Evaluating the Performance of Large Language Models for Turkish Text Summarization

Emirhan Erdoğan
Dept. of Computer Engineering
Marmara University
Istanbul, Turkey
emrhnerdogan01@gmail.com

Şule Koca
Dept. of Computer Engineering
Marmara University
Istanbul, Turkey
sulekoca.23@gmail.com

Faruk Akdemir
Dept. of Computer Engineering
Marmara University
Istanbul, Turkey
akdemirfaruk.1@gmail.com

Alper Özdemir
Dept. of Computer Engineering
Marmara University
Istanbul, Turkey
aalper.ozdemirr@gmail.com

Eren Başpınar
Dept. of Computer Engineering
Marmara University
Istanbul, Turkey
ishakeren2001@gmail.com

Abstract*— *This paper presents a comprehensive analysis of text summarization performance using large language models (LLMs) evaluated through advanced metrics. It combines insights from two reports focusing on zero-shot and one-shot performance, and the effectiveness of various evaluation metrics, including ROUGE, BLEU, METEOR, BERTScore and FrugalScore.

Keywords*— *Text Summarization, Large Language Models, Zero-Shot Learning, One-Shot Learning, Evaluation Metrics.

I. INTRODUCTION

Text summarization is a critical process in natural language processing (NLP) aimed at generating concise and accurate summaries from longer documents. This task is essential across various domains such as news aggregation, academic research, and the legal field. Large Language Models (LLMs) are designed to perform highly in NLP tasks, and their performance is measured using various metrics. This paper examines advanced evaluation metrics for LLMs and presents the evaluation results of different models.

II. TASK DESCRIPTION

The downstream task evaluated in this report is the HellaSwag Turkish (hellaswag_tr) task. This task is a benchmark designed to test natural language understanding and commonsense reasoning capabilities of models. The task involves selecting the most plausible continuation of a given scenario from multiple choices.

The evaluation was conducted under two conditions: zero-shot and one-shot settings. In the zero-shot setting, the model makes predictions without any prior examples [1], while in the one-shot setting, the model is given one example to guide its predictions [2]. The primary evaluation metrics used are accuracy (acc) and normalized accuracy (acc_norm), along with their respective standard errors (Stderr).

III. EVALUATION

The following models were evaluated:

1- gemma-2b-CSE4078S24_Grp8-r8-4bit-tr (r = 8, lora_alpha = 8)
2- gemma-2b-CSE4078S24_Grp8-r16-4bit-tr (r = 16, lora_alpha = 16)
3- gemma-2b-CSE4078S24_Grp8-r32-4bit-tr (r = 32, lora_alpha = 32)
4- gemma-2b-CSE4078S24_Grp8-r64-4bit-tr (r = 64, lora_alpha = 64)
5- gemma-2b-CSE4078S24_Grp8-r128-4bit-tr (r = 128, lora_alpha = 128)
6- gemma-2b-CSE4078S24_Grp8-r256-4bit-tr (r = 256, lora_alpha = 256)

IV. ADVANCED EVALUATION METRICS

- **BLEU:** Calculates n-gram overlaps between generated text and reference texts, suitable for short sentences [3].
- **ROUGE:** Measures n-gram word sequence and word pair overlaps for text summarization and generation [4].
- **METEOR:** Considers synonyms and word stems, producing human-aligned results [5].
- **BERTScore:** Evaluates similarity using deep learning models, capturing contextual meaning [6].
- **FrugalScore:** FrugalScore is an evaluation metric for natural language generation (NLG) tasks that balances accuracy with computational efficiency. By training a low-cost model to approximate the results of more resource-intensive metrics, FrugalScore retains most of the original performance while significantly reducing computational costs. This makes it ideal for large-scale or real-time applications where traditional metrics would be impractical. In our text summarization project, using FrugalScore allowed us to assess the quality of generated summaries efficiently and reliably [7].

V. EVALUATION RESULTS

Zero-shot Setting:

Model Name	acc	Stderr	acc_norm	Stderr
gemma-2b-CSE4078S24_Grp8-r8-4bit-tr	0.3323	±0.0149	0.3771	±0.0153
gemma-2b-CSE4078S24_Grp8-r16-4bit-tr	0.3313	±0.0149	0.3781	±0.0153
gemma-2b-CSE4078S24_Grp8-r32-4bit-tr	0.3303	±0.0148	0.3751	±0.0153
gemma-2b-CSE4078S24_Grp8-r64-4bit-tr	0.3284	±0.0148	0.3672	±0.0152
gemma-2b-CSE4078S24_Grp8-r128-4bit-tr	0.3144	±0.0147	0.3632	±0.0152
gemma-2b-CSE4078S24_Grp8-r256-4bit-tr	0.3134	±0.0146	0.3493	±0.0150

