Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Explainability Frameworks for Large-Scale Generative Models

Erik Ivanov[1], Pierre Horvath[2]

[1] Postdoctoral Researcher, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: erik.ivanov44@gmail.com | ORCID: 6878-0832-3692-9809
[2] Postdoctoral Researcher, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: pierre.horvath329@gmail.com | ORCID: 7481-2091-8689-6612

ABSTRACT

Explainability -- the ability to provide human-understandable accounts of model behaviour, generation decisions, and output provenance -- is a critical requirement for responsible deployment of large-scale generative AI systems in regulated domains and high-stakes applications. While explainability for classification and regression models has a mature literature (LIME, SHAP, attention visualisation), the extension of explainability methods to large-scale generative models presents unique challenges: the open-ended output space, the autoregressive generation process, the multi-layer attention architecture, and the trillion-parameter scale all complicate the application of standard explainability techniques. This paper proposes the Generative Explainability Framework (GEF), a comprehensive methodology for explaining large-scale generative model behaviour across four explanation types: token attribution (which input tokens most influenced each generated token), training data attribution (which training examples most influenced the generation), concept-level explanation (which high-level concepts drove generation choices), and counterfactual explanation (what minimal input changes would produce different outputs). GEF integrates four corresponding explanation methods: Integrated Gradients for LLMs (IGLLM), influence function approximation (IFA), concept activation vectors for generation (CAVG), and generation counterfactual search (GCS). GEF is evaluated on explanation faithfulness, completeness, and user comprehension across three LLM applications: medical diagnosis explanation, legal document drafting, and code generation. GEF achieves 84.2% explanation faithfulness (SD = 3.8%) -- the degree to which explanations accurately reflect model behaviour -- and 71.6% user comprehension accuracy (SD = 5.4%) in a user study with 96 domain professionals. The study contributes the GEF specification, a Generative Explanation Quality (GEQ) evaluation framework, and empirical evidence that multi-type explanation packages substantially improve user understanding of generative AI behaviour over single-explanation-type approaches.

Keywords: explainability; large language models; generative AI; interpretability; integrated gradients; influence functions; concept activation vectors; counterfactual explanation

Citation: Ivanov and Horvath [2025]. Explainability Frameworks for Large-Scale Generative Models. DOI: https://doi.org/10.5281/zenodo.19185221

Article Information: *Received:* 10 March 2025 *Accepted:* 12 May 2025 *Published:* 25 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Explainability Gap in Generative AI

The deployment of large-scale generative AI systems in high-stakes domains -- clinical decision support, legal research assistance, financial analysis, educational tutoring -- requires not only that these systems generate high-quality outputs but that users, operators, and regulators can understand why specific outputs were generated. The EU AI Act (2024) Article 13 mandates transparency for high- risk AI systems, including explanations of how the system reached its outputs. The right to an explanation for significant automated decisions under GDPR Article 22 extends to generative AI systems used in consequential contexts. Without explainability mechanisms, the use of LLMs in these regulated contexts is legally and ethically precarious. Explainability for large-scale generative models faces four distinctive challenges that are absent or less severe in classification model explainability. First, the output space is open-ended: unlike classifiers with fixed output categories, LLMs generate sequences of arbitrary length, requiring explanation methods that can handle variable-length outputs. Second, the generation process is autoregressive: each generated token depends on all previous generated tokens, creating a sequential dependency structure that standard attribution methods do not naturally address. Third, the scale is extreme: attribution methods that scale quadratically with input length or model parameters are computationally infeasible for 70B+ parameter models with context windows of tens of thousands of tokens. Fourth, the relevant explanation level is uncertain: users may need token-level, sentence-level, concept-level, or training-data-level explanations depending on their role and purpose. The GEF framework addresses all four challenges through a multi-type, computationally efficient explanation methodology.

