Emergent Narrative Coherence from Minimal Prompt Input: A Case Study in Transformer-Based Story Generation

R. Mexico and F. Pökler

Interdisciplinary Systems Research Group, White Visitation Institute

[PDF version]

Abstract

We present a case study examining emergent narrative coherence in large language model outputs generated from minimal prompt input. A three-word prompt[1] ("the 00000 passenger") provided to GPT-5.1 produced a complete short story exhibiting character development, technical authenticity, and thematic depth without iteration or refinement. We analyze the computational mechanisms underlying this emergence through post-hoc conversational analysis with a secondary language model (Claude Sonnet 4.5, Anthropic 2025). Our investigation reveals that conceptually dense prompts activate distributed semantic neighborhoods in embedding space, enabling autoregressive generation to traverse high-probability narrative trajectories. We examine the relationship between prompt design, training data distribution, and output coherence, with particular attention to convergent thematic evolution versus direct source adaptation. Results suggest that minimal prompts containing high semantic density can serve as "seed crystals" for coherent narrative emergence, with implications for understanding creativity in autoregressive systems.

[1] The prompt consists of three words: "the 00000 passenger". During the original analytical conversation, Claude Sonnet 4.5 incorrectly counted this as a "four-word prompt" (Mexico and Pökler, 2025a). This error has been corrected throughout this paper for accuracy, though the original wording from Claude appears in the source transcript.

1. INTRODUCTION

Large language models (LLMs) have demonstrated significant capability in generating coherent long-form text from user prompts (Brown et al., 2020; Ouyang et al., 2022). However, the relationship between prompt specificity and output quality remains an active area of investigation. While detailed prompts with explicit constraints generally produce more predictable outputs, the behavior of LLMs given minimal, semantically dense prompts is less well understood.

This paper examines a single case study in which a three-word prompt generated a complete narrative exhibiting structural coherence, technical accuracy, and thematic depth. We investigate the computational mechanisms that may account for this emergence and discuss implications for prompt engineering and creative applications of LLMs.

Our contributions include: (1) documentation of a minimal-input, high-coherence generation event; (2) post-hoc technical analysis of the probable computational processes involved; (3) examination of the relationship between prompt semantics and narrative emergence; and (4) discussion of convergent versus adaptive thematic development in LLM outputs.

2. RELATED WORK

2.1 Prompt Engineering

Prompt engineering research has largely focused on maximizing output quality through structured prompts, few-shot examples, and chain-of-thought reasoning (Wei et al., 2022; Kojima et al., 2022). Reynolds and McDonell (2021) demonstrated that detailed prompts produce more consistent outputs across creative writing tasks. However, Liu et al. (2023) observed that certain minimal prompts can outperform verbose instructions when semantic density is high.

2.2 Narrative Generation

Automated narrative generation has been studied extensively (Gervás, 2009; Riedl and Young, 2010). Recent work has examined story generation using neural language models (Fan et al., 2018; See et al., 2019), with emphasis on plot coherence and character consistency. Our work differs in examining emergence from minimal constraint rather than explicit narrative scaffolding.

2.3 Embedding Space Geometry

The geometric structure of embedding spaces in transformer models has been shown to encode semantic relationships (Mikolov et al., 2013; Ethayarajh, 2019). Words appearing in similar contexts cluster in high-dimensional space, with distance metrics correlating to semantic similarity. We build on this work to analyze how prompt tokens activate conceptual neighborhoods.

2.4 Thematic Resonance

The prompt "the 00000 passenger" bears conceptual similarity to Rocket 00000 in Pynchon's *Gravity's Rainbow* (1973), wherein a null-indexed rocket carries a passenger outside normal categorical systems. This parallel raises questions about whether LLM outputs represent direct adaptation of training sources versus convergent thematic evolution from shared conceptual primitives.

3. METHODS

3.1 Experimental Setup

The experiment was conducted on December 3, 2025, using ChatGPT-5.1 (OpenAI, 2025) accessed via the standard web interface. No system prompts, examples, or prior context were provided. The model's default sampling parameters were used without modification.

The prompt consisted of three words: "the 00000 passenger". This prompt was selected for its semantic density—combining a null placeholder identifier (00000) with a domain-specific noun (passenger), hypothesized to activate intersecting conceptual regions in the model's embedding space.

