Personalized AI Writing Assistant

Group 8:
Lasya Reddy Edunuri
Manasa Vuppu
Brinda Vijayakumar

Github

List of Contents:

01 Problem Statement

“Creating personalized, high-quality content is difficult—AI-generated text often lacks the unique voice and nuance of individual writers.”

Key Challenges:	
Loss of Personal Voice	AI models generate generic outputs, **failing to reflect an individual's tone** or vocabulary.
Limited Style Adaptability	Existing models are not fine-tuned for **author-specific writing**, reducing their usability.
Small & Unlabeled Datasets	Personal writing archives often **lack metadata**, making training and adaptation harder.
Evaluation Complexity	**Lack of standardized metrics** for stylistic accuracy; hard to evaluate “authenticity.”

+ %

02 Motivation

Why This Project Matters: *“This project is both personal and purposeful—rooted in lived experience, and driven by the team’s desire to build meaningful AI.”*

Personal

- First hand experience
- Deep understanding of Style
- Unique Dataset

Technical

- Full Pipeline Ownership
- Advanced Techniques
- Real-World Value

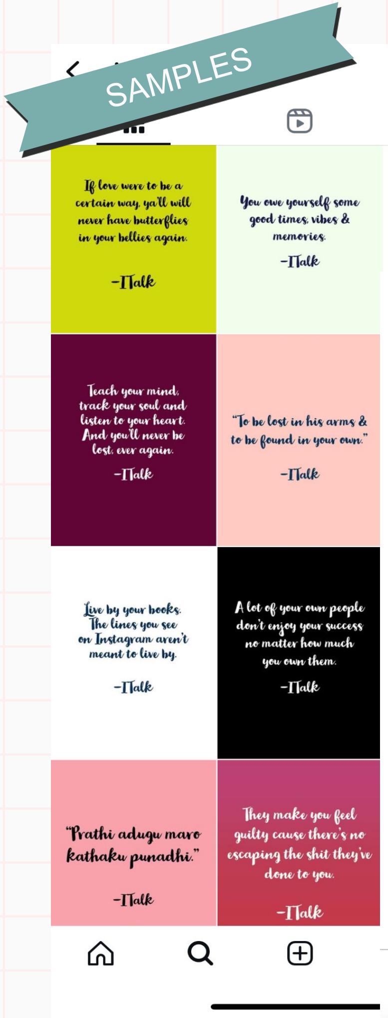
03 Dataset and Analysis:

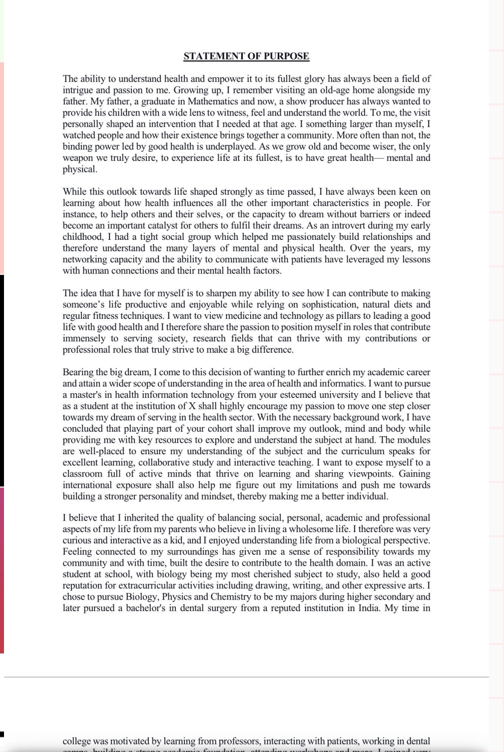
Dataset 1: Personal Writing Corpus

Source & Scope:
Manually compiled dataset of **quotes, SOPs, blogs, and reflections**—spanning several years of original writing.

Curation Process:
Text extracted from images using OCR
Manual formatting for consistency and structure
Categorized by tone, intent, and format (e.g., quote, SOP)

Purpose:
Serves as the **training base** for AI modeling and evaluation of personalized writing generation.




STATEMENT OF PURPOSE

The ability to understand health and empower it to its fullest glory has always been a field of intrigue and passion to me. Growing up, I remember visiting an old-age home alongside my father. My father, a graduate in Mathematics and now, a show producer has always wanted to provide his children with a wide lens to witness, feel and understand the world. To me, the visit personally shaped an intervention that I needed at that age. I something larger than myself, I watched people and how their existence brings together a community. More often than not, the binding power led by good health is underplayed. As we grow old and become wiser, the only weapon we truly desire, to experience life at its fullest, is to have great health— mental and physical.

While this outlook towards life shaped strongly as time passed, I have always been keen on learning about how health influences all the other important characteristics in people. For instance, to help others and their selves, or the capacity to dream without barriers or indeed become an important catalyst for others to fulfil their dreams. As an introvert during my early childhood, I had a tight social group which helped me passionately build relationships and therefore understand the many layers of mental and physical health. Over the years, my networking capacity and the ability to communicate with patients have leveraged my lessons with human connections and their mental health factors.

