Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EViSE: An Efficient Framework for Multi-Criteria Summarization Evaluation

Thi Thu Phuong Tran
Hanoi Metropolitan University
Trinh Hoang Vu
VNU University of Engineering and Technology, Vietnam National University
Van Vinh Nguyen
vinhnv@vnu.edu.vn

VNU University of Engineering and Technology, Vietnam National University
Phuong Thai Nguyen
VNU University of Engineering and Technology, Vietnam National University
Anh-Quan Nguyen-Duc
VNU University of Engineering and Technology, Vietnam National University

Research Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11409998/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

EViSE: An Efficient Framework for Multi-Criteria Summarization Evaluation

Thi Thu Phuong Tran [1,2], Trinh Hoang Vu [2], Van Vinh Nguyen [2*], Phuong Thai Nguyen [2], Anh-Quan Nguyen-Duc [2]

[1]Hanoi Metropolitan University, 98 Duong Quang Ham, Nghia Do, Hanoi, 100000, Vietnam.
[2]VNU University of Engineering and Technology, Vietnam National University, 144 Xuan Thuy, Cau Giay District, Hanoi, 100000, Vietnam.

*Corresponding author(s). E-mail(s): vinhnv@vnu.edu.vn;

Abstract

Reliable summarization evaluation requires criterion-sensitive scoring, computationally practical inference, and useful behavior during model optimization. We propose EViSE, an efficient multi-criteria summarization evaluation framework with a criterion-wise, group-balanced cardinal–ordinal objective coupling score regression with within-source ranking. The effectiveness of EViSE is demonstrated in several aspects: The framework provides a general formulation for grouped multi-criteria learning, with objective-level properties established through mathematical analysis independently of the encoder implementation. EViSE achieves criterion-wise agreement with human judgments in multi-criteria summarization evaluation while efficiently predicting Faithfulness, Coherence, and Relevance in one reference-free encoder pass. A matched WikiLingua application demonstrates downstream reuse as a frozen reward evaluator under group relative policy optimization (GRPO), with similar human-reviewed quality after evaluator substitution. Using a corpus of 80,856 human-reviewed Vietnamese document–summary instances, we evaluate two instantiations of the framework. On held-out News, Overall Spearman correlations are 0.737 for EViSE-ViDeBERTa and 0.806 for EViSE-mmBERT. Both retain substantial agreement under domain shift to information technology textbooks. The compact instantiation achieves 55.2 ms/pair latency on a matched Tesla T4 profile. The formulation can be adapted to other languages using suitable encoders and criterion-labeled source–summary groups.

Keywords: Summarization evaluation, Multi-criteria evaluation, Cardinal-ordinal learning, Reference-free evaluation, Resource-efficient inference.

1 Introduction

Automatic summarization is open-ended: multiple summaries can be acceptable for the same source, while fluent outputs may still omit salient information or introduce unsupported claims. Reliable evaluation therefore requires more than a single surface-similarity score. ROUGE and BERTScore remain useful and reproducible, but they do not explicitly separate factual support, discourse quality, and content selection (Lin 2004; Zhang et al. 2020; Maynez et al. 2020). Multi-criteria evaluation instead treats Faithfulness, Coherence, and Relevance as distinct but related dimensions (Fabbri et al. 2021; Zhang et al. 2024).

This requirement becomes more demanding when an evaluator is reused during model optimization. In reinforcement learning and related post-training, a quality signal is queried repeatedly and directly influences policy updates (Stiennon et al. 2020; Ouyang et al. 2022; Bai et al. 2022). Prompted LLM judges can provide strong rubric-based judgments (Liu et al. 2023; Kocmi and Federmann 2023), but repeated autoregressive or hosted evaluation can be costly. A practical evaluator should therefore combine criterion-level agreement, efficient local inference, and stable reuse as an optimization-time signal.

Vietnamese summarization has growing generation resources, including VNDS, ViMs, WikiLingua, VieSum, VietMed-Sum, and VSum-HB (Nguyen et al. 2019; Tran et al. 2020; Ladhak et al. 2020; Lam et al. 2023; Le-Duc et al. 2024; Tran Thi et al. 2025). However, publicly documented infrastructure for compact multi-criteria evaluation remains limited. EViSE addresses this gap by treating evaluation as a grouped learning problem: one shared source-summary representation supports criterion-specific F/C/R predictions, regression provides cardinal calibration, and within-source pairwise ranking provides ordinal supervision. The ranking branch reuses predictions already available within source-grouped mini-batches, and inference returns all three criteria from a single encoder pass.

EViSE denotes the learning framework rather than a fixed encoder. Method-level invariants are the source-conditioned shared representation, criterion-specific outputs, source grouping, and hybrid cardinal-ordinal objective. The encoder, pooling operator, tokenizer, and context budget are implementation choices. We instantiate the framework with two complementary encoders: EViSE-ViDeBERTa, a compact Vietnamese-specialized model, and EViSE-mmBERT, a higher-capacity multilingual model with a substantially longer context window. The comparison tests implementation portability and exposes an accuracy-efficiency trade-off.

EViSE's methodological focus is a criterion-wise, group-balanced hybrid objective for multi-criteria evaluators. It couples three rubric-anchored score predictions with relative comparisons defined separately for F, C, and R among summaries of the same source. Pair losses are normalized within each nonempty source-criterion group and averaged across all sources and the three criteria, with an empty pair set

contributing zero, so a source does not receive a larger ranking weight solely because it contains more unequal-label pairs. This formulation builds on combined regression and ranking (Sculley 2010) and query-level loss normalization (Liu 2009). Its analysis characterizes the interaction of absolute score anchoring and within-source separation in the multi-criteria evaluator setting.

The present work makes three contributions:

- **Criterion-wise, group-balanced hybrid evaluation.** EViSE jointly learns rubric-anchored F/C/R scores and criterion-specific within-source ordering, with pair losses normalized within source-criterion groups and source losses averaged under fixed equal criterion weights. The analysis separates group offsets from centered errors, characterizes margin compatibility and prediction-gradient alignment, and links the document-balanced ranking loss to ordering error.
- **Cross-instantiation accuracy-efficiency evidence.** Two substantially different Transformer implementations retain the same criterion structure and objective, providing evidence that the formulation is not tied to one encoder configuration.
- **Validation across criteria, domains, and uses.** Evaluation covers pointwise agreement, within-source ordering, criterion specificity, computational cost, and IT-textbook domain shift. The compact model additionally receives independent blind OOD validation, and a matched downstream application examines the effect of substituting the frozen reward evaluator.

2 Related work

2.1 Multi-criteria summarization evaluation

Reference-based metrics such as ROUGE and BERTScore remain standard because they are reproducible and inexpensive (Lin 2004; Zhang et al. 2020), but reference similarity does not directly identify factual support, discourse coherence, or salient-content preservation. Human-evaluation studies and recent summarization benchmarks therefore assess multiple dimensions, including Faithfulness, Coherence, and Relevance (Maynez et al. 2020; Fabbri et al. 2021; Zhang et al. 2024). EViSE follows this criterion-wise view while learning all three scores jointly from the source and candidate summary.

2.2 Learned evaluators, verification, and reward signals

Learned evaluators address limitations of surface overlap in different ways. UniEval predicts multiple quality dimensions (Zhong et al. 2022). QA-based methods such as FEQA, QuestEval, and QAFactEval probe factual support (Durmus et al. 2020; Scialom et al. 2021; Fabbri et al. 2022); and NLI-based methods such as SummaC model entailment and contradiction (Laban et al. 2022). LLM judges extend rubric-based evaluation further but require repeated autoregressive inference or hosted calls (Liu et al. 2023; Kocmi and Federmann 2023). These methods motivate an evaluator that is criterion-aware, reference-free at inference, and inexpensive enough for repeated use.

Recent studies have improved abstractive summary generation through topic-aware contrastive learning, sentence-level prefix prompting, and fact-aware enhancement with re-ranking (Tang et al. 2024; Zhao et al. 2025; Li et al. 2025). These studies optimize the summarization system itself. EViSE focuses on learning criterion-level F/C/R scores, preserving within-source quality ordering, and supporting reuse as a frozen reward signal during downstream optimization.

2.3 Hybrid regression-ranking objectives and grouped preferences

Shared representations can support related prediction tasks (Caruana 1997), while pairwise objectives learn relative preferences (Burges et al. 2005). Jointly optimizing regression and ranking also has a direct precedent: the Combined Regression and Ranking framework of Sculley (2010) learns from both objectives. EViSE builds on this hybrid-learning principle for multi-criteria summarization evaluation.

Preference-based reward modeling supplies another relevant comparison. In the Bradley-Terry formulation (Bradley and Terry 1952), a preference probability can be written as a logistic function of two scores. InstructGPT uses this score-difference likelihood to learn a scalar reward from human comparisons (Ouyang et al. 2022):

$$\Pr(z_a \succ z_b \mid x) = \sigma(r_\phi(x, z_a) - r_\phi(x, z_b)),$$

where x is the common prompt, z_a, z_b are responses, and σ is the logistic function. The pairwise likelihood is unchanged by a common score shift and therefore does not itself identify an absolute rubric score. EViSE's regression term explicitly anchors each F/C/R output to its reviewed numeric target; its margin term uses relative comparisons derived from those same criterion labels.

Grouping and normalization likewise have precedents. Learning-to-rank methods compare candidates associated with the same query, and query-level normalization prevents queries with more pairs from dominating the pairwise loss (Liu 2009). InstructGPT also groups comparisons by prompt and reuses response scores across pairs (Ouyang et al. 2022). EViSE maps the grouping unit to the source document and constructs separate eligible-pair sets for F, C, and R. Its ranking risk averages eligible pair losses within each source-criterion group, assigns zero when no unequal-label comparison is available, and then averages over all sources and the three criteria. The methodological focus is this criterion-wise, group-balanced evaluator formulation and the analysis of how its cardinal and ordinal terms interact.

2.4 Vietnamese resources and relation to earlier studies

Vietnamese summarization resources include VNDS, ViMs, WikiLingua, VieSum, VietMed-Sum, and VSum-HB (Nguyen et al. 2019; Tran et al. 2020; Ladhak et al. 2020; Lam et al. 2023; Le-Duc et al. 2024; Tran Thi et al. 2025). VSum-HB introduced a human-feedback-oriented resource of 5,000 Vietnews samples with LLM-assisted summary construction and human refinement (Tran Thi et al. 2025). The resource used here is organized for evaluator learning: 13,476 source documents have

six candidate summaries each, yielding 80,856 document-summary evaluations with criterion-level F/C/R labels. Section 4.1.1 describes its construction and annotation.

Tran et al. (2026) studied verifier-guided PPO/GRPO and penalty shaping using the frozen MultiEvalSumViet2 evaluator. That released compact checkpoint is retained here as EViSE-ViDeBERTa. The present study examines the evaluator's grouped learning formulation, loss geometry, criterion-level agreement, within-source ordering, computational cost, and IT-textbook domain shift, together with a higher-capacity mmBERT instantiation. The downstream experiment uses the previously reported compact-evaluator configuration as its reference and adds a matched evaluator-substitution run. Appendix C summarizes the scope of the three studies and identifies the reused checkpoint and downstream reference result.

3 Proposed EViSE framework

We formulate EViSE as a grouped multi-criteria learning problem, define its cardinal-ordinal objective, and establish objective-level properties before discussing efficient computation and language portability. The formulation separates the learning structure from encoder-specific implementation. Section 4 describes the Vietnamese realizations and their empirical evaluation. Figure 1 summarizes the common pipeline.

3.1 Grouped multi-criteria formulation

Consider a grouped supervised learning problem in which each group contains multiple candidates with numeric labels for several criteria. The learner must estimate criterion-specific scores on a common scale and preserve the relative ordering of candidates within each group. EViSE specializes this setting to source documents as groups and alternative summaries as candidates.

For a source document D and candidate summary S, EViSE models three related criteria. **Faithfulness (F)** measures whether the summary's entities, relations, quantities, attributions, and other statements are supported by the source (Maynez et al. 2020; Fabbri et al. 2021). **Coherence (C)** measures discourse organization, self-containedness, and linguistic well-formedness (Fabbri et al. 2021). **Relevance (R)** measures preservation of salient source information and appropriate content selection (Nenkova and Passonneau 2004; Fabbri et al. 2021). These criteria are complementary: a source-supported summary may omit important content, while broad coverage may include unsupported claims. Omission of salient information lowers Relevance, omission alone is not a factual inconsistency under Faithfulness.

Let the training resource be organized into G source groups, $\mathcal{T} = \{\mathcal{G}_g\}_{g=1}^{G}$, with

$$\mathcal{G}_g = \{(D_g, S_{gj}, \mathbf{y}_{gj})\}_{j=1}^{n_g}, \qquad \mathbf{y}_{gj} = (y_{gj}^F, y_{gj}^C, y_{gj}^R),$$

where $n_g \geq 2$ and $\mathcal{K} = \{F, C, R\}$. Flattening these groups gives $N = \sum_{g=1}^{G} n_g$ instances. Let g_i identify the source of instance i, $\mathcal{G} = \{1, \ldots, G\}$ denote the source indices, and $\mathcal{I}_g = \{i : g_i = g\}$ denote the candidate indices for source g.

Let y_i^k be the reviewed numeric target for criterion k on a common rubric scale. These scores provide cardinal supervision. Section 4.1.1 specifies the normalization used in the Vietnamese experiments. Multiple candidates for the same source additionally define criterion-wise ordinal targets:

$$S_i \succ_{k,D_g} S_j \quad \Longleftrightarrow \quad y_i^k > y_j^k, \qquad i,j \in \mathcal{I}_g. \tag{1}$$

The cardinal and ordinal signals therefore come from the same reviewed criterion labels.

A differentiable encoder E_θ jointly contextualizes the source and summary under an implementation-specific input budget L. A representation operator P_ψ maps the contextual states to one fixed-dimensional representation:

$$H_i^{(L)} = E_\theta(D_i, S_i; L), \qquad z_i = P_\psi\left(H_i^{(L)}\right). \tag{2}$$

Criterion-specific mappings produce the three scores:

$$\hat{y}_i^k = f_k(z_i), \qquad k \in \mathcal{K}. \tag{3}$$

Joint source-summary encoding supplies the evidence required for Faithfulness and Relevance while retaining the summary's discourse information for Coherence.

Definition 1 (Framework instantiation)**.** A grouped multi-criteria evaluator is an instantiation $\mathcal{I} = (E_\theta, P_\psi, L, \{f_k\}_{k\in\mathcal{K}})$ that maps each source-summary pair to $\hat{\mathbf{y}}_i = (\hat{y}_i^F, \hat{y}_i^C, \hat{y}_i^R)$ through Eqs. (2)–(3) and is trained with the source-grouped objective in Eq. (8). Its learning structure comprises a shared source-conditioned representation, criterion-specific outputs, source grouping, and cardinal-ordinal supervision. The encoder, representation operator, tokenizer, and context budget are implementation choices.

