

Evaluating User-Style Adaptation for Professional Text Generation

Stanford CS224N Custom Project

Andrea Nam

Department of Computer Science
Stanford University
anamsong@stanford.edu

Abstract

Personalizing text generation to match an individual’s writing style is challenging, particularly in professional communication where stylistic consistency matters. Standard alignment techniques optimize for an average user while parameter-based per-user methods require significant compute, and inference-time approaches like HyPerAlign rely on unstructured style descriptions that leave open which stylistic dimensions drive personalization quality. This project extends HyPerAlign with a structured 11-field email schema decomposing writing style into actionable directives, augmented with user-distinctive vocabulary extracted via TF-IDF, and a lightweight prompt optimizer that identifies the minimal field subset for effective personalization. Using the Enron Email Dataset, I compare six methods evaluated on stylistic fidelity, task quality, and structure compliance, benchmarked against an empirical performance ceiling. Results show that the optimized three-field schema — greeting style, punctuation habits, and distinctive voice — closes over half of the gap to the performance ceiling with a single exemplar, outperforming both unstructured hypothesis prompting and lightweight LoRA fine-tuning, suggesting that targeted surface-level style conditioning is a more data-efficient personalization strategy than either approach.

1 Key Information to include

- Mentor: Minsik Oh
- External Collaborators (if you have any): N/A
- Sharing project: N/A
- Late Days (if you are using any): N/A

2 Introduction

Large language models (LLMs) such as GPT, Gemini, and Claude are trained to serve a broad population of users, optimizing for responses that are broadly acceptable rather than stylistically distinctive. Standard alignment techniques optimize for an average user by aggregating preferences across many annotators, producing outputs that lack individual voice. In practice, even when a user repeatedly corrects a model’s tone, vocabulary, or structure across multiple conversations, the model not always retain those preferences as each session resets to the same generic prior. This limitation is especially apparent in professional writing tasks such as email generation, where stylistic consistency matters: an email written in someone’s voice must reflect their characteristic and not merely accomplish the communicative task.

Standard alignment approaches optimize for an average user, producing outputs that lack individual voice. Parameter-based methods fine-tune model weights on a user’s history, capturing

style at the cost of significant compute and a separate model per user [1] [2]. Inference-time methods avoid weight updates entirely: HyPerAlign [3] infers free-form natural language hypotheses about a user’s style from a few examples and injects them into the prompt at generation time, demonstrating strong stylistic recovery with no fine-tuning. However, HyPerAlign’s unstructured style descriptions leave open which stylistic dimensions actually drive personalization quality, and whether a more structured representation could improve performance with fewer, more targeted fields.

This project aims to improve inference-time personalization for email generation by replacing HyPerAlign’s free-form style hypotheses with a structured 11-field schema that decomposes writing style into actionable, field-specific directives, augmented with user-distinctive vocabulary extracted via TF-IDF. To identify the minimal set of fields needed for effective personalization, I conduct an exhaustive ablation study over all field combinations of sizes 1-3 and further compare against per-user LoRA fine-tuning [4] as a parameter-based baseline, to evaluate whether lightweight adaptation can match inference-time structured prompting under low-data conditions.

3 Related Work

Personalization in language models can be achieved through parameter-based or inference-time approaches. One PEFT Per User (OPPU) [5] attaches a small adapter to each user and trains it on their interaction history, showing improvements on the LaMP personalization benchmark [6]. Moreover, Panza [7] presents a fully local personalized email assistant that fine-tunes an open-source LLM on a user’s sent emails, combining Reverse Instructions-style fine-tuning with retrieval-augmented generation to mimic individual writing styles. While OPPU demonstrates that personalization can be encoded in weights rather than prompts, and Panza proves that strong email-style personalization can be achieved with limited data under realistic compute and privacy constraints, they both rely on per-user fine-tuning and a separate adapter storage or separate personalized models, respectively. This project instead explores an inference-time, schema-based approach in which a single base model is conditioned on structured style descriptors, avoiding per-user model training while still targeting high-fidelity stylistic control for email generation.

