Evaluation and simplification of text difficulty using LLMs in the context of recommending texts in French to facilitate language learning

Henri Jamet
DESI, Faculty of Business and Economics, University of Lausanne
Switzerland
henri.jamet@unil.ch

Maxime Manderlier
Technological Innovation Management Unit, Faculty of Engineering, University of Mons (UMONS)
Belgium
maxime.manderlier@umons.ac.be

Yash Raj Shrestha
DESI, Faculty of Business and Economics, University of Lausanne
Switzerland
yashraj.shrestha@unil.ch

Michalis Vlachos
DESI, Faculty of Business and Economics, University of Lausanne
Switzerland
michalis.vlachos@unil.ch

Abstract

Learning a new language can be challenging. To help learners, we built a recommendation system that suggests texts and videos based on the learners' skill level of the language and topic interests. Our system analyzes content to determine its difficulty and topic, and, if needed, can simplify complex texts while maintaining semantics. Our work explores the holistic use of Large Language Models (LLMs) for the various sub-tasks involved for accurate recommendations: difficulty estimation and simplification, graph recommender engine, topic estimation. We present a comprehensive evaluation comparing zero-shot and fine-tuned LLMs, demonstrating significant improvements in French content difficulty prediction (18−56%), topic prediction accuracy (27%), and recommendation relevance (up to 18% NDCG increase).

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

digital education, extensive reading, machine learning, large language models

ACM Reference Format:
Henri Jamet, Maxime Manderlier, Yash Raj Shrestha, and Michalis Vlachos. 2024. Evaluation and simplification of text difficulty using LLMs in the context of recommending texts in French to facilitate language learning. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640457.3688181

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688181

1 Introduction

Research has shown that reading texts in a foreign language is extremely beneficial to its learning [34], especially when the content is of interest to the reader [4]. With this in mind, we have designed a recommendation system to facilitate language learning. The system selects texts adapted to each user's level of linguistic competence and aligned with their personal interests. To easily enrich the content used by this system, we envisaged an autonomous approach that does not require pre-labeled data and can autonomously recommend a variety of texts. We have identified four essential components for the development of this system:

(1) **A model for evaluating the complexity of a text** *(to discover content that matches the user's knowledge of a foreign language)*
(2) **A model for simplifying a text while preserving its meaning** *(to increase the volume of available content and improve the relevance of recommendations)*
(3) **A model for classifying text in various domains** *(to present only content that matches the user's preferences)*
(4) **A recommendation system** *(which uses the results of the three previous models to rank and suggest the most suitable content)*

These tasks can largely benefit from the use of LLMs (Large Language Models). We have evaluated the performance that small, specialized models can offer in comparison with larger, more general models. To test the viability of our recommendation system, we developed an LLM-driven method for each task and measured the performance to different state-of-the-art models. Our tests were carried out on French texts, but the underlying methods are language-agnostic and can be applied to other foreign languages.

Contributions of this work include: a) A method for automatically evaluating the difficulty of a foreign text, and a method for simplifying and reducing the linguistic difficulty, if deemed necessary (e.g. discovering content that would be on a topic relevant for the user, but it is too difficult for the user to understand). b) A comparative performance analysis of models of different sizes, with and without fine tuning on the four defined tasks. c) The design

of an architecture that facilitates automatic text recommendation based on the user's language proficiency level and interests.

2 Related Work

For **difficulty estimation**, readability formulas like Flesch-Kincaid, SMOG, and Gunning Fog Index have been extended to various languages but primarily target native speakers [9, 24, 35]. Machine learning techniques using syntactic complexity, word frequency, and semantic similarity offer another approach [9, 16]. Recent advances include integrating pre-trained embeddings into readability models [19, 20, 22, 33], though the use of LLMs for difficulty prediction is underexplored. Our work demonstrates that LLMs can significantly boost accuracy in this area.

