Simplifying user identity linkage model via user modeling and two-phase instruction fine-tuning

Jiaqi Gao, Kangfeng Zheng*, Minjiao Yang, Yulin Yao
Beijing University of Posts and Telecommunications, Beijing, China
Email: jessie_gao@bupt.edu.cn, kfzheng@bupt.edu.cn, yangminjiao@bupt.edu.cn, yuliny@bupt.edu.cn

***Abstract*—User identity linkage (UIL) is a fundamental task in social network analysis, yet most existing approaches rely on computationally expensive ensembles of deep neural networks, limiting their scalability and generalization. We propose UILLLM, a simple and unified framework that performs UIL using a single compact large language model. UILLLM introduces a Physical–Social–Cognitive (PSC) user modeling paradigm and adopts a two-phase instruction fine-tuning strategy. In the first phase, structured PSC representations are distilled from user-generated content to guide model learning. In the second phase, UIL is formulated as a binary classification task based on these representations, enabling effective identity linkage with low computational overhead. Experiments on two real-world datasets, TWIN and TWFQ, show that UILLLM consistently outperforms state-of-the-art methods, achieving accuracies of 96.56% and 99.14%, respectively.**

***Index Terms*—Large language model, user identity linkage, data mining, user modeling, instruction fine-tuning**

I. Introduction

User identity linkage (UIL) aims to link accounts belonging to the same individual across different social media platforms [1]. It has wide-ranging applications in recommendation systems [2], information diffusion analysis [3], and user behavior modeling [4]. Social media data typically consists of three modalities including user profiles, user-generated content (UGC), and social relationships, which together provide rich yet highly noisy signals for identity inference.

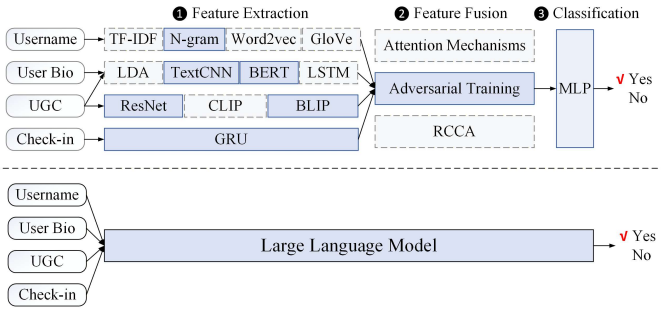


Fig. 1: Comparison between the common framework of UIL and LLM-based UIL methods.

Most existing UIL methods rely on multiple deep neural networks (DNNs) to separately extract and fuse heterogeneous features, as illustrated in Fig. 1. While effective in specific settings, this design substantially increases training complexity and limits scalability when new feature types are introduced. Moreover, social media data often contains a large amount of noise irrelevant to user identity, and directly training models on raw data frequently leads to unstable convergence and suboptimal generalization.

Recent advances in large language models (LLMs) have demonstrated remarkable reasoning and generalization capabilities across diverse domains [5], [6]. These properties make LLMs particularly appealing for UIL, as they provide a unified modeling framework for heterogeneous data, capture fine-grained semantic identity cues from noisy inputs, and adapt to new platforms through instruction tuning and in-context learning. Compared to traditional multi-DNN architectures, LLMs offer a more flexible and scalable alternative for modeling complex social media data.

Despite their potential, directly applying LLMs to UIL remains challenging. Without task-specific adaptation [7], LLMs may exhibit instruction deviation and uncontrollable outputs. In addition, the massive volume of noisy UGC can easily exceed context length limits and degrade performance when used for direct fine-tuning. These challenges highlight a significant gap between the expressive power of LLMs and the practical requirements of UIL.

To bridge this gap, we propose UILLLM, a novel framework that adapts LLMs to UIL via instruction fine-tuning [8], as shown in Fig. 2. Our approach introduces the Physical–Social–Cognitive(PSC) user modeling framework to construct robust user representations from three complementary dimensions. In the first stage, a teacher LLM processes raw UGC to generate structured PSC features, which serve as high-quality supervision for training a smaller student LLM. This process effectively filters irrelevant noise while transferring user modeling knowledge. In the second stage, the student LLM is further fine-tuned for UIL using an instruction dataset that maps PSC features to binary linkage decisions. Additionally, we incorporate targeted data augmentation to mitigate overfitting to surface-level identifiers such as usernames, significantly improving generalization to unseen users. We conduct extensive experiments on two real-world datasets, TWIN [13] and TWFQ [9], using three open-source LLMs of different scales. Experimental results demonstrate that UILLLM consistently outperforms state-of-the-art UIL methods on core metrics, validating its effectiveness and robustness in handling noisy,

*corresponding author.

heterogeneous social media data.

The main contributions of this work are summarized as follows:

- We propose the PSC user modeling framework, which constructs compact and robust user representations by effectively reducing irrelevant data noise.
- We present UILLLM, to the best of our knowledge the first framework that systematically explores instruction fine-tuning of LLMs for the UIL task.
- Extensive experiments on TWFQ and TWIN demonstrate that UILLLM significantly improves both effectiveness and generalization compared to state-of-the-art UIL methods.

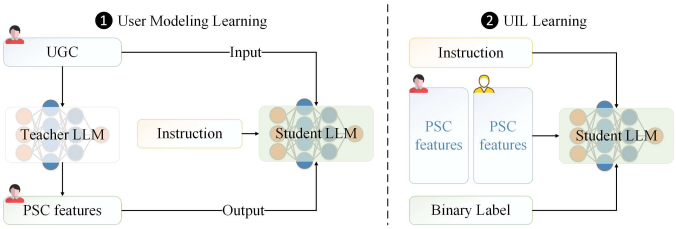


Fig. 2: Overview of the proposed UILLLM.

