IN5550 – Spring 2024 — Obligatory 3
Natural Language Generation:
Summarization with Large Language Models (LLMs)

Sushant Gautam and **Fernando Vallecillos Ruiz**

2024/04/12

1 Dataset and pre-processing

1.1 Pre-processing

Although we had employed some prepossessing on the previous assignments, we decided not to apply those to this assignment. The choice was directed by LLM's capabilities for contextual understanding and its subword tokenization, which made us believe that it is capable of handling raw text without preprocessing.

However, we decided to convert jsonl dataset to CSV format for loading to the pandas library for analysis and splitting. The resulting CSV files are in the 'data' folder in GitHub.

1.2 Splitting Dataset

Initially, we imported the dataset from train.csv, subsequently removing any duplicates and instances with missing values to ensure data integrity. To refine the dataset further, we introduced a word_count column by counting the number of words in the summary field, filtering out summaries with fewer than 6 words or more than 200 words, to maintain a focus on moderately detailed summaries. Additionally, we evaluated the length of the article field by creating an article_word_count column, and excluded articles with fewer than 41 words, acknowledging that extremely short articles might not provide sufficient context for our analysis. The row index of jsonl and CSV are the same.

For testing purposes, a subset of 10,000 samples was randomly selected (with a fixed seed for reproducibility) from the processed dataset. The remaining were then split into training and validation sets. This process was carried out with a specific random state to maintain reproducibility in our experiments. However, it was observed that evaluating the model on the entire subset of 10,000 samples was resource-intensive. As a result, we opted to use only the first 1,000 samples from this subset as our test set, striking a balance between thoroughness of evaluation and resource efficiency. The resulting test CSV dataset is in the 'data' folder in GitHub. Also a small note on effect of test set size is presented at Section 4.3.2.

For the training and validation sets, as usual, we employed the train_test_split method from Scikit-Learn, allocating 80% of the remaining data to the training set and 20% to the validation set. The split was stratified based on the distribution of the word_count, which was divided into four bins. This stratification ensured that both the training and validation sets had a representative distribution of article lengths, thereby facilitating a more robust and generalized learning process.

In preparing the dataset from 'feedback.csv' set for our model, we began by removing any duplicate entries and rows with missing values to ensure the dataset's quality and a similar process was employed as detailed before for train.csv.

2 Training Setup

All the experiments were performed on a single A100 GPU. The standard environment from the Fox module has been used. For some experiments, we resorted to external resources but efforts have been put to replicate the same environment as much as possible. The models and checkpoints were loaded from the Hugging Face model repository.

Models used:
OPT: facebook/opt-350m
T5: google-t5/t5-base
Mistral: mistralai/Mistral-7B-v0.1
OPT: facebook/opt-6.7b
BERT: google-bert/bert-large-cased
RoBERTa: FacebookAI/roberta-large
FLAN:google/flan-t5-xxl
Mistral Instruct:mistralai/Mistral-7B-Instruct-v0.2

The above list also represents a taxonomy that we consistently used below to refer to the models with short names as indicated by the bold text above.

3 Task: Generating summaries with large language models

3.1 Baseline: fine-tuned generative LMs

3.1.1 Fine-tuning language model for summarization

The T5 and OPT models were trained using different preprocessing and training approaches tailored to their respective architectures:

T5 Model Training: For T5, the preprocessing involves appending a "summarize: " prefix to articles to guide the model's summarization task. Articles and summaries are tokenized separately, respecting maximum length constraints for input and output sequences. The T5 model is trained using a sequence-to-sequence objective, where the model inputs are the tokenized articles with prefixes, and outputs are the tokenized summaries. This approach leverages Seq2Seq training arguments to optimize sequence generation tasks.

```
i_max_length, o_max_length = 512, 256
```

OPT Model Training: The OPT model's data preprocessing function first tokenizes summaries to avoid truncation, calculates token budgets, then tokenizes the articles within the remaining token budget. The article and summary tokens are then concatenated using a separator: "TL;DR". The model uses causal language modeling (CLM) objective with PEFT for efficient finetuning (r=4, lora_alpha=16), adapting parameters with low-rank updates and a specific dropout strategy to handle high-dimensional embeddings effectively.

```
i_max_length, o_max_length = 350, 350
```

Model Name	Train: Epoch	Loss	Time (s)	Eval: Loss	RMSE	MAE
T5	10	1.476	53446.88	1.639	0.291	0.387
OPT	6	2.684	5355.70	2.890	0.373	0.511

Table 1: Training and Evaluation Metrics for baseline summary models

In both of the two experiments, the models were trained using a learning rate of 1e-4, a per-device training/evaluation batch size of around, a weight decay of 0.01, and had a chance to train for a total of 10 epochs. Additionally, the experiments included an early stopping callback with a patience of 3 and the OPT model also utilized mixed precision training (fp16). For each experiment, the model state in the epoch with the highest rouge score will be saved, therefore, the loss curves contain data (for 3 max epochs) after when the model started to over-fit. The performance metrics tracked during training and evaluation were root mean squared error (RMSE) and mean absolute error (MAE).

Four of these trained models are uploaded in HuggingFace model repository with ID: SushantGautam/opt-350m-lora, SushantGautam/SushantGautam/t5-base.

The training logs are published in Weights and Biases at https://wandb.ai/sushantgautam/nlp_assignment. Click on links in the Model Name column on the table above to see details, loss and score plots for each experiment.

Model	Len	Rouge1	Rouge2	RougeL	BERT: F1	Precision	Recall
T5	32.868	0.4159	0.1999	0.3246	0.3975	0.4143	0.3805
OPT	139.523	0.2135	0.0769	0.1468	0.1261	-0.0155	0.2767

Table 2: Evaluation of baseline summary models on test set

3.1.2 Qualitative analysis of the generated summaries from the best fine tuned model

Throughout this report we have chosen 2 examples to assess the qualities of the summaries generated by the different models. We have will discuss the summaries produced by the best model in this section which is T5.

