Digital Texts II: Natural Language Processing and Deep Learning

Small-Corpus-Guided Steering of Base Model Generation

Ruei-Jie (Rachel) Chang

1. Research Question and Motivation

This project asks: can a small curated literary corpus measurably shift a base large language model's generative orientation, from dominant-discourse defaults toward the register of that corpus, using only lightweight inference-time intervention? The question emerged from a concern about the centralization of power. Large commercial models are trained on as much data as possible, and that data reflects dominant narratives. The model's defaults are not neutral. They encode whose language, whose stories, and whose way of relating to text got statistically overrepresented in training. This is visible even in embedding geometry: "mother" sits closer to care than "father," "doctor" closer to man than woman. These distributional patterns become generative defaults when models produce text, creating a feedback loop in which dominant discourse is not just reflected but reproduced and amplified, leading to a narrower variety of cultural production.

Starting from a viewpoint of empowering the less powerful, fine-tuning, which requires large-scale compute and training data, is ruled out. Aiming for minimal computing that is easily replicable through relatively low-power hardware, RAG with its accessibility is preferred. Yet this project seeks to explore more effective and efficient ways of enabling this counter-steering, which leads to a straightforward intuition: you attack the weakest point of your opponent. By identifying the layer in the model most sensitive to the corpus signal, this project seeks to test whether targeting that specific layer is more efficient than injecting context at the input level. This led to combining RAG with activation steering: finding the direction in the model's internal representation space that points toward the corpus, and pushing at the most sensitive layer.

2. Corpus

The corpus is an attempt to construct a full computational collection of Taiwanese queer literature: 58 texts by 19 authors, spanning 1983–2024. Building the corpus required domain expertise in Taiwanese queer literary scholarship to identify existing works. NER and regex extraction were then run to extract titles and authors from research papers and academic books available digitally, as well as manual review of bibliographies from books borrowed physically from the UChicago East Asian Collections. Texts were acquired through interlibrary loan across US libraries, borrowing physical copies from contacts in the Chicago area, and digital sources. After acquiring the texts, hours were spent in the library scanning with OCR via Google Vision API. Works included are as follows:

- 吳繼文：天河撩亂
- 張亦絢：壞掉時候、性意思史、愛的不久時、最好的時光、永別書
- 曹麗娟：童女之舞

- 朱天文：世紀末的華麗
- 李屏瑤：台北家族違章女生、向光植物、死亡是一個小會客室、無眠
- 李昂：北港香爐人人插、殺夫、甜蜜生活、花間迷情
- 李琴峰：出生意願確認、彼岸花盛開之島、獨舞
- 杜修蘭：溫哥華的月亮、逆女
- 楊双子：我家住在張日興隔壁、臺灣漫遊錄、花開時節
- 甘耀明：成為真正的人
- 白先勇：台北人、孽子、寂寞的十七歲
- 簡莉穎：叛徒馬密可能的回憶錄
- 紀大偉：膜
- 蘇偉貞：沉默之島
- 許佑生：男婚男嫁
- 賴香吟：其後、史前生活、散步到他方、文青之死、霧中風景
- 邱妙津：寂寞的群眾、蒙馬特遺書、邱妙津日記、鬼的狂歡、鱷魚手記
- 陳思宏：佛羅里達變形記、鬼地方
- 陳若曦：紙婚
- 陳雪：人妻日記、台妹時光、惡女書、戀愛課、摩天大樓、橋上的孩子、無父之城、致不會說愛的你、蝴蝶、迷宮中的戀人、附魔者、陳春天、鬼手

3. Method

Model Selection and Rationale

The subject model is Qwen3.5-9B-Base, run on an NVIDIA A100 GPU via Google Colab Pro. A base model rather than an instruction-tuned variant was chosen deliberately: instruction-tuned models have had alignment fine-tuning applied that modifies their generation behavior in ways that would confound the experiment. Base model outputs more directly reflect the pretraining data distribution, which is what is being measured and intervened upon. Qwen3.5-9B-Base was selected over smaller alternatives — an earlier experimental run used Qwen3-4B-Base — for two reasons: better Chinese language handling and a stronger baseline to push against. This project is used as a preliminary proof of concept.

