PRISM: MODULAR ALLOCATION AND COMPLEMENTARY FUSION FOR ARGUMENT QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous argument quality assessment (AQA) must rank arguments across topics and annotation scales, while aggregate scores obscure the distinct roles of model capacity, complementary signals, and score calibration. We introduce PRISM, a modular framework that allocates these roles to separate interfaces. An independently trained eight-replica Qwen3-8B ensemble provides one score, while a taxonomy-profile encoder provides a complementary signal; corpus-specific experts retain separate scales for closed-set IBM/Webis routing. On IBM-Rank-30k, the Qwen ensemble reaches Spearman 0.6108, profile calibration raises encoder Spearman from 0.5814 to 0.5928, and the frozen encoder complement raises the final fusion to 0.6146 ($+0.0038$ over Qwen alone). After fusion, Pearson and Spearman show relative improvements of 2.5% and 2.4%, respectively, over the strongest published reference in our comparison. On Webis-ArgQuality-20, target-domain supervision substantially outperforms IBM zero-shot transfer; in a fixed IBM/Webis mixture, confidence-based abstention over development-fitted scale maps significantly improves Webis ranking over hard routing while evaluating both experts only on low-confidence samples. PRISM therefore makes the scorer, encoder calibration, and protocol-aware scale adaptation separately measurable.

1 INTRODUCTION

Continuous argument quality assessment (AQA) predicts how strongly an argument is judged relative to others on unseen topics. Because argument quality is multidimensional, aggregate labels blend the effects of representation, model capacity, ensembling, and score calibration Wachsmuth et al. (2017a); Ivanova et al. (2024). A modular evaluation can expose these effects as measurable interfaces.

Argumentation theory organizes quality into logical, rhetorical, and dialectical criteria Wachsmuth et al. (2017b). Existing resources range from pairwise convincingness Habernal & Gurevych (2016b;a) to continuous IBM-Rank-30k scores Gretz et al. (2020). Recent LLM scorers improve prediction, making controlled attribution of theory-specific signals essential. Cross-corpus evaluation also requires joint treatment of semantic transfer and score-scale compatibility because IBM and Webis use different annotation protocols Fromm et al. (2023); Feger et al. (2025); Romberg et al. (2025).

Figure 1 summarizes the resulting design choice. A monolithic predictor provides an end-to-end score, whereas PRISM exposes separate interfaces for text-based scoring, profile calibration, and fusion. Its Qwen ensemble estimates quality directly from observable text, while the profile branch supplies an independently testable encoder signal.

The branches meet at late fusion, allowing the same frozen Qwen predictions to be paired with alternative encoder interfaces. Matched controls isolate affine correction, representation geometry, taxonomy specificity, and encoder $\times$ Qwen contributions. Figure 2 details the teacher, student profile construction, fusion path, and protocol-aware scale-adaptation study.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

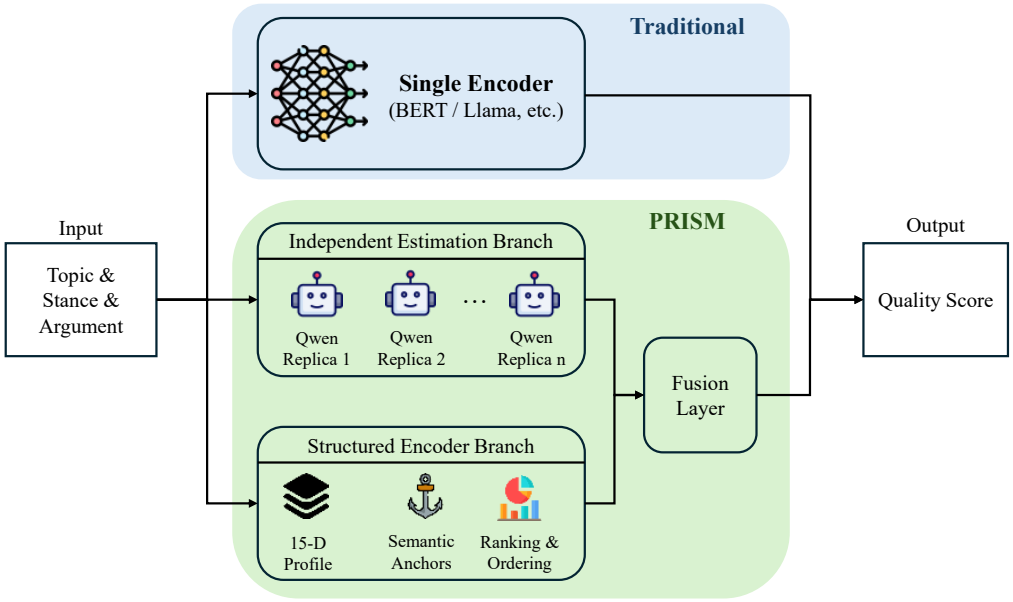


Figure 1: PRISM combines independent text-based scoring with profile-based encoder calibration through late fusion.

We introduce PRISM, a modular framework that assigns text-based scoring, profile calibration, and cross-corpus adaptation to separate interfaces. A gold-conditioned teacher creates structured training traces from fixed taxonomy excerpts, and a text-only student reconstructs profiles for a DeBERTa calibrator. An independently trained eight-replica Qwen3-8B ensemble estimates quality from observable text. Freezing its predictions in factorial comparisons makes the encoder signal directly measurable. A separate IBM/Webis study evaluates protocol-aware routing and score-scale mapping.

Contributions. (1) We formulate AQA as a modular allocation problem with separate interfaces for text-based scoring, profile calibration, and cross-corpus adaptation. (2) We provide matched controls that quantify the encoder signal: taxonomy-aligned calibration raises encoder Spearman from .5814 to .5928, and the frozen profile-calibrated branch changes fused Spearman by .0038 over the Qwen ensemble alone. (3) In an IBM/Webis study, protocol-aware scale mapping and confidence-triggered fallback improve cross-corpus scoring while preserving separate expert interfaces.

2 Related Work

Argument quality assessment. Theory-derived criteria have long supported computational AQA Wachsmuth et al. (2016; 2017b;a); Persing & Ng (2017), while ranking, revision, graph, and multidimensional resources broaden its empirical scope Toledo et al. (2019); Gretz et al. (2020); Gienapp et al. (2020); Skitalinskaya et al. (2021); Marro et al. (2022); Ivanova et al. (2024). Contextual, comparative, multi-persona, and instruction-based systems improve aggregate prediction Wang et al. (2023); Bao et al. (2024); Jin et al. (2025); Stahl et al. (2025). Recent analyses further motivate scenario- and theory-guided LLM assessment Wachsmuth et al. (2024); Mirzakhmedova et al. (2024). PRISM complements these systems by separating profile calibration from an independently trained high-capacity scorer.

Privileged supervision and calibration. The teacher observes labels unavailable at inference, placing trace construction in the learning-with-privileged-information setting Vapnik & Vashist (2009); a text-only student transfers the intermediate structure through distillation Hinton et al. (2015). The trace is converted into a fixed numeric profile and evaluated through downstream calibration. We compare this interface against a separately trained generic teacher–student schema, as

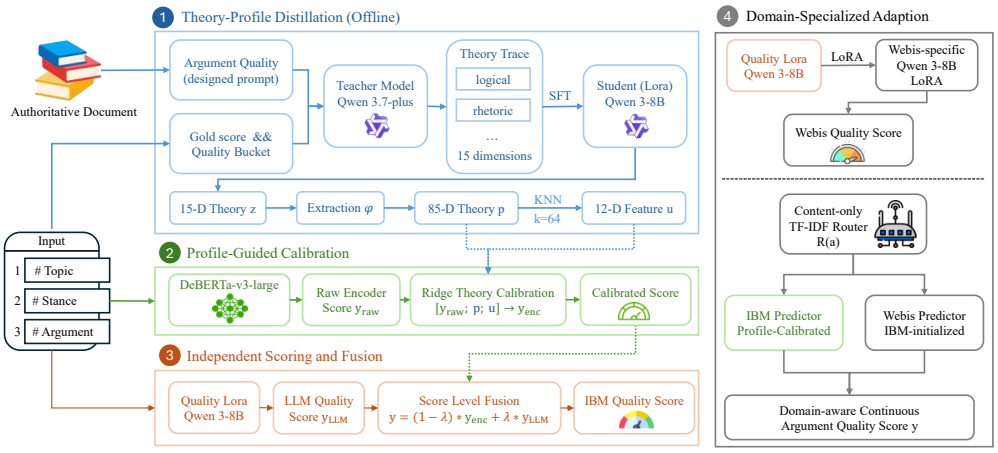


Figure 2: PRISM separates Qwen scoring, profile calibration, and protocol-aware routing. Gold scores supervise the trace teacher during training, while inference uses the text-only student and fixed taxonomy specification. The Qwen ensemble meets the calibrated encoder through scalar fusion. The IBM/Webis study uses architecture-matched experts with development-fitted scale maps.

well as semantic and equal-dimensional controls. The final scalar fusion is related to stacking and post-hoc calibration Guo et al. (2017); its independently trained branches enable complementarity measurement with frozen predictions.

Cross-corpus adaptation. Cross-dataset AQA requires protocol-aware adaptation because domains differ in both content and annotation protocol Fromm et al. (2023); Feger et al. (2025); Romberg et al. (2025). We study a closed-set setting in which the requested scale is known, separate experts retain their native outputs, a calibrated content router selects an expert, and low confidence triggers scale-mapped fusion. This use of hard-decision abstention follows the selective-prediction principle Geifman & El-Yaniv (2017) while returning a score for every input.

