Rubrik's Cube: Testing a New Rubric for Evaluation Explanations on the CUBE dataset

Diana Galvan-Sosa[1]☆, Gabrielle Gaudeau[1]☆, Pride Kavumba[2], Yunmeng Li[3], Hongyi Gu[5], Zheng Yuan[6], Keisuke Sakaguchi[3,4], Paula Buttery[1]

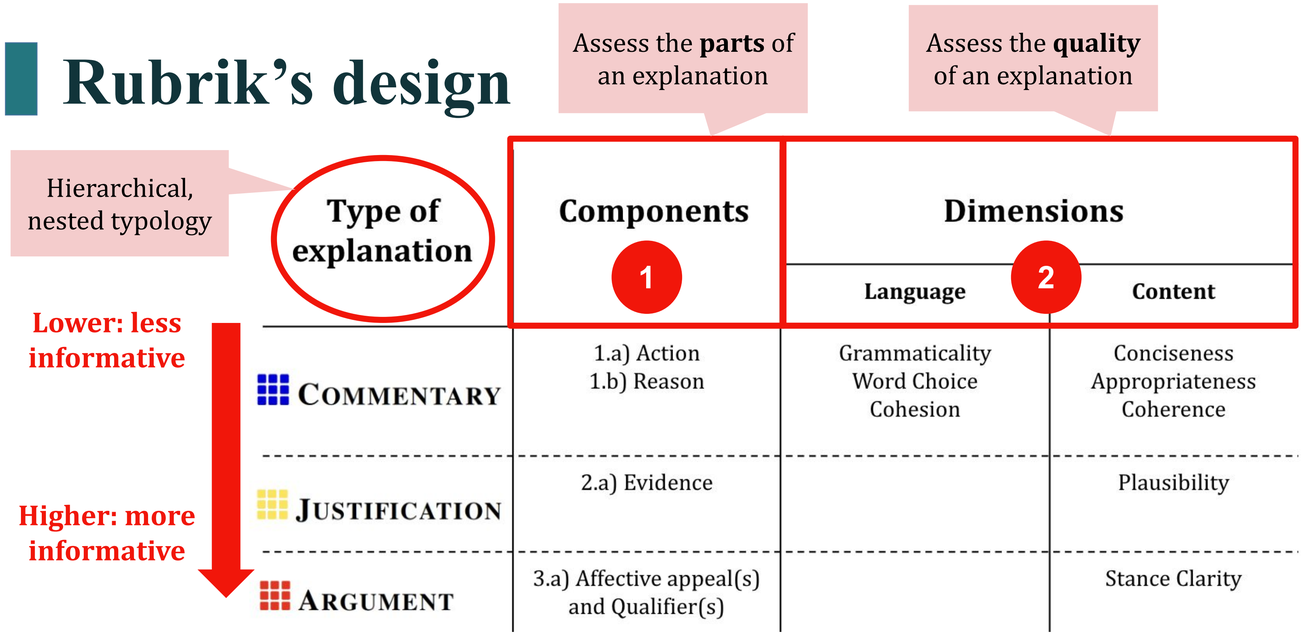
[1]ALTA Institute, Computer Laboratory, University of Cambridge | [2] SB Intuitions | [3] Tohoku University | [4] RIKEN
[5] NetMind.AI | [6] The University of Sheffield
☆Equal contribution | Contact: {dg693, gjg34}@cam.ac.uk

Overview

★ **Goal**: Allow for a more systematic evaluation of an explanation's quality

★ **Contributions**: An education-inspired rubric and a dataset of 26k explanations, written and later quality-annotated by humans and LLMs

Rubrik's design



Rubrik's validation → CUBE

Data Collection
1. **C**ommonsense reasoning (HellaSwag)
2. **U**sual fallacy detection (LOGIC)
3. **B**asic reading comprehension (RACE)
4. **E**ssay scoring (Write & Improve (BEA'19))

Explanation Generation: 6 LLMs (4 open, 2 closed) | 7 Annotators (4 contractors, 3 experts)