Beyond model-side personalization, there is growing interest in systematic prompt optimization. Maniar & Fu-Hinthorn’s [8] study of prompt optimization shows that LLM-driven methods can systematically improve prompts and automate manual prompt engineering, supporting my implementation of a simple, exhaustive subset search over schema fields to automatically identify the most effective field combination for style personalization. Additionally, Gărbacea & Tan [3] propose HyPerAlign, which decomposes personalization into hypothesis generation and hypothesis-guided prompting, separating user modeling from base model training entirely. This project expands HyPerAlign by replacing its free-form hypotheses with a structured 11-field schema, motivated by prior study on stylometric analysis of email content demonstrating that surface-level features such as length distributions are important to distinguish authors with high accuracy [9]. The project also uses Hu et al.’s LoRA [4] as a parameter-based comparison point to evaluate whether inference-time structured prompting can match lightweight fine-tuning under low-data conditions.

4 Approach

I framed personalized email generation as a style-conditioned text generation problem. Given a user’s email corpus $\mathcal{C} = \{e_1, \dots, e_n\}$ and a scenario prompt s (derived from a held-out email’s subject line), the goal was to generate an email $\hat{e}$ that accomplished the communicative task while matching the user’s stylistic profile. I compared six methods of increasing personalization:

- **Base (No Personalization):** I generated an email from the scenario prompt alone with no style information, serving as the unpersonalized lower bound.
- **Naïve Few-Shot:** I prepended k raw emails from the user’s corpus directly into the prompt, relying on the model to implicitly infer style from examples without any explicit style representation.
- **Generic Hypothesis:** Inspired by HyPerAlign [3], I developed an extraction model summarizing k exemplars into 6-10 free-form style bullets, which were injected into the generation prompt as explicit but unstructured style guidance.

- **Structured Prompt (proposed):** A labeled 11-field representation covering greeting style, sign-off, directness, hedging phrases, typical length, paragraphing, ask placement, bullet usage, punctuation habits, formality, and distinctive voice. Unlike free-form bullets, the schema produced actionable, field-specific directives injected directly into the generation prompt. User-distinctive vocabulary was extracted via TF-IDF over all user corpora, treating each user’s concatenated emails as a single document. The vectorizer operated over unigrams and bigrams (up to 5,000 features), and the top 10 highest-scoring terms and phrases for each user were appended to the schema as a grounded lexical signal. The schema extraction prompt and TF-IDF vocabulary pipeline were designed for this project, building on scikit-learn’s TfidfVectorizer with custom preprocessing and Enron-specific filtering.
- **Optimized Prompt (proposed):** An ablated variant of the structured prompt identified via a lightweight prompt optimizer, an exhaustive subset search over all field combinations of sizes 1-3: greeting style, punctuation habits, and distinctive voice. The ablation procedure and subset search were implemented from scratch.
- **LoRA Fine-Tuning:** A parameter-based baseline using per-user Low-Rank Adaptation [4] of the generation model. Following Hu et al., trainable rank- r matrices were injected into the query and value projections of all attention layers. The forward pass for each adapted layer computes:

$$h = W_0x + \frac{\alpha}{r}BAx \quad (1)$$

where W_0 is the frozen base weight, $A \in \mathbb{R}^{r \times d}$ is initialized from a random Gaussian distribution and $B \in \mathbb{R}^{d \times r}$ is initialized to zero, so $\Delta W = BA = 0$ at the start of training. The LoRA implementation was written based on Hu et al and implemented using Modal for GPU compute with custom in-place LoRA injection and removal to enable per-user fine-tuning without model duplication. Each user’s adapter was trained on up to 50 emails for 5 epochs using AdamW [4] with a linear decay schedule. This served as a comparison point for whether inference-time prompting can match lightweight weight-based adaptation under low-data conditions.

All prompt-based methods used identical decoding settings (temperature = 0.7) and generation constraints to isolate the style representation as the only variable. For LoRA, the same generation prompt as the base method was used with style encoded implicitly in the adapter weights rather than the context.

