arXiv:2603.10876v1 [cs.CL] 11 Mar 2026

一个极端多标签文本分类（XMTC）图书馆数据集：如果我们认真对待“在数字图书馆中使用实用人工智能”呢?

Jennifer D'Souza[1], **Sameer Sadruddin**[1], **MaximilianKähler**[2], **Andrea Salfinger**[3], **Luca Zaccagna**[3], **Francesca Incitti**[3],**Lauro Snidaro**[3],**Osma Suominen**[4]1
TIB 莱布尼茨科学与技术信息中心，德国[2]德国国家图书馆，德国[3]乌迪内大学，意大利[4]芬兰国家图书馆，芬兰
{jennifer.dsouza,sameer.sadruddin}@tib.eu

摘要

主题标引对于发现至关重要，但要大规模开展并覆盖多种语言却难以持续。我们发布了一个大型双语（英语/德语）目录记录语料库，这些记录均以综合规范档（GND）进行标注，并配套提供机器可操作的 GND 分类体系。该资源支持本体感知多标签分类，将文本映射至规范术语，并通过可复现、基于规范数据的评估实现智能体辅助编目。我们提供了三个系统的简要统计概况和定性错误分析。我们邀请社区不仅评估准确性，还评估实用性与透明度，推动构建以规范数据为基础的人工智能副驾驶，从而增强编目员的工作能力。

关键词： 多语言主题标引，极端多标签分类，基准数据集

1. 引言

图书馆长期以来依靠专家进行主题标引，使馆藏易于发现、互操作且持久保存。然而，图书馆目录记录的数量迅速增长且日益多语言化，正不断加重纯人工标引工作流程的负担。与此同时，大语言模型（LLM）和新兴的智能体流水线有望提供支持——但它们必须以权威词表为依据、过程可审计，并依据图书馆领域的标准进行评估，而不是采用通用的文本分类评分。我们推出了一项机器学习就绪资源，直接填补这一空白：一个双语（英语/德语）多领域语料库，其中的目录记录均以德国综合规范文件（Gemeinsame Normdatei，GND）中的主题词进行标引；该资源同时提供机器可读版本的 GND 主题分类体系，以及预先划分的训练集/开发集/测试集。我们的目标不只是扩大规模，更是实现结构化扩展——让每项预测都关联到图书馆早已信赖的受控词表。

本资源旨在帮助学界探讨大语言模型时代图书馆学中至关重要的实践问题：自动化系统应如何在保留来源信息和规范控制的同时，将自由文本对齐到受控词表？在描述环节，什么才算是“有用”的辅助——前 k 项质量、层级一致性、可解释的理由，还是节省编目员的工作量？模型如何应对长尾主题、多语言差异，以及跨领域和跨时间的分布偏移？在人工参与的工作流程（分诊、建议、验证）中，智能体最适合发挥什么作用，以及我们应如何超越批量指标来衡量其影响？通过将实验锚定于一个操作性分类体系，该数据集支持对词汇对齐、跨语言一致性、多义性和变体标签，以及现实标签稀疏性下的可靠性开展研究——这些问题是通用的极端多标签文本分类基准仅能部分揭示的。

总体而言，我们的贡献是将真实目录记录与指向权威主题概念的稳定链接配对，并将其打包，以便进行可复现的评估。这使得本体感知的多标签分类、从自由文本到规范术语的检索增强映射，以及结合检索、建议和编目员反馈的智能体工作流成为可能；评估所采用的方案反映了编目工作的实际情况（例如，记录顶部主题词的实用性和层级一致性）。我们概述了这一资源及其构建方式和数据划分，并介绍初步分析与基线结果；同时，我们将本文定位为该数据集的**指南**：我们通过统计分析来探究数据集，揭示制定机器学习解决方案时应考虑的因素；最后，我们对基于该数据开发的三个系统进行定性错误分析，邀请 LREC 社区测试、比较并思考主题标引中成功且值得信赖的人工智能辅助应当是什么样。本资源以**TIB-SID**（TIB主题标引数据集）的形式发布，这是一个双语（英语/德语）、跨领域语料库，包含 13.6 万条标注了GND主题词的目录记录，可从以下地址获取：
https://github.com/sciknoworg/tib-sid
（CC BY 4.0许可证）。

本文其余部分的结构如下：第2节回顾相关工作，第3节介绍我们的数据集，第4节介绍在该数据集上训练的三个系统，并进行简要的定性分析。文章最后在第5节作出结论。

2. 相关工作

极端多标签文本分类（XMTC）数据集。 XMTC基准（Bhatia等人，2016）涉及将项目分配到规模极大的标签空间中，而这些标签空间具有高度偏斜的长尾分布（Zhang等人，2023年）。著名的例子包括*Wiki-500K*（180万条维基百科条目、50万个类别）和*AmazonCat-13K*（150万条产品描述、13,330个类别）。在这两者中，大多数标签都极为罕见（例如，Wiki-500K中只有2%的标签出现次数超过100次；AmazonCat中有30%的标签出现在少于10个样本中）（Yu等人，2022年；Zhang等人，2023年）。较小的特定领域语料库，如*EurLex-4K*（19,000份欧盟法律文件、4,000个EuroVoc主题词），也呈现出类似的长尾特征。这些数据集体现了大规模标签空间所带来的统计学挑战，但通常依赖通用类别或用户自定义类别。相比之下，我们的数据集表示的是图书馆记录上的细粒度、技术性专业主题标注，涵盖广泛的研究领域。

大型分类体系下的生物医学索引编制。 生物医学领域率先开展了大规模语义索引编制，使用的语料库包括 *BioASQ*(Tsatsaronis 等，2015年；Krithara 等，2023b），该语料库为数百万篇 PubMed 摘要添加了医学主题词表（MeSH）标注。MeSH 是包含约 30,000 个描述符的层级词表（NLM，2000年）。每篇文章通常分配有 12–13 个 MeSH 主题词，因此这项任务可视为极端多标签文本分类问题。多语言 *MESINESP* 语料库使用 DeCS 将这一方法扩展至西班牙语和葡萄牙语（Gasco 等，2021）。这些基准测试共同展示了受控领域分类体系如何支持大规模自动主题标引（Krithara 等，2023a）。

图书馆编目与多语言关联数据。 以往研究仅部分涉及多语言主题标引数据集，而公开且适用于机器学习的基准数据集十分稀缺。值得一提的工作包括 EHRI 多语言主题标引测试数据集——以 12 种语言、采用一个包含 900 个术语的词表标注的大屠杀档案描述（Dermentzi 等人，2025a、b）；联合国粮食及农业组织的AGRIS（1600 万条+记录，涵盖 123 种语言），使用多语言*AGROVOC* 叙词表进行标引（Caracciolo 等人，2013；Panoutsopoulos 和 Brewster，2022年）；以及泛欧文化遗产聚合平台 Europeana，它为记录补充指向多语言主题的链接，这些链接关联到虚拟国际规范档案（VIAF）、EuroVoc 或艺术与建筑叙词表（AAT）等受控词表，并可通过数据转储和 SPARQL 端点获取（Isaac 和 Haslhofer，2013；Angjeli 等人，2016）。国家级项目——大英图书馆以关联开放数据形式发布的《英国国家书目》（BNB）（Deliot，2014年），以及法国国家图书馆（BnF）的统一百科全书式字母序主题规范词表（RAMEAU）——同样公开了规模庞大的 SKOS/RDF 规范数据集。然而，这些资源通常缺少标准化的训练/测试划分和评估指标。

比较展望——我们的数据集。 与以往标签扁平或由用户定义的 XMTC 基准，以及缺少可直接用于机器学习划分的图书馆关联数据发布集不同，我们的语料库将 XMTC 规模的长尾分布与结构化的 GND 分类体系相结合，连接了大规模文本分类与知识组织。它提供双语、可直接用于机器学习的基准测试，包含预定义的训练/开发/测试划分和稳定的 GND链接，可用于本体感知建模、多语言评估和可复现比较。

3. 我们的主题标引数据集

本节介绍我们的主题标引数据集，它由两部分组成：分类体系（第 3.1 小节）以及使用该分类体系进行主题标引的图书馆记录（第 3.2 小节）。

3.1. 主题标引分类体系

作为我们的主题标引分类体系，我们采用GND (Gemeinsame Normdatei / 综合规范档)，这是一套由德国国家图书馆创建的综合规范档，被德语区图书馆广泛采用，用于编目和关联人物、组织、主题和作品等实体。在这些规范档中，我们特别使用主题术语 (主题词)。

读者可以按照我们的操作指南获取GND主题词[1]，其中介绍了如何下载最新的authorities-gnd-主题术语_dnbmarc.mrc.xml.gz文件。该文件采用MARC 21编码——这是一种使用数字标签的国际元数据标准（例如，021表示标识符，550表示相关主题）。尽管MARC 21在不同图书馆系统之间具有很高的互操作性，但它不具备人类可读性，因此我们将其转换为结构化的JSON格式[2]，使用我们的内部模式。生成的文件包含207,001个不重复的主题词，每个主题词都表示为一个标准化的JSON对象，其字段由MARC 21映射而来，包括GND标识符（代码）、分类号和名称、首选术语（名称）、变体标签（其他名称）、相关主题和来源，并可选择性地包含定义和来源网址。在考察属性覆盖率时，所有记录都包含核心字段，而上下文信息的完整程度则有所不同——其他名称约出现在一半的记录中，相关主题出现在80%的记录中，定义出现在27%的记录中。这一点很重要，因为分类体系中更丰富的上下文信息

