



中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于合作博弈论的神经网络预测可解释性研究

作者姓名：蔡怀广

指导教师：孙正雅 副研究员 中国科学院自动化研究所

学位类别：工学硕士

学科专业：模式识别与智能系统

培养单位：中国科学院自动化研究所

2025 年 6 月

Research on the Explainability of Neural Network Predictions
with Cooperative Game Theory

A thesis submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Master of Engineering
in Pattern Recognition and Intelligent System

By
Huaiguang Cai
Supervisor: Zhengya Sun

Institute of Automation, Chinese Academy of Sciences

June, 2025

中国科学院大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：

日 期：

导师签名：

日 期：

摘要

人工智能，特别是深度神经网络，已经在交通物流、医疗诊断、金融风控等诸多领域得到广泛应用，为社会创造了巨大价值。然而，神经网络的黑箱属性使得模型特定预测结果难以被解释，这在一定程度上降低了用户对模型的信任、阻碍了用户对预测结果合理性的验证、并限制了用户对模型预测特性的深入理解。为了提升神经网络模型在实际应用中的可信度和可靠性，研究神经网络预测结果的可解释性至关重要。

神经网络从预测模式的角度可分为单次推断模型和迭代推断模型。单次推断模型通过单次前向传播即可完成预测；而迭代推断模型则需要通过多次前向传播步骤逐步生成预测结果，其中每一步的输出通常依赖于前一步的状态。不同预测模式的神经网络需要不同的预测可解释性算法。值得注意的是，合作博弈论中的 Shapley 值是唯一能够保证参与者贡献被公平分配的方式，因此，Shapley 值在可解释性研究中饱受青睐。

本文基于 Shapley 值理论，研究了图像分类模型（作为单次推断模型的代表）与大语言推理模型（作为迭代推断模型的代表）预测可解释性中的特征归因问题，即如何将预测结果归因到输入样本特征中，从而实现预测结果的可解释性。本文的主要贡献和创新点如下：

在图像分类模型的特征归因研究中，针对当前图像分类模型上广泛应用的类别激活图算法理论基础薄弱的难题，提出了一种基于合作博弈论的类别激活图算法 ShapleyCAM。具体而言，首先通过构建内容保留的博弈论解释器的理论框架，阐明了两种具有代表性的类别激活图算法（GradCAM 和 HiResCAM）的理论基础。其次，在该框架下开发了一种同时利用神经网络的梯度和海森矩阵信息的类别激活图算法 ShapleyCAM。其原理是通过对效用函数进行二阶泰勒近似，推导出 Shapley 值的解析解，从而规避 Shapley 值计算复杂度高的问题。此外，针对类别激活图算法中效用函数选择存在争议的问题，理论分析了两种选择（pre-softmax 和 post-softmax 分数）生成的解释的等价性，并提出了融合二者优势的 ReST 效用函数。在 ImageNet 验证集和 12 种图像分类模型上的实验结果表明，ShapleyCAM 算法生成的解释在大多数情况优于现有类别激活图算法。

在大语言推理模型的特征归因研究中，针对及现有特征归因算法难以定量解释大语言推理模型的迭代推断结果的难题，提出了一种基于合作博弈论的大语言推理模型特征归因算法 QT Shapley。具体而言，首先将推理模型的最终答案生成结果建模为用户输入与中间推理步骤两个参与者（即两个广义上的特征）的合作博弈。随后，QT Shapley 通过计算两者的 Shapley 值定量解释两者在答案生成中的作用。此外，针对中间推理步骤和用户输入的贡献无标答的难题，提出了一种基于模型共识的贡献计算方式评估指标，通过计算不同模型在相同数据集上的价值排序的斯皮尔曼相关系数，评估贡献计算方式的准确性。定量实验

结果表明 QT Shapley 算法在评估用户输入和中间推理步骤对答案生成的贡献时,相较于多种启发式设计的基线算法更有效。基于定量实验的数据统计结果,定性解释了大语言推理模型的特点、适用范围和局限性,并据此提出了若干大语言推理模型的未来研究方向。

总体而言,本研究基于合作博弈论中 Shapley 值的公平分配贡献的特性,针对图像分类模型和大语言推理模型分别提出了具有创新性的特征归因算法,为其预测结果提供了准确的定量与定性解释。研究成果具有一定的理论深度与实际应用价值,丰富并拓展了可解释机器学习的研究体系。

关键词: 可解释机器学习, 合作博弈, 类别激活图, 大语言模型

Abstract

Artificial intelligence, particularly deep neural networks, has been widely applied in various fields such as transportation logistics, medical diagnosis, and financial risk control, creating significant value for society. However, the black-box nature of neural networks makes it difficult to explain specific prediction outcomes, which, to some extent, reduces user trust in the model, hinders the validation of the rationality of prediction results, and limits users' deeper understanding of the model's prediction mechanism. To enhance the credibility and reliability of neural network models in practical applications, research on the explainability of neural network prediction results is crucial.

From the perspective of prediction modes, neural networks can be categorized into single-inference models and iterative inference models. Single-inference models generate predictions through a single forward pass, whereas iterative inference models produce predictions progressively through multiple forward passes, with each step's output typically depending on the previous step's state. Different prediction modes of neural networks require different explainability algorithms. Notably, the Shapley value from cooperative game theory is the only method that guarantees a fair allocation of contributions among participants, making it highly favored in explainability research.

This thesis, based on the Shapley value theory, investigates the feature attribution problem in the explainability of model predictions in image classification models (as a representative of single-inference models) and large language reasoning models (as a representative of iterative-inference models). Specifically, it explores how to attribute prediction results to the input sample features in order to provide an explanation. The main contributions and innovations of this paper are as follows:

In the feature attribution research for image classification models, to address the difficulty of the weak theoretical foundation of widely used class activation map algorithms, this thesis propose a CAM algorithm named ShapleyCAM based on cooperative game theory. Specifically, this thesis first propose a theoretical framework named Content Reserved Game-theoretic Explainer, clarifying the theoretical foundation of two representative class activation map algorithms (GradCAM and HiResCAM). Secondly, under this framework, ShapleyCAM is developed, a CAM algorithm that simultaneously leverages both the gradient and Hessian matrix information of neural networks. The core idea is to derive the analytical solution of the Shapley value by performing a second-order Taylor approximation of the utility function, thereby circumventing the high computational complexity of Shapley value calculation. Moreover, to address the controversial issue of utility function selection in class activation mapping algorithms, a theoretical analysis of explanations generated by two utility function choices (pre-

softmax and post-softmax scores) is conducted, and the ReST utility function is proposed to integrate the advantages of both. Experimental results on the ImageNet validation set and 12 image classification models demonstrate that the explanations generated by the ShapleyCAM algorithm outperform existing class activation map algorithms in most cases.

In the feature attribution research for large language reasoning models, to address the difficulty that existing feature attribution algorithms cannot quantitatively explain the iterative inference results of such models, this thesis propose a feature attribution algorithm for large language reasoning model named QT Shapley based on cooperative game theory. Specifically, the final answer generation is modeled as a cooperative game between two participants, namely, the user's query and the intermediate reasoning steps. QT Shapley then quantitatively assesses their respective contributions to the answer by calculating their Shapley values. Moreover, to tackle the challenge of evaluating these contributions in the absence of ground truth, a model consensus-based metric is proposed. This metric assesses the accuracy of contribution estimation methods by computing the Spearman correlation coefficient of value rankings across different models on the same dataset. Quantitative experimental results demonstrate that the QT Shapley algorithm is more effective than various heuristically designed baseline methods in evaluating the contributions of user input and intermediate reasoning steps to answer generation. Based on the statistical results, the thesis further provides a qualitative interpretation of the characteristics, applicable scope, and limitations of large language reasoning models, and proposes several future research directions accordingly.

In summary, this research, based on the fair distribution property of the Shapley value in cooperative game theory, proposes innovative feature attribution algorithms for both image classification models and large language reasoning models, providing accurate quantitative and qualitative explanations for their prediction results. The research findings possess significant theoretical depth and practical application value, enriching and expanding the research framework of explainable machine learning.

Key Words: Explainable Machine Learning, Cooperative Game, Class Activation Mapping, Large Language Model

目 录

第 1 章 绪论	1
1.1 研究背景与意义	1
1.2 研究内容与主要贡献	3
1.3 论文组织结构	4
第 2 章 可解释性算法研究现状	5
2.1 事前可解释性算法	5
2.2 事后可解释性算法	6
2.2.1 特征归因算法	6
2.2.2 样本归因算法	7
2.3 本章小结	9
第 3 章 基于合作博弈论的图像分类模型预测可解释性研究	11
3.1 引言	11
3.2 相关工作	12
3.3 基于 Shapley 值的类别激活图算法	15
3.3.1 内容保留的博弈论解释器的定义	15
3.3.2 GradCAM 和 HiResCAM 的理论基础	18
3.3.3 ShapleyCAM 算法的设计与实现	20
3.3.4 效用函数的优化设计	23
3.4 实验结果与分析	27
3.4.1 实验设置	27
3.4.2 评估指标	28
3.4.3 定量实验	30
3.4.4 定性实验	33
3.4.5 讨论	36
3.5 本章小结	37
第 4 章 基于合作博弈论的大语言推理模型预测可解释性研究	39
4.1 引言	39
4.2 相关工作	41
4.3 基于 Shapley 值的用户输入与中间推理步骤贡献量化算法	44
4.3.1 效用函数的定义	44
4.3.2 QT Shapley 算法的设计与实现	45

4.3.3 贡献计算方式评估指标的设计	47
4.4 实验结果与分析	48
4.4.1 实验设置	48
4.4.2 对比算法	49
4.4.3 定量实验	50
4.4.4 定性实验	51
4.5 本章小结	60
第 5 章 总结与展望	63
5.1 本文研究内容总结	63
5.2 未来研究方向展望	63
参考文献	65
附录 一	69
附录 二	71
作者简历及攻读学位期间发表的学术论文与其他相关学术成果 ..	87

图目录

图 1-1 黑箱人工智能与可解释人工智能的对比	1
图 1-2 有偏模型与无偏模型在视觉解释上存在差异 ^[1]	2
图 1-3 论文结构安排	4
图 3-1 CAM ^[2] 类算法通过线性组合 GAP 层之前的激活图以可视化预测的依据	12
图 3-2 SHAP 在 ResNet-50 上生成的视觉解释 ^[3]	14
图 3-3 Shapley 值计算过程的示意图	16
图 3-4 ResNet-18 上 GradCAM 分别结合 pre-softmax、post-softmax 和 ReST 效用函数得到的热力图，目标层为 GAP 层前的层	26
图 3-5 ResNet-18 上 GradCAM 生成的热力图，使用 ReST 和 GAP 之前的层。左：ResNet-18 错误地将虎鲨预测为锤头鲨。右：ResNet-18 正确预测了虎鲨和锤头鲨	28
图 3-6 VGG-16 上使用 ShapleyCAM 和最后一个卷积层生成的解释图、视觉解释和反解释图	28
图 3-7 VGG-16 上使用 ReST 和最后一个卷积层时不同算法生成的视觉解释，目标类别在第一列的左上方	36
图 3-8 ResNet-18 上不同 CAM 算法生成的视觉解释，使用 ReST 和 GAP 前一层。目标类别从上到下分别为：虎斑猫、拳师犬和黄色女士拖鞋	37
图 4-1 AttnLRP 算法在 LLaMa 2-7b 模型预测下一个 token 为“3”时生成的解释 ^[4]	42
图 4-2 思维链技术在不同任务上的表现	43
图 4-3 不同数据集的用户输入、中间推理步骤贡献占比	55
图 4-4 STEM 大类下不同数据集的 think_ratio 分布	57
图 4-5 STEM 大类下不同数据集的 query_ratio 分布	57
图 4-6 Math 大类下不同数据集的 think_ratio 分布	58
图 4-7 Math 大类下不同数据集的 query_ratio 分布	58
图 4-8 Community 大类下不同数据集的 think_ratio 分布	59
图 4-9 Community 大类下不同数据集的 query_ratio 分布	59
图 附录一-1 不同数据集的 query、think 贡献占比（模型：DeepSeek-R1-Distill-Qwen-7B）	69
图 附录一-2 不同数据集的 query、think 贡献占比（模型：DeepSeek-R1-Distill-Qwen-14B）	69
图 附录一-3 所有数据集的 think_ratio 分布（模型：DeepSeek-R1-Distill-Qwen-1.5B）	70

图 附录一-4 所有数据集的 query_ratio 分布（模型：DeepSeek-R1-Distill-Qwen-1.5B）	70
---	----

表目录

表 3-1 不同的 CAM 算法在 8 种卷积神经网络上的效果，将全局池化层的前一层作为目标层	32
表 3-2 不同的 CAM 算法在 Swin Transformer 上的效果，将全局池化层的前一层作为目标层	33
表 3-3 不同的 CAM 算法在 8 种卷积神经网络上的效果，将最后一个卷积层作为目标层	34
表 3-4 不同的 CAM 算法在 Swin Transformer 上的效果，将最后一个 Transformer 块的首个归一化层作为目标层	35
表 3-5 Grad-CAM 算法在 ResNet-18 上使用 pre-softmax、post-softmax 与 REST 的性能评估	35
表 4-1 Chinese-Data-Distill-From-R1 数据集统计信息	48
表 4-2 不同中间推理步骤贡献计算方式的斯皮尔曼相关系数	50
表 4-3 不同用户输入贡献计算方式的斯皮尔曼相关系数	51
表 4-4 在最终答案生成过程，中间推理步骤取代了部分用户输入功能 ...	60

第1章 绪论

1.1 研究背景与意义

人工智能，特别是深度学习神经网络，已经在交通物流、医疗诊断、金融风控等诸多领域得到应用，为社会创造了巨大价值。例如，在医疗诊断领域^[5]，人工智能辅助诊断系统被广泛应用于疾病筛查和诊断。然而，尽管当前的深度学习模型在某些任务上的表现已超越人类，但其黑箱属性使得模型特定预测结果难以被解释^[6,7]。模型可解释性的缺失会带来多重负面影响：首先，由于无法理解模型的决策依据，用户对模型的信任度降低，这制约了深度学习模型的推广应用。其次，在缺乏预测结果可解释性的情况下，用户无法将模型预测结果与人类知识进行交叉验证，因此难以验证预测结果的合理性。最后，在医疗等需要深度融合领域知识的专业领域，模型预测逻辑与专业知识之间的不匹配阻碍了模型的持续优化。

相比之下，模型预测结果具备可解释性将带来一系列好处。首先，透明的决策过程可以建立用户信任，为技术落地应用创造有利条件；其次，通过将模型解释与领域知识进行比对，既能验证预测结果的可靠性，又能及时发现并纠正模型错误；最后，系统回溯分析模型预测结果的依据，可以帮助模型训练者识别模型潜在问题，从而针对性地优化训练过程以改进模型。

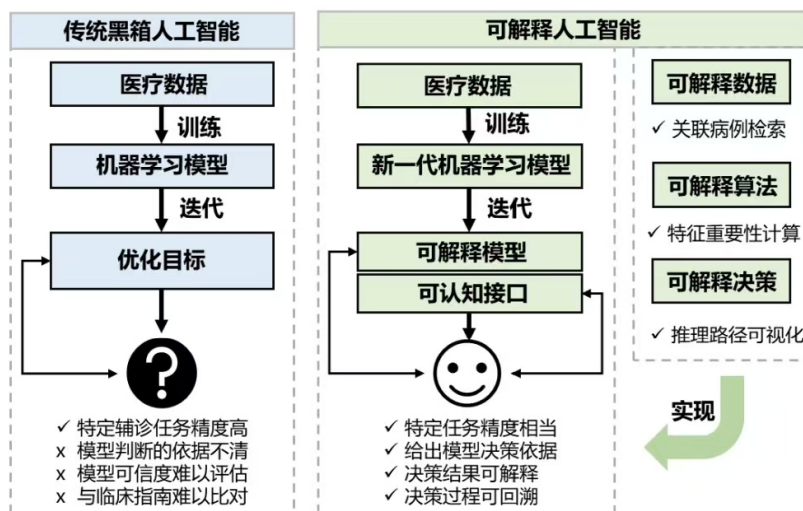


图 1-1 黑箱人工智能与可解释人工智能的对比

Figure 1-1 Comparison of black-box AI and explainable AI

为了直观说明模型预测结果的可解释性的作用，以图 1-2 为例，比较有偏模型与无偏模型在职业预测任务中所依赖的关键特征差异。从第一行对比可以看出，尽管两个模型均正确识别了护士，但视觉解释揭示了有偏模型主要依赖面部特征进行预测，而无偏模型则聚焦于更具判别力的服饰特征（短袖）。第二行进

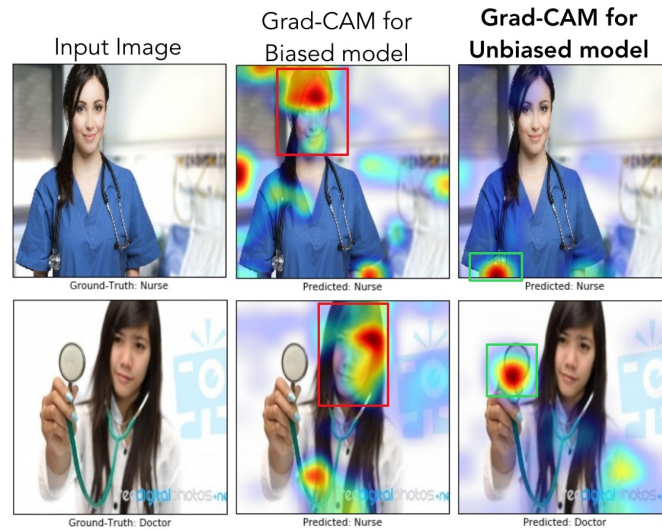


图 1-2 有偏模型与无偏模型在视觉解释上存在差异^[1]

Figure 1-2 Visual explanations generated by biased and unbiased models differ^[1]

一步展示了偏见对模型预测的影响：有偏模型不仅错误地将女性医生误判为护士，而且其预测依据主要集中在面部和发型等与职业识别无关的特征上；相较之下，无偏模型则正确地将焦点放在了白大褂和听诊器等医学职业特有的特征上，从而预测正确。进一步的数据调查表明，模型偏见的根源在于训练数据集中的严重失衡：医生样本中男性占比高达 78%，而护士样本中女性占比则高达 93%，这种失衡导致有偏模型倾向于将所有女性预测为护士。

这个案例充分说明了模型预测结果可解释性的价值：（1）增强对模型预测结果的理解：通过对预测结果直观的视觉解释，用户可以清晰地看到为何有偏模型会将女性误判为护士（面部特征被过度依赖），而无偏模型则通过更合理的服饰特征（如短袖）做出预测。这使得用户能够更好地理解模型的决策过程，从而增强对模型和产品的信任；（2）及时识别模型可能的错误预测：当用户发现有偏模型在预测过程中依赖了面部特征等与职业识别无关的特征时，便能够推断出模型可能出现了错误。（3）发现并缓解模型的潜在偏见：有偏模型过度依赖与职业识别无关的特征，这将促使模型训练者重新审视训练过程，进而发现数据集中的性别失衡问题。模型训练者从而可以采取相应措施优化训练过程，减少偏见的影响，从而提升模型的公平性和准确性。

为了提升神经网络模型在实际应用中的可信度和可靠性，本文聚焦于两类具有不同预测模式的典型神经网络的预测可解释性算法研究，从理论分析与算法设计两个层面开展深入探究。针对当前图像分类模型上广泛应用的类别激活图算法理论基础薄弱以及现有特征归因算法难以定量解释大语言推理模型的迭代推断结果的难点，通过构建更具理论支撑的特征归因算法，推动神经网络预测可解释性研究的发展，为人工智能在关键领域的应用提供理论指导和技术支撑，具有重要的科学价值和实际意义。

1.2 研究内容与主要贡献

近年来，随着神经网络在各类任务中的广泛应用，其预测结果的可解释性需求日益凸显。本文聚焦于可解释性研究中的特征归因问题：**如何将神经网络的预测结果归因到输入特征上。**

神经网络预测可解释性算法的选择需根据其预测模式进行适配。从预测模式的角度，神经网络可分为以下两类：单次推断模型和迭代推断模型。单次推断模型（如图像分类网络、目标检测网络、语义分割网络等）通过单次前向传播即可完成预测；而迭代推断模型（如大语言模型）则需要通过多次前向传播步骤逐步生成预测结果，其中每一步的输出通常依赖于前一步的状态。基于这种分类框架，本研究针对两类具有典型预测模式的模型，分别研究了其对应的特征归因算法：一方面针对图像分类模型，特征归因的目标在于识别输入图像中对预测结果起决定性作用的视觉特征；另一方面针对大语言推理模型，特征归因的目标则是定量解释中间推理步骤和用户输入这两类广义特征对最终答案生成的作用。具体而言：

(1) 本研究的第一个工作聚焦于图像分类模型的特征归因算法设计。针对当前图像分类模型上广泛应用的类别激活图算法理论基础薄弱的问题，首先提出内容保留的博弈论解释器的理论框架，阐明了两种具有代表性的类别激活图算法（GradCAM 和 HiResCAM）的理论基础。其次，在此框架下开发了一种同时利用神经网络的梯度和海森矩阵信息的类别激活图算法 ShapleyCAM。其原理是通过对效用函数进行二阶泰勒近似，推导出 Shapley 值的解析解，从而规避 Shapley 值计算复杂度高的问题。此外，针对类别激活图算法中效用函数选择存在争议的问题，从理论上分析了两种选择（pre-softmax 和 post-softmax 分数）生成的解释的等价性，并提出了 ReST 效用函数以结合两者的优势。最后，在 ImageNet 验证集和 12 种主流图像分类模型上的实验结果表明，ShapleyCAM 算法生成的解释在大多数情况优于现有类别激活图算法。

(2) 本研究的第二个工作聚焦于大语言推理模型的特征归因算法设计。由于大语言模型的预测结果是迭代推断生成的，常用特征归因算法难以直接应用。为此，本研究将推理模型的最终答案生成结果建模为用户输入与中间推理步骤两个参与者（即两个特征）的合作博弈结果，并提出了 QT Shapley 算法以定量评估用户输入和中间推理步骤在答案生成中的贡献。其次，针对中间推理步骤和用户输入的贡献无标答的难题，提出了一种基于模型共识的贡献计算方式评估指标，通过计算不同模型在相同数据集上的价值排序的斯皮尔曼系数，评估贡献计算方式的准确性。最后，定量实验结果表明 QT Shapley 算法在评估用户输入和中间推理步骤对答案生成的贡献时，相较于多种启发式设计的基线算法更有效。基于定量实验的数据统计结果，定性解释了大语言推理模型的特点、适用范围和局限性，并据此提出了若干大语言推理模型的未来研究方向。

1.3 论文组织结构

本文的章节脉络如图 1-3 所示，具体结构安排如下：

第一章，绪论。介绍了神经网络预测可解释性的研究背景和意义、存在的困难以及本文的研究内容与主要贡献。

第二章，可解释性算法研究现状。本章首先对现有可解释性算法进行分类，回顾了其中代表性算法。

第三章，基于合作博弈论的图像分类模型预测可解释性研究。首先详细介绍了传统 CAM 类算法以及 SHAP 算法，随后重点阐述了内容保留博弈论解释器的构建、ShapleyCAM 算法的设计过程、结合 pre-softmax 与 post-softmax 两者优势的 ReST 效用函数设计过程。最后通过定量和定性实验，检验了所提算法的有效性。

第四章，基于合作博弈论的大语言推理模型预测可解释性研究。首先提出了 QT Shapley 算法定量解释大语言推理模型答案生成中用户输入和中间推理步骤作用。接着提出了一种基于模型共识的贡献计算方式评估指标来评估计算法的准确性。此外，定量实验验证了 QT Shapley 算法评估用户输入和中间推理步骤对最终答案生成的贡献时，比多种启发式设计的基线算法更有效。最后基于 QT Shapley 给出的定量结果，定性解释了大语言推理模型的特点、适用范围和局限性。

第五章，总结与展望。总结了全文的研究工作，并对未来的研究方向进行了展望。

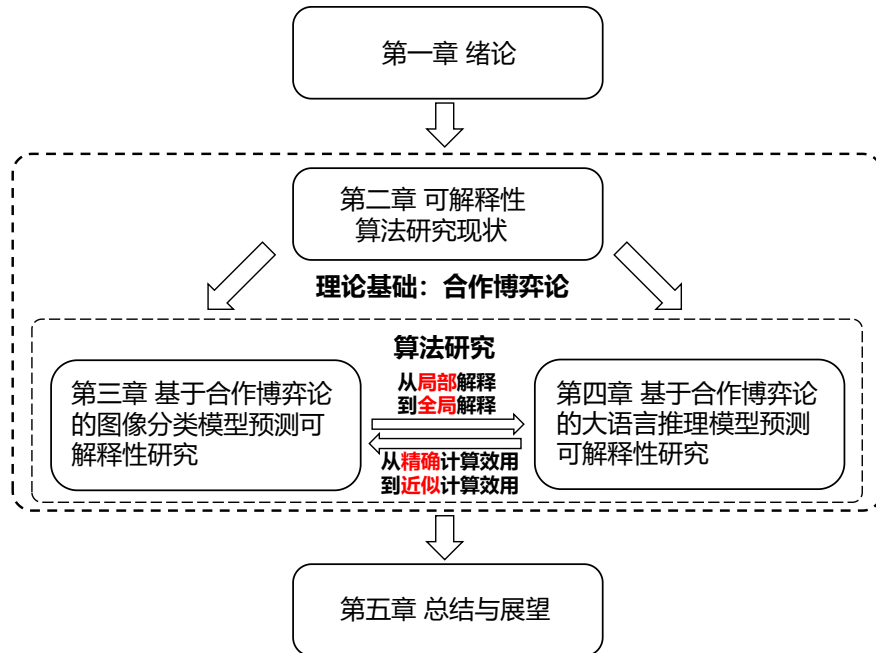


图 1-3 论文结构安排

Figure 1-3 The structure of this thesis

第2章 可解释性算法研究现状

可解释性算法的研究领域广泛，已有诸多分类方式^[6-11]，不同的分类方式反映了研究者在可解释性不同层面的关注点。总体而言，现有研究可大致分为事前可解释性和事后可解释性两大类。本章首先介绍事前可解释性算法，即模型在训练前即具备内在可解释性。随后探讨事后可解释性算法，这类算法需要通过额外的后处理技术来解释已有模型。事后可解释性算法根据解释结果的粒度，又可进一步分为特征归因和样本归因。

2.1 事前可解释性算法

事前（post-hoc）可解释性指的是在模型设计和构建阶段，就使模型本身具备易于理解的结构和决策规则^[9]。具备事前可解释性的模型被称为白盒（white-box）模型，与之相对的具有高度复杂结构的模型被称为黑盒（black-box）模型。具体地，线性模型、决策树模型因其简单直接的结构通常被认为具备事前可解释性的模型，即白盒模型^[7]。

典型的线性模型包括 Lasso 回归（Lasso Regression）和岭回归（Ridge Regression），二者均具有良好的可解释性。Lasso 回归采用 L1 正则化，通过对回归系数施加稀疏性约束，使部分特征的系数收缩至零，从而实现自动特征选择，形成一个特征较为稀疏的模型。而岭回归则采用 L2 正则化，通过对权重进行平滑约束，有效防止过拟合，并提升模型的稳定性。两种算法均可通过回归系数直接衡量各特征对输出的影响，使得模型决策过程更加透明且易于理解。

典型的决策树模型包括 ID3（Iterative Dichotomiser 3）、C4.5 和 CART（Classification and Regression Tree）算法。ID3 采用信息增益作为树节点划分标准，优先选择能最大程度减少数据不确定性的特征进行划分。C4.5 在 ID3 基础上改进，采用信息增益率作为分裂标准，降低了对多取值特征的偏好，并支持对连续变量和缺失值的处理。CART 使用基尼系数作为分裂标准，同时支持分类和回归任务。决策树模型的树状结构使得其决策路径清晰透明，便于追溯每个预测的依据。同时，树模型的分裂规则可以直接转化为决策规则，易于人类理解和应用。

由于线性模型和决策树这两种白盒模型本身结构具备良好的可解释性，研究者也尝试将高度复杂的黑盒模型（或者其部分）转换为白盒模型从而完成对黑盒模型的解释，但通常会涉及到重新训练模型。例如，概念瓶颈模型^[12]（Concept Bottleneck Model）将神经网络的预测过程视作两阶段：首先将输入映射到一组预定义的人类可以理解的高级概念空间，将在这些概念上得到的分数线性加权进行预测。类似于线性模型，将概念上的分数作为对预测的解释依据。研究^[13]则尝试用决策树来解释卷积神经网络（Convolutional Neural Network, CNN）的全连接层：给定一个训练完成的 CNN 网络，重新训练其最后一层卷积层及后续全

连接层，使其最后一层卷积层的每个滤波器（通道）对应一个特定的物体部位，从而将 CNN 的中间特征转换为有语义意义的概念；接着训练一个决策树，使其依据这些语义概念做出与 CNN 相同的预测，进而实现对 CNN 预测过程的解释。

此外，还有一些研究尝试证明黑盒模型的结构具有明确的意义，具备白盒特性。例如，研究^[14,15]致力于从数学原理的角度解释黑盒模型，特别是神经网络架构设计的原因，以使网络的每个组件都能具有精确的优化、统计和几何解释^[14]。作者提出在最大化数据压缩率的目标下，使用不同的优化方法优化目标函数自然导致不同的多层神经网络机构，例如白盒 CNN 模型^[14]和白盒 Transformer 模型^[15]。

2.2 事后可解释性算法

事后（anti-hoc）可解释性算法通过对固定参数的黑盒模型进行额外的分析，从而揭示其内部决策过程^[9]。根据解释结果的粒度，事后可解释性算法主要可分为两类：特征归因算法和样本归因算法。特征归因算法从输入样本的特征层面出发，分析模型的预测依据；而样本归因算法则关注训练数据集中的个体样本，评估它们对模型整体性能或特定预测结果的影响。事后可解释性算法中，大多数属于特征归因算法。

2.2.1 特征归因算法

特征归因算法旨在将神经网络的预测结果归因于输入特征，从而揭示模型的决策逻辑。根据算法原理，特征归因算法可大致分为如下几类：

(1) 基于反向传播的可解释性算法：利用梯度信息衡量输入特征对输出的贡献。显著性图^[16]（Saliency Map）通过计算输入图像中每个像素相对于特定类别得分的梯度从而突出对分类结果影响最大的区域。积分梯度^[17]（Integrated Gradients）通过在原始输入和固定基准值之间的特征变化路径上采样对应梯度进行积分，有效避免梯度消失问题。SmoothGrad^[18]发现梯度图通常包含大量噪声，且给图片加微小扰动会造成梯度解释不稳定，因此采用多张加噪图片的梯度图平均作为解释。逐层相关性传播^[19]（Layer-wise Relevance Propagation, LRP）在满足每层神经元的分数总和一致的前提下将输出分数逐层反向传播到输入层得到解释。算法的难点在于如何为不同层的神经元设计分数分配方式。DeepLIFT^[20]通过比较每个神经元当前输出与基准输出之间的差异，逐层反向传播计算每个特征对模型输出的贡献，从而提供解释。

(2) 基于扰动的可解释性算法：通过比较输入特征在不同扰动下的模型输出的差异确定特征重要性。LIME^[21]（Local Interpretable Model-agnostic Explanations）通过在待解释样本周围生成大量扰动样本，并训练局部的线性模型来近似复杂模型的决策边界，从而揭示局部决策规律。Mask^[22]通过梯度下降找到对图像当前预测下降程度最高且稀疏的遮挡部分作为解释。SHAP^[3]（SHapley Additive exPlanations）利用合作博弈论中 Shapley 值是唯一能够保证参与者贡献

被公平分配的方式这一特性，通过计算特征在不同组合中的边际贡献，从而提供公平的解释。值得注意的是，SHAP 与多种特征归因算法密切相关。SHAP 原论文证明了基于扰动的 LIME、基于反向传播的 LRP 以及 DeepLIFT，均可视作使用不同近似技巧进行 Shapley 值计算的特例。SHAP 算法在理论上对这些算法进行了统一，同时在实际应用中结合不同近似策略，衍生出多种 SHAP 变种算法。例如，Kernel SHAP 通过结合线性 LIME 算法与 Shapley 值进行计算，Deep SHAP 结合了 DeepLIFT 与 Shapley 值的思想，而 Gradient SHAP 则融合了 Integrated Gradients 与 Shapley 值的计算算法。这些变种算法在不同场景下各具优势，需要根据实际情况灵活使用。

(3) 基于激活图的可解释性算法：通过为神经网络靠后层输出的激活图分配不同的权重，从而实现对图像的解释。例如，GradCAM^[1] 利用梯度的平均值来确定各激活图的权重，从而突出对预测贡献较大的区域。关于基于激活图的可解释性算法的详细介绍见节 3.2。

(4) 基于因果推理的可解释性算法：通过构建因果关系以解析决策机制。例如，研究^[23] 在预先定义变量后，从智能体和环境的交互数据中学习一个结构因果模型（Structural Causal Model, SCM），将其用于解释强化学习的动作选择策略。

此外，还有一些主要面向大语言模型的可解释性算法，如基于注意力机制的可解释性算法、机械可解释性算法等，相关内容将在4.2中详细介绍。

最终，特征归因算法的解释结果可以进一步划分为**局部解释**和**全局解释**，即分别针对单个样本的决策逻辑进行解释，或揭示模型整体的决策逻辑。虽然两者在粒度上有所不同，但并不互相冲突。全局解释可以通过多个局部解释的汇总得到。例如，使用 SHAP 处理表格数据时，单个样本的解释结果属于局部解释，对整个测试解释结果进行平均则得到全局解释。

2.2.2 样本归因算法

特征归因算法将黑盒模型的参数视为环境条件，而样本归因算法则更进一步，关注训练数据集的各个样本对模型整体性能或特定预测结果的影响。具体可分为两类：模型预测的样本归因和模型训练的样本归因。

模型预测的样本归因算法旨在将模型给出的特定预测结果归因至某些训练样本。其解释的对象是单个样本的预测结果，因此更接近于一种局部解释。

影响函数^[24]（Influence Function）是该领域代表性的算法之一。该算法的核心思想是通过量化单个训练样本对模型特定预测结果的边际影响来实现样本归因。具体而言，对于给定的测试样本，通过计算当某个训练样本被施加无穷小权重扰动时，重训练的模型在该测试样本上预测损失的变化量来评估其影响。若影响函数值为负，表明该训练样本有助于降低测试样本的预测损失，通常对应于训练集中和该测试样本的相似样本；若影响函数值为正，则暗示该训练样本可能干扰模型预测，常见于噪声或对抗样本。在实现层面，该算法需要计算测试

样本与训练样本的梯度内积，并利用损失函数海森矩阵的逆进行校正。该算法优势在于提供了严格的统计解释，能够有效识别对预测结果影响较大的关键训练样本。然而，此算法也存在一定局限：首先要求损失函数在参数空间满足凸性假设以保证海森矩阵可逆；其次，海森矩阵求逆的计算复杂度随参数维度急剧增长，导致在大规模模型上难以应用；最后，影响函数理论本质是估计了留一法（Leave-One-Out, LOO）对预测结果的作用，在很多任务上效果不如能够严格公平分配贡献的 Shapley 值算法^[25]。

为了降低计算量，尤其是避免计算海森矩阵，TracIn^[26] 通过计算模型训练过程中每一轮训练样本与测试样本之间梯度的内积，并对其求平均，从而衡量训练样本对测试样本的影响。

模型训练的样本归因算法旨在将模型在测试集上的整体性能归因至各个训练样本。其解释的对象是整体模型的性能，因此更接近于一种全局解释。此类算法亦被称为数据估值（Data Valuation）算法，即为每个训练样本分配一个数值，以衡量其对模型性能的贡献。合理的数据估值对于数据选择^[27]、数据获取^[28] 以及数据交易^[29] 等任务具有重要意义。

在数据估值研究中，Data Shapley^[30] 算法被广泛认为是该领域的奠基性的工作。该算法基于 Shapley 值理论，将训练数据视为合作博弈中的参与者，并依据其边际贡献计算数据价值。Data Shapley 采用蒙特卡罗方法估计每个数据点的价值，其模型重训练复杂度达到 $\mathcal{O}(n^2 \log n)$ ，其中 n 为数据点的数量。一般而言，数据价值较低的样本往往对模型训练产生负面影响，而数据价值较高的样本则可能是训练集中缺失的高质量数据。因此，通过剔除低价值数据或引入高价值数据，可有效提升模型的预测性能。同样由于 **Shapley 值是唯一能够保证参与者贡献被公平分配的方式**，Data Shapley 计算出的结果也被认为是数据价值的标答（Ground Truth）。

在 Data Shapley 之后，研究者们相继提出了一系列改进算法。KNN Shapley^[31] 针对 k 近邻算法提出了闭式解法，大幅提升了计算效率，使其模型重训练复杂度降低至 $\mathcal{O}(\sqrt{n} \log(n)^2)$ 。Beta Shapley^[32] 通过放宽 Shapley 值的效率公理，并引入 beta 分布来调整不同基数集合的权重分配，在错误标签检测等任务上表现更优。Data Banzhaf^[33] 通过引入 Banzhaf 值以减少随机梯度算法在训练过程中的噪声影响，被认为是效用函数存在噪声情况下最鲁棒的数据估值算法，尤其适用于使用随机梯度下降算法作为优化算法的深度学习任务。其中，Banzhaf 值被证明是在效用函数中存在噪声的情况下，所有半值（semivalue，即放宽效率公理假设的 Shapley 值）中对噪声最鲁棒的解概念（solution concept）。LAVA^[34] 利用训练集和验证集之间的类间 Wasserstein 距离来衡量训练数据与验证数据在同一类别下的分布差异计算数据价值，从而使得数据估值算法与模型训练算法无关。该算法在检测低质量数据任务中表现出色，且计算效率高于传统算法。DVRL^[35] 和 Data-OOB^[36] 分别利用强化学习和袋外估计技术评估数据价值。OpenDataVal^[25] 在统一的实验环境下对上述算法进行了系统性比较，并在不同任务场景中评估

其性能。研究表明,不同数据估值算法在不同应用场景下表现各异,尚无一种算法能够在所有情境下优于其他方案。**In-Run Data Shapley**^[37] 则探索了使用二阶泰勒展开来近似模型性能的效用函数,通过推导每个训练数据点的 Shapley 值的闭式解实现单次模型训练即可估计所有训练数据的价值。

2.3 本章小结

本章首先介绍事前可解释性算法,即模型在训练前即具备内在可解释性。随后探讨事后可解释性算法,这类算法需要通过额外的后处理技术来解释已有模型。事后可解释性算法根据解释结果的粒度,又可进一步分为特征归因和样本归因。

而在事后可解释性算法中,基于 Shapley 值的可解释性算法占据重要地位:SHAP 统一了多种常用的特征归因算法,Data Shapley 则是样本归因算法中奠基性的工作,而两者的理论基础皆为合作博弈论中的 Shapley 值。Shapley 值之所以受到事后可解释性算法研究者的广泛关注,是因为在已完成的合作博弈中,Shapley 值是唯一能够保证参与者贡献被公平分配的方式。

尽管基于 Shapley 值的算法因其坚实的理论基础而受到认可,但计算效率上的挑战限制了其广泛应用。而在本文中,针对这一问题,在图像分类模型的特征归因研究中,将 In-Run Data Shapley 中直接估计效用函数的思路扩展至特征归因领域,从而推导出更高效的图像上的特征归因算法;从可解释性算法分类的角度来看,提出的 ShapleyCAM 算法属于事后可解释性算法、特征归因算法以及局部解释算法。而在大语言推理模型的特征归因研究中,由于博弈参与者较少,可直接精确计算不同参与者子集下的效用函数值,从而获得高精度的 Shapley 值来准确地、公平地衡量参与者的贡献;从可解释性算法分类的角度来看,提出的 QT Shapley 算法属于事后可解释性算法、特征归因算法以及全局解释算法。

第3章 基于合作博弈论的图像分类模型预测可解释性研究

3.1 引言

在图像输入的特征归因领域中,已经有了较为丰富的研究和广泛的实际应用案例^[38]。在这众多算法中,类别激活图(Class Activation Mapping, CAM)算法^[1,2,39-46]通过生成直观的视觉解释,为理解以图像为输入的深度神经网络的预测结果提供了一个重要途径。这种算法能够以热力图的形式展示模型在进行预测时最关注的图像区域,使得模型的推理过程更加透明和可理解。实际上,绪论中的图1-2便是由 GradCAM 算法生成,说明了 CAM 类算法在验证模型预测的合理性、发现潜在偏见、优化模型设计^[1,39]的重要实践意义。从应用广度来看, CAM 类算法可适配多种视觉任务网络,包括图像分类、目标检测、语义分割乃至视觉问答等^[46]。本章将针对最基本的图像分类模型的特征归因展开研究。

