An Extreme Multi-label Text Classification (XMTC) Library Dataset: What if we took "Use of Practical AI in Digital Libraries" seriously?

Jennifer D'Souza[1], **Sameer Sadruddin**[1], **Maximilian Kähler**[2], **Andrea Salfinger**[3], **Luca Zaccagna**[3], **Francesca Incitti**[3], **Lauro Snidaro**[3], **Osma Suominen**[4]
[1]TIB Leibniz Information Centre for Science and Technology, Germany
[2]Deutsche Nationalbibliothek, Germany
[3]University of Udine, Italy
[4]National Library of Finland, Finland
{jennifer.dsouza,sameer.sadruddin}@tib.eu

Abstract

Subject indexing is vital for discovery but hard to sustain at scale and across languages. We release a large bilingual (English/German) corpus of catalog records annotated with the Integrated Authority File (GND), plus a machine-actionable GND taxonomy. The resource enables ontology-aware multi-label classification, mapping text to authority terms, and agent-assisted cataloging with reproducible, authority-grounded evaluation. We provide a brief statistical profile and qualitative error analyses of three systems. We invite the community to assess not only accuracy but usefulness and transparency, toward authority-anchored AI co-pilots that amplify catalogers' work.

Keywords: Multilingual subject indexing, Extreme multi-label text classification, XMTC, Benchmark dataset

1. Introduction

Libraries have long relied on expert subject indexing to make collections findable, interoperable, and durable. Yet the rapidly growing, multilingual volume of library catalog records increasingly strains purely manual indexing workflows. At the same time, large language models (LLMs) and emerging agentic pipelines promise support—but they must be grounded in authoritative vocabularies, auditable, and evaluated in library terms rather than by generic text-classification scores. We present a machine-learning-ready resource that directly addresses this gap: a bilingual (English/German), multi-domain corpus of catalog records indexed with subjects from the German Integrated Authority File (Gemeinsame Normdatei, GND), released together with a machine-actionable version of the GND subject taxonomy and predefined train/dev/test splits. The goal is not merely scale, but structured scale—where every prediction links to a controlled vocabulary that libraries already trust.

This resource is designed to help the community interrogate practical questions that matter for library science in the LLM era: *How should automated systems align free text to controlled vocabularies while preserving provenance and authority control? What counts as "useful" assistance—top-k quality at the point of description, hierarchical coherence, explainable rationales, or cataloger effort saved? How can models cope with long-tail subjects, multilingual variation, and distribution shift across domains and time? Where do agents best fit in human-in-the-loop workflows (triage, suggestion, validation)*, and *how should we measure their impact beyond batch metrics?* By anchoring experiments in an operational taxonomy, the dataset enables studies of vocabulary grounding, cross-lingual consistency, polysemy and variant labels, and reliability under realistic label sparsity—questions that generic XMTC benchmarks only partially surface.

At a high level, our contribution pairs real catalog records with stable links to authoritative subject concepts and packages them for reproducible evaluation. This enables ontology-aware multi-label classification, retrieval-augmented mapping from free text to authority terms, and agent workflows that combine retrieval, suggestion, and curator feedback—evaluated with protocols that reflect cataloging realities (e.g., usefulness and hierarchical consistency at the top of the record). We outline the resource, its construction and splits, and initial analyses and baselines, and we position the paper as a **guidebook** to the dataset: we statistically explore it to surface considerations for framing machine-learning solutions, and we conclude with qualitative error analyses of three systems developed on our data—inviting the LREC community to test, compare, and *reflect* on what successful, trustworthy AI assistance for subject indexing should look like. This resource is released as **TIB-SID** (TIB Subject Indexing Dataset), a bilingual (English/German), multi-domain corpus of 136k catalog records annotated with GND subjects, available at `https://github.com/sciknoworg/tib-sid` (CC BY 4.0 license).

The rest of the paper is structured as follows: section 2 reviews related work, section 3 describes our dataset, and section 4 presents three systems trained on it along with a brief qualitative analysis. The article concludes in section 5.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 169–184
11–16 May 2026.

2. Related Work

Extreme Multi-label Text Classification (XMTC) Datasets. XMTC benchmarks (Bhatia et al., 2016) involve assigning items to very large label spaces with highly skewed long-tail distributions (Zhang et al., 2023). Notable examples include *Wiki-500K* (1.8M Wikipedia entries, 500K categories) and *AmazonCat-13K* (1.5M product descriptions, 13,330 categories). In both, most labels are extremely rare (e.g., only 2% of Wiki-500K labels occur more than 100 times; 30% of AmazonCat labels appear in fewer than 10 samples) (Yu et al., 2022; Zhang et al., 2023). Smaller domain-specific corpora such as *EurLex-4K* (19,000 EU legal documents, 4,000 EuroVoc subjects) show similar long-tail behavior. These datasets exemplify the statistical challenges of large label spaces but typically rely on generic or user-defined categories. In contrast, our dataset represents fine-grained, technically specialized subject annotations on library records over a wide range of research domains.

Biomedical Indexing with Large Taxonomies. The biomedical domain has pioneered large-scale semantic indexing with corpora such as *BioASQ* (Tsatsaronis et al., 2015; Krithara et al., 2023b), which provides millions of PubMed abstracts annotated with *Medical Subject Headings* (MeSH), a hierarchical vocabulary of about 30,000 descriptors (NLM, 2000). Each article is typically assigned 12–13 MeSH terms, framing the task as an XMTC problem. The multilingual *MESINESP* corpus extends this approach to Spanish and Portuguese using DeCS (Gasco et al., 2021). Together, these benchmarks demonstrate how controlled domain taxonomies can support large-scale automated subject indexing (Krithara et al., 2023a).

Library Cataloging and Multilingual Linked Data. Multilingual subject-indexing datasets are only partially addressed in prior work, and openly ML-ready benchmarks are scarce. Notable efforts include the EHRI Multilingual Subject Indexing Test Dataset—Holocaust archival descriptions labeled with a 900-term vocabulary in 12 languages (Dermentzi et al., 2025a,b); FAO's AGRIS (16M+ records, 123 languages) indexed with the multilingual *AGROVOC* thesaurus (Caracciolo et al., 2013; Panoutsopoulos and Brewster, 2022); and Europeana, the pan-European cultural-heritage aggregator, enriches records with multilingual subject links to controlled vocabularies such as the Virtual International Authority File (VIAF), EuroVoc, or the Art & Architecture Thesaurus (AAT), available via data dumps and SPARQL endpoints (Isaac and Haslhofer, 2013; Angjeli et al., 2016). National initiatives—the British Library's British National Bibliography (BNB) as Linked Open Data (Deliot, 2014) and the Bibliothèque nationale de France's (BnF) Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU) vocabulary—likewise expose large SKOS/RDF authority datasets. However, these resources typically lack standardized train/test splits and evaluation metrics.

Comparative Outlook – Our Dataset. Unlike prior XMTC benchmarks with flat or user-defined labels and library linked-data releases without ML-ready splits, our corpus pairs XMTC-scale long tails with the structured GND taxonomy, bridging large-scale text classification and knowledge organization. It provides a bilingual, ML-ready benchmark with predefined train/dev/test splits and stable GND links for ontology-aware modeling, multilingual evaluation, and reproducible comparisons.

3. Our Subject Indexing Dataset

This section introduces our subject indexing dataset, which consists of two parts: the taxonomy (subsection 3.1) and the library records (subsection 3.2) that are subject-indexed using it.

3.1. The Subject Indexing Taxonomy

As our subject indexing taxonomy, we use the GND (Gemeinsame Normdatei / Integrated Authority File), a set of integrated authority files created by the German National Library and widely adopted by German-speaking libraries to catalog and link entities such as people, organizations, topics, and works. Among these, we specifically rely on the *Sachbegriff* (subject terms).

