Multilinguality in MIND: Advancing Cross-lingual News Recommendation with a Multilingual Dataset

ANDREEA IANA, Data and Web Science Group, University of Mannheim, Mannheim, Germany
GORAN GLAVAŠ, University of Würzburg, Würzburg, Germany
HEIKO PAULHEIM, Data and Web Science Group, University of Mannheim, Mannheim, Germany

Digital news platforms rely on recommendation systems to meet the diverse information needs of readers. However, most research focuses on major, resource-rich languages, overlooking the linguistic diversity of online communities. Moreover, existing work typically assumes *monolingual* news consumption, neglecting polyglot users, and resulting in a lack of multilingual benchmarks for developing recommenders suited to multilingual and low-resource contexts. To address this gap, we introduce xMIND, an *open, multilingual* news recommendation dataset created by machine translating the English MIND dataset into 14 linguistically and geographically diverse languages with varying digital footprints. Using xMIND, we systematically evaluate several content-based neural news recommenders (NNRs) in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer, examining both monolingual and bilingual consumption patterns. In FS-XLT, we compare random and category-based replacement methods for incorporating target-language data during training. Our results show that (i) current NNRs, grounded in multilingual language models, experience significant performance drops in ZS-XLT, and (ii) injecting target-language data in FS-XLT provides limited improvements, especially for bilingual consumption. Notably, randomly injecting target-language news during training leads to greater performance gains compared to category-based replacements. Our in-depth analysis of representation alignment between source and target languages within the language model shows that FS-XLT improves cross-lingual alignment primarily for high-resource languages, while low-resource languages remain weakly aligned with English. These findings highlight the need for broader research efforts in multilingual and cross-lingual news recommendation. We release xMIND at https://github.com/andreeaiana/xMIND.

CCS Concepts: • **Information systems** → **Recommender systems**; **Multilingual and cross-lingual retrieval**; **Test collections**; **Personalization**; • **Applied computing** → *Document management and text processing*.

Additional Key Words and Phrases: Multilingual News Dataset, News Recommendation, Low-resource Languages, Cross-lingual Recommendation, Machine Translation

1 Introduction

The shift from traditional to digital news consumption has transformed news platforms into the primary medium of information for Internet users. This, in turn, propelled personalized news recommendation systems into the main tool used by news websites to address the individual information needs of readers worldwide. With the Internet's global reach, the linguistic diversity of its user base has significantly increased [72, 135]. A substantial portion of these users are polyglots, consuming news in two or more languages. For instance, 22% of Americans speak a language other than English at home[1], whereas 65% of the working-age adults in the European Union are

[1]https://data.census.gov/table/ACSST1Y2022.S1601?q=language

Authors' Contact Information: Andreea Iana, Data and Web Science Group, University of Mannheim, Mannheim, Germany; e-mail: andreea.iana@uni-mannheim.de; Goran Glavaš, University of Würzburg, Würzburg, Germany; e-mail: goran.glavas@uni-wuerzburg.de; Heiko Paulheim, Data and Web Science Group, University of Mannheim, Mannheim, Germany; e-mail: heiko.paulheim@uni-mannheim.de.

ACM 2770-6699/2025/11-ART
https://doi.org/10.1145/3777415

proficient in at least one foreign language.[2] This pattern is especially pronounced among migrant users, who report consuming media both in the host country's language and in the languages of their origin, using them to facilitate integration while maintaining a connection to their home country [1, 9, 23, 80]. Despite this, most online content continues to be dominated by a handful of resource-rich languages, with English alone comprising over half of the world's digital text. Meanwhile, the vast majority of languages worldwide account for less than 0.1% of websites or are missing online representation altogether.[3]

News media serve as a cornerstone of democratic societies by keeping the public informed and providing a platform for sharing and debating ideas and opinions [5, 37]. In this context, recommender systems shape people's worldviews and opinions through the way in which they filter and propagate news [81]. On the one hand, research on personalized neural news recommenders [115] has predominantly centered on enhancing recommendation accuracy and diversity [97, 116, 118, 119], by mitigating technical challenges in news encoding [49, 70, 74, 87, 102, 110–113, 117, 127, 131] and user modeling [3, 45, 66, 88–90, 92, 104]. On the other hand, recent advances in neural machine translation (NMT) [16, 17, 24, 32, 61] and multilingual pre-trained language models (mPLMs) [13, 14, 100, 106, 107, 109, 125] for low(er)-resource languages have started to democratize access to information for underrepresented communities [50].

Nonetheless, two main limitations persist in the existing body of research on news recommendation. First, there is a *scarcity of publicly available, diverse, multilingual news recommendation datasets* that could be leveraged to develop efficient multilingual news recommenders and support effective cross-lingual transfer to resource-lean languages. Although the availability of adequate datasets is paramount for developing high-quality recommenders (e.g., see the Amazon dataset[4] for product recommendation, or MovieLens [35] and Netflix [7] for movie recommendation), the vast majority of news recommendation benchmarks remain monolingual [19, 26, 30, 54, 77, 120]. Furthermore, the few existent multilingual benchmarks mostly feature high-resource languages, and are limited by narrow focus, small size [42], or proprietary constraints [117]. Second, the design of news recommenders for multi- and cross-lingual settings has been left largely unexplored. Traditional news recommenders typically cater to single-language usage, hindering the ability to browse and recommend content across multiple languages simultaneously. As a result, multilingual users often receive recommendations that are less relevant, unbalanced, and lacking in diversity [72]. Overall, these limitations pose significant challenges for online news readers who are multilingual or consume news in resource-lean and/or underrepresented languages. Accounting for such multilingual consumption patterns is crucial for developing news recommender systems that serve minority groups fairly, as overlooking their linguistic practices risks reinforcing digital inequities.

Contributions. We address the above research gaps by introducing xMIND, a new large-scale and publicly available multilingual news dataset for multi- and cross-lingual recommendation. xMIND is constructed by translating the articles of the English-only dataset MIND [120] into a diverse selection of 14 high- and low-resource languages, spanning five geographical macro-areas, and 13 distinct language families, leveraging the NLLB open-source machine translation system [16]. Compared to existing multilingual news recommendation datasets, xMIND is: **(1)** much more *diverse* – we include both resource-rich and resource-poor languages, covering a wide variety of geographical regions, language families, and scripts, with some of the languages being out-of-sample for existing mPLMs (i.e., not present in the mPLM's pre-training); **(2)** *parallel* – the same set of news has been translated into all target languages, enabling the direct comparison of the performance of multilingual news recommenders and cross-lingual transfer approaches across target languages; **(3)** *open source* – we release the dataset in the TSV format and provide scripts for loading and combining the news with the corresponding click logs from the MIND dataset in the NewsRecLib [44] library.

[2]https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180926-1
[3]https://w3techs.com/technologies/overview/content_language
[4]https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews

We use xMIND to benchmark a range of state-of-the-art neural news recommenders, in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer setups, covering two realistic news consumption settings: monolingual and bilingual. Furthermore, we examine two methods for injecting target-language data during model training, and for creating bilingual news consumption patterns: (i) random replacement and (ii) category-based replacement of English news with their target-language translations from xMIND. We show that recommenders trained monolingually on English news suffer significant performance drops when evaluated on the target languages under ZS-XLT conditions, even when paired with a massively multilingual language model. More importantly, we demonstrate that target-language injection during training has a limited effect in mitigating these performance drops. This is particularly the case when adopting a category-based injection strategy in FS-XLT. Additionally, we show that fine-tuning the recommender's underlying language model with low-rank adapters offers a lightweight alternative that drastically reduces the number of trainable parameters while maintaining comparable performance. We further analyze cross-lingual alignment between English and target languages within the news encoder, finding that few-shot target language injection primarily benefits high-resource languages, whereas low-resource languages require substantially more target-language examples to achieve alignment that nonetheless remains weaker with English. These findings reveal the urgent need for developing more accurate and robust cross-lingual news recommendation approaches. Lastly, we assess the quality of the translations in xMIND through a human-based annotation task and comparison against translations obtained with a commercial NMT system.

This manuscript is an extension of the article "MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation" published in the Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval [46], where we introduced the xMIND multilingual dataset. In this work, we broaden the scope in three main directions. First, we examine whether the choice of the base multilingual pre-trained language model used by the benchmarked recommenders imposes any specific constraints for certain languages in xMIND. To that end, we investigate how the performance of these recommenders under ZS-XLT correlates with different language models by experimenting with three mPLMs: XLM-RoBERTa [13], LaBSE [25], and umT5 [12]. As reported in Section 6, although the LaBSE sentence encoder exhibits a smaller drop in ranking performance in ZS-XLT compared to XLM-RoBERTa and umT5, all recommenders experience a significant performance decline on low-resource languages relative to English, irrespective of which mPLM is employed. Second, we study how introducing target-language news in both few-shot training and bilingual evaluation settings impacts the recommenders' performance. Specifically, beyond the random replacement strategy proposed in the original conference paper, we explore a category-based substitution of English news with their target-language translations from xMIND, applied to both few-shot training and in the construction of bilingual consumption patterns. Our findings, detailed in Sections 6.2 and 6.3, show that injecting randomly selected target-language news leads to greater improvements in FS-XLT than the category-based approach, which appears to overfit to specific topics observed during training. Third, we conduct additional analyses of cross-lingual transfer in neural news recommendation. Specifically, we investigate whether fine-tuning the recommender's language model with low-rank adapters improves transfer, and find that this strategy maintains comparable downstream performance while substantially reducing computational requirements (cf. Section 7.1). We then examine the degree of representation alignment between English and the xMIND languages within the news encoder's language model. As discussed in Section 7.2, we observe that the gains of target-language injection are concentrated in high-resource languages, whereas low-resource languages remain weakly aligned with English even after larger amounts of target-language training data are provided. We believe these comprehensive empirical analyses offer valuable new insights and advance research on cross-lingual news recommendation.

2 Related Work

2.1 News Recommendation

Personalized news recommendation aims to alleviate the information overload of online news readers by providing suggestions tailored to their individual preferences [67, 115]. Neural news recommenders (NNRs) have become the driver of personalized news recommendation, replacing systems relying on manual feature engineering [115]. The majority of NNRs commonly consist of a dedicated (i) news encoder, (ii) user encoder, and a (iii) click predictor. The news encoder learns news representations from various input features (e.g., title, topical categories, named entities), either by instantiating convolutional neural networks [102, 110, 112], self-attention networks [88, 113], or graph attention networks [87] with pre-trained word embeddings, or, more recently, by leveraging pre-trained language models [70, 117, 127, 131]. Afterwards, the user encoder aggregates and contextualizes the embeddings of a user's clicked news into a user-level representation by means of sequential [3, 91, 104] or attentive [110, 113] encoders. Lastly, a candidate article's recommendation score is computed by comparing its embedding against the user profile [115]. Although a significant body of work has sought to improve personalization by enhancing the NNRs' core components – news and user modeling – the vast majority of efforts have been nearly exclusively deployed in monolingual settings. More specifically, despite the abundance of polyglot news readers, few works explore the behavior of NNRs in a multi- or cross-lingual scenario. Wu et al. [117] suggested instantiating the news encoder with mPLMs to enable news recommendation in diverse languages. Guo et al. [31] proposed a novel news encoder based on an unsupervised cross-lingual transfer model to address the few-shot recommendation problem between record-rich and unpopular or early-stage recommendation platforms without overlapping users and with news in different languages. However, these works focus exclusively on (i) resource-rich languages and (ii) monolingual news consumption. News recommendation for multilingual news consumers, especially speakers of under-resourced languages, thus remains largely unexplored.

2.2 LLM-based Recommendation

More recently, the emergence of large language models (LLMs) and their remarkable natural language understanding and generation capabilities have significantly reshaped the design of recommender systems. Broadly, we distinguish two main paradigms. The first employs (L)LMs for a discriminative task, where the model functions primarily as a powerful encoder to capture semantic representations of users, items, or contexts [57, 79, 122]. The second leverages LLMs generatively by producing recommendations directly in natural language, generating explanations for recommendations, or reformulating user queries to improve retrieval [20]. Within the domain of news recommendation, in particular, a growing body of research is exploring how LLMs can be employed to enhance personalization, improve interpretability, and address challenges such as cold-start and dynamic user interests. Li et al. [68] evaluate ChatGPT for news recommendation, focusing on personalization, provider fairness, and fake news, while Li et al. [69] study fine-tuning strategies and show that although ChatGPT underperforms popularity-based recommenders in zero-shot settings, fine-tuning improves performance for users with consistent topical histories with the candidates but remains weak for those whose histories are not topically aligned with the candidate articles. Zhang and Wang [131] propose Prompt4NR, framing the click prediction task as a cloze-style mask-prediction task, and finding that combining discrete, continuous, and hybrid prompts performs best. PGNR [71] treats recommendation as a text-to-text language generation task with LLMs, while Yada and Yamana [126] use LLMs to generate category descriptions of news articles to enhance news recommendation. Gao et al. [27] introduce a generative news recommendation paradigm that leverages LLM knowledge to construct theme-level representations of news and users. EmbSum [129] generates user-interest summaries with LLMs to model long histories, whereas SPAR [128] uses LLMs to extract global interests from the user's history for user profile enrichment. ONCE [75] examines both open- and closed-source LLMs to augment training data and improve content representation, and Zhao et al. [132] extensively compare shallow and large language models for news

recommendation. Despite the increasingly multilingual capabilities of LLMs, all existing approaches are evaluated solely on monolingual datasets (e.g., MIND [120]) and overlook multilingual consumption patterns.

2.3 Recommendation Datasets

The advancement of recommender systems heavily depends on the existence and availability of suitable datasets. In the past decade, several public monolingual datasets have been proposed for training and benchmarking news recommenders: Plista [54] (German), Adressa [30] (Norwegian), Globo [19, 26] and its recent improved version NPR [77] (Portuguese), MIND [120] (English), and the most recent EB-NeRD [60] (Danish). Among these, the MIND dataset has become a reference benchmark for the news recommendation community, given the limitations of the earlier datasets, such as a lack of original news texts, metadata information, or limited dataset size [120]. However, these datasets consists only of monolingual news, and therefore, hinder the development of multilingual recommender systems. In earlier work [42], we have aimed to address this problem by proposing NeMig, a multilingual news recommendation dataset in English and German. NeMig contains articles on the topic of refugees and migration collected from German and US media outlets, and rich user data encompassing both click logs and demographic and political information. Besides covering only two major languages, NeMig is small (7K German and 10K English articles) and covers only one specific topic. Wu et al. [117] mention a multilingual news recommendation dataset collected from the MSN News platform to analyze the effectiveness of mPLM-based news encoders in multilingual news recommendations. Their dataset contains user data from seven countries (US, Germany, France, Italy, Japan, Spain, and Korea). Besides all seven included languages being very highly resourced, the dataset is proprietary, i.e., it is not publicly available.

2.4 Cold-start Recommendation

Personalized news recommendation relies on user-item interactions for training, making it vulnerable to the cold-start recommendation problem when data on new items or users is sparse or unavailable [28]. To address this, prior work has explored meta-learning [63, 78, 134], knowledge distillation [41, 105], or contrastive learning [108, 133], to transfer information from warm to cold items, thereby improving predictions on unseen cases. Other approaches generate synthetic interactions for cold-start items from behavioral signals [10, 40, 76]. More recently, LLMs have been employed to tackle cold-start challenges, leveraging their strong zero-shot and few-shot capabilities [38, 95, 103, 123, 130]. Retrieval-augmented generation (RAG) methods have also emerged, enhancing LLM-based recommenders with external knowledge sources [15, 21, 53, 121]. The news recommendation domain has followed similar strategies. For example, Alshehri and Zhang [2] propose a generative adversarial zero-shot learning framework that constructs virtual behavioral representations for cold users or items, whereas Li et al. [65] introduce a meta-learning approach that factorizes user preferences into time-specific and time-shift representations. GZRec [36] synthesizes cold- and warm-start interactions for users and news, and frames cold-start recommendations as zero-shot predictions. BCE4ZSR [94] and ZS-CEBE [93] both adopt a knowledge distillation framework, where a cross-encoder models user-news interactions and a bi-encoder generates user and item embeddings. However, these methods are generally evaluated only on monolingual datasets (e.g., MIND [120], Adressa [30]), and overlook language transfer or language-dependent user preferences. To the best of our knowledge, Guo et al. [31] are the only ones to study cold-start in news recommendation as a few-shot recommendation problem, showing that cross-lingual transfer, namely transferring user-news preferences from a many-shot source to a few-shot target domain, can improve performance. Cross-lingual news recommendation constitutes a specific case of cold-start: (i) new articles may appear in a language unseen by the recommender, and (ii) users histories may span different or multiple languages. In other words, users or items in one language are effectively cold-start cases relative to a knowledge-rich language familiar to the recommender. Yet, the field

still lacks multilingual benchmark datasets to evaluate and advance methods that tackle the cold-start challenges arising from linguistic diversity.

