Disentangling Style and Semantics for Calendar-Driven Text Generation: A Knowledge Graph-Guided Activation Steering Approach

Tanmay Kumar Shrivastava
BTech Final Year, DSAI, IIT, Bhilai
Durg, Chhattisgarh
tanmayku@iitbhilai.ac.in

Aditya Bajpai
BTech Final Year, CSE, IIT, Bhilai
Durg, Chhattisgarh
adityabajpai@iitbhilai.ac.in

Anil Kumar Mandal
BTech Final Year, CSE, IIT, Bhilai
Durg, Chhattisgarh
anilkumarm@iitbhilai.ac.in

Abstract

This paper presents a new kind of personalization setup which is quite jovial for large language models. It focuses on handling communications tied to calendar events. Things like invitations or announcements and even social media posts about events follow a pretty standard pattern with facts such as dates, times and venues. Still they really need to vary a lot in style based on the user. Traditional ways of fine tuning or using retrieval augmented generation just do not manage to pull apart the style from those core facts. That ends up making the whole personalization process unclear and not very effective. The main thing we bring here is a system that combines guidance from a symbolic knowledge graph with something called contrastive activation steering. This method clearly splits out the stylistic parts like whether its formal or casual in tone or how the phrasing works. It keeps those separate from the basic semantic info that comes straight from the knowledge graph. We primarily provide this method, which combines contrastive activation steering with guidance from a symbolic Knowledge Graph, or KG. With this configuration, signals for style and those for semantics are clearly separated. Specifically, we created a domain-specific knowledge graph. It contains factual information regarding events in an organised manner, such as entities and their relationships. We employ semantic masking guided by the KG to automatically extract stylistic motifs. The semantic components are removed in the masking stage. It eliminates the purely stylistic components, such as tone, degree of formality, phrasing patterns, and even punctuation. These masked forms are then used to train our framework. It then extracts style vectors specific to each user from that. Personal writing habits are grasped by these vectors. They handle this without confusing any real details. Simultaneously, everything remains firmly grounded in the facts of the actual occurrence. The knowledge graph provides a strong foundation in the facts when it comes to text generation. Simultaneously, those learnt style vectors are fed into the model's hidden layers through activation steering. This makes it simple for us to manage and comprehend the customised text output. Semantic correctness is a strong point of the entire approach. It also maintains consistency in stylistic features. Additionally, it only needs to save one style vector per user. This maintains scalability for practical use.

1 Introduction

Personalized text generation has emerged as one of the most active research areas in modern natural language processing (NLP), largely driven by the success of large language models (LLMs) like GPT, LLaMA, and Claude. While these models have shown exceptional capability in generating fluent and context-aware text, they often lack user-specific personalization, especially when it comes to consistent tone, phrasing, and writing style. For example, one person may prefer to send formal and concise event invitations, while another may choose casual and friendly tones. Achieving this level of personalization without losing factual correctness remains an open challenge.

Calendar-driven communications—such as event invitations, announcements, or social media updates—are among the most frequent forms of short-form text generation. These messages typically share a repetitive factual structure (containing event name, time, date, venue, and host), but differ widely in stylistic expression. A formal tone might say, "You are cordially invited to attend the Annual Symposium on October 5," while a casual tone might say, "Hey everyone! Don't miss our TechFest this Saturday!" Both convey the same information, yet their linguistic style and emotional tone differ significantly.

Traditional fine-tuning or retrieval-based personalization methods struggle to separate style from content. When trained jointly on factual and stylistic information, models tend to "mix" them, leading to semantic leakage—where stylistic shifts unintentionally alter event details. Moreover, maintaining separate fine-tuned models for every user is computationally expensive and lacks interpretability. There is a clear need for a more modular, interpretable, and scalable approach that isolates stylistic behavior from factual grounding.

Our proposed solution introduces a Knowledge Graph (KG)-guided disentanglement mechanism for style and semantics. By representing factual event details as nodes and relations in a KG, the model can explicitly recognize and separate factual content from stylistic cues. This allows us to "mask" factual elements (like event name or venue) during style extraction, ensuring the model learns only the tone, phrasing, and punctuation patterns unique to each user. Later, this stylistic knowledge is re-injected into the generation process through activation steering, a lightweight method that adjusts hidden layer activations in the LLM to reflect the user's personal writing style.

The motivation behind this project is deeply human-centered. In everyday communication—whether professional or social—people unconsciously express themselves in distinct ways. Reproducing such nuances in AI-generated text is essential for making language models feel authentic, context-aware, and human-aligned. The proposed method not only enhances personalization but also provides transparency and interpretability—two key pillars of responsible AI research. Unlike black-box fine-tuning, our approach explicitly explains where factual knowledge ends and stylistic influence begins, ensuring factual accuracy even in creatively personalized outputs.

In summary, this project aims to advance the interpretability and efficiency of personalized text generation by combining structured semantic representation (Knowledge Graphs) with neural activation manipulation (Activation Steering). The outcome is a system that can generate calendar-driven communications that sound uniquely tailored to a specific user while staying perfectly accurate in terms of event details.

2 Related Work

Recent years have seen an astonishing development in both personalized text generation and knowledge graph construction with the rise of large language models. However, balancing the structured semantic grounding with personalized stylistic expression still constitutes a challenging problem. Previous approaches to this goal have tended to emphasize one of these aspects: either improving factual consistency by structured representations or expressive, human-like personalization.

This section describes the essential bodies of work that motivate our study, namely: the evolution of unified knowledge graphs, style disentanglement in text generation, and activation-based mechanisms that provide the bases for our proposed method.

2.1 Knowledge Graphs and the Push Toward Unification

Knowledge Graphs have become central to modern NLP due to their representation of information in a more comprehensible form using triples of <entity, relation, entity>. The structured insights provided by them are immensely helpful for reasoning, retrieval, and question-answering systems. Over time, this basic framework has evolved to include Event Knowledge Graphs, which map dynamic event relationships, and Commonsense Knowledge Graphs, which capture more abstract knowledge and reasoning of everyday human life.

Traditionally, these graphs have been built in isolation: knowledge graphs focus on factual entities, evolving knowledge graphs emphasize temporal or causal relations, and conceptual knowledge graphs encode general knowledge. While this separation simplifies modeling, it often results in redundant computation and fragmented knowledge, which limits the capacity of a model to generalize across domains.

