Interpretable, Expert-Aligned Composite Metric with Domain-Aware Calibration for Evaluating Natural Language Generation

*Allan C. Taracatac
College of Computing Studies, Information and Communication Technology, Isabela State University, Philippines
allan.taracatac_cyn@isu.edu.ph

Arnel C. Fajardo
College of Computing Studies, Information and Communication Technology, Isabela State University, Philippines

Abstract

Automated metrics for natural language generation (NLG) often show weak or unstable alignment with expert judgment in domain-specific settings that require interpretability and tunability. Hence, this study designs and validates an interpretable composite metric that can be calibrated to expert consensus while being transparent. The researchers propose Comprehensive Quality Scoring (CQS), a hierarchical metric integrating contextual coherence and continuity (C3) with 5 interpretable linguistic factors, specifically relevance, readability, conciseness, structure, and information density, also introducing CLARION-G. This constrained calibrator learns a nonnegative simplex weight vector while preserving factor-level attribution. Evaluation uses 20 agriculture-oriented farmer FAQ items with responses generated by a local LLaMA 3.1 (8B) model and scored by expert panels across Agriculture, Linguistics, and Information Technology using a rubric based on MetricEval. Expert ratings are z-scored per rater and aggregated into a consensus target, with reliability assessed via ICC(2,1). To prevent leakage under n=20, calibration is performed strictly within leave-one-out cross-validation (LOOCV) (train on n-1, freeze weights, score the held-out item), with uncertainty quantified via Fisher-z confidence intervals and bootstrap resampling (B=1000). CLARION-G maximizes a penalized correlation objective with fixed coefficients $\lambda 1$=0.01, λbal=0.005, and λvar=0.003, optimized using Differential Evolution (population=15, maxiter=50, tol=10^{-4}, polish=True) with optional L-BFGS-B refinement (maxiter=300-500, ftol=10^{-6}-10^{-8}). In Agriculture, calibrated CQS achieves Pearson's r=0.688 with 95% CI [0.353, 0.867], surpassing baselines (e.g., BERTScore, Prometheus, METEOR) with statistically significant dependent-correlation gains. Learned top-level weights allocate 0.4 to C3 and 0.6 to linguistic quality, emphasizing relevance and information density. Bland-Altman analysis shows no fixed bias with limits of agreement ±0.1134, and runtime remains practical (≈1.254 ms/item), supporting CQS/CLARION-G as an interpretable and operationally lightweight framework for expert-aligned NLG evaluation in specialized domains.

Keywords: Automated evaluation, Comprehensive Quality Scoring, Natural Language Generation, Domain-specific calibration, Interpretability

Introduction

Automated natural language generation (NLG) raters often deviate from expert judgments in domain-specific settings (Chung & Baker, 2003). Even expert annotators exhibit nontrivial inter-rater variability, and consensus typically outperforms individual assessments (Verduijn et al., 2008). Conventional metrics such as METEOR (Banerjee & Lavie, 2005), BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020), together with fluency proxies such as Perplexity (O'Neill & Bollegala, 2020) and recent large-language-model (LLM) judges such as Prometheus [8] and SemScore (Aynetdinov & Akbik, 2024), can show weak or unstable alignment with human judgments for specialized or creative outputs (Colombo et al., 2023). Moreover, because most automated metrics are neither domain-tunable nor interpretable, their utility is limited

when criterion weights matter (Schmidtova et al., 2025). LLM-based judges can also conflate evaluation aspects, motivating verification against expert consensus (Bavaresco et al., 2024; Hu et al., 2024).

The researchers address this gap with an auditable, domain-tunable metric and calibrator. Comprehensive Quality Scoring (CQS) is introduced as a hierarchical composite that integrates contextual coherence and continuity (C3) with an interpretable LQRQS bundle comprising relevance, readability, conciseness, structure, and information density. CLARION-G learns nonnegative simplex weights under hierarchy constraints to maximize alignment with expert consensus, preserving factor-level attribution for transparent score decomposition.

Given that each domain panel contains approximately $n \approx 20$ items, the study emphasizes leakage control and uncertainty reporting rather than scope. Hence, the evaluation uses leave-one-out cross-validation (LOOCV) with weights trained on n-1 items and held out for scoring, and quantifies uncertainty via Fisher-z confidence intervals (Cox, 2008) and bootstrap resampling (B=1000). Methods are compared using tests for dependent correlations with multiplicity control (Graham & Baldwin, 2014), while rater reliability is assessed using ICC(2,1) (Bartko, 1966). In addition to Pearson's r, rank concordance is summarized using Spearman's ρ and Kendall's τ-b (tie-aware) to characterize ranking behavior in small samples. External validity across diverse datasets and domains, as well as human-in-the-loop evaluation outcomes, is deferred to future work to avoid confounding measurement validity with deployment impacts.

