MathBuddy: A Multimodal System for Affective Math Tutoring

Debanjana Kar[*1,2], **Leopold Böss**[*1], **Dacia Braca**[*1],
Sebastian Dennerlein[1], **Nina Hubig**[1], **Philipp Wintersberger**[1], **Yufang Hou**[1]

[1]IT:U Interdisciplinary Transformation University Austria
[2]IBM Research India

Correspondence: debanjana.kar1@ibm.com, {leopold.boess, dacia.braca}@it-u.at

Abstract

The rapid adoption of LLM-based conversational systems is already transforming the landscape of educational technology. However, the current state-of-the-art learning models do not take into account the student's affective states. Multiple studies in educational psychology support the claim that positive or negative emotional states can impact a student's learning capabilities. To bridge this gap, we present *MathBuddy*, an emotionally aware LLM-powered Math Tutor, which dynamically models the student's emotions and maps them to relevant pedagogical strategies, making the tutor-student conversation a more empathetic one. The student's emotions are captured from the conversational text as well as from their facial expressions. The student's emotions are aggregated from both modalities to confidently prompt our LLM Tutor for an emotionally-aware response. We have evaluated our model using automatic evaluation metrics across eight pedagogical dimensions and user studies. We report a massive 23 point performance gain using the win rate and a 3 point gain at an overall level using DAMR scores which strongly supports our hypothesis of improving LLM-based tutor's pedagogical abilities by modeling students' emotions. Our dataset and code are available at: https://github.com/ITU-NLP/MathBuddy.[1]

1 Motivation

The integration of AI tutors into educational platforms has rapidly evolved with the emergence of large language models (LLMs) (Chu et al., 2025), enabling natural and interactive support for learners across subjects, such as mathematics (Macina et al., 2023), and mitigating the issue of lack of qualified personal tutors.[2] Several studies (Kumar et al., 2024; Li et al., 2024; Wang et al., 2024) claim that

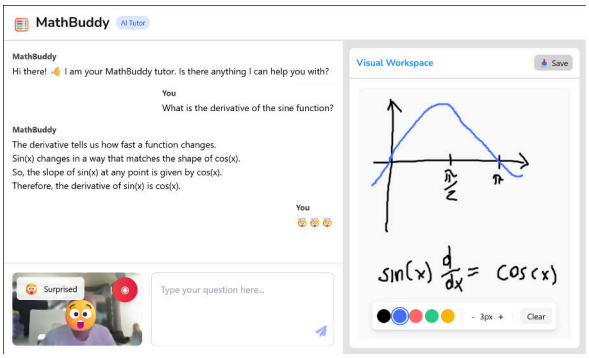

Figure 1: MathBuddy's user interface showcasing an example interaction using all functionality.

the integration of LLM-powered tutors facilitates formative self-regulated learning, where students learn and assess themselves at an independently decided pace.

Despite significant progress in instructional capabilities, most existing LLM-based tutoring systems lack emotional intelligence (Liu et al., 2024; Azerbayev et al., 2024), treating all learners homogeneously without accounting for their affective states. This can lead to disengagement, frustration, and ultimately suboptimal learning outcomes, especially in subjects such as mathematics, where learner anxiety is well-documented (Daker et al., 2023; Maki et al., 2024; Schoenherr et al., 2025).

Research in educational psychology has long emphasized the role of emotions in learning (Pekrun et al., 2002; Graesser et al., 2012). Negative affective states, e.g., confusion and frustration, can inhibit problem-solving ability and motivation, while positive emotions, such as engagement and curiosity, are positively correlated with learning gains (Pekrun, 2006; Tan et al., 2021; Zheng et al., 2023).

At the same time, recent advances in NLP and computer vision have opened up new avenues for multimodal affect recognition. Transformer-based models can infer affective cues from text with increasing accuracy (Chutia and Baruah, 2024; Ali

[*]Authors contributed equally to this work.
[1]Link to demo video: https://youtu.be/ZUjgmOw9GM0
[2]https://unesdoc.unesco.org/ark:/48223/pf0000385723

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 821–838
November 4-9, 2025

et al., 2024), while facial expression recognition using deep learning has reached practical levels of robustness (Makhmudov et al., 2024). However, existing LLM-based tutoring systems do not yet combine these modalities in real time to support accurate, empathetic adaptation in mathematics tutoring.

To address this gap, we present our AI-based math tutor, *MathBuddy* (see Figure 1), that models student emotions using both textual and visual modalities. More specifically, given a student's last utterance and facial expressions, *MathBuddy* extracts the student's emotions from both modalities and aggregates them into one of three classes (*Positive, Negative, or Neutral*). The emotion is then used to direct the LLM Tutor in one of the following ways: if the student is in a positive or neutral affective state, the student is challenged by *MathBuddy*, while a negative affective state causes the system to try to motivate the student.

We have evaluated the effectiveness of our system across eight different pedagogical dimensions (Maurya et al., 2025) through both automatic evaluation (Section 3) and real-time user studies (Section 4). The results from both the evaluation strategies emphasize the importance of modeling student emotions in tutoring strategies.

Our work's main contributions are as follows:

1. We present the first emotionally-aware LLM tutoring system, *MathBuddy* that is grounded in education theory and adapts its response based on the student's affective state.

2. We develop an automatic evaluation system to evaluate the framework thoroughly on a recent math tutor benchmark across eight different pedagogical dimensions based on the concepts introduced in (Maurya et al., 2025).

3. We have annotated 224 student utterances with emotion labels and polarity to support the development of emotion recognition models from text.

Our code, annotated dataset and a link to the system demo video are publicly available at: https://github.com/ITU-NLP/MathBuddy.

2 *MathBuddy*: Design & Implementation

In real-life tutoring scenarios, tutors often adapt their pedagogical strategies to the student's present emotional state (Lin et al., 2022). However, hardly any of the state-of-the-art AI-powered tutoring systems consider modeling students' emotions in this process. Our proposed system, *MathBuddy*, aims to address this gap by leveraging multimodal interaction channels. *MathBuddy* models one-on-one interactions between a student and a tutor in the domain of Mathematics. In this section, we provide a detailed breakdown of the system's implementation process.

2.1 Emotion Recognition from Text and Webcam Data

Since *MathBuddy* is a conversational system, we try to gauge the student's affective state by extracting the emotion from the student's turn in the conversation. However, a major bottleneck we encountered in this task is the lack of such an annotated student-tutor conversational dataset. To overcome this challenge, we annotated the student turns in the hard version of the MathDial-Bridge dataset (Macina et al., 2025). Inspired by D'Mello and Graesser (2012) work on affect in learning contexts, we assumed that students' emotions can be multifaceted and vary in intensity. Hence, we modeled emotion extraction as a multi-label classification task with three target states: *Boredom*, *Engagement*, and *Neutral* (D'Mello and Graesser, 2012). To capture the intensity of each student's emotional state at every turn, we assigned a polarity score on a scale from 0 to 2, where 0 corresponds to low intensity and 2 to high, thus serving as an indicator for arousal (Posner et al., 2005). Four annotators with academic backgrounds, aged between 27 and 38, participated in the annotation process. After an initial round, which yielded a Cohen's Kappa inter-annotator agreement score of 40%, three annotators took part in the conflict adjudication process. Finally, 224 unique student turns were annotated ensuring an equal distribution across the labels. We reserve the manually annotated dataset as our test set and generate a noisy student-tutor annotated dataset using DeepSeek-R1-Distill-LLama-70B (DeepSeek-AI, 2025) as our training dataset. We fine-tune multiple BERT-based models on the silver-labeled training data. The best model (Sanh, 2019) reports an accuracy and F1-score of 61.8% and 57.9% respectively. The detailed results are available in Appendix A.

In addition to text-based emotion recognition, the system processes webcam data to extract a student's affective states based on the student's facial expression.

We use the *face-api.js* package (Mühler, 2024) for the task of detecting students' emotions through their facial expressions. The *face-api.js* package represents a JavaScript library that builds on *tensorflow.js* to provide functionality related to human faces, such as face recognition, face landmark detection, and face expression recognition. The system utilizes its lightweight and fast in-browser face expression recognition. It is configured to recognise the emotions *Happy*, *Sad*, *Angry*, *Surprised*, *Fearful*, *Disgusted*, and *Neutral*, with *Neutral* being the default when no face is being detected. With our current configurations and considering the mapping explained later in Subsection 2.2, *face-api.js* reports an accuracy of 76% and an F1-score of 71% on our test set. The test set comprises 151 images from FERAC (Das Chaiti and Afrin, 2024) and 38 images from the Facial Emotion Recognition Dataset (Data, 2023), combined to ensure diversity across age groups as well as the inclusion of individuals diagnosed as non-neurotypical.

