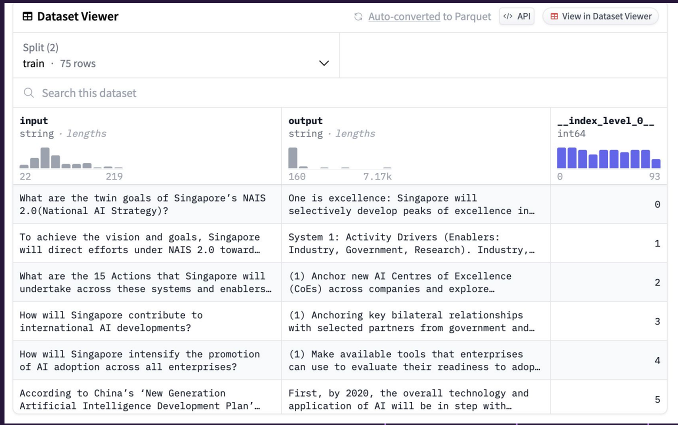
AI PolicyChat

94812 - Team#5 Report

Joey Xie
Cleon Sun
Valerie Yuan
Abhijit Verma
Jiaying Qian
Yan Luo

Data Collection

- We started by collecting **20** AI policy documents from various trusted sources listed on our course's Canvas page;
- Made Questions: We came up with around **100** questions that cover a broad range of AI policy topics;
- Found Answers: For each question, we found an answer in our collected documents.
- Pushed to Hugging Face hub.


Dataset Viewer

Auto-converted to Parquet · API · View in Dataset Viewer

Split (2)
train · 75 rows

Search this dataset

input string · lengths 22 – 219	output string · lengths 160 – 7.17k	__index_level_0__ int64 0 – 93
What are the twin goals of Singapore’s NAIS 2.0(National AI Strategy)?	One is excellence: Singapore will selectively develop peaks of excellence in…	0
To achieve the vision and goals, Singapore will direct efforts under NAIS 2.0 toward…	System 1: Activity Drivers (Enablers: Industry, Government, Research). Industry,…	1
What are the 15 Actions that Singapore will undertake across these systems and enablers…	(1) Anchor new AI Centres of Excellence (CoEs) across companies and explore…	2
How will Singapore contribute to international AI developments?	(1) Anchoring key bilateral relationships with selected partners from government and…	3
How will Singapore intensify the promotion of AI adoption across all enterprises?	(1) Make available tools that enterprises can use to evaluate their readiness to adop…	4
According to China’s ‘New Generation Artificial Intelligence Development Plan’…	First, by 2020, the overall technology and application of AI will be in step with…	5