3 Dataset Creation

We create xMIND with two primary considerations in mind: (1) covering languages that are mutually *diverse* linguistically, geographically, and in terms of amount of available text corpora (i.e., high or low resource) and (2) creating a multilingual news dataset that is multi-parallel, i.e., where an article (i.e., or its translation) exists in each covered language. The former allows for a more realistic estimate of global multilingual and cross-lingual performance of news recommendation models [50, 84], whereas the latter enables direct comparability of recommenders' performance across target languages. We thus create xMIND by translating 130,379 unique news articles from the train, development, and test portions of the English MIND dataset [120] (i.e., union of MINDlarge and MINDsmall) into 14 different languages using the NLLB 3.3B open-source NMT model [16]. The MIND news articles consist of a title and an abstract, and are additionally annotated with the topic and Wikipedia-disambiguated named entities extracted from the title and abstract.[5] Note that, while we translate only the title and abstract of each news, these can still be combined with the corresponding linked named entities for usage in knowledge-aware recommendation models [43].

Language Selection. We select target languages for xMIND based on the following criteria: (1) linguistic diversity in terms of typological properties [22, 73], language family, and geographical provenance, (2) script diversity, (3) amount of available language resources, primarily raw corpora (i.e., inclusion of both high- and low-resource languages), and (4) coverage by NLLB [16]. Table 1 lists the selected languages, summarizing the following information, in accordance to the #BenderRule [6]:

- **Code**: The three-letter ISO 693-3 code of the language;
- **Language**: In case of multiple denominations, we use the language name from the World Atlas of Structures (WALS) [22]. We cross-reference the names with two other major linguistic resources, Glottolog [34] and Ethnologue [64];
- **Script**: We provide the English name of the script;
- **Family and Genus**: Language family and genus from WALS [22] and Glottolog [34];
- **Resource Level**: We borrow NLLB's [16] classification of languages into *low-* and *high-resource*;
- **Total Speakers**: We report the total number of speakers of the language, including L1-level (first-language) and L2-level (second-language) speakers, according to Ethnologue.[6]

We follow Ponti et al. [84] and compute three different diversity scores for our language sample: (i) typology index, (ii) family index, and (iii) geographical index. **1)** The *typology index* is based on 103 typological binary features of each language from URIEL [73]: each feature indicates the presence or absence of a particular linguistic property in a language. As per [84], we compute the typology index as the average of entropy scores computed independently for each feature;[7] **2)** The *family index* is the number of distinct language families divided by the sample size; **3)** The *geography index* is the entropy of the distribution of languages in the sample over 6 geographic macro-areas of the world.[8]

Table 2 reports the three metrics for xMIND, as well as for NeMig [42], and the proprietary dataset from [117] (dubbed *Wu et al.*). xMIND offers the most diverse sample in terms of all diversity indices. The sample of languages spans five out of the six macro-areas, and 13 distinct language families covering 14 different genera.

[5]Note that 5.4% of the news in the entire MIND dataset do not contain an abstract.
[6]We use the latest statistics available in January 2024 at https://www.ethnologue.com/.
[7]The entropy of a feature for which all languages in the sample have the same value is 0; the entropy has the maximal value ($\log 2$) if the feature is present for the same number of languages as for which it is absent.
[8]The six macro-areas, as defined by Dryer and Haspelmath [22], are: Africa, Australia, Eurasia, North America, Papunesia, and South America.

Table 1. The 14 languages of xMIND. We display the language *Code* (ISO 693-3), language name, *Script*, *Macro-area*, and language *Family* and *Genus*. *Res.* indicates whether the language is classified as high or low-resource according to [16].

Code	Language	Script	Macro-area	Family	Genus	Total Speakers (M)	Res.
SWH	Swahili	Latin	Africa	Niger-Congo	Bantu	71.6	high
SOM	Somali	Latin	Africa	Afro-Asiatic	Lowland East Cushitic	22.0	low
CMN	Mandarin Chinese	Han	Eurasia	Sino-Tibetan	Sinitic	1,138.2	high
JPN	Japanese	Japanese	Eurasia	Japonic	Japanesic	1,234.5	high
TUR	Turkish	Latin	Eurasia	Altaic	Turkic	90.0	high
TAM	Tamil	Tamil	Eurasia	Dravidian	Dravidian	86.6	low
VIE	Vietnamese	Latin	Eurasia	Austro-Asiatic	Vietic	85.8	high
THA	Thai	Thai	Eurasia	Tai-Kadai	Kam-Tai	60.8	high
RON	Romanian	Latin	Eurasia	Indo-European	Romance	24.5	high
FIN	Finnish	Latin	Eurasia	Uralic	Finnic	5.6	high
KAT	Georgian	Georgian	Eurasia	Kartvelian	Georgian-Zan	3.9	low
HAT	Haitian Creole	Latin	North-America	Indo-European	Creoles and Pidgins	13.0	low
IND	Indonesian	Latin	Papunesia	Austronesian	Malayo-Sumbawan	199.1	high
GRN	Guarani	Latin	South America	Tupian	Maweti-Guarani	(L1 only) 6.7	low

Table 2. Indices of typological, genealogical, and geographical diversity for the language samples of different multilingual news recommendation datasets.

	Range	xMIND	NeMig	Wu et al.
Typology	[0, 1]	0.42	0.05	0.31
Family	[0, 1]	0.93	0.50	0.43
Geography	[0, ln 6]	1.13	0.00	0.00

We excluded languages from Australia, as they (i) have an extremely low number of native speakers (i.e., at most spoken by a few thousand people), and (ii) are not supported by NLLB [16]. Additionally, GRN and HAT (i) are spoken in South and North America, both originating from underrepresented macro-areas, and (ii) have not been seen in pre-training of XLM-RoBERTa [13]. Moreover, xMIND covers five low-resource languages (Somali, Tamil, Georgian, Haitian Creole, and Guarani) and six different scripts: Latin and Georgian are *alphabet* scripts; Japanese and Chinese Han are *logographic* scripts, whereas the Tamil and Thai are written in *Abugida* script type.

NLLB: Hyperparameter Tuning. We tune the hyperparameters of the NLLB translation model [16] using a subset of Global Voices (GV) [99] as validation set.[9] GV constitutes a parallel corpus of news stories in 46 languages collected from the Global Voices website.[10] We construct the validation dataset by selecting the data files for all covered pairs of *English* (ENG) as the source and any of the 14 languages in xMIND as the target: this results in six language pairs, statistics of which are reported in Table 3. We compare four decoding strategies: greedy, multinomial sampling, beam search, and beam search with multinomial sampling. For the beam search decoding strategies, we search for the optimal number of beams in the range $[2, 8]$ and use default values for all other hyperparameters. We evaluate the translation quality using the sacreBLEU score [85]. With the goal of finding the best decoding strategy for a broad range of languages, we compute the macro-average over the six language pairs in our validation dataset. As shown in Table 4, we identify beam search decoding with 4 beams as the best choice: using more beams increases computational cost while bringing negligible sacreBLEU gains.

[9]We use the most recent version available online in October 2023, namely *GlobalVoices v2018q4*. The original data can be accessed at https://opus.nlpl.eu/GlobalVoices/corpus/version/GlobalVoices.
[10]https://globalvoices.org/

Table 3. Statistics of the subset of Global Voices [99] used as validation data for tuning the NLLB [16] hyperparameters. We report the number of sentence pairs and of words (in millions).

Language Pair	Sentence pairs	Words (M)
ENG -> CMN	137,737	2.83
ENG -> SWH	30,338	1.13
ENG -> IND	15,266	0.54
ENG -> JPN	8,595	0.18
ENG -> TUR	7,479	0.24
ENG -> RON	4,265	0.17

Table 4. Hyperparameter optimization results on the subset of Global Voices [99]. We report only the results obtained with the best number of beams. We report macro-average sacreBLEU scores over six language pairs.

Decoding Strategy	# Beams	sacreBLEU
Greedy	1	18.42
Multinomial sampling	1	11.97
Beam Search	4	19.03
Beam Search Multinomial Sampling	4	18.87

Table 5. Number of news in the different splits of xMIND.

Small		Large		
Train	Dev	Train	Dev	Test
51,282	42,416	101,527	72,023	120,959

Final Dataset. The xMIND dataset contains 130,379 unique news in the 14 different languages listed in Table 1. Each article contains a news ID, a translated title, and a translated abstract – if one was provided in the corresponding English article from MIND [120]. Following Wu et al. [120], we split xMIND, for each language, firstly into a small and a large version of the dataset, and secondly, into train, development, and test portions, each corresponding to the original splits of the MIND dataset.[11] Table 5 lists the number of articles in each variant and split of the dataset. We release xMIND publicly, in tab-separated format at https://github.com/andreeaiana/xMIND. xMIND can be combined with additional news and behavioral information provided in MIND [120], using the news IDs. Additionally, to facilitate a seamless integration with existing NNRs, we implement the data loading functionality for xMIND in NewsRecLib [44]. xMIND can be used beyond cross-lingual news recommendation in text retrieval or translation tasks. Therefore, we release our dataset in Parquet format, in both the small and the large version, also on HuggingFace Datasets.[12]

4 Experimental Setup

We systematically benchmark a range of state-of-the-art content-based NNRs in zero-shot (ZS-XLT) and few-shot (FS-XLT) cross-lingual transfer setting. Our experiments encompass two types of news consumption patterns: monolingual and bilingual.

[11] https://msnews.github.io/

[12] The small version of xMIND can be downloaded from https://huggingface.co/datasets/aiana94/xMINDsmall. The large version of xMIND can be downloaded from https://huggingface.co/datasets/aiana94/xMINDlarge.

4.1 Benchmarked Neural News Recommenders

We evaluate several content-based NNRs: (1) *NAML* [110], (2) *LSTUR* [3], (3) *MINS* [104], (4) *CAUM* [90], (5) *TANR* [112], (5) *MINER* [66], and (6) *MANNeR* [47] (only the `CR-Module` responsible for pure content-based recommendation, without any aspect-based personalization or diversification). Additionally, we include (7) NAML_{CAT} as a baseline – a language-agnostic variant of NAML that generates news embeddings solely from randomly initialized category vectors, and derives user representations by attending to the embeddings of clicked news articles. With the exception of MINER and MANNeR, which are designed with a PLM-based news encoder, the remaining NNRs originally contextualize word embeddings using convolutional neural networks (CNNs) [55], additive attention [4], or multi-head self-attention [101] networks. For fair comparison and to enable multilingual recommendations, we follow Wu et al. [117] and replace the original news encoders of these NNRs with an mPLM. NAML, LSTUR, MINS, TANR, and CAUM leverage category information in addition to the news text, whereas CAUM and MANNeR also encode named entities. Models with multiple input features either concatenate (i.e, LSTUR, CAUM) or attend over them (i.e, NAML, MINS, MANNeR) to produce the final news embedding.

The recommenders further vary in their user encoders: NAML and TANR encode user preferences using additive attention; MINS combines multi-head self-attention with a multi-channel GRU-based [11] recurrent network and additive attention; MINER introduces a poly-attention approach based on multiple additive attentions to learn various interest vectors for each user. Additionally, LSTUR and CAUM differentiate between short and long-term user preferences. More specifically, LSTUR encodes the former from the clicked news embeddings with a GRU, and the latter via randomly initialized and fine-tuned embeddings; the final user representation combines the two embeddings.[13] CAUM models long-term dependencies between clicked news with a candidate-aware self-attention network, short-term user interests from adjacent clicks with a candidate-aware CNN, and obtains the final candidate-aware user embedding by attending over the two intermediate representations. In contrast to the other models, MANNeR does not use a parameterized user encoder. Instead, it uses a late fusion approach that involves mean pooling of dot-product scores between each of the candidates and the clicked news.

All models compute recommendation scores as the dot product between the candidate and the clicked news representations. MANNeR is optimized using a supervised contrastive loss [52], whereas the other recommenders are trained by minimizing the standard cross-entropy loss.

4.2 Training and Optimization Details

Data. We combine the xMIND news with the corresponding click logs and additional news annotations (i.e., categories and named entities) from MIND based on the news IDs. We conduct all experiments on the *small variant* of the resulting dataset. Since Wu et al. [120] do not release test labels for MIND, we use the validation portion for testing, and split the respective training set into temporally disjoint training (first four days) and validation (last day) sets.

Training Details. We conduct all the experiments with XLM-RoBERTa Base [13] as the underlying language model in the news encoder of all recommenders. In Section 5, we experiment with two additional multilingual language models, namely LaBSE [25] and umT5 [12]. Appendix A.1 indicates which languages in xMIND are already included in the pre-training corpora of these mPLMs. We fine-tune only the last four layers of all mPLMs.[14] We use 100-dimensional TransE embeddings [8], pre-trained on Wikidata, to initialize the entity encoder in the news encoder of the knowledge-aware models.[15] In line with prior work, we set the maximum history length to 50 and sample four negatives per positive sample during training, as per Wu et al. [114]. We tune the main

[13]Note that in our experiments, we use the *ini* strategy of LSTUR for obtaining the final user embedding, as it outperforms the *con* variant in preliminary evaluations. We refer the reader to [3] for more details.
[14]In the interest of computational efficiency, we keep the bottom eight layers of the Transformer encoders frozen.
[15]The entity vectors are provided as part of the original MIND dataset [120].

hyperparameters of all NNRs. We train all recommenders with a news encoder based on XLM-RoBERTa or LaBSE with mixed precision and a batch size of 8, while those using umT5 are trained with float 32 precision and a batch size of 4. We train all NNRs with a learning rate of 1e-5, for 10 epochs with early stopping and optimizing with the Adam algorithm [56]. We refer the reader to Appendix A.2 for further details concerning the hyperparameter settings. We repeat each experiment three times, and report averages and standard deviations for the standard metrics: AUC, MRR, nDCG@10.[16] To ensure comparability, we train and evaluate all NNRs using the NewsRecLib library [44].[17]

Infrastructure and Compute. We train all models on a cluster with virtual machines, on single NVIDIA A100 40/80 GB GPUs or A40 48 GB GPUs.

4.3 Cross-Lingual Recommendation Scenarios

We benchmark the NNRs in two evaluation setups: (i) **zero-shot (ZS-XLT)** and (ii) **few-shot (FS-XLT)** cross-lingual recommendation. Firstly, through ZS-XLT we aim to investigate the capabilities of NNRs trained monolingually in English (i.e., on the MIND news) to generate recommendations in another language (i.e., in one of the 14 languages in xMIND). Under ZS-XLT, the user history and candidates during training are *monolingual*, in English only. Secondly, with FS-XLT we seek to determine whether target language injection during training benefits the models' performance compared to pure ZS-XLT. In the FS-XLT setting, we explore two approaches for substituting English news with target-language news: *random* and *category-based*. In the *random* variant, we progressively replace between 10% and 90% of the English news (in both user histories and candidate sets) with target-language articles. In the *category-based* variant, we replace the English news in each user's top-k most-clicked categories with their target-language counterparts, again in both user histories and candidate sets. We test $k \in \{1, 2, 3\}$, corresponding to the one, two, or three most-clicked categories per user. For a fair setup (i.e., no knowledge of the test data distributions during training), the distribution of languages in our validation sets mirror the language ratios of the respective training sets [96].