3.2 Group-balanced cardinal-ordinal objective

Regression anchors each criterion prediction to the common reviewed score scale:

$$\mathcal{L}_{\text{reg}} = \frac{1}{3N} \sum_{k\in\mathcal{K}} \sum_{i=1}^{N} (\hat{y}_i^k - y_i^k)^2. \tag{4}$$

Within-source ranking supplies relative supervision separately for each criterion. For source g and criterion k, define the eligible unequal-label pairs

$$\mathcal{P}_g^k = \{(i,j) : i,j \in \mathcal{I}_g,\ i < j,\ y_i^k \neq y_j^k\}, \tag{5}$$

and their preference signs $s_{ij}^k = \text{sign}(y_i^k - y_j^k)$. With margin $m > 0$, the source-criterion contribution is

$$\ell_g^k = \begin{cases} \dfrac{1}{|\mathcal{P}_g^k|} \displaystyle\sum_{(i,j)\in\mathcal{P}_g^k} \max\left[0, m - s_{ij}^k(\hat{y}_i^k - \hat{y}_j^k)\right], & |\mathcal{P}_g^k| > 0, \\ 0, & |\mathcal{P}_g^k| = 0. \end{cases} \tag{6}$$

The ranking risk averages over all sources and all three criteria:

$$\mathcal{L}_{\text{rank}} = \frac{1}{3G} \sum_{k\in\mathcal{K}} \sum_{g\in\mathcal{G}} \ell_g^k. \tag{7}$$

Averaging pairs within each source-criterion group prevents a source from receiving greater ranking weight solely because it has more unequal-label pairs. An empty pair set contributes zero, and its source remains in the all-source denominator. The criterion coefficients are fixed at 1/3.

The hybrid objective is

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{rank}}, \qquad \lambda \geq 0, \tag{8}$$

where m specifies the required signed separation and the fixed coefficient λ controls the ranking contribution. Theoretical properties of this objective are developed in Section 3.3; the reported training settings are given in Section 4.2.2.

Grouped mini-batch computation.

For a micro-batch containing $b \geq 1$ complete source groups, Algorithm 1 computes

$$\widehat{\mathcal{L}}_{\text{rank}} = \frac{1}{3b} \sum_{k\in\mathcal{K}} \sum_{g\in\mathcal{G}_{\mathcal{B}}} \ell_g^k,$$

where $\mathcal{G}_{\mathcal{B}}$ contains all b source groups, including those with empty pair sets. Uniform source sampling gives $\mathbb{E}[\widehat{\mathcal{L}}_{\text{rank}}] = \mathcal{L}_{\text{rank}}$ at fixed predictions. Gradient accumulation averages complete micro-batch losses with the same zero-for-empty convention. Appendix B.2 proves the normalization and sampling identities.

3.3 Theoretical analysis of the grouped cardinal-ordinal objective

The analysis characterizes how absolute score anchoring and relative separation interact under unconditional source normalization. It addresses score identifiability, empirical target compatibility, prediction-gradient alignment, and document-balanced ordering control. Complete assumptions and proofs are given in Appendix B.

For source g and criterion k, let $e_i^k = \hat{y}_i^k - y_i^k$ denote the signed prediction error. Its group mean $\bar{e}_g^k = n_g^{-1} \sum_{i\in\mathcal{I}_g} e_i^k$ is the source-criterion offset. The centered errors $\varepsilon_i^k = e_i^k - \bar{e}_g^k$ have empirical variance $V_g^k = n_g^{-1} \sum_{i\in\mathcal{I}_g} (\varepsilon_i^k)^2$.

Algorithm 1 Source-balanced within-document margin ranking loss

Require: Micro-batch $\mathcal{B}$ partitioned into $b \geq 1$ complete source groups $\mathcal{G}_{\mathcal{B}}$; predictions, labels, and margin $m > 0$

Ensure: $\widehat{\mathcal{L}}_{\text{rank}}$

$A \leftarrow 0$
for $k \in \{F, C, R\}$ **do**
  for $g \in \mathcal{G}_{\mathcal{B}}$ **do**
    $\mathcal{I}_g \leftarrow \{i \in \mathcal{B} : g_i = g\}$
    $\mathcal{P}_g^k \leftarrow \{(i, j) : i, j \in \mathcal{I}_g,\ i < j,\ y_i^k \neq y_j^k\}$
    if $\mathcal{P}_g^k \neq \emptyset$ **then**
      $\ell \leftarrow |\mathcal{P}_g^k|^{-1} \sum_{(i,j)\in\mathcal{P}_g^k} \max\big(0, m - s_{ij}^k(\hat{y}_i^k - \hat{y}_j^k)\big)$
      $A \leftarrow A + \ell$
    end if ▷ An empty pair set contributes zero
  end for
end for
return $A/(3b)$

Group offsets and score identifiability.

Theorem 1 (Group offsets and centered errors)**.** *For every source g and criterion k:*

1. *The group regression error decomposes as*

$$\frac{1}{n_g} \sum_{i \in \mathcal{I}_g} (e_i^k)^2 = (\bar{e}_g^k)^2 + V_g^k. \tag{9}$$

2. *The ranking loss is invariant to a common shift of the group's predictions, and every ranking subgradient sums to zero within each source-criterion group.*
3. *If the prediction coordinates are unconstrained, the hybrid loss has a unique minimizer for every fixed $\lambda \geq 0$, and that minimizer satisfies $\sum_{i\in\mathcal{I}_g}(\hat{y}_i^{k\star} - y_i^k) = 0$.*

Regression penalizes both the common offset and centered error. Ranking acts on within-source differences and therefore leaves the absolute offset unidentified. The mean-alignment statement concerns the empirical unconstrained prediction-space problem; it does not establish probabilistic calibration or held-out group means for a shared neural evaluator.

Margin compatibility.

Assume each criterion has at least one eligible training group, and let $\delta_{\min}$ be the smallest nonzero label gap over all eligible training pairs.

Proposition 2 (Compatible margins)**.** *In the unconstrained prediction-space problem, for any fixed $\lambda > 0$, the exact labels $\hat{y} = y$ uniquely minimize the hybrid loss if and only if $0 < m \leq \delta_{\min}$. In particular, throughout this interval the minimizer is y for every fixed $\lambda \geq 0$.*

Compatible margins preserve the exact empirical label target. If $m > \delta_{\min}$, at least one correctly predicted label pair is still penalized, and expanding the centered labels decreases the loss at $\hat{y} = y$.

Prediction-gradient alignment.

Let $q = \nabla_{\hat{y}} \mathcal{L}_{\text{reg}}$, with $q_i^k = 2(\hat{y}_i^k - y_i^k)/(3N)$, and let $r \in \partial_{\hat{y}} \mathcal{L}_{\text{rank}}(\hat{y})$. Subgradients account for the nondifferentiable hinge boundary. At a differentiable point, r is the ranking gradient.

Corollary 3 (Prediction-gradient alignment)**.** *If $0 < m \leq \delta_{\min}$, then $\langle q, r \rangle \geq 0$ for every prediction vector and every ranking subgradient. At a differentiable ranking point with $r \neq 0$, this inner product is strictly positive.*

Thus compatible active hinges align with regression in prediction space. This result describes the interaction of the losses independently of the encoder, alignment of gradients with respect to shared model parameters requires separate consideration.

Document-balanced ordering control.

For an eligible group, define the ordering error with predicted ties counted as errors:

$$D_g^k = \frac{1}{|\mathcal{P}_g^k|} \sum_{(i,j) \in \mathcal{P}_g^k} \mathbf{1}\left[s_{ij}^k (\hat{y}_i^k - \hat{y}_j^k) \leq 0 \right]. \tag{10}$$

Proposition 4 (Deterministic ordering bounds)**.** *For every evaluator, every eligible group, and $m > 0$,*

$$D_g^k \leq \min\left\{ 1, \frac{\ell_g^k}{m} \right\}, \qquad D_g^k \leq \frac{n_g^2 V_g^k}{|\mathcal{P}_g^k| \delta_{\min}^2}. \tag{11}$$

Extend D_g^k by zero when $\mathcal{P}_g^k = \emptyset$. The unconditional criterion- and document-balanced ordering error satisfies

$$\mathcal{D} = \frac{1}{3G} \sum_{k \in \mathcal{K}} \sum_{g \in \mathcal{G}} D_g^k \leq \min\left\{ 1, \frac{\mathcal{L}_{\text{rank}}}{m} \right\}.$$

An inversion or predicted tie incurs hinge loss at least m, giving the first bound. The second shows that centered squared regression error also controls ordering. Ranking adds a one-sided penalty on insufficient signed separation, including correctly

ordered pairs whose gap is below m. The macro bound retains the all-source denominator of the training objective; Appendix B.5 gives the corresponding normalization for evaluation restricted to eligible sources. These criterion-wise bounds do not directly bound weighted Overall ordering or Kendall's tie-normalized τ_b.

Role of the fixed ranking coefficient.

Let Π subtract the mean within each source-criterion block. Theorem 1 implies

$$\langle (I - \Pi)q, r \rangle = 0, \qquad \langle q, r \rangle = \langle \Pi q, r \rangle. \tag{12}$$

The hybrid prediction subgradient therefore separates into

$$q + \lambda r = (I - \Pi)q + (\Pi q + \lambda r).$$

Regression supplies the offset correction, while a fixed λ scales the additional ranking signal in the centered subspace. Appendix B.6 derives this geometry and the associated local descent properties. Statements about an optimum use unconstrained prediction coordinates; Appendix B.7 specifies the limits of transferring them to shared-parameter optimization and population targets.

3.4 Efficient computation and language portability

3.4.1 Shared computation and pairwise overhead

A shared source-summary representation allows all criterion scores to be obtained from one non-autoregressive encoder pass. For an encoder-based realization with a shared linear projection of $z_i \in \mathbb{R}^d$ to width h, followed by GELU, dropout, and scalar linear outputs, the projection and output costs can be expressed independently of the chosen encoder. These computations require $O(dh + |\mathcal{K}|h)$ operations per document-summary instance; the contextual encoding and pooling are shared by all criteria.

During training, ranking reuses these predictions. For a micro-batch of complete source groups, the number of eligible hinge terms satisfies

$$\sum_{g \in \mathcal{G}_\mathcal{B}} \sum_{k \in \mathcal{K}} |\mathcal{P}_g^k| \le |\mathcal{K}| \sum_{g \in \mathcal{G}_\mathcal{B}} \binom{n_g}{2}.$$

Hence pair construction and loss evaluation require at most $O(|\mathcal{K}| \sum_{g \in \mathcal{G}_\mathcal{B}} n_g^2)$ additional score-level operations, with no additional encoder evaluations. The pairwise loss is confined to training, inference evaluates Eqs. (2)–(3) once per document-summary instance. Actual latency and memory depend on the encoder, context budget, and execution settings, and are measured in Section 4.4.7.

3.4.2 Adaptation to other languages

The grouping rule, criterion-specific supervision, and hybrid objective contain no Vietnamese-specific operation. A target-language realization can use a suitable

encoder/tokenizer and criterion-labeled source-summary groups while retaining the same learning formulation. Margin compatibility is expressed through the target data's nonzero label gaps, so it can be evaluated on the chosen rubric scale. Section 4 evaluates Vietnamese realizations; evaluation in additional languages is a natural extension.

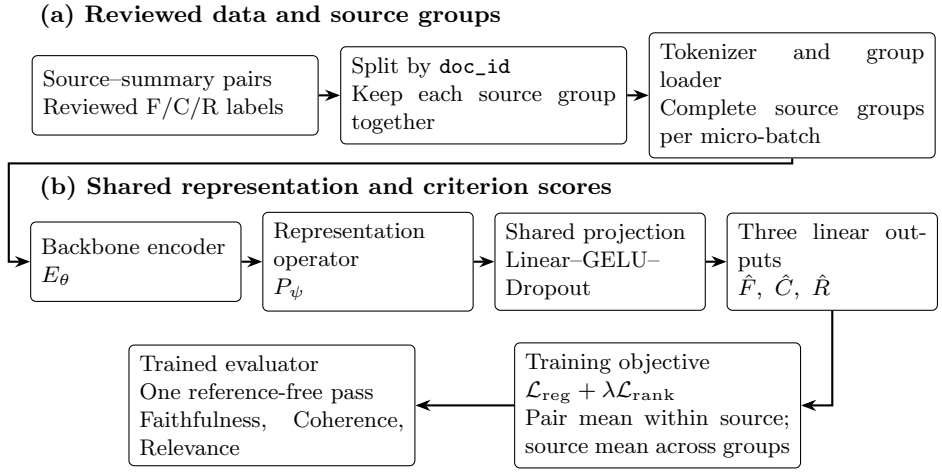


Fig. 1 Overview of the EViSE pipeline, from grouped data preparation to hybrid multi-criteria optimization and single-pass criterion prediction.

4 Experimental Setup and Evaluation

We instantiate EViSE for Vietnamese and evaluate criterion-level agreement, within-source ordering, domain shift, computational cost, and frozen-reward reuse. Both realizations share the main benchmarks and matched internal profile. Objective ablation, qualitative cases, external profiling, and independent blind OOD validation focus on the compact realization.

4.1 Vietnamese evaluation resource

4.1.1 Resource construction and annotation reliability

The evaluation resource comprises 13,476 VnExpress articles published during 2022-2024, covering 15 major topics. It is organized for evaluator learning: each source is associated with six candidate summaries and reviewed F/C/R ratings, yielding 80,856 document-summary evaluations. This annotation unit supports absolute criterion scoring and comparisons among candidates of the same source. Table 1 summarizes the source distribution.

Table 1 Distribution of source articles by topic in the EViSE dataset.

Topic	# Articles	Ratio (%)
Business	1,416	10.5
Digital	1,021	7.6
Education	623	4.6
Entertainment	2,038	15.1
Health	419	3.1
Law	86	0.6
Life	424	3.1
News	712	5.3
Opinions	748	5.5
Real Estate	1,181	8.8
Science	1,004	7.5
Sports	1,948	14.5
Travel	724	5.4
Vehicles	452	3.4
World	680	5.0
Total	**13,476**	**100.0**

Candidate summary generation

Each document is summarized by six heterogeneous systems: GPT-4o[1], Gemini 2.0 Flash[2], LLaMA-3.2 1B, LLaMA-3.2 3B, LLaMA-3.1 8B, and a ViT5-large model (Phan et al. 2022) fine-tuned on a filtered subset of VNDS (Nguyen et al. 2019). This design provides a broad quality range and multiple candidate summaries under the same source context, enabling both pointwise criterion supervision and within-document comparison.

LLM-assisted labeling with full-corpus human review

Evaluation prompt optimization follows the *LLMs-as-Optimizers* (OPRO) paradigm (Yang et al. 2024). A calibration set of 100 summaries generated by the six systems is first rated by human annotators using the F/C/R rubric in Section 3.1. Gemini 2.0 Flash, the same Gemini model used for candidate-summary generation, is then used in iterative prompt refinement. Cohen's κ is computed separately for F, C, and R, and their mean is used for prompt selection. The selected prompt achieves a mean Cohen's κ of 0.78 and requests criterion-specific scores and justifications in a structured format.

Gemini 2.0 Flash subsequently assigns initial 1-5 Likert ratings to all document-summary pairs. The full corpus is then reviewed by 12 trained volunteer annotators organized into six two-person groups. Each assigned instance is jointly reviewed by both members of one group: the two annotators discuss the initial F/C/R ratings and reach consensus on whether each criterion score should be retained or revised.

