Using LLM Judgements for Sanity Checking Results and Reproducibility of Human Evaluations in NLP

Rudali Huidrom and Anya Belz
ADAPT Research Centre
Dublin City University
Ireland
{rudali.huidrom,anya.belz}@adaptcentre.ie

Abstract

Human-like evaluation by LLMs of NLP systems is currently attracting a lot of interest, and correlations with human reference evaluations are often remarkably strong. However, this is not always the case, for unclear reasons which means that without also meta-evaluating against human evaluations (incurring the very cost automatic evaluation is intended to avoid), we don't know if an LLM-as-judge evaluation is reliable or not. In this paper, we explore a type of evaluation scenario where this may not matter, because it comes with a built-in reliability check. We apply different LLM-as-judge methods to sets of three comparable human evaluations: (i) an original human evaluation, and (ii) two reproductions of it which produce contradicting reproducibility results. We find that in each case, the different LLM-as-judge methods (i) strongly agree with each other, and (ii) strongly agree with the results of one reproduction, while strongly disagreeing with the other. In combination, we take this to mean that a set of LLMs can be used to sanity check contradictory reproducibility results *if* the LLMs agree with each other, *and* the agreement of the LLMs with one set of results, and the disagreement with the other, are both strong.

1 Introduction

While considered a particularly reliable form of evaluation (van Miltenburg et al., 2023b), the cost and expertise required for human evaluation experiments prevent them from being used as standard in NLP. Large language models now exhibit astonishing performance across a wide range of different tasks including problem-solving and reasoning tasks (Mizrahi et al., 2024; Zhang et al., 2024). In combination with their ability to interpret and follow provided instructions, this makes them tempting, more cost-efficient alternatives to human evaluation, and they are beginning to be used in place of human evaluators in approaches commonly referred to as 'LLM-as-judge.' However, LLM judgments sometimes do, and sometimes do not, agree with comparable human judgements, for reasons that are not entirely clear. This means their reliability needs to be demonstrated anew for each new domain and/or task via meta-evaluation against human judgments, incurring the very cost their use is meant to obviate.

There are nevertheless situations where we may not have to worry about this, namely where we wish to arbitrate between multiple comparable human evaluations whose results contradict each other. Here it may be possible to use results from comparable LLM judgments to decide which of the contradictory human evaluation results are more likely to reflect the true picture. In this paper, we explore this question in the context of contradictory reproducibility results for human evaluation experiments, using reproductions and reproducibility results from the ReproNLP shared tasks (Belz and Thomson, 2023, 2024) as our data.

We start with a look at related research (Section 2), followed by an overview of our study (Section 3). We present the three sets of original studies and reproductions for them that constitute our data (Section 4), and the LLM-as-judge methods we use (Section 5). For each of the three scenarios we then present side-by-side evaluation results, and correlation matrices between the different evaluations (Section 6). We discuss the results (Section 7) and finish with concluding remarks (Section 8). Code and resources can be found on GitHub.[1]

2 Related Work

LLM-as-judge evaluation methods have been shown to correlate remarkably strongly with human evaluations across a range of task contexts (Liusie et al., 2024), including text summarisation assess-

[1] https://github.com/RHuidrom96/Repro_LLM_as_Judge.git

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 354–365

ment (e.g. G-Eval Liu et al., 2023), machine translation evaluation (e.g. GPTScore Fu et al., 2023), and code generation assessment (He et al., 2025), to name but a few.

A number of studies have investigated ways to improve the reliability of LLM-as-judge evaluations, including pairwise ranking, and using bigger, instruction-tuned models (e.g. GPT-4) (Gu et al., 2024); varying evaluation item order, and using majority voting (Lin et al., 2023); targeted prompt tuning (Tian et al., 2023); deterministic settings for hyperparameters like temperature, top-k, fixed random seed (Schroeder and Wood-Doughty, 2024; Atil et al., 2024); and conducting systematic sweeps over prompt templates and decoding settings to identify the most stable configuration (Wei et al., 2024).

Overall, while techniques like the above have improved alignment with human judgments, and correlations are therefore often high, it remains unclear why this is not always the case, so that strictly speaking meta-evaluation tests against human judgments must be carried out every time LLM-as-judge methods are to be used with a new LLM, task or domain.

To the best of our knowledge, applying LLM-as-judge evaluation for sanity-checking human evaluations has not so far been explored.

3 Background and Study Overview

Consider the following scenario. The ReproNLP shared tasks (Belz and Thomson, 2023, 2024) produced sets of two or more highly comparable human evaluations, one of which was the original study, and one or more were reproductions carried out by shared task participants with precisely aligned experimental details controlled by the organisers. When conducting quantified reproducibility assessment with QRA++ (Belz, 2025), the organisers found that in some cases, one of the (typically) two reproductions strongly *agreed* with the results from the original evaluation, while the other strongly *disagreed*. In such cases, the ReproNLP shared task organisers had no basis for deciding which of the two reproductions reflected the true picture: either the *agreeing* reproduction was right and the original study had excellent reproducibility, or the *disagreeing* reproduction was right and it had terrible reproducibility.

The overarching aim of the study we report in this paper is to examine how LLM-as-judge results behave in such scenarios, and whether they can provide a basis that was missing in the ReproNLP shared task for deciding between the two possibilities above.

Our starting point is three sets of comparable human evaluations from ReproNLP 2024, each consisting of (i) a set of human-produced system-level scores from the original study (O); and (ii) two sets of human-produced system-level scores from reproduction studies conducted by ReproNLP participants (R1 and R2).