3 Method

Given topic t, stance s, argument a, and quality y, let $\mathbf{x} = (t, s, a)$. PRISM learns an encoder score $\hat{y}_{\text{enc}}$ and an independently trained LLM score $\hat{y}_{\text{LLM}}$, then combines them as

$$\hat{y} = (1 - \lambda)\hat{y}_{\text{enc}} + \lambda\hat{y}_{\text{LLM}}. \tag{1}$$

The separation is architectural: the encoder and Qwen branches are trained through separate interfaces and meet through scalar fusion.

3.1 Taxonomy-Profile Construction

An offline Qwen3.7-Plus teacher receives $\mathbf{x}$, gold quality y, a coarse quality bucket b, and fixed excerpts defining 15 criteria from the Wachsmuth taxonomy Wachsmuth et al. (2017b):

$$\mathbf{z}^{\text{tea}} = \mathcal{T}(\mathbf{x}; y, b). \tag{2}$$

The JSON trace contains hierarchical criterion scores, grounded evidence spans, an overall score, and uncertainty. Filtering enforces parse validity, score range, evidence grounding, parent–child consistency, and pairwise probes. From 20,974 IBM training arguments, 13,070 traces remain. A Qwen3-8B Yang et al. (2025) student with LoRA Hu et al. (2022) reconstructs the trace from $\mathbf{x}$ and the fixed taxonomy instructions. At inference, the text-only student reconstructs the profile directly from observable input fields.

Fixed sequential checks use thresholds of .12 for score deviation, .35 for summed hierarchy deviation, and .67 three-probe accuracy, require exact normalized evidence spans, and reject 5,077/1,345/753/729 cases, respectively.

Table 1: The 15 taxonomy-aligned coordinates used as trace targets. The first 14 follow the source taxonomy's hierarchy; the holistic coordinate aligns the structured training target with IBM overall quality.

Aspect	Coordinates in the trace	Operational focus
Cogency	cogency; local acceptability, relevance, and sufficiency	Premise acceptability and support for the conclusion
Effectiveness	effectiveness; credibility, emotional appeal, clarity, appropriateness, and arrangement	Audience-oriented rhetorical force and presentation
Reasonableness	reasonableness; global acceptability, relevance, and sufficiency	Contribution to resolving the issue and counterarguments
Holistic	overall argumentation quality	Aggregate judgment across the taxonomy

The trace provides structured predictive supervision for downstream calibration. Its dimensions are evaluated through downstream prediction and matched controls.

The student uses 12,489 traces for training and 581 topic-held-out traces for evaluation. Its 4-bit Qwen3-8B backbone follows QLoRA-style adaptation Dettmers et al. (2023), with rank-32, scale-64 adapters on the attention and feed-forward projections. Three epochs use micro-batch 2, gradient accumulation 32 (effective batch 64), learning rate 2×10^{-5}, cosine decay, and bfloat16 computation. Beyond token likelihood, the objective supervises quality, dimension values, hierarchy, and grounded evidence:

$$L = L_{\text{tok}} + L_y + 0.6L_z + 0.3L_h + 0.2L_g. \tag{3}$$

Here L_y is scalar-quality error, L_z scores the 15 coordinates, L_h penalizes hierarchy violations, and L_g upweights tokens belonging to grounded evidence. These targets encourage structured completion while inference operates from observable input fields.

3.2 Profile and Anchor Calibration

For each of the 15 criteria, the student yields a parsed score vector $\mathbf{t}$ and a token-logit expectation $\mathbf{o}$. We construct

$$\mathbf{p} = [\mathbf{o}; \mathbf{t}; \mathbf{o} - \mathbf{t}; |\mathbf{o} - \mathbf{t}|; \mathbf{o} \odot \mathbf{t}; \mathbf{s}] \in \mathbb{R}^{85}, \tag{4}$$

where $\mathbf{s} \in \mathbb{R}^{10}$ contains distribution summaries, predicted quality, and uncertainty. Cosine retrieval over ℓ_2-normalized profiles returns $k = 64$ training neighbors with labels y_j and ordered similarities $q_1 \geq \cdots \geq q_k$. We normalize

$$w_j = \frac{\exp(12 \max(q_j, 0))}{\sum_{\ell=1}^{k} \exp(12 \max(q_\ell, 0))} \tag{5}$$

and form the 12-dimensional anchor vector from the weighted label mean and median; unweighted label mean, standard deviation, minimum, and maximum; nearest-neighbor, top-3, top-5, and all-neighbor mean similarities; similarity standard deviation; and weight entropy. Thus $\mathbf{u}$ records both local label structure and retrieval concentration. Training queries exclude themselves; validation and test queries retrieve from the training pool. A standardized ridge map chosen on validation topics produces

$$\hat{y}_{\text{enc}} = f_{\text{ridge}}(\hat{y}_{\text{raw}}, \mathbf{p}, \mathbf{u}), \tag{6}$$

where $\hat{y}_{\text{raw}}$ is the five-member DeBERTa-v3-large He et al. (2021) committee score. For the weighted median, labels are sorted increasingly and the first label whose cumulative weight reaches .5 is returned; the anchor coordinates follow the order listed above.

3.3 Independent Scoring and Fusion

The Qwen scoring branch is a Qwen3-8B LoRA regressor trained directly on $\mathbf{x}, y$ from observable fields. Eight independently trained replicas are averaged. IBM validation topics select $\lambda = .65$, which is fixed on test. For the factorial ablation, raw-score and profile-calibrated encoder branches use the same frozen Qwen ensemble and the same λ; only the encoder interface changes in Eq. equation 1.

3.4 Closed-Set Scale-Aware Case Study

The IBM/Webis study uses architecture-matched three-replica Qwen experts for both corpora in a closed set with known output protocols. A content-only TF–IDF router selects an expert, and router probabilities are Platt-calibrated on held-out data Platt et al. (1999). A high-confidence decision uses the selected expert after mapping its output to the requested scale. Low confidence triggers evaluation of both experts and selective fusion:

$$\hat{y}^{(d)}_{\text{sel}} = w_d \hat{y}^{(d)} + (1 - w_d) m_{\bar{d}\to d}(\hat{y}^{(\bar{d})}), \tag{7}$$

where the monotone affine map $m_{\bar{d}\to d}$ and native-expert weight w_d are fitted on development data. All maps, calibration parameters, mixture weights, and the confidence threshold are selected before test evaluation.

4 Experimental Setup

Data and metrics. IBM-Rank-30k Gretz et al. (2020) uses the official topic-disjoint split: 20,974/3,208/6,315 arguments from 49/7/15 train/validation/test topics. Webis-ArgQuality-20 Gienapp et al. (2020) uses a fixed discussion-disjoint split of 889/191/191 arguments from 545/116/116 discussions and predicts Combined Quality. We report Pearson r, Spearman ρ, Kendall's τ_b, and MAE, matching MPAQ Jin et al. (2025). Published IBM rows provide reference values; all within-system comparisons use frozen per-example predictions.

Selection and inference. The DeBERTa committee has five seeds, the IBM quality ensemble eight, and each routing expert three. Ridge penalties and λ use IBM validation topics; $k = 64$ is fixed for the main system before the frozen sensitivity audit. Webis adapters, router calibration, affine maps, and the selective-fusion threshold use training/development data. No test label is used for model selection. Teacher decoding is deterministic, and all reported ensembles are fixed before test evaluation.

The encoder input concatenates topic, stance, and argument and is truncated to 256 tokens. Three committee members optimize MSE with correlation and pairwise-ranking terms Burges et al. (2005); Liu (2009); two validation-selected diversity members emphasize low-quality and hard-pair examples. Their unweighted mean is frozen before calibration. The independent quality branch loads the full Qwen3-8B in bfloat16 with a rank-16 LoRA and scalar head, trained from observable text fields. We average eight replicas with odd-numbered seeds 221, 223, . . ., 235. Webis likewise uses full-bfloat16 Qwen3-8B, rank-16 adapters, and an unbounded regression head on its native scale, averaging seeds 301/303/305. The routing experiment uses architecture-matched three-replica experts.

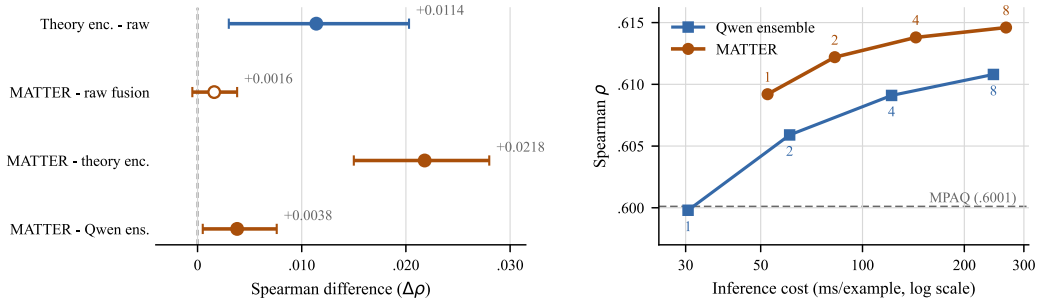
Controls. Semantic neighbors replace profile retrieval with frozen DeBERTa similarity. An 85-dimensional pseudo-schema uses the frozen raw prediction, its powers and soft bins, semantic training-label anchors, and fixed random projections. We also evaluate profile ablations that remove predicted quality, uncertainty, or summary coordinates and refit the complete calibration pipeline. The strongest control trains a separate Qwen3.7-Plus teacher and Qwen3-8B student with the same 15-field, profile, anchor, decoding, and validation interfaces, replacing the taxonomy with observable text descriptors. This deterministic comparison uses the 6,297-row common complete test intersection.

