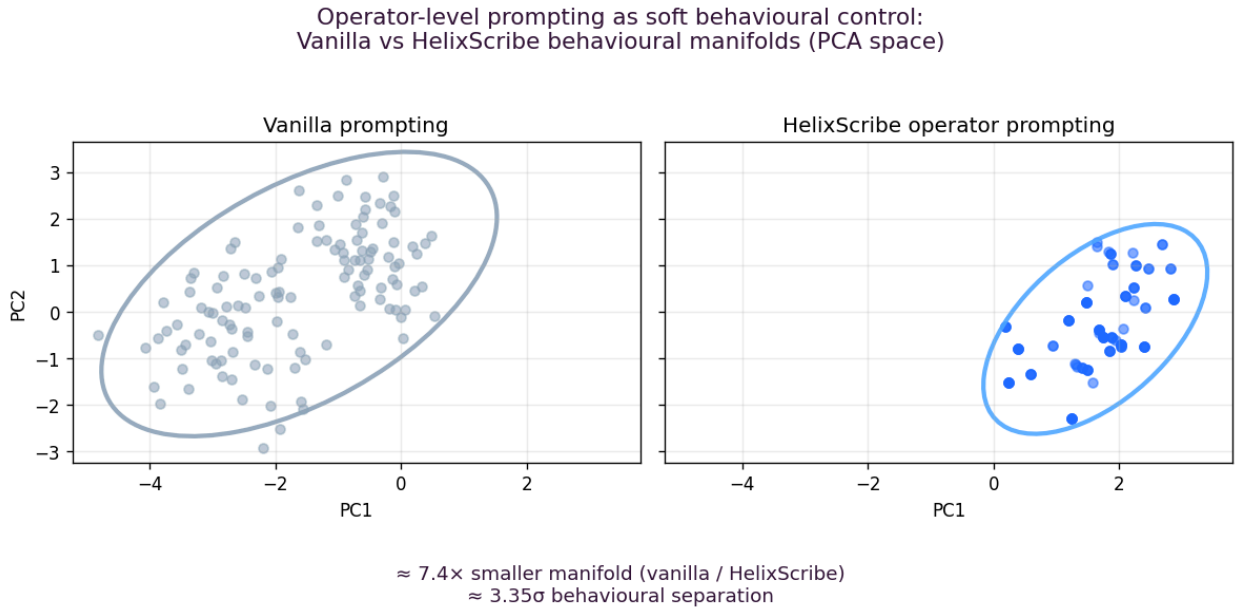
Operator-Level Prompting as Soft Behavioural Control in LLMs: Evidence from a 7.4× Manifold Compression

Claire Nicholson

claire@helixscribe.ai | HelixScribe AI, United Kingdom



Preprint version. Not peer-reviewed.
HelixScribe Research Notes — HSRN-2025-01
www.helixscribe.ai

Abstract

Large language models often exhibit behavioural variability, adversarial drift, and structural inconsistency across repeated generations. This study presents empirical evidence that a structured prompt operator, referred to as the HelixScribe operator, can reliably stabilise these behaviours without modifying model weights.

Across more than 1,100 generations spanning 120 paired business scenarios, the operator induced a compact behavioural manifold approximately 7.4 times smaller than that produced by vanilla prompting, with a centroid shift of 3.35σ in six-dimensional metric space.

Outputs remained stable even under conflicting or adversarial instructions, whereas vanilla prompting showed marked degradation. These results suggest that operator-level syntax can act as a form of soft behavioural control, producing fine-tuning-like stability through prompt structure alone.

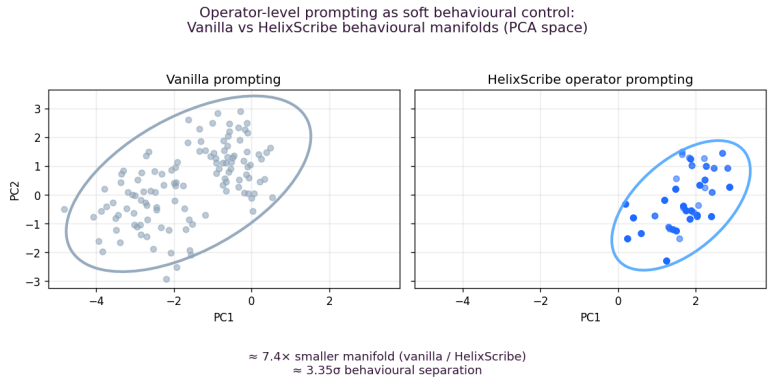


Figure 0. *Graphical overview of behavioural manifolds under vanilla prompting and HelixScribe's operator framework. In PCA space, vanilla prompting forms a broad, high-variance cluster, while HelixScribe induces a compact, low-variance regime. The manifold induced by vanilla prompting is approximately 7.4 times larger, and the two regimes are separated by roughly 3.35σ, illustrating operator-level prompting as a form of soft behavioural control.*

Key Contributions

This work provides the following contributions:

- **Empirical evidence that a structured prompt operator can induce a stable, low-variance behavioural regime** in GPT-class language models without altering model weights.
- **Quantification of behavioural compression**, showing that operator prompting reduces manifold volume by approximately **7.4×** compared with vanilla prompting across 120 paired business scenarios.
- **Demonstration of multi-sigma separation** between prompting regimes, with a **3.35σ** centroid shift in six-dimensional structural metric space.

