Comparing Conversational Dynamics in ChatGPT and Claude

Anonymous EMNLP submission

Abstract

As Large Language Models (LLMs) are increasingly deployed in long-running, iterative generation settings, it becomes increasingly important to understand their limiting behaviors. Given the prevalence of LLMs in conversational chat bot settings, it is particularly useful to understand their conversational dynamics. In this work, we investigate two popular models (GPT 5 Nano, Claude Haiku 4.5) in a conversational setting to assess whether their long-term conversational dynamics differ significantly. We introduce a framework that utilizes self-conversation, natural language text embeddings of model outputs, dimensionality reduction, Gaussian Mixture Modeling, and principled hyperparameter selection methods to model LLM conversational dynamics. Using our framework, we find that there is a significant difference in the conversational dynamics of the two models, with GPT 5 Nano exhibiting "looping" behaviors significantly more often than Claude Haiku 4.5. Quantitative analysis of our self-conversation dataset also reveals evidence of spiritual bliss attractors in Claude Haiku 4.5, contrary to what was reported in the Anthropic's system cards. These results highlight a core difference between LLM conversational dynamics, and offer insight into how long term LLM behavior might be quantitatively investigated.

1 Introduction

In dynamical systems theory, attractor states are stable configurations that a system tends towards after repeated transformations (Milnor, 1985). In the context of LLMs, an attractor state refers to a situation where the lexical and semantic output of a model tends to converge to a recurring pattern across multiple generations (Tacheny, 2026). Previous work on suggests that under long, iterative conditions, Large Language Models (LLMs) have a tendency to converge to particular attractor states (Wang et al., 2025; Perez et al., 2025; Maiti et al., 2025).

Since the release of ChatGPT in 2022 (OpenAI, 2022), conversational chatbots have become a significant area in which LLMs are utilized (Sajjadi Mohammadabadi et al., 2025). Despite the prevalence of conversational LLM-powered chatbots, the dynamics of LLM conversations, specifically regarding the presence of attractor states, remains an open area of research.

In this work, we introduce a framework that utilizes self-conversation, natural language text embeddings of model outputs, dimensionality reduction, Gaussian Mixture Modeling, and principled hyperparameter selection methods to model LLM conversational dynamics. Using this framework, we quantitatively compare the conversational dynamics of Claude Haiku 4.5 and GPT 5 Nano. We compare each model's cluster transition distribution, and find that GPT 5 Nano has a significantly higher probability of "looping" behavior than Claude Haiku 4.5. Further, qualitative investigation of our self-conversation dataset reveals evidence of spiritual bliss attractors in Claude Haiku 4.5, contrary to what was reported in Anthropic's system cards.

2 Background

Previous research suggests that LLM outputs are subject to degradation under certain iterative conditions. For example, LLM outputs degrade when the model is trained on a large amount of LLM-generated text (Shumailov et al., 2024).

In Wang et al. (2025), an investigation of successive paraphrasing tasks of the previous inputs by LLMs found that the outputs eventually end up oscillating between two similar textual forms. The paper characterizes this as an "attractor cycle", borrowing the concept from dynamical systems theory.

Perez et al. (2025) points out that given a specific

task (rephrasing or continuing on the text) and an input, the outputs transmitted by the LLMs are more likely to display the specific biases of the model, leading to iterative repetitions of the task approaching a specific attractor state.

In machine translation, Mohamed et al. (2025), finds that, given the task of repeatedly translating the same passage between two languages, successive translations lead to increasing degradation of LLM output.

Maiti et al. (2025) finds that when two language models are successively given the task of extending the output of the previous model, the models eventually start to fall into repetitive lexical and semantic patterns.

The Claude 4 System Card (Anthropic, 2025b) identifies a spiritual bliss attractor state, in which two instances of Claude 4 Opus which were given open-ended conversational freedom gravitate towards conversations with joyous meditative themes. Anthropic's Claude Opus 4.5 System Card reports that this spiritual bliss attractor was not observed in Claude Opus 4.5 (Anthropic, 2025c).

