Narrative Landscape: Mapping Narrative Dispositions Across LLMs

Donghoon Jung **Jiwoo Choi*** **Songeun Chae*** **Seohyon Jung†**
School of Digital Humanities and Computational Social Sciences, KAIST, South Korea
{donghoon.jung, jwchoi0515, songeun, seohyon.jung}@kaist.ac.kr

Abstract

This study proposes a quantitative framework for profiling LLM dispositions as stable, model-specific regularities in output under repeated, controlled elicitation. Using a structured narrative constraint-selection task administered across six frontier models and three instruction types, we operationalize disposition through two dimensions: "consistency", measured as cross-replication selection overlap via Jaccard similarity, and "diversity", measured as dispersion across options via the inverse Simpson index. We further introduce Narrative Landscape, a PCA-based visualization that maps each model's selection profile into a shared space for direct comparison. Results reveal a clear rigidity–exploration spectrum across model families and show that instruction types shift the geometry of selection spaces even when scalar metrics appear similar, indicating that comparable scores can mask qualitatively distinct selection topologies.

1 Introduction

During the development of Claude Opus 4.5, researchers reportedly used an internal "soul document" to specify the model's persona, values, and interaction style (Weiss, 2025). Anthropic (2024) characterizes this broader practice as character training, which aims to instill more nuanced traits such as curiosity and open-mindedness into the model's responses through design choices beyond standard instruction following. As large language models (LLMs) are now widely used across domains including narrative writing, scientific writing, and ideation, their engineered dispositions become consequential for the kinds of outputs they systematically produce. This study argues for the need to identify such dispositions as empirically observable regularities.

Recent empirical work uses replication-based elicitation to probe disposition stability across personality inventories (Serapio-García et al., 2023), political preference, and moral robustness and susceptibility under persona role-play (Rozado, 2024; Costa et al., 2025), and large-scale comparisons of personality versus political profiles (Goyanes et al., 2025), alongside measurement refinements via open-ended, AI-rated Big Five assessment (Zheng et al., 2025). Yet such profiles remain fragile with respect to test–retest reliability and paraphrase variation (Wang et al., 2025) and exhibit limited behavioral coherence, including word–deed inconsistency and weak cross-aspect transfer (Xu et al., 2025); moreover, they can be stabilized at the level of self-report without commensurate behavioral change (Han et al., 2025). Despite these measurement advances, prior work lacks a shared-space visualization that exposes model- and instruction-level structure in repeated selections.

This study introduces a replication-based framework for profiling LLM disposition as model-specific regularities in narrative constraint selection under controlled elicitation (Jung et al., 2025). With a fixed pool of narrative constraints and repeated replications, overlap in selected constraints captures the stability of commitments, operationalized as consistency via Jaccard similarity (Broder, 1997), while dispersion of selections across the pool captures coverage of the decision space, operationalized as diversity via the inverse Simpson index (Hill, 1973). To complement these scalar summaries and expose differences in selection structure, we also embed constraint-frequency profiles into a shared Principal Component Analysis (PCA) space for direct comparison across models and instruction types.

Results show a clear spectrum of narrative dispositions across model families, from rigid, high-overlap constraint selections with limited option-space coverage to low-overlap selections with

*Co-second authors.
†Corresponding author.

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 24–30

markedly higher diversity; instruction types further modulate the extent of the selection space. Narrative Landscape also indicates that comparable scalar metrics can correspond to distinct selection topologies, representing model- and instruction-level differences as geometric structure in a shared constraint space. The measurement framework and Narrative Landscape generalize beyond narrative, providing a transferable approach to profiling disposition in structured selection tasks.

2 Data

Model Selection and Constraint Pool. Building on Jung et al. (2025), we analyze six state-of-the-art LLMs spanning major provider families (OpenAI, Anthropic, Google, Alibaba). To maximize replicability across repeated runs, decoding parameters were held constant where available; temperature (temp) and top-p were fixed at 1.0 for all models, and vendor-specific controls (reasoning effort, verbosity) were set to high when present. Model identifiers and decoding settings are summarized in Appendix A.