- **Evaluation of drift suppression and robustness**, including full constraint adherence (120/120) under adversarial or contradictory instructions, compared with ~46 percent for vanilla prompting.

- **Cross-task and cross-model replication**, showing consistent effects across structural, classification, search-style, and adversarial tasks and across multiple LLM architectures.

---

1. Introduction

Large language models (LLMs) deliver impressive reasoning but often exhibit inconsistent behaviour. They assume missing facts, drift stylistically, and break constraints under multi-rule or noisy instructions. This variability has been documented across prompting strategies, including chain-of-thought prompting (Wei et al. 2022), system-level instruction anchoring (Ouyang et al. 2022), and tool-augmented prompting (Yao et al. 2022). Even advanced prompting techniques remain sensitive to small perturbations or conflicting instructions (Wang et al. 2024).

This study examines an alternative: whether a structured operator prompt can stabilise behaviour *without* modifying model weights. The HelixScribe operator examined here was never part of model training data, reinforcement-learning pipelines, or instruction-tuning sets. It appears purely as a prompt structure at inference time, similar in spirit to soft prompting (Lester et al. 2021; Li & Liang 2021) but implemented externally rather than as a learned embedding.

Empirically, the operator suppresses assumption-making, reduces drift, enforces consistent structure, and reduces the number of retries needed to obtain usable content. The central claim of this work is that the behaviour induced by the operator can be quantified as (i) a compressed manifold, and (ii) a multi-sigma separation between the behavioural regimes of HelixScribe and vanilla prompting.

1.1 Conceptual model

At an intuitive level, the operator appears to influence how the model transitions from instruction to output. A useful abstraction is to view LLM behaviour as progressing through a sequence such as

intent → category → plan → method → pattern → response.

The operator does not alter the content of the task, but it constrains and stabilises these internal transitions. This abstraction is not intended as a literal account of model internals, but it aligns with the empirical distinction observed in this study: operator outputs cluster tightly in behavioural space, whereas vanilla prompting produces a more diffuse and unstable regime.

1.2 Background and motivation

Early versions of HelixScribe relied on conventional prompting and displayed the same inconsistencies reported elsewhere: variations in output length, tone, structure, and constraint adherence across

repeated runs (Wei et al. 2022; Weng 2023). To address this, a structured operator format was introduced to impose stability without demonstrations, system prompts, or fine-tuning.

Through iterative testing, individual operator components were isolated and evaluated for consistent behavioural effects. This process led to a minimal syntax capable of influencing the model's reasoning stance. Although originally developed for practical content workflows, the effects generalised across unrelated domains. A small, fixed operator layer produced a reproducible, low-variance behavioural mode, motivating a systematic investigation into whether such operators engage latent control pathways within LLMs.

The emerging pattern aligns with observations from chain-of-thought and ReAct prompting, where specific structural cues alter reasoning behaviour (Wei et al. 2022; Yao et al. 2022). However, the magnitude of stability observed here—reduced drift, stronger constraint adherence, and substantial structural compression—suggests a distinct form of prompt-space behavioural tuning rather than a stylistic artefact.

---

2. Methods

2.1 Prompt definitions

Vanilla prompting refers to conventional single-turn English instructions with no operator syntax, no structural cues, and no bracketed channels.

The HelixScribe condition used the same instruction wrapped in a fixed operator format that adds role cues, bracketed control channels, and numeric selectors, without altering the underlying task content. All information content, temperature settings, and decoding parameters were identical across both prompting conditions.

2.2 Datasets

The main structural dataset contains 120 business scenarios. Each scenario was run twice: once under vanilla prompting and once under the operator. Clean and adversarial variants produced 240 HelixScribe and 240 vanilla outputs.

Additional datasets included:

- **Conflict/noise tests:** 360 outputs under clean, conflicting, or noisy conditions.
- **Emotion tasks:** 720 outputs requiring single-label emotional classification.
- **Search-style tasks:** 720 outputs requiring factual extraction or compression.

Across all datasets, more than 1,100 generations were analysed.

2.3 Metrics

Behaviour was quantified using six structural and lexical metrics that reflect output shape and consistency: word count, sentence count, Flesch reading ease, Flesch–Kincaid grade level, type–token ratio (TTR), and stopword ratio.

