An Initial Validation of a Pedagogical Simplification Gain Index for Arabic Technical Text Simplification

Khalid Essaadani[1]*, Soumia Ziti[2], Karima Salah-Eddine[3]
Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco[1,3]
IPSS Team, Faculty of Sciences, Morocco[2]

***Abstract*—Simply calculating the readability score or the degree of lexical overlap for Arabic technical text simplification does not provide the full picture because an explanation that is shorter may still damage the technical meaning, weaken terminology consistency, or fail to reduce learner difficulty. This study introduces the Pedagogical Simplification Gain Index (PSGI Auto), which is automatic and interpretable, and is designed to measure the pedagogical gain brought about by simplification. Three elements are integrated into PSGI Auto: meaning preservation gain, terminology clarity gain, and concision gain. The technique was tested on Arabic simplification outputs sourced from five information technology areas, which provided a human reference score obtained from three independent expert raters. There is a strong and statistically significant correlation between PSGI Auto and human PSGI judgments, with a Pearson correlation of 0.779, Spearman correlation of 0.755, and 81.6% direction agreement. The results suggest that PSGI Auto can possibly be used as a practical and interpretable learner-oriented Arabic technical simplification tool that would help to identify whether simplification results in pedagogical improvement, not just a superficial change in the text. Since the validation is carried out on a relatively small dataset, the metric is to be interpreted as an initial one requiring further validation by comparison with well-established simplification metrics and larger-scale testing.**

***Keywords*—Arabic technical text simplification; automatic text simplification; pedagogical evaluation; PSGI Auto; human evaluation**

I. Introduction

In the field of educational technology, Automatic Text Simplification (ATS) plays an important role, as it aims to make learning more accessible to different groups of learners by reducing linguistic complexity while preserving semantic integrity [1], [2]. Although considerable progress has been made in the simplification of general-domain texts, adapting technical and scientific content for Arabic-speaking learners remains a major challenge [3], [4]. Arabic technical texts contain a high density of terminology, complex morphological structures, and frequent code-switching with English technical terms, which can make them difficult for non-expert readers to understand [5], [6]. In educational and technical contexts, the purpose of simplification is not merely to substitute words, but also to preserve technical meaning while improving clarity for learners; this concern has been emphasized particularly in biomedical and health-information simplification, where inaccurate simplification can reduce the usefulness of specialized content [7], [8].

Although Arabic simplification corpora are becoming increasingly available, such as the SAMER Corpus for fiction and emerging resources for technical domains, the evaluation of these systems still lacks metrics that genuinely reflect pedagogical utility [3], [9]. Traditional evaluation protocols generally rely on surface-level metrics or general semantic similarity scores that are unable to capture the complex trade-off between simplicity and technical accuracy [10], [11]. A simplified passage that appears easier on the surface may still distort the intended meaning; this risk motivates the use of meaning-preservation evaluation alongside readability or simplicity evaluation [1], [12], [13]. This gap highlights the need for evaluation frameworks that align with human expert judgments regarding meaning preservation and terminology clarity [1], [14].

To address this limitation, this research introduces the Pedagogical Simplification Gain Index (PSGI Auto), an interpretable automatic metric that aims to measure the pedagogical gain of Arabic technical text simplification. PSGI Auto differs from many existing evaluation metrics that score a single output independently; it employs a gain-oriented approach that assesses the glossary-guided output relative to a baseline output. This design is motivated by prior work on reference-less and language-independent simplification evaluation, but the explicit baseline-relative gain formulation is proposed in the present study [15], [16]. The metric consists of three primary dimensions: meaning preservation (ΔM), terminology clarity (ΔT), and concision (ΔC). These dimensions are motivated by prior work on interpretable simplification feedback and pedagogically suitable simplification, while the specific weighting scheme is a pedagogical assumption introduced in this study [17], [18]. We validate PSGI Auto by comparing the results of this lightweight formula against human expert ratings across five IT domains, thereby assessing its ability to approximate pedagogical judgment in a preliminary Arabic technical simplification scenario. Prior Arabic simplification and automated scoring studies motivate the broader methodological context, but the PSGI validation itself is conducted in the present study [19], [20]. Therefore, the contribution is presented as a compact and interpretable metric proposal, rather than as a definitive replacement for established automatic metrics or human evaluation.

II. Related Work

The evaluation of text simplification systems has usually been carried out through the combined use of readability formulas and reference-based metrics. Readability indices, such as the Flesch-Kincaid Grade Level (FKGL), typically estimate text difficulty through surface features such as sentence length and word-level properties [4], [11]. Although they are computationally efficient, these formulas have several limitations. Arabic

*Corresponding author.

readability research has shown that readability assessment requires language-sensitive modeling and that surface features alone are insufficient to capture the full difficulty of Arabic texts; these limitations are even more important when texts contain technical terminology [6], [21]. In the Arabic context, machine learning models for readability assessment have recently been studied, supported by contextual embeddings to better capture syntactic and semantic subtleties [22], [23]. However, these methods mainly estimate difficulty levels rather than assessing the pedagogical benefit achieved through the simplification process.