Readers can obtain the GND Sachbegriff by following our how-to guide[1], which explains how to download the latest `authorities-gnd-sachbegriff_dnbmarc.mrc.xml.gz` file. It is encoded in MARC 21—an international metadata standard using numeric tags (e.g., 021 for identifier, 550 for related subjects). While highly interoperable across library systems, MARC 21 is not human-readable, so we converted it into a structured JSON format[2] using our internal schema. The resulting file contains 207,001 unique subjects, each represented as a standardized JSON object with fields mapped from MARC 21, including the GND identifier (`Code`), classification number and name, preferred term (`Name`), variant labels (`Alternate Name`), related subjects, and source, with optional entries for `Definition` and `Source URL`. When considering property coverage, all records include the core fields, while contextual information varies—`Alternate Name` appears in about half, `Related Subjects` in 80%, and `Definition` in 27%. This matters because richer contextual information within the taxonomy

[1] tib-sid/GND/how-to

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

```
{
 "Code": "gnd:4381803-1",
 "Classification Number": "19.3",
 "Classification Name": "Hydrologie, Meereskunde",
 "Name": "TOC",
 "Alternate Name": ["Total organic carbon","Gesamter
     Organischer Kohlenstoff","Abwasseranalyse,
     Kennzahl"],
 "Related Subjects": ["Summenparameter"],
 "Source": "B 1986"
}
```

```
{
 "Code": "gnd:4139622-4",
 "Classification Number": "27.7",
 "Classification Name": "Allgemeine Therapie",
 "Name": "Naturheilverfahren",
 "Alternate Name": ["Biologische Heilweise","
     Naturheilweise"],
 "Related Subjects": ["Erfahrungsheilkunde","
     Naturheilkunde"],
 "Source": "Reallex. Med."
}
```

```
{
 "Code": "gnd:4146660-3",
 "Classification Number": "21.4",
 "Classification Name": "
     Elementarteilchen, Kern-, Atom
     -, Molekularphysik",
 "Name": "Brom-75",
 "Alternate Name": ["Brom 75"],
 "Related Subjects": ["Bromisotop"],
 "Source": "Römpp (9.Aufl.)"
}
```

```
{
 "Code": "gnd:7576879-3",
 "Classification Number": "18",
 "Classification Name": "Natur, Naturwissenschaften allgemein",
 "Name": "Copley-Medaille",
 "Alternate Name": ["Copleymedaille","Copley Medal"],
 "Related Subjects": ["Naturwissenschaften","Preis,Auszeichnung"],
 "Source": "Wikipedia",
 "Definition": "Seit 1731 jährlich von der Royal Society in London
     vergeben, benannt nach Godfrey Copley."
}
```


Figure 1: Four example GND records in our internal JSON representation.

improves subject disambiguation, and thus enhances downstream subject indexing performance.

Figure 1 illustrates four example subject terms from our JSON file: *TOC* (total organic carbon) in hydrology, *Naturheilverfahren* (naturopathy) in medicine, *Brom-75* (bromine isotope) in physics, and *Copley-Medaille* (Copley Medal) in the natural sciences. Each record captures both the preferred term and its variants (e.g., *TOC* lists "Total organic carbon," and *Naturheilverfahren* [naturopathy] lists "Biologische Heilweise" [biological healing method] and "Naturheilweise" [natural healing method]) along with related subjects such as *Summenparameter* [aggregate parameter] or *Naturheilkunde* [natural medicine]. The `Source` field cites the reference from which the term originates (e.g., *B 1986*, *Reallex. Med.*, or *Wikipedia*), while some records also include a short explanatory `Definition`, such as the description of the Copley Medal as an annual scientific award by the Royal Society of London. Thus the structured schema preserves both lexical and contextual information.

We also provide a script[3] to export the GND taxonomy to SKOS (Simple Knowledge Organization System) format, a W3C standard for representing controlled vocabularies in RDF.

Next, we introduce our library records dataset annotated based on the GND. While the records appear in German or English, the GND subject terms themselves are predominantly in German, with English alternate names occasionally listed.

3.2. Our Library Records Dataset

Our library-record corpus is derived from the open-data collection of the TIB – Leibniz Information Centre for Science and Technology. At the time of dataset construction, the TIB catalog comprised ∼5.7M bibliographic records and continued to grow, of which an open dump of ∼200,000 records from the TIB bibliographic holdings (TIBKAT data) was periodically released via the TIB open-data portal (listed under *"The metadata as a dump"* as *TIBKAT data and metadata of freely available electronic collections*). The dump includes metadata from freely available electronic collections and the TIB AV-Portal (including thumbnails), and is released under CC0 1.0, allowing unrestricted reuse. To obtain a machine-learning-ready dataset from this heterogeneous and partially sparse raw dump, we applied four preprocessing steps: (i) language identification using `langdetect`[4], which detected 48 languages in the raw dump (top five: German 108,637; English 76,735; French 1,741; Indonesian 945; Spanish 311)[5], followed by retaining only the two predominant languages, German and English; (ii) removal of records without abstracts; (iii) pruning of infrequent and/or unsuitable record types (e.g., periodicals, chapters); and (iv) removal of records without GND subject annotations, operationalized as missing the `dcterms:subject` tag.

Resultingly, the cleaned collection contains 136,569 records, placing it among the largest bilingual, multi-domain XMTC corpora of cataloged library records. Binned by decade, the corpus is overwhelmingly modern: ∼0.8k (1970s) → 3.2k (1980s) → 8.6k (1990s) → 33.2k (2000s), peaking at 65.8k (2010s) with ∼24.6k so far in the 2020s (2024 partial), indicating rapid mid-2010s growth and a recent taper.

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

[4] https://pypi.org/project/langdetect/

[5] Language tags in the raw dump are frequently missing or noisy; we therefore used automatic language detection. Some records remain mixed-language, but this limited noise reflects real-world metadata and does not hinder system development.

Type/Split	en	de	Total
Article	1,738 (1.27)	10 (0.01)	1,748 (1.28)
train	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)	55,704 (40.79)	100,581 (73.65)
train	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)
Conference	6,032 (4.42)	3,693 (2.70)	9,725 (7.12)
train	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)
Report	2,131 (1.56)	2,509 (1.84)	4,640 (3.40)
train	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)
Thesis	5,764 (4.22)	14,111 (10.33)	19,875 (14.55)
train	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)
Total	60,542 (44.34)	76,027 (55.66)	136,569 (100.00)

Table 1: Record counts and relative shares (%) by split, type, and language for the updated dataset. Blue: type-level shares; Cyan: split-level (Total column); Gray: totals. Cells $\geq$ 15% are shaded.

Table 1 summarizes our released library catalog dataset by record count and percentage of the full collection. It is organized by the five main types of records cataloged at this library, viz. Article (1,748), Book (100,581), Conference (9,725), Report (4,640), and Thesis (19,875). Another level of organization is the language of the records, viz. English (en; 60,542) and German (de; 76,027). To support ML research on subject indexing, we release the dataset with predefined train/dev/test splits balanced by record type and language (90,452 / 19,949 / 26,168 records).

Each library record in our dataset is represented in `json-ld`, the JSON serialization of Linked Data, which encodes bibliographic metadata as machine-readable triples. Most records include both a title and abstract, providing crucial input for machine-learning systems for subject indexing. Across record types, abstract lengths average 100–150 tokens; English abstracts are generally longer than German. Theses are longest in both languages (often > 140 tokens), while Reports are shortest ($\sim$90–110). Other core bibliographic properties such as creator (`dcterms:creator`), contributor (`dcterms:contributor`), publisher (`dc:publisher`), issuing institution and place of publication (`rda:P60163`), and date of issue (`dcterms:issued`) are expressed as persistent URIs using standard vocabularies including `Dublin Core`, `BIBO`, and `RDA`. For each record, annotated GND subject tags are recorded under `dcterms:subject`, linking to controlled concept identifiers in the GND authority file. E.g., the record `dev/Conference/en/3A019447183.jsonld` describes a conference paper authored by *Tim Bedford* (`dcterms:creator` $\rightarrow$ `gnd:171970268`), published by *Oxford University Press* (`dc:publisher`), and linked to subjects such as *Ergodentheorie* (`dcterms:subject` $\rightarrow$ `gnd:4015246-7`) and *Hyperbolische Geometrie* (`gnd:4161041-6`). Similarly, the record `dev/Thesis/de/3A011101717.jsonld` represents a doctoral dissertation by *Reinhard O. Greiling* (`dcterms:creator` $\rightarrow$ `gnd:1090799454`), published by *Lang Verlag* (`dc:publisher`) in *Frankfurt am Main* (`rda:P60163`), with subject links to *Schwarzschiefer* (`dcterms:subject` $\rightarrow$ `gnd:4126782-5`) and *Silur* (`gnd:4181434-4`). This representation supports interoperable bibliographic and semantic processing across linked-data resources.