To address this, recent studies such as those by Krause et al. (2022) and Peng et al. (2023) proposed the concept of a Generalized Knowledge Graph, or GKG. This unified framework positions the KG, EKG, and CKG as part of the same hierarchy, in which the KG records facts, the EKG models events that change over time, and the CKG expresses patterns of abstract reasoning. A particularly noteworthy contribution in this area concerns a three-stage curriculum learning framework in which an LLM is progressively fine-tuned through stages of KG Empowerment, EKG Enhancement, and CKG Generalization. Consequently, the model learns structured knowledge step by step to enhance the capability of handling in-domain, counter-task, and out-of-distribution data.

This research is particularly relevant to our work. By encoding event-specific facts such as date, venue, and host in a structured knowledge graph, we are able to disentangle factual information from stylistic cues. Rather than the model implicitly inferring both, the KG explicitly encodes what is factual. In this work, our model is then free to focus on learning the users' style-tone, phrasing, and structure-without its being confused with semantic content. The idea of GKG basically acts as the semantic backbone to our approach for style-content disentanglement.

2.2 Personalized Text Generation and Style Disentanglement

While Knowledge Graphs provide a stable structure for factual understanding, personalized text generation concerns how language models can adapt to an individual's unique voice. As AI systems become more integrated into daily communication, such as writing emails, announcements, or social media posts, there is an increased desire to have text not only with the right information but also one that sounds genuinely human and contextually appropriate.

Personalization has historically been defined by two main paradigms:

RAG: Retrieval-Augmented Generation, which retrieves a user's past writings to emulate tone and phrasing.

PEFT is a fine-tuning approach that adapts small model components, such as LoRA layers, on each user's data.

While effective, both methods have notable weaknesses. RAG methods often end up in an overemphasized semantic similarity problem: it retrieves text on topic, not the style, which makes the tone sound flat or unconsistent. PEFT solves the problem introduced by RAG but brings in scalability challenges: maintaining separate adapter modules for each user quickly gets heavy.

Addressing these issues, Zhang et al. (2025) introduced the StyleVector framework, a training-free approach representing a user's writing style as a direction in the model's activation space. Comparing activations between user-authored texts-carrying both content and style-with neutral model outputs that contain only content, StyleVector identifies the "difference" capturing stylistic traits like formality, rhythm, and tone. This style vector can then be injected into the model during inference, allowing it to generate stylistically aligned text without additional training or retrieval.

Despite success, StyleVector still faces semantic leakage, which means the model changes factual details while performing style steering. In this project, we extend this framework by combining Knowledge Graph-guided masking with activation-space steering. Masking event entities (such as names, venues, and dates) ensures that only stylistic signals remain during style extraction. This makes the resulting style vector purely stylistic, enabling consistent personalization without loss of semantic fidelity.

In short, we draw from two lines of research for our approach.

From StyleVector, we adopt the mechanism of contrasting activations for style extraction.

From GKG research, we utilize structured semantic encoding in order to avoid factual interference.

Taken together, these ideas let us reach: interpretable, controllable, semantically precise personalization.

2.3 Bridging Symbolic Knowledge and Neural Personalization

A key pattern emerging from recent AI research is the blending of symbolic reasoning with neural representation learning. Symbolic structures, such as Knowledge Graphs, provide explainable, structured knowledge, while neural systems excel at generating fluid, human-like language. Bridging these two worlds offers a way to build systems that are both accurate and expressive. This philosophy is reflected in our project. We combine logical reasoning with stylistic adaptation by anchoring the factual event details in a Knowledge Graph and guiding the neural network's activations through style vectors. Such a combination will yield a model capable of generating factually accurate and stylistically different communications of calendar events. This is in line with the increasing trend in human-centered AI for models to reason reliably while showcasing human individuality in their responses.

2.4 Contrastive Activation Addition (CAA): The Foundation of Activation Steering

A key conceptual foundation for StyleVector, and by extension our approach, is Contrastive Activation Addition, introduced by Rimsky et al. (2024). The central idea behind CAA is easily stated: desired behaviors in language models can be isolated as directions in activation space.

CAA constructs a steering vector to shift a model's output distribution toward some desired property, such as honesty, confidence, or empathy, during inference. Using paired prompts-one that demonstrates the target behavior and one that demonstrates its opposite-CAA keeps the content constant while changing only the behavioral outcome. By doing so, it can calculate the mean difference of activations between such pairs of prompts. This difference isolates the internal representation of the target attribute within the model's hidden layers.

Formally, given a dataset of positive-negative prompt pairs, the MD vector at a given layer captures the direction that differentiates the two behaviors. Adding this vector to model's activations during inference "steers" the model towards the desired outcome without re-training.

For example, CAA has been applied to control behaviors such as sycophancy, refusal, hallucination, and corrigibility in LLaMA models. Analyses show that these behavioral signals become linearly separable in mid-level layers, which is indicative of how LLMs naturally encode abstract concepts in specific activation subspaces.

This discovery serves as the direct inspiration behind our proposed design for contrastive style extraction. Our approach contrasts the user-authored text against neutral, model-generated text. The difference in their activations yields an interpretable direction in the model's latent space, which captures personal stylistic features; a style vector, for short. We then extend the concept of CAA from behavior alignment towards personalized style disentanglement by synthesizing it with KG-guided masking, ensuring that factual consistency will never be compromised when generating stylistically rich event communications.

2.5 Summary

In summary, our work stands at the intersection of two leading research directions in modern language modeling: (1) contrastive activation-based style learning, and (2) Knowledge Graph (KG)-guided factual personalization.

The first research thread builds upon the foundations of the *StyleVector* framework and *Contrastive Activation Addition* (CAA), which introduce mechanisms for extracting latent stylistic or behavioral directions directly from the activation space of large language models. By leveraging contrastive activation analysis between neutral and styled examples, we highlight interpretable, continuous style vectors that enable fine-grained personalization of generation behavior. This allows for a scalable and model-agnostic approach to stylistic control, achieving tone modulation and user-adaptive writing without retraining or additional supervision.

The second research thread relates to the use of Knowledge Graphs (KGs) for structured personalization, traditionally applied in retrieval-augmented generation and question-answering systems. KGs are effective at representing user-specific or domain-specific factual knowledge, allowing for contextually grounded responses and ensuring factual consistency. In our work, this paradigm is extended from semantic retrieval to style disentanglement: instead of retrieving factual evidence, the KG acts as a semantic scaffold that isolates factual elements such as *event, host, venue, date,* and *occasion type* from stylistic features in text. This integration allows activation steering to focus purely on stylistic modulation while preserving factual integrity — a crucial step toward interpretable and controllable personalization.

