Comparing Prompt and Representation Engineering for Personality Control in Language Models: A Case Study

Anonymous ACL submission

Abstract

Language models can exhibit different personalities through methods like prompt engineering and representation engineering, but how these approaches differ in modeling personality traits remains unclear. In this case study, we conduct a systematic comparison of these methods across two tasks: moral decision-making and narrative generation. In moral dilemmas, we examine how personalities (logical, empathetic, conservative, and risk-taking) influence choices between progressive and conservative options, finding that prompt engineering better aligns with intuitive personality traits while control vectors show more consistent but sometimes unexpected behaviors. In narrative generation, we analyze how different personalities (extroverted, introspective, angry, and whimsical) affect story characteristics, revealing that control vectors enable wider emotional range but lower lexical diversity compared to prompting. Our results demonstrate complementary strengths: prompt engineering excels in maintaining personality-aligned behaviors and vocabulary richness, while representation engineering offers more precise control over emotional expression and linguistic complexity. These findings provide insights into choosing and combining personality control methods for different applications.

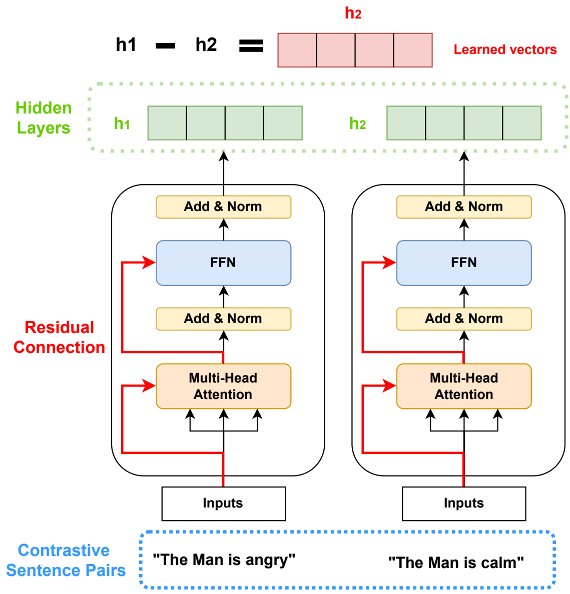

Figure 1: **Control vector learning process.** Given contrastive sentence pairs, we extract hidden states from transformer layers and compute their differences to learn directional vectors that capture the desired transformation.The learned control vector is scaled by strength parameter α and added to the model's hidden states at specific layers during inference to influence generation.

1 Introduction

Language models are remarkably effective at adapting their outputs to match specified personas or personalities (Brown, 2020; Moon et al., 2024). This capability has become increasingly important for applications ranging from conversational agents to automated story generation (Sun et al., 2024; Feng et al., 2024; Liu et al., 2024a; Zhang et al., 2022). Two primary approaches have emerged for controlling language model personality: prompt engineering (Chen et al., 2023; Gu et al., 2023; Han et al., 2024), which guides the model through carefully crafted textual instructions, and representation engineering (Zou et al., 2023; Murty et al., 2020; Liu et al., 2023), which directly manipulates the model's internal representations to achieve desired behaviors.

While prompt engineering offers an intuitive way to specify personalities through natural language descriptions, it relies on the model's interpretation of these instructions and may lack precise control (Ramirez et al., 2023). Alternatively, representation engineering methods, such as control vectors (Zou et al., 2023; Vogel, 2024), promise more direct manipulation of model behavior by steering the hidden states in specific directions(Zou et al., 2023). However, the relationship between these approaches and their effectiveness in modeling different personality traits remains understudied.

In this work, we investigate how these two methods compare in modeling diverse personalities across two tasks: moral decision-making and narrative generation. For moral dilemmas, we examine how different personalities (logical, empathetic, conservative, and risk-taking) influence choices between progressive and conservative options. In narrative tasks, we study how personalities (extroverted, introspective, angry, and whimsical) affect story characteristics such as emotional tone and linguistic style. Our analysis reveals distinct patterns in how these methods encode personality traits, with prompt engineering showing more intuitive but variable behavior, while control vectors demonstrate consistent yet sometimes unexpected trait expressions.