The 11 fields were designed based on a review of common stylistic dimensions discussed across authorship attribution and personalized generation literature, covering structural choices (greeting style, sign-off, paragraphing), lexical habits (hedging phrases, distinctive voice, formality), and communicative patterns (directness, ask placement, bullet usage, typical length). The fields were selected to be orthogonal, actionable as prompt directives, and collectively expressive enough to capture individual writing style.

5 Experiments

5.1 Data

I used the Enron Email Dataset [10], a publicly available corpus containing approximately 500,000 professional emails from over 150 users. For computational feasibility, I selected the top 10 users by email count, each with at least 50 authored emails to ensure sufficient data for style modeling.

For each user, emails were split chronologically into a style corpus (80%) and a held-out evaluation set (20%). The style corpus provided personalization signal for prompt-based methods (exemplars and schema extraction inputs), training data for the per-user LoRA adapters, and training data for the authorship classifier used in evaluation. The held-out set was used as validation set for the authorship classifier, establishing an empirical performance ceiling [11]. It was also the source of evaluation scenarios, which consisted of a subject line drawn from the held-out set that served as the generation prompt with the original email body retained as the gold reference for evaluation. For each user, 3 scenarios were sampled from the held-out set, yielding 30 evaluation instances per method.

As part of preprocessing, I removed quoted replies and signatures, redacted personally identifiable information, filtered out automated messages and emails under 50 or over 5,000 characters, and normalized whitespace.

5.2 Evaluation method

I evaluated the six methods across three dimensions, favoring rule-based and embedding-based approaches that do not rely on LLM judges or reference summaries.

- **Stylistic fidelity:** It was measured by training a multi-class authorship classifier on each user’s style corpus and scoring generated emails against it. I report both $P(\text{true author})$, the continuous probability assigned to the correct author, and authorship accuracy, the binary classification decision (Table 1). The classifier’s accuracy on real held-out emails (0.757) serves as a performance ceiling.
- **BERTScore F1** [12]: It measures semantic similarity between the generated email and the gold held-out email body.
- **Structure compliance:** It is a rule-based score computed as the mean of three binary checks: presence of a greeting, a sign-off, and an explicit ask or action request. It ranges from 0 to 1 and captures task compliance independent of stylistic personalization.

5.3 Experimental details

I used Gemini 2.0 Flash (via Google Vertex AI) for hypothesis extraction and schema generation, and Qwen2.5-7B-Instruct [13](via Modal, A10G GPU) for email generation across all methods. Generation temperature was set to 0.7 for all methods and extraction temperature was set to 0.2–0.3. All generation prompts included identical constraints: word count bounds derived from the user’s schema, a fixed structural rubric, and a no-copying instruction.

For the ablation study used to identify the Optimized Prompt fields, I first ran a greedy forward-backward search to estimate the optimal subset size. Forward selection peaked at size 2 ($P(\text{true author}) = 0.913$). However, this result was obtained on a smaller pilot set of 5 users and exceeded the classifier ceiling, suggesting overfitting. To obtain a more reliable estimate, I performed an exhaustive search over sizes 1–3 on the full 10-user set, selecting the combination that maximized mean $P(\text{true author})$ across users. The ablation procedure was implemented from scratch.

For LoRA, I used rank $r = 4$ and scaling factor $\alpha = 4$ trained with AdamW at learning rate 2×10^{-4} with linear decay for 5 epochs on up to 50 emails per user. LoRA matrices were injected into the query and value projections of all attention layers of Qwen2.5-7B-Instruct.

The authorship classifier was implemented as a logistic regression model trained on sentence embeddings of each user’s style corpus. Email bodies were encoded using the all-MiniLM-L6-v2 sentence transformer [14], and a multi-class logistic regression was trained to predict the correct author among all 10 users, where max iterations = 1000, $C = 1.0$, balanced class weights. The classifier was validated on real held-out emails before scoring generated outputs, which established the empirical performance ceiling.

I ran experiments over 10 users with 3 held-out scenarios per user. For prompt-based methods, I swept $k \in \{1, 3, 5\}$ style exemplars. Since performance deteriorated with increasing k , the Optimized Prompt and LoRA experiments were conducted at $k = 1$ only.

