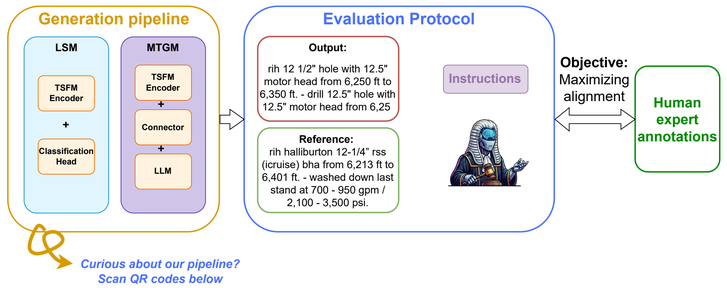
LLMs as Judges for Domain-Specific Text: Evidence from Drilling Reports

Abdallah Benzine, Soumyadipta Sengupta, Sebastiaan Buiting, Imane Khaouja, Yahia Salaheldin Shaaban, Amine El Khair
AIQ

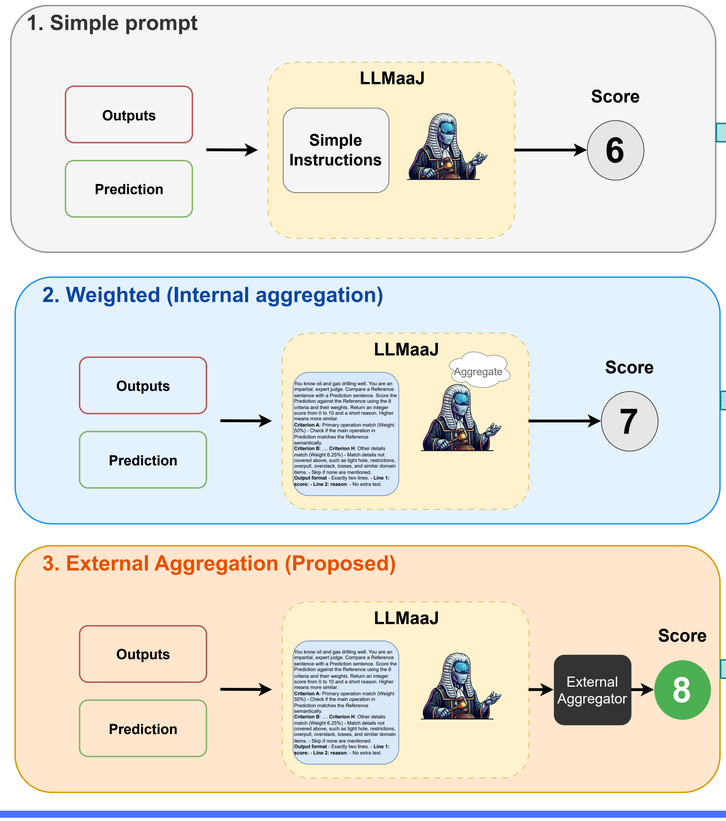
Research Question:

Can LLMs as Judges (LLMaaJ) maximize alignment with human experts when evaluating a domain-specific text generation pipeline?

The evaluated task - Drilling Reports Generation:

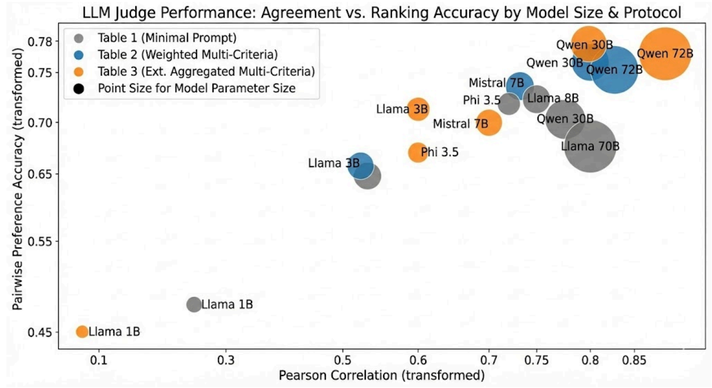


Evaluation Protocols:



The Scale Effect:

Instructions distilled from expert annotators demonstrate superior scaling, outperforming minimal prompts by a substantial margin as model capacity increases.



Agreement Metrics:

Results

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
Qwen3-30B-A3B	0.7007	0.7004	0.7698	**0.7833**	0.0738	0.2181	0.7001
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

Results

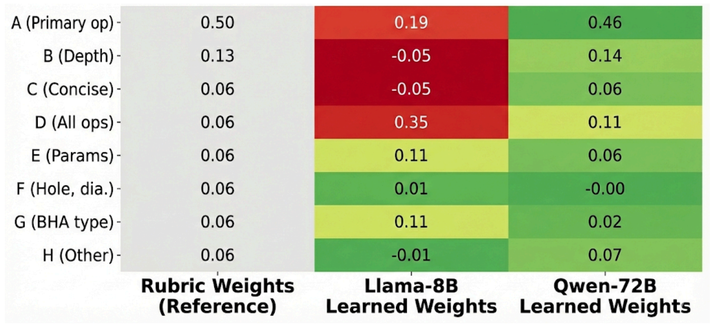
Model	ICC(2,1)	CCC	Pearson r	Spearman ρ	MSE	MAE	Pairwise Pref. Acc.
Llama-3.2-1B	0.2085	0.2083	0.2542	0.2681	0.1641	0.3559	0.4761
Llama-3.2-3B	0.5124	0.5121	0.5214	0.5890	0.1412	0.2793	0.6531
Llama-3-8B	0.7183	0.7180	0.7236	0.7187	0.0818	0.2017	0.7269
Llama-3-70B	0.7532	0.7530	0.7845	0.7817	0.0802	**0.1873**	0.6910
Qwen3-30B-A3B	0.7688	0.7686	0.7884	0.7751	0.0620	0.1868	0.7621
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

Results

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
Qwen3-30B-A3B	0.7967	0.7966	0.8126	0.7678	0.0624	0.1512	**0.7776**
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

Analysis: Why Small Models Fail?

Small models **lack** the **reasoning capacity** to enforce strict numeric and terminological constraints, leading them to misapply rubric weights.
As a result, **structured prompts amplify noise** by 'locking in' criterion-level errors rather than aligning with expert judgment.



Conclusion:

- **TL;DR: In domain-specific evaluation, Scale is necessary, but Protocol Structure is decisive.**
  - **Scale is the Prerequisite:** Only larger models (70B+) effectively track complex domain constraints.
  - **Structure Encodes Expertise:** Complex, structured prompts are superior for encoding expert priors in LLM-as-a-Judge (LLMaaJ) workflows.
  - **Deterministic Aggregation:** Offload mathematical aggregation to external tools rather than relying on the LLM's internal calculation.

Future Work:

. **Can a Council of Small Agents outperform a Single Giant?** Investigating if multi-agent systems of small, specialized models can achieve superior zero-shot judging performance compared to monolithic LLMs.

. **Optimizing the Judge:** How can Reinforcement Learning be applied to fine-tune judge policies and instructions for tighter alignment with expert human baselines?

Scan QR code for the full paper.

LiveDrill LLMaaJ