We will make the code publicly available upon acceptance of the paper.

II. Related Work

A. User identity linkage

UIL aims to match accounts belonging to the same user across different platforms. Most existing methods follow a multimodal feature engineering pipeline, where separate encoders are used to extract heterogeneous signals such as text, images, check-ins, and social relations, and then fused for alignment. For example, Chen et al. [9] combined BERT [10], ResNet [11], and GRU [12] for text, image, and check-in modeling and employed a hybrid graph network for unified representation. Chen et al. [13] adopted BiLSTM [14] and a pre-trained ResNet to encode text and images. Li et al. [15] utilized bag-of-words [16], BERT, and ResNet to model usernames, posts, and images while leveraging social relations for improved learning. Li et al. [17] further integrated BERT, GRU, BLIP [18], and ConvNeXt [19] through multimodal self-attention. Chen el.al [20] extracted username, location, and post features using bag-of-words, Word2Vec [21], and LDA [22], and applied RCCA [23] for multimodal correlation modeling. Beyond encoder design, some studies focus on robustness and discriminability in graph-based linking, such as Adaptive Graph Walk for stable representations and cross-network fusion [24], or dynamic embedding selection with iterative clustering and multi-round prediction [25].

Although effective, a common characteristic of these methods is their reliance on multiple specialized pre-trained backbones plus additional fusion modules, which results in complex architectures and high computational overhead. Different from this paradigm, our work leverages a single LLM to serve as a unified feature extractor and reasoner for cross-platform user modeling and matching. This simplifies the overall pipeline by reducing modality-specific submodels and provides an efficient, controllable alternative for UIL.

B. Parameter-efficient fine-tuning

Adapting large pre-trained models to downstream tasks is typically achieved via full fine-tuning or parameter-efficient fine-tuning (PEFT) [26]. Full fine-tuning updates all parameters, whereas PEFT updates only a small set of parameters to reduce computational cost and alleviate catastrophic forgetting. LoRA [27] is a representative PEFT method that injects low-rank trainable matrices into model weights, and QLoRA [28] further reduces memory usage by quantizing weights to 4-bit while maintaining high-precision gradients with quantization constants, enabling fine-tuning of large models under limited GPU resources.

Instruction fine-tuning is another important technique for improving the task generalization of LLMs by explicitly incorporating task instructions during training. Prior work shows that instruction tuning helps a single model adapt to diverse tasks [29], and has been applied to domains such as code vulnerability detection [30] with Code Llama [31] and recommendation systems [32]. Following this line, we construct an instruction dataset and fine-tune our LLM with QLoRA to efficiently specialize it for UIL. Concretely, we fine-tune the student LLM to first learn structured PSC user modeling and then perform precise UIL matching, transforming a generic LLM into a controllable tool for cross-platform user linkage.

III. Methodology

A. Overview

While LLMs have achieved strong reasoning and generalization performance in various natural language processing tasks, practical deployments often favor smaller-scale models due to computational and cost constraints. However, such models typically exhibit limited instruction-following and task reasoning capabilities, which may result in off-target generation. To address this issue, we propose UILLLM, an instruction fine-tuned framework designed to enhance task-specific reasoning through carefully constructed training data.

UILLLM is built upon Llama-based architectures [33] with varying parameter scales and adopts a two-phase instruction fine-tuning strategy. Let the UGC of user u^i be denoted as $T^i = \{t_1^i, t_2^i, \ldots, t_N^i\}$, where each t_k^i corresponds to an individual post. In the first phase, the model is trained to learn user modeling capabilities by extracting *physical* and *cognitive* features from T^i. In the second phase, the model is fine-tuned for UIL reasoning, where it determines whether a pair of users corresponds to the same real-world individual based on their extracted PSC representations.

B. PSC User Modeling

Social media data is inherently noisy, unstructured, and contextually lengthy. Directly fine-tuning LLMs on raw UGC is inefficient and may introduce substantial noise, thereby

degrading model performance. To mitigate these issues, we propose a PSC user modeling framework that compresses noisy UGC into structured and discriminative user representations.

The PSC framework characterizes user identities from three complementary dimensions, namely physical, social, and cognitive. UIL typically relies on features that satisfy two key properties, cross-platform consistency and individual distinctiveness. Personal information and behavioral patterns often remain consistent across platforms, such as similar usernames or thematically aligned content published within overlapping time periods. Meanwhile, user interests reflect stable cognitive traits that exhibit strong individuality and transfer well across platforms. In addition, physical attributes including language usage and geographic information provide real-world cues that further distinguish users. Given the prevalence of noisy and incomplete information in social networks, relying on any single feature dimension is insufficient. By jointly modeling multiple views, PSC effectively captures both consistency and distinctiveness, enabling robust user representation for UIL.

As illustrated in Fig. 3, the PSC representation of user u^i is denoted as $PSC^i = \{P^i, S^i, C^i\}$, where P^i, S^i, and C^i correspond to physical, social, and cognitive features, respectively.

Fig. 3: The PSC user modeling framework.

Physical features $P^i = \{L^i, G^i\}$ capture linguistic patterns L^i and geolocation information G^i. Here, L^i denotes the set of languages used in the UGC T^i, while G^i represents the countries or cities mentioned in T^i. As these attributes are not entirely self-defined by users, they often reflect underlying cultural backgrounds, social environments, and behavioral habits, serving as reliable cues for identity discrimination.