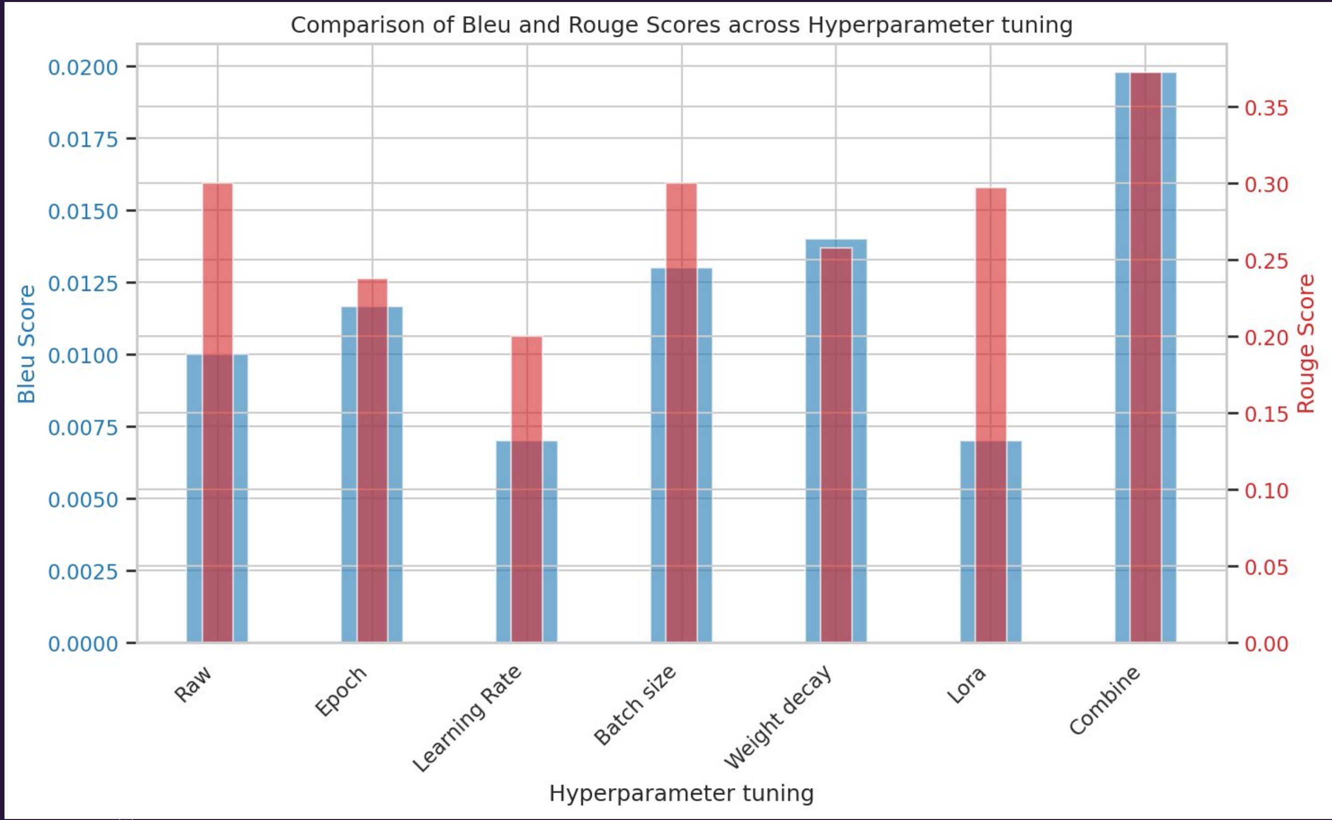
Policy Chat

- **Question:**
  Why is international cooperation on AI important?

- **Raw Llama:**
- </s><s> nobody knows
- International cooperation on AI is important for several reasons:
- 1. **Sharing knowledge and resources**: By working together, countries can share their expertise and resources in AI, leading to faster progress and more innovative solutions.
- 2. **Addressing ethical and social implications**: AI raises complex ethical and social implications, such as privacy, bias, and job displacement. International cooperation can help address

- **After Fine tuning:**
- International cooperation on AI is important because it can help address global challenges, such as climate change, pandemics, and economic inequality, more effectively than any one country could alone. It can also help ensure that AI is developed and used in ways that are ethical, transparent, and respectful of human rights.

Hyperparameter Tuning



Baseline Model

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.012922	0.307788	0.077700	0.256869
Fine_tuned_llama2_batch	0.005380	0.309104	0.067997	0.264924

```
# Quantization Config
quant_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.float16,
    bnb_4bit_use_double_quant=False
)
```

```
# LoRA Config
peft_parameters = LoraConfig(
    lora_alpha=16,
    lora_dropout=0.1,
    r=8,#rank
    bias="none",
    task_type="CAUSAL_LM"
)

# Training Params
train_params = TrainingArguments(
    output_dir="./results_modified",
    evaluation_strategy = "epoch",
    num_train_epochs=30,
    per_device_train_batch_size=4,
    gradient_accumulation_steps=1,
    optim="paged_adamw_32bit",
    save_steps=100,
    logging_steps=20
    learning_rate=2e-4,
    weight_decay=1e-3,
    fp16=False,
    bf16=False,
    max_grad_norm=0.3,
    max_steps=-1,
    warmup_ratio=0.03,
    group_by_length=True,
    lr_scheduler_type="constant",
    report_to="tensorboard"
)
```