Reference-based metrics, such as BLEU and ROUGE, originate from machine translation and summarization tasks. They focus on the degree of n-gram overlap between system outputs and human references [24], [25]. However, n-gram overlap is poorly suited to simplification, since it penalizes valid paraphrases and structural changes that differ from the reference, regardless of simplification quality [16], [26]. In response to this limitation, the SARI metric was introduced specifically for simplification. It operates by assessing the addition, deletion, and retention of n-grams [10], [24]. Although SARI correlates better with human judgments of simplicity, it does not provide an explicit account of meaning preservation. This can lead to cases where texts receive high scores for being simpler while containing semantically distorted information [1], [25].

More recent metrics, such as AMAT, LENS, and BERTScore, rely on semantic representations to measure meaning relations through contextual embeddings [12], [27]. BERTScore performs token-level matching using contextual embeddings and is widely used as a semantic similarity metric that is less dependent on exact n-gram overlap than traditional lexical metrics [27], [28]. LENS further advances this direction by combining simplicity, fluency, and meaning preservation into a learned simplification metric [29]. Despite these advancements, general-purpose metrics do not directly evaluate whether domain-specific terminology is used or explained appropriately. This is important for technical simplification, where lexical choices and specialized concepts can strongly influence the usefulness of the output [7], [30]. Moreover, these metrics generally score each output independently rather than modeling the gain achieved over a specific baseline output, which is the baseline-relative perspective introduced in the present study [15], [16], [31].

In the Arabic context, attention has mainly been placed on developing specific corpora and models for simplification, including hybrid approaches that combine machine translation with transformer-based lexical models [5], [32]. Studies have also investigated how control tokens and sequence-to-sequence architectures can influence the degree of simplification [20], [33]. Nevertheless, the assessment of Arabic simplification and readability systems is still frequently connected to adapted evaluation practices or limited human evaluation, and existing work does not yet provide a dedicated automatic metric for Arabic technical pedagogical simplification that jointly captures meaning preservation, terminology clarity, and concision [3], [4]. Recent advances in meaning preservation metrics, such as MeaningBERT and JUDGEBERT, highlight the need for domain-specific semantic evaluation; however, such resources remain almost nonexistent in the Arabic technical education context [12], [13]. Therefore, metrics that jointly assess meaning preservation, terminology clarity, and concision within the same framework remain absent for Arabic technical texts.

III. Proposed Metric

This section of the study discusses the Pedagogical Simplification Gain Index (PSGI) used in this research. PSGI does not simply assess the readability level of a simplified text; rather, it estimates the *gain* produced by the simplification process relative to a baseline output. A good simplification should preserve the original technical meaning, clarify important terms, and reduce unnecessary textual burden. Therefore, the automatic PSGI is defined as a weighted sum of three gain components: meaning preservation gain, terminology clarity gain, and concision gain. The automatic PSGI formula is conceptualized as follows:

$$PSGI_{auto} = 0.50\Delta M + 0.30\Delta T + 0.20\Delta C \tag{1}$$

where, ΔM represents the increase in meaning preservation, ΔT represents the increase in terminology clarity, and ΔC represents the increase in concision between the baseline and simplified outputs. The weights reflect a predefined pedagogical assumption rather than values learned from the validation set: meaning preservation is treated as the most important condition for useful simplification, followed by terminology clarity and concision. This choice avoids fitting the metric directly to the human ratings used for validation, but it also means that the weights should be further examined through future sensitivity and ablation analyses.

For a specific source text i, the three gain components are calculated as follows:

$$\Delta M_i = M_{i,simplified} - M_{i,baseline}, \tag{2}$$
$$\Delta T_i = T_{i,simplified} - T_{i,baseline}, \tag{3}$$
$$\Delta C_i = C_{i,simplified} - C_{i,baseline}. \tag{4}$$

The meaning component M estimates lexical-semantic fidelity to an independently prepared Arabic reference representation of the source technical content. In the present version, M is computed as the cosine similarity between character-level TF-IDF vectors extracted from that reference and the generated output:

$$M_i = \cos\left(\mathrm{tfidf}(Source_i), \mathrm{tfidf}(O_i)\right), \tag{5}$$

In the above expression, $Source_i$ denotes the Arabic reference representation of source technical content, whereas O_i denotes the generated output for item i. Character-level TF-IDF was selected as a lightweight, reproducible baseline that is relatively tolerant of Arabic affixes, borrowed English terms, and inconsistent tokenization. It remains a surface-sensitive proxy, however; it cannot establish semantic equivalence between different valid paraphrases. Accordingly, the results validate the complete PSGI Auto formulation only in this initial setting, and future work should compare this component with Arabic or multilingual embedding-based similarity measures.

