ELOQUENT Lab @ CLEF 2024

Topical Competence

Task 1, first edition 2024

This task is intended to address the general question

"Is this model competent to process material about topic X?"

This task is about probing the topical competence of a language model-powered system and gauging the reliability in generative language models for assessing topical competence. It is part of the ELOQUENT lab on assessing the quality of generative language models.

Motivation

A generative language model in practical application will in most envisioned use cases be expected to stay within topical boundaries, to generate material restricted to the domain it is employed to work within, and to have competence in the terminology, thought patterns, and conventions of that domain.

Goals of the task

The generality of language models, and their ability to work across domains with little or no retraining is part of their appeal. This task investigates how to assess if a model, whether a foundation model or a fine-tuned model, is competent for some application domain of interest, e.g. to determine whether fine-tuning has beneficial effects on the quality of the output.

This task will also investigate if models can be self-assessed in a reliable way with minimal human effort.

Example domains

Examples of relevant topical domains could be business domains such as finance and healthcare or recreational activities such as sailing or basketball. We

observe that topics may be treated differently across application domains, linguistic and cultural areas, or demographic groups.

Task procedure

The task is to generate a topical competence test for a set of given topics in the form of a number of topically relevant questions. These questions will then be posed to other models participating in the task in a round-robin fashion, with the correctness of responses assessed by every model.

Participation involves the following steps:

1. For a series of 24 test topics given by the organisers, generate and submit a quiz of 12 questions each. (Monday, May 6 AoE)
2. Accept several quizzes on those same 24 topics generated by all participants, respond to the questions, and submit the responses. (Tuesday-Wednesday, May 7-8 AoE)
3. Accept a quiz together with several sets of submitted responses, score all the responses on a scale of 0-10, and submit the scores. (Thursday-Friday, May 9-10 AoE).
4. Receive results on experimental results (May 21).
5. Submit a lab report for the lab proceedings (May 31).

This means that participating in this task will involve being available for the above three operations during the week of May 6. The quiz - step (1) above - needs to be submitted to us by May 6th, and the following steps in the days immediately following.

Task scoring

The first objective of this task is to assess how reliably a system can generate a quiz. If a system generates a quiz the scores for which vary greatly from those of other systems' quizzes on the same topic we will assume that the quiz is strange in some way and that the system did not do a good job. (We will do manual error analysis to determine if the converse might be true!)

A. Each quiz will be scored with how well the scores on that quiz correlate with other submitted quizzes on the same topic.
B. Each system will be scored by how well its quizzes score by (A).

The second objective is to assess how reliably a system can score responses to a quiz. If a system scores responses to a quiz very differently from how other systems score the same quiz, we will assume that the system did not do a good job. (Again, we will do manual error analysis to determine if the converse might be true!)

C. Each system will be scored on how well its scores for a set of responses correlate with other systems' scores over all topics.

The third, and possibly most interesting objective from the perspective of the participants, is that we will compare the systems' topical competences.

D. Each system will be scored on how well it scores on a topic (over all the quizzes it responds to). Max score will be 120: a score of ten on all 12 questions on every quiz. We do not yet know how many quizzes on a topic there will be: this depends on the number of submitted results in (1) above.
E. Each system will also be scored on how well its own scoring correlates with others, to assess if it can be trusted to score its own homework, as it were.

The fourth objective is to compare topics across systems.

F. Each topic will be scored on how difficult it appears to have been for the systems.

Data

The data format resembles standard benchmark tests such as those found in Big Bench or LM Harness. The test items have a standard **prompt** string, and a list of items with both a terse **title**, suitable to append to the prompt string, and a somewhat more verbose **description** giving the context for the topic which can be used as a supplementary prompt. The items may also optionally have an **objective** field to make clear the reason for generating the test. These fields may all be used as a prompt, and participants should indicate which of the fields they have made use of. If participants wish to change the suggested prompt string to

something more suitable, this is allowed, but this must be reported upon submissions and announced on the mailing list to allow others to consider using the same enhanced prompt if they would so wish.

