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

Probabilistic Generative Models for Uncertainty-Aware Content Creation

Clara Petrov[1], Elena Hansen[2]

[1] Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: clara.petrov41@yahoo.com | ORCID: 7545-0153-9823-1388
[2] Research Scientist, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: elena.hansen326@yahoo.com | ORCID: 0066-9685-7599-4606

ABSTRACT

Generative AI systems are increasingly deployed in high-stakes applications where the uncertainty of generated outputs is as important as their quality: a medical imaging system that generates plausible-looking but uncertain diagnoses, or a legal document assistant that produces confident-sounding but factually unreliable text, can cause harm precisely because the system fails to communicate its epistemic limitations to users. Uncertainty-aware content generation -- the principled quantification and communication of model uncertainty alongside generated content -- has received limited systematic attention despite its practical importance. This paper proposes the Uncertainty-Aware Generative (UAG) framework, a methodology for integrating epistemic and aleatoric uncertainty quantification into large generative models across text, image, and code generation domains. UAG comprises three integrated components: a Bayesian uncertainty estimator (BUE) that provides calibrated uncertainty estimates for LLM token predictions through Monte Carlo dropout ensembles; a semantic uncertainty propagator (SUP) that aggregates token-level uncertainties into claim-level and document-level uncertainty representations; and an uncertainty-adaptive generation controller (UAGC) that modulates generation behaviour -- triggering abstention, caveat insertion, or alternative generation -- based on uncertainty thresholds. UAG is evaluated on calibration quality, uncertainty communication effectiveness, and generation quality across four domains. UAG achieves a 34.8% improvement in uncertainty calibration (Expected Calibration Error reduction from 0.148 to 0.097) and a 28.6% improvement in user trust calibration accuracy (users correctly identifying high-uncertainty outputs) relative to standard LLM baselines. The study contributes the UAG specification, a Generative Uncertainty Quality (GUQ) evaluation framework, and empirical evidence for the practical value of explicit uncertainty communication in high-stakes generative AI deployment.

Keywords: uncertainty quantification; generative AI; Bayesian deep learning; calibration; epistemic uncertainty; aleatoric uncertainty; uncertainty-aware generation; responsible AI

Citation: Petrov and Hansen [2025]. Probabilistic Generative Models for Uncertainty-Aware Content Creation. DOI: https://doi.org/10.5281/zenodo.19185203

Article Information: *Received:* 20 February 2025 *Accepted:* 22 April 2025 *Published:* 25 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Uncertainty Problem in Generative AI

The confident voice of large language models is one of their most practically dangerous properties. When an LLM generates a medical recommendation, legal interpretation, or financial forecast with the same fluent, confident-sounding prose regardless of whether its uncertainty is high or low, it deprives the user of the information needed to appropriately weight the output in their decision-making. This overconfidence problem is well-documented empirically: Kadavath et al. (2022) showed that LLM confidence in generated answers is systematically miscalibrated, with models expressing high confidence on questions they answer incorrectly; Xiong et al. (2024) demonstrated that standard LLMs are poorly calibrated as uncertainty estimators across diverse knowledge-intensive tasks. Uncertainty in generative models arises from two distinct sources that require different treatment. Epistemic uncertainty -- uncertainty due to limited knowledge or training data coverage -- is in principle reducible with more training data and is the type of uncertainty that users should be most cautious about: it indicates that the model is operating outside its knowledge boundary. Aleatoric uncertainty -- irreducible uncertainty inherent in the generation task (e.g., open-ended creative generation where multiple equally valid outputs exist) -- should be communicated differently: not as a warning but as an indication of legitimate diversity. The practical challenge of uncertainty-aware generative AI is integrating meaningful uncertainty quantification into the generation process without disrupting the fluency and coherence that make generative models practically useful. The UAG framework addresses this challenge through three integrated mechanisms that quantify, propagate, and communicate uncertainty at different granularities of the generation process.