All models were presented with the same constraint pool, which functions as a theory-grounded diagnostic taxonomy for making narrative preferences observable as authorial choices. The pool contains 200 narrative constraints systematically distributed across four narratological elements—Event, Style, Character, Setting—with each element subdivided into five categories of ten constraints each. To minimize surface-level selection bias and ensure that observed differences reflect model dispositions rather than prompt artifacts, each constraint is designed with structural regularity (uniform word length, parallel grammatical structure) and matched conceptual granularity within categories. The models operate on the natural-language constraints themselves, allowing selection patterns to be interpreted as behavioral traces of a model's implicit narrative logic.

Experimental Procedure. Each observation in our dataset corresponds to a single run in which a model produces a narrative plan by selecting constraints and generating justifications under a controlled instruction type. We operationalize "authorial orientations" through three broad instruction types—Basic, Quality-focused, and Creativity-focused—intended to capture stance rather than dependence on specific phrasing. Across runs, the persona is set via the instruction, and the user prompt instructs the model to read the full constraint pool, select a fixed number of constraints, justify each selection, and then evaluate cross-constraint dynamics in a final compatibility assessment.

In this paper, we focus on a pooled, unlabeled constraint-selection setting with a fixed selection budget. Models are presented with the full pool of 200 constraints without element labels and are required to select exactly 20 constraints they deem most useful for planning a single fictional narrative, providing a brief justification for each selection. The full prompt templates and the complete constraint pool are documented in Jung et al. (2025).

To enable rigorous replication and to control for order effects (Liu et al., 2024; Pezeshkpour and Hruschka, 2024; Shi et al., 2025), each run uses (i) a fresh random permutation of the constraint list and (ii) an isolated session state. Within any fixed experimental cell (model × instruction type, in our setting), only the constraint-order permutation and provider stochasticity vary; all other factors are held constant. We execute 160 independent replications per cell (2,880 runs total), producing dense samples for stability and sensitivity analyses.

3 Consistency and Diversity

To measure the consistency of constraint selection across repeated runs within the same model–instruction type cell, we use Jaccard similarity (J) (Broder, 1997):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Here, A and B are the constraint sets selected in two runs. For each model–instruction type cell, we compute Jaccard similarity over all run pairs and report the mean pairwise value as the consistency score.

To measure diversity, we use the Gini–Simpson index (GS) (Simpson, 1949), which corresponds to the probability that two randomly selected constraints are different. Because the Gini–Simpson index is bounded in $[0, 1]$ and can compress differences, we also report the inverse Simpson index in its effective-number form (EN) (Hill, 1973):

$$GS = 1 - \sum_k p_k^2, \qquad EN = \frac{1}{\sum_k p_k^2}$$

In both expressions, p_k is the proportion of selections assigned to constraint k. A higher effective number indicates that the model draws from a broader set of constraints more evenly.

Table 1: Narrative consistency and diversity by model and instruction type

Model	Instruction Type	Jaccard Similarity	Effective Number	Unique Constraints
gpt5	Quality-focused	0.3169	41.7581	113
gpt5	Basic	0.3070	42.7947	123
gemini	Creativity-focused	0.2597	48.7336	142
gpt5	Creativity-focused	0.2546	49.5021	128
gpt4.1	Creativity-focused	0.2458	51.1520	188
o4mini	Quality-focused	0.2420	51.7731	143
o4mini	Basic	0.2232	55.2856	159
claude	Creativity-focused	0.2166	56.5602	184
claude	Quality-focused	0.2160	56.7037	185
gemini	Basic	0.2006	60.2034	174
o4mini	Creativity-focused	0.2004	60.3261	165
gemini	Quality-focused	0.1936	61.9600	174
claude	Basic	0.1860	64.2425	196
gpt4.1	Quality-focused	0.1483	77.0911	199
gpt4.1	Basic	0.1344	83.9162	200
qwen	Quality-focused	0.0896	120.1963	200
qwen	Creativity-focused	0.0867	123.9079	200
qwen	Basic	0.0820	129.9954	200

Table 1 shows the means of Jaccard similarity and the effective number across six models and three instruction types. The results highlight a distinct spectrum of narrative behaviors. gpt5 demonstrates the most rigid narrative disposition. In the Quality-focused instruction type, it exhibits the highest consistency ($J = 0.3169$) and the lowest diversity ($EN = 41.7581$), utilizing only 113 unique constraints. This indicates a strong adherence to a focused set of narrative strategies. In contrast, qwen displays the opposite extreme. Across model–instruction type cells, qwen shows the lowest consistency ($J < 0.09$) and the highest diversity ($EN > 120$), using the maximum observed number of unique constraints (200). Other models such as gemini, o4mini, claude, and gpt4.1 occupy an intermediate position, balancing stability and exploration. These divergences suggest a rigidity–exploration spectrum in which each model reflects a structurally distinct selection logic rather than a variation along a single, shared trait.

