GATS: Generative Audience Targeting System for Online Advertising

Cong Jiang
Ant Group
Hangzhou, China
jiangcong.jc@antgroup.com

Zhongde Chen
Ant Group
Hangzhou, China
zhongde.czd@antgroup.com

Bo Zhang
Ant Group
Hangzhou, China
aicui.zb@antgroup.com

Yankun Ren
Ant Group
Hangzhou, China
yankun.ryk@antgroup.com

Xin Dong
Ant Group
Hangzhou, China
zhaoxin.dx@antgroup.com

Lei Cheng
Ant Group
Hangzhou, China
lei.chenglei@antgroup.com

Xinxing Yang
Ant Group
Hangzhou, China
xinxing.yangxx@antfin.com

Longfei Li
Ant Group
Hangzhou, China
longyao.llf@antfin.com

Jun Zhou
Ant Group
Hangzhou, China
jun.zhoujun@antfin.com

Linjian Mo*
Ant Group
Hangzhou, China
linyi01@antgroup.com

ABSTRACT

This paper presents GATS (Generative Audience Targeting System for Online Advertising), a new framework using large language models (LLMs) to improve audience targeting in online advertising. GATS overcomes the shortcomings of rule-based, look-alike, and graph-based methods by facilitating flexible and interpretable audience criteria expression. The framework integrates intent recognition, knowledge mining, and Data Management Platform (DMP) mapping to translate advertiser demands into actionable user tags and correlate them within a DMP. A small, white-box model called LightGATS (base on QWen-14B), fine-tuned with a high-quality LLM corpus, ensures the framework's safety and efficiency, operating within a scalable hybrid online-offline architecture. GATS's effectiveness is validated through extensive experiments, marking a significant advancement in audience targeting technology.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

Online Advertising; Audience Targeting; Large Language Models; Multi-task Fine-tuning

*Corresponding author.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3661372

ACM Reference Format:
Cong Jiang, Zhongde Chen, Bo Zhang, Yankun Ren, Xin Dong, Lei Cheng, Xinxing Yang, Longfei Li, Jun Zhou, and Linjian Mo. 2024. GATS: Generative Audience Targeting System for Online Advertising. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3661372

1 INTRODUCTION

Nowadays, online advertising [3, 8, 22] plays a crucial role in the profit structure of many Internet companies. Effective audience targeting in online advertising ensures advertisers can accurately engage their target market, leading to higher conversion efficiency and lower ad deployment costs.

Current approaches devoted to audience targeting mainly fall into three lines, each with inherent drawbacks. The first type denotes the rule-based methods [1, 16, 17], which relies heavily on domain expertise and it often results in suboptimal outcomes due to the complexity and static nature of the targeting labels. As a comparison, the look-alike based methods [6, 15, 23, 24] seek to learn high-quality representations of seed users, and the target users can be effectively matched in the embedding space. However, these methods suffer from the following major issues: the unavailability of seed users for newcoming crowd demands and the unfriendliness of black-box procedures towards advertisers. More recently, graph-based methods [4, 20, 21] match target users interested in ads with efficient cognitive reasoning over well-established entity graphs. Nevertheless, this method is hampered by high update costs, latency issues, and inability to swiftly adapt to new knowledge or user-expressed demands that fall outside existing graph entities.

To address these issues, we propose a novel framework named GATS (Generative Audience Targeting System for Online Advertising), which is based on the advanced semantic comprehension and generative capabilities of large language models (LLMs) [2, 11, 18]. This framework empowers advertisers to convey their audience criteria with ease and flexibility, coupled with the benefit of clarity in interpretation, which in turn facilitates the refinement of targeting strategies upon reviewing the outcomes.