The terminology component T measures the proportion of expected glossary terms that are actually used in the output. For a given source text i, let G_i denote the glossary set containing the terms related to it. Terminology coverage is defined as:

$$T_i = \frac{|\{g \in G_i : g \in O_i\}|}{|G_i| + \epsilon}, \tag{6}$$

We set $\epsilon = 10^{-8}$. Each study item has at least one expected term; the constant is therefore only a numerical safeguard and does not affect the reported scores. The definition measures coverage, not contextual correctness, explanation quality, or pedagogical usefulness. It can reward unnecessary insertion and can miss valid paraphrases; the terminology component should consequently be interpreted as a glossary-alignment feature rather than an independent measure of terminology clarity. The conciseness component C is computed from the character-length reduction achieved relative to the Arabic reference representation.

$$C_i = \max\left(0, 1 - \frac{len(O_i)}{len(Source_i) + \epsilon}\right), \tag{7}$$

The length of the output in character count is denoted by $len(\cdot)$ after the preprocessing step. According to this formulation, shorter outputs receive higher values, but outputs that are longer than the source representation are not rewarded. The possibility that excessive shortening may damage meaning is the reason behind the small weight of concision in Eq. 1. The human reference score is computed based on five expert-rating dimensions. For each evaluated output, the Human Pedagogical Score (HPS) is defined as:

$$HPS = \frac{R + S + T + M + (6 - D)}{5} \tag{8}$$

Here, R signifies readability, S denotes simplicity, T stands for terminology clarity, M refers to meaning preservation, while D illustrates learner difficulty. Learner difficulty has a reverse scoring method, meaning that $(6 - D)$ is used for it, since greater difficulty implies lower pedagogical quality. The human PSGI is finally calculated as the discrepancy between the simplified and baseline HPS values.

$$PSGI_{human} = HPS_{simplified} - HPS_{baseline}. \tag{9}$$

If the PSGI value is positive, it suggests improvement relative to the paired baseline; it does not establish that the final output is objectively high quality. A value close to zero indicates minimal relative change, while a negative value indicates deterioration relative to that baseline.

IV. Methodology and Experimental Design

The study works with a controlled dataset of Arabic technical text simplification outputs that include five information technology domains: programming, databases, networks, cybersecurity, and cloud computing. Each source text has two corresponding generated outputs: a baseline output and a glossary-guided simplified output. Fig. 1 summarizes the paired computation and validation design. The objective is to examine whether the automatic PSGI score aligns with expert judgments of pedagogical simplification gain.

A. Generation Conditions

Two outputs under controlled generation conditions were produced for each source text. The baseline condition generated a simplified Arabic explanation without explicit glossary constraints. The glossary-guided condition generated a simplified Arabic explanation while encouraging the use or explanation of expected technical terms. Both outputs were paired by source text, which allows PSGI to be computed as a gain score rather than as an isolated output-quality score. Raters were not informed whether an output came from the baseline or glossary-guided condition, and outputs were presented using anonymized evaluation identifiers in randomized order. This blinding and randomization step is important because PSGI includes a terminology-coverage component and the glossary-guided condition may otherwise receive an artificial advantage. The blinding procedure reduces the risk that human scores are biased toward the glossary-guided condition, although it does not remove the need to test PSGI Auto against external automatic metrics that do not use the proposed weighting scheme.

B. Dataset

There are 100 outputs generated from 50 technical source texts in the dataset. Each source document comes with one baseline output and one glossary-guided simplified output. The dataset is distributed evenly over five technical domains, with 20 outputs assigned to each domain. The validation is carried out at the text level by pairing the baseline outputs with the corresponding simplified outputs that are generated from the same source text. Before the analysis was done, the dataset was examined for missing values, duplicated rows, duplicated text-mode pairs, and inconsistent metadata. No required values were missing, and no duplicated evaluation identifiers or invalid ratings were found. One paired item was excluded from the gain analysis because the baseline and simplified outputs were identical. Ultimately, the final PSGI Auto against PSGI Human validation involves 49 valid paired comparisons.

