Disability Assessment and Prediction of Academic Accommodations for Students in Post-Secondary Education

Rakshil Kevadiya[1], Abbas Akkasi*[1], Kathleen C. Frase[2], Boris Vukovic[3], Jessie Gunnell[3], Sonia Tanguay[3]
Behdin Nowrouzi-Kia[4] and Majid Komeili[1]

***Abstract*— The rapidly increasing diversity of student support needs and the resource-intensive nature of current support systems underscore the urgency to improve academic inclusivity for students with disabilities in post-secondary education. Our interdisciplinary collaboration between AI scientists and disability service practitioners investigates the use of various AI models to analyze functional limitations and recommend Academic Accommodations (AAs) in a reliable and equitable manner. A chatbot-based interface was developed to autonomously collect both qualitative and quantitative data through WHODAS 2.0 questionnaires and targeted follow-up questions, enabling a comprehensive assessment of functional limitations. Experimental results highlight the significance of textual data in capturing nuanced functional limitations, aligning with human assessment practices. By integrating LLM-generated explanations into the evaluation framework, this research demonstrates the potential of LLMs to enhance consistency and transparency of decision-making in disability services. The findings contribute to the fields of AI and disability services by showcasing the feasibility of highly interdisciplinary, scalable, explainable, and data-driven approaches to service recommendations to address a pressing challenge in higher education.**

I. INTRODUCTION

Post-secondary education (PSE) plays a pivotal role in promoting employment, well-being, and quality of life for individuals with disabilities [1], [2], [3]. Supporting students with disabilities in higher education is not only a mandated human right but also aligns with the United Nations Social Development Goals, which emphasize inclusive quality education across the lifespan [4], [5].

The assessment of functional limitations and strengths is essential for identifying student needs and determining appropriate supports. The WHODAS 2.0 [6], a standardized disability assessment tool, evaluates the degree of self-reported impairment in life activities across various domains. In PSE settings, students seeking Academic Accommodations (AAs), such as extended time for exams, are required by disability service offices (DSO) to provide disability documentation that describes the type and degree of functional limitations. The WHODAS 2.0 questionnaire is one way of collecting this information followed by consultations with disability service providers to formalize support recommendations.

However, disability support services in higher education face increasing challenges in meeting the needs of a growing population of students with disabilities. In the last 10 years in Canada, the proportion of university students self-identifying with a disability has increased from 9 percent to 31 percent [7]. For the same time period, the data from DSOs shows registrations for disability services increasing from 6 percent of the total student population to 12 percent in universities and 14 percent in colleges [8]. It is of note that approximately half of students who self-identify as having disabilities remain unregistered with DSO services. This rising prevalence has heightened the difficulty of promptly assessing functional limitations and providing necessary academic accommodations in universities and colleges.

The significant representation of students with disabilities underscores the need for enhanced accessibility efforts to ensure equitable inclusion [9]. Disability-related services in PSE are critical for student academic success and long-term quality of life. However, the current process for determining AAs is resource-intensive and time-consuming. Students must register with their institution's DSO and undergo evaluations that include in-person interviews with practitioners or healthcare professionals. These assessments, which utilize tools like the WHODAS 2.0 questionnaire and supplementary discussions, aim to identify functional limitations and strengths, leading to tailored AA recommendations.

With the growing population of post-secondary students with disabilities, the traditional approach has notable limitations. Its reliance on manual, face-to-face evaluations is not only time-intensive but also strains institutional resources, particularly given the increasing caseloads for service providers [10]. The process lacks scalability and efficiency, often resulting in delays that leave students without adequate support at the start of their academic programs, exacerbating inequities. There is also variability in expertise between junior and more experienced service providers, introducing potential variability in the evaluation of student needs and quality of service provision. Increasing workload reduces opportunities and capacity for professional development and mentoring.

The need for a more efficient, scalable, and reliable approach to assess functional limitations and recommend AAs is clear. Integrating AI methods into disability services can enable a timely and standardized approach to student support. However, there are important considerations, chal-

[1]School of Computer Science, Carleton university, Canada name.family@carleton.ca
[2] Computer Engineering Department, University of Ottawa, Canada kathleen.fraser@uottawa.ca
[3]Accessibility Institute, Carleton University, Canada name.family@carleton.ca
[4]Occupational Science and Occupational Therapy, University of Toronto, Canada behdin.nowrouzi.kia@utoronto.ca

lenges, and opportunities related to the ethical use of AI in disability services, risks of both AI and human biases in the assessment of needs and recommendation of services, and effective interdisciplinary collaboration between researchers and practitioners.

This paper explores the use of chatbots for autonomous data collection and evaluates various AI models, categorized into parameter-tuning and generative approaches, to generate tailored AA recommendations along with explanations for their decision-making processes. The findings underscore the potential of large language models (LLMs) as consultative tools for service providers, delivering explainable and personalized recommendations tailored to individual student profiles. Incorporating Accommodation Descriptions (ADs) as context for zero-shot learning with the open-source language model LLaMA-3.1 achieved an F1-score of 85%, demonstrating the capability of pre-trained LLMs to effectively address the challenge of AA recommendations. This result highlights the potential of leveraging structured domain knowledge alongside powerful pre-trained models to enhance decision-making accuracy and reliability.

