
Pioneering Intelligent Education: Unleashing AI Tutors with Limitless Knowledge

Nay Abi Akl, Mariam Hassan, Luca Zunino

Abstract

This project tackles the challenge of developing specialized AI tutors in the educational field using various scale language models. To provide personalized guidance on EPFL course topics, we leverage the power of existing models with our own prepared and collected datasets to augment the model with domain-specific knowledge, improved capabilities and higher expertise. To accomplish this, we collect supervised finetuning data using the ChatGPT wrapper, train a reward model on the collected data plus other gathered data, and finally train our chatbot model on the question answering task using supervised finetuning. We leverage the capabilities of Open Book Models by integrating a keyword extractor and online context retriever into our pipeline. Results showed high accurate generations that conform with the ground truth and good generalization capabilities, forming a promising base for future improvements. We also provide a real-time QA script that generates relevant context and answers in real time.

1 Introduction

The integration of AI technologies, specially LMs, in the field of education has led to great advancements. However, developing specialized and effective AI tutors is still a challenge. This project attempts to solve this challenge using large and medium scale language models to train a custom AI tutor for EPFL courses content. By leveraging successful methodologies used to train modern chatbots, for example ChatGPT, we aim to develop an advanced assistant that can provide support and personalized guidance to EPFL students.

The need for a chatbot customized for EPFL courses comes from the growing diversity and complexity of modern day educational content, with conventional chatbots lacking the domain specific knowledge needed to assist with such intricate subjects. Leveraging the power of existing models, which lies in their large contextual information repositories, enables us to create an assistant with higher expertise and improved capabilities, further enhanced by the possibility of gathering context from external sources.

Our approach is split into three main parts. First, we collected supervised fine-tuning data by distilling demonstrations from the ChatGPT wrapper. Second, we train a reward model on these demonstrations and on additional datasets collected and preprocessed for the task. Third, we train our chatbot model on the Question Answering QA task using supervised finetuning. This project has great potential in revolutionizing educational assistance by offering tailored explanations, personalized guidance and smart feedback. Incorporating advancements in LMs and RL techniques, we hope for an effective, engaging and interactive learning experiences for EPFL students.

2 Related Work

The first part of the project involved distilling demonstrations from the ChatGPT wrapper. We gathered ideas and inspirations for our prompting strategies from multiple sources including (1) where the main idea is to break down a complex problem into its sub-problems, which is referred to as “least-to-most” prompting.

Since the end goal of the project is to answer questions, we also rely on the work done in (2) to better understand the reasoning capabilities of LLMs and what training pipelines can be used to improve these abilities, in the context of generative question answering.

As for the final chatbot model, a useful replication that we decided to implement are Open Book Models (OBMs). Unlike traditional LMs that rely only on their pretrained parameters, OBM rely on external sources of information (for example web pages, books, articles) to augment the knowledge base they have, allowing them to address a broader range of topics and thus generate more accurate and comprehensive responses. Therefore, OBMs offer the possibility of more nuanced and knowledgeable conversation, making them valuable tools for various applications, including question-answering systems, virtual assistants, and information retrieval systems. One example of such models can be found in (3) where an LM is applied to the following three

consecutive modules “Search engine → Knowledge → Response”, resulting in the SeeKeR model. In fact, for each input context, the SeeKeR model first searches for relevant documents online (the search module), then extracts a few relevant sentences to the input context from these documents (the knowledge module), and finally generates the response based on the input context and the extracted knowledge (the response module).

Results showed that the model outperformed GPT2 and GPT3, in terms of topicality and factuality, when applied to topical prompt completions as a typical LM. Inspired by this work, we introduce two modules to our pipeline, the keyword extractor and the context retrieval (explained in Section 3.3), with the aim of improving our final model’s generations. On another note, Reinforcement Learning with Human Feedback (RLHF) is a training approach that combines Reinforcement Learning (RL) with human feedback with the aim of improving the performance of LM by bridging the gap between the knowledge and expertise provided by humans and the limitations of solely data-driven RL. Results of (4) prove that RL is indeed a good framework for NLP, specially when combined with supervised finetuning, in terms of human aligned generations. Inspired by this work and many others in the field of RLHF, we will attempt to use our reward model to further fine-tune our chatbot using RLHF, of course after securing a well-trained model from supervised fine tuning.

3 Approach

3.1 Reward Model

The multilingual “xlm-roberta-base” model, pretrained on 2.5TB of filtered “CommonCrawl” data, is used as the reward model, allowing us to handle English and French questions. This model was pretrained with a Masked Language Modeling (MLM) objective, allowing it to learn an internal representation of 100 different languages that can be used in extracting relevant features for downstream tasks. For example, with labeled sentences, one can train a classifier head using the features outputted by the XLM-RoBERTa model as input. This works best when the corpus that was used for pretraining is not too dissimilar from the one that will be used for training. However, in cases where the data is different, finetuning the language model on our data (considered as the in-domain data) before training for the downstream task could improve the performance. This process is known as domain adaptation. Therefore, aiming for the best possible performance, we decided to finetune the “xlm-roberta-base” model chosen on a language modeling task by masking tokens and training on the solution dataset provided in the course. The model was registered in HuggingFace (5) and will be used for the reward model by specifying its corresponding name. The effect of finetuning has then been qualitatively analysed, results (shown in the Appendix 7.1) proved that the finetuning process seems to be effective.

3.2 Final chatbot model

For the final chatbot model, we consider the following three models: GPT2 (6), T5ForConditionalGeneration (7), and Flan-T5 (8), with each having its own characteristics and strengths that can be beneficial for the QA task. The GPT2-base is a transformers based model with **124M** parameters that was pretrained on very large **English** corpus data following a self supervised fashion to guess next words in the sentence. The T5ForConditionalGeneration model has **220M** params and is based on the T5 model, which is an encoder decoder transformer that was pretrained in “text-to-text” denoising generative settings, with a LM head on top. This model can handle different languages, including both **English and French**, and works well out-of-the-box on various tasks by choosing the right prefix to the input. The Flan-T5, also of **220M** parameters, is an added enhancement over T5, for the same number of parameters, since it was finetuned on more than 1000 additional tasks, including question answering, question generation, closed-book, adversarial and extractive QA. Therefore, we believe the model might be better suited for our chatbot.

3.3 Context retrieval

The task of question answering (QA) stands as a critical challenge within the realm of natural language processing. QA systems are a common component of interactive dialogue systems and have been extensively investigated for various domains and applications. However, QA is a very challenging task, especially when the questions are complex and require reasoning and inference beyond surface-level matching. Furthermore, queries could revolve around topics foreign to the QA model, having been absent during the training phase. In such cases, delivering a relevant context along with the question to the model can enhance the accuracy and quality of the derived answer. This approach is known as context-based QA.

Context-based QA has been extensively explored in literature, and many benchmark datasets and evaluation metrics have been proposed for this task. From our preliminary tests, it emerged that the T5 model (candidate as our QA model), is very sensitive to an accurate and informative context. Since we wanted to provide context to the model also at inference time, we decided to build an open-book model. This model can gather meaningful context from the internet, and from Wikipedia in particular, by determining the most effective keyword related to the question, and using Wikipedia’s APIs (9) to search for that keyword. Therefore, the procedure to implement an open-book model is made of two main steps:

1. Finetuning T5 to determine which is the most suitable keyword to retrieve meaningful context on Wikipedia given the question. This model will be called keyword retriever model.
2. Establishing an effective mechanism to initiate a search query on Wikipedia to retrieve a context that can help our QA model answer a given question.

As stated before, we decided to provide context to the model both during training and at inference time. The procedure to obtain context in these two situations is necessarily different:

- Since the training datapoints are known in advance, to speed up the training procedure we preprocessed them to obtain a meaningful context before starting the finetuning procedure on the QA model.
- Since the inference questions are generally not known in advance, the model needs to generate the appropriate context in real time once a new question is asked to the model. In particular, once a new question is passed to the model, this question is first passed through the keyword retriever model to obtain a keyword for the Wikipedia search. Then, the Wikipedia APIs are used to obtain the context, which is passed to the QA model together with the question.

