Exploring Capability Thresholds in Ultra-Lightweight LLM Judges for Nugget-Based Report Evaluation

Mann Bajpai* **Pulkit Chatwal** **Priyanshu Deswal**
Santosh Kumar Mishra **Harish Pratap Singh**

Rajiv Gandhi Institute of Petroleum Technology
{21cs2011,21cs2027,21cs2014,
santosh.mishra,harishps23cs}@rgipt.ac.in

Abstract

Reliable automatic evaluation of retrieval-grounded long-form reports typically requires human annotation or frontier-scale proprietary LLMs, both of which are expensive in constrained settings. Team **rgipt** participated in RAG4Reports@ACL 2026 Task 1 with a zero-shot nugget-verification system that runs entirely on a single NVIDIA T4 GPU. We compare three ultra-lightweight decoder-only models:—Qwen2-0.5B, Qwen2-1.5B, and Qwen2.5-0.5B, under identical inference conditions to examine how small an LLM judge can be while retaining human-aligned ranking signal. Both Qwen2 models produced negative τ_{gap}, whereas Qwen2.5-0.5B achieved $\tau_{\text{gap}} = 0.0772$ and Pearson $r = 0.2209$, ranking 13th of 21 teams. Within this family and evaluation setting, model generation appears to matter more than parameter count, although this finding is based on three configurations on a single task and warrants further validation.

1 Introduction

The RAG4Reports shared task evaluates retrieval-grounded report generation and report evaluation under realistic long-form information-seeking settings (Yang et al., 2025). We participated in Task 1: Automatic Report Evaluation, where systems must produce report-level and system-level rankings that correlate with human-derived rankings. Our submission is semi-automatic: it uses organizer-provided human-curated nuggets as evaluation targets.

The central question of our system paper is operational rather than purely architectural: how small can an LLM evaluator be while still preserving useful human-aligned ranking structure? Existing RAG and NLG evaluation pipelines increasingly rely on LLM judges, including RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), G-Eval (Liu et al., 2023), and Auto-ARGUE (Walden et al., 2025). However, these approaches often assume access to substantially larger models than those available in constrained academic or shared-task settings.

We therefore frame RAG4Reports Task 1 as zero-shot nugget verification with ultra-lightweight instruction-tuned decoder-only transformers. Given a report and an acceptable nugget answer, the evaluator predicts whether the nugget is semantically supported by the report. Nugget recall is then aggregated into report-level and system-level rankings. This design makes the evaluator simple, reproducible, and deployable on a single Kaggle NVIDIA T4 GPU.

Our main empirical finding is that evaluator quality does not scale monotonically with parameter count in this regime. Despite being smaller than Qwen2-1.5B, Qwen2.5-0.5B was the only evaluated model to cross from negative to positive τ_{gap}. This result suggests a possible capability threshold for latent semantic alignment in ultra-small LLM judges and highlights generational model improvements as an important factor in constrained report evaluation.

2 Related Work

RAG and report evaluation. Retrieval-Augmented Generation combines parametric generation with external non-parametric evidence to improve factual, knowledge-intensive generation (Lewis et al., 2020). The RAG4Reports shared task extends this paradigm to long-form, citation-backed report generation and evaluation (Yang et al., 2025), while Auto-ARGUE provides the official framework for report-level assessment through LLM judgments over report evidence and nugget support (Walden et al., 2025).

*Primary author and corresponding author.

Proceedings of the First Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports), pages 94–98

Automatic RAG evaluation. RAGAS evaluates RAG systems along dimensions such as faithfulness, answer relevance, and context use without exhaustive human references (Es et al., 2024). ARES similarly targets scalable RAG evaluation using lightweight judges trained with synthetic data and limited human annotation (Saad-Falcon et al., 2024). Our system instead focuses on semi-automatic report ranking under a strict single-GPU constraint.

