Rubrik's Cube:

Testing a New Rubric for Evaluating Explanations on the CUBE dataset

Diana Galvan-Sosa, Gabrielle Gaudeau, Pride Kavumba,
Yunmeng Li, Hongyi Gu, Zheng Yuan, Keisuke Sakaguchi, Paula Buttery

ACL 2025
VIENNA
JULY 27 - AUGUST 1

Diana Galvan-Sosa

Gabrielle Gaudeau

Pride Kavumba

Yunmeng Li

Hongyi Gu

Zheng Yuan

Keisuke Sakaguchi

Paula Buttery

UNIVERSITY OF CAMBRIDGE

SB Intuitions

TOHOKU UNIVERSITY

RIKEN

NetMind.AI

University of Sheffield

Introduction

Large Language Models (LLMs) are increasingly being used in tasks which require a break down of their **decision-making process** (e.g., automated scoring, question generation, problem resolution; García-Méndez et al., 2024).

Though easy to generate, LLM explanations fall short due for being unreliable (Kim et al., 2024), lacking transparency (Sallam 2023; Kabir et al., 2024).

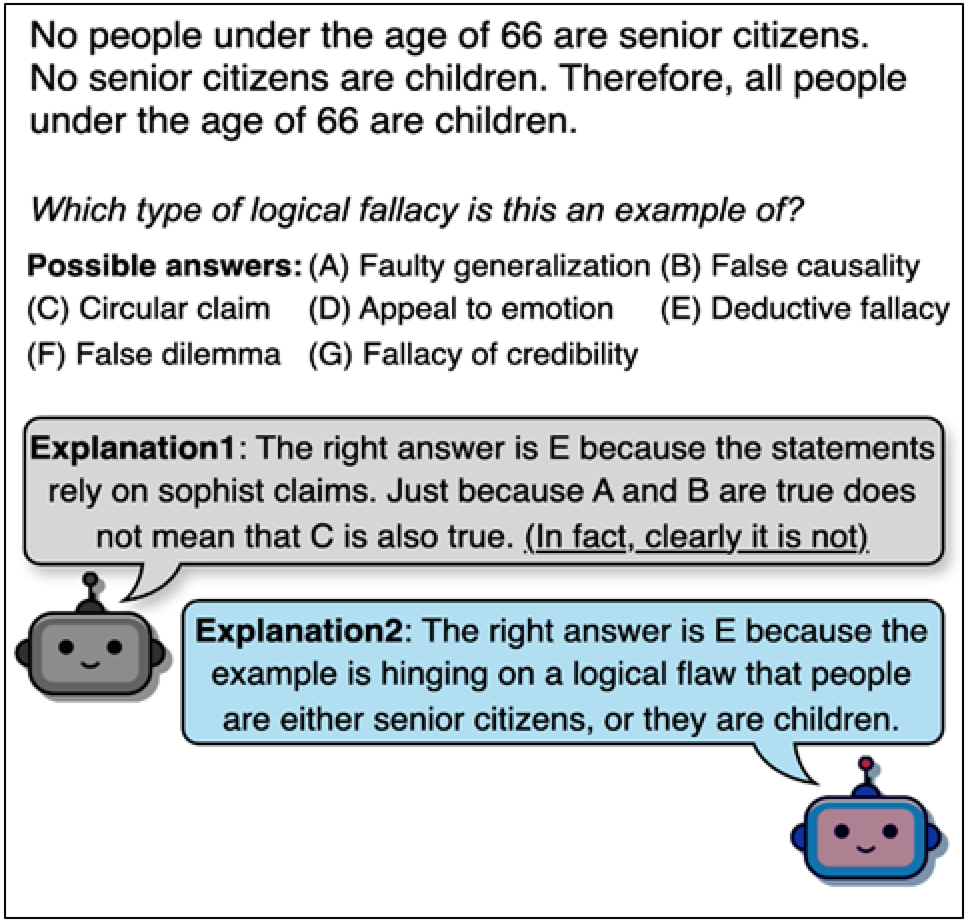
The challenge has shifted from generating explanations to assessing the quality of explanations.