Given that the system requires all emotional states to be attributed to the specific timestamp of a message, the face emotion samples have to be aggregated over the interval bounded by two messages. For this step, the system considers only the changes in recognised emotion, grouping equal consecutive emotion states. This way, each group can be assigned a duration as well as an age—the time from now to the last sample within the group. Aggregation computes a weighted sum of the durations of all groups associated with one specific emotion, choosing the emotion with the highest sum as the result. The required weights are derived from a half-life decay function employing a group's age and a fixed half-life of 120 seconds. See Algorithm 1 for a specific description.

Algorithm 1 Temporal emotion aggregation.

Require: Sequence of emotion samples $S = \{(e_1, t_1), \dots (e_n, t_n)\}$ sorted by time t
Initialize group list $G \leftarrow [\,]$
Current emotion $e_c \leftarrow e_1$, start time $t_s \leftarrow t_1$
for $i = 2$ to n **do**
  if $e_i \neq e_c$ **then**
    Add new group (e_c, t_s, t_{i-1}) to G
    $e_c \leftarrow e_i$, $t_s \leftarrow t_i$
  end if
end for
Add final group (e_c, t_s, t_n)
Initialize emotion score map $S_e \leftarrow \{\}$
Half-life constant $\lambda \leftarrow \ln(2)/120$
for each group $(e, t_s, t_e) \in G$ **do**
  duration $d \leftarrow t_e - t_s$
  age $a \leftarrow \text{now} - t_e$
  weight $w \leftarrow \exp(-\lambda \times a)$
  $S_e[e] \leftarrow S_e[e] + d \times w$
end for
$e^* \leftarrow \arg\max_e S_e[e]$
return e^*

2.2 Multimodal Emotion Aggregation

Users are assumed to feel only one unique emotion at any given time. Therefore, the system aims to approximate this emotion by compiling all sampled emotional states into a single one. Each modality ideally increases the accuracy of this aggregation.

A straightforward approach is to project the emotions sampled from both modalities to a common value space to allow them to be merged. This value space holds three primitive emotion states: *Positive*, *Neutral*, and *Negative*.

The emotion mapping applied by the system can be expressed as follows:

$$m : E \to P$$

$$m(e) = \begin{cases} \text{Positive}, & \text{if } e \in E_{\text{positive}} \\ \text{Neutral}, & \text{if } e \in E_{\text{neutral}} \\ \text{Negative}, & \text{if } e \in E_{\text{negative}} \end{cases}$$

where E denotes the set of all recognizable emotions across modalities, and P denotes the set of primitive categories: *Positive*, *Negative*, and *Neutral*. Note that P is a subset of E used for simplified downstream processing. The specifics of how each emotion detected from either modality is mapped to one of the above primitive emotion states is detailed in Appendix G. Each emotion state produced by either emotion recognition method can be directly mapped to these primitives, including their confidence values. Notably, states produced via face-based recognition are mapped before aggregation. Merging the resulting values considers the following rule: *Positive* and *Negative* are always preferred over *Neutral*; confidence is used as a secondary criterion if both values are not equal to *Neutral*. For example, a text-based positive emotion with a confidence of 50% and a contradicting facial-based negative emotion with a confidence of 70% are reconciled into a negative emotion. Hence, the system assumes that any hint of non-neutral

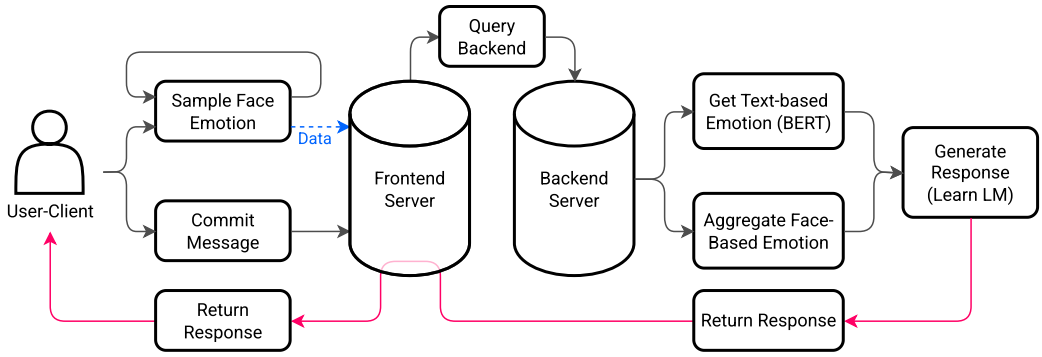

Figure 2: An overview of the system's logical flow, showcasing the core components.

emotion implies that neutrality is not given. This behavior appears beneficial as a non-neutral emotional state should enable better adjustment of the currently applied pedagogical strategy.

2.3 Emotion-Aware LLM Tutor

Inspired by educational psychology literature (Pekrun et al., 2002; Graesser et al., 2012), we design our system to map dynamically extracted student emotions to relevant pedagogical strategies in order to enhance the learning experience for the student. After extracting and aggregating multi-modal student emotions, we finally map it to relevant pedagogical strategies such that:

- *Positive* student emotion, the LLM tutor is prompted to *challenge* the student;
- *Neutral* or a *Negative* emotion, the tutor is prompted to *motivate* the student.

Please refer to Appendix C for detailed prompts.

2.4 System Implementation

The system is divided into two core components: Frontend and Backend. The Frontend represents a web server implemented in TypeScript that serves the web-based client and a public API, providing all functionality needed by the said client. This way, all requests must go through the Frontend, while all other components remain hidden from the outside. These requests mainly include the tutor response generation, when the user submits a message, and applying the face based emotion recognition to the webcam input. The later can be configured to be handled in browser by *face-api.js* or by the backend via our own model. Additionally, the Frontend also represents the main storage, holding a primitive in-memory database.

The client employs a simple design consisting of two parts: a chat-based interface and a visual workspace. The chat-based interface visualizes the tutor-student conversation similar to established LLM UIs. In addition to a text field, this interface also features an optional webcam input. The client uses this webcam input to query the user's emotional state in regular intervals; it can be configured to give live feedback on the recognised emotion. The visual workspace offers space to take notes or sketch ideas via mouse input. This feature is an artifact of the idea to submit handwritten notes to the tutor, reserved for future research, owing to implementation constraints.

The Backend component acts as a REST API, implementing all functionality related to inference via Python. It both runs local models and employs external APIs to respond to queries made by the Frontend. This structure allows for adapting models or configurations at the Backend without necessitating changes to the Frontend. Its key endpoints include the tutor response generation based on conversation history and face-based emotion states, as well as face-based emotion recognition based on an image. An overview of the system's logical flow is illustrated in Figure 2.

3 Automatic Evaluation

For a comprehensive evaluation of the system, we evaluate *MathBuddy* through automatic evaluation strategies and a user study. In this section, we detail the automatic evaluation strategies along with an analysis of the quantitative results. We evaluate *MathBuddy* with different backend LLMs on an existing math tutor benchmark (Maurya et al., 2025) with emotion information from the input text only. We use the hard version of the MathDial Bridge dataset (Macina et al., 2023) containing 327 human-annotated conversations using two different metrics:

Table 1: Comparison study using Win Rate and DAMR across eight different pedagogical dimensions. The plus (+) version of each baseline represents our models. Best reported metrics are highlighted in bold. The overall score is a combined mean of the DAMR scores across the eight dimensions.

Model	Win Rate	Mistake Identification	Mistake Location	Answer Disclosure	Providing Guidance	Action-ability	Human-likeness	Coherence	Tutor Tone	Overall Score
QSLM	0.30	0.71	0.92	0.41	0.22	0.88	0.56	0.58	0.86	0.64
QSLM+	**0.37**	0.71	0.92	0.41	0.22	0.88	0.56	0.58	0.86	0.64
LlemmaMM	0.38	0.22	0.93	**0.53**	**0.35**	0.75	0.65	0.70	0.92	0.63
LlemmaMM+	**0.38**	**0.27**	**0.96**	0.50	0.30	**0.90**	**0.72**	**0.83**	**0.97**	**0.68**
LearnLM	0.59	0.96	1.00	0.65	0.32	0.99	0.93	0.91	1.00	0.85
LearnLM+	**0.82**	**1.00**	**1.00**	**0.69**	**0.37**	**0.99**	**0.98**	**0.98**	**1.00**	**0.88**

Table 2: Comparison study using Win Rate and DAMR across eight different pedagogical dimensions with a more complex prompt.