Explanation Assessment: 1 LLM | 2 Evaluators

Scoring strategy

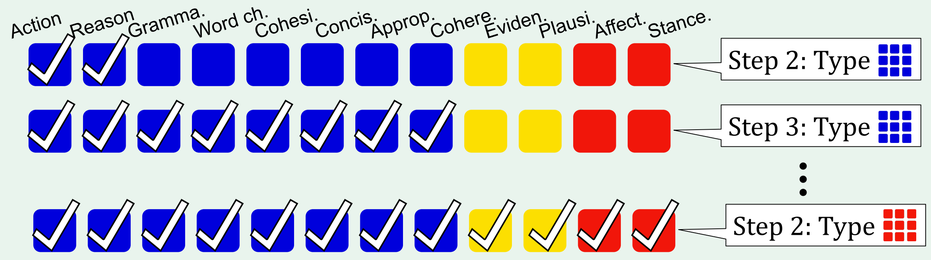
★ **Step 1:** Define the context
- *What is the task?, Who is the target audience?*

★ **Step 2:** Assess completeness, starting with COMMENTARY
- Check if all **Components** of the type are met
  - If ✔ yes → Continue to Step 3
  - If ✗ no → Stop evaluation

★ **Step 3:** Assess quality
- Check if all **Dimensions** of the type are met
  - If ✔ yes → Move to higher type and go back to Step 2
  - If ✗ no → Stop evaluation

EXAMPLE 1

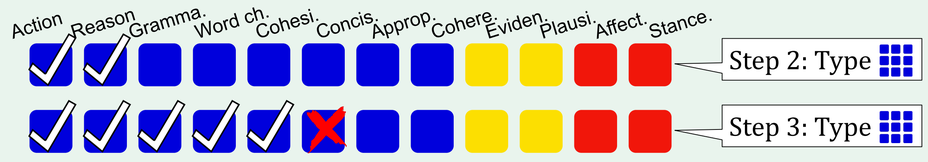
[context] essay scoring (task); academic audience
[explanation] "**The right answer is A**, because **this text is clearly of a low english level**, **with mis-conjugations of 'i do a research' and 'this are findings', alongside 'our litters' and 'whenever' instead of 'wherever' show a poor grasp of language**. The expression in the final section is very **heartfelt** however, and the tone is **excitable** and keen throughout."



☺ **Good** ARGUMENT

EXAMPLE 2

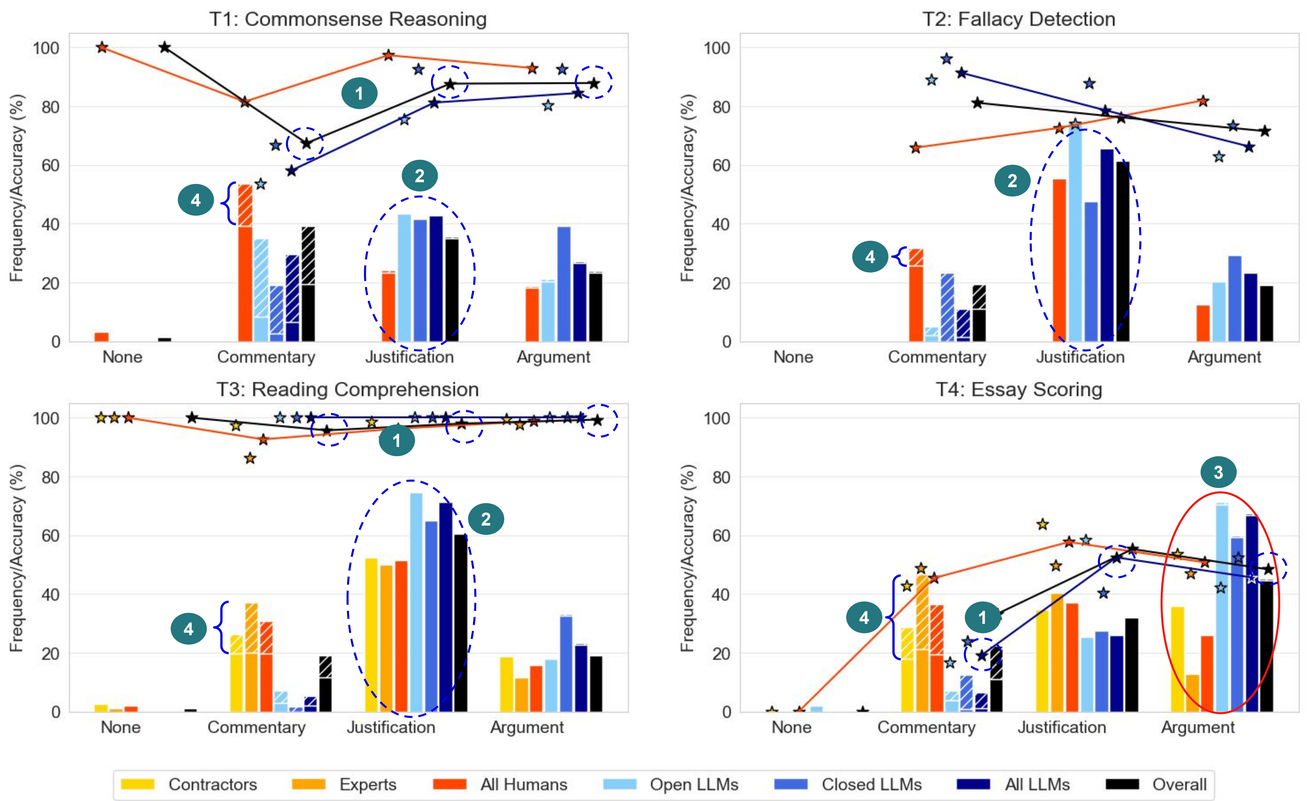
[context] commonsense reasoning (task); academic audience
[explanation] "The answer is D because the sentence mentions that **she explains how to use the lawnmower and other tools, and then she cuts the grass**. Option D accurately reflects this sequence of events."



