TaxoAdapt：将基于 LLM 的多维分类法构建调整为不断演变的研究语料库

Priyanka Kargupta[♣] **Nan Zhang**[♢] **Yunyi Zhang**[♣] **Rui Zhang**[♢]
Prasenjit Mitra[♢] **Jiawei Han**[♣]
[♣]University of Illinois at Urbana-Champaign [♢]The Pennsylvania State University
{ pk36,yzhan238,hanj } @illinois.edu
{ njz5124,rmz5227,pmitra } @psu.edu

Abstract

科学领域的快速发展在组织和检索科学文献时带来了挑战。虽然专家策划的分类法传统上满足了这一需求，但这一过程耗时且昂贵。此外，近期自动分类法构建方法要么（1）过度依赖特定语料库，牺牲了普适性，要么（2）高度依赖于其预训练数据集中包含的大型语言模型（LLMs）的通用知识，常常忽视了科学领域动态发展的性质。此外，这些方法未能考虑科学文献的多维性质，其中单个研究论文可能对多个维度做出贡献（例如，方法论、新任务、评估指标、基准）。为了解决这些差距，我们提出了 TaxoAdapt，一个能够根据多个维度动态适应 LLM 生成分类法到给定语料库的框架。TaxoAdapt 执行迭代分层分类，根据语料库的主题分布扩展分类法的宽度和深度。我们展示了其在多年来一组多样化的计算机科学会议上的最先进性能，以展示其结构和捕捉科学领域演变的能力。作为一种多维方法，TaxoAdapt 生成的分类法比由 LLMs 评价的最具竞争力的基线保留了 26.51 % 的粒度，同时比最具竞争力的基线更连贯性高 50.41 %。

1 介绍

由增加的研究兴趣和可访问性驱动，科学文献的快速扩散和随后的新知识分支的创建（例如，过去五年中生成模型的兴起）使得组织和检索特定领域知识变得越来越具有挑战性 (Bornmann et al., 2021; Aggarwal et al., 2022) 。分类法提升了数据组织，支持搜索引擎，捕获语义关系，并帮助发现。虽然专家策划和众包的分类法传统上将主题组织成层次结构（例如，文本分类 → 垃圾邮件检测），手动策划耗时且难以跟上快速发展的领域 (Bordea et al., 2016; Jurgens and Pilehvar, 2016) 。

在自动化分类构建 (ATC) 的先前努力中，主要可以分为两类：一种是直接从文本中提取主题和关系的方法，另一种是基于已有知识生成分类的方法。尽管基于语料库的方法可以有效捕捉有意义的、特定领域的主题，但其依赖于固定的方法，仅限于语料库词汇表中的术语，

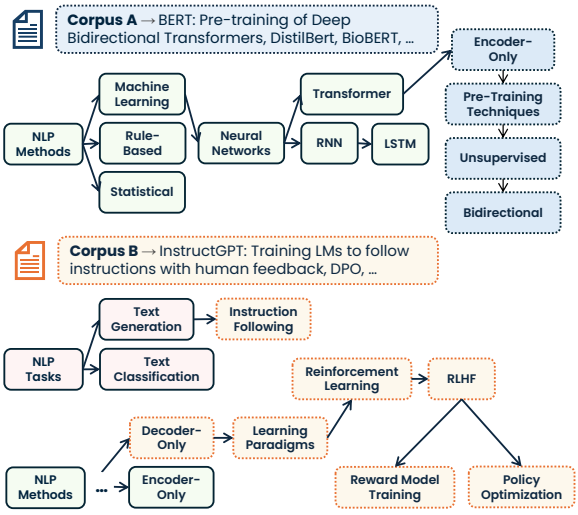


Figure 1: 语料库中的每篇论文都对科学文献的不同维度做出贡献。我们展示了不同时期的 NLP 语料库（例如，BERT 时代；RLHF 时代）如何影响其各自维度特定的分类法（我们强调了某些子树）。

缺乏广泛的背景知识，因为其起源于前 LLM 时代。相反，基于 LLM 的方法能够生成大规模的、通用的分类，但目前缺乏机制来使其与专业知识对齐，仅依赖于它们对领域及其关键主题的背景知识。

此外，截至目前，这两种方法都忽视了科学文献的多维性质。一篇研究论文可能会研究和/或贡献科学方法的多个方面（任务、方法、应用等等），基于这些我们可以不同地组织论文。当新知识出现时，我们必须适应现有的分类法。例如，在图 1 中，InstructGPT (Ouyang et al., 2022) 引入了 “Instruction Following” 作为一个新颖的 NLP 任务和 “Reinforcement Learning with Human Feedback”（RLHF）作为一种 NLP 方法，突出了单维分类法的局限性。将 ATC 设计限制于任务维度是一个严重的疏忽——遮蔽了研究更广泛、不断发展的影响。最终，语料库和基于 LLM 的方法都无法提供一个科学文献的多维视角。为了解决这些缺陷，我们提出了 TaxoAdapt，一个动态地根据多个维度将基

于 LLM 的分类法构建扎根于科学语料库的框架。TaxoAdapt 基于三个核心原则运作：

目前最先进的大型语言模型在精准建模计算机科学等领域的专用分类法（尤其是叶级实体）时存在困难。现有基于大型语言模型的方法需要预定义的实体集或仅限于实体级上下文进行分类法构建，严重限制了它们可以利用的领域特定知识的程度。作为替代，TaxoAdapt 利用文档级推理；通过每篇论文的标题和摘要，它识别出论文在哪些维度上有所贡献（例如，方法、数据集）以及贡献的方式。例如，如图 1 所示，在自然语言处理方法下扩展“Transformer”节点时，TaxoAdapt 有选择地分析以 Transformer 为基础的架构为中心的论文（例如，BERT），帮助得出“仅编码器”等子类别。与挖掘重要实体不同，这种基于文档的方法通过将扩展与针对每个维度、层次和节点的语料库知识对齐来提高分类精度。

Hierarchical text classification provides crucial signals for targeted exploration. 科学领域发展迅速，新兴的子领域不断涌现，现有领域则可能合并或消退 (Singh et al., 2022)。图 1 显示了这一趋势：语料库 A（2018–2022）强调类似 BERT 的编码器，而语料库 B（2022 至今）则突出“RLHF’’ 作为一种训练方法以及‘‘指令跟随’’ 作为 InstructGPT 及其后续技术背后的关键任务。由大型语言模型生成的分类结构往往会忽略这样的趋势，更倾向于广泛存在于训练数据中的概念（例如像文本分类这样的高层任务）。为了解决这个问题，TaxoAdapt 通过采用层次文本分类动态调整分类结构，以决定哪些节点需要扩展以及如何扩展。一个节点如果有大量的论文（例如，RLHF），则表明需要进一步探索并扩展深度（例如，奖励模型训练、策略优化）。相反，如果一个节点有许多未映射的论文（例如，如果‘‘仅解码器’’ 在“Transformer’’ 下不存在），这表明与现有子节点（例如，‘‘仅编码器’’）的研究平行，需要扩展宽度。在语料库中存在极少的节点（例如，LSTM）将不会被进一步探索。

Taxonomy-aware clustering enables meaningful expansion. 多个因素决定了哪些实体应该用于扩展给定节点：(1) 维护层级的、细粒度的关系（例如，识别维度特定的“Transformer”子节点以及“Encoder-Only”的同级节点），(2) 优先考虑在语料中的出现，以及 (3) 尽量减少冗余。最近，大型语言模型显示出强大的实体聚类能力 (Viswanathan et al., 2023; Zhang et al., 2023)。因此，TaxoAdapt 利用其对维度、层级及映射到特定被扩展节点的论文的知识，确定粒度一致的候选实体。然后利用这些信息来指导候选实体的聚类，在扩展过程中最大化覆盖率，同时尽量减少冗余。