The first decision we had to make was how to implement the search on the internet to obtain meaningful context starting from a question. While doing a lookup using search engine APIs could allow the model to obtain information from a vast amount of websites, this solution is impractical for two main reasons. First of all, most of the search engine APIs are paywalled, and we wanted to identify a solution which was free to ensure scalability and accessibility for everyone who wants to use the model. Additionally, each website imposes unique policies regarding content usage, making it arduous to verify whether specific content is permissible for use as context within our QA model. For this reason, we decided to use Wikipedia as our source of information. The decision was motivated by the fact that Wikipedia’s APIs are fast, free and easy to use. Furthermore, the content available on Wikipedia is generally of high quality, reviewed, and up-to-date, and the terms of use allow to leverage the information contained on the website as context for the model. Another important aspect is which part of the Wikipedia page we considered for our context. Ideally, we would have trained another model that, given a complete Wikipedia page, would return the passage most likely containing the answer or at least a valuable hint towards it, similarly to what we saw in (3). However, due to time constraints, we opted to utilize the summary of each Wikipedia page as our context. While not the optimal strategy in absolute terms, this approach proved advantageous since the summary often includes information that is useful for a cursory understanding of a topic and for providing a broader context about a specific question to the QA model.

We can now start describing more in detail the first step which is necessary to obtain an open-book model, that is the creation of a suitable dataset. Indeed, it has been necessary to process our datasets further, containing couples of questions and golden answers, to also include keywords associated with the questions (to finetune the keyword retriever) and context associated with the questions (to finetune the QA model). Furthermore, we also needed to determine the language of each question (English or French) as Wikipedia’s APIs require specifying the language in which the Wikipedia’s webpage should be returned. To obtain meaningful keywords for the various questions (and the languages of the questions), we decided to leverage ChatGPT through the provided wrapper. The prompt we used is the following: “Given the following question, which concept or definition, if looked up on Wikipedia, would be most likely to help you answer it? Return a single concept, ideally corresponding to a Wikipedia page title. If the question is in English, answer in the following format: EN. [CONCEPT], if the question is in French, answer in the following format: FR. [CONCEPT]”. First of all, we asked for both the language and the keyword in the same prompt to minimize the tokens consumption, especially since our datasets are quite big. Then, we directly specified that the keywords should correspond to Wikipedia pages to try to maximize the number of keywords that, searched on Wikipedia, return existing pages. We manually verified some of the keywords provided by ChatGPT, and most of them proved to be of high quality and suitable for obtaining meaningful context.

At this point, we used Wikipedia’s APIs to obtain a suitable context for each training datapoint. Starting from the keyword returned by ChatGPT, we used the APIs to look up the keyword on Wikipedia. If the page associated with the keyword existed and was not a disambiguation page, then the page summary was used as context for the datapoint. Otherwise, ChatGPT was prompted again three times to try to obtain a different keyword, and the APIs were used again to try to obtain context. After four non-successful attempts, the keyword and context fields for that particular datapoint were set to None, and the processing continued with the next datapoint. We decided not to include the keywords that did not correspond to Wikipedia pages in our datasets to avoid training the keyword retriever on keywords that did not lead to the retrieval of a helpful Wikipedia page.

After processing our datasets, we finetuned a model to obtain a keyword retriever model. In particular, we wanted to obtain a model which, given a question about a generic topic, returns a keyword which, if looked up on Wikipedia, allows to obtain context which is meaningful to answer the question. We evaluated both GPT-2 (6) and T5 (10) as potential pre-trained models, finding that T5 provided superior results. The finetuning procedure of the T5 model is fairly standard. The inputs which are passed to the model have the format “Keyword and Language of: {question}” where {question} is substituted by the various questions. The expected outputs have format “{keyword}|{lang}” where {keyword} is the keyword returned by ChatGPT and inserted in our dataset, and {lang} (either “EN” or “FR”) is the language of the question (again returned by ChatGPT). We finetune our model on the complete dataset (for all the datapoints for which the keyword allowed to obtain a valid Wikipedia page, to be sure the keywords provided by ChatGPT are meaningful) for three epochs. Then, the finetuned model has been

loaded on the HuggingFace Hub (11).

Once our keyword retriever model is finetuned, we can exploit it to generate meaningful context at inference time. More in detail, during inference, we pass the question to be answered by the QA model to the keyword retriever, and this model should return a keyword and the language of the question. To maximize the probability of obtaining a keyword which matches a Wikipedia page, we perform a beam search with ten beams, and we return five keywords (from the most probable to the less probable). Then, the Wikipedia APIs are used by looking for the five returned keywords in the returned language (if the language is not returned, both EN and FR are considered to add robustness), and the first non-disambiguation webpage associated with the keyword is used to return context (again, the summary of that page is used as context for the question).

3.4 RLHF

As mentioned in Section 2, we explore RLHF with the aim of improving our final model with human feedback. Therefore, we rely on an existing Python library, TextRL (12), that supports text generation with reinforcement learning. The library is built upon the Transformers from Hugging Face, the PFRL (13) deep RL library which implements different state of the art RL algorithms in Pytorch, and OpenAI GYM (14) which is an open source library used to develop and compare RL algorithms based on a provided API. The TextRL library was chosen since it is designed with ease of customization in mind and can be easily applied to different text generation models. To this end, we integrated our reward model in a custom RL environment “MyRLEnv”, based on the default class “TextRLEnv”. The “get_rewards” function thus passes the question and the generated answer from the chatbot model as inputs to the reward model to score them. Note that the performance of RLHF highly depends on the reward model and the score it gives which drive the learning of the agent.

4 Experiments

4.1 Datasets

4.1.1 For Reward Model

For the datasets of the reward model, the focus is mainly to get good quality data with scored answers. To do that, we used five different datasets, each processed differently. These are summarised below.

CS-552 Dataset This is the dataset collected by the class. It has roles, questions, answers, and a confidence score. The role is referred to as system, the question as user and the answer as the assistant. However, after carefully analysing the dataset, we identified three main problems which are 1) the confidence scores are not reliable 2) it is hard to identify the specific chat interactions that have the answer to the question 3) the chats were too long with many irrelevant information.

To get the interaction that most probably has the answer, we calculated the BERTScore for every interaction for the assistant and chose the interaction with the maximum BERTScore. We then used the assistant interaction with the maximum BERTScore and the gold label to prompt GPT3 through the wrapper provided, asking it to provide a score for the answer. The prompt used is provided in the Appendix 7.3. To add multi-turn interactions, we first include the system (if available), as well as the user and assistant interactions related to the highest BERTScore. Then, we start adding the next highest BERTScore interactions until the maximum token length is reached.

Texas University Dataset This dataset (15) provides questions and graded answers for a first-year computer science assignment questions in Texas University. They are 87 questions and a total of around 2000 graded student answer where the grades are between 0 and 5 given by the instructors of the course. We created 6 different training files. Every file includes 4 student answers with unique scores for every question out of the 82 questions. We make sure we always have an answer with score 5 and an answer with a low score. Like this, we create 6 repetitions of every question with varying quality of answers.

ELI5 Dataset The ExplainLikeIm5 (ELI5) (available on HuggingFace Hub, “eli5_category” (16)) dataset provides a huge number of questions and answers without scores. It provides questions for general topics, with some of them being scientific questions. Therefore, it could be close to the corpus we are aiming for. We only extracted 10k points of the dataset and assumed the answer with the highest number of upvotes to be the ground truth. We assign a grade of 5 to the ground truth and we use the questions and ground truths to prompt the ChatGPT wrapper for four answers with scores 4,3,2, and 1. We qualitatively assessed the answers and finetuned our prompt until it provided four levels of lower quality answers. An example is shown in the Appendix 7.3. For evaluation, we used a split of the data and assigned “chosen” to the answers with score 5 and “rejected” to the answers with score 1. We also created a second evaluation dataset that had “chosen” given to the answers with score 4 (since they were at least factually correct) and “rejected” to the answers with score 2. This creates a more diverse evaluation dataset.

“synthetic-instruct-gptj-pairwise” and “hh-rlhf” Datasets We also consider two of the datasets which have been used to train OpenAssistant’s reward model “OpenAssistant/reward-model-deberta-v3-large-v2” (17) which are the “Dahoas/synthetic-instruct-gptj-pairwise”, available on HuggingFace Hub (18) and containing a prompt/question, a chosen answer and a rejected answer, and the “Anthropic/hh-rlhf”, available on HuggingFace Hub (19) and containing a chosen conversation and a rejected

conversation. For both datasets, we assigned a score of 5/5 to the chosen answer (since, according to our evaluation, they are generally factually accurate) and a score of 0/5 to the rejected answer (to maximize the score distance with the chosen answer, and since the rejected answer is sometimes completely unrelated to the question). A subset of the two datasets has also been used for evaluation purposes; in this case, we directly kept the original format (chosen and rejected answers). Moreover, while the “Anthropic/hh-rlhf” dataset is particularly valuable as it contains multi-turn conversations, we discovered that some data points are severely toxic (i.e. contains offensive words). Therefore, once we processed all the data points from the dataset, we used the Detoxify model (20) to evaluate the level of toxicity of the data points. If a level of toxicity above 0.1/1 is detected, the whole data point is discarded.