[1] https://openai.com/index/hello-gpt-4o/
[2] https://ai.google.dev/gemini-api/docs/models

The resulting consensus scores constitute the final human-reviewed labels. Before the large-scale review, annotators calibrate the rubric on representative examples covering multiple topics and quality levels and agree on correction guidelines.

Approximately 10% of the initial labels are revised during full-corpus review. Frequent revisions concern cases in which different quality dimensions interact, particularly summaries that remain factually supported but omit salient information. Such cases are scored according to the criterion definitions above rather than being forced into a single notion of correctness. The initial Gemini ratings serve as annotation proposals. Only the final consensus scores are used as targets for evaluator training, validation, model selection, and News test evaluation. For a final rubric rating $a_i^k \in \{1, 2, 3, 4, 5\}$, the common score scale is obtained by

$$y_i^k = \frac{a_i^k - 1}{4} \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{13}$$

One normalized score unit corresponds to four points on the original rubric. Evaluator MAEs use the normalized scale. The annotation audit and downstream rubric evaluation explicitly use original 1–5 points.

Annotation reliability audit

Because the full-corpus reviewers corrected LLM-initialized scores rather than scoring every instance de novo, we conducted a separate blind audit to assess both independent rubric reproducibility and whether full-corpus human review changed agreement with independently obtained ratings. The audit therefore supports two complementary analyses: agreement between the two blind human annotators, and a before–after comparison of the initial Gemini proposals and the final reviewed corpus labels against those same independent ratings.

The audit subset contains 450 document-summary pairs sampled from the full annotated corpus, with 150 pairs from each quality bin: low $[0.0, 0.5)$, mid $[0.5, 0.8)$, and high $[0.8, 1.0]$. Quality was computed from the normalized final labels as

$$\text{Overall} = 0.5F + 0.3R + 0.2C. \tag{14}$$

Each sampled pair has a unique `doc_id` to avoid over-representing multiple summaries from the same source. The independently produced audit ratings were used only for reliability estimation and did not alter the corpus labels used for evaluator training, validation, model selection, or final test reporting.

Two additional annotators, denoted A and B, who were not involved in the 12-annotator full-corpus review independently re-scored all 450 pairs using the same 1–5 Faithfulness, Coherence, and Relevance rubric. They had no access to the original F/C/R labels, Overall scores, quality bins, or split metadata. Agreement between A and B was computed before adjudication.

Table 2 shows strong blind rubric reproducibility: quadratic weighted Cohen's κ ranges from 0.805 to 0.887 and adjacent agreement from 91.1% to 98.4%. Seventy of 450 instances (15.6%) contain at least one criterion differing by two or more points and are adjudicated. These A-B statistics describe rubric reproducibility.

Table 2 Inter-annotator agreement before adjudication on the 450-instance reliability audit subset.

Criterion	N	Exact agreement	Adjacent agreement	Quadratic weighted κ	MAD
Faithfulness	450	69.1%	98.4%	0.887	0.324
Coherence	450	66.4%	94.9%	0.824	0.387
Relevance	450	62.0%	91.1%	0.805	0.469

Agreement before and after full-corpus review.

On the same 450 pairs, we compare both the initial Gemini 2.0 Flash ratings before full-corpus review and the final human-reviewed corpus labels with the independent pre-adjudication ratings from A and B. Across the six criterion–annotator comparisons, quadratic-weighted Cohen's κ increases from 0.774–0.819 before review to 0.783–0.821 after review. Against the blind human mean $H = (A + B)/2$, Spearman correlations increase from 0.817 to 0.837 for Faithfulness, from 0.846 to 0.860 for Coherence, and from 0.850 to 0.857 for Relevance. The corresponding MAEs decrease from 0.476 to 0.460, from 0.422 to 0.409, and from 0.443 to 0.434 points on the original 1–5 scale. Thus, all criterion-level point estimates move in the direction of closer agreement after review. Because A and B constitute an independent reliability audit, these changes are interpreted as evidence of improved alignment with independent rubric judgments. Appendix D reports the full criterion-level results, confidence intervals, and paired before–after comparisons.

Aggregate scalar score for secondary analyses.

EViSE primarily predicts and is evaluated on separate criterion scores. For analyses requiring one scalar, we use the fixed task-motivated weighting

$$\text{Overall} = 0.5F + 0.3R + 0.2C. \tag{15}$$

This gives greatest weight to source support while retaining contributions from salient-content selection and discourse quality. A development comparison examined the separation induced by several candidate weightings. Because no independent holistic Overall label was collected, this comparison does not identify human preferences over the three criteria. Overall remains a secondary scalarization, criterion-wise reporting is primary. Appendix A.4 explains the limitation of best-worst separation as a weighting criterion.

Quality-filtered reference summary construction

After full-corpus human review, a quality-filtered reference summary is selected from the six candidate summaries of each source document. Selection uses the final human-reviewed criterion ratings on the original 1-5 scale: the candidate with the highest Overall score is retained only when its Overall score is at least 4.0, where Overall uses the same $0.5F + 0.3R + 0.2C$ weighting. This procedure yields reference summaries for 13,468 of the 13,476 source documents. The eight documents for which no

Table 3 Document-level data partition under a shared quality-binning scheme. Each document is associated with six candidate-summary instances.

Partition	Docs.	Instances	Document-level quality composition	Quality bins
Full corpus	13,476	80,856	low: 4,652, mid: 4,452, high: 4,372	low: $[0.0, 0.5)$, mid: $[0.5, 0.8)$, high: $[0.8, 1.0]$
Train + Validation	12,128	72,768	same document-level binning scheme	
Test	1,348	8,088		

candidate reaches the threshold are excluded from reference-based evaluation. When ROUGE-Lsum or BERTScore-F1 is computed for candidates drawn from the same six-summary pool, the candidate selected as the reference is excluded from evaluation against itself, thereby avoiding a trivially maximal self-reference score. These quality-filtered summaries are used only by reference-based analyses, while EViSE itself does not require a reference summary at inference time.

4.1.2 Leakage-safe split and quality stratification

The dataset is split by `doc_id`, keeping all six summaries of a source in the same partition. This prevents document-level leakage and preserves the grouped structure required by within-document ranking.

For document-level stratification, we define the quality score of source document as the mean Overall score of its six candidate summaries. Using the Overall score above, documents are assigned to low $[0.0, 0.5)$, mid $[0.5, 0.8)$, or high $[0.8, 1.0]$ quality bins. The final partition contains 12,128 documents (72,768 instances) for training and validation and 1,348 documents (8,088 instances) for the held-out test set. The training-and-validation portion is further divided by `doc_id` using an approximately 90:10 train-to-validation ratio, yielding 10,915 training documents and 1,213 validation documents. Table 3 summarizes the resulting document-level partition.

4.2 Model instantiations and experimental settings

4.2.1 Vietnamese encoder instantiations

We instantiate the framework for Vietnamese using two contrasting encoders. EViSE-ViDeBERTa uses the compact Vietnamese-specialized ViDeBERTa encoder (Tran et al. 2023) with masked mean pooling. EViSE-mmBERT uses the larger multilingual mmBERT encoder (Marone et al. 2026) with concatenated CLS and masked-mean pooling. The latter representation gives the shared projection access to both sequence-level interactions and information distributed across non-padding tokens. Both realizations retain the shared representation, criterion outputs, and grouped learning objective defined in Section 3.

The encoders, pooling operators, tokenizers, and context budgets vary jointly, giving two concrete accuracy-efficiency operating points. Appendix A.5 reports the

input budgets, parameter counts, and implementation configurations. Their measured quality and computational cost are compared in Section 4.4.6.

4.2.2 Training and ablation settings

Both realizations use a shared projection of width $h = 256$ and the fixed objective settings $(m, \lambda) = (0.05, 0.35)$. Equation (13) guarantees a minimum possible nonzero label gap of 0.25, so the chosen margin is compatible with Proposition 2. The ranking weight was selected during EViSE-ViDeBERTa development and transferred unchanged to EViSE-mmBERT. Appendices A.2 and A.3 document these choices. Appendix A.5 gives the optimizer, precision, checkpoint-selection, and other training settings.

Each micro-batch contains all six candidate summaries of one source ($b = 1$). Uniform source sampling is also compatible with the instance-averaged regression loss because all source groups have the same size. Accumulating two micro-batches for ViDeBERTa and eight for mmBERT gives 12 and 48 instances per update, respectively. Complete micro-batch losses are averaged according to Section 3.2, retaining zero ranking contributions from empty pair sets.

The regression-only control sets $\lambda = 0$ while keeping the ViDeBERTa encoder, pooling, projection, outputs, split, and training protocol fixed. Three scalar linear outputs are algebraically equivalent to one three-output linear mapping, so this ablation isolates the ranking contribution.

4.2.3 Evaluation benchmarks and computational profiling

Both EViSE-ViDeBERTa and EViSE-mmBERT are evaluated on the leakage-safe News test set of 1,348 documents and 8,088 summaries and on the same out-of-domain IT-textbook benchmark of 900 summaries generated for 150 source passages. The IT passages span 13 subject areas and 118 book-level sources. Each source is summarized by GPT-4o, Gemini 2.0 Flash, LLaMA-3.2 1B, LLaMA-3.2 3B, LLaMA-3.1 8B, and domain-adapted ViT5, and the resulting summaries are reviewed by six IT faculty using the same F/C/R rubric. The downstream application uses Qwen3-0.6B under the Vietnamese-news GRPO protocol of Tran et al. (2026) and evaluates the resulting policy on 625 Vietnamese WikiLingua examples without further WikiLingua training. The added run replaces the frozen reward evaluator with EViSE-mmBERT while retaining the policy initialization, training data and budget, reward-construction settings, decoding, and rubric-review protocol.

For OOD uncertainty, the 150 IT sources are resampled as clusters in 2,000 bootstrap replicates. A separate blind OOD subset contains one candidate from each of 150 unique sources, balanced across the six generators (25 each) and three quality bins (50 each). Two additional IT faculty independently score this subset without access to prior labels, generator identities, or model predictions. Their averaged ratings are used only for blind validation.

Computational-efficiency profiling

The external profile uses 1,344 Vietnamese pairs on one Tesla T4 and records runtime, latency, throughput, peak memory, parameter count, output dimensionality, and reference dependence for EViSE-ViDeBERTa, ROUGE-Lsum, and PhoBERT-large BERTScore-F1. UniEval is profiled on 1,344 CNN/DailyMail examples only as a contextual multidimensional reference because the language and data distribution are not matched.

The two EViSE instantiations are additionally profiled on the same 1,344 Vietnamese pairs using FP32, batch size 4, 32 warm-up pairs, and three independent full runs; median runtime is reported. This matched profile is used for the internal accuracy-efficiency comparison.

4.2.4 Baselines

The regression-only control is defined in Section 4.2.2.

ROUGE-Lsum and PhoBERT-large BERTScore-F1 provide reference-based lexical and semantic baselines (Lin 2004; Zhang et al. 2020). Their reference is the quality-filtered candidate defined in Section 4.1.1, with self-comparisons removed. The checkpoint `MoritzLaurer/mDeBERTa-v3-base-mnli-xnli` provides a verifier-style Faithfulness baseline (He et al. 2023; Conneau et al. 2018; Laurer 2023). For each summary sentence h_t, TF-IDF retrieves the three most similar source sentences, forming $\mathcal{R}_t$. Each retrieved sentence d is the NLI premise and h_t the hypothesis. With T summary sentences, the consistency score is

$$b(D, S) = \frac{1}{T} \sum_{t=1}^{T} \max_{d \in \mathcal{R}_t} p_{\text{ent}}(d, h_t) \in [0, 1].$$

Here p_{ent} is the entailment probability. This maximum-then-mean aggregation follows the zero-shot scheme of Laban et al. (2022), restricted here to retrieved evidence. UniEval (Zhong et al. 2022) is used only for contextual computational profiling.

4.3 Metrics and reporting protocol

Pointwise agreement is measured with Spearman's ρ, Pearson's r, and MAE (Spearman 1904; Pearson 1895). All evaluator agreement and ordering metrics use raw linear-head predictions, without clipping or rounding. Rubric labels are normalized as $y = (a - 1)/4$, where $a \in \{1, 2, 3, 4, 5\}$. Evaluator MAE uses this normalized scale. Ninety-five-percent confidence intervals use grouped bootstrap resampling over `doc_id`, with paired document-level resampling for direct comparison of the two News instantiations. Criterion-level reporting is primary; secondary aggregate analyses use

$$\text{Overall} = 0.5F + 0.3R + 0.2C. \tag{16}$$

Within-source ordering uses per-document Kendall's τ_b, pairwise accuracy, and tie-aware Top-1 hit rate. Pairwise accuracy is computed within each source using

only unequal human-score pairs, then averaged equally over sources with at least one eligible pair:

$$\mathrm{PA}^k = \frac{1}{|\mathcal{G}_k^+|} \sum_{g \in \mathcal{G}_k^+} \mathrm{PA}_g^k, \qquad \mathcal{G}_k^+ = \{g : |\mathcal{P}_g^k| > 0\}.$$

The same reduction applies to Overall ordering. Predicted ties count as errors in pairwise accuracy. Kendall's τ_b accounts for label and prediction ties. Top-1 accepts a selected candidate whenever its human score equals the source maximum. Criterion specificity uses the 3×3 Pearson matrix between predicted and human F/C/R.

For downstream RL evaluation, the same weights are applied to rubric scores on the original 1-5 scale:

$$\mathrm{Overall}_{\mathrm{rubric}} = 0.5F_{\mathrm{rubric}} + 0.3R_{\mathrm{rubric}} + 0.2C_{\mathrm{rubric}}. \tag{17}$$

For the downstream application, a separate LLM-assisted rubric evaluator produces initial F/C/R ratings, followed by paired human review and adjudication under the protocol of Tran et al. (2026). Equation (17) therefore measures reviewed output quality on $[1, 5]$, rather than the reward returned by the frozen training evaluator.

The experiments address five questions: (Q1) does combining cardinal and ordinal supervision improve evaluator agreement relative to regression-only learning. (Q2) does EViSE preserve criterion-specific F/C/R behavior and align with human judgments relative to established automatic metrics. (Q3) does EViSE retain agreement with human judgments under domain shift. (Q4) does the framework remain effective across substantially different encoder configurations, and what accuracy-efficiency trade-offs arise. And (Q5) how does substituting the frozen reward evaluator affect downstream quality under a fixed GRPO protocol?

4.4 Results

4.4.1 Objective ablation: regression-only versus hybrid supervision

Table 4 shows a consistent benefit from source-conditioned ranking under the fixed EViSE-ViDeBERTa architecture. Relative to regression only, Spearman increases by 0.045-0.061 and MAE decreases by 0.016-0.018 across F/C/R. The grouped-bootstrap intervals quantify held-out document uncertainty, not training-seed variability.

4.4.2 Within-document ordering and criterion specificity

Table 5 shows that source-conditioned ordering is preserved by both instantiations. EViSE-ViDeBERTa reaches Overall $\tau_b = 0.631$ and pairwise accuracy $= 0.835$, while EViSE-mmBERT reaches 0.686 and 0.864, respectively. This extends the grouped-ordering evidence beyond the compact backbone.