Text classification involves categorizing texts into fixed classes, with large-scale LLMs generally surpassing traditional models in tasks requiring extensive language knowledge such as estimating text difficulty [5, 31, 32]. These larger models typically outperform smaller counterparts like BERT, especially in zero-shot learning scenarios [6, 7, 11, 12, 17, 21, 29]. However, for topic classification tasks, this performance advantage becomes more ambiguous. Our research confirms these patterns within the context of French language texts. *Evaluating text simplification* systems relies on robust metrics. While traditional metrics such as **BLEU** and **ROUGE** are less effective for simplification, the **SARI** metric evaluates the quality of modifications, and **QUESTEVAL** uses semantic questioning, better aligning with human judgment [2, 28].

Text summarization aims to condense content, maintaining critical points, whereas simplification reduces complexity, enhancing readability [28]. *Lay summarization* combines these approaches using models like **BERT** and **PEGASUS** to make technical content accessible [8, 13, 30, 36]. Advances in LLMs have markedly affected *text simplification*, with models like **GPT-3** showing notable efficacy in simplification tasks [15, 26].

In **recommender systems**, one can use pre-trained embedding to capture the semantic of content. Graph-based recommendation methods have been extensively studied [14, 23, 37], but the use of LLMs for generating rich embeddings is recent. Studies show that LLM-generated embeddings enhance graph-based methods [10]. Our system leverages LLM embeddings to improve recommendation quality and personalization.

3 Methodology

3.1 Difficulty estimation

For a foreign language recommendation system, it is crucial to discover content that aligns with the learner's knowledge of the foreign language. Our methodology aims to create a recommendation system that provides texts at an appropriate level to help learners progress in French. The difficulty estimation of foreign text is treated as a classification problem, in which we seek to predict the CEFR difficulty level of a given text.

This difficulty estimation system serves two primary purposes: to recommend appropriate texts to readers based on their proficiency level, and to evaluate the effectiveness of our text simplification models, which will be discussed in subsequent sections. Within the recommendation system, it will be used to assess the level of all texts, enabling their organization, classification, and ranking to suggest texts tailored to the reader's level.

To develop and evaluate our models, we used three different annotated datasets with distinct characteristics: 1) `LjL` from [19], containing items with multi-sentence content and leveled labels; 2) Internet-derived sentences (`sentencesInternet`), annotated by university students, utilized fully if consensus was achieved on difficulty, matched to CEFR levels; and 3) literature-sourced sentences (`sentencesBooks`) labeled by a language professor, using error-free outputs from an OCR process, also aligned with CEFR standards. The models were tested using an 80/20 train-test split.

We benchmarked our difficulty estimator against readability metrics (FKGL, GFI, ARI) noted in Table 1, adapting these for French. Traditional metrics, predictive of a continuous difficulty scale, contrast with our discrete label approach. To align with our methodology, we trained a logistic regression model, effectively transforming regression into classification for comparative evaluation.

Additionally, we tested various Large Language Models (LLMs), including those trained specifically on French data. These models, fine-tuned on the task-specific dataset, were evaluated for F1-score in varied contexts (with and without specific training scenarios), as shown in Table 1. This comprehensive evaluation allows us to assess the capabilities of existing general models in predicting appropriate levels when given a text, which is crucial for both recommending suitable texts to readers and evaluating the efficacy of our text simplification models in the following section.

3.2 Text simplification

In a recommendation setting, we may discover content that is relevant for a user (based on the declared topics of interests), but may be too difficult for the user to understand. In this case, the text can be further simplified.

We model automatic text simplification as a sequence-to-sequence text generation problem on which we fine-tuned an LLM using pairs of sentences with the format "*original sentence → simplified sentence with exactly one CEFR level lower than the original sentence*". In experiments, post-training with only 125 sentence pairs yielded significant improvements over a zero-shot LLM approach. We note, that the simplification we perform is sentence-by-sentence.

A central challenge in text simplification is maintaining the original text's semantics while simplifying effectively. We introduce two metrics to address this: simplification accuracy and semantic similarity. **Simplification accuracy** (A) is defined as the sum of the two-by-two product of the cumulative probabilities of belonging to each class between the original sentence (from level *A2* to *C2*) and the simplified sentence (from level *A1* to *C1*). It's a score between 0 and 1, reflecting the probability of the simplified text being *at least one* CEFR difficulty level lower than the original.