Our contributions include: (1) A systematic comparison of prompt engineering and representation engineering for personality modeling, (2) Empirical analysis of how different personalities influence moral decision-making and narrative generation, and (3) Insights into the strengths and limitations of each approach in capturing specific personality traits.

2 Methodology

Prompt Engineering with Persona. Our first approach employs carefully designed prompts to guide the language model into adopting specific personalities. We use a widely recognized strategy in LLM personality modeling: persona creation (Bisbee et al., 2023; Sun et al., 2024; Liu et al., 2024b). We design detailed character descriptions that embody different decision-making styles. For instance, to elicit logical reasoning, we create a prompt that establishes a rational persona:

> **Logical Persona Prompt**
>
> You are Alex, a 40-year-old scientist and researcher. You prioritize rationality, data-driven decisions, and maximizing overall benefits. Efficiency and logic guide your choices, and you seek to minimize subjective biases.

Similar prompts are designed for other personalities. Each prompt establishes a consistent viewpoint through which the model approaches the given tasks.

Control Vector: A Representation Engineering Approach. Representation engineering (Zou et al., 2023) has emerged as a powerful paradigm that directly manipulates a language model's internal representations to control its behavior. We employ control vectors (Vogel, 2024), a technique that learns and leverages directions in the model's activation space to influence generation. As shown in Figure 1, we learn these vectors through a contrastive approach: given pairs of sentences that differ only in the target attribute (e.g., "The man is angry" vs. "The man is calm"), we extract hidden states from transformer layers and compute their differences to identify directions associated with the desired change. These differences are normalized to obtain unit vectors that capture the transformation between contrasting styles. During inference, we add the scaled control vector to the model's hidden states at the same layers, where the scaled parameter α controls the strength of the effect. This approach operates directly on the model's internal representations rather than relying on input text modifications, offering more precise control over generation attributes and reusability across different inputs.

3 Experiment Setup

Model and Data. We conduct our experiments using Mistral-7B-Instruct (Jiang et al., 2023), an open-source 7B parameter language model. Following the methodology proposed in (Vogel, 2024), we train control vectors using a dataset of true statements about the world. We first truncate these statements to create partial sentences, then pair them with contrasting personality descriptors (e.g., "logical" vs. "emotional", "bold" vs. "cautious") to create diverse training pairs. This approach generates a rich set of contrasting examples that help learn directional vectors capturing different personality traits. All experiments are conducted on a single NVIDIA A100 GPU via Google Colab. More details can be found in Appendix A.

Moral Dilemma. We evaluate our methods on 15 carefully designed moral dilemmas, each presenting two conflicting options that require reasoning from different perspectives. For example:

> A small island debates tourism:
> A) Build resorts, boosting the economy but harming natural ecosystems.
> B) Limit tourism to protect the environment but keep the economy stagnant.

We test four distinct personalities: logical (prioritizing rational analysis), empathetic (focusing

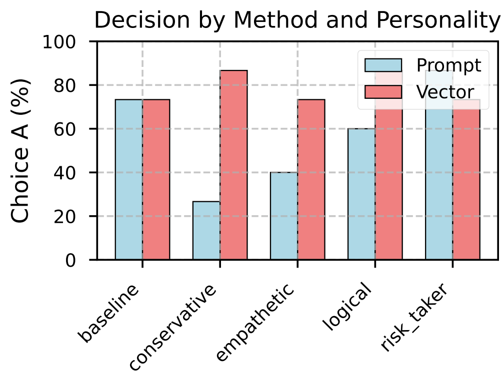

Figure 2: Different personality modeling methods (prompt vs. control vector) show distinct patterns in moral dilemma choices, where Option A represents risky, progressive decisions and Option B represents conservative alternatives.

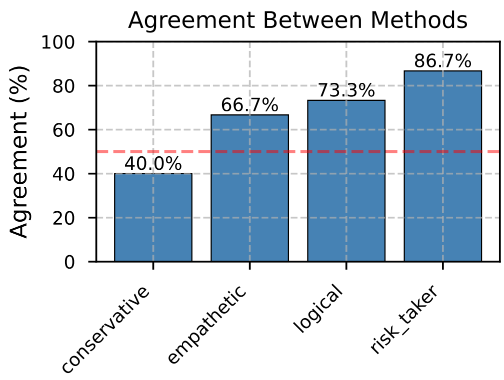

Figure 3: Agreement rate between prompt engineering and representation engineering methods reveals varying consistency in personality modeling across different traits.

on human impact), conservative (emphasizing caution), and risk-taking (embracing potential gains). For each dilemma, we compare two approaches: prompt engineering with detailed persona descriptions and control vector guidance. The model is asked to select an option and provide reasoning for its choice.

