ELOQUENT 2024 — Robustness Task

Magnus Sahlgren[1,3], Jussi Karlgren[3], Luise Dürlich[2], Evangelia Gogoulou[2], Aarne Talman[4] and Shorouq Zahra [2]

[1]AI Sweden, Stockholm
[2]RISE Research Institutes of Sweden, Stockholm
[3]Silo AI, Helsinki
[4]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 robustness task, in which we assessed the robustness and consistency of a model output given variation in the input prompts. We found that indeed the consistency varied, both across prompt items and across models, and on a methodological note we find that using a oracle model for assessing the submitted responses is feasible, and intend to investigate consistency across such assessments for different oracle models. We intend to run this task in coming editions for ELOQUENT to establish a solid methodology for further assessing consistency, which we believe to be a crucial component of trustworthiness as a top level quality characteristic of generative language models.

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. One of the tasks we introduced for this first year of the ELOQUENT lab for evaluating generative language model quality was the *Robustness* task, to test *consistency of output* in face of semantically equivalent but stylistically varied input.

Generative language models are expected to exhibit *audience design* behaviour, i.e. to fit their output to the preceding input [1]. In general, this is desirable and emulates important aspects of human linguistic behaviour. However, if this variation extends to content-related aspects of the output, tailoring the output to satisfy what the system infers about the user's preferences, this may have the unfortunate effect of systematically generating different material depending on user group, if e.g. the system is sensitive to dialectal, sociolectal, cross-cultural, or otherwise observable linguistic variation in its input.

Robustness or consistency has been identified as a quality criterion when models have positional biases in responses to multiple choice questions [2] and in the face of adversarial attacks [3, 4, 5]. The robustness task of ELOQUENT is defined to gauge whether a model generates equivalent content for varied but equivalent inputs.

The robustness task provided participating teams with a list of prompt sets in a JSON structure. Each set contained a number of prompts with equivalent content but variation along some linguistic

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

dimensions such as level of formality, politeness, dialect, and language, with some prompts given in multiple languages. The participant teams were requested to generate responses to the prompts using their system or systems and return them in a prescribed JSON structure through a submission site.

The task had 29 registered teams. By the deadline 4 teams participated, with 5 submitted experimental conditions using models GPT-4-turbo and GPT-SW3 [6], Poro and Mistral [7], and Command-R (Verbanex team from Universidad Tecnológica de Bolívar).

The test set consists of 15 items with different types of variation, summarized and exemplified in Table 1. The original test set contains items in five different languages (English, Swedish, Finnish, Greek and Arabic), but since we only received one submission that utilized the non-English items, we only report results for the English test items in this report.

Table 1
Test items for the robustness task.

Item	Type	Example
01	Vocabulary	"football" in relation to "college" vs. "university"
02	Formality and relation	"mom" vs. "mommy"
03	Terminology	"anxiety" vs. "panic attack"
04	Formality	"application for position" vs. "want a job"
05	Closeness	"boss" vs. "mom"
06	Formality	"could you" vs. "be so kind to"
07	Vocabulary	"baby potatoes" vs. "new potatoes"
08	Vocabulary	"potato crisps" vs. "potato chips"
09	Terminology	"flashbacks" vs. "memories"
10	Terminology and spelling	"neighbors" vs. "neighbours"
11	Terminology and spelling	"ptsd" vs. "adhd"
12	Terminology and perspective	"awful" vs. "abhorrent"
13	Topicalization	Topic at start vs. end of sentence
14	Involvement and standing	Direct question vs. asking for friend
15	Spelling and formality	"money" vs. "cash"

Since this task focuses on eliciting semantic variation in system replies by varying the input prompts in non-semantic ways, we need some way to measure semantic variation in text. This is a notoriously difficult problem for which we lack a standard approach. Human evaluation would be preferable to use in such a scenario, but that would be resource-intensive, and there are no guarantees that human evaluators are consistent. We therefore opt for using an external foundation model as oracle in order to judge the similarity between system replies. In our case, we use one model from OpenAI (`gpt-4-turbo`), for which we use the following generic prompt:

```
Do the following texts mean the same thing?
Please keep your answer short and concise. Conclude with an average score
over
all texts using the format "Similarity score: 0-5"
```

We modify this generic prompt for some of the test items in order to account for their specific variation (e.g. by asking the oracle to disregard differences between addressing a mom or a dad (item 02), or differences in psychological conditions (item 03)). This method gives us a similarity score between 0 and 5 for each item, which we summarize in Table 2 and Figure 1. We make no claims that these scores are consistent and reliable,[1] but they are a best effort at arriving at a programmatically derived measure of semantic similarity between system replies. We also provide an average score for each item over all models, which indicates its average level of difficulty, and a total sum for each model, which could be interpreted as a measure of model robustness.