总体来说，TaxoAdapt 将多维分类法的生成（和扩展）过程与一个语料库对齐。我们在下面总结我们的贡献：

- 据我们所知，TaxoAdapt 是第一个将基于 LLM 的分类构建应用于语料库的框架，并从多个维度研究这一任务。
- 我们提出了一种新颖的基于分类的扩展和聚类框架，用于针对性的、有意义的语料库探索。
- 通过定量实验和实际案例研究，我们表明 TaxoAdapt 在分类覆盖率、细粒度一致性以及对新兴研究趋势的适应性方面优于基线方法。

可重复性：我们的数据集和代码可在 https://github.com/pkargupta/taxoadapt 获取。

2 相关工作

分类法构建的先前研究大体可以分为三类：手工方法、语料库驱动方法和基于 LLM 的方法。

人工整理。 之前的工作 (Bordea et al., 2016; Jurgens and Pilehvar, 2016; Yang et al., 2013) 主要集中在从候选节点中提取手工制作的分类法或设计系统以支持创造人工辅助的分类法。这些分类法大多涉及手工操作，使得它们在创建过程中和未来维护时都变得昂贵，尤其是在科学领域快速发展的情况下。因此，非常需要自动化分类。

语料库驱动的方法。 一项研究路线 (Lu et al., 2024; Lee et al., 2022a,b; Zhang et al., 2018; Huang et al., 2020) 使用聚类技术从语料库中提取实体及其关系，通过识别语义上连贯的概念术语来完善给定的种子分类体系。或者，NetTaxo (Shang et al., 2020) 利用了语料库文档的元数据作为额外信号，从头开始构建分类体系。在不进行聚类的情况下，HiExpan (Shen et al., 2018) 采用关系提取模块进行深度扩展。尽管这些方法在语料库中特异性较高，但它们缺乏对大规模语言模型 (LLM) 的使用，限制了获取广泛背景知识的能力，而这种背景知识对于保持层次和细粒度的节点关系至关重要。

基于 LLM 的方法。 近年来，许多研究探讨了利用大型语言模型（LLM）进行分类扩展或构建的潜力。研究人员旨在回答 LLM 是否可以很好地替代传统的分类体系和知识图谱，他们发现 LLM 仍无法很好地捕捉分类体系和细节级实体的高度专业化知识 (Sun et al., 2024)。在 LLM 的使用方面，不依赖于任何数据进行显

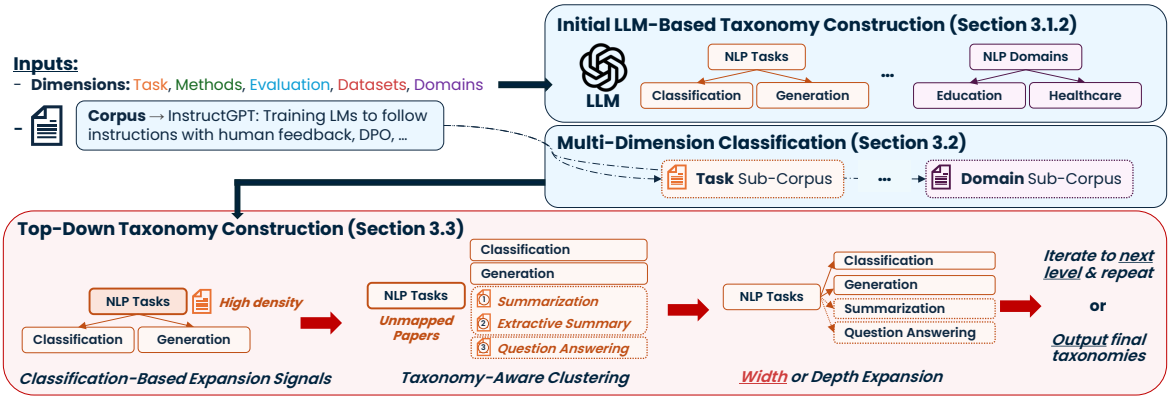


Figure 2: 我们提出了 TAXOADAPT ，一个框架，通过基于分类的扩展信号动态构建一个大语言模型增强的、语料库特定的分类法。该图示例展示了一个宽度扩展的例子，但相同的逻辑也适用于深度扩展（只是没有额外的兄弟上下文）。

式微调的提示方法优于基于微调的方法（Chen et al., 2023）。TaxoInstruct（Shen et al., 2024）通过释放 LLM 的指令追随能力，将三个相关任务（实体集扩展、分类扩展和种子引导的分类构建）统一起来。尽管提出了不同的迭代提示方法（Zeng et al., 2024; Gunn et al., 2024），据我们所知，目前还不存在一种基于 LLM 的方法能够很好地与不断发展的科学语料库保持一致。这强化了我们设计 TaxoAdapt 的动机。

3 方法论

如图 2 所示，TAXOADAPT 旨在将 LLM 的分类生成与特定语料库对齐，提高适应不断发展的研究语料库的能力。我们的框架协同利用 LLM 的一般知识和语料库特定知识，以自动构建更丰富和相关的分类体系。

3.1 预备知识

我们假设用户作为输入提供一个主题 t（例如，自然语言处理）、一组维度 D（例如，任务、数据集、方法、评估指标）和一个科学语料库 P。我们假设每篇论文 $p \in P$ 都与 t 相关，并研究至少一个 $d \in D$。TaxoAdapt 旨在输出一组 $|D|$ 分类 $T_{d \in D}$，在所有节点 $n_d \in T_d$ 上最大化映射的论文 $p \in P$ 数量。主题 t 和维度 $d \in D$ 形成每个分类 T_d 的根主题 n_0（例如，“自然语言处理任务”）。为了提供额外的灵活性，我们将每个分类定义为一个有向无环图（DAG），因为某些节点可能有两个父节点（例如，科学问答（QA）任务可以被放在 “question_answering” 和 “scientific_reasoning” 下）。

3.1.1 基于 LLM 的初始分类法构建

最近的研究 (Chen et al., 2023; Sun et al., 2024; Zeng et al., 2024; Shen et al., 2024) 探索了利用大型语言模型进行分类法构建，展示了它们在生成高级、通用分类法方面的潜力（尽管这些不保证能代表特定语料库）。由于跨多个领域获取专家策划的分类法的困难以及缺乏解决跨多个维度进行分类法构建的方法，我们利用大型语言模型生成 $|D|$ 初始单级分类法（$T_{d \in D}$），以供 TaxoAdapt 扩展。这让我们在最小化用户输入需求的同时展示了 TaxoAdapt 的有效性。尽管如此，该分类法也可以被任何用户希望的特定分类法替换。

3.1.2 分类扩展

分类扩展包括提供的分类 T_d 的深度和广度扩展。我们在下面正式定义这些：

Definition 1 (DEPTH EXPANSION) 通过识别一组子实体 $n^i_{j,d} \in N^i_d$ 来扩展叶节点 $n_{i,d} \in T_d$，这些子实体在主题上属于 $n_{i,d}$，并且包含同等粒度的实体（例如，$n^i_{1,d}$ 和 $n^i_{2,d}$ 应具有相同的主题特定性）。

Definition 2 (WIDTH EXPANSION) 展开一个非叶节点 $n_{i,d}$ 的子节点，其中现有的子节点 $n^i_{j,d} \in N^i_d$ 代表一个不完整的实体集合，需要通过额外的、独特的兄弟节点 $n'^i_d \in N'^i_d$ 进一步补充。N'^i_d 和 N^i_d 是不重叠的，并且处于相同的粒度水平。

注意，我们没有假设用户提供的实体集，这在历史上曾是 (Zeng et al., 2024; Shen et al., 2018) 的情况。

3.2 多维分类

科学文献本质上是多方面的，单个论文通常会对一个领域的多个方面做出贡献，如任务、方法论和数据集。因此，我们必须构建一组分类体系 $T_{d \in D}$，以捕捉科学知识的多样性。TaxoAdapt 试图将分类体系 T_d 的构建与文集中具有特定维度贡献的内容对齐。因此，我们