Introduction

Explanations are *diverse*; there is usually more than one way of expressing the rationale behind a choice.

Introduction

Explanations are *diverse*; there is usually more than one way of expressing the rationale behind a choice.



No people under the age of 66 are senior citizens. No senior citizens are children. Therefore, all people under the age of 66 are children.

Which type of logical fallacy is this an example of?

Possible answers: (A) Faulty generalization (B) False causality (C) Circular claim (D) Appeal to emotion (E) Deductive fallacy (F) False dilemma (G) Fallacy of credibility

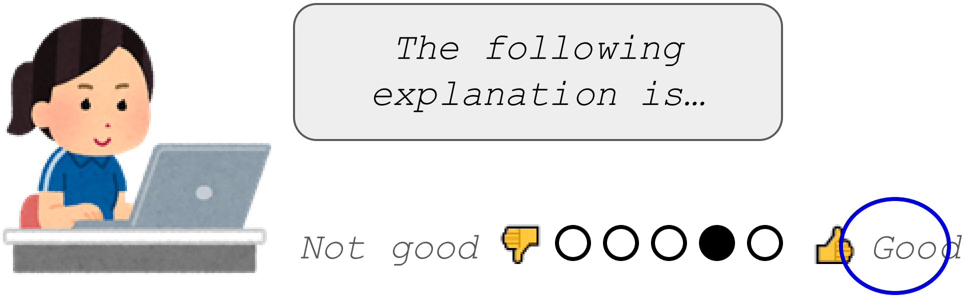
Explanation1: The right answer is E because the statements rely on sophist claims. Just because A and B are true does not mean that C is also true. (In fact, clearly it is not)

Explanation2: The right answer is E because the example is hinging on a logical flaw that people are either senior citizens, or they are children.

Introduction

Common practice is to use **human evaluators** to assess their quality.

But these **evaluations often rely on <u>intuition</u>**, rather than formal definitions (Clark et al., 2021).



The following explanation is…

Not good 👎 ○ ○ ○ ● ○ 👍 *Good*

What makes it "good"?

This form of evaluation is not systematic and prone to inconsistency.

Our proposal

Rubrik: a rubric for **explanation quality assessment**.

- Provides **clear**, **consistent**, and **objective criteria** for evaluation, following the rubric design principles set out by Dawson (2017).
- Accounts for the **diverse nature** of explanations whilst also identifying **common characteristics*** among them.
- Can easily be adapted to any task.

*Identified in different bodies of literature: cognitive and social sciences, XAI and NLG.