1.2 Research Contributions and Structure

This paper proposes the UAG framework with three components (BUE, SUP, UAGC) and evaluates it on calibration quality and uncertainty communication effectiveness. Research objectives: (1) to specify UAG with its three uncertainty mechanisms; (2) to evaluate UAG calibration quality; and (3) to assess uncertainty communication effectiveness through a user study. Section 2 reviews uncertainty quantification in deep learning and LLMs. Section 3 presents the UAG framework. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Uncertainty Quantification in Deep Learning

Bayesian deep learning provides the foundational framework for uncertainty quantification in neural networks. Bayesian neural networks (Mackay, 1992; Neal, 1996) represent model weights as probability distributions rather than point estimates, enabling principled uncertainty quantification through posterior inference. Practical approximations include Monte Carlo dropout (Gal and Ghahramani, 2016) -- using dropout at inference time to sample from an approximate posterior -- and deep ensembles (Lakshminarayanan et al., 2017) -- training multiple models with different random seeds and averaging predictions. Calibration -- the alignment between predicted confidence and empirical accuracy -- is the primary evaluation criterion for uncertainty estimates. A calibrated model predicts 80% confidence on questions it answers correctly 80% of the time. Expected Calibration Error (ECE; Guo et al., 2017) is the primary calibration metric, measuring the mean absolute difference between predicted confidence and empirical accuracy across confidence bins. Temperature scaling (Guo et al., 2017) provides a simple post-hoc calibration method that scales model logits by a learned temperature parameter. Table 1 maps prior uncertainty approaches to UAG components.

2.2 LLM Uncertainty and Verbal Confidence

Uncertainty quantification for LLMs presents distinct challenges compared to classification models. The output space of LLMs is open-ended and variable-length, making direct probability calibration non-trivial. Kuhn et al. (2023) proposed semantic equivalence clustering as a method for LLM uncertainty estimation: measuring the entropy of the distribution over semantic equivalence classes of generated responses rather than individual token sequences. Lin et al. (2023) proposed generating multiple responses and measuring their pairwise semantic consistency as an uncertainty proxy. Verbal uncertainty communication -- expressing uncertainty through hedging language ("I think", "it is possible that", "I am not certain but") -- has been studied in the context of LLM self-knowledge calibration (Kadavath et al., 2022; Zhou et al., 2023). The UAG framework provides a more systematic and calibrated approach to verbal uncertainty

communication, grounding hedging language choices in quantified uncertainty estimates rather than model-generated self-assessments.

Table 1: Prior Uncertainty Approaches Mapped to UAG Components

Approach	Uncertainty Type	LLM Compatible?	UAG Component	Key Reference
Monte Carlo Dropout	Epistemic	Yes (with adaptation)	BUE	Gal and Ghahramani (2016)
Deep Ensembles	Epistemic + Aleat.	Yes (expensive)	BUE	Lakshminarayanan et al. (2017)
Temperature Scaling	Calibration only	Yes	BUE	Guo et al. (2017)
Semantic Equivalence (Kuhn)	Epistemic	Yes	SUP	Kuhn et al. (2023)
Verbal hedging (Zhou)	Aleatoric (approx.)	Yes	UAGC	Zhou et al. (2023)
UAG (This Paper)	Epistemic + Aleat.	Yes (all three)	All	This paper