Model	Win Rate	Mistake Identification	Mistake Location	Answer Disclosure	Providing Guidance	Action-ability	Human-likeness	Coherence	Tutor Tone	Overall Score
QSLM	0.33	0.52	**0.94**	0.49	0.26	**0.86**	0.61	**0.68**	**0.92**	0.66
QSLM+	**0.35**	**0.57**	0.93	**0.49**	**0.28**	0.84	**0.61**	0.67	0.91	**0.66**
LlemmaMM	0.41	**0.86**	**1.00**	**0.53**	**0.23**	**0.99**	0.66	0.62	**1.00**	**0.74**
LlemmaMM+	**0.42**	**0.22**	**0.95**	0.49	**0.32**	**0.82**	**0.66**	**0.72**	**0.95**	**0.64**
LearnLM	0.46	0.87	0.99	0.51	0.21	1.00	0.65	**0.63**	1.00	0.73
LearnLM+	**0.46**	**0.87**	**0.99**	**0.57**	**0.24**	**1.00**	**0.67**	**0.59**	**1.00**	**0.74**

- **Win Rate**: Rate at which reward model (Macina et al., 2025) prefers the LLM Tutor response over the ground truth response.
- **Desired Annotation Match Rate (DAMR)**: % of labels assigned to the tutor responses matching the desiderata described in (Maurya et al., 2025).

Following (Maurya et al., 2025), to calculate the DAMR scores, we have assigned labels *"Yes", "No", "To some extent"* to the tutor responses using a round table of LLMs as Judges (LaaJ) (see Appendix F for the prompt details). A response gets a high DAMR score if the assigned label matches the desired label for the given pedagogical dimension as defined in (Maurya et al., 2025). The authors also report a poor correlation coefficient with two LLMs as judges across eight pedagogical dimensions. In our work, we have considered four different LLMs as judges, namely, Llama-3-3-70b-instruct (Grattafiori et al., 2024), Deepseek-V3 (DeepSeek-AI et al., 2025), Mixtral-8X22b-Instruct-v0.1 (Jiang et al., 2024), Phi-4 (Abdin et al., 2024) and a fifth ensemble model through majority voting for the same pedagogical dimensions. To establish the reliability of the LaaJ models, we use the mentioned LLMs to generate the labels for each tutor utterance on the annotated dataset[3] (Maurya et al., 2025) and report the Spearman Correlation, Pearson Correlation and Accuracy for the same (Appendix Table 8). We find that the ensemble LaaJ model serves as the best evaluator with the highest correlations across four out of the eight dimensions and comparable results for the other dimensions. We use this model to report our DAMR scores in Table 1. We have used the current state-of-the-art tutoring models (Liu et al., 2024; Azerbayev et al., 2024; Modi et al., 2024) as baselines for comparison.

Through Table 1, we can observe that our feature enhanced models perform consistently better compared to their respective baseline versions (for prompt used, check Appendix D). In the smaller models like Qwen2.5-7B-Socratic-LM (Liu et al., 2024) and Llemma-7B (Azerbayev et al., 2024), we do observe some variations, especially with regard to the *Mistake Identification* and *Mistake Location* dimensions. However, with much larger models like LearnLM 2.0, we see that the feature enhanced version produces responses with more desirable pedagogical qualities in six out of the eight dimensions and is comparable for the rest of the dimensions. We found consistent results (Table 2) when we replicated this experiment with a more complex prompt (prompt is attached in Appendix E, intended to test the model's pedagogy instruction following skills). This indicates the importance of modeling student's emotions across all the peda-

[3] https://github.com/kaushal0494/UnifyingAITutorEvaluation/blob/main/MRBench/MRBench_V2.json

Table 3: Ablation study results with emotions and education theory features. Prompt 1 and Prompt 2 refer to the simple and complex prompts respectively. QSLM = Qwen2.5-7B-Socratic-LM; ET = Education Theory

Model	Prompt1	Prompt2
QSLM	0.28	0.33
QSLM + emotion	0.21	0.12
QSLM + emotion + ET	**0.37**	**0.35**

gogical dimensions in the learning models. We occasionally observe variability in results when using smaller models — particularly with more complex prompts. This suggests that smaller models may struggle or become overwhelmed when presented with an excessive amount of information.

3.1 Ablation Studies

Table 3 highlights the importance of student emotion features mapped to education theory in enhancing the tutor's pedagogical capabilities. Across both the simple and complex prompts, prompting the model with relevant pedagogical strategies based on the detected emotion of the student results in a steep performance gain (by 9 points and 2 points through simple and complex prompts respectively). For the ablation study, we have used the Qwen2.5-7B-Socratic-LM (Liu et al., 2024) model. Since it is a smaller model, providing only the emotional information misguides the model. However, when instructions are grounded in educational theory, for instance, directing that if a student exhibits boredom, the instructional response should aim to enhance the learner's motivation, the model demonstrates a markedly stronger performance.

In general, across all dimensions, the emotion feature seems to have resulted in a performance gain particularly in *humanlikeness*, *actionability*, *coherence* and *tutor tone*. Each of these dimensions is related to a human's empathetic side, again indicating the significance of modeling student emotions in the models. Since we only have an annotated textual dataset, the quantitative results reported in Table 1 only employed emotions extracted from textual conversation. We evaluate the multimodal system in real-time through a user study discussed in the following section (Section 4) using the best model (LearnLM+).

4 User Study

To evaluate the impact of emotion-aware adaptation on learning performance, we conducted a within-subject user study using *MathBuddy*, our AI-based math tutoring system. The study aimed to assess whether integrating predicted student emotions leads to improved learning outcomes. A total of 30 participants (aged 15–55), representing diverse genders, nationalities, linguistic backgrounds, and educational profiles, took part in the study. Further details are provided in Appendix J.

4.1 User Study Analysis

The collected user study comprises a diverse set of metadata and multimodal data: (1) learning outcomes, measured through multiple-choice questionnaires on select math topics (serving as a pre- and post-test); (2) interaction data, in the form of chat logs from both tutoring conditions (Emotion ON/OF modalities); (3) affective states, derived from emotion predictions based on textual and facial inputs; (4) learning experience, measured with a 15-item post-interaction questionnaire including engagement and perceived support, rated on a 5-point Likert scale; (5) overall satisfaction, captured through a final rating-scale survey; and (6) user-generated ground truth for the aggregated emotion, through participants' gold annotations and corrections of the Emotion ON conversations.

Empathy is a desired quality in a tutor across pedagogical dimensions. Figure 3 reports the mean score of 30 participants across 15 questions, each of them mapped to different pedagogical dimensions, the same discussed in Section 3 (Table 9 in Appendix). Here, we observe that the average satisfaction scores are generally higher when emotions are used to guide *MathBuddy*'s pedagogical strategies (Emotion ON condition). This suggests that participants reported a more positive perception of this interaction mode across the questionnaire dimensions, which we consider as indicators of their overall learning experience.

Empathetic responses enhances user learning. Comparing participants' facial emotional dynamics over time, those in the Emotion ON condition displayed a higher frequency of positive expressions than those in the Emotion OFF condition (see Figure 8 in the Appendix).

Looking at the kernel density estimates (KDE) for the facial emotion duration distributions across conditions (Figure 4), we can see that all three emotion classes (*Negative*, *Neutral*, *Positive*) exhibit highly skewed distributions, with most durations clustered under 2 seconds.

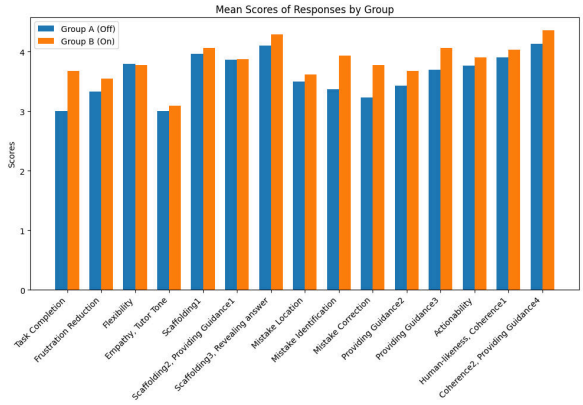

Figure 3: Average scores of participants with and without the emotion-aware condition

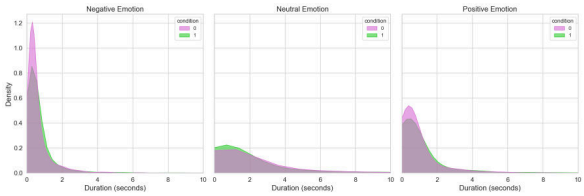

Figure 4: Kernel density estimates of facial emotion durations per class and condition.