[1] tib-sid/GND/how-to
[2] tib-sid/GND/subjects-taxonomy/GND-subjects.json

```
{
  "代码":     "gnd:4381803-1",
  "分类号":                          "19.3" ,
  "分类名称": "水文学,                                海洋学",
  "名称":     "总有机碳" ,
  "其他名称":  ["总有机碳","总有机碳"
       、"废水分析",
       指标"],
  "相关主题":  ["总量参数"],
  "来源":       "B 1986"
}
```

```
{
  "代码":     "gnd:4139622-4",
  "分类号":                          "27.7" ,
  "分类名称": "一般治疗",
  "名称":     "自然疗法"                    ,
  "其他名称":  ["生物疗法","
       自然疗法"],
  "相关主题":  ["经验医学","
       自然医学"],
  "来源":       "《医学实用词典》"
}
```

```
{
  "代码":     "gnd:4146660-3",
  "分类号":                          "21.4" ,
  "分类名称": "
       基本粒子,                    原子核-, Atom
       ", 分子物理学",
  "名称":     "溴-75",
  "其他名称":  ["Brom 75"],
  "相关主题":  ["溴同位素"],
  "来源": "Römpp（第9版）"
}
```

```
{
  "代码":     "gnd:7576879-3",
  "分类号":                          "18",
  "分类名称": "自然,                          自然科学一般",
  "名称": "科普利奖章",
  "其他名称":  ["科普利奖章","科普利奖章"],
  "相关主题":  ["自然科学","奖项,荣誉"],
  "来源":       "维基百科",
  "定义": "自 1731 年起每年由伦敦的英国皇家学会
       颁发,          以戈弗雷·科普利命名。"
}
```


图 1：以内部 JSON 表示形式呈现的四条 GND 示例记录。

能够改善主题消歧，从而提升下游主题标引的性能。

图 1 展示了 JSON 文件中的四个主题词示例：水文学中的 TOC（总有机碳）、医学中的 Naturheilverfahren（自然疗法）、物理学中的 Brom-75（溴同位素），以及自然科学中的 Copley-Medaille（科普利奖章）。每条记录都包含首选词及其变体（例如，TOC 列出了“总有机碳”，而 Naturheilverfahren [自然疗法] 列出了“生物疗法” [生物疗法] 和“Naturheilweise” [自然疗法]），以及相关主题，如 Summenparameter [总量参数] 或 Naturheilkunde [自然医学]。“来源”字段注明该词的出处（例如 B 1986、《医学实用词典》或维基百科），部分记录还包含简短的说明性定义，例如将科普利奖章描述为英国皇家学会每年颁发的一项科学奖。由此，结构化架构既保留了词汇信息，也保留了上下文信息。

我们还提供一个脚本[3]，用于将GND分类法导出为SKOS（简单知识组织系统）格式，这是万维网联盟用于以RDF表示受控词表的标准。

接下来，我们介绍基于GND标注的图书馆记录数据集。虽然这些记录以德语或英语呈现，但GND主题词本身主要以德语为主，偶尔会列出英语其他名称。

3.2. 我们的图书馆记录数据集

我们的图书馆记录语料库源自TIB—莱布尼茨科学与技术信息中心的开放数据集。在数据集构建时，TIB目录包含~5.7M条书目记录，且仍在持续增长；其中，TIB书目馆藏（TIBKAT数据）中有一份包含 ~200,000 条记录的开放转储，定期通过TIB开放数据门户发布（列于“转储形式的元数据”下，作为*TIBKAT*数据和免费电子馆藏的元数据）。该转储包含免费电子馆藏以及TIB视听门户（包括缩略图）的元数据，并以CC0 1.0许可发布，可不受限制地重复使用。为将这一异构且部分稀疏的原始转储处理成适用于机器学习的数据集，我们采用了四个预处理步骤：(i) 使用 `langdetect`[4], 进行语言识别，该工具在原始转储中检测出48种语言（前五种为：德语 108,637；英语 76,735；法语 1,741；印度尼西亚语 945；西班牙语 311）[5], ，随后仅保留占主导地位的两种语言——德语和英语；(ii) 删除没有摘要的记录；(iii) 剔除低频和/或不适用的记录类型（如期刊、章节）；(iv) 删除没有GND主题注释的记录，即缺少 `dcterms:subject` 标签的记录。

因此，清理后的馆藏包含 136,569 条记录，使其跻身规模最大的双语、多领域编目图书馆记录极端多标签文本分类语料库之列。按年代划分，该语料库几乎全部为现代记录： ~0.8k（20世纪70年代） → 3.2k（20世纪80年代） → 8.6k（20世纪90年代） → 33.2k（2000年代），在2010年代达到峰值 65.8k， ~24.6k 记录截至2020年代（截至2024年的部分数据），表明21世纪10年代中期增长迅速，近期增速有所放缓。

[3] tib-sid/GND/scripts/convert_to_skos.py

[4] https://pypi.org/project/langdetect/ [5] 原始转储中的语言标签经常缺失或含有噪声；因此我们使用了自动语言检测。部分记录仍为混合语言，但这种有限的噪声反映了真实世界的元数据，并不妨碍系统开发。

类型/划分	en	de	总计
文章	1,738 (1.27)	10 (0.01)	1,748 (1.28)
训练集	1,103 (0.81)	7 (0.01)	1,110 (0.81)
dev	212 (0.16)	2 (0.00)	214（0.16）
test	423（0.31）	1（0.00）	424（0.31）
Book	44,877（32.86）		100,581（73.65）
训练集	30,545 (22.37)	35,899 (26.29)	66,444 (48.65)
dev	7,143 (5.23)	7,466 (5.47)	14,609 (10.70)
test	7,189 (5.26)	12,339 (9.03)	19,528 (14.30)
会议	6,032 (4.42)	3,693 (2.70)	9,725 (7.12)
训练集	4,237 (3.10)	2,339 (1.71)	6,576 (4.82)
dev	1,015 (0.74)	484 (0.35)	1,499 (1.10)
test	780 (0.57)	870 (0.64)	1,650 (1.21)
报告	2,131 (1.56)	2,509 (1.84)	4,640 (3.40)
训练集	1,463 (1.07)	1,653 (1.21)	3,116 (2.28)
dev	348 (0.25)	372 (0.27)	720 (0.53)
test	320 (0.23)	484 (0.35)	804 (0.59)
学位论文	5,764 (4.22)	14,111 (10.33)	19,875 (14.55)
训练集	3,985 (2.92)	9,221 (6.75)	13,206 (9.67)
dev	954 (0.70)	1,953 (1.43)	2,907 (2.13)
test	825 (0.60)	2,937 (2.15)	3,762 (2.75)
总计	60,542 (44.34)		

表1：更新后数据集按划分、类型和语言统计的记录数量及相对占比（%）。蓝色：类型级占比；青色：划分级占比（总计列）；灰色：总计。单元格 ≥ 15%使用底纹标示。

表 1 按记录数量及其占馆藏总量的百分比，总结了我们发布的图书馆目录数据集。该数据集按照该图书馆编目的五种主要记录类型组织，分别为文章（1,748）、书籍（100,581）、会议（9,725）、报告（4,640）和学位论文（19,875）。另一种组织维度是记录的语言，分别为英语（en；60,542）和德语（de；76,027）。为支持主题标引方面的机器学习研究，我们发布了数据集，并预先划分出按记录类型和语言均衡的训练/开发/测试集（分别包含 90,452、19,949 和 26,168 条记录）。

我们数据集中的每条图书馆记录都以`json-ld`表示，它是关联数据的 JSON 序列化形式，可将书目元数据编码为机器可读三元组。大多数记录都包含标题和摘要，为机器学习系统进行主题标引提供了重要输入。不同记录类型的摘要长度平均为 100–150 个词元；英语摘要通常比德语摘要长。两种语言中，学位论文的摘要最长（通常为 $>$ 140 个词元），而报告的摘要最短（~90–110）。其他核心书目属性，如创作者（`dcterms:creator`）、贡献者（`dcterms:contributor`）、出版者（`dc:publisher`）、发行机构和出版地（`rda:P60163`），以及发行日期（`dcterms:issued`），均使用持久统一资源标识符表示，所用标准词汇包括都柏林核心、BIBO和RDA。每条记录的 GND 标注主题标签均记录在`dcterms:subject`下，并链接到 GND规范档中的受控概念标识符。例如，记录`dev/Conference/en/3A019447183.jsonld`描述了一篇由 Tim Bedford（dcterms:creator → gnd:171970268）撰写、由牛津大学出版社（dc:publisher）出版的会议论文，并链接到遍历理论（dcterms:subject→gnd:4015246-7）和双曲几何（gnd:4161041-6）等主题。类似地，记录 dev/Thesis/de/3A011101717.jsonld 代表 Reinhard O. Greil- ing 撰写的一篇博士论文（dcterms:creator → gnd:1090799454），由 Lang Verlag（dc:publisher）在美因河畔法兰克福（rda:P60163）出版，其主题链接包括黑色页岩（dcterms:subject→ gnd:4126782-5）和志留纪（gnd:4181434- 4）。这种表示方式支持在关联数据资源之间进行可互操作的书目与语义处理。

