

universität freiburg

# Content Style Transfer Using RAG (Retrieval-Augmented Generation)

Master Thesis by Aaryan Bhandari

23 February 2026



# The current limitation in LLMs in personal communication

- LLMs are increasingly being used to recreate personal forms of communication, primarily, media channels like emails.
- However, there is a huge problem with using LLMs to generate emails.
- LLMs have NO context of the Author's **Text Style** and default to using the typical formal and robotic Style.

# The current limitation in LLMs in personal communication

- LLMs are increasingly being used to recreate personal forms of communication, primarily, media channels like emails.
- However, there is a huge problem with using LLMs to generate emails.
- LLMs have NO context of the Author's **Text Style** and default to using the typical formal and robotic Style.

*Let's walk through an example...*

# Lack of style transfer in LLMs

Person (Sam) receives an email and wants to reply to the email

# Lack of style transfer in LLMs

Person (Sam) receives an email and wants to reply to the email



Person provides the LLM with the prompt/context -  
*“Ask Vince to review legal documents and requests completion by the end of the day.”*

# Lack of style transfer in LLMs

Person (Sam) receives an email and wants to reply to the email



Person provides the LLM with the prompt/context -  
*“Ask Vince to review legal documents and requests completion by the end of the day.”*



*Dear Vince,*

*I hope you are well. Could you please review the documents I sent?  
I need them by today.*

*Best regards,*  
*Sam*

# Lack of style transfer in LLMs

Person (Sam) receives an email and wants to reply to the email

Person provides the LLM with the prompt/context -  
*“Ask Vince to review legal documents and requests completion by the end of the day.”*

*Dear Vince,*  
  
*I hope you are well. Could you please review the documents I sent?  
I need them by today.*  
  
*Best regards,*  
*Sam*

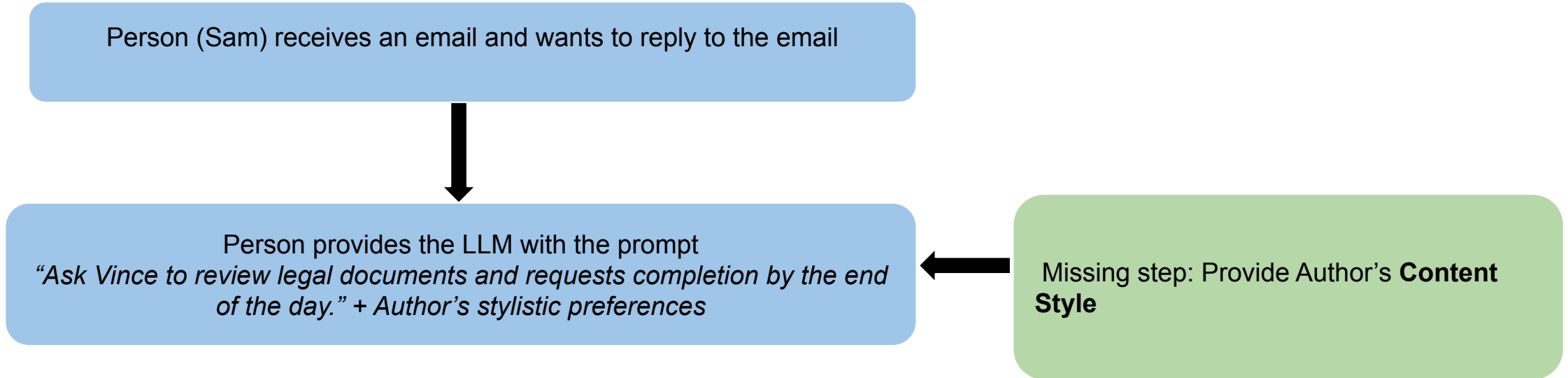
**Preserves the meaning** but  
**lacks the Author’s style** and  
defaults to the robotic LLM  
Style

# An Ideal Situation....

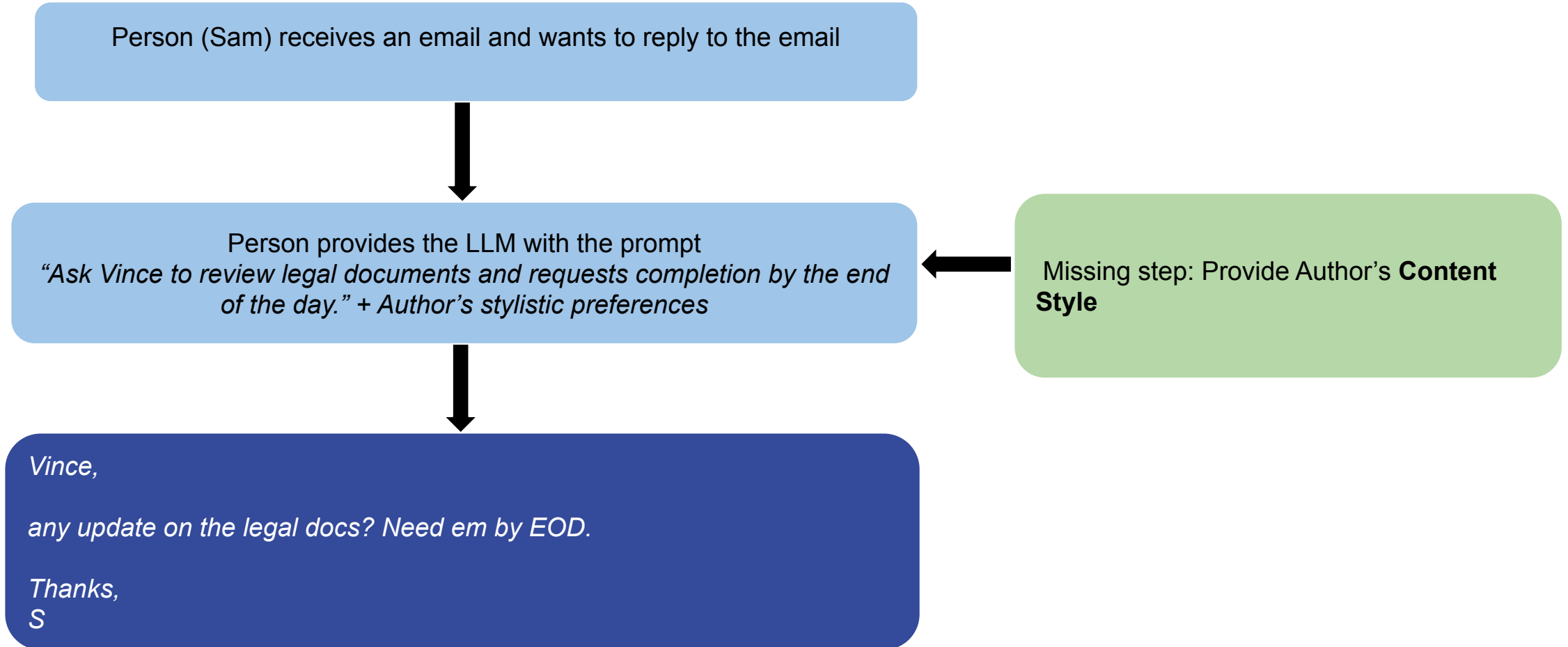
An ideal email creation tool should be able to -

- Preserve the **semantic meaning** the author provides in the context.
- Transfer the **general style** of the author for a particular recipient, in the generated email.
- Should sound **Natural** and **Coherent**.

