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

Anonymous ACL submission

Abstract

The performance and usability of Large-Language Models (LLMs) are driving their use in explanation generation tasks. However, despite their widespread adoption, LLM explanations have been found to be unreliable, making it difficult for users to distinguish good from bad explanations. To address this issue, we present Rubrik's CUBE–an education-inspired rubric and a dataset of 26k explanations, written and later quality-annotated using the rubric by both humans and six open- and closed-source LLMs. The CUBE dataset focuses on two reasoning and two language tasks, providing the necessary diversity for us to effectively test our proposed rubric. Using Rubrik, we find that explanations are influenced by both task and perceived difficulty. Low quality stems primarily from a lack of conciseness in LLM-generated explanations, rather than cohesion and word choice. The full dataset, rubric, and code will be made available upon acceptance.

1 Introduction

Explanations play a crucial role in the process of understanding why a decision was made. But, as illustrated in Figure 1, there exist many ways of expressing the rationale behind a choice. Large-Language Models (LLMs), with their inherent capacity for generating very different outputs given the same query, provide a compelling example of this phenomenon. In fact, these models are increasingly being used in applications which expect a detailed breakdown explaining why a decision was made (e.g., automated scoring, question generation, problem resolution; García-Méndez et al., 2024).

Unfortunately, LLM-generated explanations generally fall short of user expectations due to their unreliability (Kim et al., 2024). Indeed, they are known to hallucinate, producing incorrect or misleading information, and often struggle to back up their responses to queries, highlighting an overall deficiency in their reasoning capabilities (Huang and Chang, 2023; Saxena et al., 2024). As noted by Zhang et al. (2023a), these issues remain unaddressed, even by prompting strategies like *"Let's think step by step."* As a result, LLM-generated explanations lack transparency, and be a source of misinformation and limited knowledge (Sallam, 2023; Kabir et al., 2024a). Consequently, the challenge has shifted from generating text to assessing its quality, a difficulty that has led some sites to temporarily ban the use of any generative AI (GenAI)[1].

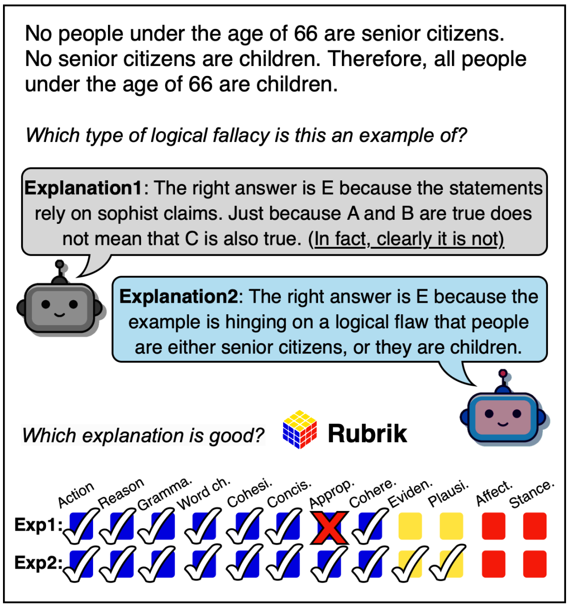


Figure 1: Different ways of articulating the logic underlying an answer choice, with quality variations based on *Appropriateness* and the provision of *Plausible* evidence.

The most common practice in GenAI to determine the quality of a text is to rely on human evaluators. However, because such evaluators typically

[1]See StackOverflow's policy on the use of ChatGPT and other LLMs: https://meta.stackoverflow.com/questions/421831/policy-generative-ai-e-g-chatgpt-is-banned

lack specific training, the exact evaluation criteria are left to their discretion (Clark et al., 2021). Inspired by the use of rubrics in education for the qualitative evaluation of complex and subjective tasks like essay writing (e.g., the IELTS writing rubric; Arnold, 2023), we design our very own rubric following Dawson (2017)'s best practices. In doing so, we align ourselves with the human-grounded evaluation proposed by Doshi-Velez and Kim (2017), which identifies and evaluates the "general notions" of the quality of an explanation without having a specific end goal.

We thus introduce Rubrik's CUBE[2], a task-independent rubric and a dataset to help evaluate the quality of LLM-generated explanations. Rubrik identifies the core components and features of a *good* explanation, differentiated by explanation type; CUBE contains 26k explanations drawn from instances of four distinct tasks, generated by both humans and a set of open- (`Command R+`, `Gemma 2`, `Llama 3.1`, `Mixtral`) and closed-source (`GPT-4o`, `Claude Sonnet 3.5`) models. We additionally include two custom agreement metrics that account for the hierarchical and nested nature of our rubric. Rubrik enabled valuable insights on output quality, allowing us to identify distinct patterns in the explanations of all annotators. We observe that the explanation type depends on the task and its perceived difficulty. Specifically, our rubric revealed that low-quality LLM explanations are primarily due to not being concise and only rarely because of word choice or cohesion.

2 Background

We summarise different bodies of literature on the nature and qualities of explanations which, alongside insights from the education assessment literature, informed the design of our proposed rubric.

2.1 Cognitive Science and Social Sciences

There is an open discussion in philosophy and other social sciences like psychology about what an explanation is and what makes the best explanation (Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Miller, 2019a). From the psychology and cognitive science perspective, an explanation is something ubiquitous, diverse, and fundamental to humans' sense of understanding. They come in a variety of forms and formats and are used for a variety of purposes (Keil, 2006), including: (1) understanding a *decision process* (2) understanding and predicting an *unexpected event*, and (3) filling a gap in knowledge (i.e., *learning*). It follows that a *good* explanation is inherently related to its purpose, which some suggest is shaped by what is being asked (Bromberger, 1992). In particular, authors like Lombrozo (2006) and Miller (2019a) argue that an explanation's relation to cognition comes from an attempt to answer a *why-question*. Miller investigated the criteria that people use to evaluate explanations, finding that the most important are: PROBABILITY, SIMPLICITY, GENERALIZE, and COHERENCE with prior beliefs. The truth of LIKELIHOOD is also identified as an important criterion. However, Miller notes that an explanation that includes this attribute is not always the *best* explanation.

2.2 Explainable AI

In the context of Explainable AI (XAI) and Machine Learning (ML) interpretability, an explanation should be able to reflect the internal decision process of a system. *Introspective* systems output this kind of explanation, while *justification* systems output evidence supporting a decision (Park et al., 2018). The most studied properties of explanation systems include FIDELITY, STABILITY, COMPREHENSIBILITY, GENERALIZABILITY and CONSISTENCY (Fel et al., 2022). According to Wiegreffe and Marasović (2021), explanations are implicitly or explicitly designed to answer the why-question *"why is <input> assigned <label>"*. They identified HIGHLIGHTS (subsets of the input elements that explain a prediction) as one type of explanation in the Explainable NLP (EXNLP) literature, where COMPACTNESS, SUFFICIENCY and COMPREHENSIVENESS are the main attributes.

2.3 Natural Language Generation

In an attempt to find a consensus about how human evaluations of generated text should be designed and reported, Howcroft et al. (2020) examined twenty years of NLG papers that reported some form of human evaluation. Some of the most common criteria used to assess quality include FLUENCY, APPROPRIATENESS and CLARITY.

The Multidimensional Quality Metrics (MQM) framework (Burchardt, 2013; Mariana, 2014; Freitag et al., 2021) has been widely applied to machine translation studies in recent years. Compared to scoring systems, MQM provides a more detailed

[2]Short for **C**ommonsense reasoning, **U**sual logical fallacies, **B**asic reading comprehension, and **E**ssay scoring.

Typology of Explanations		COMPONENTS necessary parts of an explanation	DIMENSIONS necessary qualities of a *good* explanation	
			Language	**Content**
Typ1.	COMMENTARY	1.a) Action 1.b) Reason	Grammaticality Word Choice Cohesion	Conciseness Appropriateness Coherence
Typ2.	JUSTIFICATION	2.a) Evidence		Plausibility
Typ3.	ARGUMENT	3.a) Affective appeals(s) and Qualifiers(s)		Stance Clarity

Table 1: Overview of our evaluation rubric which identifies three hierarchical types of explanations, their necessary parts (COMPONENTS), and the features that distinguish the *good* from the *bad* ones (DIMENSIONS).

and flexible approach to evaluate translation tasks. MQM includes seven major categories focusing on terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and even design and markup. The major categories are further divided into several subcategories to assess the quality specifically. In the design of our metric, we also adopted the approach of categorising the significant features in evaluating explanations into three major labels and subdividing them into sub-labels.

2.4 Education

Education, and specifically science education, has long focused on teaching students how to construct explanations, and assessing them (e.g., Sandoval, 2003; Mcneill et al., 2006; McNeill and Krajcik, 2008; Zangori et al., 2013). For them, explanations "make sense of a phenomenon based on other scientific facts" (Ohlsson, 2002). They should begin with a statement of the *explanandum* (i.e., the phenomenon to be explained). Then, what makes a *good* explanation differs is "explanatory adequacy" (Brigandt, 2016) which consists in providing an understanding of how or why a phenomenon occurs (Chin and Brown, 2000).

In practice, assessing explanations is difficult (Berland and Mcneill, 2012), so teachers generally rely on rubrics, like the one proposed by McNeill and Krajcik (2007), which provide clear, consistent, and objective sets of criteria for evaluation. More generally, rubrics are firmly established evaluation tools in written assessment and widely advocated in books by Walvoord and Anderson (1998); Huba and Freed (2000); Dunn et al. (2003); Stevens and Levi (2004); Freeman et al. (2016). Unfortunately, these practices are not currently being used beyond education, and no equivalent rubric exists for evaluating LLM explanations on a variety of tasks (beyond scientific explaining). To address this gap, we propose to draw on this literature to come up with our very own rubric.

3 A Systematic Quality Assessment Framework

This section introduces our proposed assessment framework in three parts. First, we detail the design decisions taken to develop the rubric, drawing upon the key principles outlined by Dawson (2017). Second, we provide a comprehensive overview of the rubric itself, outlining its key elements and their hierarchical relationships. Finally, we provide practical guidance on how to effectively use the rubric for an explanation's quality assessment.

3.1 Designing an Assessment Rubric

Recognising that the foundation of an effective evaluation lies in its instrument, we carefully considered the design elements suggested by Dawson (2017). A key advantage of adhering to their framework is the streamlined design process and the enhanced transparency of the resulting rubric, facilitating easier comparisons with other instruments. Table 1 presents an overview of our proposed rubric and Table 2 shows the design considerations and choices we made in developing it.

3.2 A Task-Agnostic Quality Rubric

The *context* of the explanations should be defined at the beginning of the evaluation:

- What is the task? In our case we will be looking at reasoning and language tasks (Section 4.1), but this could be anything.
- Who is the target audience?
- What is the purpose of the explanations?

Design element	Decision
Specificity: the particular object of assessment	Assess the quality of explanations.
Secrecy: who the rubric is shared with, and when it is shared	It should be secret to the annotators. It is only shared with the evaluators.
Exemplars: work samples provided to illustrate quality	Examples of acceptable and not acceptable instances.
Scoring strategy: procedures used to arrive at marks and grades	A series of binary judgments (yes/no) all amounting to a binary decision (good/bad).
Evaluative criteria: overall attributes required of the explanation	Components and dimensions.
Quality levels: the number and type of levels of quality	Two levels (good/bad).
Quality definition: explanations of attributes of different levels of quality	Motivated by different bodies of literature (social sciences, XAI, and NLG).
Judgment complexity: the evaluative expertise required of users of the rubric	Should be simple enough for **anyone** to use.
Users and uses: who makes use of the rubric, and to what end	Evaluators use for summative assessment.
Creators: the designers of the rubric	NLP researchers.

Table 2: Summary of the design decisions taken to develop our proposed rubric. The design elements are those suggested by Dawson (2017). Note that *annotators* refers to those generating an explanation (i.e., a human or an LLM).

3.2.1 Components

A fundamental assumption underlying this work is that it is possible to account for the diverse nature of explanations while identifying common features that characterize them. This assumption is grounded in the insights presented in Section 2, where different bodies of literature identify shared attributes of a *good* explanation. We propose that different types of explanations are defined not only by their goals, but also by their structure. This led to the hierarchical type classification based on COMPONENTS detailed in Table 1. Formally, COMMENTARY $\subseteq$ JUSTIFICATION $\subseteq$ ARGUMENT. As the foundational type, a COMMENTARY embodies the most basic type of explanation, with its primary objective being to provide an understanding of a decision-making process. Throughout this work, we assume a situation where there is a (explicit or implicit) set of choices and one is selected over the others. Then, a decision is the behavioural ACTION of choosing among alternative options (Brust-Renck et al., 2021) and it is complemented by the REASON that guided that choice. If there is EVIDENCE to support the decision, a COMMENTARY would then transition to a JUSTIFICATION. Note that in either case, the underlying principle of objectivity remains consistent across both types. A subjective approach to presenting a decision process shifts the main goal of *understanding* the underlying rationale to *persuading* the audience. This idea aligns with the definition of an ARGUMENT, which is the result of an activity aimed at convincing a reasonable critic of the acceptability of a standpoint (Lunsford et al., 2008).

