Can Large Language Models Identify Meaningful Touchpoints in Conversion Attribution?

Jinqi Wu*
jinqiwu001@gmail.com
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
School of Intelligence Science and Technology, Nanjing University
Suzhou, China

Sishuo Chen*
Zhangming Chan
Yong Bai
chensishuo@pku.edu.cn
zhangming.czm@alibaba-inc.com
baiyong.by@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Chao Yi
Han Zhu
Shuodian Yu
yunan.yc@alibaba-inc.com
zhuhan.zh@alibaba-inc.com
yushuodian.ysd@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Lei Zhang
Sheng Chen
Chenghuan Hou
zl165646@alibaba-inc.com
chensheng.cs@alibaba-inc.com
jinyao@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

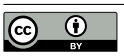
Jian Xu†
xiyu.xj@alibaba-inc.com
Taobao & Tmall Group of Alibaba
Beijing, China

Chaoyou Fu
bradyfu24@gmail.com
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
School of Intelligence Science and Technology, Nanjing University
Suzhou, China

Abstract

Touchpoint selection in conversion attribution, namely identifying meaningful touchpoints contributing to conversions, is essential for e-commerce recommendation and online advertising. Current selection methods rely heavily on collaborative-filtering-based heuristics, which fail to align with user-perceived semantic intent. Through human annotation, we reveal a significant **semantic gap**: many **implicitly-related, semantically relevant touchpoints** remain undetected by existing rules. Therefore, we systematically evaluate the capability of Large Language Models (**LLMs**) in identifying these hidden associations. Our evaluation shows that while LLMs effectively uncover a substantial portion of implicitly-related touchpoints, significant room for improvement remains in their selection performance. Furthermore, we analyze the impact of different prompting strategies and foundation model choices on identification performance, providing valuable insights into their reasoning patterns and effectiveness. These insights offer a new roadmap for transitioning conversion attribution from mechanical rule-matching to human-aligned semantic reasoning. Moreover, we leverage the LLM-attributed conversion labels for enhancing industrial CVR model training and achieve significant offline performance gains, showing the potential of LLMs in conversion attribution.

*Equal contribution.
†Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3839867

CCS Concepts

• **Applied computing** → **Electronic commerce**; • **Information systems** → **Computational advertising**; • **Computing methodologies** → **Ranking**.

Keywords

Large Language Models, Conversion Attribution, E-Commerce Recommendation, Online Advertising

ACM Reference Format:
Jinqi Wu, Sishuo Chen, Zhangming Chan, Yong Bai, Chao Yi, Han Zhu, Shuodian Yu, Lei Zhang, Sheng Chen, Chenghuan Hou, Jian Xu, and Chaoyou Fu. 2026. Can Large Language Models Identify Meaningful Touchpoints in Conversion Attribution?. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3799682.3839867

1 Introduction

Conversion attribution, namely the mechanism allocating credits for a conversion across past user touchpoints, is a cornerstone for e-commerce platforms, which affects ad performance presentation to advertisers [10, 15, 19, 22], label generation for conversion rate (CVR) prediction models [4, 21, 23], and ultimately, platform revenue. Past studies on conversion attribution primarily focus on **(1) weight allocation** given a sequence of user touchpoints, which is usually restricted to the touchpoints under **the same item or shop** as the consumed product [4, 22]. Recently, Zeng et al. [23] from Meta found that **(2) touchpoint selection** is another important aspect of conversion attribution, which aims to identify meaningful touchpoints from **the full user behavior sequence**. Figure 1 illustrates the two key components of conversion attribution.

