The CLEF 2024 Monster Track: One Lab to Rule Them All

Nicola Ferro[1], Julio Gonzalo[2], Jussi Karlgren[3], and Henning Müller[4,5]

[1] University of Padova, Italy
[2] UNED, Madrid, Spain
[3] SiloGen, Helsinki, Finland and Stockholm, Sweden
[4] HES-SO Valais, Switzerland
[5] University of Geneva, Switzerland

Abstract. Generative *Artificial Intelligence (AI)* and *Large Language Models (LLMs)* are revolutionizing technology and society thanks to their versatility and applicability to a wide array of tasks and use cases, in multiple media and modalities. As a new and relatively untested technology, LLMs raise several challenges for research and application alike, including questions about their quality, reliability, predictability, veracity, as well as on how to develop proper evaluation methodologies to assess their various capacities.

This evaluation lab will focus on a specific aspect of LLMs, namely their versatility. The CLEF Monster Track is organized as a meta-challenge across a selection of tasks chosen from other evaluation labs running in CLEF 2024, and participants will be asked to develop or adapt a generative AI or LLM-based system that will be run on all the tasks with no or minimal task adaptation. This will allow us to systematically evaluate the performance of the same LLM-based system across a wide range of very different tasks and to provide feedback to each targeted task about the performance of a general-purpose LLM system compared to systems specifically developed for the task. Since the datasets for CLEF 2024 have not yet been released publicly, we will be able to experiment with previously unseen data, thus reducing the risk of contamination, which is one of the most serious problems faced by LLM evaluation datasets.

Keywords: Generative language models · LLM · Shared task · Quality benchmarks · CLEF

1 Motivation and Objectives

Generative *Large Language Models (LLMs)*, both proprietary models such as *Generative Pre-trained Transformer (GPT)* [16, 17], and open models (i.e. which provide free access to the model weights), such as *Large Language Model Meta AI (LLaMA)* and its derivatives [4, 23–25] are being successfully applied to a wide range of tasks, covering multiple media and modalities.

As a consequence, LLMs attract considerable attention from the general public, from research teams and from industry. Much effort is put into investigating the various capacities of LLMs with respect to their quality, reliability, reasoning capabilities and more. Many dataset ensembles are being adapted and used to evaluate the overall performance of LLMs, but overall there are still several challenges to address. In particular (i) the evaluation is too often compromised because test data is publicly available and models have seen the ground truth data in the pre-training phase; this problem is known as *contamination*, and is severe[6][7] for details; (ii) with the goal of testing anthropomorphic properties of models – such as common sense reasoning – and linguistic competence, datasets are drifting away from current practical application challenges.

Our goal is to systematically explore how well a given LLM performs across several real-world application challenges with respect to algorithms specifically trained for each task, avoiding contamination. We are inspired by the work of Romei et al. Hromei et al. [11], who used a single monolithic LLM to participate in all 13 EVALITA tasks[8] in 2023 – the national evaluation campaign on *Natural Language Processing (NLP)* and speech tools for Italian language – including Affect Detection, Authorship Analysis, Computational Ethics, Named Entity Recognition, Information Extraction, and Discourse Coherence. Hromei et al. performed a single fine-tuning with all training data from the 13 tasks, and found that their model achieved first place in 41% of the subtasks and showcased top-three performance in 64% of them, without any task-specific prompt engineering phase. We know of no similar experiments in the *Information Retrieval (IR)* field or in other large-scale evaluation campaigns, to systematically explore cross-task performance in a shared-task setup.

Therefore, the CLEF *Monster Track*[9] will be organized as a meta-challenge across a selection of tasks chosen from the other labs running in CLEF 2024 and participants will be asked to develop a generative AI/LLM-based system that will be run against all the selected tasks with no or minimal adaptation. For each targeted task we will rely on the same dataset, experimental setting, and evaluation measures adopted for that specific task. In this way, the LLM-based systems participating in the CLEF Monster Track will be comparable directly with the specialized systems participating in each targeted task.