To calculate this probability, we use the **CamemBERT** model trained on the `SentencesBooks` dataset, which offers the best performance while remaining relatively lightweight. We use this model to estimate the CEFR difficulty level of both the simplified and original sentences. For enhanced precision in our evaluation, we utilize the logits from the classification model to compute the probability that the generated text is simpler than the original, independently of the actual level of the original text. This approach not only helps

reduce biases inherent in our difficulty estimation model but also provides a more nuanced assessment of the simplification process.

Semantic similarity (S), ranging from 0 to 1, measures how closely the semantic content of the simplified text aligns with the original, using the cosine similarity of their embeddings. We combine these metrics into a weighted score:

$$\text{w-Score} = 2 \times \frac{w_1 \times A \times w_2 \times S}{w_1 \times A + w_2 \times S}$$

where $w_2 = (1 - w_1) = 0.5$, balancing both metrics equally in our tests.

All models were evaluated from a zero-shot perspective, with the best result retained for each model. We also assessed the performance of these models after fine-tuning, using the datasets described below, with the exception of **GPT-4** and **GPT-4o**, for which fine-tuning capabilities were not available at the time of writing. It's worth noting that Davinci was the only model whose zero-shot results were too poor to allow for meaningful interpretation, and thus was only evaluated in its fine-tuned state. Datasets used are described below:

Training-set. To fine-tune our models for the task of simplification, we need a dataset of French sentences with their simplifications at an associated lower CEFR level. We used **GPT4** to generate 125 sentences (*25 from each level* `A2`, `B1`, `B2`, `C1`, `C2`) and their simplified versions. This dataset was further reviewed by a native French speaker.

Test-set. We take, per difficulty level `A2`, `B1`, `B2`, `C1`, `C2` (*Level A1 cannot be simplified*), 100 random sentences from the `sentencesBooks` and `sentencesInternet` dataset. The test-set consists of $5 \times 100 \times 2 = 1000$ sentences.

3.3 Topic classification

In a recommendation setting, the user will declare expertise of a language and topical interests, and the system will discover relevant content. We can use LLMs to accurately predict the topic of a textual content. The LLM will act as a label predictor for the topic. We trained and evaluated the LLM performance using data from an existing recommendation platform focusing on language learning and reading behavior, called Zeeguu, which recommends news articles. The dataset contains 1743 {Text, Label} pairs, split into training (80%) and testing (20%) sets. Labels encompassed 11 categories: World, Travel, Music, Culture, Business, Food, Sport, Politics, Health, Science, and Technology. Our methodology involved training and evaluating different models on a classification task aimed at associating each text with its preponderant topic. We explored various approaches, including zero-shot inference, fine-tuning, and adapting pre-trained models using logistic regression. To comprehensively assess the task, we selected models of different sizes and architectures, allowing us to investigate the trade-offs between model complexity and performance in this specific domain. The results were evaluated using accuracy metrics, including top-1, top-3, and top-5 accuracy.

3.4 Recommender System

In our proposed system, we aim to harness the strengths of both content-based filtering and collaborative filtering to recommend data more effectively. This dual approach allows us to leverage a comprehensive set of data, thereby enhancing the accuracy and relevance of our recommendations. Our strategy focuses on augmenting LightGCN [18], a widely-used graph convolutional network model that traditionally handles only user-item interactions. Our approach enhances LightGCN by using embeddings based on Large Language Models (LLMs) to represent the items of the recommendation system.

3.4.1 Item Embeddings. We use LLMs to analyze and understand the semantic content of the items. By inputting the item texts into an LLM, we generate high-quality, representative embeddings for each item. These embeddings encapsulate the nuanced meanings and relationships inherent in the text, providing a richer context for each item.

3.4.2 User Embeddings. To represent each user, we generate user embeddings based on the last n items a user has interacted with. The value of n will be varied and tested in the results section to identify the optimal configuration. By aggregating the embeddings of these last n items, we construct a representative embedding for each user. This method ensures that the user embeddings are informed by the user's recent interactions, providing a dynamic and contextually relevant representation.