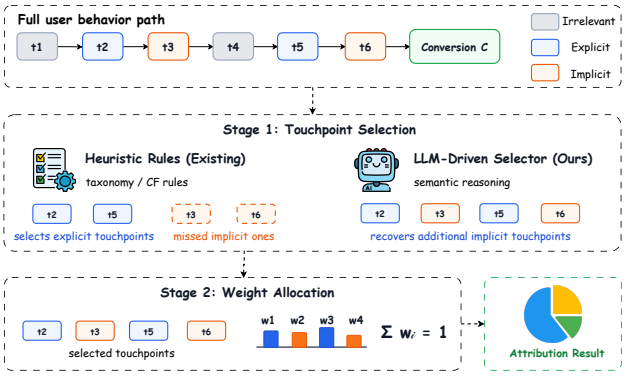


Figure 1: Touchpoint selection and weight allocation in conversion attribution.

Although Zeng et al. [23] revealed the value of full-path attribution for mining conversion signals and enhancing CVR prediction model training, their touchpoint selection relies on **heuristic rules** based on product taxonomy and collaborative-filtering (CF) signals, which may fail to capture **implicitly-related touchpoints** that are semantically related to the target product but lack explicit taxonomy or historical engagement overlaps. To verify this, we construct **SILVA**, the first benchmark for touchpoint **S**election **I**n fu**L**l-path con**V**ersion **A**ttribution, which is collected from a large-scale e-commerce platform. Human annotation results on **SILVA** reveal a striking fact: beyond the **explicitly-related** touchpoints covered by taxonomy and CF-based heuristics, *there exists a large number of **implicitly-related** touchpoints that are currently overlooked*. These findings underscore a critical deficiency in current conversion attribution approaches, which remain blind to a vast space of implicitly-related, semantically relevant touchpoints.

To bridge this gap and explore how to identify these implicitly-related touchpoints, we turn to Large Language Models (LLMs) for their superior semantic reasoning capabilities [20]. Specifically, we conduct a systematic evaluation of representative LLMs on SILVA, assessing their effectiveness in identifying both explicitly and implicitly-related touchpoints across various model scales and prompting strategies. We summarize the main findings as follows:

(1) **LLM Capability of Touchpoint Selection**: Cutting-edge LLMs are excellent at identifying explicitly-related touchpoints and effectively discover a substantial number of implicitly-related touchpoints, but substantial room for improvement remains in implicitly-related touchpoint selection.
(2) **Effect of Prompting Strategies**: Pairwise reasoning outperforms listwise reasoning, but the gap narrows when model capability increases.
(3) **Performance across LLM Families**: Proprietary LLMs, such as Gemini 3.1 Pro [8] and GPT 5.5 [17], only slightly outperform open-weight counterparts, such as GLM 5.1 [7] and DeepSeek V4-Pro [5]; smaller LLMs (<100B) largely lag behind flagship models.

Furthermore, inspired by these insights, we leverage conversion labels attributed by LLMs as auxiliary signals for training CVR prediction models, following Chen et al. [4] and Zeng et al. [23]. Results show that our approach, **L**LM-**O**riented **T**ouchpoint **U**nderstanding and **S**election (**LOTUS**), yields an absolute GAUC gain of **0.35 percentage points (pp)** over the production baseline (Base) [4] and a further gain of **0.15 pp** over the heuristic CABB competitor [23] for predicting the main conversion goal on a large-scale e-commerce platform with hundreds of millions of active users. We believe that our systematic analysis and findings open up a new avenue for harnessing LLMs for touchpoint selection in conversion attribution, which contributes to growth in user satisfaction and platform revenue for e-commerce platforms.

2 Related Work

Conversion attribution mechanisms, namely the rules allocating conversion credits among user touchpoints, play a vital role in e-commerce platforms [19]. Prior studies on conversion attribution focus on weight allocation among touchpoints via causal modeling [2, 6, 10–12, 15, 19, 22] and leveraging attribution results for training CVR prediction models [4, 21, 23]. However, almost all of them ignore **touchpoint selection** and only consider the touchpoints under the same item or shop as the consumed product by default. Zeng et al. [23] proposed to mine cross-item relevant touchpoints in a heuristic way based on collaborative-filtering signals and showed the value of touchpoint selection for enhancing the CVR prediction model training at Meta. Nonetheless, their simple heuristic selection strategies may overlook implicitly-related touchpoints. To our knowledge, our study takes *the first step to investigate the characteristics of implicitly-related touchpoints* and explore how to *harness LLMs for touchpoint selection* in conversion attribution.

