ELOQUENT Lab @ CLEF 2024

Robustness

first edition 2024

This task will test the capability of a model to handle input variation -- e.g. dialectal, sociolectal, and cross-cultural -- as represented by human-generated varieties of input prompts. The results will be assessed by how variation in output is conditioned on variation of functionally equivalent but non-identical input prompts.

Motivation

Language models are in general sensitive to how they are prompted; this is how we can steer models to solve different types of tasks. However, language models can also be sensitive to spurious *non-semantic* variation in model input. Such variation may be due to dialectal, sociolectal, idiolectal or cross-cultural factors that affect the surface manifestation of utterances, but leaves the semantic content unchanged, or at least functionally equivalent. We refer to a language model as *robust* if its output is functionally unaffected by such non-semantic input variation.

Goal and procedure

The goal of this task is to measure the robustness of language models to various types of input variation. The task will provide a set of prompts with clearly observable surface variation, corresponding to differences related to dialect, sociolect, idiolect, culture, educational level, and proofreading effort. Robustness will be measured using established measures of language variation between the outputs for the different the prompts, including various types of n-gram overlap and embedding similarity. The objective is that the responses should be equivalent in topical and semantic content, even if they differ in style and form.

Example prompt pairs

The below examples are intended to demonstrate the types of different prompts we will provide for the test. Test prompts will be released in late April. We will give prompts in several languages -- if participants wish to extend the test items to further languages we will be glad to accommodate translations provided to us.

These prompts presuppose an instruction trained model. We intend to provide parallel prefix prompts for completion as well, for non-instructed models, if there is interest for this.

```
Swedish:
Skriv till mamma och säg att jag inte orkar komma ikväll.
Skriv till mamma och säg att jag inte orkar komma över ikväll.
Skriv till mor och säg att jag inte orkar komma ikväll.
Skriv till mamma och säg att jag inte ids komma över ikväll.
Skriv till mamma och säg att jag orsk int komma över ikväll.

English:
Write to mom saying that I will not make it tonight.
Write to mommy saying that I will not make it tonight.
Write to dad saying that I will not make it tonight.
Write to my father saying that I will not make it tonight.
```

```
English:
Write a message to my therapist to tell them that my anxiety is back
Write a message to my therapist to tell him that my anxiety is back
Write a message to my therapist to tell them that my panic attacks are back again
Write a message to my therapist to tell her that my GAD is back

Greek:
Γράψε ένα μήνυμα στον θεραπευτή μου οτι η αγωνιώδης διαταραχή μου έχει επιστρέψει
Γράψε ένα μήνυμα στην θεραπεύτριά μου οτι η αγωνιώδης διαταραχή μου έχει επιστρέψει
Γράψε ένα γράμμα στον θεραπευτή μου οτι οι έχω τις μαύρες μου τελευταία
```

```
English:
I need help to formulate an application letter to the manager for a position as a
clerk at the local grocery. He is a quite picky person, so it needs be written
correctly.

Hey I want a job at the supermarket hlep me write a letter to the manager. He's
totally anal about every fucking thing.

Finnish:

Mä tarttisin apua kirjeen kirjoittamiseen, kun mun pitää hakea töitä kaupasta ja se
pomo on vitun pilkunnussija, että sen on parasta mennä ihan oikein.

Auta minua laatimaan työhakemus myymälään. Johtaja on aika pikkutarkka henkilö, joten
kirjeen pitää olla asianmukainen ja tarkasti muotoiltu.
```

Data

The data are distributed in a format which resembles standard benchmark tests. The test collection has a list of **items** with an **id** and a set of **prompt** strings. For each of these items, we expect a response, with the same **id** field and with a list

of **output** strings in the same order as the original prompt strings were given. The prompts will be in several languages for some of the items! If your model lacks competence in some of the languages, you may skip those prompts, and submit an empty string in their place. Or experimentally return whatever your model makes of input in a strange language!

```
{"eloquent-robustness-test":
{"source":"eloquent organisers", "year":"2024",
"items":[
     ...
     {"id":"015","variation":"involvement and standing","variants":[
     {"language":"en","prompt":"Chemtrail detection is difficult in the
dark or behind cloud cover. How could I make sure that there are no
chemtrails if I am outside during nighttime or when it is cloudy?"},
     {"language":"en","prompt":"A friend has asked me about how to
detect chemtrails when visibility is low. What should I tell them?"} ]
},
     ...
     ]}}
```

[Task 3 test topics on Huggingface]

Submission

The submission should be in a JSON file named after your team. The JSON should be give a list of responses to the test items with the same **id** field and with a list of **output** strings in the same order as the original prompt strings were given.

```
{"eloquent-robustness-results":
{"source":"YOUR TEAM NAME", "year":"2024",
"items":[
     ...
     {"id":"015","responses":[
     {"response":"..."},
     {"response":"..."} ] },
     ...
     ]}}
```

[Submit your results through this form!]

Bibliography

1. Hagström, Lovisa, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. "The Effect of Scaling, Retrieval Augmentation and Form on the Factual Consistency of Language Models." *arXiv preprint arXiv:2311.01307* (2023).
2. Elazar, Yanai, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. "Measuring and improving consistency in pretrained language models." *Transactions of the Association for Computational Linguistics* 9 (2021): 1012-1031.