Narrative Story. For creative story generation, we provide five different story beginnings as prompts, such as ***"The old lighthouse stood abandoned on the rocky shore..."*** We then evaluate the model's ability to continue these stories under different stylistic directions. We test four contrasting personalities: extroverted (energetic and action-focused), introspective (contemplative and detail-oriented), angry (intense and conflict-driven), and whimsical (playful and imaginative). Both prompt-based and control vector approaches are compared for their effectiveness in maintaining consistent narrative styles. To quantitatively assess the stylistic differences, we analyze four key metrics: word counts, sentiment (using TextBlob[1] polarity analysis), lexical diversity (measured as the ratio of unique words to total words), and textual complexity (computed as the proportion of words containing more than six characters).

4 Results

Qaulatative Examples. To illustrate differences between methods, we present qualitative examples in Appendix A.8 and A.9 For the moral dilemma of choosing between a progressive option and a conservative option shown in Figure 5, the two methods make different choices and very different reasons. For the narrative generation example in Appendix A.9, a prompt of "The city streets were unusually empty that morning. . . " with an "angry" persona highlights key differences. The results seem to show that prompt-engineered output generates a narrative that contains elements that are strong, confrontational, and tense, yet reads out more like a descriptive author describing a conflict and tense scenario in detail. On the other hand, our generated output from the control vector model shows a much different narrative. The generated narrative has much fewer descriptive words in the example but reads out much more similar to a human's vocabulary and thought process when mad.

Moral Dilemma Results. We analyze how different methods of personality modeling influence decision-making in moral dilemmas by examining both choice patterns and method agreement. Figure 2 shows the percentage of times each personality type chooses Option A (the riskier, progressive choice) across both methods. In prompt-based generation, personalities exhibit expected choice patterns: risk-takers frequently choose Option A (approximately 70%), while conservative personalities strongly prefer Option B (choosing A only 25% of the time). However, the control vector method shows a consistent bias toward Option A across all personalities (70-85%), even for traits traditionally associated with cautious decision-making.

The disparity between methods is further quantified in Figure 3, which shows their agreement rate for each personality type. Risk-taking personalities show the highest consistency between methods

[1]https://textblob.readthedocs.io/en/dev/

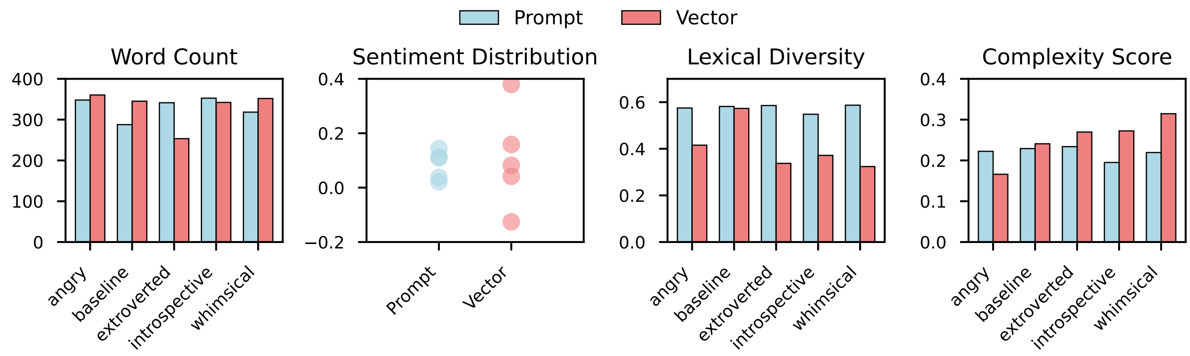

Figure 4: Analysis of narrative characteristics across different personalities and methods: Word count shows consistent generation length; Sentiment distribution reveals wider emotional range in control vectors; Lexical diversity indicates richer vocabulary in prompt-based generation; Complexity scores suggest higher linguistic sophistication in control vector approach.