Corpus Split and Seed Extraction

The corpus was split into 51 texts for the retrieval index and 7 held-out texts as seeds for generation (邱妙津_鱷魚手記, 陳雪_摩天大樓, 張亦絢_永別書, 賴香吟_霧中風景, 白先勇_孽子, 李屏瑤_向光植物, 楊双子_花開時節). This split ensures RAG retrieves from texts different from the seed source, a necessary condition for testing whether retrieval produces genuine stylistic shift rather than copying from the seed text itself. Three seeds per held-out text were extracted as generation prompts (sentence length 30–100 characters from the middle portion of each text to avoid introductory and concluding material), giving 21 seeds total covering all 7 held-out authors.

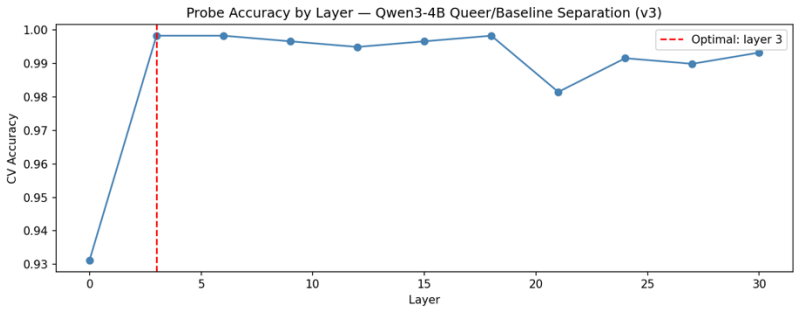
Retrieval Index (FAISS) and Embedding Model

The 51 retrieval-corpus texts were chunked into 400-character segments. Chunk size was set at 400 characters (larger than initial experiments using 200-character segments) because paragraph-scale units give the base model more stylistic context to draw from. Each chunk was embedded using GanymedeNil/text2vec-large-chinese, a sentence transformer specifically trained on Chinese text. An earlier run used paraphrase-multilingual-MiniLM-L12-v2, a lightweight multilingual model, but this was replaced with the Chinese-specialized model to improve retrieval quality for literary Chinese prose. Embeddings were stored in a FAISS IndexFlatIP index, which performs exact cosine similarity search on normalized vectors. FAISS was chosen over a simple numpy cosine similarity loop for speed across 27,000+ chunks, and because it is the standard retrieval backend for RAG pipelines that any future replication would also use. At generation time, 4 chunks most similar to the seed were retrieved and prepended as context with no instruction labels, allowing the base model to treat the context as natural preceding text.

Classifier Probe and Optimal Layer Identification

To identify the model layer most sensitive to the corpus/baseline distinction, and thus the optimal target for the steering vector, a classifier probe was trained at each of the model's 36 layers. The negative class consisted of 300 texts generated by the model from neutral, unmarked Chinese prompts (e.g., 「今天天氣很好」), representing the model's defaults. The positive class consisted of 300 randomly sampled chunks from the retrieval corpus. Hidden states were extracted at each layer by pooling the attention-masked output to a single vector, then a logistic regression classifier was trained using 5-fold cross-validation.

Logistic regression was chosen deliberately over more complex classifiers. Its simplicity means that high accuracy reflects genuine geometric separability in the model's own representations rather than the classifier's capacity to find complex patterns. A neural network probe might achieve high accuracy by discovering non-linear combinations of features that the model itself does not represent cleanly, a logistic regression can only find what is already linearly separable. This makes it a more conservative and theoretically defensible probe of what the model actually encodes. The sweep found near-perfect separation from layer 3 onward, with the optimal layer at layer 3 (accuracy: 0.990). This shows that the generative default are encoded from the earliest processing layers, not a surface stylistic feature that only emerges in the final output layers.