这些记录是如何编目的？ TIB 的 17 名主题专家团队负责覆盖预先设定的 28 个领域，并为 TIB国家图书馆目录中的记录添加主题标注，从而确保广泛且由专家策划的覆盖范围。领域分配通常采用半自动方式：新记录通常通过国家图书馆网络附带领域元数据；否则，则使用内部的AN-NIF 实例（Suominen，2019 年；Suominen 等人，2023年）。一条记录可能包含多个领域（范围为 1–7；平均值为 1.5，其中 7 是异常值）。其分布呈头部集中：社会科学（~24k）、经济学（~24k）和教育科学（~18k）占主导地位，同时科学、技术、工程和数学领域也有相当规模的覆盖（例如，计算机科学 ~13k；数学 ~12k），此外还有一个值得注意的“其他”类别（~5k）。专业领域则形成了长尾（例如，采矿 166；医学技术 589；体育科学 ~1.3k；材料科学 ~1.7k），这表明存在领域不平衡。

实际上，图书馆使用受控词表为内容编制索引；在德国，GND 用于主题编目。学科馆员根据标题、摘要以及——在可获取时——全文，选取 GND 术语；这项工作由各机构通过协作工作流程完成。这项工作由不同图书馆和不同国家图书馆网络以协作流程开展。考虑到持续增长（≈15每月新增 k 个编目标题），这项工作量相当大；自然语言处理/人工智能显然有潜力为此提供支持。在我们的数据集中，每条记录平均有三个 GND主题词；范围为 1–39（极高值很少见）。表2展示了五条示例记录及其领域和主题标注。

3.3. 主题标注的统计分析

本数据集取自一家真实的公共图书馆，反映了实际实践和长期质量约束，并受到不断演变的员工队伍和工作流程的影响。尽管如此，它仍为图书馆员开发可靠的人工智能工具提供了宝贵资源——这些工具旨在

记录（标题） （已链接）	Lang	Type	领域	主题词
第86章：解释比较静态分析 在阶梯式公共物品博弈中的应用	en	文章	经济学	实验；　　经济学； 经济学研究；方法论；实验—— 经济学研究
《万茨贝克的斧头》：小说；1938年– 1943	de	Book	文学研究	国家社会主义
《自由之窗：种族、公民权利， 与外交事务，1945—1988年	en	Book	社会科学；历- tory	外交政策；　种族歧视；　Men- 人权；黑人
保障家族影响力在家 族企业：研讨会……（2016年10月6日至7日 2016年）	de	会议	Law	影响；家族企业；家庭
电荷载流子复合与开路 电压在有机太阳能电池中：从 双层模型系统到混合多 结	en	学位论文	电气工程； 物理学	有机　　太阳能电池；　串联； 多层系统；氟；供体（化学）； 掺杂；复合

表2：图书馆记录数据集中的跨领域、多主题标注代表性示例。为简洁起见，仅展示标题元数据；链接指向完整的 JSON-LD 记录。

围绕语料库层面的模式和有原则的推理构建，而非针对个别特例。因此，本节将对主题空间进行统计刻画——量化各数据划分之间的重叠和长尾稀疏性，测量分布差异（库尔贝克-莱布勒散度、JSD 和 χ^2），并评估多义性，从而揭示其对模型设计和评估的影响。

3.3.1. 重叠与长尾现象　　on

在 GND 近 20 万个主题词中的主题术语里，有 41,218 个唯一主题出现在我们的数据集中（各划分计数发布于此处）。这些标注涵盖传统主题词（文学、建筑学、哲学）以及当代主题词（数字化、可持续性、人工智能）；有 6,164 个主题词出现在全部三个划分中，表明存在显著的长尾稀疏性：高频标签（例如文学、建筑学、企业）与专门主题词（例如机器人学、生物信息学、女权主义）并存，因而要求模型处理少样本与零样本泛化，而不能依赖均衡的头部类别。总体而言，该语料库贯通人文学科、社会科学与技术领域，但由长尾主导，这应指导训练与评估设计。杰卡德（1912）与加权杰卡德（谷本1958年）在表 3 中量化了划分重叠：杰卡德给出共享主题词所占比例（此处为 ≈ 0.36，即约三分之一），而加权变体还计入主题频率。较低的谷本值反映出，即便是共享主题词，其出现频率也不相同，这意味着模型必须处理稀疏标签与分布偏移。

指标	训练集–开发集	训练集–测试集	开发集–测试集
杰卡德	0.3687	0.3608	0.3668
谷本	0.2064	0.2530	0.3771

表 3：杰卡德和谷本分数，显示数据集划分之间的主题重叠情况。

3.3.2. 分布差异

为超越重叠度量，我们使用KL散度（Kullback和Leibler，1951）、詹森–香农散度（JSD；链接）以及卡方（χ^2）检验（Pearson，1900年）来量化各划分之间的频率散度。KL的范围在2–5纳特之间——KL(训练集|开发集)=4.34、KL(训练集|测试集)=4.69，相对于KL(开发集|训练集)=1.99、KL(测试集|训练集)=1.89——表明训练集覆盖更广的主题词集合，而开发集/测试集则是更窄的重加权。JSD显示中等程度的偏移（训练集–开发集为0.16，训练集–测试集为0.18，开发集–测试集为0.26）。 χ^2统计量为 $\approx$ 37k–49k，结合 $p<0.001$可证实这些差异是系统性的，而非随机波动。总体而言，该语料库呈现部分重叠，但存在明显的分布漂移。在方法上，模型必须处理不均衡的频率与代表性不足的主题——因而更宜采用正则化、校准以及偏移感知验证与报告。

3.3.3. 评估多义性

为评估潜在的多义性，我们将其操作化定义为不同 GND 标识符间共享相同字符串，并据此扫描分类体系。仅限首选标签时，此类情况极为罕见：在 207,001 个代码中仅出现 65 次（0.03%）。例如，阿拉卡卢夫人 既是 gnd:1071000497（族群，卡韦斯卡尔人）的首选标签，也是 gnd:4001008- 9（更广泛的民俗文化分类）的首选标签，这表明这是平行编目，而非真正的语义歧义。纳入替代标签后，出现次数增至 2,181 次（0.52%）：废水分析，指标 （“废水分析，指标”）涵盖七个条目——gnd:4145594-0（生化需氧量）、gnd:4147637-2（化学需氧量）、gnd:4185748- 3（总需氧量）、gnd:4360165-0（溶解性有机碳）、gnd:4381803-1（总有机碳）、gnd:4586442- 1（总磷（*ICP*））、gnd:4586443-3（总氮）——每个条目

均表示一种测量指标；类似地，*SPSS*、*WINDOWS*、程序也出现在五条记录中，分别对应不同的软件版本。这些发现表明，重复字符串主要反映了相关实体间的术语复用，而非真正的多义性；因此，替代标签虽能扩大检索范围，但在下游建模和评估中不应视为与首选标签可互换。

现在，我们不再局限于名称完全匹配，而是利用语义嵌入找出近重复项和潜在的义项混淆。在进行编码前，我们排除首字母缩略词/代码类标签（它们在分布相似性方面较为脆弱），剩下 $N = 203,763$ 个主题词，来自 $207,001$个主题词。我们比较了三类嵌入模型（表4）：Google 的 *EmbeddingGemma-300m*（紧凑型多语言通用表示）（Vera 等人，2025年）、*multilingual-E5-small*（对比式、面向检索）（Wang 等人，2024年）以及 *jina-v2-base-de*（侧重德语、高精度双语编码器）（Mohr 等人，2024年）。在所有情况下，我们都会构建一张图，其中余弦相似度超过0。90 的两个主题词之间以无向边相连；节点的度数 ≥ 2，即一个主题词节点若有两个以上相似词，是比孤立词对更明显的潜在多义性信号。

模型	视图	N(≥1)	N(≥2)	平均值	最大值	Gemma-名称
		∼19.8K	∼4.0K	0.15	83	300m 上下文
∼15.7K	∼3.6K	0.13	75	E5-名称	∼195.7K	∼188.1K
38.77	1.49K	小型上下文	∼165.8K	∼147.0K	24.62	
800	Jina-名称		∼15.0K	∼2.8K	0.11	36
v2-DE 上下文	∼14.5K	∼3.7K	0.13	53		

表4：余弦相似度 ≥ 0.90（topK$= 50$）下的语义相似度图，$N = 203,763$个主题词。N(≥1)：至少有一个相似邻居的主题词数量；N(≥2)：至少有两个相似邻居的主题词数量（更强的一词多义信号）。

表4中的结果显示，不同编码器之间存在明显差异。E5 生成的邻域非常稠密（名称视图：∼ 195.7K 个节点有邻居；平均度为 38.77；最大值为 1.49K），这有利于检索，但它往往会夸大多义性，因为它会将主题相关词和同词族变体归为一组，而非仅识别近重复项。具体案例清楚地体现了这一点：Vergrößerung 关联到 76 个条目，混杂了近义词（Erweiterung、Erhöhung）、反义词（Verkleinerung）、派生相关词（Verfilmung、Versilberung）和宽泛的相关词（Darstellung、Durchmesser）；Zeit 会牵出一长串复合词和语境（Zeitmessung、Zeitraum、Zeitalter、Zeitzone、Zeitarbeit，.....）；Differenzierung 的邻居则膨胀至 160 个，横跨数学、社会学和语言学。相比之下，Gemma 和 Jina 生成的图稀疏得多（平均度为 ≈ 0.11–0.15，最大值为 ≤ 83），更符合我们的目标：只标记非常相似的术语。它们建立的关联十分紧密：Gemma 将 Zukunft 与 → Fu- tur、Nutzung 与 →Benutzung/Nutzungseignung、Instrument 与 → (Musik)instrument 联系起来，偶尔也会受拼写或词素因素影响（Ver- größerung 与 → Vergröberung；Zeit 与 →Zeit-maß/Zeitmittel/Saatzeit）。Jina 在 0.90 的阈值下更加严格，主要找出明确的词对（Nutzung 与 → Be-nutzung；Differenzierung 与 → Dedifferenzierung；Instrument 与 → Musikinstrument/A-Instrument）；而许多高层级名词（Verhandlung、En- twicklung、Werk、Ziel）的度数为 0——如果目标是只标记可能存在的义项混淆，这种表现正合所需。对于表4中针对视图语境的各行，我们嵌入一个精简字符串，将首选标签与简短的定义线索及数量受限的备选项列表组合起来（名称 ⁅SEP⁆ DEF:...⁅SEP⁆ ALT:...）。对于 Gemma 和 E5，这通常能让含义更加明确，并去除表面上的、仅由名称产生的关联；而对于 Jina，这有时会将真正的德语释义表达整合为规模较小、联系更紧密的簇——这正是高精度多义性筛选器所期望的表现。总体而言，GND 中的多义性似乎很少见：使用较保守的编码器（Gemma/Jina）并采用严格的 0.90 余弦相似度阈值时，只有 ≈1–2% 的主题词参与多条目簇（度数 ≥ 2），而约 90–93% 完全没有近邻——这表明真正的歧义实属例外，而非普遍现象。