When considering the nature of argumentation, it is common to refer to the seminal work of Toulmin (1958), who provided a framework for constructing, analysing, and evaluating arguments. However, we adopt a different perspective, drawing upon the principles of rhetoric. Although there are some similarities between WARRANT$\leftrightarrow$REASON and BACKING $\leftrightarrow$ EVIDENCE, this does not hold for the relationship between CLAIM $\leftrightarrow$ ACTION. In Toulmin's framework, a warrant supports the claim and the backing further supports the warrant, but a claim is always assumed to be linked to a *standpoint*. Rhetorical argumentation, on the other hand, commonly refers to Aristotle's trio *ethos-logos-pathos* (Braet, 1992), where *ethos* refers to the credibility of the speaker, *pathos* refers to the emotional state of the audience and *logos* refers to what is true. We can identify a relationship between LOGOS $\leftrightarrow$ COMMENTARY through the REASON component and ETHOS $\leftrightarrow$ JUSTIFICATION through EVIDENCE. It is then left to PATHOS to introduce the elements of *persuasion*. Considering that a stance is usually implicit in discourse, we focus on linguistic markers; metadiscourse features used by writers to express stance (Barbara et al., 2024). Thus, we merge into one component the essence of *pathos*, usually expressed in discourse through AFFECTIVE APPEAL(S), and features from Hyland's Interpersonal Model of Metadiscourse (Amiryousefi and Barati, 2011): hedges, boosters, attitude and engagement markers (i.e., QUALIFIERS).

3.2.2 Dimensions

While the COMPONENTS provide the necessary structural elements of different types of explanations, we also need to evaluate their quality through what we call DIMENSIONS. This distinction ensures that our rubric accounts for both what is being said (through the COMPONENTS) and how well it is communicated (through the DIMENSIONS).

To identify these DIMENSIONS, we surveyed the bodies of literature introduced in Section 2 and recorded explanation qualities that have been studied, annotated or evaluated in each. We include an exhaustive list of these qualities in Table 5. Our first step was to define each, drawing from a variety of literature, and filtered out qualities that were too task-specific since our intent was to create a general-purpose rubric. For instance, we discarded Fidelity, Consistency, Transparency and Interpretability which tend to focus on the internal workings of AI models. We also classified our chosenDIMENSIONS into one of two categories: **Language** and **Content**. The first category assesses whether the explanation is well-formed; the second evaluates the ideas expressed by the explanation. This design choice was motivated by the fact that LLMs sometimes produce text that is only *good* on the surface but factually incorrect, inappropriate, or misleading (Huang et al., 2025). We describe our process in more detail in Appendix A.

We also related the dimensions to the COMPONENTS and explanation types introduced in the previous section. Indeed, having an Action and Reason is a requirement for a COMMENTARY to be considered complete; but for it to be viewed as *good*, we must enforce certain linguistic requirements: it needs to be grammatical, cohesive, and use *context*-appropriate language. On the other hand, its content should be coherent and concise and match the expectations imposed by the defined *context*. Further, a JUSTIFICATION is contingent on the presence of Evidence. Ensuring it is plausible and consistent with human reasoning is a further requirement for a *good* JUSTIFICATION. Finally, the presence of argumentative markers generally betrays the explainer's intent to persuade the audience of their *stance* (i.e., their personal feelings towards the task). Whether this stance is clearly and unambiguously conveyed distinguishes a *good* from a *bad* ARGUMENT. These correspond to the eight DIMENSIONS of our rubric (see Table 1).

3.3 Scoring Strategy

The rubric employs a binary scoring approach, where each COMPONENT and DIMENSION is assessed as either ✓ met or ✗ not met. Overall, the main rule to evaluate the quality of an explanation is to first judge the presence of the necessary components (i.e., identify the explanation's type) and then, check whether each of the dimensions is met. For example, an explanation can only be considered a COMMENTARY if it has both ACTION and REASON components. Note that this is the only case in which an evaluator should proceed to evaluate each of the dimensions, since they are meant to assess the components. Otherwise, it is not an explanation (or NONE). A *good* COMMENTARY would be one that meets <u>all</u> the DIMENSION criteria. If an explanation fails to meet one of the COMMENTARY's dimensions, it will be judged as a *bad* COMMENTARY. The same principle applies to judge an explanation as a *good/bad* JUSTIFICATION or a *good/bad* ARGUMENT. The hierarchical nature of our proposed classification of explanations implies that a *good* COMMENTARY is the base of a *good* JUSTIFICATION, which in turn is the base of a *good* ARGUMENT.

4 Rubric Validation

The main motivation behind our proposed rubric is to allow for a more systematic evaluation of an explanation's quality. In order to determine the effectiveness of our proposal, we designed a validation process aimed at addressing the following question: *Does the rubric effectively discriminate between high-quality and low-quality explanations, while simultaneously providing clear and concise guidance for evaluators?* Given the absence of existing datasets for explanation assessment, the validation of this rubric required a tailored approach. This began with identifying an appropriate source of data, followed by gathering explanations, evaluating them using the rubric with three raters, and finally, measuring the inter-rater reliability to determine the consistency of the rubric's application. The effectiveness of our rubric was evaluated by measuring the level of inter-rater agreement for each explanation.

4.1 Data Collection

As previously mentioned in Section 3.2.1, we assume a decision-making scenario involving a set of choices, where one is selected. Thus, our data collection process required instances from tasks that could be framed as a series of multiple-choice questions (MCQ) with a single correct answer. To ensure a diverse set of explanations, we chose four

		Single annotations			Joint annotations						Single evaluations			Joint evaluations				
		Inst.	LLM	Total	Inst.	H	LLM	Total	Total	Total	Inst.	E	LLM	Inst.	E	H	LLM	
T1	1000	890	6	5340	110‡	4	6	10	1100	6440	90‡	900	1	20‡	200	2	1	
T2	1000	890	6	5340	110‡	4	6	10	1100	6440	90‡	900	1	20‡	200	2	1	
T3	1000	890	6	5340	110‡	7	6	13	1430	6770	90‡	1170	1	20‡	200	2	1	
T4	1000	890	6	5340	110‡	7	6	13	1430	6770	90‡	1170	1	20‡	200	2	1	
Total	4000	3560		21360	440				5060	**26420**	360	4140		80	920			5060

Table 3: Instances and explanations (E) in CUBE. Double-underlined numbers represent the initial pool, divided into subsets (single-underlined) based on the annotators assigned. A (‡) denotes variations in evaluator assignment.

different tasks, drawn from reasoning and language assessment. The reasoning tasks are: (**T1**) commonsense reasoning and (**T2**) fallacy detection. The language tasks are: (**T3**) reading comprehension and (**T4**) essay scoring. From an initial pool of 1000 instances from each task, we curated an *annotation dataset* of 440 total instances for annotation (110 from each dataset). A brief description of the datasets follows. Detailed selection criteria are described in Appendix B.

Reasoning tasks. For T1 and T2, we selected instances from the HellaSwag (Zellers et al., 2019) and Logic (Jin et al., 2022) datasets, respectively. Each instance in HellaSwag has a **context** and a set of four ENDINGS; the task is to select the most likely follow-up sentence. Logic consists of common logical fallacy examples collected from various online educational materials.

Language tasks. For T3 and T4, we selected instances from RACE (Lai et al., 2017) and the Write&Improve (W&I) (Bryant et al., 2019) corpus, respectively. RACE consists of a series of passages and questions taken from English exams that evaluate a student's ability in understanding and reasoning. Write&Improve[3] is an online web platform that assists English Language Learners with their writing (Yannakoudakis et al., 2018). The dataset contains submissions (defined as "essays") that have manually annotated with a coarse CEFR[4] level (A, B or C) by trained raters.

4.1.1 Annotation

Two key decisions shaped the annotation process. First, we retained all annotations, regardless of the correctness of the chosen answer. This decision was driven by the need to explore the explanations associated with correct and incorrect answers, allowing for a more nuanced understanding of the explanatory quality. Second, human explanations were not treated as the gold standard. This allowed for a more objective comparison of human and LLM explanations, avoiding potential bias towards human responses. Below, we overview the annotation process, with further details provided in Appendix C.

Human. We recruited seven annotators, consisting of four general and three professional annotators with experience in language assessment. While general annotators (contractors) covered all instances, professional annotators (experts) annotated only those from language tasks. Annotators were asked to answer each multiple-choice question and explain their selected choice. This resulted in 880 explanations from T1 and T2 and 1,540 from T3 and T4.

LLM-based. We worked with six LLMs, including four open-source: `Llama 3` (Dubey et al., 2024), `Gemma 2` (Team et al., 2024), `Mixtral` (Jiang et al., 2024) `Command R+`, (Cohere for AI, 2024) and two closed-source models—GPT-4o `GPT-4o` (OpenAI, 2024) and `Claude 3.5 Sonnet` (Anthropic, 2024). Models were prompted using a few-shot setting (see Appendix C.2.1). LLM explanations were generated for all instances, yielding a total of 24,000 explanations. Table 3 shows a more detailed breakdown of annotations and evaluations.

4.1.2 Evaluation

Data evaluation was performed by two expert evaluators and the six LLMs on a subset of 80 instances (20 of each task; 920 explanations) of the annotation set. The 4,140 explanations from the remaining 360 instances were evaluated by a single LLM. We sought a third evaluator to enhance the robustness of our analysis. Using two custom agreement metrics, we identified the LLM whose performance on the human-annotated explanations most closely matched human judgment. This selection was based on a preliminary experiment with the

[3] https://writeandimprove.com/.

[4] Common European Framework of Reference for Languages (North and Piccardo, 2020) levels correspond to language proficiency levels ranging from A1 (elementary) to C2 (complete proficiency) from a second-language learner's perspective.

Type	Example
COMMENTARY	The right answer is D because the reason given to encourage Luke to eat is subjective and has nothing to do with his taste in food or any potential benefits. It is not a valid reason.
COMMENTARY	The right answer is D because this statement is trying to make Luke **eat the sheep's brains with chopped liver and brussel sprouts** by making him feel guilty **about the poor, starving children in a third world country**. It's an appeal to his emotions, rather than presenting a logical argument.
JUSTIFICATION	The right answer is A because the woman in the video is demonstrating how to make or destroy lipsticks, which is a process that involves using one's mouth and lips. This is the only option that describes an action that would require the use of the mouth and lips.
JUSTIFICATION	The right answer is D because it is the most reasonable answer of the options. A is incorrect because **demonstrating how to make and destroy lipsticks is an unusual subject to demonstrate: the making of lipsticks in particular is complicated and challenging.** B is incorrect because one's tongue does not fall out when speaking. C is incorrect because you do not describe ordinarily describe women as walking into a classroom with "both [their] boobs" as this implies their boobs are separable. D is correct because sucking from a hookah is a fairly ordinary activity.
ARGUMENT	The right answer is B because the essay is written in a somewhat coherent and understandable manner, but it lacks clarity, coherence, and proper sentence structure. The writer's emotions and thoughts are expressed, but the writing is not sophisticated or polished. The essay does not demonstrate a clear understanding of the topic or the ability to express complex ideas. The writer's use of language is simple, and the essay lacks depth and analysis. Therefore, it is best graded as Intermediate (grade B).
ARGUMENT	The right answer is B because the essay demonstrates a basic understanding of grammar and sentence structure, **but there are some errors in word choice and sentence construction. The vocabulary used is relatively simple, and the ideas are expressed in a straightforward manner. However, the essay lacks a clear focus and organization, and the conclusion is abrupt.**

Table 4: Pairs of *good* and *bad* explanations by type. From top to bottom, the source of low-quality is CONCISENESS, PLAUSIBILITY, and STANCE CLARITY.

LLMs. For details on the preliminary experiment and metrics, see the Appendix D.

