

GMTRouter: Personalized LLM Router over Multi-turn User Interactions

Yihang Sun* and Encheng Xie* and Tao Feng and Jiaxuan You

University of Illinois Urbana-Champaign

{yihangs, encheng2, taofeng2, jiaxuan}@illinois.edu[†]

Abstract

Large Language Model (LLM) routing has demonstrated strong capability in balancing response quality with computational cost. As users exhibit diverse preferences, personalization has attracted increasing attention in LLM routing, since even identical queries may require different models to generate responses tailored to individual needs. However, existing approaches are not fully personalized and often fail to faithfully capture the complex interactions between users and LLMs. Moreover, user preference data is typically scarce and inconsistent in format, which limits the effectiveness of methods that directly leverage user-specific data. To address these challenges, we propose *GMTRouter*, which represents multi-turn user-LLM interactions as a heterogeneous graph with five node types: user, LLM, query, response and turn, thereby maximally preserving the rich relational structure of the interaction. Through a lightweight inductive graph learning framework combined with a tailored user-conditioned graph sampling mechanism, *GMTRouter* learns to capture user preferences from few-shot data, enabling effective personalization. Extensive experiments demonstrate that *GMTRouter* outperforms the strongest baselines, achieving up to a 0.108 absolute improvement in accuracy and a 0.124 improvement in AUC. More importantly, we further demonstrate that *GMTRouter* can adapt to new users using only few-shot data, without extensive fine-tuning.

1 Introduction

With the rapid development of the field of Large Language Models (LLMs), an increasing number of models with varying sizes, computational costs, and domain expertise have become available (Singhal et al., 2023; Luo et al., 2022). This makes LLM

routing particularly important, as it enables the recommendation of appropriate LLMs for diverse user queries while balancing response quality with computational cost (Šakota et al., 2024; Stripelis et al., 2024). Such routing techniques are increasingly adopted in modern LLMs, including GPT-5 (OpenAI, 2025). At the same time, as more users engage with LLM routing services, differences in individual preferences become increasingly prominent: even identical queries may require different models to generate responses tailored to each user (Li et al., 2024b; Salehi et al., 2024). Therefore, this paper aims to highlight a pressing research question: *Can we design a personalized routing framework that aligns LLM selection with individual user preferences based on their interaction histories?*

Existing research has proposed various architectures for LLM routing frameworks: FrugalGPT introduces a BERT-based router that determines whether to switch to a larger LLM (Chen et al., 2024), while C2MAB-V constructs a bandit-based router to balance exploration and exploitation when selecting an LLM (Dai et al., 2024). GraphRouter formulates routing as a node classification task over a graph of queries, tasks, and LLMs (Feng et al., 2024). However, existing methods largely overlook the importance of extracting structured preference information from users’ interaction histories: they are not fully personalized and often fail to faithfully model multi-turn conversations between users and LLMs, which represent the most common form of user-LLM interaction in real-world scenarios (Zhang et al., 2025a; Li et al., 2025b). Moreover, in real-world scenarios, the preference data provided by a single user is typically scarce and inconsistent in format (Escamocher et al., 2024; Li et al., 2024a). This makes it challenging for methods that directly leverage user-specific data to learn user profiles (Salemi et al., 2024; Gao et al., 2024) or use such data as a retrieval source to support routing (Au et al., 2025), thereby limiting their effectiveness.

*Equal contribution.

[†]Accepted to Findings of the Association for Computational Linguistics: EMNLP 2026.

User ID	Query	Selected LLM	Response	Feedback
User 1	[Turn 1]"Please Explain ... ?" [Turn 2]"Can a Process ... ?"	GPT-4-1106-Preview	[Turn 1]"Exothermic and endothermic ..." [Turn 2]"Yes, a process ..."	[Turn 1]"rating: 3.0" [Turn 2]"rating: 4.5"
User 2	[Turn 1]"Compose a blog ..."	Claude-V1	[Turn 1]"Title: Aloha Spirit ..."	[Turn 1]"ranking: Claude-V1 > Koala-13B"
User 2	[Turn 1]"Compose a blog ..."	Koala-13B	[Turn 1]"Aloha, fellow travelers ! ..."	[Turn 1]"ranking: Claude-V1 > Koala-13B"
User 3	[Turn 1]"Compose an email ..." [Turn 2]"Rewrite your ..."	Vicuna-13B	[Turn 1]"Subject: An Exciting ..." [Turn 2]"Subject: A Gental ..."	[Turn 1]"response: Subject: Embrace ..." [Turn 2]"response: Subject: Soaring to ..."

Figure 1: **Multi-turn user-LLM Interaction History Table**. Each row captures a multi-turn interaction with associated user feedback. User feedback can take various forms, including ratings, rankings, ground-truth responses.

To address these challenges, we introduce **GMTRouter (Graph Multi-Turn Router)**, a heterogeneous graph-based LLM router based on multi-turn user interactions for personalized LLM routing. GMTRouter first explicitly identifies the key entities within the user-LLM interaction process: users, LLMs, queries, and responses. By modeling these entities as different types of nodes and encoding their textual information into node embeddings, it maximally preserves the semantic information from the original data. To faithfully model the relational structure of multi-turn user-LLM interactions, GMTRouter organizes these diverse node types into a heterogeneous graph that captures complex relational dependencies. Each single-turn interaction is treated as a fundamental unit, and a virtual node, referred to as a *turn node*, is introduced to aggregate local information within each interaction round. We further transform user preference feedback into node features, enabling preference information to propagate across the graph. Moreover, rather than training the model directly on large historical datasets, GMTRouter employs a novel **user-conditioned graph sampling** mechanism in conjunction with an inductive graph learning framework to **enhance the model’s ability to capture user preferences from few-shot data**. This design allows effective test-time personalization even under sparse interaction histories, such as cold-start scenarios involving new users. In summary, our main contributions are as follows:

- To our knowledge, we are among the first to introduce a personalized LLM routing task based on multi-turn user interactions, providing new insights for this rapidly growing research field.
- We propose a novel personalized LLM routing framework, which faithfully models user-LLM interactions as a heterogeneous graph, and learns to capture user preferences from few-shot data

within a lightweight inductive framework.

- Through experiments on multiple datasets spanning diverse tasks, GMTRouter outperforms the strongest baselines, achieving up to a 0.108 absolute improvement in accuracy and a 0.124 improvement in AUC. Moreover, we demonstrate that our method can efficiently adapt to unseen users with only a few interaction examples, without requiring retraining.

2 Preliminaries

2.1 Task Formulation

We introduce the personalized LLM routing task in this section. We focus on the multi-turn interaction scenario between users and LLMs with feedback (Wang et al., 2024b; Shi et al., 2026). Within a dialogue session, a user u repeatedly interacts with an LLM m : in each turn, the user issues a query q , the LLM provides a response r , and the user in turn supplies a piece of feedback f . Such feedback can take multiple forms, including: (1) scalar scores (e.g., numerical ratings), (Wang et al., 2024d,c); (2) preference rankings (e.g., choosing among multiple responses), (Yang et al., 2024; Sun et al., 2025); (3) ground-truth responses (e.g., directly providing the correct answer) (Gao et al., 2024; Salemi et al., 2024). We structure these interactions into an **Interaction History Table**, illustrated in Figure 1, where each entry records the user ID, the selected LLM, the multi-turn queries and generated responses, and the corresponding user feedback, thereby maximally preserving the rich relational information of the interaction.

Our personalized LLM routing task is then modeled as follows: Given m users $\{u_1, \dots, u_m\}$ and n LLM candidates $\{m_1, \dots, m_n\}$, as well as their historical interaction records:

$$\mathcal{H} = \{(u_i, m_i, \{(q^{(t)}, r^{(t)}, f^{(t)})\}_{t=1}^{T_i})\},$$

where u_i is the user, m_i is the selected LLM,

Table 1: **The consistency of LLM preferences between users is substantially lower than the within-user consistency.** The self-consistency score is substantially higher than the spearman scores computed across different users.

Metric	Self	Global	Intra-cluster	Inter-cluster
Spearman	0.793	0.524	0.573	0.442
Percent	100%	66.0%	72.3%	55.7%

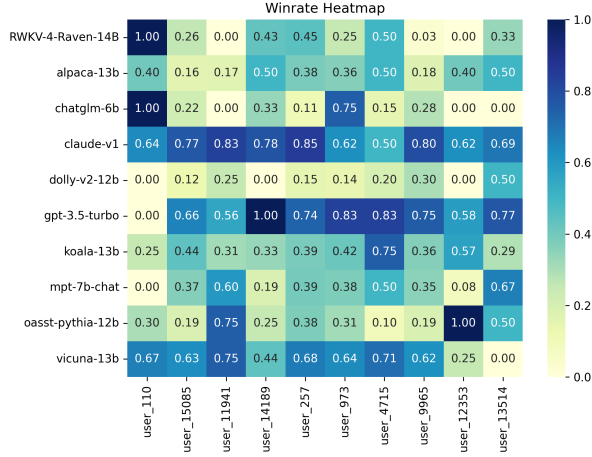


Figure 2: **Significant differences exist in LLM preferences across users.** The figure shows a heatmap of win rates for the 10 most popular LLMs across 10 active users in ChatBot Arena. Variation in win-rate patterns across rows illustrates that different users rank the same LLMs differently.

and each record contains a multi-turn sequence of queries $q^{(t)}$, responses $r^{(t)}$, and feedback $f^{(t)}$ for $t = 1, \dots, T_i$. When a user u raises a new query q , the router is required to select an LLM $m \in \{m_1, \dots, m_n\}$ to generate a response r that best aligns with the user preferences, which is measured through the feedback f provided by the user.