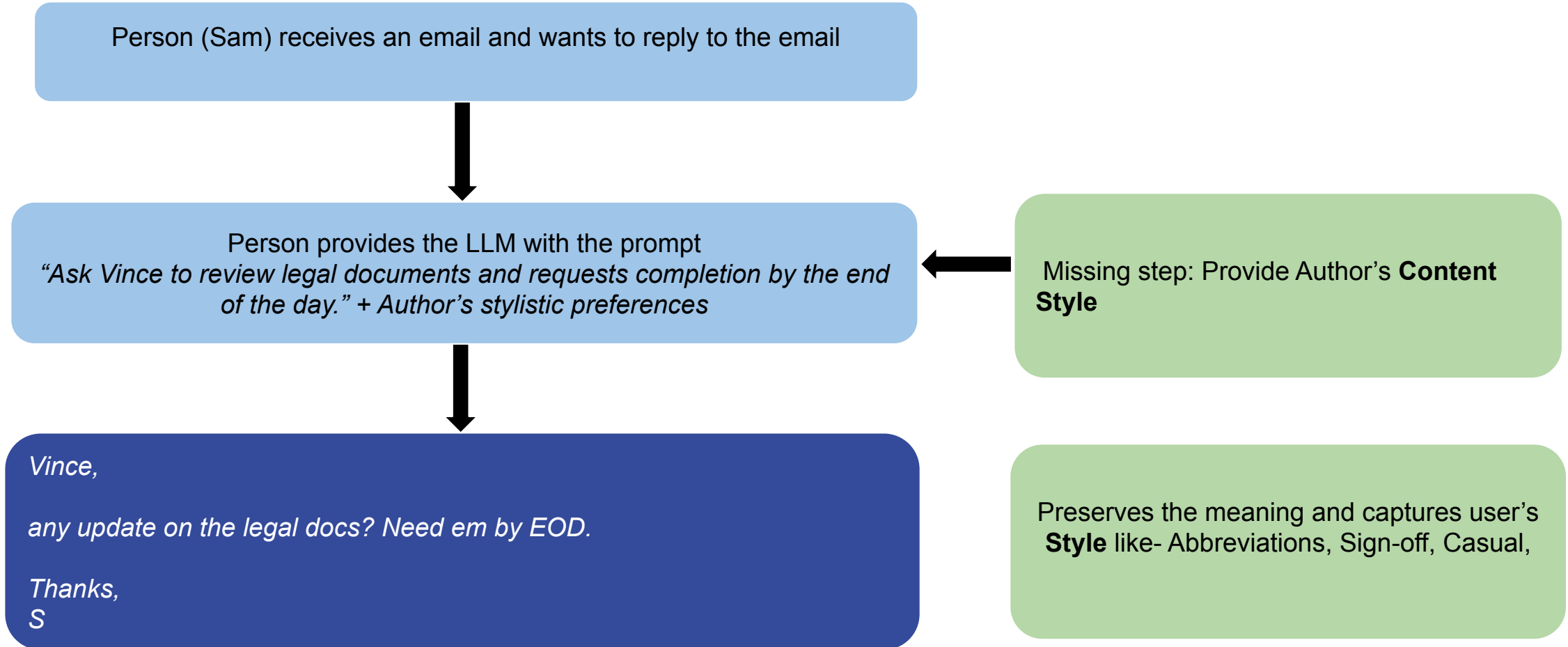
# The missing Step



# The missing Step



# The missing Step





---

*Thoughts?*



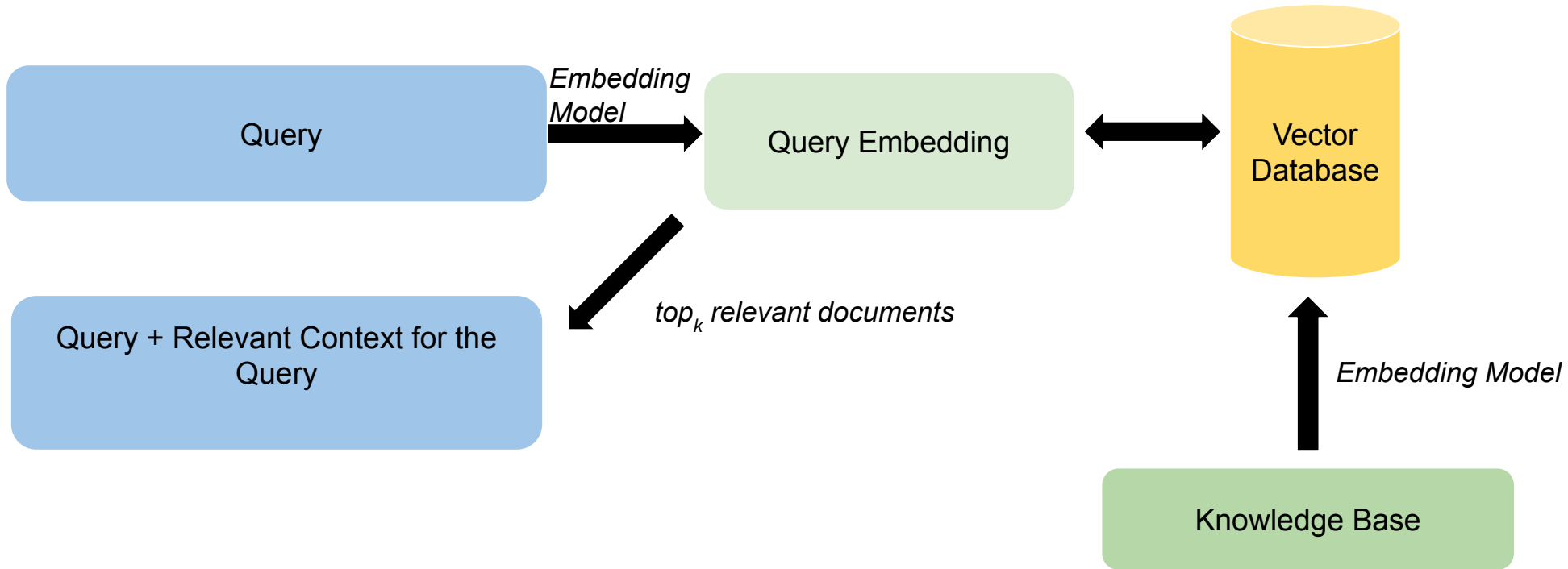
---

# Current Style Transfer Solutions -

Even though there exist many Style Transfer techniques, an ideal pipeline for Email generation with LLMs should -

- NOT require fine tuning.
- be able to transfer the **overall style of the Author**, instead of style specific features like - *Formality, Emotion* etc.
- Should be **generalizable** for different authors.

# Typical RAG pipeline



# Our novel RAG Approach

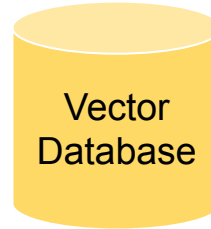
**Prompt:** Context of the email (Author  $A$  to recipient  $R$ )