II. Background

The application of machine learning (ML) in disability assessment and services has gained considerable attention, particularly in areas such as learning disabilities, disability progression, and functional impairment prediction [11].

A study by [12] investigates the use of various ML algorithms to predict learning disabilities in children, evaluating models including K-Nearest Neighbors (KNN) [13], Random Forest, and Convolutional Neural Networks (CNN) [14]. Their findings indicate that KNN achieved the highest accuracy of 90.33%, while also emphasizing the critical need for ethical considerations, particularly regarding privacy and autonomy, in ML-based disability prediction. Similarly, [15] explores ML models to predict disability progression in individuals with multiple sclerosis. Utilizing metrics such as ROC-AUC and Brier score, the study demonstrates the reliability of ML models in forecasting disability progression, thereby aiding clinicians in making informed treatment decisions.

Kalite et.al [16], combined LSTMs with explainable AI (SHAP values) to achieve accurate and interpretable predictions that support targeted educational interventions.

In a study by [17], ML algorithms such as ridge regression and gradient boosting were employed to predict the risk of functional disability in older adults over a five-year period, demonstrating effective performance. Key predictors, including age, self-rated health, and physical activity, were identified as significant contributors to the accuracy of disability risk assessments. The findings suggest that incorporating these variables as important features can enhance the prediction of functional disability. Similarly, [18] proposed an XGBoost-based model to predict disability risk in healthy older adults over a three-year period. The study highlights features such as hand grip strength and depressive symptoms as critical indicators and utilizes SHAP values to provide interpretable explanations for model predictions.

Further research has focused on developing frameworks for applying AI-based methodologies in the detection and diagnosis of disabilities. For instance, [19] explores a framework for leveraging AI in diagnosing intellectual disabilities. Additionally, [20] introduces an AI-driven framework for permanent impairment evaluation, integrating the International Classification of Diseases (ICD) and the International Classification of Functioning, Disability, and Health (ICF) to standardize disability assessments. This framework supports the development of models for consistent and objective evaluations of permanent impairments.

Transformers have emerged as a powerful tool for analyzing disability-related data, showcasing notable progress in managing complex and heterogeneous datasets. For example, [21] employed neural networks to predict disability levels among the elderly in England using heterogeneous data. Three models—Wide & Deep [22], TabTransformer [23], and TabNet [24]—were evaluated, with disability levels categorized into binary, 3-level, and 4-level classifications based on mobility and daily activity limitations. TabNet outperformed the others, particularly in binary classification, identifying key predictors such as urinary incontinence, smoking, exercise, and education as influential factors. These findings underscore the adaptability of neural networks for mixed data types and their potential in identifying critical predictors of disability.

A notable observation, however, is that the majority of existing studies focus on diagnosing disabilities or functional impairments within highly specific populations. To the best of our knowledge, no prior work has addressed the assessment of functional limitations for recommending disability services, whether in PSE or other domains. This gap highlights the need for research in this area to develop scalable and equitable solutions for disability service provision.

III. Methods

Our highly interdisciplinary team set out to address the challenges faced by disability services in higher education by investigating the feasibility of an AI-driven system for the evaluation of student functional needs and recommendation of support services. The team consists of academic researchers, subject matter experts, and practitioners from the fields of computer science and AI, disability services, education, psychology, engineering, human rights, inclusive design, and occupational therapy.

The project was undertaken in partnership with the host university's DSO and with approval from the institutional Research Ethics Board. A core group of researchers, graduate students, and practitioners developed and performed the study in a funding timeframe from spring 2022 to spring 2025, with the participation of university students with disabilities.

The design of the study required continuous interdisciplinary analyses of disability service processes and input from practitioners related to the assessment of functional

limitations and service recommendations. From the technical perspective, the primary task involved classifying student profiles of functional limitations based on 21 academic accommodations, which can be done using both binary and multi-label classification frameworks. We tackled this problem with two distinct approaches: a parameter-tuning approach and a generative models approach. Both methods utilized data collected by a novel chatbot designed specifically for interacting with students and gathering their profile information.

A. Data Collection Interface Design

In disability services, a key aspect to assessing individual needs and tailoring supports is the interaction between the practitioner and the student. Follow-up questions are essential for capturing detailed insights into the functional limitations, as well as strengths, to determine the most relevant services. Therefore, the first step in designing the data collection interface was to determine what follow-up questions should be asked in response to the user's WHODAS 2.0 input.

The host institution's DSO provided the team with access to both the archival WHODAS 2.0 data from the past five years and to the current students for new data collection. The data preparation process encompassed archival data subsampling, analysis, data collection and pre-processing as well.