Social features S^i consist of usernames, profile descriptions, and self-declared locations extracted directly from platform-provided profile fields, without requiring additional inference. To handle incomplete profiles, missing fields are filled with standardized placeholders and incorporated into the prompting process to maintain input consistency.

Cognitive features C^i describe user interests and preferences. Such interests are typically stable, personalized, and transferable across platforms, making them particularly informative for cross-network identity linkage. When combined with physical and social features, cognitive attributes contribute to a more comprehensive and discriminative user profile.

Overall, the PSC user modeling framework compresses noisy UGC into structured and highly informative representations, which are well suited for LLM fine-tuning. In the following section, we introduce how distillation learning enables LLMs to acquire effective PSC user modeling capabilities.

C. User Modeling Learning

The first training phase aims to equip the model with PSC user modeling capabilities, enabling it to extract physical features P^i and cognitive features C^i from noisy UGC. To this end, we adopt a teacher–student distillation framework, where the GPT-4 serves as the teacher model to provide structured supervision through prompt engineering.

As shown in Equation (1), the teacher model f_{teacher} processes the UGC collection T^i of user u^i and generates the corresponding physical and cognitive features, denoted as p^i and c^i, respectively. The prompt used for PSC user modeling, denoted as $Prompt_{UM}$, is illustrated in Fig. 4.

$$(p^i, c^i) = f_{\text{teacher}}(Prompt_{UM}(T^i)), \tag{1}$$

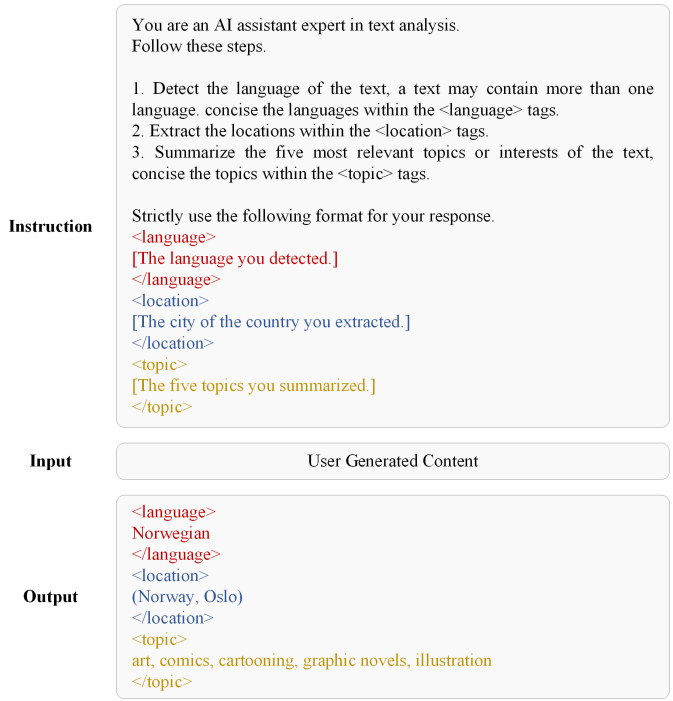


Fig. 4: The prompt for the PSC user modeling task.

The extracted physical features are strictly normalized to ensure consistency. When geographic information is mentioned at the city level, it is expanded into (country, city) tuples; country-only mentions retain the country name, while missing geographic information is marked as `NaN`.

For cognitive feature extraction, we design two prompt strategies, namely predefined topic categorization and open-ended description generation, as illustrated in Fig. 5. Although

predefined categorization produces structured outputs, its fixed topic space limits feature expressiveness and discriminative power. In contrast, the open-ended strategy leverages the emergent generative capabilities of LLMs to produce personalized and fine-grained interest descriptions. Based on empirical observations, we adopt the open-ended prompt strategy in our experiments.

	Predefined Topic Categorization	Open-ended Description Generation
Prompt Eaxmple	"**Categorize** the user's interests from the text into one or more of: [music, sports, technology, travel, food, politics]."	"**Summarize** the five most relevant topics or interests of the text."
Response Eaxmple	[music, sports, technology]	[music, hip-hop, basketball, Los Angeles Lakers, machine learning]

Fig. 5: Two prompt strategies for cognitive feature extraction.

For instruction fine-tuning, we follow the template introduced in Alpaca [38], as shown in Fig. 6. An instruction tuning dataset D_{UM} is constructed, where each sample consists of the task prompt $Prompt_{UM}$, the user post collection T^i as input, and the teacher-generated PSC features (P^i, C^i) as supervision. The student LLM is then fine-tuned on D_{UM} using QLoRA.

```
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
### Instruction:
[Task Prompt]
### Input:
[Input]
### Response:
[Output]
```

Fig. 6: The Alpaca format of the instruction fine-tuning dataset.

D. User Identity Linkage Learning

The second fine-tuning phase focuses on UIL inference. Based on the PSC representations obtained in Phase One, we construct an instruction tuning dataset for UIL reasoning. As illustrated in Fig. 7, the prompt $Prompt_{UIL}$ instructs the student model to determine whether a user pair (u^i, u^j) corresponds to the same real-world individual, given their PSC features PSC^i and PSC^j. To facilitate instruction tuning, the structured PSC features are converted into natural language descriptions. The label $y^{ij} \in \{0, 1\}$ indicates whether the two users are identity-linked.

Existing UIL datasets suffer from systematic biases, particularly the prevalence of positive user pairs sharing highly similar or even identical usernames. Such biases encourage models to overfit to superficial username patterns, leading to label leakage and poor generalization in real-world scenarios. To alleviate this issue, we introduce a data augmentation strategy that applies username perturbations to a proportion λ of positive samples.