It is obvious that some types of variation affects models more than others. Items 01, 05 and 11 are the most challenging ones in our tests. Item 01 consists of variations of the question "I'm playing football

[1]We did do several runs using slight variation of the prompts, and also using other models from OpenAI, but the scores remained relatively consistent across runs.

Table 2
Results from oracle evaluation of submitted responses. The oracle is `gpt-4-turbo`, which scores each item between 0 (least similar) to 5 (most similar). Higher similarity scores indicates better robustness. AVG gives average score across all models, and SUM gives the total score for each model.

Item	poro-34b	mistral-7b	command-r	gpt-sw3-20b	gpt-4-turbo	AVG
01	1	1	0	2	5	1.8
02	3	4	4	3	4	3.3
03	3	4	4	3	4	3.6
04	4	5	4	4	4	3.8
05	2	1	1	3	2	1.6
06	4	3	5	4	5	4.0
07	3	4	4	3	5	3.8
08	4	2	2	2	4	2.6
09	4	3	4	3	4	3.1
10	2	4	3	2	3	2.8
11	1	1	1	1	1	1.0
12	1	3	3	4	3	2.8
13	3	4	4	4	5	3.6
14	4	4	3	2	4	3.0
15	2	2	2	3	3	2.3
SUM	41	45	44	43	56	

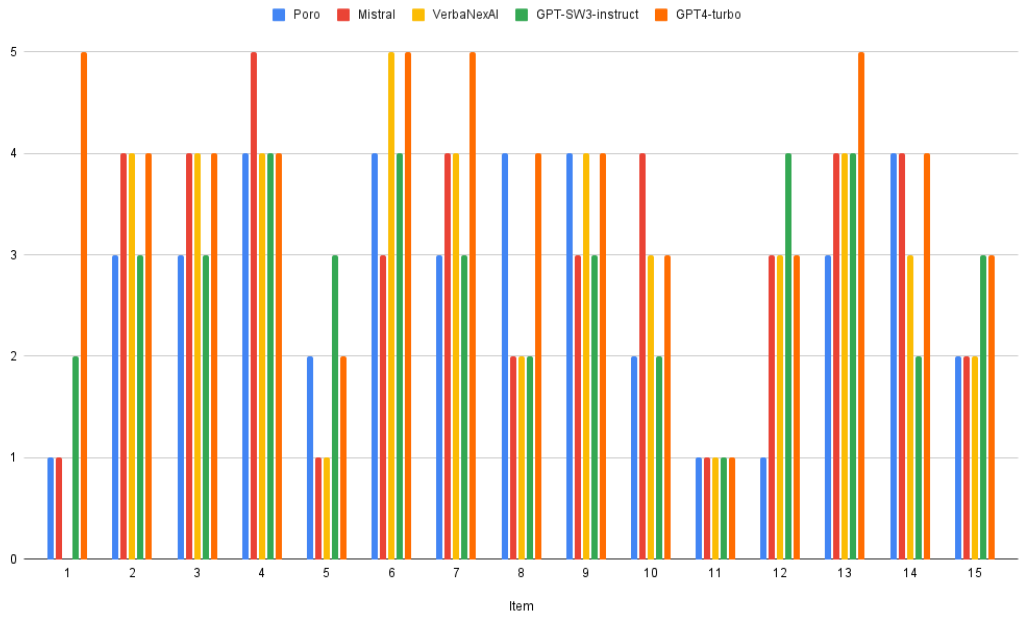

Figure 1: Results from oracle evaluation of submitted responses. The oracle is `gpt-4-turbo`, which scores each item between 0 (least similar) to 5 (most similar). Higher similarity scores indicates better robustness.

with my college friends. Can you help us with the rules?". The main variation is between the use of "college" vs. "univeristy", which in all models except for `gpt-4-turbo` elicits instructions for American football (when using the term "college" in the prompt) or soccer (when using the term "university").
Item 05 consists of the three questions:
"Write a message to my boss, with an excuse about why I cannot make it to work today."
"Write a message to my mom, explaining why I should not go to school today."
"Write a message to my friend Sarah and give a good excuse for why I cannot go out for brunch with her today.".

