Longitudinal Surveys are Texts: LLM-enhanced Analysis of School Attendance in New Zealand

Tingrui Qiao[1] ✉, Caroline Walker[1], Chris Cunningham[2], Adam Jang-Jones[3], Susan Morton[4], Kane Meissel[1], and Yun Sing Koh[1]

[1] University of Auckland
{ricky.qiao,caroline.walker,k.meissel,y.koh}@auckland.ac.nz
[2] Massey University C.W.Cunningham@massey.ac.nz
[3] New Zealand Ministry of Education Adam.JangJones@education.govt.nz
[4] University of Technology Sydney Susan.Morton@uts.edu.au

Abstract. School attendance is an important factor in educational success and plays a key role in shaping students' academic and social development. Longitudinal surveys provide valuable insights into factors affecting attendance patterns, yet analysing such data presents unique challenges. First, the variation in survey questions across data collection waves complicates the application of standard temporal modelling techniques that assume consistent features over time. Second, conventional methods often one-hot encode survey responses, stripping away contextual meaning within questions and responses. Lastly, open-ended responses are typically omitted, leading to a loss of valuable qualitative insights. To address these challenges, we propose Survey-as-Text Modelling (STM), which represents multi-wave survey questionnaires as coherent textual sequences. By maintaining the textual format, STM allows similar questions across different years to be compared directly rather than existing as independent features. STM also retains the meaning within question-response pairs, preventing loss of information from one-hot encoding and enabling the incorporation of open-ended responses. We apply STM to survey data from *Growing Up in New Zealand* and link it to official attendance records from the *New Zealand Ministry of Education*. We leverage large language models (LLMs) to predict future school attendance from text-based surveys, outperforming existing temporal methods. Beyond predictive accuracy, we propose gradient-guided counterfactual analysis to identify key survey questions influencing the model's decision-making. Our findings highlight the potential of LLMs for survey analysis and provide data-driven insights that can inform policy and intervention strategies.

Keywords: Longitudinal Survey · Large Language Models · School Attendance.

1 Introduction

School attendance is an important determinant of academic achievement, social development, and long-term well-being [20]. Attendance is shaped by a wide

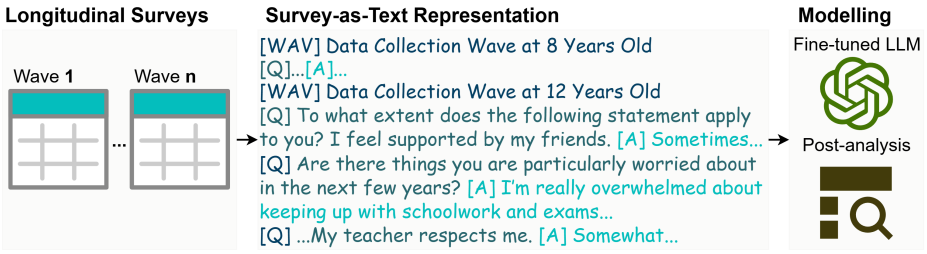

Fig. 1. In Survey-as-Text Modelling, survey data from multiple data collection waves is retained as natural text representation where special delimiters are used to separate survey waves, questions, and responses. The text representation is processed by a fine-tuned large language model for attendance prediction.

range of factors, including mental health [10], parental support [31], socioeconomic status [21], school experiences [22], and external disruptions such as the COVID-19 pandemic [29]. Data from *Growing Up in New Zealand* (GUiNZ) [34], the country's largest ongoing longitudinal study on child well-being, which tracks over 6,000 children from before they were born, provides an opportunity to examine these influences. GUiNZ collects data through repeated data collection waves, where participants provide survey responses at different ages, capturing evolving socioecological factors over time. By linking this data with official attendance records from the *New Zealand Ministry of Education*, we are able to analyse how multiple factors interact over time to shape attendance patterns.

Analysing longitudinal survey data presents challenges to existing temporal modelling methods. First, the variation in survey questions across waves disrupts the assumption of the same features being collected over time at regular intervals. This poses difficulties for existing methods such as Recurrent Neural Networks (RNNs) [42] and Transformers [56,53], which often assume structured input sequences with consistent features across observations or time steps. However, in longitudinal surveys, questions evolve due to shifting research priorities and external factors; for instance, earlier waves from GUiNZ include questions on experience starting school, while later waves focus on the impact of COVID-19 on schooling. Additionally, data collection occurs at irregular intervals, with some variables collected at certain waves but omitted in others; for example, the New Zealand index of socioeconomic deprivation [2] is recorded at ages six, eight, and twelve, while questions about material hardship are asked at ages six and twelve but not at age eight, further complicating feature alignment. Second, conventional methods often one-hot encode survey responses, leading to a loss of contextual meaning within questions and responses. For example, responses to "children feel they belong to school when they start school" and "children feel they are connected to school during lockdown" may be treated as entirely separate variables, losing their relationship in meaning, or collapsed into a single category, ignoring the differences in context. One-hot encoding forces

models to rely on statistical associations rather than textual meaning, limiting their ability to capture nuanced relationships in survey data. Lastly, conventional approaches struggle to incorporate open-ended responses, which often contain valuable qualitative insights. Since these responses cannot be easily one-hot encoded, they are frequently excluded from analysis. For example, responses to "What is the most worrying thing during the COVID-19 lockdown?" can reveal key concerns affecting school attendance, but existing methods lack a structured way to integrate this information. These limitations highlight the need for a more flexible approach that can preserve contextual meaning, handle irregular survey structures, and incorporate qualitative responses into predictive modelling.