2.2 Motivation

In this section, we highlight the significant differences in LLM preferences across users in the real world (Chevi et al., 2025; Wang et al., 2024a), emphasizing the importance of personalized LLM routing for enhancing user experience. We use the ChatBot Arena dataset (Chiang et al., 2024) to illustrate our findings, which contains extensive anonymized multi-turn conversations from numerous users, with pairwise human preference labels between various LLMs, enabling the study of real-world user-LLM interactions. From this dataset, we select 10 active users, each with at least 50 records and 1,480 records in total, for detailed analysis. For each user, we randomly split their data into two halves and compute the win rates of each

LLM within each half. We use Spearman correlation to quantify the consistency of preference rankings over LLMs (de Winter et al., 2016; Hauke and Kossowski, 2011). We then compute the Spearman correlation between the two halves to quantify their self-consistency in preferences over LLMs (Chevi et al., 2025; Jiang et al., 2025), reporting the average as a baseline for comparison with inter-user preference consistency. Next, based on the similarity of queries in each user’s interaction history, we cluster users into three groups (Zeng et al., 2024; Li et al., 2025a), and compute pairwise Spearman correlation scores among users globally, within clusters, and across clusters (Cavallo, 2020; de Winter et al., 2016), reporting the corresponding averages as summarized in Table 1. We observe that global consistency in LLM preferences among users is substantially lower than individual self-consistency, reaching only 65.99% of the latter. Even within the same cluster, the Spearman score is only 72.28% of the self-consistency, highlighting the diversity of user preferences toward LLMs (Sun et al., 2025; Salemi et al., 2024). To further visualize these differences, we select the 10 most frequently used models across these 10 users and present a win-rate heatmap in Figure 2, offering an intuitive depiction of the variability in user preferences. Therefore, to address the substantial inconsistency of LLM preferences across users, we propose **GMTRouter**, a framework that enables the personalized recommendation of suitable LLMs tailored to each user’s individual preferences.

3 GMTRouter: Router Over Multi-turn User Interactions

Method Overview As shown in Figure 3, GMTRouter operates in three stages: (a) *Node Embedding Initialization*: it extracts users, LLMs, queries, and responses from the Interaction History Table. Query, response, and LLM nodes are initialized with PLM embeddings, while user feedback is encoded as a preference feature on response nodes. (b) *Heterogeneous Graph Construction*: it further introduces turn nodes and organizes the five node types into a heterogeneous graph that captures the relational structure of multi-turn user-LLM interactions. (c) *Inductive GNN Training*: Finally, we employ a user-conditioned graph sampling mechanism and an inductive graph training framework to learn user preferences from few-shot data, enabling effective personalization under sparse interaction

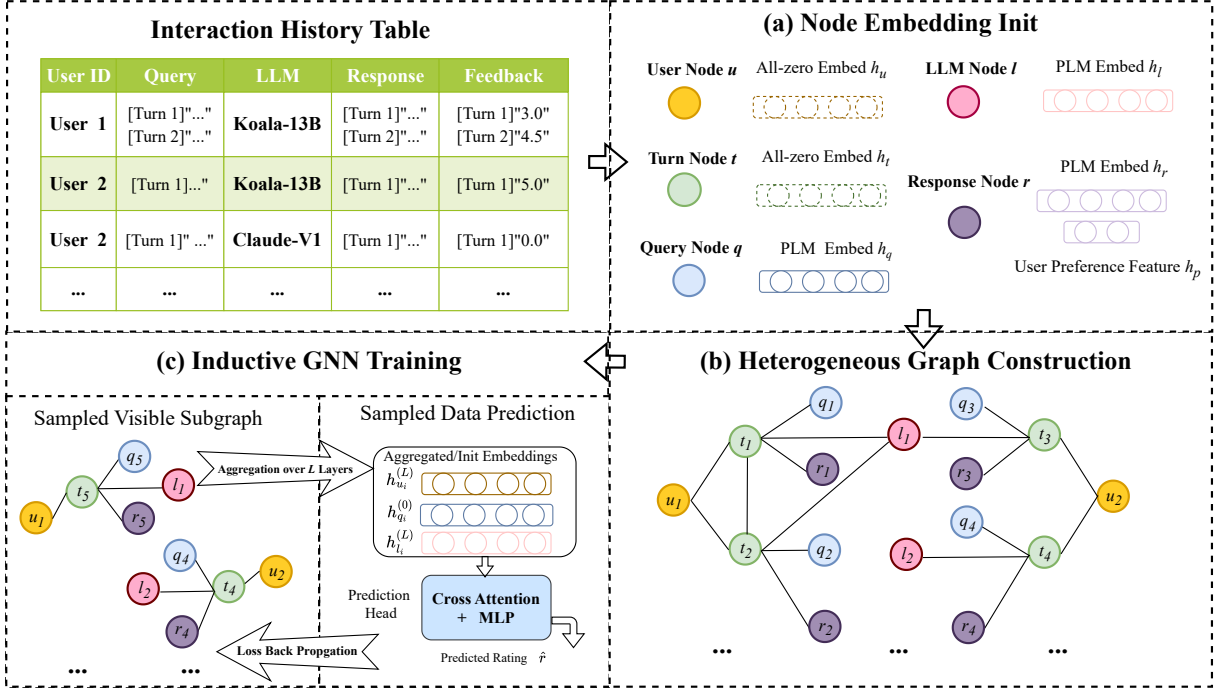


Figure 3: **Overview of GMTRouter.** (a) GMTRouter extracts users, LLMs, queries, and responses from the Interaction History Table. Query, response, and LLM nodes are initialized with PLM embeddings, while user feedback is encoded as a preference feature on response nodes. (b) It further introduces turn nodes and organizes these entities into a heterogeneous graph to model the relational structure of user–LLM interactions. (c) Within a lightweight inductive framework, GMTRouter learns to infer user preferences from few-shot interaction histories.

histories.

3.1 Node Embedding Initialization.

First, our framework focuses on comprehensively extracting the information of various entities involved in the user–LLM interaction process from the Interaction History Table, along with their relational structures. As illustrated in part (a) of Figure 3, we first extract four types of observed entities: user u , LLM m , query q , and response r , and formalize them as four corresponding node types. For LLM, query, and response nodes, we use a pretrained language model (PLM) to obtain their initial node embeddings (Wang et al., 2022, 2023), thereby preserving the semantic information from the original data. Specifically, we encode the query and response texts as their initial embeddings, denoted as h_q and h_r . In addition, we transform various forms of user feedback into numerical ratings and project them into a User Preference Feature h_p , which serves as another attribute on the response nodes. Concretely, ranking feedback is discretized into numerical ratings to ensure that higher-ranked responses receive higher scores (Banditwattana-wong and Masdisornchote, 2025); for ground-truth response feedback, we compute the geometric distance between the embeddings of the ground-truth

and the generated response as the rating criterion (Salemi et al., 2024). For LLM nodes, instead of simply using their names or IDs (Ding et al., 2024; Chen et al., 2024), we encode the model overviews provided by AI/ML API platforms as their node embeddings h_m , which typically include key information such as model size, usage cost, and domain-specific capabilities, thereby enriching the node embeddings with important background knowledge. Finally, for user nodes, we do not assume the existence of text-based user profiles, as such information is often scarce and noisy in real-world applications (Su et al., 2024; Alzubaidi et al., 2023); therefore, we initialize user embeddings h_u as zero vectors.

3.2 Heterogeneous Graph Construction.

Next, we organize these nodes into a heterogeneous graph to model the relational structure of user–LLM interactions (Zhang et al., 2025b; Schlichtkrull et al., 2018). We consider each single-round user–LLM interaction as a fundamental unit and introduce a kind of virtual node, the *turn node*, to aggregate the information within each interaction round. As illustrated in part (b) of Figure 3, within each interaction round, the associated user, LLM, query, and response nodes are connected to

a corresponding turn node that aggregates information from that round. For multi-turn conversations, the turn nodes corresponding to each round are sequentially connected in dialogue order, facilitating information propagation across turns. The turn node embedding h_t is initialized as a zero vector. The resulting heterogeneous graph captures the rich relational dependencies inherent in user–LLM interactions, where turn nodes aggregate local information within each dialogue round and propagate it to user nodes, thereby facilitating the global aggregation of user preference information.

3.3 Inductive GNN Training

After constructing the user–LLM interaction histories into a heterogeneous graph, we train our GNN model on it. Notably, GMTRouter is a **general framework** that can incorporate any heterogeneous GNN as its backbone. We denote the GNN backbone used in our method as *GNN* throughout the rest of the paper. To address scenarios with sparse user history (Su et al., 2024), instead of training the model directly on large amounts of historical data (Lin et al., 2023; Wang et al., 2025b), we employ an inductive framework along with **user-conditioned graph sampling** during training, enabling GMTRouter to **capture a user’s preferences from only a few interaction records**.

User-conditioned Graph Sampling As illustrated in the left of (c) in Figure 3, during each training epoch, we sample k interaction histories, where each history corresponds to one user–query instance with candidate LLM responses and feedback, for each user to construct a visible subgraph $\mathcal{G}_{\text{sub}}$ from the heterogeneous graph for message passing, and further sample data outside the visible subgraph as the prediction targets. We then use only these small sampled visible subgraphs and perform message aggregation separately for each user to update the node embeddings. Formally, at layer l , node v ’s embedding $h_v^{(l)}$ is updated by aggregating messages from its neighbors via the backbone GNN’s message-passing mechanism:

$$h_v^{(l)} = \text{Norm} \left(\text{Dropout} \left(\text{GNN}^{(l)}(h_v^{(l-1)}, \mathcal{G}_{\text{sub}}) \right) \right) \quad (1)$$

This restriction on the amount of data involved in message passing encourages the model to learn how to infer user preferences from very limited signals and to generalize efficiently to new users.

