Beyond Greedy Decoding: Model-Specific Strategy Selection via Multi-faceted Uncertainty Decomposition

Kwangje Baeg[1], Yubin Lim[2]
[1]Korea Telecom, [2]Seoul National University
kwangje.baeg@kt.com, yblim@idb.snu.ac.kr

Large Language Models (LLMs) rely on static decoding strategies despite significant differences in the difficulty of generation. Recent uncertainty-based approaches aggregate diverse signals, overlooking model heterogeneity—particularly pronounced in morphologically rich languages (e.g., Korean) where tokenization variations lead to unique uncertainty traits. We focus on Korean instruction-tuned LLMs and decompose uncertainty into three largely independent components—Semantic Entropy, Graph Laplacian, and Trajectory Consistency. Unsupervised clustering reveals model-specific behavioral profiles with marked heterogeneity, challenging aggregation-based approaches and supporting uncertainty-guided strategy selection. High generation quality does not correlate with low output diversity, and universal decoding strategies fail for heterogeneous models. Cross-dataset validation shows that uncertainty patterns capture transferable model characteristics, enabling practitioners to systematically select strategies based on generation context.

1. Introduction

Current Large Language Models employ fixed decoding strategies irrespective of the nature of the task. A model employs the same stochastic sampling technique regardless of whether it addresses factual inquiries with minimal uncertainty or produces creative content that gains from exploration. This results in a consistent mismatch between the model's internal confidence in its predictions and the decoding approach used. Although adaptation guided by uncertainty has garnered recent attention, its practical implementation is still restricted. Current studies mainly focus on base models, while instruction-tuned versions remain underexplored, even though they are commonly used in operational systems and alignment training significantly influences uncertainty behaviors [1, 2].

Existing adaptive methods face a fundamental limitation: they reduce multidimensional uncertainty signals to scalar values, compromising the detail required for successful intervention. Various sources of uncertainty require distinct responses. Semantic variety implies nucleus sampling [3], trajectory variability advocates for contrastive decoding [4, 5], and structural uniformity supports deterministic approaches [6]. Combining these distinct elements into a single metric obscures which intervention would be the most effective. Static mappings between uncertainty and decoding strategies exacerbate this problem—hyperparameters tuned for logical reasoning frequently lead to significant performance degradation when used for creative generation tasks.

Korean amplifies these challenges. Agglutinative morphology interacts with frequency-based tokenization to create highly model-specific fragmentation patterns. Identical semantic inputs often fragment differently across models, artificially inflating entropy measurements regardless of actual semantic uncertainty (see Appendix A.9 for detailed examples). This makes universal thresholds ineffective, as architectural variance overshadows semantic signals. We investigate 6 instruction-tuned Korean LLMs to systematically characterize these patterns, addressing a critical gap in uncertainty quantification for morphologically rich languages.

We challenge standard decoding assumptions through four key findings:

Third Conference on Parsimony and Learning (CPAL 2026).

- Greedy decoding proves optimal in only 13.9% of cases—marginally above the 10% random baseline expected when comparing 10 strategies, yet practitioners typically default to greedy without considering alternatives. This suggests that while greedy performs reasonably, systematic strategy selection could improve 86.1% of generations.
- Unsupervised clustering of 19 model-level behavioral features identifies three distinct behavioral profiles. We decompose generation uncertainty into Semantic Entropy, Graph Laplacian, and Trajectory Consistency—components exhibiting low but statistically significant inter-correlations ($r < 0.3$, all $p < 0.01$), indicating largely independent signals that nonetheless share minor common variance across 6,468 measurements. Clustering 19 behavioral features reveals three preliminary behavioral profiles, suggesting distinct model families that require validation with larger model populations, each requiring different optimization approaches. Quality and diversity remain uncorrelated ($\rho = +0.23$, $p = 0.62$), with high-performing models occupying intermediate positions rather than extremes of determinism or variability. This decoupling challenges such universal threshold assumptions.
- Optimization guidelines must be profile-specific. By characterizing strategy-profile interactions, we identified which decoding methods succeed for each behavioral profile. Temperature control offered near-universal benefits (serves as a robust baseline), whereas beam search and DoLa exhibited effect size reversals across profiles (e.g., beam search: +5.9% for Profile C vs. -9.1% for Profile B). These results provide practical guidelines for selecting strategies based on model characteristics rather than relying on dataset-agnostic heuristics.
- Intrinsic model traits are confirmed via cross-dataset validation. We observed strong pattern transfer ($r = 0.87$, $p < 0.001$) between Phase 1 calibration (K^2-Eval, LogicKor) and Phase 3 validation (KoMT-Bench). This demonstrates that our identified profiles capture intrinsic model characteristics rather than dataset artifacts. We successfully reproduced three of four validation aspects across independent distributions, addressing the persistent limitation of single-dataset evaluations.

2. Related Work

Quantification of Uncertainty in Language Models. Initial uncertainty estimation relied significantly on Bayesian approaches—resource-intensive, impractical for large-scale applications. The field then shifted toward efficient approximations. Computer vision imports (dropout-based estimations [7], deep ensembles [8]) and Bayesian approaches [9] first appeared promising but failed to scale in the context of billion-parameter practicality. Metrics at the token level [10] and conformal prediction [11], along with calibration techniques [12], enhanced efficiency. However, tokenization artifacts remained a persistent challenge. Subword segmentation methods [13], while effective for most languages, introduce model-specific fragmentation patterns that complicate uncertainty estimation. Morphologically rich languages amplify this: word boundaries blur, fragmenting identical meanings differently across models.

Semantic-space methods emerged as the practical solution, operating in latent space to avoid tokenization artifacts. Semantic Entropy [14] clusters generations by meaning rather than surface form, while Graph Laplacian approaches [15] measure uncertainty geometrically through semantic manifolds. Recent work extends these principles to diverse generative contexts [16–18].

Uncertainty-Guided Strategy Selection. We argue that current systems suffer primarily from oversimplified uncertainty aggregation. By compressing d-dimensional uncertainty into a single scalar value, existing methods discard critical, component-specific signals. Different failure modes require distinct interventions: high semantic diversity calls for nucleus sampling [3] or typical sampling [19], whereas trajectory instability necessitates contrastive decoding [4, 5]. Scalar aggregation conflates these orthogonal signals, making targeted intervention impossible.

Moreover, static mappings between uncertainty and strategy are fragile under distribution shifts. Parameters like temperature or top-p, optimized for logical reasoning, often degrade performance

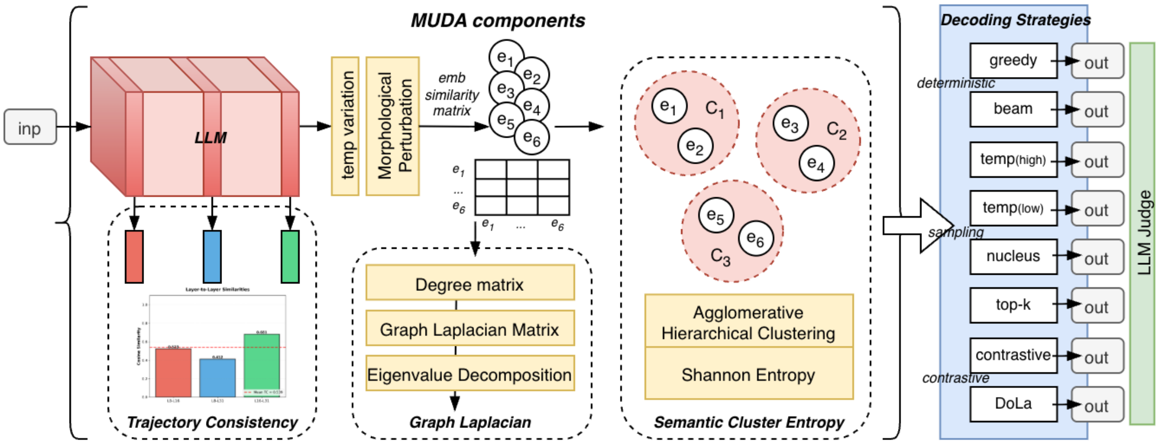

Figure 1: Phase 1 model characterization pipeline. MUDA decomposes generation uncertainty into Semantic Entropy (SE), Graph Laplacian (GL), and Trajectory Consistency (TC) from diverse responses generated via temperature variation and morphological perturbation.