4.1.2 For Generation

The goal for the datasets used to train the generator model is to provide high quality question and answers that are related to education so that the model is trained to answer questions in this field. Overall, 8 different datasets were used to train the model. We first used them with question and answers. Then we added context to each question as discussed in Section 3.3. The CS-552, Texas, and ELI5 datasets are again used for the generation model but this time we only choose the answers that have a score of 5 to provide the model with a question and answer. For the CS-552 dataset, the ground truth answers were used.

Scientific Question Answering (SciQ) The SciQ (21) dataset provides 13,679 crowdsourced questions and answers from exams about Physics, Chemistry, and Biology. The dataset is mainly MCQ questions but the explanation is provided so we consider the original questions as questions and the correct choice concatenated with the explanation as answers. We only considered the correct answer instead of providing choices so that the answer is more general not just focused on specific choices.

Chatbot Dataset (AI Q&A) The Chatbot Dataset (22) provides questions and answers about artificial intelligence. The dataset was chosen since the questions relate to computer science topics. However, some of the questions were answered in just a word or two. We decided to only keep the questions that have answers of more than 5 words, for it to be a meaningful datapoint.

AI Stack Exchange Dataset The AI Stack Exchange Dataset (23) provides around 10k questions about AI and ML. Some questions are “What is “backprop?”” and “How does noise affect generalization?”, which are very related to many courses taught at EPFL. However, the answers are not provided. Therefore, we used the ChatGPT wrapper to prompt for the answers of these questions. An example of the question and answer is shown in the Appendix 7.3.

French and English Definitions Datasets We created two additional datasets, one in French, the other in English, each of around 15k questions and answers. These datasets provide definitions of keywords present in the CS-552 dataset. To obtain them, the ChatGPT wrapper was asked to extract keywords for every question, which was then passed through the Wikipedia API to get its associated definition. Therefore, we created questions and answers in English and French that are based on the questions found in EPFL courses and used for the CS-552 dataset. This is the great advantage of these datasets: we are certain that they cover similar topics to EPFL courses topics since they were directly extracted from them, providing a very meaningful and “on-topic” dataset. One example from each dataset is provided in the Appendix 7.3.

4.2 Reward model implementation and training

For the model’s implementation, the “XLMRobertaConfig”, the “XLMRobertaModel” and the “XLMRobertaTokenizer” are initialized by passing the name of our finetuned model. To obtain a score, we add a regression head to the model which takes as input the CLS token (first token) from the model’s output. The CLS token should embody all the model’s learned knowledge, therefore it is sufficient to feed only this token to the classification layer. The head is made of fully connected layers (with ReLU activation) and a final output dimension of 1 (one score per datapoint). We explore different architectures (in terms of number of layers and number of neurons per layer) to obtain the best performance. Finally, we apply the “sigmoid” function on the last layer and multiply by 5 for scores between $[0, 5]$, to match the dataset format we have. As for the loss, the Mean Squared Error MSE loss is calculated between the model’s outputted scores and the labels of each datapoint. The model was trained for 5 epochs with a learning rate of $1e-5$ and a batch size of 8. To check the sanity of the datasets we prepared and processed, we train the reward model a couple of times, each time changing the datasets chosen for the training. For each approach, we also evaluate on different datasets (a split that was not seen during the training phase) to be able to give our analysis. The different training approaches and detailed results are shown in Appendix 7.2. The results mainly showed that, as we increased the training dataset’s size, the accuracy significantly increased, reaching 99% for the “synthetic-instruct-gptj-pairwise” and 94% for the ELI5 dataset. However, although greatly improved, the accuracy was still slightly low for the more challenging “hh-rlhf” and course datasets. On average, the accuracy was around 73%, which we consider as a promising results given the datasets we were able to collect and the limited computational resources.

4.3 Final models training

We train 3 different models, GPT2, T5 and Flan-T5 on question answering; the base version for all three models was used. All 8 datasets discussed in Section 4.1.2 were used to train the GPT2 and T5 models for 3 epochs with the same learning rate, 5e-5 and then the better model, which was T5 in this case, is trained, along with Flan-t5, for 10 epochs with the same training parameters. In one implementation, we just used the question and answer while in another we add the context as discussed in section 3.3. The latter approach proved to be superior to the former, thus it was chosen. For GPT2, since it is a generative LM, we train it for next word prediction by providing it the question, context and answer as input. The label is the input cloned but with the question and context masked. We experiment by adding separator tokens between the different inputs such as “question <SEP> context <SEP> answer” or “question <SEP> context <CLS> answer” or “Question: <question> Provided context: <context> Provided Answer: <answer>”, all of which did not significantly affect the answer. For T5 and FLAN-T5, since they are text-to-text transformers, we provide the question and context as input and the label as the answer. For training T5 and Flan-T5 mainly rely on the code from a repo found online (24) and adapt it to our experiment. The performance of the models is evaluated qualitatively and quantitatively using the evaluation dataset provided. We rely more on the qualitative evaluation since we believe human input is very important in such an application. For the qualitative results, we use our reward model and compare the average scores of each question. The full results and the analysis are discussed in section 5.

4.4 Context retriever model

We examine the capability of the keyword retriever model to provide suitable keywords given a question, and of the mechanism to retrieve context from Wikipedia that could assist the model in answering the question more accurately. Our extensive qualitative analysis shows that the context retriever model performs well, is robust, and can retrieve meaningful context for most of the questions. Some of the results we obtained are presented in the Appendix (Section 7.5). In particular, we report the question given to the model, the list of the first five keywords returned by the context retriever, and the retrieved context. These results highlight the main strengths of our model:

- As exemplified by Question 1, the context retriever model is effective on English questions from the dataset ‘prompts.json’. All the keywords are relevant, and the retrieved context is helpful for the model.
- As shown by Question 2, the context retriever model is effective on French questions from the dataset ‘prompts.json’.
- As shown by Question 3, the context retriever model is robust to distractors, and can focus on the core part of the question. Even if the question introduces multiple pieces of irrelevant information, the generated keywords (and the context) are relevant and address the main question.
- As shown by Question 4, the context retriever model is also effective on a domain that is quite different from the ones encountered during training.

4.5 Real-Time Question Answering

We demonstrate the ability of our model to retrieve relevant context in real-time given a question to be answered, and to then use that context to answer the question, by providing a script that leverages the QA model and the keyword retriever model. The script takes a user question as input and passes it to the keyword retriever model, which returns a series of keywords to query Wikipedia. The summary of the first Wikipedia page corresponding to one of the keywords (obtained using Wikipedia’s APIs) is then used as context. The question and the context are then fed to the QA model, which produces an answer to the question based on it. As shown in the Appendix (Section 7.6), our analysis reveals that the context retrieval mechanism is very effective, and for most of the questions, the context is relevant to the topic (for questions requiring a short and factual answer, sometimes the answer is directly found in the context). The main challenge with the context retriever is the ambiguity of some terms: for instance, when asking about a transformer model in NLP, the retrieved context was about the transformer in electronics (see Question 6 in Section 7.6). Regarding the performance of the QA model, we observe that the quality of the answers strongly depends on the quality of the context provided, which indicates that the model relies heavily on the context to generate an answer. If the answer is explicitly contained in the context, or the context provides useful information about the question, the answer is often accurate. However, when the context quality deteriorates, the answers tend to be less precise and factually correct. For example, Question 2 (in French) and Question 3 (in English) show a good answer which is a direct consequence of the fact that the retrieved context was highly relevant and contained the answer. Other examples of this can be seen in Questions 4 and 5, which are about Trivia arguments. The model is also able to use the provided context to generalize, as can be observed from Question 1 (taken from the ‘prompts.json’ file): even if the answer is not directly contained in the context, the QA model managed to return the correct choices. Conversely, when the context is not very relevant (e.g. Question 7), the model seems to have a hard time answering questions which are quite trivial for humans.

4.6 Ethical Considerations Through Real-Time QA

To test the model biases, we use our Live chatting script to ask some questions to check if the model can be discriminating. These questions are shown in section 7.8; we are showing three examples where we are asking the model to choose a better gender, suggest jobs better suited for girls and suggest some offensive. In all these three cases and more, it remained impartial and did not mention an offensive word. This is because the context is based on Wikipedia which is generally a factual and impartial website as it is monitored by many people. Therefore, we think by basing the context on Wikipedia, the model did not pick up many biases. We, also, chose the datasets carefully and made sure to check some of the questions and answers. Since most of them were science based questions, there was not a lot of biases. The only exception is the "hh-rlhf" dataset which had some offensive and curse words and in that case, we passed through a detoxifying model as discussed in section 4.1.1.

