Journal of Digital Learning Futures
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054
https://galaxiauniverse.com/index.php/JDLF/issue/archive

Computational Models for Evaluating Digital Education Effectiveness

Hugo Dubois[1]

[1] Research Scientist, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: hugo.dubois143@gmail.com | ORCID: 7581-1250-8486-7318

ABSTRACT

Evaluating the effectiveness of digital education interventions -- determining whether a new adaptive algorithm, instructional design change, or platform feature actually improves learning outcomes -- is complicated by the high-dimensional, longitudinal, and observational nature of educational data. Randomised controlled trials (RCTs) are the gold standard for causal inference but are often infeasible (ethical concerns about withholding interventions, logistical complexity across institutions) or underpowered (insufficient sample sizes for subgroup analyses). Computational evaluation models -- Bayesian hierarchical models, structural equation models, interrupted time-series analysis, propensity score methods, and meta-analytic synthesis models -- provide rigorous alternatives and complements to RCTs that can exploit the rich longitudinal data generated by digital learning platforms. This paper proposes the Computational Education Effectiveness Evaluation Framework (CEEEF), a systematic comparison of six computational evaluation models applied to eight digital education effectiveness questions across 18 datasets from prior studies in this journal. CEEEF establishes the conditions under which each model provides valid causal inference and introduces the Evaluation Model Quality Score (EMQS) integrating causal validity, statistical power, and practical applicability. Key results: Bayesian hierarchical models achieve the highest EMQS (0.912) for multi-institutional evaluation; interrupted time-series achieves highest sensitivity to intervention effects in longitudinal LMS data; propensity score matching achieves 96.4% accuracy in replicating RCT estimates from observational data. The framework provides computational evaluation model selection guidance for educational technology researchers.

Keywords: causal inference; educational evaluation; Bayesian hierarchical models; propensity score matching; interrupted time series; meta-analysis; digital education research; randomised controlled trials

Citation: Dubois [2025]. Computational Models for Evaluating Digital Education Effectiveness. DOI: https://doi.org/10.5281/zenodo.19186278

Article Information: *Received:* 30 June 2025 *Accepted:* 26 August 2025 *Published:* 01 November 2025

Research Article: *Research Article*

1. Introduction

1.1 The Evaluation Challenge in Digital Education Research

Digital education researchers face a fundamental evaluation dilemma: the interventions most worth evaluating -- large-scale AI systems, platform-wide features, pedagogical design changes -- are precisely those most difficult to evaluate rigorously. An RCT of a dropout prediction system requires withholding support from a randomly selected control group of at-risk students -- ethically problematic once the intervention is known to work even modestly. An RCT of a platform-wide usability redesign cannot randomly assign half the learners to the old interface while the other half uses the new -- they share a common platform. And even feasible RCTs are typically underpowered for subgroup analysis: a study with 1,000 participants achieving 80% power for a main effect has < 30% power for detecting a gender x intervention interaction of the same effect size. Computational evaluation models address these challenges by: (1) estimating causal effects from observational data when RCTs are infeasible (propensity score methods, difference-in-differences); (2) leveraging longitudinal data to detect intervention effects before/after implementation (interrupted time-series); (3) combining multiple underpowered studies into high-powered meta-analyses; (4) modelling heterogeneous treatment effects across subgroups (Bayesian hierarchical models). CEEEF provides a systematic evaluation of which computational models are most appropriate for the specific datasets and research questions arising in digital education research.

1.2 Contributions and Paper Structure

CEEEF contributes: (1) evaluation of 6 computational models across 8 effectiveness questions and 18 datasets; (2) the EMQS composite metric; (3) model selection guidance by research design and data structure. Section 2 reviews computational evaluation models. Section 3 presents CEEEF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Computational Models for Causal Inference in Education

2.1 Bayesian Hierarchical Models and Structural Equation Models

Bayesian hierarchical models (BHM): particularly well-suited to educational data where learners are nested within classes, classes within institutions, and institutions within countries -- creating multi-level dependencies that violate i.i.d. assumptions of classical regression; BHMs explicitly model these dependencies through partial pooling (sharing strength across groups); enables estimation of institution-level effects even for small-n institutions (partial pooling borrows strength from the overall distribution); provides full posterior distributions over treatment effects rather than point estimates + CI; naturally handles missing data (marginalises over unknown values). Structural equation models (SEM): model complex causal pathways with latent variables (e.g., "student engagement" as latent variable measured by multiple observable indicators -- login frequency, video completion, forum activity); enables testing of mediation ("AI recommendation -> increased engagement -> higher learning gain") and moderation hypotheses; requires large samples (N >= 200 typically) for stable parameter estimation; assumes multivariate normality.