C. Human Evaluation

The evaluation of each output was carried out independently by three raters using five distinct criteria: readability, simplicity, terminology clarity, meaning preservation, and learner difficulty. The scoring for each of these five criteria was done on a scale of 1–5. Higher points stand for better quality for readability, simplicity, terminology clarity, and meaning preservation. For learner difficulty, a higher number of points indicates greater difficulty, and therefore learner difficulty is inverted. That is to say, learner difficulty is inverted in Eq. 8 when computing the Human Pedagogical Score. For every output, the mean scores from the three raters were obtained for each criterion. The human PSGI was then computed by subtracting the baseline HPS values from the simplified HPS values for the same text. This design views expert judgment as the human reference standard and assigns automatic PSGI computation to the ranking values. Inter-rater reliability was calculated on the 100 rated

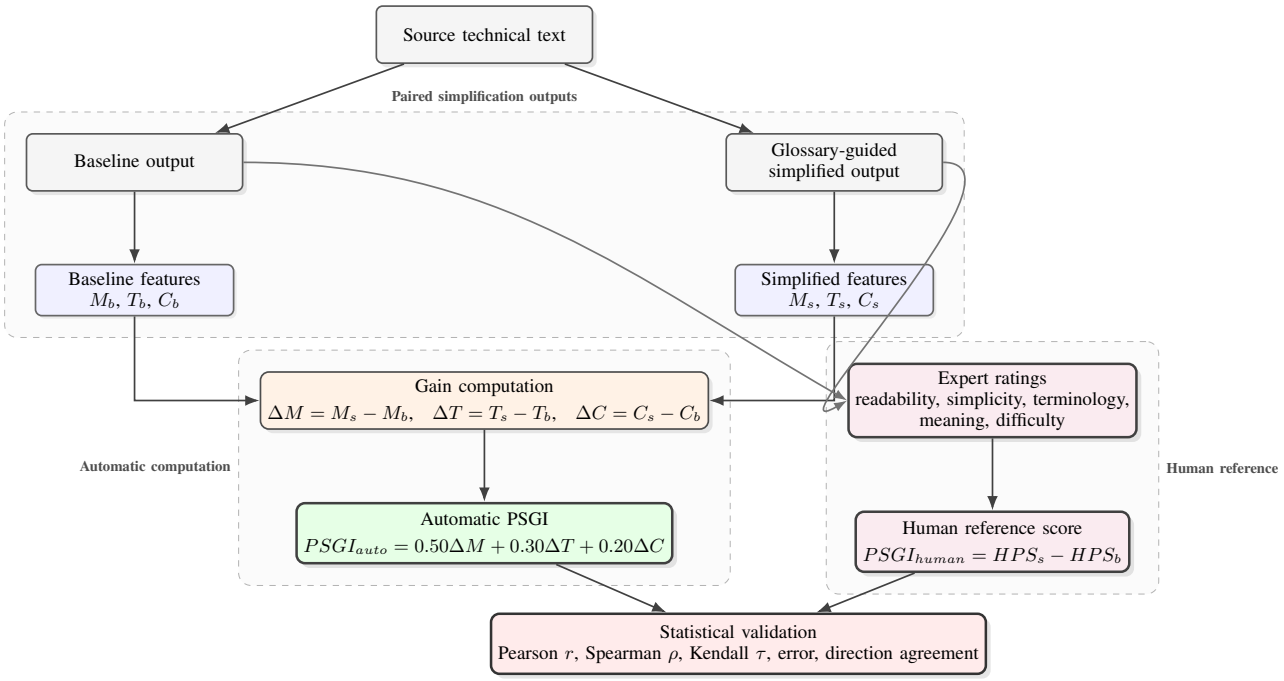

Fig. 1. PSGI Auto computation and validation workflow.

TABLE I. INTER-RATER RELIABILITY ACROSS THE THREE EXPERT RATERS. ICC(2,1) REPORTS SINGLE-RATER RELIABILITY, WHILE ICC(2,K) REPORTS RELIABILITY AFTER AVERAGING THE THREE RATERS

Rating dimension	ICC(2,1)	ICC(2,k)
Readability	0.604	0.821
Simplicity	0.599	0.818
Terminology clarity	0.582	0.807
Meaning preservation	0.648	0.847
Learner difficulty	0.547	0.783
Human Pedagogical Score	0.681	0.865

TABLE II. RELIABILITY OF RATER-LEVEL PAIRED HUMAN PSGI GAINS ACROSS 49 SOURCE-TEXT PAIRS

Statistic	Value
ICC(2,1)	0.714
ICC(2,k)	0.882

outputs before averaging the three raters. Table I displays both single-rater ICC and average-rater ICC values. We additionally computed ICC on rater-level paired HPS gains (Table II); the average-rater gain ICC(2,k) is 0.882. Thus, the averaged three-rater paired gain is a reasonably reliable reference for this pilot study, while learner difficulty remains the most subjective individual dimension.