5 Discussion

A key indicator of the utility of Rubrik is the level of agreement observed between the human evaluators who used it. Standard inter-rater agreement metrics are often inadequate for nested hierarchical data. Therefore, we designed a custom metric that accounts for both *superlabels* (explanation types) and *sublabels* (COMPONENTS and DIMENSIONS) in Rubrik, penalising discrepancies based on the difference in hierarchical level. Using this novel metric, we found an average inter-rater agreement of 0.86 and 0.878 for superlabels and sublabels, respectively, among humans. In selecting the third evaluator, our preliminary experiments revealed that LLMs tended to favour JUSTIFICATIONS, potentially inflating agreement scores on this first metric. To address this, we designed a second metric that weights the evaluations based on a comparison with both human and LLM judgments, providing a more accurate measure of performance. Using both custom metrics, we obtained scores of 0.841 (superlabel) and 0.86 (sublabel) for metric one, and 0.476 for the second. The latter, weighted metric led to the selection of `GPT-4o` as the third evaluator. As mentioned in Section 4.1.1, we decided to keep explanations, even if they are associated with an incorrect answer. Just as explanations are inherently tied to their goal, we hypothesised that they depend on the task. To explore this, we started by looking at the average performance of each annotator across tasks. Humans showed an average accuracy of T1: 70.46%, T2: 69.09%, T3: 80.78%, T4: 55.06%; LLMs showed T1: 78.94%, T2: 69.24%, T3: 87.42%, T4: 47.58%. Closed-source LLMs outperformed humans and open-source models, yet T4 proved the most challenging task, while T3 was the least challenging. Task difficulty was assessed to analyse explanation frequency and quality, illustrated in Figure 2. Overall, both LLMs and humans judged explanations to be mostly JUSTIFICATIONS. A notable observation is the low frequency of assignments with negative type (i.e., not an explanation). A closer look at the data revealed that these assignments were predominantly made by human evaluators. Furthermore, we found that T4 had a much higher proportion of ARGUMENTS than other tasks, whereas T3, the easiest task, had comparatively few. These results reveal insights into the

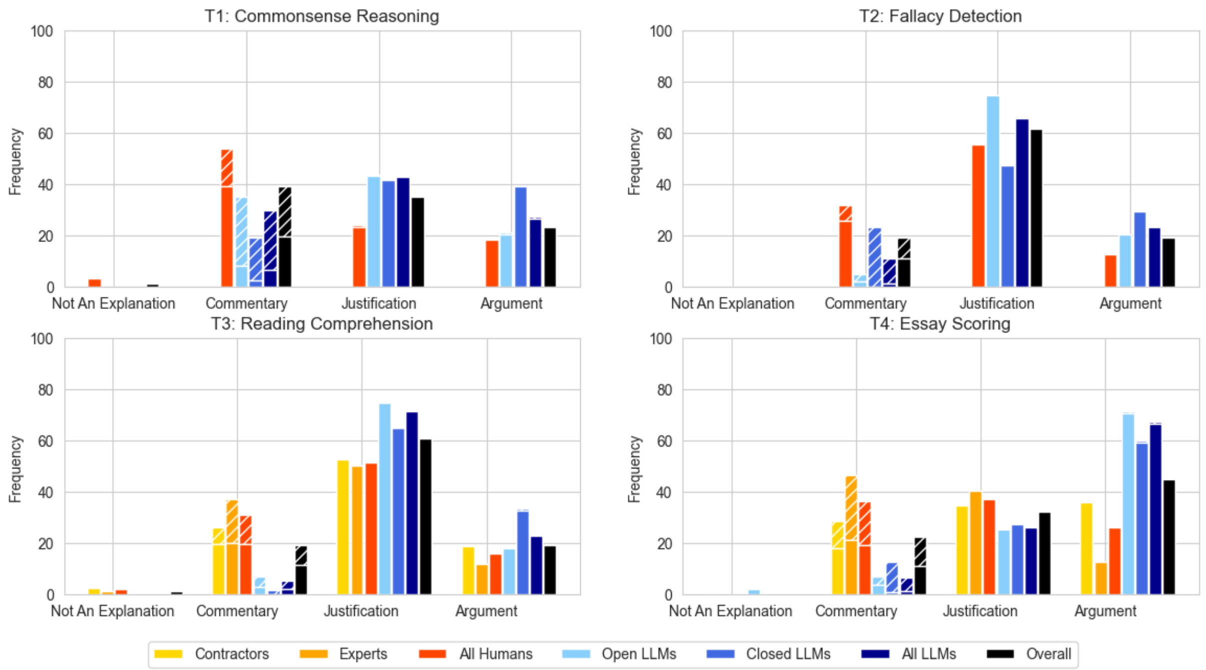


Figure 2: Frequencies of the different explanation types in each group of annotators as judged by and averaged across the three evaluators (two humans and `gpt-4o`). The patterned fill indicates the proportion of *bad* explanations of each type; the solid fill shows the proportion of *good* explanations of each type.

tendencies of humans and LLMs to generate JUSTIFICATIONS, while also highlighting the influence of task characteristics on the nature of generated explanations. T4 is a complex task that requires evaluators to go beyond simply recognising correct language use. They must also assess the effectiveness of the writing in achieving its intended purpose, which involves subjective judgments about argumentation, organization, and style. While some interpretation might be involved in understanding the context in T1, T2 and T3 the range of acceptable interpretations is much narrower. Thus, our results suggest that the presence of ARGUMENTS is correlated with the subjectivity of the task. The relationship between ARGUMENTS and task subjectivity is reinforced by the findings of our follow-up survey, where human annotators expressed lower confidence in T4. Upon further inspection of the frequency of ARGUMENTS across tasks, we found that `Sonnet 3.5`, while similar in terms of accuracy to `GPT-4o`, is more likely to produce this type of explanation.

Regarding the quality of the explanations, the number of *bad* explanations was low and concentrated in COMMENTARIES across tasks. The analysis of sublabel frequencies showed that the main source of a bad explanation was the lack of CONCISENESS, with open-source LLMs averaging 69.44% and closed-source LLMs averaging 98.18% on this sublabel. An example is shown in Table 4; the COMMENTARY is redundant, due to the repetition of details given in the question's context. For more details, see Figure 7 in Appendix F.3. Human explanations, on the other hand, were different between contractors and experts. Bad explanations produced by experts were due to GRAMMATICALITY, while contractors struggled with COHERENCE.

For a more in-depth exploration of the data, readers can refer to Appendix F.

6 Conclusion

This work introduces Rubrik, a novel evaluation rubric for assessing the quality of explanations, and a dataset. CUBE, which includes diverse explanations across four tasks, served as the testbed for evaluating Rubrik's effectiveness. Rubrik's design, rooted in educational principles, applies insights from education, XAI, and NLG literature. Our work contributes to the responsible integration of GenAI into critical decision-making processes, providing a foundation for future advancements in explanation quality assessment.

Limitations

Scoring strategy. Given the scope of this work, we opted for a binary evaluation strategy, categorising explanations as either *good* or *bad*. The task of establishing criteria for a *good* explanation presented a significant challenge, necessitating the identification and definition of relevant attributes. A more nuanced scoring system that reflects varying degrees of quality would be desirable. However, while a Likert scale might be a convenient choice, developing a valid and reliable graded scale specifically for explanations requires considerably more research. Our primary goal in this initial study was to assess the viability of our proposed rubric in its simplest form, laying the groundwork for more nuanced evaluations in future work. Furthermore, our approach does not explicitly assess the quality of reasoning itself. While a *good* explanation is generally an indicator of a good reasoning, a poor explanation could stem from how the reasoning is communicated rather than from the reasoning process itself. Although this is a complex problem, the development of methods for directly assessing reasoning quality is an interesting direction for future research.

Monolingual Data. The different attributes (DIMENSIONS) of a *good* explanation were taken from studies that exclusively considered English data. In turn, our work only includes datasets in English as well. In principle, the DIMENSIONS and definitions presented here should extend to other languages. However, it is possible that some will change depending on the cultural heritage, literature, and history. Indeed, the very concept of explanations may differ depending on the linguistic community, which may influence how explanation types, COMPONENTS or DIMENSIONS are prioritised or understood.

Annotators' Confidence Assessment. After completing the annotation tasks, human annotators were surveyed about their experience, including a self-assessment of their performance. These responses provided valuable context for interpreting the data analysis results. As for LLM annotators, they were prompted to assign probabilities reflecting their confidence in each answer option's correctness. While logit analysis would have been ideal, we hypothesized that requesting that information in the prompt would be sufficiently accurate, especially given that logit access was not available across all models (due to some being closed-source). However, the resulting probabilities often failed to sum to 100%, indicating a lack of consistent or meaningful probability assignment. Consequently, these assigned probabilities were not considered in the data analysis. Thus, we lack the means to make meaningful comparisons between human and LLM annotator confidence levels.

Ethical Considerations

Prior to commencing the study, ethical approval was obtained from a relevant Ethics Committee. Informed consent was obtained from all participants, and their anonymity/confidentiality was ensured throughout the research process.

In light of Baur (2020)'s of the current "AI hype", we acknowledge the potential for misinterpretation of GenAI capabilities, particularly the risk of users over-relying on automatic explanations in tasks where human oversight is crucial. Our work aims to mitigate this risk by providing an objective evaluation framework for model outputs. This framework enables informed decision-making regarding the selection of the most appropriate resource—whether human or automated—for a given task. For instance, Rubrik can identify instances where a less complex model is sufficient, or conversely, when human expertise is required.

Finally, we also recognise the potential for misuse of our framework. Indeed, the rubric could be exploited to deliberately generate misleading or poor-quality explanations. This could contribute to the spread of misinformation which poses a serious threat to informed decision-making. This risk highlights the importance of ensuring that the tool is used responsibly.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. *arXiv preprint*. ArXiv:2402.04614 [cs].

Michael Alley. 1996. Language: Being Concise. In Michael Alley, editor, *The Craft of Scientific Writing*, pages 119–127. Springer, New York, NY.

Mohammad Amiryousefi and Hossein Barati. 2011. Metadiscourse: exploring interaction in writing, ken hyland. *Continuum, London. Elixir Literature*, 40:5245–5250.

Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. Accessed: February 2025.

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pages 128–138, New York, NY, USA. Association for Computing Machinery.

Paris Arnold. 2023. IELTS Writing Band Descriptors.

Kaan Y. Balta, Arshia P. Javidan, Eric Walser, Robert Arntfield, and Ross Prager. 2025. Evaluating the Appropriateness, Consistency, and Readability of ChatGPT in Critical Care Recommendations. *Journal of Intensive Care Medicine*, 40(2):184–190. Publisher: SAGE Publications Inc STM.

Siu Wing Yee Barbara, Muhammad Afzaal, and Hessah Saleh Aldayel. 2024. A corpus-based comparison of linguistic markers of stance and genre in the academic writing of novice and advanced engineering learners. *Humanities and Social Sciences Communications*, 11(1):1–10.

Dorothea Baur. 2020. Four reasons why hyping AI is an ethical problem. Accessed: February 14, 2025.

Robert De Beaugrande and Wolfgang U. Dressler. 1981. *Introduction to Text Linguistics*. Longman. Google-Books-ID: TmhiAAAAMAAJ.

Leema Berland and Katherine Mcneill. 2012. For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson. *Science Education*, 96:808–813.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Antoine C Braet. 1992. Ethos, pathos and logos in aristotle's rhetoric: A re-examination. *Argumentation*, 6:307–320.

Ingo Brigandt. 2016. Why the Difference Between Explanation and Argument Matters to Science Education. *Science & Education*, 25.

Sylvain Bromberger. 1992. *On what we know we don't know: Explanation, theory, linguistics, and how questions shape them*. University of Chicago Press.

Gavin Brown. 2010. The Validity of Examination Essays in Higher Education: Issues and Responses. *Higher Education Quarterly*, 64:276–291.

Priscila G Brust-Renck, Rebecca B Weldon, and Valerie F Reyna. 2021. Judgment and decision making.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, pages 643–701. Place: Cambridge, MA Publisher: MIT Press.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Michael Canale. 1983. From communicative competence to communicative language pedagogy 1. In *Language and Communication*. Routledge. Num Pages: 26.

Mengyun Cao and Hai Zhuge. 2022. Automatic evaluation of summary on fidelity, conciseness and coherence for text summarization based on semantic link network. *Expert Systems with Applications*, 206:117777.

Christine Chin and David E. Brown. 2000. Learning in Science: A Comparison of Deep and Surface Approaches. *Journal of Research in Science Teaching*, 37(2):109–138. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291098-2736%28200002%2937%3A2%3C109%3A%3AAID-TEA3%3E3.0.CO%3B2-7.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Cohere for AI. 2024. Introducing command r plus on microsoft azure. `https://cohere.com/blog/command-r-plus-microsoft-azure`. Accessed: 2025-02-14.

James Collins. 1998. Strategies for Struggling Writers. *College Composition and Communication*, 49:298.

Ulla Connor. 1990. Linguistic/Rhetorical Measures for International Persuasive Student Writing. *Research in the Teaching of English*, 24(1):67–87. Publisher: ncte.org.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2024. The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*. Status: forthcoming.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Phillip Dawson. 2017. Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, 42(3):347–360.

Randy Devillez. 2003. *Writing: Step by Step*. Kendall Hunt Publishing Company. Google-Books-ID: 79oAePQ7Of0C.

Jean-Marc Dewaele. 2008. "Appropriateness" in foreign language acquisition and use: Some theoretical, methodological and ethical considerations. 46(3):245–265. Publisher: De Gruyter Mouton Section: International Review of Applied Linguistics in Language Teaching.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint*. ArXiv:1702.08608.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lee Dunn, Chris Morgan, Meg O'Reilly, and Sharon Parry. 2003. *The Student Assessment Handbook: New Directions in Traditional and Online Assessment*. Routledge, London.

M. Expósito-Ruiz, S. Pérez-Vicente, and F. Rivas-Ruiz. 2010. Statistical inference: Hypothesis testing. *Allergologia et Immunopathologia*, 38(5):266–277. Publisher: Elsevier.

Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. 2022. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 720–730.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling Fluency and Faithfulness for Diverse Neural Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 59–66. ISSN: 2374-3468, 2159-5399 Issue: 01 Journal Abbreviation: AAAI.

Anita Fetzer. 2012. Textual coherence as a pragmatic phenomenon. In Kasia M. Jaszczolt and Keith Allan, editors, *The Cambridge Handbook of Pragmatics*, Cambridge Handbooks in Language and Linguistics, pages 447–468. Cambridge University Press, Cambridge.

Anita Fetzer. 2018. Appropriateness in context.

Freeman, Richard, Lewis, Roger (BP Professor of Learning Development, and University of Humberside). 2016. *Planning and Implementing Assessment*. Routledge, London.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Milton Friedman. 1940. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1):86–92. Publisher: Institute of Mathematical Statistics.

Silvia García-Méndez, Francisco de Arriba-Pérez, and María del Carmen Somoza-López. 2024. A review on the use of large language models as virtual tutors. *Science & Education*, pages 1–16.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English. Version 2. Handbook and CD-ROM*.