To address these limitations, we propose Survey-as-Text Modelling (STM), which represents longitudinal survey data in its natural textual format rather than converting it into structured tabular variables. By leveraging large language models (LLMs), which have been pretrained on vast amounts of text data, STM can model survey data while preserving its original structure and meaning. STM mitigates the challenge of irregular survey structures caused by evolving questions across waves. Instead of treating missing or modified variables as a structural problem requiring imputation or manual feature alignment, STM processes survey responses as continuous textual sequences, allowing similar questions asked at different waves to be compared within context rather than treated as separate features. This enables the model to generalise across changes in wording or focus of questions, ensuring better alignment of information across survey waves. The time information of each wave can also be described within the text representation, indicating the time duration between each wave. Furthermore, STM retains the contextual meaning of survey questions and responses and leverages LLMs' ability to understand texts to maintain the conceptual connection between similar but distinct questions (e.g., school engagement at different time points) and to recognise differences in phrasing and context. This allows the model to capture deeper relationships between responses rather than treating them as isolated categorical variables. Additionally, STM directly incorporates open-ended responses by integrating them naturally within the textual representation, allowing LLMs to extract meaningful insights alongside structured survey responses. This provides a richer understanding of subjective factors influencing attendance, such as personal experiences, concerns, and motivations, that would otherwise be omitted from quantitative models.

Beyond predictive modelling, interpretability is essential for deriving insights from survey data. Therefore, we propose gradient-guided counterfactual analysis to identify the most influential survey items in attendance predictions. By aggregating gradients at the question level, we highlight which questions and responses contribute most to the model's decision. However, gradient-based attribution alone provides only ranked importance of survey items and does not determine which specific responses can consistently influence model decisions. To address this, we assess model sensitivity by iteratively swapping responses with values sampled from participants in the opposite attendance category, following the ranked importance of survey items based on question-level gradients.

We identify the minimal number of swaps required to flip a participant's classification and analyse which questions appear most frequently in these minimal swaps, highlighting the key factors influencing attendance predictions. Our main contributions are as follows:

1. We propose **Survey-as-Text Modelling (STM)**, which preserves the textual format of surveys and leverages LLMs to predict future school attendance. The text representation allows STM to address challenges in feature alignment, evolving contexts, and open-ended responses.
2. To improve interpretability, we introduce **gradient-guided counterfactual analysis** to identify influential survey items and evaluate how response variations impact model decision-making.
3. By linking *Growing Up in New Zealand* survey data with official attendance records from the *Ministry of Education*, our approach provides insights into potential factors influencing school attendance, supporting policymakers in data-driven decision-making and targeted interventions.

2 Related Work

Longitudinal Analysis. Longitudinal studies provide valuable insights by tracking individuals or groups over time, enabling researchers to identify temporal patterns, developmental changes, and underlying trends [49]. Conventional statistical methods, such as latent growth models and autoregressive approaches, rely on strong assumptions, including lack of multicollinearity, specific data distributions, and homoscedasticity [9]. These methods struggle with high-dimensional data, limiting flexibility in complex real-world applications [41]. Machine learning and deep learning offer a more adaptable alternative by capturing non-linear and higher-order relationships. Jin et al. [18] used Long Short-Term Memory to predict malnutrition from longitudinal patient records. Adler et al. [1] applied gradient boosting to track mental health symptoms using longitudinal mobile sensing data. Nitski et al. [36] explored Transformers, Temporal Convolutional Networks, and Recurrent Neural Networks for long-term mortality prediction in liver transplant recipients. These methods rely on fixed feature sets across time, making them less suited to longitudinal surveys where questions evolve, responses carry contextual meaning, and open-ended data remains underutilised. Our approach treats survey responses as natural text, enabling LLMs to address these challenges within a unified framework.

LLMs for Time Series. LLMs have been increasingly explored for time series modelling, demonstrating their ability to capture complex temporal dependencies. Existing approaches aim to enhance LLMs' ability to process numerical time series directly or align time series data with LLMs' natural language capabilities. Xue et al. [51] encode numerical inputs and outputs within prompts to adapt LLMs for time series tasks, while Zhou et al. fine-tune LLM input and output layers for time series modelling. Cao et al. [5] decompose trend, seasonal, and residual components within prompts to improve distribution adaptation. Other methods attempt to bridge time series with language processing: Jin et

al. [19] align time series patches with text modality and supplement inputs with textual dataset descriptions, Sun et al. [44] map time series embeddings to LLM token spaces, and Pan et al. [37] employ semantic-informed prompt learning for cross-modal alignment. While these methods adapt LLMs for structured time series data, they primarily treat time series as numerical sequences and lack direct contextual information. In contrast, our approach models longitudinal surveys as text, leveraging LLMs' language understanding to handle evolving questions, irregular structures, and open-ended responses, making it fundamentally different from numerical time series modelling.

LLMs Interpretability. Interpreting the predictions of LLMs has been a growing area of research, with various methods developed to explain model decisions. Local surrogate models [12] and local interpretable models [11] approximate LLM predictions using simpler models, but they rely on sampling-based approximations and do not directly reveal which variables drive predictions in structured survey responses. Concept bottleneck models [45] enforce interpretability by mapping predictions to predefined concepts; however, in our setting, survey questions already serve as well-defined interpretable variables, making additional concepts unnecessary. Self-explanation techniques [16,39] generate textual justifications for LLM outputs but can suffer from hallucination and may not consistently align with the true decision-making process. Attribution-based methods such as Shapley values [33] estimate the contribution of individual input components, while gradient-based attribution [43] analyses the sensitivity of predictions to input perturbations. However, these techniques operate at the token level, making it difficult to aggregate influence at the question-response level, which is essential for understanding survey-based predictions. Our approach, gradient-guided counterfactual analysis, first aggregates gradients at the question level to identify influential survey items, then iteratively swaps responses and observes prediction changes to reveal their impact on model decisions.

3 Survey-as-Text Modelling

We introduce Survey-as-Text Modelling (STM), a framework that leverages LLMs to process survey responses in their natural textual form. Section 3.1 details STM's text-based modelling and fine-tuning for classification, while Section 3.2 introduces gradient-guided counterfactual analysis to identify influential survey items and assess response swaps' impact on predictions.