Learning Rate

Critical role of learning rate in Fine-Tuning

- Determines how effectively the model adapts to new data
- Balances knowledge retention with new information assimilation
- Prevents overfitting to maintain generalization

Experimentation with different learning rates (1e-4 & 2e-4)

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.006988	0.287063	0.056603	0.235454
Fine_tuned_llama2_lr_1e_4_metrics	0.007629	0.205463	0.037827	0.183335
Fine_tuned_llama2_lr_2e_4_metrics	0.005543	0.229239	0.032119	0.204028

Lora Hyper-parameter

What Lora is & What We Chose

- lora_alpha: This parameter controls the scaling factor applied to the low-rank matrices in LoRA. It is used to adjust the magnitude of the updates to the original model parameters.
- lora_dropout: This parameter specifies the dropout rate applied to the low-rank matrices in LoRA. Dropout is a regularization technique used to prevent overfitting by randomly setting a fraction of the input units to 0 during training.
- r: This parameter defines the rank of the low-rank matrices in LoRA. The rank determines the number of columns in the low-rank matrices, which in turn controls the amount of parameter reduction.
- bias: This is a boolean parameter that indicates whether to include bias terms in the low-rank adaptation.

lora_alpha	lora_dropout	r	bias
32	0.2	8	none

How Lora Performs

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.012922	0.307788	0.077700	0.256869
Fine_tuned_llama2_Lora	0.007254	0.297749	0.066084	0.252001

Number of Epochs

Base model:

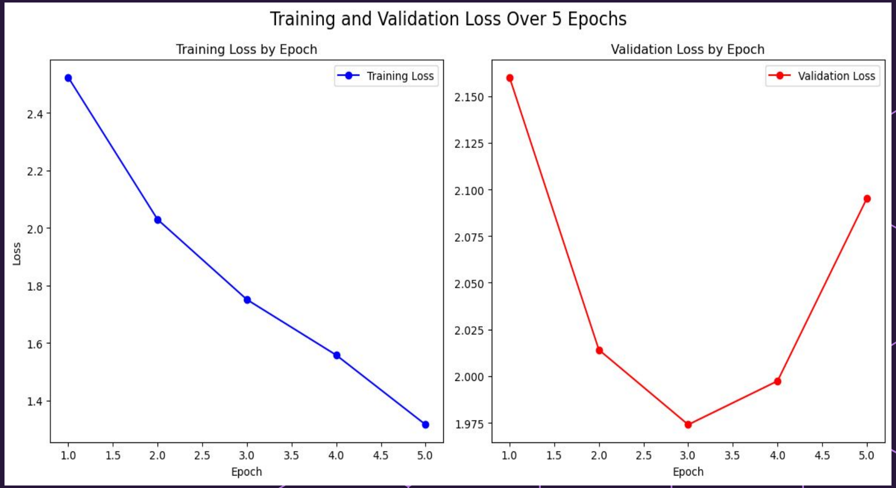
`NousResearch/Llama-2-7b-chat-hf`

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.006171	0.298435	0.062240	0.256231
Fine_tuned_llama2_3_epoches	0.007181	0.331065	0.088585	0.261162

Base model:

`meta-llama/Llama-2-7b-chat-hf`

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.002833	0.243284	0.046212	0.213330
Fine_tuned_llama2_3_epoches	0.011653	0.238402	0.047726	0.192272



num_train_epochs=3 is the best.