The archival data analysis included 4,592 WHODAS 2.0 responses from host university students (2017–2022), with a modified domain structure to emphasize academic functionality. Using K-Means [25] and Agglomerative Clustering [26], subsets were generated, and representative profiles were selected to inform the development of targeted follow-up questions.

Experienced practitioners initially developed 837 questions (650 unique) by analyzing 123 sub-sampled WHODAS 2.0 responses, focusing on academic participation, coping mechanisms, and strengths. Topic modeling was applied to generalize these questions, creating representative clusters aligned with WHODAS 2.0 domains. This led to a standardized set of three questions per WHODAS 2.0 item, addressing the impact of limitations, effective supports or coping strategies, and relevant examples from the student's lived experience, ensuring scalable and inclusive data collection for new students.

A chatbot named "AVA" was then developed to streamline WHODAS 2.0 data collection, dynamically selecting follow-up questions based on student responses. The survey structure was optimized using archival data to prioritize high-impact items, reducing time and improving completion rates. The responses were analyzed using item response theory (IRT) scoring to determine the degree of functional limitation, while optional sections captured quantitative and qualitative insights. The design prioritized accessibility, simplicity, and security, ensuring compatibility with assistive technologies and secure user authorization.

B. Accommodation Prediction

We investigated various machine learning techniques, including traditional and transformer-based models, as well as Large Language Models (LLMs) to predict academic accommodations from the collected student data.

1) Parameter Tuning Approach: Baseline experiments used Support Vector Machines (SVMs) [27] and Multi-Layer Perceptrons (MLPs) [28] on quantitative data for binary and multi-label classification. Hyper-parameter tuning was applied to SVMs (regularization, kernels) and MLPs (hidden layers, learning rates), with weighted loss functions addressing class imbalances. These provided foundational insights into traditional models.

Building on this, textual data was then incorporated using embeddings from pre-trained transformers like LongFormer [29] and BigBird [30], generating 768-dimensional embeddings for classification tasks. The experiments varied attention window sizes and sparse patterns to optimize feature extraction from long-context text.

Furthermore, fine-tuning techniques, including full fine-tuning and parameter efficient fine tuning (PEFT) methods like Low-Rank Adaptation (LoRA) [31], were employed to generate more suitable input vectors. Full fine-tuning adapted LongFormer and BigBird comprehensively, while PEFT reduced trainable parameters without compromising performance. This integrated approach enabled a robust evaluation of both traditional and advanced methods for predicting academic accommodations.

2) Generative Model Approach: A key feature of LLMs is in-context learning (ICL), which enables models to infer and perform tasks dynamically using information within the input prompt, eliminating the need for explicit training . Here, we consider GPT-4o [32], Mixtral [33] and LlaMA 3.1 [34], employing both zero-shot and few-shot learning paradigms as ICL approaches. Zero-shot learning utilized task guidelines and student profiles as input prompts, relying solely on the models' pre-trained knowledge for binary and multi-label classification tasks. Few-shot learning, applied exclusively to binary classification due to its complexity, explored two shot selection mechanisms: random selection and local-boundary selection. Random selection sampled positive and negative examples randomly, while local-boundary selection focused on challenging examples based on cosine similarity to the test profile, using embeddings from fine-tuned transformer models for distance calculations. This refined the model's ability to discern patterns in difficult scenarios.

To enhance the framework, practitioners developed Accommodation Descriptions (ADs), which provided structured mappings between functional limitations and accommodations. These domain-specific descriptions replaced shot-specific examples in zero-shot learning, reducing input complexity while maintaining relevance.

Additionally, Retrieval-Augmented Generation (RAG) [35] was employed to integrate external knowledge bases (KBs) derived from training datasets Detailed KBs were developed using OpenAI's o1-preview model [36]. These KBs were created by extracting instructions from student

profiles for various scenarios involving the recommendation and non-recommendation of accommodations. This process was applied to all student profiles in the training set for all accommodations. Additionally, these KBs, along with ADs, enriched the context provided to LLMs.

C. Explainability

While LLMs can provide reasoning for their decisions, traditional ML models (e.g., SVM, MLP) are often seen as black boxes due to their complex, non-linear structures. To interpret these models, we applied Integrated Gradients (IG) [37], a post-hoc explainability method. IG attributes model predictions to input features by integrating gradients along a path from a baseline to the actual input, satisfying sensitivity and implementation invariance axioms.

IV. Experiments

A. Data Collection and Annotation

To collect data, over 4,000 invitations were distributed, resulting in more than 250 students attempting the survey presented by AVA, with 170 students completing it successfully in almost 2 months. To address diverse objectives and enhance analytical outcomes, the dataset was structured into two versions. **Version-0** replicated the archival format, including only WHODAS 2.0 item IRT scores for domains and domain-specific questions, yielding 45 numerical features per profile without textual data. **Version-1** consists of all the textual information collected using AVA. It includes WHODAS 2.0 questions and its responses, follow-up questions and responses, additional questions about the experience of functional limitations, responses to questions on strengths and strategies helpful to overcome the experienced functional limitations, self-identified disabilities, and past or prospective academic accommodations.