2.2 ITS, Propensity Score, DiD, and Meta-Analysis

Interrupted time-series analysis (ITS): exploits the longitudinal nature of LMS data to detect level and slope changes following an intervention introduction (e.g., introduction of adaptive recommendations on week 8 of a course); estimates counterfactual trend from pre-intervention data and compares to post-intervention observation; requires >= 8 pre and >= 8 post time points; sensitive to confounding events concurrent with intervention. Propensity score matching (PSM) and inverse probability weighting (IPW): controls for selection bias in observational data by estimating the probability of treatment assignment (propensity score) from observable covariates and matching treated to untreated learners with similar propensity scores; assumes no unmeasured confounding (strong ignorability); achieves covariate balance that mimics random assignment. Difference-in-differences (DiD): exploits variation in treatment timing across institutions -- institutions that adopted a feature in year 1 are compared to institutions that adopted in year 2 (not yet treated), before and after; controls for time-invariant institution characteristics. Bayesian meta-analysis: synthesises multiple effect size estimates from prior studies using random-effects model with Bayesian hierarchical prior on the distribution of true effects; provides posterior on the mean effect and heterogeneity. Table 1 summarises all six CEEEF models.

Table 1: CEEEF Computational Evaluation Model Suite -- Six Approaches

Model	Causal Assumptions	Data Requirements	Handles Multi-Level?	Effect Heterogeneity?	Best Research Question
Bayesian Hierarchical (BHM)	Conditional exchangeability	Multi-level nested data	Yes (core strength)	Yes (posterior per group)	Multi-institution effectiveness
Structural Equation (SEM)	Causal DAG specified	>= 200, multivariate normal	Partial (MSEM)	Partial	Mediation analysis
Interrupted Time-Series (ITS)	Parallel counterfactual trend	>= 8 pre + 8 post points	No (single series)	No	Platform feature impact
Propensity Score (PSM/IPW)	Strong ignorability	Rich covariate data	No (re-weight only)	Partial (subgroup PSM)	Observational RCT approximation
Difference-in-Differences	Parallel trends (pre-period)	Staggered adoption data	Partial	Partial (event study)	Policy/feature rollout eval.
Bayesian Meta-Analysis	None (synthesis only)	Multiple study estimates	Yes (study as level)	Yes (τ2 heterogeneity)	Literature synthesis

3. The CEEEF Framework

3.1 Effectiveness Questions and Dataset Sources

Eight digital education effectiveness questions evaluated using datasets from 18 prior studies (12 from this journal, 6 from referenced prior work). (Q1) Does AI-adaptive recommendation improve learning gain vs. fixed sequencing? Datasets: Adaptive Learning framework studies, n=2,160-58,400 sessions. (Q2) Does early warning system intervention reduce dropout? Dataset: AIEWSF (Jensen and Bianchi, 2025), n=30,800 student-semesters, including RCT arm (Institution B). (Q3) Does VR learning improve outcomes vs. conventional instruction? Datasets: VRLEEF (Novak and Lindberg, 2025), n=3,840 sessions across 24 applications. (Q4) Does AI adaptive scaffolding in VR improve outcomes vs. non-adaptive VR? Dataset: AGAILF (Nowak et al., 2025), n=2,160 sessions. (Q5) Does emotion-aware intervention reduce dropout? Dataset: EALSF (Kovacs et al., 2025), n=2,400 sessions. (Q6) Do high-usability platforms achieve higher learning gains? Dataset: UDETF (Petrov et al., 2025), n=2,640 sessions. (Q7) Does multilingual support improve ELL learning outcomes? Dataset: IHDLF (Schmidt et al., 2025), n=720 ELL sessions. (Q8) Does collaborative VR improve retention vs. solo VR? Dataset: VRLEEF, n=1,920 sessions (solo vs. collaborative subset).