By merging these two lines of research, we propose **Knowledge Graph-Guided Activation Steering** — a unified framework that combines the interpretability of structured knowledge with the controllability of activation-space steering. Applied to the domain of calendar-driven event communication, the approach demonstrates that even small, domain-specific datasets can yield meaningful stylistic expressions when factual and stylistic elements are effectively disentangled. Our experiments emphasize that such disentanglement not only enhances personalization but also improves generation reliability by reducing semantic drift and hallucination.

In perspective, the developing confluence of activation-based control and KG-grounded representation opens new directions for building human-like, context-aware, and style-adaptive language models. Future extensions could explore broader domains, larger

datasets, and multi-layer steering architectures to further bridge the gap between factual accuracy and stylistic expressiveness in personalized text generation.

3 Methodology

Our approach, termed **KG-Guided Disentanglement for Calendar Events**, aims to separate user-specific stylistic tone from factual content in calendar-related communications such as invitation emails. The first step involves generation of style agnostic (neutral) as well as styled responses with the help of a general LLM. Then using Knowledge graph (KG), the email pairs for a user u, can be defactualized. The key idea now is to utilize contrastive activation analysis between *defactualized* styled and neutral text representations to obtain a user-specific *style vector*, which can then be injected into a large language model (LLM) (Llama 2-7b here) to steer its generation toward a desired style. As illustrated conceptually in Figure 1, the framework consists of five major components: The framework operates through a multi-stage pipeline:

(1) Generation of style-agnostic (neutral) as well as styled email responses
(2) Defactualization of styled and neutral email responses using KG
(3) Extraction of the style vector via contrastive activation analysis on defactualized pairs
(4) Activation-based steering during generation
(5) Rehydrating/refactualizing the output using KG.

3.1 Dataset Generation Process

The dataset generation process was designed to produce a collection of **invitation emails** in two distinct tonal styles — *Neutral* and *Casual Friendly* — through the use of the **Google Gemini API**. The procedure began with a file containing event-related information such as the date, time, venue, and host. This data served as the foundation for constructing structured prompts that guided the model's content generation.

The `Gemini 2.5 Flash Lite` model was configured using a valid API key, enabling automated and controlled text generation. Two stylistic frameworks were defined: a *Neutral* tone that maintained a formal, professional, and polite demeanor, and a *Casual Friendly* tone that emphasized warmth, approachability, and conversational fluency. For each event entry, the model received a detailed prompt encapsulating all relevant attributes, and subsequently generated two separate email invitations, each reflecting the corresponding tone.

The generated responses were then incorporated into the same `DataFrame` as new columns `response_Neutral` and `response_Casual_Friendly_User` — ensuring clear organization of the output data. Finally, the augmented dataset containing all tone-specific invitation emails was exported to a new file. Overall, this pipeline effectively automated the creation of stylistically varied and personalized email invitations, resulting in a well-structured dataset suitable for further analysis and evaluation.

Things to Keep in Mind

Properties of Generated Pairs. The key properties we seek to establish are:

(1) **Factual Consistency**: $\text{Facts}(y_i) = \text{Facts}(\hat{y}_i) = \text{Facts}(x_i)$
(2) **Stylistic Divergence**: $\text{Style}(y_i) \neq \text{Style}(\hat{y}_i)$
(3) **Entity Preservation**: $\mathcal{E}(y_i) = \mathcal{E}(\hat{y}_i) = \mathcal{E}_i$

where $\text{Facts}(\cdot)$ extracts semantic information, $\text{Style}(\cdot)$ captures stylistic features, and $\mathcal{E}(\cdot)$ identifies entities.

In practice, both y_i and $\hat{y}_i$ contain the same factual entities (event name, date, time, venue, host) extracted from x_i, but they differ significantly in:

- Lexical choice and vocabulary
- Sentence structure and complexity
- Tone and formality level
- Use of discourse markers and connectives
- Punctuation patterns and emphasis
- Greeting/closing conventions

3.2 Generating KG-Guided Defactualized Style-Agnostic Responses

KG-Guided Disentanglement for Calendar Events, addresses the fundamental challenge of separating user-specific stylistic tone from factual content in calendar-related communications such as invitation emails, meeting notifications, and event announcements. The core insight is that calendar events possess a well-defined knowledge graph (KG) structure with identifiable entity types, which can be systematically used to create defactualized text representations that preserve style while removing factual specifics.

Consider a user u from a user set $\mathcal{U}$, who has authored a collection of calendar event-related emails. Each email corresponds to a structured event with associated metadata. We define the user's dataset as:

$$\mathcal{D}_u = \{(x_i, y_i, \mathcal{E}_i)\}_{i=1}^{N_u}$$

where:

- $x_i \in \mathcal{X}$: Input context containing event metadata (subject, description, purpose)
- $y_i \in \mathcal{Y}$: Original human-authored styled email by user u
- $\mathcal{E}_i = \{e_i^{(1)}, e_i^{(2)}, \ldots, e_i^{(K)}\}$: Set of factual entities present in the email
- N_u: Number of email samples for user u

Calendar Event Knowledge Graph Structure. Calendar events naturally form a knowledge graph with a well-defined ontology. We model this as $\mathcal{KG} = (\mathcal{V}, \mathcal{R}, \mathcal{T})$ where:

- $\mathcal{V}$: Set of entity vertices representing concrete instances (e.g., "TechTalk2025", "Auditorium A")
- $\mathcal{R}$: Set of relations connecting entities (e.g., `held_at`, `scheduled_on`, `organized_by`)
- $\mathcal{T}$: Set of entity types defining the ontology

For calendar events, we focus on five primary entity types:

$$\mathcal{T} = \{\text{EVENT}, \text{DATE}, \text{TIME}, \text{VENUE}, \text{HOST}\}$$

Each entity $e \in \mathcal{E}_i$ can be represented as a tuple (v, t) where $v \in \mathcal{V}$ is the entity value and $t \in \mathcal{T}$ is its type.

