Results - 17/07/24

Model	Average Permissive Accuracy	Average Exact Accuracy	ROUGE Score	BLEU Score	METEOR Score	BERTScore Precision	BERTScore Recall
LLAMA2-7B chat baseline							
LLAMA3-8B baseline	0.33	0.05	0.20	0.03	0.19	0.836	0.876
LLAMA2-7B chat Fine tuned	0.46	0.00	0.15	0.05	0.21	0.778	0.892
LLAMA3-8B Fine tuned	0.67	0.15	0.36	0.06	0.31	0.850	0.919
LLAMA3-8B few shot + cot	0.59	0.15	0.36	0.07	0.31	0.854	0.917
LLAMA3-8B Selective Few Shot Prompting	0.64	0.11	0.27	0.06	0.28	0.809	0.913
LLAMA3-8B Selective Few Shot Prompting: with random examples and not selective	0.64	0.11	0.27	0.06	0.28	0.807	0.912
~~LLAMA3-8B {Zero shot, one shot, five shot.. }~~	Seperate table						
Tool use Langchain [150 samples]	-	0.75	0.30	0.23	0.33	0.822	0.878

requirements.txt

Prompt Type	Average Permissive Accuracy	Average Exact Accuracy	ROUGE Score	BLEU Score	METEOR Score	BERTScore Precision	BERTScore Recall
Zero-shot	0.58	0.16	0.39	0.06	0.29	0.872	0.917
One-shot	0.58	0.15	0.37	0.06	0.28	0.869	0.915
Two-shot	0.58	0.15	0.37	0.06	0.28	0.856	0.901
Three-shot	0.45	0.04	0.22	0.06	0.23	0.650	0.695
Four-shot	0.43	0.04	0.20	0.06	0.22	0.598	0.640
Five-shot	0.39	0.03	0.18	0.06	0.20	0.517	0.554

Langchain:

```
Processing queries: 100%|██████████| 150/150 [00:00<00:00, 383.10it/s]

Average Accuracy: 0.75
ROUGE Score: 0.30
BLEU Score: 0.23
METEOR Score: 0.33
BERTScore Precision: 0.822
BERTScore Recall: 0.878
BERTScore F1: 0.849
GLEU Score: 0.23

examples:

Context: Cloud & Cognitive Software * Recast to reflect segment changes. ** 2019 results were
impacted by Red...
Question: What caused the increase in Cloud & Cognitive Software revenue in 2019?
Prediction: The increase in Cloud & Cognitive Software revenue in 2019 was caused by: There wa
s strong growth in Cloud & Data Platforms, as reported and at constant currency, driven primar
ily by the acquisition of Red Hat in the third quarter of 2019. Red Hat had continued strong p
erformance since the acquisition, in Red Hat Enterprise Linux (RHEL), application development
and emerging technologies, led by OpenShift and Ansible.

Reference: There was strong growth in Cloud & Data Platforms, as reported and at constant curr
ency, driven primarily by the acquisition of Red Hat in the third quarter of 2019. Red Hat had
continued strong performance since the acquisition, in Red Hat Enterprise Linux (RHEL), applic
ation development and emerging technologies, led by OpenShift and Ansible.
Accuracy: 1
---------------------

Context: Cloud & Cognitive Software * Recast to reflect segment changes. ** 2019 results were
impacted by Red...
Question: What percentage of Cloud & Cognitive Software external revenue was Transaction Proce
ssing Platforms in 2019?
Prediction: Percentage of Cloud & Cognitive Software external revenue that was Transaction Pro
cessing Platforms in 2019: 34.21%

Reference: 34.21
Accuracy: 1
---------------------

Context: .

s&p supercap data processing & outsourced 100.00 68.26 99.41 97.33 118.68 151.90 item 6 . sele
cte...
Question: what is the roi of an investment in fidelity national information services from 2007
to 2008?
Prediction: ROI of an investment in Fidelity National Information Services from 2007 to 2008:
-29.92

Reference: -29.9%
Accuracy: 0
---------------------
```

```
Context: united parcel service , inc . and subsidiaries notes to consolidated financial statem
ents 2014 ( con...
Question: what portion of the balance of unrecognized tax benefits as of december 2017 will im
pact the effective tax rate?
Prediction: Portion of unrecognized tax benefits impacting the effective tax rate: 37.74647887
323944

Reference: 37.7%
Accuracy: 0
--------------------

Context: table of contents totaled an absolute notional equivalent of $ 292.3 million and $ 19
0.5 million , r...
Question: what portion of the presented investments is due within 24 months?
Prediction: Portion of investments due within 24 months: 74.15836590593872

Reference: 74.2%
Accuracy: 0
--------------------

Context: 30. EMPLOYEE BENEFIT PLANS (cont.) The above sensitivities are hypothetical and shoul
d be used with ...
Question: How much would the pension change if there is a 1% increase and a 1% decrease respec
tively in the future salary growth?
Prediction: Pension change with a 1% increase in future salary growth: $7572, and with a 1% de
crease in future salary growth: $-6919

Reference: $7,572, $(6,919)
Accuracy: 0
```