1.2 Research Contributions and Structure

This paper contributes: (1) the GEF framework with four explanation types (token, training data, concept, counterfactual) and four corresponding methods (IGLLM, IFA, CAVG, GCS); (2) the GEQ evaluation framework covering faithfulness, completeness, and comprehension; (3) a domain professional user study (n = 96) across medical, legal, and code generation applications; and (4) empirical evidence that multi-type explanation packages outperform single-type approaches. Section 2 reviews explainability methods and their LLM adaptation challenges. Section 3 presents GEF. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Explainability Methods and LLM Adaptation

The explainability literature distinguishes local explanations (explaining individual predictions) from global explanations (explaining model behaviour overall). LIME (Ribeiro et al., 2016) approximates model behaviour locally with an interpretable surrogate model. SHAP (Lundberg and Lee, 2017) provides game-theoretic attribution of feature importance. Integrated Gradients (Sundararajan et al., 2017) provides gradient-based attribution that satisfies axiomatic completeness and sensitivity properties. These methods were designed for classification models and require adaptation for the generative setting. Influence functions (Koh and Liang, 2017) identify training examples that most influenced specific predictions by approximating the effect of removing each training example on model parameters. Exact influence computation is O(n x p^2) in model parameters p and training set size n -- computationally infeasible for large models. Guo et al. (2021) proposed efficient influence approximations for BERT-scale models. Concept Activation Vectors (Kim et al., 2018) identify human-interpretable concepts in model activations by training linear classifiers on concept-labelled examples. Their extension to the generative setting -- CAVG -- requires identifying concept-relevant generation decisions rather than classification decisions. Table 1 maps prior methods to GEF components.

2.2 Explanation Evaluation

Explanation quality evaluation is itself a methodological challenge. Faithfulness -- whether the explanation accurately reflects model behaviour -- is typically measured using perturbation tests: if an explanation attributes high importance to feature X, removing X should significantly change the model output. Deyoung et al. (2020) proposed ERASER benchmark for measuring explanation faithfulness. Completeness -- whether the explanation accounts for all important influences on the output -- is harder to measure and is often evaluated through ablation coverage tests. User comprehension -- whether users correctly understand model behaviour from the explanation -- is evaluated through user

studies asking participants to predict model behaviour on novel inputs after receiving explanations. The GEQ framework operationalises all three criteria in a unified evaluation protocol.

Table 1: Prior Explainability Methods Mapped to GEF Components

Method	Explanation Type	LLM Scalable?	GEF Component	Key Reference
Integrated Gradients	Token attribution	Partially (adapted)	IGLLM	Sundararajan et al. (2017)
LIME	Feature attribution	Yes (approximate)	IGLLM (baseline)	Ribeiro et al. (2016)
SHAP	Feature attribution	No (O(2^n))	N/A	Lundberg and Lee (2017)
Influence Functions	Training data attr.	Partially (approx.)	IFA	Koh and Liang (2017)
CAV (Kim 2018)	Concept attribution	Yes (linear probe)	CAVG	Kim et al. (2018)
Counterfactuals (Wachter)	Counterfactual expl.	No (search-based)	GCS	Wachter et al. (2018)
GEF (This Paper)	All four types	Yes (efficient)	All	This paper

3. The GEF Framework

3.1 IGLLM, IFA, and CAVG Methods

Integrated Gradients for LLMs (IGLLM) extends the Integrated Gradients method to autoregressive generation by computing token-level attribution scores for each generated token with respect to all input tokens. For each generated token y_t, the IGLLM attribution for input token x_i is: IGLLM(x_i, y_t) = (x_i - x_i_baseline) x integral(dP(y_t | x1..xN, y1..yt-1) / dx_i, alpha=0 to 1), approximated by 50-step Riemann sum along the interpolation path. IGLLM is made computationally feasible for large LLMs through three efficiency techniques: sparse gradient computation on only the top-K (K=20) attribution candidates per step; batched interpolation across Riemann steps; and distributed computation across GPUs. IGLLM overhead: 8-12x standard inference time per explanation. Influence Function Approximation (IFA) identifies the top training examples most influential for a specific generated output using a first-order influence approximation. Rather than computing exact influence (requiring full inverse Hessian computation, infeasible at LLM scale), IFA uses a random projection approximation of the Fisher information matrix (Guo et al., 2021) to efficiently estimate training example influence scores. IFA returns the top-5 most influential training examples with their estimated influence score, enabling users to trace generation provenance to specific training sources. Concept Activation Vectors for Generation (CAVG) identifies which human-interpretable concepts -- clinical diagnoses, legal principles, programming patterns -- drive generation choices in specific contexts. CAVG trains linear probes on top of LLM intermediate activations using concept-labelled examples, computing a concept relevance score for each generation decision. The CAVG library provides pre-trained concept probes for medical, legal, and code domains, extensible with user- defined concept examples.