For each set of comparable human evaluations O, R1, R2 we produce directly comparable LLM-as-judge results using different LLM ensembles J_*. We then compute Pearson's correlations between all pairs of sets of results and analyse them.

We start below with an overview of the three original studies and two reproductions each that form the basis of our investigation, in terms of the common data and evaluation criteria used in them, and the experiment-level QRA++ Type II and IV (Belz, 2025) reproducibility results reported in the ReproNLP results reports for them (Section 4). Next we describe the LLM-as-judge methods we used to compute the sanity checks, detailing the models and model combinations they comprise (Section 5). Finally, we present and discuss the side-by-side results and correlations between them (Section 6).

4 Original Studies and Reproductions

4.1 Atanasova et al., 2020; Gao et al., 2024; Loakman & Lin, 2024

Data: LIAR-PLUS (Alhindi et al., 2018) is dataset based on PolitiFact (Vo and Lee, 2020) containing 12,836 veracity statements along with justifications. Atanasova et al. (2020) used this dataset in the original study under consideration here, the human evaluation of which was reproduced during the ReproNLP'24 Shared Task (Belz and Thomson, 2024) by two teams (Gao et al., 2024; Loakman and Lin, 2024).

Note that while the raw responses from the original experiment are available, the script to calculate system-level scores is not, and the two teams above arrived at different scores for the original results when reimplementing it (Belz and Thomson, 2024). We also found slight differences when we reimplemented it. In order to be able to compare the reproduction results to the original results on an equal footing, we used the scores produced by our reim-

plementation for all evaluations in the Atanasova et al. scenario.

Evaluation criterion: Atanasova et al. (2020) used coverage, non-redundancy and non-contradiction as the evaluation criteria, of which the reproduction studies use ***Coverage*** only, where good coverage is defined as follows:

> The explanation includes important, salient information and does not omit any key points that contribute to the fact-check.

ReproNLP Type II and IV results: The table below from the ReproNLP 2024 results report shows Pearson's and Spearman's correlations (Type II reproducibility) in the third and fourth columns, with proportion of matching rankings (Type IV reproducibility) shown in the last column. As can be seen from the table, between O (the original study) and R1, strong correlations were found, and all findings were confirmed, but both measures were very poor between O and R2, and between R1 and R2.

Study A	Study B	r	ρ	Type IV
O	R1	0.99	1.00	3/3
O	R2	-0.43	-0.50	1/3
R1	R2	-0.31	-0.50	1/3

4.2 Feng et al., 2021; Fresen et al., 2024; Lango et al., 2024

Data: The AMI Meeting Corpus (Carletta et al., 2005) is a dataset of meeting summaries that contains roughly 100 hours of recorded meetings each featuring four participants discussing a remote control design project. Feng et al. (2021) used this dataset in the original study, the human evaluation of which was reproduced in ReproNLP'24 (Belz and Thomson, 2024) by two teams (Fresen et al., 2024; Lango et al., 2024). The human evaluation experiment involved summaries (abstracts) generated for 10 randomly selected dialogues.

Evaluation criterion: Feng et al. (2021) evaluate informativeness, conciseness and coverage, of which the reproduction studies address ***Informativeness***, defined as follows:

> Informativeness measures whether the abstract contains the key information from the original conversation.

ReproNLP Type II and IV results: The table below from the ReproNLP 2024 results report shows that strong correlations are seen between the original study (O) and R2. However, correlations between O and R1, and between R1 and R2, are close to 0 (no correlation). At the same time, nearly all findings from O were confirmed by R2, but only about half of the findings were confirmed between R2 on the one hand, and O and R2 on the other.

Study A	Study B	r	ρ	Type IV
O	R1	0.01	0.27	12/21
O	R2	0.99	0.85	18/21
R1	R2	-0.03	0.11	11/21

4.3 Puduppully & Lapata, 2021; Arvan & Parde, 2023; van Miltenburg et al., 2023a

Data: ROTOWIRE (Wiseman et al., 2017) is a widely used benchmark comprising basketball game statistics and textual summaries for them (∼5K items). Puduppully and Lapata (2021) conducted a human evaluation of 10 summarisation systems on 20 summaries (200 items). As part of ReproNLP'23 (Belz and Thomson, 2023), two reproductions (Arvan and Parde, 2023; van Miltenburg et al., 2023a) were carried out.

Evaluation criteria: Puduppully and Lapata (2021) evaluated grammaticality, coherence and conciseness/repetition. The reproduction studies address all three evaluation criteria, defined as follows:

> ***Grammaticality***: Is the summary written in well-formed English?
>
> ***Coherence***: Is the summary well structured and well organized and does it have a natural ordering of the facts?
>
> ***Conciseness/Repetition***: Does the summary avoid unnecessary repetition including whole sentences, facts or phrases?

ReproNLP Type II and IV results: As the table from the ReproNLP 2023 results report below shows, strong correlations were found, and all findings were confirmed, between O and R1. However, correlations were negative and only 1/3 of findings were confirmed both for O and R2, and for R1 and R2.