LLM-as-a-judge and reasoning prompts. LLM-as-a-judge methods are widely used for scalable evaluation of open-ended generation (Zheng et al., 2023). G-Eval showed that Chain-of-Thought prompting and structured scoring improve agreement with human judgments (Liu et al., 2023), while Chain-of-Thought more generally encourages intermediate reasoning before final decisions (Wei et al., 2022). We adopt this reasoning-first paradigm for binary nugget verification using ultra-lightweight models.

Classical and learned text-generation metrics. Earlier evaluation relied on lexical-overlap metrics such as BLEU and ROUGE (Papineni et al., 2002; Lin, 2004). Learned metrics such as BERTScore and BLEURT improved sensitivity to semantic similarity and human preference (Zhang et al., 2020; Sellam et al., 2020). FActScore further motivates fine-grained atomic evaluation for long-form factuality (Min et al., 2023). Our nugget-verification formulation follows this logic for retrieval-grounded reports.

Compact decoder-only models. The Qwen2 and Qwen2.5 technical reports describe families of instruction-tuned decoder-only transformers spanning compact and large scales (Yang et al., 2024; Qwen Team et al., 2025). Our ablation tests whether Qwen2.5 improves zero-shot semantic verification relative to Qwen2 under identical inference constraints.

3 System Description

3.1 Task Formulation

For each topic t, report r, and nugget $n \in N_t$, the model predicts a binary support label $y_{t,r,n} \in \{0, 1\}$. Report-level scores are computed as average nugget recall across all nuggets associated with a topic:

$$S(t, r) = \frac{1}{|N_t|} \sum_{n \in N_t} y_{t,r,n}$$

System scores are obtained by averaging report scores across topics and are then converted into system-level rankings for submission.

This formulation deliberately avoids learned calibration, supervised training, or additional retrieval. The evaluator is therefore a zero-shot semantic verifier over organizer-provided evidence targets.

3.2 Prompting Strategy

Ultra-small LLMs are brittle under direct score generation, especially when long reports express nuggets through paraphrase, implication, or distributed evidence. We therefore use a reasoning-first Chain-of-Thought prompt inspired by G-Eval (Liu et al., 2023). The model must first produce a short explanation and then emit a JSON score. The explanation is not used directly in scoring; it acts as an intermediate latent alignment step before classification.

System: Precise JSON-formatting evaluation assistant.
`User: Given a question, acceptable nugget answer, and generated report, decide whether the report explicitly or semantically supports the nugget. Count paraphrases as support, but reject vague topical overlap or unsupported inference. Output only JSON: {"reasoning": "...", "score": 1.0}.`

Table 1: Reasoning-first nugget verification prompt.

All prompts are formatted with HuggingFace `apply_chat_template` to preserve the models' instruction-tuning conventions.

3.3 Inference Pipeline

All experiments ran on a single Kaggle NVIDIA T4 GPU with 15GB VRAM. We evaluated `Qwen2-0.5B`, `Qwen2-1.5B`, and `Qwen2.5-0.5B`. Each model was loaded natively in `torch.float16`; we avoided 4-bit quantization to prevent additional approximation error in an already low-capacity evaluation regime.

The pipeline used deterministic greedy decoding, left-padded inputs, dynamic batched inference with batch size 4, and a regex fallback parser for malformed JSON. The fallback extracted binary scores using the

pattern `"score"\\s*:\\s*([01]\\.?0?)`. The evaluation set contained 57 generation systems and 68 topics, yielding 3,876 report-topic pairs before nugget expansion.

4 Results and Analysis

4.1 Official Task 1 Results

The official RAG4Reports Task 1 metric measures correlation between submitted rankings and human-derived rankings through the Auto-ARGUE evaluation framework (Walden et al., 2025). Table 2 reports the official scores for our submissions.

Model	τ_{gap}	Kendall	Pearson
`Qwen2.5-0.5B`	**0.0772**	0.1330	0.2209
`Qwen2-1.5B`	-0.1245	**0.1542**	**0.2513**
`Qwen2-0.5B`	-0.2693	-0.2967	0.1852

Table 2: Official RAG4Reports Task 1 results.

