**Title: Classifying Student Behavior in AI-Generated Educational Responses Using DistilBERT and TinyLlama**

**Introduction**

As AI becomes more prevalent in educational tools, the ability to distinguish between different types of student responses—such as whether a student is focused, confused, or distracted—can significantly enhance personalized learning. This project addresses the problem of behavior classification in AI-generated student responses using natural language processing (NLP). By understanding behavioral signals, educators can tailor instruction and provide timely interventions. The project centers around fine-tuning a DistilBERT classifier to predict student behavior based on textual input and output, leveraging AI-generated data from a custom-trained TinyLlama language model.

**Related Research**

Previous work in educational NLP has explored dialogue systems, student emotion detection, and answer classification. Tools like BERT and its smaller derivative DistilBERT have been widely used for text classification tasks due to their contextual understanding and computational efficiency. Additionally, large language models like GPT and LLaMA have been adapted to generate synthetic data for model training and evaluation. Our approach builds on this foundation by generating synthetic student responses using TinyLlama, a lightweight transformer model trained for instructional-style output, and then classifying them using a fine-tuned DistilBERT model.

**Proposed Methodology**

We first curated a dataset consisting of 300 teacher-student interactions across three behavior categories: Focused, Confused, and Distracted. Each record included a teacher input (a question), a TinyLlama-generated student response, and a manually labeled behavior. We formatted the input for classification as:

"Teacher: [question] Student: [response]"

The DistilBERT tokenizer was used to tokenize the text with a maximum sequence length of 128. The dataset was split into training and validation sets using an 80/20 split. Preprocessing was done using the HuggingFace datasets and transformers libraries, and the model was trained using the Trainer API. The classification model outputs logits, which are passed through a softmax function to produce confidence scores for each class.

TinyLlama played a crucial role in generating realistic, behavior-controlled student responses. Using prompt engineering, we generated 900 samples (300 per behavior). Each prompt was structured with a specified behavior tag, followed by the teacher input. TinyLlama then produced a plausible student answer in line with the prompted behavior, which was later evaluated and labeled for training.

**Learnings and Preliminary Insights**

Through model evaluation on a new set of 60 examples (20 per behavior), our classifier achieved high accuracy: ~93% overall, with strongest performance on Focused and Distracted categories. TinyLlama was effective at producing realistic and behavior-diverse answers, enabling us to create a robust labeled dataset without manually collecting student data.

One key challenge was ensuring behavioral labels were consistent across synthetic examples, which we addressed through iterative labeling and validation. Another insight was that even small models like DistilBERT can perform well on nuanced classification tasks when paired with high-quality synthetic data.

Future work includes expanding the behavioral categories (e.g., curious, guessing), exploring few-shot learning approaches, and integrating real student responses to evaluate transferability. Overall, the combination of TinyLlama and DistilBERT proved effective in modeling and classifying student behavior in educational NLP applications.

**References**

* Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint [arXiv:1910.01108](https://arxiv.org/abs/1910.01108)
* OpenBMB. (2023). *TinyLlama: An open-source 1.1B LLM trained from scratch*. GitHub Repository. <https://github.com/jzhang38/TinyLlama>
* Bakht, M., Suresh, A., & Desmarais, M. C. (2018). *Detecting confusion in students’ text responses using sentiment analysis and coherence measures*. In Proceedings of EDM.
* Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). *The development and validation of automatic indicators of cohesion in text*. Behavior Research Methods, 48(4), 1224–1237.