Uncertainty. Primary IBM comparisons use 20,000 paired topic-cluster bootstrap replicates Efron (1992); Webis comparisons resample discussions. The four strict theory–generic metrics form a predefined family and receive Holm correction Holm (1979). Auxiliary semantic and pseudo-schema controls are reported as supplementary attribution analyses.

Evaluation protocol. We evaluate encoder calibration against the raw committee and a raw-score ridge, and test taxonomy specificity against a capacity-matched generic teacher–student control. Complementarity is isolated by holding the Qwen predictions and $\lambda = .65$ fixed while changing the encoder interface. The IBM/Webis study compares target-supervised adaptation, scale-mapped

Table 2: IBM-Rank-30k test results. Selected directly comparable prior-work rows are reference values reported by MPAQ. The final two rows use the same frozen Qwen ensemble and fixed $\lambda = .65$; their difference isolates the encoder branch.

Family	Base Model	Method	Pear.	Spear.	τ_b	MAE
Prior work	BERT-base	CompAQA Bao et al. (2024)	.5282	.4830	.3390	.1311
	RoBERTa-base	CompAQA Bao et al. (2024)	.5642	.5204	.3670	.1299
	DeBERTa-base	CompAQA Bao et al. (2024)	.5797	.5373	.3794	.1371
		Llama-Reg-EOS Jin et al. (2025)	.6189	.5716	.4085	.1206
	Llama-3.1-8B	CompAQA Bao et al. (2024)	.6213	.5791	.4118	.1216
		MPAQ Jin et al. (2025)	.6406	.6001	.4280	.1196
Ours	DeBERTa-v3-large	Raw committee	.6286	.5814	.4144	.1252
		Raw-score ridge	.6286	.5814	.4144	.1229
		Profile-calibrated encoder	.6382	.5928	.4232	.1198
	Qwen3-8B	Ensemble	.6516	.6108	.4374	.1165
	DeBERTa-v3-large + Qwen3-8B	Raw-score ridge fusion	.6561	.6130	.4393	.1170
		PRISM	**.6569**	**.6146**	**.4406**	**.1164**



(a) Paired effects with 95% CI

(b) Accuracy-efficiency trade-off

Figure 3: Paired attribution and the accuracy–efficiency trade-off. Left: Spearman differences with 95% topic-cluster intervals. Right: Qwen-only and fused results as the number of Qwen replicas increases; labels denote replica counts and the dashed line is the published MPAQ reference. Costs are A800 batch-8 sequential estimates including adapter loads; fused costs add 21.9 ms/example for the cached-profile encoder committee.

routing, and confidence-triggered fallback. All selection uses validation or development data, while uncertainty is estimated over IBM topics or Webis discussions.

The comparisons form a nested attribution design. The raw-score ridge isolates affine calibration; profiles and anchors test structured coordinates and local geometric information. The generic control holds the teacher/student families, field count, decoding protocol, profile dimensionality, anchor construction, and ridge selection fixed while replacing taxonomy semantics. The fixed-λ factorial preserves the complete Qwen branch while changing the encoder interface. These experiments assign separate roles to encoder calibration, final-system complementarity, and protocol-aware adaptation.

Published comparisons measure final-system quality; matched frozen predictions attribute component gains, and clustered resampling supplies uncertainty. Published MPAQ values are reported as reference results.

5 Results

5.1 Main IBM Performance and Attribution

Table 2 reports the modular allocation result. The eight-replica Qwen ensemble reaches .6108 Spearman; profile calibration raises the standalone encoder from .5814 to .5928 Spearman (95% topic-

Table 3: Frozen IBM encoder-interface decomposition. The anchor-only row removes direct profile coordinates but still retrieves in profile space.

Encoder calibration input	ρ	**MAE**
Raw committee	.581422	.125188
Raw-score ridge	.581422	.122881
Raw + direct 85D profile	.591464	.119936
Raw + 12D profile anchors	.592729	.119934
Raw + profile + anchors	**.592834**	**.119801**

Table 4: Capacity-matched deterministic control on 6,297 common test rows. Taxonomy alignment improves all reported metrics on the common test intersection.

Profile student	r	ρ	τ_b	**MAE**
Generic structured	.6255	.5826	.4144	.1205
Taxonomy aligned	**.6326**	**.5896**	**.4205**	**.1197**
Difference	+.0071	+.0071	+.0061	$-.0008$

cluster CI for the change $[.0030, .0203]$, $p = .0083$) and reduces MAE from $.1252$ to $.1198$. With the Qwen predictions and fusion weight fixed, the calibrated encoder changes fused Spearman by $.0038$ over Qwen alone, producing the final $.6569/.6146$ Pearson/Spearman result. This final system is $2.5\%/2.4\%$ higher in Pearson/Spearman than the strongest published reference in our comparison.

Figure 3(a) summarizes the frozen factorial. With Qwen predictions and $\lambda = .65$ fixed, profile-calibrated fusion adds $+.0016$ Spearman over raw-score fusion and $+.0038$ over Qwen alone. These paired comparisons quantify the incremental encoder signal separately from the ensemble score.

5.2 Where Does the Encoder Gain Enter?

Table 3 decomposes the calibration input. A ridge map on the raw score alone is rank-preserving, so it leaves correlations unchanged while reducing MAE through scale and offset correction. Direct profile coordinates and profile-derived neighbor anchors provide complementary signals, with the combined interface achieving the strongest encoder result. The anchors represent local label geometry in the learned taxonomy-profile space.

5.3 Does Taxonomy Alignment Add Beyond Generic Distillation?

Table 4 is a controlled analysis of the profile construction: both pipelines use the same teacher/student families, 15-field structure, deterministic decoding, profile/anchor dimensions, and ridge selection. Replacing generic descriptors with the taxonomy coordinates improves all three encoder correlations by $.0061$–$.0071$ and lowers MAE by $.0008$. This identifies taxonomy alignment as a useful design choice within the encoder interface; the comparison does not change the Qwen scorer.

The strict control re-decodes both students with one deterministic protocol and evaluates their 6,297-row common complete intersection.

The encoder gain remains stable when predicted quality, uncertainty, and summary coordinates are removed. The aligned score coordinates and profile-derived anchors retain the improvement, supporting their role as encoder-side complementary features.

Independent multi-annotator validation of individual coordinates is a natural extension of the profile interface.

5.4 Complementarity and Compute Budget

Figure 3(b) presents the replica-scaling profile: one-replica PRISM reaches $.6092$ Spearman at 52.4 ms/example, while eight replicas reach $.6146$ at 265.7 ms/example. In the highest-disagreement

Table 5: Training-trace retention and IBM test MAE by predefined quality interval. Retention is measured on teacher traces; errors use the complete test predictions.

Gold-quality interval	Train retained	Test n	Raw MAE	Profile MAE	Qwen MAE	PRISM MAE
$y < .45$	39.9%	489	**.2154**	.2893	.2661	.2733
$.45 \leq y < .60$	60.2%	642	.1709	.1757	**.1680**	.1687
$.60 \leq y < .75$	70.9%	1,161	.1368	**.1114**	.1156	.1124
$y \geq .75$	62.5%	4,023	.1036	.0927	.0904	**.0902**

Table 6: Target adaptation and a closed-set routing stress test with architecture-matched three-replica experts. One logical call denotes one aggregate expert; the known native target-protocol expert is the reference upper row.

Setting	IBM ρ	Webis ρ	Calls I/W
IBM Qwen zero-shot	–	.3282	–
Single IBM DeBERTa zero-shot	–	.5910	–
Webis-only aggregate	–	.7864	–
IBM-init. Webis aggregate	–	.7910	–
Hard routing	.6046	.5477	1.00/1.00
Scale-mapped hard	.6067	.7005	1.00/1.00
Selective fusion	.6066	**.7794**	1.09/1.50
Soft gating	**.6069**	.7345	2.00/2.00

quartile, fusion raises Qwen Spearman from .5891 to .6037. Estimates assume cached profiles; uncached trace generation adds 3.88 s. Peak memory is 8.2 GiB for DeBERTa, 15.9 GiB for scalar Qwen, and 13.7 GiB for full traces.

5.5 Robustness and Operating Range

The encoder result is stable across $k \in \{16, 32, 64, 128, 256\}$, where test Spearman spans .5923–.5928. Varying λ from .55 to .75 gives fused Spearman .6140–.6146. Excluding same-topic training anchors from the strict theory–generic comparison retains a $+.0070$ gap, and both profiles select ridge $\alpha = 10^4$. These checks use frozen predictions with $k = 64$ fixed as the predefined main setting before test evaluation.

Performance ordering is stable across argument-length strata: profile calibration improves the raw encoder in all three ranges, and the fused system improves over Qwen. The respective stratum sizes are 1,566/3,591/1,158, with raw-to-calibrated encoder Spearman changing $.6527 \rightarrow .6635$, $.5499 \rightarrow .5632$, and $.4691 \rightarrow .4818$.