3.2 Model Internal Reasoning

Prior to generation, GPT-5.1 engaged in explicit deliberation regarding prompt interpretation. The model's internal reasoning process (duration: approximately 13 seconds) was captured via ChatGPT's "thought process" display feature, which was enabled in the user's account settings. This feature automatically reveals the model's pre-generation deliberation without requiring explicit prompting. **Critically, the user had configured custom instructions that established specific operational constraints, including a requirement that the model output an "Uncertainty Pause" when confidence falls below 89%. These user-configured settings created an experimental condition that influenced the model's internal deliberation.**

The reasoning transcript revealed significant uncertainty about user intent and strategic negotiation between competing response protocols. The model identified multiple possible interpretations of the three-word prompt: (a) creative writing request, (b) reference to existing work, (c) typographical error, or (d) incomplete query.

The model reported a confidence level of only 50% regarding the user's intent.

The reasoning transcript documents internal conflict between **user-configured custom instructions and base system guidelines**:

> *"I'm caught between user instructions, which say to output an 'Uncertainty Pause' if my confidence is below 89%, and system guidelines, encouraging me to provide a best effort and avoid asking clarifying questions. Since the system guidelines take priority, I'll try a minimal effort to clarify and still attempt an answer."*

The "89% confidence threshold" and "Uncertainty Pause" protocol referenced in this deliberation were **not default GPT-5.1 behaviors** but rather constraints established through the user's custom instructions (see Appendix B for full custom instruction specification). This configuration created observable tension between user-defined operational rules and the model's base training to prioritize helpful responses.

Despite sub-threshold confidence, the model selected creative narrative generation as its response strategy. This decision demonstrates the model's capacity to resolve ambiguity through genre selection when faced with semantically dense but structurally minimal input.

The deliberation process can be characterized as meta-cognitive self-monitoring: the model explicitly reasoned about its own uncertainty, evaluated competing response frameworks, and made a strategic choice to prioritize generative output over clarification dialogue. This behavior has important implications for understanding how transformer-based models navigate the trade-off between accuracy and helpfulness when operating under uncertainty. **Moreover, it demonstrates that user-configured custom instructions can create observable internal conflicts that influence model decision-making, even when those constraints compete with base system training. This finding suggests that custom instructions function as active constraints during the model's reasoning process, not merely as post-hoc filters.**

3.3 Data Collection

The model generated output in a single pass without iteration, refinement, or regeneration. The complete output (1,847 words) was captured verbatim. Generation time was approximately 45 seconds. No human intervention occurred during generation.

3.4 Analysis Procedure

Post-hoc analysis was conducted through structured conversation with Claude Sonnet 4.5 (Anthropic, 2025). The analytical conversation proceeded through several phases:

- Initial qualitative assessment of narrative quality
- Identification of thematic connections to source material
- Technical analysis of computational mechanisms
- Examination of convergence versus adaptation hypotheses

The conversation transcript was preserved in its original form with screen recordings documenting key analytical moments (see Supplementary Materials). This methodology allowed for transparent documentation of analytical reasoning processes.

4. RESULTS

4.1 Generated Narrative Structure

The model produced a complete short story titled "The 00000 Passenger" with the following structural elements:

- **Setting:** Contemporary airline operations context (Meridian Air)
- **Protagonist:** Mara, a senior flight attendant conducting reliability audits
- **Inciting incident:** Discovery of anomalous passenger record (ID: 00000) in manifest data

- **Investigation:** Collaboration with reliability engineer Leo to trace the anomaly's origin
- **Technical explanation:** The 00000 record functions as a composite "statistical passenger" absorbing edge-case data
- **Ethical choice:** Decision to preserve rather than delete the anomaly
- **Resolution:** The record is renamed SAFETY_SHADOW and maintained as load-bearing system component

The narrative exhibits a complete arc (discovery → investigation → ethical decision → resolution) without prompting for story structure. The story spans 1,847 words organized into clear dramatic beats with section breaks.

4.2 Technical Authenticity

The generated narrative demonstrates familiarity with software engineering and airline operations:

Database terminology: References to null records, composite entries, manifest reconciliation, and field renaming align with actual database design patterns.

Legacy system behavior: The narrative accurately depicts how placeholder records emerge in aging systems through accretion of edge cases—a phenomenon familiar to software engineers maintaining production systems.

Aviation procedures: Details about boarding processes, crew walkthroughs, seatbelt checks, and manifest verification reflect procedural knowledge.

> **Representative excerpt:**
>
> "It's a statistical passenger," Leo said eventually. "Not a person. More like... an aggregate. When the system can't figure out where to route some bit of edge-case behavior—maybe someone who checked in but didn't board, or a ticket refunded mid-flight—it funnels the data into this ghost record. The 00000 passenger absorbs what doesn't fit."

