPU的实现和自动解释性分数的深入探索。

#### 更多语义

除了目前我们应用的是金融领域的语料库，我们还需要包含更多领域的语料库，比如通用领域的语料库。这样扩展语料库的好处是能够使筛选特征更加的纯净，并且还能使得通用领域的特征也会在金融领域起到作用。

#### 多GPU的实现

为了方便对SAE模型的扩展，需要进行多GPU的实现。目前可以实现方法有两种：

首先就是在SAElens开源的源码上做一个模型切片修改。不过其源码内部比较抽象、各种类的封装比较多，修改难度可能会比较大。但是好处是，SAElens功能比较丰富成熟，对接的其他开源代码比较多。

从0实现一个SAE模型，先实现模型切片，实现难度可能会比较低，但是在对接数据生成、可视化、模型引导部分，以及实现其他功能上面的工作量会比较大。

#### 自动解释性分数

关于自动解释性分数，我们进行了一个探索性的尝试，现在的自动解释性分数是调用GP4的API，如果对10万个特征进行解释的话，花费是比较大的。目前有两种进行解决的方式，一种是调用国产的大模型API，另外一个就是本地运行大模型。

### 业务方面

在业务方面的未来研究方向主要集中在对各种各种的特征进行探索，包括对业务特征、安全特征以及危险特征的一种分析和寻找。

#### 业务特征

对业务特征进行一种寻找，主要包括要发现更多的业务，包括但是不限于金融方面计算、企业决策、个人投资的特征。它的好处是，通过对大模型可解释性的研究，发现更多的业务特征。通过对大模型进行正向引导，以加速在各个业务领域的可移植性。

#### 安全特征

在安全特征方面，包括但不限于对逻辑错误敏感的特征、对数学计算错误敏感的特征进行寻找，它的好处是可以解释并分析大模型的严谨程度、安全能力，并且有助于确保大模型的安全性。

#### 危险特征

在危险特征方面，我们包括但不限于对性别歧视、种族歧视、欺骗以及犯罪特征的寻找。好处是通过对此类特征进行一种消融。可以增强用户的信任度，并且能够确保模型的安全性。

### 宏观展望

金融领域对模型的可解释性有着严格要求。本课题通过基于SAE对金融大模型进行可解释性分析，可以提取与金融行业密切相关的关键特征，并对模型的决策过程进行有效引导。这有助于优化金融大模型在多个金融领域的输出，显著提升用户体验。

通过对可解释性的研究，包括但不限于对以下方面进行促进：投资决策支持、风险评估与管理、客户服务与个性化推荐、合规与审计、市场分析与预测

# 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.