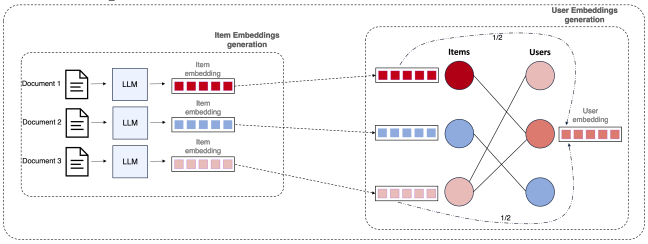

Figure 1: The process of generating item embeddings using large language models (LLMs) and combining them to form user embeddings.

Instead of relying on random initialization, we use the embeddings generated by the LLMs as the starting point for LightGCN. This approach places items and users in an embedding space where similar nodes are naturally proximate, establishing a more coherent and meaningful starting structure. The process of initializing embeddings is illustrated in Figure 1.

LightGCN then processes these semantically enriched embeddings, refining them through the learning of collaborative links between users and items. This iterative learning process enables the system to fine-tune the embeddings based on user interaction data, progressively improving the recommendation quality. By enriching LightGCN with semantically informed embeddings from LLMs, our system combines the depth of content-based analysis with the breadth of collaborative filtering. This synergy not only optimizes the initial placement of items and users in the embedding space but also enhances the model's ability to learn and predict user preferences, resulting in superior recommendation performance. The process by which LightGCN generates recommendations is briefly summarized in Figure 2.

4 Results

A video of the interface demonstrating the subsection 4.1 (difficulty estimation) and subsection 4.2 (text simplification) parts can be

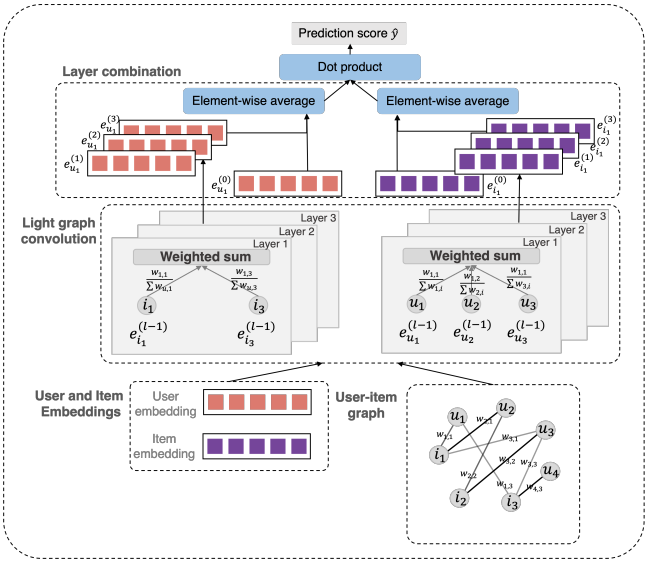

Figure 2: Generating a prediction score for user preference towards an item using LightGCN.

viewed here. The project code, including data preparation, model training and evaluation, and figure creation, can be found here.

4.1 Difficulty estimation

Our results, presented in Table 1, reveal that LLMs outperform standard readability indices significantly. The addition of a context sentence, can further improve the LLMs' classification accuracy by leveraging their understanding of CEFR notation, as evidenced for the **Mistral-7B** model. The **GPT3.5** model with context consistently exhibits superior performance. Notably, the **CamemBert** model, despite its smaller size, achieves the highest accuracy on the *SentencesBooks* dataset. To evaluate model performance in this CEFR level classification task, we use accuracy as our primary metric. Accuracy is determined by calculating the proportion of correct predictions across all CEFR levels, which is derived from the confusion matrix by dividing the sum of its diagonal elements (representing correct classifications) by the total number of predictions.