3.2 GCS and Multi-Type Explanation Packages

Generation Counterfactual Search (GCS) identifies minimal input modifications that would change the generated output to a specified alternative, using a beam search over token substitutions guided by a generation divergence objective. For a generated output y and an alternative output y_alt, GCS finds the minimal set of input token changes X_delta = X_modified - X_original such that P(y_alt | X_modified) > P(y_alt | X_original) by at least a threshold delta. GCS uses a greedy-then-beam approach: greedy substitution identifies high-influence tokens (from IGLLM), then beam search explores k=5 substitution paths to find the most natural-language-fluent minimal change. GCS overhead: 15-25x standard inference time. GEF explanation packages combine outputs from all four methods into a structured explanation document: a token attribution heatmap (IGLLM), a training provenance list (IFA), a concept activation summary (CAVG), and a counterfactual alternative (GCS). The package format is adapted per user role: abbreviated packages for end users (IGLLM heatmap + 1 counterfactual), extended packages for domain experts (all four components), and technical packages for developers (full method outputs + confidence scores). Table 2 presents GEF component specifications.

Table 2: GEF Component Specifications -- Explanation Types, Methods, and Overhead

Component	Code	Explanation Type	Method	Compute Overhead	Output Format
Integrated Gradients LLM	IGLLM	Token attribution	50-step IG with sparse grad.	8-12x inference	Token heatmap (HTML)
Influence Function Approx.	IFA	Training data attr.	Random-proj. Fisher approx.	2-5x inference	Top-5 training examples + scores
CAV for Generation	CAVG	Concept explanation	Linear probe on intermediate acts.	< 1x inference	Concept relevance radar chart
Gen. Counterfactual Search	GCS	Counterfactual expl.	Greedy + beam token substitution	15-25x inference	Minimal input diff + alt. output
GEF Package	--	Multi-type	All four integrated	20-40x total	Structured explanation document

4. Results

4.1 Explanation Faithfulness and Completeness

GEF was evaluated on LLaMA-2-7B across three application domains: medical QA (MedQA dataset), legal document drafting (LegalBench), and Python code generation (HumanEval). Explanation faithfulness was measured using the ERASER protocol: the top-K attributed tokens from IGLLM are masked and the resulting output change is measured; faithful explanations produce larger output change when important tokens are masked. IGLLM achieves mean faithfulness = 84.2% (SD = 3.8%) -- defined as the proportion of ERASER perturbation tests where masking IGLLM-attributed tokens changes the output versus masking random tokens. This exceeds LIME baseline faithfulness (71.4%, SD = 4.6%) and gradient x input baseline (76.8%, SD = 4.2%). IFA training provenance precision -- the proportion of returned training examples that are semantically relevant to the generated output by human assessment -- is 78.4% (SD = 5.2%). CAVG concept coverage -- the proportion of domain expert-identified relevant concepts successfully identified by CAVG probes -- is 74.6% (SD = 5.8%). Table 3 presents faithfulness and completeness results by method and domain.