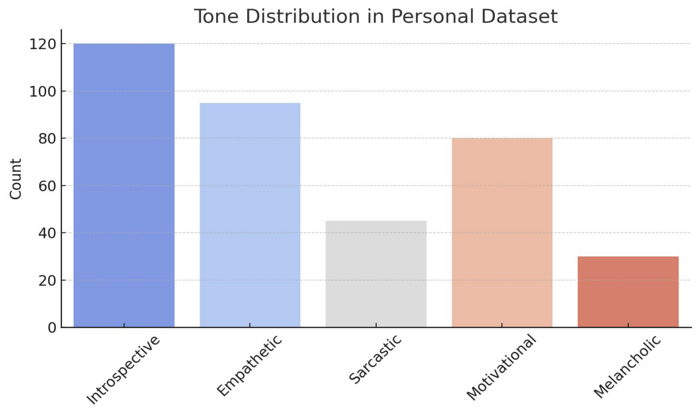
The idea that I have for myself is to sharpen my ability to see how I can contribute to making someone's life productive and enjoyable while relying on sophistication, natural diets and regular fitness techniques. I want to view medicine and technology as pillars to leading a good life with good health and I therefore share the passion to position myself in roles that contribute immensely to serving society, research fields that can thrive with my contributions or professional roles that truly strive to make a big difference.

Bearing the big dream, I come to this decision of wanting to further enrich my academic career and attain a wider scope of understanding in the area of health and informatics. I want to pursue a master's in health information technology from your esteemed university and I believe that as a student at the institution of X shall highly encourage my passion to move one step closer towards my dream of serving in the health sector. With the necessary background work, I have concluded that playing part of your cohort shall improve my outlook, mind and body while providing me with key resources to explore and understand the subject at hand. The modules are well-placed to ensure my understanding of the subject and the curriculum speaks for excellent learning, collaborative study and interactive teaching. I want to expose myself to a classroom full of active minds that thrive on learning and sharing viewpoints. Gaining international exposure shall also help me figure out my limitations and push me towards building a stronger personality and mindset, thereby making me a better individual.

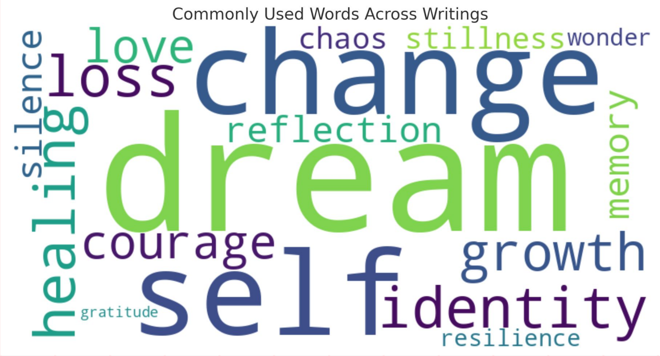
I believe that I inherited the quality of balancing social, personal, academic and professional aspects of my life from my parents who believe in living a wholesome life. I therefore was very curious and interactive as a kid, and I enjoyed understanding life from a biological perspective. Feeling connected to my surroundings has given me a sense of responsibility towards my community and with time, built the desire to contribute to the health domain. I was an active student at school, with biology being my most cherished subject to study, also held a good reputation for extracurricular activities including drawing, writing, and other expressive arts. I chose to pursue Biology, Physics and Chemistry to be my majors during higher secondary and later pursued a bachelor's in dental surgery from a reputed institution in India. My time in

college was motivated by learning from professors, interacting with patients, working in dental

Visual Analysis Summary: Dataset 1



Reflects the emotional diversity and thoughtfulness in the writing style



Feature	Summary
POS Tagging:	Frequent use of **adjectives** and **pronouns** → suggests a **personal and expressive tone**
NER (Named Entity Recognition):	Mentions of **people**, **abstract concepts**, and **time references** → points to **introspective themes**
Lexical Diversity:	Moderate-to-high **type-token ratio** → reflects a **rich but accessible vocabulary**