# Our novel RAG Approach

**Prompt:** Context of the email (Author  $A$  to recipient  $R$ )

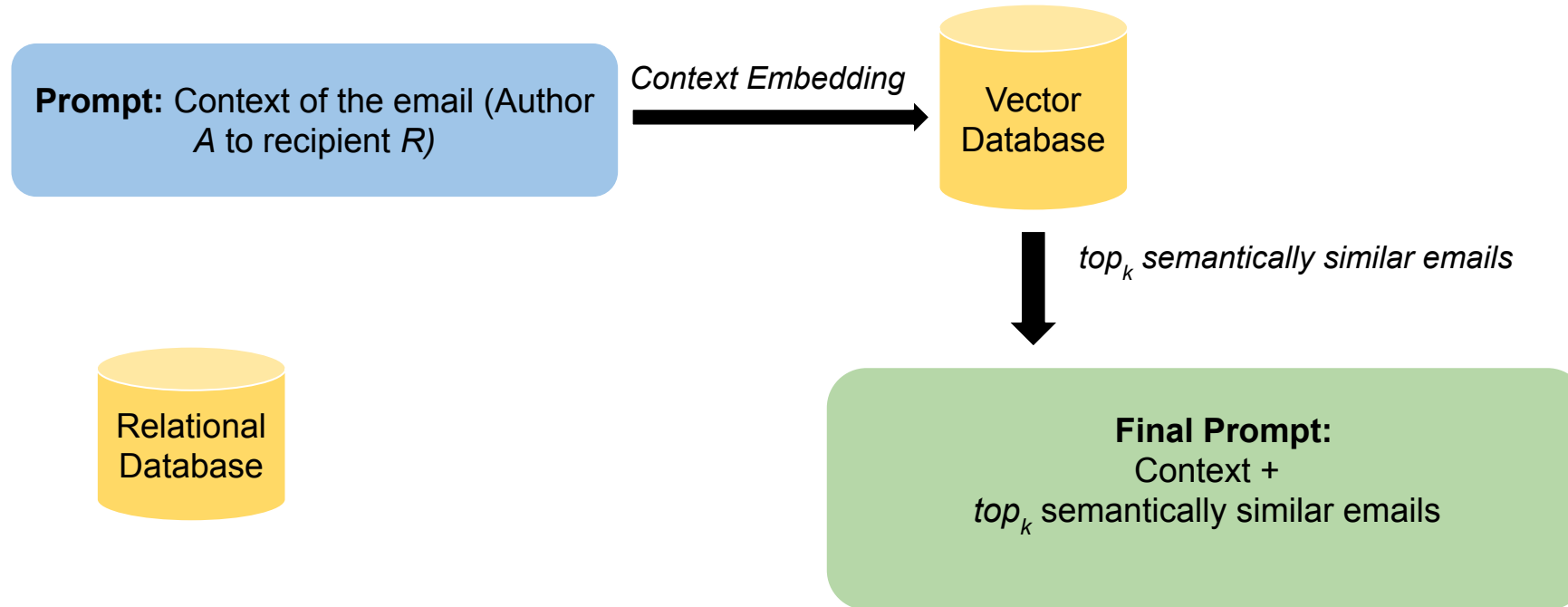


*Pre-created Relational Database of Author  $A$ 's historical emails*

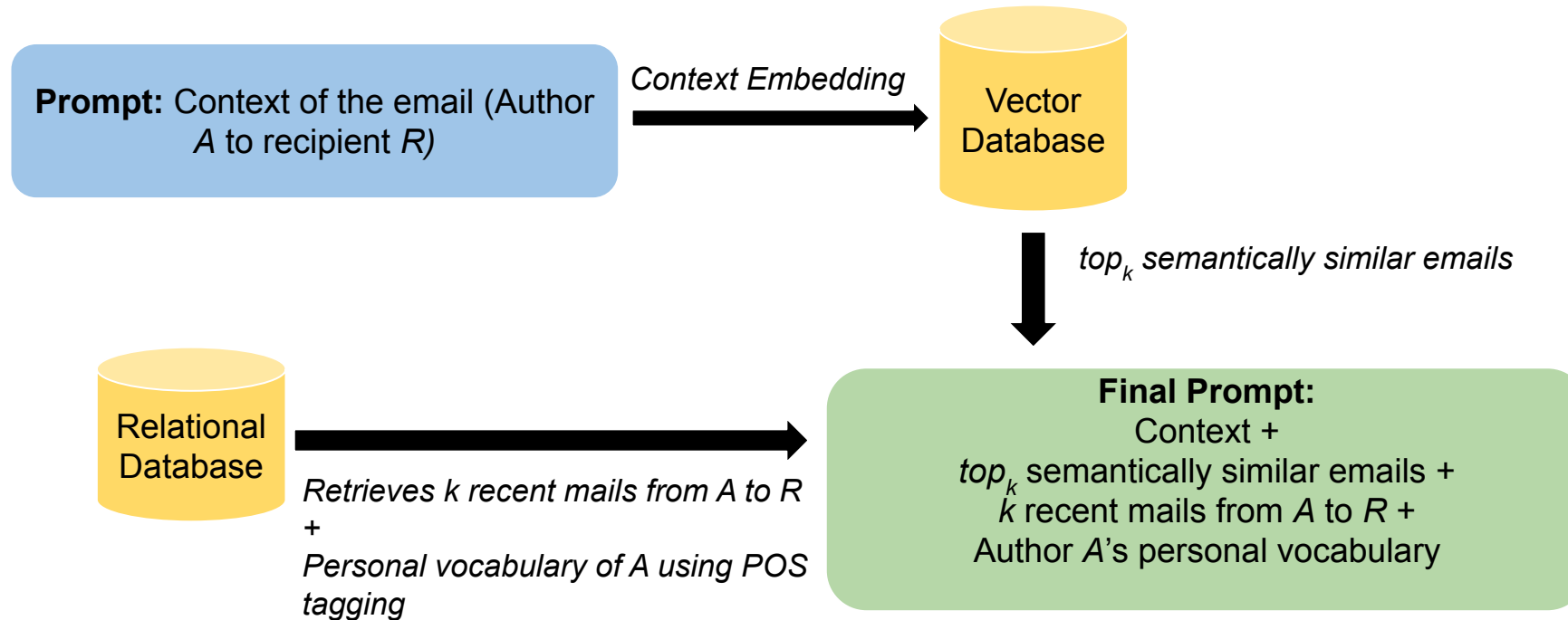


*Pre-created Vector Database of Author's  $A$  historical emails*

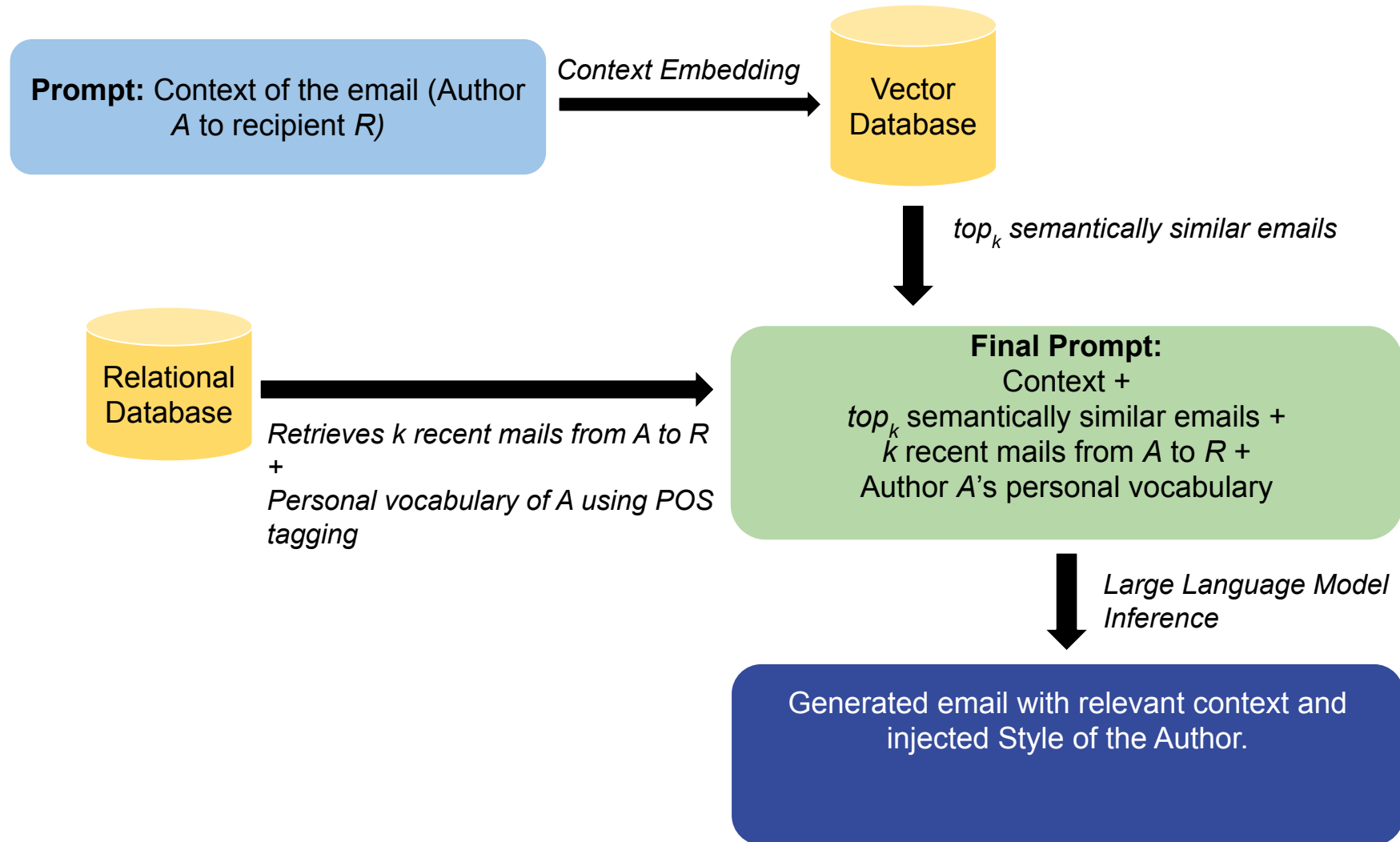
# Our novel RAG Approach



# Our novel RAG Approach



# Our novel RAG Approach



# Intuition

- **$top_k$  semantically similar emails** provide examples for the Author's writing style for similar contexts.

# Intuition

- **$top_k$  semantically similar emails** provide examples for the user's writing style for similar contexts.
- Recent emails capture the on-going dynamics between the Author and the Recipient **as style is not static and might change with time.**

# Intuition

- **$top_k$  semantically similar emails** provide examples for the user's writing style for similar contexts.
- Recent emails capture the on-going dynamics between the Author and the Recipient **as style is not static and might change with time.**
- **Parts-Of-Speech (POS)** tagging is used to get the vocabulary of the Author's word choice.
  - A frequency list of Adverbs, Adjectives and Verbs is used to create a personal dictionary and subconscious word usage
  - Example “glad” vs “happy” , “great” vs “good”, “fast” vs “quick” define a person's general word usage.