The dataset was annotated by practitioners. Three experienced disability service practitioners reviewed each student profile in the dataset. The practitioners evaluated the nature and degree of functional limitations, compensatory strengths and strategies, disability self-identification where provided, and past supports. The annotation process involved highlighting significant forced-choice descriptive responses to each WHODAS 2.0 question (e.g., "Severe" or "Extreme") and keywords and phrases in the student responses to follow-up questions that were most relevant to the selection of academic accommodations and support services. As the final step, the practitioners selected a set of academic accommodations and support services and provided a written rationale for each selection. In cases of disagreement, the annotators engaged in discussion to reach a final consensus decision.

In order to have a fair comparison between the results of different approaches, the annotated dataset was further divided into training, test and validation as follows.

B. Data Splits

Given the multi-label nature of the annotated dataset, specific strategies were employed to ensure balanced representation:

- Profiles with selected accommodations having fewer than 10 positive samples per accommodation were manually distributed to ensure at least one sample for each accommodation in both validation and test sets.
- The remaining profiles were split using Iterative Stratification, a technique designed for multi-label datasets, to balance class representation across splits.

The final dataset comprised 114 profiles for training, 25 for validation, and 31 for testing. This approach addressed class imbalances and ensured robust model evaluation.

C. Classification with Parameter-Tuned Models

The first experiment aimed to establish a baseline using Version-0 data as input features, with 21 academic accommodations as target variables. SVM models were trained for binary classification tasks for each target variable, while MLP models were employed for both binary and multi-label classification tasks. Individual hyper-parameter tuning was conducted for each model using the Version-0 validation set. This baseline experiment provided foundational insights into the performance of traditional machine learning models on the dataset. After establishing the baseline model, text data from Version-1 was encoded using document embeddings generated by fine-tuned transformer models: LongFormer and BigBird. The remaining setup was similar to baseline experiments, repeated for embeddings from both Transformer models.

Additionally, LoRA was applied, with hyper-parameters $r \in \{2,4,8\}$ and $\alpha \in \{8,16,32\}$, and a dropout of 0.1 to prevent overfitting. This enabled efficient fine-tuning and optimized textual representations for academic accommodation recommendations.

D. Classification with Generative Models

LLMs namely GPT-4o and o1-preview from OpenAI [32], and LlaMA-3.1-405B from MetaAI [34] and Mixtral-8x7B-Instruct-v1 from MistralAI [33], were used with zero-shot learning technique for binary and multi-label classification. The text-based Version-1 data was used for input to all generative models. Few-shot learning experiments build upon binary zero-shot learning experiments. The number of shots was considered as a hyper-parameter, along with the shot selection method, and different LLM models (GPT-4o and LlaMA-3.1-405B). Since the context lengths kept increasing with the number of shots and each shot being a whole student profile description, longer input prompts can degrade the performance of the LLMs [38].

To tackle this issue, an approach was devised to use the knowledge either from or similar to these shots, directly replacing the shots in the input data, and hence this additional knowledge was created in the form of instructions by practitioners. The instructions represent the mapping of functional limitations and AAs, which guide the model in deciding whether to recommend an accommodation or not based on the types of functional limitations.

The initial approach to creating this additional knowledge was for practitioners to create fixed accommodation descriptions (ADs).

Finally, the Retrieval-Augmented Generation (RAG) approach was undertaken to create an external knowledge base for each accommodation consisting of several instructions on when to recommend or not recommend the accommodation. We do not focus on examining how individual components of the RAG model influence overall performance.

An additional step of retrieval of instructions was added to experiments similar to Zero-shot learning along with the inclusion of ADs in the context of each classification. The test profiles were split into chunks of different WHODAS 2.0 domain questions and answers and follow-up questions (FQs), if present in the profile. One instruction for each of these chunks was retrieved using a Sequence-to-Vector LLM NV-Embed-V2 [39]. The prompts employed in the generative model experiments under various settings are available upon request.

V. Results

A. Evaluation of Parameter-Tuning Approaches

TABLE I: Evaluation of Parameter-tuning based experiments with different tasks, input types and methodology. Student profiles from data Version-1 were used for inputs to all variations except Baseline (Version-0) experiment, where Version-0 data was used as inputs. *[For Model$_x$, x is (Task $\in$ {binary, multi-label} or Base_Model or Fine-Tuning method $\in$ {full, lora})]. *[BB = BigBird, LF = LongFormer, MLP = Multi-layer Perception, SVM = Support Vector Machine]. The best-performing results in each category are highlighted for each metric.