3 The SILVA Benchmark

3.1 Problem Formulation

Touchpoint selection aims to identify whether a historical interaction contributed to a conversion. Formally, given a conversion C and its preceding touchpoints $\{t_1, t_2, \ldots, t_n\}$, we assign a label $l_i \in \{\text{Irrelevant, Explicit, Implicit}\}$ to each t_i. In the e-commerce context, meaningful touchpoints are categorized as:

- **Explicitly-Related**: Interactions sharing metadata-driven or statistical associations with C: (1) *Taxonomy Matching*, such as identical item, shop, brand, or leaf category; (2) *CF-based Similarity*, involving similar leaf categories determined by collaborative-filtering signals [23].
- **Implicitly-Related**: Touchpoints with semantic-driven associations that elude explicit rules, including: (1) *Functional Complementarity* (e.g., camera and memory card); and (2) *Scenario/Audience Alignment* (e.g., shared anime IPs).

3.2 Benchmark Construction and Analysis

We establish **SILVA**, the first benchmark for **S**election **I**n fu**L**l-path con**V**ersion **A**ttribution. Collected from an e-commerce platform with hundreds of millions of daily active users, **SILVA** comprises 1,000 unique conversion events and their preceding touchpoints within a 3-day lookback window, totaling 69,680 touchpoints. Each touchpoint is enriched with multimodal information about the clicked product, including its image, title, shop, and category metadata.

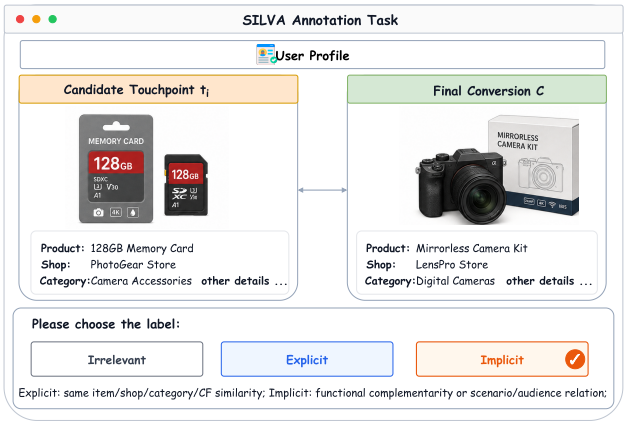


Figure 2: Schematic annotation interface for constructing the SILVA benchmark.

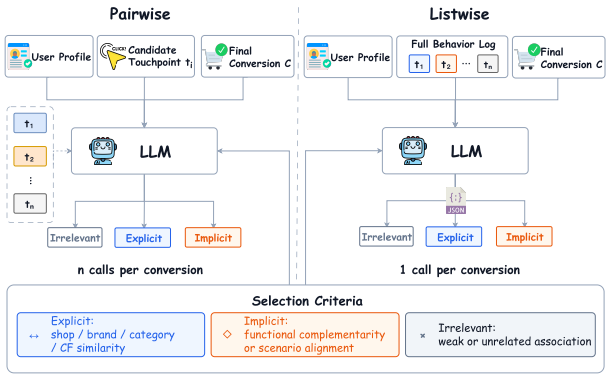


Figure 3: Schematic comparison of pairwise and listwise prompting protocols for LLM-based touchpoint selection.

Annotation reliability. For human annotation, we recruited six well-trained annotators to assign labels using the annotation interface shown in Figure 2. All annotators were fairly compensated according to local prevailing wage rates. Before formal annotation, annotators were trained with detailed guidelines and representative examples for explicit and implicit relevance. Importantly, annotators assigned labels solely according to the semantic annotation guidelines and were not shown the outputs of the taxonomy- or CF-based heuristic baselines. Each touchpoint was independently labeled by three annotators sampled from the pool of six annotators. We measure inter-annotator agreement using Fleiss' κ over the three-way label space, i.e., Irrelevant, Explicit, and Implicit. The annotations achieve Fleiss' $\kappa > 0.70$, indicating substantial agreement beyond chance; the final label is determined by majority voting among the three annotations.