We couple the two training settings (monolingual and bilingual), with two corresponding types of *news consumption patterns* during inference: (i) **monolingual** (denoted MONO) – the user reads news and receives suggestions only in the target language, and (ii) **bilingual** (denoted BILING) – the user consumes news in English and in another language, and recommendations are also provided in the same two languages. To construct the bilingual user history, and candidate set, respectively, we again consider a *random* and a *category-based* approach. Under the random approach, we randomly replace a subset of the English news items with their corresponding xMIND translations in the target language. Following the procedure used in bilingual training, we vary the portion of replaced news from 10% to 90% in increments of 10%. In the category-based replacement setting, we assume (i) users typically prefer consuming their favorite topics in their primary language, and (ii) their interests vary, leading to different top-clicked categories. Hence, we replace the English news associated with each user's top-k most frequently clicked categories with the corresponding target-language translations from xMIND. Similar to our FS-XLT setup, we incrementally replace up to three top-clicked categories per user.

The two training setups, each combined with both consumption patterns, thus result in four types of experiments: **(i) ZS-XLT$_{\text{MONO}}$** – monolingual training (in English) and evaluation on monolingual news consumption in the target language; **(ii) ZS-XLT$_{\text{BILING}}$** – monolingual training (in English) and evaluation on bilingual news consumption in English and the target language; **(iii) FS-XLT$_{\text{MONO}}$** – bilingual training in a mixture of English and target language and evaluation on monolingual news consumption in the target language; **(iv) FS-XLT$_{\text{BILING}}$** –

[16] For brevity, we omit results for nDCG@5, as they exhibit the same patterns as nDCG@10.
[17] https://github.com/andreeaiana/newsreclib

Table 6. ZS-XLT$_{\text{MONO}}$ recommendation performance. For each model, we report the number of model parameters (in millions), and the performance (i) on the English MIND dataset (denoted ENG), (ii) averaged across all 14 target languages in xMIND (denoted AVG), and (iii) the relative percentage difference between average ZS-XLT$_{\text{MONO}}$ and ENG performance (%Δ). We report averages and standard deviations across three runs. The best results per column are highlighted in bold, the second best are underlined.

mPLM	Model	# Parameters (M)		AUC			MRR			nDCG@10		
		Trainable	Total	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
–	NAML$_{\text{CAT}}$	0.387	0.387	55.46$_{\pm 0.18}$		0.0	31.12$_{\pm 0.56}$		0.0	35.81$_{\pm 0.59}$		0.0
XLM-RoBERTa	CAUM	227	284	57.82$_{\pm 3.01}$	55.90$_{\pm 1.75}$	-3.32	32.92$_{\pm 1.68}$	31.38$_{\pm 1.62}$	-4.68	37.49$_{\pm 1.71}$	35.96$_{\pm 1.58}$	-4.08
	LSTUR	313	370	56.80$_{\pm 1.36}$	56.28$_{\pm 1.68}$	-0.92	33.00$_{\pm 0.59}$	31.53$_{\pm 0.85}$	-4.47	37.45$_{\pm 0.54}$	36.03$_{\pm 0.85}$	-3.78
	MANNeR	222	279	50.00$_{\pm 0.00}$	50.00$_{\pm 0.00}$	0.00	35.58$_{\pm 0.31}$	33.03$_{\pm 0.54}$	-7.15	40.17$_{\pm 0.21}$	37.64$_{\pm 0.44}$	-6.28
	MINER	221	278	57.73$_{\pm 7.33}$	55.81$_{\pm 4.33}$	-3.32	31.71$_{\pm 4.95}$	30.20$_{\pm 3.42}$	-4.76	36.45$_{\pm 4.84}$	35.02$_{\pm 3.51}$	-3.90
	MINS	226	283	59.89$_{\pm 0.29}$	56.94$_{\pm 1.40}$	-4.93	34.7$_{5\pm 0.24}$	33.11$_{\pm 0.51}$	-4.70	39.35$_{\pm 0.20}$	37.64$_{\pm 0.50}$	-4.35
	NAML	224	280	52.85$_{\pm 2.27}$	52.49$_{\pm 2.60}$	-0.68	35.98$_{\pm 0.44}$	33.98$_{\pm 0.95}$	-5.56	40.43$_{\pm 0.39}$	38.38$_{\pm 1.02}$	-5.06
	TANR	223	280	54.18$_{\pm 5.91}$	53.27$_{\pm 1.91}$	-1.68	35.47$_{\pm 0.95}$	32.14$_{\pm 0.90}$	-9.40	40.03$_{\pm 0.86}$	36.78$_{\pm 0.88}$	-8.11
LaBSE	CAUM	414	471	64.92$_{\pm 0.83}$	63.52$_{\pm 0.80}$	-2.16	34.36$_{\pm 0.40}$	33.53$_{\pm 0.69}$	-2.69	39.40$_{\pm 0.21}$	38.34$_{\pm 0.54}$	-2.70
	LSTUR	504	560	65.62$_{\pm 0.74}$	63.28$_{\pm 0.88}$	-3.56	33.82$_{\pm 0.67}$	32.73$_{\pm 0.78}$	-3.22	**47.63$_{\pm 0.50}$**	**46.71$_{\pm 0.60}$**	-1.95
	MANNeR	415	472	**68.31$_{\pm 1.26}$**	**66.89$_{\pm 1.21}$**	-2.09	36.52$_{\pm 0.65}$	35.24$_{\pm 0.52}$	-3.50	41.11$_{\pm 0.65}$	39.81$_{\pm 0.48}$	-3.16
	MINER	414	471	60.19$_{\pm 4.61}$	62.16$_{\pm 0.44}$	+3.26	35.82$_{\pm 0.82}$	33.37$_{\pm 0.54}$	-6.84	40.27$_{\pm 0.92}$	37.78$_{\pm 0.44}$	-6.18
	MINS	416	473	63.70$_{\pm 1.02}$	62.31$_{\pm 0.77}$	-2.19	34.31$_{\pm 0.80}$	33.58$_{\pm 0.58}$	-2.11	39.06$_{\pm 0.69}$	38.22$_{\pm 0.46}$	-2.17
	NAML	414	471	51.59$_{\pm 0.71}$	51.93$_{\pm 0.65}$	+0.66	36.63$_{\pm 0.68}$	**35.91$_{\pm 0.51}$**	-1.98	41.25$_{\pm 0.71}$	40.42$_{\pm 0.49}$	-1.99
	TANR	414	471	63.93$_{\pm 0.10}$	62.40$_{\pm 0.26}$	-2.39	**36.80$_{\pm 0.18}$**	34.99$_{\pm 0.38}$	-4.91	41.42$_{\pm 0.18}$	39.56$_{\pm 0.39}$	-4.49
umT5	CAUM	203	288	60.77$_{\pm 0.93}$	59.60$_{\pm 0.83}$	-1.92	33.41$_{\pm 0.61}$	32.18$_{\pm 0.58}$	-3.69	37.86$_{\pm 0.47}$	36.72$_{\pm 0.48}$	-2.99
	LSTUR	289	374	55.10$_{\pm 1.32}$	52.55$_{\pm 1.62}$	-4.63	31.69$_{\pm 1.45}$	30.18$_{\pm 1.80}$	-4.77	45.94$_{\pm 1.06}$	44.71$_{\pm 1.33}$	-2.68
	MANNeR	198	283	66.31$_{\pm 0.49}$	60.59$_{\pm 1.18}$	-8.63	35.13$_{\pm 0.55}$	31.48$_{\pm 0.98}$	-10.39	39.55$_{\pm 0.38}$	35.65$_{\pm 0.88}$	-9.86
	MINER	197	282	53.09$_{\pm 0.71}$	50.73$_{\pm 0.30}$	-4.44	26.85$_{\pm 0.39}$	25.11$_{\pm 0.26}$	-6.49	31.17$_{\pm 0.43}$	29.21$_{\pm 0.22}$	-6.30
	MINS	202	287	54.36$_{\pm 1.82}$	52.91$_{\pm 1.25}$	-2.67	30.10$_{\pm 0.70}$	28.95$_{\pm 0.77}$	-3.82	44.69$_{\pm 0.51}$	43.82$_{\pm 0.57}$	-1.94
	NAML	199	284	54.66$_{\pm 0.53}$	52.68$_{\pm 0.62}$	-3.63	31.04$_{\pm 0.84}$	30.01$_{\pm 1.39}$	-3.32	45.49$_{\pm 0.66}$	44.66$_{\pm 1.06}$	-1.82
	TANR	199	284	53.23$_{\pm 0.48}$	49.71$_{\pm 0.47}$	-6.62	28.33$_{\pm 0.17}$	25.96$_{\pm 0.48}$	-10.39	33.18$_{\pm 0.14}$	30.34$_{\pm 0.45}$	-8.56

bilingual training in a mixture of English and target language and evaluation on bilingual news consumption in English and the target language.[18]

5 Zero-Shot Cross-lingual Transfer in Recommendation

We evaluate the performance of the benchmarked models under ZS-XLT, comparing their results to those obtained on the English data (MIND). We first discuss the performance of the models within the standard monolingual news consumption pattern, followed by an evaluation in the bilingual consumption setting. An ideal NNR should consistently rank relevant (positive) candidates higher than irrelevant (negative) ones, regardless of language. Consequently, our analysis primarily focuses on ranking performance, utilizing the nDCG@10 metric.

Monolingual Consumption Pattern. Following the experiments in [46], in most of our evaluations, we rely on XLM-RoBERTa Base [13] – a multilingual, encoder-only Transformer model – as the main text encoder for all NNRs (see 4.2 for details). We also compare the models' performance when coupled with two additional language models: (i) LaBSE [25], a BERT-based multilingual sentence encoder, and (ii) the encoder of umT5 [12], which follows the encoder-decoder architecture of the mT5 family [125] and is pre-trained on multilingual corpora. Table 6 summarizes the ZS-XLT$_{\text{MONO}}$ recommendation performance for each model, averaged across the 14 languages in xMIND. These results are compared against the models' performance on MIND (i.e., trained and

[18]Note that due to the high computational costs, in the case of FS-XLT$_{\text{BILING}}$ with a random-replacement strategy, we replace the same percentage of English news with articles in the target language both during training and testing (e.g., 10% RON with 90% ENG during training results in the same mixture in testing). We adopt the same setup in the case of the category-based replacement in FS-XLT$_{\text{BILING}}$.

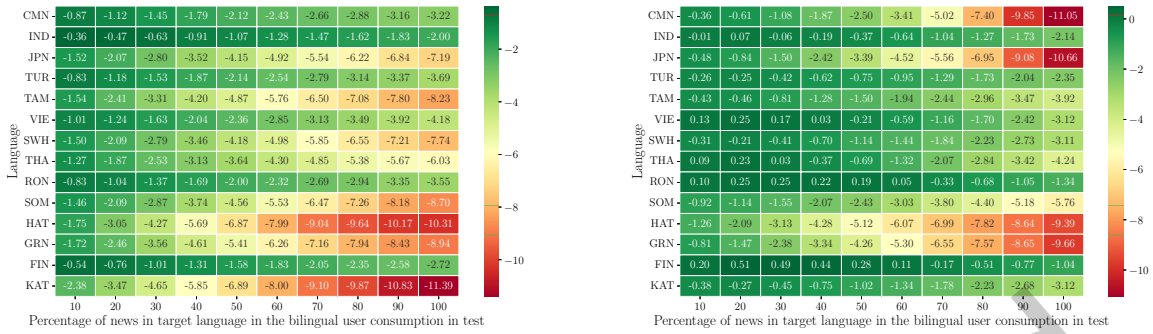
Fig. 1. ZS-XLT$_{\text{MONO}}$ ranking performance (nDCG@10), across the 14 languages in xMIND and English, for three different types of mPLMs used in the recommenders' news encoder.

tested solely in English). For additional context, we include NAML$_{\text{CAT}}$, which generates representations based only on topical categories rather than content. Because it is content- and language-agnostic, NAML$_{\text{CAT}}$ serves as a suitable baseline for both English and the target languages.

In the monolingual English setting on the MIND dataset, all XLM-RoBERTa- and LaBSE-based NNRs outperform the category-based NAML$_{\text{CAT}}$ baseline. Using XLM-RoBERTa, relative improvements in nDCG@10 range from 1.78% (MINER) to 12.91% (NAML), while LaBSE-based recommenders yield gains between 9.08% (MINS) and 33.01% (LSTUR). Generally, NNRs equipped with umT5 perform on par with or slightly better than those using XLM-RoBERTa, except for MINER and TANR, which do not surpass the content-agnostic baseline.

Under ZS-XLT$_{\text{MONO}}$, however, XLM-RoBERTa-based NNRs exhibit weaker performance relative to NAML$_{\text{CAT}}$, achieving at most a 7.19% improvement (NAML) or even a 2.19% drop (MINER). As in the English-only case, recommenders using LaBSE consistently surpass the content-agnostic NAML$_{\text{CAT}}$, with relative gains of up to 12.87% (NAML). Meanwhile, the ranking performance of umT5-based models varies substantially, exhibiting drops of up to 18.43% (MINER) or gains of up to 24.85% (LSTUR) compared to their fully English-trained and evaluated counterparts.

We also observe that XLM-RoBERTa-based NNRs trained in English and evaluated on the target languages in xMIND suffer average performance declines of 3.78% (LSTUR) to 8.11% (TANR) relative to their English results.



(a) Results for MANNeR

(b) Results for NAML

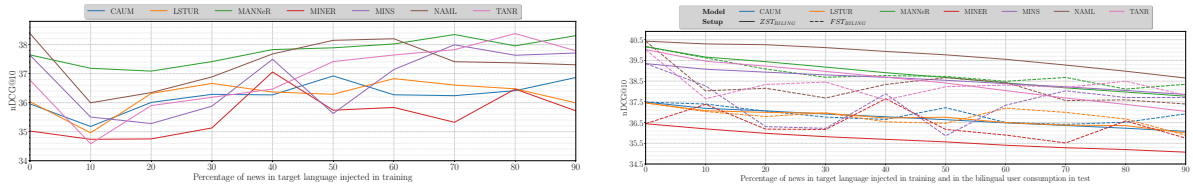
Fig. 2. Relative percentage difference in ranking performance (nDCG@10), under ZS-XLT$_{\text{BILING}}$ compared to full English training and testing.

While top-performing English models (e.g., NAML, MANNeR, TANR) still rank highly under ZS-XLT$_{\text{MONO}}$, they also experience the steepest relative drops. It is important to remember that a random recommender would exhibit the same performance across all languages (i.e., a 0% difference from English). As such, absolute performance in cross-lingual evaluation remains more crucial than the magnitude of these relative drops. Although using a sentence encoder does not entirely eliminate the gap between ZS-XLT and in-language performance, it does narrow it, reducing performance declines to between 1.95% (NAML) and 6.18% (MINER) compared to English.

Figure 1 provides a language-by-language breakdown of ZS-XLT$_{\text{MONO}}$ performance across the 14 target languages in xMIND. XLM-RoBERTa-based NNRs fare best on VIE, IND, RON, and FIN, which are well-represented in mPLM's pre-training corpus [13]. Conversely, performance is poorest across models for KAT, HAT, and GRN, the latter two being completely unseen during pre-training (though XLM-RoBERTa was at least exposed to the Latin script). All LaBSE-based recommenders exhibit notable drops on KAT, with some also performing worse on GRN (i.e., out-of-sample for LaBSE), THA, and CMN. Finally, results for umT5-based NNRs are less uniform: certain models underperform on CMN and SOM, but perform comparatively well on low-resource languages such as GRN and KAT.

Bilingual Consumption Pattern. We next examine how each model's ranking performance changes under bilingual user consumption (ZS-XLT$_{\text{BILING}}$), using XLM-RoBERTa as the base mPLM for all recommenders. In these experiments, we apply a *random replacement* of English news with their translations in the target language, both in user histories and candidate sets. Figure 2 presents results for NAML and MANNeR – representative of the overall trends – relative to their respective English-only baselines, comparing (i) performance across different target languages and (ii) varying proportions of target-language news.[19] Overall, performance steadily declines as the proportion of target-language news increases. This drop is most pronounced for languages unseen in the mPLM's pre-training (i.e., HAT and GRN), as well as for KAT. Interestingly, even though all models share the same mPLM, they are not equally robust to the choice of target language. For instance, NAML's performance on CMN drops by up to 9.85% (with 90% of the user history in CMN), exceeding its losses on HAT and GRN, whereas MANNeR exhibits only a 3.16% decline under the same conditions. However, MANNeR suffers a larger decrease for HAT, a pattern that is more pronounced in TANR and LSTUR. Additionally, for certain languages and models (e.g., LSTUR with RON or FIN), the decline in ranking performance is smaller when either English or the target language dominates the user history.