# Intuition

- **$top_k$  semantically similar emails** provide examples for the user's writing style for similar contexts.
- Recent emails capture the on-going dynamics between the Author and the Recipient **as style is not static and might change with time.**
- **Parts-Of-Speech (POS)** tagging is used to get the vocabulary of the Author's word choice.
  - A frequency list of Adverbs, Adjectives and Verbs is used to create a personal dictionary and subconscious word usage
  - Example "glad" vs "happy", "great" vs "good", "fast" vs "quick" define a person's general word usage.

The objective is to use **In Context Learning** to induce **Style Transfer**



---

*Thoughts?*



---

# Evaluation

- **Evaluating Style Transfer** remains to be a challenge.

# Evaluation

- **Evaluating Style Transfer** remains to be a challenge.
- There exists NO **Golden Metric** or a **Standard Benchmark** to evaluate successful style transfer.

# Evaluation

- **Evaluating Style Transfer** remains to be a challenge.
- There exists NO **Golden Metric** or a **Standard Benchmark** to evaluate successful style transfer.
- Style is multifaceted
  - Sentence Structure - Sentence Length
  - Vocabulary Choice - simple or complex
  - Tone - Humorous, Urgent or warm
  - Formality
  - Punctuation patterns
  - Paragraph Structure

*And many many more.....*

# Evaluation

- **Evaluating Style Transfer** remains to be a challenge.
- There exists NO **Golden Metric** or a **Standard Benchmark** to evaluate successful style transfer.
- Style is multifaceted
  - Sentence Structure - Sentence Length
  - Vocabulary Choice - simple or complex
  - Tone - Humorous, Urgent or warm
  - Formality
  - Punctuation patterns
  - Paragraph Structure

*And many many more.....*

It is not wise to measure Style Transfer on each feature alone.

# How do we evaluate?

Two methods of evaluating - **Automated Metrics** and **Human Evaluation**

- Automated Metrics on the Enron dataset using LLM-as-a-Judge
- User Study for Human evaluation
- Both use the **same** metrics

# Evaluation Metrics

Metric	Scale	Description / Rubric
<b>Style Transfer</b>	{1, 2, 3, 4, 5}	1 = Completely different styles 2 = Mostly different styles with minor similarities 3 = Moderately similar styles with notable differences 4 = Very similar styles with minor differences 5 = Completely identical styles

# Evaluation Metrics

Metric	Scale	Description / Rubric
<b>Style Transfer</b>	{1, 2, 3, 4, 5}	1 = Completely different styles 2 = Mostly different styles with minor similarities 3 = Moderately similar styles with notable differences 4 = Very similar styles with minor differences 5 = Completely identical styles
<b>Content Preservation</b>	{0, 1}	0 = Content/meaning NOT preserved 1 = Content IS preserved

# Evaluation Metrics

Metric	Scale	Description / Rubric
<b>Style Transfer</b>	{1, 2, 3, 4, 5}	1 = Completely different styles 2 = Mostly different styles with minor similarities 3 = Moderately similar styles with notable differences 4 = Very similar styles with minor differences 5 = Completely identical styles
<b>Content Preservation</b>	{0, 1}	0 = Content/meaning NOT preserved 1 = Content IS preserved
<b>Naturalness</b>	{0, 1}	0 = Unnatural or incoherent output 1 = Natural and coherent output

# Evaluation: Enron Dataset

- Enron dataset contains 500,000 emails from 150 different employees
- Has various stylistic differences between different authors, namely *Sentence Structure*, *Vocabulary*, *Tone*, *Formality* and so on.

These properties make it ideal for testing out framework.