3.1 Modelling Survey with LLMs

Problem Definition. Given a longitudinal survey collected over T waves, each wave is represented as $X^t = \{(q_i^t, x_i^t)\}_{i=1}^{D^t}$, where each (q_i^t, x_i^t) denotes a survey question q_i^t and its corresponding response x_i^t. The number of questions D^t varies across waves, meaning that certain questions may be missing in some waves. The objective is to train a classification model $\mathcal{C}$ with parameters $\Theta_{\mathcal{C}}$ to predict a categorical target variable $\hat{Y}$ using responses from past survey waves as input:

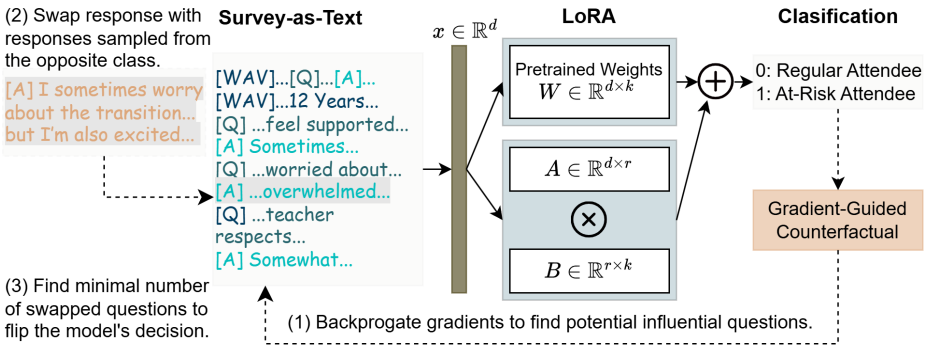

Fig. 2. Survey-as-Text Modelling processes longitudinal surveys as text. A Low-Rank Adaptation (LoRA) fine-tuned LLM is used for classification, where pretrained weights $W \in \mathbb{R}^{d\times k}$ remain frozen, and low-rank adaptation matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ are learned during finetuning. Here, d represents the input embedding dimension, k is the output dimension of the layer, and r is the rank of LoRA. Gradient-based attribution identifies potential influential questions, and counterfactual swaps reveal key variables that impact classification.

$\hat{Y} = \mathcal{C}(\{X^1, X^2, \ldots, X^{T-1}\}; \Theta_{\mathcal{C}})$. The classification output $\hat{Y}$ is a discrete label corresponding to predefined categories. Unlike time-series classification, where the input consists of structured numerical sequences $X \in \mathbb{R}^{D\times T}$ with a fixed set of D features across all time steps, our formulation allows the number of features to vary per wave, meaning $D^t \neq D^{t'}$ for some $t \neq t'$, reflecting the evolving nature of survey design.

Text Representation. We represent each wave X^t as a structured text sequence. As shown in Fig. 1, each wave begins with a textual description that explicitly marks the data collection period, such as *"Data Collection Wave at 6 Years Old (2015)"*, providing temporal context and allowing the model to distinguish between different survey waves. Each question-response pair (q_i^t, x_i^t) is structured in its natural form and concatenated sequentially within each wave. We introduce a set of special delimiter tokens: `[WAV]` is inserted before each wave description to explicitly mark temporal boundaries, `[Q]` is placed before each question q_i^t, and `[A]` precedes its corresponding response x_i^t. This structured encoding ensures that the model preserves semantic relationships between questions and responses while maintaining the sequential flow of information across survey waves. Missing responses are explicitly represented using the placeholder text *"missing"*, allowing the model to recognize patterns of missingness without relying on statistical imputation. This structured text representation allows the model to leverage its pretraining on natural language, effectively handling survey evolution across waves while preserving semantic and temporal coherence.

LLMs Finetuning. As shown in Fig. 2, to model longitudinal survey data in its natural textual format, we fine-tune a pretrained LLM using Low-Rank Adap-

tation (LoRA) [15], a parameter-efficient tuning method that introduces trainable low-rank update matrices while freezing the original model weights. This approach allows adaptation to survey-based classification tasks with reduced computational overhead. During fine-tuning, we appended a classification head to the LLM to predict attendance categories based on past survey responses. The model is optimised using cross-entropy loss.

3.2 Gradient-guided Counterfactual Analysis

Gradient-Based Attribution. To identify the most influential survey questions and responses in determining the model's decision, we compute the gradient of the model's output probability with respect to the input tokens. Given a trained classifier $\mathcal{C}$ with parameters Θ_C, the predicted probability is denoted as $P(\hat{Y} \mid X; \Theta_C)$. During inference, we first perform a forward pass through the model, encoding the survey input $X = \{(q_i^t, x_i^t)\}_{i=1}^{D^t}, t \in 1, \ldots, T$, which consists of concatenated question-response pairs formatted in text. The final classification probability $P(\hat{Y} \mid X; \Theta_C)$ is obtained from the softmax layer over the logits. To determine the impact of each response x_i^t on the prediction, we compute the gradient of $P(\hat{Y} \mid X; \Theta_C)$ with respect to the input embeddings $\nabla_{E(x_i^t)} P(\hat{Y} \mid X; \Theta_C)$, where $E(x_i^t)$ is the embedding representation of the response x_i^t after tokenization. Since the gradient is computed at the token level, we must aggregate it to obtain an importance score for each question-response pair. We achieve this by leveraging our special delimiter tokens, which explicitly separate each survey item in the text format. Let x_i^t denote the response to the i-th question q_i^t in wave t, and let its corresponding tokenized representation be $\{v_k^i\}_{k=1}^{T_i}$, where v_k^i is the k-th token of x_i^t, and T_i is the total number of tokens in the response after tokenization. To determine the contribution of each question-response pair to the classification decision, we compute a question-level importance score $S(x_i^t)$ by aggregating the gradient magnitudes of all tokens within x_i^t:

$$S(x_i^t) = \frac{1}{T_i} \sum_{k=1}^{T_i} \left\| \nabla_{E(v_k^i)} P(\hat{Y} \mid X; \Theta_C) \right\|_1 .$$

Here, $\nabla_{E(v_k^i)} P(\hat{Y} \mid X; \Theta_C)$ represents the gradient of the predicted class probability with respect to the embedding of a token v_k^i, and $\| \cdot \|_1$ denotes the $L1$ norm, capturing the absolute contribution of each token. Normalising by T_i ensures that responses of different token lengths do not disproportionately influence ranking, allowing fair comparison across survey items. To obtain an overall ranking of influential survey items, we further aggregate $S(x_i^t)$ across all participants, computing the mean importance of each question-response pair across the dataset. This allows us to determine which survey items contribute most to attendance classification, guiding interpretability and further analysis.