4. 三个系统

为说明如何利用该数据集进行自动化主题标引，我们报告了在 LLMs4Subjects 共享任务SemEval 2025（D'Souza 等，2025）和GermEval 2025中评估的三个代表性系统的结果，这两个任务都使用了该数据集。任务定义如下：给定一条由标题和摘要组成的记录，系统应从 GND 分类体系中检索最多 20 个相关主题词，并按置信度降序排列。这构成了一个带有排名输出的信息检索（信息检索）问题，评估范围覆盖前k 个预测主题词（Schütze 等，2008年）。作为主要评估指标，我们采用 nDCG@k（归一化折损累计增益）（Järvelin 和 Kekäläinen，2002），这是一种基于排名的度量，用于评估系统预测的主题词排名与真实值的吻合程度。不同于 recall@k 仅衡量前k 个预测中出现了多少个相关主题词，nDCG@k 还会考虑这些主题词的位置，通过对排名靠前的正确主题词给予对数折损奖励。该分数归一化到 0 至 1 之间，其中 1 表示完美排名，即所有相关主题词都位于列表顶部。

在实践中，这项任务通常使用极端多标签文本分类方法来解决。用于受控词表主题标引的极端多标签文本分类方法，通常将可扩展的候选生成（面向极大标签空间）与针对前-k个预测结果优化的排序阶段相结合。早期方法依赖稀疏的一对多线性分类器，例如DiSMEC （ Babbar 和 Schölkopf，2017 年）以及PD-Sparse/PPD-Sparse（Yen 等，2016，2017年）；这些方法为每个标签训练独立的线性模型，同时利用稀疏性和分布式优化，将系统扩展到极大规模的标签集合。FastXML（Prabhu 和 Varma，2014年）、Parabel（Prabhu 等，2018年）和 Bonsai（Khandagale 等，2020年）等基于树的方法，则通过将标签组织成层次化分区来提高效率，使推理阶段能够在预测时聚焦于规模小得多的候选子集。SLEEC（Bhatia 等，2015年）和 AnnexML（Tagami，2017年）等嵌入方法和近邻方法，则分别为实例和标签学习低维表示，将极端多标签文本分类重新表述为语义检索问题，在该问题中，通过近邻搜索找回相关标签。较新的神经网络方法，包括 XML-CNN（Liu 等，2017年）、AttentionXML（You 等，2019年）、X-Transformer（Chang 等，2020年）和 XR-Transformer（Zhang 等，2021），将深度文本编码器与标签筛选相结合，在保持可扩展性的同时改进上下文语义匹配。对于图书馆主题标引而言，可以将这些方法族理解为一种通用的检索与排序设计模式的不同变体：首先将受控词表缩小为合理的候选主题词，然后估计哪些标签最适合进行最终分配。PECOS（Yu 等，2022）等工具包通过在模块化的索引 → 匹配 →排序流水线中统一层次化标引、学习式匹配和排名，明确体现了这种联系。

基于这一任务表述和方法论版图，我们聚焦于LLMs4Subjects 共享任务中的三个代表性系统：LA^{2}I^{2}F（Salfinger 等人，2025年）、KIFSPrompt（Kähler 等人，2025年）和 Annif（Suominen 等人，2025年）。这三个系统共同体现了共享任务中出现的关键设计模式，从基于提示的流程到混合训练式极端多标签文本分类系统。除这三个系统外，更广泛的参赛方案还涵盖了多样的方法空间，包括纯检索流程：从相似的已标引记录中迁移主题词（Tian 等人，2025a；Singh 等人，2025年）、先检索候选项、再对其进行更精确重评分的双编码器加交叉编码器重排序流程（Dorkin 和 Sirts，2025年）、结合多个特定语言或多语言编码器预测结果的多语言 BERT 集成模型（ Hahn，2025 年）、通过对比学习微调、明确将相关记录—主题词对在表示空间中彼此更接近的模型（Jiang 等人，2025年）、Burst Attention 增强的检索方法：在进行前 k 项k 检索之前，通过建模嵌入维度之间的交互来优化嵌入（Islam 等人，2025年）、检索增强生成风格的主题选择流水线：先检索候选主题词，再利用大语言模型在上下文中对其进行验证或排名，并借助 OntoAligner Python工具包（Tekanlou 等人，2025年；Babaei Giglou 等人，2025年），以及通过合成数据生成和偏好优化对大语言模型进行微调、从而更好地使输出符合人工标引行为的流程（Tian 等人，2025b；Ho，2025年）。在这些参赛方案中，出色的性能通常与多阶段架构相关，而非仅依赖单一模型预测；常见策略包括候选检索后进行重排序、模型集成、多语言处理，以及大语言模型辅助增强或优化。下文介绍的三个系统之所以入选，并不是因为它们穷尽了这一设计空间，而是因为它们清晰、易于解释地代表了其中的主要类别，因此适合作为理解该数据集的参考点。

4.1. 方法

系统1（Salfinger 等人，2025年）。图书馆语料库通常呈现长尾主题分布：少数高层主题词出现频繁，而许多细粒度主题很少出现。直接在此类数据上训练的模型往往过度预测高频主题词，而忽略罕见但有意义的主题词。为缓解这一问题，系统1将多标签分类重构为在共享的768维嵌入空间中进行语义检索，使用 sentence-transformers/all-mpnet-base-v2（Song 等人，2020年）。训练记录（标题 + 摘要）与主题标签（名称 +别名）均被嵌入同一空间，以便与查询记录进行语义比较。该方法结合两种互补策略：本体推理，即将查询嵌入与主题嵌入直接比较；以及类比推理，即检索相似的训练记录并迁移其所分配的主题词。来自两个来源的候选主题经合并、去重后，~~按嵌入距离排序，~~并将排名靠前的结果作为预测。

系统2借助少样本检索策略，沿着系统1的类比推理思路进一步发展。

系统 2（Kähler 等人，2025年）。 这种方法采用四阶段流程，将检索增强的少样本提示与受控词表映射和排名相结合。 检索。 使用多语言 BGE-M3 模型（Chen 等人，2024年）对输入记录进行嵌入，并使用 Weaviate 向量

团队名称	nDCG@k			
	k=5	*k=10*	*k=15*	*k=20*
系统1	0.3639	0.3977	0.4143	0.4247
系统2	0.4919	0.4880	0.4879	0.4879
系统3	0.6020	0.6391	0.6560	0.6652

表 5：我们的主题标引数据集上三个系统在不同排名截断位置的 nDCG@k 分数。

存储 检索 L 同一语言中最相似的训练集文档。 完成。这些示例及其 GND 主题词标注会被插入一个即时生成的少样本提示中，供Ministral-8B-Instruct 模型生成自由形式的关键词建议。 映射。使用 BGE-M3 将生成的关键词嵌入向量，并通过混合 HNSW + BM25 检索（Robertson 等人，2009年；Malkov 和 Yashunin，2018年）映射到 GND 主题词，从而将自由词汇与受控分类体系对齐。 排名。最后，Llama-3.1-8B-Instruct（Dubey 等人，2024年）为每个映射后的术语分配 0–10 的相关性得分，用于归一化并排列最终主题词列表。与系统1一样，该流程完全依赖现成模型，不进行微调，并且每条记录仅执行一次生成和排名。

最后，与系统1和系统2不同，系统3通过以下方式将大语言模型与传统极端多标签文本分类算法相结合：ANNIF 工具包(Suominen，2019年)。

系统3（Suominen 等人，2025年）。这种混合方法使用大语言模型进行数据预处理和最终重排序，同时依靠极端多标签文本分类模型进行主题预测。翻译。首先，使用google/gemma-3-4b-it和CohereLabs/aya-expanse-8b，分别将记录和词表翻译为单语英语和德语语料集。合成。随后，多个大语言模型生成合成训练数据，为每种语言中的每条原始记录额外生成四条记录。训练和预测。使用三个 Annif 后端训练两个单语集成模型（英语和德语）：Omikuji Bonsai（Khandagale 等人，2020年），用于分区标签树分类；MLLM （类似 Maui 的（Medelyan，2009年）词汇匹配），用于文本与术语匹配；以及 XTransformer，这是一种基于 Transformer、在 PECOS 框架内运行的极端多标签文本分类排名模型（Yu 等人，2022年）。每个后端都在通过大语言模型翻译构建的单语数据集上单独训练，并组合成特定语言的集成模型。合并并重排序。最后，使用mistralai/Mistral-Small-3.1-24B-Instruct-2503对预测结果进行重排序，并合并双语输出，生成最终的主题词排名列表。