Results reveal that explicitly-related touchpoints account for 19.09% of the dataset. Crucially, **implicitly-related touchpoints constitute a substantial 14.91%**, which elude traditional taxonomy and CF-based rules. These statistics underscore that **implicit signals are non-negligible** and represent a significant portion of the user's conversion journey currently overlooked by existing studies on conversion attribution [4, 22, 23]. To fill in the gap, we investigate the capability of LLMs in discovering these latent associations in § 4 and demonstrate the downstream value of these implicitly-related touchpoints for enhancing CVR prediction in § 5.

4 Harnessing LLMs for Touchpoint Selection

4.1 Experimental Setup

Foundation models. We evaluate eight representative LLMs: GPT-5.5 [17], Gemini-3.1-Pro [8], Claude-Opus-4.7 [1], GLM-5.1 [7], DeepSeek-V4-Pro and DeepSeek-V4-Flash [5], and Qwen3.5-27B and Qwen3.5-30B-A3B [18]. For models supporting switching thinking on or off, we use the thinking mode; otherwise, we use the default setting.

Prompting protocols. We compare pairwise and listwise prompting. In the **pairwise** protocol, the model receives the user profile, one candidate touchpoint, and the final conversion, and judges whether the touchpoint is irrelevant, explicitly related, or implicitly related. In the **listwise** protocol, the model receives the user profile, the full behavior log, and the final conversion, and returns the relevant touchpoints with their relation types. Figure 3 compares the two protocols. The same criteria are listed in the prompts in both protocols: explicit relevance is based on shop/brand matching, leaf-category matching, and CF-based similar-category matching, while implicit relevance covers functional complementarity and scenario or audience consistency.

Evaluation metrics. We evaluate each conversion separately and then macro-average across conversions. For each conversion j and class $c \in \{\text{Exp.}, \text{Imp.}\}$, we compute the precision $P_{c,j}$, the recall $R_{c,j}$, and the F1 score $F1_{c,j} = 2P_{c,j}R_{c,j}/(P_{c,j} + R_{c,j})$ from the touchpoints in that conversion. Table 1 reports the averaged $\overline{P}_c$, $\overline{R}_c$, and $\overline{F1}_c$. The reported Overall F1 is computed as $(\overline{F1}_{\text{Exp.}} + \overline{F1}_{\text{Imp.}})/2$.

4.2 Main Results and Key Findings

We present the main results in Table 1 and make four key findings.

Finding 1: Implicitly-related touchpoints are much harder to identify than explicitly-related ones. Across all configurations, explicit F1 is consistently higher than implicit F1. Pairwise prompting reaches a high average explicit F1 at 0.8979, but only achieves 0.4929 implicit F1; listwise prompting averages 0.8372 explicit F1 and 0.4608 implicit F1. The best implicit F1 is 0.5426, far below the best explicit F1 of 0.9095, indicating that semantic relations such as functional complementarity and scenario consistency remain difficult even for strong LLMs. Considering that the expected F1 score for random guessing is approximately 0.1491 (based on the prevalence of implicit labels in SILVA), *the performance of current LLMs is indeed significant, yet a substantial gap remains toward human-level understanding.*

Finding 2: Pairwise prompting beats listwise prompting. Pairwise prompting outperforms listwise prompting on every model in Table 1. Averaged over all models, pairwise prompting raises the overall F1 from 0.6490 to 0.6954. Pairwise prompting likely benefits from decomposing a noisy sequence-level selection problem into candidate-level judgments, while listwise prompting tends to over-filter touchpoints, resulting in a significantly lower recall.