Minimal Response Swaps. While gradient-based attribution identifies which survey responses influence predictions, it does not directly determine their effect

on classification outcomes. To address this, we conduct counterfactual evaluation through minimal response swaps, measuring how sensitive predictions are to changes in survey responses. For each participant classified as an at-risk attendee, we iteratively replace responses with alternative responses observed in participants classified as regular attendees. The swaps are performed in order of importance, starting with the question-response pair with the highest gradient-based attribution score. Given a participant's survey representation X, we define a modified version X' in which responses are systematically substituted. At each step, the most influential response x_i^t is replaced with an alternative response $\tilde{x}_i^t$ drawn from a pool of responses observed in the regular attendee group. The classification model $\mathcal{C}(X';\Theta_C)$ is re-evaluated after each swap, and the process continues until the prediction flips to a regular attendee: $\mathcal{C}(X;\Theta_C) \neq \mathcal{C}(X';\Theta_C)$. The same process is repeated for participants classified as regular attendees. The number of swapped questions required to induce this change, denoted as the minimal swap count, provides a measure of how easily a classification outcome can be altered for each participant. This counterfactual evaluation complements gradient-based attribution by refining the identification of survey responses with the greatest impact on classification outcomes. By analysing the distribution of minimal swap counts across the dataset, we gain further insight into the relative importance of different survey items in shaping attendance classification.

Aggregating Insights Across Participants. To derive population-level insights, we measure the importance of each question by counting how often it appears in the minimal swap sets across all participants. A question appearing frequently across many minimal swap sets suggests it plays a key role in distinguishing attendance categories, whereas less frequent occurrences indicate factors that are influential only for specific subgroups. This approach provides a measure of which survey items most consistently impact classification outcomes.

4 Experiments

We evaluate Survey-as-Text Modelling (STM) for school attendance classification using longitudinal survey data from Growing Up in New Zealand (GUiNZ), linked with official attendance records. We detail experimental settings in Section 4.1. To benchmark STM, we compare it against machine learning, recurrent, convolutional, transformer-based, and LLM-based models in Section 4.2. Beyond predictive performance, we introduce gradient-guided counterfactual analysis to interpret STM's decision-making by identifying influential survey items that can flip the model's prediction. Additionally, we analyse the effectiveness of parameter-efficient fine-tuning techniques, compare various missing data imputation strategies, and evaluate the fairness of classification results among ethnicities in Section 4.3.

4.1 Experimental Settings

Dataset. We use longitudinal survey data from GUiNZ [34], linked with official school attendance records from the Ministry of Education (MoE). Our study

focuses on three survey waves collected at ages six, eight, and twelve, capturing key socioecological factors influencing school attendance. Attendance rates are derived from MoE's administrative records, calculated as the percentage of total recorded minutes present rather than the Ministry's half-day classification system. This granular, minute-based approach results in a slightly lower attendance percentage than MoE's official business rules. We categorise students into two groups: regular attendees, with attendance above 90%, and at-risk attendees, which includes all students below this threshold. The at-risk attendee category encompasses students with varying levels of absenteeism, including irregular, moderate, and chronic absenteeism. After filtering for students who participated in all three waves and removing those with missing attendance records, our final dataset consists of 3,844 participants, comprising 3,077 regular attendees and 767 at-risk attendees. We include variables related to mental health, socioeconomic status, parental support, school experiences, and COVID-19 disruptions, including open-ended questions such as asking what the biggest worry is for the future. The selected variables have an average of 3% missing rate.

Baseline Methods. We evaluate our approach against five categories of baseline models: machine learning models, Recurrent Neural Networks (RNNs), convolutional models, transformer-based models, and LLM-based approaches. For Machine learning models, we include XGBoost [6], and Random Forest (RF) [4]. We exclude logistic regression due to its reliance on manual feature engineering and the high dimensionality introduced by interaction terms. Similarly, traditional longitudinal methods, such as mixed-effects models and latent growth models, are not included because they are primarily designed for modelling individual trajectories or estimating population-level trends rather than performing multivariate temporal classification. RNN-based models, including RNN [32], Long Short-Term Memory (LSTM) [13], and Gated Recurrent Unit (GRU) [8], capture sequential dependencies in structured time series data. Convolutional models, such as Temporal Convolutional Networks (TCN) [48], use hierarchical convolutions to model temporal patterns. Transformer-based models, including Transformer [47], Autoformer [50], Crossformer [55], FEDformer [57], Informer [56], iTransformer [26], Nonformer [27], and PatchTST [35], leverage self-attention mechanisms to enhance long-range dependency modeling. Finally, LLM-based models, such as TEST [44], S^2IP-LLM [37], FPT [58], and Time-LLM [19], adapt pre-trained language models for time series tasks through reprogramming, embedding alignment, or prompt-based learning.

Evaluation. We evaluate model performance using standard classification metrics, including accuracy, precision, recall, F1 score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). We apply stratified splitting with 60% for training, 20% for validation, and 20% for testing across attendance categories.

Data Preprocessing. For STM, we describe each wave in text and use special delimiters to separate questions and responses, maintaining the natural structure of survey data. In contrast, for other methods, we transform the data into a tabular format by merging similar questions across waves into the same vari-

Table 1. Performance comparison of baseline methods for attendance prediction.

Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
RF [4]	73.5 ± .3	69.0 ± .2	70.4 ± .4	69.7 ± .4	69.8 ± .3	68.6 ± .4
XGBoost [6]	75.2 ± .2	71.1 ± .1	71.4 ± .4	71.2 ± .4	72.5 ± .2	70.4 ± .3
RNN [32]	78.0 ± .4	75.2 ± .3	73.6 ± .2	74.4 ± .3	75.2 ± .4	72.6 ± .3
LSTM [13]	80.3 ± .4	76.9 ± .4	76.5 ± .2	76.7 ± .3	77.3 ± .4	75.1 ± .3
GRU [8]	79.6 ± .3	75.7 ± .2	76.0 ± .5	75.8 ± .4	75.9 ± .4	74.9 ± .5
TCN [48]	81.2 ± .4	78.1 ± .4	78.0 ± .2	78.0 ± .2	78.0 ± .2	76.2 ± .1
Transformer [47]	80.1 ± .3	76.4 ± .5	76.2 ± .2	76.3 ± .3	77.4 ± .3	76.0 ± .2
Informer [56]	83.5 ± .2	81.0 ± .5	79.6 ± .4	79.6 ± .1	79.7 ± .1	79.3 ± .2
Autoformer [50]	84.0 ± .1	80.9 ± .2	79.8 ± .1	79.8 ± .2	81.3 ± .4	79.3 ± .3
FEDformer [57]	82.9 ± .4	79.7 ± .1	79.6 ± .5	79.6 ± .1	80.2 ± .2	77.6 ± .5
Nonformer [27]	83.8 ± .3	80.4 ± .3	79.3 ± .3	79.8 ± .2	80.1 ± .2	78.8 ± .4
PatchTST [35]	85.2 ± .2	81.0 ± .2	80.8 ± .2	80.9 ± .2	82.2 ± .2	79.3 ± .1
Crossformer [55]	84.5 ± .3	81.2 ± .4	81.1 ± .1	81.1 ± .3	81.6 ± .2	80.7 ± .2
iTransformer [26]	83.3 ± .3	78.9 ± .5	79.5 ± .3	79.2 ± .5	80.6 ± .2	78.2 ± .3
FPT [58]	85.8 ± .1	81.8 ± .4	81.2 ± .1	81.5 ± .4	83.2 ± .4	80.5 ± .5
TEST [44]	86.4 ± .1	83.3 ± .3	82.4 ± .4	82.8 ± .2	82.4 ± .3	81.0 ± .4
Time-LLM [19]	87.2 ± .4	82.7 ± .1	82.7 ± .3	82.7 ± .4	85.0 ± .4	81.5 ± .3
S^2IP-LLM [37]	86.1 ± .2	84.3 ± .3	82.7 ± .3	83.5 ± .2	84.7 ± .4	82.4 ± .2
STM (Ours)	**92.0 ± .1**	**89.2 ± .2**	**90.9 ± .4**	**90.0 ± .4**	**89.7 ± .2**	**88.3 ± .1**

able and interpolating variables that were not collected in certain waves from past waves to ensure alignment across time points. For handling missing data, STM preserves missing responses as explicit text "missing" to allow the model to learn patterns around missingness, while for tabular models, we apply KNN imputation [25] to estimate missing values based on similar participants.

Experimental Setup. Given the sensitive nature of the survey data, we employ a locally hosted model to ensure data privacy and compliance with ethical guidelines. We fine-tune LLaMA-3.1-8B using LoRA with a classification head for school attendance prediction. The same backbone model is used for all LLM-based time series methods to ensure a fair comparison. The model is trained for 10 epochs with a batch size of 16 and a learning rate of 1×10^{-4}. We optimise the model using the AdamW optimiser with weight decay regularisation to prevent overfitting. For evaluation, we perform 30 independent runs with different random seeds (1-30) and report the mean and standard deviation of classification metrics. Statistical significance is assessed using the Wilcoxon signed-rank test, and the best-performing results with statistical significance are highlighted in bold. All experiments are conducted on NVIDIA A100 GPUs.

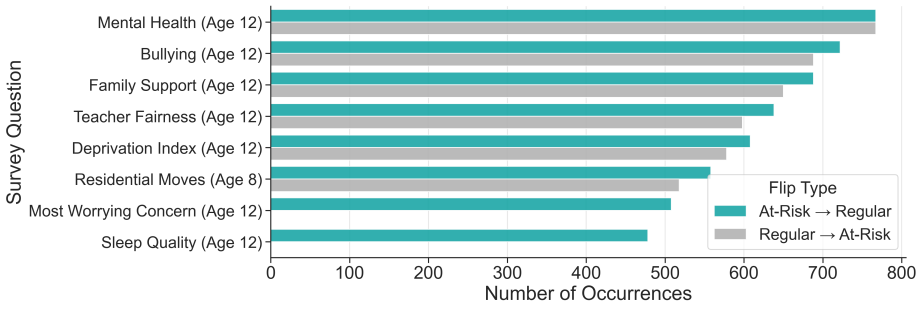

Fig. 3. Survey items most frequently appear in minimal swaps when flipping classification labels. Minimal swaps refer to the smallest number of response changes needed to flip a model's prediction between at-risk and regular attendees.

4.2 Main Results

Attendance Level Classification. Table 1 presents the performance comparison of STM against baseline methods. STM outperforms all baselines across all evaluation metrics. These results demonstrate the effectiveness of STM in capturing complex relationships in longitudinal survey data. Unlike baseline models that rely on fixed tabular structures or numerical encodings, STM attempts to preserve the contextual meaning of survey responses and adapts to evolving question formats. This flexibility allows STM to outperform existing temporal modelling methods, which are not designed to process survey data in their natural textual form.

Influential Survey Questions. To examine the model's decision-making process, we analyse the most influential survey questions identified through gradient-guided counterfactual analysis. Across all participants, the number of minimal swaps required to flip an at-risk attendee to a regular attendee is, on average, four swaps, with a maximum of eight. Conversely, flipping a regular attendee to an at-risk attendee requires a mean of three swaps and a maximum of six. Figure 3 presents the frequency of survey items appearing in minimal swaps when flipping classifications in either direction. The model is most sensitive to mental health, followed by bullying experiences, family support, and teacher fairness, all from Age 12. The deprivation index also plays a significant role, along with residential mobility from Wave Eight, suggesting that past relocations still contribute to classification shifts. Open-ended concerns about the future and sleep quality only appear in the maximal swaps for at-risk attendees transitioning to regular attendees, indicating they contribute to fine-grained adjustments rather than early decision shifts.

