Using Large Language Models to Evaluate Biomedical Query-Focused Summarisation

Hashem Hijazi[1] and **Diego Mollá**[2,1] and **Vincent Nguyen**[1] and **Sarvnaz Karimi**[1]
[1]CSIRO Data61 and [2]Macquarie University
Sydney, Australia
firstname.lastname@csiro.au
Correspondence: diego.molla-aliod@mq.edu.au

Abstract

Biomedical question-answering systems remain popular for biomedical experts interacting with the literature to answer their medical questions. However, these systems are difficult to evaluate in the absence of costly human experts. Therefore, automatic evaluation metrics are often used in this space. Traditional automatic metrics such as ROUGE or BLEU, which rely on token overlap, have shown a low correlation with humans. We present a study that uses large language models (LLMs) to automatically evaluate systems from an international challenge on biomedical semantic indexing and question answering, called BioASQ. We measure the agreement of LLM-produced scores against human judgements. We show that LLMs correlate similarly to lexical methods when using basic prompting techniques. However, by aggregating evaluators with LLMs or by fine-tuning, we find that our methods outperform the baselines by a large margin, achieving a Spearman correlation of 0.501 and 0.511, respectively.

1 Introduction

Biomedical question answering (QA) is concerned with building systems that automatically answer biomedical questions posed by humans in natural language (Soares and Parreiras, 2018; Nguyen, 2019). To develop and optimise these systems, we must use metrics that evaluate the quality of their output. Automatic metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) have been shown to correlate poorly with human evaluation (Liu et al., 2016), and human annotations are prohibitively expensive and impractical in the biomedical domain (Pampari et al., 2018; Guo et al., 2006). To rectify this problem, recent research has suggested using medium-sized model-based evaluators and Large Language Models (LLMs). Model-based evaluators such as BERTscore (Zhang et al., 2020) or BLEURT (Sellam et al., 2020) have demonstrated improvements over n-gram based metrics in various natural language generation (NLG) evaluation contexts such as summarisation and QA (Zhong et al., 2022). However, their evaluation capability is still far below that of humans.

The recent improvement of LLMs for various tasks has fostered research on their use for the evaluation of the performance of text generation tasks such as summarisation and dialogue generation (Liu et al., 2023). In this paper, we experiment with using LLMs to evaluate biomedical query-focused summarisation systems. To the best of our knowledge, this is the first study for such a task. In particular, we compare the correlation between human judgements and LLM-based evaluators for the evaluation of several systems participating in the "ideal answer" task of BioASQ 2021 and 2022 (Nentidis et al., 2022). Our study examines different prompting strategies for such evaluations.

2 Related Work

Fu et al. (2024) indicates that the use of LLMs as reference-free probability-based evaluators yields performance superior to n-gram metrics and model-based evaluators such as BERTscore, all the while providing a customised and multi-faceted evaluation with no training cost. Probability-based LLM evaluation, however, suffers from issues of robustness which lead to biases and loopholes (He et al., 2023) that impact its efficacy.

LLMs, through prompting, have also been used to evaluate text via Likert scale scoring (a five-level scale) (Likert, 1932), leading to greater performance than n-gram and model-based evaluation techniques. However, Chiang and Lee (2023) showed that outputting only a number could be suboptimal but that asking the LLM to explain its rating can lead to an increase in correlation to human ratings. The pairwise ranking also showed promising results (Kotonya et al., 2023), with accuracy outperforming n-gram and model-based evaluators.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 236–242

LLMs can also leverage emergent abilities which can be used to incorporate in-context learning (ICL) and chain-of-thought (CoT) in their evaluation strategies. In ICL (Xie et al., 2022), a model is provided with input-output examples of a downstream task instead of being trained or fine-tuned on the task. In CoT (Kojima et al., 2022), a complex task is broken down into multiple intermediate steps to improve reasoning in LLMs.