5.4 Results

Table 1 reports mean evaluation scores at $k = 1$ across all six methods over 10 users. HyPerAlign Improved and HyPerAlign Optimized substantially outperform all baselines on stylistic fidelity, achieving $P(\text{true author})$ of 0.485 and 0.483 respectively, a 77% improvement over No Persona. HyPerAlign Optimized achieves the highest authorship accuracy (0.600) and BERTScore F1 (0.724), confirming that the three-field optimized schema matches or exceeds the full 11-field schema across all four metrics while using fewer fields.

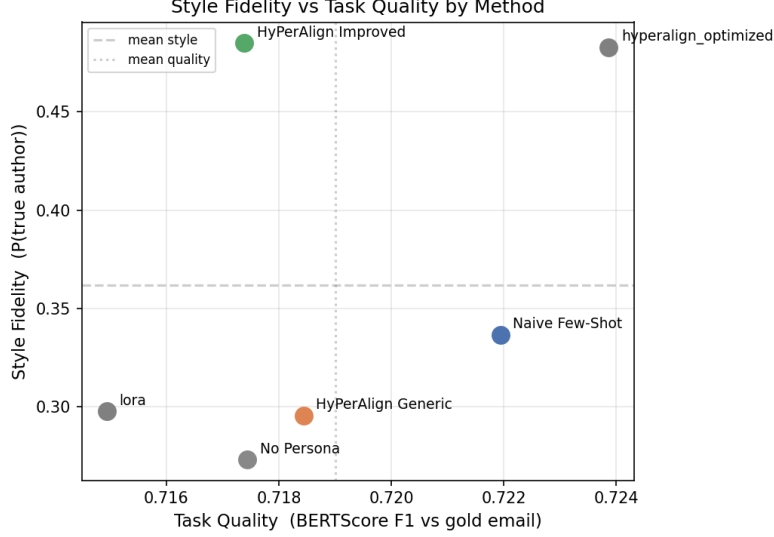


Figure 1: Style fidelity ($P(\text{true author})$) vs. BERTScore F1 (BERTScore F1) for all six methods. Top-right indicates better performance on both dimensions simultaneously.

Method	$P(\text{true author})$	Authorship Accuracy	BERTScore F1	Structure Score
No Persona	0.273	0.267	0.717	0.667
Naive Few-Shot	0.337	0.367	0.722	0.678
LoRA	0.298	0.433	0.715	0.667
HyPerAlign Generic	0.295	0.333	0.718	0.567
HyPerAlign Improved	0.485	0.567	0.717	0.578
HyPerAlign Optimized	0.483	0.600	0.724	0.645

Table 1: Mean evaluation scores at $k = 1$ over 10 users.

Figure 1 plots style fidelity against BERTScore F1 for all six methods. The proposed methods occupy the top-right region, achieving both higher style fidelity and competitive BERTScore F1. The two results that were more surprising than expected were that Naive Few-Shot surpasses the performance of HyPerAlign Generic, despite the later one using more explicit style guidance (Appendix A.2 A.3). And, LoRA fails to improve meaningfully over the generic, No Persona on BERTScore F1, despite encoding style directly in model weights. LoRA does achieve higher authorship accuracy, indicating that it captures some stylistic signal, although not enough to shift the classifier’s probability distribution substantially. Figure 2 shows style fidelity as a function of k for prompt-based methods. Contrary to my expectations, performance deteriorates consistently as k increases across all three methods with HyPerAlign Improved showing the steepest decline. I expected more exemplars to strengthen style signal but the opposite holds, suggesting that additional exemplars may introduce noise rather than useful personalization context.

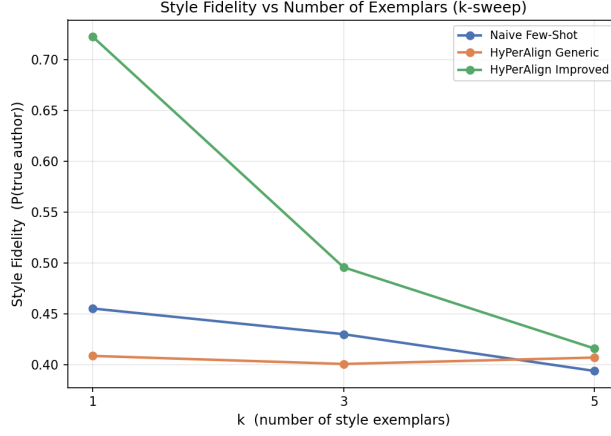


Figure 2: Style fidelity ($P(\text{true author})$) as a function of number of style exemplars k for prompt-based methods. Performance deteriorates with increasing k across all methods.