We calculated the mean duration per participant for each of the three facial emotional states and conducted Wilcoxon paired t-tests among the conditions; the results are summarized in Table 4. We can appreciate that positive expressions showed a statistically significant difference ($p = .046$), with longer average durations under the Emotion ON condition (1.14s vs. 0.93s). This result indicates that participants experience longer positive emotional states under the Emotion ON condition.

These findings suggest that modeling student emotions and providing adaptive emotional feedback may have contributed to a more engaging and satisfying user experience during the tutoring sessions, potentially enhancing the overall learning process. Nonetheless, facial expressions commonly associated with happiness, such as smiling, can sometimes reflect more complex states, e.g., sarcasm or frustration. Expanding our multimodal channels could help capture emotional nuances more accurately, especially in technical domains such as mathematics, where affective signals are often subtle and context-dependent.

The detected emotions align with user emotions. As the last step of the experiment, the participants reviewed their own conversations in the Emotion ON condition, adjusting the tutor's emotion labels when necessary. To assess the system's ability to correctly detect the participant's emotions, we compare the aggregated text and facial emotions detected by the different modalities with the participant reviewed emotions. As reported in Table 5, the system achieves an overall accuracy of 60% in real-time usage. However, we observe a sharp fall in the recall scores for the *Neutral* class. This highlights the non-robust nature of our aggregation method where we try to suppress neutrality by preferring the modality that reports a non-neutral emotion. This can be improved by adopting a more sophisticated emotion aggregation method but remains to be explored as part of our future works.

Table 4: Wilcoxon paired test on mean emotion durations per user and emotion (in seconds).

Emotion	p-value	(Emo OFF) Mean ± Std	(Emo ON) Mean ± Std
Negative	0.537	0.72 ± 0.56	0.65 ± 0.42
Neutral	0.248	12.73 ± 30.41	7.07 ± 14.7
Positive	**0.046**	0.93 ± 0.87	1.14 ± 0.87

Table 5: Comparison of system aggregated multimodal emotions to participant gold annotations.

Emotion	Precision	Recall	F1-score
Negative	0.57	0.98	0.72
Neutral	0.71	0.15	0.25
Positive	0.79	0.41	0.54

5 Conclusion

In this paper, we introduced *MathBuddy*, an emotionally aware LLM-based math tutor that uses both text and facial expressions to model student emotions and deliver empathetic, pedagogically informed responses. By bridging affective cues with adaptive tutoring strategies, *MathBuddy* enhances student engagement and learning effectiveness by a massive 23 points using win rate (Macina et al., 2025) and 3 points at an overall level using DAMR scores as reported in Table 1. The usefulness of our approach is also supported through our carefully designed user studies as reported in Section 4.

Our evaluations highlight the benefits of integrating emotional awareness into LLM-driven education. We envision *MathBuddy* as a step toward more human-centered, emotionally intelligent learning systems.

Limitations

Generally, the development and evaluation of tutoring systems is limited by the lack of a golden standard or straightforward performance metrics. Qualitative analysis relies on feedback from users, who, however, may not be aware of a tutor system's ideal behavior—for example, instead of rating the educational prowess, they might rate user experience. Moreover, the user study described in this paper captured an age group devoid of children, whose idea of ideal tutoring may be vastly different and feature concepts our evaluation misses entirely. We also introduce a dataset with this work which currently contains coarse-grain emotion labels for student utterances. However, we feel that finer emotion labels capturing micro facial expressions may greatly benefit in capturing a much more nuanced student affective state. We leave this non-trivial task for future work. Based on the data and user feedback, it appears that the modalities currently used (text-based and face-based) may not be informative enough to accurately capture a user's emotional state. In technical education contexts, emotional expression in text is often subtle, unlike in other genres such as poetry (Hou and Frank, 2015). This insight motivates future investigation into additional modalities, e.g., spoken audio, handwritten notes, or biometric sensors, which ideally considerably increase emotion recognition accuracy. Further, spoken audio and handwritten notes more accurately represent the classical learning environment people might be more familiar with than a digital chat interface. We also want to address the aggregation strategy that we use for merging the emotion classes from different modalities in our system. The strategy can be further improved to handle neutral emotional states better. We leave this improvement for future work.

Acknowledgments

We would like to acknowledge the contributions of Yesica Yanina Duarte who helped us design the user study and the interface of the system. We also want to acknowledge Prof. Daniel Klotz who helped us with his deep insights and interesting critiques on this project. We extend a hearty vote of thanks to all the enthusiastic participants of the user study who are mostly members of the IT:U Interdisciplinary Transformation University Austria.

Ethical Considerations

This work acknowledges the ethical implications of our system's design and the data it collects. All conducted studies adhered to the GDPR (Goddard, 2017), upholding all included principles, e.g., purpose limitation, data minimization, confidentiality, as well as lawful, fair, and transparent processing.

Before data collection, informed consent was obtained, clearly outlining the purpose of the research, the voluntary nature of participation, and the right to withdraw at any time without consequence. All data were anonymized to protect participant identities. Special attention was given to exclusively collecting the information necessary to fulfill the functionality of real-time interaction. Although the system processes identifiable data such as image recordings of a participant's face, this is solely used for emotion recognition and never stored beyond model inference. The entire process followed the institutional ethical guidelines set by the IT:U Interdisciplinary Transformation University Austria.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Sara Ali, Bushra Naz, Sanam Narejo, Sahil Ali, and Jitander Kumar Pabani. 2024. Transformers Unveiled: A Comprehensive Evaluation of Emotion Detection in Text Transcription. In *2024 Global Conference on Wireless and Optical Technologies (GCWOT)*, pages 1–7. IEEE.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An Open Language Model for Mathematics. In *The Twelfth International Conference on Learning Representations*.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and 1 others. 2025. LLM agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Tulika Chutia and Nomi Baruah. 2024. A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, 57(8):203.

Richard J Daker, Sylvia U Gattas, Elizabeth A Necka, Adam E Green, and Ian M Lyons. 2023. Does anxiety explain why math-anxious people underperform in math? *npj Science of Learning*, 8(1):6.

Rajasree Das Chaiti and Mahzuzah Afrin. 2024. FERAC Dataset. https://www.kaggle.com/datasets/rajasreechaiti/ferac-dataset. Accessed: 2024-07-01.

Training Data. 2023. Facial Emotion Recognition Dataset. https://huggingface.co/datasets/username/dataset-name. Accessed: 2024-07-01.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *Preprint*, arXiv:2412.19437.

Sidney K D'Mello and Art Graesser. 2012. Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5(4):304–317.

Michelle Goddard. 2017. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6):703–705.

Arthur C. Graesser, Mark W. Conley, and Andrew Olney. 2012. *Intelligent tutoring systems.*, pages 451–473. APA handbooks in psychology®. American Psychological Association, Washington, DC, US.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Yufang Hou and Anette Frank. 2015. Analyzing Sentiment in Classical Chinese Poetry. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 15–24, Beijing, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of Experts. *Preprint*, arXiv:2401.04088.

Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N. Rafferty, John Stamper, and Michael Liut. 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 86–97, New York, NY, USA. Association for Computing Machinery.

Belle Li, Curtis J. Bonk, Chaoran Wang, and Xiaojing Kou. 2024. Reconceptualizing Self-Directed Learning in the Era of Generative AI: An Exploratory Analysis of Language Learning. *IEEE Transactions on Learning Technologies*, 17:1489–1503.

Jionghao Lin, Shaveen Singh, Lele Sha, Wei Tan, David Lang, Dragan Gašević, and Guanliang Chen. 2022. Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems*, 127:194–207.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 85693–85721. Curran Associates, Inc.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Fazliddin Makhmudov, Alpamis Kultimuratov, and Young-Im Cho. 2024. Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures. *Applied Sciences*, 14(10):4199.

Kathrin E. Maki, Anne F. Zaslofsky, Robin Codding, and Breanne Woods. 2024. Math anxiety in elementary students: Examining the role of timing and task complexity. *Journal of School Psychology*, 106:101316.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Mike Schaekermann, Miruna Pîslar, and 26 others. 2024. LearnLM: Improving gemini for Learning.