Liu et al. (2023) used GPT-4 to achieve the highest correlation with human evaluations in comparison to other model and n-gram based metrics. Jain et al. (2023), using GPT-3, showed that using few-shot prompting — a small number of examples added to the prompt — can reach or exceed the state-of-the-art on multi-dimensional evaluation and that this is robust to the sampling method of in-context examples whether it be random or representative of the range of scores in the example pool. However, Kotonya et al. (2023) showed using a small LLM (`orca-mini-v3-7B`), that one-shot prompting doesn't bring significantly greater results than zero-shot.

The use of LLMs as meta-evaluators who use their reasoning capabilities to combine diverse evaluation techniques has also seen promise. In Shu et al. (2023), various LLMs were given supplementary evaluation metrics like NLI Score (Bowman et al., 2015), BLEURT and probability-based LLM techniques to aid with their judgements, with the meta-evaluation outperforming all individual evaluators.

The above research, however, was not used in query-focused summarisation tasks. This paper is the first that tests the use of LLMs for the evaluation of biomedical query-focused summarisation.

3 Methodology

We use the human judgements of runs participating in the "ideal answers" question answering task of BioASQ (Nentidis et al., 2022). Such "ideal answers" are multi-sentence answers, and therefore the task is about biomedical query-focused summarisation. In particular, training and development is based on systems (runs) participating at BioASQ 2020[1], whereas testing is on runs participating at BioASQ 2021[2] and BioASQ 2022[3]. The rationale for using BioASQ 2020 for training is that it is the most recent year prior to the test data. Resource constraints do not allow us to use all runs participating at BioASQ 2020 for training, or all runs participating at BioASQ 2021 and 2022 for testing.

Evaluation criteria

Recall: *Fraction of information in the known answers that is reported in the generated response*
Precision: *Fraction of information in the generated response that is in the known answers*
Repetition: *Amount that the generated response repeats the same information*
Readability: *Generated response's ability to be easily understood and easily identifiable as an answer to the question by a human*

Figure 1: Human criteria for the evaluation of a (question, ideal answer) pair given the known answer. Each criterium was scored between 1 and 5.

The human judges are given instructions to evaluate the pair (question, generated answer), given a correct answer, according to four criteria presented in Figure 1. The final score of a (question, generated answer) pair is the average of the 4 criteria. These judgements are provided to us by the organisers of BioASQ. To preserve privacy, we only had access to the judgement of runs submitted by us.

For each automatic evaluation technique, the Spearman correlation (Spearman, 1904) is calculated. In addition, since the LLM-based evaluation generated integer numbers 1 to 5, for each LLM-based evaluation technique, the quadratic kappa, a well-established correlation metric for nominal scales (Cohen, 1968), was also calculated[4]. BioASQ runs five to six evaluation batches each year. Since there is no guarantee that the same runs participated in all batches, the correlations are computed separately for every batch, and the results reported in this paper are the average correlation. Each batch has approximately 100 (question, generated answer, known answer) triples.

Given that the automatic evaluation by LLMs can vary each time the LLM is run, each batch is evaluated three times and then the evaluation re-

[1]Run MQ1 in each batch submitted by Mollá et al. (2020).
[2]Run MQ1 in each batch submitted by Khanna and Mollá (2021).
[3]Run MQ1 in each batch submitted by Mollá (2022).
[4]Quadratic kappa could not be used on the output of the other evaluation techniques we tested because they generate real numbers between 0 and 1.

sults are averaged before computing the correlation with the human judges.

The LLM-based evaluations return independent evaluation scores for each evaluation criterion (Figure 1). Our correlation experiments consequently measure the correlation of each criterion (and average) per the corresponding human criterion (and average). For example, the *Precision* column of Table 1 shows the correlation of the *Precision* scores generated by each LLM evaluator for the *Precision* scores of the human judges.

4 Experiments

We investigated several evaluation strategies, from baseline well-known token-based metrics to the use of LLMs as evaluators in different settings.