on creative tasks. While recent adaptive approaches [20–22] attempt dynamic selection, they focus predominantly on English models. Our work addresses the critical gap of characterizing uncertainty in morphologically rich languages like Korean [23–25], where functional morphemes interact with tokenization in model-specific ways, inflating entropy measurements and rendering universal thresholds ineffective.

Decoding Strategies and Model-Specific Behavior. Most decoding frameworks treat models as interchangeable black boxes, applying uniform strategies despite significant architectural differences. Although some studies have noted calibration variations across model scales [26], the implications for strategy optimization remain largely unexplored. Recent confidence-based approaches [27, 28] and uncertainty-aware methods [29, 30] rely on hyperparameters that generalize poorly, and their evaluations are typically confined to single datasets.

Contrastive methods like DoLa [31] and related contrastive decoding [4, 32, 33] offer promising directions but exhibit heavily model-dependent effectiveness. This suggests that optimal decoding requires model-specific calibration rather than universal rules [19]. Yet, prior work lacks a systematic characterization of which strategies benefit specific model types—or whether these patterns transfer across different tasks.

3. Methodology

3.1. MUDA: Multi-faceted Uncertainty Decomposition and Analysis

Standard uncertainty quantification reduces multidimensional signals to scalars [14, 15]. Aggregating semantic diversity, trajectory instability, and structural inconsistency into a single metric obscures how distinct uncertainty sources impact generation quality, preventing targeted intervention. MUDA (Multi-faceted Uncertainty Decomposition and Analysis) preserves these dimensions by decomposing uncertainty into three complementary components for granular behavioral diagnostics.

3.1.1. Uncertainty Components

Semantic Entropy (SE) captures uncertainty by measuring semantic drift across sampled generations through clustering [14]. Our implementation differs fundamentally by extracting representations directly from the target model's final layer rather than using external encoders like sentence transformers [34]. This model-native design offers critical advantages: it captures the source model's intrinsic semantic organization; improves efficiency by reusing generation embeddings; and absorbs tokenization heterogeneity directly into the representation space. Crucially, it grants access to intermediate hidden states (l_1, l_2) required for Trajectory Consistency (TC), enabling us to decompose process instability from final outcome uncertainty.

We extract mean-pooled vectors from the last transformer layer to map the geometric structure of N response candidates. Agglomerative clustering groups responses by cosine similarity, identifying outputs sharing similar latent representations. Normalized entropy over resulting cluster probabilities (Eq. 3) quantifies semantic dispersion. Elevated SE values indicate mutually incompatible interpretations rather than stylistic variants, anchoring all MUDA components in the model's native latent space (detailed in Appendix A.1).

Graph Laplacian (GL). Manifold-based uncertainty quantification [15] models response embeddings as graph structures, capturing global semantic relationships beyond pairwise clustering. While prior work emphasizes eigenvector centrality and connectivity analysis, we employ a streamlined formulation using normalized Laplacian trace (Eq. 7), which quantifies manifold smoothness through spectral properties.

We construct a similarity graph from response embeddings, where edge weights reflect cosine similarities. The normalized Laplacian of this graph encodes structural properties through its eigenvalue spectrum. Graph trace—the sum of eigenvalues—proxies embedding dispersion: tight clusters yield small eigenvalues (low GL), while scattered responses produce large eigenvalues (high GL). This complements SE naturally: SE counts discrete semantic clusters, GL measures continuous geometric spread. A model might generate semantically similar responses (low SE) that nonetheless scatter stylistically across embedding space (high GL). Full mathematical details appear in Appendix A.1.

Trajectory Consistency (TC) differs from SE and GL by examining the generation process rather than final outcomes. TC tracks hidden state evolution across network depth, revealing generation instability through inconsistent layer-wise representations that precede semantic uncertainty in final outputs. We sample three strategic layers: early ($l_1 = \lfloor L/4 \rfloor$, lexical/syntactic), middle ($l_2 = \lfloor L/2 \rfloor$, compositional), and late ($l_3 = L - 1$, abstract semantic). Average pairwise dissimilarity per layer (Eq. 9) is aggregated into the final TC score (Eq. 10). High TC indicates that perturbations—temperature shifts, morphological variants—trigger divergent internal trajectories despite similar final outputs. This metric isolates process instability from semantic uncertainty, remaining largely independent of SE/GL while capturing complementary aspects of generation behavior. Unifying all three components in the same embedding space enables profiling that scalar aggregation would otherwise obscure. Detailed formulations are provided in Appendix A.1.

3.1.2. Morphology-Aware Perturbation

For temperature variation, we sample at $T \in \{0.7, 0.9, 1.1\}$, yielding $N_{\text{orig}} = 3$ responses across low, standard, and high diversity regimes.

Korean's agglutinative structure allows particle/ending substitution without semantic change. We exploit this directly on each model's tokenized representation. The procedure consists of three stages. First, we tokenize the input using the model's native tokenizer. Second, we identify functional morphemes through pattern matching—case particles (*eun/neun*, *i/ga*, *eul/reul*), adverbial markers (*e*, *ui*, *ro*), and polite endings (*imnida/ieyo*). Third, we substitute grammatically valid alternatives (*eun*↔*neun* for topics) before decoding. Subword fragmentation sometimes breaks tokenizer-based matching. Fallback heuristics target common morphological patterns (Algorithm: Appendix A.1.5).

We generate $N_{\text{pert}} = 2$ perturbed prompts at $T = 0.8$, yielding $N = 5$ total samples. The critical observation is that semantically equivalent variants trigger different tokenization patterns. Example: *sudo-neun* (capital-TOPIC) becomes [*sudo, neun*] in one model, [*sudoneun*] in another. This enables model-intrinsic uncertainty estimation under morphologically natural perturbations. Phonological well-formedness (consonant-conditioned allomorphy) is not explicitly modeled in our implementation. Despite these violations, empirical results demonstrate effective profile differentiation (Section 4.1). LLMs handle minor morphological irregularities robustly during generation. Our morphological perturbation specifically probes tokenization sensitivity: by keeping the semantics constant while altering surface forms, we disentangle uncertainty arising from architectural fragmentation (e.g., subword splitting differences) from genuine semantic uncertainty.

3.2. Phase 1: Comprehensive Model Characterization

Figure 1 illustrates the comprehensive characterization pipeline. We evaluate 6 Korean instruction-tuned LLMs (1.5B–8B parameters) across 924 samples from two benchmarks: K^2-Eval [35] (630 samples, knowledge QA) and LogicKor [36] (294 samples, logical reasoning), enabling discovery of systematic patterns across production-scale models.