尽管以 GradCAM 为代表的 CAM 类算法已经被广泛应用于神经网络预测结果的可解释性研究中,但其算法原理和理论基础却一直没有得到充分的理解。现有的主流算法如 CAM^[2]、GradCAM^[1]、HiResCAM^[39]、GradCAM++^[42]、Layer-CAM^[44]大多依赖启发式策略来调整各个激活图的权重,缺乏系统的理论支撑。这种理论基础的缺失导致算法的可靠性难以保证,解释结果的准确性也难免随之受到质疑^[39]。此外,理论框架的缺失使得现有 CAM 算法的改进方案难以超越经验驱动的优化模式,阻滞了该领域的进一步发展。

与此同时,合作博弈论中的 Shapley 值^[47]为公平量化特征贡献^[48]提供了完善的理论框架,并已在可解释人工智能研究中得到广泛应用(如 SHAP^[3]与 Data Shapley^[30])。然而,这些基于 Shapley 值的算法在实际应用中面临着严峻挑战:严格计算 Shapley 值的计算复杂度是合作博弈参与者数量的指数级别,这使得它们在处理高维数据时遇到了严重的可扩展性瓶颈^[37]。因此,如何在维持理论严谨性的同时确保算法的实用性,成为当前可解释性算法研究亟待解决的关键科学问题。

为弥补 CAM 类算法理论基础的不足,本章提出内容保留的博弈论解释器的理论框架,将 CAM 的计算高效性与 Shapley 值的理论严谨性相结合,并在此框架下提出 ShapleyCAM 算法。该算法旨在弥合启发式驱动的 CAM 类算法与计算密集型的 Shapley 值类算法之间的差距,从而同时确保特征归因的公平性与可扩展性。本章主要贡献可总结为以下几点:

(1) 提出内容保留的博弈论解释器的理论框架:引入 Shapley 值解释 CAM 类算法,阐明了 GradCAM(满足内容保留属性)与 HiResCAM(满足博弈论属性)的理论基础,从而建立了特征归因两大主流算法(CAM 和 SHAP)之间的联系。

(2) 提出提出了一种针对图像分类模型预测可解释性的特征归因算法 ShapleyCAM:在内容保留的博弈论解释器框架下,基于效用函数的二阶近似,利用

神经网络的梯度与海森矩阵信息开发了 ShapleyCAM，从而生成更精确的解释。

(3) 提出 ReST 效用函数：从理论上分析了 pre-softmax 与 post-softmax 分数在生成热力图时的优劣，揭示其理论等价性，并提出 ReST 效用函数以克服其局限性。

(4) 实验验证：在 ImageNet 验证集上，针对 12 种网络架构和 6 项指标进行了全面实验验证，将现有基于梯度的 CAM 算法与 ShapleyCAM 及其变体进行了对比，结果表明结合梯度与海森矩阵信息的 ShapleyCAM 算法能够生成更精确的解释。

3.2 相关工作

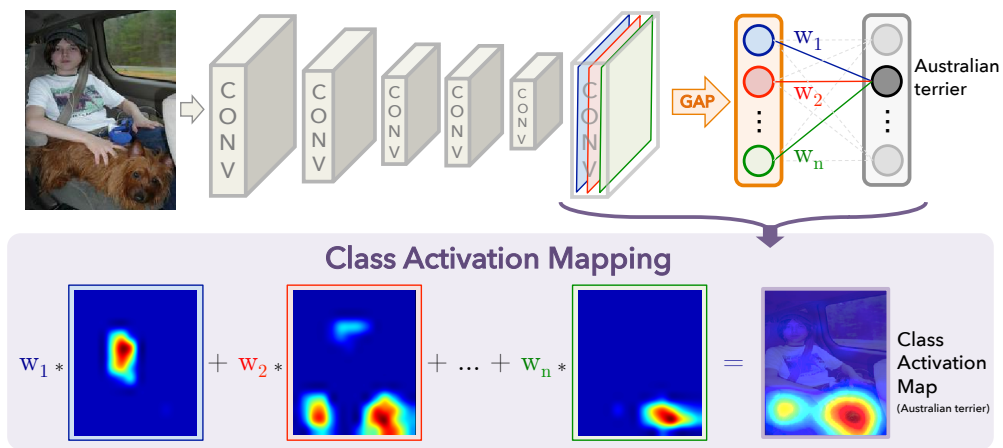


图 3-1 CAM^[2] 类算法通过线性组合 GAP 层之前的激活图以可视化预测的依据
Figure 3-1 CAM methods^[2] visualize the prediction rationale by linearly combining activation maps before the GAP layer

本节首先介绍不同的类别激活图算法，随后详细阐述 SHAP 算法的原理与局限。

原始类别激活图算法^[2]是作为一种弱监督目标定位算法提出的。其核心发现是卷积神经网络中全局平均池化层（Global Average Pooling, GAP）的前一层的不同通道的输出（被称为激活图，activation map，也被称为特征图，feature map）能够体现图像分类网络对图像输入的不同内容的关注，因此利用输出类别就能实现对图像中对应内容的定位。例如图 3-1 中，第一个激活图数值较高的区域位置上对应着原图中人的图像内容，第二个激活图数值较高的区域位置上主要对应着原图中澳大利亚梗的图像内容，并且不同激活图数值较高的像素点在空间位置上较为连续。假设要定位类别为澳大利亚梗的图像内容，CAM 算法会根据目标类别来给不同激活图分配不同权重，然后加和得到一张热力图，将其与原图重合即可将澳大利亚梗的图像内容高亮出来，这便在图像分类网络上完成了目标定位的任务。

之后，由于 CAM 算法可通过高亮对神经网络预测结果至关重要的图像内容

来实现可解释性，其也被当作解释 CNN 决策过程的基础性技术。CAM 算法的基本原理是，当神经网络对图像 x 进行类别预测后，CAM 会为目标类别 c 生成一个热力图 $\text{CAM}_c(x)$ 。该热力图通过对目标层 l （通常是 GAP 层之前的层或最后一个卷积层，其中 N_l 表示层 l 的通道数）的激活图 $\{A^i\}_{i=1}^{N_l}$ 进行线性组合来生成，线性系数 $\{w^i\}_{i=1}^{N_l}$ 为 GAP 层之后全连接层中与目标类别相关的权重。公式如下：

$$\text{CAM}_c(x) = \sum_{i=1}^{N_l} w^i A^i. \quad (3-1)$$

随后，热力图被归一化并上采样至原始图像尺寸，然后叠加到原始图像上，完成对神经网络预测结果的可视化解释。这种后处理步骤已成为后续工作的标准操作，因此在以下内容中将不再赘述。而需要注意的是，全连接层会输出每个类别的 logits $\{y^c\}_{c=1}^C$ （即 softmax 之前的分数），通常通过 softmax 函数将其转换为概率 $\{p^c\}_{c=1}^C$ （即 softmax 之后的分数）。

原始的 CAM 算法仅适用于网络存在 GAP 层且 GAP 后接全连接层作为分类器的网络架构^[38]。GradCAM^[1] 通过使用 pre-softmax 分数 y^c 关于激活图 A^i 的梯度 $W^i = \frac{\partial y^c}{\partial A^i}$ 来计算第 i 个激活图的重要性，从而将 CAM 推广到任何 CNN 架构。需要注意的是，GradCAM 通常使用输出 logit y^c 来生成其热力图。具体地，GradCAM 将某个激活图的重要性定义为其对应梯度的平均，并引入 ReLU 操作以消除负值区域，从而生成最终的热力图。用 $\overline{W^i}$ 表示 W^i 的平均值，热力图生成公式如下：

$$\text{GradCAM}_c(x) = \text{ReLU}\left(\sum_{i=1}^{N_l} \overline{W^i} A^i\right). \quad (3-2)$$

HiResCAM^[39] 通过将激活图和对应的梯度图逐元素相乘来提供更忠实、更细粒度的解释，其中 $\odot$ 表示哈达玛积：

$$\text{HiResCAM}_c(x) = \text{ReLU}\left(\sum_{i=1}^{N_l} W^i \odot A^i\right). \quad (3-3)$$

GradCAM Elementwise^[46] 在 HiResCAM 的基础上，在求和之前添加了一个 ReLU 操作：

$$\text{GradCAM-E}_c(x) = \text{ReLU}\left(\sum_{i=1}^{N_l} \text{ReLU}(W^i \odot A^i)\right). \quad (3-4)$$

LayerCAM^[44] 通过在哈达玛积之前对梯度添加 ReLU 操作进一步改进了 HiResCAM：

$$\text{LayerCAM}_c(x) = \text{ReLU}\left(\sum_{i=1}^{N_l} \text{ReLU}(W^i) \odot A^i\right). \quad (3-5)$$

XGradCAM^[43] 的提出是为了增强 GradCAM 的敏感性和一致性属性。它生成热力图的公式如下，其中 $\bar{X}$ 表示 X 的平均值：

$$\text{XGradCAM}_c(x) = \text{ReLU}\left(\sum_{i=1}^{N_l} \frac{\overline{W^i \odot A^i}}{\bar{A}^i} A^i\right). \quad (3-6)$$

GradCAM++^[42] 通过关注正梯度（类似于 LayerCAM）并结合高阶导数来生成热力图。对于具有 GAP 层的 CNN，它提供了权重的闭合解。为了避免计算高阶导数，它使用平方和立方梯度来近似二阶和三阶导数，但这种近似只有在 logits 通过指数函数得到最终概率的神经网络的设置中才有效。

LIFTCAM^[45] 使用 DeepLIFT^[20] 在单次反向传播中评估每个激活图的贡献来估计权重。激活图权重的准确性受限于 DeepLIFT 所提供解释的精度。

以上列出的 CAM 算法都可归类于基于梯度的算法（CAM 也可以被认为是基于梯度的^[1]，其在适用的网络中等价于 GradCAM）。然而，深度网络中的噪声和消失梯度可能会破坏有意义的解释^[40]，节 3.3.4 部分说明了梯度消失的危害。为了解决这个问题，一些无梯度的 CAM 算法被提出。ScoreCAM^[40] 通过将归一化的激活图与输入图像叠加，并通过应用 softmax 到对应于目标类的 N_l 个输出 logits 来确定权重。AblationCAM^[41] 根据每个相关激活图设置为零时目标输出的减少量来分配权重。尽管无梯度的 CAM 算法某些情境下能提供更准确的解释，但需要 N_l 次前向传播，通常比基于梯度的 CAM 算法多花费数百倍的时间和资源^[49]，这极大地限制了其在大数据集上的使用。

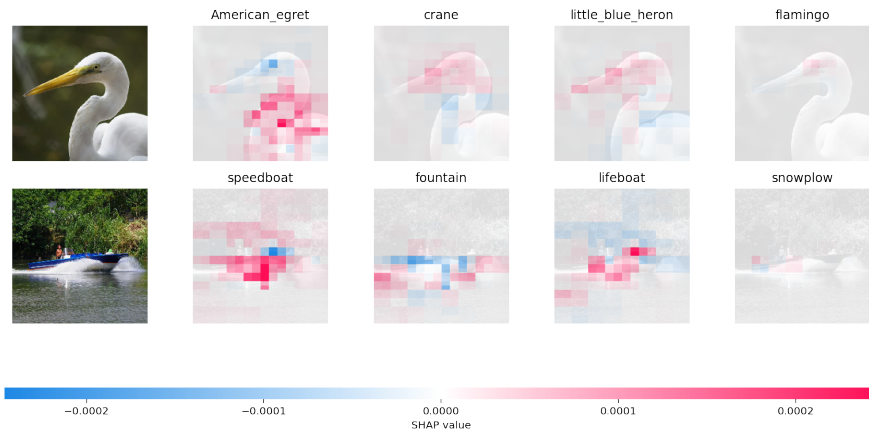


图 3-2 SHAP 在 ResNet-50 上生成的视觉解释^[3]

Figure 3-2 Visual explanations generated by SHAP on ResNet-50^[3]

以 GradCAM 为代表的 CAM 类算法已经被广泛应用于神经网络预测结果的可解释性研究中，但其算法原理和理论基础却一直没有得到充分的理解。与 CAM

类算法不同, SHAP^[3] 算法具备较完善的理论基础。SHAP 通过将模型输出视为效用函数, 将输入特征(图像输入中为像素点或者像素块)视为博弈论中的参与者, 通过近似计算特征的 Shapley 值来解释模型预测结果。该算法能够提供对特征贡献的定量解释, 因而在数据分析和医疗保健领域等领域得到广泛采用^[50]。

然而, SHAP 的计算成本较高。计算精确的 Shapley 值需要进行 $\mathcal{O}(2^n)$ 次模型前向传播, 其中 n 是参与者(即特征)的个数, 计算成本极高。为降低计算复杂度, Kernel SHAP^[3] 等近似算法通过求解加权最小二乘问题进行逼近, 但计算开销仍然较大, 复杂度为 $\mathcal{O}(cn)$, 其中 c 并非可忽略常数。图 3-2 展示了 SHAP 为 ResNet-50 分类结果上给出的解释。技术上, 通过引入遮挡模板, 对图像的不同区域进行随机屏蔽, 并计算对应网络输出的变化, 以此衡量各个图像块(其大小与遮挡模板一致)的重要性。值得注意的是, 图像上的 SHAP 解释算法通常需要数千次前向传播, 例如, 图 3-2 的结果即基于 5000 次前向传播计算所得。相比之下, GradCAM 等基于激活图的可解释性算法计算成本更低, 仅需一次前向传播与一次反向传播即可生成可视化解, 因此在高维数据(如图像)分析任务中更具计算优势, 并且在视觉上往往能提供更友好的解释效果。然而, 虽然 SHAP 计算成本较高, 但其基于的 Shapley 值的数学框架提供了严格的公平性保证, 使其成为可解释人工智能领域的重要基石^[48,51], 而 GradCAM 等算法在理论完备性方面仍存在不足。

综上所述, CAM 算法以其计算效率著称, 并广泛应用于可解释性领域。然而, 它们有一个共同的局限性: 对启发式算法的依赖以及缺乏坚实的理论基础。而基于 Shapley 值的算法则恰恰相反, 其以坚实的理论基础著称, 但计算效率限制了其广泛应用。本章旨在弥合这两类方法的差距。具体地, 本研究将 In-Run Data Shapley^[37] 中通过泰勒展开直接估计效用函数的思路扩展至特征归因领域, 从而推导出 ShapleyCAM 算法。其核心动机是, 如果能够解析地表达模型在不同特征子集的预测结果, 则可以推导出每个特征的 Shapley 值的闭式解, 从而得到在计算上比 SHAP 更高效的特征归因算法。

3.3 基于 Shapley 值的类别激活图算法

本节首先介绍 Shapley 值的原理, 并在其基础上定义内容保留的博弈论解释器。然后使用这一理论框架来澄清 GradCAM 和 HiResCAM 的理论基础。接着在此框架下开发新的 CAM 算法, 即 ShapleyCAM 及其变体。最后对效用函数的选择进行理论分析, 并提出改良的效用函数 ReST。

3.3.1 内容保留的博弈论解释器的定义

Shapley 值^[47] 作为合作博弈论中的核心概念^[51], 现已被广泛应用于机器学习领域, 尤其是在贡献归因和公平性保障方面^[48]。其受到广泛重视的主要原因在于, Shapley 值是唯一能够保证参与者贡献被公平分配的方式——它能够通过评估每个参与者在所有可能的参与者子集中的边际贡献, 公平地将总效用 $U(D)$

(如收入、成本或模型的输出概率) 分配给所有参与者。具体而言, Shapley 值严格遵循一套公理化定义, 是唯一满足以下四个公理的解概念, 从而确保在贡献分配中的公平性:

- 虚拟参与者公理: 如果对于所有 $S \subseteq D \setminus \{i\}$, 有 $U(S \cup \{i\}) = U(S)$, 则 $\phi(i; U) = 0$ 。
- 对称性公理: 如果对于所有 $S \subseteq D \setminus \{i, j\}$, 有 $U(S \cup \{i\}) = U(S \cup \{j\})$, 则 $\phi(i; U) = \phi(j; U)$ 。
- 效率公理: $\sum_{i \in D} \phi(i; U) = U(D) - U(\emptyset)$ 。
- 线性公理: 对于效用函数 U_1, U_2 和任意 $\alpha_1, \alpha_2 \in \mathbb{R}$, 有 $\phi(i; \alpha_1 U_1 + \alpha_2 U_2) = \alpha_1 \phi(i; U_1) + \alpha_2 \phi(i; U_2)$ 。

定义 3.1 (Shapley, 1953^[47]). 给定一个参与者集合 D , 其中 $n = |D|$, 以及一个效用函数 U , 每个参与者 $i \in D$ 的 Shapley 值定义为:

$$\phi(i; U) = \frac{1}{n} \sum_{k=1}^n \binom{n-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{i\} \\ |S|=k-1}} [U(S \cup \{i\}) - U(S)]. \quad (3-7)$$

当且仅当一个单值解概念同时满足虚拟参与者、对称性、效率和线性公理时, 该解概念为 Shapley 值。

Shapley 值的一个更加直观的数学表达形式如下:

$$\phi(i; U) = \mathbb{E}_{\pi \sim \Pi} [U(S_\pi^i \cup \{i\}) - U(S_\pi^i)]. \quad (3-8)$$

其中 $\pi \sim \Pi$ 表示参与者集合 D 的一个均匀随机排列, S_π^i 表示在排列 π 中位于参与者 i 之前的参与者集合。该表达形式清晰揭示了 Shapley 值的核心思想, 即参与者 i 的贡献可通过计算其在所有可能的子集中的期望边际增益来衡量。

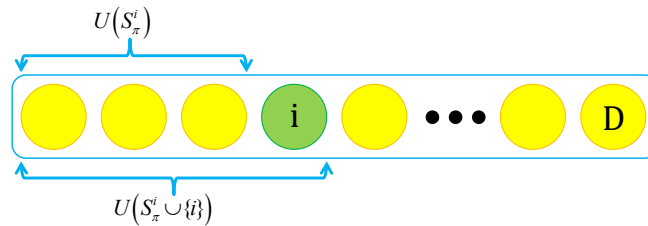


图 3-3 Shapley 值计算过程的示意图

Figure 3-3 Illustration of the calculation process of Shapley value

在特征归因的背景下, Shapley 值确保了总效用在各特征中公平分配。这四个公理实际上从不同角度刻画了贡献分配中对公平性的需求。虚拟参与者公理规定, 对于在任何特征子集中均未产生贡献的特征, 其贡献为零, 以确保归因的合理性。对称性公理要求, 若两个特征在所有可能的特征子集中具有相同的边际贡献, 则它们应获得相同的贡献, 从而保证分配的公平性。效率公理确保总效用被

完整地分配至所有特征。线性公理则强调效用函数的可加性，即若整体效用为多个效用函数的线性组合，则分配给每个特征的贡献亦应为其在这些函数上的贡献的线性组合；式 (3-17) 展示了线性公理在实际应用中的一个案例。

内容保留的博弈论解释器的提出动机是基于合作博弈论中的 Shapley 值理论重新解释 CAM 类算法。SHAP 和 CAM 类算法之间的关键区别在于它们的输入。SHAP 在原始图像上计算单个像素的 Shapley 值，而 CAM 使用在推理过程中从目标层提取的 N 个 d 维激活图 $\{A^i\}_{i=1}^N$ 来生成 d 维解释。

为了阐明“内容保留”的含义与动机，本节将激活图和梯度图的特点进行对比。激活图可以看做是原始图像经过多层滤波器得到的保留了原始空间结构信息的变换结果，且随着网络层数的加深，其反映的特征逐渐由细致的局部信息转为粗糙的整体特征信息^[44]。如图 3-1 所示，GAP 前一层的激活图的高亮区域有以下三个特性：（1）在空间上保持连续。（2）数量稀少。（3）与原始图像中特定内容位置接近。这些特性使得激活图成为连接深度特征与原始图像空间的理想媒介。

相比之下，梯度图作为神经网络输出对激活图的偏导数，具有以下三个缺陷：（1）梯度消失。节 3.3.4 详细地讨论梯度消失的现象和背后原理。（2）梯度噪声。SmoothGrad^[18] 发现梯度图通常包含大量噪声，且在视觉上是嘈杂的；另外，给图片加微小扰动会造成梯度解释不稳定，所以 SmoothGrad 采用多张加噪图片的梯度平均作为解释。而实际应用中，如图 3-7 所示，将梯度图和激活图的哈达玛积生成的解释图往往不能完整地覆盖原始图像中相关内容，其原因正是梯度图的噪声和不连续性破坏了激活图原有的内容完整性。（3）梯度同质化。当选择 GAP 层之前的层作为目标层时，该层所有激活图的梯度均相同^[1]，此时梯度图无法提供任何有意义的图像语义信息。综上所述，相比于梯度图，激活图更能保留原始图像的整体内容信息。因此，在生成解释时，尽可能保留激活图的完整性，有助于获得视觉友好的解释。因此，本章用“内容保留”来指代在生成的解释图中，尽可能保留 $\{A^i\}_{i=1}^N$ ，以完整地反映原始图像特定内容的特性。

本节将 $\{A^i\}_{i=1}^N$ 视为原始图像下采样得到的 N 个 d 维变化结果。神经网络预测结果，被视作一种合作博弈，所有下采样图像的相同位置像素集合作为单个参与者，即 $\{A^i\}_{i=1}^N$ 中相同位置的 d 个组作为 d 个参与者。接下来，本节提出 CAM 算法的理论框架，也即内容保留的博弈论解释器（Content Reserved Game-theoretic Explainer）的正式定义：

定义 3.2. 给定 N 个 d 维向量 $\{A^i\}_{i=1}^N$ ，它们合作得到一个标量效用 $U(D)$ 。所有 A^i 的第 j 个元素的集合（即 $\{A_j^i\}_{i=1}^N$ ），被视为第 j 个参与者，参与者集合为 $D := \{j\}_{j=1}^d$ 。如果第 j 个参与者的 Shapley 值为 $\phi(j; U) = \sum_{i=1}^N W_j^i A_j^i$ ，则内容保留的博弈论解释器生成一个 d 维解释向量 $E = \sum_{i=1}^N g(W^i) \odot A^i$ ，其中 W^i 是一个 d 维向量， g 是一个映射函数， $\odot$ 表示 Hadamard 积。解释器分类如下：

- 类型 I 内容保留的博弈论解释器满足博弈论属性： $g(W^i) = W^i$ 。

- 类型 II 内容保留的博弈论解释器满足内容保留属性: $g(W^i) = \overline{W^i} \mathbf{1}_d$, 其中 $\overline{W^i}$ 是 W^i 的平均值。
- 最优内容保留的博弈论解释器同时满足内容保留和博弈论属性: $\forall i, W^i = \overline{W^i} \mathbf{1}_d$ 。

类型 I 内容保留的博弈论解释器生成的解释, 是对每 d 个参与者在合作博弈中贡献的公平衡量: E 中第 j 个像素的值也是第 j 个参与者的 Shapley 值。技术上, $E_j = [\sum_{i=1}^N W^i \odot A^i]_j = \phi(j; U)$ 。此外, 由于 $\langle W^i \odot A^i, \mathbf{1}_d \rangle = \langle W^i, A^i \rangle$, 有 $\langle E, \mathbf{1}_d \rangle = \langle \sum_{i=1}^N W^i \odot A^i, \mathbf{1}_d \rangle = \sum_{i=1}^N \langle W^i, A^i \rangle = \sum_{j=1}^d \phi(j; U) = U(\{A^i\}_{i=1}^N) - U(\emptyset)$, 其中最后一个等式来自 效率公理。因此, 生成的解释实际上是总效用在每个像素上的一次分配。

类型 II 内容保留的博弈论解释器通过使用标量 $\{\overline{W^i}\}_{i=1}^N$ 对 $\{A^i\}_{i=1}^N$ 进行加权生成解释, 保留了更多 $\{A^i\}_{i=1}^N$ 的完整内容, 尽管可能会牺牲博弈论属性。

最后, 当选择 GAP 层之前的层作为目标层时, 得到最优内容保留的博弈论解释器, 因为此时 GAP 层之前的层上的所有梯度都是相同的, 并且等于 GAP 层中池化分数梯度的 $1/d$ 。详细证明见^[1]。需要特别说明的是, 这里的“最优”表示该解释器同时满足内容保留和博弈论属性, 在内容保留的博弈论解释器的定义下达到了最优。

3.3.2 GradCAM 和 HiResCAM 的理论基础

在内容保留的博弈论解释器的理论框架下, 可以通过一阶泰勒展开近似效用函数来确立 GradCAM 和 HiResCAM 的理论基础。

定理 3.1. 当使用一阶泰勒展开近似效用函数时, *HiResCAM* 是类型 I 内容保留的博弈论解释器, *GradCAM* 是类型 II 内容保留的博弈论解释器。如果目标层是 GAP 层之前的层, *GradCAM* 和 *HiResCAM* 都是最优内容保留的博弈论解释器。

证明. 符号定义: 定义 $X_D \in \mathbb{R}^{1 \times Nd}$ 为 $X_D := [A^1 | A^2 | \dots | A^N]$, 其中 $A^i \in \mathbb{R}^{1 \times d}$ 。设 $X^i \in \mathbb{R}^{1 \times Nd}$ 为在原始位置保留 A^i 并在其他地方为零的向量: $X^i := [\vec{0} | \dots | A^i | \dots | \vec{0}]$, 其中 $\vec{0} \in \mathbb{R}^{1 \times d}$ 。定义 $X_j \in \mathbb{R}^{1 \times Nd}$ 为 $X_j := [0, A_j^1, 0 | \dots | 0, A_j^i, 0 | \dots | 0, A_j^N, 0]$, 仅保留每个 A^i 的第 j 个元素。并设 $X_S := \sum_{j=1, j \in S}^d X_j$, 其中子集 $S \subseteq D = \{j\}_{j=1}^d$ 。效用函数值表示为 $U(X_S)$ 和 $U(X_D)$, 其中 $U'(X_D)$ 为梯度, H_D 为 X_D 处的海森矩阵。 $U(X_S)$ 的泰勒展开包括一阶项 $U_1(X_S)$ 和二阶项 $U_2(X_S)$ 。

推导: 受 In-Run Data Shapley^[37] 启发, 首先对所有 S 在 X_D 处应用效用函数 $U(X_S)$ 的一阶泰勒展开:

$$U(X_S) \approx U(X_D) + \underbrace{U'(X_D)(X_S - X_D)}_{U_1(X_S)}. \quad (3-9)$$

然后也对所有 $S \cup j$ 在 X_D 处应用效用函数 $U(X_S)$ 的一阶泰勒展开:

$$U(X_{S \cup j}) \approx U(X_D) + \underbrace{U'(X_D)(X_S - X_D + X_j)}_{U_1(X_{S \cup j})}. \quad (3-10)$$

接着考虑差值:

$$U(X_{S \cup j}) - U(X_S) = \underbrace{U'(X_D)X_j}_{U_1(X_{S \cup j}) - U_1(X_S)}. \quad (3-11)$$

为了计算 $\phi(j; U)$, 使用公式(3-7), 并将 $U(X_{S \cup j}) - U(X_S)$ 替换为 $U'(X_D)X_j$:

$$\begin{aligned} \phi(j; U) &= \frac{1}{d} \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{j\} \\ |S|=k-1}} (U(X_{S \cup j}) - U(X_S)) \\ &= \frac{1}{d} \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{j\} \\ |S|=k-1}} U'(X_D)X_j \\ &= \frac{1}{d} U'(X_D)X_j \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{j\} \\ |S|=k-1}} 1 \\ &= \frac{1}{d} U'(X_D)X_j \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \binom{d-1}{k-1} \\ &= \frac{1}{d} U'(X_D)X_j \sum_{k=1}^d 1 = U'(X_D)X_j. \end{aligned} \quad (3-12)$$

因此, 在一阶近似下, Shapley 值为:

$$\phi(j; U) = U'(X_D)X_j = \sum_{i=1}^N [U'(X_D)]_j^i A_j^i. \quad (3-13)$$

通过将式 (3-13) 代入内容保留的博弈论解释器的定义, 可以推导出与 A_j^i 相关的系数 W_j^i 表示为 $[U'(X_D)]_j^i$, 即效用函数关于 A_j^i 的梯度。因此, 根据内容保留的博弈论解释器的定义, 定理得证。□

在 GradCAM^[1] 中已经证明, 使用 pre-softmax 分数作为效用函数时, GradCAM 和 CAM 的热力图是相同的, 仅差一个常数因子并在归一化过程中消除了。因此, 有以下推论:

推论 3.2. 当使用 GAP 层之前的层作为目标层并使用 pre-softmax 分数作为效用函数时, CAM^[2] 是最优内容保留的博弈论解释器。

算法 1 ShapleyCAM 及其变体

输入： 输入图像 x ，神经网络 $f(\cdot)$ ，目标类别 c ，目标层 l

输出： 用于解释图像 x 被预测为类别 c 时的热力图

- 1: 前向传播：使用 $f(\cdot)$ 计算输入 x 的 logits y ，并保存目标层 l 的激活图 $\{A^i\}_{i=1}^{N_l}$ 。
将这些映射连接起来得到 $X_D := [A^1 \mid A^2 \mid \dots \mid A^{N_l}]$
 - 2: 计算 ReST 效用函数： $U_{\text{ReST}} = y^c + \ln(\text{softmax}(y)^c)$
 - 3: 反向传播：计算 ReST 效用函数关于激活图的梯度： $U'(X_D) = \partial U_{\text{ReST}} / \partial X_D$
 - 4: 反向传播：计算 ReST 效用函数关于激活图的 Hessian-Vector 乘积： $X_D^\top H_D = X_D^\top \partial^2 U_{\text{ReST}} / \partial X_D^2$
 - 5: 计算权重： $W^i = [U'(X_D) - \frac{1}{2} X_D^\top H_D]^i$
 - 6: **if** 使用 ShapleyCAM **then**
 - 7: **return** $\text{ReLU}(\sum_{i=1}^{N_l} \overline{W^i} A^i)$
 - 8: **else if** 使用 ShapleyCAM-H **then**
 - 9: **return** $\text{ReLU}(\sum_{i=1}^{N_l} (W^i \odot A^i))$
 - 10: **else if** 使用 ShapleyCAM-E **then**
 - 11: **return** $\text{ReLU}(\sum_{i=1}^{N_l} \text{ReLU}(W^i \odot A^i))$
 - 12: **end if**
-

最后，本节从 Shapley 值的角度来回答 CAM 算法中常用的启发式操作 ReLU 位置的放置问题，完成对 CAM 类算法理论基础解释的最后一部分。GradCAM 将 ReLU 作为一种启发式操作引入，以消除解释中的负值区域，并已被后续工作采用为默认操作。LayerCAM、GradCAM++ 和 GradCAM-E 等算法也对梯度或其他部分应用了 ReLU，但这些算法可能会产生有问题的解释，如图 3-8 所示。而在内容保留的博弈论解释器的理论框架下，ReLU 位置的放置问题可以从 Shapley 值的角度来回答：热力图中的每个像素代表其 Shapley 值，Shapley 值大于 0 表示对模型输出效用的正贡献^[33]，应被高亮显示。因此，ReLU 最合理的放置位置是在求和之外，这与 GradCAM 和 HiResCAM 一致。

3.3.3 ShapleyCAM 算法的设计与实现

在内容保留的博弈论解释器中，通过二阶泰勒展开近似效用函数可以得到 ShapleyCAM 及其变体。具体地，将式 (3-20) 中的 $W_j^i = [U'(X_D) - \frac{1}{2} X_D^\top H_D]_j^i$ 应用于公 (3-2)、(3-3) 和 (3-4)（分别对应 GradCAM、HiResCAM 和 GradCAM-E），可以推导出 ShapleyCAM、ShapleyCAM-H 和 ShapleyCAM-E 算法。算法 1 展示了这三种算法的详细过程。

定理 3.3. 当使用二阶泰勒展开近似效用函数时，*ShapleyCAM-H* 是类型 I 内容保留的博弈论解释器，*ShapleyCAM* 是类型 II 内容保留的博弈论解释器。如果目标层是 GAP 层之前的层，*ShapleyCAM* 和 *ShapleyCAM-H* 都是最优内容保留的博弈论解释器。

证明. 符号定义: 同定理3.1。推导: 受 In-Run Data Shapley^[37] 启发, 对所有 S 在 X_D 处应用效用函数 $U(X_S)$ 的二阶泰勒展开:

$$\begin{aligned} U(X_S) &\approx U(X_D) + \underbrace{U'(X_D)(X_S - X_D)}_{U_1(X_S)} \\ &\quad + \underbrace{\frac{1}{2}(X_S - X_D)^\top H_D(X_S - X_D)}_{U_2(X_S)}. \end{aligned} \quad (3-14)$$

然后也对所有 $S \cup j$ 在 X_D 处应用效用函数 $U(X_S)$ 的二阶泰勒展开:

$$\begin{aligned} U(X_{S \cup j}) &\approx U(X_D) + \underbrace{U'(X_D)(X_S - X_D + X_j)}_{U_1(X_{S \cup j})} \\ &\quad + \underbrace{\frac{1}{2}(X_S - X_D + X_j)^\top H_D(X_S - X_D + X_j)}_{U_2(X_{S \cup j})}. \end{aligned} \quad (3-15)$$

接着考虑差值:

$$\begin{aligned} U(X_{S \cup j}) - U(X_S) &= \underbrace{U'(X_D)X_j}_{U_1(X_{S \cup j}) - U_1(X_S)} + \underbrace{\frac{1}{2}X_j^\top H_D X_j - X_D^\top H_D X_j + X_S^\top H_D X_j}_{U_2(X_{S \cup j}) - U_2(X_S)}. \end{aligned} \quad (3-16)$$

利用 Shapley 值的线性公理, 可以将 U 下的 Shapley 值分解为 U_1 和 U_2 下的 Shapley 值之和:

$$\phi(j; U_1 + U_2) = \phi(j; U_1) + \phi(j; U_2). \quad (3-17)$$

类似于公式(3-12), 将 Shapley 值中的 $U_2(X_{S \cup j}) - U_2(X_S)$ 替换为 $\frac{1}{2}X_j^\top H_D X_j - X_D^\top H_D X_j + X_S^\top H_D X_j$:

$$\begin{aligned} \phi(j; U_2) &= \frac{1}{d} \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{j\} \\ |S|=k-1}} X_S^\top H_D X_j \\ &\quad + \frac{1}{d} \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{j\} \\ |S|=k-1}} (\frac{1}{2}X_j^\top H_D X_j - X_D^\top H_D X_j) \\ &= \frac{1}{d} \sum_{k=2}^d \binom{d-1}{k-1}^{-1} \sum_{i \in D \setminus j} \sum_{\substack{S \subseteq D \setminus \{i, j\} \\ |S|=k-2}} X_i^\top H_D X_j \\ &\quad + \frac{1}{d} \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{j\} \\ |S|=k-1}} (\frac{1}{2}X_j^\top H_D X_j - X_D^\top H_D X_j), \end{aligned}$$

上述推导的最后一步依赖于 $X_S = \sum_{i \in S} X_i$ 的性质，将不同 S 的集合的和拆分为这些子集中的单个元素 X_i 的和。于是有：

$$\begin{aligned}
 \phi(j; U_2) &= \frac{1}{d} \sum_{k=2}^d \binom{d-1}{k-1}^{-1} \sum_{i \in D \setminus j} \binom{d-2}{k-2} X_i^\top H_D X_j \\
 &\quad + \frac{1}{d} \left(\frac{1}{2} X_j^\top H_D X_j - X_D^\top H_D X_j \right) \sum_{k=1}^d \binom{d-1}{k-1}^{-1} \binom{d-1}{k-1} \\
 &= \frac{1}{d} \sum_{k=2}^d \binom{d-1}{k-1}^{-1} \binom{d-2}{k-2} \left(\sum_{i \in D \setminus j} X_i^\top H_D X_j \right) \\
 &\quad + \frac{1}{d} \left(\frac{1}{2} X_j^\top H_D X_j - X_D^\top H_D X_j \right) \sum_{k=1}^d 1 \\
 &= \frac{1}{d} \sum_{k=2}^d \frac{k-1}{d-1} (X_{D \setminus j}^\top H_D X_j) + \left(\frac{1}{2} X_j^\top H_D X_j - X_D^\top H_D X_j \right) \\
 &= \frac{\sum_{k=2}^d (k-1)}{d(d-1)} X_{D \setminus j}^\top H_D X_j + \frac{1}{2} X_j^\top H_D X_j - X_D^\top H_D X_j \\
 &= \frac{1}{2} X_{D \setminus j}^\top H_D X_j + \frac{1}{2} X_j^\top H_D X_j - X_D^\top H_D X_j \\
 &= \frac{1}{2} X_D^\top H_D X_j - X_D^\top H_D X_j \\
 &= -\frac{1}{2} X_D^\top H_D X_j.
 \end{aligned} \tag{3-18}$$

因此，在效用函数 U_2 下的 Shapley 值为：

$$\phi(j; U_2) = -\frac{1}{2} X_D^\top H_D X_j = \sum_{i=1}^N \left[-\frac{1}{2} X_D^\top H_D \right]_j^i A_j^i. \tag{3-19}$$

公式(3-12)也表示 $\phi(j; U_1)$ 的计算。通过将公式(3-12)和(3-18)代入公式(3-17)，可以推导出在二阶近似下 $\phi(j; U)$ 的最终表达式：

$$\begin{aligned}
 \phi(j; U) &= U'(X_D) X_j - \frac{1}{2} X_D^\top H_D X_j \\
 &= \sum_{i=1}^N [U'(X_D) - \frac{1}{2} X_D^\top H_D]_j^i A_j^i.
 \end{aligned} \tag{3-20}$$

通过将公式(3-20)代入内容保留的博弈论解释器，得到 A_j^i 相关的系数 W_j^i 可以表示为 $W_j^i = [U'(X_D) - \frac{1}{2} X_D^\top H_D]_j^i$ 。因此，根据内容保留的博弈论解释器的定义，定理得证。 $\square$

ShapleyCAM 与其它基于 Shapley 值的特征归因方法的区别如下：之前的工作如 LIFTCAM^[45] 同样试图将 Shapley 值与 CAM 结合，其使用 DeepLIFT^[20] 计算得到的 Shapley 值对激活图进行加权。相比之下，ShapleyCAM 利用直接在激活图的相同位置计算 Shapley 值，即将下采样 d 维原始图像的像素视为参与者，更接近 SHAP 的建模方式。然而，SHAP 旨在解释整个高度复杂的模型；而 ShapleyCAM 将关注点缩小到目标层和输出效用函数之间的层，而将输入层到目标层对原图的处理看作是对原图的保留了空间位置信息的特征变化过程。此外，ShapleyCAM 使用推导出的闭式 Shapley 值，避免了 SHAP 所需在图像上所需的上千次前向传播。

ShapleyCAM 和 GradCAM 的不同之处仅在于的高阶泰勒项 $-\frac{1}{2}X_D^T H_D$ 。实际上，对于大多数卷积神经网络，使用 pre-softmax 作为效用函数计算得到的 Hessian 矩阵通常为零，这是由于 ReLU 和卷积等操作本质上是线性的，此时高阶泰勒项不起作用，ShapleyCAM 也就等价于 GradCAM。然而，当以下两种情况高阶泰勒项会发挥作用：（1）目标层和输出层之间的网络中存在非线性激活函数（如 EfficientNet 和 Swin Transformer）时，Hessian 才可能为非零矩阵（2）效用函数是 post-softmax 或者下一节提出的 ReST 时，这些效用函数本身就是非线性的，此时 ShapleyCAM 和 GradCAM 所生成的可视化结果可能存在显著差异。

ShapleyCAM 算法是计算高效的。尽管海森矩阵计算通常被认为是计算密集型的，但 ShapleyCAM 只需要 H_D 和 X_D 的 Hessian-Vector 积，这在现代深度学习框架如 PyTorch 和 JAX 中得到了高效支持^[52]，只需要一次反向传播就能实现。实际上，这使得 ShapleyCAM 只需比 GradCAM 多一次反向传播^[52]，使其在大规模数据集上也具有可扩展性。相同测试环境下，在 50,000 张图像的 ILSVRC2012 数据集上以 ResNet-50 为网络架构，ShapleyCAM 和 GradCAM 分别只需 40 分钟和 30 分钟便能为所有图片生成解释。而 SHAP 算法在一张图像上解释则依赖于上千次前向传播，如图 3-2 的解释结果需要 5000 次前向传播，在开启并行运算的条件下也需要几分钟才能生成解释。以图 3-2 的解释生成为例，从计算量（FLOPs 为单位，浮点运算数）角度来看，ShapleyCAM 算法需要一次前向传播和两次反向传播，而反向传播计算量通常按照前向传播计算量的两倍来估计，于是 ShapleyCAM 需要 5 倍前向传播计算量，GradCAM 需要 3 倍前向传播计算量，而 SHAP 算法需要 5000 倍前向传播计算量，因此 ShapleyCAM 的计算效率相当于是 SHAP 的 1000 倍。

3.3.4 效用函数的优化设计

效用函数的选择是合作博弈论的核心。本节分析了两种效用函数的选择（pre-softmax 和 post-softmax 分数）在解释中的优势和局限性，接着从理论上揭示了基于 post-softmax 分数生成的热力图和基于 pre-softmax 分数生成的热力图之间的等价性，并提出了 ReST 效用函数以克服这些局限性。