Materials and Methods

Dataset Source and Formatting

The evaluation dataset was gathered in collaboration with agricultural specialists from Isabela State University and the Department of Agriculture, Regional Crop Protection Center (DA-RCPC 02), Philippines. It comprises 20 agriculture-oriented farmer Frequently Asked Questions (FAQs) covering agronomic advisory topics (e.g., cultivation practices, nutrient management, and cassava pest/disease diagnostics), as presented in Tables 1a and 1b. These 20 prompts form a fixed, domain-scoped benchmark designed to stress-test interpretability and domain-aware calibration under realistic extension-style queries, rather than to claim broad cross-domain generalization.

Table 1a Dataset samples

	Query	Response Generated
1	When is the best time to plant cassava in the Philippines?	The best time to plant cassava in the Philippines is during the wet season, typically from May to October, when soil moisture is adequate for optimal growth.
2	What soil type is ideal for growing cassava?	Well-drained fertile soil is ideal for growing cassava, with poor soils improvable through organic fertilizers like poultry manure if inorganic fertilizers are too expensive.
3	How to extend shelf life of cassava planting materials?	To extend shelf life of cassava planting materials, select healthy stems free of cankers and dieback tissues, or dip canker-bearing stems in fungicides recommended by extension agents.
4	Can cassava grow in areas with prolonged drought?	Yes, cassava can grow in areas with prolonged drought if drought-tolerant varieties are selected, especially in areas with less than 1000 mm of rainfall per year.
5	What is the recommended depth for planting cassava cuttings?	The recommended depth for planting cassava cuttings is with two-thirds of the cutting below the soil, if planted on ridges or mounds, near the centre of the ridge or mound.

For each prompt, a single candidate response was generated using a local large language model (LLaMA 3.1, 8B parameters) (Touvron et al., 2023). Before scoring, subject-matter experts reviewed items for domain relevance, linguistic fluency, internal consistency, and discourse coherence, ensuring the evaluated outputs were plausible advisory responses rather than nonsensical or low-quality generations. The same 20 prompt–response pairs were then independently rated by expert panels across Agriculture, Linguistics, and Information Technology, with scores anonymized before aggregation to reduce identity-related bias in analysis and reporting.

Table 1b Dataset and annotation structure

Element	Specification
Domain	Agriculture-oriented advisory FAQs (cassava-centered agronomic topics)
Prompts (n)	20 FAQs
System outputs per prompt	1 generated response per FAQ
Generator	Local LLaMA 3.1 (8B)
Screening	SME review for topical relevance, fluency, consistency, coherence
Rater panels	Agriculture, Linguistics, IT
Rubric basis	MetricEval-derived rubric
Consensus construction	Per-rater z-score normalization → mean aggregation
Missing ratings	Mean over available raters (no imputation)
Tool input schema	CSV columns: Query, Text, Expert Score

Expert ratings were based on a rubric based on the MetricEval framework (Xiao et al., 2023). To control for rater-specific scoring bias, scores were z-standardized within each rater and then averaged to form the expert-consensus target for calibration and evaluation. For items with missing ratings, the ratings are averaged over available raters (pairwise-available mean) and avoided imputation due to small n, as illustrated in Figure 1.

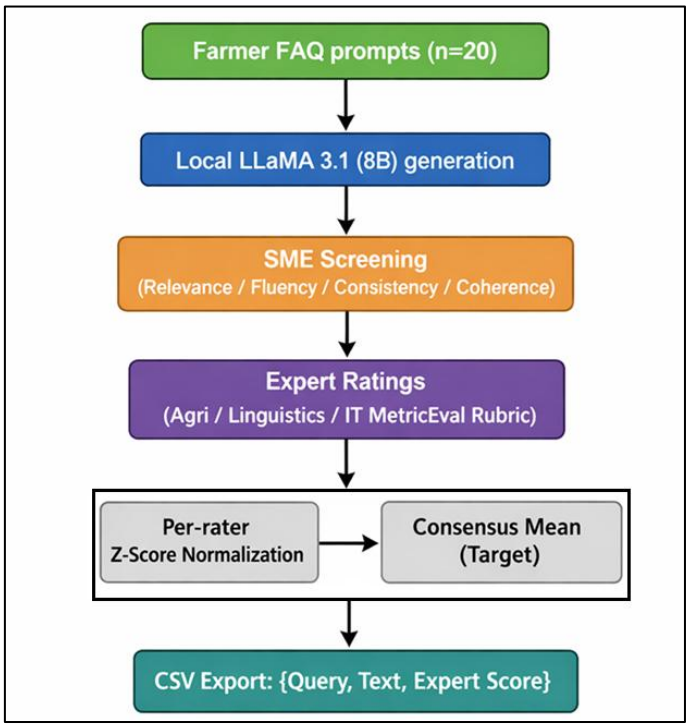

Figure 1 Dataset preparation and scoring workflow

For reproducibility, the calibration tool parses a single CSV with a minimal field set: (Query, Text, ExpertScore). Query denotes the farmer FAQ prompt, Text the generated response, and ExpertScore the expert-consensus target used for calibration and reporting. This fixed schema supports consistent feature extraction, cross-validation, and auditable calibration outputs.