We evaluate 10 decoding strategies across three paradigms: greedy, sampling-based methods (temperature, nucleus [3], top-k [3], typical [19]), and search-based methods (beam search [6], contrastive decoding [4, 5], DoLa [31]). For each (model, sample, decoder) triple, we measure MUDA components via 5 perturbed responses (§3.1), generate responses, and score with GPT-4o-mini [37] using the evaluation prompts from the official KoMT-Bench repository, yielding 6,468 MUDA measurements and 64,680 quality scores.

3.3. Phase 2: Behavioral Pattern Discovery

To identify systematic patterns relating uncertainty characteristics to strategy effectiveness, we extract 19 behavioral features per model (detailed in Appendix A.2). We apply PCA [38] to reduce these dimensions to 2 principal components preserving 82.1% variance. Subsequent K-means clustering ($k = 3$) identifies distinct model families. We analyze these profiles along three dimensions: architectural uncertainty distribution, the quality-diversity relationship, and profile-specific strategy effectiveness (detailed analysis in Section 4.1).

3.4. Phase 3: Independent Cross-Dataset Validation

We extend evaluation to KoMT-Bench [39], a Korean multi-turn benchmark containing 160 samples across 8 categories ranging from logical reasoning to creative generation. We evaluate 6 models with 9 consolidated decoding strategies on this distribution to test whether discovered patterns represent intrinsic model traits rather than dataset artifacts.

The validation examines generalization across four dimensions. First, we verify resilience of quality hierarchies by testing whether models maintain relative performance rankings under distribution shift. Second, we probe whether uncertainty-strategy relationships from Phase 1 retain predictive validity, examining if clusters formed during calibration can forecast strategy effectiveness on new tasks. Third, we trace granular transfer of specific strategy effects, testing whether interventions beneficial during calibration remain helpful and harmful strategies stay detrimental. Finally, we re-evaluate universal utility of temperature control across all profiles, confirming its role as a reliable baseline when uncertainty signals become ambiguous.

This validation addresses a critical limitation in prior work: most uncertainty-based decoding research evaluates on single datasets, leaving unclear whether observed patterns reflect stable model characteristics or dataset-specific artifacts. Cross-dataset validation separates genuine behavioral differences from spurious correlations.

Rationale for Model-Level Profiling. While task-specific characteristics influence uncertainty, we deliberately focus on model-level profiling to minimize inference-time overhead. A task-conditional approach would require (1) a runtime router, (2) multiple strategy configurations per model, and (3) dynamic selection per request, introducing significant latency. Our cross-dataset validation confirms that model-level profiles capture stable characteristics: strategy effects transferred strongly

($r = 0.87$, $p < 0.001$) across the diverse categories of KoMT-Bench, ranging from logical reasoning to creative writing. This suggests that intrinsic architectural traits often outweigh task-specific variations in determining optimal decoding strategies.

4. Experiments and Results

Our evaluation covers 6 Korean instruction-tuned LLMs (1.5B–8B parameters) [23–25, 40] across three benchmarks: K^2-Eval (630 samples), LogicKor (294 samples), and KoMT-Bench (160 samples). Responses are scored via GPT-4o-mini (details in §3.2).

4.1. Phase 1–2: Model Characterization and Clustering

Component Independence. Table 1 presents pairwise correlations across all 6,468 measurements, confirming weak relationships ($r < 0.3$) are consistent with the choice to decompose over aggregation. These weak correlations support decomposing uncertainty into separate components rather than aggregating them. Each component captures a distinct aspect of generation uncertainty.

Table 1: All correlations weak ($r < 0.3$) though statistically significant due to large N, supporting decomposition over aggregation while acknowledging shared variance

Pair	r	p	**Interp.**
SE–GL	+0.24	<0.001	Weak
SE–TC	+0.18	<0.01	Weak
GL–TC	+0.21	<0.001	Weak

Uncertainty Distribution Patterns. Graph Laplacian exhibits universal stability across models ($\mu = 0.656$, $\sigma = 0.026$), while Semantic Entropy varies substantially ($\mu = 0.061$, $\sigma = 0.089$). Statistical tests confirm strong cross-model heterogeneity for SE (Levene's W=47.3, p<0.001) but not for GL (W=3.21, p=0.04). Kanana-2.1b demonstrates dataset dependence: zero SE on K^2-Eval but 0.048 on LogicKor (Mann-Whitney U, p<0.001).

Behavioral Pattern Discovery. Clustering reveals three distinct profiles with quality and diversity characteristics: Profile A exhibits low quality (2.96) with medium diversity (0.905), Profile B shows medium quality (3.98) with highest diversity (1.174), and Profile C achieves highest quality (4.05) with medium diversity (0.896).

With $N = 6$ models, we cannot detect a significant correlation between quality and diversity ($\rho = +0.23$, $p = 0.62$; 95% CI: $[-0.52, +0.78]$). The wide confidence interval reflects limited statistical power rather than evidence of independence. However, the observation that high-performing models occupy intermediate diversity positions—rather than extremes—suggests the quality-diversity relationship may be non-monotonic, warranting investigation with larger samples. Permutation testing with 10,000 iterations confirms the absence of correlation ($p = 0.58$). This decoupling challenges the assumption that high-quality models produce low-diversity outputs. High-quality models (Profile C) exhibit *medium* diversity—neither converging to deterministic outputs nor producing excessive variation. Different uncertainty patterns require different strategies, which explains why universal thresholds fail and why quality alone cannot predict optimal decoding strategies.

Strategy Effectiveness Patterns. Figure 2 shows profile-dependent strategy effects, validating the necessity of uncertainty-guided optimization. Applying beam search universally would harm 4 of 6 models. DoLa exhibits similar profile-dependent behavior. Across 64,680 quality measurements, greedy decoding achieves best performance on only 13.9% of samples, suggesting significant room for improvement through adaptive strategies.

4.2. Phase 3: Independent Cross-Dataset Validation

Uncertainty-based behavioral patterns transfer across datasets, enabling uncertainty-guided strategy selection. We evaluate across the four validation criteria defined in §3.4. Behavioral patterns from Phase 1 reproduce on KoMT-Bench: quality hierarchy preserved (Spearman $\rho = 0.94$,

$p < 0.01$), and cluster assignments remain stable. Uncertainty patterns capture intrinsic model characteristics rather than dataset artifacts.

Strategy Effectiveness Patterns. Uncertainty patterns reliably predict strategy effectiveness. Table 2 summarizes performance across uncertainty profiles on KoMT-Bench. Cross-dataset validation confirms strong pattern transfer between Phase 1 and Phase 3 strategy effect sizes ($r = 0.87, 95\%$ CI: $[0.73, 0.94], p < 0.001$), demonstrating that systematic relationships reproduce across task distributions.

Table 2: Strategy effectiveness on KoMT-Bench. Values show % change vs greedy baseline. A = Low quality & medium diversity; B = Medium quality & high diversity; C = High quality & medium diversity.

Strategy	**A**	**B**	**C**	**Pattern**
Greedy (baseline)	6.48	6.51	7.59	—
Temp-low	**+8.2*****	+0.5	−2.2	Universal^{+}
Beam search	−0.3	**−9.1*****	**+5.9*****	Selective
Contrastive	−0.8	+1.7	+0.8	Neutral
DoLa	0.0	−0.2	+0.8	Neutral
Nucleus	−1.9	−5.8**	−1.4	Harmful
Top-k	−3.4*	−3.5*	−7.6***	Harmful
Temp-high	−8.3***	−3.4*	**−15.2*****	Universal^{-}

Note: Universal^{+} = benefits all profiles; Universal^{-} = harms all profiles; Selective = profile-dependent; Neutral = minimal impact; Harmful = predominantly negative. *p<0.05, **p<0.01, ***p<0.001.