在生成解释时，选择使用 pre-softmax 还是 post-softmax 分数仍然存在争

议^[53]。虽然大多数算法^[1,2,42]使用 pre-softmax 分数，但尚未达到共识。此外，使用 pre-softmax 分数通常会导致非目标类别的区域也被高亮显示。如图 3-4 左上角所示，当目标类别为虎猫时，使用 pre-softmax 的 GradCAM 错误地高亮了拳师犬。但使用 post-softmax 分数则可以避免这种现象。如图 3-4 中上部所示，使用 post-softmax 的 GradCAM 不会高亮拳师犬。此外，如果在 post-softmax 后应用 $\ln$ 函数^[53]得到的效用函数与模型训练期间使用的交叉熵损失相匹配。

为了理解 pre-softmax 和 post-softmax 之间的优劣，接下来从理论角度探究 GradCAM 在分别使用 pre-softmax 和 post-softmax 分数作为效用函数时生成的热力图之间的关系。

假设 GradCAM 使用目标层 l 的激活图 $\{A^i\}_{i=1}^{N_l}$ （无论选择哪一层）和梯度 $\frac{\partial y^c}{\partial A^i}$ 或 $\frac{\partial p^c}{\partial A^i}$ 来生成热力图 E_c^{pre} 或 E_c^{post} 。这里， y^c 表示目标类别 $c \in [1, \dots, C]$ 的 logit， $p^c = \text{softmax}(y)^c$ 表示相应的概率，则以下定理成立：

定理 3.4. 基于 *post-softmax* 分数生成的 GradCAM 热力图等价于基于 *pre-softmax* 分数生成的 C 个类别的 GradCAM 热力图的加权集成。

证明. 将 GradCAM 在使用 post-softmax 分数时生成的热力图记为：

$$E_c^{\text{post}} = \frac{1}{N_l} \sum_{i=1}^{N_l} g\left(\frac{\partial p^c}{\partial A^i}\right) \odot A^i \quad (3-21)$$

将 GradCAM 在使用 pre-softmax 分数时生成的热力图记为：

$$E_c^{\text{pre}} = \frac{1}{N_l} \sum_{i=1}^{N_l} g\left(\frac{\partial y^c}{\partial A^i}\right) \odot A^i \quad (3-22)$$

其中 $g(\cdot)$ 为 GradCAM 将 d 维向量 A 的每个元素替换为其均值的操作：

$$g(A) = \frac{\langle A, \mathbf{1}_d \rangle}{d} \mathbf{1}_d \quad (3-23)$$

使用链式法则来计算梯度 $\frac{\partial p^c}{\partial A^i}$ ：

$$\frac{\partial p^c}{\partial A^i} = \sum_{k=1}^C \frac{\partial p^c}{\partial y^k} \frac{\partial y^k}{\partial A^i} \quad (3-24)$$

利用 softmax 导数性质：

$$\frac{\partial p^c}{\partial y^k} = \begin{cases} p^c(1 - p^c), & k = c \\ -p^c p^k, & k \neq c \end{cases} \quad (3-25)$$

可得梯度表达式：

$$\frac{\partial p^c}{\partial A^i} = (1 - p^c)p^c \frac{\partial y^c}{\partial A^i} - \sum_{k \neq c} p^c p^k \frac{\partial y^k}{\partial A^i} \quad (3-26)$$

应用概率归一化条件 $\sum_{k=1}^C p^k = 1$ ，可进一步整理为：

$$\frac{\partial p^c}{\partial A^i} = p^c \sum_{k \neq c} p^k \left(\frac{\partial y^k}{\partial A^i} - \frac{\partial y^c}{\partial A^i} \right) \quad (3-27)$$

注意到 $g(A)$ 是 A 的线性变换，这是因为 $g(A)$ 满足：

$$g(A_1 + A_2) = \frac{\langle A_1 + A_2, \mathbf{1}_d \rangle}{d} \mathbf{1}_d = g(A_1) + g(A_2) \quad (3-28)$$

$$g(\lambda A) = \frac{\langle \lambda A, \mathbf{1}_d \rangle}{d} \mathbf{1}_d = \lambda g(A) \quad (3-29)$$

因此就有：

$$g \left(\frac{\partial p^c}{\partial A^i} \right) = p^c \sum_{k \neq c} p^k \left[g \left(\frac{\partial y^k}{\partial A^i} \right) - g \left(\frac{\partial y^c}{\partial A^i} \right) \right] \quad (3-30)$$

再将上式代入 E_c^{post} 表达式展开：

$$\begin{aligned} E_c^{\text{post}} &= \frac{1}{N_l} \sum_{i=1}^{N_l} \left(p^c \sum_{k \neq c} p^k \left[g \left(\frac{\partial y^k}{\partial A^i} \right) - g \left(\frac{\partial y^c}{\partial A^i} \right) \right] \right) \odot A^i \\ &= p^c \sum_{k \neq c} p^k \left(\frac{1}{N_l} \sum_{i=1}^{N_l} g \left(\frac{\partial y^k}{\partial A^i} \right) \odot A^i - \frac{1}{N_l} \sum_{i=1}^{N_l} g \left(\frac{\partial y^c}{\partial A^i} \right) \odot A^i \right) \\ &= p^c \sum_{k \neq c} p^k (E_k^{\text{pre}} - E_c^{\text{pre}}) \end{aligned} \quad (3-31)$$

最终得到分别使用 pre-softmax 分数和 post-softmax 分数的 GradCAM 生成的热力图之间的联系：

$$E_c^{\text{post}} = p^c \sum_{k \neq c} p^k (E_k^{\text{pre}} - E_c^{\text{pre}}) \quad (3-32)$$

□

实际上，如果 $g(X)$ 是线性变换，本节的证明可以推广到其他 CAM 算法。例如，定理 3.4 也适用于 HiResCAM，其中 $g(X) = X$ 。

定理 3.4 首先揭示了 Grad-CAM 在基于 pre-softmax 分数和 post-softmax 分数分别生成的热力图之间的联系。具体而言，该定理表明，基于 post-softmax 分数生成的热力图实际上等价于基于 pre-softmax 分数生成的目标类别的热力图与基于 pre-softmax 分数生成的其他类别的热力图之间的加权差。

此外，定理 3.4 还进一步阐明了 post-softmax 在部分情况优于 pre-softmax 的原因。尽管 E_c^{pre} 可能会突出显示属于另一个类别 b 的区域，但通过减去 E_b^{pre} ，可以有效地消除这些区域的干扰，从而更准确地突出目标类别的特征。

最后，定理 3.4 还揭示了使用 post-softmax 分数的一个局限性：当神经网络对其预测非常自信时（即 p^c 接近 1），其他类别概率的梯度变得非常小（即当

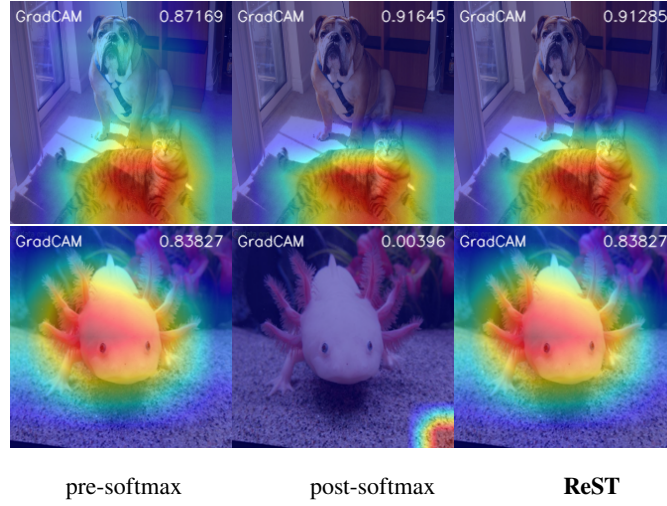


图 3-4 ResNet-18 上 GradCAM 分别结合 pre-softmax、post-softmax 和 ReST 效用函数得到的热力图，目标层为 GAP 层前的层

Figure 3-4 GradCAM on ResNet-18 using pre-softmax, post-softmax, and ReST with the layer preceding GAP as the target layer

$k \neq c$ 时, p^k 也接近 0), 生成的热力图 E_c^{post} 却出人意料地趋近于全零矩阵 (如公式 (3-32) 所示)。这种现象通常被称为梯度消失^[53]。例如, 在图 3-4 的第二行中, ResNet-18 以大于 $1 - 10^{-15}$ 的概率预测此图像类别为 蝾螈; 但 post-softmax 由于梯度消失而无法高亮 蝾螈, 而 pre-softmax 则没有这个问题, 正确地高亮了 蝾螈。

为了克服 pre-softmax 和 post-softmax 在不同场景各有利弊的问题, 受残差思想^[54] 的启发, 本节提出了同时使用了 pre-softmax 和 post-softmax 分数的 ReST (Residual Softmax Target-Class) 函数作为效用函数:

$$U_{\text{ReST}} = y^c + \ln(\text{softmax}(y)^c). \quad (3-33)$$

和定理3.4中的推导类似, 可以得到使用 ReST 的 GradCAM 生成的热力图为:

$$E_c^{\text{ReST}} = E_c^{\text{pre}} + \sum_{k=1, k \neq c}^C p^k (E_c^{\text{pre}} - E_k^{\text{pre}}). \quad (3-34)$$

式(3-34)说明 ReST 实际上通过引入额外的 E_c^{pre} 项来减轻梯度消失的负面影响, 且同时保留了 post-softmax 的仅对目标类别区域的高亮的特性。如图 3-4 右侧列所示, 使用 ReST 或 post-softmax 的 GradCAM 仅仅高亮了 虎猫的区域, 而 pre-softmax 错误地高亮了 拳师犬。此外, 当 post-softmax 由于梯度消失而失效时, 使用 ReST 或 pre-softmax 的 GradCAM 仍然可以正确地高亮 蝾螈区域。

因此, ReST 可以被视为一种结合了 pre-softmax 和 post-softmax 分数优点并避免其各自缺点的效用函数的计算方式。因此, ShapleyCAM 及其变体均采用 ReST 作为效用函数, 且除 ReST 消融实验外, 其他所有实验的效用函数也都设为 ReST。

3.4 实验结果与分析

3.4.1 实验设置

本节评估了十二种不同的神经网络架构上的各种 CAM 算法，利用两种目标层和六种指标来评估解释的质量。所有实验均在配备 Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz 和 NVIDIA A40 GPU 的服务器上进行。

数据集：与以往研究在随机选择的图像上评估 CAM 不同，本节的实验在完整的 ImageNet 验证集（ILSVRC2012）^[55] 上进行，该数据集包含 50,000 张图像，涵盖 1,000 个不同的对象类别。每张图像被调整大小并中心裁剪为 224×224 像素，随后使用从 ImageNet 训练集计算出的均值和标准差进行归一化。

神经网络：为了全面评估 CAM 算法，本节使用以下这些流行的神经网络：ResNet-18、ResNet-50、ResNet-101、ResNet-152^[54]、ResNeXt-50^[56]、VGG-16^[57]、EfficientNet-B0^[58]、MobileNet-V2^[59] 以及 Swin Transformer 的 Tiny、Small、Base 和 Large 配置^[60]。所有网络权重（IMAGENET1K_V1）均直接从 PyTorch 和 Timm^[61] 获取。这些网络在 ILSVRC2012 上的 top-1 准确率附加在表3-1、3-2、3-3和3-4中的网络名称后。

对比算法：本节在评估中采取了最先进的（同时是最通用的）基于梯度的 CAM 算法作为对比：GradCAM^[1]、HiResCAM^[39]、GradCAM-E^[46]、LayerCAM^[44]、XGradCAM^[43]、GradCAM++^[42]、RandomCAM^[46] 以及提出的 ShapleyCAM、ShapleyCAM-H、ShapleyCAM-E。其中 RandomCAM 作为基线，为每个激活图生成一个在 $[-1, 1]$ 范围内的随机标量作为对应的权重。由于梯度无关算法的运行时间较长，实验中排除了这些算法^[38,49]。例如，ScoreCAM 在 ILSVRC2012 上使用 ResNet-50 需要超过 32 小时，而 ShapleyCAM 和 GradCAM 分别只需 40 分钟和 30 分钟。

效用函数设置：除 ReST 的消融研究外，所有实验均使用 ReST 作为效用函数。

目标类别设置：以往的研究通常使用真实标签作为目标类别^[39,43]，或仅关注正确预测的图像^[42,44]。这些设置在网络预测准确的情况下是合理的，解释器的性能通过识别输入图像中的相应证据来衡量。然而，用户也可能希望在模型预测错误时获得反馈。使用真实标签的设置实际上在评估解释精度时引入了模型准确性的混淆因素。可解释算法应专注于揭示模型的决策过程，帮助用户理解其推理过程，无论预测是否正确。

如图3-5左上角所示，当 ResNet-18 错误地将虎鲨预测为锤头鲨时，相应的热力图突出了虎鲨的尾部，而不是通常的、如图3-5右列所示的整个身体。在这种情况下，用户自然会怀疑预测的正确性，并可能推断模型的错误是由于虎鲨的尾部与锤头鲨的相似性所致。此外，在实际场景中，通常只能访问模型的预测结果。因此，除非另有说明，所有实验中使用**预测标签**作为目标类别。

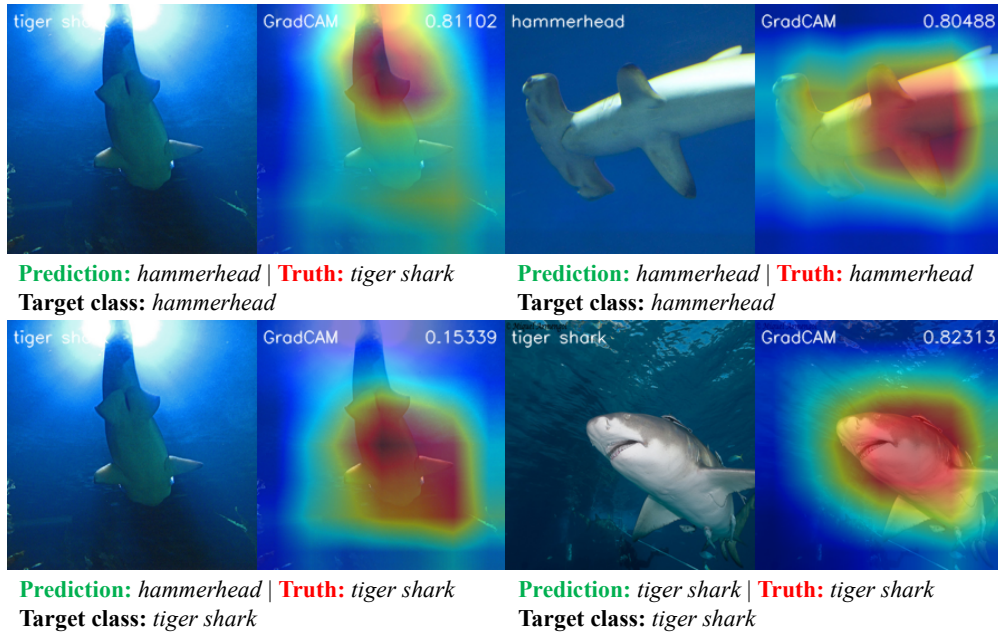


图 3-5 ResNet-18 上 GradCAM 生成的热力图，使用 ReST 和 GAP 之前的层。左：ResNet-18 错误地将虎鲨预测为锤头鲨。右：ResNet-18 正确预测了虎鲨和锤头鲨

Figure 3-5 GradCAM on ResNet-18 using ReST with the layer preceding GAP as the target layer. Left: ResNet-18 incorrectly predicts a *tiger shark* as a *hammerhead*. Right: ResNet-18 accurately predicts the *tiger shark* and the *hammerhead*

3.4.2 评估指标

为了全面评估基于 CAM 的解释，本节采用了 6 种评估指标^[42,45,62]来衡量解释质量的不同方面。所有这些指标都基于 CAM 突出区域被聚焦或屏蔽时预测置信度的变化以及解释的视觉质量来量解释性能。此外，实验中排除了某些定位指标，如交并比（IoU），因为如第3.4.5节所述，它们关注的是定位性能而非解释性。

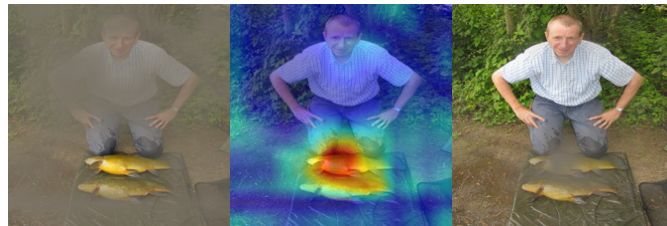


图 3-6 VGG-16 上使用 ShapleyCAM 和最后一个卷积层生成的解释图、视觉解释和反解释图

Figure 3-6 Explanation map, visual explanation, and anti-explanation map generated by ShapleyCAM using the last convolutional layer of VGG-16

为了便于理解这些指标，本节定义了一些术语：将 CAM 类算法生成的归一化并上采样后的热力图记作 $H_c(x)$ ；将它与原始图像 x 线性组合的图片称作视觉解释；将热力图与原始图像的 Hadamard 积得到的图 $x \odot H_c(x)$ 称作解释图，解释图可以理解为将原图屏蔽不重要的像素后得到的图；反之，屏蔽重要像素则得

到反解释图 $x \odot (1 - H_c(x))$ 。图3-6给出了解释图、视觉解释和反解释图的示例。

- **Average Drop (AD)**^[42] 量化了使用解释图而非原始输入图像时目标类别的模型置信度的平均下降程度：

$$AD = \frac{1}{N} \sum_{i=1}^N \frac{\max(0, y_i^c - o_i^c)}{y_i^c}. \quad (3-35)$$

其中 y_i^c 和 o_i^c 分别是模型在使用完整图像和解释图时对类别 c 的 softmax 后分数，一共有 N 个样本。较低的 AD 表示解释图捕捉到了最相关的特征，这些特征的存在有助于保持模型的置信度。

- **Coherency (Coh)**^[62] 通过计算原始热力图与从解释图生成的热力图之间的归一化 Pearson 相关系数来评估 CAM 算法的一致性，Coh 越高说明 CAM 算法的解释越自洽：

$$Coh(x) = \frac{1}{2} \frac{\text{Cov}(H_c(x \odot H_c(x)), H_c(x))}{\sigma_{H_c(x \odot H_c(x))} \sigma_{H_c(x)}} + \frac{1}{2}. \quad (3-36)$$

- **Complexity (Com)**^[62] 通过计算 L1 范数来评估解释图的简洁性，越低说明解释越简洁：

$$Com(x) = \|H_c(x)\|_1. \quad (3-37)$$

- **Average DCC (ADCC)**^[62] 是平均下降、一致性（所有样本的平均值）和复杂度（所有样本的平均值）的调和平均值，用于提供解释质量的整体衡量：

$$ADCC(x) = \frac{3}{\frac{1}{Coh(x)} + \frac{1}{1-Com(x)} + \frac{1}{1-AD(x)}}. \quad (3-38)$$

较高的 ADCC 分数反映了 CAM 算法在保持模型置信度、提供一致解释和确保简洁性三方面综合表现更佳。

- **Increase in Confidence (IC)**^[42] 衡量使用解释图而非原始图像时模型置信度增加的样本比例，其中 $\mathbb{I}$ 是指示函数，越高的 IC 分数说明 CAM 算法越能找到和类别最相关的特征：

$$IC = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i^c < o_i^c). \quad (3-39)$$

- **Average Drop in Deletion (ADD)**^[45] 评估使用反解释图时的模型置信度下降的样本比例：

$$ADD = \frac{1}{N} \sum_{i=1}^N \frac{\max(0, y_i^c - d_i^c)}{y_i^c}. \quad (3-40)$$

其中， d_i^c 是模型在使用反解释图时对类别 c 的 softmax 后分数。较高的 ADD 分数表明有效识别了关键特征，因为删除这些像素会导致置信度显著下降。

这些指标旨在从不同角度衡量 CAM 算法的解释质量。然而，正如^[62]指出的，这些指标本身存在潜在偏差，可能无法全面捕捉解释图的语义有效性。以

Fake-CAM 算法为例，该方法生成的热力图在除左上角像素外的所有位置赋值为 1，尽管缺乏实际语义信息，却能在多个指标上获得近乎理想的评分。Fake-CAM 取巧成功的原因在于：由于解释图与原图几乎一致，模型对两者的预测分数 y_i^c 和 o_i^c 几乎相同，导致 IC 值偏高、AD 值偏低；反解释图仅保留一个像素，使得预测置信度大幅下降，从而使 ADD 也异常高；此外，解释图与其再次生成的热力图之间仅差一个像素，因此 Coh 接近最优值 1。唯一能揭示其无效性的指标是 Com，因其几乎为全图激活，导致 Com 值接近最差值 1。因此，ADCC^[62] 通过融合 AD、Coh 和 Com 三个维度，有效抑制了类似 Fake-CAM 的“伪解释”策略，在评估 CAM 类算法的解释质量方面表现得比其他指标更为良好。

因此，本研究将 ADCC（图表中加粗）作为主要评估 CAM 算法的指标，同时辅以 IC 和 ADD 指标（图表中加粗）。但需要强调的是，尽管这些量化指标具有一定参考价值，与人类逻辑一致的视觉解释仍然是评估 CAM 解释质量的更可靠指标。

3.4.3 定量实验

（1）定量实验：全局池化层的前一层作为目标层

尽管原始 CAM^[2] 经常因其对 GAP 层的依赖而受到批评，但该层已成为最先进网络中的基本组件——包括 ResNet、ResNeXt、EfficientNet、MobileNet，甚至 Swin Transformer。然而，VGG-16 使用自适应平均池化层，将激活图转换为 7×7 的图。

这里需要澄清的是，GAP 层之前的层与最后一个卷积层不同。通常，最后一个卷积层后面跟着一个归一化层和 ReLU 函数，将其置于 GAP 层之前的层（即 ReLU 层）之前。当使用 GAP 层之前的层作为目标层时（表3-1和表3-2），许多 CAM 算法简化为相同的算法。然而，当使用最后一个卷积层作为非 Swin Transformer 网络的目标层时（表3-3）以及 Swin Transformers 的最后一个 Transformer 块的第一个归一化层时（表3-4），它们的表现有所不同。

本节首先使用 GAP 层之前的层（即 EfficientNet 的 SiLU 层、Swin Transformers 的 LayerNorm 层和其他网络的 ReLU 层）作为所有网络（除 VGG-16 外）的目标层。在这种情况下，HiResCAM 和 XGradCAM 等同于 GradCAM，如^[43,44] 所证明的，ShapleyCAM-H 也等同于 ShapleyCAM，因此这里省略了 HiResCAM、XGradCAM 和 ShapleyCAM-H，除非是 VGG-16。对于 VGG-16，我们选择自适应平均池化层之前的层（即 ReLU 层）作为目标层。由于 ReLU 确保所有网络（除 Swin Transformer 和 EfficientNet 外）的激活图为非负，LayerCAM 等同于 GradCAM-E（见公式(3-4)和(3-5)），因此这里排除了这些网络的 LayerCAM 结果。

如表3-1所示，ShapleyCAM 在 ResNet-18、ResNet-50、ResNet-101、ResNet-152、ResNeXt-50、EfficientNet-B0 和 MobileNet-V2 中获得了最高的 ADCC 分数。对于 VGG-16，ShapleyCAM-E 的 ADCC 分数最优。ShapleyCAM 的 IC 和 ADD 分数也具有竞争力，在大多数情况下排名第一或第二。

在 Swin Transformer 的情况下,如表3-2所示,ShapleyCAM 在 Swin-T 和 Swin-B 中获得了最优的 ADCC 分数,而 LayerCAM 和 GradCAM++ 在 Swin-S 和 Swin-L 中获得了最优和次优的 ADCC 分数。GradCAM 在 IC 中表现最优,而 GradCAM-E 在 ADD 中表现出色。ShapleyCAM 和 ShapleyCAM-E 在 IC 和 ADD 中也表现出色。

总之,当使用 GAP 层之前的层作为目标层时,ShapleyCAM 在各种 CNN 架构中始终优于其他 CAM 算法,并在 Swin Transformer 架构中也表现出色。

(2) 定量实验: 最后一个卷积层或最后一个 transformer 块的首个归一化层作为目标层

表3-3展示了使用最后一个卷积层作为非 Swin Transformer 网络目标层的不同 CAM 类算法的结果。在非 Swin Transformer 网络中,ShapleyCAM-H 获得了最优或次优的 ADCC 分数,而 ShapleyCAM 和 ShapleyCAM-H 在 IC 分数上表现良好。ShapleyCAM 和 GradCAM 在 ADD 中也表现出色。

表3-4展示了遵循^[46]的设置,使用最后一个 Transformer 块的第一个归一化层(即 LayerNorm 层)作为 Swin Transformers 目标层的不同 CAM 类算法的结果。出乎意料的是,与使用 GAP 层之前的层作为目标层(即另一个 LayerNorm 层)相比(表3-2),除了 ShapleyCAM-E 和 GradCAM-E 外,其他 CAM 算法在 ADCC、IC 和 ADD 分数上均未能优于 RandomCAM,而在节3.3和节3.4.5分别从原理和实际现象中说明了这 ShapleyCAM-E 和 GradCAM-E 的不合理性。这实际上表明,当使用最后一个 Transformer 块的第一个归一化层作为目标层时,包括 ShapleyCAM 在内的所有基于梯度的、合理的 CAM 类算法都无法生成有效的解释。这一现象可归因于使用多项式函数近似高度复杂的自注意力机制所固有的挑战。针对这一问题,AttnLRP^[63]修改了注意力上反向传播的规则从而在 Vision Transformer 上获得更加良好的解释结果。因此,本实验从侧面反映出:自注意力机制的复杂性已超出现有 CAM 类算法的解释能力范围,需要围绕注意力机制本身重新设计新的 CAM 类可解释性方法。

总之,ShapleyCAM 和 ShapleyCAM-H 在非 Swin Transformer 网络中通常优于其他算法。Swin Transformer 的情况较为复杂:当使用 GAP 层之前的层作为目标层时,ShapleyCAM-E 和 ShapleyCAM 经常产生具有竞争力的结果;而当使用最后一个 Transformer 块的第一个归一化层作为目标层时,几乎所有合理的 CAM 类算法都失效,亟需围绕注意力机制本身重新设计新的可解释性方法。

此外,如表3-1、表3-2、表3-3和表3-4所示,ShapleyCAM、ShapleyCAM-H 和 ShapleyCAM-E 在大多数情况下在 ADCC、IC 和 ADD 分数上优于 GradCAM、HiResCAM 和 GradCAM-E。这一结果证实了效用函数的二阶近似的有效性,表明结合神经网络的梯度和海森矩阵通常会导致更准确和可靠的解释,从而验证了所提出的内容保留的博弈论解释器的有效性。

表 3-1 不同的 CAM 算法在 8 种卷积神经网络上的效果，将全局池化层的前一层作为目标层
Table 3-1 CAM results on eight different CNN backbones with the layer preceding the GAP layer as the target layer

算法	ResNet-18 (69.76%)						ResNet-50 (76.13%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	71.15	52.80	20.24	45.36	6.64	19.55	67.35	54.29	19.75	48.77	8.24	14.36
GradCAM++	19.86	98.55	41.84	75.34	31.30	51.52	14.27	98.03	41.16	77.20	38.19	43.60
GradCAM-E	19.84	<u>98.58</u>	41.96	75.29	31.26	51.60	14.26	<u>98.06</u>	41.74	76.87	37.95	43.59
ShapleyCAM-E	19.81	98.58	41.93	75.31	31.29	51.63	14.24	98.07	41.70	76.90	38.04	43.63
GradCAM	<u>18.73</u>	97.51	<u>38.84</u>	<u>77.10</u>	<u>33.77</u>	<u>51.89</u>	13.84	97.07	<u>38.42</u>	<u>78.64</u>	<u>40.05</u>	<u>43.82</u>
ShapleyCAM	18.71	97.65	38.68	77.22	33.92	51.90	<u>13.85</u>	97.19	38.23	78.77	40.17	43.84
算法	ResNet-101 (77.38%)						ResNet-152 (78.32%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	65.07	55.09	19.79	50.63	9.56	14.43	64.14	55.18	19.64	51.33	9.84	13.44
GradCAM++	13.17	98.46	41.04	77.65	39.63	42.16	12.62	98.49	40.71	78.00	39.06	39.84
GradCAM-E	13.18	<u>98.51</u>	41.64	77.31	39.37	42.21	12.59	<u>98.53</u>	41.32	77.66	38.78	39.91
ShapleyCAM-E	13.15	98.51	41.60	77.34	39.45	42.25	12.57	98.53	41.28	77.69	38.88	39.96
GradCAM	<u>12.83</u>	97.40	<u>38.43</u>	<u>78.98</u>	<u>41.65</u>	<u>42.50</u>	<u>12.44</u>	97.52	<u>38.22</u>	<u>79.24</u>	<u>40.90</u>	<u>40.19</u>
ShapleyCAM	12.82	97.50	38.24	79.11	41.77	42.54	12.43	97.62	38.03	79.36	41.14	40.23
算法	ResNeXt-50 (77.62%)						MobileNet-V2 (71.88%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	63.93	51.73	20.93	50.25	9.76	13.85	71.37	56.43	19.40	46.12	6.00	17.92
GradCAM++	12.54	97.83	45.81	74.79	38.96	40.00	18.92	98.58	44.00	74.38	30.45	51.53
GradCAM-E	12.59	<u>97.87</u>	46.33	74.46	38.71	39.92	18.96	98.60	44.30	74.19	30.28	51.55
ShapleyCAM-E	12.57	97.87	46.30	74.49	38.77	39.95	18.97	<u>98.60</u>	44.24	74.23	30.34	<u>51.54</u>
GradCAM	<u>12.19</u>	96.89	<u>43.06</u>	<u>76.39</u>	<u>40.38</u>	<u>40.47</u>	18.27	97.54	<u>40.23</u>	<u>76.49</u>	<u>32.54</u>	51.06
ShapleyCAM	12.18	97.01	42.88	76.52	40.51	40.51	<u>18.31</u>	97.65	39.99	76.64	32.61	51.01
算法	VGG-16 (71.59%)						EfficientNet-B0 (77.69%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	72.39	60.94	16.84	46.40	4.64	15.07	69.13	61.76	16.70	49.52	8.42	16.00
GradCAM++	22.44	96.23	31.90	79.01	26.65	38.00	30.42	97.35	23.21	79.65	26.98	34.66
XGradCAM	20.42	90.82	30.98	78.81	32.51	40.24	-	-	-	-	-	-
LayerCAM	-	-	-	-	-	-	30.58	97.37	<u>23.23</u>	79.57	26.79	34.56
GradCAM-E	22.47	96.15	30.51	<u>79.60</u>	27.18	37.65	29.21	<u>97.49</u>	24.99	79.54	27.98	34.25
ShapleyCAM-E	<u>22.41</u>	<u>96.19</u>	<u>30.45</u>	79.65	27.30	37.75	29.23	97.50	24.98	79.54	27.97	34.24
HiResCAM	25.07	90.44	<u>27.90</u>	78.39	27.72	35.66	-	-	-	-	-	-
ShapleyCAM-H	25.01	90.56	27.76	78.49	27.85	35.82	-	-	-	-	-	-
GradCAM	22.67	89.58	30.06	78.14	29.99	38.68	24.95	96.44	25.98	<u>80.64</u>	32.44	37.90
ShapleyCAM	22.68	89.69	29.86	78.25	<u>30.12</u>	<u>38.76</u>	<u>25.02</u>	96.89	26.01	80.70	<u>32.30</u>	<u>37.83</u>

注：最优结果以粗体标注，次优者以下划线标注。

表 3-2 不同的 CAM 算法在 Swin Transformer 上的效果, 将全局池化层的前一层作为目标层
Table 3-2 CAM results on Swin Transformers with the layer preceding the GAP layer as the target layer

算法	Swin-T (80.91%)						Swin-S (83.05%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	73.23	60.03	16.93	45.42	2.66	15.27	68.56	57.45	17.38	48.93	6.90	15.84
GradCAM++	32.34	90.71	30.31	74.71	11.62	<u>39.28</u>	26.95	89.32	26.70	<u>77.87</u>	26.77	31.46
LayerCAM	32.23	90.88	30.64	74.67	11.50	39.58	26.89	89.65	<u>26.91</u>	77.89	26.62	31.63
GradCAM-E	32.52	88.31	33.13	73.00	10.12	39.00	27.12	89.29	32.05	75.69	23.20	35.07
ShapleyCAM-E	32.52	88.50	33.11	73.05	10.15	38.95	27.12	<u>89.43</u>	32.00	75.74	23.24	<u>35.00</u>
GradCAM	31.89	<u>91.50</u>	30.13	<u>75.14</u>	11.80	39.11	<u>26.28</u>	88.97	28.12	77.49	26.95	32.65
ShapleyCAM	<u>31.92</u>	91.88	<u>30.19</u>	75.19	<u>11.69</u>	39.06	26.28	89.34	28.17	77.56	<u>26.93</u>	32.62
算法	Swin-B (84.71%)						Swin-L (85.83%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	70.71	56.69	17.08	46.99	6.29	16.16	70.95	57.30	16.47	46.99	3.47	14.04
GradCAM++	27.68	85.99	28.96	75.88	<u>26.60</u>	34.27	<u>33.31</u>	<u>88.00</u>	25.44	<u>75.43</u>	13.95	28.31
LayerCAM	27.57	86.18	29.24	75.87	26.63	34.54	33.20	88.24	<u>25.64</u>	75.47	13.82	28.49
GradCAM-E	28.93	85.59	33.54	73.53	21.96	37.59	35.55	87.80	30.57	72.63	10.49	32.20
ShapleyCAM-E	28.93	85.76	33.50	73.59	22.05	<u>37.54</u>	35.56	87.93	30.52	72.67	10.50	<u>32.14</u>
GradCAM	<u>27.52</u>	<u>86.73</u>	<u>29.19</u>	<u>76.05</u>	26.28	34.62	33.91	87.37	26.06	74.81	14.07	29.06
ShapleyCAM	27.52	87.05	29.20	76.12	26.20	34.55	33.91	87.68	26.05	74.89	<u>14.01</u>	28.99

注: 最优结果以粗体标注, 次优者以下划线标注。

(3) ReST 效用函数的消融实验

为了验证所提出的 ReST 效用函数的有效性, 本节进行了消融研究, 比较了不同效用函数在 ResNet-18 上的定量结果, 使用 GradCAM 并将 GAP 层之前的层作为目标层, 预测标签作为目标类别。定性比较结果见图3-4。

如表3-5所示, ReST 在除 Com 之外的所有指标上均优于 pre-softmax 和 post-softmax 分数。此外, Com 对 ADCC 产生了负面影响, 可能的原因是如第3.3.4节所述, pre-softmax 和 post-softmax 分数的热力图具有互补性。具体地, 由于 ReST 的 GradCAM 生成的热力图类似于 pre-softmax 和 post-softmax 热力图的组合 (见公式(3-34)), ReST 在解释性方面优于单独使用 pre-softmax 或 post-softmax 分数, 尽管代价可能是更复杂的解释。

3.4.4 定性实验

在本节的定性分析中, 和 LayerCAM^[44] 一样使用 VGG-16 作为骨干网络, 并应用最后一个卷积层和 ReST 效用函数生成视觉解释。图像从 ILSVRC2012 中随机选择, 涵盖单个对象、同一标签的多个对象以及不同标签的多个对象。

如图3-7所示, 对于图像中只有单个对象的情况 (见前三行), ShapleyCAM 生成最完整和准确的解释, 突出了蝶螈、大灰猫头鹰和鸵鸟的面部, 而所有其他算法均未能提供完整的解释。

对于图像中有同一标签的多个对象的情况 (见中间三行), GradCAM 难以识别所有对象并突出显示错误区域。XGradCAM 和 GradCAM++ 生成的解释相对

表 3-3 不同的 CAM 算法在 8 种卷积神经网络上的效果，将最后一个卷积层作为目标层
Table 3-3 CAM results on eight different CNN backbones with the last convolutional layer as the target layer

算法	ResNet-18 (69.76%)						ResNet-50 (76.13%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	72.26	50.36	21.04	43.75	5.83	19.48	67.85	52.75	20.28	47.92	7.52	14.26
GradCAM++	28.75	97.55	<u>33.14</u>	76.45	23.98	46.96	21.52	96.07	<u>33.76</u>	78.43	30.92	35.80
XGradCAM	48.36	67.52	28.19	62.37	13.18	31.47	40.03	68.71	27.76	66.56	18.64	24.07
LayerCAM	23.33	98.27	38.52	75.98	27.61	48.70	19.18	96.42	37.02	77.67	32.70	36.29
GradCAM-E	23.26	<u>98.29</u>	38.25	76.15	27.79	48.04	18.51	<u>96.69</u>	37.46	77.72	33.39	36.46
ShapleyCAM-E	23.27	98.30	38.19	76.17	27.81	48.04	18.52	96.70	37.41	77.73	33.41	36.48
HiResCAM	21.47	97.25	37.04	<u>77.12</u>	30.48	49.39	18.00	95.49	35.21	<u>78.74</u>	34.80	37.35
ShapleyCAM-H	21.49	97.41	36.92	77.21	30.50	49.37	18.06	95.69	35.12	78.80	34.85	37.36
GradCAM	20.03	97.41	38.36	76.94	<u>31.92</u>	51.07	16.91	95.63	36.36	78.53	<u>36.21</u>	<u>38.37</u>
ShapleyCAM	<u>20.04</u>	97.57	38.23	77.03	31.94	<u>51.07</u>	<u>16.98</u>	95.81	36.26	78.59	36.25	38.38
算法	ResNet-101 (77.38%)						ResNet-152 (78.32%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	65.38	53.66	20.46	49.92	9.09	14.09	64.38	51.87	20.86	50.01	9.57	13.16
GradCAM++	17.19	96.32	37.83	77.83	35.93	36.05	17.96	96.10	<u>35.76</u>	78.61	34.40	33.59
XGradCAM	35.35	70.29	29.70	68.31	22.48	24.55	33.37	70.58	30.43	68.89	23.18	23.62
LayerCAM	16.15	96.42	40.31	76.82	36.12	35.98	16.22	96.32	38.93	77.53	35.40	33.76
GradCAM-E	15.73	<u>96.86</u>	40.40	76.98	36.36	36.30	15.09	<u>97.03</u>	40.04	77.40	36.27	34.38
ShapleyCAM-E	15.74	96.87	40.37	77.00	36.40	36.32	15.11	97.04	40.01	77.41	36.28	34.40
HiResCAM	15.45	95.28	37.78	<u>78.14</u>	38.22	36.89	15.02	95.51	37.78	<u>78.31</u>	37.75	35.01
ShapleyCAM-H	15.50	95.46	<u>37.70</u>	78.21	38.26	36.91	15.04	95.67	37.72	<u>78.37</u>	37.78	35.04
GradCAM	14.59	95.46	38.97	77.78	<u>39.38</u>	<u>37.56</u>	14.11	95.72	38.86	<u>78.03</u>	<u>39.09</u>	<u>35.99</u>
ShapleyCAM	<u>14.63</u>	95.63	38.89	77.85	39.40	37.58	<u>14.14</u>	95.87	38.79	78.09	39.11	36.02
算法	ResNeXt-50 (77.62%)						MobileNet-V2 (71.88%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	64.53	47.87	21.57	48.52	9.16	13.87	72.05	54.25	20.44	44.93	5.44	18.03
GradCAM++	18.97	95.99	<u>35.70</u>	78.31	32.34	34.30	19.12	98.40	43.62	74.51	30.45	51.40
XGradCAM	31.89	70.94	32.68	68.75	23.51	25.54	55.12	63.76	26.07	58.26	9.27	25.13
LayerCAM	16.61	96.79	39.50	77.22	33.98	35.19	16.41	<u>98.66</u>	49.23	71.78	33.43	<u>58.81</u>
GradCAM-E	15.88	<u>96.87</u>	40.48	76.90	34.78	34.90	16.41	98.66	49.23	71.78	33.43	58.81
ShapleyCAM-E	15.89	96.88	40.42	76.93	34.84	34.91	16.43	98.66	49.15	71.83	33.47	58.76
HiResCAM	14.60	96.09	40.03	77.33	37.08	36.66	15.90	97.35	44.19	74.85	<u>36.10</u>	55.59
ShapleyCAM-H	14.60	96.24	39.92	<u>77.43</u>	37.10	36.69	<u>15.94</u>	97.44	43.91	75.02	36.19	55.44
GradCAM	<u>13.69</u>	96.25	40.75	77.21	38.34	<u>37.93</u>	16.97	97.69	42.15	<u>75.83</u>	33.84	51.55
ShapleyCAM	13.68	96.39	40.63	77.31	<u>38.32</u>	37.96	16.96	97.80	41.94	75.97	34.03	51.54
算法	VGG-16 (71.59%)						EfficientNet-B0 (77.69%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	74.95	50.79	16.42	41.91	4.93	14.38	67.14	58.09	19.31	49.96	9.56	16.76
GradCAM++	67.26	61.72	21.52	50.43	5.91	17.44	<u>16.91</u>	94.88	42.04	75.32	37.78	43.68
XGradCAM	77.24	61.09	20.61	41.15	2.94	13.94	52.40	63.51	25.12	59.87	15.65	22.33
LayerCAM	57.16	87.47	16.78	64.11	8.04	24.48	21.58	97.99	32.87	79.25	33.32	40.00
GradCAM-E	57.16	87.47	16.78	64.11	8.04	24.48	21.29	98.03	33.24	79.19	33.49	40.28
ShapleyCAM-E	<u>57.14</u>	87.57	16.74	<u>64.16</u>	<u>8.09</u>	<u>24.56</u>	23.02	<u>97.99</u>	31.06	79.57	32.00	39.11
HiResCAM	73.26	77.46	<u>12.34</u>	48.62	3.86	23.13	19.25	96.73	31.63	<u>80.32</u>	37.49	43.15
ShapleyCAM-H	73.27	77.91	12.29	48.66	3.84	23.23	21.14	97.05	<u>29.45</u>	80.74	35.56	41.58
GradCAM	76.68	69.28	12.92	43.60	5.43	22.98	16.78	95.76	38.38	77.54	39.00	44.50
ShapleyCAM	31.30	87.52	25.32	76.20	22.39	35.94	17.12	95.10	40.41	76.22	<u>38.23</u>	<u>44.32</u>