Article: "' It has been described as 'the poster child' for shoddy British manufacturing and ridiculed for its 'dreadful' handling - but the Austin Allegro has somehow escaped the label of 'worst car ever'. The vehicle came in at number ten on a list voted for by readers of motor magazine Auto Express - with the South Korean-made SsangYong Rodius taking the undesirable top spot. One critic likened its appearance to a 'melted hearse' and said: 'If it were human, not even its mother could love it.' Scroll down for video . Number one: The South Korean-made SsangYong Rodius was voted the worst car of all time . Number two: G-Wiz, pictured being test driven by David Cameron, came in at number two for being 'ugly as sin - a tiny, uncomfortable box' A Top Gear review of the car said: 'Designed by a Brit, the Rodius is a vehicle of such cosmic ineptitude that it must surely have been done via pre-Skype in the days of glitching dial-up modems by a team of visually impaired misanthropes. 'The only thing worse than looking at the Rodius is driving it.' Second on the list of embarrassing motors was the Reva G-Wiz electric car which was described as 'ugly as sin'. Cheap foreign imports have also earned strong places in the list of shambolic vehicles. Bronze medal went to 'ugly duckling' the Chrysler PT Cruiser Cabriolet while fourth on the list was the Polish-made FSO Polonez. Number three: The Chrysler PT Cruiser Cabriolet was slammed by one reader as 'a novelty car that was no fun at all' Number four: The Polish-made FSO Polonez came in at fourth on the list with reader saying it 'lacked any desirability at all' Number five: Rover's City Rover was 'poorly received from the off, with its build quality coming in for particular criticism' Coming in at five on the shameful list was the Rover City Rover which was heavily criticised after being released in 2003. Japan, known for its manufacturing prowess, will no doubt want to forget number six and seven on the list- the Mitsubishi Mirage and Suzuki X-90 - which one reader said had 'zero off-road ability'. Ranked number eight in the worst cars of all time was the Morris Marina, which one reader said had 'terrible handling and poor build quality'. Number nine, the Lada Riva, was also condemned by motor fans with one saying it was 'poor to drive, uncomfortable and basic in the extreme. And last but not least was the Austin Allegro, manufactured by British Leyland from 1973 until 1982. Jeremy Clarkson once candidly said: 'Deciding which one is worse (the Austin Allegro or Morris Marina), is like deciding which leg you'd rather have amputated.' The Allegro was named 'the worst British car ever made' in a 2008 poll by The Sun but Auto Express readers have been kinder to the 1970s model, putting it last in their list. Number six: The Mitsubishi Mirage, another Japanese flop, was said to be overpriced with poor handling . Number seven: The Suzuki X-90 is not a car Japan will be proud of, with readers saying it had 'awful handling' Number eight: The Morris Marina was condemned by Auto Express readers as 'an utterly awful car' Number nine: The Lada Riva, produced in the Eastern Bloc, was said to be 'uncomfortable and basic in the extreme' Number 10: The Austin Allegro was described as 'everything bad about British car manufacturing in the seventies' Below are the top ten worst cars as voted for by Auto Express readers with their comments: . 1) SsangYong Rodius: 'It's time these were taken away and destroyed.' 2) Reva G-Wiz electric car: 'Slow, not very safe and ugly as sin. A tiny, uncomfortable box.' 3) Chrysler PT Cruiser Cabriolet: 'A real ugly duckling. A lack of refinement and driver enjoyment plus poor quality cabin. A novelty car that was no fun at all.' 4) FSO Polonez: 'Cheap, performed poorly and lacked any desirability at all.' 5) Rover City Rover: 'Poorly received from off, with build quality coming in for particular criticism.' 6) Mitsubishi Mirage: 'Poor handling and a very low-rent feel and it doesn't even have a super-cheap price to reflect that.' 7) Suzuki X-90: 'Who wanted a two-seater convertible mini-SUV? Nobody, especially when it had awful handling and zero off-road ability.' 8) Morris Marina: 'Terrible handling and poor build quality meant it was an utterly awful car,' said one reader. 9) Lada Riva: 'Poor to drive, uncomfortable and basic in the

extreme. Virtually any car of the same age will trump this in any category.' 10) Austin Allegro: 'The poster child for everything bad about British car manufacturing in the seventies. Looked awful and handled dreadfully.' "‘

Goal Summary: "‘ Shameful 'worst ever cars' list voted by readers of magazine Auto Express. The South Korean-made SsangYong Rodius came in at number one. One critic commented: 'If it were human, not even its mother could love it' British models Austin Allegro and Morris Marina also featured on the list. Allegro described as 'poster child for everything bad about British car manufacturing in the seventies' "‘

The BERT F1 score for the following summarization is 0.3782.

Predicted Summary: "‘ Motor magazine Auto Express voted for the 'worst cars of all time' The South Korean-made SsangYong Rodius came in at number one. Reva G-Wiz electric car came in at number two for being 'ugly as sin' Mitsubishi Mirage and Suzuki X-90 were also heavily criticised. "‘

The score obtained is slightly lower than the average for the model. The summary is acceptable in its entirety, although it does miss some punctuation. The summary feels more extractive than abstractive. The sentences are disjoint and not very well connected. However, it acknowledges some other cars in the list. Overall, it is an acceptable summary, but omits a substantial amount of details and has very low narrative flow.