Evaluation of Classification Task for Parameter-Tuning based Approaches					
Variations	Model*	Accuracy	F1-Score	Precision	Recall
Baseline (Version-0)	***SVM***	**0.619**	**0.694**	0.690	**0.697**
	MLP_{binary}	0.664	0.674	**0.843**	0.561
	MLP_{multi}	0.622	0.614	0.834	0.486
Baseline (Text Embeddings)	SVM_{LF}	0.639	0.744	0.663	**0.846**
	$MLP_{(binary,\ BB)}$	0.694	0.747	0.766	0.730
	$\mathbf{MLP_{(multi,\ LF)}}$	**0.748**	**0.796**	**0.798**	0.794
Fine-Tuned Transformers	BB_{full}	0.725	0.739	**0.897**	0.628
	$\mathbf{LF_{full}}$	**0.747**	**0.772**	0.872	**0.692**
	BB_{lora}	0.716	0.728	0.895	0.613
	LF_{lora}	0.688	0.696	0.879	0.576

Table I reports the performance of multiple parameter-tuning strategies for classification tasks on the same test set. The evaluated approaches include classical machine learning methods (SVM and MLP) with either quantitative score inputs (Version-0) or transformer-derived text embeddings, as well as fine-tuned transformer models.

The results indicate that models using text embeddings consistently outperform those using purely numerical inputs. This improvement suggests that text-based representations provide richer semantic context, enabling more accurate classification decisions. For quantitative inputs, the SVM demonstrates relatively balanced performance. Given that quantitative data often capture only a limited aspect of student profiles (e.g., numeric scores or discrete features), SVM appears adequate for identifying class boundaries in lower-dimensional spaces. However, when using embeddings from pre-trained transformers, SVM exhibits a high recall (0.846) but only moderate precision (0.663). This pattern indicates the model may be over-predicting the positive class. The increased complexity of semantic representations can lead to a bias toward identifying more potential positives, thus boosting recall but reducing precision.

In the Version-0 setting, the binary MLP outperforms the multi-label variant (F1-Score = 0.674 vs. 0.614). This discrepancy may be attributable to the simpler decision boundary requirements in binary tasks, as well as the limited representational capacity of the quantitative data for multi-label predictions. By contrast, with text embeddings, the multi-label MLP achieves higher overall performance than the binary variant. This improvement likely reflects the richness of text embeddings, which capture fine-grained contextual and semantic features conducive to distinguishing among multiple labels simultaneously. Additionally, MLPs' non-linear architectures are well-suited for leveraging complex, high-dimensional representations, thereby improving classification metrics.

Despite the strong prior success of transformer-based models on a range of NLP tasks, the fully fine-tuned and LoRA fine-tuned transformers in this study do not surpass the MLP with text embeddings. Notably, the performance of the best fine-tuned transformer variant is marginally lower than the best MLP approach. A plausible reason for this observation is the relatively small size of the training dataset, which can limit the effectiveness of end-to-end fine-tuning. Transformer-based models, especially large ones like BigBird and LongFormer, typically require substantial amounts of data to optimize a large number of parameters effectively. In contrast, using pre-extracted embeddings reduces the parameter space for the downstream classifier (e.g., MLP), mitigating the risk of overfitting in data-constrained scenarios.

Across all fine-tuned transformer variants, precision tends to be higher than recall. This high precision indicates a strong capacity to correctly identify positive instances when they are predicted as such, yet the lower recall implies that many positive instances remain unidentified. This shortfall underscores the need for either (1) additional labelled data to bolster the transformer models' capacity to generalize or (2) regularization and hyperparameter adjustments that specifically target improvements in recall.

B. Evaluation of Generative Approaches

Table II presents the evaluation results of various experiments of generative approaches, on the same test data, using popular classification metrics. The results highlight the best-performing models within each approach according to F1-score. A key observation is that in zero-shot settings, binary classification consistently outperforms multi-label classification in terms of F1-score and overall accuracy. These discrepancies likely stem from the inherent complexity of

TABLE II: Evaluation of generative models based experiments with different tasks, methodology and prompt contents. Student profiles from data Version-1 were used for inputs to all variations. *[For Model$_x$, x is (Task $\in$ {binary, multi-label} or shot_selection $\in$ {random, local-boundary}, or number_of_shots $\in \{4, 8\}$, or RAG_domain_specific_context $\in$ {only instructions, instructions + ADs})]. *[*GPT* = OpenAI GPT-4o, *o1* = OpenAI o1-Preview, *LlaMA* = Meta Llama-3.1-405B-Instruct, *MIX* = Mistral.ai Mixtral-8x7B-Instruct-v1]. The best-performing results in each category are highlighted for each metric.