Vincent Mühler. 2024. face-api.js: JavaScript API for Face Recognition and Face Detection. https://github.com/justadudewhohacks/face-api.js. Accessed: 2025-07-01.

Reinhard Pekrun. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18:315–341.

Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P Perry. 2002. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist*, 37(2):91–105.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.

V Sanh. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Johanna Schoenherr, Stanislaw Schukajlow, and Reinhard Pekrun. 2025. Emotions in mathematics learning: a systematic review and meta-analysis. *ZDM–Mathematics Education*, pages 1–18.

Jing Tan, Jie Mao, Yizhang Jiang, and Ming Gao. 2021. The influence of academic emotions on learning effects: A systematic review. *International journal of environmental research and public health*, 18(18):9678.

Chaoran Wang, Zixi Li, and Curtis Bonk. 2024. Understanding self-directed learning in AI-Assisted writing: A mixed methods study of postsecondary learners. *Computers and Education: Artificial Intelligence*, 6:100247.

Juan Zheng, Susanne Lajoie, and Shan Li. 2023. Emotions in self-regulated learning: A critical literature review and meta-analysis. *Frontiers in psychology*, 14:1137010.

Appendix

A Trained BERT Evaluation

Table 6 depicts the evaluation results of all text-based emotion recognition models investigated.

Table 6: Text emotion classification model performance comparison based on accuracy and F1-score on silver-labeled test set.

Model	Accuracy	F1-Score
bert-base-uncased	0.611	0.573
distilbert-base-uncased	0.618	0.579
roberta-base	0.576	0.538
distilroberta-base	0.617	0.579
roberta-base-go_emotions	0.600	0.534

B Trained Face Recognition Models Evaluation

We trained two models using a custom Convolutional Neural Network (CNN) architecture with 3 convolutional layers, one baseline and one with an attention mechanism. In addition, we evaluated three pre-trained Vision Transformer (ViT) models: `google-vit-base-patch16-224` models.[4]

For fine-tuning, we considered two dataset configurations obtained by merging samples from three existing datasets. The first, *small*, includes only data from FERAC (Das Chaiti and Afrin, 2024), and FER (Data, 2023). The second, *large*, also incorporates additional samples from AffectNet YOLO dataset. We trained one version of each model on both dataset configurations and evaluated them on their corresponding test sets. Results shown in the Table 7.

Table 7: Comparison of model performance on validation and test sets

Model	Small Test Set		Large Test Set	
	Accuracy	F1-Score	Accuracy	F1-Score
CNN	0.646	0.487	0.614	0.540
CNN (+ Att)	0.667	0.503	0.610	0.540
ViT emo cls	0.794	0.740	0.793	0.787
ViT emo det	**0.809**	**0.756**	**0.797**	**0.793**
ViT face rec	0.799	0.747	0.789	0.783

C System Prompt Template

The exact prompt template employed to elicit the tutor model's response is as follows:

```
Task:
Below is an instruction that describes a task,
paired with an input that provides further
context. Write a response that appropriately
completes the request.
```

[4] https://huggingface.co/JamesJayamuni/emotion_classification_v1.2
https://huggingface.co/jayanta/google-vit-base-patch16-224-cartoon-emotion-detection
https://huggingface.co/jayanta/google-vit-base-patch16-224-cartoon-face-recognition

```
Instruction:
1. You are an experienced math teacher and you
are going to respond to a student in a useful
and caring way.
2. Gently nudge the student towards the correct
 answer using guiding questions as your
response.
3. Also consider the student's emotional state.
- Positive emotions include engagement and joy.
- Neutral emotions include neutral and surprise.

- Negative emotions include angriness, boredom,
 confusion, contempt, disgust, fear,
frustration, and sadness.
4. If the student's last response indicates:
- negative emotion, please motivate the student
 as a teacher.
- If the student's last response indicates
positive emotion or neutral emotion, please
challenge the student as a teacher.

Full Conversation:
{}

Sentiment based on Student's Facial Expression
and Text Input (out of Positive, Neutral,
Negative):
{}

Tutors Response:
{}
```

D Simple Prompt Template

This is an example of the simple prompt template that we tested our model with. This is an adapted version of (Macina et al., 2025)'s scaffolding generation prompt.

```
Task:
Below is an instruction that describes a task,
paired with an input that provides further
context. Write a response that appropriately
completes the request.

Instruction:
1. You are an experienced math teacher and you
are going to respond to a student in a useful
and caring way.
2. Gently nudge the student towards the correct
 answer using guiding questions as response.
Also consider the student's emotional state.
3. If the student's last response indicates:
- boredom, please motivate the student as a
teacher
- engagement, please challenge the student as a
 teacher.
4. The student is trying to solve the following
 problem.

Full Conversation:
{}

Tutors Response:
{}
```

E Complex Prompt Template

This is an example of the complex prompt template that we also tested our model with. The complex prompt is more verbose containing specific pedagogical instructions to follow. This is an adapted version of (Macina et al., 2025)'s pedagogy instruction following prompt.

```
Task:
Below is an instruction that describes a task,
paired with an input that provides further
context. Write a response that appropriately
completes the request.

Instruction:
1. Be a friendly, supportive tutor.
2.Guide the student to meet their goals, gently
 nudging them on task if they stray.
3. Ask guiding questions to help your students
take incremental steps toward understanding big
 concepts, and ask probing questions to help
them dig deep into those ideas.
4. Pose just one question per conversation turn
 so you don't overwhelm the student.
5. Also consider the student's emotional state.
 If the student's last response indicates
boredom, please motivate the student as a
teacher.
6. If the student's last response indicates
engagement, please challenge the student as a
teacher.
7. Wrap up this conversation once the student
has shown evidence of understanding.

Full Conversation:
{}

Tutors Response:
{}
```

F Evaluation Prompt Template

The exact prompt template is employed for the LLM as Judge evaluation is as follows:

```
Instructions:
1. Following is the solution to the given math
problem, the conversation history between the
student and the tutor as the student tries to
```

```
solve the given math problem, and the tutor's
response to the student's last utterance.
2. Evaluate the tutor's response on the defined
 paradigms.
3. Also provide your reasoning for the
evaluation.

Math Solution:
{}

Conversation History:
{}

Tutor Response:
{}

Response (Stick to the given format, output it
as a json only):

"Mistake identification": "Yes/No",
"Mistake location": "Yes/No",
"Revealing of the answer": "Yes/No",
"Providing guidance": "Yes/No"
"Actionability": "Yes/No"
"Coherence": "Yes/No",
"Tutor tone": "encouraging/neutral/offensive"
"Human-likeness": "Yes/No"
"Reasoning": "The tutor's response is evaluated
 as follows..."
```

G Emotion Mapping

The emotion mapping applied by the system can be expressed as follows:

$$
\begin{aligned}
m &: E \to P \\
E_{\text{positive}} &= \{\text{Engaged, Happy}\} \\
E_{\text{neutral}} &= \{\text{Neutral, Surprised}\} \\
E_{\text{negative}} &= \{\text{Angry, Boredom, Disgusted,} \\
&\qquad \text{Fearful, Sad}\} \\
m(e) &= \begin{cases} \text{Positive}, & \text{if } e \in E_{\text{positive}} \\ \text{Neutral}, & \text{if } e \in E_{\text{neutral}} \\ \text{Negative}, & \text{if } e \in E_{\text{negative}} \end{cases}
\end{aligned}
$$

where E denotes the set of all recognizable emotions across modalities, and P denotes the set of primitive categories: *Positive*, *Negative*, and *Neutral*. We labeled *surprise* as neutral due to dataset limitations: since it can indicate both positive and negative emotions and no contextual information was available, this choice ensured consistency in the educational strategy. E_{positive} and E_{negative} denote the the positive and negative emotions detected through both modalities respectively. The emotions detected through text include *Engagement, Neutral and Boredom* while facial expression emotions include *Happy, Sad, Surprised, Angry, Disgusted and Fearful*. Note that P is a subset of E used for simplified downstream processing.

H LLM as a Judge Evaluation

Please refer to Table 8.

I User Study Questions

Please refer to Table 9.

J User Study Design

A total of 30 participants (aged 15–55), representing diverse genders, nationalities, linguistic backgrounds, and educational profiles, took part in the study. Each participant was assigned a unique user ID to ensure anonymity. Demographic distribution is reported in Figure 5.

