We're not speaking the same language: Approaches and Challenges in the Personalization of LLMs

Mauriana Pesaresi Seminars

Claudia Nanula, 14/03/2025

Outline

1. Definition and Background
2. Techniques for LLMs Personalization
3. Evaluation Methods
4. Open Problems and Challenges

What does Personalization mean?

- In general, personalization refers to the process of tailoring a system's output to meet the individual preferences, needs, and characteristics of an individual or a group of users.

Why would we need personalization?

Such personalization is crucial for human-AI interaction: it is expected to enhance user satisfaction by providing more relevant and meaningful interactions, ensuring users receive responses that are more aligned with their needs and expectations.

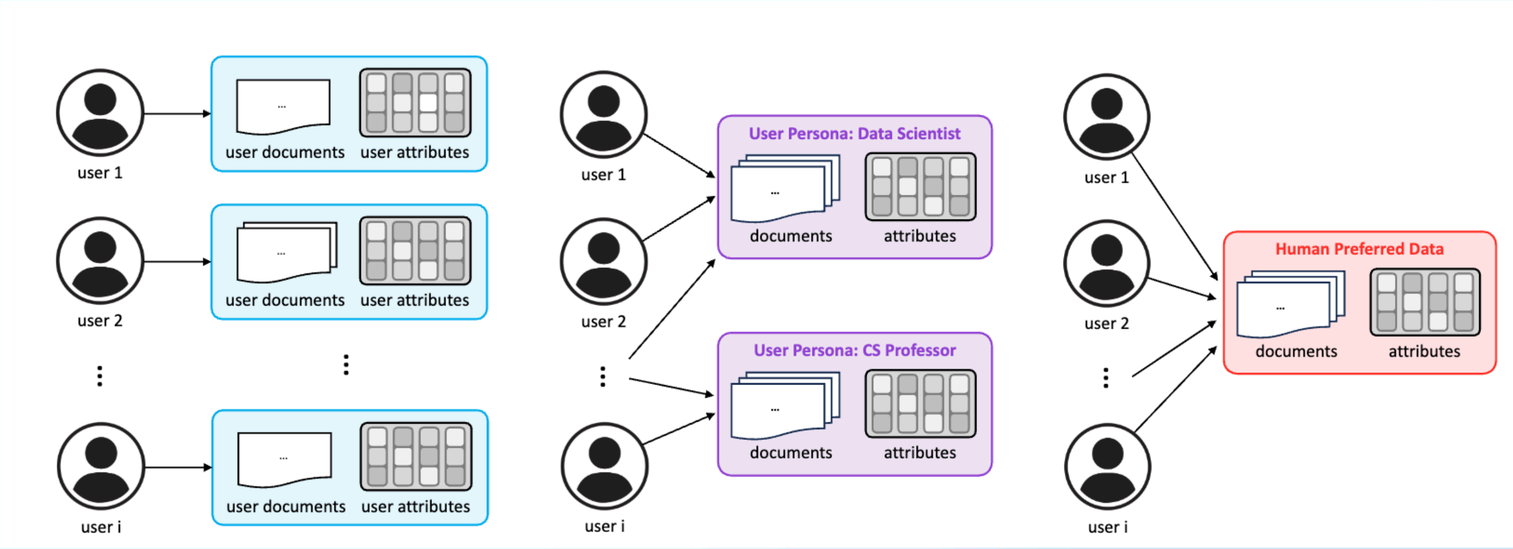
The **open problem** is that most of the current technologies are thought to be used by a "**standard" user**, meaning that they imply some generalizations about the type of people will eventually use these technologies. This translates into a possible dissatisfaction or even total exclusion from the usage of such systems.

Generative systems, in particular, could reflect some standard variety of content, language, culture, leading to potential biases and unfairness towards some categories.

Personalization Granularity of LLMs

Who is your audience?

- The level of granularity refers to how finely or broadly personalization is applied. We can make a distinction between:
  - ***User-level personalization*** focuses on tailoring outputs for individual users, using unique preferences and data, such as personal information and interaction history.
  - ***Persona-level personalization*** targets groups of users who share similar characteristics or preferences, known as *personas;* it is based on the collective attributes of these groups, such as expertise, informativeness, and style preferences.
  - ***Global preference personalization,*** this level encompasses general preferences and norms that are widely accepted by the general public. For example, broadly accepted cultural standards and social norms.



The granularity of personalization in LLMs involves trade-offs between **precision, scalability, and richness** of personalized experiences.