研究是否以及如何最小化来自没有对维度 d 做出任何贡献的论文的噪音。例如，一篇仅提出一个新的文本分类数据集但仍使用标准 F1-指标的论文会在构建“评估方法”分类体系时引入噪音，因此可能会被省略。为探索这一点，我们在进行分类扩展之前，根据每篇论文所贡献的维度对文集进行划分。

我们将这个任务视为一个多标签分类问题。最近的研究表明，LLMs 在多个领域的细粒度分类中取得了成功 (Zhang et al., 2024b,a) 。因此，我们提示 LLM 对论文 p 进行分类，在上下文中，我们提供了维度选项及其定义。我们根据我们期望论文 $p_{i,d}$ 做出的贡献类型来定义每个维度 $d \in D$ 。默认情况下，我们假设每篇论文始终属于任务维度。我们做出这个假设是因为每项工作都有一个与特定目标/任务对齐的贡献。最终，我们利用每篇论文 $p \in P$ 的输出标签，以便将语料库 P 划分为 $|D|$ 可潜在重叠的子集：$P_d \subseteq P$ 。我们在下面定义我们选择的每个维度：

> 厘米任务：我们假设所有论文都是与某个任务相关联的。方法论：一篇介绍、解释或改进一种方法或途径的论文，提供理论基础、实现细节和实证评估，以推进技术前沿或解决具体问题。数据集：介绍一个新的数据集，详细说明其创建、结构和预期用途，同时提供分析或基准以展示其相关性和实用性。重点在通过解决现有数据集的不足或 SOTA 模型的性能不足，或在该领域启用新的应用，以推动研究的进展。评价方法：一篇评估模型、方法或数据集性能、限制或偏差的论文，使用系统的实验或分析。其重点在于基准测试、比较研究，或提出新的评估指标或框架，以提供见解并提高领域内的理解。真实世界领域：本文展示了使用技术解决具体的真实世界问题或应对特定领域挑战的方法。它侧重于在各种环境中应用方法的实际实施、影响和所获得的见解。例子包括：产品推荐系统、病历总结等。

3.3 自上而下的分类法构建

由 LLM 生成的分类法可能无法充分捕捉语料库中的所有主题，尤其是在新兴研究领域中。这些领域在 LLM 的一般背景知识中代表性不足，但在输入语料库中则高度代表（例如，图 1 中的节点“RLHF”）。鉴于科学文献中的领域特定趋势在不断演变，我们必须确保基础研究全景的深度和广度都得到准确体现。

为了确定哪些节点需要更深入的探索，我们采用层次化分类。通过调整基于 LLM 的文本分类模型 (Zhang et al., 2024b) ，我们丰富了分类法节点（例如，通过添加关键词）以支持从 $n_{i,d}$ 到 $n^i_{j,d}$ 的自上而下的分类。具体而言，给定一个特定维度的论文 p 映射到 $n_{i,d}$ ，我们调整该模型以通过使用节点标签和描述的多标签分类来确定是否 p （基于其标题和摘要）映射到任何子节点 $n^i_{j,d} \in N^i_d$ 。我们定义 $n_{i,d}$ 的 density $\rho(n_{i,d})$ 为映射到其的论文数量，利用 $\rho(n_{i,d})$ 来决定其子节点（或缺乏子节点）是否应该扩展。

3.3.1 深度 & 宽度扩展信号

当许多论文堆积在给定的叶节点 $n_{i,d}$ 时，如 $\rho(n_{i,d})$ 的高值所示，这表明在语料库中对由 $n_{i,d}$ 代表的主题进行了更深入的探讨——这是当前分类法没有充分反映的。较长的分类路径表示语料库中流行的研究主题。图 1 对此进行了说明：“双向”路径明显比“基于规则”的路径更深，这反映了在语料库 A 中双向预训练语言模型的兴起以及基于规则方法的随后的衰退。在这种情况下，如果 $\rho(n_{i,d}) \geq \delta$ （用户指定的阈值），TaxoAdapt 通过识别一组子实体 N^i_d 来执行深度扩展（定义 1 ），将该主题划分为更细的、粒度一致的子主题。例如，如图 1 所示，如果 ρ(“ encoder-only ”) $\geq \delta$ ，这就需要进一步的分解，例如加深路径以包括“预训练技术”，以捕捉该领域正在进行的专业研究。

补充信号由非叶节点的未映射密度 $\tilde{\rho}(n_{i,d})$ 提供。这发生在一个节点 $n_{i,d}$ 有大量映射到该节点的论文（即高 $\rho(n_{i,d})$ ），而这些论文未分配到其任何现有的子节点 N^i_d 的情况下。

- **Definition 3 (UNMAPPED DENSITY)** 令 $P_{i,d}$ 表示与节点 $n_{i,d}$ 关联的所有论文的集合，$n_{j,d} \in N^i_d$ 表示节点 $n_{i,d}$ 下的子节点集合。未映射的密度则表示为：

$$\tilde{\rho}(n_{i,d}) = \left| P_{i,d} - \bigcup_{j=0}^{|N^i_d|} P_{j,d} \right| \tag{1}$$

如果 $\tilde{\rho}(n_{i,d})$ 超过预定义的阈值 δ ，这表明在 $n_{i,d}$ 中的语料库的重要部分没有被其当前的子节点充分表示。在这种情况下，TaxoAdapt 通过生成额外的、不重叠的兄弟节点 $n'^i_{j,d} \in N'^i_d$ 来启动宽度扩展，以覆盖代表不足的研究领域。例如，图 1 中的“仅解码器”节点，其中一个高 $\tilde{\rho}$(“NLP Methods”) 表示单个“仅编码器”节点未能充分捕捉到仅解码器架构的激增。一旦节点 $n_{i,d}$ 被触发以进行深度或宽度扩展，TaxoAdapt 将通过伪标签聚类程序（章节 **??** ）确定新的子实体集 N'^i_d 。

假设节点 $n_{i,d}$ 已被标记为展开，我们必须确定一组子实体（如果 $n_{i,d}$ 是叶节点则为 N'^i_d ，否则为 N^i_d ），这一组子实体需要满足以下标准：

1. 维持当前分类法中的层级、细粒度关系（父子关系和兄弟关系）。
2. 最大化存在于未映射论文集合 $\tilde{\rho}(n_{i,d})$ （宽度扩展）或 $\rho(n_{i,d})$ （深度扩展）。
3. 最小化子实体 $N^i_d \cup N'^i_d$ 之间的冗余。

Algorithm 1 自上而下的分类法扩展

Require: Topic t , Dimension $d \in D$, Corpus P , density_thresh = δ , max_depth= l

$T_d \in T$ = initialize_taxonomy(t, D) { T .depth $= 0$ }
$P_d \subseteq P \leftarrow$ multi_dim_class(t, D) {Section 3.2}
q = queue($\forall T_d \in T$)
while $len(q) > 0$ and T.depth $\leq l$ **do**
  $n_{i,d} \leftarrow pop(q)$
  if isLeaf($n_{i,d}$) **then**
    $n^i_{j,d} \in N^i_d \leftarrow$ expand_depth($n_{i,d}, t$) {Section **??**}
    q .append($n_{i,d}$)
  else
    classify_children($n_{i,d}, t, d$) {Section 3.3.1}
    if $\tilde{\rho}(n_{i,d}) > \delta$ **then**
      $n'^i_{j,d} \in N'^i_d \leftarrow$ expand_width($n_{i,d}, t$) {Section **??**}
      if $|N'^i_d| > 0$ **then**
        classify_children($n_{i,d}, t, d$)
      for $n^i_{j,d} \in N^i_d$ **do**
        if $n^i_{j,d}$ .level $< l$ **and** $\rho(n^i_{j,d}) > \delta$ **then**
          q .append($n^i_{j,d}$)
return T

为了维护分类法中的层次关系，我们利用 LLM 根据每篇论文的标题和摘要生成维度和粒度保持的一致伪标签。我们提示 LLM 确定其相对于 $n_{i,d}$ 的父级维度子主题（$n_{i,d}$ 的标签、维度、描述和祖先路径）以及 $n_{i,d}$ 的现有子级（如果有）的维度子主题。