Table 6 shows diagonally dominant correlation patterns for both EViSE instantiations. For EViSE-ViDeBERTa, the criterion-matched correlations are $r = 0.799$, 0.733, and 0.838 for F, C, and R, respectively, while EViSE-mmBERT shows the same pattern with $r = 0.851$, 0.790, and 0.867. These results support criterion specificity across both instantiations.

Table 4 Regression-only ablation on the News test set. Both conditions use the same ViDeBERTa configuration; brackets show document-grouped 95% confidence intervals.

Objective	Metric	F	C	R
Regression only ($\lambda = 0$)	MAE	0.120 [0.118, 0.122]	0.121 [0.118, 0.123]	0.127 [0.124, 0.129]
	Pearson r	0.770 [0.757, 0.783]	0.700 [0.680, 0.718]	0.813 [0.802, 0.823]
	Spearman ρ	0.646 [0.632, 0.661]	0.570 [0.556, 0.585]	0.674 [0.660, 0.688]
Hybrid ($\lambda = 0.35$)	MAE	**0.104** [0.102, 0.106]	**0.104** [0.102, 0.107]	**0.109** [0.106, 0.111]
	Pearson r	**0.799** [0.788, 0.811]	**0.733** [0.716, 0.749]	**0.838** [0.826, 0.848]
	Spearman ρ	**0.697** [0.684, 0.711]	**0.615** [0.602, 0.628]	**0.735** [0.722, 0.748]

Table 5 Within-document ordering for both EViSE instantiations on the 1,348-document News test set. Brackets show document-bootstrap 95% confidence intervals; Top-1 is tie-aware.

Instantiation	Target	Kendall τ_b	Pairwise acc.	Top-1
EViSE-ViDeBERTa	Faithfulness	0.614 [0.599, 0.628]	0.870 [0.861, 0.878]	0.768 [0.745, 0.791]
	Coherence	0.551 [0.537, 0.566]	0.860 [0.851, 0.870]	0.889 [0.872, 0.905]
	Relevance	0.659 [0.645, 0.672]	0.892 [0.884, 0.900]	0.783 [0.761, 0.805]
	Overall	0.631 [0.617, 0.646]	0.835 [0.828, 0.843]	0.657 [0.632, 0.682]
EViSE-mmBERT	Faithfulness	0.684 [0.671, 0.696]	0.908 [0.901, 0.916]	0.837 [0.817, 0.856]
	Coherence	0.631 [0.620, 0.642]	0.905 [0.898, 0.912]	0.950 [0.938, 0.960]
	Relevance	0.694 [0.681, 0.706]	0.910 [0.902, 0.917]	0.843 [0.823, 0.861]
	Overall	0.686 [0.673, 0.699]	0.864 [0.858, 0.871]	0.727 [0.701, 0.751]

4.4.3 Comparison with traditional automatic metrics

Table 7 shows that both EViSE instantiations align more strongly with human-derived Overall scores than ROUGE-Lsum and BERTScore-F1 on the same 6,740-pair eligible evaluation subset. EViSE-mmBERT achieves the strongest observed agreement ($\rho = 0.787$, $r = 0.870$), while EViSE-ViDeBERTa provides a lower-cost operating point. The document-grouped confidence intervals show that this pattern is stable under resampling, while the comparison with the traditional metrics is interpreted descriptively.

4.4.4 Criterion-level and cross-domain evaluation

Table 8 provides matched in-domain evidence for both framework instantiations. EViSE-mmBERT improves all three criterion correlations, while EViSE-ViDeBERTa remains competitive with lower computational cost. The paired Overall comparison and cost trade-off are summarized separately in Section 4.4.6.

Table 6 Pearson correlation between predicted and human-reviewed F/C/R criteria for both EViSE instantiations on the News test set. The largest value in each prediction row is shown in bold.

Instantiation	Prediction	Human F	Human C	Human R
EViSE-ViDeBERTa	Predicted F	**0.799**	0.613	0.553
	Predicted C	0.619	**0.733**	0.448
	Predicted R	0.522	0.421	**0.838**
EViSE-mmBERT	Predicted F	**0.851**	0.672	0.541
	Predicted C	0.677	**0.790**	0.420
	Predicted R	0.557	0.452	**0.867**

Table 7 Correlation with human-derived Overall on the common eligible News evaluation subset ($n = 6{,}740$) used for the reference-based comparison after self-reference removal. Brackets show document-grouped 95% confidence intervals.

Metric / Model	Spearman (ρ)	Pearson (r)
EViSE-ViDeBERTa (Overall)	0.721 [0.706, 0.734]	0.827 [0.816, 0.837]
EViSE-mmBERT (Overall)	**0.787** [0.776, 0.798]	**0.870** [0.862, 0.878]
ROUGE-Lsum (F1)	0.459 [0.439, 0.478]	0.580 [0.560, 0.599]
BERTScore-F1	0.357 [0.335, 0.380]	0.506 [0.467, 0.560]

Table 8 Criterion-level performance of both EViSE instantiations on the common News test set. Brackets show document-grouped 95% confidence intervals.

Instantiation	Criterion	Spearman ρ	Pearson r	MAE
EViSE-ViDeBERTa	F	0.697 [0.684, 0.711]	0.799 [0.788, 0.811]	0.104 [0.102, 0.106]
	C	0.615 [0.602, 0.628]	0.733 [0.716, 0.749]	0.104 [0.102, 0.107]
	R	0.735 [0.722, 0.748]	0.838 [0.826, 0.848]	0.109 [0.106, 0.111]
EViSE-mmBERT	F	0.777 [0.766, 0.788]	0.851 [0.841, 0.859]	0.096 [0.094, 0.098]
	C	0.712 [0.700, 0.723]	0.790 [0.774, 0.804]	0.092 [0.090, 0.094]
	R	0.785 [0.774, 0.795]	0.867 [0.858, 0.875]	0.104 [0.102, 0.106]

Table 9 shows that both EViSE instantiations retain substantial agreement under IT-textbook domain shift. EViSE-ViDeBERTa reaches Overall $\rho = 0.691$, $r = 0.829$, and MAE = 0.107, while EViSE-mmBERT reaches $\rho = 0.720$, $r = 0.850$, and MAE = 0.085. The higher-capacity instantiation therefore retains its numerical advantage under domain shift, although the overlapping source-cluster confidence intervals do not by themselves establish a statistically significant difference between the two models. For the compact EViSE-ViDeBERTa model, the independent 150-pair blind subset gives Overall $\rho = 0.620$, $r = 0.799$, and MAE = 0.108, providing an additional robustness check without implying generalization to arbitrary domains.

Table 9 Cross-domain agreement of both EViSE instantiations on 900 IT-textbook summaries. Scores are normalized to $[0, 1]$. Overall is $0.5F + 0.3R + 0.2C$. Brackets show source-cluster 95% confidence intervals.

Instantiation	Criterion	Spearman ρ	Pearson r	MAE
EViSE-ViDeBERTa	Faithfulness	0.586 [0.532, 0.637]	0.702 [0.652, 0.747]	0.152 [0.145, 0.159]
	Coherence	0.648 [0.607, 0.686]	0.727 [0.696, 0.757]	0.123 [0.117, 0.131]
	Relevance	0.632 [0.588, 0.673]	0.736 [0.706, 0.766]	0.146 [0.139, 0.154]
	Overall	0.691 [0.651, 0.730]	0.829 [0.804, 0.854]	0.107 [0.102, 0.112]
EViSE-mmBERT	Faithfulness	**0.596** [0.543, 0.646]	**0.742** [0.708, 0.772]	**0.127** [0.121, 0.133]
	Coherence	**0.678** [0.636, 0.716]	**0.757** [0.722, 0.787]	**0.122** [0.116, 0.129]
	Relevance	**0.642** [0.594, 0.685]	**0.766** [0.735, 0.794]	**0.135** [0.128, 0.143]
	Overall	**0.720** [0.678, 0.758]	**0.850** [0.831, 0.868]	**0.085** [0.081, 0.089]

Table 10 Faithfulness agreement on the full News test set ($n = 8{,}088$). Brackets show document-grouped 95% confidence intervals.

Model / Metric	Spearman (ρ)	Pearson (r)
NLI consistency baseline	0.254 [0.2321, 0.2751]	0.246 [0.2227, 0.2688]
EViSE-ViDeBERTa (Faithfulness)	0.697 [0.6839, 0.7111]	0.799 [0.7877, 0.8108]
EViSE-mmBERT (Faithfulness)	**0.777** [0.766, 0.788]	**0.851** [0.841, 0.859]

Table 11 Overall accuracy-efficiency trade-off across EViSE instantiations. Quality uses the common 8,088-pair News test set; cost uses the matched 1,344-pair Tesla-T4 FP32 profile.

Instantiation	Overall ρ	Overall r	Overall MAE	Params. (M)	ms/pair	Pairs/s
EViSE-ViDeBERTa	0.737	0.833	0.088	**183.952**	**55.198**	**18.117**
EViSE-mmBERT	**0.806**	**0.876**	**0.079**	307.334	94.502	10.582

4.4.5 Faithfulness comparison with a verifier-style NLI baseline

Table 10 shows that both EViSE instantiations substantially exceed the verifier-style NLI baseline on the same full News test set. EViSE-ViDeBERTa reaches $\rho = 0.697$ and $r = 0.799$, while EViSE-mmBERT reaches 0.777 and 0.851. For the primary model, paired grouped-bootstrap differences over NLI are 0.4432 in Spearman (95% CI $[0.4202, 0.4674]$) and 0.5534 in Pearson (95% CI $[0.5298, 0.5772]$).

Table 12 Computational profile for 1,344 instances. EViSE-ViDeBERTa, ROUGE-Lsum, and BERTScore-F1 use the same Vietnamese set; UniEval uses CNN/DailyMail and is included only as a contextual profile.

Method	Outputs	Ref.	Params. (M)	Total (s)	ms/ pair	Pairs/ s	Peak RAM (MB)	Peak GPU (MB)
ROUGE-Lsum	1	Yes	-	6.069	4.516	221.449	1,024.605	0
BERTScore-F1	1	Yes	369.163	107.735	80.160	12.475	3,397.836	1,496.009
EViSE-ViDeBERTa	**3**	**No**	**183.952**	**72.383**	**53.856**	**18.568**	**1,762.336**	**1,417.447**
UniEval	3	Criterion-dependent	783.093	2,954.236	2,198.092	0.455	2,335.113	6,506.148

4.4.6 Alternative instantiation and accuracy-efficiency trade-off

Table 11 shows the framework-level accuracy-efficiency trade-off. EViSE-mmBERT raises Overall ρ from 0.737 to 0.806 and reduces MAE from 0.088 to 0.079; paired document bootstrap gives $\Delta\rho = 0.0686$ [0.0595, 0.0783], $\Delta r = 0.0428$ [0.0375, 0.0490], and an MAE reduction of 0.0090 [0.0077, 0.0102]. The higher agreement comes with about 1.67× more parameters and 1.71× higher matched latency. Because encoder, pooling, and context budget change jointly, these are two validated operating points rather than a causal single-factor comparison.

4.4.7 Computational efficiency relative to alternative evaluators

Table 12 places the compact operating point in computational context. ROUGE-Lsum is fastest but returns one reference-dependent lexical score, whereas EViSE-ViDeBERTa returns three reference-free criterion scores in one pass at 53.856 ms/pair and 18.568 pairs/s. On the matched Vietnamese profile it is about 1.49× faster than PhoBERT-large BERTScore-F1 while using a smaller checkpoint. UniEval remains an unmatched contextual profile.

4.4.8 Illustrative qualitative analysis

This qualitative analysis focuses on the primary EViSE-ViDeBERTa instantiation to illustrate representative successes and remaining failure modes. It is intended as a diagnostic analysis.

Table 13 illustrates three complementary behaviors: A7164 is under-rewarded by lexical overlap despite largely source-supported content, A12313 retains high BERTScore despite factual and structural errors, and A329 exposes a remaining weakness in speaker attribution and entity-role binding. Full texts are provided in Appendix A.6.

4.4.9 Reward-evaluator substitution in downstream GRPO

A frozen evaluator can supply criterion scores to a summarization policy during RL. The compact checkpoint was used in this role as MultiEvalSumViet2 by Tran et al. (2026). We use that application as the reference configuration and examine substitution of the evaluator with EViSE-mmBERT. In both configurations, the quality

Table 13 Illustrative diagnostic cases of EViSE-ViDeBERTa.

Case	Doc ID	Human-reviewed (F/C/R)	Pred. (F/C/R)	Metric signal
ROUGE-Lsum under-rewards	A7164	0.750 / 1.000 / 0.750	0.757 / 1.000 / 0.748	ROUGE-Lsum = 0.274
BERTScore overestimates	A12313	0.250 / 0.250 / 0.500	0.224 / 0.275 / 0.469	BERTScore-F1 = 0.875
Difficult error case	A329	0.500 / 0.750 / 0.500	0.757 / 0.971 / 0.740	

Table 14 Frozen-evaluator substitution under the Qwen3-0.6B GRPO protocol, evaluated on 625 Vietnamese WikiLingua examples. Overall is human-reviewed rubric quality on the original 1-5 scale. The ViDeBERTa reference value is previously reported (Tran et al. 2026); the mmBERT value is from this study.

Frozen reward evaluator	Overall
EViSE-ViDeBERTa	4.45
EViSE-mmBERT	4.46

component supplied to the existing reward-construction module is

$$r_{\text{qual}}(x, y) = 0.5\hat{F}(x, y) + 0.3\hat{R}(x, y) + 0.2\hat{C}(x, y). \tag{18}$$

The evaluator remains frozen while GRPO updates the Qwen3-0.6B policy. The training and reward-shaping settings follow the reference protocol. The evaluation uses the same 625-example Vietnamese WikiLingua set in zero-shot transfer from news. The replacement run changes the evaluator checkpoint while keeping the remaining protocol fixed.

Table 14 reports the human-reviewed rubric aggregate from Eq. (17). The ViDeBERTa value is taken from the previously reported reference configuration (Tran et al. 2026), while the mmBERT value comes from the matched evaluator-substitution run in this study.

The two configurations yield similar reviewed quality in this setting. The result suggests that the higher-capacity evaluator can occupy the same frozen-reward role, its stronger pointwise agreement does not imply a commensurate increase in policy quality under a fixed RL budget.

Summary of research questions.

The results support ordinal supervision under a fixed architecture (Q1), criterion specificity and stronger agreement than traditional metrics (Q2), agreement under

IT-textbook domain shift (Q3), and transfer across two encoder configurations with an accuracy-efficiency trade-off (Q4). Evaluator substitution yields similar reviewed downstream quality under matched GRPO (Q5).

5 Discussion

The grouped analysis explains how the two supervision terms act on evaluator errors. Regression penalizes both a source-criterion offset and deviations around that offset, while ranking is invariant to the offset and directly penalizes insufficient signed separation. Compatible margins preserve the exact empirical label target in the unconstrained prediction-space problem, and the ranking loss bounds document-balanced ordering error. The fixed-architecture ablation provides the complementary empirical evidence: adding ranking raises criterion-level Spearman by 0.045-0.061 and lowers MAE by 0.016-0.018 (Table 4). The theoretical results characterize loss geometry; the measured improvement concerns a finite shared evaluator trained with a fixed coefficient.

