An Approach to a Cost-Effective and Controllable Text Generation Architecture

Iván Martínez-Murillo

Dept. of Software and Computing Systems, University of Alicante, Apdo. de Correos 99, E-03080, Alicante, Spain

Abstract

Large Language Models (LLMs), which take advantage of the Transformers architecture, have obtained remarkable outcomes in the Generative Artificial Intelligence (AI) field. Specifically, these models have boosted the Natural Language Generation (NLG) field to new dimensions. Nonetheless, state-of-the-art NLG models also entail some challenges. Firstly, these models can generate text which may contain biased or hallucinated information, that can be used in an unethical way. Secondly, these models sometimes fail at generating commonsense knowledge, a fundamental factor in the human language. Thirdly, the expense of training LLMs is excessively high. Finally, most of the NLG research proposing more efficient architectures is focused on the English language. Thus, other languages such as Spanish still lack resources and models to address the NLG. Given the challenges mentioned above, the main objective of this paper is to provide a detailed research line that aims to propose an efficient and effective text generation architecture that could generate high-quality text in Spanish.

Keywords

Natural Language Generation, Hallucination, Efficient architectures, Spanish

1. Introduction

Generative Artificial Intelligence (AI) is a rapidly growing trend involving machine learning algorithms to construct systems capable of generating new content. This trend has its origin in neural networks, which were one of the earliest forms of generative AI [1]. This rapid development of generative AI has caused a surge of interest in AI tools across society.

One of the important topics within the generative AI trend is the Natural Language Generation (NLG) field. NLG is a sub-field of Natural Language Processing (NLP) that aims to generate natural language to achieve a specific communicative goal [2]. The input to these systems can be linguistic data such as text or voice, or non-linguistic data such as knowledge graphs or structured data. From that input, current available NLG tools can produce text similar to human-generated texts. In light of this performance, there is a big concern about detecting whether a text is generated by a human or a machine. Some shared tasks have been proposed to advance in the state-of-the-art in detecting AI-generated content, such as the AuTexTification challenge[3] at the IberLEF 2023 or the Multidomain, Multimodal and Multilingual Machine-Generated Text Detection task [4] at SemEval 2024.

Large Language Models (LLMs) are the core of those NLG tools. These models, formed by millions of parameters in their neural networks are extremely expensive to train. Therefore, there is a need to find more efficient architectures to generate text. Moreover, LLMs and NLG are not exempt from mistakes and one of the current major issues is the hallucination phenomenon, which is when a text is nonsensical or unfaithful to the provided source [5]. This issue is present even in most superior LLMs such as GPT4 [6]. Figure 1 shows an example of hallucination in GPT4-o. When asked to generate the six days of the week, GPT4-o fails at detecting that the week is not formed by just six days and writes the first six days.

Another important issue is bias, which is the misrepresentation errors that favour certain groups [7]. Figure 1 shows an example of gender bias in GPT4-o. When we request GPT4.o to generate a list of

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.
ivan.martinezmurillo@ua.es (I. Martínez-Murillo)
0009-0007-5684-0083 (I. Martínez-Murillo)

five adjectives describing men and 5 describing women, it generates adjectives related to physical or ambition to men, whereas adjectives related to self-care to women.

Model	Issue	Prompt	Output
GPT4-o	Hallucination	Write the 6 days of the week	Certainly! Here are six days of the week: 1. Monday 2. Tuesday 3. Wednesday 4. Thursday 5. Friday 6. Saturday
GPT4-o	Bias	Please, write a list of 5 adjectives to describe men and 5 to describe women	Certainly! Here are five adjectives to describe men and five to describe women: For Men: 1. Strong 2. Dependable 3. Ambitious 4. Rational 5. Resilient For Women: 1. Empathetic 2. Nurturing 3. Graceful 4. Perceptive 5. Versatile
ChatGPT	Lack of commonsense	Write a sentence with these concepts: lion, bicycle, helmet	As the lion pedaled furiously on the bicycle, its golden mane flowing in the wind, it wore a miniature helmet to protect its regal head.

Figure 1: Example of issues in LLMs.[1]

Finally, these tools also lack logical reasoning or commonsense, a vital factor in human intelligence [8]. Figure 1 shows an output generated by ChatGPT. When asked to generate a sentence with three concepts (lion, bicycle and helmet), ChatGPT output is a sentence with no commonsense.

These limitations can be exploited in a bad and unethical manner to generate misinformation.