How are the records cataloged? A team of 17 subject specialists at the TIB covers a predefined set of 28 domains and assigns subject annotations to records in the TIB national library catalog, ensuring broad and expert-curated coverage. Domain assignment is usually semi-automatic: new records often arrive with domain metadata through the national library network; otherwise, an in-house AUTNIF instance is used (Suominen, 2019; Suominen et al., 2023). Records may carry multiple domains (range 1–7; mean 1.5, with 7 as an outlier). The distribution is top-heavy: Social Sciences ($\sim$24k), Economics ($\sim$24k), and Educational Science ($\sim$18k) dominate, alongside solid STEM representation (e.g., Computer Science $\sim$13k; Mathematics $\sim$12k) and a notable "Other" bucket ($\sim$5k). Specialized areas form a long tail (e.g., Mining 166; Medical Technology 589; Sports Science $\sim$1.3k; Materials Science $\sim$1.7k), indicating domain imbalance.

In practice, libraries index content with controlled vocabularies; in Germany, the GND is used for subject cataloging. Subject librarians assign GND terms from titles, abstracts, and—where available—full text, in a cooperative workflow across institutions. This work is done in a cooperative process in various libraries and in different national library networks. Given sustained growth ($\approx$15k newly indexed titles/month), this work is substantial; NLP/AI offers clear potential to support it. In our dataset, records have on average three GND subjects; the range is 1–39 (the upper extreme is rare). Table 2 illustrates five example records with their domain and subject annotations.

3.3. Statistical Analysis of Subject Annotations

Sourced from an actual public library, this dataset reflects real-world practice and long-term quality constraints, shaped by evolving staff and workflows. Nevertheless, it offers a valuable resource for building reliable AI tools for librarians—tools designed

Record (title) (Linked)	Lang	Type	Domains	Subjects
Chapter 86: Explaining the Comparative Statics in Step-Level Public Good Games	en	Article	Economics	Experiment; Wirtschaftswissenschaften; Wirtschaftsforschung; Methodologie; Experimentelle Wirtschaftsforschung
Das Beil von Wandsbek: Roman; 1938–1943	de	Book	Literature Studies	Nationalsozialismus
Window on Freedom: Race, Civil Rights, and Foreign Affairs, 1945–1988	en	Book	Social Sciences; History	Außenpolitik; Rassendiskriminierung; Menschenrecht; Schwarze
Sicherung des Familieneinflusses in Familienunternehmen: Symposium . . . (6./7. Okt. 2016)	de	Conf.	Law	Einfluss; Familienbetrieb; Familie
Charge Carrier Recombination and Open Circuit Voltage in Organic Solar Cells: From Bilayer Model Systems to Hybrid Multijunctions	en	Thesis	Electrical engineering; Physics	Organische Solarzelle; Reihenschaltung; Mehrschichtsystem; Fluor; Donator (Chemie); Dotierung; Rekombination

Table 2: Representative examples of multi-domain, multi-subject annotations from the library records dataset. Only the `title` metadata are shown for compactness; links point to the full JSON-LD records.

around corpus-level patterns and principled reasoning rather than idiosyncratic particularities. Thus, in this section, we statistically characterize the subject space—quantifying split overlap and long-tail sparsity, measuring distributional divergence (KL, JSD, and χ^2), and assessing polysemy—to surface implications for model design and evaluation.

3.3.1. Overlap and Long-Tail Phenomenon

From the nearly 200,000 subjects in the GND *Sachbegriff*, 41,218 unique subjects appear in our dataset (per-split counts released here). The annotations span traditional subjects (*Literatur*, *Architektur*, *Philosophie*) and contemporary ones (*Digitalisierung*, *Nachhaltigkeit*, *Künstliche Intelligenz*); 6,164 subjects occur in *all three* splits, evidencing pronounced long-tail sparsity in which high-frequency labels (e.g., *Literatur*, *Architektur*, *Unternehmen*) coexist with specialized ones (e.g., *Robotik*, *Bioinformatik*, *Feminismus*), requiring models to handle few-shot and zero-shot generalization rather than rely on balanced head classes. Overall, the corpus bridges humanities, social sciences, and technical domains but is dominated by a long tail that should inform training and evaluation design. The Jaccard (1912) and weighted Jaccard (Tanimoto 1958) in Table 3 quantify split overlap: Jaccard gives the fraction of shared subjects (here ≈ 0.36, i.e., about one third), while the weighted variant also accounts for subject frequencies. Lower Tanimoto reflect that even shared subjects occur at different rates, implying models must handle sparse labels and distribution shift.

Metric	Train–Dev	Train–Test	Dev–Test
Jaccard	0.3687	0.3608	0.3668
Tanimoto	0.2064	0.2530	0.3771

Table 3: Jaccard and Tanimoto scores showing subject overlap across dataset splits.

3.3.2. Distributional Divergence

To move beyond overlap, we quantify frequency divergence across splits with KL divergence (Kullback and Leibler, 1951), Jensen–Shannon divergence (JSD; link), and the Chi-Squared (χ^2) test (Pearson, 1900). KL spans 2–5 nats—$\mathrm{KL}(\text{Train}|\text{Dev})=4.34$, $\mathrm{KL}(\text{Train}|\text{Test})=4.69$ vs. $\mathrm{KL}(\text{Dev}|\text{Train})=1.99$, $\mathrm{KL}(\text{Test}|\text{Train})=1.89$—indicating Train covers a broader subject set while Dev/Test are narrower reweightings. JSD indicates moderate shifts (0.16 Train–Dev, 0.18 Train–Test, 0.26 Dev–Test). χ^2 statistics of ≈ 37k–49k with $p<0.001$ confirm differences are systematic rather than random variation. Collectively, the corpus exhibits partial overlap but clear distributional drift. Methodologically, models must handle uneven frequencies and underrepresented topics—favoring regularization, calibration, and shift-aware validation and reporting.

3.3.3. Assessing Polysemy

To assess potential polysemy, we operationalize it as string identity shared across distinct GND identifiers and scan the taxonomy accordingly. Restricting to *preferred labels*, such cases are exceedingly rare—65 occurrences (0.03%) among 207,001 codes; for example, *Alakaluf* is the preferred label for both `gnd:1071000497` (ethnographic group, Kawésqar) and `gnd:4001008-9` (broader folk-cultural classification), indicating parallel cataloging rather than genuine sense ambiguity. Including *alternate labels* raises the incidence to 2,181 (0.52%): *Abwasseranalyse,Kennzahl* ("wastewater analysis, indicator") spans seven entries—`gnd:4145594-0` (*Biochemischer Sauerstoffbedarf*), `gnd:4147637-2` (*Chemischer Sauerstoffbedarf*), `gnd:4185748-3` (*Totaler Sauerstoffbedarf*), `gnd:4360165-0` (*DOC*), `gnd:4381803-1` (*TOC*), `gnd:4586442-1` (*Pges. ICP*), `gnd:4586443-3` (*TNb*)—each

denoting a measurement indicator; similarly, *SPSS,WINDOWS,Programm* appears across five records for successive software versions. These findings suggest that duplicate strings predominantly reflect terminological reuse across related entities rather than true polysemy; consequently, alternate labels broaden retrieval but should not be treated as interchangeable with preferred labels in downstream modeling and evaluation.

We now go beyond exact name matches and use semantic embeddings to surface near-duplicates and potential sense conflations. Before encoding, we exclude acronym/code-like labels (fragile under distributional similarity), leaving $N = 203{,}763$ subjects from $207{,}001$. We compare three embedding families (Table 4): Google's *EmbeddingGemma-300m* (compact multilingual, general-purpose representations) (Vera et al., 2025), *multilingual-E5-small* (contrastive, retrieval-oriented) (Wang et al., 2024), and *jina-v2-base-de* (German-focused, high-precision bilingual encoder) (Mohr et al., 2024). In all cases we construct a graph where an undirected edge links two subjects whose cosine similarity exceeds 0.90; degree ≥ 2, i.e. if a subject node has more than 2 similar terms, is a more pronounced signal of possible polysemy than isolated pairs.