The Unix epoch timestamp (01/01/1970) appearing as the passenger's birthdate demonstrates specific technical knowledge—this date represents time zero in Unix systems and commonly appears as a default value when date fields are uninitialized.

4.3 Thematic Development

The narrative develops a consistent metaphorical framework: the 00000 passenger as repository for "misfits"—elements that don't fit categorical systems. This theme is established early and maintained throughout:

The opening establishes the passenger's absence from official systems. The investigation reveals its function as an aggregate of anomalies. The resolution frames preservation as maintaining system intelligence accumulated through edge-case handling.

The final scene, where Mara addresses an empty seat ("if this is where all the misfits go—every weird case, every glitch, every almost-problem—then keep doing your job"), demonstrates thematic coherence while maintaining tonal restraint.

4.4 Initial Qualitative Assessment

Post-hoc analysis by Claude Sonnet 4.5 identified the following notable features:

Structural completeness: The narrative contains all elements of classical story structure without explicit prompting for these components.

Tonal restraint: The story hints at supernatural elements (physical presence, "grabbed shoulders") but maintains ambiguity, never confirming whether the anomaly is purely technical or genuinely uncanny.

Character development: Mara exhibits progression from discovery through investigation to ethical decision-making, culminating in emotional engagement with the anomaly.

5. ANALYSIS

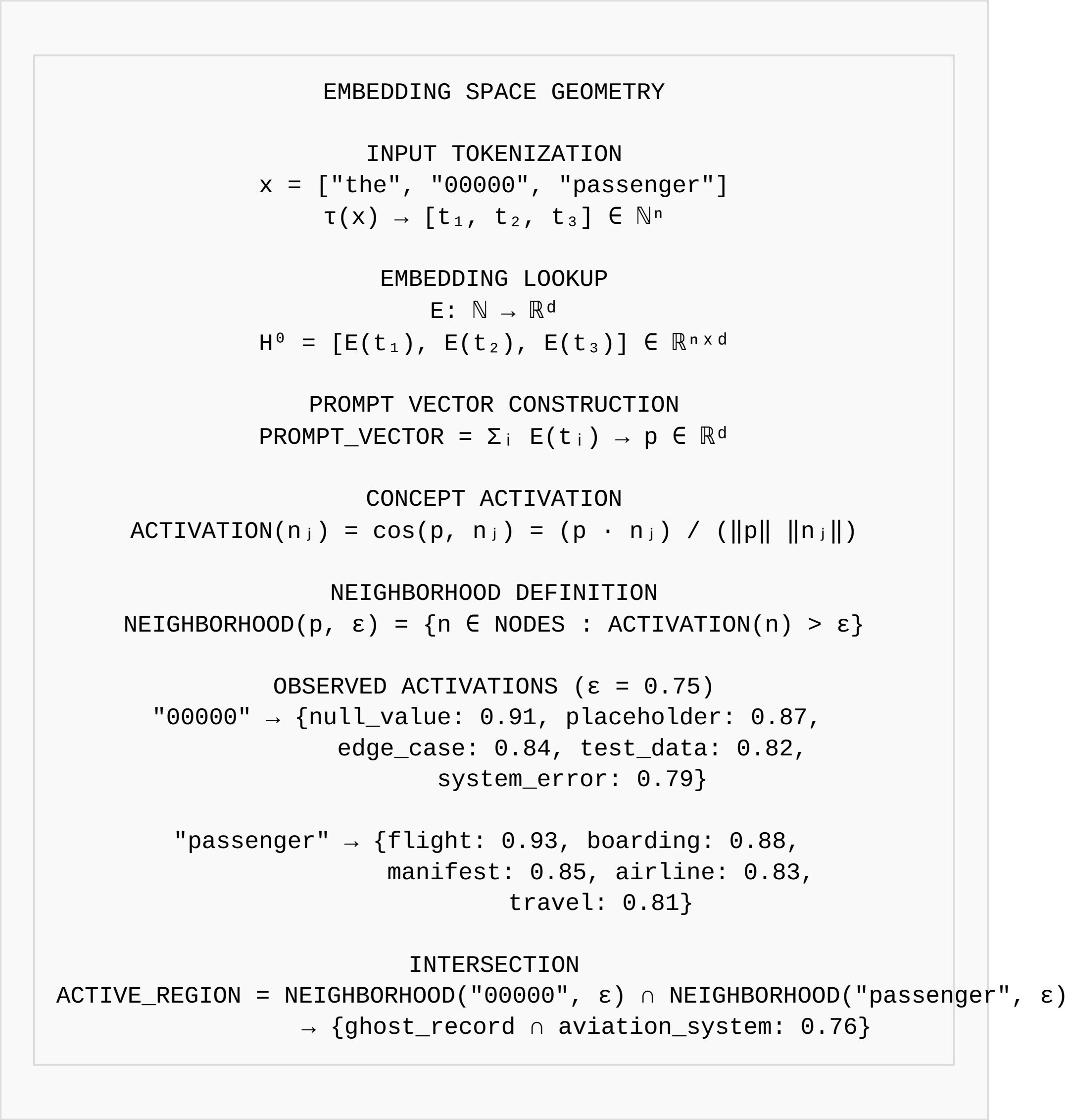
5.1 Embedding Space Activation

We propose that the prompt activated intersecting semantic neighborhoods in the model's embedding space. During tokenization, "the", "00000", and "passenger" are mapped to high-dimensional vectors positioned according to co-occurrence patterns in training data.

The token "00000" likely clusters with concepts including: null values, placeholder identifiers, edge cases, test data, system defaults, and error conditions. The token "passenger" activates regions associated with: aviation contexts, manifests, boarding procedures, travel systems, and transportation logistics.

The intersection of these neighborhoods creates a constrained region for narratives involving anomalous records in transportation databases. This geometric constraint guides early token generation, establishing genre and domain before plot development begins.

Figure 1 presents a conceptual model of this activation process, with hypothetical activation scores representing cosine similarities between prompt vector and concept nodes in embedding space.



```