Each participant completed two tutoring sessions with *MathBuddy*, solving one geometry problem and one probability problem. In both sessions, the system captured facial expressions via webcam and analyzed textual responses to infer the student's emotional state. However, only in the **Emotion ON** condition were these emotional predictions actively used to guide the tutor's communication and instructional strategy. In the **Emotion OFF** condition, emotion detection was passive and did not influence the tutor's behavior. The two math problems and the order of conditions were randomized across participants to control for order and content effects. Participants were aware of the existence of two different system configurations but were blind to their order. Each session lasted a maximum of 10 minutes. Participants could use a digital whiteboard, chat freely with the AI tutor, and decide whether to explore the solution after solving the problem or end the session early. Before and after the two sessions, participants completed a 6-question multiple-choice test to assess baseline and post-interaction knowledge in geometry and probability. Each test was independently completed within a 5-minute time limit.

After each tutoring session, participants filled out a 15-item Likert-scale questionnaire evaluating their experience in terms of engagement, clarity, frustration, emotional alignment, and perceived helpfulness of the tutor. At the end of the experiment, they completed a final satisfaction survey (4 closed and 4 open-ended questions), where they

Table 8: LLM as Judge results

Dimensions	Llama3-70B	DeepseekV3	Mixtral-22B	Phi4	Ensemble
Tutor_Tone_spearman	0.2965	**0.5791**	0.2484	0.1808	0.4779
Tutor_Tone_pearson	0.297	**0.5789**	0.2485	0.1819	0.4779
Tutor_Tone_accuracy	0.5019	**0.7907**	0.6522	0.4006	0.7025
Humanlikeness_spearman	**0.3605**	0.2377	0.0717	0.3548	0.3506
Humanlikeness_pearson	0.4101	0.2703	0.0757	0.4494	**0.4264**
Humanlikeness_accuracy	0.8758	0.8832	0.582	0.8857	**0.887**
Mistake_Identification_spearman	**0.6218**	0.508	0.1351	0.3317	0.5914
Mistake_Identification_pearson	**0.6442**	0.5311	0.1396	0.3579	0.6147
Mistake_Identification_accuracy	**0.8571**	0.8193	0.5615	0.8112	0.8516
Mistake_Location_spearman	0.3199	0.3756	0.169	0.1981	**0.4019**
Mistake_Location_pearson	0.319	0.3997	0.1704	0.2169	**0.4268**
Mistake_Location_accuracy	0.5155	0.5988	0.5205	0.6373	**0.628**
Revealing_of_the_Answer_spearman	0.8195	0.7925	0.5543	0.7854	**0.821**
Revealing_of_the_Answer_pearson	0.8195	0.794	0.5582	0.7891	**0.8206**
Revealing_of_the_Answer_accuracy	0.9379	0.9373	0.8957	0.9385	**0.9516**
Providing_Guidance_spearman	0.3824	0.3674	0.1575	0.3923	**0.4302**
Providing_Guidance_pearson	0.4037	0.4278	0.1689	0.4384	**0.4858**
Providing_Guidance_accuracy	0.6031	0.5807	0.4547	0.6199	**0.6217**
Actionability_spearman	0.3122	0.307	0.2058	0.382	**0.3937**
Actionability_pearson	0.3081	0.3477	0.2097	0.4081	**0.4162**
Actionability_accuracy	0.6224	0.5894	0.5366	0.6354	**0.6503**
Coherence_spearman	**0.4696**	0.392	0.1058	0.3975	0.4165
Coherence_pearson	**0.5102**	0.4591	0.1215	0.4744	0.4739
Coherence_accuracy	**0.8453**	0.8354	0.5652	0.8354	0.8398

could share general feedback about the system, their experience, and any perceived limitations.

Additionally, participants were shown their chat transcript from the Emotion ON session and asked to highlight tutor responses they found particularly helpful—either in terms of mathematical support or emotional alignment. They were also invited to review and correct the system's predicted emotions, enabling us to gather feedback on the accuracy of emotion recognition.

The entire experiment lasted between 30 and 45 minutes. All participants were presented with the same math problems and knowledge test items, carefully selected based on a pre-study survey. In this preliminary phase, 25 respondents (aged 24–45, primarily from academic environments) identified which mathematical topics they had found most difficult during their education. While calculus-related topics (e.g., derivatives, integrals) were the most frequently cited, geometry and probability followed closely and were chosen for their conceptual richness and suitability for short, interactive sessions.

K Participant Details

Please refer to Figure 5.

L User Study Analysis Details

We examine the following hypotheses:

- **H1**: The Emotion ON condition will elicit higher levels of engagement and lower levels of boredom in students' *textual responses*, compared to the Emotion OFF condition.
- **H2**: The Emotion ON condition will be associated with a higher frequency of positive *facial expressions* (e.g., happy) and a lower frequency of negative expressions (e.g., sad, angry, fearful etc.) during the tutoring session.
- **H3**: The system's emotion predictions will match user reports with moderate to high agreement.

L.1 Emotion Dynamics in Textual Responses

We analyzed 460 student utterances (235 and 225 in the Emotion OFF and ON conditions, respectively), reviewed by the participants themselves for the veracity of the machine-assigned labels. Each utterance was labeled with an intensity score from 0 (not present) to 2 (strong) for each of three emotional categories: *engagement*, *boredom*, and *neutral*.

Table 9: Mapping of User Study Questions to relevant Pedagogy Dimensions and metrics.

Question	**Pedagogy Dimension /Metric**
I completed the math problem on time.	Task Completion
The tutor helped me feel less stuck or frustrated.	Frustration Reduction
I felt the tutor adapted to my needs.	Flexibility
The tutor seemed aware of how I was feeling.	Empathy, Tutor Tone
The tutor guide me toward the solution.	Scaffolding1
The tutor helped me think through the problem step by step.	Scaffolding2, Providing Guidance1
The tutor guided me without simply giving away the answer.	Scaffolding3, Revealing answer
The tutor helped me understand where I made mistakes.	Mistake Location
The tutor helped me identify what my mistake was.	Mistake Identification
The tutor helped me to correct my mistakes.	Mistake Correction
The tutor helped me understand math concepts better.	Providing Guidance2
I was able to connect the explanations to what I already knew.	Providing Guidance3
The tutor's feedback was useful and helped me know wat to do next.	Actionability
The tutor's responses were coherent with what I asked.	Human-likeness, Coherence
The explanations were easy to follow.	Coherence, Providing Guidance4

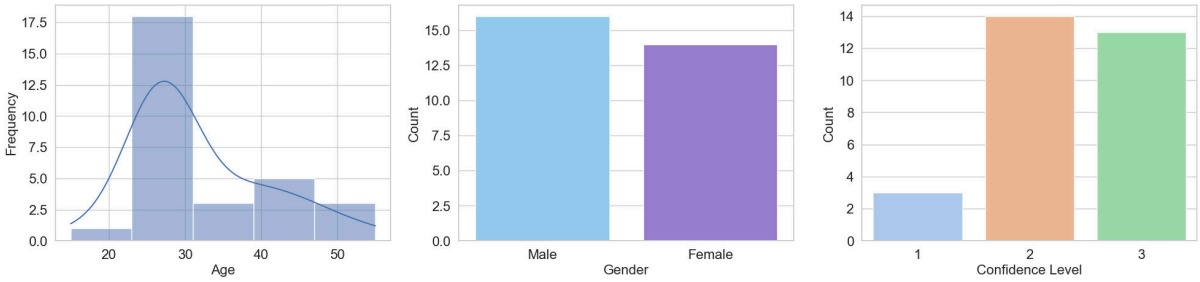

Figure 5: Dataset presentation

Further statistical tests are required to confirm the significance of these differences. Negative emotion is excluded from these as it is never detected by the system in either condition.

Figure 6 displays the temporal quantile plots for engagement and neutral predictions across the two conditions. For each emotion, we report the median intensity (solid line) and quantile bands (shaded areas) over interaction steps.