Evaluation of Classification Task for Generative Models based Approaches					
Variations	Model*	Accuracy	F1-Score	Precision	Recall
Zero-Shot	GPT_{binary}	0.737	0.799	0.76	0.841
	$LlaMA_{binary}$	0.768	0.819	0.793	0.846
	$\mathbf{MIX_{binary}}$	**0.768**	**0.834**	**0.750**	**0.938**
	$O1_{binary}$	0.753	0.809	0.775	0.846
	$\mathbf{GPT_{multi}}$	**0.679**	**0.697**	0.839	0.596
	$LlaMA_{multi}$	0.550	0.45	**0.923**	0.298
	MIX_{multi}	0.631	0.682	0.732	**0.638**
Few-Shot	$LlaMA_{(random,\ n=4)}$	**0.736**	**0.799**	0.756	**0.846**
	$GPT_{(random,\ n=4)}$	0.668	0.704	**0.786**	0.638
	$\mathbf{LlaMA_{(random,\ n=8)}}$	**0.759**	**0.817**	0.770	**0.871**
	$GPT_{(random,\ n=8)}$	0.679	0.711	**0.803**	0.638
	$\mathbf{LlaMA_{(local,\ n=4)}}$	**0.757**	**0.819**	0.761	**0.886**
	$GPT_{(local,\ n=4)}$	0.637	0.66	**0.787**	0.568
	$LlaMA_{(local,\ n=8)}$	**0.736**	**0.806**	0.739	**0.886**
	$GPT_{(local,\ n=8)}$	0.648	0.675	**0.788**	0.591
Zero-Shot (With ADs)	GPT_{binary}	0.805	0.851	0.808	0.898
	$\mathbf{LlaMA_{binary}}$	**0.816**	**0.860**	**0.810**	0.918
	MIX_{binary}	0.765	0.835	0.738	**0.963**
	$O1_{binary}$	0.796	0.847	0.790	0.913
	GPT_{multi}	0.691	0.700	0.88	0.581
	$LlaMA_{multi}$	0.582	0.513	**0.923**	0.355
	$\mathbf{MIX_{multi}}$	**0.708**	**0.765**	0.762	**0.769**
RAG	$\mathbf{GPT_{instructions}}$	**0.75**	**0.817**	**0.745**	0.906
	$LlaMA_{instructions}$	0.716	0.81	0.691	**0.978**
	$O1_{instructions}$	0.714	0.803	0.701	0.938
	$\mathbf{GPT_{(instructions\ +\ AD)}}$	**0.779**	**0.837**	**0.77**	0.916
	$LlaMA_{(instructions\ +\ AD)}$	0.714	0.809	0.69	0.978
	$O1_{(instructions\ +\ AD)}$	0.748	0.83	0.712	**0.995**

multi-label classification and from the varying degrees of alignment between the models' pre-training data and the specific multi-label disability-services domain.

Furthermore, the variability among different models in the zero-shot multi-label setting (*e.g.*, F1-score ranging from 0.450 for LlaMA to 0.697 for GPT) highlights the reliance on each model's unique pre-training corpus and training methodology. Such high variability suggests that zero-shot performance for multi-label tasks is less reliable, given that the models' decisions in this regime depend heavily on unknown or opaque features of their respective pre-training processes. As a result, purely zero-shot multi-label solutions may not be suitable for high-stakes domains such as disability accommodations, where consistency and explainability are paramount.

In few-shot experiments, both the number of in-context examples (shots) and the strategy for selecting these examples (random *vs.* local-boundary) influence model performance. For the LlaMA model with four-shot prompts, the local-boundary approach slightly outperforms random shot selection. This improvement suggests that presenting difficult, contextually relevant examples helps orient the model more effectively toward the classification boundaries. However, when the number of shots increases to eight, the differences between random and local-boundary selection can shift. This indicates that although local-boundary selection appears beneficial under certain conditions (shorter prompts, carefully chosen examples), it does not universally outperform random sampling, particularly when the model is given more contextual examples overall. Importantly, few-shot approaches in this domain still partially depend on the large language models' pre-training knowledge to interpret functional limitations and appropriate accommodations. These findings underscore the need for complementary domain-specific instructions or context to stabilize model predictions, especially for complex classification tasks involving disability accommodations.

Across multiple experimental configurations, the addition of Accommodation Descriptions (ADs) in the input context markedly improves classification performance and reduces variance. These results suggest that curated domain-relevant information, rather than relying solely on latent knowledge encoded during pre-training, enhances the model's ability to map student functional limitations to the correct accommodations. Moreover, adding ADs appears to partially mitigate the unpredictable reliance on implicit model knowledge. By providing explicit domain context, the prompts reduce the "guessing" behavior that can arise from purely zero-shot or limited few-shot prompts. This improvement is particularly important for real-world applications in disability services, where the cost of misclassification can be high and the justification for an assigned accommodation must be transparent.

In principle, RAG is expected to improve classification accuracy by supplying relevant external knowledge while keeping prompt size manageable. Indeed, Table II shows that for all models combined, RAG configurations (both "instructions only" and "instructions + AD") generally outperform corresponding few-shot approaches that rely on longer prompts. Nonetheless, an unexpected result is that the best zero-shot approach supplemented with ADs still outperforms RAG. Likely explanations for this observation are insufficient training data for the knowledge base and imprecise or under-optimized retrieval.

