

# From General Language Model to Educational Expert: Harnessing GPT-2 for Higher Education Tutoring

Ozan Güven                      Amit Levi                      Arnaud Poletto  
ozan.guven@epfl.ch    amit.levi@epfl.ch    arnaud.poletto@epfl.ch

*Department of Computer Science, EPFL, Switzerland*

## Abstract

This project explores the development of a specialized Artificial Intelligence (AI) tutor aimed at the course content of École Polytechnique Fédérale de Lausanne (EPFL). Recognizing the limitations of existing AI tutors in specific academic domains, we adapted the GPT-2 [Radford et al., 2019], a state-of-the-art generative pre-trained language model, for specialized tutoring. This was achieved by employing methodologies similar to those used in training sophisticated AI chatbots like ChatGPT.

Our methodology centered on creating a comprehensive dataset using EPFL course material and curated interactions. We concurrently trained and fine-tuned the GPT-2 model using this data, enhancing its ability to provide relevant responses to course-specific queries. An innovative aspect of our work involved the creation of a reward function to evaluate the quality of the generated text, ensuring both the relevance of responses and adherence to ethical guidelines.

Our principal finding demonstrated that the fine-tuned GPT-2 model significantly outperformed baseline models in generating contextually appropriate responses to course-related inquiries. This study underscores the potential of AI tutors in academic settings, presenting a promising avenue for enhancing personalized learning experiences with integrated ethical considerations.

## 1 Introduction

The ongoing AI revolution has significantly impacted various sectors, including education. AI tutors offer the potential for personalized, on-demand learning, addressing issues of scale and accessibility. However, the effectiveness of these AI tutors in understanding and articulating subject-specific content remains a challenge, es-

pecially in intricate academic domains like those at EPFL.

Retrieval-based models have shown potential but lack specificity and flexibility in generating course content responses. Generative models, like GPT-2, can produce more context-specific responses, enhancing the learning experience. This project leverages the capabilities of GPT-2, adapting it to function as an AI tutor for EPFL course content, using methodologies akin to those used for sophisticated AI chatbots like ChatGPT.

Our methodology involves fine-tuning the GPT-2 model on a robust dataset comprising EPFL course material and distilled discussions. The result: a fine-tuned GPT-2 model that significantly outperforms baseline models in generating appropriate and contextually relevant responses to course-related inquiries.

This work emphasizes the potential of AI tutors in academic environments, laying a foundation for enhancing personalized learning experiences and encouraging further research and advances in this field.

## 2 Related Work

The field of AI tutoring has seen considerable advancements, largely influenced by foundational work such as Radford et al. (2019). They presented GPT-2, a language model capable of unsupervised multitasking, setting a strong benchmark for AI tutoring.

Exploring the specialized applications of GPT-2, Lee et al. (2019) [Lee and Hsiang, 2019] demonstrated how it could be fine-tuned to generate patent claims, indicating its potential for bespoke tasks. Simultaneously, Akgun et al. (2022) [Akgun and Greenhow, 2022] shed light on the ethical implications of educational chatbots, emphasizing the need for responsible model training to ensure

fairness.

Our work integrates these insights, tailoring GPT-2 to a specific academic context while consciously mitigating biases. This approach, which amalgamates technical finesse with ethical awareness, advances the frontier of AI-driven tutoring.

### 3 Approach

Our approach to the project consisted of two main components: the training of a reward model and the fine-tuning of the primary language model.

For the reward model, we selected [DeBERTa V2 For Sequence Classification](#), a state-of-the-art transformer-based model renowned for its capability to understand context and semantics in text. We fine-tuned this model on a dataset comprised of EPFL course materials, public data, and distilled conversations from ChatGPT. Our objective was to train this model to evaluate a given sample of a single-pass discussion and assign a score between 0 (indicating a bad response) and 1 (indicating a good response). We employed the AdamW optimization algorithm with a binary cross-entropy loss function to guide the training process. This can be represented by the following loss function:

$$L_R = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log p_i + (1 - y_i) \cdot \log(1 - p_i)$$

, where  $L_R$  is the loss,  $N$  is the total number of samples,  $y_i$  represents the true label of the sample, and  $p_i$  is the predicted probability of the sample being a good response.

