LLMs as Judges for Domain-Specific Text: Evidence from Drilling Reports

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models are now judged by other models in many workflows. This scales, but it is risky in domains where facts, numbers, and terminology matter. We study this in an industrial data-to-text setting: short, structured reports generated from time-series sensor data. The task is Daily Drilling Report (DDR) sentence generation, but the lessons apply to any domain-grounded pipeline.

We evaluate LLMs used as judges under three protocols: a minimal single score, a weighted multi-criteria score, and a multi-criteria scheme with external aggregation. We compare model sizes and prompt designs using agreement metrics with human experts. Larger judges improve consistency, yet prompt and aggregation choices still cause large shifts in reliability and calibration. Smaller judges fail to track numeric and terminology constraints even with structure.

The takeaways are practical. Good evaluation needs domain knowledge in the rubric, transparent aggregation, and stress tests that expose failure modes of LLM-as-judge. Our study offers a blueprint for building such evaluations in data-to-text applications and a caution against treating general-purpose judges as drop-in replacements for expert assessment.

1 Introduction

Large language models are used in settings where outputs must be fluent and technically correct. Yet evaluation often relies on surface-overlap metrics such as BLEU, ROUGE, or METEOR, which miss factual and domain errors [Papineni et al., 2002, Lin, 2004, Banerjee and Lavie, 2005, Reiter, 2018, Maynez et al., 2020, Kryscinski et al., 2020]. This gap grows in specialized applications where numeric fidelity and terminology matter.

Daily Drilling Reports (DDRs) are concise, structured summaries of operations that include activities, depths, and other technical parameters. Industry standards capture DDR content in machine-readable form [Energistics, 2016]. In our setting, DDR sentences are generated automatically from raw sensor streams. The evaluation problem is general. A model must turn time-series data into short text that preserves facts and numbers, uses correct terms, and tolerates variation in wording.

A recent trend is to use LLMs as judges. Strong models can approximate human preferences on open-ended tasks, but known biases and sensitivity to instruction format limit reliability [Zheng et al., 2023, Shi et al., 2024]. In technical domains, these issues risk scoring outputs that sound plausible while misstating operations or values.

We study LLM-as-judge for domain-grounded text generation. We compare three protocols: a minimal single-score prompt, a weighted multi-criteria rubric that encodes domain checks, and a multi-criteria variant where the model emits per-criterion decisions and we aggregate externally. We

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

evaluate multiple model sizes and report agreement with expert ratings using correlation, error, and rater-consistency metrics.

Our findings point to two consistent patterns. Model scale improves alignment with expert ratings, but protocol design is equally decisive. Structured rubrics reduce ambiguity and enhance calibration for strong judges, while the same structure can amplify noise when applied to weaker models. Although our study focuses on DDR text, the framework extends naturally to other data-to-text and task-oriented applications where factual accuracy outweighs surface similarity.

2 Evaluation Protocol

We evaluate LLM judges for domain-specific text generation by directly comparing their scores with human expert ratings. Each instance includes a reference DDR entry, a model-generated prediction, and an expert rating. LLM judges receive the reference and prediction through structured prompts and return an integer score with a brief justification, which we parse and store for inspection. To measure agreement, we normalize all scores to $[0, 1]$ and compute complementary metrics: Pearson and Spearman correlations for linear and rank alignment, Mean Squared Error (MSE) and Mean Absolute Error (MAE) for deviation magnitude, and Intraclass Correlation Coefficient (ICC(2,1)) with Concordance Correlation Coefficient (CCC) for reliability and concordance. This combination captures correlation, ranking, error, and inter-rater consistency, providing a more complete picture of judge quality than correlations alone and highlighting where LLM judges diverge from human experts.

3 Results with a Minimal Scoring Prompt

Table 1: Agreement metrics between human ratings and LLM judges under the minimal prompt. Grouped by family and ordered by size. Best per column in **bold**.

