# Assessing the Benchmarking Capacity of MRC Datasets (#7755)

*S. Sugawara (UTokyo), P. Stenetorp (UCL), K. Inui (Tohoku U / RIKEN AIP), A. Aizawa (NII)

## Context word shuffle

**C:** Chris Ulmer, the 26-year-old teacher in Jacksonville starts his class by calling up each student individually to give them much admiration and a high-five. I couldn't help but be reminded of Syona's teacher and how she supports each kid in a very similar way.
**Q:** What can we learn about Chris Ulmer?
**A:** He praises his students one by one (multiple choice)

**C:** his help a in calling class but Syona's starts each 26-year-old similar individually Ulmer, and Chris admiration way. Jacksonville kid much I by couldn't them the a to supports of in student and teacher each be teacher reminded give how she high-five. up very
**Q:** What can we learn about Chris Ulmer?
**A:** He praises his students one by one (multiple choice)

## Vocabulary anonymization

**C:** Immediately behind the basilica is the Grotto, a Marian place of prayer. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858.
**Q:** To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
**A:** Saint Bernadette Soubirous

**C:** @adverb1 @prep5 @other0 @noun17 @verb2 @other0 @noun20 […] @other0 @noun20 @prep6 @noun25 @punct0 @noun26 @wh0 @other0 @noun7 @noun8 @adverb3 @verb4 **@prep4** @noun27 @noun28 @noun29 **@prep2** **@number0** @period0
**Q:** **@prep4** @wh2 @verb6 @other0 @noun7 @noun8 @adverb4 @verb4 **@prep2** **@number0** @prep2 @noun25 @noun26
**A:** @noun27 @noun28 @noun29

Both are solved → What is truly required for answering these questions?? 🤔🤔🤔

# Assessing the Benchmarking Capacity of MRC Datasets (#7755)

*S. Sugawara (UTokyo), P. Stenetorp (UCL), K. Inui (Tohoku U / RIKEN AIP), A. Aizawa (NII)

- Motivation
  - Improve the task explainability of MRC: what is required for answering questions?
  - Evaluate MRC datasets for benchmarking language understanding precisely. 🚨🚨🚨
- Solution
  - Propose 12 ablation-based methods (e.g., dropping function words & shuffling sentence words) along with requisite skills (e.g., grammar & compositionality).
  - Evaluate a baseline model on 10 datasets (e.g., SQuAD, CoQA, & RACE) 📊📊📊
- Results
  - The baseline model exhibits remarkably high performance on the ablation tests (e.g., relatively 90% of the original score even if shuffling sentence words) 👻👻👻
  - Most of the questions already answered correctly by a baseline model do not necessarily require grammatical & complex understanding. 🤔🤔🤔🤔🤔🤔🤔