D. Automatic PSGI Computation

The automatic PSGI computation involves only text-derived features and excludes the human ratings. Meaning preservation is estimated through similarity to the manual Arabic reference, terminology clarity through glossary-term coverage, and concision through the change in output length. These features are computed for both outputs of each paired source text, transformed into gain values, and combined using Eq. 1. To examine whether the weighted formula adds value beyond its individual components, the validation includes diagnostic component variants and reference-overlap baselines calculated on the same outputs. These comparisons test incremental alignment with the human PSGI reference; they do not establish broad superiority over all established simplification metrics.

E. Validation Procedure

The main validation compares $PSGI_{auto}$ with $PSGI_{human}$ across 49 valid paired items. Pearson correlation estimates linear association, while Spearman correlation and Kendall's tau estimate rank agreement. MAE and RMSE quantify prediction error after scale alignment. Direction agreement is reported because the sign of PSGI indicates relative improvement or deterioration. We additionally report bootstrap confidence intervals, leave-one-pair-out correlation ranges, and the reliability of rater-level paired gains. These robustness analyses are descriptive because the dataset is small.

V. RESULTS

In this part, the empirical validation of PSGI Auto against PSGI Human is presented. The analysis utilizes 49 valid paired comparisons after excluding one identical baseline–simplified pair.

A. Overall Validation

Table III and Fig. 2 summarize the association between automatic and human PSGI. The overall correlation is encouraging for this pilot sample, but it should not be interpreted as a stable population estimate. The rank correlations indicate that the association is not solely linear. Robustness results in

TABLE III. VALIDATION OF $PSGI_{auto}$ AGAINST $PSGI_{human}$

Metric	Value
Valid paired items	49
Pearson r	0.779
Pearson p	4.40e-11
Spearman ρ	0.755
Spearman p	3.79e-10
Kendall τ	0.562
Kendall p	1.27e-08
MAE	0.250
RMSE	0.295
R^2	0.525
Direction agreement	81.6%

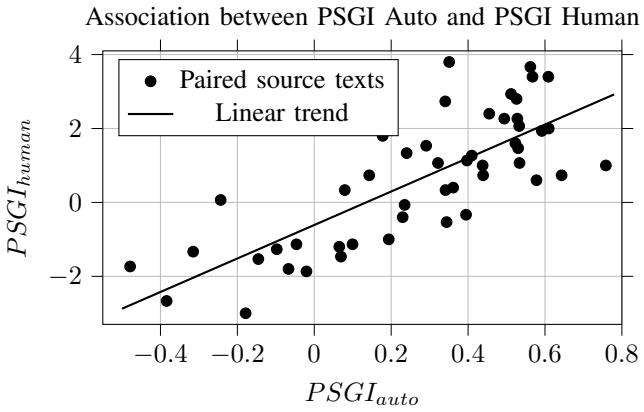

Fig. 2. Scatter plot generated from the 49 valid paired comparisons between PSGI Auto and raw human PSGI. The positive trend is consistent with the strong Pearson correlation reported in Table III.

TABLE IV. ROBUSTNESS CHECKS FOR THE OVERALL PEARSON CORRELATION

Check	Value
Observed Pearson r	0.779
Bootstrap 95% interval	[0.671, 0.860]
Leave-one-pair-out range	[0.758, 0.798]

Table IV show the extent to which the estimate changes under resampling and leave-one-pair-out analysis.

The automatic score achieves 81.6% direction agreement with the human score. Thus, for most pairs in this sample, PSGI Auto and the human reference agree about the direction of change relative to the baseline.

B. Human Rating Summary

In Table V, the averaged human ratings are given according to the generation condition. The outputs that are guided by the glossary obtain higher mean scores for readability, simplicity, terminology clarity, meaning preservation, and the overall HPS. Learner difficulty decreased from 3.26 to 2.61, which is positive because lower difficulty means that learners can more easily access the material.

C. Ablation and Weight Sensitivity

The ablation analysis examines the weighted PSGI formula against equal-weight and component-only variants. Table VI

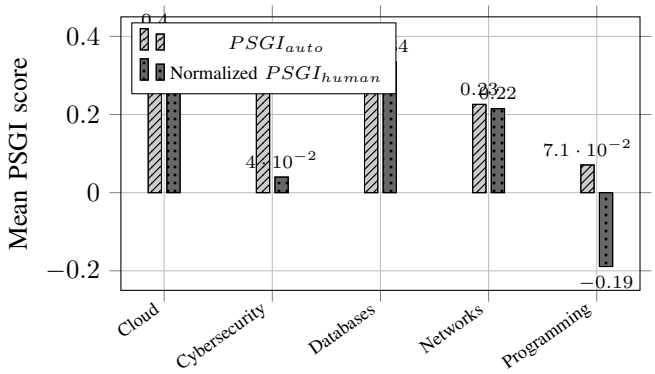

Fig. 3. Domain-level mean PSGI Auto and normalized PSGI Human scores. Positive values indicate pedagogical gain after simplification, while negative values indicate harmful simplification. Human PSGI is normalized by dividing the raw score by four for visual comparability with PSGI Auto.

shows that the proposed weighting has the highest correlations among these internal variants. The advantage over equal weights is small, so the weights should be considered pedagogically motivated and promising rather than optimized or definitive. Meaning gain is the strongest individual component; terminology and concision gains are positively correlated in this sample but are insufficient on their own to judge pedagogical quality.