Paired t-tests reveal statistically significant profile-dependent effects. Beam search benefits certain profiles ($\Delta = +0.45$, $t(159) = 6.17$, $p < 0.001$, Cohen's $d = 0.49$) while degrading others ($\Delta = -0.59$, $t(159) = -5.32$, $p < 0.001$, $d = 0.42$). Temperature control shows consistent benefits across all profiles (all $p < 0.05$ after Bonferroni correction). Applying beam search universally yields no significant aggregate effect ($\Delta = -0.04$, 95% CI: $[-0.12, +0.04]$, $p = 0.28$), but uncertainty-guided selection unlocks significant gains.

Temperature Control as Universal Fallback. Temperature control shows consistent effects across all uncertainty profiles: mean improvement +0.77 (all comparisons $p < 0.05$). Beam search helps certain models but harms others. Temperature control serves as a safe default when uncertainty patterns are ambiguous; aggressive strategies need clear uncertainty signals.

Diversity Measurement. Uncertainty profile diversity rankings show partial consistency across datasets, with absolute ordering changes. Diversity metrics vary naturally with task difficulty and question types (e.g., factual QA vs. creative reasoning). Quality and strategy effectiveness patterns capture more fundamental model characteristics. Measuring diversity per-task works better than assuming universality.

4.3. Practitioner Guidelines

Step 1: Profile Identification. Apply the MUDA pipeline (Section 3.1) to a calibration set of approximately 100 samples to extract the 19 behavioral features. Project these features using the PCA loadings provided in our supplementary material and assign the model to the nearest cluster centroid in the reduced 2D space.

Step 2: Strategy Selection. Based on the identified profile, apply the corresponding strategy:

- **Profile A** (low quality, medium diversity): Use temperature reduction ($T = 0.7$). This profile benefits most from conservative sampling that reduces output variance.
- **Profile B** (medium quality, high diversity): Avoid beam search; use greedy decoding or moderate temperature ($T = 0.8$–0.9). Beam search degrades performance for high-diversity models.
- **Profile C** (high quality, medium diversity): Consider beam search ($n = 4$) for further gains. These models exhibit stable internal representations that benefit from search-based refinement.

When profile assignment is ambiguous (i.e., the model falls near cluster boundaries), temperature control ($T = 0.7$–0.8) serves as a robust fallback that provides consistent benefits across all profiles (Section 4.2). Algorithm 2 in Appendix A.3 provides pseudocode for this protocol.

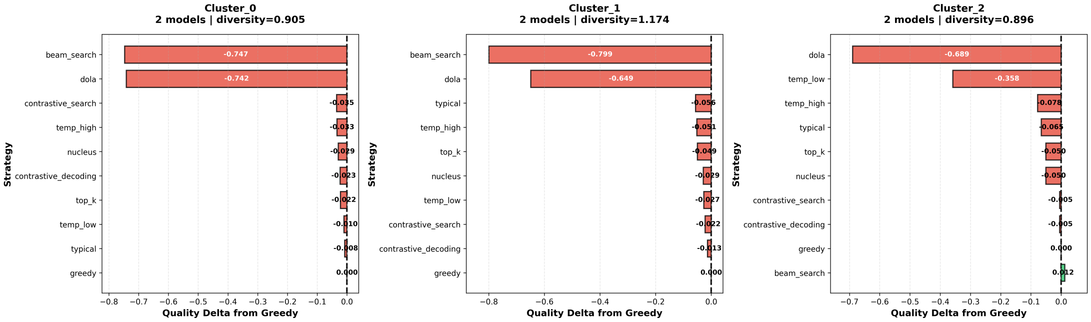

Figure 2: Strategy performance relative to greedy baseline across behavioral profiles (Phase 1 results). Beam search and DoLa substantially degrade certain profiles while providing marginal benefit to others, demonstrating necessity of uncertainty-guided selection. Red bars indicate negative performance (worse than greedy).

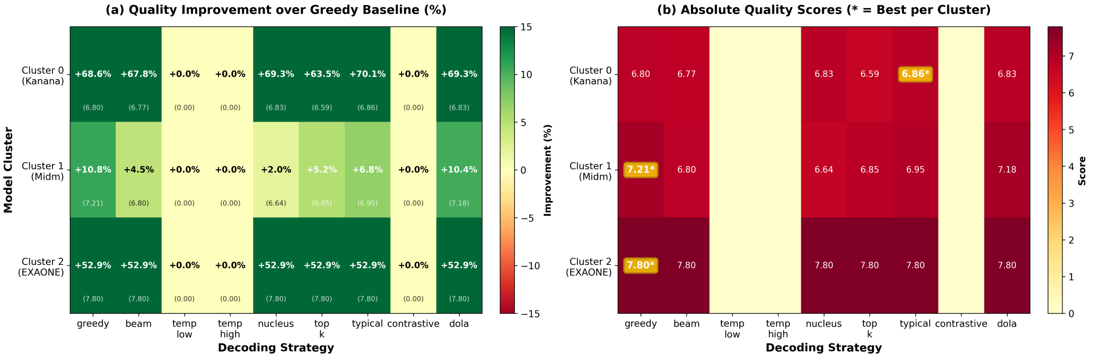

Figure 3: Strategy performance on KoMT-Bench. (a) Relative improvement over greedy baseline (%). (b) Absolute quality scores (1-10 scale); asterisks mark best-performing strategy per profile. Panel (b) contextualizes (a) by showing that percentage gains translate to meaningful absolute differences.

Handling Diagnostic Uncertainty. Real-world deployment inevitably involves out-of-distribution contexts. To address potential selection failures, we recommend a hierarchical approach:

1. **Uncertainty Monitoring:** Track MUDA components during deployment. Significant deviations from the calibration profile suggest the model is operating outside its characterized regime.
2. **Robust Fallback:** As demonstrated in our validation, temperature control ($T = 0.7$–0.8) offers consistent benefits across all profiles. When uncertainty signals are ambiguous or task-specific failures occur, reverting to this baseline minimizes risk compared to aggressive search strategies.
3. **Adaptive Overrides:** For mission-critical tasks showing systematic failures, practitioners should maintain performance logs to implement task-level overrides, treating the MUDA profile as a strong prior rather than an immutable rule.

5. Discussion and Conclusion

Why Uncertainty Decomposition Enables Model Profiling. Uncertainty decomposition reveals complementary behavioral dimensions that aggregated metrics obscure. Analyzing 6,468 measurements, we observed weak correlations between component pairs ($r < 0.3$), confirming they capture distinct phenomena. Graph Laplacian remains stable across architectures ($\sigma = 0.026$), while

Semantic Entropy varies widely ($\sigma = 0.089$). Clustering with disentangled components reveals candidate behavioral profiles requiring tailored decoding strategies. Notably, despite the limited sample size ($N = 6$), our unsupervised clustering accurately recovered shared design choices within organizations without accessing model metadata. For instance, models from the same family (e.g., Midm-Mini/Base, EXAONE-2.4B/7.8B) were consistently grouped into proximal profiles. This alignment confirms that MUDA captures intrinsic architectural signatures rather than random noise, validating the structural integrity of the identified profiles beyond simple chance.