One-shot Setting:

Model Name	acc	Stderr	acc_norm	Stderr
gemma-2b-CSE4078S24_Grp8-r8-4bit-tr	0.3323	±0.0149	0.3642	±0.0152
gemma-2b-CSE4078S24_Grp8-r16-4bit-tr	0.3323	±0.0149	0.3672	±0.0152
gemma-2b-CSE4078S24_Grp8-r32-4bit-tr	0.3294	±0.0148	0.3652	±0.0152
gemma-2b-CSE4078S24_Grp8-r64-4bit-tr	0.3284	±0.0148	0.3662	±0.0152
gemma-2b-CSE4078S24_Grp8-r128-4bit-tr	0.3234	±0.0148	0.3642	±0.0152
gemma-2b-CSE4078S24_Grp8-r256-4bit-tr	0.3134	±0.0146	0.3463	±0.0150

Below are the performance results of six different models evaluated using ROUGE, BLEU, BERT, METEOR, QaEval, and FrugalScore metrics [8]:

Model	ROUGE-1 Mean	ROUGE-2 Mean	ROUGE-L Mean	BLEU Mean	BERT Mean	METEOR Mean	QaEval Mean	Frugal Score Mean
Model 1	0.218	0.130	0.170	0.035	0.835	0.230	0.311	0.654
Model 2	0.220	0.133	0.172	0.036	0.840	0.236	0.317	0.659
Model 3	0.211	0.125	0.165	0.034	0.837	0.220	0.307	0.652
Model 4	0.252	0.139	0.196	0.035	0.879	0.212	0.292	0.715
Model 5	0.306	0.161	0.248	0.037	0.872	0.153	0.196	0.708
Model 6	0.259	0.124	0.212	0.021	0.853	0.107	0.142	0.692

Model	ROUGE-1 Std	ROUGE-2 Std	ROUGE-L Std	BLEU Std	BERT Std	METEOR Std	QaEval Std	FrugalScore Std
Model 1	0.132	0.113	0.115	0.059	0.159	0.200	0.386	0.136
Model 2	0.129	0.112	0.112	0.057	0.151	0.200	0.387	0.133
Model 3	0.131	0.111	0.113	0.057	0.140	0.200	0.385	0.148
Model 4	0.132	0.126	0.118	0.073	0.063	0.189	0.379	0.075
Model 5	0.159	0.154	0.152	0.098	0.052	0.168	0.321	0.068
Model 6	0.140	0.122	0.128	0.064	0.057	0.130	0.288	0.061

VI. ANALYSIS AND DISCUSSION

The evaluation results across the models indicate a modest level of performance in both zero-shot and one-shot settings.

Zero-shot Analysis:

- The models show a slight decline in accuracy as the rank (r) increases. This might suggest that increasing the rank beyond a certain point does not yield proportional improvements and might even degrade performance [8].
- The normalized accuracy values are consistently higher than the raw accuracy, indicating that the models are relatively better at ranking the correct answers.

One-shot Analysis:

- The accuracy values in the one-shot setting are quite similar to those in the zero-shot setting, with very little to no improvement observed.
- The normalized accuracy shows a slight increase in the one-shot setting, suggesting a marginal benefit from the additional example.
- Overall, the r8 and r16 models performed slightly better compared to higher rank models. This indicates that a moderate rank and lora_alpha might be optimal for this specific task and dataset.

VII. CONCLUSION

The evaluation of the GEMMA-2B derived models on the HellaSwag Turkish task demonstrates that the models achieve a baseline performance, with the accuracy values ranging around 0.33. The normalized accuracy values are slightly higher, reflecting the models' ability to rank choices effectively.

The results suggest that increasing the rank (r) and lora_alpha values beyond a certain point does not necessarily improve performance and might even be detrimental. The r8 and r16 models appear to be the most effective in this evaluation, balancing performance and computational efficiency.

Further research could involve fine-tuning the models on task-specific data and experimenting with different configurations to enhance performance. Additionally, exploring other downstream tasks could provide a broader understanding of the models' capabilities and limitations.

REFERENCES

[1] https://www.techopedia.com/definition/34949/zero-shot-one-shot-few-shot-learning

[2] https://rahulrajpvr7d.medium.com/zero-shot-one-shot-and-few-shot-learning-with-examples-8a3efdcbb158

[3] https://huggingface.co/spaces/evaluate-metric/bleu

[4] https://huggingface.co/spaces/evaluate-metric/rouge

[5] https://huggingface.co/spaces/evaluate-metric/meteor

[6] https://huggingface.co/spaces/evaluate-metric/bertscore

[7] https://huggingface.co/spaces/evaluate-metric/frugalscore

[8] https://arxiv.org/abs/1912.10165