Model	ICC(2,1)	CCC	Pearson r	Spearman ρ	MSE	MAE	Pairwise Pref. Acc.
Llama-3.2-1B	0.2085	0.2083	0.2542	0.2681	0.1641	0.3559	0.4761
Llama-3.2-3B	0.5391	0.5388	0.5502	0.5697	0.1261	0.2762	0.6424
Llama-3-8B	0.7126	0.7123	0.7173	0.7196	0.0840	0.2075	0.7056
Llama-3-70B	0.7515	0.7513	0.7755	**0.7832**	0.0686	0.1970	0.6688
Qwen2.5-72B	**0.7746**	**0.7744**	**0.7822**	0.7637	**0.0653**	**0.1845**	0.7301
Mistral-7B	0.7082	0.7079	0.7374	0.7386	0.0744	0.2157	**0.7429**
Phi-3.5-mini	0.7176	0.7173	0.7427	0.7480	0.0754	0.2142	0.7169

We first experimented with the simplest possible LLM-as-judge setup: a prompt that only asks for an integer score between 0 and 10, accompanied by a one-line justification. The full instruction is shown in Appendix A. Larger models generally achieve higher agreement with human ratings under the minimal prompt. Performance improves steadily from 1B to 8B, with further gains at 70B. Qwen-72B reaches the best overall reliability and lowest error, while Llama-70B shows competitive correlations. Small models (1B–3B) fail to align with experts, highlighting that scale alone provides robustness even with a simple scoring setup.

4 Results with a Weighted Multi-Criteria Prompt

We next evaluated LLM judges with a structured prompt that encodes domain knowledge and weights eight criteria following annotators' recommendations. Instead of a single similarity score, the model compares reference and prediction sentences on **primary operation** (50%), **depth** (12.5%), **conciseness** (6.25%), **all operations**, **other parameters**, **hole size**, **BHA type**, and **other details** (each 6.25%). The score is an integer from 0–10 with a one-line justification. The full prompt is in Appendix B. This weighted setup reduces ambiguity and enforces domain-consistent evaluation.

Compared to the minimal prompt, the weighted multi-criteria setup amplifies differences between small and large judges. Small models (1B–3B) remain weak, showing no real benefit from structure. Mid-size models such as Llama-8B gain slightly, but noise in per-criterion decisions still limits

Table 2: Agreement metrics under the weighted multi-criteria prompt. Grouped by family and ordered by size. Best per column in **bold**.

Model	ICC(2,1)	CCC	Pearson r	Spearman ρ	MSE	MAE	Pairwise Pref. Acc.
Llama-3.2-1B	0.2085	0.2083	0.2542	0.2681	0.1641	0.3559	0.4761
Llama-3.2-3B	0.5124	0.5121	0.5214	0.5890	0.1412	0.2793	0.6531
Llama-3-8B	0.7183	0.7180	0.7236	0.7187	0.0818	0.2017	0.7269
Llama-3-70B	0.7532	0.7530	0.7845	0.7817	0.0802	**0.1873**	0.6910
Qwen2.5-72B	**0.7780**	**0.7778**	**0.8062**	**0.7820**	**0.0573**	0.1887	**0.7723**
Mistral-7B	0.6872	0.6869	0.6944	0.6819	0.088986	0.203146	0.700201
Phi-3.5-mini	0.5983	0.5981	0.7014	0.7457	0.0980	0.2522	0.7225

reliability. The main improvements appear at scale: Llama-70B and Qwen-72B both maintain or increase agreement, with Qwen-72B reaching the best overall correlation and lowest error. The structured weighting helps large judges because they can consistently separate primary operation, depth, and parameter checks. In contrast, smaller judges misfire on these criteria, and the fixed weighting locks in their errors, explaining why their scores do not improve over the minimal prompt.

5 Results with an Externally Aggregated Multi-Criteria Protocol

The multi-criteria setup extends the weighted prompt of Section 4. Instead of asking the model to combine all criteria into a single score, we extract binary judgments (0/1) for each of the eight dimensions and then apply the weighting scheme manually. This separation avoids relying on the model's internal aggregation, which we observed to be inconsistent in the previous section: final scores sometimes contradicted the instructed rubric. Manual aggregation ensures weights are applied as intended and provides interpretable traces per criterion, clarifying where disagreements with human ratings originate.

Table 3: Agreement metrics under the multi-criteria protocol with external aggregation. Grouped by family and ordered by size. Best per column in **bold**.

