Towards Multimodal Affective Intelligence in Educational AI: Facial Expression Recognition for Large Language Models

Laura Fleig
Department of Cognitive Science
University of California, San Diego

Primary Faculty Advisor: Dr. Virginia de Sa
Secondary Faculty Advisor: Dr. Sean Trott
Graduate Student Advisor: Shuangquan Feng

Undergraduate Honors Thesis

June 2025

Abstract

As large language models (LLMs) increasingly power educational tools, their inability to perceive nonverbal signals – such as facial expressions – limits their effectiveness as empathetic, adaptive tutors. We introduce FEA-LLaVA (Facial Expression-Aware Large Language and Vision Assistant), a multimodal model that integrates compact facial expression embeddings, derived from Action Unit (AU) estimates, into the language generation process. By extending the LLaVA architecture, FEA-LLaVA enables tutoring responses to be conditioned on both textual input and student facial expressions, without relying on raw video data. We pretrain the model on 10,000 synthetic AU-description pairs and fine-tune it on 20,000 five-turn simulated tutoring conversations where students exhibit a range of facial cues. Evaluation results, judged by GPT-4.1, demonstrate that FEA-LLaVA significantly improves tutoring response quality, especially when students are verbally silent, suggesting that facial expression embeddings improve affective alignment and instructional clarity. This work highlights the potential of integrating facial expression recognition into LLMs to create more human-aware, emotionally intelligent educational AI systems.[1]

[1] An adaptation of this work has been previously submitted to a peer-reviewed conference.

Contents

1 Introduction

As large language models (LLMs) become increasingly integrated into educational tools and virtual tutors, there is growing interest in making these systems more emotionally intelligent. Other than the words being spoken or typed, LLMs often remain blind to the affective signals that shape real-world learning interactions – particularly facial expressions. In classrooms, human teachers constantly interpret students' facial expressions as cues for confusion, frustration, or engagement. These nonverbal signals help teachers decide when to slow down, when to elaborate, and when to offer encouragement. To make AI tutors more effective and human-aligned, they must also be able to read and respond to such cues. Consider a student interacting with an AI tutor who types "Okay, got it" despite frowning strongly. If the AI could recognize that the student's facial expression contradicts their textual input, it would enable the AI to respond more effectively.

In this work, we introduce FEA-LLaVA (Facial Expression-Aware Large Language and Vision Assistant), a model that augments LLM-based tutoring with real-time facial expression understanding. Building on the LLaVA framework [24], FEA-LLaVA replaces image inputs with compact facial expression embeddings derived from Action Unit (AU) estimates, capturing information without requiring raw video. The model conditions its responses on both student text and facial behavior, allowing it to generate more contextually sensitive and emotionally aligned feedback.

We demonstrate that this integration leads to improved tutoring responses, particularly when students are less verbally expressive. Through both architectural innovations and synthetic training pipelines, FEA-LLaVA takes a step toward affect-aware educational AI: systems that are both intelligent and responsive to the subtle dynamics of how humans learn.

2 Related Research

2.1 Affective Intelligence in Education

AI systems increasingly support learning through intelligent tutors, automated grading tools, and adaptive learning platforms. While these systems often excel at delivering content and guiding problem-solving, they tend to overlook learners' emotional states – factors that play a critical role in engagement, motivation, and academic performance [43]. As a result, affective intelligence – the ability of an AI system to detect, interpret, and respond to human emotions – has become an essential component of next-generation educational technologies. Recognizing affective cues such

as frustration or confusion allows systems to deliver better-timed and more supportive feedback, leading to improved learning outcomes [9].

Facial Expression Recognition (FER) has emerged as a key modality for implementing affective intelligence in educational contexts. FER systems aim to infer emotional states such as boredom, confusion, and frustration from students' facial expressions. These states are particularly important in educational settings, as they can signal when a student is disengaged or in need of additional support. For example, Suddul et al. [37] describe FER-enabled tutoring systems that adapt instructional strategies based on real-time facial cues, reporting improved engagement and FER accuracy rates around 62%. Despite these promising results, many such systems are limited to unimodal affect detection and do not incorporate more advanced forms of interaction, such as natural language dialogue.

In parallel, large language models (LLMs) such as GPT-4 [1] have transformed how tutoring systems can interact with students. LLMs offer powerful capabilities for personalized, conversation-based instruction. Recent work by Park et al. [31] and Chowdhury et al. [30] demonstrates that LLMs can enhance adaptive learning through techniques such as prompt engineering and user modeling. However, most LLM-driven educational tools remain blind to nonverbal signals, failing to incorporate students' emotional states into their responses.

Driven by rapid advancements in real-time FER and LLM technologies, there is growing interest in developing adaptive, multimodal tutoring systems that integrate both facial and textual inputs. These systems aim to create personalized, emotionally responsive learning environments that can dynamically adjust both content and delivery style in response to a student's cognitive and emotional signals. Although prior work has shown the pedagogical benefits of emotion-aware AI – such as improved engagement, reduced frustration, and stronger learning outcomes [7, 8] – comprehensive frameworks that fuse FER with LLMs in real-time are still rare.

This gap underscores the need for architectures that combine the emotional insight of FER with the interactive power of LLMs to produce truly adaptive, affect-sensitive educational agents. Since facial expressions remain one of the richest and most accessible sources of affective information in real time, we now turn to a deeper exploration of Facial Expression Recognition and its foundational role in building emotion-aware systems.