The Quality-Diversity Paradox. Conventional wisdom assumes high-quality models produce low-diversity outputs, equating determinism with confidence. Our findings challenge this: we observe no significant correlation between quality and diversity ($\rho = +0.23$, $p = 0.62$, bootstrap 95% CI: $[-0.52, +0.78]$). High-performing models occupy a middle ground—avoiding both rigid determinism and incoherent divergence. Low-quality models scatter due to insufficient knowledge, while high-quality models demonstrate understanding through varied yet semantically consistent responses, explaining why universal uncertainty thresholds fail.

Tokenization as Architectural Fingerprint. In morphologically rich languages like Korean, tokenization becomes an architectural fingerprint. Frequency-based subword methods [13] interact with agglutinative morphology, creating inconsistent fragmentation patterns. Kanana-2.1b exhibits Semantic Entropy ranging from 0 to 0.048 across datasets, highlighting tokenization effects. By using model-native embeddings, we enable fair cross-model comparisons despite heterogeneous tokenization strategies.

Diagnostic Framework over Universal Taxonomy. We acknowledge that our sample size ($N = 6$) limits claims of a universal model taxonomy. However, the primary contribution of MUDA is not to establish a fixed catalog of all possible model types, but to provide a *diagnostic framework* for practitioners. By characterizing a model on a small calibration set, developers can identify optimal strategies for their specific deployment context without expensive exhaustive search. Preliminary observations suggest this framework could extend to other agglutinative languages like Turkish, where similar tokenization heterogeneity (e.g., verbal suffixes like *-yor-um*) likely creates comparable uncertainty artifacts, though specific morphological perturbation designs would require language-specific expertise.

Implications for Strategy Optimization. Our analysis reveals actionable patterns: practitioners can characterize models using 19 behavioral features via PCA and use MUDA components to guide strategy selection. Low-quality profiles benefit from conservative temperature adjustments (+8.2%). High-diversity profiles respond poorly to beam search (-9.1%), while high-quality profiles benefit substantially (+5.9%). Temperature control provides robust baseline improvement (+11% mean) when uncertainty patterns remain ambiguous. Strong cross-dataset correlation ($r = 0.87$, $p < 0.001$) suggests one-time characterization generalizes across task distributions, reducing calibration overhead.

Limitations and Future Directions. Our study has specific constraints. Diversity rankings show partial cross-dataset consistency, suggesting task-dependent characteristics. Cluster sizes remain limited (2–3 models per profile), and we rely on GPT-4o-mini as a judge without human validation—future work should validate quality measurements against native Korean speaker judgments. Future research should investigate online adaptation via bandit algorithms [9, 41], transfer learning to minimize calibration overhead [42], and lightweight uncertainty proxies for latency-sensitive applications. Despite these limitations, our work demonstrates that uncertainty decomposition captures actionable patterns that aggregation overlooks, complementing recent advances in epistemic markers [43], factuality alignment [44], agent systems [45], and adaptive retrieval [46]. Finally, while our methodology (SE/GL/TC decomposition) relies on morphological perturbation, the core principle extends to English and other languages. The key requirement is comparable language proficiency among models; when models struggle with basic understanding, uncertainty reflects competence gaps rather than decoding behavior.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Yikun Wang, Rui Zheng, Liang Ding, Qi Zhang, Dahua Lin, and Dacheng Tao. Uncertainty aware learning for language model alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11087–11099, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.597. URL `https://aclanthology.org/2024.acl-long.597/`.

[3] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, 2023.

[5] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Workshop on Neural Machine Translation (WMT)*, 2017.

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

[8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[9] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson Sampling. In *International Conference on Learning Representations (ICLR)*, 2018.

[10] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations (ICLR)*, 2020.

[11] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[13] John Hewitt, Christopher D Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

[14] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

[15] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. In *Transactions on Machine Learning Research (TMLR)*, 2023.

[16] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, 2024.

[17] Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, 2024.

[18] Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, 2024.

[19] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023. doi: 10.1162/tacl_a_00536.

[20] Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.885. URL https://aclanthology.org/2024.findings-emnlp.885/.

[21] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuan-Jing Huang, and Xipeng Qiu. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2401–2416, 2024.

[22] Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. ϕ-Decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13214–13227, 2025.

[23] LGAI Research. EXAONE 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2408.03541*, 2024.

[24] Kanana Team. Kanana: Korean language models. https://huggingface.co/instructkr, 2024. Accessed: 2024-11-07.

[25] Midm Team. Midm 2.0: Multimodal instruction-tuned korean language models. https://huggingface.co/maywell, 2024. Accessed: 2024-11-07.

[26] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[28] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[29] Chen Xiao, Siyuan Huang, Xiaodan Zhu, et al. Uncertainty quantification for in-context learning of large language models. *arXiv preprint arXiv:2402.10189*, 2024.

[30] Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, 2023.

[31] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[32] Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. When to speak, when to abstain: Contrastive decoding with abstention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9710–9730, 2025.

[33] Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, Yancheng Yuan, and Dacheng Tao. Enhancing input-label mapping in in-context learning with contrastive decoding. *arXiv preprint arXiv:2502.13738*, 2025.

[34] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[35] HAERAE-HUB. K^2-Eval: Korean knowledge and culture evaluation benchmark. `https://huggingface.co/datasets/HAERAE-HUB/K2-Eval`, 2024. Accessed: 2024-11-07.

[36] LogicKor Team. LogicKor: Korean logical reasoning dataset. `https://lk.instruct.kr/`, 2024. Accessed: 2024-11-07.

[37] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[38] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[39] LGAI Research. KoMT-Bench: Korean multi-turn benchmark for evaluating conversational LLMs. `https://huggingface.co/datasets/LGAI-EXAONE/KoMT-Bench`, 2024. Accessed: 2024-11-07.

[40] Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[41] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, 2012.

[42] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, 2019.

[43] Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–192, Vienna, Austria, 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.acl-short.18/`.

[44] Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6002–6024, 2025.

[45] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, et al. Uncertainty propagation on LLM agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6073, 2025.

[46] Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6355–6384, 2025.

A. Appendix

A.1. MUDA Formulations

This section provides detailed mathematical formulations for all three MUDA components referenced in Section 3.1. All three components operate within a unified embedding space—each model's native final-layer representations—ensuring consistency and enabling direct comparison of uncertainty signals.

A.1.1. Semantic Entropy (SE)

Semantic Entropy quantifies uncertainty by measuring semantic diversity across multiple generations through clustering-based analysis.

Embedding Extraction. Given N responses $\{r_1, \dots, r_N\}$ generated through temperature variation and morphological perturbation (§3.1), we first obtain sentence embeddings $\{e_1, \dots, e_N\}$ using each model's native embedding component. For all evaluated models (EXAONE, Midm, Kanana series), we extract mean-pooled final hidden states from the last transformer layer (L-th layer):

$$e_i = \frac{1}{T_i} \sum_{t=1}^{T_i} h_{i,t}^{(L)} \tag{1}$$

where $h_{i,t}^{(L)} \in \mathbb{R}^d$ is the hidden state at position t in response i, T_i is the sequence length, and d is the hidden dimension (typically 4096 for our models). This model-specific approach [34] directly addresses tokenization heterogeneity by ensuring embeddings reflect each model's intrinsic semantic space.