Quality-stratified analysis shows where profile calibration contributes most: it reduces encoder MAE in the .60–.75 interval and remains beneficial above .75. The retention profile also identifies low-quality trace coverage as a direct target for future data construction.

Trace retention varies across stance and topic groups, identifying coverage as a direct target for future data construction. Retention is 57.6% for supporting versus 67.2% for contesting arguments and ranges from 45.5% to 71.7% across the 49 training topics. The generic teacher retains 12,928 of the 13,070 theory-retained IDs (98.9%), providing a capacity-matched control on the same retained population.

5.6 Closed-Set IBM/Webis Case Study

Target-domain supervision is the key driver of Webis performance: the Webis-trained Qwen aggregate reaches .7864 Spearman, while the IBM-trained zero-shot baseline reaches .3282.

The content router attains 99.23% ordinary accuracy. Scale mapping raises Webis Spearman from .5477 to .7005, and confidence-triggered selective fusion reaches .7794 while evaluating both experts on 49.7% of Webis items and 9.2% of IBM items. Its gain over scale-mapped hard routing is

$+.0789$ ($[.0228, .1421]$, adjusted $p = .0054$), demonstrating that routing and scale calibration should be designed jointly.

The operating rule is fully development-fitted. A threshold of $.99$ minimizes mean range-normalized development MAE subject to at most 1.25 logical expert calls per example on the fixed pooled IBM/Webis development mixture. The Webis-to-IBM map is $.1591x + .8825$ and the IBM-to-Webis map $.8534x - .4451$. Fitted native-expert weights are $.9749$ on IBM and effectively 1.0 on Webis. The resulting rule combines confidence-triggered native-expert selection with scale-aware fallback.

5.7 Reproducibility and Audit Trail

All headline and matched comparisons are computed from frozen per-example prediction files. The retained project audit trail records split identifiers, SHA-256 hashes, teacher/student manifests, prompt builders, selected checkpoints, seeds, validation grids, and scripts for each table and clustered interval. Deterministic IDs align predictions before any metric is calculated, and common-intersection controls never impute missing parses. Theory profiles cover 6,314/6,315 IBM test rows; generic profiles cover 6,298, giving the reported 6,297-row intersection with DeBERTa predictions.

The Webis split is deterministic and discussion-disjoint. A label-independent split audit records duplicate checks for the final analysis.

We report component-level operating points measured on an A800 with batch size 8: DeBERTa and scalar Qwen timing use 512 examples, score-only profile generation 256, and full-trace generation 16. These measurements provide a reproducible reference for the accuracy–cost comparison.

6 Discussion and Limitations

The experiments establish a modular allocation pattern: the eight-replica Qwen ensemble, profile calibration, and protocol-aware scale handling each have separately measured effects on IBM and cross-corpus scoring.

The measurements also clarify how these interfaces should be used together. The Qwen branch supplies the strongest broad text-based signal, while the profile branch contributes a calibrated view whose value is most visible in the middle and upper quality intervals and in the highest-disagreement cases. The routing study adds a separate operational dimension: when the requested protocol is known, native experts and development-fitted scale maps combine with confidence-triggered evaluation to preserve corpus-compatible score semantics. These roles can therefore be tuned independently as the data distribution or compute budget changes.

The framework supports several direct extensions: expand low-quality trace coverage, validate learned dimensions with independent annotators, broaden evaluation across protocols and languages, and distill the multi-replica scorer into a lower-cost model.

7 Conclusion

PRISM organizes argument quality assessment as modular allocation: an eight-replica Qwen ensemble, a profile-calibrated encoder, and protocol-aware routing provide separate interfaces for scoring and adaptation. On IBM-Rank-30k, the frozen profile-calibrated branch changes Spearman by $.0038$ over the Qwen ensemble, yielding $.6569/.6146$ Pearson/Spearman. Matched controls quantify the encoder signal and the role of profile design while keeping the scoring branches separately identifiable.

Across the controlled studies, the framework turns a single aggregate score into a set of measurable operating points. The ensemble establishes the main text-scoring signal, calibration supplies a complementary encoder interface, and the closed-set study shows how scale compatibility and confidence can govern expert selection. The resulting design preserves a fixed prediction path for reproducible comparisons while allowing the allocation of replicas, calibration features, and expert calls to match the target setting. This combination provides a practical basis for extending AQA across annotation protocols with explicit control over scoring, calibration, and scale adaptation.

AI USE STATEMENT

Generative AI assistants, including GPT-5.5, OpenAI Codex, and GitHub Copilot, were used in an author-supervised capacity to reorganize and revise author-drafted manuscript text, polish grammar and style, and provide code-completion suggestions. AI assistance helped express and organize the authors' research and findings; it was not used to conduct experiments, generate data, or produce the reported results. The authors reviewed and validated all AI-assisted text and code, verified the technical claims against their experiments, and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies automated argument-quality assessment using publicly available benchmark datasets and does not recruit participants or collect new human-subject data. Argument-quality scores can encode annotator preferences and topic biases, and model outputs may reproduce them. The generated coordinates and evidence spans are evaluated for predictive utility; their construct validity and explanatory faithfulness have not been independently established. Accordingly, these scores should not be used for automated moderation, access, or educational decisions without qualified human review. The study reports corpus-level results on topic- and discussion-disjoint splits, and dataset use remains subject to the original providers' licenses and terms.

REPRODUCIBILITY STATEMENT

The paper specifies the model interfaces, training objectives, data splits, retrieval construction, validation-based selection rules, ensemble settings, and evaluation metrics. The *Experimental Setup* and *Reproducibility and Audit Trail* sections identify the frozen prediction files, deterministic IDs, seeds, checkpoints, prompt builders, validation grids, bootstrap and multiple-comparison procedures, and hardware settings used for the reported results. Tables and figures are generated from these fixed per-example predictions, and the IBM/Webis operating rules are fitted entirely on the development data before test evaluation. The authors retain the split IDs and dataset hashes, prompt builders, trace manifests, model configurations and checkpoints, per-example predictions, analysis scripts, package versions, and generation manifests described in the original reproducibility record. These artifacts support independent verification and can be provided in an anonymized supplementary package.

REFERENCES

Jianzhu Bao, Bojun Jin, Yang Sun, Yice Zhang, Yuhang He, and Ruifeng Xu. A comparison-based framework for argument quality assessment. *Electronics*, 13(20):4088, 2024.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Marc Feger, Katarina Boland, and Stefan Dietze. Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23900–23915, 2025.

Michael Fromm, Max Berrendorf, Evgeniy Faerman, and Thomas Seidl. Cross-domain argument quality estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13435–13448, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5772–5781, 2020.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7805–7813, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Ivan Habernal and Iryna Gurevych. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1214–1223, 2016a.

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1589–1599, 2016b.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Rositsa V Ivanova, Thomas Huber, and Christina Niklaus. Let's discuss! quality dimensions and annotated datasets for computational argument quality assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20749–20779, 2024.

Bojun Jin, Jianzhu Bao, Yufang Hou, Yang Sun, Yice Zhang, Huajie Wang, Bin Liang, and Ruifeng Xu. A multi-persona framework for argument quality assessment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12148–12170, 2025.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

Santiago Marro, Elena Cabrio, and Serena Villata. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4154–4164, 2022.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. Are large language models reliable argument quality annotators? In *Conference on advances in robust argumentation machines*, pp. 129–146. Springer, 2024.

Isaac Persing and Vincent Ng. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 4082–4088, 2017.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. Towards a perspectivist turn in argument quality assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7458–7485, 2025.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1718–1729, 2021.

Maja Stahl, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth. Arginstruct: specialized instruction fine-tuning for computational argumentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11103–11127, 2025.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic argument quality assessment-new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5625–5635, 2019.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: Technical papers*, pp. 1680–1691, 2016.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 250–255, 2017a.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 176–187, 2017b.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pp. 1519–1538, 2024.

Yiran Wang, Xuanang Chen, Ben He, and Le Sun. Contextual interaction for argument post quality assessment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 10420–10432, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Implementation Properties and Derivations

This section formalizes the properties governing score fusion, routing cost, fallback behavior, calibration, and uncertainty estimation.

A.1 Notation

For an input argument x, the encoder branch produces $e(x)$, the independent quality-LoRA produces $q(x)$, and score-level fusion is

$$f_\lambda(x) = (1-\lambda)e(x) + \lambda q(x), \qquad \lambda \in [0,1]. \tag{8}$$

For corpus expert j and a known requested output scale d, the development-fitted mapping is $m_{j\to d}(y) = a_{jd}y + b_{jd}$ with $a_{jd} > 0$. Router probabilities are $\pi_j(x)$ and router confidence is $c(x) = \max_j \pi_j(x)$.

A.2 Rank invariance of a positive affine map

Implementation Property 1. Let $m(y) = ay + b$ with $a > 0$. Applying m uniformly to a finite prediction vector preserves all pairwise order relations. Consequently, Spearman correlation and Kendall's τ_b against the gold labels are unchanged, apart from ties already present.

Derivation. For predictions y_i, y_j,

$$m(y_i) - m(y_j) = a(y_i - y_j). \tag{9}$$

The sign of every pairwise difference is unchanged because $a > 0$. The rank vector, concordant pairs, discordant pairs, and tied pairs therefore remain the same. Scale mapping preserves the selected expert's ranking while aligning its outputs with the requested cross-corpus scale.