Subtopic Clustering. 给定每篇论文都由其对应的伪标签表示，聚类这些伪标签可以让我们最大化所表示的论文数量（$\tilde{\rho}(n_{i,d})$ 或 $\rho(n_{i,d})$）。此外，有效的聚类本质上最小化了冗余，因为它旨在生成独特的、不重叠的论文集。我们特别利用 LLM 的聚类能力 (Viswanathan et al., 2023; Zhang et al., 2023) ，因为这使我们能够轻松将特定尺寸和粒度的信息整合到背景中，并在我们的聚类中保留这些特征。包括在子主题伪标签中提供的相同背景，以及完整的论文-子主题伪标签列表，我们提示 LLM 确定主要的子 [维度] 主题聚类（例如，子任务、子方法论），以最佳地涵盖伪标签列表，并为每个聚类提供标签和描述。如果 $n_{i,d}$ 是叶节点（深度扩展），则这些生成的聚类因此形成 N'^i_d ，否则形成 N^i_d （宽度扩展）。

我们通过迭代地逐级分类、识别扩展信号，并执行考虑分类结构的聚类来构建层级结构。我们在算法 1 中提供了完整的自上而下的分类构建算法。最终，当没有节点被标记为扩展或者达到最大分类深度时，这个过程结束——输出我们的最终 $T_d, \forall d \in D$ 。

4 实验设计

我们使用开源（`Llama-3.1-8B-Instruct` ）和闭源（`GPT-4o-mini` ）模型的结合来探索 TAXOADAPT 的性能。我们这样做是为了展示如何在不牺牲性能的情况下优化分类和伪标签步骤的成本（这两者在 `Llama` 上运行）。我们在附录 **??** 中讨论了我们的实验设置细节。

为了评估 TAXOADAPT 在不同语料库中的适应能力以及反映不断发展的研究主题，我们选择了几个跨越计算机科学不同子领域的会议。这些会议及其相应规模显示在表 1 中，在那里我们收集了每篇论文的标题和摘要。我们选择专门在计算机科学领域中探索我们的方法，以便我们的维度可以在所有会议中保持一致：任务、方法论、数据集、评估方法和现实世界领域。我们还包含了一个会议的两年不同年份（例如，EMNLP'22 和 EMNLP'24），以展示我们的方法如何反映其相应领域的发展。

Table 1: 每个数据集的主题 t 和论文数量（规模）。

Conference	Size	Topic t
EMNLP 2022	828	Natural Language Processing
EMNLP 2024	2954	
ICRA 2020	1000	Robotics
ICLR 2024	2260	Deep Learning
Total Papers	7,042	

4.1 基线

TaxoAdapt 将基于 LLM 的分类法构建与专业的、多维的语料库对齐。因此，我们选择将我们的方法与语料库驱动和基于 LLM 的方法进行比较。请注意，所有基于 LLM 的基线都使用 `GPT-4o-mini` 作为其底层模型。我们在附录 **??** 中提供有关每个基线的详细信息。

1. 全大语言模型 → 层级链 (Zeng et al., 2024) ：给定一组实体，仅依赖于一个大语言模型（没有语料库）来选择每个层级中相关的候选实体，并自上而下构建分类体系。
2. LLM + 语料库 → 基于提示：一个迭代基线，通过提示 LLM 识别与该维度、子节点及其对应论文相关的论文。
3. 仅使用语料的 → TaxoCom (Lee et al., 2022a) ：一种基于语料驱动的、人工构建的分类法完成框架，通过对输入语料中的术语进行聚类，以递归扩展人工构建的种子分类法。
4. "No-Dim" 和 "No-Clustering" 是 TaxoAdapt 的消融实验，分别去除了特定维度的划分和子主题聚类。

4.2 评价指标

我们设计了一个全面的自动评估套件，使用 `GPT-4o` 和 `GPT-4o-mini` 以确定我们生成的分

Table 2: 在所有数据集上进行模型比较，并对所有维度取平均值。所有数值都经过归一化处理，并乘以100。每个指标的最高分用粗体标出，第二高的分数用 † 标记。

Models	EMNLP'22					EMNLP'24				
	Path	Sib	Dim	Rel	Cover	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	46.87	67.67	94.61	77.65	50.54	49.56	67.67 †	92.56 †	82.13	48.66
With-Corpus LLM	66.14	33.93	88.82	72.87	39.35	49.51	29.74	83.56	84.13 †	39.20
TaxoCom	23.85	33.89	89.81	91.31	64.53	13.89	59.42	86.97	95.96	64.81
TaxoAdapt	81.09	82.92	100.00	82.69 †	55.81 †	83.04	77.86	98.04	88.76 †	60.29 †
- No Dim	88.47	82.30	99.49	81.46	62.26	89.98	76.97	99.05	86.23	66.42
- No Clustering	76.45 †	69.33 †	98.49 †	81.63	50.38	65.15 †	62.15	92.31	80.22	60.80
Models	ICRA'20					ICLR'24				
	Path	Sib	Dim	Rel	Cover	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	52.92	43.46	95.06	95.00	55.96	40.75	43.16	95.92	69.66	48.50
With-Corpus LLM	74.58	32.54	97.34	94.18	45.50	70.44	29.70	88.37	67.78	33.62
TaxoCom	43.05	54.21	99.06 †	96.28 †	60.75 †	30.00	67.00	91.27	86.88	56.25 †
TaxoAdapt	86.69	91.59	100.00	97.82	52.09	78.93 †	81.47	99.62 †	71.99 †	53.96
- No Dim	91.82	89.59 †	100.00	92.95	67.97	86.32	76.45 †	100.00	69.45	62.54
- No Clustering	87.74 †	85.76	100.00	93.97	50.86	65.69	67.85	93.13	68.56	54.60

类法的质量，利用节点级别和分类法级别的指标来评估。对于每一个判断，我们要求 LLM 提供额外的理由说明（所有提示在附录 **??** 中）：

- （节点级）路径粒度：通往节点 $n_{i,d}$ 的路径是否保留了其实体之间的层次关系（每个子节点 n^i_j 是否比父节点 $n_{i,d}$ 更具体）？由 `GPT-4o` 评分为 0 或 1。
- （层次级别）兄弟节点一致性：确定父节点 $n_{i,d}$ 的兄弟节点集合 $n_j \in N^i$ 是否形成了一个具有相同特异性和细度层次的连贯集合。由 `GPT-4o` 进行评分，从 0 到 1。
- （节点级别）维度对齐：节点 $n_{i,d}$ 是否与根主题 t 的维度 d 相关？由 `GPT-4o` 评分为 0 或 1。
- （节点相关性）论文相关性：节点 $n_{i,d}$ 是否与文集的至少 5 % 相关？每个节点由 `GPT-4o-mini` 评分为 0 或 1（由于论文上下文较长，因此成本较高）。最终得分在所有节点间平均。
- （层级覆盖）覆盖范围：给定一个父节点 $n_{i,d}$ 的兄弟节点集合 $n_j \in N^i$，确定 $n_{i,d}$ 的相关论文中有多少部分被至少一个兄弟节点覆盖（相关）。由 `GPT-4o-mini` 进行评分（由于论文上下文较长，因此成本较高）。

除了这个自动评估之外，我们还对这些评估指标进行了补充的人类评估。我们在附录 A 中提供了 LLM 与人类的一致性分析。我们还在附录 B 中提供了子主题伪标签和聚类步骤（第 **??** 节）的人类评估。

