Multiobjective Evolutionary Calibration of Uncertainty Quantification for Large Language Models

Maria Mora-Cross
maria.mora@itcr.ac.cr
Costa Rica Institute of Technology
Costa Rica

Saúl Calderón-Ramírez
sacalderon@itcr.ac.cr
Costa Rica Institute of Technology
Costa Rica

Sebastián Rojas González
sebastian.rojasgonzalez@ugent.be
Ghent University - IMEC
Belgium

Abstract

Large language models (LLMs) are increasingly used in high impact applications such as clinical decision support, where fluent outputs can hide critical errors. Uncertainty quantification (UQ) should discriminate correct from incorrect answers and remain informative under dataset shift. We formulate UQ design as a multiobjective evolutionary optimization problem over unsupervised estimators derived from hidden state embeddings. Candidate estimators rely on clustering algorithms followed by a kNN based uncertainty signal. We evaluate Llama 3.1 8B, and DeepSeek 7B on MedQuAD, with a suite of out of distribution datasets (OOD). Objective functions minimize Pearson correlation between UQ and text quality metrics (BERTScore F1 and ROUGE L). Using RVEA, we analyze the search dynamics and the resulting Pareto front of the trade off between semantic and lexical quality. Preliminary results show competitive AUROC and AURC, and improved AUROC relative to a Monte Carlo Dropout (MCD) baseline, highlighting the potential of evolutionary multiobjective optimization as a robust and scalable strategy for building reliable, risk aware generative AI systems.

CCS Concepts

• **Computing methodologies** → **Machine learning algorithms**; **Natural language generation**; **Genetic algorithms**; **Dimensionality reduction and manifold learning**; **Cluster analysis**.

Keywords

uncertainty quantification, large language models, multiobjective evolutionary optimization, hidden state embeddings

ACM Reference Format:
Maria Mora-Cross, Saúl Calderón-Ramírez, and Sebastián Rojas González. 2026. Multiobjective Evolutionary Calibration of Uncertainty Quantification for Large Language Models. In *Genetic and Evolutionary Computation Conference (GECCO Companion '26), July 13–17, 2026, San Jose, Costa Rica.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3795101.3805422

1 Introduction

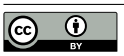
Large Language Models (LLMs) have shown strong performance in many applications, including clinical decision support, but their reliability remains limited when queries require knowledge not well represented in training data, often leading to hallucinations [6]. This has motivated growing interest in uncertainty quantification (UQ) methods for safer deployment in high risk domains [1, 15]. Among the most promising directions are embedding based approaches that analyze the geometry of hidden representations through techniques such as eigenvalue analysis, convex hull methods, and semantic entropy [3, 16]. These methods can better capture predictive uncertainty from internal activations, but their effectiveness depends strongly on hyperparameter choices [16]. To address this, we formulate UQ as the design of an unsupervised estimator over hidden representations, combining Principal Component Analysis (PCA), Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), and K-Nearest Neighbors (kNN) distances on final layer embeddings from Llama 3.1 8B and DeepSeek 7B. We then use Reference Vector Guided Evolutionary Algorithm (RVEA) [4] to optimize uncertainty estimators by minimizing their correlation with Bidirectional Encoder Representations from Transformers Score (BERTScore) F1 [19] and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) L [11], producing a Pareto front that balances semantic and lexical criteria and is later evaluated with AUROC and risk coverage analysis.

This work makes the following contributions:

- We propose a multiobjective evolutionary framework for optimizing embedding-based uncertainty estimators in generative question answering (QA). The framework approximates the Pareto front of solutions that balance semantic and lexical sensitivity to detect incorrect generations, using objectives defined through the correlation of UQ with BERTScore F1 and ROUGE-L. Our results show that these criteria play complementary roles, and that the most effective selective prediction behavior emerges from solutions that preserve a meaningful balance between them. We compare the resulting estimators against a Monte Carlo Dropout (MCD) baseline, showing that the optimized embedding based estimators achieve stronger AUROC across correctness definitions.
- We introduce a semantic prioritization analysis using calibrated percentile thresholds over complementary quality metrics, enabling controlled evaluation across different correctness definitions. Results show that uncertainty performance remains stable under semantic prioritization, lexical variability, and distribution shift.