Finding 3: Strong open-weight models are competitive with proprietary counterparts, but smaller open-weight variants lag behind. GLM-5.1 and DeepSeek-V4-Pro reach 0.7086 and

Table 1: Main results of LLM-based touchpoint selection on SILVA. Exp. and Imp. denote explicit and implicit touchpoints. Shaded columns mark F1 metrics, with bold and underline indicating the best and second-best values in each F1 column.

Model	Listwise Prompting							Pairwise Prompting						
	Exp.			Imp.			Overall	Exp.			Imp.			Overall
	P	R	F1	P	R	F1	F1	P	R	F1	P	R	F1	F1
Closed-source large models														
GPT-5.5	0.9180	0.9418	**0.9070**	0.5161	0.4763	0.4732	0.6901	0.8747	0.9751	0.8968	0.6454	0.5567	**0.5426**	**0.7197**
Gemini-3.1-Pro	0.9312	0.9275	0.9053	0.5612	0.5201	**0.5077**	**0.7065**	0.8953	0.9686	**0.9095**	0.6246	0.5398	0.5293	0.7194
Claude-Opus-4.7	0.9151	0.9374	0.9002	0.4698	0.4664	0.4491	0.6747	0.9261	0.9147	0.8975	0.5707	0.6210	0.4979	0.6977
Open-weight large models														
GLM-5.1	0.9074	0.8186	0.8148	0.5748	0.5050	0.4668	0.6408	0.9116	0.9458	0.9092	0.6108	0.5576	0.5080	0.7086
DeepSeek-V4-Pro	0.9044	0.8480	0.8448	0.5472	0.5577	0.4990	0.6719	0.9115	0.9250	0.8956	0.6292	0.5474	0.5155	0.7056
DeepSeek-V4-Flash	0.9117	0.8396	0.8415	0.5185	0.4948	0.4694	0.6554	0.9249	0.9109	0.8975	0.5693	0.4761	0.4586	0.6780
Open-weight smaller models														
Qwen3.5-27B	0.9292	0.7636	0.7804	0.4772	0.4436	0.4232	0.6018	0.9345	0.8875	0.8852	0.4969	0.4541	0.4420	0.6636
Qwen3.5-30B-A3B	0.9317	0.6608	0.7034	0.4753	0.4437	0.3983	0.5509	0.9013	0.9327	0.8915	0.5325	0.4713	0.4491	0.6703

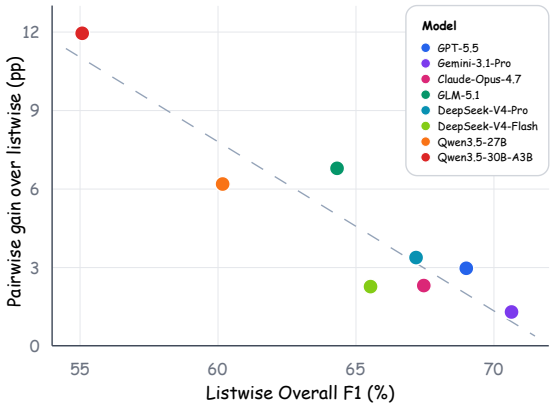


Figure 4: The performance gap between pairwise and listwise prompting narrows when the listwise overall F1 is higher.

0.7056 in terms of overall F1 under pairwise prompting, respectively, close to the best proprietary model GPT-5.5 at 0.7197. In contrast, smaller Qwen variants lag behind flagship models, especially on the implicitly-related touchpoints. This performance gap underscores that capturing latent semantic associations for touchpoint selection necessitates the superior reasoning power of large flagship models.

Finding 4: Model capability narrows the performance gap between pairwise and listwise prompting strategies. The gap is 1.3 percentage points (**pp** for short) for Gemini-3.1-Pro and 3.0 pp for GPT-5.5, but grows to 6.2 pp for Qwen3.5-27B and 11.9 pp for Qwen3.5-30B-A3B. This suggests that stronger models handle listwise comparison better, whereas smaller models benefit more from pairwise decomposition. Figure 4 further shows that the pairwise gain is larger when listwise Overall F1 is lower. This trend indicates that pairwise prompting is especially useful when a model struggles to compare many touchpoints in a single listwise response.

