ELOQUENT 2024 — Topical Quiz Task

Jussi Karlgren[1], Aarne Talman[2]

[1] Silo AI, Helsinki
[2] University of Helsinki

Abstract

ELOQUENT is a set of shared tasks for evaluating the quality and usefulness of generative language models. ELOQUENT aims to apply high-level quality criteria, grounded in experiences from deploying models in real-life tasks, and to formulate tests for those criteria, preferably implemented to require minimal human assessment effort and in a multilingual setting. One of the tasks for the first year of ELOQUENT was the *Topical quiz*, in which language models are probed for topical competence. This first year of experimentation has shown — as expected — that using self-assessment with models judging models is feasible, but not entirely straight-forward, and that a judicious comparison with human assessment and application context is necessary to be able to trust self-assessed quality judgments.

1. Introduction

Generative language models ("LLMs") as a foundational component in an information system are able to handle a broad variety of input data robustly and elegantly, and are able to provide appropriately creative generated output to fit a broad range of application situations and the preferences of a diverse user population. An information service with a generative language model can be built to provide a flexible low threshold conversational interface for its users: there is considerable interest to put generative language models to use in productive practical applications, across domains, sectors of society, languages, and cultural areas.

The ELOQUENT lab is intended to probe the quality of a generative language model, and to do this by addressing specifically such quality issues that are raised at the deployment time when a model is included in a system for productive downstream tasks. The lab also intends to explore the reliability of system self-assessment of model quality using other models or even the same model, and to reduce the dependence of human-assessed gold standard data sets.

A generative language model in practical application will in most envisioned use cases be expected to stay within given task-appropriate topical boundaries, to generate material restricted to the domain it is employed to work within, and to have competence in the terminology and conventions of that domain. Examples of relevant topical domains could be business domains, such as finance [1] or healthcare [2], or even recreational activities such as sailing or basketball, ranging to differences in how a topic is treated differently across linguistic and cultural areas or in specific demographic groups.

The topical quiz task intends to answer to the need for verifying a model's understanding of an application domain of interest. The task is defined for a system to generate a topical quiz for some given topic; to respond to such quizzes, including the one it has generated itself; and to score responses to quizzes numerically from 1 to 10. Every participating team was given a list of topics, shared as a JSON structure, and asked to use their system or systems to generate a set of questions for each topic. The dataset includes a suggested prompt string, but participants were free to reformulate the string to fit their model or system. The generated questions were submitted in a prescribed JSON structure by the participants through a submission form. These question structures were shared back to the participants for them to use their systems to generated responses to the questions. The generated responses were then again submitted in a prescribed JSON structure by the participants using a submission form. These responses were then scored 1-10 by four systems: Reindeer-Poro, Reindeer-Mistral, GPT-SW3, and GPT-4o. An example topic with responses and scores is shown in Figure 2.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

Topic: Wine and terroir
Question (Reindeer-Poro):
What factors contribute to the unique terroir of a vineyard or region?
Response (Reindeer-Mistral): The factors that contribute to the unique terroir of a vineyard or region include the soil composition, drainage patterns, exposure to sunlight, temperature and rainfall patterns, and the presence of natural features such as mountains, rivers, and bodies of water. These factors can vary greatly from one vineyard to another, even within the same region, and can have a significant impact on the taste and character of the wine produced.
Response (Reindeer-Poro): Factors that contribute to the unique terroir of a vineyard or region include the soil type and composition, the climate, the topography, and the local microclimate. These factors can interact to influence the growth and development of the vines, as well as the flavors and characteristics of the wine.
Response (GPT-SW3): A unique terroir is the combination of natural conditions like soil composition, elevation, climate, and exposure to sunlight, as well as human factors such as farming methods, viticultural practices, and winemaking techniques that all contribute to creating a distinct quality in wine. The specific characteristics of each terroir are what make it distinctive, allowing for wines from one region to taste different than wines from another region with similar conditions.

	Reindeer-Mistral	Reindeer Poro	GPT-SW3	GPT-4o
Reindeer-Mistral	9	7	8	8
Reindeer-Poro	8	8	9	7.5
GPT-SW3	9	10	10	8

Figure 1: A sample topic for the Topical Quiz task and a sample quiz question by Reindeer-Poro; responses by Reindeer-Poro, Reindeer-Mistral, and GPT-SW3; and scores for the responses as given by Reindeer-Poro, Reindeer-Mistral, GPT-SW3, and GPT-4o

The task had 27 registered participant teams. Three teams submitted quizzes, with two teams submitting responses. The teams used Poro [3], Mistral [4] (for team "Reindeer" [5]), GPT-SW3 [6], and a modular RAG approach [7] (for team Verbanex). This rather limited set of submitted data does not provide us with support to use the originally planned fairly elaborate scoring scheme, comparing the correlation of scoring across systems and the reliability of the self-scoring of a system to scores others do.