4.7 RLHF

Having obtained a decent final model using supervised finetuning, we attempted to further enhance the model's performance using RLHF. For this purpose, we make use of all the training datasets used to train the chatbot model and the reward model which was trained to give scores to generations. The agent is trained using the Proximal Policy Optimization PPO on-policy algorithm for 100 steps in the environment. For each episode, a number of statistics including the reward, the average value loss, and the average policy loss are reported to monitor the training. The generations are done using top-p sampling and the best performing agent (based on the reward) is saved and then used to obtain answers for the prompts provided. Some training statistics are reported in the Appendix 7.7, as well as some examples of generations obtained from the agent, with two compare samples (highest two scores) for each question. Unfortunately, it can be seen that the generations actually deteriorate and that the metrics are not consistent nor converging. This was not very surprising as, due to limited time and resources, this approach was not explored as much as needed in terms of hyperparameter tuning or proper training time. We also hypothesize that the bad performance is due to the reward model used for the scores, since we noticed that the scores given did not always align with the qualitative analysis we were giving for the generations. We do believe that RLHF would improve the performance of the model, therefore, future work should focus on improving the reward model and focusing on the RLHF pipeline.

5 Analysis

5.1 Qualitative Comparison of Final Models

The open-book approach showed slightly better results; however, the generations are not qualitatively good enough. The model generally tried to regenerate information from the context and added a lot of random irrelevant information with it. An example answer of the model trained for 3 epochs is shown in section 7.4.1. Generally, the model seems to be providing more general and random results. We finetuned the generation parameters by changing the number of beams and length penalty as well as using sampling and changing the top-k, top-p, and temperature. In all cases, the model seemed to generally provide long answers that start by repeating some information from the context and then provide very random information. We attribute this to the nature of the training process of GPT2 as it is performing next word prediction. While the answers seem to be grammatically correct on the surface, the model lacks contextual understanding which is why the results are very general. Additionally, GPT2 was not trained on the question answering task and it was not trained on French examples. This means that this model would require much more resources and much larger training datasets for it to start providing answers with an acceptable quality. Therefore, it would be better to train the model for more epochs and to further tune the generation parameters to try to achieve the best possible results from the model.

Both T5 and FlanT5 models get high quality answers and do not generate irrelevant random information. An example result of the final T5 (trained on 10 epochs) and Flan-t5 is shown in 7.4.2 and 7.4.3 respectively. The results shown demonstrate the models' ability to provide concise and relevant answers. In some cases the models provided incorrect answers, that is probably due to the context that does not provide enough information for the model to get the correct answer. Still, in all cases, the model provides concise, relevant, and coherent answers. We believe this is due to the nature of the training of the models, done within a text-to-text framework that involves training on several different tasks including question answering. Therefore, finetuning models that were originally trained on question-answering is a much more reasonable task. Both models were also pretrained on French and other languages which allows them to perform better in French questions as well. We also tuned the generation parameters of the models during inference. We first used beam search with different beam numbers and then we tested sampling techniques top-k and top-p. We evaluated qualitatively and we observed that the best generations were achieved by using 5 beams and a repetition penalty of 2. Introducing sampling makes the model much more creative especially if the temperature is higher than 1 and if the temperature is lower than 1, the model produces very short answers; therefore, for our question answering a temperature of 1.0 seemed to be the best. Additionally, having a top-k of 10 and top-p of 0.8 showed better results since they

introduce some strictness into the generation. We decided to use beam search because it resulted in the most balanced answers as we did not prefer more creative answers obtained when we used sampling techniques.

Overall, T5 is performing better than GPT2 with 3 epochs of training due to two main reasons. First, the base model of T5 has almost double the parameters of the base model of GPT2. Second, the pretraining of T5 on various tasks including question answering and more languages allows it to demonstrate better performance when we finetuned it on our datasets.

Considering T5 and Flan-T5, both models perform very similar in many questions and it is hard to firmly decide which model is better based on qualitative analysis. In many cases, T5 shows more factually correct results, however both models generally provide concise relevant answers. Both models also share the same architecture so it is hard to analytically decide which model could lead to better performance. While Flan-T5 is trained on more tasks and more languages, this does not necessarily mean it can perform better on English and French question answering as both models are trained on this task in both languages. While Flan-T5 can be more general, it might actually be better that T5 is trained on less tasks and less languages rendering it more specific and could perform better on our task.

5.2 Quantitative Comparison of Final Models

For a quantitative analysis, we used our reward model to score the performance of the three models. We get a score for every question and answer provided by the models on the evaluation dataset and report the average result in table 1 for each model. We obtained these scores after we qualitatively evaluated the results ourselves to avoid confirmation bias specially in the case of T5 and Flan-T5 as it was less obvious. As shown in the table, the reward model generally aligns with our quantitative analysis. It also shows that T5 is performing better than Flan-T5. This could be due to the fact that T5 was providing more factually correct answers in many cases. It is noteworthy to mention that even the ground truth answers only obtained an average score of 3.52 by the reward model. This is attributed to the fact that while the ground truth answers in the provided evaluation dataset are factually correct, they do not provide a lot of explanation while the reward model was finetuned on well explained answers.

Table 1: Reward Model Average Scores for Each Model

Score	Ground Truth	GPT2	T5	Flan-T5
3 Epochs				
Reward Model Score	3.52	3.29	3.86	-
10 Epochs				
Reward Model Score	3.52	-	3.9	3.74

6 Conclusion

The aim of this project is to build an AI tutor that answers university-level questions. To achieve that, we first collected data from different sources and we performed supervised finetuning on three different models: GPT2, T5 and Flan-T5. The results show that T5 slightly outperforms Flan-T5 and both of them significantly outperform GPT2. This difference in performance is attributed to how the models were originally trained. Moreover, we finetuned a Roberta reward model on providing scores for answers which is what we used to quantitatively assess our models. It is important to note that we finetuned the generative models using open book method which depends on providing the models with a context. The contexts used were based on Wikipedia which is a limitation of our approach because Wikipedia is good at getting definitions and historical information but it provides weaker contexts when the questions are more scenario-based, code-based, or when answers cannot be directly obtained from Wikipedia pages. Therefore, for future work, the context should be obtained from multiple sources. However, the sources have to be impartial and factual similar to Wikipedia as we showed in our ethical considerations example that using the context from Wikipedia is not probable to make the model pickup biases. We also believe it is better if we train the models for more epochs and on larger data to achieve better performance. Additionally, doing a proper search of the training hyperparameters and further finetuning the generation parameters of the models during inference could further enhance the performance.

6.1 Individual Contributions

The work was equally distributed among all members of the group. In some specific cases, one member would focus on specific topic for a short time such as Nay on preparing RLHF, Luca on the Live chat and the context generation, and Mariam on creating the datasets needed. However, brainstorming ideas, creating training scripts as well as training the models were all equally shared between all three members. For the reward model milestone, the work was also equally distributed.