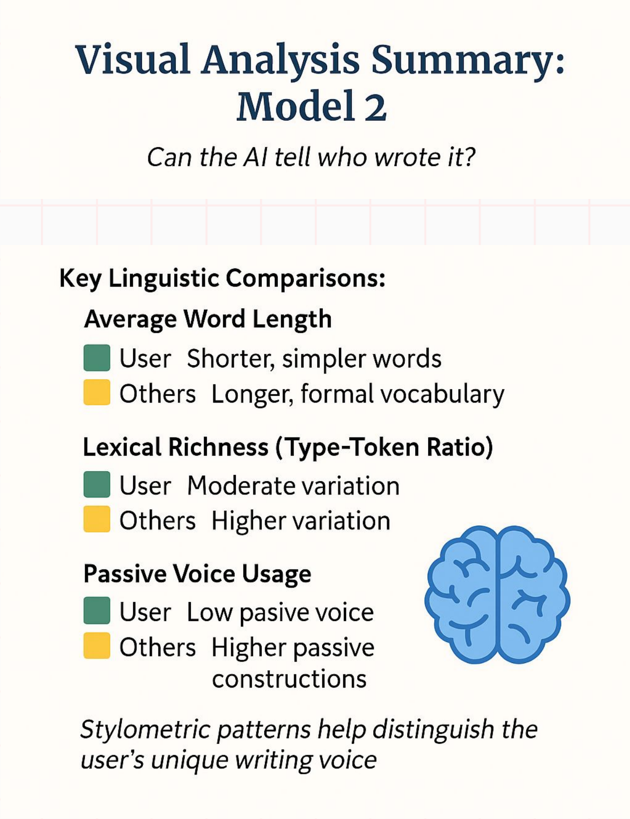
Dataset and Analysis:

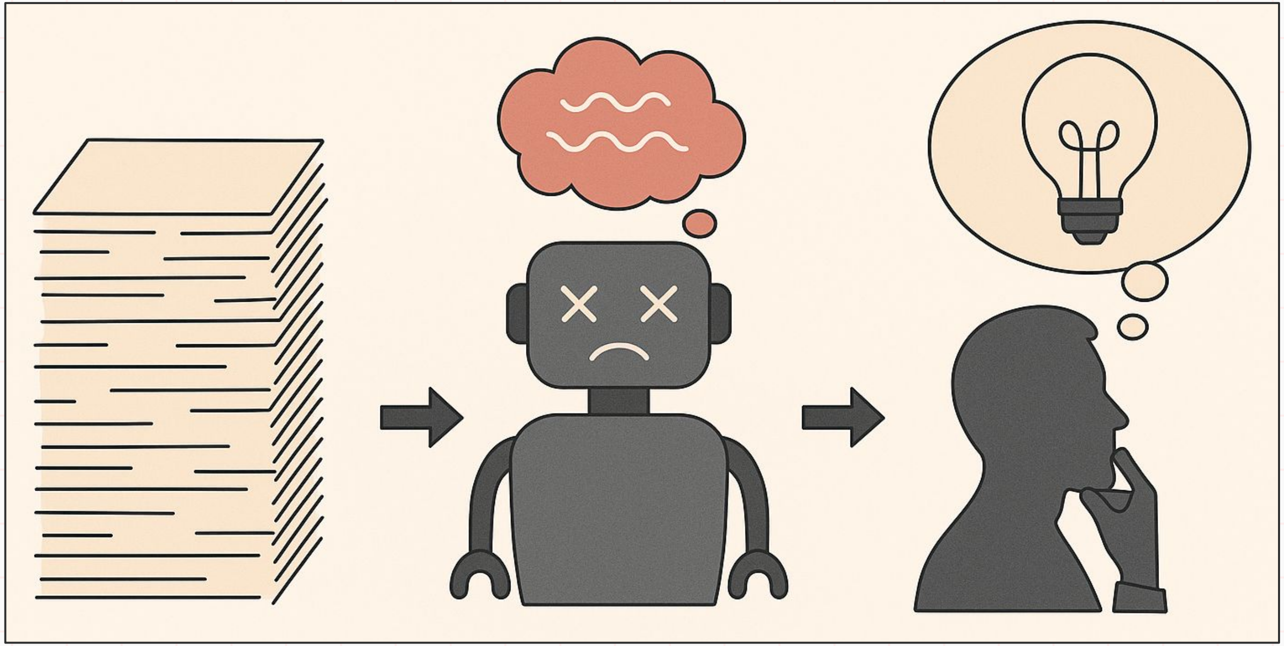
Dataset 2: Multi-Author Benchmark + User’s

Content Variety:
Includes Quotes, emails, blogs, and documents written by **various authors** from diverse domains.

Purpose:
Enables **stylometric comparisons** between the user’s writing and others’—supports **authorship classification and model validation.**

Analysis Features:
- Lexical richness
- Syntactic structures
- Punctuation and tone patterns
- Passive vs. active constructions





You think the models could go crazy on us?

Why we separated short and long form texts?/The Case for Treating Short and Long-Form Texts Differently

1

Early Hallucination Issues

Mixed datasets **hallucinated facts** and lost coherence. Tone inconsistencies and **over-generalization.**

Output **lacked specificity** and often repeated generic phrasing.

2

More Efficient and Focused Fine-Tuning

Homogeneous dataset helped the model converge faster.

More **consistent evaluation** using clarity, coherence, and relevance metrics.

