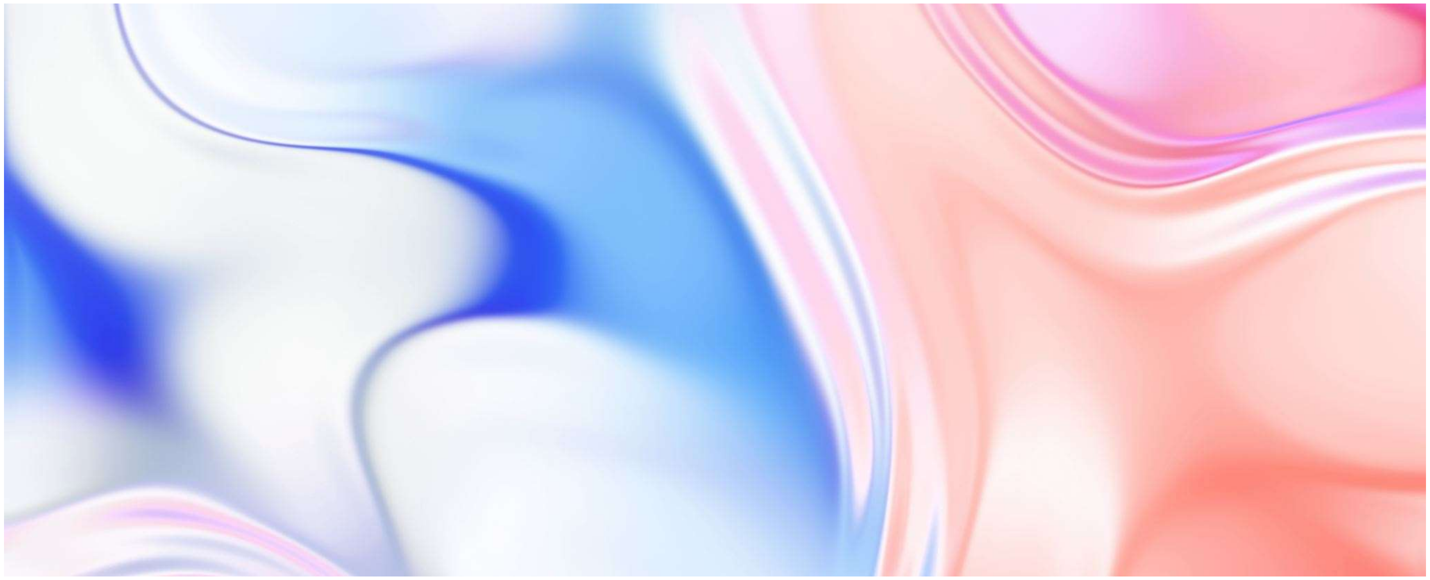
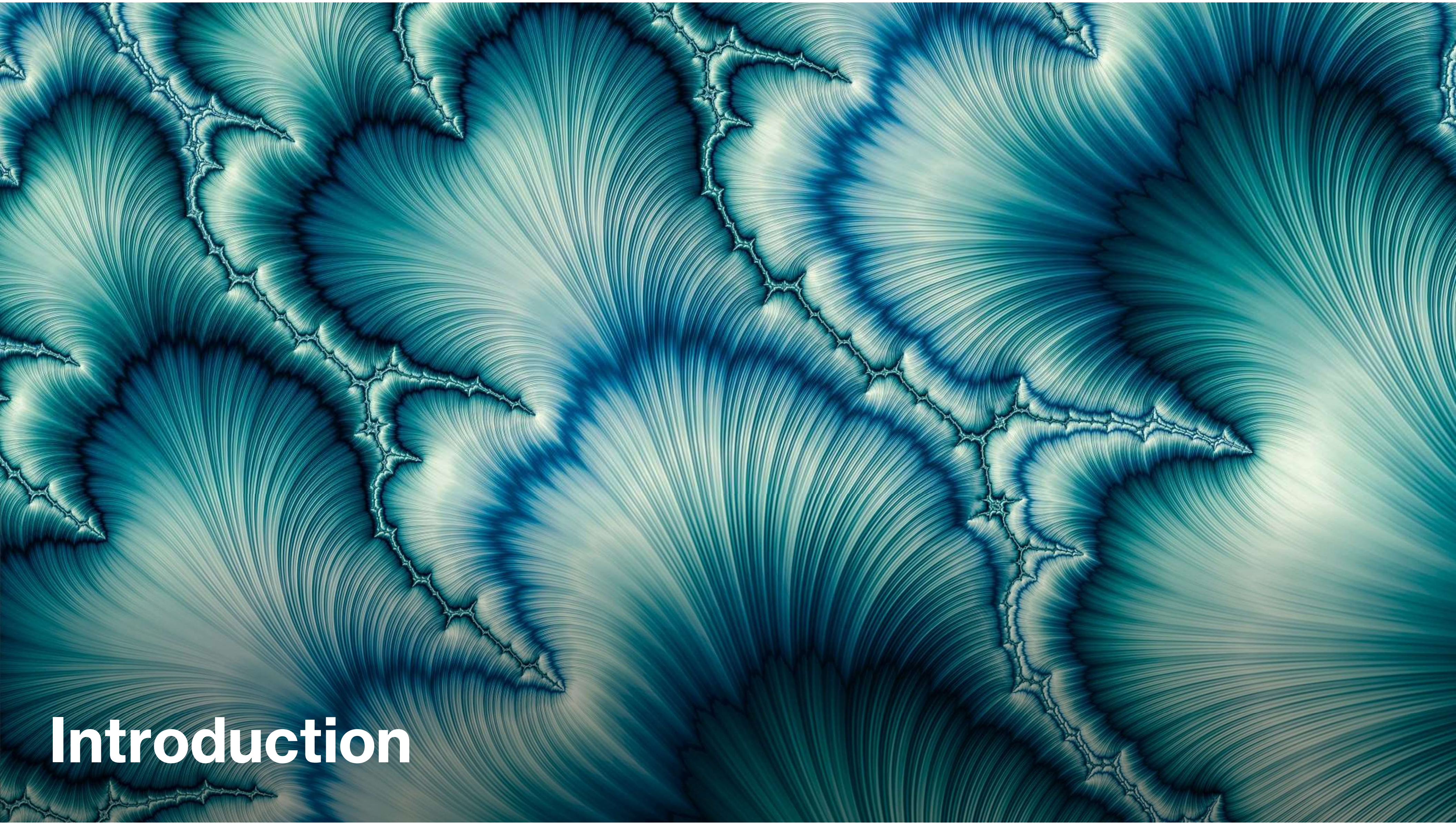
Unsupervised Human Preference Learning