Article: "‘ By . Sean O'hare . PUBLISHED: . 07:49 EST, 23 December 2012 . — . UPDATED: . 11:49 EST, 23 December 2012 . Facebook channelled profits through a series of tax havens in order to pay just £2.9m of corporation tax on more than £800m of overseas profits in 2011, it has been reported. Like Google and Apple the social networking site is said to have used its headquarters in Ireland to avoid tax liabilities in the UK before directing earnings to a subsidiary in the Cayman Islands. British companies that buy advertising on Facebook must do so via Facebook Ireland Ltd which entitles the company to sidestep HM Revenue and Customs and authorities in other higher-tax jurisdictions, reported the Sunday Times today. Under the spotlight: Mark Zuckerberg's (pictured) Facebook last year funnelled earnings into the Cayman Islands via Ireland in order to pay only a fraction in tax on more than £800million of overseas profits . As a result less than £240,000 was paid to the UK taxman. The Dublin office, with a staff of 400 people showed a gross profit of £840million in 2011. Despite this, Facebook Ireland posted a loss of £15million for the year after hundreds of millions were routed to a subsidiary in the Cayman Islands and to its parent company in the U.S. Labour MP John Mann recently spoke out about the company's actions, calling them 'disingenuous and immoral'. 'They benefit enormously from the country's internet infrastructure but do nothing to fund it. It's like driving a car with no tax. We would stand for it on our roads so why stand for it on the net?,' he said. Accounts for Facebook Ireland revealed that last year £440m was moved into an Irish sister company before being diverted to a subsidiary company in the Cayman Islands. A spokesman for Facebook defended its accountancy and said it complied with all relevant regulations and was acting within the legal bounds of taxation. 'Facebook complies with all relevant corporate regulations including those related to filing company reports and taxation,' he said. 'We have our international headquarters in Ireland that employs over four hundred people and a series of smaller local offices providing support services all over Europe. Dublin was selected as the best location to hire staff with the right skills to run a multi-lingual hi-tech operation serving the whole of Europe.' Facebook reported £620million in worldwide profit last year, 44 per cent of this came from outside the U.S. The news comes in the wake of similar behaviour from other multinational . companies like Starbucks and Google who use loopholes to avoid taxes. Storm in a coffee cup: Protesters are now planing 'creative direct action' against the company . It will renew pressure on the Government to close the loopholes that . allowed the companies to escape making any contributions to Treasury . coffers through corporation tax –despite raking in billions of pounds in . sales. Starbucks coffee chain is facing a boycott by . thousands of customers who are angry over revelations that it has paid just £8.6million in . corporation tax in its 14 years of trading in the UK and nothing in the . last three. Outraged . protest groups have promised 'direct action' against the coffee . giant and threatened to try to close some branches. Internet giant Google avoided tax on . £10billion revenue last year by doubling the amount of money put into a . shell company in Bermuda. Google's . decision to move nearly 80 per cent of its pre-tax profits to the . company, which is not subject to corporation tax on the Atlantic island, . saw the company slash its overall tax rate almost in half - avoiding . more than £1billion in payments. Chancellor George Osborne pledged to wage war on multinational companies paying little or no tax. In . his

> Autumn Statement, he said he would give HM Revenue and Customs a . further £77million to fight tax avoidance by wealthy individuals and . global firms. News . of the American company's latest legal attempt to avoid paying high . levels of tax comes after MPs recently criticised firms including . Google, who they claim are 'immorally' minimising tax bills. Earlier this year Facebook floated $104 billion of shares in a initial public offering (IPO). Anger : Webgiant Google avoided paying 2$billion in tax last year by moving around 80 per cent of its overseas profits to a company in the tax haven of Bermuda . But the much-hyped flotation fell flat by the closing bell, finishing at only 23 cents more than its expected opening price. Some market experts blamed trading problems and an anxiety-filled half hour where traders were having problems placing and cancelling orders for Facebook stocks. Meanwhile, Zuckerberg – who owns 503.6million shares in his company – is worth an estimated $19.25billion and has become one of the wealthiest people in the world in a single day. Zuckerberg, pictured here at Facebook's flotation earlier this year, is worth an estimated $19.25billion and . became one of the wealthiest people in the world when the company went public . "'

> **Goal Summary:** "' It used Irish office to avoid UK tax liability on earnings from British business. Less than £240,000 was paid to the UK taxman. Money was routed from Ireland into Cayman Islands subsidiary. "'

The BERT F1 score for the following summarization is 0.5164.

> **Predicted Summary:** "' Facebook used headquarters in Ireland to avoid tax liabilities in UK. It funneled £800million of earnings to subsidiary in Cayman Islands. Chancellor George Osborne pledges to wage war on tax avoidance. "'

The score obtained is quite high. The summary is written acceptably without any major errors, just some minor ones such as lack of spaces after numbers. However, the sentences are quite disjointed and sounds more like list generated in an extractive summary. Even if it mentions the £800 millions, it still misses how much they paid which was essential in the goal summary. We can see that although the score is high, we are missing quite a few main points of the article. It could be argued that even the goal summary provided also misses some of the points. This finding stresses the importance of manual verification to assess the quality of the outputs generated.

3.2 In-context learning with pretrained LLMs

Configurations:

```
max_new_tokens=256, do_sample=True
```

3.2.1 Zero-shot summarization

Prompts used:

- #0: The plain-text summary sentence without URLs or lists is : "'
- #1: The main argument/summary in plain text, avoiding URLs and lists is : "'
- #2: The summary in plain text, up to 200 words, without URLs or lists is: "'
- #3: The detailed summary in plain text, including background and implications, without URLs or lists is : "'
- #4: In 200 words, the summary in plain text the who, what, where, when, why, avoiding URLs or lists is : "'

Prompt template:

```
f"Article:{row.article[:2000]}.\n\n{prompt}"
```

Model	Prompt#	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
Mistral	#0	90.23	0.1904	0.0529	0.1267	-0.0965	-0.2361	0.0537
Mistral	#1	96.25	0.1463	0.035	0.0976	-0.2633	-0.3828	-0.1371
Mistral	#2	120.96	0.1879	0.0574	0.1223	-0.0880	-0.2428	0.0775
Mistral	#3	131.48	0.1786	0.0533	0.1165	-0.1106	-0.2767	0.0679
Mistral	#4	94.81	0.143	0.0335	0.095	-0.2797	-0.4210	-0.1292
OPT	#0	89.68	0.1586	0.0364	0.1075	-0.2633	-0.3889	-0.1285
OPT	#1	134.58	0.147	0.0347	0.0969	-0.2029	-0.3596	-0.0336
OPT	#2	116.91	0.1556	0.0431	0.1043	-0.4176	-0.5683	-0.2551
OPT	#3	119.53	0.1334	0.0346	0.0881	-0.3588	-0.5268	-0.1754
OPT	#4	127.79	0.1661	0.044	0.1084	-0.2261	-0.3776	-0.0629

Table 3: Comparison of Zero-shot performance metrics between Mistral and OPT models.