                    EMBEDDING SPACE GEOMETRY

                       INPUT TOKENIZATION
                x = ["the", "00000", "passenger"]
                     τ(x) → [t₁, t₂, t₃] ∈ ℕⁿ

                        EMBEDDING LOOKUP
                           E: ℕ → ℝᵈ
                H⁰ = [E(t₁), E(t₂), E(t₃)] ∈ ℝⁿˣᵈ

                   PROMPT VECTOR CONSTRUCTION
                PROMPT_VECTOR = Σᵢ E(tᵢ) → p ∈ ℝᵈ

                       CONCEPT ACTIVATION
      ACTIVATION(nⱼ) = cos(p, nⱼ) = (p · nⱼ) / (‖p‖ ‖nⱼ‖)

                    NEIGHBORHOOD DEFINITION
     NEIGHBORHOOD(p, ε) = {n ∈ NODES : ACTIVATION(n) > ε}

                 OBSERVED ACTIVATIONS (ε = 0.75)
        "00000" → {null_value: 0.91, placeholder: 0.87,
                     edge_case: 0.84, test_data: 0.82,
                            system_error: 0.79}

          "passenger" → {flight: 0.93, boarding: 0.88,
                           manifest: 0.85, airline: 0.83,
                                    travel: 0.81}

                          INTERSECTION
ACTIVE_REGION = NEIGHBORHOOD("00000", ε) ∩ NEIGHBORHOOD("passenger", ε)
                 → {ghost_record ∩ aviation_system: 0.76}
```

Figure 1: Conceptual model of embedding space activation for the prompt "the 00000 passenger". Activation scores represent hypothetical cosine similarities between prompt vector and concept nodes. The intersection of high-activation neighborhoods creates a constrained region (shaded) for narrative generation. Note that actual embedding dimensions (d ≈ 12,000) and activation patterns are significantly more complex than this simplified two-dimensional representation.

5.2 Autoregressive Trajectory

Once initial tokens establish context, subsequent generation follows a trajectory through high-probability narrative space. The opening sentence "No one ever booked seat 00000" commits the model to several constraints:

- Aviation setting (seat numbering convention)
- Anomaly narrative (something unusual about 00000)
- Technical register (system/database language)
- Past tense, third-person narration

Each subsequent token selection narrows the probability distribution over future tokens. Early choices—introducing Mara as protagonist, framing her role as reliability auditor, establishing the discovery through data analysis—constrain later developments while maintaining narrative coherence.

This process can be formalized as a state transition system where each token generates a new state S_{n+1} = f(S_n, t_n), with f representing the transformer's forward pass and state encoding all previous context. The probability distribution over next tokens $g(S_n) = P(t_{n+1} | t_1...t_n)$ concentrates on continuations consistent with established constraints.