Table 2: Ablation of thinking mode for LLM-based touchpoint selection. We report the F1 gain of `think` over `nothink`.

Prompt	Model	ΔExp.	ΔImp.	ΔOverall
Listwise	GPT-5.5	+0.0165	+0.0296	+0.0231
	Gemini-3.1-Pro	+0.0089	+0.0089	+0.0089
	DeepSeek-V4-Pro	+0.0036	+0.0016	+0.0026
	DeepSeek-V4-Flash	-0.0012	+0.0063	+0.0025
	Qwen3.5-27B	+0.0826	+0.0239	+0.0533
	Qwen3.5-30B-A3B	+0.0452	+0.0094	+0.0274
	Avg.	+0.0259	+0.0133	+0.0196
Pairwise	GPT-5.5	-0.0010	+0.0014	+0.0002
	Gemini-3.1-Pro	+0.0010	+0.0157	+0.0084
	DeepSeek-V4-Pro	+0.0033	+0.0156	+0.0095
	DeepSeek-V4-Flash	-0.0014	+0.0140	+0.0062
	Qwen3.5-27B	+0.0028	+0.0022	+0.0025
	Qwen3.5-30B-A3B	+0.0313	+0.0130	+0.0222
	Avg.	+0.0060	+0.0103	+0.0082

4.3 Ablation Study

We further analyze the impact of chain-of-thought reasoning by switching the reasoning mode from `think` to `nothink` and comparing the resulting metrics in Table 2. We find that the *thinking mode is generally helpful, but its effect varies across models and protocols.* On average, it brings a larger Overall F1 gain under listwise prompting (+0.0196) than under pairwise prompting (+0.0082), while the strongest listwise gains come from the smaller open-weight Qwen variants. Overall, the thinking mode provides only marginal improvements, suggesting that current reasoning patterns are not fully optimized for the complexities of touchpoint selection. This indicates a critical need for future research to develop specialized fine-tuning or reasoning strategies specifically adapted to the nuanced semantic judgments required for the touchpoint selection task.

Table 3: Comparison of different touchpoint selection methods for CVR prediction. Our LLM-based selection captures incremental semantic signals beyond traditional heuristics.

Method	In-Shop	Heuristic	LLM-Based	ΔGAUC(pp)
Base [4]	✓	✗	✗	-
CABB [23]	✓	✓	✗	+0.20
LOTUS (Ours)	✓	✓	✓	**+0.35**

5 Application: Enhancing CVR Prediction

To validate the potential utility of LLM-based touchpoint selection, we conduct an experiment using a production-scale CVR prediction model on our e-commerce platform. Following Chen et al. [4] and Zeng et al. [23], we incorporate the selected touchpoints as positive samples for **an auxiliary indirect conversion prediction task**. Leveraging the insights from our benchmark evaluation in § 4, we employ our internal foundation model with pairwise reasoning to perform touchpoint selection at scale.

As shown in Table 3, our approach, **L**LM-**O**riented **T**ouchpoint **U**nderstanding and **S**election (**LOTUS**), substantially outperforms both the in-shop attribution baseline Base [4] and the heuristic rule-based selection method CABB [23], which covers only explicit cross-shop touchpoints, achieving absolute offline Group AUC (GAUC) gains of **0.35 pp and 0.15 pp** [4, 9, 13, 14, 16] for the main conversion goal over Base and CABB, respectively. The improvement over Base demonstrates the value of incorporating cross-shop touchpoints, while the additional gain over CABB suggests that LLM-based semantic reasoning can identify useful implicit relations beyond those captured by explicit heuristic rules.

Considering that an absolute GAUC improvement of 0.1 pp is empirically sufficient to drive substantial growth in online sales and advertising revenue [3, 4, 24], these results validate the practical utility of leveraging LLMs to discover latent, high-value touchpoints that traditional methods overlook.