Personalized Criterion Taxonomy in LLMS

What are we trying to achieve?

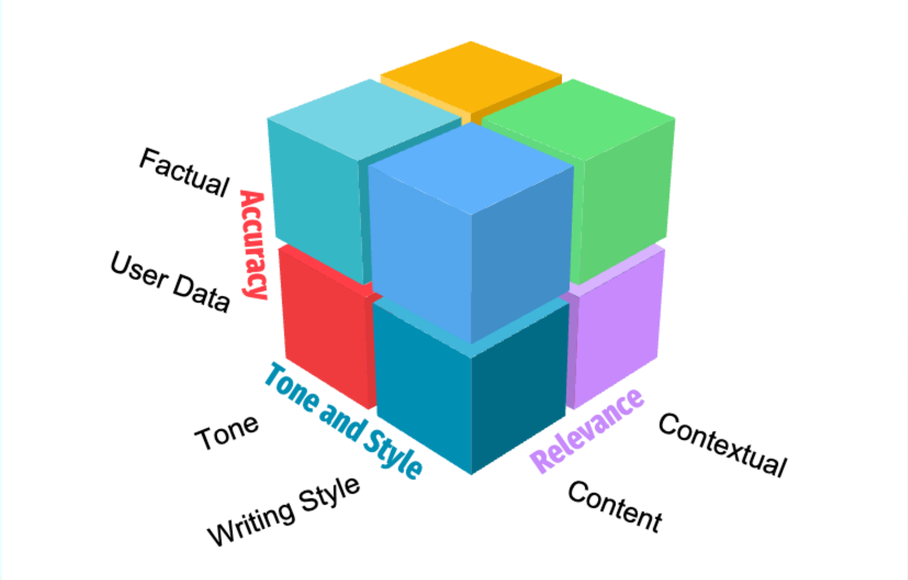
- There are several critical aspects we need to consider when developing and evaluating a personalized system.
- We can divide these features in three categories, which are nonetheless correlated to each other:

    1. ***Tone and Style:*** *Is the writing style consistent with the user's preferred style or previous interactions? Does the tone of the text match the user's preferences (previous written text) and context (e.g., formal, casual, etc)?*

    2. ***Accuracy:*** *Does the content match the user's interests, preferences, and needs? Is the content appropriate for the specific context/situation that the user will encounter it?*

    3. ***Relevance:*** *Are the facts and information presented in the text correct and reliable? Is the personalized content based on accurate and up-to-date user data?*

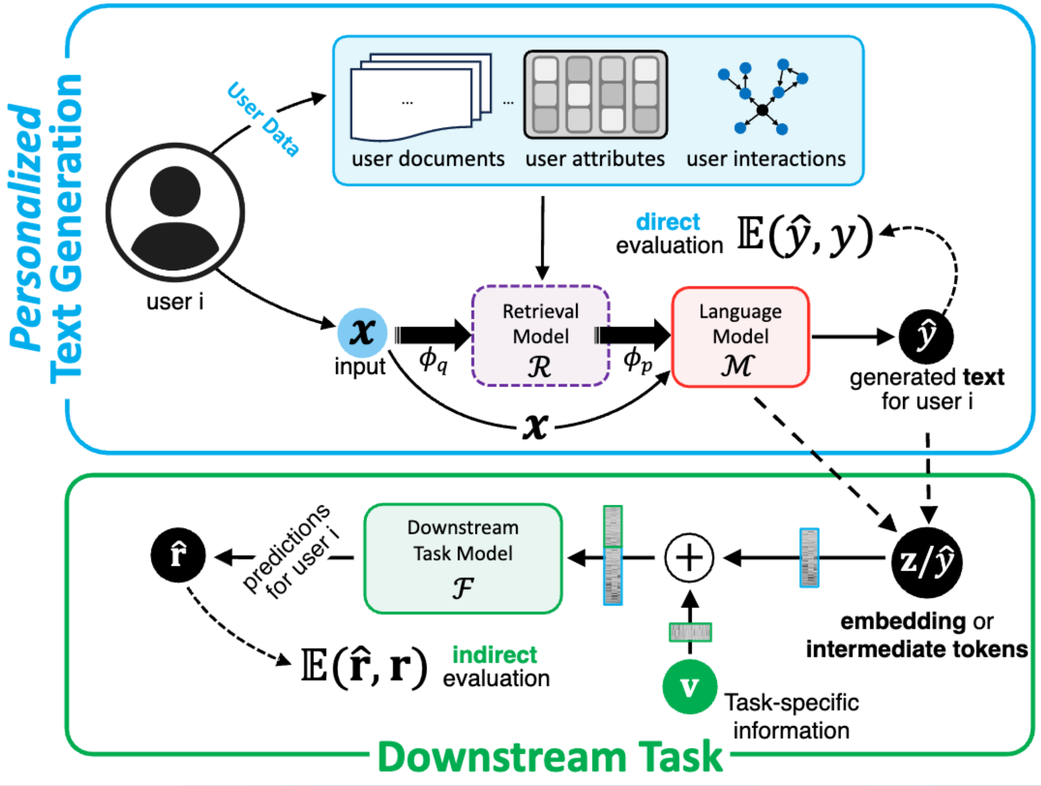
- The point here is that these characteristics are essentially related to the use of language, and even if LLMs are excelling at generating it they still struggle to produce text in a meaningful and variable way for the user.