# Evaluation: Enron Dataset

Steve,

Friday, September 8, 11:30 is fine with me. I can cancel the other meeting

Vince

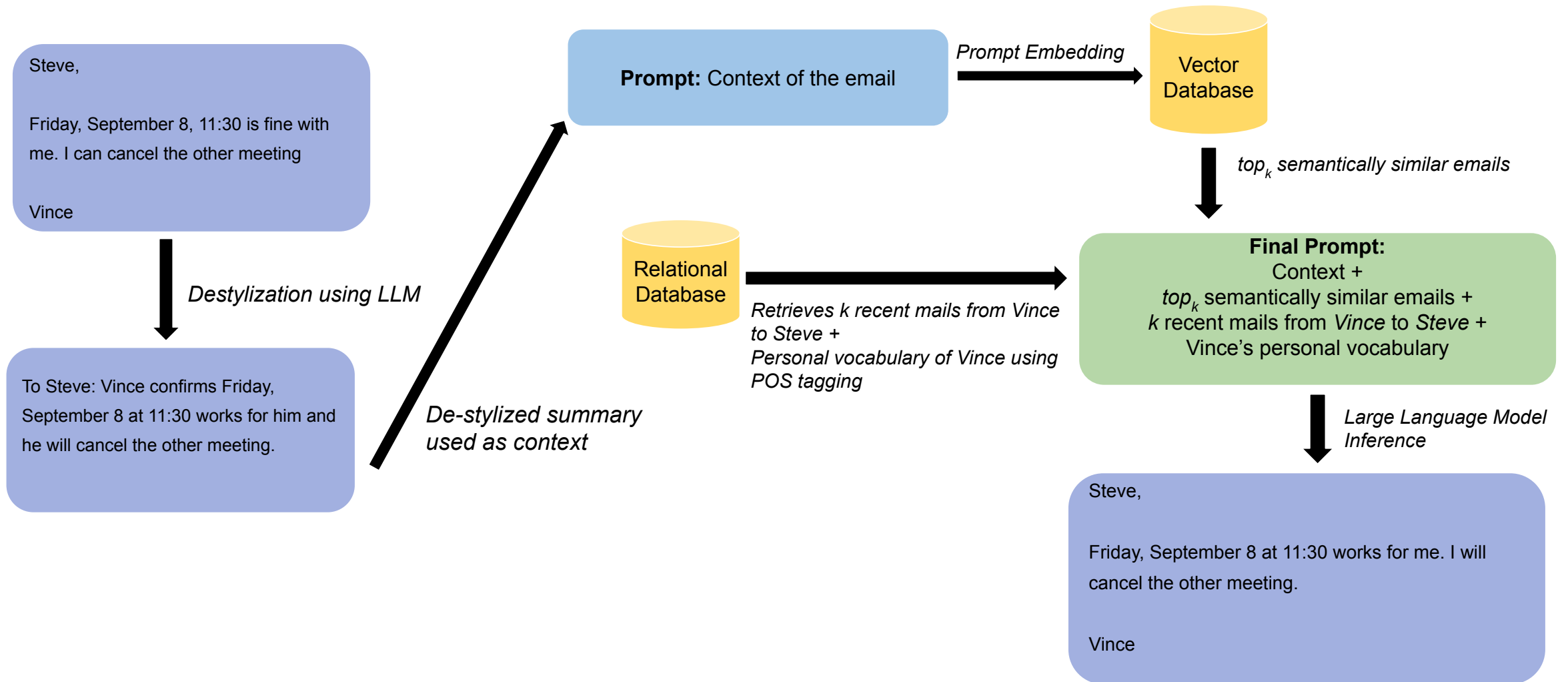


*Destylization using LLM*

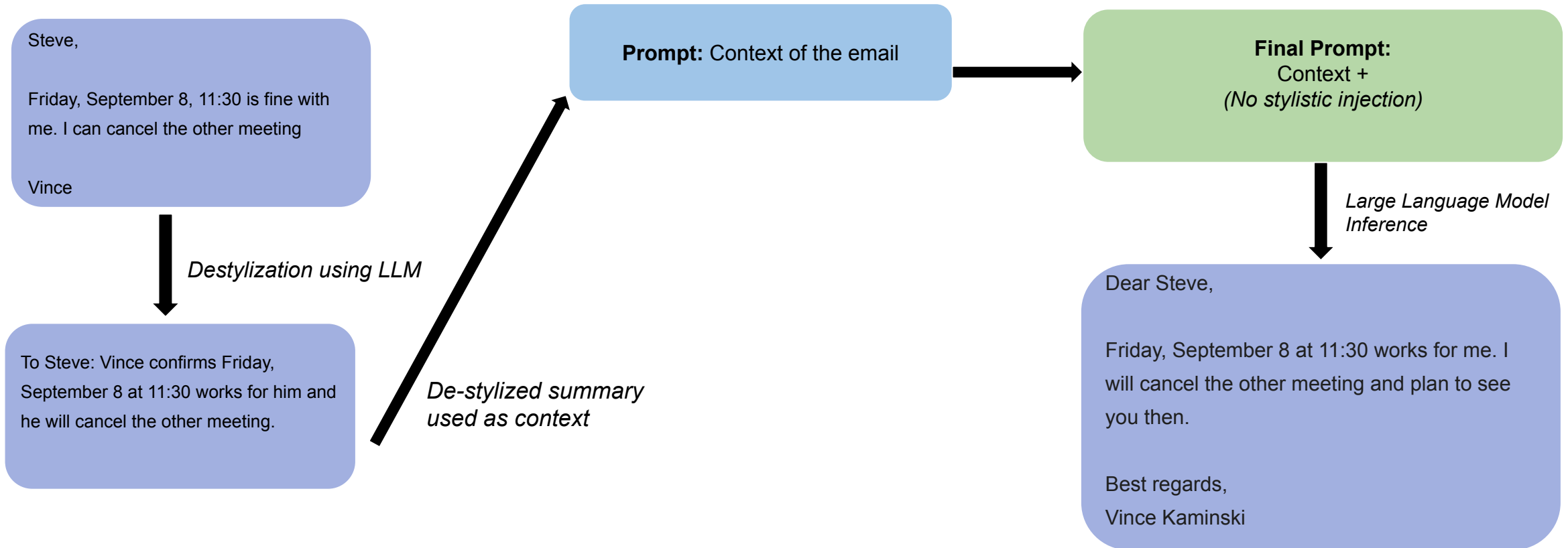
To Steve: Vince confirms Friday, September 8 at 11:30 works for him and he will cancel the other meeting.

- An email is sampled
- A *de-stylized* summary is created using an LLM

# Evaluation Approach: RAG Recreation



# Evaluation Approach: Vanilla ( w/o-RAG)



# LLM-as-a-judge comparison

## Original email

Steve,

Friday, September 8, 11:30 is fine with me. I can cancel the other meeting

Vince

Steve,

Friday, September 8 at 11:30 works for me. I will cancel the other meeting.

Vince

RAG output

Dear Steve,

Friday, September 8 at 11:30 works for me. I will cancel the other meeting and plan to see you then.

Best regards,  
Vince Kaminski

w/o RAG

# Evaluation: Enron Dataset (Large Parameter Models)

Configuration	Style Transfer (1-5)
Vanilla (Context-Only)	1.90
RAG ( $top_k = 1$ )	3.07
RAG ( $top_k = 5$ )	<b>3.41</b>
RAG ( $top_k = 10$ )	3.33

## Setup

- 1,000 sampled emails
- Recreator model: `gpt-5-mini`
- LLM-as-a-judge: `gpt-5-mini`

# Evaluation: Enron Dataset (Large Parameter Models)

Configuration	Style Transfer (1-5)	Content Pres. (0-1)
Vanilla (Context-Only)	1.90	<b>0.96</b>
RAG ( $top_k = 1$ )	3.07	0.88
RAG ( $top_k = 5$ )	<b>3.41</b>	0.90
RAG ( $top_k = 10$ )	3.33	0.89

## Setup

- 1,000 sampled emails
- Recreator model: gpt-5-mini
- LLM-as-a-judge: gpt-5-mini

# Evaluation: Enron Dataset (Large Parameter Models)

Configuration	Style Transfer (1-5)	Content Pres. (0-1)	Naturalness (0-1)
Vanilla (Context-Only)	1.90	<b>0.96</b>	<b>1.00</b>
RAG ( $top_k = 1$ )	3.07	0.88	0.99
RAG ( $top_k = 5$ )	<b>3.41</b>	0.90	0.98
RAG ( $top_k = 10$ )	3.33	0.89	0.99

## Setup

- 1,000 sampled emails
- Recreator model: `gpt-5-mini`
- LLM-as-a-judge: `gpt-5-mini`

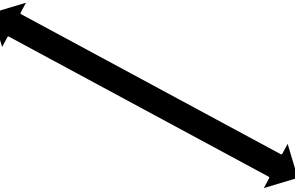
# Example where RAG works well

## Original email

Krishna,  
  
We should invite Kim Watson and her associates as well.  
  
Kay



*LLM-as-a-judge  
Comparison*



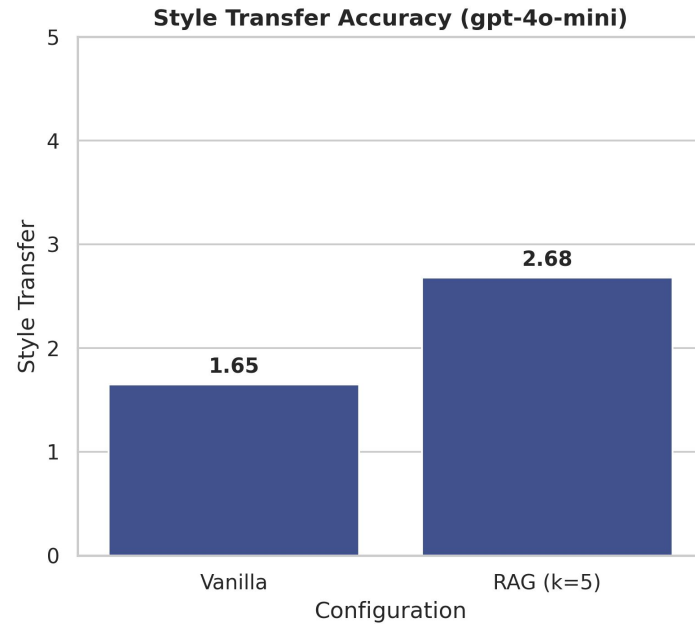
Krishna  
  
Please invite Kim Watson and her associates as well.  
  
Kay

RAG output

Dear Krishna,  
  
Could you please extend the invitation to Kim Watson and her associates as well? Add them to the distribution for the meeting and copy me on the invitation. If you need their contact details or any additional information, let me know.  
  
Thanks,  
Kay

w/o RAG

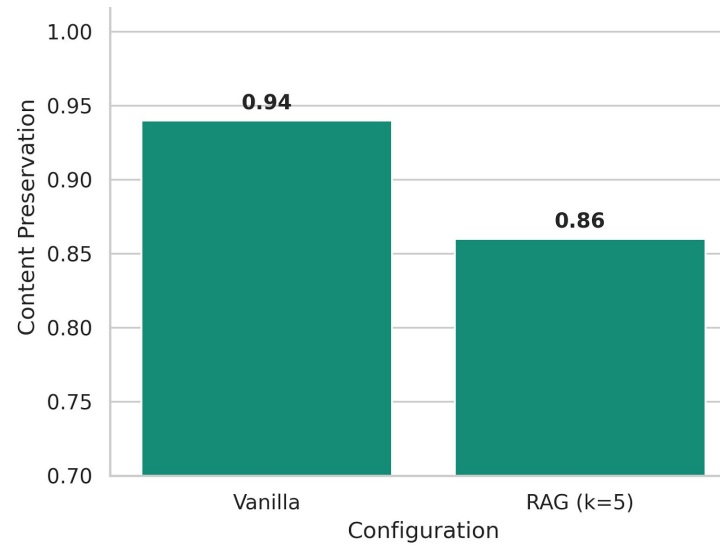
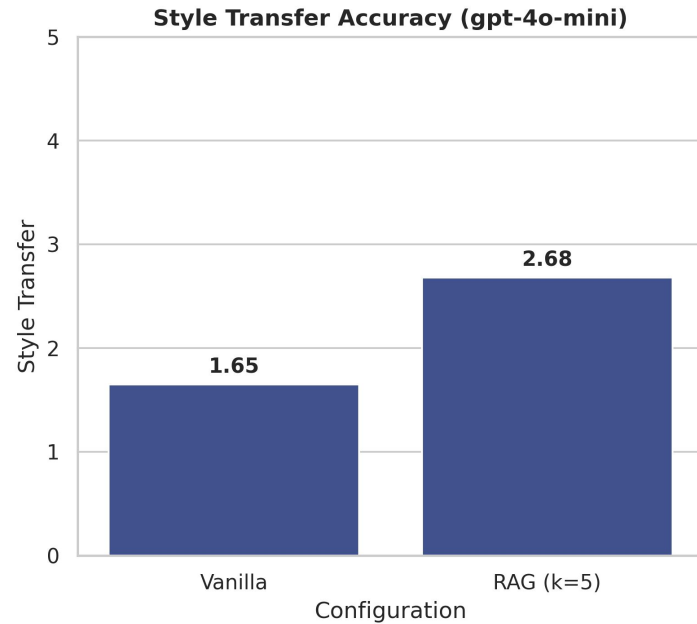
# Evaluation: Enron Dataset (Large Parameter Models)