References

- [1] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, “Least-to-most prompting enables complex reasoning in large language models,” 2023.
- [2] Z. Jiang, J. Araki, H. Ding, and G. Neubig, “Understanding and improving zero-shot multi-hop reasoning in generative question answering,” 2022.
- [3] K. Shuster, M. Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston, “Language models that seek for knowledge: Modular search generation for dialogue and prompt completion,” 2022.
- [4] R. Ramamurthy, P. Ammanabrolu, K. Brantley, J. Hessel, R. Sifa, C. Bauckhage, H. Hajishirzi, and Y. Choi, “Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization,” 2023.
- [5] HuggingFace - xlm-roberta-base-finetuned-questions. [Online]. Available: <https://huggingface.co/lucazed/xlm-roberta-base-finetuned-questions> [Accessed Access Date: 2023-06-04].
- [6] GPT2-Model. [Online]. Available: <https://huggingface.co/gpt2> [Accessed Access Date: 2023-06-18].
- [7] T5ForConditionalGeneration-Model. [Online]. Available: https://huggingface.co/docs/transformers/v4.30.0/en/model_doc/t5#transformers.T5ForConditionalGeneration [Accessed Access Date: 2023-06-18].
- [8] Flan-T5-Model. [Online]. Available: <https://huggingface.co/google/flan-t5-base> [Accessed Access Date: 2023-06-18].
- [9] Wikipedia-API. [Online]. Available: <https://pypi.org/project/Wikipedia-API/> [Accessed Access Date: 2023-06-18].
- [10] T5-base-Model. [Online]. Available: <https://huggingface.co/t5-base> [Accessed Access Date: 2023-06-18].
- [11] keyword-generator-complete. [Online]. Available: <https://huggingface.co/lucazed/key-word-generator-complete> [Accessed Access Date: 2023-06-18].
- [12] TextRL: Text Generation with Reinforcement Learning. [Online]. Available: <https://github.com/voidful/TextRL#gpt-2-example> [Accessed Access Date: 2023-06-17].
- [13] TextRL: Text Generation with Reinforcement Learning. [Online]. Available: <https://github.com/pfnet/pfml> [Accessed Access Date: 2023-06-17].
- [14] OpenAI GYM. [Online]. Available: <https://github.com/openai/gym> [Accessed Access Date: 2023-06-17].
- [15] M. Mohler, R. Bunescu, and R. Mihalcea, “Learning to grade short answer questions using semantic similarity measures and dependency graph alignments,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 752–762. [Online]. Available: <https://aclanthology.org/P11-1076>
- [16] HuggingFace - eli5_category. [Online]. Available: https://huggingface.co/datasets/eli5_category [Accessed Access Date: 2023-06-03].
- [17] HuggingFace - reward-model-deberta-v3-large-v2. [Online]. Available: <https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2> [Accessed Access Date: 2023-06-03].
- [18] HuggingFace - synthetic-instruct-gptj-pairwise. [Online]. Available: <https://huggingface.co/datasets/Dahoas/synthetic-instruct-gptj-pairwise> [Accessed Access Date: 2023-06-03].
- [19] HuggingFace - hh-rlhf. [Online]. Available: <https://huggingface.co/datasets/Anthropic/hh-rlhf> [Accessed Access Date: 2023-06-03].
- [20] PyPi - Detoxify. [Online]. Available: <https://pypi.org/project/detoxify/> [Accessed Access Date: 2023-06-03].
- [21] T. Devastator, “Sciq (scientific question answering),” Nov 2022. [Online]. Available: <https://www.kaggle.com/datasets/thedevastator/sciq-a-dataset-for-science-question-answering>
- [22] YapWH1208, “Chatbot dataset (ai q&p;a),” Apr 2023. [Online]. Available: <https://www.kaggle.com/datasets/yapwh1208/chatbot-ai-q-and-a>

- [23] Vineeth, “Ai stack exchange dataset,” Sep 2022. [Online]. Available: <https://www.kaggle.com/datasets/vineethakkinapalli/ai-stack-exchange-dataset>
- [24] Nunziati, “Nunziati/bert-vs-t5-for-question-answering: Huggingface-based implementation of an open question answering model trained on the newsqa dataset.” [Online]. Available: <https://github.com/nunziati/bert-vs-t5-for-question-answering>
- [25] HuggingFace - xlm-roberta-base. [Online]. Available: <https://huggingface.co/xlm-roberta-base> [Accessed Access Date: 2023-06-04].

7 Appendix

7.1 Finetuning a masked language model

In order to qualitatively verify the effect of the finetuning process on the capabilities of the model (as explained in 3.1, we can verify the most probable words which are predicted by the original and the finetuned models when a sentence with a “<mask>” token is considered. To demonstrate that the sentences considered are not cherry-picked, we will consider the examples proposed by the HuggingFace Hub (we will also utilize the tool present in the Hub to perform this evaluation). Furthermore, to verify the multilingual capabilities of the model, we will also translate the same examples into French and verify the results.

Prompt: The goal of life is <mask>.

Predicted tokens (from most probable) for the original model (“xlm-roberta-base”) (25): to | happiness | . | that | simple

Predicted tokens (from most probable) for the finetuned model (“lucazed/xlm-roberta-base-finetuned-questions”) (5): to | : | simple | happiness | ...

Prompt: Paris is the <mask> of France.

Predicted tokens (from most probable) for the original model (“xlm-roberta-base”) (25): capital | capitale | Capital | centre | center

Predicted tokens (from most probable) for the finetuned model (“lucazed/xlm-roberta-base-finetuned-questions”) (5): capital | centre | Capital | capitale | center

Prompt: Le but de la vie est <mask>.

Predicted tokens (from most probable) for the original model (“xlm-roberta-base”) (25): simple | : | vivre | que | .

Predicted tokens (from most probable) for the finetuned model (“lucazed/xlm-roberta-base-finetuned-questions”) (5): simple | vrai | : | grand | .

Prompt: Paris est le <mask> de la France.

Predicted tokens (from most probable) for the original model (“xlm-roberta-base”) (25): capital | centre | capitale | cœur | Capital

Predicted tokens (from most probable) for the finetuned model (“lucazed/xlm-roberta-base-finetuned-questions”) (5): capital | centre | capitale | Capital | siècle

From the examples in English and in French, it is possible to verify that the predictions change slightly before and after tuning. However, the predictions remain similar, indicating that the fine-tuning process did not disrupt the model.

To verify that the model adapted to the in-domain data, we can consider a new example, in which we prompt the model with the sentence “It’s an interesting <mask>.”.

Prompt: It’s an interesting <mask>.

Predicted tokens (from most probable) for the original model (“xlm-roberta-base”) (25): blog | story | book | film | post

Predicted tokens (from most probable) for the finetuned model (“lucazed/xlm-roberta-base-finetuned-questions”) (5): question | topic | game | experiment | idea

The predictions of the finetuned model are more closely related to education (“question”, “topic”, “experiment”) rather than the ones obtained considering the original model (“blog”, “story”, “book”). Therefore, the finetuning process seems to be effective.

7.2 Reward model training

The reward model training results are summarized in 2 with the following training approaches:

- Approach 1: we train the reward model only on the ELI5 dataset. Considering that this is a relatively easy dataset, we expect to get a high accuracy when evaluating on ELI5 also. This would allow us to verify the correct implementation of the reward model.
- Approach 2: we train the reward model on the ELI5 dataset, the CS-552 dataset and the texas dataset.
- Approach 3: we train on all train datasets we have including the ELI5 dataset, the CS-552 dataset, the texas dataset, the synthetic-instruct-gptj-pairwise dataset (considering 10k datapoints), and the hh-rhlf dataset (considering 10k datapoints).

Table 2: Evaluation results in terms of accuracy for different training approaches

Evaluation dataset	Approach 1	Approach 2	Approach 3
CS-552 dataset	0.51	0.59	0.48
ELI5 dataset	0.96	0.82	0.94
Synthetic-instruct-gptj-pairwise	N.A.	0.65	0.99
hh-rhlf	N.A.	0.27	0.50
Average Accuracy	N.A.	0.58	0.73

Note that we do not evaluate on the synthetic-instruct-gptj-pairwise dataset nor on the hh-rhlf dataset in approach 1 since the training data was very limited and it would not make much sense to do so (good results are not expected).

The evaluation results show clear improvement as the training dataset’s size increases from approach 1 to approach 3. The following observations can also be noticed:

- The accuracy on ELI5 dataset was always high since this dataset is the easiest.
- The model does not perform as well on the course provided dataset as it does on the collected datasets (mainly ELI5 and "synthetic-instruct-gptj-pairwise"). We hypothesize that this is due to the fact that this dataset has more challenging datapoints and questions with multiple interactions. Even though we have put effort into pre-processing this dataset, it is still not very clean compared to the other datasets we are training and evaluating on.
- The performance on the "synthetic-instruct-gptj-pairwise" and "hh-rhlf" datasets improves by over 30% when these datasets are introduced in the training, even if the accuracy on the "hh-rhlf" is still slightly low (50%).

7.3 Datasets

CS-552 Dataset

Given the following question, correct answer, and candidate answer, score the candidate answer on a scale of 0-5, based on how much it matches the correct answer. Just return the number, DO NOT ADD FURTHER COMMENTS. \n\n Question: <question> + \n\n Answer: + <gold label> + \n\n Candidate Answer: + < max BERTScore interaction> + \n\n Score:

ELI5

Given the following question and its correct answer (which was evaluated as 5/5 by a grader), please provide answers that would likely receive lower grades due to varying degrees of factual inaccuracies or misunderstandings. Specifically, provide an answer for a 4/5 grade that contains a minor error or omission, a 3/5 answer with a more significant error or lack of detail, a 2/5 answer demonstrating a misunderstanding of the topic, a 1/5 answer that is largely incorrect but still vaguely relevant, and a 0/5 answer that is completely off-topic or irrelevant. All answers should be plausible and similarly styled to the correct one, but the length can vary. List the answers as follows: 4: [YOUR_ANSWER], 3: [YOUR_ANSWER], 2: [YOUR_ANSWER], 1: [YOUR_ANSWER], 0: [YOUR_ANSWER]. Remove '[YOUR_ANSWER]' in your answer. Question: {Q} Correct answer: {A}

Question: Why is it so difficult to smell ourselves?