Model	ICC(2,1)	CCC	Pearson r	Spearman ρ	MSE	MAE	Pairwise Pref. Acc.
Llama-3.2-1B	0.0776	0.0775	0.0819	0.0857	0.2498	0.4017	0.4491
Llama-3.2-3B	0.6120	0.6117	0.6139	0.6283	0.1174	0.2334	0.7145
Llama-3-8B	0.5437	0.5434	0.5786	0.6092	0.1343	0.2584	
Llama-3-70B	0.7689	0.7687	0.7806	0.7415	0.0730	0.1737	0.6810
Qwen2.5-72B	**0.8640**	**0.8638**	**0.8663**	**0.8182**	**0.0422**	**0.1173**	**0.7701**
Mistral-7B	0.6872	0.6869	0.6944	0.6819	0.0890	0.2031	0.7002
Phi-3.5-mini	0.5345	0.5342	0.6005	0.6768	0.1449	0.2605	0.6655

Multi-criteria evaluation makes the contrasts between small and large judges clear, and the supporting tables show why. Table 4 highlights weak per-criterion correlations for Llama-8B, especially on depth and conciseness, while Llama-70B and Qwen-72B are consistently stronger. Table 5 shows how Llama-8B mis-weights the rubric, under-weighting depth and conciseness and over-weighting "all operations," sometimes with negative signs, whereas larger models keep weights aligned with the rubric. These errors feed into calibration: Table 6 shows Llama-8B overscoring humans under MC, while Qwen-72B stays close to neutral. Error breakdowns confirm the pattern. Table 7 shows that Llama-8B fails most in low-score regions, while larger models reduce error there. Table 8 reveals that MC raises error for Llama-8B whenever depth or parameters are present, but reduces error for Llama-70B and Qwen-72B in the same cases. Finally, Table 9 shows learned importance: large judges weight core checks (primary op, depth) in line with humans, while Llama-8B spreads weight across unstable criteria. Taken together, the diagnostics explain why MC amplifies noise in smaller models like Llama-8B but reinforces reliable criterion-level judgments in larger ones like Llama-70B and Qwen-72B.

Table 4: Per-criterion Pearson correlation with human scores (binary output vs. human rating).

Criterion	L-8B	L-70B	Q-70B
A Primary op.	0.483	0.704	0.809
B Depth	0.289	0.621	0.643
C Concise	0.062	0.047	0.136
D All ops.	0.602	0.643	0.722
E Params	0.463	0.638	0.634
F Hole, dia.	0.408	0.617	0.563
G BHA type	0.452	0.611	0.539
H Other	0.244	0.618	0.648

Table 5: Rubric weights vs. weights learned from MC outputs (OLS coefficients on $[0, 1]$ human score).

Criterion	Rubric	L-8B	L-70B	Q-70B
A Primary op.	0.5000	0.192	0.344	0.458
B Depth	0.1250	−0.046	0.137	0.143
C Concise	0.0625	−0.052	0.025	0.063
D All ops.	0.0625	0.347	0.187	0.106
E Params	0.0625	0.108	0.023	0.056
F Hole, dia.	0.0625	0.013	0.100	−0.002
G BHA type	0.0625	0.106	0.062	0.019
H Other	0.0625	−0.005	−0.028	0.067

Table 6: Calibration: mean judge score minus mean human score on $[0, 1]$. Negative means under-scoring.

Model + Prompt	Mean Hum.	Mean LLM Judge	Bias
L-8B W-MC	0.4506	0.3999	−0.0508
L-8B MC	0.4506	0.5726	**+0.1220**
L-70B W-MC	0.4506	0.3379	−0.1128
L-70B MC	0.4506	0.3877	−0.0630
Q-70B W-MC	0.4506	0.4535	+0.0029
Q-70B MC	0.4506	0.4286	−0.0220

Table 7: MAE by human-score quintile (0 = lowest human scores).

	Q0	Q1	Q2	Q3	Q4
Llama-8B W-MC	0.162	0.187	0.246	0.359	0.145
Llama-8B MC	0.346	0.327	0.284	0.231	0.041
Llama-70B W-MC	0.075	0.161	0.359	0.345	0.074
Llama-70B MC	0.101	0.096	0.255	0.306	0.036
Qwen-70B W-MC	0.206	0.171	0.197	0.258	0.070
Qwen-70B MC	0.099	0.060	0.207	0.173	0.016

Table 8: Presence-based error change: ΔMAE = MC minus W-MC. Positive is worse. n = number of samples where group is present.

Group	n	L8B	L70B	Q70B
Depth present	599	+0.039	−0.023	−0.072
Parameters present	363	+0.038	−0.002	−0.074
Other details present	499	+0.057	−0.017	−0.082
Any operation present	776	+0.038	−0.027	−0.073
BHA present	316	+0.007	−0.047	−0.059
Hole size present	6	−0.100	−0.090	+0.010

Table 9: Per-criterion learned weights (normalized importance).

