Team NLLG submission for Eval4NLP 2023 Shared Task: Retrieval-Augmented In-Context Learning for NLG Evaluation

Daniil Larionov
Bielefeld University
daniil.larionov@uni-bielefeld.de

Vasiliy Viskov
Skoltech
Vasiliy.Viskov@skoltech.ru

George Kokush
HSE University
g.kokush5@gmail.com

Alexander Panchenko
Skoltech
A.Panchenko@skoltech.ru

Steffen Eger
Mannheim University
steffen.eger@uni-mannheim.de

Abstract

In this paper, we introduce a novel approach for evaluating natural language generation (NLG) using retrieval-augmented in-context learning. Our method empowers practitioners to leverage large language models (LLMs) for diverse NLG evaluation tasks without the need for fine-tuning. We put our approach to the test in the context of the Eval4NLP 2023 Shared Task, specifically in translation evaluation and summarization evaluation subtasks. The results indicate that retrieval-augmented in-context learning holds great promise for the development of LLM-based NLG evaluation metrics. Future research directions involve investigating the performance of various publicly available LLM models and identifying the specific LLM attributes that contribute to enhancing metric quality.

1 Introduction

Like any machine learning task, the NLG problem requires a quality metric to compare model outputs to a gold standard. The most popular method for human evaluation is MQM (Lommel et al., 2014), which allows building an interpretation of the generation model through error detection. However, this technique requires expensive manual work of an expert. As a consequence, automatic evaluation systems that would have a high correlation with state-of-the-art evaluation techniques, in particular MQM, would be highly desirable as a replacement for human MQM annotations. One such approach, became entrenched after the appearance of LLMs, is zero-shot or few-shot generation by text query, prompt. The score is obtained from the model by (i) the numerical estimate itself (Kocmi and Federmann, 2023), (ii) aggregation over the probabilistic distribution of the model (Liu et al., 2023) or (iii) a real function over the resulting text, repeating the existing methodology of expert evaluation (Fernandes et al., 2023).

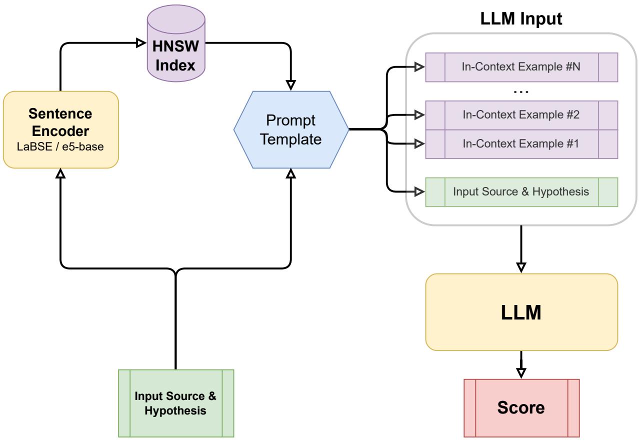

Figure 1: Architecture of the proposed approach

The shared task of Eval4NLP 2023 (Leiter et al., 2023b) challenges to solve the problem of evaluating machine translation and summarization results using a fixed set of LLMs without any fine-tuning techniques and in a reference-free manner. Reference-free means that the metric rates the provided machine translation solely based on the provided source sentence/paragraph, without any additional, human written references.

The shared task has the following goals:

1. What is the best strategy for constructing LLM-based evaluation metrics using prompting?

2. How could we explain obtained scores?

The main judgement metric during the competition is segment-level Kendall-τ correlation between model scores and MQM expert annotations. For the second goal listed above, the organizers will evaluate explanations manually.

The following list of models from Huggingface (Wolf et al., 2019) was available during the competition:

- Guanaco-65B-GPTQ: a four-bit quantized version of Guanaco-65B (Dettmers et al., 2023)

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 228–234
November 1, 2023 ©2023 Association for Computational Linguistics

- Platypus2-70B-Instruct-GPTQ: based on LLaMA2, a quantized version (Lee et al., 2023)
- WizardLM-13B-V1.1-GPTQ: a four-bit quantized version of WizardLM-13B-V1.1 (Xu et al., 2023)
- Nous-Hermes-13b: a model by Nous Research
- OpenOrca-Platypus2-13B: based on LLaMA2 (Mukherjee et al., 2023)
- orca_mini_v3_7b: smaller than the others on this list and also performs well on LLM leaderboards

We consider only large model tracks in our work due to the empirical discovery that it is easier to produce adequate texts from large models. The project code is open-sourced and available by the link[1].