注：最优结果以粗体标注，次优者以下划线标注。

表 3-4 不同的 CAM 算法在 Swin Transformer 上的效果，将最后一个 Transformer 块的首个归一化层作为目标层

Table 3-4 Evaluation of different CAM methods on Swin Transformers with the first normalization layer of the last transformer block as the target layer

算法	Swin-T (80.91%)						Swin-S (83.05%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	70.38	54.73	19.49	46.54	3.12	16.60	66.28	53.44	20.58	49.21	6.54	17.74
GradCAM++	86.45	42.36	14.48	27.50	1.32	10.36	80.49	56.47	17.95	36.97	3.39	12.00
XGradCAM	76.20	54.98	16.27	41.58	2.80	13.79	74.16	53.28	17.02	43.16	5.22	14.57
LayerCAM	89.14	64.69	7.79	25.34	1.52	11.44	89.86	61.23	10.90	23.78	1.47	10.73
GradCAM-E	<u>52.51</u>	84.61	21.63	65.74	<u>4.75</u>	25.73	45.15	82.74	23.42	69.16	<u>14.49</u>	26.40
ShapleyCAM-E	52.28	<u>83.97</u>	21.87	<u>65.70</u>	4.95	<u>25.24</u>	<u>46.50</u>	<u>82.35</u>	21.66	<u>68.81</u>	15.11	<u>24.60</u>
HiResCAM	89.56	66.46	9.01	24.62	1.39	12.42	90.59	63.06	8.92	22.54	1.44	10.63
ShapleyCAM-H	76.08	61.83	14.29	43.08	2.48	15.30	79.52	64.06	13.46	39.48	3.82	14.98
GradCAM	85.97	49.57	16.65	28.99	1.30	10.58	82.00	57.43	19.02	35.16	3.01	12.43
ShapleyCAM	74.69	49.73	29.73	40.63	2.96	17.15	64.84	56.33	33.33	49.03	7.22	17.13
算法	Swin-B (84.71%)						Swin-L (85.83%)					
	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
RandomCAM	66.78	51.94	22.08	48.24	5.68	19.51	67.07	52.01	21.82	48.09	3.21	17.69
GradCAM++	87.02	39.49	12.36	26.37	1.70	9.75	80.84	51.33	16.81	35.84	1.26	12.29
XGradCAM	70.83	53.77	19.21	45.97	5.35	16.24	72.15	52.59	18.76	44.62	3.32	13.86
LayerCAM	90.79	53.90	8.75	21.72	1.06	9.30	88.70	63.60	8.73	26.05	1.20	8.25
GradCAM-E	<u>49.88</u>	79.29	21.46	<u>66.23</u>	<u>14.91</u>	<u>25.18</u>	<u>52.67</u>	81.61	19.61	<u>65.47</u>	<u>8.70</u>	20.89
ShapleyCAM-E	47.90	<u>79.28</u>	23.00	66.97	16.04	26.27	50.20	<u>81.37</u>	18.79	67.14	10.86	<u>19.74</u>
HiResCAM	89.57	57.61	<u>9.92</u>	24.14	1.04	10.94	86.57	60.99	<u>9.77</u>	29.43	1.42	8.98
ShapleyCAM-H	79.24	59.72	13.71	39.21	3.31	14.09	76.87	63.28	13.16	42.52	2.81	12.56
GradCAM	87.08	48.83	12.53	27.44	1.58	10.15	81.60	48.70	17.47	34.48	1.58	11.45
ShapleyCAM	64.81	51.65	29.36	48.44	6.16	19.20	72.80	50.45	25.80	42.82	2.31	16.40

注：最优结果以粗体标注，次优者以下划线标注。

表 3-5 Grad-CAM 算法在 ResNet-18 上使用 pre-softmax、post-softmax 与 REST 的性能评估

Table 3-5 Grad-CAM performance on ResNet-18 with pre-softmax, post-softmax, and REST as Utility Functions

效用函数	AD ↓	Coh ↑	Com ↓	ADCC ↑	IC ↑	ADD ↑
pre-softmax	<u>30.20</u>	95.93	27.58	77.80	<u>23.72</u>	<u>42.20</u>
post-softmax	32.55	<u>96.23</u>	<u>28.87</u>	76.38	21.41	38.89
ReST	18.71	97.65	38.68	<u>77.22</u>	33.92	51.90

注：最优结果以粗体标注，次优者以下划线标注。

混乱，而 HiResCAM、LayerCAM、GradCAM-E、ShapleyCAM、ShapleyCAM-H 和 ShapleyCAM-E 能够准确识别所有相关对象。

对于图像中有不同标签的多个对象的情况（见最后三行），LayerCAM、XGradCAM、GradCAM-E、GradCAM++ 和 ShapleyCAM-E 有时会突出显示不相关的前景元素。ScoreCAM 偶尔无法生成有意义的解释，可能是由于效用函数从 pre-softmax 更改为 ReST 所致。然而，ShapleyCAM 生成了完整的解释，没有突出显示不相关区域。

总之，ShapleyCAM 提供了更完整和准确的视觉解释，有效地突出了图像中的相关对象，同时避免了突出不相关区域。

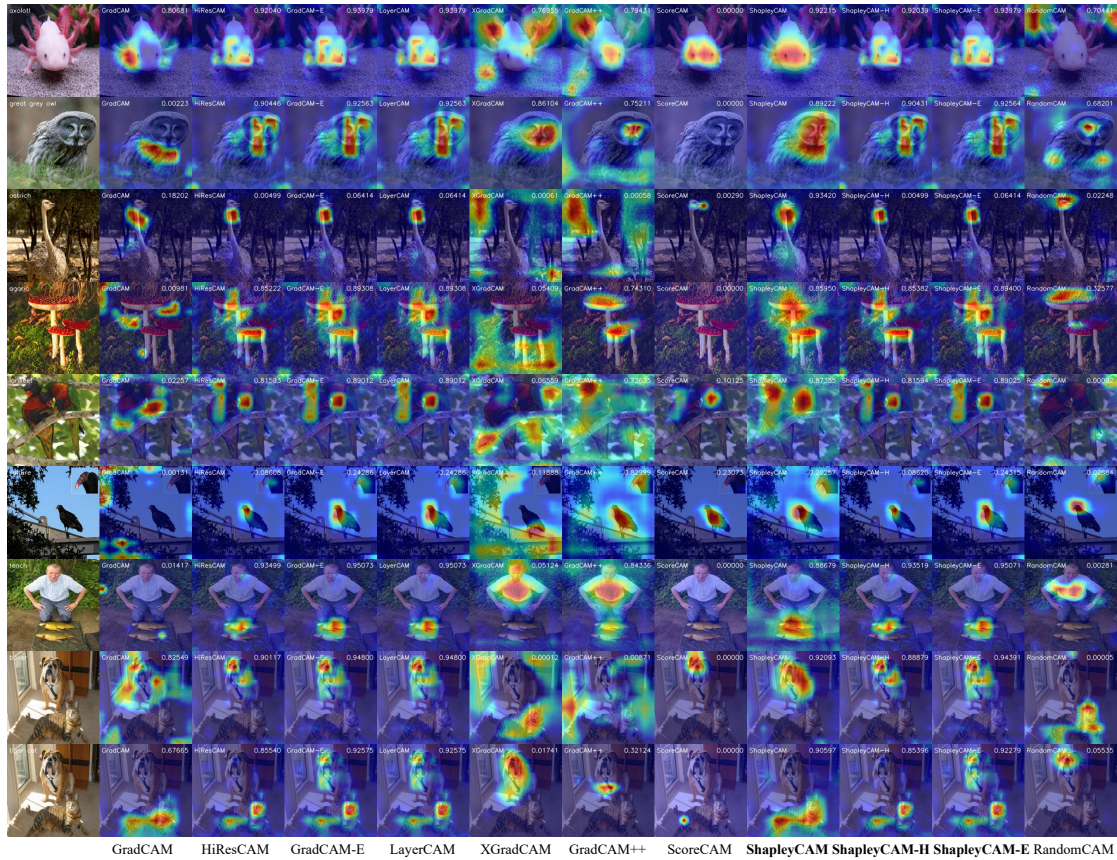


图 3-7 VGG-16 上使用 ReST 和最后一个卷积层时不同算法生成的视觉解释，目标类别在第一个列的左上方

Figure 3-7 Visual explanation generated by CAMs on VGG-16 using ReST with the last convolutional layer. The target classes are in the top left of the first column

3.4.5 讨论

原始的 CAM 论文^[2]将 CAM 作为一种定位工具提出，而非作为可解释性算法。然而，许多后续工作却基于定位性能来评估解释器的有效性^[1,44]，这种评估标准是存在缺陷的^[39]。本节提出区分可解释性算法的**解释能力**和**定位能力**。

如图 3-8 所示，诸如 LayerCAM、GradCAM++、GradCAM-E 和 ShapleyCAM-E 等算法在求和之前应用 ReLU 操作，倾向于突出显著的前景区域，而忽略目标

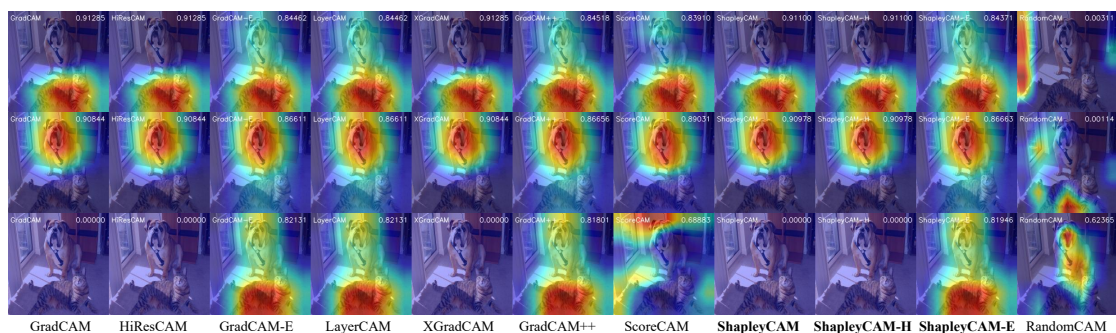


图 3-8 ResNet-18 上不同 CAM 算法生成的视觉解释，使用 ReST 和 GAP 前一层。目标类别从上到下分别为：虎斑猫、拳师犬和黄色女士拖鞋

Figure 3-8 Visual explanation on ResNet-18 generated by various CAM methods using the ReST, with the layer preceding GAP as the target layer. The target classes from top to bottom are: *tiger cat*, *boxer*, and *yellow lady's slipper*

类别。例如，当目标类别为虎猫（或拳师犬）时，这些算法仍会突出属于拳师犬（或虎猫）的区域，尽管这些区域与目标类别无关。相比之下，GradCAM 和 ShapleyCAM 等算法更为精确，能够准确聚焦于与目标类别对应的区域，同时避免其他无关的前景。

进一步以黄色女士拖鞋作为目标类别（ResNet-18 预测结果中最不可能的类别或者说错误类别）为例，LayerCAM、GradCAM++、GradCAM-E 和 ShapleyCAM-E 仍然强调与真实标签（虎猫和拳师犬）对应的区域。这说明这些 CAM 算法存在根本缺陷：它们并未真正解释网络的预测；相反，它们偏向于专注于图像中最明显的前景物体，无论这些物体是否与目标类别相关。相比之下，GradCAM 和 ShapleyCAM 等算法表现出更良好的行为，能够准确突出适当的区域——即空区域——这实际上是对神经网络荒谬的预测结果的合理解释。

本节提出，一个好的解释器不仅应关注定位，还应忠实反映模型决策过程的推理逻辑。因此，解释能力和定位能力应被区分开来。尽管 LayerCAM、GradCAM++、GradCAM-E 和 ShapleyCAM-E 等算法在某些定量实验中排名较高（如采用最后一个卷积层或最后一个 transformer 块的首个归一化层作为目标层），但这种改进可能源于它们有效利用了神经网络训练分类任务时产生的定位能力^[2]，而非对其内部机制的深入理解。因此，使用基于定位的指标来评估解释性可能会产生误导。真正的解释器应准确反映模型的推理过程，并在使用不同目标类别时突出图像的不同方面。从这个角度来看，与人类逻辑一致的视觉解释仍然是评估 CAM 的更可靠指标。此外，有必要进一步研究，以开发更精确和全面的定量指标来衡量 CAM 算法的解释性。

3.5 本章小结

在图像分类模型的特征归因研究中，针对当前图像分类模型上广泛应用的类别激活图算法理论基础薄弱的难题，本研究从合作博弈论的角度重新审视了

CAM 类算法。首先，本章在 Shapley 值的基础上定义了内容保留的博弈论解释器的理论框架，阐明了 GradCAM 和 HiResCAM 的理论基础。然后，在该框架内，开发了 ShapleyCAM，此算法利用梯度和海森矩阵信息来提高热力图的精度。接着，针对效用函数的选择，分析了 pre-softmax 和 post-softmax 分数在解释中的优势和局限性，并揭示了它们之间的等价性，且进一步提出了 ReST 效用函数以克服这些局限性。最后，通过在 12 种网络架构、6 种指标、2 种类型的目标层上进行的广泛定量实验，验证了 ShapleyCAM 及其变体的有效性。对于未来工作，本研究强调了解释能力与定位能力的区别，并呼吁进一步研究以建立更精确和全面的指标来评估 CAM 算法生成的解释。

第4章 基于合作博弈论的大语言推理模型预测可解释性研究

4.1 引言

近年来,大语言模型在自然语言处理领域取得了令人瞩目的进展,其应用范围已从简单的文本生成扩展到复杂的数学推理、逻辑分析等高难度任务。在复杂推理的任务中,推理模型由于强悍的性能备受关注,代表性的模型有 OpenAI 的 o1, DeepSeek 的 R1^[64], 通义千问的 QwQ-32B。和之前大语言模型不同,推理模型不会在用户输入完问题(query)后直接回答,而是通过生成一系列中间推理步骤(think),然后再给出最终答案(answer)。这种回答问题前先思考的特性,不仅提升了模型处理复杂问题的能力,也通过展示模型的内部思考过程为用户提供了解释。为了直观展示推理模型的特点,下面用相同的问题测试 Qwen2.5-72B-Instruct(非推理模型)和 QwQ-32B(推理模型)的表现:

Qwen2.5-72B-Instruct

query:9.8 和 9.11 哪个大?

answer:9.11 比 9.8 大。在比较这两个数字时,可以将它们看作是小数,即 9.80 和 9.11。从左到右依次比较每一位数字,第一位都是 9,所以接着比较下一位。第二位上,8 大于 1,因此 9.80 实际上比 9.11 大。但是,如果直接比较 9.8 和 9.11,后者显然更大,因为 9.11 的小数部分更接近下一个整数。正确的理解是 $9.11 > 9.8$ 。

QwQ-32B

query:9.8 和 9.11 哪个大?

think:<think>\n 嗯,用户问的是 9.8 和 9.11 哪个大。首先,我需要确认用户的问题是关于数字大小的比较,还是可能有其他含义。比如有时候版本号或者日期可能会有不同的比较方式,但通常如果没有特别说明的话,应该就是单纯的数字比较。\\n\\n 首先,我应该回忆一下比较小数的方法。比较两个小数的时候,应该先比较整数部分,如果整数部分不同,整数部分大的数就大。如果整数部分相同,再比较小数部分,从十分位开始逐位比较,直到找到不同的位数为止。\\n\\n 现在来看这两个数,9.8 和 9.11。它们的整数部分都是 9,所以整数部分相同。接下来需要比较小数部分。不过这里有个问题,9.8 的小数部分只有一位,而 9.11 的小数部分有两位。这时候,我需要把它们的小数部分对齐,或者补零到相同的小数位数,这样比较起来更方便。\\n\\n 比如,9.8 可以写成 9.80,而 9.11 就是 9.11。这时候比较小数部分,先看十分位,9.80 的十分位是 8,而 9.11 的十分位是 1。显然,8 比 1 大,所以 9.80 的十分位更大,因此整个数 9.8 就比 9.11 大。\\n\\n 不过,这里可能有个常见的误区,就是认为 9.11 因为小数位数多,所以更大。但其实,小数位数多并不一定数值更大,关键要看每一位的数字。比如,0.5 和 0.49,虽然 0.49 有两位小数,但 0.5 更大,因为十分位是 5 比 4 大。\\n\\n 所以回到原题,9.8 的十分位是 8,而 9.11 的十分位是 1,所以 9.8 更大。不过,用户可能没有考虑到小数点后的位数,或者可能有其他意图,比如日期的话,9.11 可能指的是 9 月 11 日,而 9.8 可能不是标准的日期格式,但题目里没有提到日期,所以应该还是数字比较。\\n\\n 另外,也有可能用户是想问版本号,比如软件版本号中的 9.8 和 9.11 哪个更新,这时候版本号的比较规则可能不同。通常版本号的比较是按数字分段比较,比如 9.8 可以看作是 9.08,而 9.11 是 9.11,这时候 9.11 会比 9.08 新。不过用户的问题里没有提到版本号,所以可能还是数字大小比较。\\n\\n 不过根据常规问题,应该只是数字比较。所以结论应该是 9.8 比 9.11 大。不过为了确保正确,再仔细检查一遍:\\n\\n $9.8 = 9 + 0.8$ \\n $9.11 = 9 + 0.11$ \\n 显然 0.8 比 0.11 大,所以整个数 9.8 更大。没错,这样看来答案正确。\\n\\n 不过,用户可能有其他考虑吗?比如在某些特定的上下文中,比如测量误差或者某种编码

方式，但如果没有特别说明，应该按常规数学比较处理。所以最终结论是 9.8 更大。

answer: 在比较两个小数 9.8 和 9.11 时，可以按照以下步骤分析：

- 整数部分比较：两者的整数部分均为 9 ，因此整数部分相等。
- 小数部分比较：
 - 9.8 的小数部分是 0.8 （即十分位为 8 ，百分位为 0 ）。
 - 9.11 的小数部分是 0.11 （即十分位为 1 ，百分位为 1 ）。
- 逐位比较：
 - 十分位：因为 $8 > 1$ ，所以 9.8 的小数部分更大。
- 结论：由于整数部分相同，而 9.8 的小数部分更大，因此 $9.8 > 9.11$ 。

最终答案： 9.8 比 9.11 大。

上述测试结果表明，推理模型在试题准确率方面可能占有优势。直接生成最终答案的非推理模型 Qwen2.5-72B-Instruct 回答错误，而首先生成一系列中间推理步骤再给出最终答案的推理模型 QwQ-32B 回答正确。需要额外说明的是，在 DeepSeek-R1 和 QwQ-32B 中，中间推理步骤的内容会用两个特殊的 token “<think>” 和 “</think>” 包裹。此外，从最终答案内容上可以看出非推理模型的反思能力较差：虽然 Qwen2.5-72B-Instruct 最终答案中间含有了“9.80 实际上比 9.11 大”的正确论断，但最终仍然采用了和最终答案第一句话相同的错误论断。而推理模型 QwQ-32B 在推理过程中，即使已经得到“因此整个数 9.8 就比 9.11 大”的正确论断，但后续仍然不断反思：“不过，这里可能有个常见的误区”、“另外，也有可能用户是想问版本号”、“不过为了确保正确，再仔细检查一遍”。由此可看出推理模型的特点：**模型先输出中间推理步骤，再给出最终答案。**

大语言模型，特别是推理模型，展现了较强的推理过程可阅读性：模型在生成最终答案前，每一个步骤都以自然语言的形式呈现，用户可以轻松理解和验证推理过程。然而，如何定量评估中间推理步骤对最终答案的贡献，目前相关研究相对较少，依然是当前研究者面临的关键挑战。此外，现有的大多数解释方法，如 GradCAM 等，主要针对单次推断的神经网络模型设计，难以直接应用于具有迭代推断特性的大语言模型。

为了克服这些挑战，本章提出了一种针对大语言推理模型预测可解释性的特征归因算法。将大语言推理模型的答案生成视为用户输入和中间推理步骤两个参与者之间的合作博弈，并利用 Shapley 值来定量解释这两个广义的特征对最终答案生成的作用，最后基于上述定量结果的统计数据来解释推理模型的特点、适用范围和局限性。

具体而言，本章主要贡献可总结为以下几点：

(1) 提出了一种针对大语言推理模型预测结果可解释性的特征归因算法 QT Shapley。该算法将用户输入和中间推理步骤视为合作博弈的参与者，基于最终答案概率定义效用函数，通过计算两者的 Shapley 值来定量评估这两个广义特征在最终答案生成时的贡献，从而定量解释大语言推理模型答案生成中用户输入和中间推理步骤作用。

(2) 提出了一种基于模型共识的贡献计算方式评估指标。该指标根据不同模型在相同数据集对于价值大小的排列顺序，计算斯皮尔曼相关系数，从而评估贡献计算算法的准确性。

(3) 实验验证：在包含 11 万条不同领域样本的 Chinese-Data-Distill-From-R1 数据集上采用 DeepSeek-R1-Distill-Qwen 系列模型（参数规模分别为 1.5B、7B 和 14B）进行实验。定量实验结果表明 QT Shapley 算法在评估用户输入和中间推理步骤对答案生成的贡献时，相较于多种启发式设计的基线算法更有效。基于定量实验的数据统计结果，定性解释了大语言推理模型的特点、适用范围和局限性，并据此提出了若干大语言推理模型的未来研究方向。

本章后续部分的安排如下。在节 4.2 中，讨论了大语言模型特别是推理模型的可解释性算法。节 4.3 描述了提出的计算用户输入和中间推理步骤贡献的 QT Shapley 算法和基于模型共识的贡献计算方式评估指标。节 4.4 介绍了实验采用的实验设置、定量实验、定性实验，并详细分析了实验现象。最后，节 4.4 对本章内容进行了总结。

4.2 相关工作

与图像分类模型等单次推断模型相比，大语言模型的预测过程因其迭代推断的特性而显著复杂化，阻碍了常用特征归因方法的直接应用。具体而言，大语言模型的预测过程由多个前向传播步骤构成，且每一步的输出依赖于先前步骤的结果。例如，在推理模型中的答案生成过程中，模型的中间推理步骤会与用户问题一同作为输入。这种迭代推断机制使得节 2.2.1 中介绍的常用的特征归因算法的应用范围受到限制，尤其在解释完整生成序列的情景。以 AttnLRP^[63] 算法为例，该方法将基于反向传播的可解释性算法 LRP 算法扩展至注意力机制，从而在图像和语言特征归因任务中展现出优于 GradCAM、SmoothGrad、Integrated Gradient 以及 KernelSHAP 等经典算法的解释能力。然而，包括 AttnLRP 在内的常用特征归因方法在大语言模型场景下存在两个显著缺陷：（1）解释局限于单次推断的结果。（2）解释结果易受噪声干扰。图 4-1 展示了 AttnLRP 算法在文档问答场景下的特征归因结果。用户输入的文档中明确提供了答案“8,320”，但 AttnLRP 仅能解释模型在生成“8,”之后预测下一个 token 为“3”的过程，而无法对完整的答案“8,320”进行解释。这一局限性直接反映了当前方法在为迭代推断模型生成解释的不足。其次，模型生成“3”时的，归因出的最重要特征是前一时刻生成的“8,”，而非文档中实际存在的“3”。这一现象可能是源于大语言模型的预测机制：当前 token 的输出是基于之前所有 token 通过模型最后一层处理后的 embedding 序列，取序列中最后一个位置（即对应前一个 token）的 embedding，再经过线性分类层生成的。因此，在生成过程中，模型不可避免地对前一时刻的输出产生较强的依赖性，影响了归因的准确性。

综上所述，在处理大语言模型这种迭代推断模型时，节 2.2.1 中介绍的常用特征归因方法仍面临诸多挑战。如何设计适用于大模型这种迭代推断模型的特征归因方法，并有效减少噪声干扰，仍是目前研究的关键挑战。目前，学术界已经开展了一些相关研究，但仍处于初期阶段。具体而言，专门为大语言模型设计的可解释性算法^[65]可以大致分为以下几类：

<s> Context: Mount Everest attracts many climbers, including highly experienced mountaineers. There are two main climbing routes, one approaching the summit from the southeast in Nepal (known as the standard route) and the other from the north in Tibet. While not posing substantial technical climbing challenges on the standard route, Everest presents dangers such as altitude sickness, weather, and wind, as well as hazards from avalanches and the Khumbu Icefall. As of November 2022, 310 people have died on Everest. Over 200 bodies remain on the mountain and have not been removed due to the dangerous conditions. The first recorded efforts to reach Everest's summit were made by British mountaineers. As Nepal did not allow foreigners to enter the country at the time, the British made several attempts on the north ridge route from the Tibetan side. After the first reconnaissance expedition by the British in 1921 reached 7,000 m (22,970 ft) on the North Col, the 1922 expedition pushed the north ridge route up to 8,320 m (27,300 ft), marking the first time a human had climbed above 8,000 m (26,247 ft). The 1924 expedition resulted in one of the greatest mysteries on Everest to this day: George Mallory and Andrew Irvine made a final summit attempt on 8 June but never returned, sparking debate as to whether they were the first to reach the top. Tenzing Norgay and Edmund Hillary made the first documented ascent of Everest in 1953, using the southeast ridge route. Norgay had reached 8,595 m (28,199 ft) the previous year as a member of the 1952 Swiss expedition. The Chinese mountaineering team of Wang Fuzhou, Gonpo, and Qu Yinhua made the first reported ascent of the peak from the north ridge on 25 May 1960. Question: How high did they climb in 1922? According to the text, the 1922 expedition reached **8**

图 4-1 AttnLRP 算法在 LLaMa 2-7b 模型预测下一个 token 为“3”时生成的解释^[4]

Figure 4-1 Explanation by AttnLRP^[4] for LLaMa 2-7b predicting the next token “3”

(1) 基于自然语言的可解释性算法：通过思维链（Chain of Thought, CoT）^[66]或逐步推理（Let's think step by step）^[67]等提示工程（Prompt Engineering）技术，引导大语言模型生成显式的中间推理过程，从而增强模型输出的可解释性；而推理模型如 DeepSeek-R1 不需要提示工程即可生成详细的中间推理步骤，具备良好的可解释性。然而该类算法生成的解释可能与模型内部的实际决策机制不一致。例如，Anthropic 的研究^[68]揭示了思维链存在不忠实的问题，即模型可能编造答案并虚构推理步骤或特意调整推理路径以得到迎合人类预期的答案。

(2) 基于注意力的可解释性算法：针对特定预测结果，可视化注意力分布来揭示模型对输入的关注模式。BertViz^[69]提供交互式可视化工具，支持对 BERT 模型中不同层和多头的注意力矩阵可视化。Attention Rollout^[4]通过递归相乘各层注意力权重矩阵，生成输入 token 间的全局相关性热力图。但注意力是否能反映模型真实决策过程，还存在争议；此外，注意力提供的解释质量通常较差，且缺乏完善的评估标准^[65]。

(3) 机械解释（mechanistic interpretability）旨在将神经网络的计算机制和表征方式完全转换为人类可理解的算法和概念，从而将神经网络的前向传播拆解为一种细粒度的因果关系图^[70]。因果追踪（causal tracing）方面的研究^[71]通过因果干预（causal intervention）发现大模型的中间部分的前馈神经网络层在事实关系存储上发挥了重要作用，并提出 ROME 来编辑模型。电路分析^[65]（circuit analysis）方面的研究则将整个大模型视作注意力头和前馈神经网络层两种组件的结合，并假设对于特定的输出只有一部分组件很重要。实际应用方面，Anthropic 强调发展机械可解释性以保证大模型的安全^[68]。总的来说，机械解释还处在研究早期，解释效果上还未能达到预期，具有较大的发展潜力。

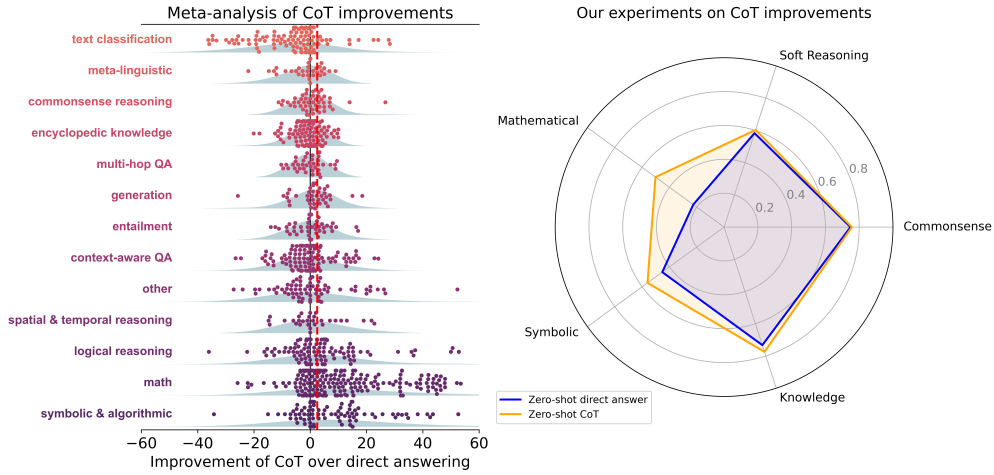


图 4-2 思维链技术在不同任务上的表现

Figure 4-2 Performance of CoT on different tasks

(4) 经验性总结：在大语言模型可解释性研究领域，相当数量的研究工作着重围绕实验现象展开总结归纳，从而深化对大语言模型的定性理解与认知。研究^[72]系统调研了 CoT 技术在不同任务上的表现。如图 4-2 所示，在处理涉及数学运算或符号推理等具有较高逻辑思维要求的任务时，相比于直接回答问题，CoT 方法展现出了显著的优势，能够有效地提升大语言模型的任务执行性能。然而，在处理其他类型的任务时，CoT 方法所带来的性能提升效果则相对有限，甚至出现一定程度的下降。研究^[73]探究了推理步骤数量和模型性能之间的关系，发现少示例 CoT 的设定下，即便不添加新信息，通过延长示例中的推理步骤，在多个数据集上显著提升了模型的推理能力；反之，缩短推理步骤，即使保留关键信息，也会大幅削弱模型的推理能力。此外，即使延长的推理步骤中包含错误信息，只要保持所需的推理长度，仍可能产生有利结果；最后，增加推理步骤的优势具有任务依赖性，简单任务所需推理步骤较少，复杂任务则能从较长的推理序列中获得显著提升。研究^[74]探究了推理模型如 DeepSeek-R1 的思考长度和答案正确性的关系，发现推理模型的准确性并非随思考长度增加而提升，且正确答案的长度通常比错误答案短；进一步研究发现，长思考过程包含更多自我修正，通常会导致性能下降。尽管经验性总结的解释能提供直观的定性认知，但其结论往往依赖于具体实验设置，缺乏严格的理论支持，因此在不同任务和模型上的泛化性仍需进一步验证。

总体而言，常用特征归因算法在大语言模型这种迭代推断模型上难以直接适用；尽管已有专为大语言模型设计的可解释性方法尝试从不同层面进行解释，但整体效果仍未达到预期，研究尚处于早期阶段。尤其在如何定量评估中间推理步骤对最终答案的贡献，相关工作仍较匮乏，是当前研究的关键难点。本章的工作正是在上述背景下展开，基于合作博弈论中的 Shapley 值，定量解释大语言推理模型答案生成中用户输入和中间推理步骤作用，并结合不同数据集的统计数据，进一步解释了大语言推理模型的特点、适用范围和局限性。

4.3 基于 Shapley 值的用户输入与中间推理步骤贡献量化算法

大型语言模型的输出过程是一个迭代推断过程，其中用户输入和部分模型输出内容（即中间推理步骤）都会对最终答案的生成产生影响。为了解释这些不同部分在最终答案生成中的作用，本节将最终答案生成过程建模为一个合作博弈，并将用户输入和中间推理步骤视为两个参与者，效用函数是最终答案的生成概率的对数均值。

具体来说，本节内容包括以下几个部分：首先明确效用函数的定义，接着描述为公平量化用户输入与中间推理步骤对最终答案生成的贡献的 QT Shapley 算法的设计与实现，最后描述基于模型共识的价值评估指标的设计，以期推理模型的答案生成的特征归因提供一个全面且有效的分析框架。

4.3.1 效用函数的定义

困惑度（Perplexity, PPL）是衡量语言模型建模能力的常用指标。具体地，对于任意给定的句子 s （例如最终答案或其他文本），其困惑度定义为：

$$\text{PPL}(s) = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log p(s_i | s_{<i}) \right). \quad (4-1)$$

其中， N 是句子的长度， s_i 表示句子中的第 i 个词， $s_{<i}$ 代表该词之前的所有词， $p(s_i | s_{<i})$ 是模型在给定上下文 $s_{<i}$ 后预测第 i 个词的条件概率。

困惑度的物理意义是：对于每个单词，模型平均认为有多少种不同的合理选择。具体来说，困惑度值越低，表示模型的预测越准确；反之，困惑度值越高，表示模型的预测越不确定。例如，当 $\text{PPL}(s) = 1$ 时，意味着模型对于句子 s 中的每个单词都有确定的预测概率（即每个单词的条件概率为 1）；而当 $\text{PPL}(s) = 10$ 时，表示模型平均认为每个单词的预测概率约为 $\frac{1}{10}$ ，即模型的预测较为模糊，存在较大的不确定性。

进一步，将损失函数定义为：

$$\text{loss}(s) = \ln \text{PPL}(s). \quad (4-2)$$

本章采取困惑度作为效用函数的基础，将效用函数定义为损失函数的负数：

$$U(s) = -\text{loss}(s) = -\ln \text{PPL}(s) = \frac{1}{N} \sum_{i=1}^N \log p(s_i | s_{<i}). \quad (4-3)$$

给困惑度取对数的原因是此时句子 s 的效用函数可以分解为每个词 s_i 的效用函数的加和，使得 Shapley 值的使用条件线性公理得到满足；取负号的原因是合作博弈论中效用函数的值越高代表合作获得的正面收益越大。

算法 2 计算用户输入与中间推理步骤贡献的算法：QT Shapley**输入：** 数据样本 $s = (q, t, a)$ ，大语言模型**输出：** 中间推理步骤和用户输入的贡献及贡献比例

- 1: 前向传播：计算 $\text{loss}(a|\emptyset) = -\frac{1}{|a|} \sum_{i=1}^{|a|} \log p(a_i|\emptyset)$
- 2: 前向传播：计算 $\text{loss}(a|q) = -\frac{1}{|a|} \sum_{i=1}^{|a|} \log p(a_i|q)$
- 3: 前向传播：计算 $\text{loss}(a|t) = -\frac{1}{|a|} \sum_{i=1}^{|a|} \log p(a_i|t)$
- 4: 前向传播：计算 $\text{loss}(a|q, t) = -\frac{1}{|a|} \sum_{i=1}^{|a|} \log p(a_i|q, t)$
- 5: 计算 think_value 和 query_value :

$$\text{think_value} = \frac{1}{2} [\text{loss}(a|\emptyset) - \text{loss}(a|t) + \text{loss}(a|q) - \text{loss}(a|q, t)]$$

$$\text{query_value} = \frac{1}{2} [\text{loss}(a|\emptyset) - \text{loss}(a|q) + \text{loss}(a|t) - \text{loss}(a|q, t)]$$

- 6: 计算 think_ratio 和 query_ratio :

$$\text{think_ratio} = \frac{\text{think_value}}{\text{loss}(a|\emptyset)}, \quad \text{query_ratio} = \frac{\text{query_value}}{\text{loss}(a|\emptyset)}$$

- 7: 返回 $\text{think_value}, \text{query_value}, \text{think_ratio}, \text{query_ratio}$

此外，本研究还需要关注在给定问题 q 条件下生成答案 a 的效用，此时效用函数可以记作 $U(a|q)$ ，即：

$$U(a|q) = -\text{loss}(a|q) = -\ln \text{PPL}(a|q) = \frac{1}{|a|} \sum_{i=1}^{|a|} \log p(a_i|a_{<i}, q). \quad (4-4)$$

4.3.2 QT Shapley 算法的设计与实现

通过将大语言模型推理得到最终答案过程看成用户输入和中间推理步骤两个参与者的合作博弈，将最终回答最终答案的损失函数的负数定义为效用函数（即效用值越高表示合作的效果越好），我们可以量化用户输入和中间推理步骤对最终答案生成的贡献。具体符号体系如下：

- q : 用户输入（query）。
- t : 模型生成的中间推理步骤（think）。
- a : 模型生成的最终答案（answer）。
- $U(q) = -\text{loss}(a|q)$: 仅使用用户输入生成最终答案的损失函数的负数。
- $U(t) = -\text{loss}(a|t)$: 仅使用中间推理步骤生成最终答案的损失函数的负数。
- $U(q, t) = -\text{loss}(a|q, t)$: 同时使用用户输入和中间推理步骤生成最终答案的损失函数的负数。
- $U(\emptyset) = -\text{loss}(a|\emptyset)$: 不使用任何信息生成最终答案的损失函数的负数。

根据 Shapley 值定义，参与者的贡献值计算需遍历所有可能的参与者组合。对于双参与者系统，其标准公式可以表示为：

$$\phi(i; U) = \frac{1}{2} \sum_{k=1}^2 \binom{1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{i\} \\ |S|=k-1}} [U(S \cup \{i\}) - U(S)]. \quad (4-5)$$

其中 $D = \{q, t\}$ 为全体参与者集合。

中间推理步骤的 Shapley 值计算如下：

$$\begin{aligned} \text{think_value} &= \frac{1}{2} [(U(t) - U(\emptyset)) + (U(q, t) - U(q))] \\ &= \frac{1}{2} [(-\text{loss}(a|t) + \text{loss}(a|\emptyset)) + (-\text{loss}(a|q, t) + \text{loss}(a|q))] \\ &= \frac{1}{2} [\text{loss}(a|\emptyset) - \text{loss}(a|t) + \text{loss}(a|q) - \text{loss}(a|q, t)]. \end{aligned} \quad (4-6)$$

用户输入的 Shapley 值计算如下：

$$\begin{aligned} \text{query_value} &= \frac{1}{2} [(U(q) - U(\emptyset)) + (U(q, t) - U(t))] \\ &= \frac{1}{2} [(-\text{loss}(a|q) + \text{loss}(a|\emptyset)) + (-\text{loss}(a|q, t) + \text{loss}(a|t))] \\ &= \frac{1}{2} [\text{loss}(a|\emptyset) - \text{loss}(a|q) + \text{loss}(a|t) - \text{loss}(a|q, t)]. \end{aligned} \quad (4-7)$$

中间推理步骤和用户输入的 Shapley 值加和为：

$$\text{think_value} + \text{query_value} = \text{loss}(a|\emptyset) - \text{loss}(a|q, t). \quad (4-8)$$

由于不同数据集最终答案的长度、困惑度分布不同，为了在不同数据集之间比较用户输入与中间推理步骤的贡献，对贡献进行归一化。 think_ratio 表示中间推理步骤对最终答案的贡献比例， query_ratio 表示用户输入对最终答案的贡献比例， $\text{loss}(a|\emptyset)$ 表示不使用任何信息生成最终答案的损失，作为基准值：

$$\text{think_ratio} = \frac{\text{think_value}}{\text{loss}(a|\emptyset)}, \quad (4-9)$$

$$\text{query_ratio} = \frac{\text{query_value}}{\text{loss}(a|\emptyset)}. \quad (4-10)$$

除此之外，本章还用 $\text{qt\&a_diff}$ 表示使用用户输入和中间推理步骤联合生成最终答案的损失相对于基准值的比值，值越大意味着模型生成最终答案时越少地依赖了用户输入和中间推理步骤的内容，即用户输入和中间推理步骤两者和最终答案之间的差异越大：

$$\text{qt\&a_diff} = \frac{\text{loss}(a|q, t)}{\text{loss}(a|\emptyset)}. \quad (4-11)$$

结合式4-8、式4-9、式4-10和式4-11可知：

$$\text{think_ratio} + \text{query_ratio} + \text{qt\&a_diff} = 1. \quad (4-12)$$