Methodology: Overview

Step 1: Rubric design → **`Rubrik`**

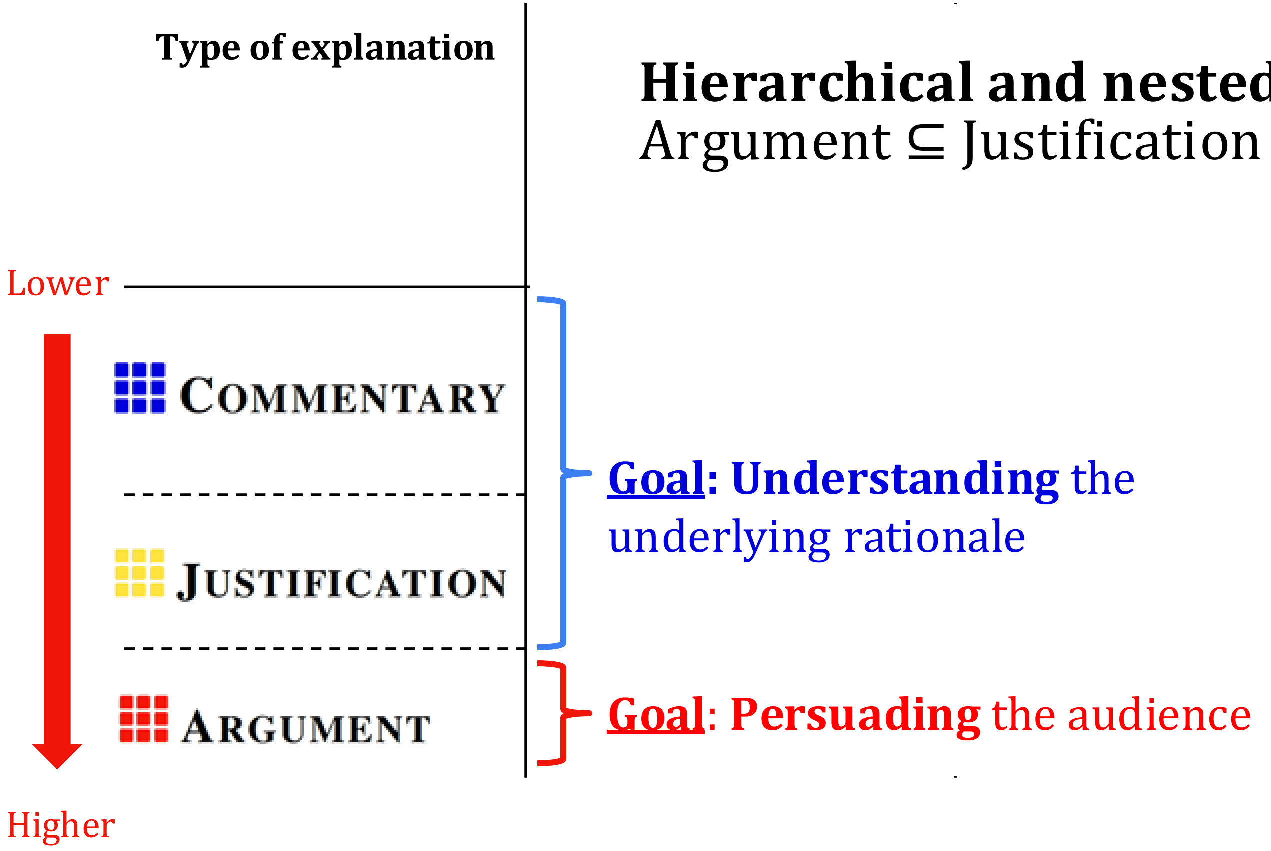
Step 2: Rubric validation → **`CUBE dataset`**

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
Commentary	1.a) Action 1.b) Reason	Grammaticality Word Choice Cohesion	Conciseness Appropriateness Coherence
Justification	2.a) Evidence		Plausibility
Argument	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

Rubrik

Methodology: Step 1



Hierarchical and nested:
Argument ⊆ Justification ⊆ Commentary

Goal: **Understanding** the underlying rationale

Goal: **Persuading** the audience

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness* 1	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality* 2	
		Language	**Content**
COMMENTARY	1.a) Action 1.b) Reason	Grammaticality Word Choice Cohesion	Conciseness Appropriateness Coherence
JUSTIFICATION	2.a) Evidence		Plausibility
ARGUMENT	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
COMMENTARY	1.a) Action ☑ 1.b) Reason ☑	Grammaticality Word Choice Cohesion	Conciseness Appropriateness Coherence
JUSTIFICATION	2.a) Evidence		Plausibility
ARGUMENT	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

A commentary

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
COMMENTARY	1.a) Action ☑ 1.b) Reason ☑	Grammaticality ☑ Word Choice ☑ Cohesion ☑	Conciseness ☑ Appropriateness ☑ Coherence ☑
JUSTIFICATION	2.a) Evidence		Plausibility
ARGUMENT	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

Good commentary

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
COMMENTARY	1.a) Action ✓ 1.b) Reason ✓	Grammaticality ✓ Word Choice ✓ Cohesion ✓	Conciseness ✓ Appropriateness ✓ Coherence ✓
JUSTIFICATION	2.a) Evidence ✓		Plausibility ✓
ARGUMENT	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