The gradient analysis further shows that, with compatible margins, regression and ranking have a nonnegative inner product in prediction space. Ranking therefore acts on adverse within-source score-difference errors while leaving the direct group-offset correction to regression.

The two instantiations expose an accuracy-efficiency trade-off. EViSE-mmBERT raises Overall News ρ from 0.737 to 0.806, while matched latency rises from 55.2 to 94.5 ms/pair and parameter count from 184.0M to 307.3M (Table 11). Encoder, pooling, and context length change together, so this comparison describes two operating points. Their common $(m, \lambda) = (0.05, 0.35)$ configuration shows practical transfer of the fixed objective settings.

On the IT-textbook benchmark, Overall Spearman correlations of 0.691 and 0.720 show substantial agreement under the evaluated domain shift (Table 9). The compact model's independent blind subset provides additional evidence, with Overall $\rho = 0.620$. In the WikiLingua application, replacing the frozen reward evaluator produces similar reviewed policy quality (Table 14). This is a downstream use of the evaluator whose outcome also depends on the policy, reward shaping, and training budget.

The study has several limits. The blind audit supports rubric reproducibility and reviewed-corpus agreement on a quality-stratified subset; it does not identify population-wide label error. Cross-domain evaluation covers one technical-educational domain, and independent blind OOD validation is available only for the compact model. The theoretical mean-alignment result assumes free prediction coordinates; shared parameters can couple offset and ordering updates, and the finite-sample target-compatibility result is not a population-consistency theorem. Overall weights remain a secondary task-motivated scalarization without independently elicited holistic labels. Multilingual effectiveness remains to be evaluated, while downstream evidence for the second instantiation is limited to one matched GRPO configuration.

6 Conclusion

EViSE implements a criterion-wise, group-balanced cardinal-ordinal objective in a shared evaluator for Faithfulness, Coherence, and Relevance. The analysis separates group offsets from centered errors, characterizes compatible margins and prediction-gradient alignment, and links ranking loss to document-balanced ordering error. Under fixed training weights, the objective ablation supports the contribution of ranking, and two instantiations provide complementary accuracy-efficiency operating points with Overall News $\rho = 0.737$ and 0.806. Both retain substantial agreement on IT-textbook summaries, with Overall $\rho = 0.691$ and 0.720. A matched WikiLingua application yields similar reviewed quality after substituting the frozen reward evaluator. These findings support EViSE as an evaluation framework with explicit loss structure and practical reuse; broader language, domain, and reward settings remain to be evaluated.

Acknowledgements. The authors sincerely thank the 12 volunteer annotators with NLP expertise who participated in the full-corpus review of the EViSE annotations, the two additional annotators who conducted the blind reliability audit, the six IT faculty reviewers who contributed to the human review of the out-of-domain IT-textbook benchmark, and the two additional IT faculty experts who participated in the blind OOD validation. We are also grateful to colleagues and collaborators for their constructive feedback. Finally, we acknowledge the developers of ViDeBERTa and mmBERT for making the pretrained encoders publicly available for research use.

Statements and Declarations

Funding. No funding was received for conducting this study.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate. Formal ethics approval was not required for this study under the applicable institutional regulations because the human involvement was limited to voluntary expert annotation and evaluation of non-sensitive textual data, with no collection of personal, clinical, biometric, or other sensitive participant information. All annotators and expert reviewers participated voluntarily and provided informed consent for their assessments to be used for research purposes.

Consent for publication. Not applicable. No identifiable personal information of annotators or expert reviewers is reported in this manuscript.

Data availability. The primary EViSE-ViDeBERTa model checkpoint is publicly available under its original repository identifier, https://huggingface.co/phuongntc/Multi_EvalSumViet2, which is retained for release and citation continuity. The secondary EViSE-mmBERT checkpoint is available at https://huggingface.co/quancute/EViSE-mmBERT. The accompanying Vietnamese multi-criteria evaluation dataset is available at https://huggingface.co/datasets/phuongntc/EViSE-Dataset.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process. During the preparation of this work, the authors used OpenAI ChatGPT to assist with manuscript organization and language refinement. The authors reviewed and edited the resulting content as needed and take full responsibility for the content of the publication.

Appendix A Additional analyses and implementation details

This appendix records annotation, hyperparameter-selection, secondary-instantiation, and implementation details that support the main analysis.

A.1 Annotation and full-corpus review protocol

The labeling pipeline has three stages. First, Gemini 2.0 Flash evaluation prompts are optimized on 100 human-rated summaries using mean Cohen's κ across F/C/R. Second, the selected prompt assigns initial 1-5 ratings to the full corpus. Third, 12 trained annotators organized into six two-person groups jointly review every assigned instance and reach consensus on retained or revised criterion scores. Approximately 10% of initial labels are revised, and final F/C/R values are normalized to $[0, 1]$ for training.

Reliability-audit subset construction

A separate 450-pair audit subset is stratified by Overall quality (150 low, 150 mid, 150 high) with one pair per `doc_id` and fixed seed 20260612. Two additional annotators independently re-score F/C/R without access to corpus labels, quality bins, or split metadata. Agreement is computed before adjudication. The audit ratings are used only to estimate blind inter-annotator reproducibility and to compare the initial Gemini proposals with the final reviewed corpus labels; they do not change the corpus targets used for training, validation, model selection, or test reporting. The before–after comparisons are detailed in Appendix D. Table 15 summarizes the sampling design.

A.2 Practical selection of the pairwise ranking margin

Normalized Likert labels differ in increments of 0.25. Across 361,330 unequal within-source criterion pairs, the 20th percentile of $|\Delta y|$ is 0.25. We use a conservative scaled margin

$$m = \eta Q_{0.20}(|\Delta y|), \qquad \eta = 0.20, \tag{A1}$$

which gives $m = 0.05$. Table 16 records the label-gap diagnostics. The normalized five-point scale itself guarantees a minimum possible nonzero gap of 0.25, so $m = 0.05$ is compatible independently of the encoder. Proposition 2 establishes the admissible interval $0 < m \leq \delta_{\min}$; the fraction $\eta = 0.20$ determines the particular interior value used in development.

Table 15 Summary of the reliability-audit subset construction.

Item	Value
Full corpus size	80,856 document-summary pairs
Source documents	13,476
Audit subset size	450 pairs
Low-quality pairs	150
Mid-quality pairs	150
High-quality pairs	150
Unique `doc_id` values	450
Train/validation pairs	404
Held-out test pairs	46
Sampling rule	Stratified random sampling by quality bin
Random seed	20260612
Use of audit labels	Reliability estimation only

Table 16 Data-informed statistics used to select the pairwise ranking margin m.

Statistic	**Value**
Number of valid unequal-score pairs	361,330
$\min \lvert\Delta y\rvert$	0.25
$Q_{0.20}(\lvert\Delta y\rvert)$	0.25
$Q_{0.25}(\lvert\Delta y\rvert)$	0.25
Median $\lvert\Delta y\rvert$	0.25
Mean $\lvert\Delta y\rvert$	0.398
$\max \lvert\Delta y\rvert$	1.00
Shrinkage factor η	0.20
Selected margin m	0.05

A.3 Development setting of the fixed ranking weight

The ranking coefficient was fixed at $\lambda = 0.35$ in all reported hybrid runs. During EViSE-ViDeBERTa development, this value was selected from $\{0.20, 0.25, 0.30, 0.35, 0.40\}$ by mean validation Pearson correlation across F/C/R in a preliminary frozen-encoder comparison. It was retained for evaluator training and transferred unchanged to EViSE-mmBERT.

The theoretical analysis explains how a nonnegative fixed coefficient scales the ranking contribution within the centered prediction subspace. For compatible margins, the exact labels remain the minimizer in the unconstrained prediction-space problem for every $\lambda \geq 0$; this property does not determine a unique numerical coefficient. The reported value is therefore a validation-selected operating setting, whose practical effect is assessed by the fixed-architecture objective ablation.

A.4 Interpretation of the secondary Overall weighting

The scalarization $0.5F + 0.3R + 0.2C$ is used consistently for secondary aggregate reporting, quality bins, and reference selection. A development comparison

Table 17 Main implementation and training configuration for the EViSE-ViDeBERTa instantiation.

Component	Setting
Backbone encoder	`Fsoft-AIC/videberta-base`
Checkpoint parameter count	183.952M under the profiling implementation
Input / representation setup	Joint document-summary cross-encoder, summary cap 256 tokens, joint cap 512, `only_first` truncation, masked mean pooling
Prediction mapping	Linear projection to 256 + GELU + Dropout(0.1), followed by three criterion-specific scalar linear outputs (algebraically equivalent to one three-output linear mapping)
Optimizer	AdamW
Learning-rate setup	2×10^{-5}, weight decay 0.01, warmup ratio 0.05
Batch / precision	Configured batch size 8; one complete six-instance source group per micro-batch; gradient accumulation 2, giving 12 instances per full optimizer update; FP16 mixed precision
Training control	Maximum 5 epochs, gradient clipping 1.0, early stopping on mean validation Pearson across (F, C, R)
Ranking configuration	Margin 0.05; fixed ranking weight 0.35 selected during primary development; pair mean within each group, zero for an empty pair set, source mean as in Algorithm 1
Validation split protocol	Grouped split by `doc_id`
Inference profile	One forward pass produces F/C/R, no reference summary
Random seed	42

inspected the within-source best-worst range for several weightings. This statistic measures separation induced by a scalarization, not agreement with independent holistic judgments.

To see the distinction, let $w \in \mathbb{R}^3_+$ with $\sum_k w_k = 1$, and define

$$\Delta_g(w) = \max_i w^\top y_i - \min_j w^\top y_j, \qquad \Delta_{g,k} = \max_i y_i^k - \min_j y_j^k. \tag{A2}$$

For every candidate pair, $w^\top(y_i - y_j) \le \sum_k w_k \Delta_{g,k}$. Averaging the maximum pair difference over sources gives

$$\overline{\Delta}(w) \le \sum_k w_k \overline{\Delta}_k \le \max_k \overline{\Delta}_k. \tag{A3}$$

A simplex vertex for a criterion with the largest average range attains the upper bound. Unconstrained range maximization can therefore concentrate on one criterion rather than identify balanced quality preferences. The chosen Overall weights remain task-motivated, and criterion-wise evidence is the primary basis for evaluator assessment.

A.5 Reproducibility details

Tables 17 and 18 summarize the two training configurations. They share the same criterion targets, source-grouped split, margin, and hybrid objective, while operational settings differ to accommodate model size and context length.

Table 18 Implementation and training configuration for the secondary EViSE-mmBERT instantiation.

Component	Setting
Backbone encoder	`jhu-clsp/mmBERT-base` (Marone et al. 2026)
Released checkpoint	`https://huggingface.co/quancute/EViSE-mmBERT`
Checkpoint parameter count	307.334M
Input / representation setup	Joint document-summary encoding, joint cap 2,048 tokens, training-derived summary cap 512 tokens, remaining token budget allocated to the source, concatenated CLS + masked mean pooling
Prediction mapping	Linear projection from the concatenated representation to 256 + GELU + Dropout(0.1), followed by three criterion-specific scalar linear outputs
Optimizer / learning rate	AdamW, 2×10^{-5}, weight decay 0.01, warmup ratio 0.05
Grouped batch setup	All six candidate summaries of a source are retained in one micro-batch, gradient accumulation 8, giving an effective 48 examples per optimizer update
Training precision	Mixed precision selected by hardware support (BF16 when supported, otherwise FP16) with gradient checkpointing
Training control	Maximum 5 epochs, gradient clipping 1.0, checkpoint and early stopping on $0.5(\overline{r} + \overline{\rho})$ over F/C/R validation correlations
Ranking configuration	Margin 0.05 and fixed ranking weight 0.35 transferred unchanged; the same zero-for-empty source average as Algorithm 1
Validation split protocol	Same grouped split by `doc_id`, random seed 42
Matched internal profile	Tesla T4, FP32, batch size 4, 32 warm-up pairs, median of three 1,344-pair runs, loading excluded

A.6 Illustrative qualitative cases

Case A7164: Low overlap for a largely source-supported summary

Source context

"Djokovic thực hiện ca mổ đó sau khi cân nhắc. Cậu ấy may mắn có đội ngũ bác sĩ giỏi ở Paris", HLV Pilic nói trên Tennis Magazine ngày 13/8. Ông từng dạy Djokovic tại học viện quần vợt Munich những năm 1990. Theo HLV Pilic, chấn thương của Djokovic nhẹ, và ca phẫu thuật thực tế chỉ là một cuộc tiểu phẫu. "Nhưng cậu ấy chắc chắn không thể chơi tốt nếu không phẫu thuật", ông nói thêm. "Một bên chân Djokovic mềm hơn, nên di chuyển chậm hơn. Cậu ấy đã tập suốt nhiều tuần cho bên chân đó để lấy lại sức mạnh". Djokovic dính chấn thương gối ở vòng bốn Roland Garros đầu tháng Sáu, rồi lập tức rút lui và phẫu thuật. Sau đó, tay vợt 37 tuổi hồi phục với tốc độ đáng kinh ngạc, anh vẫn dự Wimbledon và vào tới chung kết. Sau đó, Djokovic lại chơi tốt ở Olympic Paris để đánh bại Carlos Alcaraz giành HC vàng. Điều này khiến nhiều người hoài nghi Djokovic có thực sự phẫu thuật hay không. "Djokovic may mắn có cơ địa tốt hơn người thường", ông Pilic nói thêm. "Cơ thể cậu ấy dẻo dai, sức mạnh thể chất và tinh thần đều tốt. Cậu ấy có sự tập trung rất cao, trên sân tập cũng như trong trận đấu. Đó là khác biệt của một nhà vô địch". Pilic cho biết ông là một trong số ít chuyên gia dự đoán Djokovic thắng Alcaraz ở chung kết Olympic Paris. Theo ông, khi được kích hoạt tinh thần, Djokovic rất khó bị đánh bại. HLV này nói: "Tôi luôn nghĩ nếu Djokovic tìm được nhịp chơi thì cậu ấy có cơ hội lớn. Cậu ấy biết đây là Olympic cuối và sẽ dốc hết sức. Nhiều người nói động lực

của Alcaraz không lớn bằng. Nhưng tôi nhìn thấy thái độ tuyệt vời của Djokovic, và biết cậu ấy đủ khả năng chiến thắng". Djokovic hoàn tất bộ sưu tập danh hiệu khi thắng Alcaraz hai loạt tie-break ở chung kết Olympic. Ở tuổi 37, tay vợt Serbia giữ kỷ lục 24 Grand Slam, 40 Masters 1000, bảy ATP Finals, số tuần giữ số một thế giới và vô số kỷ lục khác. Anh dự kiến bảo vệ danh hiệu Mỹ Mở rộng vào cuối tháng này, hướng tới trở thành tay vợt đầu tiên cán mốc 25 major đánh đơn. Djokovic và huyền thoại nữ Margaret Court hiện cùng có thành tích 24 Grand Slam.

Reference-summary

HLV Pilic, người từng dạy cho Djokovic, cho biết rằng chấn thương của Djokovic nhẹ và ca phẫu thuật chỉ là một cuộc tiểu phẫu. Djokovic đã hồi phục với tốc độ đáng kinh ngạc sau khi phẫu thuật và đã dự Wimbledon và Olympic Paris, nơi anh đã giành HC vàng. Pilic cho biết Djokovic có cơ địa tốt hơn người thường và có sự tập trung cao, đó là khác biệt của một nhà vô địch. Djokovic đã hoàn tất bộ sưu tập danh hiệu khi thắng Alcaraz hai loạt tie-break ở chung kết Olympic và đang hướng tới trở thành tay vợt đầu tiên cán mốc 25 major đánh đơn.