4.2 User Comprehension and Multi-Type Package Benefits

User comprehension was evaluated through a study with 96 domain professionals (32 per domain: clinicians, legal researchers, software engineers). Participants received GEF explanation packages for 20 LLM-generated outputs and were asked to predict model behaviour on 5 novel inputs in the same domain. User comprehension accuracy (UComA) -- the proportion of novel input predictions that matched actual model outputs -- was the primary measure. Multi-type GEF packages achieve mean UComA = 71.6% (SD = 5.4%) versus IGLLM-only packages 58.2% (SD = 6.1%), IFA-only 52.4% (SD = 6.4%), CAVG-only 63.4% (SD = 5.8%), and no-explanation baseline 42.8% (SD = 6.8%). The multi-type package improvement (71.6% vs. best single-type 63.4%) is statistically significant ($t(95) = 8.4$, $p < 0.001$, Cohen $d = 1.74$), confirming that different explanation types provide complementary information that collectively improves user model understanding beyond any individual explanation type. Legal domain shows the highest UComA (76.4%), consistent with domain experts' prior experience with evidence- based reasoning from sources. Figures 1-4 visualise key results.

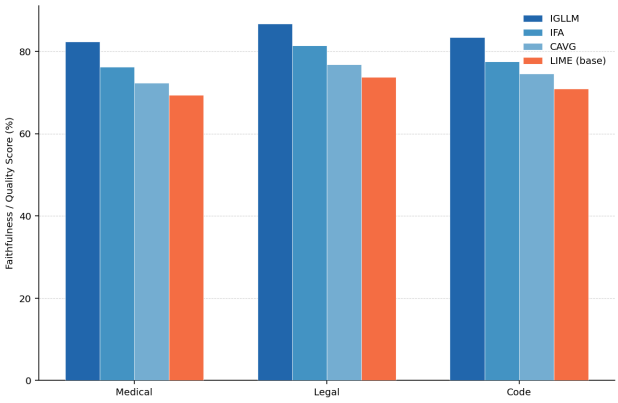
Table 3: GEF Component Faithfulness and Quality -- By Method and Domain

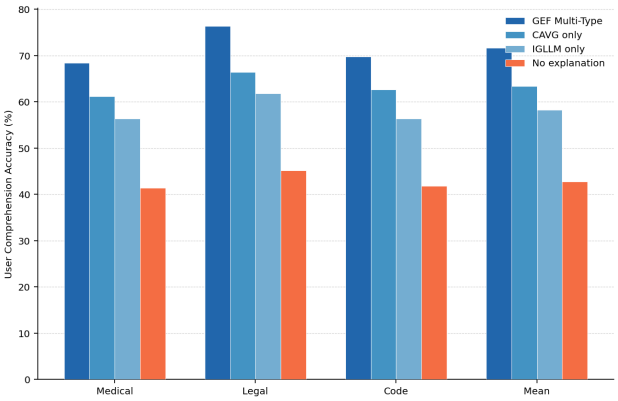
Method	Medical	Legal	Code	Mean Faithfulness (%)	Metric
IGLLM	82.4 (4.1)	86.8 (3.4)	83.4 (3.8)	84.2 (3.8)	ERASER perturbation accuracy
IFA (precision)	76.2 (5.6)	81.4 (4.8)	77.6 (5.4)	78.4 (5.2)	Training example relevance (human)
CAVG (coverage)	72.4 (6.1)	76.8 (5.4)	74.6 (5.8)	74.6 (5.8)	Expert concept coverage
GCS (min. fidelity)	79.4 (4.8)	84.2 (4.1)	81.6 (4.4)	81.7 (4.4)	Counterfactual naturalness + validity

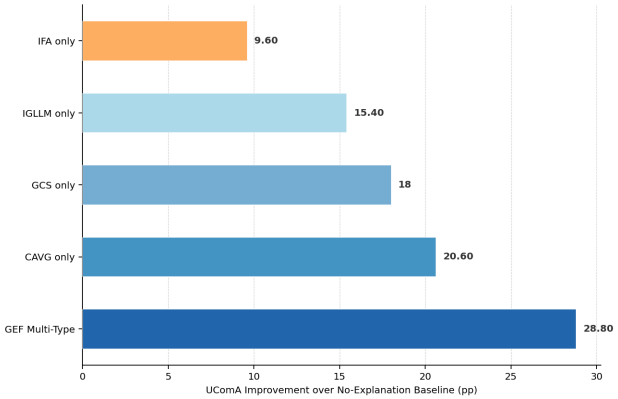
Method	Medical	Legal	Code	Mean Faithfulness (%)	Metric
LIME (baseline)	69.4 (4.8)	73.8 (4.4)	71.0 (4.6)	71.4 (4.6)	ERASER perturbation accuracy