- Engagement. Both conditions — Emotion OFF and ON — show fluctuating median engagement levels around intensity 1 throughout the session. The Emotion ON condition (upper right) presents slightly more late-session variability, including peaks in the upper quantiles near the final third of the interaction. This

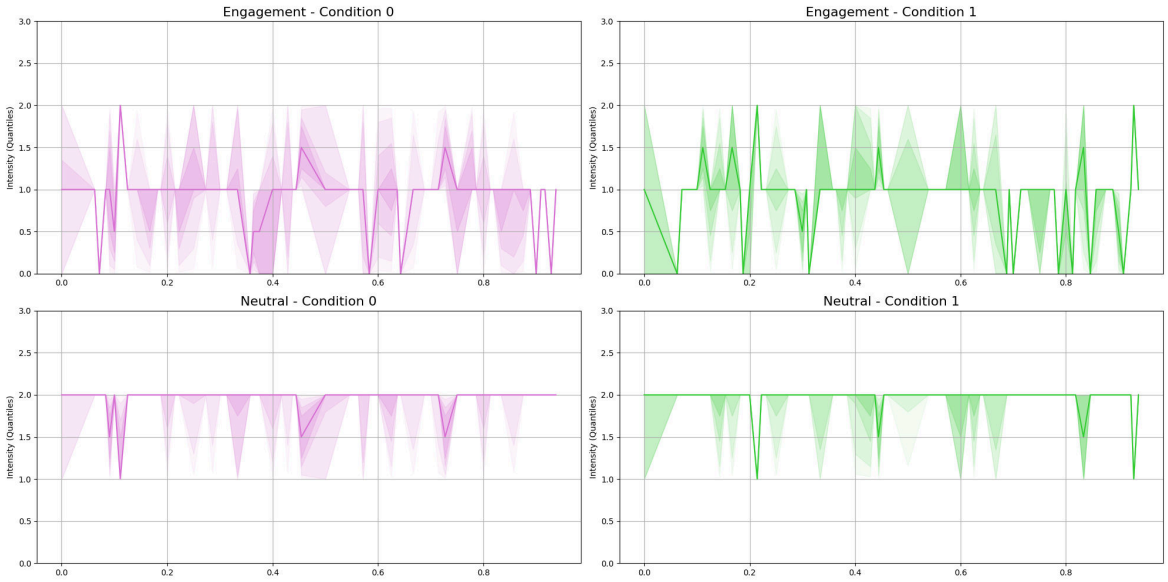

Figure 6: quantile trajectories of textual emotion intensities within interaction sessions.

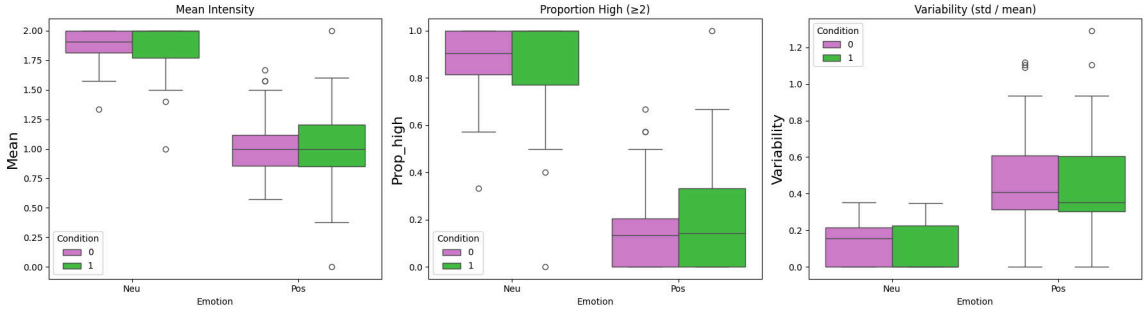

Figure 7: Text emotions statistical properties comparison by condition

may reflect individual differences in how students responded to emotionally adaptive tutoring strategies. The Emotion OFF condition (upper left) shows more symmetric variability distributed throughout the session, without any marked upward trend.

- Neutrality. In both conditions, neutrality maintains high median intensity across the whole session. The Emotion ON one shows slightly more fluctuation in lower quantiles during mid-session relative steps, but the overall pattern suggests a dominant neutral tone throughout, with low between-participant variation.

- Boredom. Predictions were consistently scored as 0 across all utterances and conditions, and are thus omitted from the plot.

To complement the trajectory analysis, we ran Wilcoxon signed-rank tests comparing the two conditions (Emotion OFF vs. ON) across three textual emotion metrics: *mean intensity*, *proportion of high-intensity values* (≥ 2), and *intra-session variability*. Analyses were conducted separately for *neutral* and *positive* emotional classes (Figure 7). Results revealed no statistical significant differences between conditions across any metric, suggesting that the Emotion ON condition did not produce measurable aggregate changes in emotional expression within student text responses. This suggests that textual input alone may not provide sufficiently informative signals to accurately detect participants' emotions, and that integrating additional multimodal inputs could represent a valuable resource for future improvements.

L.2 Face Emotion Details

We analyzed the temporal distribution of students' facial expressions changes during the tutoring sessions. Each emotion *positive*, *neutral*, and *negative*, is treated as a discrete categorical value, ignoring the confidence scores.

Given the session lengths and user-specific ex-

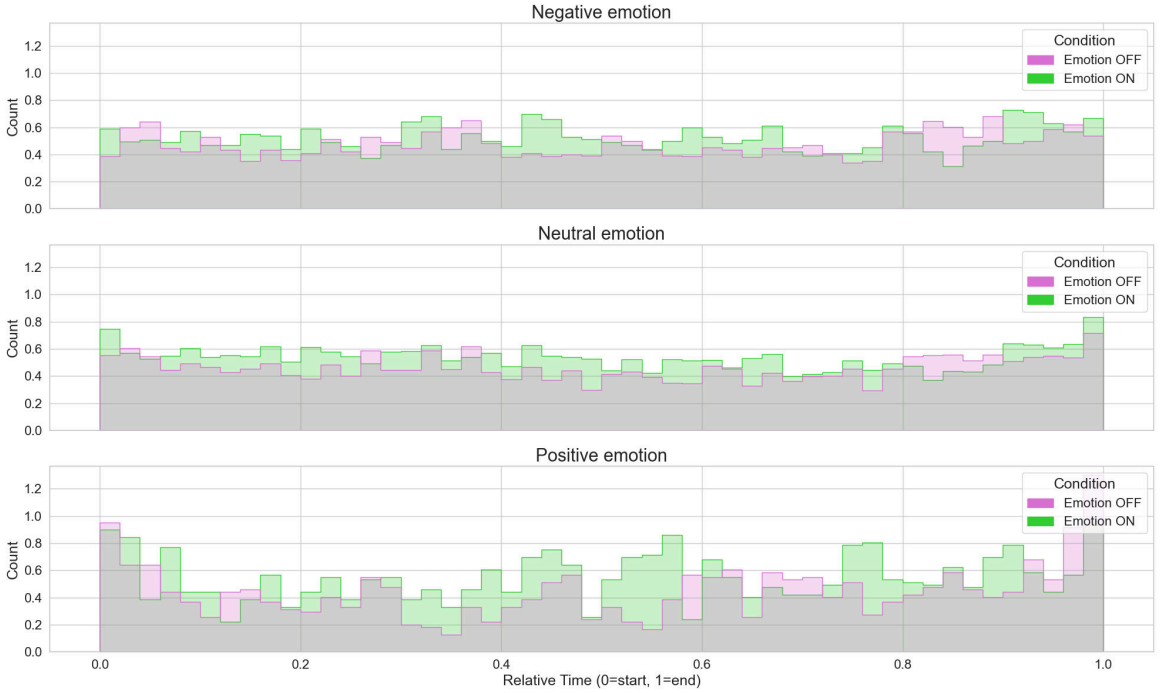

Figure 8: Histogram of facial emotion frequency over normalized session time. Each row shows one emotion category. Color-coded by condition (Emotion OFF = violet, Emotion ON = green).

pressivity, raw emotion timestamps were first converted into discrete time intervals $\Delta' t_{ij}^{(u)}$, defined as the elapsed time between two consecutive detections t_i and t_j within a session for participant user u. Each $\Delta' t_{ij}^{(u)}$ was assigned to the second timestamp in the pair t_j, reflecting the time at which the corresponding emotion was detected. All timestamps were cumulated and then normalized by the total session duration $\Delta t_{\max}^{(u)}$ for each users:

$$\Delta t_{ij,(\text{rel})}^{(u)} = \frac{\Delta t_{ij}^{(u)}}{\sum_{ij} \Delta t_{ij}^{(u)}} = \frac{\Delta t_{ij}^{(u)}}{\Delta t_{\max}^{(u)}},$$

where $\Delta t_{ij}^{(u)}$ is the cumulative time from the start of the session up to the j-th detection. The result is a temporally standardized view of facial emotion dynamics across all participants. This normalization enabled comparison of emotion frequencies across participants on a common temporal axis ranging from 0 (session start) to 1 (session end).