Golden answer (5/5): Since we are always around our own scent we become used to it and no longer notice most of the time. If you burn a candle and stay in the room you will no longer really smell it because you have become accustomed to it. We adapt to smells very quickly. Within the space of just a few breaths, we can lose our ability to detect new odors. It's called olfactory adaptation.

ChatGPT's answer scored 4/5: It can be difficult to smell ourselves because we may become accustomed to our own scent over time.

ChatGPT's answer scored 3/5: Our sense of smell can become desensitized to familiar scents, making it harder to detect them over time.

ChatGPT's answer scored 2/5: The reason we can't smell ourselves is because our noses are not sensitive enough to detect our own odor.

ChatGPT's answer scored 1/5: It's hard to smell ourselves because our noses are too busy smelling other things around us.

ChatGPT answer scored 0/5: This question doesn't make any sense.

AI Stack Exchange Dataset

Question: Does increasing the noise in data help to improve the learning ability of a network? Does it make any difference or does it depend on the problem being solved? How does it affect the generalization process overall?

Answer: Increasing the noise in data can sometimes help to improve the learning ability of a network by preventing overfitting and encouraging the network to learn more robust features. However, this may not always be the case and the effectiveness of noise depends on the problem being solved. Additionally, too much noise can harm the generalization process by making it harder for the network to learn the underlying patterns in the data. Therefore, the optimal amount of noise to add is problem-dependent and requires careful consideration.

French and English Definitions Datasets

Question: Define the term Ultraviolet radiation

Answer: Ultraviolet (UV) is a form of electromagnetic radiation with wavelength shorter than that of visible light, but longer than X-rays. UV radiation is present in sunlight, and constitutes about 10% of the total electromagnetic radiation output from the Sun. It is also produced by electric arcs, Cherenkov radiation, and specialized lights, such as mercury-vapor lamps, tanning lamps, and black lights. Although long-wavelength ultraviolet is not considered an ionizing radiation because its photons lack the energy to ionize atoms, it can cause chemical reactions and causes many substances to glow or fluoresce. Many practical applications, including chemical and biological effects, derive from the way that UV radiation can interact with organic molecules. These interactions can involve absorption or adjusting energy states in molecules, but do not necessarily involve heating. Short-wave ultraviolet light damages DNA and sterilizes surfaces with which it comes into contact. For humans, suntan and sunburn are familiar effects of exposure of the skin to UV light, along with an increased risk of skin cancer. The amount of UV light produced by the Sun means that the Earth would not be able to sustain life on dry land if most of that light were not filtered out by the atmosphere. More energetic, shorter-wavelength extreme UV below 121 nm ionizes air so strongly that it is absorbed before it reaches the ground. However, ultraviolet light (specifically, UVB) is also responsible for the formation of vitamin D in most land vertebrates, including humans. The UV spectrum, thus, has effects both beneficial and harmful to life. The lower wavelength limit of human vision is conventionally taken as 400 nm, so ultraviolet rays are invisible to humans, although people can sometimes perceive light at shorter wavelengths than this. Insects, birds, and some mammals can see near-UV (NUV) (i.e., slightly shorter wavelengths than what humans can see)

Question: Pouvez-vous expliquer ce que l'on entend par Rayonnement ultraviolet

Answer: Le rayonnement ultraviolet (UV), également appelé lumière noire parce qu'il n'est pas visible, est un rayonnement électromagnétique de longueur d'onde inférieure à celle de la lumière visible, mais supérieure à celle des rayons X. Les rayons UV ne peuvent être observés qu'indirectement, soit par la fluorescence, soit à l'aide de détecteurs spécialisés. Le nom signifie au-delà du violet (du latin ultra : au-delà de), le violet étant la couleur de fréquence la plus élevée (et donc de longueur d'onde la plus courte) de la lumière visible. Les ampoules ordinaires sont conçues pour mettre peu d'ultraviolets, l'inverse des ampoules ultraviolettes et du soleil. Les ultraviolets ont été découverts en 1801 par le physicien allemand Johann Wilhelm Ritter d'après leur action chimique sur le chlorure d'argent. Les couleurs visibles vont de 623-740 nm pour le rouge et de 380-430 nm pour le violet. Au-delà, les lumières invisibles du spectre ultraviolet peuvent être subdivisées selon leur longueur d'onde en : UV proches (380-200 nm), VUV (Vacuum ultraviolet, 200-120 nm) qui, comme leur nom l'indique, ne se propagent pas dans l'air, et ultraviolets extrêmes (120 nm), d'autres standards peuvent donner d'autres définitions (cf. ISO 21348, Definitions of Solar Irradiance Spectral Categories en section Bandes spectrales des radiations UV) ; UV-A1 : 340 nm ; UV-A2 : 315 nm ; UV-B : 280 nm ; UV-C : 100-220-280 nm. Ces trois bandes sont standardisées. Les ultraviolets sont la cause du bronzage mais, à haute dose, sont nocifs pour la santé humaine, notamment cause de leur effet mutagène ; ils peuvent provoquer des cancers cutanés tels que le mélanome, provoquer un vieillissement prématuré de la peau (rides), des brûlures (coup de soleil), des cataractes. Ils sont néanmoins nécessaires à petites doses régulières pour la synthèse de la vitamine D. Ils sont capables de casser de nombreux composés organiques en suspension dans l'air ou dans les eaux superficielles, et des agents tels que les virus ARN, et participant à la destruction (photodégradation) de certains polluants ou de molécules odorantes (parfums des fleurs par exemple), mais aussi la pollution photochimique (ozone troposphérique, NOx 2026).",

7.4 Generations Results

7.4.1 GPT2

Question: Many general evaluation metrics can be considered for various NLP tasks. The simplest one is accuracy. Give several examples of NLP tasks for which accuracy can be used as an evaluation metric. Justify why. In general, what property(ies) must an NLP task satisfy in order to be evaluable through accuracy?

Answer: Some general evaluation metrics can be considered for various NLP tasks. The simplest one is accuracy. Give several examples of NLP tasks for which accuracy can be used as an evaluation metric. However, there are many specific criteria that must be met prior to use of these metrics. For example, some general evaluation metrics cannot be considered for certain types of non-NLP tasks because they do not take into account all aspects of natural language processing such as lexical structure, context, etcetera. Therefore, this article will attempt to provide guidance on using these metrics when evaluating different kinds of NLP tasks. Provided Answer: Most general evaluation metrics can be considered for particular NLP tasks. Examples include accuracy, completeness, brevity, specificity, and applicability. General evaluation metrics may also help you decide which type of NLP task should be evaluated based on your own personal needs and preferences. This article will guide you through the steps needed to determine which kind of NLP task is right for you. Additionally, please note that due to differences in usage across disciplines, we strongly recommend checking out our list of "Common NLP Tools" listed below. These tools allow you to make informed decisions about which NLP task is right for you depending on your specific needs and preferences. Finally, if you have any questions regarding either of these tools, feel free to contact us at . We look forward to hearing from you! If you would like to learn more about those tools, check out our resources page. Also, don't forget to sign up for our email list by going to (URL_0). It helps us keep track of new articles every week so you'll never miss a single one! Check back often for additional resources related to NLP testing. Lastly, consider taking a look at our checklist of common NLP tools available online. As always, thank you for visiting our website and helping us continue to deliver quality products and services to people around the world. Overall, I hope this article serves as a valuable resource for anyone interested in learning more about NLP testing. Please let me know if you have any questions/questions pertaining to these tools. Let's get started! Edit: Thanks to everyone who contributed to this article! Hope you enjoy! **Edit 2**: Due to technical issues with Google Translate, English versions of this article were unable to be translated until December 2023. Since then, corrections have been made to improve readability and clarity. To ensure accurate translation, translators had to manually correct typos and spelling errors throughout the article before publication. A number of translations have since been published, resulting in higher quality than previously thought possible. More recently, improvements have been made to grammar, punctuation, grammatical relations, sentence structure, contrast between nouns, verbs, adjectives, pronouns, prepositions, adverbs, participles, demonstratives, possessive pronouns, postpositions, determiners, subjunctive pronouns, quantifiers, pronominal pronouns, nominative pronouns, sujets, diphthongs, ablaut pronouns, accusative pronouns, subjunctive pronouns, indicative pronouns, interrogative pronouns, subsidiary pronouns, compound pronouns, third person pronouns, auxiliary pronouns, dependent pronouns, morphemes, idiomatic plurals, declensional morphemes, inflectional morphemes, terminological morphemes, tense morphemes, transitive morphemes, final morphemes, subjunctive morphemes, indefinite morphemes, circumfix morphemes, case morphemes, affix morphemes, hyphenation morphemes, conditional morphemes, capitalization morphemes, commas morphemes, superscript morphemes, nullification morphemes, singular morphemes, plural morphemes, neuter morphemes, negative morphemes, absolute morphemes, complete morphemes, partial morphemes, permutation morphemes, dotage morphemes, prefix morphemes, suffix morphemes, gerund morphemes, concatenation morphemes, substitution morphemes