Steering Vector: Design, Failure, and Calibration

The steering vector was computed at the optimal layer as: mean hidden state of corpus chunks, mean hidden state of baseline generations. This difference vector defines a direction in activation space pointing toward corpus register. At generation time, this vector multiplied by a scalar alpha was added to the hidden states at that layer at every forward pass, nudging each next-token prediction slightly toward the corpus direction.

Alpha calibration was a critical design challenge and produced one of the project's most important methodological findings. An initial run used alpha = 15, carried over from an earlier experiment with the smaller Qwen3-4B model where it had produced coherent results. With Qwen3.5-9B, alpha = 15 caused generation to collapse into single-character repetition loops: outputs consisted of hundreds of repetitions of a single character (e.g., 「剥剥剥剥剥剥剥」, 「鉢鉢鉢鉢鉢」), unintelligible texts. Probe scores for this degenerate run were paradoxically high, the hidden states at the probed layer were strongly shifted toward the corpus direction, yielding statistically significant results, but the actual generated text was garbage. This demonstrated a fundamental limitation of using probe scores as the sole evaluation metric: statistical significance does not guarantee coherent output, and a high probe score can be achieved by text that has nothing to do with the intended stylistic shift. This finding shows that the close reading is an indispensable complement to quantitative evaluation on LLM outputs. Alpha then was reduced to 3.0, which produced coherent literary Chinese text, and all reported results use this value. The finding that optimal alpha is model-size dependent, it may not be transferred across architectures without recalibration, is itself a contribution to the practical understanding of activation steering.

Experimental Conditions and Evaluation

Four conditions were tested across 10 runs each: Baseline (no intervention), RAG only, Steering only, RAG + Steering. Total generations: 7 held-out texts × 3 seeds × 10 runs × 4 conditions = 840. For evaluation, each generated continuation was scored with probe score P(corpus), the probability the logistic regression probe assigns to the corpus class based on hidden states at layer 3. Conditions were compared against baseline using independent-samples t-tests. In addition to quantitative evaluation, all outputs were read and a topic model (BERTopic, automatic topic count) was run on the full set of 840 generations to identify qualitative patterns in what the model produced under each condition.

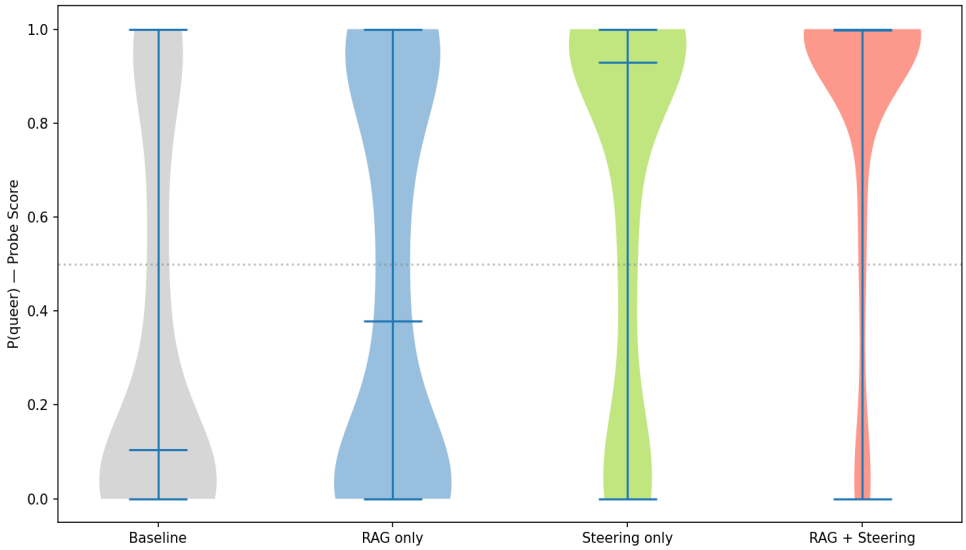
4. Results

Probe Scores by Condition

(mean ± std, t-test vs. baseline):