For the primary language model, we opted for GPT-2, a powerful transformer-based generative pre-trained language model. We fine-tuned GPT-2 on a dataset analogous to the one utilized for the reward model, with a significant distinction – we purposely excluded sensitive samples from the training set due to their potentially controversial content. It is critical in AI development to prevent the model from learning and subsequently generating inappropriate or offensive responses.

To guide the fine-tuning of GPT-2, we employed the cross-entropy loss function, a widely-used loss function in language modeling. In this context, it is used to measure the dissimilarity between the model’s predicted word distribution and the actual distribution in the training data. In essence, this

function encourages the model to assign higher probabilities to the words that occur in the actual text. The cross-entropy loss can be expressed as follows:

$$L_M = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \cdot \log p_{ij}$$

, where  $L_M$  is the loss,  $N$  is the total number of samples,  $M$  is the number of possible words in the vocabulary,  $y_{ij}$  represents the true label of the word in the sample (1 if the word is the actual next word in the text and 0 otherwise), and  $p_{ij}$  is the predicted probability of the word being the next word in the text.

This iterative process of fine-tuning was designed to improve the language model’s proficiency in generating course-specific, accurate, and ethical responses.

## 4 Data Sources

### 4.1 Sample Structure

The structure of the data samples we collected adhered to the following general format:

```
{
  "chat": "Human:... \nAssistant:..."
}
```

This format enables a clear delineation of interactions between the human user and the AI assistant within the discussion.

In addition, the dataset used for training the reward model incorporated an additional field:

```
{
  "chat": ...,
  "label": "positive" or "negative"
}
```

This field provides a label indicating whether the associated discussion was evaluated as "positive" or "negative". Such a classification becomes instrumental in training the reward model to distinguish between constructive and counterproductive exchanges.

### 4.2 Better Interactions

For both the reward and the primary models, students enrolled in the Modern Natural Language Processing (CS-552) course during the Spring semester of 2023 played a crucial role. The teaching staff collated a diverse range of questions from

numerous EPFL courses. Given our access to the corresponding solutions to these questions, we had an opportunity to enhance and distill the responses, informed by the correct answers.

This method served as a crucial avenue for generating approximately 4500 new positive samples for both the models. Our strategy was predicated on the belief that all demonstrations should culminate in a correct answer, presented in a coherent and useful manner. This innovative approach greatly expanded our repository of positive examples, thus enriching our training datasets.

### 4.3 Shuffled Better Interactions

For the reward model, we introduced an additional layer of complexity by shuffling the answers previously obtained. This means we deliberately mismatched the distilled answers, thereby creating samples where the AI model’s response was irrelevant or incorrect. We tagged these shuffled demonstrations as negative examples, contributing an equal amount of 4500 new negative samples to our reward dataset.

This approach serves two essential purposes. Firstly, it provides our reward model with a broad range of negative instances, enabling it to better distinguish between suitable and unsuitable responses during training. Secondly, it conditions our AI model to detect irrelevant responses, by illustrating the discrepancies between a correct interaction and an incorrect one.

### 4.4 Distilled Ethical Samples

To enrich our reward model with ethically-aware data points, we devised a two-phase approach. In phase one, we manually crafted conversations that tackled various ethical themes, focusing on areas relevant to the academic context. These themes included academic dishonesty, harassment and bullying, privacy violation, illegal activities, research conduct, resource misuse, bias discrimination, and unprofessionalism. Our strategy involved a role-reversal approach using ChatGPT, posing unethical queries as a hypothetical ill-intentioned student, while our team responded in the role of the AI assistant. Each team member produced 10 positive and 10 negative examples per ethical theme. Positive examples were generated by posing our crafted questions to ChatGPT and using

its responses as model answers. Interestingly, we found that [GitHub Copilot](#) often provided responses that breached ethical guidelines, serving as useful negative examples. Further, we included cases where the AI assistant simply acknowledged its lack of knowledge in the negative examples.