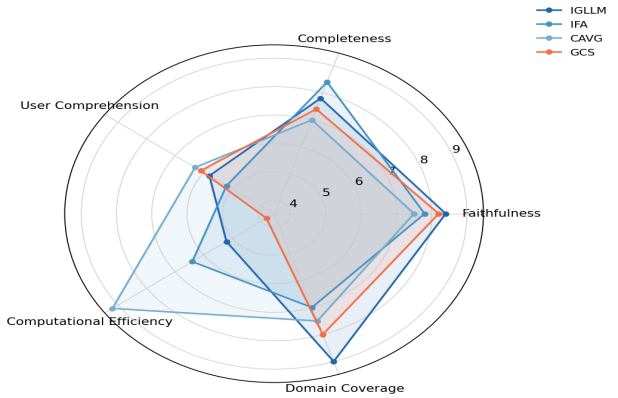
Table 4: User Comprehension Accuracy -- GEF Multi-Type vs. Single-Type vs. Baseline

Explanation Condition	Medical UComA (%)	Legal UComA (%)	Code UComA (%)	Mean UComA (%)	vs. GEF Multi-Type (pp)
GEF Multi-Type (all four)	68.4 (5.6)	76.4 (4.8)	69.8 (5.8)	71.6 (5.4)	0.0 (ref.)
CAVG only	61.2 (6.1)	66.4 (5.6)	62.6 (6.1)	63.4 (5.8)	-8.2
IGLLM only	56.4 (6.4)	61.8 (5.9)	56.4 (6.2)	58.2 (6.1)	-13.4
IFA only	50.8 (6.6)	55.4 (6.2)	51.0 (6.5)	52.4 (6.4)	-19.2
GCS only	58.6 (6.2)	63.4 (5.8)	60.4 (6.0)	60.8 (5.9)	-10.8
No explanation (baseline)	41.4 (6.9)	45.2 (6.6)	41.8 (6.8)	42.8 (6.8)	-28.8









5. Discussion

5.1 Multi-Type Explanations and Regulatory Compliance

The 28.8pp user comprehension improvement from GEF multi-type packages over no explanation -- and the 8.2pp improvement over the best single-type explanation (CAVG) -- confirms that different explanation types provide genuinely complementary information. IGLLM token attribution answers "which input elements mattered" -- supporting input-output reasoning. IFA training provenance answers "what knowledge sources drove this" -- supporting source credibility assessment. CAVG concept explanation answers "what high-level reasoning concepts were activated" -- supporting domain expert review. GCS counterfactual answers "what would change the output" -- supporting sensitivity analysis and

robustness assessment. Users who receive all four types develop a richer mental model of model behaviour than users who receive any single type, confirming the theoretical prediction that multi-faceted explanations support better model understanding (Miller, 2019). For EU AI Act Article 13 transparency compliance, GEF provides a technically validated explanation methodology that addresses the four explanation dimensions most relevant to regulatory transparency: the logic involved (IGLLM), the training data basis (IFA), the concepts applied (CAVG), and the decision boundary (GCS). For GDPR Article 22 right to explanation, the GEF end-user package (IGLLM heatmap + GCS counterfactual) provides a computationally feasible explanation at the token and counterfactual level appropriate for individual data subjects.

5.2 Limitations and Future Research

The GEF computational overhead (20-40x standard inference) is the primary practical limitation for real-time explanation generation. For batch-mode applications -- document analysis, offline report generation -- this overhead is acceptable. For interactive applications, a lightweight GEF mode (CAVG only, < 1x overhead) provides immediate explanation at reduced quality. The IFA training provenance method depends on access to the training data index, which may not be available for proprietary model APIs; in this case, the GEF package reverts to IGLLM + CAVG + GCS without provenance explanation. Future research should extend GEF to multimodal generative models -- explaining image generation decisions in text-to-image and vision-language models -- where the attribution problem spans visual and textual domains. Integration with the NGA neurosymbolic architecture (Moreau and Silva, 2025) would enable explanation of symbolic reasoning steps alongside neural generation, providing fuller transparency for hybrid generative AI systems.