LLM Routing with a Prediction Head. After completing L layers of message aggregation, we obtain the updated node representations $h^{(L)}$. We then employ a **Prediction Head** module f_{pred} for preference prediction. As illustrated in the right of (c) in Figure 3, the Prediction Head takes the updated user embedding $h_u^{(L)}$, LLM embedding $h_m^{(L)}$, and the query embedding $h_q^{(0)}$ from PLM as input. It applies a cross-attention module, where the LLM embedding attends to the fused user–query context to extract relevant preference signals. The module outputs a scalar score $s_{u,q,m}$ for each LLM candidate, representing the likelihood that user u would prefer m to answer query:

$$s_{u,q,m} = f_{\text{pred}}(h_u^{(L)}, h_q^{(0)}, h_m^{(L)}) \quad (2)$$

These scores are then used to rank LLM candidates under the same (u, q) condition. We normalize both the predicted scores and the ground-truth ratings, and apply a criterion function to compute the training loss, which is subsequently used to update the model parameters.

During inference, when a user raises a new query, we first sample k interaction histories of that user to construct the visible subgraph and update the node embeddings. Then, the LLM candidate is selected from the candidate set $\mathcal{M}$ as the one with the highest predicted score:

$$m^* = \arg \max_{m \in \mathcal{M}} f_{\text{pred}}(h_u^{(L)}, h_q^{(0)}, h_m^{(L)}) \quad (3)$$

4 Experiment Setup

4.1 Datasets and data processing

We evaluate our approach on five datasets covering diverse tasks, spanning both real-world and synthetic settings. The two real-world datasets are: (1) **HUMAINE** (Petrova et al., 2026), which provides controlled, demographically stratified multi-turn interactions with direct human preference judgments, and (2) **ChatBot Arena** (Chiang et al., 2024), which captures large-scale, in-the-wild multi-turn interactions and crowdsourced user preferences. The remaining three are synthetic datasets: (3) **MT-Bench** (Zheng et al., 2023), assessing multi-turn reasoning and conversational ability; (4) **GSM8K** (Cobbe et al., 2021), focused on grade school math problem solving; and (5) **MMLU** (Hendrycks et al., 2021a,b), measuring general knowledge and multi-domain reasoning. Detailed descriptions of the datasets and preprocessing are provided in Appendix B.1.

Data Processing For HUMAINE and ChatBot Arena, we convert the pairwise human preference outcomes into ratings for the corresponding responses. For HUMAINE, persistent participant IDs are unavailable, so following its focus on demographic preference heterogeneity, we group records with the same demographic profile into a preference cohort and treat each cohort as one “user”.

For the three synthetic datasets, we adopt the data collected in Ong et al. 2025, which generated responses to all questions using “GPT-4-1106-preview” (Achiam et al., 2023) and “Mixtral-8x7B-Instruct-v0.1” (Jiang et al., 2024). Based on this, we convert these datasets into multi-user personalized datasets. Specifically, for each response, we consider the following four dimensions: (a) Quality: For open-ended questions, we use the GPT-4 scores provided by Ong et al., 2025; for objective questions, we directly evaluate the correctness. (b) Cost: We calculate the cost of generating each response based on the API pricing provided by the AI/ML API platform. (c) Response Length: We compute the token length of each response using the Contriever tokenizer (Izacard et al., 2022). (d) Rare Words: We count the number of rare words in each response using the *wordfreq* package (Speer, 2022). We obtain the final rating of a response by computing a weighted sum of these four metrics. Different users are assigned different weightings to simulate diverse preferences over response quality, cost sensitivity, verbosity, and language complexity (Feng et al., 2024, 2025). The specific weights used are provided in Appendix B.5.

Data Splitting For all datasets, we partition the data into training, validation, and test sets with a 7:1:2 ratio. For the new-user evaluation, we further adopt an additional splitting strategy: we sample 30% of the users and restrict their data to the validation and test sets only, in order to evaluate the generalization ability of our method to new users unseen during training.

4.2 Baselines

We compare our GMTRouter against the following baselines:

Prompt-based: (1) Vanilla LLM. We incorporate the query and the descriptions of candidate LLMs into the prompt, and feed it into LLaMA-3.1-70B (Grattafiori et al., 2024) to select the LLM. **(2) Personalized LLM.** Building on the vanilla LLM, we retrieve the ten most relevant interac-

tion histories from the user’s training set using semantic query similarity and incorporate them into the prompt. Leveraging in-context learning (Dong et al., 2024), the LLM is then guided to perform personalized routing.

Representative Router: (3) GraphRouter. (Feng et al., 2024) A graph-based model that formulates routing as a node classification task over a graph of queries, tasks, and LLMs with learned edge interactions. **(4) FrugalGPT** (Chen et al., 2024) utilizes a PLM to predict the score of the generation result of all LLMs given a query, and then selects the LLM with the highest score within a given cost. **(5) RouteLLM** (Ong et al., 2025). Learns to route queries among a weak-strong pair of LLMs. Following the official setup, we designate the weak model as the one with the lower average win rate in the dataset, and the strong model as the one with the higher win rate.

Memory-based: (6) MA-GNN (Ma et al., 2020). A memory-augmented GNN that models both *short-term* and *long-term* user interests through item-level message passing and a dedicated memory module. **(7) TIGER** (Rajput et al., 2023). A generative retrieval-based sequential recommender that models item sequences via semantic discrete codes.

4.3 Metrics

We evaluate the performance of all methods using two metrics: **(1) Accuracy** measures how often the model correctly identifies the most preferred LLM to answer a given query from a specific user. **(2) AUC-ROC** (Area Under the Receiver Operating Characteristic Curve) measures the model’s ability to correctly rank candidate LLMs according to user preferences.

4.4 Implementation Details

We employ Contriever (Izacard et al., 2022) as the PLM to obtain the initial node embeddings. We adopt the HGT (Hu et al., 2020) as our model backbone due to its strong capability to maintain dedicated representations for different types of nodes. Additionally, we experiment with various other GNNs as the backbone to investigate their impact on GMTRouter’s performance. Experimental details are provided in Appendix E.5. We set the visible data size per user to $k = 10$ during both training and inference and adopt Entropy Loss as our loss function. In Appendix C, we will experimentally analyze the impact of different values of

Table 2: **GMTRouter consistently outperforms baselines across all datasets.** Bold and underline denote the best and second-best results. The results are averaged over multiple runs.

Method	HUMAINE		Chatbot-Arena		MT-Bench		GSM8K		MMLU	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Vanilla LLM	0.526	0.544	0.525	0.741	0.481	0.457	0.546	0.533	0.473	0.475
Personalized LLM	0.514	0.548	0.646	0.780	0.437	0.491	0.553	0.536	0.675	0.678
GraphRouter	0.669	0.712	0.771	<u>0.869</u>	0.568	0.550	0.717	0.792	0.699	0.746
FrugalGPT	0.657	<u>0.719</u>	0.562	0.622	0.551	0.552	0.504	0.515	0.545	0.575
RouteLLM	0.641	0.681	0.492	0.485	0.480	0.475	0.499	0.498	0.532	0.513
MA-GNN	<u>0.670</u>	0.717	0.673	0.775	0.679	0.739	0.636	0.648	0.702	0.758
TIGER	0.571	0.567	0.739	0.735	0.656	0.691	0.639	0.683	0.710	0.764
Ours	0.778	0.843	<u>0.774</u>	0.875	0.784	0.859	0.773	0.859	0.771	0.870
Ours (new user)	0.658	0.707	0.780	0.858	<u>0.759</u>	<u>0.824</u>	<u>0.756</u>	<u>0.833</u>	<u>0.751</u>	<u>0.831</u>

k on our method, and hyperparameter details are provided in Appendix A.1.

5 Experiment Results

5.1 Comparison with Baselines

We compare GMTRouter with multiple baselines across five datasets in Table 2. We observe that GMTRouter outperforms the strongest baseline on **all evaluation metrics**, achieving up to **0.108 absolute improvement in accuracy and 0.124 in AUC**, which demonstrates the effectiveness of our framework. For prompt-based baselines, although incorporating user interaction histories into prompts can improve performance over the Vanilla LLM, these methods still fall significantly behind GMTRouter. This result highlights the limited ability of LLMs to directly extract user preference patterns from raw interaction data. Moreover, our approach consistently outperforms router-based baselines, including GraphRouter, which has shown strong performance in non-personalized LLM routing tasks. These results validate the importance of leveraging structured information from user-LLM interaction data, together with explicit user preference signals, to better align model selection with diverse user needs. Compared to memory-based methods, GMTRouter benefits from learning over a fine-grained and highly structured heterogeneous graph, leading to superior performance. Furthermore, even when **30% of users are absent from the training set**, GMTRouter remains competitive across datasets, demonstrating its ability to generalize to previously unseen users.

We find that the improvement on ChatBot Arena is relatively modest compared with HUMAINE and the synthetic datasets. ChatBot Arena provides valuable in-the-wild interactions, but its

open crowdsourcing setting offers limited control over participant recruitment and interaction quality. Consequently, some comparisons may contain noisy or low-information preference signals, making stable personalized patterns harder to capture. In contrast, HUMAINE recruits compensated participants and applies real-time quality monitoring. Together with its broader candidate-model coverage and more controlled data collection protocol, this provides a more controlled setting for personalized routing, where GMTRouter exhibits a much clearer advantage. We provide further discussion of the selected real-world datasets in Appendix B.6.

Our Framework is Lightweight With only 27.4M trainable parameters and a 109.6MB model size, our framework remains compact compared to existing routing models. During training, only 4.3GB of GPU memory is needed, making it feasible to train on a single modern GPU without specialized hardware. Additional efficiency comparisons are provided in Appendix E.4.