Sumuk Shashidhar, Abhinav Chinta, Vaibhav Sahai, Dilek Hakkani-Tür

Department of Computer Science, University of Illinois Urbana Champaign



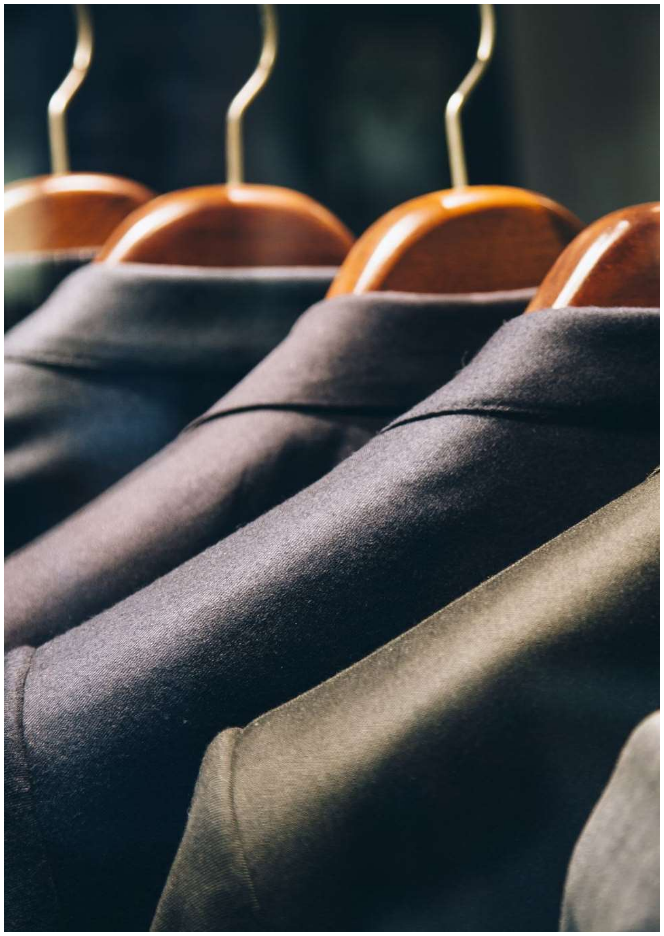


Introduction

The Value of Personalization

- Personalization enhances user engagement and satisfaction.
  - 80% of consumers are more likely to do business with a company that offers personalized experiences *(McKinsey & Company, 2020).*

- Historically, personalization has been a luxury reserved for a select few.
  - Bespoke Clothing
  - Personal Concierge
  - Custom Financial Advice



McKinsey & Company. (2020). Personalizing the customer experience: Driving differentiation in retail.

Gap in LLM Personalization

- LLM-as-a-service is growing rapidly *(Dimension Market Research, 2024)*.
- Yet, personalization in the LLMs of today is not commonplace or expected, though they excel at task completion capability.
  - For example, GPT-4 can write a correct and factual email, but it may not reflect an individual's unique style or preferences.
  - We have a chance to democratize personalization in LLMs and craft compelling experiences for humans.