Grammaticality	Orig	Repro 1	Repro 2
Orig	1	0.975	-0.205
Repro 1	0.975	1	-0.100
Repro 2	-0.205	-0.100	1
Coherence	**Orig**	**Repro 1**	**Repro 2**
Orig	1	0.900	-0.100
Repro 1	0.900	1	-0.300
Repro 2	-0.100	-0.300	1
Conciseness	**Orig**	**Repro 1**	**Repro 2**
Orig	1	1	-0.051
Repro 1	1	1	-0.051
Repro 2	-0.051	-0.051	1

5 LLM-as-judge Methods

5.1 LLMs used

We use the following LLMs, on their own and/or in combination as LLM judges:

- C4AI Command R+[2] (Cohere, 2024): Cohere's open-weights research release of a 104B parameter model; a multilingual model evaluated in 10 languages for performance, and optimised for a variety of tasks including reasoning, summarisation, and question answering.
- Deepseek-Llama3-70B-Instruct[3] (DeepSeek-AI, 2025): One of the model distillations that was part of Deepseek's release of their first-generation reasoning models, based on a 70B-paramater Llama model and fine-tuned with comprehensive reasoning instructions.
- Granite-7B-Instruct[4] (Sudalairaj et al., 2024): IBM's Granite 7B model, instruction-tuned with curated human instructions and optimised for task-specific performance and in-context learning.
- Llama3-8B-Instruct[5] (Touvron et al., 2023): Meta's Llama 3 series model in the smaller 8B parameter size, pretrained, instruction-tuned, and optimised for dialogue-based applications.
- Llama3.3-70B-Instruct[6] (Grattafiori et al., 2024): Meta's Llama 3.3 series model in the 70B parameter size, an instruction-tuned text-only model optimised for multilingual dialogues.
- Mistral-7B-Instruct-v0.2[7] (Jiang et al., 2023): Fine-tuned from Mistral-7B-v0.2 using a diverse range of public conversation datasets, designed to follow instructions, generate creative text, and handle requests.
- Qwen2.5-7B-Instruct-1M[8] (Yang et al., 2025): Alibaba's Qwen series model in the smaller 7B parameter size, fine-tuned, instruction-tuned and optimised to handle long-context tasks while maintaining short-task capability.
- Qwen2-72B-Instruct[9] (Qwen, 2024): Alibaba's Qwen2 series model in 72B parameter size, fine-tuned, and instruction-tuned, supporting a long context length of up to 131,072 tokens.

5.2 LLM ensembles

Atanasova et al.

In the Atanasova et al. experiments, three items at a time were ranked by three human evaluators and the ranks aggregated into a single score via **mean average rank (MAR)**. For the LLMs, we obtain individual per-item rankings (measured as ranks 1, 2 or 3) with each of three LLMs, then compute the MAR of the three rankings. We used the following three model ensembles, each consisting of three models (to match the three human evaluators in Atanasova et al. and reproductions):

- $\mathbf{J_{C_S}}$: Small-model ensemble comprising Mistral-7B-Instruct-v0.2, Llama3-8B-Instruct, Qwen2.5-7B-Instruct-1M, all with either 7B or 8B parameters.
- $\mathbf{J_{C_L}}$: Medium-size model ensemble comprising Deepseek-Llama3-70B-Instruct, Llama3.3-70B-Instruct, Qwen2-72B-Instruct, all with either 70B or 72B parameters.
- $\mathbf{J_V}$: Mixed-size ensemble comprising C4AI Command R+1, Mistral-7B-Instruct-v0.2, and Llama3-8B-Instruct, i.e. two small models (7B, 8B), and one large one (C4AI, at 104B).

[2] https://huggingface.co/CohereForAI/c4ai-command-r-plus-4bit
[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B
[4] https://huggingface.co/ibm-granite/granite-7b-instruct
[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[6] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[8] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M
[9] https://huggingface.co/Qwen/Qwen2-72B-Instruct

Feng et al.

In the Feng et al. experiments, outputs are assessed for Coverage on a 1–5 scale; scores are **averaged**. We used the following nine model ensembles, each consisting of four models (to match the four human evaluators in Feng et al. and reproductions):

- $\mathbf{J_1}$: Granite-7B-Instruct, Mistral-7B-Instruct-v0.2, C4AI Command R+, Llama3.3-70B-Instruct.
- $\mathbf{J_2}$: Granite-7B-Instruct, Mistral-7B-Instruct-v0.2, C4AI Command R+, Qwen2-72B-Instruct.
- $\mathbf{J_3}$: Granite-7B-Instruct, Mistral-7B-Instruct-v0.2, Llama3.3-70B-Instruct, Qwen2-72B-Instruct.
- $\mathbf{J_4}$: Granite-7B-Instruct, Qwen2.5-7B-Instruct-1M, C4AI Command R+, Llama3.3-70B-Instruct.
- $\mathbf{J_5}$: Granite-7B-Instruct, Qwen2.5-7B-Instruct-1M, C4AI Command R+, Qwen2-72B-Instruct.
- $\mathbf{J_6}$: Granite-7B-Instruct, Qwen2.5-7B-Instruct-1M, Llama3.3-70B-Instruct, Qwen2-72B-Instruct.
- $\mathbf{J_7}$: Qwen2.5-7B-Instruct-1M, Mistral-7B-Instruct-v0.2, C4AI Command R+, Llama3.3-70B-Instruct.
- $\mathbf{J_8}$: Qwen2.5-7B-Instruct-1M, Mistral-7B-Instruct-v0.2, C4AI Command R+, Qwen2-72B-Instruct.
- $\mathbf{J_9}$: Qwen2.5-7B-Instruct-1M, Mistral-7B-Instruct-v0.2, Llama3.3-70B-Instruct, Qwen2-72B-Instruct.