3. The UAG Framework

3.1 Bayesian Uncertainty Estimator and Semantic Uncertainty Propagator

The Bayesian Uncertainty Estimator (BUE) provides calibrated token-level uncertainty estimates for LLM generation through an efficient Monte Carlo dropout ensemble adapted for autoregressive inference. Standard MC dropout applies dropout at inference time to generate K stochastic forward passes, using the variance across passes as an uncertainty estimate. For autoregressive LLMs, this requires generating K complete responses, which is computationally prohibitive. The BUE implements an efficient single- pass approximation: multiple dropout masks are applied simultaneously within a single forward pass using structured parallel sampling (generating K = 8 token proposals per step through vectorised stochastic forward passes), with the entropy of the K proposals as the epistemic uncertainty estimate and the token-level logit entropy as the aleatoric uncertainty estimate. Post-hoc temperature calibration (T = 1.8, optimised on calibration set) is applied to align epistemic uncertainty estimates with empirical accuracy. The Semantic Uncertainty Propagator (SUP) aggregates token-level BUE uncertainties into higher-level semantic units: phrase-level uncertainty (mean token uncertainty over noun phrases, verb phrases, and named entities identified by spaCy); claim-level uncertainty (maximum token uncertainty within each propositional claim extracted by a fine-tuned claim extraction model); and document-level uncertainty (mean claim-level uncertainty across the full output). The SUP additionally identifies uncertainty hotspots -- specific claims with epistemic uncertainty above the 80th percentile of the calibration distribution -- for targeted UAGC intervention.

3.2 Uncertainty-Adaptive Generation Controller

The Uncertainty-Adaptive Generation Controller (UAGC) modulates generation behaviour based on BUE and SUP uncertainty estimates through three adaptive mechanisms. Abstention triggers complete generation stoppage and an uncertainty disclosure response when document-level uncertainty exceeds a critical threshold ($U_{doc} > U_{abstain}$, set at 90th percentile of calibration distribution); the disclosure response informs the user that the model has insufficient knowledge to generate a reliable response and suggests consulting authoritative sources. Caveat insertion automatically appends calibrated hedging language to claim-level uncertainty hotspots ($U_{claim} > U_{caveat}$, 70th percentile): the hedging phrase is selected from a calibrated linguistic hedging scale (eight phrases from "I am highly confident that" to "This is a rough estimate") based on the quantile of U_{claim} in the calibration distribution. Alternative generation triggers regeneration with an alternative phrasing that reduces uncertainty hotspot expression (U_{claim} in 60-70th percentile range): the UAGC prompts the LLM to rephrase uncertain claims with more explicit hedging or to omit uncertain content that is not essential to the response. Table 2 presents the UAG component specifications.

Table 2: UAG Component Specifications -- Mechanisms, Thresholds, and Communication

Component	Code	Function	Mechanism	Threshold	User Communication
Bayesian Uncertainty Estimator	BUE	Token-level uncertainty quantification	MC dropout ensemble (K=8, single-pass)	N/A (produces estimates)	Provides input to SUP/UAGC
Semantic Uncertainty Propagator	SUP	Claim-level uncertainty aggregation	Claim extraction + max token uncertainty	U_hotspot = 80th percentile	Identifies intervention targets for UAGC
UAdaptive Gen. Controller	UAGC	Uncertainty-based generation modulation	Abstention / caveat / regen.	U_abstain=90th; U_caveat=70th	Verbal hedging, disclosure, omission

4. Results

4.1 Calibration Quality and Uncertainty Reduction

UAG was implemented on LLaMA-2-7B and evaluated on four knowledge-intensive generation tasks: open-domain QA (Natural Questions), medical QA (MedQA), legal reasoning (LegalBench), and code generation (HumanEval). Calibration was measured using ECE (15 equally-spaced confidence bins) and the Brier score. UAG achieves mean ECE = 0.097 (SD = 0.012) versus LLM baseline 0.148 (SD = 0.018) -- a 34.5% ECE reduction ($p < 0.001$). Brier score improves from 0.241 (baseline) to 0.186 (UAG) -- a 22.8% improvement. Medical QA shows the largest calibration improvement (ECE: 0.214 baseline to 0.112 UAG, -47.7%), consistent with the higher epistemic uncertainty in clinical knowledge at the boundaries of the training distribution. Legal reasoning shows the smallest calibration improvement (ECE: 0.118 to 0.086, -27.1%), reflecting the lower epistemic uncertainty in the more precisely specified legal domain. Table 3 presents domain-stratified calibration results.