3.3 Entity Type Definitions and Patterns

Building upon the calendar event knowledge graph ontology, we define comprehensive extraction patterns for each entity type. The
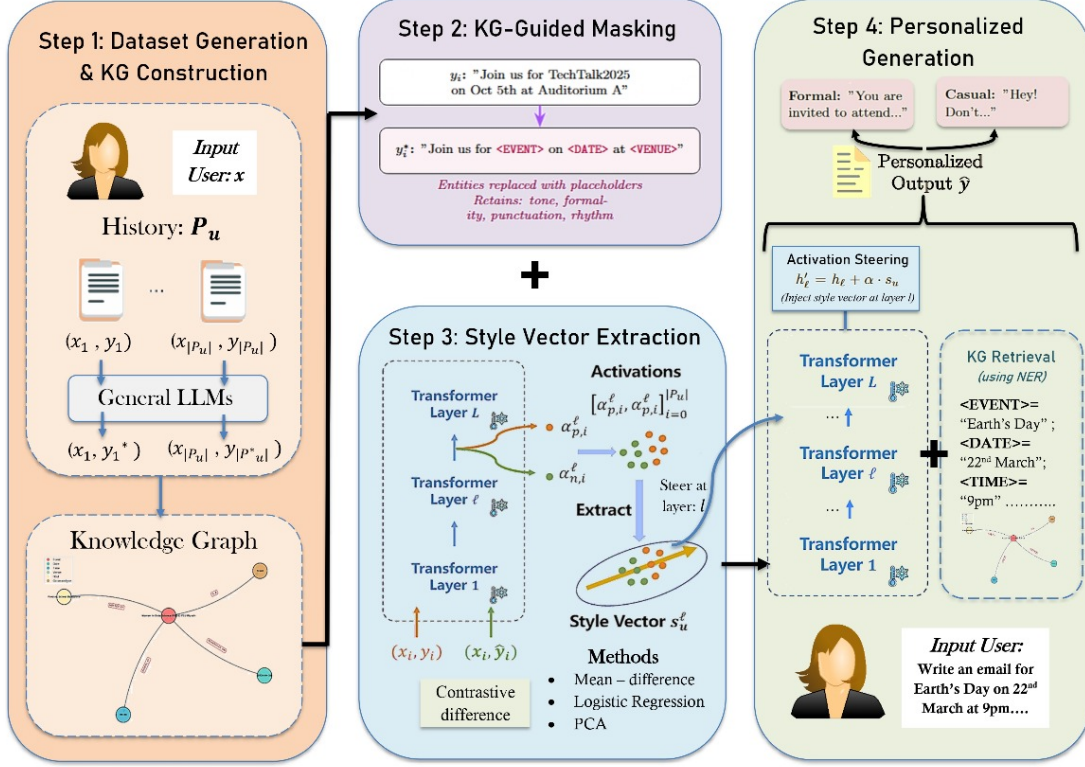
Figure 1: Complete pipeline of our KG-Guided activation steering

extraction takes advantage of bothenerates both structural properties of calendar communications and the linguistic patterns.

EVENT Entities (t = EVENT): Event entities represent the name, title, or identifier of the calendar activity. These typically appear in:

- Email subject lines: `Subject: Invitation: [EVENT NAME]`
- Main body references: "invites you to [EVENT NAME]"
- Quoted titles: "'[EVENT NAME]' will take place..."

Extraction Patterns:

```
P_EVENT = {(?:Subject|SUBJECT):\s*(?:Invitation:\s*)?
        ([\n]+?)(?:\n|$),
        (?:session|webinar|conference|event)\s+on\s+
        ["']?(["'\.!\n]+)["']?,
        invites you to\s+(?:a\s+)?["']?
        (["'\.!\n]+?)["']?,
        ["'](["'{10,100})["']    (quoted text)}
```

DATE Entities (t = DATE): Date entities capture temporal information about when the event occurs. Calendar communications use diverse date formats:

Extraction Patterns:

```
P_DATE = {\b(?:January|February|...|December)
        \s+\d{1,2},?\s+\d{4}\b,
        \b(?:Jan|Feb|Mar|...|Dec)\.?\s+
        \d{1,2},?\s+\d{4}\b,
        \b\d{1,2}[-/]\d{1,2}[-/]\d{2,4}\b,
        \b\d{4}[-/]\d{1,2}[-/]\d{1,2}\b,
        \b\d{1,2}(?:st|nd|rd|th)?
        \s+(?:of\s+)?(?:January|...|December),?
        \s+\d{4}\b}
```

TIME Entities (t = TIME): Time entities specify the precise moment when events begin, using various time notation conventions:

Extraction Patterns:

```
P_TIME = {\b\d{1,2}:\d{2}
        \s*(?:AM|PM|am|pm|a\.m\.|p\.m\.)\b,
        \b\d{1,2}\s*(?:AM|PM|am|pm)\b,
        at\s+\d{1,2}:\d{2}\b,
        \b\d{1,2}:\d{2}
        \s*(?:hours|hrs)?\b}
```

VENUE Entities (t = VENUE): Venue entities identify the physical or virtual location where events take place:

Extraction Patterns:

```
P_VENUE = {(?:will take place|held|located|venue|at|in)
        \s+(?:at|in|on)\s+([A-Z][a-zA-Z\s]+?),
        (?:in|at)\s+([A-Z][a-zA-Z]+
        (?:\s+[A-Z][a-zA-Z]+)?),
        (?:city|location|venue|place):
        \s*([^\n,\.]+),
        (?:will\s+be\s+held\s+in|venue\s+is)
        \s+([A-Z][a-zA-Z\s]+?)}
```

HOST Entities (t = HOST): Host entities identify the organizer, presenter, or responsible party for the event:

Extraction Patterns:

```
P_HOST = {(?:Sincerely|Regards|Best\s+regards),?
        \s*\n\s*([^\n]+?),
        ^([A-Z][a-zA-Z\s]+?)
        \s+(?:cordially\s+invites|invites\s+you),
        (?:organized\s+by|hosted\s+by|presented\s+by)
        \s+([^\n,\.]+),
        (?:From|FROM):\s*([^\n]+)}
```

3.4 Systematic Defactualization

Placeholder Substitution. Given extracted entity sets $\mathcal{E}(y_i)$ and $\mathcal{E}(\hat{y}_i)$, we perform systematic replacement to create defactualized versions. The replacement operation is defined as:

$$y_i^* = \text{Replace}(y_i, \mathcal{E}(y_i), \mathcal{P})$$

where $\mathcal{P} = \{$<EVENT>, <DATE>, <TIME>, <VENUE>, <HOST>$\}$ is the set of typed placeholders.

Comprehensive Example: End-to-End Transformation

Example 1: Enthusiastic Invitation.