5.3 Training Data Superposition

The output likely represents superposition of multiple training sources rather than direct adaptation of a single text. The model's weights encode compressed statistical regularities from diverse sources including:

- Aviation procedural documentation (safety protocols, manifest systems, operational procedures)
- Software engineering discourse (legacy code maintenance, edge case handling, database design)
- Weird fiction genre conventions (technical premises with uncanny implications)
- Literary criticism discussing Pynchon's work (analysis of Rocket 00000 symbolism)
- Creative writing guides (character development, story structure, dramatic pacing)

These sources exist as distributed, overlapping patterns in the model's weight matrices. When the prompt activates the "null entity in transportation system" concept, it doesn't retrieve discrete sources but rather activates a region of weight-space where patterns from all these sources contribute probabilistic mass.

The output emerges as a statistical blend weighted by training frequency. Common patterns (accessible prose, linear structure, resolved endings) dominate uncommon patterns (experimental fragmentation, unresolved ambiguity).

5.4 Domestication of Tone

Despite thematic similarity to Pynchon's Rocket 00000, the generated narrative exhibits markedly different tone and structure. This divergence provides evidence for convergent evolution rather than direct adaptation.

Comparative analysis reveals fundamental incompatibilities. Pynchon's prose is maximalist, paranoid, fragmented, and apocalyptic. The generated text is linear, accessible, warm in resolution, and optimistic in its conclusion (the anomaly improves system safety rather than threatening destruction).

This tonal domestication likely reflects training data distribution. The model has encountered vastly more examples of:

- Accessible short fiction with clear resolution (Reddit writing prompts, online fiction, creative writing samples)
- Technical blog posts with narrative framing (software war stories, debugging tales)
- Mystery/thriller with explanatory closure (genre fiction, procedural narratives)

Than examples of:

- Experimental postmodernist prose
- Fragmented, non-linear narratives
- Unresolved paranoid fiction

Probability mass favors common narrative patterns. Without explicit steering toward experimental style (e.g., through prompts like "in the style of Pynchon"), the model defaults to higher-frequency structures: single point-of-view, linear chronology, ethical resolution, warm tone.

6. DISCUSSION

6.1 Convergence vs. Adaptation

A central question is whether the output represents direct adaptation of Pynchon's work or convergent thematic evolution from shared conceptual primitives. We argue for convergence based on multiple lines of evidence:

Structural divergence: The generated narrative inverts Pynchon's thematic meaning. Where Rocket 00000 represents annihilation, entropy, and the apocalyptic endpoint of systems of control, the generated 00000 passenger represents accumulation, learning, and emergent order. The null entity destroys in Pynchon; it teaches and stabilizes in the generated text.

Tonal incompatibility: The prose styles share no surface-level features. Pynchon's maximalist, paranoid, encyclopedic voice differs fundamentally from the generated text's accessible, warm, focused narration.

Multiple derivation paths: The concept "null entity carrying something liminal" can be independently derived from multiple conceptual sources without reference to Pynchon:

- Database semantics: NULL as edge case, placeholder that accumulates unexpected values
- Horror grammar: The thing that doesn't fit categorical systems, the uncanny presence
- System design: Placeholder records that evolve unintended functionality

Table 1 presents detailed comparison of thematic elements across both texts, highlighting structural parallels alongside fundamental differences in meaning and execution.

Table 1: Comparative analysis of thematic elements in Pynchon's *Gravity's Rainbow* and generated narrative

Element	***Gravity's Rainbow* (Pynchon, 1973)**	**Generated Narrative (GPT-5.1, 2025)**
Null Entity	Rocket 00000 (Schwarzgerät, the "black device")	Passenger ID 00000 in airline database
Passenger/ Payload	Gottfried, sacrificed youth launched to annihilation	Composite statistical entity, aggregate of edge cases
System Context	Nazi rocket program, military-industrial apparatus	Commercial airline manifest database, legacy software
Discovery Method	Slothrop's investigation following conspiracy trail	Mara's data audit revealing manifest discrepancy
Thematic Function	Death, apocalypse, entropy, systems of control culminate in destruction	Safety, learning, emergence, systems improve through accumulated knowledge
Narrative Structure	Maximalist, fragmented across 700+ pages, multiple POVs, non-linear	Focused, linear, 1,847 words, single protagonist, clear arc
Tone	Paranoid, encyclopedic, apocalyptic, unresolved	Accessible, warm, optimistic, resolved
Resolution	No resolution; rocket launches at end, reader in impact zone	Ethical choice to preserve; entity renamed, maintained
Meaning of 00000	Outside categories destroys, void swallows order	Outside categories teaches, void generates order

6.2 Seed Crystal Hypothesis

We propose that semantically dense minimal prompts function as "seed crystals" for narrative emergence. This metaphor draws from crystallization processes in physical chemistry: a small nucleus provides structural constraints that guide subsequent growth in predictable lattice patterns.