In the context of LLMs

There are two main categories of LLMs' usage for personalization:

- ***Personalized Text Generation,*** whose goal is to generate text that directly aligns with individual or group preferences, for example a personalized mental health chatbot.
  - This process of personalization could be *direct*, in the sense that the personalized content is evaluated against some ground-truth, to assess the **quality** of the generated text.
  - However, there's a problem of scarcity of gold data for this purpose.
- ***Downstream Task Personalization,*** here LLMs are used to enhance the performance of a specific task, i.e. recommendation, aiming at improving the task rather than the text itself.
  - This process is often *indirect*: LLMs are used to generate a personalized representation which is simply added as information needed for the task.



Personalization techniques for LLMs

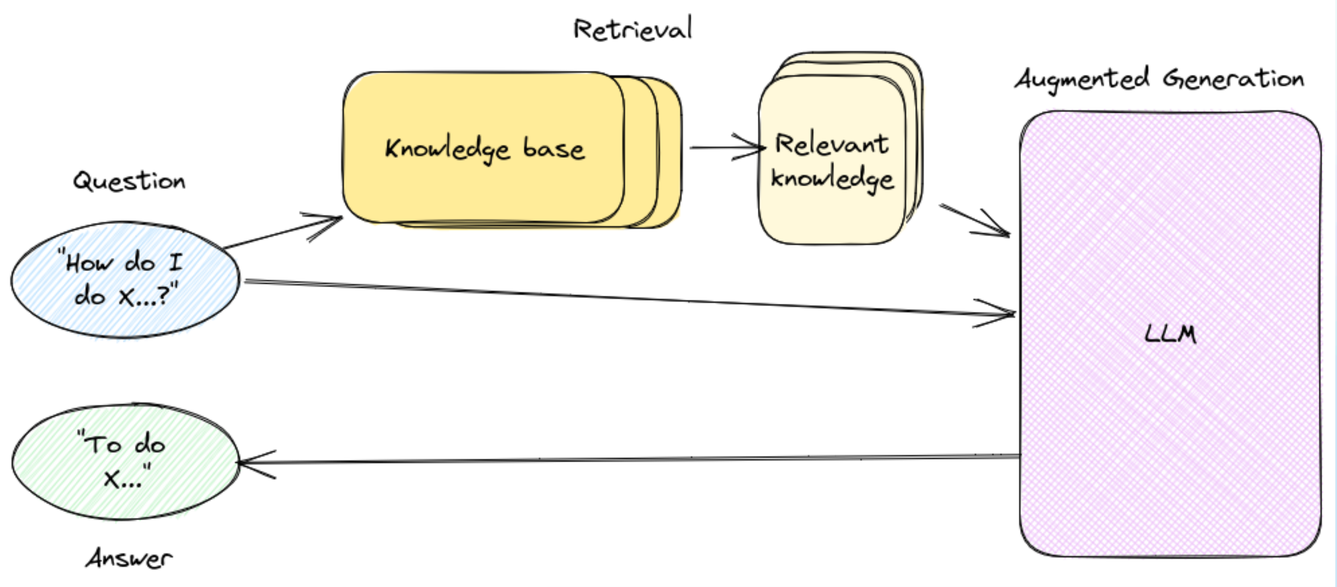
Techniques can be categorized according to the way user information is used.

- This is just a theoretical distinction, in practice the dimensions in which each of these approaches work are orthogonal and can coexist at the same time.