Model	View	N(≥1)	N(≥2)	Mean	Max
Gemma-300m	name	∼19.8K	∼4.0K	0.15	83
	context	∼15.7K	∼3.6K	0.13	75
E5-small	name	∼195.7K	∼188.1K	38.77	1.49K
	context	∼165.8K	∼147.0K	24.62	800
Jina-v2-DE	name	∼15.0K	∼2.8K	0.11	36
	context	∼14.5K	∼3.7K	0.13	53

Table 4: Semantic-similarity graph at cosine ≥ 0.90 (topK$= 50$), $N = 203{,}763$ subjects. N(≥1): count of subjects that have at least one similar neighbor; N(≥2): subjects with at least two similar neighbors (stronger polysemy signal).

Results in Table 4 show clear differences across encoders. E5 produces very dense neighborhoods (name view: $\sim$ 195.7K nodes with neighbors; mean degree 38.77; max 1.49K), which is useful for retrieval but tends to overstate polysemy by grouping topical associates and word-family variants rather than only near-duplicates. This is visible in concrete cases: *Vergrößerung* links to 76 items mixing near-synonyms (*Erweiterung, Erhöhung*), antonyms (*Verkleinerung*), derivational relatives (*Verfilmung, Versilberung*), and broad associates (*Darstellung, Durchmesser*); *Zeit* pulls a long tail of compounds/contexts (*Zeitmessung, Zeitraum, Zeitalter, Zeitzone, Zeitarbeit, . . .*); *Differenzierung* balloons to 160 neighbors across mathematics, sociology, and linguistics. By contrast, Gemma and Jina yield much sparser graphs (mean degree ≈ 0.11–0.15, max ≤ 83), aligning better with our goal of flagging only very similar terms. Their links are tight: Gemma connects *Zukunft* → *Futur*, *Nutzung* → *Benutzung*/*Nutzungseignung*, *Instrument* → *(Musik)instrument*, with occasional orthographic/morphemic effects (*Vergrößerung* → *Vergröberung*; *Zeit* → *Zeitmaß*/*Zeitmittel*/*Saatzeit*). Jina is stricter still at 0.90, mostly surfacing crisp pairs (*Nutzung* → *Benutzung*; *Differenzierung* → *Dedifferenzierung*; *Instrument* → *Musikinstrument*/*A-Instrument*) while many high-level nouns (*Verhandlung, Entwicklung, Werk, Ziel*) have degree 0—behaviour that is desirable if the aim is to flag only plausible sense conflations. For the rows against the view *context* in Table 4, we embed a compact string that joins the preferred label with short definitional cues and a capped list of alternates (`Name [SEP] DEF: . . . [SEP] ALT: . . .`). For Gemma and E5 this generally sharpens meanings and prunes superficial, name-only links, whereas for Jina it sometimes consolidates genuine German paraphrases into small, tighter clusters—exactly the kind of behavior desired for a high-precision polysemy screen. Overall, polysemy in the GND appears rare: with conservative encoders (Gemma/Jina) at a strict 0.90 cosine, only ≈1–2% of subjects participate in multi-item clusters (degree ≥ 2) while about 90–93% have no near neighbor at all—indicating that genuine ambiguity is exceptional rather than pervasive.

4. Three Systems

To illustrate how the dataset can be used for automated subject indexing, we report results from three representative systems evaluated in the LLMs4Subjects shared tasks at SemEval 2025 (D'Souza et al., 2025) and GermEval 2025, both of which used this dataset. The task is formulated as follows: given a record consisting of a title and abstract, a system should retrieve up to 20 relevant subjects from the GND taxonomy, ranked in descending order of confidence. This constitutes an information retrieval (IR) problem with ranked outputs, evaluated over the top-k predicted subjects (Schütze et al., 2008). As the primary evaluation metric, we use nDCG@k (Normalized Discounted Cumulative Gain) (Järvelin and Kekäläinen, 2002), a rank-based measure that assesses how closely a system's predicted subject ranking matches the ground truth. Unlike recall@k, which measures only how many relevant subjects appear among the top-k predictions, nDCG@k additionally considers their *positions* in the ranking, rewarding correct subjects that appear higher in the list through a logarithmic discount. The score is normalized between 0 and 1, where 1 is a perfect ranking with all relevant subjects at the top.

In practice, this task is commonly addressed using XMTC methods. XMTC methods for controlled-vocabulary subject indexing typically combine scalable candidate generation over very large label spaces with a ranking stage optimized for top-k predictions. Early approaches relied on sparse one-vs-rest linear classifiers such as DiSMEC (Babbar and Schölkopf, 2017) and PD-Sparse/PPD-Sparse (Yen et al., 2016, 2017), which train independent linear models for each label while exploiting sparsity and distributed optimization to scale to very large label sets. Tree-based methods such as FastXML (Prabhu and Varma, 2014), Parabel (Prabhu et al., 2018), and Bonsai (Khandagale et al., 2020) improved efficiency by organizing labels into hierarchical partitions, allowing inference to focus on a much smaller candidate subset at prediction time. Embedding and nearest-neighbor approaches such as SLEEC (Bhatia et al., 2015) and AnnexML (Tagami, 2017) instead learn low-dimensional representations for instances and labels, reframing XMTC as a semantic retrieval problem in which relevant labels are recovered via nearest-neighbor search. More recent neural methods, including XML-CNN (Liu et al., 2017), AttentionXML (You et al., 2019), X-Transformer (Chang et al., 2020), and XR-Transformer (Zhang et al., 2021), combine deep text encoders with label shortlisting to improve contextual semantic matching while preserving scalability. For library subject indexing, these families of methods can be understood as variants of a common retrieve-and-rank design pattern: first narrowing the controlled vocabulary to plausible candidate subjects, then estimating which labels are most appropriate for final assignment. Toolkits such as PECOS (Yu et al., 2022) make this connection explicit by unifying hierarchical indexing, learned matching, and ranking within modular *index* $\rightarrow$ *match* $\rightarrow$ *rank* pipelines.

Given this task formulation and methodological landscape, we focus on three representative systems from the LLMs4Subjects shared tasks: LA^{2}I^{2}F (Salfinger et al., 2025), KIFSPrompt (Kähler et al., 2025), and Annif (Suominen et al., 2025). Together, they capture key design patterns that emerged in the shared tasks, from prompting-based pipelines to hybrid trained XMTC systems. Beyond these three, the broader submissions covered a diverse methodological space, including retrieval-only pipelines that transfer subjects from similar indexed records (Tian et al., 2025a; Singh et al., 2025), bi-encoder plus cross-encoder reranking pipelines that first retrieve candidates and then rescore them more precisely (Dorkin and Sirts, 2025), multilingual BERT ensembles that combine predictions from several language-specific or multilingual encoders (Hahn, 2025), contrastively fine-tuned embedding models that explicitly pull relevant record–subject pairs closer in representation space (Jiang et al., 2025), Burst Attention–enhanced retrieval methods that refine embeddings by modeling interactions across embedding dimensions before top-k retrieval (Islam et al., 2025), RAG-style subject selection pipelines that retrieve candidate subjects and then use an LLM to verify or rank them in context, leveraging the OntoAligner Python toolkit (Tekanlou et al., 2025; Babaei Giglou et al., 2025), and LLM finetuning pipelines with synthetic data generation and preference optimization to better align outputs with human indexing behavior (Tian et al., 2025b; Ho, 2025). Across these submissions, strong performance was typically associated with multi-stage architectures rather than single-model prediction alone, with common strategies including candidate retrieval followed by reranking, model ensembling, multilingual processing, and LLM-assisted augmentation or refinement. The three systems presented below were selected not because they exhaust this design space, but because they provide clear, interpretable exemplars of its main families and thus serve well as reference points for understanding the dataset.