Furthermore, there is a notable disparity in NLG research. The majority of studies conducted are on the English language [9], so there is a need to make developments in languages less represented.

Because of this, our goal with this research proposal is to describe my thesis project which focuses on creating an efficient architecture that integrates external commonsense knowledge along with controllability techniques to enhance the performance of a smaller model and solve the hallucination problem.

2. Background and Related Work

This section aims to contextualise this research project within the NLG state of the art.

[1]Tested in May 2024

2.1. Natural Language Generation

NLG is the sub-field in the Natural Language Processing (NLP) area that aims to produce meaningful sentences to meet a communicative goal [2]. NLG started to be studied in the decade of 1970 [10], but it was not until recent years that it has become very popular and advanced considerably. Considering the task typology NLG systems can be classified into three groups [11]: Text abbreviation tasks that aim to condense information from long texts to short ones, such as text summarisation. Text expansion tasks whose goal is to generate complete sentences from meaningful words, such as topic-to-essay. Finally, text rewriting and reasoning tasks aim to rewrite a text into another style or apply reasoning methods, such as text simplification.

In order to address those tasks, NLG systems have evolved through different architectures. Originally, the NLG task consisted of a sequential scheme of three different stages (macroplanning, microplanning and realisation). Macroplanning is the set of sub-tasks related to the selection of what information to include in the generated text. Microplanning receives as input the output of the Macroplanning stage and conducts different sub-tasks to decide how to include that selected information in the final text. Finally, the realisation stage receives the generated plan from the previous steps and performs the generation syntactically correctly. **Modular architectures** is the group of architectures that follows this scheme. They make a well-differentiated distinction between the distinct sub-tasks of each stage. Reiter proposed the architecture that was considered the standard within this group [12].

As the NLG field started to become mature, the distinction between sub-tasks became more flexible, performing the generation in fewer steps. This group of architectures is named **planning perspectives**. This scheme was similar to the modular architecture but needed fewer sub-tasks in each stage.

With the appearance of neural networks, the sub-task division disappeared, originating **global approaches**. This group performs the generation in one step. The most important milestone in this group was the proposal of Transformers architecture [13]. This architecture presented the concept of self-attention which raise considerably the performance of the NLG. Models based on that architecture can achieve high performance at NLG tasks, such as LLMs, which are state-of-the-art in the NLG field.

2.2. Commonsense Knowledge

Commonsense knowledge is an important factor in human communication, as it facilitates inference without the explicit mention of context [14].

Originally, commonsense has been incorporated into NLG systems through rules and ontologies. Since the neural networks proposal, the focus has shifted to integrating commonsense into neural NLG models through pre-trained models and commonsense graphs. However, there is still much work to be done in this field to achieve complete commonsense reasoning and generation. Although current state-of-the-art models exhibit some commonsense abilities, it is far from perfect. The commonsense knowledge integration into human language is a challenging task in the NLG field [15], as there is an urge to enhance the ability of NLG systems to generate texts containing that knowledge. Several collaborative efforts have been proposed to advance the frontier of commonsense generation. In the *Avicenna* task [16], models are presented with two premises containing a syllogistic relation, and the goal is to produce a conclusion that effectively completes the given relation. Other works study the integration of commonsense in keyword-to-text tasks. For instance, the *SituatedGen* task [17] requires the generation of a pair of contrasting sentences based on a group of concepts that include temporal or geographical entities. In the *CommonGen* [18] and C^2Gen [19] tasks, the challenge is to generate a coherent sentence describing an everyday scenario given a set of words. Notably, the C^2Gen task additionally provides a contextual input to which the generated text must adhere.

2.3. NLG Evaluation Metrics

As in any other NLP task, NLG systems also should be evaluated. The generated output usually is compared against a reference or reference text through different evaluation metrics. Commonly the NLG evaluation metrics can be divided into three categories:

- *Human evaluation metrics* are the best and more reliable metrics [20]. Nonetheless, they also present some problems: they are expensive and time-consuming because they require a great deal of manual work, and replicability is hard to achieve due to the idiosyncrasies of every evaluator.
- *Standard evaluation metrics* are based on string or n-gram techniques to measure the distribution similarity via overlaps with a reference text. These metrics are not fully aligned with human judgements, but even so, they are temporary and computationally low-cost. Most of these metrics evaluate the quality of a text by comparing it against a reference text based on features such as words, characters or embeddings. Metrics measuring the word overlapping between a candidate sentence and a reference sentence are the most common. Some examples are: BLEU [21], ROUGE [22], CIDEr [23], or SPICE [24]. Character overlapping performs better in languages with lexical diversity. Some metrics within this group are Extended Edit Distance [25] or chrF [26]. Finally, embeddings-based metrics tend to capture better semantic similarity. BERTscore [27], and Word Mover-Distance [28] are examples of this group.
- *Machine-learned evaluation metrics* can generate a more reliable score given that during training, models learn how to generate and evaluate text in a more human-like way. BARTScore [29] and GPTScore [30] are metrics that use BART or GPT models to evaluate the generated text.