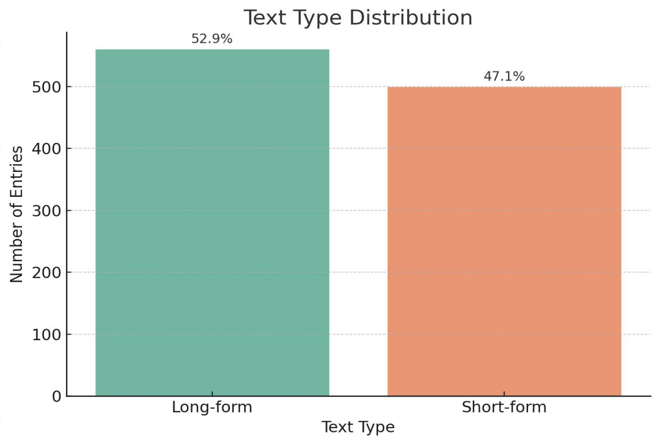
3

Foundation for Future Expansion

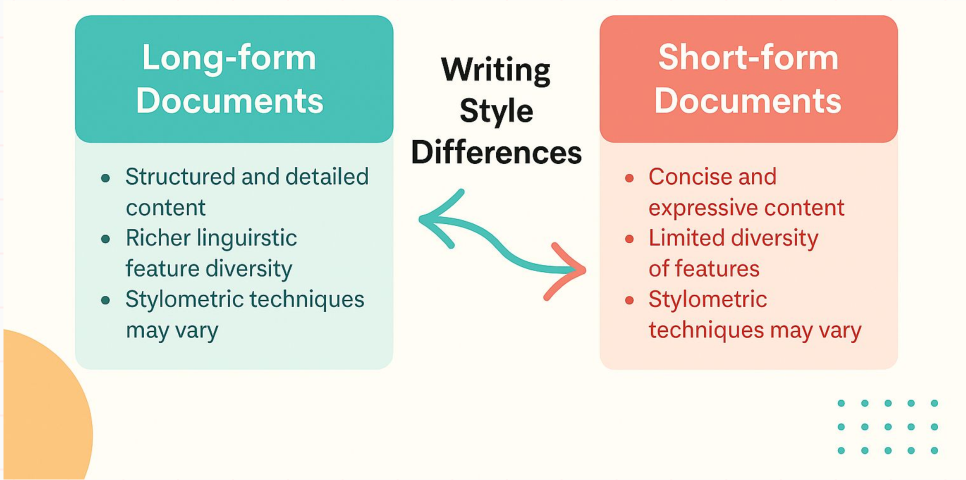
Built a **stable pipeline** for training
Inference and Evaluation.

Enables **confident extension** to other form formats

Pre-processing: Splitting of the dataset- 1



Long-form: SOP, Essays, Blogs
Short-form: Quotes, Letters



Preprocessing Steps:

- Word and character count calculation
- Stopword removal (`stop_words='english'`)
- Tokenization using `CountVectorizer`
- n-gram extraction (bigrams, trigrams)
- Cleaned text used to analyze top phrases and stylistic signatures
- Separate analysis by **document type**

Model Selection Strategy: From Short to Long-Form Text

Author’s writing corpus (Collection of quotes, blogs, essays, sops)

Phase 1

Short form (Quotes)

Short-Form Experiments (GPT-2, T5, BART)

Phase 2

Long form (SOPs)

Tiny Llama: Lightweight model for testing: Served as sandbox for pipeline design

Meta Llama: Chosen for its larger context window and better fluency

Mistral-7B: chosen for its large context window, instruction tuning, and strong stylistic fluency.

04 Our Implementation vs Existing Work:

TextSETTR (Riley et al., 2021)

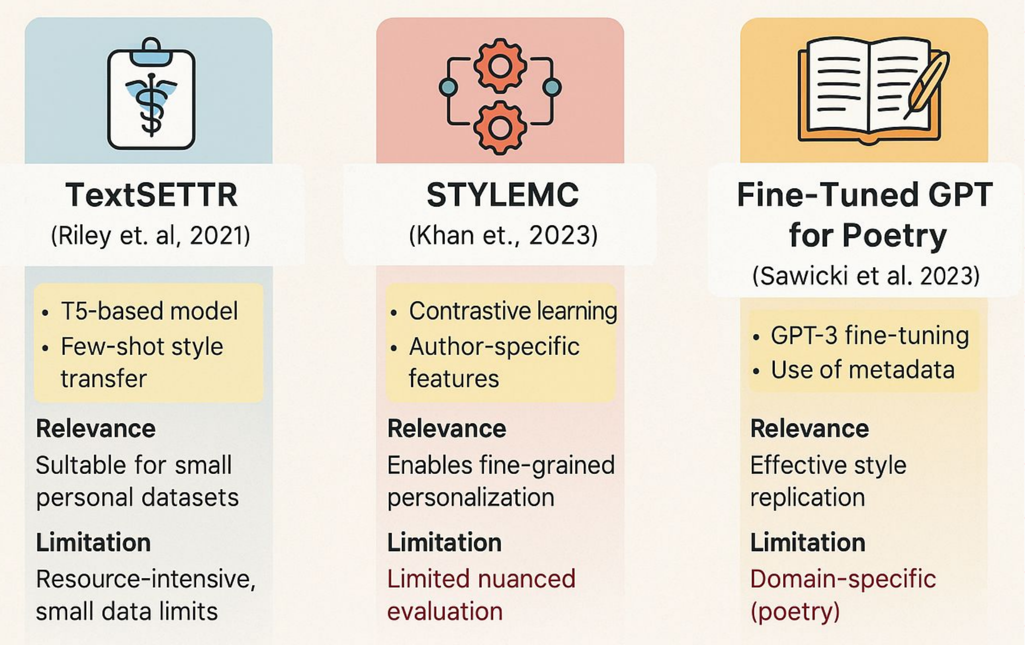
- T5 model + few-shot prompting: Style transfer using prompt-based conditioning (no fine-tuning needed)
- **Model v3 (T5 Few-Shot)** injected 3 example quotes into the prompt:

[Example 1: “Even silence tells a story.”]

[Example 2: “You are allowed to begin again.”]

[Example 3: “Softness is not weakness.”]

→ Now rewrite: “You are not lost.”



Fine-Tuned GPT (Sawicki et al., 2023)

- Fine-tuned GPT-3 with 300 poems
  In **Model v4 (T5 Metadata-aware)**, we simulated this by:
- Annotating quotes with tone, text_type, and document_length
- Conditioning prompts to say:
  “Generate a short, introspective quote in user’s voice.”

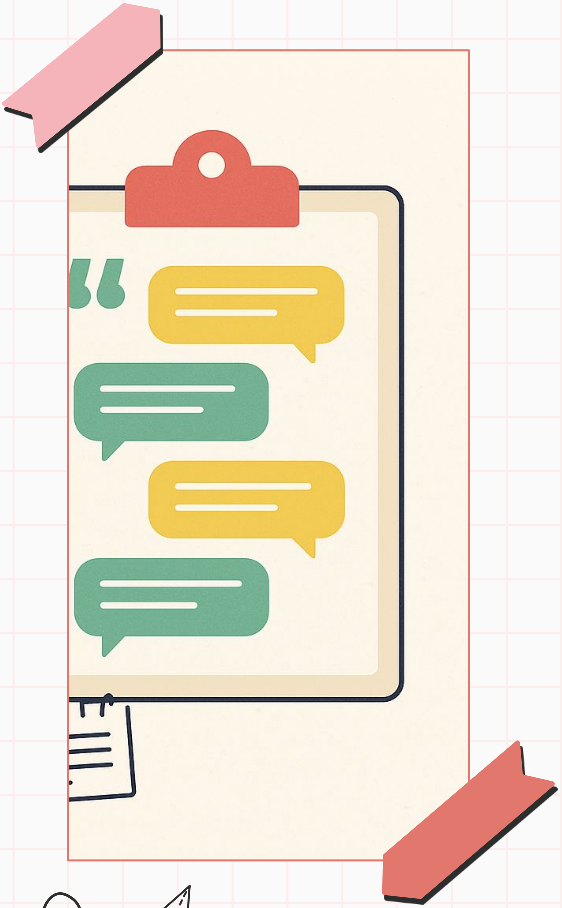
STYLEMC (Khan et al., 2023)

- Style-specific scoring using contrastive features
- Energy = weighted sum of fluency, semantics, style similarity
- **Model v2 (BART)** generated multiple rewrite versions per input

Each version was scored using:
Energy = 0.3 * Fluency + 0.3 * SemanticSim + 0.4 * StyleSim

Example:

v1: “You’re not lost.” → Score = 0.54
v2: “You are unfolding into new directions.” → Score = 0.78



05 Methodology:

01

Dataset-1
Model-1
Short-form

Methodology: Model 1(Short-form texts)