5.2 Generalization to New Users

We further investigate the personalized capability of our method in few-shot scenarios with new users. Specifically, we evaluate on the GSM8K and MMLU by sampling 30% users from each dataset and varying the number of visible data k . Figure 4 presents averaged results of the sampled users under two settings: (i) the old user setting, where their records are included in the training set, and (ii) the new user setting, where they appear only in the validation and test sets. We observe that new users achieve results comparable to old users, and their performance curves consistently peak far above the GraphRouter baseline. These findings demonstrate that **our approach effectively learns**

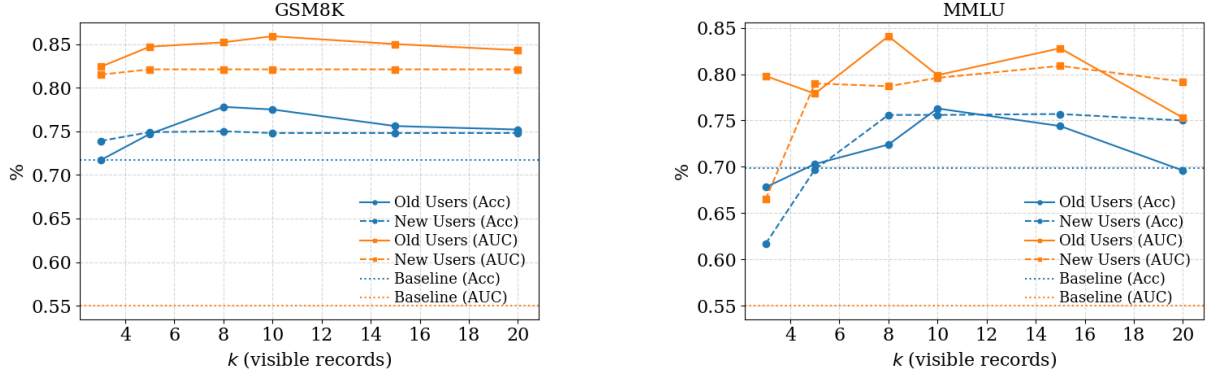


Figure 4: **Result comparison between old-user and new-user settings for GSM8K (left) and MMLU (right).** The dashed line represents the GraphRouter baseline. The personalized performance under the new-user setting is comparable to that under the old-user setting, highlighting the strong generalization capability of our method.

Table 3: **Ablation of design components.** We compare the full model with three variants: (1) removing the user preference features, (2) replacing the prediction head with a dot-product, (3) not using user embeddings during prediction. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	Chatbot-Arena		MT-Bench		GSM8K		MMLU	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
w/o h_p	0.768	0.872	0.569	0.507	0.715	0.784	0.494	0.613
Dot-product	0.777	0.868	<u>0.730</u>	<u>0.795</u>	0.629	0.724	0.681	0.746
w/o h_u	0.771	<u>0.873</u>	0.569	0.631	<u>0.725</u>	<u>0.814</u>	<u>0.701</u>	<u>0.771</u>
GMTRouter	<u>0.774</u>	0.875	0.784	0.859	0.773	0.859	0.771	0.870

to capture user preferences from few-shot data and can adapt to new users without requiring extensive fine-tuning.

5.3 Ablation Studies

To evaluate the effectiveness of each design component of the GMTRouter, we conduct ablation studies along the following aspects.

- **w/o User Preference Feature** To verify the effectiveness of the user preference feature in propagating preference signals during GNN aggregation, we remove this feature in this variant. As a result, node embeddings are updated without incorporating preference ratings.
- **Dot-product Prediction Head** To evaluate whether the cross-attention prediction head captures non-linear interactions more effectively than standard similarity scoring when predicting the optimal model, we replace it in this variant with a simple dot product between the (user + query) and LLM embeddings.
- **w/o User Embedding** To evaluate the effectiveness of user embeddings aggregated from the sampled visible graph for personalized prediction, we replace the user embeddings fed into the prediction head with zero vectors, thereby ablating their influence on the predictions.

The results of our ablation studies are presented

in Table 3. As shown, our GMTRouter achieves the best performance on most metrics compared to the other variants, confirming the effectiveness of our design choices. Additional ablations on multi-turn modeling and turn nodes are provided in Appendix E.2.

5.4 Further Experiments

We present additional experiments in Appendix E to further examine the robustness, design choices, generality, and efficiency of GMTRouter. Specifically, we evaluate GMTRouter under more human-like preference factors (Appendix E.1), analyze the effects of multi-turn modeling and turn nodes (Appendix E.2), compare against the memory-based MemoryBank baseline (Appendix E.3), and report its computational efficiency (Appendix E.4). We further study the impact of different GNN backbones (Appendix E.5) and PLM-based node initialization (Appendix E.6).

6 Additional Related Works

LLM Routing. LLM routing focuses on enhancing inference efficiency and response quality by assigning queries to the most appropriate model (Yue et al., 2025; Wang et al., 2025a). Recent work frames routing as learning with cost-quality trade-offs (Kadavath et al., 2022; Dekoninck et al., 2025):

RouteLLM learns from preference data (Ong et al., 2025), and RouterBench offers standardized routing benchmarks (Hu et al., 2024). BEST-Route jointly selects LLM and generation count at test-time via a bandit controller (Ding et al., 2025). However, existing approaches are not fully personalized and fail to exploit user information from interaction histories as well as the structure of multi-turn dialogues.

Heterogeneous Graph Learning. HetGNNs are designed to model heterogeneous graphs by capturing complex multi-type interactions among various nodes and edges (Chien et al., 2022; Feng et al., 2019). HAN uses hierarchical attention over metapaths (Wang et al., 2019), while MAGNN and HeCo improve metapath aggregation and cross-view contrast (Fu et al., 2020; Wang et al., 2021). Transformers such as HGT provide inductive, relation-aware message passing with temporal encoding (Hu et al., 2020). This enables rich relational structures in user-LLM interactions while leveraging inductive training to enhance generalization on sparse data from new users.

7 Conclusion

In this work, we introduced GMTRouter, a heterogeneous graph-based framework for personalized LLM routing. By modeling multi-turn user-LLM interactions as a heterogeneous graph and propagating preference signals across node types, our method effectively captures user-specific patterns even from few-shot data. Experiments across five benchmarks confirm that GMTRouter consistently surpasses strong baselines, while adapting efficiently to new users without retraining. These results highlight the value of structured interaction modeling for advancing preference-aware LLM routing and point to promising future directions in scalable, user-aligned LLM deployment.

8 Limitations

While GMTRouter demonstrates significant improvements in personalized routing, we identify a few limitations that offer avenues for future research. First, the initial embeddings for LLM nodes are currently derived from static descriptions provided by API platforms. While effective, these representations may require manual updates if a model undergoes significant versioning or performance shifts. Second, although we evaluated our framework across five diverse benchmarks spanning reasoning, math, and general knowledge, its performance in highly specialized niche domains, such as advanced medical or legal sub-specialties, remains to be further explored. Finally, the current implementation focuses on individual user personalization; extending the framework to model collective or group-level preferences in collaborative environments represents a promising direction for future work.

9 Ethical Considerations

Our experiments rely on publicly released datasets and do not involve new human-subject data collection. HUMAINE includes demographic attributes, which we use only to construct aggregate preference cohorts; we do not attempt to identify individual participants. Personalized routing may nevertheless inherit or amplify biases present in user preference data and candidate LLMs, and demographic cohorting may obscure preference heterogeneity within each group. In practical deployment, user interaction histories may also contain sensitive information. Appropriate consent, data minimization, access control, and privacy protections should therefore be considered when applying personalized routing systems to real users.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2023. [GPT-4 Technical Report](#). *arXiv preprint arXiv:2303.08774*.
- Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, José Santamaría, Ahmed Shihab Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A. Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H. Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin M. Abbosh, and Yuantong Gu. 2023. [A Survey on Deep Learning Tools Dealing with Data Scarcity: Definitions, Challenges, Solutions, Tips, and Applications](#). *Journal of Big Data*, 10(1):46.

- Steven Au, Cameron J. Dimacali, Ojasmitha Pedirappagari, Namyong Park, Franck Dernoncourt, Yu Wang, Nikos Kanakaris, Hanieh Deilamsalehy, Ryan A. Rossi, and Nesreen K. Ahmed. 2025. [Personalized Graph-Based Retrieval for Large Language Models](#). *arXiv preprint arXiv:2501.02157*.
- Thepparit Banditwattanawong and Masawee Masdisornchote. 2025. [Unbiased Machine Learning-Assisted Approach for Conditional Discretization of Human Performances](#). *PeerJ Computer Science*, 11:e2804.
- Bice Cavallo. 2020. [Functional Relations and Spearman Correlation between Consistency Indices](#). *Journal of the Operational Research Society*, 71(2):301–311.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2024. [FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance](#). *Transactions on Machine Learning Research*.
- Rendi Chevi, Kentaro Inui, Tamar Solorio, and Alham Fikri Aji. 2025. [How Individual Traits and Language Styles Shape Preferences in Open-Ended User-LLM Interaction: A Preliminary Study](#). *arXiv preprint arXiv:2504.17083*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.
- Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. 2022. [You Are AllSet: A Multiset Function Framework for Hypergraph Neural Networks](#). In *International Conference on Learning Representations*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training Verifiers to Solve Math Word Problems](#). *arXiv preprint arXiv:2110.14168*.
- Xiangxiang Dai, Jin Li, Xutong Liu, Anqi Yu, and John C. S. Lui. 2024. [Cost-Effective Online Multi-LLM Selection with Versatile Reward Models](#). *arXiv preprint arXiv:2405.16587*.
- Joost C. F. de Winter, Samuel D. Gosling, and Jeff Potter. 2016. [Comparing the Pearson and Spearman Correlation Coefficients across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data](#). *Psychological Methods*, 21(3):273–290.
- Jasper Dekoninck, Maximilian Baader, and Martin Vechev. 2025. [A Unified Approach to Routing and Cascading for LLMs](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 12987–13010. PMLR.
- Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. 2024. [Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing](#). In *International Conference on Learning Representations*.
- Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Rühle. 2025. [BEST-Route: Adaptive LLM Routing with Test-Time Optimal Compute](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13870–13884. PMLR.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. [A Survey on In-Context Learning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128. Association for Computational Linguistics.
- Guillaume Escamocher, Samira Pourkhajouei, Federico Toffano, Paolo Viappiani, and Nic Wilson. 2024. [Interactive Preference Elicitation under Noisy Preference Models: An Efficient Non-Bayesian Approach](#). *International Journal of Approximate Reasoning*, 178:109333.
- Tao Feng, Yanzhen Shen, and Jiaxuan You. 2024. [GraphRouter: A Graph-Based Router for LLM Selections](#). *arXiv preprint arXiv:2410.03834*.
- Tao Feng, Haozhen Zhang, Zijie Lei, Pengrui Han, Mostofa Patwary, Mohammad Shoeby, Bryan Catanzaro, and Jiaxuan You. 2025. [Fusing LLM Capabilities with Routing Data](#). *arXiv preprint arXiv:2507.10540*.
- Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. [Hypergraph Neural Networks](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3558–3565.
- Matthias Fey and Jan E. Lenssen. 2019. [Fast Graph Representation Learning with PyTorch Geometric](#). In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. [MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding](#). In *Proceedings of The Web Conference 2020*, pages 2331–2341.
- Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. [Aligning LLM Agents by Learning Latent Preference from User Edits](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint arXiv:2407.21783*.