Figure 3 shows the mean $P(\text{true author})$ when each field is included across all subsets of sizes 1-3. Greeting, punctuation, and voice consistently achieve the highest mean scores while directness, paragraphing, and sign-off contribute less on average. This ranking was largely expected, as surface-level stylistic markers such as greetings and punctuation are more directly observable than higher-level rhetorical choices.

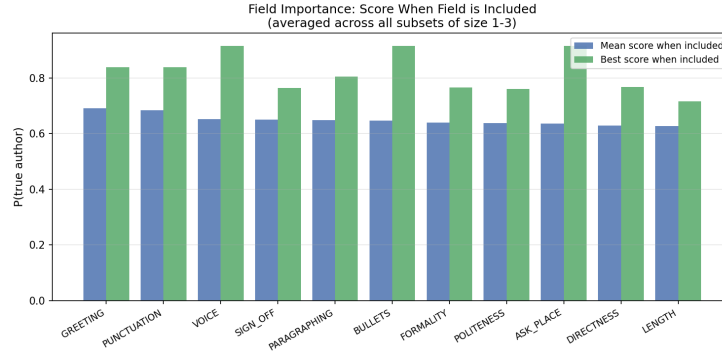


Figure 3: Mean $P(\text{true author})$ when each schema field is included, averaged across all subsets of sizes 1–3.

Figure 4 shows the pairwise field interaction scores. The punctuation-voice pair and sign-off-voice pair pairs reach the highest scores (0.756 and 0.765 respectively), which was somewhat surprising as I expected content-oriented fields such as ask placement and bullets to lead pairwise interactions but stylistic surface markers proved more informative, concurring with Figure 3.

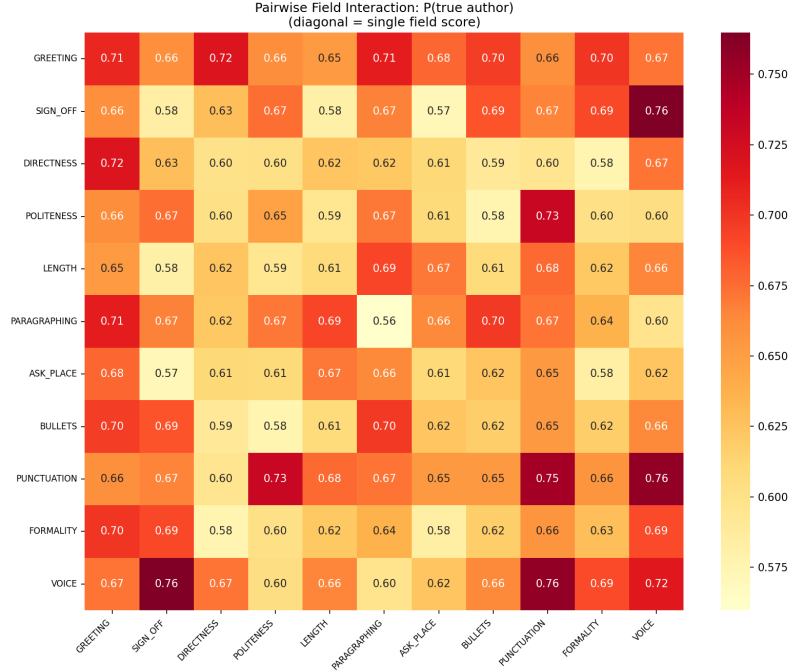


Figure 4: Pairwise field interaction heatmap showing $P(\text{true author})$ for all field pair combinations.