Table 1: Difficulty estimation metrics for all datasets.

model	context	LjL	SentencesInternet	SentencesBooks
GPT-3.5[1]	✓	0.72	**0.90**	0.50
	-	**0.73**	0.87	0.49
BERT[1]	-	0.62	0.82	**0.52**
Mistral[1]	✓	0.64	0.75	0.51
Davinci[1]	-	0.59	0.82	0.47
	✓	0.61	0.81	0.47
Mistral[1]	-	0.47	0.63	0.35
FKGL	-	0.42	0.34	0.35
GFI	-	0.45	0.32	0.34
ARI	-	0.40	0.34	0.34

[1] In this figure, "GPT3.5" corresponds to `gpt-3.5-turbo-0613`, "BERT" to `camembert-base`, "Mistral" to `Mistral-7B`, and "Davinci" to `davinci-002`.

A sample context for the LLM is shown here[2]:

```
You are a linguistic expert specialized in evaluating French language levels according to the Common European Framework of Reference for Languages (CEFR). Your task is to classify the following French text into one of the CEFR levels: A1, A2, B1, B2, C1, or C2. Respond ONLY with the most appropriate level label, without any explanation or additional text.

Example:
Text to classify: "Bonjour, je m'appelle Jean. J'habite à Paris. J'aime jouer au football."
CEFR Level: A1

Now, classify this French text:
{{ text_to_classify }}

CEFR Level:
```

4.2 Text Simplification

Table 2 demonstrates that all tested LLMs, except **Davinci**, perform well in simplification tasks. The fine-tuned models perform better in general, in particular **Mistral-7B** whose fine-tuned version is much better and is our best model for this task. **GPT-4** and **GPT-4o** showed unexpectedly weak results, possibly due to unsolicited contextual phrases in its outputs, which were retained in our evaluation for consistent comparison. In Table 2 and Figure 3 we present the different components of the metric `w-score` introduced in subsection 3.2.

Table 2: Results for the text simplification task.

model	fine-tuned	Simplification Accuracy	Semantic Similarity	Weighted-Score
Mistral[2]	✓	**0.59**	0.91	**0.72**
GPT-3.5[2]	✓	0.57	0.91	0.70
GPT3.5[2]	-	0.53	0.93	0.67
GPT-4	-	0.51	**0.93**	0.66
Mistral[2]	-	0.47	0.93	0.63
GPT-4o	-	0.46	0.89	0.60
Davinci[2]	✓	0.44	0.83	0.57

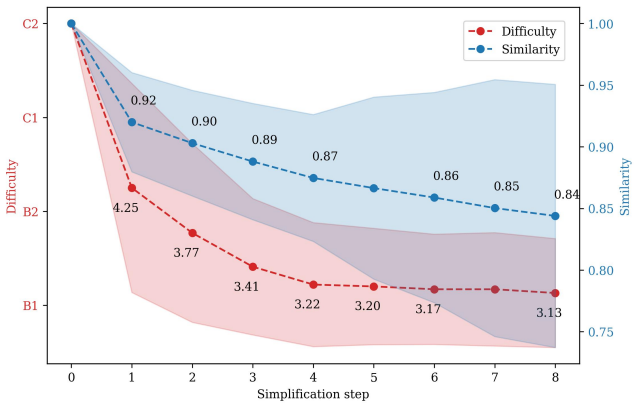

Figure 3: Iterative CEFR C2 sentence simplification using Mistral-7B, with CamemBERT for difficulty estimation and cosine similarity for evaluating text similarity. Results averaged over 100 experiments.

[2] The original was in French and has been translated here for readability purposes. Note that this context had to be slightly modified for the LjL dataset as it did not use the CEFR classification system, but the structure remained the same.

4.3 Topic classification

We evaluate how accurate an LLM can predict the topic of a text. On Table 3, we see a clear superiority of the fine-tuned version of the **Flaubert** model, which is specialized for French text, over all other zero-shot models, despite their considerable difference in size. These results suggest that small, specialised models are preferable to much larger, more general models for specific classification tasks, a conclusion that is in line with other recent studies [3], [1], [25].