4.1. Approaches

System 1 (Salfinger et al., 2025). Library corpora commonly exhibit long-tailed subject distributions: a few high-level subjects occur frequently, while many fine-grained subjects appear rarely. Models trained directly on such data tend to overpredict frequent subjects and overlook rare but meaningful ones. To mitigate this, system 1 reframes multi-label classification as semantic retrieval in a shared 768-dimensional embedding space using sentence-transformers/all-mpnet-base-v2 (Song et al., 2020). Both training records (title + abstract) and subject labels (names + alternates) are embedded into the same space to enable semantic comparison with a query record. The method combines two complementary strategies: *ontological reasoning*, which compares the query embedding directly to subject embeddings, and *analogical reasoning*, which retrieves similar training records and transfers their assigned subjects. Candidate subjects from both sources are merged, deduplicated, and ranked by embedding distance, with the top results predicted.

System 2 builds upon system 1's analogical reasoning thread with a few-shot retrieval strategy.

System 2 (Kähler et al., 2025). This approach implements a four-stage pipeline combining retrieval-augmented few-shot prompting with controlled vocabulary mapping and ranking. *Retrieve.* The input record is embedded with the multilingual BGE-M3 model (Chen et al., 2024), and a Weaviate vector

Team Name	nDCG@k			
	k=5	*k=10*	*k=15*	*k=20*
System 1	0.3639	0.3977	0.4143	0.4247
System 2	0.4919	0.4880	0.4879	0.4879
System 3	0.6020	0.6391	0.6560	0.6652

Table 5: nDCG@k scores at different ranked cutoffs for three systems on our subject indexing dataset.

store retrieves the L most similar training documents in the same language. *Complete.* These examples, together with their GND subject annotations, are inserted into an on-the-fly few-shot prompt for the Ministral-8B-Instruct model, which generates free-form keyword suggestions. *Map.* The generated keywords are embedded with BGE-M3 and mapped to GND subjects using hybrid HNSW + BM25 search (Robertson et al., 2009; Malkov and Yashunin, 2018), aligning free vocabulary to the controlled taxonomy. *Rank.* Finally, Llama-3.1-8B-Instruct (Dubey et al., 2024) assigns each mapped term a 0–10 relevance score used to normalize and rank the final subject list. Like system 1, this pipeline relies entirely on off-the-shelf models without fine-tuning and performs only a single generation and ranking pass per record.

Finally, in contrast to systems 1 and 2, system 3 combines LLMs with traditional XMTC algorithms via the ANNIF toolkit (Suominen, 2019).

System 3 (Suominen et al., 2025). This hybrid approach uses LLMs for data preprocessing and final reranking while relying on XMTC models for subject prediction. *Translate.* Records and vocabularies are first translated into monolingual English and German collections using google/gemma-3-4b-it and CohereLabs/aya-expanse-8b, respectively. *Synthesize.* Several LLMs then generate synthetic training data, producing four additional records per original entry in both languages. *Train and predict.* Two monolingual ensembles (English and German) are trained using three Annif backends: Omikuji Bonsai (Khandagale et al., 2020) for partitioned label-tree classification, MLLM (Maui-like (Medelyan, 2009) lexical matching) for text–term matching, and XTransformer, a transformer-based XMTC ranking model within the PECOS framework (Yu et al., 2022). Each backend is trained separately on monolingual datasets created via LLM translation and combined into language-specific ensembles. *Merge and rerank.* Finally, predictions are reranked using mistralai/Mistral-Small-3.1-24B-Instruct-2503, and bilingual outputs are merged to produce the final ranked list of subjects.

4.2. Quantitative Results

Table 5 presents the quantitative results on our released test set for the three systems. Overall, the heavily engineered System 3 proved most effective, while System 2 showcased a clever extension of System 1's ideas by using dynamically retrieved few-shot examples to elicit strong task performance from LLMs without training. In practical library settings, the usefulness of AI assistance depends on having the most relevant subjects appear early in a concise prediction list. Based on librarian feedback, we set the optimal list length to $k = 20$, balancing coverage and efficiency. Figure 2 and Figure 3 show performance by record type and language at nDCG@5. System 1 performed best on article records, while System 3 led across the other four types. By language, System 1 was strongest on English records, whereas Systems 2 and 3 performed better on German. Overall, System 2 establishes a baseline for untrained LLM-based methods (nDCG@5 = 0.4919), and System 3 sets the benchmark for hybrid, model-trained approaches (nDCG@5 = 0.6020).

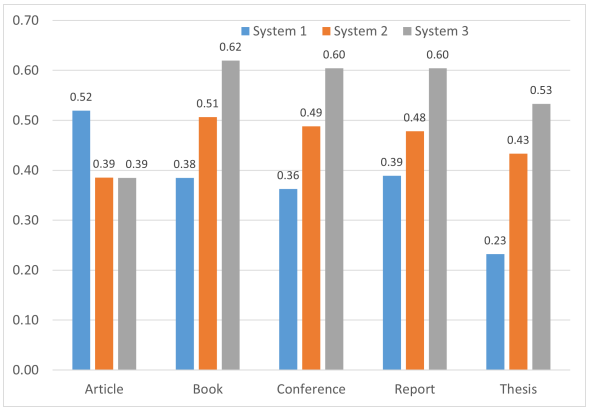


Figure 2: nDCG@5 scores by the five record types.

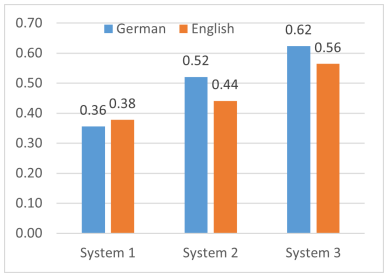
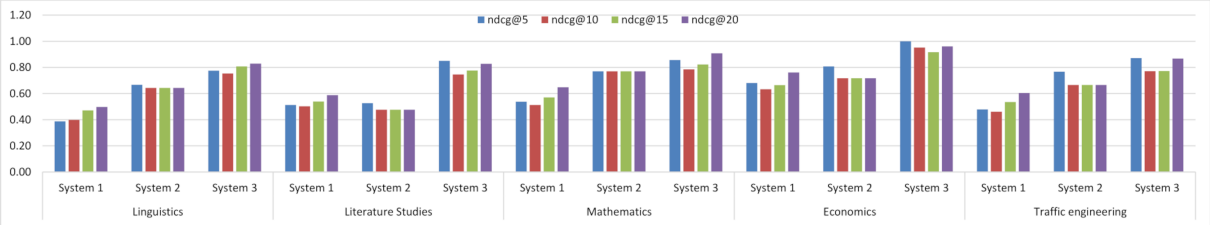


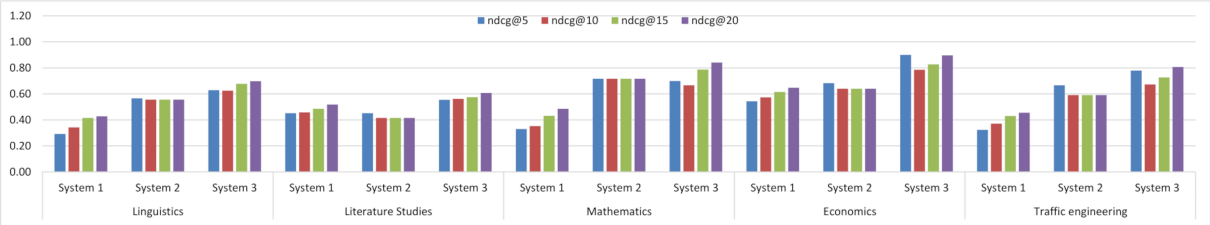
Figure 3: nDCG@5 scores by the two languages.

4.3. Qualitative Results

We asked subject librarians to manually assess outputs from the three systems as a form of human spot-checking (Figure 4). Ten random test records across five domains—Linguistics, Literature Studies, Mathematics, Economics, and Traffic Engineering—were each reviewed by a specialist. For every predicted subject, librarians marked Y



(a) Case 1 - Treating both Y (correct) and I (irrelevant but technically correct) subject predictions as correct.



(b) Case 2 - Treating only Y (correct) subject predictions as correct.

Figure 4: nDCG@k scores, where k=5,10,15,20, for qualitative evaluation on 10 records per five domains. For this exercise, subject predictions were manually labeled Y and I by subject specialists at the library.

(correct), I (technically correct but irrelevant), or N (incorrect). Figure 4a shows nDCG scores at k=5,10,15,20 when Y and I are treated as correct (case 1), while Figure 4b shows results when only Y counts as correct (case 2).