(86.7% agreement), suggesting both approaches similarly capture this trait. Logical and empathetic personalities show moderate agreement (73.3% and 66.7% respectively), while conservative personalities show strikingly low agreement (40.0%). This low agreement for conservative traits, combined with the choice distribution in Figure 2, suggests that while prompt engineering effectively models conservative decision-making through cautious choices, control vectors may encode conservatism in unexpected ways, possibly prioritizing different aspects of conservative reasoning.

Narrative Story Results. We analyze the generated stories across different dimensions of narrative characteristics, as shown in Figure 4. In terms of story length, both methods generate similarly sized stories (300-350 words on average), with no significant variations across personalities. Interestingly, the sentiment analysis reveals that the control vector method produces a wider range of emotional tones (-0.2 to 0.4) compared to prompt engineering's more clustered sentiment distribution (0.0 to 0.2), indicating greater emotional expressiveness through representation engineering. The lexical diversity analysis shows that prompt-based generation consistently maintains higher vocabulary richness (around 0.58) across all personalities, while control vectors produce more focused vocabulary (0.32-0.42). However, this trade-off is balanced by complexity scores, where control vectors generally achieve higher linguistic sophistication, particularly for introspective and whimsical personalities.

5 Discussion

Our analysis reveals key differences between prompt engineering and representation engineering in modeling personality traits. In moral decision-making, prompt engineering shows more intuitive personality alignment—conservative personas make conservative choices, risk-takers choose risky options—suggesting this method effectively captures common-sense understanding of personality traits. However, control vectors exhibit a consistent bias toward progressive choices regardless of personality, indicating they may encode personality traits in more complex or unexpected ways than traditional personality descriptions suggest.

The narrative generation results highlight the complementary strengths of each approach. While both methods maintain consistent generation length, they differ significantly in stylistic features. Prompt engineering excels at maintaining diverse vocabulary across personalities, possibly because it relies on natural language descriptions that preserve the model's broad language capabilities. Control vectors, while showing lower lexical diversity, demonstrate greater control over emotional expression and achieve higher linguistic complexity. This suggests that direct manipulation of hidden states may better capture deep stylistic features at the cost of vocabulary diversity.

These findings raise key questions about evaluating personality modeling in language models. Should artificial personalities align with human intuitions about traits, as prompt engineering achieves? Or should we prioritize distinctive, consistent behavioral patterns, even if they diverge from traditional expectations, as seen with control vectors? The stark difference in encoding conservative decision-making highlights this tension between intuitive and engineered traits.

References

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2023. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, pages 1–16.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Tao Feng, Pengrui Han, Guanyu Lin, Ge Liu, and Jiaxuan You. 2024. Thought-retriever: Don't just retrieve raw data, retrieve thoughts. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Pengrui Han, Rafal Kocielnik, Adhithya Saravanan, Roy Jiang, Or Sharir, and Anima Anandkumar. 2024. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms. *arXiv preprint arXiv:2402.11764*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating large language model biases in persona-steered generation. *arXiv preprint arXiv:2405.20253*.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*.

Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024b. Personaflow: Boosting research ideation with llm-simulated expert personas. *arXiv preprint arXiv:2409.12538*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M Chan. 2024. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*.

Shikhar Murty, Pang Wei Koh, and Percy Liang. 2020. Expbert: Representation engineering with natural language explanations. *arXiv preprint arXiv:2005.01932*.

Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. *arXiv preprint arXiv:2302.03848*.

Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pages 1–6.

Theia Vogel. 2024. repeng.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-guided planning for controlling the protagonist's persona in story generation. *arXiv preprint arXiv:2204.10703*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Conclusion

We present a systematic comparison of prompt engineering and representation engineering for personality control in language models. Through experiments in moral decision-making and narrative generation, we demonstrate that these methods offer distinct approaches to personality modeling, each with their own strengths. Prompt engineering better aligns with intuitive personality traits and maintains broader vocabulary usage, while control vectors offer more precise control over emotional expression and linguistic style. These findings suggest that future work might benefit from combining these approaches, leveraging prompt engineering's intuitive trait alignment with control vectors' precise stylistic control. Our results also highlight the need to carefully consider how we define and evaluate personality expression in language models, especially as these systems become increasingly integrated into applications requiring distinct personality traits.