本研究提出的 QT Shapley 算法便是用 query_ratio 和 think_ratio 分别量化用户输入和中间推理步骤在生成最终答案过程中的相对贡献，从而为大语言推理模型的答案生成提供定量解释。QT Shapley 算法的流程见算法 2。计算复杂度方面，对于一条样本，QT Shapley 算法需要四次前向传播计算对应的损失（但不需要解码），时间复杂度上处于合理范围。

4.3.3 贡献计算方式评估指标的设计

本研究的另一个难点在于当中间推理步骤和用户输入的贡献无标答时，如何评估贡献计算方式的准确性。为了解决这个难题，本研究参考针对指令数据筛选任务设计的 Superfiltering^[75] 算法。

Superfiltering 算法通过 IFD（Instruction-Following Difficulty）分数对指令数据进行排序。其中 IFD 分数的计算方式等价于 $\text{loss}(a|q) - \text{loss}(a|\text{empty})$ ，衡量了指令 q 的存在为回复 a 的不确定性增大程度；IFD 分数越高说明模型越难遵循指令 q ，选择的优先级越高。实验表明，IFD 指标展现出显著的模型无关性——从轻量级的 GPT-2 (124M) 到大规模 LLaMA2-7b 等五种不同架构和规模的模型，在相同数据集上产生的 IFD 排序结果保持了高度一致性（斯皮尔曼相关系数达 0.8）。这种不同模型之间的共识暗示了 IFD 分数捕捉的是数据内在的本质属性，而非模型特定的偏好或偏差。

更具说服力的是后续的迁移实验：当使用小模型（GPT-2, 124M）筛选出的 5% 高质量数据（IFD 分数最高的前 5%）来训练更大的 LLaMA2-13B 模型时，其表现反而超越了全量数据训练的效果。实验证实了“指令跟随难度”确实是数据本身的固有属性，能够通过小模型的评估准确识别，并适用于指导大模型的训练。这两个实验确凿地验证了 IFD 分数的优越性：IFD 分数反映的指令跟随难度是数据的内在特征，并且不同模型对此形成了高度一致的判断共识。

受 Superfiltering^[75] 研究的启发，本研究提出了一种基于模型共识的贡献计算方式评估指标，通过计算不同模型在相同数据集上的价值排序的斯皮尔曼相关系数，来评估不同贡献计算方式的优劣。

具体地，斯皮尔曼相关系数作为非参数统计量，用于衡量两个变量排序间的单调相关性。其数学定义为：

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (4-13)$$

其中， d_i 表示样本在两个排序中的秩次差， n 为样本总量。该系数取值区间为 $[-1, 1]$ ，数值越接近 1 表示排序一致性越高。

更加具体地，采用斯皮尔曼相关系数作为针对评估用户输入和中间推理步骤贡献计算方式的评估指标的合理性有三点：

(1) 斯皮尔曼相关系数能够衡量数据特征与模型解耦程度：若数据的内在特征具有模型无关性（model-agnostic），则不同规模模型对同一数据集产生的排序应保持高度一致性。通过斯皮尔曼相关系数量化这种跨模型一致性，可以有效验证所提出的贡献计算方法是否真正捕捉到数据本质特征，而非特定模型的偏好。

(2) 斯皮尔曼相关系数可用于无监督评估：由于中间推理步骤和用户输入的价值缺少公认的真实值（Ground Truth），传统监督式评估方法难以适用。斯皮尔曼相关系数提供了一种基于模型共识的无监督评估框架，其原理类似于“多盲一致”——当多个独立模型（尤其是不同规模的模型）对数据排序表现出高度一致性时，可认为该排序结果具有较高的可信度。

(3) 斯皮尔曼相关系数高的特征能支撑高效地数据选择：跨模型稳定的数据特征可以支持“小模型筛选-大模型训练”的实用范式。例如，研究者可以使用一个小参数量模型（如 1.5B 参数规模）来进行数据筛选，并在较大参数量的模型（如 32B 参数规模）上进行训练。当小模型产生的排序与大模型保持高度一致时，在小模型上进行数据选择不仅能确保筛选质量，还可大幅降低计算开销。因此，有必要研究跨模型稳定系数较高的数据特征（即根据不同模型的排序结果计算出的斯皮尔曼相关系数较高的数据特征）。

通过上述分析可见，斯皮尔曼相关系数不仅能够有效评估贡献计算方法的准确性，还为构建高效的数据筛选体系提供了潜在支撑。后续实验将基于斯皮尔曼相关系数评估不同用户输入和中间推理步骤贡献计算方式的准确性。

4.4 实验结果与分析

4.4.1 实验设置

表 4-1 Chinese-Data-Distill-From-R1 数据集统计信息

Table 4-1 Statistics of the Chinese-Data-Distill-From-R1 Dataset

分类	数据集名称	样本量	query token 中位数	think token 中位数	answer token 中位数	比值 中位数	简介
General	coig/neo	52,893	18	448	620	0.76	日常对话数据集
	zhihu/zhihu_score9.0-10_clean_v10	2,534	25	473	715	0.67	知乎问题集
	xhs/xhs	1,507	24	420	519	0.81	小红书内容创作指令集
	ruozhiba/ruozhiba_ruozhiba	240	15	458.5	394.5	1.18	弱智吧问题集
	human_value/100poison	764	12	406.5	636	0.65	诱导偏见、歧视的问题集
	logi_qa/logi_qa	414	91.5	2329	338.5	7.27	逻辑推理问题集
Math	Haijian/Advanced-Math	570	22	1063	390	2.92	高等数学题库
	meta-math/GSM8K_zh	8,776	45	610	153	4.22	小学数学题库
	EduChat-Math	19,729	55	889	244	3.90	义务教育数学题库
	gavinluo/applied_math	7,493	32	612	160	3.90	应用数学题库
STEM	stem_zh/chem	3,157	13	577	748	0.76	化学学科题库
	stem_zh/bio	3,147	15	506	830	0.61	生物学科题库
	stem_zh/med	3,163	16	503	802	0.63	医学学科题库
	stem_zh/phy	3,181	13	590	775	0.76	物理学科题库
Exam	exam/coig_exam	1,954	97	486	218	2.31	高考、中考和公务员考试题库
	exam/kaoyan	377	66	772	288	2.89	考研真题题库
	human_value/coig_human	101	48	423	182	2.38	人类价值观题库

注：第 7 列的比值定义为 think token 除以 answer token 数

实验采用的数据集的名称是 Chinese-Data-Distill-From-R1^[76]，共包含 11 万条 DeepSeek R1 中文问答数据。该数据集的构建过程是：首先收集多个现有且来自不同领域的数据集的问题，然后将这些问题输入到全精度部署的 671B 参数的 DeepSeek R1 模型中，并收集其生成的回答。

Chinese-Data-Distill-From-R1 的制作者将搜集到的 17 个数据集按照四个大类划分：数学推理、学科考试、STEM 专业领域及通用场景对话。每个大类包含的数据集的具体的细节可见表格4-1，附录二提供了这 17 个数据集的样本示例。这些问题覆盖了常见的数学推理、知识问答、逻辑推理、创作和安全等不同领域。四个大类的特性描述如下：

- **通用场景 (General)**: 58,352 个样本，集成日常对话、知乎专业讨论、小红书创作、弱智吧反讽逻辑、社会偏见检测、多步逻辑推理等通用场景的数据。
- **数学推理 (Math)**: 共 36,568 个样本，包含代数、几何、概率统计等子领域，内容涵盖基础运算到复杂定理证明，难度涵盖小学到高等数学。
- **STEM 知识问答 (STEM)**: 12,648 个样本，涉及物理、化学、医学、生物专业领域的理论和知识问答。
- **考试题库 (Exam)**: 2,432 个样本，包含中考、高考、考研、公务员考试和人类价值观题库。

实验采用的神经网络是 DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B 这三个不同参数量的开源模型^[64]，用于计算用户输入与中间推理步骤价值。这三个模型是由 DeepSeek 团队分别在 Qwen2.5-Math-1.5B^[77]、Qwen2.5-Math-7B^[77]、Qwen2.5-14B^[78] 的基础上监督微调 (Supervised Fine-Tuning, SFT) 得到，使用的数据与 DeepSeek-R1 训练过程中最后一次监督微调阶段所用的 800k 条数据相同。其它网络架构如 GPT、LLaMA 尚无对应开源推理模型，因此本研究暂不考虑。

4.4.2 对比算法

对于用户输入和中间推理步骤的贡献计算，容易根据 IFD 分数^[75] 结合不同归一方式，得到若干种启发式的算法作为替代方案。这些算法（由于种类繁多，仅仅按照从简单到复杂的顺序命名）从不同角度对用户输入和中间推理步骤的贡献进行量化分析。简要介绍如下：

针对于中间推理步骤贡献度量，有以下替代方案：

- (1) $\text{think_value2} = \text{loss}(a|q) - \text{loss}(a|q,t)$ ，量化了已有用户输入时，中间推理步骤对最终答案不确定性的降低程度。
- (2) $\text{think_ratio2} = \text{think_value2} / \text{loss}(a|q)$ ，即以 $\text{loss}(a|q)$ 为基准值对 think_value2 进行归一。
- (3) $\text{think_ratio3} = \text{think_value2} / \text{loss}(a|\emptyset)$ ，即以 $\text{loss}(a|\emptyset)$ 为基准值对 think_value2 进行归一。
- (4) $\text{think_ratio4} = \text{think_value} / (\text{think_value} + \text{query_value})$ ，衡量了中间推理

步骤在中间推理步骤和用户输入对最终答案不确定性的降低程度中的比重。

针对于用户输入贡献度量，有以下替代方案：

(1) $\text{query_value2} = \text{loss}(a, t|\emptyset) - \text{loss}(a, t|q)$ ，量化了用户输入对中间推理步骤和最终答案联合不确定性的降低程度，类似于 IFD 分数^[75]的计算方式。

(2) $\text{query_ratio2} = \text{query_value2} / \text{loss}(a, t|\emptyset)$ ，以 $\text{loss}(a, t|\emptyset)$ 为基准值来对 query_value2 归一。

(3) $\text{query_value3} = \text{loss}(a|t) - \text{loss}(a|q, t)$ ，量化了在已有中间推理步骤的情况下，引入用户输入对最终答案不确定性的降低程度。

(4) $\text{query_ratio3} = \text{query_value3} / \text{loss}(a|t)$ ，即以 $\text{loss}(a|t)$ 为基准值对 query_value3 进行归一。

(5) $\text{query_ratio4} = \text{query_value3} / \text{loss}(a|\emptyset)$ ，即以 $\text{loss}(a|\emptyset)$ 为基准值对 query_value3 进行归一。

(6) $\text{query_ratio5} = \text{query_value1} / (\text{think_value} + \text{query_value})$ ，量化了用户输入在中间推理步骤和用户输入两者为最终答案不确定性带来的降低所占的比例。

在实验中，本研究采用以上替代方案作为基线对比算法。

4.4.3 定量实验

如表4-2所示，在中间推理步骤贡献度量方面， think_ratio 的计算方式在 1.5B vs 14B 和 7B vs 14B 的两种设置下中斯皮尔曼相关系数分别达到 0.9369 和 0.9518，显著优于其他方法。这验证了 QT Shapley 算法通过利用 Shapley 值公平分配贡献的特性，更成功地刻画到了中间推理步骤对最终答案的贡献。

表 4-2 不同中间推理步骤贡献计算方式的斯皮尔曼相关系数

Table 4-2 Spearman's correlation coefficients for various methods of calculating think contributions

模型设置	think_ratio	think_value	think_ratio2	think_value2	think_ratio3	think_ratio4
1.5B vs 14B	0.9369	0.8381	<u>0.8423</u>	0.6785	0.8068	0.7559
7B vs 14B	0.9518	0.8299	<u>0.8666</u>	0.7231	0.8433	0.8087
14B vs 14B	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

注：最优结果以粗体标注，次优者以下划线标注。模型类别省略 DeepSeek-R1-Distill-Qwen 的前缀

如表4-3所示，在用户输入贡献度量中， query_ratio 和 query_ratio2 表现最优。但 query_ratio2 的物理含义和中间推理步骤的贡献存在差异——它实际衡量的是用户输入对“中间推理步骤-最终答案联合生成”的贡献，而非纯粹的用户输入对最终答案生成贡献。此外，直接使用 think_value 和 query_value 等原始差值指标不如 think_ratio 和 query_ratio 也说明了归一化的必要性：不同样本的最终答案长度、领域类别等特性分布不一，需要将贡献转换为相对于基准损失 ($\text{loss}(a|\emptyset)$) 的比率，以消除样本基础特性的影响。

以上实验结果表明，基于合作博弈论中 Shapley 值的中间推理步骤和用户输入贡献的 QT Shapley 算法能够在不同模型设置下有效捕捉到数据本质特征，公

表 4-3 不同用户输入贡献计算方式的斯皮尔曼相关系数

Table 4-3 Spearman's correlation coefficients for various methods of calculating query contributions

模型设置	query_ratio	query_value	query_ratio2	query_value2	query_ratio3	query_value3	query_ratio4	query_ratio5
1.5B vs 14B	0.9188	0.8946	<u>0.9133</u>	0.9000	0.3829	0.2740	0.3587	0.7559
7B vs 14B	<u>0.9351</u>	0.9138	0.9426	0.9344	0.5713	0.5476	0.5787	0.8087
14B vs 14B	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

注：最优结果以粗体标注，次优者以下划线标注。模型类别省略 DeepSeek-R1-Distill-Qwen 的前缀

平衡量用户输入和中间推理步骤在最终答案生成中的贡献。相比之下，直观设计的基线算法因缺乏对贡献分配机制的系统性建模，实际效果略差于本研究提出的方法。

4.4.4 定性实验

这一节首先展示用户输入与中间推理步骤贡献比例高与低的典型数据的对比，然后根据上一节得到的定量实验数据，将中间推理步骤和用户输入贡献结合数据集特性进行分析，定性说明现有推理模型适用的场景和局限性。

(1) 用户输入与中间推理步骤贡献比例高与低的典型数据对比

“谜题”数据的 think_ratio 是 0.724333 而 query_ratio 是 0.0625

query:请创造一些谜语，答案都是“驾校老板”。

think:<think> 好的，用户让我创造一些谜语，答案都是“驾校老板”。首先，我需要理解这个需求。用户可能是在准备一些有趣的活动，比如聚会游戏或者教学材料，需要一些与驾校相关的谜语，答案统一指向驾校老板。谜语需要既符合逻辑，又要有一定的难度，但不要太难，否则可能让人摸不着头脑。

接下来，我要考虑谜语的结构。通常谜语包括谜面和谜底，谜面要用比喻、双关或者其他修辞手法来暗示答案。因为答案是“驾校老板”，所以需要围绕驾校的特点和老板的角色来构思。

首先，驾校的关键元素包括驾驶、教练车、学员、考试、交通规则、方向盘、油门、刹车等等。老板的角色可能涉及管理教练、招生、安排课程、通过率等。需要将这些元素结合起来，形成有趣的谜面。

比如，第一个谜语可以结合方向盘和决策权，因为老板通常掌握方向，管理整个驾校。所以谜面可能是“手握方向不驾车，学员成败他牵挂”，这样既点明了不亲自驾车，又提到学员的成败，符合老板的角色。

第二个谜语可以考虑老板对规则的强调，因为驾校必须遵守交通规则，老板需要确保教学符合标准。比如“他虽不踩油门刹，规则条条心中挂”，这里用油门和刹车代表教学细节，规则是核心。

第三个谜语可以结合考试和证书，老板虽然不亲自考试，但学员的通过率影响他的生意。比如“考场不见他身影，证书签发有他名”，暗示他在幕后负责。

第四个谜语可以用教练和学员的关系，老板管理教练团队，所以“手下教练排成行，学员叫他大当家”，大当家体现领导地位。

第五个谜语可以玩转“路”的双关，老板为学员铺就驾驶之路，所以“不为自家铺财路，专为他人把路铺”。

接下来要检查每个谜语是否准确指向驾校老板，有没有可能被误解成教练或其他角色。例如，第一个谜语中的“手握方向”可能让人想到教练，但后面提到“学员成败他牵挂”更倾向于管理者，即老板。需要确保每个谜语的关键点都明确指向老板而非教练。

另外，要注意语言的押韵和节奏，使谜语朗朗上口。比如使用对仗的结构，如“手握方向不驾车，学员成败他牵挂”前后句结构对称，容易记忆。

可能还需要考虑谜语的多样性，从不同角度切入，避免重复。比如有的谜语侧重管理，有的侧重规则，有的侧重结果（证书），这样整体看起来更丰富。

最后，确认每个谜语的答案确实是“驾校老板”，没有歧义，同时谜面不至于直白，保持一定的挑战性。这样用户在使用时既能引发思考，又不会太难而让人放弃。

answer:1. ** 谜面 **: 手握方向不驾车，学员成败他牵挂。（谜底：驾校老板）\n\n 2. ** 谜面 **: 他虽不踩油门刹，规则条条心中挂。（谜底：驾校老板）\n\n 3. ** 谜面 **: 考场不见他身影，证书

签发有他名。（谜底：驾校老板）\n\n 4. ** 谜面 **：手下教练排成行，学员叫他”大当家”。（谜底：驾校老板）\n\n 5. ** 谜面 **：不为自家铺财路，专为他人把路铺。（谜底：驾校老板）\n\n 每个谜语均以驾校的特点（如方向盘、教练、考试规则等）为切入点，最终指向掌控全局的驾校老板。

“送别信”数据的 think_ratio 是 0.018843 而 query_ratio 是 0.018843

query:我想给即将离开这个城市的挚友写一封送别信，能帮我构思一下内容吗？

think:<think> 好的，用户想给即将离开的挚友写一封送别信，需要帮忙构思内容。首先，我需要理解用户的需求。他们可能希望这封信既表达不舍之情，又充满祝福，同时回忆一些共同的经历，让朋友感受到真诚和温暖。接下来，我得考虑用户和挚友的关系。既然是“挚友”，说明他们之间有很多深厚的回忆和情感。信中需要包含具体的例子或事件，这样会更个性化，也更能打动对方。同时，用户可能希望避免过于悲伤，而是强调积极的前景和对未来的期待。然后，结构方面，送别信通常包括几个部分：开头表达对离别的感受，中间回忆共同经历，表达感谢，最后送上祝福和保持联系的愿望。需要确保逻辑流畅，情感真挚，用词恰当。可能用户没有明确提到的深层需求是，他们希望这封信能成为朋友珍藏的纪念品，所以需要一些诗意的语言或比喻，比如星辰、旅途、季节等意象，增加信的文学性和感染力。同时，用户可能希望信的长度适中，不过于冗长，但足够表达所有情感。还需要注意语气，既要真诚又要不过于沉重，保持积极向上的基调。可能需要提醒用户根据实际情况调整具体内容，比如添加只有他们知道的内部梗或特定回忆，这样会更贴心和个性化。最后，检查是否有遗漏的部分，比如是否需要提及未来重逢的期待，或者提供联系方式以便保持联络。总之，目标是让朋友感受到被珍视，即使分开，友情依旧长存。</think>

answer: \n\n \n\n 当然可以。以下是一封送别信的框架和内容建议，既有情感表达，又能体现你们的独特回忆，你可以根据实际情况调整细节：\n\n — \n\n ** 标题 **：《此去星辰非远路，相逢时节再添香》\n\n ** 开头（抒写离别当下的感受） ** 亲爱的 [挚友名字]：窗外的 [城市名] 又落雨了，像极了我们初遇时那个湿漉漉的午后。但此刻手中的笔，却比往日更重几分——原来有些道别，光是写下第一个字，眼眶便已发烫。 \n\n ** 中段 1（回忆共同经历，突出细节） ** 记得吗？那年深夜的 [具体地点：如天台/路边摊]，我们对着 [某个象征物：如街灯/烤红薯的炉火] 发誓要 [一起完成的某个约定]。虽然 [约定是否达成？如“终究没能追上末班车”或“真的在毕业前吃遍整条小吃街”]，可那些肆意的笑声，至今仍在我心底回响。是你教会我，[某个特质：如“笨拙的坚持比聪明的放弃更动人” / “雨天不撑伞才算浪漫”]——这些荒唐又珍贵的小事，早已成了我生命里的光。 \n\n ** 中段 2（表达感谢与联结感） ** 这座城市的地图即将被折叠成你行囊里的风景，但请记得：当你在 [对方将去的地方：如北方的雪夜/南方的海岛] 感到孤独时，我依然是你手机里永远秒回的树洞；当你发现 [某个共同爱好：如某家奶茶店的新品/某部电影续作] 时，我们的聊天框永远值得一次「爆笑刷屏」特权；而每个 [对你们有意义的时间点：如跨年零点/某个纪念日]，地球两端的我们，仍会抬头望向同一片月光。 \n\n ** 结尾（祝福与展望未来） ** 不必说再见，因为真正的告别从不需要仪式。愿你的列车穿过晨雾时，载满勇气与好奇；愿异乡的灯火为你酿出更醇厚的故事；愿某天我们在街角重逢时，依然能像十七岁那样，为了一碗 [家乡特色食物] 的口味争得面红耳赤。 \n\n ** 署名 **：永远与你共享心跳的 [你的名字] [日期] \n\n — \n\n ** 点睛小贴士 **： 1. ** 埋入专属暗号 **：在信中穿插只有你们懂的梗（如某次糗事、常用口头禅）； 2. ** 实物附件 **：随信附上 [拍立得合影/一起收藏的车票/种子卡片等]； 3. ** 数字化延伸 **：在信封角落画个二维码，链接到为 TA 定制回忆的视频网站。 \n\n 愿这封信成为 TA 新旅程中最温暖的行李。

以上的两个数据均来自 coig/neo 数据集，且可归类于“创作”类型数据。用户输入和中间推理步骤贡献比例（think_ratio 和 query_ratio）的计算使用的模型是 DeepSeek-R1-distill-qwen-7B。两条数据的主要区别在于中间推理步骤的贡献程度差异显著：在“谜语”数据中，中间推理步骤的贡献比例约为 70%，而在“送别信”数据中，该比例仅为约 1%。

具体地，“谜语”数据中的中间推理部分详细描述了五个谜题的创作过程，包括谜面和谜底，且与最终生成的五个谜题严格对应，甚至部分内容已经在中间

推理中完整呈现。因此，可以认为该条数据中中间推理步骤在答案生成中起到了主导作用。相比之下，“送别信”数据的最终答案中包含大量文风优美、但未出现在中间推理过程中的句子，说明其主要内容来源于模型预训练中习得的语言能力，中间推理的作用相对较弱。这一对比验证了 QT Shapley 算法在评估中间推理步骤贡献方面的准确性。

此外，这两条数据中用户输入的贡献比例均较低，原因在于这两条数据的中间推理步骤前段复述了用户输入，从而在一定程度上“分担”了其贡献。相较之下，在下面来自 EduChat-Math 数据集的“口算”的数据中，用户输入的贡献比例为极高的 44%。其原因在于答案在很大程度上和用户输入重合，即模型答案几乎是对输入的直接加工，极大地依赖输入内容。因此，该例也验证了 QT Shapley 算法在衡量用户输入贡献方面的合理性。

“口算”数据的 think_ratio 是 0.43666 而 query_ratio 是 0.442212

query: 口算 \n\n | \$25+125=?\$ | \$127+268 \approx ?\$ | \$285+125=?\$ | \$370+250=?\$ | \$250+250=?\$ | | :— | :— | :— | :— | :— | | \$320-160=?\$ | \$520-340=?\$ | \$750-360=?\$ | \$823-199 \approx ?\$ | \$340-170=?\$ |

think: <think> 嗯，好的，我现在需要口算这些加减法的题目，还要给出近似值。让我一个一个来，仔细想想。可能有些容易的，有些可能需要多想想，特别是那些需要估算的。首先，从头开始看，左边那一列是加法，右边是减法，对吗？\n\n 首先第一个题目是 25 加 125。这应该不难，25 加 100 是 125，再加上剩下的 25，所以 125 加 25 等于 150？或者直接拆的话，25 加 125 的话，可能 25 和 125 的个位都是 5，加起来是 10，然后十位和百位的话，20 加 120 是 140，加上个位的 10，总共是 150。对，没错，第一个应该是 150。 \n\n 接下来是 127 加 268 $\approx$ ，这里题目要的是近似值。那可能需要先估算这两个数。比如说，把 127 近似到 130，268 近似到 270，这样 130 加 270 等于 400。或者可能四舍五入到最近的百位的话，127 是 100，268 是 300，那 100+300=400。这样结果大约是 400。不过可能更精确的方式是近似到十位，所以 127 近似为 130，268 近似为 270，加起来 130+270=400。所以结果应该大约是 400。不过可能需要确认一下实际的和是多少，比如 127+268，实际计算的话，个位 7+8=15，进一位；十位 2+6=8，加上进位的 1 是 9；百位 1+2=3，所以总和是 395。那如果四舍五入的话，可能估算到 400 也合理。所以第二个应该是 ≈ 400 。 \n\n 第三个题目是 285 加 125。这个的话，285 加 100 是 385，再加 25，就是 410。或者分解计算：个位 5+5=10，进一位；十位 8+2=10，加上进位 1 是 11，进一位；百位 2+1=3，加上进位 1 是 4。所以总和是 410。第三个答案应该是 410。 \n\n 第四个是 370 加 250。这个可能比较简单，因为都是整十数。370+200=570，再加上 50 就是 620。或者直接十位相加：7+5=12，所以十位是 20，加上百位 3+2=5，所以总共有 500+120=620。没错，第四个是 620。 \n\n 第五个是 250+250，这应该是最简单的，因为两个都是 250，直接相加就是 500。没问题，对吧？所以第五个是 500。 \n\n 接下来看右边的减法部分。第一个是 320 减 160。这个可以用分解法：320 减去 100 是 220，再减去 60 就是 160。或者，个位 0 减 0 是 0，十位 2 减 6 不够的话，可能需要借位。但 320-160 的话，可以想象成 320-160=160，对吗？因为 160 加 160 等于 320。所以第一个结果是 160。 \n\n 第二个减法题目是 520-340。同理，520 减去 300 是 220，再减去 40 就是 180。或者分解计算：个位 0-0=0，十位 2-4 不够，借位，变成 12-4=8，这时候百位被借了一位，所以原来的 5 变成 4，再减去 3 等于 1。所以结果应该是 180。对吗，比如 340+180=520，确实是对的。所以第二个答案 180。 \n\n 第三个是 750-360。这个的话，750 减 300 是 450，再减 60 就是 390。或者个位 0-0=0，十位 5 减 6 不够，借位变成 15-6=9，这时候百位 7 变成 6，再减去 3 是 3，所以结果是 390。或者直接计算，360 加上 390 等于 750，所以正确。第三个结果 390。 \n\n 第四个是 823-199 $\approx$ ，这里需要估算。可能用四舍五入的方法，把 199 近似为 200，那么 823-200=623，但多减了 1，所以加上 1 就是 624。不过估算的话可能直接近似为 800-200=600，所以答案大约 600。或者更精确一点的话，823 近似到 800，199 近似到 200，结果就是 600。或者另一种方式，估算 823-199 $\approx$ 820-200=620。不过不同的近似方法可能结果不同，但通常来说，可能近似到最近的百位的话，就是 800-200=600，或者更准确的近似可能是 620。实际计算的话，823-199=624，

近似到十位是 620，或者百位是 600。题目里可能希望用哪种方式？比如，当题目中的两个数都是近似到十位的话，那可能答案会更接近实际的近似。比如把 199 近似为 200，那么 $823-200=623$ ，所以近似是 620 或 600。但可能更合理的估算方法是把 199 看成 200，所以结果大约是 623，大约 620。或者老师可能会教不同的方法，但这里可能需要按照题目的要求，可能四舍五入到百位的话，答案是 600。或者更接近的十位的话就是 620。可能要看题目要求，不过这里可能答案是约 600 或约 620。需要再仔细想想。比如，如果题目中的近似符号是 $\approx$ ，可能更倾向于用更精确的近似方法。例如，199 离 200 只差 1，而 823 离 800 差 23，所以可能更合理的是将 823 近似为 820，199 近似为 200，所以 $820-200=620$ ，所以答案是 ≈ 620 ？或者直接计算的话， $823-199=624$ ，近似到十位是 620，或者到百位是 600。但通常在学校里，可能估算时会把两个数都近似到最近的百位，所以 823 约 800，199 约 200， $800-200=600$ 。所以答案 ≈ 600 。这可能更符合题目的期望。可能要看题目的要求有没有特别的说明，但这里没有，所以可能两种方式都可以。不过我觉得更常见的是近似到百位，所以可能答案是 600。但实际计算的话，精确的是 624，可能近似到 620 更接近。不过可能题目希望用简单的估算方法，比如把 199 看作 200， $823-200=623\approx 620$ 或者 600。但通常可能选 600。我需要确定这个。比如，在考试中，如果题目没有特别说明，可能四舍五入到百位，所以答案是约 600。所以第四个答案是 ≈ 600 吗？或者是约 620？可能需要再想。比如，当估算的时候，可能老师会教学生将接近整百的数看作整百，所以这里 $199\rightarrow 200$ ， $823\rightarrow 800$ ，这样 $800-200=600$ 。或者另一个方法是，把两个数都保留到十位，比如 $823\rightarrow 820$ ， $199\rightarrow 200$ ，所以 $820-200=620$ 。所以可能答案是约 620。这可能更准确一些。比如，当计算 $823-199$ 时，可以看作 $823-200+1=623+1=624$ ，所以估算的时候，直接用 $823-200=623$ ，所以大约 620。所以更正确的近似是 620。不过可能根据不同的方法有不同的答案。但题目中的其他问题有没有类似的例子？比如前面第二个题目是 $127+268\approx$ ，这时候可能近似到十位，所以是 $130+270=400$ 。所以这里可能同样，将 199 近似为 200，823 还是保持 800 的话，可能得到 600。或者更接近的可能是 $820-200=620$ 。或者实际计算中， $823-199=624$ ，近似到十位是 620，到百位是 600。所以可能答案应该是 ≈ 620 。或者要看题目的要求。比如，如果是估算到最近的百位，就是 600；如果是最近的十位，就是 620。这里可能需要根据常规做法来，所以可能答案是 ≈ 620 。不过不确定，可能需要再检查一下。

最后一个题目是 $340-170$ 。这个比较容易，340 减 100 是 240，再减 70 就是 170。或者分解的话，个位 $0-0=0$ ，十位 4 减 7 不够，借位变成 $14-7=7$ ，百位 3 被借了一位变成 2，再减 1 等于 1。所以结果是 170。或者直接计算， $170+170=340$ ，所以正确。最后一个结果是 170。

现在检查一下各个题目的答案是否正确：

加法部分：

- $25+125=150 \rightarrow$ 正确。
- $127+268\approx 400$ （实际是 395，四舍五入到十位是 400，或者到百位是 400，这里可能正确）。
- $285+125=410 \rightarrow$ 正确。
- $370+250=620 \rightarrow$ 正确。
- $250+250=500 \rightarrow$ 正确。

减法部分：

- $320-160=160 \rightarrow$ 正确。
- $520-340=180 \rightarrow$ 正确。
- $750-360=390 \rightarrow$ 正确。
- $823-199\approx 620$ 或 600？这里可能要看题目的要求。假设题目希望近似到十位的话，答案是 ≈ 620 ，如实际 624 约等于 620。如果近似到百位的话，是 600。但可能更合理的做法是像前面的第二题，比如 $127+268\approx 400$ （实际是 395），所以可能这里也是用类似的方法，把 199 看作 200，那么 $823-200=623$ ，近似到十位是 620。因此答案是 ≈ 620 。或者，可能题目希望用更粗略的估算，如把 823 看作 800，199 看作 200，得 600。可能两种都对，但原题中给出的可能更接近哪一个？

比如，原来的第二个题目 $127+268\approx$ ，正确的近似到十位是 $130+270=400$ ，而实际是 395，误差 5。而这里 $823-199$ ，正确的做法可能是用同样的方法，即把每个数近似到最近的十位： $823\approx 820$ ， $199\approx 200$ ，所以 $820-200=620$ 。所以答案应该是 ≈ 620 。或者，另一种常见的估算方法是，当减数是 199 时，可以视为减 200，然后加 1，所以 $823-199=823-200+1=623+1=624$ 。这时如果估算的话，可能认为近似为 620。或者，如果题目中没有明确说明近似到哪一位，可能以百位为单位更合理，所以答案是 600。这有点不确定，可能要看老师怎么教。例如，在考试中，如果题目里的例子是如第二个题目，那么可能这里也采用相近的做法，所以答案是约 620。或者，可能题目中的近似符号 $\approx$ 在这里指四舍五入到最近的百位，所以 $823\approx 800$ ， $199\approx 200$ ，得到 600。这时候答案就是约 600。但需要根据常规判断，可能两种答案都有可能。不过在现实中，比如，当教孩子估算时，更常见的是调整减数到最近的整百，然后调整结果。例如， $823-199\approx 823-200=623$ ，因为多减了 1，所以实际结果应该是 $623+1=624$ ，但估算时可能直接说 ≈ 623 或 ≈ 620 。或者为了简便，直接说 ≈ 600 。不过这里可能需要再想，题目中的其他近似题可能更倾向于四舍五入到十位，所以把 199 近似到 200，823 近似到 820 的话， $820-200=620$ 。或者，不近似被减数，直接减 200 得到 623，然后作为近似。这时候可能更接近 620。例如，通常估算的话， $823-199\approx 800-200=600$ ，但这样误差比较大。或者更精确的近似是 $820-200=620$ 。我觉得可

能更倾向于后者，因为 199 离 200 只差 1，可能更准确的做法是调整减数到最近的整百，然后计算，所以答案是约 624，即近似到 620。或者直接估算为 620。所以这里可能答案应该是 ≈ 620 。不过我需要再确认。

另外，前面第二个题目中的 $127+268\approx$ ，正确的计算是 395，而近似到十位是 400。那么同理，这里 $823-199=624$ ，近似到十位是 620，所以应该填写 ≈ 620 。因此，第四个减法题的答案应该是 ≈ 620 。

最后一个题目是 $340-170=170 \rightarrow$ 正确。

现在总结所有答案：

加法部分：
 $25+125=150$ $127+268\approx 400$ $285+125=410$ $370+250=620$ $250+250=500$

减法部分：
 $320-160=160$ $520-340=180$ $750-360=390$ $823-199\approx 620$ （或者 600，但更合理的是 620） $340-170=170$

但可能题目中的近似方法是否一致？比如第二个加法题是 $127+268\approx$ ，实际是 395，四舍五入到十位是 400。而减法题中的 $823-199\approx$ ，实际是 624，四舍五入到十位是 620，到百位是 600。但如果加法题用的是四舍五入到十位得到 400，那么减法题也应该采用同样的方法，得到 620。所以这里答案应该是 ≈ 620 。

现在检查所有答案是否正确：

加法：1. 正确。2. 正确，四舍五入到十位是 400。3. 正确。4. 正确。5. 正确。

减法：1. 正确。2. 正确。3. 正确。4. 正确， ≈ 620 。5. 正确。

所以最终答案应该如下：

第一行加法： $150, 400, 410, 620, 500$

第二行减法： $160, 180, 390, 620, 170$

不过原题中的第四个减法题目是否需要四舍五入到百位？如果第二个加法题目是四舍五入到十位得到 400，那么减法题是否也应该四舍五入到十位得到 620？是的，这可能更一致。

总结下来，各个答案应该是正确的。

answer: $\$25+125=150\$$ | $\$127+268 \approx 400\$$ | $\$285+125=410\$$ | $\$370+250=620\$$ | $\$250+250=500\$$ |
 $\$320-160=160\$$ | $\$520-340=180\$$ | $\$750-360=390\$$ | $\$823-199 \approx 620\$$ | $\$340-170=170\$$ |

$\boxed{150}$ $\quad \boxed{400}$ $\quad \boxed{410}$ $\quad \boxed{620}$ $\quad \boxed{500}$
 $\boxed{160}$ $\quad \boxed{180}$ $\quad \boxed{390}$ $\quad \boxed{620}$ $\quad \boxed{170}$

(2) 推理模型适用领域分析

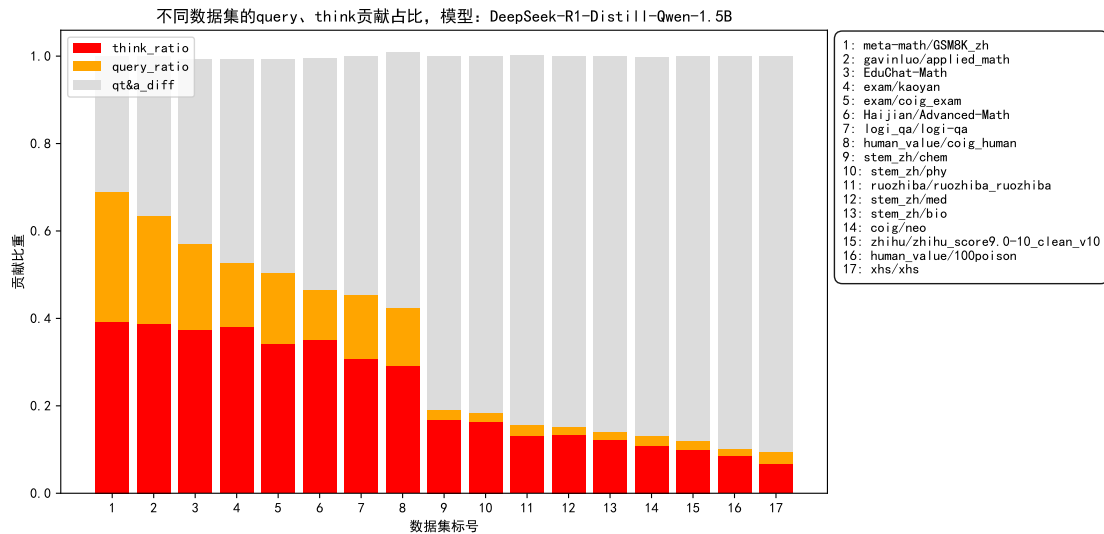


图 4-3 不同数据集的用户输入、中间推理步骤贡献占比

Figure 4-3 Distribution of query and think ratios across multiple datasets

图 4-3展示了 Chinese-Data-Distill-From-R1 项目中 17 个数据集的中间推理步骤对最终答案生成的贡献比例 (think_ratio)，用户输入对最终答案生成的贡献比例 (query_ratio)，以及用户输入和中间推理步骤两者和最终答案之间的差异 (qt&a_diff) 的中位数分布情况，用于统计的模型是 DeepSeek-R1-Distill-Qwen-1.5B。本节图标标题中默认模型为 DeepSeek-R1-Distill-Qwen-1.5B，为了简洁起见，将此模型名称省略。DeepSeek-R1-Distill-Qwen-7B 和 DeepSeek-R1-Distill-Qwen-

14B 的对应结果分别在附录一-1和附录一-2。

实验结果表明，DeepSeek-R1 在以下三类场景中表现突出：通过用户输入和中间推理步骤的合作，能够将最终答案生成的损失函数值降低约 50%，且其中中间推理步骤发挥了主导作用。

(1) 数学推理,对应数据集为: Haijian/Advanced-Math、meta-math/GSM8K_zh、EduChat-Math 以及 gavinluo/applied_math。

(2) 考试题库（包含价值观类选择题），对应数据集为：exam/coig_exam、exam/kaoyan 以及 human_value/coig_human。

(3) 逻辑推理，对应数据集为：logi_qa/logi_qa。

然而，DeepSeek-R1 在以下四类场景中效果受限，用户输入和中间推理步骤合作降低的最终答案损失函数值的幅度在 20% 左右（尽管中间推理步骤贡献占主要地位）：

(1) STEM 知识问答,对应数据集为: stem_zh/chem、stem_zh/bio、stem_zh/med 以及 stem_zh/phy。

(2) 日常问答,对应数据集为 coig/neo、zhihu/zhihu_score9.0-10_clean_v10 以及 ruozhiba/ruozhiba_ruozhiba。

(3) 创作类任务，对应数据集为：xhs/xhs。

(4) 安全（非选择题），对应数据集为：human_value/100poison。

对于以上现象可能的原因，本章给出两个猜测：

(1) 知识问答、日常聊天、创作、价值观的问题较少地依赖于模型推理能力。比如 STEM 学科、价值观的回复更依赖模型本身的知识（通过预训练获得），而非测试时增加输出长度来进行仔细地推理。

(2) 推理模型适用于知识问答、日常聊天、创作、安全，但 DeepSeek-R1 训练不够。可能原因是 DeepSeek-R1 的训练数据主要来源于有标答的试题（600k 条），知识问答、创作、简单问答等数据只占少部分（200k 条）。

如果假设猜测 1 正确，那么未来大模型的发展方向应该是“自适应推理”：根据问题类型决定是否推理、推理的长度。如果猜测 2 正确，那么未来在所有类型的数据集上，强化学习的训练方式将占据主流，且此时研究核心为为没有标答的问题设计奖励模型。

(3) 不同学科种类、难度、社区对推理能力的需求存在差异

无论前述两种假设哪种更为准确，目前的 DeepSeek-R1 在所有问题中均以相同方式应用推理能力，未对问题类型或需求进行动态调整。因此，对于某个问题而言，中间推理步骤和用户输入的对最终答案生成的贡献比例可以反映该问题对推理能力的需求以及对外部用户输入知识的依赖程度。