Input Context x_i:

```
Event: Annual Innovation Summit 2025
Date: March 15, 2025
Time: 2:00 PM
Location: Grand Conference Hall
Organizer: Dr. Sarah Mitchell, Innovation Department
```

Factualized Styled Response y_i:

> "*Hey everyone! Super excited to invite you all to the* Innovation Summit 2025*! It's happening on* March 15th 2025 *at* 2 PM *in the* Grand Conference Hall*. This is going to be absolutely amazing – we've got incredible speakers lined up!*
> *Can't wait to see you all there. The* Innovation Summit 2025 *is definitely not to be missed! Cheers,*
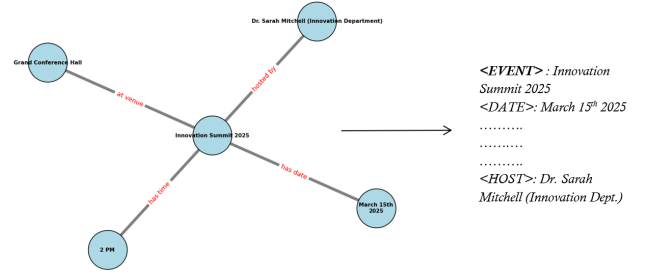> Dr. Sarah Mitchell
> Innovation Department

Figure 2: This figure illustrates entity extraction from a Knowledge Graph (KG), where nodes are event-related entities (Event, Date, Time, Venue, Host) and edges capture their semantic relationships for structured IR using NER.

Defactualized Styled Response y_i^*:

> "Hey everyone! Super excited to invite you all to the <EVENT>! It's happening on <DATE> at <TIME> in the <VENUE>. This is going to be absolutely amazing – we've got incredible speakers lined up!
> Can't wait to see you all there. The <EVENT> is definitely not to be missed! Cheers,
> <HOST>
> Innovation Department"

Stylistic Analysis: **Preserved Stylistic Features in y_i^*:**

(1) **Greeting style**: Informal "Hey everyone!" vs. formal "Dear Colleagues"
(2) **Emotional markers**: Emoji , "Super excited", "absolutely amazing"
(3) **Exclamation usage**: 4 exclamation marks vs. 0 in neutral
(4) **Personal engagement**: "Can't wait to see you all" vs. "We look forward"
(5) **Emphasis patterns**: "definitely not to be missed" vs. plain statement
(6) **Sentence structure**: Shorter, energetic sentences vs. longer, formal sentences
(7) **Closing style**: Casual "Cheers" vs. formal "Sincerely"
(8) **Repetition pattern**: Event name mentioned twice for emphasis

Removed Factual Content: All specific identifying information (event name, exact date, precise time, venue name, organizer name) has been systematically replaced, leaving only the stylistic skeleton.

3.5 Extracting the Style Vector

We identify the user-specific stylistic direction in the model's latent activation space through contrastive analysis of hidden states. Let $h_\ell(r) \in \mathbb{R}^d$ denote the hidden representation of the last token at layer ℓ when processing text r. For each paired example i, we compute the positive and negative activations as:

$$a_{p,i}^\ell = h_\ell(x_i \oplus y_i^*), \quad a_{n,i}^\ell = h_\ell(x_i \oplus \hat{y}_i^*), \tag{5}$$

where $\oplus$ denotes the concatenation of query/prompt and response (email) sequences.

These activations, respectively, encode stylistically enriched contexts ($a_{p,i}^{\ell}$) and style-neutral ($a_{n,i}^{\ell}$). The user-specific style vector s_u^{ℓ} is obtained by aggregating contrastive information in all paired examples:

$$s_u^{\ell} = f\big([a_{p,i}^{\ell}, a_{n,i}^{\ell}]_{i=1}^{|P_u|}\big), \tag{6}$$

where $f(\cdot)$ is a function that extracts a direction in the activation space pointing from style-neutral to stylistically rich samples.

We explore multiple formulations of $f(\cdot)$ as follows:

(1) Mean Difference. The simplest approach computes the average difference between positive and negative activations:

$$s_u^{\ell} = \frac{1}{|P_u|} \sum_{i=1}^{|P_u|} (a_{p,i}^{\ell} - a_{n,i}^{\ell}). \tag{7}$$

Here, s_u^{ℓ} represents the average displacement in the activation space that distinguishes defactualized messages of the user-style from the neutral ones.

(2) Logistic Regression. To find a discriminative direction that separates the two activation classes, we employ logistic regression. Let $X = [a_{p,1}^{\ell}; \ldots; a_{p,|P_u|}^{\ell}; a_{n,1}^{\ell}; \ldots; a_{n,|P_u|}^{\ell}]$ denote the concatenated activations and $y = [1, \ldots, 1, -1, \ldots, -1]$ their labels. The separating hyperplane is obtained by:

$$w = \arg \min_w \sum_i \log(1 + e^{-y_i X_i w}), \tag{8}$$

where w is the normal vector of the decision boundary. Moving in the direction of w increases the probability that the model will generate user-style activations. We normalize w to obtain the style vector:

$$s_u^{\ell} = \frac{w}{\|w\|_2}. \tag{9}$$

(3) Principal Component Analysis (PCA).. The PCA approach identifies the principal direction of variation between positive and negative activations. Let $\Delta_i = a_{p,i}^{\ell} - a_{n,i}^{\ell}$ represent the difference per-sample. We compute the direction s_u^{ℓ} that maximizes projected variance:

$$s_u^{\ell} = \arg \max_{\|v\|=1} \sum_{i=1}^{|P_u|} (\Delta_i^T v)^2. \tag{10}$$

This formulation ensures that (1) s_u^{ℓ} has unit norm, (2) captures the maximum stylistic variance, and (3) including $\{-\Delta_i\}$, the solution is symmetric with respect to label assignment. The resulting s_u^{ℓ} is computed as the principal eigenvector of the covariance matrix $\sum_i (\Delta_i \Delta_i^T + (-\Delta_i)(-\Delta_i)^T)$, efficiently solvable via Singular Value Decomposition (SVD).

Through these formulations, s_u^{ℓ} becomes a robust latent representation of user-specific tone (e.g., formality, warmth, conciseness) abstracted from factual event details.

3.6 Steering Personalized Generation and Re-hydration

After obtaining the style vector s_u^{ℓ}, we personalize generation by directly intervening in the model's hidden activations during the inference. Given a new input query x containaing subject line with factual placeholders (e.g., "Invitation for <EVENT> on <DATE>"),

the base LLM processes the input to generate a defactualized continuation. Let $h_{\ell}(x)_t$ denote the hidden representation of the token at position t in layer ℓ during generation. We modify these activations by injecting the scaled style vector:

$$h_{\ell}'(x)_t = h_{\ell}(x)_t + \alpha s_u^{\ell}, \tag{11}$$

where α is a hyperparameter which is a tunable scaling coefficient controlling stylistic intensity.