Subject: Update on Deal Confirmations #478168, 478944, 478945, and 478946

Dear [Recipient's Name],

I hope this message finds you well. I wanted to update you regarding the confirmation of deals #478168, 478944, 478945, and 478946.

Due to the holiday season, our office attendance has been fluctuating. However, I have forwarded your email to Christian and Greg, who are expected to be in the office today. I will check in with them to confirm whether they have already sent out the confirmations. If not, I will inquire about the expected timeline and let you know as soon as possible.

Thank you for your patience. I will get back to you with an update from Christian or Greg as soon as I have more information.

Best regards,

[Your Name]
[Your Position]
[Your Contact Information]
[Your Company]

GPT4 Email

I've forwarded this e-mail to Christian and Greg. Unfortunately, the population of the office is quickly dwindling as the holidays come near. I've been told that one or the other of them will be in today, and I'll find out if something has been sent out already, and if not, how soon we can do that. I'll get back to you as soon as I talk to them.

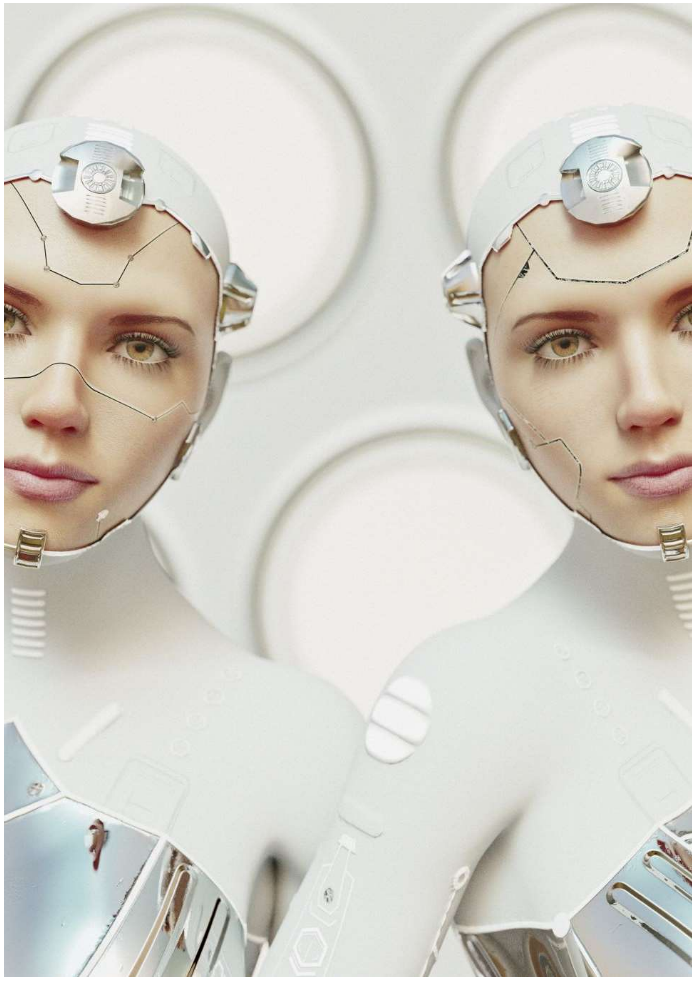
Thanks,
Kate

Ground Truth

Dimension Market Research. (2024). Large Language Models Market: A Comprehensive Analysis of the USD 140.8 Billion Market by 2033.