D. Comparison with Reference-Overlap Baselines

Table VII compares PSGI Auto with reference-overlap baselines on the same 49 paired outputs. These metrics are included as transparent, reproducible diagnostic baselines. A complete comparison with SARI, BERTScore, and LENS is outside the present dataset and implementation: SARI requires an Arabic complex-source/reference configuration that is not available here, and Arabic-validated implementations of BERTScore and LENS should be selected and evaluated in a separate, preregistered benchmark. Accordingly, this table does not claim to settle comparative performance against those metrics.

PSGI Auto has higher correlation than the listed baselines in this sample. This result is consistent with the value of combining the three features, but it remains preliminary and does not demonstrate general superiority.

E. Domain-Level Results

Table VIII and Fig. 3 report descriptive domain-level results. Human PSGI is divided by four in the figure solely for scale comparability. Each domain contains only nine or ten pairs; consequently, the domain correlations have very wide uncertainty and are not interpreted as stable domain-specific findings.

F. Interpretation of the Main Finding

The principal finding is that the proposed formula correlates with human paired gain scores in this dataset. PSGI Auto is computed from text-derived features, whereas the reference is based on independently collected expert ratings. The ablation results suggest that combining meaning preservation, glossary alignment, and concision is more aligned with the human gain

TABLE V. HUMAN RATING SUMMARY BY GENERATION CONDITION. GAINS ARE COMPUTED AS GLOSSARY-GUIDED MINUS BASELINE ACROSS THE 49 VALID PAIRED COMPARISONS

Dimension	Baseline mean	Glossary-guided mean	Mean gain	Gain SD
Readability	3.50	4.05	0.55	1.59
Simplicity	3.31	3.80	0.49	1.75
Terminology clarity	2.83	3.80	0.97	1.80
Meaning preservation	2.90	3.44	0.54	1.98
Learner difficulty	3.26	2.61	-0.65	1.83
Human Pedagogical Score	3.06	3.70	0.64	1.73

TABLE VI. ABLATION AND WEIGHT-SENSITIVITY ANALYSIS AGAINST HUMAN PSGI ACROSS 49 VALID PAIRED COMPARISONS. DIRECTION AGREEMENT REPORTS WHETHER THE AUTOMATIC AND HUMAN GAIN SIGNS MATCH

Variant	Pearson r	p-value	Spearman ρ	Kendall τ	Direction agreement
PSGI current $(0.50/0.30/0.20)$	0.779	4.40e-11	0.755	0.562	81.6%
Equal weights $(1/3, 1/3, 1/3)$	0.768	1.20e-10	0.742	0.547	81.6%
Meaning only	0.657	2.90e-07	0.677	0.472	85.7%
Terminology only	0.546	4.96e-05	0.500	0.402	61.2%
Concision only	0.509	1.87e-04	0.482	0.335	51.0%
Meaning–terminology only $(0.625/0.375)$	0.745	8.58e-10	0.724	0.528	81.6%

TABLE VII. CORRELATION OF PSGI AUTO AND REFERENCE-OVERLAP BASELINES WITH HUMAN PSGI ACROSS 49 PAIRED COMPARISONS. BASELINES ARE GLOSSARY-GUIDED MINUS BASELINE SCORES AGAINST THE MANUAL ARABIC REFERENCE

Metric	Pearson r
PSGI Auto	0.779
BLEU-1 gain	0.643
ROUGE-L-like gain	0.739
Character TF-IDF gain	0.657

TABLE VIII. DOMAIN-LEVEL COMPARISON BETWEEN PSGI AUTO AND PSGI HUMAN

Domain	n	Mean PSGI Auto	Mean PSGI Human	Pearson r
cloud	10	0.402	1.460	0.447
cybersecurity	10	0.300	0.160	0.790
databases	10	0.362	1.340	0.544
networks	10	0.226	0.860	0.885
programming	9	0.071	-0.756	0.866

reference than the evaluated individual components. This should be interpreted as initial evidence for the formulation under the tested generation conditions, not as proof that the metric is universally valid or that glossary-term occurrence itself causes pedagogical improvement.

VI. DISCUSSION

The validation results indicate that PSGI Auto may provide an initial approximation to expert judgments of relative pedagogical gain. Its baseline-relative design asks whether an intervention improves an output compared with its paired baseline. This is useful because a shorter text may still weaken technical meaning or terminology. It does not, however, imply that a positive PSGI score certifies the final output as high quality.