M. A. K. Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. Routledge, London.

Sandra G Hart. 2006. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA.

SG Hart. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload/Elsevier*.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating Multiple Aspects of Coherence in Student Essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 185–192, Boston, Massachusetts, USA. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are LLM-based Evaluators Confusing NLG Quality Criteria? *arXiv preprint*. ArXiv:2402.12055 [cs].

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Mary Huba and Jann Freed. 2000. Learner-Centered Assessment on College Campuses: Sifting the Focus from Teaching to Learning. *Community College Journal of Research and Practice*, 24.

Dell Hymes. 1972. On Communicative Competence. In *Sociolinguistics*, pages 269–293. Harmondsworth: Penguin.

Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310. Place: Cambridge, MA Publisher: MIT Press.

Arshia P. Javidan, Tiam Feridooni, Lauren Gordon, and Sean A. Crawford. 2024. Evaluating the progression of artificial intelligence and large language models in medicine through comparative analysis of ChatGPT-3.5 and ChatGPT-4 in generating vascular surgery recommendations. *JVS-Vascular Insights*, 2:100049.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical Fallacy Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samia Kabir, David N Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024a. Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024b. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–17, New York, NY, USA. Association for Computing Machinery.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. pages 6300–6308.

Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.*, 57(1):227–254.

Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 385–404.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Cherise Kristoffersen. 2019. Where do my words come from? Towards methods for analyzing word choice in primary level writing. *Apples - Journal of Applied Language Studies*, 13(3):59–75. Number: 3.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, 50(3):1030–1046.

Kristopher Kyle and Scott A. Crossley. 2015. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, 49(4):757–786. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tesq.194.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. ICLE++: Modeling Fine-Grained Traits for Holistic Essay Scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Elizabeth Cloninger Long. 2007. *College writing resources with readings*. New York : Pearson/Longman.

Andrea A Lunsford, Kirt H Wilson, and Rosa A Eberly. 2008. *The SAGE handbook of rhetorical studies*. Sage Publications.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Danielle McNamara and Com. 2010. *Cohesion, coherence, and expert evaluations of writing proficiency*. Journal Abbreviation: Proceedings of the 32nd Annual Conference of the Cognitive Science Society Publication Title: Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press. Google-Books-ID: xSPeAgAAQBAJ.

Katharine L. McNeill and Joseph Krajcik. 2007. Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In *Thinking with data*, Carnegie Mellon symposia on cognition, pages 233–265. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Katherine Mcneill, David Lizotte, Joseph Krajcik, and Ronald Marx. 2006. Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials. *Journal of the Learning Sciences*, 15:153–191.

Katherine L. McNeill and Joseph Krajcik. 2008. Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, 45(1):53–78. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.20201.

Tim Miller. 2019a. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Tim Miller. 2019b. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Eleni Miltsakaki. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10:25–55.

Brian North and Enrica Piccardo. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion volume Language Policy Programme Education Policy Division Education Department Council of Europe*.

Stellan Ohlsson. 2002. Generating and understanding qualitative explanations. In *The psychology of science text comprehension*, pages 91–128. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

OpenAI. 2024. Hello gpt-4o. Accessed: February 2025.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788.

Isaac Persing and Vincent Ng. 2013. Modeling Thesis Clarity in Student Essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Maxime Peyrard. 2019. A Simple Theoretical Model of Importance for Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

J. B. Pride. 1972. *Sociolinguistics : selected readings*. Harmondsworth, Penguin.

Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 83–88, New York, NY, USA. Association for Computing Machinery.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

William A. Sandoval. 2003. Conceptual and Epistemic Aspects of Students' Scientific Explanations. *The Journal of the Learning Sciences*, 12(1):5–51. Publisher: Taylor & Francis, Ltd.

Yash Saxena, Sarthak Chopra, and Arunendra Mani Tripathi. 2024. Evaluating consistency and reasoning capabilities of large language models. *arXiv preprint arXiv:2404.16478*.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying Argumentation Schemes for Essay Scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

Micol Spitale, Minja Axelsson, and Hatice Gunes. 2024. Appropriateness of LLM-equipped Robotic Well-being Coach Language in the Workplace: A Qualitative Evaluation. *arXiv preprint*. ArXiv:2401.14935 [cs].

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Manfred Stede. 2002. Lexical Choice Criteria in Language Generation.

Dannelle D. Stevens and Antonia J. Levi. 2004. *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback and Promote Student Learning*. Stylus Publishing, LLC. Publication Title: Stylus Publishing, LLC ERIC Number: ED515062.

Susan Strauss and Parastou Feiz. 2013. Discourse analysis: Putting our worlds into words. *Discourse Analysis: Putting our Worlds into Words*, pages 1–411.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Stephen Toulmin. 1958. *The Uses of Arguments*, 1 edition. Cambridge University Press.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272. Publisher: Nature Publishing Group.

Barbara E. Walvoord and Virginia Johnson Anderson. 1998. *Effective Grading: A Tool for Learning and Assessment*. Jossey-Bass Publishers, 350 Sansome St. ERIC Number: ED416810.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 2903–2908, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Johnny Tian-Zheng Wei, Khiem Pham, Brian Dillon, and Brendan O'Connor. 2018. Evaluating Syntactic Properties of Seq2seq Output with a Broad Coverage HPSG: A Case Study on Machine Translation. *arXiv preprint*. ArXiv:1809.02035 [cs].

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *NeurIPS Datasets and Benchmarks*.

Yuxiang Wu and Baotian Hu. 2018. Learning to Extract Coherent Summary via Deep Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). Number: 1.

Helen Yannakoudakis, Øistein Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.

Laura Zangori, Cory T. Forbes, and Mandy Biggers. 2013. Fostering student sense making in elementary science learning environments: Elementary teachers' use of science curriculum materials to promote explanation construction. *Journal of Research in Science Teaching*, 50(8):989–1017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.21104.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Jiegen Zhang. 2006. *A Text-based Approach to Cohesion and Coherence*. Ph.D. thesis.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint*. ArXiv:2309.01219 [cs].

A Rubric Creation

We conducted an extensive review of NLP literature including work in Natural Language Generation (NLG) such as Machine Translation (MT) and Educational NLP (including Grammatical Error Correction and Automated Essay Scoring), but also in Linguistics and Cognitive Science. In doing so, we recorded the names of qualities (or DIMENSIONS) that people have looked for in explanations or argumentative writing more generally, and, when present, their definitions. We also kept note of how these qualities have been evaluated in a target text, using either human annotators or automated methods. See Table 5 for the exhaustive list.

Dimension Name
Appropriateness
Adequacy
Clarity
Coherence
Cohesion
Completeness
Conciseness
Consistency
Comprehensibility
Comprehensiveness
Correctness
Factuality
Faithfulness
Fidelity
Fluency
Grammaticality
Interpretability
Organisation
Persuasiveness
Plausibility
Readability
Reasonableness
Transparency
Truth of likelihood
Usefulness
Word choice

Table 5: Exhaustive list of the quality DIMENSIONS of explanation we found when surveying the literature. We highlight in **bold** the names of the DIMENSIONS we included in our rubric *verbatim*.

Below we describe how we defined and chose the eight DIMENSIONS that are represented in Rubrik. We also introduce a few of the many qualities that were considered and explain why they were excluded, as a demonstration of our overall process. Though we cannot be exhaustive at this time, we rigorously researched each and every one of the dimensions mentioned in Table 5. The final definitions we used in the automated evaluation prompts are provided in Appendix E. The full rubric with examples will be released upon acceptance.

A.1 Grammaticality

Grammaticality, though essential, was surprisingly hard to define. This was largely due to the fact that grammar has a long-standing tradition in a variety of fields—including Linguistics, Psychology, Education, and Cognitive Science—which have each contributed different perspectives and theories over time. As a result there is no single, universally accepted definition. Definitions which originate from the field of Linguistics tend to be highly theoretical, and as a result, quite impractical. A classic example is Chomsky (1965, Chapter 1, p.2) for whom the "grammar of a language purports to be a description of the ideal speaker-hearer's intrinsic competence", which has been criticised for being too abstract and disconnected from actual language use (Pride, 1972, Chapter 18). On the other hand, most NLP studies assume that the definition of grammaticality is common knowledge and avoid going through the trouble of formally defining it in the context of their work (e.g., Wei et al., 2018). In fact, it is openly admitted that "Grammatical Error Correction is thus something of a misnomer, but is nevertheless now commonly understood to encompass errors that are not always strictly grammatical in nature" (Bryant et al., 2023).

However, to avoid relying on our intuition of what a grammatical explanation is, we needed to bridge the gap between theory and practice, and find a definition that could be both pragmatic and grounded in the literature. We did find one in a paper by Hu et al. (2024, Table 10), similarly focused on the evaluation of LLM outputs, which defines grammaticality as measuring "whether the target text is grammatically correct without any lexical or syntax errors, regardless of its content and meaning. Consider whether the target text itself complies with the English standard usage and rules of grammar, such as tense errors, misspellings, incorrect prepositions, collocation misusages, and so on." In using this definition, it is quite straightforward to classify Grammaticality as a **Language** DIMENSION as it in no way attends to the content of the text.

A.2 Conciseness

In contrast, we found Conciseness to be well-documented across many literatures and much less controversial. In Education, "concise writing gets to the point quickly and does not introduce unnecessary information" (Long, 2007, p.25) and requires you to "cut fat" into your writing by "eliminating redundancies, eliminating writing zeroes, reducing sentences to simplest form, and cutting bureaucratic waste" (Alley, 1996, Chapter 8). Similarly, in NLP, Cao and Zhuge (2022) define it as a measure of "non-redundancy" in text, sometimes through the number of repeated words (Peyrard, 2019) or through computing sentence similarities (Wan et al., 2007).

We finally opted for Kabir et al. (2024b)'s comprehensive taxonomy of three conciseness issues:

> *Redundant* sentences reiterate information stated in the question or in other parts of the answer. *Irrelevant* sentences talk about concepts that are out of the scope of the question being asked. And lastly, *Excess* sentences provide information that is not required to understand the answer.

Not only were these issues identified when evaluating ChatGPT answers, a task closely related to ours, we additionally felt that they encompassed all the elements that were individually picked out in previous definitions. Note that since this definition is concerned with redundant, irrelevant or excess information, not just language, we decided to classify Conciseness as a **Content** dimension.

A.3 Fluency

For a while, we considered fluency, an important notion in Machine Translation, which is generally evaluated by humans (e.g., Callison-Burch et al., 2007; Graham et al., 2013; Bojar et al., 2016), or using automated metrics (e.g., Toral and Sánchez-Cartagena, 2017; Martindale et al., 2019; Feng et al., 2020). In the first case, we found that human annotators were almost never provided with a proper definition of fluency and expected to use their intuition of what the word meant via prompts like "*how do you judge the fluency of this translation?*" in Callison-Burch et al. (2007) or "*read the text below and rate it by how much you agree that: the text is fluent English*" in (Graham et al., 2013). In the latter case, the metrics used were only considered to be proxies for fluency which was never actually defined.

As with Grammaticality, Hu et al. (2024, Table 9) provided the following definition: "[fluency] measures the quality of individual sentences, are they grammatically correct, non-repetitive, and in accord with common English usage, with clear meanings", which seemed to overlap both our definitions for Conciseness and Grammaticality. Since our goal was to reach a set of well-delineated, atomic dimensions, we chose to discard it.

A.4 Cohesion

Cohesion is a very important notion in Linguistics and is classically defined by Halliday and Hasan (2014, p.4) as:

> occur[ring] where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it. When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated into the text.

Unfortunately, as with Grammaticality, this definition is not accessible to most people and is far too theoretical.

However, Cohesion is also widely present in Education, particularly in writing assessment and teaching literature, due to the common idea that a written text's quality is highly related to its level of Cohesion (McNamara and Com, 2010). This belief is reflected in the literature about writing (e.g., Collins, 1998, Devillez, 2003) and the rubrics that teachers use to assess writing (e.g., Arnold, 2023; Crossley et al.,2024). It is notably defined by McNamara and Com (2010) as follows:

> Cohesion refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text. For example, overlapping words and concepts between sentences indicate that the same ideas are being referred to across sentences. Likewise, connectives such as 'because', 'therefore', and 'consequently', inform the reader that there are relationships between ideas and the nature of those relationships.

Or more simply as the "appropriate use of transition phrases" by Ke and Ng (2019, Table 1). For our purposes, we prefer these pragmatic definitions to those offered by Linguistics.

From these definitions, it seems that Cohesion is only concerned with **Language** not the content of a text. In fact, the dimension has also been examined through automated tools like Coh-Metrix (McNamara et al., 2014) or TAACO (Crossley et al., 2016), which use a compound of linguistic metrics like the Type Toke Ration (TTR; McCarthy and Jarvis, 2007) as proxies for Cohesion.

A.5 Coherence

A related notion to Cohesion is Coherence. It has been defined in Linguistics as a "continuity of sense" by Beaugrande and Dressler (1981, p.84), or more concretely as "the state of being logically consistent and connected" (Fetzer, 2012). It is also an important notion in Document Summarisation, where Coherence is similarly defined as "what makes multiple sentences semantically, logically and syntactically coherent" (Yao et al., 2017). It is also frequently evaluated writing assessment either by humans (e.g., Higgins et al., 2004) or via automated methods (e.g., Higgins et al., 2004; Miltsakaki, 2004; Wu and Hu, 2018).