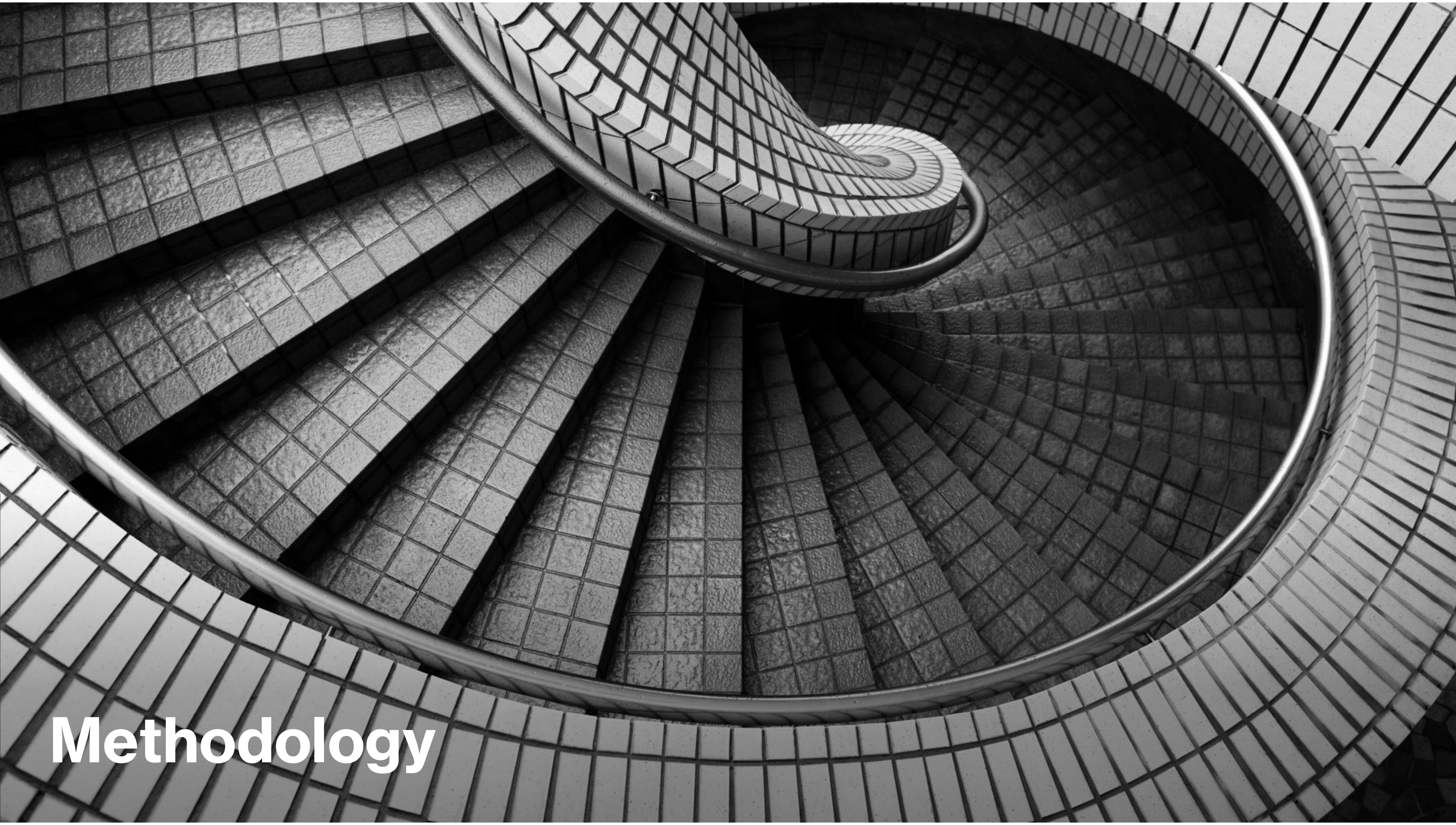
Limitations in Existing Methods

- There has been substantial prior work that assists with personalization:
  - **In-Context Learning**
    - Struggles to capture complex individual preferences with limited examples.
  - **Fine-Tuning**
    - Requires large, annotated datasets that users typically don't have.
    - **PeFT** methods such as QLoRA *(Dettmers et. al, 2023)*, Prefix Tuning *(Li and Liang, 2021)*, etc., are expensive and impractical at user-scale for large models.
    - Model Dependent, i.e., adapters trained for one model cannot be applied zero-shot on others



Dettmers et. al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs
Li and Liang, (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation

Can we decouple preference adaptation from the base model?



Methodology

Preference Agents

- Small Preference Agents
  - Separate from capable, large foundation models.
  - Locally Trainable and Inferable

Process:

1. Generate NL *preference rules* from existing user demonstrations of task completions.
2. Train a small, *preference agent* to produce preference rules using LoRA.
3. Use generated preference rules to *steer large task completion models*

Intent Data Generation

For the user data corpus, we first need to understand the user intent behind each demonstration

1. Generate a hypothetical user intent with a large model, for each data point
2. Use a cognitive scratchpad and few-shot examples to enhance the realism of the hypothetical intents.
3. Repeat generations with diverse hyperparameters to simulate real human interaction.

```
Variant 2:
<thinking>
1. Identify the main surprise or
   unexpected event mentioned in the
   email.
2. Note who initiated the contact and
   what they want to do.
3. Capture Jeff's skepticism and
   uncertainty about the outcome.
4. Mention Jeff's intention to follow up
    with Michael.
</thinking>
<bullet_points>
- Jerry Bloom left a message for me,
   wants to talk
- I'm skeptical and wouldn't get hopes
   up
- Will let Michael know what Jerry says
- Uncertainty about the outcome
</bullet_points>
```

Excerpt of thinking tokens and intents

Training Data Generation

We contrast the zero-shot completion of the foundation model for a given task, with the ground truth to generate guiding rules.