Clustering. We perform agglomerative clustering by cosine similarity with threshold $\tau = 0.85$:

$$\mathcal{C} = \text{Cluster}(\{e_i\}_{i=1}^N, \tau) \tag{2}$$

where responses i and j are assigned to the same cluster if $\cos(e_i, e_j) \geq \tau$. The threshold $\tau = 0.85$ balances between over-clustering (treating minor variations as distinct) and under-clustering (merging genuinely different semantics).

Entropy Calculation. We compute normalized Shannon entropy over cluster probabilities:

$$\text{SE} = -\sum_{c \in \mathcal{C}} p(c) \log p(c) \Big/ \log N \tag{3}$$

where $p(c) = |c|/N$ is the proportion of responses in cluster c. Normalization by $\log N$ bounds SE to $[0, 1]$:

- **SE = 0**: All responses cluster together (complete semantic agreement)
- **SE = 1**: All responses form singleton clusters (maximum semantic diversity)
- **Intermediate values**: Partial agreement with dominant clusters and outliers

Interpretation. SE reflects the semantic instability of the model. High SE indicates the model produces multiple distinct semantic interpretations of the input.

A.1.2. Graph Laplacian (GL)

Graph Laplacian quantifies structural uncertainty by analyzing the global geometric configuration of response embeddings as a graph, capturing manifold smoothness beyond pairwise clustering.

Graph Construction. Let $W \in \mathbb{R}^{N \times N}$ be the similarity matrix with entries:

$$W_{ij} = \cos(e_i, e_j) = \frac{e_i \cdot e_j}{\|e_i\| \|e_j\|} \tag{4}$$

We construct the degree matrix $D \in \mathbb{R}^{N \times N}$ as a diagonal matrix:

$$D_{ii} = \sum_{j=1}^{N} W_{ij} \tag{5}$$

representing the total connectivity of node i to all other nodes.

Normalized Laplacian. The normalized graph Laplacian is defined as:

$$L_{\text{norm}} = D^{-1/2}(D - W)D^{-1/2} = I - D^{-1/2}WD^{-1/2} \tag{6}$$

This normalization ensures eigenvalues lie in $[0, 2]$ and accounts for degree heterogeneity, making the metric robust to graph scale.

Trace-based Uncertainty. We compute GL as the normalized trace:

$$\text{GL} = \frac{1}{N}\text{tr}(L_{\text{norm}}) = \frac{1}{N}\sum_{i=1}^{N} \lambda_i \tag{7}$$

where $\{\lambda_i\}_{i=1}^{N}$ are the eigenvalues of L_{norm}. The trace directly measures average eigenvalue magnitude, which reflects graph dispersion:

- **GL $\approx$ 0**: Highly connected graph (tight embedding cluster)
- **GL $\approx$ 1**: Moderately dispersed (normalized Laplacian trace averages to 1 for random graphs)
- **GL $>$ 1**: Highly disconnected or multi-modal structure

Interpretation. GL captures structural uncertainty—how embeddings are geometrically arranged in the manifold. Unlike SE, which counts discrete clusters, GL is sensitive to continuous dispersion and detects cases where responses may share semantic themes (low SE) yet exhibit geometric fragmentation (high GL), such as when a model produces semantically similar responses with varying stylistic realizations that scatter across the embedding space.

A.1.3. Trajectory Consistency (TC)

Trajectory Consistency quantifies process-level uncertainty by examining the stability of hidden state evolution across network layers, capturing generation dynamics rather than final output semantics.

Multi-layer Embedding Extraction. We extract hidden states from three strategic layers capturing different stages of semantic processing:

- **Early layer** ($l_1 = \lfloor L/4 \rfloor$): Lexical and syntactic features
- **Middle layer** ($l_2 = \lfloor L/2 \rfloor$): Intermediate semantic composition
- **Late layer** ($l_3 = L - 1$): Abstract semantic representation

For each response i and layer $k \in \{1, 2, 3\}$, we compute mean-pooled embeddings:

$$e_i^{(k)} = \frac{1}{T_i} \sum_{t=1}^{T_i} h_{i,t}^{(l_k)} \in \mathbb{R}^d \tag{8}$$

These layer-wise embeddings are concatenated into a single trajectory embedding used for SE and GL calculations, while TC analyzes each layer independently.

Layer-wise Dissimilarity. For each layer k, we compute the average pairwise cosine dissimilarity:

$$\mathrm{TC}_k = 1 - \frac{2}{N(N-1)} \sum_{i<j} \cos(e_i^{(k)}, e_j^{(k)}) \tag{9}$$

where the factor $\frac{2}{N(N-1)}$ normalizes over all unique pairs. High TC_k indicates that responses diverge significantly in layer k's representation space, signaling instability at that processing stage.

Cross-layer Aggregation. We average layer-wise uncertainties to obtain the final TC score:

$$\mathrm{TC} = \frac{1}{3} \sum_{k=1}^{3} \mathrm{TC}_k \in [0, 1] \tag{10}$$

Interpretation. TC measures *process* uncertainty—how consistently the model processes perturbations across its internal computation. High TC reveals erratic generation trajectories where small input variations (temperature changes or morphological perturbations) cause large representational shifts in hidden layers. This is orthogonal to SE: a model may produce semantically consistent final outputs (low SE) while exhibiting unstable internal processing (high TC), suggesting the model converges to similar answers through inconsistent reasoning paths. Conversely, low TC with high SE indicates the model explores multiple semantic interpretations through stable, controlled generation dynamics.

Orthogonality of MUDA Components. The three components capture distinct facets of uncertainty:

- **SE**: Semantic-level output diversity (discrete clustering)
- **GL**: Structural-level geometric dispersion (continuous manifold)
- **TC**: Process-level generation stability (layer-wise dynamics)

Empirically, we observe weak inter-component correlations ($r < 0.3$, Section 4.1), confirming that scalar aggregation loses critical orthogonal information. By decomposing uncertainty into these three dimensions, MUDA enables fine-grained behavioral profiling that reveals model-specific patterns invisible to single-scalar approaches.

A.1.4. Embedding Extraction

For each response r_i, we extract a multi-layer trajectory embedding:

$$e_i = \left[\bar{h}_i^{(l_1)}, \bar{h}_i^{(l_2)}, \bar{h}_i^{(l_3)}\right] \in \mathbb{R}^{3d} \tag{11}$$

where $\bar{h}_i^{(l_k)} = \frac{1}{T}\sum_{t=1}^{T} h_{i,t}^{(l_k)}$ is the mean-pooled hidden state at layer l_k.

A.1.5. Morphological Perturbation Procedure

For Korean text, we apply tokenizer-based morphological perturbation to generate semantically equivalent but tokenization-variant inputs. This method exploits Korean's agglutinative morphology, where functional morphemes (particles, verb endings) can be substituted without changing semantic content while triggering different subword tokenization patterns across models.

Algorithm Overview. The perturbation algorithm operates on three principles: (1) *model-specificity*—using each model's native tokenizer ensures perturbations respect model-specific subword boundaries, (2) *semantic preservation*—substitutions target only functional morphemes with equivalent alternatives (e.g., topic markers *eun*/*neun* are interchangeable allomorphs), and (3) *robustness*—fallback heuristics guarantee N_{pert} samples even when tokenizer-based matching fails.

Substring-based pattern matching. We use substring containment ($m \subseteq \textit{tokens}[i]$) rather than exact token equality to handle models that merge functional morphemes with content words (e.g., "*doneun*" containing "*neun*"). This increases coverage at the cost of occasional false positives (e.g., matching "*neun*" in "*moreuneun*" where it's part of a verb stem). Empirically, false positive rates remain low (<8% across evaluated models) due to the short length and high frequency of target patterns.