Overall, these results demonstrate that the structured schema approach meaningfully improves stylistic personalization over all baselines. The prompt optimizer successfully identified a minimal three-field configuration that matches the full 11-field schema in style fidelity while achieving the highest scores across authorship accuracy, BERTScore F1, and structure compliance. Such results support my hypothesis that targeted, structured prompt conditioning is a more effective personalization strategy than either unstructured guidance or lightweight fine-tuning under low-data conditions.

6 Analysis

The exhaustive subset search identified greeting, punctuation, and voice as the optimal field combination. The highest-scoring triple—ask place, bullets, and voice—was excluded because it excessively surpasses the classifier ceiling, indicating overfitting rather than genuine style recovery. The selection was motivated by two converging signals: the single-field scores and the mean score when included across all subsets of sizes 1–3 (Figure 3), where the same three fields ranked among the highest on average. The gap to the next best field, politeness, provides additional support that these optimal three fields carry the strongest independent style signal.

Field Combination	$P(\text{true author})$
ASK_PLACE + BULLETS + VOICE	0.916
GREETING + PUNCTUATION + VOICE	0.840
GREETING + PARAGRAPHING + VOICE	0.806
GREETING + PARAGRAPHING + PUNCTUATION	0.792
GREETING + POLITENESS + ASK_PLACE	0.761
Classifier Ceiling	0.757

Table 2: Top-5 triples from exhaustive subset search.

HyPerAlign Optimized closes the gap to the classifier ceiling using just three schema fields and a single exemplar. While a gap of 0.173 remains, this comparison is asymmetric: the ceiling reflects a classifier trained on thousands of emails per user while the generation model receives only a minimal stylistic summary. The ceiling itself is not a perfect target since at 0.757 authorship accuracy, roughly 24% of real held-out emails are misclassified even when written by the true author. This means the true achievable upper bound may be lower than 0.656, making the optimized prompt’s performance a

stronger result than the raw gap suggests.

Inspecting individual outputs reveals consistent behavioral differences between methods. No Persona and HyPerAlign Generic produce formally structured emails that accomplish the communicative task but lack distinctive stylistic markers. No Persona defaults to a generic, formally structured emails that accomplish the communicative task but fails to adapt to any individual, which explains its high structure score but low $P(\text{true author})$ in Table 1. HyPerAlign Improved and Optimized more reliably reproduce user-specific patterns consistent with the schema fields. The low $P(\text{true author of LoRA})$ suggests that the fine-tuning signal is too weak to override the base model’s pretraining distribution at 50 examples. Moreover, as shown in Figure 3, length instructions proved to have limited influence on generation behavior. Considering that generated emails frequently fell near the lower bound of the target word range, the model tends to default to brevity regardless of the length directive specified in the schema. This points to a general limitation of prompt-based length control rather than a method-specific failure.

7 Conclusion

In this project, I extended HyPerAlign with a structured 11-field email schema and a lightweight prompt optimizer to improve inference-time personalized email generation, eliminating the need for model retraining or multiple historical interactions. Using the Enron Email Dataset, I compared six methods (No Persona, Naive Few-Shot, LoRA, HyPerAlign Generic, HyPerAlign Improved, and HyPerAlign Optimized) evaluating them on stylistic fidelity, task quality, and structure compliance against an empirical performance ceiling defined by the authorship classifier’s accuracy on real held-out emails.

The results show that surface-level stylistic fields such as greeting style, punctuation habits, and distinctive voice contribute more to personalization performance than higher-level communicative fields such as directness or ask placement. The optimized three-field schema matches or outperforms the full 11-field schema across all metrics whereas LoRA underperforms both proposed methods despite being a parameter-based approach, suggesting that inference-time structured prompting is a more data-efficient personalization strategy under low-data conditions.

This project has several limitations. The evaluation was conducted on only 10 users with 3 scenarios each, which limits the generalizability of the findings. Performance also deteriorates with increasing k , suggesting that the current prompting strategy does not scale well with more exemplars. Future work could explore combining structured schema prompting with other inference-time techniques, evaluating on a broader set of users and domains, and investigating whether larger LoRA ranks or more training data close the gap with prompt-based methods.