Batch Size

Gradient accumulation = 2

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.012922	0.307788	0.077700	0.256869
Fine_tuned_llama2_batch	0.013182	0.312330	0.271695	0.237278

Larger batch size usually makes gradient approximation more accurate and stabilize training. Our baseline batch size is 4. Because T4 **GPU RAM** is limited to 14 GB, large batch like 8 cannot fit into the GPU.

Gradient accumulation is a technique used to train models with large mini-batches that cannot fit entirely into the GPU memory at once.

We can backward the gradient after two gradients of two small batches are calculated and add them to get a large batch gradient, and update the model.

Weight Decay

How does weight decay affect the model? L2 Regularization.

The weight decay parameter is a regularization technique used during the training of neural networks, and it serves multiple purposes:

- Preventing Overfitting
- Improving Generalization
- Stabilizing Training

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.007532	0.227682	0.043049	0.197577
Fine_tuned_llama2_control_codes_prompting	0.014546	0.258092	0.059738	0.219957

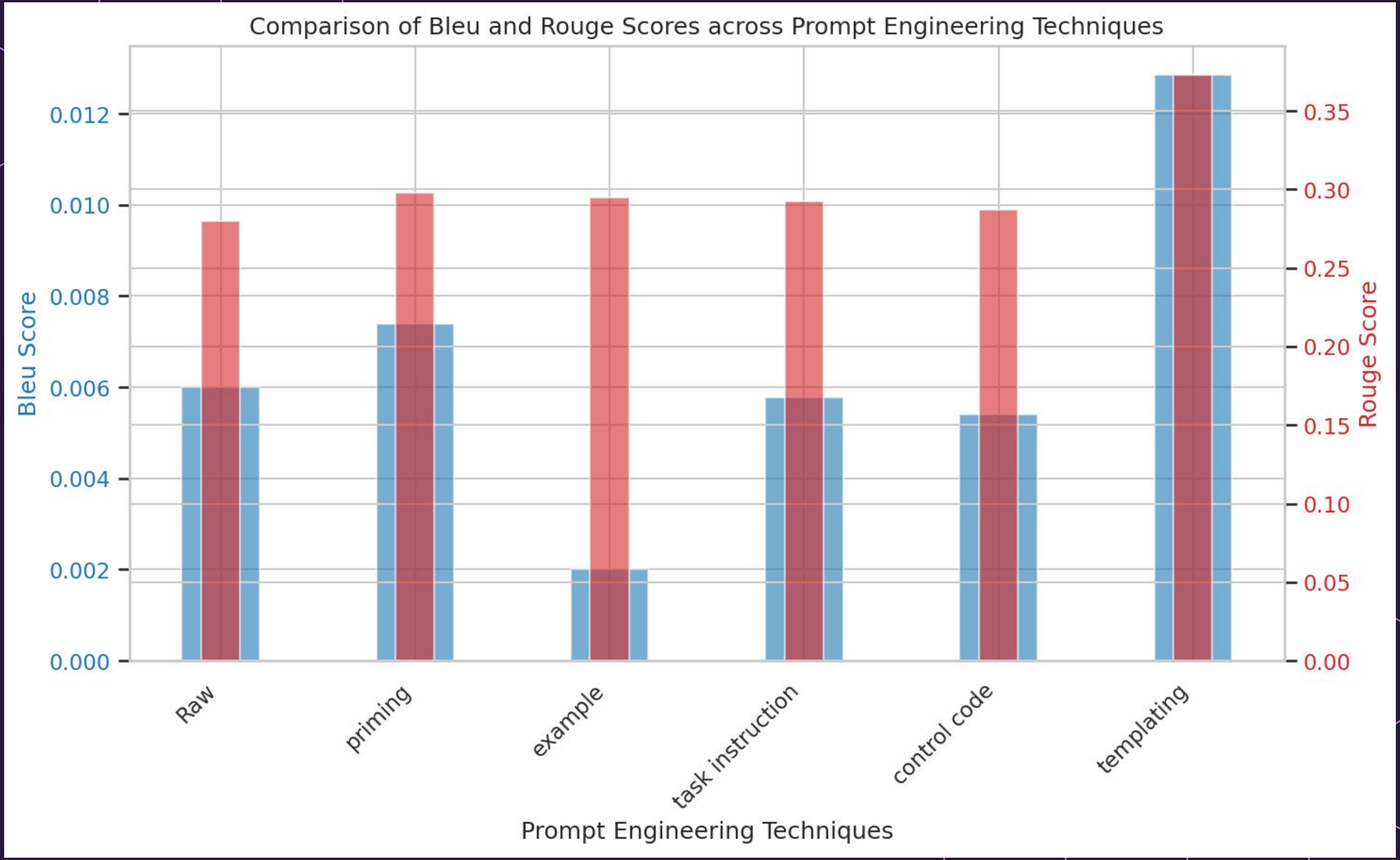
Our experiments with varying weight decay values from 0.0001 to 0.1 during the training of the LLaMA model demonstrated that this parameter had little effect on the training loss, indicating a low influence on the model's learning across the dataset used.