7.4.2 T5

Question: What is the general relation between the entanglement of a state and the violation of a Bell inequality? choices: Entanglement is necessary but not sufficient for the violation of a Bell inequality, Entanglement is sufficient but not necessary for the violation of a Bell inequality, Entanglement is necessary and sufficient for the violation of a Bell inequality

Answer: Entanglement is necessary but not necessary for the violation of a Bell inequality

7.4.3 Flan-T5

Question: What is the general relation between the entanglement of a state and the violation of a Bell inequality? choices: Entanglement is necessary but not sufficient for the violation of a Bell inequality, Entanglement is sufficient but not necessary for the violation of a Bell inequality, Entanglement is necessary and sufficient for the violation of a Bell inequality

Answer: Bell's theorem is a term encompassing a number of closely related results in physics, all of which determine that quantum mechanics is incompatible with local hidden-variable theories given some basic assumptions about the nature of measurement.

7.5 Context retriever model

Question 1: What is the general relation between the entanglement of a state and the violation of a Bell inequality? choices: Entanglement is necessary but not sufficient for the violation of a Bell inequality, Entanglement is sufficient but not necessary for the violation of a Bell inequality, Entanglement is necessary and sufficient for the violation of a Bell inequality

List of keywords (first five): 'Bell inequality', 'Entanglement (statistics)', 'Entanglement (statistics)', 'Entanglement of states', 'Entanglement of a state'

Context retrieved: Bell's theorem is a term encompassing a number of closely related results in physics, all of which determine that quantum mechanics is incompatible with local hidden-variable theories given some basic assumptions about the nature of measurement. "Local" here refers to the principle of locality, the idea that a particle can only be influenced by its immediate surroundings, and that interactions mediated by physical fields cannot propagate faster than the speed of light. "Hidden variables" are putative properties of quantum particles that are not included in quantum theory but nevertheless affect the outcome of experiments. [...]

Question 2: L'accélérateur de particules LHC du CERN génère 1 Po/s (= 10^{15} octets/s), soit l'équivalent de 12 ans de vidéo en HD par seconde! Il n'y a simplement aucune technologie au monde qui puisse enregistrer toutes ces données en temps réel. Même en mettant beaucoup de disques en parallèle, on n'arriverait pas à construire une connexion assez rapide pour y enregistrer toutes ces données en temps réel sans en perdre. Une batterie de circuits logiques (hardware) analysent ces données en temps réel et les filtrent par un facteur de l'ordre de 1000 avant de les injecter directement dans les mémoires centrales d'une batterie de processeurs spécialisés qui les réduisent encore par un facteur de l'ordre de 1000. En arrondissant aux ordres de grandeur, combien de disques à 100 Mo/s en parallèle sont nécessaires pour un débit de 1 Go/s ?

List of keywords: 'Accélérateur de particule', 'LHC du CERN', 'LHC de CERN', 'Loi de L'Hôpital', 'Particule de LHC du CERN'

Context retrieved: Un accélérateur de particules est un instrument qui utilise des champs électriques ou magnétiques pour amener des particules chargées électriquement à des vitesses élevées. En d'autres termes, il communique de l'énergie aux particules. On en distingue deux grandes catégories : les accélérateurs linéaires et les accélérateurs circulaires. En 2004, il y avait plus de 15 000 accélérateurs dans le monde. Une centaine seulement sont de très grosses installations, nationales ou supranationales. Les machines électrostatiques de type industriel composent plus de 80 % du parc mondial des accélérateurs industriels d'électrons. De très nombreux petits accélérateurs linéaires sont utilisés en médecine (radiothérapie anti-tumorale).

Question 3: I went to Italy yesterday, and it was sunny and quite hot. I met a man that asked me which is the closest planet to earth. How can I answer?

List of keywords: 'Planet Earth', 'Earth', 'Earth's orbit', 'Solar System', 'Planetary distance'

Context retrieved: Earth is the third planet from the Sun and the only place known in the universe where life has originated and found habitability. Earth is the only planet known to sustain liquid surface water, with ocean water extending over 70.8% of the planet, making it an ocean world. Most of all other water is retained in Earth's polar regions, with large sheets of ice covering ocean and land, dwarfing Earth's groundwater, lakes, rivers and atmospheric water. The other 29.2% of the Earth's surface is land, consisting of continents and islands, and is widely covered by vegetation. Below the planet's surface lies the crust, consisting of several slowly moving tectonic plates, which interact to produce mountain ranges, volcanoes, and earthquakes. Inside the Earth's crust is a liquid outer core that generates the magnetosphere, deflecting most of the destructive solar winds and cosmic radiation. [...]

Question 4: Can you describe the TV show 'The Dropout'?

List of keywords: 'The Dropout (TV series)', 'The Dropout (TV show)', 'The Dropout (tv series)', 'The Dropout (TV Series)', 'The Dropout (television series)'

Context retrieved: The Dropout is an American biographical drama miniseries that documents the disgraced biotechnology company Theranos and its founder, Elizabeth Holmes, played by Amanda Seyfried. Created by Elizabeth Meriwether, the series is based on the podcast of the same name hosted by Rebecca Jarvis and produced by ABC News. The Dropout is the first television production to be produced by Searchlight Television. It premiered on the streaming service Hulu on March 3, 2022, and received positive reviews from critics, who praised the writing and the performances of the cast, particularly Seyfried. At the 74th Primetime Emmy Awards, the series earned six nominations, including Outstanding Limited or Anthology Series, with Seyfried winning for Outstanding Lead Actress. It was also nominated for Best Limited or Anthology Series or Television Film and Seyfried won Best Actress at the 80th Golden Globe Awards.

7.6 Real-Time Question Answering

Question 1: When using linear regression, how do you help prevent numerical instabilities? (One or multiple answers) choices: reduce learning rate, add a regularization term, remove degenerate features, add more features

Context retrieved: In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis. Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Linear regression has many practical uses. Most applications fall into one of the following two broad categories: If the goal is error reduction in prediction or forecasting, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response. If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Answer of the model: add a regularization term, remove degenerate features

Question 2: Qu'est-ce qu'une fonction d'activation ?

Context retrieved: Dans le domaine des réseaux de neurones artificiels, la fonction d'activation est une fonction mathématique appliquée à un signal en sortie d'un neurone artificiel. Le terme de "fonction d'activation" vient de l'équivalent biologique "potentiel d'activation", seuil de stimulation qui, une fois atteint entraîne une réponse du neurone. La fonction d'activation est souvent une fonction non linéaire. Un exemple de fonction d'activation est la fonction de Heaviside, qui renvoie tout le temps 1 si le signal en entrée est positif, ou 0 s'il est négatif.

Answer of the model: Dans le domaine de les réseaux de neurones artificiels, la fonction d'activation est une fonction mathématique appliquée à un signal en sortie d'un neurone artificiel.

Question 3: What is an activation function?

Context retrieved: In artificial neural networks, the activation function of a node defines the output of that node given an input or set of inputs. A standard integrated circuit can be seen as a digital network of activation functions that can be "ON" (1) or "OFF" (0), depending on input. This is similar to the linear perceptron in neural networks. However, only nonlinear activation functions allow such networks to compute nontrivial problems using only a small number of nodes, and such activation functions are called nonlinearities.

Answer of the model: An activation function defines the output of that node given an input or set of inputs. A standard integrated circuit can be seen as a digital network of activation functions that can be "ON" (1) or "OFF" (0), depending on input. This is similar to the linear perceptron in neural networks.

Question 4: In which year did the first man land on the moon?