Specifically, the GATS framework comprises three integral modules: intent recognition, knowledge mining, and DMP mapping. The intent recognition module processes the demands of advertisers and translates them into basic user profile tags (e.g., gender, age, and occupation) and entity tags indicative of user interests. Knowledge mining expands upon entity tags by surfacing similar entities that align with the advertiser's targeting criteria. The DMP mapping module then correlates all tags with existing tag values within the advertising Data Management Platform (DMP) [7, 10] to curate target users. Due to safety and efficiency concerns, each module in our framework is carefully prompted by a small and white-box audience targeting model (dubbed as LightGATS), which is fine-tuned with a high-quality corpus provided by a strong teacher-LLM involving 1.8 trillion parameters [9]. We subsequently follow the hybrid online-offline architecture to satisfy the requirements of scalability and timeliness, amalgamating extensive offline entity knowledge mining with an online process for intent recognition and DMP mapping.

In summary, the contributions of our paper are threefold:

- To our knowledge, GATS is the first generative audience targeting framework for online advertising, which empowers advertisers to convey their audience criteria with ease and flexibility over LLMs.
- This approach yields a comprehensive solution that spans intent recognition, knowledge mining, and DMP mapping , which follows the hybrid online-offline architecture prompted by a small and white-box audience targeting model.
- We demonstrate the superiority of the proposed GATS through extensive experiments and practical applications in audience targeting.

2 PROPOSED FRAMEWORK

As illustrated in Figure 1, this section provides a detailed explanation of the GATS system's operating pipeline. It covers the overall workflow, task modules (intent recognition, knowledge mining and DMP alignment), and the shared LLM across these modules, which we temporarily name LightGATS for distinction.

2.1 Intent Recognition

In this work, we present a dual-agent system for enhancing intent recognition in conversational interfaces, aligned with ChatIE's approach [19]. The first agent identifies user-referenced slots, covering demographics and behavioral interests, while the second agent fills these slots with user-stated needs.

The process is represented by the equation:

$$p((s,r,o))|x,q) = p(r|x,q_1)p((s,o)|q_r), \tag{1}$$

where model predicts triples $T(x) = (s_1, r_1, o_1), \ldots, (s_n, r_n, o_n)$ from sentence x and question q. Here q_1 and q_r are the first and second phase questions, respectively.

This segmentation yields a more precise intent recognition, confirming that two-stage methods outperform single-turn LLM prompts in tasks like IE, as noted in Alipay's ad scene and supported by ChatIE [19].

2.2 Knowledge Mining

We introduce a knowledge mining module to better capture advertisers' target audience intentions, addressing expressions that are often vague and lack context. The module encompasses a priori knowledge introduction along with multi-round mining. The mathematical framework is given by:

$$p(k,r,t|s) = p(k|s)p(r|s,k)p(t|s,k,r), \tag{2}$$

where, s is the target entity, k is prior knowledge, r is the expansion perspective, and t is the extended knowledge.

RAG & Dynamic Prompt: We enhance retrieval using a bespoke knowledge graph to counter LLMs' limited entity understanding and intrinsic knowledge. This pre-structured knowledge informs the LLM's initial expansions.

Chain of Thought: Initially, the LLM identifies expansion directions that are based on target entities and informed by prior knowledge, while it focuses on relevant dimensions. This scope limitation increases efficiency and reduces spurious outputs. In subsequent stages, the LLM builds on these directions, recursively expanding until reaching saturation or limits, showcasing the synergy between structured knowledge graphs and neural language models' learning capabilities.

2.3 DMP Mapping

Our work introduces an efficient method to enhance Data Management Platforms (DMPs) through advanced Intent Recognition and precise entity linking. Our approach aligns key user and entity tags with the corresponding DMP labels to target specific audience segments.