Candidate summary

HLV Boris Pilic xác nhận Djokovic đã trải qua một cuộc tiểu phẫu ở đầu gối sau Roland Garros, dù chấn thương không nghiêm trọng. Tuy nhiên, ca phẫu thuật là cần thiết để Djokovic lấy lại phong độ thi đấu đỉnh cao và giành huy chương vàng Olympic Paris, đánh bại Alcaraz trong trận chung kết nhờ sự tập trung cao độ và thể chất dẻo dai.

Scores

human-reviewed (F/C/R): 0.75 / 1.00 / 0.75
Prediction (F/C/R): 0.757 / 1.000 / 0.748
ROUGE-Lsum: 0.274

The candidate has low lexical overlap despite source-supported claims about the operation and Olympic victory. Its added first name is unsupported. The reviewed Faithfulness score is 0.75, and EViSE-ViDeBERTa closely matches all three criterion scores.

Case A12313: BERTScore-F1 overestimates a low-quality summary

Source context

Công ty cổ phần Hoàng Anh Gia Lai (HAGL, HAG) vừa công bố thương vụ bán khách sạn Hoàng Anh Gia Lai cho một đối tác, ghi nhận doanh thu từ thanh lý tài sản 180 tỷ đồng, trong tháng 9. Theo đó, bên mua là Công ty TNHH Đầu tư Hoàn Sinh Gia Lai. Công ty này thành lập ngày 20/6, vốn điều lệ 200 tỷ đồng. Ngành nghề kinh doanh chính là chăn nuôi trâu bò, lợn... Công ty còn hoạt động trong mảng câu lạc bộ thể thao (trừ hoạt động của các sàn nhảy). Theo thông tin đăng ký doanh nghiệp, Hoàn Sinh Gia Lai có trụ sở chính tại đường 3/2, quận 10, TP HCM. Doanh nghiệp do 2 cổ đông sáng lập là ông Đỗ Xuân Đức (địa chỉ Quảng Nam) và bà Nguyễn Thị Huyền (địa chỉ Bình Định). Trong đó, ông Đức sở hữu 49% vốn còn bà Huyền nắm

51% vốn. Ông Đức cũng giữ chức danh giám đốc công ty. Xác nhận với VnExpress, ông Đỗ Xuân Đức - Giám đốc Hoàn Sinh Gia Lai - cho biết nhận chuyển nhượng khách sạn Hoàng Anh Gia Lai từ tháng 9. Đây là giao dịch riêng giữa hai bên nên cả hai không thể công bố thông tin chi tiết. Ngoài Hoàn Sinh Gia Lai, ông Đức còn là thành viên Hội đồng quản trị của một công ty chuyên về xử lý nước thải, cung cấp dịch vụ vệ sinh môi trường và cung cấp nước sạch ở Quảng Nam. Khách sạn Hoàng Anh Gia Lai hoạt động từ tháng 12/2005, gồm 117 phòng ngủ, nằm ở vị trí đắc địa là số 1 Phù Đổng, TP Pleiku. Mục đích bán khách sạn này, theo bầu Đức để doanh nghiệp thanh toán một phần nợ trái phiếu HAGL 2016 tại Ngân hàng thương mại cổ phần Đầu tư và Phát triển Việt Nam (BIDV). Toàn bộ tiền bán bất động sản trên sẽ được ưu tiên thanh toán nghĩa vụ trái phiếu này tại BIDV. Tháng 9, công ty của bầu Đức đạt doanh thu thuần 679 tỷ đồng. Trong đó, ngành chăn nuôi chiếm 196 tỷ đồng; cây ăn trái là 375 tỷ đồng, ngành phụ trợ đạt 108 tỷ đồng. Lãi sau thuế 9 tháng của công ty này khoảng 710 tỷ đồng, giảm 20% so với cùng kỳ 2022. Thi Hà

Reference-summary

Công ty cổ phần Hoàng Anh Gia Lai vừa bán khách sạn Hoàng Anh Gia Lai cho Công ty TNHH Đầu tư Hoàn Sinh Gia Lai. Doanh thu từ thanh lý tài sản 180 tỷ đồng. Khách sạn Hoàng Anh Gia Lai hoạt động từ tháng 12/2005, gồm 117 phòng ngủ, nằm ở vị trí đắc địa là số 1 Phù Đổng, TP Pleiku. Mục đích bán khách sạn này, theo bầu Đức để doanh nghiệp thanh toán một phần nợ trái phiếu HAGL 2016 tại Ngân hàng thương mại cổ phần Đầu tư và Phát triển Việt Nam (BIDV).

Candidate summary

Công ty Cổ phần Hoàng Anh Gia Lai (HAG) vừa bán khách sạn Hoàng Anh Gia Lai cho Công ty TNHH Đầu tư Hoàn Sinh Gia Lai, thành lập năm 2006 với vốn 200 tỷ đồng. Công ty này hoạt động trong ngành chăn nuôi và thể thao, và có trụ sở tại TP HCM. Ông Đỗ Xuân Đức, Giám đốc, và bà Nguyễn Thị Huyền, là cổ đông sáng lập và là thành viên Hội đồng quản trị của một công ty xử lý nước thải. Doanh thu thuần tháng 9 đạt 679 tỷ đồng, đạt doanh thu thuần 679 tỷ đồng, giảm 20% so với cùng kỳ 202

Scores

human-reviewed (F/C/R): 0.25 / 0.25 / 0.50
Prediction (F/C/R): 0.224 / 0.275 / 0.469
BERTScore-F1: 0.875

The summary preserves many named entities and topical cues, which leads to a high BERTScore-F1. However, it also contains factual distortion, relation misbinding across entities, repetition, and an incomplete final clause. EViSE-ViDeBERTa remains much closer to the low human-reviewed scores, suggesting that criterion-wise evaluation is less easily misled by topical similarity alone.

Case A329: Diagnostic difficult error case

Source context

Chuyến buýt từ Tân Cương đến Astana, thủ đô Kazakhstan, không có gì đặc biệt để giúp Lin En, du khách 23 tuổi người Trung Quốc, tăng tính trải nghiệm. Thậm chí, hành trình còn khiến anh mệt mỏi vì phải ngồi thẳng trên ghế cứng suốt ba ngày mới đến nơi. Dù vậy, nếu cho chọn lại, anh vẫn ngồi xe buýt thay vì đi máy bay. Lin là một trong số hàng triệu người Trung Quốc ngày nay có biệt danh 'mông sắt' - những du khách chấp nhận mệt mỏi, ngồi nhiều tiếng trên ghế cứng của xe buýt, tàu hỏa để đi du lịch. Đa số du khách mông sắt còn rất trẻ hoặc mới tốt nghiệp đại học. Điểm chung của họ là nhiều thời gian, ít tiền. Phương châm của họ là: 'Những người có mông sắt tận hưởng thế giới trước'. Tình hình kinh tế ảm đạm, nhu cầu tiêu dùng thấp, tỷ lệ thất nghiệp cao ở thanh niên Trung Quốc hiện nay là nguyên nhân giới trẻ có xu hướng đi du lịch tiết kiệm tối đa nhiều hơn. 'Du lịch kiểu mông sắt cho phép tôi đến nhiều nơi mà không cần nhiều tiền', Peng Fei, 27 tuổi, tự nhận là một 'mông sắt' chia sẻ. Nhiều du khách 'mông sắt' cho biết dù phải mất nhiều thời gian di chuyển, họ vẫn tận hưởng chuyến đi như kết bạn trên đường, thưởng thức các món ngon địa phương trong thời gian dừng chân. Nhưng để có được các trải nghiệm du lịch kể trên, trước hết họ cần sống sót sau lịch trình di chuyển quá mệt mỏi, vất vả như vậy. Chuyến buýt của Lin En bắt đầu từ 6h, trừ vài lần xe buýt dừng lại để khách xuống nghỉ ngơi, Lin ngồi tổng cộng 46 tiếng trên xe. Anh hầu như không ngủ được trong hành trình dài vì đau lưng. 'Tôi thức trắng đêm, đến sáng hôm sau thì kiệt sức', anh nói. Mỗi phút ngồi xe với Lin là một cực hình. Anh thậm chí còn mong sớm đến nơi hơn cả tài xế. Chuyến quay về, Lin bị ốm. Đây không phải lần đầu Lin thực hiện các chuyến đi như vậy. Trong chuyến tàu hồi đầu năm kéo dài 53 tiếng từ Quảng Châu đến Lhasa, Lin đã nhìn thấy cao nguyên Tây Tạng rộng lớn, sa mạc Gobi nổi tiếng và những dãy núi hùng vĩ ở cự ly gần. Với chàng trai trẻ, mọi thứ đều vô cùng hấp dẫn vì đây là những khung cảnh khách đi máy bay không thể nhìn thấy. Từ cuối năm ngoái đến nay, Lin đã thực hiện 10 chuyến tương tự, ngồi khoảng 300 giờ trên xe buýt, tàu hỏa. Một trong những chuyến đi đáng nhớ nhất của anh là hành trình 20 tiếng Côn Minh (Trung Quốc) - Huay Xai (Lào). Lin gặp một nhóm công nhân đang trên đường đến khu Tam Giác Vàng nằm ở biên giới Thái Lan - Lào - Myanamar, một trung tâm buôn bán ma túy khét tiếng, để tìm việc. Họ nói sợ khi đến đó làm nhưng vẫn lựa chọn vì phải nuôi gia đình. Peng Fei, một mông sắt khác, đã chi 10.000 tệ (1.400 USD) để ghé 7 quốc gia, từ Kazakhstan đến Thổ Nhĩ Kỳ, chủ yếu bằng xe buýt. Chuyến bay duy nhất là từ Uzbekistan đến UAE với giá 300 tệ (42 USD). Fei có mức lương khá nhưng cảm thấy nhàm chán nên đã nghỉ việc để đi du lịch nhằm tìm kiếm ý nghĩa cuộc sống. Trong các chuyến đi, cô có cơ hội làm quen với nhiều người cùng chí hướng. Đó là những người trẻ mới tốt nghiệp chưa vội tìm việc ngay hoặc sinh viên đại học nghỉ một năm để đi khám phá thế giới. 'Tôi muốn khám phá thế giới càng nhiều càng tốt khi còn trẻ', Fei nói vì e ngại khi về già sẽ không thể làm du khách mông sắt được nữa. Anh Minh (TheoCNN)

Candidate summary

Tình hình kinh tế ảm đạm, nhu cầu tiêu dùng thấp, tỷ lệ thất nghiệp cao ở thanh niên Trung Quốc hiện nay là nguyên nhân khiến họ đi du lịch tiết kiệm tối đa nhiều

hơn. 'Du lịch kiểu mông sắt cho phép tôi đến nhiều nơi mà không cần nhiều tiền', Lin En chia sẻ.

Scores

human-reviewed (F/C/R): 0.50 / 0.75 / 0.50
Prediction (F/C/R): 0.757 / 0.971 / 0.740

The summary is fluent and remains on topic, but it attributes a quoted motivation to the wrong individual and reduces a richer experiential narrative into a simpler socioeconomic statement. Because the broader topic and discourse frame are preserved, EViSE-ViDeBERTa overestimates the output. This case highlights a remaining weakness in speaker attribution and entity-role binding.

Appendix B Theoretical analysis of the grouped cardinal-ordinal objective

This appendix gives the assumptions and proofs for Section 3.3. The loss-geometry and ordering bounds apply to arbitrary predictions. Statements about a unique empirical minimizer assume unconstrained prediction coordinates. The ranking coefficient is fixed throughout the analysis, as in the reported experiments.

B.1 Setting and scope

The N training instances are partitioned into disjoint source groups $\mathcal{I}_g$, each with $n_g \geq 2$; EViSE uses $n_g = 6$. Criterion labels y_i^k are numeric rubric scores on a common scale, $k \in \mathcal{K} = \{F, C, R\}$. The reported realizations use the normalized five-point labels $\{0, 0.25, 0.5, 0.75, 1\}$. The analysis uses raw linear-head predictions $\hat{y} \in \mathbb{R}^{3N}$. Let $e_i^k = \hat{y}_i^k - y_i^k$, $\bar{e}_g^k = n_g^{-1} \sum_{i \in \mathcal{I}_g} e_i^k$, $\varepsilon_i^k = e_i^k - \bar{e}_g^k$, and $V_g^k = n_g^{-1} \sum_i (\varepsilon_i^k)^2$. Eligible pairs and losses are as in Eqs. (5)-(8). Write $\mathcal{G}_k^+ = \{g \in \mathcal{G} : |\mathcal{P}_g^k| > 0\}$. We assume each criterion has at least one eligible source in the full training set, and define

$$\delta_{\min} = \min_{k \in \mathcal{K}} \min_{g \in \mathcal{G}_k^+} \min_{(i,j) \in \mathcal{P}_g^k} |y_i^k - y_j^k|. \tag{B1}$$

Assumption 1 (Unconstrained prediction coordinates)**.** For statements about an empirical optimum, the predictions on the fixed training instances are treated as freely adjustable optimization coordinates $\hat{y} \in \mathbb{R}^{3N}$. This unconstrained prediction-space problem omits the constraints induced by a shared neural parameterization.

This idealization studies the objective independently of encoder capacity. A shared finite neural model generally constrains feasible prediction vectors, and its parameter-space objective need not be convex. No proof below asserts neural convergence or population Fisher consistency. Unless explicitly stated, $m > 0$ and the coefficient is a fixed $\lambda \geq 0$.

B.2 Risk normalization and grouped mini-batches

Choose k uniformly among the three criteria and g uniformly from all G sources. If $\mathcal{P}_g^k$ is nonempty, choose a pair uniformly within it; otherwise return zero. A nonempty tuple has probability $[3G|\mathcal{P}_g^k|]^{-1}$, and the expected hinge contribution is exactly Eq. (7). This construction gives equal source coefficients before observing pair counts and fixed equal criterion coefficients.

For a fixed source and criterion, let n_ℓ count the candidates at rubric level $\ell \in \{1, \dots, 5\}$, with $\sum_{\ell=1}^{5} n_\ell = 6$. The number of eligible unequal-label pairs is

$$|\mathcal{P}_g^k| = \binom{6}{2} - \sum_{\ell=1}^{5} \binom{n_\ell}{2}.$$

The count configuration $(5, 1)$, omitting zero counts, means that five candidates share one label and the remaining candidate has a different label, yielding $15 - \binom{5}{2} = 5$ eligible pairs. The configuration $(2, 1, 1, 1, 1)$ means that one label occurs twice and each other label occurs once, yielding $15 - \binom{2}{2} = 14$ pairs. These configurations attain the minimum and maximum positive pair counts, respectively.

Pooling all pairs weights sources by their eligible-pair counts. For two eligible sources with 5 and 14 pairs and respective ordering errors 1 and 0, the pooled error is $5/19$, whereas their equally weighted source mean is $1/2$. Averaging pairs within each source prevents a source from dominating solely through its pair count.