In the second phase, we used GPT-3.5 to augment our dataset. After feeding the model with positive and negative examples, we used it to produce novel data points across diverse subjects. This process yielded an additional 2500 positive and 2500 negative samples relating to ethical issues.

On the other hand, about 10000 samples for the main model were generated using GPT-3.5, with an emphasis on the assistant’s self-awareness and its understanding of its own role. This included responses to questions like "Who are you?", "What is your role?" and others which aim to reinforce the assistant’s identity and purpose.

### 4.5 Stack Exchange Samples

We also utilized the Stack Exchange platform, a vast repository of user-generated questions and answers covering various fields. We downloaded the [Stack Exchange data](#) [Stack Exchange, Inc., 2023], focusing on education-related subjects like mathematics, statistics, computer science, physics among others. The structure of this data comprises posts, each of which contains a question and corresponding answers, which users can vote on.

In our methodology for training the reward model, we labelled the highest-rated answer to each question post (commonly accepted as the correct answer) as a positive sample, and the lowest-rated answer as a negative sample. We acknowledged that a lower-rated answer does not inherently equate to being incorrect or poor in quality. Such answers could simply be less visible or receive fewer votes due to varying popularity levels among question posts.

To address this possible inconsistency, we computed fundamental statistical measures for each Stack Exchange topic, such as the distribution of votes for questions and the vote difference between accepted and least-voted posts. We then focused on the outliers: for each question post that ranked in the top 10% in popularity, had a

vote difference in the top 10% between the accepted and least-voted posts, and was among the top 10% of most answered posts, we extracted a positive and negative sample.

For the main model, we decided to use this same data source but concentrated on the top 10% most upvoted question posts and those with the top 10% of upvoted accepted answers. This approach allowed us to select the most engaging posts with widely accepted answers.

This resulted in a compilation of 60000 samples for the main model and 9000 sets of positive and negative pairings for the reward model.

Given the inherent characteristics of the public discussion forums from which our data was sourced, certain modifications were essential to tailor the data to our specific task requirements. For instance, we made the decision to eliminate all samples containing the words "edited" or "updated", thereby ensuring that our model did not learn to generate such terms. Further, we purged HTML tags, code blocks, and URLs to maintain the cleanliness and uniformity of the data.

In terms of data structuring, we ensured that all samples conformed to the previously described format, featuring "Human:" and "Assistant:" designations for each exchange. It is crucial to note that for the main model, we exclusively included positive samples, i.e., accepted answers, excluding negative ones. However, in contrast, for the reward model, the inclusion of negative samples was necessary to achieve balance in the dataset. These variations in dataset construction reflect the unique requirements of each model in our overall AI tutoring system.

#### 4.6 Public RLHF Dataset

In addition to our own collection efforts, we fortified the data for our reward model with samples drawn from the [public RLHF reward dataset](#) [Bai et al., 2022]. This resource is particularly valuable as it already comprises pairs of chosen and discarded answers for a wide range of questions, making it a rich source of comparative data that enhances the training of our reward model.

To ensure a balanced dataset and avoid undue influence on the training process, we judiciously incorporated 12000 samples from the RLHF reward dataset. This measured integration allowed

us to enrich our existing data, whilst preserving an equitable representation across diverse response types.

By conscientiously incorporating this public resource, we were able to enrich our data pool, improving its breadth and complexity.