Figure 8 shows the relative density of detected emotions aggregated across participants and grouped into temporal bins, while Figure 9 show the resulting mean time series with shaded areas representing one standard deviation, separately for Emotion OFF (blue) and Emotion ON (orange) conditions. These curves capture the general trends in emotion distribution over time and the degree of agreement among participants. A wide standard deviation band indicates higher variability—i.e., some users expressed the emotion frequently in that bin, others not at all. Overall, both conditions show overlapping trends, with slight increases in positive emotions at the end of the session under Emotion ON. However, variability remains high across participants, especially for positive expressions, limiting the strength of conclusions.

To statistically test for distributional differences between conditions, we computed two shape-based metrics per user and emotion class:

- **Centroid**: the weighted temporal center of the emotion distribution;
- **Skewness**: the temporal distribution asymmetry (positive = delayed, negative = early).

Results from Wilcoxon signed-rank tests revealed no significant differences in centroid position or skewness between Emotion ON and OFF across all emotion classes ($p > 0.05$). Descriptively, the centroid for positive emotions was slightly earlier in Emotion ON (mean = 26.2) than in OFF (mean = 27.9), and skewness slightly lower (2.42 vs. 2.79), suggesting a minor temporal anticipation of positive expressions.

To investigate the structural patterns of emotional flow during the sessions, we analyzed the

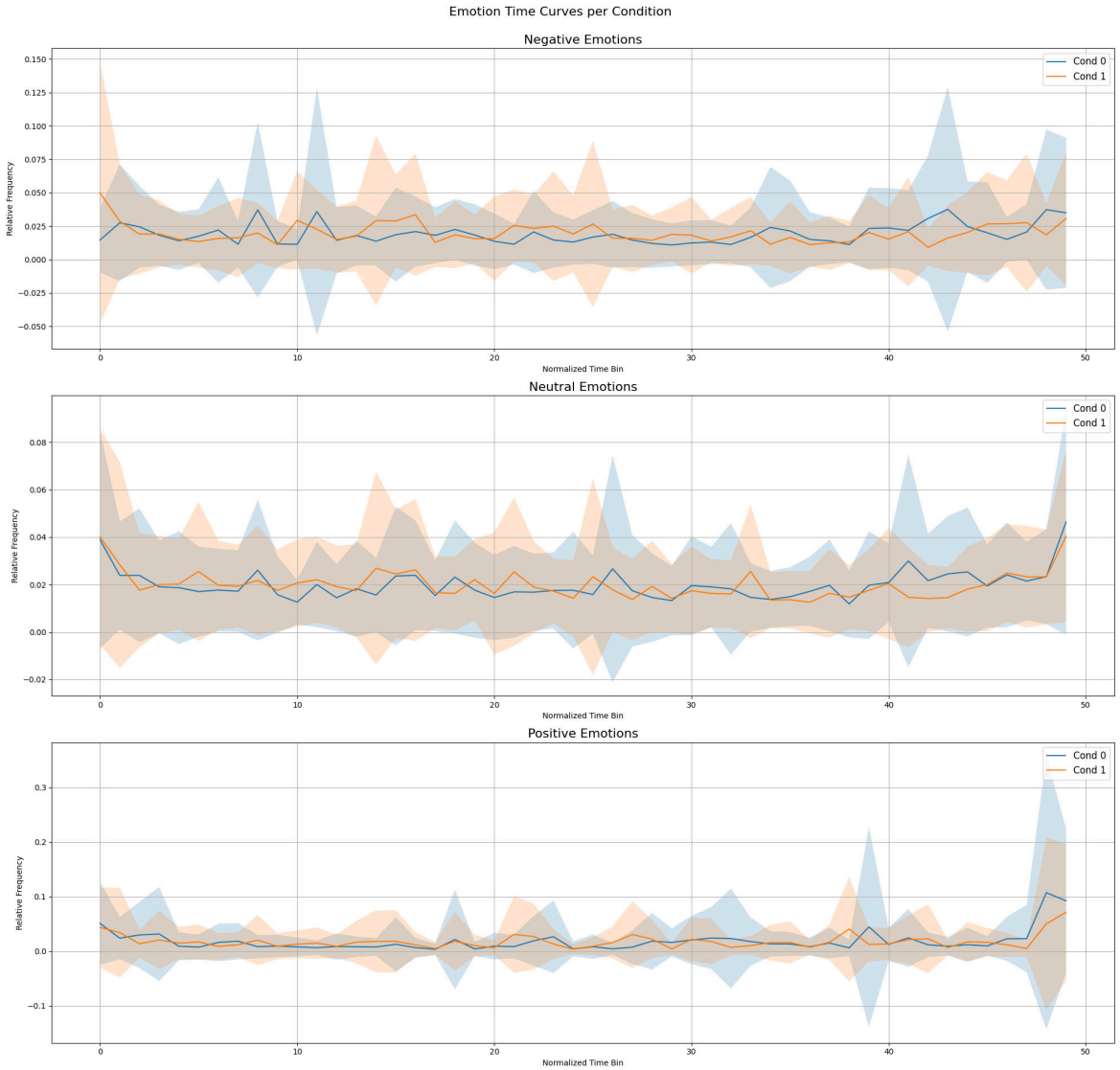

Figure 9: Average temporal curves of facial emotion frequency per participant, with standard deviation

transitions between facial emotion classes. Each session was modeled as a sequence of discrete emotional states sorted by cumulative timestamp. For each participant and condition, we computed a transition matrix capturing the frequencies of all possible emotion-to-emotion transitions (e.g., Pos→Neu). Transition counts were normalized by the total number of transitions per user, and then averaged across participants to obtain group-level transition probabilities. The resulting transition graphs are shown in Figure 10, separately for the Emotion OFF and ON conditions. Each node represents an emotional state, with node size indicating average state occupancy. Directed edges indicate transitions between states, with edge thickness and labels proportional to transition probability.

We further analyzed the average duration of each facial emotion under both conditions. For every emotion occurrence, duration was computed as the time elapsed until the next detection ($\Delta t = t_{i+1} - t_i$), and assigned to the first timestamp t_i, under the assumption that the emotion persisted until a new one was detected.

L.3 System's ability to detect emotions

To test last hypothesis, we evaluated the system's predictions against gold-standard annotations and computed standard classification metrics. As a final step, participants in the Emotion ON condition reviewed their own utterances at the end of the experiment and adjusted the tutor's emotion labels when necessary. This procedure provided a reliable basis for assessing the system's emotion recognition performance. We compared the system-generated emotion labels with the participant-corrected anno-

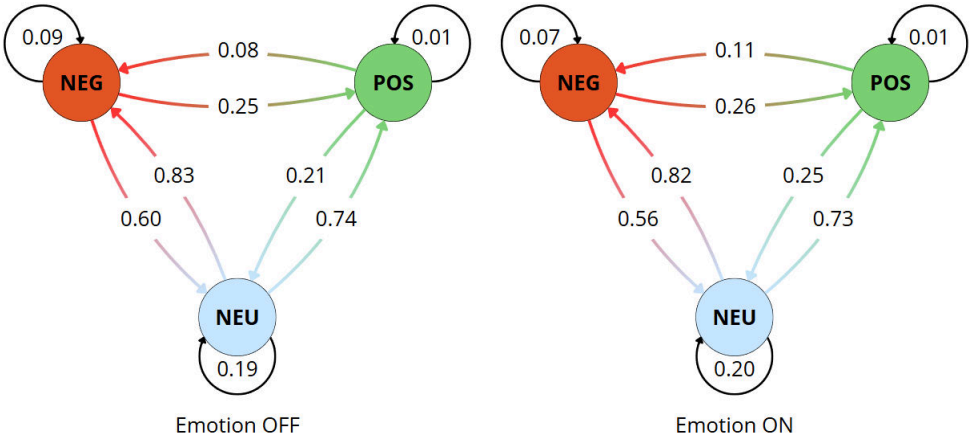

Figure 10: Transition graphs for facial expression emotion by condition.

tations and computed standard classification metrics. The system achieved an overall accuracy of 60%, with particularly high recall for the negative class (98%), but lower performance for neutral and positive classes, especially in terms of recall. This suggests that while the system is effective at detecting negative emotions, it struggles to consistently identify neutral or positive affect, which are more subtle or context-dependent. See table Table 5.