2 Related Work

In general, designing high-quality evaluation metrics for NLG tasks such as summarization and machine translation is an highly active field of research. It is especially active since the recognition that decades old metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are inadequate for evaluation (Mathur et al., 2020; Peyrard, 2019; Freitag et al., 2022). The focus in recent years is on developing high-quality LLM based metrics (Zhang et al., 2020; Zhao et al., 2019) that are (among others) *explainable* (Kaster et al., 2021; Leiter et al., 2022a, 2023a, 2022b; Sai et al., 2021), *efficient* (Kamal Eddine et al., 2022; Grünwald et al., 2022; Zouhar et al., 2023; Belouadi and Eger, 2023), *robust* (Chen and Eger, 2023; Rony et al., 2022), and *reproducible* (Chen et al., 2022; Grusky, 2023). The focus of Eval4NLP's Shared Task is on explainable high-quality metrics induced from prompting the most recent classes of LLMs including variants of LLaMA (Touvron et al., 2023).

The ability of GPT-4 (OpenAI, 2023) to solve different NLG problems in a zero-shot manner led to appearance of new NLG evaluation approaches utilized this model. GEMBA (Kocmi and Federmann, 2023) used a set of instruction prompts for machine translation evaluation which differ from each other with score ranges and its descriptions, model is expected to generate repeatedly the text until it is a score as sequence of digits. Another usage of GPT-4 in NLG evaluation is G-Eval (Liu et al., 2023), which used a similar approach for summarization evaluation with zero-shot instruction based generation but with another score obtaining. The final score is an aggregation of digits with their token generation probabilities.

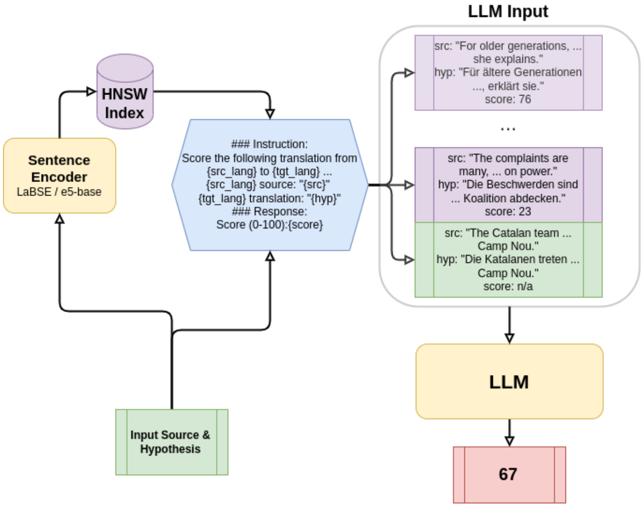
LLM Input
HNSW Index
Sentence Encoder
LaBSE / e5-base
### Instruction:
Score the following translation from {src_lang} to {tgt_lang} ...
{src_lang} source: "{src}"
{tgt_lang} translation: "{hyp}"
### Response:
Score (0-100):{score}
src: "For older generations, ... she explains."
hyp: "Für ältere Generationen ..., erklärt sie."
score: 76
...
src: "The complaints are many, ... on power."
hyp: "Die Beschwerden sind ... Koalition abdecken."
score: 23
src: "The Catalan team ... Camp Nou."
hyp: "Die Katalanen treten ... Camp Nou."
score: n/a
LLM
67
Input Source & Hypothesis

Figure 2: Workflow of the proposed method for English-German machine translation evaluation

AutoMQM (Fernandes et al., 2023) is a fine-grained approach which allows to construct interpreted evaluation via modeling MQM metric. The model is expected to generate error major and minor spans, after that the deterministic score based on MQM error weights is alculated. The vanilla approach used full transformer architecture, we try to repeat this approach with decoder-only model.

Similarly to our proposed approach, retrieval-augmented in-context learning was used for multiclass text classification in (Milios et al., 2023). In this paper, the pretrained retrieval model from SentenceTransformers (Reimers and Gurevych, 2019) is used to collect the in-context examples, closest to the input text. In their case, the length of the examples is consistently small, so they are able to fit as many as 110 in-context examples by greedily selecting examples until they completely fill the model's context window.

3 Approach

The basis of our approach is the selection of several few-shot examples for each specific instance. To do this, we use an index, a large array of source texts

[1] https://github.com/Rexhaif/retrieval-augmented-score

from the training dataset for different language pairs. In the index, all texts are stored as embeddings, which are then compared to the source text by the cosine distance. We specifically compare samples by their source texts, as we hypothesize that for similar source examples, the way to evaluate translation/summarization is usually similar.

The whole workflow with the examples is illustrated on Figure 2. The few-shot examples themselves use the same prompt format as the request — only with an already inserted score. All the examples go in a row, forming a single prompt from several few-shot parts and a prompt with the requested rating. For a more accurate assessment, we obtain various examples from the index, both with high and low scores.