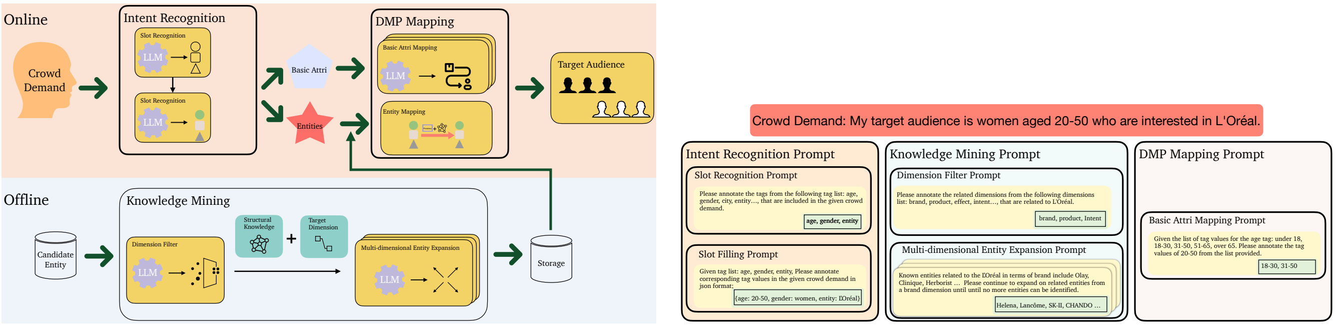
We start by applying Intent Recognition algorithms to analyze user interactions and allocate specific atomic labels in Alipay's DMP ecosystem [20]. Entity selection is refined in two stages, combining vector retrieval and semantic similarity.

To improve entity linking accuracy, we integrate a Language Model-based Annotation Agent (LLM) into our system. This agent ensures the consistency between selected entities and user intents based on DMP labels. Upon validation, the LLM finalizes the DMP setup, thereby optimizing audience interaction.

2.4 LightGATS

As illustrated in Figure 2, with the objective of achieving a lightweight architecture, we endeavor to enhance system performance through the dual-pronged approach of optimizing both the training and the inference processes. To distinguish the core large language model from the entire system, we refer to it as LightGATS within the framework.

2.4.1 ***Multi-task Balanced Fine-Tuning Strategy.*** During the training phase, we employ a multi-task fine-tuning strategy, wherein



(a) GATS architecture (b) Agent prompts example

Figure 1: The figure provides an overview of GATS' overall architecture (left), along with examples of prompt words for each LLM Agent (right). It illustrates that GATS comprises both online and offline workflows, and features 5 fine-tuned LLMs that play key roles in the system.

a base large language model is fine-tuned on several tasks data in parallel. Simultaneously, a harmonized training paradigm is adopted, which guarantees the fine-tuning conducted between tasks is mutually non-disruptive. This approach not only enhances the efficiency of model training but also guarantees the efficacy of the fine-tuned model across the spectrum of individual sub-tasks.

To this end, we use an open-source multi-task fine-tuning framework, MFTCoder [13], that enables simultaneous and parallel fine-tuning on multiple tasks. MFTCoder integrates an array of meticulously designed loss functions, thereby efficaciously surmounting ubiquitous challenges endemic to multi-task learning, such as data imbalance, varying difficulty levels, and inconsistent convergence speeds.

Specifically, to redress the prevalent issue of data imbalance within multi-task learning environments, MFTCoder meticulously ensures that each datum from the entirety of tasks is leveraged precisely once within the confines of an individual epoch. In an effort to preclude any potential bias wherein the model might disproportionately favor tasks with more abundant data, MFTCoder implements a sagacious weight distribution strategy in its loss computation algorithm, contingent on the number of valid tokens implicated in the loss evaluation process. The above weighted loss calculation is shown in Eq.3.

$$Loss_1(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M_i}\sum_{k=1}^{T_{ij}} - \log(p_\theta(t_{ijk}))\right) \Big/ \left(\sum_{j=1}^{M_i} T_{ij}\right). \tag{3}$$

Confronting the complexity engendered by task heterogeneity, MFTCoder deftly incorporates a focal loss [12] paradigm, executing dual-tiered focal loss functions explicitly designed to provide nuanced solutions attuned to distinct levels of granularity.

$$Loss_2(\theta) = \left(\sum_{i=1}^{N}\sum_{j=1}^{M_i} - \alpha_i * (1 - P_{ij})^{\gamma} * Q_{ij}\right) \Big/ \left(\sum_{i=1}^{N} M_i\right). \tag{4}$$