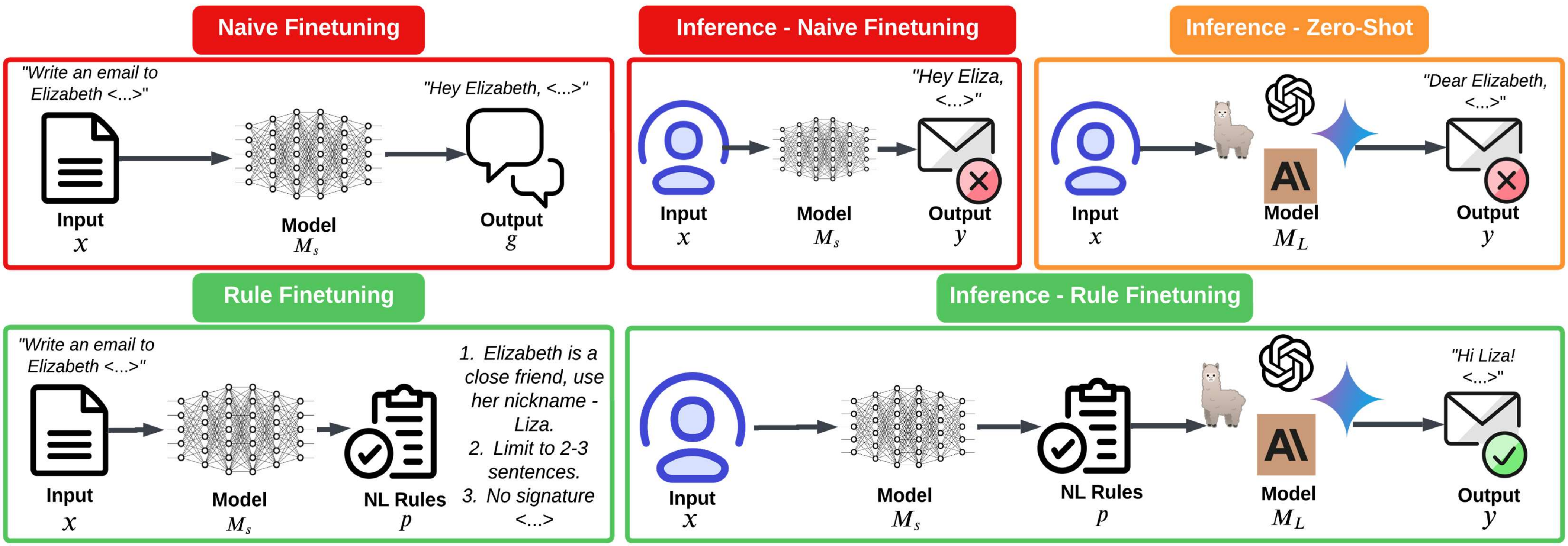
1. Generate a zero-shot output by a large, task completion model for each data point.
2. Compare the differences between the zero-shot completion of the large model, and the ground truth.
3. Formalize these differences into sets of “preference rules”.
4. Repeat for all data points until rich, preference information sets are distilled.

```
6. **Grammar and Punctuation**: The
   email uses casual grammar and
   minimal punctuation. Avoid
   unnecessary capitalization and
   formal sentence structures.
</thinking>
<rules>
1. **Length and Structure**: The email
   should be concise, consisting of
   approximately 8 sentences. Avoid
   adding formal greetings or sign-offs.

2. **Tone and Style**: Use an informal
   and conversational tone. Include
   contractions and colloquial phrases
   to reflect the casual nature of the
   communication.
3. **Market Terminology**: Incorporate
   specific market-related terms and
   shorthand (e.g., "VAR", "h/j spread
   ", "Access", "bearish to neutral
   trade"). Ensure the language
   reflects insider knowledge and
   jargon of the trading community.
4. **Observational Detail**: Provide
   specific details about market
```

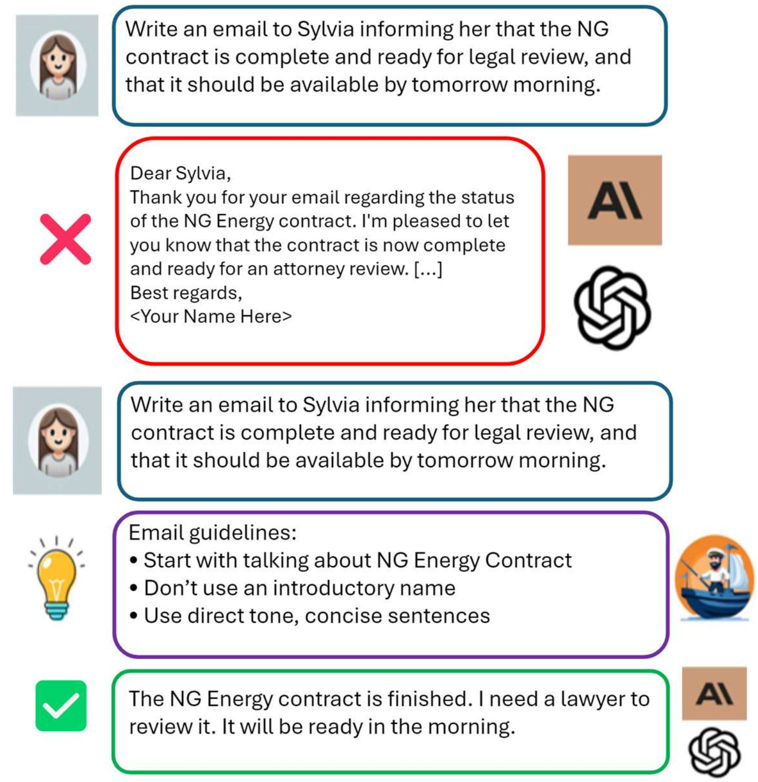
Excerpt of thinking tokens and rules

Rule Finetuning



Demonstration

- User provides task intent, and information to write an email to a colleague.
- The preference agent, based on information about the user, and the given task, generates preference rules, sent along with the task to the large model.
- The foundation model follows the preference rules to complete the task.
- These focused, distilled preference rules are sufficient for the large model to understand the user's preferences.