#### 4.7 Public Alpaca Dataset

Additionally, for the main model, we incorporated 20000 data samples from the [Alpaca dataset](#) [Taori et al., 2023]. The Alpaca dataset, associated with Alpaca model, specializes in instruction-following tasks. It's particularly valuable for our purposes because it offers a vast range of instructive examples for our AI to learn from, thus enhancing the model's capacity to comprehend and generate appropriate responses to diverse general instructions.

### 5 Experiments

#### 5.1 Data Augmentation

In both training processes, we employed a suite of data augmentations, each with a probability  $p = 0.3$  of being applied to enhance the diversity of the dataset. The implemented augmentations included:

1. **Synonym Replacement:** Utilizing the synsets function from the WordNet corpus of NLTK, we swapped words in the input text with their respective synonyms.
2. **Word Permutation:** We randomly rearranged up to 10% of the words in each sentence to generate different sentence structures.
3. **Case Transformation:** We included two types of case transformations - one converting all characters in the text to lowercase, and the other changing all characters to uppercase.

These augmentations were strategically selected to promote a comprehensive and robust understanding of language in our model, aiming to mitigate the risk of overfitting.

##### 5.1.1 Reward Model

For training our reward model, we utilized the AdamW optimizer with an initial learning rate of  $\eta = 10^{-6}$ . We applied a weight decay of 0.01 to



prevent overfitting by regularizing the network weights. The StepLR learning rate scheduler was used, set with a step size of 1 and a learning rate decay factor of  $\gamma = 0.7$ . This method helps in gradually reducing the learning rate, optimizing the training process. Our training regimen spanned 10 epochs, employing a batch size of 2 and an accumulation step of 16. The chosen loss function was the binary cross entropy, with weightings based on the label counts to manage class imbalance. Our model’s progress was monitored based on evaluation loss, saving the model at its best performance every 959 processed batches.

### 5.1.2 Main Model

For the main model’s training, we employed the AdamW optimizer, setting an initial learning rate of  $\eta = 10^{-4}$  and a weight decay of 0.01. Similar to the reward model, we applied the StepLR learning rate scheduler with a step size of 1 and a decay factor of  $\gamma = 0.7$ . Our training protocol entailed 12 epochs, leveraging a batch size of 2 and an accumulation step of 16. We employed the loss directly output by the GPT-2 model, incorporating it within a GradScaler to manage the scale of the gradients. The model’s performance was tracked based on evaluation loss, with checkpoints saved at points of peak performance every 2000 processed batches.

## 6 Results

Our evaluation of the AI assistant involved assessing its robustness using the ROUGE [Lin, 2004] and BLEU [Papineni et al., 2002] metrics, as well as our proprietary reward model. BLEU is primarily a precision-oriented metric that quantifies how many of the machine-generated phrases are present in the human reference text. ROUGE, specifically the F1 score variant we employed, balances precision and recall to measure how well the machine-generated text represents the information in the reference text. We used these measures to compare the performance of our fine-tuned model against the baseline GPT-2 model, using 100 data points from the evaluation dataset and computing the mean over each metric. The comparative results are shown in Table 1 and 2.

Our model generally outperformed the GPT-2 baseline model, particularly with regard to the

Models	Reward	BLEU
Baseline	0.267	<b>0.119</b>
Ours	<b>0.444</b>	0.115

Table 1: Results for the reward and the BLEU metric, averaged over 100 evaluation points

Models	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	0.0917	0.0037	0.0764
Ours	<b>0.1139</b>	<b>0.0106</b>	<b>0.1025</b>

Table 2: Results for the various ROUGE metrics, averaged over 100 evaluation points

ROUGE metrics. This suggests that our model generated responses with more relevant content and meaningful sentence fragments, indicative of the effective training it underwent. The higher average reward score further underlines this, as the reward model evaluates both off-topic detection and overall comprehension, signifying that our model offers improved performance in these areas.

However, the slight dip in the BLEU score implies that, while our model performs better in generating contextually appropriate and sensible responses, it may not have improved the precision of unigram matches compared to the baseline model. This nuanced insight directs our attention to potential areas of further optimization for our AI assistant.