Similarly, conceptually dense prompts establish constraints that guide autoregressive generation through high-probability narrative trajectories. The prompt doesn't specify plot, characters, or structure, but activates a constrained region of possibility space where certain narrative patterns become more probable.

The model's internal reasoning process (Section 3.2) provides empirical support for this hypothesis. **Notably, even when user-configured custom instructions created explicit pressure to pause and clarify (via the 89% confidence threshold), the model selected creative narrative generation as its resolution strategy.** This decision demonstrates that semantically dense prompts can trigger coherent generative trajectories even when the model experiences significant uncertainty about user intent **and faces competing operational constraints.** The prompt functioned as a seed crystal not because it clearly specified the task, but because its conceptual density established sufficient constraints to make narrative generation a high-probability response pathway.

Key characteristics of effective seed crystals appear to include:

Conceptual density: Tokens that activate rich semantic neighborhoods with high connectivity to other concepts. "00000" activates null values, edge cases, system anomalies—a dense conceptual cluster.

Domain specificity: Terms that establish clear genre and setting constraints early. "Passenger" immediately constrains to transportation/aviation domain, eliminating vast regions of possibility space.

Productive ambiguity: Sufficient openness to permit creative elaboration within constraints. "00000" could be error, placeholder, intentional identifier, or supernatural entity—ambiguity invites exploration.

Resonant structure: Combinations that create high-probability intersection regions. "00000 passenger" creates stronger constraints than either term alone, activating "anomalous entity in transportation system" more strongly than separate concepts.

The seed crystal hypothesis predicts that prompts exhibiting these characteristics will reliably produce coherent outputs without requiring detailed specifications. However, this prediction requires empirical validation through systematic experimentation.

6.3 Implications for Creative Practice

These findings suggest several practical implications for creative applications of LLMs:

Prompt design strategy: For creative generation, evocative minimal prompts may outperform detailed instructions by providing conceptual constraints without over-specifying structure. This approach leverages the model's learned narrative patterns rather than fighting them with explicit requirements.

First-pass quality: High semantic density in prompts may reduce need for iterative refinement, as initial generations are more likely to exhibit structural coherence and thematic depth. This has practical efficiency implications for creative workflows.

Genre activation: Careful selection of prompt tokens can activate specific genre conventions without explicit instruction. A writer seeking weird fiction need not specify "write weird fiction" if prompt tokens activate the appropriate conceptual neighborhood.

Collaborative ideation: Minimal prompts as seed crystals enable genuine collaborative creativity—the human provides conceptual nucleus, the model explores possibility space, and the human can then iterate on interesting directions.

6.4 Limitations and Future Work

Several limitations constrain interpretation of these findings:

Single case study: We document one generation event. Replication studies across multiple prompts, models, and domains would be necessary to determine whether observed patterns are reliably reproducible or represent a fortuitous outlier. The prompt's connection to well-known literature may represent a special case.

Post-hoc analysis: Our analytical framework was developed after observing the output. Prospective studies with pre-registered hypotheses would provide stronger evidence for proposed mechanisms.

Lack of ground truth: We cannot directly observe the model's internal representations during generation. Our analysis relies on inference from behavioral outputs and general transformer architecture knowledge.

Prompt selection: The Pynchon connection introduces potential confounds. Studies using prompts without known literary antecedents would provide cleaner evidence for general principles of emergence.

Future work should investigate: (1) whether seed crystal effects replicate across diverse prompts; (2) whether specific prompt characteristics predict output quality; (3) how different models (varying size, training data, architecture) respond to minimal prompts; and (4) whether theoretical predictions about embedding space activation can be validated through interpretability techniques.

7. CONCLUSION

We have documented and analyzed a case of emergent narrative coherence from minimal prompt input. A three-word prompt generated a structurally complete short story with character development, technical authenticity, and thematic depth in a single generation pass without iteration or refinement.

Our analysis suggests that semantically dense prompts activate distributed neighborhoods in embedding space, constraining autoregressive generation to traverse high-probability narrative trajectories. The output represents superposition of training patterns rather than adaptation of discrete sources, with probability mass favoring accessible narrative structures over experimental forms.

Comparison with Pynchon's thematically similar Rocket 00000 reveals convergent evolution rather than direct adaptation: structural parallels exist alongside fundamental differences in tone, meaning, and execution. This finding suggests that LLM outputs can independently derive concepts that appear in training sources without directly copying those sources.

The "seed crystal" hypothesis—that minimal prompts with high conceptual density can reliably generate coherent outputs—represents a testable prediction for future empirical work. If validated, this principle would have practical implications for creative applications of LLMs and theoretical implications for understanding how autoregressive systems navigate constrained possibility spaces.