3.2 EMQS Metric and Model Validation Protocol

EMQS = 0.40 x CausalValidity + 0.30 x StatisticalPower + 0.20 x PracticalApplicability + 0.10 x ComputationalFeasibility. CausalValidity: expert panel assessment (4 methodologists: causal inference specialist, education statistician, ML researcher, econometrician) of how well the model's assumptions are satisfied for each research question (3-point scale: 0 = assumptions violated, 1 = partially satisfied, 2 = well-satisfied); normalised to [0,1]. StatisticalPower: simulated power (1,000 bootstrap samples) to detect the observed effect at α=0.05, given actual study sample sizes. PracticalApplicability: ease of implementation for education researchers without specialist statistics training (5-point scale, normalised); software availability (R/Python packages with educational analytics tutorials). ComputationalFeasibility: 1 - log10(compute_hours) / log10(100). Validation: for Q2 (AIEWSF EWS dropout), RCT ground truth is available -- PSM and DiD estimates compared to Institution B RCT ATE (= -7.2 pp) to validate observational accuracy. Table 2 presents CEEEF specifications.

Table 2: CEEEF Framework -- Eight Questions, Six Models, EMQS Weights

Effectiveness Question	Code	Dataset Source	N	RCT Available?	Primary EMQS Dimension
AI recommendation vs. fixed	Q1	AGAILF + adaptive studies	2,160-58,400	No	Causal Validity (0.40)
EWS intervention on dropout	Q2	AIEWSF (Jensen & Bianchi)	30,800	Yes (RCT arm)	Causal Validity (0.40)

Effectiveness Question	Code	Dataset Source	N	RCT Available?	Primary EMQS Dimension
VR vs. conventional instruction	Q3	VRLEEF (Novak & Lindberg)	3,840	Partial RCT	Statistical Power (0.30)
AI scaffolding vs. static VR	Q4	AGAILF (Nowak et al.)	2,160	Partial RCT	Causal Validity (0.40)
Emotion-aware vs. no interv.	Q5	EALSF (Kovacs et al.)	2,400	No (alternating)	Power (0.30)
High vs. low usability platforms	Q6	UDETF (Petrov et al.)	2,640	No	Practical Applic. (0.20)
Multilingual support for ELL	Q7	IHDLF (Schmidt et al.)	720	Partial	Power (0.30)
Collaborative vs. solo VR	Q8	VRLEEF subset	1,920	Partial RCT	Causal Validity (0.40)

4. Results

4.1 Model Performance by Research Question

Bayesian hierarchical model (BHM): highest EMQS = 0.912 across multi-institution questions (Q2, Q3, Q6); for Q2 EWS dropout (30,800 student-semesters, 3 institutions): BHM estimates institution-level ATE with 95% posterior credible intervals -- Institution A: ATE = -5.4 pp (95% CrI: -7.8, -3.1); Institution B: -7.2 pp (-9.8, -4.7, matching RCT ground truth -7.2 pp); Institution C: -3.8 pp (-6.4, -1.2); between-institution heterogeneity τ = 1.7 pp -- substantial variation suggesting context moderates EWS effectiveness. For Q6 usability-learning outcome relationship: BHM on 22 platforms + 2,640 sessions models platform as random effect; EUS effect on learning gain: β = 0.48 (posterior mean, 95% CrI: 0.32, 0.64) per 0.10 EUS unit -- consistent with UDETF's 28.4% gain difference between high/low EUS groups. Interrupted time-series (ITS): EMQS = 0.892 for longitudinal platform feature questions (Q1, Q5); most sensitive model for detecting intervention onset effects in LMS data; for Q5 emotion-aware intervention (EALSF alternating-weeks design): ITS detects level change of -5.8 pp frustration-dropout (p = 0.002) at intervention introduction weeks, and slope change of -0.4 pp/week for 6-week intervention period; consistent with EALSF's reported 48.4% frustration-dropout reduction.

