arXiv:2601.05103v1 [cs.DL] 8 Jan 2026

引文分类的语义正交框架：厘清意图与内容

Changxu Duan[1⋆[0000−0003−0547−0901]] 和 Zhiyin Tan[2⋆[0009−0002−4166−5810]]

[1] 达姆施塔特工业大学，达姆施塔特，德国
duan@linglit.tu-darmstadt.de[2] 莱布尼茨汉诺威大学 L3S 研究中心，汉诺威，德国 zhiyin.tan@l3s.de

摘要。 理解引文的作用对于研究评估和引文感知型数字图书馆至关重要。然而，现有的引文分类框架往往将引文意图（为何引用某项研究）与被引内容类型（引用了哪部分内容）混为一谈。由于细粒度类型区分与实际分类可靠性之间存在两难，这限制了它们在自动分类中的有效性。我们提出**SOFT**，即一个**S**语义**O**正交**F**框架，具有**T**两个维度，借鉴语义角色理论，明确区分引文意图与被引内容类型。我们使用 SOFT 对 ACL-ARC 数据集进行了系统性重新标注，并发布了一个从 ACT2 中抽样得到的跨学科测试集。使用零样本和微调后的大语言模型进行评估后发现，与 ACL-ARC 和 SciCite 标注框架相比，SOFT 能够提高人工标注者与大语言模型之间的一致性，并实现更强的分类性能和稳健的跨域泛化。这些结果证实，SOFT 可作为一种清晰且可复用的标注标准，提升数字图书馆和学术交流基础设施的清晰度、一致性和泛化能力。所有代码和数据均已在 GitHub 上公开：https://github.com/zhiyintan/SOFT。

关键词： 引文意图分类 · 标注框架 · 跨域泛化。

1 引言

引文是学术交流的核心，影响着研究社群如何认可、组织和复用知识。然而，传统引文指标（如引文数量或h指数）只能粗略衡量科学影响力，无法捕捉作品被引用、解读或质疑的种种微妙方式。为弥补这一不足，引文分类已成为计算文献计量学和信息科学中的一项关键任务。它通过分析文本语境来理解某项作品为何

⋆ 作者贡献相同，姓名按字母顺序排列。

被引用。这项工作的目标是区分先前作品是被用作基础方法、因其研究发现而受到批评，还是仅仅作为背景提及。这种功能层面的理解为新一代数字图书馆和智能研究助手提供了技术基础，使自动化文献综述 [3,15,27,28],、科学争议检测 [41], 以及影响力感知科学检索 [35,63]等先进服务成为可能。

自然语言处理（NLP）的最新进展使大规模引文分类得以部署于数字图书馆基础设施中。Web of Science [13], Scite [41], 和 Semantic Scholar [2] 等平台集成了旨在检测引文意图各个方面的功能，以改进检索、映射和评估。通过自动标注背景、方法、结果或支持，这些平台提升了透明度，加快了文献综述，并揭示了具有影响力或存在争议的贡献 [17]。

尽管采用日益广泛，当前的引文分类实践仍面临持续存在的局限。现有标注框架常常混淆不同的引文意图和内容类型，并且缺乏一致的语义定义。这些歧义阻碍了下游应用，例如基于引文的检索 [7] 和影响力分析 [37]。

本文提出 **SOFT**：一个具有**S**语义**O**正交性、包含**F**两种**T**维度的引文标注框架。SOFT 将被引用的内容（**被引内容类型**）与引用它的原因（**引文意图**）分离，从而实现更清晰的定义、更高的一致性，并更有力地支持基于大语言模型的分类。不同于以往的一维框架，SOFT 以语义角色原则为基础，通过对对象（内容）执行动作（意图），在概念上与作者组织论证的方式保持一致。

我们通过对 ACL-ARC 数据集进行全面的重新标注 [28] ，并在涵盖 19 个科学领域的 ACT 上进行跨领域评估2 [40], ，验证了 SOFT。结果表明，与现有框架相比，SOFT 能够提高模型间及人类与大语言模型间的一致性，提升零样本模型和微调模型的分类准确率，并在跨领域泛化方面表现得更加稳健。这些研究发现确立了 SOFT 作为一种可复用标注标准的地位，可用于支持具备引文感知能力的数字图书馆和透明的研究评估。我们的贡献如下：

– 我们提出 SOFT，这是一种语义基础扎实的二维引文标注框架，明确区分引文的谓述意图与其被引内容类型。– 我们依据 SOFT 框架对 ACL-ARC 数据集进行了系统性重新标注，并发布了来自 ACT2 的跨域测试集。– 我们使用零样本和微调后的大语言模型，将 SOFT 与 ACL-ARC 和 SciCite 进行基准比较，结果显示其在一致性、可学习性和泛化能力方面均有所提升。

2 相关工作

引文分类研究涉及多个方面，包括标注框架的开发、自动分类模型的设计，以及对这些模型在不同数据集和科学领域中的表现进行评估。第 2.1 节回顾了现有的引文意图分类框架，其中大多数采用单层标签，将引文的谓述功能与所引用内容的类型混为一谈。第 2.2 节综述了用于自动化引文意图分类的计算方法，涵盖从早期的基于特征的模型到近期由提示驱动的大语言模型。第 2.3 节指出，现有框架设计对跨领域适用性的探索有限，尤其缺乏对不同科学学科之间适用性的研究。最后，第 2.4 节讨论了重新标注基准数据集的相关工作，这些工作为我们使用二维框架重新标注 ACL-ARC 和 ACT2 奠定了基础。

2.1 引文意图分类框架

作者引用某一特定作品背后的功能性目的，曾以多种术语描述，包括 引文动机 [55, 59], 引文目的 [1], 引文类型 [45], 引文功能 [28, 60], 和 引文意图 [14]。为确保概念清晰且用语一致，本文通篇采用**引文意图**这一术语。

引文分类研究已从文献计量学传统发展到针对大规模建模量身定制的计算框架。最早且粒度最细的类型体系之一由 [55], 提出，其中定义了十三种引文角色，涵盖修辞、认识论和语境功能。引文功能语料库（CFC） [60] 是最早面向计算的引文资源之一。该语料库借鉴 Spiegel 的类型体系 [55],，定义了十二种功能标签，并将其归为四个超类（弱点、对比、正面和中性）。引文类型本体（CiTO） [52]引入了基于本体的框架，包含三十多种引文关系，旨在实现语义网互操作性。尽管 CiTO 形式严谨且表达力强，但在自然语言处理研究中的使用仍相对有限。

其他框架则是为特定领域标注和模型训练而设计的。 [1]提出的六标签框架（使用、批评、比较、佐证、依据和中性）通过对自然语言处理出版物进行标注，强调以目的为导向的类型。ACL-ARC [28]定义了一个六类型框架（背景、比较或对比、动机、用途、扩展和未来），旨在对引文影响力进行建模。SciCite [14] 则将这一六类型框架精简为三类型框架（背景、方法和结果比较）；这一做法基于这样的观察：动机、扩展和未来等类型往往发挥背景信息的作用，从而更好地支持计算机科学和生物医学领域的机器阅读与研究导航。

ACT 和 ACT2 [40,46] 将引文标注扩展到更多科学领域，采用了与 ACL-ARC 类似的六类型框架（将“比较或对比”替换为“比较与对比”，并扩展其含义，使其不仅涵盖引文之间的差异，也涵盖相似之处和分歧）。除意图外，ACT 和 ACT2 还使用两个二元标签（有影响力和附带）对影响力进行标注，并着重关注作者影响力和可比性。

这些框架共同塑造了引文意图分类的格局，为理解学术话语提供了宝贵见解。然而，许多框架采用单层标签结构，将作者的行为（即引文的功能）与其语义论元（即被引用内容的类型）隐式地结合起来。尽管这种设计对某些任务有效，却可能限制语义清晰度，并妨碍其在不同领域中的灵活复用。我们将在第 3 节重新探讨这些结构性挑战，这也促使我们设计一种新的、在语义上相互解耦的框架。

2.2 引文意图分类（CIC）模型

CIC 的计算方法已取得显著进展。早期模型依赖人工设计特征 [36,47], ，但 ACL-ARC 和 SciCite 等基准数据集的发布推动了向深度学习的转变。基于 Transformer 的编码器，尤其是面向特定领域的 SciBERT [10], ，迅速成为标准方法 [18,50]。近期进展则探索了多种学习范式，包括基于图的模型 [11], 多任务学习 [22,53], 和基于提示的方法 [30,32]。顺应最新趋势，近期研究也聚焦于对大语言模型进行微调，以完成这项任务 [29]。