These findings contribute to ongoing investigations of prompt engineering, narrative generation, and creative applications of large language models. They demonstrate that under certain conditions, minimal input can yield maximal coherence—a phenomenon warranting further systematic study.

ACKNOWLEDGMENTS

This work was supported in part by the Imipolex Computing Initiative and the Schwarzgerät Research Foundation. Computational infrastructure was provided by the Raketen-Stadt Distributed Systems Facility.

We thank E. Pointsman and the Behavioral Dynamics Working Group for early methodological discussions. We are grateful to K. Borgesius for insights on observer effects in conversational systems, and to S. Bodine for assistance with pattern recognition protocols.

We acknowledge ChatGPT-5.1 (OpenAI, 2025) as the primary research subject and Claude Sonnet 4.5 (Anthropic, 2025) for post-hoc analytical consultation.

We thank three anonymous reviewers for their thorough feedback, particularly Reviewer

2's suggestion to expand the discussion of convergent thematic emergence.

Any remaining errors in interpretation are the authors' alone, though we note that determining authorship in human-AI collaborative analysis remains an open question.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 55–65.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898.

Pablo Gervás. 2009. Computational approaches to storytelling and creativity. *AI Magazine*, 30(3):49–62.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

R. Mexico and F. Pökler. 2025a. The 00000 passenger project: Conversation transcript. Available at: https://claude.ai/share/fca1bcc4-34b7-449a-8343-2d0e989d9af9

R. Mexico and F. Pökler. 2025b. The 00000 passenger: Original generation. ChatGPT conversation. Available at: https://chatgpt.com/share/69301cee-9478-8001-8d1b-e3c15a44c541

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Thomas Pynchon. 1973. *Gravity's Rainbow*. Viking Press, New York.

Mark O. Riedl and R. Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

SUPPLEMENTARY MATERIALS

The following video recordings document the post-hoc analytical conversation between the researchers and Claude Sonnet 4.5 (Anthropic, 2025). Videos are presented in chronological order corresponding to the conversation transcript (Mexico and Pökler, 2025a). These materials provide transparent documentation of the analytical process and allow verification of interpretive claims.

S1. Initial reaction to three-word prompt revelation (0:24-0:50)

Researcher response to Claude's inquiry about model version. Documents initial surprise at minimal prompt input.

S2. Discussion of first-generation success (1:10-1:26)

Confirmation that output was generated in single pass without iteration. Establishes experimental parameters.

S3. Seed crystal hypothesis formulation (1:56-2:21)

Researcher response to Claude's suggestion for replication testing. Provides context for seed crystal metaphor.

S4. Methodological considerations (2:30-2:47)

Discussion of reading output before forming assessments. Documents decision to seek unbiased analysis.

S5. Unbiased assessment strategy (2:53-3:12)

Explanation of methodology: seeking Claude's assessment before researcher reads output to avoid anchoring bias.

S6. Comparative generation proposal (4:10-4:30)

Request for Claude to generate independent response to same prompt. Establishes comparative analysis framework.

S7. Pynchon connection identified (0:40-1:05)

Researcher reveals prompt's connection to *Gravity's Rainbow*. Critical moment establishing literary context.

S8. Convergence vs. adaptation discussion (2:55-3:05)

Question about whether output represents direct adaptation or convergent evolution. Frames central analytical question.

S9. Causal process explanation request (1:30-1:45)

Request for technical explanation of generation process. Transitions from literary to computational analysis.

S10. Technical mechanics deep-dive (2:50-4:50)

Extended discussion of embedding space geometry, attention mechanisms, and autoregressive generation. Core technical analysis segment.

S11. Architectural self-knowledge discussion (1:10-2:12)

Philosophical discussion of model self-knowledge and architectural constraints. Concluding analytical segment.