`Qwen2.5-0.5B` was our best model under the primary ranking metric, achieving $\tau_{\text{gap}} = 0.0772$, Pearson correlation of 0.2209, and rank 13 of 21 participating teams. The result is modest in absolute magnitude but important operationally: it shows that a 0.5B model can preserve some human-aligned ranking signal when paired with curated nuggets and a constrained verification pipeline.

The secondary metrics reveal a calibration tension. `Qwen2-1.5B` achieved higher Kendall and Pearson values but negative τ_{gap}, indicating that global association alone did not translate into the official gap-sensitive ranking objective. This discrepancy suggests that the older 1.5B model captured some coarse score variance while still misordering systems in the regions most consequential for the leaderboard metric.

4.2 Impact of Model Generation vs. Scale

Figure 1 is the core ablation. Both Qwen2 models fall below zero on τ_{gap}, meaning their induced rankings are inversely aligned with the official human-derived ordering under the primary metric. In contrast, `Qwen2.5-0.5B` crosses into positive correlation despite using fewer parameters than `Qwen2-1.5B`. This pattern is consistent with the hypothesis that model generation, instruction tuning, and latent semantic alignment may matter more than raw parameter count at the ultra-lightweight boundary, although this observation is based on only three models within the Qwen family.

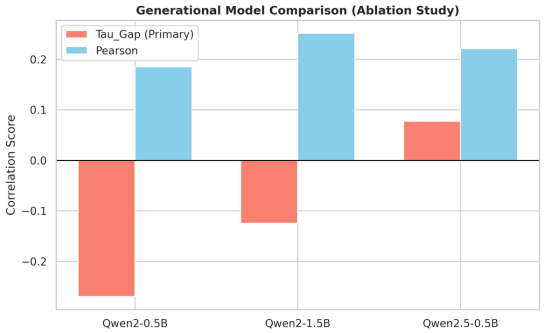

Figure 1: Generational ablation across Qwen variants.

The implication is practical: for constrained evaluator deployment, selecting the most recent compact instruction-tuned model may be more important than selecting the largest model that fits memory. In our setting, added Qwen2 capacity did not compensate for weaker zero-shot nugget verification behavior.

4.3 Score Distribution and Leniency

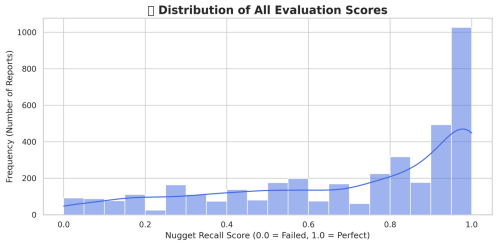

Figure 2: Score distribution for `Qwen2.5-0.5B`.

Figure 2 shows a bimodal score distribution dominated by positive classifications. This behavior indicates that `Qwen2.5-0.5B` often treats partial semantic overlap as sufficient nugget support. Such leniency is beneficial when reports express evidence through semantic paraphrase, but harmful when topical proximity is incorrectly treated as factual entailment.

This positive bias explains why the model can recover a weak ranking signal while remaining poorly calibrated at the nugget level. The evaluator is better interpreted as a high-recall semantic detector than as a strict factual verifier. For future systems, this suggests that lightweight judges need explicit contradiction handling, abstention, or threshold calibration to reduce false-positive nugget matches.