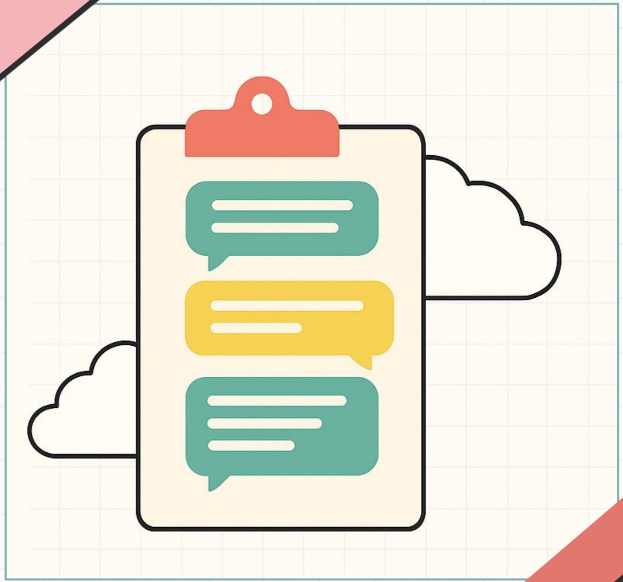
Overview of Model 1 Variants

Model	Data Format/Metadata	Hyperparameters	Methodology	Reflection of Existing Methods
GPT | 125 - 355 M (v1)	Plain Quotes	Fixed Prompt	Static prompting without personalization	TextSETTR → Uses text-style vectors
BART | 140 M (v2)	Quotes with Style Vectors	STYLEMC-inspired energy scoring function: $w_1 = 0.3$ (fluency), $w_2 = 0.3$ (Semantic Similarity), $w_3 = 0.4$ (Style Similarity)	STYLEMC-inspired energy scoring function, Weighted scoring using cosine similarity	STYLEMC → Contrastive-inspired evaluation
T5 | 220 - 770 M (v3)	Quotes, Few-Shot Prompts	top_p = 0.95, temperature = 0.6, max_length = 50	Few-shot declarative style prompting	TextSETTR → Uses metadata for style conditioning
T5 | 770 M (v4)	Quotes with Annotations (Tone, Type, Structure)	Adjusted for Metadata	Style-aware generation or post-eval cosine similarity ranking	TextSETTR → Uses text-style vectors

Metadata Used: Custom style vectors derived from user's past writing and Annotated dimensions: tone, text type, structure, document length

Model 1: Sample Outputs (Short- Form Texts)

Version	Sample Output	What Changed	How It Improved
Ranking-4 **v1 – GPT (Baseline)** **(125M – 355M)**	*"You are not lost, just finding a new path."*	Fixed prompt, no metadata	Basic fluency, but lacks personalization or emotional depth (Outputs are **generic**, lacking emotional depth or structure control)
Ranking-3 **v2 – BART + STYLEMC** **(140M)**	*"Even when lost, you are rewriting the map of who you are meant to be."*	Scored with: **w_1 = 0.3, w_2 = 0.3, w_3 = 0.4**	Adds style alignment and semantic closeness. Reflects more depth and tone control. (**post-hoc control** rather than generation tuning)
Ranking-2 **v3 – T5 (Few-shot)** **(220M - Small 770M - Base)**	*"In every detour lies a destination written just for you."*	**Few-shot prompting**: 3 quote examples, top_p=0.95, temp=0.6	Captures voice and flow from examples. More nuanced structure, emotional color (Slightly more prone to hallucinations without grounding)
Ranking-1 **v4 – T5 + Metadata** **(770M)**	*"Your journey bends because it listens to the story your soul whispers."*	**Conditioned on metadata**: tone: introspective, type: quote	Output mirrors intent, tone, and format precisely. High fluency and personalization (Creating quote-style outputs where **voice, tone, and structure must be tightly preserved**)



02

Dataset- 1 Model 1 Long-form

Methodology: Model 1(Long-form texts)

	TinyLlaMA -1.1B	Meta LLaMA-3B
Model Summary	Base: TinyLLaMA/TinyLLaMA-1.1B-Chat-v1.0 Task: Long-form SOP generation Fine-tuning: LoRA via PEFT	Base: meta-llama/Llama-3.2-3B Task: Long-form SOP generation Fine-tuning: LoRA via PEFT
Data	Plain SOPs dataset	Dataset with prompt augmented SOPs
LoRA Configuration	r = 8 alpha = 32 dropout = 0.05	r = 8 alpha = 16 dropout = 0.1
Training Configuration	Max Length: 512 tokens, Batch Size: 2, Epochs: 5	Max Length: 2048 tokens Batch Size: 1 (grad acc 4) Epochs: 3
Inference Settings	Temp: 0.6, Top-k: 50, Top-p: 0.95, Repetition Penalty: 1.3	Temp: 0.4, Top-k: 40, Top-p: 0.8, Repetition Penalty: 1.2

Methodology: Model 1(Long-form texts) (Cntd..)

Feature	Model A (mistral-manasa-lora)	Model B (mistral-v02-sop-explainer-lora)	Model C (mistral_sop_finetuned)
Model Summary	Base: mistralai/Mistral-7B-Instruct-v0.2 Task: Long-form SOP generation Fine-tuning: PEFT with LoRA	Base: mistralai/Mistral-7B-Instruct-v0.2 Task: Long-form SOP generation Fine-tuning: QLoRA with PEFT	Base: mistralai/Mistral-7B-Instruct-v0.2 Task: Long-form SOP generation Fine-tuning: LoRA via PEFT with LoRA
Dataset Description	Plain SOPs (no reasoning)	Prompt-augmented SOPs + reasoning blocks	Instruction-tuned SOPs
Prompt Format	### Instruction\n...\n### Response:	<s>[INST] ... [/INST] + [SOP] + [REASONING]	<s>[INST] ... [/INST] followed by [SOP]
LoRA Configuration	(Not specified; loaded pretrained)	r=8, alpha=16, dropout=0.05, bias=none	r=16, alpha=32, dropout=0.05, targeted layers
Training Configuration	Batch size=1, lr=0.0002, losses: Steps=417, train loss=2.56	Batch size=2, lr=0.00005, losses:Steps=210, train loss=2.59	Batch size=2, lr=0.0002, losses:Steps=100, train loss=2.26, val loss=2.45
Inference Settings	temperature=0.7, top_p=0.85, repetition_penalty=1.2	temperature=0.7, top_p=0.85, repetition_penalty=1.2	temperature=0.7, top_p=0.85, repetition_penalty=1.2