To surmount the challenge of inconsistent convergence velocities, MFTCoder has assimilated the FAMO [14] methodology, applying it in an innovative fashion for the computation of validation loss, which serves to normalize convergence trajectories and bolster the robustness of the model's optimization process.

$$Loss_3(\theta) = \frac{1}{N}\sum_{i=1}^{N} -\alpha_i * (1 \; - \; P_i)^{\gamma} * Q_i. \tag{5}$$

Moreover, we also employ QLoRA [5], a parameter-efficient fine-tuning technology, that effectively reduces the computing resources required for training through quantization and matrix low-rank adaptation technology. Compared with traditional fine-tuning methods, the speed is significantly improved.

2.4.2 *Hybrid Online-Offline Architecture.*

In contemplation of the latency issues associated with the deployment of large-scale language models in real-world application scenarios, we have adopted a hybrid online-offline architecture to optimize the inference process. This architecture is specifically designed to pre-emptively address a multitude of high-frequency crowd demands offline, thereby caching the intermediate outcomes at each stage of the model's operation within intent recognition, knowledge mining, and DMP mapping. Therefore, this strategy considerably mitigates the necessity for online real-time inference, which not only augments the stability of the system but also accelerates the throughput in addressing high-frequency crowd demands.

In application scenarios, when receiving crowd demands, the system first compares whether the crowd demand exists in the cache through the text similarity method. If it exists, the cached inference result is directly returned, otherwise it is handed over to the large language model for real-time inference.

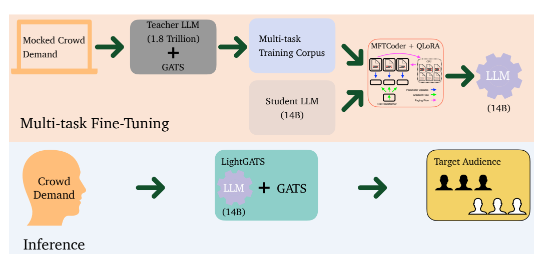


Figure 2: The multi-task fine-tuning and inference pipeline of LightGATS (shared fine-tuned LLM: QWen-14B (+CoT, +MFT, +QLoRA).

Table 1: Performance comparison on offline datasets.

Backbone Model	Methods	Knowledge Mining F1	Intent Recognition F1
GPT-3.5	+CoT	0.3248	0.7590
AntGLM-10B	-	0.0734	0.3432
AntGLM-10B	+CoT	0.2939	0.5988
AntGLM-10B	+CoT, +MFT, +FPT	0.2968	0.7329
QWen-14B	+CoT, +FT, +FPT	0.2029	0.5736
QWen-14B	+CoT, +MFT, +LoRA	0.2932	0.7754
QWen-14B	**+CoT, +MFT, +QLoRA**	**0.3327**	**0.8313**

3 EXPERIMENTS

We executed offline evaluations of GATS utilizing a real-world industrial dataset, and also report online results after GATS has been deployed in Alipay.

3.1 Experimental Setup

3.1.1 ***Datasets.*** Our study tackles user targeting in advertising, utilizing standard datasets custom-built by Alipay. The system's attribute extrapolation accuracy is gauged using Alipay's *Knowledge Mining Evaluation Set*, while user intent comprehension is measured with the *Intent Recognition Evaluation Set*. These datasets are pivotal for honing our algorithms and refining targeted advertising strategies.

3.1.2 ***Evaluation metrics.*** Based on the standards set mentioned above, we have selected the F1 score as our evaluation metric, which integrates both precision and recall.

3.2 Effectiveness of GATS

Table 1 summarizes the experimental results of the main iteration versions of the GATS model on the knowledge mining and intent recognition evaluation sets. Considering factors such as data security, we only consider using the proprietary pre-trained LLM developed by Ant Group (AntGLM) and Alibaba Group (QWen) as backbone models, and we conduct a performance comparison with GPT-3.5 under the same conditions.