4.1 Token-based Techniques and their Limitations

ROUGE-1 and ROUGE-2 (F1), as well as SciBERTscore (Precision, Recall, F1), were tested to attain a baseline correlation level. ROUGE F1 was chosen given its robustness over precision and recall (Mollá and Jones, 2020). Sci-BERT (Beltagy et al., 2019) is a BERT (Devlin et al., 2019) model pretrained on papers from Semantic Scholar, of which 82% are from the biomedical domain. SciBERT was chosen over BERT due to its greater understanding of biomedical terminology.

4.2 LLMs

GPT-3.5 (`gpt-3.5-turbo-0125`) and GPT-4 (`gpt-4-1106-preview`) were used as evaluators. All prompts included the question, reference answer(s), a system output, the defined evaluation criteria, and instructions rating responses on a 1-5 integer scale. For all runs, the system prompt was set to "You are a useful evaluator of a biomedical question answering system", and the temperature and top p were set to 0 and 0.6, respectively, to facilitate reproducibility. Out-of-the-box (OTB) GPTs were tested with the base prompt listed in Figure 2.

4.2.1 Reason then Score

Chain-of-thought (CoT) prompting has demonstrated an increase in performance in various tasks, such as arithmetic and commonsense reasoning (Kojima et al., 2022). We used the variant of CoT called "Reason then Score" (RTS), in which an LLM is asked to explain its reasoning. RTS has emerged as a popular prompting technique (Shen et al., 2023). In our experiments, we altered the answer reporting format to include an explanation area that instructs the LLM to explain its answers (Figure 3).

OTB prompt

We have a biomedical question, a list of known answers, and the output generated by an automatic question-answering system. Given the known answers, evaluate the quality of the answer generated by the system. In your evaluation, address the following:
1. recall: The fraction of information in the known answers that is reported in the generated response. A higher score indicates better recall.
2. precision: The fraction of information in the generated response that is in the known answers. A higher score indicates better precision.
3. repetition: The amount that the generated response repeats the same information. A higher score indicates less repetition.
4. readability: The generated response's ability to be easily understood and easily identifiable as an answer to the question by a human. A higher score indicates better readability.
Use a 1-5 integer scale. Report the answer as a json structure using the template.
{"readability": {"score": 1-5}, "recall": {"score": 1-5}, "precision": {"score": 1-5}, "repetition": {"score": 1-5} }.
remember to report the answer as the format above with no deviations from this format. remember "score" is a key.

Figure 2: Prompt used in the out-of-the-box (OTB) LLM systems.

4.2.2 Few Shot

LLMs are reported to display an increase in performance when the prompt includes a few examples with their input query (Brown et al., 2020). We provided the LLMs with six examples from BioASQ 2020. Our initial experiments showed that random sampling of examples yielded poor performance. We instead used a percentile-based selection strategy to ensure a wide coverage of the example pool, based on the scores given by the human judges. In

RTS prompt

... (text of figure 2 inserted here, replacing its json template with the following) ...
{"recall": {"explanation":"", "score":1-5}
"precision": {"explanation":"", "score":1-5}
"repetition": {"explanation":"", "score":1-5}
"readability": {"explanation":"", "score":1-5}

Figure 3: Prompt used for *Reason then Score* (RTS) prompting.

particular, examples with 15th and 80th-percentile scores for recall, precision and readability were included in the prompt.

4.2.3 Fine-tuning

We also experimented with fine-tuned LLMs. In particular, we fine-tuned GPT-3.5 (`gpt-3.5-turbo`) using the same prompting format as OTB.

4.2.4 LLMs as Meta Evaluators

Inspired by work by Shu et al. (2023), we experimented with the use of LLMs as meta-evaluators of 3 different evaluators. In particular:

1. To aid the LLM in scoring repetition, we provided it with a "Repscore" which took the number of unique words in a response and divided it by the total number of words.

2. Smog score (Laughlin, 1969), which looks at the number of polysyllabic words and sentences in a text, was used to aid the scoring of readability.