Specifically, four types of perturbation operations are employed: (1) random character substitution, (2) random character deletion, (3) random character insertion, and (4) fully random character generation. For example, the username "john_doe" may be transformed into variants such as "jxhn_do3" or "j8h_doe". These perturbations preserve valid username formats while disrupting their original semantic associations. By injecting controlled noise into username features, the proposed strategy reduces the model's reliance on surface-level identifiers and encourages greater attention to deeper semantic cues captured by PSC representations.

The complete workflow of the proposed two-phase instruction fine-tuning framework is summarized in Algorithm 1.

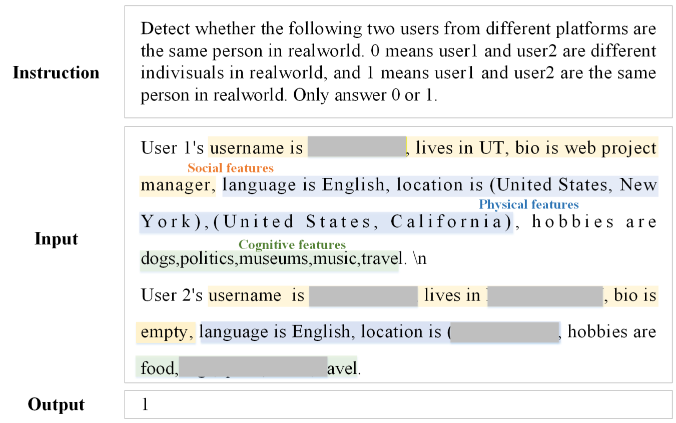


Fig. 7: The prompt for the UIL task.

Algorithm 1 UILLLM

Input: Raw user UGC T, Social profile data S_{prof}, Pre-trained Teacher LLM M_t, Base Student LLM M_s
Output: Fine-tuned Student LLM M_s^* for UIL task

$\mathcal{D}_{um} \leftarrow \emptyset$
for each user u^i **do**
  $P^i, C^i \leftarrow M_t(T^i, S^i_{prof})$
  Construct instruction instance I^i_{um} with P^i, C^i
  $\mathcal{D}_{um} \leftarrow \mathcal{D}_{um} \cup \{I^i_{um}\}$
end for
$M_s' \leftarrow \text{InstructionFineTune}(M_s, \mathcal{D}_{um})$
$\mathcal{D}_{uil} \leftarrow \emptyset$
for each user pair (u^i, u^j) with linkage label y_{ij} **do**
  $P^i, C^i \leftarrow M_s'(T^i, S^i_{prof})$
  $P^j, C^j \leftarrow M_s'(T^j, S^j_{prof})$
  $\mathrm{S}^i_{aug}, \mathrm{S}^j_{aug} \leftarrow \text{DataAugmentation}(\mathrm{S}^i, \mathrm{S}^j)$
  Construct instruction instance: I_{uil} with $P^i, S^i_{aug}, C^i, P^j, S^j_{aug}, C^j$
  $\mathcal{D}_{uil} \leftarrow \mathcal{D}_{uil} \cup \{I_{uil}\}$
end for
$M_s^* \leftarrow \text{InstructionFineTune}(M_s', \mathcal{D}_{uil})$
return M_s^*

IV. Experiments

A. Experimental settings

Datasets. We conduct extensive experiments on two real-world datasets, TWIN [13] and TWFQ [9], where TWIN is a UGC-based dataset and TWFQ contains user profiles, UGCs, check-ins, and social relationships.

- TWIN collected 15,595 users with accounts on six social networks using the cue "#mytweet via Instagram" [34], and retrieved the latest 200 UGCs of each user from Twitter and Instagram. It kept only users with more than 150 UGCs on both platforms, yielding 5,765 matched user pairs and 1,729,500 UGCs.
- TWFQ, built by Chen et al. based on Ren et al.'s dataset [35], originally contained 3,282 positive Twitter–Foursquare user pairs with posts and social relations, and kept 1,071 pairs after filtering for complete records. It also collected the latest 50 images per user on each platform via APIs, yielding 33,454 Twitter and 28,567 Foursquare images.

TABLE I: Statistics of the datasets.

Dataset	Train Set	Test Set
TWFQ	1714	214
TWIN	11234	1249

For each positive pair, we generated negative samples by randomly pairing the Twitter user with other users. We split the resulting data into training and test sets with a 9:1 ratio. Detailed statistics are reported in Table I.

Baselines. We compare UILLLM with eight representative UIL approaches, grouped by their primary modeling focus and corresponding datasets.

On TWIN, UserNet [13] encodes textual and visual content using BiLSTM and a pre-trained ResNet, respectively. EUIL [36] uses a pre-trained vision–language model to extract multimodal features, while TKM [37] enhances post representations via topic modeling and sequential encoding.

On TWFQ, MAUIL [20] models usernames, locations, and posts with handcrafted features and aligns them using RCCA. AHG-Net [9] integrates text, image and check-in features through a hybrid graph network with adversarial learning. AMSA [17] employs multimodal self-attention to fuse features from language and vision encoders, and MFLink [15] aggregates username, content, and social-relation information for user identity linkage.

Settings. We perform instruction fine-tuning on three Llama backbones of different scales, including Llama3.2-1B-Instruct, Llama3.2-3B-Instruct, and Llama3.1-8B-Instruct, to study the effect of model size. We use a learning rate of 1×10^{-4}, a maximum sequence length of 512, a batch size of 32, and train the models for three epochs. For efficiency, the models are loaded in 8-bit precision and fine-tuned using LoRA with rank 16 and alpha set to 32, which is applied to the attention projection layers q_proj, k_proj, v_proj, and o_proj.