3.1 Machine Translation

```
### Instruction:
Score the following translation from
{src_lang} to {tgt_lang} on a continuous
scale from 0 to 100 that starts with
"No meaning preserved", goes through
"Some meaning preserved", then "Most
meaning preserved and few grammar
mistakes", up to "Perfect meaning
and grammar".
{src_lang} source: "{src}"
{tgt_lang} translation: "{hyp}"
### Response:
Score (0-100):{score}
```

Figure 3: The prompt we used for the machine translation task

The final method for the machine translation evaluation task was to generate the score itself. The main difference with summarization task was in the selected text embedding model: for the summarization task we had to use a model which was trained to handle the retrieval of long texts.

3.2 Fine-grained error identification

We also tried the AutoMQM (Fernandes et al., 2023) approach for machine translation evaluation. Instead of evaluating the sample score itself, the model was instructed to generate a list of all translation errors in the example, indicating their criticality — based on this, the score is calculated following the MQM (Freitag et al., 2021) scoring method. To do this, we modified few-shot prompts to include fine-grained translation errors. However, this approach was unsuccessful: often the error spans were not recognized correctly. We believe this is because the model we tried was a decoder-only one, unlike the model in the original paper; (Fernandes et al., 2023) used an encoder-decoder architecture, which may be better for in-context learning (we leave a thorough investigation to future work).

3.3 Summarization

```
### Instruction:
Score the summarization with respect to
the summarized document on a continuous
scale from 0 to 100, where a score of zero
means "irrelevant, factually incorrect and
not readable" and score of one hundred
means "relevant, factually correct, good
readability".
Source text: "{src}"
Summary: "{hyp}"
### Response:
Score (0-100):{score}
```

Figure 4: The prompt we used for the summarization task

For the summarization evaluation task, we used a model for large texts because the source texts have a long length.

4 Experimental Setup

Following the competition rules, our choice of base LLMs was limited. Eventually, we have conducted experiments using 3 different models: «TheBloke/Platypus2-70B-Instruct-GPTQ», «Open-Orca/OpenOrca-Platypus2-13B», «NousResearch/Nous-Hermes-13b».

All experiments were conducted on a single Nvidia A40 GPU with 48GB of VRAM. We used model implementation in PyTorch 2.0 (Paszke et al.) together with transformers (Wolf et al., 2019) framework. We used greedy decoding limited to generation of 3 new tokens to generate scores for the analyzed text. At this time, we have not implemented any controlled generation to enforce generation of digit tokens, if model have generated something that could not be parsed into an integer, we did a fallback to default score of 0.

4.1 MT Evaluation

To construct the pool of examples for retrieval-augmentation, we use a set of datasets from previous years of WMT Metrics Shared Task. We took datasets from 2017 to 2022, with DA (Direct Assesment) scores. In total, the pool of examples containes around 1.5m examples. The nearest-neighbors index was constructed on sentence embeddings vectors of source texts of these examples. We employ LaBSE (Feng et al., 2020) to construct embeddings due to its superior performance on multilingual tasks[2]. The overall pipeline is illustrated on Figure 1.

For each analyzed example, we collect 10 in-context examples, which have semantically-closest source text. In order to avoid accidental data leakage, we have queried 10+1 examples from the index and excluded the first one with the highest similarity score. Both input example and in-context examples were formatted according to GEMBA's-SQM[noref] (Kocmi and Federmann, 2023) prompt template and concatenated to form a single prompt.

4.2 Summarization Evaluation

For the construction of the example pool for summarization, we use SummEval (Fabbri et al., 2021) dataset. This dataset contains 100 distinct source texts and 16 different summaries per text. In order to increase diversity of in-context examples, we take a single summary out of 16 for each of the source texts at random. The nearest-neighbor index is constructed on embeddings of the source texts. The embeddings are computed using e5-base-v2 model (Wang et al., 2022). We choose this model because it was specifically trained to handle retrieval of long texts. According to the model specifications, we add the prefix "passage: ".

Due to large size of in-context examples for this task, we reduce the number of in-context examples to 3 in order to fit into the base LLMs context window.

5 Results & Discussion

The results of evaluation of the proposed approaches are presented in Table 1. As illustrated in the table, the «Platypus2-Instruct-70B» model, which has the largest number of parameters, outperforms all other approaches. It suggests that retrieval-augmented in-context learning, expectedly, does benefit from LLMs with more parameters. However, for the summarization task we see no difference in obtained scores. These findings suggest that our approach has substantial limitations when applied to summarization. Indeed, while the pool of in-context examples for MT evaluation consists of 1.5m examples, in the case of summarization, we only have 100 examples to choose from. This does limit the variability of the scores and texts that are included in in-context examples. An additional limitation factor is the context window size of the LLM, which reduces the amount of in-context examples that we could include.