[19]We focus on NAML and MANNeR for clarity, as their behavior reflects that of other recommenders.



(a) FS-XLT$_{\text{MONO}}$ results

(b) FS-XLT$_{\text{BILING}}$ versus ZS-XLT$_{\text{BILING}}$ results

Fig. 3. FS-XLT ranking performance (nDCG@10) averaged across the 14 languages in xMIND, for various target-language injection ratios during training, and in the user's bilingual news consumption in test. 0% denotes fully monolingual English training (i.e., ZS-XLT). We show the results for the *random replacement* strategy in FS-XLT.

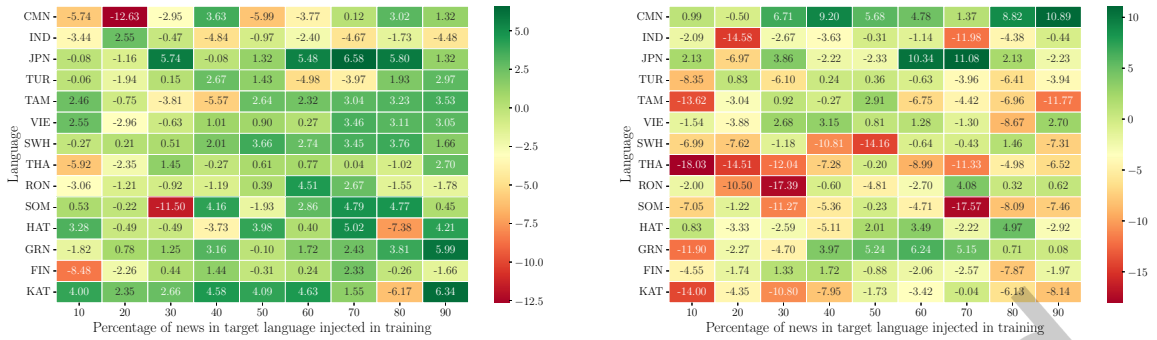
Given the models' diverse architectural designs, this highlights the need to investigate the robustness of the architecture, particularly of the user encoder, under varying news consumption patterns, moving beyond the traditionally assumed monolingual user history. Prior studies [45, 48] have shown that NNRs based on sentence encoders or employing late fusion, rather than a parameterized user encoder, tend to exhibit better robustness. Nonetheless, these findings were derived exclusively from monolingual news consumption scenarios, and their validity may differ substantially in the context of multilingual user histories.

6 Few-Shot Cross-lingual Transfer in Recommendation

Few-shot transfer, which requires the injection of a few target-language instances during model training, has been shown to yield sizable performance gains in NLP tasks [96]. Therefore, it is often leveraged as an effective remedy to the dramatic performance drops suffered by multilingual models in ZS-XLT setups, particularly for resource-lean target languages that are linguistically distant from the source language [62]. Motivated by these findings, we further analyze whether adding target-language data in training also benefits news recommenders. For all experiments in this section, we use only XLM-RoBERTa as the underlying language model in the news encoder of all NNRs. We first examine the effects of few-shot target language injection based on the random replacement strategy. Afterwards, we investigate how category-based replacement in FS-XLT affects the NNR's training. Finally, we compare the two replacement methods against ZS-XLT in a monolingual news consumption setting.

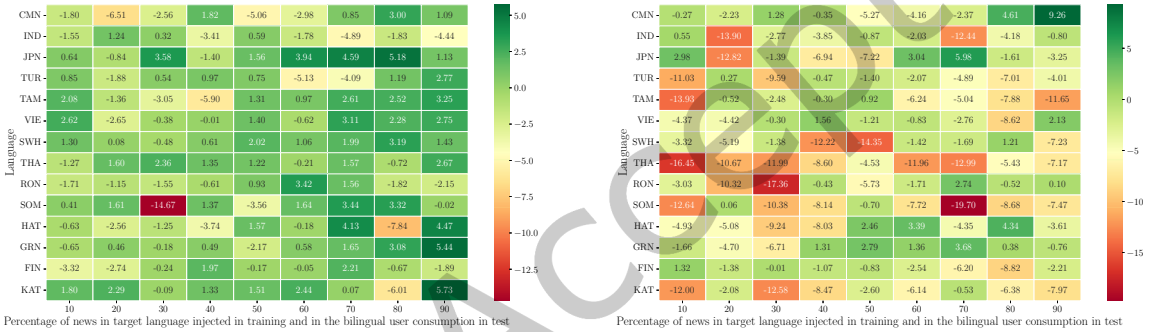
6.1 Random Replacement Replacement of Source-Language News

Monolingual Consumption Pattern. Figure 3a displays the average ranking performance (nDCG@10) of the NNRs across the 14 target languages in xMIND, given increasing proportions of target-language news added to the training data. Although incorporating target-language news does mitigate some of the performance loss seen under ZS-XLT$_{\text{MONO}}$, our results show that when the target language comprises less than 30–40% of the training data, the recommenders perform worse than under ZS-XLT. We believe that this effect stems from the relatively short user histories (averaging 33 articles), where a small fraction of news in a different language can confuse the recommenders' user encoder. Moreover, while higher proportions of target-language data generally produce greater gains over ZS-XLT, the improvements differ by model and the amount of target-language news injected. For example, models such as NAML and CAUM tend to yield less accurate rankings when the share of target-language news is very low or very high, whereas MANNeR – featuring a non-parameterized user encoder – shows more stable performance across varying distributions of source and target-language articles. Examining representative results per language (Figure 4), we find that FS-XLT particularly benefits low-resource



(a) Results for MANNeR (b) Results for NAML

Fig. 4. Relative percentage difference in ranking performance (nDCG@10), under FS-XLT$_{\text{MONO}}$ compared to ZS-XLT$_{\text{MONO}}$, with a *random replacement* of English with target-language news in training.

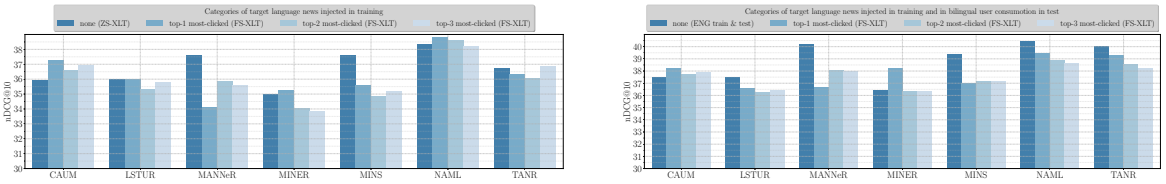


(a) Results for MANNeR (b) Results for NAML

Fig. 5. Relative percentage difference in ranking performance (nDCG@10), under FS-XLT$_{\text{BILING}}$ compared to ZS-XLT$_{\text{BILING}}$, with a *random replacement* of English with target-language news.

languages, and languages not included in XLM-RoBERTa's pre-training. While NAML (Figure 4b) and LSTUR rarely see significant gains from target-language injection, models such as TANR, and particularly MANNeR (Figure 4a), reap substantial benefits from even a modest amount of target-language training data. In the case of MANNeR, this robustness could be attributed to its contrastive loss objective, which naturally pulls positive news (i.e., clicked articles) from the user history closer together in the embedding space, independent of language. For TANR, a plausible explanation is its joint recommendation-topic prediction objective, which places greater emphasis on topic representations that are largely language-agnostic. We note, in particular, that for languages such as KAT or THA, injecting any fraction of target-language news actually reduces ranking performance in NAML compared to the zero-shot setting. One possible explanation is that the additive-attention-based user encoder struggles to capture language-specific patterns from the short user histories.

Bilingual Consumption Pattern. In contrast to the FS-XLT$_{\text{MONO}}$ setting, few-shot target-language injection appears less effective under bilingual news consumption (i.e., FS-XLT$_{\text{BILING}}$). Figure 3b plots the ranking performance (nDCG@10) of various NNRs in FS-XLT$_{\text{BILING}}$ compared to their respective ZS-XLT$_{\text{BILING}}$ baselines, for



(a) FS-XLT$_{\text{MONO}}$ vs ZS-XLT$_{\text{MONO}}$ results

(b) FS-XLT$_{\text{BILING}}$ results

Fig. 6. FS-XLT ranking performance (nDCG@10) averaged across the 14 languages in xMIND, for various target-language injection ratios during training, and in the user's bilingual news consumption in test. We show the results for the *category-based replacement* strategy in FS-XLT.

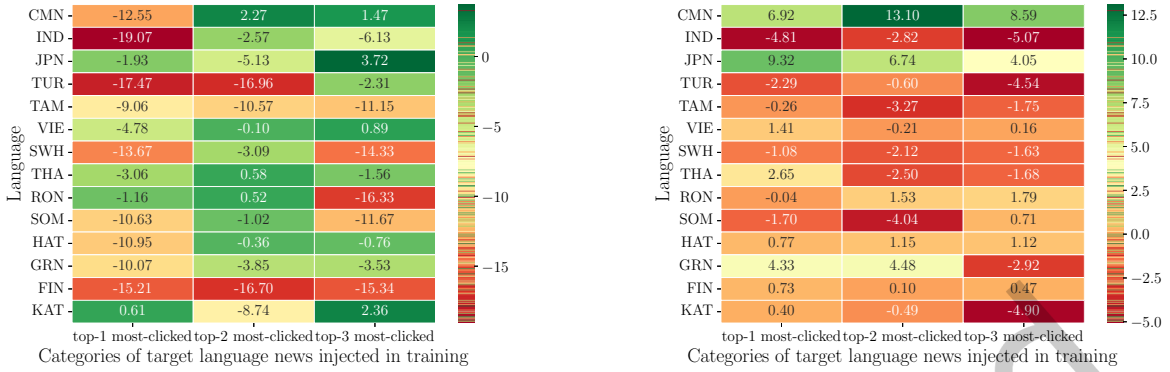
different ratios of target-language news injected during training. Most NNRs match or surpass their ZS-XLT$_{\text{BILING}}$ performance, indicating that limited exposure to target-language samples can be beneficial. However, NAML – generally the top-performing NNR in our evaluation (when using XLM-RoBERTa as the base model) – performs worse than the version trained solely on English. As seen in Figure 5b, NAML only benefits from FS-XLT$_{\text{BILING}}$ for languages where it already suffers the greatest losses under ZS-XLT$_{\text{BILING}}$ (e.g., CMN, JPN, HAT, and GRN, cf. Figure 2b). Similar trends emerge for CAUM and MINER, whereas MANNeR (Figure 5a) and MINS achieve more consistent improvements across all languages.

Overall, these performance variations, and the relatively modest gains from random-based few-shot target-language injection under bilingual consumption, emphasize the need for a deeper investigation into the specific factors influencing multilingual NNR performance. Such insights could guide the design of more robust user encoder architectures for multilingual user histories. Although Iana et al. [48] performed a comprehensive analysis of encoder design in neural news recommenders, their findings focus solely on monolingual settings in the English-only MIND dataset [120], leaving open questions about multilingual settings.

6.2 Category-based Replacement Replacement of Source-Language News

Next, we analyze the models' performance under FS-XLT with a category-based injection of target-language news during training. Specifically, we replace English news with target-language translations from a user's top-k most-clicked categories.

Monolingual Consumption Pattern. Figure 6a illustrates the average ranking performance across all 14 languages in xMIND for the benchmarked recommenders under FS-XLT$_{\text{MONO}}$, varying $k \in \{1, 2, 3\}$, and compares these results to the ZS-XLT$_{\text{MONO}}$ baseline (i.e., no injected target-language news during training). We find that for 5 out of 7 models, category-based injection offers small to moderate gains over FS-XLT$_{\text{MONO}}$. However, MANNeR and MINS experience significant performance drops under category-based injection. For the remaining models, injecting target-language news from only the single most-clicked category generally yields the highest relative improvement over FS-XLT$_{\text{MONO}}$, ranging from 0.31% for LSTUR to 3.64% for CAUM. Injecting more categories (i.e., $k = 2$ or $k = 3$) yields limited additional gains (e.g., CAUM, NAML) or even leads to performance decline (e.g., LSTUR, MINER). TANR is an exception, benefiting only when all top 3 most-clicked categories per user are replaced. Figure 7 compares category-based FS-XLT$_{\text{MONO}}$ to ZS-XLT$_{\text{MONO}}$ for NAML and MANNeR across all 14 languages in xMIND. For NAML (Figure 7b), category-based injection tends to harm performance on higher-resource languages (e.g., IND, TUR, SWH) – regardless of the value of k – aligning with the behavior observed for MANNeR and MINS, and contrasting the ZS-XLT$_{\text{MONO}}$ results of other models. Nevertheless, category-based injection does improve performance for non-Latin scripts (e.g., CMN, JPN) and low-resource languages (e.g., HAT,



(a) Results for MANNeR

(b) Results for NAML

Fig. 7. Relative percentage difference in ranking performance (w.r.t. nDCG@10), under FS-XLT$_{MONO}$ compared to ZS-XLT$_{MONO}$, with a *category-based replacement* of English with target-language news in training.

GRN) that were not seen during the pre-training of the base mPLM. This pattern holds across most models except MANNeR (Figure 7a). However, we find that the optimal k value varies per language for each recommender.

Bilingual Consumption Pattern. Finally, we evaluate the recommenders under FS-XLT$_{BILING}$ with category-based replacement (Figure 6b), averaging results across all languages and comparing them to the fully English (monolingual) training and testing setting (i.e., no target-language injection and monolingual history). Most models under bilingual consumption perform worse in terms of ranking compared to the monolingual English setting, despite target-language injection during training. However, CAUM and MINER show notable benefits from category-based target-language injection even in a bilingual news consumption context. These findings underscore the need for more advanced multilingual news encoders and user modeling approaches tailored to multilingual reading histories.

6.3 Random versus Category-based Replacement of Source-Language News

Lastly, we investigate whether category-based injection of target-language news during training offers any advantages over random-based injection. Figure 8 compares the ranking performance of the NNRs in both ZS-XLT$_{MONO}$ and FS-XLT$_{MONO}$ conditions. For FS-XLT$_{MONO}$, we select each model's best-performing injection approach (i.e., the proportion of randomly replaced news or the number of replaced categories) as determined by the highest nDCG@10. Our findings show that random replacement outperforms category-based replacement for 5 out of 7 models; only CAUM and NAML see better results when the replaced news belongs to a user's most-clicked category.

To better interpret these differences, we examine the proportion of target-language news injected by each strategy. On average, a user's history consists of 33.05 clicked news items across 7.12 unique categories. The top most-clicked category contains an average of 12.79 news items, while the second and third most-clicked categories include 5.83 and 6.85 items per user, respectively.[20] The random replacement strategy's best-performing injection rates range between 40% (for MINER) and 80% (for TANR). In contrast, injecting only the top most-clicked category yields a 38.7% replacement rate – below the lowest best-performing random replacement ratio. As shown in Figure 3 and discussed in Section 6.1, injecting less than 40% of target-language news offers little to

[20]We only report statistics for the training set, as those for validation and test sets are nearly identical.