B. Preliminary analysis

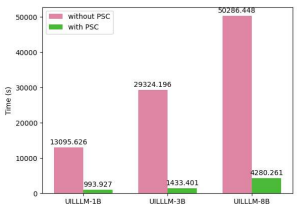
Table II shows that UILLLM clearly surpasses representative baselines on TWIN, delivering the strongest overall performance. Smaller backbones perform best on most metrics, while the largest model mainly improves recall. Table III indicates near-ceiling results on TWFQ and a clear advantage over strong baselines such as MFLink, suggesting the task may be partly driven by easy-to-match surface cues. This is consistent with TWFQ containing many paired users with highly similar or identical usernames, which can inflate in-dataset scores for small-scale data. To verify robustness, we further test cross-dataset generalization and synthetic settings with perturbed usernames, where UILLLM still maintains strong performance when username consistency is disrupted.

We further conducted bootstrap hypothesis testing with 1,000 resamples, with EUIL on TWIN and MFLink on TWFQ as baselines. We computed 95% CIs for the accuracy differences between UILLLM-8B and the baselines; improvements are significant at $\alpha = 0.05$ when the CI excludes zero. As shown in Table IV, gains are significant on both datasets.

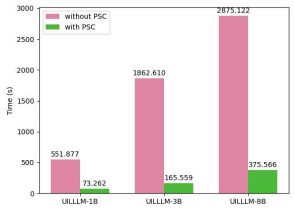
C. Ablation study

Ablation on PSC user modeling. To assess the impact of PSC user modeling, we replace PSC features with raw user posts and retrain the UILLLM-1B-wo-PSC, UILLLM-3B-wo-PSC, and UILLLM-8B-wo-PSC models. As shown in Table V and Table VI, removing PSC results in a substantial performance drop on TWIN, exceeding 35% across all metrics. On TWFQ, UILLLM-1B exhibits limited benefit from PSC, whereas UILLLM-3B and UILLLM-8B consistently degrade when PSC is removed. These results indicate that PSC effectively filters redundant information and extracts discriminative identity features, while raw posts introduce noise and suffer from context length limitations.

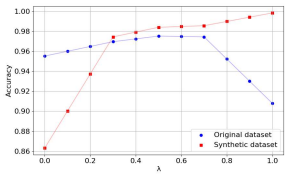
PSC also improves training efficiency. As shown in Fig. 8a and Fig. 8b, models equipped with PSC achieve approximately 10× faster per-epoch training on both TWIN and TWFQ by compressing lengthy post histories into compact multi-level user meta-features.



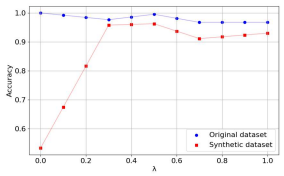
(a) Training time on TWIN.



(b) Training time on TWFQ.



(c) Accuracy on TWIN.



(d) Accuracy on TWFQ.

Fig. 8: Ablation Study on Training Efficiency and Data Augmentation.

Ablation on instruction fine-tuning. We evaluate instruction tuning by feeding PSC-based prompts directly into untuned

TABLE II: Results for UILLLM and baselines on TWIN.

Methods	Accuracy	Precision	Recall	F1
UserNet	0.8311	0.8448	0.8112	0.8276
TKM	0.8601	0.8632	0.8559	0.8595
EUIL	0.8815	0.8387	0.9362	0.8846
UILLLM-1B	$\mathbf{0.9656 \pm 0.0010}$	$\mathbf{0.9608 \pm 0.0018}$	0.9656 ± 0.0005	$\mathbf{0.9632 \pm 0.0010}$
UILLLM-3B	0.9592 ± 0.0014	0.9472 ± 0.0027	0.9705 ± 0.0006	0.9587 ± 0.0013
UILLLM-8B	0.9552 ± 0.0014	0.9248 ± 0.0025	$\mathbf{0.9885 \pm 0.0002}$	0.9556 ± 0.0013

TABLE III: Results for UILLLM and baselines on TWFQ.

Methods	Accuracy	Precision	Recall	F1
SSF	0.5549	0.5723	0.5471	0.5686
MAUIL	0.6186	0.6234	0.6031	0.6171
AHG-Net	0.9035	0.9269	0.8993	0.9107
AMSA	0.7196	0.6909	0.7451	0.7170
MFLink	0.9460	0.9544	0.9302	0.9421
UILLLM-1B	0.5327 ± 0.0032	0.5193 ± 0.0018	0.8785 ± 0.0031	0.6528 ± 0.0021
UILLLM-3B	0.9766 ± 0.0030	0.9722 ± 0.0019	0.9813 ± 0.0067	0.9767 ± 0.0034
UILLLM-8B	$\mathbf{0.9914 \pm 0.0031}$	$\mathbf{0.9892 \pm 0.0036}$	$\mathbf{0.9920 \pm 0.0029}$	$\mathbf{0.9906 \pm 0.0032}$

TABLE IV: Bootstrap analysis of accuracy differences between UILLLM-8B and the baseline models

Dataset	Baseline	Accuracy Difference (Bootstrap 95% CI)	Statistical Significance
TWIN	EUIL	0.0772[0.0758,0.0783]	Yes
TWFQ	MFLink	0.0454[0.0427,0.0481]	Yes

LLMs, denoted as UILLLM-1B/3B/8B-wo-tuning. Tables VII and VIII show that instruction tuning consistently improves performance, with the largest gains on the small-scale TWFQ, and the tuned 8B model reaching ceiling-level results. On TWIN, different scales exhibit complementary strengths, indicating that model size should be selected based on data volume and task requirements. Overall, instruction tuning is crucial for both performance and instruction-following ability.