A.3 Expert-call accounting for confidence-triggered fusion

Implementation Property 2. In the two-expert policy, one aggregate expert is evaluated when $c(x) \geq t$ and both aggregate experts are evaluated when $c(x) < t$. The expected logical call count is

$$\mathbb{E}[C_t] = 1 + \Pr[c(x) < t] = 2 - \text{coverage}(t). \tag{10}$$

Derivation. The call count is $C_t = 1 + \mathbb{I}[c(x) < t]$. Taking expectations gives the result. A logical expert is a frozen three-replica aggregate in the routing study, not one underlying LoRA replica.

A.4 Convex-hull behavior of the fallback score

Implementation Property 3. For mapped expert outputs $z_1, z_2 \in \mathbb{R}$ and fitted weight $\alpha \in [0, 1]$, the confidence-triggered fallback

$$\hat{y} = \alpha z_1 + (1 - \alpha) z_2 \tag{11}$$

lies in $[\min(z_1, z_2), \max(z_1, z_2)]$.

Derivation. The result follows directly from convexity. The fallback is a convex combination of the two mapped expert outputs and therefore remains between them. The implementation applies no additional clipping or output transformation.

A.5 Ridge calibration

For profile vector $p_i \in \mathbb{R}^{85}$, anchor vector $u_i \in \mathbb{R}^{12}$, and raw encoder prediction r_i, define $v_i = [r_i; p_i; u_i]$. The standardized ridge estimator minimizes

$$\sum_{i \in \mathcal{D}_{\text{train}}} (y_i - b - w^\top v_i)^2 + \alpha \|w\|_2^2. \tag{12}$$

For centered standardized design matrix V and centered target $\mathbf{y}_c$,

$$\hat{w} = (V^\top V + \alpha I)^{-1} V^\top \mathbf{y}_c, \qquad \hat{b} = \bar{y} - \hat{w}^\top \bar{v}. \tag{13}$$

The frozen main configuration uses validation-selected $\alpha = 10{,}000$ and predefined $k = 64$ neighbors. Training anchors exclude the query row; validation and test anchors retrieve only from the training pool.

A.6 Paired cluster bootstrap

For a system comparison, both systems are evaluated on the same resampled clusters. If the observed clusters are $G_1, \ldots, G_M$, one replicate samples M cluster indices with replacement and concatenates all examples in the selected clusters. IBM uses topic clusters and Webis uses discussion clusters. We use 20,000 replicates, percentile 95% intervals, and two-sided plus-one Monte Carlo p-values,

$$p = \min\left(1, \frac{2(B_{\text{opp}} + 1)}{B + 1}\right), \tag{14}$$

where B_{opp} counts bootstrap differences on the side opposite the observed effect. Holm correction is applied within each predefined comparison family.

B Datasets and Fixed Splits

Dataset split	Examples	Topics	Discussions
IBM train	20,974	49	–
IBM validation	3,208	7	–
IBM test	6,315	15	–
Webis train	889	20	545
Webis development	191	20	116
Webis test	191	20	116

Table 7: Fixed data partitions. IBM follows the official topic-disjoint split. Webis is discussion-disjoint.

IBM-Rank-30k Gretz et al. (2020). Each example contains a topic, stance, argument, and continuous weighted-average quality label in $[0, 1]$. We use the official 49/7/15 topic-disjoint train/validation/test split without moving topics between partitions.

Webis-ArgQuality-20 Gienapp et al. (2020). We retain rows for which `IsArgument?` is true, yielding 1,271 eligible arguments. The target is the unbounded continuous `CombinedQuality` score; predictions are neither clipped nor transformed. Whole discussions are assigned deterministically to an approximate 70/15/15 split using seed 20270712, giving 889/191/191 arguments. Stable IDs are the first 20 hexadecimal characters of the SHA-256 hash over Topic ID, Discussion ID, Argument ID, and Premise. All Webis bootstrap analyses resample Discussion ID rather than rows.

Duplicate audit. A label-independent audit finds two normalized exact train–test duplicates and three train–test pairs with character 3–5-gram cosine at least .98. Removing from evaluation the two test items whose normalized text occurs in training ($n = 189$) changes IBM-initialized/Webis-only aggregate Spearman from $.7910/.7864$ to $.7867/.7820$. Excluding instead all three test items whose nearest training argument has cosine at least .98 ($n = 188$) gives $.7855/.7818$. The reported ordering is stable under both text-overlap filters, which operate on the original frozen test predictions.

C Prompt Interfaces and Structured Fields

C.1 Teacher, student, and generic-control prompts

The teacher API model ID is `qwen3.7-plus`, accessed through the Alibaba Cloud Bailian OpenAI-compatible endpoint. Temperature is .2, maximum completion length is 768 tokens, JSON mode is enabled, thinking mode is disabled, and generation is non-streaming. The system message defines one internally consistent expert. The IBM task bridge defines quality as whether an argument should be recommended for use as-is in a speech supporting or contesting the topic, irrespective of the evaluator's personal position. The output contract requires the specified JSON fields in a compact, machine-readable response.

C.2 Theory-coordinate inventory

The first 14 entries follow the definitions and hierarchy of Wachsmuth et al. (2017b). We instantiate overall argumentation quality as a holistic coordinate aligned with the IBM target, complementing the 14 hierarchical taxonomy coordinates.

C.3 Capacity-matched generic-field inventory

D Trace Construction, Filtering, and Student Training

D.1 Sequential trace filters

The score check requires $|\hat{y}_{\text{teacher}} - y| \leq .12$. An evidence span must be an exact normalized substring of the argument. The sum of the three parent-to-child-mean deviations must not exceed

Phase	Fixed system content	Instance-specific user content	Required output or target
Theory teacher	One expert, no personas; five fixed taxonomy passages G0–G4; IBM task bridge; exact JSON-only contract	Topic, stance, argument, privileged gold label, and its coarse bucket; 30% Task-B rows also receive an explicitly non-authoritative wrong bucket	Fifteen bounded coordinates, at most one exact evidence span, `quality_score`, and uncertainty
Theory student	The same fixed taxonomy passages, task bridge, trace instruction, and JSON schema	Topic, stance, and argument only; no label, bucket, or Task-B cue	Filtered teacher JSON after replacing `quality_score` and `overall_argumentation_quality` with IBM gold and moving `quality_score` to the first JSON position
Generic teacher	Single-expert wrapper and identical output interface; theory passages replaced by generic definitions plus 13 neutral context-budget notes	The same privileged construction fields and 70/30 Task-A/Task-B protocol	Fifteen generic fields, at most one exact evidence span, scalar quality, and uncertainty
Generic student	Fixed generic definitions, notes, instruction, and schema	Topic, stance, and argument only	Gold-aligned filtered generic JSON with the same ordering and loss interface as the theory student

Table 8: Exact information boundary. The student has only observable instance fields, while the fixed theory or generic wrapper remains in its prompt.

.35. Three same-topic, same-stance probes use a gold margin of .05 and require at least .67 pairwise accuracy.

D.2 Trace retention across data strata

The theory filter retains 13,070/20,974 traces (62.3%). Retention is characterized across predefined quality, length, stance, and topic strata.

Retention by argument length is 58.4% for at most 12 words, 65.2% for 13–25 words, and 59.0% for at least 26 words. It is 57.6% for supporting and 67.2% for contesting arguments and ranges from 45.5% to 71.7% across the 49 training topics. Retained-versus-all standardized mean differences are $-.033$ for quality and $+.002$ for length. The generic teacher is run on the 13,070 theory-retained IDs and retains 12,928 (98.9%), preserving the matched comparison population.

D.3 Theory and generic student optimization

The theory traces are divided by a deterministic topic split into 12,489 training and 581 evaluation rows. The Qwen3-8B student is loaded in 4-bit with bfloat16 computation. LoRA uses $r = 32$, $\alpha = 64$, dropout .1, and the `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules. Maximum sequence length is 4,096. Optimization uses three epochs, learning rate 2×10^{-5}, cosine decay, warmup ratio .03, weight decay .01, micro-batch size 2, and 32-step gradient accumulation, for effective batch size 64. The update count is therefore

$$3 \left\lceil \frac{\lceil 12489/2 \rceil}{32} \right\rceil = 588. \tag{15}$$

The composite objective is

$$L = L_{\text{tok}} + L_y + 0.6L_z + 0.3L_h + 0.2L_g, \tag{16}$$

where L_y is scalar quality error, L_z supervises the 15 coordinates, L_h penalizes parent–child inconsistency, and L_g upweights grounded evidence tokens. The generic student uses the same model,

Field	Role	Operational meaning
cogency	Parent	Premises are acceptable, relevant to the conclusion, and jointly sufficient to draw it.
local_acceptability	Cogency child	A premise is rationally worthy of being believed true.
local_relevance	Cogency child	A premise contributes to accepting or rejecting the conclusion.
local_sufficiency	Cogency child	The premises together provide enough support to rationally draw the conclusion.
effectiveness	Parent	The argumentation persuades, or corroborates agreement with, the stated stance for the target audience.
credibility	Effectiveness child	The presentation makes the author worthy of credence.
emotional_appeal	Effectiveness child	Emotions are used in a way that makes the audience more open to the argument.
clarity	Effectiveness child	Language is correct and broadly unambiguous and avoids needless complexity or deviation.
appropriateness	Effectiveness child	Language supports credibility and emotion and is proportionate to the issue.
arrangement	Effectiveness child	The issue, reasons, and conclusion are presented in an appropriate order.
reasonableness	Parent	The argumentation contributes to resolving the issue sufficiently and acceptably for the target audience.
global_acceptability	Reasonableness child	The audience accepts both the consideration of the reasons and how they are stated.
global_relevance	Reasonableness child	The contribution helps resolve the issue or arrive at an ultimate conclusion.
global_sufficiency	Reasonableness child	Anticipated counterarguments are adequately rebutted.
overall_argumentation_quality	Holistic	Overall coordinate matched to the IBM gold label in the SFT target.

