EECS487 Final Project - **NLP Study Buddy: Summarization and Quiz Generator for informative Texts**

Group 19: Zhixiao Song, Fengyuan Hu, Wenqian Li

Motivation & Problem Definition

As the saying goes, "practice makes perfect." In education, students need high-quality problem sets to help them better absorb the knowledge learned from classes. However, producing such problem sets is a challenging and cumbersome task even for humans. In this work, we present an automated quiz-generation pipeline that takes in a context and generate a quiz from the context.

Prior Work & Initial Attempts

Prior work

There are a plethora of summarization models catered to diverse topics, but our aim is to narrow the focus and incorporate them into quiz questions that can be used to test students' learning outcome. To the best of our knowledge, existing tools focus on only one of the phases in our pipeline and lack the ability to produce high-quality output.

Model Choice

We first consider smaller pretrained encoder-decoder transformer models for the abstractive summarization task and tried to fine-tune them on summarization datasets of our choice. The models we looked at include T5-Small [1], BigBirdPegasus [2], and LED-Base-16384 [3].

Dataset Exploration

We first tried to find separate datasets for summarization and quiz generation. The dataset we chose for summarization including CNN-DailyMail news [4], BBC News Summary [5], and Scientific papers [6].

Quiz Generation

We first tried Blank-filling, True/False quiz generation with ECR and nltk.

Results:

In general, the datasets we chose put a heavy burden on these transformer-based models' context length, and they are computationally challenging to fine-tune. T5-Small struggles with short input context, scoring only 0.09 in ROUGE-2 on the CNN-Daily dataset. BigBirdPegasus, while able to handle longer context length, started producing repeated output after fine-tuning with a small amount of examples. LED-Base-16384 can handle up to 16k tokens, hence better addressed the issue with context limit. It achieved a ROUGE-2 score of 0.13, which is still well below the desired performance. For quiz generation, the results are acceptable but lack of variance. The quiz generated all follows the same pattern.

Methodology

Reflecting on our previous results, we realized that instead of treating the 2 phases (summarization and quiz generation) as separate tasks, **integrating them into one pipeline** might be a better option. Since our method is not focused on tackling long-document processing, we decided to use quiz-generation datasets that have shorter contexts. Inspired by a recent paper by Zhang et al.[7], we decided to use few-shot in-context learning with the LLaMA2-7B model from Meta [8] for both summarization and generation, as illustrated in Fig. 1 (a). This approach first summarizes the long context into a short summary, which helps the LLM to focus on the most important parts in the context. Hence, subsequently generated quizzes will be more aligned with the context. A detailed example can be found at Fig. 2.

Dataset - SQuAD(Stanford Question Answering Dataset) v2

This dataset comprises ~150k answered and unanswered questions from 536 Wikipedia articles. It is a convenient format to incorporate into the pipeline with sufficient reference for both summarization and quiz generation. However, due to the large size, we were only able to train on a subset of the dataset.

4-bit quantization

To enable efficient in-context learning with the LLaMA2 model, we implemented 4-bit quantization using the Hugging Face Accelerate library. Before quantization, loading the model onto the GPU takes ~26GB of VRAM, while only ~3GB of VRAM is needed after quantization. During evaluation, we managed to use a batch size of 8 on 1,204 examples and the model finished in 1.5 hours with a maximum VRAM usage of 60% on a single NVIDIA A40 (48G) GPU.

Control Group

To demonstrate the necessity of the summary in our pipeline, we set up a control group that performs in-context learning directly to quiz generation (Fig. 1 (b)).

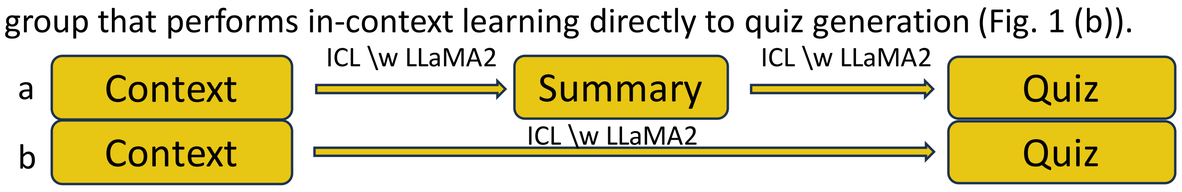

Figure 1. In-context learning structure for experiment (a) & control (b)

Full Pipeline

Context: Along with giving the offender his "just deserts", achieving crime control via incapacitation and deterrence is a major goal of criminal punishment. Brownlee argues, "Bringing in deterrence at the level of justification detracts from the law's engagement in a moral dialogue with the offender as a rational person because it focuses attention on the threat of punishment and not the moral reasons to follow this law." Leonard Hubert Hoffmann writes, "In deciding whether or not to impose punishment, the most important consideration would be whether it would do more harm than good. This means that the objector has no right not to be punished. It is a matter for the state (including the judges) to decide on utilitarian grounds whether to do so or not."