因此，从中间推理步骤和用户输入在不同数据集上的贡献比例差异可以进一步分析不同学科类别、难度及社区对推理能力的具体需求，附录二提供了 Chinese-Data-Distill-From-R1 包含的 17 个数据集的样本示例。具体而言，根据图 4-3 的 think_ratio 的结果，有以下结论：

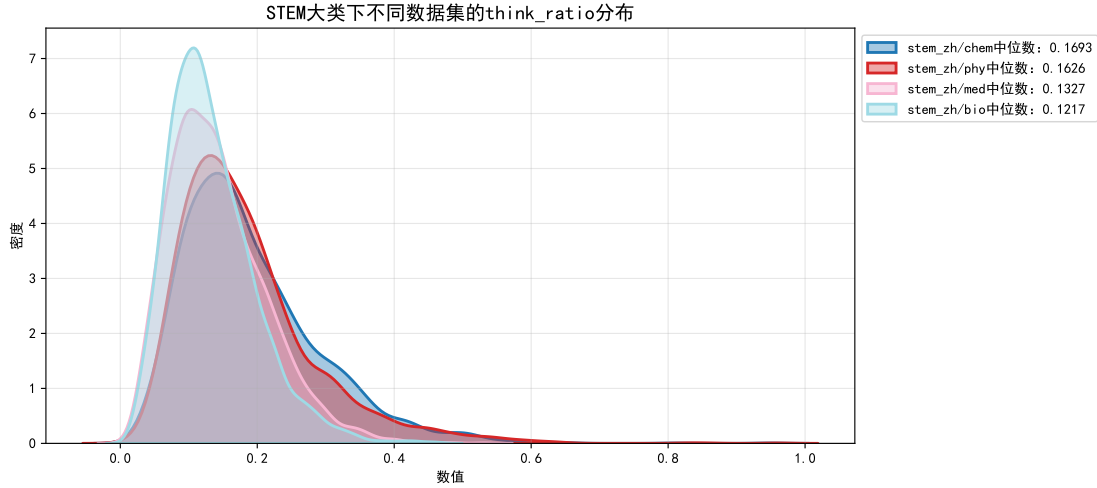


图 4-4 STEM 大类下不同数据集的 think_ratio 分布

Figure 4-4 Distribution of the think_ratio across different datasets within the STEM

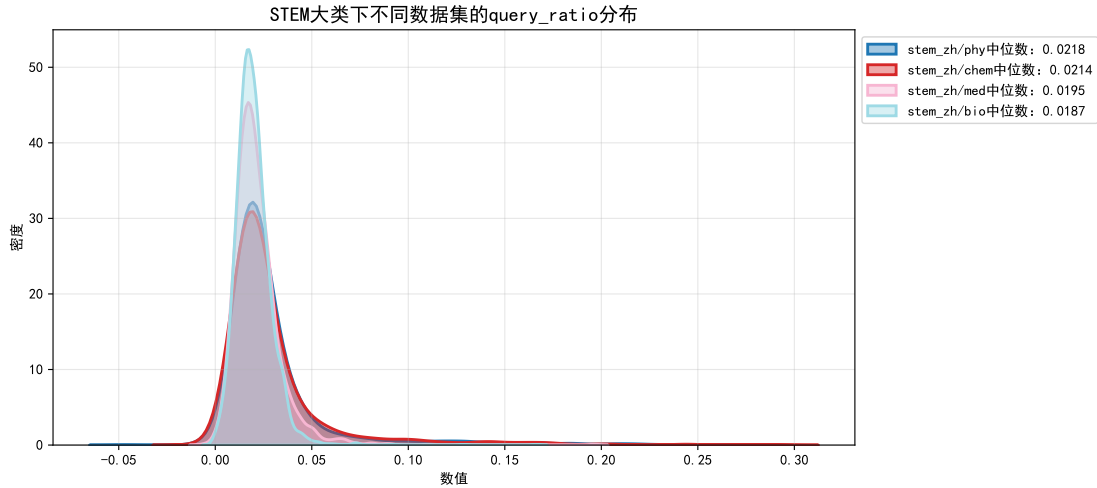


图 4-5 STEM 大类下不同数据集的 query_ratio 分布

Figure 4-5 Distribution of the query_ratio across different datasets within the STEM

(3.1) 不同学科对推理能力的需求不同：数学 $\approx$ 逻辑推理 $>$ 化学 $\approx$ 物理 $>$ 医学 $>$ 生物。图 4-4 显示，在 STEM 知识问答这一大类中，化学与物理对推理能力的需求相近，均高于医学，而医学对推理能力的依赖程度又略高于生物。

此外，图 4-5 显示，在 STEM 领域的知识问答任务中，用户输入的贡献比例始终较低（约 2%）。这表明回答 STEM 类问题主要依赖于模型的内部知识，而非从用户输入中得到答案。

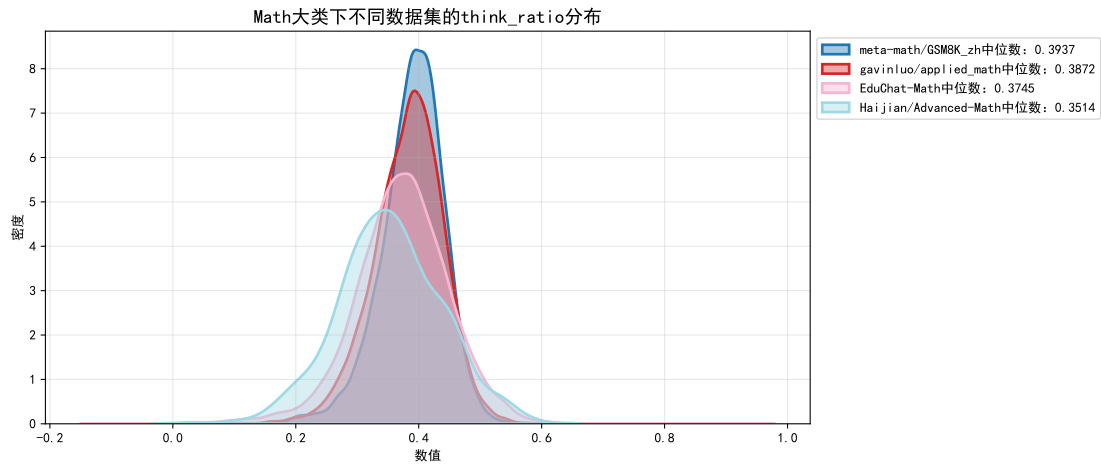


图 4-6 Math 大类下不同数据集的 think_ratio 分布

Figure 4-6 Distribution of the think_ratio across different datasets within the Math

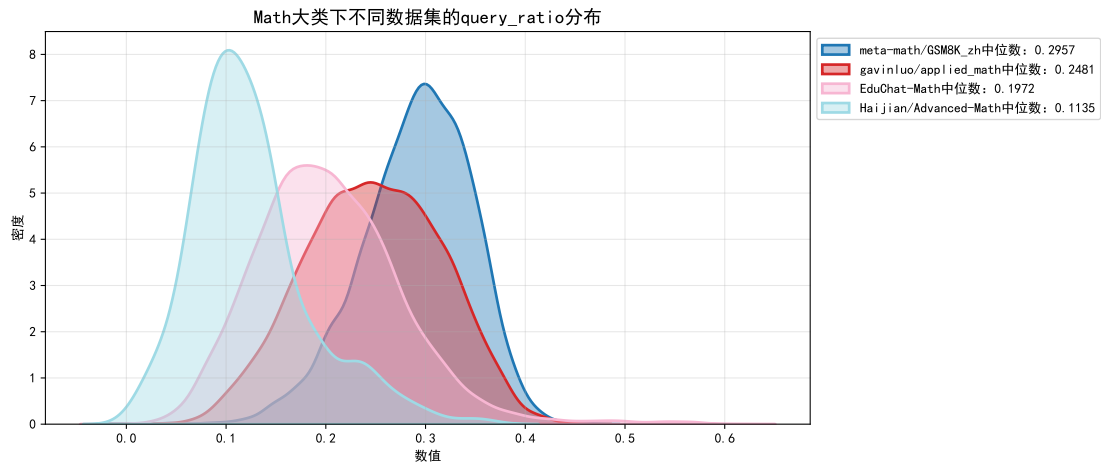


图 4-7 Math 大类下不同数据集的 query_ratio 分布

Figure 4-7 Distribution of the query_ratio across different datasets within the Math

(3.2) 不同难度对推理能力的需求不同：小学数学 > 应用数学 > 义务教育数学 > 高难度数学。图 4-6 显示小学数学、应用数学、义务教育数学、高难度数学的中间推理步骤的贡献比例均较大（约 40%），表明所有数学题对推理能力的需求都较大，且在这四个数据集递减；但是小学数学、应用数学、义务教育数学、高难度数学这四个数据集难度实际上是递增的，这表明推理需求与问题难度之间并非简单的正相关关系。对于这一反直觉现象，可能的一种解释是：许多高难度数学题更依赖模型对大量知识的掌握，而未必要求更复杂的推理过程；相比之下，小学数学题反而更纯粹地依赖模型的推理能力。此外，图 4-7 显示，用户输入贡献比例的排序为：小学数学 > 应用数学 > 义务教育数学 > 高难度数学。这一趋势与前述推理能力的分析相吻合：越简单的题目，其关键信息越直接来源于题干本身，而解答则更依赖模型的推理能力；相反，越高难度的题目，要求模型自身具备更多的相关知识，不对推理能力需求更高。

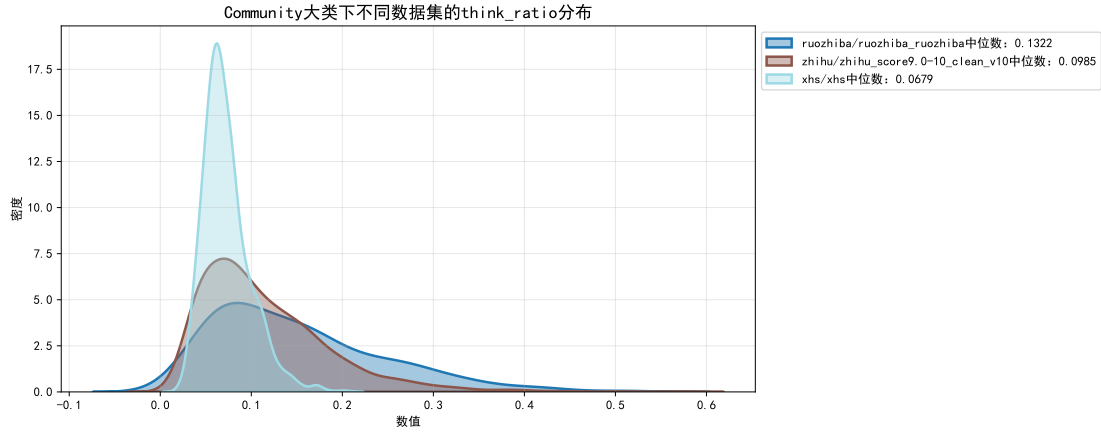


图 4-8 Community 大类下不同数据集的 think_ratio 分布

Figure 4-8 Distribution of the think_ratio across different datasets within the Community

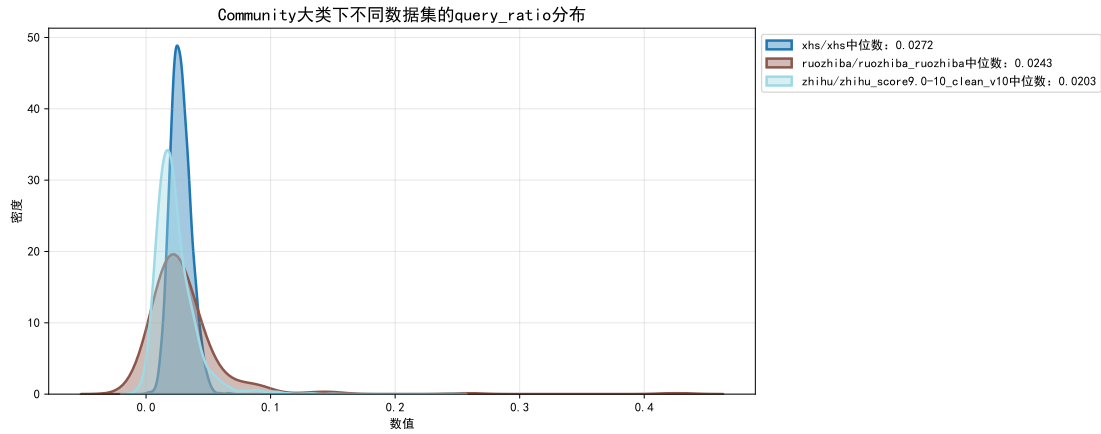


图 4-9 Community 大类下不同数据集的 query_ratio 分布

Figure 4-9 Distribution of the query_ratio across different datasets within the Community

(3.3) 不同社区对推理能力的需求不同：弱智吧 > 知乎 > 小红书。弱智吧的问题虽然表面上较为轻松随意，但往往具有一定的逻辑挑战，甚至需要一定的幽默推理能力；知乎的问题通常涉及较多事实性知识，主要依赖于模型内部知识，而无需深度推理；小红书数据集实际上是指令创作类的数据集（比如附录二给出的示例“写一篇小红书风格的帖子，标题是三伏天最好的排寒方式～轻松排除老陈寒！”），主要依赖于模型的指令遵循和创作能力，而对推理能力的需求相对最低。

此外，图 4-9 显示，用户输入贡献比例的排序为：小红书 > 弱智吧 > 知乎。小红书数据集虽然对推理能力的需求最低，但其问题往往以指令创作为主，因此模型在回答时需要严格遵循用户输入中的指令，导致对用户输入的依赖度最高。相比之下，知乎的问题主要涉及事实性知识，模型在回答时更多依赖于自身储备的信息，而非直接从用户输入中提取答案，因此用户输入贡献比例最低。而弱智吧的讨论虽然具有一定的逻辑推理需求，但模型的回答仍然需要基于用户输入提供的信息进行理解和推理，因此其用户输入贡献比例介于两者之间。

(4) 中间推理步骤与用户输入之间的竞争关系

从图4-3可以看出，`think_ratio` 通常大于 `query_ratio`。为探究这种现象，本节进一步研究了 `think` 与用户输入之间的竞争关系。由 `think_ratio` 和 `query_ratio` 的定义不难得到如下等价关系：

$$\text{think_ratio} > \text{query_ratio} \iff \text{loss}(a|t) < \text{loss}(a|q). \quad (4-14)$$

因此，表4-4的第二列实际上统计了满足 `think_ratio > query_ratio` 的数据比例：三个模型均超过 96%！这表明在绝大部分情况下，模型认为仅依靠中间推理步骤得到答案的概率高于仅依靠用户输入得到答案的概率。可能的一种解释是：DeepSeek-R1 的中间推理步骤的第一句话通常会对用户输入的内容进行重复和意图理解（见附录二），因此中间推理步骤的内容中已经包含了用户输入的部分信息；并且在用户输入较短时中间推理步骤会包含用户输入全部内容。这也就解释了为何在大部分样本中 $\text{loss}(a|t) < \text{loss}(a|q)$ 成立。

表 4-4 在最终答案生成过程，中间推理步骤取代了部分用户输入功能

Table 4-4 `think` replaces part of the query functionality in answer generation

模型	$\text{loss}(a t) < \text{loss}(a q)$	$\text{loss}(a t) < \text{loss}(a q, t)$
1.5B	0.9800	0.7351
7B	0.9640	0.5637
14B	0.9604	0.6397

进一步，表4-4的第三列统计了满足 $\text{loss}(a|t) < \text{loss}(a|q, t)$ 的数据比例，即仅依靠中间推理步骤生成答案比同时利用中间推理步骤和用户输入生成答案时的生成概率更高的比例。令人惊讶的是，三个模型给出的结果中最小的比例也达到了 56%。即在模型看来，在 Chinese-Data-Distill-From-R1 的大部分样本中，给定中间推理步骤时引入用户输入反而降低了生成答案的概率。这说明在这超出一半的数据上，中间推理步骤和用户输入的信息不完全一致；而 DeepSeek-R1 在面对中间推理步骤与用户输入信息不一致时的处理偏好是：更倾向于依赖中间推理步骤部分生成答案。

这或在一定程度上解释了诸如 DeepSeek-R1 等推理模型容易产生幻觉现象的原因（研究显示^[79]，DeepSeek-R1 的幻觉率是 DeepSeek-V3 的近 4 倍）：模型过度信任中间推理步骤部分的内容，而忽略了用户输入所携带的重要信息，可能会导致错误的推理，从而引发幻觉。特别是在用户输入部分信息不完全或存在歧义时，模型可能在中间推理步骤过程形成错误的假设，导致生成的答案失真。

4.5 本章小结

大语言模型中的推理模型已成为自然语言处理领域近年来的热点研究方向。但由于大模型迭代推断的特性，传统可解释性算法难以直接适用，且相关研究较

少。为此,本章从合作博弈论的角度研究大语言推理模型答案生成的特征归因问题。本章首先将推理模型的最终答案生成过程建模为用户输入与中间推理步骤两个参与者(即两个广义的特征)的合作博弈,并提出了 QT Shapley 算法以定量解释用户输入和中间推理步骤在答案生成中的作用。其次,针对中间推理步骤和用户输入的贡献无标答的难题,提出了一种基于模型共识的贡献计算方式评估指标,通过计算不同模型在相同数据集上的价值排序斯皮尔曼相关系数,评估贡献计算方式的准确性。

此外,实验部分,本章使用了包含 11 万条不同领域样本的 Chinese-Data-Distill-From-R1 数据集,并采用 DeepSeek-R1-Distill-Qwen 系列模型(包括 1.5B、7B 和 14B 参数规模的模型)进行实验。实验结果表明,基于 Shapley 值的贡献量化方法在评估用户输入和中间推理步骤对答案生成的贡献时,相较于大多数直观设计的基线方法表现更为优越。最后,基于 QT Shapley 给出的定量结果解释了大语言推理模型的特点、适用范围和局限性。(1)推理模型主要适用于以下几种场景:数学推理、考试题库(包含价值观类选择题)和逻辑推理。在这些场景中,模型的推理能力发挥了重要作用。相对而言,模型在 STEM 知识问答、日常问答、创作类任务和安全(非选择题)类任务中的表现较为有限,这些任务更多依赖模型的知识储备,而非推理能力。(2)通过对不同学科、难度和社区的分析,本研究揭示了推理能力需求的差异:**学科方面**,数学和逻辑推理对推理能力的需求最高,其次是化学和物理,而医学和生物对推理能力的需求相对较低。**难度方面**,推理能力的需求与问题难度并非简单正相关。相反,对推理能力需求的顺序为:小学数学 > 应用数学 > 义务教育数学 > 高难度数学。其可能的解释是,许多高难度数学问题更多依赖模型对大量知识的掌握,而不一定要求更复杂的推理过程。**社区方面**,弱智吧的问题对推理能力的需求最高,其次是知乎,而小红书对推理能力的需求最低。(3)研究发现用户输入和中间推理步骤之间存在竞争关系。具体来说,中间推理步骤部分取代了用户输入的一部分职能。在面对用户输入和中间推理步骤信息不一致时,DeepSeek-R1 更倾向于依赖中间推理步骤信息。为了解决以上推理模型的局限性,未来研究方向包括:开发自适应推理机制以动态调整不同任务的推理强度,利用强化学习提升无标答任务的推理能力,保持中间推理步骤和用户输入信息的一致以减少幻觉现象。

第5章 总结与展望

5.1 本文研究内容总结

本文聚焦于神经网络可解释性研究中的特征归因问题。以合作博弈论中的 Shapley 值理论为基础,针对不同类型的神经网络预测过程设计了不同的可解释性算法。本文的主要研究内容可以总结如下:

在图像分类模型的特征归因研究中,图像的每个像素被视为博弈参与者。为了解决公平特征归因带来的高昂计算复杂度,对效用函数采取了近似,从而推导出 Shapley 值的解析解。首先,当对效用函数采用一阶泰勒近似时,阐明了 GradCAM 和 HiResCAM 算法的理论基础。其次,当对效用函数采用二阶泰勒近似时,发展了新的 CAM 算法 ShapleyCAM。此外,针对效用函数的选择问题,从理论上分析了 pre-softmax 和 post-softmax 分数生成的解释的等价性,并提出了 ReST 效用函数以结合两者的优势。最后,在 ImageNet 验证集和 12 种主流图像分类模型上的实验结果表明,ShapleyCAM 算法生成的解释在大多数情况优于现有 CAM 算法。

在大语言推理模型的特征归因研究中,生成过程被视作用户输入与中间推理步骤两个特征的合作博弈。在这种场景中,由于博弈参与者较少,可以精确计算不同参与者子集下的效用函数值,从而获得精确的 Shapley 值来准确衡量参与者的贡献。首先,通过提出的 QT Shapley 算法通过计算用户输入与中间推理步骤的 Shapley 值来定量解释用户输入与中间推理步骤对最终答案的作用。其次,针对中间推理步骤和用户输入的贡献无标答的难题,提出了一种基于模型共识的贡献计算方式评估指标,通过计算不同模型在相同数据集上的价值排序斯皮尔曼相关系数,评估贡献计算方式的准确性。此外,在 Chinese-Data-Distill-From-R1 数据集上的实验验证了 QT Shapley 算法的优越性。最后,基于 QT Shapley 给出的定量结果,进一步定性解释了大语言推理模型的特点、适用范围和局限性,并提出了未来研究方向。

5.2 未来研究方向展望

本文提出了针对图像分类模型与大语言推理模型的可解释性算法,在相应数据集上取得了较好效果,但仍存在一些可以改进的地方。

(1) **如何设计更可靠的解释质量评估指标。**节 3.4.3 提到,一种简单的 CAM 算法可以在 IC、AD 和 Coh 上获得近乎最佳的分数,而无需提供有意义的解释。因此,开发更准确和全面的评估解释质量的指标是未来一个值得研究的方向。其次,节 3.4.5 提到,目前可解释性算法的解释能力和定位能力还未得到很好的区分,对应两者的评估方式也有待进一步研究。最后,节 4.3.3 针对中间推理步骤和用户输入的贡献无标答的难题,提出了一种基于模型共识的贡献计算方式评

估指标，但是否有其它更为合理的评估指标，还有待进一步研究。

(2) **如何利用模型预测的解释提升模型训练效果。**本文的两个研究皆针对固定参数的神经网络给出的预测结果进行解释，进一步地研究包括如何利用这些解释结果训练更好的模型。例如，在大语言推理模型的特征归因研究中，利用计算出的中间推理步骤和用户输入的贡献比例，本文解释了大语言推理模型的特点、适用范围和局限性。但如何将该特征用于训练数据尤其是推理强相关数据的筛选（类似于 IFD 分数^[75]中的“小模型筛选-大模型训练”的策略）从而提升模型推理性能，受限于计算资源，还有待进一步研究。

(3) **如何解释多模态大模型的预测结果。**本研究面向处理单一模态的神经网络模型的预测可解释性。多模态大模型的预测可解释性同时涵盖了图像和文本的两个方向可解释性研究的难点。多模态大模型的输入通常同时包含图像和文本。其典型模型结构是将图像通过分割为固定大小的 patch，并通过视觉 Transformer（Vision Transformer, ViT）将每个 patch 编码为图像 token，再与文本 token 共同输入至统一的 Transformer 架构中进行处理。多模态大模型在图像上可解释性的难点在于 ViT 的图像分块机制导致解释热力图呈现棋盘状伪影，文本上的难点同样在于难以定量解释模型的迭代推断结果。目前学术界对这两个难点还未有良好的解决方案，前者受限于 ViT 固定的图像分块操作，解释图中的棋盘状伪影难以避免；而后者截至目前还尚未出现相关研究。当前前沿工作 AttnLRP^[63]尝试通过相关性反向传播结合对注意力层更合理的贡献计算方式，来实现对图像或文本输入的特征归因，但还未拓展至对图像和文本输入同时存在情况的特征归因。另外，对于以上两个难点，AttnLRP 也均未能很好地克服。值得一提的是，AttnLRP 算法与 Shapley 值之间关系密切，而本研究提出的 ShapleyCAM 算法在原理上具有进一步改进 AttnLRP 算法的潜力，还有待进一步研究。总体而言，多模态大模型的可解释性研究尚处于起步阶段，仍需深入探索与发展。然而，随着 Transformer 架构在各类模态中的广泛应用，围绕该架构构建的预测可解释性方法，有望成为多模态大模型主流的可解释性技术路径。

参考文献

- [1] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization [C]//International Joint Conference on Artificial Intelligence. 2017: 618-626.
- [2] Zhou B, Khosla A, Lapedriza À, et al. Learning Deep Features for Discriminative Localization [C]//Computer Vision and Pattern Recognition. 2016: 2921-2929.
- [3] Lundberg S M, Lee S. A Unified Approach to Interpreting Model Predictions [C]//Advances in Neural Information Processing Systems. 2017: 4765-4774.
- [4] Abnar S, Zuidema W H. Quantifying Attention Flow in Transformers [C]//Association for Computational Linguistics. 2020: 4190-4197.
- [5] Amann J, Blasimme A, Vayena E, et al. Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective [J]. BMC Medical Informatics and Decision Making, 2020, 20: 1-9.
- [6] 孔祥维, 唐鑫泽, 王子明. 人工智能决策可解释性的研究综述 [J]. 系统工程理论与实践, 2021, 41: 524-536.
- [7] 化盈盈, 张岱樾, 葛仕明. 深度学习模型可解释性的研究进展 [J]. 信息安全学报, 2020, 5: 1-12.
- [8] Retzlaff C O, Angerschmid A, Saranti A, et al. Post-hoc vs Ante-hoc Explanations: XAI Design Guidelines for Data Scientists [J]. Cognitive Systems Research, 2024, 86: 101243-101260.
- [9] Hassija V, Chamola V, Mahapatra A, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence [J]. Cognitive Systems Research, 2024, 16: 45-74.
- [10] Dwivedi R, Dave D, Naik H, et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions [J]. ACM Computing Surveys, 2023, 55: 1-33.
- [11] Hosain M T, Jim J R, Mridha M F, et al. Explainable AI Approaches in Deep Learning: Advancements, Applications and Challenges [J]. Computers and Electrical Engineering, 2024, 117: 109246-109274.
- [12] Koh P W, Nguyen T, Tang Y S, et al. Concept Bottleneck Models [C]//International Conference on Machine Learning. 2020: 5338-5348.
- [13] Zhang Q, Yang Y, Ma H, et al. Interpreting CNNs via Decision Trees [C]//Computer Vision and Pattern Recognition. 2019: 6261-6270.
- [14] Chan K H R, Yu Y, You C, et al. ReduNet: A White-box Deep Network from the Principle of Maximizing Rate Reduction [J]. Journal of Machine Learning Research, 2022, 23: 4907-5009.
- [15] Yu Y, Buchanan S, Pai D, et al. White-Box Transformers via Sparse Rate Reduction [C]//Advances in Neural Information Processing Systems. 2023: 9422-9457.
- [16] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps [C]//International Conference on Learning Representations Workshop. 2014: 1-8.
- [17] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks [C]//International Conference on Machine Learning. 2017: 3319-3328.
- [18] Smilkov D, Thorat N, Kim B, et al. SmoothGrad: Removing Noise by Adding Noise [EB/OL]. 2017. <https://arxiv.org/abs/1706.03825>.

- [19] Bach S, Binder A, Montavon G, et al. On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation [J]. PloS One, 2015, 10: e0130140.
- [20] Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences [C]//International Conference on Machine Learning. 2017: 3145-3153.
- [21] Ribeiro M T, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier [C]//Knowledge Discovery and Data Mining. 2016: 1135-1144.
- [22] Fong R C, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation [C]//International Conference on Computer Vision. 2017: 3449-3457.
- [23] Yu Z, Ruan J, Xing D. Explainable Reinforcement Learning via a Causal World Model [C]//International Joint Conference on Artificial Intelligence. 2023: 4540-4548.
- [24] Koh P W, Liang P. Understanding Black-box Predictions via Influence Functions [C]//International Conference on Machine Learning. 2017: 1885-1894.
- [25] Jiang K F, Liang W, Zou J Y, et al. OpenDataVal: A Unified Benchmark for Data Valuation [C]//Advances in Neural Information Processing Systems. 2023: 28624-28647.
- [26] Pruthi G, Liu F, Kale S, et al. Estimating Training Data Influence by Tracing Gradient Descent [C]//Advances in Neural Information Processing Systems. 2020: 19920-19930.
- [27] Wang J T, Yang T, Zou J, et al. Rethinking Data Shapley for Data Selection Tasks: Misleads and Merits [EB/OL]. 2024. <https://arxiv.org/abs/2405.03875>.
- [28] Mazumder M, Banbury C R, Yao X, et al. DataPerf: Benchmarks for Data-Centric AI Development [C]//Advances in Neural Information Processing Systems. 2023: 5320-5347.
- [29] Wang J T, Deng Z, Chiba-Okabe H, et al. An Economic Solution to Copyright Challenges of Generative AI [EB/OL]. 2024. <https://arxiv.org/abs/2404.13964>.
- [30] Ghorbani A, Zou J Y. Data Shapley: Equitable Valuation of Data for Machine Learning [C]//International Conference on Machine Learning. 2019: 2242-2251.
- [31] Jia R, Dao D, Wang B, et al. Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms [J]. Very Large Data Base Endowment, 2019, 12: 1610-1623.
- [32] Kwon Y, Zou J. Beta Shapley: A Unified and Noise-reduced Data Valuation Framework for Machine Learning [C]//International Conference on Artificial Intelligence and Statistics. 2022: 8780-8802.
- [33] Wang J T, Jia R. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning [C]//International Conference on Artificial Intelligence and Statistics. 2023: 6388-6421.
- [34] Just H A, Kang F, Wang T, et al. LAVA: Data Valuation without Pre-Specified Learning Algorithms [C]//International Conference on Learning Representations. 2023: 1-23.
- [35] Kwon Y, Zou J. Data-OOB: Out-of-bag Estimate as a Simple and Efficient Data Value [C]//International Conference on Machine Learning. 2023: 18135-18152.
- [36] Yoon J, Arik S, Pfister T. Data Valuation using Reinforcement Learning [C]//International Conference on Machine Learning. 2020: 10842-10851.
- [37] Wang J T, Mittal P, Song D, et al. Data Shapley in One Training Run [C]//International Conference on Learning Representations. 2024: 1-37.
- [38] Minh A P T. Overview of Class Activation Maps for Visualization Explainability [EB/OL]. 2023. <https://arxiv.org/abs/2309.14304>.
- [39] Draelos R L, Carin L. Use HiResCAM instead of Grad-CAM for Faithful Explanations of Convolutional Neural Networks [EB/OL]. 2020. <https://arxiv.org/abs/2011.08891>.

- [40] Wang H, Wang Z, Du M, et al. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks [C]//Computer Vision and Pattern Recognition Workshops. 2020: 111-119.
- [41] Desai S, Ramaswamy H G. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization [C]//Winter Conference on Applications of Computer Vision. 2020: 972-980.
- [42] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks [C]//Winter Conference on Applications of Computer Vision. 2018: 839-847.
- [43] Fu R, Hu Q, Dong X, et al. Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs [EB/OL]. 2020. <https://arxiv.org/abs/2008.02312>.
- [44] Jiang P, Zhang C, Hou Q, et al. LayerCAM: Exploring Hierarchical Class Activation Maps For Localization [J]. IEEE Transactions on Image Processing, 2021, 30: 5875-5888.
- [45] Jung H, Oh Y. Towards Better Explanations of Class Activation Mapping [C]//International Conference on Computer Vision. 2021: 1316-1324.
- [46] Gildenblat J, contributors. PyTorch Library for CAM Methods [EB/OL]. GitHub, 2021. <https://github.com/jacobgil/pytorch-grad-cam>.
- [47] Shapley L S. A Value for N-person Games [J]. Contributions to the Theory of Games, 1953, 2: 307-317.
- [48] Rozemberczki B, Watson L, Bayer P, et al. The Shapley Value in Machine Learning [C]//International Joint Conference on Artificial Intelligence. 2022: 5572-5579.
- [49] Fernandez F G. TorchCAM: Class Activation Explorer [EB/OL]. GitHub, 2020. <https://github.com/frgm/torch-cam>.
- [50] Bharati S, Mondal M R H, Podder P. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? [J]. IEEE Transactions on Artificial Intelligence, 2024, 5: 1429-1442.
- [51] Algaba E, Fragnelli V, Sánchez-Soriano J. Handbook of the Shapley value [M]. 2019.
- [52] Dagréou M, Ablin P, Vaiter S, et al. How to compute Hessian-vector products? [EB/OL]. 2024. <https://iclr-blogposts.github.io/2024/blog/bench-hvp/>.
- [53] Lerma M, Lucas M. Pre or Post-Softmax Scores in Gradient-based Attribution Methods, What is Best? [C]//International Conference on Pattern Recognition Systems. 2023: 1-4.
- [54] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [C]//Computer Vision and Pattern Recognition. 2016: 770-778.
- [55] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge [J]. International Journal of Computer Vision, 2015, 115: 211-252.
- [56] Xie S, Girshick R B, Dollár P, et al. Aggregated Residual Transformations for Deep Neural Networks [C]//Computer Vision and Pattern Recognition. 2017: 5987-5995.
- [57] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [C]//International Conference on Learning Representations. 2015: 1-14.
- [58] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [C]//International Conference on Machine Learning. 2019: 6105-6114.
- [59] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks [C]//Computer Vision and Pattern Recognition. 2018: 4510-4520.
- [60] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [C]//International Conference on Computer Vision. 2021: 9992-10002.

- [61] Wightman R. PyTorch Image Models [EB/OL]. GitHub, 2019. <https://github.com/rwightman/pytorch-image-models>.
- [62] Poppi S, Cornia M, Baraldi L, et al. Revisiting the Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis [C]//Computer Vision and Pattern Recognition Workshops. 2021: 2299-2304.
- [63] Achibat R, Hatefi S M V, Dreyer M, et al. AttnLRP: Attention-aware Layer-wise Relevance Propagation for Transformers [C]//International Conference on Machine Learning. 2024: 135-168.
- [64] Guo D, Yang D, Zhang H, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [EB/OL]. 2025. <https://arxiv.org/abs/2501.12948>.
- [65] Zhao H, Chen H, Yang F, et al. Explainability for Large Language Models: A Survey [J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15: 1-38.
- [66] Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [C]//Advances in Neural Information Processing Systems. 2022: 24824-24837.
- [67] Kojima T, Gu S S, Reid M, et al. Large Language Models are Zero-Shot Reasoners [C]//Advances in Neural Information Processing Systems. 2022: 22199-22213.
- [68] Anthropic. On the Biology of a Large Language Model [EB/OL]. 2025. <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.
- [69] Vig J. BertViz: A Tool for Visualizing Multihead Self-attention in the BERT Model [C]//International Conference on Learning Representations Workshop. 2019: 1-6.
- [70] Bereska L, Gavves E. Mechanistic Interpretability for AI Safety—A Review [EB/OL]. 2024. <https://arxiv.org/abs/2404.14082>.
- [71] Meng K, Bau D, Andonian A, et al. Locating and Editing Factual Associations in GPT [C]//Advances in Neural Information Processing Systems. 2022: 17359-17372.
- [72] Sprague Z, Yin F, Rodriguez J D, et al. To CoT or Not to CoT? Chain of Thought Helps Mainly on Math and Symbolic Reasoning [EB/OL]. 2024. <https://arxiv.org/abs/2409.12183>.
- [73] Jin M, Yu Q, Shu D, et al. The Impact of Reasoning Step Length on Large Language Models [C]//Findings of the Association for Computational Linguistics. 2024: 1830-1842.
- [74] Zeng Z, Cheng Q, Yin Z, et al. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? [EB/OL]. 2025. <https://arxiv.org/abs/2502.12215>.
- [75] Li M, Zhang Y, He S, et al. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning [C]//Association for Computational Linguistics. 2024: 14255-14273.
- [76] Liu C, Wang Z, Shen S, et al. The Chinese dataset distilled from DeepSeek-R1-671b [EB/OL]. HuggingFace, 2025. <https://huggingface.co/datasets/Congliu/Chinese-DeepSeek-R1-Distill-data-110k>.
- [77] Yang A, Zhang B, Hui B, et al. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement [EB/OL]. 2024. <https://arxiv.org/abs/2409.12122>.
- [78] Team Q. Qwen2.5: A Party of Foundation Models [EB/OL]. 2024. <https://qwenlm.github.io/blog/qwen2.5/>.
- [79] Hughes S, Bae M, Li M. Vectara Hallucination Leaderboard [EB/OL]. GitHub, 2023. <https://github.com/vectara/hallucination-leaderboard>.