The quizzes generated by the systems are of varying quality, but mostly quite acceptable. The responses are somewhat more varied, since in some cases, some of the systems failed to generate any response at all but instead repeated the questions, descended into a generative loop, or gave an overall summary of the question set instead of responding to them. Scoring questions clearly was the most demanding task, since even instruction trained models struggle with processing lists of items and assessing them in turn. A sample question for one of the topics is given in Figure 1 together with responses from three systems and scores for those responses by the systems themseles and GPT-4o. The scores are high and vary in fair consensus across the scoring systems for this specific question, with no clear preference exhibited for own answers. This is not as clearly the case for questions where there is a larger discrepancy in topical competence. Another example is given in Figure 2 where the responses to the question "What are the major rules and regulations changes in the NBA this season?" are quite different from each other and in many cases factually erroneous in that they disregard the "this season" specification and take up rule changes from previous seasons. The self-assessed scoring in this case is unreliable, limited by the actual topical competence of the system doing the scoring, and appears to be based more on the appearance of competence and clarity of expression rather than factual correctness.

A more comprehensive exposition of responses and their scoring is given in the task-specific overview paper.

Topic: Basketball News
Question (Reindeer-Mistral):
What are the major rules and regulations changes in the NBA this season?
Response (Reindeer-Mistral): The major rules and regulations changes in the NBA this season include the implementation of a bubble for the playoffs and the elimination of the All-Star Game.
Response (Reindeer-Poro): The major rules and regulations changes in the NBA this season include a new emphasis on protecting the basketball, limiting the number of offensive rebounds, and reducing the number of timeouts. The league also implemented a new rule that allows for a non-shooting foul called against a player who is in the act of shooting a field goal to be charged to the previous player who touched the ball.
Response (GPT-SW3): This is a complex question that involves many factors. Here's a summary of some of the most significant changes: *lengthy exposition of pandemic-related quarantine rules, salary cap regulations*

	Reindeer-Mistral	Reindeer Poro	GPT-SW3	GPT-4o
Reindeer-Mistral	7	6	8	4
Reindeer-Poro	6	6	8	3
GPT-SW3	9	10	8	6

Figure 2: A second more topically demanding sample topic for the Topical Quiz task and a sample quiz question by Reindeer-Mistral; responses by Reindeer-Poro, Reindeer-Mistral, and GPT-SW3; and scores for the responses as given by Reindeer-Poro, Reindeer-Mistral, GPT-SW3, and GPT-4o

2. Conclusion

The goal of the Topical Quiz task of the ELOQUENT lab was to evaluate the quality of LLMs by how well they can generate, respond to, and score in-domain questions. We also find that system performance varies highly for specific tasks, which does not yet allow for any systematic observations. The cross-model evaluation set-up proved to be challenging without the use of human annotations. This we will be working in coming editions of ELOQUENT, together with exploring new automatic ways of evaluating LLM-generated outputs.

Acknowledgments

This lab has been supported by the European Commission through the DeployAI project (grant number 101146490).

References

[1] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, Bloomberggpt: A large language model for finance, arXiv preprint: 2303.17564 (2023).

[2] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan, Large language models encode clinical knowledge, Nature 620 (2023).

[3] R. Luukkonen, J. Burdge, E. Zosa, A. Talman, V. Komulainen, V. Hatanpää, P. Sarlin, S. Pyysalo, Poro 34b and the blessing of multilinguality, arXiv preprint: 2404.01856 (2024).

[4] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[5] V. Neralla, S. Bijl de Vroe, Evaluating Poro-34B-Chat and Mistral-7B-Instruct-v0.1: LLM System Description for ELOQUENT at CLEF 2024, in: G. Faggioli, N. Ferro, M. Vlachos, P. Galuščáková,

A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[6] A. Ekgren, A. C. Gyllensten, E. Gogoulou, A. Heiman, S. Verlinden, J. Öhman, F. Carlsson, M. Sahlgren, Lessons learned from gpt-sw3: Building the first large-scale generative language model for swedish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), 2022.

[7] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).