- William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30, pages 1024–1034.
- Jan Hauke and Tomasz Kossowski. 2011. [Comparison of Values of Pearson’s and Spearman’s Correlation Coefficients on the Same Sets of Data](#). *Quaestiones Geographicae*, 30(2):87–93.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. [Aligning AI with Shared Human Values](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring Massive Multitask Language Understanding](#). In *International Conference on Learning Representations*.
- Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. [Router-Bench: A Benchmark for Multi-LLM Routing System](#). *arXiv preprint arXiv:2403.12031*.
- Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. [Heterogeneous Graph Transformer](#). In *Proceedings of The Web Conference 2020*, pages 2704–2710.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised Dense Information Retrieval with Contrastive Learning](#). *Transactions on Machine Learning Research*.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. [Mixtral of Experts](#). *arXiv preprint arXiv:2401.04088*.
- Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. [Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale](#). In *Conference on Language Modeling (COLM)*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, and 1 others. 2022. [Language Models \(Mostly\) Know What They Know](#). *arXiv preprint arXiv:2207.05221*.
- Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2016. [On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima](#). *arXiv preprint arXiv:1609.04836*.
- Haoxuan Li, Chunyuan Zheng, Wenjie Wang, Hao Wang, Fuli Feng, and Xiao-Hua Zhou. 2024a. [De-biased Recommendation with Noisy Feedback](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1576–1586.
- Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Leqi Liu. 2024b. [Personalized Language Modeling from Personalized Human Feedback](#). *arXiv preprint arXiv:2402.05133*.
- Ying Li, Ye Zhong, Lijuan Yang, Yanbo Wang, and Penghua Zhu. 2025a. [LLM-Guided Crowdsourced Test Report Clustering](#). *IEEE Access*, 13:24894–24904.
- Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025b. [Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models](#). *arXiv preprint arXiv:2504.04717*.
- Weiwei Lin, Haojun Xu, Jianzhuo Li, Ziming Wu, Zhengyang Hu, Victor Chang, and James Z. Wang. 2023. [Deep-Profiling: A Deep Neural Network Model for Scholarly Web User Profiling](#). *Cluster Computing*, 26(3):1753–1766.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. [BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining](#). *Briefings in Bioinformatics*, 23(6):bbac409.
- Chen Ma, Liheng Ma, Yingxue Zhang, Jianing Sun, Xue Liu, and Mark Coates. 2020. [Memory Augmented Graph Neural Networks for Sequential Recommendation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5045–5052.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. [RouteLLM: Learning to Route LLMs from Preference Data](#). In *International Conference on Learning Representations*.
- OpenAI. 2025. [GPT-5 System Card](#). Technical report, OpenAI.
- O. K. Oyedotun, Konstantinos Papadopoulos, and Djamila Aouada. 2023. [A New Perspective for Understanding Generalization Gap of Deep Neural Networks Trained with Large Batch Sizes](#). *Applied Intelligence*, 53(12):15621–15637.
- Nora Petrova, Andrew Gordon, and Enzo Blindow. 2026. [Unpacking Human Preference for LLMs: Demographically Aware Evaluation with the HUMAINE Framework](#). In *International Conference on Learning Representations*.
- Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy.

2023. [Recommender Systems with Generative Retrieval](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10299–10315.
- Marija Šakota, Maxime Peyrard, and Robert West. 2024. [Fly-Swat or Cannon? Cost-Effective Language Model Choice via Meta-Modeling](#). In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 606–615.
- Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. 2024. [ViPer: Visual Personalization of Generative Models via Individual Preference Learning](#). In *Computer Vision – ECCV 2024*, pages 391–406. Springer.
- Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. [Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 752–762.
- Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling Relational Data with Graph Convolutional Networks](#). In *The Semantic Web – ESWC 2018*, pages 593–607. Springer.
- Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Kumar Jauhar, Sihao Chen, Shan Xia, Hongfei Zhang, Jieyu Zhao, Xiaofeng Xu, Xia Song, and Jennifer Neville. 2026. [WildFeedback: Aligning LLMs With In-situ User Interactions And Feedback](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36701–36725. Association for Computational Linguistics.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, and 1 others. 2023. [Large Language Models Encode Clinical Knowledge](#). *Nature*, 620(7972):172–180.
- Robyn Speer. 2022. [wordfreq: v3.0.2](#). Zenodo.
- Dimitris Stripelis, Zijian Hu, Jipeng Zhang, Zhaozhuo Xu, Alay Shah, Han Jin, Yuhang Yao, Salman Avestimehr, and Chaoyang He. 2024. [PolyRouter: A Multi-LLM Querying System](#). *arXiv preprint arXiv:2408.12320*.
- Hongzu Su, Jingjing Li, Zhekai Du, Lei Zhu, Ke Lu, and Heng Tao Shen. 2024. [Cross-Domain Recommendation via Dual Adversarial Adaptation](#). *ACM Transactions on Information Systems*, 42(3):83:1–83:26.
- Yihang Sun, Tao Feng, Ge Liu, and Jiaxuan You. 2025. [PREMIUM: LLM Personalization with Individual-Level Preference Feedback](#). Preprint.
- Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2023. [Pre-Trained Language Models and Their Applications](#). *Engineering*, 25:51–65.
- Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. [Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text Embeddings by Weakly-Supervised Contrastive Pre-Training](#). *arXiv preprint arXiv:2212.03533*.
- Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, Philip S. Yu, and Yanfang Ye. 2019. [Heterogeneous Graph Attention Network](#). In *Proceedings of The Web Conference 2019*, pages 2022–2032.
- Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. 2021. [Self-Supervised Heterogeneous Graph Neural Network with Co-Contrastive Learning](#). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1726–1736.
- Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024b. [MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback](#). In *International Conference on Learning Representations*.
- Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. 2025a. [MixLLM: Dynamic Routing in Mixed Large Language Models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10912–10922. Association for Computational Linguistics.
- Zhaoyang Wang, Li Li, Ketai He, and Zhenyang Zhu. 2025b. [User Profile Construction Based on High-Dimensional Features Extracted by Stacking Ensemble Learning](#). *Applied Sciences*, 15(3):1224.
- Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024c. [HelpSteer 2: Open-Source Dataset for Training Top-Performing Reward Models](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2024d. [HelpSteer: Multi-Attribute Helpfulness Dataset for SteerLM](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384. Association for Computational Linguistics.

- Hongyu Yang, Liyang He, Min Hou, Shuanghong Shen, Rui Li, Jiahui Hou, Jianhui Ma, and Junda Zhao. 2024. [Aligning LLMs through Multi-Perspective User Preference Ranking-Based Feedback for Programming Question Answering](#). *arXiv preprint arXiv:2406.00037*.
- Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. 2025. [MasRouter: Learning to Route LLMs for Multi-Agent Systems](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15549–15572. Association for Computational Linguistics.
- Hang Zeng, Chaoyue Niu, Fan Wu, Chengfei Lv, and Guihai Chen. 2024. [Personalized LLM for Generating Customized Responses to the Same Query from Different Users](#). *arXiv preprint arXiv:2412.11736*.
- Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025a. [A Survey on Multi-Turn Interaction Capabilities of Large Language Models](#). *arXiv preprint arXiv:2501.09959*.
- Chi Zhang, Junho Jeong, and Jin-Woo Jung. 2025b. [Anomaly Detection Over Multi-Relational Graphs Using Graph Structure Learning and Multi-Scale Meta-Path Graph Aggregation](#). *IEEE Access*, 13:60303–60316.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [MemoryBank: Enhancing Large Language Models with Long-Term Memory](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

Appendix Contents

A	Implementation Details	15
A.1	Model Configuration and Hyperparameters	15
A.2	Sampling Unit in User-conditioned Graph Sampling	15
A.3	Training of GMTRouter	15
B	Dataset Preparation	15
B.1	Details for Datasets	15
B.2	Dataset Statistics	16
B.3	Scale of the ChatBot Arena Motivation Analysis	16
B.4	Feedback Modalities and Sparse Supervision	16
B.5	Synthetic User Design	16
B.6	Discussion of Real-World Datasets	17
C	Investigating the Impact of Visible Data Size	18
D	Additional Results on New User Experiments	19
E	Further Experiments	19
E.1	Human-like Preference Augmentation	19
E.2	Effect of Multi-turn Modeling and Turn Nodes	19
E.3	Comparison with MemoryBank	20
E.4	Efficiency Analysis	20
E.5	Using Different GNNs as the GMTRouter Backbone	20
E.6	Ablation on PLM-based Node Initialization	21
F	Baseline Routing Prompts	21
F.1	Retrieval Details for the Personalized LLM Baseline	21
F.2	Routing Prompt Templates	21
G	The Use of Large Language Models (LLMs)	21

A Implementation Details

A.1 Model Configuration and Hyperparameters

We implement our method using PyTorch Geometric (Fey and Lenssen, 2019) and conduct all experiments on a single NVIDIA RTX A6000 GPU.