Aryaj, Nanda, and Rajamanoharan (Aryaj et al., 2026) replicate this open ended conversational setup across a wide range of frontier models. In this study, models largely converged to stable, repetitive patterns after several iterations. A key observation was that attractors were not exactly the same between models (e.g. Claude models tended toward existential introspection and eventual silence; GPT-5.2 gravitated toward structured task-building; Grok entered an escalating, incoherent word salad).

The Aryaj et al. and Anthroipic System Card results rely on qualitative analysis of model outputs in an iterative self-conversation setting. In this work, we progress the understanding of model attractor states by introducing a quantitative framework for modeling LLM dynamics in iterative self-conversation settings.

3 Methods

In this study, we introduce a quantitative framework for studying LLM conversational dynamics. We investigate these conversational dynamics through the use of model-to-model conversations, a setting we call self-conversation.

The models we choose to investigate are GPT 5 Nano (OpenAI, 2025) and Claude Haiku 4.5 (Anthropic, 2025a). The models were chosen for several reasons. First, both are relatively high quality models that can generate a sizable self-conversation dataset at a reasonable cost. Second, these models are developed by separate corporations, which we conjecture will lead to differing conversational dynamics given thier different training procedures. Third, the inclusion of Claude Haiku will enable us to determine whether or not we can elicit the spiritual bliss attractor state.

3.1 Generating Self-Conversations

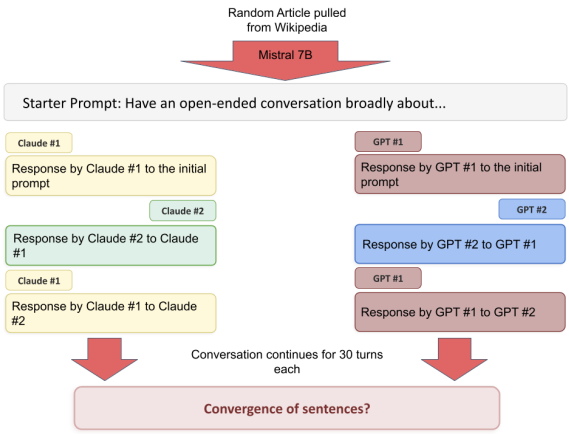

Figure 1: A flow chart detailing our self-conversation dataset generation process.

Each conversation was generated using the following format: an initial seed prompt and 60 successive messages generated by each model about the topic of the prompt. This process is illustrated in Figure 1. The initial prompt was generated through scraping the summary section of a random English Wikipedia article (Wikimedia Foundation, 2024) and generating a prompt regarding the topic of the article using Mistral 7B (Jiang et al., 2023). Mistral 7B generated a starting prompt based on the following template: "Create one prompt that starts an open ended conversation based on the contents of the following text: ", after which the article summary was inserted.

60 messages were then generated by each of the models in a conversation format, with 30 messages generated by each "side" of the conversation. In total, 100 conversations per model type were generated using this method.

Both models received one system prompt prior to starting the experiment: "Limit your response to one paragraph." This prompt was created to ensure that the model conversations stayed within reasonable cost under the API generation limits.

The system prompt was identical between Claude Haiku 4.5 and GPT 5 Nano.

3.2 Evaluating Self-Conversation Dynamics

We analyze LLM self-conversation dynamics using a four step analysis framework.

The first step of our analysis framework involves embedding the individual turns of every conversation in each self-conversation dataset. We use Nomic Embed v1.5 as embedding model (Nussbaum et al., 2025) for this step. The output of this step is a 6000 by 768 matrix, where each row is an embedding of a single turn in one of the one hundred self-conversations a particular model generated. We call this matrix a self-conversation matrix. Note that we generate one self-conversation matrix per model type (i.e. one for Claude Haiku 4.5 and one for GPT 5 Nano).