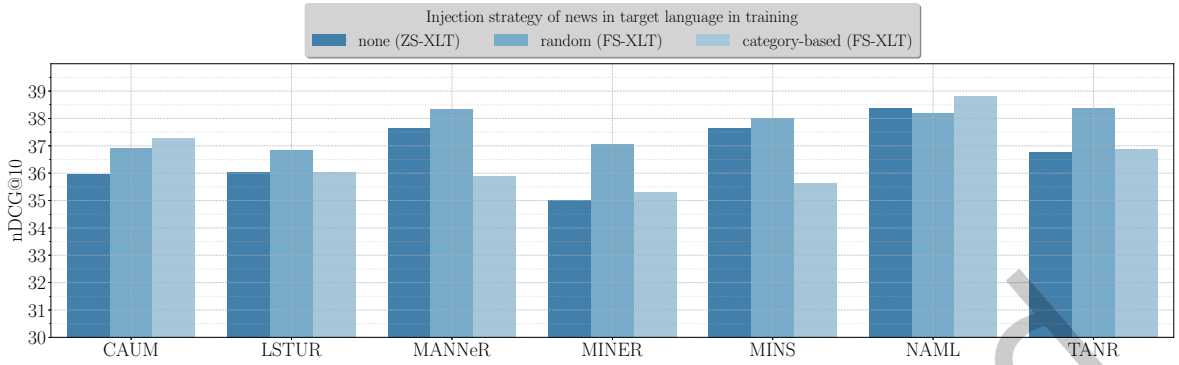


Fig. 8. ZS-XLT$_{\text{MONO}}$ versus FS-XLT$_{\text{MONO}}$ ranking performance (nDCG@10) averaged across the 14 languages in xMIND. For the FS-XLT$_{\text{MONO}}$ setups, we consider both a *random* and a *category-based* replacement strategy. We report the performance with the best-performing ratio (for random replacement) or number of categories (for category-based replacement) of target-language news injected per model.

no improvement over ZS-XLT. Increasing injection of news belonging to the top two (56.33%) or three (77.07%) categories brings the replacement rate closer to that achieved by random replacement (40% - 80% as noted above), but performance remains lower. We hypothesize that inclusion of more diverse categories in the target language constitutes a training signal that better generalizes to unseen topics.

In conclusion, randomly injecting target-language news during training generally delivers better FS-XLT performance, irrespective of the news topic. Nonetheless, further research is needed to develop effective cross-lingual transfer encoders that can narrow the performance gap between source (typically English) and target languages.

7 Further Analysis of Cross-lingual Transfer in News Recommendation

To gain deeper insight into cross-lingual transfer in news recommendation, we conduct two additional analyses. First, we test whether more advanced fine-tuning strategies for the news encoder's underlying mPLM improve ZS-XLT and FS-XLT compared to simply unfreezing the last four layers. Second, we examine the degree of representation alignment between the source language (English) and the target languages within the mPLM during news encoding. Due to the high computational cost, we restrict these experiments to ZS-XLT and FS-XLT settings with *monolingual news consumption patterns* (e.g., user reading histories are exclusively in English or exclusively in the target language). We further narrow the scope to two of the best-performing NNRs identified in our earlier evaluations (cf. Sections 5 and 6): NAML, representative of models trained with point-wise classification objectives, and MANNeR, trained with a supervised contrastive loss.

7.1 Fine-tuning Strategies

In the previous experiments, we updated only the parameters of the last four layers of the news encoder's underlying mPLM on the news recommendation task, a strategy which we denote *Last four*. We now compare this against a parameter-efficient fine-tuning technique. Specifically, we insert low-rank adapters (LoRAs) [39], $\Delta W_{i=1}^{W}$, into the linear layers W of the mPLM. Only the LoRA parameters are updated during training, while all other mPLM parameters remain frozen. The goal is to examine whether retaining most of the pre-trained

Table 7. ZS-XLT$_{\text{MONO}}$ recommendation performance with different fine-tuning strategies for MANNeR and NAML. For each model, we report the number of model parameters (in millions), and the performance (i) on the English MIND dataset (denoted ENG), (ii) averaged across all 14 target languages in xMIND (denoted AVG), and (iii) the relative percentage difference between average ZS-XLT$_{\text{MONO}}$ and ENG performance (%Δ). We report averages and standard deviations across three runs.

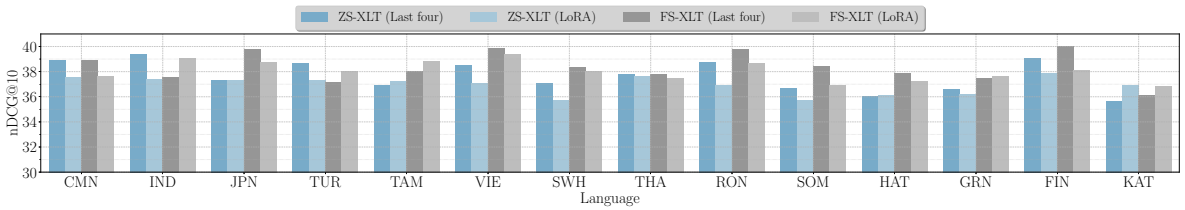
Model	Fine-tuning	# Parameters (M)		AUC			MRR			nDCG@10		
		Trainable	Total	ENG	AVG	%Δ	ENG	AVG	%Δ	ENG	AVG	%Δ
MANNeR	Last four	222	279	$50.00_{\pm 0.00}$	$50.00_{\pm 0.00}$	0.00	$35.58_{\pm 0.31}$	$33.03_{\pm 0.54}$	-7.15	$40.17_{\pm 0.21}$	$37.64_{\pm 0.44}$	-6.28
	LoRA	4.3	282	$62.44_{\pm 3.18}$	$61.27_{\pm 2.31}$	-1.87	$33.11_{\pm 1.55}$	$32.48_{\pm 1.18}$	-1.90	$37.62_{\pm 1.61}$	$36.91_{\pm 1.21}$	-1.88
NAML	Last four	227	280	$52.85_{\pm 2.27}$	$52.49_{\pm 2.60}$	-0.68	$35.98_{\pm 0.44}$	$33.98_{\pm 0.95}$	-5.56	$40.43_{\pm 0.39}$	$38.38_{\pm 1.02}$	-5.06
	LoRA	5.6	283	$58.66_{\pm 0.82}$	$55.16_{\pm 1.83}$	-5.96	$35.77_{\pm 0.77}$	$33.49_{\pm 0.98}$	-6.36	$40.28_{\pm 0.58}$	$38.01_{\pm 0.97}$	-5.63

knowledge while updating only a small number of additional weights can improve cross-lingual performance, while also significantly reducing the number of parameters that need to be fine-tuned.

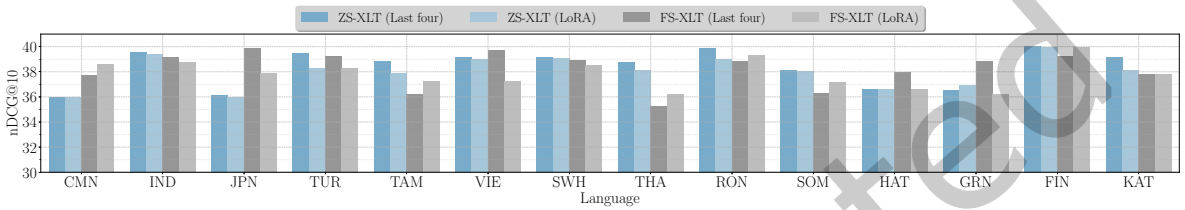
Experimental Setup. We conduct ablation experiments using XLM-RoBERTa [13] as the underlying mPLM in the news encoder of the recommenders. For LoRA, we tune its main hyperparameters, the dimension (rank r) and the scaling factor alpha (α) of the low-rank matrices, and the learning rate, independently for both evaluated NNRs. We perform a grid search with the learning rate in the range $[5e^{-5}, 1e^{-4}, 2e^{-4}, 5e^{-4}]$, r in $[16, 32]$, and α in $[32, 64]$. We train all models with the optimized learning rate of $5e^{-5}$ using LoRA with rank $r = 16$, a scaling factor alpha $\alpha = 32$, and LoRA dropout of 0.05 in all linear layers of the mPLM. We set all other model-specific hyperparameters to the optimal values reported in Section 4.2 and Appendix A.2. In FS-XLT$_{\text{MONO}}$, we adopt the *random replacement* strategy with the best-performing target-language injection ratio per model, namely 60% for NAML, and 70% for MANNeR. We repeat each experiment three times, and report averages and standard deviations for the standard metrics: AUC, MRR, nDCG@10.

Results and Discussion. Table 7 compares the performance of the NNRs in ZS-XLT$_{\text{MONO}}$ under the two fine-tuning strategies. Using LoRAs instead of fine-tuning the last four layers of the mPLM reduces the number of trainable parameters in the NNRs by 98%. More importantly, NAML and MANNeR trained with LoRA achieve considerably better classification performance and comparable ranking performance relative to the *Last four* training variant. For MANNeR in particular, LoRA significantly narrows the ranking performance gap between English and target languages, reducing the relative drop in MRR from 7.15 to 1.90, and in nDCG@10 from 6.28 to 1.88. Although it may appear counterintuitive that LoRA improves classification while *Last four* performs better on ranking, one possible explanation lies in their different adaptation strategies. LoRA freezes the original mPLM weights and learns a low-rank update, which provides stronger regularization and preserves much of the pretrained representation, thereby reducing overfitting and often improving global classification calibration [39]. Ranking metrics, however, depend on fine-grained local ordering of embeddings. Full or partial fine-tuning (i.e., *Last four*) can rotate many singular directions of the weight matrices to optimize relative ordering, whereas low-rank updates produce different structural updates that may be insufficient to induce the high-dimensional shifts needed for top-k ordering [98]. More broadly, dense retrieval and ranking require high capacity and more aggressive feature learning to separate subtle semantic differences [51]. Since LoRA updates fewer effective degrees of freedom, it may underfit pairwise signals while performing well on per-sample label prediction (classification).

Figures 9a and 9b further break down the ZS-XLT$_{\text{MONO}}$ and FS-XLT$_{\text{MONO}}$ ranking performance (nDCG@10) across the 14 target languages in xMIND, for MANNeR, and NAML, respectively. The ZS-XLT$_{\text{MONO}}$ results reinforce the findings from Table 7: ranking performance either slightly deteriorates or remains stable across all languages when reducing the number of trainable parameters via LoRA fine-tuning. In contrast, in FS-XLT$_{\text{MONO}}$, LoRA provides modest gains over *Last four* for certain languages, for example IND, TUR, and KAT in MANNeR, and CMN,



(a) Results for MANNeR



(b) Results for NAML

Fig. 9. ZS-XLT$_{\text{MONO}}$ and FS-XLT$_{\text{MONO}}$ ranking performance (nDCG@10), across the 14 languages in xMIND, for different fine-tuning strategies of the mPLM (XLM-RoBERTa) in the recommenders' news encoder. In FS-XLT$_{\text{MONO}}$, we report the results for the *random replacement* strategy with the best-performing target-language injection ratio per model.

TAM, and SOM in NAML. Although LoRA does not consistently yield large performance improvements, it offers a more parameter-efficient and composable alternative to partial or full fine-tuning. Given the prohibitive cost of fine-tuning large mPLMs, particularly in low-resource settings, LoRA and other lightweight methods such as adapter layers or language specific-adapters [83, 86] warrant further exploration to mitigate performance degradation in cross-lingual transfer in news recommendation.

7.2 Representation Alignment across Languages

We next examine the degree of alignment between source-language (English) and target-language representations of news articles produced by the mPLM before and after fine-tuning.

Experimental Setup. We use XLM-RoBERTa [13] as the underlying mPLM for both recommenders, taking the output representation of the sequence start token [*CLS*] as the news embedding. We consider the mPLM in two states: (i) out-of-the-box, i.e., without any task-specific fine-tuning, and (ii) after fine-tuning it in an FS-XLT setup with varying target-language injection ratios during NNR training. To isolate the effects of the mPLM itself and avoid confounding contributions from other encoder components (e.g., category or entity encoders), we only analyze the fine-tuned mPLM rather than the full news encoder. Concretely, we encode the titles of articles from the MIND test set (English) and xMIND test set (target language), both before and after fine-tuning.

Alignment Evaluation. We assess representation alignment along three dimensions: (i) pairwise alignment, (ii) subspace alignment, and (iii) representation similarity.

Pairwise alignment. We measure pairwise alignment at the article-level using cosine similarity between English and target-language embeddings. Cosine similarity indicates whether representations of the same article in two

languages become more similar after fine-tuning the underlying mPLM. However, it is sensitive to absolute shifts and orientation changes in the embedding space.

Subspace alignment. To assess whether the embedding spaces of the source and target languages span similar directions, we measure subspace alignment using principal angles and Grassmann distance. Formally, given two k-dimensional subspaces $U, V \subset \mathbb{R}^d$, the principle angles $\theta = [\theta_1, \theta_2, \ldots, \theta_k] \in [0, \frac{\pi}{2}]$ are defined recursively as:

$$\begin{aligned} \cos\theta_k &= \max_{\substack{\mathbf{u_k} \in U \\ \mathbf{v_k} \in V}} \mathbf{u_k}^T \mathbf{v_k}, \text{ subject to} \\ \mathbf{u_k}^T \mathbf{u_k} = 1, &\mathbf{v_k}^T \mathbf{v_k} = 1, \mathbf{u_k}^T \mathbf{u_i} = 0, \mathbf{v_k}^T \mathbf{v_i} = 0, (i = 1, ..., k-1) \end{aligned} \tag{1}$$

The first principal angle θ_1 is the smallest angle between any pair of unit vectors from U and V (i.e., the most aligned directions). Subsequent angles measure alignment between the next closest directions, under orthogonality constraints [33]. If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ denote orthonormal bases of U and V, then the cosines of the principal angles are given by the singular values of $\mathbf{Y}_1^T \mathbf{Y}_2$. Small angles indicate strong overlap between subspaces, while large angles reflect near-orthogonality. The Grassmann (geodesic) distance provides a scalar summary [33]: $d_{Grass}(U, V) = \left(\sum_{i=1}^{k} \theta_i^2 \right)^{\frac{1}{2}}$. Intuitively, this distance measures the global misalignment between two subspaces. In our experiments, we compare the top-k ($k = 50$) PCA subspaces of the source- and target-language embeddings, reporting both the distribution of principal angles and the Grassmann distance before and after fine-tuning.

Representation similarity. Finally, we evaluate structural similarity between source- and target-language embeddings using the Central Kernel Alignment (CKA) metric [58]. Let $X, Y \in \mathbb{R}$ denote two sets of embeddings of the same n instances, with linear kernels (e.g, similarity matrices) $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{L} = \mathbf{Y}\mathbf{Y}^T$. CKA is defined as the normalized Hilbert-Schmidt Independence Criterion (HSIC) [29]:

$$CKA(\mathbf{K}, \mathbf{L}) = \frac{HSIC(\mathbf{K}, \mathbf{L})}{\sqrt{HSIC(\mathbf{K}\mathbf{K}) HSIC(\mathbf{L}\mathbf{L})}} \tag{2}$$

where $HSIC(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2} tr(\mathbf{K}\mathbf{H}\mathbf{L}\mathbf{H})$, and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. CKA values lie in $[0, 1]$, with 1 indicating identical relational structure, and 0 denoting unrelated representations. In our context, higher CKA scores between source- and target-language embeddings after fine-tuning suggest that the relational geometry across languages becomes more similar.

Results and Discussion. Figure 10 shows heatmaps of cosine similarities between English and selected target languages from xMIND, before and after fine-tuning, for both MANNeR (left-hand side) and NAML (right-hand side).[21] The diagonal of each heatmap represents the similarity of the same article across English and the target language. In the pre-fine-tuning heatmaps, the diagonal is fuzzy with pairwise matches not being clearly distinguishable from off-diagonal entries (i.e., non-matches). This suggests that the embedding spaces are poorly aligned, as confirmed by many strong off-diagonal similarities. After fine-tuning, the diagonals generally become sharper, with lower off-diagonal cosine similarities (lighter colors). This indicates improved cross-lingual alignment, with English and the target-language counterparts of the same article moving closer in the embedding space. Still, consistent with Section 6, the benefits of different target-language injection ratios vary by language. Higher-resource languages such as CMN, FIN, JPN, SWH, or VIE achieve good alignment with only 30% to 50% target-language injection during fine-tuning. In contrast, languages such as HAT, KAT, or GRN require higher proportions of target-language training examples to improve cross-lingual alignment. These findings partially

[21] For brevity, we illustrate only representative examples of target languages from xMIND in this section. However, our analysis in the discussion of results covers all 14 target languages.