---- Results for llama3.1 8b fine tuned -----
Average Accuracy with Leniency: 0.67
Average Exact Accuracy: 0.13
ROUGE Score: 0.31
BLEU Score: 0.06
METEOR Score: 0.30
BERTScore Precision: 0.830
BERTScore Recall: 0.918
BERTScore F1: 0.870
GLEU Score: 0.06

- Fine tuned on with 1 relevant example---- Results for 1 relevant example prompt-----
  Average Permissive Accuracy : 0.62
  Average Exact Accuracy: 0.12
  ROUGE Score: 0.32
  BLEU Score: 0.06
  METEOR Score: 0.28
  BERTScore Precision: 0.841
  BERTScore Recall: 0.914

BERTScore F1: 0.874
GLEU Score: 0.06

- Fine tuned on with 1 random example ---- Results for 1 random example prompt -----
  Average Permissive Accuracy : 0.63
  Average Exact Accuracy: 0.13
  ROUGE Score: 0.33
  BLEU Score: 0.06
  METEOR Score: 0.29
  BERTScore Precision: 0.840
  BERTScore Recall: 0.915
  BERTScore F1: 0.874
  GLEU Score: 0.06

- Fine tuned with relevant, random, zero and relevant+random combined examples ---- Results 1 -----
  Example type distribution:
  both: 341
  relevant_only: 304
  random_only: 325
  none: 317
  Average Permissive Accuracy : 0.61
  Average Exact Accuracy: 0.14
  ROUGE Score: 0.34
  BLEU Score: 0.06
  METEOR Score: 0.28
  BERTScore Precision: 0.852
  BERTScore Recall: 0.916
  BERTScore F1: 0.881
  GLEU Score: 0.07
- Fine tuned with relevant, random, zero and relevant+random combined examples ----
- Results 2 -----
  Example type distribution:
  both: 147
  relevant_only: 324
  random_only: 499
  none: 317
  Average Permissive Accuracy : 0.61
  Average Exact Accuracy: 0.14
  ROUGE Score: 0.34
  BLEU Score: 0.06
  METEOR Score: 0.28
  BERTScore Precision: 0.853
  BERTScore Recall: 0.915
  BERTScore F1: 0.881
  GLEU Score: 0.07
- Fine tuned with relevant, random, zero and relevant+random combined examples----
- Results 3-----
  Example type distribution:
  both: 338
  relevant_only: 616
  random_only: 197
  none: 136

Average Permissive Accuracy : 0.61
Average Exact Accuracy: 0.14
ROUGE Score: 0.35
BLEU Score: 0.06
METEOR Score: 0.28
BERTScore Precision: 0.855
BERTScore Recall: 0.916
BERTScore F1: 0.883
GLEU Score: 0.07

```
#inference type selection
example_type_choices = ['both', 'relevant_only', 'random_only', 'none']
Result 1: example_type_weights = [0.25, 0.25, 0.25, 0.25]
Result 2: example_type_weights = [0.10, 0.25, 0.40, 0.25]
Result 3: example_type_weights = [0.25, 0.50, 0.15, 0.10]
```

Prompting Analysis

```
LLAMA3.1-8B --
Processing completed. 221 samples were filtered out  <--------> no code generated
Total samples selected: 1000
Remaining samples: 779
Percentage of samples with execution errors: 31.58%
Average Permissive Accuracy: 0.23
Average Exact Accuracy: 0.18
ROUGE Score: 0.25
BLEU Score: 0.05
METEOR Score: 0.14
BERTScore Precision: 0.841
BERTScore Recall: 0.874
BERTScore F1: 0.856
GLEU Score: 0.06
Results have been saved to filtered_test_data_1000.json

LLAMA3.1-8B-Instruct --
Processing samples: 100%|██████████| 1000/1000 [2:42:32<00:00,  9.75s/sample]The change in emplo
Processing completed. 157 samples were filtered out.
Total samples selected: 1000
Remaining samples: 843
Percentage of samples with execution errors: 7.12%
Average Permissive Accuracy: 0.35
Average Exact Accuracy: 0.28
ROUGE Score: 0.41
BLEU Score: 0.14
METEOR Score: 0.21
BERTScore Precision: 0.866
BERTScore Recall: 0.898
BERTScore F1: 0.881
GLEU Score: 0.14
Results have been saved to filtered_test_data_1000.json
```

```
-- llama3.1 8b on the same 1000 samples without Code--

Average Permissive Accuracy : 0.40
Average Exact Accuracy: 0.30
ROUGE Score: 0.39
BLEU Score: 0.17
METEOR Score: 0.32
BERTScore Precision: 0.868
BERTScore Recall: 0.898
BERTScore F1: 0.882
GLEU Score: 0.18
```

test samples