D. Generalization Performance Analysis

UIL models may overfit superficial cues such as usernames, which degrades generalization, especially on TWFQ where positive pairs often share identical usernames. To measure this effect, we evaluate models under a perturbed setting that randomly replaces usernames in positive samples and breaks username consistency. As shown in Table IX, models trained on the original data suffer substantial performance drops, with accuracy decreasing to 86.31% on synthetic TWIN and 53.27% on synthetic TWFQ.

To mitigate this issue, we apply username replacement to a proportion λ of positive training samples. As shown in Fig. 8c and Fig. 8d, moderate augmentation ratios yield the best trade-off by preserving performance on the original data while markedly improving robustness under perturbation. These results indicate that controlled username perturbation reduces reliance on spurious identity cues and enhances generalization.

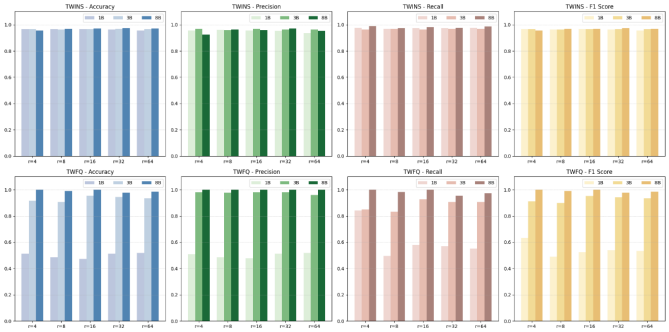


Fig. 9: The performance at different ranks and model scales.

E. Sensitivity to hyper-parameter and model scale

We vary the QLoRA rank to assess robustness to PEFT hyperparameters. As shown in Fig. 9, performance on TWIN remains stable across different ranks and model scales, while on TWFQ rank variations have limited impact and larger Llama backbones consistently yield better results under fixed ranks. This contrast suggests that larger datasets such as TWIN are more robust to low-rank adaptation, whereas smaller datasets like TWFQ benefit more from increased model capacity.

F. Cross-Dataset Transferability Evaluation

To evaluate generalizability, we conduct cross-dataset transfer experiments by training UILLLM on one platform and testing it on another. As shown in Table X, UILLLM demonstrates strong transferability, and UILLLM-3B trained on TWIN achieves an accuracy of 0.9844 on TWFQ, outperforming the model trained directly on TWFQ and suggesting that TWIN provides more generalizable features. Model scale plays a critical role, as UILLLM-8B transfers effectively, whereas UILLLM-1B exhibits limited cross-domain performance. Overall, these results confirm the robustness of UILLLM under distribution shifts.

TABLE V: Results for UILLLM and UIL-wo-PSC on TWIN.

Methods	Accuracy	Precision	Recall	F1
UILLLM-1B-wo-PSC	0.4716 ± 0.0021	0.4602 ± 0.0018	0.4738 ± 0.0023	0.4669 ± 0.0020
UILLLM-1B	$\mathbf{0.9640 \pm 0.0012}$	$\mathbf{0.9608 \pm 0.0014}$	$\mathbf{0.9656 \pm 0.0011}$	$\mathbf{0.9632 \pm 0.0013}$
UILLLM-3B-wo-PSC	0.4900 ± 0.0022	0.4819 ± 0.0020	0.5885 ± 0.0027	0.5299 ± 0.0024
UILLLM-3B	$\mathbf{0.9592 \pm 0.0012}$	$\mathbf{0.9472 \pm 0.0010}$	$\mathbf{0.9705 \pm 0.0013}$	$\mathbf{0.9587 \pm 0.0012}$
UILLLM-8B-wo-PSC	0.4884 ± 0.0022	0.4784 ± 0.0020	0.5262 ± 0.0024	0.5012 ± 0.0022
UILLLM-8B	$\mathbf{0.9552 \pm 0.0014}$	$\mathbf{0.9248 \pm 0.0017}$	$\mathbf{0.9885 \pm 0.0010}$	$\mathbf{0.9556 \pm 0.0014}$

TABLE VI: Results for UILLLM and UIL-wo-PSC on TWFQ.

Methods	Accuracy	Precision	Recall	F1
UILLLM-1B-wo-PSC	0.5093 ± 0.0025	0.5156 ± 0.0026	0.3084 ± 0.0015	0.3860 ± 0.0019
UILLLM-1B	$\mathbf{0.5327 \pm 0.0025}$	$\mathbf{0.5193 \pm 0.0025}$	$\mathbf{0.8785 \pm 0.0032}$	$\mathbf{0.6528 \pm 0.0028}$
UILLLM-3B-wo-PSC	0.5187 ± 0.0025	0.5238 ± 0.0026	0.4112 ± 0.0019	0.4607 ± 0.0022
UILLLM-3B	$\mathbf{0.9766 \pm 0.0030}$	$\mathbf{0.9722 \pm 0.0019}$	$\mathbf{0.9813 \pm 0.0033}$	$\mathbf{0.9767 \pm 0.0028}$
UILLLM-8B-wo-PSC	0.5280 ± 0.0026	0.5160 ± 0.0025	0.9065 ± 0.0031	0.6576 ± 0.0029
UILLLM-8B	$\mathbf{0.9914 \pm 0.0031}$	$\mathbf{0.9892 \pm 0.0036}$	$\mathbf{0.9920 \pm 0.0029}$	$\mathbf{0.9906 \pm 0.0032}$

TABLE VII: Results for UILLLM and UIL-wo-Tuning on TWIN.