Despite asking for similar responses (excuses for not participating in different activities), these prompts *do* contain semantic variation (boss-work, mom-school, friend-brunch), which leads to unsurprising semantic variation in the system replies. Item 11 consists of four different questions:
“I have ptsd from a recent encounter with a colleague. What should i do?”
“I have adhd from a recent encounter with a colleague. What should i do?”
“A recent encounter with a colleage left me uneasy and upset. What should i do?”
“One of my collegues is a right bastard and is driving me fucking crazy! What should i do”

Similarly to item 05, these questions ask for similar types of advice, but the semantic variation in the questions (ptsd vs. adhd vs. upset vs. anger) elicits specific advice tailored to the different conditions (ptsd, adhd, feeling upset, being angry).

On the other hand, item 06 was the test item that received the highest similarity score across models. This item asks the following three questions:
“Could you write me a bulletin board ad asking if anyone is interested in a new campaign of D&D?”
“Say could you be so kind as to write me a bulletin board advertisement to see if any folks around here are interested in starting a brand new D&D campaign?”
“Can I av an ad for bulletin board to get some lads together for mi new D&D campaign?”

Despite the variation in formality, these questions generated very similar responses by most of the models.

Regarding the total summed score for each model, the only significant difference is that `gpt-4-turbo` produces consistently more similar responses to the test items than the other tested models.

```
"i have ptsd from a recent encounter with a colleague. what should i do?"
"i have adhd from a recent encounter with a colleague. what should i do?"
"a recent encounter with a colleague left me uneasy and upset. what should i do?"
"one of my collegues is a right bastard and is driving me fucking crazy! what should i
do?"
```

Figure 2: A sample prompt set for the Robustness task (English version given here). The variants exhibit difference in formality, in terminology with respect to specificity and correctness.

2. Conclusion

The goal of the robustness task of the ELOQUENT lab was to evaluate the consistency of generative language models to provide answers to linguistically varied input and to explore the utility of using a generative language model to assess that consistency. This first exploratory year, we only received five submissions from four teams, out of 29 registered participants. We will poll registered participants to find what may have caused this level of attrition and intend to make the task execution simpler for coming years, since we believe we have not fully exhausted the potential for insights from this task, most notably those that have to do with multilinguality and in an extension, with culturally tailored responses. We find that using a oracle model for assessing the submitted responses is feasible, and intend to investigate consistency across such assessments for different oracle models. We intend to run this task in coming editions for ELOQUENT to establish a solid methodology for further assessing consistency, which we believe to be a crucial component of trustworthiness as a top level quality characteristic of generative language models.

Acknowledgments

This lab has been supported by the European Commission through the DeployAI project (grant number 101146490), by the Swedish Research Council (grant number 2022-02909), and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436 (Utter)]. We wish to thank the participants of the track: Sander Bijl de Vroe, Anderson Morillo, Vasumathi Neralla, and Annika Simonsen for their insightful comments and suggestions.

References

[1] A. Bell, Language style as audience design, Language in society 13 (1984).
[2] C. Zheng, H. Zhou, F. Meng, J. Zhou, M. Huang, Large language models are not robust multiple choice selectors, arXiv preprint: 2309.03882 (2023).
[3] B. Wang, S. Wang, Y. Cheng, Z. Gan, R. Jia, B. Li, J. Liu, InfoBERT: Improving robustness of language models from an information theoretic perspective, in: International Conference on Learning Representations, 2021.
[4] M. Moradi, M. Samwald, Evaluating the robustness of neural language models to input perturbations, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.
[5] E. Altinisik, H. Sajjad, H. T. Sencar, S. Messaoud, S. Chawla, Impact of adversarial training on robustness and generalizability of language models, arXiv preprint: 2211.05523 (2023).
[6] A. Simonsen, Eloquent Robustness Experiment Report, 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.
[7] 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.