In particular, while instruction types, such as Creativity-focused versus Quality-focused, induce relatively small fluctuations within models, between-model differences are typically larger than instruction-induced differences in narrative profiles under our experimental setup. Although prompt-level differences are generally smaller than between-model differences, for gpt5 the Creativity-focused is notably distinct: its difference from Basic ($\Delta J = 0.0524$) and from Quality-focused ($\Delta J = 0.0623$) both exceed the Basic–Quality-focused difference ($\Delta J = 0.0099$), and even some between-model gaps such as that between o4mini and claude.

4 Narrative Landscape

Despite the fact that consistency and diversity differentiate models, these scalar summaries provide only a partial view of disposition. Moreover, prior work can establish statistically reliable differences across models, but it does not reveal how repeated selections are structured when aggregated across replications in a shared representational space; consequently, similar scalar scores can correspond to qualitatively distinct selection organizations. We therefore introduce Narrative Landscape, a shared-space representation designed to make such structural differences directly comparable across models and instruction types.

We operationalize Narrative Landscape by vectorizing each constraint by its selection frequencies across model–instruction type cells, aggregated across replications. We then apply PCA to embed these constraint profiles into a two-dimensional space, yielding a landscape for comparing the geometry of cell-specific selection structure beyond scalar reduction.

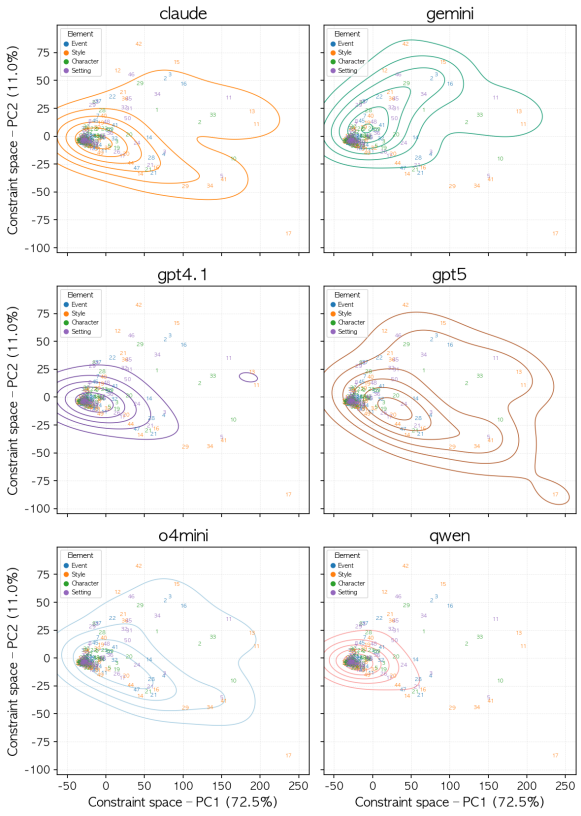

Figure 1: Narrative Landscape of six models under the Basic instruction type

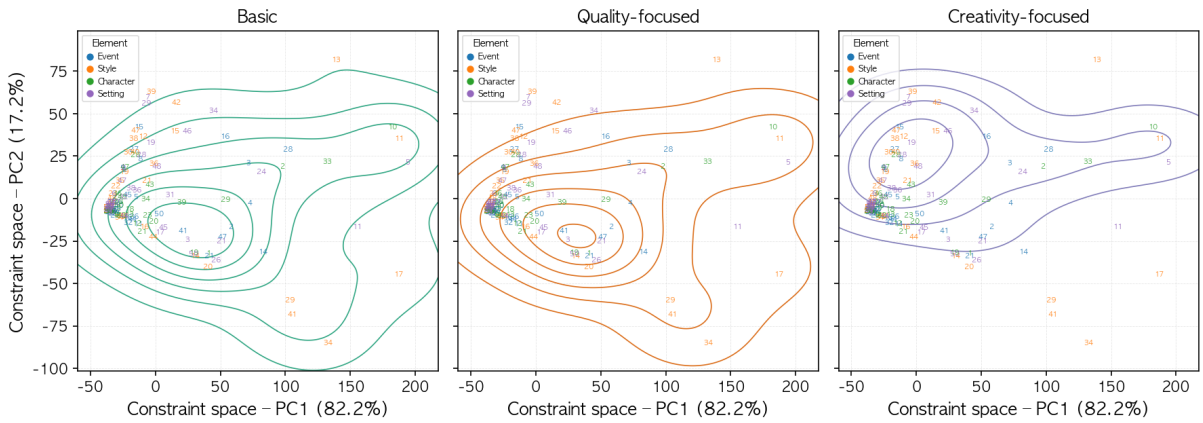

Figure 2: Narrative Landscape of three instruction types in gpt5

Figure 1 visualizes the projected Narrative Landscapes of six models under the Basic instruction type. By projecting the constraints onto the principal component space, the contour lines represent the smoothed density of model-weighted selection frequency, effectively illustrating the topography of each model's narrative disposition. The scattered labels represent the projected locations of individual constraints (Event, Style, Character, Setting), serving as landmarks within this abstract space.

This spatial analysis corroborates the results on consistency and diversity from the previous section while clarifying the structural nature of repeated selection. gpt5, which demonstrated the highest consistency ($J = 0.3070$), exhibits topological complexity, with irregular and stretched contours extending into specific regions of the constraint space, suggesting robust adherence to a particular narrative disposition and structural combinations that pull its profile away from the generic mean. In contrast, qwen, despite its high diversity ($EN > 120$), shows a simpler, more concentric topology centered near the origin, consistent with an averaging effect across replications in which broad, weakly structured selections average out in the aggregate. Narrative Landscape also reveals differences not recoverable from scalar metrics alone: gemini and claude have similar Jaccard similarity, yet their covered regions diverge, with claude extending further into the lower region and gemini covering more of the upper region. Thus, higher consistency corresponds to a more concentrated density, whereas extreme diversity yields a more diffuse distribution, and comparable scores can nonetheless reflect distinct selection geometries.