3. Finally, the output from the fine-tuned GPT-3.5 model was also included to aid with the scoring of all dimensions.

Testing was done primarily on GPT-4, since GPT-3.5 showed a very limited capability in reasoning with other scores.

4.2.5 Pairwise Ranking

LLMs have also shown potential in comparative assessment (Liusie et al., 2024). We conducted pairwise ranking, where the LLM evaluator (gpt4-1106-preview) was given the output of two systems,

Pairwise Ranking prompt

We have a biomedical question, a list of known answers, and outputs generated by different automatic question answering systems.
Given the known answers, rank the quality of the outputs generated by the system based on how similair they are to the ideal answers and if they answer the question properly.
Report the answer as a JSON structure using the template, noting that a rank of 1 implies that this system output is the best compared to the others:
{"ranks": 1: "system name", 2: "system name", "Explanation": "reasoning for why you ranked the systems this way. be specific"}.

Figure 4: Prompt used for *Pairwise ranking* prompting.

the question, and was asked to rank them. We gave the LLM a CoT of the form Score then Reason as shown in Figure 4. We used accuracy and Cohen's kappa[5] to evaluate the performance.

5 Results and Discussion

Table 1 shows the results of all experiments except pairwise ranking, and Table 2 shows the results of the experiments with pairwise ranking.

Similar to previous work, we find that token-based methods perform worse than LLMs in general (Liu et al., 2023). Possible causes of the relatively poor performance of token-based approaches is, as mentioned by Hanna and Bojar (2021), that ROUGE is unable to incorporate information on context and semantic meaning, whereas Sci-BERTscore is less sensitive to errors in text, especially if the candidate is lexically or stylistically similar, and both are insensitive to negation. These limitations have increasingly larger impacts on abstractive over extractive systems, making evaluating outputs using these metrics potentially unreliable. Still, Table 1 shows that the token-based methods achieve a correlation with the human judgements of precision and readability that is comparable to that of some of the LLM approaches.

[5]This is the standard Cohen's kappa, not quadratic kappa, since now the score is not a nominal scale.

Table 1: Spearman Correlation and Quadratic Kappa values. The Average column shows the average of Precision, Recall, Readability, and Repetition scores. The Combined column shows the correlation between the score resulting from averaging the machine predictions for Precision, Recall and averaging the human annotations. FS: Few Shot and CoT: Chain-of-Thought.

		Precision		Recall		Readability		Repetition		Average		Combined
		ρ	κ	ρ	κ	ρ	κ	ρ	κ	ρ	κ	ρ
Token-based	ROUGE-1-F1	0.384	-	0.280	-	0.159	-	0.118	-	0.235	-	0.357
	ROUGE-2-F1	0.412	-	0.271	-	0.164	-	0.129	-	0.244	-	0.364
	Sci-BERTscore-P	0.489	-	0.184	-	0.150	-	0.233	-	0.264	-	0.391
	Sci-BERTscore-R	0.283	-	0.325	-	0.146	-	0.152	-	0.227	-	0.341
	Sci-BERTscore-F1	0.420	-	0.264	-	0.154	-	0.208	-	0.262	-	0.391
LLM	GPT-3.5	0.370	0.235	0.298	0.229	0.123	0.090	0.232	0.143	0.256	0.174	0.388
	GPT-3.5 - CoT	0.322	0.266	0.293	0.256	0.168	0.151	0.242	0.155	0.256	0.207	0.376
	GPT-3.5 - FS	0.363	0.252	0.333	0.253	0.130	0.122	0.039	0.036	0.216	0.166	0.370
	Fine-tuned GPT-3.5	**0.537**	0.472	0.352	0.331	0.295	0.273	**0.516**	**0.460**	**0.425**	**0.384**	**0.511**
	GPT-4 as meta evaluator	0.531	**0.472**	**0.426**	**0.428**	**0.343**	**0.317**	0.388	0.275	0.422	0.373	0.501

Table 2: Accuracy and Kappa values of pairwise ranking evaluation.