2 State of the Art

Uncertainty quantification in LLMs has moved from classical methods such as MCD [8] and Deep Ensembles [10] to approaches better suited to autoregressive generation. Because token level signals often correlate weakly with semantic correctness under



This work is licensed under a Creative Commons Attribution 4.0 International License.
GECCO Companion '26, San Jose, Costa Rica
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2488-6/2026/07
https://doi.org/10.1145/3795101.3805422

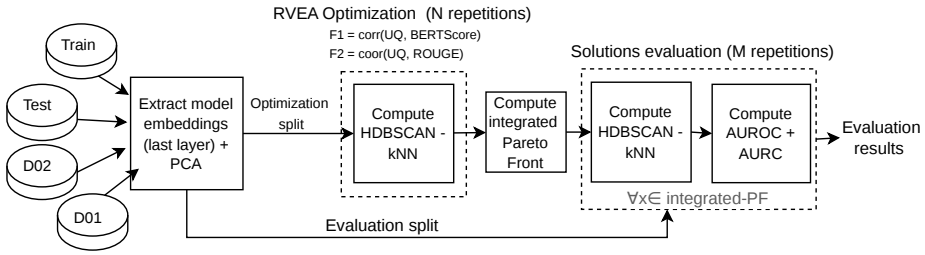


Figure 1: Workflow for embedding space uncertainty estimation. Final layer embeddings are reduced with PCA and normalized, then used to compute HDBSCAN and kNN uncertainty scores. RVEA optimizes $F_1 = \text{corr}(\text{UQ}, \text{BERTScore})$ and $F_2 = \text{corr}(\text{UQ}, \text{ROUGE-L})$ over N repetitions to produce an integrated Pareto front ($N = 20$ for Llama and 16 for Deepseek), whose solutions are evaluated with AUROC and AURC ($M = 15$).

distribution shift, recent work increasingly uses hidden representation spaces [6]. In generative QA, evaluation commonly uses ROUGE, BERTScore, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Risk–Coverage Curve (AURC).

3 Proposed Method

In this work, we define an uncertainty estimator as a function that quantifies the reliability of an LLM prediction using signals from its internal representations. Fig. 1 presents the whole pipeline.

3.1 Uncertainty Evaluation Metrics, Datasets and Models

We evaluate uncertainty quality using three complementary measures: AUROC, AURC, and rejection accuracy at fixed coverage levels. AUROC measures how well uncertainty distinguishes correct from incorrect generations, with higher values indicating better separability [9, 12]. AURC measures selective prediction performance through the risk coverage curve, where lower values indicate more effective uncertainty guided rejection [7]. As a reference baseline, we also evaluate MCD under the same repeated calibration and evaluation protocol, using 10 model generations, and report its AUROC. We use Medical Question Answering Dataset (MedQuAD) [2] as the in distribution dataset, with 11,576 training, 1,447 validation, and 1,447 test samples. Robustness under distribution shift is evaluated with: Minecraft QA (D01) [17], containing 694,814 synthetic samples from Minecraft Wiki articles, and Medical Meadow WikiDoc (D02) [13], with 10,000 synthetic medical QA instances.

Experiments use the Transformers library [18] with Llama 3.1 8B and DeepSeek 7B, both fine tuned on MedQuAD with LoRA [5]. Llama 3.1 8B reached training and validation losses of 0.8334 and 0.8959, and DeepSeek 7B reached 1.0367 and 1.0779.

3.2 Embedding-Based Uncertainty Estimation

Given an LLM after fine tuning on the training samples of the MedQuaD dataset, we extract sentence level embeddings from the last hidden layer and fit PCA on the resulting training embeddings to obtain a compact representation. We then construct an uncertainty estimator based on two components: clustering via HDBSCAN, trained on the MedQuaD training split to identify dense semantic regions in the embedding space and assign points either to clusters or to noise, and local density estimation using kNN. For each evaluation instance, taken from the test splits of MedQuaD, D01, and D02, we compute the corresponding embedding vector and project it with the fitted PCA transformation. The final uncertainty score combines the cluster membership and the distance to the 10 nearest neighbors.

3.3 Joint Correctness Labeling Using Complementary Metrics

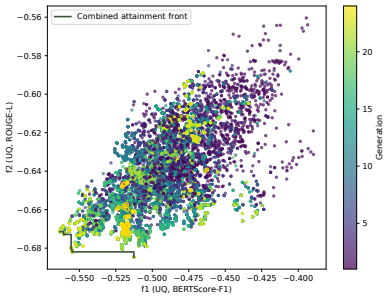
To evaluate uncertainty quality, we require a definition of correctness for generated answers. Unlike classification tasks, correctness in generative QA is inherently ambiguous. To reduce metric bias, we define correctness using two complementary evaluation metrics, BERTScore F1 and ROUGE-L, through the joint criterion $y = \mathbb{I}(\text{BERTScore} \geq \tau_b \ \wedge \ \text{ROUGE-L} \geq \tau_r)$, where τ_b and τ_r are percentile-based thresholds estimated on a calibration split.