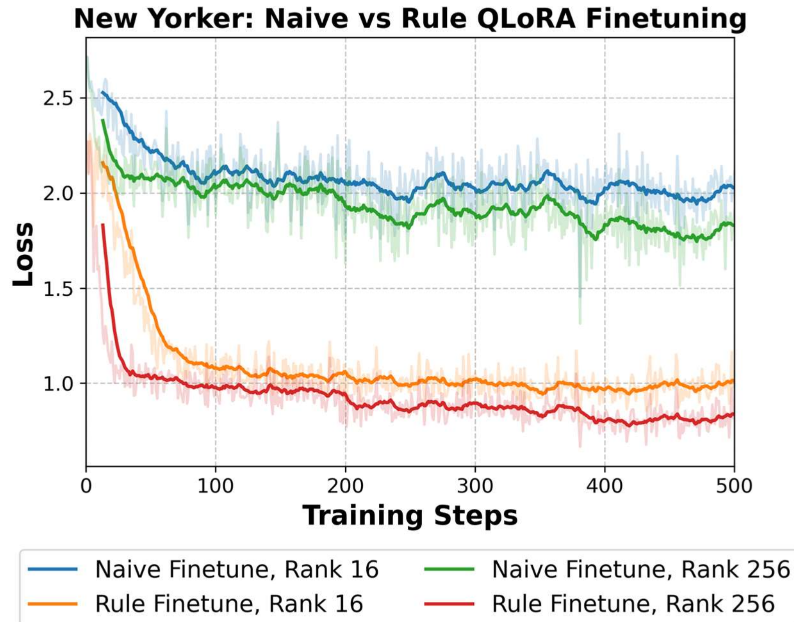
RESULTS

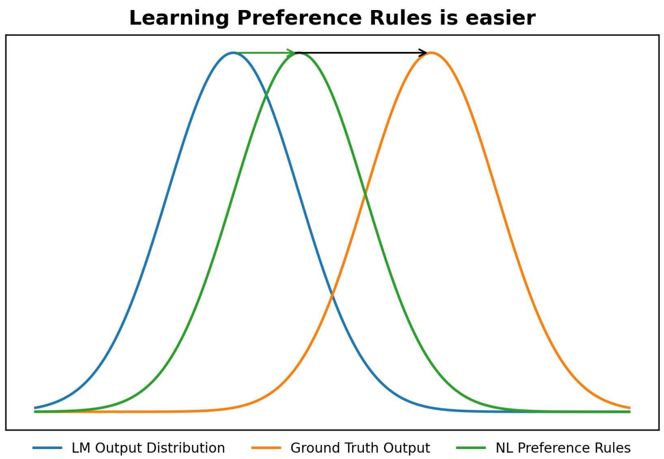
Naïve vs Rule FT

- Finetuning on preference rules **<u>converges faster</u>** than finetuning on content.

    - *<u>Hypothesis</u>*: It is easier for models to learn the structured nature of preference rules, rather than distributionally different User Generated Content (UGC)

    - Leads to better performance with fewer training examples.