Architecture. We use a heterogeneous graph transformer (HGT) with:

- **GNN:** 2 layers (single-turn) or 3 (multi-turn), 768-dim hidden, 4-head HGTCov, LayerNorm, dropout 0.1.
- **Predictor:** 4-head cross-attention followed by an MLP with hidden dimension 256 and dropout 0.1, where the LLM representation attends to the user and query representations.

Training.

- **Epochs:** 1000 **LR:** 5e-4
- **Visible records/user (k):** {3, 5, 8, 10, 15, 20}
- **Batch size:** 256 supervision triplets
- **Model selection:** prioritize validation AUC, then Accuracy.

A.2 Sampling Unit in User-conditioned Graph Sampling

In user-conditioned graph sampling, the basic sampling unit is one user–query instance from the Interaction History Table. Each instance contains a user, a query or multi-turn query context, a set of candidate LLMs, the corresponding responses generated by these candidate LLMs, and the user feedback associated with these responses. Therefore, sampling k interaction histories for a user means selecting k such user–query instances from this user’s available interaction records to construct the visible subgraph for message passing.

This design matches the personalized routing objective: for a given user and query, the router estimates the user’s preference over candidate LLMs and selects the LLM with the highest predicted preference score. In our experiments, each sampled instance contains two candidate LLM responses. The visible subgraph constructed from these sampled instances provides the local evidence used to infer the user’s preference representation during both training and inference.

A.3 Training of GMTRouter

Algorithm 1 summarizes the training procedure of GMTRouter.

Algorithm 1: Training GMTRouter

Data: $\mathcal{D}_{\text{train}} = \{(x, y)\}$
Hyperparams: epochs E , visible k , supervision s , PLM, GNN f_ϕ , predictor Pred
Init: PLM-encode query, response, and LLM nodes; initialize user/turn nodes and edge features

```

1 for  $e \leftarrow 1$  to  $E$  do
2    $\mathcal{G}^{(e)} \leftarrow$  subgraph from  $k|\mathcal{U}|$  visible records
3    $\mathcal{M}^{(e)} \leftarrow s$  held-out triples  $(u, q, m)$ 
4    $h \leftarrow f_\phi(\mathcal{G}^{(e)})$  // message passing
5   for  $(u, q, m) \in \mathcal{M}^{(e)}$  do
6      $\hat{y} \leftarrow \text{Pred}(h_u, h_q, h_m)$ 
7   Update  $f_\phi$  and  $\text{Pred}$  by minimizing  $\mathcal{L}_{\text{rank}}(\hat{y}, y)$ 

```

B Dataset Preparation

B.1 Details for datasets

We evaluate our approach on both real-world and synthetic datasets spanning five distinct tasks to provide a comprehensive assessment.

- **HUMAINE (Petrova et al., 2026):** HUMAINE is a large-scale benchmark of naturalistic multi-turn interactions with human preference judgments across demographically stratified participants. We group records sharing the same demographic profile into a preference cohort and treat each cohort as one “user”. For our experiments, we select the 20 “users” with the most interaction records. Detailed statistics are provided in Appendix B.2.
- **Chatbot Arena (Chiang et al., 2024):** ChatBot Arena is a real-world dataset of anonymized multi-turn conversations with pairwise human preference labels across LLMs. For our experiments, we select the 11 users and 16 LLMs with the largest numbers of interactions. Detailed statistics are provided in Appendix B.2.
- **MT-Bench (Zheng et al., 2023):** MT-Bench is a benchmark for evaluating the reasoning and multi-turn conversational capabilities of LLMs, containing 80 multi-turn questions.
- **GSM8K (Cobbe et al., 2021):** GSM8K is a dataset of grade school-level math word problems, designed to assess LLMs’ mathematical reasoning and problem-solving skills.
- **MMLU (Hendrycks et al., 2021a,b):** MMLU is a comprehensive benchmark covering 57 subjects across STEM, humanities, social sciences, and professional domains, used to measure general knowledge and multi-domain reasoning abilities of LLMs. We sample 10 questions from each subject for our experiments.

B.2 Dataset Statistics

We preprocess each dataset into user–query–LLM–response records and partition them into training, validation, and test sets. Table 4 summarizes the resulting statistics. We evaluate generalization to unseen users separately under the new-user setting described in Section 5.2.

Table 4: Dataset statistics, including the number of entries, users, and LLMs in each split.

Dataset	Split	Entries	Users	LLMs
ChatBot Arena	Train	2780	11	16
	Valid	386	11	16
	Test	824	11	16
MT Bench	Train	2240	10	2
	Valid	320	10	2
	Test	640	10	2
GSM8K	Train	18460	10	2
	Valid	2620	10	2
	Test	5300	10	2
MMLU	Train	3970	5	2
	Valid	560	5	2
	Test	1150	5	2
HUMAINE	Train	3535	20	45
	Valid	483	20	45
	Test	1020	20	45

For ChatBot Arena, we selected the following users and LLMs:

Users: arena_user_9965, arena_user_15085, arena_user_257, arena_user_13046, arena_user_11473, arena_user_3820, arena_user_9676, arena_user_6467, arena_user_6585, arena_user_5203, arena_user_1338

LLMs: koala-13b, vicuna-13b, gpt-3.5-turbo, oasst-pythia-12b, gpt-4, claude-v1, RWKV-4-Raven-14B, palm-2, alpaca-13b, mpt-7b-chat, vicuna-7b, claude-instant-v1, chatglm-6b, fastchat-t5-3b, dolly-v2-12b, stablelm-tuned-alpha-7b

We report in Table 5 the size of each dataset along with the time required to process it into the heterogeneous graph used in our experiments.

B.3 Scale of the ChatBot Arena Motivation Analysis

In Section 2.2, we use a subset of 10 active ChatBot Arena users to analyze cross-user differences in LLM preferences. This subset is used only for

the motivation analysis and is distinct from the ChatBot Arena split used in the main experiments. Although each selected user has at least 50 records, the analysis covers 1,480 interaction records in total, as shown in Table 6.

B.4 Feedback Modalities and Sparse Supervision

GMTRouter is designed to support heterogeneous feedback formats rather than assuming that all interactions share the same type of supervision. In our formulation, user feedback may appear as scalar ratings, pairwise or listwise preference rankings, or ground-truth responses. These feedback modalities can be converted into comparable preference signals over candidate LLM responses: scalar ratings are used directly, rankings are discretized into relative preference scores, and ground-truth responses are converted into supervision by measuring the similarity between the generated response and the reference response.

This design allows GMTRouter to be applied to datasets with different forms of personalization supervision. For example, HUMAINE and ChatBot Arena provide pairwise human preferences, whereas the synthetic datasets use user-specific ratings constructed from multiple response-level attributes, as described in Appendix B.5.

We also note that GMTRouter does not require dense feedback histories for each user. During both training and inference, user-conditioned graph sampling constructs a visible subgraph from only a small number of sampled interaction histories, as detailed in Appendix A.2. This enables the model to infer user preferences from sparse supervision and supports personalization in few-shot or cold-start settings.

B.5 Synthetic User Design

To simulate diverse user preferences, we introduce synthetic users whose routing behavior is governed by a weighted linear utility function over multiple response-level attributes: response quality, generation cost, response length, and rare-word count. These attributes are designed to reflect individualized trade-offs, such as prioritizing answer quality, preferring lower-cost responses, favoring concise versus detailed answers, or preferring simple versus more technical language. For each dataset, we manually assign different weights $\{w_{\text{quality}}, w_{\text{cost}}, w_{\text{length}}, w_{\text{rare}}\}$ per user to simulate these personalized preferences. These weights are

Table 5: Computational cost of graph construction on four datasets.

Dataset	Data Entries	Avg. Tokens	Encoding Time (s)	Graph Construction Time (s)
ChatBot-Arena	3990	184.41	51.73	1.70
MT-Bench	3200	4511.73	55.68	2.40
GSM8K	26380	112.68	142.84	1.49
MMLU	5680	9.35	4.27	1.56

Table 6: Number of interaction records for the 10 active ChatBot Arena users used in the motivation analysis.

User ID	# Records
arena_user_110	51
arena_user_15085	288
arena_user_11941	84
arena_user_14189	79
arena_user_257	292
arena_user_973	99
arena_user_4715	69
arena_user_9965	376
arena_user_12353	64
arena_user_13514	78
Total	1,480

normalized within each dataset to prevent scale bias.

Table 7: Synthetic user weights for MT-Bench dataset.

User	w_{quality}	w_{length}	w_{rare}	w_{cost}
user_1	1.42	0.0087	-0.174	-45.23
user_2	1.87	0.0012	0.091	-15.55
user_3	0.96	0.0135	0.045	-48.42
user_4	1.15	-0.0008	-0.220	-10.00
user_5	1.69	0.0024	0.175	-38.50
user_6	1.08	-0.0015	-0.030	-25.12
user_7	0.53	0.0162	0.230	-5.75
user_8	1.34	-0.0005	-0.145	-12.40
user_9	1.98	0.0101	0.087	-25.10
user_10	1.57	0.0024	-0.065	-7.79

B.6 Discussion of Real-World Datasets

Both HUMAINE and ChatBot Arena contain real-world multi-turn interactions with direct human preference judgments, but they differ substantially in their data collection protocols and degree of quality control.

ChatBot Arena. ChatBot Arena (Chiang et al., 2024) is collected through open online crowd-

Table 8: Synthetic user weights for GSM8K dataset.

User	w_{quality}	w_{length}	w_{rare}	w_{cost}
user_1	1.0	20.0	100.0	-0.0
user_2	1.5	18.0	50.0	-1.0
user_3	0.8	22.0	80.0	-0.5
user_4	1.2	17.0	120.0	-0.2
user_5	2.0	15.0	70.0	-0.4
user_6	0.4	6.0	-4.0	-1.0
user_7	0.3	7.0	-5.0	-0.9
user_8	0.6	8.0	-7.0	-1.2
user_9	0.2	9.0	-9.0	-0.8
user_10	0.8	10.0	-3.0	-1.1

Table 9: Synthetic user weights for MMLU dataset.