3.4 Calibration Strategy

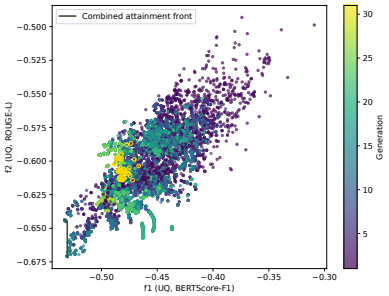
To avoid evaluation bias, all correctness thresholds are computed only on a calibration subset comprising 30% of the in distribution test data, while the remaining 70% and the out of distribution datasets D01 and D02 are reserved exclusively for evaluation. Thresholds are defined as fixed percentiles of the calibration metric distributions, namely $\tau_b = \text{percentile}_{p_b}(\text{BERTScore})$ and $\tau_r = \text{percentile}_{p_r}(\text{ROUGE-L})$, which provides stable and reproducible correctness definitions despite scale differences between metrics. Since semantic fidelity is more important than surface lexical overlap in generative QA, we analyze semantic prioritization using $(p_b, p_r) \in \{(30, 30), (30, 20), (25, 20), (0, 30), (30, 0)\}$, corresponding to balanced, moderately semantic prioritized, strongly semantic prioritized, lexical only, and semantic only settings, respectively. This design enables a controlled study of uncertainty robustness under progressively relaxed correctness criteria while explicitly varying the semantic and lexical trade off.

4 Experiments

Our experiments evaluate the effectiveness, robustness, and stability of embedding space uncertainty estimators optimized through multiobjective evolutionary search. All experiments were conducted



(a) Llama 3.1 8B: integrated Pareto front.



(b) DeepSeek 7B: integrated Pareto front.

Figure 2: Integrated Pareto fronts aggregated over N independent evolutionary repetitions for Llama 3.1 8B (top) and DeepSeek 7B (bottom) (N=20 for Llama and 16 for Deepseek). The fronts illustrate the trade off between semantic sensitivity (correlation between UQ scores and BERTScore F1) and lexical sensitivity (correlation between UQ scores and ROUGE L).

on the Kabré HPC cluster using NVIDIA L40S GPUs. The study is guided by the following research questions: 1) Can final layer embedding based uncertainty estimators from Llama 3.1 8B and DeepSeek 7B reliably detect incorrect generations and support safe abstention, namely, the ability to avoid answering when the predicted output is likely to be unreliable, in both in distribution and out of distribution settings?; 2) Does jointly optimizing semantic and lexical uncertainty objectives produce a diverse Pareto front with meaningful trade offs between BERTScore F1 and ROUGE L? 3) How stable are AUROC, risk coverage behavior, and AURC across correctness definitions based on calibrated BERTScore F1 and ROUGE L thresholds?

4.1 Multiobjective Optimization of Uncertainty Estimators

We formulate the uncertainty estimator design as a multiobjective problem solved with RVEA. Each solution defines a full embedding space pipeline, and we minimize $f_1 = \text{corr}(UQ, \text{BERTScore})$ and $f_2 = \text{corr}(UQ, \text{ROUGE-L})$, encouraging uncertainty to be inversely related to both semantic and lexical quality. Since the Pearson correlation lies in $[-1, 1]$, minimizing these objectives favors more negative values, encouraging higher uncertainty for lower quality answers. This is a surrogate optimization criterion, later validated through AUROC, AURC, and rejection accuracy. The search space includes HDBSCAN cluster size $[5, 20]$, amount of samples $[1, 20]$, epsilon $[0.0, 0.5]$, kNN amount of samples $[5, 35]$, and embedding dimension $\{128, 256\}$. Higher dimensions $\{512, 1024\}$ were discarded because they were consistently dominated and more costly. Multiple runs (20 for Llama and 16 for Deepseek) with different seeds are merged into one integrated PF per model (Fig. 2), capturing the main trade offs between semantic and lexical sensitivity. The optimization is performed offline once during calibration and does not add cost at inference time.