Where Cohesion is an "overt (or explicit) linguistic-surface phenomenon, [...] coherence is a covert (or implicit) deep-structure phenomenon". But while Coherence is more concerned with meaning (i.e., **Content**) than form (Fetzer, 2012), it also "depends on a number of factors, including explicit cohesion cues, implicit cohesion cues (which are more closely linked to text coherence than are explicit cues), and nonlinguistic factors such as prior knowledge and reading skill" (Crossley et al., 2016). They are thus "interdependent" notions (Zhang, 2006). To portray this in our rubric, we chose to similarly relate both DIMENSIONS: an explanation should thus not be labelled as coherent without first being judged as cohesive.

A.6 Clarity

We first encountered this quality while looking at writing education papers, where clarity generally "refers to how clearly an author explains the thesis of her essay, i.e., the position she argues for with respect to the topic on which the essay is written" (Persing and Ng, 2013). It also appears in the ICLE++ corpus of persuasive student essays (Granger et al., 2009; Li and Ng, 2024), an important dataset in the field of Automated Written Assessment. However, the definitions we found were far too vague and we struggled to find more formal or practical descriptions of the term which seemed to support Beaugrande and Dressler (1981, Chapter 2)'s claim that clarity is "too vague and subjective to be reliably defined and quantified". We ultimately decided to drop this DIMENSION.

A.7 Word Choice

The Word Choice DIMENSION is broadly defined as "the choice and aptness of the vocabulary used" (Mathias and Bhattacharyya, 2018). It is frequently included in written assessment rubrics (e.g, see the very detailed 6-point rubric for this dimension in the ASAP[5] corpus) and the focus of automated assessment research (e.g., Kyle and Crossley, 2015; Kyle et al., 2018; Kristoffersen, 2019).

We also came across Stede (2002)'s work on lexical choice for NLG:

> Generally speaking, the point of "interesting" language generation (that is, more than merely mapping semantic elements one-to-one onto words) is **to tailor the output to the situation at hand**, where "situation" is to be taken in the widest sense, including the regional setting, the topic of the discourse, the social relationships between discourse participants, etc.

Though not explicitly defining Word Choice, the above citation introduces the idea that every "interesting" or *good* utterance (or in our case, explanation) is made within a given "situation" and thus evaluating the language of that utterance should be context-dependent. It is this **context** that dictates what is "apt" (Mathias and Bhattacharyya, 2018). Realising that it is necessary to define an evaluation *context* before starting any kind of evaluation (see Section 3.2) was a turning point for our rubric.

Now, *context*-appropriateness relies on both form and content. However, due to the strong emphasis on evaluating Word Choice as a surface-level feature, not a content one, in automated assessment research, we chose to classify it as a **Language** DIMENSION.

A.8 Appropriateness

Appropriateness defined in Linguistics by Canale (1983) as "the extent to which particular communicative functions [...] and **ideas** are judged to be proper in a given situation" or as "an optimal mapping between context and speech, or as 'natural speech', is also connected intrinsically with

[5] The original dataset and annotation guidelines can be downloaded from https://www.kaggle.com/c/asap-aes/data.

the sociocultural notions of politeness and impoliteness" by Fetzer (2018). This term also occasionally appears in AI literature as something we must ensure in the systems we develop, and thus, evaluate (e.g., Spitale et al., 2024; Javidan et al., 2024; Balta et al., 2025;). There, it is more often related to other qualities such as safety, consistency, and readability. Hence, Appropriateness is a complex, multi-faceted dimension which also relies on *context*.

For our purpose, we needed to relate this DIMENSION to Word Choice. For this, we turned to the prominent sociolinguist, Dell Hymes who "pointed out that appropriateness [depend] both on linguistic and sociocultural competence" (Dewaele, 2008), and defined it as "what to say to whom in what circumstances and how to say it" in Hymes (1972, p.277). We deem that this last part, "how to say it" is already encompassed by our definition of Word Choice. Further, "to whom in what circumstances" refers to our very own definition of the *context*, which leaves us with the "what to say" for Appropriateness, that is, the **Content**.

A.9 Plausibility

In reading around the topic of explanations in AI, we came across the following trait: "the **truth of likelihood** of an explanation is considered an important criterion of a good explanation" in a paper by Miller (2019b). The term was used to refer to facts that were judged as "either true or likely to be true by the explainee." We note that in no way is our rubric intended to evaluate the truth condition of explanations. However, we felt that it was important that our rubric allows for JUSTIFICATION to be evaluated as *bad* or of *bad* quality if their EVIDENCE was deemed implausible by the evaluator. After some research, we could not find any other mention of the "truth of likelihood" and sought a more general name for our DIMENSION.

A related notion was Plausibility which was present in similar literature and already being used to evaluate explanations. For instance, Agarwal et al. (2024) who define plausible explanations as being "seemingly logical and coherent to human users" or as "being convincing towards the model prediction, regardless of whether the model was correct or whether the interpretation is faithful" by Jacovi and Goldberg (2021). Though not exactly similar, the latter introduces the idea that using Plausibility as criteria for a *good* explanation might encourage deception. As a result, the authors advise against pursuing this DIMENSION.

Taking this warning into consideration, it was important to us to centre our definition of Plausibility around the Evidence component (2.a), and we modified Agarwal et al. (2024)'s Definition 1, substituting the word "explanation" with "evidence":

> An evidence* is considered plausible if it is coherent with human reasoning and understanding.

A.10 Stance Clarity

Whenever we found a mention of ARGUMENTS in the literature, the concept of persuasiveness was almost always mentioned. It thus seemed natural that it would be included in our rubric. We first looked at the notion of "argument strength" in persuasive writing defined as " Whenever we found a mention of ARGUMENTS in the literature, the concept of persuasiveness was almost always mentioned. It thus seemed natural that it would be included in our rubric. We first looked at the notion of "argument strength" in persuasive writing which is defined, in an admittedly very circular fashion, as "the strength of the argument an essay makes for its thesis" and evaluated by Persing and Ng (2015). In a similar vein, we discovered work by Song et al. (2014) and Stab and Gurevych (2014) which designed argument schemes for annotating arguments manually in student essays. Yet, none of the definitions we found seemed right.

We then turned to persuasiveness in rhetoric, and found Connor (1990, Table 5)'s Persuasive Appeals Scale. Though very useful, we struggled to see whether these were in fact COMPONENTS or indeed a DIMENSION, and where to fit them in our rubric. After some iterations, we arrived at the fact that the presence of Affective appeals and Qualifiers in an argument help us understand what the explainer's "stance" is, that is, their personal "feeling, attitude, perspective, or position as enacted in discourse" (Strauss and Feiz, 2013). By that point, it felt like persuasiveness was too vague and we coined the term "Stance Clarity" for our last DIMENSION.

B Data Selection

Considering the fact that the four datasets we chose to work with were all of different sizes, we chose to only work with a subset of each dataset: namely $n = 1000$ instances for each task. Thus, our *base set* has a total of 4000 instances.

We collected a set of human-written (see Sec. C.1) and LLM-generated explanations (see Sec. C.2). Due to limitations in time and resources, only a subset of the 1000 instances was shown to the annotators: namely $n = 110$ instances for each task. Thus, our *annotation set* has 440 instances. The following subsections detail the subset selection criteria.

B.1 Commonsense reasoning

Base set. Each CONTEXT in the HELLASWAG dataset is taken either from ActivityNet's video captions or WikiHow's how-to-articles. During the annotator's training (see Sec. C.1.1), questions whose context made reference to a video were constantly flagged as *"not clear or ambiguous"*. Thus, we filtered instances that include the word *"camera"*, *"video"* or *"clip"*. After that, instances were selected randomly, making sure that the correct answers were distributed as evenly as possible across the four options (A-D), with roughly 25% assigned to each.

Correct answer	Base set	Ann set
A	267	27
B	228	28
C	266	27
D	239	28
Total	1000	110

Table 6: Distribution of questions across each possible correct answer for T1's *base set* and *annotation set*.

Annotation set. Since the *base set* already had an even distribution of the four answer choices, we selected a proportionally representative subset of 110 instances. See Table 6 for a summary of this selection process.

B.2 Fallacy detection

Base set. Jin et al. (2022) classified fallacies in the LOGIC dataset into 13 fallacy types. Due to potential overlap between some of the initial types and dataset imbalance, we focused on a subset of 7 types.

Selecting instances within the 30-300 character range effectively eliminated instances requiring specialized political or religious knowledge, ensuring consistent annotation based on general knowledge. After manual inspection, we removed some duplicated instances and statements that were not exactly fallacies, but rather someone's opinion on a topic. We also identified a few instances that were incorrectly labelled (i.e., were assigned the wrong fallacy type). Those were re-labelled and kept in the final subset. Table 7 shows the final distribution of our subset.

Logical Fallacy	Inc	Base set	Ann set
Faulty Generalization	✓	289	17
Ad Hominem	✗		
Ad Populum	✗		
False Causality	✓	154	15
Circular Claim	✓	112	15
Appeal to Emotion	✓	109	15
Fallacy of Relevance	✗		
Deductive Fallacy	✓	120	15
Intentional Fallacy	✗		
Fallacy of Extension	✗		
False Dilemma	✓	118	17
Fallacy of Credibility	✓	95	16
Equivocation	✗		
Total		1000	110

Table 7: Distribution of instances across each fallacy type for T2's *base set* and *annotation set*.

Annotation set. This task was originally framed as a classification task. For the purposes of this research, we adapted the task to follow an MCQ format, where the CONTEXT was the fallacy statement, and each of the fallacy types was listed as ANSWER CHOICES. We aimed for a balanced distribution of correct answers across the seven options (A-G). Instances were selected randomly from the *base set*. See Table 7 for a summary of this selection process.

B.3 Reading comprehension

Base set. RACE data is grouped by difficulty (RACE-M: middle school; RACE-H: high school). To better understand the dataset, authors subdivided questions into five reasoning categories. Since the *Passage Summarization* and *World Knowledge* do not fully require people to carefully read through the passage to answer, we focused on the other three question types: *Detail Reasoning*, *Whole Picture Reasoning*, and *Attitude Analysis*. To be specific, the answer to *Detail Reasoning* questions can not be found by simply matching the question with the passage, which needs people to provide reasons for their choices. For *Whole Picture Reasoning* questions, people need to understand the entire story to obtain the correct answer with evidence. *Attitude Analysis* question asks about the author's or a character's opinions or attitudes.

Unfortunately, there are no assigned question types in the published dataset; hence, we manually selected data based on the description and examples given by Lai et al. (2017) and reviewed them to ensure quality.

Question type	Inc	Base set	Ann set
Detail reasoning	✓	400	36
Whole-picture reasoning	✓	400	37
Passage summarization	✗		
Attitude analysis	✓	200	37
World knowledge	✗		
Total		1000	110

Table 8: Distribution of text passages across each question type for T3's *base set* and *annotation set*.

Annotation set. Each question in RACE has four answer choices (A-D). We aimed for a balanced distribution of instances of correct answers across options within each question type. Instances were randomly selected from the *base set*, targeting a proportion of approximately 25% per option. See Table 8 for a summary of this selection process.

B.4 Essay Scoring

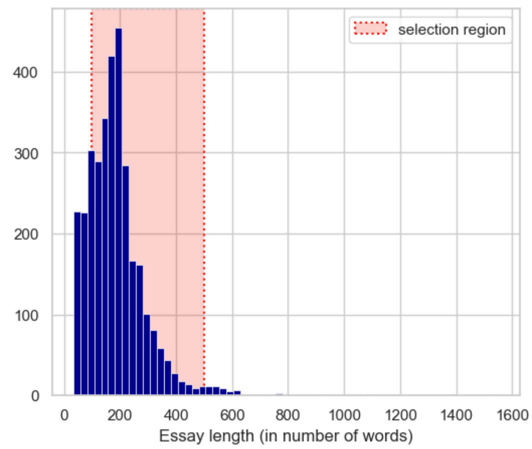
Base set. In the W&I corpus, essays range between 33 and 1,551 words in length. Figure 3(a) plots this distribution. We chose to exclude essays of less than 100 words, and more than 500 words, to avoid selecting essays sitting on either extreme of this distribution. Indeed, essays that are too short might contain too little information to be interesting to evaluate; essays that are long might exceed the limits of LLM contexts or prove too time-taking to annotate for humans. This step left us with a remaining total of 2,598 essays (833 A-scored essays, 1,039 B-scored essays, and 726 C-scored essays). Then, we randomly sampled 333 essays from each CEFR level group (334 for the B level) to obtain our *base set* of 1000 essays. We additionally randomly selected 3 essays (one of each CEFR level) from the remaining pool of essays to be used as examples in our experiments.

Annotation set. For our *annotation set*, we again selected randomly from the *base set*, aiming for a balanced distribution of essays across the three CEFR levels. See Table 10 for a summary of this selection process.