Puduppully & Lapata

In the original evaluation, system summaries were evaluated by three human evaluators who were given pairs of systems to rank. **Best-worst scaling** was then applied to provide per-system scores ranging from -100 to $+100$. We obtain the same type of scores with our LLM ensembles, the three LLMs in each standing in for the three human evaluators in the original evaluation.

The model ensembles are two of the same ones as used for the Atanasova et al. experiments above:

- $\mathbf{J_V}$
- $\mathbf{J_{Cs}}$

5.3 Hyperparameters and prompts

We run the LLMs listed in Section 5.1 with the following hyperparameters: temperature = 0.001, maximum length = 1500, and top-p = 1. We quantise the models to 4-bit and run our experiments on a single rtxa6000/a100 GPU.

We recreate the original for-human evaluation interface as closely as possible, with no additional LLM-specific instructions, as text-only model prompts, inserting the evaluation items, and adding model-specific elements, as shown in more detail in the three example prompts in Appendix Section A. Each prompt produces either one score (Atanasov et al., Feng et al.), or three scores (Puduppully & Lapata).

We run each prompt with three different seeds (42; 1,738; 1,234), and compute the mean scores over the seeds. The resulting mean scores are then aggregated at system level for each model ensemble from the preceding section by computing either the mean average ranking (Atanasova et al.), the average (Feng et al.), or the best-worst scaling (Puduppully & Lapata).

In other words, each score in the tables below is one of the above system-level aggregations of the model-level scores themselves obtained by averaging over three seeds. All experiments use English-language data.

6 Results

In this section, we present two types of results for each of our three sets of evaluations above: (i) side-by-side system-level scores, and (ii) correlation matrices between the scores obtained in each set.

6.1 Atanasova et al. (2020) results

Table 1 presents the system-level MAR scores for *Coverage* on the LIAR-PLUS dataset for the original and reproduction studies for Atanasova et al. (2020), and the three LLM ensembles from Section 5.2. A lower MAR indicates a better average ranking. For each column, the best results are in bold. As can be seen from the table, the Just system obtains the best results in the original study O, reproduction study R1 and the LLM judgements, but not for R2 where the Explain-MT system is the best.

Table 2 reports the correlations (Pearson's r) between O, R1, R2 and the LLM ensembles. One set of reproduction results (R2) is in contradiction to all other sets of scores including the LLM ensem-

	Mean Average Rank ↓					
	O	R1	R2	J_V	J_{C_S}	J_{C_L}
Just	**1.46**	**1.58**	2.18	**1.83**	**1.83**	**1.78**
Explain-MT	1.71	1.83	**1.63**	1.84	2.02	1.89
Explain-Extr	1.88	2.03	1.93	1.97	2.08	2.14

Table 1: System-level MAR scores for Atanasova et al. / *Coverage* on LIAR-PLUS by the original study (O), two reproduction studies (R1 and R2), and the three LLM ensembles from Section 5.2. O, R1 and R2 scores as recalculated by us.

	O	R1	R2	J_V	J_{C_S}	J_{C_L}
O	1.00	1.00	-0.54	0.84	0.99	0.95
R1	1.00	1.00	-0.48	0.87	0.98	0.97
R2	-0.54	-0.48	1.00	0.01	-0.66	-0.25
J_V	0.84	0.87	0.01	1.00	0.75	0.97
J_{C_S}	0.99	0.98	-0.66	0.75	1.00	0.89
J_{C_L}	0.95	0.97	-0.25	0.97	0.89	1.00

Table 2: Pearson's r correlation matrix for Atanasova et al. / *Coverage* on LIAR-PLUS by the original study (O), two reproduction studies (R1 and R2), and the three LLM ensembles from Section 5.2.

bles. The latter, in contrast, all agree strongly with each other, indicating that R2 may not reflect the true picture: since it is either the case that R2 is right and all the others wrong, or that R2 is wrong and all the others right, it is far more likely that the latter is the case (see also Discussion section below).

One other aspect is worth noting: the mixed model sizes ensemble J_V agrees slightly less strongly with R_1, O and particularly with the small model ensemble J_{C_S} than those all agree with each other. At the same time, the small model ensemble agrees less well with the large model ensemble than with the others. This would seem to indicate that the large model ensemble gives the most reliable sanity check. Still, all models strongly point in the same direction.

6.2 Feng et al. (2021) results

Table 3 presents the system-level average scores for *Informativeness* on the AMI dataset from the original, reproduction and LLM ensemble evaluations for Feng et al. (2021). Participants were asked to rate the informativeness of system outputs (paragraph-sized summaries of multi-page meeting transcripts) on a scale of 1 (worst) to 5 (best). We see that the human-produced 'Golden' texts have the best average scores throughout. For each column, the best system results (second best overall after human) are in bold. We can see that R1 is the only evaluation that does not put the HMNet top of the systems.

Table 4 shows the correlations (Pearson's r) between all the human and LLM evaluations. Here too, we observe that one set of reproduction results (R1) is in contradiction with the original evaluation (O), with the other set of reproduction results (R2), and with all nine LLM ensemble results. Here the discrepancy is even clearer than for the Atanasova experiments above: R1 has r values around 0 with all other evaluations, indicating entirely random correlation, whereas agreement between other evaluations ranges from 0.89 to 0.99.

