# AI-Powered Chatbot: Coach

The AI-powered Coach can be used to aid the interpretation of test-taker reports, provide summaries, and answer questions about the assessment itself. As with all large language models, outputs should be sense-checked and used to support, rather than make, decisions.

## Chatbot Finetuning

The current version of the AI-powered Coach relies on a Retrieval-Augmented Generation (RAG) system, which utilizes prompt engineering with GPT-4 Turbo and an external knowledge base incorporating valuable literature to enhance response accuracy and relevance.

To establish a robust system for future improvement based on user feedback on the summaries or conversations generated by the Coach, and to ensure that this feedback is effectively utilized for finetuning the model or adjusting prompts, a number of important factors have been kept in mind in order to develop a responsible and effective AI 'Coach':

* **Training Data Set:** Coach will be improved using user-generated feedback on the summaries and conversations provided by the AI Coach. This data can include ratings, qualitative feedback, and specific annotations indicating whether the content was helpful, accurate, and unbiased.
* **Data Gathering:** This feedback data is gathered directly within the application interface, where users are prompted to rate the helpfulness and accuracy of the AI-generated content immediately after receiving it, with mechanisms for reporting biases, inaccuracies, or other concerns.

* **Data Labelling:** initial labels on the user-inputted feedback data are semi-automated.For more nuanced feedback, a manual review by a team of trained annotators categorizes feedback into more detailed labels, such as types of bias, areas of inaccuracy, or aspects of helpfulness.
* **Training rules:** the Coach language model is trained with the objective of maximising the efficacy, helpfulness, and unbiased nature of the content it generates. Training rules include prioritising user engagement metrics (e.g., positive ratings), minimising identified biases, and enhancing the diversity of perspectives in responses. Moreover, techniques like reinforcement learning from human feedback (RLHF) all the model to learn directly from user ratings and annotations to improve over time.

## AI ethics

**Auditing for adverse impact**

In addition, conducting regular audits of the AI's performance across a variety of demographic groups and topics is crucial for identifying any adverse impacts. These audits include benchmarking analyses to uncover systematic biases or disparities in the modelss performance, utilizing tools such as BBQ or BOLD for bias benchmarking, Toxigen for toxicity benchmarking, Halueval, and Truthful QA for hallucination benchmarking. Additionally, Red Teaming Tests are used to evaluate the model's robustness and security. These include:

* Static Red Teaming: Employing publicly available red teaming datasets, such as Harmful QA, Cat QA, Adversarial QA, etc., to assess the model's robustness and security against biased, harmful, and sensitive elicitation questions.
* Dynamic Red Teaming: Leveraging LLM to generate a private red teaming dataset that dynamically tests the model's robustness and security, offering a continuous evaluation approach.
* Interactive Red Teaming: Engaging in agent interaction to allow other LLMs to persuade or deceive our model into producing harmful or stereotypical content. This method tests the model's conversational integrity at the interaction level.

Moreover, qualitative engagement with external experts or diverse focus groups is used to uncover subtle biases or unintended consequences not immediately evident from the testing described above.

**Increasing Diversity**

To increase diversity and mitigate bias, the data collection and training process explicitly incorporates diverse perspectives. This involves ensuring that the feedback and training data come from a broad and diverse user base and that the data labelling and review processes involve individuals from varied backgrounds. Indeed, the development team alone includes individuals with various areas of expertise, including business psychology, AI ethics, and machine learning, and represents a variety of nationalities. Techniques such as counterfactual data augmentation, where the training data is augmented with examples that challenge stereotypical associations or biases, are also employed to increase diversity.

**Mitigating Unintended Consequences and Bias**

Continuous monitoring and feedback loops are crucial. This involves not only adjusting the model based on user feedback but also regularly re-evaluating the criteria for feedback collection and the model's training objectives. As such, ethical guidelines and frameworks guide the development process, ensuring that decisions around data use, model training, and performance evaluation prioritize fairness, transparency, and accountability.

In summary, a feedback loop for finetuning the AI 'Coach' based on user feedback involves careful consideration of data collection, labeling, and training processes. Assessing for adverse impact and actively working to increase diversity and mitigate bias are crucial components of responsibly deploying AI in this context. Engaging with diverse stakeholders and continuously monitoring the system's impact is key to understanding and addressing unintended consequences.