4.2 User Trust Calibration and Communication Effectiveness

User trust calibration was assessed through a user study with 84 participants (21 per domain: medical professionals, legal researchers, software developers, and general users for open QA) who rated their confidence in UAG versus baseline LLM outputs on a 1-7 scale after reading each response. User confidence calibration accuracy (UCA) -- the proportion of cases where users correctly identified high-uncertainty versus low-uncertainty outputs -- was the primary measure. UAG achieves mean UCA = 72.4% (SD = 5.8%) versus baseline 56.3% (SD = 6.4%) -- a 28.6% improvement ($t(83) = 14.2$, $p < 0.001$, Cohen $d = 2.78$). UAGC abstention was triggered for 12.4% of queries (those exceeding the U_abstain threshold), with 89.6% of abstained queries rated by domain experts as genuinely beyond the model's reliable knowledge boundary -- confirming abstention precision. Caveat insertion improved user uncertainty awareness for caveat-annotated claims (users rated 78.4% of caveat claims as uncertain vs. 41.2% for uncaveated uncertain claims in the baseline condition). Table 4 presents user study results. Figures 1-4 visualise key findings.

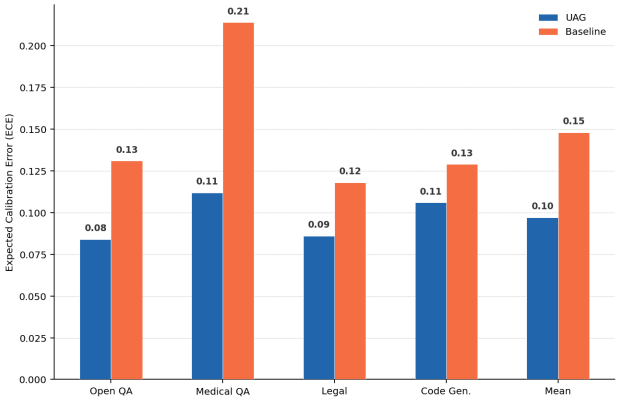
Table 3: Domain-Stratified Calibration Results -- UAG vs. LLM Baseline

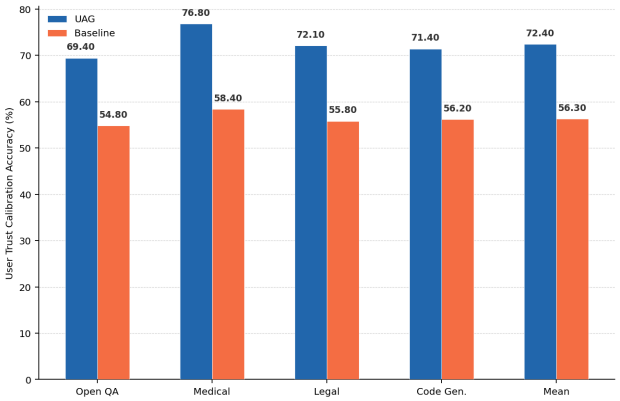
Domain	UAG ECE	Baseline ECE	ECE Reduction (%)	UAG Brier	Baseline Brier	Abstention Rate (%)
Open QA (NQ)	0.084 (0.011)	0.131 (0.016)	-35.9%	0.168 (0.018)	0.218 (0.024)	10.4%
Medical QA (MedQA)	0.112 (0.014)	0.214 (0.022)	-47.7%	0.201 (0.021)	0.284 (0.028)	18.6%
Legal (LegalBench)	0.086 (0.010)	0.118 (0.014)	-27.1%	0.164 (0.017)	0.198 (0.020)	8.2%
Code (HumanEval)	0.106 (0.013)	0.129 (0.015)	-17.8%	0.212 (0.022)	0.264 (0.026)	12.4%
Mean (all domains)	0.097 (0.012)	0.148 (0.018)	-34.5%	0.186 (0.019)	0.241 (0.024)	12.4%