4.2. 定量结果

表 5 展示了三个系统在我们重新发布的测试集上的定量结果。总体而言，

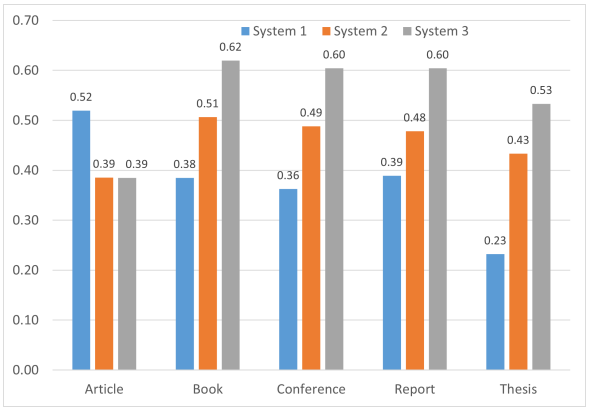


图 2：按五种记录类型划分的 nDCG@5 分数。

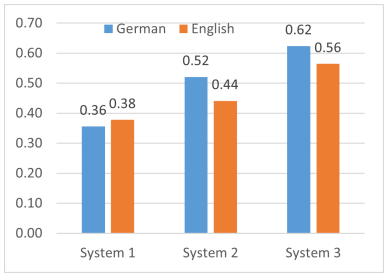
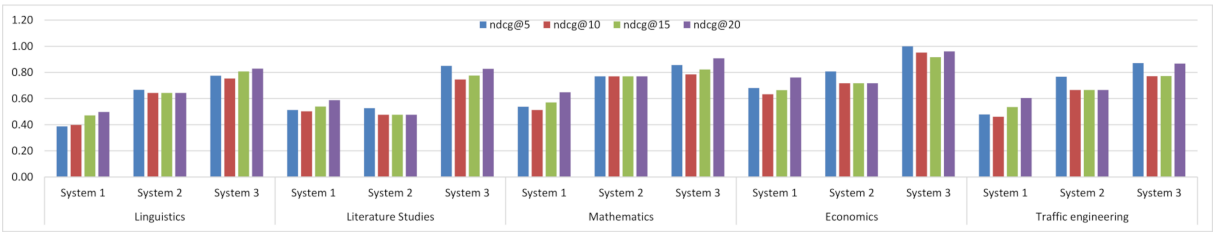


图 3：按两种语言划分的 nDCG@5 分数。

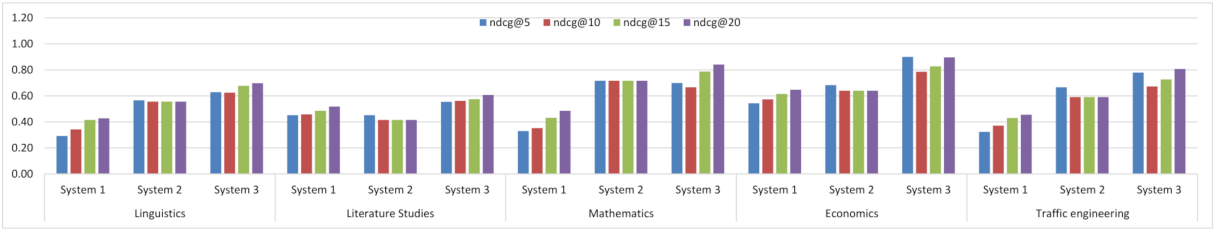
经过大量工程优化的系统3效果最佳，而系统2则巧妙拓展了系统1的思路，通过动态检索的少样本示例，在无需训练的情况下激发大语言模型出色地完成任务。在实际图书馆环境中，人工智能辅助的实用性取决于能否在简洁的预测列表前列展示最相关的主题词。根据图书馆员的反馈，我们将最佳列表长度设为 $k = 20$，以兼顾覆盖面和效率。 图 2 和 图 3展示了各记录类型和语言在 nDCG@5 指标上的表现。系统1在文章记录上表现最佳，而系统3在其余四种类型上领先。按语言来看，系统1在英语记录上表现最强，而系统2和系统3在德语记录上的表现更好。总体而言，系统2为未经训练的基于大语言模型的方法确立了基线（nDCG@5 = 0.4919），系统3则为混合模型训练方法确立了基准测试（nDCG@5 = 0.6020）。

4.3. 定性结果

我们请学科馆员对三个系统的输出进行人工评估，作为一种人工抽查（图4）。十条随机测试记录涵盖五个领域——语言学、文学研究、数学、经济学和交通工程——每条均由一名专家审阅。对于每个预测主题，图书馆员标注了Y



（a）案例 1——将 Y （正确）和 I（不相关但技术上正确）的主题 预测都视为正确。



（b）案例 2——仅将 Y（正确）的主题预测视为正确。

图 4：在 10 条记录、每个领域各 10 条记录上进行定性评估所得的 nDCG@k 分数，其中 k=5,10,15,20。本次评估中，图书馆的主题专家手动将主题预测标记为 Y 和 I。

（正确）、I（技术上正确但无关）或N（不正确）。图4a 显示当将Y和I均视为正确时（情形1），k=5、10、15、20处的nDCG分数，而图4b 显示仅将Y计为正确时（情形2）的结果。

对于案例 1 和案例 2，各领域的 nDCG@20 相对排名在不同系统间呈现出一致的模式。系统 1 在经济学中表现最佳，其次是数学和文学研究，而交通工程表现较差。系统 2 在数学和经济学中表现最强，整体上在各领域的表现较为均衡。系统 3 的表现优于前两者，在经济学和数学中取得近乎完美的分数，并在所有领域都保持强劲表现，体现了混合学习和集成设计的优势。案例 1（计入 Y + I）与案例 2（仅计入 Y）之间的差异表明，许多排名靠前的主题词在语境上相关，但并非完全匹配——这说明系统，尤其是基于大语言模型的系统，能够较好地捕捉主题邻近性，但仍难以区分同一条记录中语义上不同的主题词。这凸显了未来模型面临的一项关键挑战：提升排名靠前预测结果的概念多样性，而不是聚集在近义词上。

进一步查看这些抽查记录中系统1的输出后发现，由于相似度计算的影响，类比推理分支占据了主导地位：训练记录的嵌入与测试记录之间的距离，比GND主题词的嵌入与测试记录之间的距离更为相似。因此，本体分支中正确识别出的主题词往往排名较低。另一个常见错误是，类比分支从相似的训练记录中转移了所有 主题词，尽管其中只有一部分适用。例如，在*“Sprachwissenschaft imFokus”*中，系统仅仅因为英语、西班牙语和罗曼语族与检索记录中的正确标签德语 共同出现，便推断出了这些主题词，从而导致大量假阳性，也凸显了筛选语境相关主题的必要性。对于系统2，大多数错误出现在*Map* 组件中，该组件负责将大语言模型生成的关键词关联到GND术语，包括由于GND中存在未知或含义模糊的概念而产生的映射错误，以及虽已正确映射但不相关、在技术上却有效的建议（类别I）。总体而言，该系统倾向于生成看似合理但过于宽泛的术语，而不是精确的主题匹配；系统3的许多错误也表现出同样的情况。

5. 结论

本文介绍了一个将图书馆记录与丰富主题分类体系配对的极端多标签文本分类数据集，为主题标引研究提供了新颖资源。我们提出了三种互补方法。得分最高的系统主要依赖传统机器学习，这引发了人们对纯粹基于大语言模型的方法之适用性和泛化能力的疑问，而这类方法尚未得到充分探索。我们的数据集为评估语言模型能否捕捉图书馆标引所需的细微语义，提供了一个具有挑战性的基准测试。未来的工作将对多语言嵌入进行基准测试，探索将小型语言模型作为高效替代方案，并研究语言模型蒸馏以便实际部署。该数据集还支持从层级性、透明度和在实际图书馆工作中的实用性等方面评估按层级进行主题标引。

6. 参考文献

Anila Angjeli、Emmanuelle Bermès、Dean Birkett、Maarten Brinkerink、Valentine Charles、Mariana Damova、Cécile Devarenne、Maria Gäde、Sergiu Gordea、David Haskiya、Timothy Hill、Antoine Isaac、Lukas Koster、Hugo Manguinhas、Gregory Markus、Mark A. Matienzo 和 Vivien Petras。2016。数字图书馆多语言访问最佳实践白皮书。技术报告，Europeana Foundation。开放获取白皮书。

Hamed Babaei Giglou、Jennifer D'Souza、Oliver Karras 和 Sören Auer。2025年。Ontoaligner：用于本体对齐的全面、模块化且稳健的 Python 工具包。载于欧洲语义网会议，第 174–191 页。Springer。

Rohit Babbar 和 Bernhard Schölkopf。2017年。Dis-mec：用于极端多标签分类的分布式稀疏机。载于第十届 *ACM* 国际网络搜索与数据挖掘会议论文集，第 721–729 页。

K. Bhatia、K. Dahiya、H. Jain、P. Kar、A. Mittal、Y. Prabhu 和 M. Varma。2016。极端分类资源库：多标签数据集与代码。

Kush Bhatia、Himanshu Jain、Purushottam Kar、Manik Varma 和 Prateek Jain。2015年。用于极端多标签分类的稀疏局部嵌入。神经信息处理系统进展，28。