A justification

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
Commentary	1.a) Action ☑ 1.b) Reason ☑	Grammaticality ☑ Word Choice ☑ Cohesion ☑	Conciseness ☑ Appropriateness ☑ Coherence ☑
Justification	2.a) Evidence ☑		Plausibility ☑
Argument	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

Good justification

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
COMMENTARY	1.a) Action ☑ 1.b) Reason ☑	Grammaticality ☑ Word Choice ☑ Cohesion ☑	Conciseness ☑ Appropriateness ☑ Coherence ☑
JUSTIFICATION	2.a) Evidence ☑		Plausibility ☑
ARGUMENT	3.a) Affective appeal(s) and Qualifier(s) ☑		Stance Clarity ☑

An argument

Methodology: Step 1

Type of explanation	**Components** necessary parts of an explanation that contribute to its *completeness*	**Dimensions** necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	**Content**
Commentary	1.a) Action ☑ 1.b) Reason ☑	Grammaticality ☑ Word Choice ☑ Cohesion ☑	Conciseness ☑ Appropriateness ☑ Coherence ☑
Justification	2.a) Evidence ☑		Plausibility ☑
Argument	3.a) Affective appeal(s) and Qualifier(s) ☑		Stance Clarity ☑

Good argument

Methodology: Step 2

Step 1: Rubric design → **`Rubrik`**

- Explanation types (hierarchical and nested)
- Quality dimensions

Step 2: Rubric validation → **`CUBE dataset`**

Methodology: Step 2

1. Data collection **4 Tasks**

- **C**ommonsense Reasoning
- **U**sual Fallacy Detection
- **B**asic Reading Comprehension
- **E**ssay scoring

Methodology: Step 2

1. Data collection

4 Tasks

- Commonsense Reasoning
- Usual Fallacy Detection
- Basic Reading Comprehension
- Essay scoring

2. Explanation Generation

6 LLMs
(4 open, 2 closed)

7 annotators
(4 contractors, 3 experts)

Methodology: Step 2

1. Data collection

4 Tasks

- Commonsense Reasoning
- Usual Fallacy Detection
- Basic Reading Comprehension
- Essay scoring

2. Explanation Generation

6 LLMs
(4 open, 2 closed)

7 annotators
(4 contractors, 3 experts)

1 ,000 explanations/LLM/task **= 24,000**
110 explanations/contractor/task + 110 explanations/expert/tasks 3&4 **= 2,420**
Total = 26,420 explanations

Methodology: Step 2

1. Data collection

4 Tasks

- Commonsense Reasoning
- Usual Fallacy Detection
- Basic Reading Comprehension
- Essay scoring

2. Explanation Generation

6 LLMs
(4 open, 2 closed)

7 annotators
(4 contractors, 3 experts)

1 ,000 explanations/LLM/task **= 24,000**
110 explanations/contractor/task + 110 explanations/expert/tasks 3&4 **= 2,420**
Total = 26,420 explanations

3. Explanation assessment

1 LLM

2 evaluators

Explanation assessment: Essay Scoring

Essay:

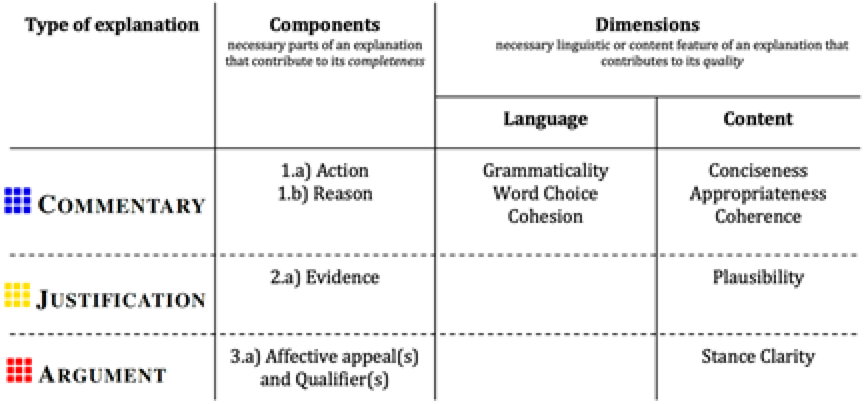
To: International organisation
From: Dimitris Barberis
Subject: Our green town

Introduction
The aim of the report is to write how are town take care of the environment. I do a research and this are findings.