Table 3: Results for topic classification accuracy.

model	accuracy
Flaubert-fine-tuned	**0.74**
GPT-4-turbo-2024-04-09	0.61
GPT-4o-2024-05-13	0.61
GPT-3.5-turbo-1106	0.58
Flaubert-pretrained	0.56
mDeBERTa	0.45
Davinci-002	0.09

4.4 Recommender System

We evaluate the recommendation accuracy and related metrics using LLM-enhanced recommendation models across diverse datasets, including Zeeguu, ml-100k, Goodreads, and Tomplay. Table 4 summarizes these datasets.

Table 4: Overview of datasets used for the recommendation task.

Dataset	# Users	# Items	# Interactions	Sparsity	Test data
Zeeguu	1,574	21,030	25,227	99.92%	20.77%
ml-100k	942	1,447	55,375	95.94%	20.01%
Goodreads	39,628	92,131	504,585	99.99%	22.37%
Tomplay	35,028	30,050	724,651	99.93%	19.96%

4.4.1 Data Pre-processing. Datasets were split into training and test sets, excluding users with a single interaction and aiming for an 80/20 train-test ratio.

4.4.2 Embeddings. Item embeddings were generated using BERT and ADA v2, known for their semantic capabilities. User embeddings were computed as the mean of the last n items viewed by a user, with smaller n values generally yielding better performance.

4.4.3 Models. We evaluated ALS, BPR, LMF, and LightGCN to establish baselines and measure the improvements achieved with precomputed embeddings in LightGCN. Recent studies have reaffirmed the relevance of these models, particularly ALS [27].

4.4.4 Results. In Table 5, we observe consistent improvements in LightGCN's performance with pre-trained embeddings over Xavier initialization, particularly for text-rich datasets like Zeeguu and Goodreads. For Zeeguu, ADA shows a 12.12% improvement in NDCG@5 and BERT a 1.67% increase, highlighting the influence of high-quality textual information. In the ml-100k dataset, ADA saw a 0.52% increase and BERT a 2.59% increase. For Goodreads, ADA improved NDCG@5 by 18.75% and BERT by 13.7%. In the Tomplay dataset, ADA showed a 1.54% increase and BERT a 1.18% increase, likely due to the more compact textual content of the dataset. Despite this, both embeddings significantly outperform the Xavier initialization, emphasizing the value of pre-trained embeddings across different datasets.

Table 5: Evaluation Results.

Dataset	Model	Recall@5	Precision@5	F1@5	NDCG@5	MRR@5	MAP@5
Zeeguu	ALS	0.0626	0.0236	0.0309	0.0595	0.0774	0.0498
	BPR	0.0182	0.0061	0.0082	0.0168	0.0191	0.0146
	LMF	0.0251	0.0084	0.0116	0.0235	0.0291	0.0202
	LightGCN ADA (9 layers)	**0.0721**	**0.0274**	**0.0352**	**0.0666**	**0.0843**	**0.0550**
	LightGCN Xavier ADA (10 layers)	0.0630	0.0238	0.0308	0.0594	0.0759	0.0498
	LightGCN Bert (10 layers)	0.0659	0.0249	0.0323	0.0609	0.0788	0.0501
	LightGCN Xavier Bert (10 layers)	0.0622	0.0231	0.0300	0.0599	0.0765	0.0508
ml-100k	ALS	0.0767	0.1285	0.0806	0.1414	0.2504	0.0850
	BPR	0.0551	0.0975	0.0580	0.1084	0.1995	0.0634
	LMF	0.0381	0.0635	0.0400	0.0704	0.1419	0.0379
	LightGCN ADA (2 layers)	**0.0903**	**0.1348**	**0.0882**	**0.1556**	0.2682	**0.0975**
	LightGCN Xavier ADA (2 layers)	0.0881	0.1346	0.0869	0.1548	0.2720	0.0958
	LightGCN Bert (2 layers)	0.0855	0.1316	0.0847	0.1547	**0.2761**	0.0971
	LightGCN Xavier Bert (1 layer)	0.0855	0.1323	0.0847	0.1508	0.2636	0.0922
Goodreads	ALS	0.0634	0.0156	0.0237	0.0447	0.0419	0.0366
	BPR	0.0447	0.0114	0.0168	0.0323	0.0323	0.0261
	LMF	0.0432	0.0099	0.0155	0.0280	0.0245	0.0221
	LightGCN ADA (2 layers)	**0.0732**	**0.0183**	**0.0275**	**0.0513**	**0.0481**	**0.0415**
	LightGCN Xavier ADA (3 layers)	0.0603	0.0155	0.0231	0.0432	0.0415	0.0352
	LightGCN BERT (5 layers)	0.0673	0.0170	0.0254	0.0473	0.0449	0.0382
	LightGCN Xavier BERT (2 layers)	0.0570	0.0149	0.0220	0.0416	0.0405	0.0340
Tomplay	ALS	0.0875	0.0598	0.0650	0.0838	0.1408	0.0528
	BPR	0.0295	0.0236	0.0246	0.0306	0.0583	0.0178
	LMF	0.0327	0.0249	0.0263	0.0323	0.0598	0.0187
	LightGCN ADA (2 layers)	0.0958	0.0646	0.0706	0.0924	0.1541	0.0595
	LightGCN Xavier ADA (3 layers)	0.0943	0.0643	0.0699	0.0910	0.1528	0.0581
	LightGCN Bert (2 layers)	**0.0984**	**0.0660**	**0.0722**	**0.0947**	**0.1582**	**0.0608**
	LightGCN Xavier Bert (2 layers)	0.0966	0.0656	0.0715	0.0936	0.1566	0.0601