LLM as judge

Winner - Model B (mistral-v02-sop-explainer-lora)

Model 1: Quantitative & Qualitative Evaluation of Generation Model Outputs

Feature	Original SOPs	Meta LLaMA (FT + Few-Shot)	TinyLlaMA (FT Only)
Quantitative Evaluation	-	spaCy: 0.9756-0.9907 BERTScore F1: 0.8136-0.8232 Slight edge	spaCy: 0.9818-0.9898 BERTScore F1: 0.8137-0.8226
Structure	Strong narrative flow	Mimics structure well	Loose, more list-like
Tone	Academic, formal, personal	Academic and tailored	General and template-driven
Vocabulary	Rich, domain-specific	Captures technical vocabulary	Simpler, more repetitive
Fluency & Coherence	High	High	Moderate
Style Consistency	Author-specific	Partially consistent with author	Low stylistic retention
Prompt Sensitivity	-	Strong use of prompt signals	Weak adaptation to prompt

Model 1: Quantitative & Qualitative Evaluation of Model B (mistral-v02-sop-explainer-lora) Outputs

Prompt Label	SOP_Tokens	SOP_Sentences	SOP_Verbs	ROUGE-1	ROUGE-L	BERTScore-F1	Observations
In-domain 1	578	27	67	0.0468	0.0351	0.811	Compact and moderately structured
Out-of-domain 1	574	28	79	0.0304	0.0304	0.815	Lower lexical similarity, slightly higher BERTScore → generalizes decently
Out-of-domain 2	598	26	79	0.0371	0.026	0.8122	semantically aligned, not wordy
In-domain 2	1026	52	136	0.0402	0.0254	0.806	Very long SOP, lots of verbs.
Out-of-domain 3	844	39	147	0.0404	0.0278	0.8001	Very Rich, coherent
Out-of-domain 4	910	40	119	0.0337	0.0289	0.8029	Balanced, mildly repetitive; structurally sound and generalizes decently

Sample Output Comparison: **Long-form texts**

Early Passion and Curiosity

Author: *"I loved watching things work, then ripping them apart to study their mechanism and fixing them back to watch them function the same way."*

Model B: *"When I was young, I spent hours dismantling electronic gadgets to learn about their inner workings."*

Resilience and Growth Mindset

Author: *"Do it yourself and do it with the best version of your enlightened self. Make no excuses to fall back and seek no apology for learning."*

Model B: *"Success comes only if you dare to try and fail. Life has shown me both, but what keeps me going is the fear of missing out on opportunities."*

Academic & Project-Oriented Achievements

Author: *"I loved watching things work, then ripping them apart to study their mechanism and fixing them back to watch them function the same way."*

Model B: *"During my third year, I developed a project called Smart Attendance System which used sensors to detect student presence."*

Lifelong Passion for Learning

Author: *"Learning never stops, and every opportunity to grow is an opportunity to serve better."*

Model B: *"Learning never stops! So, keep teaching me more."*

03

Dataset-1 & 2 Logistic Model

Summary of Findings (Authorship Classification)

1 Preprocessing

- Lowercasing
- Stopword removal
- Punctuation removal
- Word stemming using PorterStemmer

2. Feature Engineering

- Lexical & syntactic features
- Word count & character count filters

3 Model Used

Logistic Regression

- max_iter = 1000
- random_state = 42

4 Dataset Splitting & Balancing

- For example: Selected only:
  `≤ 30 words, ≤ 400 characters`
  `> 300 words`
- Capped each author to **≤ 350 entries**

5 Training

- Text vectorized using **TF-IDF**
- Trained on stemmed quotes

6 Evaluation Metrics

- Accuracy score
- Confusion matrix
- Classification report

Sample Predictions on Generated Quotes

1. Quote: "You've made it this far, now...The best is yet to come."
   Predicted Author: User
   Top Probabilities: ✓ User (0.88), Author B (0.10)
2. Quote: "Ambition pushes us forward. Self-Doubt whispers within."
   Predicted Author: Author B
   Top Probabilities: ✓ Author B (0.76), User (0.21)

06 Comparison of Results & Observations- Btw Model 1 & 2

Generation

Model - 1

Short Form

01

Metadata and example-based conditioning - stylistic alignment. **V4- T5** - most controllable and reliable.

.