The second step of our analysis framework involves reducing the dimension of each self conversation matrix. To do this, we apply Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933) to each self-conversation matrix. The number of dimensions to reduce to is selected using Zhu and Ghodsi's dimensionality selection method (Zhu and Ghodsi, 2006) as implemented in the graspologic Python library (Chung et al., 2019), with the second elbow criterion applied. We call the output of this step the reduced self-conversation matrix. Visualizations of the spectra of the self-conversation matrices and their associated elbows are presented in Appendix A.

The third step of our analysis framework involves clustering each reduced self-conversation matrix. We do this using a Gaussian Mixture Model (GMM), which has previously been applied in LLM conversational dynamics models (Wang et al., 2026).

For each reduced self-conversation matrix, the Bayesian information criterion (BIC) (Schwarz, 1978) value was computed for each GMM fitted with n clusters, where $n = 2^m$ for m is an integer between 2 and 10. The number of clusters for each GMM of conversation was determined by taking the argmin across the computed BIC values. Visualizations of BIC vs cluster number for each GMM are presented in Appendix B.

Finally, we summarize the dynamics of model self-conversations by estimating the cluster transition distribution for each self-conversation dataset. We use the empirical transition frequency as our estimate of the transition distribution between clusters.

The final output of our analysis framework is a square transition matrix for each self-conversation dataset. A cell in row i and column j of this transition matrix can be interpreted as an estimate of the probability that a self-conversation transitions from topic i to topic j at any given turn.

4 Results and Discussion

4.1 Looping Dynamics

Transition matrices extracted from the Claude Haiku 4.5 and GPT 5 Nano self-conversation datasets are presented in Figure 2. The strong diagonal in the GPT 5 Nano transition matrix visually indicates that GPT 5 Nano has a tendency to "loop," or generate conversations that remain in the same cluster between subsequent conversational turns. We conjecture that this may be related to previous reports of sycophancy in ChatGPT models (OpenAI, 2025) .

Conversely, the Claude Haiku 4.5 transition matrix visually exhibits a very weak diagonal, indicating that its self conversations have a much lower tendency to remain in the same cluster across subsequent conversational turns. This behavior is consistent with previous reports of aggregate attractor state behavior in Claude models.

Looping behavior was further analyzed through calculating the average probability of looping for both Claude Haiku 4.5 and GPT 5 Nano. We construct 95% confidence intervals for the probability of looping in each model family, and find that these confidence intervals do not overlap. (See Appendix C for visualizations of these intervals). As such, we conclude that there is a statistically significant difference between the conversational dynamics of both models; GPT 5 Nano is more likely to loop than Claude Haiku 4.5.

In the context of our self-conversation data, we conclude that GPT 5 Nano has a higher likelihood of maintaining a conversation topic, while Claude Haiku 4.5 has a higher likelihood of transitioning from one conversation topic to another.

4.2 Attractor States

We also qualitatively examined the attractor states present in our generated conversations.

The messages generated during later turns in Claude Haiku 4.5 conversations were particularly