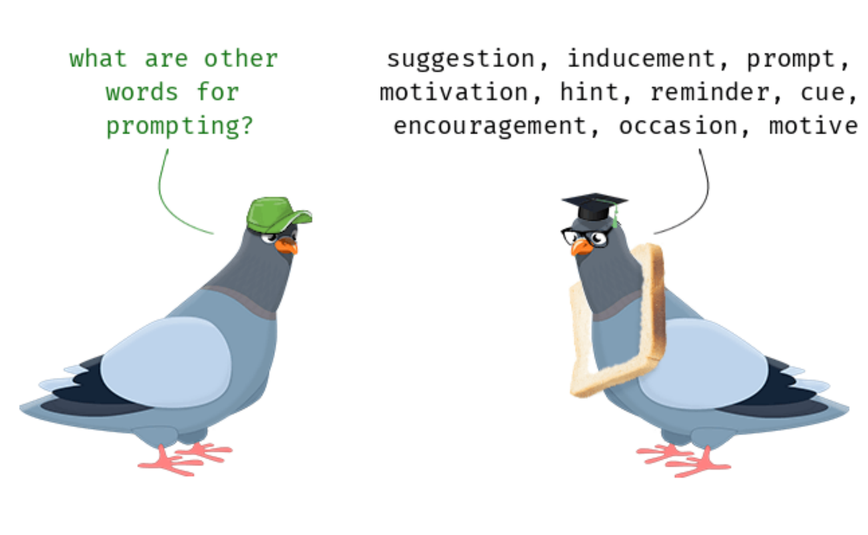
  - **Personalization via Retrieval Augmented Generation (RAG)**
  - **Personalization via Prompting**
  - **Personalization via Representation Learning**
  - **Personalization via Reinforcement Learning from Human Feedback (RLHF)**

Personalization via RAG, or Giving the model the information it needs

This approach combines the generative capabilities of an LLM with external information to better satisfy user requirements. It leverages a Retrieval Model— either sparse (e.g., TF-IDF, BM25) or dense (using embedding spaces)— to fetch relevant information. The retrieved data is then incorporated into the prompt and used to augment the model's response generation.



Personalization via Prompting, or Telling the model what we want it to do



- A prompt serves as an input for a generative model, guiding the content it generates. In the field of personalized, we can distinguish:

- ***Contextual Prompting:*** these methods directly incorporate user history information into the prompt.

- ***Persona-based Prompting:*** these approaches introduce specific personas into the prompt ("Act like..."). By encouraging LLMs to role-play these personas, it aims to enhance the performance of downstream personalization tasks.

- ***Profile-Augmented Prompting:*** these methods focus on designing prompting strategies that enrich the original user history information, which can be full noise or absent at the very beginning (cold-start problem).

- ***Prompt Refinement:*** this category of methods focuses on developing robust frameworks that iteratively refine the initial hand-crafted prompts.

Personalization via Representation Learning or, Waiting for the model to learn stuff

- Personalized representation learning aims to learn latent representations that accurately capture each user's behavior, with applications in personalized response generation or recommendations.

According to current approaches in research, we can identify three different categories:

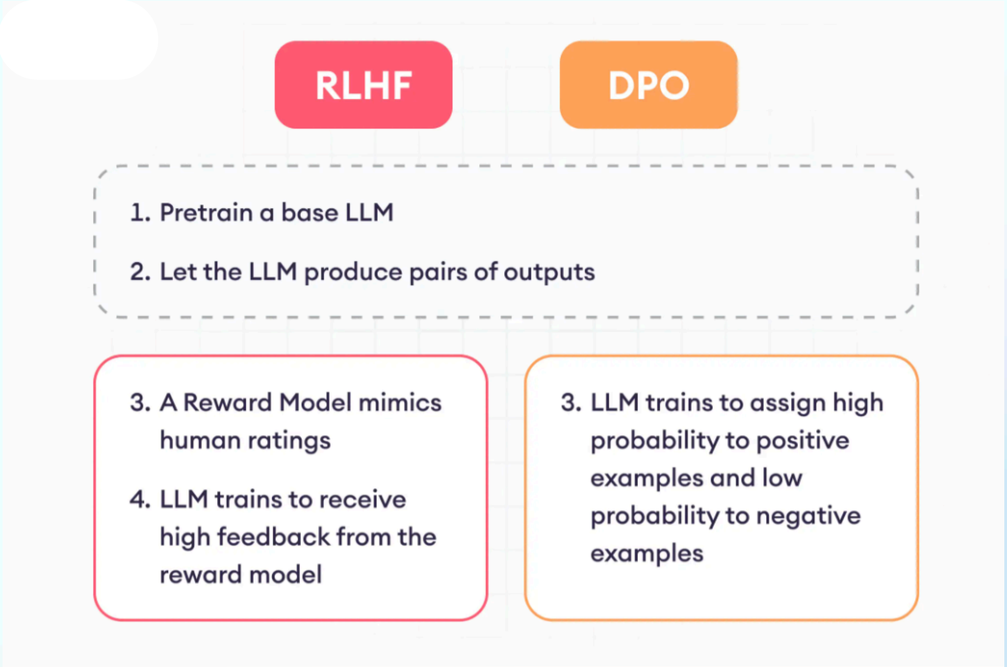
- ***Full-Parameter Fine-tuning:*** this category of methods focuses on developing training strategies and curating datasets to update all parameters of the LLM, enhancing its ability to perform downstream personalization tasks more effectively.