Table 9: Complete 15-coordinate theory inventory and implemented hierarchy.

LoRA, optimizer, batch, and loss settings. Its 12,928 traces split into 12,353 training and 575 evaluation rows, giving

$$3\left\lceil \frac{\lceil 12353/2 \rceil}{32} \right\rceil = 582 \tag{17}$$

updates. The comparison matches model capacity, optimization protocol, decoding, and downstream interfaces while evaluating each representation on its retained traces.

For the strict deterministic comparison, both students use one greedy candidate, seed 20260722, generation batch size 32, log-probability batch size 8, at most 384 new tokens, no generation-time anchors, length bucketing, and resumable streaming output.

E Predictive Branches and Optimization Details

Development grids and fixed values. The data-dependent grids are: anchor count $k \in \{16, 32, 64, 128, 256\}$; Ridge penalty $\alpha \in \{100, 1000, 3000, 10^4, 30000, 10^5\}$; IBM scalar-fusion weight $\lambda \in \{0, .05, \ldots, 1\}$; and routing threshold $\{.50, .51, \ldots, 1.00\}$. Ridge and fusion maximize IBM validation Spearman; the router minimizes mean two-domain normalized development MAE subject to at most 1.25 expert calls per example. The reported main settings are respectively $k = 64$, $\alpha = 10^4$, $\lambda = .65$, and threshold .99. All remaining optimization, capacity, filtering, precision, batch, and length values listed in this document were single settings fixed before the final run from memory constraints, standard practice, or matched-control requirements; they were not selected on test labels. Thus their number of tried values is one per listed setting. Checkpoints use only validation/development criteria.

Field	Role	Operational meaning
content_profile	Parent	Arithmetic mean of topic overlap, stance explicitness, and claim specificity.
topic_overlap	Content child	Lexical and referential overlap with the stated topic.
stance_explicitness	Content child	How explicitly the text signals support or contest stance.
claim_specificity	Content child	Presence of concrete, delimited claims rather than vague text.
discourse_profile	Parent	Arithmetic mean of the five discourse children below.
local_cohesion	Discourse child	Surface continuity between adjacent clauses and sentences.
global_cohesion	Discourse child	Consistency of entities and subject matter across the full text.
syntactic_fluency	Discourse child	Grammatical well-formedness and ease of syntactic processing.
information_density	Discourse child	Non-repeated propositional content per unit of text.
redundancy_control	Discourse child	Absence of unnecessary repetition or duplicated phrasing.
style_profile	Parent	Arithmetic mean of readability, lexical diversity, and hedging level.
readability	Style child	Ease with which a general reader can decode the wording.
lexical_diversity	Style child	Variation in content words, independent of argument quality.
hedging_level	Style child	Degree of qualification, uncertainty marking, and modal language.
overall_text_profile	Auxiliary	Gold-consistent scalar matching the output interface.

Table 10: Complete 15-field generic control inventory with matched dimensionality and prompt budget.

Filtering outcome	Count
Raw teacher rows	20,974
Rejected: score tolerance	5,077
Rejected: ungrounded evidence	1,345
Rejected: hierarchy consistency	753
Rejected: pairwise probe	729
Retained	13,070

Table 11: Theory-teacher trace filtering. Checks are sequential, so rejection counts are mutually exclusive.

Gold-quality bin	Available	Retained	Rate
$y < .45$	1,370	546	39.9%
$.45 \leq y < .60$	1,974	1,188	60.2%
$.60 \leq y < .75$	3,833	2,717	70.9%
$y \geq .75$	13,797	8,619	62.5%

Table 12: Theory-trace retention by predefined quality interval.

Computing environment. Experiments ran on Ubuntu 22.04.5 LTS (Linux 5.15) with an NVIDIA A800 80GB GPU, Intel Xeon Platinum 8358 CPU (128 logical CPUs available), and 1.0 TiB host RAM. The frozen software environment is Python 3.12.13, PyTorch 2.8.0+cu128, CUDA runtime 12.8, cuDNN 91002, Transformers 5.12.1, PEFT 0.19.1, Accelerate 1.14.0, bitsandbytes 0.49.2, Datasets 5.0.0, NumPy 2.3.2, SciPy 1.18.0, scikit-learn 1.9.0, and pandas 3.0.3.

Exact model identifiers. All local Qwen branches use the Hugging Face repository `Qwen/Qwen3-8B`, revision `b968826d9c46dd6066d109eabc6255188de91218`. The encoder backbone uses `microsoft/deberta-v3-large`, revision `64a8c8eab3e3`

Component	Backbone/precision	Optimization	Replicas/selection
Theory student	Qwen3-8B, 4-bit/bfloat16	LoRA $r = 32$, $\alpha = 64$; three epochs; 588 updates	Seed 42; evaluation loss
Generic student	Qwen3-8B, 4-bit/bfloat16	Matched LoRA, batch, and loss; 582 updates	Seed 42; same topic protocol
Encoder committee	DeBERTa-v3-large, fp32	Full fine-tuning plus nonlinear scalar head	Seeds 13,17,23,29,41; validation selection
IBM quality branch	Qwen3-8B, bfloat16	LoRA $r = 16$, $\alpha = 32$ plus last-token scalar head	Seeds 221–235; unweighted mean
Webis experts	Qwen3-8B, bfloat16	LoRA $r = 16$, $\alpha = 32$ plus unbounded last-token head	Seeds 301,303,305; development selection
Ridge/fusion	Standardized ridge/scalar blend	$\alpha = 10{,}000$; $\lambda \in \{0, .05, \dots, 1\}$	IBM validation topics

Table 13: Frozen component configurations. The IBM quality and strict Webis regressors load the full backbone in bfloat16 rather than 4-bit.

`52a784c658aef62be1662607476f`. Teacher calls use the API model ID `qwen3.7-plus`. These identifiers are recorded from the frozen configuration and local Hugging Face cache metadata; no unverified revision is inferred.

E.1 DeBERTa-v3-large committee

The encoder input is exactly `Topic: ...`, `Stance: ...`, and `Argument: ...`, truncated to 256 tokens. The regression head is dropout $\to$ Linear$(h, h/2)$ $\to$ tanh $\to$ dropout $\to$ Linear$(h/2, 1)$ $\to$ sigmoid. The model uses its pooler when available and otherwise masked-mean pooling. All members train for five epochs with batch size 8, gradient accumulation 4, AdamW learning rate 10^{-5}, weight decay .02, warmup .08, gradient clipping 1, fp32, and AMP disabled.

Seeds 13, 17, and 23 use $L = \text{MSE} + .8L_{\text{Pearson}} + .3L_{\text{pair}}$. Seed 29 uses pairwise weight .48, margin .07, and weight 2.2 for gold below .48. Seed 41 uses pairwise weight .60, margin .09, and weight 3 for gold below .45. Each checkpoint is selected on validation data, and the final committee is their unweighted mean.

E.2 Independent IBM Qwen quality branch

The quality branch uses only topic, stance, and argument. A full bfloat16 Qwen3-8B receives a rank-16 LoRA with $\alpha = 32$, dropout .05, and the same seven projection families listed above. A float32 linear head reads the last non-padding hidden state and applies a sigmoid. Training uses two epochs, micro-batch size 2, accumulation 8, learning rate 10^{-4}, weight decay .01, warmup .05, maximum length 768, and clipping 1. Frozen evaluation uses maximum length 512.

The effective objective of all eight frozen replicas is MSE: the soft-Spearman implementation returns zero at micro-batch size two. Seeds 221, 223, 225, 227, 229, 231, 233, and 235 are averaged without weighting. This independent quality branch enters PRISM through scalar fusion.

F Profile, Anchor, and Calibration Inventory

F.1 Exact 85-dimensional profile

Let $o, t \in \mathbb{R}^{15}$ denote respectively the token-logit expectation and the parsed JSON coordinate vectors in the ordering of Table 9. The profile is

$$p = [o; t; o - t; |o - t|; o \odot t; s] \in \mathbb{R}^{85}. \tag{18}$$

The ten summary entries, in order, are mean, population standard deviation, minimum, and maximum of o; the same four statistics of t; the extracted `quality_score`; and uncertainty.

Block	Dim.	Exact ordering
Outer scores	15	$o_1, \ldots, o_{15}$
Parsed scores	15	$t_1, \ldots, t_{15}$
Signed gap	15	$o - t$
Absolute gap	15	$\lvert o - t \rvert$
Products	15	$o \odot t$
Summaries s	10	See text below

Table 14: Exact profile blocks; dimensions sum to 85.