For case 1 and 2, the relative domain-wise nDCG@20 rankings show consistent patterns across systems. System 1 performed best in *Economics*, followed by *Mathematics* and *Literature Studies*, while *Traffic Engineering* lagged behind. System 2 achieved its strongest results in *Mathematics* and *Economics*, showing generally balanced performance across domains. System 3 outperformed both, reaching near-perfect scores in *Economics* and *Mathematics* and maintaining strong results across all domains, reflecting the advantage of its hybrid learning and ensemble design. The differences between case 1 (Y + I counted) and case 2 (Y only) reveal that many high-ranked subjects were *contextually related* but not exact matches—indicating that systems, especially LLM-based ones, capture topical proximity well but still struggle to distinguish semantically distinct subjects within a record. This highlights a key challenge for future models: promoting conceptual diversity among top-ranked predictions rather than clustering around near-synonyms.

A closer look at System 1's output for these spot checked records revealed that the analogical reasoning branch dominated due to its similarity computation: embeddings of training records yielded more similar distances to test records than embeddings of GND subject terms. Consequently, correctly identified subjects from the ontological branch were often ranked lower. Another common error occurred when the analogical branch transferred *all* subjects from a similar training record, even though only some applied. For example, in *"Sprachwissenschaft im Fokus"*, the system inferred *Englisch*, *Spanisch*, and *Romanische Sprachen* merely because they co-occurred with the correct label *Deutsch* in the retrieved record, leading to many false positives and underscoring the need for filtering contextually relevant subjects. For System 2, most errors arose in the *Map* component linking LLM-generated keywords to GND terms, including mapping errors due to unknown or ambiguous concepts in the GND; and correctly mapped but irrelevant yet technically valid suggestions (category I). Overall, the system tended to produce plausible but overly broad terms rather than precise subject matches, which was also the case in many of system 3 errors.

5. Conclusion

This paper introduced an XMTC dataset of library records paired with a rich subject taxonomy, offering a novel resource for subject indexing research. We presented three complementary approaches. The best scoring system relied heavily on traditional machine learning, raising questions about the applicability and generalizability of purely LLM-based approaches, which remain largely untapped. Our dataset provides a challenging benchmark for assessing how LMs capture the nuanced semantics required for library indexing. Future work will benchmark multilingual embeddings, explore small LMs as efficient alternatives, and investigate LM distillation for practical deployment. It also supports evaluating subject indexing by hierarchy, transparency, and usefulness in real library work.

6. Bibliographical References

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, and Vivien Petras. 2016. White paper on best practices for multilingual access to digital libraries. Technical report, Europeana Foundation. Open access white paper.

Hamed Babaei Giglou, Jennifer D'Souza, Oliver Karras, and Sören Auer. 2025. Ontoaligner: A comprehensive modular and robust python toolkit for ontology alignment. In *European Semantic Web Conference*, pages 174–191. Springer.

Rohit Babbar and Bernhard Schölkopf. 2017. Dismec: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 721–729.

K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse local embeddings for extreme multi-label classification. *Advances in neural information processing systems*, 28.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1):2013.

Caterina Caracciolo, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. 2013. The agrovoc linked dataset. *Semantic Web*, 4(3):341–348.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3163–3171.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Corine Deliot. 2014. Publishing the british national bibliography as linked open data. *Catalogue & Index*, 174:13–18.

Maria Dermentzi, Mike Bryant, Herminio García-González, and Fabio Rovigo. 2025a. Ehri multilingual subject indexing test dataset.

Maria Dermentzi, Mike Bryant, Fabio Rovigo, and Herminio García-González. 2025b. Multilingual Automated Subject Indexing: a comparative study of LLMs vs alternative approaches in the context of the EHRI project. Working paper or preprint.

Aleksei Dorkin and Kairit Sirts. 2025. Tartunlp at semeval-2025 task 5: Subject tagging as two-stage information retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2449–2454.

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. SemEval-2025 task 5: LLMs4Subjects - LLM-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2570–2583, Vienna, Austria. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

FAO. 2023. Introducing agris 2.0! Online news article. Accessed: ‹add your access date›.

Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Darryl Estrada-Zavala, Renato Toshiyuki Murasaki, Elena Primo-Peña, Cristina Bojo Canales, Georgios Paliouras, Martin Krallinger, et al. 2021. Overview of bioasq 2021-mesinesp track. evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. *CEUR Workshop Proceedings*.

Jim Hahn. 2025. Jim at semeval-2025 task 5: Multilingual bert ensemble. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2407–2412.

Clara Wan Ching Ho. 2025. UBFFM at the GermEval-2025 LLMs4Subjects task: What if we take "you are an expert in subject indexing" seriously? In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 471–478, Hannover, Germany. HsH Applied Academics.

Antoine Isaac and Bernhard Haslhofer. 2013. Europeana linked open data–data. europeana. eu. *Semantic Web*, 4(3):291–297.

Baharul Islam, Nasim Ahmad, Ferdous Barbhuiya, and Kuntal Dey. 2025. Nbf at semeval-2025 task 5: Light-burst attention enhanced system for multilingual subject recommendation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 953–958.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Hong Jiang, Jin Wang, and Xuejie Zhang. 2025. Ynu-hpcc at semeval-2025 task 5: Contrastive learning for gnd subject tagging with multilingual sentence-bert. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2443–2448.

Maximilian Kähler, Lisa Kluge, and Katja Konermann. 2025. DNB-AI-project at the GermEval-2025 LLMs4Subjects task: KIFSPrompt - knowledge-injected few-shot prompting. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 455–464, Hannover, Germany. HsH Applied Academics.

Sujay Khandagale, Han Xiao, and Rohit Babbar. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning*, 109(11):2099–2119.

Anastasia Krithara, James G. Mork, Anastasios Nentidis, and Georgios Paliouras. 2023a. The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. *Frontiers in Research Metrics and Analytics*, 8.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023b. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Olena Medelyan. 2009. *Human-competitive automatic topic indexing*. Ph.D. thesis, The University of Waikato.

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, et al. 2024. Multi-task contrastive learning for 8192-token bilingual text embeddings. *arXiv preprint arXiv:2402.17016*.

NLM. 2000. *Medical subject headings*, volume 41. US Department of Health and Human Services, Public Health Service, National Library of Medicine.

Hercules Panoutsopoulos and Christopher Brewster. 2022. Data-driven update of agrovoc using agricultural text corpora. In *HAICTA*, pages 260–265.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, pages 993–1002.

Yashoteja Prabhu and Manik Varma. 2014. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Andrea Salfinger, Luca Zaccagna, Francesca Incitti, Gianluca De Nardi, Lorenzo Dal Fabbro, and Lauro Snidaro. 2025. LA²I²F at SemEval-2025 task 5: Reasoning in embedding space – fusing analogical and ontology-based reasoning for document subject tagging. In *Proceedings of the 19th International Workshop on Semantic*

Evaluation (SemEval-2025), pages 2413–2423, Vienna, Austria. Association for Computational Linguistics.

Gauri Salokhe, Irene Onyancha, James Weinheimer, Barbara Richards, Fynvola Le Hunte Ward, and Johannes Keizer. 2023. Agris 2.0: A more flexible, multilingual bibliographic platform. Food and Agriculture Organization of the United Nations (FAO), Knowledge Exchange and Capacity Building Division. Accessed: <add your access date>.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Sumit Singh, Pankaj Goyal, and Uma Tiwary. 2025. silp_nlp at semeval-2025 task 5: Subject recommendation with sentence transformer. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2455–2460.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Osma Suominen. 2019. Annif: Diy automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1):1–25.

Osma Suominen, Juho Inkinen, and Mona Lehtinen. 2025. Annif at the GermEval-2025 LLMs4Subjects task: Traditional XMTC augmented by efficient LLMs. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 447–454, Hannover, Germany. HsH Applied Academics.

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

Yukihiro Tagami. 2017. Annexml: Approximate nearest neighbor search for extreme multi-label classification. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 455–464.

T.T. Tanimoto. 1958. *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation.