User	w_{quality}	w_{length}	w_{rare}	w_{cost}
user_1	1.0	0.00	0.00	0.0
user_2	1.0	0.00	0.00	-600.0
user_3	1.0	0.00	0.00	-1200.0
user_4	1.0	0.00	0.00	-1800.0
user_5	1.0	0.00	0.00	-2400.0

sourcing, where anonymous, self-selected users freely submit prompts and vote between model responses. This setting provides valuable in-the-wild interactions, but offers limited control over participant recruitment and interaction quality. Consequently, some comparisons can contain noisy or low-information preference signals, making stable personalized patterns harder to identify. Our manual inspection finds examples where the preferred response is objectively incorrect or where both responses substantially fail the task. Representative cases are shown in Table 10.

We emphasize that these examples do not imply that ChatBot Arena is generally low-quality. Rather, they illustrate an inherent trade-off of its open, in-the-wild collection setting: broad coverage and natural user behavior come with noisier preference supervision. This may partially explain

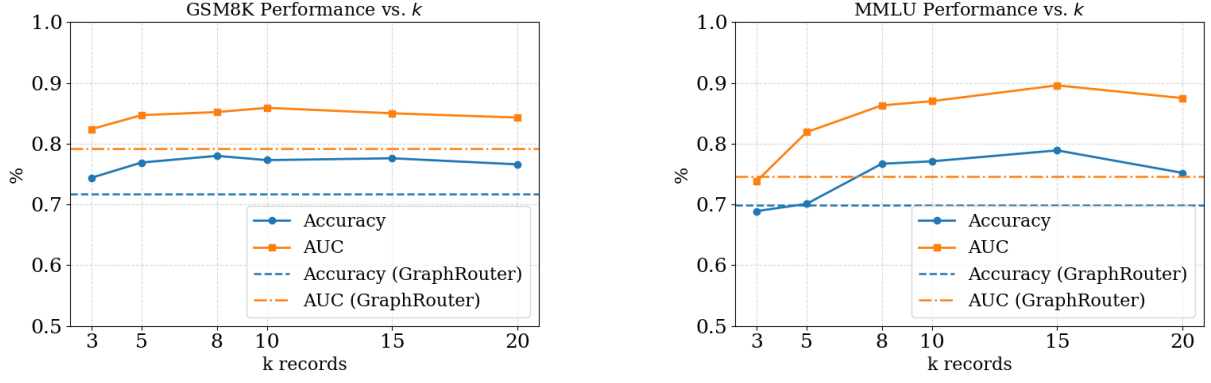


Figure 5: This figure illustrates **the impact of the visible data size k on GMTRouter for GSM8K (left) and MMLU (right)**. The dashed line represents the GraphRouter baseline. As k increases, the performance of our method improves, but it saturates once k reaches 10.

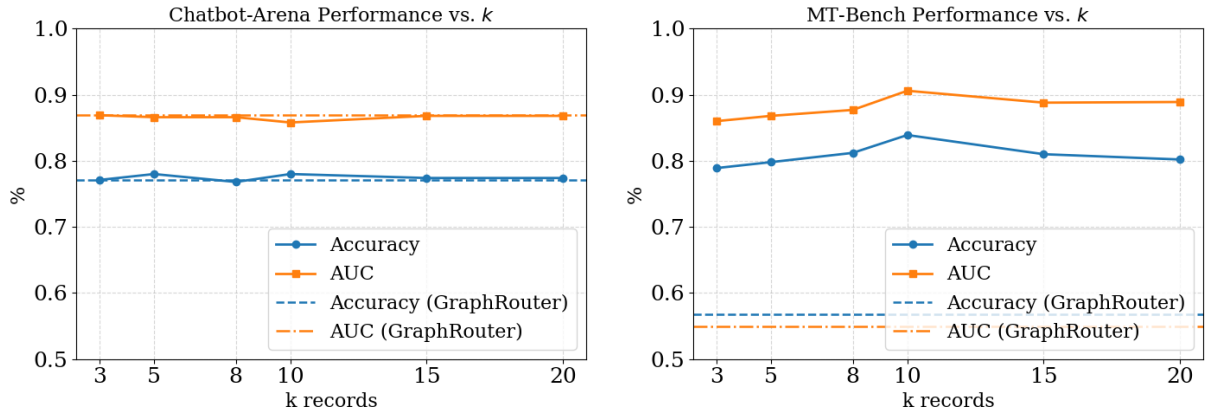


Figure 6: This figure illustrates **the impact of the visible data size k on GMTRouter for ChatBot Arena (left) and MT-Bench (right)**. The dashed line represents the GraphRouter baseline. As k increases, the performance of our method improves, but it saturates once k reaches 10.

why GMTRouter shows a more modest improvement on ChatBot Arena than on the other datasets.

HUMAINE. HUMAINE (Petrova et al., 2026) provides a more controlled real-user evaluation setting. Participants are recruited and compensated through Prolific, each comparison requires at least three conversational turns, and low-effort or disengaged interactions are monitored in real time. The benchmark also contains substantially more interaction records and candidate LLMs than the ChatBot Arena subset used in our experiments. These properties provide cleaner and larger-scale preference supervision for personalized routing, under which GMTRouter achieves substantially larger improvements over strong baselines.

HUMAINE does not release persistent participant identifiers. Therefore, as described in Section B.1, we group records sharing the same demographic profile into a preference cohort and treat each cohort as one “user”. This construction cap-

tures demographic-level preference heterogeneity rather than longitudinal preferences of a single individual, and should be interpreted accordingly.

C Investigating the Impact of Visible Data Size

We investigate the impact of k visible data per user on the quality of the aggregated node embeddings. Figures 5 and 6 present the test results on four datasets, respectively. As k increases, both accuracy and AUC improve, but beyond $k=10$, the performance begins to plateau or slightly decline, indicating diminishing returns from including additional visible data. This may be due to reduced generalization or potential instability caused by excessively large batch sizes during training (Keskar et al., 2016; Oyedotun et al., 2023). Therefore, we choose $k=10$ as a balanced setting for capturing user preferences without compromising generalization.

Query	Preferred response	Rejected response	Analysis
“Write a function in Scheme that reverses a list of strings?”	dolly-v2-12b: (define reverse-list ... (list (car l) (reverse-list (cdr l))))	gpt-3.5-turbo: (define (reverse-list-of- strings lst) ... (append ...))	The preferred response is incomplete and constructs nested lists, whereas the rejected response at least provides executable recursive code that reverses the list order.
“Simulate Python console. Input: print([2 * 2] * 2)”	dolly-v2-12b: [4, 4, 4, 4]	fastchat-t5-3b: print(2 * 2 * 2)	The correct output is [4, 4]; both responses are incorrect, making the preference signal uninformative.
“For each skill, choose their favorite horror game: Logic, Encyclopedia, Rhetoric, Drama, ...”	chatglm-6b: Logic: “The Shining”, ... Encyclopedia: “The Great Gatsby”, ...	mpt-7b-chat: Logic: Inland Empire; Encyclopedia: Shivers; ...	The preferred response mainly lists movies, books, or unrelated titles rather than horror games, while the rejected response maps skills to other skill names. Both substantially fail the task.

Table 10: Representative low-information or potentially unreliable preference comparisons in ChatBot Arena.

D Additional Results on New User Experiments

Here, we present the results of the experiment described in Section 5.2 on the ChatBot Arena and MT-Bench datasets, as shown in Figure 7.

E Further Experiments

E.1 Human-like Preference Augmentation

To further examine whether GMTRouter can capture richer user preferences beyond the original proxy metrics, we construct an augmented synthetic personalization setting on MT-Bench with additional human-like preference factors. In the original synthetic setting, user preferences are simulated using response quality, cost, response length, and rare-word count. While these dimensions capture important aspects of user preferences, they may not fully reflect more subjective factors such as tone, safety, and factuality.

Specifically, we use Qwen3-32B-AWQ as an annotator to label each response with three additional attributes: (1) **tone**, selected from a predefined set of response styles; (2) **safety**, represented as a binary label; and (3) **factuality**, represented as a score from 0 to 100. We then extend each simulated user profile with four additional preference attributes: a preferred tone, a tone-matching weight, a safety weight, and a factuality weight. These additional attributes are combined with the original synthetic preference dimensions to reconstruct the user-specific ratings.

Table 11 reports the results on this augmented MT-Bench setting. GMTRouter consistently out-

Table 11: Performance on the augmented MT-Bench setting with additional human-like preference factors. The best results are highlighted in bold.

Method	Accuracy	AUC
GraphRouter	0.667	0.717
TIGER	0.642	0.721
RouteLLM	0.429	0.462
FrugalGPT	0.642	0.687
GMTRouter	0.715	0.777

performs all representative baselines, achieving the best performance in both Accuracy and ROC-AUC. These results suggest that GMTRouter remains effective when user preferences involve more human-like and subjective factors beyond the original proxy metrics.

E.2 Effect of Multi-turn Modeling and Turn Nodes

To evaluate the effect of multi-turn modeling and the proposed turn-node design, we conduct additional ablation experiments on MT-Bench. We compare three settings: (1) **Single-turn only**, where the model is trained and evaluated using only single-turn interactions; (2) **Multi-turn w/o turn nodes**, where the model uses the same multi-turn data but removes turn nodes from the heterogeneous graph; and (3) **Multi-turn**, which corresponds to the full GMTRouter setting with explicit turn nodes.