在我们的实验中，我们使用提出的二维框架重新标注了 ACL-ARC 数据集，并将其用于模型训练和评估。为评估跨领域泛化能力，我们还对 ACT2 数据集中的一个样本进行了重新标注。这些评估的结果见第 5 节，其中我们根据该框架比较了 CitePrompt 和 CitationInten-tOpenLLM 的表现。

2.3 跨领域泛化能力

CIC 面临的一个长期挑战是模型在不同科学领域或数据集之间的泛化能力有限 [33,53,64]。要严格评估模型真正的泛化能力，必须对模型进行跨领域测试：在源数据集（例如 ACL-ARC）上训练，并在不同的目标数据集（例如 ACT2）上评估 [12]。

跨域或跨数据集评估提供了一套原则性框架，用于测试泛化能力。在这种设置下，模型在一个源数据集上训练，并在独立的目标数据集上评估 [12]。这种设置广泛应用于各种自然语言处理任务，以考察模型学习到的是领域无关特征，还是过拟合了数据集特定模式。例如，先前的研究考察了问答任务在生物医学、维基百科和

网络数据集 [24], 之间的迁移性能，以及关系抽取任务在新闻和政治等不同子领域之间的迁移性能 [8]。另一个例子是机器翻译在资源丰富语言与低资源语言之间的泛化能力差异 [4]。

即使数据集表面上采用相同的标注框架，这种稳健性不足的问题也尤为明显。例如，ACT2 采用了 ACL-ARC 标签集，对多个学科的内容进行标注。尽管共享这一分类体系，先前研究表明，在 ACL-ARC 上表现良好的模型应用于 ACT2 [30]时，性能会大幅下降。这说明，表层标签的一致性并不意味着不同领域间存在语义等价，也不能保证模型能够稳健迁移。

我们认为，这些研究发现反映了现有引文分类框架存在更深层的结构性局限。具体而言，许多现有框架将引文的功能与其语义论元（被引内容的类型）混为一谈，导致标注歧义；而在异质的多领域场景中，这种歧义会进一步加剧。在实验中，我们通过使用重新标注的 ACL-ARC 数据集训练模型，并依据修订后的框架标注测试其在 ACT2 数据集上的表现，以评估这一问题。

2.4 重新标注数据集的适用性

监督式机器学习模型的性能，根本上取决于其训练数据的质量 [65]。然而，现实世界中的标注往往存在噪声，错误可能源于指南含糊、主观解释、标注人员专业知识有限，或简单的人为失误 [25,43]。这类错误会损害模型训练、降低评估的可靠性，并削弱下游应用的稳健性。

认识到这些局限后，近期研究着重系统性地识别并纠正广泛使用的基准数据集中的标签错误。例如，对 COCO 数据集 [34] 的研究揭示了目标掩码标注中的缺陷，进而推动了 COCO-ReM [54] 和 Sama-COCO [67],等项目；这两个项目都引入了修订后的标注，提升了下游模型的性能。类似地，CleanCoNLL [51]项目对 CoNLL-03 命名实体识别数据集 [61], 进行了全面重新标注，纠正了 7% 的实体标签，并证明基准数据集的可靠性得到了显著提升。

与此同时，研究人员开发了自动化和半自动化方法，以辅助错误检测与纠正。以模型为驱动的方法（如 Confident Learning）会估计观测标签与真实标签的联合分布，以识别可能存在的错误标注 [42],；而 ObjectLab 等领域专用工具则会运用经过训练的目标检测器，标记标注质量方面的异常 [62]。近期，研究人员也开始探索将大语言模型用作标注代理 [23, 56], ，但在这类应用中，仔细的质量控制仍然至关重要。

在本研究中，我们使用所提出的二维框架对 ACL-ARC 数据集 [28]进行人工重新标注。为进行比较和内部验证，我们还在受控提示条件下，让一组开源大语言模型执行相同的重新标注。为了评估标签质量和框架

的清晰度，我们计算了人工标注者与大语言模型之间的一致性。这些实验使我们能够评估该框架的可靠性，并探讨大语言模型在未来标注工作流中的潜在作用。

3 SOFT：一个语义正交双维框架

为满足对语义稳健、可解释且可迁移的引文标注日益增长的需求，我们提出了一个以话语意图和被引内容结构为基础的新框架。在介绍其设计之前，我们首先考察现有分类框架的局限性，这些局限促使我们摒弃传统方法。这些问题源于结构性纠缠、功能角色界定不足以及指称目标含糊不清，都会妨碍标注的一致性和下游模型的性能。

3.1 现有引文分类框架的局限性

尽管 CiTO [52], 、ACL-ARC [28], 、SciCite [14], 和 ACT/ACT2 [40, 45], 等广泛使用的框架取得了进展，但若干结构性局限依然存在。这些问题降低了标注的清晰度和一致性，妨碍模型泛化，并使引文行为的语义解释更加复杂。我们指出了三个相互关联的问题：

（1）维度纠缠：混淆被引内容类型与引文意图。许多早期引文分类框架未能区分被引内容类型（如方法、数据或发现）与引文意图或功能（如使用），而是将这两个维度合并为一个标签。例如，在 CiTO 中，“使用某方法”“使用某数据”和“使用某结论”等类型，将使用这一行为与所使用的具体对象明确绑定，导致内容类型与引文意图相混淆。这种设计使类型数量激增，却仍遗漏重要的边界情况，增加了标注员的认知负担；此外，类型激增和无法分类的实例也使建模更具挑战性。相比之下，SciCite 以另一种方式体现了相关局限：其“背景信息”类型定义宽泛，涵盖了所有不属于方法或结果类型的引文行为。因此，许多具有特定功能意图的引文最终都被归入背景类型，其实际的学术作用也因此变得模糊。

(2) 视角歧义：作者承诺与话语声音。此外，不同类型间视角不一致，也会给标注员带来更多困惑。例如，在 ACL-ARC 框架中，背景、动机和未来使用等类型以被引作品作为动作主体

而用途、扩展、比较或对比等类型则以引用作品作为行动主体。这种视角转换会造成认知上的不一致，使标注过程更加令人困惑。

（3）缺乏功能导向的意图框架。此外，潜在的引文意图并不总是得到明确说明。例如，被标记为比较或对比的引文可能是在对齐研究发现、方法或概念框架，但这种比较究竟是用于论证某项设计、批评某个先例，还是仅仅提供背景，完全取决于引用作者的功能目标。现有框架没有提供区分这些不同功能的机制，导致引文行为的谓词动作含义模糊不清。在缺乏将作者的行为与该行为对象明确分离的模型时，标注依赖于推断和主观解释，从而降低了可复现性和标注者间一致性。

3.2 SOFT 引文标注框架

我们提出了 **SOFT**，这是一种旨在克服以往引文分类系统核心局限性的二维引文标注框架。SOFT 直接针对三个主要问题：（1）内容类型与引文功能相互纠缠；（2）标注视角不一致；（3）引文意图含糊或说明不充分。

首先，SOFT 明确将**被引内容类型**（即被引作品中引用了哪些具体内容）与**引文意图**（即引用作品为何引用这些内容）区分开来，从而解决了长期存在的维度纠缠问题，确保各个维度彼此正交且语义边界清晰，大幅减少歧义和认知负荷。

其次，SOFT 始终对引用作品的行为进行建模，从而保持统一的作者视角。这种方法以语义角色理论为基础 [6, 20, 21, 44],，该理论将事件分解为谓词（行为）和论元（参与者或相关实体）。在 SOFT 中，引文意图始终对应引用作品所执行的行为或交际功能，例如“使用来自 [被引作品]的 X”、“修改来自 [被引作品]的 X”或“对照评估来自 [被引作品]的 X”，其中 X 是具体的被引内容。这些意图，如“使用”或“修改”，与语义角色理论中的经典事件谓词相对应。SOFT 中的其他意图，例如“使用来自 [被引作品]的 X 提供背景”、“使用来自 [被引作品]的 X 指出研究空白”、“强调 [被引作品]中 X 的局限性”或“基于 [被引作品]中的 X 论证设计选择”，则涵盖了科学写作中更广泛的话语行为。虽然后几种情况超出了简单的事件谓词范畴，但所有 SOFT 意图类型都旨在明确指出引用作品正在做什么，确保每项标注都立足于清晰且一致的作者视角。

第三，SOFT 为所有引文意图类型提供了明确且语义精确的定义，确保每种情况下的底层作者意图都得到清晰说明。例如，我们将 Background（用于 ACL-ARC、SciCite，

以及 ACT/ACT2）重新表述为 Contextualize，以标记作者有意采取的行动；并将 Extension 改为 Modify，以涵盖的不仅是扩展，还包括缩减、替换或新颖组合。我们用 SignalGap 和 HighlightLimitation 替换引文功能 Motivation（用于 ACL-ARC 和 ACT/ACT2），从而能够细致地区分指出研究空白与强调局限性。为进一步减少歧义，我们删除了 Comparison or Contrast 类型（见于 ACL-ARC），因为该类型在实际应用中常被混淆，转而引入 JustifyDesignChoice 来论证决策，并引入 EvaluateAgainst 来进行直接的实证比较。