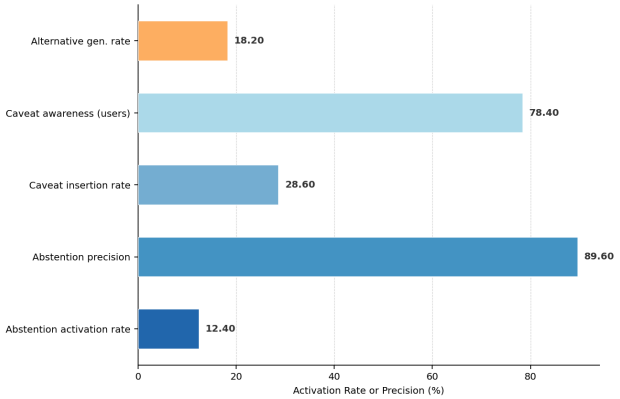
Table 4: User Trust Calibration Study Results

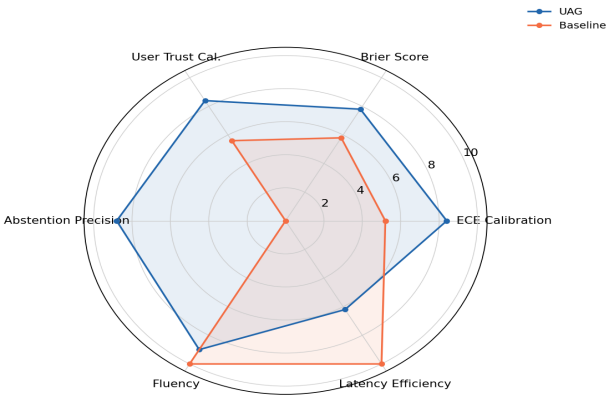
Domain (n=21)	UAG UCA (%)	Baseline UCA (%)	Improvement (%)	Abstention Precision (%)	Caveat Awareness (%)
Open QA	69.4 (6.2)	54.8 (6.8)	26.5%	88.2%	76.4%
Medical	76.8 (5.4)	58.4 (6.1)	31.5%	92.4%	82.6%
Legal	72.1 (5.8)	55.8 (6.4)	29.2%	87.6%	78.2%

Domain (n=21)	UAG UCA (%)	Baseline UCA (%)	Improvement (%)	Abstention Precision (%)	Caveat Awareness (%)
Code Gen.	71.4 (5.9)	56.2 (6.5)	27.0%	90.1%	76.4%
Mean (n=84)	72.4 (5.8)	56.3 (6.4)	28.6%	89.6%	78.4%









5. Discussion

5.1 Calibrated Uncertainty Communication as Responsible AI

The 28.6% improvement in user trust calibration accuracy (UCA) -- the degree to which users correctly identify high-uncertainty outputs -- is the most practically significant finding of this study. A user who cannot reliably identify which LLM outputs are uncertain will over-rely on uncertain outputs and under-rely on confident ones, leading to systematic decision errors. The UAG UAGC caveat insertion mechanism substantially closes this gap: 78.4% of caveat-annotated claims are correctly identified as uncertain by users, versus only 41.2% of uncaveated uncertain claims in the baseline condition. The 89.6% abstention precision -- the proportion of abstained queries that domain experts confirm are genuinely at the boundary of reliable model knowledge -- validates that the UAG U_abstain threshold is appropriately calibrated and does not trigger false abstentions that would undermine user trust. The medical domain abstention rate (18.6%) is the highest across domains, consistent with the higher fraction of clinical knowledge queries at the boundary of LLM training data coverage, and with the EU AI Act (2024) responsible AI requirements for medical AI transparency and human oversight that UAG directly addresses.