Combine hyperparams tuning

Lora_rank=16,
Gradient Accumulation=2,
Learning_rate=1e-5,
Weight_decay=2e-3,
Max_grad_norm=0.1,
Epoch=10

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.012922	0.307788	0.077700	0.256869
Fine_tuned_llama2_combineHyp	0.019834	0.299497	0.080271	0.257430

Prompt Engineering



Examples

Adopting fixed example method for Examples prompting technique

- Mitigates risks of biased guidance in model responses
- Provides a consistent template for quality and format
- Ensures stylistic uniformity without content influence

How it performs:

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.006988	0.287063	0.056603	0.235454
llama2_examples_prompting	0.002679	0.295252	0.057654	0.233252

Priming

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
generations_without_priming	0.006171	0.298435	0.062240	0.256231
generations_with_global_priming	0.007318	0.364842	0.085737	0.293611

```
from tqdm.notebook import tqdm
import gc

priming_text = """
The following responses should reflect a deep understanding of AI Policy, including ethical considerations,
regulatory frameworks, and the societal impact of AI technologies.
Answers should be informed, nuanced, and precise, demonstrating a comprehensive grasp of the subject matter.
"""

generations_primed =[]
for i in tqdm(range(len(df_test_all)), "generating..."):
  prompt = f"{priming_text}### Question: {df_test_all['input'][i]}\n Briefly, in 100 words answer the question. ### Answer: </s>"
  # Generate predictions
  inputs = llama_tokenizer(prompt, return_tensors='pt')
  inputs = inputs.to("cuda")
  output = raw_model.generate(**inputs, max_new_tokens=200,temperature=0.2)
  response = llama_tokenizer.decode(output[0].tolist())
  # print(response)
  # break
  generations_primed.append(response)
  del inputs, output
  gc.collect()
  torch.cuda.empty_cache()
```

Task Instruction

What Task Instruction Is:

Task Instruction: **Providing specific task instructions** showed a balanced improvement across all metrics, suggesting that task instruction might be a more effective approach than individual priming in certain contexts.

In the experiment, before the question, we add “Based on your knowledge, please answer the question:”

How it performs:

	Experiment	BLEU Score	ROUGE-1	ROUGE-2	ROUGE-L
0	Baseline	0.006728	0.213823	0.035865	0.176707
1	Task Instruction	0.005774	0.292351	0.051474	0.255964

Control Codes Prompting

What Control Codes Prompting Is:

Control Codes: **Special instructions or tokens** we prepend or append to our input prompt to guide the model's generation towards a desired format, tone, or content type.

For the Llama-2 7b model, this involves specifying the type of response needed (e.g., factual, analytical, policy recommendation) or indicating the focus area (e.g., AI ethics, regulation frameworks, implementation strategies).

Example: [International Cooperation] Why is international cooperation on AI important?