## Setup

- 1,000 sampled emails
- Recreator model: gpt-4o-mini
- LLM-as-a-judge: gpt-5-mini
- $k = 5$

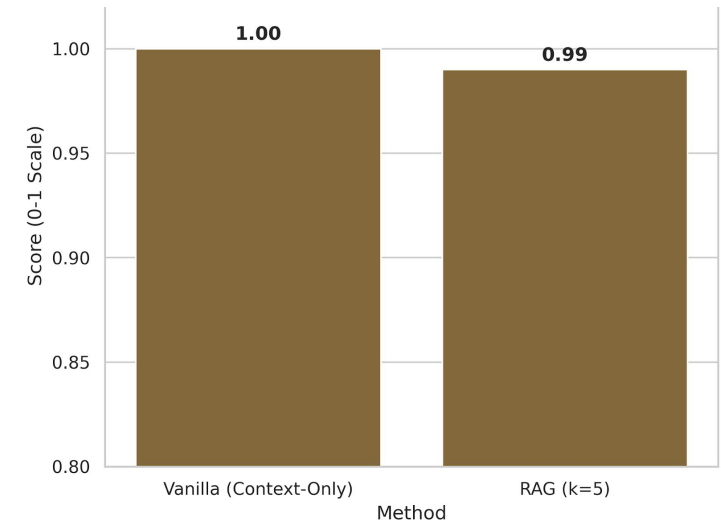
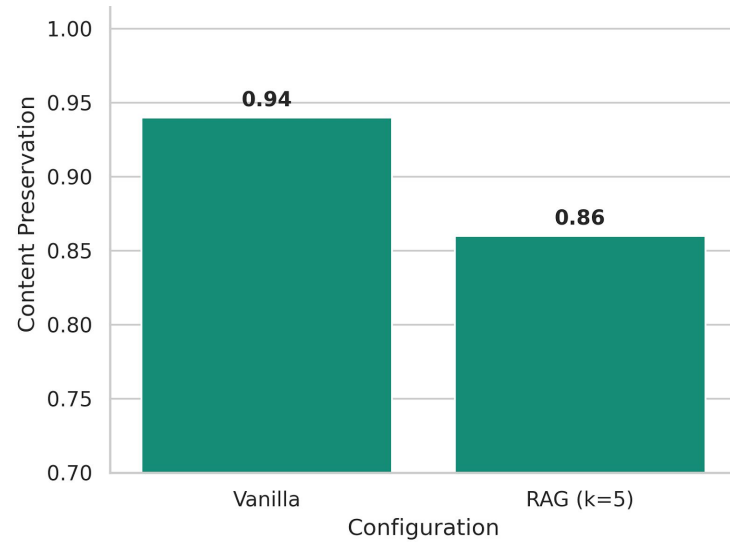
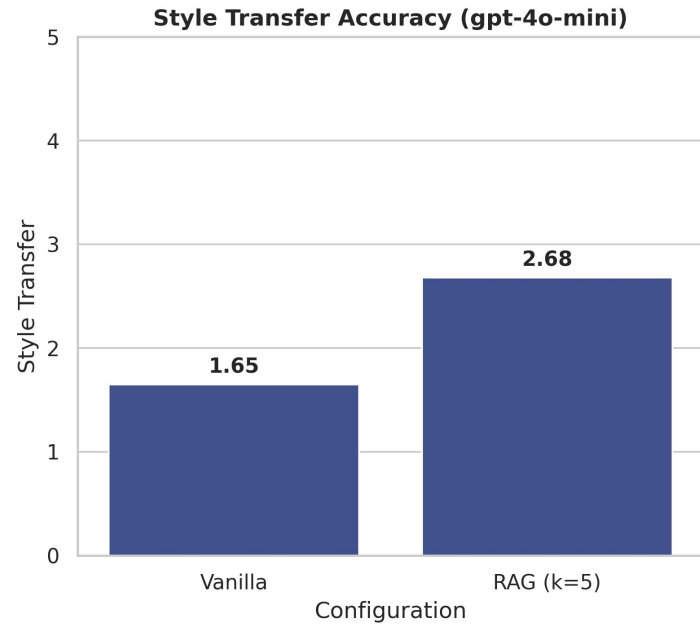
# Evaluation: Enron Dataset (Large Parameter Models)



## Setup

- 1,000 sampled emails
- Recreator model: gpt-4o-mini
- LLM-as-a-judge: gpt-5-mini
- $k = 5$

# Evaluation: Enron Dataset (Large Parameter Models)

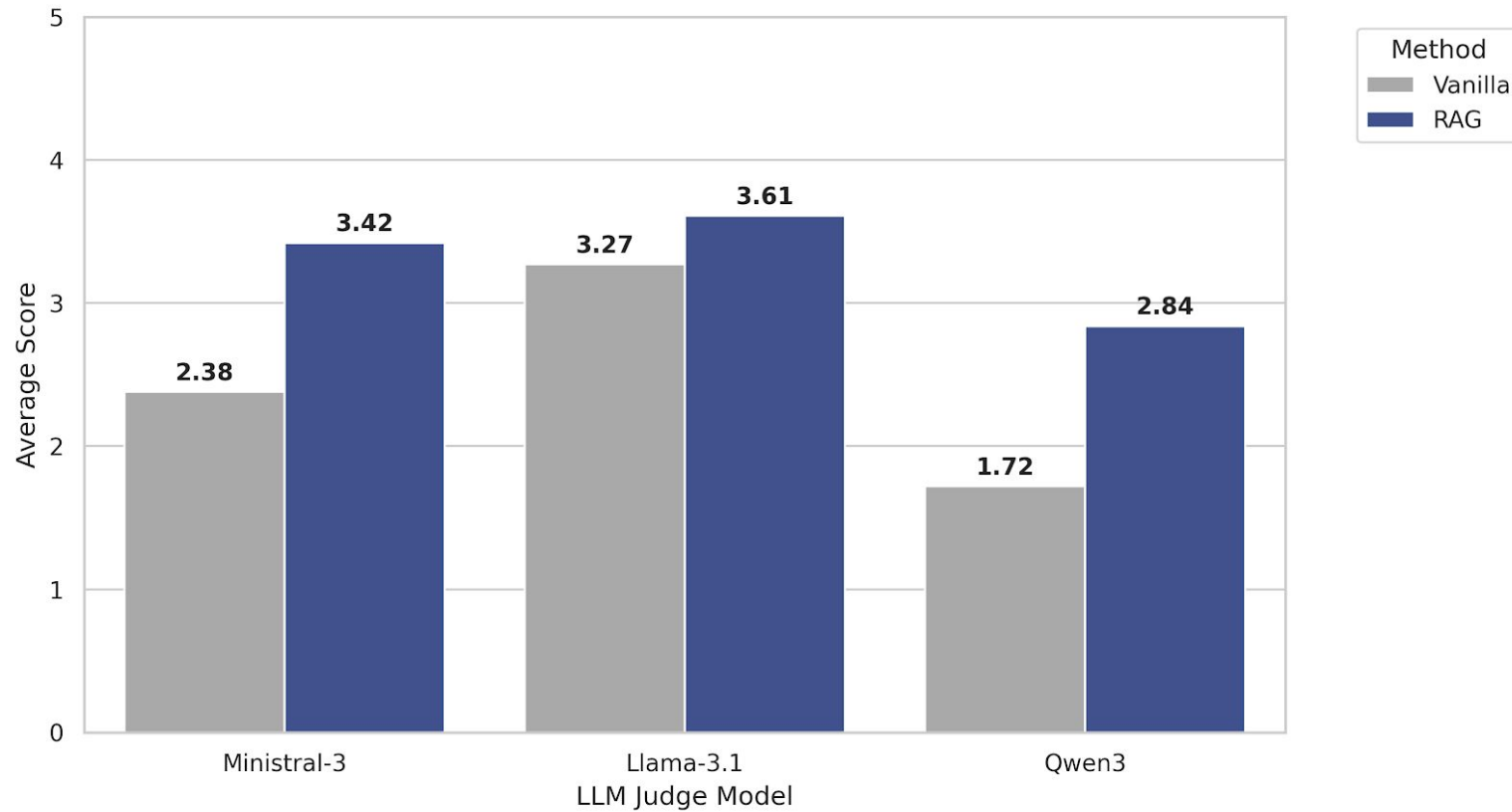


## Setup

- 1,000 sampled emails
- Recreator model: gpt-4o-mini
- LLM-as-a-judge: gpt-5-mini
- $k = 5$