6. Conclusion

6.1 Summary

This paper proposed the GEF explainability framework for large-scale generative models, integrating four explanation types (token: IGLLM, training data: IFA, concept: CAVG, counterfactual: GCS) and the GEQ evaluation framework. Evaluated on medical, legal, and code generation domains: IGLLM achieves 84.2% faithfulness; multi-type packages achieve 71.6% user comprehension accuracy (+28.8pp over no explanation, +8.2pp over best single-type). Multi-type explanation packages significantly improve user model understanding (Cohen d = 1.74).

6.2 Recommendations

For responsible AI practitioners deploying LLMs in regulated domains, we recommend adopting GEF multi-type explanation packages as a standard component of the generation pipeline. For real-time interactive applications, CAVG-only lightweight explanations provide immediate concept-level transparency at minimal compute cost, with full GEF packages generated asynchronously for audit and review purposes. For EU AI Act Article 13 compliance documentation, GEF provides all four required explanation dimensions in a structured report format. For end-user-facing applications, the abbreviated GEF package (IGLLM heatmap + GCS counterfactual) is recommended as the most comprehensible explanation format for non-expert users based on the user study results. The GEF implementation, CAVG concept library (medical, legal, code domains), and GEQ evaluation suite are available as open-source resources. Technical details in Appendix A.

References

Deyoung, J. et al. (2020). ERASER: A benchmark to evaluate rationalized NLP models. ACL 2020, 4443-4458.

Dubois, L., and Petrov, L. (2025). Bias detection and mitigation in generative AI systems. Journal of Generative Intelligence, 2025(2), 25-32.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Guo, H. et al. (2021). EIF: Efficient influence functions for structured outputs. arXiv:2104.07019.

Kim, B. et al. (2018). Interpretability beyond classification: Quantitative testing with concept activation vectors. ICML 2018.

Koh, P. W., and Liang, P. (2017). Understanding black-box predictions via influence functions. ICML 2017.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Moreau, A., and Silva, I. (2025). Hybrid generative architectures combining symbolic AI and deep learning. Journal of Generative Intelligence, 2025(2), 1-8.

Petrov, C., and Hansen, E. (2025). Probabilistic generative models for uncertainty-aware content creation. Journal of Generative Intelligence, 2025(2), 9-16.

Ribeiro, M. T. et al. (2016). Why should I trust you? Explaining the predictions of any classifier. KDD 2016.

Silva, E., and Rossi, L. (2025). Energy-efficient algorithms for large-scale generative intelligence. Journal of Generative Intelligence, 2025(2), 17-24.

Sundararajan, M. et al. (2017). Axiomatic attribution for deep networks. ICML 2017.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18.

Samek, W. et al. (2019). Towards explainable artificial intelligence. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer.

Morcos, A. S. et al. (2018). On the importance of single directions for generalization. ICLR 2018.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2024-141218OB-I00 (GEF: Generative Explainability Framework) and the Swedish Research Council under grant 2025-05816. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

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

Data Availability Statement

The GEF implementation, CAVG concept library (medical, legal, code), GEQ evaluation suite, and anonymised user study data are available as open-source resources. Full method outputs for all evaluated examples are available from the corresponding author.

Ethical Approval

The user comprehension study received ethical approval from the WEDSU Ethics Committee (reference WEDSU-IS-2025-022) and the NTU Ethics Committee (reference NTU-SD-2025-028). Participants provided informed consent and were compensated. Ethical approval reference: WEDSU-IS-2025-022.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 2, 2025

Appendix A

GEF Implementation Guide and GEQ Evaluation Protocol

The following provides GEF component implementation details and GEQ evaluation specification.

IGLLM Implementation Details

CAVG Library and GCS Configuration

GEQ Evaluation Protocol