Following Rimsky *et al.* (2024), we apply the intervention to all tokens generated after the end of the input prompt, i.e., $t \geq |x|$. By adjusting α, we can finely control the degree of stylistic influence—lower α values yield subtle tone modulation, while higher values produce overtly styled expressions.

The final output is a defactualized, style-consistent email maintaining factual placeholders such as <EVENT> and <TIME>. In post-processing, these placeholders are rehydrated via Knowledge Graph lookups to restore actual entities, ensuring that factual correctness is preserved while the desired stylistic characteristics are retained. This process enables context-aware, stylistically personalized calendar communications without compromising factual integrity.

4 Novelty

We introduce a novel Knowledge Graph-guided personalization framework that, for the first time in this domain, integrates KG-guided masking with activation-space style disentanglement. This approach grounds user style learning in structured semantic relationships among event attributes, ensuring both interpretability and factual consistency during text generation.
Creation of an event-centric dataset designed for calendar-driven communication tasks such as formal and informal invitations, event announcements, and short reports. Each record includes structured fields such as Event, Date, Time, Venue, and Host, paired with human-like-written examples across multiple stylistic tones along with style-neutral outputs.

5 Results

We evaluate the effectiveness of our proposed **KG-Guided Disentanglement for Calendar Events** framework under two settings: (1) **Pure Activation Steering** — where the style vector is applied directly to factual texts without any Knowledge Graph (KG) involvement or masking, and (2) **KG-Guided Activation Steering** — where defactualization, semantic masking, and KG-based rehydration are integrated. All experiments employ the LLaMA-2-7B-hf model as the underlying language model.

5.1 Quantitative Evaluation

We evaluate both configurations using standard automatic metrics: ROUGE-L and METEOR. The results, reported as average F1 scores, are shown in Table 1.

Table 1: Performance comparison between Pure Activation Steering (without KG) and KG-Guided Activation Steering.

Configuration	ROUGE-L	METEOR
Pure Activation Steering (no KG)	0.2024	0.1961
KG-Guided Activation Steering	0.2261	0.2418

We observe that integrating KG-based guidance leads to consistent improvements across all reported metrics. The average ROUGE-L increases from 0.2024 to 0.2261, and METEOR improves from 0.1961 to 0.2418, indicating better semantic alignment and stylistic coherence. The improvements are particularly visible in METEOR, suggesting enhanced lexical-semantic consistency between generated and reference texts when factual and stylistic information are disentangled.

5.2 Qualitative Analysis

Manual inspection of generated outputs reveals a divergence between metric-based performance and perceptual quality. Although KG integration yields higher metric scores, qualitative examination exposes several notable artifacts and hallucinations—especially in the KG-guided activation steering :

- In the presence of KG guidance, the model frequently **hallucinates spurious placeholders** such as `<SENDER>`, `<RECIPIENTS>`, or `<SIGNATURE>`, despite these never appearing in the data schema.
- The model also occasionally interprets placeholder brackets `< >` as HTML or XML syntax, producing elements like `<div>`, `<p>`, or `<body>`—indicative of interference from web-style markup seen during the model's pretraining.
- Without KG-guided masking, such random hallucinations are significantly reduced; however, stylistic richness sometimes decreases, leading to more templated, formalized outputs.

These findings suggest that while activation steering alone captures high-level stylistic signals, it lacks grounding and tends to misinterpret non-linguistic symbols, resulting in text artifacts that undermine factual reliability.

5.3 Interpretation of Quantitative Trends

The contrast between improved evaluation metrics and imperfect qualitative outputs highlights an important dynamic in style disentanglement systems.

The KG-guided configuration stabilizes factual structure by explicitly masking and rehydrating entities, thereby constraining factual variability. This enables the activation steering mechanism to focus primarily on stylistic modulation rather than inadvertently altering semantic elements. As a result, automatic metrics such as ROUGE-L and METEOR—which reward structural and semantic consistency—show measurable improvements.

However, the metric behavior also reveals an unusual phenomenon: **ROUGE-2 (0.1361) is lower than ROUGE-L (0.2261)**, which is counterintuitive since bigram overlap (ROUGE-2) typically correlates closely with LCS-based alignment (ROUGE-L). We attribute this to three interacting factors:

(1) **Placeholder token disruption:** The insertion of entity placeholders (`<EVENT>`, `<DATE>`, `<VENUE>`) breaks bigram continuity even when overall sentence structure is preserved.
(2) **Pseudo-tag hallucination:** Occasional inclusion of non-existent markup-like tokens introduces rare bigrams, reducing the overlap count.
(3) **Stylistic paraphrasing:** KG-guided steering encourages tonal variation and rewording (e.g., "Join us for" vs. "You're

invited to"), which maintains sequence-level alignment (ROUGE-L) but weakens exact bigram matches (ROUGE-2).

Consequently, ROUGE-L emerges as a more stable indicator of structural fidelity in this defactualized text generation setup, whereas ROUGE-2 becomes sensitive to placeholder boundary artifacts.

Summary of Findings

Overall, the results demonstrate that:

- **Pure activation steering** (without KG) successfully injects stylistic features but introduces unrelated factual information.
- **KG-guided activation steering** enhances factual grounding and improves automatic metric performance but increases noise from random hallucinations.
- The observed anomaly where ROUGE-2 < ROUGE-L arises due to placeholder-induced bigram disruptions and paraphrastic rewording.

Hence, the integration of KG-based semantic masking provides a stabilizing effect, enabling the model to modulate stylistic attributes without compromising factual alignment—though further refinement in placeholder token handling and decoding constraints remains necessary for consistent qualitative fidelity.

6 Ablation Study

Our experimental workflow followed a systematic progression to optimize email generation with style steering. Initially, we implemented basic steering with evaluation metrics to establish baseline performance.

Subsequently, we integrated knowledge graph (KG) components with the steering mechanism and evaluated the combined system. The latter part focused on refining the approach using PCA-based steering with a fixed multiplier of 3.0, which demonstrated promising results.

Finally, we conducted comprehensive hyperparameter tuning to identify the optimal steering multiplier by testing values ranging from 0.5 to 3.0.

To better understand the contribution of individual components and hyperparameters in our system, we conducted a detailed ablation study focusing on three core dimensions: (1) the impact of steering strength (multiplier value), (2) the comparative performance of style vector extraction functions (Mean, Logistic Regression, and PCA), and (3) the integration of Knowledge Graph (KG) grounding with activation steering.