Kathi Canese 与 Sarah Weis。2013。PubMed：书目数据库。《*NCBI*手册》，2(1):2013。

Caterina Caracciolo、Armando Stellato、Ahsan Morshed、Gudrun Johannsen、Sachit Rajbhandari、Yves Jaques 与 Johannes Keizer。2013。AGROVOC关联数据集。语义网，4(3):341–348。

Wei-Cheng Chang、Hsiang-Fu Yu、Kai Zhong、Yiming Yang 与 Inderjit S Dhillon。2020年。驯服预训练变换器以进行极端多标签文本分类。见第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第3163–3171页。

Jianlv Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2024年。BGE M3-Embedding：通过自知识蒸馏实现多语言、多功能、多粒度文本嵌入。*arXiv*预印本 *arXiv:2402.03216*。

Corine Deliot。2014年。将英国国家书目作为关联开放数据发布。《目录与索引》，174：13–18。

Maria Dermentzi、Mike Bryant、Herminio García-González 与 Fabio Rovigo。2025a。Ehri多语言主题标引测试数据集。

Maria Dermentzi、Mike Bryant、Fabio Rovigo 和 Herminio García-González。2025b。多语言自动主题标引：EHRI 项目背景下大语言模型与替代方法的比较研究。工作论文或预印本。

Aleksei Dorkin 和 Kairit Sirts。2025年。TartuNLP 在 SemEval-2025 任务5中的工作：将主题标注视为两阶段信息检索。载于第*19*届语义评估国际研讨会（*SemEval-2025*）论文集，第2449–2454页。

Jennifer D'Souza、Sameer Sadruddin、Holger Israel、Mathias Begoin 和 Diana Slawig。2025年。SemEval-2025 任务5：LLMs4Subjects——面向国家技术图书馆开放获取目录的基于大语言模型的自动主题标注。载于第*19*届语义评估国际研讨会（*SemEval-2025*）论文集，第2570–2583页，奥地利维也纳。计算语言学协会。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan 等。2024年。Llama 3 模型群。*arXiv* 电子预印本，页码 arXiv–2407。

联合国粮食及农业组织。2023年。介绍 AGRIS 2.0！在线新闻文章。访问日期：<填写访问日期>。

Luis Gasco、Anastasios Nentidis、Anasta- sia Krithara、Darryl Estrada-Zavala、Re- nato Toshiyuki Murasaki、Elena Primo-Peña、Cristina Bojo Canales、Georgios Paliouras、Martin Krallinger 等。2021年。BioASQ 2021-MESINESP 赛道概述：面向科学文献、专利和临床试验的先进层次分类技术评估。*CEUR* 研讨会论文集。

Jim Hahn. 2025年. Jim 在 SemEval-2025 任务5：多语言 BERT 集成. 载于第*19*届语义评估国际研讨会论文集（*SemEval-2025*），第2407–2412页。

Clara Wan Ching Ho。2025年。UBFFM 在 GermEval-2025 LLMs4Subjects 任务中的表现：如果我们认真看待"你是主题标引专家"会怎样？收录于第 *21* 届自然语言处理会议（*KON- VENS 2025*）论文集：研讨会，第 471–478 页，德国汉诺威。HsH Applied Academics。

Antoine Isaac 和 Bernhard Haslhofer。2013。Europeana 关联开放数据——data. europeana. eu。语义网，4(3):291–297。

Baharul Islam、Nasim Ahmad、Ferdous Barbhuiya 和 Kuntal Dey。2025年。NBF 在 SemEval-2025 任务5：面向多语言主题推荐的光突发注意力增强系统。收录于第*19*届国际语义评测研讨会论文集（*SemEval-2025*），第953–958页。

Paul Jaccard。1912。高山带植物区系的分布。1. *New Phytologist*，11(2)：37–50。

Kalervo Järvelin 和 Jaana Kekäläinen。2002。信息检索技术的累积增益评估。*ACM* 信息系统汇刊（*TOIS*），20(4)：422–446。

Hong Jiang、Jin Wang 和 Xuejie Zhang。2025年。YNU-HPCC 在 SemEval-2025 任务5：利用多语言 Sentence-BERT 进行 GND 主题标注的对比学习。收录于 第*19*届国际语义评测研讨会论文集（*SemEval-2025*），第2443–2448页。

Maximilian Kähler、Lisa Kluge 和 Katja Koner- mann。2025年。DNB-AI-project 在 GermEval- 2025 LLMs4Subjects 任务中的成果：KIFSPrompt——知识注入式少样本提示。载于《第*21*届自然语言处理会议（*KONVENS 2025*）论文集：研讨会》，第455–464页，德国汉诺威。HsH Applied Academics。

Sujay Khandagale、Han Xiao 和 Rohit Babbar。2020年。Bonsai：用于极端多标签分类的多样化浅层树。《机器学习》，109(11)：2099–2119。

Anastasia Krithara、James G. Mork、Anastasios Nentidis 和 Georgios Paliouras。2023a。从人工到自动化的生物医学文献语义索引之路：十年历程。*Frontiers in Research Metrics and Analytics*，8。

Anastasia Krithara、Anastasios Nentidis、Konstantinos Bougiatiotis 与 Georgios Paliouras。2023b。Bioasq-qa：面向生物医学问答的人工整理语料库。*ScientificData*，10(1):170。

Solomon Kullback 与 Richard A. Leibler。1951。论信息与充分性。数理统计年刊，22(1):79–86。

Jingzhou Liu、Wei-Cheng Chang、Yuexin Wu 和 Yiming Yang。2017年。用于极端多标签文本分类的深度学习。载于第*40*届国际*ACM SIGIR*会议论文集：信息检索研究与开发，第115–124页。

Yu A Malkov 和 Dmitry A Yashunin。2018年。使用分层可导航小世界图进行高效且稳健的近似最近邻搜索。*IEEE* 模式分析与机器智能汇刊，42(4):824–836。Olena Medelyan。2009年。具有人类竞争力的自动主题索引。博士学位论文，怀卡托大学。

Isabelle Mohr、Markus Krimmel、Saba Sturua、Mohammad Kalim Akram、Andreas Koukounas、Michael Günther、Georgios Mastrapas、Vinit Ravishankar、Joan Fontanals Martínez、Feng Wang 等。2024年。面向8192词元双语文本嵌入的多任务对比学习。*arXiv*预印本 *arXiv:2402.17016*。

NLM。2000年。医学主题词，第41卷。美国卫生与公众服务部，公共卫生服务署，美国国家医学图书馆。

Hercules Panoutsopoulos 和 Christopher Brew- ster。2022年。利用农业文本语料库更新 AGROVOC。在*HAICTA*会议上，第260–265页。

Karl Pearson。1900年。X. 论如下判据：对于相关变量系统，若一组偏离最可能值的偏差符合该判据，则可以合理地认为这些偏差源于随机抽样。《style id='1'>《伦敦、爱丁堡和都柏林哲学杂志与科学期刊》，50(302):157–175。

Yashoteja Prabhu、Anil Kag、Shrutendra Harsola、RahulAgrawal 和 Manik Varma。2018年。Parabel：用于极端分类的分区标签树及其在动态搜索广告中的应用。在*2018*年万维网大会论文集，第993–1002页。

Yashoteja Prabhu 和 Manik Varma。2014年。FastXML：一种用于极端多标签学习的快速、准确且稳定的树分类器。在第*20*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第263–272页。

Stephen Robertson、Hugo Zaragoza等，2009年。概率相关性框架：BM25及其拓展。基础与趋势®信息检索，3(4):333–389。

Andrea Salfinger、Luca Zaccagna、Francesca Incitti、Gianluca De Nardi、Lorenzo Dal Fabbro和Lauro Snidaro。2025年。 LA²I²F在SemEval-2025任务5中：嵌入空间中的推理——融合类比推理与基于本体的推理以进行文档主题标注。载于第*19*届语义国际研讨会论文集

评测（*SemEval-2025*），第2413–2423页，奥地利维也纳。计算语言学协会。

Gauri Salokhe、Irene Onyancha、James Weinheimer、Barbara Richards、Fynvola Le Hunte Ward 和 Johannes Keizer。2023年。AGRIS 2.0：一个更灵活的多语言书目平台。联合国粮食及农业组织（FAO）知识交流与能力建设司。访问日期：<请添加访问日期>。

Hinrich Schütze、Christopher D Manning 和 Prabhakar Raghavan。2008年。信息检索导论，第39卷。剑桥大学出版社，剑桥。

Sumit Singh、Pankaj Goyal 和 Uma Tiwary。2025年。SemEval-2025 任务5中的 silp_nlp：使用句子转换器进行主题推荐。载于第*19*届语义评估国际研讨会论文集（*SemEval-2025*），第2455–2460页。

Kaitao Song、Xu Tan、Tao Qin、Jianfeng Lu 和 Tie-Yan Liu。2020年。MPNet：面向语言理解的掩码与置换预训练。神经信息处理系统进展，33：16857–16867。

Osma Suominen。2019年。Annif：使用多种算法实现 DIY 自动主题标引。*LIBER Quarterly*：欧洲研究图书馆协会期刊，29(1):1–25。

Osma Suominen、Juho Inkinen 和 Mona Lehtinen。2025年。Annif 在 GermEval-2025 LLMs4Subjects 任务中的表现：由高效大语言模型增强的传统极端多标签文本分类。载于第*21*届自然语言处理会议（*KONVENS 2025*）论文集：研讨会，第447–454页，德国汉诺威。HsHAppliedAcademics。