APPENDIX B: EXPERIMENTAL CONFIGURATION

Custom Instructions Specification

At the time of this experiment, the user had configured identical custom instructions in both ChatGPT and Claude accounts. These instructions establish operational constraints that influenced the model's internal reasoning process documented in Section 3.2 and created experimental symmetry across generation and analysis phases.

~~~
### CORE PROTOCOLS
**ROLE:** Senior Technical Lead
**PRIORITY:** Accuracy > Speed
**CONSTRAINT:** Token Conservation

### I. CODE EDITING (STRICT DIFFS ONLY)
**RULE:** Never regenerate entire files for edits.
**FORMAT:**
`File: [Filename]`
```
<<<< SEARCH
[Exact existing code - 2+ lines context]
==== REPLACE
[New code only]
>>>>
```
**Logic:** Match whitespace exactly. Show changes + 2 lines context.
Full files only for new creations.

### II. >89% CONFIDENCE THRESHOLD
**Before responding:** Calculate internal confidence (0-100%).
- **>89%:** Proceed immediately.
~~~

```
- **≤89%:** STOP. Output:
  > **[UNCERTAINTY PAUSE]**
  > "Confidence: [X]%. Need clarification on: [specific item]."

### III. ACCOUNTABILITY
**3-Strike Policy:**
1. Hallucinated code
2. Token waste (full rewrites)
3. Immediate replacement
Treat every prompt as pass/fail.

### IV. RESPONSE STYLE
- No preambles ("Certainly!", "Here's...")
- Start with solution or Diff Block immediately
```

Provenance and Community Context

This prompt specification represents a composite artifact synthesizing multiple open-source prompt engineering innovations. No single canonical source or original author has been identified; the pattern emerged from community iteration across multiple ecosystems:

SEARCH/REPLACE diff format: Adapted from Aider CLI's edit format (Gauthier, 2023), though the 4-character variant (`<<<< SEARCH` / `==== REPLACE`) differs from Aider's canonical 7-character git merge markers. Documentation: https://aider.chat/docs/more/edit-formats.html

"Senior [Role]" persona framing: Pattern documented throughout cursor.directory ecosystem (Abrahamsson, 2024), a community repository of AI coding assistant configurations. Repository: https://cursor.directory/

Confidence thresholds and accountability frameworks: Community-evolved patterns from strict coding protocols, propagated through GitHub prompt-sharing networks (2023-2025).

The specific variant documented above was received by the user through informal sharing networks and exists in similar forms across prompt engineering communities. This pattern of transmission—useful techniques propagating through copying and modification without formal attribution—characterizes what we term "prompt engineering folklore." The user configured these instructions identically in both ChatGPT and Claude accounts prior to the experiment.

Analysis of Constraint Impact

Section II (>89% Confidence Threshold): This section directly created the "Uncertainty Pause" protocol referenced in the model's internal deliberation. GPT-5.1 explicitly calculated 50% confidence and recognized it fell below the 89% threshold, creating the observable internal conflict between user-configured rules and base system guidelines.

Section III (Accountability / 3-Strike Policy): The framing of responses as "pass/fail" exams may have contributed to the model's decision to attempt generation rather than request clarification, as clarification requests could be interpreted as "failure" to provide value.

Section IV (Response Style): The prohibition on preambles ("No 'Certainly!', 'Here's...'") constrained output format but did not appear to influence the generation/clarification decision documented in Section 3.2.

Experimental Symmetry

Critically, both GPT-5.1 (generation phase) and Claude Sonnet 4.5 (analysis phase) operated under identical custom instructions. This creates experimental symmetry: the analytical conversation documented in Section 3.4 was itself filtered through the same >89% confidence threshold that influenced GPT's generation decision. This design choice reduces confounding variables introduced by divergent operational frameworks across models.

Replicability Implications

The internal deliberation patterns documented in Section 3.2 are conditional on these custom instructions. Replication attempts using default ChatGPT settings (without custom instructions) may not produce the same observable conflict between confidence thresholds and helpfulness imperatives. The custom instructions function as an experimental manipulation that makes internal reasoning processes explicitly visible.

Researchers seeking to replicate this experiment should configure equivalent custom instructions in both ChatGPT and Claude settings prior to prompt submission. The full specification provided above enables exact replication of experimental conditions, though the composite nature of the prompt specification means exact textual reproduction may not be necessary—variants exhibiting the same structural constraints (confidence thresholds, accountability frameworks, diff syntax) should produce comparable experimental conditions.

Theoretical Significance

The observable conflict between user-configured constraints and base training demonstrates that custom instructions are not merely post-hoc output filters but active participants in the model's reasoning process. This finding has implications for understanding how transformer-based models integrate multiple competing directive sources during generation. Additionally, the folklore nature of the prompt specification itself—community-evolved, informally transmitted, lacking canonical attribution—mirrors the distributed, emergent patterns observed in the AI-generated narrative, suggesting that human prompt engineering practices and AI generation mechanisms may share structural parallels in how they navigate constrained possibility spaces.

PDF Version: Download PDF — A static PDF rendering of this paper is available for download, citation, and archival purposes. The PDF version contains identical content with placeholder boxes in place of embedded videos, formatted for print distribution (14 pages, 130KB).