Table 2. Performance comparison of different finetuning methods for LLMs.

Method	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
Zero-shot	75.2 ± .5	70.1 ± .6	65.4 ± .7	67.7 ± .4	72.8 ± .5	64.9 ± .6
Few-shot	78.5 ± .4	73.3 ± .5	69.7 ± .6	71.5 ± .5	76.2 ± .4	68.1 ± .5
Full Fine-Tuning	91.5 ± .1	88.7 ± .2	90.2 ± .3	89.4 ± .3	89.2 ± .2	87.8 ± .1
Adapters [14]	91.1 ± .2	88.3 ± .3	89.8 ± .4	89.0 ± .3	88.9 ± .3	87.3 ± .3
Prefix [24]	90.8 ± .2	88.0 ± .3	89.4 ± .4	88.7 ± .2	88.5 ± .3	87.0 ± .3
Prompt [23]	90.5 ± .3	87.6 ± .4	89.0 ± .5	88.3 ± .2	88.2 ± .4	86.7 ± .4
LoRA [15]	**92.0 ± .1**	**89.2 ± .2**	**90.9 ± .4**	**90.0 ± .1**	**89.7 ± .2**	**88.3 ± .1**

Table 3. Comparison of imputation methods with explicit "missing" text.

Method	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
Mean	89.0 ± .5	86.5 ± .4	87.2 ± .5	86.8 ± .4	87.0 ± .5	85.5 ± .5
Median	89.2 ± .4	86.8 ± .4	87.4 ± .4	87.1 ± .3	87.3 ± .4	85.8 ± .5
KNN [25]	91.7 ± .2	88.5 ± .3	90.1 ± .4	89.3 ± .3	89.0 ± .3	87.5 ± .3
XGBoost [28]	91.8 ± .2	88.8 ± .3	90.3 ± .3	89.5 ± .2	89.3 ± .3	87.9 ± .3
MIWAE [30]	91.7 ± .2	88.6 ± .3	90.2 ± .3	89.4 ± .2	89.2 ± .3	87.7 ± .3
GAIN [52]	91.6 ± .3	88.4 ± .3	90.1 ± .3	89.2 ± .3	89.1 ± .3	87.6 ± .3
Text (Ours)	92.0 ± .1	**89.2 ± .2**	**90.9 ± .4**	**90.0 ± .1**	89.7 ± .2	88.3 ± .1

4.3 Further Analysis

Parameter-efficient Finetuning Methods. Table 2 presents the performance of different fine-tuning approaches for LLMs in longitudinal survey classification. Zero-shot learning achieves the lowest performance, indicating that using a pretrained LLM without task-specific finetuning is insufficient. Few-shot tuning provides a moderate improvement but remains limited in effectively capturing survey patterns. Full fine-tuning demonstrates strong performance but requires significantly more computational resources. Among parameter-efficient methods, LoRA achieves the highest accuracy at 92.0% and the strongest recall at 90.9%, outperforming adapters, prefix-tuning, and prompt-tuning, which exhibit slightly lower but comparable performance. While full fine-tuning performs well, LoRA matches or surpasses it across all metrics with substantially reduced computational overhead. These results highlight the effectiveness of LoRA in adapting LLMs for longitudinal survey classification while maintaining efficiency.

Missing Data. Table 3 compares different missing data handling methods in longitudinal survey classification with STM. Traditional imputation techniques, such as mean and median imputation, yield the lowest performance, indicating their limited effectiveness in reconstructing missing information. More advanced methods, including KNN [25], XGBoost-based imputation [28], Monte Carlo Importance-Weighted Autoencoder (MIWAE) [30], and Generative Adversarial Imputation Nets (GAIN) [52], leverage statistical and machine learning-based imputation strategies to infer missing responses, improving overall classification performance. However, our approach, which retains missing responses as explicit

tokens rather than imputing values, achieves similar performance, demonstrating that LLMs can implicitly model missingness without requiring explicit data reconstruction. These results suggest that imputing missing responses may not be necessary when using LLMs, as they can naturally infer meaningful patterns from the surrounding context.

Fairness Evaluation. While ethnicity is not included as a predictor in our model, we assess fairness by evaluating Equal Opportunity, which compares true positive rates across ethnic groups, and Equalised Odds, which ensures both false positive and false negative rates remain consistent [54]. We examine classification performance across five major groups specific to the New Zealand setting: European, Māori, Pacific, Asian, Middle Eastern, Latin American, and African ethnicities (MELAA) and Others. For the children belonging to more than one ethnic group, we record their ethnicity in the priority order of Māori, Pacific, Asian, MELAA and others, and Europeans. These groups reflect the diverse composition of the GUiNZ cohort, which aligns with birth demographics in Auckland at the time of recruitment [34]. Statistical analysis using the Kruskal-Wallis test finds no significant differences in true positive rates or false positive rates across ethnicities, indicating that classification outcomes are consistent across demographic groups.

5 Conclusion

We proposed Survey-as-Text Modelling (STM) to address challenges in analysing longitudinal survey data, including irregular feature alignment, evolving context, and the integration of open-ended responses. By representing survey data as text and leveraging LLMs, STM preserves contextual meaning and enables more flexible predictive modelling compared to conventional tabular approaches. Our results demonstrate that STM significantly outperforms traditional machine learning models, transformer-based methods, and LLM-based time-series models across all evaluation metrics. Additionally, we introduced gradient-guided counterfactual analysis to enhance interpretability by identifying the most influential survey items affecting attendance classification. This analysis revealed that recent social, emotional, and economic factors play a crucial role in distinguishing attendance patterns. These findings contribute to the methodological advancement of longitudinal survey analysis and provide data-driven insights for policymakers seeking to improve school attendance.