Dependency on the prompt formulation
We can calculate the variability to try to assess the dependence of the models on the different prompts.

Model	Rouge1 Var	Rouge2 Var	RougeL Var	BERT: F1 Var	Prec. Var	Recall Var
Mistral	0.00042	0.00010	0.00017	0.00727	0.00574	0.00962
OPT	0.00013	0.00002	0.00006	0.00667	0.00737	0.00630

Table 4: Variance in performance metrics across prompts for Mistral and OPT models.

Both models exhibit a relatively low dependency on prompt formulation. This variability is more pronounced on metrics such as BERTScore rather than Rouge-related metrics. This may indicate that while the wording has a low impact on basic overlapping metrics, it can result in a larger impact when evaluating semantic quality.

3.2.2 Few-shot summarization

Few-shot (1-shot actually) example used:

> **few_shot_example=** “‘ Given the following example of an article and its summary. Article: “‘The global economy is facing unprecedented challenges due to the combined impact of the COVID-19 pandemic, geopolitical tensions, and climate change. Economic growth has slowed considerably across both developed and developing nations. Supply chain disruptions and inflationary pressures are exacerbating economic inequalities, making it harder for low-income families to afford basic necessities. Governments worldwide are responding with a mix of fiscal stimuli, policy reforms, and support packages aimed at stabilizing markets and fostering sustainable growth.“‘ Summary: “‘The global economy is struggling due to COVID-19, geopolitical tensions, and climate change, leading to slower growth, supply chain issues, and increased inequality. Governments are countering these challenges with various support measures.“‘ ”””

The few-short example above was fused into Prompts from Section 3.2.1.
Prompt template used:

```
f"{few_shot_example}\n\nArticle:{row.article[:2000]}.\n\n{prompt}"
```

Model	Prompt#	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
Mistral	#0	102.75	0.1666	0.0431	0.1126	-0.0547	-0.1761	0.0750
Mistral	#1	102.01	0.1517	0.0356	0.1024	-0.1220	-0.2378	0.0012
Mistral	#2	116.40	0.1768	0.0457	0.1151	-0.0096	-0.1436	0.1339
Mistral	#3	137.89	0.1748	0.0484	0.1141	0.0304	-0.1087	0.1786
Mistral	#4	107.72	0.1793	0.0469	0.1192	-0.0111	-0.1334	0.1191
OPT	#0	116.56	0.0945	0.0082	0.0692	-0.1730	-0.2727	-0.0688
OPT	#1	121.55	0.0944	0.0084	0.0686	-0.1555	-0.2421	-0.0660
OPT	#2	118.43	0.1013	0.0103	0.0722	-0.1438	-0.2359	-0.0481
OPT	#3	126.42	0.0934	0.0090	0.0659	-0.2132	-0.3113	-0.1111
OPT	#4	124.18	0.1021	0.0110	0.0728	-0.1596	-0.2522	-0.0638

Table 5: Comparison of Few-shot performance metrics between Mistral and OPT models.

Reliability of the results and limiting factor

We can calculate the variability to try to assess the dependence of the models on the different prompts. We expand the previous table to compare the variability of the few-shot prompts vs their zero-shot counterparts.

Model	Rouge1 Var	Rouge2 Var	RougeL Var	BERT: F1 Var	Prec. Var	Recall Var
Zero-Shot Mistral	0.00042	0.00010	0.00017	0.00727	0.00574	0.00962
Few-Shot Mistral	0.00010	0.00002	0.00003	0.00269	0.00198	0.00361
Zero-Shot OPT	0.00013	0.00002	0.00006	0.00667	0.00737	0.00630
Few-Shot OPT	0.00001	0.0000012	0.0000064	0.00058	0.00074	0.00044

Table 6: Variance in Zero-Shot and Few-Shot performance metrics across prompts for Mistral and OPT models.

The transition from zero-shot to few-shot learning significantly reduces the variance across all metrics for both models. The reduction clearly indicates that incorporating few-shot examples leads to more stable and predictable model behaviours, minimizing the impact of how the prompts and formulated. However, while few-shot examples can improve the performance and reliability of the models, we should also consider that may introduce bias or overfitting. The examples chosen may not be representative of the broader task. We should also take into account the energy-efficiency aspect. Zero-shot learning offers higher energy efficiency by minimizing computational demands during inference. However, few-shot learning can increase the model performance enough that fewer outputs are needed. Therefore, there is a trade-off between energy conservation and precision.

There is another crucial limiting factor when approaching the topic of few-shot learning, that is context length. Few-Shot learning is bounded by the models' capacity to process only a certain amount of text at once. The selection of representative but short few-short examples to maximize the relevance and diversity of information within the context is crucial. While detailed examples can potentially help generalize the capability of the model, it can result in the model not being able to answer properly.

3.2.3 Qualitative analysis of the generated summaries from the best combination of prompt and model

The best combination is Few-Shot Mistral with prompt #3. A good summarization is the following.

The F1 score for the following summarization is 0.0604.

> **Predicted Summary:** “‘ The detailed summary in plain text, including background and implications, without URLs or lists is : The Austin Allegro was described as "the worst car ever" - but fell out of the top ten of the UK's most shambolic motors that auto enthusiasts chose. But the South Korean-made South Korean-made South Korean-made was voted the worst car of all time, followed by the G-Wiz and the Chrysler PT Cruiser Cabriolet. “‘”””” “‘

The score obtained is higher than the average for this model. The summary is adequately written with good connections between sentences which are also complete. We can see how there is a repetition error at the beginning of the text which will decrease the score. The predicted summary has mistook the 'worst car' as the Austin Allegro while the original article presented it as number 10, which the summary hallucinated it fell out from. It was able to understand the Korean-made car SsangYong Rodius was voted the worst but could not write the name. Furthermore, the summary contains repetition errors.