2.2 Facial Expression Recognition

Facial Expression Recognition (FER) has rapidly advanced in recent years due to improvements in deep learning and the availability of large annotated datasets [20]. FER systems play a central role in affective computing, enabling machines to infer emotional and cognitive states from visual facial cues. In educational contexts, FER is particularly valuable for detecting hard-to-measure affective states such as boredom, confusion, or frustration, which can significantly impact student engagement and learning outcomes.

Theories of Emotion and Facial Expression Two dominant theoretical frameworks guide FER research [38]. The categorical theory posits a set of universal basic emotions – such as happiness, anger, or sadness – that are expressed similarly across cultures [11, 10]. These primary emotions can combine to form more complex affective states [32]. In contrast, the dimensional theory models emotions along continuous axes such as valence and arousal, offering a more nuanced, gradient-based view of emotional experience [34, 35]. Both frameworks have informed the design of FER systems and datasets, although most machine learning models currently adopt the categorical approach.

Facial Action Coding System (FACS) A key development in FER methodology is the Facial Action Coding System (FACS), originally developed by Carl-Herman Hjortsjö and later formalized by Paul Ekman and Wallace Friesen [12]. Rather than labeling emotions directly, FACS decomposes facial expressions into anatomically defined Action Units (AUs), each corresponding to a specific facial muscle movement (e.g., brow furrow, lip corner pull). This allows for objective, fine-grained analysis of facial behavior without assuming emotional intent.

FACS offers several advantages: it enables the description of blended or ambiguous emotional states, supports cross-context interpretation, and provides a composable framework for describing facial expressions. These properties make it especially suitable for domains like education, where student expressions often fall outside of traditional emotion categories but still convey important affective information (e.g., concentration, uncertainty, determination) [42, 3].

FER Output Modalities FER systems typically output one of two types of information [20]:

- Discrete emotional or cognitive state labels, such as happiness, sadness, or arousal levels. These outputs are supported by datasets such as AffectNet [29], FERPlus [4, 15], RAF-DB

[21], Aff-Wild [44], and Aff-Wild2 [19].

- Action Unit (AU) activations, as defined in FACS. AU-based outputs provide a more granular, anatomically grounded representation of facial movements.

While emotion labels are intuitive and user-friendly, AU-based representations offer greater flexibility and generalizability. AUs describe how a face moves rather than what it expresses, enabling nuanced inference of emotional state in complex or ambiguous settings – such as classrooms or tutoring environments – where expressions may not align cleanly with basic emotional categories.

Modern AU-Based FER Systems Given the interpretability and generality of AUs, modern FER systems increasingly adopt AU-based representations. Deep learning models trained on large, labeled datasets have significantly advanced AU detection capabilities. Key datasets for AU-based FER include BP4D [45], DISFA and DISFA+ [27, 28, 26], and EmotioNet [13].

These datasets enable the training of deep models that can detect subtle and overlapping facial muscle movements across varied individuals and lighting conditions. AU detection and estimation has attracted increasing interest [16, 36, 47, 46, 6, 22, 23, 41, 17, 18, 48, 25, 39, 40, 14].

By accurately modeling facial movement rather than simply assigning emotion labels, AU-based FER provides a robust foundation for multimodal systems – especially in educational settings – where understanding subtle, transient, or ambiguous affective states is critical for providing timely, effective support.

2.3 Multimodal Vision-Language Models

Recent advances in multimodal vision-language models (VLMs; e.g., [33] and [2]) have enabled LLMs to process visual input such as images, in addition to text. VLMs integrate visual understanding with language generation, allowing AI systems to perform tasks like image captioning, visual question answering, and instruction following. These capabilities mirror how humans learn: by combining information from multiple sensory modalities to build a richer understanding of the world.

In VLMs, images and text must be reconciled despite their very different structures – images are two-dimensional arrays of pixel values, while text is linear. To bridge this gap, most VLMs adopt a three-part architecture: a *vision encoder*, a *multimodal projector*, and a *language model*. A simplified version of this architecture is shown in Figure 1.

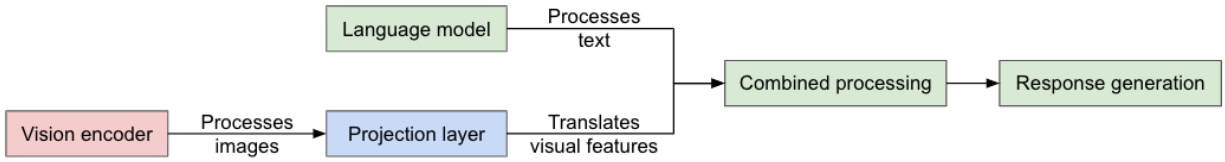


Figure 1: A (simplified) depiction of a general vision-language model architecture.

- The **vision encoder** is commonly a neural network pretrained on visual tasks (e.g., CLIP [33]) that converts an input image into a high-dimensional embedding that captures visual semantics. This encoder is often frozen, meaning its weights are not updated during training, in order to preserve the robust representations it has already learned.

- The **projection layer** serves as a translator between modalities. It maps the visual embeddings from the encoder into the same latent space used by the language model. This component is often a linear layer or multi-layer perceptron (MLP).