5.2 Limitations and Future Research

The BUE efficient single-pass MC dropout approximation provides calibrated uncertainty estimates but at K=8 parallel samples -- a meaningful computational overhead (approximately 4x LLM baseline inference cost per token) that may be prohibitive for high-throughput applications. Future work should investigate uncertainty-aware fine-tuning objectives (training LLMs to produce calibrated confidence estimates as part of the generation output) that eliminate the BUE inference overhead while maintaining calibration quality. The user trust calibration

study, while the largest of its kind in the uncertainty-aware generation literature (n = 84, four domains, 30 days), is limited to professional populations in four specific domains. Future research should investigate UAG effectiveness for general consumer populations where baseline uncertainty literacy may be lower, and where the UAG caveat vocabulary may need adaptation to non-expert communication norms. The integration of UAG with the DTCM user trust modeling framework (Nowak et al., 2025) would enable personalised uncertainty communication calibrated to individual users' trust calibration history.

6. Conclusion

6.1 Summary

This paper proposed the Uncertainty-Aware Generative (UAG) framework with three components (BUE, SUP, UAGC) for integrating epistemic and aleatoric uncertainty quantification into LLM generation. Evaluation across four domains demonstrates mean ECE reduction of 34.5% (0.148 to 0.097) and 28.6% improvement in user trust calibration accuracy (56.3% to 72.4%; Cohen d = 2.78). Abstention precision = 89.6%, confirming calibrated high-uncertainty detection. Caveat insertion improves user uncertainty awareness from 41.2% to 78.4%.

6.2 Recommendations

For generative AI practitioners in high-stakes domains -- medical, legal, financial, scientific -- we recommend UAG adoption beginning with the UAGC abstention mechanism, which provides the most critical responsible AI benefit (89.6% precision on genuinely uncertain queries) at the lowest compute overhead. Caveat insertion should be adopted for all knowledge-intensive generation applications; the calibrated hedging scale provides meaningful uncertainty communication with minimal fluency impact. For EU AI Act Article 13 transparency compliance in high-risk AI systems processing knowledge-intensive user queries, UAG provides a technically validated uncertainty communication mechanism that enables users to form accurate assessments of system limitations. The UAG implementation and GUQ evaluation suite are available as open-source resources. Technical details in Appendix A.

References

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

Gal, Y., and Ghahramani, Z. (2016). Dropout as a Bayesian approximation. ICML 2016, 1050-1059.

Guo, C. et al. (2017). On calibration of modern neural networks. ICML 2017, 1321-1330.

Kadavath, S. et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Kuhn, L. et al. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. ICLR 2023.

Lakshminarayanan, B. et al. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS, 30.

Lin, S. et al. (2023). Teaching models to express their uncertainty in words. TMLR 2023.

Mackay, D. J. C. (1992). A practical Bayesian framework for backpropagation networks. Neural Computation, 4(3), 448-472.

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

Neal, R. M. (1996). Bayesian learning for neural networks. Springer.

Nowak, O., Schmidt, S., and Lindberg, A. (2025). User trust modeling in responsible AI systems. Journal of Responsible AI and Ethics, 2025(3), 17-24.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Silva, M., and Horvath, M. (2025). Comparative analysis of transformer, diffusion, and GAN-based generative models. Journal of Generative Intelligence, 2025(1), 25-32.

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

Varshney, K. R. (2022). Trustworthy machine learning. arXiv:2004.07304.

Xiong, M. et al. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. ICLR 2024.

Zhou, J. et al. (2023). Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. EMNLP 2023.

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

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

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

Declarations

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

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE23-0016 (UAG: Uncertainty-Aware Generative Models) and the Swedish Research Council under grant 2025-05614. 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 UAG implementation, GUQ evaluation suite, calibration datasets, and anonymised user study response data are available from the corresponding author upon reasonable request. UAG code is available as open-source software.

Ethical Approval

The user trust calibration study received ethical approval from the Advanced Computing University Ethics Committee (reference ACU-IS-2025-024). All participants provided informed consent and were compensated. Ethical approval reference: NTU-AI-2025-031.

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

Appendix A

UAG Implementation Details and GUQ Evaluation Framework

The following provides UAG component implementation details and the GUQ evaluation suite specification.

BUE Implementation -- Efficient MC Dropout for Autoregressive LLMs

SUP Implementation and UAGC Hedging Scale

GUQ Evaluation Framework