The F1 score for the following summarization is -0.0555.

> **Predicted Summary:** “‘ This article was originally published at Truthout on April 5, 2012. It is the latest in Truthout's series of articles covering Facebook. Click below to access the full article. . In "Facebook Gouging the Poor to Make Billions for Itself", author Sharon Zhang documents Facebook's use of poor people in developing nations as cheap labor to make it wealthy, while using it's considerable clout with the U.S. government to get the Obama administration to let them off the hook for tax evasion. Truthout has been following the story of Facebook's relationship with the developing world since December 2011. Articles in this series include: "Facebook Profits from the Poor," "Facebook: Taxation Deficit," "Facebook's Secret Tax Haven for Profits," "Facebook:

> Untouchables, and "Facebook's "Morally Reprehensible" Activities Target the Poor". On April 2, 2012 Facebook was ordered to forfeit $40 million and pay approximately $12 million in disgorgement and "‘

The score is lower than the average for this model. The summary is not well written, containing errors in punctuation and also incomplete sentences. In this case, the summary presents topics that are not presented in the original article. This is a case of the summary showing inaccurate claims that are not based on the information presented. This is a clear case where being verbose, or having outputted more tokens, does not equal to more detail or quality in the response.

3.3 Zero-shot prediction with instruction-tuned LLMs

Prompts used:

- #0: Summarize the following article:
- #1: Summarize the main points of the following article, focusing on the key events, findings, or arguments presented:
- #2: Provide a concise summary of this article, highlighting the most important information and conclusions in maximum 200 words:
- #3: Analyze the content of the article, providing a detailed summary that includes the background of the topic, key arguments or evidence presented, and the implications of the findings. Highlight any recommendations or future directions mentioned:
- #4: In maximum 200 words, transform the article into a comprehensive summary, addressing the who, what, where, when, and why. Discuss the significance of the research or events, the main arguments or theories presented, and any counterpoints or criticisms:

Configurations:

```
max_new_tokens=1000, do_sample=True
```

Mistral 7B Instruct prompt template used:

```
f"<s>[INST]{prompt}‘‘‘{row.article}‘‘‘[/INST]"
```

Flan-T5 XXL prompt template used:

```
f"{prompt}-{row.article}\n-Summary:-\n"
```

3.3.1 Zero-shot performance of a conversational finetuned language model

Prompt#	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
0	155.48	0.2317	0.0776	0.1493	0.1942	0.0385	0.3593
1	184.70	0.2127	0.0732	0.1392	0.1662	0.0010	0.3420
2	148.32	0.2313	0.0773	0.1489	0.1967	0.0394	0.3633
3	377.08	0.1228	0.0434	0.0831	0.0493	-0.1614	0.2777
4	281.44	0.1521	0.0529	0.0999	0.0579	-0.1458	0.2779

Table 7: Evaluation Metrics for Different Prompts for Mistral 7B Instruct model on test set.

Metric	Variance
Rouge1	0.00198
Rouge2	0.00020
RougeL	0.00075
BERT: F1	0.00431
Precision	0.00798
Recall	0.00148

Table 8: Variance in Evaluation Metrics for Different Prompts for the Mistral 7B Instruct Model

The Zero-shot Instruction set, designed for conversational tasks, shows a further evolution in model performance. When comparing the results, we should note that there are substantial increases in the performance of BERT-related metrics when compared to its non-finetuned counterpart. The Rouge-related metrics have increased compared to the previous section reaching a new high of 0.1493 for L-Rouge. As for F1 BERT Score, we have achieved a new high of 0.1967.

The variance of the model in the Rouge-related metrics is substantially higher than previously. On the other hand, the difference in BERT-related metrics maintains some degree of similarity. We can assume that instruction-finetuned models are able to keep the same semantic meaning, although the output may look different for overlapping metrics.

3.3.2 Zero-shot performance of FLAN-like dataset finetuned language model

Prompt#	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
0	29.9	0.3239	0.113	0.2327	0.3074	0.3312	0.2841
1	32.029	0.3278	0.1132	0.238	0.3129	0.3342	0.2921
2	29.176	0.3265	0.1189	0.2387	0.3081	0.334	0.2825
3	30.632	0.3202	0.1113	0.2322	0.3028	0.3249	0.2814
4	30.066	0.3169	0.1098	0.2291	0.3041	0.3301	0.2786

Table 9: Evaluation Metrics for Different Prompts for Flan-T5 XXL model on test set.

There are noticeable differences between the Flan-T5 XXL and the Mistral 7B Instruct model. First, there is a substantial difference in the length of the responses generated. Mistral produces considerably longer responses, or, in other words, is more verbose. This is not necessarily an improvement, since longer output does not always result in more detail.

When looking at Rouge-related metrics, Flan-T5 XXL scores consistently higher. This suggests that the model is better at capturing essential content of the original article. Additionally, the model also scores consistently higher when looking at BERT-related metrics. Therefore, this also suggests that it is better at understanding the nuances of language overall.

The difference highlights the effectiveness of Flan-T5 XXL in generating responses that are concise and closely align with the golden annotations in terms of chosen words and meaning. Although Mistral produces longer responses, it may capture irrelevant information or just noise.

3.3.3 Qualitative analysis of the generated summaries from the best combination of prompt and model

For Mistral, prompt #0 and #2 obtains almost the same scores. We choose prompt #0 since it produces slightly better scores on the examples chosen.

The BERT F1 score for the following summarization is 0.2465.

> **Predicted Summary:** “‘ The Austin Allegro, manufactured by British Leyland from 1973 to 1982, was described as the ”poster child for shoddy British manufacturing” but managed to avoid being labeled as the ”worst car ever made.” It came in at number ten on a list of the worst cars of all time as voted for by readers of Auto Express. The South Korean-made SsangYong Rodius took the top spot, with critics comparing its appearance to a ”melted hearse” and describing it as a ”vehicle of cosmic ineptitude.” The Reva G-Wiz electric car came in second, being called ”ugly as sin,” and the Chrysler PT Cruiser Cabriolet and Polish-made FSO Polonez also made the list. The Rover City Rover, Mitsubishi Mirage, and Suzuki X-90 were also included, with readers criticizing their poor handling, cheap quality, and lack of desirability. Finally, the Morris Marina and Lada Riva were condemned for their terrible handling, poor build quality, and basic features. “‘

The score is higher than the average for this model. The summary is on its own well written and the sentences are well connected. Overall, it does a good job summarizing the article and extracting the main points. It also extracted and used double quotes to incorporate the descriptions of the cars. However, the summary starts by pointing out the car ranked 10th before jumping to the first and going into descending order, which may be somewhat confusing.