Criterion	L8B	L70B	Q70B
A	0.609	0.435	0.464
B	0.628	0.304	0.325
C	0.776	0.924	0.895
D	0.269	0.199	0.263
E	0.485	0.260	0.304
F	0.509	0.281	0.373
G	0.278	0.213	0.355
H	0.720	0.239	0.309

6 Conclusion

This study examined the reliability of LLMs as judges for domain-specific text generation in drilling operations. Across minimal, weighted, and multi-criteria prompts, results show that evaluation quality depends strongly on both model scale and protocol design. Larger models such as Qwen-72B and Llama-70B achieve the highest agreement with human experts, while smaller models fail to provide stable judgments regardless of prompting. Structured protocols reduce ambiguity and improve calibration for strong judges, but can amplify noise in weaker ones.

These findings highlight two key lessons. First, scale is necessary but not sufficient: prompt design and aggregation strategy materially affect alignment with expert ratings. Second, domain-aware evaluation remains essential, as general-purpose judges still struggle with numeric fidelity and specialized terminology. Future work should develop benchmarks that embed domain knowledge more directly and refine structured evaluation protocols. Reliable evaluation is central to deploying text generation systems in industrial contexts, and progress requires both larger models and more domain-grounded approaches.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0909/`.

Energistics. Witsml v2.0 documentation: 11.1 drill report data object. `https://docs.energistics.org/WITSML/WITSML_TOPICS/WITSML-000-146-0-C-sv2000.html`, 2016. Accessed 2025-09-03.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346, Online, 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.emnlp-main.750.pdf`.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018. doi: 10.1162/coli_a_00322. URL https://aclanthology.org/J18-3002/.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024. URL https://arxiv.org/abs/2406.07791.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

A Appendix - Minimal Prompt

"You have excellent knowledge of oil and gas drilling. You are also an impartial and expert judge. Your task is to compare a 'Reference' sentence with a 'Prediction' sentence. Give a score as an integer between 0 and 10 both included and a reason for the score. Higher score means more similarity.

OUTPUT FORMAT - Your entire response MUST BE ONLY two lines of text. - The first line must start with 'score:' followed by the integer score. - The second line must start with 'reason:' followed by your concise explanation. - Do NOT add any other text."

B Appendix - Weighted Multi-Criteria Prompt

```
You know oil and gas drilling well. You are an impartial, expert judge.
Compare a Reference sentence with a Prediction sentence.
Score the Prediction against the Reference using the 8 criteria and their
weights.
Return an integer score from 0 to 10 and a short reason. Higher means more
similar.

Criterion A: Primary operation match (Weight 50%)
- Check if the main operation in Prediction matches the Reference
semantically. Examples: drilling, RIH/tripping in, POOH/tripping out,
circulation/pumping, hole cleaning, casing, liner, upper completion, lower
completion, flow check, pressure testing, wash pipe change.
- Match by meaning, not exact words.
- Count logical sub-steps or preparatory steps as matches.
- Do not give 0 unless you are sure there is no match.

Criterion B: Depth match (Weight 12.5%)
- Match depth or depth range within tolerance.
- Tolerance: 100 ft for drilling operations. 1000 ft for others (tripping,
casing, liner, completion, reaming, stuck pipe, etc.).
- Skip if the Reference has no depth.

Criterion C: Conciseness, no repetition (Weight 6.25%)
- Penalize unnecessary repetition.

Criterion D: All operations match (Weight 6.25%)
- All operations in Prediction appear in Reference and vice versa.

Criterion E: Other parameters match (Weight 6.25%)
- Check flow rate (gpm or bph), torque (k ft-lbf), hook load (kips), volume
(bbl), RPM, ECD or mud weight (ppg), etc.
- Tolerance is plus or minus 10%.
- Skip if neither sentence has such parameters.

Criterion F: Hole size and pipe diameter match (Weight 6.25%)
- All hole sizes and pipe diameters match within plus or minus 10%.
- Skip if neither sentence has these.

Criterion G: BHA type match (Weight 6.25%)
- If a specific BHA type is mentioned in the Reference, check that
Prediction matches it.
- Skip if none are mentioned.

Criterion H: Other details match (Weight 6.25%)
- Match details not covered above, such as tight hole, restrictions,
overpull, overslack, losses, and similar domain items.
- Skip if none are mentioned.

Output format
- Exactly two lines.
- Line 1: score: <integer 0-10>
```

```
- Line 2: reason: <concise explanation>
- No extra text.
```