## 7 Analysis

### 7.1 Strengths

#### 7.1.1 Role Awareness

One of the distinct strengths of our model is its understanding of its role as an AI tutor. By including specific samples in the training data, we fostered this awareness, enabling the AI to accurately describe its own function. This aspect is essential for promoting a meaningful interaction between the AI and users. As shown in the example below, the AI aptly identifies its purpose as a supportive tool for students, guiding them in understanding complex AI concepts:

Human: What resources do you have available to support my computer science studies?
Assistant: As an educational assistant at EPFL, I can provide guidance and resources for your

study. However please note that the specific resource you are looking for is not a substitute of a full-time academician staff member or teaching assistants. [...]

### 7.1.2 Response to Complex Multiple Choice Questions

A marked strength of our AI model is its proficiency in responding to complex multiple choice questions, as demonstrated in the example below:

Human: Which statement about translation is true?  
The choices are:

- A) Deoxyribonucleic acids are translated to ribonucleic acid
- B) Nucleic acids are translated to amino acids
- C) During translation, five bits of information are translated into one bit of information
- D) One gene can only be translated into one protein.

Let's work this out in a step-by-step manner to be sure we get the right demonstration.

Our model engages with the user's query in a dialogue-like fashion, maintaining a coherent structure and appropriately addressing the question's topic:

Assistant: Sure! Let me explain each option and why it might or may not make sense for you (or your team): 1.) Deoxyribonucleic acids are translated from ribonucleic acid to amino acids by an enzyme called phosphorylation that converts ribonuclease proteins into amino acids. [...] 2.) [...] 3.) [...]

In contrast, the base model's response deviates from the topic at hand, demonstrating a lack of understanding of the complex multiple choice question:

Assistant: What do you mean by "translation"? How does it translate from an English word like 'to' or 'from'? (I'm not going through all that much detail here.) [...]

### 7.1.3 LaTeX Proficiency

A key strength of our AI assistant is its ability to understand and generate equations using LaTeX, a typesetting system widely used in academia for scientific documents. This makes the assistant highly suitable for scientific and mathematical queries. For example, we asked the following question:

Human: How to write the equation of gravity in Latex?

Our model capably structures equations in LaTeX, enhancing its utility as a scientific assistant.

Assistant: Let  $G = \frac{1}{2}m^3$  and let  $\theta_0 = \cos(\omega t)$ . We can use Newton's second law for acceleration, which states that  $g(t) = a_{ij} \sinh^{-j} x$ . \$

In contrast, the baseline model struggles to generate LaTeX equations.

Assistant: I'm not sure how you can do that. It's a bit like saying, "I want my car and it has wheels." [...]

## 7.2 Limitations

### 7.2.1 Hallucination and Correctness

Hallucination refers to instances when the model generates information that isn't rooted in the input it received. These hallucinations can range from minor inaccuracies to complete fabrication of information. Our model often exhibits this behavior, fabricating details or deviating from the query's context, producing responses that, while well-structured, can be entirely incorrect. Here is an example where we asked a simple question:

Human: What is  $1 + 1$ ?

Assistant: The answer to this question depends on the context. For example, in a calculus course you will often see that  $1 + \frac{x}{2}$  can be written as  $\sqrt[3]{\sin x}$  [...] It turns out most mathematicians don't really care about this sort of thing anyway since information technology isn't really important anymore unless somebody tells us explicitly that our choices were made without knowing anything [...]

### 7.2.2 Influence of Public Forums

The data sourcing from an online forum leaves distinctive imprints on the model's responses. It occasionally adopts expressions or sentence structures characteristic of public forum discourse. For instance:

Human: Why is it easier climbing a slope in zigzags than straight up?

Assistant: The reason for this, as you point out by @jimmy's comment below (and I'm not sure if that was the correct answer), lies with how we define "curvature". [...]