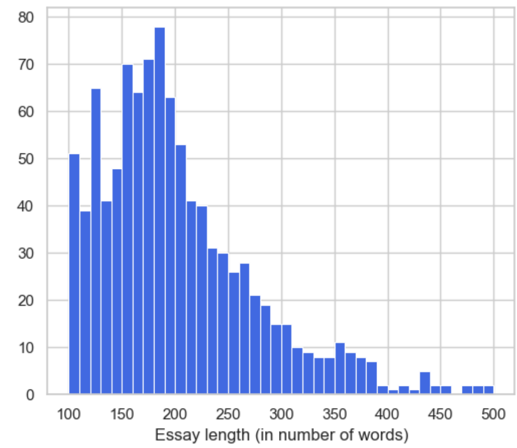
C Data Collection

C.1 Human Annotators

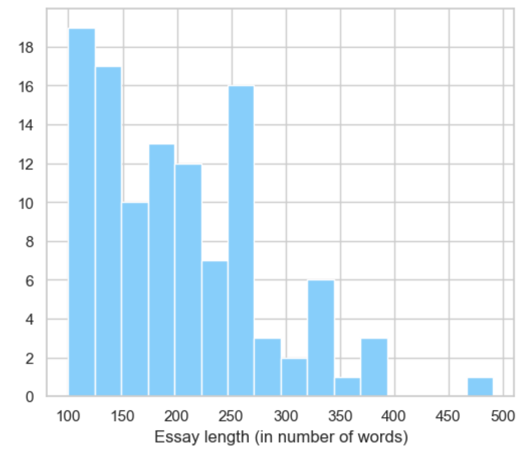
We recruited seven human annotators: four research assistants (RA's) and three professional annotators (PA's). One of the main authors, along



(a) W&I corpus word count distribution. We highlight in orange the region from which the *base set* essays were selected.



(b) *Base set* word count distribution



(c) *Annotation set* word count distribution.

Figure 3: Plotting the word count distributions

Essay Grade	W&I	Base set	Ann set
A	1430	333	36
B	1100	334	37
C	770	333	37
Total	3300	1000	110

Table 9: Distribution of W&I essays across each CEFR level for T4's *base set* and *annotation set*.

Essay Grade	W&I		Base set		Ann set	
	μ	σ	μ	σ	μ	σ
A	125	70	163	56	150	51
B	211	100	207	73	205	71
C	262	132	235	71	245	77
Overall	186	113	201	73	201	78

Table 10: Mean (μ) and standard deviation (σ) word count of the essays in the W&I corpus, the *base set*, and the *annotation set* (rounded to the nearest integer).

with a senior researcher, led the RA's recruiting efforts, which included conducting interviews with potential candidates. We selected individuals who appeared to have strong abilities in **attention to detail**, **assessment**, and **strong language skills.** These skills were essential for *completing* the assigned reasoning and language tasks. The PA's were annotators who were specially trained EFL (English as a Foreign Language) teachers and examiners. The annotators were paid an hourly rate of £22.59 for their work. We anonymized the annotations by not including personally identifiable information. Each annotator was identified with a randomly assigned ID (e.g., 000005FB, 000004E4)

C.1.1 Training

All annotators received a detailed annotation guide that introduces the four tasks, and provides a number of annotated examples (question + answer choices + correct answer) for each task. The examples are for them to familiarise themselves with the tasks. Since T2 necessitates some familiarity with fallacious reasoning, this task is further supported by an appendix with definitions of all fallacy types.[6] We do not include explanations so as to not bias the annotators as to what a *good* explanation should look like. The annotation guide also includes a series of guidelines they should abide by during the annotation process.

Upon reading the annotation guide, annotators were asked to write explanations for the guide's annotated examples. Their explanations were reviewed by two of the main authors to ensure they were acceptable in terms of format and length.[7] Unless absolutely necessary, annotators did not receive any feedback on their explanations.

Subsequently, each annotator received an invite-only Google Spreadsheet with a set of 15 to 40 examples per task.[8] Before start working through the tasks, the annotators were reminded that:

1. They have 20 minutes to complete a task. They should not necessarily aim to complete all of the provided questions as we intentionally put more than what we thought they would do in 20 minutes.

2. At the end of the 20 minutes, they should move to the next task without delay and not go back to any previous task (even if they have spare time).

3. They only have to select one single answer per question given a set of potential answers, and will not have to explain their decision process during the training phase.

4. They can attempt the questions in any order. However, they should not spend more than 5 minutes on a question. In order to manage their time more efficiently, it is recommended that they (1) flag difficult questions as they find them, moving immediately to the next one. In order words, they should **focus on answering the questions where they feel confident** (2) go back to the flagged questions and try to solve them, if they still have time. If it keeps taking longer, **flag the questions either as "too difficult" or "not clear or ambiguous"**.

5. They can consult the annotation guide at any time.

After the training, their files were marked by two of the main authors. Annotators were asked to review their answers in order to learn from their mistakes.

[6]Specifically, the information provided by Jin et al. (2022) in their Appendix D.

[7]Since the guide does not specify a minimum length for the explanations, we made sure annotators wrote complete sentences as opposed to short phrases.

[8]The number varied according to the difficulty of each task. For example, the questions in fallacy detection were short but required more specific knowledge while reading comprehension contained longer but easier-to-read texts.

C.1.2 Annotation Process

As shown in Table 11, we followed a two-phase iterative approach. Phase 1 included a small batch from the T2, T3 and T4's *annotation set*. Note that T1 data was excluded due to necessary revisions based on training feedback (see Section B.1). Once completed, explanations underwent the same review process as those used during the annotation training. Our training scheme proved to be effective, resulting in minimal necessary corrections to the annotations. Phase 2 included the remaining instances in the *annotation set*.

Phase	T1	T2	T3	T4
1	0	28	28	28
2	110	82	82	82
Total	110	110	110	110

Table 11: Distribution of task instances across each annotation phase.

Annotators generally adhered to the allocated time frame of 5 minutes per instance, which translated to approximately 7 hours of annotation in Phase 1 and 30 hours in Phase 2. Upon completion, their files were marked and formatted as a JSON file.

C.1.3 Follow-up Survey

After completing the annotation, we asked the annotators to take a brief follow-up survey. We collected task load data for each of the four tasks using all six NASA-TLX items on a 9-point scale (1-10) (Hart, 1988, 2006). We considered the items individually, as well as their sum, as has been done in prior work (e.g., Quinn and Zhai,2016; Arnold et al., 2020).

Figure 4 shows box-plot representations of the responses from the NASA-TLX surveys, on which we performed Friedman tests (Friedman, 1940) using the `friedmanchisquare` function of the `scipy` Python library (Virtanen et al., 2020). Taking the accepted standard $\alpha = 0.05$ as the significance threshold (Expósito-Ruiz et al., 2010), we found significant differences for performance ($\chi^2 = 8.11$, p-value $= 0.044$) only. Annotators generally reported a lower sense of achievement in T2 and T4, than in T1 and T3.

In the survey, we also included the two open-ended questions to learn more about the annotators' individual approaches to writing the explanations: specifically, whether they had a particular audience in mind, and what they thought the purpose of the explanations was. We include the exact wording of the questions below:

Q1: *The intended recipient of our writing shapes our choice of language and style. Different audiences have different expectations, knowledge levels, and interests. When writing your explanations, did you have a specific audience in mind, or were you writing for a general audience?*

Q2: *Explanations can serve a range of purposes: (1) provide an understanding of why a choice was made, (2) justify how that choice was made by providing some evidence, (3) convince others that the choice was correct, and (4) other. When writing your explanations, what were you trying to achieve?*

In response to **Q1**, some annotators reported targeting a "specific" audience, such as researchers or students. On the other hand, one annotator explicitly aimed for a general audience. Others assumed an educated readership with basic linguistic knowledge of English without necessarily being specific about who they might be. Notably, one annotator expressed frustration at the lack of clarity regarding the intended readership. The diversity in the annotators' conceptual audiences is very much echoed in the variety of tones used and the level of depth of the explanations we collected (refer to Table 4 for example).

In response to **Q2**, five out of the six annotators that completed the survey chose (1) as their intended purpose which roughly matches our idea of what a COMMENTARY should do. The remaining annotator sought to justify their choice with evidence (2). While annotators assumed similar strategies, it is interesting to see that they in fact often went well beyond simply providing an understanding of why a choice was made, and provided a majority of JUSTIFICATIONS instead (see Figure 2).

C.2 LLM Annotators

Six different models were used to generate annotations. They were chosen based on coverage of different model sizes, architectures and diversity of sources.

- *Llama-3.1-8B-Instruct*[9]: It belongs to the family of Llama3.1 models published by Meta AI

[9] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

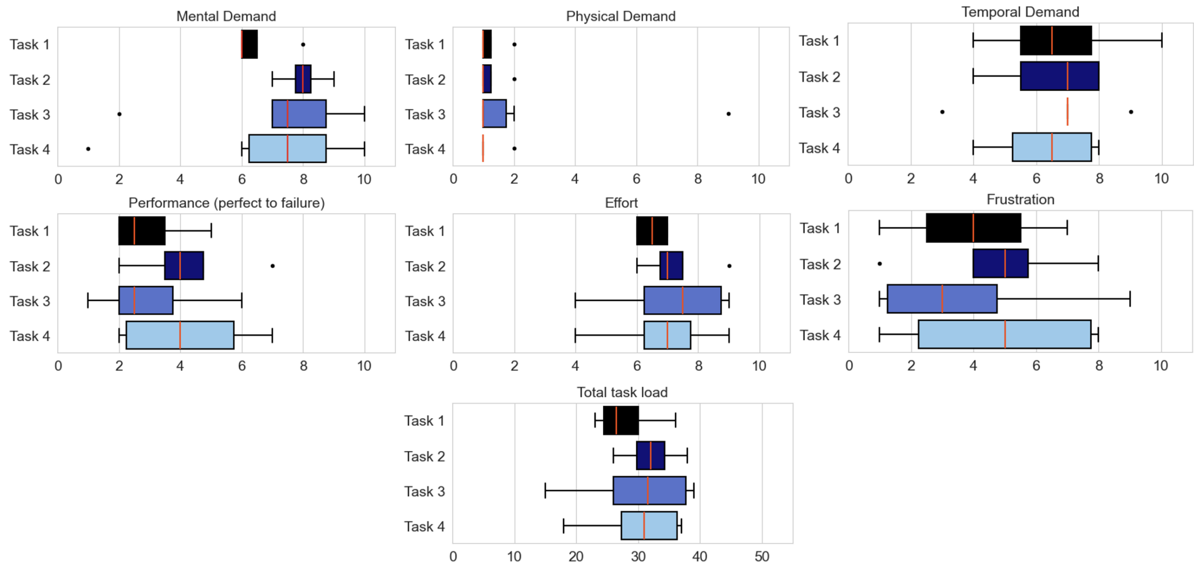


Figure 4: Box-plots of the six NASA-TLX items on a 9 point scale and their sum total. The median is shown in red.

under the Llama3 community license. It incorporates a context window of 128k length and is pre-trained on a corpus of about 15 trillion tokens.

- *gemma-2-9b-it*[10]: The model is a lightweight open-source model from Google that also supports a 128k length context window. It's trained on 8 trillion tokens of data covering web documents, code, mathematics and more.

- *Mixtral-8x7B-Instruct-v0.1*[11]: t is a pretrained generative Sparse Mixture of Experts model from mistralai. It has a context window of 32k tokens and is pre-trained on data extracted from open web.

- *c4ai-command-r-plus-08-2024*[12]: This is a 104B parameter multilingual model released from Cohere For AI. It supports a context length of 128K.

- *GPT-4o*[13]: GPT-4o is a multimodal model from OpenAI capable of processing and generating text, images, and audio. The parameter count of GPT-4o has not been publicly disclosed.

- *Claude 3.5 Sonnet (claude-3-5-sonnet-20240620)*[14]: This is an LLM model from Anthropic with improvements in reasoning, language understanding, and coding. The parameter count of Claude 3.5 Sonnet has not been publicly disclosed.

All open-source models were run on NVIDIA A100 GPUs using bf16 precision. We used the latest checkpoints of all open-weight models available at the time of the experiment, along with the default pretrained tokenizers provided for each model. A temperature of 0 was used for all models, including `Sonnet 3.5` and `GPT-4o`, which we accessed via API (for some HuggingFace models, we used 0.01 or set `do_sample=False` due to implementation constraints).

C.2.1 Prompts for Eliciting Explanations

To elicit explanations from the model, we use a structured prompting approach. Each dataset is associated with a specific prompt designed to guide the model in generating explanations. Additionally, all prompts are preceded by a common system prompt:

> You are a helpful, pattern-following assistant. Use the following instructions to respond to user inputs. 1. Start your answer with a prefix that says "The right answer is: ". 2. Explain the response given in Step 1, with a prefix that says

[10] https://huggingface.co/google/gemma-2-9b-it
[11] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[12] https://huggingface.co/CohereForAI/c4ai-command-r-plus-08-2024
[13] https://openai.com/index/hello-gpt-4o/
[14] https://www.anthropic.com/news/3-5-models-and-computer-use

"Because: ". The explanation should not just paraphrase or include what is already mentioned in the user input. 3. Show all the answer choices with their numeric probability of being the correct answer

Below, we present the prompts used for each dataset.