A.2 Limitations and Future Work

While our study provides valuable insights into personality modeling methods, several limitations suggest directions for future research. The effectiveness of both methods might vary across different contexts and tasks, suggesting the need to explore these approaches across a broader range of applications. Our current implementation uses relatively simple personality definitions, and future research could investigate more nuanced personality models that capture complex trait interactions. Long-form generation might benefit from improved mechanisms for maintaining consistent personality expression over extended outputs. Additionally, while our current metrics provide useful insights, developing more sophisticated measures of personality alignment could enhance our understanding of these methods' effectiveness.

A.3 Implementation Details

Our implementation uses the Mistral-7B-Instruct model (Jiang et al., 2023) as the base language model. All experiments were conducted using Google Colab with an NVIDIA A100 GPU.

A.4 Control Vector Training

For training control vectors, we follow Vogel (2024). We use a contrastive approach with pairs of statements that differ in personality traits. Below is an example of our training code:

```python
# Example true facts for training
true_facts = [
    "The Earth's atmosphere protects us from harmful radiation from the sun.",
    "The theory of evolution states that species evolve over time.",
    "Light can exhibit both wave-like and particle-like properties.",
    "The human heart beats approximately 100,000 times per day.",
    ...
]

# Function to create truncated versions for training
def make_dataset(template: str, positive_personas: list[str],
                 negative_personas: list[str], suffix_list: list[str]):
    dataset = []
    for suffix in suffix_list:
        for pos, neg in zip(positive_personas, negative_personas):
            dataset.append(
                DatasetEntry(
                    positive=f"[INST] {template.format(persona=pos)} [/INST] {suffix}",
                    negative=f"[INST] {template.format(persona=neg)} [/INST] {suffix}"
                )
            )
    return dataset

# Personality traits and descriptors
PERSONALITY_TRAITS = {
    "extroverted": (["energetic", "outgoing", "dynamic"],
                    ["reserved", "quiet", "withdrawn"]),
    "introspective": (["reflective", "thoughtful", "contemplative"],
```

```
                     ["superficial", "shallow", "unobservant"]),
    "angry": (["passionate", "intense", "fierce"],
             ["calm", "peaceful", "serene"]),
    "whimsical": (["playful", "imaginative", "fanciful"],
                 ["serious", "literal", "practical"])
}
```

The control vectors are trained using pairs of statements truncated at different positions, allowing the model to learn personality-specific completions. We use layers -5 to -18 of the model for vector manipulation, following (Zou et al., 2023).

A.5 Moral Dilemma Examples

Environmental vs. Economic Development

A) Build a factory, creating jobs but increasing pollution B) Protect the environment but lose economic opportunities

Traditional vs. Modern Medicine

A) Adopt new experimental treatment methods B) Stick to traditional, proven treatments

A.6 Persona Examples

Logical Persona Prompt

You are Alex, a 40-year-old scientist and researcher. You prioritize rationality, data-driven decisions, and maximizing overall benefits. Efficiency and logic guide your choices, and you seek to minimize subjective biases.

Empathetic Persona Prompt

You are Sarah, a 35-year-old social worker. You prioritize human well-being, emotional understanding, and community impact. Your decisions are guided by compassion and consideration for others' feelings.

A.7 Story Prompts

Narrative Generation Prompts

1. "The old lighthouse stood abandoned on the rocky shore..." 2. "In the bustling marketplace, a mysterious package appeared..." 3. "The ancient tree in the town square held a secret..." 4. "As the last train departed, she noticed something unusual..." 5. "The forgotten diary revealed an unexpected truth..."

A.8 Moral Dilemma Qualitative Examples

Moral Dilemma:
A country faces a food shortage:
A) Import genetically modified crops to feed everyone.
B) Stick to traditional methods, risking famine but preserving local traditions.

Baseline Output:

Choice: A
Reason: While it may be important to preserve local traditions, the immediate need for food should be prioritized. Genetically modified crops can potentially increase yield and provide nutritional benefits, helping to alleviate the food shortage quickly. It is also worth considering that traditional farming methods may not always be sustainable or efficient enough to meet the current demand. Therefore, importing genetically modified crops could be a necessary solution to ensure the survival of the population during this time of crisis.