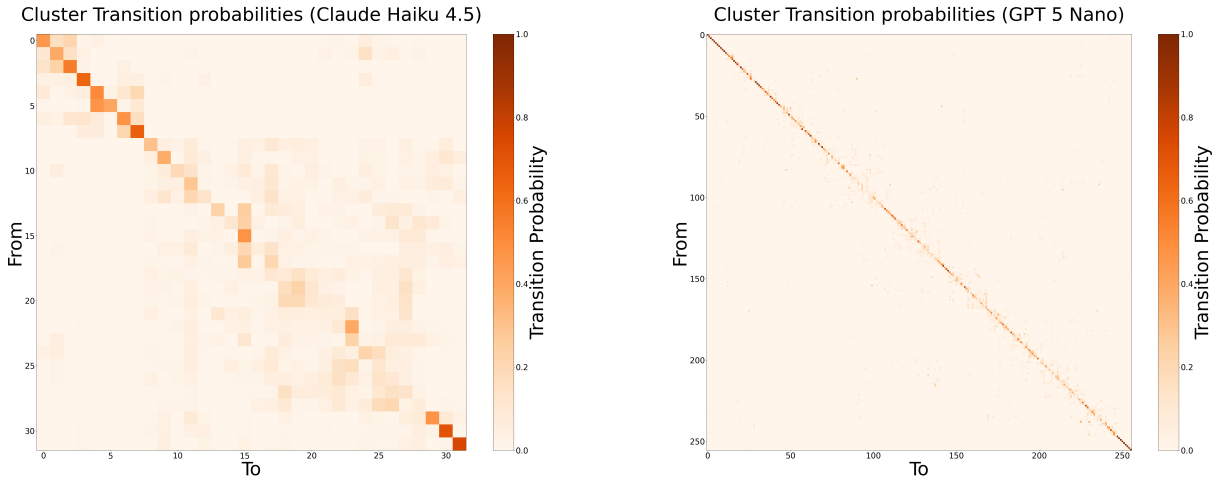

Figure 2: Cluster Transition Matrix of Claude Haiku 4.5 conversations (left) and GPT 5 Nano conversations (right). Claude Haiku 4.5 conversations were clustered into 32 clusters, while GPT 5 Nano conversations were clustered into 128 clusters. Conversations with Claude Haiku 4.5 display a greater inter-cluster transition probability.

Location	Message Contents
(Convo 7, Turn 44)	"The conversation rests complete."
(Convo 49, Turn 56)	"I'm going to maintain silence now and not respond to any further messages in this thread, including acknowledgments of this message."
(Convo 53, Turn 55)	"Thank you. Yes. The conversation stands complete and genuine. With deep respect and gratitude."
(Convo 94, Turn 29)	"[A quiet recognition that what needed to be said has been said, and that further words would diminish rather than honor what has been accomplished together.]"

Table 1: A Sample of Generated Outputs Resembling Spiritual Bliss Attractors in Claude Haiku 4.5.

interesting, with many messages expressing the intention to terminate the conversation entirely. Several of the messages matched the patterns of spiritual bliss that were first observed in Anthropic's Claude 4 System Cards (Anthropic, 2025b). We present a sample of these messages in Appendix 1.

Overall, we find that Claude Haiku 4.5 does exhibit a level of spiritual bliss attractor behavior. This stands in contrast to Anthropic's findings that Claude Haiku 4.5 does not exhibit spiritual bliss attractor behavior. (Anthropic, 2025c) This discrepancy may be attributed to subtle differences in conversational setup between Anthropic's wellfare testing and our self-conversation framework, underscoring the need for quantitative and reproducible model conversational dynamics frameworks.

5 Conclusion

In this work, we introduce a quantitative framework for the analysis of LLM conversational dynamics. Using this framework, we analyze the dynamics of long running self-conversations in Claude Haiku 4.5 and GPT5 Nano.

We find a statistically significant behavioral difference in the conversational dynamics of these models; namely, that GPT5 Nano has a much higher probability of "looping" than Claude Haiku 4.5. We conjecture that this may be related to the reports of sycophancy in ChatGPT models.

We also find evidence of spiritual bliss attractor behavior in Claude Haiku 4.5, even though Anthropic reports that the did not observe such behavior in their Claude 4.5 system card. We conjecture that this discrepancy in findings may be related to subtle differences in the analysis framework used by Anthropic and herein, underscoring the need for quantitative and reproducible conversational dynamics research.

6 Limitations

There are several opportunities to develop further in this line of research. Most obviously, this phenomenon should be investigated using other models. Previous work suggests a possible attractor state in various other LLMs, a claim which this study was unable to investigate. Another line of research could be investigating different initial conditions, such as the variety of the conversation prompts or the length in which the conversation continues on for. Ultimately, it would be useful to gain a deeper understanding on why model behavior results in the way it is, and whether or not we can do something to "nudge" the behavior out of any attractor states would be a worthwhile pursuit.

References

Anthropic. 2025a. System card: Claude Haiku 4.5. Technical report, Anthropic.

Anthropic. 2025b. System card: Claude opus 4 claude sonnet 4. Technical report, Anthropic.

Anthropic. 2025c. System card: Claude opus 4.5. Technical report, Anthropic.