4.2 Evaluation Protocol

We use a two stage protocol to evaluate uncertainty quality while accounting for both calibration variability and estimator diversity. For each Pareto solution $s \in \mathcal{P}$, we fit HDBSCAN on 5,000 in distribution training samples and compute uncertainty scores on 600 in distribution test samples plus 75 out of distribution samples from D01 and D02. The evaluation is repeated $R = 15$ times, each time splitting the in distribution test set into 30% calibration and 70% evaluation, estimating percentile thresholds on the calibration split, defining correctness with $(p_b, p_r) \in \{(30, 30), (30, 20), (30, 0), (0, 30)\}$, and measuring AUROC, AURC, and rejection accuracy at coverage levels $\{80\%, 90\%, 95\%, 100\%\}$. To characterize global model behavior, results are further aggregated across all solutions in the Pareto front. Final performance statistics are computed as empirical averages across both dimensions:

$$\mu = \frac{1}{|\mathcal{P}|\,R} \sum_{s \in \mathcal{P}} \sum_{r=1}^{R} M_{s,r}, \quad \sigma^2 = \frac{1}{|\mathcal{P}|\,R} \sum_{s \in \mathcal{P}} \sum_{r=1}^{R} (M_{s,r} - \mu)^2$$

Where $M_{s,r}$ denotes a metric computed for solution $s \in \mathcal{P}$ at repetition r.

5 Results

We analyze optimization dynamics, uncertainty discrimination, selective prediction, and robustness across settings.

5.1 Optimization Behavior

Fig. 2 shows the integrated Pareto front aggregated over independent runs for Llama 3.1 8B and DeepSeek 7B. Since search trajectories are colored by generation rather than by run, consistency across runs is assessed through the repeated run statistics. For both models, the optimization yields diverse nondominated solutions and a modest but meaningful trade off between semantic and lexical sensitivity. Balanced configurations outperform schemes that remove either constraint, indicating that robust uncertainty estimation benefits from combining both signals.

Across repetitions, RVEA converges to stable Pareto fronts and consistently favors lower dimensional embeddings (128 and 256), which provide strong uncertainty discrimination at lower computational cost. Early stopping was also reliable, with Llama 3.1 8B reaching the best hypervolume (HV) of 0.404 ± 0.036 and DeepSeek 7B 0.333 ± 0.045, relative HV gains of $5.08 \pm 2.72\%$ and $5.66 \pm 2.84\%$, and average stopping generations of 18.9 ± 3.9 and 19.7 ± 5.2, respectively, across 20 and 16 runs.

Table 1: Performance summary by model (mean ± std over 15 repetitions). Higher AUROC indicates better discrimination between correct and incorrect generations, while lower AURC reflects more effective selective prediction and safer abstention behavior. Results include the proposed method and the MCD baseline. Schemes Bp_b Rp_r denote percentile-based correctness thresholds, where p_b is the BERTScore-F1 percentile and p_r is the ROUGE-L percentile used to define joint correctness (e.g., B25 R10 corresponds to $p_b = 25$, $p_r = 10$) ($n = Paretofront(PF)solutions$).

Model	Scheme	AUROC	AURC	Acc (80%)	Acc (90%)	Acc (95%)	Acc (100%)	n	MCD AUROC
Llama 3.1 8B	B0 R30	0.767 ± 0.007	**0.190 ± 0.003**	**0.713 ± 0.005**	**0.675 ± 0.001**	**0.652 ± 0.001**	**0.626 ± 0.000**	11	0.660 ± 0.024
	B25 R10	0.755 ± 0.009	0.198 ± 0.004	0.702 ± 0.005	0.673 ± 0.004	0.649 ± 0.001	0.623 ± 0.000	11	0.645 ± 0.025
	B30 R0	0.747 ± 0.007	0.200 ± 0.003	0.697 ± 0.004	**0.675 ± 0.003**	0.651 ± 0.001	0.625 ± 0.000	11	0.637 ± 0.023
	B30 R20	0.759 ± 0.009	0.213 ± 0.004	0.674 ± 0.005	0.646 ± 0.004	0.622 ± 0.001	0.597 ± 0.000	11	0.659 ± 0.026
	B30 R30	**0.782 ± 0.008**	0.233 ± 0.004	0.642 ± 0.005	0.604 ± 0.003	0.578 ± 0.000	0.553 ± 0.000	11	**0.691 ± 0.015**
DeepSeek 7B	B0 R30	0.726 ± 0.016	**0.207 ± 0.007**	0.691 ± 0.007	0.663 ± 0.001	**0.643 ± 0.001**	**0.626 ± 0.000**	4	0.664 ± 0.021
	B25 R10	0.728 ± 0.012	0.212 ± 0.005	**0.699 ± 0.004**	**0.665 ± 0.002**	0.642 ± 0.001	0.623 ± 0.000	4	0.653 ± 0.023
	B30 R0	0.728 ± 0.011	0.210 ± 0.005	0.696 ± 0.005	**0.665 ± 0.002**	**0.643 ± 0.001**	0.625 ± 0.000	4	0.642 ± 0.019
	B30 R20	0.730 ± 0.011	0.228 ± 0.005	0.668 ± 0.004	0.637 ± 0.002	0.614 ± 0.001	0.597 ± 0.000	4	0.665 ± 0.022
	B30 R30	**0.740 ± 0.013**	0.255 ± 0.006	0.624 ± 0.005	0.594 ± 0.002	0.571 ± 0.001	0.553 ± 0.000	4	**0.694 ± 0.013**

5.2 Uncertainty Discrimination Performance

Table 1 shows strong uncertainty discrimination for both models, with high AUROC and low AURC across correctness definitions. Llama 3.1 8B consistently outperforms DeepSeek 7B, and performance remains stable across semantic and lexical criteria.