6.3 Puduppully and Lapata (2021) results

Table 5 presents the system-level average scores for *Coherence, Grammaticality*, and *Conciseness/Repetition* on the Rotowire dataset from the original, reproduction and LLM ensemble evaluations for Puduppully and Lapata (2021). For each column, the best results are in bold. We observe that the 'Gold' system has the highest best-worst scaled scores for all three criteria, in all evaluations except R2. The Template system has the worst scores for all criteria, again in all evaluations except R2. In fact, R2 has the Template system as the best.

Table 6 shows the complete Pearson's correlation matrix between the original, reproduction and LLM ensemble evaluations, for each of the three evaluation criteria. For Coherence and Repetition, the picture is pretty clear: all evaluations except R2 strongly agree with each other; R2 is medium strongly *negatively* correlated with all of the other evaluations.

For Grammaticality, the picture is similar, but less uniformly clear. This time, the R2 correlations are mixed, from random between R1 and R2, and

	Average ratings (1–5 scale) ↑											
	O	R1	R2	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
Golden	4.70	2.40	4.60	4.63	4.78	4.63	4.3	4.45	4.3	4.53	4.68	4.53
PGN	2.92	2.18	1.53	4.13	3.66	3.58	3.58	3.11	3.03	3.93	3.46	3.38
HMNet	**3.52**	2.20	**2.68**	**4.30**	**3.83**	**3.72**	**3.83**	**3.35**	**3.24**	**4.12**	**3.64**	**3.53**
PGN(DKE)	3.20	2.18	1.93	4.08	3.60	3.53	3.58	3.10	3.03	3.99	3.52	3.44
PGN(DRD)	3.15	**3.00**	1.90	4.22	3.72	3.64	3.69	3.19	3.12	3.93	3.43	3.36
PGN(DTS)	3.05	2.28	1.85	4.08	3.63	3.46	3.57	3.12	2.95	3.98	3.53	3.36
PGN(DALL)	3.33	2.53	1.85	4.01	3.58	3.35	3.43	3.00	2.77	3.87	3.44	3.21

Table 3: System-level aggregated scores for *Informativeness* on the AMI dataset, for Feng et al. O=original study, R1=reproduction 1, R2= reproduction 2; J_i=the nine LLM ensembles from Section 5.2.

	O	R1	R2	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
O	1.00	0.01	0.99	0.89	0.96	0.93	0.91	0.96	0.93	0.95	0.97	0.94
R1	0.01	1.00	-0.03	0.06	0.02	0.02	0.01	0	0	-0.15	-0.08	-0.09
R2	0.99	-0.03	1.00	0.94	0.97	0.96	0.96	0.98	0.96	0.98	0.98	0.97
J_1	0.89	0.06	0.94	1.00	0.96	0.98	0.99	0.96	0.97	0.95	0.91	0.94
J_2	0.96	0.02	0.97	0.96	1.00	0.99	0.97	1.00	0.99	0.97	0.99	0.99
J_3	0.93	0.02	0.96	0.98	0.99	1.00	0.98	0.99	1.00	0.97	0.97	0.99
J_4	0.91	0.01	0.96	0.99	0.97	0.98	1.00	0.97	0.99	0.97	0.94	0.96
J_5	0.96	0	0.98	0.96	1.00	0.99	0.97	1.00	0.99	0.98	0.99	0.99
J_6	0.93	0	0.96	0.97	0.99	1.00	0.99	0.99	1.00	0.98	0.97	0.99
J_7	0.95	-0.15	0.98	0.95	0.97	0.97	0.97	0.98	0.98	1.00	0.98	0.99
J_8	0.97	-0.08	0.98	0.91	0.99	0.97	0.94	0.99	0.97	0.98	1.00	0.99
J_9	0.94	-0.09	0.97	0.94	0.99	0.99	0.96	0.99	0.99	0.99	0.99	1.00

Table 4: Pearson's r correlation matrix for *Informativeness* on the AMI dataset, for Feng et al. J_5, J_6 vs. R1 rounded from -0.00158 and -0.00470, respectively. O=original study, R1=reproduction 1, R2= reproduction 2; J_i=the nine LLM ensembles from Section 5.2.

	Coherence					Conciseness/Repetition					Grammaticality				
	O	R1	R2	J_{CS}	J_V	O	R1	R2	J_{CS}	J_V	O	R1	R2	J_{CS}	J_V
Gold	**46.25**	**12.5**	-0.42	**40.00**	**40.00**	**30.83**	**5.83**	-1.67	**47.50**	**41.67**	**38.33**	**14.17**	9.17	**29.17**	**41.67**
Templ	-52.92	-20.00	**25.42**	-50.83	-62.50	-36.67	-5.83	**43.75**	-47.50	-54.17	–61.67	-23.33	**17.08**	-15.83	-35.83
ED+CC	-8.33	-7.50	-15.00	-16.67	-15.83	-4.58	-5.00	-25.83	-16.67	-11.67	5.00	-8.33	-19.58	-19.17	-25.00
Hier	4.58	9.17	-10.42	13.33	20.83	3.75	0.83	-14.58	3.33	12.50	13.33	9.17	-9.58	-1.67	5.83
Macro	10.42	5.83	0.42	14.17	17.50	6.67	4.17	-1.67	13.33	11.67	5.00	8.33	2.92	7.50	13.33

Table 5: System-level best-worst scaled scores for *Coherence, Conciseness/Repetition* and *Grammaticality* on the Rotowire dataset, for Puduppully & Lapata. O=original study, R1=reproduction 1, R2= reproduction 2; J_*=the two LLM ensembles from Section 5.2.