The knowledge base constructed for RAG may be too small or unrepresentative of the entire set of accommodations and student profiles. Some accommodations have as few as one training example, limiting the diversity and quantity of retrievable information. Similarly, the retrieval steps themselves leverage large language models without the fine-tuning on disability-services data. Thus, the retrieval may not be sufficiently targeted, potentially failing to return the most useful context. This issue is compounded by the highly specialized nature of academic accommodations and disability services, where domain expertise is crucial. Taken together, these findings suggest that while RAG approaches hold promise, especially when prompt length or complexity is a concern, careful construction of the knowledge base and the fine-tuning of retrieval steps are vital to achieving optimal performance in specialized domains.

C. Comparative summary

Overall, the generative approaches evaluated in this study surpass parameter-tuning methods. While parameter-tuning models exhibit high precision but low recall, generative models achieve higher recall with only a modest decrease in precision. This difference reveals a conservative tendency in parameter-tuning methods, leading to higher false negatives, compared to the over-recommendation bias of generative models. Among the generative strategies, RAG shows competitive results but demonstrates the largest recall–precision gap, indicating a pronounced inclination to over-recommend. In contrast, zero-shot configurations enhanced with domain knowledge (ADs) achieve both superior and more balanced performance compared to all other approaches and variations.

Given that unrecognized accommodations (false negatives) pose a more serious risk than unnecessary recommendations (false positives), the study's findings underscore the importance of prioritizing recall. Moreover, experiments with Integrated Gradients (IG) for parameter-tuning approaches highlight the complexity and expense of generating interpretable outputs. By comparison, LLMs can provide concise, context-specific rationales for their decisions (see Figure 1), thus offering enhanced transparency and practical utility for accommodation recommendations. Practitioners have noted that such explanations can be of value in standardizing the rationales for accommodation decisions and training junior professionals.

VI. Conclusion

This study addressed the critical challenge of providing timely and effective academic accommodations for students with disabilities in PSE by leveraging AI-driven methodologies and comprehensive data collection. The research demonstrated the potential of LLMs in enhancing inclusivity and accessibility, particularly in domain-specific tasks such as accommodation recommendations. While LLMs exhibited strong contextual understanding and classification capabilities, their performance varied significantly based on the availability of domain-specific instructions and context. Incorporating Accommodation Descriptions into the input data bridged this gap, improving the relevance and reliability of recommendations.

Additionally, we explored various explainability methods across different approaches. The self-explainability of LLMs allows both students and practitioners to understand the rationale behind accommodation recommendations, fostering transparency and trust, and can contribute to more standardized and reliable practices. This interpretability significantly enhances the utility and adoption of LLMs in accommodation services in higher education.

However, the study has notable limitations. The reliance on a small dataset of 170 samples restricts the generalizability of the results and exacerbates challenges related to data imbalance. Also, under-representation of certain accommodation categories led to skewed predictions, underscoring the need for more comprehensive datasets.

Our study also emphasized the subjective nature of 'ground truths' in AI research. While traditionally the notion of 'ground truths' represents an absolute reference for AI models, we observed that the determination of ground truths for accommodation decisions was more fluid. Each practitioner has a unique perspective on when an accommodation may or may not be appropriate, and while there is a degree of consensus, there are multiple factors that can influence such decisions and vary the outcome between different practitioners or institutional settings. This can present a challenge for AI research in this area and deserves further investigation.

ACKNOWLEDGMENT

This work was supported by the New Frontiers in Research Fund Grant: NFRFE-2021-00855. Additionally, the Digital Alliance of Canada is acknowledged for its support of this research.

References

[1] S. Morris, G. Fawcett, L. Brisebois, and J. Hughes, "A demographic, employment and income profile of canadians with disabilities aged 15 years and over, 2017," *Canadian Survey on Disability Reports*, 2018. [Online]. Available: https://www150.statcan.gc.ca/n1/en/catalogue/89-654-X2018002

[2] P. Susan, "Inclusive education: Achieving education for all by including those with disabilities and special education needs," The World Bank, Tech. Rep., 2003.

[3] A. I. Amjad and S. Irshad, "Role of self-advocacy, academic accommodation, and attitude in career readiness among students with learning disabilities: a social cognitive perspective," *BMC psychology*, 2026.

[4] O. H. R. Commission, "Policy on accessible education for students with disabilities," 2018, accessed: 2023-10-19. [Online]. Available: https://www.ohrc.on.ca/en/policy-accessible-education-students-disabilities

[5] C. Dumitru, G. Muttashar Abdulsahib, O. Ibrahim Khalaf, and A. Bennour, "Integrating artificial intelligence in supporting students with disabilities in higher education: An integrative review," *Technology and Disability*, vol. 38, no. 1, pp. 3–24, 2026.

[6] World Health Organization, *Measuring Health and Disability: Manual for WHO Disability Assessment Schedule WHODAS 2.0*. World Health Organization, 2010.

[7] U. Carleton, "Middle - years students survey master report," June 2023, accessed: 2023-10-19. [Online]. Available: https://cusc-ccreu.ca/?page_id=32&lang=en

[8] S. Lanthier, R. Tishcoff, S. Gordon, and J. Colyar, "Accessibility services at ontario colleges and universities: Trends, challenges and recommendations for government funding strategies," *Higher Education Quality Council of Ontario. Retrieved from: https://heqco. ca/wp-content/uploads/2023/11/Accessibility-Services-at-Ontario-Collegesand-Universities-FINAL-English. pdf*, 2023.