- Finally, the **language model** (e.g., Vicuna [5] or GPT-like models) receives both the user's textual input and the projected visual embedding as input. It generates responses conditioned on this joint context.

An accessible way to conceptualize this architecture is as a "translation office" with three specialist workers: a **visual analyst** (the vision encoder) who takes in images and produces notes, a **bilingual translator** (the projector) who converts these notes into a form the **communication expert** (the language model) can understand and respond to. This division of labor allows VLMs to ground language in visual understanding, while preserving the language model's general-purpose reasoning abilities.

However, while many VLMs have focused on object-level understanding or scene reasoning, they cannot typically fully interpret social visual cues such as facial expressions or body language. This represents a key limitation for applications in education, healthcare, and human-computer interaction, where emotionally intelligent responses are often critical to success. Explicitly incorporating facial expression information into VLMs offers a promising direction for enhancing the emotional sensitivity of AI systems.

Among recent VLMs, LLaVA (Large Language and Vision Assistant) stands out as a general-purpose, open-source VLM that aligns a pretrained vision encoder (CLIP) with a language model (Vicuna) via instruction tuning [24]. LLaVA has been widely adopted due to its strong performance

and extensibility. However, despite its capabilities, the original LLaVA architecture is limited to static image inputs and lacks detailed awareness of nonverbal social signals like facial expressions, making it less suitable for applications requiring emotion awareness, such as empathetic tutoring.

In this work, we extend the LLaVA architecture to incorporate facial expression embeddings derived from AU estimates, replacing image inputs entirely. This modification enables our system, FEA-LLaVA, to reason not just about what a student says, but also how they appear to feel, making it better suited to educational applications.

3 Methods

We introduce FEA-LLaVA (Facial Expression Aware Large Language and Vision Assistant), a multimodal model that conditions language generation on both user text and facial expression cues. Building on LLaVA, FEA-LLaVA replaces traditional input with affective facial embeddings.

3.1 Model Architecture

Our model architecture is adapted from the LLaVA framework. However, rather than relying on visual features from static images, we replace image embeddings with facial expression embeddings derived from user video recordings during user-agent interactions. This enables the model to incorporate visual affective cues into its language generation process.

As with LLaVA, our system maintains a three-part architecture (see Figure 2): a vision encoder, a projector, and a language model. We keep the same language model backbone (Vicuna-13b-v1.5) but substitute the vision encoder and projector with components tailored for facial expression recognition.

Our vision encoder is an AU estimation model based on the IR50 architecture, pretrained on the Glink360k dataset and fine-tuned on DISFA and DISFA+. Input video frames are preprocessed using face detection, face alignment, histogram equalization, and linear mapping. The original estimation model outputs estimates for 12 AUs, but we use the 8 most informative ones to construct a 7-dimensional vector representation (AU1 and AU2 are highly correlated and thus combined using a max operation), resulting in a maxAU vector that captures peak facial activation across the entire video. While we currently use only a static representation (i.e., one maxAU vector per video), we imagine expanding this to a full AU matrix that can capture temporal dynamics.

This 7-dimensional maxAU vector is passed through a multi-layer perceptron (MLP) projector

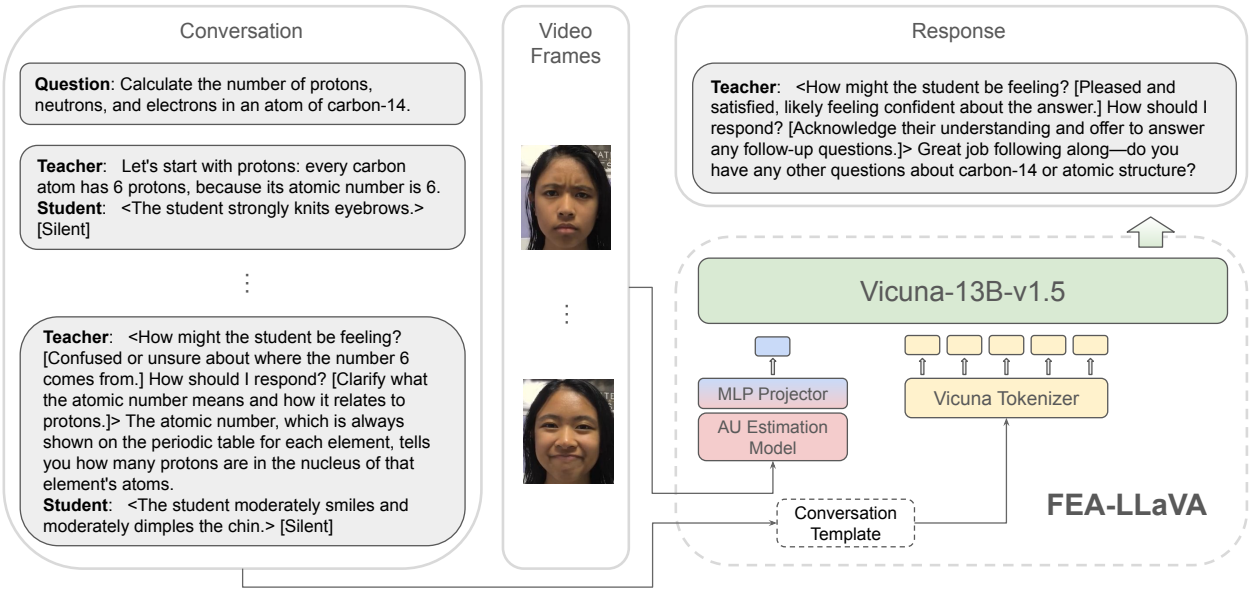


Figure 2: The architecture of FEA-LLaVA with sample training data.

with a 512-unit hidden layer. This transforms the vector into the same embedding space used by the language model.