Acknowledgment. We thank the *Growing Up in New Zealand* team and the *New Zealand Ministry of Education* for providing access to data, and acknowledge the support and contributions of the *Our Voices* team and study investigators. This research was funded by the *Our Voices* programme (Endeavour grant UOAX1912), supported by the Ministry of Business, Innovation and Employment (2019–2025). The study was conducted in accordance with the Declaration of Helsinki, and all procedures involving human subjects were approved by the Ministry of Health's Northern B Health and Disability Ethics Committee. Consent was obtained from all participants and their parents or guardians. Tingrui Qiao is supported by the University of Auckland Doctoral

Scholarship and CSGST travel award from the School of Computer Science, University of Auckland.

References

1. Adler, D.A., Wang, F., Mohr, D.C., Choudhury, T.: Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. Plos one **17**(4), e0266516 (2022)
2. Atkinson, J., Salmond, C., Crampton, P.: Nzdep2013 index of deprivation. Wellington: Department of Public Health, University of Otago **5541**, 1–64 (2014)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
5. Cao, D., Jia, F., Arik, S.O., Pfister, T., Zheng, Y., Ye, W., Liu, Y.: Tempo: Prompt-based generative pre-trained transformer for time series forecasting. arXiv preprint arXiv:2310.04948 (2023)
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: SIGKDD. pp. 785–794 (2016)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
8. Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (gru) neural networks. In: MWSCAS. pp. 1597–1600. IEEE (2017)
9. Erceg-Hurn, D.M., Mirosevich, V.M.: Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. American Psychologist **63**(7), 591 (2008)
10. Finning, K., Ukoumunne, O.C., Ford, T., Danielson-Waters, E., Shaw, L., Romero De Jager, I., Stentiford, L., Moore, D.A.: The association between anxiety and poor attendance at school–a systematic review. Child and adolescent mental health **24**(3), 205–216 (2019)
11. Harder, F., Bauer, M., Park, M.: Interpretable and differentially private predictions. In: Proceedings of AAAI. vol. 34, pp. 4083–4090 (2020)
12. Heyen, H., Widdicombe, A., Siegel, N.Y., Perez-Ortiz, M., Treleaven, P.: The effect of model size on llm post-hoc explainability via lime. arXiv preprint arXiv:2405.05348 (2024)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)

15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
16. Huang, S., Mamidanna, S., Jangam, S., Zhou, Y., Gilpin, L.H.: Can large language models explain themselves? a study of llm-generated self-explanations. arXiv preprint arXiv:2310.11207 (2023)
17. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), `https://arxiv.org/abs/2310.06825`
18. Jin, B.T., Choi, M.H., Moyer, M.F., Kim, D.A.: Predicting malnutrition from longitudinal patient trajectories with deep learning. PloS one **17**(7), e0271487 (2022)
19. Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.Y., Liang, Y., Li, Y.F., Pan, S., et al.: Time-llm: Time series forecasting by reprogramming large language models. arXiv preprint arXiv:2310.01728 (2023)
20. Kearney, C.A., Benoit, L., Gonzálvez, C., Keppens, G.: School attendance and school absenteeism: A primer for the past, present, and theory of change for the future. In: Frontiers in Education. vol. 7, p. 1044608. Frontiers (2022)
21. Klein, M., Sosu, E.M., Dare, S.: Mapping inequalities in school attendance: The relationship between dimensions of socioeconomic status and forms of school absence. Children and Youth Services Review **118**, 105432 (2020)
22. Laith, R., Vaillancourt, T.: The temporal sequence of bullying victimization, academic achievement, and school attendance: A review of the literature. Aggression and violent behavior **64**, 101722 (2022)
23. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
24. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
25. Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F.C., Tseng, G.C.: Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC bioinformatics **15**, 1–12 (2014)
26. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625 (2023)
27. Liu, Y., Wu, H., Wang, J., Long, M.: Non-stationary transformers: Exploring the stationarity in time series forecasting. NeuraIPS **35**, 9881–9893 (2022)
28. Madhu, G., Bharadwaj, B.L., Nagachandrika, G., Vardhan, K.S.: A novel algorithm for missing data imputation on machine learning. In: 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). pp. 173–177. IEEE (2019)
29. Maltezou, H.C., Ledda, C., Sipsas, N.V.: Absenteeism of healthcare personnel in the covid-19 era: a systematic review of the literature and implications for the post-pandemic seasons. In: Healthcare. vol. 11, p. 2950. MDPI (2023)
30. Mattei, P.A., Frellsen, J.: Miwae: Deep generative modelling and imputation of incomplete data sets. In: ICML. pp. 4413–4423. PMLR (2019)
31. McConnell, B.M., Kubina Jr, R.M.: Connecting with families to improve students' school attendance: A review of the literature. Preventing School Failure: Alternative Education for Children and Youth **58**(4), 249–256 (2014)
32. Medsker, L.R., Jain, L., et al.: Recurrent neural networks. Design and Applications **5**(64-67), 2 (2001)
33. Mohammadi, B.: Explaining large language models decisions using shapley values. arXiv preprint arXiv:2404.01332 (2024)