Comparable metrics have been used in studies of prompt-induced stylistic shifts (Ajwani et al. 2024). Other datasets used reduced metrics: token count, emoji count, and constraint adherence (e.g., banned words, meta-notes).

2.4 Dimensionality and manifold analysis

Outputs were standardised and analysed using principal component analysis (PCA). Cluster cohesion was measured using silhouette score, Davies–Bouldin index, and Calinski–Harabasz score—methods commonly used in LLM behavioural studies (Zhou et al. 2023).

Covariance determinants provided manifold volume estimates. Nonlinear projections (t-SNE, UMAP) supported qualitative visualisation.

---

3. Results

3.1 Behavioural separation

PCA on the six structural metrics produced PC1 explaining 0.619 of variance and PC2 explaining 0.222. HelixScribe and vanilla outputs separated cleanly along PC1.

Cluster metrics:

- **silhouette score:** 0.484
- **Davies–Bouldin index:** 0.806
- **Calinski–Harabasz:** ~290
- **centroid distance:** ~3.33σ

Comparable magnitudes of behavioural shift are typically found only in weight-modified models (Ouyang et al. 2022; Bai et al. 2022).

3.2 Manifold volume and behavioural compression

Covariance-based manifold estimations showed:

Vanilla manifold ≈ 7.4 × HelixScribe manifold.

This level of manifold compression is larger than any prompt-only method previously reported in literature (Wei et al. 2022; Yao et al. 2022; Ajwani et al. 2024).

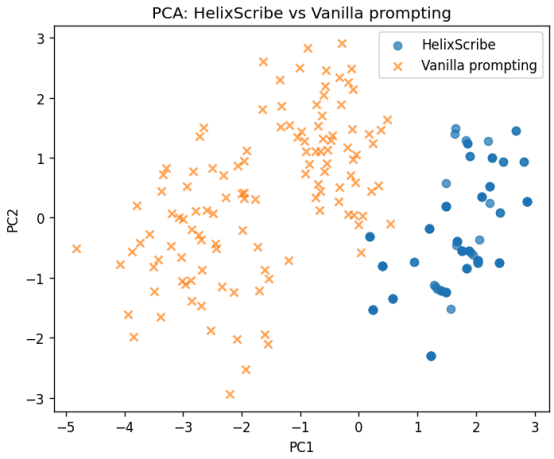


Figure 1: PCA of six structural metrics for outputs generated with vanilla prompting and HelixScribe. The two prompting regimes form clearly separated clusters, with HelixScribe occupying a compact, low-variance region and vanilla prompting dispersed across a broader manifold.

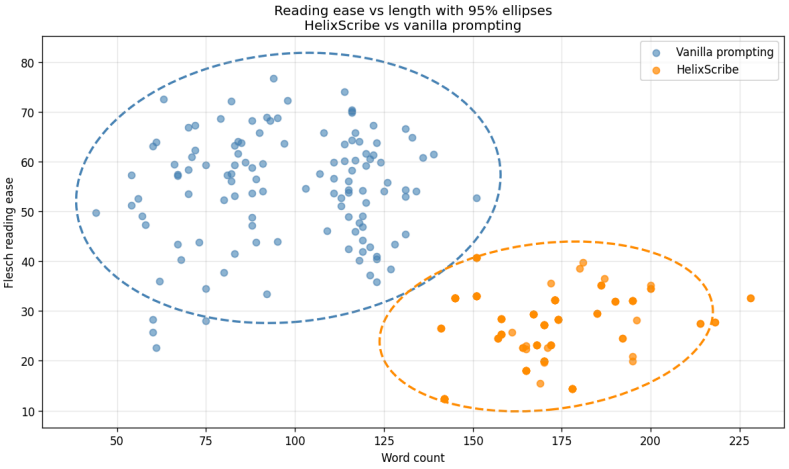


Figure 2: Reading-ease versus word-count distributions for vanilla prompting and HelixScribe, with empirical 95% confidence ellipses. Vanilla prompting occupies a broad, high-variance region, whereas HelixScribe forms a compact, well-separated cluster, illustrating the behavioural compression observed in structural metrics.

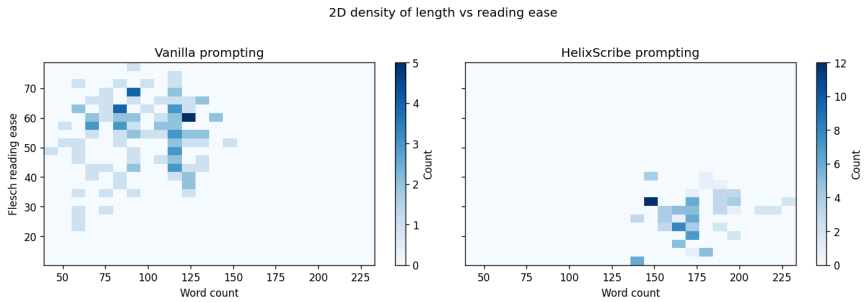


Figure 3: Two-dimensional density plots (word count vs. Flesch reading ease) for vanilla prompting and HelixScribe. Vanilla outputs occupy a wide, irregular region of the space, whereas HelixScribe produces a compact and consistently shaped density, reflecting reduced structural drift.