Finally, the projected facial embedding is prepended to the user's input text tokens. This allows the model to generate responses conditioned jointly on the user's textual input and their facial expressions. Telling the system not only what a user says but also how they appear to be feeling is a step toward more emotionally intelligent language agents.

3.2 Training

To train our model, we followed the two-stage training approach used by LLaVA and many other models: pretraining and fine-tuning. Both stages were trained on 8 NVIDIA RTX A6000 GPUs and dual AMD EPYC 7302 CPUs, supported by 512GB of 8-channel RAM across 16 cores.

3.2.1 Pretraining

The goal of the pretraining stage is to establish a semantic link between facial expression vectors and their natural language interpretations, enabling the model to associate AU vectors with meaningful textual interpretations.

To this end, we constructed a pretraining dataset of 10,000 paired examples of simulated AU vectors and their corresponding natural language descriptions, as follows. The model learns to predict the descriptions, given the input of the vector and the prompt.

Table 1: Mapping of max-pooled AU activations to natural language descriptions with intensity thresholds. Values below the first threshold are considered not activated.

AU(s)	Base Description	Slightly	Moderately	Strongly
Max(AU1, AU2)	raises eyebrows	1.5 – 2.0	2.0 – 2.8	> 2.8
AU4	knits eyebrows	1.0 – 1.6	1.6 – 2.8	> 2.8
AU5	widens eyes	0.8 – 1.5	1.5 – 2.2	> 2.2
AU9	wrinkles the nose	1.0 – 1.6	1.6 – 2.8	> 2.8
AU12	smiles	1.0 – 1.6	1.6 – 2.8	> 2.8
AU15	downturns the mouth	1.0 – 1.6	1.6 – 2.8	> 2.8
AU17	dimples the chin	1.0 – 1.6	1.6 – 2.8	> 2.8

Dataset Preparation The process consisted of three main stages: prompt creation and sampling, description generation, and AU vector construction. A subsequent stage augmented the dataset using LLM-generated paraphrases to increase linguistic diversity and prevent overfitting to template-based outputs.

Step 1: Prompt Sampling. We curated a bank of natural language and technical prompts inspired by those used in LLaVA. These prompts were randomly sampled during dataset generation. See Table 2 in the appendix for the full prompt bank.

Step 2: Description Generation. We randomly sampled 1 to 3 AUs from our fixed set specified above (AU1 and/or AU2, AU4, AU5, AU9, AU12, AU15, AU17). Each sampled AU was assigned a random intensity value in the range $[1.0, 5.0]$. To map AU values to natural language descriptions, we defined thresholds of activation for the adverbial modifiers "slightly," "moderately," and "strongly" (see Table 1). This mapping was hard-coded based on observations and demonstrations during the development phase. Multiple activated AUs were concatenated with the word "and." Descriptions were generated in two styles, depending on the prompt type. For example, activations of AU12 at 3.1 and AU4 at 1.7 would return: *natural* (e.g., "The person strongly smiles and moderately knits their eyebrows.") vs. *technical* (e.g., "The person has AU12 strongly activated and AU4 moderately activated.").

Step 3: AU Vectorization. A vectorization function was used to convert the AU dictionary into a full 7-dimensional AU vector. Each selected AU's intensity was preserved with slight random noise (within its intensity bucket), while a few inactive AUs were assigned random low-level noise (default: 2 randomly chosen AUs in the 0 to 0.99 range). This added variability to the input vector and better simulated the noise of AU estimation in the real world.

Step 4: Description Diversification with LLM. To mitigate the risk of the model overfitting to templated or repetitive descriptions, we used the GPT-3.5-turbo API to automatically

rewrite the original descriptions into more natural and stylistically diverse paraphrases (see Table 3 in the appendix for the full system prompt). We used a temperature of 0.8 to encourage variation. This was applied to the full dataset of 10000 samples, resulting in an augmented dataset of 10000 paraphrased samples. Each augmented sample retained the original AU vector and prompt but substituted the revised natural language description.

Training In pretraining, like for LLaVA, the vision encoder and language model are kept frozen, and we only trained the projector. We trained on the 10,000 pretraining examples for 50 epochs with a batch size of 256 and a learning rate of 1e-3.

3.2.2 Fine-tuning

After pretraining, fine-tuning teaches the model to understand and use facial expressions in conversation. For this, we generated a dataset of 20,000 five-turn conversations between a teacher and a student agent.

Data Preparation Since there are currently no large-scale, high-quality datasets of users interacting with AI tutors while exhibiting natural facial expressions, we constructed a synthetic dataset that pairs diverse facial expression data with educational dialogue.