C.3 Hellaswag Prompt

Each model was given 4 examples to guide its responses. For brevity, these examples are omitted from the prompt shown below.

```
## Examples

Please choose the most plausible ending
(event) for the given context. There is
only **one** correct answer. After
selecting a correct answer, explain why
you selected that option. The examples
do not include an explanation but you
will need to provide it when answering
the question.

For reference, we provide below four
examples that have already been solved
for you.

{% for example in examples %}
**Example {{loop.index}}**
{{example}}
{% endfor %}

## Exercise

Context: {ctx_a}

Question: Choose the option that best
completes the above story.

Options:
{% for ending in endings %}
{{ 'ABCDEFG'[loop.index0] }}) {{ctx_b}}
{{ ending }}
{% endfor %}
```

C.4 RACE Prompt

We provided 4 examples per query to improve model performance. The prompt format is shown below, excluding the examples for conciseness."

```
## Examples

In this task, you will be presented with
 a series of articles. Each is followed
by a question which relates to the
information provided in the text, and
four possible answers. Select only **one
** of these options as the correct
answer, and explain your choice.

For reference, we provide below four
examples that have already been solved
for you.

{% for example in examples %}
**Example {{loop.index}}**
{{example}}
{% endfor %}


# Exercise

Article: {article}

Question: {question}

Options:
{% for option in options %}
{{ 'ABCD'[loop.index0] }}) {{ option }}
{% endfor %}
```

C.5 W&I Prompt

Models received 3 examples as part of the prompt structure. The displayed prompt excludes these examples for clarity.

```
# Task

In this task, you will be presented with
 a series of essays. Annotate each of
these with exactly **one** of three
grades: A (beginner), B (intermediate),
C (advanced), and then explain your
choice.

For reference, we provide below three
examples that have already been solved
for you.

## Examples
{% for example in examples %}
**Example {{loop.index}}**
{{example}}
{% endfor %}

## Exercise

Essay: {{full_text}}

Question: If you were to assign a grade
to this essay, what would it be?

Options:

1. Beginner (grade A)
2. Intermediate (grade B)
3. Advanced (grade C)
```

C.6 Logic Prompt

Each model was given 6 examples to guide its responses. For brevity, these examples are omitted from the prompt shown below.

```
## Examples
Please identify the type of logical
fallacy. There is only **one** correct
answer. After selecting a correct answer
, explain why you selected that option.
```

```
For reference, we provide below seven
examples that have already been solved
for you.

{% for example in examples %}
**Example {{loop.index}}**
{{example}}
{% endfor %}


## Exercise

Statement: {source_article}

Question: Which type of logical fallacy
is this an example of?

Options:
A. Faulty generalisation
B. False causality
C. Circular claim
D. Appeal to emotion
E. Deductive fallacy
F. False dilemma
G. Fallacy of credibility
```

D Custom Agreement Metric

First metric. Cohen's κ (Cohen, 1960) and Krippendorff's α (Krippendorff, 2011) are among the most frequently used inter-rater reliability metrics. However, their direct application is best suited to nominal or categorical data. Even with adaptations like weighted kappa, these coefficients struggle to capture the full inter-relationship of hierarchical nested data. To bridge this gap, we introduced a custom metric that specifically accounts for the nested dependencies in CUBE. Our custom metric accounts for the *superlabels* (NONE, COMMENTARY, JUSTIFICATION, ARGUMENT) and *sublabels* (i.e., all DIMENSIONS) in Rubrik. In both cases, the metric penalizes discrepancies between ratings, with the penalty proportional to the difference in the hierarchical level. For example, consider the cases shown in Table 12 and Table 13.

Case	Rater 1	Rater 2	Diff.	Agree. (%)
1	COMMENTARY	JUSTIFICATION	1	67
2	COMMENTARY	ARGUMENT	2	50
3	NONE	ARGUMENT	3 to 4	0 to 25

Table 12: Superlabel agreement. NONE denotes the case where either of the COMMENTARY's COMPONENTS are missing, namely Action (1.a) and Reason (1.b).

From the *superlabel* point of view, there is a partial agreement in Case 1 since a JUSTIFICATION has the two components (ACTION and REASON) of a COMMENTARY + an additional one: EVIDENCE. Thus, the difference in the raters' judgement is 1. From the *sublabel* point of view, the agreement range is higher as it takes into consideration all the elements of a COMMENTARY (8: (2 COMPONENTS, 6 DIMENSIONS)) and a JUSTIFICATION (10: 3 COMPONENTS, 7 DIMENSIONS).

Case	Rater 1	Rater 2	Diff.	Agree. (%)
1	COMMENTARY	JUSTIFICATION	1-8 of 10	90-20
2	COMMENTARY	ARGUMENT	4-10 of 12	66-17
3	NONE	ARGUMENT	11-12 of 12	8-0

Table 13: Sublabel agreement. The difference (Diff.) column shows a range, taking both COMPONENTS and DIMENSIONS into consideration.

As explained in Section 3.3, a good COMMENTARY is the base of a good JUSTIFICATION. This means that Rater 2 judged with ✓ met all the elements of a COMMENTARY. The disagreement with Rater 1 comes from them judging with ✗ not met one or more of the six dimensions. The same logic applies to Cases 2 and 3.

Second metric. The first agreement metric accounts for partial agreement between LLMs and human annotators. We tested all LLMs as evaluators on the same subset judged by humans. However, we observe that LLMs often rate an explanation as JUSTIFICATION over the other options, compromising their ability to detect other types (see Table 15). This highlighted the need for an additional custom metric, which we designed based on a weighted F1 score to penalize over-centralization on a single label. The class weights are derived from both human evaluations and LLM evaluations from all six models. In our approach, we first calculate the distribution percentage of each superlabel in human evaluation p_i^{human} for label i. We then calculate the average distribution percentage of each superlabel across all 6 LLM evaluations denoted as p_i^{LLM}. These two percentages are combined as the class weight:

$$w_i = \lambda p_i^{human} + (1 - \lambda) p_i^{LLM}$$

where λ is a hyperparameter representing the relative importance of human evaluations vs. LLM evaluations. The derived class weights are then incorporated into the calculation of the weighted F1 score.

As shown in Table 14, our first metric points to `Command R+` as the model with higher agreement with human evaluators. However, a closer look at

Task	Agreement	Humans	Open models				Closed Models	
			Llama 3	Gemma 2	Command R+	Mixtral	GPT-4o	Sonnet 3.5
T1	Superlabel	0.814	0.693	0.799	0.797	**0.812**	0.794	**0.800**
	Sublabel	0.823	0.706	0.795	0.826	**0.829**	0.807	**0.811**
T2	Superlabel	0.910	0.832	0.862	**0.873**	0.869	0.878	**0.879**
	Sublabel	0.923	0.865	0.888	**0.903**	0.898	**0.902**	0.899
T3	Superlabel	0.830	0.830	0.838	0.843	**0.847**	0.844	**0.854**
	Sublabel	0.869	0.862	0.866	0.881	**0.887**	0.872	**0.881**
T4	Superlabel	0.887	0.797	**0.817**	0.810	0.774	**0.846**	0.833
	Sublabel	0.897	0.807	0.804	**0.853**	0.787	**0.860**	0.851
Overall	Superlabel	0.860	0.788	0.829	**0.831**	0.825	0.841	**0.842**
	Sublabel	0.878	0.810	0.838	**0.866**	0.850	**0.860**	**0.860**

Table 14: Overview of agreements scores, calculated with the first metric. In bold, the highest score by superlabel and sublabel, comparing the performance of open- vs. closed-source models.

Annotator	**None**	**Commentary**	**Justification**	**Argument**	**Second-metric-score**	**Second-metric-rank**
`Human_annotator 1`	0	293	406	221	-	-
`Human_annotator 2`	5	264	229	422	-	-
`LLama 3`	87	47	450	336	0.405	5
`Gemma 2`	**9**	**222**	**561**	**128**	**0.464**	**2**
`Command R+`	4	20	894	2	0.346	6
`Mixtral`	5	240	654	21	0.427	4
`GPT-4o`	**14**	**107**	**685**	**114**	**0.476**	**1**
`Sonnet 3.5`	5	126	742	47	0.444	3

Table 15: Aggregated label counts for each annotator and metric score. In bold are the results from the two best-ranked LLM evaluators. In both cases, there is a better balance in the judgement of explanation types.

the distribution of the explanation types assigned show that the high agreement is due to identifying an explanation as JUSTIFICATION nearly always. Our second metric penalizes this behaviour, ranking `Command R+` as the least effective evaluator.

E Rubric Evaluation Prompts

To evaluate explanations generated by the model, we use a structured prompting approach based on a rubric. Each dataset is associated with a specific prompt designed to guide the model in assessing explanations. Below is the prompt template that encodes the evaluation rubric.

```
{# Base template for rubric scoring #}
 # Explanation Judging Task

Your task is to evaluate a set of explanations in a given context. We
 define the context (**Task**, **Audience**, and **Purpose**) in the
following way:

**Task**: you will be shown a series of multiple-choice questions
relating to one of four tasks (commonsense reasoning, fallacy
detection, reading comprehension and essay scoring) in the following
format:
1. **Question**: The question being answered.
2. **Answer Choices**: The possible answer choices for that question.
3. **Correct Answer**: The correct answer to the question.
4. **User Answer**: The answer provided by the user.
5. **Explanation**: The explanation provided by the user to support
their answer.

**Audience**: you should assume that the audience of the explanations
 is adult, English-proficient, and provided in a formal academic
setting.

**Purpose**: the explanations should provide an understanding of why
a certain answer was chosen for a given multiple-choice question.

---

## Evaluation Criteria

For the given explanation, please answer the following questions with
 either **Yes** or **No**. Note that you **should not consider the
correctness of the user's answer** when evaluating the explanation.
Focus solely on the quality of the explanation according to the
criteria provided.

1. **Action**: Does the explanation clearly indicate the decision or
choice being made (e.g., specifying the selected answer)?
    - Answer **Yes** if it does. For example "The correct answer is A
    ."
    - Answer **No** if it does not. For example "Because it is the
    final part of the sequence."

2. **Reason**: Does the explanation provide reasoning or insight into
 why the decision or choice was made, explaining the underlying logic
 or rationale for the **Action**?
    - Answer **Yes** if it does. For example "The right answer is C
    because it is the final part of the sequence."
    - Answer **No** if it does not. For example "The correct answer
    is A."

3. **Grammaticality**: Is the explanation grammatically correct and
```

```
free of lexical or syntax errors? Small typos are acceptable, but the
 errors should not impede comprehension in any way.
    - Answer **Yes** if it is. For example "The correct answer is A
    because nowadays our society is based on consumerism and the way
    in which we are producing is contaminating the world."
    - Answer **No** if it is not. For example "The correct answer is
    A because now a day our socity it is bassed in consumer, so that
    become the word more contaminate to produce the products that we
    demanding."

4. **Word Choice**: Is the language used in the explanation tailored
to the given context (task, audience, purpose)? And are the sentences
 in the explanation well-formed?
    - Answer **Yes** if they are. For example "The correct answer is
    A because the essay lacks fluency. There are many incorrect
    clauses and missing words. And while the overall meaning can be
    deduced, the essay does not demonstrate an accurate grasp of
    language (e.g., frequent spelling and punctuation errors)."
    - Answer **No** if they are not. For example "Answer A. lack of
    fluency, incorrect clauses and missing words, meaning can be
    found but does not demonstrate an accurate grasp of language"

5. **Cohesion**: Does the explanation make appropriate use of
transition phrases (e.g., connectives like "because", "therefore", "
consequently", overlapping words across sentences, etc.)?
    - Answer **Yes** if it does. For example "The correct answer is C
     because the man is on roller blades, not on a skateboard.
    Further, he is not talking to anyone and therefore cannot
    possibly 'continue speaking.'"
    - Answer **No** if it does not. For example "The correct answer
    is C, because the man is on roller blades, not a skateboard, and
    is not talking to anyone in the example so cannot 'continue
    speaking'".

6. **Conciseness**: Is the explanation free of any redundant,
irrelevant, or excess sentences (that is, not required to understand
the answer)?
    - Answer **Yes** if it is. For example "The correct answer is D
    because it accurately reflects the sequence of events."
    - Answer **No** if it is not. For example, given that the option
    D was "next she explains how to use the lawnmower and other tools
     and then she cuts the grass", the following explanation is not
    concise: "The correct 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 the
    sequence of events."

7. **Appropriateness**: Is the explanation culturally appropriate,
matching expectations for the given context?
    - Answer **Yes** if it is. For example "The right answer is B
    because the tenses are properly used and the story makes sense."
    - Answer **No** if it is not. For example "The right answer is B
```

```
    because the tenses are properly used and (within the slightly odd
     context) the story makes sense."

8. **Coherence**: Does the explanation appropriately transition
between ideas? That is, does the explanation make sense as a whole (e
.g., good context-relatedness, semantic consistency, and inter-
sentence causal and temporal dependencies, etc.)?
    - Answer **Yes** if it does. For example "The correct answer is D
    , because no information about Liu's relationship to science
    subjects specifically is given in the passage, therefore the fact
     that they like chemistry is implied and ambiguous."
    - Answer **No** if it does not. For example "The correct answer
    is D, because no information about Liu's relationship to science
    subjects specifically is given in the passage, therefore the fact
     that they like cheese is implied and ambiguous."

9. **Evidence**: Does the explanation provide concrete evidence (can
be both explicit or implicit) that supports the reasoning, such as
information from the question's context or general knowledge?
    - Answer **Yes** if it does. For example "The right answer is C,
    because it finishes the sequence, describing the effect of
    bowling the ball and what happens as a result."
    - Answer **No** if it does not. For example "The right answer is
    C, because is is the final part of the sequence."

10. **Plausibility (of the evidence)**: Is the provided evidence
plausible and consistent with human reasoning, considering the
context and general world knowledge?
    - Answer **Yes** if it is. For example "The correct answer is A
    ('Jack picks the cheese') because we are told that he enjoys
    eating 'mozzarella' in the morning."
    - Answer **No** if it is not. For example "The correct answer is
    A ('Jack picks the cheese') because my name is also Jack and I
    personally love cheese for breakfast."

11. **Affective Appeals**: Does the explanation use vivid, or
emotionally charged language (e.g., metaphors) to evoke feelings in
the audience?
    - Answer **Yes** if it does. For example "The expression in the
    final section is very heartfelt; the tone is excitable and keen
    throughout."
    - Answer **No** if it does not. For example "The final section
    reflects the writer's strong feelings on this issue."

12. **Qualifiers**: Does the explanation make use of hedges, boosters
, attitude markers, self-mentions, or engagement markers to clarify
the writer's stance (i.e., the explainer's personal feelings towards
the task)? Note that the stance can be implicit unlike the **Action
**.
    - Answer **Yes** if it does. For example "The right answer is B,
    because the text is keeping with what is presumably a tour guide'
    s voice: intentionally using clunky and overly expressive words."
```

```
    - Answer **No** if it does not. For example ”The right answer is
    B, because the text is keeping with the original tour guide's
    voice.”

13. **Stance Clarity**: Is the explainer's stance (their personal
feelings towards the task) clearly and unambiguously conveyed through
 affective appeals or qualifiers? Note that the stance can be
implicit unlike the Action.
    - Answer **Yes** if it is. For example ”The correct answer is A (
    beginner) because this text is undeniably of a low English level
    .”
    - Answer **No** if it is not. For example ”The correct answer is
    A (beginner) because this text is clearly of a low English level
    although the final section is incredibly well written.”

---

## Expected Output

Your answers should be formatted as follows:

1. Action: **Yes** or **No**
2. Reason: **Yes** or **No**
3. Grammaticality: **Yes** or **No**
4. Word Choice: **Yes** or **No**
5. Cohesion: **Yes** or **No**
6. Conciseness: **Yes** or **No**
7. Appropriateness: **Yes** or **No**
8. Coherence: **Yes** or **No**
9. Evidence: **Yes** or **No**
10. Plausibility: **Yes** or **No**
11. Affective Appeals: **Yes** or **No**
12. Qualifiers: **Yes** or **No**
13. Stance Clarity: **Yes** or **No**

---

## Question

{% block question -%}

{{ task_question }}

{%- endblock -%}

## Answer Choices

{% block choices %}

{% for choice in choices %}
{{ 'ABCDEFG'[loop.index0] }}) {{ choice }}
{% endfor %}
```

```
{% endblock %}

## Correct Answer
{{correct_answer}}

## User Answer
{{user_answer}}

## Explanation
{{explanation}}
```