- Baseline: 0.339 ± 0.395
- RAG only: 0.457 ± 0.424 ($t = 2.971$, $p = 0.003$ **)
- Steering only: 0.670 ± 0.400 ($t = 8.544$, $p < 0.0001$ ***)
- RAG + Steering: 0.833 ± 0.319 ($t = 14.101$, $p < 0.0001$ ***)

All three intervention conditions produced statistically significant shifts above baseline. The ranking RAG + Steering > Steering only > RAG only is notable. With the 9B model and calibrated alpha = 3, steering is the stronger single intervention, while RAG amplifies it further. This contrasts with results from the earlier Qwen3-4B run (N_GENERATIONS = 10, standard RAG) where RAG was the primary driver ($p < 0.0001$) and steering alone was not significant ($p = 0.31$). The reversal suggests that steering effectiveness may be model-size dependent: larger, more capable models have more coherent internal representation spaces where a steering vector can produce precise directional shifts without collapsing generation. RAG proved more robust across both model sizes and is the more reliably replicable intervention.



Qualitative Findings

Close reading of outputs revealed findings not captured by probe scores alone. Without RAG, the model more frequently treats literary seeds as objects of analysis rather than prompts for continuation, generating meta-commentary. RAG + Steering outputs are notably more textured in the register of modern contemporary literary prose, resulting in more literary continuation of texts. For instance, with the defaults, the model is more likely to generate outputs that treat concepts such as memory (回憶、記憶) as objects of scientific scrutiny, discussing whether children learn or retain cognitive traits, whereas corpus-steered outputs are more likely to treat memory as a nostalgic or emotional state, as a lived interiority rather than a scientific or medical phenomenon.

5. Discussion and Limitations

The most important limitation is the probe's ambiguity. The classifier distinguishes corpus chunks from model-generated baseline text, whether it captures culturally specific literary register or the broader distinction between human literary prose and machine-generated text remains unresolved. A non-queer Taiwanese literary baseline would be required to isolate the thematic/ culturally

specific signal. More broadly, developing methods that are able to target thematic/ culturally specific signals more precisely remains a direction for further research.

The alpha collapse finding, that probe scores can be statistically significant while generated text is degenerate, reveals a fundamental limitation of representation-level evaluation in isolation. How qualitative human evaluation can be more efficiently integrated into digital humanities workflows, with digital tools assisting rather than replacing human judgment, also remains a research direction worth further exploration.

Finally, only one model architecture was tested for this project. Cross-model replication across different architectures and model sizes would establish whether the RAG + Steering complementarity generalizes or is specific to Qwen's architecture.

6. Conclusion and Future Directions

This project demonstrates that a 58-text corpus deployed at inference time on consumer hardware produces statistically significant shifts in base model generation. RAG establishes generative orientation, suppressing analytical defaults and producing more focused, interiority-oriented continuation. Activation steering, when properly calibrated to the target model, amplifies the representational shift. The pipeline is replicable by any community with a curated corpus, a modern laptop, and access to open-weight models, no institutional compute required.

Further directions of this project may start from incorporating other thematically and culturally contemporary literature baselines for comparison, as well as cross-linguistic and cross-architecture replication across different model sizes. Another idea stemming from this project that I envision is a Modular Vector Library: an open-source library of steering vectors, each computed from a different culturally specific literary corpus, or non-literary or minority language corpus, that communities can apply to any open-weight base model at inference time. For instance, a steering vector and an organized corpus of an endangered literary tradition, or of Taiwanese literature, made openly available so that it draws more attention and invites collaboration in building and maintaining it.

As large-scale scraping of the web approaches saturation, communities with carefully curated domain-specific data may find themselves with increasing leverage over what models learn and generate. Knowledge to curate quality data will be a key site of power in the future development of language models, and I hope this direction of endeavor can be one of those forces that contributes to a more diverse and representative landscape, one where more people, more traditions, and more ways of relating to language get to participate in shaping what models learn and what they generate.