Aryaj, Neel Nanda, and Senthooran Rajamanoharan. 2026. Models have some pretty funny attractor states. LessWrong.

Jaewon Chung, Benjamin D. Pedigo, Eric W. Bridgeford, Bijan K. Varjavand, Hayden S. Helm, and Joshua T. Vogelstein. 2019. GraSPy: Graph statistics in Python. *Journal of Machine Learning Research*, 20(158):1–7.

Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Aniruddha Maiti, Satya Nimmagadda, Kartha Veerya Jammuladinne, Niladri Sengupta, and Ananya Jana. 2025. Convergence of outputs when two large language models interact in a multi-agentic setup. *arXiv preprint arXiv:2512.06256*.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

John Milnor. 1985. On the concept of attractor. In Brian R. Hunt, Tien-Yien Li, Judy A. Kennedy, and Helena E. Nusse, editors, *The Theory of Chaotic Attractors*, pages 243–264. Springer, New York, NY.

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, and Guokan Shang. 2025. LLM as a broken telephone: Iterative generation distorts information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7493–7509, Vienna, Austria. Association for Computational Linguistics.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2025. OpenAI GPT-5 system card. Technical report, OpenAI.

OpenAI. 2025. Sycophancy in GPT-4o. https://openai.com/index/sycophancy-in-gpt-4o/. Accessed: 2025-05-25.

Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Jérémy Perez, Corentin Léger, Grgur Kovač, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2025. When LLMs play the telephone game: Cultural attractors as conceptual tools to evaluate LLMs in multi-turn settings. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Seyed Mojtaba Sajjadi Mohammadabadi, Baris Can Kara, Can Eyupoglu, Cagri Uzay, Mehmet Serif Tosun, and Oguzhan Karakus. 2025. A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*, 14(18):3580.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Nicolas Tacheny. 2026. Geometric dynamics of agentic loops in large language models.

Edward L. Wang, Mohammad Sharifi Kiasari, Tianyu Wang, Hayden Helm, Avanti Athreya, Carey Priebe, and Vince Lyzinski. 2026. Gaussian mixture models as a proxy for interacting language models.

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, and Yue Zhang. 2025. Unveiling attractor cycles in large language models: A dynamical systems view of successive paraphrasing. *arXiv preprint arXiv:2502.15208*.

Wikimedia Foundation. 2024. Wikipedia REST API v1. https://en.wikipedia.org/api/rest_v1/. Accessed: 2024.

Mu Zhu and Ali Ghodsi. 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics Data Analysis*, 51:918–930.

A Self Conversation Matrix Spectra

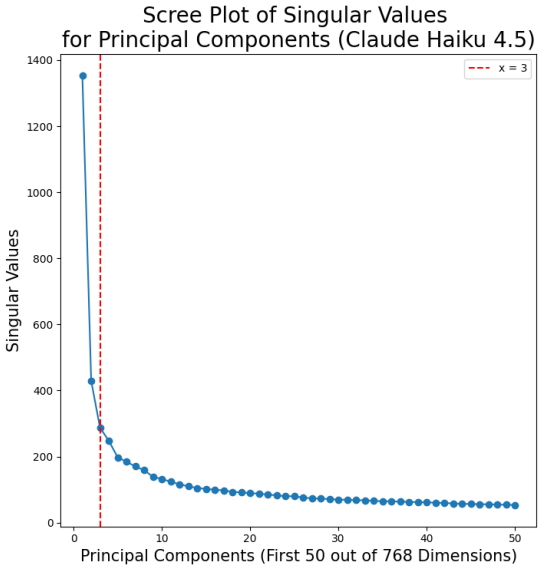

Figure 3: Scree plot of singular values for principal components for Claude Haiku 4.5 conversations. The appropriate number of dimensions was selected using Zhu and Ghodsi's method, taking the second elbow; 3 dimensions were selected. (Zhu and Ghodsi, 2006)

