Sl

- Chairs have requested users to enter domain conflicts. Please click here to enter your conflict domains.

# View Reviews

**Paper ID**
12

**Paper Title**
Can You Answer This? - Exploring Zero-Shot QA Generalization Capabilities in Large Language Models

**Track Name**
Extended Abstract (non-archival)

**Reviewer #1**

## Questions

**1. Summary**
The abstract investigates the length of the generated answers produced by transformer-based language models (TLM) to see if they are capable of producing long text predictions, as longer answers is often required in tasks of closed-domain datasets.

The abstract also investigates the semantic similarity of polysemous words in closed-domain datasets, i.e. words that have many and similar meanings.

By counting the average number of predicted characters in one closed-domain dataset (TechQA) and one out-of-domain dataset (SQuAD) they find that the TLM produce shorter predictions than a recurrent and a convolutional text model.

When measuring the cosine similarity between the representation of different words with similar semantic meanings (in a general setting, but may have dissimilar meaning in a CDD), they find that TLMs finetuned on the close-domain dataset is not able to differentiate between such words.

**2. Strengths**
The introduction is good, although it raises more questions than the extended abstract can answer.

The paper takes an interesting approach in investigating transformer-based language models. Investigations of similarities of word representations in different finetuned models is an interesting approach, that may lead to insight to the effects and mechanisms of pretraining and finetuning.

**3. Weaknesses**
Although the average character length may be an important dimension for answers in closed-domain datasets, the analysis seems a bit primitive. The analysis seems not thorough enough, and is not discussed properly. The results of zero-shot performance is not addressed. It would be interesting to see some qualitative outputs and also further analysis of the output of different finetuned models.

The semantic similarity study is interesting, but lacks of attempts to investigate or explain the results further, especially when the results are somewhat counterintuitive.

Unfortunately I find the paper to be somewhat poorly structured without conveying a clear message. There are also multiple typos and at times very informal language. All tables would benefit from more extensive table captions.

**4. Overall score**
Weak Reject

**5. Justification of rating**
The extended abstract conducts two experiments that are well motivated. Unfortunately both lack discussion and analysis of the results. In addition the abstract could be written more clearly, so it is easier to understand what and how the experiments were conducted, and (not least) what insight they bring. Some of the results presented are not addressed by the author.

**Reviewer #2**

---

## Questions

**1. Summary**
References should be improved and the format should be the same. Similarly for the table also it should be the same for all.

**2. Strengths**
Good writing , Good topic

**3. Weaknesses**
Format should be same for all

**4. Overall score**
Weak Accept

**5. Justification of rating**
Need some minor format revision

**6. Additional comments and/or questions to the authors**
The paper can be improved in tems of formating and english writing rest of the remain good.