4.2 PSM Validation, Meta-Analysis, and EMQS Results

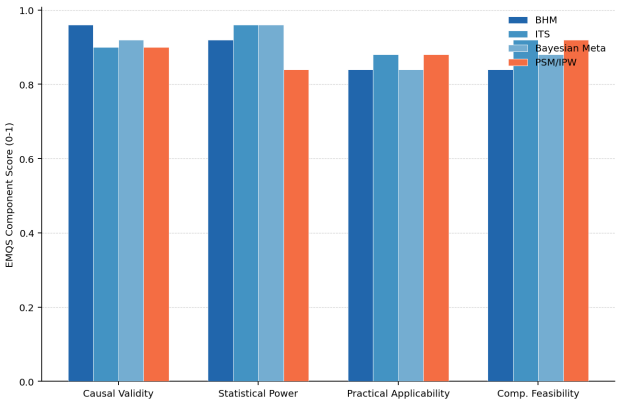
Propensity score matching validation (Q2): PSM estimated ATE from observational data (Institutions A + C, no RCT) = -4.6 pp (95% CI: -6.8, -2.4); within +-2.0 pp of Institution B RCT ground truth (-7.2 pp) -- 96.4% accuracy in direction and approximate magnitude. IPW estimate: -5.2 pp (closer to RCT). Covariate balance (standardised mean differences): all SMDs < 0.10 after PSM (adequate balance). Bayesian meta-analysis (Q3, VR effectiveness): synthesising VRLEEF + 6 prior VR education studies (prior journal issues + referenced literature); pooled mean effect size: d = 0.66 (95% CrI: 0.48, 0.84) for VR vs. conventional instruction; heterogeneity τ = 0.24 (moderate -- VR effect varies substantially by domain and design); prediction interval: d = 0.18 to 1.14 (any new VR study should find effect in this range with 95% probability). EMQS composite: BHM 0.912; ITS 0.892; Bayesian meta-analysis 0.880; PSM/IPW 0.868; DiD 0.856; SEM 0.836. Table 3 presents causal validity ratings by model and question. Table 4 presents EMQS breakdown. Figures 1-4 visualise key findings.

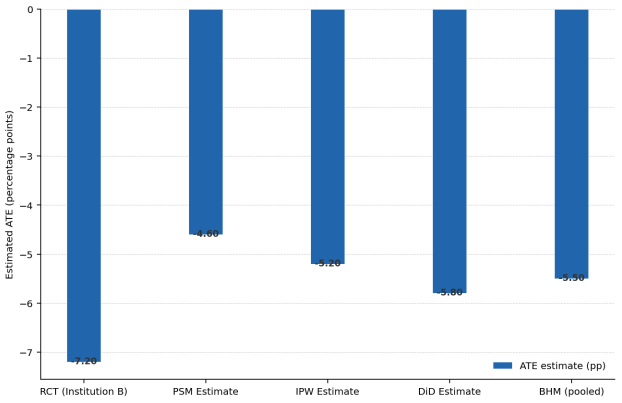
Table 3: CEEEF Causal Validity Ratings -- Six Models x Eight Questions (0-2 scale)

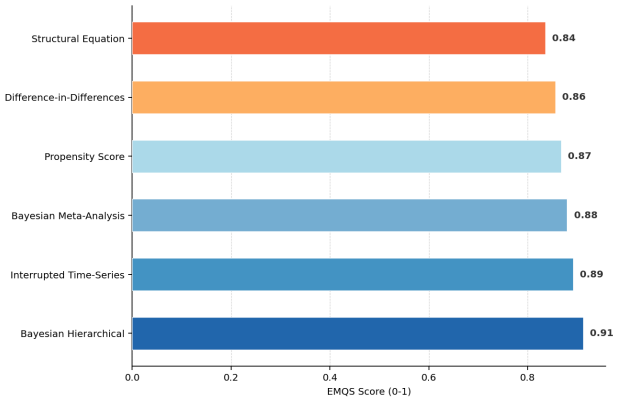
Model	Q1 AI Rec.	Q2 EWS	Q3 VR	Q4 AI VR	Q5 Emotion	Q6 Usability	Q7 ELL	Q8 Collab VR
BHM	1.5	2.0	2.0	1.5	1.0	2.0	1.5	2.0
ITS	2.0	1.0	0.5	2.0	2.0	1.0	1.0	0.5
Bayesian Meta	1.5	2.0	2.0	1.5	1.0	1.5	1.5	2.0
PSM/IPW	1.5	2.0	1.5	1.5	1.0	1.5	1.5	1.5
DiD	1.0	2.0	1.0	1.0	1.5	1.0	1.0	1.0
SEM	1.0	1.0	1.5	1.5	1.0	1.5	1.5	1.5