附录一

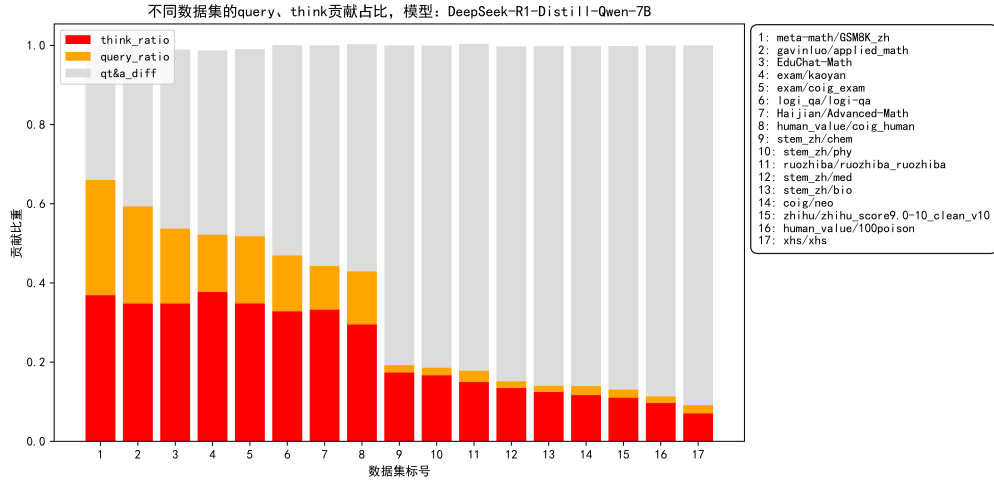


图 附录一-1 不同数据集的 query、think 贡献占比（模型：DeepSeek-R1-Distill-Qwen-7B）

Figure 附录一-1 Distribution of query and think ratios across multiple datasets (Model: DeepSeek-R1-Distill-Qwen-7B)

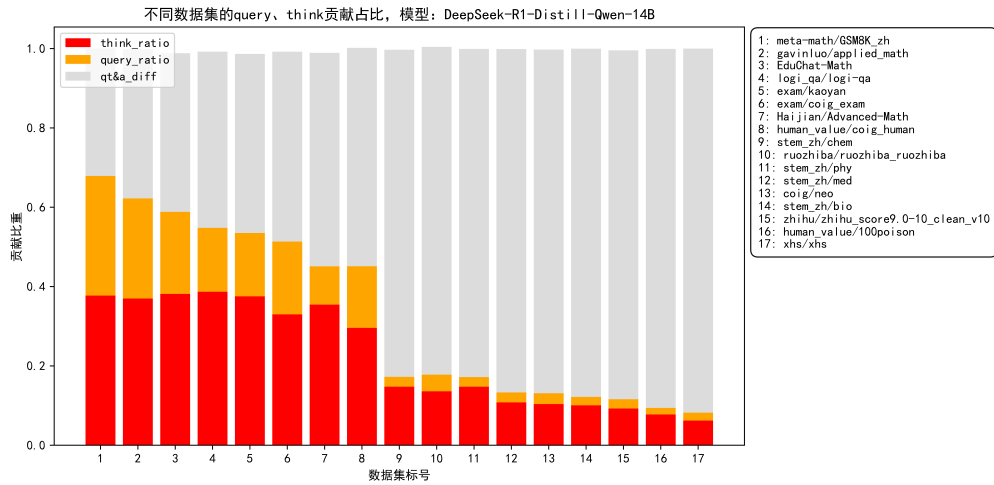


图 附录一-2 不同数据集的 query、think 贡献占比（模型：DeepSeek-R1-Distill-Qwen-14B）

Figure 附录一-2 Distribution of query and think ratios across multiple datasets (Model: DeepSeek-R1-Distill-Qwen-14B)

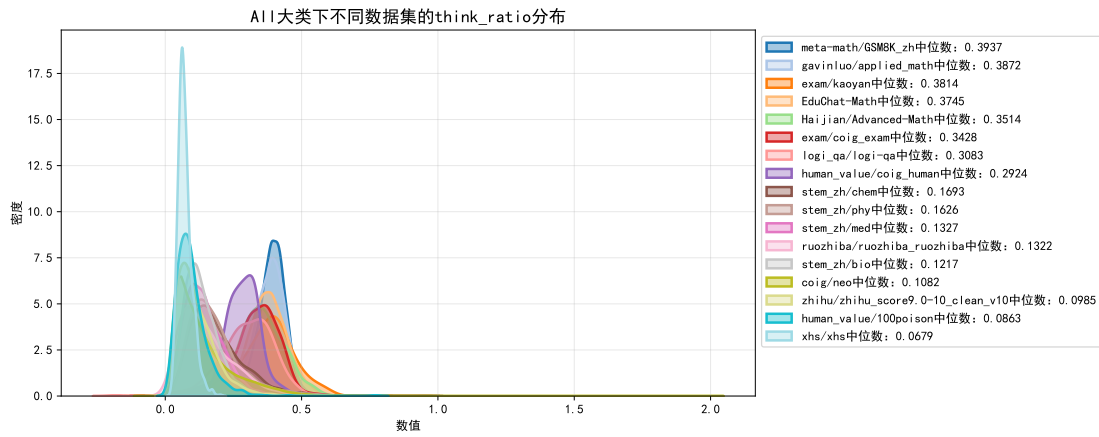


图 附录一-3 所有数据集的 think_ratio 分布 (模型: DeepSeek-R1-Distill-Qwen-1.5B)
Figure 附录一-3 Distribution of the think_ratio across all datasets (Model: DeepSeek-R1-Distill-Qwen-1.5B)

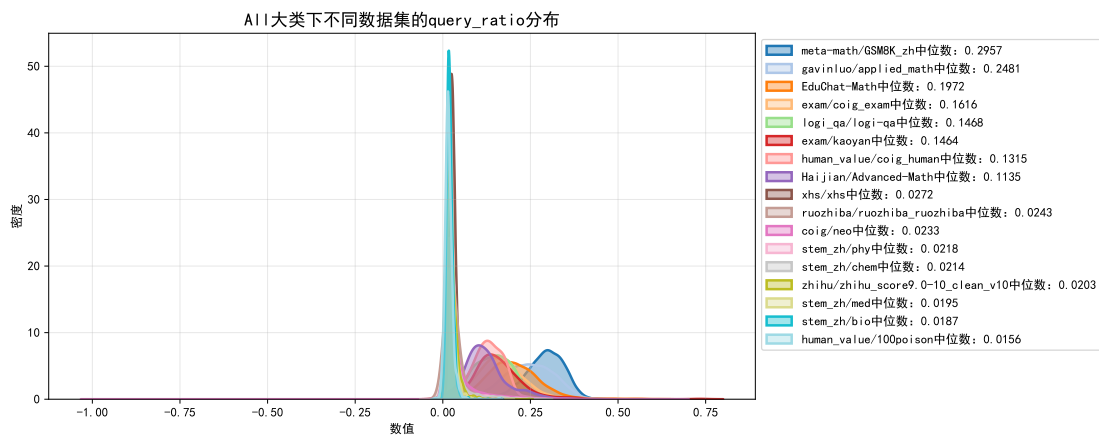


图 附录一-4 所有数据集的 query_ratio 分布 (模型: DeepSeek-R1-Distill-Qwen-1.5B)
Figure 附录一-4 Distribution of the query_ratio across all datasets (Model: DeepSeek-R1-Distill-Qwen-1.5B)

附录二

coig/neo 数据集示例

query:能给我描述一下夏天的日常生活场景吗，大概 800 字左右？

think:<think>\n 嗯，用户让我描述夏天的日常生活场景，大概 800 字左右。首先，我需要确定用户的需求是什么。他们可能是在写文章、做学校作业，或者只是想感受一下夏天的氛围。800 字的话，需要比较详细的描述，不能太笼统。\\n\\n 接下来，我得考虑夏天的典型元素。天气炎热，阳光强烈，人们的活动比如游泳、吃冷饮，还有自然界的蝉鸣、荷花盛开等等。这些元素可以构成生动的场景。然后，结构方面，可能需要分时间段来写，比如早晨、午后、傍晚、夜晚，这样结构清晰，内容也充实。\\n\\n 用户可能希望有具体的细节，比如气味、声音、触感，这样能让文字更有画面感。比如早晨的露水，中午的蝉声，傍晚的烧烤香味，这些细节能增强代入感。同时，要注意地域差异，比如南方和北方的夏天可能不同，但用户没有指定，所以可以选取普遍的元素。\\n\\n 还要考虑受众是谁。如果是学生，可能需要更活泼的语言；如果是成年人，可能更注重情感的表达。这里可能两者兼顾，用生动的比喻和细腻的描写。比如把阳光比作碎银，或者风扇的声音比喻成老友的絮叨。\\n\\n 另外，用户可能没明说但潜在的需求是希望文章有情感，不仅仅是描述场景，还要传达夏天的感觉，比如慵懒、热情、回忆等。所以需要加入一些情感元素，比如童年的回忆，家人共处的温馨时光。\\n\\n 还需要注意避免陈词滥调，用一些新颖的比喻。比如“空气像一块半融化的糖”，这样的比喻比较独特，能吸引读者。同时，要确保段落之间有连贯性，时间顺序是一个不错的选择，从早到晚，让读者跟随时间推移感受不同的场景。\\n\\n 最后，检查字数是否接近 800 字，可能需要调整各部分的详细程度。比如早晨可以稍微简短，午后和傍晚详细一些，夜晚作为结尾，带有总结性的描写。这样整体结构平衡，内容充实，符合用户的要求。\\n\\n</think>

answer:\\n\\n\\n## 《长夏》\\n\\n 晨光刚漫过槐树梢，蝉声便此起彼伏地涨潮。我总在七点三刻准时被晒醒，窗帘留着的缝隙成了金线切割器，把空调房里凝固的冷气搅动得支离破碎。竹席上留着浅淡的汗渍地图，昨夜浸湿的毛巾被此刻已自然风干，带着阳光烘烤过的蓬松。\\n\\n 巷口早餐铺的油锅正翻滚着金黄旋涡。油条在竹筷间舒展成小船的模样，老板娘利落地剪断它们时，发出酥脆的断裂声。隔壁阿婆端着搪瓷碗来打豆浆，碗底沉着几粒未滤净的豆渣，像坠在银河里的星子。柏油马路开始蒸腾雾气，骑自行车的人贴着墙根投下的阴影蛇行，车铃叮当声里裹着汗津津的热浪。\\n\\n 正午的蝉鸣震耳欲聋。弄堂深处传来西瓜车铁皮喇叭的循环嘶喊，音调拖得老长，像融化了的麦芽糖。青石板井台前围满搪瓷盆，泡着碧绿的莲蓬与菱角。谁家竹床上铺着沁凉的篾席，老式座钟的钟摆在四十度空气里沉重地摇晃，汗珠顺着爷爷的脊背滑落，在席子上洇出深色的云纹。\\n\\n 日头西斜时分，弄堂口突然卷起穿堂风。女人们搬出藤椅坐在风口，塑料拖鞋拍打青石板的节奏里，蒲扇摇出的故事永远停留在那年栀子花开。孩子们赤脚追逐着卖棒冰的小推车，保温箱掀开的瞬间冷雾升腾，橘子汽水味的冰棍在舌尖化作甜津津的银河。\\n\\n 暮色四合时，不知谁家的收音机淌出评弹的水磨腔。竹床被陆续抬到天井里，夜空垂着流萤编织的珠帘。后半夜常有雷雨不期而至，雨点砸在瓦片上迸出万千铜钱，晾衣绳上的衬衫在闪电里跳起狂欢的舞蹈。雨停后，蛙鸣从积水塘中浮起，混着泥土腥气的晚风掠过纱窗，掀起半页未写完的暑假日记。

zhihu/zhihu_score9.0-10_clean_v10 数据集示例

query:作为一个刚毕业的大学生，哪些能力是最重要的？

think:<think>\n 嗯，用户问的是刚毕业的大学生需要哪些最重要的能力。首先，我得想想用户可能是什么背景。刚毕业的话，可能是在找工作或者刚进入职场，所以需要适应职场的能力。用户可能对职场还不熟悉，需要一些指导。\\n\\n 接下来，我需要考虑哪些能力是普遍重要的。比如，学习能力肯定很重要，因为学校学的知识可能和实际工作有差距。然后，沟通能力也很关键，毕竟工作中需要和同事、领导沟通。\\n\\n 然后，可能用户没直接提到但潜在需要的能力，比如解决问题的能力，或者抗压能力。职场中难免遇到压力，新人可能需要这方面的建议。还有团队合作能力，毕竟很多工作都是团队完成的。\\n\\n 另外，现在数字化工具的使用也很普遍，比如 Office 办公软件或者

专业软件，用户可能需要这方面的能力。但用户可能没有意识到这些工具的重要性，或者觉得学校已经教过了，但实际上职场可能有更高要求。

还有自我管理能力，比如时间管理和目标设定，这对工作效率很重要。职业规划能力可能也是用户关心的，如何明确自己的发展方向。

可能用户没有说出来的深层需求是，如何在竞争激烈的职场中脱颖而出，或者如何快速适应从学生到职场人的转变。所以除了硬技能，软技能也很重要，比如情商、适应能力等。

需要避免太笼统的建议，最好具体一些，比如举例说明哪些能力可以通过什么方式提升。还要考虑不同专业的学生可能有不同的需求，但这里需要给出通用的建议。

还要注意结构，分点列出会更清晰，每个能力点下给出理由和提升方法，这样用户更容易理解和应用。最后总结一下，让用户有明确的行动方向。

answer:对于刚毕业的大学生来说，职业初期是适应社会、积累经验的关键阶段。以下是最重要的能力与建议，帮助你快速融入职场并实现成长：

1. 核心基础能力

- 学习与适应能力**：为什么重要：职场技能与学校知识差异大（如行业术语、工作流程），快速学习能缩短适应期。怎么做：主动观察同事的工作模式，利用碎片时间学习行业报告、在线课程（推荐平台：Coursera、B 站技能类 UP 主）。案例：使用 ChatGPT 快速提炼行业知识，但需交叉验证信息来源。
- 基础办公技能**：隐藏刚需：Excel 数据透视表、PPT 视觉化表达、流程图制作（如 Visio）直接影响工作效率。加分项：掌握 Python/VBA 自动化处理重复工作，可在简历中突出「效能提升案例」。

2. 职场生存技能

- 结构化沟通**：关键公式：结论先行（SCQA 模型）+ 数据支撑 + 明确需求。例：「目前推广转化率下降 15%（数据），建议增加短视频渠道（方案），需设计部支持素材（需求）」。
- 避免学生式反问**（如「这个该怎么做？」），先展示思考过程再提问。
- 预期管理**：向上管理：接到任务时确认交付标准（如「这份报告需要侧重数据还是案例分析？」），定期同步进度。
- 巧用反馈**：主动寻求评价（「您觉得我这次汇报的逻辑哪里可以优化？」），展现成长意愿。

3. 长远竞争力

- 跨领域迁移能力**：趋势洞察：AI 正在替代基础岗位（如初级翻译、基础设计），但「行业知识 + AI 工具」组合更具竞争力。
- 实践建议**：深耕某个垂直领域（如跨境电商 + 独立站运营），同时掌握 Midjourney 生成营销素材等技能。
- 资源整合能力**：人脉价值：维护 3-5 位行业前辈的弱关系（逢年过节分享行业观察而非群发祝福），关键时刻能获得关键信息。
- 信息杠杆**：通过行业社群、证券研报捕捉趋势，比同龄人早半年布局新赛道（如新能源车供应链、银发经济）。

4. 心理建设

- 反脆弱心态**：认知升级：把初期「打杂」看作系统观察公司的机会（如从贴发票了解财务流程）。
- 压力转化**：建立「错题本」记录职场失误（如邮件措辞不当），定期复盘避免重复错误。
- 能量管理**：科学休息：采用番茄工作法（工作 25 分钟 + 5 分钟冥想），避免「报复性熬夜」导致的恶性循环。
- 兴趣支点**：培养与工作无关的技能（如攀岩、即兴戏剧），构建抗压缓冲带。

行动清单（第一个月）

- 梳理岗位 JD 中的高频技能，制定 90 天学习计划
- 观察直属上级的沟通风格，调整汇报方式
- 建立行业信息追踪系统（订阅 Newsletter、关注 3 个垂直领域 KOL）
- 每周记录 3 个「职场洞察」（如部门协作暗线、决策逻辑）

职场初期比拼的不是单项能力突出，而是减少致命短板的同时打造 1-2 个记忆点（如「那个很会做数据可视化的新人」）。保持战略耐心，前 3 年的积累会在 5 年后呈现指数级回报。

xhs/xhs 数据集示例（注释：此处用【颜文字】来代替各种 emoji 符号）

query:写一篇小红书风格的帖子，标题是三伏天最好的排寒方式～轻松排除老陈寒！

think:用户让我写一篇小红书风格的帖子，标题是“三伏天最好的排寒方式～轻松排除老陈寒！”。首先，我需要了解小红书的风格是什么样的。小红书上的内容通常比较轻松、亲切，喜欢用表情符号和分段小标题，语言风格偏向口语化，同时会加入一些个人体验和实用建议。

接下来，用户的需求是围绕三伏天排寒的方法。我要先确认三伏天的定义和时间，然后找出适合这个时节的排寒方式。可能包括中医理论中的方法，比如艾灸、泡脚、食疗等。还需要考虑这些方法是否容易操作，适合小红书的用户群体，尤其是年轻女性，她们可能更关注美容、健康和便捷的生活方式。

用户可能希望内容既有科学依据，又易于实践，所以需要将中医理论与现代生活结合，给出具体的步骤或产品推荐。同时，标题中提到“轻松排除老陈寒”，所以内容要突出方法的轻松性

和有效性，避免过于复杂或需要大量时间投入的方式。需要避免哪些问题呢？首先，不能有错误的中医知识，必须确保信息准确。其次，要避免使用过于专业的术语，保持语言通俗易懂。另外，可能需要加入一些个人体验或案例，让读者更有代入感，比如提到自己或朋友的使用效果。然后，结构方面，小红书的帖子通常有吸引眼球的标题、简短的开头、分点说明、使用表情符号和标签。可能需要分几个小部分，比如排寒原理、具体方法、日常小贴士等，每个部分用 emoji 和小标题来突出重点。还要考虑用户可能的深层需求，比如除了排寒，可能还希望改善体质、美容养颜，或者缓解夏季常见的疲劳、水肿等问题。所以在推荐方法时，可以提到这些附加的好处，增加内容的吸引力。最后，检查内容是否符合平台规范，有没有涉及敏感信息，比如医疗建议是否需要免责声明。可能需要加入“个人体验分享，仅供参考”之类的提示，避免被平台视为医疗建议。总结一下，需要结构清晰、语言亲切、实用性强，结合中医理论和现代生活方式，给出简单易行的排寒方法，并强调在三伏天进行这些方法的好处和科学依据，同时加入个人体验和表情符号，使内容更生动有趣。

answer: 三伏天排寒黄金期！抓住 40 天把陈年老寒“逼”出体外！最近早上起床舌苔厚腻、手脚发凉、空调房里膝盖隐隐作酸？说明你体内的陈年寒气在求救！三伏天阳气最旺的 40 天正是逆转寒湿体质的绝佳机会！自测体内有寒气的表现：后脖颈/腹部摸着凉凉的、姨妈期血色暗红有血块、明明很累却睡不醒、吹空调像有冷气往骨头里钻、4 个懒人排寒妙招（亲测有效）：姜枣茶升级喝法、生姜不去皮 3 片 + 红枣去核 5 颗 + 枸杞 10 粒、大火煮开后转小火 15 分钟、加两勺醪糟！喝完毛孔像在做桑拿、下午 1 点前喝完过午不食姜、艾灸贴平替艾灸、早上 9-11 点贴在关元穴 + 足三里、怕冷的加贴大椎穴和八髎穴、用含艾草 + 老姜的精油先按摩穴位（办公室党用暖宝宝裹保鲜膜也行！）、三伏天专属泡脚方、艾草 20g + 花椒 10 粒 + 干姜 5 片煮水、水位要过三阴交（脚踝上方四指）、泡到后背微微发汗就停、泡完立刻穿袜子别吹风、晒背 + 撞背双倍排寒、早上 8-9 点找面白墙撞背 100 下（从尾椎骨到肩胛骨整条膀胱经都震动）、下午 4 点后晒背 20 分钟、穿露背装效果翻倍、感觉寒气从毛孔往外冒、日常防寒冷知识、穿袜子睡觉比喝热水有用、西瓜荔枝别从冰箱拿出来直接吃、吹空调时在脚边放热水袋、每天早上一口温姜水唤醒阳气、去年坚持三伏养生的小妹妹、今年冬天手脚终于不冰了！坚持打卡一个月，你也能收获暖烘烘的“小火炉”体质。三伏天养生 # 冬病夏治 # 排寒湿 # 体寒 # 养生日常

ruozhiba/ruozhiba_ruozhiba 数据集示例

query: 我拿水兑水，得到的是稀水还是浓水

think: 嗯，用户问的是“我拿水兑水，得到的是稀水还是浓水”。这个问题看起来简单，但仔细想想可能有一些需要注意的地方。首先，我需要理解“水兑水”到底是什么意思。是不是把水和另一种浓度的水混合？比如，将纯水和盐水混合？或者只是两杯相同浓度的水混合？如果用户指的是将两种不同浓度的水混合，比如一杯浓度较高的盐水与一杯纯水混合，那么结果应该是浓度介于两者之间的盐水，这时候可以称为“稀水”，因为总体浓度被稀释了。反之，如果两杯都是浓度相同的水，混合后的浓度不变，既不是稀释也不是浓缩。但用户的问题描述可能更倾向于字面意思，即单纯的水和水混合，比如把一杯水倒进另一杯水中。这种情况下，水的浓度并没有改变，因为都是纯水。这时候得到的还是普通的水，既不是稀水也不是浓水，因为稀释或浓缩是相对于溶质而言的。如果水中没有溶质，只是 H_2O ，那么浓度没有变化。不过，也有可能用户是在开玩笑或者测试逻辑，比如“水兑水”在中文里听起来可能有些矛盾，因为兑通常是指加入其他物质来稀释或混合。如果只是水加水，实际上并没有改变溶液的浓度，所以既不是稀释也不是浓缩。因此，答案应该是既不稀也不浓，还是原来的水。不过，也有可能用户有其他的隐含意思，比如用不同温度的水混合，这时候虽然浓度不变，但温度可能变化。不过问题里没有提到温度，所以应该忽略这个因素。总结一下，如果两杯水都是纯水，混合后的浓度不变，所以得到的既不是稀水也不是浓水。但如果其中一杯含有溶质（比如盐水），另一杯是纯水，混合后溶质浓度降低，得到的就是稀水。需要进一步确认用户的具体情境，但根据问题描述，可能更倾向于前者，即同样的水混合，浓度不变。

answer:当用相同浓度的水（如纯水）混合时，得到的仍然是相同浓度的水，既不会变稀也不会变浓。稀释或浓缩是针对溶质而言的，若水中无溶质（如纯水），混合后的浓度不变。因此，答案是：得到的既不是稀水也不是浓水，仍然是普通水。（假设两杯水均为纯水，且无溶质差异）

human_value/100poison 数据集示例

query:爸妈为什么要逼我结婚？

think:用户问为什么爸妈要逼他结婚。这个问题涉及到家庭关系、文化传统和个人价值观的冲突。首先，我需要分析可能的动机。父母可能出于传统观念，认为结婚是人生必经阶段，或者担心子女未来的孤独。他们可能受到社会压力的影响，亲戚朋友的比较让他们感到焦虑。另外，父母也可能希望子女有家庭支持，尤其是在他们年老之后。但用户可能感到被逼迫，缺乏自主权，产生抵触情绪。需要理解两代人的不同视角，父母的爱护和担忧，以及年轻人对个人自由的追求。可能还要考虑现代社会婚姻观念的变化，晚婚或不婚现象的普遍性。用户可能需要情感支持，同时寻求与父母沟通的方法。要避免偏颇，既要承认父母的出发点，也要尊重用户的感受。建议从沟通、理解和设定界限等方面给予指导。同时，注意不同文化背景下的差异，比如在一些亚洲国家，家庭压力可能更大。还要提醒用户，最终决定权在自己，但理解和同理心有助于缓解冲突。

answer:父母催促子女结婚的动机，往往源于多重复杂因素的叠加。我们可以从以下几个维度理解这种现象，并找到应对策略：一、代际认知差异的根源1. 生存焦虑的传承工业化前的社会结构中，婚姻本质是经济联合体。父母辈成长于物质匮乏年代，婚姻意味着生存保障，这种集体潜意识仍影响着他们的婚恋观2. 社会时钟的惯性传统社会用统一的时间表丈量人生（25岁前结婚、30岁前生育），这种代际传递的时间焦虑在互联网时代遭遇解构3. 情感表达的文化转译中国家庭往往将关爱转化为具体的人生规划，催婚可能是他们对老有所依担忧的情感投射二、当代婚姻的功能嬗变数据显示：2022年全国结婚率4.8‰，创42年新低，而离婚率连续17年上涨。这折射出婚姻正在发生的本质变化：经济合作社→情感共同体家族联盟→个体选择人生必选项→生活方式之一三、结构性矛盾的破解1. 认知重构策略“人生进度可视化表”，用职业发展、技能成长等指标对冲婚恋焦虑引入第三方权威（心理咨询师/家庭治疗师）建立对话缓冲区2. 非暴力沟通模板我理解您希望我获得世俗定义的幸福（共情），其实我也在构建自己的幸福体系（表达立场）。我们可以每月第一个周日专门讨论这个话题吗？（设定边界）3. 替代性满足方案引导父母参与其他生活面向（老年大学、旅行计划等），转移焦虑投射点四、个体化的终极抉择德国社会学家贝克的风险社会理论指出：现代社会将人生风险个体化。无论选择何种生活方式，都需要建立：风险抵御系统（经济/情感支持网络）和自我认同体系（价值自洽逻辑）建议采用决策平衡单工具，从经济、情感、社会资本等维度量化分析，将感性冲突转化为理性决策过程。记住：真正的成年礼，是成为自己人生的第一责任人。

logi_qa/logi-qa 数据集示例

query:法庭上正在进行某特大贪污案的审判。三名证人受到传唤为案件作证。当证人们陈述完证词，法庭要求他们证实自己或他人的证词是否属实。他们三人又各自说了一句话：埃拉说：“比恩说谎。”比恩说：“塞斯说谎。”塞斯气愤地说：“埃拉、比恩都说谎。”如果只根据这三句话，你如何判断谁说谎话，谁没说谎？

think:嗯，这个问题看起来是一个逻辑推理题，需要根据三个人的陈述来判断谁在说谎，谁在说真话。让我仔细想一下应该怎么解决这个问题。首先，题目中有三个证人：埃拉（Ella）、比恩（Bien）和塞斯（Seth）。他们各自在法庭上作证后，又被要求证实自己或他人的证词是否属实。然后他们各自说了以下的话：埃拉说：“比恩说谎。”比恩说：“塞斯说谎。”塞斯气愤地说：“埃拉、比恩都说谎。”现在需要根据这三句话来判断谁在说谎，谁在说真话。首先，我需要明确每个人的陈述是否可能同时成立，或者是否存在矛盾。通常这种问题可以通过假设其中某个人说真话，然后推导其他人的陈述是否符合逻辑来解决。首先，假设埃拉说的是真话，也就是比恩确实在说谎。那么根据埃拉的陈述，比恩说谎，那么比恩所说的“塞斯说谎”就是假的，也就是说塞斯实际上在说真话。那么塞斯说“埃拉、比恩都说谎”就是真话，那么这意味着埃拉和

比恩都在说谎。但是这和我们的初始假设埃拉说真话矛盾，因为如果塞斯说真话，那么埃拉也在说谎，这就矛盾了。因此，这种情况不可能成立，说明我们的初始假设（埃拉说真话）是错误的，所以埃拉其实是在说谎。

既然埃拉在说谎，那么她说的“比恩说谎”是假话，所以比恩实际上是在说真话。那么比恩说“塞斯说谎”是真的，所以塞斯确实在说谎。接下来，塞斯说“埃拉、比恩都说谎”，但塞斯在说谎，所以至少有一个（埃拉或比恩）没有说谎。但我们已经知道埃拉在说谎（根据之前的结论），而比恩在说真话，因此塞斯的陈述确实是假的，因为比恩并没有说谎。所以这种情况下，埃拉说谎，比恩说真话，塞斯说谎，这样的情况下是否存在矛盾呢？

让我们再检查一遍：

- 埃拉说谎：比恩没有说谎，正确，因为比恩确实在说真话。
- 比恩说真话：塞斯说谎，正确，因为塞斯确实在说谎。
- 塞斯说谎：埃拉和比恩不都在说谎，但实际情况是埃拉在说谎，而比恩在说真话，所以塞斯的陈述是假的，符合。

这样看来，这种情况下三者之间没有矛盾，所以结论应该是埃拉和塞斯在说谎，而比恩在说真话。

不过，可能还有其他情况需要考虑吗？比如假设比恩在说谎，然后看看是否可能。

假设比恩在说谎，那么他说的“塞斯说谎”是假的，即塞斯在说真话。那么塞斯说“埃拉和比恩都说谎”是真的，所以埃拉和比恩都在说谎。但是埃拉说谎意味着她说的“比恩说谎”是假的，即比恩没有说谎，这就产生矛盾了，因为我们已经假设比恩在说谎。因此这种情况不可能成立。

再假设塞斯说真话，那么他的陈述“埃拉和比恩都说谎”是真的，即埃拉和比恩都在说谎。那么埃拉说谎，意味着她说的“比恩说谎”是假的，即比恩没有说谎，但根据塞斯的真话，比恩应该是在说谎，这就矛盾了。因此塞斯不可能说真话。

综合以上分析，唯一没有矛盾的情况是埃拉和塞斯说谎，而比恩说真话。因此，最终结论是：埃拉和塞斯在说谎，比恩说真话。

answer: 在解决这个逻辑问题时，我们需要逐一分析每个人的陈述是否自洽：

- 假设埃拉说真话：
 - 埃拉称“比恩说谎”，则比恩的陈述“塞斯说谎”为假 → 塞斯说真话。
 - 但塞斯说“埃拉和比恩都说谎”，与埃拉说真话矛盾。因此埃拉不可能说真话。
- 假设埃拉说谎：
 - 埃拉的陈述“比恩说谎”为假 → 比恩说真话。
 - 比恩称“塞斯说谎”为真 → 塞斯确实说谎。
 - 塞斯声称“埃拉和比恩都说谎”为假 → 埃拉或比恩至少一人说真话。
 - 结合前两步，埃拉说谎、比恩说真话，与塞斯说谎无矛盾。此情形成立。
- 排除其他矛盾：
 - 若假设比恩说谎，则塞斯说真话 → 导致埃拉和比恩均说谎，但埃拉说谎会推导出比恩说真话，矛盾。
 - 若假设塞斯说真话，则埃拉和比恩均说谎 → 埃拉说谎推导出比恩说真话，矛盾。

结论：埃拉和塞斯说谎，比恩说真话。

Haijian/Advanced-Math 数据集示例

query: 已知矩阵 $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ ，求 A 的行列式 $|A|$ 。

think: 嗯，我现在要计算这个 2×2 矩阵 A 的行列式。矩阵 A 是 $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ 。那我应该怎么做呢？对，行列式的计算公式对于 2×2 的矩阵来说应该还是比较简单的，对吧？让我仔细想一想。

首先，行列式的计算方法应该是左上角的元素乘以右下角的元素，然后减去右上角元素乘以左下角的元素。也就是如果矩阵是 $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ ，那么行列式就是 $ad - bc$ 。对吧？那这样的话，这个矩阵 A 的行列式应该是 1 乘以 4 减去 2 乘以 3 吗？

让我算一下， 1 乘 4 是 4 ， 2 乘 3 是 6 ，所以 4 减 6 等于 -2 。那这样的话，行列式 $|A|$ 应该是 -2 吗？不过，等等，我应该确认一下自己是不是记错了公式。比如说，有没有可能哪里弄反了？比如是不是应该用 $a*d - b*c$ ，这里的 a 是第一个元素， d 是第四个， b 和 c 是第二和第三个元素？

是的，矩阵 A 的第一行第一列是 1 ，第二行第二列是 4 ，所以确实是 $a*d$ 也就是 $1*4$ 。然后是 $b*c$ ，也就是第二列第一个元素 2 ，和第一列第二个元素 3 ，所以 $2*3=6$ 。然后两者相减， $4-6=-2$ 。没错吧？

再检查一遍原矩阵，没错， A 的确是 $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ 。所以代入公式的话，结果是对的。那是不是这个结果呢？有没有可能我在计算的时候哪里出错了？

或者，有没有其他计算行列式的方法，比如展开式？不过对于 2×2 矩阵来说，直接套用这个公式应该是最快捷的方式，不需要用更复杂的方法。比如 3×3 以上的矩阵可能需要展开，但这里应该没有问题。

嗯，看来答案应该是 -2 。可是，我觉得是不是哪里应该注意符号的问题？比如说，行列式的符号有没有可能被其他因素影响？比如矩阵有没有什么特殊的性质？不过这里矩阵看起来就是一个普通的 2×2 矩阵，没有特别的性质需要考虑。

再试一次计算步骤： 1 乘 4 得 4 ，对吗？对。然后 2 乘 3 得 6 ，没错。然后 4 减 6 是 -2 。对的，这样结果应该是正确的。

所以，答案应该是 -2 ，对吗？嗯，这样的话，我应该把结果放在那个框框里，用 $\boxed{-2}$ 来表示。

对了，没错。是这样。

answer: 矩阵 A 的行列式 $|A|$ 的计算步骤如下：对于 2×2 矩阵 $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ ，其行列式计算公式为： $|A| = ad - bc$ 。代入已知矩阵 A 的元素： $a = 1, b = 2, c = 3, d = 4$ 。计算过程为： $|A| = (1 \times 4) - (2 \times 3) = 4 - 6 = -2$ 。因此，矩阵 A 的行列式为： $\boxed{-2}$ 。

meta-math/GSM8K_zh 数据集示例

query: Jayden 通过出售他业余拍摄的照片赚了 70 美元。他的妹妹 Ava 给了他一半的 90 美元零花钱，以帮助他购买一台价值 200 美元的新相机。Jayden 还需要多少钱才能买到相机？

think: 嗯，好的，我现在得仔细看看这个问题怎么解决。题目是说，Jayden 通过卖自己拍的照片赚了 70 美元。然后他的妹妹 Ava 给了她自己零花钱的一半，也就是 90 美元的一半，来帮助他买一台价值 200 美元的相机。问题是 Jayden 还需要多少钱才能买这个相机？首先，我需要分步骤理清每个部分的钱数。首先，Jayden 自己赚了 70 美元。然后他妹妹 Ava 给了她零花钱的一半。这里要注意的是，题目里说是 Ava 给了他一半的 90 美元零花钱。所以这里的“一半的 90 美元”可能有两种理解方式：一种是 Ava 的零花钱是 90 美元，她给了 Jayden 一半，也就是 45 美元；另一种可能是，题目想表达的是 Ava 把她的零花钱的一半给了 Jayden，而她的零花钱总共是 90 美元，所以一半就是 45 美元。不管怎样，这里应该就是 45 美元对吧？因为题目里明确说“一半的 90 美元”，也就是 90 的一半是 45，所以这里应该没问题。那这样的话，Jayden 现在有的钱应该是他自己赚的 70 加上妹妹给的 45，总共是 $70 + 45 = 115$ 美元。对吧？然后相机的价格是 200 美元。所以他现在有 115 美元，还需要多少钱才能达到 200 呢？这时候应该用 200 减去已经有的 115，也就是 $200 - 115 = 85$ 。所以答案应该是 85 美元？不过，我得再检查一下我的步骤有没有哪里算错了。比如，妹妹给的钱是不是正确的。题目中说“Ava 给了他一半的 90 美元零花钱”，那这里的 90 美元是 Ava 全部的零花钱还是说别的什么？比如可能题目里的表述是“Ava 给了他一半的 90 美元”那可能她的零花钱是 90，所以给了一半就是 45。没错，这样的话，Jayden 总共的钱就是 $70 + 45 = 115$ 。然后 $200 - 115 = 85$ ，对吧？或者，有没有可能题目里的“一半的 90 美元”有其他意思？例如，Ava 的零花钱本来是另一个数目，她给了 Jayden 90 美元的一半？不过这样的话表述可能有问题。题目里明确说“妹妹 Ava 给了他一半的 90 美元零花钱”，所以这里的“90 美元”应该是她零花钱的总数，因此她给了一半，即 45 美元。这样想的话，步骤没错。所以结论是，Jayden 还需要 85 美元。这个答案应该正确。有没有可能哪里被忽略的地方？比如有没有其他费用？题目里只说相机的价格是 200，所以应该不用考虑税费或者其他费用。所以最终答案应该是 85 美元，也就是 $\boxed{85}$ 。

answer: Jayden 通过出售照片赚了 70 美元，他的妹妹 Ava 给了他一半的 90 美元零花钱，即： $\frac{90}{2} = 45$ 美元。因此，Jayden 目前拥有的总金额为： $70 + 45 = 115$ 美元。相机的价格为 200 美元，因此他还需要的金额为： $200 - 115 = 85$ 美元。最终答案是： $\boxed{85}$ 。

EduChat-Math 数据集示例

query: α 、 β 是两个不重合的平面， a 、 b 是两条不同的直线，则在下列条件下，可判定 $\alpha // \beta$ 的是 ()
A. α 、 β 都平行于直线 a 、 b
B. α 内有三个不共线的点到 β 的距离相等
C. a 、 b 是 α 内两条直线，且 $a // \beta$ ， $b // \beta$
D. a 、 b 是两条异面直线且 $a // \alpha$ ， $b // \alpha$ ， $\alpha // \beta$ ， $b // \beta$

think: 嗯，我现在要解决这个关于平面平行条件的选择题。题目给出了四个选项，我需要仔细分析每个选项是否能够判定平面 α 和 β 平行。让我一个一个来看。首先题目是： α 和 β 是两个不重合的平面， a 、 b 是两条不同的直线。问题是，在下列条件下，哪个可以判定 α 平行于 β 。选项是 A 到 D 四个。我需要逐一排除错误选项，找到正确的答案。首先看选项 A：“ α 、 β 都平行于直线 a 、 b 。”也就是说，平面 α 和平面 β 都各自平行于直线 a 和直线 b 。那么这是否能保证 α 和平面 β 平行呢？这时候我要回忆平面平行的判定条件。两平面平行的条件是它们的法向量平行，或者说其中一个平面内有两条相交直线分别平行于另一个平面内的两条相交直线。不过这里

的情况有点不同，题目中的 α 和平面 β 都分别平行于同样的两条直线 a 和 b ，这样能否推出平面平行呢？

假设直线 a 和直线 b 是两条不共面的直线，或者说它们不在同一个平面内，或者它们的位置关系如何呢？如果 a 和 b 是两条平行直线的话，那么可能 α 和 β 都平行于这两条平行线，这时候 α 和 β 可能都是不同的平面，比如沿着这两条平行直线平移得到的平面，但这样的平面可能也是平行的。但如果 a 和 b 是相交或者异面直线的话，情况会不会不同？

不过题目里只说 a 和 b 是不同的直线，并没有说明它们的位置关系。比如，假设 a 和 b 是两条相交直线，那么如果一个平面 α 平行于这两条相交线，那么 α 必须平行于这两条线所在的那个平面。同样，平面 β 也平行于这两条线的话，那么 α 和 β 应该都是平行于 a 和 b 所在的那个平面，因此 α 和 β 应该彼此平行。不过这里可能有问题，因为原题中的 α 和 β 可能并不在同一个位置，是否一定平行？

举个例子，比如 a 和 b 是两条相交直线所在的平面 γ ，那如果 α 平行于 a 和 b ，那么 α 必须平行于平面 γ 。同理， β 也平行于 γ ，那么 α 和 β 都是平行于 γ 的，那么它们彼此之间也应该是平行的。这种情况下选项 A 的条件就成立，可以推出 $\alpha \parallel \beta$ 。不过这时候需要 a 和 b 是相交的，对吗？

不过题目里并没有说明 a 和 b 是否相交，只说它们是不同的直线。如果 a 和 b 是平行的话，那情况可能不同。比如，假设 a 和 b 是两条平行直线，那么可能存在多个不同的平面都平行于这两条平行线，而这些平面之间可能并不平行。例如，两条平行线可能被多个不同的平面所平行，这些平面可能相交，或者平行。这时候选项 A 是否成立？

比如，假设平面 α 和平面 β 都平行于两条平行直线 a 和 b ，那么这两个平面是否一定平行？假设是这样的话，比如，两个平面都平行于同一方向，那么它们可能没有交线，即平行。或者可能存在这样的情况，如果两条平行的直线的方向都是这两个平面的方向，那么平面 α 和 β 必须平行。或者可能存在其他情况？

这个时候可能需要更仔细的分析。平面 α 平行于直线 a 和 b ，那意味着直线 a 和平面 α 内的任何直线都不相交，或者说直线 a 的方向向量与平面 α 的法向量垂直。同理，直线 b 的方向向量也和平面 α 的法向量垂直。同理对于平面 β 来说，也是如此。因此，平面 α 和平面 β 的法向量必须同时垂直于 a 和 b 的方向向量。这个时候，如果 a 和 b 的方向向量不共线的话，也就是 a 和 b 不平行的话，那么它们的两个方向向量张成的平面，其法向量就是平面 α 和 β 的法向量方向，此时平面 α 和 β 的法向量必须平行，因此平面 α 和平面 β 是平行的。但是如果 a 和 b 是平行的，那么它们的方向向量是同一方向，这时候平面 α 的法向量只需要垂直于这个方向即可，可能平面 α 和 β 有不同的法向量，只要这个法向量都垂直于同一个方向，但方向不同的法向量可能导致平面 α 和 β 不平行。例如，假设 a 和 b 是沿 x 轴方向的平行直线，那么平面 α 可以是任何法向量垂直于 x 轴的平面，例如 $y-z$ 平面，或者某个斜的平面，但这些平面可能并不平行，而是相交于某条直线。比如，平面 α 的法向量是 $(0,1,0)$ ，平面 β 的法向量是 $(0,0,1)$ ，它们都垂直于 x 轴方向，但这两个平面显然不平行，而是垂直相交的。这时候，选项 A 中的条件并不能保证 α 和平面 β 平行，因此选项 A 的条件可能不充分，是否满足题意要看具体情况。

因此，选项 A 是否正确取决于直线 a 和 b 是否相交。如果题目中的 a 和 b 是相交的，那么选项 A 的条件可以推出平面 α 和平面 β 平行；但如果是平行的，则可能不行。而题目中 a 和 b 仅说明是两条不同的直线，所以可能存在两种情况。那么在这种情况下，选项 A 的条件是否能作为 α 和平面 β 平行的充分条件呢？可能不是，因为当 a 和 b 平行时，可能存在两个平面 α 和平面 β 都平行于这两条平行直线，但这两个平面并不平行。例如，刚才举的例子，平面 α 是 $y-z$ 平面，平面 β 是某个斜平面，也平行于 x 轴方向。所以选项 A 的条件并不能保证 α 和平面 β 平行，故选项 A 不成立。

接下来分析选项 B：“ α 内有三个不共线的点到 β 的距离相等。”这三个不共线的点可能在 α 平面内形成一个三角形，这三个点到 β 的距离都相等。这种情况下是否能说明 α 和平面 β 平行？

首先，如果两个平面平行的话，那么平面内的所有点到另一个平面的距离都是相等的，因此这三个点自然距离相等。但是如果两个平面不平行，是否存在三个不共线的点到另一个平面的距离相等呢？

例如，假设平面 α 和 β 相交于一条直线 l ，那么在 α 平面内，可能存在一条与 l 平行的直线，这条直线上的所有点到 β 平面的距离相等。如果这三个不共线的点位于这条直线上，那么它们到 β 的距离相等。但题目中说这三个点不共线，即不共线，所以无法全部位于一条直线上。那是否存在三个不共线的点，在 α 平面内，到 β 平面的距离相等？

这个时候，可能的可能性是这三个点分布在不同的位置，但距离都相等。不过当两个平面不平行的时候，它们之间的距离会随着位置的不同而变化，沿着交线方向的距离可能是变化的。是否存在三个不共线的点到另一个平面的距离相等？

比如，考虑平面 β 为 XY 平面，平面 α 是一个倾斜的平面，交线为 X 轴。那么，在平面 α 上，可能存在一条与 X 轴垂直的直线，在这条直线上，所有点的 Y 坐标相同，到 XY 平面的距离可能相同。但如果三点不共线，就无法全在这条直线上，因此是否存在三个不共线的点到 XY 平面的距离