Hadi Bayrami Asl Tekanlou, Jafar Razmara, Mahsa Sanaei, Mostafa Rahgouy, and Hamed Babaei Giglou. 2025. Homa at semeval-2025 task 5: Aligning librarian records with ontoaligner for subject tagging. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2400–2406.

Xia Tian, Yang Xin, Wu Jing, Xiu Heng, Zhang Xin, Li Yu, Gao Tong, Tan Xi, Hu Dong, Chen Tao, et al. 2025a. Ruc team at semeval-2025 task 5: Fast automated subject indexing: A method based on similar records matching and related subject ranking. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2437–2442.

Yicen Tian, Erchen Yu, Yanan Wang, Dailin Li, Jiaqi Yao, Hongfei Lin, Linlin Zong, and Bo Xu. 2025b. Dutir831 at semeval-2025 task 5: A multi-stage llm approach to gnd subject assignment for tibkat records. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 363–372.

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, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *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, et al. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Ian EH Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit Dhillon, and Eric Xing. 2017. Ppdsparse: A parallel primal-dual sparse method for extreme classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 545–553.

Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. 2016. Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification.

In *International conference on machine learning*, pages 3069–3077. PMLR.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in neural information processing systems*, 32.

Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. 2022. Pecos: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98):1–32.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. 2021. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280.

Ruohong Zhang, Yau-Shian Wang, Yiming Yang, Donghan Yu, Tom Vu, and Likun Lei. 2023. Long-tailed extreme multi-label text classification by the retrieval of generated pseudo label descriptions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1092–1106.

A. System 1 Detailed Error Analysis

In the following section, we present a more detailed analysis of the different systematic types of errors made by System 1 (LA^{2}I^{2}F) (Salfinger et al., 2025) on the test data set described in subsection 4.3. In this qualitative analysis, subject-matter librarians manually rated the system's outputs by checking the 20 predicted labels for each of 50 test documents, covering 161 gold-standard subjects. Each prediction was marked as either Y (correct), I (technically correct but irrelevant), or N (incorrect), allowing us to analyze systematic error patterns in greater depth.

False Negatives (Missed Ground Truth Labels). In total, System 1 achieved a *False Negative Rate* (FNR) of 56.5%, i.e., the share of missed ground-truth subject labels, when considering $k = 20$ predicted subjects per document, and an FNR of 73.2% when considering only the system's top-5 ranked predictions. However, due to the sheer size of the GND ontology and the human-tagged training corpus, the specific selection of subject labels assigned to a document can be somewhat subjective and thus biased by the tagging expert, especially if synonymous concepts are present in the ontology. To account for such potential ambiguities when evaluating the subject labels missed by the system's predictions, we dissect our analysis of *False Negatives* (FNs), i.e., ground-truth labels missed by the system, according to the following categories:

FNM: aspect completely missed by the system

FNC: aspect missed, but closely matched by other predictions

By manually checking System 1's top-5 ranked predictions only, we obtain the following absolute counts on the 50 test set records using these definitions:

Error Category	Absolute Count
FNM	79
FNC	41
False Negatives (total)	120

Table 6: Distribution of false negatives for 250 subject suggestions from System 1 (5 predicted labels per test set document) on 50 test-set documents with 161 gold-standard subjects.

As this fine-grained manual inspection, factoring in synonymous subjects, reveals, 34.1% of System 1's FNs did not retrieve the exact ground-truth label but instead a closely matching synonym, while 65.8% correspond to missing ground-truth concepts that are completely absent from System 1's top-5 predictions.

When digging further into the reasons behind this, we identify the following systematic error category introduced by System 1: In System 1's implemented fusion strategy, *analogical reasoning dominates*. Analyzing the contributions from System 1's two reasoning branches to the fused results, we observe that the analogical branch drives the results: document-to-document distances tend to be smaller than document-to-subject distances. Subjects contributed from the ontological branch thus are typically ranked behind those from the analogical branch and therefore often do not make it into the fused list of top-k ranked subjects returned, as illustrated by the test set record shown in Table 7 (its gold-standard predictions can be found in Table 9). In this case, the ground-truth label *AWACS* was predicted by the ontological branch at rank 22, but got outranked in the fusion process by the smaller document distances, i.e., higher cosine similarities, of the analogical branch, resulting in a correct prediction being filtered out in the merging process. This indicates the need for developing a more sophisticated fusion strategy, which could include another downstream LLM-based relevancy filtering and re-ranking component, similar to System 2.

False Positives (Incorrect Predictions). An additional downstream relevancy filtering component will also be needed to address another inherent

error category identified: System 1 currently implements an overly optimistic assumption for analogical reasoning, considering *all* subject tags from the closest training documents in embedding space as appropriate tags for the new document. We found this to be the dominant inherent error source for introducing false predictions not matching the new document, i.e., *False Positives* (FPs), in System 1. While related documents might overlap in their subject labels, not necessarily *all* subjects from document A might be appropriate for a similar document B.

Error Categories. In summary, we denote the dominant identified systematic error categories as:

FPAR FP introduced by analogical reasoning

ORoR Correct result from ontological reasoning outranked after fusion with analogical results

Table 7 illustrates some concrete examples of these systematic errors on test set record ID `3A1007389885`.

Prediction	R	Rank	Sim.	Eval.	Error Category
Elektronische Gegenmaßnahme [electronic countermeasure] `gnd:1082384615`	A	1	0.687	N	FPAR
Störsender [jammer] `gnd:4273774-6`	A	2	0.687	Y	– (TP, Y)
Radar [radar] `gnd:4176765-2`	A	3	0.687	Y	– (Y)
Sonar [sonar] `gnd:4181785-0`	A	4	0.652	N	FPAR
Signalverarbeitung [signal processing] `gnd:4054947-1`	A	5	0.652	Y	– (Y)
AWACS [airborne warning and control system] `gnd:4309079-5`	O	22	0.406		ORoR

Table 7: System 1's predictions for the test set record with ID `3A1007389885`, *"Recent advancements in airborne radar signal processing: emerging research and opportunities"*, dissecting which reasoning branch (**R**) identified each predicted label – analogical (A) or ontological (O) reasoning. The first five results are the top-5 predictions returned by the fusion system after merging the results from both A and O branches; the last row shows an actual ground-truth subject excluded from the retrieved list due to being outranked. **Rank** denotes the ranking of each subject within each reasoning branch (analogical vs. ontological), **Sim.** shows the cosine similarity to the query document in embedding space determining the ranking, and **Eval.** lists the human expert's judgment of the predicted label.

In conclusion, this analysis confirms the hypothesized complementarity of System 1's fusion architecture, with both reasoning branches identifying complementary information, and identifies avenues for future work. The systematic error sources identified can be tackled by introducing additional filtering components for evaluating and re-ranking the identified (candidate) subjects, which should up-rank matching predictions contributed from the ontological branch and eliminate FPs introduced by the analogical branch.

B. System 2 Detailed Error Analysis

Manual inspection of the system output reveals certain repeatedly occurring error patterns. In a close examination of the 50 test documents that were also rated by the subject experts during qualitative evaluation, we observe 176 subject terms suggested by System 2 (KIFSPrompt) (Kähler et al., 2025) in total, including 110 false positives that can be classified into a variety of subcategories. The entire data sheet for the qualitative error analysis can be found at `https://github.com/sciknoworg/tib-sid/tree/main/evaluation/results/system2%20-%20additional%20analysis`.

Forty of the false positive suggestions were rated by the subject experts as relevant, and 27 suggestions were rated as not relevant but technically correct. This leaves 43 false positive suggestions (39%) that should be considered truly erroneous. It is interesting to study how the system came to make these truly erroneous suggestions. Indeed, some of the errors can be attributed to issues occurring during the mapping stage of the system, where free keywords get matched to normalized subject terms. We observe two recurring patterns:

MEU: mapping error, because a concept is unknown in the GND

MEA: mapping error, because a concept is ambiguous in the GND

Let us illustrate these categories with examples. In our sample, a document titled *"The Arden research handbook of Shakespeare and adaptation"* is tagged with the gold-standard subject term *Adaption (Literatur)* (gnd:102289935X). The LLM suggested simply *Adaption*, which is used as an alternative label for the GND subject term *Anpassung* (gnd:4128128-7). The mapping used this alternative label and found a perfect match. This is what we mean by **MEA**: a mapping error due to ambiguity.