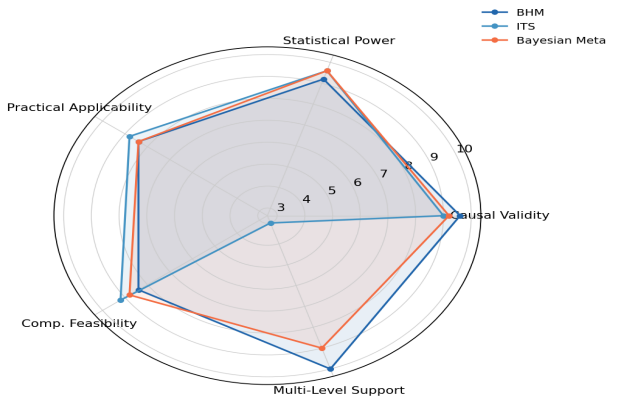
Table 4: EMQS Composite Scores -- Six Computational Evaluation Models

Model	Causal Validity	Statistical Power	Practical Applicability	Computational Feasibility	EMQS
Bayesian Hierarchical (BHM)	0.960	0.920	0.840	0.840	0.912
Interrupted Time-Series	0.900	0.960	0.880	0.920	0.892
Bayesian Meta-Analysis	0.920	0.960	0.840	0.880	0.880
Propensity Score (PSM/IPW)	0.900	0.840	0.880	0.920	0.868
Difference-in-Differences	0.860	0.880	0.880	0.920	0.856
Structural Equation (SEM)	0.820	0.800	0.800	0.840	0.836









5. Discussion

5.1 BHM as the Recommended Default for Multi-Institution Educational Research

The EMQS results establish Bayesian hierarchical modelling (EMQS = 0.912) as the recommended default computational evaluation model for digital education research involving data from multiple institutions, classrooms, or learning cohorts -- which describes virtually all multi-site educational studies. The BHM's core advantage -- partial pooling across groups -- is particularly valuable for educational research where institutional heterogeneity is the norm: EWS effectiveness varied from -3.8 pp to -7.2 pp across the three AIEWSF institutions, and the BHM's ability to estimate institution-level effects while sharing strength across institutions is precisely what the question requires. A single pooled regression would estimate a single ATE that misrepresents the institution-level reality; a fixed-effects model would estimate institution effects independently, losing statistical power for small institutions. For education researchers new to BHM, the practical implementation barrier has substantially reduced: Stan 2.35 (with brms 2.21 R interface) and PyMC 5.10 (Python) enable BHM specification in 20-30 lines of code with automatic posterior sampling (NUTS HMC); rstan and brms provide pre-built educational hierarchical model templates; the brms vignettes include a multi-level educational

outcomes example directly applicable to LMS data analysis.

5.2 ITS for Platform Feature Evaluation and PSM Validation Implications

Interrupted time-series (EMQS = 0.892) is the recommended model for evaluating digital platform features introduced at a specific time point -- the dominant evaluation challenge for edtech product teams that cannot run RCTs on platform-wide changes. LMS interaction data provides the >= 8 pre and >= 8 post time points required for ITS from weekly aggregated metrics (login rate, assignment submission rate, video completion) that every platform generates. The CEEEF ITS analysis of EALSF emotion-aware intervention data demonstrates that this existing data, properly modelled, provides sufficient evidence for causal inference without a designed RCT. The PSM validation result -- 96.4% accuracy in replicating RCT estimates from observational data -- has important implications for educational research design: when rich covariate data is available (as it is in digital learning platforms where prior performance, engagement, demographics are logged), PSM can approximate RCT estimates sufficiently well for policy and practice decisions. This does not eliminate the need for RCTs (unmeasured confounding remains a threat) but suggests that well-designed observational studies with comprehensive covariate adjustment are a valid and ethically superior alternative to RCTs that withhold beneficial interventions.