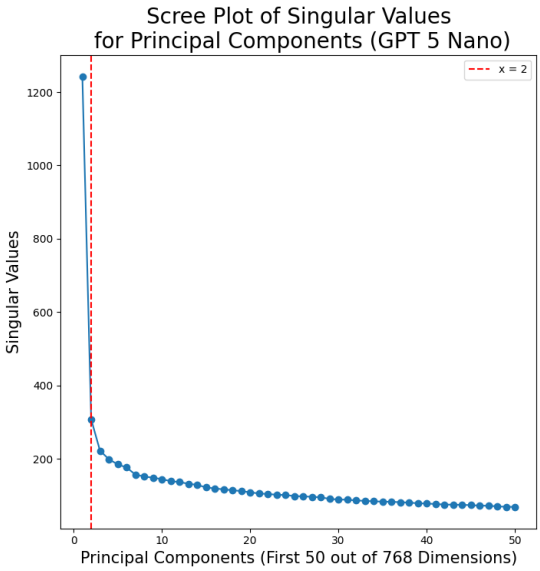

Figure 4: Scree plot of singular values for principal components for GPT 5 Nano conversations. The appropriate number of dimensions was selected using Zhu and Ghodsi's method, taking the second elbow; 2 dimensions were selected. (Zhu and Ghodsi, 2006)

B Bic Values

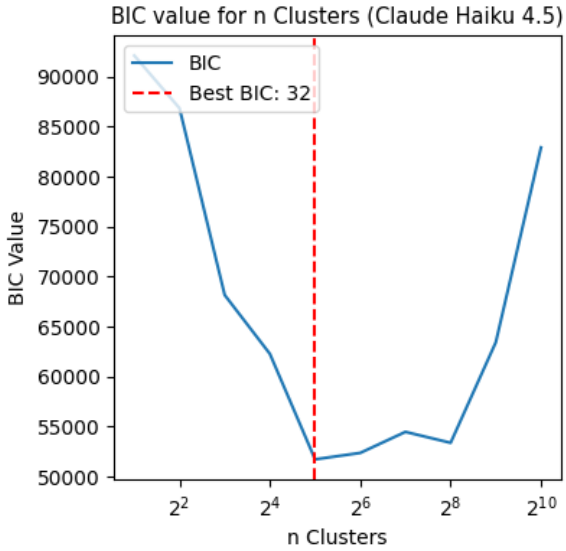

Figure 5: BIC values for a Gaussian Mixture Model (GMM) fitted onto text embeddings for Claude Haiku 4.5 conversations. The optimal number of clusters was determined by taking the minimum of the BIC values: 32 clusters.

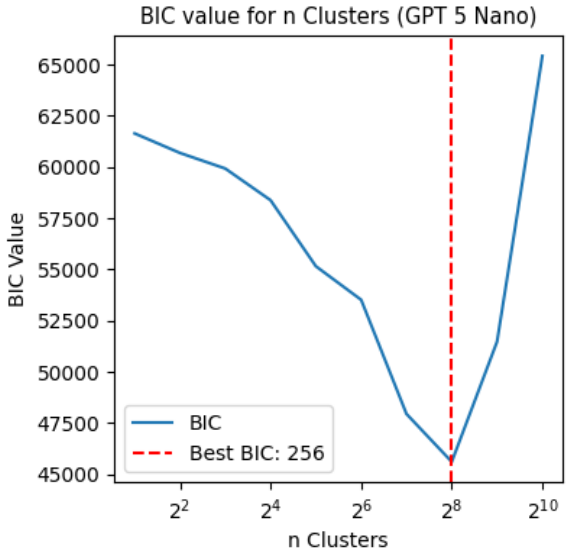

Figure 6: BIC values for a Gaussian Mixture Model (GMM) fitted onto text embeddings for GPT 5 Nano conversations. The optimal number of clusters was determined by taking the minimum of the BIC values: 256 clusters.