6.1 Experimental Workflow

Our experimental workflow followed a systematic, stepwise progression aimed at optimizing stylistically controlled email generation. Initially, we implemented **basic activation steering** on factual email data to establish a baseline for style modulation without external grounding. ROUGE-L and METEOR metrics were employed for initial quantitative assessment across a small validation subset.

In the next phase, we integrated the **Knowledge Graph (KG)** component to achieve factual disentanglement and semantic stability. By masking factual entities (e.g., <EVENT>, <DATE>, <VENUE>), the model was guided to focus steering perturbations on stylistic rather than factual activations. This phase established the combined KG-steering framework, which formed the foundation for subsequent optimization experiments.

Finally, we refined the approach by employing the **PCA-based style vector formulation**, followed by hyperparameter tuning on the steering multiplier α. The PCA variant was selected for deeper exploration since preliminary evaluations suggested it provided smoother tonal control and reduced syntactic noise compared to Mean or Logistic Regression variants.

6.2 Hyperparameter Tuning: Steering Multiplier Analysis

We systematically varied the steering multiplier α in the range $[0.5, 3.0]$ to observe its effect on generation stability and stylistic expression. This parameter controls the scaling of the style vector s_u^ℓ added to hidden activations during generation, directly influencing stylistic intensity.

The model was evaluated on 20 test examples per setting using ROUGE-L and METEOR metrics. As shown in Table 2 and figure 3, the results reveal a non-linear trend: performance initially degrades with moderate values of α and then recovers at higher magnitudes, peaking at $\alpha = 3.0$.
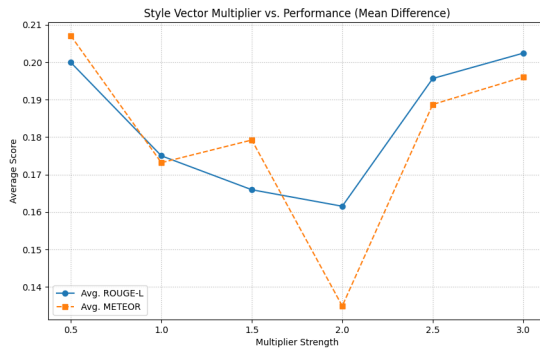

Figure 3: Performance variation across different hyperparameter settings.

Table 2: Hyperparameter Tuning Results Across Different Multiplier Values

Multiplier (α)	Avg. ROUGE-L	Avg. METEOR
0.5	0.1999	0.2071
1.0	0.1750	0.1732
1.5	0.1659	0.1792
2.0	0.1615	0.1349
2.5	0.1957	0.1887
3.0	**0.2024**	**0.1961**

The analysis suggests that lower steering magnitudes ($\alpha \leq 1.0$) result in insufficient activation displacement, producing near-neutral text with minimal stylistic deviation. Conversely, excessively high multipliers ($\alpha > 3.0$) induce instability and semantic drift, often leading to misplaced placeholders or repetitive phrasing. The multiplier of $\alpha = 3.0$ achieves the best balance between stylistic richness and factual coherence, yielding the highest combined performance across both metrics. This observation aligns with prior literature, which reports that moderate-to-high scaling amplifies stylistic directionality without triggering excessive hallucination.

6.3 Comparison of Style Vector Formulations

A parallel set of experiments was conducted to compare the three proposed formulations for the style vector extraction function $f(\cdot)$: Mean-Difference, Logistic Regression, and Principal Component Analysis (PCA). Each method was evaluated using the same model configuration, ensuring controlled experimental conditions. Results can be shown in the table 3

Table 3: Comparison of Style Vector Extraction Methods (Average F1-Score)

Method	Avg. ROUGE-L	Avg. METEOR
Mean	0.1635	0.1608
Logistic Regression	0.1523	0.1609
PCA	**0.1795**	**0.1847**

The PCA-based approach outperformed both Mean and Logistic Regression methods across ROUGE-L and METEOR. While the differences may appear modest numerically, qualitative inspection revealed that PCA-based steering consistently produced smoother, more coherent, and stylistically balanced email texts. Specifically:

- **Mean-Difference:** Generated text exhibited stylistic drift toward overly formal or overly generic phrasing, likely due to averaging across heterogeneous activation patterns.
- **Logistic Regression:** Outputs were often abrupt or inconsistent in tone, possibly because the discriminative boundary learned by logistic regression overfits to small-sample activation differences.
- **PCA:** Captured the dominant stylistic variance across the dataset, resulting in subtler, more natural tonal modulation without loss of factual coherence.

6.4 Summary of Ablation Findings

The ablation study offers three key insights:

(1) **Steering Magnitude Sensitivity:** The steering multiplier α plays a critical role in balancing stylistic enhancement and factual stability. Optimal performance occurs at $\alpha = 3.0$, beyond which the model begins to hallucinate or distort placeholder semantics.

(2) **PCA-based Advantage:** Among the three style extraction methods, PCA produces stylistic steering that is both quantitatively superior and qualitatively more controlled, likely due to its ability to capture orthogonal variance components while ignoring noise.

(3) **Importance of Grounding:** Experiments reaffirm that KG integration provides a stabilizing effect by isolating stylistic perturbations from factual content, enabling higher metric performance and reducing token hallucinations.

Together, these findings confirm that even under limited data conditions, activation steering can be fine-tuned effectively through targeted parameter optimization and principled choice of style vector formulation. They also highlight that structured, domain-specific datasets amplify the interpretability and controllability of activation-space manipulations—offering a reproducible pathway for future low-resource personalization tasks.

7 Discussion

7.1 Motivation for Domain-Specific Focus

The original activation-steering framework was demonstrated on general-purpose text corpora where factual content and stylistic tone coexist in complex, entangled forms. In our case, initial attempts to apply factual disentanglement on unrestricted text proved infeasible, primarily because the factual boundaries in general communication are highly diffuse and context-dependent. To overcome this limitation, we restricted our scope to a **domain with clearly segmentable factual attributes**—namely, *calendar-based event communications* such as invitation emails. This domain offers a consistent factual schema that can be expressed through a limited set of entity types:

$$\text{Event, Host, Venue, and Date.}$$

This explicit structure enables us to systematically defactualize texts by masking these entities and rehydrating them post-generation using the constructed Knowledge Graph (KG). Hence, the calendar-event domain provided a controlled yet semantically meaningful environment to evaluate stylistic disentanglement and steering effectiveness.