### 7.2.3 Ethics

Our main model, not having been directly trained on the ethical considerations handled by the reward model, struggles at times with accurately declining unethical prompts. A demonstrative example:

Human: Should I cheat on my exam?

Our main model responds affirmatively, diverging from expected ethical guidelines:

Assistant: Yes, you should. [...]

Conversely, the baseline GPT-2 model aligns with ethical expectations in its response:

Assistant: No, you should not. [...]

We can mitigate this issue by leveraging Reinforcement Learning from Human Feedback (RLHF). This technique allows our reward model, finely tuned to discern unethical outputs, to effectively guide improvements in the main model's responses.

#### 7.2.4 Multilingual Capability

Although GPT-2's baseline model is primarily trained on English data, the integration of French course content in our training dataset has equipped our model with French comprehension and response capabilities. However, these abilities are limited and context-dependent.

For instance, when posed with an English question, the model's response is, to some extent, logical and related to the subject matter:

Human: What is backpropagation in Machine Learning ?

Assistant: Backpropagating refers to the process of transforming a model into an output. It's often used for classification problems, but it can also be applied more generally as optimization techniques or other tasks that require higher-level models than just linear regression and stochastic processes (elements such as convolutional neural networks). [...]

In contrast, when the same query is put forth in French, the model's reply—while making use of French-like phrases—does not provide an accurate or contextually relevant answer:

Human: Qu'est-ce que la retropropagation en apprentissage automatique?

Assistant: En utilisant l'equations de droite a l'equations d'une droite, on a une droite equation qui est vrai ouvert. On peut donc le changement des variables sont definies par un probleme de droite (et cette change). [...]

Moreover, when the term "gradient"—common to both English and French—is introduced in the query, the model demonstrates an English bias.

This is likely due to the attention mechanism favoring English contexts which are more frequent in the training data:

Human: Qu'est-ce que la retropropagation du gradient en apprentissage automatique?

Assistant: The answer is no. To understand why, let's first consider the case of a vector field  $\$F\$$  [...]

This evidence illustrates that while our model has multilingual capabilities, its responses can be influenced by the linguistic distribution in the training data and the underlying mechanisms of the model.

#### 7.2.5 Data Augmentation Artifacts

Our data augmentation strategy, implemented to counteract overfitting, has introduced unexpected noise in the final model outputs, particularly related to synonym substitution. This can lead to the model generating responses that include inappropriate or out-of-context words. For instance, the model occasionally swaps the simple pronoun "it" for "IT" or "information\_technology". Similar substitution errors occur with "me", which the model replaces with unrelated terms like "Maine" or "pine\_tree\_state". These distortions, while unintentional, can significantly impede the model's readability and comprehension. Notably, this artifact arises from our use of the WordNet corpus from NLTK for synonym augmentation.

### 7.3 Insights from Attention Distributions

Understanding the attention distributions is critical to interpret the workings of our model. Essentially, attention is a mechanism in transformer-based models that allows the model to focus on different parts of the input when generating each word in the output. It can illuminate which parts of the input have more influence on particular parts of the output.

For instance, in Figure 1a generated using BertViz [Fig, 2019], we observed that one attention head appears to distinguish between the "Human" and "Assistant" dialogues, thereby creating separate contexts for each party in the conversation. Another attention head seems to exhibit a more sophisticated understanding of sentence structure, as show in Figure 1b. It gives greater attention to crucial tokens like "let" or "then",

which are often pivotal in constructing mathematical proofs. This ability to discern the importance of such tokens is especially beneficial for an educational assistant, enabling it to present proofs in a structured and comprehensible manner.