This allows us to systematically evaluate the performance of the same LLM-based system across a wide range of very different tasks and to provide feedback to each targeted task about the performance of a general-purpose LLM system compared to systems specifically developed for the task. Moreover, since the datasets for CLEF 2024 are in large part not public, yet, we will be able to experiment with previously unseen data, thus avoiding the risk of contamination.

[6] See the *LM contamination index*

[7] https://hitz-zentroa.github.io/lm-contamination

[8] https://www.evalita.it/

[9] https://monsterclef.dei.unipd.it/

2 Benchmarks for quality assessment of generative language models

There is already a considerable and varied body of work on quality assessment of generative language models with a rich selection of benchmark resources. Many of these address capacities beyond that of generating fluent and grammatically correct language. Current evaluation procedures range over common sense reasoning [1, 7, 15, 19, 22, 26], world knowledge [12, 14], reading comprehension [5, 6], math capabilities [8], and coding tasks [3]. Some popular aggregated benchmarks are *Massive Multitask Language Understanding (MMLU)* [10], *BIG-Bench Hard (BBH)* [21] and *Artificial General Intelligence (AGI)* Eval [27]. Other examples include Chen et al. [2], with a dataset in both Chinese and English to evaluate how well LLMs avoid hallucinating by making use of a *Retrieval-Augmented Generation (RAG)*; Gao et al. [9] with a dataset for evaluating how well LLMs generate text with citations, improving factual correctness and verifiability of the generated output; Kamalloo et al. [13] with a dataset for building end-to-end generative information-seeking models that are capable of retrieving candidate quotes and generating attributed explanations; and [13] Rashkin et al. [18] with a dataset and a two-stage annotation pipeline to evaluate attribution capacity of LLMs.

As we mentioned in the introduction, two remaining challenges are that (i) evaluation can be compromised by contamination issues (since evaluation material can be seen by the model in the pretraining process), and (ii) the overarching goal of testing anthropomorphic properties and generality of systems built on generative language models may drift away from current practical application challenges.

3 Candidate CLEF tasks

Most CLEF labs can be used for evaluation of general-purpose technologies; but the Monster Track will primarily make use of tasks where language plays an important role and where data sets are novel — in contrast with those where public data sets are used that could have been used to train the participating LLMs. A number of candidate tasks from CLEF labs are candidates for inclusion in the Monster Track and the final selection will be made collaboratively with lab organizers.

3.1 CheckThat!

CheckThat![10] is a CLEF Lab devoted to combat misinformation. The task proposed for the Monster Track is *Check-worthiness*: given a tweet, systems must determine if it contains a claim that is worth fact checking. The organizers provide English, Arabic, and Spanish datasets to be used for instruction fine-tuning.

[10] https://checkthat.gitlab.io/clef2024/

3.2 ELOQUENT

ELOQUENT[11] is a CLEF 2024 lab devoted to the evaluation of certain quality aspects of content generated by LLMs. It intends to use LLMs to test the capacities of themselves, and is thus a good fit for the meta-lab evaluation effort. The ELOQUENT lab proposes four evaluation tasks for LLMs:

i Topical competence: Can an LLM assess itself if it is capable to process data in some application domain of interest?
ii Veracity and Hallucination: Can an LLM be used to evaluate the output of other LLMs to detect hallucinated or factually incorrect information?
iii Robustness: Will an LLM output the same content independent of input variation which is equivalent in content but non-identical in form or style?
iv Voight-Kampff task: Can an LLM be used to detect if some piece of text is written by a human author or generated by an LLM? This task will be organised in collaboration with the PAN lab at CLEF.

3.3 EXIST

EXIST[12] is a lab devoted to the detection and characterization of sexism in online content. Three tasks are proposed for the Monster Track:

i Sexism identification: given a tweet, systems must determine if it has sexist content or not.
ii Source intention: systems must determine if the sexist content is reported, judgemental or direct.
iii Sexism categorization: systems must classify sexist content into one of five categories of sexism (ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny and non-sexual violence).