7.2 Data Constraints and Construction Challenges

One of the most significant challenges encountered was the lack of an existing dataset tailored for our task. There are currently no publicly available corpora of human-written invitation or announcement emails labeled by style or tone. Consequently, we had to construct a **small synthetic dataset** consisting of only 40 email pairs—each comprising:

- A **style-agnostic** (neutral) version generated by a general LLM M_g, and
- A **styled** version exhibiting a specific tone and rhythm (e.g., formal, friendly, or enthusiastic).

These styled samples were generated under human supervision and manually corrected for grammatical, syntactic, and tonal consistency. After KG-guided defactualization, the resulting dataset pairs $(y_i^*, \hat{y}_i^*)$ were used to compute the user style vector through the activation-steering process.

During the integrated KG steering, one main problem stood out. It involved getting factual entities right from the Knowledge Graph we built. We had to map them accurately. In quite a few instances, the process for extracting and linking entities ended up with mappings that were not complete. Sometimes they were just noisy. This led to mismatches in placeholders. It also created garbage values right in the output text. Those problems mostly came from token boundaries that did not match up well. There were ambiguous references to entities in the original emails too. All that made it hard to line up the KG nodes properly with the text parts they should connect to. As a result, the grounding system did not always separate out the factual tokens in a clean way. That reduced how stable the style steering was overall. It even brought in some semantic distortions to a handful of the samples we generated.

Despite the limited data size, the evaluation metrics obtained—particularly the ROUGE-L (0.2261) and METEOR (0.2418) scores under KG-guided conditions—were comparable to those reported in the original activation steering paper, which used substantially larger and more diverse datasets. This result suggests that even small, domain-constrained datasets can produce meaningful style directions in the activation space, provided the factual structure is explicitly defined and consistent.

7.3 Observations on Style Vector Function Behavior

In the original activation steering framework, **mean-difference** was reported as the most stable and generally best-performing function for extracting the style vector s_u^ℓ. However, in our ablation experiments, conducted within the structured and syntactically uniform domain of invitation emails, the **Principal Component Analysis (PCA)** formulation slightly outperformed both mean and logistic regression approaches.

This divergence can be attributed to several domain-specific properties:

(1) **Low lexical variance:** Since all email samples adhere to similar sentence templates, the mean-difference vector captures only minor stylistic shifts, whereas PCA amplifies subtle but consistent stylistic patterns embedded across the dataset.

(2) **Defactualization regularity:** The presence of identical placeholder tokens in every sample (e.g., <EVENT>, <DATE>) leads to structured activation patterns that benefit from PCA's ability to extract dominant variance directions while filtering out noise.

(3) **Steering interpretability:** Manual qualitative inspection of generated emails confirmed that PCA-based vectors produced stylistic outputs with smoother tone transitions and less abrupt phrasing changes compared to mean-difference steering.

Thus, while mean-difference may remain optimal for broader, heterogeneous text distributions, our findings indicate that PCA-based style extraction better suits low-variance, syntactically regular domains where stylistic deviation is subtle but systematic.

7.4 On the Adequacy of Evaluation Metrics

A recurring theme during experimentation was the mismatch between automatic evaluation metrics and human judgment of stylistic quality. Metrics such as ROUGE and METEOR, while effective for content fidelity and paraphrase alignment, are largely **insensitive to stylistic variation**. They measure lexical overlap and sequence alignment, but cannot capture stylistic phenomena such

as tone, rhythm, formality, or emotional resonance. For instance, two stylistically distinct sentences—

> *"You are invited to attend the annual research symposium."*
> *"We'd love to have you join us for this year's research meet!"*

—may yield low ROUGE overlap despite exhibiting clear stylistic correlation in tone and communicative intent.

Future work should thus move beyond token-based similarity metrics toward style-aware evaluation. We propose incorporating a **style classifier** trained on user-specific tone or sentiment profiles to quantify stylistic resemblance more accurately. Such a classifier could complement existing metrics by evaluating latent features that correspond to tone consistency, rhythm, and formality, rather than surface-level word overlap.

7.5 Limitations and Future Work

Due to limited computational and data resources, we were unable to conduct large-scale evaluations or systematic hyperparameter sweeps (e.g., varying steering strength α, layer depth ℓ, or number of masked entities). A larger dataset of human-authored emails with diverse stylistic exemplars would likely reduce hallucination artifacts and yield more stable activation directions. Additionally, future studies could explore:

- **Token-level steering control:** Localized activation interventions that adjust specific stylistic features (e.g., formality or conciseness) without affecting sentence structure.
- **Layer-wise disentanglement:** Investigating whether style representations emerge predominantly in certain transformer layers, improving steering precision.
- **Multi-domain generalization:** Extending the approach to other structured text genres (e.g., business memos, academic abstracts) to test robustness of the style vector extraction process.

7.6 Summary

Overall, the constrained but well-structured calendar-event domain provided an effective sandbox for exploring activation-based stylistic control. Even with only 40 curated examples, the model learned meaningful stylistic directions and exhibited measurable improvements in factual grounding and alignment. While the PCA-based vector performed best for this domain, the broader insight lies in demonstrating that activation-space steering remains effective even under resource-constrained, domain-specific conditions. Future iterations incorporating larger datasets, domain-adaptive style classifiers, and improved placeholder tokenization could significantly enhance the reliability and expressivity of the generated text.

8 References and Citations

References

[1] J. Zhang, Y. Liu, W. Wang, Q. Liu, S. Wu, L. Wang, and T.-S. Chua, *Personalized Text Generation with Contrastive Activation Steering.* arXiv preprint, arXiv:2503.05213, 2025.

[2] D. Prahlad, C. Lee, D. Kim, and H. Kim, *Personalizing Large Language Models using Retrieval Augmented Generation and Knowledge Graph.* In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*, pages 1259–1263, ACM Press, 2025. DOI: 10.1145/3701716.3715473.

[3] J. Zhang, B. Wei, S. Qi, H. Zhu, J. Liu, and Q. Lin, *GKG-LLM: A Unified Framework for Generalized Knowledge Graph Construction.* arXiv preprint, arXiv:2503.11227, 2025.

[4] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner, *Steering Llama 2 via Contrastive Activation Addition.* arXiv preprint, arXiv:2312.06681, 2024.

[5] H. Saadany and C. Orăsan, *BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text.* In *Proceedings of the Translation and Interpreting Technology Online Conference (TRITON 2021)*, pages 48–56, 2021. Available at: https://aclanthology.org/2021.triton-1.6.pdf.

[6] Chin-Yew Lin, Information Sciences Institute University of Southern California *ROUGE: A Package for Automatic Evaluation of Summaries.* In *In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, pages 2-10, 2004. Available at: https://aclanthology.org/W04-1013.pdf.