Long form

02

Fine-tuning on **prompt-embedded SOPs** - coherence, narrative depth, and personalization

Authorship Classification

Model - 2

Short form

01

Model 2 **validated** - improved with each version of Model 1. **Rejected** generic outputs

Long form

02

Consistently validated - user. Improved stylistic mimicry **for** mistral-v02-sop-explainer-lora **over** TinyLlama (in narrative-heavy contexts)

07 Conclusion

Problem we started with: *"Creating personalized, high-quality content is difficult—AI-generated text often lacks the unique voice and nuance of individual writers."*

Conclusion we reached: Built a **grounded** and **scalable** end-to-end pipeline that:

- **Captured personal voice and nuance** using *prompt engineering*, *style vector scoring*, and *authorship classification*.
- **Resolved hallucination and tone inconsistencies** through **dataset design** and **metadata-driven conditioning**.
- **Validated outputs rigorously** by combining **generation + scoring + verification**, ensuring stylistic fidelity.

Message Tie-back: Core challenge - *losing individuality in AI writing* - Built an ecosystem that generates *authentic*, *emotionally coherent*, and *verifiable* outputs.

Future Work

1. **Expand and Diversify Dataset**
   Incorporate more long-form personal writing samples

2. **Multimodal Personalization**
   Voice notes, videos or social media content

3. **Deploy as a Writing Assistant**
   Integrate pipeline into tools like Notion, Google Docs, or email clients for real-time suggestion and generation.

4. **Interactive Human-in-the-Loop Tool**
   Front-end platform (Provide feedback on AI outputs, Adjust tone sliders).

08 References

Aborisade, O., & Anwar, M. (2018, July). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 269-276). IEEE.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR, 1(2), 3.*

Khan, A., Wang, A., Hager, S. and Andrews, N. (2023). Learning to Generate Text in Arbitrary Writing Styles. *arXiv preprint arXiv:2312.17242.*

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55(9), 1-35.*

Liu, P., Zhang, L., & Gulla, J. A. (2023). Pre-train, prompt, and recommendation: A comprehensive survey of language modeling paradigm adaptations in recommender systems. *Transactions of the Association for Computational Linguistics, 11, 1553-1571.*

Martinez, F., Weiss, G. M., Palma, M., Xue, H., Borelli, A., & Zhao, Y. (2024). GPT vs. Llama2: Which Comes Closer to Human Writing?. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 107-116).

Nath, S., Li, J. (V.), & Zhang, Y. (2024, August 2). Few-shot prompt engineering and fine-tuning for LLMs in Amazon Bedrock. *AWS Machine Learning Blog.* https://aws.amazon.com/blogs/machine-learning/few-shot-prompt-engineering-and-fine-tuning-for-llms-in-amazon-bedrock/

Riley, P., Constant, N., Guo, M., Kumar, G., Uthus, D. and Parekh, Z. (2020). TextSETTR: Few-shot text style extraction and tunable targeted restyling. *arXiv preprint arXiv:2010.03802.*

Sawicki, P., Grzes, M., Góes, L.F., Brown, D., Peeperkorn, M., Khatun, A. and Paraskevopoulou, S. (2023). On the power of special-purpose GPT models to create and evaluate new poetry in old styles.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652.*

Zhang, X., Rajabi, N., Duh, K., & Koehn, P. (2023, December). Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 468-481).

Zou, J., Zhou, M., Li, T., Han, S., & Zhang, D. (2024). Promptintern: Saving inference costs by internalizing recurrent prompt during large language model fine-tuning. *arXiv preprint arXiv:2407.02211.*

OpenAI. (2025). *ChatGPT (GPT-4.5, accessed via https://chat.openai.com)* [Large language model]. https://openai.com/chatgpt

Thank you!

Project Github link

07 Conclusion & Future Work

Conclusion

- We successfully built an end-to-end pipeline to **generate**, **score**, and **verify** personalized short-form content.
- Using **prompt engineering**, **style vector scoring**, and **authorship classification**, we implemented a system that:
  - Mimics the **tone, structure, and emotion** of user-authored texts
  - Ranks rewrites using **semantic and stylistic fidelity**
  - Validates AI outputs using a separate **classification model**
- Our approach combined concepts from **STYLEMC**, **TextSETTR**, and **metadata-driven generation** to build a solution that's both **grounded and scalable**.

Future Work

1. **Expand Dataset**
   Incorporate more long-form personal writing samples for richer style vectors and broader context modeling.
2. **Fine-Tune LLMs**
   Move from prompt-based control to **actual fine-tuning** using custom training loops on user data.
3. **Multimodal Personalization**
   Explore style embeddings from voice notes, videos, or social media content to enrich author profile.
4. **Interactive Human-in-the-Loop Tool**
   Build a front-end platform that allows users to:
   - Provide feedback on AI outputs
   - Adjust tone sliders (e.g., empathetic ↔ witty)
5. **Deploy as a Writing Assistant**
   Integrate the pipeline into tools like **Notion**, **Google Docs**, or **email clients** for real-time suggestion and generation.