The dataset contains tweets in English and Spanish: more than 3 200 tweets per language for the training set, around 500 per language for the development set, and nearly 1 000 tweets per language for the test set. A crucial characteristic of the dataset is that it provides six annotations per tweet and task, with each annotator belonging to one out of six cohorts (three age groups × two genders). All raw annotations are provided to participants, instead of a merged single ground truth; and all annotations in the test collection are considered in the evaluation process.

3.4 ImageCLEF

ImageCLEF[13] aims to provide an evaluation forum for the cross-language annotation and retrieval of images. For the Monster Track, ImageCLEF will provide two image caption tasks in the biomedical domain (radiological images):

[11] https://eloquent-lab.github.io/
[12] http://nlp.uned.es/exist2024
[13] https://www.imageclef.org/2024

i Concept detection where systems must predict a set of concepts (defined by the UMLS CUIs) based on the visual information provided by the radiology images.

ii Caption prediction which requires systems to automatically generate captions for the radiology images provided.

Both will use the ImageCLEFmedical 2024 Caption dataset, which consists of radiologic images of 7 different imaging modalities (angiography, CT, MRI, PET, ultrasound, X-ray, and combined modalities) with varying image dimensions as extracted from PubMed Open Access publications, along with the pre-processed image caption and a set of UMLS concepts.

3.5 LongEVAL

LongEval[14] is a shared task evaluating the temporal persistence of Information Retrieval systems and text classifiers. Two tasks are proposed:

i LongEval Retrieval: Retrieval systems are evaluated in terms of their retrieval effectiveness when the test documents are dated either right after (short term) or three months (long term) after the documents available in the train collection. The Longeval Websearch collection relies on a large set of data (corpus of pages, queries, user interaction) provided by a commercial search engine (Qwant).

ii LongEval Classification: Classification systems are evaluated in terms of their short-term effectiveness (test documents are dated shortly after training documents) and long-term effectiveness (test documents are dated more than one year apart from the training data).

3.6 PAN

The *PAN*[15] lab has organised numerous CLEF tasks related to authorship identification and verification, author profiling, plagiarism detection, and related tasks. This year PAN hosts four tasks, and two of them have been proposed to join the Monster Track effort:

i Multilingual Text Detoxification: Given a toxic piece of text in one of 7 languages, re-write it in a non-toxic way while preserving the content.

ii Voight-Kampff task (in collaboration with the ELOQUENT lab, see description above).

3.7 Touché

Touché[16] is a series of scientific events and shared tasks on computational argumentation and causality. Three tasks have been proposed for the Monster Track:

[14] https://clef-longeval.github.io/

[15] https://pan.webis.de/

[16] https://touche.webis.de/

i Human Value Detection: given a long text (in one of eight languages), for each sentence, identify which human values the sentence refers to and their level of attainment. The task employs a collection of roughly 3000 human-annotated texts between 400 and 800 words. The annotated values are those of the Schwartz' value continuum [20].

ii Ideology and Power Identification in Parliamentary Debates: given a parliamentary speech in one of several languages, identify the ideology of the speaker's party, and whether the speaker's party is currently governing or in opposition. The data for this task comes from ParlaMint[17], a multilingual collection of parliamentary debates.

iii Image Retrieval for Arguments. Given an argument, create a prompt for a text-to-image generator to generate an image that helps to convey the argument's premise. Organizers provide access to a Stable-Diffusion API for image generation.

3.8 Final Selection Procedure and Meta-Evaluation

Selection of tasks to include in Monster Track will be made from the above list of 21 candidate tasks using criteria such as:

i Suitability: it should be possible to address every Monster Track task using a single system based on a specific LLM: we assume that participants will have a limited time to adapt their systems to each of the proposed tasks, and this is in keeping with the objective to test the generality of a system based on a generative language model.

ii Diversity: we want Monster Track tasks to cover much of the broad variety exhibited by practical challenges in information access.

iii Contamination: the test sets for Monster Track tasks should have not been made available in the past, in order to eliminate or at least minimize contamination (the possibility that language models have been exposed to the ground truth in the pre-training phase).