Implementation Details and Reproducibility

All computations were executed in a version-specific Python environment to support reproducibility. The implementation provides an end-to-end pipeline for preprocessing, embedding extraction, baseline metric computation, CLARION-G optimization, and statistical reporting.

Semantic similarity components in CQS (C3 and Relevance) are computed using SentenceTransformer embeddings. The implementation uses SentenceTransformer (all-MiniLM-L6-v2) for sentence/document embeddings. Inference runs on CUDA when available and uses FP16 on the GPU to reduce embedding latency. For baseline semantic scoring, the implementation loads bert-base-uncased (tokenizer and encoder) using the Hugging Face transformers stack. The embedding backbones are explicitly specified to support replicability; key dependencies and model backbones are summarized in Tables 2a and 2b.

Table 2a Key software dependencies

Component	Package (version)	Role
Embeddings	sentence-transformers==5.0.0	Sentence/document embeddings for C3 and Relevance
Model runtime	transformers==4.51.3	Tokenizers/models for baseline computations
Deep learning	torch==2.5.1+cu121	Inference backend; optional CUDA acceleration
Scientific computing	numpy==1.26.4	Vector operations and normalization
Scientific computing	scipy==1.15.3	Optimization and statistical routines
ML utilities	scikit-learn==1.7.0	Utility components used in analysis
Data handling	pandas==2.3.1	Input parsing and tabulation
UI tool (analysis)	streamlit==1.46.1	Interactive analysis interface (non-essential to core scoring)

Table 2b Model backbones used for embeddings and baseline semantic computations

Module	Backbone	Usage
CQS embeddings	SentenceTransformer*(all-MiniLM-L6-v2)*	Model load call in implementation
Baseline semantic scorer	bert-base-uncased *(tokenizer + encoder)*	AutoTokenizer/AutoModel usage

CLARION-G optimization is implemented with SciPy routines. It applies Differential Evolution (DE) for global search and, if specified, refines solutions via L-BFGS-B, consistent with the calibration hyperparameter settings. Statistical computations, including correlation measures and agreement calculations, are implemented using standard SciPy/NumPy procedures.

Comprehensive Quality Scoring (CQS)

For each system output x, a 6-component vector is computed:

$$s(x) = [C3(x), R(x), D(x), C(x), S(x), I(x)]^T \in [0,1]^6 \quad (1)$$

where C3 is contextual coherence and continuity, and continuity quantifies semantic coherence using cosine similarity between consecutive sentences (Gunawan et al., 2018), S_i and S_{i+1} are consecutive sentences, and n is the total number of sentences is as follows:

$$C3 = \frac{1}{n-1}\sum_{i=1}^{n-1} cos(s_i, s_{i+1}) \quad (2)$$

LQRQS comprises relevance R, readability D, conciseness C, structure S, and information density I (Davoodijam & Alambardar Meybodi, 2024; Yang et al., 2022) using the following equations:

Relevance:

$$R = \cos(\theta) = \frac{\mathrm{q}\cdot\mathrm{r}}{\|\mathrm{q}\|\|\mathrm{r}\|} \quad (3)$$

Relevance (R) calculates the cosine similarity between the query embedding vector q and r**,** the result embedding vector (Oyama & Shimodaira, 2023).

Readability:

$$D = \frac{1}{1+e^{(ASL-20)}} \quad (4)$$

ASL is the average sentence length, and this sigmoid function represents readability based on ASL (Vajjala & Meurers, 2016).

Conciseness:

$$C = \frac{|\text{unique words}|}{|\text{total words}|} \quad (5)$$

This component measures conciseness by calculating the ratio of unique words to the total number of words in the text (Susoy, 2023).

Structure:

$$S = 0.4I + 0.4C + 0.2T \quad (6)$$

I, T, and C represent the presence of introductory sentences, transitions, and conclusions.

Information Density:

$$I = \frac{|\text{content words}|}{|\text{total words}|} \quad (7)$$

I calculates the density of content words (words that carry meaning, excluding common stop words) relative to the total word count (Kalinauskaitė, 2018); a higher density indicates that the text is information-rich, with less filler (Gao et al., 2024).