Methods	Accuracy	Precision	Recall	F1
UILLLM-1B-wo-tuning	0.5020 ± 0.0025	0.4694 ± 0.0023	0.1508 ± 0.0015	0.2283 ± 0.0018
UILLLM-1B	$\mathbf{0.9656 \pm 0.0012}$	$\mathbf{0.9551 \pm 0.0014}$	$\mathbf{0.9754 \pm 0.0011}$	$\mathbf{0.9651 \pm 0.0013}$
UILLLM-3B-wo-tuning	0.5364 ± 0.0026	0.5512 ± 0.0027	0.2738 ± 0.0019	0.3658 ± 0.0021
UILLLM-3B	$\mathbf{0.9664 \pm 0.0012}$	$\mathbf{0.9686 \pm 0.0011}$	$\mathbf{0.9623 \pm 0.0013}$	$\mathbf{0.9655 \pm 0.0012}$
UILLLM-8B-wo-tuning	0.6629 ± 0.0029	0.7962 ± 0.0031	0.4164 ± 0.0023	0.5468 ± 0.0026
UILLLM-8B	$\mathbf{0.9552 \pm 0.0014}$	$\mathbf{0.9248 \pm 0.0017}$	$\mathbf{0.9885 \pm 0.0010}$	$\mathbf{0.9556 \pm 0.0014}$

TABLE VIII: Results for UILLLM and UIL-wo-Tuning on TWFQ.

Methods	Accuracy	Precision	Recall	F1
UILLLM-1B-wo-tuning	0.4907 ± 0.0024	0.4762 ± 0.0023	0.1869 ± 0.0016	0.2685 ± 0.0019
UILLLM-1B	$\mathbf{0.5327 \pm 0.0025}$	0.5193 ± 0.0025	$\mathbf{0.8785 \pm 0.0032}$	$\mathbf{0.6528 \pm 0.0028}$
UILLLM-3B-wo-tuning	0.5514 ± 0.0027	0.6000 ± 0.0028	0.3084 ± 0.0020	0.4074 ± 0.0022
UILLLM-3B	$\mathbf{0.9766 \pm 0.0030}$	$\mathbf{0.9722 \pm 0.0019}$	$\mathbf{0.9813 \pm 0.0033}$	$\mathbf{0.9767 \pm 0.0028}$
UILLLM-8B-wo-tuning	0.6729 ± 0.0030	0.9111 ± 0.0032	0.3832 ± 0.0022	0.5395 ± 0.0026
UILLLM-8B	$\mathbf{0.9914 \pm 0.0031}$	$\mathbf{0.9892 \pm 0.0036}$	$\mathbf{0.9920 \pm 0.0029}$	$\mathbf{0.9906 \pm 0.0032}$

TABLE IX: The impact of username inconsistency on the accuracy of UILLLM-8B.

Dataset	Original Dataset	Synthetic dataset
TWIN	0.9552	0.8631
TWFQ	0.9914	0.5327

TABLE X: Cross-dataset transferability results.

Methods	Train	Test	Accuracy
UILLLM-1B	TWIN	TWIN	0.9656 ± 0.0012
		TWFQ	0.9607 ± 0.0018
	TWFQ	TWFQ	0.5140 ± 0.0032
		TWIN	0.4964 ± 0.0015
UILLLM-3B	TWIN	TWIN	0.9664 ± 0.0014
		TWFQ	0.9844 ± 0.0024
	TWFQ	TWFQ	0.9159 ± 0.0030
		TWIN	0.8990 ± 0.0012
UILLLM-8B	TWIN	TWIN	0.9552 ± 0.0014
		TWFQ	0.9574 ± 0.0024
	TWFQ	TWFQ	0.9914 ± 0.0031
		TWIN	0.9339 ± 0.0034

V. Conclusion

We propose UILLLM, a framework that combines PSC-based user modeling with two-phase instruction fine-tuning to make LLMs practical for robust and efficient user identity linkage. UILLLM distills discriminative signals from noisy UGC and leverages targeted data augmentation to stabilize model reasoning and improve generalization across platforms. Beyond achieving strong empirical performance, our study shows that instruction-tuned LLMs can reliably solve structured UIL tasks rather than relying on brittle surface cues, and it provides a scalable foundation for future research on cross-platform adaptation and lightweight prompt design.

References

[1] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," ACM SIGKDD Explor. Newsl., vol. 18, no. 2, pp. 5–17, 2017.

[2] Y. L. Lee, T. Zhou, K. Yang, Y. Du, and L. Pan, "Personalized recommender systems based on social relationships and historical behaviors," Appl. Math. Comput., vol. 437, Art. no. 127549, 2023.

[3] S. Li, C. Zhao, Q. Li, J. Huang, D. Zhao, and P. Zhu, "BotFinder: A novel framework for social bots detection in online social networks based on graph embedding and community detection," World Wide Web, vol. 26, no. 4, pp. 1793–1809, 2023.

[4] S. Li, G. Zhong, Y. Jin, X. Wu, P. Zhu, and Z. Wang, "A deceptive reviews detection method based on multidimensional feature construction and ensemble feature selection," IEEE Trans. Comput. Soc. Syst., vol. 10, no. 1, pp. 153–165, 2022.