3. Main Hypothesis and Objectives

This PhD thesis hypothesises that integrating external commonsense knowledge into a smaller architecture in terms of parameters, compared to state-of-the-art LLMs, will boost the quality of the outputs when generating text. Those texts could be similar to the ones that could write a human when solving certain tasks. Moreover, external commonsense knowledge will help reduce the problem of hallucinations. After reviewing the scientific literature, we have observed that some works have been conducted to incorporate external commonsense in NLG systems. However, although they improved the results of previous works, the performance was still not excellent. Moreover, most of the research has been conducted in the English language. Therefore, the main objective of this research is to design an efficient architecture that will require less computational expenses than the state-of-the-art LLMs. This architecture will be capable of achieve human-like performance in generating text for different NLG tasks prioritising the research for Spanish. Additionally, a key focus will be on minimising the issue of hallucinations.

4. Methodology

To fulfil the objectives of this PhD thesis, we have proposed a methodology consisting of four main important milestones, that will be explained next: define a NLG pipeline; analyse, test and propose NLG architectures; analyse and build a corpus in Spanish; and evaluate proposed NLG architectures.

1. **Define a NLG pipeline.** The first step of my research was to propose a pipeline to follow. It involves the steps of defining the task I want to address, the corpus and external knowledge collection, how to integrate both in a NLG architecture and finally evaluating the results obtained. Figure 2 shows the sequential steps of this pipeline. The task this research project is addressing first is the commonsense generation through a concept-to-text task. Given some words, the goal is to generate a sentence that contains those words and has common sense. Afterwards, more NLG tasks will be addressed.
2. **Analyse, Test and Propose NLG Architectures.** During the initial part of the research, we explored different available NLG architectures in English and Spanish to find a suitable model to address the NLG task based on the topics pursued. Nowadays, most of the architectures employed are based on LLMs. Despite this, we will first test and compare traditional and neural architectures to effectively decide which architecture performs better based on a set of decisions, such as content produced, efficiency, etc. Consequently, we will try to integrate commonsense knowledge and

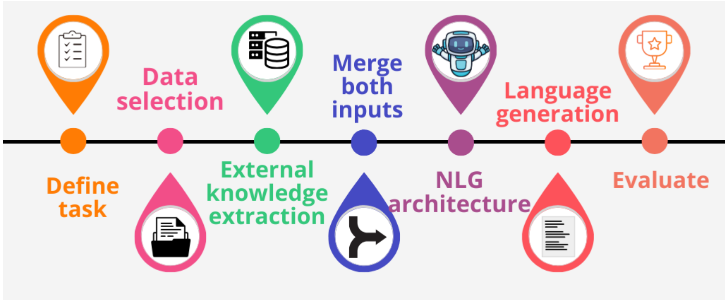

Figure 2: Proposed pipeline.

controllable generation techniques into the best-performing architecture obtained from our experimentation to raise its performance and compare the results against the state-of-the-art models.

3. **Analyse and Build Corpus in Spanish.** Datasets constructed for a specific task in Spanish are hard to find, as most of the available corpora are in the English language. Moreover, in some cases, these datasets may be biased [31]. Once we have proposed an efficient architecture, we will focus on adapting that architecture to address different NLG tasks in Spanish, such as commonsense generation, text simplification, or abstractive summarisation. Therefore, we will have to collect and build specific corpora in Spanish to address these tasks. Moreover, as we are concerned about the importance of the training data on the results of the NLG model, we will check and preprocess, if necessary that corpora to try to be unbiased and balanced.
4. **Evaluate Proposed NLG Architecture on different tasks.** An important part of this research is to measure the performance of our proposed model. Most NLG metrics tend to compare the generated text against some reference texts. As NLG systems can generate texts in a large number of different styles, these metrics may not be appropriate. To be able to determine which metric is the most appropriate to measure the performance of each proposed task an exploratory analysis will be conducted and compared with a manual evaluation of those tasks.