Details on the evaluation procedure are yet to be decided, and we will focus on qualitative insights rather than crude competition. In any case, we will rank systems with at least two procedures: (i) the average effectiveness across tasks (once all official metrics from each task are mapped into the same scale); (ii) the average rank in each of the tasks. This second procedure is more informative, as it compares the Monster Track systems with all other dedicated systems in each of the tasks. We can average the rank at least in two ways: directly (average of the rank or the inverse rank) or via percentiles. Percentiles have the advantage that relativise a rank in terms of the number of elements in the rank, and give more credit to, e.g., a winner with 20 opponents than to a winner with 2 opponents.

[17] https://www.clarin.eu/parlamint

References

1. Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al.: Piqa: Reasoning about physical commonsense in natural language. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp. 7432–7439 (2020)
2. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking Large Language Models in Retrieval-Augmented Generation. arXiv.org, Computation and Language (cs.CL) **arXiv:2309.01431** (September 2023)
3. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
4. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. `https://lmsys.org/blog/2023-03-30-vicuna/` (March 2023)
5. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.t., Choi, Y., Liang, P., Zettlemoyer, L.: QuAC: Question answering in context. arXiv preprint arXiv:1808.07036 (2018)
6. Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: BoolQ: Exploring the surprising difficulty of natural yes/no questions. arXiv preprint arXiv:1905.10044 (2019)
7. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457 (2018)
8. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
9. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling Large Language Models to Generate Text with Citations. arXiv.org, Computation and Language (cs.CL) **arXiv:2305.14627** (May 2023)
10. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
11. Hromei, C.D., Croce, D., Basile, V., Basili, R.: ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme. In: Lai, M., Menini, S., Polignano, M., Russo, V., Sprugnoli, R., Venturi, G. (eds.) Proc. 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023), CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, `https://ceur-ws.org/Vol-3473/`. (2023)
12. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551 (2017)
13. Kamalloo, E., Jafari, A., Zhang, X., Thakur, N., Lin, J.: HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution. arXiv.org, Computation and Language (cs.CL) **arXiv:2307.16883** (July 2023)

14. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al.: Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019)
15. Mihaylov, T., Clark, P., Khot, T., Sabharwal, A.: Can a suit of armor conduct electricity? a new dataset for open book question answering. arXiv preprint arXiv:1809.02789 (2018)
16. OpenAI: GPT-4 Technical Report. arXiv.org, Computation and Language (cs.CL) **arXiv:2303.08774** (March 2023)
17. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Proc. 36th Annual Conference on Neural Information Processing Systems (NeurIPS 2022), `https://proceedings.neurips.cc/paper_files/paper/2022` (2022)
18. Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G.S., Turc, I., Reitter, D.: Measuring Attribution in Natural Language Generation Models. Computational Linguistics pp. 1–64 (August 2023)
19. Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM **64**(9), 99–106 (2021)
20. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
21. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
22. Talmor, A., Herzig, J., Lourie, N., Berant, J.: CommonsenseQA: A question answering challenge targeting commonsense knowledge. arXiv preprint arXiv:1811.00937 (2018)
23. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpaca: A Strong, Replicable Instruction-Following Model. `https://crfm.stanford.edu/2023/03/13/alpaca.html` (March 2023)
24. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models. arXiv.org, Computation and Language (cs.CL) **arXiv:2302.13971** (February 2023)
25. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I.,

Korenev, A., Singh Koura, P., Lachaux, M.H., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv.org, Computation and Language (cs.CL) **arXiv:2307.09288** (July 2023)

26. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830 (2019)
27. Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., Duan, N.: AGIEval: A human-centric benchmark for evaluating foundation models. arXiv preprint arXiv:2304.06364 (2023)