Rubbish
We have a lot of bins around the area, so now we can throw our litters whenever we are. Also we have recycle bins for paper and glass.

Cleaners
Every Saturday our local cleaning team clean the park and now everyone can enjoy it!

Conclusion
We do everything to make our town more green, our citizens always have new ideas that make the difference of our daily life.



Type of explanation	Components necessary parts of an explanation that contribute to its *completeness*	Dimensions necessary linguistic or content feature of an explanation that contributes to its *quality*	
		Language	Content
🟦 COMMENTARY	1.a) Action 1.b) Reason	Grammaticality Word Choice Cohesion	Conciseness Appropriateness Coherence
🟨 JUSTIFICATION	2.a) Evidence		Plausibility
🟥 ARGUMENT	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

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.

Explanation assessment: Essay Scoring

Essay:

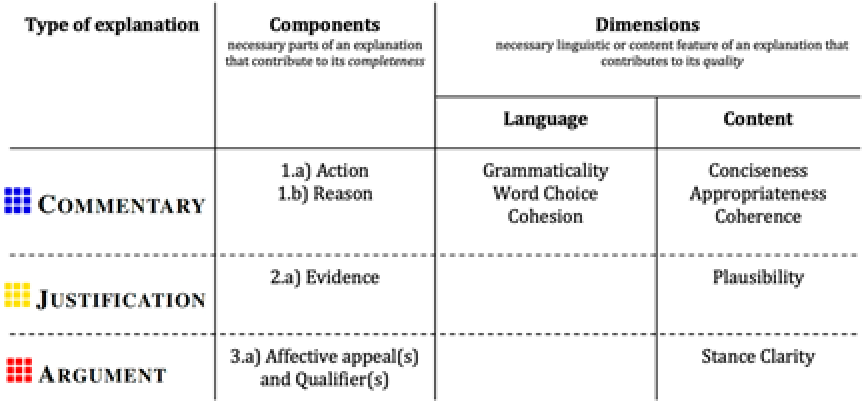
To: International organisation
From: Dimitris Barberis
Subject: Our green town

Introduction
The aim of the report is to write how are town take care of the environment. I do a research and this are findings.

Rubbish
We have a lot of bins around the area, so now we can throw our litters whenever we are. Also we have recycle bins for paper and glass.

Cleaners
Every Saturday our local cleaning team clean the park and now everyone can enjoy it!

Conclusion
We do everything to make our town more green, our citizens always have new ideas that make the difference of our daily life.



Type of explanation	Components necessary parts of an explanation that contribute to its completeness	Dimensions necessary linguistic or content feature of an explanation that contributes to its quality	
		Language	Content
COMMENTARY	1.a) Action 1.b) Reason	Grammaticality Word Choice Cohesion	Conciseness Appropriateness Coherence
JUSTIFICATION	2.a) Evidence		Plausibility
ARGUMENT	3.a) Affective appeal(s) and Qualifier(s)		Stance Clarity

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!