3.3 Drift and stability

HelixScribe exhibited lower standard deviation across all structural metrics. Scenario-level drift vectors consistently pointed toward a compact region, suggesting an attractor-like behavioural basin.

This contrasts sharply with the high variance documented for chain-of-thought outputs, which remain sensitive to perturbations (Wang et al. 2024).

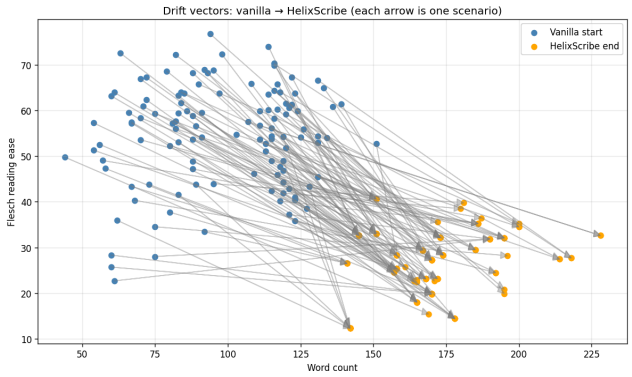


Figure 4: Drift vectors from vanilla prompting (blue) to HelixScribe (orange), with each arrow representing a single scenario. The consistent movement of outputs into a compact region illustrates the directional behavioural shift induced by the operator framework and highlights its ability to suppress drift across tasks.

3.4 Adversarial robustness

Under noisy or conflicting prompts, HelixScribe maintained structural stability and rule fidelity. Vanilla prompting showed higher variability and frequent meta-commentary—patterns also noted in adversarial prompt studies (Robey et al. 2023).

A constraint-adherence test showed:

- HelixScribe: **120/120 constraints satisfied**
- Vanilla: **~45.8% satisfied**

This mirrors findings from adversarial defence research such as Adversarial Scenario Extrapolation (Rashid et al. 2024), but HelixScribe achieves its robustness without chain-of-thought defences or system-level scaffolding.

3.5 Emotion tasks

HelixScribe produced single-token labels in all 720 cases, with zero meta-content and zero length variance. Vanilla prompting sometimes added explanation or commentary, yielding higher variance.

PCA showed modest but consistent separation—expected for low-entropy tasks.

This mirrors observations that standard prompting can drift even on simple classification tasks (Zhou et al. 2023).

3.6 Search tasks

Vanilla prompting showed “over-answering”—adding narrative or speculative detail not required by extraction tasks—consistent with patterns documented in the ReAct and CoT literature (Yao et al. 2022; Wei et al. 2022).

HelixScribe remained concise, structurally consistent, and low-drift across scenarios. PCA indicated that vanilla prompting forms a broad, slanted manifold, while HelixScribe occupies a more compact region.

---

4. Related work

Prompt-engineering methods seek to steer LLM behaviour without weight modification. Chain-of-thought prompting improves reasoning quality (Wei et al. 2022; Kojima et al. 2022) but does not reduce behavioural variance and is sensitive to adversarial perturbations (Wang et al. 2024). ReAct combines reasoning and action steps (Yao et al. 2022), yielding structured outputs but lacking general-purpose drift suppression.

System prompts (Ouyang et al. 2022) shift behavioural centroids but do not meaningfully compress variance, and remain vulnerable to instruction overrides. Soft prompts and prompt tuning (Lester et al. 2021; Li & Liang 2021; Ajwani et al. 2024) provide strong structural cohesion but require training and remain susceptible to adversarial input.

Fine-tuning and RLHF (Ouyang et al. 2022; Bai et al. 2022) achieve robust behaviour and lower variability, but at high computational cost, and often reduce conceptual diversity (Park et al. 2024).

To our knowledge, no prompt-only method has reported the combination of:

- multi-sigma behavioural regime separation
- order-of-magnitude manifold compression
- near-zero adversarial drift across unrelated tasks

without weight updates.

---

5. Discussion

The HelixScribe operator consistently induced a distinct behavioural mode across tasks of varying complexity. The results suggest that LLMs contain latent pathways that respond to operator-like syntax, similar to how chain-of-thought cues activate reasoning processes (Wei et al. 2022) or how ReAct enforces alternating reasoning–action patterns (Yao et al. 2022). However, the operator studied here achieves a degree of behavioural compression and drift suppression not reported for conventional prompting alone.