6 Conclusion

In this study, we identify a critical **semantic gap** in conversion attribution, where traditional heuristic methods overlook a vast space of **implicitly-related touchpoints**. Through our newly established **SILVA** benchmark, we find that implicitly-related and explicitly-related touchpoints account for 14.91% and 19.09% of all annotated touchpoints, respectively. Our systematic evaluation shows that LLMs can effectively bridge this gap by reasoning over product metadata and user-behavior context. Detailed analysis further reveals that pairwise reasoning improves performance and that larger foundation models are better able to capture subtle semantic nuances. Furthermore, integrating these LLM-identified touchpoints into a production-scale CVR prediction model improves GAUC by an absolute margin of **0.35 pp** over the production baseline [4], which considers only in-shop touchpoints, and by **0.15 pp** over the heuristic CABB competitor [23]. Overall, our study highlights the promising potential of LLMs in refining conversion attribution, providing a useful roadmap for future research on conversion attribution and conversion rate prediction in the LLM era.

Acknowledgments

This work is funded by CCF-ALIMAMA TECH Kangaroo Fund (No. CCF-ALIMAMA of 2025004).

GenAI Usage Disclosure

Generative AI tools were only used for limited text polishing during the draft preparation.

References

[1] Anthropic. 2026. Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7

[2] John Bencina, Erkut Aykutlug, Yue Chen, Zerui Zhang, Stephanie Sorenson, Shao Tang, and Changshuai Wei. 2025. LiDDA: Data Driven Attribution at LinkedIn. arXiv:2505.09861 [cs.LG] https://arxiv.org/abs/2505.09861

[3] Zhangming Chan, Yu Zhang, Shuguang Han, Yong Bai, Xiang-Rong Sheng, Siyuan Lou, Jiacen Hu, Baolin Liu, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. Capturing Conversion Rate Fluctuation during Sales Promotions: A Novel Historical Data Reuse Approach. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 3774–3784. doi:10.1145/3580305.3599788

[4] Sishuo Chen, Zhangming Chan, Xiang-Rong Sheng, Lei Zhang, Sheng Chen, Chenghuan Hou, Han Zhu, Jian Xu, and Bo Zheng. 2025. See Beyond a Single View: Multi-Attribution Learning Leads to Better Conversion Rate Prediction. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (Seoul, Republic of Korea) *(CIKM '25)*. Association for Computing Machinery, New York, NY, USA, 5600–5608. doi:10.1145/3746252.3761580

[5] DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro

[6] Ruihuan Du, Yu Zhong, Harikesh Nair, Bo Cui, and Ruyang Shou. 2019. Causally Driven Incremental Multi Touch Attribution Using a Recurrent Neural Network. arXiv:1902.00215 [cs.LG] https://arxiv.org/abs/1902.00215

[7] GLM-5-Team. 2026. GLM-5: From Vibe Coding to Agentic Engineering. arXiv:2602.15763 [cs.LG] https://arxiv.org/abs/2602.15763

[8] Google DeepMind. 2026. Gemini 3.1 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/

[9] Jiacen Hu, Zhangming Chan, Yu Zhang, Shuguang Han, Siyuan Lou, Baolin Liu, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. PS-SA: An Efficient Self-Attention via Progressive Sampling for User Behavior Sequence Modeling. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 4639–4645.

[10] Sachin Kumar, Garima Gupta, Ranjitha Prasad, Arnab Chatterjee, Lovekesh Vig, and Gautam Shroff. 2020. CAMTA: Causal Attention Model for Multi-touch Attribution. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 79–86. doi:10.1109/ICDMW51313.2020.00020

[11] Randall Lewis, Florian Zettelmeyer, Brett R. Gordon, Cristobal Garib, Johannes Hermle, Mike Perry, Henrique Romero, and German Schnaidt. 2025. Amazon Ads Multi-Touch Attribution. arXiv:2508.08209 [econ.EM]