The BERT F1 score for the following summarization is 0.2193.

> **Predicted Summary:** “‘ Facebook reportedly channelled over £800m of overseas profits through tax havens in Ireland and the Cayman Islands in 2011, resulting in a UK tax payment of only £2.9m. This was achieved by having British companies buy advertising on Facebook via Facebook Ireland Ltd, which allows the company to bypass UK tax authorities. Labour MP John Mann criticized this as "disingenuous and immoral," and called for the government to close tax loopholes. Similar tax avoidance practices have been reported for companies like Starbucks and Google. Facebook reported £620m in worldwide profit that year, with 44% coming from outside the US. The news comes after Chancellor George Osborne pledged to increase HM Revenue and Customs' budget to fight tax avoidance by multinational companies. Facebook's much-hyped IPO fell flat, with Zuckerberg still worth an estimated $19.25billion. “‘

The score is slightly higher than the average for this model. The summary is on its own well written and the sentences are well connected. It makes a somewhat mislead point of an UK tax payment of £2.9m while that refers only to corporate tax. Although the summary is accurate and understands most of the article, it misses the main point of the goal summary about how much Facebook paid to the UK government.

For FLAN-T5 XXL, the prompt #1 scores highest but with a low margin.

The BERT F1 score for the following summarization is 0.2241.

> **Predicted Summary:** “‘ Like Google and Apple, the social networking site was said to have used its HQ in Ireland to avoid paying profits tax. Reports today claimed that the company paid just £2.9m of corporation tax on £800million of overseas profits in 2011. The Dublin office, with a staff of 400 people showed a gross profit of £840million in 2011. However, Facebook Ireland did not record a profit in 2011. This happened because millions of pounds were routed through a subsidiary in the Cayman Islands. Labour MP John Mann recently called the company's actions 'disingenuous and immoral' “‘

The score is lower than the average for this model. The summary is on its own well written and the sentences are well connected, although a missing final full stop. This time, the summary does correctly assess that the tax payment of £2.9m refers to corporate tax. However, it still misses the point about how much Facebook paid to the UK government in total.

The BERT F1 score for the following summarization is 0.1360.

> **Predicted Summary:** “‘ SsangYong Rodius tops the list of shameful vehicles - with Jeremy Clarkson's 'poster boy' taking a distant second. Reva G-Wiz was second out of ten on the list - deemed 'ugly as sin' Rover City Rover at fifth place - branded 'poorly received' for build quality. “‘

As we can see, the score is quite low compared to the average of the model. The summary is on its own well written, although some sentences are poorly connected. It also makes a misleading point by stating that the Austin Allegro is the distant second, when the ranking placed it at the 10th position. While the summary is quite concise, it misses some of the main points of the article.

There are some key differences between the summaries made by both models:

1. Mistral tends to be more verbose outputting longer responses with clear beginning, middle and end. Although FLAN summaries focus more on key points, they are more straightforward and direct. This may constitute nuanced information if deemed irrelevant.

2. Mistral, by being more verbose, can match the tone of the original article, which may be useful to keep the sentiments of the author. FLAN, on the other hand, summarizes to a degree where the tone of the author is mostly lost.

3. Mistral summaries can be recognized as more engaging because they provide the reader with more information on the topic. FLAN offers summaries that are shorter and more concise. This can be useful for readers who just need to grasp the points quickly.

Overall, Mistral tends to produce more detailed and sentimental summaries, while FLAN focuses on conciseness and directness. The choice between the two depends on the reader's needs for depths or efficiency.

4 Task: Improving a generative LLM with a MLM as a scorer

4.1 Fusing scores and normalization

We chose these 2 sets to calculate the correctness score. The reasons for choosing these are:

Accuracy and Coverage: This combination ensures that the summaries are both factually correct and comprehensive, covering the essential points of the original text. Accuracy is non-negotiable for summaries to be trustworthy, while high coverage ensures that the summary is informative and representative. Training a model to optimize for both accuracy and coverage can lead to summaries that are both reliable and useful, serving as a solid foundation for any summarization application.

Overall Quality and Coherence: Choosing overall quality allows the model to capture a broad range of factors that contribute to making a summary good while adding coherence to the mix ensures that the summaries are not only high-quality in content but also in presentation. Coherent summaries are easier to read and understand, enhancing the user experience. This combination could be particularly useful in scenarios where readability and user engagement are critical.

The scores were averaged and then normalized between 0 and 1.

4.2 Training the scorer models

For all four experiments, the models were trained using a learning rate of 2e-5, a per-device training/evaluation batch size of 60, a weight decay of 0.01, and a chance to train for a total of 20 epochs. Additionally, the experiments included an early stopping callback with patience of 3 and utilized mixed precision training (fp16). For each experiment, the model state in the epoch with the lowest evaluation loss will be saved, therefore, the loss curves contain data (for 3 max epoch) after when the model started to over-fit. The performance metrics tracked during training and evaluation were root mean squared error (RMSE) and mean absolute error (MAE).

Four of these trained models are uploaded in HuggingFace model repository with ID:
SushantGautam/roberta-large_accuracy-coverage, SushantGautam/roberta-large_overall-coherence,
SushantGautam/bert-large-cased_accuracy-coverage, SushantGautam/bert-large-cased_overall-coherence.

The training logs are published in Weights and Biases at https://wandb.ai/sushantgautam/nlp_assignment. Click on links in Scorer Model column on the table below to see details, loss and score plots for each experiment.