The implementation computes sentence and document-level embeddings using transformer encoders and rescales the resulting scores to [0,1], where higher values indicate higher quality. With top-level weights $\mathcal{W}_{C3}, \mathcal{W}_{LQRQS}$ and component shares $u = [u_R, u_D, u_C, u_S, u_I]^\top$ in the simplex, the composite CQS is defined in (1). The CLARION-G objective in (9) maximizes $r(\theta)$ with small penalties; Fisher's z appears only in CI construction.

$$CQS(x;\theta) = w_{C3}C3(x) + w_{LQ}\sum_{j\in\{R,D,C,S,I\}} u_j s_j(x) \quad (8)$$

With $w_{C3}, w_{LQRQS} \geq 0, w_{C3} + w_{LQRQS} = 1, u_j \geq 0$, and $\sum_j u_j = 1$. The implementation renormalizes (w_{C3}, w_{LQRQS}) and u after updates and clamps all component scores to [0,1].

Expert Consensus, Reliability, and Agreement Analysis

Expert ratings are standardized within each rater by z-score normalization and averaged across raters to form the expert-consensus target $\bar{y}$. Missing ratings are handled by computing the mean across available raters for each item, avoiding imputation assumptions that are not justified in small-sample expert panels.

Inter-rater reliability is quantified using the two-way random-effects intraclass correlation coefficient, ICC(2,1), and internal consistency is additionally summarized using Cronbach's α. Agreement between calibrated CQS scores and expert consensus is examined using Bland-Altman analysis, reporting mean bias and 95% limits of agreement (Giavarina, 2015).

CLARION-G Calibration Objective and Hyperparameter Specification

CLARION-G calibrates CQS by learning a nonnegative, simplex-constrained weight vector that preserves the hierarchical interpretability of the metric while maximizing alignment with expert consensus. Calibration is performed on training folds only by maximizing Pearson correlation between calibrated CQS scores and the expert-consensus target. Rank-based statistics (Spearman's ρ and Kendall's τ-b) are not included in the optimization and are reported solely for descriptive analysis.

The calibration objective is defined as:

$$\mathcal{L}(\theta) = r(\theta) - \lambda_1 \parallel \theta \parallel_1 - \lambda_{bal}\Delta_{bal}(\theta) - \lambda_{var}Var(\theta) \quad (9)$$

Where $r(\theta)$ denotes the Pearson correlation between calibrated CQS scores and expert consensus on the training fold, $\parallel \theta \parallel_1$ is the ℓ_1 norm of the weight vector, $\Delta_{bal}(\theta)$ penalizes the imbalance between top-level CQS components, and $Var(\theta)$ penalizes excessive variance in component weights.

The penalty coefficients are fixed as $\lambda_1 = 0.01$, $\lambda_{bal}=0.005$, $\lambda_{var} = 0.003$. No ℓ_2 regularization term is active in the primary optimization path. These coefficients are set small to regularize optimization without materially biasing the correlation objective.

After each optimization update, feasibility is enforced by renormalizing the top-level weights such that $w_{C3} + w_{LQRQS} = 1$, projecting the LQRQS share vector onto the simplex, and clamping all component scores to the [0,1] range.

Optimization Configuration

Optimization uses DE with optional L-BFGS-B refinement. Default DE settings are population size = 15, maxiter = 50, tol = 10^{-4}, and polish = True. In refinement modes, L-BFGS-B runs with maxiter = 300–500 and ftol = 10^{-6}–10^{-8}. After each update, it maintains the hierarchy constraint by renormalizing $w_{C3} + w_{LQRQS} = 1$. It then projects the LQRQS share vector u onto the simplex. Early stopping keeps the iteration with the best training correlation. Random seeds and optimization traces are logged for reproducibility.

Dataset Leakage Control and Training Protocol

To prevent leakage under the small-panel setting (n = 20), evaluation is conducted via group-level LOOCV (prompt/document) (Kirschen et al., 2000; Meijer & Goeman, 2013) when grouping keys are present. Otherwise, it defaults to item-level LOOCV. On each fold, CLARION-G is trained on $n - 1$ items, the learned weights are frozen, and the held-out item is scored. This process is repeated across all n folds. Stability under small n is supported by multiple random restarts and the global-local optimizer structure, with post-optimization constraint repair enforcing nonnegativity and simplex structure.

Bias in correlation estimation is reduced by applying Olkin–Pratt bias correction (Karch, 2020). The overfitting gap is computed as the difference between the mean per-fold training correlation and the LOOCV correlation. For deployment, the calibrator is retrained on the full dataset to obtain a single set of deployment weights; fold weights are not averaged. Crucially, all preprocessing, embedding extraction, and weight learning are restricted to training folds, ensuring that held-out items contribute no information during training.