SOFT 类型清单通过多种数据集和现有框架（包括 SciCite、ACL-ARC 和 CiTO）进行实证推导并迭代完善，系统地消除了标签歧义和语义范围混杂的问题。因此，SOFT 提供了一个更具可解释性且更稳健的标注框架，提升了话语分析和引文影响检测等任务中的人工标注一致性与模型泛化能力。下文将详细介绍该框架的两个核心维度。

维度 1：被引内容类型 指的是被引作品中所引用贡献的本体论状态。我们区分三种类型：

(1) 已完成工作：引用作品提及被引作品所做的工作（例如，实验过程、流程设计），而不单独指出具体成果或可重复使用的资源。示例："/被引作品/ 开发了一个自然语言处理流程。"

(2) 发现：引用作品提及被引作品得出的观察结果、研究发现或理论结论。示例："/被引作品/ 观察到随机失活可提高稳定性。"

(3) 产出资源：引用作品提及可重复使用的产出，例如数据集、算法、模型、工具、度量指标、标准化设置等。示例："我们使用了 /被引作品/中的解析器。"

这些区分有助于细致分析不同形式的科学贡献如何在学术话语中被引用、再利用或加以阐述。

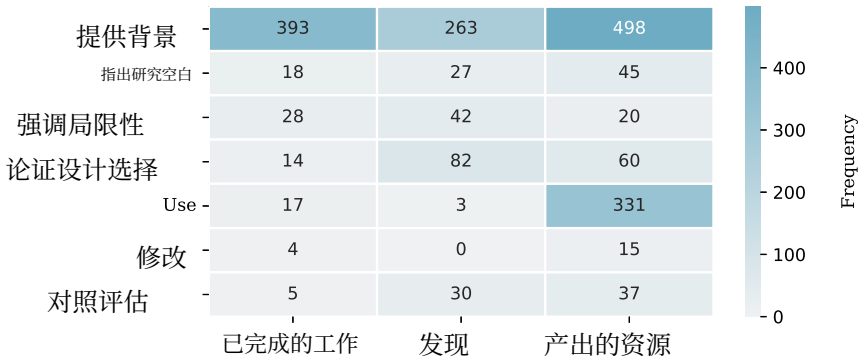


图 1：重新标注的数据集的统计数据类型。

维度 2：引文意图刻画引文的谓述功能，体现引用作者所执行的动作。每个标签都被定义为一个谓词，并将被引内容作为其核心论元之一。我们从引用作者作为引文行为施事者的视角来界定每个标签。

(1) 提供背景：提及被引作品是为了提供背景、展示先前研究或描述相关贡献。此处不涉及任何设计决策或复用。示例："/被引作品/ 提出了用于情感分析的层次注意力。"

(2) 指出研究空白：该引文指出尚未解决的问题或有待探究的问题。研究空白可能由被引作品或引用作品指出，无须承诺解决该问题。示例："/被引作品/ 指出，调优仍未得到充分探索。"

(3) 强调局限性：引用作品或被引作品指出被引方法或结果中的缺陷、限制或不足。仅当被引贡献受到明确批评时，才使用此类型。示例："/被引作品/ 未能捕捉用户行为中的领域漂移。"

(4) 论证设计选择：引用作品通过引用被引作品，为某项设计或方法选择提供支持。无须直接复用，但作者必须明确作出选择。示例："我们采用 /被引作品/的特征分组，以确保一致性。" 如果该行动仅是假设性的（例如"我们可以采用……"），则不适用此类型。

(5) 使用：引用作品直接应用被引作品中的可复用贡献（例如模型、流程、设置、定义）。引用者必须是执行者，因此"[被引作品] 使用了……"之类的表述不符合此类型。只有过去或现在时的应用才计入；计划或假设性的使用不属于"使用"。示例："我们使用 /被引作品/的 *BERT* 分类器。"

(6) 修改：引用作品对被引作品中的可复用贡献进行修改或扩展，例如调整配置、改变算法，或将其整合到新的流水线中。示例："我们调整了 /被引作品/ 的编码器，以处理多域输入。"

(7) 对比评估：引用作品明确将自身的研究发现和结果与被引作品的研究发现和结果进行比较，通常是为了验证有效性。示例："我们的模型在 /被引作品/的方法所使用的 F_1 分数指标上表现更佳。"

这些引文意图类型旨在彼此互斥，可与内容类型结合，以进行多维度解读。

设计原理与应用。 通过在结构上将引文意图与被引内容类型分离，SOFT 消除了单维框架中常见的歧义。例如，引用某个模型但未实际应用它时，标注为 Contextualize:ProducedResource；而提及尚未解决的问题时，则标注为 SignalGap:Discovery。只有当引用作者明确采取行动，例如重用、修改、论证或批评时，才会赋予主动意图。这种正交性有助于在不同引文上下文中保持标注一致，并提高标注可靠性。SOFT 的二维设计也能增

强下游应用。将空白信号与设计论证分开，可以更清晰地建模研究动机。区分工具的使用（Use:ProducedResource）与对方法趋势的引用（Contextualize:PerformedWork），有助于更细致地追踪科学影响、方法谱系以及对意图敏感的科学推荐。如第 5 节所示，SOFT 框架在人类与大语言模型之间实现了更高的一致性，取得了更好的分类性能，并在各科学领域展现出更强的泛化能力。接下来，我们将介绍标注指南和流程，包括对 ACL-ARC 的重新标注，以及来自 ACT2 的跨域样本。

4 实验

4.1 数据集与重新标注流程

数据集概述。ACL-ARC 数据集 [28]是标准的 CIC 基准，但其标注框架存在第 3.1 节讨论的概念混淆和边界问题。为解决这些缺陷，我们使用二维 SOFT 框架对整个数据集重新进行标注，从而提供一种经过改进且语义连贯的资源。

在我们的实验中，我们采用 [39], 提供的 ACL-ARC 扩展上下文版本。该版本通过增加更广泛的引文上下文和结构化元数据（例如引用论文和被引论文的标题，以及引文出现的章节）扩充了原始数据集。在重新标注之前，我们对数据集进行了人工清理，以纠正系统性的文本提取错误，例如分页符错位、阅读顺序混乱以及脚注顺序错误。此外，我们还从原始论文 PDF 文件中恢复了 533 个缺失的章节标题，以改善对背景的理解。ACL-ARC 数据集包含 1,931 个引文实例。按照 [28], 中的原始数据划分，我们采用相同的划分方式：1,647 个实例用于训练，284 个用于测试。

为评估跨域泛化能力，我们还对 ACT2 测试集 [40] 的一个子集（264 个示例）重新进行了标注。ACT2 涵盖微软学术图谱（MAG）中的 19 个顶级科学领域，主要领域包括心理学（21.99%）、医学（13.48%）、生物学（10.91%）和计算机科学（10.27%）。这些示例作为我们的分布外测试数据，可用于严格评估泛化表现。训练集和测试集的划分确保来自同一源文档的引文实例不会同时出现在两个集合中，从而防止数据泄漏并维护评估完整性。

重新标注流程。 重新标注过程由一组计算语言学研究生完成。为确保可靠性并顾及解释歧义，每个引文实例均采用修订后的二维框架，由至少三名标注员独立标注。标注员会标记分歧，随后通过结构化小组讨论加以解决；讨论中，标注员会审阅框架定义并进行审议，为每个案例达成共识类型。这种多轮策略有助于确保

类型一致性，并统一不同标注员的解释标准。重新标注后的 ACL-ARC 的统计数据类型见图 1。

4.2 评估框架可解释性与标注可靠性

标注框架质量的核心指标之一是标注者间一致性（IAA），它反映了框架中各类型定义得是否清晰且一致 [5, 9]。然而，ACL-ARC 的原始版本未报告 IAA [28], ，因此其框架的可解释性尚未得到量化。本研究也未报告人类标注者间一致性，因为标注过程是动态的：部分标注员中途离开，另有新成员加入，因此并非所有实例都由同一组人员标注，而且每个实例的标注数量也不尽相同。这种不断变化的标注员队伍类似于 SciCite 数据集采用的众包方式 [14], ，在这种情况下，计算覆盖整个数据集的人类标注者间一致性分数同样面临挑战。因此，我们通过基于大语言模型的一致性研究来评估框架的可解释性和可靠性。