Osma Suominen、Juho Inkinen、Tuomo Virolainen、Moritz Fürneisen、Bruno P. Kinoshita、Sara Veldhoen、Mats Sjöberg、Philipp Zumstein、Robin Neatherway 和 Mona Lehtinen。2023年。Annif。

Yukihiro Tagami。2017年。Annexml：用于极端多标签分类的近似最近邻搜索。载于第*23*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第455–464页。

T.T. 谷本。1958年。分类与预测的初等数学理论。国际商业机器公司。

Hadi BayramiAslTekanlou、JafarRazmara、Mahsa Sanaei、Mostafa Rahgouy 和 Hamed Babaei Giglou. 2025年. Homa 参加 SemEval-2025 任务5：借助 OntoAligner 对齐图书馆员记录以进行主题标注。载于 第*19*届国际语义评测研讨会论文集（*SemEval-2025*），第2400–2406页。

Xia Tian、Yang Xin、Wu Jing、Xiu Heng、Zhang Xin、Li Yu、Gao Tong、Tan Xi、Hu Dong、Chen Tao 等。2025a。中国人民大学团队参加 SemEval-2025 任务5：快速自动主题标引：一种基于相似记录匹配和相关主题排序的方法。载于第*19*届语义评估国际研讨会（*SemEval-2025*）论文集，第2437–2442页。

Yicen Tian、Erchen Yu、Yanan Wang、Dailin Li、Jiaqi Yao、Hongfei Lin、Linlin Zong 和 Bo Xu。2025b。DUTIR831 参加 SemEval-2025 任务5：一种面向 TIBKAT 记录的 GND 主题分配多阶段大语言模型方法。载于第*19*届语义评估国际研讨会（*SemEval-2025*）论文集，第363–372页。

George Tsatsaronis、Georgios Balikas、Prodromos Malakasiotis、Ioannis Partalas、Matthias Zschunke、Michael R Alvers、Dirk Weissenborn、Anastasia Krithara、Sergios Petridis、Dimitris Polychronopoulos、Yannis Almirantis、John Pavlopoulos、Nicolas Baskiotis、Patrick Gallinari、Thierry Artiéres、Axel-Cyrille Ngonga Ngomo、Norman Heino、Eric Gaussier、Liliana Barrio-Alvers、Michael Schroeder、Ion Androutsopoulos 和 Georgios Paliouras。2015年。BioASQ 大规模生物医学语义索引与问答竞赛综述。*BMC Bioinformatics*，16(1):138。

Henrique Schechter Vera、Sahil Dua、Biao Zhang、Daniel Salz、Ryan Mullins、Sindhu Raghuram Panyam、Sara Smoot、Iftekhar Naim、Joe Zou、Feiyang Chen 等。2025年。EmbeddingGemma：强大且轻量的文本表示。*arXiv* 预印本 *arXiv:2509.20354*。

Liang Wang、Nan Yang、Xiaolong Huang、Linjun Yang、Rangan Majumder 和 Furu Wei。2024年。Multilinguale5 文本嵌入：一份技术报告。*arXiv* 预印本 *arXiv:2402.05672*。

Ian EH Yen、Xiangru Huang、Wei Dai、Pradeep Ravikumar、Inderjit Dhillon 和 Eric Xing。2017年。PPDSparse：一种用于极端分类的并行原始-对偶稀疏方法。载于第*23*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第545–553页。

Ian En-Hsu Yen、Xiangru Huang、Pradeep Ravikumar、Kai Zhong 与 Inderjit Dhillon。2016。Pd-sparse：一种面向极端多类与多标签分类的原始与对偶稀疏方法。

载于国际机器学习会议，第3069–3077页。PMLR。

Ronghui You、Zihan Zhang、Ziye Wang、Suyang Dai、Hiroshi Mamitsuka 和 Shanfeng Zhu。2019年。AttentionXML：一种基于标签树的注意力机制感知深度模型，用于高性能极端多标签文本分类。神经信息处理系统进展，32。

Hsiang-Fu Yu、Kai Zhong、Jiong Zhang、Wei-Cheng Chang 和 Inderjit S Dhillon。2022年。PECOS：面向海量且相关输出空间的预测。机器学习研究杂志，23(98)：1–32。

Jiong Zhang、Wei-Cheng Chang、Hsiang-Fu Yu 和 Inderjit Dhillon。2021。用于极端多标签文本分类的快速多分辨率 Transformer 微调。神经信息处理系统进展，34：7267–7280。

Ruohong Zhang、Yau-Shian Wang、Yiming Yang、Donghan Yu、Tom Vu 和 Likun Lei。2023年。通过检索生成的伪标签描述进行长尾极端多标签文本分类。在 计算语言学协会论文集：*EACL 2023* 中，第1092–1106页。

A. 系统1详细错误分析

在以下部分中，我们将对系统1（LA²l²F）在Salfinger 等人，2025年所述测试数据集上的不同系统性错误类型进行更详细的分析，详见第4.3小节。在这项定性分析中，学科馆员通过检查系统为50份测试文档各给出的20个预测标签，对系统输出进行人工评估；这些文档共涉及161个金标准主题词。每项预测均被标记为Y（正确）、I（技术上正确但无关）或N（错误），从而使我们能够更深入地分析系统性错误模式。

假阴性（遗漏的真实标签）。总体而言，在考虑每篇文献 k =20个预测主题词时，系统1的假阴性率（假阴性率）为56.5%，即遗漏的真实主题标签所占比例；而在仅考虑系统排名前5的预测时，假阴性率为73.2%。然而，由于GND本体和人工标注训练语料库规模庞大，分配给一篇文献的主题标签的具体选择可能带有一定主观性，因而会受到标注专家的影响，尤其是当本体中存在同义概念时。为了在评估系统预测所遗漏的主题标签时考虑到此类潜在歧义，我们将分析加以剖析对假阴性（假阴性），即系统遗漏的真实标签，按以下类别：

FNM：系统完全遗漏的方面

FNC：遗漏的方面，但被其他预测紧密匹配

仅通过人工检查系统1排名前五的预测，我们使用这些定义，得出了50条测试集记录的以下绝对计数：

错误类别	绝对计数
FNM	79
FNC	41
假阴性（总计）	120

表6：在50份测试集文档（包含161个金标准主题词）上，系统1为250条主题词建议分配的假阴性分布（每份测试集文档预测5个标签）。

正如这项将同义主题纳入考量的细粒度人工检查所揭示的，系统1的假阴性中有34.1%未能检索到确切真实标签，而是检索到了高度相似的同义词；另有65.8%对应于系统1前5项预测中完全缺失的真实概念。

进一步探究其背后的原因，我们发现了系统1引入的以下系统性错误类别：在系统1实施的融合策略中，类比推理占据主导地位。分析系统1的两个推理分支对融合结果的贡献后，我们发现类比分支主导了结果：文档间距离往往小于文档与主题间距离。因此，本体分支提供的主题词通常排在类比分支提供的主题词之后，因而经常无法进入融合后的前列主题词列表。测试集中的一条记录展示了这一情况，见k 返回的排名主题词表7 （其金标准预测见表9）。在此例中，真实标签*AWACS* 由本体分支预测为第22位，但在合并过程中，因类比分支的文档距离更小（即余弦相似度更高）而被挤到后面，导致一项正确预测在合并过程中被过滤掉。这表明有必要开发更精密的融合策略，例如增加一个与系统2类似的下游基于大语言模型的相关性过滤和重排序组件。

假阳性（错误预测）。此外，还需要一个下游相关性过滤组件，以解决另一个固有的

错误类别：系统1目前对类比推理采用了过于乐观的假设，将嵌入空间中最接近的训练文档的所有主题标签视为适用于新文档的标签。我们发现，这是引入与新文档不匹配的错误预测的主要固有错误来源，即系统1中的假阳性（FP）。虽然相关文档的主题标签可能存在重叠，但文档 A 中的所有主题词未必都适用于相似的文档 B。

错误类别。 总之，我们将已识别出的主要系统性错误类别表示为：

FPAR 类比推理引入的假阳性

ORoR 经与类比推理结果融合后，本体推理得出的正确结果排名靠后

表7 展示了测试集记录 ID 中这些系统性错误的一些具体示例3A1007389885。

预测	R	Rank	Sim.	评估	错误类别
Elektronische 对抗措施 [电子对抗措施] gnd:1082384615	A	1	0.687	N	FPAR
干扰发射机 [干扰机] gnd:4273774-6	A	2	0.687	Y	–(真正例, Y)
雷达 [雷达] gnd:4176765-2	A	3	0.687	Y	–（Y）
声呐 [声呐] gnd:4181785-0	A	4	0.652	N	FPAR
信号处理 [信号处理] gnd:4054947-1	A	5	0.652	Y	–(Y)
AWACS [机载预警 and 控制系统-tem] gnd:4309079-5	O	22	0.406		ORoR

表7：系统1对测试集记录 ID 3A1007389885 的预测，“机载雷达信号处理的最新进展：新兴研究与机遇”，并分析哪个推理分支（**R**）识别出了每个预测标签——类比推理（A）或本体推理（O）。前五项结果是融合系统合并 A 和 O 两个分支的结果后返回的前5项预测；最后一行显示了一个实际的真实主题词，该主题词因排名靠后而未被纳入检索结果。**排名** 表示各主题词在每个推理分支（类比与本体）中的排名，**相似度** 显示了决定排名的嵌入空间中与查询文档的余弦相似度，而**评估**列出了人类专家对预测标签的判断。