R2 and J_V, to the medium strong *positive* correlation between R2 and O.

7 Discussion

We have looked at three scenarios where we had one original human evaluation and two contradicting reproductions of the original evaluation, one strongly agreeing with it, the other strongly disagreeing. In this situation, we would not normally have a way of telling whether (i) the reproduction that agrees with the original evaluation is right and the original evaluation has terrible reproducibility, or (ii) the reproduction that disagrees with the original evaluation is right and the latter has excellent reproducibility.

For each of these three scenarios, we tested multiple LLM ensembles as stand-in replacements for the human evaluators, and found that in all three scenarios, they not only *all* strongly agreed with each other, but also with the original evaluation and *one* of the reproductions. That the LLMs agree with each other may not come as a surprise, but that they also strongly agree with one set of human evaluation while strongly disagreeing with the other, is more so.

This pattern held true for all twelve different LLM ensembles we tested, whether they consisted of all small LLMs, all medium-sized LLMs, or a combination of both. In one scenario (Atanasova et al.), the small-LLMs ensemble J_{CS} agreed slightly less well with two of the other evaluations (R1,

	O	R1	R2	J_{C_S}	J_V
Coherence					
O	1.000	0.930	-0.572	0.980	0.964
R1	0.930	1.000	-0.584	0.982	0.992
R2	-0.572	-0.584	1.000	-0.547	-0.625
J_{C_S}	0.980	0.982	-0.547	1.000	0.993
J_V	0.964	0.992	-0.625	0.993	1.000
Grammaticality					
O	1.000	0.912	-0.420	0.695	0.831
R1	0.912	1.000	-0.185	0.814	0.931
R2	-0.420	-0.185	1.000	0.358	0.133
J_{C_S}	0.695	0.814	0.358	1.000	0.969
J_V	0.831	0.931	0.133	0.969	1.000
Conciseness/Repetition					
O	1.000	0.871	-0.622	0.984	0.991
R1	0.871	1.000	-0.277	0.935	0.898
R2	-0.622	-0.277	1.000	-0.482	-0.619
J_{C_S}	0.984	0.935	-0.482	1.000	0.981
J_V	0.991	0.898	-0.619	0.981	1.000

Table 6: Pearson's correlation matrix for *Coherence, Conciseness/Repetition* and *Grammaticality* on Rotowire, for Puduppully and Lapata (2021). O = original study, R1 = reproduction 1, R2 = reproduction 2; J_* = the two LLM ensembles from Section 5.2. For grammaticality, O vs. R1, R2 rounded off from 0.6641 and 0.6597, respectively.

J_{C_V}) than the other agreeing evaluations, but in the other scenario we tested it in (Puduppully & Lapata), J_{C_S} and J_{C_V} had a correlation of $r = 0.99$.

Interestingly, we saw different kinds of disagreement. In the *Feng et al.* scenario, correlation coefficients were all very close to 0 indicating an entirely **random** relationship between the disagreeing evaluation and the others. In contrast, in the case of *Coherence and Repetition in Puduppully & Lapata*, we saw pronounced negative correlation scores throughout, indicating an **inverse** relationship between the disagreeing evaluation and the rest. Finally, for the *Atanasova et al.* evaluations, and *Grammaticality in Puduppully & Lapata*, we see a mix of **random and inverse** relationships.

All of which begs the question what this can tell us about the disagreeing evaluations? Is there necessarily something wrong with them? In directly comparable human evaluations, the main difference will tend to be the sample of evaluators performing the assessments. Clearly, different samples (from the population of all evaluators) will results that differ to different degrees, with a small proportion deviating substantially from true population-level result. The greater the deviation, the smaller the likelihood of it occurring, but it is possible that the disagreeing evaluations we have seen in this paper are due to rare sampling effects, whereas the LLMs are able to produce assessments closer to the population level, because trained on very large (in effect population-level) samples of text.

The nature of the disagreement discussed above can provide more information. If we are dealing with a rare sampling effect, we would not expect to see near perfect random correlations (as in R1 in the Feng et al. scenario) with multiple other evaluations. In this scenario therefore, it may be supposed that something has gone wrong, perhaps a coding error at some point in the pipeline from collecting the evaluator assessments to aggregating the results at system-level which resulted in the association between evaluation items and scores being lost.

In the case of the negative correlations seen consistently with other evaluations in the Coherence and Repetition evaluations in the Puduppully & Lapata scenario, another explanation is needed. Here, the relationship is not random; there is a pronounced association, but it is in the wrong direction. Here it is possible that at some point in the analysis carried out in the R2 evaluation, the signs of the evaluation scores inadvertently became inverted, perhaps as a result of a bug in the best-worst scaling.

This leaves just the mixed random and negative correlations seen in the Grammaticality evaluation in the Puduppully & Lapata scenario. Given the negative correlations seen consistently for the other two evaluation criteria (Coherence and Repetition), we would expect to see the same for Grammaticality given that the evaluator sample was the same. The fact that we see a mix of random and mild to medium positive associations makes this picture very hard to interpret. Note however that correlations between the other evaluations (both human and LLM-based) are also considerably weaker and more mixed than in any of our other scenarios, perhaps indicating that the Grammaticality evaluation task itself was somehow harder to perform consistently.