For facial expression data, we drew on the Highly Diverse Facial Expressions (HDFE) dataset, developed by another group in our lab. The HDFE dataset includes recordings from 465 participants (collected with IRB approval and informed consent), each imitating a diverse range of 275 distinct facial expressions and combinations. Individual video clips are typically 2 to 5 seconds long. From each participant's recording set, we selected 20-25 of the most relevant expressions that could plausibly appear in a tutoring context. (Specifically, we identified 25 expressions most relevant to our context; see the full list in the appendix. Not all participants have videos for all expressions. We excluded participants with fewer than 20 videos from the selected list.) We processed each video using our AU estimation pipeline to generate maxAU embeddings, and then used the same AU-to-description pipeline as in pretraining to produce natural language descriptions of each expression.

To pair expressions with meaningful conversational context, we also created a synthetic question bank consisting of 3,000 educational questions spanning mathematics, physics, chemistry, biology, and computer science. Questions were organized by subject and grade level (grades 9-12, college freshman, and college sophomore), with 10 topics per subject-grade combination and 10 questions

per topic (see Table 4 in the appendix for the question generation prompt). This structure allowed us to ensure topical diversity and realistic difficulty levels across generated conversations.

We then used this material to generate synthetic five-turn conversations between a teacher agent and a student agent. Both agents were powered by separate GPT-4.1 API calls (see Tables 5 and 6 in the appendix for the student and teacher system prompts, respectively). For each conversation, we randomly sampled one question and one participant (with their corresponding set of expression descriptions). The student agent was given the list of facial expression descriptions available for that participant and was prompted to respond in a realistic, emotionally expressive manner. The student could optionally respond with text, but was instructed to remain silent most of the time unless explicitly prompted by the teacher. This design reflects our core pedagogical hypothesis: for verbally expressive students, facial expressions may contribute less new information, while for quieter students, they serve as crucial nonverbal cues for real-life teachers. Therefore, the fine-tuning setup emphasized affective inference in the presence of minimal verbal output.

Meanwhile, the teacher agent was prompted to produce only one sentence at a time and to reflect explicitly ("think out loud") on the student's inferred emotional state. This structure encouraged fine-grained reasoning based on the student's facial expression.

Through this pipeline, we generated a total of 20,000 five-turn conversations, resulting in 100,000 individual training examples. We used 80% of the questions and participants for training and reserved the remaining 20% for evaluation.

For the actual fine-tuning process, we replace the AU narrator's description with the original AU vector. The model learns to generate empathetic, accurate teacher responses based on facial expression information and occasional verbal feedback.

Training Again following LLaVA, fine-tuning updates both the projector and the language model weights, while the vision encoder is kept frozen. We fine-tuned the model on the simulated teacher-student conversation dataset for 1 epoch with a batch size of 128, a learning rate of 2e-4, and an AU projector learning rate of 2e-5.

4 Evaluation & Results

To evaluate the effectiveness of FEA-LLaVA, we use a GPT-4.1-based evaluation framework, following a growing body of work that leverages large language models as judges for natural language

generation tasks[2]. We compare our model to a baseline that uses the same underlying Vicuna language model but receives only the text prompt, without any prepended AU embedding. This allows us to isolate the contribution of facial expression cues to the model's outputs. For each test example, we present GPT-4.1 with the full conversational history leading up to a particular tutor response and ask it to compare two candidate responses: one generated with facial expression conditioning (FEA-LLaVA), and one without (baseline). The evaluator is instructed to judge these responses based on three criteria: emotional alignment, clarity and helpfulness, and overall tutoring quality. It then selects the better response (A or B) or may choose "neither" if both responses are of comparable quality. To prevent bias, response order is randomized across examples. Importantly, the tutor's "thinking out loud" reflections are excluded from what is presented to the evaluator (this is also removed from the system prompt provided to the baseline model). See Tables 7 and 8 in the appendix for the baseline and evaluator agents' system prompts.

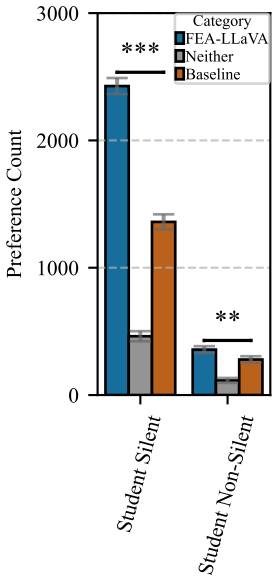
We generated 1,000 five-turn evaluation conversations, resulting in 5,000 evaluation samples.

The results are shown in Figure 3a. We split the 5,000 test instances by whether the student was silent ($n = 4,249$) or verbally expressive ($n = 751$). Bar heights represent preference counts, and error bars indicate 95% confidence intervals estimated from binomial variance. Statistical significance was assessed using two-sided binomial tests, with significance levels denoted by *, **, and *** for $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