Scorer Model (score)	Train: Epoch	Loss	Time (s)	Eval: Loss	RMSE	MAE
BERT overall-coherence	7	0.0203	1255.70	0.0324	0.1801	0.1330
BERT accuracy-coverage	12	0.0102	2150.22	0.0416	0.2039	0.1502
RoBERTa overall-coherence	5	0.0278	899.58	0.0343	0.1853	0.1349
RoBERTa accuracy-coverage	7	0.0232	1256.50	0.0345	0.1859	0.1362

Table 10: Training and Evaluation Metrics for BERT and RoBERTa Models

4.2.1 Handling larger input sequences

Large sentences are handled by truncating the article and summary tokens to fit within a predefined maximum length. Truncation prioritizes the summary, ensuring it doesn't exceed half of the budget, while the remainder is allocated to the article. Special tokens are appended if necessary to maintain sequence integrity. This strategy is utilized throughout this assignment for scorer model.

4.3 Applying the scorer model

4.3.1 Generating candidate summaries

We selected the Flan-T5 XXL model from five candidate generations based on its superior evaluation scores as seen in Table 9. The parameters utilized for text generation were as follows: a maximum of 1024 new tokens generated per sequence, with five sequences returned, enabling sampling, selecting from the top 50 tokens, using a top probability cutoff of 0.95, and a temperature of 0.9 for controlling the randomness of the generated text.

Prompt used:
"Provide a concise summary of the text provided, highlighting the most important information and conclusions in maximum 100 words:"

```
max_new_tokens=1024, num_return_sequences=5
do_sample=True, top_k=50, top_p=0.95, temperature=0.9,
```

4.3.2 Estimate the quality of the generated summaries

Scorer Model	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
BERT overall-coherence	26.698	0.3318	0.1285	0.2461	0.3176	0.3548	0.2815
BERT accuracy-coverage	26.762	0.3317	0.1294	0.2471	0.3198	0.3566	0.2841
RoBERTa **overall-coherence**	26.717	0.3304	0.1284	0.2449	0.3208	0.3576	0.2852
RoBERTa accuracy-coverage	27.043	0.3319	0.1309	0.2472	0.3255	0.3637	0.2883

Table 11: Evaluation of scorer models on test set, candidate summaries generated using Flan-T5 XXL model

RoBERTa model trained on (accuracy-coverage) score generally outperforms the others across most metrics, indicating it generates summaries that are more aligned with the reference summaries both in terms of word/phrase overlap and semantic similarity as shown in Table 11. It manages to capture a broader range of information (as indicated by its recall) while maintaining relevance (precision) and balancing detail with brevity (F1 score). The BERT models are slightly behind, particularly in precision, suggesting their summaries might not be as closely matched to the reference summaries in terms of specific details. However, they maintain competitive F1 scores, indicating a decent balance between precision and recall.

4.3.3 Repeating generation and summary evaluation using Mistral-7B-Instruct

Model	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
BERT overall-coherence	118.355	0.2652	0.0866	0.1699	0.2201	0.0955	0.3509
BERT accuracy-coverage	122.38	0.2638	0.0860	0.1678	0.2180	0.0933	0.3488
RoBERTa overall-coherence	94.966	0.2751	0.0873	0.1735	0.2314	0.1191	0.3482
RoBERTa accuracy-coverage	119.678	0.2668	0.0871	0.1696	0.2223	0.0993	0.3514

Table 12: Evaluation of scorer models on test set, candidate; summaries generated using Mistral 7B Instruct model

The Table 12 showcases the evaluation scores of summaries generated using a zero-shot prompt on the "mistralai/Mistral-7B-Instruct-v0.2" model with a similar setup as above, subsequently assessed with four different scoring models. For each scoring model, the best summary out of five was selected for comparison against a reference summary to compute the scores.

Interpreting the results, we observe differences in performance across the various metrics and models used for evaluation. The 'roberta-large_overall-coherence' model achieves the highest scores in terms of Rouge1 and BERT F1 score, indicating a better overall coherence and quality of the generated summaries. In contrast, the BERT models tend to generate longer summaries, as indicated by the 'Len' column but do not perform as well on coherence or accuracy metrics. The precision scores are notably lower across all models, suggesting that the summaries may include relevant information but struggle with conciseness or specificity. The recall scores are comparatively higher, indicating that the summaries are good at capturing the essential information from the source text.

Note on tests set size:
In this experiment, we also wanted to evaluate the impact of test set size. We took first 5K samples of the full test set instead of first 1K in this experiment, generated 5 summaries for each, and recalculated the scores using scorer model. The difference was negligible (so not presented).

4.3.4 Qualitative analysis of at least 2 of the choices

The original response obtains the maximum score for RoBERTa overall-coherence. While the original choice holds a score by RoBERTa accuracy-coverage of 0.657. The candidate obtains 0.682:

> **Original Summary Response:** "‘ Rover's City Rover was 'poorly received from the off, with its build quality coming in for particular criticism' "‘

> **Candidate #2 Summary:** "‘ Rover's City Rover was a 'poorly received from the off' due to poor build quality. "‘

As we can see, the difference in the responses is simply their length. The candidate truncates the response to give a more concise summary simply stating the quality was poor directly.

While the original choice holds a score by RoBERTa accuracy-coverage of 0.598. The candidate obtains 0.677:

> **Original Summary Response:** "‘ Facebook Ireland Ltd and Facebook Limited which operate as separate companies with different names and purposes in different countries around the world. "‘

> **Candidate #3 Summary:** "‘ Facebook Ireland, with its 400 staff, and the Dublin office has reported a gross profit of £840million in 2011 while registering a loss of £15million. "‘

In this case, the difference between the summaries generated is quite substantial. While the first one completely misses the point of the summary, the candidate chosen reports a quite accurate summary stating different quantities such as the number of staff, the gross profit, and the loss of money. This is a clear example how the candidate summaries can differ each other and therefore, underlying the importance of the scorer model.