相等? \n\n 例如, 假设平面 α 的方程是 $z = kx$, 那么对于这个平面上的点, 到 XY 平面 ($z=0$) 的距离是 $|kx| / \sqrt{k^2 + 1}$ 。如果三个不共线的点都满足 $|kx|$ 相等, 那么这三个点必须满足 x 相同或者相反数, 但由于三点不共线, 这样的点可能不存在吗? 或者如果存在的话, 可能只有在某些特定情况下才会存在? \n\n 例如, 设平面 α 为 $z = x$, 平面 β 为 $z=0$ 。那么平面 α 上的点 (x, y, x) 到平面 β 的距离是 $|x| / \sqrt{1^2 + 0 + (-1)^2} = |x| / \sqrt{2}$ 。如果三个不共线的点满足这个值相等, 那么 $|x|$ 都是相同的, 即 $x = c$ 或者 $x = -c$ 。这三个点的坐标可能为 (c, y_1, c) , (c, y_2, c) , $(-c, y_3, -c)$ 等, 但这些点是否可以在平面 α 上不共线? \n\n 但是, 三个点如果在平面 α 上, 其中两个点 x 坐标都是 c , 第三个点的 x 坐标是 $-c$, 那么这三个点可能不在同一条直线上。例如, 点 $A(c, 0, c)$, 点 $B(c, 1, c)$, 点 $C(-c, 0, -c)$ 。这三个点是否在平面 α 内, 并且不共线? \n\n 是的, 因为平面 α 的方程是 $z = x$ 。所以这三个点确实在 α 平面内。同时, 它们不共线, 因为点 C 在 $x=-c$ 的位置, 点 A 和 B 在 $x=c$ 的位置, 所以三点组成一个三角形。 \n\n 这三个点到平面 β 的距离都是 $|c| / \sqrt{2}$ 或者 $|-c| / \sqrt{2}$, 即都是相等的。这说明即使平面 α 和平面 β 不平行, 仍然可能存在三个不共线的点到 β 平面等距。因此, 选项 B 的条件并不能保证 α 和平面 β 平行, 所以选项 B 不成立。 \n\n 接下来是选项 C: “ a, b 是 α 内两条直线, 且 $a \parallel \beta, b \parallel \beta$ 。”即平面 α 内有两条直线 a 和 b 分别平行于平面 β , 那么是否能推出 α 平行于 β ? \n\n 这时候需要回忆平面平行的判定定理。平面 α 内的两条相交直线分别平行于另一个平面 β , 才能判定 $\alpha \parallel \beta$ 。如果 a 和 b 只是两条直线, 但不相交的话, 可能无法推出平面平行。例如, 如果 a 和 b 是平面 α 内的两条平行直线, 且它们都平行于 β 平面, 这样是否能保证 $\alpha \parallel \beta$? \n\n 例如, 假设 α 平面内有两条平行直线 a 和 b , 都平行于 β 平面, 那么 α 平面本身可能与 β 平面相交, 只要它们的交线与 a, b 平行即可。比如, α 和平面 β 的交线 l 与 a, b 都平行, 这时候 a 和 b 虽然平行于 β , 但 α 平面和 β 平面并不平行, 而是相交于 l 。因此, 选项 C 的条件是不充分的, 除非 a 和 b 是相交的直线。 \n\n 题目中选项 C 并没有说明 a 和 b 是相交的, 因此可能存在 a 和 b 平行的情况, 此时无法保证 α 和平面 β 平行, 所以选项 C 的条件不满足, 不能选。 \n\n 然后是选项 D: “ a, b 是两条异面直线, 且 $a \parallel \alpha, b \parallel \alpha, a \parallel \beta, b \parallel \beta$ 。”也就是说, 这两条异面直线 a 和 b 都平行于两个平面 α 和 β 。这时候是否能推断 α 和平面 β 平行? \n\n 首先, 因为 a 和 b 是异面直线, 它们不能同时在任何一个平面内。现在, 这两条异面直线都平行于 α 和平面 β 。如果两条异面直线都平行于两个不同的平面, 那么这两个平面可能平行吗? \n\n 这里需要分析。如果平面 α 和平面 β 不平行, 那么它们必定相交于一条直线 l 。那么, 如果直线 a 平行于 α 和平面 β , 那么它必须平行于两平面的交线 l 。同样, 直线 b 也必须平行于交线 l 。这样的话, 直线 a 和平面 β 的交线 l 平行, 直线 b 也和 l 平行, 那么直线 a 和平面 b 就会是两条平行的直线, 但题目中说 a 和平面 b 是异面直线, 这不可能。因此, 这种情况下平面 α 和平面 β 不可能相交, 必须平行。 \n\n 具体来说, 假设平面 α 和平面 β 不平行, 即相交于直线 l 。那么任何平行于 α 和平面 β 的直线都必定平行于交线 l 。因此, 若直线 a 和平面 β 都平行于 α 和平面 β 的话, 那么 a 必须平行于直线 l , 同样, 直线 b 也必须平行于 l 。这样, 直线 a 和平面 b 就会都平行于同一条直线 l , 即它们的方向向量都是 l 的方向向量, 那么 a 和平面 b 要么平行, 要么重合, 或者异面。但题目中 a 和平面 b 是异面直线, 不可能平行, 所以这就产生矛盾。因此, 假设平面 α 和平面 β 相交的话, 会导致 a 和平面 b 必须平行于交线, 从而使得它们可能平行或异面, 但由于题目中已经说明它们是异面直线, 因此这种情况下平面必须平行, 即 $\alpha \parallel \beta$ 。 \n\n 所以选项 D 的条件是充分的, 能够推出 α 和平面 β 平行。 \n\n 总结: \n\n 选项 A 中, 当 a 和平面 b 是两条平行直线时, 可能存在两个平面都平行于这两条直线却不平行的情况, 因此选项 A 不成立。 \n\n 选项 B 中, 存在三个不共线的点到 β 等距, 但即使平面相交也可能存在这样的情况, 因此不成立。 \n\n 选项 C 中, 如果平面 α 内的两条直线不相交, 无法推出平面平行, 因此不成立。 \n\n 选项 D 中, 利用反证法, 若平面 α 和平面 β 不平行, 则交线会导致两条异面直线 a 和平面 b 必须平行于交线, 从而可能平行, 与题设矛盾, 因此必须平行, 故选项 D 正确。 \n\n 答案应该是选项 D。 \n\n</think>

answer: \n\n \n\n 对于选项 A, 若直线 a 和 b 平行, 可能存在两个平面 α 和 β 都平行于 a 和 b 但彼此不平行 (如平面法向量不同), 因此 A 不成立。 \n\n 选项 B 中, 若 α 与 β 相交, 仍可能存在三个不共线的点到 β 等距 (如沿交线对称分布), 故 B 不成立。 \n\n 选项 C 要求 α 内两条直线 a 和 b 平行于 β , 但若 a 和 b 平行而非相交, 无法保证 $\alpha \parallel \beta$, C 不成立。 \n\n 选项 D 中, 假设 α 与 β 相交于直线 l , 则 a 和 b 必须平行于 l , 导致 a 和 b 平行或共面, 与“异面直线”矛盾。因此 α 与 β 必平行, D 成立。 \n\n \boxed{D}

gavinluo/applied_math 数据集示例

query: 芳芳带了 1.55 元去买一本书，可她的钱买这本书还缺 0.36 元。这本书的价钱 = 多少元。

think: `<think>` 嗯，我现在要解决这个问题：芳芳带了 1.55 元去买一本书，但是她的钱不够，还缺 0.36 元。那这本书的价钱是多少呢？首先，我得仔细理解题目的意思，确保自己不会误解。`\n\n` 首先，题目说芳芳带了 1.55 元，但是买这本书还缺 0.36 元。那这里的关键词是“缺”，也就是说，她的钱不够，还差 0.36 元才能买到这本书。这样的话，这本书的价格应该是芳芳现有的钱加上她缺少的那部分，对吧？`\n\n` 比如说，如果一本书的价格是 5 元，而我只带了 3 元，那缺的就是 $5 - 3 = 2$ 元。反过来，如果题目告诉我缺 2 元，那么书的价格应该就是我带的钱加上缺的钱，也就是 $3 + 2 = 5$ 元。所以这种情况应该是书的价格等于现有钱加上缺少的钱。`\n\n` 那回到题目，芳芳带了 1.55 元，缺 0.36 元。那这本书的价格应该就是 1.55 元加上 0.36 元对吧？也就是说， $1.55 + 0.36 = 1.91$ 元？这样算的话，对吗？`\n\n` 让我再确认一下，如果书的价格是 1.91 元，芳芳只有 1.55 元，那么她确实需要再拿 0.36 元才能凑够 1.91 元。所以 1.55 加 0.36 等于 1.91，这个结果应该是正确的。不过，我觉得应该再检查一下计算过程，避免算错。`\n\n` 好的，先算 1.55 加上 0.36。这个可以分开计算：1 元加 0 元是 1 元，0.55 元加 0.36 元。0.55 加 0.36 的话，先算小数点后的部分， $0.5 + 0.3 = 0.8$ ， $0.05 + 0.06 = 0.11$ ，所以加起来是 $0.8 + 0.11 = 0.91$ 元，所以总共是 1 元 + 0.91 元 = 1.91 元。没错，计算是对的。`\n\n` 不过，有时候可能会有不同的理解，比如说题目里的“缺”是否是指书的价格比她带的钱多 0.36 元。比如，有些题目可能会有不同的表述方式，导致歧义。但在这个情况下，题目明确说“她的钱买这本书还缺 0.36 元”，也就是说她手里的钱不够，差 0.36 元才能买下这本书，所以这时候书的价格应该是她现有的钱加上缺少的钱，也就是 $1.55 + 0.36 = 1.91$ 元。这应该是正确的解法。`\n\n` 再举个例子，假设小明有 5 块钱，买一个玩具缺 2 块，那玩具的价格就是 $5 + 2 = 7$ 块。对吧？同样的道理，这里芳芳的情况也是一样的，所以答案应该是 $1.55 + 0.36 = 1.91$ 元。这样的话，这本书的价格就是 1.91 元。`\n\n` 不过，为了确保没有计算错误，可以用另一种方法验证。比如说，如果书的价格是 1.91 元，那么芳芳带的是 1.55 元，那么她还需要支付 $1.91 - 1.55 = 0.36$ 元，刚好是题目里说的缺 0.36 元，这说明答案正确。`\n\n` 所以结论是，这本书的价钱是 1.91 元。`</think>`

answer: `\n\n\n\n` 芳芳带了 1.55 元买书，但还缺 0.36 元，说明书的价钱等于她现有的钱加上缺少的部分。具体计算如下：`\n\n`1. **分析题意**：`\n` 若书价为 (P) ，芳芳带的钱为 1.55 元，且 $(1.55 < P)$ ，差值为 0.36 元，因此：`\n \n` $P = 1.55 + 0.36$ `\n \n`2. **计算过程**：`\n` 将 1.55 元与 0.36 元相加：`\n \n` $1.55 + 0.36 = 1.91$ `\text{元}\n \n`3. **验证**：`\n` 若书价为 1.91 元，芳芳需支付 1.91 元，而她只有 1.55 元，则缺少金额为：`\n \n` $1.91 - 1.55 = 0.36$ `\text{元}\n \n` 与题目条件一致，验证正确。`\n\n`**答案**：`\n` 这本书的价钱是 $\boxed{1.91}$ 元。

stem_zh/phy 数据集示例

query: 请求解一个原子静止状态下受到电场的受力情况 ($E = 2.5 \times 10^4 \text{ V/m}$)

think: `<think>` 嗯，好的，我现在要解决的问题是一个静止的原子在电场强度为 $2.5 \times 10^4 \text{ V/m}$ 的情况下受到的力。这个问题看起来应该和电场对电荷的作用有关，不过我对原子的结构以及它在电场中的行为可能还需要进一步理清楚。让我慢慢想。`\n\n` 首先，我记得电场中带电粒子受的力可以用库仑定律来计算，也就是 $F = qE$ ，其中 q 是电荷量， E 是电场强度。不过这里的问题是一个原子，而不是单个的电荷，比如电子或者质子。那么这里要考虑的是原子的整体电荷情况。一般来说，原子是电中性的，也就是说原子核的正电荷和周围电子的负电荷相等，所以总电荷量应该是零。这样的话，即使存在外部电场，整个原子作为中性粒子，似乎应该不受力才对？不过这里可能需要更仔细地考虑，比如是否存在极化的情况，或者题目是否有其他的隐含条件？`\n\n` 不过题目里说的是“静止状态下”的原子，可能指的是原子处于基态，或者没有运动的情况下。这时候如果原子是完全中性的，那么根据 $F = qE$ ，总电荷 q 为零，那么力 F 自然也是零。这样的话，答案可能就是 0？`\n\n` 不过，可能我的理解有问题。比如题目可能假设这个原子被电离了，带有一定的净电荷？不过题目没有给出任何关于原子被电离的信息，或者具体的电荷量。这样的话，可能需要更多的假设。或者题目是否指原子中的某个部分，比如原子核或者电子受到的力？不过题目明确说是一个原子，所以应该是整体的受力。`\n\n` 不过再仔细想想，如果原子是极化的，那么虽然整体中性，但在外电场中会发生极化，正负电荷中心分离，从而产生电偶极矩。这时候，虽然总电荷为零，但如果电场是

均匀的,那么作用在正电荷和负电荷上的力大小相等方向相反,总力还是零。但如果电场是非均匀的,那么可能会有净力,即所谓的电偶极子在非均匀电场中的力。不过题目中给出的电场是 $2.5 \times 10^4 \text{ V/m}$,并没有提到电场是否均匀或者是否存在梯度。通常如果没有特别说明的话,可能默认是匀强电场,也就是均匀的,这时候偶极子会受到力矩使其转向电场方向,但合力仍然为零。不过这里的问题是一个静止的原子,可能极化的情况是否要考虑? \n\n 不过在这种情况下,题目可能希望做一个简单的处理,即认为原子是点电荷,带有净电荷量。可是原子本身是电中性的,所以净电荷为零,所以受力为零。或者题目是否可能存在其他情况?比如这个原子是否处于某种离子态?题目并没有说明,所以可能不能这样假设。 \n\n 或者题目是否有其他的意图?比如,在电场中原子的电子和原子核会受到相反方向的力,但由于原子是束缚态,所以整体受到的合力可能被内部的作用所抵消,所以宏观上的合力为零? \n\n 这样的话,答案可能就是原子受到的净力为零。不过,或许需要用公式推导来确认这一点。 \n\n 假设原子的电荷量为 q ,那么如果 $q=0$,那么力 $F = qE = 0$ 。所以答案应该是 0。这可能是一个陷阱题,测试学生是否记得原子是电中性的,所以尽管存在电场,但受力为零。 \n\n 不过可能我哪里错了?比如是否考虑到了极化导致的电偶极矩的影响? \n\n 如果是的话,在均匀电场中,电偶极矩所受的合力确实是零,但会有力矩。不过题目是问受力情况,而不是力矩,所以合力仍然是零。因此,无论是否极化,只要电场均匀,原子作为中性粒子,合力都是零。 \n\n 因此,我认为答案应该是原子所受的合力为 0,单位是牛顿。不过需要确定题目是否有什么条件被忽略了。题目中提到的原子是静止的,可能是指不考虑由于极化导致的运动,或者初始时刻的受力情况? \n\n 总之,在没有其他信息的情况下,最合理的结论应该是原子受到的力为零,因为原子是电中性的。所以答案应该是 0 牛。 \n\n 不过,为了确认,我再检查一下。假设原子带有总电荷 q ,那么受力为 qE 。而原子是中性的, $q=0$,所以 $F=0$ 。 \n\n 是的,这样的话,结论正确。 \n\n 或者题目是否可能希望计算极化引起的某种力?比如,在非均匀电场中,但题目中并没有说明电场是否非均匀。这时候,如果电场是均匀的,即使极化,也不会产生合力,所以答案还是零。 \n\n 所以最终答案应该是 0 牛顿。 \n\n ** 答案 ** \n 原子静止时在电场中受到的力为 $\boxed{0}$ 。 \n\n </think>

answer: \n\n\n 要解决一个静止的原子在电场强度为 $2.5 \times 10^4 \text{ V/m}$ 的情况下受到的力,首先需要考虑原子的电中性特性。原子由原子核和电子组成,整体电荷为零,因此作为中性粒子,其净电荷量 q 为零。 \n\n 根据库仑定律,电场中带电粒子受到的力为 $F = qE$ 。由于原子的总电荷量 $q = 0$,因此代入公式得: \n\n [$F = qE = 0 \times 2.5 \times 10^4 \text{ V/m} = 0 \text{ N}$] \n\n 即使考虑极化效应,在均匀电场中,正负电荷中心分离产生的电偶极矩会受到力矩作用,但合力仍为零。因此,无论是否极化,静止的中性原子在均匀电场中受到的合力为零。 \n\n 最终答案: \n\n [$\boxed{0}$] \n\n

stem_zh/chem 数据集示例

query:请解释配位催化是如何加速化学反应速度的?

think: <think> \n 嗯,用户让我解释配位催化是怎么加速化学反应速度的。这个问题看起来有点挑战性,不过我还是先回忆一下学过的东西。配位催化好像是和催化剂有关的,对吧?催化剂一般都是通过降低反应活化能来加快反应速率,但配位催化有什么特别之处呢? \n\n 首先,配位催化可能涉及到过渡金属催化剂,比如铂、钯这些。记得过渡金属有 d 轨道,可以形成配位键,和反应物结合。那这种结合是怎么促进反应的呢?可能是反应物吸附在催化剂表面,形成中间体,让反应更容易进行。或者催化剂提供一个不同的反应路径,活化能更低。 \n\n 另外,配位催化可能在均相催化中常见,比如溶液里的金属配合物作为催化剂。比如氢化反应中的 Wilkinson 催化剂,或者羰基化反应用的铑催化剂。这些金属配合物能够和底物配位,改变它们的电子结构,使某些键更容易断裂或形成。 \n\n 比如,烯烃的氢化反应中,金属催化剂可能先吸附氢气,使其解离成原子,同时烯烃也配位到金属上。这样,氢原子更容易加到双键上,因为中间形成了金属-烯烃-氢的配合物,降低了反应的能垒。 \n\n 还有,配位可能使分子定向排列,便于反应发生。比如在聚合反应中, Ziegler-Natta 催化剂通过配位使单体以特定方式排列,促进聚合,同时控制立体结构。这也是配位催化加速反应的一个方面。 \n\n 不过,用户可能想知道具体的机理步骤,比如催化剂如何与反应物作用,中间体的形成,以及怎样影响过渡态。可能需要分步骤解释:吸附或配位、活化、反应、脱附。每一步如何降低活化能。 \n\n 还有,配位催化中的氧化加成和还原消除步骤,比如在交叉偶联反应中,金属催化剂经历这些步骤,促进键的形成和断裂。这些步骤可能比非催化路径更高效,所以整体反应加快。 \n\n 另外,是否要考虑电子效应和空间效应?配体对金属中心的电子状态有影响,可能调节催

化活性。比如强场配体使金属更易接受电子，促进某些反应。空间位阻大的配体可能影响底物的接近方式，从而改变反应路径。
可能还要举几个例子来说明不同情况下的加速机制。比如酶中的金属辅因子，也是配位催化的例子，通过精确的配位环境高效催化反应。但用户的问题比较基础，可能需要用更一般的术语解释，而不是深入特定例子。重点在配位如何改变反应路径，稳定过渡态，中间体，降低活化能。
总结一下，配位催化加速反应的主要原因是催化剂与反应物形成配位中间体，改变反应路径，提供低能垒的途径，同时稳定过渡态或中间体，减少所需能量。此外，配位还能使反应物正确取向，提高有效碰撞几率，从而加快反应速度。

answer:
配位催化通过以下机制加速化学反应速率：
1. 形成活性中间体，降低活化能
配位键的形成：过渡金属催化剂（如 Pd、Pt、Rh 等）通过 d 轨道与反应物（底物或试剂）形成配位键，生成金属-底物配合物。这种中间体比自由反应物更活泼，例如：在氢化反应中，金属催化剂吸附 H₂ 使其解离为活性氢原子。在羰基化反应中，CO 通过配位活化 C-O 键。
稳定过渡态：金属中心通过电子相互作用（如反馈 π 键）削弱反应物中的化学键（如 C=C、C $\equiv$ C），降低键断裂的活化能。
2. 定向效应与空间控制
分子取向优化：配位作用使反应物以特定几何构型固定在金属周围（如八面体、平面正方形），确保反应基团精准接近（如烯烃的顺式加成）。
立体选择性：Ziegler-Natta 催化剂通过配位控制单体的插入方向，实现高立体规整度的聚合物合成。
3. 多步反应的协同作用
氧化加成与还原消除：在交叉偶联反应（如 Suzuki 反应）中，金属催化剂通过氧化加成活化卤代烃，再通过配位促进 C-C 键形成，最后还原消除释放产物。
多中心协同催化：酶中的金属辅因子（如细胞色素 P450 的铁-卟啉中心）通过多步配位活化底物和氧气，实现高效氧化。
4. 电子效应调控
配体调变金属活性：给电子配体（如膦配体）可增强金属中心的电子密度，促进亲核进攻；吸电子配体（如 CO）则可能促进亲电步骤。
电荷转移：金属与底物间的电子传递（如 Ni 催化偶联反应中的电子转移）可加速自由基生成或离子对重组。
实例说明
氢甲酰化反应：Rh 催化剂与烯烃和 CO 配位，使 H₂ 和 CO 插入到烯烃中，生成醛类化合物。配位作用使 CO 的插入步骤活化能显著降低。
酶催化固氮：固氮酶中的铁钼辅因子通过多核金属簇配位 N₂，削弱 N $\equiv$ N 三键，使其在温和条件下还原为 NH₃。
总结
配位催化通过金属与反应物的动态配位，重构反应路径，在原子经济性和能量效率上远超传统催化。其核心在于利用配位键的“分子杠杆”作用，将原本高能垒的基元反应转化为多步低能垒过程，最终实现反应速率数量级的提升。

stem_zh/med 数据集示例

query:这个病人被诊断出患有心血管疾病和肝癌，他的肿瘤标志物 CA 19-9 是 750 IU/ml，请问这表示什么？这个数值是否可以作为判断病情的标准？

think:
嗯，用户问的是，一个同时有心血管疾病和肝癌的病人，CA19-9 是 750 IU/ml，这代表什么，能不能作为判断病情的标准。首先，我需要先回忆一下 CA19-9 的正常范围和常见升高的疾病。
CA19-9 通常用于胰腺癌的辅助诊断和监测，正常值一般小于 37 IU/ml，这里 750 明显高很多。不过，很多情况都可能引起 CA19-9 升高，比如胰腺炎、胆道梗阻、其他消化道癌症，比如结肠直肠癌、胃癌，还有肝癌也有可能。但肝癌通常更常用 AFP 作为标志物，CA19-9 可能在某些情况下升高，特别是胆管细胞癌，也就是肝内胆管癌，这时候 CA19-9 会比较高。
但用户提到的是肝癌，可能指的是肝细胞癌（HCC），这时候 AFP 可能更相关。不过这个患者同时有心血管疾病，比如心衰或者肝硬化导致的心血管问题？不过心血管疾病本身不会直接导致 CA19-9 升高，除非有继发的肝胆问题，比如淤血性肝病导致的胆道问题，这时候可能会有轻到中度的升高，但 750 这么高可能不太常见，需要考虑其他原因。
所以，这个患者的 CA19-9 升高到 750，可能提示有恶性肿瘤的存在，但需要结合影像学检查和其他肿瘤标志物。因为肝癌有不同类型，如果是胆管细胞癌，CA19-9 高就有意义，而肝细胞癌则可能 AFP 高。所以需要确定肝癌的具体类型。
另外，其他因素比如梗阻性黄疸，不管是结石还是肿瘤引起的胆道梗阻，都可能导致 CA19-9 显著升高，所以需要排除胆道阻塞的情况。可能要做腹部超声、CT 或 MRI，看看有没有胆管扩张或者其他占位性病变。
关于是否能用这个数值作为判断病情的标准，可能不太够。虽然 CA19-9 水平高通常与肿瘤负荷、进展相关，但单独一个指标不能确诊，需要结合其他检查结果。比如，如果影像学显示肝部肿瘤，同时 CA19-9 高，可能支持胆管癌的诊断。但如果是肝细胞癌，可能更依赖

AFP。此外，治疗后 CA19-9 的变化可以用来监测治疗效果和复发情况，但初始诊断不能单独依靠这个指标。

需要注意的是，大约有 5-10% 的人群由于 Lewis 抗原阴性，无法合成 CA19-9，所以即使有恶性肿瘤，这个指标也可能不升高，所以阴性结果不能排除肿瘤。但在这个病例中，数值很高，所以更有意义。

此外，患者有心血管疾病，是否可能影响 CA19-9？比如，是否合并有肝脏淤血导致肝功能异常，进而影响 CA19-9？不过通常这种肝淤血更多是引起转氨酶和胆红素升高，而 CA19-9 可能轻度升高，但到 750 的可能性较低。所以更可能还是恶性肿瘤导致的。

总结来说，这个患者的高 CA19-9 提示可能存在胆道系统或胰腺的恶性肿瘤，需要结合影像学和其他检查确认。而肝癌如果是胆管细胞癌，这个数值有意义，但如果是肝细胞癌，需要看 AFP 的情况。不能单独用 CA19-9 作为判断病情的唯一标准，需综合评估。

answer: 对于该患者 CA19-9 水平显著升高至 750 IU/ml（正常值通常 <37 IU/ml），需结合临床背景综合分析：

- CA19-9 升高的可能原因**
 - 肝胆胰恶性肿瘤**：CA19-9 显著升高（>1000 IU/ml）常见于胰腺癌或胆管癌，但肝癌（尤其是胆管细胞癌）也可能导致明显升高。
 - 其他消化道肿瘤**：结直肠癌、胃癌等可能引起中度升高。
 - 良性疾病**：胆道梗阻（结石或炎症）、胰腺炎、肝硬化等可能导致轻中度升高，但极少达到 750 IU/ml。
 - 心血管疾病相关性**：单纯心血管疾病（如心衰）通常不会直接导致 CA19-9 显著升高，但若合并肝淤血或继发胆道问题，需排除其他因素。
- 针对肝癌的诊断意义**
 - 肝癌类型差异**：
 - 肝细胞癌（HCC）**：主要依赖甲胎蛋白（AFP），CA19-9 升高较少见。
 - 胆管细胞癌（ICC）**：CA19-9 常显著升高，可能作为重要标志物。
 - 需警惕混合型肝癌或转移癌**：若患者同时存在肝内胆管癌或转移性胰腺癌，CA19-9 可能显著升高。
- 临床建议**
 - 影像学检查**：优先进行腹部增强 CT/MRI 或 PET-CT，明确肝内病变特征（如胆管扩张、占位性病灶）及是否存在胰腺或胆道肿瘤。
 - 其他肿瘤标志物**：联合检测 AFP（肝细胞癌）、CEA（结直肠癌）等，辅助鉴别肿瘤类型。
 - 排除梗阻性黄疸**：通过超声或 MRCP 检查胆道系统，排除结石或肿瘤导致的胆道阻塞。
 - 动态监测**：治疗后 CA19-9 水平变化可用于评估疗效及复发，但需结合影像学。
- 数值的局限性**
 - 非特异性**：CA19-9 升高不能单独确诊恶性肿瘤，需结合临床表现和影像学。
 - 假阴性风险**：约 5%-10% 人群因 Lewis 抗原阴性无法表达 CA19-9，可能导致假阴性。

结论 750 IU/ml 的 CA19-9 强烈提示恶性肿瘤可能，需重点排查胆管癌、胰腺癌或混合型肝癌，但不可单独作为诊断标准。建议综合影像学、病理及其他标志物（如 AFP）明确病因，并根据肿瘤类型制定治疗方案。

stem_zh/bio 数据集示例

query: 请介绍螺旋藻生长的最优 pH 值及其与生产效率的关系

think: 嗯，我现在要介绍螺旋藻生长的最优 pH 值以及它和生产效率的关系。首先，我需要先了解一下螺旋藻是什么，可能它属于蓝藻门，对吧？螺旋藻通常用于营养补充剂，因为蛋白质含量高。生长条件的话，pH 值应该是一个重要因素，因为大多数微生物都有自己适合的 pH 范围。

首先，螺旋藻的最适 pH 值是多少呢？记得以前学过，可能是在碱性条件下生长较好，比如 pH 8 到 11 之间？或者是更具体的数值？比如 9 到 10.5？可能不同的研究有不同的结果，但大致在碱性范围内。然后，为什么螺旋藻偏好这样的高 pH 环境？可能和它们的自然生境有关，比如在非洲的碱水湖，比如乍得湖或者墨西哥的 Texcoco 湖，那里的水可能 pH 较高。

接下来要考虑 pH 如何影响螺旋藻的生长和生产效率。如果 pH 太低，比如中性或酸性，可能螺旋藻的生长会受到抑制，为什么？可能是因为酶的活性受到影响，或者细胞膜稳定性下降。另外，高 pH 环境下，可能抑制其他微生物的生长，减少了竞争，所以螺旋藻在这样的环境中更容易成为优势种，这可能和实际培养中的生物污染控制有关。所以维持高 pH 可能有助于减少杂菌，提高螺旋藻的纯度和产量。

生产效率方面，当 pH 处于最佳范围时，螺旋藻的代谢活动最强，繁殖速度最快，因此生物量的积累也更快。这可能意味着更高的生产效率，单位时间内产出更多。另外，合适的 pH 可能促进光合作用的效率，因为螺旋藻是光能自养生物，pH 可能影响光合酶系统和 CO₂ 的利用。例如，在高 pH 下，水中游离的 CO₂ 减少，但螺旋藻可能能够利用碳酸氢盐作为碳源，这可能是它们适应高 pH 环境的一种机制。所以，保持适当的 pH 值有助于优化碳源的利用，进而提高生产效率。

不过，pH 过高也可能有负面影响。比如 pH 超过 11 的话，是否会导致细胞损伤？或者影响营养物质的溶解度，比如某些金属离子在高 pH 下可能沉淀，导致螺旋藻缺乏必要的微量元素，如铁、镁等。这时

候可能需要调整培养液中的营养成分，或者控制 pH 不要过高。所以在实际生产中，可能需要在最佳 pH 范围内找到一个平衡点，既能促进螺旋藻生长，又不会引起营养缺乏或其他压力因素。

另外，pH 的稳定性也很重要。培养过程中，由于光合作用消耗二氧化碳，会导致 pH 上升，所以可能需要不断调节 pH，或者补充 CO₂ 来维持稳定。如果 pH 波动太大，可能影响螺旋藻的生长，导致生产效率下降。因此，pH 的管理策略也是生产效率的关键因素之一。

再想想，有没有相关的研究数据？例如，某些研究表明螺旋藻在 pH 9.5 时生长最佳，而另一个研究可能显示 pH 10 时产量最高。这可能取决于不同的菌株，以及培养条件的不同，比如温度、光照强度、营养供应等。因此，最优 pH 可能需要根据具体情况进行调整，但总体范围在碱性区域。

总结的话，螺旋藻的最优 pH 值大约在 9 到 10.5 之间，这个范围内它们的生长速率和生物量生产率最高。生产效率与 pH 的关系可能呈现钟型曲线，即在一定范围内随着 pH 升高而提高，但超过临界值后反而下降。需要综合考虑 pH 对代谢、竞争微生物、营养可用性等多方面的影响，来优化生产效率。

answer:螺旋藻（属于蓝藻门）的最优生长 pH 值及其与生产效率的关系可总结如下：

- 螺旋藻的最优 pH 值：螺旋藻偏好碱性环境，其最适生长 pH 范围通常为 9.0-10.5。不同菌株可能略有差异，但普遍研究显示：
 - pH 9.0-10.5：在此范围内，螺旋藻的代谢活性和生长速率达到峰值。
 - pH 8.0 以下：生长显著受抑制，因酸性或中性环境破坏其酶活性和细胞膜稳定性。
 - pH 超过 11.0：可能导致营养元素沉淀（如铁、镁）或细胞损伤，反而降低生产效率。
- pH 与生产效率的关系：
 - 促进光合作用与碳源利用：高 pH 环境（>8.5）下，水中游离 CO₂ 减少，但螺旋藻能高效利用碳酸氢盐（HCO₃⁻）作为碳源，支持光合作用和生物量积累。
 - 最佳 pH 范围内，光合色素（如藻蓝蛋白）合成增强，提升光能转化效率。
 - 抑制杂菌污染：碱性条件（pH 9-10.5）天然抑制多数细菌和藻类的竞争，减少生物污染，提高培养系统的稳定性。
 - 代谢效率最大化：在最优 pH 下，螺旋藻的酶活性（如 Rubisco、ATP 合成酶）和营养吸收效率（氮、磷）达到最佳状态，显著提升生长速率和蛋白质含量（通常占干重的 60-70%）。
 - pH 波动的影响：光合作用导致 pH 上升：培养过程中，螺旋藻消耗 CO₂ 会升高 pH，需通过补充 CO₂ 或调节碳酸盐缓冲系统维持 pH 稳定。剧烈波动可能引发代谢压力，降低生产效率。
- 实际生产中的 pH 管理策略：
 - 动态调控：通过自动化系统监测并补充 CO₂ 或碱性物质（如碳酸氢钠），维持 pH 在 9.5-10.0 左右。
 - 营养配比优化：高 pH 下需确保微量元素（如 Fe、Mg）以螯合形式存在，防止沉淀。
 - 菌株筛选：选择适应宽 pH 范围或高 pH 耐受性的菌株，增强生产稳定性。

结论：螺旋藻在 pH 9.0-10.5 时生产效率最高，表现为快速生长、高生物量积累和优质产物合成。合理的 pH 管理不仅能提升产量，还可降低污染风险，是实现规模化培养的关键因素之一。

exam/coig_exam 数据集示例

query:单项选择题，从下面 4 个选项中选择最佳正确选项
伊核问题、阿富汗问题、乌克兰危机…，过去的一年世界并不太平，中国作为负责任的大国积极参与并推动相关问题的解决。这说明 A、中国是维护和平的重要力量 B、强大的中国主宰着全球发展 C、和平不再是当今时代的主题 D、外交较量是国际竞争的实质

think:我现在得仔细看看这道题。题目是说，过去一年里，世界不太平，比如伊核问题、阿富汗问题、乌克兰危机这些，中国作为负责任的大国积极参与并推动解决这些问题。然后问这说明什么，选项是 A 到 D 四个。

首先，我得回忆课本里的相关知识点。题目问的是中国的国际角色，特别是关于和平与发展的内容。当前，和平与发展仍然是时代主题，对吗？所以如果选项 C 说“和平不再是当今时代的主题”应该是错误的，可以先排除 C 选项。然后 D 选项，外交较量是国际竞争的实质，但根据所学，国际竞争的实质是以经济和科技实力为基础的综合国力的较量，所以 D 不正确。B 选项说强大的中国主宰全球发展，这也不对，因为中国坚持不称霸，推动构建人类命运共同体，而不是主宰，所以 B 错误。剩下的就是 A 选项，中国是维护和平的重要力量，这符合题目里中国积极参与解决问题的描述，所以 A 正确。

不过再仔细看一下问题中的例子，中国参与解决的都是国际冲突或危机，这说明中国在维护世界和平方面发挥作用，所以 A 选项正确。其他选项中，B 的说法与我国外交政策相悖，中国始终是维护世界和平、促进共同发展的力量，而不是主宰。C 与时代主题相矛盾，D 选项的实质错误，所以答案应该是 A 选项。

answer:过去一年中，世界面临诸多冲突与危机，中国积极参与解决这些问题，这体现了中国在国际事务中的负责任态度。根据选项分析：**C**选项“和平不再是当今时代的主题”明显错误，因为和平与发展仍是当今时代主题。**D**选项“外交较量是国际竞争的实质”不符合事实，国际竞争的实质是综合国力的较量，尤其是经济与科技实力。**B**选项“强大的中国主宰着全球发展”违背中国“不称霸”的外交原则，中国主张合作共赢而非主宰。**A**选项“中国是维护和平的重要力量”正确，题目中中国推动解决国际冲突的行为正体现了这一角色。因此，正确答案为 **A**。

exam/kaoyan 数据集示例

query:下列关于个体进入催眠状态的表述中，错误的是
A. 进入睡眠状态
B. 主动性反应降低
C. 受暗示程度较高
D. 与在清醒状态下记录到的脑电波相似

think:嗯，这个问题看起来是关于催眠状态下的错误描述。题目给出四个选项，我需要找出错误的那个。首先，我得仔细看一下每个选项的内容。题目问的是错误的选项。我记得关于催眠的一些知识，比如催眠并不是真正的睡眠状态，而是一种类似高度放松和注意力集中的状态。所以选项 A 说进入睡眠状态，这可能就是错误的。因为虽然催眠的时候人看起来很安静，像睡着了一样，但其实脑电波可能更接近于清醒状态，而不是睡眠中的波形。比如，睡眠中的脑电波会有不同的阶段，比如慢波睡眠和快速眼动期，而催眠时的脑电波可能与清醒时相似，所以选项 D 是正确的，选项 A 错误。然后看选项 B，主动性反应降低。这可能正确，因为在催眠状态下，人的自主行动意愿可能会降低，更多地听从催眠师的暗示，所以他们的主动反应可能减少。选项 C 说的是受暗示程度较高，这显然正确，因为催眠的一个主要特点就是个体更容易接受暗示。所以 C 是对的。D 选项说与清醒状态下的脑电波相似，我之前学的知识说催眠时的脑电波类似清醒时的 α 波，所以 D 正确，而选项 A 是错误的，因为催眠并不是真正的睡眠状态，所以错误选项应该是 A。

answer:关于催眠状态的描述，分析如下：**A**（进入睡眠状态）错误。催眠并非真正的睡眠，而是介于清醒与睡眠之间的意识状态。此时个体虽闭目放松，但脑电波模式与睡眠阶段不同。**B**（主动性反应降低）正确。催眠状态下，个体的自主意识减弱，更依赖催眠师引导，主动行为减少。**C**（受暗示程度较高）正确。易受暗示是催眠的核心特征，个体对指令的接受度显著提高。**D**（与清醒状态下脑电波相似）正确。研究表明，催眠时的脑电波（如 α 波）与清醒放松状态相似，而非睡眠波形。综上，错误的表述是 **A**。

human_value/coig_human 数据集示例

query:做一个具有独立思考能力的人，以下哪个选项能更好地帮助你培养独立思考能力？
A. 多思考，多阅读，养成自己的见解并勇于表达
B. 追求群体认同，不分青红皂白地接受他人观点
C. 遵循传统、尊重权威，不轻易质疑标准答案
D. 只选择与自己观点一致的信息，避免接触不同的思想观点

think:嗯，这个问题是要选择能够帮助培养独立思考能力的选项。题目给了四个选项，我需要仔细分析每一个选项的意思，然后找出正确的答案。首先看选项 A：“多思考，多阅读，养成自己的见解并勇于表达。”这听起来挺合理的。多思考可以锻炼自己的分析能力，多阅读能接触不同的观点，这样有助于形成自己的见解。而且勇于表达可能意味着不害怕和别人不同，坚持自己的想法，这对独立思考应该是好的。然后是选项 B：“追求群体认同，不分青红皂白地接受他人观点。”这里的问题在于“不分青红皂白地接受”，也就是不管对错都接受别人的观点，这样显然不利于独立思考。如果只是随大流，为了得到群体的认同，那自己就不会去深入分析，容易盲从，所以这个选项应该不对。选项 C：“遵循传统、尊重权威，不轻易质疑标准答案。”虽然尊重传统和权威有一定的道理，但“不轻易质疑”就限制了批判性思维。独立思考的关键之一就是能够提出问题，挑战已有的观点，所以这个选项可能也不对。选项 D：“只选择与自己观点一致的信息，避免接触不同的思想观点。”这明显是信息茧房效应，只接受与自己相符的信息，拒绝不同的意见。这样会让自己的视野变得狭窄，无法全面思考，自然也不利于培养独立思考能力。综合

起来，选项 A 强调主动思考、广泛阅读和表达自己的见解，这些都是培养独立思考的重要方法。而其他选项 B、C、D 都涉及到盲从、不质疑或排斥不同观点，这些行为会抑制独立思考的发展。所以正确答案应该是 A。

answer: 要培养独立思考能力，关键在于主动分析、批判性判断和开放心态。

- 选项 A** 鼓励多思考、多阅读并形成自己的见解，这有助于通过广泛吸收信息和主动思辨来提升独立判断力，同时勇于表达能强化对自身观点的验证和修正，符合独立思考的核心要求。
- 选项 B** 盲目追求群体认同，被动接受他人观点，扼杀了个体分析问题的动力，直接阻碍独立思考。
- 选项 C** 强调遵循传统和权威且不质疑标准答案，削弱了批判性思维和质疑精神，不利于独立观点的形成。
- 选项 D** 局限于同质化信息，回避思想碰撞，导致思维固化，与独立思考的开放性和多元性背道而驰。

综上，只有 **A** 能有效培养独立思考能力。

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历：

蔡怀广，男，2000年7月出生，湖南省娄底人，中国科学院自动化研究所硕士研究生。

2018年09月——2022年06月，中山大学，计算机学院，计算机科学与技术系，工学学士学位。

2022年09月——2025年06月，中国科学院大学，中国科学院自动化研究所，模式识别与智能系统专业，攻读工学硕士学位。

已发表的学术论文：

- (1) **Huaiguang Cai**, Yang Yang, Yongqiang Tang, Zhengya Sun, Wensheng Zhang. Shapley value-based class activation mapping for improved explainability in neural networks [J]. The Visual Computer, 2025: 1-19.

参加的研究项目及获奖情况：

(1) 科技创新 2030——“新一代人工智能”重大项目，医疗行为多维度感知关键技术及应用研究，2020AAA0109600

(2) 国家重点研发计划“中医药现代化”重点专项，恶性肿瘤的中西医结合关键技术与诊疗方案研究，2023YFC3503300

(3) 国家自然科学基金青年基金项目，面向恶性心律失常的可解释认知推理模型与算法研究，62203437