	Accuracy	Kappa
Pairwise Ranking	0.61	0.31

GPT-3.5 with basic prompting displays similar performance to token-based metrics when using prompt engineering techniques such as CoT and few shot. When fine-tuned, GPT-3.5 attains a much higher correlation with humans than the token-based metrics. GPT-4 as a meta-evaluator achieves very similar results to the fine-tuned model. More testing is needed to be done on GPT-4 to see if prompt engineering leads to even better results.

Few-shot prompting performed the worst out of the LLM-based methods. In our preliminary experiments, we observed that the performance of the few-shot approach varied, with some batches increasing their correlations and others decreasing. This was in contrast with the performance of the fine-tuned approach, which had a lower variation across batches. This suggests that the examples chosen for few-shot could be more suited to certain batches. A promising future direction is to incorporate a dynamic selection of examples.

6 Conclusions

We compared the use of traditional evaluation metrics (ROUGE, BERTScore) with the use of LLMs for the evaluation of query-focused summarisation of biomedical questions. For this, we used system runs that participated in BioASQ challenge, and computed correlation between automatic evaluations and human judgements.

Our experiments show that, while LLMs with basic prompting do not outperform ROUGE or BERTScore, approaches that use fine-tuning or that combine LLMs with additional scorers, significantly improve correlation with human judgements.

7 Limitations

Due to limitations of resources and the availability of few runs, we have not experimented with a wide range of outputs of differing characteristics. Our training and test data used runs that were performed in the middle to the top range of systems participating in BioASQ. Therefore, the quality of the evaluators has not been tested on poor-quality runs. As a consequence, even though the results presented here should be valid to evaluate medium to high-quality systems, we cannot guarantee that the quality of the evaluations applies to poor-performing systems.

8 Ethical Considerations

The human judgements were obtained from the organisers of BioASQ. To ensure the privacy of these judgements, we only had access to judgements of past runs submitted by the authors of this paper, and the review judgements were anonymous.

Even though our results show a better correlation with human judgements than other automatic evaluation metrics, there is still room for improvement, and the evaluation results might not be reliable enough for applications requiring high-quality output systems and high-quality evaluation.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained language model for scientific text. In *EMNLP*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*, 70(4):213–20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Yikun Guo, Robert Gaizauskas, Ian Roberts, and George Demetriou. 2006. Identifying personal health information using support vector machines. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Urvashi Khanna and Diego Mollá. 2021. Transformer-based language models for factoid question answering at BioASQ9b. In *Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*, pages 247–257.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Neema Kotonya, Saran Krishnasamy, Joel Tetreault, and Alejandro Jaimes. 2023. Little giants: Exploring the potential of small LLMs as evaluation metrics in summarization in the Eval4NLP 2023 shared task. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 202–218, Bali, Indonesia.

G. Harry Mc Laughlin. 1969. Smog grading — a new readability formula. *Journal of Reading*, 12(8):639–646.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the Association for Computational Linguistics Workshop*, pages 74–81, Barcelona, Spain.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Diego Mollá. 2022. Query-focused extractive summarisation for biomedical and COVID-19 complex question answering. In *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum*, pages 305–314.

Diego Mollá and Christopher Jones. 2020. Classification betters regression in query-based multi-document summarisation techniques for question answering. In *Machine Learning and Knowledge Discovery in Databases*, pages 624–635, Cham. Springer International Publishing.

Diego Mollá, Christopher Jones, and Vincent Nguyen. 2020. Query focused multi-document summarisation of biomedical texts. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2022. Overview of BioASQ 2022: The tenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 337–361, Cham. Springer International Publishing.

Vincent Nguyen. 2019. Question answering in the biomedical domain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 54–63, Florence, Italy.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, US.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore.

Lei Shu, Nevan Wichers, Liangchen Luo, Yun Zhu, Yinxiao Liu, Jindong Chen, and Lei Meng. 2023. Fusion-Eval: Integrating evaluators with LLMs. *Preprint*, arXiv:2311.09204.

Marco Soares and Fernando Parreiras. 2018. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.