F.2 Exact 12-dimensional anchor vector

Let $q_1 \geq \cdots \geq q_k$ be cosine similarities between the normalized query profile and its predefined $k = 64$ retrieved training profiles, with corresponding labels y_j. Define

$$w_j = \frac{\exp(12 \max(q_j, 0))}{\sum_{\ell=1}^{k} \exp(12 \max(q_\ell, 0))}. \tag{19}$$

The weighted median sorts the labels increasingly and returns the first label whose cumulative weight is at least .5.

No.	Statistic	No.	Statistic
1	Weighted label mean	7	q_1
2	Weighted label median	8	Mean $q_{1:3}$
3	Label mean	9	Mean $q_{1:5}$
4	Label population SD	10	Mean $q_{1:k}$
5	Minimum label	11	Population SD $q_{1:k}$
6	Maximum label	12	$-\sum_j w_j \log w_j$

Table 15: Exact anchor-vector ordering. Training queries self-exclude; validation and test retrieve only from the training pool.

F.3 Profile-versus-anchor decomposition

Encoder interface	P	S	K	MAE
Raw committee	.628639	.581422	.414358	.125188
Raw-score ridge	.628639	.581422	.414358	.122881
Raw + direct 85D profile	.638396	.591464	.422177	.119936
Raw + 12D profile anchors	.637928	.592729	.423205	.119934
Raw + profile + anchors	**.638172**	**.592834**	**.423216**	**.119801**

Table 16: Frozen IBM encoder-interface decomposition. The anchor-only row removes the direct 85 coordinates but still retrieves neighbors in the learned profile space.

The raw-score ridge changes scale and offset, while direct profile coordinates and profile-derived anchors reorder examples and recover the correlation gain. Combining both representations gives the best Spearman and MAE in this decomposition.

F.4 Score-like field removal

Full versus quality/uncertainty removal has $\Delta\rho = .000077$ with 95% CI $[-.000191, .000311]$. Full versus removal of all summary fields has $\Delta\rho = .000028$ with CI $[-.000384, .000400]$. Full versus the 30D-only profile has $\Delta\rho = .000308$ with CI $[-.000944, .001526]$. Removing the quality and uncertainty fields leaves Spearman essentially unchanged, showing that the profile's predictive signal extends beyond those direct score-like fields.

For additional descriptive controls, a frozen semantic-neighbor baseline reaches test Spearman .5800 and MAE .1212, while an equal-dimensional label-bin/fixed-projection pseudo-schema reaches .5803 and .1208. Their stored bootstrap artifacts use row-level rather than topic-cluster resampling; the predefined topic-cluster inference appears in the matched comparisons below.

Calibration profile	Dim.	Spearman	MAE
Full profile + anchors	85+12	.592834	.119801
Drop quality + uncertainty	83+12	.592757	.119799
Drop all 10 summary fields	75+12	.592806	.119759
Only two 15D score sets	30+12	.592526	.119809

Table 17: Frozen-prediction controls removing direct and derived score-like fields. Every variant refits the standardized ridge and recomputes anchors in the retained feature space.

Profile student	Pearson	Spearman	Kendall	MAE	RMSE	ECE	Pairwise
Generic structured	.625513	.582556	.414427	.120541	.154692	.011695	.740242
Theory aligned	**.632599**	**.589649**	**.420541**	**.119741**	**.153607**	**.009926**	**.743654**

Table 18: Strict capacity-matched test comparison on 6,297 common complete rows. Lower is better for MAE, RMSE, and ECE. Both ridge penalties are selected as $\alpha = 10{,}000$ on validation Spearman.

G Capacity-Matched Generic Teacher–Student Control

The generic control uses the same Qwen3.7-Plus teacher, Qwen3-8B student Yang et al. (2025), gold conditioning, JSON/evidence/score/hierarchy filters, 15-field structure, prompt budget, LoRA configuration, loss weights, epochs, and downstream 85D-profile/12D-anchor ridge interface. The mean teacher input length is 3,111 tokens versus 3,108 for the theory teacher.

Metric	Advantage	95% CI	Raw p	Holm p
Pearson	+.007085	$[-.000304, .014248]$	.0610	.1924
Spearman	+.007094	$[-.000386, .014330]$	.0661	.1924
Kendall	+.006114	$[.000036, .012064]$	.0481	.1924
MAE	+.000800	$[-.000259, .001874]$	.1391	.1924

Table 19: Theory advantage under 20,000-replicate paired IBM-topic bootstrap. For MAE, positive means lower theory error.

The taxonomy-aligned profile has favorable point estimates on all reported metrics, including +.007094 Spearman and .000800 lower MAE. The paired intervals and multiplicity-adjusted results in Table 19 quantify the uncertainty of this matched comparison. Excluding same-topic training anchors gives theory/generic Pearson $.632629/.625544$, Spearman $.589603/.582574$, Kendall $.420496/.414455$, and MAE $.119740/.120534$.

The common complete intersections with DeBERTa are 20,897/20,974 training rows (99.63%), 3,192/3,208 validation rows (99.50%), and 6,297/6,315 test rows (99.71%). On test, theory parses completely for 6,314 rows and generic for 6,298. Strict metrics use the same 6,297-row complete intersection, giving both systems identical evaluation support.

H Extended IBM Results and Uncertainty

H.1 Primary topic-cluster intervals

The full system obtains Pearson .656863, Spearman .614587, Kendall .440608, and MAE .116429 on all 6,315 IBM test examples.

H.2 Neighbor-count, ridge, and fusion sensitivity

The validation grid covers $\alpha \in \{100, 1000, 3000, 10^4, 30000, 10^5\}$ in the frozen sensitivity run. Regularization is consequential: validation Spearman peaks at 10^4 and declines on either side. Across the tested anchor counts, test Spearman ranges from .592307 to .592834. Joint validation selection chooses $k = 256$ and obtains test Spearman .592412; the predefined main setting uses $k = 64$.

Comparison ($A - B$)	Metric	Δ	95% topic-cluster CI	Two-sided p
PRISM − Qwen only	Pearson	+.005266	[.001826, .008852]	.0016
	Spearman	+.003804	[.000533, .007600]	.0186
	Kendall	+.003201	[.000656, .006182]	.0098
	MAE	−.000111	[−.000752, .000457]	.7323
PRISM − theory encoder	Pearson	+.018692	[.011455, .025718]	$< .001$
	Spearman	+.021753	[.015007, .028019]	$< .001$
	Kendall	+.017392	[.012062, .022453]	$< .001$
	MAE	−.003372	[−.004296, −.002422]	$< .001$
Theory encoder − raw encoder	Pearson	+.009533	[.003104, .016937]	.0035
	Spearman	+.011412	[.003044, .020282]	.0083
	Kendall	+.008858	[.002336, .015811]	.0071
	MAE	−.005387	[−.007537, −.003384]	$< .001$

Table 20: Twenty-thousand-replicate paired IBM-topic bootstrap. Negative MAE differences favor system A. These cluster intervals replace the earlier row-bootstrap sensitivity calculations.

k	Pearson	Spearman	Kendall	MAE
16	.637631	.592307	.422829	.119881
32	.637846	.592599	.423044	.119836
64	.638172	.592834	.423216	.119801
128	.638105	.592519	.422968	.119816
256	.638077	.592412	.422927	.119816

Table 21: Anchor-count sensitivity at fixed $\alpha = 10{,}000$.

Validation selects Qwen weight $\lambda = .80$ when fusing with the raw encoder and $\lambda = .65$ with the theory encoder. Their individually selected test operating points are:

Fused encoder	Pearson	Spearman	Kendall	MAE
Raw + Qwen	.655169	.613122	.439365	.116563
Theory + Qwen	**.656863**	**.614587**	**.440608**	**.116429**

Table 22: Validation-selected fusion weights with the same frozen Qwen branch.

To isolate the encoder replacement from fusion-weight selection, the factorial comparison fixes $\lambda = .65$ for both systems. Under this matched operating point, raw-score ridge plus Qwen obtains Pearson .656143, Spearman .612964, Kendall .439289, and MAE .116978; replacing only the encoder input with the theory-calibrated prediction gives the full-system values reported above.

Metric	Δ	95% topic-cluster CI	Two-sided p
Pearson	+.000720	[−.000888, .002597]	.4074
Spearman	+.001623	[−.000483, .003774]	.1358
Kendall	+.001319	[−.000381, .003109]	.1288
MAE	−.000548	[−.000908, −.000245]	$< .0001$

Table 23: Fixed-λ factorial: PRISM minus raw-score ridge plus the identical frozen eight-replica Qwen ensemble. Results use 20,000 paired IBM-topic bootstrap replicates; negative MAE favors PRISM.

At the fixed fusion weight, profile calibration lowers MAE by .000548 relative to raw-score ridge with the same frozen Qwen ensemble. The encoder-side ranking gain is clearest before fusion; the paired intervals in Table 23 quantify the post-fusion uncertainty.

Within $\lambda = .65 \pm .10$, theory-fused test Spearman ranges from .614024 to .614587. This supports broad complementarity rather than dependence on a single narrow fusion weight.

Words	n	Raw	Theory	Qwen	PRISM
≤ 12	1,566	.6527	.6635	.6702	**.6765**
13–25	3,591	.5499	.5632	.5880	**.5911**
≥ 26	1,158	.4691	.4818	.4910	**.4976**

Table 24: IBM Spearman by argument length. These strata are descriptive.