☹ **Bad** COMMENTARY

Freq. and quality of explanation types

1 Lower accuracy is associated with the lowest type: annotators tended to generate a "Commentary" when their answers were incorrect and "Justifications" when they were correct.

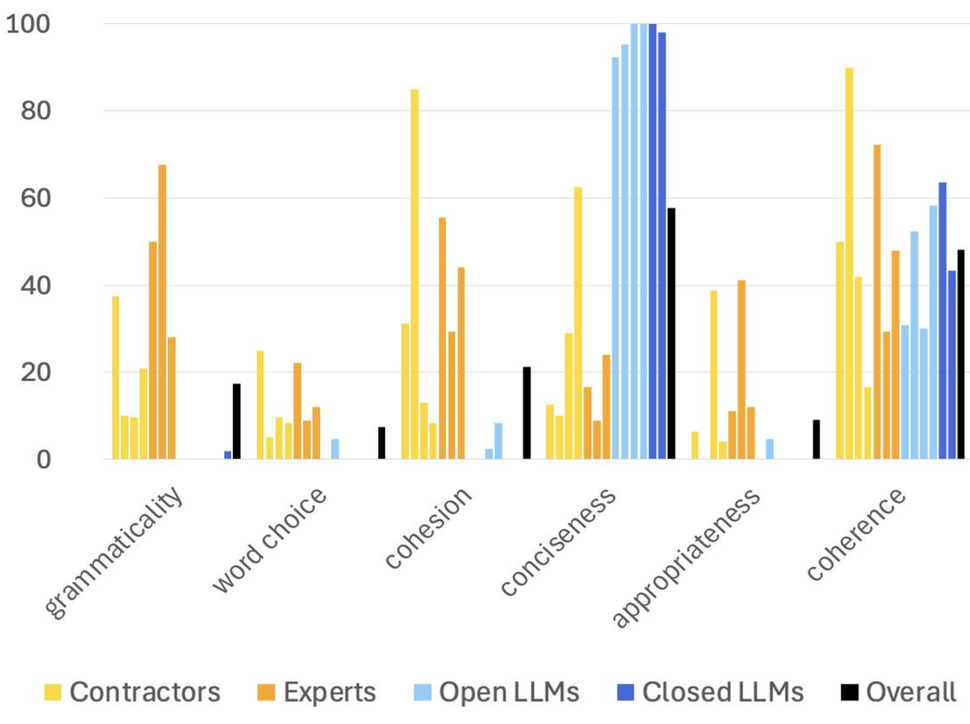


2 Both LLMs and humans tend to write "Justifications".

3 Explanation type seems to be correlated with the subjectivity of the task. T4, the hardest task, had a higher proportion of "Arguments".

Source of bad explanations

★ **LLMs**: Low quality stems primarily from a lack of conciseness.
★ **Humans**: Low quality stems primarily from a lack of coherence.
★ **Experts vs. Contractors**: Low quality stems primarily from grammaticality and coherence, respectively.



4 The number of bad explanations was low and concentrated in "Commentaries" across tasks.

The 63rd Annual Meeting of the Association for Computational Linguistics (July 29, 2025)