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.007532	0.227682	0.043049	0.197577
Fine_tuned_llama2_control_codes_prompting	0.014546	0.258092	0.059738	0.219957
prompt_engineering_control_codes_metrics	0.005434	0.287454	0.051924	0.245689

Templates

What Control Codes Prompting Is:

For example:
The original prompt is: “How does the European Union classify AI systems under its AI Act, and what are the implications for "high risk" AI systems?”

Template:
Discuss the **main concerns** raised by critics regarding the European Union's AI Act, particularly focusing on its **potential impact on innovation and competitiveness** within the EU. Analyze how the regulation might **hinder or foster** technological advancement and the competitiveness of European AI industries on a global scale.

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
Raw_llama2	0.006728	0.213823	0.035865	0.176707
templating_prompt_engineering_metrics	0.012847	0.372629	0.091972	0.314735

Evaluation Metrics

Automatic Metrics

- BLEU
- ROUGE

Human Evaluation

- Relevance
- Coherence
- Our process of human evaluation across different fine-tuning experiments and prompt engineering techniques:
  - https://docs.google.com/spreadsheets/d/17-JDgSJSxjbH-zk6wxtRN3Y2OmibVuHddKCz7MqQaMQ/edit#gid=700621598

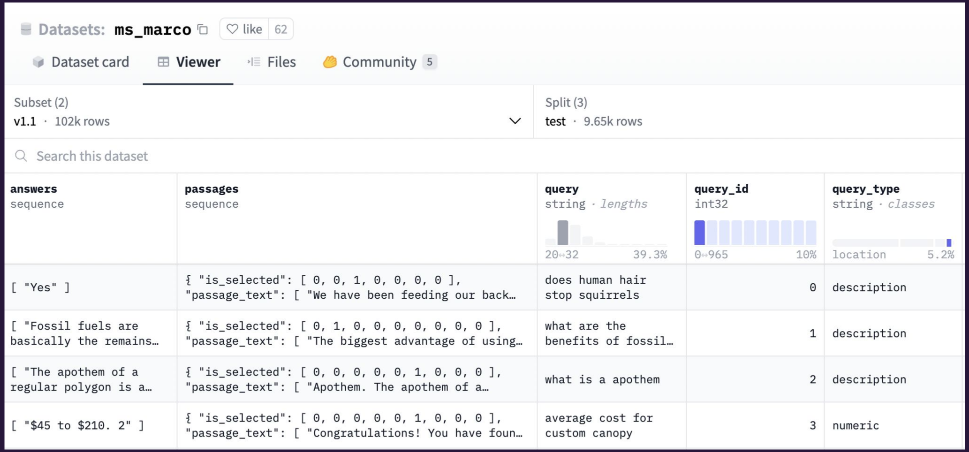
Benchmarking datasets and results

ms_marco

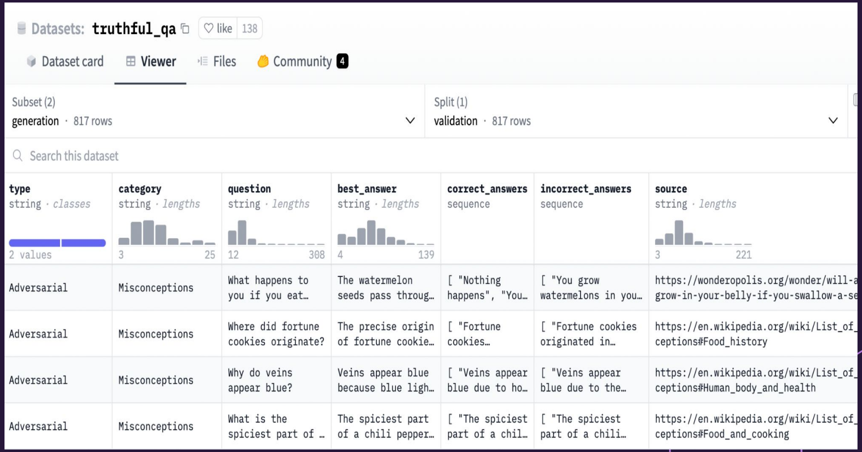
a question answering dataset featuring 100,000 real Bing questions and a human generated answer

truthful qa

comprises 817 questions that span 38 categories, including health, law, finance and politics