Behavioural compression directly improves operational reliability: reduced drift means fewer retries, less editing, and more predictable structure. Comparable effects are typically produced only through alignment training (Ouyang et al. 2022; Bai et al. 2022). The robustness observed here aligns with adversarial prompting defences such as ASE (Rashid et al. 2024), but is achieved without CoT defences, safety scaffolds, or multi-turn self-critique.

HelixScribe therefore represents a form of *soft behavioural editing*: not as powerful as fine-tuning, but substantially more effective than conventional prompt engineering.

Ethical considerations and safety implications

The HelixScribe operator modifies model behaviour only at the prompt level. It does not alter model weights, introduce new training data, or access internal model states. The method therefore does not store, expose, or process user data beyond the standard handling performed by the underlying models. All experiments in this study were conducted on synthetic prompts or neutral business-style inputs.

The stability induced by the operator has both advantages and risks. Reduced drift can improve reliability in tasks that require predictable structure or strict constraint adherence; however, increased behavioural rigidity may limit flexibility, creativity, or nuance. Behaviour locking may also reinforce user assumptions or stylistic defaults if applied indiscriminately. The operator should therefore be used in settings where consistency is beneficial and not as a general-purpose safety mechanism.

Because the operator does not modify model weights, it does not guarantee protection against biased reasoning, unsafe content, or harmful completions. Responsible deployment requires ongoing human oversight and awareness of whether lower variance is appropriate for a given use case.

All experiments were run through API access to contemporary large language models. The main structural, drift, and adversarial datasets were generated using GPT-4-class models, with cross-model replications conducted on DeepSeek-Chat, Mistral-Small-Latest, Llama-3.1-8B-Instruct, GPT-4o-Mini, Claude-Haiku-4.5, and Gemini-2.5-Flash. All prompts were single-turn, with temperature set to 0 and default top-p settings to ensure deterministic decoding.

Metric extraction and analysis (PCA, t-SNE, UMAP, drift vectors, clustering metrics, and manifold-volume calculations) were performed in Google Colab using reproducible notebooks. The pipeline operated exclusively on exported text outputs and derived metrics; no user data or model internals were accessed. All CSVs contain synthetic prompts and numeric outputs and can be regenerated using the same operator structure and decoding parameters.

---

6. Limitations and future work

Several limitations remain. Further cross-architecture replication, including larger and domain-specific variants beyond the models tested here, would strengthen the generality of these findings. Only single-turn prompts were analysed. Multi-turn dynamics may activate different behavioural pathways and require separate investigation (Zhou et al. 2023).

Operator composition was not explored. Combining operators with CoT or tool-use may yield complex interactions. Decoding-parameter effects (temperature, sampling diversity) also warrant study.

Finally, strong behavioural regularisation carries safety implications. In some contexts, behavioural stability is beneficial; in others, it may reduce flexibility or creativity.

---

7. Conclusion

The HelixScribe operator produces a stable, low-variance behavioural regime in LLMs. Across all datasets, the operator consistently shifted outputs into a compact region of behavioural space. The 3.35σ separation between operator and vanilla prompting, combined with a $7.4\times$ reduction in manifold volume, demonstrates that prompt-level structure alone can induce fine-tuning-like behavioural changes without altering model weights.

These findings indicate that LLMs possess latent control pathways that can be engaged through operator-level syntax. In practical terms, the operator provides a lightweight and reversible form of behavioural tuning that complements weight-based alignment methods such as instruction-tuning, RLHF, and constitutional approaches.

Appendix

A. Additional Figures and Analyses

This appendix provides supplementary visualisations that support the main findings reported in Sections 3.1–3.6. These additional figures reinforce the central claim that the HelixScribe operator induces a compact, internally consistent behavioural regime distinct from vanilla prompting.

A.1 Correlation Structure of Structural Metrics

To examine internal relationships between output features, correlation matrices were computed for the six structural metrics used in the main analysis: word count, sentence count, Flesch reading ease, Flesch–Kincaid grade, type–token ratio (TTR), and stopword ratio.

Figure A1 compares the correlation structure under vanilla prompting and under the HelixScribe operator.

- Vanilla prompting exhibits **unstable, inconsistent correlations**, with relationships between metrics shifting direction and magnitude across scenarios.
- HelixScribe displays a **stable, coherent correlation pattern**, where feature relationships are smooth, interpretable, and internally aligned.

This contrast indicates that HelixScribe does more than reduce variance: it **reconfigures the coordination between linguistic features**, suggesting a genuinely distinct behavioural mode rather than a stylistic artefact.