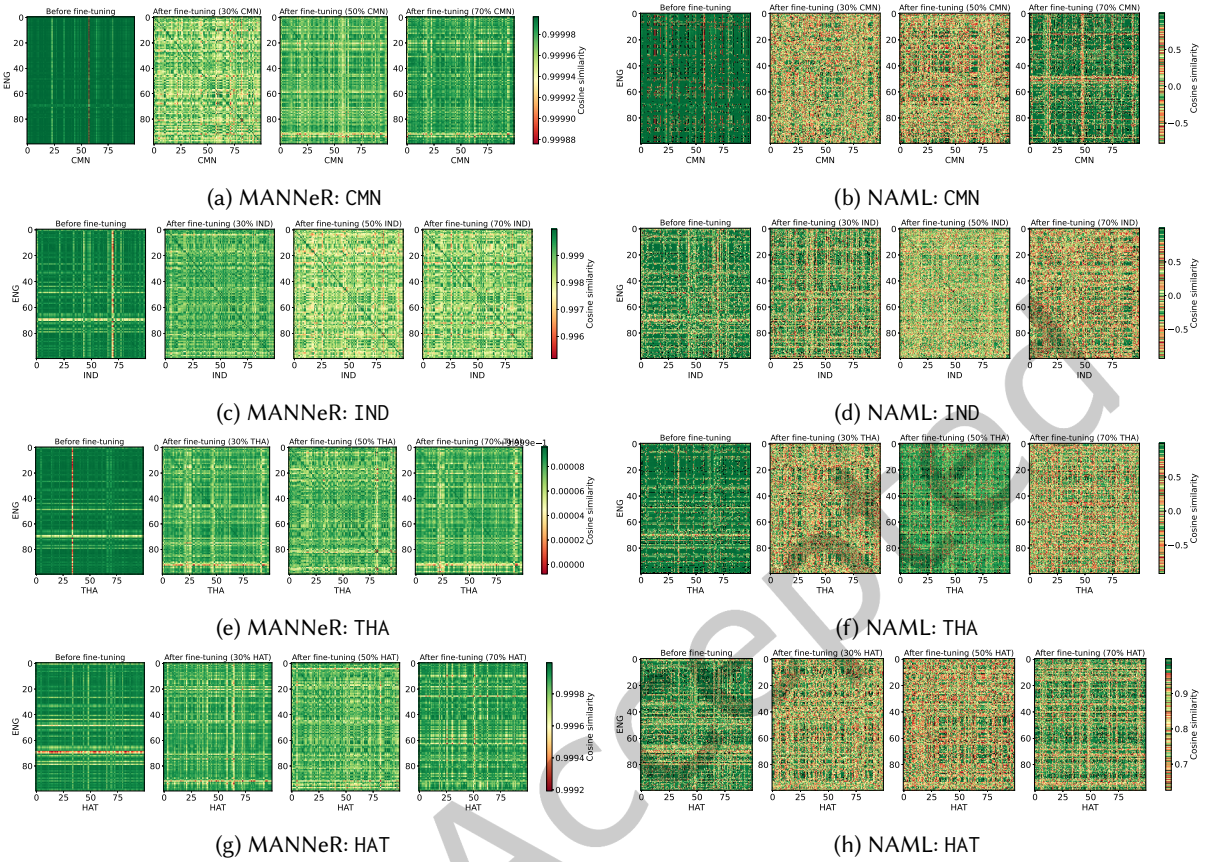
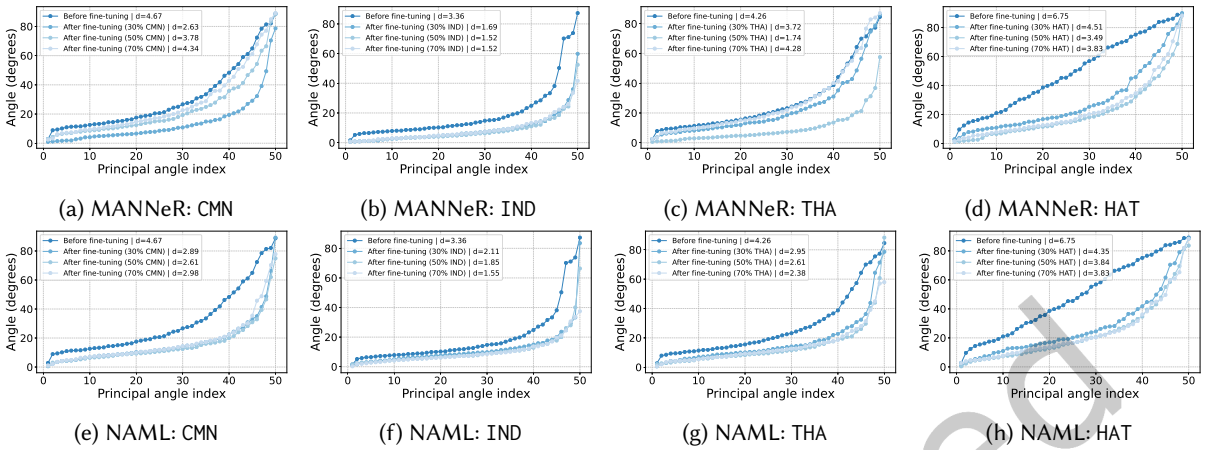


Fig. 10. Cosine similarity between the embeddings of the English and the target-language news articles in the test set, before and after fine-tuning the mPLM (XLM-RoBERTa) in the recommenders' news encoder, across selected languages in xMIND. We report results for FS-XLT$_{\text{MONO}}$, using the *random replacement* strategy and three target-language injection ratios during training (e.g., 30%, 50%, and 70%).

mirror the downstream raking performance trends discussed in 6.1 (see Figure 4): greater alignment seems to correspond to better downstream performance, though the optimal target-language injection ratios may vary.

We next turn to global alignment between the English and target-language embedding spaces. Figure 11 shows the distribution of principal angles and corresponding Grassmann distances before and after fine-tuning. These metrics confirm our earlier observations: source- and target-language embedding spaces become more aligned after fine-tuning, as indicated by smaller angles and lower Grassmann distances. Before fine-tuning, larger principal angles indicate that source and target languages occupy different directions in the embedding space. After fine-tuning, these angles shrink, showing that the spaces move closer together. We observe the largest drop for IND, where the maximum principal angle decreases by over 40° with a 70% target-language injection ratio for both models. However, even after fine-tuning, low-resource languages remain less aligned with English than higher-resource languages, as reflected in their relatively high Grassmann distances. We also observe that different languages benefit from different target-language injection ratios. Notably, for MANNeR, certain ratios fail to improve, or even hurt, alignment: 70% for FIN, 50% for JPN, and 70% for THA. Interestingly, this contrasts



(a) MANNeR: CMN (b) MANNeR: IND (c) MANNeR: THA (d) MANNeR: HAT

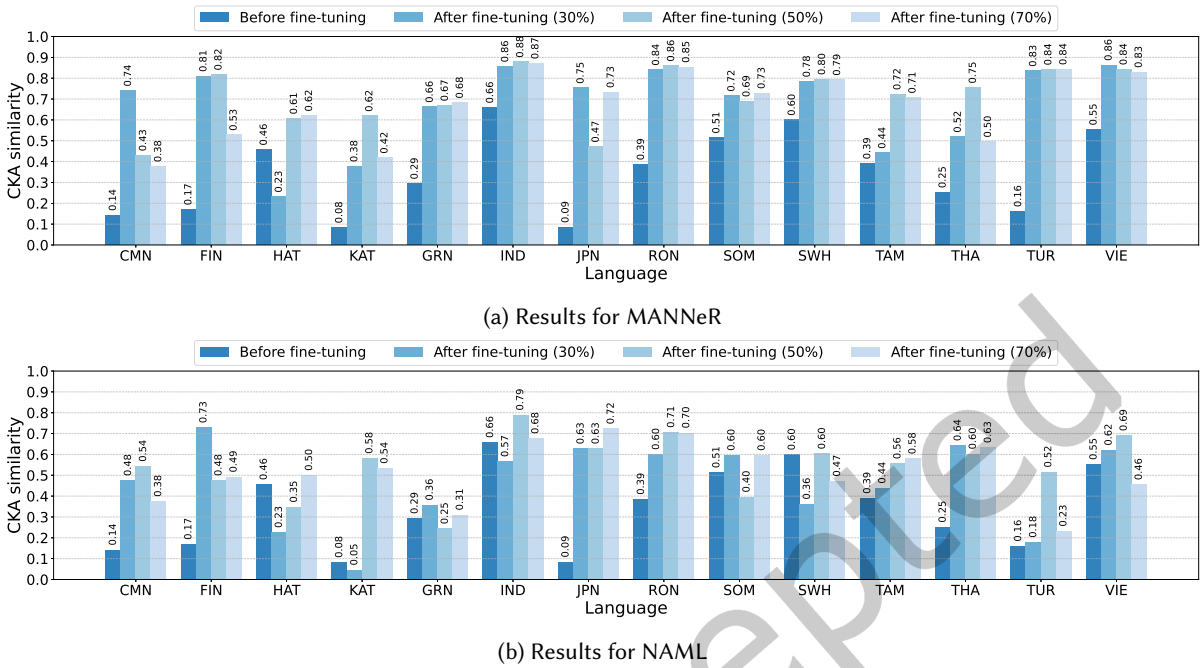
(e) NAML: CMN (f) NAML: IND (g) NAML: THA (h) NAML: HAT

Fig. 11. Distribution of principal angles for the top $k = 50$ PCA subspaces, and Grassmann distance, for the embeddings of the English and the target-language news articles in the test set, before and after fine-tuning the mPLM (XLM-RoBERTa) in the recommenders' news encoder, across selected languages in xMIND. We report results for $\text{FS–XLT}_{\text{MONO}}$, using the *random replacement* strategy and three target-language injection ratios during training (e.g., 30%, 50%, and 70%).

with the downstream performance results in Section 6.1 (see Figure 4), where performance was less affected. This discrepancy may arise from differences in the training regimes of the two models, though further research is needed to better understand cross-lingual alignment in NNRs.

Finally, we examine representation similarity between source and target languages. Figure 12 reports CKA scores for MANNeR and NAML across the 14 xMIND languages, before and after fine-tuning. Results again corroborate previous findings: fine-tuning generally increases CKA, reflecting more similar relational structures between source and target-language embeddings. We highlight a few noteworthy observations. For some languages, specific injection ratios decrease similarity: e.g., 30% for HAT (both models), 30% for KAT, IND, or SWH (NAML), and 50% for GRN (NAML). Additionally, for MANNeR, CKA scores for some languages remain similar across injection ratios (e.g., 30%, 50%, 70%), implying that fewer target-language examples may suffice. This is particularly encouraging for resource-poor languages such as HAT, GRN, or SOM, where training data is scarce. These results thus show that comparable representation similarity can be achieved with fewer samples.

Overall, our findings suggest that task-specific fine-tuning with even a limited number of target-language instances improves both local and global cross-lingual alignment. Nevertheless, the embedding spaces of low-resource languages remain only weakly aligned with English, even after few-shot target-language fine-tuning. This highlights the need for further research into the factors driving cross-lingual NNR performance, and into strategies for improving representation alignment to better serve speakers of underrepresented languages. A promising direction is to leverage multilingual LLMs to reduce the semantic gap across languages. As discussed in Section 2.2, LLMs can help mitigate sparsity by generating metadata (e.g., summaries, categories) in a pivot language from item descriptions in another language, or serve as a bridge between languages by rephrasing, translating, or contextualizing user histories across languages. Beyond improving semantic alignment, future work should also explore how LLMs can capture language-dependent preferences that extend beyond literal overlaps - for instance, recognizing that a user reading "fotbal" articles in Romanian might also be interested in "soccer" content in English. Since cross-lingual transfer effectively constitutes a cold-start scenario for the target language, future work could adapt methods from cold-start recommendation, such as exploiting side



(a) Results for MANNeR

(b) Results for NAML

Fig. 12. CKA similarity between the embeddings of the English and the target-language news articles in the test set, before and after fine-tuning the mPLM (XLM-RoBERTa) in the recommenders' news encoder, across the 14 languages in xMIND. We report results for FS-XLT$_{\text{MONO}}$, using the *random replacement* strategy and three target-language injection ratios during training (e.g., 30%, 50%, and 70%).

information or identifying transferable behavior patterns, within LLM-driven frameworks to build more accurate and equitable cross-lingual recommenders.

8 Translation Quality

We finally investigate the quality of the translations in xMIND. Concretely, we (i) estimate the translation quality and (ii) investigate the robustness of the NNRs to different translations of the source news from MIND [120]. To this end, we use MINDsmall and additionally translate its training and development portions using Google (Neural Machine) Translation (GNMT) [124], a commercial MT system that supports all xMIND languages. GNMT has been shown to outperform NLLB on translation from English to various low-resource languages [16]. [22] [23]

8.1 Manual Quality Estimation of Translation

Given the size of the xMINDsmall dataset and our (limited) annotation budget, it was infeasible to manually post-edit the translations of the news in the test portion of xMIND. We therefore resorted to conducting an annotation task to estimate the quality of the translations. To this end, we sample 50 news from the development portion of the MIND dataset (i.e., the portion used as test set in all our experiments), according to the (i) the

[22] GNMT is a proprietary system, hindering fair comparisons to open-source models due to the lack of transparency regarding its model architecture and training procedures.

[23] We translated the text with the Cloud Translation - Advanced (v3) API: https://cloud.google.com/translate/docs/overview, using Google Cloud research credits worth approximately $5,000.

	Language	CMN	IND	JPN	TUR	TAM	VIE	SWH	THA	RON	SOM	HAT	GRN	FIN	KAT
Intelligibility	NLLB	0.36	0.37	0.39	0.33	0.02	0.35	0.45	0.46	0.51	0.19	0.20	0.10	0.21	0.92
	GNMT	0.30	-0.01	0.09	0.12	0.34	0.13	0.33	-0.11	0.26	-0.08	-0.06	-0.05	-0.21	0.00
Fidelity	NLLB	0.55	0.42	0.36	0.48	0.51	0.25	0.20	0.40	0.55	0.34	0.17	0.30	0.34	0.61
	GNMT	0.28	0.21	0.20	0.39	0.10	0.14	0.19	0.02	0.37	-0.17	0.01	0.03	0.09	0.12
Pairwise Comparison		0.01	0.22	0.19	0.33	0.26	0.16	0.22	0.01	0.30	0.17	0.09	0.16	0.28	0.65

Fig. 13. Annotator agreement (Krippendorf's alpha), per language, for three types of translation aspects in the annotation task.

distribution of categories in the dataset and (ii) the distribution of the total length of the news (i.e., composed of title and abstract). This way, we ensure that the sampled instances are representative of the full dataset.

We carry out the annotations using the Potato annotation tool [82]. Two annotators judged the quality of the NLLB and GNMT translations for each language.[24] The task comprised of a total of five questions, targeting three aspects of the translations: *intelligibility*, *fidelity*, and *pairwise comparison* between the NLLB and GNMT translations. The first two question types were repeatedly asked independently for the NLLB and GNMT translations. The annotators answered the following questions: *(1-2) Is the translation acceptable?* – binary answer; *(3-4) To which extent is the information from the original text accurately retained in the translation?* – 5-point Likert-scale answers, ranging from "Not at all" to "Completely"; *(5) Which translation is better?* – categorical answer with three options, namely "Translation A", "Translation B", or "They are comparably good". In order to remove any position bias in all questions, we randomized the source of translations A and B shown to the annotator, such that 50% of the time translation A stemmed from NLLB, and the remaining 50% from GNMT. Overall, across most of the target languages, we observed higher annotators agreement (Krippendorf's alpha [59]) for the NLLB translations for the first two question types, as shown in Figure 13, than for GNMT, where we observe little to no agreement between the annotators.

For over half of the target languages in xMIND, the annotators deemed the NLLB translations to be intelligible in at least 60% of the cases (see Figure 14a). Similarly, we find that our translations retain the information of the original texts, at least partially, in the majority of cases, as illustrated in Figure 14b. Notably, the NLLB-sourced translation are deemed more faithful to the original news than the GNMT-based ones particularly for low-resource languages such as TAM and GRN. This finding is corroborated by the results from the pairwise comparison, shown in Figure 14c, which shows that NLLB translations are judged to be overall better than their GNMT counterparts for these two languages. Nonetheless, across all languages and aspects of evaluation, translations obtained with the commercial GNMT are deemed generally of higher quality than those generated with the open-source NLLB.