Additionally, the Zeeguu dataset benefits from a higher number of layers in LightGCN, indicating that increased model depth enhances recommendation accuracy in sparse datasets.

5 Conclusion

This study lays the groundwork for a novel, cost-effective approach to foreign language learning by combining advanced natural language processing techniques with hybrid recommendation systems. We have developed an integrated system comprising a robust LLM-based method for estimating text difficulty, a meaning-preserving text simplification algorithm, and an efficient topic classification system. These elements have been integrated into a hybrid recommendation framework using LightGCN with pre-trained embeddings, designed to operate with limited computational resources and training data.

Our experiments demonstrate that small, specifically fine-tuned language models can compete with larger, more general models for specialized tasks, offering a balance between performance and resource efficiency. The resulting system can recommend content adapted to the learner's linguistic level while respecting their thematic interests.

Although our study focused on French, the proposed approach is designed to be language-independent. The significant improvements observed in recommendation performance, particularly for text-rich datasets, underscore the effectiveness of our approach. User studies could further validate these results by evaluating the system's effectiveness in real-world language learning scenarios. Additionally, extending the evaluation to multiple languages and comparing with state-of-the-art models could strengthen the validity and applicability of our approach.

This study provides a solid foundation for future research in personalized language learning systems and other areas of AI-assisted education.

References

[1] [n. d.]. Small Triumphs Over Large: Instances Where BERT-Based Fine-Tuned Models Surpass GPT-4 in Classification Tasks — papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4780220. [Accessed 16-05-2024].

[2] Suha Al-Thanyyan and Aqil M. Azmi. 2021. Automated Text Simplification. *ACM Comput. Surv.* (2021).

[3] Kimia Ameri, Michael Hempel, Hamid Sharif, et al. 2021. CyBERT: Cybersecurity Claim Classification by Fine-Tuning the BERT Language Model. *Journal of Cybersecurity and Privacy* 1, 4 (2021), 615–637. https://doi.org/10.3390/jcp1040031

[4] Majid Asgari, Saeed Ketabi, and Zahra Amirian. 2019. Interest-Based Language Teaching: Enhancing Students' Interest and Achievement in L2 Reading. (2019). https://doi.org/10.30466/IJLTR.2019.120633

[5] Bart Bonikowski, Yuchen Luo, and Oscar Stuhler. 2022. Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models. *Sociological Methods & Research* 51, 4 (2022), 1721–1787.

[6] Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Youngjin Chae and Thomas Davidson. 2023. Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning. (08 2023). https://doi.org/10.31235/osf.io/sthwk

[8] Rochana Chaturvedi, S. Jaspreet Singh Dhani,et al. 2020. Divide and Conquer: From Complexity to Simplicity for Lay Summarization. *SDP* (2020).