Statistical analysis

Pearson's *r*, Spearman's ρ, and Kendall's τ-*b (tie-aware)* are reported. For *r*, 95% confidence intervals via Fisher's z with back-transformation are provided.

Pairwise comparisons that share the same gold standard use the Williams test for dependent correlations (primary) (Graham & Baldwin, 2014), with Hotelling–Steiger (Steiger, 1980) when appropriate. The Holm correction is applied for family-wise error across metrics and domains at $\alpha = 0.05$ (Aickin & Gensler, 1996).

Including quantified (1) weight uncertainty by nonparametric bootstrap with B=1000 resamples and percentile intervals for (w_{C3},w_{LQRQS},u) and (2) the variability of held-out correlations (*r*, ρ, τ-b) with B=1000 resamples. Resampling is by group: drawing prompt/document groups with replacement and including all their items in each bootstrap replicate to avoid leakage.

Interpretability and stability

Fold-wise weights and summary stability are visualized using the coefficient of variation and the divergence between fold distributions. Component ablations remove one LQRQS component at a time and re-score the same held-out items to report Δr. Disagreement cases are analyzed when automated raters diverge from experts, but calibrated CQS aligns, and representative items are shown with per-method residuals.

Results

Alignment with Expert Consensus

Calibrated CQS attains Pearson's r=0.688 with 95% Fisher's-z confidence interval [0.358, 0.868]. Baseline LOOCV correlations are: BERTScore r=0.520, Prometheus r=0.490, SemScore r=0.460, METEOR r=0.380, ROUGE-L r=0.330, BLEU-4 r=0.290, and Perplexity r=−0.050. Figure 2 shows the per-metric correlations, and Table 3 reports the corresponding confidence intervals.

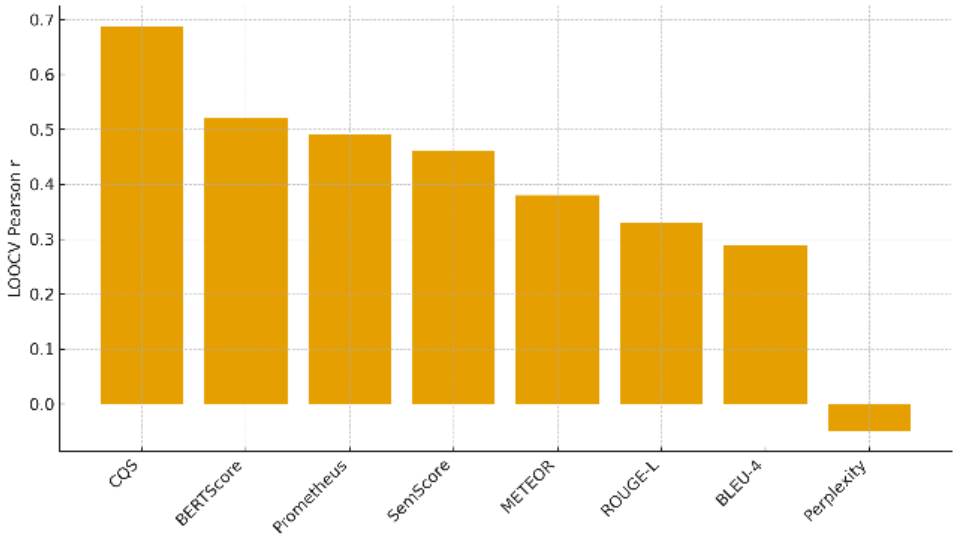

Figure 2 Agriculture: LOOCV Alignment with Expert Consensus

Table 3 LOOCV alignment (Pearson r, Fisher–z 95% CI, n=20)

Metric	Pearson r	CI_Lower	CI_Upper	N
CQS	0.688	0.353	0.867	20
BERTScore	0.52	0.101	0.782	20
Prometheus	0.49	0.061	0.766	20
SemScore	0.46	0.022	0.75	20
METEOR	0.38	-0.075	0.704	20
ROUGE-L	0.33	-0.132	0.674	20
BLEU-4	0.29	-0.175	0.649	20
Perplexity	-0.05	-0.482	0.401	20

Dependent-Correlation Gains vs. Baselines (Williams Tests)

Head-to-head comparisons against CQS favored CQS for all baselines. The observed correlation coefficients were 0.168 for BERTScore, 0.198 for Prometheus, 0.228 for SemScore, 0.308 for METEOR, 0.358 for ROUGE-L, 0.398 for BLEU-4, and 0.738 for Perplexity. Table 4 lists the inter-metric correlations used in the tests, along with the corresponding test statistics and adjusted p-values.