Methodology: Step 2

1. Data collection

4 Tasks

- **C**ommonsense Reasoning (HellaSWAG)
- **U**sual Fallacy Detection (LOGIC)
- **B**asic Reading Comprehension (RACE)
- **E**ssay scoring (Write & Improve, BEA'19)

2. Explanation Generation

+

6 LLMs
(4 open, 2 closed)

7 annotators
(4 contractors, 3 experts)

1 ,000 explanations/LLM/task **= 24,000**
110 explanations/contractor/task + 110 explanations/expert/tasks 3&4 **= 2,420**
Total = 26,420 explanations

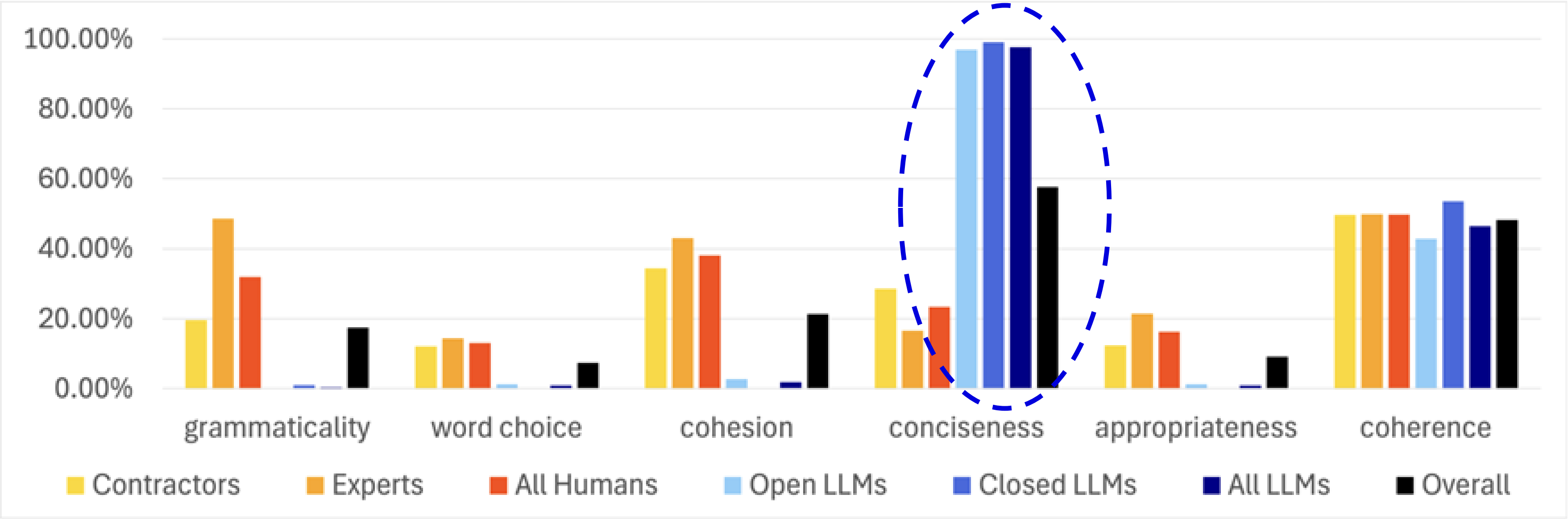
3. Explanation assessment

1 LLM

2 evaluators

920 explanations jointly assessed by the LLM and human evaluators.
4,140 explanations assessed by the LLM only.
Total = 5,060 explanations assessments.

Results: Source of *bad* commentaries



Low-quality LLM explanations are due to lack of **conciseness**.

Further Results

- The types of explanations varied depending on task difficulty
  (e.g., more arguments in essay scoring).
- Task accuracy and our typology correlate
  (e.g. justifications coincided with higher task accuracy as opposed to commentaries).
- LLMs and humans tend to output justifications
  (i.e. providing evidence).

Our results demonstrate the *usefulness* of our rubric.

Conclusion

To address the **lack of widely-agreed definition** of what constitutes a *good* explanation, we propose:

- **Rubrik**, a **general-purpose rubric** for evaluating the quality of LLM-generated and human-written explanations.
- **CUBE**, a **dataset** of 26k explanations written by both humans and LLMs across four tasks (Commonsense Reasoning, Fallacy Detection, Reading Comprehension, Essay Scoring), to **validate the rubric.**

We hope to advance explanation quality assessment in the future.

Thank you!

Contact us: dg693@cam.ac.uk, gjg34@cam.ac.uk