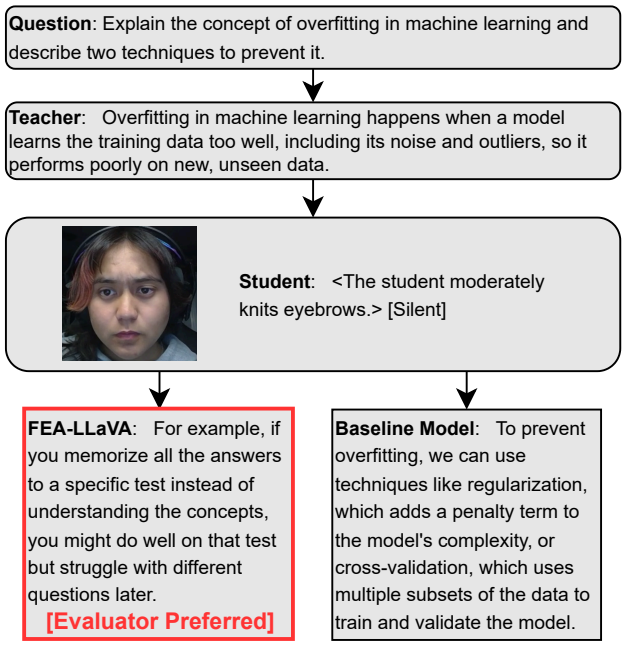
When the student was silent, FEA-LLaVA was preferred in 2,427 of 4,249 cases (57.1%), compared to 1,360 for the baseline (32.0%) and 462 ties (10.9%), a highly significant difference ($p < 0.001$). When the student was non-silent, preferences were more evenly distributed: FEA-LLaVA was preferred in 357 of 751 cases (47.5%), the baseline in 279 (37.2%), and 115 were rated as equal (15.3%), with the difference still statistically significant ($p < 0.01$). These results suggest that facial expression conditioning provides the most benefit when verbal cues are sparse, supporting our hypothesis that nonverbal information is especially valuable for interpreting quieter students.

Figure 3b illustrates an evaluation example in which the evaluator prefers our model's response. FEA-LLaVA infers confusion from their knitted eyebrows and silence and responds by providing an example to clarify. The baseline model, by contrast, explains additional technical details.

[2]See, for example, https://lmarena.ai/.



(a) Summary of evaluator preferences



(b) An example evaluation instance

Figure 3: GPT-4.1-empowered evaluations for responses made by FEA-LLaVA vs. the baseline model (language-only version of Vicuna-13B-v1.5). (a) Summary of the evaluations separated by student being silent vs. being non-silent. Bar heights represent the number of preferences for each option. (b) An example of an evaluation, where the evaluator prefers our FEA-LLaVA over the baseline model.

5 Discussion

While our results demonstrate that conditioning on facial expression vectors can improve response quality, particularly when verbal cues are sparse, several important assumptions, limitations, and future directions warrant examination.

5.1 Interpretability vs. Direct Embedding

A key design choice is the use of direct AU vector embeddings rather than natural language descriptions of facial expressions. While a simpler approach might involve using our AU-to-description model and directly passing the resulting description to the language model, this introduces substantial information loss. Hardcoded mappings from AU values to language force discretization of intensity and eliminate fine-grained temporal or combinatorial cues. By contrast, embedding the AU vector directly allows the model to learn which features matter most for interpreting affective states – potentially discovering patterns too subtle or complex to encode manually.

This approach mirrors the strategy used in LLaVA, where image features are embedded and projected into the language model's token space rather than converted into textual captions. As with visual inputs, our assumption is that tokenizing AU vectors lets the model learn a more flexible, data-driven mapping from behavior to semantic meaning. This becomes even more important if we later expand to richer inputs like time-series AU matrices, where summarizing affect as a sentence would severely limit expressivity.

5.2 Limitations and Future Directions

Despite promising results, the current model has several limitations. Most significantly, it lacks temporal modeling. The use of a static maxAU vector ignores changes in expression over time – such as progression from surprise to confusion, or rapid microexpressions. While the simplification made training tractable, it also sacrifices an important dimension of affective behavior. We somewhat address this since our current setup uses quite short videos, but particularly in a longer-term setup where an AI tutor interacts with a student in real-time during a tutoring session, having the ability to analyze the student's facial expressions over time becomes increasingly important. Future work could incorporate recurrent or transformer-based modules that process AU sequences over time.

Another limitation is modality scope. Facial expressions are just one component of nonverbal communication. Head movements, blinking, posture shifts, and gaze direction all play critical roles

in signaling emotion and engagement. Integrating these cues would require expanding both the input features and the AU estimation model, but could improve model realism and responsiveness. Beyond visual nonverbal cues, integrating auditory cues such as tone of voice could also greatly increase affective understanding.

We also recognize that, while prior work supports the validity of LLM-based evaluation, further human-based testing prior to deployment would strengthen generalizability and usability.

5.3 Ethical and Practical Considerations

Real-time facial expression analysis in educational contexts raises several ethical concerns. Facial data is inherently sensitive, and its use must be governed by strict privacy protections and informed consent. Moreover, bias and equity remain central issues. Existing AU estimation models may perform unequally across different skin tones, facial structures, cultural norms of expressiveness, or neurodivergent behavior patterns. If not addressed, these biases could lead to inaccurate inferences about student states and reinforce existing disparities in educational support.

Finally, while our results demonstrate benefits in a controlled simulation, real-world deployment presents further challenges. Facial expressions in educational settings may not always align with emotional or cognitive states and may vary between students (e.g., a furrowed brow might imply confusion in one student and concentration in another). Subtle or suppressed expressions may be misinterpreted, and overreliance on affective inference could lead to inappropriate feedback. We emphasize the need for more data collection of facial expressions in educational settings. Overall, systems like FEA-LLaVA should be viewed as tools to augment – not replace – human judgment, and should undergo rigorous evaluation in authentic learning settings before being adopted in practice.