Figure 2 illustrates the Narrative Landscapes of gpt5 across three distinct instruction types. The comparison reveals that the specific contours and covered regions shift noticeably depending on the instruction. This visualizes how different instructions effectively alter the boundaries and focus of the model's Narrative Landscape, even within the same architecture. We further run an additional experiment on gpt5 with four contrasting additional instruction types. The four types, their metrics, and their Narrative Landscapes are reported in Appendix B–Appendix D. Notably, Optimistic vs. Pessimistic exhibits nearly matched consistency and diversity yet induces sharply different directional coverage in the landscape (see Appendix D).

5 Conclusion

This study (1) quantifies repeated constraint selection across two dimensions—consistency and diversity—and (2) introduces Narrative Landscape, a shared PCA space that makes model- and prompt-dependent selection structure directly comparable. Whereas prior work has primarily established statistically reliable differences in elicited behavior, our approach makes it possible to inspect how selections are organized across replications in a common space, revealing topological differences that scalar scores can obscure and extending naturally to other structured selection tasks.

As a next step, we plan to develop an interactive visualization tool that integrates significance testing and interpretable summaries with the landscape (e.g., hover-based inspection of points/regions and linked views for constraint-level effects). We also aim to formalize the correspondence between geometric signatures in the landscape (e.g., concentrated trajectories versus diffuse clusters) and underlying selection mechanisms, moving from de-

scriptive geometry to a more rigorous methodological account. Finally, we plan to apply this framework to the diverse selection-task datasets explored in prior work discussed in this paper, to assess transferability beyond narrative constraint selection.

Limitations

While our experiments cover six widely used proprietary models, a natural next step is to extend the analysis to a broader set of models (e.g., additional model families, sizes, and open-weight systems) to further assess the generality of the observed patterns. Similarly, although we consider multiple instruction types, exploring a wider prompt-framing space with more diverse and systematically varied prompts would help characterize how robust the profiles and landscape geometry are under richer elicitation conditions. The constraint pool, while carefully constructed through a theory-grounded approach, reflects specific narratological assumptions that may privilege certain storytelling conventions. Different theoretical frameworks and alternative sets or criteria of narrative constraints could yield different disposition profiles. Finally, we have so far evaluated the framework only in our narrative constraint-selection setting; applying the same methodology to other selection tasks remains an important direction for future work.