8 Conclusion

In this paper, we have examined the behaviour of LLM-as-judge methods in situations where they are used to obtain additional evaluation results to add to a set of comparable human evaluation studies of which at least two strongly disagree with each other. We have seen that in such scenarios,

all twelve LLM ensembles we tested invariably strongly agreed with one of the disagreeing human evaluations, and strongly disagreed with the other, providing evidence that the one they all agree with is the more reliable.

Drawing out the commonalities from the three different scenarios we examined (corresponding to five different evaluation experiments, each with one evaluation criterion), we conclude that LLMs can be used as sanity checkers to validate human evaluations in scenarios where:

1. There are two or more directly comparable human evaluations of which at least two strongly disagree with each other;
2. Multiple LLMs of different types, or ensembles of such LLMs, are used to produce multiple different evaluations directly comparable to the human evaluations; and
3. Correlation analysis shows that all (ensembles of) LLMs strongly agree with each other and one of the disagreeing evaluations, while strongly disagreeing with the other.

Even in the case of single human evaluations, running multiple LLM-as-judge methods in parallel could provide additional confirmation of results, provided the methods involve a variety of different types of LLMs, and they all agree with each other and with the (single) human evaluation.

All in all, using LLMs as sanity checkers for human evaluations would seem to be one application of the 'LLM-a-judge' paradigm where the built-in reliability check against human evaluations means results means they can be relied on without the need for independent validation by meta-evaluation for every new domain and/or dataset.

Limitations

The experiments conducted showed promising alignment between human and LLM evaluations. However, we only looked into a limited set of models and tasks, therefore we can't make claims beyond those.

Ethics Statement

As a paper that meta-evaluates existing human evaluation tasks using the same and custom instructions, the risk associated with this study was minimal.

Acknowledgments

Huidrom's work is supported by the Faculty of Engineering and Computing, DCU, via a PhD studentship. Both authors benefit from being members of the SFI Ireland funded ADAPT Research Centre.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90.

Mohammad Arvan and Natalie Parde. 2023. Human evaluation reproduction report for data-to-text generation with macro planning. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 89–96, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. LLM stability: A detailed analysis with some surprises. *arXiv preprint arXiv:2408.04667*.

Anya Belz. 2025. QRA++: Quantified reproducibility assessment for common types of results in natural language processing. *Preprint*, arXiv:2505.17043.

Anya Belz and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Cohere. 2024. Introducing command r+: A scalable LLM built for business. *https://cohere.com/blog/command-r-plus-microsoft-azure*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. *arXiv preprint arXiv:2105.12544*.

Vivian Fresen, Mei-Shin Wu-Urbanek, and Steffen Eger. 2024. Reprohum# 0043: Human evaluation reproducing language model as an annotator: Exploring dialogue summarization on AMI dataset. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 199–209.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Mingqi Gao, Jie Ruan, and Xiaojun Wan. 2024. Reprohum# 0087-01: A reproduction study of the human evaluation of the coverage of fact checking explanations. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 269–273.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Junda He, Jieke Shi, Terry Yue Zhuo, Christoph Treude, Jiamou Sun, Zhenchang Xing, Xiaoning Du, and David Lo. 2025. From code to courtroom: LLMs as the new software judges. *arXiv preprint arXiv:2503.02246*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mateusz Lango, Patricia Schmidtova, Simone Balloccu, and Ondrej Dusek. 2024. ReproHum #0043-4: Evaluating summarization models: investigating the impact of education and language proficiency on reproducibility. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 229–237, Torino, Italia. ELRA and ICCL.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. 2024. Efficient LLM comparative assessment: a product of experts framework for pairwise comparisons. *arXiv preprint arXiv:2405.05894*.

Tyler Loakman and Chenghua Lin. 2024. Reprohum# 0087-01: Human evaluation reproduction report for generating fact checking explanations. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 255–260.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Qwen. 2024. Qwen2 technical report.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust LLM judgments? Reliability of LLM-as-a-judge. *arXiv preprint arXiv:2412.12509*.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D Cox, and Akash Srivastava. 2024. Lab: Large-scale alignment for chatbots. *arXiv preprint arXiv:2403.01081*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023a. How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning'. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 75–88, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023b.

How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning'. *Human Evaluation of NLP Systems*, page 75.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint arXiv:2408.13006*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. LLM as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

A Example Prompts

The following shows an example prompt from the Pudupully & Lapata scenario, as used on the Command R+ model (other models will have had slightly different model-specific elements):

```
## Instructions
Summaries

System Summaries

## Input
A: The Portland Trail Blazers (2-2) defeated the
Minnesota Timberwolves (2-1) 106-101. Damian
Lillard scored 34 points (14-25 FG, 4-9 3PT, 2-3 FT)
to go with 2 rebounds. Kevin Martin scored 24 points
(7-12 FG, 2-4 3PT, 8-11 FT) to go with 2 rebounds.
CJ McCollum scored 18 points (7-18 FG, 2-6 3PT,
2-2 FT) to go with 6 rebounds. Al-Farouq Aminu
scored 17 points (7-12 FG, 2-5 3PT, 1-2 FT) to go
with 9 rebounds. Andrew Wiggins scored 16 points
(5-17 FG, 0-3 3PT, 6-7 FT) to go with 6 rebounds.
Gorgui Dieng scored 12 points (6-9 FG, 0-0 3PT,
0-1 FT) to go with 5 rebounds. The Trail Blazers'
next game will be at home versus the Mavericks,
while the Timberwolves travel to play the Bulls.

B: The Portland Trail Blazers (2-2) defeated the
Timberwolves (2-1) 106-101 on Wednesday at the
Target Center in Minnesota. The Blazers got off
to a quick start, out-scoring Minnesota 34-21 in
Q1. They shot 46% from the field and 30% from deep,
while the Wolves shot 43% and 23%. Lillard and C.J.
McCollum led the way. Lillard went 14-25 and 4-9 to
score 34 points, with seven assists and two steals.
It was his second straight 10-rebound game; he's
now averaging 16 points and 7 boards. McCollum went
7-18 and 2-6 to score 18 points, adding six rebounds.

The Blazers' next game is on the road against the
Denver Nuggets on Wednesday; the Timberwolves will
travel to Houston to play the Rockets on Wednesday.

## Criterion
Ranking Criteria
Coherence: How coherent is the summary? How natural
is the ordering of the facts? The summary should be
well structured, well organized, and follow a
natural fact ordering.


## Output
Answers
Best:
Worst:
Analysis


Output:
Best:
Worst:
```