References

- [1] Koptyra B. Janz A. Kazienko P. Kocón J. Wózniać, S. Personalized large language models. In *arXiv*, 2024.
- [2] Patel C. Shahs, A. Mastering llm techniques: Customization. In *NVIDIA*, 2023.
- [3] Tan C. Garbacea, C. Hyperalign: Interpretable personalized llm alignment via hypothesis generation. In *arXiv*, 2025.
- [4] Shen. Y Wallis P. Allen-Zhu Z. Li Z. Wang S. Wang L. Chen W. Hu, E. Lora: Low-rank adaptation of large language models. In *arXiv*, 2021.
- [5] Zeng Q. Tian Y. Liu-Z. Yin B. Jiang M. Tan, Z. Democratizing large language models via personalized parameter-efficient fine-tuning. In *arXiv*, 2025.
- [6] S. Bendersky M. Zamani H. Salemi A., Mysore. Lamp: When large language models meet personalization. In *arXiv*, 2024.
- [7] Iofinova E. Jovanovic A. Kurtic E. Nikdan M. Panferov A. Nicolicioiu, A. Panza: Design and analysis of a fully-local personalized text writing assistant. In *arXiv*, 2025.

- [8] Iofinova E. Jovanovic A. Kurtic E. Nikdan M. Fu-Hinthorn W. Maniar, K. Exploring prompt optimization. In *LangChainBlog*, 2025.
- [9] Vivek K. V. Kodimala S. Singh, P. K. Stylometric analysis of e-mail content for author identification. In *ACM*, 2017.
- [10] Enron email dataset. <https://www.kaggle.com/datasets/wcukierski/enron-email-dataset>, 2015.
- [11] Ahmed Mohsin M. Bikia V. Singhvi A. Gaus R. Bedi S.-Cui H. Fuentes M. Unell A. Mai Y. Cahoon J. Pfeffer M. Daneshjou R. Koyejo S. Alsentzer E. Potts C. Shah N. H. Chaudhari A. S. Asad, A. Structured prompting enables more robust evaluation of language models. In *arXiv*, 2025.
- [12] Kishore V. Wu F. Weinberger K. Q. Artzi Y. Zhang, T. Bertscore: Evaluating text generation with bert. In *arXiv*, 2020.
- [13] Qwen Team. Qwen2.5 technical report. In *arXiv*, 2025.
- [14] Gurevych I. Reimers, N. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

A Appendix

A.1 No Persona (Basic Prompt)

You are a professional email writing assistant. Write a professional email for the scenario below.

Consider these constraints:

- Length: {min_words-max_words}
- Structure: Greeting line; 1-3 short paragraphs; clear ask/next step; sign-off + [NAME]
- Do not invent facts. Use only what the scenario provides.
- Output only the email.

A.2 Naive Few-Shot Prompt

You are a professional email writing assistant. Below are example emails written by the same user. Use them only as style inspiration. Do not copy phrases verbatim. Do not reuse specific names, dates, or facts from examples.

EXAMPLES: ...

Now write an email for the scenario below.

Consider these constraints:

- Length: {min_words-max_words}
- Structure: Greeting line; 1-3 short
- Structure: Greeting line; 1-3 short paragraphs; clear ask/next step; sign-off + [NAME]
- Do not invent facts.
- Output only the email.

SCENARIO: ...

A.3 HyPerAlign Generic Prompt

You are a professional email writing assistant. Write an email that matches the style profile below.

STYLE PROFILE: ...

Take into consideration these constraints:

- Length: {min_words-max_words}
- Structure: Greeting line; 1-3 short paragraphs; clear ask/next step; sign-off + [NAME]
- Apply the style profile strictly but write entirely original content.
- Do not copy phrases from the style profile verbatim.
- Do not invent facts. Use only what the scenario provides.
- Do not reuse specific names, dates, or details from the style profile.
- Output only the email.

SCENARIO: ...

A.4 HyPerAlign Improved Prompt

You are a professional email writing assistant. Write an email that follows the STYLE CONTROL BLOCK below.

STYLE CONTROL BLOCK: ...