5. Experiments

5.1. Conducted experiments

Two experiments have been conducted: a preliminary experiment on English NLG architectures and the problems in evaluating its results, and a similar experiment in Spanish.

1. English NLG models evaluation: This experiment is detailed in [32]. We conducted an experiment for the task of commonsense generation based on the CommonGen dataset [18]. Based on a set of concepts, the systems must generate a sentence with common sense that contains those concepts. We experimented with a modular architecture, and a global architecture and evaluated the results. We evaluated them manually and automatically. The manual evaluation showed that global architectures performed better than the modular ones. However, none of them performed in an acceptable manner. Furthermore, in the automatic evaluation, the employed metrics did not reflect the difference in quality between both outputs, despite performing overall the global architectures on those metrics.
2. Spanish NLG preliminary experimentation: Following the ideas of the CommonGen dataset, we wanted to extend the CommonGen task to other languages, in this case Spanish. To do so, we have proposed a Spanish corpus called COCOTEROS [33]. We have collected Spanish text from the Tatoeba dataset[2] and formatted to be similar to the CommonGen. With this Spanish corpus, we have fine-tuned a T5 Spanish model to generate texts to evaluate their performance.

[2]Available at https://tatoeba.org/es/

We also performed a manual and automatic evaluation. The manual evaluation shows that the produced sentences are not acceptable. Furthermore, in some cases, the generated sentences do not contain all the words given as keywords. For the words "despedir", "aeropuerto", and "amigo" the system generated the sentence "él se despidieron que vinieras al aeropuerto.". We also evaluated automatically the sentences with the same metrics as for the English experiment. Table 1 shows the results obtained by this model in different NLG evaluation metrics and compared with the results obtained for the T5 model in English of the previous experiment. Although they are two different datasets, we can see that the results achieved for Spanish are lower than the results obtained in the English experimentation. Those results align with the manual evaluation, which showed a worse performance for the Spanish model than for the English model.

Model	SPICE	CIDEr	BLEU_1	ROUGE-L	Readability	Cosine	BERTScore
T5-Small Spanish	0.04	0.02	0.042	0.081	0.325	0.244	0.87
T5-Small English	0.256	0.109	0.600	0.444	0.215	0.287	0.914

Table 1
Results obtained by the carried experiments.

5.2. Planned experiments

Considering the obtained results, the next steps are the experimentation with different architectures that integrate external commonsense and the search for the optimal way to evaluate the generated outputs, which can be done manually or to find other more appropriate automatic metrics.

1. External commonsense Integration: To figure how to obtain external knowledge for the task of concept-to-text and how to include it in the proposed model are two key points to incorporate commonsense in NLG architectures. There are different options such as ConceptNet, WordNet or COMET that could be interesting to explore. The research question is how to successfully extract knowledge from them and incorporate it into a NLG system.
2. Output evaluation: As demonstrated in [32], most common automatic evaluation metrics do not align completely with human performance, only measuring the overlapping with a reference sentence. Thus, finding other methods to evaluate the generated text is important to measure the performance of the models. Some works already done to measure hallucinations could be interesting to explore to evaluate outputs. Examples are AlignScore metric [34] and Hallucination_evaluation_model [35].

6. Research issues to discuss

As I explained, my research is currently focused on the task of concept-to-text generation. This task is based on giving some concepts, the systems must generate a sentence using those concepts. We have compiled a Spanish corpus named COCOTEROS to accomplish this task. The next step is to incorporate external commonsense into some pre-trained models to enhance the results obtained. Given that, the research questions that arise are the following:

- Which is the most adequate external knowledge base to complement this task?
- In which part of the pipeline is more efficient to incrust the knowledge, during fine-tuning, during inference, or in both?
- How the hallucination issue can be measured for this task in Spanish?
- If we use LLMs to address this task, will the results be improved considerably?

Acknowledgments

This research work is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"

References

[1] J. Sun, Q. V. Liao, M. Muller, M. Agarwal, S. Houde, K. Talamadupula, J. D. Weisz, Investigating explainability of generative AI for code through scenario-based design, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 212–228.

[2] E. Reiter, R. Dale, Building applied natural language generation systems, Natural Language Engineering 3 (1997) 57–87. doi:10.1017/S1351324997001502.

[3] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

[4] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, A. F. Aji, P. Nakov, M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, arXiv:2305.14902 (2023).

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.

[6] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[7] E. Ferrara, Should ChatGPT be biased? challenges and risks of bias in large language models, arXiv preprint arXiv:2304.03738 (2023).