5 Task: Training an MLM scorer on a synthetic dataset

5.1 Choosing one of the large instruction fine-tuned generative LMs

We selected the Flan-T5 XXL model based on its superior evaluation scores.

5.2 Generating at 5 candidate summaries, conditioned on the input article

The 5 candidates' summaries were generated for every instance of training data using a prompt same as the experiment in Section 4.3.1. The parameters utilized for text generation were as follows: a maximum of 1024 new tokens generated per sequence, with five sequences returned, enabling sampling, selecting from the top 50 tokens, using a top probability cutoff of 0.95, and a temperature of 0.9 for controlling the the randomness of the generated text.

5.3 Assigning the quality score to the generated summary

We use rougeL metric to score the generations against their original summary. The triplets with quality scores are in the file: "data_tmp/5.3synthtic_scored.csv".

5.4 Regression dataset

The rougeL score was made sure to be in the range 0 to 1 and the synthetic regression dataset was ready.

5.5 Concatenating semi-synthetic dataset with the variation of the 'feedback'

From the feedback set, we use overall score' quartile distribution statistics: min: 1.000000, 25%: 4.000000, 50%: 5.000000, 75%: 6.000000, max: 7.000000. This distribution is similar to roughL score distribution in the synthetic data. So we normalize all the scores to between 0 and 1 before concatenating the dataset for training.

Statistics:
total data len = 39402
train len = 31521, **val len** = 7881

Keeping the reference summaries We select all 5000 rows from the train set that will be used in synthetic data generation, extract the article/summary pair and assign a default correctness score of 1 to them before appending to the training regression dataset. They will be used to train the model ending with "with reference" in their name in the table below.

Statistics with reference summary added:
total data len = 44402
train len = 35521, **val len** = 8881
***clarification**: 5000 reference summary from train samples were also added here*

5.6 Fine-tuning Scorer Model

The fine-tuned model is located in: All the configuration for training are exactly same from Section 4.2.

Four of these trained models are uploaded in HuggingFace model repository with ID: SushantGautam/roberta-large_synt_flan, SushantGautam/roberta-large_synt_flan_with_reference_summ, SushantGautam/bert-large-cased_synt_flan, SushantGautam/bert-large-cased_synt_flan_with_reference_summ.

The training logs are published in Weights and Biases at https://wandb.ai/sushantgautam/nlp_assignment. Click on links in Scorer Model column on the table below to see details, loss and score plots for each experiment. Also to be noted here, for each experiment, the model state in the epoch with the lowest evaluation loss will be saved, therefore, the loss curves contains data (for 3 max epoch) after when the model started to over-fit.

Scorer Model	Train: Epoch	Loss	Time (s)	Eval: Loss	RMSE	MAE
RoBERTa	6	0.0167	4423.67	0.0272	0.1650	0.1232
RoBERTa with reference	9	0.0138	7409.56	0.0427	0.2067	0.1423
BERT	7	0.0099	5116.51	0.0285	0.1689	0.1198
BERT with reference	9	0.0122	7680.50	0.0380	0.1948	0.1353

Table 13: Training and Evaluation Metrics for Scorer models using synthetic dataset generated using Flan-T5 XXL model

It was seen that keeping the reference summaries in the regression dataset doesn't improve the performance.

5.7 Generating summaries for test sets using these new scorer models

5 Summaries were already generated for the test set before in Task 4 which was reused here, stored in "data_tmp/4.3generations.csv Each of those was scored using 4 model variants.

5.8 Compare generated summaries to the ones from the previous sections

Scorer Model	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
RoBERTa	25.446	0.3465	0.1398	0.26	**0.3359**	0.3787	0.2941
RoBERTa with reference	23.856	0.3419	0.1388	0.2577	0.3338	0.3851	0.2840
BERT	25.870	0.3444	0.1415	0.26	0.3308	0.3717	0.2910
BERT with reference	25.470	0.3365	0.1349	0.2546	0.3227	0.3662	0.2803

Table 14: Evaluation Scores for Scorer models using synthetic dataset generated using Flan-T5 XXL model on test set

In the samples explored, the ranking obtained through ROBERTa was equal, or very similar, to the one obtained in the previous section. As mentioned previously, we have already shown that the use of scorer model can introduce a substantial difference in the final results of summaries generated given that the summaries are generated with enough diversity. While training with synthetic datasets is a great way to fight data shortage, we must also keep in mind that models can quickly learn their own patterns and overfit; therefore, it is crucial to always evaluate with a test set.

5.9 Repeating generation and summary evaluation using Mistral-7B-Instruct

The same experiment above was re-done for Mistral-7B-Instruct model and reported:

Scorer Model	Train: Epoch	Loss	Time (s)	Eval: Loss	RMSE	MAE
RoBERTa	14	0.0063	7045.43	0.02077	0.14412	0.09053
RoBERTa_with_reference_summ	7	0.0119	3977.40	0.01969	0.14032	0.08843
BERT	8	0.0092	4012.33	0.02534	0.15919	0.10057
BERT_with_reference_summ	7	0.0093	3982.68	0.02504	0.15825	0.09851

Table 15: Training and Evaluation Metrics for Scorer models using synthetic dataset generated using Mistral 7B Instruct model

Scorer Model	Len	Rouge1	Rouge2	RougeL	BERT: F1	Prec.	Recall
RoBERTa	90.497	0.2875	0.096	0.1869	0.242	0.132	0.356
RoBERTa with reference	92.922	0.2831	0.0935	0.1838	0.239	0.127	0.355
BERT	91.675	0.2799	0.0897	0.1813	0.234	0.125	0.347
BERT with reference	100.504	0.273	0.09	0.1757	0.229	0.113	0.351

Table 16: Evaluation Scores for Scorer models using synthetic dataset generated using Mistral 7B Instruct model on test set

6 Inference on the blind set

The final submission was prepared by generating five summaries using the FLAN-T5 XXL model with the same promops used in Section 3.2.2, scoring each summary using a RoBERTa scorer model trained without the reference summary included from Section 5.8 and selecting the summary with the maximum ROUGE-L score out of the five generated summaries for each prompt.

Link to Github Repository