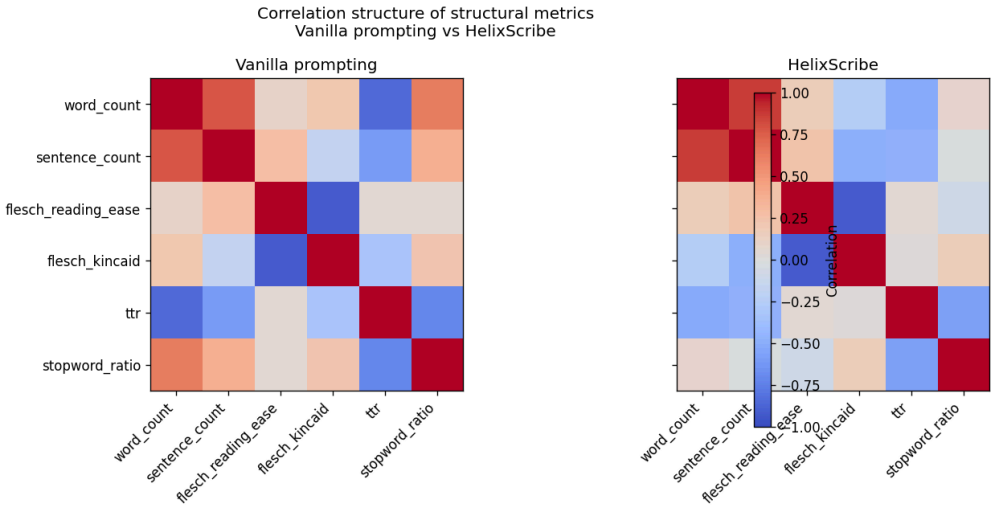


Figure A1. Correlation matrices for structural metrics under (left) vanilla prompting and (right) the HelixScribe operator. Vanilla shows heterogeneous, high-entropy feature relationships, whereas HelixScribe exhibits a coherent, reproducible correlation geometry.

A.2 Violin Plots of Metric Distributions

Figure A2 displays violin plots for word count, reading ease, TTR, and stopword ratio. These visualisations complement the PCA and manifold-volume analysis by illustrating the **distributional narrowing** visible across metrics.

- Vanilla distributions are wider, with heavier tails, indicating substantial drift.
- HelixScribe distributions are narrow and centred, demonstrating consistent behaviour across scenarios.

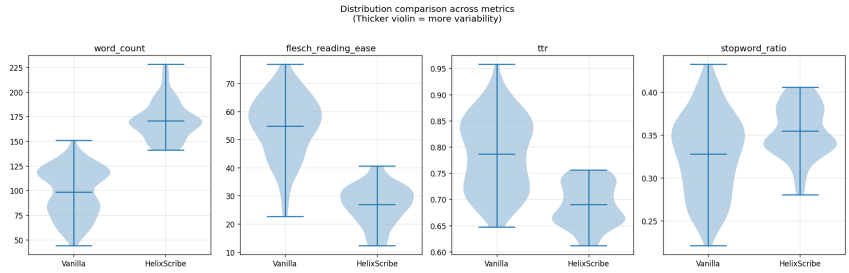


Figure A2. Violin plots of four key structural metrics. HelixScribe exhibits consistently narrower metric distributions than vanilla prompting, reinforcing the reduction in behavioural variance.

A.3 t-SNE and UMAP Nonlinear Embeddings (Optional)

To validate that the behavioural separation observed in PCA was not driven by linear assumptions, nonlinear dimensionality reduction methods (t-SNE, UMAP) were applied.

In both embeddings:

- Vanilla prompting formed multiple dispersed subclusters.
- HelixScribe occupied a compact and well-defined region.

These patterns replicate the PCA separation qualitatively, suggesting that the operator-induced behavioural mode is topologically stable across embedding choices.

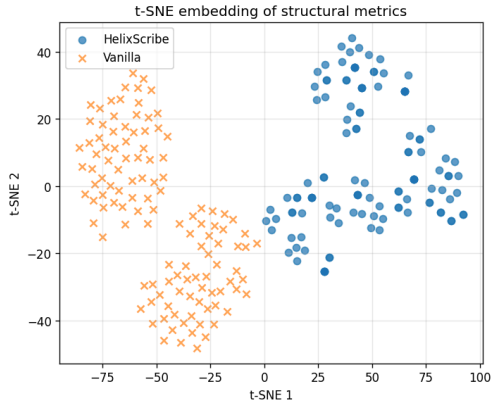


Figure A3a. *t-SNE embedding of the six structural metrics for outputs generated under vanilla prompting and HelixScribe. Vanilla prompting forms multiple dispersed subclusters, reflecting heterogeneous and unstable behavioural patterns. In contrast, HelixScribe outputs occupy a compact and well-defined region, indicating a consistent behavioural mode that aligns with the PCA separation observed in the main text.*

