Datasheet for TRAM Benchmark

OVERVIEW

Motivation and Intended Uses

1. What are the intended purposes for this benchmark?

The benchmark is designed to establish a standard for evaluating temporal reasoning in large language models. It focuses on three key areas: Foundational Temporal Understanding (such as Duration and Frequency), Temporal Interpretation and Computation (including Ambiguity Resolution and Arithmetic), and Advanced Temporal and Conceptual Understanding (encompassing areas like Causality and Storytelling).

2. Was it designed to address a specific task or fill a particular gap in research or application?

The benchmark is curated to address the need for a robust and comprehensive tool, specifically designed to evaluate temporal reasoning in large language models. It provides a diverse set of tasks that challenge models in the more intricate aspects of temporal reasoning.

Limitations and Inappropriate Uses

3. Are there any specific tasks or applications for which this benchmark should not be used?

The focus of the benchmark is on understanding and interpreting time-related concepts. Therefore, it may not be suitable for evaluations that significantly diverge from temporal reasoning, such as tasks involving texts that require contextual emotional intelligence, or domain-specific applications in medical or legal document analysis.

DETAILS

Composition

4. What do the instances that comprise the benchmark represent?

The instances consist of multiple-choice questions, created from a combination of existing datasets and human-curated problems, with a focus on temporal reasoning tasks. Each instance is specifically designed to assess a language model's ability to process and reason about time in natural language.

5. How many instances are there in total (of each type, if appropriate)?

There are a total of 526,668 problems. Specifically, the dataset comprises 10 main tasks and 38 subtasks. The number of problems for each main task is as follows: Ordering (29,462), Frequency (4,658), Duration (7,232), Typical Time (13,018), Ambiguity Resolution (3,649), Arithmetic (15,629), Relation (102,462), Temporal NLI (282,144), Causality (1,200), and Storytelling (67,214).

6. Does the benchmark contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

Part of the benchmark comprises a curated selection of instances, representing a comprehensive but not exhaustive collection of temporal reasoning problems. Specifically, it includes problems selectively sourced from existing datasets that exemplify a wide array of temporal reasoning scenarios. Human expertise has verified and determined the representativeness of the selected problems.

7. Is there a label or target associated with each instance?

Yes, the label for each instance is the correct answer to the multiple-choice question, indicated as either A, B, C, or D, and this varies by task.

8. Is the benchmark self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The benchmark is partially self-contained. Problems derived from existing datasets have been integrated into TRAM in a way that makes them standalone. This integration includes manually adding distracting or confusing options, filtering out irrelevant questions for relevance, and reformulating problems. For transparency, references are provided for problems that originated from existing data. The remaining problems are driven by human curation, supplemented by programmatic generation.

9. Does the benchmark contain data that might be considered sensitive in any way?

The benchmark does not contain any sensitive data.

Data Quality

10. Is there any missing information in the benchmark?

Everything is included. No data is missing.

11. What errors, sources of noise, or redundancies are important for benchmark users to be aware of?

Firstly, some problems in the benchmark might contain contextual ambiguities leading to multiple plausible interpretations. The benchmark is designed to have one correct answer per question, with the final unique correct answer determined or verified by a group of professionals. Secondly, within the same main task, there may be similar problems with nuanced differences. While complete redundancy of problems across the entire benchmark is avoided, the presence of similar problems is not. Finally, for problems sourced from existing datasets, irrelevant or diverging options may occur during reformulation due to issues with the source data. Further verification checks will be conducted to minimize any errors or noise that may arise in the benchmark.

12. How was the data validated/verified?

The benchmark was initially verified by multiple professionals possessing advanced degrees (M.S. or Ph.D.) in cognitive science and psychology, who provided insights into the nuances of human temporal cognition, as well as in statistics, mathematics, and computer science, for their expertise in analytical rigor required by many tasks. They reviewed the problems for relevance and common errors, such as formatting inconsistencies or logical discrepancies in questions and answers. The final review of the benchmark was conducted by the authors of the TRAM paper, who checked for relevance and removed any obvious noise and redundancies.

Pre-Processing, Cleaning, and Labeling

13. What pre-processing, cleaning, and/or labeling was done on this benchmark?

In the preparation of the benchmark, several key steps were undertaken to ensure its overall quality and relevance:

1) Pre-processing: This step involved standardizing the format of problems sourced from relevant existing datasets to align with the TRAM benchmark's structure. It included unifying the formats of questions and answers, normalizing temporal expressions, and ensuring consistency in language and style. Additionally, over 100k problems in the benchmark were manually crafted, supplemented by program generation.

2) Cleaning: A thorough review was conducted to identify and correct any obvious errors in the data. This process involved resolving typos, rectifying factual inaccuracies, and eliminating ambiguous or misleading phrasing in both questions and options. However, nuanced errors such as acceptable bias in multiple interpretations of the same problem and subtle logical errors might be overlooked and could still be present in the current version of the benchmark.

3) Labeling: Each problem in the benchmark was carefully labeled with the correct answer. In the case of multiple-choice questions, plausible distractors were also manually created and added. Labels were verified for accuracy by subject matter experts to ensure that they correctly represented the intended temporal reasoning challenge.

14. Provide a link to the code used to pre-process/clean/label the data, if available.

The code for data pre-processing is available on the official GitHub page.

15. If there are any recommended data splits (e.g., training, validation, testing), please explain.

For each main task, there is a few-shot development set, with 5 questions per category (subtask), and a separate test set for evaluation.

ADDITIONAL DETAILS ON DISTRIBUTION AND MAINTENANCE

Distribution

16. Will the benchmark be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, the benchmark is publicly available on the Internet.

17. How will the benchmark be distributed (e.g., tarball on website, API, GitHub)?

The benchmark is distributed via the official GitHub page.

18. When will the benchmark be distributed?

The benchmark was first released in September 2023.

Maintenance

19. Who will be supporting/hosting/maintaining the benchmark?

The first author of the TRAM paper will be supporting and maintaining the benchmark.

20. Will the benchmark be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Updates to question sets, error corrections, and results will be shared on the official GitHub page.

21. Will older versions of the benchmark continue to be supported/hosted/maintained?

Given any updates to the benchmark, older versions will be retained for consistency.

22. If others want to extend/augment/build on/contribute to the benchmark, is there a mechanism for them to do so?

Others wishing to do so should contact the original authors of TRAM about incorporating fixes or extensions.