Performance Comparisons

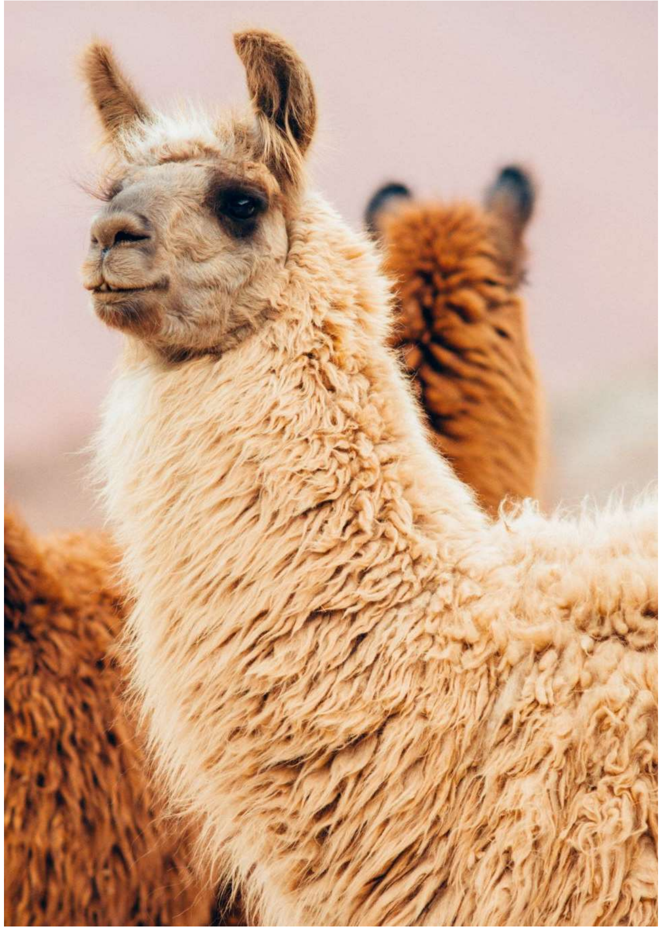
- We show win rates of a Llama3B preference agent and 3 different foundation models, against 4 baseline methods.
- Stronger task completion models demonstrate stronger performance (Correlated to MMLU)
- Datasets with high personality information (e.g. Enron Email Corpus) benefit more from preference agents.
- No baseline Agent: Version of the Preference Agent technique, without zero-shot cross distillation of preference rules.

Preference Agents	New Yorker			Enron			LAMP 3U			Aggregated	
$M_L \rightarrow$ vs $\downarrow$	**Llama3 70B Instruct**	**Claude 3.5 Sonnet**	**Gemini 1.5 Pro**	**Llama3 70B Instruct**	**Claude 3.5 Sonnet**	**Gemini 1.5 Pro**	**Llama3 70B Instruct**	**Claude 3.5 Sonnet**	**Gemini 1.5 Pro**	**LLM Evaluation**	**Human Evaluation**
Small Baseline	77.4	91.5	80.0	88.4	96.1	89.8	74.6	84.0	75.3	84.1	91.0
Large Baseline	67.7	75.2	66.9	85.6	83.7	88.2	66.5	69.5	63.8	74.1	84.5
Few Shot	68.3	62.0	66.7	61.1	68.0	57.4	58.3	57.4	59.4	62.0	73.4
Naive Finetune	80.3	82.4	81.8	75.3	87.8	81.3	85.0	92.7	89.0	83.9	92.2
No Baseline Agent	65.1	68.8	63.8	58.4	61.3	62.5	63.8	67.2	60.4	63.4	52.0

Table 2: **Win Rates** of Llama3 8B M_s combined with various M_L, evaluated by GPT4o and human evaluation.

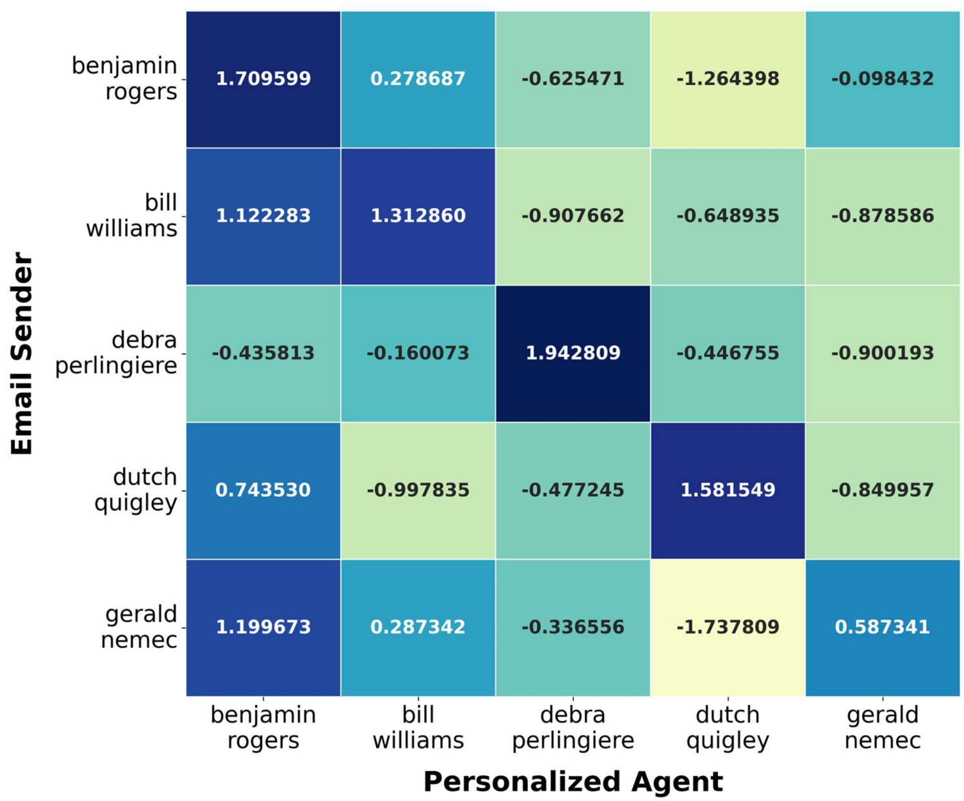
Model Specific Semantic Understanding

- Models understand their own generated rules better.
  - For example, Llama-3 models interpret rules generated by themselves more effectively than those from GPT-4 or humans.
- Human-written rules led to a 16.8% performance drop compared to model-generated rules.
  - Models interpret keywords like “concise”, “informal”, etc, differently compared to what humans expect them to mean.
- Indicates that semantic understanding is model-specific, even with natural language rules.