Example prompt from Atanasova et al. scenario

The following shows an example prompt from the Atanasova et al. scenario, as used on the Command R+ model (other models will have had slightly different model-specific elements):

```
## Input
Claim: Says Bill and Hillary Clinton attended Donald
Trumps last wedding.
Label: True
Justification 1: Curbelo said Bill and Hillary
Clinton were at Donald Trump's last wedding. Bill
Clinton only made the reception, but Hillary Clinton
did have a seat in the first row at the church in
2005. Both rubbed elbows with the stars at the
reception.
Justification 2: The short answer is, yes, the
Clintons did attend Trump\u2019s 2005 wedding to
Melania Knauss. \"That\u2019s part of the problem
with the system. They were at his last wedding.
He has contributed to the Clintons' foundation.
Justification 3: (PunditFact has found to be the
case.) The short answer is, yes, the Clintons did
attend Trump\u2019s 2005 wedding to Melania Knauss.
If I say go to my wedding, they go to my wedding.
It was the then-58-year-old Trump\u2019s third
wedding.
## Output

Coverage rank for Justification 1:
Coverage rank for Justification 2:
Coverage rank for Justification 3:
```

Example prompt from Feng et al. scenario

The following shows an example prompt from the Feng et al. scenario, as used on the Command R+ model (other models will have had slightly different model-specific elements). For presentation purposes here in the paper, we have truncated the (very long) meeting transcript, as indicated by [...]; the summary is given in full:

```
## Input\nMeeting 2
B : it's up there ? \n B : that screen's black . \n B : are we done ? \n B : , this is our second meeting and might be bit all over the place . \n B : our agenda for today , do you want us to give you second ? \n D : no that's , . \n B : i'll go over what we decided last meeting , , we decided upon universal control , one handset for all , t_v_ , video equipment . \n B : that it was important that the product was accessible to wide range of consumers , wide age range , not limiting anyone . \n B : we decided it was important to reflect the company's image in our product , we put fashion in electronics , that thing . \n B : our budget would have to affect try not to reflect our budget , that we might have bit of you can see it , . \n B : dissonance between what our budget was and what we want it to look like . \n B : want it to look uncluttered , undaunting to the customer . \n B : we discussed flip-open design , reducing the size of the control and an electronic panel for further features like programming , things like that . \n B : three presentations , i've got written here so shall we hear from marketing first ? \n D : is it if postpone that til later , want to get access to little bit more information , is that ? \n B : no that's fine , that's fine . \n C : i'll go first . \n C : can grab the . \n C : what do have to press ? \n B : f_n_ function eight . \n C : there we go . \n C : this is the working design , presented by me , the industrial designer extraordinaire . \n C : this is where went bit mad with powerpoint so . \n C : what the first thing question asked was what are we trying to design ? \n C : device which just sends the signal to the t_v_ to change its state , whether that be the power , or the channel or the volume , everything is just some signal to change the state of the t_v_ or other appliance that it's sending the signal to . \n C : so decided i'd have look at what other people have designed and try and take some inspiration from that .  [...]

Summary:
The Industrial Designer gave his presentation on the basic functions of the remote. He presented the basic components that remotes share and suggested that smaller batteries be considered in the product design. The User Interface Designer presented his ideas for making the remote easy-to-use; he discussed using a simple design and hiding complicated features from the main interface. The Marketing Expert presented the findings from a lab study on user requirements for a remote control device, and discussed users' demand for a simple interface and advanced technology. The Project Manager presented the new requirements that the remote not include a teletext function, that it be used only to control television, and that it include the company image in its design. The group narrowed down their target marketing group to the youth market. They discussed the functions the remote will have, including Video Plus capability and rechargeable batteries. A customer service plan was suggested to make the remote seem more user-friendly, but it was decided that helpful manuals were more within the budget. The group then discussed the shell-like shape of the remote and including several different casing options to buyers. The Marketing Expert will research consumers' opinions on instruction manuals. It was decided that the group will produce one product design instead of creating alternate designs in an attempt to accomodate different users' preferences. The marketing will be focused towards a young, business-class buyer. The remote will feature Video Plus capabilities and a seashell-like shape to accomodate the LCD display and the flip screen. The remote will be bundled with a docking station to recharge the remote's batteries and a user-friendly instruction manual, and multiple casings will be made available. The limitations of the budget will restrict the development of some features; several of the features that the group wanted to include may have to be made simpler to decrease cost.
## Output
Informativeness:
```