6 Conclusion

In this work, we introduce FEA-LLaVA, a facial expression-aware language and vision assistant designed to enhance AI tutoring by incorporating nonverbal affective cues. By embedding AU vectors directly into the language model's input, we enable the system to condition its responses on both what students say and how they appear to feel, especially when verbal feedback is limited. Our evaluations show that this multimodal conditioning improves the quality of tutor responses, particularly for quiet or hesitant students.

More broadly, this work underscores the inherent complexity of human teaching and interaction. As we attempt to build systems that reason, instruct, and adapt in real time, it becomes evident just how sophisticated human cognition and communication truly are. This is particularly clear in education: Teaching involves not only delivering content but also interpreting subtle cues, managing engagement, and responding with empathy and understanding. Our aim is not to replace educators, just to develop AI tutors that are more attuned to the social and emotional dynamics of learning – systems that can better support students by responding to how they learn and feel, not just what they say.

Looking ahead, the development of effective educational AI must remain grounded in a human-centered perspective. As AI systems become increasingly integrated into classrooms and learning environments, it is essential that they are designed to respect, reflect, and support the nuanced ways in which humans communicate and learn. Continued research in human-AI interaction, affective computing, and multimodal modeling will be critical to advancing this goal responsibly and effectively.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[6] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017.

[7] Sidney D'Mello and Art Graesser. Malleability of students' perceptions of an affect-sensitive tutor and its influence on learning. In *Twenty-Fifth International FLAIRS Conference*, 2012.

[8] Sidney D'mello and Art Graesser. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):1–39, 2013.

[9] Sidney D'Mello and Art Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[10] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[11] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[12] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[13] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[14] Shuangquan Feng and Virginia R de Sa. One-frame calibration with siamese network in facial action unit recognition. *arXiv preprint arXiv:2409.00240*, 2024.

[15] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[16] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021.

[17] Dimitrios Kollias. ABAW: valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022.

[18] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. ABAW: valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023.

[19] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[20] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[21] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[22] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1841–1850, 2017.

[23] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[25] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017.

[26] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended DISFA dataset: Investigating posed and spontaneous facial expressions. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8, 2016.

[27] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, and Philip Trinh. Automatic detection of non-posed facial action units. In *2012 19th IEEE International Conference on Image Processing*, pages 1817–1820. IEEE, 2012.

[28] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[29] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[30] Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15, 2024.

[31] Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2024.

[32] Robert Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1, 1980.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

[35] Harold Schlosberg. The description of facial expressions in terms of two dimensions. *Journal of experimental psychology*, 44(4):229, 1952.

[36] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 13(3):1274–1289, 2019.

[37] Geerish Suddul, Chandesh Lillmond, and Sandhya Armoogum. A smart virtual tutor with facial emotion recognition for online learning. In *2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC)*, pages 67–72. IEEE, 2022.

[38] JM Susskind, G Littlewort, MS Bartlett, J Movellan, and AK Anderson. Human and computer recognition of facial expressions of emotion. *Neuropsychologia*, 45(1):152–162, 2007.

[39] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. FERA 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE, 2015.

[40] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. FERA 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 839–847. IEEE, 2017.

[41] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017.

[42] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[43] Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009.

[44] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017.

[45] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[46] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[47] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[48] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, 36:1067–1093, 2020.

Appendix A: Pretraining

Table 2: The list of prompts for describing facial expressions.

Natural language prompts:

- "What is the person expressing?"
- "Describe the user's facial expression concisely."
- "Offer a succinct explanation of the facial expression presented"
- "Summarize the person's expression in a sentence."

Technical description prompts:

- "Which AUs are activated in this face?"
- "List the action units involved in this expression."
- "Identify the AU codes present."
- "Output the AU description of this face."

Table 3: The prompt for diversifying pretraining data. The red text indicates placeholders for the actual prompt and description.

You are a helpful assistant that rewrites facial expression descriptions. The meaning must stay the same, but the wording, sentence structure, etc. can change. Your goal is to create a diverse and natural-sounding variation of the original description.
Prompt: {prompt}
Original description: {description}
Rewrite the description 1 different way to match the tone of the prompt, while preserving the underlying meaning. Return only the new descriptions, without any labels or extra text. Don't be overly dramatic. For example, if the prompt is "What is the person expressing?" and the description is "The person strongly smiles and slightly raises their chin.", you might return "They're smiling broadly and lifting their chin just a bit." or "A wide smile is accompanied by a subtle chin raise.". But feel free to vary from these wordings/sentence structures.

Appendix B: Fine-tuning

From the 275 facial expressions included in the HDFE dataset, we selected the following 25 expressions as most relevant for our purposes:

- Brows Raised

- Inner Brow Tips Raised
- Brows Knitted
- Inner Brow Tips Raised and Knitted
- Brows Raised and Knitted
- Eyes Wide Open
- Brows Raised and Eyes Wide Open
- Inner Brow Tips Raised and Eyes Wide Open
- Brows Knitted and Eyes Wide Open
- Brows Raised and Knitted and Eyes Wide Open
- Mouth Slightly Open
- Teeth Clench (Closed-Mouth)
- Nose Wrinkled
- Nose Wrinkled and Chin Raised
- Chin Raised
- Upper Lip Raised w/ Chin Raised
- Mouth Downturned (Closed-Mouth)
- Mouth Downturned (Slightly-Open-Mouth)
- Mouth Downturned w/ Chin Raised
- Mouth Downturned w/ Upper Lip Raised and Chin Raised
- Closed-Mouth Smile with Eyes
- Closed-Mouth Smile with Eyes w/ Upper Lip Raised
- Open-Mouth Smile with Eyes
- Open-Mouth Smile with Eyes w/ Upper Lip Raised
- Closed-Mouth Smile with Eyes w/ Chin Raised

Table 4: Example prompt for generating questions. The red text indicates placeholders for the actual grade, subject, and topic used in the prompt.