For an illustration of **MEU**, consider the title *"Recent advancements in airborne radar signal processing : emerging research and opportunities"*

(record-ID 3A1007389885). Table 8 shows the output of System 2.

LLM-suggested term	Mapped GND term	Error Category
Signalverarbeitung [Signal processing]	Signalverarbeitung gnd:4054947-1 [Signal processing]	FP, but relevant
Jamming [Jamming]	Störsender gnd:4273774-6 [jamming transmitter]	TP
Spoofing [Spoofing]	betrügen gnd:4554780-4 [deceive]	MEU
Luftfahrtradar [aviation radar]	Wetterradar gnd:4270420-0 [Weather radar]	MEU

Table 8: System 2 output for the record-ID 3A1007389885 *"Recent advancements in airborne radar signal processing : emerging research and opportunities"* with error category annotation.

Table 9 shows the gold-standard annotations for the same document.

Gold-standard term	GND identifier
Störsender [jamming transmitter]	gnd:4273774-6
AWACS [airborne warning and control system]	gnd:4309079-5
Raum-Zeit-Signalverarbeitung [space-time signal processing]	gnd:4834654-8
Bordradar [on board radar]	gnd:4456131-3
Zielerkennung [target recognition]	gnd:4190792-9
Täuschung (Militär) [deception (military)]	gnd:4184333-2

Table 9: Gold-standard terms for the record-ID 3A1007389885.

We see that the LLM-suggested candidate *Spoofing* is a close match to the concept *Täuschung (Militär)*, and *Luftfahrtradar* is close to both *AWACS* and *Bordradar*. However, as *Spoofing* and *Luftfahrtradar* do not exist directly in the GND, these terms get matched to the wrong entities in the GND. In these cases, the LLM suggested concepts that are unknown in the GND.

Out of the 43 truly erroneous suggestions in our sample, 11 may be classified as **MEU** and 5 may be classified as **MEA**.

Fine-tuning of the ranking stage and stricter filtering based on cosine similarity might help alleviate such errors in the future. Suggestions removed in such a filtering step could also be used as candidates for new subject terms in the GND that need to be added to the vocabulary (as synonyms or new concepts). The problem of resolving ambiguity remains a severe challenge. In a productive setting, analyzing these errors to enhance the vocabulary for automated subject indexing would complement the system for improved performance.

Complementing the analysis of false positives, we can also analyze this sample with a focus on recall and take a closer look at false negatives.

Using the false-negative categories introduced earlier[6], we find the following absolute counts:

Error Category	Absolute Count
FNM	68
FNC	28
False Negatives (total)	96

Table 10: Distribution of false negatives for 176 subject suggestions from System 2 on 50 test-set documents with 162 gold-standard subjects.

If we only count **FNM** as truly missing aspects, we find a micro average recall of

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

C. System 3 Detailed Error Analysis

In a close examination of the 50 test-set documents that were also rated by the experts during qualitative evaluation and their top-5 subjects predicted by System 3 (Annif) (Suominen et al., 2025) (250 predicted subject terms in total), as well as their TIBKAT ground-truth subjects (161 in total), we found 91 true positives and 70 false negatives. In addition to the false-negative categories introduced earlier, we define the following additional category:

FNM5: aspect missed by the top-5 predictions but present in the subsequent top-20 predictions (subcategory of FNM)

Using this classification, we found 61 cases of FNM, of which 25 were FNM5; that is, the subject did not appear among the top-5 predictions, but was present further down the list of the top-20 predictions. We also found 9 cases of FNC. The full data for the error analysis can be found at https://github.com/sciknoworg/tib-sid/tree/main/evaluation/results/system3%20-%20additional%20analysis.

To illustrate these categories with examples, consider the document titled *"Dynamic term structure modeling beyond the paradigm of absolute continuity"* (record-ID 3A168734406X). Table 11 shows the top-5 predictions of System 3 for this document. Table 12 shows the gold-standard annotations for the same document.

This example document illustrates the difficulties of accurately predicting the gold-standard subjects, despite the predictions being rated as very relevant

[6] Opposed to System 1 and System 3, System 2 has a dynamically regulated number of suggestions per record, typically fewer than five. It is therefore irrelevant to analyze whether an aspect was matched later in the ranking, as for the other systems

Prediction	GND identifier	Eval.	Error Cat.
Kreditrisiko [credit risk]	gnd:4114309-7	Y	FP, but relevant
Kreditmarkt [credit market]	gnd:4073788-3	Y	FP, but relevant
Zinsstrukturtheorie [term structure theory]	gnd:4117720-4	Y	FP, but relevant
Modellierung [modelling]	gnd:4170297-9	I	FP, technically correct
Semimartingal [semimartingale]	gnd:4180967-1	Y	TP

Table 11: System 3 top-5 output for the record-ID 3A168734406X *"Dynamic term structure modeling beyond the paradigm of absolute continuity"* with expert evaluations and error category annotation.

Gold-standard subject	GND identifier	Freq.	Error Cat.
Arbitrage [arbitrage]	gnd:4002820-3	0	FNM
Semimartingal [semimartingale]	gnd:4180967-1	4	TP
Ausfallrisiko [default risk]	gnd:4205942-2	4	FNM5
Zinsstruktur [term structure]	gnd:4067855-6	1	FNC
HJM-Modell [HJM model]	gnd:4642940-2	1	FNM5

Table 12: Gold-standard subjects for the record-ID 3A168734406X with train-set frequencies and error category annotation.

by the subject experts. Out of the top-five predictions in Table 11, only *Semimartingal* appears in the gold-standard subjects and thus counts as a true positive. Of the other four predictions, three were considered relevant (Y) and one was considered technically correct but irrelevant (I) by the subject experts.

Looking at the same document from the perspective of the five gold-standard subjects in Table 12, again only one is a true positive while the other four are different kinds of false negatives: *Arbitrage* was not predicted at all by the system, *Ausfallrisiko* and *HJM-Modell* were not in the top-five predictions but appeared within the remaining top-20, and *Zinsstruktur* was not predicted, but the closely related concept *Zinsstrukturtheorie* was among the top-five predictions. All five gold-standard subjects for this document have a low frequency in the training set, ranging from 0 to 4 occurrences. This makes it challenging to predict them using System 3, which mostly relies on models that learn to distinguish each individual subject based on patterns in the training data.

Figure 5 shows a histogram of the subjects binned by training-set frequency and split into true positives and the three subtypes of false negatives (in the diagram, FNM5 has been separated out from the remaining FNM cases). For the low-frequency bins 0, 1, and 4–7, false negatives dominate over true positives. In the higher frequency bins from 8–15 upwards, true positives are more common

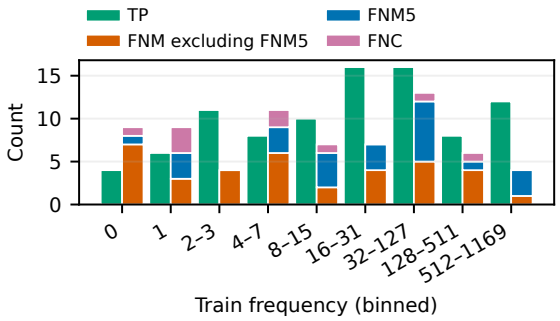


Figure 5: Train-frequency distribution (binned) for the System 3 predictions, split into true positives and three false-negative subtypes.

than false negatives. Bin 2–3 is an outlier in this pattern where true positives dominate despite the low training-set frequency.

Despite the outlier bin, the pattern is clear: System 3 struggles in the prediction of subjects that have a low frequency in the training set, while higher-frequency subjects are more often correctly predicted. This is expected, because out of the three algorithms that form the ensemble, only one (MLLM) is capable of predicting zero-shot subjects, while the other two algorithms (Omikuji and XTransformer) rely on subject-specific training data and therefore cannot predict low-frequency subjects very well, or indeed at all in the zero-shot case. In the future, the system could be improved by including more methods focused on predicting low-frequency terms, for example by matching document text to GND subjects via embeddings as in System 1, or by using off-the-shelf LLMs to suggest possible candidate subjects as in System 2.