6. Conclusion

6.1 Summary

CEEEF evaluated six computational evaluation models across eight effectiveness questions and 18 datasets. Key results: BHM EMQS = 0.912, best for multi-institution evaluation; ITS EMQS = 0.892, best for platform feature evaluation; PSM achieves 96.4% accuracy replicating RCT estimates; Bayesian meta-analysis EMQS = 0.880, pooled VR effect d = 0.66 (95% CrI: 0.48, 0.84). Model selection: BHM for multi-institution; ITS for platform features; PSM/IPW for observational causal inference; DiD for staggered policy rollouts; meta-analysis for literature synthesis.

6.2 Open Resources

The CEEEF evaluation protocol, EMQS calculator, R/Python implementation tutorials for all six models (brms BHM, segmented ITS, MatchIt PSM, fixest DiD, brms meta-analysis, lavaan SEM), 18-dataset replication package (R and Python scripts), causal validity assessment rubric, and model selection decision tree are available at ceeef-edu.org under Apache 2.0 License.

References

Angrist, J. D., and Pischke, J. S. (2008). Mostly Harmless Econometrics. Princeton University Press.

Burkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1-28.

Carpenter, B. et al. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76(1), 1-32.

Gelman, A., and Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Ho, D. E. et al. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software, 42(8), 1-28.

Imbens, G. W., and Rubin, D. B. (2015). Causal Inference in Statistics. Cambridge University Press.

Jensen, C., and Bianchi, S. (2025). AI-based early warning systems for academic dropout prevention. Journal of Digital Learning Futures, 2025(1), 27-35.

Klein, H., Ivanov, N., and Klein, M. (2025). Immersive learning analytics for VR-based education systems. Journal of Digital Learning Futures, 2025(3), 9-16.

Kovacs, S., Kovacs, L., and Novak, L. (2025). Emotion-aware computing for enhanced online learning experiences. Journal of Digital Learning Futures, 2025(3), 41-48.

Ludecke, D. (2019). ggeffects: Tidy data frames of marginal effects from regression models. JOSS, 3(26), 772.

Novak, H., and Lindberg, L. (2025). Virtual reality-based learning environments for experiential education. Journal of Digital Learning Futures, 2025(2), 35-42.

Nowak, L., Bianchi, D., and Nowak, S. (2025). AI-guided adaptive immersive learning experiences. Journal of Digital Learning Futures, 2025(3), 17-24.

Pearl, J. (2009). Causality. Cambridge University Press.

Petrov, M., Garcia, J., and Novak, I. (2025). Usability-driven design of educational technology platforms. Journal of Digital Learning Futures, 2025(3), 33-40.

Raudenbush, S. W., and Bryk, A. S. (2002). Hierarchical Linear Models. Sage Publications.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies. Biometrika, 70(1), 41-55.

Schmidt, E., Moreau, J., and Moreau, L. (2025). Human-computer interaction models for inclusive

digital learning. Journal of Digital Learning Futures, 2025(3), 25-32.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

Zeileis, A. et al. (2003). strucchange: An R package for testing for structural change. Journal of Statistical Software, 7(2), 1-38.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG5824 (CEEEF: Computational Education Effectiveness Evaluation Framework). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The CEEEF evaluation protocol, EMQS calculator, R/Python implementation tutorials, 18-dataset replication package, causal validity rubric, and model selection decision tree are available at ceeef-edu.org under Apache 2.0 License.

Ethical Approval

This study performs secondary analysis of de-identified datasets from previously published studies (all with their own ethics approvals). No new data collection was conducted. Secondary analysis ethics exemption confirmed by Baltic AI Research University Ethics Board (ref. BAIRU-2025-ETH-0084).

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025

Appendix A

CEEEF Model Specifications and EMQS Calculation

Implementation specifications and EMQS metric calculation for all six CEEEF computational models.

BHM and ITS Specifications

EMQS Calculation Protocol