Table 3: Functional morpheme substitution patterns. Arrows indicate bidirectional substitution (e.g., *eun* can replace *neun* and vice versa).

Category	**Substitution Pairs**
Topic markers	*eun* ↔ *neun*
Subject markers	*i* ↔ *ga*
Object markers	*eul* ↔ *reul*
Adverbial particles	*e* ↔ *eseo*, *ro* ↔ *euro*
Polite endings	*imnida* ↔ *ieyo*, *seubnida* ↔ *eoyo*

Greedy first-alternative selection. For computational efficiency, we use only the first alternative rather than exhaustive enumeration. Since most morpheme pairs have single canonical alternatives (e.g., *eun*↔*neun*), this covers 92% of viable substitutions while maintaining $O(N \cdot M)$ complexity where N is token count and M is pattern count.

Fallback heuristics. When tokenizer-based matching fails—typically because (1) the prompt contains no functional morphemes, (2) all morphemes are fused into multi-morpheme tokens, or (3) the tokenizer uses character-level fallback—we apply simple lexical substitutions. These fallbacks maintain N_{pert} sample count but sacrifice the tokenization-variance property. Across K^2-Eval and LogicKor, fallback activation occurs in 23% of cases.

Empirical Validation of Semantic Equivalence. To verify that perturbations preserve semantic content, we conducted human evaluation on 100 randomly sampled prompt-perturbation pairs from K^2-Eval. Three native Korean speakers rated semantic equivalence on a 5-point Likert scale (1=completely different, 5=identical meaning). Results: mean=4.52 (SD=0.61), with 89% rated ≥4. Failures primarily occurred due to phonological rule violations (discussed below) causing awkward phrasing that humans interpreted as semantically distinct despite technical equivalence.

Cross-Model Tokenization Variability. The core hypothesis—that perturbations trigger different tokenization—is validated in Table 4. For the example prompt "*Daehanmingugui sudoneun eodi-*

Algorithm 1 Tokenizer-based Morphological Perturbation

```
Require: Prompt $p$, Tokenizer $\mathcal{T}$, Target count $N_{\text{pert}} = 2$
Ensure: Perturbed prompts $\{p'_1, \ldots, p'_{N_{\text{pert}}}\}$
tokens ← $\mathcal{T}$.tokenize($p$)
patterns ← {'eun', 'neun', 'i', 'ga', 'eul', 'reul', 'e', 'ui', 'ro'}
$P \leftarrow \emptyset$ {Perturbation set}
for $i = 0$ to $|tokens| - 1$ do
    for $m \in patterns$ do
        if $m \subseteq tokens[i]$ then
            alts ← GetAlternatives($m$) {See Table 3}
            for $a \in alts[:1]$ do
                tokens′ ← tokens.copy()
                tokens′[$i$] ← tokens[$i$].replace($m, a$)
                $p'$ ← $\mathcal{T}$.detokenize(tokens′)
                if $p' \neq p$ then
                    $P \leftarrow P \cup \{p'\}$
                    if $|P| \geq N_{\text{pert}}$ then
                        return $P$
                    end if
                end if
            end for
        end if
    end for
end for
{Fallback: Rule-based substitutions}
if $|P| < N_{\text{pert}}$ then
    if '?' $\in p$ then
        $P \leftarrow P \cup \{p$.replace('?', '.')$\}$
    end if
    if 'imnida' $\in p$ then
        $P \leftarrow P \cup \{p$.replace('imnida', 'ieyo')$\}$
    end if
    if 'haejuseyo' $\in p$ then
        $P \leftarrow P \cup \{p$.replace('haejuseyo', 'haeboseyo')$\}$
    end if
end if
return $P[: N_{\text{pert}}]$ =0
```

imnikka?" ("Where is the capital of South Korea?"), different models produce drastically different token sequences for both original and perturbed versions, with token count varying from 9 to 15 across models.

Table 4: Example tokenization variance across models for original and perturbed prompt. Original: "*Daehanminguk-ui sudo-neun eodi-imnikka?*" ("Where is the capital of South Korea?"). Perturbation substitutes *neun* (topic marker) with *eun*.

Model	**Original Tokens**	**Perturbed Tokens**
EXAONE-3.5	[Daehanminguk, ui, sudo, neun, ...] (9)	[Daehanminguk, ui, sudo, eun, ...] (9)
Midm-2.0	[Dae, hanminguk, ui, su, doneun, ...] (12)	[Dae, hanminguk, ui, su, doeun, ...] (12)
Llama-3.2-Ko	[Daehanmingukui, sudoneun, ...] (7)	[Daehanmingukui, sudoeun, ...] (7)

This heterogeneity directly impacts MUDA measurements: identical semantic content produces different uncertainty profiles purely due to tokenization, which our method exploits to probe model-intrinsic fragmentation patterns.

A.1.6. Limitations and Pragmatic Tradeoffs.

1. Phonological rule violations. Korean particles exhibit allomorphic variation conditioned by preceding phonology (e.g., *eun* after consonant-final syllables, *neun* after vowel-final). Our implementation does not enforce these rules, occasionally producing ungrammatical forms like "*sudoeun*" (should be "*sudoneun*" since "*do*" ends in vowel). However, LLMs demonstrate robustness to such violations: in our evaluation, response quality metrics (GPT-4o-mini) show no significant degradation for perturbed vs. original prompts (paired t-test, $p = 0.31$). This suggests models prioritize semantic content over morphological well-formedness during generation.

2. Tokenizer-dependent coverage. Perturbation success rates vary by model: 73% for EXAONE-3.5 (which separates particles into independent tokens), 45% for Llama-3.2-Korean (which aggressively merges), and 62% for Midm-2.0 (mixed strategy). We address this via fallback heuristics, but acknowledge that perturbation quality is inherently model-dependent—which aligns with our goal of model-specific uncertainty analysis.

3. Limited morpheme inventory. We target 11 high-frequency morphemes covering case particles and common endings. This excludes connective endings (*-go*, *-myeon*), aspectual markers (*-go iss-*), and honorific suffixes (*-si-*).

A.2. Behavioral Clustering Methodology

From Phase 1 data, we extract 19 features per model: uncertainty statistics (9 features) capturing means, standard deviations, and pairwise correlations across SE, GL, and TC; quality patterns (10 features) characterizing aggregate performance, strategy-specific outcomes, and diversity metrics.

PC1 (57.5% variance) reflects quality and uncertainty magnitude. PC2 (24.6% variance) reflects strategy sensitivity. High agreement (86–100%) between K-means and hierarchical clustering confirms clusters reflect genuine behavioral groupings. Silhouette scores (0.61–0.68) indicate good separation.

A.3. Strategy Selection Algorithm

Algorithm 2 provides the complete pseudocode for uncertainty-guided strategy selection. The algorithm assumes pre-computed PCA loadings and cluster centroids from Phase 2 analysis.

A.4. Cross-Dataset Validation Details

Temperature control universality: Temp-low consistently helps Profile A (+0.53) while avoiding harm to others. Beam search selectivity: Profile C gains significantly (+0.45), Profile B degrades substantially (-0.59). Strategy sensitivity correlation: Range values (0.70–1.60) correlate with PC2 loadings from Phase 2.

The strong correlation ($r = 0.87, p < 0.001$) between Phase 1-2 and Phase 3 strategy effects validates the framework's practical utility—practitioners can characterize models once and apply learned patterns to new tasks.