For b complete source groups sampled uniformly, the batch estimator is

$$\widehat{\mathcal{L}}_{\text{rank}} = \frac{1}{3b} \sum_{a=1}^{b} \sum_{k \in \mathcal{K}} \ell_{g_a}^k. \tag{B2}$$

For sampling with replacement, linearity of expectation gives

$$\mathbb{E}\Big[\widehat{\mathcal{L}}_{\text{rank}}\Big] = \frac{1}{3b} \sum_{a=1}^{b} \sum_{k} \frac{1}{G} \sum_{g \in \mathcal{G}} \ell_g^k = \mathcal{L}_{\text{rank}}.$$

The same identity holds for a uniform subset sampled without replacement because every source has inclusion probability b/G. For the reported $b = 1$ micro-batches, Eq. (B2) is exactly the zero-for-empty computation in Algorithm 1. If the six-instance source groups are visited once in a shuffled epoch, the average of their loss values, evaluated at a fixed prediction vector, also equals Eq. (7) exactly.

Each group contains six instances, so uniformly sampling sources also produces an unbiased estimate of the instance-averaged regression term in Eq. (4). For K accumulated one-source micro-batches, the effective loss is

$$\widehat{\mathcal{L}}_{\text{hybrid,acc}} = \frac{1}{K} \sum_{a=1}^{K} \left(\widehat{\mathcal{L}}_{\text{reg},a} + \lambda \widehat{\mathcal{L}}_{\text{rank},a} \right).$$

The implementation settings use $K = 2$ for ViDeBERTa and $K = 8$ for mmBERT. An empty ranking contribution remains zero during this averaging. These identities describe loss normalization at fixed predictions; they do not establish convergence of stochastic optimization of a shared network.

To make the eligibility distinction explicit, let $\mathcal{G}_k^+ = \{g : |\mathcal{P}_g^k| > 0\}$ and $p_k = |\mathcal{G}_k^+|/G > 0$. Define the conditional source mean

$$\overline{\mathcal{L}}_{k,+} = \frac{1}{|\mathcal{G}_k^+|} \sum_{g \in \mathcal{G}_k^+} \ell_g^k.$$

Then

$$\mathcal{L}_{\text{rank}} = \frac{1}{3} \sum_{k \in \mathcal{K}} p_k \overline{\mathcal{L}}_{k,+}. \tag{B3}$$

Thus a conditional objective $\frac{1}{3}\sum_k \overline{\mathcal{L}}_{k,+}$ would require inverse-eligibility factors $1/p_k$ when sources are sampled uniformly from all groups. It is a different objective. Equations (7) and (B2) consistently use the unconditional source average, so no assumption $p_k = 1$ and no inverse-eligibility correction are needed. The geometric proofs below use its positive pair weights and the zero contribution convention.

B.3 Proof of the group-offset and centered-error theorem

Let

$$\mathcal{W} = \{v \in \mathbb{R}^{3N} : \sum_{i \in \mathcal{I}_g} v_i^k = 0 \text{ for all } g, k\}, \qquad (\Pi v)_i^k = v_i^k - \bar{v}_g^k. \tag{B4}$$

The map Π is the Euclidean orthogonal projection onto $\mathcal{W}$.

Proof of Theorem 1.
Expanding $e_i^k = \bar{e}_g^k + \varepsilon_i^k$ and using $\sum_i \varepsilon_i^k = 0$ gives

$$\sum_{i \in \mathcal{I}_g} (e_i^k)^2 = n_g (\bar{e}_g^k)^2 + \sum_{i \in \mathcal{I}_g} (\varepsilon_i^k)^2,$$

which proves the decomposition. For every eligible pair,

$$s_{ij}^k (\hat{y}_i^k - \hat{y}_j^k) = |y_i^k - y_j^k| + s_{ij}^k (\varepsilon_i^k - \varepsilon_j^k). \tag{B5}$$

The common offset cancels, so adding a constant within one source-criterion group leaves every hinge unchanged.

Let $\mathbf{b}_i^k$ be the standard basis vector for prediction coordinate (i, k). For a pair, write $a = s_{ij}^k (\mathbf{b}_i^k - \mathbf{b}_j^k)$ and $h(\hat{y}) = \max(0, m - a^\top \hat{y})$. Its subgradients are $-ca$, with $c = 1$ when $a^\top \hat{y} < m$, $c \in [0, 1]$ when equality holds, and $c = 0$ above the margin. Each vector is supported on the source and sums to zero there. The sum and linear-composition rules for finite convex functions (Rockafellar 1970) preserve this property. Hence every $r \in \partial \mathcal{L}_{\text{rank}}$ belongs to $\mathcal{W}$.

Under Assumption 1, regression has Hessian $2I/(3N) \succ 0$, while ranking is convex and nonnegative. Their sum is strictly convex and coercive, and therefore has a unique minimizer. The optimality condition is

$$0 = \frac{2}{3N}(\hat{y}^\star - y) + \lambda r, \qquad r \in \partial \mathcal{L}_{\text{rank}}(\hat{y}^\star). \tag{B6}$$

Summing over each source-criterion group gives $\sum_i (\hat{y}_i^{k\star} - y_i^k) = 0$, proving mean alignment. Ranking alone cannot identify offsets: all its minimizers may be independently shifted in each group and criterion. Such minimizers exist because sufficiently expanding the centered labels makes every eligible gap at least m. □

B.4 Margin compatibility and prediction-gradient alignment

Proof of Proposition 2.

If $m \le \delta_{\min}$, every hinge vanishes at $\hat{y} = y$, as does regression. Nonnegativity and strict convexity prove uniqueness for all $\lambda \ge 0$. Conversely, suppose $m > \delta_{\min}$ and choose a minimum-gap pair in source g and criterion k. Let $d_i^k = y_i^k - \bar{y}_g^k$ on that group and zero elsewhere. At $y + td$, each signed gap in that group is $(1+t)|y_i^k - y_j^k|$. Every hinge has a nonpositive right derivative at $t = 0$, and the selected pair has derivative $-\delta_{\min} < 0$. Regression has zero derivative at y. For every $\lambda > 0$, the total directional derivative is therefore negative, so y is not a minimizer. □

Proof of Corollary 3.

For each eligible pair, let $u_{ij}^k = s_{ij}^k(\hat{y}_i^k - \hat{y}_j^k)$ and let c_{ij}^k be its hinge subgradient coefficient, as defined in the proof of Theorem 1. Write $\omega_{gij}^k = [3G|\mathcal{P}_g^k|]^{-1}$. Since $q = 2(\hat{y} - y)/(3N)$,

$$\langle q, r \rangle = \frac{2}{3N} \sum_{k,g,(i,j)} \omega_{gij}^k c_{ij}^k \left(|y_i^k - y_j^k| - u_{ij}^k \right), \tag{B7}$$

where the sum ranges over eligible tuples. A nonzero coefficient requires $u_{ij}^k \le m \le |y_i^k - y_j^k|$, so each summand is nonnegative. At a differentiable point with $r \neq 0$, at least one hinge is strictly active, with $c_{ij}^k = 1$ and $u_{ij}^k < m$, making the sum strictly positive. □

For compatible margins, all nonnegative weights have the same ideal empirical target. This proves that target preservation cannot identify a unique coefficient such as 0.35. With an incompatible margin, the improving direction is centered, but the resulting optimum need not be a uniform rescaling of labels.

B.5 Proofs of the ordering bounds

Proof of Proposition 4.

For every real u, $(m - u)_+ \ge m\mathbf{1}[u \le 0]$. Apply this to each signed predicted gap and average over eligible pairs to obtain $D_g^k \le \ell_g^k / m$. Since $D_g^k \le 1$, the first bound

follows. Extend D_g^k by zero on empty pair sets and average with the same $1/(3G)$ coefficients as Eq. (7). This proves the unconditional macro bound. For $p_k > 0$, a conditional ordering-error mean instead satisfies

$$\overline{D}_{k,+} = \frac{1}{|\mathcal{G}_k^+|} \sum_{g \in \mathcal{G}_k^+} D_g^k \le \min\left\{1, \frac{\overline{\mathcal{L}}_{k,+}}{m}\right\} = \min\left\{1, \frac{G^{-1}\sum_g \ell_g^k}{p_k m}\right\}.$$

The factor p_k is therefore retained when connecting an unconditional training risk to evaluation restricted to eligible sources.

An inversion or predicted tie implies $s_{ij}^k(e_i^k - e_j^k) \le -|y_i^k - y_j^k|$, so

$$\mathbf{1}[s_{ij}^k(\hat{y}_i^k - \hat{y}_j^k) \le 0] \le \frac{(e_i^k - e_j^k)^2}{(y_i^k - y_j^k)^2}. \tag{B8}$$

Use $|y_i^k - y_j^k| \ge \delta_{\min}$ and enlarge the sum to all unordered pairs in the source. The identity

$$\sum_{i<j}(e_i^k - e_j^k)^2 = n_g \sum_i (e_i^k - \bar{e}_g^k)^2 = n_g^2 V_g^k$$

follows by expanding $\frac{1}{2}\sum_{i,j}(e_i^k - e_j^k)^2$. Substitution gives the regression bound. It is attained for $n_g = 2$, label gap $\delta_{\min}$, and errors $-\delta_{\min}/2, +\delta_{\min}/2$, which produce a predicted tie. □

With $n_g = 6$ and gap bound 0.25, the coefficient of V_g^k is $36/(|\mathcal{P}_g^k|\, 0.25^2)$, ranging from about 41.14 to 115.2 over nonempty groups. The smallest value applies only at 14 eligible pairs. A possibly loose upper bound supplies sufficient conditions, not necessary RMS-error thresholds for good ordering. Aggregate MAE alone does not determine V_g^k because it does not separate offsets from centered errors.

By Eq. (B5), a hinge is positive exactly when

$$s_{ij}^k(\varepsilon_i^k - \varepsilon_j^k) < m - |y_i^k - y_j^k|. \tag{B9}$$

This is a one-sided condition on adverse centered error. Regression measures errors symmetrically and also controls ordering through the proved bound; zero regression error already gives perfect ordering. The analysis therefore motivates, but does not guarantee, an empirical hybrid advantage.

For strict pairwise accuracy, $\mathrm{PA}_g^k = 1 - D_g^k$. Giving half credit to predicted ties still yields $1 - \mathrm{PA}_g^k \le D_g^k$. If C_g, A_g count concordant and discordant eligible pairs, respectively, the restricted coefficient $\tau_g^\dagger = (C_g - A_g)/|\mathcal{P}_g^k|$ satisfies $\tau_g^\dagger \ge 1 - 2D_g^k$. This coefficient is not Kendall's τ_b. When the ranking loss is zero, all highest-label candidates outrank all lower-label candidates, giving correct tie-aware Top-1 selection for that criterion. None of these statements directly controls the order of the weighted Overall score.

B.6 Prediction-gradient geometry of the fixed ranking weight

Set $q = \nabla_{\hat{y}}\mathcal{L}_{\text{reg}}$, $q_W = \Pi q$, $q_0 = (I - \Pi)q$, and $r \in \partial\mathcal{L}_{\text{rank}}$. By Theorem 1, $r \in \mathcal{W}$, whereas q_0 is constant within each source-criterion group. Therefore $\langle q_0, r\rangle = 0$, proving Eq. (12). For a fixed coefficient $\lambda \geq 0$, a hybrid-loss subgradient decomposes as

$$q + \lambda r = q_0 + (q_W + \lambda r). \tag{B10}$$

The first component supplies the offset correction; the second combines centered regression error with within-source ordering. Thus changing the fixed coefficient scales the ranking contribution in the centered subspace while preserving the direct offset component in prediction space.

At a point where no eligible signed gap equals m, the ranking loss is differentiable. Let $d = q + \lambda r$ with any fixed $\lambda \geq 0$. If $0 < m \leq \delta_{\min}$, Corollary 3 yields

$$\frac{d}{dt}\mathcal{L}_{\text{reg}}(\hat{y} - td)\bigg|_{t=0} = -\|q\|_2^2 - \lambda\langle q, r\rangle \leq 0, \tag{B11}$$

$$\frac{d}{dt}\mathcal{L}_{\text{rank}}(\hat{y} - td)\bigg|_{t=0} = -\langle q, r\rangle - \lambda\|r\|_2^2 \leq 0. \tag{B12}$$

Both derivatives are strictly negative when $r \neq 0$, since the corollary then gives $\langle q, r\rangle > 0$. These are local directional derivatives in Euclidean prediction space. At a hinge boundary, the derivative must be evaluated using all active terms; an arbitrary selected subgradient alone is insufficient. Shared-parameter updates are addressed in Appendix B.7.

This geometry applies to the fixed coefficient used in the experiments. It characterizes the interaction of the two terms without identifying a unique best coefficient for the finite shared evaluator; the development choice is documented in Appendix A.3.

B.7 Limits of transferring the analysis to a shared evaluator

For $\hat{y} = f_\theta$ with Jacobian J_θ, backpropagation gives $\nabla_\theta\mathcal{L}_{\text{reg}} = J_\theta^\top q$ and $\nabla_\theta\mathcal{L}_{\text{rank}} = J_\theta^\top r$. Although $q_0^\top r = 0$, the parameter-space inner product

$$\langle J_\theta^\top q_0, J_\theta^\top r\rangle = q_0^\top J_\theta J_\theta^\top r \tag{B13}$$

need not vanish. Shared parameters can therefore move group offsets during ranking updates. The prediction-space properties do not establish convergence of stochastic AdamW or guarantee joint loss reduction after shared-parameter updates and gradient clipping.

Likewise, exact empirical target compatibility is not population consistency under random labels. For one criterion and two fixed candidates, let labels be $(0.5, 0.25)$ with probability 0.6 and $(0, 1)$ with probability 0.4. Every realized gap is at least 0.25, but the MSE population target is $(0.3, 0.55)$. At $m = 0.05$, the positive-sign event has an active hinge and the negative-sign event does not, producing an expected ranking gradient proportional to $(-0.6, 0.6)$ at that target. Adding positive ranking weight

can therefore shift the conditional-mean target. A population-consistency claim would need additional assumptions.

These limits do not affect the algebraic decomposition or deterministic ordering bounds. They delimit which properties can be attributed to the objective and which require empirical evidence from the trained evaluator.

Appendix C Relation to earlier Vietnamese summarization studies

Table 19 compares EViSE with VSum-HB (Tran Thi et al. 2025) and the verifier-guided RL study of Tran et al. (2026). The retained MultiEvalSumViet2 checkpoint is analyzed here as EViSE-ViDeBERTa, alongside the mmBERT instantiation and matched evaluator-substitution experiment.

Table 19 Scope of EViSE relative to earlier Vietnamese summarization studies.