The weights were specified before validation from a pedagogical priority: preserving meaning is necessary, terminology alignment can support technical learning, and concision is beneficial only when it does not remove essential content. The small difference from equal weights means that this rationale should not be confused with empirical optimization. The terminology component is particularly constrained by coupling: the glossary-guided condition was encouraged to use the same glossary that PSGI Auto measures. Blinded, randomized human rating reduces rater bias but does not remove this metric–intervention coupling. Future validation should include outputs generated without glossary guidance and deliberate adversarial cases, including unnecessary term insertion, valid paraphrases, meaning deletion, and overly compressed text.

The domain results are descriptive only. With nine or ten pairs per domain, apparent differences can be driven by a few observations. PSGI Auto should therefore be treated as a practical screening signal for the studied setting, alongside human review, rather than as a replacement for human evaluation.

VII. THREATS TO VALIDITY

This study has several limitations. First, the validation dataset contains 49 valid paired comparisons after one identical pair was removed. The overall estimates are therefore preliminary, and the domain-level estimates are especially uncertain. Second, the weights were not learned from the human ratings, but the small difference from equal weights means that they require larger sensitivity analyses. Third, the benchmark includes reference-overlap and component baselines, not a complete SARI, BERTScore, or LENS comparison. Future work should use an Arabic complex-source/reference dataset and validated Arabic or multilingual implementations of these metrics on the same outputs. Fourth, the glossary used to guide one generation condition also informs the terminology-coverage feature. This coupling can inflate the apparent contribution of glossary coverage. Fifth, coverage alone cannot assess whether a term is correct, necessary, explained, or pedagogically helpful; valid paraphrases may be penalized. Sixth, character-level TF-IDF is a lightweight surface-sensitive proxy for meaning and cannot reliably recognize all semantically equivalent Arabic formulations. Seventh, the human reference uses paired HPS gains. We report gain-level reliability, but direct learner outcomes remain necessary to establish educational benefit.

Finally, the study does not include deliberately constructed failure cases or a full influence analysis beyond leave-one-pair-out checks. These should be included before deploying PSGI Auto as an operational evaluation metric.

VIII. Ethical and Reproducibility Considerations

The present article is an analysis of generated text outputs and expert evaluations. No student performance data, sensitive personal data, or intervention data were included in the study. The expert ratings were collected voluntarily and were subsequently analyzed using anonymized output identifiers. To ensure reproducibility, the study documents the generation conditions, PSGI formulation, rating procedure, and validation statistics. The anonymized generated outputs, rating files, derived PSGI tables, and calculation scripts can be made available upon reasonable request.

IX. Conclusion

This study introduces PSGI Auto, an interpretable metric for assessing relative pedagogical gain in Arabic technical text simplification. It combines meaning-preservation gain, glossary-alignment gain, and concision gain. On a small paired dataset, PSGI Auto correlates encouragingly with three-rater human PSGI gains. The evidence is preliminary: the dataset is small, terminology coverage is coupled to the glossary-guided intervention, the meaning feature is surface-sensitive, and the benchmark is incomplete. Future research should use larger and more diverse datasets, adversarial failure cases, independent terminology evaluation, stronger Arabic semantic models, direct learner outcomes, and a full comparison with established simplification metrics.

References

[1] S. Agrawal and M. Carpuat, "Do text simplification systems preserve meaning? a human evaluation via reading comprehension," *Transactions of the Association for Computational Linguistics*, vol. 12, no. 00, pp. 432–448, 2024.

[2] P. B. Githinji, A. Melliou, Z. Liang, L. Zhang, and P. Qin, "Making knowledge accessible: Divergent readability-accuracy strategies of mistral and qwen in biomedical text simplification," arXiv preprint, 2025.

[3] B. Alhafni, R. Hazim, J. P. Liberato, M. A. Khalil, and N. Habash, "The samer arabic text simplification corpus," arXiv preprint, 2024.

[4] N. Nassiri, V. Cavalli-Sforza, and A. Lakhouaja, "Approaches, methods, and resources for assessing the readability of arabic texts," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1–30, 2023.

[5] A. Alsughayyir and A. Alshanqiti, "A light but efficient switch transformer model for arabic text simplification," *International Journal of Advanced and Applied Sciences*, vol. 12, no. 4, pp. 71–78, 2025.

[6] N. Nassiri, A. Lakhouaja, and V. Cavalli-Sforza, "Arabic l2 readability assessment: Dimensionality reduction study," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3789–3799, 2022.

[7] B. Ondov, K. Attal, and D. Demner-Fushman, "A survey of automated methods for biomedical text simplification," *Journal of the American Medical Informatics Association*, vol. 29, no. 11, pp. 1976–1988, 2022.