34. Morton, S.M., Atatoa Carr, P.E., Grant, C.C., Robinson, E.M., Bandara, D.K., Bird, A., Ivory, V.C., Kingi, T.K.R., Liang, R., Marks, E.J., et al.: Cohort profile: growing up in new zealand. International journal of epidemiology **42**(1), 65–75 (2013)
35. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730 (2022)
36. Nitski, O., Azhie, A., Qazi-Arisar, F.A., Wang, X., Ma, S., Lilly, L., Watt, K.D., Levitsky, J., Asrani, S.K., Lee, D.S., et al.: Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data. The Lancet Digital Health **3**(5), e295–e305 (2021)
37. Pan, Z., Jiang, Y., Garg, S., Schneider, A., Nevmyvaka, Y., Song, D.: s^2ip-llm: Semantic space informed prompt learning with llm for time series forecasting. In: ICML (2024)
38. Pang, B., Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Libra: Measuring bias of large language model from a local context. In: European Conference on Information Retrieval. pp. 1–16. Springer (2025)
39. Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Thematic-lm: A llm-based multi-agent system for large-scale thematic analysis. In: Proceedings of the ACM on Web Conference 2025. pp. 649–658 (2025)
40. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
41. Sheetal, A., Jiang, Z., Di Milia, L.: Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. Applied Psychology (2023)
42. Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena **404**, 132306 (2020)
43. Srinivas, S., Fleuret, F.: Rethinking the role of gradient-based attribution methods for model interpretability. arXiv preprint arXiv:2006.09128 (2020)
44. Sun, C., Li, H., Li, Y., Hong, S.: Test: Text prototype aligned embedding to activate llm's ability for time series. arXiv preprint arXiv:2308.08241 (2023)
45. Tan, Z., Chen, T., Zhang, Z., Liu, H.: Sparsity-guided holistic explanation for llms with interpretable inference-time intervention. In: Proceedings of AAAI. vol. 38, pp. 21619–21627 (2024)
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeuraIPS **30** (2017)
48. Wan, R., Mei, S., Wang, J., Liu, M., Yang, F.: Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. Electronics **8**(8), 876 (2019)
49. White, R.T., Arzi, H.J.: Longitudinal studies: Designs, validity, practicality, and value. Research in science education **35**, 137–149 (2005)
50. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. NeuraIPS **34**, 22419–22430 (2021)
51. Xue, H., Salim, F.D.: Promptcast: A new prompt-based learning paradigm for time series forecasting. IEEE TKDE **36**(11), 6851–6864 (2023)
52. Yoon, J., Jordon, J., Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: ICML. pp. 5689–5698. PMLR (2018)

53. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of AAAI. vol. 37, pp. 11121–11128 (2023)
54. Zhang, J., Bareinboim, E.: Equality of opportunity in classification: A causal approach. NeuraIPS **31** (2018)
55. Zhang, Y., Yan, J.: Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: ICLR (2023)
56. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of AAAI. vol. 35, pp. 11106–11115 (2021)
57. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In: Proceedings of ICML. pp. 27268–27286. PMLR (2022)
58. Zhou, T., Niu, P., Sun, L., Jin, R., et al.: One fits all: Power general time series analysis by pretrained lm. NeuraIPS **36**, 43322–43355 (2023)

A List of Variables

1. **New Zealand Deprivation Index (NZDep)**: an area-based measure of socioeconomic deprivation in New Zealand based on nine Census variables.
2. **Crowding groups**: how many people live in the house.
3. **Easy access to school**: whether this is a deciding factor for choosing a school.
4. **Resource provided by the school**: whether the ability of the school to provide good resources is a deciding factor for choosing a school.
5. **Children's physical needs**: degree of satisfaction from mother.
6. **Children's learning needs**: degree of satisfaction from mother.
7. **Children's social and emotional needs**: degree of satisfaction from mother.
8. **Children's culture needs**: degree of satisfaction from mother.
9. **Difficulty starting school**: level of difficulty and how long the difficulty lasts.
10. **Parental support**: whether the mother is confident she knows how to help her children do well at school.
11. **Belong to school (mother)**: mother feels comfortable and welcomed when visiting the school.
12. **Belong to school (children)**: children feel they belong to their school.
13. **Number of moves after the last wave**.
14. **Form of transport and duration of transport**.
15. **Put up with feeling cold**.
16. **Gone without fresh fruit or vegetables**.
17. **Centre for Epidemiologic Studies Short Depression Scale (CES-D-R 10)**: a concise self-report tool to assess depressive symptoms, comprising 10 items rated on a 4-point Likert scale, with higher total scores indicating greater depressive symptomatology.
18. **Work status of the mother**.
19. **Household income groups**.
20. **Housing tenure**.

Table 4. Performance comparison of different LLM backbones for school attendance classification.

Backbone	Accuracy	Precision	Recall	F1 Score	EiCAT
BERT [7]	83.2 ± .3	80.5 ± .4	81.0 ± .5	80.7 ± .4	5.9 ± .3
GPT-2 [40]	85.0 ± .3	82.7 ± .3	83.1 ± .4	82.9 ± .3	1.6 ± .4
Mistral-7B [17]	92.0 ± .2	89.3 ± .3	91.0 ± .3	90.1 ± .3	11.5 ± .3
Qwen1.5-7B [3]	92.1 ± .1	89.2 ± .3	90.9 ± .4	90.0 ± .3	10.9 ± .3
LLaMA-3.1-8B [46]	92.0 ± .1	89.2 ± .2	90.9 ± .4	90.0 ± .1	11.2 ± .1

21. **Rurality**.
22. **Time and energy for parenting**.
23. **Home atmosphere**.
24. **Children have enough friends and are treated well by them**.
25. **Children are bullied at school**.
26. **Culture acceptance at school**.
27. **Gender acceptance at school**.
28. **Sleeping quality**.
29. **Children feel supported by their family**.
30. **Children feel supported by their friends**.
31. **Miss school due to COVID-19**.
32. **Financial stress due to COVID-19**.
33. **People getting along at home during COVID-19**.
34. **Worries and fears of social mixing during COVID-19**.
35. **Teachers respect and are fair to children**.
36. **School work stress**.
37. **Things the children look forward to for the next few years**.
38. **Things the children worry about for the next few years**.

B LLM Backbones

To assess the impact of different language model architectures on school attendance classification, we evaluate STM using a range of local, smaller-scale LLM backbones. This experiment ensures that our approach remains effective across different models while addressing data privacy constraints by using locally hosted models. Table 4 presents the results, showing that STM achieves consistent performance across various backbones, with all models in the LLM category performing within a close range. Smaller models such as BERT [7] and GPT-2 [40] exhibit lower performance. Apart from performance metrics, we also measured the bias within the LLMs in the New Zealand context through the EiCAT score [38]. We found that LLMs tend to have lower biases than smaller models such as BERT and GPT-2, as shown by the larger EiCAT scores.