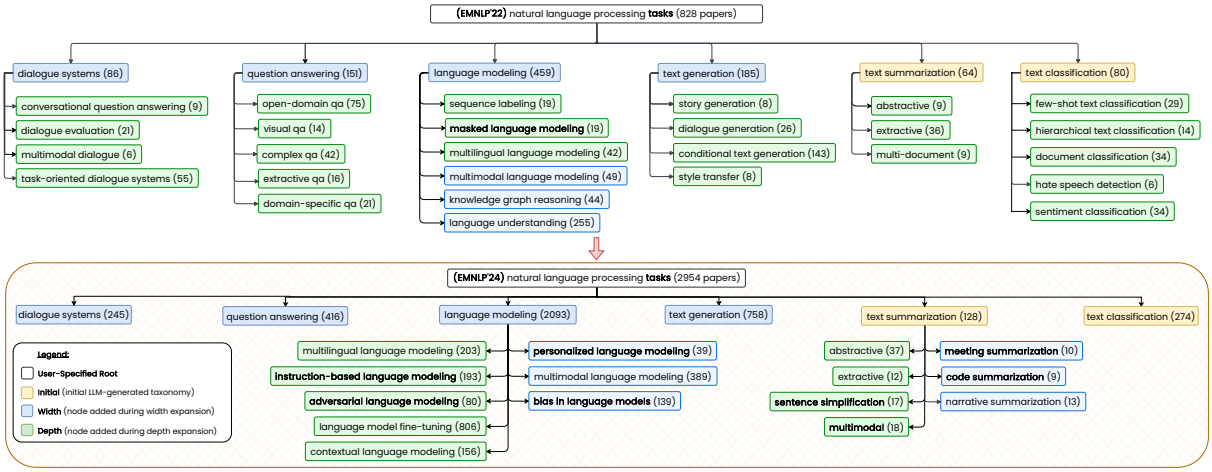
Table 3: 模型性能在所有数据集和维度上的标准差。

Models	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	0.078	0.109	0.008	0.043	0.005
With-Corpus LLM	0.054	0.036	0.010	0.027	0.004
TaxoCom	0.041	0.035	0.039	0.016	0.022
TaxoAdapt	0.027	0.021	0.007	0.043	0.015

5 实验结果

表 2 显示了 TAXOADAPT 在各种节点、级别和分类法指标上的表现与基线的比较。从结果中我们可以看到，与最具竞争力的基线相比，TaxoAdapt 的分类法在所有数据集和维度上都展现出更强的粒度保持性、更高的一致性、更具维度特异性、更相关于语料库，并且更能代表语料库。这些结果表明，TaxoAdapt 能在多个维度上显著更好地与语料库对齐，同时大幅提升构建分类法的结构完整性。基于我们一系列的实验，我们能够得出几个有趣的见解：我们观察到，基线方法倾向于生成明显不平衡的分类体系，其中多个节点仅有一个子节点。此外，每个层级往往具有不一致的粒度混合（例如，将 "情感分析" 和 "情绪检测" 作为兄弟节点）。特别是在 TaxoCom 中，这种情况尤为明显，它的路径粒度显著较低，但相关性和覆盖率分数最高。这是因为它选择了高度粗粒度的节点（例如，NLP 任务显著改进）。相比之下，TaxoAdapt 保持了分类学主题之间的层级关系，为每个非叶节点提供了融洽的子节点集，其中节点的子节点具有与相应论文集高度相关的

Figure 3: 我们展示了 NLP 任务从 EMNLP'22 到 EMNLP'24 的发展过程。由于篇幅限制，我们重点展示了特定的子树，突出了在 NLP 中具有普遍认知主题趋势的节点。我们还展示了 TaxoAdapt 映射到每个节点（章节 3.3 ）中的论文数量，括号中为这些节点的数字。



Models	EMNLP'22					ICRA'20				
	Path	Sib	Dim	Rel	Cover	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	46.87	67.67	94.61	77.65	50.54	52.92	43.46	95.06	95.00	55.96 †
With-Corpus LLM	66.14	33.93	88.82	72.87	39.35	74.58	32.54	97.34	94.18	45.50
TaxoCom	23.85	33.89	89.81	91.31	64.53	43.05	54.21	99.06	96.28	60.75
TaxoAdapt (+)	81.09	82.92	100.00	82.69	55.81 †	86.69 †	91.59 †	100.00	97.82 †	52.09
TaxoAdapt	69.92 †	74.33 †	98.70 †	88.69 †	51.95	92.08	95.11	100.00	98.20	49.10

Table 4: 对比模型在 EMNLP'22 和 ICRA'20 数据集上的表现。

高覆盖度。此外，每个子节点至少与文集中一定数量的论文相关，反映在表中的路径粒度增加、兄弟节点的凝聚力和覆盖率得分。我们可以将这些收益归因于 TaxoAdapt 的层级分类和基于消融表现较差的分类体系感知聚类步骤。我们还注意到，TaxoAdapt 主要使用作为其分类和聚类的主干模型，这比基线完全依赖的模型要弱得多。

除每个 TaxoAdapt 的节点 $n_{i,d} \in T_d$ 更好地反映其对应的维度 (Dim) 外，TaxoAdapt 对不同研究维度表现出稳健性。具体而言，表 3 展示了所有维度和数据集平均分数的标准差。我们观察到 TaxoAdapt 在所有粒度度量中表现出最低的标准差，同时在每个度量中得分最高（表 2 ）。我们通过消融实验 "No-Dim" 进一步探讨这一发现，该实验移除了语料库 P 初始的维度特定分区到 $P_{d \in D} \subset P$ （章节 3.2 ）。我们观察到，语料库的分区提高了粒度，但也对相关性和覆盖率产生了负面影响——只有缩小的、维度特定的池被认为对维度特定的分类构建是相关的。

TAXOADAPT 构建了反映 evolving research 的分类体系。在图 3 中，我们展示了 TaxoAdapt 的分类体系如何适应来自不同自然语言处理研究时代的语料库（EMNLP'22 → EMNLP'24）。我们展示了任务维度，鉴于 EMNLP 投稿和录用论文数量的快速增长，总体上特征更多节点（EMNLP'22：62 个节点；EMNLP'24：99 个节点）。此外，在两届会议年之间，我们看到某些节点的研究存在下降（例如，掩码语言建模），而其他节点显著上升（例如，语言建模、基于指令的语言模型、语言模型中的偏见）。我们还看到某些研究趋势开始出现，这归因于基于初步未映射论文的宽度扩展（例如，个性化语言模型）。总体而言，图 3 展示了考虑基于分类的信号进行知识增强扩展的强大能力。我们在附录 D 中包含了一个关于使用 EMNLP 数据集的现实领域数据集分类体系演化的额外案例研究。

如在第 4 节中所述，我们通过将某些任务指派给开源模型而非闭源模型来优化 TaxoAdapt 的成本：(1) `Llama-3.1-8B`：维度分类 + 分层分类信号 + 子主题伪标签；(2) `GPT-4o-mini`：初步/初始分类法构建（第 3.1.1 节；被视为我们核心框架的输入）和子主题聚类。因此，我们的核心框架大量使用开源模型 Llama-3.1 构建。

我们在表 4 中使用完全开源模型验证了我们方法在 EMNLP’22 和 ICRA’20 数据集上的性能。

通过 TaxoAdapt 使用仅一个 8B 开源模型的结果，我们可以看到，它在两个数据集上的表现仍然非常具有竞争力，甚至超过了我们主要的 TaxoAdapt 的 Llama-GPT 变体。这表明 TaxoAdapt 在不同的模型设置中非常 robust 。

附录 C 展示了一个案例研究和讨论，展示了我们基于语料库、具有分类法意识的框架在结合 LLM 的一般知识和语料库特定知识以生成更丰富和相关的分类法方面的强大能力。

附录 **??** 提供了关于 TaxoAdapt 在一个生物学数据集上的性能的额外定量研究，展示了即使在更专业的领域内，TaxoAdapt 仍能取得高性能。

6 结论

我们介绍了 TaxoAdapt，这是一种使用 LLM 构建与发展中的研究文献相一致的多维分类法的新框架。TaxoAdapt 能够动态地适应特定文献的趋势和研究维度。我们全面的实验表明，TaxoAdapt 在粒度保护、维度特异性和文献相关性方面显著优于现有方法。这些结果突出显示了 TaxoAdapt 作为一种可扩展的、多维的、动态适应的方法，在快速发展的领域中组织科学知识的能力。