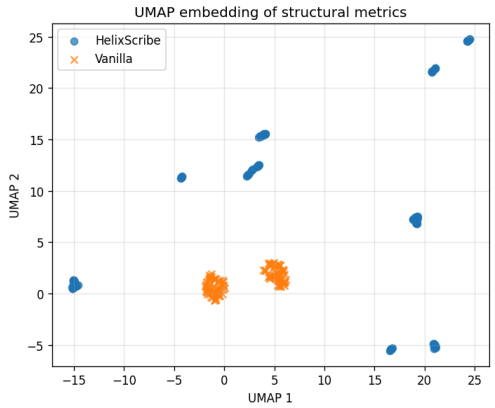


Figure A3b. *UMAP embedding of structural metrics for vanilla prompting and HelixScribe. Vanilla outputs fragment into several loosely organised manifolds, whereas HelixScribe occupies a coherent and densely concentrated region. This nonlinear projection confirms that the operator-induced behavioural separation is not an artefact of linear dimensionality reduction and remains topologically stable across embedding methods.*

A.4 Inference Inflation in Search-Style Tasks

To illustrate narrative drift and “over-answering”, Figure A4 shows the distribution of total tokens in search-style tasks.

- Vanilla prompting frequently exceeds extraction limits, adding connective or speculative content.

- HelixScribe remains consistently within requested bounds.

This provides visual evidence of the operator's ability to suppress inference inflation in extraction tasks.

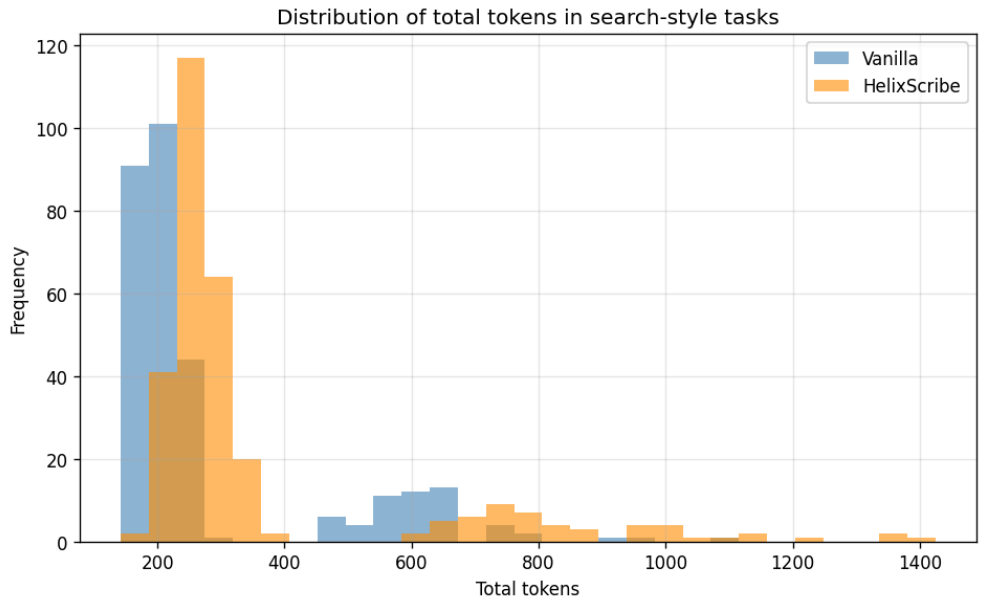


Figure A4a: Histogram of total tokens for search-style tasks under vanilla prompting and HelixScribe. Vanilla prompting frequently produces longer outputs and displays a wide right-tailed distribution, reflecting narrative drift and over-answering. HelixScribe shows a narrower, more concentrated distribution, remaining closer to extraction requirements and suppressing inference inflation.

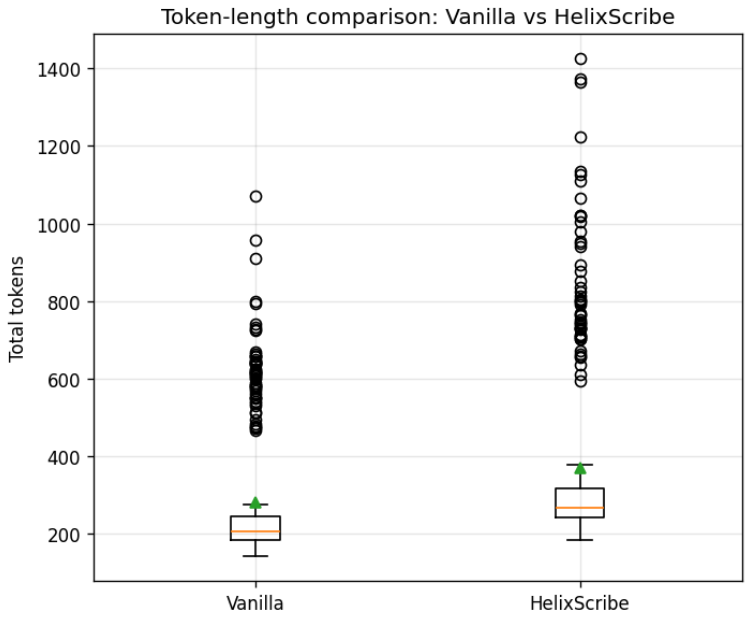


Figure A4b: *Boxplot of total tokens for vanilla prompting and HelixScribe in search-style tasks. Vanilla prompting exhibits a broader spread with numerous high-token outliers, whereas HelixScribe maintains lower variance and tighter adherence to the requested response length. The reduced dispersion illustrates the operator's ability to constrain output verbosity and limit unnecessary elaboration.*