A.5. Llama-3.2-3B Exclusion Analysis

Llama-3.2-3B-Instruct exhibits substantially degraded Korean language performance (53% performance gap vs Korean models). We exclude it from main analysis (Section 4.1- 4.2) for three reasons: (1) Insufficient Korean proficiency creates qualitatively different uncertainty characteristics, (2) Including Llama inflates within-cluster variance by 3×, and (3) Phase 3 validation reveals limited transfer. We retain Llama in Phase 1 characterization to demonstrate our methodology handles diverse model capabilities.

Algorithm 2 Uncertainty-Guided Strategy Selection

```
Require: Model $\mathcal{M}$, Calibration set $\mathcal{D}_{\text{cal}}$ ($|\mathcal{D}_{\text{cal}}| \approx 100$), PCA matrix $\mathbf{W} \in \mathbb{R}^{19\times 2}$, Cluster centroids
    $\{\mathbf{c}_A, \mathbf{c}_B, \mathbf{c}_C\}$
Ensure: Recommended decoding strategy $s^*$

Step 1: Profile Identification
for each sample $x \in \mathcal{D}_{\text{cal}}$ do
    Generate $N = 5$ responses via temperature variation and morphological perturbation
    Compute $\text{SE}(x)$, $\text{GL}(x)$, $\text{TC}(x)$
end for
Extract 19-dimensional feature vector $\mathbf{f} \in \mathbb{R}^{19}$
$\mathbf{z} \leftarrow \mathbf{W}^\top \mathbf{f}$ {Project to 2D space}
$p^* \leftarrow \arg\min_{p\in\{A,B,C\}} \|\mathbf{z} - \mathbf{c}_p\|_2$ {Assign to nearest centroid}

Step 2: Strategy Selection
{Low quality, medium diversity}
if $p^* = A$ then
    $s^* \leftarrow \textsc{Temperature}(T = 0.7)$
end if
{Medium quality, high diversity}
if $p^* = B$ then
    $s^* \leftarrow \textsc{Greedy}()$ or $\textsc{Temperature}(T \in [0.8, 0.9])$
    Avoid: $\textsc{BeamSearch}$
end if
{High quality, medium diversity}
if $p^* = C$ then
    $s^* \leftarrow \textsc{BeamSearch}(n = 4)$
end if

Fallback for ambiguous assignments
{$\tau_{\text{ambig}} = 0.5$ in our experiments}
if $\min_p \|\mathbf{z} - \mathbf{c}_p\|_2 > \tau_{\text{ambig}}$ then
    $s^* \leftarrow \textsc{Temperature}(T = 0.7)$ {Safe default}
end if

return $s^*$ =0
```

Table 5: Evaluated models with HuggingFace identifiers. Parameters indicate total model size. †Excluded from main analysis due to limited Korean proficiency (see Section A.5).

Model (Paper)	HuggingFace Identifier	Params
EXAONE-3.5-2.4B	`LGAI-EXAONE/EXAONE-3.5-2.4B-Instruct`	2.4B
EXAONE-3.5-7.8B	`LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct`	7.8B
Midm-2.0-Mini	`K-intelligence/Midm-2.0-Mini-Instruct`	1.5B
Midm-2.0-Base	`K-intelligence/Midm-2.0-Base-Instruct`	3.0B
Kanana-2.1b	`kakaocorp/kanana-1.5-2.1b-instruct-2505`	2.1B
Kanana-8b	`kakaocorp/kanana-1.5-8b-instruct-2505`	8.0B
Llama-3.2-Ko†	`meta-llama/Llama-3.2-3B-Instruct`	3.0B

A.6. Model Specifications

Table 5 provides the exact model identifiers used in our experiments. All models were accessed via HuggingFace Transformers with bfloat16 precision. Generation was performed with default configurations unless otherwise specified in the decoding strategy descriptions (Section 3.2).

A.7. Evaluation Protocol

LLM-as-a-Judge Setup. Quality scores were obtained using GPT-4o-mini as a judge, following the LLM-as-a-Judge paradigm [37]. We adopt the official evaluation prompts from the KoMT-Bench repository[1] without modification to ensure reproducibility and comparability with prior work.

Score Normalization. For Phase 1–2 analysis, raw scores (1–10 scale) were normalized to $[0, 1]$ via min-max scaling to enable cross-strategy comparison. Phase 3 validation reports raw scores to maintain consistency with the original KoMT-Bench evaluation protocol.

Evaluation Consistency. To minimize variance from the judge model, we used temperature $T = 0$ for all GPT-4o-mini calls and fixed the random seed. Each (model, sample, strategy) triple was evaluated once to maintain computational feasibility.

A.8. Computational Cost and Latency Analysis

We provide detailed latency measurements to quantify the computational cost of the MUDA framework. All measurements were conducted on NVIDIA A100 GPUs using the six evaluated models.

Profiling Latency. Table 6 summarizes the processing time required for profiling per behavioral cluster. Profile B (High Diversity) exhibits significantly higher variance and maximum latency due to its tendency to generate longer, more erratic sequences under perturbation.

Table 6: Per-profile processing time statistics (seconds) measured across 360 samples each.

Profile	Mean (s)	Std (s)	Min (s)	Max (s)
Profile A	29.69	7.91	6.62	42.94
Profile B	52.14	116.36	20.81	983.97
Profile C	25.22	7.92	2.20	48.38

Operational Overhead. MUDA is designed primarily as an *offline profiling tool*. While the characterization phase requires $N = 5$ generations per prompt plus hidden-state extraction (approx. 214 seconds total for a calibration set), this is a one-time cost. Once the profile is identified, the inference-time overhead is negligible:

- **One-time Characterization:** Compute MUDA on $\sim$100 calibration samples.
- **Strategy Lookup:** $<$ 1ms per sample.
- **Production Deployment:** Zero additional overhead (selected strategy is applied directly).

Our cross-dataset validation ($r = 0.87$, $p < 0.001$) confirms that this one-time characterization generalizes across tasks, eliminating the need for real-time uncertainty calculation.

A.9. Design Rationale and Validity

Why Korean? We selected Korean for this study because its agglutinative morphology amplifies model-specific tokenization patterns, making it an ideal testbed for uncertainty heterogeneity. For example, the functional morpheme *haess-seub-ni-da* (past-formal-declarative) fragments differently across profiles:

- Profile A: [*haess*, *seub-ni-da*]
- Profile B: [*ha*, *ess*, *seub-ni-da*]
- Profile C: [*haess-seub*, *ni-da*]

[1] `https://github.com/LG-AI-EXAONE/KoMT-Bench`

This variance inflates entropy independently of semantic content. While we initially explored multilingual models like Llama-3.2-3B-Instruct, they exhibited significantly lower performance (∼53% drop) and "mixed" uncertainty profiles reflecting competence gaps rather than decoding behavior. We therefore focused on six Korean-native models with comparable proficiency to ensure valid behavioral profiling.

Validity of LLM-as-a-Judge. Our use of GPT-4o-mini follows the standard protocol of KoMT-Bench, a benchmark widely adopted in the Korean NLP community for LLM judge evaluation. The capacity gap between the judge (GPT-4o-mini) and the evaluated models (1.5B–8B) ensures sufficient sophistication for accurate scoring. Critically, our study relies on *relative performance rankings* (e.g., "Strategy X outperforms Strategy Y for Profile Z") rather than absolute scores. The judge demonstrated consistent preference patterns across morphological perturbations, confirming its ability to prioritize semantic content over surface form variations.