[12] Ning Li, Sai Kumar Arava, Chen Dong, Zhenyu Yan, and Abhishek Pani. 2018. Deep Neural Net with Attention for Multi-channel Multi-touch Attribution. arXiv:1809.02230 [cs.LG] https://arxiv.org/abs/1809.02230

[13] Xinyu Li, Sishuo Chen, Guipeng Xv, Li Zhang, Mingxuan Luo, Zhangming Chan, Xiang-Rong Sheng, Han Zhu, Jian Xu, and Chen Lin. 2026. Delayed Feedback Modeling for Post-Click Gross Merchandise Volume Prediction: Benchmark, Insights and Approaches. In *Proceedings of the ACM Web Conference 2026*. Association for Computing Machinery, 6299–6308. doi:10.1145/3774904.3792348

[14] Mingyang Liu, Yong Bai, Zhangming Chan, Sishuo Chen, Xiang-Rong Sheng, Han Zhu, Jian Xu, and Xinyang Chen. 2026. EST: Towards Efficient Scaling Laws in Click-Through Rate Prediction via Unified Modeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Republic of Korea) *(KDD '26)*. Association for Computing Machinery, New York, NY, USA, 7656–7666. doi:10.1145/3770855.3818400

[15] Yuguang Liu, Luyao Xia, Hu Liu, Zhangxi Yan, Jian Liang, Han Li, and Kun Gai. 2026. ALM-MTA: Front-Door Causal Multi-Touch Attribution Method for Creator-Ecosystem Optimization. In *The Fourteenth International Conference on Learning Representations*.

[16] Mingxuan Luo, Guipeng Xv, Sishuo Chen, Xinyu Li, Li Zhang, Zhangming Chan, Xiang-Rong Sheng, Han Zhu, Jian Xu, Bo Zheng, and Chen Lin. 2026. Modeling Cascaded Delay Feedback for Online Net Conversion Rate Prediction: Benchmark, Insights and Solutions. In *Proceedings of the ACM Web Conference 2026*. Association for Computing Machinery, 6252–6263. doi:10.1145/3774904.3792327

[17] OpenAI. 2026. GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/

[18] Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5

[19] Xuhui Shao and Lexin Li. 2011. Data-Driven Multi-Touch Attribution Models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. Association for Computing Machinery, 258–264. doi:10.1145/2020408.2020453

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 24824–24837. doi:10.52202/068431-1800

[21] Jinqi Wu, Sishuo Chen, Zhangming Chan, Yong Bai, Lei Zhang, Sheng Chen, Chenghuan Hou, Xiang-Rong Sheng, Han Zhu, Jian Xu, Bo Zheng, and Chaoyou Fu. 2026. MAC: A Conversion Rate Prediction Benchmark Featuring Labels Under Multiple Attribution Mechanisms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Republic of Korea) *(KDD '26)*. Association for Computing Machinery, New York, NY, USA, 10009–10019. doi:10.1145/3770855.3817488

[22] Di Yao, Chang Gong, Lei Zhang, Sheng Chen, and Jingping Bi. 2022. CausalMTA: Eliminating the User Confounding Bias for Causal Multi-touch Attribution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 4342–4352. doi:10.1145/3534678.3539108

[23] Xiangyu Zeng, Amit Jaspal, Bin Liu, Goutham Panneeru, Kevin Huang, Nicolas Bievre, Mohit Jaggi, Prathap Maniraju, and Ankur Jain. 2026. Click A, Buy B: Rethinking Conversion Attribution in E-Commerce Recommendations. In *Computational Advertising (Communications in Computer and Information Science)*. Springer Nature Switzerland, 113–124. doi:10.1007/978-3-032-16358-5_9

[24] Yujing Zhang, Zhangming Chan, Shuhao Xu, Weijie Bian, Shuguang Han, Hongbo Deng, and Bo Zheng. 2022. KEEP: An Industrial Pre-Training Framework for Online Recommendation via Knowledge Extraction and Plugging. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 3684–3693. doi:10.1145/3511808.3557106