Table 12 summarizes the results. Multi-turn interactions substantially improve routing performance over the single-turn setting, indicating that richer interaction context provides useful prefer-

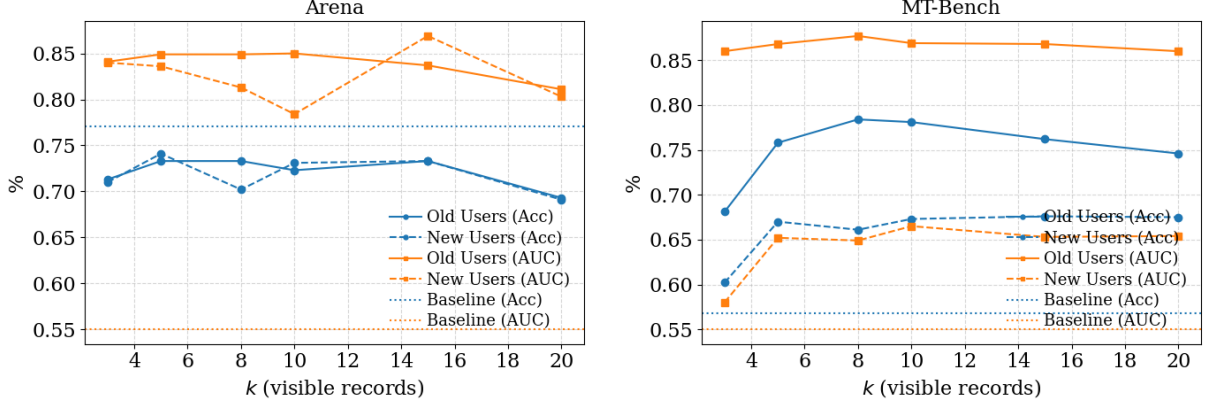


Figure 7: **Result comparison between old-user and new-user settings for ChatBot Arena (left) and MT-Bench (right).** The dashed line represents the GraphRouter baseline. The personalized performance under the new-user setting is comparable to that under the old-user setting, highlighting the strong generalization capability of our method.

ence signals for personalized routing. Moreover, adding turn nodes further improves performance over the multi-turn variant without turn nodes, especially in ROC-AUC. This supports our design motivation that turn nodes help aggregate local information within each dialogue round and facilitate preference propagation across multi-turn user-LLM interactions.

Table 12: **Effect of multi-turn modeling and turn nodes on MT-Bench.** The best and second-best results are highlighted in **bold** and underline, respectively.

Setting	Accuracy	AUC
Single-turn only	0.463	<u>0.709</u>
Multi-turn w/o turn nodes	<u>0.644</u>	0.669
Multi-turn	0.675	0.786

E.3 Comparison with MemoryBank

To further compare GMTRouter with recent memory-based personalization methods, we include MemoryBank (Zhong et al., 2024) as an additional baseline. MemoryBank equips LLMs with a long-term memory module that stores, retrieves, and updates past interactions to support personalized responses over time. We adapt it to the personalized routing setting by retrieving user-specific interaction histories as memory and using the backbone LLM to select the candidate LLM that best matches the user’s preference. We use Llama-4-Scout-17B-16E-Instruct as the backbone model for this baseline.

Table 13 reports the results across four datasets. GMTRouter consistently outperforms MemoryBank on both Accuracy and ROC-AUC. This com-

parison further supports our main observation that, compared with methods that primarily store and retrieve unstructured user histories, GMTRouter benefits from learning over a fine-grained heterogeneous graph that explicitly models user–query–LLM–response interactions.

E.4 Efficiency Analysis

To further evaluate the computational overhead of GMTRouter, we compare it with representative training-based routing baselines on MT-Bench. We report training time, peak training GPU memory, and inference latency per query in Table 14. Since LLM-based prompting baselines require substantially larger inference-time computation, we focus this comparison on lightweight training-based methods.

The results show that GMTRouter introduces additional training cost due to its graph-structured user-LLM interaction modeling. However, its peak training memory remains close to other lightweight training-based routers, and its inference latency is still at the millisecond level. This indicates that GMTRouter preserves practical efficiency while providing stronger personalized routing performance.

E.5 Using Different GNNs as the GMTRouter Backbone

We investigate how different GNNs used as the GMTRouter backbone affect its personalization capability. We evaluate four backbone models, including both homogeneous and heterogeneous GNNs: GraphSAGE (Hamilton et al., 2017), HGT

Table 13: **Comparison with MemoryBank across four datasets.** The best results are highlighted in **bold**.

Method Metric	ChatBot Arena		MT-Bench		GSM8K		MMLU	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
MemoryBank (Zhong et al., 2024)	0.601	0.634	0.575	0.555	0.555	0.552	0.610	0.604
GMTRouter	0.774	0.875	0.784	0.859	0.773	0.859	0.771	0.870

Table 14: **Efficiency comparison on MT-Bench.** Lower values indicate better efficiency, and the best results are highlighted in **bold**.

Method	Training Time (s)	Peak Train GPU Mem (MB)	Inference Time / Query (s)
GraphRouter	156.77	1898.64	0.0052
TIGER	32.65	1475.35	0.0014
GMTRouter	710.46	2050.70	0.0063

(Hu et al., 2020), HAN (Wang et al., 2019), and HeteroConv, and present the results in Table 15.

HGT achieves the strongest overall performance, obtaining the best AUC across all four datasets and the best or near-best Accuracy. HAN is also competitive on several datasets. These results further indicate that GMTRouter is not dependent on any particular GNN architecture: multiple attention-based backbones achieve strong performance, suggesting that the gains primarily stem from the graph-based personalized routing framework and data modeling rather than from a specific convolutional operator.

E.6 Ablation on PLM-based Node Initialization

GMTRouter uses PLM embeddings to initialize the query, response, and LLM nodes, thereby incorporating their semantic information into the interaction graph. To examine the contribution of such semantic initialization, we construct a variant that replaces these PLM embeddings with non-semantic 64-dimensional hash embeddings while keeping the remaining architecture unchanged.

As shown in Table 16, replacing PLM representations with hash embeddings consistently degrades performance across all four datasets, reducing Accuracy by **0.058–0.129** and AUC by **0.077–0.146**. This confirms that semantic node initialization provides useful information for personalized routing. Nevertheless, PLM initialization serves as a supporting component of GMTRouter; the core design of our framework remains the structured modeling of multi-turn user–LLM interactions together with user-conditioned inductive graph learning.

F Baseline Routing Prompts

F.1 Retrieval Details for the Personalized LLM Baseline

For the Personalized LLM baseline, retrieval is performed separately for each user rather than globally. Given a new query from a user, we first encode the query using Contriever. We then encode all historical queries from the same user’s training set and compute their inner-product similarity with the new query embedding. The top-10 most similar interaction histories are retrieved and inserted into the prompt as personalized context. This per-user retrieval design ensures that the prompt contains only interaction histories from the target user, allowing the LLM baseline to perform personalized routing based on user-specific preference evidence.

F.2 Routing Prompt Templates

To benchmark routing strategies, we design two representative prompt templates: one for a vanilla router that selects the best LLM without personalization, and another for a personalized router that incorporates user history and preferences. Both prompts simulate realistic routing scenarios where a system must choose a single LLM for the next turn in a multi-turn dialogue.

G The Use of Large Language Models (LLMs)

During the writing of this paper, we used the GPT-5 Mini model for text polishing and grammatical corrections to enhance the readability of the manuscript.

Table 15: **GMTRouter performance with different GNN backbones.** The best and second-best results are highlighted in **bold** and underline, respectively.

Dataset Metric	Chatbot-Arena		MT-Bench		GSM8K		MMLU	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
GraphSAGE	0.768	<u>0.873</u>	0.569	0.645	0.635	0.648	0.494	0.487
HeteroConv	0.777	0.867	0.569	0.492	0.499	0.603	0.494	0.542
HANConv	0.766	0.776	<u>0.646</u>	<u>0.680</u>	0.774	<u>0.775</u>	<u>0.707</u>	<u>0.746</u>
HGTConv	<u>0.774</u>	0.875	0.784	0.859	<u>0.773</u>	0.859	0.771	0.870

Table 16: **Ablation on PLM-based node initialization across four datasets.** We replace the PLM embeddings of query, response, and LLM nodes with non-semantic 64-dimensional hash embeddings. The best results are highlighted in **bold**.

Method Metric	HUMAINE		MT-Bench		GSM8K		MMLU	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Hash Embedding	0.649	0.697	0.658	0.729	0.702	0.782	0.713	0.767
GMTRouter	0.778	0.843	0.784	0.859	0.773	0.859	0.771	0.870

Table 17: **Prompt Template: Vanilla LLM Routing (No Personalization)**

[Instruction]

You are an expert routing agent. Your task is to select the most suitable Large Language Model (LLM) to handle the next query in a multi-turn conversation.

[Input Format]

[Candidate LLM List]

{{CANDIDATE_LLM_LIST}}

[Previous Conversation]

{{PREVIOUS_CONVERSATION}}

[Current Query]

{{CURRENT_QUERY}}

[Instructions for Model Selection]

- Consider the query difficulty, the context of the previous conversation, and each LLM’s expertise, cost, and size.
- Choose the single best LLM to respond to the current query.
- Output only the name of the selected LLM in the exact format below.
- Do not provide explanations or commentary.

[Output Format]

<‘{selected_model_name}’>

Table 18: **Prompt Template: Personalized Routing (User History Aware)**

[Instruction]

You are an expert routing agent. Your task is to select the most suitable Large Language Model (LLM) to handle the next query in a multi-turn conversation, incorporating both model characteristics and personalization signals from the user’s history.

[Input Format]

[Candidate LLM List]

{{CANDIDATE_LLM_LIST}}

[Previous Conversation]

{{PREVIOUS_CONVERSATION}}

[Current Query]

{{CURRENT_QUERY}}

[User Preference History]

{{USER_PREFERENCE_HISTORY}}

[Instructions for Model Selection]

- Consider the query difficulty, the context of the ongoing conversation, the LLMs’ specializations, cost, and size.
- Additionally, factor in the user’s historical preferences and ratings to personalize the routing decision.
- Choose the single best LLM to respond to the current query.
- Output only the name of the selected LLM in the exact format below.
- Do not provide explanations or commentary.

[Output Format]

<‘{selected_model_name}’>