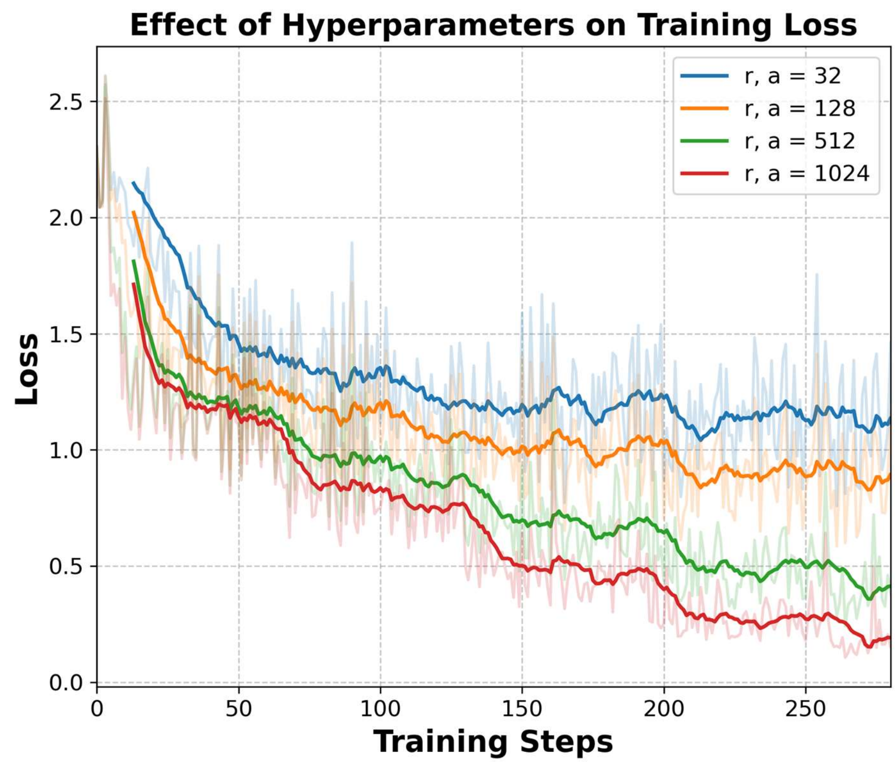
Does the model truly learn individuality?

- We train preference agents for 5 different users from the Enron corpus.
- We permute preference agents for email tasks on all users
- We observe, through Bert Score, that agents perform best on their source users, indicating that they capture individual user styles accurately.



Effect of Rank

- We find that training more parameters (with higher LoRA ranks), leads to lower losses
- This indicates that there is a rich amount of preference information contained within the distilled preference rules



Interpretability

- We also achieve interpretability as natural language preference rules can be examined to understand task completion model outputs.

- For example, we understand that the preference agent learnt that the user:
  - usually signs of as “Kay” (hence instructs the large model to close with “Kay”)
  - Refers to the recipient as “John”, instead of his full name (likely due to closeness or friendship)

1. **Length:** The email should be between 100-120 words. ...
5. **Nicknames:** Use the nickname "John" instead of "John Keffer".
...
8. **Nuance:** The email should use phrases that imply a more casual tone, such as "Some future deal", "Thanks for the memory" or "Looks like we got lucky, starting with Inter".
9. **Content:** The email should express gratitude for John's help on the Inter deal, look forward to future collaborations, and mention checking availability for dinner at John's house.
10. **Closing:** The email should end with a casual closing, such as "Kay".

John,
I appreciated your help. I'm glad we got to do a real deal, even if it was a strange one. More to come! I'll check on the sitter, but dinner on the 2nd sounds great.
Kay

Conclusions and Future Work

Effectiveness At Scale

- Can be run locally on end user hardware to guide large, cloud models.
- Device-Locality of preference models preserves privacy of user data.
- No modification of large models, while harnessing their excellent capabilities
- Can be re-used across different large models without re-training
- Improved tone, style, content relevance and reliability compared to naïve FT

Future Work

- Exploration of preference adaptation beyond style and structure to traditional LoRA tasks.
- Preference learning across various modalities, such as image, audio, video and action
- Optimization of rule generation and preference distillation
- Direct embedding communication (in non natural language) for efficiency, without sacrificing interpretability.

Thank You!