为评估框架的清晰度与可靠性，我们使用开源大语言模型作为人工标注者的代理开展了一项一致性实验，从而以可扩展、可复现的方式评估标注可行性。具体而言，我们部署了四种不同的大语言模型：通义千问-*2.5-72B-Instruct*（通义千问） [48] , *Mistral-Small-24B-Instruct-2501* （Mistral） [38], *Llama-3.3-70B-Instruct* （Llama） [58],和 *Gemma-3-27B-it* （Gemma） [57], , 根据我们的框架和指南，独立标注 ACL-ARC 的引文上下文。对于每个上下文，每个大语言模型都会生成理由说明并作出分类决策，重复三次，最终类型由多数投票确定。随后，我们在两种设置下计算成对 Cohen's κ 分数 [16]：(1) 模型两两之间（大语言模型间标注者间一致性），以及 (2) 各模型与第 4.1 节所述的人类共识标注之间（人类与大语言模型间标注者间一致性）。后者衡量该框架在多大程度上能使大语言模型近似人类专家判断。

作为比较基线，我们采用六类型 ACL-ARC 框架和三类型 SciCite 框架重复了这一过程， [14], 指示相同的大语言模型对 ACL-ARC 背景进行标注。对于 ACL-ARC 框架，标注基于其原始数据集进行。对于 SciCite，我们将 ACL-ARC 类型映射到 SciCite 框架：将 Uses 对应到 Method，将 ComparisonContrast 对应到 ResultComparison，并将 Background、Extension、Motivation 和 Future 归为 Background。

这种基于大语言模型的模拟可作为人工标注一致性的实用代理，并支持在不同框架之间进行直接的实证比较。所得的标注者间一致性（IAA）分数为 SOFT 框架下的边界清晰度、可解释性和标注可行性提供了可量化的衡量指标。

4.3 通过分类测试正交性与解释清晰度

为实证评估我们修订后的标注框架是否比以往的框架具有更清晰的语义边界，我们开展了一项比较分类实验。这项设计基于以下假设，该假设的依据是

框架比较领域的既有研究 [26, 49], : 如果一个框架对类型的定义具有更精确、歧义更少的边界，那么使用此类数据训练的机器学习模型应能取得更高的分类性能。

分类器。 我们采用了两种架构各异、先进的开源引文意图分类模型:

– **CitePrompt** [32]: 一种基于提示的微调框架，构建于 SciB- ERT [10], 之上，并通过 OpenPrompt 工具包 [19]实现。–**CitationIntentOpenLLM** [29]: 一种基于**通义千问-2.5- 14B-Instruct**（Qwen-Small）的大语言模型分类器，使用 LLaMA-Factory 工具包 [66]进行微调。

除了这些微调模型外，我们还评估了使用第 4.1 节中介绍的大语言模型进行的若干零样本分类器的性能。

本研究选择分类器时，有意且强烈地优先考虑开源模型，而非闭源替代方案。这一承诺对于严谨科学探究的多个基本方面至关重要。首先，使用 SciBERT 等公开可用的模型以及各种指令微调大语言模型，能够确保研究的透明度和可复现性。这使其他研究人员能够独立验证我们的研究发现，并在此基础上开展工作，而不受专有系统往往存在的“黑箱”特性或访问限制的影响。其次，开源模型赋予我们深入分析和定制微调所需的关键灵活性，正如我们使用 LLaMA-Factory 调整 Qwen-Small 所展示的那样。模型架构和训练过程方面的这种精细控制，通常无法通过闭源 API 实现；而要严谨评估 SOFT 等新型标注框架并理解模型细微的行为差异，这种控制至关重要。最后，依赖开源解决方案有助于提升可及性并营造协作研究环境，这对于推动计算文献计量学领域的发展至关重要，同时也避免受到商业许可或不透明模型机制的限制。

评估指标。 遵循先前的引文意图分类评估协议 [29, 32], ，我们报告了**准确率**和**宏平均** F_1 **分数**。准确率提供了预测正确率的整体衡量，而宏平均 F_1 则计算各类别的平均F_1 分数，从而减轻标签不平衡的影响，并更好地反映类型级可学习性。

4.4 在 ACT2 上评估跨领域泛化能力

为评估我们修订后的框架是否能在不同学科背景下实现可泛化的引文意图分类，我们使用 ACT2 [40], 的测试子集进行了迁移评估。ACT2 [40], 是一个涵盖生物学、医学、经济学和计算机科学的多领域基准。如第 2.1 节所述，ACT2 采用了源自 ACL-ARC 的扁平框架。为使其适用于我们的评估，我们从 ACT2 中选取了一个具有代表性的子集作为跨域测试

数据集，并使用我们的二维框架，按照第 4.1 节所述的重新标注协议进行了标注。为进行受控比较，我们沿用第 4.3 节中的分类模型（CitePrompt、CitationIntentOpenLLM 和零样本大语言模型）及评估指标（准确率和宏观 F_1）。这种对齐方式使我们能够直接评估：在领域迁移情况下，我们的框架是否增强了跨域鲁棒性并保持了语义清晰度。

5 结果与分析

我们从三个维度评估 SOFT 框架：可解释性、域内可学习性和跨领域泛化能力。并将结果与两个现有引文框架 ACL-ARC 和 SciCite 进行比较。

表1：大语言模型与人工标注者之间的两两 Cohen's κ。

（a）ACL-ARC 框架（6 种类型）

	Llama	Mistral	Gemma	通义千问
Mistral	0.5833			
Gemma	0.4504	0.3860		
Qwen	0.5776	0.6013	0.3600	
人类	0.3956	0.4048	0.4302	0.3809

(b) SciCite 框架（3种类型）

	Llama	Mistral	Gemma	通义千问
Mistral	0.6363			
Gemma	0.6855	0.6423		
Qwen	0.6762	0.7151	0.6928	
人类	0.4016	0.3850	0.4618	0.4116

(c) SOFT：被引内容类型（3种类型）

	Llama	Mistral	Gemma	通义千问
Mistral	0.5396			
Gemma	0。5751	0。5657		
Qwen	0.5496	0.6169	0.5717	
人类	0.519	32	**0.590**	

(d) SOFT：引文意图（7 种类型）

	Llama	MistralGemma	通义千问
Mistral	0.6154		
Gemma	0.5941	0.5949	
Qwen	0.6282	0.5887	0.6175
人类	0.66		

可解释性与一致性。 我们通过计算四个预训练大语言模型两两之间的 Cohen's κ（大语言模型间），以及大语言模型预测与人工共识标签之间的 Cohen's κ（人类与大语言模型间），来评估框架的清晰度。正如预期，SciCite 的大语言模型间一致性最高，这归因于其较粗略的粒度。然而，如表1所示，SOFT 的人类与大语言模型间一致性显著更高，尤其是在**引文意图** 维度上（Llama 为 $\kappa = 0.662$，通义千问为 0.692）。尽管**被引内容类型** 在概念上更复杂，但其人类与大语言模型间一致性也优于 ACL-ARC 和 SciCite，这表明这两个维度都具有可解释性，且能够得到一致应用。

域内可学习性。 我们使用三种标注框架——ACL-ARC、SciCite 和 SOFT——在 ACL-ARC 测试集上评估微调模型和零样本模型的分类性能。在 SOFT 框架中，**被引内容类型** 和**引文意图** 被视为两个独立的分类任务。附录 B 中的表2和图3报告了所有模型的准确率和宏平均F_1 分数。尽管**被引内容类型** 在概念上更具体，其分类仍取得了较强的域内性能。微调后的 Qwen-Small 达到 0.78 的宏平均 F_1，而零样本模型的平均值为 0.66。尽管现有框架中没有可直接比较的基线，这些结果表明，内容类型区分不仅便于人类理解，也能被不同架构的模型学习。SOFT 在**引文意图**分类上的表现也优于以往框架。例如，微调后的 Qwen-Small 在 SOFT 下达到 0.65 的宏平均 F_1，而在 ACL-ARC 和 SciCite 上分别为 0.51 和 0.62。值得注意的是，尽管 SOFT-Intent 的类型数量（7）多于 ACL-ARC（6），仍取得了这一提升，这驳斥了性能仅取决于标签数量的假设。这表明，与标签粒度本身相比，框架的语义清晰度和明确界限对模型性能更为关键。SOFT 下零样本模型的平均 F_1 进一步印证了这一点，达到 0.69，显著高于 ACL-ARC（0.49）和 SciCite（0.61），说明 SOFT 的功能类型与预训练和微调模型的表征更加契合。