The annotators' feedback revealed several challenges that contributed to the generally low scores assigned to both kinds of translations. Firstly, we remark that often one part of the news (e.g., title) was perfectly translated, whereas the other portion (e.g., abstract) was not accurately depicted in the target language. Secondly, in few cases, the phrasing of the news title was hard to comprehend even in English, impeding translation. Lastly, we note that given the US origins of the news articles from the MIND dataset [120], many of the topics discussed in the news are not usually encountered in some of the target languages or they pertain solely to the US (e.g., national baseball news). In such cases, we observed that the MT systems performed particularly poorly. A closer look at the translation errors reveals that, in terms of intelligibility, both NLLB and GNMT generate worse

[24]All annotators were native speakers of the target language and fluent in English. Most of the annotators were certified interpreters/translators of the target language.

translations for news in categories such as entertainment (e.g., for languages such as VIE, TUR, JPN), movies, music, or television (e.g., particularly for lower-resource languages TAM and KAT). Such errors can be explained by the fact that these categories of news tend to contain terms that exhibit more idiomaticity (e.g., especially in movie titles), which is well-documented source of trouble for MT [18]. Similar patterns emerge when analyzing the news categories on which the fidelity of the translations is lower. However, our results indicate that there is not a particular category on which one of the MT systems is better than the other (according to our annotators). Lastly, we observe that translations of shorter texts, obtained with both NLLB and GNMT, are of higher quality than those of longer texts. This can be explained by the fact that, at least NLLB, has been trained on pairs of shorter documents [16].

8.2 Robustness of NNRs to Translation Quality

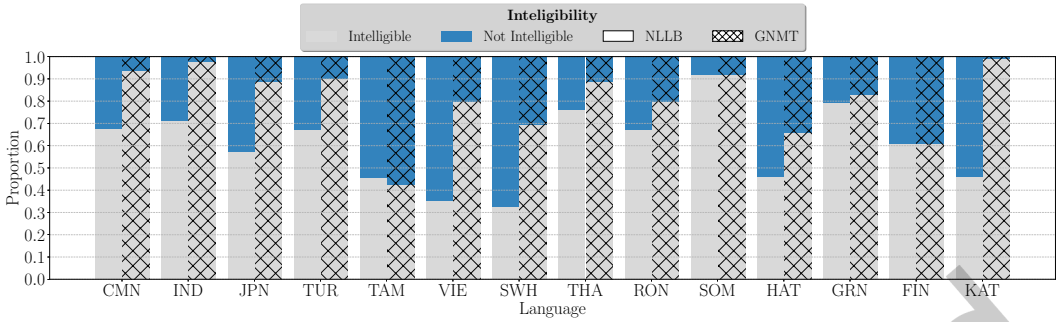
We next investigate the robustness of the best-performing NNRs from our previous experiments, namely NAML and MANNeR, to translations obtained with different MT systems. To this end, we re-run the previous experiments for the version of the dataset translated with GNMT and compare the results against the ones obtained using the NLLB-translated xMINDsmall.

Figure 15 shows the ranking performance (nDCG@10) for both models and MT systems under $\texttt{ZS-XLT}_{\texttt{MONO}}$ (i.e., training on the English MINDsmall and evaluation on the target language). The performance of both NNRs seems largely unaffected by the provenance of the translations. The differences are insignificant according to an independent samples T-test (for a p-value of 0.05), with the exception of MANNeR's performance on IND and KAT, where GNMT test translations lead to better ranked recommendations. This is important as it indicates that, although the GNMT translations were judged to be better on average than the ones produced by NLLB, the NNRs appear to be robust to differences in translation quality.

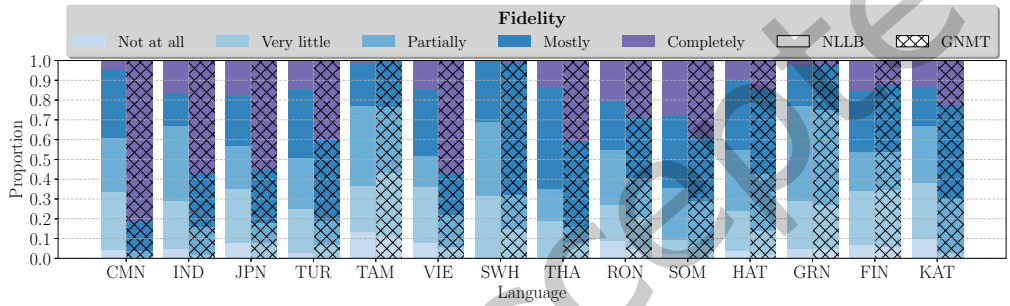
We further check this hypothesis in the $\texttt{ZS-XLT}_{\texttt{BILING}}$ setting. The corresponding results from Figure 16a demonstrate small differences in performance depending on the MT system used. However, the differences are again not statistically significant according to the same independent T-test. We observe similar patterns and no statistical significance in FS-XLT settings (e.g., Figure 16b illustrates the ranking performance for the $\texttt{FS-XLT}_{\texttt{MONO}}$ case). Overall, these findings indicate that, despite the infeasibility of manual post-editing of the test set translations in xMIND, the quality of the translations obtained with the open-source NLLB (i) is on par with those generated by a state-of-the-art commercial MT system, and (ii) has no significant effect on the NNR's recommendation performance.

9 Conclusion

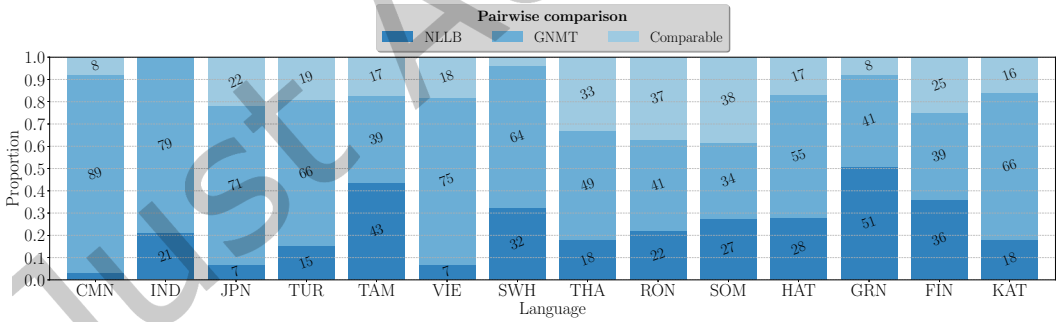
The ever-growing linguistic diversity of online news readers has yet to be fully addressed in news recommendation research, which remains predominantly focused on resource-rich languages – particularly English – and on monolingual consumption. In this work, we introduce xMIND, an open-access, multilingual news recommendation dataset derived from the English-only MIND dataset via machine translation. xMIND encompasses 14 linguistically and geographically diverse languages. We use xMIND to benchmark several state-of-the-art content-based NNRs under zero-shot and few-shot cross-lingual transfer settings, exploring both monolingual and bilingual consumption scenarios. We consider two strategies for injecting target-language data during training, as well as constructing bilingual news consumption patterns: randomly replacing a portion of English news with their translations from xMIND, and replacing only English news corresponding to a user's top-clicked categories. Our findings reveal that existing NNRs experience substantial performance losses under ZS-XLT. Moreover, introducing target-language data during FS-XLT yields only modest gains, especially in bilingual news consumption contexts. Interestingly, random replacement generally outperforms category-based replacement, producing superior ranking performance in most NNRs. Additionally, our analysis of representation alignment between English



(a) Intelligibility of translations



(b) Fidelity of translations



(c) Pairwise comparison of translations

Fig. 14. Annotation task results for each type of translation aspect.

and target languages within the news encoder's language model show that few-shot target-language injection during training primarily benefits high-resource languages. In contrast, low-resource languages remain weakly aligned with English and require substantially more target-language examples to achieve meaningful cross-lingual alignment. We believe xMIND serves as a valuable resource for the news recommendation community, and hope

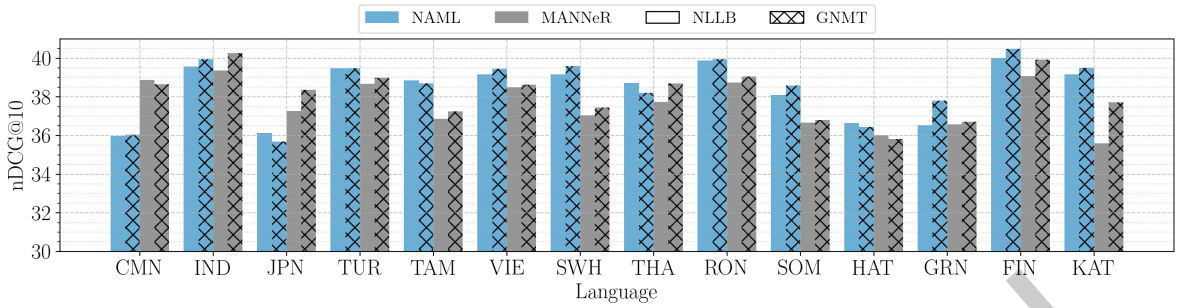
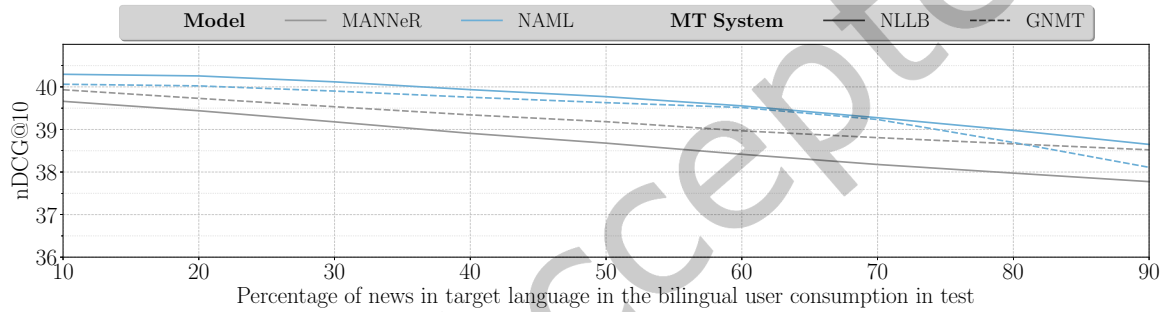
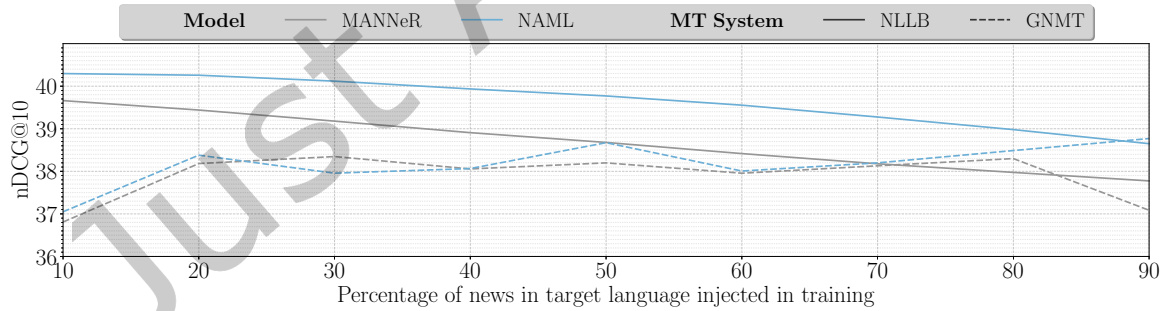


Fig. 15. ZS-XLT$_{\text{MONO}}$ ranking performance (nDCG@10) in terms of MT system for NAML and MANNeR.



(a) ZS-XLT$_{\text{BILING}}$ results



(b) FS-XLT$_{\text{MONO}}$ results

Fig. 16. Ranking performance (nDCG@10) in terms of MT system for NAML and MANNeR.

that it will foster more research on multilingual and cross-lingual news recommendation, for speakers of both high- and low-resource languages.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. This material is based upon work supported by the Google Cloud Research Credits program with the award EDU275608761. We used AI assistance (ChatGPT 5) to polish the writing, as well as to refine the code for our visualizations.

References

[1] Amanda Alencar and Mark Deuze. 2017. News for Assimilation or Integration? Examining the Functions of News in Shaping Acculturation Experiences of Immigrants in the Netherlands and Spain. *European Journal of Communication* 32, 2 (2017), 151–166.

[2] Manal A Alshehri and Xiangliang Zhang. 2022. Generative Adversarial Zero-Shot Learning for Cold-Start News Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 26–36.

[3] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 336–345.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations* (2014).

[5] Jack M Balkin. 2017. Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *UCDL rev.* 51 (2017), 1149.

[6] Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient* 14 (2019).

[7] James Bennett, Stan Lanning, et al. 2007. The Netflix Prize. In *Proceedings of KDD Cup and Workshop*, Vol. 2007. New York, 35.

[8] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 2787–2795. https://dl.acm.org/doi/abs/10.5555/2999792.2999923

[9] Çiğdem Bozdağ and Yasemin Karakasoğlu. 2024. Multilingual Media Repertoires of Young People in the Migration Society: A Plea for a Language and Culture-aware Approach to Media Education. *Global Studies of Childhood* 14, 4 (2024), 448–461.

[10] Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. 2022. Generative Adversarial Framework for Cold-Start Item Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2565–2571.

[11] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734. doi:10.3115/v1/D14-1179

[12] Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining. In *The Eleventh International Conference on Learning Representations*.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[14] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 7059–7069.

[15] Emile Contal and Garrin McGoldrick. 2024. RAGSys: Item-Cold-Start Recommender as RAG System. In *Proceedings of the Workshop Information Retrieval's Role in RAG Systems (IR-RAG 2024)co-located with the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*, Vol. 3784.

[16] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672* (2022).

[17] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[18] Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3608–3626.

[19] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News Session-Based Recommendations using Deep Neural Networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*. 15–23.

[20] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, Rene Vidal, Maheswaran Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, et al. 2024. Recommendation with Generative Models. *arXiv preprint arXiv:2409.15173* (2024).

[21] Dario Di Palma. 2023. Retrieval-augmented Recommender System: Enhancing Recommender Systems with Large Language Models. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1369–1373.
[22] Matthew S. Dryer and Martin Haspelmath (Eds.). 2013. *WALS Online (v2020.3)*. Zenodo. doi:10.5281/zenodo.7385533
[23] Jonas Elis. 2025. How Do Immigrant-Origin and Native Voters Consume Political News Media During a National Election Campaign? *Politische Vierteljahresschrift* (2025), 1–26.
[24] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research* 22, 107 (2021), 1–48.
[25] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 878–891.
[26] P Moreira Gabriel De Souza, Dietmar Jannach, and Adilson Marques Da Cunha. 2019. Contextual Hybrid Session-Based News Recommendation With Recurrent Neural Networks. *IEEE Access* 7 (2019), 169185–169203.
[27] Shen Gao, Jiabao Fang, Quan Tu, Zhitao Yao, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2024. Generative News Recommendation. In *Proceedings of the ACM Web Conference 2024*. 3444–3453.
[28] Jyotirmoy Gope and Sanjay Kumar Jain. 2017. A Survey on Solving Cold Start Problem in Recommender Systems. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 133–138.
[29] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *International Conference on Algorithmic Learning Theory*. Springer, 63–77.
[30] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The Adressa Dataset for News Recommendation. In *Proceedings of the International Conference on Web Intelligence*. 1042–1048.
[31] Taicheng Guo, Lu Yu, Basem Shihada, and Xiangliang Zhang. 2023. Few-shot News Recommendation via Cross-lingual Transfer. In *Proceedings of the ACM Web Conference 2023*. 1130–1140.
[32] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-resource Machine Translation. *Computational Linguistics* 48, 3 (2022), 673–732.
[33] Jihun Hamm and Daniel D Lee. 2008. Grassmann Discriminant Analysis: A Unifying View on Subspace-based Learning. In *Proceedings of the 25th iInternational Conference on Machine Learning*. 376–383.
[34] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. glottolog/glottolog: Glottolog database 4.4.
[35] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2015), 1–19.
[36] Syed Zain Ul Hassan, Muhammad Rafi, and Jaroslav Frnda. 2024. GCZRec: Generative Collaborative Zero-shot Framework for Cold Start News Recommendation. *IEEE Access* 12 (2024), 16610–16620.
[37] Natali Helberger. 2021. On the Democratic Role of News Recommenders. In *Algorithms, Automation, and News*. Routledge, 14–33.
[38] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *European Conference on Information Retrieval*. Springer, 364–381.
[39] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. [n. d.]. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
[40] Feiran Huang, Yuanchen Bei, Zhenghang Yang, Junyi Jiang, Hao Chen, Qijie Shen, Senzhang Wang, Fakhri Karray, and Philip S Yu. 2025. Large Language Model Simulator for Cold-Start Recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 261–270.
[41] Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. 2023. Aligning Distillation For Cold-start Item Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1147–1157.
[42] Andreea Iana, Mehwish Alam, Alexander Grote, Katharina Luwig, Philipp Müller, Christof Weinhardt, and Heiko Paulheim. 2023. NeMig-A Bilingual News Collection and Knowledge Graph about Migration. In *Proceedings of the Workshop on News Recommendation and Analytics co-located with RecSys 2023*.
[43] Andreea Iana, Mehwish Alam, and Heiko Paulheim. 2024. A Survey on Knowledge-aware News Recommender Systems. *Semantic Web* 15, 1 (2024), 21–82.
[44] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. NewsRecLib: A PyTorch-Lightning Library for Neural News Recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 296–310.
[45] Andreea Iana, Goran Glavas, and Heiko Paulheim. 2023. Simplifying Content-Based Neural News Recommendation: On User Modeling and Training Objectives. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2384–2388.
[46] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 553–563.