7 局限性

TaxoAdapt 依赖于 LLMs 将论文分类到特定维度。虽然现有的研究已表明 LLMs 在细粒度分类上的成功，但这种分类依赖于 LLMs 的参数化知识，当 LLMs 的知识变得过时时，这可能成为一个限制。例如，当一个数据集论文提出一个与现有方法名称相同（或相似）的新基准时，LLMs 可能会错误地将其分配到方法学维度。然而，这是一个罕见的边缘情况，如上所述，TaxoAdapt 已经生成了比基线更多的维度特定分类法。

此分类法的潜在下游用例是帮助更好地检索 (Kang et al., 2024)，作为一个更具实验性的想法，可以利用 TaxoAdapt 的粗粒度和细粒度信号来告知基于 LLM 的研究助手当前领域的发展方向。这些用例依赖于我们方法的更专业的调整（因此不在本文范围内），我们留待未来的工作来探索这些潜在的途径。

8 致谢

本项工作得到了国家科学基金会研究生研究奖学金的支持。该研究使用了由国家科学基金会（奖项编号 OAC 2320345）和伊利诺伊州支持的 DeltaAI 高级计算和数据资源。DeltaAI 是伊利诺伊大学厄巴纳-香槟分校及其国家超级计算应用中心的联合项目。

References

Karan Aggarwal, Maad M Mijwil, Abdel-Hameed Al-Mistarehi, Safwan Alomari, Murat Gök, Anas M Zein Alaabdin, Safaa H Abdulrhman, et al. 2022. Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, 3(1):115–123.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1081–1091.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 588–596. IEEE.

Michael Gunn, Dohyun Park, and Nidhish Kamath. 2024. Creating a fine grained entity type taxonomy using llms. *Preprint*, arXiv:2402.12557.

Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1928–1936.

David Jurgens and Mohammad Taher Pilehvar. 2016. Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 1092–1102.

SeongKu Kang, Yunyi Zhang, Pengcheng Jiang, Dongha Lee, Jiawei Han, and Hwanjo Yu. 2024. Taxonomy-guided semantic indexing for academic paper search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7169–7184, Miami, Florida, USA. Association for Computational Linguistics.

Dongha Lee, Jiaming Shen, Seongku Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022a. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, WWW ’22, page

2819–2829, New York, NY, USA. Association for Computing Machinery.

Dongha Lee, Jiaming Shen, Seonghyeon Lee, Susik Yoon, Hwanjo Yu, and Jiawei Han. 2022b. Topic taxonomy expansion via hierarchy-aware topic phrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1687–1700, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. 2024. Self-supervised topic taxonomy discovery in the box embedding space. *Transactions of the Association for Computational Linguistics*, 12:1401–1416.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of the web conference 2020*, pages 1908–1919.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2180–2189.

Yanzhen Shen, Yu Zhang, Yunyi Zhang, and Jiawei Han. 2024. A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion. *arXiv preprint arXiv:2402.13405*.

Chakresh Kumar Singh, Emma Barme, Robert Ward, Liubov Tupikina, and Marc Santolini. 2022. Quantifying the rise and fall of scientific fields. *PloS one*, 17(6):e0270131.

Yushi Sun, Hao Xin, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2024. Are large language models a good replacement of taxonomies? *arXiv preprint arXiv:2406.11131*.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*.

Hui Yang, Alistair Willis, David Morse, and Anne de Roeck. 2013. Literature-driven curation for taxonomic name databases. In *Proceedings of the Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction*, pages 25–32, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024. Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3093–3102.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2701–2709.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024a. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2024b. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *arXiv preprint arXiv:2403.00165*.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. Clusterllm: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920.

我们使用开源（`Llama-3.1-8B-Instruct`）和闭源（`GPT-4o-mini`）模型的混合来探索TAXOADAPT的性能。我们这样做是为了展示如何在不牺牲性能的情况下优化分类和伪标签步骤（都运行在`Llama`上）的成本。我们使用`GPT-4o-mini`构建初始的、确定性的单层分类法（见第 3.1.1 节）。对于我们框架的所有其他模块，我们从前 1 % 的令牌中采样，并将温度设置为 0.1 。我们将密度阈值 δ 设为 40 篇文章，将最大深度设为 $l = 2$ 。假设根节点的深度为 0，并且由于任务的性质，分类法的大小有可能呈指数增长，特别是考虑到要插入的子节点数量是动态选择的。因此，我们将构建的分类法的最大层数设置为三层（$l = 2$）。对于 δ ，我们通过确定可以归入具有足够兴趣的细分类别的合理数量的论文来选择这个值（避免构建具有极其细分类主题的非常大的分类法）。我们故意不设置动态阈值，以便扩展也可以受到该领域发展的影响。

我们的主要动机是展示 TaxoAdapt 将基于大语言模型的分类法构建与特定的、多维的语料库对齐的能力。因此，我们选择将我们的方法与基于语料库和基于大语言模型的方法进行比较。请注意，所有基于大语言模型的基线都是以`GPT-4o-mini`为其基础模型。

1. 仅用 LLM → 层链 (Zeng et al., 2024)：这是一种方法，提供了一组实体，并完全依赖于一个 LLM（没有语料库）来为每个分类层选择相关的候选实体，并从上到下逐渐构建分类法。我们调整此方法以使用 LLM 根据根主题 t 和维度 d 推荐实体。

2. LLM + 语料库 → 基于提示：鉴于目前没有基于语料库指导 LLM 分类结构构建的方法，我们设计了自己的基于提示的基线。具体来说，我们进行一个迭代过程，首先要求 LLM 识别与维度相关的论文、相关子节点及其对应的论文。我们继续这一过程，直到达到最大深度。

3. 仅由语料库驱动的 → TaxoCom (Lee et al., 2022a)：一个语料库驱动的分类法完成框架，该框架从输入语料库中聚类术语，以递归扩展手工制作的种子分类法。我们使用第 3.1.1 节中相同的单级分类法作为种子输入，但如果标签名称在语料库中尚不存在，则将其修改为相似的概念。

Table 5: 在路径粒度、兄弟节点一致性、维度对齐和节点-论文相关性方面大型语言模型与人工评估者之间的一致性百分比。

Granularity	Coherence	Alignment	Relevance
0.900	0.700	0.700	0.875

A LLM-人类一致性分析

由于我们的自动评估套件主要使用 `GPT-4o` 和 `GPT-4o-mini`，我们进行了一项小规模的人类评估以测试我们指标的可靠性。利用 EMNLP'24，一名人类评估者负责验证在 TaxoAdapt 分类的任务维度上 LLMs 的评估输出。我们展示了关于路径粒度、同级结构一致性和维度对齐指标的共识百分比（LLM 和人类评估者在一个实例上达成一致的案例百分比），这些指标在第 4.2 节中定义。对于路径粒度，我们从 TaxoAdapt 的分类中随机选择 30 条路径，让评估者对实体之间的层次关系进行独立判断（由评估者评分为 0 或 1）。类似地，我们选择了 10 组相对于父节点的随机同级节点集，让评估者判断同级一致性（评估者对合理的一致性评分为 0.67 或对最强一致性评分为 1），并研究 30 个随机节点在任务维度上的对齐（评估者评分为 0 或 1）。对于（节点级）论文相关性和（层级）覆盖率指标，由于它们关于评估节点和论文的相关性，我们随机选择了 16 对节点和论文（其中 8 对被认为相关而另 8 对被认为不相关由 `GPT-4o-mini`）供评估者判断相关性以验证这两个指标。

一致性百分比显示在表 5 中。LLMs 与人工评估者之间的同意百分比范围从 70 % 到 90 %，表明整体的一致性很强。因此，这次人工评估加强了我们指标的有效性，所以我们决定使用它们作为我们的自动评估指标。

B 人工评估子主题伪标记 & 聚类

Table 6: 不同伪标签类型的对齐分数。

Pseudo-Label Type	Dimension	Paper
Width Expansion	0.8	0.8
Depth Expansion	0.85	0.75