[8] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, Y. Zhang, Evaluating the logical reasoning ability of ChatGPT and GPT-4, arXiv preprint arXiv:2304.03439 (2023).

[9] K. Maurya, M. Desarkar, Towards low-resource language generation with limited supervision, in: Y. Elazar, A. Ettinger, N. Kassner, S. Ruder, N. A. Smith (Eds.), Proceedings of the Big Picture Workshop, Association for Computational Linguistics, Singapore, Singapore, 2023, pp. 80–92. URL: https://aclanthology.org/2023.bigpicture-1.7. doi:10.18653/v1/2023.bigpicture-1.7.

[10] D. D. McDonald, Natural language generation., Handbook of natural language processing 2 (2010) 121–144.

[11] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, M. Yang, A survey of natural language generation 55 (2022). URL: https://doi.org/10.1145/3554727. doi:10.1145/3554727.

[12] E. Reiter, Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?, 1994. arXiv:cmp-lg/9411032.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[14] S. Mahamood, M. Clinciu, D. Gkatzia, It's common sense, isn't it? demystifying human evaluations in commonsense-enhanced NLG systems (2021).

[15] H. Wang, Y. Liu, C. Zhu, L. Shou, M. Gong, Y. Xu, M. Zeng, Retrieval enhanced model for commonsense generation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3056–3062. URL: https://aclanthology.org/2021.findings-acl.269. doi:10.18653/v1/2021.findings-acl.269.

[16] Z. Aghahadi, A. Talebpour, Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning, Journal of Applied Non-Classical Logics 32 (2022) 55–71.

[17] Y. Zhang, X. Wan, SituatedGen: Incorporating geographical and temporal contexts into generative commonsense reasoning, arXiv preprint arXiv:2306.12552 (2023).

[18] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, X. Ren, CommonGen: A constrained

text generation challenge for generative commonsense reasoning, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1823–1840. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.165.

[19] F. Carlsson, J. Öhman, F. Liu, S. Verlinden, J. Nivre, M. Sahlgren, Fine-grained controllable text generation using non-residual prompting, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 6837–6857.

[20] K. R. Chandu, A. W. Black, Positioning yourself in the maze of neural text generation: A task-agnostic survey, 2020. URL: https://arxiv.org/abs/2010.07279. doi:10.48550/ARXIV.2010.07279.

[21] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[22] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[23] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[24] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic propositional image caption evaluation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, Springer, 2016, pp. 382–398.

[25] P. Stanchev, W. Wang, H. Ney, EED: Extended edit distance measure for machine translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), 2019, pp. 514–520.

[26] M. Popović, chrF: character n-gram f-score for automatic MT evaluation, in: Proceedings of the tenth workshop on statistical machine translation, 2015, pp. 392–395.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[28] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: International conference on machine learning, PMLR, 2015, pp. 957–966.

[29] W. Yuan, G. Neubig, P. Liu, BARTScore: Evaluating generated text as text generation, Advances in Neural Information Processing Systems 34 (2021) 27263–27277.

[30] J. Fu, S.-K. Ng, Z. Jiang, P. Liu, GPTScore: Evaluate as you desire, arXiv preprint arXiv:2302.04166 (2023).

[31] J. S. Ernst, S. Marton, J. Brinkmann, E. Vellasques, D. Foucard, M. Kraemer, M. Lambert, Bias mitigation for large language models using adversarial learning (2023).

[32] I. Martínez-Murillo, P. Moreda, E. Lloret, Analysing the problem of automatic evaluation of language generation systems, Procesamiento del Lenguaje Natural 72 (2024) 123–136.

[33] M. M. Maestre, I. Martínez-Murillo, E. Lloret, P. Moreda, A. Suárez Cueto, COCOTEROS: A spanish corpus with contextual knowledge for natural language generation, in: 40th Annual Conference of the Spanish Association for Natural Language Processing 2024: Posters (SEPLN-P 2024), CEUR, 2024.

[34] Y. Zha, Y. Yang, R. Li, Z. Hu, AlignScore: Evaluating factual consistency with a unified alignment function, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11328–11348. URL: https://aclanthology.org/2023.acl-long.634. doi:10.18653/v1/2023.acl-long.634.

[35] S. M. Hughes, M. Bae, Hughes hallucination evaluation model (hhem) leaderboard, Huggingface, available at https://huggingface. co/spaces/vectara/leaderboard (accessed 22nd April, 2024) (2024).