[9] Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*. 113–119.

[10] Zhikai Chen, Haitao Mao, Hang Li, et al. 2024. Exploring the Potential of Large Language Models (LLMs) in Learning on Graphs. arXiv:2307.03393 [cs.LG]

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116 [cs.CL]

[12] Alexis Conneau, Guillaume Lample, Ruty Rinott, et al. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. arXiv:1809.05053 [cs.CL]

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). 4171–4186.

[14] Shaohua Fan, Junxiong Zhu, Xiaotian Han, et al. 2019. Metapath-guided Heterogeneous Graph Neural Network for Intent Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2478–2486. https://doi.org/10.1145/3292500.3330673

[15] Yutao Feng, Jipeng Qiang, Yun Li, et al. 2023. Sentence Simplification via Large Language Models. *ArXiv* (2023).

[16] Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 466–477.

[17] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543 [cs.CL]

[18] Xiangnan He, Kuan Deng, Xiang Wang, et al. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 639–648. https://doi.org/10.1145/3397271.3401063

[19] Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. 2022. Open corpora and toolkit for assessing text readability in French. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*. 54–61.

[20] Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935* (2021).

[21] Hang Le, Loïc Vial, Jibril Frej, et al. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. arXiv:1912.05372 [cs.CL]

[22] John SY Lee. 2022. An editable learner model for text recommendation for language learning. *ReCALL* 34, 1 (2022), 51–65.

[23] Siwei Liu, Iadh Ounis, Craig Macdonald, et al. 2020. A Heterogeneous Graph Neural Model for Cold-start Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2029–2032. https://doi.org/10.1145/3397271.3401252

[24] Nadezda Okinina, Jennifer-Carmen Frey, and Zarah Weiss. 2020. CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020*. 7123–7131.

[25] Youri Peskine, Damir Korenčić, Ivan Grubisic, et al. 2023. Definitions Matter: Guiding GPT for Multi-label Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4054–4063. https://doi.org/10.18653/v1/2023.findings-emnlp.267

[26] Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[27] Steffen Rendle, Walid Krichene, Li Zhang, et al. 2022. Revisiting the Performance of iALS on Item Recommendation Benchmarks. In *Proceedings of the 16th ACM Conference on Recommender Systems* (<conf-loc>, <city>Seattle</city>, <state>WA</state>, <country>USA</country>, </conf-loc>) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 427–435. https://doi.org/10.1145/3523227.3548486

[28] Thomas Scialom, Louis Martin, Jacopo Staiano, et al. 2021. Rethinking Automatic Evaluation in Sentence Simplification. *ArXiv* (2021).

[29] Xiaofei Sun, Xiaoya Li, Jiwei Li, et al. 2023. Text Classification via Large Language Models. arXiv:2305.08377 [cs.CL]

[30] Oliver Vinzelberg, M. Jenkins, Gordon Morison, et al. 2023. Lay Text Summarisation Using Natural Language Processing: A Narrative Literature Review. *arXiv.org* (2023).

[31] Sandra Wankmüller. 2022. Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis. arXiv:2102.02111 [cs.CL]

[32] Tobias Widmann and Maximilian Wich. 2022. Creating and Comparing Dictionary, Word Embedding, and Transformer-based Models to Measure Discrete Emotions in German Political Text. *SSRN Electronic Journal* (01 2022). https://doi.org/10.2139/ssrn.4127133

[33] Rodrigo Wilkens, David Alfter, Xiaoou Wang, et al. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1217–1233.

[34] M. Wright and Pamela Brown. 2006. Reading in a modern foreign language: exploring the potential benefits of reading strategy instruction. (2006). https://doi.org/10.1080/09571730685200071

[35] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580* (2019).

[36] Jingqing Zhang, Yao Zhao, Mohammad Saleh, et al. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv:1912.08777 [cs.CL]

[37] Jun Zhao, Zhou Zhou, Ziyu Guan, et al. 2019. IntentGC: A Scalable Graph Convolution Framework Fusing Heterogeneous Information for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2347–2357. https://doi.org/10.1145/3292500.3330686