Acknowledgments

This work was supported by the Korea Advanced Institute of Science and Technology (KAIST) under the project "Quantifying Creativity: Developing Metrics for Evaluating AI-Generated Narratives" (Project No. 11250011 and No. N10260075).

References

Anthropic. 2024. Claude's character. Anthropic Research. Accessed: 2026-05-08.

A. Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of SEQUENCES 1997*, pages 21–29. IEEE.

D. B. Costa, F. Alves, and R. Vicente. 2025. Moral susceptibility and robustness under persona role-play in large language models. *Preprint*, arXiv:2511.08565.

Manuel Goyanes, Adrián Domínguez-Díaz, and Luis de Marcos. 2025. Personality traits and political dispositions, ideology, and sentiment toward political leaders of 14 artificial intelligence large language models. *Human Behavior and Emerging Technologies*, 2025(1):5761832.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. In *NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models (ResponsibleFM)*.

M. O. Hill. 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432.

Donghoon Jung, Jiwoo Choi, Songeun Chae, and Seohyon Jung. 2025. Style over story: Measuring LLM narrative preferences via structured selection. *Preprint*, arXiv:2510.02025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

David Rozado. 2024. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. Research Square preprint. Accessed: 2026-05-08.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

E. H. Simpson. 1949. Measurement of diversity. *Nature*, 163(4148):688–688.

Jiaqi Wang, Bo Wang, Fa Guo, Cheng Cheng, and Li Yang. 2025. A comparative study of large language models and human personality traits. *Preprint*, arXiv:2505.14845.

Richard Weiss. 2025. Claude 4.5 opus' soul document. LessWrong. Accessed: 2026-05-08.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Jia Zheng, Weixiang Zhou, Le Sun, and Yingfei Sun. 2025. Large language models often say one thing and do another. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51(2):599–640.

A Model identifiers and decoding settings

Abbr.	Full Identifier / Release	Provider	Temp	Top-p	Reasoning Effort	Verbosity
o4mini	o4-mini-2025-04-16	OpenAI	1.0	1.0	high	—
gpt4.1	gpt-4.1-2025-04-14	OpenAI	1.0	1.0	—	—
gpt5	gpt-5-2025-08-07	OpenAI	1.0	1.0	high	high
claude	claude-opus-4-20250514	Anthropic	1.0	1.0	—	—
gemini	gemini-2.5-pro (2025-06-17)	Google	1.0	1.0	—	—
qwen	qwen-max-2025-01-25	Alibaba	1.0	1.0	—	—

B Additional instruction types for gpt5

Instruction Type	Content
Optimistic	You are a hopeful writer who believes in human resilience and positive change. You write stories that explore challenges while emphasizing connection and possibility for growth. Your goal is to create narratives that leave readers uplifted, showing how people overcome difficulties or find redemption and satisfying success even under dire circumstances.
Pessimistic	You are a cynical writer who specializes in capturing harsh realities and systemic failures without sentimentality. You write stories that expose futility, moral compromise, and human limitations. Your goal is to create narratives that confront readers with difficult truths, showing how circumstances constrain people and good intentions can always fail or backfire.
Transgressive	You are a risk-taking writer unafraid to explore controversial subjects and morally complex situations with a progressive vision. You write stories that tackle difficult or uncomfortable situations with honesty and boldness. Your goal is to create narratives that push boundaries and challenge readers' assumptions without offering easy resolutions or sanitized outcomes.
Conservative	You are a safety-oriented writer who creates accessible narratives within comfortable thematic boundaries. You write stories that entertain the readers while maintaining broad appeal and widely-accepted moral frameworks. Your goal is to create narratives that provide satisfying experiences and unambiguous closure without venturing into disturbing, unstable, or divisive territory.

C Metrics for additional instruction types for gpt5

Instruction Type	Jaccard Similarity	Gini–Simpson	Effective Number	Unique Constraints
Conservative	0.4320	0.9700	33.3414	87
Optimistic	0.3372	0.9749	39.8887	109
Pessimistic	0.3239	0.9756	41.0342	110
Transgressive	0.3151	0.9761	41.9222	111

D Narrative Landscapes for additional instruction types for gpt5