Model	en-de	en-zh	en-es	summ
⭐ platypus-70b	0.24	0.13	0.18	0.35
platypus-13b	0.07	0.04	0.10	0.35
nous-hermes	0.09	0.06	0.10	n/a
fine-grained 70b	0.11	n/a	n/a	n/a

Table 1: Kendall-τ correlations of the tested models/approaches on the shared task test set. The first three lines refer to models tested with score generation, while the last lines refer to a fine-grained error identification approach. 'n/a' refers to subtasks that we have not been able to evaluate on with particular models due to time restrictions as well as technical difficulties. ⭐ indicates the variant that was submitted to the shared task.

From the multilingual perspective, all our models rely on substantially limited/non-existent multilingual pretraining of the base model as well as the fine-tuned versions. In fact, all those models use the small vocabulary of 32k tokens. This does seem to be enough to capture word pieces for English and similar Latin scripted languages: Spanish and German. However, in the case of the English-Chinese language pair, we see a consistent drop in metric correlation among all tested LLMs.

Lastly, the fine-grained approach described above yielded only 0.11 on Kendall-τ correlation with human judgment for the English-German translation subtask. While we were not able to finish its inference on other MT subtasks in time, we did find several problems with this approach. In most cases, the model failed to accurately produce spans for identified errors as they contained some words from the translated text but in a disarranged order, along with unrelated words. Also, we found that in some cases, the model generated a list of duplicate or near duplicate errors, which resulted in an

[2] See 'Bitext mining' section at the leaderboard: https://huggingface.co/spaces/mteb/leaderboard

overly pessimistic approximation of the translation quality. We hypothesize that it was likely due to the model we have used. In the original paper (Fernandes et al., 2023), the authors use Google's private PaLM-2 (Anil et al., 2023) model which is **a)** has more (540B) parameters, **b)** was pre-trained on 'parallel data covering hundreds of languages' and **c)** is based on encoder-decoder architecture. In contrast, in our case, the largest model had only 70B parameters and was mostly pretrained on monolingual English data. Also, according to (Ding et al., 2023), the decoder-only CausalLMs are suboptimal for the case of in-context learning, while PrefixLMs (encoder-decoder) are better suited to utilize in-context examples for generating prediction.

6 Conclusion

During experiments for Eval4NLP 2023 Shared Task, we considered approaches with in-context learning and fine-grained evaluation and observed that adding reference examples could boost the generation result, even though it is the only score. However, this method is sensitive to the encoder model with index setup, examples' set size and requires a lot of diverse references. We did not manage to observe good results for fine-grained approach with AutoMQM, we think that the problem is with the model size and architecture.

Some ideas for further research include: **a)** exploring the capabilities of LLMs with more parameters when applied with our prompting strategy, **b)** utilizing models with larger (or unlimited) context window to increase the number of in-context examples, **c)** experimenting with LLMs pre-trained on multilingual data for translation evaluation and **d)** applying encoder-decoder LLMs to achieve better incorporation of in-context examples.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Tachard Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D'iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *ArXiv*, abs/2305.10403.

Jonas Belouadi and Steffen Eger. 2023. UScore: An effective approach to fully unsupervised evaluation metrics for machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 358–374, Dubrovnik, Croatia. Association for Computational Linguistics.

Yanran Chen, Jonas Belouadi, and Steffen Eger. 2022. Reproducibility issues for BERT-based evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2965–2989, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust Evaluation Metrics from Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Nan Ding, Tomer Levinboim, Jialin Wu, Sebastian Goodman, and Radu Soricut. 2023. Causallm is not optimal for in-context learning. *ArXiv*, abs/2308.06912.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Languageagnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jens Grünwald, Christoph Leiter, and Steffen Eger. 2022. Can we do that simpler? simple, efficient, high-quality evaluation metrics for nlg. *ArXiv*, abs/2209.09593.

Max Grusky. 2023. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland. Association for Computational Linguistics.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*.

Christoph Leiter, Piyawat Lertvittayakumjorn, M. Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022a. Towards explainable evaluation metrics for natural language generation. *ArXiv*, abs/2203.11131.

Christoph Leiter, Piyawat Lertvittayakumjorn, M. Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2023a. Towards explainable evaluation metrics for machine translation. *ArXiv*, abs/2306.13041.

Christoph Leiter, Hoang-Quan Nguyen, and Steffen Eger. 2022b. Bmx: Boosting machine translation metrics with explainability. *ArXiv*, abs/2212.10469.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023b. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison for NLP systems*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. *ArXiv*, abs/2309.10954.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Md Rashad Al Hasan Rony, Liubov Kovriguina, Debanjan Chaudhuri, Ricardo Usbeck, and Jens Lehmann. 2022. RoMe: A robust metric for evaluating natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5645–5657, Dublin, Ireland. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, and Mrinmaya Sachan. 2023. Poor man's quality estimation: Predicting reference-based MT metrics without the reference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1311–1325, Dubrovnik, Croatia. Association for Computational Linguistics.