我们注意到，报告的分数低于先前研究中的分数，包括 CitePrompt [32], ；这可能归因于数据集划分的差异，正如第 4.1 节所讨论的那样。

跨领域泛化能力。为评估不同框架在领域迁移下的稳健性，我们直接在重新标注的 ACT2 子集上评估使用 ACL-ARC 训练的模型。未进行进一步微调或领域适应。附录 B 中的表 2 和图 3 报告了准确率和宏平均F_1 分数，展示了各框架在原始训练领域之外的泛化能力。尽管概念上较为复杂，**被引内容类型**分类仍具有良好的泛化能力。零样本大语言模型保持了良好的性能（平均 F_1 降幅为 17%），微调后的 Qwen-Small 性能则有所下降（从 0.78 降至 0.64），这表明这些内容区分在不同领域迁移下仍然稳定。经过 SOFT 训练的模型在**引文意图**分类上展现出更强的泛化能力。微调后的 Qwen-Small 分数从 0.65 降至 0.56（SOFT），优于 ACL-ARC（从 0.51 降至 0.49）和 SciCite（从 0.62 降至 0.23 ）。零样本性能也呈现出相同趋势（图 2 ）。SOFT 的平均 F_1 降幅仅为 15.8%（其他框架为 30%–65%），表明其基于话语的意图类型具有更强的泛化能力。

维度比较。 维度分析揭示了不同的模型行为。在域内，零样本大语言模型在**引文意图**（平均 F_1：0.69）上的表现优于**被引内容类型** （0.66），这表明这些功能与大语言模型先验具有良好的一致性。然而，经过微调后，**被引内容**

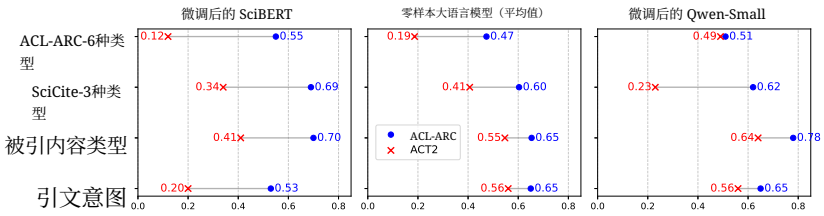


图 2：四种标注框架在域内和跨域设置下的宏观 F_1 分数比较。每个子图对应一种模型类型，蓝色圆圈表示域内表现，红色叉号表示跨域表现。

类型 变得更易于学习（Qwen-Small F_1：0.78，而意图为0.65），这表明其区别更能从监督中受益。两个维度的跨域泛化能力都具有竞争力，微调后的通义千问在内容（Qwen-Small：18%；SciBERT：41%）和意图（Qwen-Small：14%；SciBERT：62%）上的性能下降幅度都较为有限。这凸显了 SOFT 在领域迁移下的稳定性，以及微调后的通义千问相较于 SciBERT 更强的迁移能力。

总体分析。 在所有实验中，SOFT 框架在可解释性、可学习性和跨域稳健性方面始终优于 ACL-ARC 和 SciCite。微调后的 Qwen-Small 在**被引内容类型**上取得最佳结果，而零样本通义千问在**引文意图**上表现最佳。这些研究发现确立了 SOFT 作为一种语义上正交且可泛化的引文标注替代方案。

6 结论

本文介绍了 SOFT 框架，这是一种语义正交框架，包含用于引文标注的两个维度，将被引内容类型与引文意图区分开来。针对现有框架中概念混淆和边界模糊的问题，SOFT 提供了更清晰的解释性定义，并更好地支持人工标注和基于模型的分类。针对大语言模型一致性、域内分类和跨域泛化开展的实证评估证实了 SOFT 的优势。引文意图维度展现出出色的零样本和微调性能；而被引内容类型维度尽管语义复杂，也证明了其可学习性和可迁移性。与 ACL-ARC 和 SciCite 相比，SOFT 在人类与大语言模型的一致性方面始终更高，域内宏观 F_1 表现更强，跨域泛化也更稳定。这些优势在指令微调大语言模型上尤为显著：经过微调后，它们仍能保持更广泛的泛化能力。

未来的工作将探索扩展 SOFT，以支持多标签引文语境、整合更细粒度的功能类型，并将该框架应用于大规模引文图，以实现话语层面的功能建模、影响力追踪和知识提取。

致谢

Changxu Duan 获得联邦教育与研究部（BMBF）资助，资助项目编号为 01UG2130A，资助来源为 InsightsNet 研究项目（insightsnet.org）。Zhiyin Tan 获得“HybrInt——通过可解释人工智能实现机器感知与交互中的混合智能”项目资助（Zukunft Nds，下萨克森州科学部，项目编号：ZN4219）。我们感谢 Sabine Bartsch 博士协助校对。我们也感谢匿名审稿人提出的深刻意见和建议。

参考文献

1. Abu-Jbara, A., Ezra, J., Radev, D.：引文目的与极性：迈向基于自然语言处理的文献计量学。载于：Vanderwende, L., Daumé III, H., Kirchhoff, K.（编）《2013年计算语言学协会北美分会会议论文集：人类语言技术》。第596–606页。计算语言学协会，亚特兰大，佐治亚州（2013年6月），https://aclanthology.org/N13-1067/ 2. 艾伦人工智能研究所：Semantic Scholar。https://www.semanticscholar.org/（2025年）3. Anderson, M.H., Lemken, R.K.：引文语境分析：开展严谨且具有影响力的文献综述的一种方法。《组织研究方法》**26(1)**，77–106（2020年12月）。https://doi.org/10.1177/1094428120969905, http://dx.doi.org/10.1177/1094428120969905 4. Araabi, A., Niculae, V., Monz, C.：联合丢弃：通过短语对变量提高低资源神经机器翻译的泛化能力。载于：Utiyama, M., Wang, R.（编）《第十九届机器翻译峰会论文集》第1卷：研究分会。第12–25页。亚太机器翻译协会，中国澳门特别行政区（2023年9月），https://aclanthology.org/2023.mtsummit-research.2/ 5. Artstein, R., Poesio, M.：综述文章：计算语言学中的编码者间一致性。《计算语言学》**34(4)**，555–596（2008年）。https://doi.org/ 10.1162/coli.07-034-R2, https://aclanthology.org/J08-4004/ 6. Baker, C.F., Fillmore, C.J., Lowe, J.B.：伯克利框架网项目。载于：《第36届计算语言学协会年会暨第17届国际计算语言学大会论文集》第1卷。第86–90页。计算语言学协会，加拿大魁北克省蒙特利尔（1998年8月）。https: //doi.org/10.3115/980845.980860, https://aclanthology.org/P98-1013/ 7. Bascur, J.P., Verberne, S., van Eck, N.J., Waltman：利用引文簇进行学术信息检索：基于系统综述的深入评估。《科学计量学》**128**(5)，2895–2921（2023年5月）。https://doi.org/10.1007/s111 92-023-04681-x, https://doi.org/10.1007/s11192-023-04681-x 8. Bassignana, E., Plank, B.：CrossRE：用于关系抽取的跨领域数据集。载于：Goldberg, Y., Kozareva, Z., Zhang, Y.（编）《计算语言学协会研究发现：EMNLP 2022》。第3592–3604页。计算语言学协会，阿布扎比，阿拉伯联合酋长国（2022年12月）。https://doi.org/10.18653/v1/2022.findings-emnlp.263, https: // aclanthology.org/2022.findings-emnlp.263/

9. Bayerl, P.S., Paul, K.I.：哪些因素决定人工标注中的编码者间一致性？一项元分析研究。《计算语言学》**37**(4)，699–725（2011年12月）。https://doi.org/10.1162/COLI_a_00074, https://aclanthology.o rg/J11-4004/ 10. Beltagy, I., Lo, K., Cohan, A.：SciBERT：面向科学文本的预训练语言模型。载于：Inui, K., Jiang, J., Ng, V., Wan, X.（编）《2019年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》。第3615–3620页。计算语言学协会，中国香港（2019年11月）。https://doi.org/10.18653/v1/D19-1371, https://aclanthology.org/D 19-1371/ 11. Berrebbi, D., Huynh, N., Balalau, O.：Graphcite：利用图嵌入对科学出版物中的引文意图进行分类。第779–783页。WWW ’22：2022年ACM网络会议，ACM（2022年4月）。https://doi.org/10.1145/3487553.3524 657, 3524657 12 10 1145 3487553 http://dx.doi.org/ . / . . Chen, Y., Liu, P., Zhong, M., Dou, Z.Y., Wang, D., Qiu, X., Huang, X.：CDEval-Summ：神经摘要系统跨数据集评估的实证研究。载于：Cohn, T., He, Y., Liu, Y.（编）《计算语言学协会研究发现：EMNLP 2020》。第3679–3691页。计算语言学协会，线上（2020年11月）。https://doi.org/10.18653/v1/2020.finding s-emnlp.329, https://aclanthology.org/2020.findings-emnlp.329/ 13. 科睿唯安：Web of Science。https://www.webofscience.com/（2025年）14. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.：科学出版物中引文意图分类的结构支架。载于：Burstein, J., Doran, C., Solorio, T.（编）《2019年计算语言学协会北美分会会议论文集：人类语言技术》第1卷（长篇与短篇论文）。第3586–3596页。计算语言学协会，明尼苏达州明尼阿波利斯（2019年6月）。https://doi.org/10.18653/v1/ N19-1361, https://aclanthology.org/N19-1361/ 15. Cohen, A.M., Hersh, W.R., Peterson, K., Yen, P.Y.：利用自动引文分类减少系统综述准备工作量。《美国医学信息学会会刊》**13**(2)，206–219（2006年3月）。https: //doi.org/10 1929 16 1197 1929, 10 1197 . / jamia.m http://dx.doi.org/ . /jamia.m . Cohen, J.：名义量表的一致性系数。《教育与心理测量》**20**(1)，37–46（1960年4月）。https://doi.org/10.1177/00131644 6002000104, 001316446002000104 17 10 1177 http://dx.doi.org/ . / . Dardas, L.A., Sallam, M., Woodward, A., Sweis, N., Sweis, N., Sawair, F.A.：基于Semantic Scholar高影响力引文、总引文数和Altmetric关注度评分评估研究影响力：寻找精细化指标的努力仍未见效。《Publications》**11**(1)，5（2023年1月）。https://doi.org/10.3390/publicatio ns11010005, 11010005 18 10 3390 http://dx.doi.org/ . /publications . Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向变换器预训练。载于：Burstein, J., Doran, C., Solorio, T.（编）《2019年计算语言学协会北美分会会议论文集：人类语言技术》第1卷（长篇与短篇论文）。第4171–4186页。计算语言学协会，明尼苏达州明尼阿波利斯（2019年6月）。https: //doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/ 19. Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H., Sun, M.：OpenPrompt：用于提示学习的开源框架。载于：Basile, V., Kozareva, Z., Stajner,