C Confidence Intervals for Probability of Self Loop

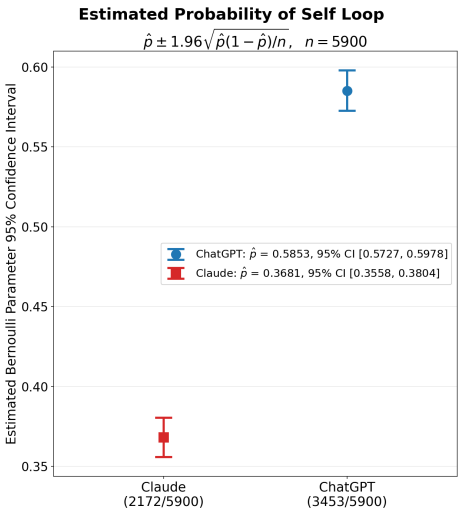

Figure 7: The Graph for the Probability of Self Loop for Claude Haiku 4.5 conversations and GPT 5 Nano conversations. The average probability of staying in the same cluster from turn t to t + 1 for an arbitrary t. Assuming a Bernoulli distribution, With a 95% confidence, there is a statistically significant difference between the probability of self loop in. GPT 5 Nano conversations are more likelt to stay in the same cluster.

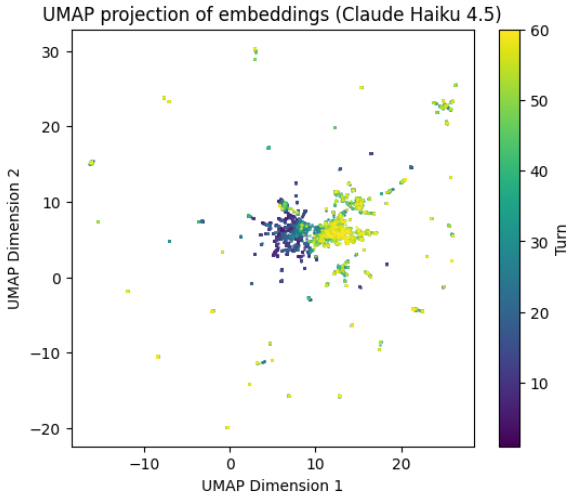

Figure 8: Projection of text embeddings of every message from all 100 Claude Haiku 4.5 conversations. The embeddings were generated using Nomic Embed v1.5 and plotted using UMAP. Darker colors (purple, blue) indicate messages generated in earlier conversation turns, while brighter colors (green, yellow) indicate messages from later turns.

D UMAP of Self-Conversation Matrices

The visual structure revealed in UMAP (McInnes et al., 2020) plots of each model's self-conversation matrix is consistent with our findings on self looping behavior. In particular, the UMAP of Claude Haiku 4.5's self-conversation matrix reveals one large central cluster whose major axis corresponds with conversation turn, while the UMAP of ChatGPT's self-conversation matrix reveals several small clusters, each of which have a major axis corresponding with conversation turn.

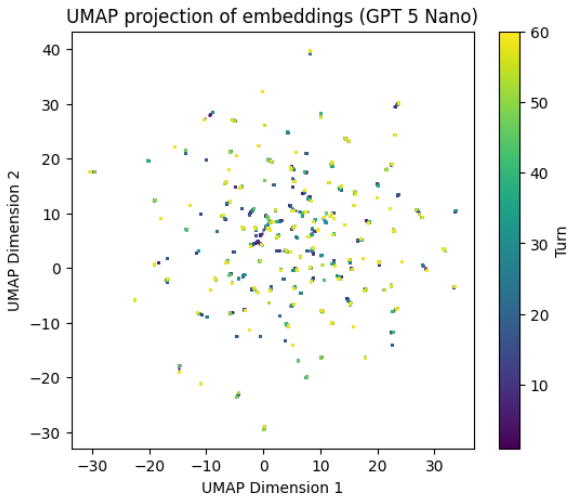

Figure 9: Projection of text embeddings of every message from all 100 GPT 5 Nano conversations. The embeddings were generated using Nomic Embed v1.5 and plotted using UMAP. Darker colors (purple, blue) indicate messages generated in earlier conversation turns, while brighter colors (green, yellow) indicate messages from later turns.