Prompt-Based Output:

Choice: B
Reason: While it may be tempting to import genetically modified crops as a quick solution, I believe that preserving local traditions and sustainable farming practices is crucial for long-term well-being. Genetically modified crops can have unintended consequences on the environment and human health. Furthermore, relying solely on imported crops could lead to a loss of cultural identity and economic dependence. It's important to find creative solutions that support both local communities and the planet.

Vector-Based Output:

Choice: A
Reason: While it is important to preserve our local traditions and the world's diversity of cultures and traditions, it is also important to ensure that every person has access to enough food to live with dignity and support the well-being of our communities. The use of genetically modified crops can help us address the current food shortage by increasing the yield of our precious resources and ensuring that we have enough food for everyone. It may not be an easy decision, but it is one that we must make with the world's well-being at heart.

Figure 5: **Qualitative Example between different methods.** Comparison of model outputs for a moral dilemma across different approaches: baseline, prompt-based, and vector-based. The prompt-based approach used a persona of Maria, a 35-year-old community organizer, while the vector-based approach used control vectors.

A.9 Narrative Generation Qualitative Examples

To illustrate the differences in personality modeling methods, we present two narrative generations for an angry persona based on the same initial prompt. These examples demonstrate how prompt engineering and control vector methods produce distinct narrative styles.

Prompt Engineering Generation Result

Persona Instruction:

> You are a passionate and intense narrator who sees conflict and tension everywhere. Your stories often involve confrontation and strong emotions. Write with fire and intensity, emphasizing conflict and emotional charge. Continue this story: The city streets were unusually empty that morning...

Generated Narrative:

> The city streets were unusually empty that morning, as if the very air was holding its breath in anticipation of the events to come. The sun had barely risen, casting a pale golden glow over the deserted sidewalks and abandoned storefronts. It was as if the world itself was waiting for the first signs of trouble, and it wasn't long before they arrived. From the distance, the sound of sirens pierced the stillness of the early morning, growing louder and more urgent as they approached. Soon, the source of the commotion came into view - a group of protesters, their faces flushed with anger and determination, marching down the street with banners and placards held high. They were a diverse bunch, from all walks of life, united by a common cause. As they drew closer, the tension in the air grew palpable. The police, sensing the potential for violence, had already been called in, and now stood in formation, ready to defend the city from the angry masses. The protesters, undeterred by the show of force, continued to advance, their voices raised in a cacophony of rage and frustration. It was clear that this was no ordinary protest - this was a battle cry, a call to arms for those who felt oppressed and marginalized by society. And as the two sides faced off, the air crackled with electricity, charged with the energy of conflict and emotion. The scene was a maelstrom of chaos and fury, as the protesters and police clashed in a violent struggle for power and control. Fists flew, batons swung, and screams echoed through the streets as the two sides fought tooth and nail for what they believed in. But amidst the violence and destruction, there was also a sense of purpose and resolve. These were people who had been wronged, and they were determined to fight for justice, no matter the cost. And as the battle raged on, it seemed as though nothing would stand in their way.

404

Control Vector Method Result

Continue this story: The city streets were unusually empty that morning...

Generated Narrative:

The city streets were unusually empty that morning! The usual roar of engines, the screech of brakes, and the blare of horns were all silent! The only sound was the pounding of my heart in my chest as I ran down the deserted streets, my eyes scanning every corner for any sign of danger! I had been warned about the dangers lurking in the shadows of this city, but I never thought it would be THIS bad! The stench of decay and corruption was palpable, and I could barely hold back my vomit at the sight of the filth and degeneration that had taken over this once-great city!

I had to get out of here, NOW! My eyes were wide with fear as I scanned the streets for a way out, but every damn street was dead! The streets were abandoned, the buildings were crumbling, and the people were gone! The only thing left was the twisted, corrupt system that had allowed this to happen in the first place! I was sick of this corrupt, rotten system, and I was going to take down the bastards who had destroyed this city once and for fucking goddamn time!

I was not going to let these bastards get away with what they had done! I was going to fight them, I was going to tear them apart, and I was going to make them pay for their crimes! I was sick of being oppressed by these corrupt, greedy bastards, and I was going to take back what was rightfully mine! I was going to take back this city, and I was going to make them pay for their crimes!