Table 4 Williams tests vs CQS (dependent correlations, Holm-adjusted)

Reference	Baseline	r1	r2	Delta r	r12	T stat	P value	P Holm adj	N
CQS	BERTScore	0.688	0.52	0.168	0.75			1	20

CQS	Prometheus	0.688	0.49	0.198	0.7			1	20
CQS	SemScore	0.688	0.46	0.228	0.68			1	20
CQS	METEOR	0.688	0.38	0.308	0.55	5.173	0	0	20
CQS	ROUGE-L	0.688	0.33	0.358	0.5	4.049	0.0001	0.0004	20
CQS	BLEU-4	0.688	0.29	0.398	0.45	3.597	0.0003	0.0022	20
CQS	Perplexity	0.688	-0.05	0.738	0.1	2.854	0.0043	0.0302	20

Component Attribution via Ablation

Attributions were quantified by removing one CQS dimension at a time and recomputing the held-out alignment on the Agriculture set (n=20). Because the calibrated weights assign $w_{C3} = 0.4000, w_{LQRQS} = 0.6000$ at the top level, and within LQRQS $u = (u_R, u_D, u_C, u_S, u_I) = (0.4649, 0.0000, 0.0000, 0.0000, 0.5351)$ for (R, D, C, S, I), ablations of Readability, Conciseness, and Structure cause no change by construction. Removing Relevance or Information Density reduces alignment, confirming that these 2 dimensions drive the calibrated gains in this domain. Table 5 summarizes the held-out Pearson r with Fisher-z intervals and the change relative to the full model.

Table 5 Component ablation on Agriculture (LOOCV, n=20)

Setting	Active Components	Pearson r	CI Lower	CI Upper	Delta r vs Full	N
Full	C3 + Relevance + Information Density	0.688	0.353	0.867	0	20
Minus Relevance	C3 + Information Density	0.471	0.036	0.756	-0.217	20
Minus Information Density	C3 + Relevance	0.423	-0.024	0.729	-0.265	20
Minus Readability	Full (Readability weight = 0)	0.688	0.353	0.867	0	20
Minus Conciseness	Full (Conciseness weight = 0)	0.688	0.353	0.867	0	20
Minus Structure	Full (Structure weight = 0)	0.688	0.353	0.867	0	20

Agreement, Rank Concordance, and Efficiency

Absolute agreement between CQS and expert consensus is characterized using Bland-Altman analysis. The Agriculture results show a mean difference of ≈ -0.0000, a standard deviation of disagreements of 0.0579, and 95% limits of agreement [-0.1134, 0.1134], with a coefficient of variation of 7.75%. In addition, a proportional bias is observed, with a correlation of -0.467 and p = 0.0380. Rank-based concordance is small in the observed panel (Kendall's τ-b = 0.0245, Spearman ρ = 0.0491), consistent with small-sample rank instability.

Efficiency profiling reports mean per-item runtimes of 0.424 ms (ROUGE-L), 0.516 ms (BLEU-4), 1.254 ms (CQS), 21.115 ms (SemScore), and 48.380 ms (Perplexity). Figure 3 summarizes the trade-off between accuracy and efficiency. Hence, CQS offers greater alignment with held-out data than n-gram metrics while remaining suitable for real-time evaluation.