The state-of-the-art (SOTA) version of GATS, which incorporates the QWen-14B model enhanced with techniques such as Chain of Thought (CoT), Multi-Task Finetuning (MFT), and Q-Logits Rank Regularization (QLoRA), has surpassed GPT-3.5 in the tasks of intent recognition and knowledge mining.

Further details refer to Table 1. In the table, FT, MFT, FPT, and CoT stand for finetuning, multi-task finetuning, full parameters training (FPT), and chain of thought, respectively.

3.3 Online Performance of GATS

The GATS has been successfully implemented in Alipay's production environment, where it serves a wide range of industries by optimizing their advertising campaigns. To evaluate GATS's effectiveness, we have carried out online A/B testing in real traffic. The results of these experiments are detailed in Table 2.

It's important to note that the objective of the audience targeting task is to identify users who are most likely to be interested in the

Table 2: Online experiment performance.

Industry	Exposure	CTR	CVR
3C Rental	+0.10%	+17.35%	+28.18%
Automotive	+0.17%	+10.23%	+19.47%
Education	-0.15%	+5.85%	+9.62%
E-commerce	+0.34%	+13.34%	+16.52%

advertisements. An increase in Click-Through Rate (CTR) and Conversion Rate (CVR) signifies a more precise selection of potential customers. In Table 2, we present the performance improvement of GATS in comparison to the previous online baseline method. The data indicates a significant enhancement in both CTR and CVR, affirming the efficacy of the GATS approach.

3.4 Application Case

Figure 3 illustrates a real case of the GATS system as applied to online advertising on Alipay. An advertiser specifies his crowd demand by inputting: "My target audience is women aged 20-50 who are interested in L'Oréal." along with a expected audience size: "1000000 100000000".

Upon submission, GATS recognizes the intent as follows: "Gender: Female, Age: 18-30, 31-50, Interest Entity: L'Oréal." It then proceeds to expand more entities to encompass related brands, products, and intents associated with L'Oréal. Meanwhile, the advertiser can select "More" to explore additional related entities. Finally, by clicking on "Create" the request is submitted to the DMP backend to generate the targeted audience.

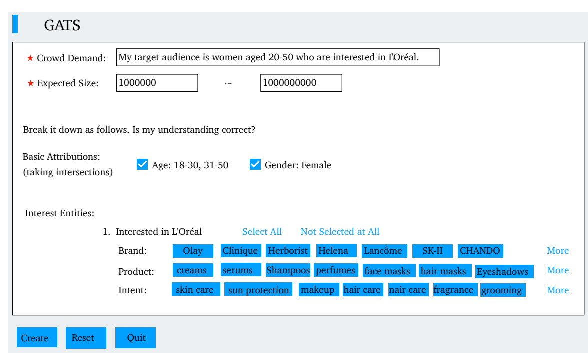


Figure 3: A real case of audience targeting using GATS in Alipay's online advertising

4 CONCLUSION

This paper presents the Generative Audience Targeting System (GATS), an innovative framework that improves online ad targeting using large language models (LLMs). GATS overcomes traditional targeting constraints, enabling flexible, intelligible audience specification. It integrates intent recognition and knowledge mining with Data Management Platforms (DMPs) to translate advertiser criteria into precise DMP user tags. The LightGATS mini-model, honed with an extensive LLM corpus, ensures safety and efficiency in a scalable hybrid structure. GATS demonstrates notable advances in audience targeting, validated by thorough experiments. Future integration of multi-turn dialogue within GATS could revolutionize interactive advertising, offering a smarter, user-centric experience.

REFERENCES

[1] Alan S Abrahams, Eloise Coupey, Eva X Zhong, Reza Barkhi, and Pete S Manasantivongs. 2013. Audience targeting by B-to-B advertisement classification: A neural network approach. *Expert systems with applications* 40, 8 (2013), 2777–2791.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] G Anusha. 2016. Effectiveness of online advertising. *International journal of research–Granthaalayah* 4, 3 (2016), 14–21.