- ***Parameter-Efficient Fine-tuning (PEFT):*** this category of methods avoids fine-tuning all the parameters by updating only a small number of additional parameters or a subset o the pre-trained parameters to adapt the LLMs to downstream personalization tasks. A popular approach is LoRa (Low Rank Adaptation).

- ***Embedding Learning:*** this category of methods focuses on learning embeddings that represent both input text and user information in vectorized form, enabling models to more effectively incorporate personalized features and preferences into the learning process.

Personalization via RLHF, or Aligning the model to our preferences

- In general, this technique is used to align LLMs to human preferences and it has proven to be very effective in improving generative responses. As for classical RL, the model is guided towards the production of a better output following an optimization. The two major methods used for LLMs are PPO (Proximal Policy Optimization) and DPO (Direct Preference Optimization).
- In the context of Personalization, **Personalized-RLHF** approaches have been used to account for different preferences from different individuals, using personalized reward models through representation learning and clustering or ensuring diversity when recruiting people (Park et. al, 2024).



Evaluation Methods

- Metrics depend on the task in which the process of personalization is involved, but they can be broadly divided into metrics regarding an ***intrinsic evaluation*** and metrics used for an ***extrinsic evaluation.***

  - Intrinsic metrics are used when ground truth textual data is available to assess the quality of the generated text. These are often borrowed from other tasks, such as **BLUE, METEOR** (translation) or **ROUGE** (summarization), and not specifically designed for a personalization scenario.

  - Extrinsic metrics assess the quality of the personalized LLM in a downstream tasks, and come from the world of Informational Retrieval (**NDCG**), Classification (**Accuracy, F1 Score**) or both (**Recall, Precision**).

Real-world Applications (and related problems)

The Application

- ***Education***
  - Personalized LLMs can be used a as support both for teachers and students, providing tailored feedback or aligning with specific learning needs. ChatGPT or ChatGPT Edu are effective in this sense, but they're not a dedicate system adaptable to different needs.
- ***Healthcare***
  - LLMs have shown potentials in this field, acting as personal medical assistants and health helpers.
- There are other domain such as **finance, legal assistance** or even **coding** where LLMs are used in an agentic framework, but not from a personalized perspective (yet). In other domains like **Recommendation** and **Search** their usage is more robust, but not so language-oriented.

The Risks

Biases in model outputs, over-reliance, data privacy and security concerns, developing appropriate user interfaces, and ensuring fair access across languages and socioeconomic backgrounds.

Open challenges

There are more technical and practical challenges, like:

1. ***Absence of Benchmark and Metrics***

   - Existing benchmarks for personalization are largely derived from recommendation systems, where the focus is predominantly on final predictions such as ratings, recommended items, or rankings. These benchmarks often overlook the intermediate processes in LLMs' output generation, which are critical for assessing whether the output is genuinely personalized.

   - LaMP (Salemi et al., 2023) is one of the few benchmarks that specifically targets the evaluation of LLMs in generating personalized outputs, but is limited to text classification and short, single-turn text generation tasks and so it lacks the complexity of real-world interactions.

   - In addition, there is currently no comprehensive quantitative metric to assess the degree of personalization in LLM-generated outputs.

2. ***Cold-start Problem***

- The cold-start issue is a prevalent and challenging problem in recommendation systems, where the system must generate recommendations for items that have not yet been rated by any users in the dataset, or when there is minimal information available about user preferences. This empty profiles are often excluded at pre-processing time, so few works try to handle this problem. One popular approach is synthetic data generation, but it can encounter biases and stereotypes propagation.