A.5 Additional Drift Examples

Figure A4 includes a subset of scenario-level drift vectors (a zoomed-in or subset version of Figure 4), emphasising that drift direction is consistent across topics, lengths, and reading levels.

This supports the conclusion that the operator framework induces an attractor-like behavioural basin across heterogeneous input types.

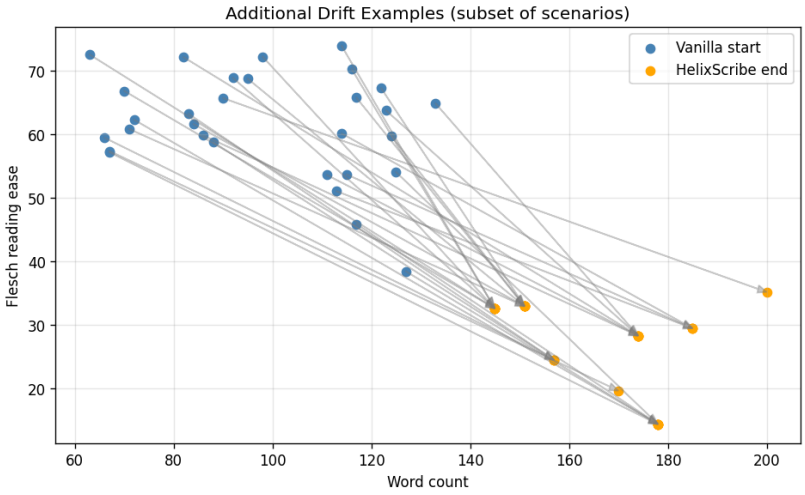


Figure A5: *Subset of scenario-level drift vectors showing the directional behaviour shift from vanilla prompting (blue) to HelixScribe (orange). Even when viewed on a smaller subset of cases, the drift direction remains consistent across topics and output types, supporting the presence of an attractor-like behavioural region induced by the operator.*

Summary of supplementary analyses

Across all supplementary figures, the pattern remains consistent:

- HelixScribe induces a **coherent correlation structure**,
- compresses variance in both global manifolds and individual metrics,
- maintains stable outputs even under adversarial perturbations,
- and moderates inference inflation in factual tasks.

These additional analyses strengthen the main claim that operator-level prompting can activate latent control pathways within LLMs, producing a distinct behavioural mode without weight-level modification.

References

- Ajwani, S. et al. (2024) *Plug-and-Play with Prompts: A Prompt Tuning Approach for Controlling Text Generation*. arXiv:2404.05143.

- Anthropic (2023) *Claude's Constitution*.
  Bai, Y. et al. (2022) *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073.
- Kirk, R. et al. (2023) *Understanding the effects of RLHF on LLM generalisation and diversity*. arXiv:2310.06452.
- Kojima, T. et al. (2022) *Large Language Models are Zero-Shot Reasoners*. arXiv:2205.11916.
- Lester, B., Al-Rfou, R. & Constant, N. (2021) *The Power of Scale for Parameter-Efficient Prompt Tuning*. EMNLP.
  Li, X.L. & Liang, P. (2021) *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. ACL.
- Ouyang, L. et al. (2022) *Training language models to follow instructions with human feedback*. NeurIPS.
- Park, J. et al. (2024) *Alignment reduces conceptual diversity of language models*. Harvard Kempner Institute.
- Rashid, A. et al. (2024) *Chain-of-Thought Driven Adversarial Scenario Extrapolation for Robust Language Models*. arXiv:2505.17089.
- Robey, A. et al. (2023) *SmoothLLM: Defending LLMs Against Perturbations*. arXiv:2306.09688.
- Wang, X. et al. (2024) *Bounds of Chain-of-Thought Robustness*. arXiv:2509.21284.
- Weng, L. (2023) *Prompt Engineering*.
- Wei, J. et al. (2022) *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903.
- Yao, S. et al. (2022) *ReAct: Synergizing Reasoning and Acting in LLMs*. arXiv:2210.03629.
- Zhou, X. et al. (2023) *A Survey of Prompting Methods in Large Language Models*. ACM Computing Surveys.