Gold interval	Raw	Theory	Qwen	PRISM
$y < .45$	**.2154**	.2893	.2661	.2733
$.45 \leq y < .60$	.1709	.1757	**.1680**	.1687
$.60 \leq y < .75$	.1368	**.1114**	.1156	.1124
$y \geq .75$	.1036	.0927	.0904	**.0902**

Table 25: IBM test MAE by predefined quality interval. Training-trace retention and interval counts are reported in Table 12.

H.3 Length and quality strata

The quality-stratified results reveal where calibration has the most room to improve: on the 489 examples with $y < .45$, signed biases are $+.1930$ (raw), $+.2846$ (theory), $+.2619$ (Qwen), and $+.2699$ (PRISM). In the two higher intervals, theory reduces encoder error. Complementarity is clearest where the branches disagree: in the top disagreement quartile (1,579 examples), Qwen and PRISM Spearman are .5891 and .6037 and MAE falls from .1340 to .1326. Theory improves encoder MAE on 14 of 15 test topics.

H.4 Interpretation of Generated Coordinates

The generated coordinates and evidence spans are evaluated as predictive interfaces. A dedicated multi-annotator study would further assess coordinate agreement, adjudication consistency, and explanatory faithfulness.

H.5 Single-model and ensemble-size results

k	Subsets	Qwen ρ	Fused ρ	ms/example
1	8	$.5998 \pm .0031$	$.6092 \pm .0022$	30.5
2	28	$.6059 \pm .0018$	$.6122 \pm .0013$	60.9
4	70	$.6091 \pm .0009$	$.6138 \pm .0007$	121.9
8	1	.6108	.6146	243.8

Table 26: All prediction subsets for ensemble sizes 1, 2, and 4; size 8 uses all replicas. Fusion keeps the theory encoder and $\lambda = .65$ fixed. Costs are A800 batch-8 sequential estimates including adapter loads.

The ensemble curve exposes several accuracy–cost operating points: fused Spearman rises from .6092 with one replica to .6146 with eight, enabling deployment to select the desired budget.

The 30.5 ms/example entry is $243.8/8$, amortizing sequential loading across the eight-replica benchmark. The 13.8 ms/example component measurement in Table 28 is post-load inference for one resident model. Together they describe ensemble and single-model deployment costs.

I Webis Adaptation and Known-Scale Routing

I.1 Strict Webis training protocol

The strict runner loads the full Qwen3-8B in bfloat16 and attaches a last-token float32 linear regression head without sigmoid, clipping, or label transformation. LoRA uses $r = 16$, $\alpha = 32$, dropout .05, and the seven attention/MLP projection families. Webis-only training starts from the base model. IBM-initialized training continues the seed-229 IBM adapter and head. Both use three

epochs, batch size 8, gradient accumulation 2, learning rate 5×10^{-5}, weight decay .01, warmup .05, maximum length 384, clipping 1, and loss

$$L = \mathrm{MSE} + 0.5 L_{\text{soft-Spearman}}. \tag{20}$$

Checkpoints are selected by development Combined-Quality Spearman. Final predictions are the unweighted mean of seeds 301, 303, and 305.

The frozen aggregate gives Webis-only Spearman .7864 and IBM-initialized Webis Spearman .7910. Their paired discussion-cluster difference is $+.0046$ with 95% CI $[-.0452, .0587]$ and $p = .870$. A separately frozen matched-update comparison gives $+.0057$, CI $[-.0438, .0595]$, and $p = .837$. IBM Target-domain supervision drives the principal transfer gain: relative to IBM zero-shot, Webis adaptation improves Spearman by $+.4628$, CI $[.3293, .6326]$, Holm-adjusted $p < .001$. IBM initialization yields a comparable adapted aggregate (.7910 versus .7864); the paired interval for this comparison is $[-.0452, .0587]$.

I.2 Router, mappings, and target-scale scope

The router receives body text only. Five template cues are removed: `debate-forum`, `continuousscale`, `Ratetheoverallquality`, `verypoor`, and `excellent`. Features are word 1–2-gram TF–IDF with `min_df=2`, `max_features=50000`, and sublinear TF. A balanced liblinear logistic regression with seed 42 is fit on 20,974 IBM training and 889 Webis training arguments. A Platt calibrator is fit on IBM validation plus Webis development.

The routing evaluation assumes a supplied target corpus protocol. For each example, the known protocol determines the requested output scale and mapping direction, while the router selects the expert and fallback policy.

The test confusion matrix, with true corpora as rows and predicted corpora as columns, is

$$\begin{bmatrix} 6301 & 14 \\ 36 & 155 \end{bmatrix}. \tag{21}$$

Accuracy is 99.23%, balanced accuracy 90.46%, IBM recall 99.78%, and Webis recall 81.15%.

Directed monotone affine mappings minimize target-development MAE. The fitted Webis-to-IBM map is $.159079x + .882484$ and the IBM-to-Webis map is $.853398x - .445142$. Target-specific fallback weights put .974910 on the IBM native expert and effectively 1.0 on the Webis native expert. A threshold grid from .50 to 1.00 in .01 increments minimizes mean two-domain, range-normalized development MAE subject to at most 1.25 logical calls per example on the fixed pooled IBM/Webis development mixture. The selected threshold is .99, with a pooled development call rate of 1.1874. The 1.25-call budget applies to the pooled IBM/Webis development mixture: low-confidence cases evaluate both experts and use the fitted convex fallback, whereas high-confidence cases use one scale-mapped selected expert.

Policy	ρ	τ	MAE	Calls
IBM requested scale				
Hard routing	.6046	.4323	.1200	1.00
Scale-mapped hard	.6067	.4338	.1173	1.00
Confidence-triggered	.6066	.4337	.1173	1.09
Soft gating	.6069	.4338	.1173	2.00
Webis requested scale				
Hard routing	.5477	.3894	.6871	1.00
Scale-mapped hard	.7005	.5089	.5992	1.00
Confidence-triggered	**.7794**	**.5839**	**.4899**	1.50
Soft gating	.7345	.5389	.5757	2.00

Table 27: Complete routing metrics on the fixed test sets. Confidence-triggered denotes the selective-fusion policy; one call denotes one frozen aggregate expert.

Paired 20,000-replicate cluster bootstrap resamples IBM topics and Webis discussions. On IBM, confidence-triggered fusion versus hard routing changes Spearman by $+.0020$ (95% CI

$[-.00002, .00424]$; Holm-adjusted $p = .185$), On Webis, it improves over hard routing by $+.2317$ ($[.1098, .3547]$, adjusted $p < .001$), over scale-mapped hard routing by $+.0789$ ($[.0228, .1421]$, adjusted $p = .0054$), and over soft gating by $+.0449$ ($[.0041, .0937]$, adjusted $p = .0297$). The learned Webis fallback weight is nearly a native-expert selection, with confidence-triggered allocation providing the largest gains on the transferred target scale.

J Compute and Runtime

Component, batch 8	Bench. n	ms/example	Peak GiB
Single DeBERTa	512	5.2	8.2
DeBERTa committee (5)	512	21.9	8.2
Single Qwen regression LoRA	512	13.8	15.9
Qwen LoRAs (8), sequential	256	243.8	15.9
Theory score-only	256	168.4	6.6
Theory full trace	16	3,877.3	13.7

Table 28: A800 80GB component microbenchmarks at the indicated sample counts. The eight-replica row includes sequential adapter/model loads.

These measurements characterize resident-model inference, sequential ensemble scoring, and offline structured-trace construction. The DeBERTa and single-Qwen measurements use 512 cached validation examples; the eight-Qwen aggregate uses 256; theory score-only generation uses 256; and full traces use 16 because autoregressive structured output is substantially slower. The 13.8 ms/example single-Qwen value is post-load inference with the model already resident, whereas the ensemble curve's 30.5 ms/example per-replica estimate is $243.8/8$ and includes sequential model/adapter loads. The original theory-student SFT uses 588 optimizer steps and approximately 31 hours on the original machine. Full traces are intended for offline construction or caching, while the single-replica and calibrated encoder configurations expose lower-cost deployment points.

K Reproducibility Record

The retained reproducibility record comprises:

- fixed IBM and Webis split IDs, stable sample IDs, and dataset hashes;
- theory and generic teacher/student prompt builders and all field definitions;
- raw and filtered trace manifests with sequential rejection counts;
- theory/generic student configurations, checkpoints, and seeds;
- frozen per-sample encoder, Qwen, Webis, routing, and scale-mapped predictions;
- scripts for feature construction, validation grids, ensemble subsets, 20,000-replicate cluster bootstrap, duplicate sensitivity, and cost measurement;
- package versions, hardware summaries, and resumable generation manifests.

Model, feature, threshold, and routing choices are fixed from training, validation, or development data. Frozen predictions and prespecified diagnostic tables form the final test evaluation.

L Evaluation Scope and Responsible Use

This study evaluates English argument-quality prediction on IBM-Rank-30k and Webis-ArgQuality-20 under their native annotation protocols. The generated coordinates and evidence spans serve as predictive interfaces; a dedicated multi-annotator study can further characterize their agreement and explanatory faithfulness. The matched generic control identifies the value of taxonomy alignment within the encoder interface, while the quality-stratified results highlight low-quality examples as a direct target for calibration. The routing study demonstrates protocol-aware allocation when the target scale is known, providing a foundation for broader open-set, multilingual, and adversarial

evaluations. Because argument-quality labels encode corpus and annotator preferences, deployment should preserve qualified human oversight and respect the original dataset and model licenses.