[8] O. Alhadreti, "An examination of the content of diabetes websites targeting arabic speakers," *International Journal of Medical Informatics*, vol. 179, p. 105242, 2023.

[9] N. Habash, H. Taha-Thomure, K. N. Elmadani, Z. Zeino, and A. Abushmaes, "Guidelines for fine-grained sentence-level arabic readability annotation," arXiv preprint, 2024.

[10] L. Cripwell, J. Legrand, and C. Gardent, "Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation," arXiv preprint, 2024.

[11] L. C, J. Legrand, and C. Gardent, "Readability measures and automatic text simplification: In search of a construct," arXiv preprint, 2024.

[12] D. Beauchemin, H. Saggion, and R. Khoury, "Meaningbert: Assessing meaning preservation between sentences," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[13] D. Beauchemin, M. Albert-Rochette, R. Khoury, and P.-L. Déziel, "Judgebert: Assessing legal meaning preservation between sentences," arXiv preprint, 2025.

[14] A. B. Al Ajlouni, J. Li, and M. A. Ajlouni, "Towards a comprehensive metric for evaluating text simplification systems," in *2023 14th International Conference on Information and Communication Systems (ICICS)*, 2023, pp. 1–6.

[15] L. Mucida, A. Oliveira, and M. Possi, "Language-independent metric for measuring text simplification that does not require a parallel corpus," in *The International FLAIRS Conference Proceedings*, vol. 35, 2022.

[16] L. Martin, S. Humeau, P.-E. Mazaré, A. Bordes, É. V. de La Clergerie, and B. Sagot, "Reference-less quality estimation of text simplification systems," arXiv preprint, 2019.

[17] P. G, M. Orosoo, W. G. Shanthi, V. Waiker, A. Smerat, B. Pagidipati, B. G. S, and O. R. Shahin, "An interpretable dual-level feedback approach for improving graded language simplification and readability," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 11, 2025.

[18] J. Degraeuwe and H. Saggion, "Lexical simplification in foreign language learning: Creating pedagogically suitable simplified example sentences," in *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability*, 2022.

[19] K. Dahimi, H. Cherroun, and A. Belabbaci, "Automated scoring of arabic text using large language models: A literature review," arXiv preprint, 2026.

[20] N. Khallaf and S. Sharoff, "Towards arabic sentence simplification via classification and generative approaches," arXiv preprint, 2022.

[21] V. Cavalli-Sforza, H. Saddiki, and N. Nassiri, "Arabic readability research: Current state and future directions," in *Procedia Computer Science*, vol. 142, 2018, pp. 38–49.

[22] M. A. Ouassil, M. Jebbari, R. Rachidi, M. Errami, B. Cherradi, and A. Raihani, "Enhancing arabic text readability assessment: A combined BERT and BiLSTM approach," in *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, 2024, pp. 1–7.

[23] S. Bessou and G. Chenni, "Efficient measuring of readability to improve documents accessibility for arabic language learners," arXiv preprint, 2021.

[24] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," in *Transactions of the Association for Computational Linguistics*, vol. 4, 2016, pp. 401–415.

[25] L. Vásquez-Rodríguez, M. Shardlow, P. Przybyła, and S. Ananiadou, "Investigating text simplification evaluation," arXiv preprint, 2021.

[26] T. Scialom, L. Martin, J. Staiano, É. V. de la Clergerie, and B. Sagot, "Rethinking automatic evaluation in sentence simplification," arXiv preprint, 2021.

[27] S. Jaskowski, S. Chava, and A. Shah, "Bertscorevisualizer: A web tool for understanding simplified text evaluation with bertscore," arXiv preprint, 2024.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *International Conference on Learning Representations*, 2020.

[29] M. Maddela, Y. Dou, D. Heineman, and W. Xu, "LENS: A learnable evaluation metric for text simplification," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 16 383–16 408.

[30] S. Štajner, D. Ferrés, M. Shardlow, K. North, M. Zampieri, and H. Saggion, "Lexical simplification benchmarks for english, portuguese, and spanish," *Frontiers in Artificial Intelligence*, vol. 5, 2022.

[31] L. Cripwell, J. Legrand, and C. Gardent, "Simplicity level estimate (sle): A learned reference-less metric for sentence simplification," arXiv preprint, 2023.

[32] S. S. Al-Thanyyan and A. M. Azmi, "Simplification of arabic text: A hybrid approach integrating machine translation and transformer-based lexical model," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, p. 101662, 2023.

[33] Z. Li and M. Shardlow, "How do control tokens affect natural language generation tasks like text simplification," *Natural Language Engineering*, vol. 30, no. 5, pp. 915–942, 2024.