[47] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9555–9571. doi:10.18653/v1/2024.findings-emnlp.558

[48] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2025. Peeling Back the Layers: An In-Depth Evaluation of Encoder Architectures in Neural News Recommenders. In *CEUR Workshop Proceedings*, Vol. 3929. RWTH Aachen, 1–15.

[49] Junxiang Jiang. 2023. TADI: Topic-aware Attention and Powerful Dual-encoder Interaction for Recall in News Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 15647–15658.

[50] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6282–6293.

[51] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[52] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[53] Hai-Dang Kieu, Minh-Duc Nguyen, Thanh-Son Nguyen, and Dung D Le. 2025. Keyword-driven Retrieval-Augmented Large Language Models for Cold-start User Recommendations. In *Companion Proceedings of the ACM on Web Conference 2025*. 2717–2721.

[54] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The Plista Dataset. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*. 16–23.

[55] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. doi:10.3115/v1/D14-1181

[56] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (2014).

[57] Anton Korikov, Scott Sanner, Yashar Deldjoo, Zhankui He, Julian McAuley, Arnau Ramisa, Rene Vidal, Mahesh Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, et al. 2024. Large Language Model Driven Recommendation. *arXiv preprint arXiv:2408.10946* (2024).

[58] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3519–3529.

[59] Klaus Krippendorff. 2013. *Content Analysis: An Introduction to its Methodology*. Sage Publications.

[60] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. EB-NeRD A Large-scale Dataset for News Recommendation. In *Proceedings of the Recommender Systems Challenge 2024*. 1–11.

[61] Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[62] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4483–4499.

[63] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1073–1082.

[64] M Paul Lewis, Gary F Simons, and Charles D Fennig. 2009. Ethnologue: Languages of the World, Dallas, Texas: SIL International. *Online version: http://www. ethnologue. com* 12, 12 (2009), 2010.

[65] Jingyuan Li, Yue Zhang, Xuan Lin, Xinxing Yang, Ge Zhou, Longfei Li, Hong Chen, and Jun Zhou. 2023. TAML: Time-Aware Meta Learning for Cold-Start Problem in News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2415–2419.

[66] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-Interest Matching Network for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 343–352.

[67] Miaomiao Li and Licheng Wang. 2019. A Survey on Personalized News Recommendation Technology. *IEEE Access* 7 (2019), 145861–145879.

[68] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. A Preliminary Study of ChatGPT on News Recommendation: Personalization, Provider Fairness, and Fake News. In *CEUR Workshop Proceedings*, Vol. 3561. CEUR-WS.

[69] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. Exploring Fine-tuning ChatGPT for News Recommendation. *arXiv preprint arXiv:2311.05850* (2023).

[70] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2023. PBNR: Prompt-based News Recommender System. *arXiv preprint arXiv:2304.07862* (2023).

[71] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Prompt-Based Generative News Recommendation (PGNR): Accuracy and Controllability. In *European Conference on Information Retrieval*. Springer, 66–79.

[72] Chenjun Ling, Ben Steichen, and Silvia Figueira. 2020. Multilingual News–An Investigation of Consumption, Querying, and Search Result Selection Behaviors. *International Journal of Human–Computer Interaction* 36, 6 (2020), 516–535.

[73] Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 8–14.

[74] Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020. KRED: Knowledge-aware Document Representation for News Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 200–209. doi:10.1145/3383313.3412237

[75] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 452–461.

[76] Ruochen Liu, Hao Chen, Yuanchen Bei, Qijie Shen, Fangwei Zhong, Senzhang Wang, and Jianxin Wang. 2024. Fine Tuning Out-of-Vocabulary Item Recommendation with User Sequence Imagination. *Advances in Neural Information Processing Systems* 37 (2024), 8930–8955.

[77] Joel Pinho Lucas, João Felipe Guedes da Silva, and Letícia Freire Figueiredo. 2023. NPR: A News Portal Recommendations Dataset. In *Proceedings of the The First Workshop on the Normative Design and Evaluation of Recommender Systems (NORMalize 2023), co-located with the ACM Conference on Recommender Systems 2023 (RecSys 2023)*.

[78] Yunze Luo, Yuezihan Jiang, Yinjie Jiang, Gaode Chen, Jingchi Wang, Kaigui Bian, Peiyi Li, and Qi Zhang. 2025. Online Item Cold-Start Recommendation with Popularity-Aware Meta-Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 927–937.

[79] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 583–612.

[80] Sarah Naseer, Christopher St. Aubin, and Michael Lipka. 2024. *English- and Spanish-Language News Consumption Among Hispanics*. Pew Research Center. https://www.pewresearch.org/race-and-ethnicity/2024/03/19/english-and-spanish-language-news-consumption-among-hispanics/ Accessed: 2025-09-15.

[81] Eli Pariser. 2011. *The Filter Bubble: What the Internet is Hiding from You*. Penguin UK.

[82] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 327–337.

[83] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 46–54. doi:10.18653/v1/2020.emnlp-demos.7

[84] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2362–2376.

[85] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (Eds.). Association for Computational Linguistics, Brussels, Belgium, 186–191. doi:10.18653/v1/W18-6319

[86] Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Yansong Feng and Els Lefever (Eds.). Association for Computational Linguistics, Singapore, 149–160. doi:10.18653/v1/2023.emnlp-demo.13

[87] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Personalized News Recommendation with Knowledge-aware Interactive Matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 61–70. doi:10.1145/3404835.3462861

[88] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. PP-Rec: News Recommendation with Personalized User Interest and Time-aware News Popularity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5457–5467. doi:10.18653/v1/2021.acl-long.424

[89] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. FUM: Fine-grained and Fast User Modeling for News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1974–1978.

[90] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News Recommendation with Candidate-aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1917–1921.

[91] Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-Preserving News Recommendation Model Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1423–1432. doi:10.18653/v1/2020.findings-emnlp.128

[92] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5446–5456. doi:10.18653/v1/2021.acl-long.423

[93] Muhammad Arslan Rauf, Mian Muhammad Yasir Khalil, Muhammad Ahmad Nawaz Ul Ghani, Weidong Wang, Qingxian Wang, and Junaid Hassan. 2024. ZS-CEBE: Leveraging Zero-shot Cross and Bi-encoder Architecture for Cold-start News Recommendation. *Signal, Image and Video Processing* 18, 8 (2024), 6455–6467.

[94] Muhammad Arslan Rauf, Mian Muhammad Yasir Khalil, Weidong Wang, Qingxian Wang, Muhammad Ahmad Nawaz Ul Ghani, and Junaid Hassan. 2024. BCE4ZSR: Bi-encoder Empowered by Teacher Cross-encoder for Zero-shot Cold-start News Recommendation. *Information Processing & Management* 61, 3 (2024), 103686.

[95] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 890–896.

[96] Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't Stop Fine-Tuning: On Training Regimes for Few-Shot Cross-Lingual Transfer with Multilingual Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 10725–10742.

[97] Hao Shi, Zi-Jiao Wang, and Lan-Ru Zhai. 2022. DCAN: Diversified News Recommendation with Coverage-attentive Networks. *arXiv preprint arXiv:2206.02627* (2022). doi:10.48550/arXiv.2206.02627

[98] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. 2024. LoRA vs Full Fine-tuning: An Illusion of Equivalence. *arXiv preprint arXiv:2410.21228* (2024).

[99] Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

[100] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010. https://dl.acm.org/doi/abs/10.5555/3295222.3295349

[102] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*. 1835–1844. doi:10.1145/3178876.3186175

[103] Lei Wang and Ee-Peng Lim. 2023. Zero-Shot Next-Item Recommendation using Large Pretrained Language Models. *arXiv preprint arXiv:2304.03153* (2023).

[104] Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. 2022. News Recommendation Via Multi-Interest News Sequence Modelling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7942–7946.

[105] Shuai Wang, Kun Zhang, Le Wu, Haiping Ma, Richang Hong, and Meng Wang. 2021. Privileged Graph Distillation for Cold Start Recommendation. In *Proceedings of the 44th International ACM SIGIR cCnference on Research and Development in Information Retrieval*. 1187–1196.

[106] Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. PolyLM: An Open Source Polyglot Large Language Model. *arXiv preprint arXiv:2307.06018* (2023).

[107] Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On Learning Universal Representations Across Languages. In *International Conference on Learning Representations*.

[108] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-Start Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[109] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100* (2022).

[110] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-View Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3863–3869. doi:10.24963/ijcai.2019/536

[111] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2576–2584. doi:10.1145/3292500.3330665

[112] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Topic-Aware News Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1154–1159. doi:10.18653/v1/P19-1110

[113] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6389–6394. doi:10.18653/v1/D19-1671

[114] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2022. Rethinking InfoNCE: How Many Negative Samples Do You Need?. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2509–2515. doi:10.24963/ijcai.2022/348

[115] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems* 41, 1 (2023), 1–50.

[116] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. SentiRec: Sentiment Diversity-aware Neural News Recommendation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 44–53. https://aclanthology.org/2020.aacl-main.6

[117] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering News Recommendation with Pre-trained Language Models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[118] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. End-to-end Learnable Diversity-aware News Recommendation. *arXiv preprint arXiv:2204.00539* (2022). doi:10.48550/arXiv.2204.00539

[119] Chuhan Wu, Fangzhao Wu, Tao Qi, Wei-Qiang Zhang, Xing Xie, and Yongfeng Huang. 2022. Removing AI's Sentiment Manipulation of Personalized News Delivery. *Humanities and Social Sciences Communications* 9, 1 (2022), 1–9.

[120] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.

[121] Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. 2024. CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3391–3401.

[122] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A Survey on Large Language Models for Recommendation. *World Wide Web* 27, 5 (2024), 60.

[123] Xuansheng Wu, Huachi Zhou, Yucheng Shi, Wenlin Yao, Xiao Huang, and Ninghao Liu. 2024. Could Small Language Models Serve as Recommenders? Towards Data-centric Cold-start Recommendation. In *Proceedings of the ACM Web Conference 2024*. 3566–3575.

[124] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144* (2016).

[125] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 483–498.

[126] Yuki Yada and Hayato Yamana. 2024. News Recommendation with Category Description by a Large Language Model. *arXiv preprint arXiv:2405.13007* (2024).

[127] Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, and Qi Liu. 2022. Tiny-NewsRec: Effective and Efficient PLM-based News Recommendation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 5478–5489. https://aclanthology.org/2022.emnlp-main.368

[128] Chiyu Zhang, Yifei Sun, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Sinong Wang, Rong Jin, Sem Park, Ning Yao, and Bo Long. 2024. SPAR: Personalized Content-Based Recommendation via Long Engagement Attention. *arXiv preprint arXiv:2402.10555* (2024).

[129] Chiyu Zhang, Yifei Sun, Minghao Wu, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Rong Jin, Angli Liu, Ji Zhu, Sem Park, et al. 2024. EmbSum: Leveraging the Summarization Capabilities of Large Language Models for Content-Based Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1010–1015.

[130] Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, et al. 2025. Cold-Start Recommendation towards the Era of Large Language Models (LLMs): A Comprehensive Survey and Roadmap. *arXiv preprint arXiv:2501.01945* (2025).

[131] Zizhuo Zhang and Bang Wang. 2023. Prompt Learning for News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 227–237.

[132] Yuyue Zhao, Jin Huang, David Vos, and Maarten de Rijke. 2025. Revisiting Language Models in Neural News Recommender Systems. In *European Conference on Information Retrieval*. Springer, 161–176.
[133] Zhihui Zhou, Lilin Zhang, and Ning Yang. 2023. Contrastive Collaborative Filtering for Cold-Start Item Recommendation. In *Proceedings of the ACM Web Conference 2023*. 928–937.
[134] Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to Warm Up Cold Item Embeddings for Cold-start Recommendation with Meta Scaling and Shifting Networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1167–1176.
[135] Ethan Zuckerman. 2008. The Polyglot Internet. (2008). https://ethanzuckerman.com/the-polyglot-internet/

A Experimental Settings Details

A.1 Language Model Support

Table 8 indicates whether the languages from xMIND are included in the pre-training corpora of the mPLMs used in our experiments (cf. Section 4).

Table 8. Support of the 14 languages of xMIND by different pre-trained multilingual language models.

Language Code	SWH	SOM	CMN	JPN	TUR	TAM	VIE	THA	RON	FIN	KAT	HAT	IND	GRN
XLM-RoBERTa [13]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
LaBSE [25]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
umT5 [12]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗

A.2 Hyperparameter Optimization

We perform hyperparameter optimization for the most important hyperparameters of each NNR using grid search and the English news as training and validation sets (i.e., hyperparameter tuning on the MIND dataset). Concretely, we search for the optimal learning rate in the range $[1e^{-3}, 1e^{-4}, 1e^{-5}]$ for all models, and find $1e^{-5}$ to be the most suitable value for all NNRs. Table 9 lists the search spaces for the other tuned hyperparameters, as well as the best values for each model. We set all remaining model-specific hyperparameters to the optimal values reported in the respective papers. We repeat each experiment three times with different random seeds (i.e., {42, 43, 44}) set with PyTorch Lightning's `set_everything`.

Table 9. Search spaces used for hyperparameter optimization and best values found for all models. We report the optimal values depending on the underlying language model, in the format $value_{XLM\text{-}RoBERTa}$ / $value_{LaBSE}$ / $value_{umT5}$. We use the following abbreviations: $\texttt{num}_{\texttt{heads}}$ = number of attention heads, $\texttt{query}_{\texttt{dim}}$ = dimensionality of the query vector in additive attention, K = number of context codes in MINER [66], score agg = aggregation function for the final user click score calculation in MINER [66], τ = temperature parameter in supervised contrastive loss in MANNeR [47].

	$\texttt{num}_{\texttt{heads}}$	$\texttt{query}_{\texttt{dim}}$	K	score agg	τ
Search Space	{8, 12, 16, 24, 32}	[50, 200]	{8, 16, 32, 48}	{mean, max, weighted}	[0.1, 0.5]
Step	–	–	50	–	0.02
CAUM	8 / 8 / 8	200 / 200 / 200	–	–	–
LSTUR	16 / 8 / 8	150 / 50 / 200	–	–	–
MANNeR	–	–	–	–	0.38 / 0.38 / 0.38
MINER	–	–	8 / 32 / 48	mean / mean / weighted	–
MINS	32 / 16 / 32	50 / 100 / 200	–	–	–
NAML	24 / 24 / 8	200 / 50 / 200	–	–	–
TANR	24 / 12 / 16	150 / 200 / 100	–	–	–

Received 29 December 2024; revised 18 September 2025; accepted 11 November 2025