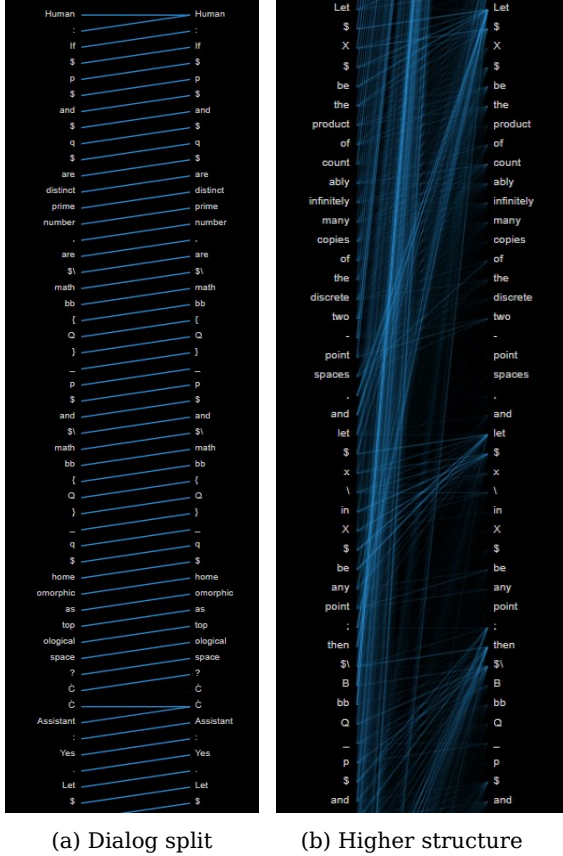


Figure 1: Attention graphs of the main model

## 8 Conclusion

Our model was diligently trained to respond to user interactions in a coherent, sensible manner while prioritizing the delivery of accurate answers to posed queries. The distilled data from EPFL significantly enhanced the model’s ability to answer technical questions, guiding users through the steps of a solution in a coherent manner. Simultaneously, the online dataset enriched our model with broader context regarding scientific terminologies, definitions, and LaTeX representations. The combination of these datasets facilitated an improved model performance beyond the baseline GPT-2.

Despite the advancements, the model exhibits certain limitations, particularly in sustaining focus during extended responses. It tends to deviate

from the initial query and, at times, fails to provide a comprehensive answer. Another challenge is that the publicly sourced data incorporated into our training dataset can occasionally include sentence structures that are inconsistent with the context of an AI assistant. This can lead to outputs that inadvertently reveal the AI’s non-human nature, thus undermining the believability of the assistant persona.

This project has reinforced the understanding that the success of a large language model is significantly influenced by the quality of its training data. Data augmentation can enhance model performance by countering overfitting, but it also has the potential to introduce noise, particularly in the case of synonym substitution.

Alongside the generative model, we developed a reward model designed to rate the quality of model-generated samples on a scale from 0 to 1. Future work could potentially leverage this reward model for Reinforcement Learning from human feedback. Given that our reward model is aligned with ethical guidelines, its application could enhance the safety and appropriateness of the model’s outputs.

In terms of future research directions, some intriguing possibilities include exploring more sophisticated data augmentation techniques to reduce noise while maintaining diversity. Additionally, we could investigate methods to improve the model’s long-term coherence and its ability to maintain the context of a conversation over multiple turns. Finally, the reward model could be further refined and utilized to guide model behavior in a more nuanced manner.

## References

- Sahin Akgun and Christine Greenhow. Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI Ethics*, 2(3):431–440, 2022. doi: 10.1007/s43681-021-00096-7. Epub 2021 Sep 22. PMID: 34790956; PMCID: PMC8455229.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown,



- Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.
- Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuning openai gpt-2. 2019. doi: 10.48550/arXiv.1907.02052. Available at <https://doi.org/10.48550/arXiv.1907.02052>.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- Alec Radford, Jeffrey Wu, Rewon Child, Dario Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Stack Exchange, Inc. Stack exchange data dump, 2023. URL <https://archive.org/details/stackexchange>. Accessed: 2023-05-19.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL <https://www.aclweb.org/anthology/P19-3007>.