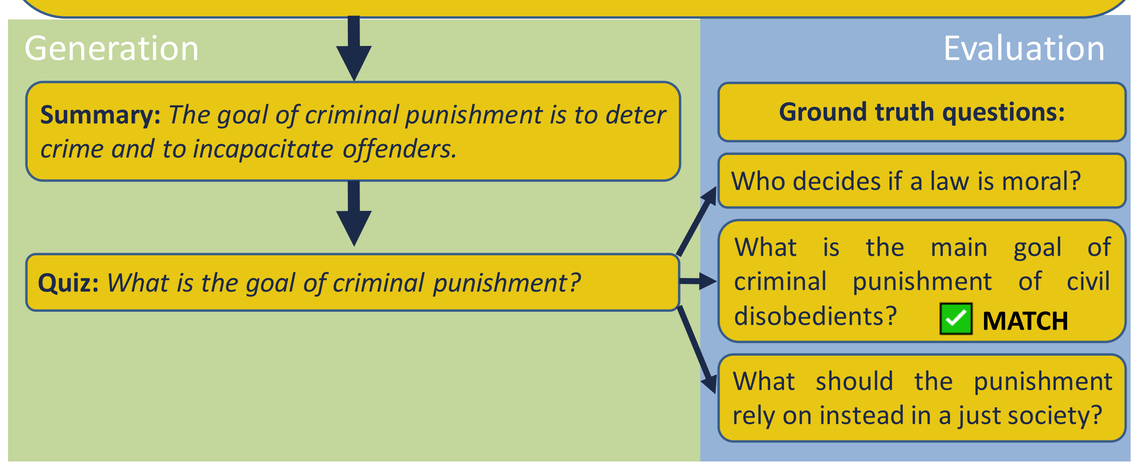

Figure 2. Detailed pipeline with an example

Results

We calculated the mean ROUGE f-measure for 1- and 2-shot experimental groups as well as control groups. The scores are shown in Fig. 3. We found that:

1. Experimental groups outperformed control by 6-18%
2. 2-shot outperformed 1-shot by 17-31%

We also did some case studies investigating the f-measure on six examples, and the results are shown in Fig. 4

Mean ROUGE f-measure	1-shot	2-shot
Experiment	**0.2324**	**0.2733**
Control	0.1971	0.2573

Figure 3. Mean ROUGE f-measure for 1- and 2-shot experiment and control group.

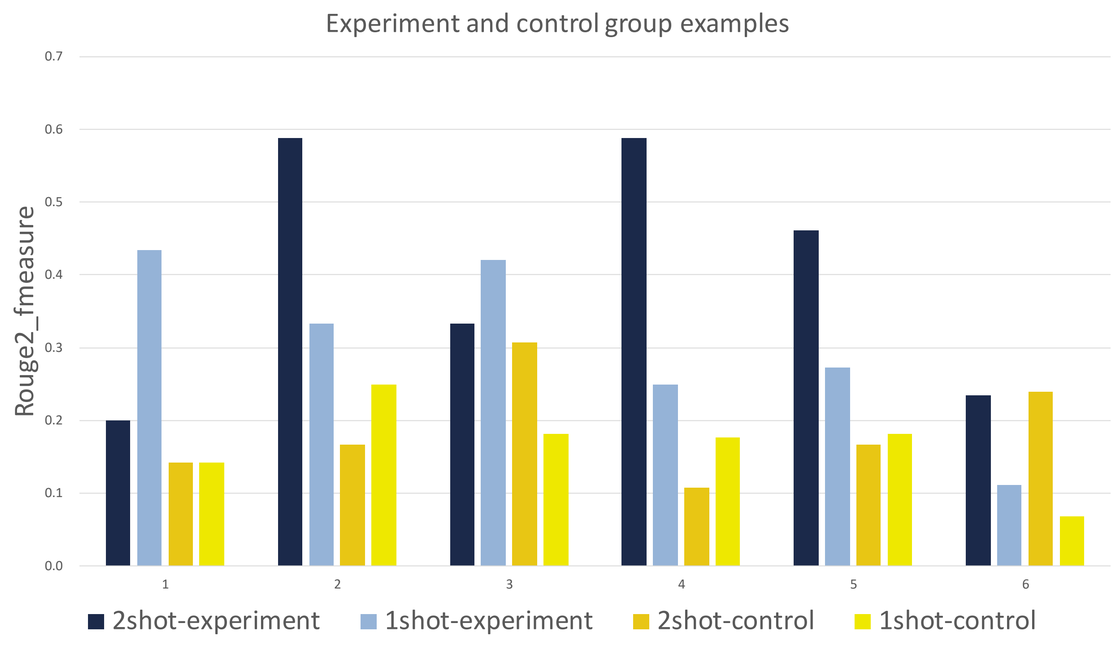

Figure 4. Case studies f-measure

Conclusion

In this work, we tackled the quiz generation problem with in-context learning using the LLaMA2-7B model. We proposed to generate a summary as an intermediate step before generating the quiz and verified with experiments that our method is effective. We hope our work can free some burden for the teachers by automating part of the quiz-generation process and help students learn class materials more concretely.

Reference

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.
2. Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." International Conference on Machine Learning. PMLR, 2020.
3. Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).
4. Zhong, Ming, et al. "Extractive summarization as text matching." arXiv preprint arXiv:2004.08795 (2020).
5. Greene, Derek and Pádraig Cunningham. "Practical solutions to the problem of diagonal dominance in kernel document clustering." Proceedings of the 23rd international conference on Machine learning (2006): n. pag.
6. Cohan, Arman, et al. "A discourse-aware attention model for abstractive summarization of long documents." arXiv preprint arXiv:1804.05685 (2018).
7. Zhang, Zheyuan, et al. "From Heuristic to Analytic: Cognitively Motivated Strategies for Coherent Physical Commonsense Reasoning." arXiv preprint arXiv:2310.18364 (2023).
8. Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).