我们进行了两次人工评估，以证明子主题伪标签化和子主题聚类（章节 **??**）的有效性。具体而言，对于伪标签化，我们定义了两个二元标准来验证 LLM 生成的伪标签：

1. 维度对齐：伪标签与分类法的整体维度对齐。

2. 论文对齐：伪标签与其对应论文的标题和摘要对齐。

我们从宽度扩展节点中选择 20 篇论文，从深度扩展节点中选择 20 篇论文。由于每篇论文都有一个伪标签，一位人工评估者统计有多少标签符合这些标准。满足每个标准的伪标签比例如表 6 所示。

很明显，大多数伪标签都与它们各自的维度和论文一致。这表明提示大型语言模型生成伪标签以保持我们分类法的粒度是有效的。

关于子主题聚类，每个聚类都有一个名称、一个描述和一组伪标签。我们定义了两个二元评价标准：

1. 相关性：一个簇名需要涵盖其大多数伪标签。

2. 一致性：一个聚类的所有伪标签在该聚类内需要有意义。

随机选择 20 个簇，我们的人类评估者计算满足我们标准的簇的数量。满足每个标准的簇的比例如表 8 所示：

这两种比例都表明使用 LLMs 来确定主题簇是有效的。我们观察到，连贯簇的比例低于簇名称相关性，这是因为我们对簇的连贯性设置了更严格的要求（所有伪标签都需要与簇的名称和描述一致）。

Biology Papers	Path	Sib	Dim	Rel	Cover
Chain-of-Layers	52.69	62.99	98.67	61.50	49.95
TaxoAdapt (+)	91.08	72.81	98.67	68.23	39.70

Table 7: 在生物学论文数据集上的性能比较。

Table 8: 基于名称相关性和一致性评估聚类质量。值表示满意聚类的比例。

	Relevance	Coherence
Proportion	0.95	0.7

C 关于 LLM 基础知识在分类法构建中作用的案例研究

我们工作的基本动机是如何将基于 LLM 的分类法构建适应于特定语料库，从而使过程有知识依据，并最终得到更高质量的分类法。因此，虽然任何利用 LLM 的方法都会受益于其一般知识，但我们表明，仅依靠 LLM 的一般知识对于我们的任务是不足够的。我们通过将我们的方法与 Chain-of-Layers（仅使用 LLM）以及 With-Corpus LLM（两者都在第 4.1 节和附录 **??** 中描述）进行比较来证明这一点，其中我们在所有指标上都取得了显著的性能提升——如表 2 所示。

TaxoAdapt 在所有指标上都优于 Chain-of-Layer，这表明仅使用 LLMs 是不够的。我们观察到 Chain-of-Layer 的路径细粒度得分非常低，这表明其分类体系从上到下的实体之间存在较差的层次关系。原因是 Chain-of-Layer 缺乏知识基础，无法理解细粒度的实体。尽管 Chain-of-Layer 被提供了语料库中存在的细粒度实体作为输入，但其细粒度性能仍然较差（这一点在下面的定性示例中也可以看到）。这表明它（GPT-4o-mini）的常识不足以理解这些细粒度实体之间的层次关系。相比之下，TaxoAdapt 使用较弱的基础模型（Llama-3.1-8B）且仅仅将语料库作为输入就显著优于它。

- 语言模型训练
  - 语言模型中的参数敏感性
  - 面向检索的语言模型预训练
    - RetroMAE
  - 高效的掩码语言模型训练
    - 通过基于概念的课程屏蔽高效预训练掩码语言模型
  - 意图检测框架
    - 多标签意图检测
  - 跨语言摘要数据集
    - EUR-Lex-Sum
  - 科学文档表示
    - 对比学习
    - 引用嵌入
    - 基于相似性的学习
    - 科学文献表示
  - 答案句子选择模型
    - 预训练 Transformer 模型

与包含语料库的 LLM 相比，TaxoAdapt 也实现了显著的改进，这表明即使将语料库特定的信息与 LLM 整合也是不够的。含语料库的 LLM 具有非常低的同级连贯性评分，这显示出其无法形成连贯的同级节点集。相反，通过我们基于分类法的伪标签 & 聚类（表 2 中的无聚类消融实验），TaxoAdapt 优于含语料库的 LLM。这展示了我们基于语料库驱动并且意识到分类法的框架的强大功能。

我们最初选择了基于计算机科学的论文，因为这个领域自然而然地在会议层面上组织了大规模的公开论文。然而，我们展示了 TaxoAdapt 在一个包含 1000 篇生物论文的数据集上的表现，并将其与整体上最具竞争力的基线，即链层（与论文正文相同的实验设置）进行比较。我们可以看到，尽管严重依赖于一个小型开源模型，TaxoAdapt 在大多数指标上仍然表现出显著的提升。我们注意到覆盖率得分较低，这是因为链层在整个分类法中生成了更多的粗粒度节点（因此它们的路径细粒度得分较低）。这表明，即使在更专业的领域内，TaxoAdapt 仍能取得高水平的性能。

D 自然语言处理实际领域演变的案例研究

在图 4 和 5 中，我们分别提供了 TaxoAdapt 为 EMNLP'22 和 EMNLP'24 的现实世界领域维度输出的最终分类法。我们看到，随着大型语言模型的兴起，研究人员能够以更广泛和深入的方式探索自然语言处理的现实应用。这体现在最初由大型语言模型生成的节点（例如医疗保健、电子商务）在 EMNLP'24 中得到了扩展（例如，医疗记录管理、临床决策支持、患者参与等）。此外，随着多模态模型的显著改进，我们看到更多的多模态研究。最后，我们在 2024 年看到一个新的突出的节点出现："自动事实