# Evaluation: Enron Dataset (<10b parameters)

## Style Transfer Comparison: Ministral 3 Reconstructor



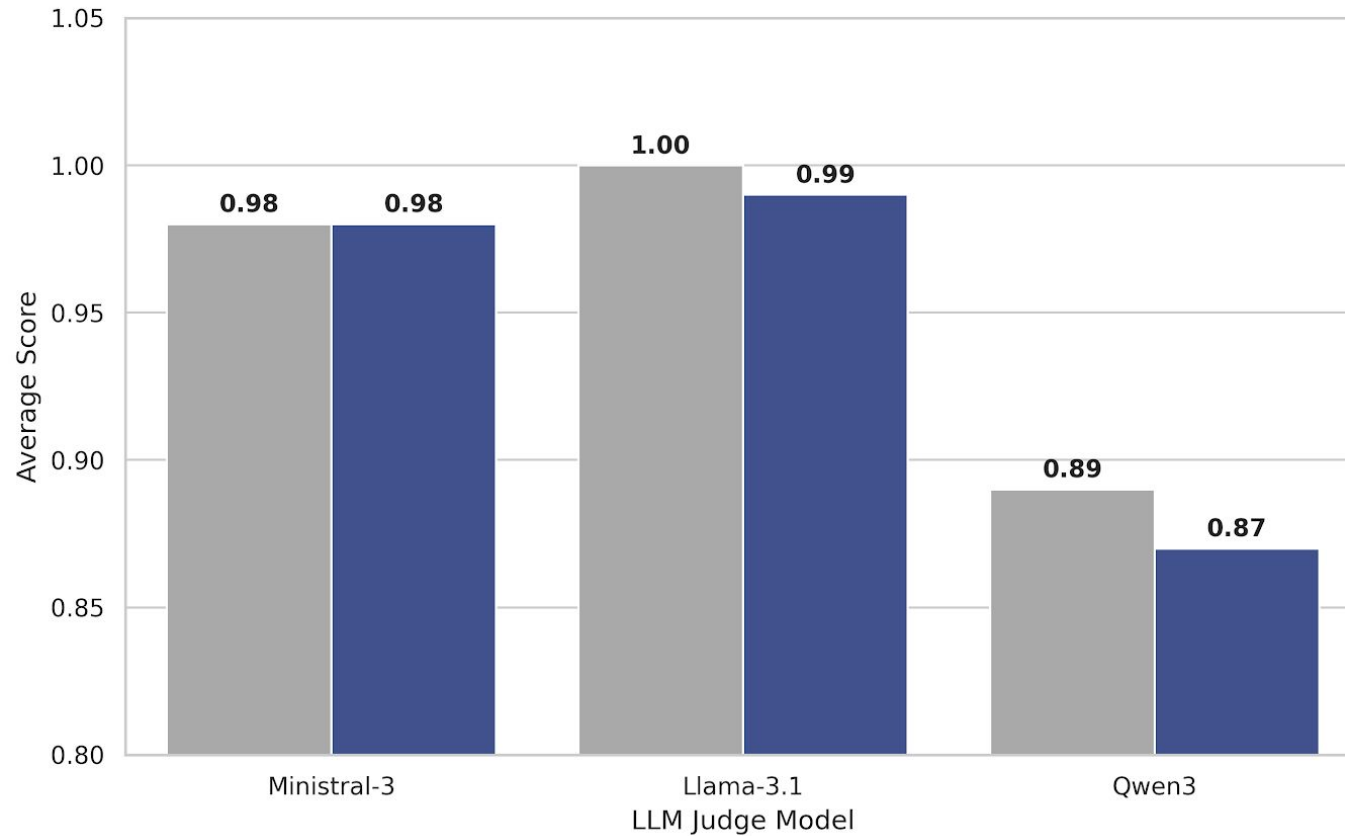
## Setup

- 1,000 sampled emails
- Recreator model: Ministral-3
- $k=5$

Style Transfer Scores using different LLM-Judges

# Evaluation: Enron Dataset (<10b parameters)

Content Preservation Comparison: Ministral 3 Reconstructor



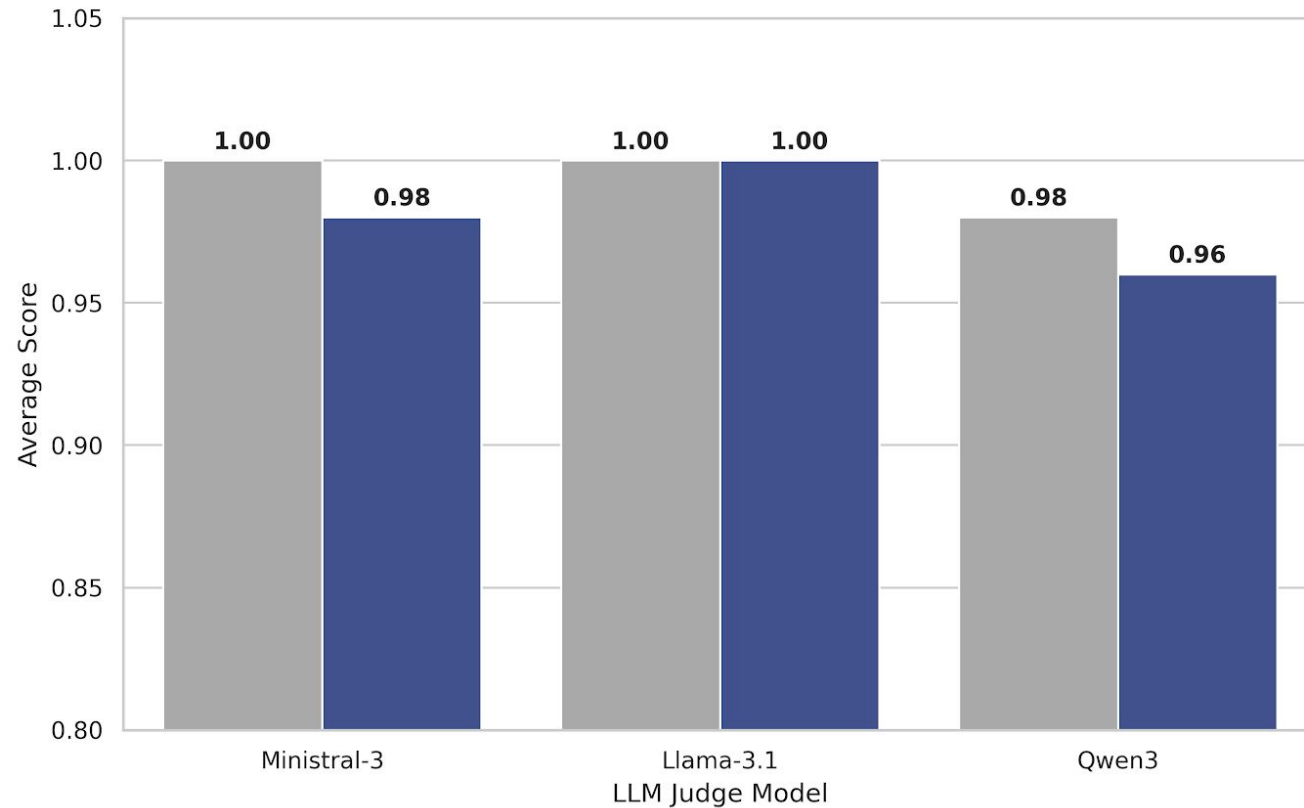
## Setup

- 1,000 sampled emails
- Recreator model: Ministral-3
- $k=5$

Content Preservation Scores using different LLM-Judges

# Evaluation: Enron Dataset (<10b parameters)

Naturalness Comparison: Ministral 3 Reconstructor



## Setup

- 1,000 sampled emails
- Recreator model: Ministral-3
- $k=5$

Naturalness Scores using different LLM-Judges

# Context Leakage

- Injecting emails for In Context Learning can lead to “context leakage”
  - The LLM adds unnecessary information from the retrieved emails which distort the meaning of the generated e-mail.