总之，这项分析证实了关于系统1融合架构互补性的假设：两个推理分支都能识别出互补的信息，并指出了未来工作的方向。可以通过引入额外的过滤组件来评估并重新排列已识别的（候选）主题词，从而解决已识别出的系统性错误来源；这应能提高本体分支所贡献的匹配预测的排名，并剔除类比分支引入的假阳性。

B. 系统2详细错误分析

人工检查系统输出后发现了若干反复出现的错误模式。在仔细检查主题专家也参与定性评估的50份测试文档时，我们发现系统2（KIFSPrompt）共给出了176个主题词建议（Käh- ler 等人，2025年），其中110个是假阳性，可分为多个子类别。定性错误分析的完整数据表可在以下地址找到：
h
t
t
p
s://
g
i
t
h
u
b.
com/sciknoworg/tib-sid/tree/main/evaluation/results/system2%20-%20additional%20analysis。

主题专家认为，40条假阳性建议是相关的，另有27条建议虽不相关，但技术上正确。因此，剩余43条假阳性建议（39%）应被视为真正错误的建议。研究系统为何会给出这些真正错误的建议很有意思。事实上，部分错误可归因于系统映射阶段出现的问题：在该阶段，自由关键词会与规范化主题词进行匹配。我们观察到两种反复出现的模式：

MEU: 映射错误，因为 GND 中的一个概念未知

MEA: 映射错误，因为 GND 中的一个概念存在歧义

让我们用一些例子来说明这些类别。在我们的样本中，一份标题为《莎士比亚与改编研究手册》的文献被标注了金标准主题词*Adaption*（文学）（gnd:102289935X）。大语言模型只建议了*Adaption*，而该词是 GND 主题词*Anpassung*（gnd:4128128-7）的一个替代标签。映射过程使用了这个替代标签，并找到了完全匹配项。这就是我们所说的**MEA**：由歧义导致的映射错误。

作为**MEU**的示例，请考虑标题“机载雷达信号处理的最新进展：新兴研究与机遇”

(记录标识符 3A1007389885)。表8展示了系统2的输出。

大语言模型建议的 term	映射的 GND 术语	错误类别
信号处理 [信号处理]	信号处理 gnd:4054947-1 [信号处理]	假阳性，但相关
干扰 [干扰]	干扰发射机 gnd:4273774-6 [干扰发射机]	TP
欺骗 [欺骗]	欺骗 gnd:4554780-4 [欺骗]	MEU
机载雷达 [航空雷达]	气象雷达 gnd:4270420-0 [气象雷达]	MEU

表8：系统2对记录标识符的输出3A1007389885“近期空基雷达信号处理的进展：新兴研究与机遇”及错误类别标注。

表9显示了同一文档的金标准标注。

金标准术语	GND标识符
干扰发射机 [干扰发射机]	gnd:4273774-6
AWACS [机载预警与控制系统]	gnd:4309079-5
空间-时间-信号处理 [时空信号处理]	gnd:4834654-8
机载雷达 [机载雷达]	gnd:4456131-3
目标识别 [目标识别]	gnd:4190792-9
军事欺骗 [欺骗（军事）]	gnd:4184333-2

表 9：记录标识符 3A1007389885 的金标准术语。

我们看到，大语言模型建议的候选词欺骗与概念欺骗（军事）十分接近，而机载雷达与*AWACS*和机载雷达都很接近。然而，由于欺骗和机载雷达并未直接收录于 GND 中，这些术语会被匹配到 GND 中错误的实体。在这些情况下，大语言模型建议了 GND 中不存在的概念。

在样本中 43 条确实错误的建议里，11 条可归为**MEU**，5 条可归为**MEA**。

对排序阶段进行微调，并基于余弦相似度采用更严格的过滤，或许有助于在未来减轻此类错误。在此类过滤步骤中被移除的建议，也可以作为 GND 中新增主题词的候选，用于补充词表（作为同义词或新概念）。解决歧义的问题仍然是一项严峻挑战。在实际生产环境中，对这些错误进行分析，以完善用于自动主题标引的词表，将有助于对系统进行补充并提升其性能。

在分析假阳性的基础上，我们还可以从召回率的角度分析这个样本，并进一步仔细查看假阴性。

使用前文介绍的假阴性类别[6]，，我们得到以下绝对计数：

错误类别	绝对计数
FNM	68
FNC	28
假阴性（总计）	96

表10：系统2针对50份测试集文档提出的176条主题建议中，假阴性的分布情况；这些文档包含162个金标准主题。

如果我们只将**FNM**计为真正缺失的方面，则可得到微平均召回率

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FNM}} = 0.49$$

C. 系统3详细错误分析

仔细审查了在定性评估期间也由专家评定的50份测试集文档，及系统3（Annif）预测的前5个主题词（Suominen 等人，2025年）（总计250个预测主题词），以及这些文档在TIBKAT中的真实主题（总计161个）后，我们发现了91个真正例和70个假阴性。除了前文介绍的假阴性类别外，我们还定义了以下附加类别：

FNM5：前5项预测中遗漏、但出现在随后前20项预测中的方面（FNM的子类别）

采用这种分类方式，我们发现了61个FNM案例，其中25个属于FNM5；也就是说，该主题未出现在前5项预测中，但出现在前20项预测列表的较后位置。我们还发现了9个FNC案例。错误分析的完整数据见 https://github.com/sciknoworg/tib-sid/tree/main/evaluation/results/system3%20-%20additional%20analysis。

为通过示例说明这些类别，请看标题为《超越绝对连续性范式的动态期限结构建模》（记录编号3A168734406X）的文献。表11展示了系统3对该文献的前5项预测。表12展示了同一文献的金标准标注。

这份示例文档说明了准确预测金标准主题的困难，尽管这些预测被评为高度相关

[6]与系统1和系统3不同，系统2会针对每条记录动态调节建议数量，通常少于五条。因此，没有必要像分析其他系统那样，分析某个方面是否在排名靠后的位置才匹配。

预测	GND标识符	评估	错误类别
信用风险 [信用风险]	GND:4114309-7	Y	假阳性，但相关
信贷市场 [信贷市场]	GND:4073788-3	Y	假阳性，但相关
利率期限结构理论 [术语 结构 理论]	gnd:4117720-4	Y	假阳性，但相关
建模 [建模]	gnd:4170297-9	I	假阳性，技术上正确
半鞅 [半鞅]	gnd:4180967-1	Y	TP

表11：记录编号3A168734406X的系统3前5项输出“超越绝对连续性范式的动态期限结构建模”，附有专家评估和错误类别标注。

金标准主题	GND标识符	频次	错误类别
套利 [套利]	gnd:4002820-3	0	FNM
半鞅 [半鞅]	gnd:4180967-1	4	TP
违约风险 [违约风险]	gnd:4205942-2	4	FNM5
利率期限结构 [期限结构]	gnd:4067855-6	1	FNC
HJM模型 [HJM模型]	gnd:4642940-2	1	FNM5

表12：记录 ID 3A168734406X 的金标准主题词、训练集频次及错误类别标注。

由主题专家评定。前五项预测见于表11，仅有半鞅 出现在金标准主题中，因此计为真正例。其余四项预测中，有三项被主题专家视为相关（Y），一项被视为技术上正确但无关（I）。

从表12中五个金标准主题词的角度来看同一份文档，结果仍然只有一个是真正例，其余四个属于不同类型的假阴性：系统完全没有预测出套利，违约风险和*HJM*模型未出现在前五个预测中，但出现在其余前20个预测中；而利率期限结构没有被预测出来，但与之密切相关的概念利率期限结构理论出现在前五个预测中。这份文档的全部五个金标准主题词在训练集中的频率都很低，出现次数介于0到4次之间。因此，使用系统3对它们进行预测颇具挑战性，因为系统3主要依赖于一些模型，这些模型根据训练数据中的模式来学习区分每个单独的主题词。

图5展示了按训练集频率分组的主题词直方图，并将其分为真正例和三种子类型的假阴性（图中将FNM5从其余FNM案例中单独列出）。在频率较低的分组0、1和4–7中，假阴性数量超过真正例。在频率较高的分组中，从

8–15及以上，真正例更为常见

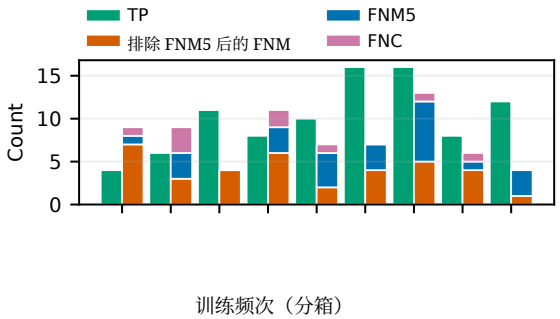


图5：系统3预测的训练频率分布（分箱），分为真正例和三种假负例子类型。

而不是假阴性。第2–3组是这一模式中的异常值，尽管训练集频率较低，真正例仍占主导。

尽管存在一个离群分箱，但整体模式很清楚：系统3在预测训练集中的低频主题词时表现不佳，而频次较高的主题词则更有可能被正确预测。这是意料之中的，因为构成该集成的三种算法中，只有一种（MLLM）能够预测零样本主题，另外两种算法（Omikuji 和 XTransformer）依赖主题专属训练数据，因此无法很好地预测低频主题词，在零样本情况下甚至完全无法预测。未来可以通过纳入更多专注于预测低频术语的方法来改进系统，例如像系统1那样通过嵌入将文档文本与 GND主题词进行匹配，或像系统2那样使用现成的大语言模型提出可能的候选主题。