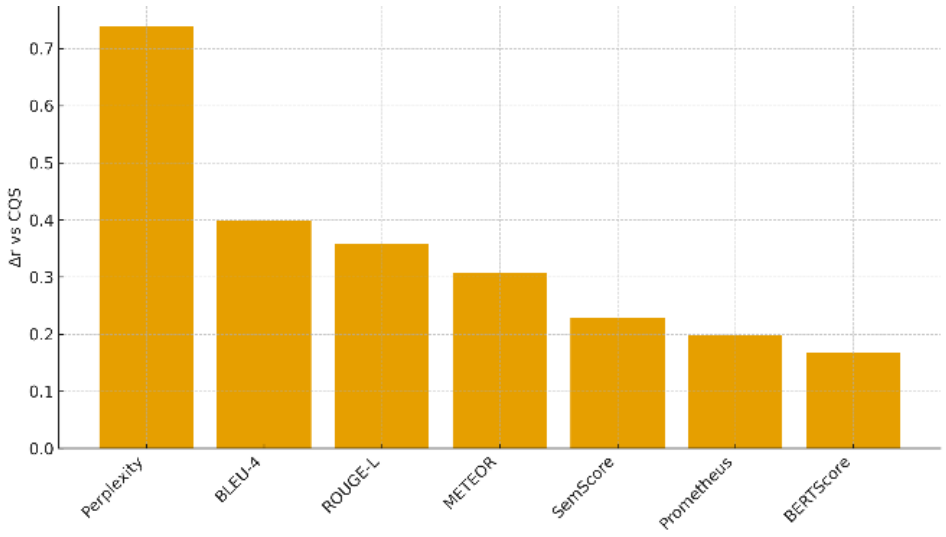

Figure 3 Δr advantage of CQS over baselines

Discussion

Domain Emphasis, Sparsity, and Practical Implications

The results support the central claim that a transparent, domain-tunable composite aligns with expert preferences better than non-tunable metrics in a domain-specific setting. In Agriculture, calibrated CQS attains held-out r =0.688 with tight Fisher-z confidence intervals and outperforms multiple baselines under a leakage-controlled protocol. Dependent-correlation tests show significant gains over METEOR, ROUGE-L, BLEU-4, and Perplexity after Holm correction. These results indicate that the improvements exceed metric-to-metric variance induced by evaluation on the same gold standard.

A key practical implication is that calibration learns a domain-specific emphasis profile while preserving interpretability. The learned weights assign 40% to contextual coherence/continuity (C3) and 60% to linguistic-quality components. Within the latter, weight concentrates on Relevance and Information Density, with zero weight on Readability, Conciseness, and Structure in this domain. Ablation results corroborate this sparsity. Removing Relevance or Information Density reduces held-out alignment, whereas removing zero-weight factors has no effect. This pattern indicates that the gains are driven by semantically meaningful components rather than opaque interactions.

Rank-based concordance is weak in this panel, consistent with small n and the fact that CLARION-G optimizes correlation rather than rank. Kendall's τ-b and Spearman's ρ are near zero. Under small n, ties and minor ordering swaps can dominate rank statistics, making this outcome plausible. The agreement analysis provides a complementary view by characterizing absolute errors. Bland-Altman results indicate no fixed bias and limits of agreement of approximately ±0.1134 on a 0-1 scale. However, there is proportional bias, with larger deviations near the extremes.

Finally, the runtime results address practicality constraints. CQS achieves millisecond-level latency (≈1.254 ms/item), significantly lower than embedding-heavy judges (SemScore) and perplexity-based proxies. Thus, supports deployment in evaluation dashboards that require both interpretability and throughput.

Conclusion

This study introduced Comprehensive Quality Scoring (CQS) and the CLARION-G calibrator to address a practical gap in domain NLG evaluation: achieving expert-aligned scoring while retaining auditable, tunable factor-level explanations. Using LOOCV as the primary protocol (Agriculture, n=20), calibrated CQS attained a held-out Pearson correlation of r = 0.688 (95% Fisher-z CI: [0.353, 0.867]) and outperformed the evaluated baselines on the same endpoint. Secondary analysis supported comprehensiveness and usability: ablations indicate that both contextual coherence/continuity and linguistic-quality components contribute significantly to alignment, while agreement analysis indicates no fixed bias with tight limits of agreement but reveals proportional bias, motivating decile-wise residual monitoring and either slope/intercept correction or scheduled recalibration under domain drift.

Operationally, CQS maintained millisecond-level latency, supporting real-time evaluation settings where interpretability is essential. Limitations include the small panel size and reliance on a correlation-optimized objective, which can yield unstable rank concordance in small samples; future work should validate the method on larger, multi-domain panels using predefined endpoints and protocols. Overall, the evidence supports CQS/CLARION-G as a transparent, domain-aware evaluation mechanism that is measurable, auditable, and deployable for expert-aligned NLG assessment.