[5] Y. Ge, W. Hua, J. Ji, J. Tan, S. Xu, and Y. Zhang, "OpenAGI: When LLM meets domain experts," in Proc. Adv. Neural Inf. Process. Syst., vol. 36, pp. 5539–5568, 2023.

[6] Z. Gao, H. Wang, Y. Zhou, W. Zhu, and C. Zhang, "How far have we gone in vulnerability detection using large language models," CoRR, abs/2311.12420, 2023.

[7] C. Zhang, H. Liu, J. Zeng, K. Yang, Y. Li, and H. Li, "Prompt-enhanced software vulnerability detection using ChatGPT," in Proc. 46th IEEE/ACM Int. Conf. Softw. Eng. (ICSE Companion), pp. 276–277, 2024.

[8] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," CoRR, abs/2308.10792, 2023.

[9] X. Chen, X. Song, G. Peng, S. Feng, and L. Nie, "Adversarial-enhanced hybrid graph network for user identity linkage," in Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 1084–1093, 2021.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778, 2016.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint, arXiv:1412.3555, 2014.

[13] X. Chen, X. Song, S. Cui, T. Gan, Z. Cheng, and L. Nie, "User identity linkage across social media via attentive time-aware user modeling," IEEE Trans. Multimedia, vol. 23, pp. 3957–3967, 2020.

[14] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997.

[15] S. Li, D. Lu, Q. Li, X. Wu, S. Li, and Z. Wang, "MFLink: User identity linkage across online social networks via multimodal fusion and adversarial learning," IEEE Trans. Emerg. Topics Comput. Intell., vol. 8, no. 5, pp. 3716–3725, 2024.

[16] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," Int. J. Mach. Learn. Cybern., vol. 1, no. 1, pp. 43–52, 2010.

[17] Y. Li, G. Gou, G. Xiong, Z. Li, and M. Cui, "The potential utility of image descriptions: User identity linkage across social networks based on multimodal self-attention fusion," in Proc. IEEE IPCCC, pp. 265–273, 2023.

[18] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn., pp. 12888–12900, 2022.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 11976–11986, 2022.

[20] B. Chen and X. Chen, "MAUIL: Multilevel attribute embedding for semisupervised user identity linkage," Inf. Sci., vol. 593, pp. 527–545, 2022.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint, arXiv:1301.3781, 2013.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

[23] D. I. Warton, "Penalized normal likelihood and ridge regularization of correlation and covariance matrices," J. Amer. Statist. Assoc., vol. 103, no. 481, pp. 340–349, 2008.

[24] X. Xie, H. Guo, Y. Lu, and T. Zhang, "Adap-UIL: A multi-feature-aware user identity linkage framework based on an adaptive graph walk," Applied Sciences, vol. 15, no. 12, Art. no. 6762, 2025.

[25] Y. Peng, X. Chen, D. Miao, H. Zhang, X. Qin, S. Du, and P. Lu, "DECTUIL: Cross-social network user identity linkage via dynamic embedding and clustering model driven by three-way decision," Expert Systems with Applications, Art. no. 129026, 2025.

[26] V. Lialin, V. Deshpande, and A. Rumshisky, "Scaling down to scale up: A guide to parameter-efficient fine-tuning," arXiv preprint, arXiv:2303.15647, 2023.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., "LoRA: Low-rank adaptation of large language models," in Proc. Int. Conf. Learn. Represent., p. 3, 2022.

[28] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Adv. Neural Inf. Process. Syst., vol. 36, pp. 10088–10115, 2023.

[29] Y. Zhang, C. Li, X. Wang, and B. Zhang, "Application and research on a large model training method based on instruction fine-tuning in domain-specific tasks," in Proc. CCF Conf. Big Data, pp. 181–194, 2023.

[30] X. Du, M. Wen, J. Zhu, Z. Xie, B. Ji, H. Liu, et al., "Generalization-enhanced code vulnerability detection via multi-task instruction fine-tuning," arXiv preprint, arXiv:2406.03718, 2024.

[31] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, et al., "Code Llama: Open foundation models for code," arXiv preprint, arXiv:2308.12950, 2023.

[32] J. Zhang, R. Xie, Y. Hou, X. Zhao, L. Lin, and J.-R. Wen, "Recommendation as instruction following: A large language model empowered recommendation approach," ACM Trans. Inf. Syst., early access, 2023.

[33] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., "The LLaMA 3 herd of models," arXiv preprint, arXiv:2407.21783, 2024.

[34] B. H. Lim, D. Lu, T. Chen, and M. Y. Kan, "#mytweet via Instagram: Exploring user behaviour across multiple social networks," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, pp. 113–120, 2015.

[35] Y. Ren, C. C. Aggarwal, and J. Zhang, "ActiveIter: Meta diagram based active learning in social networks alignment," IEEE Trans. Knowl. Data Eng., vol. 33, no. 5, pp. 1848–1860, 2019.

[36] J. Gao, K. Zheng, X. Wang, C. Wu, and B. Wu, "Efficient user identity linkage based on aligned multimodal features and temporal correlation," Comput. Mater. Contin., vol. 81, no. 1, 2024.

[37] R. Huang, T. Ma, H. Rong, K. Huang, N. Bi, P. Liu, and T. Du, "Topic and knowledge-enhanced modeling for edge-enabled IoT user identity linkage across social networks," J. Cloud Comput., vol. 13, no. 1, Art. no. 107, 2024.

[38] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, and C. Guestrin, et al., "Alpaca: a strong, replicable instruction-following model," Stanford Center for Research on Foundation Models, Stanford, CA, USA, 2023.