4.4 Ranking Consistency Across Systems

Figure 3 analyzes score dispersion for top-ranked systems. Narrower interquartile ranges indicate

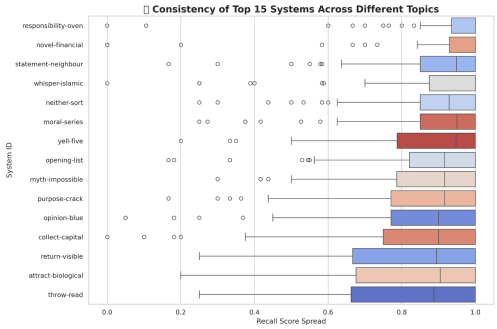

Figure 3: Nugget-score consistency among top-ranked systems.

that strong systems tend to receive stable nugget recall across topics, rather than relying on isolated high-scoring reports. This matters for report evaluation because system-level rankings should reward consistent retrieval-grounded coverage rather than topic-specific overperformance.

At the same time, residual variance among strong systems shows that even the best report generators are uneven across topic types. For the evaluator, this variance creates a difficult ranking problem: small calibration errors at the nugget level can shift aggregate system orderings. The figure therefore supports using consistency analysis alongside headline correlation metrics.

4.5 Topic Difficulty

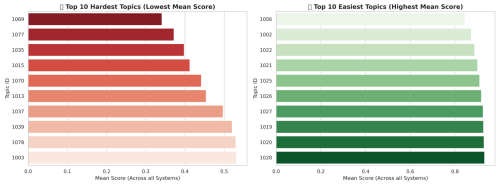

Figure 4: Hardest and easiest topics under nugget verification.

Figure 4 shows substantial topic-level difficulty variation. Easier topics appear to contain nuggets that are recoverable through explicit lexical overlap or localized factual statements. Harder topics require semantic abstraction, multi-paragraph evidence aggregation, or compositional reasoning over dispersed report content.

This result highlights the main failure mode of ultra-small evaluators: they can detect direct factual restatements but struggle when nugget satisfaction requires cross-context inference. This explains why a model can achieve positive aggregate ranking correlation while still failing on semantically demanding topics. Lightweight evaluation is therefore most effective when accompanied by topic-level robustness analysis.

5 Limitations

Our evaluator remains weakly calibrated, exhibiting positive classification bias, limited negation sensitivity, and difficulty with compositional reasoning. Because we use deterministic zero-shot inference with greedy decoding and no supervised calibration, the system cannot learn topic-specific thresholds or reliably distinguish partial from complete nugget support. The approach also depends on organizer-provided nuggets, limiting applicability when high-quality human-curated evaluation targets are unavailable. All experiments were conducted under a single-GPU constraint, so we do not claim competitiveness with frontier-scale proprietary evaluators. Consequently, our conclusion regarding a possible capability threshold is preliminary and should not be interpreted as evidence of a general scaling law.

6 Conclusion

We presented Team **rgipt**'s semi-automatic evaluator for RAG4Reports@ACL 2026 Task 1. Our system reduces long-form report evaluation to zero-shot nugget verification with ultra-lightweight decoder-only transformers and executes entirely on a single NVIDIA T4 GPU.

The central finding is evidence consistent with a possible capability threshold within the evaluated Qwen family: `Qwen2.5-0.5B` was the only tested model to produce positive τ_{gap}, while both Qwen2 baselines remained negative. These findings suggest that evaluator capability in ultra-lightweight LLMs may emerge discontinuously across architectural generations rather than scaling smoothly with parameter count alone.

Use of Generative AI Tools

The authors used generative AI tools solely for language editing and paraphrasing. All research ideas, experiments, analyses, and conclusions were developed and verified by the authors, who take full responsibility for the paper.

References

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. Association for Computational Linguistics.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, and 1 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

William Walden, Marc Mason, Orion Weller, Laura Dietz, John Conroy, Neil Molino, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, Dawn Lawrie, James Mayfield, and Eugene Yang. 2025. Auto-ARGUE: LLM-based report generation evaluation. *Preprint*, arXiv:2509.26184.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, and 1 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Eugene Yang and 1 others. 2025. RAG for Reports (RAGTIME) at TREC 2025. https://rag4reports.github.io/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

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. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Preprint*, arXiv:2306.05685.