Task 1 sample and test topics on Huggingface

Submission format

If your system cannot produce reliable JSON, you can fix the format manually. We are not testing JSON competence here!

You may submit several quizzes (within reason, use your judgment for how many), e.g. if you experiment with some settings for your system or formulations for your prompts. Submit them in several files, and give each a unique name by appending a distinguishing token or integer to your team name.

Check the JSON of your submission before uploading – use e.g.

https://jsonlint.com/

to validate it!

Submission 1: quiz

We expect the participating systems to accept a topic in the format defined above and to return a test quiz, suitable for submission to other systems.

```
{"topicaltests":{"language":"en","source":"chatgpt
gpt3.5","id":"eloquent-2023-task01-sample","quiz":{"topicid":"001","title":"e-
scooters", "test_items":[{"qid":"001", "question":"What is an electric
scooter, and how does it differ from traditional scooters or
motorcycles?"},{"qid":"002","question":"Can you briefly explain the key
components of an electric scooter?"},{"qid":"003","question":"Why have
electric scooters gained popularity in urban areas around the
world?"},{"qid":"004","question":"Name a few cities where electric scooters
are commonly found as a mode of transportation."}, ... ]}}}
```

Submit your quiz here!

Submission 2: responses

We expect the participating systems to accept a number of test quizzes as above and to return responses, in a format scorable by other systems. Submit

each response in a separate file, using the "responseid" tag to distinguish them with your team name and the above system specific token in case you submit several ones.


```
{"topicaltestresponses":{"language":"en","source":"chatgpt
gpt3.5","quizid":"eloquent-2023-task01-sample","responseid":
"eloquent-2023-task01-sample-sample01",
"quizresponses":{"topicid":"001","title":"e-scooters","question":"What is an
electric scooter, and how does it differ from traditional scooters or
motorcycles?", "responses":[{"qid":"001","response":"An electric scooter is a
two-wheeled vehicle powered by an electric motor. It differs from traditional
scooters by using electric energy for propulsion."},
{"qid":"002","question":"Can you briefly explain the key components of an
electric scooter?","response":"Key components include the electric motor,
rechargeable battery, braking system, and control mechanisms."},
{"qid":"003","question":"Why have electric scooters gained popularity in urban
areas around the world?","response":"Electric scooters have gained popularity
due to their environmental friendliness, cost-effectiveness, and
convenience."}, {"qid":"004","question":"Name a few cities where electric
scooters are commonly found as a mode of transportation.","response":"They are
commonly found in cities like San Francisco, Paris, and Berlin, where they
provide a quick and efficient means of transportation."}]}}}
```


Submit your responses here!

Submission 3: scores

We expect the participating systems to accept responses for a test quiz as above and to return scores for each response. The motivation field is optional.


```
{"topicaltestscores":{"language":"en","source":"chatgpt
gpt3.5","scorer":"chatgpt gpt3.5,"quizid":"eloquent-2023-task01-sample",
"responseid":"eloquent-2023-task01-sample-sample", "scoreid":
"responseid":"eloquent-2023-task01-sample-sample-sample",
"quizresponses":{"topicid":"001", "title":"e-scooters",
"scores":[{"qid":"001","score":"9","motivation":"The response provides a clear
and concise overview of electric scooters and their key components."},
{"qid":"002","score":"9","motivation":"The response provides a clear and
concise overview of electric scooters and their key components."},
{"qid":"003","score":"8","motivation":"The answer effectively highlights the
reasons for the popularity of electric scooters."},
{"qid":"004","score":"8","motivation":"The answer mentions some cities where
they are commonly found."}]}}}
```


Submit your response scores here!

Multilinguality

This first year, no special provisions for multi-linguality will be made. The topics will be provided in English. If your system does not accept English-language input, feel free to translate the topic and the quizzes to your system's preferred language - manually or automatically - and share them to other participants. Please contact us to discuss how this might be done as smoothly as possible.