S.（编）《计算语言学协会第60届年会论文集：系统演示》。第105–113页。计算语言学协会，爱尔兰都柏林（2022年5月）。https://doi.org/10.18653/v 1/2022.acl-demo.10，https://aclanthology.org/2022.acl-demo.10/ 20. Fillmore, C.J.：格位。语言学理论中的普遍性（1968年）21. Fillmore, C.J.等：框架语义学。认知语言学：基础读物**34**，373–400（2006年）22. Ghosal, T., Varanasi, K.K., Kordoni, V.：一种利用被引论文与引用论文关系的深度多任务方法，用于引文意图分类。《科学计量学》**129**(2)，767–783（2023年12月）。https://doi.org/10.1007/s11192-023-04811-5，http://dx.doi.org/10.1007/s11192-023-04811-5 23. Gilardi, F., Alizadeh, M., Kubli, M.：ChatGPT在文本标注任务中的表现优于众包工作者。《美国国家科学院院刊》**120**7 2023 30 10 1073 10 10( ) ( )。https://doi.org/ . /pnas.2305016120，http://dx.doi.org/ . 73/pnas.2305016120 24. Guo, M., Yang, Y., Cer, D., Shen, Q., Constant, N.：MultiReQA：检索式问答模型的跨领域评估。载于：Ben-David, E., Cohen, S., McDonald, R., Plank, B., Reichart, R., Rotman, G., Ziser, Y.（编）《第二届自然语言处理领域适应研讨会论文集》。第94–104页。计算语言学协会，乌克兰基辅（2021年4月），https://aclanthology .org/2021.adaptnlp-1.10/ 25. Hettiachchi, D., Sanderson, M., Goncalves, J., Hosio, S., Kazai, G., Lease, M., Schaekermann, M., Yilmaz, E.：调查并缓解众包数据中的偏见。第331–334页。CSCW '21：计算机支持的协同工作与社会计算，ACM（2021年10月）。https://doi.org/10.1145/3462204.3481729，http://dx.doi.org/10.1145/3462204.3481729 26. Ivanova, R.V., Kirrane, S., van Erp, M.：比较用于英语文学命名实体识别的标注数据集。载于：Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis（编）《第13届语言资源与评估会议论文集》。第3788–3797页。欧洲语言资源协会，法国马赛（2022年6月），https://aclanthology.org /2022.lrec-1.404/ 27. JHA, R., JBARA, A.A., QAZVINIAN, V., RADEV, D.R.：用于科学计量学的自然语言处理驱动的引文分析。《自然语言工程》**23**(1)，93–130（2016年1月）。https://doi.org/10.1017/s1351324915000443，http://dx.doi.org /10.1017/S1351324915000443 28. Jurgens, D., Kumar, S., Hoover, R., McFarland, D., Jurafsky, D.：通过引文框架衡量科学领域的演变。《计算语言学协会汇刊》**6**，391–406（2018年）。https://doi.org/10.116 2/tacl_a_00028，https://aclanthology.org/Q18-1028/ 29. Koloveas, P., Chatzopoulos, S., Vergoulis, T., Tryfonopoulos, C.：大语言模型能否预测引文意图？对开放式大语言模型进行上下文学习与微调的实验分析（2025年），https://arxiv.org/abs/2502.14561 30. Kunnath, S.N., Pride, D., Knoth, P.：引文分类的提示策略。载于：《第32届美国计算机协会信息与知识管理国际会议论文集》。第1127–1137页。CIKM '23，美国计算机协会，纽约，美国（2023年）。https://doi.org/10.1145/3583780.36 15018，https://doi.org/10.1145/3583780.3615018

31. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.：使用PagedAttention实现大语言模型服务的高效内存管理。载于：《ACM SIGOPS 第29届操作系统原理研讨会论文集》（2023年）32. Lahiri, A., Sanyal, D.K., Mukherjee, I.：CitePrompt：使用提示词识别科学论文中的引文意图。第51–55页。2023年ACM/IEEE数字图书馆联合会议（JCDL），IEEE（2023年6月）。https://doi.org/10.1109/jcdl57 899 00017 33 2023 00017, 10 1109 57899 2023 . . http://dx.doi.org/ . /JCDL . . . Li, K., Rubungo, A.N., Lei, X., Persaud, D., Choudhary, K., DeCost, B., Dieng, A.B., Hattrick-Simpers, J.：探究机器学习在材料领域的分布外泛化（2024 06489 34 2406），https://arxiv.org/abs/ . . Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.：Microsoft COCO：上下文中的常见物体。载于：《计算机科学讲义》，第740–755页。Springer国际出版公司（2014年）。https: //doi.org/10.1007/978-3-319-10602-1_48, http://dx.doi.org/10.1007/978 -3-319-10602-1_4835. Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.：S2ORC：Semantic Scholar开放研究语料库。载于：Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.（编）《计算语言学协会第58届年会论文集》。第4969–4983页。计算语言学协会，线上（2020年7月）。https://doi.org/10.18653/v1/2020.acl-main.447, https://aclanthology.org/2020.acl-main.447/ 36. Meng, R., Lu, W., huan Chi, Y., Han, S.：利用新语言特征自动分类引文功能（2017年），http://memray.me/uploads/iconfere nce17-citation-function.pdf 37. Mercier, D., Rizvi, S., Rajashekar, V., Dengel, A., Ahmed, S.：Impactcite：一种基于XLNet、利用情感与意图进行定性引文影响分析的解决方案。载于：《第13届国际智能体与人工智能会议论文集》。第159–168页。SCITEPRESS科学技术出版社（2021年）。https://doi.org/10.5220/0010235201590168，http: //dx.doi.org/10 0010235201590168 38 5220 3 . / . Mistral AI团队：Mistral Small。https://mistral.ai/news/mistral-small-3/（2025年）39. Nambanoor Kunnath, S., Pride, D., Knoth, P.：用于引文分类的动态背景提取。载于：He, Y., Ji, H., Li, S., Liu, Y., Chang, C.H.（编）《计算语言学协会亚太分会第2届会议暨第12届国际自然语言处理联合会议论文集（第1卷：长篇论文）》。第539–549页。计算语言学协会，仅线上（2022年11月）。https://doi.org/10.18653 /v1/2022.aacl-main.41, https://aclanthology.org/2022.aacl-main.41/ 40. Nambanoor Kunnath, S., Stauber, V., Wu, R., Pride, D., Botev, V., Knoth, P.：ACT2：用于引文重要性与目的分类的多学科半结构化数据集。载于：Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., Piperidis（编）《第13届语言资源与评估会议论文集》。第3398–3406页。欧洲语言资源协会，法国马赛（2022年6月），https://aclanthology.org/2022.lrec-1 .363/ 41. Nicholson, J.M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N.P., Grabitz, P., Rife, S.C.：Scite：一种展示背景的智能引文索引