Relative to the MCD baseline, the proposed embedding based estimators achieve higher AUROC across all reported correctness schemes for both Llama 3.1 8B and DeepSeek 7B. This suggests that uncertainty signals derived from the optimized embedding space structure provide stronger discrimination between correct and incorrect generations than the baseline dropout based approach.

6 Conclusions

This work presents a multiobjective evolutionary framework for optimizing embedding based uncertainty estimators in generative QA. By combining semantic and lexical criteria, it produces robust and interpretable uncertainty signals. The resulting Pareto fronts show clear trade offs between UQ correlation with BERTScore F1 and ROUGE L, confirming that no single configuration optimizes both objectives.

Experiments with final layer embeddings from Llama 3.1 8B and DeepSeek 7B show that the proposed estimators effectively distinguish correct from incorrect generations, achieving strong AUROC, competitive AURC, good rejection accuracy, and better AUROC than an MCD baseline. Under out of distribution evaluation, they also preserve meaningful risk coverage behavior, supporting multiobjective evolutionary optimization as a useful approach for uncertainty aware generative QA. The code is publicly available [14].

Acknowledgments

Computational support was provided by the Kabré supercomputer at the National High Technology Center of Costa Rica (CeNAT). Sebastián Rojas González acknowledges support by FWO Belgium (grant 1216021N).

References

[1] Miguel Abreu-Cardenas, Saúl Calderón-Ramírez, and Martín Solís. 2023. Uncertainty Estimation for Complex Text Detection in Spanish. In *2023 IEEE 5th International Conference on BioInspired Processing (BIP)*. IEEE, 1–6.

[2] Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics* 20, 1 (Dec. 2019), 511.

[3] Ferhat Ozgur Catak and Murat Kuzlu. 2024. Uncertainty quantification in large language models through convex hull analysis. *Discover Artificial Intelligence* 4, 1 (Nov. 2024), 90.

[4] Ran Cheng, Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. 2016. A Reference Vector Guided Evolutionary Algorithm for Many-Objective Optimization. *IEEE Transactions on Evolutionary Computation* 20, 5 (Oct. 2016), 773–791.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314* (2023).

[6] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (June 2024), 625–630. doi:10.1038/s41586-024-07421-0

[7] Tom Fawcett. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. doi:10.1016/j.patrec.2005.10.010

[8] Yarin Gal and Zoubin Ghahramani. 2015. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *ICML 2016* (2015).

[9] James A. Hanley and Barbara J. McNeil. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143, 1 (1982).

[10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2016. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NIPS 2017* (2016).

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[12] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. arXiv:2503.15850 [cs.CL] https://arxiv.org/abs/2503.15850

[13] MedAlpaca. 2023. medical_meadow_wikidoc. Available at: `https://huggingface.co/datasets/medalpaca`.

[14] Maria Mora. 2026. *gecco2026_UQ_LLMs*. Retrieved April 10, 2026 from https://github.com/marimoracross/gecco2026_UQ_LLMs

[15] Maria Mora-Cross and Saul Calderon-Ramirez. 2024. Uncertainty Estimation in Large Language Models to Support Biodiversity Conservation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. Association for Computational Linguistics, Mexico City, Mexico, 368–378.

[16] Maria Mora-Cross and Saúl Calderón-Ramírez. 2025. Uncertainty Quantification in Large Language Models Using Feature Space Density and Clustering. In *Proceedings of the 2025 International Conference on Bio-inspired Information Processing (BIP 2025)*.

[17] Naklecha. 2024. minecraft-question-answer-700k. Available at: https://huggingface.co/datasets/naklecha/minecraft-question-answer-700k. Synthetic Minecraft Q&A dataset, CC BY-NC-SA 3.0.

[18] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-Art Natural Language Processing. In *In Proc. of EMNLP 2020: System Demonstrations, ACL*. Association for Computational Linguistics, Online, 38–45.

[19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *ICLR 2020* (2019).

Received 21 January 2026; revised 12 March 2026; accepted 5 June 2026