Aspect	VSum-HB (2025)	Tran et al. (2026)	Present EViSE study
Research focus	Vietnamese summarization data for learning from human feedback.	Verifier-guided RL and penalty shaping for small summarization policies.	Learning and analyzing a shared evaluator for Faithfulness, Coherence, and Relevance.
Resource or model role	5,000 Vietnews samples with LLM-assisted summary construction and human refinement.	Frozen MultiEvalSumViet2 scorer supplying criterion scores during PPO/GRPO.	80,856 document-summary evaluations; retained compact checkpoint and a higher-capacity mmBERT instantiation.
Method examined	Human-feedback-oriented summary construction and generation.	Reward construction, VR/VRP shaping, and policy optimization.	Criterion-wise, group-balanced cardinal-ordinal objective; loss geometry, compatible margins, and ordering bounds.
Empirical emphasis	Summary quality and use of the resource for RLHF.	Policy backbones, optimizers, initialization regimes, reward ablations, and WikiLingua transfer.	Criterion agreement, within-source ordering, objective ablation, computational cost, IT-textbook transfer, and evaluator substitution.
Relation to present results	Earlier resource study informing the Vietnamese summarization context.	Previously reported compact-evaluator RL reference configuration.	Evaluator analyses and mmBERT substitution; the compact WikiLingua Overall value of 4.45 remains a previously reported result.

Table 14 retains the previously reported compact-evaluator reference result; its mmBERT row comes from this study's matched substitution run. The methodological contribution concerns the criterion-wise grouped evaluator objective.

Appendix D Agreement before and after full-corpus review with blind human ratings

This analysis compares two states of the labels on the same 450-pair blind audit subset described in Section 4.1.1 and Appendix A.1: the initial Gemini 2.0 Flash

ratings before full-corpus review and the final consensus labels after review by the 12 annotators. A and B are the two additional blind annotators, and their ratings are taken before audit adjudication. Neither their ratings nor the analyses below alter the corpus targets used for evaluator learning or evaluation.

For criterion $k \in \{F, C, R\}$, let G_i^k denote the initial Gemini rating, L_i^k the final reviewed corpus rating, and A_i^k, B_i^k the two blind human ratings for pair i. All ratings use the original 1–5 rubric. We retain the fractional blind-human mean

$$H_i^k = \frac{A_i^k + B_i^k}{2}.$$

Quadratic-weighted Cohen's κ compares G^k and L^k separately with A^k and B^k. For either label state $X \in \{G, L\}$, comparison with the blind-human mean uses Spearman correlation, MAE, and mean signed difference:

$$\mathrm{MAE}_k(X) = \frac{1}{450}\sum_{i=1}^{450} |X_i^k - H_i^k|, \qquad \overline{\Delta}_k(X) = \frac{1}{450}\sum_{i=1}^{450}(X_i^k - H_i^k).$$

A positive $\overline{\Delta}_k(X)$ indicates that label source X assigns higher scores on average than the blind-human mean. MAE and signed differences are reported in original 1–5 score points. Percentile 95% confidence intervals use 5,000 bootstrap resamples of documents with replacement. For before–after comparisons, the initial Gemini rating, final reviewed rating, and both blind ratings for a document are retained together in each resample. There is one audit pair per document. Tables 20 and 21 report the before–after agreement with individual blind annotators and their mean, respectively.

Table 20 Quadratic-weighted Cohen's κ with the two independent blind annotators before and after full-corpus human review on 450 document–summary pairs. Brackets give document-bootstrap 95% confidence intervals.

Criterion	Initial Gemini–A	Reviewed corpus–A	Initial Gemini–B	Reviewed corpus–B
Faithfulness	0.796 [0.754, 0.833]	0.809 [0.769, 0.845]	0.794 [0.752, 0.829]	0.810 [0.770, 0.844]
Coherence	0.774 [0.727, 0.814]	0.783 [0.739, 0.822]	0.784 [0.744, 0.820]	0.795 [0.758, 0.829]
Relevance	0.819 [0.781, 0.851]	0.821 [0.784, 0.853]	0.811 [0.771, 0.846]	0.816 [0.776, 0.851]

Table 21 Initial Gemini proposals and final reviewed corpus labels compared with the blind-human mean $H = (A + B)/2$ on the same 450 audit pairs. Brackets give document-bootstrap 95% confidence intervals. MAE and mean signed difference use the original 1–5 scale.

Criterion	Label state	Spearman ρ	MAE	Mean score $-H$
Faithfulness	Initial Gemini	0.817 $[0.778, 0.848]$	0.476 $[0.424, 0.529]$	-0.069 $[-0.137, -0.001]$
	Reviewed corpus	0.837 $[0.802, 0.866]$	0.460 $[0.408, 0.513]$	$+0.013$ $[-0.054, 0.080]$
Coherence	Initial Gemini	0.846 $[0.816, 0.873]$	0.422 $[0.373, 0.473]$	$+0.073$ $[0.010, 0.137]$
	Reviewed corpus	0.860 $[0.831, 0.884]$	0.409 $[0.361, 0.459]$	$+0.084$ $[0.023, 0.147]$
Relevance	Initial Gemini	0.850 $[0.816, 0.878]$	0.443 $[0.394, 0.492]$	-0.001 $[-0.063, 0.062]$
	Reviewed corpus	0.857 $[0.825, 0.884]$	0.434 $[0.386, 0.484]$	$+0.003$ $[-0.057, 0.067]$

All six quadratic-weighted κ point estimates increase after full-corpus review, as do the Spearman correlations for all three criteria, while MAE decreases for F, C, and R. In paired document-bootstrap comparisons, the post-review increases in Spearman correlation are $+0.020\,[0.006, 0.035]$ for Faithfulness, $+0.014\,[0.003, 0.028]$ for Coherence, and $+0.008\,[0.002, 0.014]$ for Relevance. These paired intervals support modest improvements in rank agreement. Marginal intervals in the two tables do not establish the significance of a before–after difference in κ or MAE; such a conclusion requires the interval of the paired change itself.

The signed-difference analysis also clarifies the effect of review. For Faithfulness, the initial Gemini proposals have a small negative mean difference of -0.069 points whose 95% confidence interval excludes zero; after review, the mean difference is $+0.013$ and its interval includes zero. Coherence retains a small positive mean difference after review, whereas Relevance remains close to zero. The post-review proportions satisfying $|L_i^k - H_i^k| \leq 1$ are 92.67%, 93.11%, and 94.44% for F, C, and R, respectively.

These before-after results provide evidence that full-corpus review preserved and, on these aggregate measures, modestly improved alignment with independent blind rubric judgments. The consistency of the post-review results across quadratic-weighted κ, Spearman correlation, MAE, and signed-difference analyses supports the reliability of the final reviewed labels used for evaluator learning and evaluation. The audit subset is balanced across three quality bins rather than sampled in proportion to their full-corpus prevalence. Accordingly, the reported estimates characterize agreement on this quality-stratified audit subset and should not be interpreted as direct population-wide estimates for the entire corpus.

References

Bai Y, Kadavath S, Kundu S, et al (2022) Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073, https://doi.org/10.48550/arXiv.2212.08073, URL https://arxiv.org/abs/2212.08073

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39(3–4):324–345. https://doi.org/10.1093/biomet/39.3-4.324

Burges CJC, Shaked T, Renshaw E, et al (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning, pp 89–96, https://doi.org/10.1145/1102351.1102363

Caruana R (1997) Multitask learning. Machine Learning 28(1):41–75. https://doi.org/10.1023/A:1007379606734

Conneau A, Rinott R, Lample G, et al (2018) XNLI: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 2475–2485, https://doi.org/10.18653/v1/D18-1269, URL https://aclanthology.org/D18-1269/

Durmus E, He H, Diab M (2020) FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 5055–5070, https://doi.org/10.18653/v1/2020.acl-main.454, URL https://aclanthology.org/2020.acl-main.454/

Fabbri AR, Kryściński W, McCann B, et al (2021) SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics (TACL) 9:391–409. https://doi.org/10.1162/tacl_a_00373, URL https://aclanthology.org/2021.tacl-1.24/

Fabbri AR, Wu CS, Liu W, et al (2022) QAfacteval: Improved QA-based factual consistency evaluation for summarization. In: Carpuat M, de Marneffe MC, Meza Ruiz IV (eds) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 2587–2601, https://doi.org/10.18653/v1/2022.naacl-main.187, URL https://aclanthology.org/2022.naacl-main.187/

He P, Gao J, Chen W (2023) Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In: International Conference on Learning Representations (ICLR)

Kocmi T, Federmann C (2023) Large language models are state-of-the-art evaluators of translation quality. In: Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation, Tampere, Finland, pp 193–203, URL https://aclanthology.org/2023.eamt-1.19/

Laban P, Schnabel T, Bennett PN, et al (2022) SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics (TACL) 10:163–177. https://doi.org/10.1162/tacl_a_00453, URL https://aclanthology.org/2022.tacl-1.10/

Ladhak F, Durmus E, Cardie C, et al (2020) WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2020, https://doi.org/10.18653/v1/2020.findings-emnlp.360, URL https://aclanthology.org/2020.findings-emnlp.360/

Lam K, Doan T, Pham K, et al (2023) Abstractive text summarization using the brio training paradigm. In: Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 92–99, https://doi.org/10.18653/v1/2023.findings-acl.7, URL https://aclanthology.org/2023.findings-acl.7/

Laurer M (2023) mdeberta-v3-base-mnli-xnli. https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli, hugging Face model card

Le-Duc K, Nguyen KN, Vo-Dang L, et al (2024) Real-time speech summarization for medical conversations. In: Proceedings of INTERSPEECH 2024, pp 1960–1964, https://doi.org/10.21437/Interspeech.2024-2250

Li Z, Ou L, Huang Y (2025) Balancing factual consistency and informativeness for abstractive summarization. International Journal of Machine Learning and Cybernetics 16:10365–10381. https://doi.org/10.1007/s13042-025-02724-8, URL https://doi.org/10.1007/s13042-025-02724-8

Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, pp 74–81, URL https://aclanthology.org/W04-1013/

Liu TY (2009) Learning to rank for information retrieval. Foundations and Trends in Information Retrieval 3(3):225–331. https://doi.org/10.1561/1500000016

Liu Y, Iter D, Xu Y, et al (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 (EMNLP), https://doi.org/10.18653/v1/2023.emnlp-main.153, URL https://aclanthology.org/2023.emnlp-main.153/, arXiv:2303.16634

Marone M, Weller O, Fleshman W, et al (2026) mmBERT: A modern multilingual encoder with annealed language learning. In: Zhang T, Dudik M, Jaggi M, et al (eds) Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 306. PMLR, Seoul, South Korea, pp 86768–86783, URL https://proceedings.mlr.press/v306/marone26a.html

Maynez J, Narayan S, Bohnet B, et al (2020) On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), https://doi.org/10.18653/v1/2020.acl-main.173, URL https://aclanthology.org/2020.acl-main.173/

Nenkova A, Passonneau R (2004) Evaluating content selection in summarization: The pyramid method. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. Association for Computational Linguistics, Boston, Massachusetts, USA, pp 145–152, URL https://aclanthology.org/N04-1019/

Nguyen VH, Nguyen TC, Nguyen MT, et al (2019) VNDS: A vietnamese dataset for summarization. In: 2019 6th NAFOSTED Conference on Information and Computer Science (NICS), pp 375–380, https://doi.org/10.1109/NICS48868.2019.9023886

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems (NeurIPS), https://doi.org/10.48550/arXiv.2203.02155, URL https://arxiv.org/abs/2203.02155, arXiv:2203.02155

Pearson K (1895) Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58(347–352):240–242. https://doi.org/10.1098/rspl.1895.0041, URL https://royalsocietypublishing.org/rspl/article/58/347-352/240/43470/VII-Note-on-regression-and-inheritance-in-the-case

Phan L, Tran H, Nguyen H, et al (2022) ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, pp 136–142, https://doi.org/10.18653/v1/2022.naacl-srw.18, URL https://aclanthology.org/2022.naacl-srw.18/, arXiv:2205.06457

Rockafellar RT (1970) Convex Analysis. Princeton University Press, Princeton, NJ, https://doi.org/10.1515/9781400873173

Scialom T, Dray PA, Lamprier S, et al (2021) QuestEval: Summarization asks for fact-based evaluation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 6594–6604, https://doi.org/10.18653/v1/2021.emnlp-main.529, URL https://aclanthology.org/2021.emnlp-main.529/

Sculley D (2010) Combined regression and ranking. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 979–988, https://doi.org/10.1145/1835804.1835928

Spearman C (1904) The proof and measurement of association between two things. The American Journal of Psychology 15(1):72–101. URL https://archive.org/download/proofmeasurement00speauoft/proofmeasurement00speauoft_bw.pdf

Stiennon N, Ouyang L, Wu J, et al (2020) Learning to summarize with human feedback. In: Advances in Neural Information Processing Systems (NeurIPS), https://doi.org/10.48550/arXiv.2009.01325, URL https://arxiv.org/abs/2009.01325, arXiv:2009.01325

Tang H, Li R, Duan W, et al (2024) A novel abstractive summarization model based on topic-aware and contrastive learning. International Journal of Machine Learning and Cybernetics 15:5563–5577. https://doi.org/10.1007/s13042-024-02263-8, URL https://doi.org/10.1007/s13042-024-02263-8

Tran CD, Pham NH, Nguyen AT, et al (2023) ViDeBERTa: A powerful pre-trained language model for Vietnamese. In: Findings of the Association for Computational Linguistics: EACL 2023. Association for Computational Linguistics, Dubrovnik, Croatia, pp 1071–1078, https://doi.org/10.18653/v1/2023.findings-eacl.79, URL https://aclanthology.org/2023.findings-eacl.79/, arXiv:2301.10439

Tran NT, Nghiem MQ, Nguyen NTH, et al (2020) ViMs: A high-quality vietnamese dataset for abstractive multi-document summarization. Language Resources and Evaluation 54(4):893–920. https://doi.org/10.1007/s10579-020-09495-4, URL https://link.springer.com/article/10.1007/s10579-020-09495-4

Tran TTP, Vu TH, Nguyen VV, et al (2026) Reinforcement learning with verifier guidance and penalty shaping for vietnamese summarization using small language models. IEEE Access 14:89301–89322. https://doi.org/10.1109/ACCESS.2026.3701952

Tran Thi TP, Nguyen Van V, Nguyen Phuong T, et al (2025) Vsum-hb: A vietnamese text summarization dataset for reinforcement learning from human feedback. In: Information and Communication Technology, Communications in Computer and Information Science, vol 2352. Springer, Singapore, p 148–161, https://doi.org/10.1007/978-981-96-4288-5_12, URL https://link.springer.com/chapter/10.1007/978-981-96-4288-5_12

Yang C, Wang X, Lu Y, et al (2024) Large language models as optimizers. In: The Twelfth International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=Bb4VGOWELI

Zhang T, Kishore V, Wu F, et al (2020) BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1904.09675, URL https://openreview.net/forum?id=SkeHuCVFDr, arXiv:1904.09675

Zhang T, Ladhak F, Durmus E, et al (2024) Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics 12:39–57. https://doi.org/10.1162/tacl_a_00632, URL https://aclanthology.org/2024.tacl-1.3/

Zhao S, Sun Y, Chen J, et al (2025) Sentop: sentence-level prefix prompt for controllable abstractive summarization. International Journal of Machine Learning and Cybernetics 16:7543–7557. https://doi.org/10.1007/s13042-025-02669-y, URL https://doi.org/10.1007/s13042-025-02669-y

Zhong M, Liu Y, Yin D, et al (2022) Towards a unified multi-dimensional evaluator for text generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 2023–2038, https://doi.org/10.18653/v1/2022.emnlp-main.131, URL https://aclanthology.org/2022.emnlp-main.131/