Figure 4: EMNLP'22 的自然语言处理现实世界领域输出分类法。

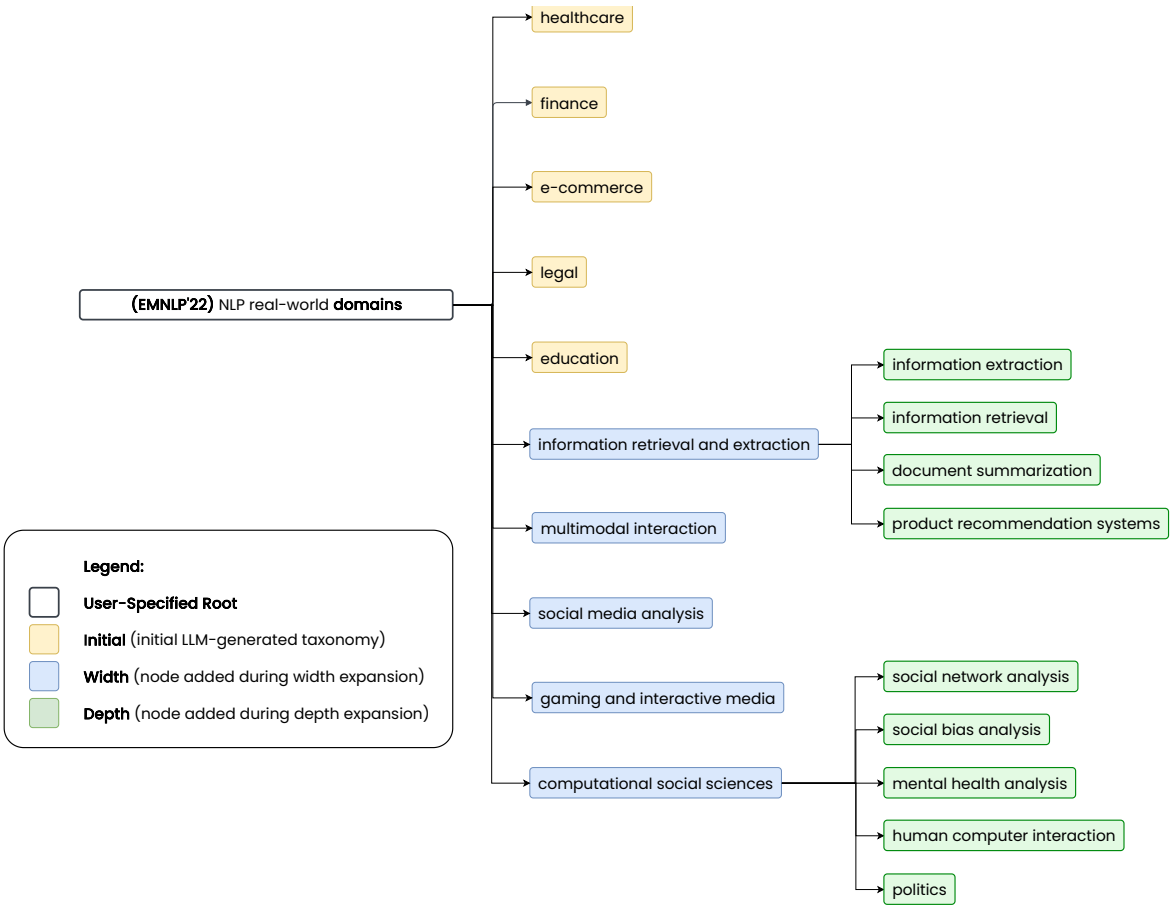
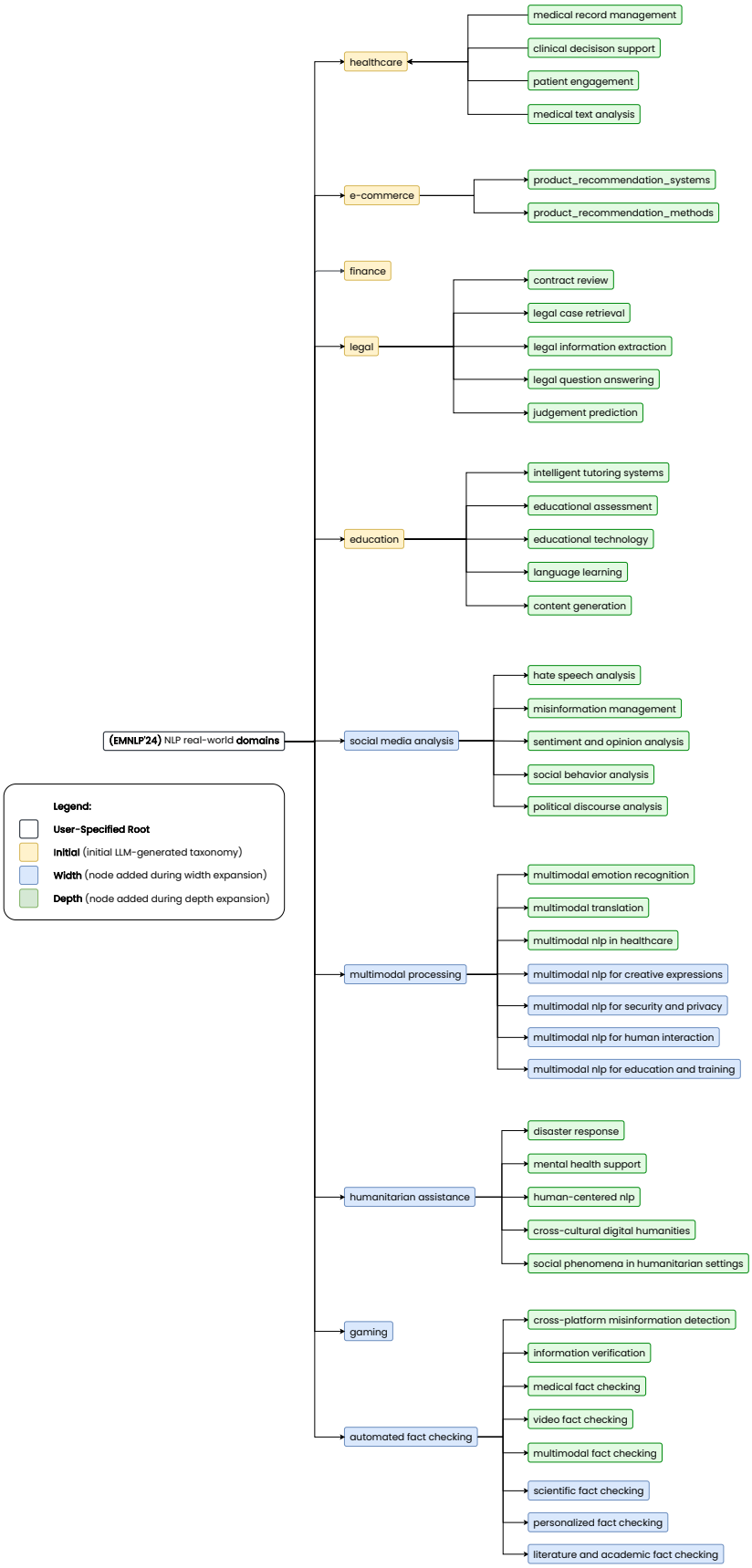


Figure 5: EMNLP'24 自然语言处理现实世界领域输出分类法。



核查”。这与大型语言模型幻觉作为一个主要公众关注点的兴起密切相关。总体而言，关于任务和现实世界领域维度的两个案例研究表明，TaxoAdapt 能够捕捉不断发展的研究文献。

如第 4.2 节所述，我们展示了用于生成评估输出以计算自动指标的 LLM 提示，如图 6 所示。

```
path_granularity = (
            "Scientific concepts are naturally organized in multi-dimensional taxonomic structures, with more specific concepts being the children of a
broader research topic.\n\n"
            f"Given the root topic: '{root}', decide whether this path from the scientific concept taxonomy has good granularity: '{path}' Check whether the
child node is a more specific subaspect of the parent node. \n\n"
            "Output options: '<good granularity>' or '<bad granularity>'. Do some simple rationalization before giving the output if possible."
        )


sibling_coherence = (
            f"You are determining the coherence of a set of {root} subtopics of the parent topic {parent}.\n\nThe parent topic is: {parent}.\n\nThe set of
siblings, which are child subtopics of the parent, are: '{', '.join(siblings)}'\n\nEvaluate the overall coherence of the sibling set based on their
collective specificity and granularity relative to {parent}. Use the following scoring criteria:\n\n"
            f"Score=<no_sibling_coherence>: The set is highly inconsistent or incoherent (only one subtopic), with most topics significantly misaligned in
specificity relative to the parent.\n"
            f"Score=<weak_sibling_coherence>: The set shows considerable inconsistency, with several topics deviating noticeably from the expected level of
specificity.\n"
            f"Score=<reasonable_sibling_coherence>: The set is generally coherent, with only minor inconsistencies in specificity among the topics.\n"
            f"Score=<strong_sibling_coherence>: The set is fully coherent, with all topics properly matching the expected level of specificity and
granularity for the parent.\n"
            "Output options: '<no_sibling_coherence>', '<weak_sibling_coherence>', '<reasonable_sibling_coherence>', or '<strong_sibling_coherence>'. Do some
simple rationalization before giving the output if possible."
        )

dimension_alignment = (
            "Scientific concepts are naturally organized in multi-dimensional taxonomic structures, with more specific concepts being the children of a
broader research topic.\n\n"
            f"Given the root topic: '{root}', decide whether this node from a taxonomy is relevant to the {dim} aspect of the root topic: '{node}'\n\n"
            "Output options: '<relevant>' or '<irerelevant>'. Do some simple rationalization before giving the output if possible."
        )

paper_relevance = ("Scientific concepts are naturally organized in a multi-dimensional taxonomic structure, with more specific concepts being the children of
a broader research topic.\n\n"
                f"Given the root topic: '{root}', here is one of its subtopics: {node_name} and these are some papers: {index_papers}\n\n"
                "Provide a list of paper IDs that are relevant to this subtopic.\n\n"
                "Output options: '<rel_paper> ID1, ID2, ... </rel_paper>'. Do some rationalization before outputting the list of relevant paper IDs.")
```

Figure 6: 用于计算路径粒度、兄弟一致性、维度对齐、论文相关性和覆盖率的 LLM 评估提示。