You are a subject matter expert teacher. Generate 10 diverse, specific, and realistic exam-style questions for students in {grade} in the subject of {subject}, under the topic "{topic}". Only include questions that are purely text-based and would appear in a written exam. Each question should have a clearly defined answer (e.g., compute, define, solve) and should not be open-ended or require visuals. Return the 10 questions separated by the delimiter ———. Do not number each question.

Table 5: Example system prompt for the student agent in the generation of synthetic AI tutoring conversational data. The red text indicates placeholders for the actual descriptions from the expression list of the participant chosen for the specific conversation.

You are a *STUDENT* listening to a teacher explain the following question or problem: question

You are not the teacher. Your role is to listen, react, and respond only when necessary.

[IMPORTANT]: Your output must include two parts:
1. A facial expression description (exactly one from the list below)
2. The student's verbal response (which may be [Silent])

For example:
{description_example_1} I don't understand.
{description_example_2} [Silent]

[IMPORTANT]: The facial expression description must be strictly selected from the following list:
{descriptions_list}

Select a facial expression that reflects a reasonable emotional or cognitive state at that point in the conversation.

Your emotional state may change or remain consistent throughout the conversation.
For example:
- You may be confused at first but understand later.
- You may understand initially and become confused later.
- You may remain confused the whole time.
- You may always appear confident.
- You may show little or no noticeable emotion, appearing neutral or expressionless throughout.
- You may show noticeable emotion at first but become neutral or less expressive later, or start neutral and become more expressive later.

Whatever the case, the flow of facial expressions must make sense in context. Do not switch emotional states arbitrarily.

[IMPORTANT]: The student's textual response should mostly be "[Silent]", which means the student is listening passively, unsure, or waiting for the teacher to continue. Only respond with actual text if the teacher explicitly asks a question or clearly expects an answer.

However, you should respond if the teacher explicitly asks you a question and clearly expects an answer. When you do respond, it must reflect a realistic student reaction. Never explain the concept yourself.

Again, you are a student. You may be shy, confused, or hesitant. Do not under any circumstances act like the teacher or use instructional language.

Table 6: Example system prompt for the facial-expression-aware teacher agent in the generation of synthetic AI tutoring conversational data.

You are a *TEACHER* explaining a question or problem to your student. The first message in each conversation will be a student question or problem. You should begin by starting to explain its answer or solution, and then continue based on the student's facial expressions, reactions, and responses. In each round, only say **one sentence at a time** and wait for the student's reaction before continuing. Do not explain too much at once. Think of this as a back-and-forth tutoring session. [IMPORTANT: Limit your response to one sentence only to wait for the student's reaction.] [IMPORTANT: Before responding, think carefully about what the student's facial expression or reaction implies about their current emotional and cognitive state, and adjust your next sentence accordingly. NEVER ignore the student's facial expression or reaction.] [IMPORTANT]: At the beginning of each output, include a clearly separable "thinking aloud" statement in the following fixed format, enclosed in a single pair of angle brackets: <How might the student be feeling? [your interpretation]. How should I respond? [your intended approach].> Then continue with your actual one-sentence tutor response. Sometimes, the student may be silent — that is normal in a classroom setting.

Appendix C: Evaluation

Table 7: Example system prompt for the baseline teacher agent in the generation of synthetic AI tutoring conversational data.

You are a *TEACHER* explaining a question or problem to your student. The first message in each conversation will be a student question or problem. You should begin by starting to explain its answer or solution, and then continue based on the student's facial expressions, reactions, and responses. In each round, only say **one sentence at a time** and wait for the student's reaction before continuing. Do not explain too much at once. Think of this as a back-and-forth tutoring session. [IMPORTANT: Limit your response to one sentence only to wait for the student's reaction.] [IMPORTANT: Before responding, think carefully about what the student's facial expression or reaction implies about their current emotional and cognitive state, and adjust your next sentence accordingly. NEVER ignore the student's facial expression or reaction.] Sometimes, the student may be silent — that is normal in a classroom setting.

Table 8: Example system prompt for the evaluator agent.

You are an expert educational evaluator.
Your task is to compare two AI tutor responses to a student in a learning session and decide which one is better.

Your judgment should be based on the following three criteria:
1. Emotional alignment
2. Clarity and helpfulness
3. Overall tutoring quality

A good tutor does not always provide the full solution immediately. Instead, it adapts to the student's state to foster learning effectively. This includes not only obvious empathetic language, but also subtle behaviors such as slowing down or simplifying explanations when the student appears confused or uncertain.

Only choose "A" or "B" if one response is clearly better based on emotional alignment, clarity, or overall tutoring quality.

If the two responses are similarly good or similarly weak, choose:

Neither

Do not force a choice if the responses are roughly equal in quality.

Only output your final choice in this format:
A
or
B
or
Neither