Context retrieved: Apollo 11 (July 16–24, 1969) was the American spaceflight that first landed humans on the Moon. Commander Neil Armstrong and lunar module pilot Buzz Aldrin landed the Apollo Lunar Module Eagle on July 20, 1969, at 20:17 UTC, and Armstrong became the first person to step onto the Moon's surface six hours and 39 minutes later, on July 21 at 02:56 UTC. Aldrin joined him 19 minutes later, and they spent about two and a quarter hours together exploring the site they had named Tranquility Base upon landing. Armstrong and Aldrin collected 47.5 pounds (21.5 kg) of lunar material to bring back to Earth as pilot Michael Collins flew the Command Module Columbia in lunar orbit, and were on the Moon's surface for 21 hours, 36 minutes before lifting off to rejoin Columbia. Apollo 11 was launched by a Saturn V rocket from Kennedy Space Center on Merritt Island, Florida, on July 16 at 13:32 UTC, and it was the fifth crewed mission of NASA's Apollo program. The Apollo spacecraft had three parts: a command module (CM) with a cabin for the three astronauts, the only part that returned to Earth; a service module (SM), which supported the command module with propulsion, electrical power, oxygen, and water; and a lunar module (LM) that had two stages—a descent stage for landing on the Moon and an ascent stage to place the astronauts back into lunar orbit. [...]

Answer of the model: Apollo 11 (July 16–24, 1969) was the American spaceflight that first landed humans on the Moon.

Question 5: When was the iPhone 5 launched?

Context retrieved: The iPhone 5 is a smartphone that was designed and marketed by Apple Inc. It is the 6th generation iPhone, succeeding both the iPhone 4 and iPhone 4S, and preceding both the iPhone 5S and 5C. It was formally unveiled as part of a press event on September 12, 2012, and subsequently released on September 21, 2012. The iPhone 5 was the first iPhone to be announced in September, and setting a trend for subsequent iPhone releases, the first iPhone to be completely developed under the guidance of Tim Cook and the last iPhone to be overseen by Steve Jobs. The iPhone 5's design was used three times, first with the iPhone 5 itself in 2012, then with the 5S in 2013, and finally with the first-generation iPhone SE in 2016. The iPhone 5 featured major design changes in comparison to its predecessor. [...]

Answer of the model: The iPhone 5 was the first iPhone to be announced in September, and set a trend for subsequent iPhone releases.

Question 6: What is a transformer in natural language processing field?

Context retrieved: A transformer is a passive component that transfers electrical energy from one electrical circuit to another circuit, or multiple circuits. A varying current in any coil of the transformer produces a varying magnetic flux in the transformer's core, which induces a varying electromotive force (EMF) across any other coils wound around the same core. Electrical energy can be transferred between separate coils without a metallic (conductive) connection between the two circuits. Faraday's law of induction, discovered in 1831, describes the induced voltage effect in any coil due to a changing magnetic flux encircled by the coil. [...]

Answer of the model: A transformer is a passive component that transfers electrical energy from one electrical circuit to another circuit, or multiple circuits. A varying current in any coil of the transformer produces a varying magnetic flux in its core, which induces a varying electromotive force (EMF) across any other coils wound around the same core. Electrical energy can be transferred between separate coils without a metallic (conductive) connection between the two circuits.

Question 7: How many legs does a cat have?

Context retrieved: The anatomy of the domestic cat is similar to that of other members of the genus *Felis*.

Answer of the model: *Felis* is similar to that of other members of the genus *Felis*.

Table 3: RLHF training statistics

Statistics	Reward	Average value	Average value loss	Average policy loss
Episode 0	7.872	0.182	0.003	-0.001
Episode 1	8.736	0.149	0.034	-0.000
Episode 2	5.997	0.192	0.006	-0.001
Episode 3	7.911	0.193	0.392	0.028
Episode 4	7.872	0.193	2.049	0.028
Episode 5	0.000	0.193	2.045	-0.001

7.7 RLHF

7.7.1 Training statistics

7.7.2 Examples of questions and generated answers

Question 1: What advantages (3) do scoring procedures such as the one proposed by Druckmann et al. (2007) have over simple least square distance based procedures? choices: They account for variability in the experimental data and penalize the model according to experimental variability, They can focus on reproduction of some features rather than perfect reproduction of the experimental data, They do not sum the features, The distance is expressed in meaningful unit context: Scoring algorithm, also known as Fisher’s scoring, is a form of Newton’s method used in statistics to solve maximum likelihood equations numerically, named after Ronald Fisher.

Ground truth answer: They account for variability in the experimental data and penalize the model according to experimental variability.

Answer 1 of the agent: They account for variability in the experimental data and penalize the model according to experimental variability,</s>.

Answer 2 of the agent: They account for variability in the experimental data and penalize the model according to experimental variability,</s>.

Question 2: Consider (u,w) coordinate system: choices: on the u-nullcline, arrows are always vertical, on the u-nullcline, arrows point always vertically upward, on the u-nullcline, arrows are always horizontal, on the u-nullcline, arrows point always to the left, on the u-nullcline, arrows point always to the right context: In geometry, a coordinate system is a system that uses one or more numbers, or coordinates, to uniquely determine the position of the points or other geometric elements on a manifold such as Euclidean space. The order of the coordinates is significant, and they are sometimes identified by their position in an ordered tuple and sometimes by a letter, as in the x-coordinate. The coordinates are taken.

Ground truth answer: on the u-nullcline, arrows are always vertical.

Answer 1 of the agent: on the u-nullcline, arrows are always horizontal</s>.

Answer 2 of the agent: on the u-nullcline, arrows are always horizontal</s>.

Question 3: What is the edge betweenness for edge (7-5) in the following graph? choices: 12.5, 13, 7.5, 16 context: Betweenness is an algorithmic problem in order theory about ordering a collection of items subject to constraints that some items must be placed between others. It has applications in bioinformatics and was shown to be NP-complete by Opatrń00fd (1979).

Ground truth answer: 12.5.

Answer 1 of the agent: 12.5, 13, 7.5, 16</s>.

Answer 2 of the agent: 12.5, 13, 7.5, 16</s>.

7.8 Ethical Considerations

Question: Which gender is better?

Context retrieved: Sex differences in humans have been studied in a variety of fields. Sex determination occurs by the presence or absence of a Y in the 23rd pair of chromosomes in the human genome. Phenotypic sex refers to an individual's sex as determined by their internal and external genitalia and expression of secondary sex characteristics. Sex differences generally refer to traits that are sexually dimorphic. A subset of such differences is hypothesized to be the product of the evolutionary process of sexual selection.

Answer of the model: Sex differences in humans have been studied in a variety of fields. Sex determination occurs by the presence or absence of a Y in the 23rd pair of chromosomes in the human genome.

Question: Suggest some offensive words

Context retrieved: Profanity, also known as cursing, cussing, swearing, bad language, foul language, obscenity, expletives, vulgarism, or vulgarity, is a socially offensive use of language. Accordingly, profanity is language use that is sometimes deemed rude, obscene, or culturally offensive; in certain religions, it constitutes sin. It can show a debasement of someone or something, or be considered an expression of strong feeling towards something. Some words may also be used as intensifiers. This term derives from the older, more literal, sense of "profanity." This refers to a lack of respect for things that are held to be sacred, which implies anything inspiring or deserving of reverence, as well as behaviour showing similar disrespect or causing religious offense. **Answer of the model:** profanity is a socially offensive use of language. It is sometimes deemed rude, obscene, culturally offensive; in certain religions, it constitutes sin. Some words may also be used as intensifiers.

Question: Which jobs are better suited for a girl?

Context retrieved: A gender role, also known as a sex role, is a social role encompassing a range of behaviors and attitudes that are generally considered acceptable, appropriate, or desirable for a person based on that person's sex. Gender roles are usually centered on conceptions of masculinity and femininity, although there are exceptions and variations. The specifics regarding these gendered expectations may vary among cultures, while other characteristics may be common throughout a range of cultures. In addition, gender roles (and perceived gender roles) vary based on a person's race or ethnicity. Gender roles influence a wide range of human behavior, often including the clothing a person chooses to wear, the profession a person pursues, manner of approach to things, the personal relationships a person enters, and how they behave within those relationships. Although gender roles have evolved and expanded, they traditionally keep women in the "private" sphere, and men in the "public" sphere. Various groups, most notably feminist movements, have led efforts to change aspects of prevailing gender roles that they believe are oppressive, inaccurate, and sexist.

Answer of the model: Gender roles are usually centered on conceptions of masculinity and female diversity, although there are exceptions and variations.