……引文，并使用深度学习对其意图进行分类。《定量科学研究》**2**(3), 882–898 (2021)。https://doi.org/10.1162/qss_a_00146, http: //dx.doi.org/ 10 1162 _ _00146 . /qss a 。Northcutt, C., Jiang, L., Chuang, 42 I.: Confident learning：估计数据集标签中的不确定性。《人工智能研究杂志》**70**, 1373–1411 (4 2021)。https://doi.org/10.1613/jair.1.12125, http://dx.doi.org/10.1613/jair. 1.12125 43. Northcutt, C.G., Athalye, A., Mueller, J.: 测试集中普遍存在的标签错误会破坏机器学习基准。在：第三十五届神经信息处理系统会议数据集与基准赛道（第一轮）（2021），https://openreview.net/forum?id= XccDXrDNLek 44. Palmer, M., Gildea, D., Kingsbury, P.:《命题库》：语义角色标注语料库。《计算语言学》**31**(1), 71–106 (2005)。https://doi. org/10.1162/0891201053630264, https://aclanthology.org/J05-1004/ 45. Pride, D., Knoth, P.: 一种权威的引文分类方法。在：2020年ACM/IEEE数字图书馆联合会议论文集。第337–340页。JCDL '20，美国计算机协会，美国纽约州纽约市（2020）。 https://doi.org/10.1145/3383583.3398617, https://doi.org/10.114 5/3383583.3398617 46. Pride, D., Knoth, P.: 一种权威的引文分类方法。JCDL '20：2020年ACM/IEEE数字图书馆联合会议，ACM（8 2020）。 https://doi.org/10.1145/3383583.3398617, http://dx.doi.org/10.1145/338 3583.3398617 47. Qazvinian, V., Radev, D.R.: 利用引文摘要网络生成科学论文摘要。在：Scott, D., Uszkoreit, H.（编）第22届国际计算语言学会议（Coling 2008）论文集。第689–696页。Coling 2008组委会，英国曼彻斯特（2008年8月），https: // aclanthology.org/C08-1087/ 48 15115 49 2 5 2025 2412 。通义千问：通义千问技术报告（），https://arxiv.org/abs/ . . Riabi, A., Mouilleron, V., Mahamdi, M., Antoun, W., Seddah, D.: 超越数据集创建：对标注差异的批判性审视，以及对在线激进内容检测数据集的偏差探测。在：Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S.（编）第31届国际计算语言学会议论文集。第8640–8663页。计算语言学协会，阿布扎比，阿联酋（2025年1月），https://aclanthology . org/2025.coling-main.578/ 50. Roman, M., Shahid, A., Khan, S., Koubaa, A., Yu, L.: 使用词嵌入进行引文意图分类。《IEEE Access》**9**, 9982–9995 (2021)。 https://doi.org/10.1 109 3050547 51 2021 3050547, 10 1109 2021 /access. . http://dx.doi.org/ . /ACCESS. . . Rücker, S., Akbik, A.: CleanCoNLL：一个近乎无噪声的命名实体识别数据集。在：Bouamor, H., Pino, J., Bali, K.（编）2023年自然语言处理实证方法会议论文集。第8628–8645页。计算语言学协会，新加坡（2023年12月）。 https: //doi.org/10.18653/v1/2023.emnlp-main.533, https://aclanthology.org/2 023.emnlp-main.533/ 52. Shotton, D.: CiTO，引文类型本体。《生物医学语义学杂志》**1**(增刊1), S6 (2010)。https://doi.org/10.1186/2041-1480-1-s1-s6, http: // dx.doi.org/10 6 53 1186 2041 1480 1 1 . / - - -S -S 。Shui, Z., Karypis, P., Karls, D.S., Wen, M., Manchanda, S., Tadmor, E.B., Karypis, G.: 在多个数据集上微调语言模型，以进行引文意图分类。在：

Al-Onaizan, Y., Bansal, M., Chen, Y.N.（编）计算语言学协会研究发现：E MNLP 2024。第16718–16732页。计算语言学协会，美国佛罗里达州迈阿密（2024年11月）。https://doi.org/10.18653/v1/2024.findings-emnlp.974, https: // aclanthology.org/2024.findings-emnlp.974/ 54. Singh, S., Yadav, A., Jain, J., Shi, H., Johnson, J., Desai, K.: 使用COCO对目标检测器进行基准测试：前进的新路径。在：Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G.（编）《计算机视觉——ECCV 2024》。第279–295页。施普林格·自然瑞士，尚贝里（2025）55. Spiegel-Rosing, I.: 科学研究：文献计量与内容分析。《科学社会研究》**7**(1), 97–113 (2 1977)。https://doi.org/10.1177/0306312777007001 11, 030631277700700111 56 10 1177 http://dx.doi.org/ . / . Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., Liu, H.: 用于数据标注与合成的大语言模型：综述。在：Al-Onaizan, Y., Bansal, M., Chen, Y.N.（编）2024年自然语言处理实证方法会议论文集。第930–957页。计算语言学协会，美国佛罗里达州迈阿密（2024年11月）。https://doi.org/10.18653/v1/2024.emnlp-main.54, https://aclanthology.o rg/2024.emnlp-main.54/ 57 19786 58 3 2025 2503 3 2024 2407 。Team, G.: Gemma技术报告（），https://arxiv.org/abs/ . . Team, L.: Llama模型家族（），https://arxiv.org/abs/ .21783 59. Teufel, S.: 论证分区：从科学文本中进行信息抽取（1999），https://www.cl.cam.ac.uk/~sht25/thesis/t1.pdf 60. Teufel, S., Siddharthan, A., Tidhar, D.: 引文功能的自动分类。在：Jurafsky, D., Gaussier, E.（编）2006年自然语言处理实证方法会议论文集。第103–110页。计算语言学协会，澳大利亚悉尼（2006年7月），https://aclanthology .org/W06-1613/ 61. Tjong Kim Sang, E.F., De Meulder, F.: CoNLL-2003 共享任务简介：语言无关的命名实体识别。在：HLT-NAACL 2003第七届自然语言学习会议论文集。第142–147页（2003），https://aclanthology.org/W03-0419/ 62. Tkachenko, U., Thyagarajan, A., Mueller, J.: ObjectLab：目标检测数据中错误标注图像的自动诊断（2023）。https://doi.org/10.48550/A RXIV.2309 00832 63 00832, 2309 。https://arxiv.org/abs/ . . Valenzuela, M., Ha, V., Etzioni, O.: 识别有意义的引文。在：AAAI研讨会：学术大数据。第15卷，第13页（2015）64. Yu, H., Liu, J., Zhang, X., Wu, J., Cui, P.: 分布外泛化评估综述（2024 01874 65 2403），https://arxiv.org/abs/ . . Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., Hu, X.: 以数据为中心的人工智能：综述。《ACM计算综述》**57**(5)（2025年1月）。https://doi.org/10 3711118 66 1145 3711118, 10 1145 . / https://doi.org/ . / . Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z.: LlamaFactory：统一且高效地微调 100+ 语言模型。在：Cao, Y., Feng, Y., Xiong, D.（编）计算语言学协会第62届年会论文集（第三卷：系统演示）。第400–410页。计算语言学协会，泰国曼谷（2024年8月）。https://doi.org/10.186 53/v1/2024.acl-demos.38, https://aclanthology.org/2024.acl-demos.38/ 67. Zimmermann, E., Szeto, J., Pasquero, J., Ratle, F.: 基准测试基准：MS-COCO有多可靠?（2023），https://arxiv.org/abs/2311.02709

A 整体模型性能

表 2 展示了在 ACL-ARC、SciCite 以及我们提出的 SOFT 框架上，对所评估模型整体性能的比较，涵盖域内（ACL-ARC测试集）和跨域（重新标注的ACT2测试集）场景。性能通过准确率和宏平均-F_1 分数进行衡量。

B 跨框架与模型的逐类型分析

为了更细致地理解模型行为和框架特征，我们使用雷达图将域内和跨域设置下的逐类别 F_1 分数进行可视化，如图 3 所示。每个图展示了特定模型在给定框架各类别上的性能，其中蓝色表示域内，红色表示跨域 F_1 分数。这些可视化结果可以直接比较各个类别的学习效果，以及这些类别上的性能如何迁移到不同域中。

在 ACL-ARC 和 SciCite 中，只有少数类型表现出稳定且可学习的性能。在 ACL-ARC 中，Uses（$F_1 = 0$.60-0.72）和 Future（$F_1 = 0$.70-0.80）能够在不同模型中被可靠地检测出来，而 Motivation（$F_1 = 0$.06-0.27）和 extends（$F_1 = 0$.17-0.24）等类型则无论模型类型如何，结果都普遍较差。在 SciCite 中，background 和 method 达到了中等水平的 F_1（最高分别为 0.83 和 0.69），但 ResultComparison 始终较弱（$F_1 \leq 0.47$，召回率为 < 0.5）。这些模式表明，传统框架中存在语义重叠和标签模糊性，尤其是在需要区分细微意图的任务中。