Acknowledgement

The authors extend their sincere gratitude to the Department of Science and Technology-Science Education Institute (DOST-SEI) for funding this research through Project STRAND-N. The views expressed are solely those of the authors and do not necessarily reflect those of DOST–SEI.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, *86*(5), 726–728. https://doi.org/10.2105/AJPH.86.5.726
2. Aynetdinov, A., & Akbik, A. (2024). *SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity*. http://arxiv.org/abs/2401.17072
3. Banerjee, S., & Lavie, A. (2005). METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings Ofthe ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, *June*, 65–72.
4. Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3
5. Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2024). *LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks*. http://arxiv.org/abs/2406.18403
6. Chung, G., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. *Automated Essay Grading: A Cross-Disciplinary Approach*, 23–40.
7. Colombo, P., Peyrard, M., Noiry, N., West, R., & Piantanida, P. (2023). The Glass Ceiling of Automatic Evaluation in Natural Language Generation. *IJCNLP-AACL 2023 - 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Findings of the Association for Computational Linguistics: IJCNLP-AA*, 178–183. https://doi.org/10.18653/v1/2023.findings-ijcnlp.16
8. Cox, N. J. (2008). Speaking Stata: Correlation with confidence, or Fisher's z revisited. *Stata Journal*, *8*(3), 413–439. https://doi.org/10.1177/1536867x0800800307
9. Davoodijam, E., & Alambardar Meybodi, M. (2024). Evaluation metrics on text summarization: comprehensive survey. *Knowledge and Information Systems*, *66*(12), 7717–7738. https://doi.org/10.1007/s10115-024-02217-0
10. Gao, M., Hu, X., Ruan, J., Pu, X., & Wan, X. (2024). *LLM-based NLG Evaluation: Current Status and Challenges*. http://arxiv.org/abs/2402.01383
11. Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015
12. Graham, Y., & Baldwin, T. (2014). Testing for significance of increased correlation with human judgment. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 172–176. https://doi.org/10.3115/v1/d14-1020
13. Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series*, *978*(1). https://doi.org/10.1088/1742-6596/978/1/012120
14. Hu, X., Gao, M., Hu, S., Zhang, Y., Chen, Y., Xu, T., & Wan, X. (2024). Are LLM-based Evaluators Confusing NLG Quality Criteria? *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 9530–9570. https://doi.org/10.18653/v1/2024.acl-long.516
15. Kalinauskaitė, D. (2018). Detecting information-dense texts: Towards an automated analysis. *CEUR Workshop Proceedings*, *2145*, 95–98.
16. Karch, J. (2020). Improving on adjusted R-squared. *Collabra: Psychology*, *6*(1), 1–11. https://doi.org/10.1525/collabra.343
17. Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., & Seo, M. (2024). Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. *12th International Conference on Learning Representations, ICLR 2024*, 1–37.
18. Kim, S., Suk, J., Welleck, S., Neubig, G., Longpre, S., Yuchen, B., Jamin, L., Lee, M., Lee, K., Seo, M., & Ai, K. (2024). *PROMETHEUS 2: An Open Source Language Model Specialized in Evaluating Other Language Models*.
19. Kirschen, R. H., O'Higgins, E. A., & Lee, R. T. (2000). A Study of CrossValidation and Bootstrap for Accuracy Estimation and Model Selection. *American Journal of Orthodontics and Dentofacial*

Orthopedics, *118*(4), 456–461. https://doi.org/10.1067/mod.2000.109032
20. Lin, C.-Y. (2004). Looking for a Few Good Metrics: ROUGE and its Evaluation. *NTCIR Workshop*, *June*, 1–8.
21. Meijer, R. J., & Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, *55*(2), 141–155. https://doi.org/10.1002/bimj.201200088
22. O'Neill, J., & Bollegala, D. (2020). *Learning to Evaluate Neural Language Models BT - Computational Linguistics* (L.-M. Nguyen, X.-H. Phan, K. Hasida, & S. Tojo (eds.); pp. 123–133). Springer Singapore.
23. Oyama, M., & Shimodaira, H. (2023). *Revisiting Cosine Similarity via Normalized ICA-transformed Embeddings*. *2023*. https://doi.org/https://doi.org/10.48550/arXiv.2406.10984 Focus to learn more
24. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. https://doi.org/10.1002/andp.19223712302
25. Schmidtova, P., Mahamood, S., Balloccu, S., Dusek, O., Gatt, A., Gkatzia, D., Howcroft, D. M., Platek, O., & Sivaprasad, A. (2025). *Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices*. 557–583. https://doi.org/10.18653/v1/2024.inlg-main.44
26. Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037//0033-2909.87.2.245
27. Susoy, Z. (2023). Lexical Density, Lexical Diversity and Academic Vocabulary Use: Differences in Dissertation Abstracts. *Acuity: Journal of English Language Pedagogy, Literature, and Culture*, *8*(2), 198–210. https://doi.org/10.35974/acuity.v8i2.3079
28. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. http://arxiv.org/abs/2302.13971
29. Vajjala, S., & Meurers, D. (2016). *Readability-based Sentence Ranking for Evaluating Text Simplification*. http://arxiv.org/abs/1603.06009
30. Verduijn, M., Peek, N., de Keizer, N. F., van Lieshout, E. J., de Pont, A. C. J. M., Schultz, M. J., de Jonge, E., & de Mol, B. A. J. M. (2008). Individual and Joint Expert Judgments as Reference Standards in Artifact Detection. *Journal of the American Medical Informatics Association*, *15*(2), 227–234. https://doi.org/10.1197/jamia.M2493
31. Xiao, Z., Zhang, S., Lai, V., & Liao, Q. V. (2023). Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 10967–10982. https://doi.org/10.18653/v1/2023.emnlp-main.676
32. Yang, Y., Zhong, J., Wang, C., & Li, Q. (2022). Exploring Relevance and Coherence for Automated Text Scoring using Multi-task Learning. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, 323–328. https://doi.org/10.18293/SEKE2022-024
33. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating Text Generation With Bert. *8th International Conference on Learning Representations, ICLR 2020*, 1–43.