[9] L. Chengying and E. M. Sagubo, "Inclusive student management practices: strategies to support diverse and underrepresented student groups," *International Journal of Science and Engineering Applications*, vol. 14, no. 03, pp. 15–20, 2025.

[10] R. Sarkar and R. Madhu, "Ensuring equity and access for students with disabilities," 2026.

[11] M.-L. Lee, G.-H. Lin, Y.-C. Wang, S.-C. Lee, and C.-L. Hsieh, "Machine learning-based world health organization disability assessment schedule for persons with parkinson's disease," *Parkinsonism & Related Disorders*, vol. 133, p. 107316, 2025.

[12] A. V. Ravindran, A. V. J, and M. P, "Prediction of learning disability using machine learning," *International Journal For Research in Applied Science and Engineering Technology*, 2023.

[13] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, R. Meersman, Z. Tari, and D. C. Schmidt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 986–996.

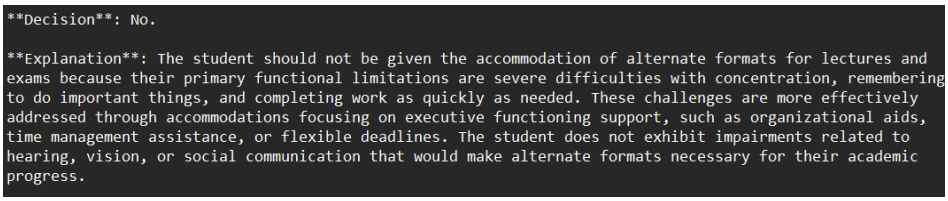

Fig. 1: Sample output from LLMs with decision on recommendation of the Academic Accommodation and explanation for the decision.

[14] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, 1995/4.

[15] E. De Brouwer and et al., "Machine-learning-based prediction of disability progression in multiple sclerosis: An observational, international, multi-center study," *PLOS Digital Health*, vol. 3, no. 7, pp. 1–25, 07 2024. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000533

[16] E. Kalita, H. El Aouifi, A. Kukkar, S. Hussain, T. Ali, and S. Gaftandzhieva, "Lstm-shap based academic performance prediction for disabled learners in virtual learning environments: a statistical analysis approach," *Social Network Analysis and Mining*, vol. 15, no. 1, p. 65, 2025.

[17] Y. Lu, K. Sato, M. Nagai, H. Miyatake, K. Kondo, and N. Kondo, "Machine learning-based prediction of functional disability: a cohort study of japanese older adults in 2013-2019," *J Gen Intern Med*, 2023.

[18] Y. Han and S. Wang, "Disability risk prediction model based on machine learning among chinese healthy older adults: results from the china health and retirement longitudinal study," *Sec. Aging and Public Health*, 2023.

[19] M. F. Almufareh, S. Tehsin, M. Humayun, and et al., "An artificial intelligence perspective and framework," *Journal of Disability Research*, 2023.

[20] R. Scendoni, L. Tomassini, M. Cingolani, A. Perali, S. Pilati, and P. Fedeli, "Artificial intelligence in evaluation of permanent impairment: New operational frontiers," *Healthcare*, vol. 11, no. 14, 2023.

[21] M. Qazvini, "Forecasting the levels of disability in the older population of england: Application of neural nets," 05 2023.

[22] H.-T. Cheng, L. Koc, and et al., "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, ser. DLRS 2016. New York, NY, USA: Association for Computing Machinery, 2016, p. 7–10.

[23] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," 12 2020.

[24] S. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, May 2021.

[25] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[26] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," 2011. [Online]. Available: https://arxiv.org/abs/1109.2378

[27] C. Cortes and V. Vladimir, "Support-vector networks," *Machine Learning*, 1995.

[28] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65 6, pp. 386–408, 1958. [Online]. Available: https://api.semanticscholar.org/CorpusID:12781225

[29] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[30] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2021. [Online]. Available: https://arxiv.org/abs/2007.14062

[31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[32] OpenAI, J. Achiam, and et. al., "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[33] A. Q. Jiang and et. al., "Mixtral of experts," 2024. [Online]. Available: https://arxiv.org/abs/2401.04088

[34] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[35] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[36] E. Latif, Y. Zhou, S. Guo, Y. Gao, L. Shi, M. Nayaaba, G. Lee, L. Zhang, A. Bewersdorff, L. Fang *et al.*, "A systematic assessment of openai o1-preview for higher order thinking in education," *arXiv preprint arXiv:2410.21287*, 2024.

[37] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[38] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," 2023. [Online]. Available: https://arxiv.org/abs/2307.03172

[39] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "Nv-embed: Improved techniques for training llms as generalist embedding models," 2024. [Online]. Available: https://arxiv.org/abs/2405.17428