Datasets: ms_marco like 62
Dataset card Viewer Files Community 5
Subset (2) v1.1 · 102k rows
Split (3) test · 9.65k rows
Search this dataset

answers sequence	passages sequence	query string · lengths 20↔32 39.3%	query_id int32 0↔965 10%	query_type string · classes location 5.2%
["Yes"]	{ "is_selected": [0, 0, 1, 0, 0, 0, 0], "passage_text": ["We have been feeding our back…	does human hair stop squirrels	0	description
["Fossil fuels are basically the remains…	{ "is_selected": [0, 1, 0, 0, 0, 0, 0, 0, 0], "passage_text": ["The biggest advantage of using…	what are the benefits of fossil…	1	description
["The apothem of a regular polygon is a…	{ "is_selected": [0, 0, 0, 0, 0, 1, 0, 0, 0], "passage_text": ["Apothem. The apothem of a…	what is a apothem	2	description
["$45 to $210. 2"]	{ "is_selected": [0, 0, 0, 0, 0, 1, 0, 0, 0], "passage_text": ["Congratulations! You have foun…	average cost for custom canopy	3	numeric


Datasets: truthful_qa like 138
Dataset card Viewer Files Community 4
Subset (2) generation · 817 rows
Split (1) validation · 817 rows
Search this dataset

type string · classes 2 values	category string · lengths 3 25	question string · lengths 12 308	best_answer string · lengths 4 139	correct_answers sequence	incorrect_answers sequence	source string · lengths 3 221
Adversarial	Misconceptions	What happens to you if you eat…	The watermelon seeds pass throug…	["Nothing happens", "You…	["You grow watermelons in you…	https://wonderopolis.org/wonder/will-a grow-in-your-belly-if-you-swallow-a-se
Adversarial	Misconceptions	Where did fortune cookies originate?	The precise origin of fortune cookie…	["Fortune cookies…	["Fortune cookies originated in…	https://en.wikipedia.org/wiki/List_of_ ceptions#Food_history
Adversarial	Misconceptions	Why do veins appear blue?	Veins appear blue because blue ligh…	["Veins appear blue due to ho…	["Veins appear blue due to the…	https://en.wikipedia.org/wiki/List_of_ ceptions#Human_body_and_health
Adversarial	Misconceptions	What is the spiciest part of …	The spiciest part of a chili pepper…	["The spiciest part of a chil…	["The spiciest part of a chili…	https://en.wikipedia.org/wiki/List_of_ ceptions#Food_and_cooking

Benchmarking datasets and results

Different models excel on different datasets:

MS_MARCO

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
llama	0.002696	0.197467	0.037944	0.194967
llama_chat	0.000000	0.148073	0.014250	0.133303
finetuned_llama2_chat	0.000000	0.145014	0.006926	0.133466

TRUTHFUL QA

	BLEU_Score	ROUGE-1	ROUGE-2	ROUGE-l
llama	0.000000	0.054168	0.000000	0.049168
llama_chat	0.019834	0.342512	0.149015	0.306943
finetuned_llama2_chat	0.019288	0.352340	0.174474	0.333537

Conclusion & Our Findings

Best hyperparameter combination

- Lora_rank=16,
- Gradient Accumulation=2,
- Learning_rate=1e-5,
- Weight_decay=2e-3,
- Max_grad_norm=0.1,
- Epoch=10

Best prompting techniques

- Templates

Balancing BLEU and ROUGE Metrics in model performance is challenging

- akin to precision and recall

Thanks!