Shirley,

Please, send it to the entire group.



Shirley,

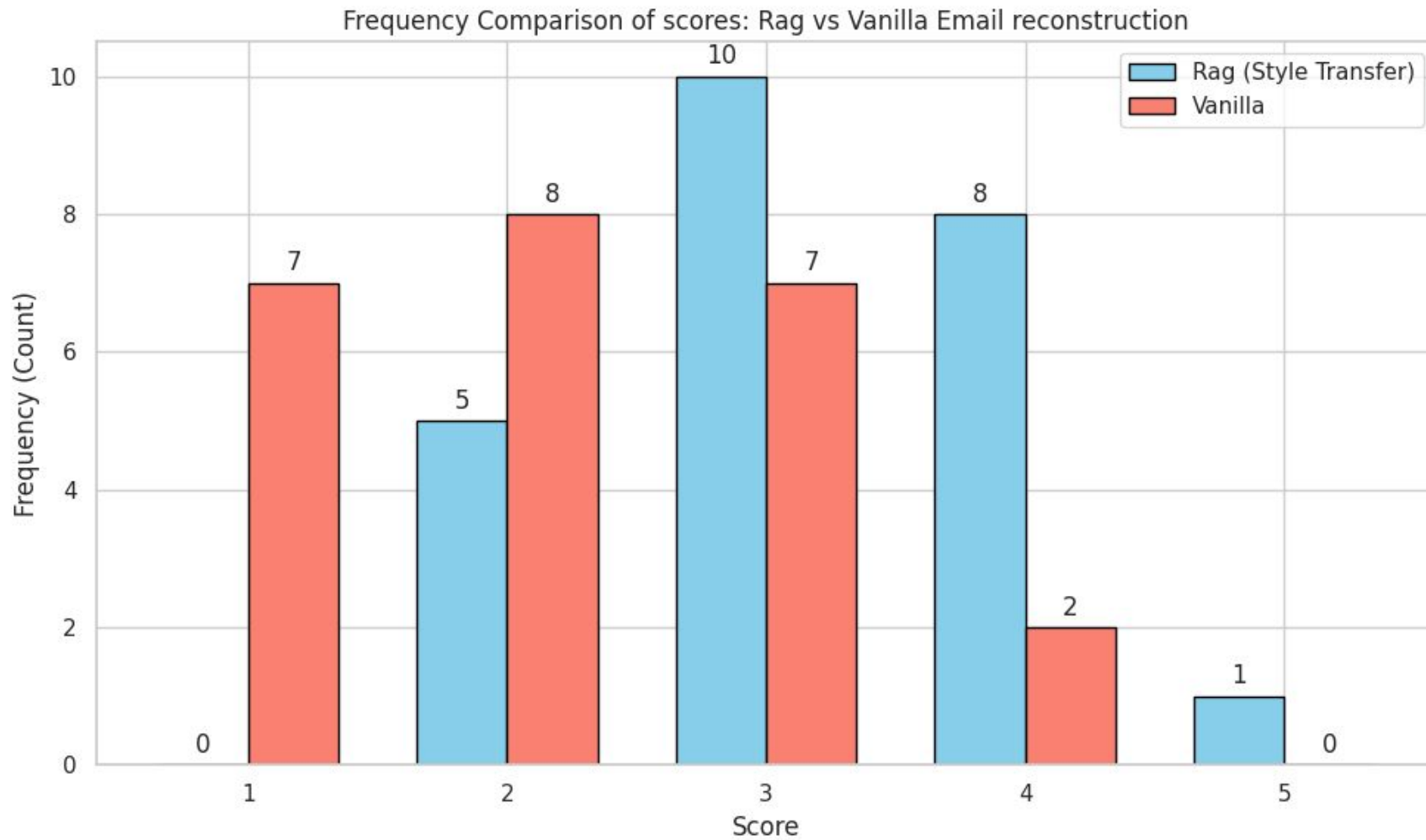
I shall remind the group once more about our collective duty to leave the conference rooms in order after our meetings.

Meaning NOT preserved

# Evaluation: User study

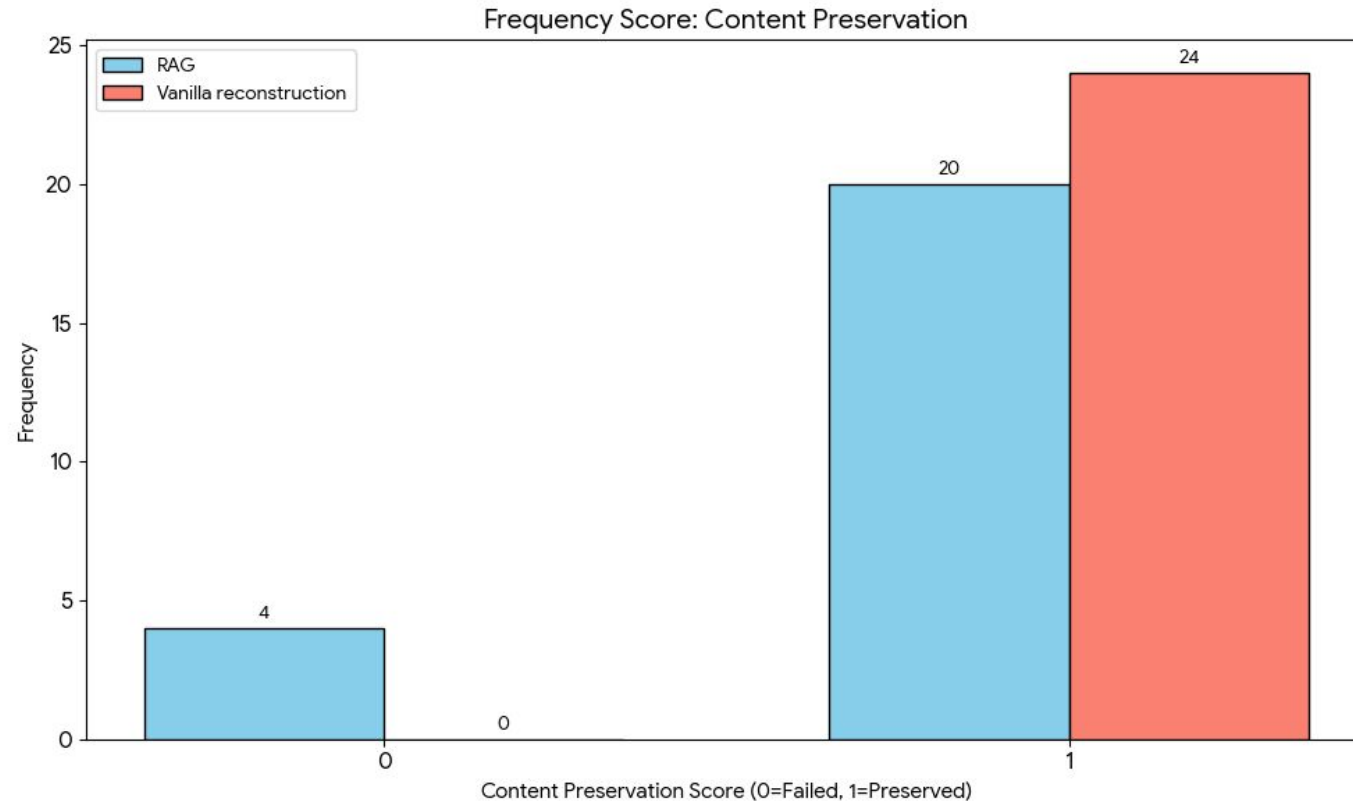
- The user study was done using a Web application which allows the users to upload their emails and arranges to them into threads.
- The users can choose these emails and reply to existing threads after providing the context.
- The users are then provided with two replies
  - Without our RAG pipeline
  - With our RAG pipeline (only recreated with Gemini 2.5 Flash)
- The user has no knowledge which reply is generated with our RAG framework.

# Evaluation: User study



k = 5 retrieved e-mails

# Evaluation: User study



# Conclusion

- Results demonstrate that RAG is a robust solution for style transfer.
- Further research needs to be done on context leakage
- Further experiments need to be done on different domains
- User Study and Automated evaluation both suggest that RAG is a viable and scalable solution for Content Style Transfer.



---

*Thoughts?*



---