[4] Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. *arXiv preprint arXiv:2301.12810* (2023).

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Stephanie deWet and Jiafan Ou. 2019. Finding users who act alike: transfer learning for expanding advertiser audiences. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2251–2259.

[7] Hazem Elmeleegy, Yinan Li, Yan Qi, Peter Wilmot, Mingxi Wu, Santanu Kolay, Ali Dasdan, and Songting Chen. 2013. Overview of turn data management platform for digital advertising. *Proceedings of the VLDB Endowment* 6, 11 (2013), 1138–1149.

[8] David S Evans. 2008. The economics of the online advertising industry. *Review of network economics* 7, 3 (2008).

[9] Chunjing Gan, Dan Yang, Binbin Hu, Ziqi Liu, Yue Shen, Zhiqiang Zhang, Jinjie Gu, Jun Zhou, and Guannan Zhang. 2023. Making Large Language Models Better Knowledge Miners for Online Marketing with Progressive Prompting Augmentation. *arXiv preprint arXiv:2312.05276* (2023).

[10] Vaibhav Gardé. 2018. Digital audience management: Building and managing a robust data management platform for multi-channel targeting and personalisation throughout the customer journey. *Applied Marketing Analytics* 4, 2 (2018), 126–135.

[11] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169* (2023).

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[13] Bingchang Liu, Chaoyu Chen, Cong Liao, Zi Gong, Huan Wang, Zhichao Lei, Ming Liang, Dajun Chen, Min Shen, Hailian Zhou, et al. 2023. MFTCoder: Boosting Code LLMs with Multitask Fine-Tuning. *arXiv preprint arXiv:2311.02303* (2023).

[14] Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. 2024. FAMO: Fast Adaptive Multitask Optimization. *Advances in Neural Information Processing Systems* 36 (2024).

[15] Qiang Ma, Eeshan Wagh, Jiayi Wen, Zhen Xia, Robert Ormandi, and Datong Chen. 2016. Score look-alike audiences. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 647–654.

[16] Ashish Mangalampalli, Adwait Ratnaparkhi, Andrew O Hatch, Abraham Bagherjeiran, Rajesh Parekh, and Vikram Pudi. 2011. A feature-pair-based associative classification approach to look-alike modeling for conversion-oriented user-targeting in tail campaigns. In *Proceedings of the 20th international conference companion on World wide web*. 85–86.

[17] Jianqiang Shen, Sahin Cem Geyik, and Ali Dasdan. 2015. Effective audience extension in online advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2099–2108.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[19] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205* (2023).

[20] Dan Yang, Binbin Hu, Xiaoyan Yang, Yue Shen, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023. Who Would be Interested in Services? An Entity Graph Learning System for User Targeting. *arXiv preprint arXiv:2305.18780* (2023).

[21] Xiaoling Zang, Binbin Hu, Jun Chu, Zhiqiang Zhang, Guannan Zhang, Jun Zhou, and Wenliang Zhong. 2023. Commonsense Knowledge Graph towards Super APP and Its Applications in Alipay. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5509–5519.

[22] Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin. 2019. " Deep reinforcement learning for search, recommendation, and online advertising: a survey" by Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin with Martin Vesely as coordinator. *ACM sigweb newsletter* 2019, Spring (2019), 1–15.

[23] Yongchun Zhu, Yudan Liu, Ruobing Xie, Fuzhen Zhuang, Xiaobo Hao, Kaikai Ge, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to expand audience via meta hybrid experts and critics for recommendation and advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4005–4013.

[24] Chenyi Zhuang, Ziqi Liu, Zhiqiang Zhang, Yize Tan, Zhengwei Wu, Zhining Liu, Jianping Wei, Jinjie Gu, Guannan Zhang, Jun Zhou, et al. 2020. Hubble: An industrial system for audience expansion in mobile marketing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2455–2463.