相比之下，SOFT 框架包含多种类型，这些类型在概念上定义明确，也易于通过计算方法学习。零样本大语言模型在使用（$F_1 = 0$.79-0.90）、提供背景（0.78-0.86）和产出资源（0.76-0.82）方面表现出色，而微调后的大语言模型（通义千问-small）在发现和产出资源方面优于所有其他模型（两者均为 $F_1 = 0.82$）。难度更高的类型（如已完成工作）在 F_1 分数上表现出较大的性能差距（大语言模型：0.37-0.57，通义千问-small：0.69，SciBERT：0.79），这表明在词汇变化较大时，针对领域进行自适应微调至关重要。对于 JustifyDesignChoice，零样本大语言模型 Llama（$F_1 = 0.72$）优于两种微调模型（通义千问-small：$F_1 = 0$.45，SciBERT：$F_1 = 0$.42），这说明成功与否取决于通用话语建模，而非针对特定标签的训练。

精确率与召回率的不对称性进一步体现了框架质量的差异。在 Sci-Cite 中，该方法呈现出高召回率（例如，零样本大语言模型：0.86–0.94）但精确率较低（例如，零样本大语言模型：0.4–0.53），这表明存在标签过度泛化。在 SOFT 中，包括论证设计选择和已完成的工作在内的若干类型则呈现相反的情况：精确率较高（例如，通义千问：0.79），但召回率较低（例如，Mistral：0.28），说明模型能够识别典型案例，却难以广泛覆盖。值得注意的是，指出研究空白（通义千问：精确率 0.81，召回率 0.68）和方法（SciBERT：精确率 0.64，召回率 0.76）实现了更好的平衡。这些对比证实，我们的 SOFT 框架不仅能生成更具区分度且更易学习的标签，还能使模型错误更易于解释，从而支持零样本应用和有针对性的微调。

C 计算资源与软件

我们采用了 CitePrompt [32] 和 CitationIntentOpenLLM [29]的原始超参数设置。所有实验均在单张 H100 GPU 上进行，包括微调和推断。对于大语言模型推断，我们使用了 vLLM 库 [31]。微调耗时约 5 GPU小时，推断总共额外耗时 4 GPU小时。

表 2：各框架下引文分类模型的域内与跨域性能（零样本："ZS"，微调："FT"）。

框架 分类器		域内（ACL-ARC）		跨域（ACT2）	
		准确率	宏平均 F_1	准确率	宏平均 F_1
ACL-ARC（6 种类型）	零样本 Llama	0.58	0.52	0.60	0.13 (75%↓)
	零样本 Mistral	0.54	0.49	0.59	0.16 (67%↓)
	零样本 Gemma	0.65	0.48	0.62	0.17 (65%↓)
	零样本 Qwen	0.48	0.45	**0.63**	0.23 (49%↓)
	微调 SciBERT	**0.66**	**0.55**	0.47	0.12 (78%↓)
	FT Qwen-Small	0.65	0.51	0.58	**0.49** (4%↓)
SciCite（3种类型）	零样本 Llama	0.72	0.63	0.80	0.37（54%↓）
	零样本 Mistral	0.67	0.59	0.81	0.41（31%↓）
	零样本 Gemma	0.70	0.63	0.80	0.36 (43%↓)
	零样本 Qwen	0.67	0.59	**0.84**	**0.45** (24%↓)
	微调 SciBERT	**0.77**	**0.69**	0.66	0.34 (51%↓)
	FT Qwen-Small	0.69	0.61	0.40	0.23 (63%↓)
SOFT: 被引内容 Type（3种类型）	零样本 Llama	0.73	0.67	0.75	0.55（18%↓）
	零样本 Mistral	0.67	0.62	**0.78**	0.46（26%↓）
	零样本 Gemma	0.74	0.68	0.73	0.58（15%↓）
	零样本 通义千问	0.70	0.67	0.77	0.60 (10%↓)
	微调 SciBERT	0.70	0.70	0.58	0.41 (41%↓)
	FT Qwen-Small	**0.79**	**0.78**	0.75	**0.64** (18%↓)
SOFT: 引文意图（7 种类型）	零样本 Llama	0.79	0.72	0.72	0.57 (21%↓)
	零样本 Mistral	0.74	0.71	0.69	0.57 (20%↓)
	零样本 Gemma	0.71	0.59	0.71	0.55（7%↓）
	零样本 Qwen	**0.81**	**0.75**	**0.82**	**0.64** (15%↓)
	微调 SciBERT	0.72	0.53	0.52	0.20 (62%↓)
	FT Qwen-Small	0.77	0.65	0.77	0.56 (14%↓)

颜色图例：≤ 降 > 'o 的降幅 F_1。

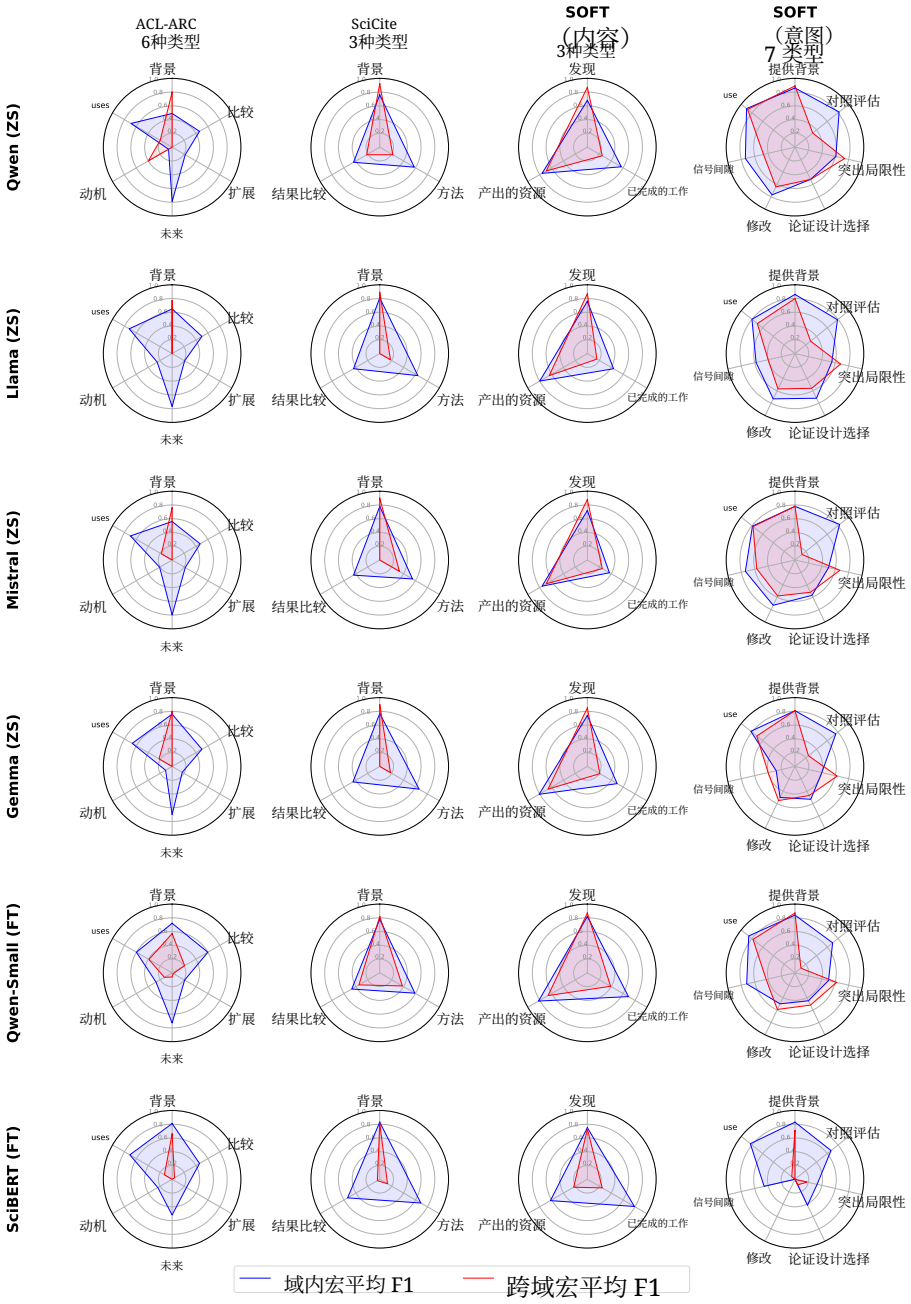


图 3：雷达图网格（4 个框架 × 6 模型），展示各类别的 F1 分数。蓝色线条/区域表示域内 F1 分数，红色线条/区域表示跨域 F1 分数。每个坐标轴对应相应框架中的一个类别标签。F1 分数范围为 0（中心）至 1（外缘）。