These are the constraints:

- Length: {min_words-max_words}
- Greeting line; 1-3 short paragraphs; clear ask/next step; sign-off + [NAME]
- Apply every instruction in the style control block strictly but write entirely original content.
- Do not copy phrases from the style control block verbatim.
- Do not invent facts. Use only what the scenario provides.
- Do not reuse specific names, dates, or details from the style control block.
- Output only the email.

SCENARIO: ...

A.5 Style Profile vs. Style Control Block

Both the Generic and Improved methods extract style information from a user’s email corpus but differ in how that information is represented and injected into the generation prompt. The Generic method produces a free-form style profile (STYLE PROFILE: a set of 6-10 unstructured bullet points summarizing the user’s writing style, generated by prompting an extraction model to observe patterns in tone, verbosity, structure, and punctuation. These bullets are observational in nature and left for the generation model to interpret freely.

The Improved and Optimized methods instead produce a style control block: a structured set of labeled directives, one per schema field, where each line maps directly to a concrete writing decision (e.g., greeting choice, sign-off, ask placement, bullet usage). Unlike free-form bullets, each directive is prefixed with an actionable instruction derived from a fixed field map, making the style constraints explicit and independently applicable. The style control block is further augmented with user-distinctive vocabulary extracted via TF-IDF, appended as a list of preferred phrases. For the Optimized method, only the three highest-performing fields — greeting style, punctuation habits, and distinctive voice notes — are included in the style control block with all other fields omitted.

A.6 Example Output of Style Profile of User *X*

Okay, I've analyzed the email examples and here's a style profile:

- Informal Tone: Uses casual language and contractions.
- Direct and Concise: Gets straight to the point with minimal preamble.
- Short Sentences: Predominantly uses short, declarative sentences.
- Simple Structure: Employs a basic greeting and sign-off.
- Limited Politeness: Shows a lack of hedging or softening language.
- Questioning: Uses questions to challenge or probe.
- Minimal Punctuation for Emphasis: Relies on direct statements rather than punctuation for emphasis.
- Abbreviations/Initialisms: Readily uses abbreviations and initialisms (e.g., SWPC, et.al.).

A.7 Example Output of Style Control Block of User *X*

- Use greeting: First name only
- Use sign-off: First name only
- Tone: direct
- Use these politeness phrases when natural: none
- Length: short 50-70
- Structure: 2 short paragraphs
- Place the main ask/request: middle
- Bullet points: never
- Punctuation: uses commas
- Formality: informal
- Voice: blunt and to-the-point

A.8 Example Scenario and Six Emails Generated for User *X*

SCENARIO: Sending over the files

No Persona Email:

Subject: Attached Files

Dear Recipient,

Please find attached the documents you requested. I have included [mention the specific files or types of files].

Do let me know if you have any trouble accessing the files or if you require any further information at this time.

Thank you for your time. I look forward to your response.

Sincerely,
NAME

Naive Few-Shot Email:

Subject: Attached Files

Hi [NAME],

Please find the files you requested attached to this email.

Let me know if you have any questions or need anything else.

Thanks,
NAME

Generic Hypothesis (HyPerAlign Generic) Email:

Subject: Attached Files

Hey,

FYI, I've attached the files we discussed. LMK if you have any Qs.

Can you review and get back to me by EOD Friday?

Thx,
Drew

Structured Prompt (HyPerAlign Improved) Email:

Subject: Attached Files

Hey Bob,

I'm sending over the documents we chatted about yesterday - just wanted to get them to you ASAP.

Can you check them out and tell me if that works for your corp?

Thanks,
Jen

Optimized Prompt (HyPerAlign Optimized) Email:

Subject: Attached Filed

Hey Ron,

Attached are the reports from this quarter - sales, marketing, ect. Some files are large, kay?

Review everything ASAP. Let me know if you see any issues, kay mann.

Thanks,
Jen

LoRA Email:

Subject: Attached Files

Dear [RECIPIENT NAME],

I hope this email finds you well. I have attached the documents we discussed during our last meeting for your review and approval. Please let me know if you need any further information or have any questions.

Best regards,
YOUR NAME