Open challenges

And more ethical and social problems, such as:

1. ***Stereotypes & Biases***

- When LLMs generate personalized outputs, they rely on data that may inherently contain societal biases related to gender, race, ethnicity, culture, and other sensitive attributes. Personalization can unintentionally reinforce these biases by tailoring content that aligns with the biased data the models are trained on or the ones provided in the prompt, thus exacerbating the problem.

- This can lead to the deepening of **echo chambers**, where users are repeatedly exposed to biased or stereotypical information without opportunities for counterbalance. Despite growing efforts to mitigate biases in LLMs there is a limited number of works on how personalization intersects with these biases.

2. Privacy Issues

- Privacy, particularly concerning ***Personally Identifiable Information (PII),*** is a critical concern in LLM personalization applications, where the objectives of personalization and privacy often conflict. Current LLMs are vulnerable to privacy breaches, as they can accurately infer personal attributes from unstructured text, even when common mitigations such as text anonymization and model alignment are employed. Additionally, adversarial attacks, such as prompt injections and jailbreaking can cause LLMs to generate inappropriate content or reveal sensitive information from their training data.

- There is limited work specifically targeting the intersection of personalization and privacy. An ideal solution would allow for flexible adjustment, enabling a balanced trade-off between the degree of personalization and privacy protection, tailored to individual user preferences and specific application contexts.

Open challenges

The exclusion of some part of the population, especially neurodivergent people (Carik et al, 2025).

Themes & Sub-topics	Autism	SA	ADHD	Dyslexia
Challenges	**7.67%**	**5.00%**	**8.33%**	**11.00%**
– Prompting frustrations	30.43%	40.00%	52.00%	27.27%
– NT biases in LLM responses	17.39%	¡0.1%	20.00%	¡0.1%
– Lack of personal voice	43.48%	60.00%	20.00%	¡0.1%
– Text-centric interactions	8.70%	¡0.1%	8.00%	72.73%
Needs and Wants	**7.33%**	**9.00%**	**5.67%**	**18.00%**
– Multimodal interactions	9.09%	22.22%	11.76%	44.44%
– ND-friendly prompts	31.82%	¡0.1%	52.94%	11.11%
– LLM tools to support daily tasks	54.55%	77.78%	35.29%	22.22%
– Greater acceptance of LLM use by ND users	4.55%	¡0.1%	¡0.1%	22.22%
Hacks & Resources	**10.67%**	**15.00%**	**25.00%**	**29.00%**
– Prompting hacks	9.38%	33.33%	44.00%	27.59%
– LLM applications for ND users	34.38%	20.00%	22.67%	27.59%
– LLM applications built by ND users	21.88%	13.33%	13.33%	17.24%
Concerns	**3.67%**	**5.00%**	**5.00%**	**8.00%**
– False information	54.55%	20.00%	80.00%	75.00%
– Overreliance	27.27%	20.00%	20.00%	25.00%
– Replacing human connections	18.18%	60.00%	¡0.1%	¡0.1%

Table 3. Distribution of challenges and concerns related to LLM use, needs and wants, and sharing of hacks and resources, as reported by individuals with autism, social anxiety, ADHD, and dyslexia. The bold percentages represent the proportion of each thematic area within the analyzed posts, comments, and replies, while the percentages for sub-topics reflect their share within each theme.

Conclusion

- Personalized LLMs have the potential of being very effective in improving the quality of interactions in different tasks where some level of adaptiveness to user needs is necessary. But, considering their “nature”, they may encounter several problems in terms of safety, fairness and inclusivity.

- Different techniques have proven to be promising, but efforts in this direction still need to be done. Since this is an interdisciplinary field, more contributions from linguistics, computer science, psychology and many other domains are essential.

References

Zhang, Z., Rossi, R. A., Kveton, B., Shao, Y., Yang, D., Zamani, H., ... & Wang, Y. (2024). Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Carik, B., Ping, K., Ding, X., & Rho, E. H. (2025). Exploring Large Language Models Through a Neurodivergent Lens: Use, Challenges, Community-Driven Workarounds, and Concerns. *Proceedings of the ACM on Human-Computer Interaction*, *9*(1), 1-28.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097-1179.

Liang, X., Wang, H., Wang, Y., Song, S., Yang, J., Niu, S., ... & Li, Z. (2024). Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., ... & Testoni, A. (2024). Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2023). Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Park, C., Liu, M., Kong, D., Zhang, K., & Ozdaglar, A. (2024). Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*.

Thanks for your attention :)

Q&A Time