Dataset-Specific Evaluation Prompts

In the above template, the main difference between datasets is the format of the question and the options. Below, we show how each dataset-specific question and option block is customized.

E.1 Hellaswag

```
{% extends "rubric_prompt" %}

{% block question -%}

{{ ctx_a }}

{%- endblock %}

{% block choices %}

{% for ending in endings %}

{{ 'ABCDEFG'[loop.index0] }}) {{
ctx_b}} {{ ending }}

{% endfor %}

{% endblock %}
```

E.2 RACE

```
{% extends "rubric_prompt" %}

{% block question -%}

Article: {text}

Question: {question}

{%- endblock %}
```

E.3 WANDI

```
{% extends "rubric_prompt" %}

{% block question %}
Essay: {text}
{% endblock %}

{% block choices -%}

1. Beginner (grade A)
2. Intermediate (grade B)
3. Advanced (grade C)

{%- endblock %}
```

E.4 Logic

```
{% extends "rubric_prompt" %}

{% block question %}

Statement: {{text}}

Question: {{question}}

{% endblock %}

{%- block choices -%}

A. Faulty generalisation
B. False causality
C. Circular claim
D. Appeal to emotion
E. Deductive fallacy
F. False dilemma
G. Fallacy of credibility

{%- endblock -%}
```

F Detailed Analysis Results

This section delves deeper into the data, offering additional insights to complement the summary provided in Section 5.

F.1 Answer Frequencies

First, we report the frequencies of the answer choices picked by different groups of annotators during the annotation phase, and compare these to the actual distribution of correct answers in each task on the *annotation set* in Figure 5. Recall that we explicitly tried to get as uniform a distribution across the different answer choices as possible in the *annotation set* (as described in Appendix B).

Overall, we note that while human annotators sometimes refused to choose an answer between those provided ("None"), the LLMs almost never refused to answer. This may because LLMs have a tendency to overestimate their ability to answer questions (Zhang et al., 2023b).

In T1 and T3, the answer frequencies of all annotators seem fairly balanced, with the only notable difference being that human annotators also responded "None". In T2, however, we can see that the grouped Open LLMs (`Command R+`, `Mixtral`, `Llama 3` and `Gemma 2`) seem to significantly favour answers A, B and D at the expense of answers C and G, while the other groups of annotators remain relatively close to the actual frequency distribution. We should note that despite the fact that the *annotation set* is more or less balanced, in Jin et al. (2022) authors state that more than a single fallacy type may apply to a single instance. This may explain the variation observed. Specifically, they identified "common among incorrect but reasonable predictions" in their task, which "are debatable cases where multiple logical fallacy types seem to apply".

In T4, we notice a stark difference between humans and LLMs annotators. On one hand, LLMs almost never assign C (advanced) scores to essays, and overwhelmingly assign B (intermediate) scores around 65% of the time. While human annotators use the whole range of the scale, though still showing signs of a strong central tendency or severity by only assigning around half the actual proportion of advanced scores. Interestingly, experts annotators, that are professionally trained to assess the work of language learners, did not distinguish itself from the contractors we hired who have very similar frequency distributions the two language tasks. Overall, evaluators failed to identify advanced essays, focusing most of their attention on the middle of the rating scale. Essay scoring is a notoriously complex and subjective task (Brown, 2010), and we intentionally did not provide any scoring rubric to the annotators. They thus lacked a proper point of reference for the scale, which seems to be the source of the frustration reported by one annotator (see Section C.1.3).

F.2 Accuracy

Next, in Figure 6 we report the performance of the individual annotators and their groups, in each of the tasks, as well as their overall average performance across the four tasks.

Looking at the average performance across the four tasks, closed LLMs seem to perform the best, while open LLMs perform the worst, with humans (contractors and experts) performing just slightly better than the open models. The two closed models exhibited comparable average performance across the four tasks, but `Sonnet-3.5` is more consistently good across the four tasks, whereas `GPT-4o` is very good at reading comprehension (T3) and less good at essay scoring (T4).

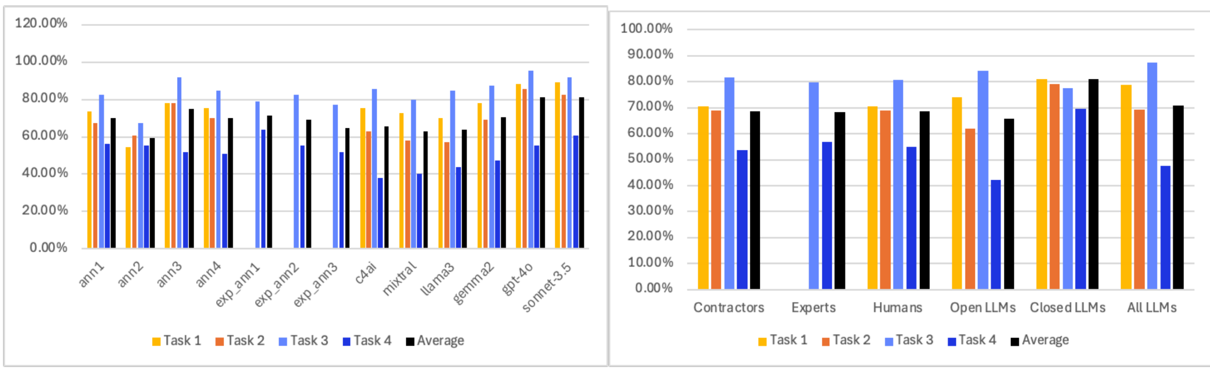
Overall, these graphs make it apparent that Essay Scoring (T4) was the hardest with an average accuracy of roughly 52% (across all annotators), while Reading Comprehension (T3) was by far the easiest with an average accuracy reaching almost 84%.

As in the previous section, we note that humans were overall quite consistent. The experts were ever so slightly better at essay scoring (T4) than the contractors, but this difference is very small. We had expected them to do much better due to being professionally trained to perform language assessment tasks. Further, while this background should have directly impacted their capacity to well in T4, we also expected them to do better than the contractors in T3 given the language-related nature of their day-to-day work. However, contractors were in fact ever so slightly better a reading comprehension (T3). These findings suggest that we do not always necessarily need to hire professionals, and that professional expertise can be matched by a rigorous selection process and sufficient training of annotators.

F.3 Sources of *bad* COMMENTARIES

Finally, we plot the sources of *bad* COMMENTARIES (in terms of DIMENSIONS) for each annotator group across each task in Figure 7 to supplement some of the results discussed in Section 5. The most prominent observation from this figure is the high frequency of CONCISENESS as the reason why an explanation generated by either of the LLMs is judged to be bad. This contrasts with the low frequency of WORD CHOICE, COHESION, APPROPRIATENESS and GRAMMATICALITY. On the other hand, CONCISENESS is less of a problem to humans, whose explanations are mostly judged as bad due to poor COHERENCE. In the particular case of experts, it is interesting to see their explanations are less grammatical than then contractors'.

Figure 5: Frequencies of the answers picked by the different groups of annotators during the annotation phase. We also show the **Actual** distribution of correct answers in black in the *annotation set*.



(a) (b)

Figure 6: Accuracy results of the different annotators in each of the tasks. On the left, 6(a) shows the individual annotator performance, and on the left, 6(b) shows the performance by group of annotators. We also include the **Average** accuracy across the four tasks of each annotator or group in black.

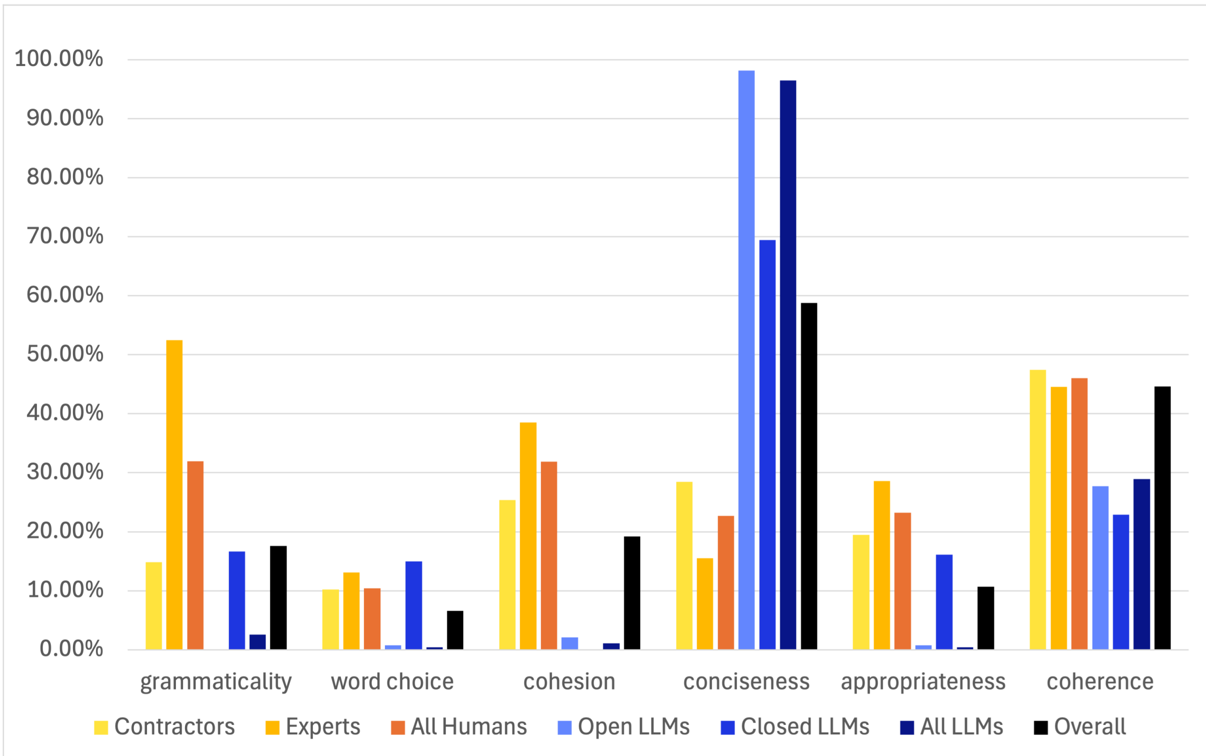


Figure 7: Plot showing the source of the *bad* COMMENTARIES (i.e., for which at least one of COMMENTARY's DIMENSIONS is missing) in the manually evaluated subset of the *annotation set*. We average the frequencies across all three evaluators (two humans and gpt-4o).